Silesian University of Technology

SCIENTIFIC DISCIPLINE CIVIL ENGINEERING, GEODESY AND TRANSPORTATION

PHD DISSERTATION

Forecasting prices for road freight transport services using machine learning

MSc Artur Budzyński

Discipline: Civil Engineering, Geodesy and Transport

PhD Supervisor
Professor Aleksander Sładkowski, DSc.

PhD Assistant Supervisor
Associate Professor Marcin Michalak, DSc. Eng.

Katowice, 2025

1. INTRODUCTION

1.1. Significance of Road Freight Transport

Road freight transport is a cornerstone of global logistics, driving economic growth, fostering social development, and ensuring political stability through its unparalleled door-to-door accessibility. As the primary mode of transporting goods over short to medium distances, road freight transport is essential for maintaining supply chain continuity and adapting to changing market demands.

Regional studies provide valuable insights into the dynamics of this sector. A statistical analysis of Croatia's road freight transport system from 2002 to 2012 revealed significant shifts in transport demand and operational efficiency, reflecting the evolving nature of logistics in this region [1]. Similarly, another study highlighted the impact of these trends on freight transport operations, emphasizing the sector's adaptability to changing economic and logistical conditions [2].

Service quality in road freight transport is shaped by multiple factors. Research on Lithuanian carriers underscored the critical role of safety, specifically cargo loading control, as the most important determinant of service quality [3]. In Austria, the practice of flagging out (i.e. registering vehicles abroad to reduce operational costs) emerged as a dominant strategy for cost management between 2003 and 2012, despite governmental efforts to limit this practice [4]. These examples underscore the importance of balancing cost efficiency with service quality to sustain competitive advantage in the sector.

Innovative strategies have also proven effective in addressing logistical challenges. For instance, in Pakistan, the integration of intermodal freight transport reduced operational costs by over 60% compared to conventional road freight methods, as confirmed by quantitative analyses from leading industry stakeholders [5]. These findings were further validated by a case study involving a prominent paper and board business, highlighting the economic benefits of intermodal strategies in freight transport management [6].

Transport exchanges have emerged as a promising solution for managing road freight operations. These platforms facilitate efficient coordination between shippers and carriers, helping to reduce empty vehicle runs and optimize fleet utilization. A comparative analysis highlighted the effectiveness of transport exchanges in improving operational efficiency and reducing costs [7].

Technology adoption continues to transform road freight transport. Electronic consignment notes have been shown to enhance both operational efficiency and sustainability in logistics processes [8]. Furthermore, administrative data from weigh-in-motion systems in Australia revealed a significant rise in road freight activities along key interoccipital routes, driven by the widespread adoption of B-double trucks [9]. The integration of Industry 4.0 technologies, including advanced business strategies and smart operations, has also been identified as a critical driver of service quality improvement in road freight logistics [10].

According to [11], transport exchange users benefit from utilizing transport exchanges, as doing so reduces empty vehicle runs. The authors of a previous study [12] introduced the strategy of leveraging dependencies between freight prices and economic factors, such as GDP, to enhance the efficiency of road transport management.

Collectively, these findings underscore the importance of road freight transport in contemporary logistics and highlight the ongoing advancements and strategies that enhance its

efficiency and quality. This foundation sets the stage for exploring the role of machine learning in transport in the subsequent section.

1.2. Applications of Machine Learning in Transport Systems

Machine learning has emerged as a transformative technology in transportation systems that provides innovative solutions to improve efficiency, sustainability, and overall performance. Its applications span multiple domains, ranging from operational optimization to intelligent transportation systems (ITSs) and infrastructure monitoring.

Machine learning techniques have been used to enhance data analytics in transportation. For example, an introductory study [13] demonstrated that MATLAB can be used to predict roadway speed, showcasing the ability of machine learning to analyze large datasets and improve system performance. Another innovative application involved a pair-wise attention-based pointer neural network, which optimized last-mile delivery routes by predicting driver stop sequences based on historical data, reducing route disparities by 15% [14].

Machine learning also plays a crucial role in extracting knowledge from complex systems. A previous study [15] proposed an eight-stage process for urban rail control, leveraging machine learning to derive decision rules from performance attributes such as travel time, energy consumption, and passenger comfort. This methodology enhances the understanding of train motion and stopping regimes, providing valuable insights for urban rail systems.

Despite its advantages, machine learning in transportation faces challenges related to fairness and bias. Studies have shown that both deep neural networks (DNNs) and discrete choice models (DCMs) can exhibit prediction disparities across social groups, potentially reinforcing inequalities [16]. To address this issue, researchers have proposed absolute correlation regularization, a method that reduces biases in both synthetic and real-world datasets. This approach enables models to balance prediction accuracy with fairness, ensuring that machine learning applications promote equity in transport systems.

Hybrid approaches that combine interpretability and predictive power have been developed to overcome the limitations of traditional models. The Theory-Based Residual Neural Network (TB-ResNet) framework integrates utility specifications from DCMs with the flexibility of DNNs, enhancing both prediction accuracy and robustness [17]. This synergy allows for the more effective modelling of complex transportation behaviours while maintaining interpretability—a critical feature for practical implementation in transport planning and policy-making.

Machine learning has transformed ITSs by improving traffic management, enhancing safety, and optimizing operations. Context-aware machine learning techniques have proven particularly effective in enhancing traffic prediction capabilities [18]. These methods incorporate real-time contextual data to improve decision-making and operational efficiency. Machine learning methods have been extensively applied to address critical transportation challenges, such as truck fuel consumption, demand forecasting, and price prediction in road freight management. These applications leverage Python-based tools and libraries, emphasizing data preprocessing, model training, validation, and real-world implementation, as detailed in a previous study [19].

An integrated machine learning framework for urban transportation optimization in smart cities combines advanced algorithms, such as hybrid ANN–RNN techniques, to achieve adaptability and sustainability [20]. This approach utilizes a multilayer objective function that

considers interaction costs, energy consumption, and environmental impact, contributing to the development of resilient and sustainable ITSs.

In structural health monitoring, machine learning enables the analysis of sensor data to detect structural changes in transportation infrastructure, such as bridges and tunnels. However, a review [21] highlighted the limited use of these techniques in condition assessment, indicating a need for further research to develop robust methods.

Emerging trends in machine learning also highlight the integration of citizen science with urban transportation systems. By processing data collected from urban communities, researchers have proposed models to enhance system responsiveness to demographic changes [22]. Additionally, machine learning in maritime transportation has shown the potential to optimize voyages and forecast maintenance needs despite challenges related to limited data availability [23].

However, the integration of machine learning requires the management of practical challenges (e.g. ensuring data quality, managing computational complexity, and developing frameworks that prioritize ethical considerations). A comprehensive review [24] emphasized the importance of balancing technological advancements with practical implementations to maximize the benefits of machine learning in transport systems.

1.3. Price Forecasting in Transport

Accurate price forecasting is essential for the transportation industry, as it directly influences strategic planning, budgeting, and operational efficiency. By anticipating future price trends, transportation companies can make informed decisions regarding pricing strategies, cost management, and resource allocation. In reviewing various methodologies and models used in price forecasting in the transport sector, this section highlights the integration of econometric techniques and machine learning approaches to enhance prediction accuracy and reliability.

Econometric methods have long been employed to analyze price trends and changes in transportation. A notable example is the development and validation of a three-layer mathematical model for the Freight Price Index on Highways [25], which enhances market supervision by accurately reflecting price trends and revealing discrepancies between supply and demand. This model provides transportation stakeholders with actionable insights into market dynamics.

Seasonal adjustment methods and Hodrick-Prescott filters have also been applied to analyze road freight pricing in China, identifying fluctuation patterns across vehicle types, haul distances, and routes [26]. These methods provide recommendations for integrating road freight pricing strategies with rail freight markets, offering a holistic view of intermodal transportation pricing.

Another study using a vector error correction GARCH-in-mean model demonstrated the significant impact of Mississippi River barge freight price volatility on grain prices and marketing margins [27]. The research underscored the importance of mitigating barge rate risks to stabilize international grain markets and reduce marketing margins, offering valuable insights for pricing strategies in multi-modal transport chains.

Behavioural modelling in price forecasting has also gained traction, with researchers employing a bi-level programming model based on cumulative prospect theory to analyze the behaviour of customers in freight markets [28]. This approach revealed that incorporating customer utility and network interactions enhances equilibrium freight prices and company profits to a greater extent than traditional logit-based models.

Empirical studies further highlight the structural factors influencing freight prices. An analysis of the French ECHO survey demonstrated that road freight pricing is shaped not only by technical attributes, such as load weight, travel duration, and distance, but also by non-traditional factors, such as shipper-operator relationships and company size [29]. These findings emphasize the multifaceted nature of freight pricing, necessitating comprehensive modelling frameworks.

Machine learning techniques are increasingly adopted to improve price prediction accuracy and reliability in transportation. For example, an early warning index system for railroad bulk freight price risk was developed by integrating macroeconomic, transportation market, freight owner, and enterprise-specific factors [30]. Using methods such as entropy weight, TOPSIS, k-means clustering, and machine learning models like BP neural networks and LSTM, the study found that LSTM networks significantly outperformed traditional models in predicting freight price risks. This advancement aids railway departments in formulating adaptive pricing strategies and mitigating potential market disruptions.

The integration of econometric and machine learning methodologies shows promise for advancing price forecasting in the transportation sector. Econometric approaches provide robust theoretical frameworks for understanding price volatility and its impact on market segments, while machine learning models excel in handling complex, high-dimensional data and generating actionable predictions.

Future research should prioritize the development of hybrid models that leverage the complementary strengths of these methodologies. Such models could integrate real-time data streams, enabling dynamic adaptability to market conditions and improving forecasting precision. Additionally, exploring the role of big data and advanced analytics in enriching price forecasting frameworks could further enhance the competitiveness and stability of the transportation industry.

Recent advancements in price forecasting methodologies underscore the critical roles of econometric and machine learning models in the transportation industry. By combining the interpretive strength of econometrics with the computational power of machine learning, stakeholders can achieve a deeper understanding of price dynamics, enabling more informed decision-making and strategic planning. These innovations not only improve operational efficiency but also strengthen the industry's resilience to market fluctuations, ensuring long-term sustainability and competitiveness.

2. STUDY CONCEPT

This study aims to address the significant gaps in the literature and practice of forecasting road freight transport service prices by developing a comprehensive machine learning methodology. It focuses on creating a robust framework that includes data collection, preprocessing, model selection, training, validation, and deployment tailored to the specific needs of the road freight transport sector. By systematically evaluating and selecting appropriate machine learning algorithms, this research seeks to enhance the accuracy and reliability of price predictions. Additionally, this study emphasizes the integration of various models into existing business processes to improve decision-making and operational efficiency. This approach provides practical benefits for transportation companies by optimizing resource allocation and adapting to dynamic market conditions.

2.1. Justification

This study is justified by the growing complexity of price forecasting in the road freight transport sector. Current forecasting methods, often based on intuition or basic statistical models, fail to address the dynamic and multifaceted nature of the market, including fluctuating fuel prices, regulatory changes, and logistical constraints.

This research introduces machine learning techniques tailored to the unique challenges of this sector, thus enabling more accurate and reliable price predictions. By integrating advanced methodologies, this study provides tools that carriers and logistics companies can use to enhance operational efficiency, optimize resource allocation, and make informed strategic decisions.

Ultimately, this study addresses practical needs within the industry by bridging the gap between traditional approaches and the potential of data-driven, machine learning-based solutions.

2.2. Research Gap

A review of the literature indicates a significant gap concerning the development of a comprehensive methodology for building machine learning models to forecast prices in road freight transport services, particularly within the European Union (EU). Although machine learning techniques have been widely adopted across various sectors, current studies often overlook the distinct data challenges found in spot transactions for full truckload (FTL) services—where short-term contracts, rapid market fluctuations, and route-specific supply-demand factors substantially influence real-time pricing.

Moreover, a notable deficiency exists in examples of fully replicable experimental setups, including open-source code and standardized datasets. This lack of transparency hinders the reproducibility of findings and the broader advancement of research, as other investigators are unable to verify or refine existing models effectively.

Altogether, these gaps underscore the pressing need for a structured, domain-specific framework that guides the entire model lifecycle, from data collection and preprocessing to validation and deployment, while ensuring that the work is repeatable and verifiable. Addressing these issues could drive innovation in designing robust, scalable models that deliver accurate, real-time price forecasts for the road freight transport sector in the EU. Such advancements would yield tangible benefits for transport businesses, offering more precise

pricing insights, improved operational efficiency, and a solid foundation for ongoing research in a rapidly evolving market.

2.3. Research Problem

It is challenging to forecast road freight transport prices due to the large number of interdependent variables and the time-consuming nature of the forecasting process.

The complexity of data, which stems from the diversity of transport characteristics such as cargo type, organizational requirements, transport relations, and seasonality, increases the risk of modelling errors and limits the effectiveness of traditional methods. Additionally, the time-intensive nature of current forecasting practices diverts operational staff from other tasks. This study aims to address these issues by developing a methodology that leverages machine learning to efficiently handle complex dependencies and optimize the forecasting process.

2.4. Research Question

This research aims to answer the following question:

I. **What factors influence the pricing of road freight transport services, how can a methodology for processing transport offer data be developed to effectively train machine learning models, and can the application of these methods achieve greater forecasting accuracy compared to expert-based approaches?**

This research question integrates three critical aspects: identifying key factors influencing road freight transport pricing, developing a methodology for processing transport offer data to effectively train machine learning models, and assessing whether these methods can achieve greater forecasting accuracy than expert-based approaches.

2.5. Objective

The primary objective of this study is to develop a robust methodology for forecasting road freight transport service prices using machine learning techniques. The study involves creating a systematic framework that encompasses data collection, preprocessing, model training, validation, and implementation. Special emphasis is placed on addressing issues related to incomplete, diverse, and variable data in transport pricing. The proposed techniques will include methods for cleaning, normalization, and feature engineering to enhance model performance.

This research systematically evaluates and compares various machine learning algorithms, such as gradient boosting, random forests, and neural networks, to identify the most effective models for price forecasting. A key objective is to ensure that the developed methodology can be seamlessly integrated into existing business processes in the road freight transport industry, supporting decision-making and operational optimization.

The methodology is designed to adapt dynamically to changing market conditions, including real-time model updates and scalability across different segments of the transport market. Empirical validation is carried out using real-world datasets to ensure that the forecasts are accurate and practically applicable.

The results of this study are expected to contribute to the academic understanding of machine learning applications in transport economics and provide a practical tool to enhance pricing strategies and operational efficiency within the transport industry.

2.6. Hypothesis

Based on a comprehensive review of the literature, the following hypothesis has been proposed:

Implementing machine learning techniques to forecast road transport service prices will improve forecast accuracy and enable transport companies to adjust their operational decisions in response to changing conditions.

Machine learning makes price forecasting more precise through the analysis of complex patterns in large and diverse datasets, surpassing the limitations of traditional methods. These techniques allow models to be updated in real time as new data becomes available, enabling continuous adjustment of pricing and operational strategies to reflect current market dynamics.

Accurate forecasts support decision-making in critical areas such as route planning, resource allocation, and contract negotiations, leading to improved operational efficiency and cost optimization. Additionally, machine learning facilitates the identification of key market trends and potential risks, helping companies anticipate and respond to changes in the business environment.

By integrating machine learning-based forecasting tools with transport management systems, companies can automate processes, enhance consistency, and reduce their reliance on manual operations. This approach allows transport businesses to manage operations more flexibly and effectively so they can adapt to volatile market conditions and maintain a competitive edge.

2.7. Expected Benefits

The development of machine learning-based forecasting models for road freight transport service prices is anticipated to bring numerous benefits to transportation companies, freight forwarders, customers, and the academic community, as well as future specialists in the field of transport organization and management.

Machine learning models offer improved precision and accuracy in price forecasting. Traditional methods often rely on human expert judgment, which may be insufficient when dealing with complex and dynamic datasets. Incorporating a greater number of variables and analyzing intricate patterns enables machine learning to offer more accurate forecasts and better responsiveness to changes in the market.

Improved price forecasting enhances the operational efficiency of transportation companies since accurate predictions allow for better resource planning, cost optimization, and the adjustment of offerings to meet demand. As a result, companies can make more informed business decisions, boosting their competitiveness and profitability.

The application of machine learning in price forecasting also supports improved risk management in the transportation industry. By analyzing a variety of factors and variables, machine learning models help identify risks and predict potential market changes. This empowers companies to refine their pricing strategies, minimize risks, and take proactive actions in a timely manner.

Furthermore, the development and implementation of machine learning-based forecasting models can contribute to the advancement of the transport and logistics sector. Research in this area has the potential to foster innovations that lead to the creation of tools and methodologies that can be applied across various transportation sectors. The findings of this research can serve as a foundation for educating future specialists in the organization and management of road freight transport. This knowledge could lead to the development of modern tools and methods, thus equipping future professionals with the skills needed to operate effectively in the dynamic transport industry. Furthermore, integrating machine learning into educational programs can enhance the competencies of professionals and their ability to implement innovative solutions in practice.

The results of this research are expected to improve operational efficiency, enhance competitiveness, and contribute to the development of the transportation industry while supporting the growth of future leaders in this field.

3. METHODOLOGY AND TECHNIQUES

This chapter presents the theoretical basis for the model developed for forecasting the prices of road freight transport services using machine learning. It discusses the key theories and concepts that influenced the construction and functioning of the predictive model.

3.1. Theoretical Assumptions

This subsection discusses key theoretical assumptions related to forecasting prices in the transport of goods by road freight. The theoretical foundations of supply and demand theory and their impact on the price of transport services are presented in the context of modelling prices using machine learning techniques.

The concept underlying this work is based on the machine learning model lifecycle depicted in Fig. 1. The figure illustrates the sequence of key stages that constitute the processes of creating and deploying a predictive model of prices within road freight transport services using machine learning techniques.

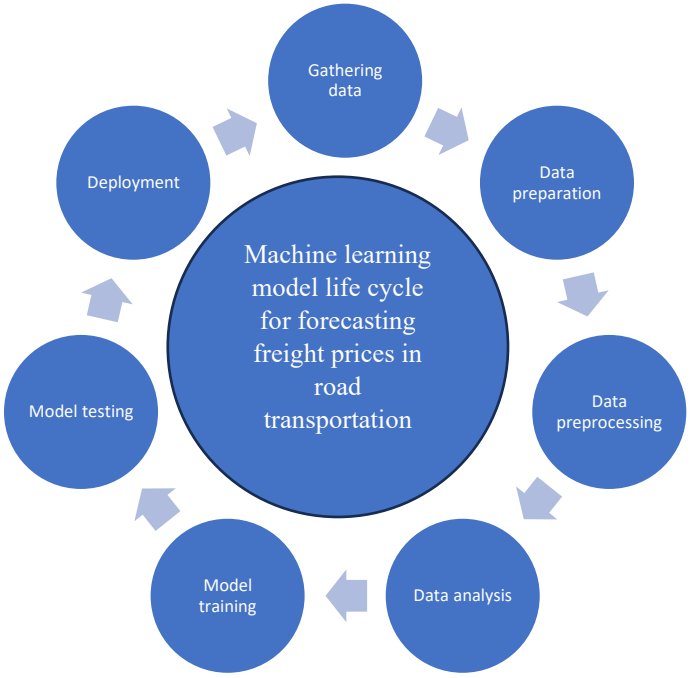


Fig. 1. Lifecycle of a machine learning model for predicting the price of a road freight transport service. Source: The author's own work

In the initial phase of the cycle, data collection involves gathering relevant information concerning prices of road freight transport services and variables related to them. We then proceed to data preparation, during which the data undergoes cleansing, error elimination, and the removal of redundant observations (duplicate shipments from different sources) to ensure data consistency and quality.

Data transformation involves converting and shaping the data into a format suitable for machine learning modelling. After data transformation, we move on to data analysis, where exploratory data analysis techniques are applied to understand the relationships between variables and detect potential patterns.

After appropriately preparing the data, we proceed to model training. In this stage, we select the appropriate machine learning algorithm and utilize training data to teach the model

to predict prices based on available information. After completing the model training process, we move on to model testing, during which the model is evaluated based on test data that were not used during the training process.

After successful testing, deployment takes place, during which the model is implemented in a production environment. This stage considers various aspects such as scalability, performance, and model accessibility to users.

This cycle of creating a predictive model for forecasting prices for road freight transport services using machine learning is iterative, meaning that after the model is deployed, its results and forecast quality are monitored, and the model may be adjusted to changing conditions and data. The entire process aims to deliver precise and effective forecasting tools in the context of road freight transport services.

3.2. Tools and Libraries Used for Modelling

This subsection focuses on the crucial stage of selecting the appropriate machine learning techniques for price forecasting in road freight transport services. The choice of these techniques is of paramount importance for the effectiveness of the forecasts and the quality of the results achieved.

Python has emerged as an exceptionally suitable choice for forecasting the price of road freight transport services due to its versatile and advanced tools for data analysis and machine learning. Python offers a wide range of libraries and frameworks dedicated to these domains, such as NumPy, pandas, scikit-learn, TensorFlow, and PyTorch. These tools enable efficient data processing and analysis, as well as the creation of advanced forecasting models.

The Python language is characterized by its simplicity and readability of syntax, facilitating efficient work and the rapid prototyping of models. This feature is particularly relevant in complex price analysis scenarios in road freight transport services, where code readability translates into an easier understanding of processes.

Moreover, Python's immense popularity across scientific and business communities ensures access to abundant online resources, including documentation, tutorials, and forums. This accessibility significantly aids learning and problem-solving, offering a substantial advantage in implementing forecasting models in road freight transport services.

For instance, the graph in Fig. 2 depicts the number of queries related to the most popular programming languages on Stack Overflow. Python's popularity in such queries underscores its feasibility for data analysis, modeling, and machine learning. Python-related questions cover diverse topics, including image processing, deep learning, and price forecasting, demonstrating the language's versatility for addressing various research issues.

It was shown in [31] that programming microcontrollers in Python simplifies and accelerates the development of embedded systems due to the simplicity and elegance of the language. According to [32], owing to its simplicity and powerful tools, Python is the ideal language for effectively teaching programming to beginners and is currently the most desirable and fastest-growing programming language. Researchers at PyVerDetector [33], a Chrome extension that lets users check the compatibility of Python code snippets on Stack Overflow with their chosen Python version, solved the incompatibility problem between Python 2 and Python 3.

A book [34] describes six key concepts needed to learn Python 3.10, offering practical examples and advanced advice to provide a solid understanding of modern Python features for beginners and experienced developers. Researchers in [35] showed how Python, as a multi-

paradigm programming language, accelerates software development and stands out in the Big Data era with its rich data structures, standard libraries, and applications in sentiment analysis and data science, outperforming other languages in many areas.

According to these resources, Python is not only a popular choice but also a substantively justified one for forecasting the price of road freight transport services. Its versatility, ease of use, rich community, and integration capabilities make it an suitable tool for creating and deploying forecasting models in the context of this dissertation.

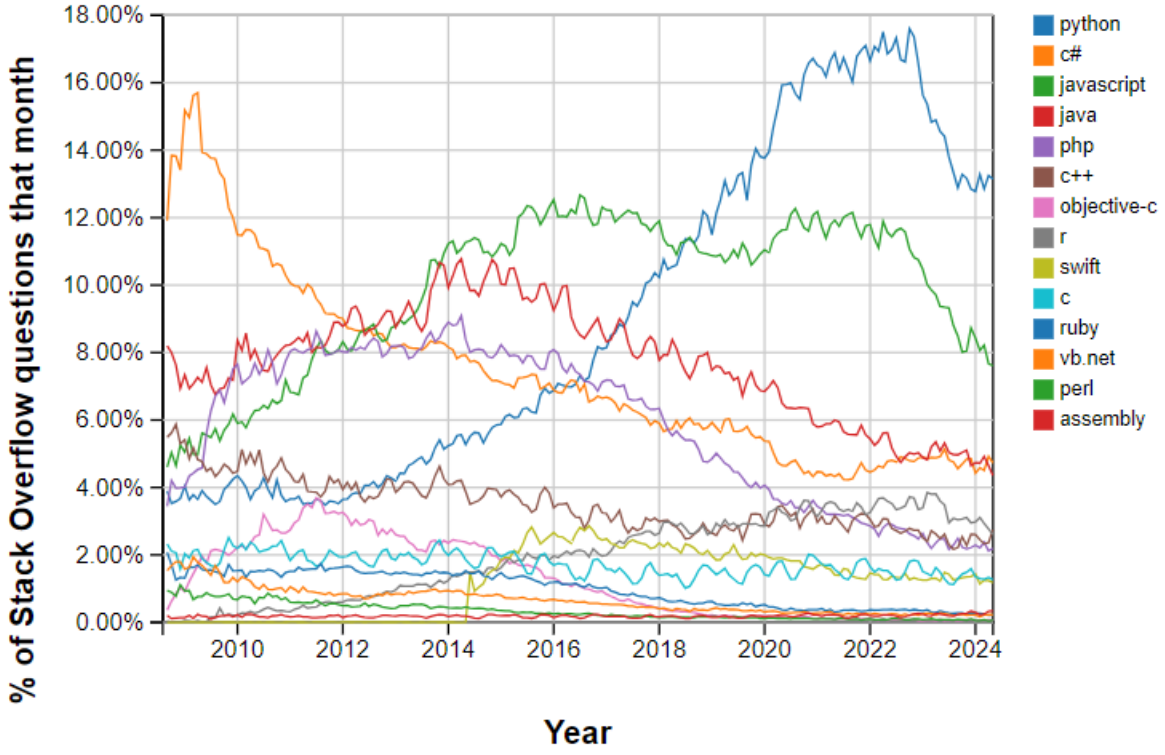


Fig. 2. The most popular programming languages on Stack Overflow [36]

According to researchers [7], with the growing demand for research in all areas related to writing computer code, publishing this code becomes essential for the repeatability of research results. Jupyter Notebooks is presented as a document format for publishing code, results, and explanations in a readable and executable form. Various tools and use cases of notebook documents are discussed. Jupyter Notebook is an exceptional choice for conducting analyses and forecasting prices for road freight transport services due to its technical advantages, interactivity, and ability to document the research process.

Jupyter Notebook allows users to create dynamic analyses that allow for interactive data exploration. This is particularly important in the analysis of prices in road freight transport services, as access to different perspectives on the data can help identify trends and deviations from the norm. Jupyter Notebook also makes it possible to combine code with result visualization, facilitating the understanding of analyses and the presentation of results. Creating interactive charts and graphs allows for better communication of price forecast results both within the research project and in the context of publication.

Jupyter Notebook allows researchers to document the research process directly and clearly. The ability to input comments, descriptions, and step-by-step analysis helps maintain transparency in the price forecasting process and enables analysis repeatability by other

researchers. Jupyter Notebook enables the easy sharing and reproduction of research. Other researchers can analyze the code, data, and analysis, allowing for result verification and further research development based on existing material.

Owing to its interactivity, visualization, documentation, and language versatility, Jupyter Notebook is an excellent tool for forecasting prices for road freight transport services. Its role in conducting research analyses allows for efficient data exploration, result visualization, and documentation of the research process, resulting in the quality and reliability of results obtained in this research field.

The pandas library stands out as a useful choice for analyzing and forecasting prices for road freight transport services due to its technical properties, functionality, and ability to process data. Pandas offers tabular data structures, such as Data Frame, which are well-suited for storing and processing data related to prices and road freight transport services. By enabling the manipulation and analysis of data in tabular form, pandas significantly facilitates data exploration and building predictive models. This process is crucial in the analysis of prices in the domain of road freight transport services, where data may contain gaps, duplicates, and other irregularities. Pandas allows for precise data preparation for modelling.

The pandas library enables the merging of different data sources and the performance of transformations, such as aggregations and grouping. In the case of price forecasting, these operations allow for the creation of consistent datasets that can be used to build effective models. Pandas offers a range of data analysis functions, such as calculating statistics, grouping, and filtering. These operations allow users to examine price dependencies and identify trends. Additionally, the ability to combine pandas with visualization tools allows for the creation of clear charts presenting price analyses.

Pandas can be effectively used in the data preparation process for modelling. The ability to transform, filter, and analyze data using pandas enables efficient preparation of datasets for training and testing predictive models. Owing to its functionalities, pandas is an indispensable tool in the analysis and forecasting of prices for road freight transport services. Its ability to process data, perform analysis operations, and prepare data for modelling makes it an ideal tool for researchers.

Researchers analyzed the practical issues of working with common datasets in finance, statistics, and related fields [38]. The pandas library facilitates work with these datasets and provides fundamental elements for implementing statistical models. Specific design issues encountered during the creation of the pandas library are discussed, along with relevant examples and comparisons with the R language. The article concludes with a discussion of possible directions for the development of statistical computations and data analysis using the Python language.

The NumPy library stands out as a strong choice for forecasting prices for road freight transport services due to its technical advantages, computational efficiency, and capabilities in processing numerical data. NumPy offers efficient and optimized numerical operations, which are crucial in price analysis and forecasting. Owing to its implementation of low-level operations, NumPy allows users to quickly process large datasets, which is essential in analyzing prices for road freight transport services. NumPy provides data structures such as Nd array, which are efficient in data management. In price forecasting, these data structures allow data to be stored and manipulated in tabular form, which is important in time series analysis and price modelling. The NumPy library offers a wide range of numerical functions, including mathematical, algebraic, and statistical operations. This allows researchers to perform various calculations, analyses, and data transformations, which are crucial in price analysis and

modelling. NumPy is often used as the foundation for creating more advanced libraries for data analysis and machine learning, such as pandas and scikit-learn. The use of these tools enables the development of more complex price forecasting models. The NumPy library allows for high-level abstraction numerical computations, resulting in clear and understandable code. This facilitates price analysis and model building, which is important for understanding and reproducing results. In the Python environment, NumPy arrays are the standard way of representing numerical data, enabling efficient implementation of numerical computations in a high-level abstraction language. As demonstrated in a previous study [39], NumPy's performance can be enhanced using three techniques: vectorization of computations, avoiding data copying in memory, and minimizing the number of operations. Owing to its technical advantages, computational efficiency, and versatility in numerical functions, the NumPy library is an excellent tool in price analysis and forecasting for road freight transport services. Its ability to process data, perform numerical operations, and integrate with other tools makes it a valuable choice for researchers in this field.

The Matplotlib library is a practical choice for forecasting prices for road freight transport services due to its visualization advantages, ability to create clear plots, and customizability. Matplotlib enables the creation of various plots and graphs, which is crucial in presenting price data in road freight transport analysis. The ability to represent data graphically allows for a quick understanding of price dependencies and trends, which is important in forecasting. Matplotlib offers advanced tools for customizing plots, allowing the creation of clear and aesthetic visualizations. In price analysis, the ability to adjust plot parameters allows users to focus on relevant information and highlight key aspects of forecasting. The Matplotlib library provides a wide range of plot types, such as line plots, scatter plots, histograms, and box plots. This enables researchers to choose the most suitable visualization type for presenting price data and gaining a fuller picture of the situation. Matplotlib is often used in combination with other data analysis tools, such as pandas or NumPy. This integration allows users to effectively present analysis results and forecasts based on data prepared using other tools. A previous article [40] presents Matplotlib as a 2D graphics package used in the Python language for creating applications and interactive scripts and generating high-quality images for publication in various user interfaces and operating systems. Owing to its ability to create clear visualizations, customize plots, and versatility in visualization types, the Matplotlib library has become a valuable tool in analyzing and forecasting prices for road freight transport services. Its ability to present data in a clear and understandable manner makes it a significant choice for researchers in this field.

The seaborn library is an expedient choice for forecasting the price of road freight transport services due to its advanced visualization features, ability to create attractive charts, and built-in statistical analysis tools. Seaborn offers tools for creating a variety of statistical plots, which are crucial for presenting price data. The ability to create elegant and informative visualizations allows for the depiction of price trends and other key data related to road freight transport services.

Seaborn provides built-in functions for statistical analysis, such as creating distribution, correlation, and regression plots. In price forecasting, these functions enable researchers to analyze relationships between variables and identify price patterns. The seaborn library allows the advanced customization of plots, enabling the creation of aesthetic and professional visualizations. In price analysis, the ability to customize colours, styles, and other graphical parameters helps highlight important information.

Seaborn is easily integrated with the pandas and NumPy libraries, which are often used for numerical data analysis. This integration facilitates the smooth processing and visualization of data related to the pricing of road freight transport services. According to [41], seaborn can be used to create statistical graphics in Python. It provides a high-level interface integrated with the Matplotlib library and tight integration with pandas data structures. The functions available in the seaborn library present a declarative, dataset-oriented API that simplifies transforming data questions into graphics that can answer them.

For a given dataset and plot specification, seaborn automatically maps data values to visual attributes such as colour, size, or style, internally computes statistical transformations, and adds descriptive axis labels and legends. Many functions in the seaborn library can generate multi-panel plots, allowing comparisons between conditional subsets of data or different variable sets within the dataset. Seaborn is designed to be useful at various stages of a scientific project. By generating complete graphics with a single function call and minimal arguments, seaborn facilitates quick prototyping and exploratory data analysis. Since it offers extensive customization options and exposes underlying Matplotlib objects, it can be used to create sophisticated, publication-quality graphics.

Owing to its advanced visualization features, support for statistical analysis, and customization capabilities, the seaborn library is a valuable tool for forecasting the prices of road freight transport services. Its ability to create attractive and informative visualizations enhances the understanding and interpretation of transport price data.

The scikit-learn library is a good choice for forecasting the price of road freight transport services due to its advanced machine learning tools, variety of predictive models, and flexibility in adapting to specific analysis needs. Scikit-learn provides a rich spectrum of advanced machine learning techniques, including regression, classification, and clustering. In price forecasting, the ability to select the most appropriate model and machine learning technique is crucial for achieving accurate forecast results.

Scikit-learn offers a wide selection of predictive models, such as linear regression, random forests, neural networks, and support vector machines. In price analysis, the ability to experiment with different models allows for identifying the best fit for the specific forecasting problem. The scikit-learn library enables the customization of analyses to meet the specific needs of the study. It includes tools for tuning model hyperparameters, performance evaluation, and cross-validation. This allows models to be optimized for price forecasting accuracy.

Scikit-learn is often used in conjunction with other data analysis libraries, such as NumPy and pandas. This integration facilitates the efficient processing and analysis of data related to the pricing of road freight transport services. According to [42], scikit-learn is a Python module that integrates a wide range of advanced machine learning algorithms for medium-scale supervised and unsupervised problems. This package focuses on making machine learning accessible to non-specialists using a high-level, general-purpose programming language. Ease of use, performance, documentation, and the consistency of the programming interface are emphasized. The library has minimal dependencies and is distributed under the BSD license, encouraging its use in both academic and commercial environments.

Owing to its advanced machine learning techniques, variety of predictive models, and flexibility in analysis customization, the scikit-learn library is an indispensable tool in forecasting the price of road freight transport services. Its ability to select optimal models and tailor analyses to specific research requirements contributes to accurate price forecasts and supports the decision-making process.

The XGBoost algorithm is an excellent choice for forecasting the price of road freight transport services due to its ability to efficiently model complex dependencies, regularization techniques, and outstanding performance in analyzing large datasets. XGBoost is based on gradient boosting of decision trees and is well-suited for modelling intricate and nonlinear relationships in data. In price forecasting, multiple factors can influence price values, and this capability allows various factors involved in the forecasting process to be considered accurately.

XGBoost offers advanced regularization techniques, such as limiting tree depth and leaf weights. In price analysis, this approach helps mitigate overfitting, which improves the generalization ability of the model. The XGBoost algorithm is characterized by excellent performance and scalability, which is crucial when analyzing large datasets related to prices. Its optimized implementation enables fast model training, which is essential when dealing with dynamic and changing price data.

XGBoost can flexibly handle numerical, categorical, and temporal variables, among others. In price forecasting, in which data can encompass different types of information, this feature allows for the comprehensive consideration of available data.

Owing to its ability to model complex dependencies, regularization techniques, and outstanding performance, the XGBoost algorithm is the best choice for forecasting the prices of road freight transport services. Its ability to account for multiple factors and efficiently analyze large datasets results in accurate price forecasts, thus supporting the decision-making process.

3.3. Success Metric

The right way to compare predicted prices with actual prices is to calculate the mean percentage error (MPE), as shown in Equation (1). MPE measures the average percentage error between actual and predicted values, which allows the user to better understand how wrong a model is in percentage terms.

$$\mathrm{MPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right) \times 100, \quad (1)$$

where: y_i – real value; $\hat{y}_i$ – predicted value; n – number of observations.

3.4. Cross Validation

Cross-validation is a fundamental technique in machine learning and statistical modelling that is used to evaluate model performance, optimize hyperparameters, and enhance the generalizability of predictive algorithms. Various studies have explored its applications and adaptations to address specific challenges across different domains. For instance, k-fold and leave-one-out cross-validation methods are commonly employed for evaluating classification algorithms. One study [43] highlighted these methods' limitations by addressing inconsistencies in their application and introducing independence assumptions to derive sampling distributions for their point estimators, offering new insights into performance evaluation methodologies.

Another study [44] applied cross-validation to optimize hyperparameter selection for extreme learning machine models, focusing on reducing the computational costs associated with repeated model training. Similarly, the comparison of nested and flat cross-validation approaches for binary classification demonstrates their effectiveness in evaluating algorithms and hyperparameters, revealing that flat cross-validation can provide computational efficiency while maintaining comparable performance to nested approaches [45].

Cross-validation also plays a crucial role in assessing the generalizability of machine learning models for multi-source datasets. A previous study [46] utilized k-fold and leave-source-out cross-validation to evaluate cardiovascular disease classification models using multi-source electrocardiogram data, highlighting the limitations of traditional methods in capturing inter-source variability.

In kernel regression, cross-validation has been shown to optimize bandwidth selection when combined with penalty functions, such as generalized cross-validation, Shibata's model selector, Akaike's information criterion, and Akaike's finite prediction error. One study [47] applied these methods within the Nadaraya-Watson kernel estimator framework, offering practical guidelines for bandwidth optimization. Similarly, another study adapted cross-validation for ridge regression and showed that a modified approach using repeated fold assignment and quantile-based selection improved the stability of shrinkage parameter estimation [48].

Efficiency in cross-validation has been further enhanced through innovative methods such as complexity-based efficient cross-validation, which has been developed for binary classification problems. By optimizing training data size and experiment runs, this method reduces evaluation time while maintaining performance comparable to traditional techniques like k-fold and repeated random sub-sampling cross-validation [49]. In building energy assessment, repeated cross-validation has been employed to validate machine learning models, thus reducing accuracy variations and providing stable results with sufficient data and an optimal fold number of 10 [50].

Lastly, cross-validation has been effectively applied to select the number of components in principal component analysis. A study [51] introduced computationally efficient approximation criteria, including the smoothing approximation of cross-validation and generalized cross-validation, to reduce the computational cost associated with traditional leave-one-out cross-validation.

These diverse applications underscore the adaptability and importance of cross-validation in achieving reliable, efficient, and robust model evaluation across various fields.

3.5. Forecasting Methods

The historical average method is a simple forecasting technique that predicts future values based on the average values of past data. It calculates the forecast as the arithmetic mean of all past observations and is particularly useful when the data lacks clear trends or seasonality. This method assumes that past data are a reliable predictor of future values and is most effective in stable environments where data fluctuate around a constant mean, as shown in Equation (2). While it is easy to implement and requires minimal computational resources, the historical average method has limitations, as it ignores trends, seasonality, and recent changes in data, making it somewhat unsuitable for datasets with such patterns. Despite these limitations, it is often used as a baseline for more complex forecasting models or in situations when data are relatively stable.

$$\hat{y}_i = \frac{\sum_{i=1}^{n} y_i}{n}. \quad (2)$$

Regression is a statistical technique used to model and analyze relationships between variables. The main goal of regression is to understand how a dependent variable (output, target) is influenced by one or more independent variables (input, predictor). Regression allows for

predicting the value of the dependent variable based on the known values of the independent variables.

Linear regression assumes that there is a linear relationship between the independent variable X and the dependent variable, which can be described by Equation (3).

$$\hat{y}_i = \beta_0 + \beta \times X, \tag{3}$$

where: β_0 – intercept; β – coefficient slope.

In linear regression, values for β_0 (intercept) and β (slope) are sought to minimize the sum of squared errors, which are the differences between the actual values Y and the values predicted by the model Y. The sum of squared errors is given by Equation (4)

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - (\beta_0 + \beta \times X))^2, \tag{4}$$

β is calculated using Equation ((5).

$$\beta_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \tag{5}$$

where: $\overline{X}$ – mean X value; $\overline{Y}$ – mean Y value.

After β has been calculated, β_0 can be calculated using Equation (6),

$$\beta_0 = \overline{Y} - \beta_1 \times \overline{X}. \tag{6}$$

Ridge regression is a regularization technique used in linear regression to stabilize and improve a model in situations when there are many correlated independent variables (multicollinearity) or when the number of variables is close to the number of observations. Unlike standard linear regression, ridge regression adds an additional term to the cost function that penalizes large values of the regression coefficients, thereby preventing the model from overfitting the data.

Ridge regression seeks to minimize both the prediction errors and the magnitude of the regression coefficients, leading to more stable and less overfitted models. This technique is particularly useful in analyses in which the variables are highly correlated or where modelling a large number of variables risks overfitting the training data.

Least absolute shrinkage and selection operator (LASSO) is a regression method that introduces L1 regularization, aimed at simultaneously estimating regression coefficients and selecting variables. Unlike ridge regression, which applies L2 regularization, LASSO penalizes the sum of the absolute values of the regression coefficients, which can lead to some coefficients being shrunk exactly to zero. LASSO minimizes the sum of squared errors with an additional L1 regularization term, which is the sum of the absolute values of the coefficients.

Due to L1 regularization, LASSO can shrink some coefficients to 0, effectively excluding the corresponding variables from the model. This makes LASSO an effective tool for variable selection, identifying which variables are most important to the model. By zeroing out some coefficients, LASSO results in simple models that are easy to interpret and often generalize better to new data. LASSO is particularly useful when dealing with a large number of variables, many of which may be irrelevant. It is widely used in genomic analysis, variable selection in high-dimensional data, and other scenarios where modelling with a large number of predictors is necessary. While ridge regression only shrinks the coefficient values, LASSO can eliminate some variables. This means that LASSO provides both regularization and variable selection, which is a key advantage.

In the LASSO objective function, the α parameter controls the strength of the regularization, determining how much the model penalizes large coefficient values of β_j. When α is close to 0, the regularization has little effect, and the model behaves similarly to ordinary

linear regression, including all variables. When α becomes significant, some coefficients can be set to zero, thereby eliminating the corresponding variables from the model.

ElasticNet is a linear regression method that combines the regularization techniques of L1 (used in LASSO) and L2 (used in ridge regression). It is particularly useful in situations involving many features that are highly correlated or when the number of predictors exceeds the number of observations. By blending the strengths of LASSO and ridge regression, ElasticNet allows for simultaneous variable selection and stabilization of coefficient estimates. ElasticNet introduces a mix of L1 and L2 penalties.

Like LASSO, ElasticNet can shrink some coefficients to 0, effectively selecting a subset of the available variables. The L2 component, however, allows it to handle situations when LASSO might struggle, such as when predictors are highly correlated. In cases where many predictors are correlated, LASSO tends to select one variable from a group and discard the others, which may lead to suboptimal models. ElasticNet mitigates this by spreading the penalty across all correlated predictors, allowing it to select groups of correlated variables together. ElasticNet is more flexible than LASSO or ridge regression alone, as it can balance both types of regularization. This is particularly useful in complex datasets with multicollinearity or when there is a need for both variable selection and stable coefficient estimation. ElasticNet is widely used in fields where the number of predictors is large, such as genomic analysis, financial modelling, and machine learning, particularly when variable selection and model interpretability are critical.

ElasticNet offers a flexible and powerful approach to linear regression by combining the benefits of LASSO and ridge regression. It is especially useful when predictors are highly correlated or when the number of predictors exceeds the number of observations. By tuning λ and ρ, ElasticNet can be adjusted to balance between variable selection and coefficient stability, making it an essential tool in modern regression analysis.

A decision tree is a non-parametric, supervised learning algorithm that is used for both classification and regression tasks. It works by recursively splitting the dataset into subsets based on the feature values that result in the highest information gain (or lowest error for regression), creating a tree-like structure where each internal node represents a decision based on a feature, each branch represents the outcome of the decision, and each leaf node represents the predicted value or class. Decision trees are intuitive, easy to interpret, and capable of modelling complex relationships in data. Decision trees recursively divide the data into smaller and smaller subsets by selecting the feature and split point that minimize the error or maximize the information gain. For regression, this process minimizes the variance of the target variable within each subset. The reduction in the variance (or mean squared error) of the target variable is the most common criterion for selecting a split for regression tasks. The algorithm searches for the feature and threshold that best split the data, reducing the variability in the target values in each resulting subset. The prediction is made at the terminal nodes (leaves) of the tree. In regression trees, the value at a leaf is typically the mean of the target values within that leaf. Decision trees are simple to understand and visualize, making them interpretable. They can model non-linear relationships and do not require feature scaling or normalization. Decision trees can also handle both numerical and categorical data. A key drawback of decision trees is that they can easily overfit training data, especially if the tree is allowed to grow deep. This overfitting occurs when the tree becomes too complex and starts capturing noise in the data. Regularization techniques such as limiting the tree depth or pruning can help mitigate this issue. Decision trees are used in many fields, including finance, healthcare, and marketing, where the

ability to interpret and explain decisions is important. They are also used as building blocks in complex models, such as random forests and gradient boosting machines.

The Random Forest Regressor is an ensemble learning method used for regression tasks, which builds multiple decision trees and combines their predictions to produce a more accurate and stable result. It is an extension of decision trees that addresses some of their limitations, such as overfitting. The random forest algorithm uses a technique known as "bagging" (bootstrap aggregating) and random feature selection to create a collection of trees that form a robust model. The Random Forest Regressor creates multiple decision trees by randomly sampling the training data with replacement (bootstrap sampling). Each tree is built on a different random subset of the data, which helps to reduce the variance and prevents the model from overfitting. At each split of a decision tree, the random forest algorithm selects a random subset of features to determine the best split. This added randomness reduces the correlation between individual trees, improving the overall accuracy and stability of the model. For regression tasks, the final prediction is the average of the predictions made by all the decision trees in the forest. This aggregation reduces the overall variance of the model, making it more robust to overfitting and better able to generalize to new data. The Random Forest Regressor can effectively handle datasets with a large number of features and is robust to noisy data. The random feature selection ensures that not all features are used in each tree, which helps handle irrelevant or redundant features. The algorithm can provide a ranking of feature importance based on how often a feature is used to split the data across the trees. This insight is useful for understanding the influence of different variables on the target and for feature selection in model optimization. Individual decision trees tend to overfit, especially if they are deep. By averaging the predictions from multiple trees, the Random Forest Regressor reduces the overfitting that typically occurs with a single decision tree, leading to a model that generalizes better.

The Extra Trees Regressor, also known as the Extremely Randomized Trees Regressor, is a machine learning algorithm from the ensemble methods group used for regression problems. This method constructs a large number of decision trees, which are then aggregated to obtain the final prediction. Unlike the classic random forest, the Extra Trees Regressor introduces additional randomness during the tree-building process, which can lead to better model generalization and increased computational efficiency.

The fundamental premise of the Extra Trees Regressor is the use of the entire dataset to build each tree without applying sample bootstrapping. At each node split of the tree, a subset of features is randomly selected, and then, for each of these features, a split threshold is randomly chosen from the range of values of that feature in the training set. Among these random splits, the one that optimizes a specific split quality criterion—such as minimizing variance in the case of regression—is selected.

Introducing additional randomness in selecting split thresholds and features causes individual trees in the ensemble to be more diverse. As a result, aggregating the predictions of many such trees leads to a model with lower variance and better generalization ability on test data. Increased randomness also reduces the risk of overfitting the model to the training data.

The Extra Trees Regressor is particularly useful in forecasting prices of road freight transport services, where data are characterized by high complexity and nonlinearity of relationships between variables. This algorithm can effectively model complex relationships among various features, such as cargo parameters, transport routes, economic conditions, or seasonality. Moreover, this method is robust to outliers and noise in the data, which is important in practical applications with real market data.

The authors of a previous study [52] demonstrated the Extra Trees Regressor's effectiveness compared to other ensemble methods, such as random forest. The application of this method in various fields confirms its universality and effectiveness in modelling complex regression problems [53].

The main difference between the Extra Trees Regressor and random forest is how trees are built. Random forest bootstraps samples and searches for optimal split thresholds, whereas the Extra Trees Regressor uses the entire dataset without bootstrapping, and the split thresholds are selected randomly. These differences can lead to faster model training and better generalization of results on unseen data.

The Extra Trees Regressor is available in popular machine learning libraries such as scikit-learn [3], which facilitates its practical application. In this research, this algorithm is used to build a predictive model of road freight transport service prices, and its effectiveness is compared with other methods in subsequent chapters of this work.

The Adaptive Boosting Regressor (AdaBoost Regressor) algorithm belongs to the ensemble methods group and is specifically designed for regression tasks. Introduced by [54], AdaBoost combines multiple weak learners to form a strong predictive model. The fundamental idea is to train weak learners sequentially, typically using simple decision trees known as stumps, where each subsequent model focuses more on the data points that previous models predicted incorrectly.

In regression, AdaBoost aims to minimize a suitable loss function for continuous outputs, such as the squared error loss. The algorithm begins by assigning equal weights to all training instances. During each iteration, a weak learner is trained on the weighted dataset. After training, the model's error is calculated, and the weights of the training instances are updated so that instances predicted poorly receive higher weights, causing subsequent learners to focus more on them. This process continues for a predetermined number of iterations or until convergence.

AdaBoost constructs a strong predictive model by combining the predictions of all weak learners, typically through a weighted sum where each learner's weight is proportional to its accuracy. This ensemble approach allows AdaBoost to achieve higher predictive accuracy than individual weak learners.

The AdaBoost Regressor offers several advantages. It improves accuracy by focusing on difficult-to-predict instances, allowing the ensemble model to achieve better performance. It is flexible, as it allows for the use of various types of weak learners, providing versatility in modelling different types of data. Moreover, the underlying concept is straightforward, and the algorithm is relatively easy to implement.

However, some limitations must be considered. AdaBoost is sensitive to noisy data and outliers because it emphasizes mispredicted instances, which can lead to overfitting to noise present in the data. The sequential training process can also be computationally expensive, especially when large datasets and many iterations are involved.

In forecasting prices for road freight transport services, the AdaBoost Regressor is particularly useful because it captures complex and nonlinear relationships between features and target variables. Variables such as distance, cargo type, economic indicators, and time-related factors can interact in intricate ways, and AdaBoost can effectively model these interactions by combining multiple weak learners that focus on different aspects of the data. By leveraging this algorithm, the predictive model can provide more accurate price forecasts, which is crucial for decision-making in the transportation industry.

Implementation of the AdaBoost Regressor is available in popular machine learning libraries like scikit-learn, facilitating its application in practical scenarios. In this research, the AdaBoost Regressor is employed to build a predictive model for road freight transport service prices, and its performance will be compared with other methods in subsequent chapters.

The Gradient Boosting Regressor (GBR) is a robust machine learning algorithm belonging to the ensemble methods specifically tailored for regression tasks. Introduced by Friedman [55], it builds predictive models by sequentially combining weak learners, typically decision trees, and optimizing a loss function through gradient descent. This iterative approach enables the GBR to correct previous errors and achieve high accuracy, making it suitable for various applications.

In modern agriculture, the GBR has been employed to predict crop nutrient requirements using historical rainfall data and soil fertility levels [56]. By optimizing fertilizer use, this application enhances crop yields and promotes sustainability. Similarly, the GBR significantly outperformed the RANSAC Regressor in solar power generation forecasting, achieving an accuracy of 93.3% compared to 79%, with a statistically significant improvement ($p = 0.001$) [57]. The GBR has also proven valuable in supply chain management, where it improves daily demand forecasting for a Brazilian logistics company [58]. Compared to artificial neural networks, the GBR reduced error rates by an average of 0.86% and a maximum of 1.44%, addressing overfitting and complexity issues often faced with limited datasets.

In the field of epigenetics, the GBR has been applied to predict individual ages using DNA methylation markers, achieving a strong correlation ($R^2 = 0.97$) and a mean absolute deviation of 2.72 years on training datasets [59]. This model also demonstrated robust performance on independent datasets, proving effective across various conditions, including diseased samples.

Hydrological forecasting has similarly benefited from the GBR, as shown in a study predicting water levels at Lago de Chapala Dam [60]. Incorporating a 1-day lag feature for temporal context significantly reduced RMSE from 886.37 to 32.49, highlighting the GBR's effectiveness in leveraging temporal dependencies for water resource management.

In engineering, the GBR has been proposed as a faster alternative to the finite element method (FEM) for predicting stress intensity factors in small bore piping [61]. The correlation between GBR and FEM predictions was high ($R^2 = 0.977$), and the GBR drastically reduced computation time from 30 minutes to one second, demonstrating its efficiency in structural integrity assessments.

Chemical engineering applications include using the GBR to predict the infinite dilution activity coefficient ($\gamma\infty$) of dichloromethane (DCM) in ionic liquids (ILs) [62]. By combining ionic fragment contribution and the GBR, the model achieved a high R^2 of 0.9703 and strong generalization capability, providing critical data for IL-based absorption and recovery processes.

The GBR's capability in aerodynamic predictions was demonstrated in a study analyzing force distribution on horizontal axis wind turbine blades under varying wind speeds and angles of attack [63]. The model achieved variance scores of 0.933 and 0.917 for 4° and 8° attack angles, respectively, offering a faster alternative to computational fluid dynamics simulations for untested conditions.

Finally, the GBR has been shown to outperform linear regression in airfare prediction, achieving significantly higher accuracy (82.5% vs. 62.5%), with a statistically significant difference ($p < 0.05$) [64]. This highlights the GBR's utility in optimizing ticket pricing strategies.

These examples collectively demonstrate the GBR's versatility and effectiveness across diverse domains. Its ability to achieve high accuracy while addressing challenges like overfitting, computational time, and data limitations establishes it as a critical tool for modern predictive modelling.

The XGBoost library [65] efficiently implements its algorithm and is seamlessly integrated with platforms like scikit-learn, making it both accessible and user-friendly. Its widespread adoption and demonstrated performance across various domains establish its suitability for developing predictive models in this research area.

An article [66] investigated the application of the XGBoost Regressor in predicting transmittance values for graphene-based refractive index sensors. The results showed that regression analysis can reduce simulation time and computational resources by over 70%, thereby enhancing the efficiency of sensor design and optimization.

Other research [67] explored feature selection methods to improve ozone prediction using regression models, emphasizing that the XGBoost Regressor outperformed other approaches in accuracy and generalizability, underscoring the significance of machine learning in advancing environmental forecasting and public health policy.

Another study [68] focusing on academic performance prediction of immigrant students using PISA data highlighted the XGBoost Regressor's superior predictive accuracy and low error rates when optimized through hyperparameter tuning, supporting the development of informed educational policies.

Another paper [69] described a system combining N-Stream ConvNets and XGBoost Regressor with embedding and lexicon-based features, achieving state-of-the-art performance in Arabic valence intensity regression and ordinal classification tasks in the Affect in Tweets challenge.

In financial risk estimation, a study [70] examined fractional differencing and logarithmic transformations, revealing that contrary to De Prado's claims, logarithmic transformation outperformed fractional differencing in volatility prediction accuracy and profitability when using LSTM and XGBoost Regressor models.

In research on household power consumption prediction, the XGBoost Regressor surpasses other models in accuracy, measured by mean absolute error, root mean square error, and coefficient of determination, confirming its reliability in this application [71].

In airline ticket price prediction, a study [72] demonstrated that the XGBoost Regressor achieved the highest accuracy ($R^2 \approx 84\%$) and lowest RMSE (≈ 1807.59) among tested models, confirming its effectiveness in forecasting price fluctuations.

A study [73] addressing ride fare prediction shows that the XGBoost Regressor achieved top accuracy with an R^2 of 0.9734. Additionally, subsequent classification of fare ranges identified the support vector machine as the best performer, achieving an F1 score of 99.53.

In agricultural forecasting involving a machine learning-based system developed for sorghum yield prediction in Sudan's El-Gadarif State [74], the XGBoost Regressor, with an R^2 of 0.87, significantly outperformed the Multi-Layer Perceptron model, highlighting its efficacy in data-scarce contexts and the importance of careful calibration.

Finally, a study on song popularity prediction using Spotify data [75] revealed that XGBoost Regressor, Light Gradient Boosting Machine (LightGBM) Regressor, and random forest models achieved high correlation values of 99.85%, 99.87%, and 99.84%, respectively, with LightGBM slightly leading in predictive accuracy.

The LightGBM Regressor is a highly efficient and scalable machine learning algorithm developed by [76]. It is a gradient boosting implementation optimized for performance on large-

scale datasets, and it is particularly noted for its speed, low memory usage, and high accuracy, making it ideal for real-time applications and large data volumes.

One paper [77] introduced an ensemble learning approach that combines extreme gradient boosting (XGBoost) and a light gradient boosting machine (LightGBM) to forecast electrical load behaviour. The hybrid model outperformed individual methods, reducing key error metrics, including MAPE, RMSPE, and MAE, by over 1%, as validated using data from the Pennsylvania-New Jersey-Maryland interconnection power grid.

Another study [78] highlighted the application of the Light Gradient Boosting Machine (LightGBM) Regressor in predicting hospital length of stay (LoS), achieving an outstanding R^2 score of 96.14%. The study also proposed a unified framework for generalized prediction by critically evaluating existing methods, exploring routinely collected data, and offering recommendations for robust knowledge modelling.

Another study [79] investigated machine learning-based daily reference evapotranspiration (ET0) predictions for Taif, Saudi Arabia. The results demonstrated that the Light Gradient Boosting Machine Regressor (LightGBMR) outperformed other models, achieving an R^2 of 0.998 with minimal error metrics (MSE: 0.016 mm day^{-1}, RMSE: 0.128 mm day^{-1}, MAE: 0.093 mm day^{-1}). The study further revealed that the top five weather parameters identified through principal component analysis can provide predictions as accurate as those attained using all parameters, offering an optimized approach for ET0 estimation.

Another study [80] proposed a novel methodology using the Light Gradient Boosting Regressor (LightGBR) to predict engine-out emissions of NOx, HC, and CO with high precision. The model achieved mean absolute percentage errors (MAPE) of 5.2% for CO, 5.7% for HC, and 6.8% for NOx across 47 experimental driving cycles. Thus, the study demonstrated this regressor's ability to evaluate powertrain calibration changes and significantly reduce experimental testing costs in a virtual environment.

The Gaussian Process Regressor (GPR) is a non-parametric, probabilistic model used for regression tasks, providing a flexible and robust way to model complex relationships between variables. Unlike traditional regression methods that fit a fixed function to data, the GPR treats the function being modelled as a distribution over possible functions, allowing it to capture a wide variety of behaviours. The core of the GPR is its use of a Gaussian process, which is defined by a mean function and a covariance function, also known as the kernel. The kernel function plays a key role in determining how input points are related to one another, with commonly used kernels like the radial basis function and the Matérn kernel allowing the GPR to model both smooth and more complex, non-linear functions [81].

One of the main strengths of the GPR is its ability to quantify uncertainty in predictions, making it particularly useful in scenarios where understanding the confidence in a model's output is crucial. This uncertainty estimation is especially important in cases where the data may be sparse, noisy, or incomplete, as the GPR reflects higher uncertainty in regions with little information. This feature sets the GPR apart from other regression methods, which typically provide only point estimates without indicating how confident the model is in its predictions.

However, the GPR has limitations. The computational cost of the GPR grows cubically with the number of data points, making it impractical for very large datasets. The model also requires careful tuning of the kernel and its hyperparameters, as the choice of kernel significantly impacts the model's performance. Despite these challenges, the GPR is highly flexible and can model various relationships between variables, making it well-suited for problems where the underlying data patterns are complex and difficult to capture with simpler methods.

In forecasting prices for road freight transport services, the GPR's ability to model non-linear relationships while providing uncertainty estimates is highly valuable. Transport pricing is influenced by various factors, including distance, cargo type, fuel prices, and fluctuating economic conditions. The flexibility of the GPR allows it to capture these intricate relationships, and its ability to estimate the confidence of its predictions offers additional insights for decision-making in an industry that is often subject to high levels of uncertainty. Nevertheless, due to the computational challenges associated with the GPR, it may be best suited for medium-sized datasets where interpretability and uncertainty quantification are priorities.

The Multi-Layer Perceptron (MLP) Regressor is an artificial neural network used for regression tasks. It is a feedforward neural network model composed of multiple layers of nodes, or neurons, which are organized into an input layer, one or more hidden layers, and an output layer. Each neuron in the MLP Regressor is connected to neurons in adjacent layers through weighted connections, and these weights are adjusted during the training process to minimize the error between predicted and actual outputs. This makes the MLP Regressor highly flexible and capable of capturing complex, non-linear relationships between input variables and the target output [82].

The MLP Regressor operates by passing the input data through a series of transformations using activation functions, typically the rectified linear unit (ReLU) or the sigmoid function. Each layer processes the inputs, computes a weighted sum, and applies the activation function to produce the output, which is then passed to the next layer. The final layer produces the predicted output. The training process is performed using backpropagation, where the model adjusts the weights in the network to minimize the error, usually measured by a loss function such as the mean squared error (MSE). Gradient descent—or one of its variants, such as stochastic gradient descent—is used to update the weights iteratively until convergence is achieved or a stopping criterion is met.

One of the main strengths of the MLP Regressor is its ability to model highly complex, non-linear functions, making it particularly useful for regression tasks where traditional linear models may struggle. The presence of multiple hidden layers allows the model to learn hierarchical representations of the input data, capturing intricate patterns that are difficult to model with simpler approaches. The MLP Regressor can handle both continuous and categorical input data, although it often requires normalization or scaling of the input features to ensure efficient training and convergence.

However, the MLP Regressor has certain limitations. The model is prone to overfitting, especially when many hidden layers or neurons are used without appropriate regularization techniques, such as dropout or L2 regularization. Training MLPs can also be computationally expensive, particularly when dealing with large datasets or deep architectures. The model's performance is highly sensitive to the choice of hyperparameters, such as the number of hidden layers, the number of neurons per layer, the learning rate, and the activation function. These hyperparameters often require careful tuning through grid or random search methods to achieve optimal performance.

Concerning prices for road freight transport services, the MLP Regressor is particularly valuable for capturing the non-linear and complex relationships between various factors influencing transport pricing. Variables such as distance, cargo type, fuel prices, and economic conditions can interact in ways that are difficult to model using traditional regression techniques. The MLP Regressor's capacity to learn complex mappings from inputs to outputs makes it an excellent choice for this type of predictive modelling. However, given its

susceptibility to overfitting, it is important to ensure that regularization techniques are applied and that proper cross-validation is performed during model training to prevent overfitting and improve generalization.

3.6. Expert Calculations

In machine learning and predictive modelling, expert calculations involve integrating domain knowledge and manual computations to complement or validate the performance of automated models. While machine learning algorithms provide powerful tools for prediction, expert insights are often necessary to ensure alignment with real-world phenomena, industry standards, and regulatory requirements. This approach bridges the gap between purely data-driven models and the nuanced realities of complex systems like road freight transport.

Expert calculations rely on well-established principles, empirical formulas, and industry rules derived from years of experience. These are particularly valuable when data are sparse or noisy or when certain aspects of the problem demand a deeper understanding that might not be captured fully by machine learning models.

Expert calculations are critical in adjusting or fine-tuning model predictions to forecast prices for road freight transport services. Freight pricing is influenced by a complex interplay of factors such as fuel costs, vehicle types, cargo characteristics, distance, and prevailing economic conditions. Experts with a deep understanding of these dynamics can provide insights or apply specific formulas to enhance the accuracy and reliability of predictions.

4. THEORETICAL METHODOLOGY FOR MODEL DEVELOPMENT

This chapter presents a comprehensive theoretical methodology for developing a machine-learning model aimed at forecasting prices for road freight transport services. The chapter is divided into sections, each addressing a critical component of the methodology, from data collection to integration with external databases. The detailed exploration of these components provides a robust framework for constructing an effective predictive model.

The initial focus is on the methodology for gathering data, emphasizing the importance of collecting accurate and comprehensive data for model development. Various attributes are considered, including distance, location, date, time, trailer type, body characteristics, vehicle type, loading/unloading method, and cargo securing method. Each of these attributes is crucial for building a reliable dataset that reflects the complexities of road freight transport services.

Following data collection, the focus shifts to data transformation. This involves processes that convert raw data into a format suitable for analysis and modelling. Specific transformations are necessary for distance features, relation features, date features, time, body type, body characteristics, vehicle type, loading/unloading method, load securing method, and other features. These transformations are essential for normalizing data and enhancing their utility in predictive modelling.

Data analysis methods are then presented, providing techniques for extracting meaningful insights from the transformed data. Various analysis methods tailored to different data types, such as distance, relations, and body type, are discussed. The goal is to identify patterns and correlations that can inform the development of predictive models.

The chapter also addresses the integration of the model with external databases, such as those containing fuel prices and economic data. This integration is vital for ensuring the model's accuracy and relevance, as it allows for the inclusion of external factors that significantly impact road freight transport pricing.

These components form a cohesive methodology for developing a machine-learning model that accurately forecasts road freight transport prices. The structured approach ensures that each aspect of data handling and analysis is meticulously addressed, providing a solid foundation for model development.

4.1. Data Gathering Method

This section focuses on the data recording method, which is crucial for effectively collecting and storing information related to price forecasting for road freight transport services. In this context, a scientific analysis of various data recording approaches and their impact on the price forecasting process will be discussed. The data recording method is essential for maintaining the accuracy and completeness of information in price forecasting. Applying an appropriate method can significantly impact the quality of forecast results and the efficiency of the analysis. According to [83], one of the important principles is that the information provided by a variable should be reflected in the encoded data.

It is proposed that transport order data be recorded in comma-separated values (CSV) format. The CSV format is characterized by a simple structure by which each data record is separated by commas or other delimiters. In the case of transport order data, which includes information on the origin, destination, goods, delivery time, and other variables, the CSV format allows for the clear and understandable representation of these data. The CSV format is widely accepted by data analysis tools and spreadsheets. This allows for easy processing, filtering,

sorting, and analyzing data related to transport orders. Additionally, this format allows for the quick loading and saving of data, which is crucial in the dynamic environment associated with transport.

In transport orders, data can change over time, and new attributes can be added in response to evolving needs. The CSV format enables flexible data structure expansion by adding new columns without complex structural modifications. Data recorded in CSV format is easily understandable, which facilitates data sharing among different teams, departments, or partners in the transport industry. This contributes to the efficient flow of information across the entire transport-related ecosystem. The CSV format is an appropriate choice for recording data related to transport orders. Its simple structure, ease of processing and analysis, flexibility in data expansion, and the ability to easily share data translate into an effective and practical method of storing information related to transport orders.

In price forecasting for road freight transport services, clean code plays a key role in effectively conducting analyses and developing predictive models. “Clean code” refers to computer code that is understandable, consistent, well-formatted, and free of unnecessary elements. There are several critical scientific reasons for maintaining clean code in this field. Price forecasting for road freight transport services is a complex task that requires manipulating and analyzing large amounts of data. Clean code makes it easier to understand what operations are being performed on the data, what transformations are being applied, and what models are being created. This makes analyses more transparent, which is essential for the credibility of the results and the ability to replicate studies. Clean code enhances the readability and understandability of the research team's work. Working together on a price forecasting project requires collaboration among various experts, such as programmers, statisticians, and analysts. Clean code facilitates communication among team members, enabling quick understanding and verification of the code by other project participants. In price forecasting, accuracy and repeatability are crucial. Clean code minimizes the risk of errors and ambiguities in data analysis and model creation. Owing to its consistent and readable code structures, it makes it easier to detect potential errors and minimize the risk of introducing inaccuracies into the forecasting process. Clean code is essential for price forecasting for road freight transport services due to increased transparency of analyses, facilitated team collaboration, error minimization, and ensuring accuracy and repeatability in the forecasting process. All these aspects contribute to efficient and reliable work on creating and implementing predictive models in this field.

A previous article [84] presented an approach that positively impacted the quality of source code from the very beginning of software development. The main goal was to achieve readable code and ensure its long-term usefulness. It is assumed that using the discussed toolset—such as linters (e.g., ESLint, Pylint), code formatters (e.g., Prettier, Black), and static code analysis tools (e.g., SonarQube, Checkmarx)—will help prevent the deterioration of software maintainability. The choice of appropriate procedures and practices, including adhering to established coding standards and leveraging automated testing frameworks, is more of a cultural than a technical decision. Numerous tools with a long history and proven effectiveness are available. However, their effective use requires an open approach from developers, enabling the acceptance of these solutions and their integration into the development workflow.. It is necessary to achieve a common understanding of the importance of adhering to specific code quality standards. It is worth noting that issues related to clean code can deteriorate over time, and quality attributes such as performance must be maintained in long-term software.

It is proposed to use English for naming features. The use of English feature names in the dataset is justified for several important reasons. The global nature of science and research means that much of the terminology, especially in the fields of technology, science, and data analysis, is used in English. Introducing feature names in English creates an understandable communication bridge between the doctoral dissertation and the international community of scientists, facilitating knowledge exchange and discussions about research results. English is widely used in scientific literature and professional publications related to data analysis and data science. Using English in feature names makes it easier to integrate new results with the existing scientific corpus and allows for the comparison and verification of findings. Using English terminology ensures consistency with industry terminology used in the international scientific environment and in professional practice. Consistent feature names are crucial for precise communication and understanding, especially in globally oriented fields such as transport and logistics. From a technical perspective, many tools, libraries, and frameworks for data analysis and machine learning are available in English. Therefore, choosing English feature names aligns with the convention used in programming and data analysis, facilitating work with data and their analysis.

Recording distance data with country segmentation is particularly important when considering tolls, whose costs are closely related to the specific country of occurrence. Introducing this aspect allows for a more accurate assessment of the impact of tolls on shaping the price of road freight transport services, taking into account differences between regions. Tolls, such as vignettes or highway fees, vary significantly by country. Each country may have its own regulations and toll rates based on local conditions and road infrastructure investments. Including these costs in the distance analysis between countries allows the specific aspects of a region to be considered and the total transport costs to be estimated more accurately. Analyzing distance data considering tolls adapted to the specific country enables precise forecasting of transport costs on a given route. This allows for realistic estimations of road freight transport service prices, depending on the destination and the anticipated route. Applying such an approach, which takes into account country-specific costs, allows for a more comprehensive and accurate analysis of transport costs and their impact on the final service price.

Recording distance data with country segmentation, considering various tolls characteristic of specific regions, is vital for reliable analysis and forecasting of road freight transport service prices. This approach enables the consideration of variable transport costs resulting from local conditions and allows for a more realistic estimation of prices that considers the significant factor of tolls in the pricing process of transport services. Despite the Polish language in the doctoral dissertation, the use of English feature names in the dataset is justified due to the global nature of science, as it facilitates communication and integration with the international scientific community, consistency with industry terminology, and work with data analysis tools and libraries. Feature names are written in uppercase letters, with words separated by underscores. Feature values are proposed to always be written in lowercase, separating words with underscores. When there are multiple values for a given case, individual values are separated by commas, and features are listed in alphabetical order.

Below is the classification of various transport-related features. Categories are divided into five main areas: "distance," "relation," "date," "goods feature," and "organizational features." Each category lists specific features related to that area. In the "distance" category, features related to distances are divided by country. In the "goods feature" category, details about goods are listed, such as "GOODS_TYPE" and "BODY_TYPE," while the "organizational features" category provides information related to the organization, such as

"OTHER_COSTS" and "QTY_LOADS." The list serves as a useful tool for classifying and analyzing various aspects of transport logistics:

- Distance: AT_KM, BE_KM, CZ_KM, DE_KM, DK_KM, EE_KM, ES_KM, FI_KM, HR_KM, FR_KM, HU_KM, IT_KM, LT_KM, LV_KM, NL_KM, PL_KM, RO_KM, SE_KM, SI_KM.
- Relation: COD_LP, COD_DP, ROUTE_TYPE.
- Date: START_LOAD_DATA, START_LOAD_TIME, END_LOAD_DATA, END_LOAD_TIME, START_DELIVERY_DATA, START_DELIVERY_TIME, END_DELIVERY_DATA, END_DELIVERY_TIME, TIME_OF_ENTRY.
- Cargo features: GOODS_TYPE, BODY_TYPE, VEHICLE_TYPE, LOAD_UNLOAD_METHOD, REQUIREMENTS, EPALE, LDM, TONS, M3.
- Organizational features: OTHER_COSTS, QTY_LOADS, QTY_DELIVERIES, PAYMENT_TERM, DOCUMENTS_BY, CUSTOMS.

Below are more detailed discussions. Each category has its own subsection. The choice of the kilometre as the basis for distance analysis in forecasting the price of road freight transport services is justified. The kilometre is a universally accepted unit of measurement in the field of transportation, especially for road transport.

The kilometre is a recognized unit of distance in most countries around the world. It is understandable and intuitive for most people, both in professional contexts and everyday life.

The kilometre allows for accurate distance measurements, which are crucial in transport routes. This is particularly important for cost forecasting, where even small differences in distances can impact the final transport costs.

Many roads and highways have distance markers in kilometres, facilitating navigation and route planning. This increases the accuracy and relevance of distance analysis in transport cost forecasting. Using the kilometre unit simplifies the comparison of different routes and distances in a consistent manner. This is essential when analyzing various transport routes and their impact on service pricing. The kilometre is an internationally accepted unit, which eases comparisons and data analysis between different countries and on a global scale. Many transport operators, navigation systems, and distance databases use the kilometre as a unit of measure. This enables easy utilization of existing data sources. The kilometre is an appropriate unit for distance analysis in forecasting the price of road freight transport services due to its universality, accuracy, standardization, and ease of comparison and data analysis.

The International Organization for Standardization (ISO) has developed a country coding standard known as ISO 3166, which is widely used in many fields, including road transport analysis. ISO 3166 defines unique codes for countries, allowing for the unambiguous identification of countries worldwide. It is proposed to record countries in accordance with the ISO 3166-1 standard, using two-character codes. The Online Browsing Platform [85] was utilized to obtain the country codes. Tab. 1 illustrates the creation of features for individual countries. All countries that were part of the European Union (EU) as of August 15, 2023, are included. All features are expressed in kilometres. There is potential to extend the scope to countries outside the European Union (EU) and to use other units, such as miles.

Postal codes, also known as address codes, constitute a system of numerical or alphanumeric identifiers used to precisely determine geographical locations for postal and logistical purposes. They are used to mark specific areas, districts, towns, or streets, enabling effective and efficient delivery of parcels, correspondence, and goods. Postal codes play a

crucial role in transportation and logistics fields, including forecasting the prices of road freight transport services.

Tab. 1
Names of features in the distance category

COUNTRY	ISO	FEATURE
Austria	AT	AT_KM
Belgium	BE	BE_KM
Bulgaria	BG	BG_KM
Croatia	HR	HR_KM
Cyprus	CY	CY_KM
Czech Republic	CZ	CZ_KM
Denmark	DK	DK_KM
Estonia	EE	EE_KM
Finland	FI	FI_KM
France	FR	FR_KM
Greece	GR	GR_KM
Spain	ES	ES_KM
Ireland	IE	IE_KM
Lithuania	LT	LT_KM
Luxembourg	LU	LU_KM
Latvia	LV	LV_KM
Malta	MT	MT_KM
Netherlands	NL	NL_KM
Germany	DE	DE_KM
Poland	PL	PL_KM
Portugal	PT	PT_KM
Romania	RO	RO_KM
Slovakia	SK	SK_KM
Slovenia	SI	SI_KM
Sweden	SE	SE_KM
Hungary	HU	HU_KM
Italy	IT	IT_KM

By incorporating postal codes in analyses and predictive models, it is possible to more accurately determine delivery locations and loading and unloading points, as well as to analyze the differences in transport costs depending on the location. Using postal codes, researchers and professionals in the transport industry can analyze price variability by region, identify key areas affecting costs, consider differences in tolls, and adjust logistical and pricing strategies to specific geographical areas.

Incorporating postal codes into analyses related to forecasting the prices of road freight transport services contributes to more precise and realistic forecasts, which is essential for making informed decisions in logistics, route planning, and cost assessment in road transport.

Tab. 2 presents the proposed method for recording location information. The records consist of two letters in line with the ISO standard (similar to distance encoding) and five digits.

For countries with shorter postal codes, the remaining places are filled with zeros. In analyses related to forecasting the prices of road freight transport services, accurately determining the location is crucial for achieving reliable and precise results. In this context, recording the location using postal codes demonstrates advantages over city names.

Postal codes constitute a system of unique identifiers that precisely specify a particular location on the geographical map. Unlike city names, which can be confusing or ambiguous since different towns often have similar names, postal codes are unique to a specific area. This allows for the precise identification of loading points, delivery locations, and other transport-related sites. Additionally, the use of postal codes facilitates the automation of logistical processes and data analyses.

Incorporating postal codes into price forecasting analyses for road freight transport enhances precision, eliminates errors associated with the inaccuracy of city names, and enables more effective logistical management. This results in more reliable forecasts, which are crucial for making business strategies and operational decisions in the field of transport.

Tab. 2

Location recording method

PL	12345
Two-letter country code	Five-digit location code

Tab. 3 presents a list of features related to location, encompassing two primary variables: the first loading place and the final unloading place. The relationship between these origin and destination points is a critical factor in analyzing the pricing of road freight transport services. This relationship sheds light on the structure and nature of freight movements and facilitates the identification of key trends and patterns in freight traffic, contributing to more informed decision-making.

In road freight transport, understanding the geographical relationship between loading and unloading locations is crucial for several reasons. Cost analysis enables the assessment of financial implications by examining the distance between loading and unloading points, helping to identify the most cost-effective routes. Efficiency optimization benefits from analyzing these relationships to streamline logistical operations, ensuring resources are utilized effectively and delivery times are minimized. Trend identification allows commonly used origin-destination pairs to be discovered, revealing market behaviours such as popular routes or regions with high demand for transport services. Strategic planning is supported by leveraging location data to tailor services, meeting the specific needs of regions and customer bases.

The use of postal codes for this analysis enhances precision and clarity, providing a granular and accurate examination of transport patterns. This methodology facilitates automation in data processing, resulting in more reliable and actionable insights. These insights empower transportation companies to make data-driven decisions, optimize operations, and strategically plan for evolving market demands.

Tab. 3
List of location features

Feature name:	Meaning of the feature:
CODE_LOAD_PLACE	Starting place of loading
CODE_DELIVERY_PLACE	Final unloading place
ROUTE_TYPE	Single or round trip

When analyzing and forecasting prices for road freight transport services, recording date data is a critical element. The consideration of dates in analyses has crucial implications for the accuracy and relevance of forecast results. The date provides a chronological context for understanding the sequence of events, changes, and trends.

In price analysis for transport, including dates allows seasonality, cyclicality, and price evolution over time to be identified. Awareness of price changes during specific periods can aid in making informed business decisions, such as cost optimization or adjusting pricing strategies. Additionally, analyzing date data enables the identification of regular patterns and exceptional events that may affect transport prices. These events can include holidays, periods of increased commercial activity, or unpredictable external factors such as regulatory changes or significant economic events.

Thus, gathering data is fundamental to the accuracy of analyses and price forecasts for road freight transport services. The accurate recording of dates allows for the proper consideration of temporal and sequential factors, leading to more precise and relevant analysis and forecast results.

Tab. 4 presents the method for recording date data. This table will outline the structure for capturing essential date-related information, ensuring that all relevant temporal aspects are appropriately considered in the analysis.

Tab. 4
Date gathering method

DD/	MM/	YYYY
Day	Month	Year

Dates should be recorded in a way that provides an understanding of the loading and unloading time frames. Loadings with strictly specified dates may incur higher costs due to limited flexibility and the need to precisely align vehicle fleets and personnel with a fixed schedule. Preparing for a specific delivery date may require more effort in planning and organization, impacting operational costs.

On the other hand, loadings within a range of dates that allow for greater flexibility in scheduling may enable more efficient use of available resources. Vehicle fleets can be more flexibly adjusted to variable demand, minimizing the risk of empty runs and potentially lowering operational costs.

Tab. 5 lists features related to date data, detailing how to capture essential information for loading and unloading time frames to ensure that all relevant temporal aspects are appropriately considered in the analysis.

Tab. 5
List of date features

Feature name:	Feature meaning:
START_LOAD_DATA	Start date of loading
END_LOAD_DATA	End date of loading
START_DELIVERY_DATA	Start date of unloading
END_DELIVERY_DATA	End date of unloading

In transport data analysis, time-related data are of paramount importance. Appropriate methods for recording these data are essential to accurately analyze and model temporal changes. Time can be divided into loading and unloading times, each of which can be further categorized into start and end times. The requirements for loading and unloading can be specified as either a range of hours or a fixed time, known as "FIX."

The method with an hourly range can correspond to the operational hours of a given business and is more commonly used by smaller enterprises. This allows for flexibility within business hours and can accommodate variable schedules.

On the other hand, a strictly specified loading time is more commonly practiced in large enterprises to streamline operations. In such cases, time slots are scheduled, known as hourly notifications. Failure to adhere to these scheduled time slots can result in financial penalties. Additionally, fixed loading or unloading times may require the special preparation of other resources, such as hiring additional personnel or machinery, like cranes.

Accurate recording of time data, including both the start and end times for loading and unloading, is crucial for efficient logistics and cost management, as is the distinction between flexible hourly ranges and fixed times. This level of detail allows for a comprehensive understanding of the temporal aspects of transport operations, leading to better planning, optimization, and decision-making in the transport sector.

Tab. 6
List of features related to time

Feature name:	Feature meaning:
START_LOAD_TIME	Start time of loading
END_LOAD_ TIME	End time of loading
START_DELIVERY_ TIME	Start time of unloading
END_DELIVERY_ TIME	End time of unloading

In this context, it is essential to understand whether a given road transport vehicle has a specific body type, which is particularly important for analysis purposes. “Body type” refers to the characteristic structural layout and equipment of a vehicle or trailer, which determines its functionality and intended use. In studies on forecasting the price of road freight transport services, considering the body type can significantly impact costs, capacity, efficiency, and the specificity of transport operations.

The impact of body type on price forecasting arises from differences in operational costs, transport capacity, logistical requirements, and adaptation to specific types of cargo. Ultimately, analyzing body type in road freight price forecasting aims to consider the specific characteristics of vehicles concerning the nature of the transported goods. Having precise knowledge about the body type is crucial for a fuller understanding of the factors that influence the pricing of transport services and for constructing more accurate and realistic price

forecasting models. Tab. 7 presents the types of bodies used in transport vehicles. Assuming that these types can occur in any combination, there are numerous possible combinations.

Tab. 7
Trailer types with descriptions

BODY_TYPE	Description
box	This type is characterized by an enclosed space resembling a box designed for transporting various types of goods. It is a popular choice in the logistics and transport sectors as it protects cargo from weather conditions and other external factors, which is crucial for transporting delicate or valuable goods. Box bodies can come in various sizes and configurations depending on the type of cargo and logistical requirements. They may feature doors for easy loading and unloading and systems to secure the cargo during transport.
car_transporter	This specialized road transport unit is designed for transporting cars. It has a structure and mechanisms that allow the simultaneous loading and securing of multiple cars. These trailers are used in the automotive industry to deliver new or used cars to various locations, such as dealerships or auctions.
coil	This is a specialized trailer or transport set designed to transport steel coils. These are widely used in the steel industry and require a special type of trailer with increased stability and security measures for transport.
container_chassis	This is a specialized undercarriage used for transporting containers. It is a structure tailored to place and secure containers on the road, allowing for safe road transport.
drop_side	These trailers feature unique construction with sides that can be lowered, allowing for the easy loading and unloading of goods from the side of the vehicle. This flexibility is particularly useful for goods that cannot be easily loaded on traditional trailers. They also provide security as loaded goods can be securely fastened and prevented from moving during transport.
extendable trailer	This type of trailer can be lengthened to accommodate various types of loads. It allows both short and long loads to be transported and is often used to transport long items such as steel beams, pipes, or building prefabricates. The extending mechanism increases its flexibility and adapts it to different load dimensions.
flatbed	This type of truck is characterized by a flat, open body with no sides or roof, enabling it to transport a wide range of cargo.
inloader	This type of trailer is specifically designed for transporting glass and other sensitive materials. It features advanced technological solutions for the safe and precise transport of fragile cargo. Inloaders are equipped with special

	mechanisms and securing systems to stabilize and protect flat glass surfaces during transport.
jumbo	This type of trailer is characterized by a large cargo space and is designed to hold a greater volume of cargo compared to standard trailers. It is ideal for transporting large-volume goods like lightweight, volumetric products. Jumbo trailers can also feature adjustable roof heights to accommodate specific dimensions of the cargo.
low_loader	This type of trailer has a low profile, enabling the loading and transport of tall and heavy loads, such as construction machinery, agricultural equipment, or large structures. The low loading platform allows the load to be placed closer to the road surface, facilitating the transport of oversized loads.
mega	These trailers feature exceptionally large cargo spaces and high bodies, enabling the transport of taller and larger loads. They typically have a body height of 3 m, making them suitable for transporting tall items like pallets or large structures.
panel_van	This delivery vehicle type has an enclosed cargo area with solid walls, making it ideal for transporting goods that require protection from weather or theft. It is widely used in logistics and delivery services for safe and secure transport.
platform trailer	This type of trailer is used in road transport and is characterized by an open cargo space without fixed bodywork. It is versatile and used for transporting various types of loads, such as pallets, construction equipment, or large items. It is valued for its flexibility in transporting loads of different sizes and shapes.
refrigerator	These trailers are equipped with refrigeration units to maintain a controlled low temperature inside the cargo space, which is crucial for transporting perishable goods or pharmaceuticals. They are essential in the food and pharmaceutical industries to ensure products arrive in proper conditions.
roll_on_roll_of_tipper	This type of vehicle can load and unload itself using a tipping mechanism and is suitable for transporting loose materials. It facilitates efficient unloading by sliding or dumping the cargo, which is useful for materials like construction aggregates or waste.
semi_trailer_with_inclided table	This type of trailer features a tilting mechanism for the cargo platform, aiding in loading and unloading. It is used for transporting equipment or vehicles that need to be loaded at an angle.
silo	These are used for transporting bulk materials like grains, cement, or sand. They feature a specially designed tank that allows for efficient storage and transport of bulk loads.

skip_loader	This trailer type has a system for lifting and transporting skips or containers. It is commonly used for waste management and construction material transport, with hydraulic arms for efficient loading and unloading.
standard	These trailers have flexible sides, often referred to as "curtains," allowing for easy access to the cargo space from the sides. They are versatile and suitable for transporting various types of goods.
swap_body	This vehicle type features interchangeable bodies that can be swapped depending on the type of cargo. It is useful in intermodal transport, as it allows quick changes between different modes of transport.
tank	These trailers are used for transporting liquids, such as fuels, chemicals, or food products. They are sealed and equipped with safety systems to ensure secure transport.
thermo	These trailers are equipped with advanced temperature control systems to maintain specific conditions, which are essential for transporting temperature-sensitive goods.
tipper	These trailers, which are used for transporting loose materials, have a tipping mechanism for easy unloading. They are commonly used in construction and road maintenance.
tractor	This is a truck unit without a trailer. It can be used independently or attached to different types of trailers as needed.
walking_floor	This trailer has a moving floor mechanism for the controlled loading and unloading of bulk or loose materials. It is used for transporting items like grains, wood, or recyclables.
walking_floor_bulk_materials	This is similar to a walking floor and is used for bulk materials.

In road freight transport, the characteristics of the body play a crucial role in determining the vehicle's transport capacity, safety, and functionality.

Tab. 8 presents a list of body characteristics. The body is the structure or construction mounted on a vehicle, which is designed to adapt it to a specific type of cargo or application. Understanding and analyzing body characteristics is a key step in designing and selecting the appropriate vehicle for transporting goods.

It is essential to ensure that the chosen body characteristics are precisely matched to the nature of the cargo and the type of transport activity. An improperly selected body can lead to inefficiencies, difficulties in loading and unloading, and the risk of cargo damage. Moreover, special requirements for the body can affect the cost of the transport service, which is an important economic aspect.

Understanding and accurately saving these body characteristics are crucial for effective logistics management, as they ensure safe and efficient transport and optimize transport costs.

Tab. 8
Body characteristics with descriptions

BODY_CHARACTERISTIC	Description
air_suspension	This is an advanced form of technology that is becoming increasingly popular in the transport industry. This system uses compressed air to assist the vehicle's suspension and shock absorption. The main advantage of pneumatic suspension is its ability to adjust the vehicle's height depending on the load. This means that the vehicle can be adapted to different transport conditions, which is particularly important when transporting loads of varying weights or dimensions. It helps minimize the risk of cargo damage and improves overall ride comfort. Pneumatic suspension is especially useful for transporting goods sensitive to shocks, such as food products or delicate materials. The use of compressed air ensures that the forces acting on the vehicle are evenly distributed when the body is driven on uneven roads or terrain. Additionally, pneumatic suspension can improve the vehicle's traction and stability, which is significant for safety.
back tipper	This is a specific type of body with a distinctive construction. This type of vehicle is mainly used in the construction industry and for transporting bulk materials. The rear tipper is characterized by its ability to tilt and raise the rear container, allowing for the efficient and effective unloading of loose materials such as sand, gravel, rubble, or other bulk materials. The key element of a rear tipper is its controlled tilting mechanism (usually hydraulic), which allows for the even distribution of materials at the construction site or other destination.
code xl	This certificate relates to the structural strength of trailer bodies in road transport. It is an important document certifying compliance with specific standards regarding the structural strength of trailers used for transporting various loads. The certificate ensures that the trailer body design has been tested and meets the minimum strength requirements specified in EN 12642. This standard defines the minimum requirements for trailer superstructures, which must bear a portion of the forces generated by the cargo's mass. If the superstructures cannot fully secure the load, additional measures such as straps, wedges, or anchoring points are necessary to prevent cargo movement.
curtainsider	This sophisticated system allows for precise and controlled opening and closing of the side walls of a trailer or vehicle body. This system uses special tracks and drive mechanisms that enable the panels or curtains to slide along the sides of the vehicle. The main purpose of the curtainsider mechanism is to facilitate quick and

	convenient loading and unloading, especially for goods that do not require any lifting equipment. When the curtain is open, the side walls of the vehicle are rolled up or moved on special tracks, creating a wide side opening. This allows for the efficient loading of goods and access to the cargo inside the vehicle.
double_deck	This solution enhances transport efficiency, especially for goods with specific requirements. This innovative system creates two layers of flooring inside the vehicle that are typically made from lightweight and durable composite materials. The primary purpose of the double floor is to enable the transport of goods under controlled conditions, particularly those requiring specific temperature, humidity, or protection from mechanical damage.
dual evaporator	This cooling system is used in certain types of refrigerated trailers and refrigeration units. This system is equipped with two separate evaporators responsible for cooling two different compartments or sections within a single refrigeration unit. The main advantage of the dual evaporator is its ability to independently control the cooling conditions in both compartments. This means that humidity, temperature, and evaporation cycles can be adjusted separately in each evaporator.
lifting roof	This solution enhances the flexibility and efficiency of goods transport by allowing the roof over the cargo space to be raised or lowered, significantly impacting the loading and unloading process and the overall cargo capacity. The liftable roof is usually controlled via a hydraulic mechanism, allowing for smooth raising and lowering of the structure. This is particularly important for irregularly shaped loads requiring a personalized approach to transport.
side tipper	This mechanism enables the dumping of cargo from the side of the vehicle. This system is particularly useful for transporting bulk materials such as sand, gravel, soil, or similar loads. The side dumping process is precise and controlled, allowing for the accurate unloading of the cargo.
sliding roof	Sliding roofs in trucks are innovative solutions that significantly increase the flexibility of the loading and unloading process. The mechanism operates like a concertina structure, allowing the roof of the vehicle to open. The main purpose of the sliding roof is to facilitate easy access to and loading of cargo through the top opening.
widenable	Also known as width-expandable trailers, these innovative solutions in heavy transport allow for the regulation of the body width, which is particularly useful for transporting goods of varying dimensions and during loading and unloading.

An important aspect of forecasting prices for road freight transport services is understanding the diversity of vehicle types used in the transport industry. Various vehicle types, such as rigid trucks (solo vehicles) and tractor-trailers, encompass vehicles with different transport capacities and purposes. These differences in vehicle types affect operating costs, fuel efficiency, and transport capacities. As a result, the prices of transport services depend on the characteristics of each vehicle type. Below, the key features and parameters of each vehicle type are discussed to understand how their differences impact the process of forecasting prices for road freight transport services. Tab. 9 presents the types of vehicles and their descriptions.

Tab. 9

Vehicle types with descriptions

VIHECLE_TYPE	Description
articulated	Among the various vehicle types used in the transport industry, one important category is the articulated truck. This type of vehicle consists of two main parts: a tractor unit and a trailer. These two parts are connected by a special mechanism that allows for pivoting movement, which provides greater maneuverability. This type of vehicle is characterized by its large transport capacity, making it commonly used for long-distance freight transport. It also has a significant payload capacity, allowing it to carry more cargo than other vehicle types.
rigid_with_trailer	In the analysis of different vehicles used in the transport sector, an essential element is the rigid truck with a trailer, also known as a combination vehicle. This form of vehicle consists of two parts: a rigid truck and a trailer, which are connected. This configuration provides increased transport capacity and flexibility in road transport. Combination vehicles are often used for longer routes and the transport of larger loads, making them an important element in forecasting prices for road freight transport services.
<12.5t.	Also known as a light truck, this is one of the important vehicle types used in the transport sector. It is characterized by a total weight that does not exceed 12.5 tons, including the weight of the vehicle and the cargo being transported. This type of vehicle is used in various logistical operations, both in freight transport and delivery services. Due to their relatively low weight, these vehicles are often preferred in situations requiring flexibility and maneuverability, especially in urban areas.
<7.5t.	Often referred to as a light truck, this is a key element in the field of transport and logistics. It is characterized by a limited total weight that does not exceed 7.5 tons, considering the weight of the vehicle and the cargo being transported. This type of vehicle is used in various applications, both in goods delivery and logistical operations. Trucks with a total weight of up to 7.5 tons

	are particularly popular in urban traffic, where greater maneuverability and flexibility are required. Due to their low weight, they often do not fall under strict regulations for heavy vehicles, allowing for more flexible logistical operations in areas with limited space.
<3.5t.	Also known as a light truck, this type of vehicle plays a significant role in the field of transport and logistics. This type of vehicle is characterized by a limited total weight that does not exceed 3.5 tons, including the weight of the vehicle and the cargo being transported. Light trucks are used in many different sectors, including goods delivery, courier services, and logistical operations. An important aspect of trucks with a total weight of up to 3.5 tons is their maneuverability and ability to move freely in areas with limited space. Due to their low weight, some of these vehicles can be driven by drivers with a standard category B driving license. The limited total weight also affects the ability to transport goods, which is significant for urban deliveries and courier services.

The loading/unloading method is a crucial component of the transportation process, impacting the operational efficiency and safety of the transported goods. This section discusses various loading/unloading methods in road freight transport. Available options include top loading, side loading, and rear loading.

Tab. 10 presents various loading/unloading methods, which are dictated by the type of vehicle body or the infrastructure of the loader or unloader.

Each loading/unloading method has specific advantages and is suitable for different types of cargo and operational requirements. The choice of method can significantly influence the efficiency and safety of the transportation process. Understanding and selecting the appropriate method based on the vehicle type and infrastructure can lead to more effective logistics operations and optimized transport costs.

Tab. 10
Loading/unloading methods

LOAD_UNLOAD_METHOD	Description
back	Backloading involves loading and unloading goods through the rear of the vehicle. It is one of the most commonly used methods, particularly for standard trailers. This method allows for efficient and quick operations but may require access to suitably adapted facilities. Rear loading is advantageous for streamlined loading processes, especially in warehouses with designated loading docks.
side	Side loading entails loading and unloading goods through the side doors or walls of the vehicle. This method is preferred for large goods or when easy access to the cargo is necessary during loading/unloading operations. Side

	loading allows for efficient use of space and facilitates the handling of bulky items, making it ideal for operations where rapid access and frequent handling of goods are required.
top	Top loading involves loading and unloading goods through the top of the vehicle. This method is popular for transport units with roofs that can be opened or lifted. It is often used for cargo that has a non-standard shape or requires special protection. Top loading offers the advantage of utilizing the vertical space of the vehicle, making it suitable for oversized or irregularly shaped items.

An essential aspect of road transport is properly securing cargo to ensure safety during transit and minimize the risk of damage or loss. Tab. 11 presents the available values for the cargo securing attribute. Cargo security is a key priority in the transport process. Securing cargo aims to ensure that transported goods do not get damaged during transit and to minimize the risk of road accidents related to improper load fastening or arrangement. Safe cargo securing is not only a responsibility of the carrier but also builds customer trust in transport services.

The diversity of goods that can be transported necessitates various securing techniques and tools. Each type of cargo may require different methods of fastening, shock protection, or weatherproofing. Properly tailoring securing methods to the specific characteristics of the cargo is crucial for maintaining its integrity.

Additionally, special requirements for securing cargo can impact transport costs. The need for advanced techniques or materials for securing can increase operational costs. However, this investment in security can prevent losses and maintain a reputation as a reliable and professional transport service provider:

Choosing the right stakes is crucial for effectively securing cargo and minimizing damage risk. Stakes are particularly important in preventing bulk or loose loads from shifting during transit. They are also essential for irregularly shaped or sized loads, ensuring stability and integrity. Using stakes can significantly impact the loading and unloading process, requiring proper placement and fastening to ensure effective securing. The decision to use stakes should be based on the type of goods being transported, their characteristics, and safety requirements. The cost impact of this decision on transport costs should also be considered in the planning process.

Tab. 11

Features of load securing

LOAD_SECURING	Description
anti_slip mats	One of the key methods for securing cargo during road transport is using anti-slip mats. Anti-slip mats provide an effective strategy to prevent the movement of goods during transit, which can lead to damage and pose a safety hazard. These mats are typically made of rubber, plastics, or other materials with anti-slip properties. Their surface features a special pattern or texture that provides traction and grip between the mat, the cargo surface, and the vehicle's floor. The main advantage of anti-slip mats is their versatility. They can be used for various types of goods, regardless of their shape or size. These mats are particularly useful for transporting goods with smooth or

	slippery surfaces that may easily shift and become damaged during transport. Ensuring the proper placement and securing of anti-slip mats minimizes the risk of cargo shifting, which is crucial for protecting the goods and ensuring road safety and compliance with transportation regulations. The proper choice and use of anti-slip mats depend on the type of cargo, weather conditions, and vehicle characteristics. Incorporating anti-slip mats into the cargo-securing process enhances safety and minimizes the risk of damage and loss during transit.
edge_protection	Edge protectors, also known as corner protectors, are effective for protecting goods from damage due to friction and impacts during transit. These protectors are typically made of durable materials such as metal, plastic, or wood, which can shield the sharp edges of goods from abrasions and dents. Their design is usually tailored to the shape and dimensions of the cargo, ensuring a secure fit between the protectors and the cargo's surface. The primary goal of using edge protectors is to safeguard goods with sharp or delicate edges that are prone to damage from friction or improper placement on the transport surface. Edge protectors also protect against accidental impacts and vibrations that may occur during road transport. They prevent the cargo from coming into direct contact with the vehicle's surface or other goods, thus minimizing the risk of damage and deformation. Properly planning and securing edge protectors is essential for effectively protecting the cargo and enhancing overall transport safety. For certain types of goods, where maintaining the integrity of edges is crucial, using edge protectors can influence transport costs, which will be further discussed in the context of road freight transport pricing.
lashing_chains	Effective cargo securing during road transport involves chain lashings. These flexible elements are made of durable metal, such as stainless or hardened steel, and are used to fasten goods to the vehicle or trailer floor. Chain lashings allow for adjustable tension, stabilizing the cargo and minimizing its movement during transit. They are particularly useful for securing irregularly shaped or dimensioned loads that may not fit standard securing systems. The proper application of chain lashings helps prevent cargo shifting, deformation, or overturning, ensuring both the safety of the goods and road safety. It is crucial to choose the appropriate length, quantity, and fastening method of chain lashings based on the cargo's characteristics and vehicle type. Properly used chain lashings are indispensable for effective cargo securing and can impact the costs and safety of the entire transport process.
lashing_straps	These straps provide a versatile and effective method for securing various types of goods on trucks and trailers. They are made of durable materials like polyester, nylon, or steel and feature adjustable tension mechanisms. Depending on the type of cargo, different types of tie-down straps can be used, such as straps with hooks, buckles, or ratchet mechanisms. Ensuring the stability and safety of cargo during transport is critical, and the proper securing of goods with tie-down straps is the legal, ethical, and professional responsibility of

	the carrier. Different types of cargo require specific approaches and appropriate strap selection to ensure stability, prevent damage, and minimize the risk of accidents during transit. The choice and proper use of tie-down straps, along with their quantity and fastening method, can influence transport costs. Some loads may require advanced straps or additional accessories, impacting the overall cost.
locking_bar	These are steel structures mounted around the pallet on the vehicle or trailer platform. Their main function is to lock the pallet in place during maneuvers and transit. These bars are designed to effectively hold the pallet, minimizing the risk of damage or overturning. Ensuring the safety of palletized loads is a priority in road transport, and pallet-blocking bars provide an effective solution to minimize the risk of accidents and cargo damage. Properly installing and adjusting the bars to the specific load is crucial for their securing function. The cost of transport often depends on the securing method used; pallet blocking bars can affect transport costs due to their securing function.
pallet_retaining_bar	These are metal or plastic elements equipped with numerous regularly spaced holes that are mounted on the walls of the trailer or vehicle. Their main purpose is to provide flexible securing of various loads. E-track systems are highly versatile, and they adapt to the shape and size of the transported goods. By enabling the attachment of straps, ropes, or tie-downs in different positions, E-track systems create a flexible securing system that minimizes the risk of shifting or damaging the load during transport. The adjustable nature of E-track systems allows for variable attachment points, ensuring secure load fastening to prevent movement during transit. Their durability and load-bearing capacity make them indispensable for ensuring the integrity of transported goods. The use of E-track systems provides carriers with the flexibility to adjust securing to the specific requirements of each load, maintaining stability and integrity during transport. However, E-track systems, like other securing methods, can impact transport costs, as installation and additional materials involve extra effort and time.
perforated_batten	These are used in specialized transport applications, and they incorporate battens (narrow strips of material) with a series of perforations or holes. This design offers unique advantages for specific types of cargo, particularly those requiring enhanced ventilation or drainage.
side_panels	Sideboards are liftable or removable side walls of the vehicle that secure the load during transport. Depending on the type and characteristics of the load, sideboards can be adjusted in height or position, enabling the vehicle's structure to match the transported goods' dimensions. Sideboards can be opened or removed, facilitating loading and unloading from the sides of the vehicle. Securing cargo with sideboards ensures stability and minimizes the risk of damage during transit. Sideboards act as physical barriers, preventing the load from falling off or shifting within the vehicle. For certain types of loads, such as bulk materials or loose goods, sideboards are indispensable for preventing loss and damage. The

	use of sideboards as a securing method also impacts the loading and unloading process, with opening or lifting sideboards facilitating access to the cargo area from the sides of the vehicle.
stanchions	These special metal or plastic rods, beams, or rails are placed in the cargo area of the trailer or vehicle. Their main purpose is to lock the load in place, preventing it from shifting, tipping, or being damaged during transit. Stakes are adjustable in length and height depending on the load type and the cargo area's dimensions.

Tab. 12 lists other characteristics that do not fall under any of the previously mentioned categories, including the height of cargo (in meters), the required trailer width (in meters), the required loading meters, the required pallet exchange quantity, and the required minimum and maximum temperature in the vehicle during transport.

Tab. 12
List of other features

Feature:	Description
TONS	This is a crucial parameter in road freight transport. It refers to the mass or weight of the cargo transported by heavy vehicles. Weight in tons plays a vital role as it directly impacts several aspects of transportation. Primarily, it determines how much cargo can be transported by a given vehicle, considering its payload capacity. This is essential for drivers, who must adhere to regulations regarding overloading, and for transport companies, which must optimize the utilization of their vehicles. Moreover, weight in tons is significant from a cost and fee perspective. Many roads, especially highways and bridges, implement toll systems often based on the vehicle's weight, which can significantly impact transport costs. Additionally, fuel consumption by heavy vehicles is largely influenced by the load, thus affecting operational costs. Weight in tons is also a standard measure internationally, facilitating international cargo transport, where uniform measurement standards are required. It is essential for road safety, as overloaded vehicles can pose a hazard to other road users and infrastructure.
HEIGH	This term refers to the vertical distance between the base and the top of the cargo or its packaging, expressed in meters. It is an important criterion affecting various aspects of the transport process. Height is crucial in determining whether a particular load will fit within a vehicle or trailer. Road transport regulations and infrastructure, such as overpasses, tunnels, and bridges, impose height restrictions. Thus, considering cargo height is essential when planning transport routes. Additionally, height impacts transport safety. High-placed cargo can affect the stability of the vehicle, especially during maneuvers or cornering. The load must be properly secured and stabilized to avoid the risk of vehicle tipping or cargo damage. Height also influences the vehicle's load capacity. Volume and height determine how much cargo can be transported. Transport companies must optimize vehicle load space utilization to increase transport efficiency and minimize operational costs.
WIDTH	This refers to the minimum distance between the trailer's side walls, measured in meters. Required trailer width is crucial for proper cargo transport and compliance with road transport regulations and standards. A trailer’s width must comply with regulations for legal use on public roads.

	Violating these standards can lead to legal issues and pose a risk to road safety.
LDM	"Required loading meters" refers to the amount of available space or length in a trailer or transport vehicle designated for the cargo. This parameter is crucial in road transport, influencing carrying capacity and transport efficiency. Measuring the available space in loading meters ensures the cargo has adequate room and can be transported safely and legally. The required loading meters can vary based on cargo type and its size and dimensions. Transport companies must accurately consider this parameter for efficient transport planning. Available loading meters can be restricted by transport norms and regulations, including maximum permissible load and safety requirements. Therefore, proper determination and consideration of this value are essential for legal and safe transport operations.
EPALE	This refers to the number of empty pallets that the carrier must leave at the loading site as part of the transport process. This is a significant element of transport and logistics organization, ensuring smooth and efficient loading and unloading processes. The required pallet exchange quantity varies based on the type of goods, their dimensions, and the industry's logistic standards. Carriers must provide the necessary number of pallets to facilitate smooth and trouble-free exchange of goods and contribute to transport efficiency.
TEMP_MIN	This refers to the temperature range that must be maintained inside the transport vehicle during cargo transit. This is especially crucial for transporting temperature-sensitive goods such as food products, medicines, or chemicals. Maintaining the required temperature range ensures the integrity and quality of transported goods, preventing spoilage or damage. For food products, specific temperature control is necessary to ensure safety and compliance with health regulations. In pharmaceuticals, precise temperature control is essential to maintaining the efficacy and stability of medications. Adhering to the required temperature range often necessitates specialized equipment, such as refrigerated trailers or climate-controlled containers. This can impact transport costs, as maintaining specific temperature conditions requires additional energy and resources. Considering the required temperature range is essential for ensuring the safe and effective transport of temperature-sensitive goods. Transport companies must ensure that their vehicles meet these temperature requirements, safeguarding cargo quality and compliance with relevant regulations.
TEMP_MAX	

4.2. Data Transformation Methods

Data processing methods are a cornerstone in developing a price forecasting model for road freight services, as they prepare raw data for analysis and modelling. This process involves transforming, standardizing, cleaning, and structuring the data to ensure it is suitable for predictive modelling. Data standardization focuses on ensuring consistency and comparability, particularly when working with information from diverse sources or different measurement units. This includes scaling and normalization to adjust feature scales, unit conversion to standardize measurements such as distances into kilometres, and categorical encoding, which transforms qualitative data into numerical formats using techniques like one-hot or label encoding.

Data cleaning identifies and corrects errors or inconsistencies to ensure a dataset is accurate and reliable. This process involves handling missing values through methods like mean substitution, removing duplicate records to prevent biased analysis, and resolving discrepancies such as variations in spelling, date formats, or location entries. Feature engineering plays a critical role in improving the model's predictive power by creating new variables or modifying existing ones. Key techniques include generating interaction terms, extracting temporal patterns from date fields, and aggregating data to summarize trends over specific periods such as weeks or months.

Data integration combines information from various sources into a unified dataset. This involves merging datasets based on common identifiers, enriching datasets by joining supplementary information, and addressing schema differences to ensure seamless integration of data with diverse structures. Data validation ensures the processed data are accurate, reliable, and ready for modelling. This includes consistency checks to confirm that transformations and integrations are correctly applied, statistical validation to verify that the data aligns with modelling assumptions, and cross-validation to evaluate the model's performance and prevent overfitting.

The effectiveness of data processing directly impacts the accuracy and reliability of a price forecasting model. By meticulously standardizing, cleaning, engineering features, integrating, and validating the data, researchers can build models on a solid foundation of high-quality information. These carefully executed steps enable the model to capture the complexities of pricing in the road freight transportation industry, leading to more precise and actionable forecasts.

Features related to distance are numerical, so the process of handling these data is not complex. The processing of these data is outlined in Tab. 13. Values that are already assigned remain unchanged.

Filling in all distance values for each country would be time-consuming. Therefore, when creating raw data, only non-zero values need to be filled in. During the data processing stage, empty values are filled using the following function: pandas.DataFrame.fillna(0).
A new feature, the sum of all kilometres, is generated by summing all columns related to distance. This aggregated feature will be crucial for developing further features that are important in subsequent processes.

The feature "OTHER_COSTS" is a significant variable in forecasting road freight transport service prices. It encompasses various costs associated with transportation that are not directly related to the distance between the loading and unloading points but still impact the final price of the service. These costs may include tolls for tunnels, bridges, or ferry crossings.

Tolls for tunnels, bridges, and ferries are not typically considered in standard price forecasting models based solely on distance. However, they are significant, especially in regions where transport involves crossing geographical barriers that can only be traversed via such facilities. In the analysis and forecasting of prices, including these additional costs can lead to more accurate predictions and better reflect actual transportation expenses. This is crucial for transport companies, which need to price their services correctly, and clients, who want accurate estimates of their transport costs.

Considering the cost per kilometre [€/km] when forecasting road freight transport service prices is preferred over using the total amount in euros (€) for several reasons. The cost per kilometre is a more precise indicator of costs, as it directly relates to the distance travelled. Expressing transport costs only in euros does not account for variations in route lengths, potentially leading to inaccuracies in cost estimation.

Incorporating the cost per kilometre allows for the differentiation of routes. Not all routes are identical—some may be longer, more complex, or require more effort. The cost per kilometre helps adjust the pricing according to the route's specific parameters. Expressing costs in euros per kilometre increases transparency, enabling a clearer understanding of the expenses associated with covering a particular distance. This transparency helps clients and transport companies accurately estimate service costs.

By considering the cost per kilometre, transport companies and clients can better plan their routes and transport strategies. Doing so makes the management of resources and costs more efficient. Such an approach objectively compares how other features impact the price, as described in subsequent sections. Before calculating the cost per kilometre, other costs are subtracted from the total cost. Equation (7) presents the formula for calculating this feature.

Tab. 13

Distance data processing

Feature (raw data)	Feature (transformed)
AT_KM	AT_KM
BE_KM	BE_KM
BG_KM	BG_KM
HR_KM	HR_KM
CY_KM	CY_KM
CZ_KM	CZ_KM
DK_KM	DK_KM
EE_KM	EE_KM
FI_KM	FI_KM
FR_KM	FR_KM
GR_KM	GR_KM
ES_KM	ES_KM
IE_KM	IE_KM
LT_KM	LT_KM
LU_KM	LU_KM
LV_KM	LV_KM
MT_KM	MT_KM
NL_KM	NL_KM
DE_KM	DE_KM
PL_KM	PL_KM
PT_KM	PT_KM
RO_KM	RO_KM
SK_KM	SK_KM
SI_KM	SI_KM
SE_KM	SE_KM
HU_KM	HU_KM
IT_KM	IT_KM
	KM
EURO	EURO
OTHER_COSTS	OTHER_COSTS
	NETTO_EURO_FOR_KM

$$\text{NETTO\_EURO\_FOR\_KM} = \frac{(\text{EURO} - \text{OTHER\_COSTS})}{\text{KM}}. \quad (7)$$

In freight transport, the term "relations" refers to the unique combination of the loading and unloading countries. This feature captures the pairwise connection between the origin and destination, making it a crucial component for understanding and modelling the pricing structure within the industry.

The transformation of relationship data involves several steps designed to standardize, encode, and enhance these features for effective use in machine learning models.

Categorical encoding converts the categorical data representing loading and unloading countries into numerical formats using techniques such as one-hot encoding or label encoding, ensuring the model can process these features effectively and recognize their influence on pricing.

Combination features are created by concatenating the encoded values of the loading and unloading countries, generating a new variable that uniquely identifies each relationship and serves as a distinct identifier for every route.

Properly handling missing data ensures that all potential relationships are represented in the dataset, even if some are initially absent, by imputing missing values through predefined strategies like the mean or median of existing relationships.

Standardization normalizes the relationship features to ensure consistent scaling, which contributes to improved model performance and comparability across diverse relationships.

By applying these transformation steps, the relationship features are systematically structured to accurately represent the unique combinations of loading and unloading locations, facilitating the development of a robust and reliable pricing model for road freight transport. The detailed process is presented in Tab. 14.

Tab. 14
Processing of relationship data

Feature (raw data)	Feature (transformed)
CODE_LOAD_PLACE	COUNTRY_LOAD_PLACE_FACTORIZED
	MEAN_LOAD_PLACE
	MEDIAN_LOAD PLACE
	STD_LOAD_PLACE
CODE_DELIVERY_PLACE	CONTRY_DELIVERY_PLACE_FACTORIZE
	MEAN_DELIVERY_PLACE
	MEDIAN_DELIVERY_PLACE
	STD_DELIVERY_PLACE
CODE_LOAD_PLACE +CODE_DELIVERY_PLACE	RELATION_FACTORIZED
	RELATION_MEAN
	RELATION_MEDIAN
	RELATION_STD

Tab. 15 presents the process of transforming date-related data, which is crucial for analyzing historical price trends and understanding the factors influencing transport service pricing. Examining historical data helps identify whether transport prices are subject to seasonal fluctuations or long-term trends, providing valuable insights that support more accurate future forecasts. In industries such as agricultural goods transport, in which prices can vary

significantly with the seasons, date information is essential for pinpointing these periods of increased demand, which often lead to higher service prices.

Date-related data are divided into loading and unloading dates, each further split into start and end dates, resulting in four primary features. From these, additional attributes are derived to capture temporal patterns, including the day of the month, day of the year, day of the week, week of the year, month, and year. These transformed features are numerical, meaning no factorization is applied.

The transformation process involves precise formatting and parsing of loading and unloading dates to ensure consistency. Temporal features are systematically extracted from the original date fields, enabling the identification of specific trends and patterns. Missing date values are handled using imputation methods such as forward-fill, backward-fill, or median imputation, ensuring no gaps remain in the temporal dataset. These steps collectively enhance the dataset's robustness, supporting the development of accurate and reliable pricing models for road freight transport.

Tab. 15
Date data processing

Feature (raw data)	Feature (transformed data)
START_LOAD_DATA	START_LOAD_DATA_DAY
	START_LOAD_DATA_DAY_IN_YEAR
	START_LOAD_DATA_WEEKDAY
	START_LOAD_DATA_WEEK_IN_YEAR
	START_LOAD_DATA_MONTH
	START_LOAD_DATA_YEAR
END_LOAD_DATA	END_LOAD_DATA_DAY
	END_LOAD_DATA_DAY_IN_YEAR
	END_LOAD_DATA_WEEKDAY
	END_LOAD_DATA_WEEK_IN_YEAR
	END_LOAD_DATA_MONTH
	END_LOAD_DATA_YEAR
START_DELIVERY_DATA	START_DELIVERY_DATA_DAY
	START_DELIVERY_DATA_DAY_IN_YEAR
	START_DELIVERY_DATA_WEEKDAY
	START_DELIVERY_DATA_WEEK_IN_YEAR
	START_DELIVERY_DATA_MONTH
	START_DELIVERY_DATA_YEAR
END_DELIVERY_DATA	END_DELIVERY_DATA_DAY
	END_DELIVERY_DATA_DAY_IN_YEAR
	END_DELIVERY_DATA_WEEKDAY
	END_DELIVERY_DATA_WEEK_IN_YEAR
	END_DELIVERY_DATA_MONTH
	END_DELIVERY_DATA_YEAR

The method described above is the first step, after which the process must proceed to the proper feature creation using statistical methods. presents data analysis based on the mean, median, and standard deviation of the price per kilometre, particularly in the context of the initial loading date, which holds significant importance and implications. The calculation of statistical measures begins with the mean, which represents the average price per kilometre for

transport services starting on specific dates, providing insights into typical pricing trends. The median offers the middle value in the distribution of prices per kilometre, making it less sensitive to outliers and more reflective of central tendencies in skewed datasets. The standard deviation captures the variability of the price per kilometre around the mean, highlighting how much prices fluctuate over time and offering an indicator of the reliability of the mean as a measure of central tendency.

Flexibility in loading dates is an additional factor considered in this analysis. When the exact loading date is not critical, a flexible date can be selected, allowing transport companies to adapt operations more effectively to dynamic circumstances. This flexibility has a direct impact on pricing, as it often results in lower costs due to optimized fleet utilization and reduced expenses associated with empty runs. Together, these statistical insights and operational strategies contribute to a more accurate and efficient pricing model for road freight transport.

Tab. 16 presents data analysis based on the mean, median, and standard deviation of the price per kilometre, particularly in the context of the initial loading date, which holds significant importance and implications. The calculation of statistical measures begins with the mean, which represents the average price per kilometre for transport services starting on specific dates, providing insights into typical pricing trends. The median offers the middle value in the distribution of prices per kilometre, making it less sensitive to outliers and more reflective of central tendencies in skewed datasets. The standard deviation captures the variability of the price per kilometre around the mean, highlighting how much prices fluctuate over time and offering an indicator of the reliability of the mean as a measure of central tendency.

Flexibility in loading dates is an additional factor considered in this analysis. When the exact loading date is not critical, a flexible date can be selected, allowing transport companies to adapt operations more effectively to dynamic circumstances. This flexibility has a direct impact on pricing, as it often results in lower costs due to optimized fleet utilization and reduced expenses associated with empty runs. Together, these statistical insights and operational strategies contribute to a more accurate and efficient pricing model for road freight transport.

Tab. 16

Data processing for the initial loading date

Feature (data after 1st transformation)	Feature (data after 2nd transformation)
START_LOAD_DATA_DAY	START_LOAD_DATA_DAY_MEAN
	START_LOAD_DATA_DAY_MEDIAN
	START_LOAD_DATA_DAY_STD
START_LOAD_DATA_DAY_IN_YEAR	START_LOAD_DATA_DAY_IN_YEAR_MEAN
	START_LOAD_DATA_DAY_IN_YEAR_MEDIAN
	START_LOAD_DATA_DAY_IN_YEAR_STD
START_LOAD_DATA_WEEKDAY	START_LOAD_DATA_WEEKDAY_MEAN
	START_LOAD_DATA_WEEKDAY_MEDIAN
	START_LOAD_DATA_WEEKDAY_STD
START_LOAD_DATA_WEEK_IN_YEAR	START_LOAD_DATA_WEEK_IN_YEAR_MEAN
	START_LOAD_DATA_WEEK_IN_YEAR_MEDIAN
	START_LOAD_DATA_WEEK_IN_YEAR_STD
START_LOAD_DATA_MONTH	START_LOAD_DATA_MONTH_MEAN
	START_LOAD_DATA_MONTH_MEDIAN

	START_LOAD_DATA_MONTH_STD
START_LOAD_DATA_YEAR	START_LOAD_DATA_YEAR_MEAN
	START_LOAD_DATA_YEAR_MEDIAN
	START_LOAD_DATA_YEAR_STD

Tab. 17 outlines the process of transforming date-related data. Time-related data enable the analysis of past price trends. Historical data can be used to identify whether the prices of transport services are subject to seasonal changes or if long-term price trends can be observed. This facilitates a better understanding of the factors influencing prices and helps to generate future forecasts. In some cases, such as in the transportation of agricultural products, prices can significantly fluctuate across seasons. The date is crucial for identifying these seasonal periods when the demand for transport services might be higher, and consequently, prices may increase.

The data are divided into loading and unloading dates. Each of these is further divided into start and end dates, resulting in four features. New features are created from each of these four by deriving the day of the month, day of the year, day of the week, week of the year, month, and year. Since the transformed data are numerical, factorization is not applied to date features.

Tab. 17

Processing of data regarding the final loading date

Feature (data after 1st transformation)	Feature (data after 2nd transformation)
END_LOAD_DATA_DAY	END _LOAD_DATA_DAY_MEAN
	END _LOAD_DATA_DAY_MEDIAN
	END _LOAD_DATA_DAY_STD
END _LOAD_DATA_DAY_IN_YEAR	END _LOAD_DATA_DAY_IN_YEAR_MEAN
	END _LOAD_DATA_DAY_IN_YEAR_MEDIAN
	END _LOAD_DATA_DAY_IN_YEAR_STD
END _LOAD_DATA_WEEKDAY	END _LOAD_DATA_WEEKDAY_MEAN
	END _LOAD_DATA_WEEKDAY_MEDIAN
	END _LOAD_DATA_WEEKDAY_STD
END _LOAD_DATA_WEEK_IN_YEAR	END _LOAD_DATA_WEEK_IN_YEAR_MEAN
	END _LOAD_DATA_WEEK_IN_YEAR_MEDIAN
	END _LOAD_DATA_WEEK_IN_YEAR_STD
END _LOAD_DATA_MONTH	END _LOAD_DATA_MONTH_MEAN
	END _LOAD_DATA_MONTH_MEDIAN
	END _LOAD_DATA_MONTH_STD
END _LOAD_DATA_YEAR	END _LOAD_DATA_YEAR_MEAN
	END _LOAD_DATA_YEAR_MEDIAN
	END _LOAD_DATA_YEAR_STD

Tab. 18 outlines the processing of data related to the initial unloading date. The initial unloading date can be influenced by various factors, such as the reopening of a factory after a shutdown. Another example is the requirement to free up enough warehouse space to accommodate the unloading.

Tab. 18

Data processing for the initial unloading date

Feature (data after 1st transformation)	Feature (data after 2nd transformation)
START_DELIVERY_DATA_DAY	START_DELIVERY _DATA_DAY_MEAN
	START_DELIVERY _DATA_DAY_MEDIAN
	START_DELIVERY _DATA_DAY_STD
START_DELIVERY_DATA_DAY_IN_YEAR	START_DELIVERY _DATA_DAY_IN_YEAR_MEAN
	START_DELIVERY_DATA_DAY_IN_YEAR_MEDIAN
	START_DELIVERY _DATA_DAY_IN_YEAR_STD
START_DELIVERY _DATA_WEEKDAY	START_DELIVERY _DATA_WEEKDAY_MEAN
	START_DELIVERY _DATA_WEEKDAY_MEDIAN
	START_DELIVERY _DATA_WEEKDAY_STD
START_DELIVERY_DATA_WEEK_IN_YEAR	START_DELIVERY _DATA_WEEK_IN_YEAR_MEAN
	START_DELIVERY_DATA_WEEK_IN_YEAR_MEDIAN
	START_DELIVERY _DATA_WEEK_IN_YEAR_STD
START_DELIVERY _DATA_MONTH	START_DELIVERY _DATA_MONTH_MEAN
	START_DELIVERY _DATA_MONTH_MEDIAN
	START_DELIVERY _DATA_MONTH_STD
START_DELIVERY _DATA_YEAR	START_DELIVERY _DATA_YEAR_MEAN
	START_DELIVERY _DATA_YEAR_MEDIAN
	START_DELIVERY _DATA_YEAR_STD

Tab. 19 outlines the processing of data related to the final unloading date. The methodology for transforming these data is similar to that used for loading dates. The final unloading date can be influenced by factors such as the depletion of products or semi-finished goods needed for production.

Tab. 19
Processing of unloading end date data

Feature (data after 1 transformation)	**Feature (data after 2nd transformation)**
END_DELIVERY_DATA_DAY	END _DELIVERY _DATA_DAY_MEAN
	END _DELIVERY _DATA_DAY_MEDIAN
	END _DELIVERY _DATA_DAY_STD
END_DELIVERY_DATA_DAY_IN_YEAR	END_DELIVERY _DATA_DAY_IN_YEAR_MEAN
	END_DELIVERY _DATA_DAY_IN_YEAR_MEDIAN
	END _DELIVERY _DATA_DAY_IN_YEAR_STD
END_DELIVERY_DATA_WEEKDAY	END_DELIVERY _DATA_WEEKDAY_MEAN
	END_DELIVERY _DATA_WEEKDAY_MEDIAN
	END _DELIVERY _DATA_WEEKDAY_STD
END_DELIVERY_DATA_WEEK_IN_YEAR	END_DELIVERY _DATA_WEEK_IN_YEAR_MEAN
	END_DELIVERY_DATA_WEEK_IN_YEAR_MEDIAN
	END_DELIVERY _DATA_WEEK_IN_YEAR_STD
END _DELIVERY _DATA_MONTH	END_DELIVERY _DATA_MONTH_MEAN
	END_DELIVERY _DATA_MONTH_MEDIAN
	END _DELIVERY _DATA_MONTH_STD
END _DELIVERY _DATA_YEAR	END _DELIVERY _DATA_YEAR_MEAN
	END _DELIVERY _DATA_YEAR_MEDIAN
	END _DELIVERY _DATA_YEAR_STD

Transforming data related to loading and unloading times is a critical step in analyzing transport data and forecasting transportation service prices. Tab. 20 outlines the methodology for transforming time-related data, focusing on the calculations of loading time, unloading time, and cumulative time windows for both processes. These raw values are directly utilized to ensure accuracy, as transportation service prices can vary significantly depending on the

specific time of day or time window. Transforming these data enables precise analysis of variations, facilitating a better understanding of when the highest or lowest prices occur.

The significance of time data transformation lies in its ability to provide an accurate analysis of price variations by allowing a detailed examination of price fluctuations based on specific hours or time windows. This information helps identify optimal times for scheduling transport to minimize costs. Improved forecasting is achieved by understanding temporal patterns in transportation prices, leading to more precise predictions and enhanced logistics planning and cost management. Time data transformation also promotes operational efficiency by optimizing scheduling and resource allocation, reducing idle times, and enhancing overall efficiency. Furthermore, accurate time predictions improve service reliability and customer satisfaction by ensuring timely deliveries.

Tab. 20
Detailed time data transformation process

Feature (Raw Data)	Description	Transformation Method
START_LOAD_TIME	Start time of the loading process	Direct extraction from raw data
END_LOAD_TIME	End time of the loading process	Direct extraction from raw data
START_DELIVERY_TIME	Start time of the unloading process	Direct extraction from raw data
END_DELIVERY_TIME	End time of the unloading process	Direct extraction from raw data

Transforming data related to the body type of freight vehicles is a crucial step in analyzing transportation data. Initially, body type is represented as a textual variable, which is unsuitable for direct use in machine learning models that require numerical input. Converting this variable into a numerical format is essential for its integration into analysis and forecasting models.

The transformation process involves factorization, where each unique body type is assigned a corresponding integer, enabling its inclusion in analytical and predictive models. Additionally, statistical measures such as mean, median, and standard deviation are introduced as new features to describe the historical behaviour of specific vehicle groups with a given body type. These measures provide valuable insights into how body types impact transportation costs and efficiency, aiding in modelling future pricing and operational patterns.

A detailed understanding of the influence of vehicle body types on transportation prices can be achieved by incorporating these statistics. This transformation enhances the predictive power of models by capturing historical trends and variances in transportation costs. Furthermore, it supports comprehensive analysis, enabling logistics companies to make informed decisions about vehicle selection and routing, which can optimize operations, reduce costs, and improve service quality.

This systematic transformation of body type data significantly enhances the analytical and predictive capabilities of transportation models, contributing to more effective logistics management. Tab. 21 summarizes the process and demonstrates how raw textual data are converted into numerical and statistical features for modelling.

Tab. 21
Body type data transformation

Feature (Raw Data)	Feature (Transformed Data)
BODY_TYPE	BODY_TYPE_FACTORIZED
	BODY_TYPE_MEAN
	BODY_TYPE_MEDIAN
	BODY_TYPE_STD

Processing data related to body characteristics in freight transport is critical in the analysis and optimization of logistical operations. The methods for processing these characteristics aim to extract significant features that can impact transport efficiency, cargo safety, and operational costs. Tab. 22 outlines the process of feature creation related to body characteristics. From each raw feature, four new features are derived: factorized feature, mean, median, and standard deviation within the group.

Tab. 22
Transformation of body characteristic data

Feature (Raw Data)	Feature (Transformed Data)
BODY_CHARACTERISTIC	BODY_CHARACTERISTIC _FACTORIZED
	BODY_CHARACTERISTIC _MEAN
	BODY_CHARACTERISTIC _MEDIAN
	BODY_CHARACTERISTIC _STD

Creating features related to vehicle type is a crucial step in analyzing data concerning freight transport. It provides a deeper understanding of the differences among various vehicle types and allows for more advanced analyses. Tab. 23 outlines the process of creating new features related to vehicle type. From each raw feature, four new features are derived: factorized feature, mean, median, and standard deviation within the group.

Tab. 23
Transformation of vehicle type data

Feature (Raw Data)	Feature (Transformed Data)
VEHICLE_TYPE	VEHICLE_TYPE _FACTORIZED
	VEHICLE_TYPE _MEAN
	VEHICLE_TYPE _MEDIAN
	VEHICLE_TYPE _STD

The data processing related to the loading and unloading methods for freight vehicles resulted in the creation of four new numerical features, which are crucial for transport analysis. These features provide an in-depth view of the characteristics of loading and unloading. They are detailed in Tab 24. From each raw feature, four new features are derived: factorized feature, mean, median, and standard deviation within the group.

Tab 24
Transformation of Load/Unload Method Data

Feature (Raw Data)	Feature (Transformed Data)
LOAD_UNLOAD METHOD	LOAD_UNLOAD METHOD _FACTORIZED
	LOAD_UNLOAD METHOD _MEAN
	LOAD_UNLOAD METHOD _MEDIAN
	LOAD_UNLOAD METHOD _STD

The data processing shown in

Tab. 25 related to the method of securing the cargo resulted in the creation of four new numerical features, which are crucial for transport analysis. These features provide an in-depth view of the characteristics of cargo securing methods. The new features derived from the raw data are factorized feature, mean, median, and standard deviation within the group.

Tab. 25
Transformation of load securing method data

Feature (Raw Data)	Feature (Transformed Data)
LOAD_SECURING	LOAD_SECURING _FACTORIZED
	LOAD_SECURING _MEAN
	LOAD_SECURING _MEDIAN
	LOAD_SECURING _STD

Additional data features that do not fit into the previously classified categories are introduced without modifications, retaining their original form. These features include TONS (the weight of the cargo in tons), HEIGHT (the height dimension of the cargo or vehicle), WIDTH (the width dimension of the cargo or vehicle), LDM (load meters, indicating the length of cargo space occupied), EPALE (pallet space, representing the number of standard pallets), TEMP_MIN (the minimum temperature requirement for transporting temperature-sensitive goods), and TEMP_MAX (the maximum temperature requirement for such goods).

Maintaining the original form of these features ensures consistency and accuracy in subsequent analyses, allowing for the direct interpretation of values without additional transformations. This approach promotes consistency in analysis by preserving the data in its raw form, facilitating straightforward interpretation and comparison. Additionally, these features are directly relevant to transportation cost models; variables such as weight, dimensions, load meters, pallet space, and temperature requirements are integral to logistics and cost calculations. Using them without transformation maintains the integrity of the data and enhances the relevance of the models.

Keeping these features in their raw form also simplifies the data processing pipeline, reducing the risk of errors and misinterpretations from unnecessary transformations. This simplicity and clarity are essential for efficient data handling. Moreover, these raw features can be directly applied in various analytical models, facilitating seamless integration into machine learning algorithms used for predictive analysis and optimization in transport logistics.

Therefore, maintaining the original form of certain key features ensures clarity, consistency, and relevance in the data analysis process. This approach simplifies the transformation pipeline and preserves the integrity of essential data points crucial for accurate modelling and prediction in transportation logistics.

4.3. Data Analysis Methods

Distance is a fundamental factor in developing a model to predict the cost of road freight transport services. The analysis focuses on the distance between the loading and unloading points as a critical determinant of transport costs. The process begins with the calculation of basic statistics related to distance, such as the mean, median, standard deviation, and range. These metrics provide an understanding of the overall distribution of distances within the dataset. Histograms and distance distribution plots are created to visualize how distances are spread across the data, revealing whether the dataset predominantly contains short, long, or evenly distributed routes. Correlation analyses between distances and transport service prices are conducted to determine whether longer distances correlate with higher prices and whether this relationship is linear or more complex. This analysis helps to identify key patterns and informs the development of more accurate pricing models.

The concept of relation is understood as the unique combination of the country of loading and the country of unloading. Analyzing the relationship between different loading and unloading locations is a crucial aspect of price forecasting for road transport services.

Significant correlations between various location combinations and transport service prices can be identified by analyzing relations. Such analyses help show which geographical connections have the most substantial impact on pricing, detect seasonal patterns related to specific routes, and provide insights into geographical factors that influence transport costs.

Analyzing date-related data plays a vital role in understanding seasonal and periodic patterns in logistics and transport. By examining these data, significant trends related to transportation can be identified. Statistical analysis allows for determining which days or periods affect fleet workload and operational rhythms.

Date analysis helps identify critical temporal dependencies, such as seasonal spikes in demand for transport services during holidays or pre-holiday periods. These insights are invaluable for transport companies that need to adjust their operations to changing demand. In the long term, date analysis can also aid in forecasting transport trends, which is crucial for strategic business decision-making.

Analyzing the dependency of road freight transport service prices on body type is crucial for understanding the factors influencing pricing in this industry. Body type is one of the primary factors determining the costs associated with transporting goods. Differences in operating and maintenance costs for various body types, such as refrigerated trailers, flatbed trailers, or curtain-sided trailers, directly impact the service price.

For example, higher energy consumption can make transporting temperature-controlled goods significantly more expensive. This analysis enables transport service providers to better understand which body types are most cost-effective and competitive. Additionally, this analysis can help identify market trends and price variability depending on body type, contributing to more informed pricing strategies and financial stability.

Integration into external databases aims to improve the model's quality and investigate the extent to which external data affects pricing. When considering integration with external databases, the implementation method must be considered, requiring a common key feature

present in both the project's data and the external database. An example is the transport execution date and the fuel price on the same date.

It is hypothesized that fuel prices significantly impact the cost of road freight transport services. Historical wholesale fuel prices from Orlen [86], specifically focusing on eco-diesel fuel prices, were analyzed to investigate this relationship.

Fuel is one of the primary operational costs for road freight transport companies, with fluctuations in fuel prices directly influencing their profitability and pricing strategies. An increase in fuel prices typically raises transportation costs, prompting companies to implement fuel surcharges, which are passed on to customers. Conversely, a decrease in fuel prices can reduce transportation costs, offering freight operators potential competitive pricing advantages.

The focus on eco-diesel fuel prices is particularly relevant due to its growing importance in the transportation sector. Eco-diesel, a biodiesel blend, provides a cleaner alternative to traditional diesel fuel by emitting lower levels of pollutants such as sulphur oxides and particulate matter. This aligns with increasing regulatory pressures and societal expectations for environmental responsibility within the transport industry. Adopting eco-diesel is expanding as companies aim to improve their sustainability profiles and comply with stricter emissions standards.

Orlen, one of Central and Eastern Europe's leading oil refiners and fuel retailers, supplies a significant portion of the region's fuel needs. Known for its commitment to innovation and sustainability, Orlen produces eco-diesel as part of its product portfolio. Eco-diesel is a blend of traditional diesel and biodiesel derived from renewable sources, offering high performance and efficiency while meeting stringent environmental standards. This makes it a popular choice among freight operators striving to balance performance with sustainability.

Orlen’s wholesale prices for eco-diesel provide an industry benchmark, reflecting a combination of global oil prices, regional market dynamics, and production costs. These wholesale prices, which are typically lower than retail prices due to the exclusion of taxes and retail margins, provide a reliable indicator of underlying fuel costs for large-scale consumers such as freight companies.

The factors influencing Orlen’s wholesale eco-diesel prices include crude oil prices, production costs, market demand, regulatory environments, and logistics and distribution costs. Crude oil prices, driven by global supply and demand, geopolitical events, and decisions by major oil producers, have a significant impact. Production costs are determined by refining processes and the blending of biodiesel components, with technological advancements and economies of scale contributing to cost fluctuations. Regional and global demand for eco-diesel affects its pricing, with higher demand driving prices upward. The regulatory environment, including policies promoting renewable energy and reducing emissions, influences production incentives and penalties, thereby affecting pricing. Logistics and distribution costs, which cover the transportation and delivery of eco-diesel, further shape wholesale prices.

By analyzing Orlen’s wholesale eco-diesel prices, broader trends and underlying factors influencing the cost of road freight transport services can be understood. This analysis provides critical insights into the relationship between fuel prices and freight companies’ operational costs, supporting strategic decision-making and pricing model development.

Eurostat, the statistical office of the European Union (EU), is a primary source of statistical data related to the EU. It collects, processes, and disseminates a wide range of economic, social, and environmental information from member states and other institutions. The data collected by Eurostat cover production, trade, employment, unemployment, inflation, income, social expenditure, health, education, and the natural environment.

Eurostat adheres to rigorous data collection standards to ensure the quality, consistency, and comparability of its data across the EU. This reliability suggests that Eurostat data are suitable for inclusion in the predictive model. These statistics are vital for policymakers, economists, researchers, and the public to make informed decisions, analyze trends, and monitor economic and social development in Europe.

5. PRACTICAL APPLICATION OF THE METHOD ON A STATISTICAL SAMPLE

This section demonstrates the practical application of the developed forecasting methodology using a real-world dataset. The focus is on evaluating the model's effectiveness in predicting the prices of road freight transport services.

5.1. Analysis

The information obtained from df.info() is presented in Tab. 26. The DataFrame consists of 45,668 entries and 54 columns of various data types, including float64, int64, datetime64, and object. There are no missing values in the columns from AT_KM to SK_KM, but many other columns contain missing data (e.g., START_LOAD_TIME, END_LOAD_TIME, START_DELIVERY_TIME, END_DELIVERY_TIME, GOODS_TYPE, CARGO_TYPE, TEMP_MIN, TEMP_MAX, HEIGHT, REQUIREMENTS, QTY_LOADS, QTY_DELIVERIES, PAYMENT_TERM, DOCUMENTS_BY, CARGO_VALUE_EURO, CUSTOMS, and TIME_OF_ENTRY). Some columns, like TEMP_MIN and TEMP_MAX, have only 20 non-empty values, indicating sparse data. Moreover, some columns have very few non-empty values (e.g., REQUIREMENTS, CARGO_VALUE_EURO, DOCUMENTS_BY, CUSTOMS, and TIME_OF_ENTRY). The total memory usage of the DataFrame is 18.8+ MB.

Tab. 26

Information from DataFrame

Column	Non-Null Count	Dtype
AT_KM	45,569	float64
BE_KM	45,569	float64
CZ_KM	45,569	float64
DE_KM	45,569	float64
DK_KM	45,569	float64
EE_KM	45,569	float64
ES_KM	45,569	float64
FI_KM	45,569	float64
HR_KM	45,569	float64
FR_KM	45,569	float64
HU_KM	45,569	float64
IT_KM	45,569	float64
LT_KM	45,569	float64
LV_KM	45,569	float64
NL_KM	45,569	float64
PL_KM	45,569	float64
RO_KM	45,569	float64
SE_KM	45,569	float64
SI_KM	45,569	float64
SK_KM	45,569	float64
COD_LP	45,569	object

COD_DP	45,569	object
ROUTE_TYPE	45,569	object
START_LOAD_DATE	45,569	datetime64[ns]
START_LOAD_TIME	42,760	object
END_LOAD_DATE	45,569	datetime64[ns]
END_LOAD_TIME	42,754	object
START_DELIVERY_DATE	45,569	datetime64[ns]
START_DELIVERY_TIME	42,422	object
END_DELIVERY_DATE	45,569	datetime64[ns]
END_DELIVERY_TIME	42,426	object
VEHICLE_TYPE	45,569	object
BODY_TYPE	45,569	object
LOAD_UNLOAD_METHOD	45,569	object
EPALE	45,569	int64
GOODS_TYPE	45,394	object
CARGO_TYPE	42,851	object
TEMP_MIN	20	float64
TEMP_MAX	20	float64
EUR	45,569	float64
LDM	45,569	float64
M3	45,569	float64
HEIGHT	45,569	float64
WIDTH	45,569	float64
TONS	45,569	float64
REQUIREMENTS	2	object
OTHER_COSTS	45,569	float64
QTY_LOADS	45,569	int64
QTY_DELIVERIES	45,569	int64
PAYMENT_TERM	45,45,1	float64
DOCUMENTS_BY	2486	object
CARGO_VALUE_EURO	2	float64
CUSTOMS	45,569	int64
TIME_OF_ENTRY	2463	datetime64[ns]

Tab. 27 below presents a set of statistics for selected features. The rows of the DataFrame correspond to various statistics, such as the mean, standard deviation, median, minimum and maximum values, percentiles (q1 and q3), as well as V and Vq values. The values in the DataFrame have been rounded to two decimal places for easier reading and analysis. This DataFrame can be used to quickly summarize and analyze numerical data from the DataFrame df, providing essential statistical information for the selected columns.

Tab. 27
Basic statistical data

	mean	std	V	median	min	max	q1	q3	q	Vq
EUR	818.92	409.98	50.06	753	54.88	4300	500	1025	262.5	34.86

KM	739.15	436.73	59.09	704.7	1	3062	410.2	937.7	263.75	37.43
TONS	24.04	0.53	2.18	24	1.52	25.7	24	24	0	0
EUR_FOR_KM	1.34	2.1	157.18	1.1	0.28	87.53	0.97	1.46	0.25	22.52
EPALE	0	0.34	8941.54	0	0	34	0	0	0	
LDM	13.6	0	0.01	13.6	13.2	13.6	13.6	13.6	0	0
M3	96.94	3.1	3.2	97.72	84.68	120	97.72	97.72	0	0
HEIGHT	2.97	0.12	4	3	2.5	3	3	3	0	0
WIDTH	2.4	0	0	2.4	2.4	2.4	2.4	2.4	0	0
QTY_LOADS	1.5	0.5	33.27	2	1	4	1	2	0.5	25
QTY_DELIVERIES	1.5	0.5	33.39	2	1	6	1	2	0.5	25
OTHER_COSTS	0.24	11.32	4758.83	0	0	898.71	0	0	0	
CUSTOMS	0	0.01	10,672.98	0	0	1	0	0	0	

Fig. 3 presents a correlation that depicts the linear relationships between pairs of numerical features. The values in the matrix are Pearson correlation coefficients, which range from -1 to 1. Values on the main diagonal (from the top left to the bottom right) are equal to 1, representing the correlation of each feature with itself, which is obvious.

Values off the diagonal represent the correlations between different pairs of features. The numbers in the matrix indicate the degree of correlation between features. The closer the value is to 1, the stronger the positive correlation (i.e., both features increase together); the closer the value is to -1, the stronger the negative correlation (one feature increases while the other decreases). Values close to 0 indicate no linear relationship between the features. Features “EUR” and “KM” show a high positive correlation (around 0.92), suggesting that the longer the route, the higher the price. “EUR_FOR_KM” shows a slight negative correlation with “KM,” which may indicate that the price per kilometre is lower for longer routes. “M3” and “HEIGHT” have a strong positive correlation (almost 1), suggesting that these two features are practically identical. “QTY_LOADS” and “QTY_DELIVERIES” have a very high positive correlation (around 0.99), indicating that the number of loads is strongly correlated with the number of deliveries. “OTHER_COSTS” and “CUSTOMS” show little correlation with other features, suggesting their weak linear dependence on the remaining features.

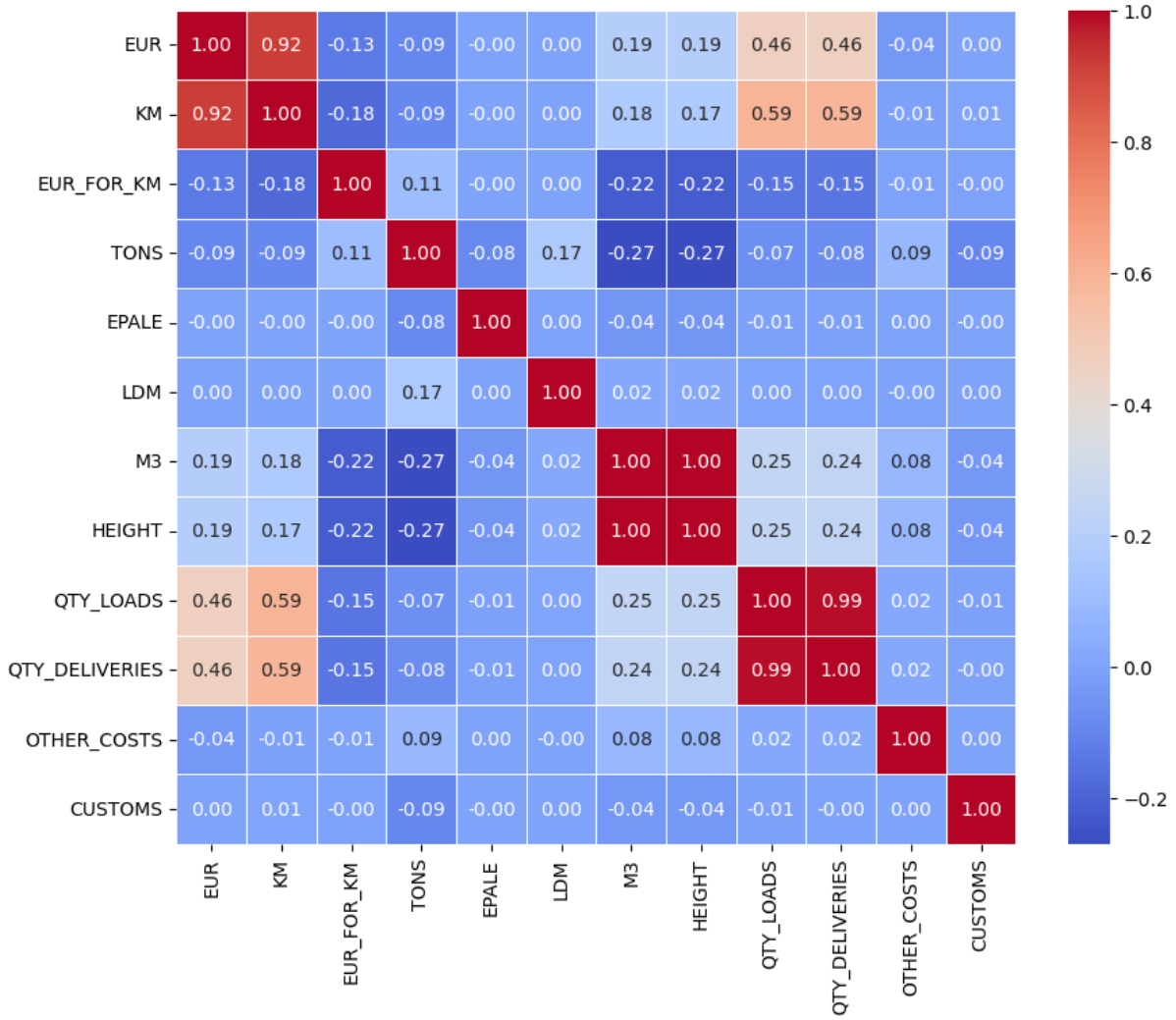


Fig. 3 Correlation matrix of numerical features

Fig. 4 is a histogram depicting the distribution of kilometres (km) in the data. The X-axis represents ranges of kilometre values, divided into 100-km intervals, starting at 0 km and ending at the nearest whole number above the maximum value contained in the data. The Y-axis represents the frequency, indicating how many times a certain number of kilometres occurs in the data. Each bar on the histogram represents one group of values (i.e., the range of values on the X-axis). The height of the bar corresponds to the number of occurrences of values within that range. Additionally, every second label on the X-axis is marked with a value every 200 km, facilitating easier reading of the data. Such a histogram is useful for understanding the distribution of data related to the number of kilometres, allowing us to identify how often routes of various lengths occur.

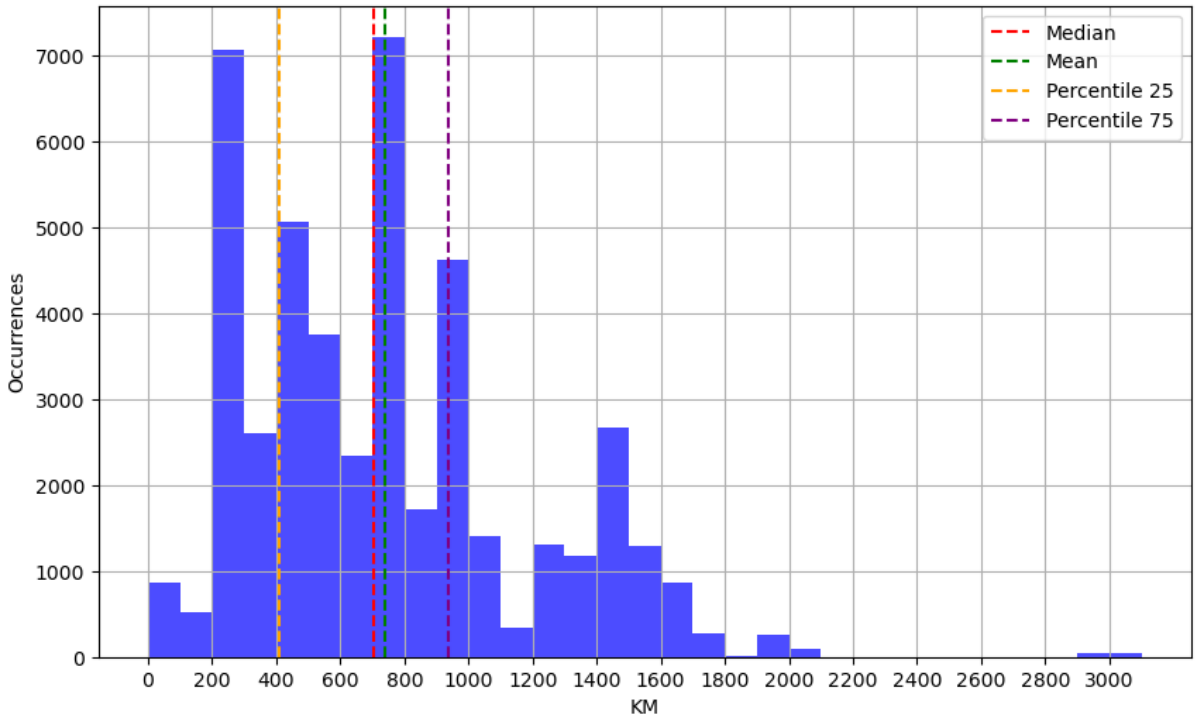


Fig. 4 Histogram of distance variable distribution

Fig. 5 shows the relationship between the price in euros (EUR) and the distance in kilometres (km). The X-axis represents distances in kilometres, while the Y-axis represents prices in euros. The chart includes three lines representing (1) the average EUR price per km in 100-km intervals, marked by a solid line, (2) the average price of 1.34 EUR, marked by a dashed line, representing the value of 1.34 EUR multiplied by each distance in kilometres, and (3) the median price of 1.1 EUR, marked by a dotted line, representing the value of 1.1 EUR multiplied by each distance in kilometres. The trend shows that the average price for distances up to 400 km is below the overall dataset average, confirming the hypothesis that higher prices are associated with shorter transport distances.

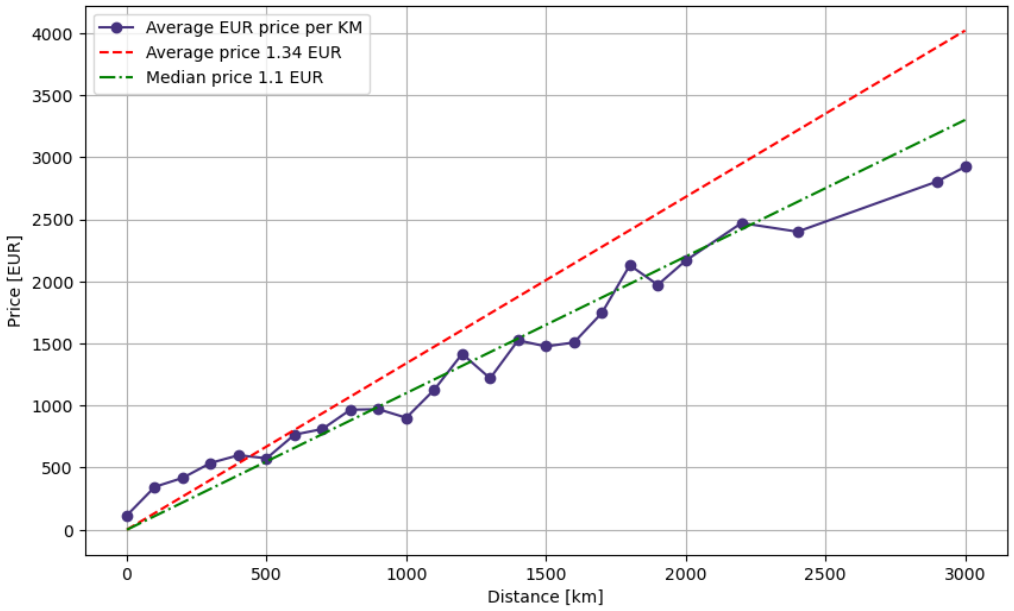


Fig. 5 Price–distance relationship

Fig. 6 shows the relationship between the average EUR/km price and the distance in kilometres. The X-axis represents distance in kilometres, with labels every 500 km. The Y-axis represents the EUR/km price. The points on the chart represent the average EUR/km price in a given distance range, where the ranges are grouped every 100 km. Additionally, the chart includes two horizontal lines: one indicating the average value (1.34) and the other representing the median value (1.1) of the EUR/KM price. The chart allows for an easy visualization of the changes in EUR/km price depending on the distance.

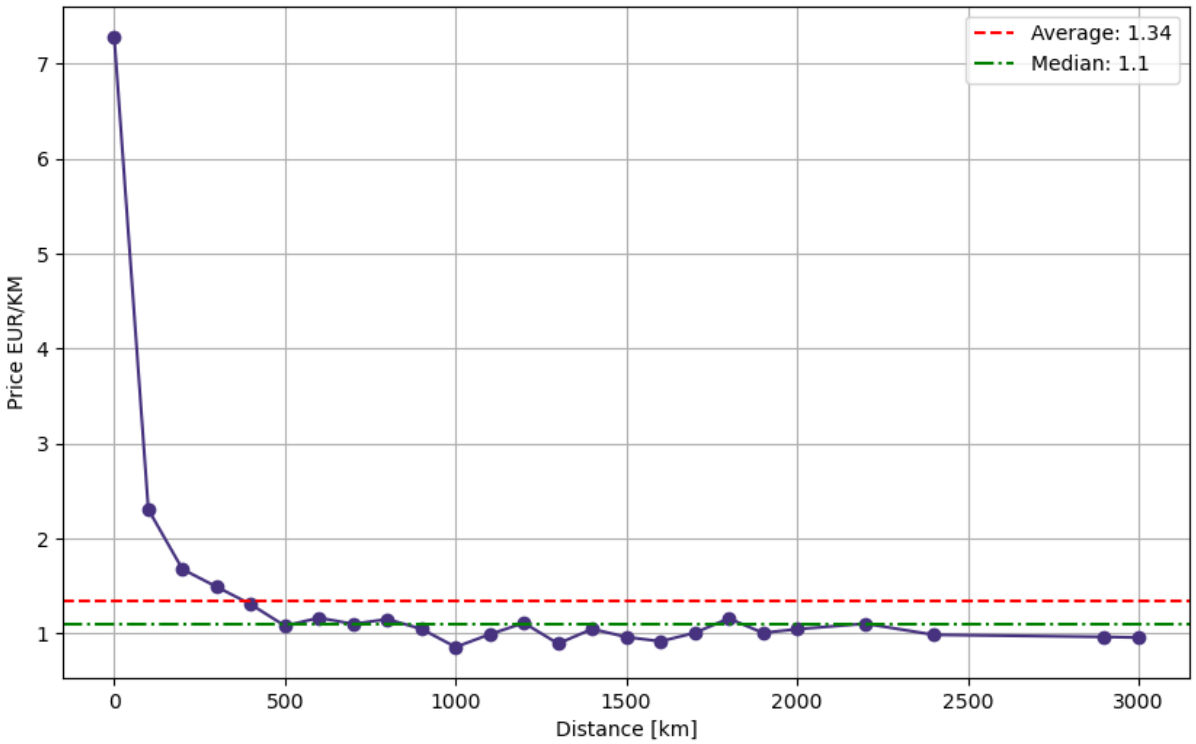


Fig. 6 Price per kilometre vs. distance relationship

Tab. 28 presents data on the number of kilometres for individual countries and their percentage share in the total number of kilometres. The first row provides information about all countries, with the value of "KM" indicating the sum of all kilometres. Then, data for individual countries is presented, with "PL_KM" denoting the number of kilometres for Poland, "DE_KM" for Germany, etc. The third column shows the percentage share of each country in the total number of kilometres.

Tab. 28
Distribution of distance by country

Country	Kilometres	Percentage Share
Total	33,682,440	100.00 %
PL_KM	14,909,043	44.26 %
DE_KM	13,217,990	39.24 %
CZ_KM	3,835,888	11.39 %
SK_KM	1,340,260	3.98 %
IT_KM	93,167	0.28 %
HU_KM	79,063	0.23 %
FR_KM	57,085	0.17 %
BE_KM	32,929	0.10 %
AT_KM	31,595	0.09 %
LT_KM	30,699	0.09 %
NL_KM	15,077	0.04 %
SE_KM	14,735	0.04 %
LV_KM	11,324	0.03 %
EE_KM	8280	0.02 %
FI_KM	2674	0.01 %
RO_KM	1198	0.00 %
DK_KM	973	0.00 %
ES_KM	216	0.00 %
HR_KM	140	0.00 %
SI_KM	104	0.00 %

The distribution in Fig. 7 represents the number of transports loaded in various countries. Poland (PL) has the highest number of transports, totalling 38,426. Germany (DE) ranks second with 4801 transports, while the Czech Republic (CZ) is third with 2023 transports. Other countries have significantly fewer loaded transports, with most having fewer than 100. For example, Slovakia (SK) has 116 transports, Italy (IT) has 69, Lithuania (LT) has 39, and Hungary (HU) has 38. Austria (AT) has the fewest loaded transports, with just one.

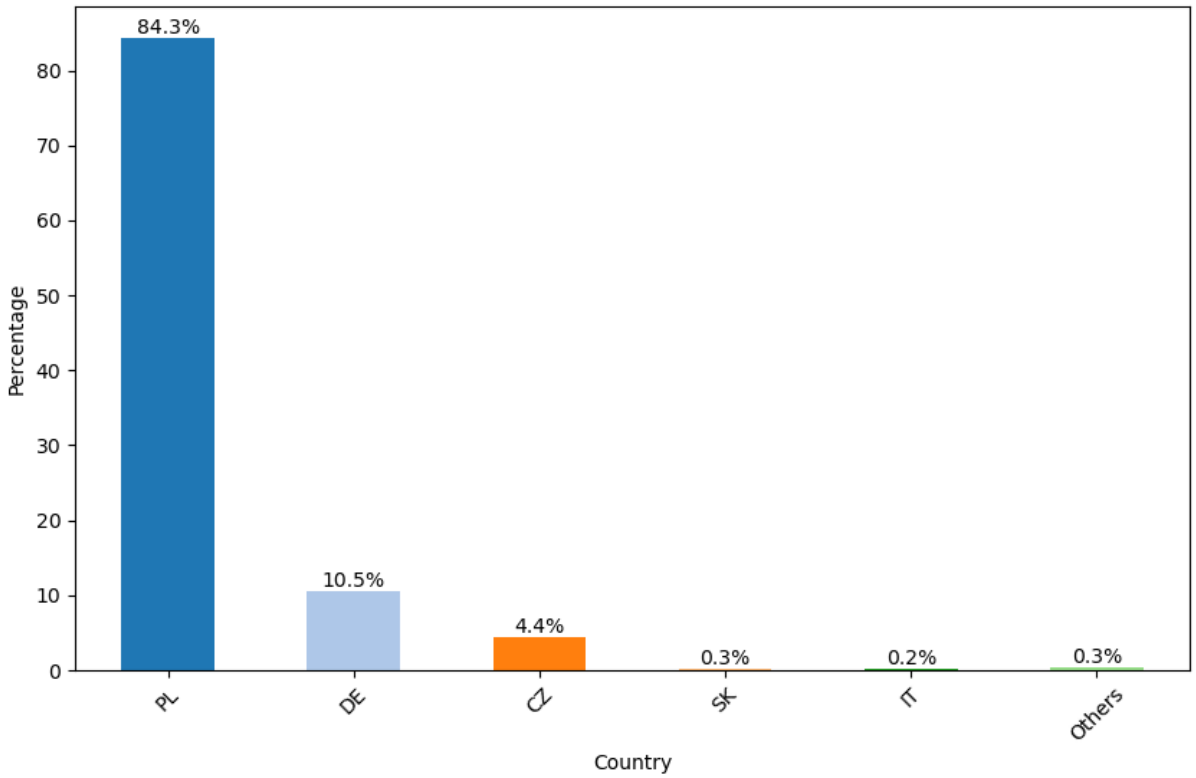


Fig. 7 Distribution of data by loading country

The analysis of the distribution of data by the country of unloading shown in Fig. 8 highlights the variation in the number of deliveries to different countries. The highest number of deliveries is recorded in Poland, which is also the loading country for many transports. This may be due to its central location and large domestic market. The Czech Republic and Germany follow, perhaps owing to their central position in Europe and logistical significance. Slovakia, Hungary, and Belgium also have a significant number of deliveries, possibly due to their developed transport and trade infrastructure. Other countries, such as the Netherlands, Lithuania, and Sweden, have fewer deliveries, which may be related to a smaller target market or specific trade relations. It is worth noting that the number of deliveries to individual countries reflects their economic importance, geographical location, and trade relations with Poland.

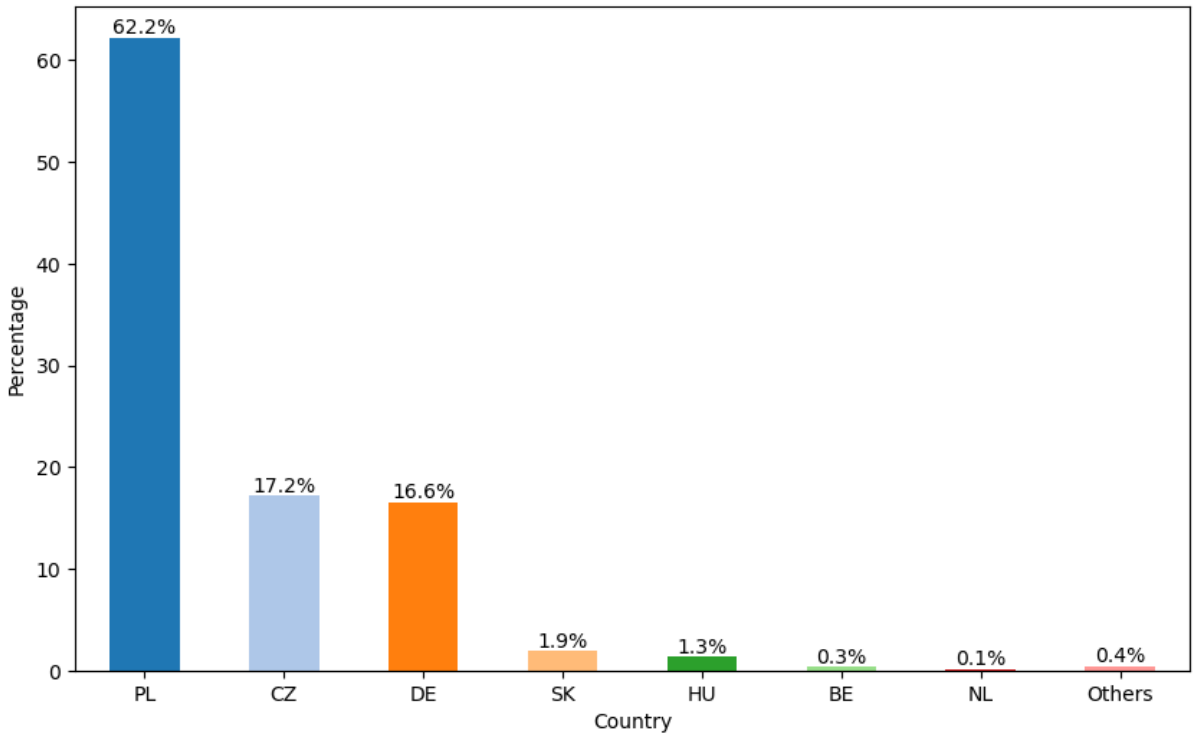


Fig. 8 Distribution of data by unloading country

Fig. 9 shows the average price per kilometre for different loading countries. The highest average price per kilometre is recorded for transports loaded in Poland (PL), at €1.39. Conversely, the lowest average price per kilometre is observed for transports loaded in Hungary (HU), at €0.79.

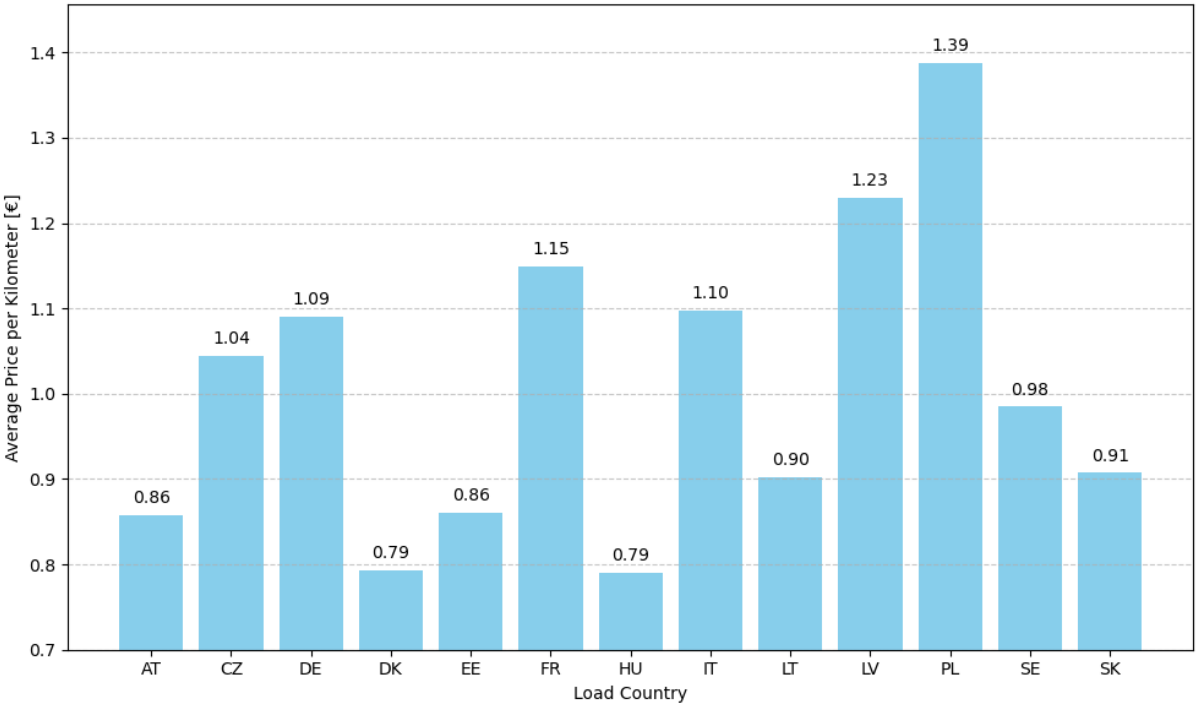


Fig. 9 Average prices by loading country

Fig. 10 shows that the highest average price per kilometre for unloading countries is achieved in Sweden (SE), with a value of €2. Meanwhile, the lowest price of €0.98 is observed in Spain (ES). This information can be crucial for logistical decisions and transport cost management.

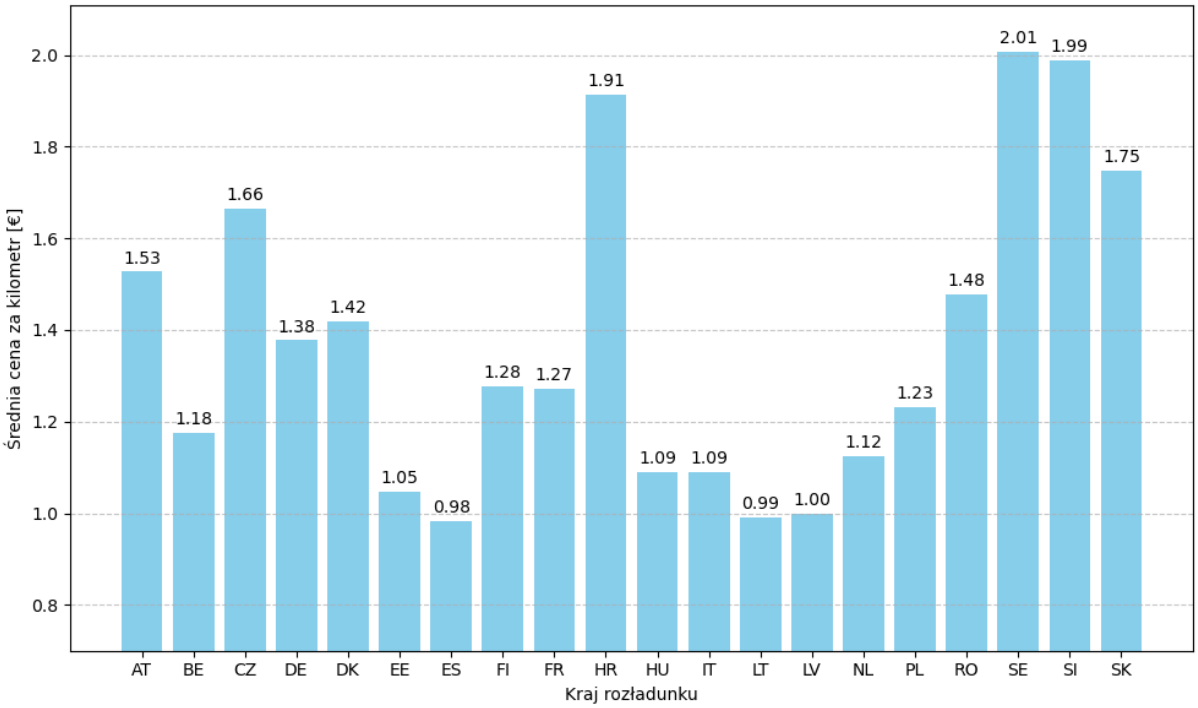


Fig. 10 Average prices by unloading country

Fig. 11 presents the average price per kilometre (EUR/km) for transports between different pairs of loading and unloading countries. The key insights gained from this table are as follows:

High prices for Poland: Poland (PL) consistently shows high average prices for transport to various countries. For instance, the price for transports loaded in Poland and delivered within Poland is exceptionally high at €4.85 per kilometre; this is significantly higher than other routes. This high value for domestic transport could be due to shorter average distances, leading to a higher price per kilometre, or higher demand for internal logistics.

Significant variations: There is notable variation in prices between different country pairs. For example, the average price per kilometre for transport from Poland to the Czech Republic (CZ) is €1.70, while from Poland to Germany (DE), it is €1.48. Such variations could be influenced by factors such as the distance, demand for transport services, and economic relationships between the countries.

High costs for specific routes: Certain routes exhibit particularly high costs. For instance, transport from Latvia (LV) to Sweden (SE) shows an extremely high price of €6.72 per kilometre. This could be due to logistical challenges, a low frequency of transport, or high costs associated with crossing the Baltic Sea.

Low average prices: Hungary (HU) shows relatively low prices for transport to various destinations, such as €0.79 from Hungary to Sweden (SE). This suggests that pricing is competitive pricing or that operational costs are low for these routes.

Cross-border transport costs: Cross-border transport costs vary significantly. For instance, the cost of transporting goods from the Czech Republic to Germany is €1.73 per

kilometre, whereas from Germany to the Czech Republic, it is €2.00. This asymmetry might reflect different demand and supply conditions or logistical complexities.

Sparse data for some routes: some country pairs have missing data (NaN), indicating either a lack of recorded transports or insufficient data points to calculate a reliable average. For example, there is no data available for transports from Austria (AT) to most other countries except Poland, suggesting that these routes are not common or are underrepresented in the dataset.

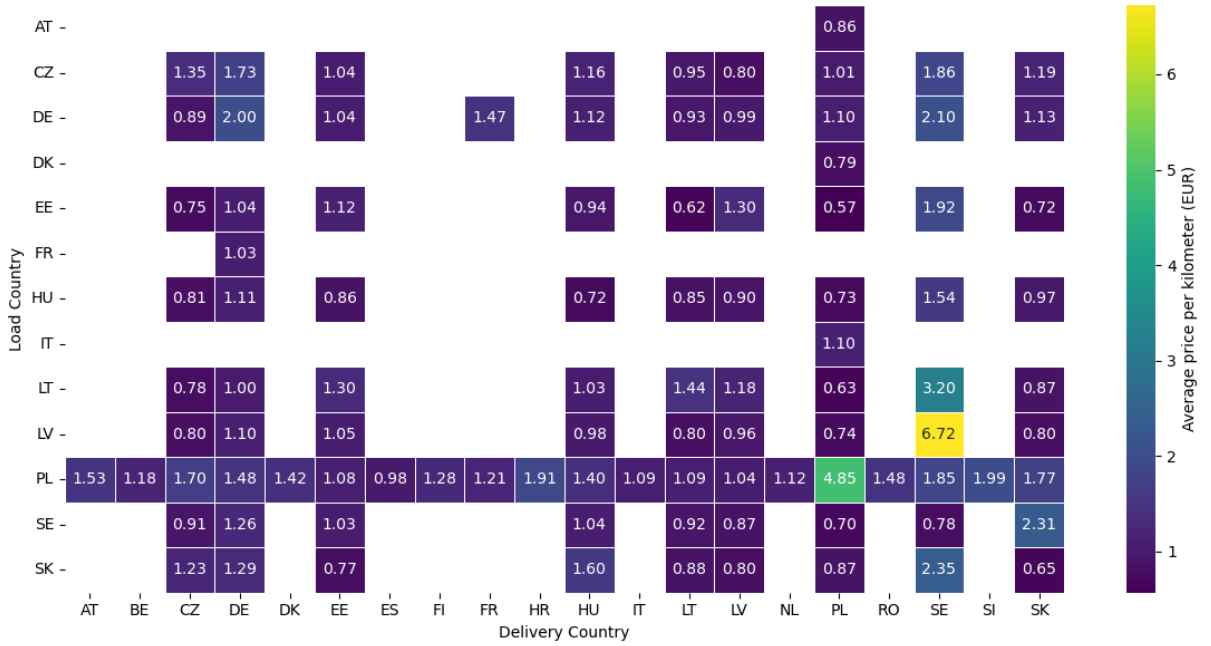


Fig. 11 Average price by relation

The bar chart shown in Fig. 12 presents the number of transports in each month. The horizontal axis (X-axis) represents the consecutive months of the year, while the vertical axis (Y-axis) shows the number of transports in each month. Each bar in the chart corresponds to one of the 12 months, and its height represents the number of transports in that month.

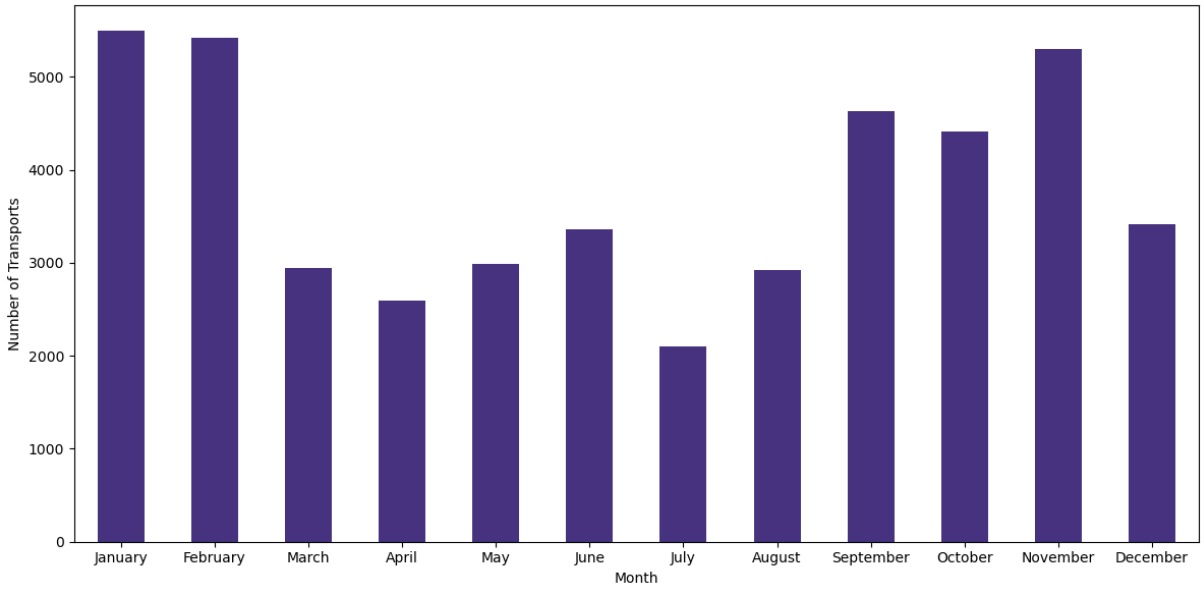


Fig. 12 Number of transport offers summed by month

Fig. 13 illustrates the average prices per kilometre over 48 months. The data show the variability of average prices per kilometre in different months and years. In some industries, seasonality can affect price variability. For example, during summer or pre-holiday months, demand for transport might increase, leading to higher prices. The overall situation in the transport market, including supply and demand, can affect prices. When supply is limited or demand is high, prices may increase. Similarly, costs associated with vehicle operation, such as fuel prices, can influence transport prices. Events such as changes in transport regulations, economic crises, natural disasters, or pandemics can also impact transport prices.

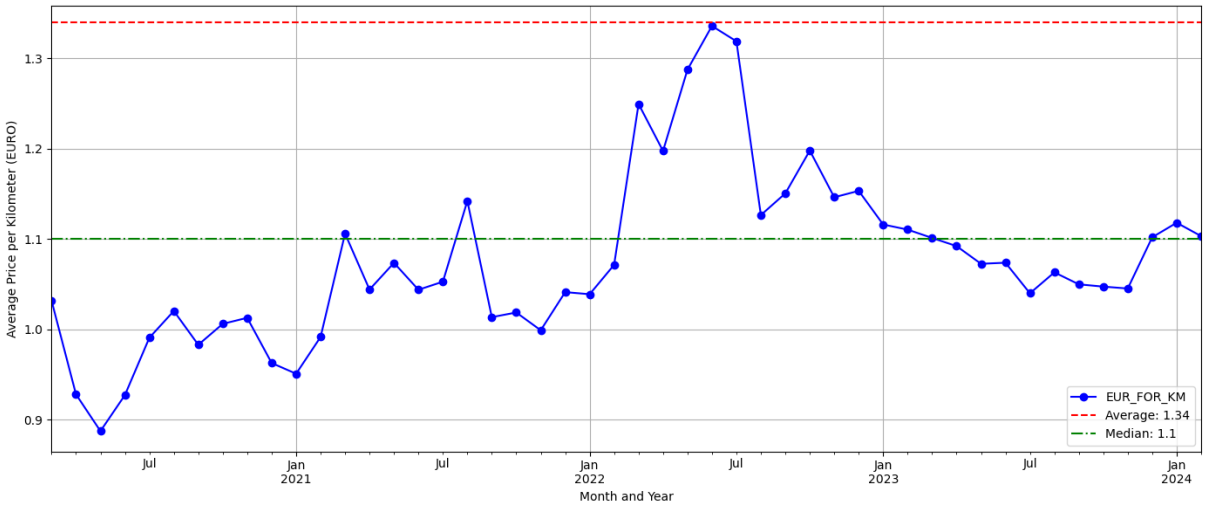


Fig. 13 Average prices by month over four years

Fig. 14 presents the average price per kilometre (€) for each month. Each month is represented by a month number (from 1 to 12), with the corresponding value indicating the average price per kilometre for that month. The average price per kilometre in different months can vary for several reasons. Changes in weather conditions can impact transport costs. For instance, in winter months, transport costs might increase due to poor road conditions, as additional expenses are required for safety and vehicle maintenance. Periods such as Christmas, New Year's, or other significant events can lead to increased demand for transport services, potentially raising prices during those months.

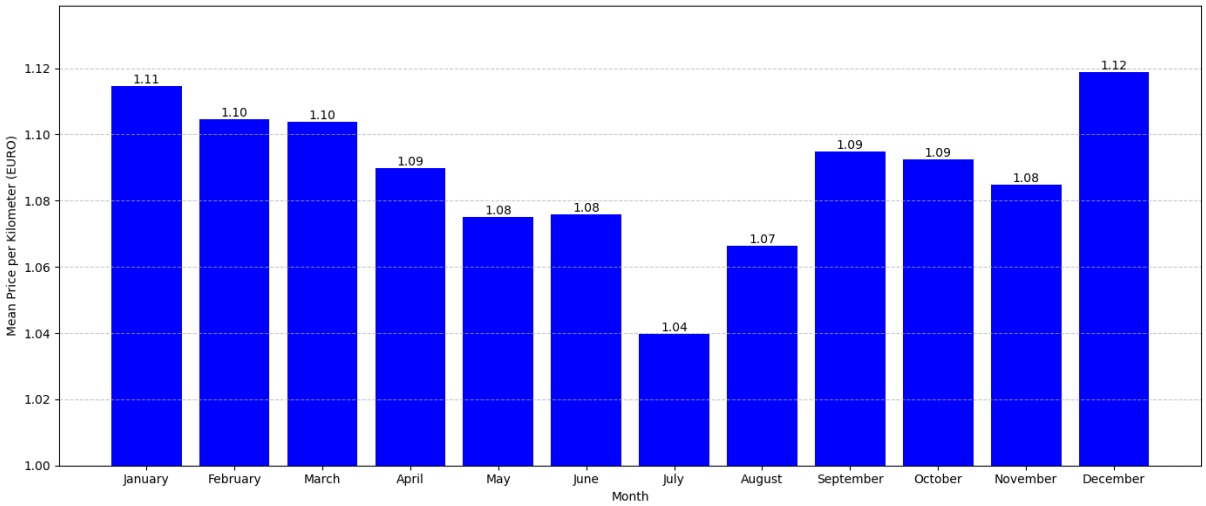


Fig. 14 Average prices by month

Fig. 15 presents the average price per kilometre for the types of dates (START_LOAD_DATE, END_LOAD_DATE, START_DELIVERY_DATE, and END_DELIVERY_DATE) for each day of the week, expressed in euros (€). The average prices have been calculated based on transport data and represent the average rate per kilometre for each type of date on a specific day of the week. The average prices for the start and end of the loading day are similar. A similar situation is observed for the start and end of the delivery day. In most cases, the price of loading matches the price of the subsequent delivery day.

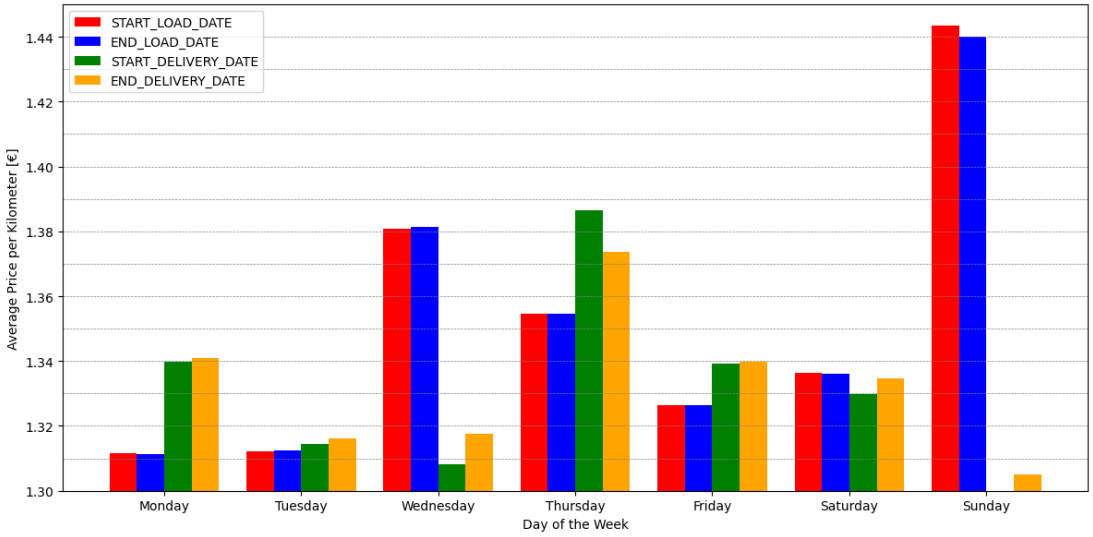
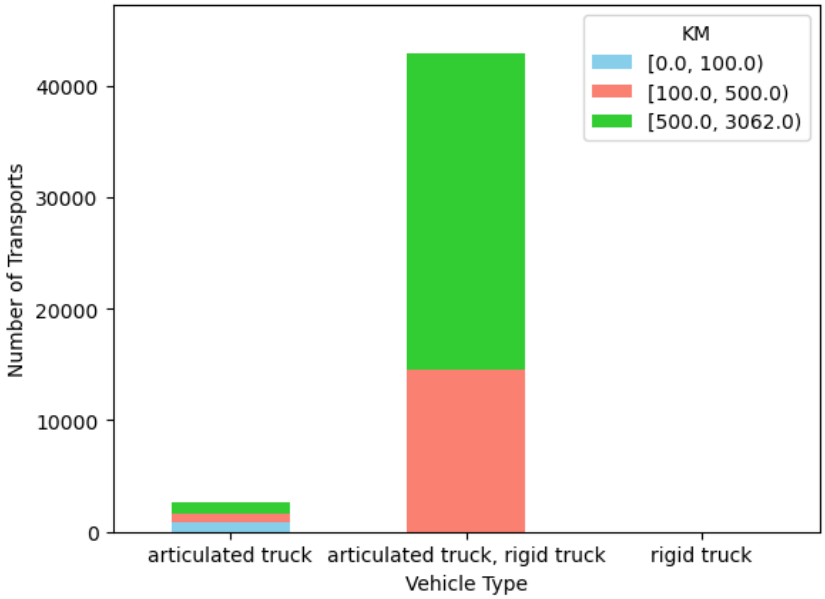


Fig. 15 Average price by day of the week

The analysis of transport data over time reveals significant trends and patterns. Examining the number of transport offers and their average prices allows the identification of seasonal variations, economic impacts, and other factors influencing the transport market. Understanding these dynamics is crucial for logistics planning and cost management, as it enables better decision-making and optimization of transport operations.

Fig. 16 reveals the number of vehicles categorized by type and range (km). The table includes three categories of mileage intervals: [0.0, 100.0), [100.0, 500.0), and [500.0, 3062.0). The results are grouped by vehicle type, including articulated truck, articulated truck with an additional rigid truck, and rigid truck.



	(0 km, 100 km)	[100 km, 500 km)	[500 km, 3062 km]
articulated truck	872	771	968
articulated truck, rigid truck	1	14486	28430
rigid truck	0	0	1

Fig. 16 Frequency of vehicle types by mileage ranges

Fig. 17 presents data on vehicle groups by travel distance in three categories: up to 100 km, from 100 to 500 km, and above 500 km. It shows differences in average prices per kilometre for different vehicle types. For articulated trucks, the average price per kilometre is 7.29 currency units for trips up to 100 km, which significantly exceeds the values for journeys ranging from 100 to 500 km (1.54€) and above 500 km (1.05€). In contrast, the combination of articulated and rigid trucks does not occur for trips up to 100 km. Their average price per kilometre for journeys from 100 to 500 km is 1.54€, while for trips over 500 km, it is 1.06€. It is noteworthy that there is no data (NaN) for articulated trucks in trips up to 100 km combined with a rigid truck.

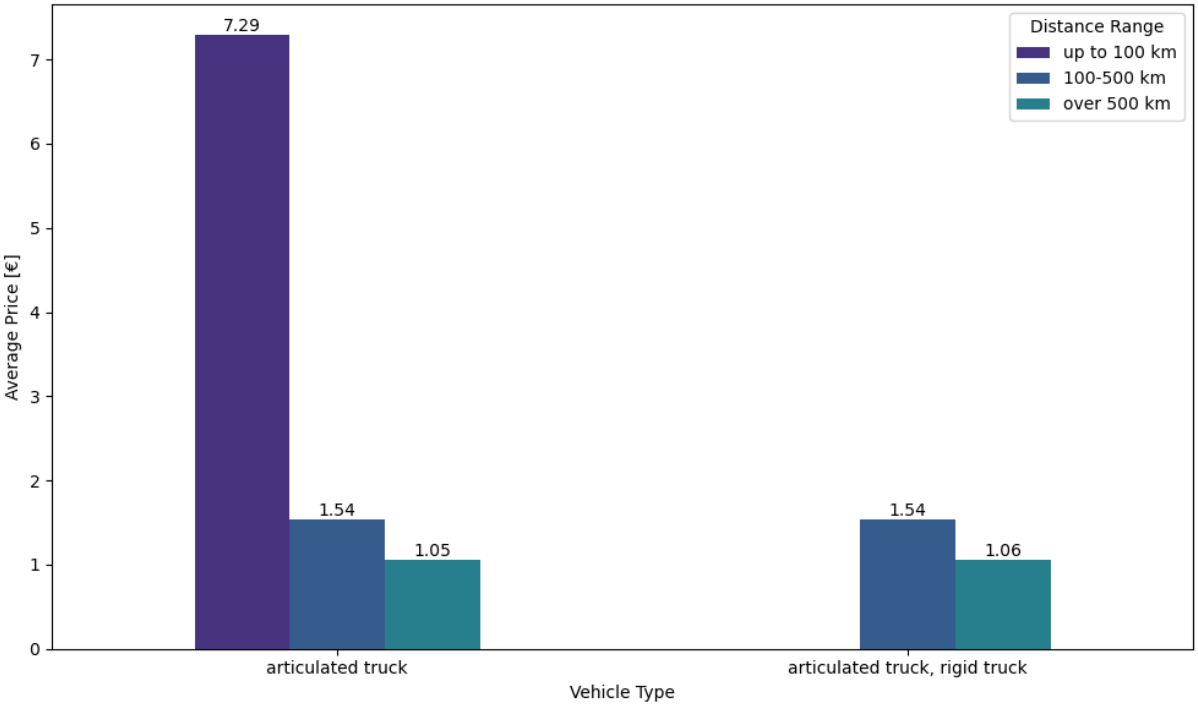


Fig. 17 Average price per kilometre by vehicle type and mileage range

Tab. 29 presents data on the body type of vehicles depending on the distance travelled, divided into three categories: from 0 to 100 km, from 100 to 500 km, and above 500 km. The table shows the variation in the frequency of occurrences of different body type combinations. For body type combinations such as box, coil, mega, refrigerator, standard, and thermo, we observe a variable number of occurrences depending on the travel distance, with values for the three intervals being 1, 4, and 7, respectively. For the combination of box, coil, mega, and standard, the number of occurrences is 0, 4, and 3, respectively. Similarly, for other combinations like box, mega, refrigerator, standard, and thermo or box, refrigerator, and standard, the number of occurrences varies depending on the distance category. It is worth noting the high number of occurrences of the "mega" body type, especially for journeys between 100 and 500 km and above 500 km, for which the numbers reach 14,447 and 28,345, respectively. This data analysis can help explain preferences and demand for different body types depending on the travel distance.

Tab. 29
Frequency of body types by distance

BODY_TYPE	[0.0, 100.0)	[100.0, 500.0)	[500.0, inf)
box, coil, mega, refrigerator, standard, thermo	1	4	7
box, coil, mega, standard	0	4	3
box, jumbo, refrigerator, standard	0	1	0
box, mega, refrigerator, standard, thermo	0	6	2
box, refrigerator	0	0	1
box, refrigerator, standard	0	1	0
cooler, isotherm, standard	0	0	1
flatbed truck	0	1	0
isotherm, standard, refrigerator	0	0	1

izotherm, standard, refrigerator	0	1	0
jumbo, mega, standard, refrigerator	0	0	1
jumbo, standard	0	0	1
mega	0	14,447	28,385
mega, standard	0	2	6
refrigerator	0	6	21
refrigerator, standard	0	0	4
standard	872	782	995
standard, refrigerator	0	2	1
standard, thermo	0	0	3
standard, thermo, refrigerator	0	0	2
standard, thermo, refrigerator,	0	0	1
standard	0	0	2
thermo, refrigerator	0	0	1
tipper	0	0	1

Fig. 18 presents data on prices per kilometre depending on the travel distance, divided into three ranges (up to 100 km, from 100 to 500 km, and above 500 km) and vehicle type (BODY_TYPE). There are noticeable differences in average prices per kilometre between vehicle types and travel distances. The lack of data (NaN) for trips up to 100 km for “mega” type vehicles suggests that this type of vehicle may be less common than others for short-distance journeys. The average price per kilometre for “mega” in the range of 100 to 500 km is 1.542, which shows a significant difference compared to the prices for trips over 500 km (1.055). For “standard” type vehicles, short trips up to 100 km are more common, with an average price per kilometre of 7.29. However, in the further distance ranges, the average prices per kilometre for “standard” are relatively low (1.54 and 1.057 for trips from 100 to 500 km and over 500 km, respectively).

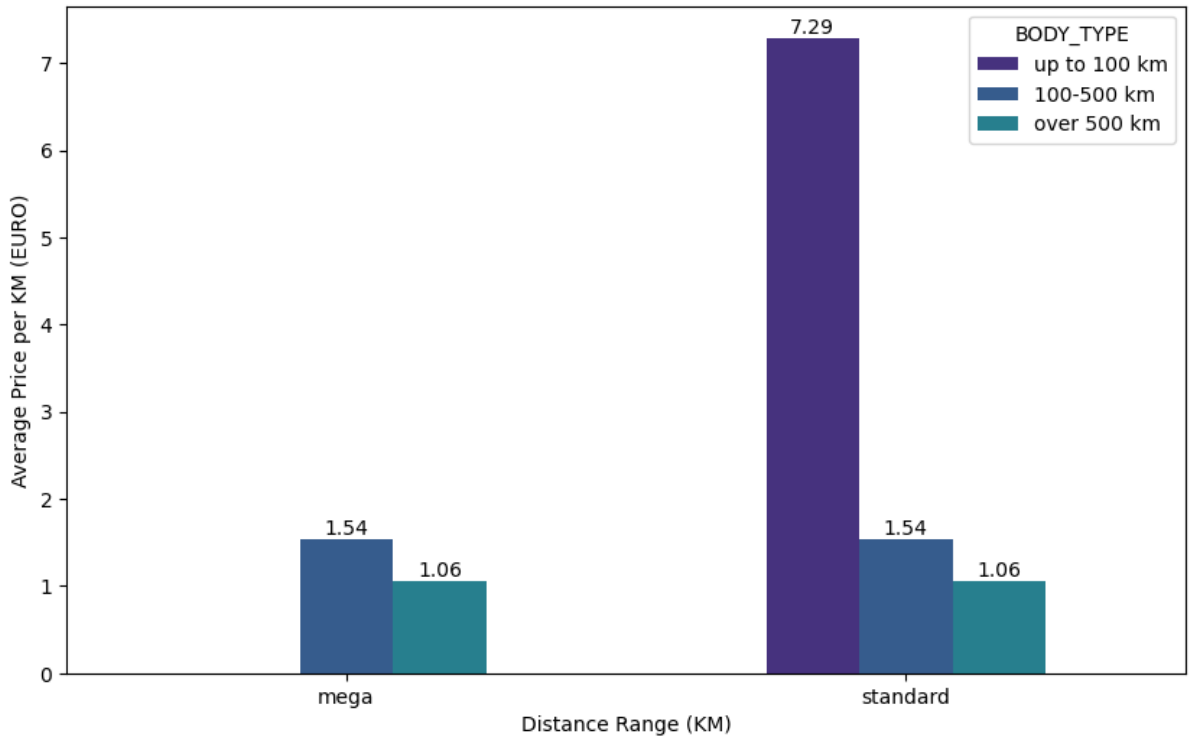


Fig. 18 Average prices by body type in mileage ranges

Tab. 30 shows unique values and counts for the OTHER_COSTS feature. The value 0 is the most frequent, occurring 45,544 times. This indicates that for the majority of records, there are no additional costs beyond the primary transport charges.

The positive values represent additional costs, such as the fees for ferries, tunnels, and bridges. These relatively infrequent costs range from small amounts (e.g., 18) to large (e.g., €796.12). Costs above €300 are relatively rare, each appearing only once or a few times. These may represent specific and possibly less common additional charges. The additional costs vary widely, which suggests variability in the types and amounts of additional charges that might be incurred during transportation.

Tab. 30
OTHER_COSTS unique values and counts

OTHER_COSTS [€]	Count
0	45,544
300	9
298	3
384.83	3
643	1
468	1
570.97	1
776.5	1
662.07	1
840.09	1
796.12	1

18	1
420.26	1
898.71	1

Tab. 31 shows unique values and counts for the QTY_LOADS and QTY_DELIVERIES features. The values of 2 and 1 are the most frequent for QTY_LOADS (appearing 22,937 and 22,628 times, respectively) and QTY_DELIVERIES (appearing 22,801 and 22,759 times). This indicates that the majority of transactions involve either 1 or 2 loads or deliveries. In contrast, values 3 and 4 for loads and 3, 4, 5, and 6 for deliveries appear only a few times. This suggests that transactions involving more than two loads or deliveries are rare. The high counts for 1 and 2 in both QTY_LOADS and QTY_DELIVERIES indicate that single and double load/delivery transactions are predominant in the dataset. The very low counts for higher values indicate that transactions involving more than two loads or deliveries are rare, possibly due to logistical constraints or the nature of the transported goods.

Tab. 31
QTY_LOADS and QTY_DELIVERIES unique values and counts

QTY_ LOADS	QTY_LOADS_ COUNT
2	22,937
1	22,628
3	3
4	1

QTY_ DELIVERIES	QTY_DELIVERIES _COUNT
2	22,801
1	22,759
3	5
4	2
6	1
5	1

Tab. 32 shows unique values and counts for the PAYMENT_TERM feature. The analysis of the PAYMENT_TERM feature reveals that 60 is overwhelmingly the most frequent value, appearing 45,313 times. This suggests that a 60-day payment term is standard for most transactions. The next most common payment terms are 45, 30, and 55 days, but these are much less frequent, appearing 79, 23, and 13 times, respectively. Other terms, such as 35, 14, 40, and 50 days appear only a few times (less than 15 occurrences each). Rare terms like 21, 5, 49, and 0 days appear once. The high frequency of the 60-day payment term indicates a strong standardization in payment terms, with most clients adhering to this timeframe. Despite the dominance of the 60-day term, there is some variability, as indicated by the presence of other payment terms ranging from 0 to 55 days. This suggests flexibility in payment agreements in certain situations. Terms like 0, 5, 21, and 49 days are extreme outliers, possibly representing special agreements or errors in data entry. It is recommended to validate the outlier values to ensure they are correct and not data entry errors, as well as to confirm the reasons for these rare payment terms, thus determining whether they represent special agreements. Given the dominance of the 60-day term, it might be beneficial to review if this standard term aligns with the company's cash flow and financial strategies and consider whether the less frequent terms are necessary or if a more standardized approach would be beneficial.

Tab. 32
PAYMENT_TERM unique values and counts

PAYMENT_TERM [DAY]	Count
60	45,313
45	79
30	23
55	13
35	11
14	3
40	3
50	2
21	1
5	1
49	1
0	1

Further analyses could investigate the conditions under which non-standard payment terms are used to reveal insights into customer preferences or specific contractual agreements and analyze the impact of different payment terms on financial metrics such as receivables turnover and cash flow. This detailed understanding of the PAYMENT_TERM feature could provide insights into payment practices and inform strategic decisions regarding payment policies and customer agreements.

Tab. 33 shows unique values and counts for feature DOCUMENTS_BY. The analysis of the DOCUMENTS_BY feature reveals that the majority of transactions are documented by POST (with a count of 2,479), while only seven transactions are documented by MAIL. This significant disparity suggests that POST is the predominant method for handling documents in this dataset, indicating a strong preference or standard practice for using postal services for documentation. The minimal use of MAIL might reflect a limited need for electronic mail documentation or a specific policy favouring postal services. Notably, there is no information on other possible methods for handling documents in the dataset, as all 2,486 non-null entries out of 45,569 entries are either POST or MAIL. This means the DOCUMENTS_BY field is missing for the majority of the entries, highlighting a significant gap in the documentation process. Understanding the reasons behind the heavy reliance on POST and the absence of data for other methods could provide insights into the operational practices and potential areas for increasing efficiency or adopting more modern documentation methods.

Tab. 33
DOCUMENTS_BY unique values and counts

DOCUMENTS_BY	Count
POST	2479
MAIL	7

Tab. 34 shows unique values and counts for feature CUSTOMS. The analysis of the CUSTOMS feature reveals that an overwhelming majority of transactions (45,565 out of 45,569

entries) do not involve customs procedures, as indicated by the value of 0. Only four transactions involve customs, as indicated by the value of 1. This significant disparity suggests that customs procedures are extremely rare in this dataset and that most transactions likely occur within regions where customs clearance is not required, such as domestic transport or movement within a free trade area. The near absence of customs involvement highlights a streamlined logistics process, minimizing delays and complexities associated with customs procedures. This streamlined process likely contributes to operational efficiency and faster turnaround times for the majority of transactions.

Tab. 34
CUSTOMS unique values and counts

CUSTOMS	Count
0	45,565
1	4

5.2. Models Preselection

The mean absolute percentage error (MAPE) values presented in Tab. 35 demonstrate significant differences in predictive performance among the tested regression models. Ensemble tree-based methods, such as gradient boosting (8.11%), XGBoost (8.44%), and LightGBM (8.67%), achieved the lowest MAPE values, indicating their strong ability to capture complex nonlinear relationships and interactions between input features and the target variable. Random forest (9.11%) and extra trees (9.29%) also performed well, though they fell slightly short of the top-performing boosting models.

In contrast, simple approaches like single decision trees (9.86%) and the K-Neighbors Regressor (12.41%) exhibited high MAPE scores. While the decision tree model remained below the 10% threshold, the K-Neighbors Regressor exceeded it, suggesting that models without robust variance reduction mechanisms or sophisticated distance-based strategies struggle in more complex regression tasks.

Models yielding MAPE values above 10%, including linear techniques (e.g., ElasticNet with 25.56%, LASSO with 26.26%, ridge with 30.97%) and kernel-based methods (e.g., Support Vector Regressor with 21.67%, kernel ridge with 30.61%), are excluded from further analysis, as their inferior performance likely reflects an inability to fully capture nonlinearities or an inadequate parameterization under the current configuration.

Tab. 35
Preselected models comparison by MAPE

Model	MAPE [%]
Gradient Boosting	8.11
XGBoost	8.44
LightGBM	8.67
Random Forest	9.11
Extra Trees	9.29
Decision Tree	9.86
K-Neighbors Regressor	12.41

AdaBoost	19.74
Support Vector Regressor	21.67
ElasticNet Regression	25.56
Lasso Regression	26.26
Kernel Ridge	30.61
Ridge Regression	30.97
Gaussian Process Regressor	99.40
MLP Regressor	482.54

The Gaussian Process Regressor (99.40%) and MLP Regressor (482.54%) also demonstrated exceptionally poor fits, possibly due to suboptimal parameterization or the need for more complex feature engineering. These models are likewise excluded from subsequent examination.

In summary, the findings highlight the superior performance of ensemble boosting techniques and some tree-based models (i.e., those with MAPEs below 10%). These models are the focus of further analysis and optimization, given their efficacy in this regression context.

5.3. Cross-Validation Rationale

Fig. 19 compares the mean absolute percentage error (MAPE) values obtained from two validation strategies: cross-validation (CV) and time-based splitting. The primary goal is to demonstrate that the results derived from these two approaches are sufficiently comparable, indicating that the temporal dimension of the data does not strictly preclude standard cross-validation. Notably, the best-performing model (gradient boosting) shows only a minor difference in MAPEs between CV (8.11%) and time-split validation (7.62%), suggesting that cross-validation provides a robust estimate of model performance even when the data have a temporal ordering. While some models exhibit slightly higher MAPE values under time-split validation, these differences are not substantial enough to undermine the applicability of CV. The comparable outcomes imply that temporal dependencies within the dataset are not so pronounced as to invalidate cross-validation, thus enabling the efficient utilization of available data and maintaining methodological flexibility. In conclusion, this analysis supports the notion that cross-validation can be successfully employed, complementing more time-sensitive validation approaches without sacrificing reliability in performance assessment.

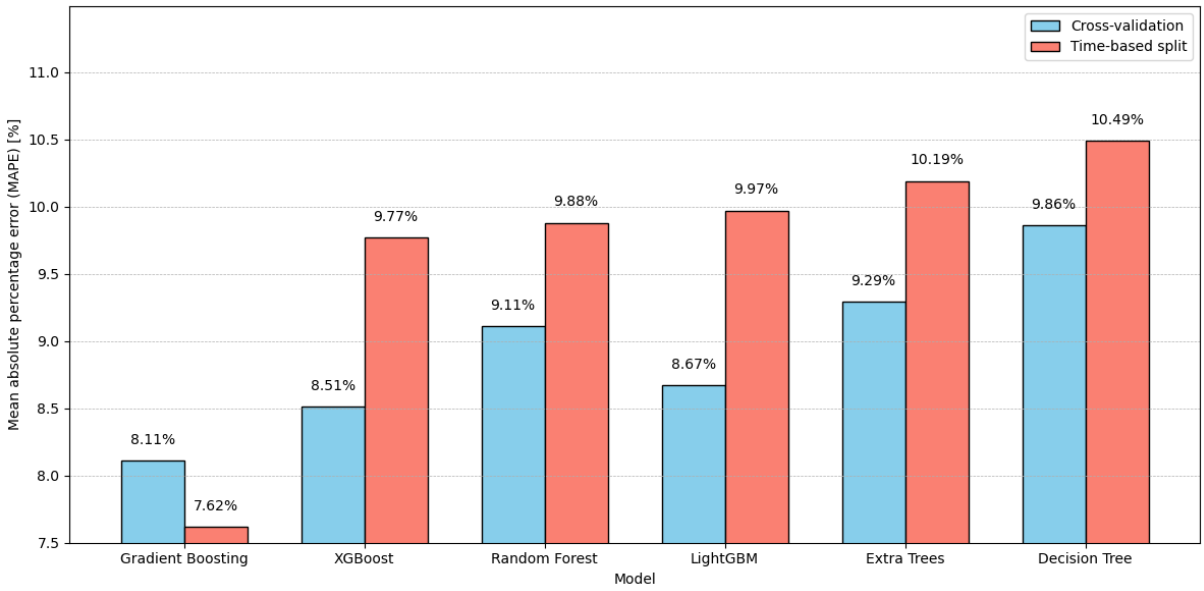


Fig. 19 Model comparison: cross-validation vs. time-based split

5.4. Impact of Training Dataset Size on Model Performance

Fig. 20 demonstrates the relationship between training dataset size and mean absolute percentage error (MAPE) across several machine learning models, including decision tree, random forest, gradient boosting, extra trees, and XGBoost. Initially, all models exhibit high MAPE values when trained on smaller datasets, with the decision tree model exceeding 25%, highlighting its limitations in capturing complex patterns with limited data. As the training dataset size increases, MAPE decreases significantly for all models, stabilizing below 10% for most once the dataset exceeds 10,000 samples. Among the models, random forest and gradient boosting consistently deliver superior performance, achieving MAPE values close to 5% with larger datasets, while the decision tree model demonstrates the highest variance and lowest overall accuracy. Beyond 20,000 samples, most models show diminishing returns in performance improvement, suggesting a saturation point in their learning capacity. This finding underscores the importance of sufficient training data in reducing prediction error while also indicating that further enhancements could focus on optimizing model complexity and feature engineering rather than solely increasing dataset size.

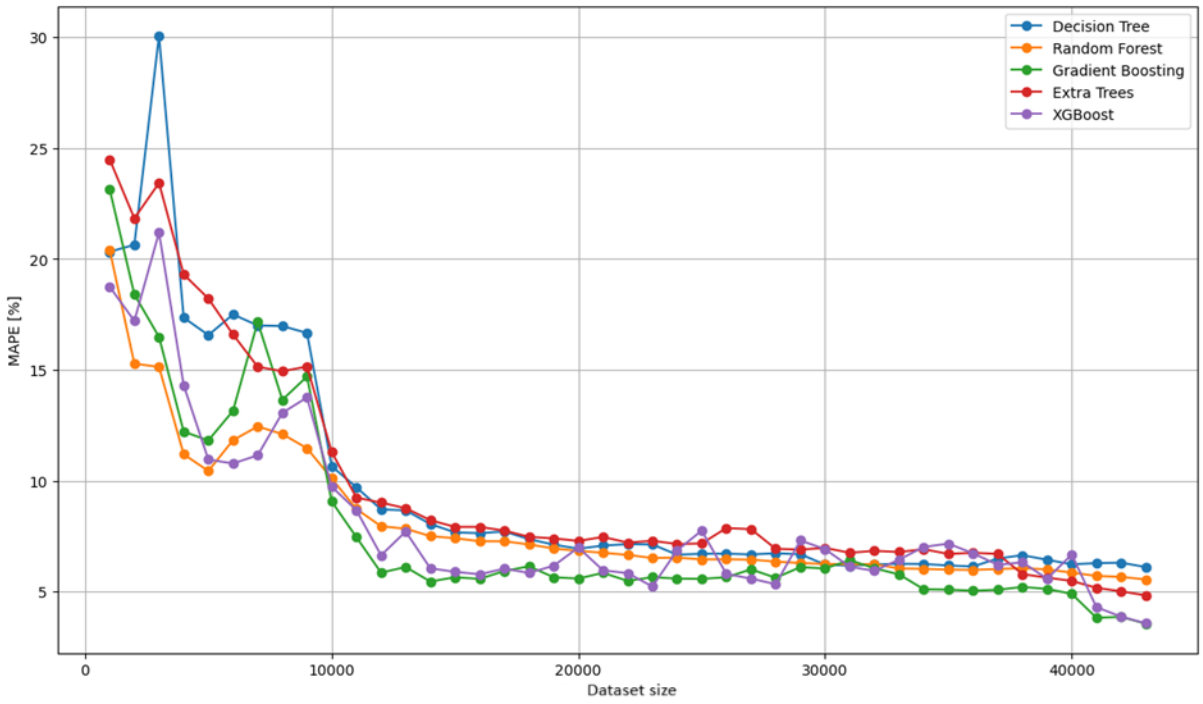


Fig. 20 Training dataset for different models comparison

5.5. Comparison of the Results from the Model and from Experts

A group of 172 experts estimated the pricing for five distinct transport scenarios. Fig. 21 illustrates the distribution of specialists by the type of enterprise in which they are employed. The largest groups of participants work in logistics and forwarding companies, each contributing nearly 40 specialists, highlighting the central role of these enterprises in road freight transport operations. Carrier companies follow with slightly fewer participants, reflecting their significant operational involvement. Trading firms represent a smaller share of the sample, while production companies account for the fewest participants, with fewer than 30 specialists. This distribution emphasizes the focus on enterprises directly engaged in managing and executing transport operations.

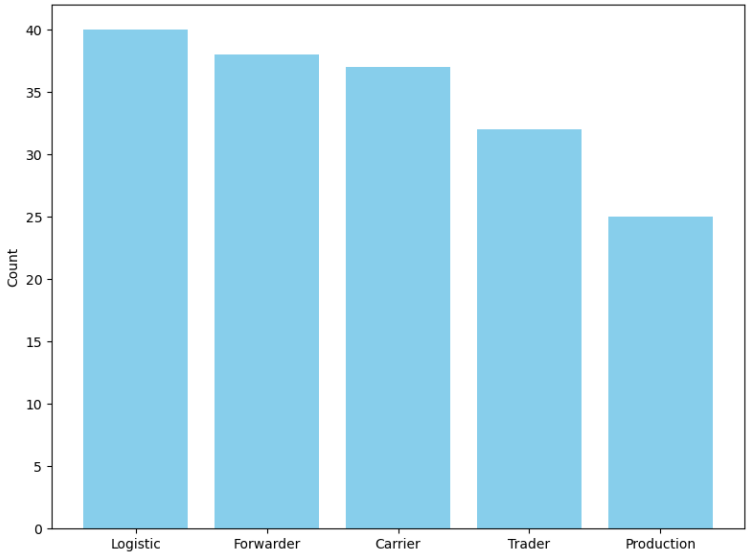


Fig. 21 Company type distribution

Fig. 22 demonstrates the distribution of professionals by their functional roles within the logistics and transport sector. The highest representation is observed among forwarders, with over 50 participants, emphasizing their pivotal role in coordinating freight operations. Logistic specialists and dispatchers closely follow, each contributing around 45 individuals, reflecting their integral involvement in managing and executing transport activities. Managers constitute a smaller group, with fewer than 35 participants, underscoring a relatively leaner layer of decision-makers. This distribution highlights the operational focus of the workforce, providing insights into the industry's role structure and the functions driving its efficiency.

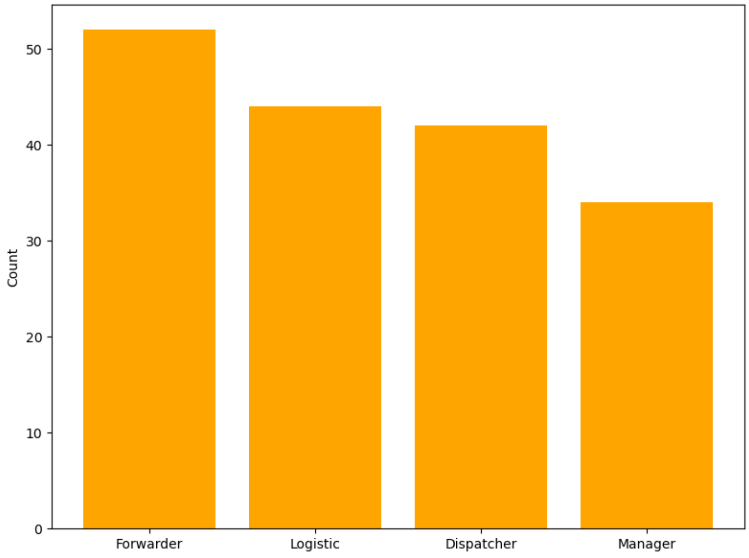


Fig. 22 Function distribution

Fig. 23 illustrates the distribution of specialists based on the size of the companies in which they are employed. Large companies, with more than 250 employees, and medium-sized companies, with 50 to 250 employees, have the highest representation, each accounting for over 40 participants. Small companies, employing 10 to 50 people, and micro-enterprises, with fewer than 10 employees, show slightly lower but comparable representation. This distribution highlights a balanced contribution from organizations of various sizes, providing diverse insights into how company scale may influence transport operations and pricing strategies.

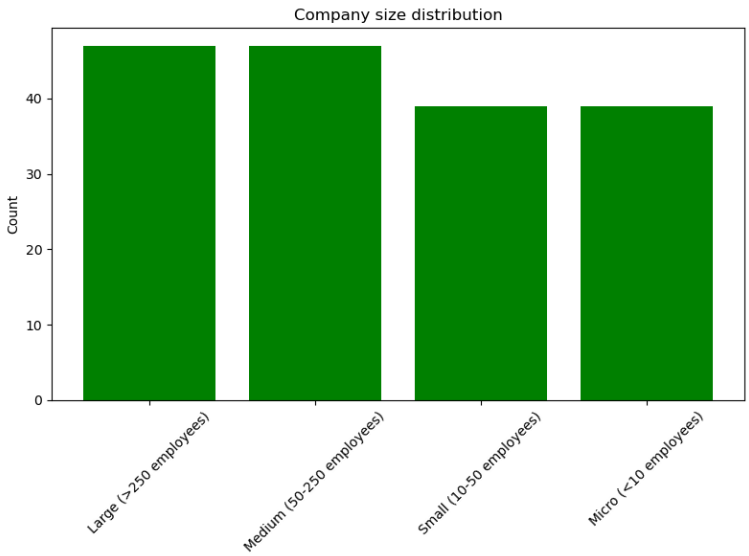


Fig. 23 Company size distribution

Fig. 24 illustrates the age distribution of specialists who participated in the study. The age groups with the highest representation are those between 20–25 and 55–60 years old, each exceeding 20 participants, suggesting significant contributions from both young and experienced professionals. The groups between 30–35 and 45–50 years old show noticeably lower representation, with fewer than 15 participants, indicating less engagement from mid-career specialists. Overall, the distribution reflects a diverse range of ages, with a balance between younger participants likely bringing innovative perspectives and older participants contributing extensive experience in the field. This diversity enhances the comprehensiveness and reliability of the study's findings.

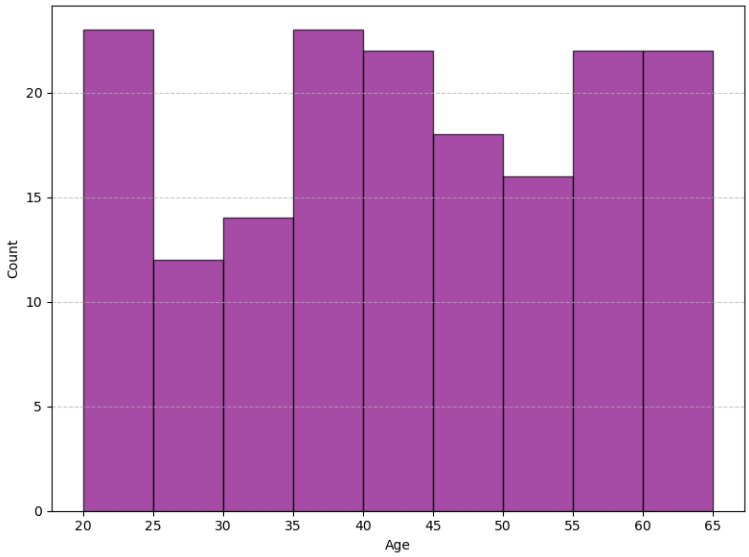


Fig. 24 Age distribution

Fig. 25 illustrates the distribution of observations based on years of experience. The X-axis represents experience ranges divided into five-year intervals (e.g., 0–5, 6–10 years), while the Y-axis shows the number of cases in each range. The highest number of observations is recorded in the range of 0–5 years, indicating that the majority of individuals in the analyzed group have relatively little professional experience. Subsequent intervals (6–10 and 11–15

years) also have significant representation, but the number of observations decreases as experience increases. In groups with over 20 years of experience, a stable but low level of observations is evident. The histogram reveals an uneven distribution of years of experience within the sample, with a dominance of individuals in the early stages of their careers.

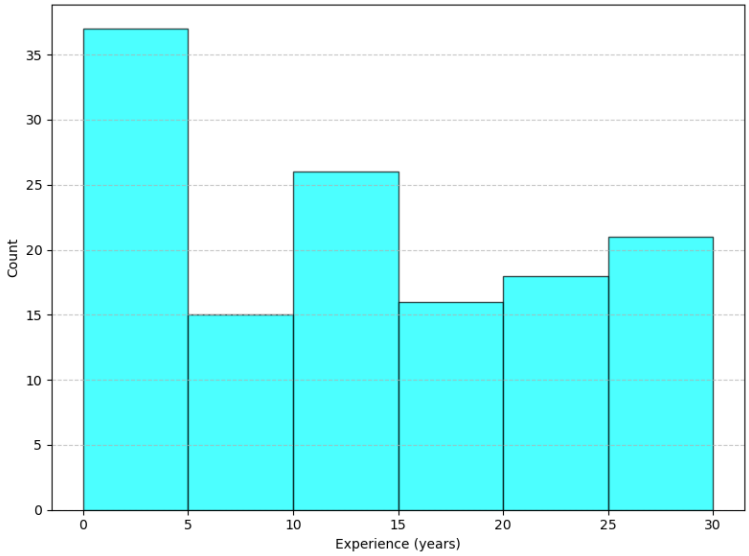


Fig. 25 Experience in transport distribution

In this study, a validation dataset was created comprising five transport offers reported after the training data period. This dataset serves as a benchmark to evaluate the performance of various forecasting methods. The comparison of methods is based on the mean absolute percentage error (MAPE) metric, ensuring a consistent and interpretable measure of accuracy.

Tab. 36 presents the features of the five transport offers used for validation. The dataset includes key features such as the distances covered in different countries, total kilometres travelled, and detailed information about the transport process, including loading and unloading dates, times, vehicle types, body types, and cargo characteristics. These features provide a comprehensive view of the transport offers, allowing for robust validation of the predictive models. The data highlight the diversity of transport scenarios, which is critical for testing the generalizability and effectiveness of the proposed methods.

By integrating expert calculations into the validation process, the models' predictions can be cross-verified and refined, ensuring they align closely with industry practices and real-world conditions. This hybrid approach of combining machine learning with expert insights enhances the credibility and practical applicability of the forecasting methodology.

Tab. 36
Collection of five offers for validation

Feature	Transport 1	Transport 2	Transport 3	Transport 4	Transport 5
CZ_KM	0	0	348.1	410	0
DE_KM	542.6	395.4	0	0	0
PL_KM	438.4	439.8	336.1	333.9	487.3
SK_KM	0	0	0	0	99.4
KM	981	835.2	684.2	743.9	586.7
COD_LP	PL96200	DE98673	PL86300	CZ38732	PL98100
COD_DP	DE59174	PL32400	CZ39101	PL37100	SK07101

START_LOAD_DATE	18.07.2024	22.07.2024	22.07.2024	23.07.2024	23.07.2024
START_LOAD_TIME			06:00:00		
END_LOAD_DATE	18.07.2024	04.08.2024	22.07.2024	23.07.2024	24.07.2024
END_LOAD_TIME			13:00:00		
START_DELIVERY_DATE	22.07.2024	23.07.2024	23.07.2024	24.07.2024	24.07.2024
START_DELIVERY_TIME			08:00:00		
END_DELIVERY_DATE	22.07.2024	05.08.2024	23.07.2024	24.07.2024	25.07.2024
END_DELIVERY_TIME			21:00:00		
VEHICLE_TYPE	articulated truck	articulated truck	articulated truck, rigid truck	articulated truck	articulated truck
BODY_TYPE	standard	standard	box, refrigerator	jumbo, mega, standard	standard
LOAD_UNLOAD_METHOD	back, side	back, side	back	back, side	back, side
GOODS_TYPE	neutral	neutral	neutral	neutral	neutral
LDM	13.6	13.6	13.6	13.6	13.6
M3	85	85	85	85	85
HEIGHT	2.6	2.6	2.6	2.6	2.6
WIDTH	2.4	2.4	2.4	2.4	2.4
TONS	24	24	9.5	24	24

The results of these estimations are summarized in the presents data analysis based on the mean, median, and standard deviation of the price per kilometre, particularly in the context of the initial loading date, which holds significant importance and implications. The calculation of statistical measures begins with the mean, which represents the average price per kilometre for transport services starting on specific dates, providing insights into typical pricing trends. The median offers the middle value in the distribution of prices per kilometre, making it less sensitive to outliers and more reflective of central tendencies in skewed datasets. The standard deviation captures the variability of the price per kilometre around the mean, highlighting how much prices fluctuate over time and offering an indicator of the reliability of the mean as a measure of central tendency.

Flexibility in loading dates is an additional factor considered in this analysis. When the exact loading date is not critical, a flexible date can be selected, allowing transport companies to adapt operations more effectively to dynamic circumstances. This flexibility has a direct impact on pricing, as it often results in lower costs due to optimized fleet utilization and reduced expenses associated with empty runs. Together, these statistical insights and operational strategies contribute to a more accurate and efficient pricing model for road freight transport.

Tab. 16

Tab. 37 shows predicted values, expressed in euros (EUR), varied across the transports, with corresponding absolute percentage errors (APE) indicating the accuracy of these predictions. The experts' pricing predictions range from €600 for Transport 4 (which exhibited the highest percentage error of 17.83%) to €1200 for Transport 1 (which had the lowest error at 7.39%). The mean percentage error (MPE) across all transports was 11.86%, reflecting the overall accuracy and consistency of the expert estimations. These findings highlight the variability in expert judgment and underscore the challenges associated with accurately predicting transport costs.

Tab. 37
Predictions with the experts method

Transport	EUR	Absolute percentage error
1	1200	7.39%
2	700	11.45%
3	800	10.25%
4	600	17.83%
5	750	12.33%
	MPE	11.86%

The data present a comparative analysis of the prediction errors between human estimators and a predictive model over five distinct transport instances. The error metrics used are the mean percentage error for human estimators and the percentage error for the model. The results are summarized in Fig. 26.

Across all five transports, the model consistently exhibits lower percentage errors than human estimators. For Transport 1, the model's error (3.77%) is significantly lower than the human mean error (7.40%), suggesting a notable efficiency of the model in this instance. For Transport 2, both human and model errors increase compared to Transport 1, but the model maintains a lower error margin (6.67%) relative to human estimators (11.47%). For Transport 3, similar to Transport 2, the model's error (6.44%) remains lower than the human error (10.26%), demonstrating consistent model performance. Transport 4 exhibits the highest errors for both humans (17.80%) and the model (13.19%). Despite the increase, the model still shows better accuracy. The errors for Transport 5 decrease for both humans (12.36%) and the model (11.73%) compared to Transport 4, with the model maintaining a slight edge in accuracy.

The analysis reveals that the predictive model outperforms human estimators in all examined transport instances. The reduced error margins of the model suggest higher reliability and precision in forecasting transport outcomes. This performance can be attributed to the model's ability to systematically process and analyze large datasets, reducing the influence of biases and errors inherent in human judgment.

The comparative study underscores the efficacy of predictive models in transport estimation tasks. The consistently lower error percentages of the model across all transports highlight its potential as a tool for prediction accuracy that can enhance decision-making in logistics and transport management. Future research could further explore the factors contributing to human errors and refine the model to address the challenges observed in higher-error transports, such as Transport 4. This analysis demonstrates the practical advantages of integrating predictive models in operational environments, ultimately leading to more accurate and efficient outcomes.

The average error of the experts was 11.86%. The model’s average was 9.28%, indicating an improvement of 2.58 percentage points.

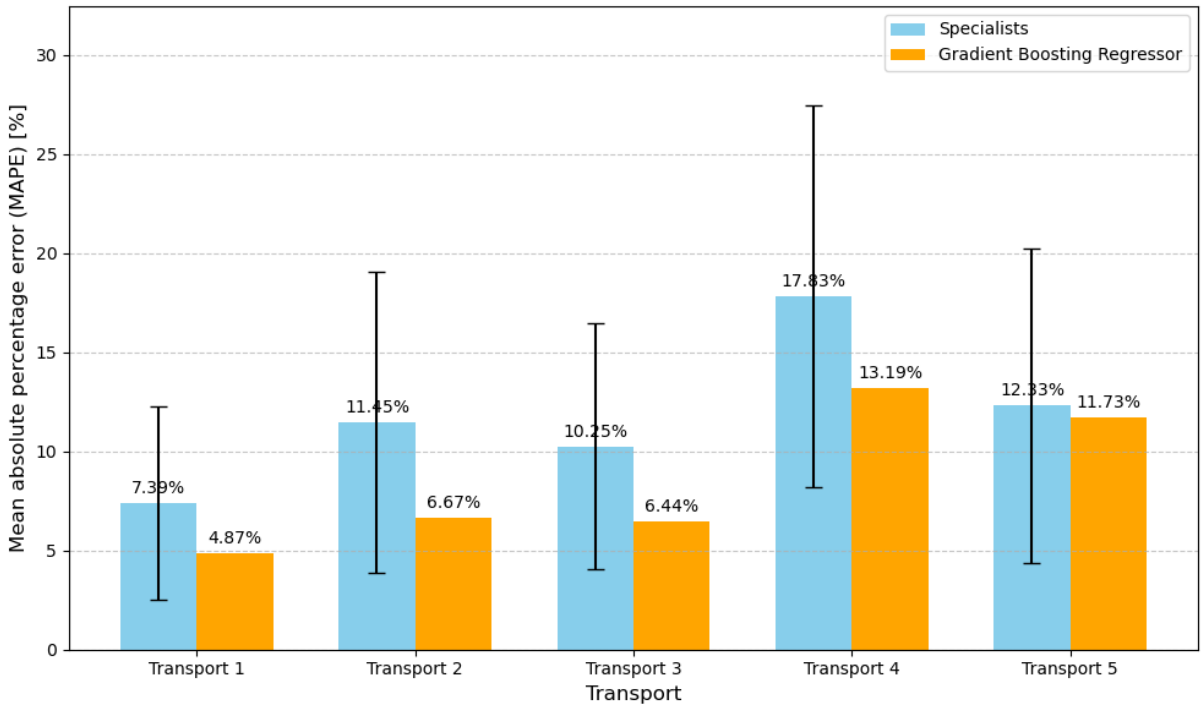


Fig. 26 Comparison of mean absolute percentage error for human predictions and model predictions

5.6. Implementation of External Databases

The data analysis presented in Fig. 27 shows a clear upward trend in both fuel prices and transportation costs over the examined period. There is a positive correlation between fuel prices and transportation costs, with a correlation coefficient of 0.14. This suggests that increases in fuel prices can directly impact the rise in transportation costs, which is typical in the transport industry, where fuel price is a significant operational cost factor. Additionally, periodic price spikes can be observed, which may be due to various factors such as global market fluctuations, changes in energy policy, or seasonal changes in transportation demand.

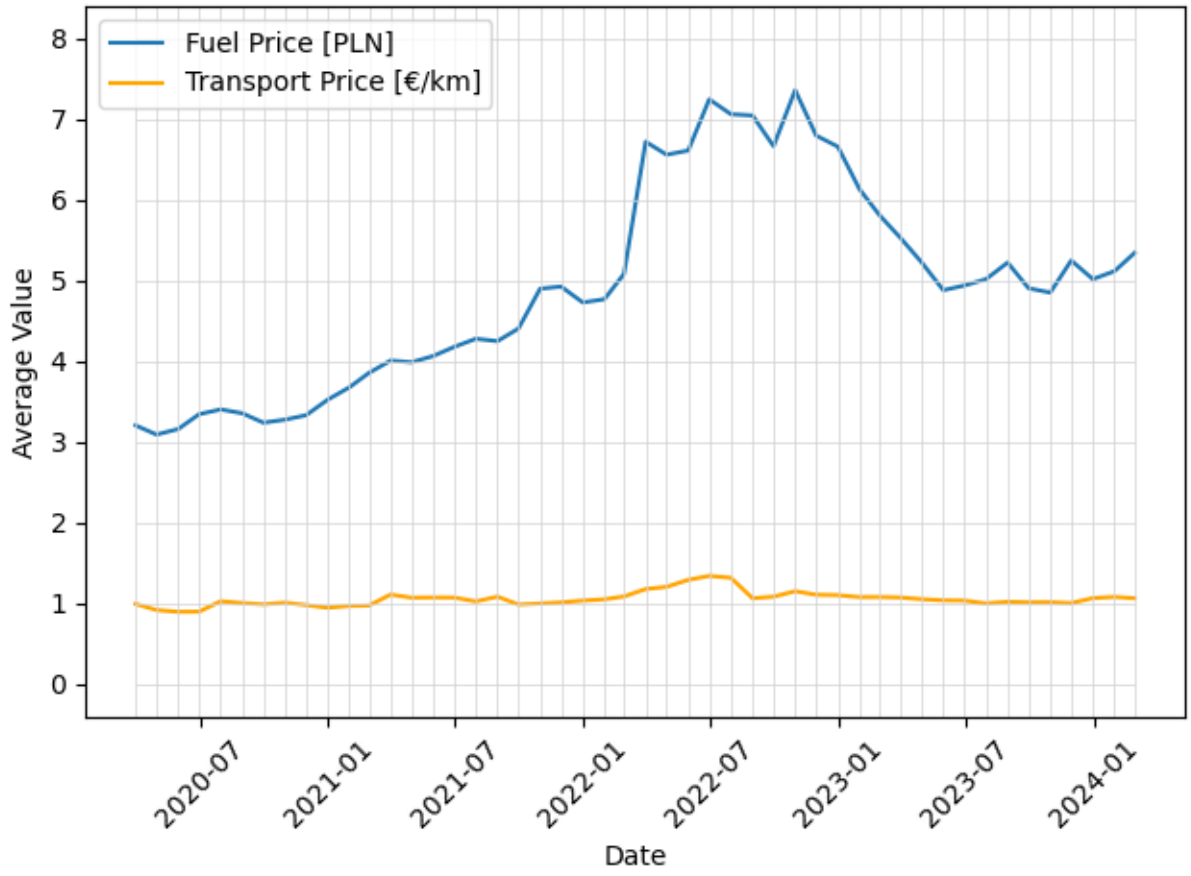


Fig. 27 Relationship between transportation costs and fuel prices

The correlation analysis (Tab. 38) between the rate per kilometre (EUR_FOR_KM) and fuel price (FUEL_PRICE) for different combinations of route and transport type provides important insights. It is noteworthy that some combinations exhibit perfect correlations of 1.00. Examples include routes from CZ to CZ, FR to DE, and HU to CZ, along which single trips show a correlation of 1.00, with a small sample size of only three observations per combination, which is too small to establish a reliable dependence.

High correlations close to 1.00, such as the relationship from PL to DK (0.99) and PL to IT (0.93), also suggest a strong positive dependency between fuel price and the rate per kilometre. Conversely, some combinations show lower correlations, such as PL to EE (-0.02) and CZ to SK (-0.50), indicating either no noticeable dependency or even a negative relationship between fuel price and the rate per kilometre for these combinations. Furthermore, combinations with larger sample sizes tend to exhibit more stable correlation results than those with fewer observations. The average correlation result for the overall relationship is 0.34, which is higher than the correlation examined in the entire dataset.

Tab. 38
Relationship between transport costs and fuel prices by route

Route	Type	Correlation Coefficient	Sample Size
CZ to CZ	Single	1	3
FR to DE	Single	1	3
HU to CZ	Single	1	3
PL to DK	Single	0.99	6
LV to CZ	Single	0.94	6
PL to IT	Single	0.93	21
PL to BE	Single	0.83	147
PL to FR	Single	0.76	9
LT to PL	Single	0.73	12
PL to NL	Single	0.72	47
SK to LT	Single	0.71	6
LT to HU	Single	0.7	5
PL to LT	Single	0.69	19
PL to AT	Single	0.61	20
IT to PL	Single	0.6	69
PL to CZ	Round Trip	0.6	242
PL to HU	Round Trip	0.56	413
LT to SK	Single	0.55	9
CZ to SE	Single	0.53	5
SE to SE	Single	0.5	3
DE to EE	Single	0.5	3
PL to HR	Single	0.5	3
LT to CZ	Single	0.45	7
HU to PL	Single	0.4	27
PL to DE	Single	0.4	5776
PL to SE	Single	0.37	9
FR to DE	Round Trip	0.33	6
PL to SK	Single	0.32	833
DE to DE	Round Trip	0.32	49
PL to DE	Round Trip	0.31	1621
PL to FI	Single	0.28	11
DE to PL	Single	0.24	4514
PL to LV	Single	0.23	9
PL to HU	Single	0.2	184
CZ to CZ	Round Trip	0.19	74
PL to PL	Single	0.19	1409
DE to PL	Round Trip	0.16	174
PL to PL	Round Trip	0.16	20,199
CZ to LT	Single	0.11	7
SK to PL	Single	0.1	99
EE to PL	Single	0.1	5
DE to CZ	Single	0.09	49

CZ to PL	Single	0.07	1840
CZ to DE	Single	0.06	82
PL to CZ	Single	0.06	7437
PL to EE	Single	-0.02	7
CZ to SK	Single	-0.5	5
LV to DE	Single	-0.5	3
SE to PL	Single	-0.5	3
DE to FR	Single	-0.5	3
CZ to EE	Single	-0.6	4
SK to CZ	Single	-0.7	5

The analysis of GDP data in Tab. 39, spanning from Q4 2019 to Q4 2023 for EU countries, provides significant insights into the economic growth dynamics across different regions. These data are organized in a tabular format; the columns represent successive quarters starting from Q4 2019, and the rows correspond to individual countries identified by unique country codes. The values in the table cells represent GDP figures.

The analysis reveals variability in economic growth dynamics among countries. For instance, countries like Austria (AT), Belgium (BE), and Denmark (DK) exhibited stable GDP growth over the period with minor fluctuations. In contrast, other countries such as Greece (GR) and Italy (IT) experienced more significant variations in their economic growth. Moreover, certain pan-European trends can be observed, such as the GDP decline in Q2 2020, likely associated with the impact of the COVID-19 pandemic on the region's economy. However, as national and international adaptive strategies were implemented, many countries began to show economic recovery in subsequent quarters, demonstrating the European economy's resilience and flexibility in facing challenges.

Tab. 39
EU countries GDP

	PKB [thousand €]																
	2019	2020				2021				2022				2023			
Country	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
DE	892	864	786	863	890	865	882	923	948	943	950	977	1007	1011	1008	1035	1067
FR	628	581	527	588	622	603	618	624	656	638	657	6e55	690	682	700	691	730
IT	477	405	376	425	456	420	449	460	493	459	487	488	529	499	518	517	551
ES	324	289	250	281	299	282	303	304	333	315	337	334	360	350	367	360	385
NL	208	201	192	197	207	203	220	218	229	227	242	238	252	250	263	255	266
PL	149	129	117	133	146	131	137	146	162	148	156	165	185	170	180	188	211
BE	128	116	106	114	124	119	126	125	138	132	139	135	148	142	146	142	153
SE	125	117	115	117	131	126	136	130	148	133	145	138	146	136	139	130	144
AT	103	95	89	97	100	92	101	104	109	104	113	112	118	116	120	118	124
DK	80	76	74	78	82	79	85	86	93	89	96	97	99	93	94	90	97
RO	68	46	47	60	68	47	56	66	73	55	67	78	85	64	75	88	97
FI	63	58	58	59	63	58	63	63	68	63	67	67	71	67	70	69	72
CZ	60	53	50	55	57	53	60	62	64	63	69	71	74	72	79	77	78
PT	55	51	46	52	53	49	54	56	57	56	61	62	63	63	67	68	68
GR	46	41	38	44	42	40	44	50	48	46	51	57	53	49	55	60	56

HU	40	33	31	35	38	31	38	40	44	37	42	42	46	40	50	51	56
SK	24	22	22	25	25	22	25	26	26	25	27	29	29	28	30	32	32
BG	17	13	14	17	17	15	17	20	21	17	20	23	25	21	22	25	27
LU	17	15	15	16	18	17	18	18	20	18	19	19	20	19	20	19	21
HR	14	12	12	14	13	12	14	17	15	14	17	20	17	16	19	22	19
LT	13	11	12	14	13	12	14	15	15	15	16	19	18	16	18	19	19
SL	13	11	11	12	12	12	13	14	14	13	14	15	15	14	16	16	16
LV	8	7	7	8	8	7	8	9	9	8	10	10	10	9	10	10	11
EE	7	6	6	7	7	7	8	8	9	8	9	9	10	9	9	10	10
CZ	6	6	5	6	6	6	6	7	7	6	7	7	7	7	7	8	8
M	4	3	3	3	3	4	4	4	4	4	4	5	5	4	5	5	5

The analysis of the territorial surface area of various European countries, as presented in Tab. 40, allows for a deeper understanding of their geographical size and spatial proportions. The table lists the country code and its corresponding area in km². These countries represent a geographical and demographic diversity across Europe, from large continental nations like France (FR), Spain (ES), and Sweden (SE) to smaller countries with limited areas such as Luxembourg (LU) and Malta (M). The largest country in terms of area is France (over 600,000 km²), while the smallest, Malta, covers less than 400 km². This analysis aids in comprehending the geographical scale and spatial distribution of European countries.

Tab. 40
Surface area of EU countries

Country	Area [k²]
FR	633,886
ES	502,654
SE	407,300
DE	353,296
PL	307,236
FI	304,316
IT	297,825
RO	234,270
GR	130,048
BG	110,001
HU	91,248
PT	90,996
AT	82,519
CZ	77,212
LV	63,290
LT	62,643
HR	55,896
SK	48,702
EE	43,110
DK	41,987
NL	34,188
BE	30,452

SL	20,145
CY	9213
LU	2586
M	313

The above data are used to calculate each country's GDP per square kilometre. The formulas are described as follows:

Eq. (8): GDP per square kilometre (RGFKL) for the loading country. The GDP of the loading country is divided by its land area (LA) in km², resulting in GDP per unit area.

$$\mathrm{RGFKL} = \frac{\mathrm{LG}}{\mathrm{LA}}, \tag{8}$$

where: LG – GDP of the loading country; LA – land area of the loading country (km²); RGFKL – GDP per square kilometre.

Eq. (9): GDP per square kilometre (RGFKD) for the unloading country. The GDP of the unloading country is divided by its land area (DA) in km², resulting in GDP per unit area.

$$RGFKD = \frac{DG}{DA}, \tag{9}$$

where: DG – GDP of the unloading country; DA – land area of the unloading country (km²); RGFKD – GDP per square kilometre

Eq. (10): The ratio between the GDP per square kilometre of the loading country (RGFKL) and the GDP per square kilometre of the unloading country (RGFKD). This value is used to compare the economic parameters of the loading and unloading countries.

$$RGFKLTD = \frac{RGFKL}{RGFKD}, \tag{10}$$

where: RGFKLTD – the ratio between the GDP per square kilometre of the loading and unloading countries.

The hypothesis that transport service price depends on the GDP difference between the loading and unloading countries is tested. Data were collected on 31 bilateral relations. For each transport, the GDP ratio for the respective quarter was calculated, and the average for the entire relation was computed. The results are shown in Tab. 41. Relations where transport from a country with a higher GDP per km² to a country with a lower GDP per km² is more expensive than the reverse are marked in green. Relations showing the opposite trend are marked in red. In most cases (68% of bilateral relations), the hypothesis that transport from a more developed to a less developed country is more expensive than vice versa is confirmed.

Tab. 41

Dependency of transport rates on GDP

Relation	GDP Ratio (Loading to Unloading)	Rate [€/km]	Number of Occurrences
PL to DE	0.21	1.26	5004
PL to CZ	3.47	1.17	1454
PL to HU	1.14	1.02	458
PL to SK	0.97	1.31	186
DE to CZ	16.71	0.86	86
IT to PL	2.88	1.11	61
PL to LT	2.04	1.07	15
CZ to LT	1.87	0.95	14
LV to CZ	0.92	0.80	12
CZ to SE	1.28	1.86	10

Relation	GDP Ratio (Loading to Unloading)	Rate [€/km]	Number of Occurrences
DE to PL	4.82	0.89	2513
CZ to PL	0.90	0.68	1182
HU to PL	0.90	0.66	7
SK to PL	1.05	0.74	7
CZ to DE	0.18	0.86	4
PL to IT	0.32	1.09	21
LT to PL	0.49	0.63	11
LT to CZ	1.58	0.78	14
CZ to LV	3.12	0.80	4
SE to CZ	2.22	0.91	2

LT to SK	0.45	0.87	9	SK to LT	2.17	0.88	6
PL to LV	3.63	1.04	9	LV to PL	0.26	0.74	2
PL to SE	1.41	1.85	9	SE to PL	0.67	0.65	2
CZ to EE	2.34	1.04	8	EE to CZ	1.24	0.75	2
PL to EE	2.58	1.08	7	EE to PL	0.39	0.57	5
HU to CZ	2.81	0.81	6	CZ to HU	1.10	1.16	2
LT to HU	0.54	1.03	5	HU to LT	2.02	0.85	2
LV to DE	0.05	1.10	3	DE to LV	17.91	0.99	2
EE to LT	0.79	0.62	2	LT to EE	1.26	1.30	1
DE to LT	10.43	0.93	2	LT to DE	0.09	1.00	2
DE to HU	7.10	1.12	1	HU to DE	0.15	1.11	1
DE to SE	7.88	2.10	1	SE to DE	0.12	1.26	1
DE to SK	4.88	1.01	1	SK to DE	0.19	1.29	1
EE to HU	0.45	0.94	1	HU to EE	2.23	0.86	1
EE to SE	0.59	1.92	1	SE to EE	1.73	1.03	1
EE to SK	0.37	0.72	1	SK to EE	2.68	0.77	1
HU to LV	3.08	0.90	1	LV to HU	0.32	0.98	1
HU to SE	1.30	1.54	1	SE to HU	0.77	1.04	1
LT to SE	0.74	3.20	1	SE to LT	1.39	0.92	1
LV to SK	0.27	0.80	1	SK to LV	3.70	0.80	1
SE to SK	0.63	2.31	1	SK to SE	1.55	2.35	1

The analysis of the number of registered trailers in various countries (Tab. 42) can be interpreted as an indicator of economic activity and the dynamics of the transport sector in a given region. An increase in the number of registered trailers between the studied years may suggest the development of transport infrastructure and increased demand for transport services due to intensified economic activity. Countries with stable economic growth and a central geographical location may tend to have more registered trailers, which play a key role in international trade and freight transport. Additionally, government policy, transport regulations, and investments in road infrastructure can influence the dynamics of trailer registration growth in individual countries. Factors such as seasonality, consumer trends, and changes in international trade can significantly affect the number of registered trailers over different periods. Therefore, analyzing these data can provide valuable information for policymakers, researchers, and businesses in the transport and logistics sector.

Tab. 42

Number of registered trailers [87]

Country	Number of registered trailers	
	2021	2022
AT	38,598	39,552
BE	114,399	117,485
BG	60,772	61,984
HR	16,516	17,342
CY	8,46	8933
CZ	40,666	37,052
DK	48,229	50,848
FI	39,924	40,790

FR	322,785	327,181
DE	383,027	398,452
LV	125,609	129,762
LT	16,306	17,273
LU	45,630	49,865
M	4696	4756
NL	177,294	185,262
PL	503,812	532,762
PT	38,816	39,932
RO	147,513	156,501
SK	28,412	28,880
SL	13,563	14,024
ES	286,474	293,266
SE	32,173	34,365

The analysis of registered trailers per square kilometre in various countries (Fig. 28) shows significant variations among regions. Several factors may contribute to these differences. Population density and the economic structure of individual countries can affect transport needs and trailer usage. Countries with lower population densities, such as Finland and Sweden, may exhibit lower values due to a reduced demand for freight transport relative to the number of inhabitants. Conversely, countries with high population density, like Luxembourg and the Netherlands, may have higher transport demands and, consequently, higher values. This dependence may also result from more favourable tax regulations for businesses in countries like Luxembourg and Malta. Furthermore, these differences may also stem from variations in transport infrastructure, transport policy, and the availability and cost of freight transport. Countries with well-developed highway networks and logistics infrastructures may be more attractive to transport companies, which can contribute to higher values.

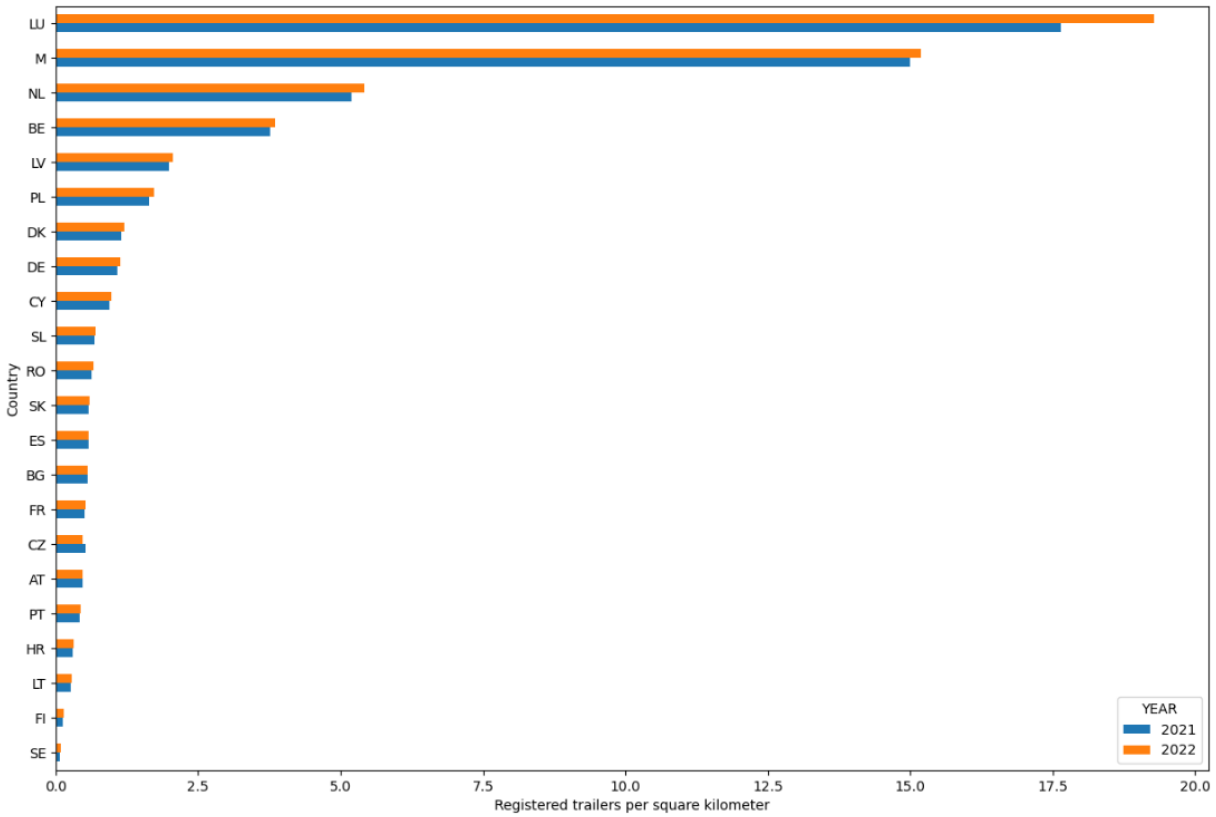


Fig. 28 Number of registered trailers per km² by country

The above data are used to calculate the ratio of the number of registered trailers per square kilometre for each country. Eq. (11): Number of registered trailers per square kilometre (TFAL) for the loading country. The number of registered trailers in the loading country is divided by its land area (LA) in km², resulting in the number of trailers per unit area.

$$\mathrm{TFAL} = \frac{\mathrm{LT}}{\mathrm{LA}}, \tag{11}$$

where: LT – number of registered trailers in the loading country; LA – land area of the loading country (km²); TFAL – number of trailers per square kilometre.

Eq. (12): Number of registered trailers per square kilometre (TFAD) for the unloading country. The number of registered trailers in the unloading country is divided by its land area (DA) in km², resulting in the number of trailers per unit area.

$$\mathrm{TFAD} = \frac{\mathrm{DT}}{\mathrm{DA}}, \tag{12}$$

where: DT – number of registered trailers in the unloading country; DA – land area of the unloading country (km²); TFAD – number of trailers per square kilometre.

Eq. (13): The ratio between the number of registered trailers per square kilometre in the loading country (TFAL) and the number of registered trailers per square kilometre in the unloading country (RTFALTD). This value is used to compare the economic parameters of the loading and unloading countries.

$$\mathrm{RTFALTD} = \frac{\mathrm{TFAL}}{\mathrm{TFAD}}, \tag{13}$$

where: RTFALTD – the ratio between the number of registered trailers per square kilometre of the loading and unloading countries.

The analysis (Tab. 43) of the ratio of the number of registered trailers per square kilometre of the loading and unloading countries, along with the corresponding euro rates per kilometre, reveals the complex dynamics of international road transport in Europe. Green indicates

relations where transport from a country with a higher trailer ratio to a country with a lower ratio is more expensive than the reverse. The table presents 34 bilateral relations. In 28 cases (88%), the price from a country with a higher ratio to a country with a lower ratio is higher. Poland (PL) shows a higher ratio of trailers per square kilometre than other countries, especially Sweden (SE) and Lithuania (LT), where these ratios are exceptionally high. For transport from Poland to Germany (DE), the Czech Republic (CZ), and Slovakia (SK), the euro rates per kilometre are relatively high, reflecting long distances or other factors affecting transport costs. Conversely, relations between countries with relatively low trailer ratios per square kilometre, such as Denmark (DK), Latvia (LV), and Sweden (SE), often exhibit lower euro rates per kilometre. The relatively low trailer ratios per square kilometre in the Czech Republic (CZ) and Slovakia (SK) may result from the small areas of these countries, which can affect traffic intensity. However, despite differences in trailer ratios, euro rates per kilometre do not always reflect these differences, suggesting that additional factors such as operational costs or market competition influence transport pricing.

Tab. 43

Dependency of transport rates on the number of registered trailers

Relation	RTFALTD	[€/km]	Number of Occurrences	Relation	RTFALTD	[€/km]	Number of Occurrences
PL to DE	1.53	1.25	5708	DE to PL	0.65	0.90	2958
PL to CZ	3.58	1.18	835	CZ to PL	0.28	0.68	638
PL to SK	2.91	1.31	206	SK to PL	0.35	0.74	7
CZ to DE	0.43	0.86	2	DE to CZ	2.35	0.87	47
PL to LT	6.29	1.07	15	LT to PL	0.16	0.63	11
LT to SK	0.45	0.87	9	SK to LT	2.18	0.88	6
CZ to LT	1.94	0.95	7	LT to CZ	0.52	0.78	7
PL to SE	20.76	1.85	9	SE to PL	0.05	0.65	2
LV to PL	1.20	0.74	2	PL to LV	0.84	1.04	9
CZ to LV	0.27	0.80	2	LV to CZ	4.02	0.80	6
DK to PL	0.70	0.79	2	PL to DK	1.43	1.42	6
CZ to SE	6.67	1.86	5	SE to CZ	0.18	0.91	1
DE to LV	0.55	0.99	2	LV to DE	1.82	1.10	3
DE to LT	4.13	0.93	2	LT to DE	0.24	1.00	2
DE to SE	13.73	2.10	1	SE to DE	0.07	1.26	1
DE to SK	1.86	1.01	1	SK to DE	0.53	1.29	1
LT to SE	3.27	3.20	1	SE to LT	0.31	0.92	1
LV to SK	3.46	0.80	1	SK to LV	0.29	0.80	1
SE to SK	0.14	2.31	1	SK to SE	7.03	2.35	1

The distance-based road toll data for Euro VI 4-axle vehicles was collected. The data sources are the Austrian system GO BOX, Belgian Viapass (prices are different for each of the three regions, and the price for the largest in terms of area, Wallonia, was adopted), Czech myto, and German Toll Collect. Denmark, the Netherlands and Sweden use the Eurovignette. Other systems are free or have varying prices per kilometre depending on the section.

Tab. 44 presents the performance results of several machine learning models used for forecasting road freight transport prices, evaluated based on various external datasets. The models included are gradient boosting, XGBoost, LightGBM, random forest, extra trees, and decision tree. The evaluation metric is the mean absolute percentage error (MAPE), with models showing less than 10% error included in this analysis.

Basic Model: The basic model without external databases serves as the baseline for comparison. Gradient boosting exhibits the lowest error at 8.11%, followed by XGBoost (8.44%) and LightGBM (8.67%). Random forest, extra trees, and decision tree models show higher errors, with the decision tree model performing the worst at 9.86%.

Impact of Road Tolls: Adding road toll data slightly improves the performance for some models, such as LightGBM (8.47% vs. 8.67% basic), but the improvement is not significant across the board. The gradient boosting model's error remains almost unchanged at 8.12%.

Impact of Fuel Prices: Incorporating fuel prices generally leads to better performance. The gradient boosting model achieved the lowest error (7.98%), indicating that fuel prices are a significant factor in forecasting. Most models experienced a reduction in error, with the mean for this feature configuration being 8.78%, the lowest among all individual feature additions.

Impact of GDP: Adding the GDP coefficient increased the error for all models, indicating that this feature might introduce noise or is not as relevant for this specific forecasting task. The mean error for this configuration is the highest at 9.13%.

Impact of Registered Trailers: Including information about registered trailers improves the performance of most models. The gradient boosting model benefits the most, achieving an error of 7.94%, the lowest for this feature set.

Combined Features Analysis: The fuel and trailer combination provides robust performance, with the gradient boosting model achieving an error of 7.91%. The overall mean for this configuration is 8.71%.

Adding road tolls, fuel prices, and registered trailers yields the lowest errors across all configurations, with a mean error of 8.64%. The gradient boosting model leads with an error of 7.90%. The combination of road tolls, fuel prices, registered trailers, and GDP had an increase in errors, with a mean of 9.08%, suggesting that the addition of GDP does not contribute positively.

Model Comparison: Gradient boosting consistently outperformed other models across all configurations, achieving the lowest errors in most cases.

XGBoost and LightGBM also performed well but slightly lagged behind gradient boosting.

Random forest and extra trees showed moderate performance, with errors generally around 9%.

Decision tree performed the worst among the selected models, indicating it may not be suitable for this forecasting task without further optimization.

The analysis highlights that combining road tolls, fuel prices, and registered trailers provides the best model performance, with gradient boosting emerging as the best performer. The inclusion of GDP does not appear beneficial, potentially due to the specific nature of the transport pricing model. This detailed evaluation underscores the importance of selecting relevant external features and robust algorithms to improve predictive accuracy in road freight transport pricing.

Tab. 44
Mean absolute percentage error (MAPE) of machine learning models

model	basic	toll	fuel	GDP	trailers	fuel, trailers	toll, fuel, trailers	toll, fuel, trailers, GDP	mean
Gradient Boosting	8.11	8.12	7.98	8.57	7.94	7.91	7.9	8.59	8.14
XGBoost	8.44	8.66	8.21	8.62	8.23	8.31	8.25	8.48	8.40
LightGBM	8.67	8.47	8.39	8.75	8.09	8.2	8.18	8.78	8.44
Random Forest	9.11	9.12	9.13	9.42	8.93	8.92	8.89	9.26	9.10
Extra Trees	9.29	9.13	9.34	9.4	9.29	9.28	9.09	9.33	9.27
Decision Tree	9.86	9.89	9.65	10.01	9.53	9.65	9.5	10.02	9.76
Mean	**8.91**	**8.90**	**8.78**	**9.13**	**8.67**	**8.71**	**8.64**	**9.08**	

Tab.45 presents the training times (in seconds) for various machine learning models across different feature configurations. The models evaluated include gradient boosting, XGBoost, LightGBM, random forest, extra trees, and decision tree. Below is a scientific commentary on the results.

Gradient boosting showed the highest mean training time at 231.63 seconds. The training time varied significantly with different feature sets, from 168 seconds for the basic model to 299 seconds when road toll data were added. This indicates that gradient boosting is computationally intensive and its training time is sensitive to the complexity and number of features.

XGBoost demonstrated a relatively low mean training time of 25.27 seconds. It showed remarkable efficiency, with training times ranging from 14 to 49 seconds across different feature sets. This highlights XGBoost's ability to handle complex datasets efficiently.

LightGBM had the lowest mean training time among all models at 7.98 seconds. Its training times were impressively short, ranging from two to 24 seconds. This efficiency is particularly evident in feature-rich configurations like toll, fuel, and trailers, where the training time remains minimal.

Random forest had a mean training time of 143 seconds, with times ranging from 107 seconds for the basic model to 183 seconds for the combination of fuel and trailers. This indicates moderate computational requirements, though higher than XGBoost and LightGBM. Extra trees showed a mean training time of 85 seconds. Training times varied from 60 to 126 seconds, indicating that while it is more efficient than gradient boosting and random forest, it still requires significant computational resources.

Decision tree was the most efficient in terms of training time, with a mean of only 3.25 seconds. Training times were consistently low across all feature sets, highlighting its simplicity and speed. However, this efficiency may come at the cost of lower predictive performance compared to more complex models.

The basic model training times ranged from two seconds (decision tree) to 168 seconds (gradient boosting), with a mean of 59.83 seconds. This indicates that initial feature sets are relatively easy to handle for most models, except for gradient boosting.

Adding toll data increased the mean training time to 112.33 seconds. Gradient boosting and random forest exhibited the most significant increases, suggesting that toll data adds complexity.

Including fuel data resulted in a mean training time of 80.33 seconds. Gradient boosting and random forest again showed high training times, while LightGBM and XGBoost remained efficient.

Including GDP data resulted in a mean training time of 81.83 seconds. The consistent performance across models indicates that GDP data are moderately complex to process.

The addition of trailer data led to a mean training time of 108 seconds. Gradient boosting and random forest showed significant increases, indicating that this feature set adds substantial complexity.

Combining fuel and trailer data resulted in a mean training time of 92.67 seconds. LightGBM and XGBoost maintained low training times, demonstrating their efficiency with combined features.

Adding toll, fuel, and trailer data slightly reduced the mean training time to 62.47 seconds. This counterintuitive result suggests that certain combinations may streamline the training process for some models.

The most complex feature set, including toll, fuel, trailer, and GDP data, resulted in a mean training time of 64.03 seconds. The slight increase from the previous combination indicates that GDP data does not add significant complexity when combined with other features. The analysis demonstrates that LightGBM and XGBoost consistently offered the shortest training times across various feature sets, making them highly efficient for practical applications. Gradient boosting, while providing robust performance, requires significantly more computational resources, making it less efficient for scenarios with time constraints. Random forest and extra trees offer moderate efficiency, while decision tree excels in training speed but may lack the predictive power of more complex models. These findings highlight the importance of balancing computational efficiency with model performance in selecting machine learning algorithms for road freight transport price forecasting.

Tab.45

Training time (in seconds) for machine learning models

model	basic	toll	fuel	GDP	trailers	fuel, trailers	toll, fuel, trailers	toll, fuel, trailers, GDP	mean
Gradient Boosting	168	299	220	227	288	291	180	180	231.63
XGBoost	17	49	23	24	46	15	14	14.18	25.27
LightGBM	3	21	6	2	24	3	2.8	2	7.98
Random Forest	107	175	135	138	168	183	115	123	143.00
Extra Trees	62	126	95	97	118	60	60	62	85.00
Decision Tree	2	4	3	3	4	4	3	3	3.25
Mean	**59.83**	**112.33**	**80.33**	**81.83**	**108.00**	**92.67**	**62.47**	**64.03**	

Tab. 46 presents the 30 most important features from the full dataset included in the model for forecasting road freight transport prices. The weights assigned to these features reflect their relative impact on model accuracy and help identify the key factors influencing price formation.

The highest weight is assigned to the KM (transport distance) feature, which aligns with expectations. Transport distance is a fundamental factor directly affecting operational costs, such as fuel consumption, driver working hours, and vehicle depreciation. Its significance, with

a weight of 2.023848, far exceeds that of the other features. However, this research focuses on analyzing the impact of the remaining factors, which influence transport prices in more nuanced ways and whose roles are less apparent.

The second most influential feature is RELATION_MEDIAN (median price for a specific transport relation), with a weight of 0.010231. This indicates that historical data on specific transport routes plays an important role in price forecasting. Other significant features include PL_KM (distance in Poland), TOLL (road tolls), and VEHICLE_TYPE (type of vehicle), suggesting the influence of geographical route specifics and cargo characteristics on price formation.

Operational and time-related variables, such as START_LOAD_DATE_YEAR (year of loading) and END_DELIVERY_DATE_DAY (delivery day), were assigned lower weights. While their overall influence on the model is limited, they should not be overlooked, as they contribute to specific aspects of seasonality, demand fluctuations, and local market trends.

Among the less significant variables are FUEL_PRICE (fuel price) and LOAD_COUNTRY_MEAN (average for the country of loading), whose weights are relatively low compared to features like transport distance or transport relation. This likely reflects that the influence of fuel prices and country-specific data is indirect and often correlated with other, more impactful features.

Tab. 46

Feature importance analysis for the Gradient Boosting Regressor

FEATURE	WEIGHT
KM	2.023848
RELATION_MEDIAN	0.010231
PL_KM	0.006093
END_DELIVERY_DATE_DAY	0.004354
RELATION_MEAN	0.002871
TOLL	0.002546
CZ_KM	0.002407
RATIO_TFK2_LOAD_DELIVERY	0.001988
DE_KM	0.001856
VEHICLE_TYPE	0.00183
LOAD_COUNTRY_MEAN	0.001768
START_LOAD_DATE_YEAR	0.001756
FUEL_PRICE	0.001185
GOODS_TYPE	0.001126
END_DELIVERY_DATE_DAY_OF_YEAR	0.001088
START_DELIVERY_DATE_YEAR	0.00099
END_DELIVERY_DATE_YEAR	0.000651
COUNTRY_DELIVERY_MEDIAN	0.000628
END_LOAD_DATE_YEAR	0.000591
LOAD_COUNTRY_MEDIAN	0.000538
END_LOAD_DATE_WEEK_OF_YEAR	0.000532
END_LOAD_DATE_DAY_OF_YEAR	0.000517
START_LOAD_DATE_DAY_OF_YEAR	0.000486

START_DELIVERY_DATE_DAY_OF_YEAR	0.000479
COUNTRY_DELIVERY_MEAN	0.000427
COUNTRY_DELIVERY_STD	0.000325
START_DELIVERY_DATE_MONTH	0.000285
GOODS_TYPE_STD	0.000222
START_LOAD_DATE_WEEK_OF_YEAR	0.000212
QTY_LOADS	0.000205

6. CONCLUSION

6.1. Answers to the Research Question

I. What factors influence the pricing of road freight transport services, how can a methodology for processing transport offer data be developed to effectively train machine learning models, and can the application of these methods achieve greater forecasting accuracy compared to expert-based approaches?

The analysis reveals that the pricing of road freight transport services is predominantly influenced by transport distance (KM), historical data on transport relations (RELATION_MEDIAN), and specific route characteristics, such as distances within particular countries (PL_KM, CZ_KM, DE_KM).

As expected, transport distance (KM) has the most significant impact since it directly correlates with operational costs such as fuel consumption, driver working hours, and vehicle wear and tear. The weight of this variable far exceeds that of other factors, underscoring its dominant role in the forecasting model.

Historical pricing patterns, represented by RELATION_MEDIAN, play a critical role in reflecting market-specific conditions for certain routes. Time-related variables, such as the day or year of loading (END_DELIVERY_DATE_DAY, START_LOAD_DATE_YEAR), while not highly influential, provide valuable insights into seasonality and fluctuations in demand.

Operational characteristics, including vehicle type (VEHICLE_TYPE) and cargo type (GOODS_TYPE), are less significant than macroeconomic factors such as distance and transport relations. Nevertheless, these variables remain relevant for more detailed operational analyses.

The research also successfully developed an effective methodology for processing transport offer data to facilitate the efficient training of machine learning models. This methodology encompasses key steps such as data cleaning, normalization, feature selection, managing missing values, and ensuring high-quality and consistent input data.

A central element of this methodology is handling the diverse data formats commonly found in transport offers. By applying techniques for data standardization and transformation, the methodology enables the seamless integration of disparate data sources into a unified dataset. Furthermore, advanced feature engineering techniques were used to identify key variables and extract significant patterns, which enhanced the accuracy and robustness of the machine learning models.

The study further confirmed that machine learning methods outperform expert-based approaches in forecasting road transport prices. A comparative analysis showed that the average forecast error for expert-based methods was 11.86%, while machine learning models achieved an average error of 9.28%, representing an improvement of 2.58 percentage points.

This improvement highlights the superior capability of machine learning models to capture complex patterns and relationships in data, which are often overlooked or underestimated by expert-based methods. Beyond improving accuracy, machine learning provides scalability and flexibility, enabling models to adapt dynamically to evolving market conditions and incorporate new data in real time.

In addition to reducing forecast errors, machine learning enhances the speed and consistency of predictions, offering transport companies a reliable tool for decision-making.

This supports improved route planning, resource allocation, and pricing strategies, ultimately increasing competitiveness and operational efficiency.

In conclusion, the reduction of the average forecast error from 11.86% to 9.28% demonstrates that machine learning methods deliver greater accuracy and practical applicability compared to expert-based approaches, solidifying their position as a superior solution for the transportation industry.

6.2. Discussion

This section examines the implications of the research findings and evaluates the effectiveness of the implemented machine learning models in forecasting road freight transport costs. A key achievement of this study is that it demonstrates that the model trained using the developed methodology can outperform industry specialists in forecasting accuracy. The model achieved a lower mean absolute percentage error (MAPE) than seasoned professionals. This validates the efficacy of the proposed approach and highlights the value of integrating advanced analytics into decision-making processes in the transport industry. The average error that the experts had was 11.86%, while the model's average was 9.28%, reflecting an improvement of 2.58 percentage points.

It is proposed to ask here whether this is a lot or little. Eurostat data should be considered in answering this question [88]. According to the data, in 2022, road freight transport in the European Union (EU) amounted to 1,866,295,000,000 tonne-kilometres. Assuming that each truck has a payload of 24 tons, we have approximately 77,762,291,667 kilometres. Assuming the average rate of €1.34 per kilometre given in Tab. 27, we are talking about transport worth approximately €104,201,470,834. If we assume that this error translates into the payment for the transport service of the above-mentioned shipments, then applying the method from this doctoral thesis will allow for savings of €2,688,397,948.

This study confirms that incorporating variables like fuel prices and regional economic indicators significantly enhances the predictive performance of forecasting models. Among the algorithms tested, gradient boosting, XGBoost, and LightGBM consistently outperformed other approaches, highlighting their suitability for practical, industry-oriented applications.

Although this research focused on the European Union (EU) market, the methodology is highly adaptable. By using ISO country codes, the models can be reconfigured for deployment in non-EU regions. Moreover, when applying these models globally, it may be necessary to adjust units of distance measurement (e.g., miles instead of kilometres) to maintain accuracy and relevance in different geographical contexts.

The current findings offer several avenues for future research:

- Adaptation to Alternative Fuels: Researchers could investigate how the forecasting models and feature engineering strategies can be extended or modified to accommodate new client requirements for alternative fuels such as electricity or hydrogen. This may involve the inclusion of additional data sources and adjustments to model parameters, as well as rigorous testing to confirm their effectiveness in this evolving landscape.
- Global Expansion: Researchers could investigate model performance and adaptability in non-EU countries, employing ISO country codes and adjusting units of measurement as needed.
- Inclusion of Additional Factors: Researchers could explore the impact of other potential determinants—such as geopolitical events, seasonal trends, and technological innovations—on transportation costs.

- Model Enhancement and Optimization: Researchers could continuously refine the machine learning models by incorporating updated datasets to improve feature engineering techniques.

These insights are valuable for both academic research and practical applications. Transport companies can leverage the proven accuracy advantage of the model to refine their pricing strategies, improve cost management, and enhance operational efficiency. Policymakers and regulatory authorities can also draw on these findings to inform infrastructure development and policy frameworks.

In summary, this study provides a comprehensive analysis of the factors influencing road freight transport pricing and sets a new benchmark in forecasting accuracy. Given that it outperformed human experts, the developed methodology exemplifies the power of machine learning in tackling complex, data-driven challenges, laying a solid foundation for future research and practical implementations in the field of transportation cost forecasting.

6.3. Final Conclusions

This section presents the most important final conclusions and argumentation confirming the thesis of the dissertation.

I. **A new method for forecasting the price of road freight transport services using machine learning has been developed.**
   The effectiveness of the model trained using the proposed method has been demonstrated. An experiment comparing the forecast errors of the model (MAPE = 9.28%) against those of human experts (MAPE = 11.86%) indicates that implementing machine learning techniques can improve the accuracy of price forecasts. This improvement, in turn, enables transport companies to better adjust their operational decisions in response to changing market conditions. **Therefore, the hypothesis of this dissertation has been confirmed.**

II. An experiment was conducted to compare different machine learning algorithms. Six models achieved an average percentage error (MAPE) below 10%: Gradient boosting (8.11%), XGBoost (8.44%), light GBM (8.67%), random forest (9.11%), extra trees (9.29%), and decision tree (9.86%).

III. A solution was proposed to implement external databases into the model. Experiments were carried out with four databases: fuel, road tolls, the number of registered truck trailers in each of the countries of loading and unloading, and the gross domestic product (GDP) of each of the countries of loading and unloading.
   a. Improvement in the model's performance was observed when three databases (fuel, trailers, and road tolls) were used simultaneously. Under this configuration, the average percentage error decreased from 8.91% to 8.64%.
   b. For the best gradient boosting algorithm, an improvement from 8.11% to 7.90% was achieved.

Conclusions II and III show that the proposed method is an original solution that applies the research findings to economic contexts.

IV. Methods for implementing the method presented in solving problems of various types of enterprises and the educational process were proposed.

Bibliography

[1] M. Rajsman, G. Luburić, and M. Muhin, "Dynamics and trends of the development of transport relations in road freight transport," *Teh. Vjesn. - Tech. Gaz.*, vol. 24, no. 2, Apr. 2017, doi: 10.17559/TV-20151222091742.

[2] J. R. Daduna, "Aspects of Information Management in Road Freight Transport," in *Computational Logistics*, vol. 6971, J. W. Böse, H. Hu, C. Jahn, X. Shi, R. Stahlbock, and S. Voß, Eds., in Lecture Notes in Computer Science, vol. 6971. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 29–43. doi: 10.1007/978-3-642-24264-9_3.

[3] I. Meidutė-Kavaliauskienė, D. Stanujkic, A. V. Vasiliauskas, and V. Vasilienė-Vasiliauskienė, "Significance of Criteria and Resulting Significance of Factors Affecting Quality of Services Provided by Lithuanian Road Freight Carriers," *Procedia Eng.*, vol. 187, pp. 513–519, 2017, doi: 10.1016/j.proeng.2017.04.408.

[4] S. Kummer, M. Dieplinger, and E. Fürst, "Flagging out in road freight transport: a strategy to reduce corporate costs in a competitive environment," *J. Transp. Geogr.*, vol. 36, pp. 141–150, Apr. 2014, doi: 10.1016/j.jtrangeo.2014.03.006.

[5] R. Shoukat, "Economic Impact, Design, and Significance of Intermodal Freight Distribution in Pakistan," *Eur. Transp. Eur.*, no. 88, pp. 1–14, Sep. 2022, doi: 10.48295/ET.2022.88.6.

[6] Đ. Stojanović, "Road freight transport outsourcing trend in Europe – what do we really know about it?," *Transp. Res. Procedia*, vol. 25, pp. 772–793, 2017, doi: 10.1016/j.trpro.2017.05.457.

[7] A. Sładkowski and A. Budzyński, "[In rus. 'Transport exchanges, as one of the promising solutions for the problems of transport logistics']," presented at the Transport bridge Europe-Asia. IV Polish-Georgian scientific and technical conference, Tbilisi, Georgia, 10.10 2018.

[8] M. Poliak and J. Tomicová, "Transport document in road freight transport - paper versus electronic consignment note CMR," *Arch. Automot. Eng. – Arch. Motoryz.*, vol. 90, no. 4, pp. 45–58, Jan. 2021, doi: 10.14669/AM.VOL90.ART4.

[9] D. Mitchell, "Trends in non-urban road freight using weigh-in-motion (WIM) data," presented at the Australasian Transport Research Forum 2010, Canberra, Australia.

[10] K. Noijaroen, N. Bua-In, and T. Thawornsujaritkul, "Approaches to Develop Service Quality of Road Freight Transport Service Providers in Industrial 4.0," *Pak. J. Life Soc. Sci. PJLSS*, vol. 22, no. 2, 2024, doi: 10.57239/PJLSS-2024-22.2.00217.

[11] A. Budzyński, P. Malinowski, A. Zaworska, and P. Błaszczyk, "[In polish: Reduction of empty runs and optimization of cargo space utilization using transport exchanges - case study]," *Logistyka*, pp. 51–52, 2019.

[12] A. Budzyński, "Road transport price : Correlation of rates for road transport services in domestic transport in Poland," *Int. Verkehrswesen*, vol. 73, no. 3, pp. 65–67.

[13] P. Bhavsar, I. Safro, N. Bouaynaya, R. Polikar, and D. Dera, "Machine Learning in Transportation Data Analytics," in *Data Analytics for Intelligent Transportation Systems*, Elsevier, 2017, pp. 283–307. doi: 10.1016/B978-0-12-809715-1.00012-2.

[14] B. Mo, Q. Wang, X. Guo, M. Winkenbach, and J. Zhao, "Predicting drivers' route trajectories in last-mile delivery using a pair-wise attention-based pointer neural network," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 175, p. 103168, Jul. 2023, doi: 10.1016/j.tre.2023.103168.

[15] T. Arciszewski, S. Khasnabis, S. K. Hoda, and W. Ziarko, "Machine learning in transportation engineering: a feasibility study," *Appl. Artif. Intell.*, vol. 8, no. 1, pp. 109–124, Jan. 1994, doi: 10.1080/08839519408945434.

[16] Y. Zheng, S. Wang, and J. Zhao, “Equality of opportunity in travel behavior prediction with deep neural networks and discrete choice models,” *Transp. Res. Part C Emerg. Technol.*, vol. 132, p. 103410, Nov. 2021, doi: 10.1016/j.trc.2021.103410.
[17] S. Wang, B. Mo, and J. Zhao, “Theory-based residual neural networks: A synergy of discrete choice models and deep neural networks,” *Transp. Res. Part B Methodol.*, vol. 146, pp. 333–358, Apr. 2021, doi: 10.1016/j.trb.2021.03.002.
[18] G.-L. Huang, A. Zaslavsky, S. W. Loke, A. Abkenar, A. Medvedev, and A. Hassani, “Context-Aware Machine Learning for Intelligent Transportation Systems: A Survey,” *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 17–36, Jan. 2023, doi: 10.1109/TITS.2022.3216462.
[19] A. Budzyński and A. Sładkowski, “Machine Learning in Road Freight Transport Management,” in *Using Artificial Intelligence to Solve Transportation Problems*, vol. 563, A. Sładkowski, Ed., in Studies in Systems, Decision and Control, vol. 563. , Cham: Springer Nature Switzerland, 2024, pp. 485–565. doi: 10.1007/978-3-031-69487-5_9.
[20] G. Chen and J. W. Zhang, “Intelligent transportation systems: Machine learning approaches for urban mobility in smart cities,” *Sustain. Cities Soc.*, vol. 107, p. 105369, Jul. 2024, doi: 10.1016/j.scs.2024.105369.
[21] N. Sholevar, A. Golroo, and S. R. Esfahani, “Machine learning techniques for pavement condition evaluation,” *Autom. Constr.*, vol. 136, p. 104190, Apr. 2022, doi: 10.1016/j.autcon.2022.104190.
[22] Inayatulloh, I. K. Hartono, and P. Cahya S, “Conceptual Model of Citizen Science with Machine Learning to Increase The Effectiveness of Land Transportation of Urban Communities,” in *2023 International Conference on Informatics, Multimedia, Cyber and Informations System (ICIMCIS)*, Jakarta Selatan, Indonesia: IEEE, Nov. 2023, pp. 458–461. doi: 10.1109/ICIMCIS60089.2023.10348969.
[23] E. Akyuz, K. Cicek, and M. Celik, “A Comparative Research of Machine Learning Impact to Future of Maritime Transportation,” *Procedia Comput. Sci.*, vol. 158, pp. 275–280, 2019, doi: 10.1016/j.procs.2019.09.052.
[24] R. Santhiya and C. GeethaPriya, “Machine Learning Techniques for Intelligent Transportation Systems-An overview,” in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India: IEEE, Jul. 2021, pp. 1–7. doi: 10.1109/ICCCNT51525.2021.9579970.
[25] H. Zhong, L. An, X. Huang, and W. Cai, “A Three-Layer Model for Freight Price Index on Highways,” in *Logistics*, Chengdu, China: American Society of Civil Engineers, Jan. 2009, pp. 655–661. doi: 10.1061/40996(330)94.
[26] M. Zhang and X. Du, “An Empirical Study of the Road Freight Prices in China,” in *Proceedings of the Asia-Pacific Conference on Intelligent Medical 2018 & International Conference on Transportation and Traffic Engineering 2018*, Beijing China: ACM, Dec. 2018, pp. 155–160. doi: 10.1145/3321619.3321640.
[27] M. Haigh, “The effect of barge and ocean freight price volatility in international grain markets,” *Agric. Econ.*, vol. 25, no. 1, pp. 41–58, Jun. 2001, doi: 10.1016/S0169-5150(00)00103-1.
[28] X. Li and J. Li, “A freight transport price optimization model with multi bounded-rational customers,” *Transportation*, vol. 48, no. 1, pp. 477–504, Feb. 2021, doi: 10.1007/s11116-019-10064-0.
[29] F. Combes, J. Harache, M. Koning, and E. Morau, “Empirical Analysis of Freight Transport Prices Using the French Shipper Survey ECHO,” in *Commercial Transport*, U. Clausen, H. Friedrich, C. Thaller, and C. Geiger, Eds., in Lecture Notes in Logistics. , Cham: Springer International Publishing, 2016, pp. 321–335. doi: 10.1007/978-3-319-21266-1_21.

[30] J. Zeng, S. Guo, J. Zeng, and K. Jin, “Deep Learning Based Method for Early Warning of Price Risk in Railroad Bulk Freight Transport with Consideration of Time Series of Comprehensive Impact Index,” *Transp. Res. Rec. J. Transp. Res. Board*, vol. 2678, no. 7, pp. 434–449, Jul. 2024, doi: 10.1177/03611981231207842.
[31] A. Subero, *Programming Microcontrollers with Python: Experience the Power of Embedded Python*. Berkeley, CA: Apress, 2021. doi: 10.1007/978-1-4842-7058-5.
[32] A. Javed, M. Zaman, M. M. Uddin, and T. Nusrat, “An Analysis on Python Programming Language Demand and Its Recent Trend in Bangladesh,” in *Proceedings of the 2019 8th International Conference on Computing and Pattern Recognition*, Beijing China: ACM, Oct. 2019, pp. 458–465. doi: 10.1145/3373509.3373540.
[33] S. Yang, T. Kanda, D. Pizzolotto, D. M. German, and Y. Higo, “PyVerDetector: A Chrome Extension Detecting the Python Version of Stack Overflow Code Snippets,” in *2023 IEEE/ACM 31st International Conference on Program Comprehension (ICPC)*, Melbourne, Australia: IEEE, May 2023, pp. 25–29. doi: 10.1109/ICPC58990.2023.00013.
[34] G. Guta, *Pragmatic Python Programming: Learning Python the Smart Way*. Berkeley, CA: Apress, 2022. doi: 10.1007/978-1-4842-8152-9.
[35] A. Kumar and Supriya. P. Panda, “A Survey: How Python Pitches in IT-World,” in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, Faridabad, India: IEEE, Feb. 2019, pp. 248–251. doi: 10.1109/COMITCon.2019.8862251.
[36] “Tag Trends.” Accessed: Nov. 11, 2024. Available: https://trends.stackoverflow.co/
[37] T. Kluyver, B. Ragan-Kelley, F. Perez, B. Granger, M. Bussonier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdallah, C. Willing and P. Jupyter, “Jupyter Notebooks - a publishing format for reproducible computational workflows,” in *International Conference on Electronic Publishing*, 2016.
[38] W. McKinney, “Data Structures for Statistical Computing in Python,” presented at the Python in Science Conference, Austin, Texas, 2010, pp. 56–61. doi: 10.25080/Majora-92bf1922-00a.
[39] S. van der Walt, S. C. Colbert, and G. Varoquaux, “The NumPy Array: A Structure for Efficient Numerical Computation,” *Comput. Sci. Eng.*, vol. 13, no. 2, pp. 22–30, 2011, doi: 10.1109/MCSE.2011.37.
[40] J. D. Hunter, “Matplotlib: A 2D Graphics Environment,” *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, 2007, doi: 10.1109/MCSE.2007.55.
[41] M. Waskom, “Seaborn: statistical data visualization,” *J. Open Source Softw.*, vol. 6, no. 60, p. 3021, Apr. 2021, doi: 10.21105/joss.03021.
[42] F. Pedregosa*,* G. Varoquaux, A. Gramfort, V. Michel and B. Thirion “Scikit-learn: Machine Learning in Python,” 2012, doi: 10.48550/ARXIV.1201.0490.
[43] T.-T. Wong, “Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation,” *Pattern Recognit.*, vol. 48, no. 9, pp. 2839–2846, Sep. 2015, doi: 10.1016/j.patcog.2015.03.009.
[44] Y. Kokkinos and K. G. Margaritis, “Managing the computational cost of model selection and cross-validation in extreme learning machines via Cholesky, SVD, QR and eigen decompositions,” *Neurocomputing*, vol. 295, pp. 29–45, Jun. 2018, doi: 10.1016/j.neucom.2018.01.005.
[45] J. Wainer and G. Cawley, “Nested cross-validation when selecting classifiers is overzealous for most practical applications,” *Expert Syst. Appl.*, vol. 182, p. 115222, Nov. 2021, doi: 10.1016/j.eswa.2021.115222.
[46] T. Leinonen, D. Wong, A. Vasankari, A. Wahab, R. Nadarajah, M. Kaisti, A. Airola,, “Empirical investigation of multi-source cross-validation in clinical ECG classification,”

Comput. Biol. Med., vol. 183, p. 109271, Dec. 2024, doi: 10.1016/j.compbiomed.2024.109271.
[47] Y. Zhang, "Bandwidth Selection for Nadaraya-Watson Kernel Estimator Using Cross-Validation Based on Different Penalty Functions," in *Machine Learning and Cybernetics*, vol. 481, X. Wang, W. Pedrycz, P. Chan, and Q. He, Eds., in Communications in Computer and Information Science, vol. 481. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 88–96. doi: 10.1007/978-3-662-45652-1_10.
[48] Z. Y. Algamal, "Shrinkage parameter selection via modified cross-validation approach for ridge regression model," *Commun. Stat. - Simul. Comput.*, vol. 49, no. 7, pp. 1922–1930, Jul. 2020, doi: 10.1080/03610918.2018.1508704.
[49] D.-C. Li, Y.-H. Fang, and Y. M. F. Fang, "The data complexity index to construct an efficient cross-validation method," *Decis. Support Syst.*, vol. 50, no. 1, pp. 93–102, Dec. 2010, doi: 10.1016/j.dss.2010.07.005.
[50] X. Li, B. Yin, W. Tian, and Y. Sun, "Performance of Repeated Cross Validation for Machine Learning Models in Building Energy Analysis," in *Proceedings of the 11th International Symposium on Heating, Ventilation and Air Conditioning (ISHVAC 2019)*, Z. Wang, Y. Zhu, F. Wang, P. Wang, C. Shen, and J. Liu, Eds., in Environmental Science and Engineering. , Singapore: Springer Singapore, 2020, pp. 523–531. doi: 10.1007/978-981-13-9528-4_53.
[51] J. Josse and F. Husson, "Selecting the number of components in principal component analysis using cross-validation approximations," *Comput. Stat. Data Anal.*, vol. 56, no. 6, pp. 1869–1879, Jun. 2012, doi: 10.1016/j.csda.2011.11.012.
[52] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.
[53] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. in Springer Series in Statistics. New York, NY: Springer New York, 2009. doi: 10.1007/978-0-387-84858-7.
[54] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.
[55] J. H. Friedman, "Greedy function approximation: A gradient boosting machine.," *Ann. Stat.*, vol. 29, no. 5, Oct. 2001, doi: 10.1214/aos/1013203451.
[56] D. S. Kumar and N. P. G. Bhavani, "Improving the Accuracy for Predicting Solar Power Using the Novel Gradient Boosting Regressor Algorithm in Comparison With the RANSAC Regressor Algorithm," in *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, Chennai, India: IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/ICCEBS58601.2023.10448971.
[57] A. A and L. V, "Nutrigrow Using Gradient Boosting Regressor," in *2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST)*, Kochi, India: IEEE, Apr. 2024, pp. 1–4. doi: 10.1109/ICTEST60614.2024.10576157.
[58] T. Anzar, "Forecasting of Daily Demand's Order Using Gradient Boosting Regressor," in *Progress in Advanced Computing and Intelligent Engineering*, vol. 1299, C. R. Panigrahi, B. Pati, B. K. Pattanayak, S. Amic, and K.-C. Li, Eds., in Advances in Intelligent Systems and Computing, vol. 1299. , Singapore: Springer Singapore, 2021, pp. 177–186. doi: 10.1007/978-981-33-4299-6_15.
[59] X. Li, W. Li, and Y. Xu, "Human Age Prediction Based on DNA Methylation Using a Gradient Boosting Regressor," *Genes*, vol. 9, no. 9, p. 424, Aug. 2018, doi: 10.3390/genes9090424.
[60] J. D. López-Barrios, I. K. De Anda-García, R. Jimenez-Cruz, L. A. Trejo, G. Ochoa-Ruiz, and M. Gonzalez-Mendoza, "Predicting Water Levels Using Gradient Boosting Regressor

and LSTM Models: A Case Study of Lago de Chapala Dam," in *Advances in Computational Intelligence*, vol. 15246, L. Martínez-Villaseñor and G. Ochoa-Ruiz, Eds., in Lecture Notes in Computer Science, vol. 15246. , Cham: Springer Nature Switzerland, 2025, pp. 101–120. doi: 10.1007/978-3-031-75540-8_8.

[61] A. Keprate and R. M. C. Ratnayake, "Using gradient boosting regressor to predict stress intensity factor of a crack propagating in small bore piping," in *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, Singapore: IEEE, Dec. 2017, pp. 1331–1336. doi: 10.1109/IEEM.2017.8290109.

[62] K. Li *et al.*, "A new method of Ionic Fragment Contribution-Gradient Boosting Regressor for predicting the infinite dilution activity coefficient of dichloromethane in ionic liquids," *Fluid Phase Equilibria*, vol. 564, p. 113622, Jan. 2023, doi: 10.1016/j.fluid.2022.113622.

[63] N. Bagalkot, A. Keprate, and R. Orderløkken, "Combining Computational Fluid Dynamics and Gradient Boosting Regressor for Predicting Force Distribution on Horizontal Axis Wind Turbine," *Vibration*, vol. 4, no. 1, pp. 248–262, Mar. 2021, doi: 10.3390/vibration4010017.

[64] N. S. S. V. S. Rao, S. J. J. Thangaraj, and V. S. Kumari, "Flight Ticket Prediction Using Gradient Boosting Regressor Compared With Linear Regression," in *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, Chennai, India: IEEE, Apr. 2023, pp. 1–6. doi: 10.1109/ICONSTEM56934.2023.10142428.

[65] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[66] S. K. Patel, J. Surve, V. Katkar, J. Parmar, F. A. Al-Zahrani, K. Ahmed, F. M. Bui, "Encoding and Tuning of THz Metasurface-Based Refractive Index Sensor With Behavior Prediction Using XGBoost Regressor," *IEEE Access*, vol. 10, pp. 24797–24814, 2022, doi: 10.1109/ACCESS.2022.3154386.

[67] M. A. K. Jailani N and G. C. Mara, "Feature Selection in Ozone Feature Space Impacts Performance in Gradient Boosting, Random Forest, Xgboost and Adaptive Boosting Regressors," in *2024 International Conference on Current Trends in Advanced Computing (ICCTAC)*, Bengaluru, India: IEEE, May 2024, pp. 1–6. doi: 10.1109/ICCTAC61556.2024.10581262.

[68] S. Jeganathan, A. R. Lakshminarayanan, N. Ramachandran, and G. B. Tunze, "Predicting Academic Performance of Immigrant Students Using XGBoost Regressor:," *Int. J. Inf. Technol. Web Eng.*, vol. 17, no. 1, pp. 1–19, Jun. 2022, doi: 10.4018/IJITWE.304052.

[69] M. Apidianaki, Association for Computational Linguistics, and Association for Computational Linguistics, Eds., *EiTAKA at SemEval-2018 Task 1: An Ensemble of N-Channels ConvNet and XGboost Regressors for Emotion Analysis of Tweets*. Stroudsburg, PA: Association for Computational Linguistics (ACL), 2018.

[70] A. Raudys and E. Goldstein, "Forecasting Detrended Volatility Risk and Financial Price Series Using LSTM Neural Networks and XGBoost Regressor," *J. Risk Financ. Manag.*, vol. 15, no. 12, p. 602, Dec. 2022, doi: 10.3390/jrfm15120602.

[71] E. Abdelfattah and K. Bowlyn, "Application of Machine Learning Models on Individual Household Electric Power Consumption," in *2023 IEEE World AI IoT Congress (AIIoT)*, Seattle, WA, USA: IEEE, Jun. 2023, pp. 0143–0146. doi: 10.1109/AIIoT58121.2023.10174456.

[72] A. Kumar, "Airline Price Prediction Using XGBoost Hyper-parameter Tuning," in *Advanced Network Technologies and Intelligent Computing*, vol. 1798, I. Woungang, S. K. Dhurandher, K. K. Pattanaik, A. Verma, and P. Verma, Eds., in Communications in

Computer and Information Science, vol. 1798. , Cham: Springer Nature Switzerland, 2023, pp. 239–248. doi: 10.1007/978-3-031-28183-9_17.
[73] K. Jashwanth, K. L. Sai Praneeth Reddy, M. Sai Snehitha, N. Sampath, and P. C. Nair, "Analyzing Urban Transportation Services using RideShare Data Insights," in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, Pune, India: IEEE, Apr. 2024, pp. 1–7. doi: 10.1109/I2CT61223.2024.10543505.
[74] E. Alnayer, K. Elsheikh, A. Alawad, O. Husain, I. Basheir, and A. Siddig, "Enhancement of Sorghum Forecasting Models Using Machine Learning in the rain-fed sector in Sudan," in *2023 International Conference on Electrical, Communication and Computer Engineering (ICECCE)*, Dubai, United Arab Emirates: IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/ICECCE61019.2023.10442132.
[75] K. J. Medows, L. R, and M. M. Thiruthuvanathan, "Predicting Song Popularity Using Data Analysis," in *2024 IEEE International Conference on Contemporary Computing and Communications (InC4)*, Bangalore, India: IEEE, Mar. 2024, pp. 1–6. doi: 10.1109/InC460750.2024.10649267.
[76] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*.
[77] E. Nziyumva, R. Hu, C.-Y. Hsu, and J. Niyogisubizo, "Electrical Load Forecasting Using Hybrid of Extreme Gradient Boosting and Light Gradient Boosting Machine," in *The International Conference on Image, Vision and Intelligent Systems (ICIVIS 2021)*, vol. 813, J. Yao, Y. Xiao, P. You, and G. Sun, Eds., in Lecture Notes in Electrical Engineering, vol. 813. , Singapore: Springer Nature Singapore, 2022, pp. 1083–1093. doi: 10.1007/978-981-16-6963-7_95.
[78] M. Suchithra, K. Shashwat, and M. S. Khan, "Predicting Hospital Length of Stay Using Light Gradient Boosting Machine Regression," in *Computational Intelligence in Data Science*, vol. 718, M. L. Owoc, F. E. Varghese Sicily, K. Rajaram, and P. Balasundaram, Eds., in IFIP Advances in Information and Communication Technology, vol. 718. , Cham: Springer Nature Switzerland, 2024, pp. 487–498. doi: 10.1007/978-3-031-69986-3_37.
[79] R. N. Bashir, O. Mzoughi, M. A. Shahid, N. Alturki, and O. Saidani, "Principal Component Analysis (PCA) and feature importance-based dimension reduction for Reference Evapotranspiration (ET 0) predictions of Taif, Saudi Arabia," *Comput. Electron. Agric.*, vol. 222, p. 109036, Jul. 2024, doi: 10.1016/j.compag.2024.109036.
[80] E. Giovannardi, A. Brusa, B. Petrone, N. Cavina, E. Corti, and M. Barichello, "An Enhanced Light Gradient Boosting Regressor for Virtual Sensing of CO, HC and NOx," in *2023 IEEE International Workshop on Metrology for Automotive (MetroAutomotive)*, Modena, Italy: IEEE, Jun. 2023, pp. 1–6. doi: 10.1109/MetroAutomotive57488.2023.10219122.
[81] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, 3. print. in Adaptive computation and machine learning. Cambridge, Mass.: MIT Press, 2008.
[82] C. M. Bishop, "Neural Networks for Pattern Recognition".
[83] G. D. Hutcheson, "Data coding, management and manipulation," *J. Model. Manag.*, vol. 6, no. 1, 2011, doi: https://doi.org/10.1108/jm2.2011.29706aab.001.
[84] B. Latte, S. Henning, and M. Wojcieszak, "Clean Code: On the Use of Practices and Tools to Produce Maintainable Code for Long-Living Software."
[85] "Online Browsing Platform." Available: https://www.iso.org/obp/ui
[86] "Archive of wholesale Orlen fuel prices." Available: https://www.orlen.pl/pl/dla-biznesu/hurtowe-ceny-paliw#paliwa-archive

[87] “Semi-trailers and their load capacity, by permissible maximum gross weight.” Available: https://ec.europa.eu/eurostat/databrowser/view/road_eqs_semit/default/table?lang=en&category=road.road_eqs
[88] Eurostat, “Territorialised road freight transport, by transport coverage – annual data.” Mar. 04, 2024. Accessed: Dec. 22, 2024. [Online]. Available: https://doi.org/10.2908/ROAD_TERT_GO