Modelling & Analysis of Predictive Pricing in the world of RFQs and Big Data

Janssen, Nick

Award date:
2022

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Master Thesis Project

Department of Industrial Engineering & Innovation Sciences: *Operations Planning, Accounting & Control group*

Modelling & Analysis of Predictive Pricing in the world of RFQs and Big Data

Ewals Cargo Care

Supervisors	Name
1st	*Mutlu, Nevin*
2nd	*Jaarsveld, W.L.V.*
3rd	*Martagan, Tugce*
Company	*Heesen, Freek*

In partial fulfillment of the requirements of the degree of Master of Science in **Operations Management & Logistics**

Author
Janssen. Nick

Student ID
0913644

Eindhoven, June 6, 2022

Eindhoven University of Technology (TU/e), School of Industrial Engineering
Series Master Theses Operations Management and Logistics

Keywords: request for quotation, machine learning, supervised learning, prediction, regression, support vector regression, artificial neural network, multi layer perceptron, bayesian optimization.

Confidentiality

Due to confidentiality reasons, some information in this report cannot be presented in this published version of the master thesis report. Therefore, some information in the text, figures or tables is replaced with letters from the alphabet (Customer X). If such a replacement is performed, this will be mentioned in the text, figure- or table caption.

Preface

There it is, the project that marks the end of my student life. I am grateful for the time I spent at the Eindhoven University of Technology, who managed to constantly challenge me. Throughout my bachelor in Industrial Engineering & Innovation Sciences, I always wondered where I would end up and was constantly evaluating my interests. During my master's studies in Operations, Management & Logistics, I finally got a grasp of the things that fascinate me. Therefore, I cannot wait to take the next step in my professional career.

There are a number of people I would like to thank. First of all, I would like to thank my first supervisor Nevin Mutlu, who helped to reflect on my work critically and helped me to maintain a helicopter view on the project. I also would like to thank my second supervisor, Willem van Jaarsveld for being my second supervisor and for his valuable insights, particularly in the final- and most critical part of the project. From Ewals Cargo Care, I would like to thank Freek Heesen for being my company supervisor and his consistent involvement in the project. Because of you, I quickly found my place within the company and I felt welcome from the first moment I arrived. Also, your contribution to my professional development in such a short time is valued deeply.

I am proud to have overcome the challenges during my education, in particular during my master thesis project. Additionally, the circumstances with respect to the Covid-19 virus were especially demanding, which forced me to take all of my master's courses online. I persevered, because of the unconditional support of my friends, family and girlfriend.

Nick Janssen
Tegelen, June 2022

Abstract

Machine-learning is an area which attracted a lot of attention and was subject to rapid progress in recent years (Jordan & Mitchell, 2015). The concept of machine-learning is explained by building computers that are capable of improving automatically through experience. The increasingly growing amount of data available online and the developments in computational technologies increases both the relevance and applications of machine-learning. In this research, machine-learning is embraced in the price-setting for transportation services from the perspective of a logistics service provider. Multiple machine-learning models (support vector regression, artificial neural network) are developed in this research. It is shown that a support vector regression model can reach a Mean Absolute Error-score of 0.032 (€/kilometer).

Executive summary

Business problem

For a Logistic Service Provider (LSP), the price-setting of a Request For Quotation (RFQ) is of utmost importance. Although it is not the only factor in determining whether a LSP is awarded with business or not, it is the factor weighing the heaviest because of the cost-minimizing nature of the companies in need for transportation. The price-calculation is currently being realized with the help of a cost-based calculation tool. Due to the increased awareness of the importance of data and the processing of this data in the manual, the tool is becoming larger and larger, increasing the computational burden of the manual. Furthermore, the accuracy of such a tool is being questioned because of the substantial investments of time and resources that are paired with the calculation, while overall having limited amount of returns in the form of business being awarded to the company. Therefore, the need to investigate a new model arose. This new model should account for the computational burden and should also include more intelligence to the pricing process. From the available literature, it became clear that there is a gap to be filled when it comes to price-setting in RFQ's. At the moment, the available literature with respect to price-setting of RFQ's is mainly focused on strategies, whereas exploration of optimization techniques or algorithms remain unaddressed. From this discovery, the need to investigate the potential of Machine-Learning (ML) algorithms to predict competitive prices arose. Because of the different nature of a ML model in comparison to the cost-based calculation which is currently reality, the following main research question is answered by this research:

What kind of Machine-Learning algorithm can be developed to set competitive prices in the RFQ process at Ewals Cargo Care?

Data understanding & Analysis

The collection and analysis of the data started by considering the available data sources for this research. The available data sources comprised of RFQ data, client feedback data, market data and actuals (invoice) data. From these sources, it was found that all but the actuals data source were deemed important by this research. This way, internal factors, client factors and market factors could be included in the modelling part of this research. An analysis of the factors in the included data sources lead to the discovery of the following set of variables that were relevant for price-setting of RFQ's:

Selection of included variables in this research

Variables
Lane
Number of rounds
Customer
Strategic End-Market
Lead-time
Modality
Equipment type
Initiative
Current business
Contract rejection rate

Modelling

From the literature review, it appeared that Machine-Learning (ML) models are not yet embraced in the price-setting process of a RFQ. Making use of this opportunity, two ML algorithms, a Support

Vector Regression (SVR) and a Artificial Neural Network (ANN) were developed to bridge the gap in the literature. The models were optimized with the help of two prominent hyper-parameter tuning algorithms, after which the models were validation by means of 5-fold Cross-Validation (CV) and evaluated on a simulation based on previously unseen data. It appeared that the SVR model was most performant.

Results

A deep dive in the results lead to the conclusion that the outcome variable in this research (the price per kilometer) is most sensitive to the country relationship, price of the current business and the contract rejection rate. The effects of the repositioning problem related to the prevention of empty backhaul became apparent in terms of the predicted prices. Countries that focus more on production reflect higher prices for transport leaving the country, while countries that are focused largely on consumption reflect higher prices entering the country. Different Strategic End-Markets (SEM) were investigated thoroughly, and it appeared that prices in the e-commerce and the automotive sector were ranked favorably (reflecting high prices) when considering the top-10 country relationships of the company. The relatively high prices in these sectors were explained by the low lead-times and the fit of the equipment type (truck) with automotive market. The chemicals, building construction and industrial markets showed less favorable price predictions, which was explained by the miss fit of the equipment type in many cases. Moreover, it was discussed that investigating the difference in prices between import- and export related transports leads to interesting insights with respect to network optimization. Therefore, the output of this research can be used as input for a project related to network optimization.

Potential savings and Recommendations

Tackling the problems related to the accuracy of the current pricing process and the invested resources, lead to a model which is able to predict the prices in a RFQ competitively. On the short-term, the prediction intervals can be used to offer higher prices in earlier rounds, while adjusting the price offers closer to the lower bound of the interval based on the feedback of the client. On the long-term, the model could develop into a tool which enables automatic pricing of RFQ's, reducing the considerate investments in resources related to the current pricing calculation process.

The recommendations of this research can be distinguished between company-specific recommendations (1 & 2) and recommendations from a broader perspective (3 & 4). (1) Improving the quality of data by incorporating a uniform data management process. This way, the data is stored in a uniform way and connection of all data sources is simplified, which will improve the modelling capabilities of the company. (2) Incorporate the prediction together with a prediction interval in the current pricing process. With this, the specialists occupied with the price calculations are backed by a data-driven model, which will help decision-making in the pricing process. (3) Extending the investigation and acquisition of relevant factors influencing the prices in a RFQ. With this, more macro-level factors can be included in the research that were not investigated before, maximizing the model accuracy. Some of the factors to be investigated are the difference between spot- and contract rates, the capacity index, demand of trucks, fuel data and driver contracts. (4) Investigate the potential of time series models for time-sensitive information. Because some data that is used by the model depends on the factor time, while RFQ's are typically implemented in the future, it is important to account for the factor time. Time series models can be employed to forecast macro-level factors and therefore (partially) account for the accompanied uncertainty. (5) Explore other Machine-Learning model which may flourish in the world of RFQ's and transportation.

Contents

List of Abbreviations

AI Artificial Intelligence.

ANN Artificial Neural Network.

BO Bayesian Optimization.

BU Business unit.

CV Cross-Validation.

ECC Ewals Cargo Care.

FTL Full Truck Load.

KAM Key Account Manager.

KPI Key Performance Indicator.

LASSO Least Absolute Shrinkage & Selection Operator.

LSP Logistic Service Provider.

MAD Mean Absolute Deviation.

MAE Mean Absolute Error.

ML Machine-Learning.

MLP Multilayer Perceptron.

MR Milk-Run.

MSE Mean Squared Error.

OEM Original Equipment Manufacturer.

OLS Ordinary Least Squares.

PCT Pricing Calculation Tool.

ReLu Rectified Linear Unit.

RFQ Request For Quotation.

RMSE Root Mean Squared Error.

RSS Residual Sum of Squares.

RT Round-Trip.

SEM Strategic End-Market.

SRM Sales Rate Manual.

ST Single-Trip.

SVM Support Vector Machine.

SVR Support Vector Regression.

TOHR Turnover Hit-Rate.

VIF Variance Inflation Factor.

Glossary

backhaul The return movement of a transportation vehicle from the unloading location back over part or all of the route.

Business Unit A Business Unit is an entity owned by Ewals Cargo Care (ECC).

Carrier The party realizing transportation of a pre-determined set of goods.

ceteris paribus Implying that all other things are kept equal.

consignment Equivalent to a client order to move goods between a collection and delivery location.

Full Truck Load (FTL) A transportation is called FTL, when a truck is loaded to its maximum extent. The maximum is expressed in "payload" and differs between equipment types.

Geo-coding Geo-coding is the process of translating addresses to specific coordinates (latitude/-longitude). Accurate distance calculations require coordinates rather than addresses.

Hyper-parameter Hyper-parameters are explicitly defined parameters that control the tuning process. They are used in order to tune (or optimize) the model.

Key A key refers to a column in a dataset. Datasets can be joined/merged on keys: where values in both keys of the dataset match, the data from the other dataset is joined.

Key Account Manager An Account Manager is allocated to an account, where each account may consist of multiple clients. This person is responsible for managing the accounts to which (s)he is allocated. The Account Managers who are responsible for a top-35 account are called Key Account Managers.

Lane A trajectory with a defined start and end point (city, postal code, ...).

LSP A Logistic Service Provider (LSP) is an organization that provides its transportation services to another organization.

Milk-Run In a Milk-Run, multiple orders are consolidated, which results in a sub-optimal route in terms of distance, but increases cost-efficiency of combined orders. An example of a Milk-Run is a transport from A to C to B, whereas one order needs transport from A to B and another order needs transport from C to B.

payload Payload is measured in Units. The maximum payload of a truck depends on the equipment type and legislation.

product With "Product", a combination of zone-zone relationship with equipment type, modality and execution possibility in which it has a competitive advantage is meant.

Request For Quotation A customer request for transportation. Interchangably used with Tender.

right to rin When a business unit has "right to win", the RFQ contains lanes that are part of the business units "products", implying the business unit is eligible to offer.

Round-Trip A Round-Trip is a trip from A to B to A. The difference with a Single-Trip is that a Round-Trip returns to the origin.

Tender A tender is equivalent to a Request For Quotation.

Trip-leg A trip can be subdivided in trip-legs. When an intermodal transport option is involved in a trip, multiple trip-legs are a fact. A trip-leg may be satisfied by a truck, a train or a ferry.

Turnover Hit-Rate The Turnover Hit-Rate is explained by the percentage of the sales rates being offered that are awarded business.

1 Introduction

This master thesis project is the final part of the master's degree in the Operations Management & Logistics study program. The project will be performed under the Operations, Planning, Accounting & Control group. The project is conducted at Ewals Cargo Care, which is a logistics company located in Tegelen, the Netherlands. The introduction of this research will start with a problem definition in which the problem will become tangible. The problem statement will lead to a set of research questions, that will be addressed in this research. Furthermore, the introduction will comprise of a company description.

1.1 Problem definition

The problem at hand is rooted at the offering process of a logistics company. Companies in need of transportation will send a Request For Quotation (RFQ). The concept of RFQ's can be seen as a reverse auction, where the seller is a Logistic Service Provider (LSP), which can bid for the prices at which they are willing to sell their goods and service for (Chen & Kindnes, 2021). The buyer in this reverse auction is the customer in need for the transportation service. The logistics company has a Business unit (BU) in multiple countries, whereas the pricing process can be different between the different business units. Additionally, calculating the prices (for which the LSP sells his service) is currently an outdated process. The Sales Rate Manual (SRM), which is the tool that is employed in order to calculate the sales prices, is created 20 years ago. Although the parameters used for the calculation are updated regularly (daily), the calculation itself is in need for improvement. A lot of data is involved in the calculation process, while the calculation is performed in Excel (Corporation, 2018). The tool is currently limited by the capabilities of Excel, while the accuracy can still be improved: a lot of time is invested in the pricing process, while the Key Performance Indicator (KPI) are not convincing in terms of performance. Regarding large RFQ's, the performance is becoming especially problematic: the RFQ has to be split into parts, such that the SRM can handle the calculation. The reason for the expensive computation of the sales price, is because the calculation is currently cost-based, implying that the total cost calculation is fixed and the calculation which is executed is determined by pre-determined cost aspects. These costs are determined by a set of underlying parameters that are updated regularly. Although the tool is currently transparent (it is possible to analyze how the price is derived), the prices are often not competitive, implying the SRM is not performant enough to satisfy company needs. Next to the computational expense, an increased level of accuracy is desired for the calculation of the sales price. Some future-proof alternative must be found in order to improve the accuracy and capture other developing complexities, while still controlling for computational costs. In addition, more intelligence should be added to the tool, which will improve decision-making and potentially the accuracy of the price calculations. In order to tackle the performance-related issues in the SRM, there is a need to investigate the potential of a new model. While the SRM is currently the backbone of the calculation and is still internally perceived to be a reliable tool to calculate the prices, the performance issues suggest otherwise: somewhere in the process exists a substantial amount of non-added value effort. In essence, this means that most of the offers submitted are too expensive and the competitiveness of the company is jeopardy.

A deeper dive into the problem mess, leads to the discovery of yet another important problem. Although the cost calculation in the SRM is transparent and domain knowledge is applied to it, this information is not logged in a database. This complicates calculation of similar RFQ's with should lead to similar results. Therefore, a foundation has to be build such that improved modelling is possible in the future and a contribution can be made to the accuracy of a future performant model.

1.1.1 Research Questions

Following from the problem definition, this research treats the following research questions:

> **Main research question**: *What kind of Machine-Learning algorithm can be developed to set competitive prices in the RFQ process at Ewals Cargo Care?*
>
> > **Subquestion 1**: *What data could be used for the development of a price-setting model?*
> >
> > **Subquestion 2**: *What is the current quality of the available data and how to enhance data management in order to improve modelling capabilities in the future?*
> >
> > **Subquestion 3**: *What feedback mechanisms should be incorporated into the model in order to establish a benchmark that includes internal and external factors that contribute to competitive quotations?*
> >
> > **Subquestion 4**: *How to measure the performance of the algorithm in order to capture its effectiveness in practise and when is the algorithm performant enough for price-setting?*

1.2 Company description

Ewals Cargo Care (ECC) is a Dutch Logistic Service Provider (LSP) (transportation company) with 22 Business Units located across 15 different countries. The largest BU is called the Network BU, which is responsible for directing the own assets of ECC and is repsonsible for about 60% of the overall turnover residing in the entire Ewals Group. Next to the Network BU, other BUs exist which are mainly distinguishable by means of their local presence and entrepreneurship, taken together covering a pan-European presence. The company is founded by Alfons Ewals in the year 1906 under the name "Ewals Transport" and employs currently about 2,400 employees. In 1994 the companies "Ewals Transport" and "Cargo Care" merged into one company: "Ewals Cargo Care (ECC)". The fleet of ECC consists of 550 trucks (which is expandable to about 1400 trucks from partners) and 3,600 trailers. Currently, ECC is a family-owned organization led by the fourth-generation of the Ewals family. The organizational hierarchy becomes clear in the organizational chart of ECC in Section 11.1. ECC is responsible for directing its own assets. For this sake, it participates in a Request For Quotation (RFQ), in which it sells its service to clients that are in need of transportation. ECC is mainly active in the following Strategic End-Market (SEM): Automotive Original Equipment Manufacturer (OEM), Automotive After Sales, Automotive 1st, 2nd & 3rd tier, Aerospace, Industrial, Consumer Goods, Consumer Electronics, Chemicals, (Lead) Logistics Service Provider (LSP), Paper & Packaging, Fashion, Waste-recycling, Agricultural, Events-projects, Building Construction, E-commerce. At the moment, the automotive SEM is the largest market for ECC: 65% of the total turnover relies on this sector. However, this market is highly volatile. Therefore, ECC is trying to broaden the other markets in order to decrease its reliance on the Automotive sector.

ECC is constantly trying to adjust to the fast changing environment in which we are surrounded, such as: self-driving vehicles, data connectivity & Artificial Intelligence (AI). Rather than accepting a new situation, Ewals is tempted to take the lead: discover new territories in order to become the front runner in the logistics industry. In order to succeed, "Next Generation Logistics" are introduced. One of the topics within next generation logistics is especially relevant to this research: "Business systems to perform", as the research is aimed at using business intelligence to enhance offering capabilities of the company, ultimately reducing time waste that is currently involved with the offering process. The new hierarchical structure within ECC adds to the next generation logistics by integrating Product Intelligence within Product Management.

2 Theoretical Framework

In this section, the existing literature will be reviewed. Directions of research are summarized and the relevance with respect to this research is discussed. The literature review is divided into three main streams. First, the developments with respect to price-setting in a Request For Quotation (RFQ) are discussed. Next, the developments in the area of Machine-Learning are discussed. Lastly, the applications of Machine-Learning in the transportation market are discussed. At the end of the literature review, the implications for this research are concluded.

2.1 Price-setting for RFQ's

According to Gudehus and Kotzab (2009), a Logistic Service Provider (LSP) utilizes a marketing strategy that aims to maximize the total profit by providing optimal sales prices for maximal sales volumes. The source distinguishes strategies on three different time-horizons: long-term, medium-term and short-term. The book suggests that LSP's maintain a value leadership strategy on the long-term dimension, implying that the efforts are focused on the value of the offer for the customer. The focus of Marketing and R&D (research and development) is at providing as much as unique selling points to the customer as possible and to reduce the amount of competitors. Next to these long-term strategies (Porter & Strategy, 1980), LSP's consider customer attraction and demand investigation on the medium-term horizon. Lastly, competition price investigation and expectation price investigation follow on the short-term horizon, implying that before delivering its own price quotation, LSP's attempt finding out the competitors' prices and the price expectation of the customer. With respect to the full price calculation (which is the most important step of pricing), several price differentiation strategies are possible. Some examples are: regional price differentiation; Temporal price segmentation; customer group segmentation; Product & service differentiation; quantity differentiation. The strategy is the result of the unique selling points of LSP's. Related to Ewals Cargo Care (ECC), the focus is on Single-Trip (ST), Full Truck Load (FTL) business. The strategies discussed, seem to comply with the current pricing strategy at ECC. More specifically, the Sales Rate Manual (SRM) is used in order to price a Request For Quotation (RFQ). Investigation of competitor prices and price expectations are taking into account in this process.

From the research by Gudehus and Kotzab (2009), no detailed solution was proposed apart from a general, profit-maximizing strategy. In more detail, a development from the perspective of the Carrier (the LSP) is discussed by Raychaudhuri and Veeramani (2003). Here, an algorithm is proposed in order to determine the best price. According to the source, no literature exists in this direction specifically applying to the transportation market. The research distinguishes between different rounds in the RFQ process. While the algorithm proposes solely to make use of historical data before participating in the first rounds, the next rounds use both historical data and a minimum price decrement. Lastly, the final rounds will follow a more aggressive style of pricing in order to win the business. While this research treats the pricing process with respect to individual Lanes (transport trajectories), Ueasangkomsate, Lohatepanont, et al. (2012), suggests the possibility of pricing multiple different combinations of lanes simultaneously. The concept is introduced by Elmaghraby and Keskinocak (2004) and is called a Combinatorial Auction (CA). The objective of this type of auction is to minimize the empty backhaul- and repositioning costs. In the perspective of ECC, this objective is minimized by means of Round-Trip (RT)'s and Milk-Run (MR)'s where both concepts consolidate transports. In order to do this, a hard decision has to be made in order to decide which lanes should be prices (An, Elmaghraby, & Keskinocak, 2005). This latter research introduces a simulation technique in order to tackle the profit-maximizing stochastic optimization problem. The method used for this sake is a Monte-Carlo simulation in a setting of an incomplete information game. This type of "game" represents a situation where asymmetric information exists between the perspectives of the customer and the LSP. The research succeeded in improving the expected profit over traditional quoting strategies.

The next development discovered in the research of Y. Zhang, Luo, and He (2015) is the importance

of big data in the evaluation of RFQ's. Although the research is focused at RFQ's in the building construction market, the markets show some overlap because they have a similar RFQ process. The research showed that big data had a significant contribution to RFQ price evaluation by proving a reasonable cost range. Another research related to big data describes more clearly how the data could be utilized for customized pricing. Although the research is not specifically tailored towards the logistics market, customized pricing is also applicable in the RFQ process, where pricing for customers is a separate, stand-alone process which is unique for each customer. The road to intelligent decision-making related to customized pricing is conceptualized with the help of Figure 2.1 below.

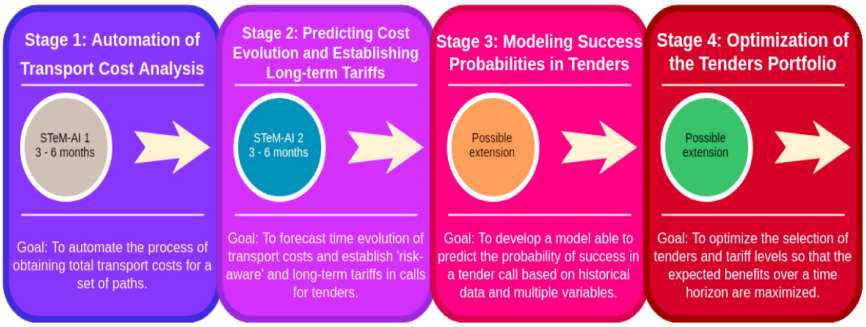

Figure 2.1: The stages towards intelligent decision-making (Nataraj et al., 2020)

First, the cost-based part of the quoting process is automated. Secondly, a prediction (forecast) is added to the cost-based calculation performed in step 1. Then the success probabilities of RFQ's are modelled with the help of statistical models, after which the RFQ portfolio can be optimized such that the expected benefits on the long-term is maximized. According to Nataraj et al. (2020), the methodology is introduced to handle the competitive environment which is largely explained by commerce globalization.

2.2 Machine-Learning algorithms

The trends, perspectives and prospects of Machine-Learning (ML) are discussed in the research by Jordan and Mitchell (2015). According to the source, ML is explained by building computers that are capable of improving automatically through experience. The rapidly growing technical field has its foundation in the areas of computer science and statistics, while already being widely integrated in the fields of Artificial Intelligence (AI) and computer science. The ongoing explosion of data availability online and low-cost computation increases the relevance and applications of ML, leading to more evidence-based decision-making across many walks of life. Nowadays, the implications of ML are already found in a wide variety of markets that have to deal with data-intensive problems. Among them, the control of logistics chains is one of them. The most widely integrated form of ML is supervised learning, where an output y^* is predicted from a set of inputs x. In all kinds of supervised learning problems, progress is visible. A variety of learning algorithms were proposed in recent years, where neural networks and support vector machines are one of the many examples among them. Additionally, different learning algorithms were combined in order to exploit the capabilities of multiple algorithms. The availability of a wide variety of algorithmic architectures and their approaches to trade-off complexity, the amount of data and the performance demonstrate the high need of ML in a diversity of applications.

One of the highest impact areas in which significant progress was realized, is the area of deep networks. This type of Artificial Neural Network (ANN) is called deep because it consists of multiple layers. Neural networks are named after their function which is similar to biological neurons in the human brain and are widely used for predictive modelling. Exploiting modern parallel computing architectures, nowadays it is possible to build deep learning systems based on a very large set of parameters and a large set of (online) data. The data feeding the networks may appear as images, videos and speech samples. In the areas of computer vision (Krizhevsky, Sutskever, & Hinton,

2012) and speech recognition (G. Hinton et al., 2012) the most progress was found. However, more and more applications are being explored and actively pursued with ML. Other than supervised learning, progress was also made in the areas of unsupervised learning and reinforcement learning. With unsupervised learning, learning without labeling the output is meant while for reinforcement learning, only an indication is provided to the training data. Recent progress in machine learning has been driven both by the development of new learning algorithms, theory and by the ongoing explosion in the availability of online data and low-cost computation.

2.3 Applications of Machine-Learning in the transportation market

Nowadays, some innovative ML concepts can already be found in the transportation market. A paper by Woschank, Rauch, and Zsifkovits (2020) discusses the advances and further directions for Artificial Intelligence (AI), Machine-Learning (ML) and deep learning in logistics. In order to do so, the areas of impact can be subdivided in multiple clusters. Clusters may involve strategic and tactical topics to support management decision-making, but may also involve smart logistic opportunities, predictive maintenance or production planning & control systems. The cluster most relevant to this research is related to improvements of operational processes in logistics. Advances (and challenges) in this sector relate for example to swarm robotics for warehouse delivery management, reducing response delays and enhancing quality control (Wen, He, & Zhu, 2018). "*In swarm robotics multiple robots collectively solve problems by forming advantageous structures and behaviors similar to the ones observed in natural systems, such as swarms of bees, birds, or fish*" (Schranz, Umlauft, Sende, & Elmenreich, 2020). A next application is related to object tracking in the field of transportation and attempts to create a new localization system that is able to learn the ability to locate the origin of sounds. The need for such a system arises due to the high complexity of the current methodology, which requires more computational complexity due to its requirements of having to learn spatial characteristics (Laux et al., 2018). Teschemacher and Reinhart (2017) proposes another application of ML by introducing an approach to solve the vehicle routing problem where different materials have to be delivered to different stations in little time such that routes cannot be planned in advance anymore. The solution employs an ant-colony optimization algorithm which is inspired by the behavior of real ants, employing chemicals in order to signal other agents of the solution found.

2.4 Conclusion

Concluding the literature above, some algorithms were found that help at utilizing a pricing strategy in RFQ's. Moreover, a customized pricing approach was found that was not specifically designed for the RFQ process, but has overlap with the process. The stages towards intelligent decision-making were composed, specifying at what stages the concepts of automation, prediction an modelling could be applied. Next, the development of ML algorithms were discussed. It became clear that with the growth of data-intensiveness in a growing amount of markets and companies, more and more awareness is created for ML and the possibilities are being exploited. Progress was found in all areas, where supervised learning is one of them. Also in the transportation market, ML and AI already have found their place being employed in order to tackle a variety of problems related to warehouse delivery management, localization systems and consolidation of transports in Milk Runs.

What is still missing in the literature available is the implementation of ML algorithms for the prediction of prices. With a lot of RFQ's available for quoting on a wide variety of platforms, there is a need to automate the pricing procedure of RFQ's. While there is an abundance of RFQ's ready to quote, the quoting procedure is time-consuming. On top of this, often business is not even won, which means the consumption of time is wasted. By collecting as much data as possible from the market and the previously participated RFQ's, the applicability of ML in the context of RFQ's becomes interesting. Whether or not ML algorithms are effective in automated pricing, a price prediction could always help human resources that are currently occupied with pricing in order to support their decisions data-driven.

3 As is situation

In this section, the current situation at Ewals Cargo Care (ECC) is discussed. The relevant departments within the company are introduced along with the business processes that relate to the price-setting of a Request For Quotation (RFQ).

3.1 RFQ & Solutions Desk

In order to get a better grasp of the problem at hand (Section 1.1), it is important to understand the business process of the Request For Quotation (RFQ) desk well, as the research is conducted under the supervision of this department. A visualization is made in order to capture the most important business processes of the RFQ desk, which is shown in Figure 3.1 below.

The concept of RFQ's can be seen as a reverse auction, where the sellers (a logistics service provider such as ECC) can bid for the prices at which they are willing to sell their goods and service for (Chen & Kindnes, 2021). The buyer in this reverse auction is the customer in need for the transportation service. As the name suggests (Request for Quotation), a customer requests for a quotation (a bid from a LSP), which can be done via multiple ways. Some companies prefer to send RFQ's to transportation companies directly, whereas others make use of a platform for this purpose. An example of a widely integrated platform is Ticontract (*About Ticontract | Transporeon*, 2021), which is a cloud-based logistics solution platform for the procurement of transportation and freight cost management by Transporeon. Some larger companies have their own procurement platform. When a RFQ is collected at the RFQ desk, it is decided whether or not the RFQ will be handled by the company. The desk has a tool (Intelligent Tender Selection tool developed by de Roeck (2022)) available that helps determining the potential of a RFQ. Around 15 to 20% of the RFQ's are processed at the RFQ desk. The purpose of the RFQ desk is to process the RFQ's centrally. From a strategic point of view, it is believed that restructuring the pricing process on a more central level (RFQ level) will improve the business processes of ECC. Before the RFQ & Solutions desk came into existence, all BU's (business units) processed RFQ's by themselves, which lead to a lot of inefficiencies. With the RFQ-desk, ECC has one uniform process in the communication to its external stakeholders. Other benefits of centralization of the RFQ process is the enhancement of productivity, transparency and synergy. For example, by centralization only one central department prepares the RFQ's. A summary is made which can be read by the interested business, which may offer subsequently. By implementing the Product Database, BU's are not able to offer on each RFQ anymore: only the BU's which are likely to offer a competitive solution for the client. The strength of each BU is determined by a predefined set of zone-relationships and are referred to as "products". If a RFQ contains lanes that are part of a BU's product, than it is decided that the concerning BU has "right to rin" and is eligible for offering. An overview of the zone setup can be found in Section 5.3, which will be explained more thoroughly in Section 5.

Currently, every BU is responsible for their own sales price calculation. These BU's respond to the RFQ by either declining or proposing an offer for the first round. In most cases, a RFQ consists of transportation from multiple locations to multiple destinations. One combination of location-destination is called a "lane" in the transportation sector. The calculated sales prices are lane-specific. A BU may only respond to the lanes that match their product portfolio, but it may also offer on multiple lanes in order to offer a package to the customer. Dependent on the requirements of the customer, it might be useful to offer packages to the customer, or just "cherry pick" the interesting lanes. With cherry-picking, it is meant that a BU only selects the lanes of interest and ignores the remaining lanes.

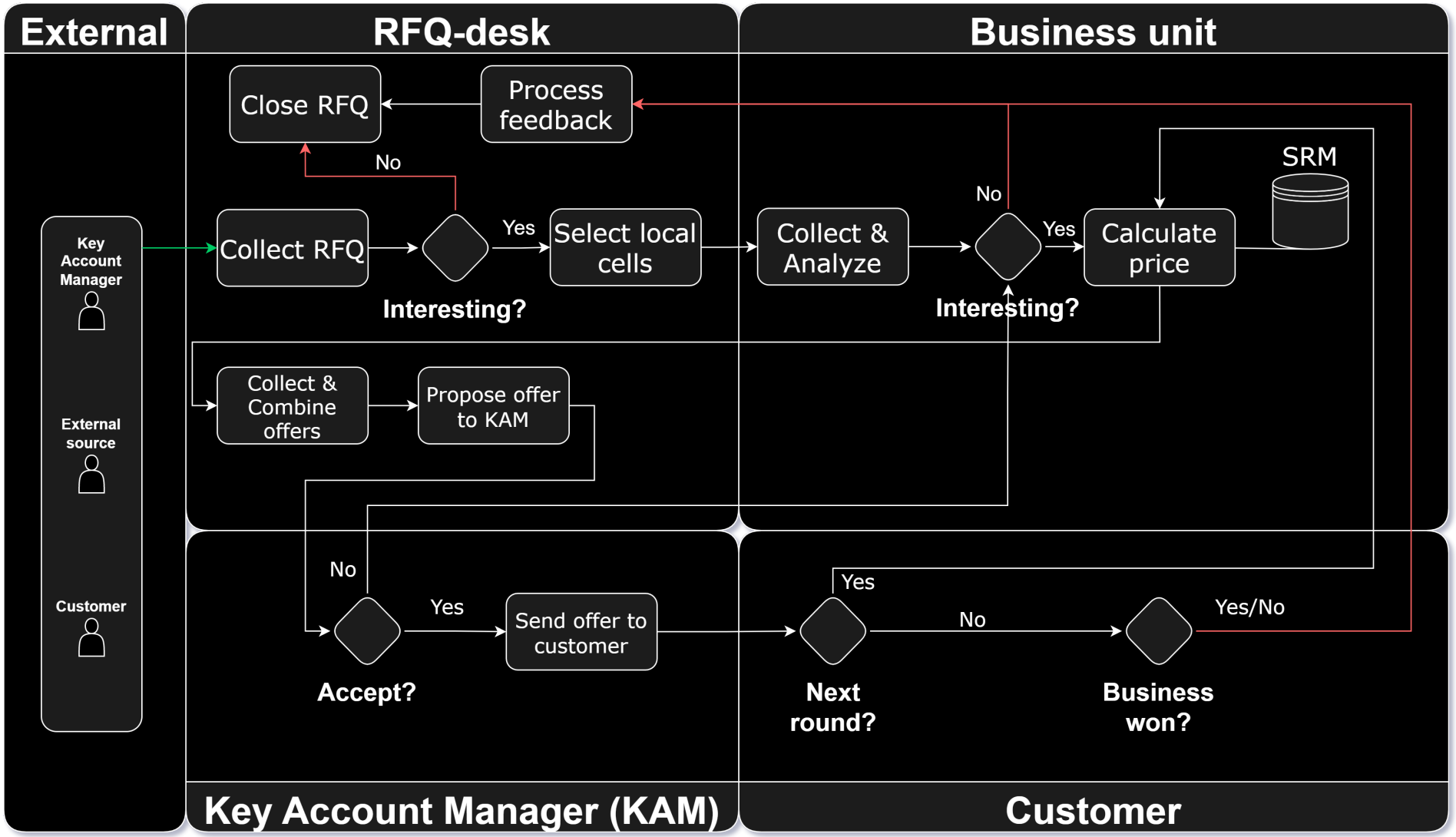

Figure 3.1: Visual representation of the business processes of the RFQ Desk

After selecting and calculating prices for the interesting lanes, the RFQ is sent back to the RFQ desk. Here, all offers are collected from all BU's. In collaboration, negotiation and reconciliation with a Key Account Manager (KAM), the offers are ultimately sent to the client on a corporate behalf. An Account Manager is responsible for the communication with the customer, whereas a Key Account Manager is responsible for the communication with the top-35 accounts. If a RFQ contains multiple rounds, the customer provides feedback on the offers. The level of feedback varies strongly between customers, but in essence it becomes clear if the offers are competitive or not. If the LSP proceeds to the next round, the information along with feedback (if provided) is transferred to the relevant BU's and the pricing process is repeated. If not, a competitor won the RFQ and the end of the RFQ process is reached. The process is repeated until business is either won, not won or lost. Losing business means that ECC has current business with the client, but the business will not be continued. Not winning business implies losing the RFQ to a competitor. The number of rounds of a RFQ varies greatly. Some RFQ's take only one round, whereas others may take up to seven rounds to complete. Most of the time, the number of rounds in a RFQ process is stated before the RFQ starts. The performance of the offering process is measured by the Key Performance Indicator (KPI) Turnover Hit-Rate (TOHR). This KPI measures the percentage of total offers (by turnover) that represents the business won.

Next to the RFQ's, the RFQ-desk is occupied with a lot of project work. The product database and the Intelligent Tender Selection tool (de Roeck, 2022) are just one of the few examples among them. The desk is highly motivated to increase the efficiencies and effectiveness of its processes.

3.2 The pricing process

In the RFQ process that is explained in Section 3.1, the pricing itself does not occur. This is because this procedure does not happen on the level of the RFQ-desk, but is performed by all the BU's themselves. After the RFQ-desk sends invitations to several BU's, the BU's are requested by the RFQ desk to offer a price on the lanes. For the Network BU (by far the largest BU), the Sales Engineering Team is responsible for this price calculation. After the calculation, the team returns the prices to the RFQ desk. Then, the RFQ desk is responsible for the communication to the stakeholders (customers or KAM).

Additionally, calculating the prices is currently an outdated process. The Sales Rate Manual (SRM), which is the tool that is employed in order to calculate the prices, is created 20 years ago. Although the parameters used for the calculation are updated regularly (daily), the calculation itself is in need for improvement. A lot of data is involved in the calculation process, while the calculation is performed in Excel. The tool is currently limited by the capabilities of Excel, while the accuracy can still be improved (a lot of time is invested in the pricing process, while the TOHR is still too low). Regarding large RFQ's (existing of 30+ lanes), this is becoming especially problematic. Currently, when a RFQ exists of more than 30 lanes, the SRM is used multiple times such that at most 30 lanes in the SRM are assessed simultaneously. When more lanes are added, the calculation becomes computationally too expensive. Next to the computational difficulty, an increased level of accuracy is desired. A simple example of this is the use of postal codes. Currently, the SRM uses a 2-digit postal code to calculate the distance between two locations, which are then involved in the calculation of the sales prices. A three-digit postal code is necessary, because the dispersion of a two-digit postal code is too broad. Especially in areas where few people live, the distance calculation will be inaccurate relative to the actual distance. Currently, manual corrections for this are applied to the calculation. Another platform (TLN, 2021) is used to check the actual distance calculation and with this, a lot of time is involved. Moreover, for a dataset that uses 2-digit postal codes, around 60,000 data-points are included. Using a 3-digit postal code means that the dataset will be expanded to around 600,000 datapoints, increasing the file size and concurrently slowing down the calculation. These findings are based on actual experiments with the SRM by the Sales Engineering team (the team responsible for the offers made and perform the calculations with the SRM). An attempt was made to extend the distance dataset, but this resulted in the SRM to freeze and being impossible

to work with. Some future-proof alternative must be found in order to improve the accuracy and capture other developing complexities, while still controlling for computational costs. In addition, more intelligence should be added to the tool, which will improve decision-making and therefore the accuracy of price calculations. Before diving into detail of a new model, it is important to understand the concept of the SRM. For the sake of explaining how the SRM works, a simplified visualization is created which is shown in Figure 3.2 below. It should be noted that the factors influencing the calculation is a hot topic at ECC and the SRM is continuously under construction.

The total cost calculation is fixed, implying that a calculation is executed based on pre-determined cost aspects. These costs are determined by a set of underlying parameters that are updated regularly. After the total cost calculation, the final sales price is calculated by adding a sales margin, surcharges and corrections to the cost calculation. This leads to a sales price that is subsequently being assessed from a practical point of view by NEC Engineers and sometimes Product Managers. From experience and the current company performance, measured by the KPI TOHR, it appears that the current calculation performed by the SRM is not performant enough to satisfy company needs: only for 6% of the sales prices that are offered, business is being awarded. This suggests that somewhere in the process exists a substantial amount of non-added value effort. Therefore, ECC wants to broaden its knowledge with respect to the competitiveness of its offers and eventually bring more competitive prices to the market. On the other hand, it wants to know beforehand if it is not going to be competitive, such that it prevents wasting resources to the offering process.

Moreover, the data relevant to the cost calculations are not being logged to a database, which complicates the spread of knowledge in the company. This means that is currently hard to track down the reason for a price within a RFQ. For example, a market correction is applied in every price calculation, which is a measure of the market situation at a particular moment. The market situation is a relevant cost factor for the price calculation. With a larger emphasis on data management and quality by ECC, the company wants to combine big data and data intelligence in the new model. This means, analyzing the large amount of data at the company and make a definition from it. Separating the useful data from the less useful data creates room for modelling and doing predictions. Based on historical data, client feedback, external market information and RFQ data, the company is inclined to benchmark the price in an improved and future-proof way. The new model should be maintainable in terms of computational cost, easy to understand and written in a structured way, such that knowledge could be transferred easily and the model is controllable without wasted effort.

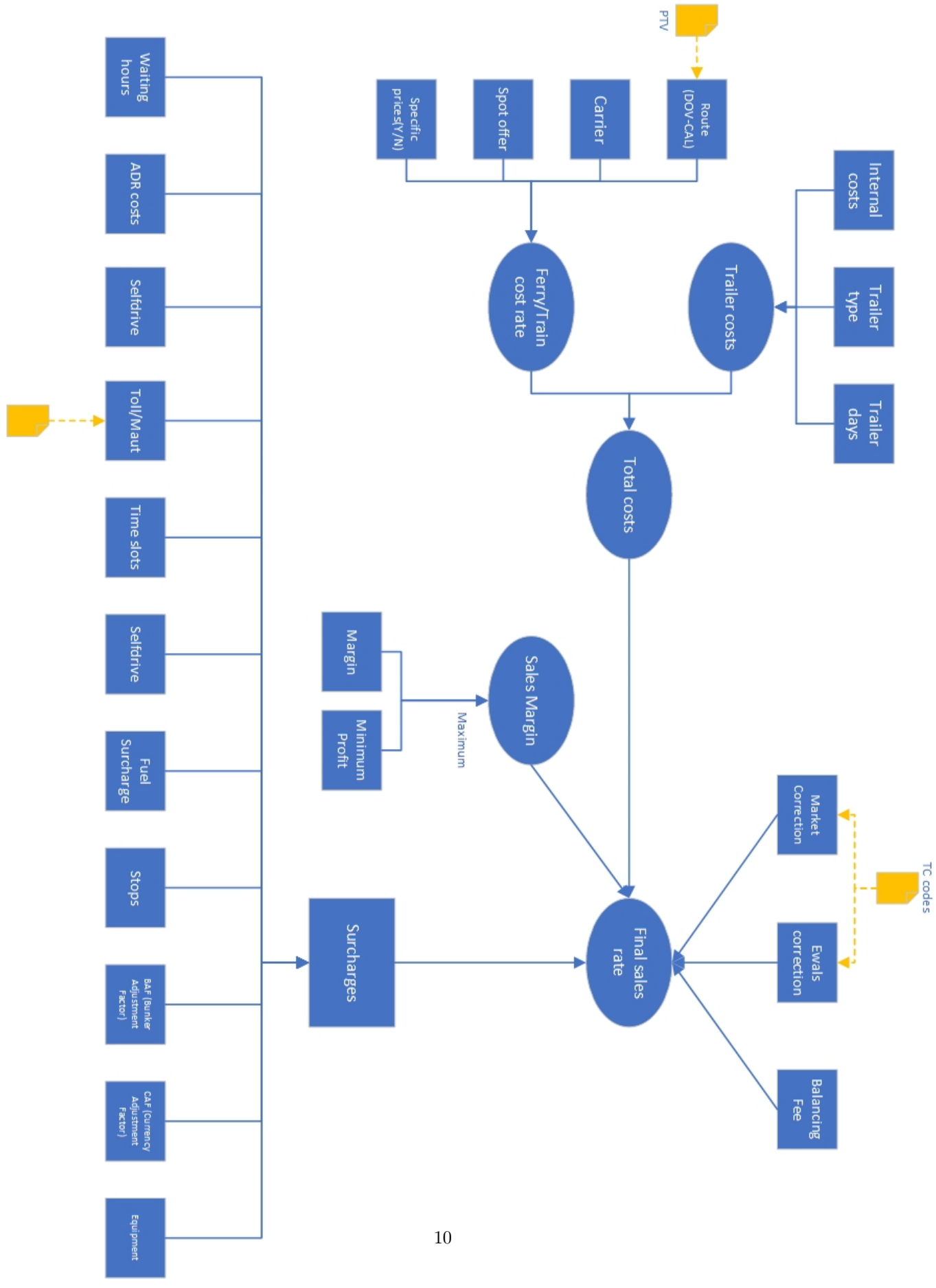

Figure 3.2: Concept of the Sales Rate Manual (SRM)

4 Data Understanding

In this section, a deep analysis of the data is performed. This involves an analysis of the data structure, the available data and the feedback mechanisms that can be used for price-setting. Finally, a conclusion of the available data and definition of the data is found at the end of this section.

Because of the growing company awareness in importance of the collection of data and maintaining the quality of data, the data management within Ewals Cargo Care (ECC) made some significant steps in recent years. As a model is only as good as the data, it becomes especially important in relation with this project, which is aimed at developing a Machine-Learning algorithm and of which the performance is highly dependent on the (quality of the) available data. The data relevant to the price-setting of a Request For Quotation (RFQ) may be originating from four different sources, which are visualized in Figure 4.1 below. The four different sources are referred to as the "pillars" towards predictive pricing. The pillars will be explained in this chapter and the data available in these pillars will be discussed.

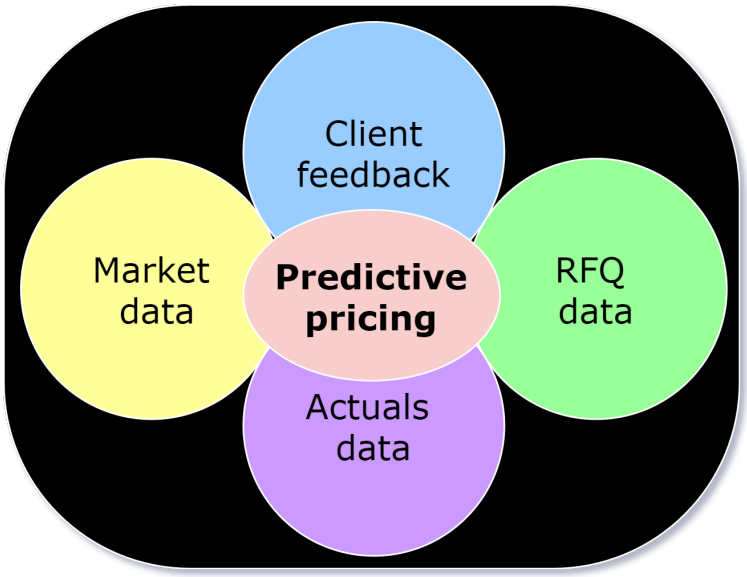

Figure 4.1: The four pillars towards predictive pricing

4.1 Pillar 1: Client Feedback

From June 2021 onwards, all offered prices are being tracked, together with feedback. In order to increase the competitiveness of the prices, feedback is valued highly by ECC. Worth mentioning is that the client feedback by itself is not satisfying enough according to ECC specialists. The reason for this is that the intention of clients using this feedback mechanism is not clear. Clients might use the feedback mechanism to their own advantage: by providing feedback in the form of unrealistic prices/percentages, the company might offer more competitive prices to the client in the future. Therefore, it is difficult to gather feedback with a high level of quality. Nevertheless, ECC is inclined using this knowledge in an intelligent way. An important drawback of the client feedback is that it is only being logged from June 2021. This means that only a scarce amount of the complete dataset is accompanied with feedback: only 8.3% of the lanes in the complete dataset are provided with feedback.

The feedback structure at ECC is explained with the help of Table 4.1 below.

Table 4.1: Feedback concept

Code	Explanation
0	RFQ open/not closed
1	Closed by the RFQ-desk
2	Rejected by client - no information
3	Rejected by client - general feedback
4	Rejected by client - detailed feedback
5	Closed by RFQ-desk - more detailed feedback in earlier round
6	Business won

First of all, code 0 is applied to business that is still in process. Code 1 is applied to RFQ's that are not of interest and are therefore closed by the RFQ-desk. Business accompanied with code 2, 3 or 4 indicates that the price offer is rejected by the client. The distinction between these code is determined by the quality of the feedback given by the client. When code 2 is applied, no feedback is given in any form. Code 3 implies that general feedback is given to the offer. According to the specialists active at the RFQ-desk, this means that vague feedback is received. Currently, it is hard to make sense of this data as the data is not clean. An example of this type of feedback is that ECC is in the top-30 of carriers offering a price to the RFQ. Code 4 is applied to RFQ's that received detailed feedback from the client. Furthermore, code 5 is applied to RFQ's that are closed by the RFQ-desk, but received detailed feedback in an earlier round. Practically, this means that it is decided to not proceed with the RFQ because of the feedback of the client. Lastly, code 6 is applied to RFQ's that result in awarded business. From the different codes, especially codes 4 & 6 seem relevant for modelling purposes in the way that provides feedback in sufficient quality with respect to the calculated prices.

A histogram, which is shown in Figure 4.2 below, is made in order to see the distribution of the feedback codes and to see how the feedback codes are used within ECC.

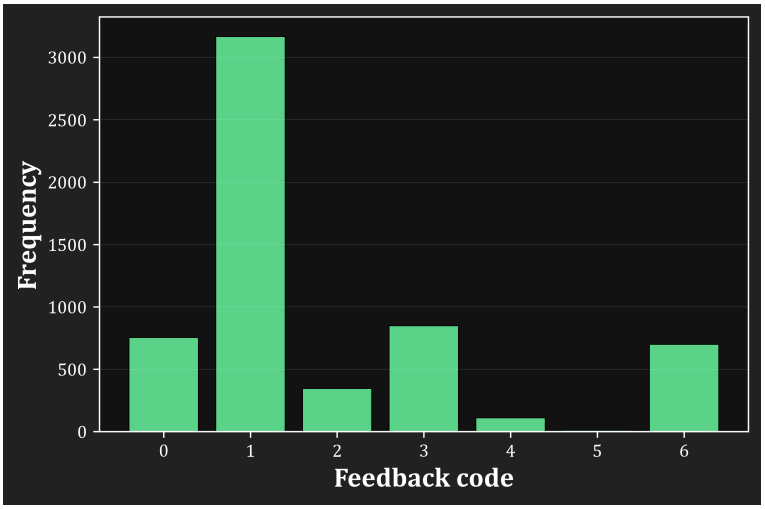

Figure 4.2: Histogram of feedback codes

From the figure, it appears that code 5 is never applied in practise. It also becomes clear that feedback code 1 is applied to most RFQ's, implying that most RFQ's are closed because they are not interesting enough. Another considerable amount of data is accompanied with code 0, implying that the RFQ is not closed yet. According to an ECC specialist, this is not strange as the duration of a RFQ may take up to six months. From all types of feedback, it appears that detailed feedback (code 4) is the least prevalent one, albeit that the RFQ-desk chases its client for this type of feedback in particular, bringing to heart that the feedback may help both parties in the future: ECC would be able to offer more competitive prices, while the client will receive better prices in the future.

From the feedback-codes discussed, data accompanied with either code 2, 3, 4, or 6 are included in this research, which provides this research with information about the competitiveness of the price.

In addition to the feedback codes which are measured on RFQ-level, there is a dataset available that contains the gathered feedback on lane-level. The detailed feedback mentioned before (code 4) can either be in the form of a price, or in a percentage difference from the offered price. In this dataset, there is a distinction between the two: when a numerical value is collected this is seen as quantitative feedback, which should represent the best price offer collected by the client. When a percentage is acquired, this is defined as qualitative feedback, which represents the difference with the best price offering received by the client. Furthermore, the feedback related to awarded business (RFQ's with code 6) are equal to the price offered by ECC. In the dataset, this awarded value is visible as quantitative value. The qualitative feedback in this dataset is not clean enough in order to make sense from it modelling-wise. This means that in this research, only quantitative feedback is used. How this feedback will be included in the research in more detail, will be discussed in Section 5.

4.2 Pillar 2: RFQ Data

Furthermore, it is investigated what parameters are currently being logged at ECC that influence the price offerings. As mentioned in Section 3, a majority of the parameters that are included in the current price calculation (Figure 3.2) of the sales price are not being logged to a database. Although the calculation is executed with the help of the SRM at the Sales Engineering department, only the output of the calculation (the offered prices) is being logged, in absence of the parameters that lead to the final price calculation. At the RFQ-desk, the importance of data storage is more widely acknowledged. In this department, as much data as possible is being stored, of which the most important ones are shown in Table 4.2 below.

Table 4.2: Available variables in RFQ data

Country-relationship	**Zone-relationship**	**Expected payload**	**Annual nr of shipments**	**Distance (KM)**	**Planned implementation date**
Transport type	**Lead time**	**Equipment type**	**Transport mode**	**Value calculation**	

The most detailed level at which these parameters are being measured, is at lane-level. Because a RFQ consists of multiple lanes and rounds, and ECC has multiple business units, it is easy to get lost in the data. The data structure maintained at the RFQ-desk is explained with the help of one-to-many relationships shown in Figure 4.3 below. The RFQ-desk has full control of this data structure.

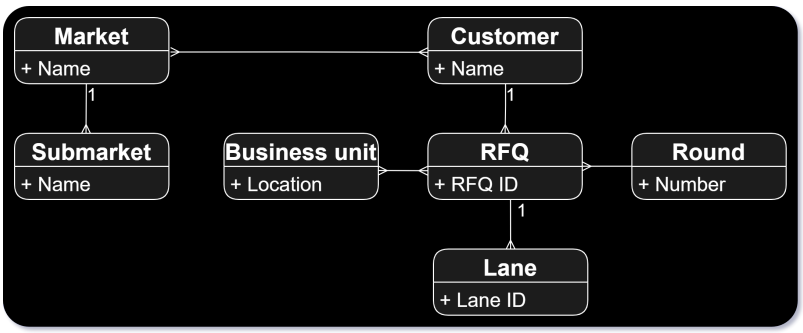

Figure 4.3: Data structure of RFQ's

A RFQ is initiated by a customer, which is active in a (or more) particular market and sub-market. When the RFQ-desk does not close the RFQ, the RFQ enters the business process of ECC, where

the RFQ consists of a "RFQ ID". Within a RFQ, multiple lanes exist, which are the trajectories on which price offers are collected. These lanes contain a unique "Lane ID". Multiple business units are allowed to participate in RFQ's, whereas they offer (only) to lanes of interest to the business unit. Therefore, it might occur that at lane-level, multiple offers are being collected. Moreover, the offering process consists of multiple rounds. In order to track down the price offers, the offers are composed of a "RFQ ID", a "Lane ID", a "Business unit" and a "Round number". A combination of these four factors is always referring to a unique offer in the database.

In order to gather as much data as possible, while maximizing the cleanness of data, all price offers related to RFQ's in which ECC participated are included in this research. These price offers are reviewed by the RFQ-desk and are assumed to have a sufficient level of data quality.

4.3 Pillar 3: Market data

Following the vision of Predictive Pricing in Figure 4.1, another data source available is related to external market information. This market information originates from a platform that is widely used for the RFQ process: TI-contract (*About Ticontract | Transporeon*, 2021). Here, many companies publish their RFQ's at which organizations like ECC are able to respond in order to win business. This platform is up-to-date in terms of competitive prices on trajectories as it has all the data available of business won by RFQ's that were made available via this platform. The company behind TI-contract (Transporeon) sells this information to interested parties. The level of granularity in which the data is available is at country relationship. Data is available for each week of the year. However, there is a considerable amount of data missing the dataset, which is explained with the help of Table 4.3 below.

Table 4.3: Observations in the market dataset by date (year-week) and their frequency (count) in the dataset

2019	Count
2019-W01	194
2019-W05	1
2019-W14	194
2019-W27	194
2019-W40	194

2020	Count
2020-W01	194
2020-W14	194
2020-W27	194
2020-W40	194

2021	Count
2021-W13	194
2021-W26	194
2021-W39	194
2021-W53	194

2022	Count
2022-W09	194
2022-W12	2
2022-W13	194
2022-W14	3
2022-W15	3
2022-W16	2
2022-W17	2
2022-W52	194

In the table, the observations per year are displayed. As mentioned before, the data in the market dataset is available on a weekly basis. Therefore, the observations are formatted in a year-week fashion. For example, there is only data available for weeks 1, 14, 27 and 40 in the year 2020, which implies that a lot of weeks remain absent in the market data. Moreover, it becomes clear that even for some dates, only information for one country relationship is available (e.g.: week 5 in 2019). Therefore, in this setup it is not possible to completely add market data to all the samples in the RFQ dataset.

Investigating in greater detail what information is exactly available in the market dataset, leads to the following discovery: Contract prices (€/km); Contract Rejection Rate. The contract price is the price/km value on a particular country relationship, whereas the contract rejection rate is a measure of the market capacity and will explained in more detail in Section 5. Including this market information, ECC generates insights in the competitiveness of the offers it brings to the customers and it is therefore only logical to include this information in the predictive pricing model. Other information, such as capacity-index, is also available but unfortunately, too much data is missing in order to be useful. In the future, it is likely that an increased level of clean data will be available from the market data source which may positively impact the models in this research.

4.4 Pillar 4: Actuals Data

Lastly, there is a dataset available, which is related to the invoices at ECC. This data source is referred to as the "actuals data", because in this data, the actual realizations can be found. When a transportation is completed, details will be registered in this dataset. The data included in the actuals dataset involves a variety of aspects, where trucking costs or ferry costs (related to oversea transportation) are just a few of them. Investigation of this data-source leads to the conclusion that the actual bookings do not always coincide with the agreements made in the RFQ. These deviations may be explained by both the carrier-side and the client-side. With respect to the client-side, it may occur that no capacity is available in order to realize the transportation. Naturally, the fault lies with the carrier in such a situation. On the contrary, because of production issues, it may occur that transportation is cancelled, while the capacity of the carrier-fleet is reserved for this particular transport. In this case, the fault lies with the client. Nevertheless, by including this data in this research, not only market information, client feedback and RFQ data is included, but also the company's realizations. As discussed before, the feedback from clients may be biased, while the invoices of ECC do not lie: it represents the current situation of transports.

Opposed to the RFQ data (which stores data under a "RFQ ID"), the actuals data is logged under a trip-number. In an intermodal transport option (with only one modality), a trip consists of multiple Trip-legs. For example, a first leg involves trucking from the loading location to the intermodal terminal. The second leg concerns the intermodal transport, whereas the third leg involves the trucking to the destination. Trips may consolidate multiple consignments, whereas a consignment is equivalent to a client order. When ECC decides to consolidate multiple consignments, this means that the trip is not unique anymore. Related to this project, the data on consignment-level is most relevant. As mentioned before, trips are not logged under a RFQ ID, which complicates the direct connection between RFQ's and actuals data: these sources do not have any overlapping Key. However, in the trips it is mentioned what the origin and destination is, just like in RFQ's. An attempt is made in order to connect the RFQ dataset with the actuals dataset via this way, but complexities with the postal-codes restricted this procedure. The most relevant challenge was the amount of digits in the postal-code. In the RFQ process, which happens well in advance of the actual realization, the full postal-code is not yet available (mainly 2-digit postal-codes) while in the actuals dataset, the full postal-code is available. Because of these complexities, it is decided to exclude the actuals dataset from this research. In the future, when it is possible to connect the two datasets, the information in this dataset can be included in the analysis, potentially contributing to the models employed for the prediction of prices.

4.5 Conclusion

In conclusion, the datasets that are available in all four pillars (Figure 4.1) contain information that is likely to be relevant for modelling. However, linking all sources is not always straightforward. Important to mention is that most observations in the dataset do not contain information from all available datasets. From the market data, some weeks and country relationships were missing, while actuals data may be missing because transportation is not yet executed. Because of the exclusion of the actuals dataset, the RFQ-data, feedback data and market data remain as data sources relevant to this research. Bringing together these data sources, a "truth" has to be derived. This truth should represent the price of transport on lane-level accurately. A decision has to be made on how to combine the RFQ truth (business won), the market truth and the client truth (feedback). Moreover, the variables influencing this truth should be investigated. This subject is elaborated in more detail in the upcoming section (Section 5).

5 Data Preparation

This section describes how the data is prepared for modelling. From Section 4, it followed that three data sources are relevant to this research. In this section, the relevant data from these data sources are pre-processed, analyzed and transformed. By pre-processing, the data is cleaned to eliminate outliers and missing data. From the analysis, it becomes clear what factors in the data sources are relevant to this particular research and how they relate to the price per kilometer that is to be predicted. Finally, the data is transformed, such that the model interprets the data in the right way. From the scheme in Figure 5.1, all phases but "Modelling" are treated in this section.

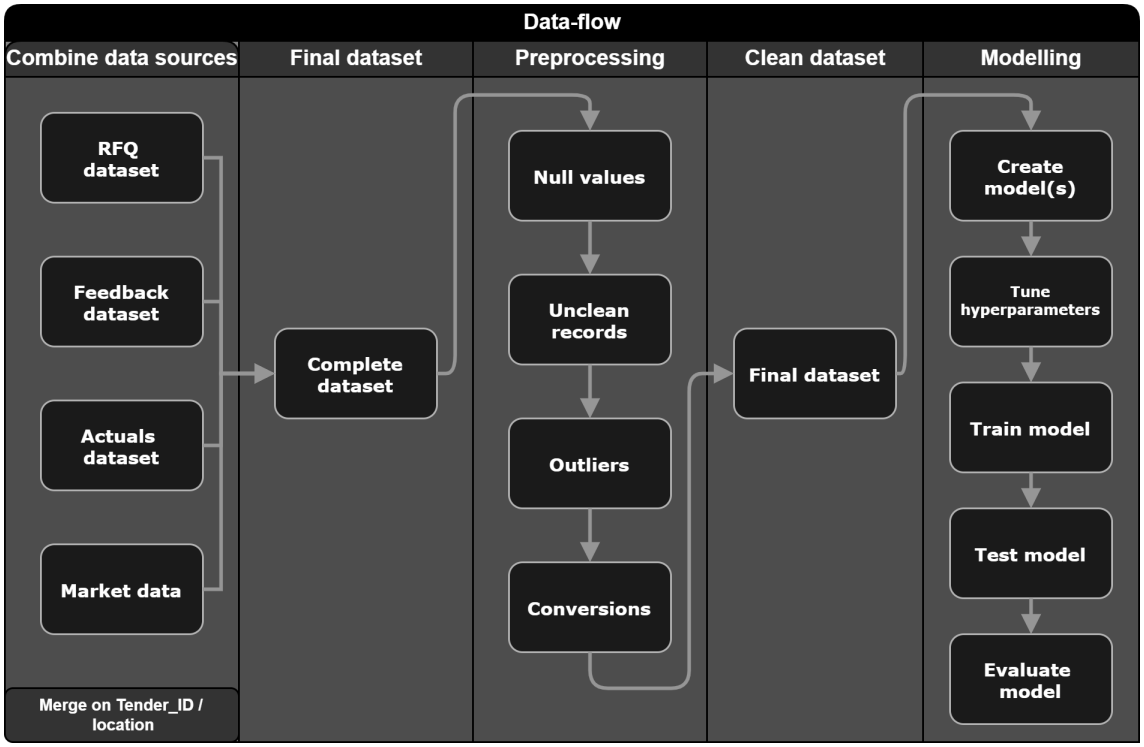

Figure 5.1: Data flow

5.1 Final dataset

From Section 4, it appeared that four data sources are relevant to this research. However, it appeared that in its current state, the actuals data source is not useful and therefore excluded in the research. This means that the basis of the final dataset is formed from the RFQ-, feedback- and market data source.

Following from subsection 4.2, it should be clear that the granularity of data should be chosen carefully: for a specific lane on a given time, multiple samples may exist. This is possible because a RFQ may consist of multiple rounds, whereas multiple business units (BU) are able to offer on the same lane. Ultimately, there is one truth for this lane at a given time. Therefore, this research is interested in one sample for each lane-date combination in the dataset. Next, from a discussion with important stakeholders, the choice has been made to filter on the samples that are accompanied with an allocated price. A price is allocated when the offer proposed by a BU is accepted by the RFQ-desk. Therefore, the allocations are in control of the RFQ-desk and this way, a minimum level of data quality can be ensured as the RFQ-desk is responsible for treating the data. Moreover, when filtering on these samples, only the last rounds are included in the resulting dataset, which is ensured by the logic of allocations. Noteworthy is that this filtering still involves awarded business and business

that is not won. Because the project's goal is to predict competitive prices on lane-level, it is further investigated what aggregations should be made. From the final dataset, an analysis is done which helps in understanding the variables more thoroughly. The variables are treated individually and cleaning steps are explained in detail. From the RFQ dataset, the variables displayed in Table 5.1 below are worth investigating. The scale of the variables is also shown in the table. A ratio scale refers to a numerical variable with a meaningful zero, while an interval scale refers to a numerical variable with equal distances between the values. A nominal variable cannot be ordered in any meaningful way and is (in this research) not a numerical value, but a categorical one.

Table 5.1: Classification of variables in the RFQ dataset

Variable	**Scale**
Lane	Nominal
Distance	Ratio
Date	Nominal
Number of rounds	Interval
Customer	Nominal
Strategic End-Market	Nominal
Lead-time	Interval
Annual number of shipments	Ratio
Modality	Nominal
Equipment	Nominal
Initiative	Nominal
Current business	Ratio
Margin	Ratio
Expected payload	Ratio
Transport type	Nominal
Contract rejection rate	Ratio

5.2 Variables

5.2.1 Price per kilometer

It is decided that the price per kilometer (€/km) is the only dependent variable in the dataset. In other words: this is the variable which this research attempts to predict. For each sample in the population, one truth exists in the form of this price. However, the price is strongly dependent on the distance of transport. The most important thing according to ECC specialists it is to be able to predict the price per kilometer, which can be translated to a lane-specific price afterwards. After choosing the dependent variable, the next step is to choose from what data sources this price can be derived. In Section 5, it came to light that the price per kilometer may originate from the RFQ dataset, from the market dataset, or from clients' feedback. ECC wants to exploit its own knowledge to the fullest extend. Therefore, the RFQ is preferred over the other data sources when this is possible. However, as mentioned before, not all samples in the population represent business that is won by ECC. Therefore, RFQ data cannot always be used for the price per kilometer variable. For these samples, the clients' feedback or market data should be used instead. Feedback is not always provided by the client. Therefore, sometimes, only market data can be used. For this research, it is decided to combine the datasets according to the following priority rule: RFQ data → feedback data → market data.

Although the price in the market data is already provided in a €/km unit of measurement, prices in the RFQ- and feedback data is provided with full price (€). For these observations, the prices are divided with the distance in kilometers. This distance variable will be explained more thoroughly later in this section. After applying the priority rule just defined, the price per kilometer values of the samples are investigated in more detail. Some samples contain an unrealistically high value for

price per kilometer. Therefore, the following rule is defined in consultation with ECC specialists: when $price\_km > 5$ & $distance > 500km$, the sample should be dropped from the dataset. These samples should not be considered clean. After this procedure, the price per km variable can be summarized with the help of Table 5.2 below.

Table 5.2: Descriptives of variable: price per km

price_per_km	
count	13421
mean	1.28
std	0.41
min	0.10
25%	0.98
50%	1.25
75%	1.55
max	4.94

Next, the samples are displayed in a boxplot, which is shown in Figure 5.2 below.

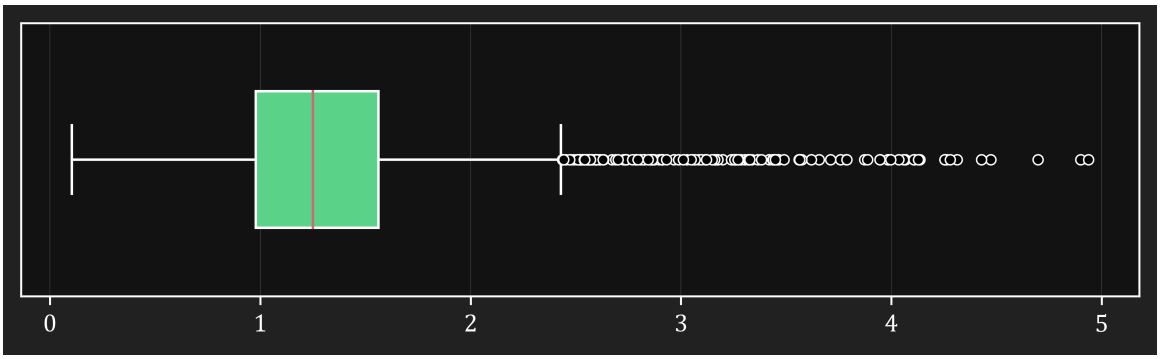

Figure 5.2: Boxplot of the price per kilometer (whiskers set at $1.5 \cdot IQR$ and the median in red)

The sample points $price/km > 1.5 \cdot IQR$ are not considered as outliers, because they can be realistic according to ECC specialists and the set of rules that were defined above. The observations outside of the whiskers can be explained by transports with a relatively low distance. Distance from the Netherlands to the Ruhr-area in Germany is an example of this. While it is possible that the trajectory is only 50 kilometers in length, the costs are high in comparison to the amount of kilometers. This is due to the (un)loading costs involved in the process.

5.2.2 Lane

A lane corresponds to a trajectory from A to B. Transports are typically stored by postal-codes: for each transport a collection country, collection postal-code, delivery country and a delivery postal-code are stored in the system. The level of detail in postal-codes is high, which complicates capturing patterns from the data because too few samples exist from the lane under investigation. Therefore, the lane information is aggregated in this research. The aggregation levels available are: zone-relationship; country-relationship. A country-relationship is straightforward, whereas the zone-relationship requires some more clarification. Nowadays, the zone-relationship is common practise at ECC. From Section 3.1, it appeared that a zone relationship is implemented in order to develop a product portfolio. With this, ECC is able to allocate products to its business units. The goal of this is to allocate all business units only to business in which the BU has a competitive edge. ECC subdivides each country in one or multiple zones. Small countries (such as the Netherlands) contain one or few zones, whereas bigger countries (such as Germany) contain multiple zones. The zones can be found in Figure 5.3 below.

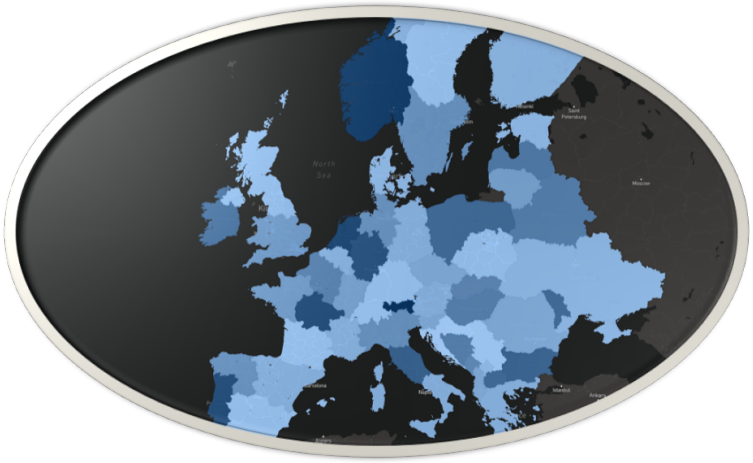

Figure 5.3: An overview of the zone-relationship as a result of the Product Database

In order to investigate the different aggregation levels in more detail, the pie-charts in Figure 5.4 below are created.

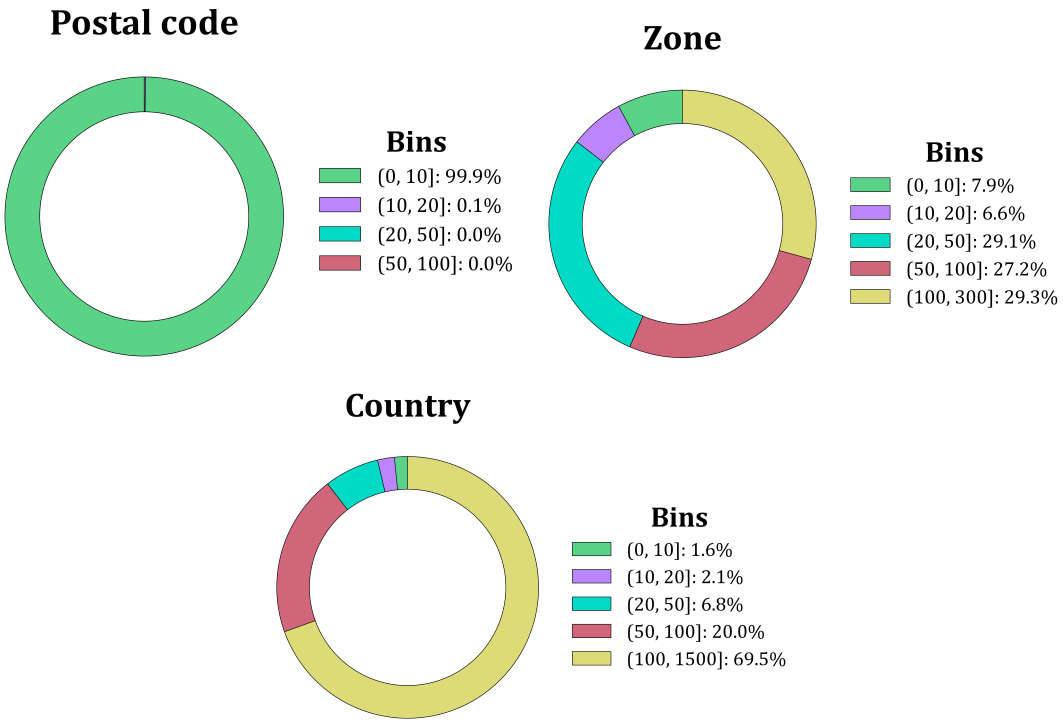

Figure 5.4: Aggregation results: frequency in percentages

In these charts, the fractions reflect the presence of the bins. Bins are created which refer to the amount of observations in the data. It appears that when aggregating on postal-code, 99.9% can be subdivided in the bin $(0, 10]$, which implies that 99.9% of the complete dataset contains a postal-code relationship which only occurs between 0 and 10 times. This bin range is chosen, because it represents a situation in which it will be difficult to capture any patterns from this scarce amount of data. With respect to the zone-relationship, this already seems improved: only 7.9% of the complete dataset is accompanied with a zone-relationship that appears between 0 and 10 times. The rest of the zone-relationships appear more often. Most observations are retrieved when aggregating on country-relationship: only 1.6% of the complete dataset contains a country-relationship which occurs

less than 10 times. A careful consideration has to be made on the aggregation level: a sufficient amount of observations has to be present in the final dataset, while not losing too much information related to the price per kilometer. According to ECC specialists, a big difference may exist in prices between different zones. When there is too much aggregation, the level of competitiveness will be reduced. To help decide on the aggregation level, the other data sources are considered as well. Looking at the market data (Section 4.3), the highest level of detail is at country relationship. For observations where business is not won by ECC, the truth of the price is defined by the market price. For these market prices, the zone relationship remains unknown. Therefore, it is only logical to set the aggregation-level to country relationship. In the future, market data might be available in more detail and this decision can be reconsidered. Additional information related to the country relationship is shown in Table 5.3 below.

Table 5.3: Data analysis: Country-relationship

Country relationship	
count	13,518
unique	166
top	DE-DE
freq	1,433

Every observation is concerned with a country-relationship. Therefore, this is no missing data related to this variable. According to the analysis, 166 unique country-relationships are present in the dataset and the domestic transport "DE-DE" (transport from Germany to Germany) is the most prevalent one, which appeared 1,433 times in the dataset. The country codes comply with the Alpha-2 codes, which can be derived from Section 11.2.

5.2.3 Distance

The distance considerably impacts the price per kilometer variable according to ECC specialists. However, in order to collect accurate distances for every sample, some challenges need to be addressed. The first challenge is that the RFQ dataset does not contain a distance for a lot of data: only 50,095 data entries in the RFQ dataset is accompanied with a distance, which comprises about 50% of the data. An additional challenge for this data is that the distance calculation is often not correct.

ECC has access to tools in order to calculate distances. The most important one is PTV (*PTV Group*, 2021), which offers the tools in order to Geo-coding and retrieve distances between coordinates. Geo-coding is the process of translating addresses into coordinates. This step is necessary in order to retrieve accurate distances from the PTV server. In order to translate addresses into coordinates, the PTV server requires some level of accuracy, whereas this accuracy is country-dependent: some countries require less accurate input (e.g.: a two-digit postal-code) and some countries require more accurate input (e.g.: a four-digit postal-code). As mentioned in Section 4, a full postal-code is often lacking in the RFQ data source. In order to maximize the amount of observations that can be filled with a distance, the two-digit postal-codes are matched to a similar, full-digit postal-code originating from the actuals data source. In this source, the full postal-codes are available and based on the actual transport realizations by ECC. In consultation with ECC, it is decided that this procedure is currently the best in order to calculate distances of lanes. Subsequently, these full-digit postal-codes are send to the PTV server, returning a distance for all the requests, accompanied with a matching score. The samples with a matching score higher than a threshold value are retained in the dataset. This matching score indicates the accuracy of the returned distance and the threshold is set at 80%.

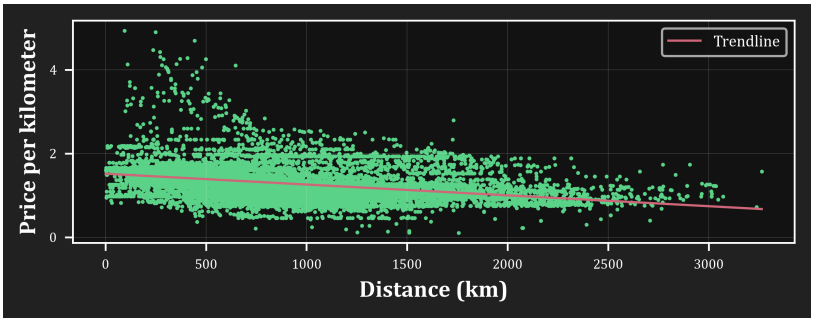

Figure 5.5: Distance versus price per kilometer

From Figure 5.5 above, it appears that there is a trend in the data: the higher the distance, the lower the price per kilometer. Moreover, some transports below 500 kilometers seem to have an exceptionally high price. According to ECC specialists, this can be explained due to the relative high prices for low distance transportation. An example of this is a transport for the Netherlands to the Ruhr-area in Germany. It is possible that a transport on this trajectory is only 50 kilometers in length. However, due to the relatively high amount of time involved in the (un)loading process, the price per kilometer will be high in this case.

The distance variable is not present in the final dataset. As mentioned before, the distance is solely used to compose the price per kilometer variable for the RFQ and feedback dataset. The market data is already available in the form of a price per kilometer.

5.2.4 Date

In order to account for seasonality, the date is investigated as well. A monthly level of detail is chosen for this purpose, because contracts typically start at the beginning of a month. Furthermore, at the RFQ-desk, it is busier at the end of the year compared to the beginning of the year. Interesting to investigate is whether or not ECC has a more powerful position in the negotiations at the end of the year. The distributions from the dates are visualized in Figure 5.6 (upper) below. The x-axis represent a year-month observation.

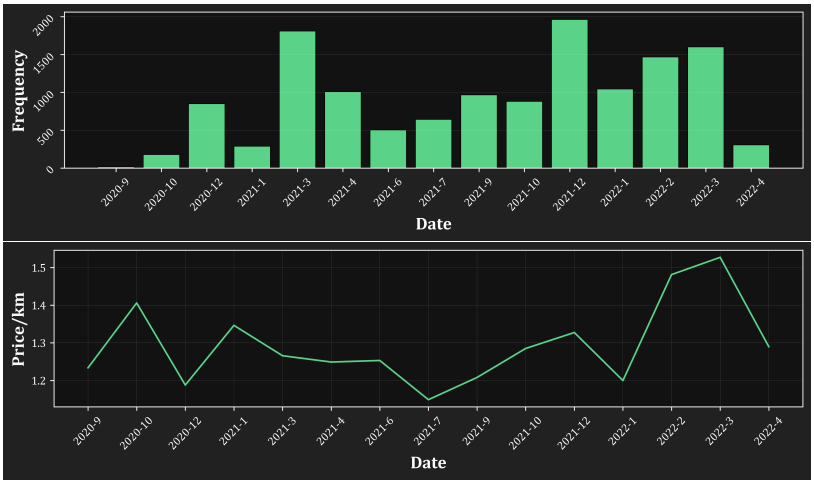

Figure 5.6: Frequency of transports (upper) and price development (lower)

From this figure, it becomes clear that dates range from September, 2020 until April, 2022. This can be explained by the improvements in data management at ECC since 2020. Before, data quality was not a hot topic. Because of this, it is hard to accurately get a grasp of the seasonality as the data only accounts for one full year (2021). However, it is still interesting how the dates relate to the

price per kilometer variable. For this matter, a one-way ANOVA analysis of variance is conducted in order to investigate whether some groups differ significantly from each other, by comparing the data for equal means. This analysis is build on the assumption that the standard deviations of all groups are equal. In order to test this, Levene's test (Levene, 1961) for homoskedasticity is employed because in Box (1953) it is stated that this test for equal variances is more applicable than Bartlett's test (Bartlett, 1937), specifically for an ANOVA analysis. The results of these tests are displayed in Table 5.4 below.

Table 5.4: Results of the statistical tests related to seasonality

Levene's test	
p-value	0.57
statistic	0.84

ANOVA	
p-value	0.36
statistic	1.10

From these tests, it appears that both null hypotheses can be accepted ($p > 0.05$). For Levene's test, this indicates that the populations for each unique month have equal variances. Therefore, performing an ANOVA analysis is justified. Next, in the ANOVA test, a p-value of $p > 0.05$ implies that the investigated populations have equal means. Therefore, it is concluded that no seasonality is detected in the data and it is therefore decided to exclude the date variable into the dataset. In the future, when the population is larger, the assumption can be reconsidered.

5.2.5 Number of rounds

The next variable under consideration is the number of rounds a RFQ contains. When multiple rounds exist, the prices may be driven down because the transport companies are competing with each other in order to win the business. On the other hand, when problems on the market exist, such as capacity problems, the client might harm its own interests. Namely, when capacity is low, transport will be relatively expensive. To see the summary and how the number of rounds relate to the price per km, Table 5.5 below is created.

Table 5.5: Summary of the variable: number of rounds

rounds	
count	13421
mean	1.52
std	0.57
min	1
25%	1
50%	1
75%	2
max	4

Number of rounds	**1**	**2**	**3**	**4**
count	6812	6315	197	97
mean	1.33	1.24	1.17	1.13
std	0.44	0.37	0.3	0.25

From the left table it becomes clear that more than 50% of the RFQ's exist of one round only. Moreover, the RFQ with the most rounds contained four rounds. The mean and standard deviation are not meaningful because it concerns an interval scale. From, the right table, it becomes clear that by far the most RFQ's exist of only 1 or 2 rounds. Noteworthy is that it seems that the more rounds a RFQ contains, the lower the price per km. A logical reasoning for this would be the already mentioned competition between LSP's in order to stay in the race for winning the business. Nevertheless, from this analysis, it becomes clear that including the number of rounds in the final dataset would be wise thing to do because of the implications from this analysis.

5.2.6 Customer

Another interesting variable in the dataset is the customer. Generally speaking, the higher the service level required, the higher the eventual price for transport will be. Therefore, some customers typically pay more for transport than others. Some examples of requirements are the type of truck

they want for their transport, or the need for some additional accessories in the truck. Moreover, some customers request a higher percentage of on-time deliveries of transport than others. When this percentage is relatively high, the cost of transportation will be higher. The analysis of the different customers can be found in Table 5.6 below.

Table 5.6: Summary of the variable: customer (actual customers are confidential)

customer	
count	13421
unique	347
top	Customer A
freq	797

Customer A (actual customers are confidential) appears to be the largest customer in terms of number of lanes. Furthermore, the customer base currently exists out of 347 unique customers. To see the distribution of the prices for the customers, the histogram shown in Figure 5.7 below is drawn. Here, the mean prices/km of all customers are plotted.

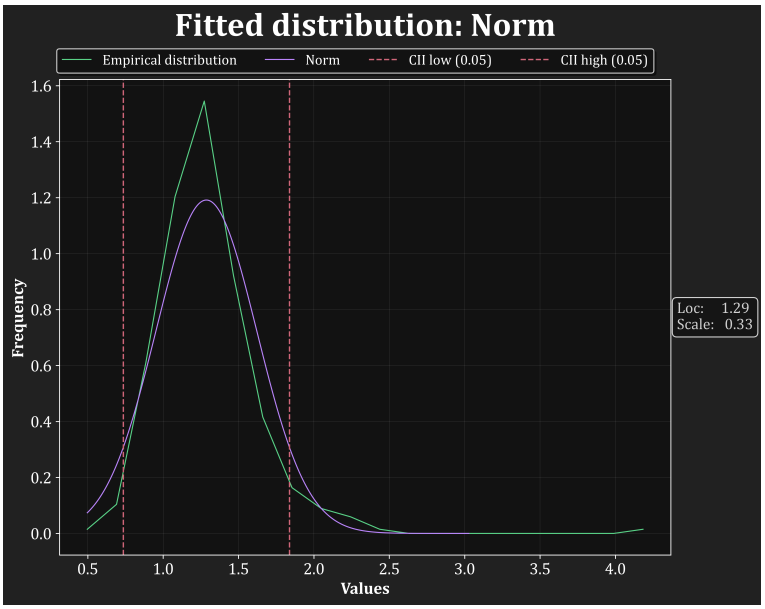

Figure 5.7: A visualization of the distribution of the data related to the price per kilometer

The purple line seems to fit the distribution of the data quite well, which represents a normal distribution ($\mu = 1.29$ & $\sigma = 0.33$). This normal distribution is 95% sure that the mean of the population falls between the two red lines. The 5 customers with the highest prices are shown in Table 5.7 below.

Table 5.7: Top-5 most expensive customers in terms of price per kilometer

Customer	**Price/km**
Customer B	4.28
Customer C	2.35
Customer D	2.33
Customer E	2.29
Customer F	2.2

Customer B appears to be paired with a high price per kilometer. However, this customer appears only once in the dataset, which complicates drawing a conclusion from it.

5.2.7 Strategic End-Market

According to ECC, the concerning SEM is a relevant factor for price-setting as well. Practically, this means that when calculating the rates, corrections are applied for particular SEM's by the people who have the knowledge to do so. In the future, ECC want to structure this "discounting" process by maintaining all the discounts in a robust way. All SEM's are listed in Table 5.8 below, along with the activity per market and the mean prices.

Table 5.8: Activity and prices per strategic end-market

SEM				
Market	**frequency**	**mean**	**std**	**Distance**
Paper-packaging	3325	1.21	0.34	858
Automotive1st2ndand3rdsupplier	2297	1.36	0.42	766
Industrial	1551	1.33	0.6	947
Consumer electronics	1354	1.3	0.31	1134
Chemical	1161	1.21	0.38	918
Logistics provider	1131	1.21	0.41	922
Consumer goods	1018	1.32	0.37	874
Building construction	808	1.45	0.28	608
Automotive OEM	419	1.31	0.44	1224
Agricultural	211	1.23	0.3	1008
Fashion	72	1.24	0.34	813
E-commerce	67	1.14	0.35	816
Waste recycling	5	0.95	0.02	774

From the table, it becomes clear that there is a lot of variation in prices between SEM's. While Waste recycling may not be representative because of only five occurrences, transport in the Builing construction SEM seem to have a high price in comparison to other SEM's. This does not seems strange at all as this SEM often concerns transport with a low distance compared to other markets according to ECC specialists. This is indeed confirmed with the help of the table above. The Automotive OEM market contains relatively large-distance and highly priced transportation. This is explained with the value of the goods that are being transported. Transportation in the Automotive OEM- and Consumer electronics markets are typically paired with high-valued goods. Paper-packaging products are typically low-values goods. The value of goods and the distance of transport are the main reason for the price differences in the markets according to ECC specialists.

5.2.8 Lead-time

The next variable to be analyzed is the lead-time. The lead-time has a lot of effect on the eventual price, because for longer lead-times, cheaper options become available. For road options, this may imply that toll routes can be avoided. With respect to the modalities, this may imply that a ferry or train can be used, rather than the road option. Intermodal options are often cheaper than road options, so this variable seems an important one.

First of all, it is important to understand the lead-time structure at ECC, which can be done with the help of Table 5.9 below.

Table 5.9: Lead-time structure

Leadtime	Mon	Tue	Wed	Thu	Fri	Total	Numerical lead-time
A-A	1	1	1	1	1	5	1
A-B	2	2	2	2	4	12	2,4
A-C	3	3	3	5	5	19	3,8
A-D	4	4	6	6	6	26	5,2
A-E	5	7	7	7	7	33	6,6
A-F	8	8	8	8	8	40	8
A-G	9	9	9	9	11	47	9,4
A-H	10	10	10	12	12	54	10,8
A-I	11	11	13	13	13	61	12,2
A-J	12	14	14	14	14	68	13,6

A lead-time of "A-A" corresponds to a same-day delivery, which equals a lead-time of 1 in days. Likewise, "A-B" corresponds to a lead-time of two days. However, when the transport starts on Friday, the weekend is in-between the transport and the delivery will be finalized on Monday, which results in a four-day delivery rather than a two-day delivery. Therefore, the actual lead-times depend on the day on which the transport has started. In order to reduce the complexity of the lead-time, it is decided to introduce a numerical lead-time, which is also displayed in the table above. This numerical lead-time is an average of the lead-times on each of day of the week. For lead-time "A-B" this implies that the total lead-time of 12 days is divided by 5 days, resulting in an average lead-time of 2.4 days. After investigating the data in more detail, it became clear that a fraction of the data was contaminated. After deletion of these observations, the distribution of the lead-times can be found in in Figure 5.8. In order to draw the histogram, 10 bins are created because 10 unique lead-times exist in the dataset (see Table 5.9).

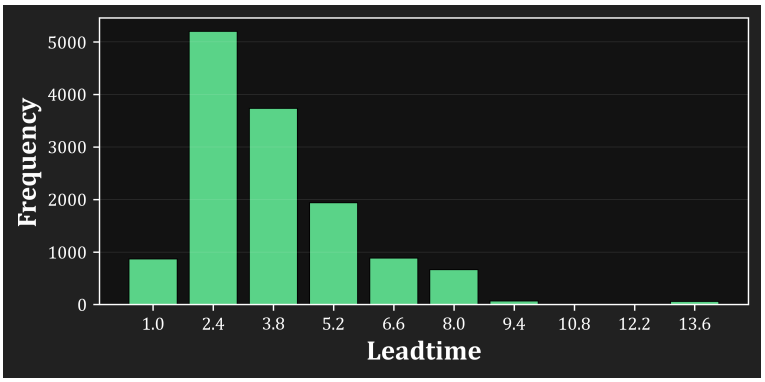

Figure 5.8: Lead-time distribution

From the bar-chart, it looks like a distribution can be fitted to the lead-times. After applying the fitting procedure explained in Appendix B (Section 11.3), Figure 5.9 is drawn below.

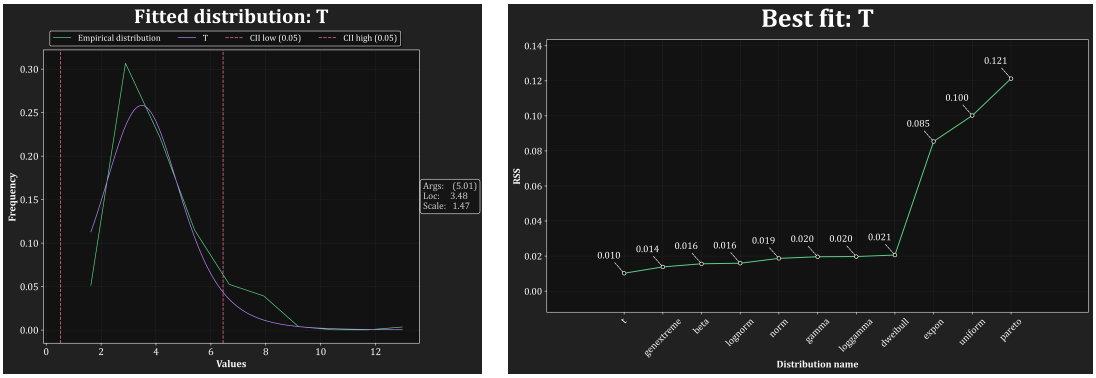

Figure 5.9: Best fitted distribution (left) and estimates (right)

From the figures, it appears that a T-distribution fits the data best. This is confirmed by the Kolmogorov-Smirnov test (Massey Jr, 1951) which cannot reject the null hypothesis of the data distribution being identical to the T-distribution ($p = 0.20$).

5.2.9 Annual number of shipments

The next variable in the RFQ dataset to be analyzed is the annual number of shipments. When the type of business is contract-business, multiple transports will be executed in order to fulfill the need of transportation from the customer. In this case, the annual number of shipments will satisfy the condition: $nr_{shipments} > 1$. On the contrary, with respect to spot business, only one transport is executed. Moreover, spot business mainly refers to spontaneous transport on the short term. Before performing the analysis, the samples containing $nr_shipments = 0$ are removed from the dataset, because these samples are contaminated. According to field experts at the RFQ desk, these cases exist in order to bypass the data validation fields in the dataset that is operational (it is mandatory to register a value under the annual number of shipments field). The variable's analysis is shown in Table 5.10 below.

Table 5.10: Summary of the variable: nr_shipments

Annual number of shipments	
count	13,380
mean	48.2
std	134.3
min	1
25%	1
50%	8
75%	42
max	4180

From the table, it appears that the mean annual number of shipments equals 48.2, while there is a lot of spread according to the standard deviation. Moreover, more than 25% corresponds to spot-business. The remaining part of the business concerns contract-business, with the largest RFQ in terms of annual number of shipments corresponding to 4,180 shipments, which boils down to around 12 shipments a day. After inspection of the annual number of shipments in relation with the price per kilometer variable, it is concluded that no correlation exists between the two factors. Therefore, the annual number of shipments is not included in this research.

5.2.10 Modality

For the transport mode it is meant what transport modality is operated. Before diving into deeper detail, the contaminated samples which do not contain a modality are removed from the dataset.

This contained a total of 8 samples. The different modalities are shown in Table 5.11 below.

Table 5.11: An overview of the different transport modalities

Modality	
Modality	**count**
Road	9403
Intermodal, short-sea	2770
Intermodal, train	823
Multi-modal	376

While the road option is self-explanatory, some records need some more attention. With intermodal, ECC means transports that do not only contain road transports. Examples of this may be an oversea transport or train transport. Air transportation is not included, as this is not among the core business of ECC. With multi-modal transport, the use of multiple modalities is meant. This concept excludes road transport, as this is a modality that is used for every transportation in some way. Even for intermodal transportation, transportation to the terminal needs to be realized with the help of road transportation. Considering the unique values in the dataset, with multi-modal, a combination of short-sea and train transportation is assumed. Within ECC, road transport is the most prevalent transport modality. Next, a substantial amount of transportation is realized by ferry. The amount of transports by train and multi-modality are small compared to road and short-sea, but they are still valued within ECC. The relation of transport modality to the price per kilometer is visualized in Figure 5.10 below.

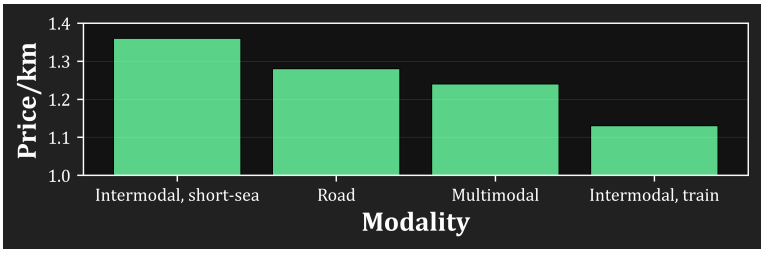

Figure 5.10: Mean prices/km for each transport modality

It appears that the short-sea option has the highest price per kilometer, whereas the train option has the lowest. It is hard to draw any conclusion from these numbers, because an alternative modality may be even more expensive on a particular lane. Additionally, it is interesting to investigate the prevalence of transport modalities in the different strategic end-markets. A deeper analysis showed that most markets conform to the order (in frequency) shown in Table 5.11 above. However, one market stands out: the automotive-OEM market makes use of the intermodal (short-sea) option most frequently, by quite some margin on road transportation.

5.2.11 Equipment type

With respect to the equipment type, it is meant what trucks/trailers are used for transport. For this variable, some pre-processing is performed before diving into deeper detail. Namely, unclean samples and samples that had no equipment type specified were dropped from the dataset. This concerned 149 samples in total. Next, some values were filled with different values while referring to the same aspect. More specifically, both "Remaining" and "Other" were present in the dataset. These values were merged into one value: "Other". Lastly, as a rule of thumb, the cut-off for the minimum sample size per independent variable is set at 10. This means the following equipment types are dropped from the final dataset: Other, Coil trailer, Box-trailer, 45ft Container, Temperature controlled, 46t Container. As a result, 8 different equipment types exist, which are shown in Table 5.12 below along with their frequencies and mean prices.

Table 5.12: Frequency and mean price of all the equipment types

equipment	count	equipment	count
Standard tautliner	5989	Jumbo road train	313
Mega xl	3963	3,5 ton	80
Mega	1978	7,5 ton	68
Mega xls	758	Van	39

The top three equipment types by frequency are: Standard tautliner; Mega XL; Mega. The Mega XL and Mega trailers are part of the ECC fleet, while the Standard Tautiliner is an equipment type that is widely used in the transport market.

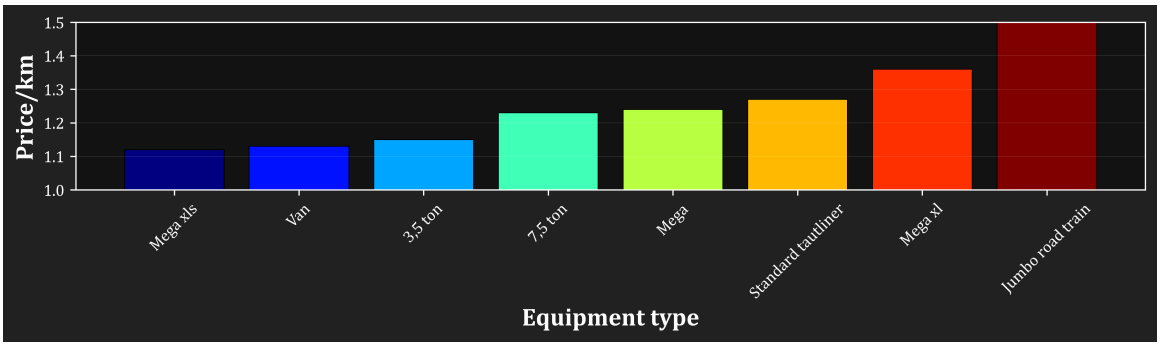

Figure 5.11: Mean prices by equipment type

From a first look, it seems that the Jumbo road train is the most expensive equipment type. This makes sense because it concerns a truck with two trailers, increasing the capacity of the truck. While the price per km may be larger than other equipment types, a larger volume can be transported, reducing the $price/m^3$. The means of the groups seem to differ substantially from each other. Unfortunately, this cannot be tested statistically, because Levene's test for homoskedasticity shows that it cannot be proven that all samples are from populations with equal variances ($p = 0.000$ according to Levene (1961)). Therefore, one of the assumptions of an ANOVA test is violated and we cannot compare the means of the groups via this way. Nevertheless, equipment type seems to be an important factor influencing the prices/km. Therefore, equipment types is included in the final data to be used in the modelling phase.

5.2.12 Initiative

According to stakeholders within ECC, it is important to distinguish between local and group initiatives. A group initiative is when the transportation is performed for the network BU and thus considers a transportation with value for the Ewals Group as a whole. On the contrary, a local initiative concerns a transportation which only has added value for a specific BU. These transports are only reported to the RFQ-desk on a central level, whereas the transport itself is managed by the individual BU. These individual BU may have access to a fleet which may be more competitive on specific lanes. The frequencies and mean prices of the initiatives are shown in Table 5.13 below.

Table 5.13: Group versus local initiative

Initiative	count	Price/km
group	10344	1.26
local	2844	1.38

From the table, it becomes clear that 78% of the samples refer to group initiatives. Moreover, the mean price of group initiatives tend to be substantially lower than local initiatives. In order to test whether the two groups are statistically different, Welch's t-test is performed which takes into

account groups with different sample sizes and variances under the normality assumption. From this, the two groups appear to have statistically different means ($p = 0.000$), which increases the interest in this variable.

Furthermore, interesting to investigate is the strategic end-markets for each of these groups. In Table 5.14 below, it seems that while the Ewals group has more activity in the consumer electronics market, the local initiatives do not have much activity in this sector. Furthermore, the industrial and logistics provider sectors seem to be more relevant for the local initiatives in comparison to the Ewals group.

Table 5.14: Activity in strategic end-markets by initiative

Group initiative	
Market	**count**
Paperpackaging	2874
Automotive1st2ndand3rdtiersupplier	1761
Consumerelectronics	1297
Chemicals	1024
Industrial	871
Consumergoods	761
Logisticsprovider	761
Buildingconstruction	617
Automotiveoem	194
Agricultural	118
Ecommerce	66
Fashion	0
Wasterecycling	0

Local initiative	
Market	**count**
Industrial	627
Paperpackaging	440
Automotive1st2ndand3rdtiersupplier	438
Logisticsprovider	363
Automotiveoem	217
Consumergoods	211
Buildingconstruction	191
Chemicals	133
Agricultural	92
Fashion	72
Consumerelectronics	54
Wasterecycling	5
Ecommerce	1

5.2.13 Current business

The next variable that is considered is the amount of current business ECC has related to the RFQ's. Interesting to see is the relation between the current business variable and the price per kilometer. This relation is visualized in Figure 5.12 below.

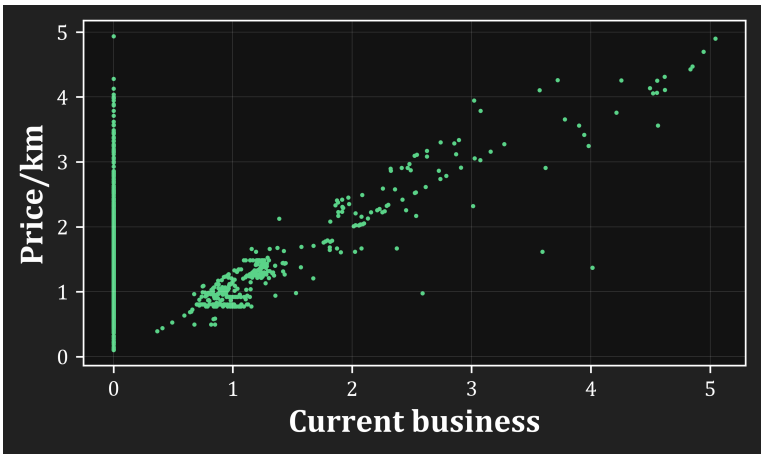

Figure 5.12: The relation between current business and price per kilometer

Apart from the cases where ECC has no current business, the relationship appears to be linear, which is something one would expect: if the price of the current business is high, the price of the new business will likely also be high. Moreover, interesting to investigate is in what markets ECC has the most business. However, this is something which is not possible with price per kilometer as

unit of measure. In the RFQ dataset, a "value calculation" variable is also available. This variable indicates the total value of the RFQ. When only the awarded business is considered, this is equivalent to the current business. The formula is shown in Equation 5.1 below.

$$Value\ calculation = \sum_{l \in L} (\text{Awarded-price})_l \cdot (\text{shipments}) \tag{5.1}$$

Here, L is the set of lanes belonging to a particular RFQ. The awarded price is a price which is only larger than zero when the business is awarded to ECC. The shares of value calculation per market is shown in in Figure 5.13 below.

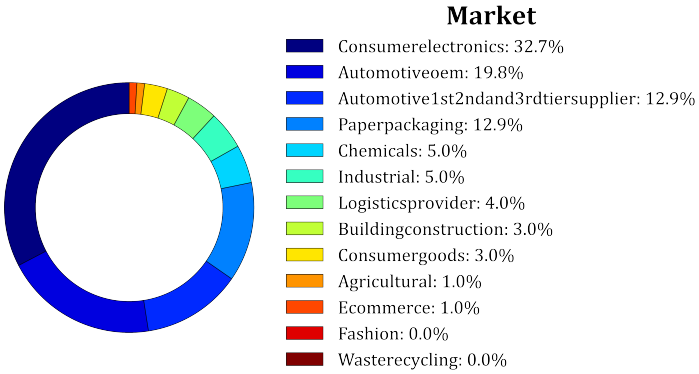

Figure 5.13: Business activity per market

From the figure, it is clear that more than 75% of the current business activity originates from the top-4 markets: Consumer electronics, Automotive and Paper-packaging.

5.2.14 Margin

Moreover, the margin is investigated. The margin is defined as the percentage of the agreed price, which reflects the profit for ECC. Interesting to investigate is how the margin relates to the price per kilometer. This relationship is visualized in Figure 5.14 below.

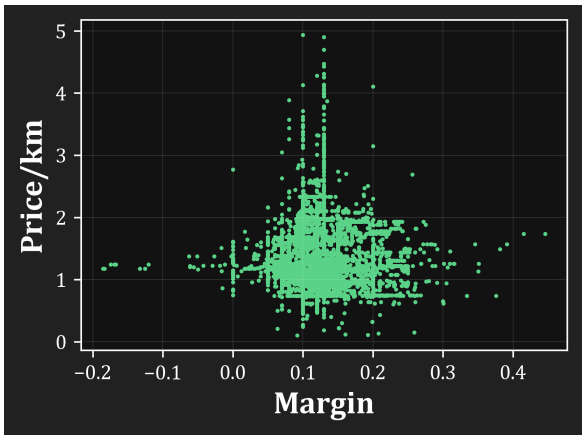

Figure 5.14: The relationship of margin versus price per kilometer

For a given price, the margin is all over the place: no clear relationship can be derived from the figure. Therefore, the margin is excluded from the dataset. The margin is something which is internally decided within ECC. When the market is favorable, a higher margin can be set in favor of ECC. It is logical that profit-organizations seek to maximize this margin, while still remaining competitive. It makes sense, that the margin percentage is unrelated to the price per kilometer variable.

5.2.15 Expected Payload

The next variable considered is the expected payload, which reflects the utilization level of the capacity of a truck. Because the core business of ECC is related to Full Truck Load (FTL) (FTL) transportation, only FTL transport is included in this research. The remaining observations are dropped from the final data.

5.2.16 Transport type

A transport type within ECC can be either be Single-Trip (ST), Round-Trip (RT) or Milk-Run (MR). The difference between the three concepts is explained with the help of Figure 5.15 below. A ST is the simplest form of transport which only involves transport from A to B. A RT differs from a ST, because it returns to the origin (A-B-A). A MR collects multiple orders before directing to the destination. If A and C are orders that should be transported to delivery location B, the transport occurs from A-C-B or from C-A-B.

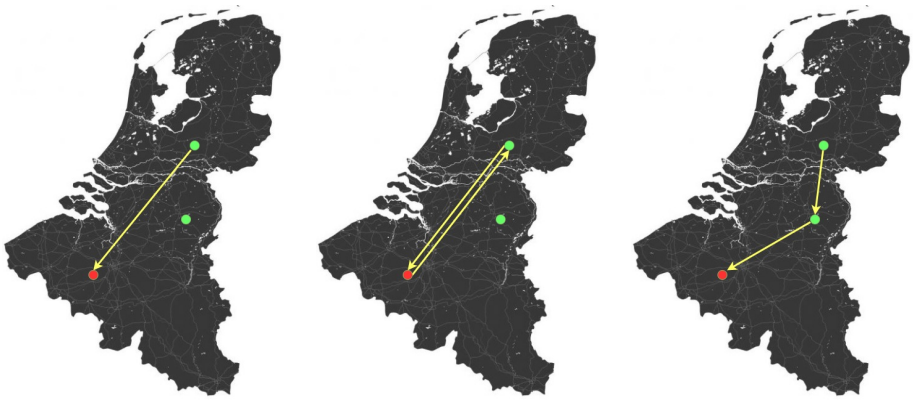

Figure 5.15: Single-Trip (left), Round-Trip (center), Milk-Run (right)

The core business of ECC exists of ST transport. Therefore, only ST transports are considered in this research.

5.2.17 Contract rejection rate

The contract rejection rate originates from the market data, corresponding to pillar 3 in Section 4. As the name suggests, this rate indicates the rate at which contracts get rejected in the perspective of carriers. Therefore, it is a measure of capacity of the market. When the contract rejection rate is high, many transport requests are rejected by carriers, which have two primary reason according to Pahulje (2021). The first reason is because carriers have no capacity left to accept any additional transport requests, whereas the second relates to carriers thinking the price is too low to gain an acceptable margin on the transport. In Figure 5.16 below, th relation between the contract rejection rate and price per kilometer is plotted.

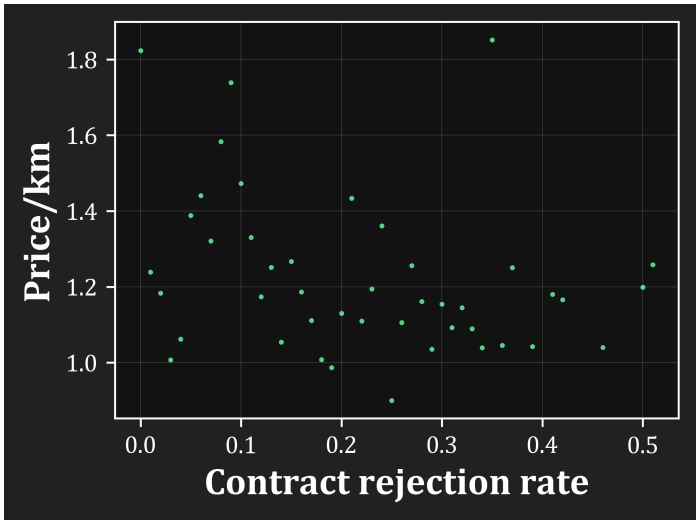

Figure 5.16: The relationship of contract rejection rate versus price per kilometer

Some relationship seems to be present in the plot above. In general, it seems that higher contract rejection rates mean lower prices/km. From Pahulje (2021) it seems that the first reason is not true, because one would expect higher prices when there is not a lot of capacity left for the new contract. However, the second reason could still hold: when clients want too low prices, carriers may also reject the contract and look for other opportunities (for example: transports on the spot-market). When the alternatives are not better in terms of profitability, they still have to conform to the lower prices. Nevertheless, there seems to be a relationship between the two factors and therefore, the contract rejection rate is included in the final data.

5.2.18 Final set of variables

Concluding the analysis of all the variables, Table 5.1 is reduced to Table 5.15 below. This the selection of variables that are included in the final dataset that will be the input for modelling.

Table 5.15: Classification of variables in the RFQ dataset

Variables	**Category**
Lane	Nominal
Number of rounds	Interval
Customer	Nominal
Strategic End-Market	Nominal
Lead-time	Interval
Modality	Nominal
Equipment type	Nominal
Initiative	Nominal
Current business	Ratio
Contract rejection rate	Ratio

5.3 General analysis

As the name suggests, the relation between the independent variables should be independent. In order to test this, the ratio variables are plotted in a Pearson correlation matrix below (Table 5.16).

Table 5.16: Pearson correlation matrix of the variables in the final dataset (***: $p \leq 0.001$, **: $p \leq 0.01$, *: $p \leq 0.05$)

Correlation-matrix	Rounds	Leadtime	Current business	Contract rejection rate	price/km
Rounds	1	-0.06***	-0.07***	0.08***	-0.11***
Leadtime	-0.06***	1	0.07***	0.1***	-0.12***
Current business	-0.07***	0.07***	1	-0.05***	0.25***
Contract rejection rate	0.08***	0.1***	-0.05***	1	-0.25***
Price per kilometer	-0.11***	-0.12***	0.25***	-0.25***	1

Following the correlation guide by Akoglu (2018), only weak correlations are found. The highest correlations are found between independent variables and the dependent variable (price per kilometer). Current business and contract rejection rate appear to have the highest correlations with the price per kilometer variable. In Section 5.2.13, it already looked like current business has a positive linear relationship with the price per kilometer variable, which complies with the positive correlation found in Table 5.16. The absence of moderate and strong relationships in the Pearson correlation matrix comply with the difficulty in collecting relevant data to the prediction of the price per kilometer. At the moment, ECC does not have extensive access to much external information. Moreover, while making considerable steps in data quality, the process is certainly not finished yet.

From Table 5.16, the independent variables do not seem to be highly correlated with each other: there is no collinearity between the independent variables. In order to confirm this, the Variance Inflation Factor (VIF)-scores are computed and can be found in Table 5.17 below. Because this procedure tests the collinearity between independent variables, the dependent variable (price per kilometer) is excluded from this analysis.

Table 5.17: Variance Inflation Factor for the ratio independent variables

	VIF
number of rounds	4.29
Contract rejection rate	3.82
Leadtime	3.74
Current business	1.03

From the table, it becomes clear that all Variance Inflation Factor (VIF)-scores are below a cut-off score of VIF < 5, which is commonly used according to Sheather (2009). Therefore, it can be concluded that there is not collinearity between the independent variables and all the variables can be retained in the final data.

5.4 Conversions

After careful investigation of the variables, the data has to be prepared before it is fed to the models. From the available data, variables can be divided into three groups: nominal, ratio and interval variables. For the model, nominal variables are transformed into dummy variables, where the process is displayed in Table 5.18 below. The classification is shown in Table 5.1 below.

Table 5.18: Transformation nominal variable into dummy variable

Sample	Country_relationship
1	NL-BE
2	DE-IT
3	ES-GB
4	NL-BE

Sample	NL-BE	DE-IT	ES-GB
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0

Models are only able to use numerical data. Therefore, the nominal variables that contain text-values need to be converged to dummy variables. For each unique value related to this variable,

a column is added to the dataset. Where the value equals a particular occurrence of the nominal variable, the value will be filled with a 1. Otherwise, the value in this new column will equal 0. As a result, for each nominal variable in the dataset the number of columns is equal to the number of unique values in that nominal variable.

Ratio- and interval variables also need a conversion. In order to account for the different scales, these variables are standardized, of which the equation is shown in Equation 5.2 below. The dummy variables are not standardized, because re-scaling them will change the definition of these variables. The dummy variables have a clear meaning, indicating whether or not the value occurred in the sample.

$$z = \frac{X - \mu}{\sigma} \tag{5.2}$$

Standardization has the goal of scaling the ratio variables to a common scale, with mean zero ($\mu = 0$) and a variance equal to 1 ($\sigma^2 = 1$). This way, the ability of the model to learn the appropriate weights is improved. While the re-scaled variables now have a common scale making the data internally consistent in content and format, the difference in range of the values is still intact. In the formula, z refers to the standardized value, whereas μ is the population mean and σ the population's standard deviation. X refers to the raw, unconverted sample.

6 Modeling

In the previous chapter, data was prepared and an analysis was performed. After discussion of the set of variables, some were already dropped from the final dataset (distance, date, annual number of shipments, margin, payload, transport type). However, in order to test the importance of the variables quantitatively, another phase is included in this research: the feature-selection phase. In this phase, the final dataset is reduced even more: only the variables which are deemed important by the feature selection model are included in the final models. After the features are selected definitively, the models can be build in the model selection phase. Here, a multitude of models are build which help answer the main research question. Next, each model has a set of hyper-parameters, which impact the performance of the model. This research makes a distinction between a parameter and a Hyper-parameter. Whereas a parameter is internal to the model and is part of the configuration of the model, hyper-parameters are explicitly defined parameters that control the tuning process. Parameters are essential for a model to do a prediction, whereas hyper-parameters are used in order to "tune" (optimize) the model. Configuring the hyper-parameters can be major task, dependent on the amount and the ranges of the hyper-parameters. After selection of the hyper-parameters, the configuration has to be validated. Hence, the next phase is the validation phase. In short, the performance of the model is evaluated against unseen data in this phase. After validation, the best hyper-parameter setting is chosen and the model is trained on almost the entire dataset. A small part of the dataset is excluded in the training process and will be used for evaluation of the model. After training, the model is used to predict the outcome variable (price per kilometer) for samples in the test set. The samples in the test set are used to simulate the real-world: the trained model is used to make predictions on previously unseen data. Because the true values are already known for this dataset, the performance metrics can be easily computed.

The five different phases of modelling are visualized in Figure 6.1 below. The structure of this section is aligned with the different phases, and the phases are elaborated more thoroughly throughout this section.

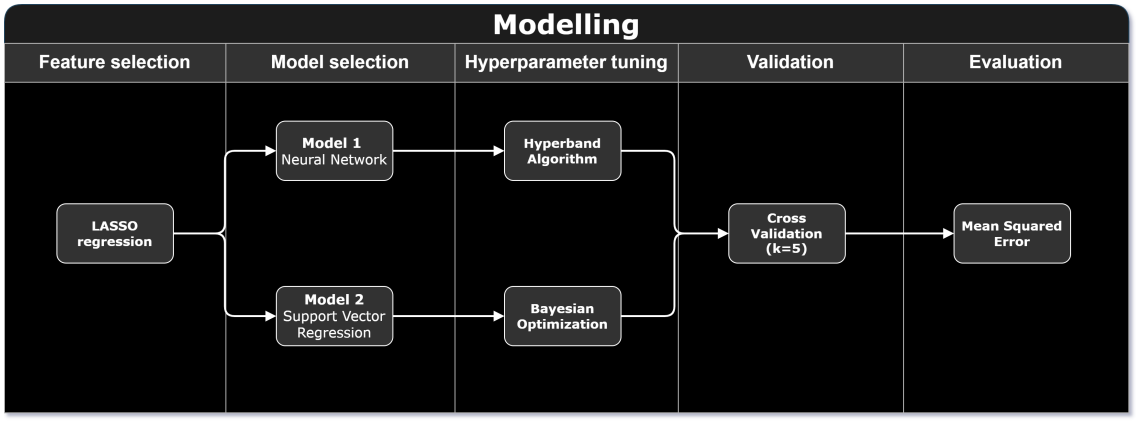

Figure 6.1: A visualization of the modelling phase

6.1 Feature selection

Following from Figure 6.1 above, the first phase corresponds to the feature selection phase. According to J. Li et al. (2017), feature selection has been proven to be effective in preparing data in machine-learning problems. The objective is to build simpler, more understandable models based on clean and understandable data. In order to justify the features selected in this research quantitatively, the Least Absolute Shrinkage & Selection Operator (LASSO) regression method is employed, which proved to be a helpful tool in choosing a model with the most relevant features (Fonti & Belitser, 2017; Muthukrishnan & Rohini, 2016). According to the latter source, traditional feature selection procedures such as (OLS) regression, Stepwise regression and Partial Least Squares regression are

"very sensitive to random errors", which was the motivation behind investigating alternatives. The strength of the algorithm lies in shrinking the weights of irrelevant features to exactly zero, which makes the identification of irrelevant features a straightforward process: when the features with coefficients equal to zero are eliminated from the feature set, we end up with the set of selected features. The objective function of the LASSO-regression to be minimized is shown below in Equation 6.1.

$$\textit{Cost function} = \sum_{i \in I} \underbrace{(y_i - \sum_{j \in J} x_{ij}\beta_j)^2}_{\text{OLS objective}} + \underbrace{\alpha \sum_{j \in J} |\beta_j|}_{\text{l1-penalty}} \tag{6.1}$$

The first part of the equation corresponds to the objective to be minimized by Ordinary Least Squares (OLS) and is nothing more than the Residual Sum of Squares (RSS) (sum of squared differences) between the predicted- and observed values. Here, y refers to the true value of the outcome variable and x to the value of the predictor variable. The β-coefficient refers to the degree of change in the outcome variable for every 1-unit change in the predictor variable. J is the set of variables, whereas I is the set of observations. In addition to the first part, a penalty is added, which is called the *l1-penalty*. To put it simply, a penalty is added for assigning weights to the coefficients. Therefore, the LASSO regression method has a trade-off to make for each variables: minimizing the first part of the objective function (which requires positive coefficients if the observed value is larger than zero) or minimizing the right side of the objective function by shrinking the coefficients to zero. The magnitude of the *l1-penalty* is determined by α, which is a hyper-parameter to be configured in the model. When set to $\alpha = 0$, the objective function is nothing more than the objective function of OLS: minimizing RSS. However, when $\alpha > 0$, shrinkage will occur and the coefficients of the less important variables will be shrunk to zero, thereby making the detection of irrelevant variables an easy and straightforward process.

In order to set α, grid-search is employed. Grid-search is nothing more than testing all possible combinations of values. The process is explained with the help of Figure 6.2 below.

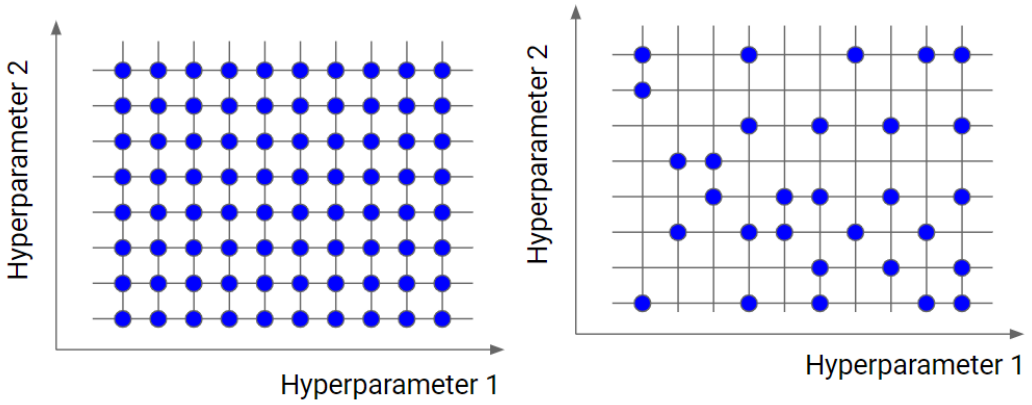

Figure 6.2: A visualization of choosing parameter-values with the help of grid-search (left) and random-search (right)

The parameter combinations at all the intersections of the graph (the blue dots) are chosen for the model. For LASSO-regression, only one hyper-parameter exists (α). While x and y are fixed, β-values are estimated by the model and are therefore no hyper-parameters. Therefore, the grid relevant for feature selection in the LASSO model is one-dimensional. The boundaries for α are set at $[0.001,\ 1]$ in steps of 0.001. This range is chosen after exploration of α in the range between $[0.01, 10]$ in steps of 0.1. The LASSO model with the best model fit is selected as the feature selection model with α_{best}. In order to determine the best model fit, k-fold Cross-Validation (CV) is implemented, which in essence scores the parameter configuration k time on a different part $(\frac{1}{k})$ of

the dataset. The scoring is determined by the coefficient of determination (R^2), which is the default scoring metric for LASSO models. The coefficient represents the proportion of the total variation of outcomes explained by the model. The equation is shown in Equation 6.2 below.

$$R^2 = 1 - \frac{\sum_{i \in I}(y_i - \hat{y})^2}{\sum_{i \in I}(y_i - \overline{y})^2} = 1 - \frac{SS_{res}}{SS_{tot}} \tag{6.2}$$

In this formula, y and I should be clear following the same terminology as in Equation 6.1 above. $\hat{y}$ refers to the predicted value, whereas $\overline{y}$ refers to the average observed value: $\overline{y} = \frac{1}{n}\sum_{i \in I} y_i$. SS_{res} and SS_{tot} refer to the sum of squared residuals ($SS_{res} = \sum_{i \in I}(y_i - \hat{y}_i)^2$) and the total sum of squares ($SS_{tot} = \sum_{i \in I}(y_i - \overline{y}_i)^2$) respectively. In the best case, when the predicted value is exactly equal to the observed value $SS_{res} = 0$ and therefore $R^2 = 1$.

A visualization of the k-fold CV process can be found in Figure 6.9, which will be clarified later in this section. For the implementation of the LASSO algorithm, the steps treated in Malato (2021) are applied, which are visualized in Figure 6.3 below.

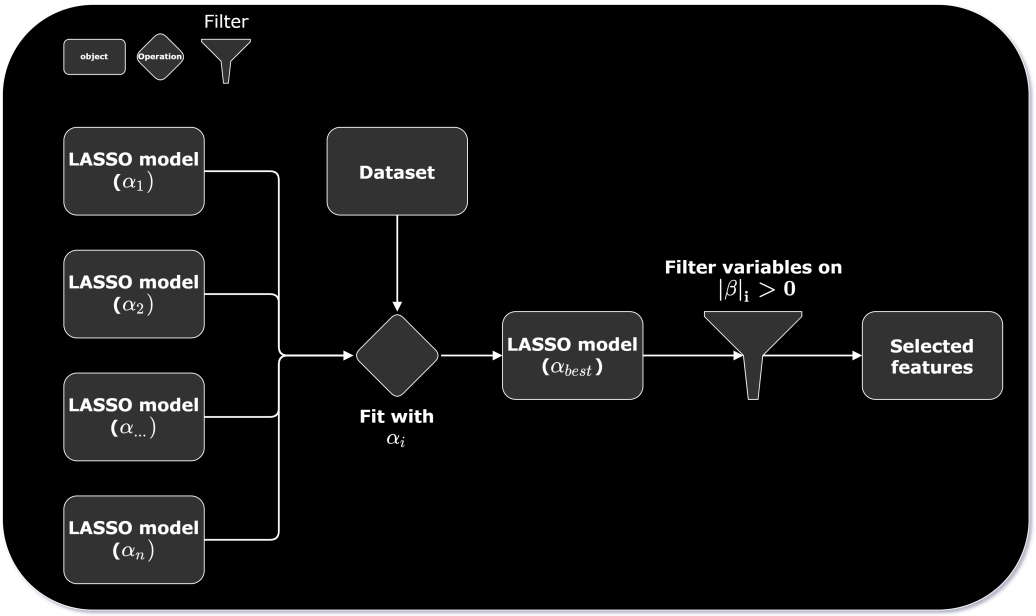

Figure 6.3: Features selection process

Based on this grid-search, n LASSO models are created. After fitting the data to the models, the results are displayed in Table 6.1 below. According to the table, the model with $\alpha_{best} = 0.001$ is chosen as the feature selection model based on CV with R^2 as the scoring metric.

Table 6.1: Results of the top-3 LASSO models

Rank	α	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean R^2	Standard deviation
1	0.001	0.4	0.68	0.73	0.63	0.58	0.6	0.11
2	0.002	0.37	0.63	0.68	0.54	0.55	0.55	0.1
3	0.003	0.36	0.56	0.64	0.45	0.53	0.51	0.09

After analysis of the coefficients β-coefficients, the features with $|\beta_i > 0|$ are selected for modelling. With respect to the LASSO model with $\alpha = 0.001$, this implies that from the 543 variables (after creating dummy variables following the procedure discussed in Section 5.4), only 132 variables should be selected. From this selection, all non-nominal variables (Table 5.1) are included. This means, that only dummy columns are deemed unimportant by the model. However, this is a known limitation

of the LASSO model, which tends to select one variable from a group, ignoring the others (Fonti & Belitser, 2017). A group in this case is the set of dummy variables measuring the same nominal variable (see Table 5.18). Because of this, it is decided to include the entire group of variables when the coefficients of one of the variables has not shrunk to zero. Following this logic, all the grouped variables are deemed important by the LASSO feature selection technique. Therefore, all the variables are included in the models, which will be next topic.

6.2 Model selection

The second step in the modelling phase is to create the models that should test the main research question (*"What kind of Machine-Learning algorithm can be developed to set competitive prices in the RFQ process at Ewals Cargo Care?"*). Two models are chosen for this research, following from the literature review treated in section 2. The first model is a Artificial Neural Network (ANN) and the second model under investigation is a Support Vector Regression (SVR). In this section, the selected models are explained and their hyper-parameters are discussed. Configuration of the hyper-parameters is discussed in the next phase: hyper-parameter tuning.

6.2.1 Model 1: Multilayer Perceptron (MLP)

The first model to be tested is an Artificial Neural Network (ANN). More specifically, a fully connected class of feed-forward ANN. The most common types of ANN are feed-forward neural networks and recurrent neural networks. In feed-forward neural networks, signals travel in one direction only. These type of networks are widely used to model relationships between a set of predictor variables and outcome variables. In recurrent neural networks, signals may travel bidirectionally. This makes it possible to use its internal state (memory) to process variable-length inputs in sequential data. From these general definitions, the feed-forward neural network is most applicable to this research problem, where an attempt is made to model the relationship between a set of predictor variables and the price per kilometer (outcome variable). For feed-forward neural networks, the most widely studied network is the Multilayer Perceptron (MLP), which simply refers to a feed-forward neural network with at least three layers containing nodes: an input layer, a hidden layer and an output layer. The input layer catches the features of the dataset in a layer, whereas the hidden layer performs operations to the previous layer. The output layer combines all of the operations done previously into one last layer, consisting of one node in this research (referring to the one outcome variable: price per kilometer). The more layers and neurons the network has, the higher the complexity of the model. Complexity may decrease its interpretability and generalizing capability, which is something that should be prevented as much as possible. When a model is not able to generalize well, the model is likely overfitted. This concept means that while the training loss is reduced as much as possible, the model is still not able to perform well on unseen data, which corresponds with a high validation loss. A MLP model is visualized in Figure 6.4 below.

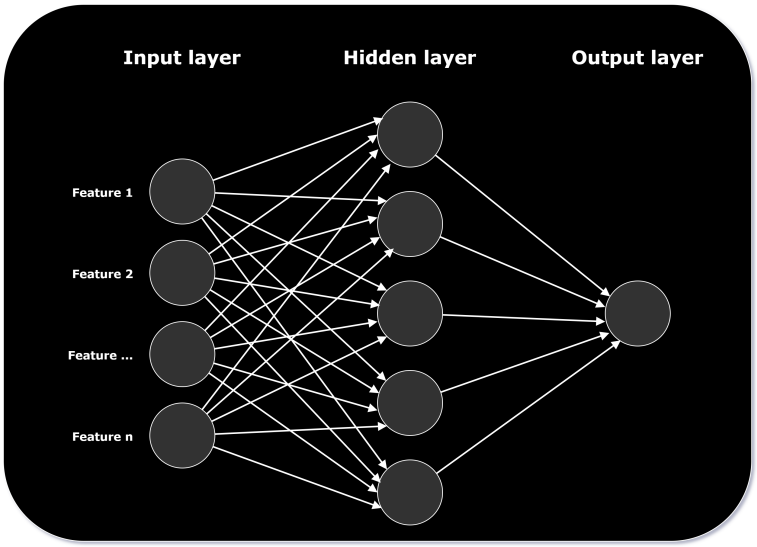

Figure 6.4: An abstraction of a Multilayer Perceptron (MLP)

Apart from the input nodes in the input layer, nodes refer to neurons using a non-linear activation function. This function transforms the input that is fed to the neurons (non-linearly). While the most common activation functions are relatively simple non-linear functions, the model is capable of detecting non-linear relationships between the predictor and outcome variables. Moreover, MLP uses backpropagation in order to learn the weights of the graphs, where graphs are equivalent to the connections between nodes. Backpropagation is a supervised learning technique which computes the gradient with respect to each weight individually. Therefore, it is able to update the individual weights towards their local optimum in order to minimize the objective (or the loss-function).
The MLP chosen for this research consists of multiple parameters. An overview is given in Table 6.2 below.

Table 6.2: An overview of the parameters in a MLP model

MLP parameters			
Parameter	**Name**	**Description**	**Range**
b	Batch-size	The number of samples after which updating of the weights occurs	$choice[16, 32, 64, 128]$
e	Epochs	The number of times the entire dataset is seen by the model during training	100
l	Layers	The number of hidden layers in the neural network	$int[1, 10]$
$n_i,\ i \in \{1, 2, ..., l\}$	Neurons	The number of neurons in a layer	$int[1, 20]$
a_l	Activation-function	The activation function used in a specific layer	$choice[\text{ReLu}, \text{Sigmoid}]$
lr	Learning-rate	The rate at which learning occurs	$choice[1e^{-4}.1e^{-3}, 1e^{-2}, 1e^{-1}]$
d	Dropout-rate	The fraction of neurons being ignored in all layers	0.5
p	Early-stopping patience	The amount of epochs during which no improvement of the validation loss-function is observed, before terminating the training procedure	4

The amount of neurons and activation-functions are set for each layer individually. It is possible that in a network with two hidden layers 5 neurons are chosen in the first hidden layer with a *sigmoid* activation function, while in the second hidden layer, 3 neurons are chosen with a *ReLu* activation function (these functions are explained later). The learning rate specifies the magnitude of weight updates after evaluation of the batch. If the learning-rate is high, the weights are changed dramatically when the error of the batch is high. When set too low, it takes a long time for the model to reduce the loss.

It is possible that the model overfits on the data fed to the model. The concept of overfitting is explained with the help of Figure 6.5 below.

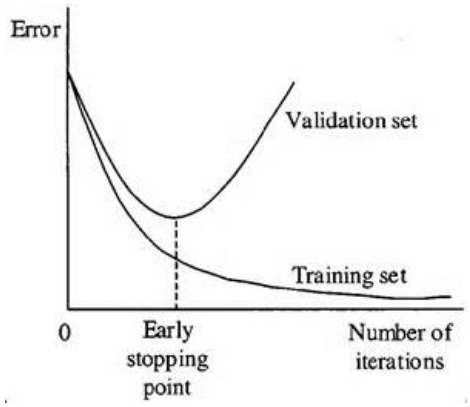

Figure 6.5: The concept of overfitting and early-stopping

Overfitting occurs when the data is trained on a dataset, such that the training loss is still being reduced while the performance on previously unseen data is decreased. In the graph this corresponds to the dotted line. Early-stopping attempts to the detect this point by keeping track of the number of epochs the validation loss did not improve, denoted with a patience p. When during p epochs, no improvement in validation loss is detected, training is terminated. Another procedure to tackle overfitting is drop-out regularization. This procedure attempts to prevent making complex connections by excluding a fraction d of the inputs. Thus, this fraction of neurons is left out in the computation of the weights.

6.2.2 Model 2: Support Vector Regression (SVR)

The next model under investigation is a Support Vector Regression (SVR), which is the regression version of the well known Support Vector Machine (SVM) for classification problems. In order to help understand the concept of the model, a visualization is made which can be found in Figure 6.6 below.

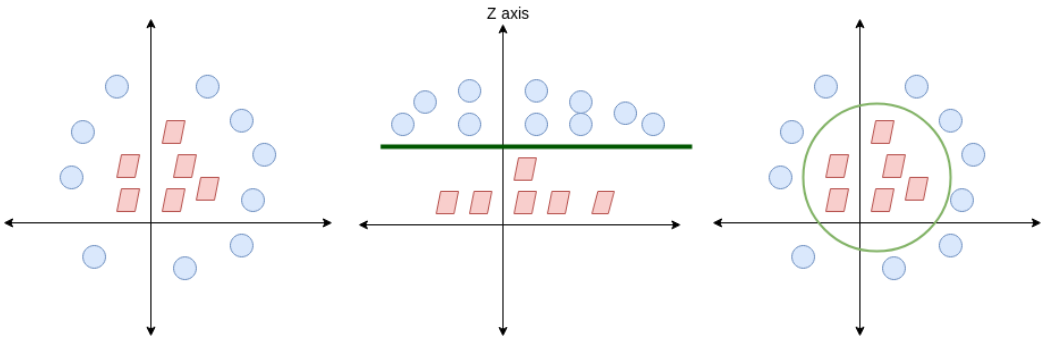

Figure 6.6: Class separation by the SVM model

In the picture, the goal of a SVR is to separate the blue and the red classes from each other by drawing a line. It is not possible to draw a straight line, such that both classes can be separated from each other. Therefore, another dimension (the z-axis) is added to the graphs (middle). Now, it is possible to draw the green line, which separates the red and blue classes from each other perfectly. If the dimensions are reduced to the dimensions from the left figure, A circle can be observed which correctly classifies the samples belonging to the red class. However, when the data is becomes more complex, the samples are not easily classifiable. Consider the left picture in Figure 6.7 below.

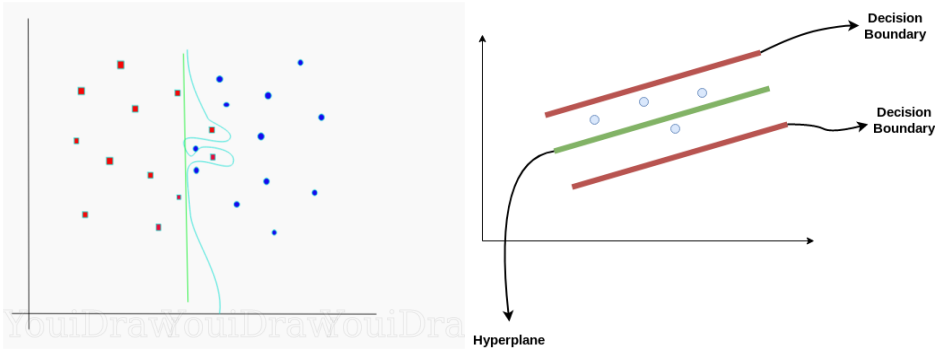

Figure 6.7: Classifying more complex samples in a SVM (left) and a Support Vector Regression (right)

Here, a trade-off has to be made between a smooth decision boundary and classifying all the samples correctly.

In a SVR model, a similar principal holds. In SVR, the model tries to find the best fit line (the so-called hyperplane), such that the samples that fall within the decision boundaries is maximized. However, the distance to the decision boundaries is penalized and therefore, the model will trade-off between correctly predicting the samples and increasing the error margin. This concept is explained by the right picture in Figure 6.7 above. The hyperplane and decision boundaries are set such that the a trade-off between the number of points that fall within the decision boundaries and the error margin is optimized.

The parameters of a SVR model are displayed in Table 6.3 below.

Table 6.3: An overview of all the hyper-parameters in the SVR model

SVR parameters			
Parameter	**Name**	**Description**	**Range**
k	Kernel	A kernel helps finding the hyper-plane in a higher dimension space	*choice* [Linear, Polynomial, Radial-basis function]
ϵ	Epsilon	The distance between the hyper-plane and the decision boundaries	$[0, 1]$
C	Regularization	Regularization parameter whereas the value is inversely proportional to the regularization factor	$[1e^{-5}, 1]$

The most important parameter is the kernel-function, which will help finding the right hyperplane. ϵ defines the margin tolerance where no penalty is given to errors and corresponds to the distance between the green- and the red line in Figure 6.7 (right). The regularization parameter (C) specifies the amount of regularization, which is equal to the squared *l2-penalty*. Regularization adds a penalty to the cost-function in order to account for model complexity. Unlike the *l1-penalty* (discussed earlier in this chapter) which equaled the absolute value of magnitude of coefficients ($\sum_{j=1}^{p} |\beta_j|$), the *l2-penalty* equals the squared value of magnitude of coefficients ($\sum_{j=1}^{p} \beta_j^2$).

6.3 Hyper-parameter Tuning

For each of the models that are chosen, the hyper-parameters displayed in Table 6.2 and Table 6.3 have to be specified. The process of optimizing the parameters for the models is called hyper-parameter tuning. According to Snoek, Larochelle, and Adams (2012), there is a high need for automatic approaches to tune the hyper-parameters of models, as the process requires expert experience, rules of thumb, or sometimes brute-force search. Several techniques exist in order to help choose the right configuration of parameters. In order to optimize the hyper-parameter settings of the models, the following hyper-parameter-setting techniques are considered: grid-search, random-search, Bayesian Optimization (BO) and the Hyperband-algorithm.

Grid-search is the technique employed in the LASSO feature selection model earlier in this section in order to optimize α. While grid-search is still widely employed in order to optimize the model parameters, it seems strange with so many algorithms to choose from (Bergstra & Bengio, 2012). However, the implementation of grid-search is simple relative to the more sophisticated algorithms and the code infrastructure is not that extensive.

For random-search, the same principle applies and implementation is rather similar to grid-search in terms of code structure and difficulty. The technique is shown in Figure 6.2 (right). Similar to grid-search, a range of the values the parameters can assume is pre-specified. Then, a random selection of parameter combination is chosen, as long as the parameters fall within the specified ranges. In terms of performance, random-search seems to outperform the grid-search strategy (Bergstra, Bardenet, Bengio, & Kégl, 2011; Bergstra & Bengio, 2012). While grid-search and random-search are exhaustive optimization techniques, more sophisticated algorithms attempt to optimize the parameters more intelligently.

Bergstra and Bengio (2012) suggest exploration of more sophisticated techniques, particularly the Bayesian Optimization algorithm. Although being more difficult to implement, Snoek et al. (2012) demonstrated that implementation of the Bayesian Optimization methodology outperformed the other employed methodologies and surpassed the performance of human-experts selecting the hyper-parameters. Bayesian Optimization (Snoek et al., 2012) attempts to use all of the information from previous computations to make an informed decision on what parameters to choose for the next trial (the next hyper-parameter configuration). For this, it constructs a probabilistic model and its strength lies in finding the minimum of difficult non-convex functions with relatively few trials at the cost of taking more computational time in deciding what parameters to evaluate next. However, when evaluations of parameters are expensive to perform in terms of resources, it may be beneficial to take some extra time choosing the next set of parameters, in order to reduce the total amount of trials. The technique makes use of a prior, which captures believes about the behavior of a function. The prior chosen for this process is the Gaussian Process prior, which is known for its flexibility and tractability.

Lastly, the Hyperband algorithm proposed by L. Li, Jamieson, DeSalvo, Rostamizadeh, and Talwalkar (2017) assumes the tuning process is *"a pure-explorative, non-stochastic infinite-armed bandit problem where a predefined resource like iterations, data samples, or features is allocated to randomly sampled configurations"*. The Hyperband algorithm (L. Li et al., 2017) is an extension of the Successive Halving algorithm proposed by Jamieson and Talwalkar (2016). The idea behind this algorithm is to uniformly allocate a budget to a set of hyper-parameter configurations. After evaluation, the worst half is thrown away and the procedure is repeated until one configuration is left. The most relevant problem for this algorithm is the trade-off between budget and resources: a large amount of configurations can be evaluated with small average training times or few configurations can be trained for a longer time. The Hyperband algorithm addresses this problem by grid-searching the number of configurations with a fixed budget. The outer loop varies between the number of configurations and the amount of resources, whereas the inner-loop, the bracket, has a fixed number of configurations and amount of resources. In essence, the algorithm adapts resource allocation and

early-stopping specifically for deep-learning problems. With resource allocation, the algorithms focuses on the more promising hyper-parameter configurations, while early-stopping stops the training of models early when the hyper-parameter configuration is not promising. In terms of performance, the Hyperband algorithm showed an order of magnitude improvement in comparison to the Bayesian Optimization technique.

Following the evolution of the literature above, the Hyperband algorithm is chosen for the MLP model. Because the Hyperband algorithm is unavailable to the SVR model (it is solely available for deep learning problems), the Bayesian Optimization methodology is selected for the SVR model. This is also visible in Figure 6.1. Implementation is realized with the help of O'Malley et al. (2019). They managed to develop a library which enables implementation of both algorithms, while being able to manually adjust the properties. For example, by default, implementation of cross-validation is not incorporated in the algorithms. The reason for this is the extensive amount of datatypes that should be handled by the library. However, by subclassing the Tuner- and Hypermodel classes, implementation of k-fold CV can be realized. An example of adding CV into the Bayesian Optimization Tuner-class is shown in the code in Section 11.4. The Tuner-class takes a Hypermodel as input, which is simply a function which creates and returns a model with hyper-parameters (that are to be optimized).

After explaining the hyper-parameter tuning techniques, the choices of the techniques have to be specified. For all the models, the choices are displayed in Table 6.4 below.

Table 6.4: hyper-parameters ranges of the models

MLP hyper-parameters		
Parameter	**Name**	**Range**
b	Batch-size	$choice[16, 32, 64, 128]$
e	Epochs	100
l	Layers	$int[1, 2, ..., 10]$
$n_l,\ i \in \{1, 2, ..., l\}$	Neurons	$int[1, 2, ..., 20]$
a_l	Activation-function	$choice[\text{ReLu}, \text{Sigmoid}]$
lr	Learning-rate	$choice[1e^{-4}.1e^{-3}, 1e^{-2}, 1e^{-1}]$
d	Dropout-rate	0.5
p	Earl-stopping patience	4

SVR hyper-parameters		
Parameter	**Name**	**Range**
k	Kernel	$choice[\text{Linear}, \text{Polynomial}, \text{Radial-basis function}]$
ϵ	Epsilon	$[0, 1]$
C	Regularization	$[1e^{-5}, 1]$

In this table, integer values are only allowed when *int* is specified. When a parameter range is denoted with *choice*, the values that follow are the only values that can be assumed. When a range ([ ]) is denoted without a specification, any value between the numbers can be assumed. Lastly, when a value is denoted without a range, the parameter value is fixed.

With respect to the MLP model, in Bengio (2012) the range of the batch-size is chosen between 1 and a few hundreds, whereas often a batch-size of 32 is a performing value. Because there is some discussion about whether or not to tune the batch-size, it is decided to only include a sub-selection of the suggested range of the hyper-parameter, including a batch-size of 32. Batch-sizes are typically chosen in a power of 2, because of computational reasons related to the processors. The maximum number of epochs is set at 100. By trial and error it was found that in most instances, this amount of epochs was not even reached because of the early-stopping rule ($p = 4$). With this amount of patience, overfitting the model on the training data is prevented and the speed of the training process

is improved. The amount of neurons and layers capped at 20 and 10 respectively, in order to limit the complexity of the model. Only two activation functions are included in this research, which are part of the most common/popular activation functions for deep learning problems (H. Zhang, Weng, Chen, Hsieh, & Daniel, 2018). These activation functions are explained with the help of Baheti (2022) and Figure 6.8 below. The equations of the functions can also be found in the figure.

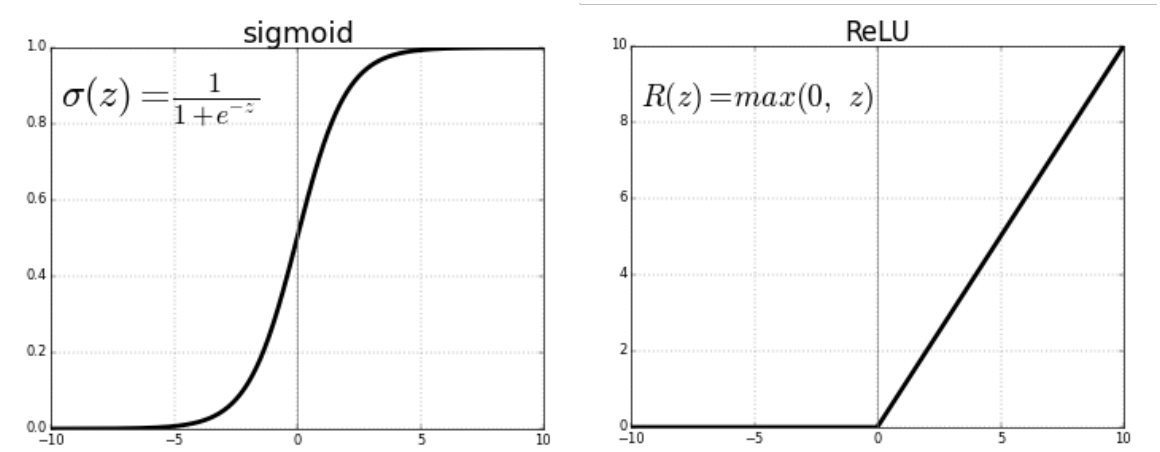

Figure 6.8: The ReLu and Sigmoid activation function

On the left, the sigmoid function can be found. This functions takes any value as input and transforms the value in a value within a range of 0 and 1. The larger the value, the closer the output will be to 1, whereas the lower the value, the closer the output will be to 0. A benefit of the sigmoid function is that the function is differentiable and provides a smooth gradient. In other words, the function prevents "jumps" in the output value. On the right, the Rectified Linear Unit (ReLu) activation function can be found. ReLu also has a derivative and allows for backpropagation. Benefits of the ReLu are that only a number of neurons are activated, which makes it computationally more efficient than the sigmoid function. Additionally, ReLu is relatively fast in finding the global minimum of the loss function. Next, for the learning rate, it is common practise to choose a value on a logarithmic scale (Goodfellow, Bengio, & Courville, 2016), whereas for dropout-rate, a value of $d = 0.5$ is a common but performing choice in deep learning problems and is used in the popular Imagenet classification problem treated in Krizhevsky et al. (2012).

Related to the SVR model, all available kernels in the library by Pedregosa et al. (2011) are used in the hyper-parameter tuning process apart from the *sigmoid* kernel, which showed computational difficulties in the implementation. For ϵ, a range from zero to one is employed, based on the research by Smets, Verdonk, and Jordaan (2007), but including 0 as a possible choice. For the regularization parameter C, multiple values are tested in order to find the optimal amount of regularization, preventing overfitting on the training data.

6.4 Validation

After selecting the parameters with the help of the chosen algorithm, the performance of the configuration has to be validated. The procedure chosen for validating the hyper-parameters is Cross-Validation (CV), being the simplest and most widely used method for estimating prediction error (Hastie, Tibshirani, Friedman, & Friedman, 2009). The type of CV chosen is K-fold CV, which implies there is no need to holdout one part of the dataset for validation, which cannot be used for training. The concept is visualized in Figure 6.9 below.

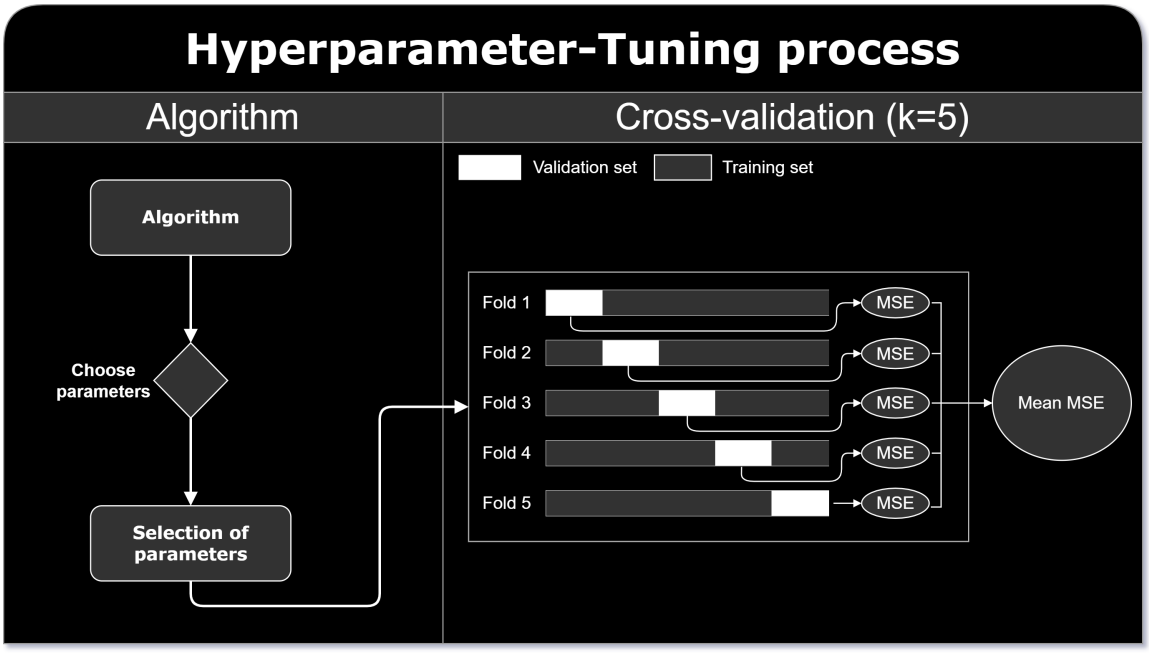

Figure 6.9: The hyper-parameter tuning and validation process

First, the hyper-parameter tuning algorithm chooses the set of hyper-parameters. Then, the dataset is split into k (the amount of folds) parts and the model will also be trained k times, each time using $k-1$ part of the dataset (gray) as training set and $\frac{1}{k}$ part of the dataset (white) as validation set. Each time the hyper-parameter configuration is trained, it uses a different $\frac{1}{k}$ part of the dataset as the validation-set. The most common values for k are five or ten. Therefore, in this research the amount of folds is set to $k = 5$.

Each model has a loss-function, which measures the performance of the model. As performance metric in the validation phase, the Mean Squared Error (MSE) is selected, which is a common performance metric for regression problems. According to Wallach and Goffinet (1989), MSE is a reasonable criterion of model quality when the main purpose of the model is prediction. The formula can be found in Equation 6.3 below, where y represents the observed value, whereas $\hat{y}$ represents the predicted value. The metric is shown in Equation 6.3 below.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{6.3}$$

For each fold, the MSE will be computed. After training and validating the hyper-parameter-setting for all the k folds, the average MSE is calculated which represents the performance of the hyper-parameter-setting. The mathematical formula is shown in Equation 6.4 below (where hp represents

the hyper-parameter configuration).

$$MSE_{hp} = \frac{1}{k} \sum_{i=1}^{k} MSE_{hp}^{i} \tag{6.4}$$

The results of the hyper-parameter tuning process are displayed in Table 6.5 below. The results are derived within a pre-defined resource budget. The output in the table below is the result from a hyper-parameter tuning process with a maximum resource-budget of 24 hours.

Table 6.5: Hyper-parameter tuning results

MLP											
Rank	**MSE**	**l**	$\mathbf{n_1}$	$\mathbf{a_1}$	$\mathbf{n_2}$	$\mathbf{a_2}$	$\mathbf{n_3}$	$\mathbf{a_3}$	**lr**	**b**	**e**
1	0.031	3	19	sigmoid	9	relu	19	sigmoid	0.01	64	34
2	0.032	3	17	relu	16	sigmoid	16	sigmoid	0.001	16	34
3	0.032	3	17	relu	16	sigmoid	16	sigmoid	0.001	16	100

SVR				
Rank	**MSE**	**k**	ϵ	**C**
1	0.022	rbf	0.0	1.0
2	0.022	rbf	0.0	1.0
3	0.022	rbf	0.0	1.0

From the results, it appears that the Hyperband algorithm decided that three layers were the optimal amount of layers within the resource budget for the whole top-three MLP models. With respect to the amount of neurons, no clear pattern is recognized. Apart from the best model, the amount of neurons per layer appear to be consistent in the model. While the activation function for layer 1 and layer 2 is not evident, the top-three agrees on using a sigmoid activation function for layer 3. The top two models do not need the maximum epochs of $e = 100$, which means the training is stopped preliminary because the validation loss did not decrease anymore during four epochs (remember that $p = 4$ from Table 6.4).

6.5 Evaluation

From the validation phase, the best hyper-parameter configurations became clear. However, the results of the validation phase are not definite yet. Remember that in the hyper-parameter tuning phase, a training set and validation set were used in order to tune the hyper-parameters. The models were trained on the training set, while the validation set (which was different in each fold) was employed in order to validate the hyper-parameter configuration. In the evaluation phase, the model is evaluated against another part of the dataset: the test set. How this set is different from the other sets becomes clear with the help of Figure 6.10 below.

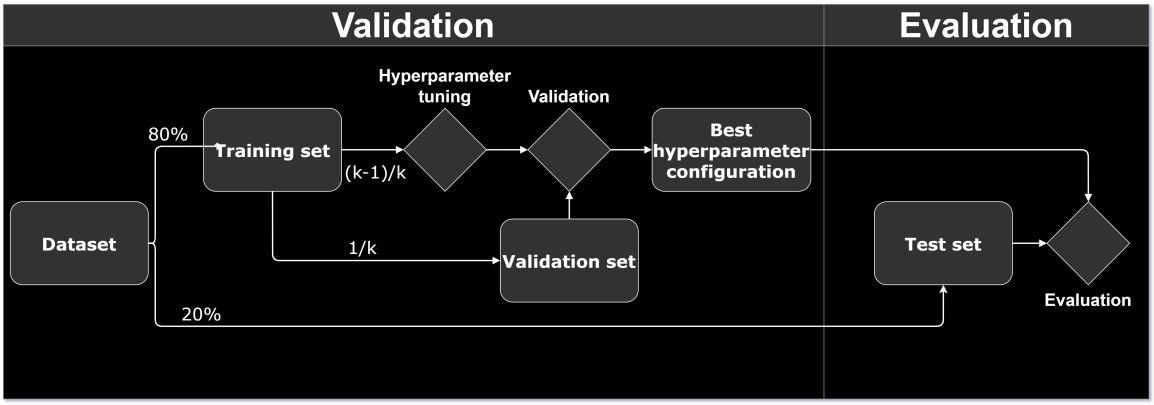

Figure 6.10: The test set explained

Between the feature selection phase and the hyper-parameter tuning phase, the test set is subdivided from the training set. In this research, the fraction of the dataset that represents the test set equals 20% of the entire dataset. Unlike the training set which is used in the hyper-parameter tuning procedure, the test set is kept separate and is relevant for the evaluation phase only. This is important because the test set is used in order to simulate new, previously unseen data. This is necessary because we are interested in predicting a price per kilometer when a new RFQ arrives at the RFQ-desk.

Following from the hyper-parameter tuning process, the best configurations for both models are now fitted on the entire training set (the full 80%). This means, the models trained in the hyper-parameter tuning phase can be discarded and only the hyper-parameter configurations of the best models are remembered. With these configurations, the models are trained on the full training set, after which they are evaluated on the test set. As test-metrics, Mean Absolute Error (MAE), Mean Squared Error (MSE) and R^2 are chosen. The MSE was explained in the validation phase already and was proven to be a reasonable criterion of model quality when the main purpose of the model is prediction Wallach and Goffinet (1989). The MAE criterion is included to allow for easy interpretation of the results and is simply the absolute difference between the prediction and the observed value (see Equation 6.5 below).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6.5}$$

The next criterion, the R^2, is a well established metric in classical regression analysis (Nagelkerke et al., 1991; Rao, Rao, Statistiker, Rao, & Rao, 1973). The metric represents the proportion of variance explained by the regression model and is therefore a useful metric in predicting the outcome variable from a set of predictor variables. The formula can be found in Equation 6.2, which was already discussed in the feature selection phase.

For all the models with their best hyperparameters configuration, the performance on the test set is displayed in Table 6.6 below.

Table 6.6: Final results of all investigated models

Rank	Model	MAE	MSE	R^2
1	SVR	0.032	0.019	0.888
2	MLP	0.083	0.028	0.828

It appears that the SVR model is the best model with respect to predicting the price per kilometer variable. The prediction error, measured by the MSE and MAE is lowest, while the variance explained (R^2) is highest. Based on the MAE, the mean error with respect to the predictions is only 0.032€, implying the predictions are actually considerably accurate. Nevertheless, both models

have low prediction errors, while having a substantial amount of variance explained.

With respect to the MLP model, intermediary results are saved while training the model. As mentioned before, the amount of times the entire training set is seen by the model are called *epochs*. At every epoch, the training loss and validation loss are reported. From this report, Figure 6.11 can be drawn.

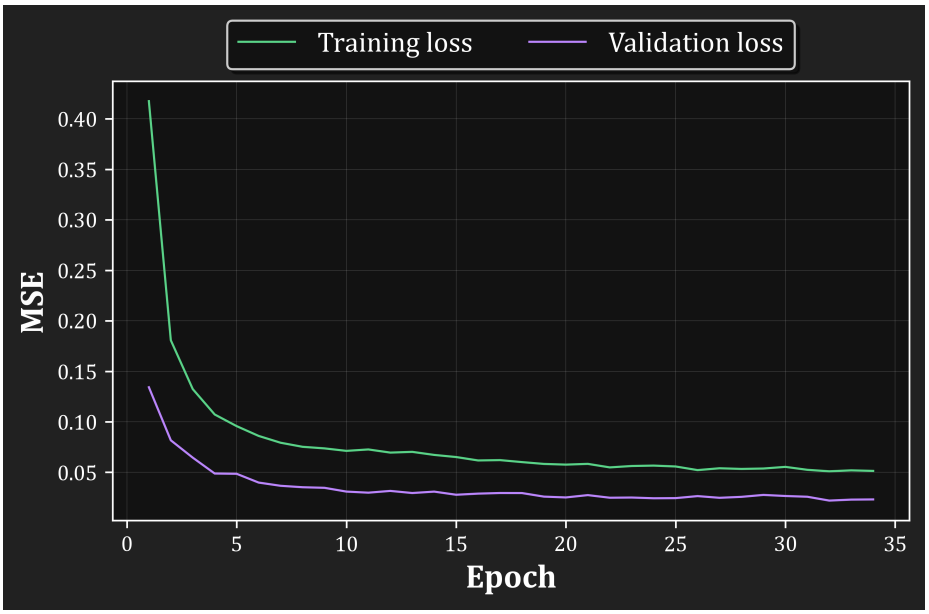

Figure 6.11: Training loss verses validation loss

In this graph, the effect of early-stopping is clearly visible. The validation loss seems to be close to a minimum, while no increase is in validation loss is visible. From Figure 6.5, it became clear that overfitting occurs when the validation loss starts increasing. When only the left side of the early-stopping point in Figure 6.5 is inspected, a similar graph can be observed in comparison to Figure 6.11. The only major difference is that the validation loss in the results is lower than the training loss, which seems strange. Apparently, the model performs better on previously unseen data, than the training data itself. This can be explained by regularization (the strategy to tackle overfitting). Regularization methods have the objective of sacrificing performance on the training set in order to improve the performance on the validation set. In the MLP model, drop-out regularization is applied. This type of regularization is only applied during training and not during validation, which also contributes to the validation loss being lower than the training loss.

7 Results

The input for this section is determined by the output of Section 6. Here, it became clear that the Support Vector Regression (SVR) model is the most performing model for the prediction of prices in a Request For Quotation (RFQ). In this section, the performance of the model is evaluated by a deep dive in the data.

7.1 Sensitivity analysis

In order to evaluate the sensitivity of the model by its underlying variables, a default scenario is created. The nominal- and interval variables (Table 5.15) are chosen based on their most prevalent observation, whereas the ratio variables are set based on their averages. The default scenario is shown in Table 7.1 below. This default scenario is used to evaluate the variables from the model individually (ceteris paribus).

Table 7.1: Default scenario for the sensitivity analysis (actual customers are confidential)

Lane	Number of rounds	Customer	Strategic End-market	Lead-time	Modality	Equipment	Initiative	Current business	Contract rejection rate
DE-GB	1	Customer X	Paperpackaging	2.4	ROAD	Standard tautliner	group	0,05	0,14

With respect to the nominal- and interval parameters, all observed and unique values in the training data are included in the analysis. The assumed values for the ratio variables are chosen by choosing a range from the minimum- to the maximum observed value in the training data with an interval equal to 0.01 €. The result is shown in Figure 7.1 below.

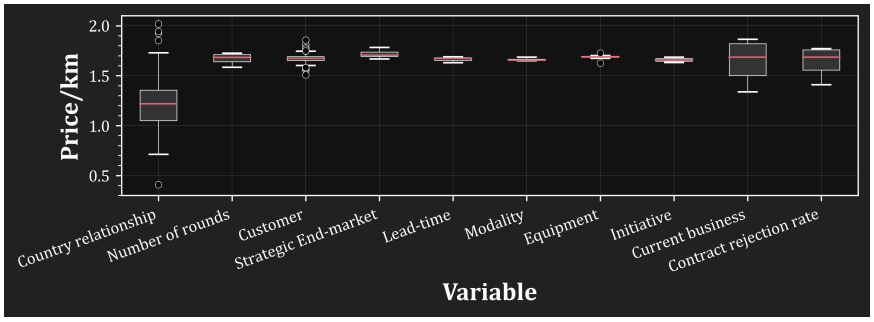

Figure 7.1: A visualization of the sensitivity analysis, from all variables combined

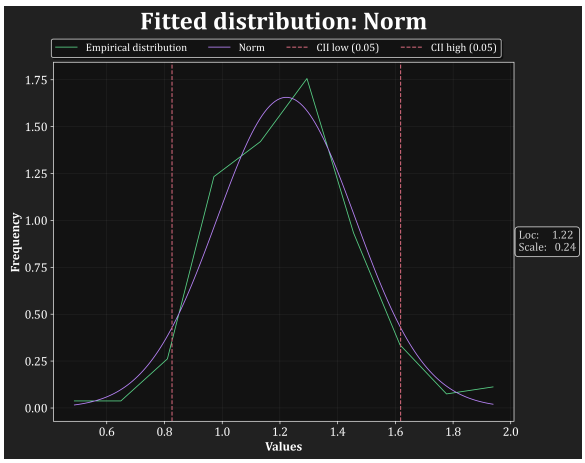

Figure 7.2: A normal distribution fitted to the predictions of all unique country relationship

Based on the figure, it appears that the country relationship, current business and contract rejection rate affect the predicted price in this default scenario the most, as the price-range is most divergent for these variables. In particular, the country relationship strongly affects the prediction. The 166 country relationships under investigation appear to be normally distributed ($\mu = 1.22, \sigma = 0.24$) in terms of the predicted price according to the procedure described in Section 11.3 (see Figure 7.2 on the right).

The direction of current business and contract rejection rate appear to be the same: the higher the value, the lower the price (see Figure 7.3 below).

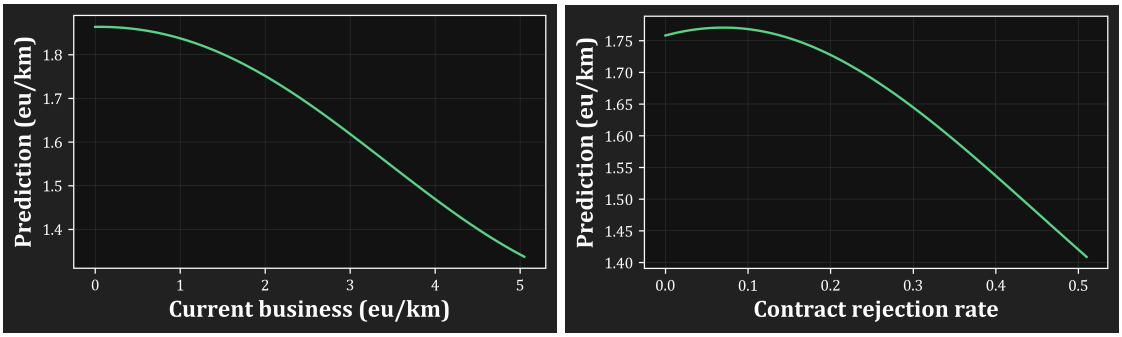

Figure 7.3: The effects of current business (left) and contract rejection rate (right) on the predicted price per kilometer

The curve of the contract rejection rate is likely explained by the price-levels: when the price for transport is low, there is a higher probability of the contract being rejected by the transportation company. The trucks and trailers are likely allocated to business that is more favorable in terms of prices. The curve related to the current business is interesting as well. The models predict lower prices when the price of the current business is higher. It appears that in order to maintain business from a particular customer, lower prices should be offered in a next RFQ.

7.2 Predictions results

7.2.1 Prediction interval

With respect to the predictions, a prediction interval $(l, u])$ can be constructed by Equation 7.1 below. This prediction interval is based on the assumption that the sample is from a normal distribution.

$$\begin{aligned} l &= \hat{y} - z \cdot \sigma \\ u &= \hat{y} + z \cdot \sigma \end{aligned} \tag{7.1}$$

Here, z refers to the standard score, where choosing $z = 1.96$ results in a 95% prediction interval. There are various way to measure the standard deviation σ of a model prediction $\hat{y}$, where the Root Mean Squared Error (RMSE) is a popular one according to Shmueli, Bruce, Yahav, Patel, and Lichtendahl Jr (2017). The RMSE is simply equal to $RMSE = \sqrt{MSE}$ (see Equation 6.3 for the MSE-equation). Therefore, the prediction intervals are set at $[l, u] = [\hat{y} - 1.96 \cdot RMSE,\ \ \hat{y} + 1.96 \cdot RMSE\,]$ in this research. From all observations, more than 98% of the true price per kilometer fell within these boundaries when the RMSE is set equal to $RMSE = 0.138$, which followed from Table 6.6.

7.2.2 Analysis

Investigating the predictions of the model in greater detail, lead to the interest in studying the distribution of the prices across Europe. For this purpose, Figure 7.4 below is created, which shows the prices for both origin and destination.

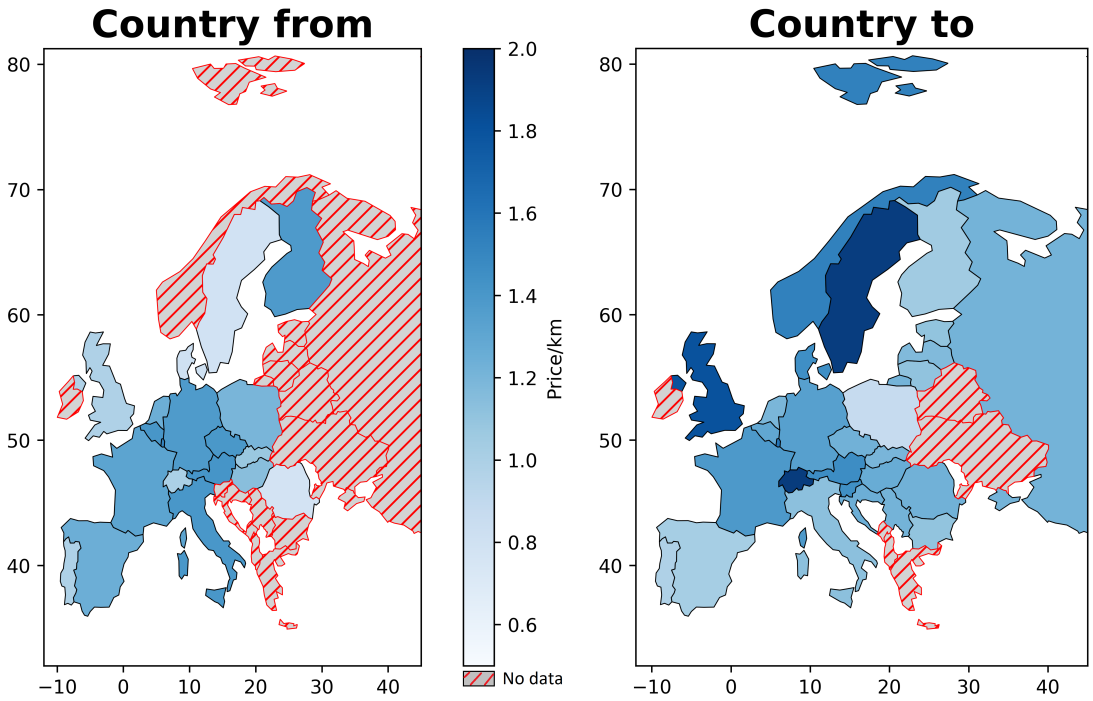

Figure 7.4: A visualization of the prices per country (left: prices by origin, right: prices by destination)

In this figure, the aggregated prices per country are subdivided into a "country from" (left) and a "country to" (right). From the picture, it becomes clear that on average, transports from Sweden are cheaper than transports from Finland. It also becomes clear that the data in the left figure is more dense than the data in the right figure. This is because in the left figure, it appears that a lot of countries are missing in the dataset. The absence of a Business unit (BU) near these countries may partly affect this missingness, but the main effect is explained due commercial and operational reasons, as a transport company's presence on the map is determined by its competitive edges over other competitors. Sweden is an example which is present in both maps while there is no existence of a BU in Sweden. Preferably, there is another loading location close to the unloading location of transport because this way, empty backhaul can be minimized, which is the (empty) return movement of vehicles over part or all of the route. By minimizing the empty backhaul, the utilization of trucks is maximized. The undesired alternative is that the truck has to drive unloaded (empty) to the next loading location, which may become costly dependent on the distance. The strategy of preventing empty backhaul is exactly what is observed in the example of Sweden: transport occurs to Sweden and consecutively, trucks are loaded close to the unloading location. More specifically, it is observed that prices for transport from Sweden are cheaper than transport to Sweden. By offering lower prices from Sweden, Ewals Cargo Care (ECC) is able to prevent the empty backhaul problem. Profit on these lanes will be lower, but this is compensated by the higher prices for transport to Sweden. Sweden is not the only country in which this balancing problem of transports is visible. In the United Kingdom and Swiss, the same observation can be done. According to ECC, these are all consumption countries. With consumption countries, it is meant that these countries mainly focus on the consumption of goods, rather than the production. Practically, in the world of logistics this means that there is a higher need for transport entering than to leave the country (import>export). Production countries follow the reverse logic: there is a lot of demand for the transportation of moving goods produced in this type of country. Related to production countries (e.g.: Spain, Italy, Poland), there is also presence of a pattern: prices for transport from these locations are higher than the prices to these locations, which can be explained with the same repositioning-logic explained before: because of the high demand for loading in the production countries, it is beneficial for a

Logistic Service Provider (LSP) to acquire transports with said production country as destination. For this, it is willing to pay a price: offer lower prices to prevent empty repositioning costs. Lastly, the vast majority of countries in the figure have stable prices, which is the result of contract business. Demand of transport for this type of business is consistent. Therefore, transports leaving and entering these countries can be arranged in advance. Therefore, the LSP is able to solve the repositioning/backhaul problem well in advance.

In a discussion with Product Design & Product Development Network Fleet Specialist Mr. Simons, it came forward that the price alone is not the determining factor whether or not the trajectory is interesting or not. He mentioned that the costs involved in transportation highly affects the prices. Toll-prices, ferry-prices or toll-bridges are all factors influencing the price. A starting point in the analysis to investigate what trajectories are interesting ones and which not is a comparison between the import and export prices related to individual country relationships. This makes comparison grounded, as it is likely that the same costs apply in both export and import to the same country. The top-10 most differing country-relationships are displayed in Table 7.2 below.

Table 7.2: Differences between export and import on the level of country relationship (top-10)

Export country relationship	**Import country relationship**	**Prediction difference (export-import)**	**Prediction difference (absolute)**
BE-GB	GB-BE	1.22	1.22
PL-RU	RU-PL	1.18	1.18
CH-DE	DE-CH	-1.12	1.12
DE-GB	GB-DE	1.02	1.02
GB-NL	NL-GB	-0.93	0.93
PL-SE	SE-PL	0.93	0.93
DE-SE	SE-DE	0.9	0.9
DE-DK	DK-DE	0.83	0.83
FR-GB	GB-FR	0.8	0.8
PL-RO	RO-PL	0.63	0.63

In this table, only the top-10 is shown. From this, it becomes clear that a lot transports directed to Great-Britain (GB) appear competitive. However, returning from GB appears to be an issue. The output of the predictions in this research are the input for a problem that follows after this research: network optimization. This problem is related to the arrangement of transportation in such a way that the amount of empty backhaul is minimized and the profit is maximized.

Another interesting aspect to investigate is whether the type of goods transported affect the prediction of prices, which is measured by the Strategic End-Market (SEM). In an interview (Section 11.5) with Product Design & Product Development Network Fleet Specialist Mr. Simons, it came forward that aggregating transportation data by the market alone is not sufficient. The reason for this is that the predicted price per kilometer highly depends on the country relationship and aggregating on a higher lever will likely skew the data. In order to account for this, the top-10 country relationships are investigated in deeper detail. This top-10 comprises of the country relationships in which ECC currently has the most business, which is defined by the total value of transport according to Equation 7.2 below.

$$\text{Value}_c = \sum_{r \in R} \sum_{l \in L} (s_{r,l,c} \cdot p_{r,l,c}),\ \forall c \in C \tag{7.2}$$

In this formula, R refers to the set of RFQ's, L to the set of lanes and C to the set of country relationships. For each country relationship, the value can be calculated by summing over the number of shipments s multiplied by the price for transport p of all the lanes in RFQ's which concern transport on a particular country relationship c. These top-10 countries account for roughly

30% of all the business ECC is involved in and can be found in the right part of Table 7.4. Because data related to one country-relationship is missing, only 9 top-10 country relationships are displayed in the table. The notation of country relationships is based on the Alpha-2 country codes which can be found in Section 11.2. Next to defining these top-10 country relationships, rather than averaging the prices per kilometer, a ranking is given to the markets for each country relationship. For example, suppose that in a particular country relationship, the predicted price per kilometer is highest in the automotive market, than a ranking of 1 is assigned to this market. The average ranking of the top-10 country relationships are shown in Figure 7.5 below.

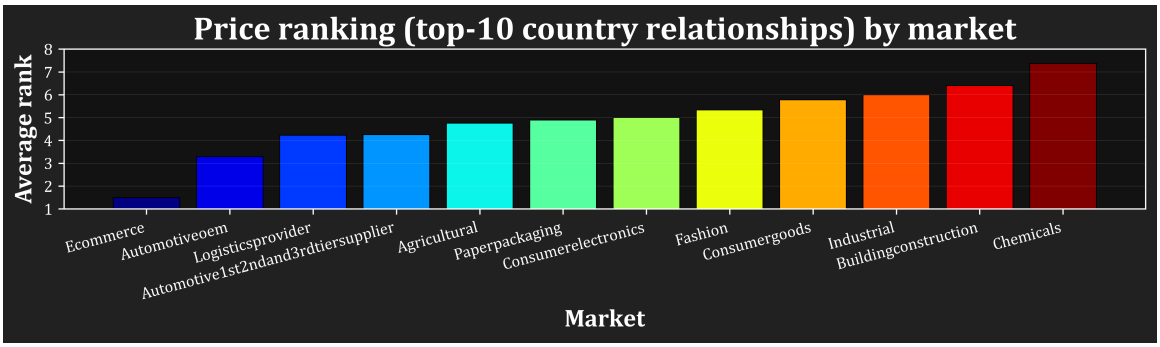

Figure 7.5: The average ranking related to the predictions for the top-10 country relationships by market

From the figure, it can be seen that the e-commerce market on average has the lowest ranking with quite some margin and is therefore on average accompanied with the highest price predictions relative to the other markets in the same country relationship. The e-commerce sector puts great value on rapid deliveries and is happy to be a price for this. These rapid deliveries can also be derived from the data, where the e-commerce is on the fourth spot in terms of lead-time (from all 13 SEM's). The presence of the automotive sector in this chart can be explained due to the fit between the equipment type and the automotive customers. Because of this fit, customers are able to maximize their load per shipment and therefore reduce their costs per unit. However, the fit between equipment type and its customers is not always there: in the chemicals, building construction and industrial sectors, the fit is considerable less, which reduces the competitiveness of ECC. The logistics provider and agricultural sectors are accompanied with a varied set of goods, which makes it difficult to draw any conclusion. The type of goods in the paper-packaging sector is low-value, while the requirements for transport are not demanding at all. Therefore, it is expected that the prices in this sector are not that high. In an interview (Section 11.5) with Product Design & Product Development Network Fleet Specialist Mr. Simons, the paper-packaging market is a market in which a reasonable margin on the price is scarce. He also mentioned that business in this market is particularly useful in solving the repositioning problem of trucks.

7.2.3 Comparison to the market price

Next, it is interesting to see whether or not the predictions deviate from the market prices acquired from Transporeon (*About Ticontract | Transporeon*, 2021). For this, Figure 7.6 is drawn below, which shows the number of times the deviations to the market price occur. The deviation to the market price is measured by the absolute difference of the predicted price per kilometer $\hat{y}$ and the market price per kilometer y_{market}.

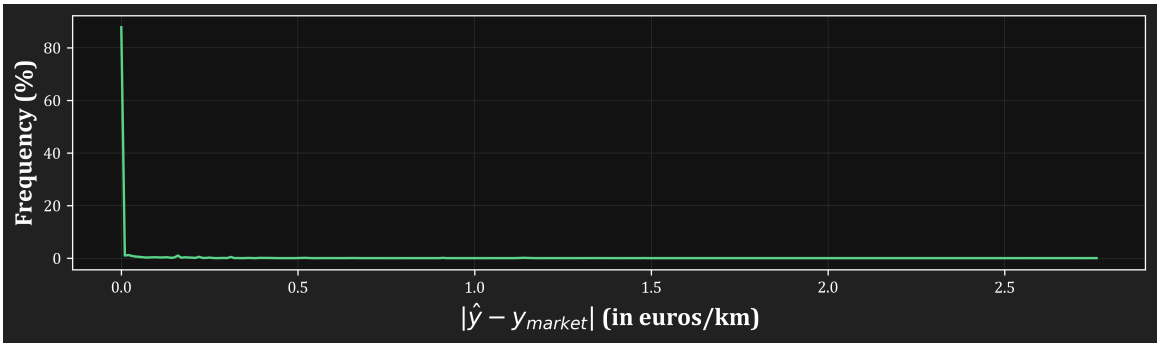

Figure 7.6: The distribution of deviations between the predicted prices $\hat{y}$ and the market prices y_{market} (rounded to 2 decimals)

From the figure, it appears that more than 80% of the predictions comply with the market price (when rounding on 0.01 €). From the left table in Table 7.4, the mean prices also appear to be close to each other. Nevertheless, statistical comparison of the means with the help of Welch's t-test (Welch, 1947) brings evidence that the mean prices are different from each other ($p = 0.002$) and the mean market price is lower than the mean price of the prediction ($p = 0.001$). Moreover, it was found that in 6,988 out of 13,127 samples (53%), the prediction was higher than the market price.

Analysis of the deviations by country relationships, leads to the top-5 and bottom-5 country relationships in terms of deviation ($\hat{y} - y_{market}$) between the predicted price $\hat{y}$ and the market price y_{market} in Table 7.3 below.

Table 7.3: Top-5 countries (left) and bottom-5 countries in terms of accuracy in comparison with market data ($|\hat{y} - y_{market}|$)

Country relationship	$\hat{\mathbf{y}} - \mathbf{y_{market}}$	**Nr of observations**
FI-SE	0,69	152
SK-SK	0.22	4
DE-AT	0.24	72
IT-GB	0.13	140
PL-ES	0.10	154

Country relationship	$\hat{\mathbf{y}} - \mathbf{y_{market}}$	**Nr of observations**
LU-NL	-0.57	1
AT-AT	-0.47	1
LU-DE	-0.25	1
DE-DK	-0.12	3
IT-FR	-0.12	1

From this, it appears that the highest deviations exist on transports between Finland and Sweden, which could be explained by the large amount of awarded business on this trajectory. Because of the amount of business awarded on this trajectory, the true values for the model will be pushed towards the RFQ price, rather than the market price. There seems to be a relation between the share of business awarded on a trajectory with the deviation to the market price. Therefore, the relationship between the two is shown in Figure 7.7 below.

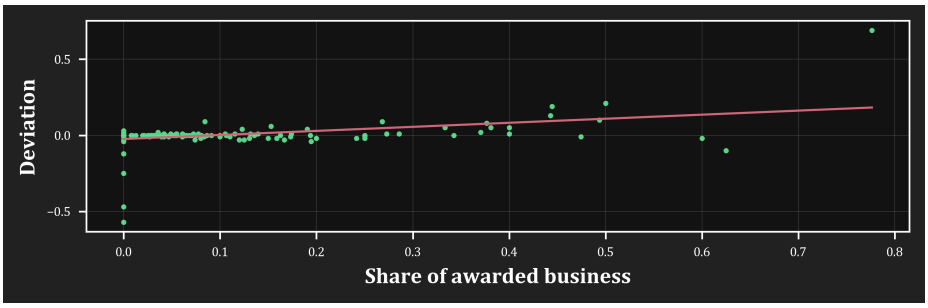

Figure 7.7: The relationship between the deviation of the predicted price and the market price ($\hat{y} - y_{market}$) with the share of business awarded

From the trend-line (defined by minimizing the sum of squared errors for linear linear regression), it appears that there is indeed a relation between the two: the higher the share of awarded business as the true outcome variable, the higher the deviation from the market price. This also explains the highest deviations from the top-5 country relationships in Table 7.3. With respect to the bottom-5 countries from the table, it appears that only few observations in the data are accompanied with a top-5 country. By only observing few observations for this country relationship, the model is not able to learn the appropriate relationships in order to predict a price per kilometer accurately. This complies with the book by Hair (2009), in which it is suggested that a minimum of 10 observations per predictor variable should be used as a rule of thumb. Next to the top-5 and bottom-5 country relationships (in terms of deviation) discussed, 75 country relationships had a deviation of 0.00 € in comparison to the market data. This accounted to roughly 41% of all observations. Next, the top-10 countries defined before are investigated in more detail. The deviations from these country relationships in comparison to the market price are shown in the right part of Table 7.4 below.

Table 7.4: Summary of the market price versus prediction (left) and a comparison between the market prices and the prediction between 9 top-10 country relationships

Metric	**Market price (€/km)**	**Prediction price (€/km)**
count	13127	13127
mean	1.27	1.28
std	0.34	0.38
min	0.46	0.42
25%	0.98	0.98
50%	1.25	1.25
75%	1.56	1.56
max	2.6	4.43

Country relationship	**Market price (€/km)**	**Prediction price (€/km)**
DE-GB	1.88	1.90
NL-DE	1.41	1.39
FR-GB	2.10	2.10
DE-SE	1.71	1.70
ES-GB	1.69	1.69
NL-GB	1.93	1.93
NL-FR	1.60	1.59
BE-DE	1.35	1.35
CZ-DE	1.46	1.46

From this table, it appears that the previous finding (prediction price higher than the market price) does not hold for the top-10 country relationships at ECC. Only in the trajectory "DE-GB", a prediction higher than the market price is the result. In four country relationships, a more competitive price was predicted than the market price, whereas in five top-10 countries, the prices correspond with each other. In 98.4% of all observations, the market prices fall within the prediction intervals. A serious deviation from the market price would lead to questioning the model in terms of accuracy according to ECC specialists.

7.2.4 Performance on business not won

Specifically for all the observations in the historic data for which business was not won, it is investigated how the model behaves. For this, a dataset of 3,370 samples is made available. 76% of these price predictions were lower than the price which was offered in the name of the company, indicating the model is capable of predicting a more competitive price than the sales specialists. However, 794 samples in this dataset were predicted higher than the price offered. From these 794 samples, 89% of the predicted prices are also higher than the market prices. These are the cases where the model is the least performing, because prediction of a price higher than both the offered price and the market price, does not seem competitive at all. However, 91 % of these samples still remain within the boundaries of the prediction interval defined in Equation 7.1.

8 Deployment

In this section, it is explained how the predictive pricing tool is incorporated within the business processes of Ewals Cargo Care (ECC). In Section 1.1, it became clear that the Sales Rate Manual (SRM) is an outdated tool developed with Microsoft Excel (Corporation, 2018). Because of this, the tool is currently being refurbished in Python (Van Rossum & Drake, 2009), which was also the main resource for modelling in this project.

8.1 Pricing Calculation Tool

The name of the new tool in development is called the Pricing Calculation Tool (PCT), which will replace the existing SRM. The amount of available data influencing the prices of transport is increasingly growing in both quantity and quality and therefore, a solution had to be found in order to deal with the computational issues of the SRM. The new tool is able to collect data from a variety of sources that were not included in the SRM before, such as market data and data from the ECC's data warehouse. By incorporating these data sources, the price calculation will be increasingly more data-driven. A lot of development is currently in process at ECC and for this purpose, Python is growing in importance. In the PCT, not only the calculation of the prices is realized in Python, also the front-end is developed with the help of the programming language. A variety of libraries were involved in the process, where libraries created by PyQT (2012) were most contributing with respect to the front-end development and software by McKinney et al. (2010) was the most contributing factor in developing the back-end. The realization of the tool in Python simplifies the implementation of the predictive pricing project into the new tool, because all data processing steps in this research were also realized with Python. On the short-term, the new tool will be used by the network Business unit (BU) only, whereas on the long-term, ECC wants a uniform pricing process where all BU's have to conform to using the PCT for their price calculations.

8.2 Preparation

In order to make use of the model developed in this research, the information required by the model should be considered. In Section 5, the relevant factors influencing the prices per kilometer were determined and this is exactly the input to the model which should be provided by the PCT. The information required by the model is collected from a variety of sources: ratecard, actuals- and market data source. Most information can be collected from the "ratecard", which is the summary prepared by the RFQ-desk when a RFQ enters the business process of ECC and the RFQ is of interest to the company. The actuals and market data source were already explained in Section 4. Some information, such as lead-time, modality and equipment type can be decided by sales specialists, whereas lead-time and modality can also be a requirement set by the customer. For all the variables relevant to the model, the data sources are listed in Table 8.1 below.

Table 8.1: The sources for required information by the model

Variable	Source
Lane	Ratecard
Number of rounds	Ratecard
Customer	Ratecard
Strategic end market	Ratecard
Lead-time	Ratecard / sales specialist
Modality	Ratecard / sales specialist
Equipment type	Sales specialist
Initiative	Ratecard
Current business	Actuals
Contract rejection rate	Market

In order to simplify re-training the model, the following steps are maintained: collect data; pre-process data; train model; save model. In the first step, data from all sources is collected and com-

bined into one dataset representing the predictor variables defined in Table 5.15 and the outcome variable "price per kilometer", defined in Section 5.2.1. In the next step, the data is pre-processed in order to account for contamination, missingness and standardization of data, explained in Section 5. The last step is to train the best model (a Support Vector Regression (SVR)) concluded from Section 6.5 on all data, making use of the hyper-parameter configuration described in Section 6.4. Lastly, when the model completed the training phase, the model can be saved. This is done in a *joblib* format (Joblib Development Team, 2020), which can be loaded in any Python application.

Because the amount of data increases every day, the model should be re-trained periodically. Because the amount of data on which the model can be trained is still limited, there are no computational difficulties which affect the training phase yet. Therefore, the model can be trained easily on a daily basis. In the future, possibilities with respect to online training can be investigated. This way, the model does not have to be re-trained manually, but will automatically do so when new data is available to the model.

8.3 Launch

When the joblib library is installed in the PCT environment, the model can be loaded in the tool. Because during the calculation, the sales specialists should be able to manually adjust some of the parameters in the price-calculation (see Table 8.1), it is possible that some of the inputs to the model are changed during the pricing process. Therefore, the predictive pricing tool cannot be used before sending the RFQ to the sales department, but should be incorporated in the price calculation instead. A preview of the PCT can be found in Section 11.6, whereas in Figure 11.4 the calculation screen is previewed. When a cell referring to an input of the predictive pricing model is adjusted, the model should be able to re-initiate a price prediction. From Table 8.1, it follows that this should be the case when the lead-time, modality or equipment type is adjusted by the sales specialist.

The prediction of the price is implemented in the PCT by incorporating two new columns in the tool. These two columns represent the predicted price and the prediction interval that is defined in Equation 7.1. The final implementation is previewed in Figure 8.1 below.

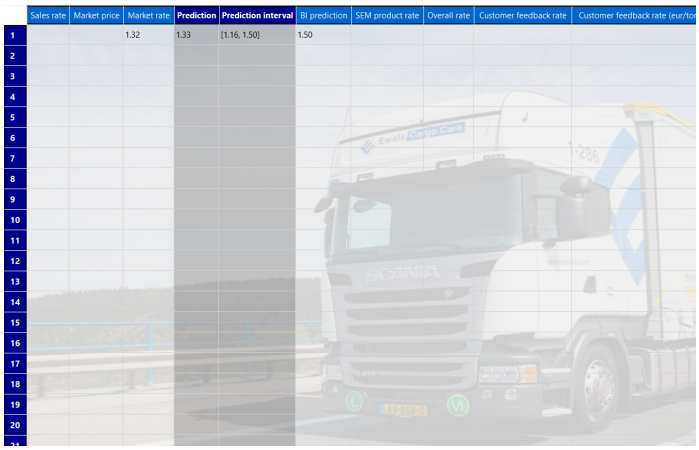

Figure 8.1: A preview of the predictive pricing project implemented in the Pricing Calculation Tool

8.4 Data storage

In Section 1.1, it became clear that the output of the SRM was not logged to database. In the PCT, this issue is solved. In collaboration with the team responsible for data management at ECC, the output of all calculations is uploaded periodically to the data warehouse, which enables to quickly access all data relevant to sales specialists.

9 Conclusion & Discussion

In this section, the findings of the research are discussed with respect to the research questions formulated in Section 1. In order to do this, the section is subdivided into two topics: data management and modelling, as the research questions relate to these topics. Following from the results and the discussion, a set of recommendations to Ewals Cargo Care (ECC) is composed, after which the limitations and the direction for future research is discussed. The section ends with a conclusion.

9.1 Discussion

9.1.1 Data management

With respect to data management, it appeared that ECC has access to multiple data sources. From this extensive amount of possibilities, it is essential to make a right selection such that the prediction of prices can be realized accurately. Next to selecting the right data sources, it is important that the data found in the sources were treated with care. In Section 4, it was concluded that relevant data can be found from the following four data sources (Figure 4.1): RFQ-data; feedback data; actuals data; market data.

Taking into account the quality of the data, lead to the founding that data in control of the RFQ-desk and the market data have a sufficient quality of the data and is useful for modelling purposes. The RFQ-desk is careful with treatment of data as the desk aims at using business intelligence to enhance offering capabilities of the company. The data sources that are within the control of the RFQ-desk are the RFQ- and the feedback data sources. The market data source is controlled by an external organization, which maintains a platform (*About Ticontract | Transporeon*, 2021) that is widely used for requesting transportation services via a Request For Quotation (RFQ). The data became available only in the final phase of this research and because ECC is one of the first to acquire this data commercially, some problems with the data were identified. Although the data is still assumed to be clean, the data is not complete. Nevertheless, part of the market data shows its relevance in this research. With respect to the actuals data (the invoice data), some critical problems were discovered. When an attempt was made on linking the RFQ-data to the actuals data (Section 4.4), a variety of problems were encountered. The main problems were three-fold:

1. Absence of an overlapping key with other data sources
2. The absence of a uniform logging approach
3. Operational deviations from the RFQ

Concluding the data management topic, the four pillars towards predictive pricing of RFQ's in Figure 4.1 are now reduced to the three pillars that can be found in Figure 9.1 below.

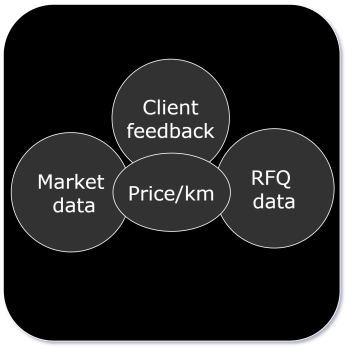

Figure 9.1: The three pillars towards predictive pricing

9.1.2 Modelling

In order to answer which Machine-Learning models are most appropriate for price-setting of RFQ's, the answer is not unambiguously. From the evaluation in Section 6, it appears that the Support Vector Regression (SVR) model outperforms the Multilayer Perceptron (MLP) model on all performance metrics. From Table 6.5, it already became clear that the top-three SVR models all outperformed the best MLP model. Therefore, it is safe to say that the SVR is better in predicting the target variable. It appeared that for both models, a considerate amount of variance was explained, implying that the problem under investigation is suitable for application of Machine Learning algorithms. Moreover, the mean (absolute) error of only 0.02 € indicates a relatively low error margin. The performance on observations related to business that was not won (Section 7.2.4) showed that in the majority of the observations (76%), a more competitive price than the price offered was advised. Assuming that for all of these samples business would have been won when offering this price prediction would be premature. Nevertheless, the price suggested by the predictions are closer to winning business than the one that was offered historically. The results are promising and therefore, it would be interesting to investigate its performance in practise.

9.2 Savings potential

Furthermore, it is interesting to put thought to the potentials in terms of savings. In Section 3, it was mentioned that the need of a new model arose from performance-related issues in the current Sales Rate Manual (SRM). A Turnover Hit-Rate (TOHR) of 6% indicated a lack of performance and on top of this, a lot of time is invested in the quoting process. By introducing the new model to the price calculation, time investments can be significantly reduced. In order to discuss this, an important distinction has to be made between short-term- and long-term savings. On the short-term, the predictive capability has to be proven in practise: specialists need to be convinced that the model predictions are close enough to their perception of competitive prices. On the long-term, the model could evolve into a model that is capable to automatically price RFQ's. The long-term is where the potential of the model is most promising.

Short-term
On the short-term, the process of pricing RFQ's remains the same to the as-is situation. However, the price predictions can be used to benchmark the prices developed by Sales Engineers. This is realized by integrating the price prediction into the tool which is employed by the Sales Engineers. With the help of these predictions, the amount of time involved in the pricing process can be reduced. Nowadays, the competitiveness of the calculated prices is often a question mark, invoking a lot of discussion about the eventual price as a result. With the benchmark, it is possible to indicate the competitiveness of the prices developed by the Sales Engineers, reducing the time required for discussion.

Long-term
In the long-term, when the model is consistently being evaluated and improved, the model may be used to automatically price the incoming RFQ's. RFQ's enter the business processes of a transportation company from a variety of sources and for these requests, resources have to allocated efficiently in order to maximize the profit. Therefore, with a strong model, the entire calculation process can be revised in the long-run. While at the moment, a lot of time is invested in the calculation process, by automating this process, the selection process for participating in RFQ's can be less strict. This way, ECC is able to participate in an increased amount of RFQ's while spending less time in the calculation process. In addition to the RFQ's that are clearly of interest, participation in other RFQ's can be realized with a new model as well because of the automation of the process. However, this brings some more implications to the table: by automating the process and subsequently increasing the amount of participations in RFQ's, the Key Performance Indicator (KPI) introduced in Section 3 (Turnover Hit-Rate (TOHR)) will not be a valid performance indicator anymore. In the new process, it is possible that ECC will also participate in RFQ's in which ECC has no competitive advantage and consequently has a high chance of not winning any business at all.

9.3 Recommendations

In the short-term, it is recommended to integrate the Support Vector Regression (SVR) algorithm in the current pricing calculation process. It is suggested to make use of the prediction interval proposed in Equation 7.1, such that sales specialists can use the upper bound in earlier rounds and a price closer to the lower side of the interval can be offered in a later stadium based on the feedback of the client. By means of this prediction model, the specialists responsible for calculation of the prices are backed with a data-driven model, and do not solely have to rely on their own intuition and information anymore. The world is evolving into a data-driven world and it is important that knowledge could be transferred easily. By means of consistently revising the model and sharing information that could help the model towards more competitive prices, the perceived added value of the model should increase. With revising the model, the following aspects are suggested:

1. Improving the training data
    - The completeness and correctness of the data should consistently be subject for improvement. While the RFQ data is considered as data of sufficient quality, other data sources are in need of improvement. The actuals data is an example of this. By improving data from this source, options to include actuals data to the model is enabled. The actuals data can be improved by introducing a uniform data logging process at the company, such that all the data sources can be connected with each other and the quality of the data is ensured.
    - The amount of training data can be increased by including archived data of the RFQ desk. This data is not clean by nature and should therefore be cleaned carefully before using it for training the model.
    - The outcome value should constantly be revised. In this research, it was decided that the true value is a combination of the RFQ data, feedback data and market data. Although this is currently a best practise according to this research, the quality of the data is still an area of improvement. Whereas the business won observed in the RFQ data is perceived as the truth, the feedback data and market data can still be improved. By pro-actively chasing the clients of ECC for quantitative feedback, the quality of this data source can be improved. Secondly, by improving the granularity of the market data, data from this source can be improved in terms of quality as well. This will align with the zone setup discussed in Section 3. The company is already convinced that aggregation on zone-level has added value for the company. When the market data is made available in this aggregation level, the accuracy will likely be improved, because from Section 7.1 it already appeared that the price is sensitive to the country relationship.
2. Improving the variables
    - The variables selected for modelling should not be assumed definitive. From the sensitivity analysis in Section 7.1, it appeared that some variables are less sensitive to the price than others. Moreover, the majority of variables are micro-level, implying company-specific data is included to predict a price on macro-level: the profit-maximizing but business winning price at a particular moment. It is recommended to investigate other sources that could be of use in predicting competitive prices. Some suggestions are:
        - **Difference between spot and contract rate**: from an interesting discussion with Product Design & Product Development Network Fleet Specialist Mr. Simons and Manager Product Intelligence Freek Heesen (Section 11.5) it came forward that the difference between the spot price and the contract price is likely to be a factor influencing the future prices of transport. When spot price are higher than contract prices, it is likely that contract prices will increase because of the growing demand. Moreover, when the prices for contract business are less favorable, transport companies will engage more in spot business.

  - **Competitor information**: any information about the capacity of the closest competitors could improve the chances of winning business.
  - **Truck demand**: the demand for trucks in the automotive market. When demand is less booming, capacity could be a problem in the future.
  - **Driver contracts**: information about the conditions of driver contracts in a certain country. It would be beneficial to predict shortages of drivers in a certain country, such that anticipation on this problem is possible. Additionally, prices for transport will when shortages of drivers are a fact.
  - **Fuel data**: at the moment, no fuel information is included in the model, while this likely impacts the prices considerably. Fuel impacts the prices for transport directly, and the demand for transport indirectly. It is interesting to investigate whether or not fuel information will benefit the predictions of prices.

3. Improving the model

   - Because the world of Machine Learning is one of the most rapidly evolving technical fields (Jordan & Mitchell, 2015), it is of utmost importance to keep track of the developments in the field. The type of models, the hyper-parameter tuning-, training-, validation- and evaluation methods are all subject for improvement and the potential gains in terms of performance should be exploited whenever possible.

On the long-term however, it is recommended to investigate the possibilities to deploy a model which is capable of automatically pricing RFQ's. This way, significant savings can be realized in terms of resource investment. Additionally, dependency on human intuition can be reduced while decisions are increasingly data-driven by nature. Moreover, it is recommended to reset the Turnover Hit-Rate (TOHR) as a performance measure. Because the entire process is revised, the performance should not be compared to a situation in which ECC only participates in RFQ's in which chances of winning business are highest. This research still suggests the TOHR KPI as the main performance metric, as it is still perceived to be a useful tool in measuring the performance of the model. When the process of automatically pricing RFQ's is deployed, the performance can be monitored over time with the help of this KPI. In order to compare performance of the new process with the old process, it is recommended to introduce new KPI's. This research suggests the following KPI's:

1. **Product turnover hit-rate**: this KPI measures the total amount of business won (in turnover) belonging to ECC's product portfolio.

2. **New business turnover hit-rate**: this KPI measures the total amount of business (in turnover) that does not belong to ECC's product portfolio.

With product portfolio, the trajectories are meant in which the company is perceived to be most competitive. The first KPI measures whether or not the new model will increase (decrease) the turnover on the portfolio transports, whereas the second KPI measures if the model is able to increase the business won on transports which are not considered in the old situation.

9.4 Limitations & Future research

It is important to understand what aspects influence the price of transport. While this research achieved the prediction of prices with the currently available data, it is important to extend the boundaries and investigate what data is not yet present in the current model, but should be in the future. Because models are only as good as the data that is fed to the model, the data should be evaluated constantly. Currently, the data at Ewals Cargo Care (ECC) has its limitations. These limitations can be related to data at Ewals Cargo Care (ECC) itself (actuals data), but can also be related to the data sources that are maintained externally (market data). The limitations do not solely relate to the data sources, but also to the variables. Whereas this research mostly includes micro-level data, the importance of macro-level data should not be neglected. Where micro refers to the data internally at Ewals Cargo Care (ECC), macro-level data refers to data from a wider perspective. An example of this is the market data, which is comprised of data that measures the transportation market as a whole. Currently, this market data is the only macro-level data, which has its limitations. First, the aggregation level is a problem. Currently, the market data is available by country relationship, whereas prices may differ between countries on a more detailed aggregation level.

Next, this research was limited to the investigation of Machine-Learning models in order to predict a competitive price in a RFQ. However, some of these variables may better be investigated with time series models. The reason for this is that when RFQ's enter the business process of a Logistic Service Provider (LSP), the transport itself is being executed in the future. While the market may seem favorable at the moment the RFQ was introduced to the LSP, the situation can be changed completely over the course of a couple months (Covid-19, Suez-canal barrier, Ukrainian war). By using time-series to forecast time-sensitive variables, there can be (partly) accounted for the accompanied uncertainty in these variables by using the forecasts as input to the Machine-Learning models.

This research only considered a pre-specified set of models in order to set competitive prices at a LSP. Both the Support Vector Regression (SVR)- and the Multilayer Perceptron (MLP) models are difficult to investigate in terms of analyzing why a particular prediction is made, because they are known for their black box nature. This research therefore suggests the exploration of a wider set of models, which may even show an improvement in performance. In particular the Artificial Neural Network (ANN) model did not perform as good as expected in this research. The amount of qualitative data is likely to be a cause for this. Neural Networks are capable of utilizing thousands, if not millions of training data. Because the amount of data in both quantity and quality is limited in this research, the entire capability of a Neural Network may not be maximized in this particular research.

In summary, the following directions for future research are suggested:

1. Analysis of macro-level factors affecting the prices in a RFQ
2. Opportunities to exploit time-series analysis in order to forecast time-sensitive variables in the transportation sector
3. Investigate which other Machine Learning algorithms may flourish in the world of RFQ's

10 References

About ticontract / transporeon. (2021). https://www.transporeon.com/en/about-us/ticontract/. ((Accessed on 09/29/2021))

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish journal of emergency medicine*, *18*(3), 91–93.

An, N., Elmaghraby, W., & Keskinocak, P. (2005). Bidding strategies and their impact on revenues in combinatorial auctions. *Journal of Revenue and Pricing Management*, *3*(4), 337–357.

Baheti, P. (2022, Apr). *12 types of neural networks activation functions: How to choose?* Retrieved from https://www.v7labs.com/blog/neural-networks-activation-functions

Ban, G.-Y., & Keskin, N. B. (2021). Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science*.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, *160*(901), 268–282.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade* (pp. 437–478). Springer.

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, *24*.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Box, G. E. (1953). Non-normality and tests on variances. *Biometrika*, *40*(3/4), 318–335.

Chen, J., & Kindnes, J. (2021). *Reverse auction definition.* https://www.investopedia.com/terms/r/reverse-auction.asp. ((Accessed on 09/29/2021))

Corporation, M. (2018). *Microsoft excel.* Retrieved from https://office.microsoft.com/excel

D'Agostino, R., & Pearson, E. S. (1973). Tests for departure from normality. empirical results for the distributions of b 2 and$\sqrt{}$ b. *Biometrika*, *60*(3), 613–622.

de Roeck, J. (2022). Smart evaluation tool for incoming requests for quotation at ewals cargo care.

Elmaghraby, W., & Keskinocak, P. (2004). Combinatorial auctions in procurement. In *The practice of supply chain management: Where theory and application converge* (pp. 245–258). Springer.

Fonti, V., & Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam research paper in business analytics*, *30*, 1–25.

Girden, E. R. (1992). *Anova: Repeated measures* (No. 84). Sage.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT press.

Gudehus, T., & Kotzab, H. (2009). Logistic pricing and marketing. In *Comprehensive logistics* (pp. 157–184). Springer.

Hair, J. F. (2009). Multivariate data analysis.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., ... others (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, *29*(6), 82–97.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

HM-Group. (2021, 12). *Paying weight.* https://www.logisticsglossary.com/term/paying-weight/. ((Accessed on 12/14/2021))

Jamieson, K., & Talwalkar, A. (2016). Non-stochastic best arm identification and hyperparameter optimization. In *Artificial intelligence and statistics* (pp. 240–248).

Joblib Development Team. (2020). *Joblib: running python functions as pipeline jobs.* Retrieved from https://joblib.readthedocs.io/

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. CRC Press.

Laux, H., Bytyn, A., Ascheid, G., Schmeink, A., Kurt, G. K., & Dartmann, G. (2018). Learning-based indoor localization for industrial applications. In *Proceedings of the 15th acm international conference on computing frontiers* (pp. 355–362).

Levene, H. (1961). Robust tests for equality of variances. *Contributions to probability and statistics. Essays in honor of Harold Hotelling*, 279–292.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), 1–45.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2017). Hyperband: A novel bandit-based approach to hyperparameter optimization. *The Journal of Machine Learning Research*, *18*(1), 6765–6816.

Malato, G. (2021, May). *Feature selection in machine learning using lasso regression.* Towards Data Science. Retrieved from `https://towardsdatascience.com/feature-selection-in-machine-learning-using-lasso-regression-7809c7c2771a`

Massey Jr, F. J. (1951). The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68–78.

McKinney, W., et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference* (Vol. 445, pp. 51–56).

Muthukrishnan, R., & Rohini, R. (2016). Lasso: A feature selection technique in predictive modeling for machine learning. In *2016 ieee international conference on advances in computer applications (icaca)* (pp. 18–20).

Nagelkerke, N. J., et al. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.

Nataraj, S., Alvarez, C., Sada, L., Juan, A., Panadero, J., & Bayliss, C. (2020). Applying statistical learning methods for forecasting prices and enhancing the probability of success in logistics tenders. *Transportation Research Procedia*, *47*, 529-536.

O'Malley, T., Bursztein, E., Long, J., Chollet, F., Jin, H., Invernizzi, L., et al. (2019). *Kerastuner.* `https://github.com/keras-team/keras-tuner`.

Pahulje, M. (2021). *5 tips to fight the tender rejection rate.* Retrieved from `https://blog.flexis.com/5-tips-to-fight-the-tender-rejection-rate`

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Porter, M. E., & Strategy, C. (1980). Techniques for analyzing industries and competitors. *Competitive Strategy. New York: Free*.

Ptv group. (2021, Nov). Retrieved from `https://company.ptvgroup.com/nl`

PyQT. (2012). Pyqt reference guide. Retrieved from `http://www.riverbankcomputing.com/static/Docs/PyQt4/html/index.html`

Rao, C. R., Rao, C. R., Statistiker, M., Rao, C. R., & Rao, C. R. (1973). *Linear statistical inference and its applications* (Vol. 2). Wiley New York.

Raychaudhuri, S., & Veeramani, D. (2003). Carrier bidding strategies for iterative auctions for transportation services.

Schranz, M., Umlauft, M., Sende, M., & Elmenreich, W. (2020). Swarm robotic behaviors and current applications. *Frontiers in Robotics and AI*, *7*, 36.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shapiro, S. S., & Wilk, M. B. (1965, dec). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591–611. Retrieved from `https://doi.org/10.1093/biomet/52.3-4.591` doi: 10.1093/biomet/52.3-4.591

Sheather, S. (2009). *A modern approach to regression with r*. Springer Science & Business Media.

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl Jr, K. C. (2017). *Data mining for business analytics: concepts, techniques, and applications in r*. John Wiley & Sons.

Smets, K., Verdonk, B., & Jordaan, E. M. (2007). Evaluation of performance measures for svr

hyperparameter selection. In *2007 international joint conference on neural networks* (pp. 637–642).

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, *25*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, *15*(1), 1929–1958.

Taskesen, E. (2019). *distfit.* urlhttps://github.com/erdogant/distfit.

Teschemacher, U., & Reinhart, G. (2017). Ant colony optimization algorithms to enable dynamic milkrun logistics. *Procedia CIRP*, *63*, 762–767.

TLN. (2021). *Dé online routeplanner voor de logistieke branche - tlnplanner.* `https://www.tlnplanner.nl/`. ((Accessed on 09/29/2021))

Ueasangkomsate, P., Lohatepanont, M., et al. (2012). Bidding strategies for carrier in combinatorial transportation auction. *International Journal of Business Research and Management*, *3*(1), 1–17.

Van Aken, J. E., & Berends, H. (2018). *Problem solving in organizations.* Cambridge university press.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual.* Scotts Valley, CA: CreateSpace.

Wallach, D., & Goffinet, B. (1989). Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecological modelling*, *44*(3-4), 299–306.

Welch, B. L. (1947). The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, *34*(1-2), 28–35.

Wen, J., He, L., & Zhu, F. (2018). Swarm robotics control and communications: Imminent challenges for next generation smart logistics. *IEEE Communications Magazine*, *56*(7), 102–107.

Woschank, M., Rauch, E., & Zsifkovits, H. (2020). A review of further directions for artificial intelligence, machine learning, and deep learning in smart logistics. *Sustainability*, *12*(9), 3760.

Yamashita, T., Yamashita, K., & Kamimura, R. (2007). A stepwise aic method for variable selection in linear regression. *Communications in Statistics—Theory and Methods*, *36*(13), 2395–2403.

Yang, C., Feng, Y., & Whinston, A. (2021). Dynamic pricing and information disclosure for fresh produce: An artificial intelligence approach. *Production and Operations Management*.

Yuen, K. K., & Dixon, W. (1973). The approximate behaviour and performance of the two-sample trimmed t. *Biometrika*, *60*(2), 369–374.

Zhang, H., Weng, T.-W., Chen, P.-Y., Hsieh, C.-J., & Daniel, L. (2018). Efficient neural network robustness certification with general activation functions. *Advances in neural information processing systems*, *31*.

Zhang, J., Nault, B. R., & Tu, Y. (2015). A dynamic pricing strategy for a 3pl provider with heterogeneous customers. *International Journal of Production Economics*, *169*, 31–43.

Zhang, Y., Luo, H., & He, Y. (2015). A system for tender price evaluation of construction project based on big data. *Procedia Engineering*, *123*, 606–614.

11 Appendices

11.1 Appendix A: Hierarchical structure

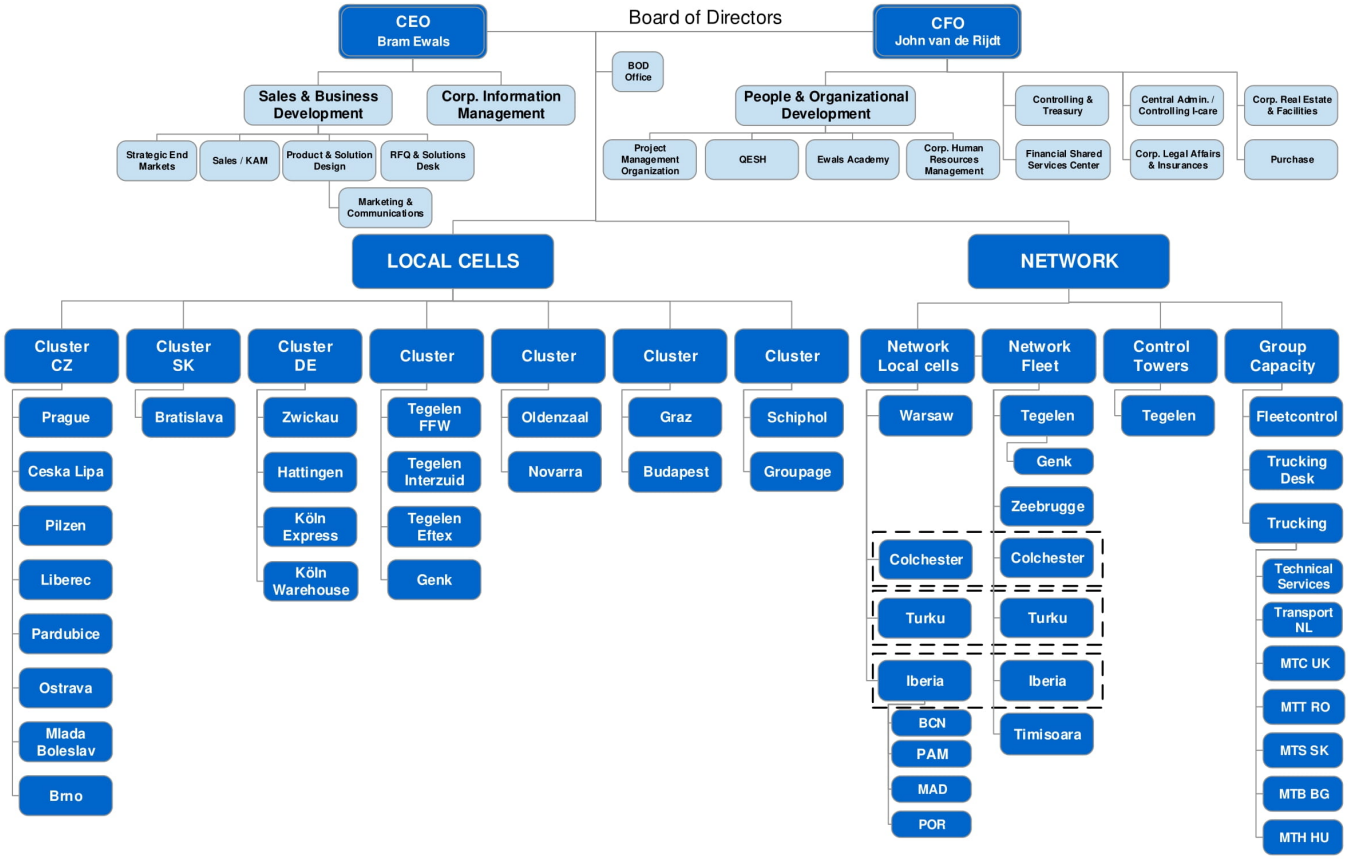

Figure 11.1: Organization chart of Ewals Cargo Care

11.2 Appendix B: Countries and their Alpha-2 codes

Table 11.1: A list of countries with their Alpha2 code

Country	**Alpha2**
Austria	AT
Belgium	BE
Bulgaria	BG
Croatia	HR
Cyprus	CY
Czech Republic	CZ
Denmark	DK
Estonia	EE
Finland	FI
France	FR
Germany	DE
Greece	GR
Hungary	HU
Ireland	IE
Italy	IT
Latvia	LV
Lithuania	LT
Luxembourg	LU
Malta	MT
Netherlands	NL
Poland	PL
Portugal	PT
Romania	RO
Slovakia	SK
Slovenia	SI
Spain	ES
Sweden	SE
United Kingdom (Great Britain)	GB

11.3 Appendix C: Distribution fitting

In order to fit distributions to empirical data, the distfit library (Taskesen, 2019) is employed. The library tests several distributions, but the distributions shown in Table 11.2 are included in this research.

Table 11.2: Considered distributions from the distfit library

norm	genextreme
expon	gamma
pareto	lognorm
dweibull	beta
t	uniform

The considered distributions are fitted with the help of the Residual Sum of Squares (RSS), of which the formula is depicted in Equation 11.1 below.

$$RSS = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{11.1}$$

After estimating the distributions, the results are visualized in two figures: the first figure shows the empirical distribution, the best fitted distribution and the 95-% confidence-intervals. The second figure plots the RSS-values on the y-axis and the tested distributions (Table 11.2) on the x-axis. In the latter plot, the differences in estimates of the tested distributions can easily be derived.

11.4 Appendix D: Kerastuner subclassing

```
import keras_tuner as kt
from sklearn import model_selection
import numpy as np
import tensorflow as tf

class MyBayesianTuner(kt.BayesianOptimization):
    def run_trial(self, trial, x, y):
        cv = model_selection.KFold(kwargs.pop('kfolds'))
        val_losses = []
        fold = 1
        for train_indices, test_indices in cv.split(x):
            x_train, x_test = x[train_indices], x[test_indices]
            y_train, y_test = y[train_indices], y[test_indices]
            model = self.hypermodel.build(trial.hyperparameters)
            self.hypermodel.fit(trial.hyperparameters, model, **kwargs)
            val_losses.append(model.evaluate(x_test, y_test))
            fold += 1
        self.oracle.update_trial(trial.trial_id, {'val_loss': np.mean(val_losses)})

class MyHyperModel(kt.HyperModel):
    def __init__(self):
        kt.HyperModel.__init__(self)

    def build(self, hp):
        # BUILD MODEL
        model = tf.keras.Sequential()

        # INPUT LAYER
        input_layer = tf.keras.Input(shape=self.nr_features)
        model.add(input_layer)

        # HIDDEN LAYERS
        for i in range(hp.Int('layers', 2, 10)):
            layer = tf.keras.layers.Dense(hp.Int(f'neurons_lay{i+1}', 1, 20),
                                                    activation=hp.Choice(f'
                                                    activation_function_lay{i+1}',
                                                     ['relu', 'sigmoid']))
            model.add(layer)
            dropout_layer = tf.keras.layers.Dropout(**self.dropout_params)
            model.add(dropout_layer)
        model.compile(
            optimizer=tf.keras.optimizers.Adam(learning_rate=hp.Choice('lr', [1e-4,
                                                    1e-3, 1e-2, 1e-1])),
            loss='mean_squared_error',
            )

        # OUTPUT LAYER
        layer = tf.keras.layers.Dense(1)
        model.add(layer)
        return model

    def fit(self, hp, model):
        return model.fit(
            batch_size=hp.Choice('batch_size', [16, 32, 64, 128]))
```

11.5 Appendix E: Interview Product Design & Product Development Network Fleet Specialist (Henk Simons) & Manager Product Intelligence (Freek Heesen)

This interview was planned in order to discuss the results of the prediction model and acquire some additional insights on the implications and possibilities of the model. For this matter, Henk Simons was invited because of his practical knowledge of fleet and price-setting. He is responsible for the current cost-based Sales Rate Manual (SRM), so he has awareness of the competitiveness of the prices suggested by the new model. Moreover, Freek Heesen was invited to the meeting because of his extensive analytical skills, practical know-how and his overall involvement in the process of this research. In the interview, the persons involved are referred to as Henk (Henk Simons), Freek (Freek Heesen) and Nick (Nick Janssen).

Date: 25-05-2020
Nick: Thank you for joining. I prepared a presentation for this meeting which is subdivided into a problem definition, data sources, variables and modelling. In short, a lot of time is invested in RFQ process while the KPI is still low. Because of this, we are in search of methods in order to improve the process. For this purpose, I developed Machine-Learning models which predicts a price based on a set of inputs. This development is now completed and they were tested on a dataset which was comprised of cases in which we offered, but did not win any business. This way, we can see if the model is able to predict more competitive prices than the ones we offered. Additionally, a comparison against the market data is conducted, which will be discussed later. First, I will show you the modelling results. The performance metric of the model which is the most interpretable is the Mean Absolute Deviation (MAD), which we can see here and boils down to 0.03 €, implying that the mean deviation from the true price is 0.03 € on average.

Freek: It would be powerful if you could back this error margin, by showing how to use this error in practise. We work with a RFQ process, which is often distinguished into multiple rounds. I propose using this error margin to produce a confidence interval. For example, suppose our predicted price is 1.00 € for a transport of 1000 kilometers. For a first round, we can use the upper bound for the offer ($1000 * (1 + 0.03) = 1030.00$ €) and see what the feedback of this offer will be.

Nick: Would you not be afraid to lose the business if you would offer the highest price?

Freek: No, because you would always get an additional chance in the next round.

Henk: The carrier which already realizes transportation of the company always has a last-call option.

Nick: What is a last-call option?

Freek: This means that the company in need for transportation always informs the current Logistic Service Provider (LSP) that they got a better offer for transportation. Then, the LSP can still choose to lower the price and keep the business consequently.

Henk: This is preferential because this LSP knows the business and does not want to lose it because this also costs money and resources.

Nick: Okay, this is good to know. Then I made a visualization which shows in what countries the highest prices were present. Here, a distinction between country-from and country-to was made. From the picture it becomes visible that the consumption countries are relatively expensive, because a lot of transport will go to these countries, but not from there. Therefore, LSP's compensate for the possibility of empty backhaul asking a higher price. For the production countries, the logic is the other way around: a lot of transport is requested from these locations, but not to these location.

Freek: How do you explain Ireland (there is no data for Ireland)?

Nick: North-Ireland is present, but the rest of Ireland is abscent in the dataset. This may be due to the cleaning of the dataset.

Freek: Yes this may be due to the different postal-code setup of Ireland (which is based on EIR-codes instead of the traditional postal-codes). The method you used in order to calculate distances for the lanes in your dataset, does not hold for Ireland. Therefore, you probably eliminated these observations because they were not accompanied with a distance. I would revise this if I were you.

Henk: Yes and North-Ireland is coded as "GB", so this is probably linked to the data from the rest of England. Furthermore, we can see that the country-from prices are way lower than the country-to prices, which is due to the imbalances.

Nick: I will look into the Ireland issue. Moreover, I compared the mean prices of the prediction with the mean prices of the market. Here we can see that both means are close to each other. On the right, I displayed the top-10 countries by value calculation, where the domestic transports are kept out. The prices shown are prices/km.

Henk: How do I determine the price/km to for example Great Britain? Is this based on the kilometers from TLN?

Freek: It is also important to be aware of the difference from our process to that of TLN/PTV in calculating distances, because we also do a lot of intermodal, unaccompanied transport. TLN/PTV does not account for this. They only account for a road solution. If you want to compare these prices, you should benchmark the road observations only. However, to use this as a benchmark, the model will be of use, because on some trajectories, we compete with intermodal solutions to road solutions of competitors. Ultimately, the customer decides if they want to transport their goods by means of intermodal transport or by road. So for this matter, the model will be right. However, when we consider the example trajectory of Germany to Great Britain, TLN/PTV will calculate the distance by means of using the tunnel. In our process, we will use unaccompanied intermodal transport, which means we do not account for these kilometers. Here, the model will provide an inaccurate indication of a price per km.

Henk: If I start in Hamburg, and I would end in Liverpool, than PTV/TLN will calculate the distance between Hamburg and Calais, use the tunnel to Folkestone and consequently drive to Liverpool. No LSP company is interested in doing this, because intermodal transport is way more interesting and the price is way lower. So for trajectories between Great britain and the continent, this procedure will not be valid.

Freek: On the continent, this model will certainly be useful, definitely in order to benchmark the prices, but to Great Britain this is a lot more complicated.

Nick: Do you see a way to account for this in the model in the future?

Freek: In principal, you already benchmark with the market, so you already account for this in the model. Additionally, a substantial amount of observations are already present in the dataset. A lot of business is already present on this trajectory, so it should provide an indication.

Nick: Furthermore, I received a dataset which contained offers from business that is not won. Interesting to investigate would be to test the performance of the model to this dataset. 76% of the observations, a lower price was predicted, which means that prices were more competitive than the prices offered.

Freek: Assuming that the price is the driving force for winning is the price?

Nick: Yes. From the cases in which a higher price was predicted, 89% was also higher than the market price. From these inaccurate price prediction, 32% originate from the Paper packaging market, 22% from the automotive OEM market and 17% from the LogisticsServiceProviders.

Freek: What are your takeaways from these results?

Nick: That the weaknesses of the model mainly come to light in the top-3 markets just mentioned.

Freek: But to get a good grasp of this, you should also show the successes in the markets: in which markets does the model perform well? Otherwise, these numbers do not mean anything. It is too premature to tell from these numbers that the predictions in the paperpackaging market are not accurate. Maybe most observations are distributed in the paperpackaging market and the distribution is large because of this. How do you see this practically?

Nick: Practically, I would say that when a RFQ enters our process and it contains transport in the paper-packaging market, you should not blindly follow the model prediction because it is not that competitive.

Henk: The paperpackaging market does not pay well, so I think it is strange that the automotive market is so close to the paperpackaging market in these results.

Freek: Maybe you can zoom in in which customers this may have been in the automotive market. The paperpackaging market is not a high-value product and costs are often lower. Delivery time is not as important here and insurance is also less applicable. The reason why we do a lot of business in this market is because of repositioning. The reason why prices are often lower in this market is that we do not have to gain a lot of profit in this market. By repositioning, we already benefit from the transport over empty backhaul.

Henk: It is also possible that we offer higher prices when we want to get rid of some business, or because we have low capacity and spot rates are more interesting than contract rates.

Nick: And it is also possible that the results show that the model is competitive in the automotive market, because I do not show this in the picture here. Moreover, when I zoom in on a similar observation from the training set to a observation in the test set, I see that a higher price was predicted than the one in the training set, based on a different contract rejection rate and type of initiative.

Henk: I know this lane and what happened here, is that the ferry rates exploded significantly because of the diesel prices. The rate increased with 125 € from diesel alone. Moreover, nowadays everything has to be transported with local traction in Sweden and we ourselves are responsible for getting the trailers to the east op Europe, increasing the cost even more. I believe immediately that the prediction of the model is higher when these factors are accounted for. There is a huge difference between month 3 or 4 in this month compared to month 3 or 4 last year.

Nick: The diesel price is something I do not account for in the model, so maybe this is still room for improvement.

Henk: But not everything is diesel.

Freek: At least you have to account for time. An indication of the time is so important in impacting the prices.

Nick: I already considered time, but because I only have data from 1.5 years I could not really discover a real pattern.

Henk: I think you can already spot this in the market data, when you observe a particular country-country relationship a lot of what is happening can be explained with the business cycle at a particular moment. The incident at Suez Canal and Covid also impact this business cycle and therefore also impact the prices. At the moment, we also see an explosion of diesel prices which are the caused by the Ukraine war.

Nick: That is why I wondered if I should include time in the model, or that it is better explained by another factor, such a diesel price, the capacity of the market, etcetera.

Freek: The development of prices with time is explainable by other factors, such as diesel prices and business cycle. In a stable market, time is less important, but in a volatile market, this is of utmost importance.

Nick: But if we would have access to data which measures the business cycle and external factors which actually impact the prices, rather than just a "time" component, wouldn't this be be better?

Freek: Yes for sure, we need macro economical factors in such a model. For example, you can include the purchase manager index, which measures the German economy. There several of such factors which could benefit the model.

Nick: Yes , I agree that this would benefit the model and that this is something which is still missing in the research.

Freek: Even more so in a volatile market. In a stable market, it is a question how to fill the trailers, but a market like nowadays is completely different and we need to account for such things in order to realize a sustainable business.

Nick: For example, the crypto-market is also a highly volatile market, in which Machine-Learning is being used in order to predict the market.

Freek: The crypto market is a little bit a different story, but indeed, here it is also relevant to find out what are the causes of the volatility of the market. We can think something of this, because diesel, changed legislation, covid, inflation etcetera, all impact(ed) the prices in our market. Anyways, when we look at your comparison, I am curious how you compensate for time in your comparison? How do you know these RFQ's are different from each other for example?

Nick: I do not really account for time, I account for the market. The contract rejection rates in these examples are different from each other, which indicates these RFQ's are different from each other and RFQ's were in a different period. Maybe this is the same RFQ in a different year.

Freek: Do you perform a time-series?

Nick: No, in Machine-Learning I have a set of input variables which are imputed to the model, after which the model predicts a price. I thought about including time in the model, but I only have 1.5 years of data.

Freek: Yes but this will only become bigger and bigger. In ten years, we want to be able to look back and learn from this. Henk performs well in his job because he has the knowledge in his head. I want to include this knowledge in such techniques. But I think it is clear now.

Nick: So this is what I prepared for the meeting and what I had in mind for the results and discussion. I am still looking for ways to extend the insights I could get from the results of the model. For example, with this picture over here, I wanted to show in which markets the model would not perform well.

Henk: But are the predictions worse based on the market, or is the root cause the country relationship? If we have loading filled with paper, we know this does not pay well. But the imbalance in Sweden is only solved if we make use of these paper loadings. Therefore, the loading to Sweden has to compensate for the low margin we make from transporting from Sweden. If we do not do this, we have a problem. We have one paperpackaging client which pays well to Sweden, but for the rest we do not do a lot of business with the paperpackaging industry. Therefore, it is logical that the prediction accuracy will be bad. But if you decide to include this client to the prediction, this is not a good representation of the paperpackaging industry based on what they request for the transport.

Nick: So you think it will be beneficial to perform an analysis based on country-country relationship?

Henk: Yes I think so. Suppose transport from country 1 to country 2 is 1.10 €/km and I will not be successful in the paper-packaging market, but I will be in other markets like Automotive, machinery, etcetera, and I have limited resources, then I will not even consider the paper packaging market. On the other side, when I still have capacity left, then I will rather do business with paperpackaging than with another LSP.

Freek: It is still interesting to look into the markets in order to find out what the prices are, but only for a specific country-country relationship.

Henk: Yes but only for transport on the continent, because of the kilometer issue we discussed before.

Nick: Maybe it is an idea to look at the top-10 countries I have here, to select only those which are on the continent. So then we have DE-SE, NL-FR, BE-DE, CZ-DE and NL-DE?

Freek: Yes, and to add to this, you can add CZ-DE as a showcase. I am very curious to the performance on this country relationship. Because your prediction looks accurate here, We could say to our colleagues in Prague to use this model and analyze how it performs. For example, predict 20 RFQ's and see how it compares to the market. Do you know enough now?

Nick: I can at least extend my results with new information, so I am helped with this.

Henk: But how should I see this? The market price suggests that the price between Spain and Great-Britain is 1.69 € and we also predict 1.69 €?

Freek: The prediction is based on the four datasets he mentioned before.

Henk: And in these datasets, what is the weight of the market price? Because this is the most relevant one.

Freek: Yes partly, but we also have our offers that lead to business won, which is also a reflection of the truth of the prices.

Nick: The truth in the model is determined as follows: when we won any business, this is seen as the truth. When we did not win but we have feedback data, than this feedback will reflect the truth. When both are missing, the market data is seen as the truth by the model. When all are missing, this data is eliminated from the dataset, as we do not know a truth.

Henk: So the market price is a prediction from *PTV Group* (2021), and our prediction price reflect something in our historic data?

Freek: The prediction is the outcome from his algorithm he developed.

Nick: Yes, and this algorithm sometimes observes the market data as the truth, that's why I think they are close. Also, it includes the contract rejection rate, which is also data related to the market. Moreover, the prices you see here are averages from all observations on this trajectory. It is possible that the market data and the prediction deviate from each other, albeit the means are close to each other.

Freek: So this could be vary between markets on the same trajectory. Every market will have his own flavour of the price.

Nick: Is it maybe a suggestion to create ten inputs for the model, based on which you decide is interesting, and see what the model spits out?

Henk: I think this is indeed something that has added value for you. Then we can see immediately if these are interesting transports in the current market. If we have a lot of capacity, then we want as much business as possible to utilize this capacity. If we achieved this, we can see what transports are better utilizes elsewhere.

Freek: This has to do with supply and demand.

Nick: Wouldn't the capacity of the Ewals fleet be a nice factor to include in the model?

Freek: I think not because we are dependent on the market.

Henk: I agree, with a measure of the business cycle rather than a measure of our own capacity. If all transport companies are busy, it is hard to acquire any capacity for your transport. When you still have some capacity left in your own fleet, you can ask high prices. So we need a measure of the business cycle.

Freek: Another complexity in this story is that what we price now, is related to transport in a couple months (the future). In a stable market, this time factor is not as important, but in a volatile market this can be very relevant.

Henk: I would want to know in advance what the market is going to do. You can offer a price which is 5% lower than what you actually want. But if the economy is bad in a couple months, you will be the king with this price.

Freek: We sat down recently with a company which tries to predict such a thing. However, you have to assume that this prediction is right and that is speculative.

Nick: In order to process what you just mentioned about the capacity and the price we can ask: is it maybe an idea to include a factor of the market capacity relative to our own capacity?

Henk: But how are you going to capture our own capacity?

Nick: Can't we think of something for this?

Henk: The only thing I can think of is the amount of trailers which are empty. We can also calculate how much payload we can theoretically move, and compare this to what we actually move. We also have a list of trailers that are empty in Europe, maybe we can use this. When this list decreases,

capacity is becoming scarce. This can be temporary, but also structurally. At the moment, the market is very booming but this is captures nowhere. I think you just have to measure the business cycle.

Freek: I agree, this reflects if we have any work for our trailers.

Henk: And then the question is what source we can use for this.

Freek: In principal, we have this available: we have the market barometer, purchase manager index and capacity index. The latter one is probably the strongest indicator and we have this available. Apart from this, it is also relevant to look at the efficiency of the process. The low KPI you mentioned which is achieved by investing a lot of time can be improved. If you know with a confidence interval of 0.03€ that a price within this range is competitive, then we also know that prices offered above this threshold do not have any chance to win business, or are lucky hits. In terms of quick wins, we can also decide to not do any ridiculous offers anymore based on this model. This way, a lot of time is saved already. However, if we can automatically offer prices based on the model, then we can still decide to bring those offers because no time is invested anymore, and we can still have lucky hits. You should understand that 99% of our business coincides with 25% of our calculations. This means that 75% of our calculations coincide with only 0.02% of our business in terms of turnover. Therefore, the quick wins are gained in these 75%.

Henk: With respect to the prediction of the price, we just talked about the business cycle. But you can also decide to see how far the spot rate is above the contract rate. If you use the market inforamation for this, and you see the spot rate is above the contract rate, then you will know the contract rates will go up soon.

Freek: This one is good. If there is a positive difference between the spot rate and the contract rate, then there is more demand than supply. Then you know for the future, that the contract rates will increase.

Henk: Exactly, why would you do any contract business if the spot market is more attractive? You have to invest time for making the contracts, etcetera, while more money is made elsewhere.

Freek: And if you see a trend in this every year, then we can use this information to our advantage. Anyways, do you know enough now?

Nick: I got some interesting insights indeed. Thanks for this.

Henk: I am still interested in seeing the model in practise and explain why the predictions. It would be nice if you could show that for a specific market and country-country relationship, the rates will be A and on another trajectory it will be B. However, you have to perform this analysis in a trajectory where you can make a decent comparison. For example, NL-CZ if this is in your dataset. Or NL-SI, NL-HU or something alike. In any case, two small countries such that the kilometers have little impact. Suppose you pick DE-NL for a transport of 50 kilometers and you have to load for two hours, unload for 2 hours, then I want a lot of money for this relatively to the amount of kilometers. Therefore, it would be nice to see the amount of kilometers of the transport next to the average amount of kilometers in a country-country relationship.

Freek: This average distance is available in the market data. I suggest you look into the spot rates versus contract rates and additionally look into the kilometers versus average kilometers. When the kilometers are below average, your price will likely be higher.

Henk: You will see that the prediction will be higher when you investigate NL-NL, NL-BE, NL-DE (Ruhr-area), BE-BE, because these are transports within a range of 35-300 kilometers.

Freek: Therefore he left the domestic transport out of this picture.

Henk: But I still think we can use this information, because the higher prices are a fact.

Nick: I can add the average distance information to the data and perform an analysis with it.

Freek: Yes you can split this data in two by a dummy variable which indicates when the distance is higher than average and when not?

Nick: Thanks for you time and interesting insights.

11.6 Appendix F: Preview of the new Corporate Pricing Calculation Tool

In this appendix, a limited preview of the new Pricing Calculation Tool (PCT) is presented with the help of four windows. In window 1, the tool is initialized and a user can be selected. In window 2, the calculation window is shown. In this screen, the lanes that have to be priced are uploaded and selected. Moreover, the parameters of the calculation can be configured. After configuring the calculation, the calculation can be initialized. All the selected lanes will be priced accordingly, which is shown in the table-like screen in window 3.1. In this window, the sales specialists are able to manually adjust the calculation parameters, until they are satisfied with the calculation. In window 3.2, the market- and a prediction rate is visible, at the end of the calculation window. With these parameters, the sales specialists can benchmark their calculation with the market and the so-called BI-prediction. Among these parameters, the price prediction from this research will be implemented. At first, the prediction will be used by sales specialists in order to benchmark the prices, while the ultimate goal of the tool is automatic pricing. Therefore, the tool will be under development constantly, until the prediction is deemed sufficiently accurate and the company is ready for change in the current pricing procedure.

Window 1

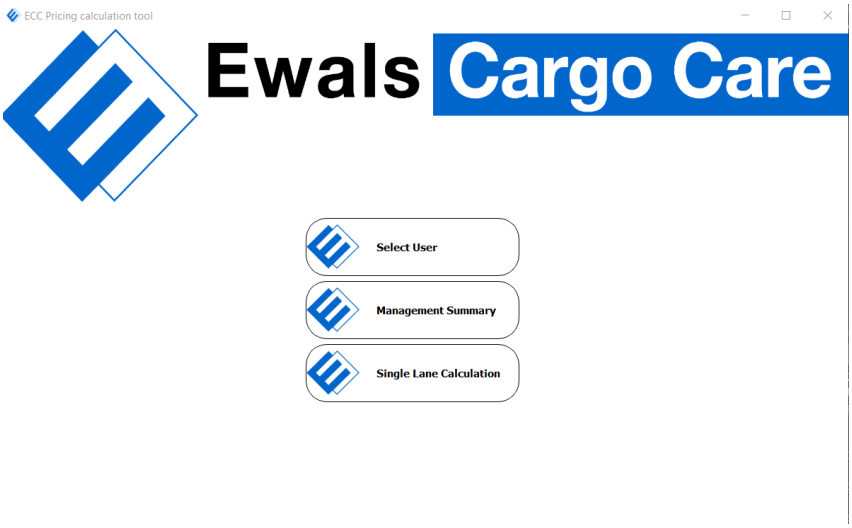

Figure 11.2: A preview of the Pricing Calculation Tool (PCT) (1/4)

Window 2

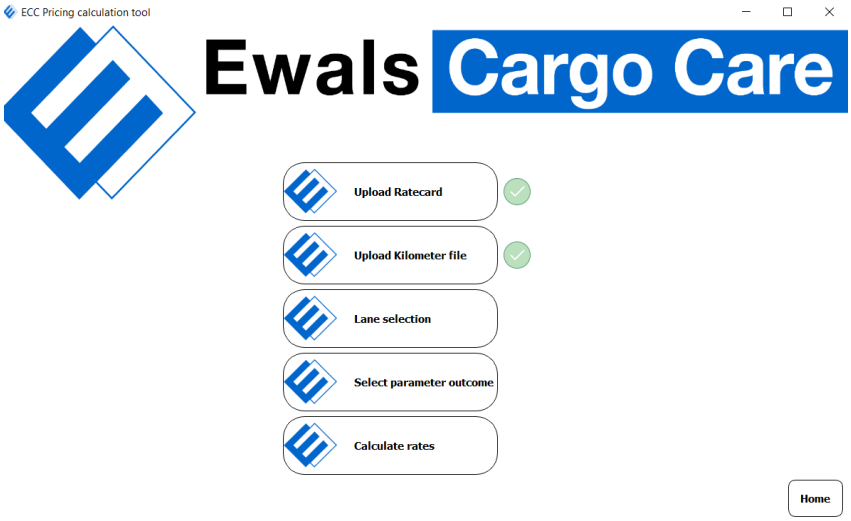

Figure 11.3: A preview of the Pricing Calculation Tool (PCT) (2/4)

Window 3.1

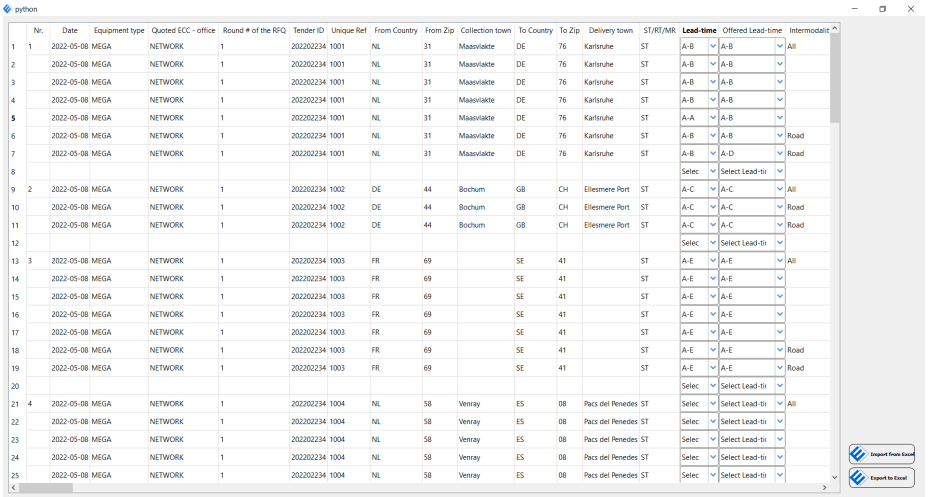

Figure 11.4: A preview of the Pricing Calculation Tool (PCT) (3/4)

Window 3.2

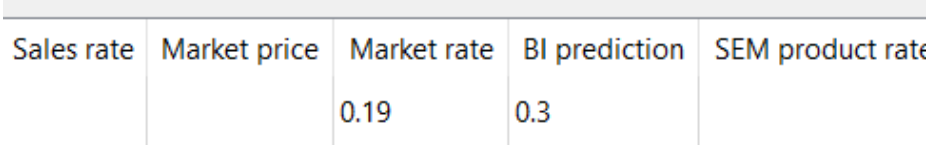

Figure 11.5: A preview of the Pricing Calculation Tool (PCT) (4/4)