TILBURG UNIVERSITY

FORECASTING INSURANCE CLAIM AMOUNTS IN THE PRIVATE AUTOMOBILE INDUSTRY USING MACHINE LEARNING ALGORITHMS

TARA JONKHEIJM

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

Thesis committee:
Dr. S. Khoshrou
Dr. N.Ranković

STUDENT NUMBER

2005791

COMMITTEE

dr. Samaneh Khoshrou
dr. Nevena Ranković

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

May 23th, 2023

WORD COUNT

8773

ACKNOWLEDGMENTS

> "The important thing is to never stop questioning"- Albert Einstein

You are about to read the thesis "Forecasting Insurance Claim Amounts In The Private Automobile Industry Using Machine Learning Algorithms", which symbolizes the end of an exciting journey in obtaining my master's degree in Data Science & Society. The above-cited quotation from Albert Einstein is an excellent metaphor that captures my journey over the past year and certainly during this thesis project.

I have found this year extremely valuable and exciting, and I have grown tremendously both intellectually and personally. In addition, I have become aware of my own tenacity and persistence, and I have learned life lessons that I will cherish and carry with me forever.

That being said, the production of this thesis also signifies the effort of several people. First of all, I express my gratitude to my thesis supervisor Dr. Samaneh Khoshrou, who has offered valuable patience and guidance throughout the period. Additionally, I want to thank Jurgen Schaap (Raetsheren) and Your Benefits Assuradeuren for their encouragement and advice at a time when it was most needed.

Most importantly, this adventure would never have come this far without the immense support of a number of people who I would like to name explicitly. First of all, I would like to express my gratitude to my beloved Tijn Kapteijns for his everlasting support and faith in me, even when I lost confidence. Additionally, thank you to my family and friends for your encouragement, interest, and kind words. As always, your assistance has been deeply valued.

I hope you find enjoyment in reading this work.

Tara Jonkheijm
Breda, June 23, 2023

CONTENTS

FORECASTING INSURANCE CLAIM AMOUNTS IN THE PRIVATE AUTOMOBILE INDUSTRY USING MACHINE LEARNING ALGORITHMS

TARA JONKHEIJM

Abstract

The accurate forecasting of insurance claim amounts at an individual level in the private automobile industry using machine learning algorithms is crucial for insurers to fine-tune their rate-making strategies and ensure healthy competition. Machine learning algorithms could potentially add value to this area by estimating dependencies between customer information and claim amounts via historical claim data. This study focuses on a dataset from a Dutch private automobile insurer and evaluates the performance of four machine learning algorithms in predicting individual insurance claim amounts. These machine learning algorithms (linear regression, regression tree, random forest regression, and XGBoost regression) have not been extensively applied within the private automobile insurance industry. This study's evaluation of prediction accuracy and investigation of feature influences were conducted using various evaluation metrics. The results indicate that the XGBoost regressor achieved the lowest mean absolute error (MAE) in Experiment 2, demonstrating the highest accuracy with a value of 943.72 compared to the other algorithms. Nonetheless, the differences in performance among the algorithms in the experiments were marginal, suggesting relatively similar predictive capabilities across the models. While the MAE scores provide a reasonable measure of the average deviation of the models, a more comprehensive analysis of the graphical representations reveals a more nuanced perspective. Further research is required to evaluate whether advanced feature selection/engineering, algorithm adjustments, or alternative methods can improve prediction accuracy.

1 DATA SOURCE/ CODE / ETHICS STATEMENT

The data for this study has been acquired from Your Benefits Assuradeuren B.V., an authorized agent in the Netherlands. The organization explicitly gave its consent to use this data for this thesis, and the use of this data complies with the GDPR. The original owner of the data used in this thesis retains ownership of the data during and after the completion of this thesis.

The obtained data is anonymous and not available to researchers. Work on this thesis did not involve collecting data from human participants or animals. This thesis's author acknowledges that they have no legal claim to the data. All figures and tables in this study are the author's creation. The thesis code is currently not released to the public.

2 INTRODUCTION

2.1 *Context*

The insurance industry, built on data-driven risk assessment, plays a vital role as the world's population continues to increase. With the global population reaching 8 billion people, a significant percentage of individuals are expected to acquire one or more insurance policies (United Nations, n.d.). To accurately assess risks, insurance companies must analyze extensive amounts of data due to the large number of insured individuals. Machine learning (ML) can be used to process this data more effectively.

The principle of insurance is ancient and requires data about the customer and/or the object to be insured before an insurance policy can be issued. The insurer must estimate the probability of loss and determine how much premium is needed to pay for any benefits. Although there is a relationship between the amount of data and the accuracy of the insurer's estimate, this relationship is not directly proportional.

Recent research shows that the insurance industry is gradually adopting ML approaches (Grize et al., 2020; Herrmann & Masawi, 2022; Krasheninnikova et al., 2019; Zheng & Guo, 2020). Particularly in recent years, using ML in insurance domains has provided opportunities to strive for data-driven procedures that enable effective and generalizable prediction models. Based on various studies, it can be observed that technological improvements have significantly changed the insurance sector, resulting in the creation and maintenance of large databases with more information than ever before (Vandrangi, 2022; Zheng & Guo, 2020). Nonetheless, despite recent technological advancements, many insurers still need to widely integrate ML approaches into their business (Grize et al., 2020). Consequently, it becomes clear that further action is necessary to fully leverage this data's potential.

In addition to the overall transformations observed in the insurance industry, similar changes have been witnessed specifically within the automobile insurance sector. This particular industry has been challenging for insurers for several decades, and they thus must continually fine-tune their rate-making to differentiate themselves from all industry competitors (Selvakumar et al., 2021; Spedicato et al., 2018). With the rising availability of data and advances in data science techniques, insurers are increasingly turning to predictive modeling to forecast insurance claims which has been steadily increasing over the past few years.

Forecasting insurance claims is not a new area of expertise and has a long-standing history in the insurance industry. A precise and accurate prediction of claim amounts is crucial for the private automobile insurance industry to determine policy prices and effectively manage risks (Fauzan & Murfi, 2018). ML techniques are a suitable fit for this field because these algorithms can analyze large volumes of data to accurately assess risk. By incorporating a combination of various individual data features into the forecasting model, insurers can learn from their customers' behavior (Zheng & Guo, 2020).

The main objective of this research is to examine the accuracy of predicting claim amounts on an individual level for private automobile insurance policyholders in the upcoming year. Machine learning techniques (specifically linear regression, regression tree, random forest regression, and XGBoost regression) will be utilized to assess the predictive performance of these models. Evaluation metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) will be employed to compare the effectiveness of the models in each experiment. This study aims to fill a gap in knowledge within the private automobile insurance industry by offering insights that can aid insurers in refining their rate-making strategies and enhancing risk management through data-driven approaches.

2.2 *Research Questions*

This study addresses the following research questions:

RQ1 *To what extent can ML techniques like linear regression and ensemble methods for regression purposes be relied upon to forecast claim amounts in the private automobile insurance industry?*

RQ2 *To what extent do actuarial profile features and non-actuarial features contribute to the predictive performance of the ML techniques?*

2.3 *Gap in Knowledge*

This study contributes to the scientific literature in multiple ways:

- First, ML techniques hold significant potential for the insurance industry to forecast future patterns and behaviors. Previous research has examined traditional linear regression and/or econometric models; however, despite the unprecedented popularity of machine learning techniques, the ensemble methods examined in this study have not been widely used in the automobile insurance industry. Nevertheless, the automobile insurance industry presents an ideal context for such methods due to the abundance of available data. This research aims to provide a scientific contribution toward reducing the gap in the literature on this topic.
- Second, the research topic of this thesis currently pertains to a gap in the literature. Relevant studies in this field are mainly conducted in the United States, the United Kingdom, and China; however, research with a similar scope has yet to be conducted in the context of the Netherlands.
- Third, although some overlap in the predictor variables can be observed in the literature, no other study has used the exact composition of features in scientific research. By integrating this unique composition of features, this work seeks to create a more accurate and nuanced prediction model for automobile claims.

2.4 *Societal Relevance*

Accurate claim forecasting has considerable value for the automobile insurance industry, as it enables more effective pricing and broader coverage from insurers (Fauzan & Murfi, 2018). In the past, it was common practice to predict automobile insurance claims using complicated actuarial calculations, which was a time-consuming and labor-intensive process. Additionally, this method was not the most accurate method of prediction; therefore, a comprehensive and modern pricing system is crucial for healthy competition and preventing adverse selection (Yu et al., 2021).

Wider society benefits from insurance organizations' security by offering protection against financial losses. By transferring the risk to the insurer in return for a set premium enables people to swap uncertainty for certainty. Consequently, it is crucial that an insurer accurately assesses the risks in its portfolio and can produce a "correct" premium concerning the societal interest they should pursue. Tools for predictive modeling are helpful in this regard.

3 LITERATURE REVIEW

Adoption of Machine Learning in the Insurance Industry

The adoption of machine learning techniques within the insurance sector has gained considerable attention due to their potential to enhance prediction accuracy and address the challenges encountered by conventional methods of risk assessment (Blier-Wong et al., 2020). One noteworthy advantage of machine learning lies in its capacity to extract information from extensive datasets in an automated fashion, and this ability is particularly relevant in the context of this study. Moreover, machine learning algorithms possess the capability to handle both linear and nonlinear relationships, thereby facilitating the extraction of insights from complex data patterns within the overall insurance industry (Goulet Coulombe et al., 2022). Nonetheless, caution must be exercised in ensuring data accuracy when employing data-driven algorithms to prevent the introduction of bias and discrimination (Diepgrond, 2020).

The private insurance industry has undergone rapid transformation in recent years, thereby necessitating quick and efficient solutions for decision-making (Krasheninnikova et al., 2019; Wang et al., 2017). The utilization of traditional calculations for estimating claim costs, which typically rely on aggregate claims data, is often both costly and complex. Furthermore, these traditional approaches tend to overlook individual claims behavior and, thus result in a rigid methodology (Wüthrich, 2018). Such traditional techniques are also ill-suited for addressing forecasting challenges. Consequently, there is a growing demand for alternative approaches, like the application of machine learning algorithms, as the industry embraces digital transformation (Dhieb et al., 2019). Moreover, it is evident that the insurance industry is shifting further towards data-driven strategies, emphasizing the crucial role of machine learning techniques and their forecasting capabilities for the sector (Grize et al., 2020).

Machine Learning in Claim Amount Predictions
In the meantime, the adoption of machine learning techniques in the insurance sector has gained significant momentum, offering the potential for enhancing claims prediction. While linear regression and autoregressive integrated moving average models (ARIMA) have traditionally been popular due to their simplicity and interpretability, recent advancements in machine learning have introduced more accurate and sophisticated algorithms (Cummins & Griepentrog, 1985). These recent advancements are known as ensemble methods, which have emerged as valuable alternative prediction algorithms due to their ability to balance bias and variance (Quan & Valdez, 2018). Nevertheless, despite demonstrating promising results, the application of these methods in the private automobile industry has been relatively limited. Consequently, there is a scarcity of information regarding the performance of ensemble methods in this specific context, highlighting the need for further research and analysis to which this study aims to contribute.

Machine Learning in Insurance Sectors Other Than the Automobile Industry
Nonetheless, a few studies outside of the automobile insurance industry provide valuable insights that can serve as an effective approximation for evaluating the performance of linear regression and ensemble methods in the insurance claim prediction context; for instance, Duval and Pigeon (2019) conducted a case study using a North American dataset of claims within liability insurance. In their study, the authors employed linear regression, XGBoost regressor, and regression trees as weak learners to directly predict claim amounts. In order to predict the total claim amount, a collaborative effort was conducted with experts within the insurance company to select 20 features. This selection process was based on the understanding that the insurance sector is a highly specialized domain and therefore, require domain-specific knowledge for feature selection. This expertise is crucial in the context of feature selection, which is commonly classified as the filter method within the data science domain. Duval and Pigeon (2019)'s subsequent comparison of the results of these different models, highlighted the instability of linear regression compared to XGBoost regressor in estimating individual claim amounts in the liability insurance industry. Additionally, Duval and Pigeon (2019) did not utilize commonly used machine learning evaluation metrics such as MAE, MSE, RMSE, and R2 due to the presence of numerous zero claims in the dataset. Instead, Duval and Pigeon (2019) evaluates the performance of the models

through the mean and variance of the residuals. Since this thesis eliminates the zero claims from the full dataset to reduce noise, working under the conviction that these conventional evaluation metrics are the best option for this investigation, it employs the aforementioned evaluation metrics.

The perspective offered by Duval and Pigeon (2019) differs from the approach of Wüthrich (2018), who proposed a machine learning method for individual claim amounts using regression trees. The difference in these approaches lies in their prediction strategy, as Duval and Pigeon (2019) directly predicted a claim amount based on all available information at a given time, while Wüthrich (2018) used regression trees to calculate the total amount paid for all claims. Wüthrich (2018) employed a dimensionality reduction technique to minimize the number of input features while maximizing the retention of pertinent information. Additionally, the study incorporates expert knowledge from actuaries within the insurance company by utilizing the filter method for feature selection. This thesis follows the approach taken by Wüthrich (2018) and Duval and Pigeon (2019) because not all information in the dataset is equally valuable for prediction.

Wüthrich (2018)'s study highlighted the flexibility of regression trees in incorporating various types of feature information, as is often the case in the insurance industry. The findings of Wüthrich (2018)'s study recommended the exploration of machine learning techniques, including random forests and neural networks. Nonetheless, it is important to acknowledge that neural networks have a significant drawback; namely, the presence of hidden layers, which can be considered as black boxes lacking comprehensive theoretical guarantees. The results obtained from neural networks often involve prediction uncertainty and potential discrimination, issues that could potentially be mitigated with a sufficient amount of data. Given that addressing these challenges is often complex in the insurance industry, a deliberate decision was made not to employ the neural network algorithm in this thesis. Moreover, Wüthrich (2018)'s study focused on the total number of payments for all claims and did not examine individual claims, thus necessitating further investigation. This thesis adopts the recommendation to employ a more robust model, such as random forest, and also addresses the desired follow-up research at the individual claim level to fill the existing gap in knowledge. Similarly, Guelman (2012) presented the theory of gradient boosting and its application in auto insurance loss cost modeling; the study showcased the predictive accuracy and model interpretation capabilities of gradient boosting, which are crucial in business environments where decision-makers without statistical training require an understanding of model outputs. This is also the case for the organization who provided the dataset for this study.

Guelman (2012)'s study lacks a discussion of feature selection and only offers a brief mention that this process was performed as an integral part of the research. The specific method used is not explicitly stated, making it challenging for other researchers to reproduce and validate the process. Transparently stating the methods used is crucial for promoting transparency and reproducibility in scientific research. The study of Guelman (2012) has the drawback of having more category predictors than numerical predictors. As a result, a large number of dummy variables had to be included, which complicates how the results should be interpreted. The results of Guelman (2012)'s study indicated higher prediction accuracy compared to conventional approaches, such as linear regression; however, Guelman (2012)'s study focused on the numerical value of claim frequency which despite having a similar scope, has a different target variable. Nonetheless, the information regarding the used predictor variables and models employed in Guelman (2012)'s research is deemed interesting for the purposes of this study as there is some overlap in the context.

Given the limited availability of relevant information regarding the scope of this research, some additional branches within the insurance industry were also examined. In addition to the private automobile insurance industry and the liability insurance industry, studies in the healthcare insurance industry were similarly considered, as machine learning approaches are also increasingly being utilized in this sector. The study by Kaushik et al. (2022) closely aligns with the approach and regression focus of this thesis. Kaushik et al. (2022) employed regression techniques such as neural networks and linear regression to predict future health insurance premiums and, although the study targeted a different variable than the present thesis, they share a similar scope. Nevertheless, it is noteworthy that, although this study also obtained the predictor variables through dimensionality reduction, once again, no explicit explanation is provided regarding the specific implementation of this process. The dataset used in Kaushik et al. (2022)'s study encompasses various predictor variables, contains no null values, and utilizes key regression evaluation metrics (MAE, MSE, RMSE, and R2) to assess model performance. The MAE value takes precedence in demonstrating accuracy due to its interpretability and relevance in predicting future expenses. Remarkably, the evaluation metrics and graphical visualization of the actual and predicted values are highly accurate, yielding an MAE value of 0.249, MSE value of 0.344, RMSE value of 0.499, and an R2 value of 0.75. This finding is somewhat surprising as the correlation matrix reveals relatively low correlation values between all the predictor variables themselves (values between 0.00041 and 0.3) where it can be questioned how these low correlation values affects the results. The correlation with the target variable in this study is not reported, which is yet another remarkable finding. Additionally, Kaushik et al. (2022) recommends conducting thorough research to further investigate the low values of the identified correlations.

Current Research
Each of the aforementioned related studies employs a distinct combination of predictor variables that may be relevant in the insurance industry. The predictor variables in this study were obtained using the filter method of feature selection, as suggested by Duval and Pigeon (2019). All previous studies focused on traditional variables, while this thesis aims to provide a new insight by additionally incorporating the discovery of additional services such as legal aid and (extra) breakdown service. In addition to the traditional approach of linear regression, this thesis incorporates ensemble methods, as there is limited or no research available in the related literature; by doing so, the thesis aims to address the existing knowledge gap. Furthermore, ensemble methods have demonstrated promising results in terms of accuracy, as indicated by Quan and Valdez (2018). Key regression evaluation metrics (namely MAE, MSE, RMSE, and R2) are employed as they offer the best fit for the dataset, industry domain, and chosen machine learning algorithms. Nevertheless, no prior research has employed the exact composition of variables used in this study, thus emphasizing the unique relevance of this research.

4 METHODOLOGY AND EXPERIMENTAL SETUP

4.1 *ML Algorithms*

Linear Regression
A straightforward linear regression model serves as a forecasting method for estimating the claim amount by a specific predictor variable time (Selvakumar et al., 2021). For estimating the claim amount based on historical data, the procedure for fitting a simple linear model with an error term is denoted as:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

Where Xi is the independent variable, Yi is the dependent variable, and ei is a residual error term (Selvakumar et al., 2021).

Ensemble Methods for Regression

The ensemble modeling approach was chosen due to its ability to include both linear and non-linear interactions in data (Ramzai, 2019). Ensemble methods are appropriate for examining larger and more complicated data sets (Breiman, 2001). Furthermore, the several ensemble technique approaches are exceedingly reliable and flexible, contributing to their recent increase in popularity. Despite these various advantages, there is also a drawback to these techniques: it is computationally expensive to create the algorithms. Nonetheless, ensemble methods usually generate accurate results once they are assembled, and this cost can thus be considered as a one-time investment made in this study.

Regression Trees

Regression trees are a variant of decision trees and are considered a classic machine learning algorithm. According to the literature, this strategy is known as classification and regression trees (CART). It is an informative technique since it prepares candidate predictors and represents information in a way that it can be logically and simply visualized. Furthermore, regression trees may tolerate missing data in predictor factors by utilizing surrogates and are insensitive to outliers (Breiman, 1996). Furthermore, unlike classification trees, regression trees work with numerical target variables, which makes them well-suited to examining financial data like the dataset used in this study. Regression models aim to establish the connection between a single response variable and a set of independent variables that are extracted from the original dataset. Unless the node is a terminal node to which the data points are assigned, a regression tree is created using a tree-like structure. It is formed iteratively by dividing each node into child nodes depending on specific criteria (Breiman, 2001).

The algorithm partitions explanatory variable space into regions and assigns a constant value to each of them, thus allowing predictions for observations within. The response variable of the new data is then predicted by fitting a regression model to each terminal node. The mathematical formula of the regression tree algorithm is denoted as:

$$f(x_i|\Theta) = \sum_{m=1}^{M} c_{m,1} I(x_i \in R_m) \tag{2}$$

Regression trees in this study will be used to predict the claim amounts. A regression tree is not an ensemble method by itself but will be used as a building block in the ensemble models, as mentioned before (Quan & Valdez, 2018).

Random Forest Regressor

On numerous benchmark datasets, Breiman (2001)'s ensemble technique for classification and regression problems demonstrated outstanding performance in prediction accuracy. This thesis uses random forest for regression problems, an alternative approach to the random forest for classification problems. The method received its name because the ensemble's fundamental components are tree-structured predictors, and each of them is built using an injection of randomness.

Numerous regression trees are developed by random forest regressor. Each successive split for all of the trees is performed using only the part of the last split to which it corresponds rather than the entire dataset. Random forest lowers the variance of individual trees by generating many trees, which clarifies why it is frequently used. Furthermore, non-linear relationships between factors can also be taken into account, which could be an advantage for this thesis (Breiman, 2001). The random forest regressor is a collection of unpruned regression trees. The mathematical formula is denoted as:

$$f_B(x|\Theta) = \frac{1}{B}\sum_{b=1}^{B} f(x|\Theta_b) \tag{3}$$

where x represents the input, and the prediction is the unweighted average of the collection (Breiman, 2001).

XGBoost Regressor

The second ensemble model employed in this study is the XGBoost algorithm developed by Chen and Guestrin (2016). The XGBoost regressor algorithm is expressed as the sum of regression trees and is a scalable machine-learning approach for tree boosting (Chen & Guestrin, 2016; Fauzan & Murfi, 2018). This method is particularly suitable for large datasets and is frequently utilized in financial forecasting. XGBoost regressor is designed to be both efficient and fast (Pesantez-Narvaez et al., 2019).

XGBoost regression has achieved remarkable success due to its exceptional scalability, surpassing existing solutions by operating more than 10 times faster on a single machine. Moreover, it possesses the capability to handle billions of samples in memory-constrained or distributed environments. This scalability is made possible through the implementation of innovative techniques for handling sparse data, instance weights, and approximate tree learning. By leveraging parallel and distributed computing, XGBoost accelerates the learning process, thereby facilitating efficient exploration of models.(Chen & Guestrin, 2016). The mathematical formula is denoted as:

$$F_S(x|\Theta) = \sum_{s=1}^{S} f_s(x|\Theta_s) \tag{4}$$

where the result is a set of S regression trees. It is expressed as the sum of trees, where Os are the corresponding models produced by the trees. S is referred to as the number of iterations (Quan & Valdez, 2018).

4.2 *Dataset Description*

Predictive learning refers to a process that involves a set of input (predictor) variables, denoted as x = x1, ..., xp, and an output (target) variable. The datasets used for this research were supplied by a Dutch-authorized agent in the automobile insurance industry (non-life insurance data).

The organization provided two anonymized datasets with different attributes that can have both categorical and continuous characteristics:

- **Dataset 1** consists of claims data, based on a unique policy number, with 27 different columns and 42,929 rows. The dataset starts in 2012 and ends in February 2023.
- **Dataset 2** consists of more comprehensive information on the claims figures. This dataset has 47 columns and 60,508 rows. The 60,508 rows represent the number of policies in the portfolio of the automobile branch of Your Benefits. The dataset comprise policy numbers ranging from 1968 to February 2023. Each customer record consists of 47 attributes.

The above-mentioned datasets were merged using a common identifier; namely, the policy number. In the insurance industry, a policy number is the primary factor for data collection and processing. A policy number is a combination of digits generated for each insured object. A policyholder (who is assigned one customer number for all insurance products during their collaboration with an insurer) can have multiple policy numbers. In the event of a significant change to the insured object, a new policy number is generated along with new data and information. Consequently, no information is overwritten, and incorrect merging cannot occur because the unique policy number takes precedence over the underlying data. Furthermore, all claims are reported and recorded under the respective unique policy number. If a policy number reports multiple claims in the same year, the total claim cost amount is not cumulatively added under the same policy number.

The policy numbers in both datasets were anonymized before delivery by using pseudorandom number generators to generate unique and random policy numbers. These policy numbers cannot be traced back to any specific policyholder in any way.

A verbatim example of the merging of the two datasets is provided below:

"Policy number X has a driver who is 29 years old and resides in Region 4. The driver has operated a Citroen CX (petrol) since 2002 with a catalog value of 16,519 euros and a weight of 1,284 kilograms. The policy number has a total of three consecutive claim-free years and does not include any additional services such as legal aid, breakdown service, or extra breakdown service in the policy. The driver claimed 571.41 euros for car damage in 2018, which is recorded under a unique policy number."

The example above is translated from the selected features of the dataset. To offer a comprehensive understanding of the dataset, two additional examples are provided in Table 15 in Appendix A (page 48).

4.3 *General Workflow of the Current Project*

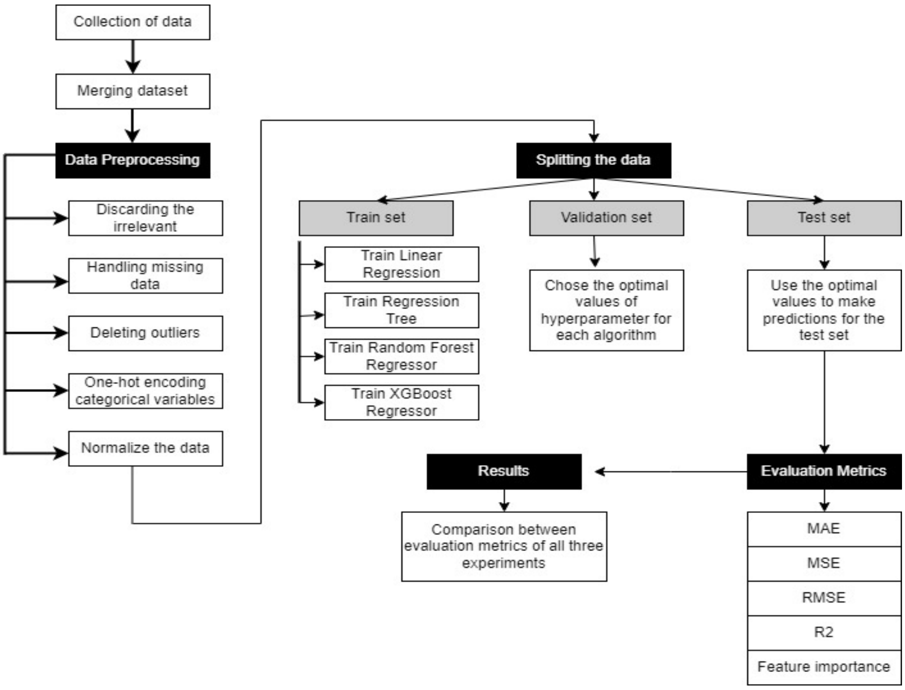


Figure 1: General Workflow of the Current Project

The data was split individually for each experiment. Furthermore, the subsequent process steps were performed independently for each experiment. The evaluation metrics are compared at the end of the process.

4.4 *Data Cleaning and Preprocessing*

4.4.1 *First Impression Dataset*

As has been stated, a unique policy number combined the two datasets (customer data and claims data). A dataset with 74 columns and 103,437 entries was produced after the merge; such datasets typically contain both categorical and continuous variables (Dal Pozzolo et al., 2011). Many columns in the dataset were immediately apparent as being unimportant to this study, and feature selection was thus applied. The most significant of the original features were chosen using this filter method, as well as being based on expert knowledge and the studies of Bian et al. (2018), Duval and Pigeon (2019), Guelman (2012), and Samson and Thomas (1987). This approach takes into account generalization beyond the training set of data; as a result, the model's variance and overfitting are minimized as much as possible. Moreover, feature selection increases model accuracy while decreasing computational costs and training time.

Eighteen relevant columns and 96,765 entries are left, specifically (Table 1):

Table 1: Relevant columns of the dataset

0	Policynumber	Object
1	Type of insurance	Category
2	Manufacturing year	Int64
3	Pure claim-free years	Int64
4	Region	Category
5	BMpercentage	Float64
6	Net premium	Float64
7	Fuel type	Category
8	Legal aid	Bool
9	Breakdown service in the Netherlands	Bool
10	Extra breakdown service in the Netherlands	Bool
11	Driver age	Float64
12	Catalog value	Float64
13	Weight of the car	Float64
14	Credit score	Float64
15	Collection amount	Float64
16	Car brand	Category
17	Total	Float64

4.4.2 *Cleaning*

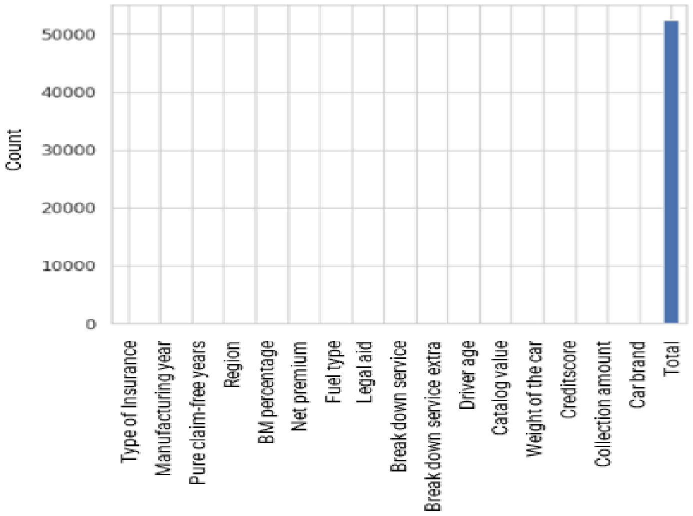
After the data was extracted, certain columns were incorrectly formatted by Excel. For instance, Region 5 was subsequently given more weight than Region 1 in the "Region" column, which has an ordinal distribution; this is not the intention of this column. Additionally, some variables were seen as strings, which required the data type to be corrected. After the columns were defined, the values of the columns were initially assessed. As shown in Table 2 and Figure 2, it is noteworthy that there are many NaN values (missing values) present. In particular, the dependent variable "Total" has a significant number of missing values (57,551); this is because not every policy number is required to have incurred damages and have subsequently filed them with the insurer during the period from 2012 to 2023.

Table 2: Percentage of missing values in the dataset

Policynumber	0.000099%
Type of insurance	0.000099%
Manufacturing year	0.000099%
Pure claim-free years	0.000099%
Region	0.000307%
BMpercentage	0.000099%
Net premium	0.000099%
Fuel type	0.000099%
Legal aid	0.000121%
Breakdown service in the Netherlands	0.000121%
Extra breakdown service in the Netherlands	0.000121%
Driver age	0.000099%
Catalog value	0.000099%
Weight of the car	0.000099%
Credit score	0.000099%
Collection amount	0.000099%
Car brand	0.000263%
Total	0.575515%

To prevent regression models from being significantly disrupted by the presence of missing values, there are several ways to handle this issue; for example, columns or rows can be deleted, or gaps can be filled with the column average. To avoid introducing noise and ensure that the columns serve as inputs for the regression models, it was decided to eliminate all rows with missing values. Finally, this process resulted in a total of 18 columns and 39,214 entries.

Figure 2: Barplot missing values



4.4.3 *Outliers*

The numerical values were transposed into a matrix to identify any outliers in this new dataset. Multiple columns were affected by this problem and require a deeper cleaning; for example, a manufacturing year of construction appeared to be in the future (2102). In addition, pure claim-free years had a negative minimum value and a maximum value of 89; however the pure claim-free years were capped at 15, as this is the maximum value the organization works with. Additionally, many values in the "Total" column of the dataset were given a negative value, which is an error after data extraction; these values were converted to positive values.

Some large standard deviations could be seen, indicating that there were extreme differences in the spread caused by outliers. These were columns with typing errors that were adjusted in the cleaning phase.

Nonetheless, in the "Net premium" column, there were significant differences that where also found to be correct after consultation with the organization. These premiums are so high because this policy has been considered high-risk (e.g., ancient car; high damage). Usually, de Vereende, an insurer for special risks, receives these policies; all rows with annual premiums above 2,500 euros were therefore removed from the dataset as they are outliers that occur very rarely. Furthermore, it is important to note that regular insurers are not allowed to keep or hold on to such items.

4.4.4 *Basic Statistics of the Dataset*

After the second cleansing session where outliers were removed, a clean and final dataset was obtained with 38,243 entries and 18 columns. The bas ic statistics of the dataset are presented in Table 3.

Table 3: Basic statistics of the dataset

	Count	Mean	Std	Min	25%	50%	75%	Max
Manufacturing Year	38.243	2007.47	6.08	1975.00	2003.00	2007.00	2012.00	2022.00
Pure claim-free years	38.243	4.98	4.46	0.00	1.00	4.00	8.00	15.00
BMpercentage	38.243	57.22	21.89	0.00	45.00	65.00	75.00	75.00
Net premium	38.243	624.45	384.65	109.56	360.12	508.00	757.56	2493
Driver age	38.243	42.97	11.13	18.00	35.00	40.00	52.00	92.00
Catalog value	38.243	23.693,19	12579.58	6095.00	14270.00	19644.00	29866.50	69980
Weight of the car	38.243	1124.34	227.87	625.00	950.00	1100.00	1285.00	2345.00
Credit score	38.243	18.59	26.01	0.00	12.00	31.00	49.00	80.00
Collection amount	38.243	756.96	465.48	132.60	438.00	618.00	1500.00	3016.68
Total	38.243	1111.70	1518.26	10.00	160.36	600	1428.08	10000.00

To gain a comprehensive understanding of the dataset, the distribution of the data was also examined by visualizing it in a manner that illustrates its spread. This graphical representation aids in understanding the data's characteristics and underlying patterns. In this study, density plots were used to visualize the distribution. Only the most important features are expected to be presented in the main text (page 25) for representation purposes. This selection was based on both the domain expert's knowledge and the findings of David (2015) and Guelman (2012) where the features "Driver age", "Pure claim-free years" and "Catalog value" were found to be statistically relevant. Additional density plots can be found in Figure 15 and Figure 16 in Appendix C (page 49).

4.4.5 *One-Hot Encoding*

Data preprocessing is required due to the dataset's combination of continuous and categorical variables with a continuous expected outcome to test a predictive model for insurance claims. Dummy variables must accurately encode categorical features (often referred to as one-hot encoding in the machine learning domain). Dummy variables are simply another way of representing categorical features. They do not create or extrapolate new signals or information from the data; for example, the following categorical variables were encoded: "Insurance type", "Region", "Fuel type", "Legal aid", "Breakdown service in the Netherlands", "Extra breakdown service in the Netherlands", and "Car brand". After one-hot encoding, the dataset contained 38,243 entries and 91 columns, representing an increase of 73 columns.

4.4.6 *Z-Score Normalization*

In all experiments involving a linear regression model, z-score normalization was implemented. This normalization technique was applied using the MinMaxScaler function from the sklearn.preprocessing module, as recommended by a study conducted by Anggoro and Supriyanti (2019). The purpose of z-score normalization is to potentially enhance the accuracy of the model. The z-score was calculated using the formula provided below:

$$z = \frac{x - \mu}{\sigma} \tag{5}$$

Nonetheless, it should be noted that normalization is not necessary for ensemble methods since these models do not rely on distance calculations; therefore, before computing error measurements, the data was rescaled to ensure comparability and consistency across results on the same scale.

4.5 *Experimental Setup*

4.5.1 *Software Packages and Applications*

The experiments were conducted using the Python programming language (version 3.10.11), and the following Python packages and libraries were used:

- Scikit-learn (Pedregosa et al., 2011)
- Pandas (McKinney et al., 2010)
- Matplotlib (Hunter, 2007)
- Tensorflow (Abadi et al., 2016)
- NumPy (Harris et al., 2020)
- Seaborn (Waskom et al., 2017)

The training of the models was carried out using Google Colaboratory Pro, a research project that provides access to powerful hardware options such as Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs). The user-friendly interface of Jupyter Notebook was utilized for conducting the training process (Bisong et al., 2019).

4.5.2 *Train/Validation/Test Split*

It is crucial to arrange the data in a way that makes learning possible in order to adequately prepare it for ML. To accomplish this goal, the dataset is frequently divided into subsets containing sufficient data to develop the model, optimize its hyperparameters, and assess its performance on new and untested data. Consequently, the train/validation/test split (70:20:10) remained the same for all individual experiments.

4.5.3 *Experiments*

- Experiment 1 concerned the relationship between the actuarial profile and the dependent variable "Total" as the claim amount. The features (columns) of the actuarial profile were “Credit score", “Fuel type", “Catalog value", “Driver age", “Manufacturing year", “Region" and “Weight of the car".

- Experiment 2 concerned the relationship between the features not included in the actuarial profile and the dependent variable "Total" as the claim amount. Features (columns) outside the actuarial profile were "Type of insurance", "BM percentage", "Net premium", "Legal aid", "Breakdown service in the Netherlands", "Extra breakdown service in the Netherlands","Pure claim-free years" and "Car brand".

- Experiment 3 merged the features (columns) from Experiments 1 and 2. The relationship between this combination and the dependent variable "Total" as the claim amount was examined.

  Each experiment implemented all four algorithms (Chapter 4). All the experiments had the same number of rows, and there were different numbers of columns for Experiments 1 and 2. The number of columns for the third experiment was equal to the sum of the number of columns in the first two experiments.

4.5.4 *Hyperparameter Tuning*

Grid search was specifically chosen over random search in this study due to its practicality and effectiveness in low-dimensional spaces. The manageable number of hyperparameters (Table 4) allowed for an exhaustive exploration of all possible combinations without significant computational burdens. Grid search is particularly recommended in domains where the accuracy of models is crucial, as noted by Siami-Namini et al. (2019). In contrast, random search can occasionally rely on luck, which is not a desirable approach for this study.

Table 4: Hyperparameter tuning

Algorithm	Hyperparameters	Experimented values
Linear regression	'alpha'	0.1, 1.0, 10.0, 100
	'loss'	'squared_loss,' 'Huber,' 'epsilon_insensitive.'
	'penalty'	'l2', 'l1', 'elastic net'
	'learning rate'	'constant', 'optimal', 'invscaling'
Regression tree	'max_depth'	4, 6, 8, 10
	'min_samples_split'	10,20,30
	'min_samples_leaf'	10, 16, 20
Random Forest Regressor	'n_estimators'	90, 100
	'max_depth'	1,7,1
	'min_samples_leaf'	1,7,1
	'min_samples_split'	1,7,1
	'max_features'	'auto', 'log2'
XGBoost Regressor	'n_estimators'	50, 100, 120
	'max_depth'	5, 7, 9
	'min_samples_split'	40, 50
	'learning_rate'	0.1, 0.3, 0.5

4.5.5 *Evaluation Methods*

Various prediction accuracy metrics are available, but no singular optimal measure can assess prediction accuracy across all scenarios. Several common metrics were employed to ensure a reasonable comparison between distinct models.

First, the evaluation of the model's quality of fit and accuracy of future predictions included the R-squared metric (R2). This metric can be seen as an effect size and measures the proportion of variability in the dependent variable that is accounted for by the model, with a value ranging from 0 to 1. A value closer to 1 indicates a better fit of the model to the data.

The R-squared is denoted as:

$$R^2 = \sum(\hat{yi} - \hat{y})^2 / \sum(yi - \hat{y})^2 \tag{6}$$

A popular evaluation statistic for regression models is RMSE, a modification of MSE that takes the error's square root. Since it uses the same units as the quantity represented on the vertical axis, it is favored over MSE since it is simpler to comprehend. A perfect RMSE score of 0.0 means that all forecasts perfectly matched the anticipated values.

Conversely, large discrepancies between the predicted and actual values are indicative of poor model performance, which is indicated by a greater RMSE. The RMSE is denoted as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

Finally, the mean squared error (MSE) and mean average error (MAE) were measured. MSE evaluates the quality of a forecasting model, considering both the variance and bias; a lower MSE value indicates a better model. The MAE ascertains how well the optimal set of parameters can generalize on validation data, showing the expected inaccuracy of the forecast on average values. A lower MAE value suggests a lower prediction error and indicates a better model.

The mean average error is denoted as:

$$MAE = \sum_{n=1}^{N} abs(y_i - \hat{y}_i) \tag{8}$$

The mean squared error is denoted as:

$$MSE = \sum_{n=1}^{N} (y_i - \hat{y}_i)^2 \tag{9}$$

4.5.6 *Managing Overfitting*

Overfitting is a common problem in machine learning, and no one-size-fits-all solution exists for managing it. Consequently, a combination of techniques was chosen to enhance the performance and generalization of the models. In addition to employing various preprocessing techniques like feature selection, z-score normalization, and outlier handling, a five-fold nested cross-validation strategy was implemented. This approach served the dual purpose of mitigating overfitting and improving model robustness. To execute the nested cross-validation, the dataset was partitioned into a training set (80%) and a test set (20%) for each iteration. Furthermore, each training set was subdivided into five inner folds. The resulting output displays the R-squared scores for each of the five components, which are stored in an array. These scores provide insight into the overall model performance and facilitate comparisons with other models.

4.5.7 *Feature Importance*

Feature importance was calculated for each model used in this study to determine the contribution of features to the target variable. The scikit-learn library provided the feature importance attributes for ensemble regression models based on Gini impurity, while for linear regression, feature importance was determined using the absolute values of coefficients. This analysis yielded 12 feature importance tables.

5 RESULTS

5.1 *Exploratory Data Analysis*

By utilizing exploratory data analysis and visualization techniques, valuable insights can be derived from the raw dataset. First, individual feature statistics were examined separately; for instance, the person insured's age (Figure 3), car brands (Figure 4), catalog values (Figure 5), pure claim free years (Figure 6) and the dependent variable total (Figure 7) were analyzed. As stated in paragraph 4.4.4, only the most important features are expected to be presented in the main text for presentation purposes.

Figure 3: Distribution Driver Age

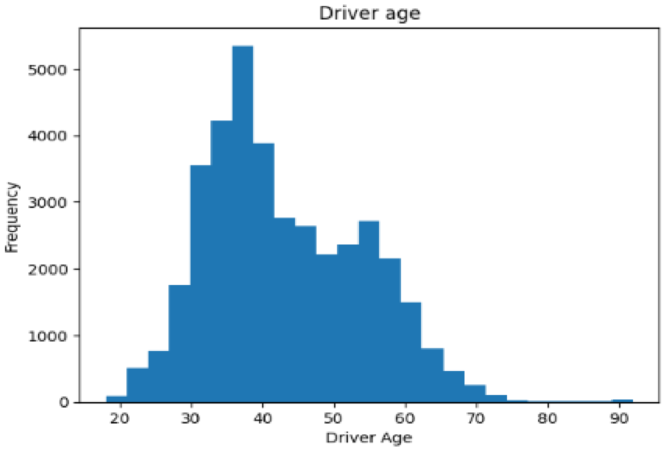


Figure 4: Distribution Car brands

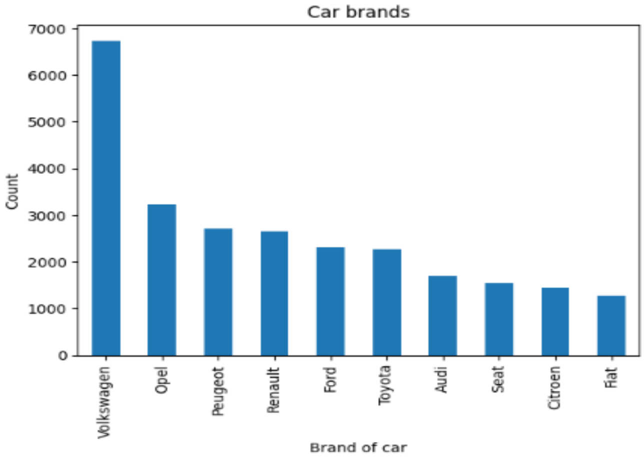


Figure 5: Distribution Catalog Values

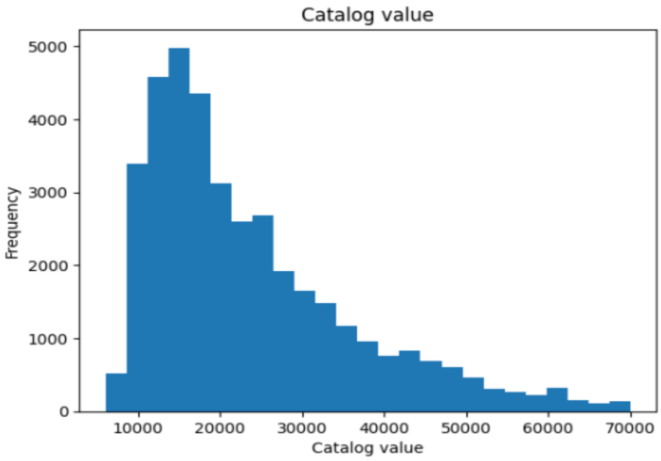


Figure 6: Distribution Pure claim free years

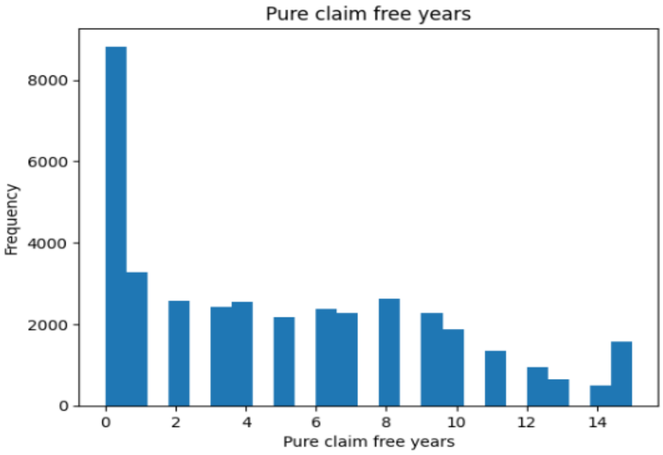
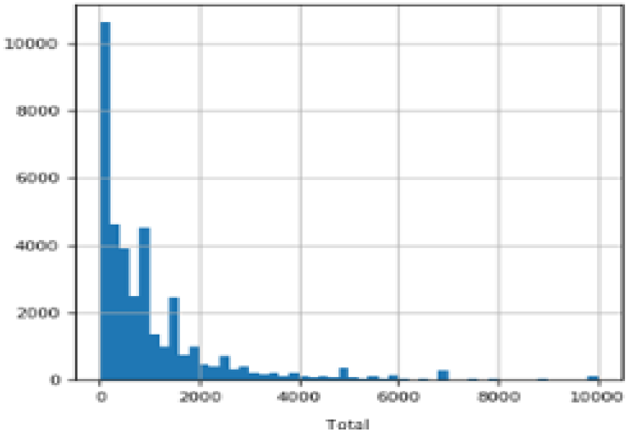


Figure 7: Distribution dependent variable Total



Furthermore, exploratory data analysis and visualization techniques were employed to examine the relationship between individual predictor variables and the dependent variable. Correlation plots were generated for the float and integer features, while boxplots were utilized for the categorical features. It is worth noting that all correlation plots (as can also be seen in Figure 17 and Figure 18 in Appendix C (page 49)) exhibited no discernible linear or nonlinear relationship between the individual feature and the dependent variable. Additionally, the presented boxplots in both Figure 8 and Appendix C (page 49) revealed that the means were lower than the target variable's mean. This finding suggests a relatively weaker predictive relationship between these two variables.

Figure 8: Boxplot Pure claim-free years

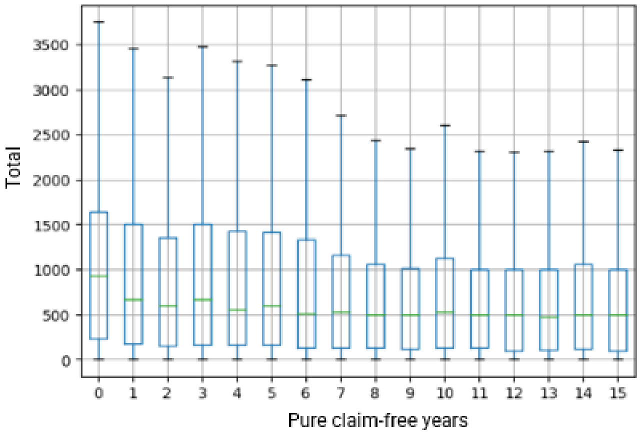
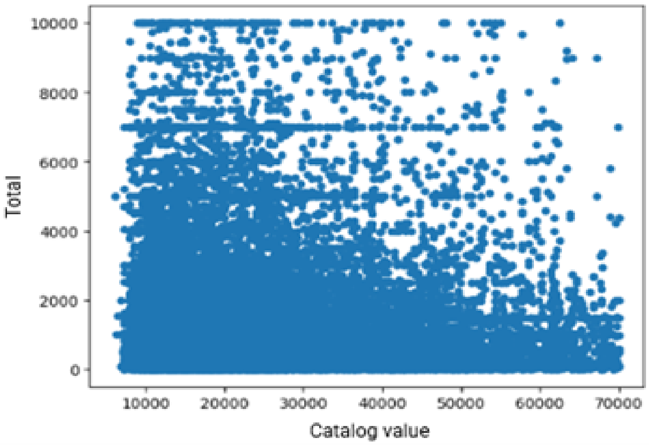
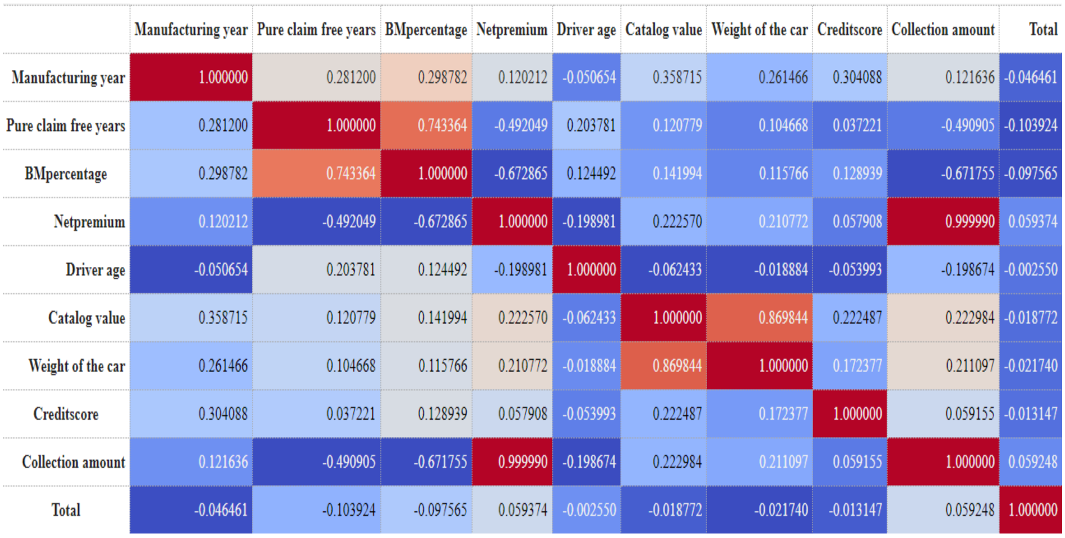


Figure 9: Correlation plot Catalog Value



To investigate the level of correlation based on the previously generated plots, two correlation matrices were constructed. Nevertheless, for presentation, only the correlation matrix (excluding categorical features) is presented in the main text (Figure 10). This visual representation illustrates that all features exhibited either low or slightly negative correlations with the dependent variable.

Figure 10: Correlation matrix



	Manufacturing year	Pure claim free years	BMpercentage	Netpremium	Driver age	Catalog value	Weight of the car	Creditscore	Collection amount	Total
Manufacturing year	1.000000	0.281200	0.298782	0.120212	-0.050654	0.358715	0.261466	0.304088	0.121636	-0.046461
Pure claim free years	0.281200	1.000000	0.743364	-0.492049	0.203781	0.120779	0.104668	0.037221	-0.490905	-0.103924
BMpercentage	0.298782	0.743364	1.000000	-0.672865	0.124492	0.141994	0.115766	0.128939	-0.671755	-0.097565
Netpremium	0.120212	-0.492049	-0.672865	1.000000	-0.198981	0.222570	0.210772	0.057908	0.999990	0.059374
Driver age	-0.050654	0.203781	0.124492	-0.198981	1.000000	-0.062433	-0.018884	-0.053993	-0.198674	-0.002550
Catalog value	0.358715	0.120779	0.141994	0.222570	-0.062433	1.000000	0.869844	0.222487	0.222984	-0.018772
Weight of the car	0.261466	0.104668	0.115766	0.210772	-0.018884	0.869844	1.000000	0.172377	0.211097	-0.021740
Creditscore	0.304088	0.037221	0.128939	0.057908	-0.053993	0.222487	0.172377	1.000000	0.059155	-0.013147
Collection amount	0.121636	-0.490905	-0.671755	0.999990	-0.198674	0.222984	0.211097	0.059155	1.000000	0.059248
Total	-0.046461	-0.103924	-0.097565	0.059374	-0.002550	-0.018772	-0.021740	-0.013147	0.059248	1.000000

Furthermore, an examination was performed to identify the features with the greatest impact on the dependent variable, based on the total correlation matrix including categorical variables. Table 5 presents an overview of the top three features. Moreover, it can be observed that the absolute values of these correlations are relatively small, indicating weak correlations between these variables and the dependent variable.

Table 5: Most important features relative to target

Feature	Correlation
Form of insurance WA	0.080698
Netpremium	0.059374
Region 5.0	0.025664

Finally, an analysis of the correlation matrix reveals that certain features exhibited high correlations with each other, indicating redundancy in their predictive power. In such cases, it is advisable to either merge or remove these features to improve the predictor's performance. Highly correlated features tend to contain redundant information, and a threshold of 0.95 was therefore applied to identify variables with strong correlations, suggesting that they may be measuring the same underlying phenomenon. This threshold led to the identification of a correlation between the collection amount and the net premium. The decision was taken to include net premiums in the study's scope.

5.2 *Results*

5.2.1 *Results of Experiment 1*

Experiment 1 used only the features of the actuarial profile in all algorithms. The optimal values from the hyper-parameter tuning were used to achieve the optimal score in the test set (Table 6).

Table 6: Optimal values experiment 1 Hyperparameter tuning

Algoritme	Optimal value
Linear regression	10
	'Huber'
	'elastic net'
	'constant'
Regression tree	6
	30
	10
Random Forest Regressor	100
	7
	7
	7
	log2
XGBoost Regressor	50
	7
	40
	None

The evaluation metrics in table 7 show that the random forest regressor had the lowest MAE, MSE, and RMSE scores and was the most accurate. Linear regression had the highest MAE, MSE, and RMSE of all models. However, all models have a low R2 score which prevents the models from adequately predicting the observed variability. Although cross-validation shows some improvement in the R2 values, the values remain close to 0 (Table 16 in Appendix B, page 9).

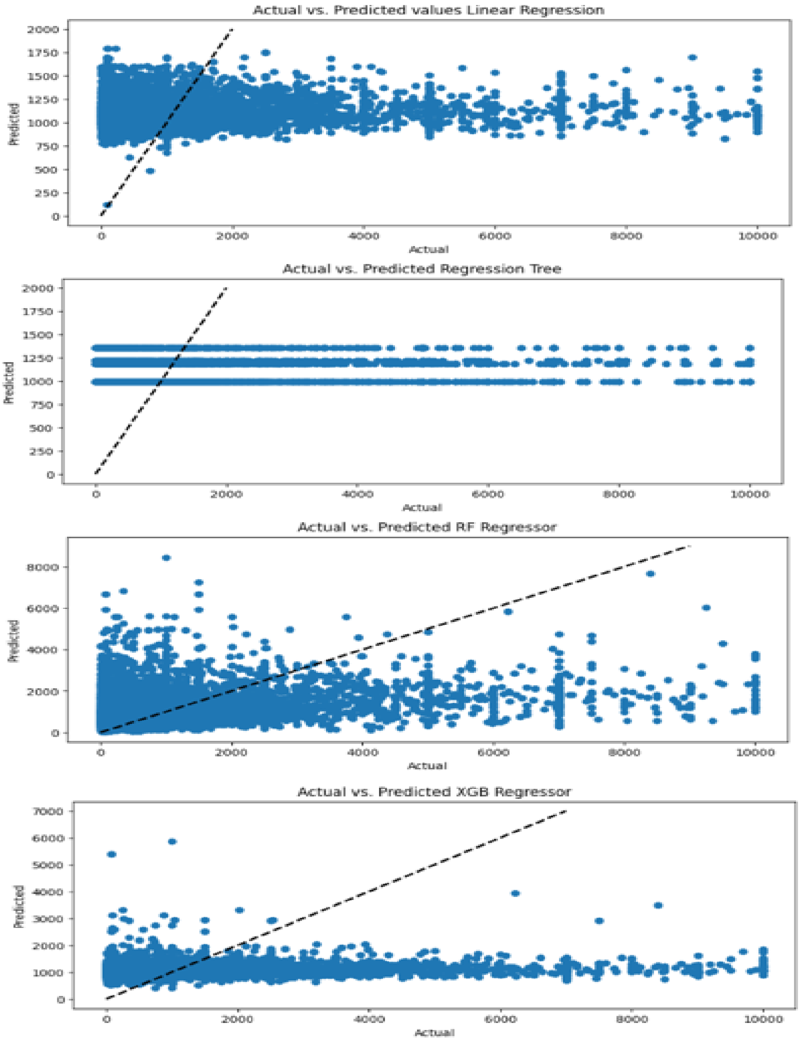
Table 7: Results experiment 1

	MAE	MSE	RMSE	R2
Linear regression	978.66	2256389.04	1502.12	0.005
Regression Tree	978.14	2249012.08	1499.67	0.005
RF Regressor	949.79	2195759.32	1481.80	0.032
XGB Regressor	957.03	2206515.66	1485.43	0.027

When solely examining the absolute values from Experiment 1, it appears that the random forest regressor yielded the most accurate predictions based on the actuarial profile, as indicated by its relatively low MAE (the error metric employed in all experiments). Nonetheless, upon closer examination of the graphical representations depicting the plotted predictions against the actual values (Figure 11), several noteworthy observations emerged.

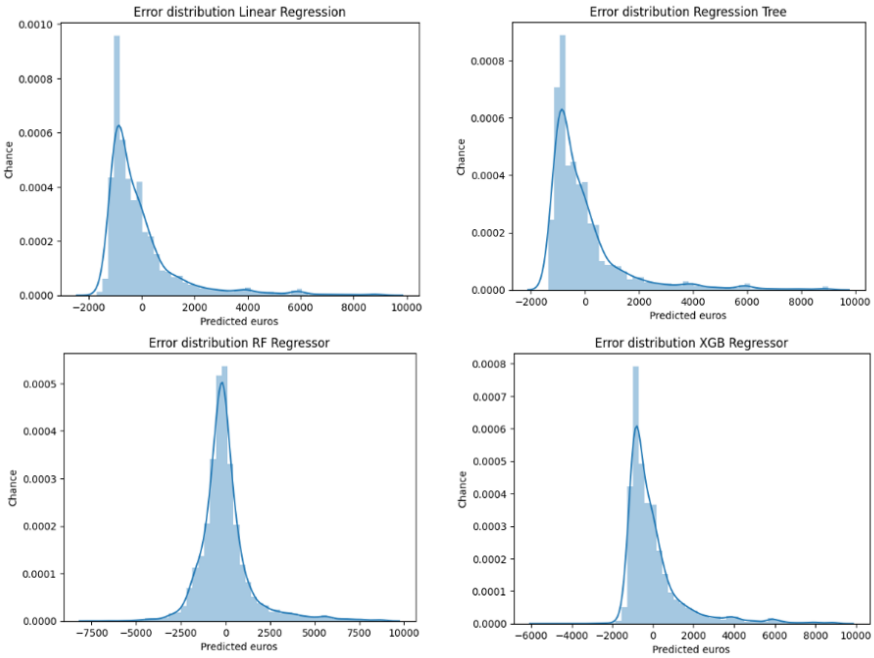
First, it is evident that random forest regressor exhibited substantially higher predictions for the claim rate than linear regression, regression tree, and XGBoost regression. The random forest regressor demonstrated a wider dispersion of predicted values, starting from zero. The predictions of linear regression, regression tree and XGBoost regressor tended to cluster around the mean of the dependent variable, while the actual costs displayed a broader range. Additionally, the visualizations revealed a long tail for nearly all predicted values, with none of the algorithms displaying predictions tightly clustered around the identity line. Consequently, the initial accuracy suggested by the MAE was contradicted by the divergent patterns observed in the plots (Figure 11 and Figure 12).

Figure 11: Actual versus Predicted plots experiment 1



Furthermore, an error distribution was generated to assess the disparities between the algorithms' actual and predicted values (Figure 12). In this distribution, the y-axis denotes the probability of overestimation or underestimation of the euro amount, while the x-axis represents the respective deviations. Although all models exhibited a similar pattern in this regard, the random forest stood out as appearing more precise than the other algorithms.

Figure 12: Error distribution between actual and predicted values experiment 1



Feature Importance
The most impactful features of the target variable were determined for each model; table 8 displays the top three features for each. Notably, the ensemble methods and regression tree identified catalog value (F2) and weight of the car (F3) as the most important features (albeit with different values). In contrast, linear regression had a completely different set of most important features for the target variable.

Table 8: Top 3 features ranked by importance experiment 1

Linear Regression	Regression Tree	RF Regressor	XGB Regressor
F7= 0.17	F2 = 0.25	F3 = 0.16	F2 = 0.12
F6 = 0.10	F3 = 0.21	F2 = 0.15	F3 = 0.08
F15 = 0.1	F0 = 0.07	F1 = 0.13	F1 = 0.07

F0 = Manufacturing year, F1 = Driver age , F2 = Catalogvalue, F3 = Weight of the car, F6 = Type of fuel A, F7 = Type of fuel AG, F15 = Type of fuel G

5.2.2 *Results of Experiment 2*

Experiment 2 did not use the features of the actuarial profile in all algorithms. The optimal values from the hyper-parameter tuning are used to achieve the optimal score in the test set (Table 9).

Table 9: Optimal values experiment 2 Hyperparameter tuning

Algoritme	Optimal value
Linear regression	10
	'squared_loss'
	'l2
	'constant'
Regression tree	4
	10
	10
Random Forest Regressor	90
	7
	7
	7
	Log2
XGBoost Regressor	50
	5
	40
	None

The evaluation metrics in table 10 show that the XGBoost regressor achieved the lowest MAE, MSE, and RMSE scores in this experiment. The random forest regressor was second in terms of accuracy of the MAE value. The MAE of the regression tree and linear regression were very similar, with regression tree narrowly showing a better score.

Table 10: Results Experiment 2

	MAE	MSE	RMSE	R2
Linear regression	968.31	2285351.93	1511.73	0.015
Regression Tree	967.27	2280302.10	1510.06	0.015
RF Regressor	961.00	2310938.72	1520.17	0.004
XGB Regressor	943.72	2241125.38	1497.03	0.034

Nonetheless, all models again had low R2 scores in this second experiment, which means that the models could not adequately predict the observed variability. Despite cross-validation showing some improvement in the R2 values, the values also remained close to 0 in this experiment (Table 17 in Appendix B, page 49).

Upon examining the MAE values from the second experiment, it became evident that they were similar to the MAE values obtained in the first experiment; however, the XGBoost regressor demonstrated a marginally higher level of accuracy in this case. Nonetheless, when considering the graphical representations (Figure 13), a contrasting picture emerged compared to the absolute values presented in table 10. Furthermore, it is notable that the plots exhibited a remarkable similarity to those generated in the first experiment; this resemblance also extended to the error distributions (Figure 14).

Figure 13: Actual versus Predicted plots experiment 2

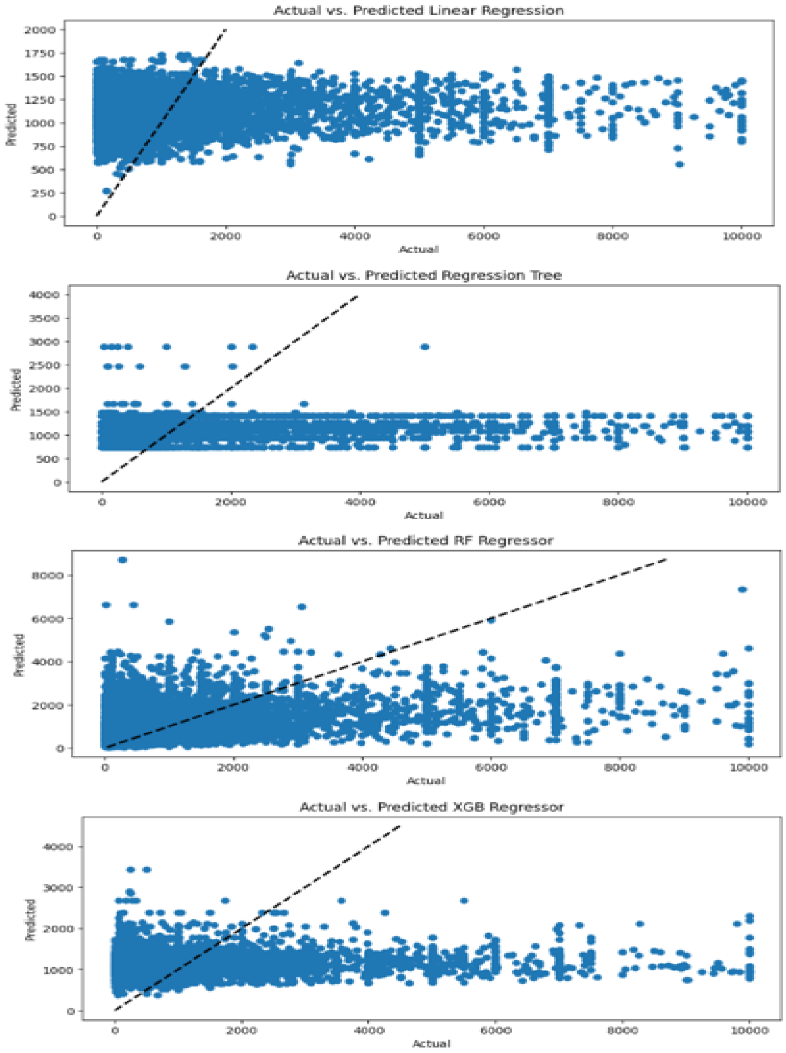
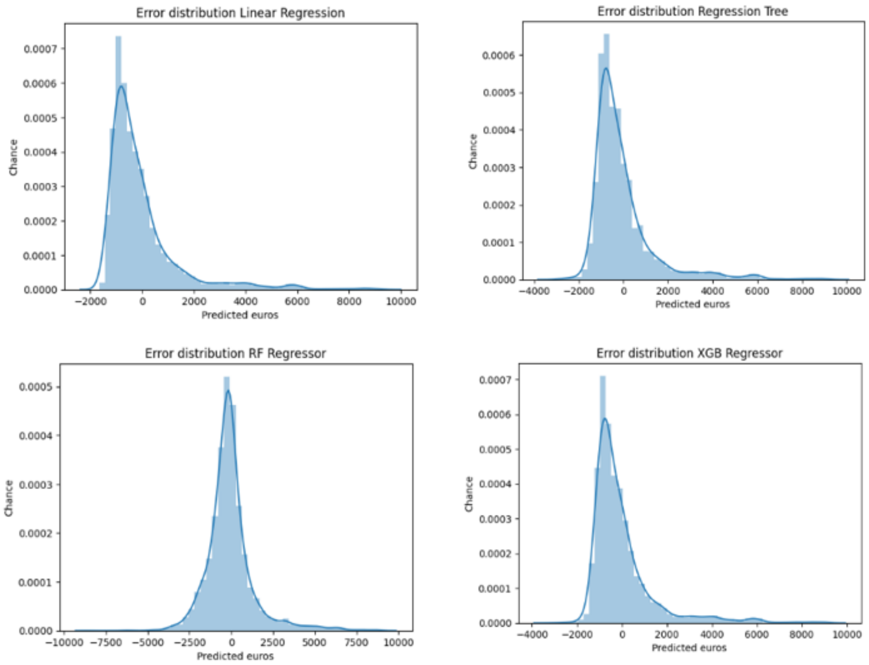


Figure 14: Error distribution between actual and predicted values experiment 2



Feature Importance

The influential characteristics of each model on the target variable were identified and summarized in table 11. Notably, the order of feature importance was identical for the regression tree and random forest regressor, although the values differed. In contrast, the linear regression and XGBoost regressor had a distinct set of most important features compared to the other methods.

Table 11: Top 3 features ranked by importance experiment 2

Linear Regression	Regression Tree	RF Regressor	XGB Regressor
F63 = 0.16	F2 = 0.26	F2 = 0.28	F7 =0.18
F64 = 0.16	F1 = 0.07	F1 = 0.08	F6 = 0.14
F60 = 0.10	F0 = 0.07	F0 = 0.07	F41 = 0.13

F0 = Pure claim-free years, F1 = BM percentage, F2 = Net premium, F6 = Type of Insurance Limited Comprehensive (Casco), F7 = Car brand Alfa, F41 = Car brand Nissan, F60 = Breakdown service in the Netherlands No, F63 = Legal aid Yes, F64 = Legal aid No

5.2.3 *Results of Experiment 3*

Experiment 3 was a combination of the features from Experiments 1 and 2 with the target variable. The optimal values obtained from the hyperparameter tuning were used to achieve the best performance on the test set (Table 12).

Table 12: Optimal values experiment 3 Hyperparameter tuning

Algoritme	Optimal value
Linear regression	10
	'Huber'
	'elastic net'
	'constant'
Regression tree	4
	30
	20
Random Forest Regressor	90
	7
	7
	7
	Log2
XGBoost Regressor	50
	5
	40
	None

The evaluation metrics, as summarized in table 13, demonstrate that the XGBoost regressor achieved the lowest scores for MAE, MSE, and RMSE in this experiment, followed by the RF regressor. The regression tree ranked third, while the linear regression model exhibited the highest MAE value. All models in this final experiment have poor R2 scores, and cross-validation barely makes an impact (Table 18 in Appendix B, page 49). It is noteworthy that, even though these evaluation metric values were very similar to the previous two experiments, upon analyzing the graphical representations in this final experiment (Figure 15), it again became apparent that a divergent pattern emerged compared to the absolute values reported in table 13. Moreover, it is worth noting that the plots exhibited a striking resemblance to those observed in the two preceding experiments (Figure 15), which was likewise reflected in the error distributions (Figure 16).

Table 13: Results Experiment 3

	MAE	MSE	RMSE	R2
Linear regression	966.06	2204267.39	1484.67	0.012
Regression Tree	965.23	2206055.12	1485.27	0.012
RF Regressor	958.48	2241058.78	1497.01	0.004
XGB Regressor	944.50	2168744.31	1472.66	0.028

Figure 15: Actual versus Predicted plots experiment 3

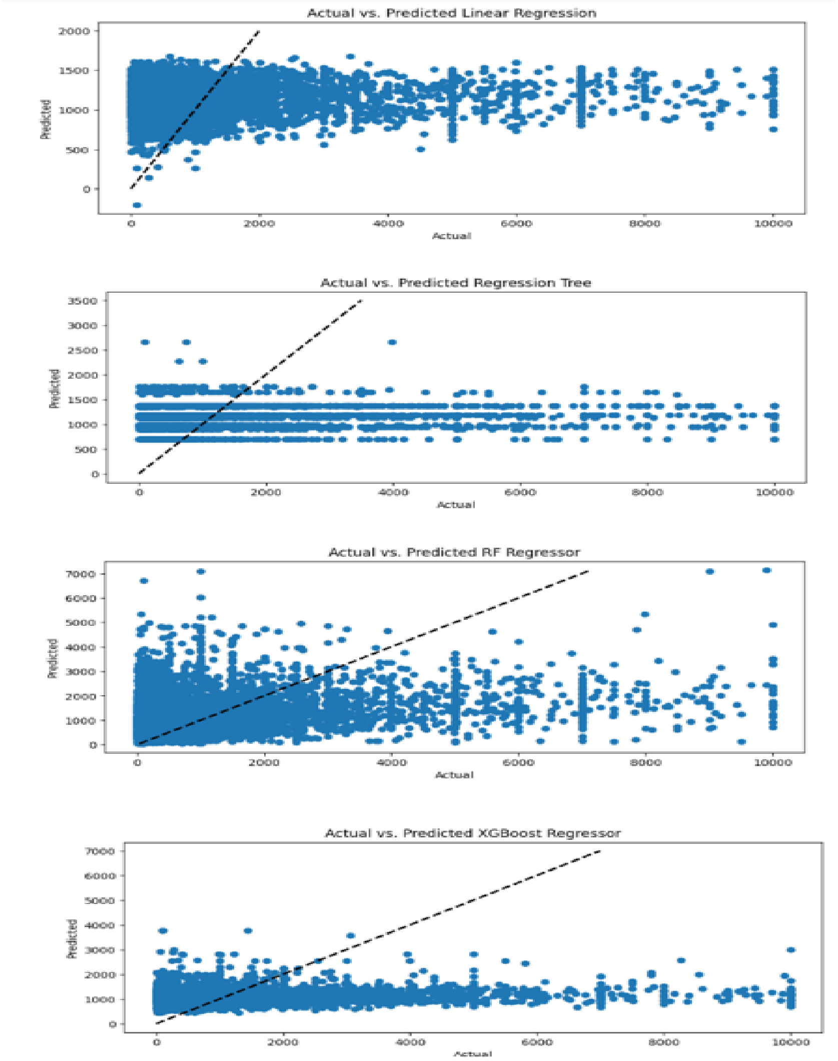
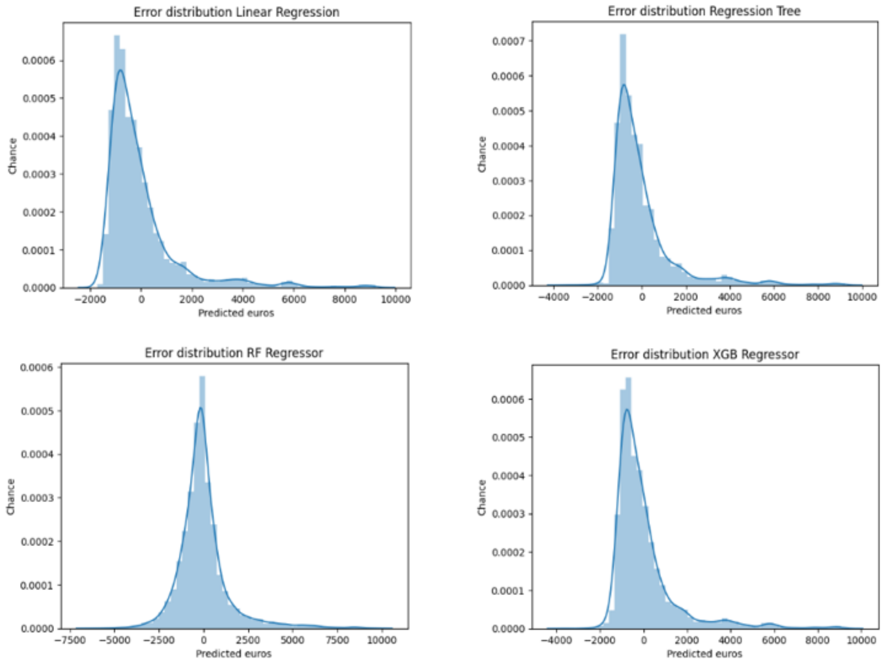


Figure 16: Error distribution between actual and predicted values experiment 3



Feature Importance

In addition to evaluating the models' performance, an analysis of feature importance was conducted to identify the variables with the most significant impact on the target variable. The top three most important features for each model are presented in table 14. Interestingly, the regression tree and random forest regressors showed the same order of feature relevance, albeit with varying magnitudes; however, the linear regression and XGBoost regressor models prioritized features beyond the actuarial profile, suggesting a shift in their focus toward other variables.

Table 14: Top 3 features ranked by importance experiment 3

Linear Regression	Regression Tree	RF Regressor	XGB Regressor
F26 = 0.16	F4 = 0.13	F4 = 0.10	F25 = 0.19
F60 = 0.06	F5 = 0.11	F5 = 0.09	F37 = 0.14
F52 = 0.04	F2 = 0.10	F2 = 0.08	F60 = 0.12

F2 = BM percentage, F4 = Driver age, F5= Catalogvalue, F25 = Type of Insurance Full Comprehensive (Casco), F26 = Car brand Alfa, F37 = Car brand Dacia, F52 = Car brand MCC , F60 = Car brand Nissan

5.3 *Comparison of Results*

In Experiment 2, which incorporates features beyond the actuarial profile, the XGBoost regressor emerged as the algorithm with the lowest mean absolute error (MAE) and the best prediction accuracy in two out of three experiments. Nonetheless, the difference in performance between the algorithms in the experiments was minimal, indicating a relatively similar predictive capability across the models.

While the MAE scores provided a useful measure of the average deviation of the models, a more comprehensive analysis of the graphical representations revealed a distorted picture. In all experiments, there was little to no apparent relationship between the independent variables and the dependent variable. The predictions of all algorithms in each experiment tended to cluster around the mean of the target variable, rather than accurately capturing the correlation and composition of the independent features. Moreover, the predictions demonstrated a narrow variance and did not align well with the desired identity line.

Additionally, when examining the feature importance results, it is notable that in the final experiment, five out of the nine features (which were also included in the previous two experiments) were identified as being most important.

6 DISCUSSION

The main goal of this thesis, **RQ1**, was to assess the effectiveness of machine-learning techniques in forecasting claim amounts in the private automobile industry. Various algorithms were employed to effectively address this research question.

Notably, this thesis aligns with recent research suggesting that ensemble methods hold promise for such investigations (Baker et al., 2020; Quan & Valdez, 2018), as the experiments consistently demonstrated superior performance by one of the two ensemble methods for regression. On average, the XGBoost regressor yielded the best MAE results, with a minimum score of 943.72 (Experiment 2), initially appearing to be accurate. While this thesis primarily focused on using MAE, as recommended by similar studies in the field of health insurance Kowshalya and Nandhini (2018), the results are promising at first glance, with higher MSE and RMSE values. These metrics assign greater weight to significant errors in their mean calculation, which is a common phenomenon when handling financial data expressed in thousands of euros.

Nevertheless, upon examining the graphical visualizations and observed R2 values, it became apparent that the algorithms' predictive performance was inadequate. It appears that the models rely predominantly on values around the mean of the dependent variable (between 750 and 1,750 euro) rather than incorporating a combination of independent variables. Given the low correlations observed during the exploratory data analysis phase, one possible explanation is that the models overlook the independent variables, regardless of their combination, leading to an anticipated improvement in predictive performance. The low correlations with the dependent variable (below 0.15) may be attributed to a shortage of data points and/or the inclusion of irrelevant independent variables.

No research to date exists on the specific scope of this thesis that pays attention to limited interactions; however, this aspect is of significant interest, considering that the dataset comprises historical information and rather dated features. It indirectly relates to societal changes caused by national and global challenges, such as the rise of remote work due to the pandemic, which has had an impact on mobility and driving behavior. Currently, this data is neither collected nor monitored.

By investigating the contribution of actuarial profile features and non-actuarial features to the predictive performance of the machine learning techniques **RQ2**, notable insights were gained. While previous studies such as Bian et al. (2018), Dal Pozzolo et al. (2011), Duval and Pigeon (2019), and Guelman (2012) have explored related variables, this study incorporated a combination of features, including specific elements unique to the Netherlands (such as the breakdown service) that have not been extensively examined in prior research. Although Experiment 2 demonstrated the best MAE results, the overall findings across experiments did not exhibit significant variations, as indicated by the visual representations in the previous results section. Moreover, the feature importance analysis did not reveal substantial discrepancies between the impact of actuarial and non-actuarial features.

6.1 *Limitations and Implications for Further Research*

Firstly, this study recommends future research that takes a new critical look at feature engineering and current society. The dataset used for this research comprises only historical features from 11 years, which is no longer fully representative of the modern-day world. This is also seen in the exploratory data analysis and result sections of this thesis. Due to feature engineering, performance may increase.

Furthermore, it is recommended to gain insights into telematics data. By incorporating information on driving behavior (such as distance traveled, road types, and time spent on the road) insurers can strive for a 'clean portfolio' and effectively reward safe driving practices. This measure would align with current societal trends and leverage advancements in car technology, as newer vehicles equipped with automated safety features (for example, signal lights or sound) can potentially reduce accidents and claim costs.

Finally, it is recommended to obtain more individual-level datasets in the private automobile industry for training complex models, improving variability, and achieving accurate results. The availability of sufficient and additional data allows for exploring the performance impact of deep learning models, like neural networks, in forecasting real-world financial data. Deep learning models also offer the advantage of data augmentation, enabling exhaustive utilization of synthetic and real data (Shah & Shroff, 2021). In addition, investigating autoregressive models such as ARIMA would be worthwhile for financial data forecasting.

6.2 *Implications for Broader Society*

The impact of the machine learning algorithms will save time and money for both policyholders and the authorized agent. This research can guide insurance companies in making informed decisions about implementing machine learning algorithms for claim prediction and help them allocate resources more efficiently. In addition, the findings regarding the contribution of actuarial profile features and non-actuarial features have implications for insurers in terms of their data collection and risk assessment practices. Although the importance of individual features might not significantly vary, it is crucial to consider other independent variables such as emerging societal factors and technological advancements.

Consequently, this thesis contributes to the emerging field of machine learning approaches in the private automobile industry, as no comparable research has been conducted in the Netherlands to date.

7 CONCLUSION

In conclusion, this thesis aimed to assess the reliability of machine learning techniques in forecasting claim amounts in the private automobile insurance industry. Despite conducting various experiments, the results obtained are specific to the dataset used in this study and may not be generalizable to other datasets. While the mean absolute error (MAE) values initially provided a realistic view of the predictive performance, the R2 metric and graphical representations did not accurately reflect the models' performance. Additionally, the similarity of results across all the experiments can be attributed to the limited influence of independent variables, both within and outside the actuarial profile, on the dependent variable. The restricted availability of data and low interactions among the independent variables hindered the complex algorithms' ability to achieve optimal results.

One of the most challenging aspects of this study was preprocessing the data and creating a suitable dataset for different algorithms. The dataset contained missing values, outliers, and noise, and the algorithms struggled to handle both categorical and continuous variables simultaneously. Consequently, categorical variables had to be transformed, and various preprocessing techniques (including feature selection, z-score normalization, handling outliers, and five-fold nested cross-validation) were employed to manage overfitting.

Given the immense potential of machine learning techniques in the insurance industry, the conclusions of this thesis highlight the crucial need for further research and development of machine learning algorithms tailored specifically to the complexities of the private automobile insurance industry. Future studies should explore alternative models, innovative feature engineering techniques, and data augmentation approaches to overcome the limitations observed in this research and enhance the architecture of the algorithms.

8 REFERENCES

REFERENCES

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 265–283.

Anggoro, D., & Supriyanti, W. (2019). Improving accuracy by applying z-score normalization in linear regression and polynomial regression model for real estate data. *International Journal of Emerging Trends in Engineering Research*, *7*(11), 549–555.

Baker, R., Forrest, D., & Pérez, L. (2020). Modelling demand for lotto using a novel method of correcting for endogeneity. *Economic Modelling*, *84*, 302–308.

Bian, Y., Yang, C., Zhao, J. L., & Liang, L. (2018). Good drivers pay less: A study of usage-based vehicle insurance models. *Transportation research part A: policy and practice*, *107*, 20–34.

Bisong, E., et al. (2019). *Building machine learning and deep learning models on google cloud platform*. Springer.

Blier-Wong, C., Cossette, H., Lamontagne, L., & Marceau, E. (2020). Machine learning in p&c insurance: A review for pricing and reserving. *Risks*, *9*(1), 4.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*, 123–140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Cummins, J. D., & Griepentrog, G. L. (1985). Forecasting automobile insurance paid claim costs using econometric and arima models. *International Journal of Forecasting*, *1*(3), 203–215.

Dal Pozzolo, A., Moro, G., Bontempi, G., & Le Borgne, D. Y. A. (2011). Comparison of data mining techniques for insurance claim prediction. *Universita degli Studi di Bologna*.

David, M. (2015). Auto insurance premium calculation using generalized linear models. *Procedia Economics and Finance*, *20*, 147–156.

Dhieb, N., Ghazzai, H., Besbes, H., & Massoud, Y. (2019). Extreme gradient boosting machine learning algorithm for safe auto insurance operations. *2019 IEEE international conference on vehicular electronics and safety (ICVES)*, 1–5.

Diepgrond, D. (2020). *Can prediction explanations be trusted? on the evaluation of interpretable machine learning methods* (Doctoral dissertation).

Duval, F., & Pigeon, M. (2019). Individual loss reserving using a gradient boosting-based approach. *Risks*, *7*(3), 79.

Fauzan, M. A., & Murfi, H. (2018). The accuracy of xgboost for insurance claim prediction. *Int. J. Adv. Soft Comput. Appl*, *10*(2), 159–171.

Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, *37*(5), 920–964.

Grize, Y.-L., Fischer, W., & Lützelschwab, C. (2020). Machine learning applications in nonlife insurance. *Applied Stochastic Models in Business and Industry*, *36*(4), 523–537.

Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, *39*(3), 3659–3667.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2

Herrmann, H., & Masawi, B. (2022). Three and a half decades of artificial intelligence in banking, financial services, and insurance: A systematic evolutionary review. *Strategic Change*, *31*(6), 549–569.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, *9*(3), 90–95.

Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums. *International Journal of Environmental Research and Public Health*, *19*(13), 7898.

Kowshalya, G., & Nandhini, M. (2018). Predicting fraudulent claims in automobile insurance. *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 1338–1343.

Krasheninnikova, E., Garcıa, J., Maestre, R., & Fernández, F. (2019). Reinforcement learning for pricing strategy optimization in the insurance industry. *Engineering applications of artificial intelligence*, *80*, 8–19.

McKinney, W., et al. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011).

Scikit-learn: Machine learning in python. *Journal of machine learning research*, *12*(Oct), 2825–2830.

Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—xgboost versus logistic regression. *Risks*, *7*(2), 70.

Quan, Z., & Valdez, E. A. (2018). Predictive analytics of insurance claims using multivariate decision trees. *Dependence Modeling*, *6*(1), 377–407.

Ramzai, J. (2019). Simple guide for ensemble learning methods. *Towards Data Science*.

Samson, D., & Thomas, H. (1987). Linear models as aids in insurance decision making: The estimation of automobile insurance claims. *Journal of Business Research*, *15*(3), 247–256.

Selvakumar, V., Satpathi, D. K., Praveen Kumar, P., & Haragopal, V. (2021). Predictive modeling of insurance claims using machine learning approach for different types of motor vehicles. *Univers J Acc Finance*, *9*(1), 1–14.

Shah, V., & Shroff, G. (2021). Forecasting market prices using dl with data augmentation and meta-learning: Arima still wins! *arXiv preprint arXiv:2110.10233*.

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2019). A comparative analysis of forecasting financial time series using arima, lstm, and bilstm. *arXiv preprint arXiv:1911.09512*.

Spedicato, G. A., Dutang, C., & Petrini, L. (2018). Machine learning methods to perform pricing optimization. a comparison with standard glms. *Variance*, *12*(1), 69–89.

United Nations. (n.d.). World population to reach 8 billion on 15 november 2022.

Vandrangi, S. K. (2022). Predicting the insurance claim by each user using machine learning algorithms. *Journal of Emerging Strategies in New Economics*, *1*(1), 1–11.

Wang, Z., et al. (2017). *Loss pattern recognition and profitability prediction for insurers through machine learning* (Doctoral dissertation). Massachusetts Institute of Technology.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., Martin, M., . . . Qalieh, A. (2017). *Mwaskom/seaborn: V0.8.1 (september 2017)* (Version v0.8.1). Zenodo. https://doi.org/10.5281/zenodo.883859

Wüthrich, M. V. (2018). Machine learning in individual claims reserving. *Scandinavian Actuarial Journal*, *2018*(6), 465–480.

Yu, W., Guan, G., Li, J., Wang, Q., Xie, X., Zhang, Y., Huang, Y., Yu, X., & Cui, C. (2021). Claim amount forecasting and pricing of automobile insurance based on the bp neural network. *Complexity*, *2021*, 1–17.

Zheng, L., & Guo, L. (2020). Application of big data technology in insurance innovation. *International conference on education, economics and information management (ICEEIM 2019)*, 285–294.

9 APPENDICES

APPENDIX A

Table 15: Examples from the selected features of the dataset

Example	Policynumber	Type of insurance	Manufacturing year
1)	5160100539	WA	2009
2)	5160113338	WA casco	2010

Example	Region	BM percentage	Net premium (€)
1)	3	75	520,08
2)	5	75	228,24

Example	Legal aid	Breakdown NL	Extra Breakdown	Driver age
1)	No	No	No	36
2)	No	Yes	No	53

Example	Catalog value (€)	Weight of the car	Credit score	Collection amount
1)	44.729	1785	53	629,28
2)	10.850	840	0	280,44

Example	Car brand	Total
1)	Mitsubishi	3250,00
2)	Ford	1864,58

APPENDIX B

Table 16: Cross-validation experiment 1

	R2	CV1	CV2	CV3	CV4	CV5
Linear regression	0.01	0.01	0.01	0.00	0.00	0.01
RT	0.01	0.06	0.01	0.07	0.01	0.08
RF Regressor	0.04	0.06	0.04	0.08	0.001	0.08
XGB Regressor	0.03	0.07	0.04	0.08	0.09	0.09

Table 17: Cross-validation experiment 2

	R2	CV1	CV2	CV3	CV4	CV5
Linear regression	0.01	0.01	-0.09	0.05	0.01	-0.07
RT	0.01	-0.02	0.06	0.01	0.04	0.04
RF Regressor	0.06	0.02	0.07	-0.11	0.06	0.04
XGB Regressor	0.03	0.04	0.07	-0.09	0.06	0.05

Table 18: Cross-validation experiment 3

	R2	CV1	CV2	CV3	CV4	CV5
Lineare regressie	0.01	0.01	0.05	0.04	-0.10	-0.10
RT	0.01	0.02	0.03	0.01	-0.10	-0.20
RF Regressor	0.04	0.03	0.03	0.03	0.01	0.01
XGB Regressor	0.02	0.03	0.04	0.04	0.01	0.01

APPENDIX C

Due to presentation purposes, only a snapshot of the correlation plots and boxplots is displayed in Appendix C. This is because the dataset contains 91 columns, and all the plots exhibit a similar graphical representation.

Figure 17: Additional correlation plots EDA phase

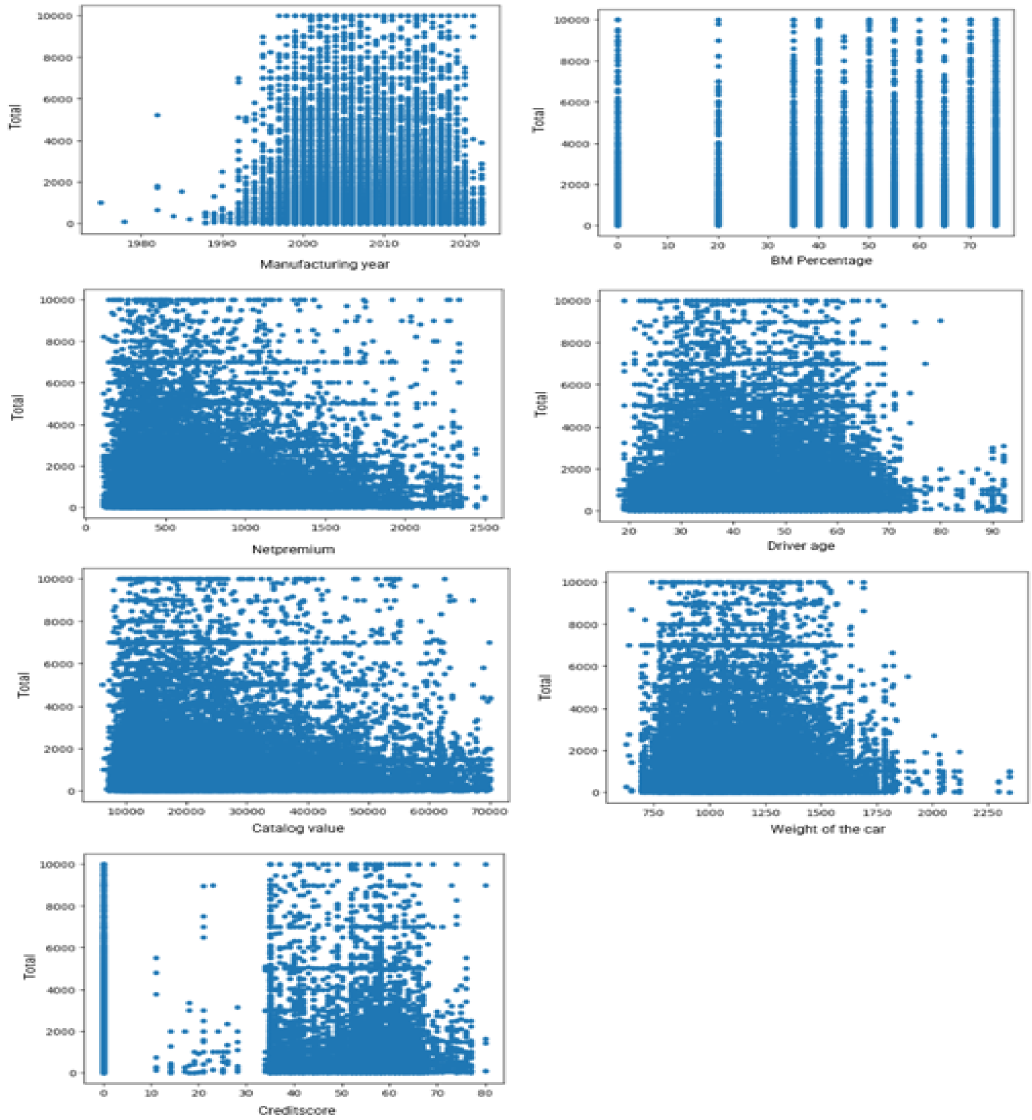
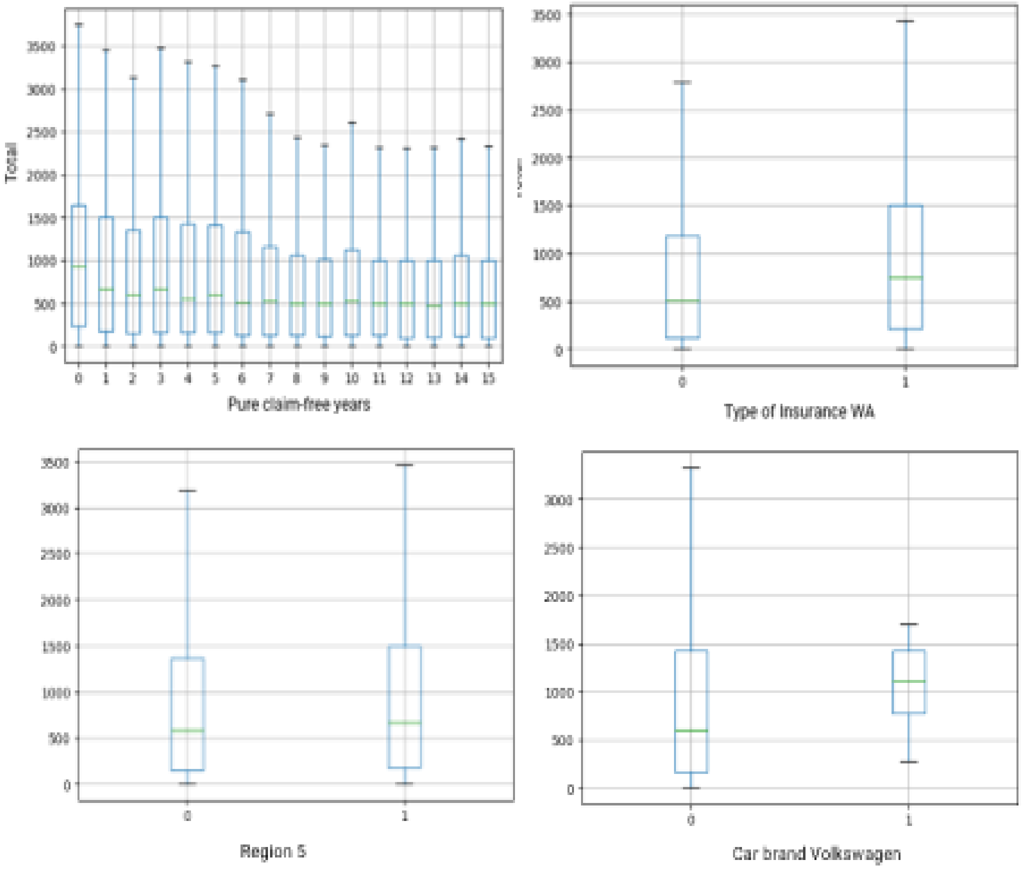


Figure 18: Additional boxplots EDA phase


Totall
Pure claim-free years
Type of Insurance WA
Region 5
Car brand Volkswagen