MASTER

Smart evaluation tool for incoming requests for quotation at Ewals Cargo Care

de Roeck, J.

Award date:
2022

Eindhoven University of Technology
Department of Industrial Engineering & Innovation Sciences
Operations, Planning, Account and Control Group

Smart evaluation tool for incoming requests for quotation at Ewals Cargo Care

J. de Roeck
Master Thesis
0861709

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

Eindhoven, December 2021

Project Supervisor: Dr. N. Mutlu
Second Supervisor: Dr. A. Akcay
Third Supervisor: Dr L. Genga
Company supervisor: F. Heesen

Preface

This thesis marks the coming end of my Masters in Operations Management and Logistics with a specialization in Data Intensive industries at the Eindhoven University of Technology. A time, which provided a lot of great memories and in which I have developed myself both personally as professionally.

First of all, I would like to thank Nevin Mutlu, my mentor and supervisor of my master thesis project. I would like to thank her for the provided feedback and guidance during my thesis project. Secondly, I would like to thank my second supervisor Alp Akcay for the final feedback providing guidance for completing this project.
Furthermore, I would like to thank my company supervisor Freek Heesen. The discussions and conversations about the data and the decision making within Ewals really helped me developing the tool. I am glad to have had the opportunity to start my professional career at Ewals as well.
Lastly, I would like to thank my parents, brother and sister, for the amazing support during my study period. They are always available to talk and I could always count on their help for overcoming challenges.

Abstract

The growth of demand for logistic activities leads to an intensifying level of procurement activities for transport requests, by setting out RFQ's through tendering. Tendering is defined as the procurement process of a Logistics Service User (LSU), whereby the Logistics Service User (LSU) requests an (price) offer (RFQ) for a transportation service and afterwards the LSU assigns business to one or multiple Logistics Service Provider (LSP) candidates. Participation in this process is very time and resource consuming for the LSP. Therefore, this research focuses on supporting the LSP in deciding in which tendering to participate. In this research, a binary classification model is developed for predicting the outcome of tendering for a LSP. For this prediction task, a logistic regression, artificial neural networks and a support vector machines model are evaluated. The prediction performance for all three models is close to each other. Both the logistic regression model and support vector machines model score slightly better than the other artificial neural networks. These models provide a 68% prediction accuracy.

Management summary

Business Problem

The growth of demand for logistic activities leads to an intensifying level of procurement activities for transport requests, by setting out RFQ's through a tendering process. Tendering is defined as the procurement process of a Logistics Service User (LSU), whereby the Logistics Service User (LSU) requests an (price) offer (RFQ) for a transportation service and afterwards the LSU assigns business to one or multiple LSP candidates. Generating offers for tendering takes much time for the LSP's, while the probability of winning business in tendering could be small, when many LSP's are participating in tendering. Mostly, the LSU assigns multiple winners, meaning each winner will be responsible for specific transport requests within the RFQ.

Ewals Cargo Care is interested in predicting the outcome of tendering, because it helps Ewals Cargo Care to assign resources to promising RFQ's, therefore reduce waste in the main process of the RFQ & Solutions Desk, the department involved in the decision making and preparation of the RFQ offer.

This leads to the following research question:

- *How could Ewals Cargo Care predict in which tendering they should participate to increase potential business winning?*

Analysis

The analysis started with evaluating the available data. Within Ewals data is available on RFQ and product level. On product level less data is available, thus first a model will be build based on the data on RFQ level. First a logistic regression model will be designed. After checking the model fit for the potential model variables, a final model is found with five input variables.

The five input variables are Stragetic End Markets (SEM), current business, Key Account Manager (KAM) customer, FTL/LTL truckload and high potential. The SEM variable represents the strategic end market of the cient. The current business variable represents of the transport business, which is present within the new arrived RFQ, is conducted by Ewals in the year before the arrival RFQ. The third variable KAM customer provides information about the relationship with the customer (good relationship when a KAM is assigned to this customer). The fourth variable indicates if Full Truck Load (FTL) or Less Then Load (LTL) transports are requested. The last variable, high potential indicates if the customers business is perceived interesting by Ewals. Afterwards, the model fit is evaluated for additional variables on product level. Two variables can be added, namely the number of transport lanes and the number of shipments within a RFQ. The model with 5 variables and the model with 7 variables will be evaluated further in the next step.

In the next step of the analysis, the three models: logistic regression, artificial neural networks and support vector machines model, are compared on model performance for both the models named above. The models with five variables outperform the models with seven variables. The performance of the models with five variables is close to each other, however, the logistic regression model and support vector model score slightly better than the artificial neural networks. These models achieve a predicting accuracy of 68% relative to 67% for the artificial neural networks.

Conclusion

The final step is the implementation of the prediction model within Ewals. A decision framework is created to include potential business value in the decision making. The potential business value within a RFQ can differ much. Therefore, a final recommendation for each RFQ will provide by combining the outcome of the prediction model with the possible monetary value that can be gained in a RFQ based on historical performance of products.

Contents

List of Figures

List of Tables

Acronyms

AIC Akaike's entropy-based Information Criterion.

ANN Artificial Neural Network.

BIC Bayesian Information Criterion.

CAM Carrier Assignment Model.

FTL Full Truck Load.

KAM Key Account Manager.

LSP Logistics Service Provider.

LSU Logistics Service User.

LTL Less Then Load.

NDA Non Disclosure Agreement.

QESH Quality Environment Safety Health.

RFI Request For Information.

RFQ Request For Quotation.

SEM Stragetic End Markets.

SMOTE Synthetic Minority Over-sampling Technique.

SVC Support Vector Classification.

SVM Support Vector Machines.

VIF Variance Inflation Factor.

1 Introduction

1.1 Topic Introduction

The demand on the logistics function has grown and will keep growing tremendously, through the ongoing growth in the industry following Bask (2001). This growth is automatically positively correlated with an intensifying level of procurement (a concept known in logistics as tendering). The growth of demand for logistic activities in combination with a shortage of specific logistics knowledge of a Logistics Service User (LSU) leads to an increase in use of a Logistics Service Provider (LSP) (Bask (2001) & Y. Chen, Goan, and Huang (2011)). Therefore, there is a likelihood that procurement activities keep increasing too, causing that LSP's should be more critical in which tendering they participate.

The concept tendering needs to be defined. In general, a Logistics Service User (LSU) has a need for transportation. This LSU sends out a Request For Information (RFI) to get information about a potential Logistics Service Provider (LSP) or a Request For Quotation (RFQ) to get (price) offers from a selection of LSP's. After receiving these offers, they will assign the transport business within the RFQ to one or more LSP's. This complete process will be referred to as tendering. The concepts, Request For Quotation (RFQ) and tender have the same meaning, the term RFQ will be used in the further part of this report. Thus, tendering is defined as the procurement process of a LSU, whereby the LSU requests an (price) offer for a transportation service and afterwards the LSU assigns business to one or multiple LSP candidates. Generating offers for tendering takes much time for the LSP's, while the probability of winning business in tendering could be small, when many LSP's are participating in tendering. Mostly, the LSU assigns multiple winners, meaning each winner will be responsible for specific transport requests within the RFQ. On this moment, LSP's are investing much time in generating the offers for participating in tendering without success, while still little attention for this is given in the literature following Oeser (2020).

One of the LSP's, participating in tendering is Ewals Cargo Care. Mostly, Ewals Cargo Care does not win business at the end of tendering. Data from the start of April 2020 till the end of January 2021 shows that Ewals Cargo Care participated in 633 RFQ's, of which only 165 RFQ's have been successful. This project aims to enhance the opportunities to win business by selecting the right RFQ's, hence introducing the concept of intelligent RFQ selection. This project will therefore focus on how to predict successful tendering for Ewals Cargo Care. This project is conducted for Eindhoven University of Technology and Ewals Cargo Care.

Ewals Cargo Care is interested in predicting the outcome of tendering, because it helps Ewals Cargo Care to assign resources to promising RFQ's, therefore reduce waste in the main process of the RFQ & Solutions Desk, the department involved in this project. This department will be described further in section 4. Furthermore, the prediction model could help to improve the performance of the process by scoring better on the key performance indicators determined by Ewals Cargo Care. Ewals Cargo Care has historical and current data available for building a prediction model. The available data is not used for decision making, currently. Designing a prediction model, could provide advantages in decision making by supporting the decisions with the available data. In the status-quo, the experience of employees is predominantly used to determine to what extent Ewals Cargo Care should participate in a received RFQ. However, making the transitioning to data driven decision making by using a prediction model will help to improve decision making about participation in received RFQ's, which will improve the main process of the RFQ & Solutions Desk.

First, the literature study of this project will help identify criteria which could predict winning business in tendering, provide an overview of current research to tendering from the view of the LSP and provide available prediction models, so that the available data can be used for designing a prediction model. Afterwards, the model will be designed and implemented, which should improve the main process of the RFQ & Solutions Desk.

The following research questions will be investigated in order to study the concept of predicting the outcome of tendering from the perspective of a LSP.

- *How could Ewals Cargo Care predict in which tendering they should participate to increase potential business winning?*

The output of the main research question is the design and implementation of prediction model, which makes a prediction for each incoming RFQ. The expectation arises that there is a huge likelihood that Ewals will participate in RFQ's which are predicted as winning RFQ's. Thus, the prediction model will help Ewals in decision making and focus more on RFQ's, which have a higher probability of being a winning RFQ. Ewals could therefore focus more on RFQ's which will yield revenues instead of no win in tendering and get nothing in return. For answering this research question, first the following subquestions will be answered. These questions help to identify input for the model, help to shape the input data to the right form and to choose the right model.

1. *Which internal and external variables influence tendering?*
2. *Which kind of models are best suited for the available data?*
3. *Which model provides the best predictions?*
4. *Does the decision making supported by the prediction model increase the percentage of winning RFQ's?*
5. *How could the outcome of the model be transformed into a recommendation for Ewals*

1.2 Ewals Cargo Care

Ewals Cargo Care is a logistic service provider, which was founded in 1906 as Ewals by Alfons Ewals. In 1994 two companies, Ewals and Cargo Care, merged into Ewals Cargo Care (further referred to as Ewals).

Ewals was first founded as logistic service provider, mainly occupied with transportation of goods. Besides this main task, nowadays Ewals offers also logistic solutions for customers and the market. Ewals has divided its provided services in six product lines. These are Full Truck Loads, Part Truck Loads, ProjectLeads, ControlTowers, ValueAddedLogistics and Global Forwarding.

Ewals has grown to a lead logistics provider with a turnover of more 600 million with 2400 employees and 37 locations. Ewals offers intermodal transportation, multiple transport modes are involved, and they do have a fleet of 3600 Huckepack trailers, 550 trucks and more than 850 subcontracters to complete transport services. Moreover, Ewals has thirty two offices in Europe, and is therefore locally present.

Ewals is active in the following Stragetic End Markets (SEM):

- Automotive Original Equipment Manufacturer
- Automotive 1^{st}, 2^{nd} and 3^{rd} Tier Supplier
- Agricultural
- Building & Construction
- Industrial
- Consumer Goods
- Consumer Electronics
- Chemicals
- (Lead) Logistics Service Provider (LSP)
- Paper & Packaging
- Aerospace
- Fashion

The biggest strategic end market for Ewals is the automotive market. This market covers around 65% of the total turnover.

The group responsible for RFQ management in whole Europe is the Request For Quotation (RFQ) & Solutions Desk. The RFQ & Solutions Desk delivers service to all commercial oriented branches, residing in the Ewals Group, hence they can be defined as a corporate entity. The main process of this department will be explained in more detail in Section 4. Briefly, a request from the customer arrives, this request is evaluated whereby it will be handled by the RFQ & Solutions Desk or participation will be declined. If the RFQ is handled by the RFQ & Solutions Desk, the processing starts. The outcome of the process is business won or business not won. Business won implies that business has been nominated by Ewals' customer to the designated Ewals entity quoting on the business. Within a RFQ, a second level op depth is found, namely transport lanes. If business is won, minimal one lane is allocated to the Ewals Group. Lanes are subrequests for a specific transport service within a RFQ, detailed description is provided in section 4. When no lane from the requested lanes is allocated to the Ewals group, no business is won. Business lost is defined as the withdrawal of current business.

1.3 Problem Context Request For Quotation & Solutions Desk

This project is executed for the RFQ & Solutions Desk. This desk is mainly responsible for the connection and coordination of the RFQ process in Europe with main stakeholders: Customers / Key Account Manager (KAM) / local cells / legal / SEM leaders / Quality Environment Safety Health (QESH) / Management. These main stakeholders are provided in figure 1. The department is responsible for handling all European RFQ's. 80% Percent of their time is used for processing RFQ's and 20% of the time is used for projects to improve processes within the department or company. Processing RFQ's can also be described as the production process of the RFQ & Solutions Desk. The five main steps will be explained in detail in section 4. The five steps are arrival of a RFQ, validation, preparation & distribution, internal allocation & decision making and feedback. This project focuses mainly on the validation phase.

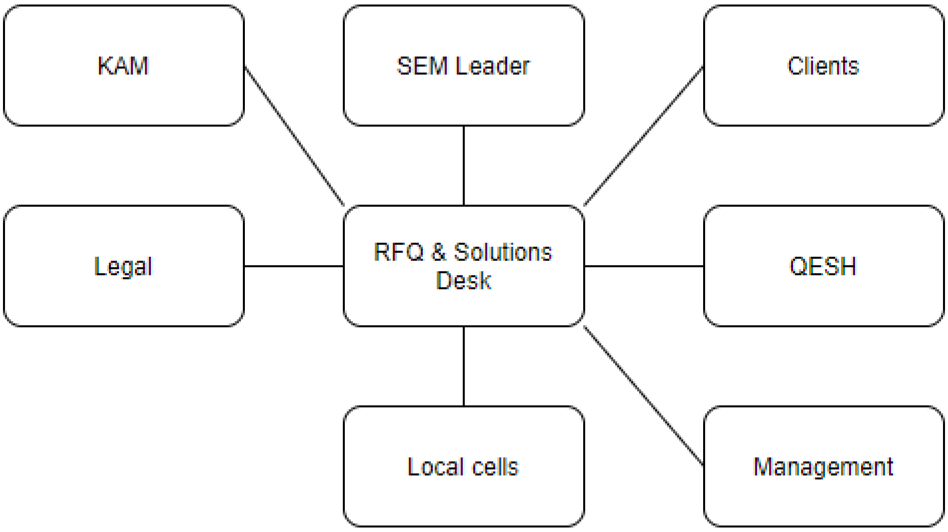

Figure 1: Stakeholders RFQ & Solutions Desk

1.4 Scope and Problem Statement

This project is focused on step two of the process of the RFQ & Solutions Desk, namely the validation. The validation phase consists of three parts, being the validation of the RFQ inquiry, validation of the Non Disclosure Agreement (NDA) and status management.

The most important part for this project is the validation of the RFQ. During the registration, the incoming RFQ's are registered in the RFQ master registration file, this can be seen as the RFQ list, which is the data backbone of the process. A daily stand up is introduced to reconcile the RFQ list each day. This daily stand up meeting is held at the same time and place to decrease complexity, thus the structure remains always the same.

First, the list of external deadlines for the designated day are discussed. New arrived

RFQ's are considered and afterwards these RFQ's are assigned to a team member of the RFQ & Solutions Desk or rejected and closed. Rejection is based on the experience of the team members. Factors influencing this decision are the ability to compete; matching the supply and demand side, strategic fit and feasibility of the deadline. If the RFQ is not fitting for the products or end markets of Ewals, then Ewals would not be able to be competitive or provide a reliable offer, thus it is not logical to participate. The decision is based on the RFQ information, however documentation and analysis of the volumes takes place in step three of the main process of the RFQ and & Solutions Desk.

The decision making of participation in a customers tendering is now done manual and this is very resource demanding. Employees of the RFQ & Solutions Desk determine based on experience and other available information as named before if incoming RFQ's from the customers are interesting. Furthermore, a maximum is reached of handling requests for quotations with the current resources.

The evaluation of interesting RFQ's could be found in figure 2. This could also be described as a two step gate way process. Figure 2 is the second gate. The first gate is checking the Ewals lead list. This list is created by the sales team of Ewals. customers are scored on potential fit with Ewals' services. High potential customers are placed as high potentials on the lead list. Ewals intends always participate in tendering of high potential customers.

The second gate is the complete process as provided in figure 2. Ewals will always participate in tendering of customers with whom Ewals has current business. The scoring model is the decision of a request for quotation engineer, who makes the decision based on experience. However, in the future, a tool should score the RFQ's on potential. This should make the decision making more correct and unambiguous.

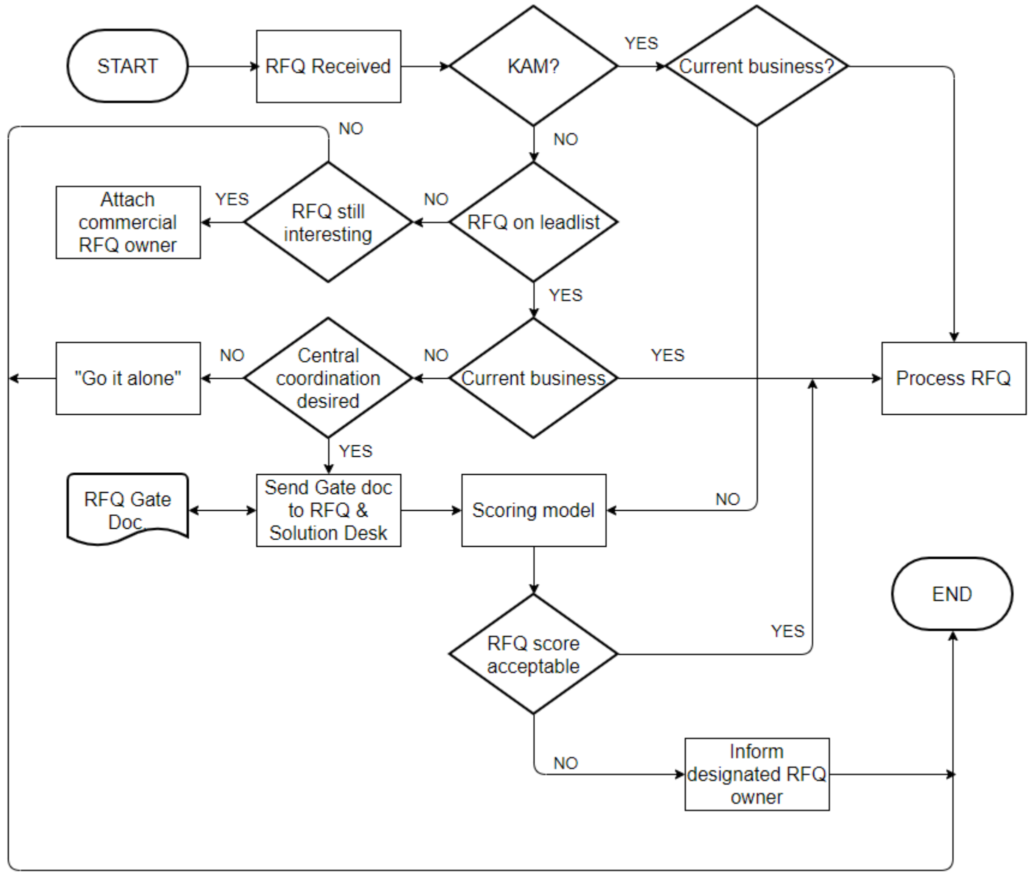

Figure 2: RFQ evaluation

The output of the participation in tendering is measured with two key performance indicators the RFQ hit rate and the hit rate turnover. The first key performance indicator is calculated as follows:

- RFQ hit rate $= \frac{number\ of\ RFQ's\ won}{number\ of\ positively\ evaluated\ RFQ's}$

The variables named in this calculations need additional explanation. The variable number of positively evaluated RFQ's is a subset of the number of RFQ's arrived at Ewals. RFQ's declined immediately in step two of the main process of the RFQ & Solutions Desk are not included in the variable positively evaluated number of RFQ's. Only RFQ's, which are further processed in the RFQ cycle, are counted for the variable number of positively evaluated RFQ's. Thus, the RFQ hit rate provides the percentage of RFQ's won of the total number of RFQ's in which Ewals participated.

The second key performance indicator is calculated as follows:

- Hit rate turnover $= \frac{Awarded\ revenue}{Value\ offered\ to\ customer:\ (allocated)}$

This performance indicator provides the proportion of revenue won by tendering of the total revenue offered to customers. Within Ewals multiple local offices could offer on the same customer request. Only the best offer is send to the customer, this offer is allocated within Ewals. The allocated revenue is thus equal to the revenue offered to customers. The RFQ hit rate provides a percentage of RFQ's won, while the business value of RFQ's varies. Therefore, the second variable hit rate turnover is introduced to compare value of RFQ's won.

The problem to be solved in this thesis is the following. The RFQ hit rate from April first 2020 till the January thirty first 2021 is only 26%. So time and resources are mostly spend to lost tendering. This could be seen as waste of resources. The process should therefore more efficient by implementation of the tool, resources should be dedicated to higher potential requests and by spending less time to RFQ evaluation. Furthermore, the process should also be more effective, a higher hit rate should be achieved.

As named before, the decision making is influenced by certain factors like the ability to be competitive, SEM and feasibility to the deadline. Feasibility of the deadline could be checked quickly. However, checking requested products, SEM, missing information and current business/past experiences takes more time for the employees to investigate. The proposed tool should help to automate and speed up the decision process. Increasing the efficiency and productivity of the process, could also be seen as improving the productivity/performance of the RFQ & Solutions Desk.

Secondly, the decision making based on the experience is not substantiated. By implementing a trained model based on historical and current data, then decision making should be better supported. An example of better decision making, is a supported decision when making a choice between several local cells providing offers for the same lanes based on historical performance. Furthermore, actual market information should be included too. This external variable influences also the outcome of tendering.

Lastly, the decision making should be improved. By using the implemented model, it should be easier to indicate high potential RFQ's. With high potential is meant potential to win business and also high potential in the way of a high profit rate/high margin, which simultaneously is aligned with Ewals' corporate diversification strategy.

1.5 Report outline

The report is structured as follows. Section 2 discusses the current literature relevant for this research. The section contains literature about the context of tendering, literature from the perspective of the LSP and lastly literature from the perspective of the LSU. In section 3, the methodology used in this research is provided. Section 4 provides a detailed description of the main process of the RFQ & Solutions Desk, involved as company department in this research. Section 5 contains the first analysis of the data. Data is evaluated at RFQ and product level. Section 6 provides the different prediction models.

2 Literature background

In this section, relevant literature for this research topic is provided. This section is a summary of the literature study conducted before this research. In the literature, two main directions could be found about this topic. This research project is mainly focused on the transport procurement process. In the literature one research direction focuses on the perspective of the LSU, while the second research direction focuses on the perspective of the LSP.

2.1 Context transport procurement

Following Caplice and Sheffi (2003), three stage could be identified in the procurement process of a LSU searching for transport. These stages are: offer preparation, offer execution and the offer analysis & assignment.

During the preparation stage, the LSU decides which transport lanes will be included in the auction, how to auction will be set up and which LSP's will be invited. Thus, during this stage a pre selection of LSP's is already made. In the literature articles could be found for indicating selection criteria used for selecting the LSP's. Gibson, Sink, and Mundy (1993), Aguezzoul (2014), Oeser (2020) described these criteria. Much criteria where found, therefore ranking methods based on input of experts are used for searching the most important criteria. Important criteria found are historical performance, service, prices, Information Systems, financial stability and handling of special needs.

The second stage is the execution stage which focuses on the communication of the information and collecting the offers provided by the LSP. Ewals is involved in this part of the process. The prediction model designed in this research should help Ewals decide, if they should participate in the procurement process yes or no. Research focused on the perspective of the LSP's is provided in section 2.2.

The last stage is the analysis of the offers and the business nomination of LSP's. Research focused on the perspective of the LSU is about this last stage. Models are designed, which help LSU's assign the winners of the procurement process. Section 2.3 provides a summary of this research.

2.2 Perspective LSP

From the perspective of the LSP two main research directions could be indicated. In one direction research is focused on offer creation, whereby support is provided to come up with the right price, see section 2.2.1. On the other hand research is conducted to prediction models to assign the auction winner, see section 2.2.2.

2.2.1 Offer creation

Following Leung, Luk, Choy, Lam, and Lee (2019), the LSP should first decide if they can fulfill the requested quantity and quality requirements. Afterwards, the LSP should decide which price to offer, which is indicated as the most crucial decision in the RFQ management process indicated Leung et al. (2019). Based on the input from the LSP, the LSU can assign the winner of procurement process. The decision making of the LSP is time consuming and depends mainly on human knowledge and experience. The price offered depends on many considerations and trade-offs like the purchasing frequency, credibility, and the business environment. Therefore, Leung et al. (2019) developed a Smart-Quo system, which is a fuzzy association rule mining approach, which supports in decision making about the pricing.

Other articles focused on offer generation can be found. Song and Regan (2005) highlighted also the shortage of decision support tools for the LSP. They introduce a optimization-based approximation method to solve the offer generation problem. The objective of their method is based on the assumption of minimizing empty movement costs when choosing lanes. This assumption holds only for transport with own assets. The problem is described as the offer construction problem and is formulated as follows:

$$Min \sum_{j=1}^{J} e_j \ * y_j$$

$$s.t. \sum_{j=1}^{J} b_{ij} \ * y_j \ = u_i \quad \forall i \ \epsilon \ I$$

$$y_j = 0,1 \quad \forall j \ \epsilon \ J$$

$$b_{ij} = \begin{cases} \dfrac{1 \ if \ lane \ i \ is \ in \ offer \ j}{0 \ otherwise} \end{cases}$$

The variable y_j is a binary variable representing a offer belonging to set of offer J and the variable e_j indicates the empty return costs associated with a offer. y_j could also be an integer if the lane contains multiple loads. The variable i is a new lane belonging to set of lanes I. The variable u_i represents the number of loads on lane i.

2.2.2 Prediction model winning

Following Adjavor, Al-Hihir, and Wang (2020) few research is executed about the prediction of auction winners with machine learning models. The authors think that the Carrier Assignment Model (CAM) could also be solved with machine learning models to indicate the auction winner instead of optimization methods. Adjavor et al. (2020) focuses on the United States food industry.

The variables provided for logistic regression by Adjavor et al. (2020) are lane volume, rate per distance, manual award amount and lane capacity. A Gini Index is used to indicate the amount of incorrect classifications for each variable. Adjavor et al. (2020)

conclude that the variables rate per distance, manual award amount and volume are statistical significant predictors for winning a auction. Furthermore, he recommends to add more service performance variables to improve the model.

2.3 Perspective LSU

Finally in the procurement process, all transport lanes should be assigned to one or more LSP's. This assignment is solved as a mathematical problem, whereby the the problem is described as a combinatorial auction. Mostly, it is referred to as the winning determination problem. Following Vries and Vohra (2003), LSP's have preference for sets of lanes instead of individual lanes, because of complementary and substitution effects of the lanes offered. In combinatorial auctions of transport lanes, bidders are able to submit bids on different combinations of transport lanes.

Different mathematical models are designed, each with its own extensions to include except price also factors like service/ non monetary requirements (Beil and Wein (2003), Engelbrecht-Wiggans, Haruvy, and Katok (2007), Rekik and Mellouli (2012), Guo, Lim, Rodrigues, and Zhu (2006)) , penalty costs (Zhang, Ding, Li, Wang, and Yao (2014), combining lanes to sets (R. Chen, Gi, Cohn, Beil, and Sinha (2009)) and volume uncertainty (Zhang et al. (2014). One example of the mathematical models will be described below.

Caplice and Sheffi (2003) focus on the analysis of the offers send by a LSP. Caplice and Sheffi (2003) explains the Carrier Assignment Model (CAM), which supports the LSU when analyzing the offers received in tendering. In the paper of Caplice and Sheffi (2003) the cost of the shipper is adjusted for reflecting the service level. Afterwards the minimal value is searched. Guo et al. (2006) extends the CAM with non-price objectives and transit point costs.

The general CAM is defined as follows:

$$min \sum_{k \in K} c^k \ x^k$$

$$subject\ to: \sum_{k \in K} x^k = D$$

$$x^k \ \epsilon \ X$$

Where:

- x^k : represents the volume vector measured in full load trailers for each lane for LSP k
- $C^k(x^k)$: represents the costs for LSP k for transporting the vector of lanes x
- D: indicates a vector of the demanded volume expected for each lane

- X: set of allocations feasible in terms of volume and lane assignments
- K is the set of LSP's

The model is extended with conditional offer, to be able to assign partial lanes, single lanes or set of lanes to a LSP. The extended model is defined as follows:

$$min \sum_{k \in K} \sum_{p \in P} c_p^k \ x_p^k$$

$$s.t. : \sum_{k \in K} \sum_{p \in P} \delta_{ij}^{pk} \ x_p^k = 1 \quad \forall ij$$

$$x_p^k = [0, 1]$$

where:

- i: shipping location origin
- j: shipping location destination
- p: set of lanes p in complete set of lanes P in a conditional offer
- k: LSP identification in set of LSP's K

Decision variable:

- $x_p^k = 1$ if LSP k is assigned all lanes in set p, $= 0$ otherwise

Data:

- $\delta_{ij}^{pk} = 1$ if LSP k's offer set p contains lane i to j, $= 0$ otherwise
- c_p^k Total annual cost for LSP k to service lane set p

2.4 Conclusion

The literature provided focused on two perspectives. From the perspective of the LSU, the literature focused on the assignment of the winner of the procurement process. Optimization is based on price, however, extensions can be found to include factors like service/ non monetary requirements (Beil and Wein (2003), Engelbrecht-Wiggans et al. (2007), Rekik and Mellouli (2012), Guo et al. (2006)) , penalty costs (Zhang et al. (2014), combining lanes to sets (R. Chen et al. (2009)) and volume uncertainty (Zhang et al. (2014). One example of the mathematical models will be described below.

From the perspective of the LSP two directions of research could be found. The first direction is focused on offer creation, mainly on the creation of a price offer. Leung et al. (2019) developed a Smart-Quo system, which is based on a fuzzy association rule mining approach, which supports in decision making about the pricing. However, this research focuses on the decision making of the LSP before the offer preparation.

The most fitting example could be found in the research of Adjavor et al. (2020). Adjavor et al. (2020) come up with machine learning models to indicate the auction winner instead of using an optimization methods. Adjavor et al. (2020) conclude that the variables rate per distance, manual award amount and volume are statistical significant predictors for winning a auction. They suggest to add other service variables into the model.

3 Methodology

As discussed in section 1.3, the main objective of Ewals for this research, is reducing waste in the main process of the RFQ & Solutions Desk. This reduce in waste can be obtained by selecting the RFQ's with the highest probability of winning business by designing a prediction tool. Currently, the available data is not used for decision making. Thus, this project is also aimed at implementing data driven decision making for decision making about RFQ participation. Experienced employees decide on participating in a RFQ on this moment. The main research question and problem explanation is already provided in section 1.1.

The problem solving cycle by Aken and Berends (2018) will be used for this research project. The problem solving cycle can be used for solving field problems. Thus, eventually the focus is on solving the main problem, reducing waste by participating only high potential RFQ's instead of gaining knowledge. The concept high potential RFQ should be defined in this thesis before conclusions can be made, because the output percentage of the prediction model can not be used as final recommendation. The output of the prediction model could be combined with the concept high potential to provide final recommendations.

The steps of the problem solving cycle are discussed in section 3.1. These steps will be linked to the sub questions. After selecting the best model, this model will be used for making predictions for incoming RFQ's. However, not all RFQ outcomes will be known before the ending of this projects. Some RFQ's consist of multiple rounds and these take multiple months to end. The evaluation of these predictions can not be included in this project. However, a number of RFQ's will be selected and evaluated in detail to evaluate the performance of the model. For those tenders the output of the prediction model will be evaluated by comparing it with the decision of the employees of the RFQ & Solutions Desk.

3.1 Sub-question and methodology

Step one of the problem solving cycle, the problem definition is already provided in section 1. This section contains the background information and problem context. Furthermore, the research question and sub questions can be found in section 1. First step two and three will be discussed, namely the analysis and diagnosis phase and the Solutions design phase. After this, the implementation and evaluation will be discussed and this will lead to the final recommendations. An overview of the problem solving cycle is provided in figure 3.

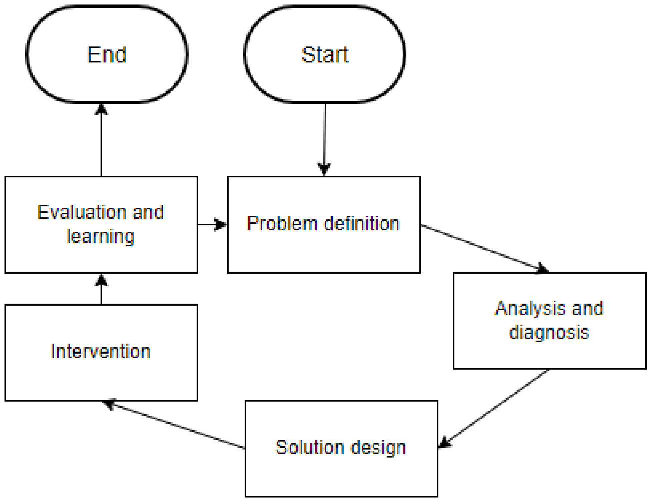

Figure 3: Problem solving cycle by Aken and Berends (2018)

During the analysis and diagnosis phase, the problem is analyzed in detail. The available data will be viewed and evaluated. Summary statistics will be provided to understand the data. First, the data files needed for the prediction model will be selected. Afterwards, the data will be cleaned and filtered.

Furthermore, the information in these data files will be evaluated. This part belongs to research question one and two. Which internal and external variables influence tendering. The context of tendering will be viewed in detail and possible predictors will be explained. The output of this part will be used for the Solutions design. The data will be the input for the prediction models.

The Solutions design phase is also an regulative cycle. The preliminary selected variables will be used for model building. Based on performance metrics and significance, the best model variables could be found, which will lead to the final model. The final model variables could be used in different models to come up with the best prediction model. In this research phase the second and third sub question will be answered, which models are best suited for the data and which model provides the best prediction.

Lastly, the prediction model will be implemented in the business process of Ewals, so that the model supports in the decision making about participating in RFQ's. During this implementation, the output of the model will be evaluated to be able to provide recommendations for decision making based on the numerical output of the model. This answers the last sub question.

4 RFQ and Solutions Desk

In this section the complete production process executed by the RFQ & Solutions Desk will be explained. In handling RFQ's, multiple concepts, namely product, product option and lane, are used to define transport requests. On holistic level, from a commercial point of view, Ewals is selling transport products. However, the concept of products requires specific elaboration as to the validity of its concept throughout this report. A product within the Ewals internal organization's jargon is referred to as a country-country relationship, e.g., " transport from Germany to The Netherlands is the product DE - NL". Thus, country abbreviations are used for naming products for which ISO coding is used as a general process alignment rule (increasing standardization within corporate processes).

Products are not to be confused with "lanes", which have a second level of depth. In this second level of depth, products may consist of second layer specifications, referred to as product options (e.g., Accord Dangereux Routier (ADR) transports, this means that the company complies with government regulations to transport dangerous goods) or lanes could also difference on postal code. Multiple lanes described as the same product could be present multiple times in a RFQ, see figure 4. The difference between the DE-NL products in figure 4 is the delivery postal code. Thus, this difference provides a variation of lanes for the same product. Hence, within this report we refer to products as an unique country relationship.

Unique ref	Client ID	Ewals office	Collection country	Col postal code	Col town	Delivery country	Del postal code	Del town	Country relationship
1422	K0422	ECC Germany	DE	24149	KIEL	NL	10	-	DE-NL
1423	K0423	ECC Germany	DE	24149	KIEL	NL	13	-	DE-NL
1424	K0424	ECC Germany	DE	24149	KIEL	NL	20	-	DE-NL
1425	K0425	ECC Germany	DE	24149	KIEL	NL	21	-	DE-NL
1426	K0426	ECC Germany	DE	24149	KIEL	NL	32	-	DE-NL
1427	K0427	ECC Germany	DE	24149	KIEL	NL	33	-	DE-NL
1428	K0428	ECC Germany	DE	24149	KIEL	NL	35	-	DE-NL
1429	K0429	ECC Germany	DE	24149	KIEL	NL	36	-	DE-NL
1430	K0430	ECC Germany	DE	24149	KIEL	NL	38	-	DE-NL

Figure 4: Example product with multiple lanes

The complete production process / RFQ processing cycle of the RFQ & Solutions desk is provided in figure 5. The five main steps will be explained.

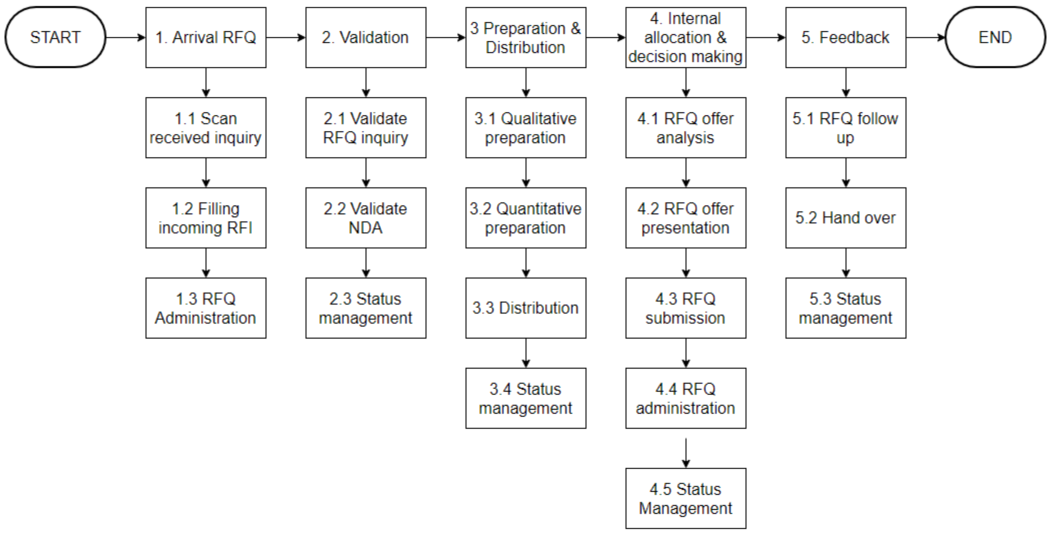

Figure 5: Production process RFQ & Solutions desk

4.1 Arrival RFQ

The first step in the production process is the arrival of a RFQ. In this step the registration of an incoming RFQ takes place. Furthermore, administration of the RFQ in the RFQ master registration file occurs in this step (referred to as the RM_File). The first step is executed by the RFQ & Solutions support team. The QESH department could support when information is missing when a request for information comes in.

4.2 RFQ validation

The following step is validation. During the validation phase, qualitative and quantitative aspects of the RFQ's are considered. In this step RFQ's can be declined, in case there is no fit with the Ewals product portfolio, send to one local cell (status referred to as Go it alone, see Appendix A) or go further in the RFQ processing cycle of the RFQ and Solutions Desk, as described below. These local cells are the locally present parties through Europe belonging to the Ewals Group. No mediation of the RFQ & Solutions desk is necessary if only one local cell is involved. This local cell can send an offer directly to the customer. This happens for example when the customer is a direct contact of the local cell. The RFQ desk keeps status management in all cases. If the RFQ goes further in the processing cycle of RFQ & Solutions Desk, then multiple local cells are involved, therefore, the RFQ & Solutions Desk takes the lead. Different types of statuses could be assigned for RFQ's in progress or closed, these statuses are provided in Appendix A. This validation step is within the scope of this project.

4.3 RFQ preparation & distribution

The third step in the production process is the preparation & distribution phase. If a RFQ is deemed interesting, then the further processing in the RFQ processing cycle will take place. The further preparation covers contractual details and additional qualitative and quantitative documentation in Ewals sheets for further processing and analysis. When the preparation is finished, the RFQ will be shared with the designated local cells of Ewals by the RFQ & Solutions Desk. These designated local cells are the ones, which are able to fulfil the customer requests and can be competitive on pricing. On this moment, the local cells could decide for themselves, if they are able to be competitive. The local cells invited by the RFQ & Solutions Desk during the internal distribution process are able to share an offer with the RFQ & Solutions Desk. By sharing the RFQ only with the designated local cells, it is easier to analyze the responses afterwards (invalid response: no reaction or valid response: declination (with reason code included) or an offer).

4.4 RFQ allocation & decision making

Step four is internal allocation & decision making. After sharing the RFQ with the local cells: the local cell declines participation or sends an offer for the shared RFQ. The RFQ desk will analyze all offers and prepare a proposal and shares this proposal with the Key Account Manager (KAM), responsible for the customer. The KAM will decide which internal parties will be included in the final offer for the customer, thus the KAM makes the final decision. After decision making the RFQ & Solutions Desk will share the proposal with the customer.

4.5 RFQ hand over

Step five is the follow up by the RFQ & Solutions Desk after getting feedback by the customer. When business is won the RFQ will be handed over to the operational units responsible for execution of the nominated business. When business is not won, the feedback will be documented and the designated shareholders will be informed. When the customer tendering exists out of multiple rounds, then the process will go back into the RFQ processing cycle after receiving feedback.

Now the complete process and terminology within Ewals are described. Furthermore, the problem to be investigated is described in detail. This completes the first phase of the problem cycle, namely the problem definition. In the next section, section 5, the available data within Ewals will be analysed to come up with input for the prediction model designed in section 6.

5 Analysis and diagnosis

5.1 Data understanding

The data available at Ewals is divided in multiple datasets. Within Ewals two levels of detail are distinguished: data on RFQ level and data on product level. Monetary or performance statistics could be measured on both levels. A RFQ can consist out of many products, as explained in section 4.

The datasets used in this research project are the Registration Master File and Product Performance File. All incoming RFQ's within Ewals are stored in the Registration Master file. The Registration Master File contains only information on RFQ level, no product or lane information can be found. The Registration Master File contains all RFQ's separated in rows.

The Product Performance File contains the performance on product level, thus performance on country relation level. This data is stored in the data warehouse of Ewals. This information can be linked to the RFQ level information, however the number of observations will decrease. The registration master file was already in use before the implementation of the data warehouse. Therefore, the information on RFQ and product level is analysed separately in section 5.2 and section 5.3.

For both data files the following information will be excluded. First, RFQ's which are still open/in progress should be excluded from the dataset. Open data can not be used for model training/testing. Furthermore, RFQ's which are closed immediately by for example bad previous experience will be excluded, no prediction is needed for these customers. Lastly, RFQ's which are conducted not following the default RFQ process will be excluded. Those are exceptions, which can influence the prediction.

5.2 RFQ level

First, the data will be described on the most general level, namely on RFQ level. The original Registration Master File contains 4621 RFQ's. After conducting the data cleaning as described above, 1431 RFQ's are left. For each RFQ, three numerical variables are stored. The first variable is current business. Current business is defined as the value of the transports conducted by Ewals in the year before the arrival of the new RFQ, while this business is again requested in the new RFQ for the coming year. The current business can be retained or lost. Besides of this current business, new business can be won. However, for most RFQ's, the current business variable has a value of zero, because Ewals is not yet conducting business requested in the RFQ.

The second variable is offered revenue. This variable represents the monetary value, which Ewals offered to the customer depending on the offered prices and the number of transports. The third variable is the assigned revenue. This variable describes the monetary value assigned to Ewals at the end of the procurement process. This value can also be zero, when the customer assigns no transports to Ewals. The statistics of these variables are provided in table 1.

Statistic	**Current Business**	**Offered Revenue**	**Assigned Revenue**
Mean	59.887	2.710.820	118.546
Min	0	158	0
Max	9.838.890	141.130.800	7.124.990
25%	0	79.632	0
50%	0	371.456	0
75%	0	1.675.670	0

Table 1: Summary statistics RFQ data in euros

As can be seen in table 1. The mean value of the variable current business is 59.887 euros. The average value of offered revenue to customers is 2.7 million euros and the average assigned revenue is 118.546 euros. The average value assigned to Ewals in RFQ's is thus much lower than the value offered in most cases.

The minimum value of current business is zero, which could indicate that this request is provided by a new customer. The minimum value of the assigned revenue is also zero, thus no business is assigned to Ewals in this case. The minimal value offered by Ewals in a RFQ is 158 euros.

The maximum value of business assigned to Ewals is 7.1 million eruos. The maximum value offered by Ewals is 141 million euros. The difference between the minimum and maximum value is really big. The maximum value of the current business variable is 9.8 million euros.

Lastly, looking at the quartiles, the 75% quartile of current business and assigned revenue are zero. Thus 75% of the observations is equal to zero. In most RFQ's in which Ewals participated no current business is found. Furthermore, in most RFQ's in which Ewals participated no business is assigned to Ewals.

Figure 6 shows the ten best customers after aggregation of the assigned business value in different RFQ's.

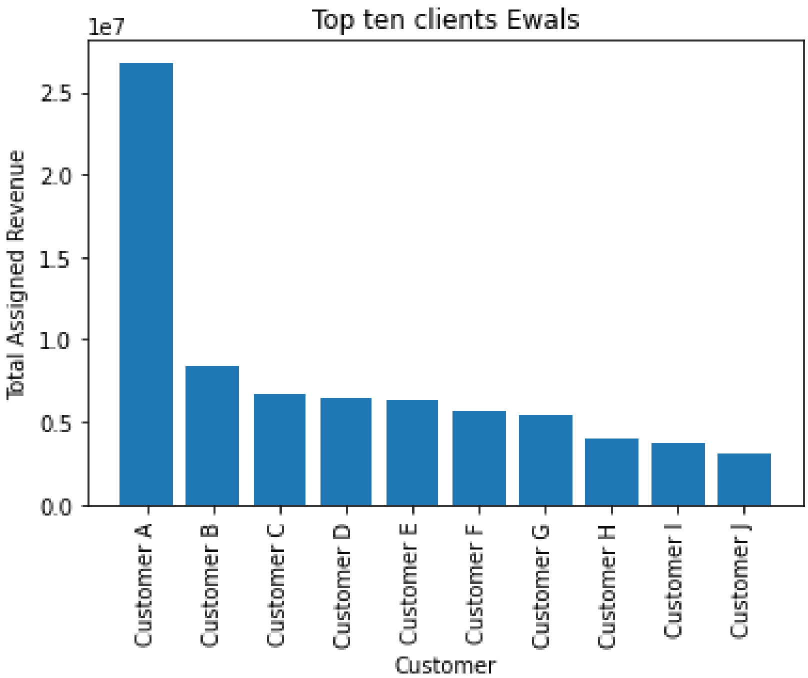

Figure 6: Best customers

Figure 7 provides, the individual RFQ's with the highest assigned value. As can be seen, the two RFQ's with the highest assigned value to Ewals belong to the same customer, which is also the best customer of Ewals overall, see figure 6.

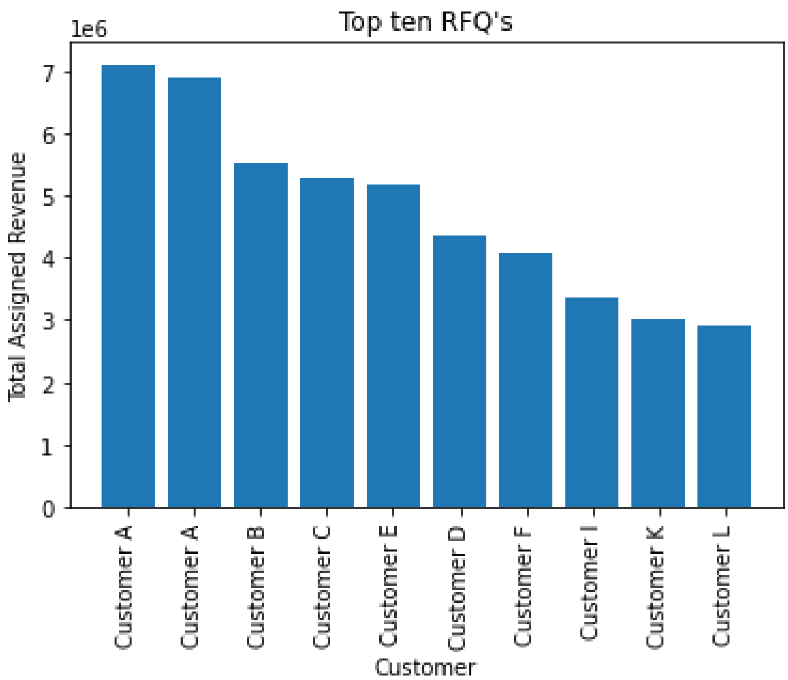

Figure 7: RFQ's with highest value

Figure 8 provides the bottom ten customers of Ewals, which have assigned business value to Ewals. The customers which have never assigned business to Ewals are excluded.

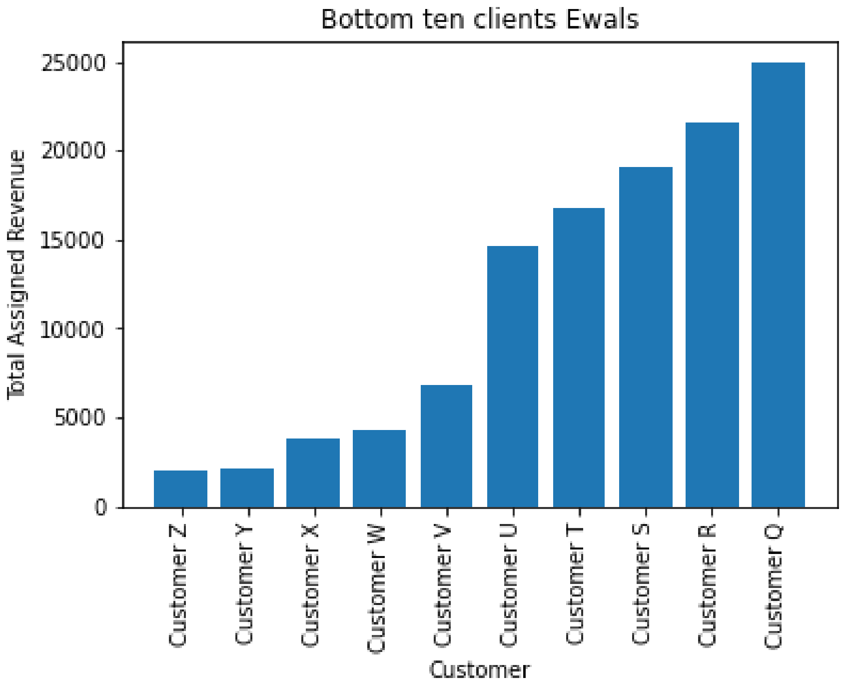

Figure 8: Bottom ten customers Ewals

Figure 9 shows the ten RFQ's in which the lowest value is assigned to Ewals. Again the RFQ's in which no value is assigned to Ewals are excluded. Looking at the values assigned to Ewals in figure 9, it could be that just one or a few transport trips are assigned to Ewals in one year. The most important conclusion drawn for this graphs is the difference in business value assigned. Not only the win or no win outcome of the tendering process is important, but also the value won.

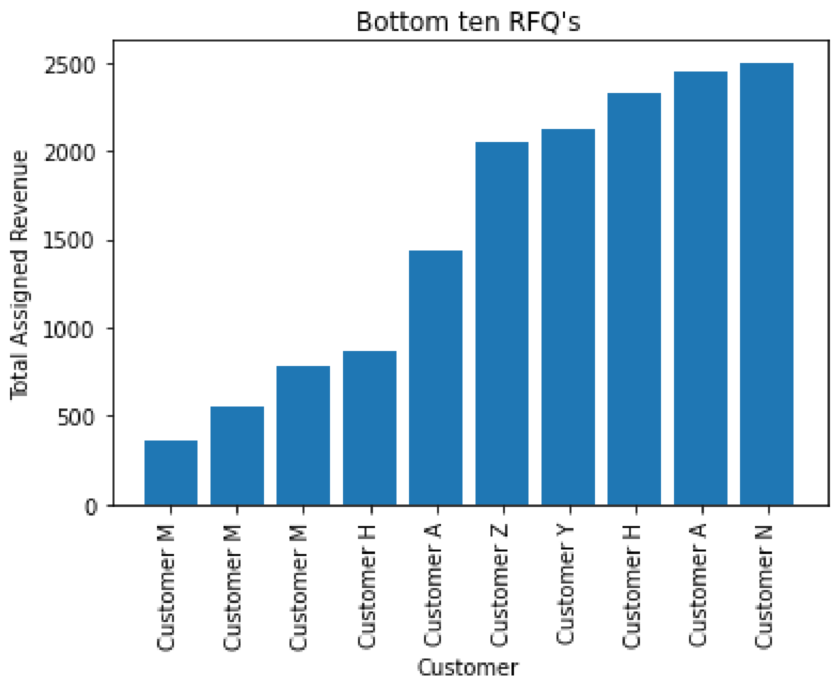

Figure 9: RFQ's with lowest value

Now, input variables for the prediction model will be searched and described. The experience of the employees of the RFQ & Solutions Desk is used to select the variables. Furthermore, the following reasoning is used to exclude variables. Before starting the preparation and analysis of a RFQ, the choice of possible participation in a RFQ (RFQ validation) should already by made. Therefore, the values of the input variables for the prediction model should already be known before the analysis starts. Referring back to the complete process of the RFQ & Solutions Desk, the analysis is step three in the process, see section 4.3, while the choice of participation is made in step 2, see section 4.2. The variable offered revenue could therefore not be used as input variable for the prediction model. The offered revenue value is known after analyzing the responses of the local entities in step four of the product process of RFQ & Solutions Desk, see section 4.4.

Furthermore, some variables are not suitable for the prediction model, because the number of categories keeps increasing. For example, a customer could be a indicator of winning a RFQ. When a good relationship with the customer exists, the customer will quicker do business with Ewals. However, still new customers are requesting transport service of Ewals, while new customers have few or no observations, so no prediction could be made.

Eventually six potential variables are selected from the data on RFQ level based on the above named criteria and the experience of employees making the decisions on this moment. These variables are Key Account Manager (KAM), RFQ received from, Full Truck Load (FTL)/Less Then Load (LTL), Stragetic End Markets (SEM), current business and high potential. An explanation about each variable will be provided in this section. Furthermore, for each variable a graph will be provided showing the proportion of business won/no business won for each value of the variable.

The first variable used for the model is the variable KAM, indicating if a Key Account Manager (KAM) is assigned to this customer. A customer with a KAM is a good contact, based on this relationship the probability of getting business assigned is higher. If regular contact exists between Ewals and the customer, an account manager is assigned, however only for strategic important customers a KAM is assigned. When first looking at the, two other categories are found, namely AM CZ and KAM CZ. Before the process alignment within Ewals, the local entity in the Czech Republic had its own process, therefore these RFQ's are excluded from the data. As can be seen in figure 10, more business is won for RFQ's, for which a KAM is assigned to the customer.

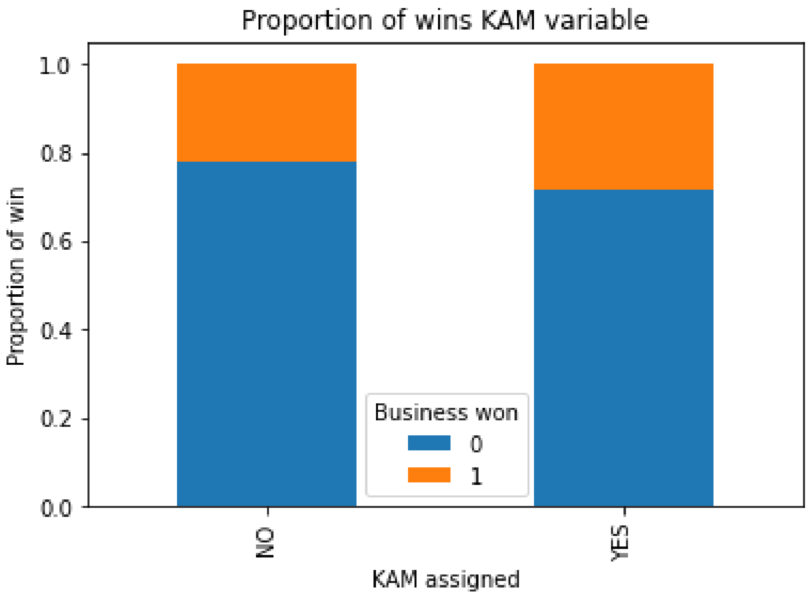

Figure 10: KAM variable

The second variable is RFQ Received From, this variable indicates how Ewals did receive the RFQ. As can be seen if figure 11, five groups can be distinguished: RFQ received direct from customer, RFQ received from an external source, RFQ received from KAM, RFQ received from local entity (office in Europe) and RFQ received from the sales team. A RFQ received directly, should have a higher winning probability, so this could be an indicator of winning business. As can be found in figure 11, the least business is won from RFQ's received from an external source.

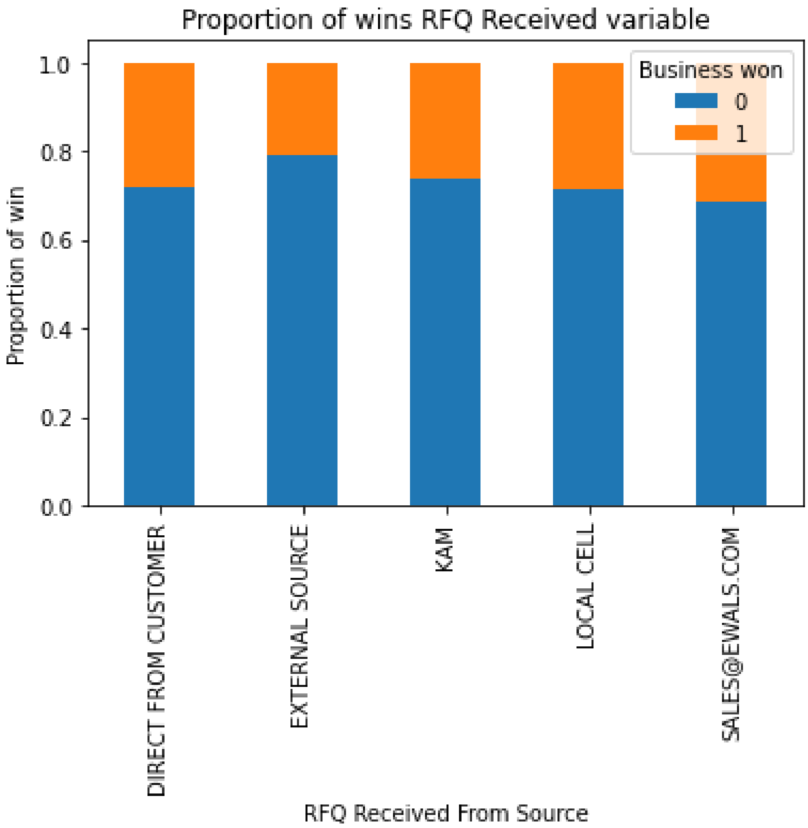

Figure 11: RFQ Received From variable

The third variable FTL/LTL indicates the kind of truck load, which can be found in the RFQ. Ewals respond most of the time on FTL RFQ's, which could be found more frequent in the data. Now RFQ's consisting of both FTL and LTL are always registered as two separate RFQ's in the registration master file, while before they could be registered as one. Variables not divided in two seperate RFQ's are therefore excluded from the data. Furthermore, rows can be found, whereby the registration is not yet conducted, these are excluded too from the dataset. As can be seen in figure 12, the proportion of wins for LTL business is higher than for FTL. As indicated before, Ewals participates mostly in RFQ's with requests for FTL transports. LTL requests are selected better and therefore, the proportion of win could be higher.

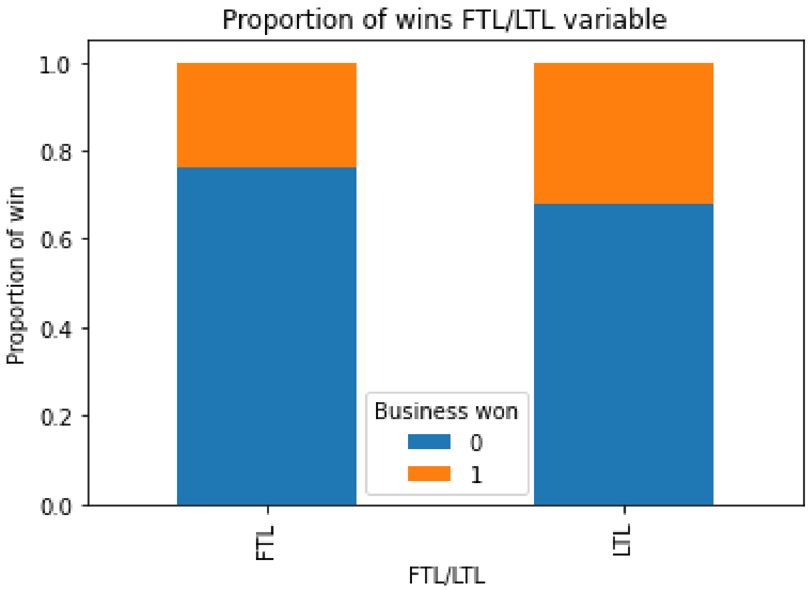

Figure 12: FTL or LTL variable

The fourth variable represents the Stragetic End Markets (SEM). Ewals has divided its business in groups based on the products transported. Ewals' trailers are better suited for transporting products from a certain market based on requirements obligated by the customer. Therefore, this could be an indicator for winning business. Fifteen SEM's are distinguished in the data. Looking a the SEM's, some of them have 100% or 0% win percentage. Taking a close look at the markets Aerospace, Automotive Aftersales, Fashion, Ecommerce and WasteRecycling, only a few observations could be found, which are exceptions or wrongly labeled RFQ's. Therefore, rows containing these SEM are deleted from the data. The remaining SEM's are provided in figure 13.

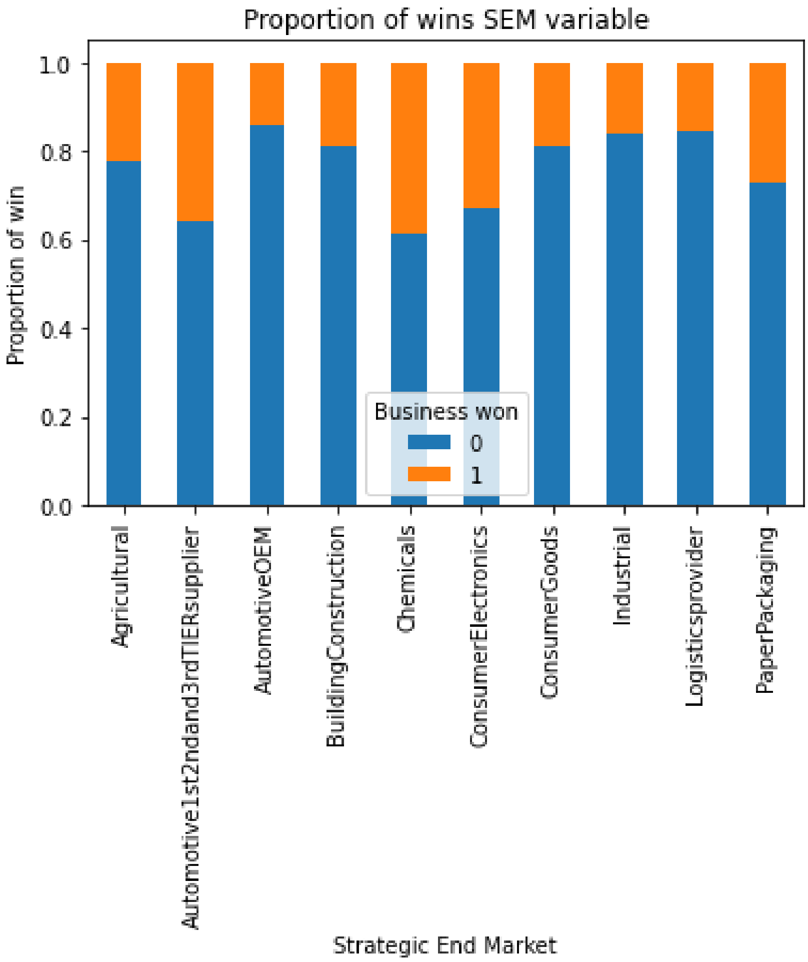

Figure 13: SEM variable

The fifth variable is current business which represents transport business within the new arrived RFQ executed by Ewals in the year before the arrival of the new RFQ. If Ewals is already conducting business, a certain relationship is already build, therefore current business could be a good predictor of winning business. As can be seen in figure 14, RFQ's with current business are won in most of the cases. This variable is a better indicator than customer, because this variable has only two values and deals with the change in relationship. The difference between the values yes and no is also really big, thus this could be a very strong predictor. For the model building in the next section this variable will therefore be used at the start of the model development.

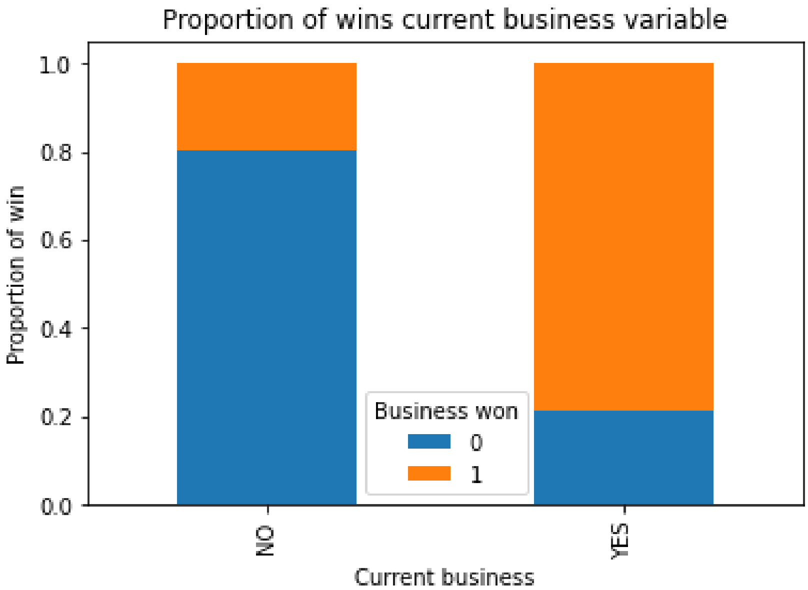

Figure 14: Current business variable

The last variable is high potential. Within Ewals customers are analyzed based on fit with Ewals' capabilities, market information and customer characteristics. This variable indicates if they see the RFQ's from a specific customer as high potential yes or no. As can be seen in figure 15, for RFQ's described as high potential a bit more RFQ's are won than for RFQ's not defined as high potential.

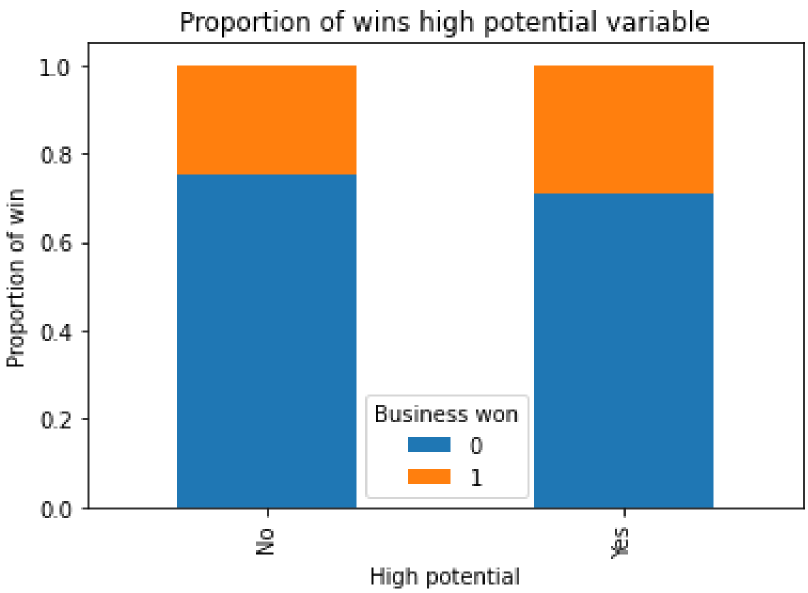

Figure 15: High potential variable

5.3 Product level

On product level, as defined before country relationship level, performance information is available. The product observations come from products found in the RFQ's in which Ewals has participated. This information is not on RFQ level, but on general lane level (product level), meaning not all specific lane level details are considered, only the country relationship of a lane. A product could therefore be present multiple times in the same RFQ. Figure 16 provides the number of observations of each product. The number of observations between products differs much. The products with a very high number of observations are not shown in this figure to better evaluate to other products.The top products will be discussed based on figure 19. The products most present in the data come above the 1000 observations. As can be seen, many products are present in the data, 843 products are found. The variety in number of observations could make it difficult to calculate certain performance statistics for all products.

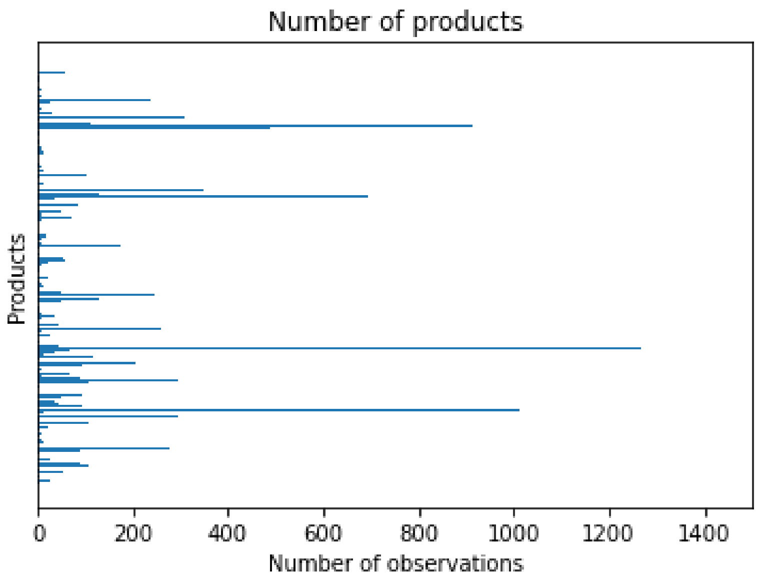

Figure 16: Overview product observations

Figure 17 provides the products in order from least frequent to most frequent observed. Most products have zero till two hundred observations, while even the biggest part seems to have less than fifty observations.

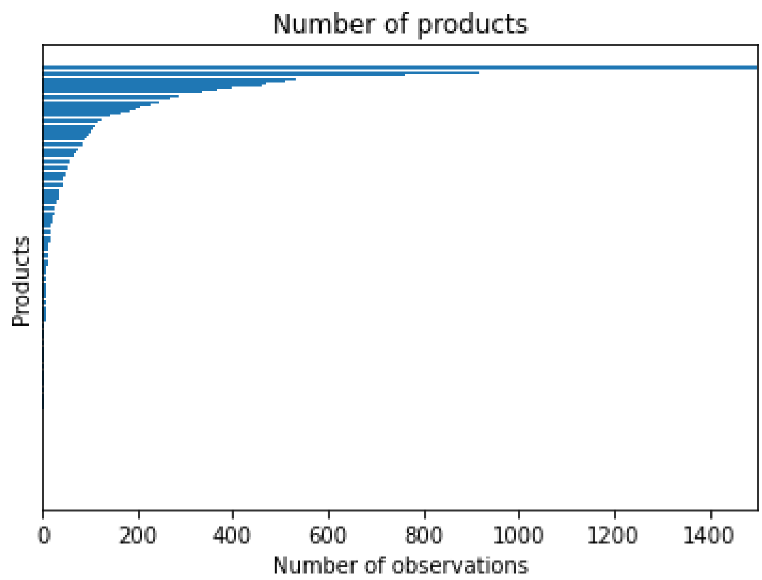

Figure 17: Overview product observations ascending

To get a better overview, the products will be distributed in bins, based on the number of observations. This provides a better overview in which bin, the most products can be found. Looking at figure 18 many products can be found in the bin with less than fifty observations. The influence of the number of observations on modeling should be viewed in more detail, because products with few observations could provide uncertain predictions.

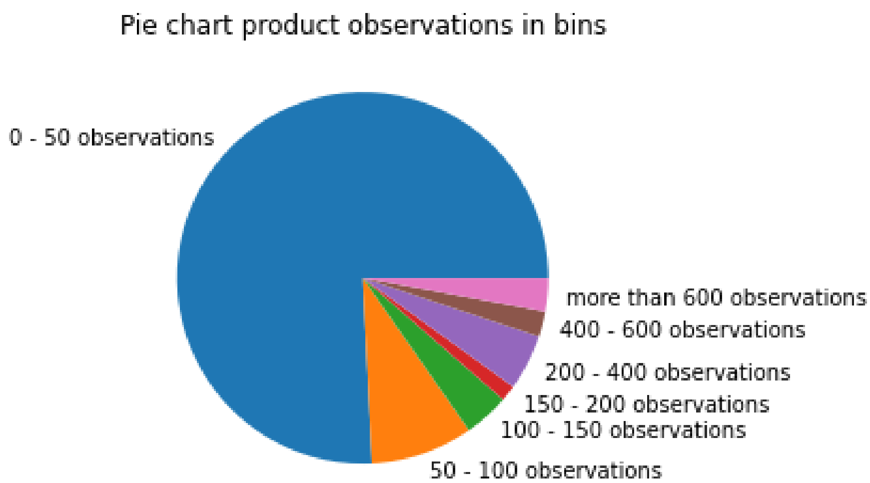

Figure 18: Overview product observations pie chart bins

Now the products most present in the data will be provided in figure 19. As can be seen most transports are conducted in domestic Germany. This product is far more present in the data as other products. However, the other most demanded product have also more than 1000 observations.

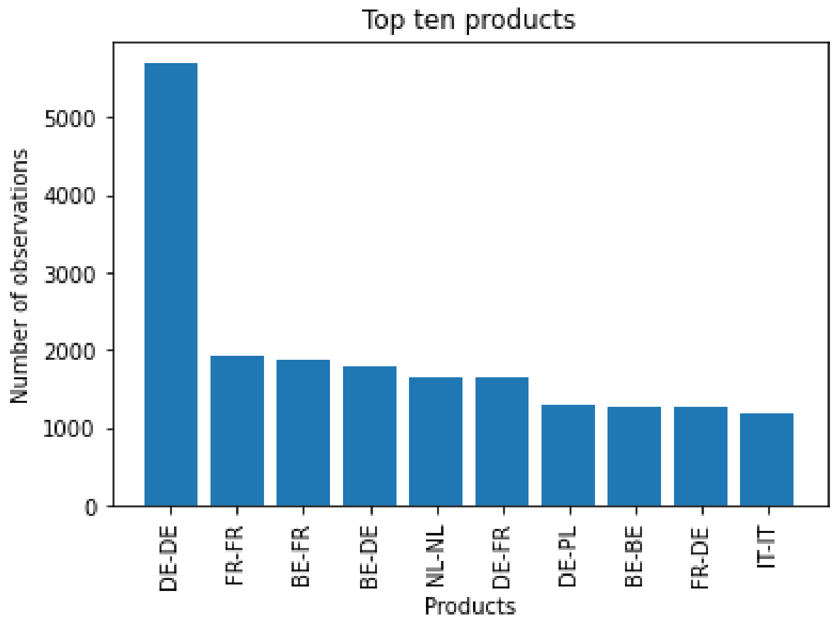

Figure 19: Most demanded products

As can be seen in figure 20, the products least demanded have all one observation. It would be difficult to make predictions based on one observation.

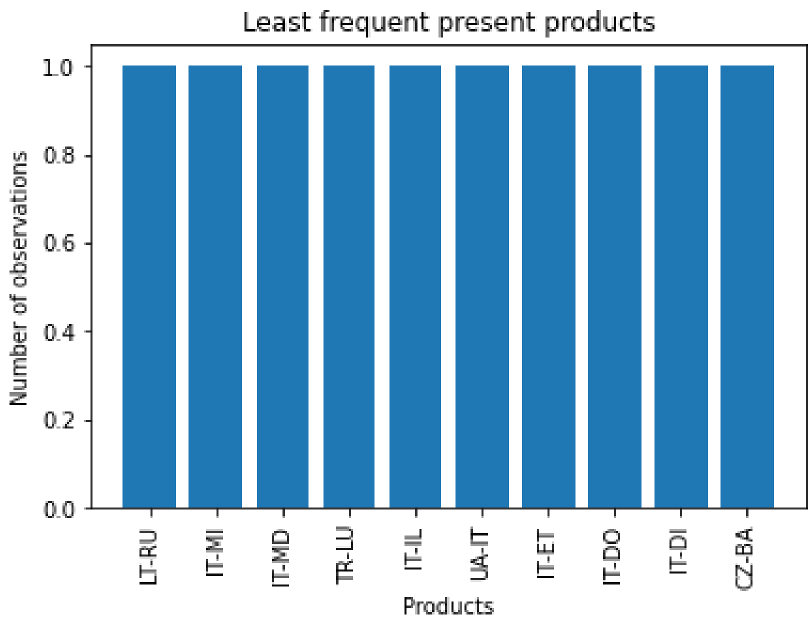

Figure 20: Least demanded products

Now the summary statistics of the numerical variables of the product data will be provided in table 2. The first variable is the percentage of the number of times an product is offered. Ewals does not offer on all products present in a RFQ. Thus, this variable provides the proportion of the number of times Ewals does offer. The second variable provided is the offered revenue, which provides the monetary value offered to customers for each product. This variable provides the total value offered for each product. The third variable provides the number of times a product is assigned to Ewals by customers, while the fourth variable represents the total monetary value assigned to Ewals for each product. The fifth variable provides the percentage of the number of times a products is assigned to Ewals from the number of times the products was offered to the customer. The last variable provides the percentage of assigned monetary value to Ewals from the value offered to the customer by Ewals.

Statistic	**Percentage of product offerings by Ewals**	**Offered value to customers**	**Number of times product assigned to Ewals**
Mean	25%	1.189.205	3
Min	0	0	0
Max	100%	66.958.644	187
Statistic	**Assigned value to Ewals**	**% number of times assigned to Ewals**	**% of monetary value assigned to Ewals**
Mean	68.631	11%	9%
Min	0	0	0
Max	3.098.546	100%	100%

Table 2: Summary statistics product level

As can be seen in table 2, the average product offer percentage is 25%. The average offered value for a product is 1.2 million euros. The average number of times a product is assigned to Ewals is three, while the average total product value assigned to Ewals is 68.631 euros. The average percentage of the number of times a product is assigned to Ewals from the number of times offered is 11%. The average percentage of the total monetary product value assigned to Ewals is a bit lower with a value of 9%. The value of 11% is lower than the RFQ hitrate of Ewals (26%), because a RFQ is a winning RFQ if the number of lanes assigned to Ewals is higher than zero, while the performance here is measured for each product (general lane level) separated.

All minimum values provided in table 2 are equal to zero. For the first variable, this means that products are present in the RFQ's in which Ewals participated, however Ewals did not offer on these specific products (zero is provided if no percentage can be calculated). Thus for these products the offered value to customers is also zero. The variable representing the number of times products assigned to Ewals could also be zero, in that scenario Ewals offered the product to the customer, but never got business assigned for this specific product. For the last two variables also a value of zero is provided if no percentage can be calculated.

The maximum value for the percentage of products offerings by Ewals is 100%. Ewals provided always an offer for this product. The maximum total value offered for a product is 67 million euros. Furthermore, the best performing product is assigned 187 times to Ewals and the maximum total assigned value assigned to Ewals is 3.1 million. The maximum values of the last two percentages are 100%. However, this can also be that for this product Ewals participated one time and Ewals got the business assigned for that one time. The values could be misleading when the number of observations is low.

As can be seen in table 3, the total number of products is 863. The total number of shipments offered by Ewals is 996.185. The average number of observations for products is 84, while the average number of offered shipments for the different products is 1155. As seen before, the product which is observed the most times is the product DE-DE, which is observed 5693 times. The total highest number of shipments offered for a product is 107.135.

Statistic	Number of Products	Number of shipments
Total	863 Unique products	996185 Shipments
Mean	84	1155
Min	1	1
Max	5693	107.135

Table 3: Summary statistics 2 product level

As indicated at the start of section 5, the data on product level could be combined with the data on RFQ level. Products differ on performance (percentage of number of times/value assigned to Ewals), thus the presence of a good performing product in the RFQ could predict winning business. However, as indicated before the high diversity in number of observations could make it difficult to include all products in a prediction model. Furthermore, the difference in number of observations of the product data can provide uncertainty. A second possibility is to include delivery and collection country separately, because a combination of these two is a product. Other possible model variables are the number of products and the number of shipments in a RFQ. As explained before, first a model will be designed with the variables on RFQ level in section 6.2.2, because of the higher number of observations, which can be used for model building. Afterwards, the model can be extended with the product level variables in section 6.2.4. First, the outcome variable of the model will be described in section 5.4.

5.4 Output variable

The variable which should be predicted is the binary variable business won yes or no. This variable has the value 1 (yes), as the monetary value assigned to Ewals in a RFQ is higher than zero, thus a lane within the RFQ is awarded. As can be seen in figure 21, the number of observations of winning RFQ's is much lower than the number of observations of not winning RFQ's.

Following Ramyachitra and Manikandan (2014), this type of imbalance has become a critical problem within machine learning, because the minority class is more important and this minority class will be ignored during the predictions. One of the solutions proposed by Ramyachitra and Manikandan (2014) is the Synthetic Minority Over-sampling Technique (SMOTE). This solution can be conducted on data level and can be used flexible. Synthetic Minority Over-sampling Technique (SMOTE) is known as the best oversampling technique. The minority class is over sampled with synthetic examples instead of oversampling by alternation. This technique will be conducted to balance the dataset.

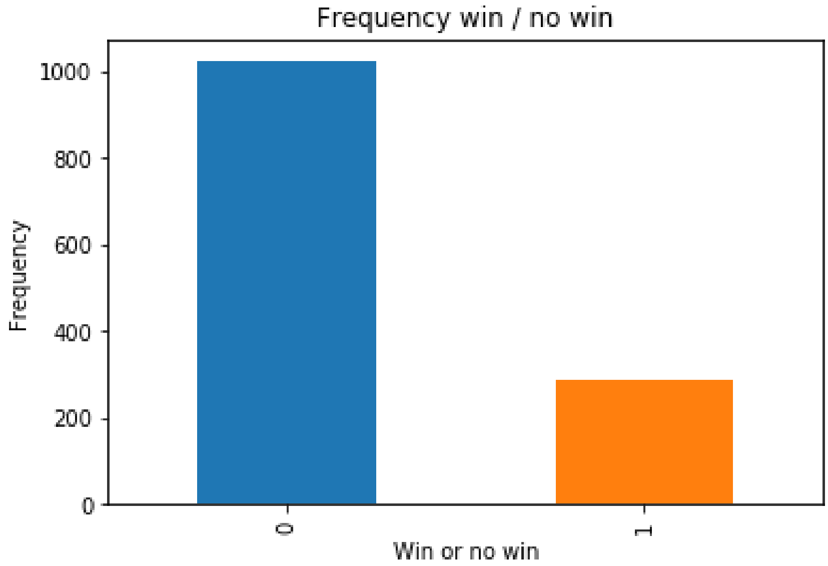

Figure 21: Win or no win overview

6 Solution design

6.1 Model selection

First, prediction models must be chosen for this research. The outcome variable of the prediction model is a binary variable, therefore binary classification models should be evaluated. K.Kirasich, Smith, and Sadler (2018) compare the performance of a logistic regression model and a random forest model for binary classification. They conclude that the logistic regression model scores better on overall accuracy than the random forest model when the variance of the explanatory variables is changed. The random forest model scores better on the false positive rate, while no significant difference can be found for the true positive rate. Therefore, the logistic regression model will be used.

Dreiseitl, Ohno-Machado, and Ohno-Machado (2002) evaluate the use of a logistic regression model and a Artificial Neural Network (ANN), because those two are mostly chosen as prediction model. They found seventy two papers comparing the ANN and logistic regression model. In most cases the ANN model outperformed the logistic regression model. However, the model building of a logistic regression is easier and the model variables can be interpreted, which makes the model more popular than the ANN, which is a black box model. First, a logistic regression model will be designed to evaluate the model parameters, afterwards a ANN model will be designed with the same model parameters to find the better performing model.

Lastly, the Support Vector Machines (SVM) model is chosen as third model. This model is indicated as reliable model for classification tasks by Gaspar, Carbonell, and Oliveira (2012). Besides different kernels functions can be used to separate the data easily, which provides flexibility.

6.2 Model design Logistic Regression

6.2.1 Model metrics Logistic regression

The next step is selecting performance metrics for comparing the models with different input parameters. The first performance metric is the Akaike's entropy-based Information Criterion (AIC). This metric is used relatively, a lower AIC score indicates a better model fit. The AIC parameter has a low penalty for complex models, therefore the AIC metric depends more on the performance on the training dataset, more complex models are selected earlier. An second performance metric used is the Bayesian Information Criterion (BIC) metric, which is used to prevent overfit. In contradiction to the AIC, the BIC assigns a higher penalty to complex models to prevent overfitting.

The AIC and BIC metric are calculated as follows:

$AIC = \frac{-2}{N} * LL + 2 * \frac{k}{N}$
N = number of samples in the training dataset

LL = log-likelihood of the model on the training dataset
k = number of parameters in the model

$BIC = -2 * LL + log(N) * k$

The AIC and BIC metrics indicate that the addition of a new variable leads to a higher model fit or not. A lower value indicates an increase in model fit, which is thus a recommendation for adding the variable to the model.

6.2.2 Variable selection

First the model fit score of the variables found in the Registration Master file are measured for each variable separately in table 4. In section 5.2 it seemed that the variable current business could be the strongest variable based on the big difference in RFQ's won for RFQ's in which current business is found in comparison with RFQ's with no current business. The six variables involved are current business in the new RFQ, Full Truck Load (FTL)/LTL, Key Account Manager (KAM) customer, RFQ typified as high potential, RFQ received from and Stragetic End Markets (SEM). The first four variables can be added as one variable. For the variables RFQ received from and SEM a group of variables will be added.

Model	**Variable**	**Beta**	**AIC**	**BIC**	**P-value**
Model 1	Current Business	2.36	1565	1570	P <0.01
Model 2	FTL/LTL	-0.1103	1701	1706	P <0.0766
Model 3	KAM	0.2663	1696	1701	P <0.01
Model 4	High Pot	-0.2877	1704	1708	P <0.4513
Model 5	RFQ received from		1703	1729	
	Direct from customer	0.0451			P <0.7948
	From external source	-0.1379			P <0.1264
	From KAM	0.0547			P <0.6012
	From locall cell	0.2151			P <0.1095
	From general sales	0.1311			P <0.4255
Model 6	SEM		1635	1686	
	Agricultural	-0.2877			P <0.2585
	Automotive 1st, 2nd, and 3rd Tier Supplier	0.4940			P <0.01
	Automotive (OEM)	-0.7340			P <0.01
	Building & Construction	-1.2528			P <0.0271
	Chemicals	1.0245			P <0.01
	Consumer Electronics	0.5653			P <0.0240
	Consumer Goods	-0.3665			P <0.0278
	Industrial	-0.3102			P <0.1760
	Logistics Provider	-0.5878			P <0.01
	Paper & Packaging	0.2693			P <0.1326

Table 4: Overview one variable models

As can be seen in table 5, the strongest predictor with the best model fit is the variable current business. This variable has a significant positive influence on the outcome of the RFQ participation. When current business is present in the RFQ the probability of getting value awarded is higher. In table 5 variables will be combined searching for the best model, starting with the variable current business. For the following design steps, each individual variable is added. The variable, which provides the highest difference in model fit will be added. These step will be repeated till the best model is found.

Model	**Cur bus**	**SEM**	**KAM**	**FTL**	**High Potential**	**Received From**	**AIC**	**BIC**
Model 1	X						1565	1570
Model 7	X	X					1511	1567
Model 8	X		X				1567	1577
Model 9	X			X			1534	1544
Model 10	X				X		1564	1574
Model 11	X					X	1546	1577
Model 12	X	X	X				1496	1558
Model 13	X	X		X			1513	1575
Model 14	X	X			X		1506	1567
Model 15	X	X				X	1486	1568
Model 16	X		X	X			1533	1549
Model 17	X			X	X		1540	1555
Model 18	X			X		X	1522	1558
Model 19	X	X	X	X			1495	1561
Model 20	X	X	X		X		1486	1552
Model 21	X	X	X			X	1440	1527
Model 22	X	X		X		X	1490	1577
Model 23	X	X			X	X	1495	1582
Model 24	X	X	X	X		X	1426	1518
Model 25	X	X	X		X	X	1440	1532
Model 26	X	X	X	X	X	X	1397	1494

Table 5: Model fit overview

As can be seen in table 5, model 6 is the best scoring model on model fit.

6.2.3 Multicollinearity

Following Midi, Sarkar, and Rana (2010), the phenomenon where the prediction variables of a logistic regression are correlated or associated is called multicollinearity. In this situation, strong correlation could be found between two or more explanatory variables. This strong correlation causes the coefficients and p-values to be unstable. Small changes in the dataset or the model design can drastically change the values of the coefficients and p-values. Perfect correlation exists when, a correlation with a value of 1 or -1 is found between variables. An important note about multicollinearity is provided by Midi et al. (2010), the reliability or statistical power of the entire model does not decrease by multicollinearity, however coefficients or p-values calculated for individual predictors are influenced by multicollinearity. Only, the influence of the set of variables on the model can be found in a model with multicollinearity. Multicollinearity is caused by use of variables depending on each other or by using variables which measure the same thing.

As first check for multicollinearity, the correlation matrices are provided. A rule of thumb provided by Midi et al. (2010) is to search for a correlation value greater than 0.8 or below -0.8. These values indicate that you have a problem caused by multicollinearity. The correlation matrix for model 6 is provided in figure 22. No correlation values above 0.8 or below -0.8 can be found.

	KAM	FTL	Cur Bus	High Pot	Agricult	Auto Tier	Auto OEM	Build Cons	Chemic	Cons El	Cons Good	Industr	Logi Provi	Paper	Customer	External	From KAM	Local Cell	Sales
KAM	1.00	-0.09	0.10	-0.07	-0.19	0.25	0.58	-0.12	-0.12	0.07	-0.32	-0.17	-0.23	-0.20	-0.06	0.21	0.08	-0.31	-0.05
FTL	-0.09	1.00	-0.05	0.02	0.01	-0.24	0.15	0.01	-0.05	0.05	0.04	-0.03	0.07	0.04	0.01	-0.15	0.08	0.09	-0.04
Cur Bus	0.10	-0.05	1.00	0.02	0.00	0.18	-0.08	0.00	-0.01	-0.02	-0.06	-0.06	-0.03	0.01	0.01	0.03	0.02	-0.06	-0.03
High Pot	-0.07	0.02	0.02	1.00	-0.03	-0.05	-0.07	-0.02	0.01	0.14	0.06	0.01	0.00	0.02	-0.05	-0.03	-0.01	0.08	-0.02
Agricult	-0.19	0.01	0.00	-0.03	1.00	-0.15	-0.12	-0.03	-0.04	-0.05	-0.10	-0.06	-0.07	-0.07	-0.06	-0.06	0.00	0.11	0.04
Auto Tier	0.25	-0.24	0.18	-0.05	-0.15	1.00	-0.31	-0.09	-0.10	-0.14	-0.25	-0.16	-0.18	-0.20	-0.09	0.30	-0.09	-0.19	-0.01
Auto OEM	0.58	0.15	-0.08	-0.07	-0.12	-0.31	1.00	-0.07	-0.08	-0.11	-0.20	-0.12	-0.14	-0.15	-0.16	0.01	0.22	-0.14	-0.02
Build Cons	-0.12	0.01	0.00	-0.02	-0.03	-0.09	-0.07	1.00	-0.02	-0.03	-0.06	-0.04	-0.04	-0.04	-0.03	-0.01	-0.01	0.05	-0.02
Chemic	-0.12	-0.05	-0.01	0.01	-0.04	-0.10	-0.08	-0.02	1.00	-0.04	-0.06	-0.04	-0.05	-0.05	-0.02	-0.03	0.03	0.01	-0.02
Cons El	0.07	0.05	-0.02	0.14	-0.05	-0.14	-0.11	-0.03	-0.04	1.00	-0.09	-0.06	-0.06	-0.07	0.17	-0.14	0.03	-0.01	0.04
Cons Goods	-0.32	0.04	-0.06	0.06	-0.10	-0.25	-0.20	-0.06	-0.06	-0.09	1.00	-0.10	-0.12	-0.13	0.32	-0.14	-0.05	-0.01	0.00
Industr	-0.17	-0.03	-0.06	0.01	-0.06	-0.16	-0.12	-0.04	-0.04	-0.06	-0.10	1.00	-0.08	-0.08	-0.09	-0.08	-0.02	0.20	-0.03
Logi Provi	-0.23	0.07	-0.03	0.00	-0.07	-0.18	-0.14	-0.04	-0.05	-0.06	-0.12	-0.08	1.00	-0.09	0.01	-0.13	0.06	0.09	-0.01
Paper	-0.20	0.04	0.01	0.02	-0.07	-0.20	-0.15	-0.04	-0.05	-0.07	-0.13	-0.08	-0.09	1.00	0.00	0.06	-0.14	0.08	0.04
Customer	-0.06	0.01	0.01	-0.05	-0.06	-0.09	-0.16	-0.03	-0.02	0.17	0.32	-0.09	0.01	0.00	1.00	-0.27	-0.23	-0.16	-0.04
External	0.21	-0.15	0.03	-0.03	-0.06	0.30	0.01	-0.01	-0.03	-0.14	-0.14	-0.08	-0.13	0.06	-0.27	1.00	-0.54	-0.38	-0.09
From KAM	0.08	0.08	0.02	-0.01	0.00	-0.09	0.22	-0.01	0.03	0.03	-0.05	-0.02	0.06	-0.14	-0.23	-0.54	1.00	-0.32	-0.08
Local Cell	-0.31	0.09	-0.06	0.08	0.11	-0.19	-0.14	0.05	0.01	-0.01	-0.01	0.20	0.09	0.08	-0.16	-0.38	-0.32	1.00	-0.05
Sales	-0.05	-0.04	-0.03	-0.02	0.04	-0.01	-0.02	-0.02	-0.02	0.04	0.00	-0.03	-0.01	0.04	-0.04	-0.09	-0.08	-0.05	1.00

Figure 22: Correlation matrix model 6

Secondly, the Variance Inflation Factor (VIF) is used for searching the presence of multicollinearity. The VIF score indicates if the variance of the coefficients are influenced by multicollinearity. Senaviratna and Cooray (2019) indicate that a VIF value above ten indicates the presence of multicollinearity, in weaker models like logistic regression it could be that for lower values multicollinearity is present. Robinson and Schumacker (2009) indicate that as general rule can be used that VIF values should not exceed the value ten. In case the VIF value exceeds ten, then the following solutions are provided for multicollinearity. In some cases, the variables with correlation could be combined into one variable. If they could not be combined into one variable, one of the variables should be dropped. No theoretical ground is available to choose, which variable to delete from the model. Before omitting variables, data could be added to decrease multicollinearity. Another solution is using the model with the multicollinearity included. In some situations it is not possible to delete the multicollinearity completely, the model can still be used after carefully considering the variables. Each variable should measure different input.

Variable	**VIF-value**
KAM	2.66
FTL	1.11
Current business	1.06
High Potential	1.05
Agricultural	5.29
Automotive Tier	18.77
Automotive OEM	15.58
Building & Construction	2.64
Chemicals	3.11
Consumer Electronics	4.85
Consumer Goods	11.34
Industrial	5.81
Logistics Provider	7.19
PaperPackaging	7.83
From customer	11.08
External source	38.05
From KAM	30.98
From local cell	17.51
From Sales	2.20

Table 6: VIF values model variables

As can be seen in table 6, VIF values above ten are found. A remarkable observation is that all VIF values from the group RFQ received from are above ten. Therefore, this variable will be deleted. Table 7 provides the new VIF values and table 8 provides the model fit.

Variable	**VIF-value**
KAM	4.16
FTL	7.21
Current business	1.15
High Potential	1.05
Agricultural	1.36
Automotive Tier	3.62
Automotive OEM	4.87
Building & Construction	1.13
Chemicals	1.12
Consumer Electronics	1.56
Consumer Goods	1.95
Industrial	1.35
Logistics Provider	1.61
PaperPackaging	1.64

Table 7: VIF values with the variable RFQ received from deleted

As can be seen in table 7, no VIF values above ten are found. However, the VIF value of the variable FTL is above five. Furthermore, the VIF values of the KAM, and two Automotive SEM markets are near five. However, an explanation for this can be provided. Ewals is doing most of its business in the Automotive market. Furthermore, most of its better customers are active in this market. Therefore, more KAM customers come from this market. Lastly, most automotive customers prefer FTL Loads. Thus, that explain the higher VIF values, however the variables measure other input. Using model twenty seven, is therefore acceptable, see table 8.

Model	**Cur bus**	**SEM**	**KAM**	**FTL**	**High Potential**	**Received From**	**AIC**	**BIC**
Model 27	X	X	X	X	X		1486	1557

Table 8: Model fit without variable RFQ received from

6.2.4 Adding more variables

As explained before, coming up with model twenty seven in the previous chapter, the Registration Master File is used as data source. Combining this data with the product data in the data warehouse, more variables can be added. However, one main disadvantage is the decrease in available data. The storage of data in the data warehouse is implemented later, therefore after cleaning the data (closed data as before) only 319 RFQ observations are left. Model twenty seven is taken as start model and new variables are added in this model, see table 9.

The first two new variables introduced are the number of shipments in a RFQ and the number of lanes in a RFQ. These variables are related to the possible monetary value in a RFQ. Secondly, three variables will be introduced related to the countries involved in the transports. These variables will be grouped variables, the groups consist of all countries present in the RFQ. These variables will be added as follows; X_i (which equals 1 if country i is present in the RFQ otherwise 0). The variables are collection country, delivery country and country relationship (which equals 1 if country relationship i is present in the RFQ otherwise 0).

Before adding the country variables to the model, additional cleaning is needed. Country variables, which are nearly not present in the data must be deleted, because these variables will provide infinite VIF values or provide predictions with very high standard errors / uncertainty. At the start of the cleaning process, 37 collection countries are found, after cleaning 20 collection countries are left. For delivery countries, first 49 countries are available, after cleaning 30 countries are added in the model. Lastly, for the country relationship variable, first 318 country combinations are present. After cleaning, 31 country combinations are left.

Model	Model 27	shipments	lanes	Collection	Delivery	Country Rel	AIC	BIC
Model 27.2	X						297	345
Model 28	X	X					270	322
Model 29	X		X				277	329
Model 30	X	X	X				267	323
Model 31	X			X			280	399
Model 32	X				X		284	438
Model 33	X					X	305	458
Model 34	X	X	X	X			272	397
Model 35	X	X	X		X		277	438

Table 9: Model fit with additional variables

As can be seen in table 9, adding the country variables leads to an increase in model complexity based on the increasing BIC values. Model thirty has the best model fit scores, thus it seems to be the best model in the new scenario. Now, the VIF scores will be evaluated for model thirty, see figure 10. No VIF values above ten can be found. The VIF value of FTL of six is the highest value, which could again indicate the presence of multicollinearity. In the next section, the parameters and model performance will be evaluated in more detail.

Variable	VIF-value
KAM	3.75
FTL	6.03
Current business	1.35
High Potential	1.19
Number of shipments	3.35
Number of lanes	3.53
Agricultural	1.38
Automotive Tier	4.20
Automotive OEM	2.32
Building & Construction	1.31
Chemicals	1.39
Consumer Electronics	1.57
Consumer Goods	1.79
Industrial	1.51
Logistics Provider	1.63
PaperPackaging	1.60

Table 10: VIF values model 30

6.2.5 Logistic regression

Figure 11 provides the betas and p-values of the model variables of model twenty seven and model thirty. As can be seen, insignificant values can be found, however deleting insignificant values lead to major changes of the values of the other variables, which indicates that multicollinearity still is present. However, this is acceptable, because the input variables measure different information as explained above. As stated before, the reliability or statistical power of the entire model does not decrease by multicollinearity, however coefficients or p-values calculated for individual predictors are influenced by multicollinearity and these can not be interpreted. Therefore, the performance of the models is evaluated in table 12.

	Model 27		Model 30	
Variable	**Beta**	**P-value**	**Beta**	**P-value**
Current Business	2.39	P <0.01	2.71	P <0.01
FTL	-0.46	P <0.01	1.62	P <0.01
KAM	1.10	P <0.01	1.20	P <0.01
High Pot	-1.02	P <0.05	-0.28	P <0.80
Number of lanes			0.002	P <0.36
Number of shipments			-0.000	P <0.80
Agricultural	-0.04	P <0.88	-2.33	P <0.01
Automotive 1st, 2nd, and 3rd Tier Supplier	-0.26	P <0.27	-2.09	P <0.01
Automotive (OEM)	-1.58	P <0.01	-3.09	P <0.01
Building & Construction	-1.03	P <0.10	-1.97	P <0.05
Chemicals	1.31	P <0.01	-1.84	P <0.03
Consumer Electronics	0.22	P <0.53	-2.85	P <0.01
Consumer Goods	-0.08	P <0.71	-3.32	P <0.01
Industrial	0.03	P <0.92	-3.32	P <0.01
Logistics Provider	-0.32	P <0.26	-3.45	P <0.01
Paper & Packaging	0.36	P <0.15	-3.04	P <0.01

Table 11: Final overview variables logistic regression

The available data is split in 80% training data and 20% test data. The test data is used to calculate the performance of the prediction models. The precision measures the number of correct positive predictions. The recall measures the number of correct positive prediction predictions as proportion of all positive cased in the data. The F1-score combines the precision and the recall to one value.

As can be seen in figure 12, the accuracy of model twenty seven is higher. Furthermore, the F1-score of model twenty seven is higher than the score of model thirty. For the logistic regression model, model twenty seven is the best performing model.

	Test accuracy	**F1 score**	**Precision**	**Recall**
Model 27	68%	65%	71%	59%
Model 30	63%	62%	61%	63%

Table 12: Performance overview logistic regression

6.3 Neural Network

As can be seen in table 13, the accuracy of model twenty seven is higher than the accuracy of model thirty. In contradiction to the accuracy, the F1-score for model thirty is higher than the score of model twenty seven. Model twenty seven has more positive correct prediction observations. Based on the bigger difference in accuracy, model twenty seven is the best performing model. Comparing to the logistic regression model. The logistic regression is performing better on both the accuracy and the F1-score metric. Input parameters for the model are:

- Hidden layer sizes: (100), representing one hidden layer with 100 units in this layer.
- Activation function: rectified linear unit function (relu)
- Solver: stochastic gradient based optimizer proposed by Kingma and Ba (2014) (adam solver)
- Alpha (penalty parameter): 0.0001
- Learning rate: constant

	Test accuracy	**F1 score**	**Precision**	**Recall**
Model 27	67%	64%	69%	60%
Model 30	57%	67%	54%	87%

Table 13: Overview performance Neural Network

Now a grid search will be conducted for model twenty seven to find the best parameters for a Neural Network model. The following input parameters and values will be used in the evaluation:

- Hidden layer sizes: (50) /(100) / (100,100) / (200) / (200,200) / (100,100,100)
- Activation function: hyperbolic tan function (tanh) / rectified linear unit function (relu) / logistic sigmoid function (logistic)
- Solver: stochastic gradient descent (sgd) / stochastic gradient based optimizer (adam solver) / optimizer in quasi newton methods (lbfgs)
- Alpha: 0.0001, 0.001, 0.05
- Learning rate: constant / adaptive / invscaling

The best combination of parameters found for these input variables are:

- Hidden layer sizes:(100,100)
- Activation function:logistic function

- Solver: stochastic gradient based optimizer proposed (adam)
- Alpha: 0.0001
- Learning rate: constant

An overview of the performance metrics of the final Neural Network model is provided in table 14.

	Test accuracy	**F1 score**	**Precision**	**Recall**
Model 27	67%	65%	68%	62%

Table 14: Overview performance optimized Neural Network

Still, the accuracy is slightly less than the accuracy of the logistic regression model.

6.4 SVM Model

Three methods are available within Support Vector Machines (SVM) for binary classification, namely Support Vector Classification (SVC), Nu-SVC and linear SVC. The SVC and NuSVC methods differ slightly. Those methods have other mathematical approximations and SVC is based on parameter C for misclassification, while NuSVC is based on the parameter Nu, which represents the upper bound of the margin errors. The SVC and NuSVC are more complex and for those non-linear kernels are available. The linear SVM has as main assumption linearity.

As named before kernel functions could be used to include degrees of non linearity. Those kernel functions are mathematical functions, which conduct a transformation on the data points provided as input. For the Support Vector Classification (SVC) models, the following kernel functions will be used and compared on accuracy: linear, polynomial, sigmoid and a Guassian radial basis function (rbf). For both model twenty seven and model thirty, the available kernels will be evaluated to find the best suiting kernel for this dataset.

SVM Model	**Test accuracy**
SVC rbf	67%
SVC poly	68%
SVC linear	66%
SVC sigmoid	56%
NuSVC rfb	52%
NuSVC poly	59%
NuSVC linear	41%
NuSVC sigmoid	47%
Linear SVM	67%

Table 15: Model 27 SVM

As can be seen in table 15, model twenty seven performs best using a SVC poly SVM model.

SVM Model	**Test accuracy**
SVC rbf	49%
SVC poly	48%
SVC linear	61%
SVC sigmoid	48%
NuSVC rfb	49%
NuSVC poly	51%
NuSVC linear	62%
NuSVC sigmoid	49%
Linear SVM	48%

Table 16: Model 30 SVM

As can be seen in table 16, model thirty performs best using a NuSVC linear SVM model. Table 17 provides the model performance metrics for the SVM models.

For the SVM model, the accuracy of model twenty seven is also higher than the accuracy of model thirty. However, the F1-score of model thirty is higher than the score of model twenty seven. However, the difference is quite big between precision and recall. To get a more balanced model, model twenty seven will be a better choice.

	Test accuracy	**F1 score**	**Precision**	**Recall**
Model 27	68%	65%	69%	62%
Model 30	61%	68%	58%	83%

Table 17: Performance overview SVM

The logistic regression model and SVM model score the same values on both accuracy and F1-score. Both models can be used as final model. The most important conclusion to be made is that model twenty seven is a better prediction model than model thirty in all cases. Both models, score slightly better than the SVM model.

7 Implementation

The final model from the previous chapter provides as outcome an percentage indicating what the probability of winning business is for an incoming RFQ with a cutoff score of 50% for separating the two groups. Now the question is how can the outcome of this model used within the RFQ & Solutions Desk. This part is focused on the last research question, how could the outcome of the model be transformed into a recommendation for Ewals? To come up with a final recommendation the outcome of the prediction model should be combined with potential monetary value in a RFQ. Ewals wants to participate in RFQ's in which the potential monetary value gained is as high as possible. Therefore, the performance of the product data will be used to calculate the possible monetary value in an incoming RFQ. However, as indicated before not all products have much observations. This can lead to high errors in the calculations of the monetary value.

The value in a RFQ is calculated by multiplying the number of shipments awarded with the offered price. The number of shipments requested by the customer and the historical average price will be combined with the historical product performance on product level and the offered percentage to compensate for the times Ewals does not offer and the other times no business is awarded. Before this calculation is conducted, statistical knowledge will be used to calculate a confidence interval for the product performance of each product to prevent errors in the performance data.

To deal with uncertainty caused by the difference in number of observations, a confidence interval for each product is calculated. By using a confidence interval, the performance values will be more reliable. Following Donnelly and Abdel-Raouf (2016) the central limit theorem can be used as follows: despite of the distribution of the population itself, when the sample is large enough, being thirty or higher then the distribution of the sample will be normally distributed. For calculating the confidence interval of products meeting this requirement, the z score can be used for the calculation. For products with less than thirty observations a t score can be used following Donnelly and Abdel-Raouf (2016).

As described before, the prediction models conduct a binary classification task. Based on this output, the estimator sample proportion theory is used for making further calculations. An additional check for the usability of the Central Limit Theorem for proportional sampling is the use of the 5-rule. This rule will be provided with the formulas. N represents the number of observations. σ represents the standard deviation. The following formulas are used for calculating the confidence intervals. Observations that do not meet this 5-rule requirement will be described as not statistical significant.

$E(\widehat{P}) = p$ as estimator for the sample mean

$\sigma(\widehat{P}) = \sqrt{\frac{p*(1-p)}{n}}$ as estimator for σ

$\widehat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$ calculation estimator

Lower bound meeting 5-rule: $\widehat{P} - z_{\alpha/2}\sqrt{\frac{\widehat{P}*(1-\widehat{P})}{n}}$

Upper bound meeting 5-rule: $\widehat{P} + z_{\alpha/2}\sqrt{\frac{\widehat{P}*(1-\widehat{P})}{n}}$

Lower bound not meeting 5-rule: $\widehat{P} - t_{\alpha/2}\sqrt{\frac{\widehat{P}*(1-\widehat{P})}{n}}$

Upper bound not meeting 5-rule: $\widehat{P} + t_{\alpha/2}\sqrt{\frac{\widehat{P}*(1-\widehat{P})}{n}}$

The following conditions should be met to apply the Central Limit Theorem following the 5-rule.

$n * p \geq 5$

$n * (1 - p) \geq 5$

The 5-rule is combined with practical implications from Ewals' side. Multiple products could belong to one RFQ, therefore two practical rules are added. The number of RFQ's for a product should be higher than ten RFQ's. Furthermore, winning products should come from more than three winning RFQ's. By adding these two rules the reliability of the performance calculations will increase, because the data observations come from multiple sources (multiple RFQ's). Finally, two groups are distinguished as named above, namely statistical significant products and not significant products. Products labeled as statistical significant meet the following requirements:

- Number of observations $>$ 30 (number of product observations $>$ 30)
- Number of RFQ's $>$ 10
- Number of awarded RFQ's $>$ 5 (for a winning product)

After the calculations of the confidence interval for each product, the performance value should be combined with other information to be practical applicable. As explained above, RFQ's consist out of multiple lanes, whereby the lanes can be linked with products. Thus, now a performance value can be assigned to each row/lane in a RFQ. The performance will be combined with the percentage indicating the number of times offered by Ewals, the historical average price and the number of shipments requested to get a potential monetary value. The average price will be calculated based on the total offered value divided by the number of shipments belonging to that lane.

The output of the calculations is provided in figure 23. The sum of the value of each group is presented with the number of shipments. This potential value can be used to compare RFQ's, to prioritize RFQ's with a higher potential gain. The minimum value will be used for the decision making representing the most negative scenario.

II. PRODUCT ASSESSMENT				
	Minimum	Maximum		
BUSINESS POTENTIAL	**€ 1.456.992,00**	**€ 4.090.190,00**	[redacted]	
	Value minimum	Value maximum	# of shipments minimum	# of shipments maximum
- STATISTICAL SIGNIFICANT	€ 1.260.602,00	€ 3.032.257,00	€ 1.348	3032
- NOT STATISTICAL SIGNIFICANT	€ 196.391,00	€ 1.057.934,00	€ 140	1058

Figure 23: Example potential

The outcome of the prediction model is a good recommendation for going further with the RFQ preparation to come to an offer for the customer. However, a high probability of winning the RFQ is not enough. A win probability of 90% for a RFQ with a potential of only 10.000 euros could still not be interesting.

Therefore, the prediction outcome and the potential value should be combined for final decision making. As explained before, two categories are created, namely significant potential and non significant potential, which are the products which have a low frequency in the data, this low frequency causes uncertainty. The borderline for the potential business value is based on the average of the assigned value to Ewals in historical RFQ's. The potential value is high when the assigned value is more than the average assigned value, namely 108522 euros, otherwise the potential is low.

Win probability	**Significant potential**	**Non significant potential**	**Decision**
High	High	High	Go
High	High	Low	Go
High	Low	High	Go
High	Low	Low	Employee
Middle	High	High	Go
Middle	High	Low	Go
Middle	Low	High	Employee
Middle	Low	Low	No Go
Low	High	High	No Go
Low	High	Low	No Go
Low	Low	High	No Go
Low	Low	Low	No Go

Table 18: Final decision making

The decision making based on the prediction model and the potential value is shown in table 18. In this table, four columns can be found, namely win probability, significant potential, non significant potential and decision. Within the win probability category, three values are distinguished: high, middle and low. The middle category is introduced, because the prediction model is not perfect, prediction errors can still be found (accuracy

68%). If a percentage of 53% is found, this is really close to the border of 50%. The middle category is introduced to capture incorrect predictions by better evaluate predictions close to the cut off score of 50%. The categories are defined as follows:

- Low win probability: $x < 33\%$
- Middle win probability: $33\% \leq x \leq 67\%$
- High win probability: $x > 67\%$

Now the guidelines for the decision making will be explained. In the scenario of a high win probability, Ewals should proceed with participation in tendering, except for low potential business value. In this situation, the preparation and execution are time and resource consuming, while the gain is low. The decision description employee is added to this scenario. It could be that Ewals received the RFQ from a regular customer, in this situation the employee can evaluate if they should proceed.

In the scenario of middle win probability, Ewals should proceed if both the potential values are high or if the significant value is high. In these scenarios, Ewals could be certain that the potential gain is high. In the scenario, of a high non significant potential value, a second check should be conducted by a Ewals' employee. The uncertainty in the data could influence the decision in this scenario. If both potential values are low, then Ewals should not proceed with the RFQ.

Lastly, in all scenarios of a low win probability, Ewals should not proceed with the received RFQ. The implementation of the prediction model is introduced to filter the waste of time and resources in this situation.

8 Conclusion and recommendations

8.1 Conclusion

In this section an overall conclusion about this thesis will be drawn based on the main research question, which is provided below.

- *How could Ewals Cargo Care predict in which tendering they should participate to increase potential business winning?*

For this research project a prediction model is designed. This model is used to predict the outcome of participation in the tendering process of a received RFQ before processing it. At the beginning of the project, the data was evaluated to come up with input variables for the prediction model. This part answered research question one, indicating the internal and external factors influencing tendering.

Within Ewals, two levels of depth for data registration are found, data is available on RFQ level and on product level (country relationship). First, variables on RFQ level are used to build a prediction model, because the number of observations on product level are more limited than the variables on RFQ level.

The next step is the selection of the prediction models. For this binary classification task, the following three models are selected: logistic regression, artificial neural networks and a support vector machines model. Model building is started with the logistic regression model. The selection of the variables for the prediction model is based on model fit. After measuring the model fit for different variable combinations, the five variables indicating the SEM of the customer, current business, which represents transport business within the new arrived RFQ executed by Ewals in the year before the arrival of the new RFQ, the relationship with the customer (good relationship when a KAM is assigned to this customer), FTL or LTL truckload and lastly an indication of high potential within Ewals, are selected for the final model

Secondly, model fit is checked again for possibly adding variables on product level. This step provides two additional variables, namely the number of transport lanes and the number of shipments within a RFQ. For both models, the model performance is checked for the three selected prediction models.

After comparing the performance of the three models, the model with the product level variables is outperformed by the model containing five variables on RFQ level. The logistic regression and support vector machine models score slightly better than the artificial neural networks. These two model achieve a prediction accuracy of 68% relative to a accuracy of 67% for the neural networks. Both models could be implemented, whereby the logistic regression model could be recommended to have a better overview of the influence of the model parameters on the prediction score.

Lastly, the outcome of the prediction model is combined with business potential. This step is conducted to come up with recommendations for Ewals, which answers research question four. A statistical analysis is used to come up with a confidence interval for product performance. This confidence interval of product performance is combined with the historical price, number of shipments and the number of times Ewals offered products to the customers. Eventually, an interval is provided providing an overview of the potential monetary gain in a RFQ.

Finally, for different outcomes of the prediction model and potential business winning in a RFQ, a decision framework is provided. This framework provides a recommendation about proceeding or not. This decision framework, tries to capture wrong predictions by better evaluating prediction close to the cut off score of 50%, because of the prediction accuracy of 68%.

8.2 Recommendations

The first recommendations are about extending the prediction model to increase the predicting accuracy. First, an variable linked to the requested equipment type can be added. Equipment type has an important role in assigning the business, because of scale advantages achieved by using certain equipment types for specific products. Furthermore, additional requirements in the field of load securing, tracking of the transports, pallet exchange and dangerous good transports can be added. For example, in most situations Ewals does not conduct pallet exchange, thus Ewals has a lower win probability for RFQ's containing the pallet exchange request.

On this moment the product level data is limited. The product level data contains many products with few observations. If more product level data is available, these variables should be considered again. Possibly, the performance of the models with product level data included is better than the current models.

In the current models some customer information is included, however, more extensive customer information can be considered to include in the model. Qualitative customer information like requirements in the field of payments terms, penalties and liability can be considered.

Furthermore, the decision framework should be evaluated after some time. The percentages and monetary value guidelines could be adjusted after evaluation next year. By changing these values, the decision making could be improved.

9 Discussion

During this research some assumptions and decisions are made to be able to design the model. In this section, these limitations are discussed.

Firstly, data unavailability influences the choices of the model parameters. In the literature, price per kilometer was proposed as possible input parameter for a prediction model. Furthermore, the literature about the decision making of the LSU's is based on pricing. Other variables as service are added as costs as found in the literature. First, the idea was to include market information about pricing/costs in the model. Secondly, market variables like capacity and regulations can be added. Market information was not yet available during this project.

In this research, the model design was limited by two cases of data unavailability. For the product level data, a limited number of observations was available on this moment. Furthermore, pricing information of Ewals could not be included, because the prediction model is already before prices are calculated in the preparation process of the offer.

References

Adjavor, H., Al-Hihir, S., & Wang, Y. (2020). A classification of winners and losers of shippers' weekly transportation capacity procurement. In *Iie annual conference. proceedings* (p. 346 - 351).

Aguezzoul, A. (2014). Third-party logistics selection problem: A literature review on criteria and methods. *Omega*, *49*, 69 - 78.

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

Bask, A. (2001). Relationships among tpl providers and members of supply chains – a strategic perspective. *Journal of Business and Industrial Marketing*, *16(6)*, 470 - 486.

Beil, D., & Wein, L. (2003). An inverse-optimization-based auction mechanism to support a multiattribute rfq process. *Management Science*, *49*(11), 1529 - 1545.

Caplice, C., & Sheffi, Y. (2003). Optimization-based procurement for transportation services. *Journal of Business Logistics*, *24*(2), 109 - 128.

Chen, R., Gi, S., Cohn, A., Beil, D., & Sinha, A. (2009). Solving truckload procurement auctions over an exponential number of bundles. *Transportation Science*, *43*(4), 493 - 510.

Chen, Y., Goan, M., & Huang, P. (2011). Selection process in logistics outsourcing - a view from third party logitstics provider. *Production Planning & Control*, *22(3)*, 308 - 324.

Donnelly, R., & Abdel-Raouf, F. (2016). *Statistics*. O'Reilly Media, Inc.

Dreiseitl, S., Ohno-Machado, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*(5-6), 352 - 359.

Engelbrecht-Wiggans, R., Haruvy, E., & Katok, E. (2007). A comparison of buyer-determined and price-based multiattribute mechanisms. *Marketing Science*, *26*(5), 629 - 641.

Gaspar, P., Carbonell, J., & Oliveira, J. (2012). On the parameter optimization of support vector machines for binary classification. *Journal of integrative bioinformatics*, *9*(3), 33–43.

Gibson, B., Sink, H., & Mundy, R. (1993). Shipper-carrier relationships and carrier selection criteria. *Logistics and Transportation Review*, *29*(4), 371 - 382.

Guo, Y., Lim, A., Rodrigues, B., & Zhu, Y. (2006). Carrier assignment models in transportation procurement. *Journal of the Operational Research Society*, *57*(12), 1472 - 1481.

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

K.Kirasich, Smith, T., & Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogeneous datasets. *SMU Data Science Review*, *1*(3), article 9.

Leung, K., Luk, C., Choy, K., Lam, H., & Lee, C. (2019). A b2b flexible pricing decision support system for managing the request for quotation process under e-commerce business environment. *International Journal of Production Research*, *57*(20), 6528 - 6551.

Midi, H., Sarkar, S., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics*, *13*(3), 253 - 267.

Oeser, G. (2020). Segmenting logistics service users based on their provider selection criteria to succeed in outsourcing. *International Journal of Retail & Distribution Management*.

Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, *5*(4), 1 - 29.

Rekik, M., & Mellouli, S. (2012). Reputation-based winner determination problem for combinatorial transportation procurement auctions. *Journal of the Operational Research Society*, *63*(10), 1400 - 1409.

Robinson, C., & Schumacker, R. E. (2009). Interaction effects: centering, variance inflation factor, and interpretation issues. *Multiple linear regression viewpoints*, *35*(1), 6 - 11.

Senaviratna, N., & Cooray, T. (2019). Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics*, 1 - 9.

Song, J., & Regan, A. (2005). Approximation algorithms for the bid construction problem in combinatorial auctions for the procurement of freight transportation contracts. *Transportation Research Part B: Methodological*, *39*(10), 914 - 933.

Vries, S. D., & Vohra, R. (2003). Combinatorial auctions: A survey. *INFORMS Journal on computing*, *15*(3), 284 - 309.

Zhang, B., Ding, H., Li, H., Wang, W., & Yao, T. (2014). A sampling-based stochastic winner determination model for truckload service procurement. *Networks and Spatial Economics*, *14*(2), 159 - 181.

Appendices

A Status Management RFQ's

Request For Information Invited - Action Desk: Request For Information need to be filled

Request For Information Closed - Waiting Feedback: Waiting upon customer response to filled Request For Information

RFQ Closed - Not invited: Request For Information has been filled, but Ewals did not make it to the eventual RFQ

RFQ R1,2,X - Preparation Documentation & Volume analysis: Preparation by the RFQ & Solutions Desk

RFQ R1,2,X - Inventarize interest local cell: Identify whether a local cell is interested to participate in a given RFQ

RFQ R1,2,X - Waiting feedback local cell: Waiting for local cells to respond to the RFQ

RFQ R1,2,X - Prepare offer: The internal deadline has been expired, the RFQ & Solutions Desk will prepare the offer before it eventually will be shared with the customer

RFQ R1,2,X - Waiting feedback customer: Waiting for local cell to respond to the RFQ

RFQ R1,2,X - Preparation feedback to local cells: Feedback has been obtained by the customer, the desk will prepare and analyze the feedback and revert back to designated participants

RFQ Closed - No Business won: customer feedback obtained, no business won

RFQ Closed - Business Won - Prepare implementation: customer feedback obtained, business won: RFQ knowledge conveyance from RFQ & Solutions process owner to KAM to be executed

RFQ Closed - Business Won - Handover complete: customer feedback obtained, business won: all (operational) details have been shared with designated KAM

RFQ Closed - No Participation - Commercial / Technical / Operational / Other: No participation due to commercial, technical, operational or general reason(s).

RFQ Closed - No Offer Received: RFQ has been distributed, but no response offers obtained from Ewals local cells.

RFQ Closed - Go It Alone Ewals Local Cell: RFQ & Solutions Desk will not insource the RFQ and (a) local cell may handle the RFQ independently

RFQ - On Hold RFQ received, but the RFQ will not processed pro-actively. After RFQ deadline expiration, the RFQ will be closed as "RFQ Closed - No Participation - Other: No proactive pursue"