Machine learning methods vs. traditional methods in forecasting loss reserves

Niitta Kotsalo

Master's Thesis

Master of Engineering - Big Data Analytics

2021

MASTER'S THESIS	
Arcada University of Applied Sciences	
Degree Programme:	Master of Engineering - Big Data Analytics
Identification number:	Degree Thesis ID: 8274
Author:	Niitta Kotsalo
Title:	Machine learning methods vs. traditional methods in forecasting loss reserves
Supervisor (Arcada):	Leonardo Espinosa Leal
Commissioned by:	Chubb European Group SE

Abstract:

The purpose and topic of the study is to investigate can machine learning methods provide better estimations of loss reserves compared to traditional chain-ladder method. The aim is to provide the commissioning organisation own automated tools to create the prediction without expert knowledge and reduce manual work of future data analysis. The study is aiming to answer can machine learning method predict outstanding loss reserves and ultimate claims to be paid and can the ML method create better estimates or not. The limitation in insurance industry is that creating estimations of the loss reserves can be challenging due to the random instabilities of claims data. The data used in the research turned out to be smaller than expected, is highly imbalanced and biased, therefore forecasting errors are possible. The dataset includes real-life individual claims data collected by the commissioning company. Quantitative research methods are applied. The ML algorithms used are linear ridge regression and traditional chain-ladder to predict the loss reserves. Logistic regression and random forest for multiclassification are trained to predict the development delay. Model accuracy to actual results and AUC are used to evaluate the models. New research applying modern machine learning methods to address the loss reserving problem is reviewed, and the framework for chain-ladder theory by Mack (1994) presented. It is found that chain-ladder method is simple and can provide accurate predictions on ultimate claims data. Ridge regression results were inaccurate to make predictions from the data. It was able to provide individual claims predictions, therefore the method was not directly comparable with chain-ladder. Logistic regression was able to provide the best result to predict the development delay. In conclusion, the chain-ladder is accurate and easy to apply into actual usage. The machine learning methods can bring new insight from the data as they consider more variables. Data automation and collection of more historical data to make predictions is recommended for future.

Keywords:	chain-ladder, machine learning, loss reserving, insurance, imbalanced dataset
Number of pages:	68
Language:	English
Date of acceptance:	

CONTENTS

1 Introduction 7

1.1 Definitions and table of abbreviations 7

1.2 Background and statement of the problem 8

1.3 Purpose of the study 9

1.4 Research questions 10

1.5 Significance to the field 10

1.6 Limitations 11

1.7 Ethical considerations 12

2 Literature review 14

2.1 Introduction 14

2.2 Related work on the field 14

2.3 Chapter summary 17

3 RESEARCH METHODOLOGY 19

3.1 Methodology 19

3.1.1 Quantitative research methodology 19

3.1.2 Research design 20

3.2 Methods 22

3.2.1 The Chain-Ladder method 22

3.2.2 Linear regression and Ridge regression 24

3.2.3 Logistic regression 25

3.2.4 Random forest classifier 26

3.3 Setting 27

3.4 The dataset 27

3.4.1 Feature engineering 28

3.5 Descriptive data analysis 32

3.5.1 Issues with the data 41

3.5.2 Conclusion 42

4 Results 43

4.1 Predicting loss reserves with chain-ladder method 43

4.2 Predicting the claims amount with Ridge regression 45

4.2.1 Model evaluation and improvement 47

4.2.2 Comparison of chain-ladder vs. ridge regression results 48

4.3 Logistic regression to predict development month 49

4.4 Random forest classifier results 53

4.4.1 Model evaluation and selection 55

5 Discussion 57

5.1 Recommendations for future research 59
6 Conclusion 60
References 61
Appendices 66

LIST OF TABLES AND FIGURES

Table 1. Open source libraries used in data analysis 21
Figure 1. A run-off triangle 23
Figure 2. Chain-Ladder run-off triangle with cumulative payment data with predictions 24
Figure 3: Reserves per occurence period and for total Claims 24
Table 2. Variables of the datasets 28
Table 3. Modified data frame variables and data types 29
Table 4. Analysis of null values in dataset 30
Table 5. Final dataset variables and data types 32
Table 6. Description and basic statistics of the dataset 2016-2019 34
Figure 4. Distribution of Paid euros 2016-2019 35
Figure 5. Payment delay distribution in months 2016-2019 36
Table 7. Value counts of payment_delay per accident year. 36
Figure 6. Payment delay by accident year 37
Figure 7. Boxplot to visualize the value of paid claims and the payment delay 2016-2019 38
Figure 8. Scatter plot of paid euros vs. settlement delay 2016-2019 distribution 39
Figure 9. Scatter plot of paid euros vs. reporting delay 2016-2019 distribution 40
Figure 10. Distribution of paid euros vs. payment delay 2016-2019 41
Table 8. Total claims paid until 11/2020 43
Table 9. Cumulative claims paid until 11/2020 44
Table 10. Completed run-off triangle Chain-Ladder method 44
Table 11. Prediction of outstanding claims and measure of accuracy 45
Table 12. Linear ridge regression test set results 46
Figure 10. Learning curve of ridge regression on the 2016 claims dataset 47
Table 13. Cross-validation results 48
Table 14. Payment delay value counts 2016-2019 49
Table 15 and 16. Training and train-test set scores 50
Figure 11. Coefficients learned by the logistic regression model 51
Figure 12. Confusion matrix predicted vs. actual results 52
Table 17. Classification report logistic regression 53

Table 18. Random forest classifier train-test scores 53
Figure 13. Random forest classifier feature importance 54
Table 19. Grid search with cross validation results 55

1 INTRODUCTION

1.1 Definitions and table of abbreviations

As the topic of the thesis is insurance industry related, the abbreviations and definitions of the industry related terms are introduced first to provide clearer understanding since they are being used in the text frequently.

Accident year: The period when all the claims relating to accidents that occurred for 12 months period are grouped together (Forfar & Raymont, 2002).

Chain-Ladder (CL) method:

> "*The chain ladder or development method is a prominent actuarial loss reserving technique. The chain ladder method is used in both the property and casualty and health insurance fields. Its intent is to estimate incurred but not reported claims and project ultimate loss amounts. The primary underlying assumption of the chain ladder method is that historical loss development patterns are indicative of future loss development patterns*." (The ActuarialClub.com, 2019).

Claim: A payment requested by a policyholder followed by an insured event, but it does not always mean a payment is made to the insured (Forfar & Raymont, 2002).

Incurred but Not Reported (IBNR): IBNR means a reserve account which is used in insurance industry as the facility for claims that are already known, however they have not yet been reported to an insurance company. As these are latent responsibilities, the actuary needs to calculate adequate funds to hold. (Kagan, 2020b).

Loss: The occurrence suffered by the policyholder and what the insurance is intended to cover (Forfar & Raymont, 2002).

Loss reserve: A reserve left aside by insurer with the cedant to cover in part outstanding claims (Forfar & Raymont, 2002). *Outstanding loss reserves (OSLR)*: OSLR means the claims that will be paid but it is unknown how much claims will be paid out on top of this. The final costs of the total amount of claims is therefore unknown until the reserves are released (IRMI, 2020).

1.2 Background and statement of the problem

The thesis topic came as a work-related assignment concerning the need for more thorough analysis of a client accounts portfolio and claims data. The company does not have their own control over the maintenance of the claims data they receive. The accuracy of the data and unpredictability had been issues for several years and there exists a knowledge gap regarding the profitability of the account. It was considered to have enough collected historical data by 2020 to perform deeper analysis of the account. A deeper analysis was needed in order to investigate the product pricing and possibilities to build a better pricing model, study the claims reserving accuracy based on previously paid claims and the frequency of the losses. In addition, it is wanted to automize the data flow and analytical processed in order to reduce manual work from the underwriting department. The problem that was interesting in machine learning point of view and was selected for the thesis topic is the prediction of the loss reserves. It was wanted to examine if machine learning methods and big data analytics can provide a solution to the problem. After discussion with the underwriting and actuary's department of the commissioning organization, it was considered to be useful to create the company's own prediction of the loss reserving which could be then used to estimate and calculate similar account type loss reserving prediction and performance.

Based on the preliminary research around loss reserving in insurance, it was noticed finding the suitable formula to calculate any of the loss reserving related acronyms (IBNR, OSLR, loss reserves) has frequently been one of the hardest challenges in the insurance industry. The reason for this is usually the insurance claim variables are non-normally distributed, which makes their estimation problematic. Not getting them correct has consequences for the insurer as inaccurate estimates can provide incorrect view of the company's health which might lead to harmful actions for the company (Kagan, 2020).

1.3 Purpose of the study

The purpose of this study is to build a development thesis where specific ML-framework method will be used to solve a problem of the commissioning organization.

Since the topic of the thesis relates strongly to actuarial science and insurance, the focus was wanted to be more on the data scientific and the data engineering point of view of the problem. The aim is to keep the predictive part simple and create easily understandable, automated and accurate method for the business and underwriters to use in the future. In addition, the objective in the commissioning organization is to replicate the same method to be used for similar type of data that is handled by external partners.

The purpose is to compare the traditional statistical model with simple ML algorithm to see whether they can produce more accurate prediction of the loss reserves. The aim is to understand the traditional forecasting model called the chain-ladder method which is the most used calculation method within the industry.

The predictive model used to predict the claims monetary amount is done with linear regression model Ridge. Logistic regression and random forest classifier are used to predict the payment delay or development delays (the time between the claim is reported until it is finally paid to the claimant). These approaches are chosen for the research as they are simple to understand and create, the end user in real business setting is not data scientist or actuary. The dataset has claims observations from five which was expected to be adequate amount of data in order to create the chain-ladder calculations and build the fore mentioned ML methods in Python. Once the models have been trained a comparison of the accuracy between the results will be viewed and the most appropriate model selected.

1.4 Research questions

1. Can machine learning method predict outstanding loss reserves and ultimate claims to be paid? Will the ML method create better estimates, and can it be used in real world business setting?
2. What are the advantages/disadvantages of ML methods vs. statistical chain-ladder method in forecasting loss reserves and ultimate claims?

1.5 Significance to the field

Based on the literature review and previous studies in the field, the actuary science in insurance relies on traditional mathematical models in predicting the loss reserves. The traditional calculations are usually based on combined claims data which are structured in triangular shape, the most used methods are the chain-ladder and Borhuetter-Ferguson method (Wüthrich, 2018). Studying and using the modern machine learned predictive methods based on individual claims data have been introduced during the past few years (Wüthrich, 2018).

Most of the research papers referred in this thesis have been trained with artificial datasets. The significance of this research paper to the industry is that the predictive model will be trained with real life dataset. Usually the studies discuss the P&C (property and casualty) insurance (meaning car or home insurances for example), this thesis is investigating the reserving problem in Accidental & Health insurance, which means the type of insurances individuals can take for themselves through an employer, an association or individually to cover from injuries or illnesses. The predictive models and the statistical method will be compared to the actual results to investigate the accuracy of each method.

1.6 Limitations

The main limitation of this research was the small dataset. There were 8351 rows of data and many missing datapoints or blank values. The dataset was also highly imbalanced; therefore, many algorithms were tested to see which ones produce the most accurate outcome. In several cases, there was no useful outcome. In order to overcome these issues, lot of feature engineering and data clean-up were made, and ML methods changed. Before feature engineering, further information was needed to be requested from the stakeholders in order to understand certain values and why there was significant amount of null values in the data. Data cleaning was done several times throughout the different phases of the analysis and model training process, different data issues was faced throughout the project. It is worth mentioning in the limitations that in insurance industry it is commonly known that making estimations can be challenging due to the random fluctuations in claims data (Kagan, 2020a). To overcome the challenge of the data's random fluctuations, Friedland (2010) wrote about evaluation of the basic techniques used to estimate unpaid claims. Friedland (2010) used in their research numerous methodologies for the same examples, they state actuaries should use more than one method when analysing unpaid claims as no single method can produce the best estimation in every situation. For this reason and due to the imbalanced and small dataset, more than one algorithm was chosen to be trained. The dataset was also needed to be split into subsets. Therefore, it was decided to train the simpler models, and not the more sophisticated ones as suggested by the literature, in order to find solution to the problem.

In addition, as it was mentioned by Wüthrich (2018) the CL method takes into account the cumulative amount of claims. The data set used in the thesis includes individual claims entries, therefore the results comparison was difficult as the ML predictions are based on individual entries and the CL method considers cumulative claims combined. Therefore, another limitation is the available data does not include all the above-mentioned elements.

In order to be able to provide the business and underwriting department most accurate estimation and calculation techniques, further data will need to be collected to support the findings, it will be unfortunately then out of the scope of this thesis. The author was also unfamiliar with insurance actuarial calculations, which is also a limitation as there is

possibilities not all relevant factors were taken into consideration. Before the results can be applied in real world setting the calculations need to be verified by the commissioning organization's professionals.

1.7 Ethical considerations

The research has been done in a manner that follows the responsible conduct of research guidelines (Hyvä tieteellinen käytäntö, 2012). Research integrity has been adopted while doing the research and writing the thesis. This research follows the principles endorsed by the research community, which are:

- accuracy in conducting research and recording, presenting and evaluation the research results
- The data acquisition methods follow scientific criteria and they are ethically sustainable
- The research results are communicated openly and responsibly
- The work of other researchers has been acknowledged and cited in appropriate manner
- It has been ensured before the research with the employer on the researcher's rights, responsibilities and obligations, questions regarding data access and archiving
- Sources of finance and conflicts of interest are reported if there are any
- The research organisation follows good personnel and financial administration practices and takes into account the data protection legislation

(Hyvä tieteellinen käytäntö, 2012).

In general, big data analytics has raised a lot of discussion on the ethical aspects of data-analysis. The developers and data engineers can reveal from data patterns and new knowledge that was not been able to be accessed couple of years ago. The legal and ethical guidelines have not yet been able to adapt to this scale of new usage of data (Uria-Recio,

2018). The experts in the ethical issues around big data, agree on the following principles that should be considered;

- private customer data and identity should remain private
- shared private information should be treated confidentially (there should be restrictions if and how the information can be shared further)
- customers should have transparent view how their data is handled and processed in third-party analytical systems
- big data can determine and make decision based on data for human beings, it should not affect human will
- machine learning should not absorb unconscious biases based on race or gender via training samples

(Uria-Recio, 2018)

The ethical research guidelines have been followed from the beginning of the project. Since the data collection process, it has been ensured that all the data follows GDPR regulations (Yleinen tietosuoja-asetus, 2021). The dataset has been anonymized ensuring no sensitive information are shared from any individuals and no one can be recognized from the analysis. The dataset contains claims details from individuals; therefore, it has been needed to process anonymously. Any identification details of any individuals have been removed before processing the data and running any algorithms. Maintaining a good quality research has been ensured by storing the data on commissioning organization secured computer. In addition, the analysis and running the algorithms have been conducted on company secured servers and laptop in order to secure data security. The secrecy and anonymity of the company and the third party has needed to be ensured, by masking part of the data and results in this research with dummy variables. The consent for the thesis work has been given by the commissioning organization in order to find solution to the business problem. The consequences of the results in this research have been taken into consideration and the study has been written in a manner that will not harm any individuals or any organization.

2 LITERATURE REVIEW

2.1 Introduction

There is vast amount of research, literature and articles done regarding prediction of outstanding loss reserves in insurance industry. By searching from Google Scholar with the search words “outstanding loss reserves” 252 000 results can be found (20.9.2020). Narrowed down by concerning the insurance industry only, the same search brings 104 000 results. The model used to predict the outstanding loss reserves is commonly calculated with the chain-ladder method. References to this can be found from every introduction chapter of any paper one might read, for example from one of the latest papers by Kuang & Nielsen (2020) in Scandinavian Actuarial Journal. As mentioned, there are several researches in the field regarding this method, and even more about using alternative predictive models aiming to predict the problem more accurately and comparing them to the traditional methods.

2.2 Related work on the field

There exists several actuaries’ studies, where researchers have used traditional mathematical methods to calculate the outstanding loss reserves. In addition, there are more recent studies done in the field by using machine learning (ML) methods in predicting individual claims reserving, as by Wüthrich in Scandinavian Actuarial Journal (2017). Merz & Wüthrich (2008) wrote a book concentrating on exclusively in several different stochastic models used in claims reserving. Stochastic modelling is a form of financial modelling used to make investment decisions, it forecasts the probability of various outcomes in different circumstances and it uses random variables. The method is used in several industries in order to improve business practices and profitability of their portfolios. The insurance industry relies on this modelling method to predict how the company balance sheet will look like in a certain point in the future (Kenton, 2020). Merz & Wüthrich (2008) explain in detail the traditional models and the mathematical measures used in

them. In the beginning of their book they explain the basic methods and definitions used in claims reserving. One of the most cited ones is the study done by Mack who wrote in Insurance: Mathematics and Economics Journal in 1994 article about "*Which Stochastic Model is Underlying the Chain Ladder Method?*" (Mack, 1994). In addition, they talk about the literature in the field and mention that choosing the model for specific dataset is one of the most difficult questions to answer and there is limited amount of literature regarding this topic available. Due to this they use different methods mechanically always on the same dataset (Merz & Wüthrich, 2008). In addition to the chain-ladder models they write about Bayesian models, distributional models like log-normal model, gamma model and Poisson model, general linear models and bootstrap methods (Merz & Wüthrich, 2008).

De Alba (2002) argues in their research paper there is a demand for improved ways to estimate the loss reserves and find measures of their variability and information on their future behaviour. He introduces Bayesian forecasting methods. Bayesian forecasting methods are used to predict data points from historical data in order to understand future behaviour. According to Wikipedia these methods are used for event base data and it means interpretation of probabilities, that the behaviour is something that happens randomly in different points in time (Wikipedia, 2020).

"*Bayes' theorem describes the conditional probability of an event based on data as well as prior information or beliefs about the event or conditions related to the event. For example, in Bayesian inference, Bayes' theorem can be used to estimate the parameters of a probability distribution or statistical model. Since Bayesian statistics treats probability as a degree of belief, Bayes' theorem can directly assign a probability distribution that quantifies the belief to the parameter or set of parameters."* (Wikipedia, 2020).

De Alba (2002) presents some previous researches done in the field using a full Bayesian Model, and introduces two models, one to forecast the number of outstanding claims and one for total aggregate claims. He explains the chain-ladder method was used in the previous works by other researchers as a reference to benchmark other models due to it being easy to apply and generalize (De Alba, 2002).

The latest study in Scandinavian actuarial journal by Kuang & Nielsen (2020), presents new methods for distribution forecasting of general insurance reserves by using a log-normal model. Their study provides alternative method to the traditional chain-ladder. According to the authors, this study and their results make the actuaries choose if the traditional, log-normal chain -ladder or a third method should be used in a reserving triangle. As their third method, they used bootstrapping which did not give too accurate results. However, they stated it has become more popular method to use in recent years. The definition of bootstrapping is;

"*Bootstrapping is a statistical procedure that resamples a single dataset to create many simulated samples. This process allows for the calculation of standard errors, confidence intervals, and hypothesis testing*" (Forst, cited in Joseph, 2020).

Kuang & Nielsen (2020) discuss also another method, an asymptotic theory which is an analytical tool for evaluating forecasting errors and building inferential measures and specification test for the model. They adapted the infinitely divisible framework to the log-normal chain-ladder and present an asymptotic theory for the distribution forecasts and model evaluation. These should cover a wide range of reserving triangles (Kuang & Nielsen, 2020). They also mentioned in the conclusion part of their article about the seasonality of the occurred claims.

"*In particular, if seasonal patterns are common from year to year or across triangles there may be scope for improving the performance of the asymptotic theory*." (Kuang & Nielsen, 2020).

They argue the generalized log-normal model distribution can improve the actuarial process for a corporation and that it is used to simulate attritional reserve. This is then combined with the bootstrap method for the traditional chain-ladder. Apparently, it is possible it can cause inconsistencies between reserving and capital modelling risk (Kuang & Nielsen, 2020). Limitation to the method is the log-normal model fits positive incremental values, however in reality values can also be negative due to reinsurance recoveries, salvage or data issues for example, misallocation of monies between different business classes. In such event, they recommend further research should look at how to provide other statistical tools to tackle this kind of limitation (Kuang & Nielsen, 2020).

Harej, et al. 2017 argue that modern machine learning techniques can offer more accurate estimates of provisions. Neural Networks and the methods capabilities of pattern recognition could provide new insight to claims reserving and pricing on the individual claim's prediction. They found out the chain-ladder methods estimated the reserves accurately however, it performed badly with individual claim predictions and state that the Artificial Neural Networks (ANNs) made better predictions regarding the paid claims. (Harej, et al. 2017). They found out in their study the ANN methods and model should be used on real data instead of synthetic data. For further research, they suggest using support vector machines or random forest (Harej et al., 2017).

Kuo (2019) introduced in his work the DeepTriangle which is a deep learning framework for forecasting paid losses. Their attempt was based on the available data to predict future development of the accident year's paid losses and claims outstanding based on the observed history. They found out their DeepTriangle was able to improve the chain-ladder, common machine learning models and Bayesian stochastic models. They conclude their model can match with modern stochastic reserving techniques without expert knowledge. The model used neural networks which allows the usage of multiple heterogenous inputs and train multiple object simultaneously, it also allows customization based on the available data and variables. Neural networks can help extend machine learning methods to incorporate additional features those are not able to handle (Kuo, 2019).

2.3 Chapter summary

The literature review presented selection of the recent studies in the field that compare the traditional methods vs. ML algorithms around the claims reserving problem. The recent literature around the topic in actuarial science is investigating the usage of decision tree and random forests and artificial neural networks (Wüthrich, 2018, and Harej, et al. 2017). As stated in the beginning of the literature review, it is not very clear which model should be used and which one is the most appropriate method to forecast the outstanding reserves or ultimate loss reserves, and most of the research uses more than one method.

The complexity of the task was noticed during the data analysis process, therefore more than one ML method was investigated and compared.

The claims reserving problem in this research was approached by using the traditional statistical method theory chain-ladder as commonly used in the industry and the simple ML algorithms appropriate for the data and the problem. As suggested by the previous research linear regression, random forest and logistic regression were chosen. Afterwards, it will be evaluated the accuracy of each methods and their advantages and disadvantages.

3 RESEARCH METHODOLOGY

In this chapter the chosen research method will be reviewed. The structure of the thesis is a development thesis, the purpose is to build accurate and automated loss reserving method for the commissioning organisation. The chosen research methods, the chosen ML algorithms, and their basic theories, terminology, definition and calculation methods will be presented. Linear ridge regression, logistic regression and random forest were decided to be used as the ML methods and compared with the chain-ladder method.

3.1 Methodology

3.1.1 Quantitative research methodology

The methodology used in the thesis are quantitative research methods.

Quantitative research is defined as research strategy that highlights quantification of the collection and analysis of data (Bryman, 2015).

Mathematical models were used as the methodology of data-analysis, the method involves usually collection of numerical data. It typically includes research design, test and measurement procedures and statistical analysis (Williams, 2007). More closely the area of quantitative research in this thesis is descriptive research which means identifying attributes of a specific phenomenon on observational basis or exploring correlation between two or more phenomena (Williams, 2007). There are many research methods to conduct descriptive quantitative research for example, descriptive, correlational, developmental design, observational studies and survey research (Williams, 2007). For this thesis, the method used is mainly correlational research method, which studies the differences between two characteristics of the study group. According to Creswell (2002, cited in Williams, 2007) the purpose of the correlation is a statistical test to establish patterns between

two variables. However, Bold (2001, cited in Williams, 2007) notified the aim of correlational study is to find out if two or more variables are related. Since the nature of the variables in the dataset, part of the research is a multiclass classification problem, therefore it is assumed that the approach is mixture of correlational and development design. In the development design it is explored how characteristics vary over time in the study group, and the researcher investigates in the cross-sectional study two different groups with similar constraints (Williams, 2007).

In this thesis the archival data collection method is used. It means using already gathered data about the variables in the correlational research. Usually this type of approach uses previous studies or historical records as a basis of the variables that are analysed. The benefits of this approach are less expensive, saves time since the data already exists and provides the researcher more previously available data. The disadvantages are data accuracy since the researcher does not have control over the data collection process (Formplus, 2020).

3.1.2 Research design

The data was collected by external party who is responsible of maintaining the data for the company. The dataset contains claims from accident years 2016 until end of November 2020.The dataset is stored on the commissioning organizations computer and secured servers.

The data analysis and the machine learning algorithms are done with Python's Jupyter notebook and Microsoft Excel. The chain-ladder method, which is a statistical mathematics model, was calculated with Excel. There is also available chain-ladder library in Python to calculate chain-ladder, which was tested as well, however the findings are not included in this research.

For the data-analysis and building the chosen predictive models, the Python's available open source libraries were used instead of creating own libraries. The Python libraries used in the thesis and for the data-analysis are listed in table 1.

Table 1. Open source libraries used in data analysis

Name	**Purpose**	**Version**
Pandas	Data analysis	0.25.1
Numpy	Mathematical computing	1.16.5
Matplotlib	Plot graphs and data visualization	3.1.3
Seaborn	Plot graphs and data visualization	0.9.0
Scikit-learn	Machine learning library	0.2.41

The linear ridge regression and traditional chain-ladder method are used to predict the price of the claim. Linear ridge regression accuracy is measured with scikit learn built-in metrics and cross validation score. The chain-ladder method is compared with the results received from the linear ridge regression model. Both predictions are measured against the company's own results and predictions. In the KPM white paper by Golfin & Kuo (2016), they have handled the loss reserving also as a regression problem. They used as the predictor's accident year, development period and incremental loss (Golfin & Kuo, 2016).

Random forest classifier and logistics regression are trained to predict when the claim will be paid. The performance and accuracy are measured with the scikit learn libraries' own methods. The *receiver operating curves (ROC curve)* is a common tool to analyse the behaviour of classifiers at different thresholds and show the false positive rate against the true positive rate. The *area under the ROC curve* is referred to as AUC which is used to summarize the ROC curve, and it is a better metric to use with analysis of imbalanced dataset than accuracy (Guido & Müller, 2016), the methods were analysed using the AUC.

However, as mentioned already in the thesis before, the correct amount of outstanding loss reserve and the amount of ultimate claims changes constantly over time due to the nature of insurance business. The reserving is not done in certain point in time, but ongoing basis, therefore it will be difficult to know the exact figure to make the comparisons. A decision was made, the data collected until November 2020 used in this thesis is compared to data collected from March 2021 to see how accurately the outstanding claims were predicted and if any changes had occurred in the research period and during the

writing of the thesis. The dataset is analysed first using descriptive data analysis, the descriptive analysis defines the most important characteristics and basic information of the dataset (Singh, 2018). The descriptive part is followed by the feature engineering and then training of the predictive models.

3.2 Methods

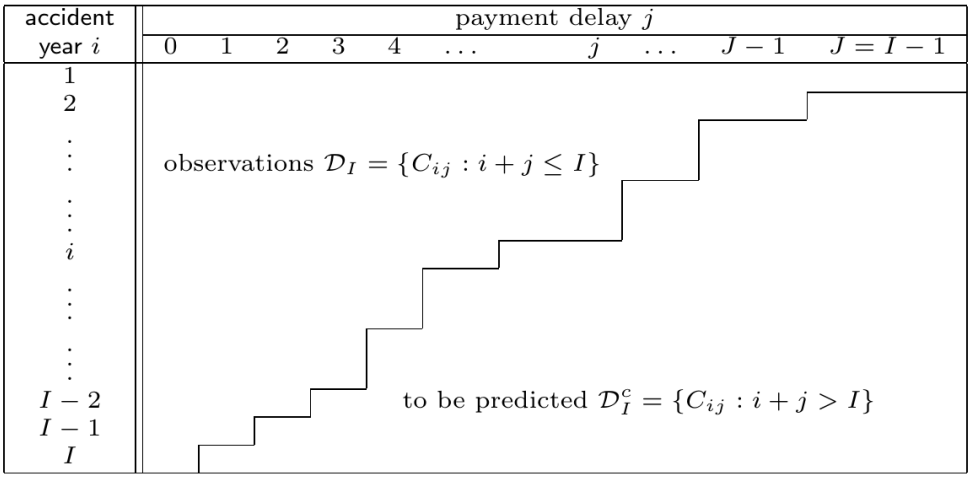
3.2.1 The Chain-Ladder method

Insurance companies need to set aside a portion of premiums they receive from the active policies within the policy period to pay for claims occurred by accidents that may be submitted in the future by the insured person or organization. The common problem is the accuracy of the claims predictions and reserves due to the time between the accident and the final claims payment. The estimation of claims during the financial year with the amount of actual paid claims regulate the amount of profit the insurer will publish in the financial documents (Kagan, 2020a).

The chain-ladder method is the most used and the easiest claims reserving method due to its simplicity and accuracy. The actuarial literature defines the chain-ladder method as computational algorithm for estimating claims reserves (Wüthrich and Merz, 2008). It means chaining a sequence of year-to-year development factors into a ladder, meaning undeveloped losses climb towards maturity which are multiplied by the chaining of ratios (Amin.et al., 2020). A run-off triangle is presented in figure 1.

Wüthrich and Merz (2008) define the equation for calculating stochastic chain-ladder as it is assumed the last development period is given by J, hence $X_{i,j} = 0$ for $j > J$, and the last accident year is given by I, they refer to their general assumption that $I=J$.

Figure 1. A run-off triangle



Source: Wüthrich and Merz, 2008 (cited in Amin, Z., et al. 2020).

These run-off triangles are two-dimensional matrices that are generated by accumulating claim data over a period (Kagan, 2020a).

Amin, et al. (2020) present in their open source book the Taylor (1986, cited in Amin, et al., 2020) deterministic chain-ladder run-off triangles in cumulative form. The triangle presents the cumulative amount paid until development period *j* for claims that occurred in year *i*. The chain-ladder method needs development factors *fj*. The development factor defines the growth of cumulative amount *j* to cumulative amounts in year *j+1*. The formula to calculate the development factors is:

$$C_{i,j+1} = fj \times C_{i,j}.$$

The ratio of the cumulative amounts for the upcoming years can be estimated from the first development factor f_0 that describes the cumulative claim amount from development period 0 to 1. The chain-ladder method calculates with the development factor estimator the cumulative amounts for all the years, this prediction is calculated by multiplying the most recent cumulative claim with the development factor for the occurrence period. After the development factor for period 0 to 1 is calculated, the similar method is used to calculate the development methods from period 1 and 2. It measures how the cumulative paid amount grows from one period to the other, then it is averaged across all occurrence periods which they are being observed. The second development factor is used to predict the missing information in development period 2 (Amin, et al., 2020).

Figure 2. Chain-Ladder run-off triangle with cumulative payment data with predictions

accident year	payment delay (in years)									
	0	1	2	3	4	5	6	7	8	9
1	5,947	9,668	10,564	10,772	10,978	11,041	11,106	11,121	11,132	11,148
2	6,347	9,593	10,316	10,468	10,536	10,573	10,625	10,637	10,648	
3	6,269	9,245	10,092	10,355	10,508	10,573	10,627	10,636	*10,647*	
4	5,863	8,546	9,269	9,459	9,592	9,681	9,724	*9,735*	*9,745*	
5	5,779	8,524	9,178	9,451	9,682	9,787	*9,837*	*9,848*	*9,858*	
6	6,185	9,013	9,586	9,831	9,936	*10,005*	*10,057*	*10,067*	*10,078*	
7	5,600	8,493	9,057	9,282	*9,420*	*9,485*	*9,534*	*9,545*	*9,555*	
8	5,288	7,728	8,256	*8,445*	*8,570*	*8,630*	*8,675*	*8,684*	*8,693*	
9	5,291	7,649	*8,243*	*8,432*	*8,557*	*8,617*	*8,661*	*8,671*	*8,680*	
10	5,676	*8,471*	*9,130*	*9,339*	*9,477*	*9,543*	*9,592*	*9,603*	*9,613*	
$\hat{f}^{CL}$	1.493	1.078	1.023	1.015	1.007	1.005	1.001	1.001		

Source: Wüthrich and Merz (2008), *Table 2.2.* (cited in Amin, et al., 2020).

Figure 2 shows the completed run-off triangle using the chain-ladder method. The numbers in the last column are the estimation of the ultimate claims amounts for each accident year. The estimate is calculated by the difference between the ultimate claim amount and the cumulative amount (Amin, et al., 2020). The chain-ladder estimates for the claims reserves that are needed for the future commitments to be fulfilled during the occurrence period is presented in figure 3 by Wüthrich and Merz (2008, cited in Amin, et al., 2020):

Figure 3: Reserves per occurence period and for total Claims

	$C_{i,I-i}$	Dev.To.Date	$\hat{C}_{iJ}^{CL}$	$\hat{\mathcal{R}}_i^{CL}$
1	11,148,123	1.000	11,148,123	0
2	10,648,192	0.999	10,663,317	15,125
3	10,635,750	0.998	10,662,007	26,257
4	9,724,069	0.996	9,758,607	34,538
5	9,786,915	0.991	9,872,216	85,301
6	9,935,752	0.984	10,092,245	156,493
7	9,282,022	0.970	9,568,142	286,120
8	8,256,212	0.948	8,705,378	449,166
9	7,648,729	0.880	8,691,971	1,043,242
10	5,675,568	0.590	9,626,383	3,950,815
totals	92,741,332.00	0.94	98,788,390.50	6,047,058.50

Source: Wüthrich and Merz (2008, cited in Amin, Z.et al., 2020)

3.2.2 Linear regression and Ridge regression

Linear regression is a linear model, which adopts a linear relationship between input and output variables (Brownlee, 2016). It is one of the basic and most easily understood supervised regression machine learning algorithms (Guido & Müller, 2016). In the case of multiple input variables, the method can be also called as multiple linear regression

(Brownlee, 2016). For this thesis, the used method is multilinear regression due to multiple input variables.

Linear regression makes better predictions if the relationship between the input and output variables is linear and if the data has Gaussian distribution (Brownlee, 2016). Therefore, the linear method chosen in the thesis to predict the value of paid claims is ridge regression, which is another linear model. Ridge regression considers in addition to fitting well on training data, to fit additional constraints, which means regularization that should avoid overfitting the data (Guido & Müller, 2016). Since the dataset has few datapoints in some of the input variables, it was noticed in the data-analysis and model building phase, ridge regression is a better approach since it is more restricted than the simple linear regression (Guido & Müller, 2016).

3.2.3 Logistic regression

Logistic regression is one of the most used classifications algorithms, it can also be used for multiclass classification problems (Guido & Müller, 2016). Therefore, it was the chosen method for this thesis to predict the year (the development month) the claim is going to be paid.

Logistic regression is a linear classifier that uses the following function:

$$f(\mathbf{x}) = b_0 + b_1 x_1 + \cdots + b_r x_r.$$

The variables b_0, b_1, …, b_r are the estimators of the regression coefficients, which are called as the predicted weights or coefficients (Stojiljković, 2021).

Multiclass logistic regression makes results in one coefficient vector and capture it per class, it uses the same method when making the predictions on the test set (Guido & Müller, 2016). The regularization parameter C is the main regularization parameter of logistic regression model, small values of C mean simple models and it is important to tune these parameters. In addition, it is needed to be decided if L1 regularization or L2 regularization will be used, if not specified separately the model uses as default the L2

(Guido & Müller, 2016). In this thesis, the default parameter was used since it was not wanted to rule out any of the features, instead it was left for the model to decide them.

The classification performance of logistic regression is evaluated with the results of true negatives, true positives, false negatives and false positives. The evaluation happens by comparing the actual and predicted outputs by counting the correct and incorrect predictions (Stojiljković, 2021). In addition, the model will be improved with using grid search and evaluated with the AUC score and compared with the random forest classifier since accuracy cannot be relied on due to the imbalanced dataset issue.

3.2.4 Random forest classifier

The random forest classifier was chosen to be trained as was recommended in the literature review by (Harej et al. 2017).

Random forest classifier is one of the most used machine learning methods as they are powerful, they do not require intense parameter tuning and the method does not require scaling of data. The advantage of random forest is to avoid overfitting which is the downside of decision trees. When starting to build a random forest, first it needs to be decided the number of trees wanted to be built by setting *n_samples*. The function randomly creates equally sized datasets as the original one (Guido & Müller, 2016). Random forest is a collection of decision trees, where the idea is to build many decision trees based on subsamples of the dataset with slightly different selection of features, then by averaging their results overfitting can be avoided and the predictive accuracy improved. When then random forest is built, during the process each node is split, and the best split is found from the input features or from the random subset (Pedregosa et al., 2011). This number of features is controlled by *max_features* parameter. These two methods make sure the trees in the random forest are different. The algorithm then makes a prediction for each tree in the forest, the predictions of all the trees are averaged, and from that the class with the highest probability is predicted (Guido & Müller, 2016). Averaging the predictions cancels out errors (Pedregosa et al., 2011). After training the random forest it was improved with grid search and measured with AUC score.

3.3 Setting

The implementation and training process to apply the chosen machine learning methods followed the same steps.

First the needed python libraries and functions were imported, the dataset was uploaded and the needed feature engineering steps and transforming of the data into correct format was done with NumPy or Pandas, training of the classification or regression model with using the modified dataset, model evaluation and cross validation or grid search to observe the performance, and application of the model to make predictions. Finally, the results of the ML method were compared with the outcome of the actual claims data as of March 2021.

The chain-ladder and linear ridge regression were used to predict the loss reserves. The chain-ladder method and the linear ridge regression method were compared with each other and with the actual results to compare the accuracy.

The logistic regression and random forest were compared to see which can better predict the development year (development month) when the claim will be fully paid. As accuracy is not a good evaluation technique with imbalanced dataset, the AUC scores were compared to select the best model.

3.4 The dataset

The dataset 1 size is 581kb and it has 8351 individual claims from accident years 2016-2019. The dataset 2 includes all the claims data since 2015 until November 2020. The two datasets contain mainly the same variables, however there are some differences for example, the dataset 1 contains information of the insurance product codes and the gender of the claimants which is missing from the dataset 2. Both datasets have the figures needed for make the predictions of the outstanding loss reserves.

Table 2. Variables of the datasets

Dataset 1
Policy number
Claim number
Insured
Date of loss
Date of notice to TPA
Cause of loss
Body part
Period
Paid & Not Paid
OSLR€
Status
Combined Paid & Not paid & OSLRE

Dataset 2
Claim number
Insured gender
Policy period
Limits
Reserves
Paid
Loss type
Reported date
Loss date
Denial
Date of decision or denial
Payment date
Manual reserves
Status
Injury type

The data cleansing and modification is the biggest part of any analytics project. It was needed to revisit the data several times and review the needed parameters to be used in order to have the correct predictors for the ML models.

3.4.1 Feature engineering

The two datasets were combined and cleaned up first using Excel's V-LOOKUP function and other formulas. The variable names were translated from Finnish into English and some new variables were created in order to perform the predictions. The clean-up and modification of data was started from the Injury type variable, which was translated, and the title changed to Injured_body_part, for python script to read the feature names better. The Cause of loss and Injury type columns have been written by different claim handlers by several different ways using different wordings, therefore these two variables were

cleaned and unified in Excel to reduce the number of classes. Obstacle was that the VLOOKUP function did not find many variables since the words included letters that are used instead of the Finnish alphabets *ä* or *ö* by a system that does not recognise the Finnish alphabets. In addition, the wordings had extra blank space after the end of the word, which was replaced by special symbols, which needed to be deleted. These were changed by using Excel's find and replace function. Some of the clean-up had to be done manually as well. In table 3 below are the final modified columns in the dataset and their datatypes.

Table 3. Modified data frame variables and data types

<class 'pandas.core.frame.DataFrame'>		
RangeIndex:	8353 entries 0 to 8352	
Data columns	20	
dtypes:	datetime64[ns](4), int64 (4), object (8), float64 (4)	
memory usage:	1.3+ MB	
Columns	**Entries**	**Data type**
Claim_number	8353	object
Insured_gender	8353	int64
Policy_period	8353	int64
Limits	8339	object
Reserves	8353	int64
Paid	8353	float64
Loss_type	8353	object
Reported_date	8351	datetime64[ns]
Loss_date	8351	datetime64[ns]
Denial	8353	object
Date_of_decision_or_denial	7059	datetime64[ns]
Paid_date	6659	datetime64[ns]
Manual_reserve	8353	int64
Status	8351	object
Injured_body_part_2	8351	object
Injured_body_part	8353	object
Paid_Not_Paid	7786	float64
OSLR	7156	float64
Status_2	7156	object
Comb_Paid_Not paid_OSLRE	7156	float64

The dataset has 20 columns and 8353 rows. The datatypes of the columns are in datetime, floats, integers and objects. It was analysed how many null values the data entries contain. It was noted many of those variables that were considered to be the most important variables for making prediction included blank values as illustrated in table 4.

Table 4. Analysis of null values in dataset

Columns	**Count of null values**
Claim_number	0
Insured_gender	0
Policy_period	0
Limits	14
Reserves	0
Paid	0
Loss_type	0
Reported_date	2
Loss_date	2
Denial	0
Date_of_decision_or_denial	1294
Paid_date	1694
Manual_reserve	0
Status	2
Injured_body_part_2	2
Injured_body_part	0
Paid_Not_Paid	567
OSLR	1197
Status_2	1197
Comb_Paid_Not paid_OSLRE	1197

Therefore, it was needed to go back to the company who produces and records the data to request clarifications on the blank cells especially for the Paid_Not_Paid, OSLR, Comb_Paid_Not paid_OSLRE variables since the volume of null values were extreme. It was important information to be able to handle the data in an appropriate manner and have the valid prediction outcomes. After the answers were received from the company with the explanation of the null values, they were filled in according to the instructions with value 0.

The data was uploaded into Jupyter notebook again and any remaining null values were deleted from the dataset with a code function. The count of rows dropped from 8352 rows

to 5905, which was 30% of the data. As the dataset was originally including many blank cells and unpopulated fields, it gave the indication of not very good quality dataset.

The columns containing object variables and the variables not needed in the model were removed. The datetime values, financial and numerical values were transformed into integers in order to be able to work with the model. The column names were also handled and changed into readable format in Python. The rest of the data clean-up and modification was done in Python Jupyter Notebooks, in order to prepare the dataset in correct format and have all the predictors available for building the models.

In chain-ladder theory the accident year is needed in order to formulate the triangle. Therefore, from the variable Loss_date was created new feature Accident_year containing only the year data instead of dates. In order to predict the development year when the payment was made, a new dummy variable Paid_y_n was created. The name of the column Paid which includes the paid claims amounts was changed to Paid_eur to describe the column information better. Other new variables created from the original datasets were payment_delay (calculated between paid_date and loss_date), settlement_delay and reporting_delay. The final dataset and the datatypes and size of the dataset is illustrated in table 5.

Table 5. Final dataset variables and data types

<class 'pandas.core.frame.DataFrame'>		
RangeIndex:	0 to 5904	
Data columns	12	
dtypes:	int64(12)	
memory usage:	553.7 KB	
Columns	**Entries**	**Data type**
Insured_gender	5905	int64
Policy_period	5905	int64
Reserves	5905	int64
Paid_euros	5905	int64
Paid_1_0	5905	int64
accident_year	5905	int64
Manual_reserve	5905	int64
Paid_Not_Paid	5905	int64
OSLR	5905	int64
reporting_delay_days	5905	int64
settlement_delay_days	5905	int64
payment_delay	5905	int64

3.5 Descriptive data analysis

The descriptive analytics and statistics of the final dataset was analysed first. The table 6 shows the basic statistics from python describe-function.

In the final dataset, there are 5905 rows of each 12 variables. The main variables that are Paid_euros, accident_year, OSLR, reporting_delay_days, settlement_delay_days and payment_delay and their statistics. The actual figures have been masked with dummy figures.

From the table 6 it can be seen the Paid_euros and Paid_not_paid variables are identical. The mean of the paid claim is around 60 euros, standard deviation 311 euros, minimum value 0 euros and maximum 667,2 euros. The lowest 25% of the Paid_euros are 13,2

euros, 50% are 98,2 euros and 75% 370,2 euros, which means that majority of the individual values of a claim are within this price range.

The OSLR value counts for 0 was 5888 and only 17 datapoints holding some financial values remained in the dataset. The OSLR variable on its own did not contain enough data in order to make valid predictions based on solely on this variable. Due to the low number of datapoints the variable was not included in the training dataset. Based on the methodology of how the chain-ladder and the claims prediction is calculated, it was possible to use the other variables of the dataset such as the accident year and the claims euro amount. Same was noticed from the Reserves variable, 5879 datapoints have the value 0 in the dataset, 26 datapoints have some claims reserves marked in the dataset. This variable was decided to be dropped as well since it will not bring additional value to the predictions. The decision of leaving some of the variables such as the OSLR and Reserves out from the final training dataset, was based on earlier attempts to include them in the model. Based on those findings that they do not have predictive value, they were dropped.

The variables reporting_delay_days and settlement_delay_days were left as date-time values, but the payment_delay was changed into months, since this is the commonly used way when calculating OSLR in insurance industry. The reporting_delay_days and settlement_delay_days were decided to leave as they are for the training dataset.

The variables needed to predict the outstanding claims are the payment delay and the amount of paid claims. The rest of the descriptive analysis is focused on analysing and bringing understanding of these variables.

Table 6. Description and basic statistics of the dataset 2016-2019

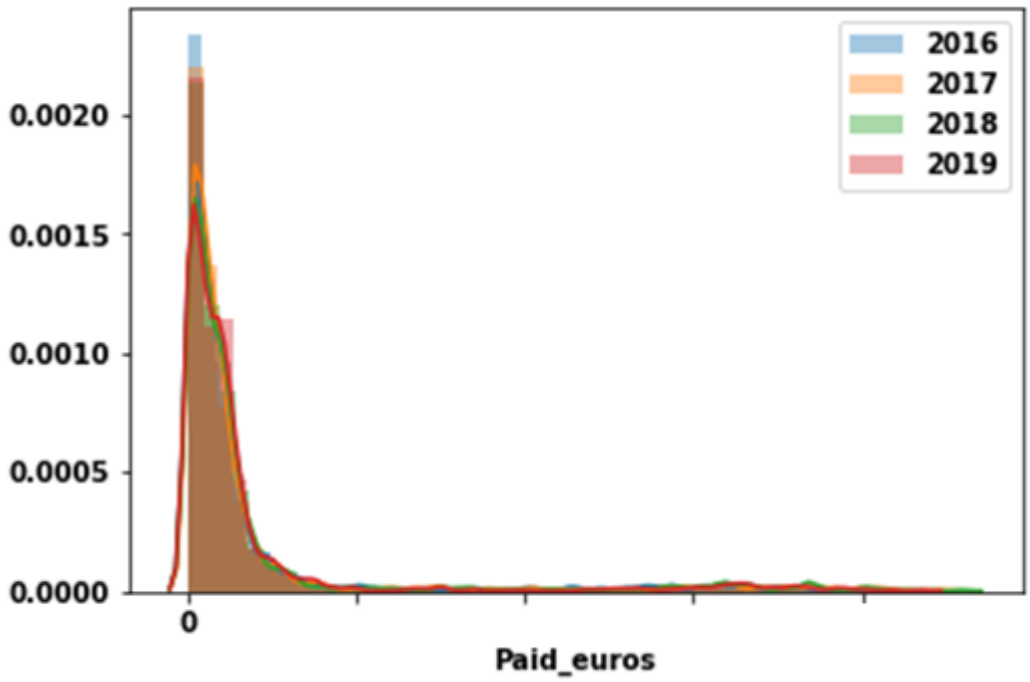
Describe	count	mean	std	min	25%	50%	75%	max
Insured_ gender	5905	0,7	0,5	0,0	0,0	1,0	1,0	1,0
Policy_period	5905	2017,5	1,1	2016	2017	2018	2019	2019
Reserves	5905	0,7	33,3	0,0	0,0	0,0	0,0	435,4
Paid_euros	5905	60,0	311,0	0,0	13,2	98,2	370,2	667,2
Paid_1_0	5905	0,9	0,3	0,0	1,0	1,0	1,0	1,0
accident_year	5905	2017,5	1,1	2016	2017	2018	2019	2019
Manual_reserve	5905	0,7	33,3	0,0	0,0	0,0	0,0	435,4
Paid_Not_Paid	5905	60,0	311,0	0,0	13,2	98,2	370,2	667,2
OSLR	5905	0,5	30,7	0,0	0,0	0,0	0,0	435,4
reporting_delay _days	5905	15,6	26,4	0,0	2,0	7,0	17,0	292,0
settlement_ delay_days	5905	81,9	113,3	0,1	16,0	40,0	92,0	1143,0
payment_ delay	5905	12,4	2,3	12,0	12,0	12,0	12,0	48,0

The payment_delay mean is at 12.4 months, from the table it can be seen majority of claims are paid within 12 months. By analysing the value counts of the payment_delay variable, the distribution of the development months was 12: 5717, 24:178, 36: 9, 48:1. As this is the main variable, it shows the dataset is imbalanced.

The reporting delay is on average between 7 to 17 days, which means the time between the accident occurred and the insured notified the company about the accident (loss date). Settlement delay means the days between the decision the claim was accepted or declined and the loss date.

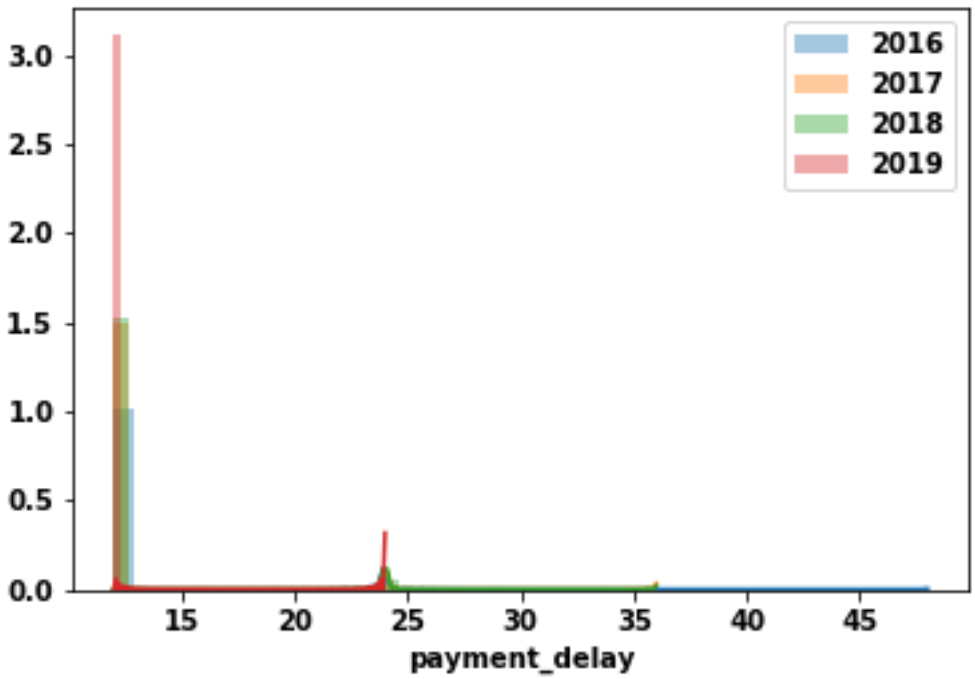
The distribution of the variables needed in the predictions between 2016-2019 were investigated in more detail with visualisations.

Figure 4. Distribution of Paid euros 2016-2019



The Paid_euros in figure 4 follow the standard distribution, however the peak of the curve if heavily shifted to the left and towards 0 euros. Here it can be seen very few claims are more than 370 euros in each accident year which was the sum 75% of the pair euros. In each accident year the amount of paid claims follow the same trend.

Figure 5. Payment delay distribution in months 2016-2019



The figure 5 illustrates the payment delay distribution, the highest peaks in the data for each accident year can be seen at the month 12. It means most of the claims are paid within 12 months from the reported date. The red trendline shows the next peak on the 24th month mark. The value counts analysis in Jupyter provides the following counts of these classes showed in table 7.

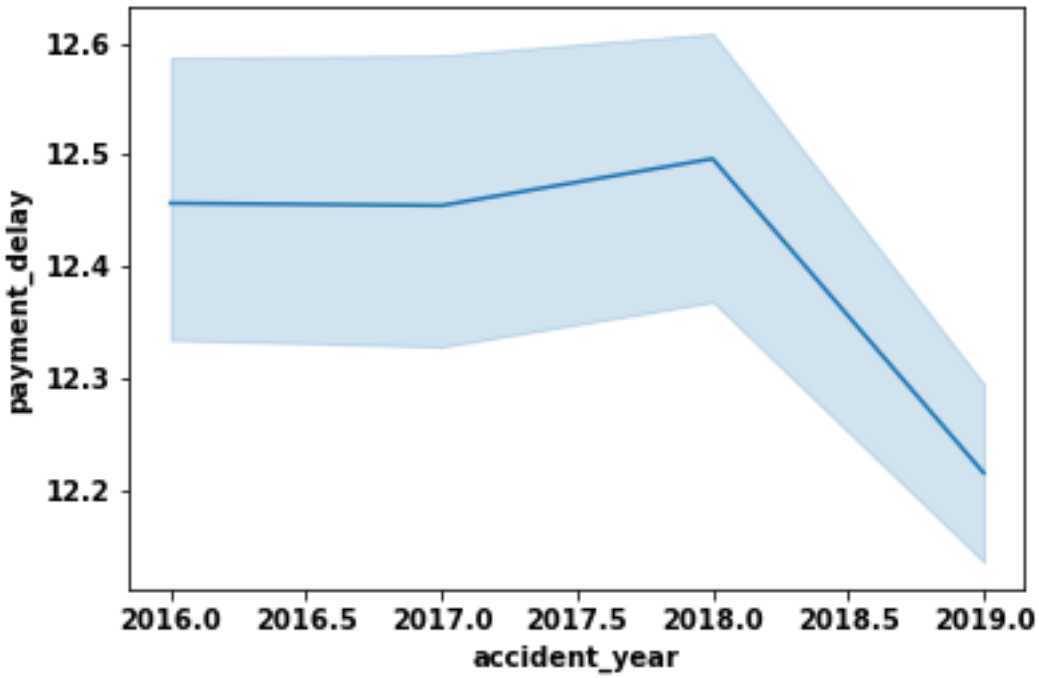
Table 7. Value counts of payment_delay per accident year.

payment_delay	**12**	**24**	**36**	**48**
2016	1419	51	1	1
2017	1377	44	55	0
2018	1440	56	3	0
2019	1481	27	0	0

The analysis of these variables shows again the dataset is highly imbalanced which needs to be taken into account with the predictions. The prediction of payment delay will be a multiclass classification task and the prediction of the claims amount will be a regression task.

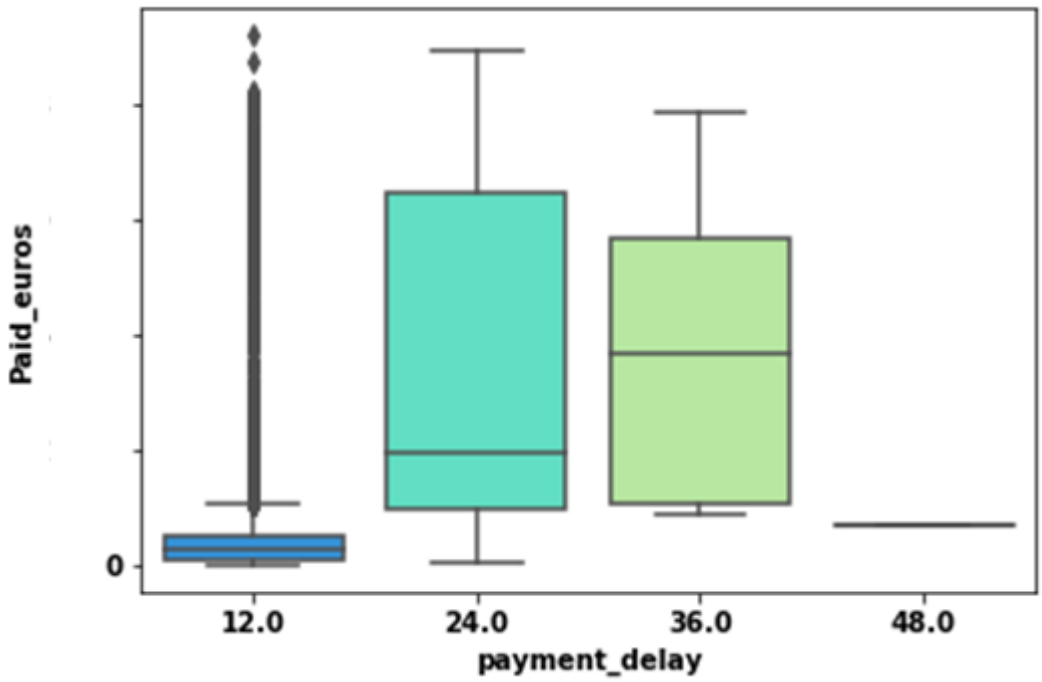
Further visualizations of the needed variables in the predictive analytics were done in order to gain better understanding of the relationship of the variables. The payment delay, paid euros and how they are distributed between the accident years were looked into. The dataset was decided to be divided into four subsets based on the accident year.

Figure 6. Payment delay by accident year



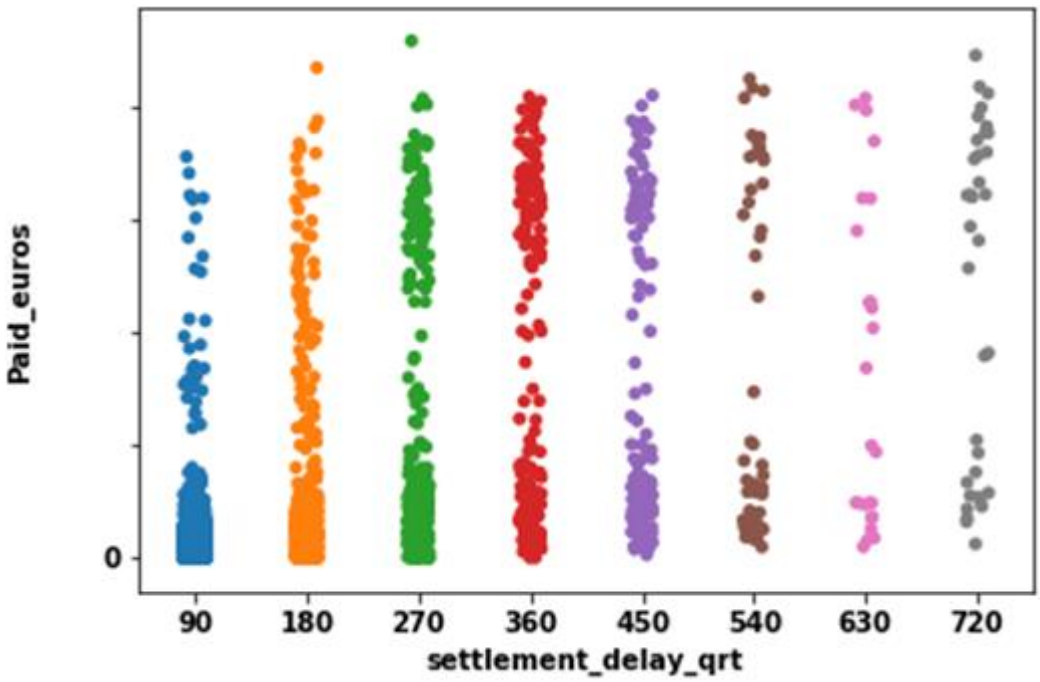
The figure 6 shows downward trend towards 2019 as longer payment delays than 24 months have not yet been reported. It shows the company can close and finalize the open claims incident is within or soon after 12 months, which is good.

Figure 7. Boxplot to visualize the value of paid claims and the payment delay 2016-2019



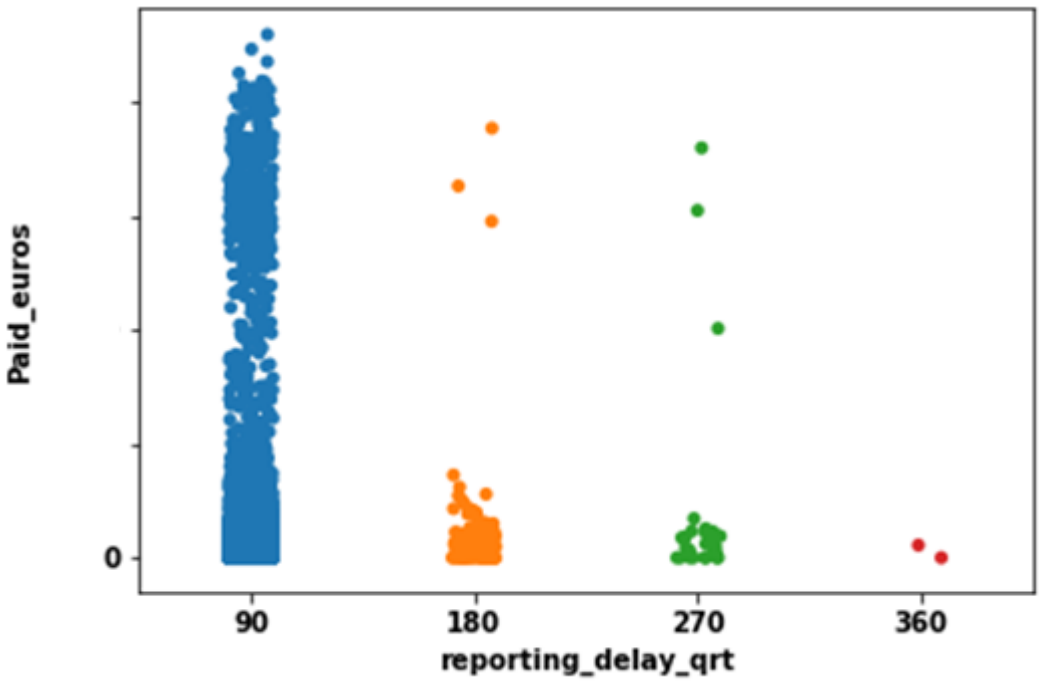
The figure 7 with boxplot shows the month 12 has many claims but the amounts paid are small. The box size with months 24 and 36 illustrates greater financial values of the claims, but they have fewer datapoints. The payment delay in the industry usually depends on the type of the claim and the injury that has occurred. If the claimant was in a severe accident and a doctor's appointment or hospitalisation has been required, the payment delay is longer depending on for example, some external factors such as the surgery queues, need for more treatments or other similar factors compared to smaller injuries which do not require similar care and can be treated sooner. This is the random fluctuations of data in the industry the literature was referring to.

Figure 8. Scatter plot of paid euros vs. settlement delay 2016-2019 distribution



In order to illustrate the settlement delay in days better and gain more valuable insight, the days were divided into quarters (three-month periods). The first class 90 in the figure 8 is including those datapoints which had value smaller than 90 in the data, the rest of them were divided using the same principal. The settlement delay has more datapoints they are distributed more evenly than the other variables. The settlement delay is scattered between the larger claim values towards the right; therefore, it can be concluded larger claims take more time to be settled.

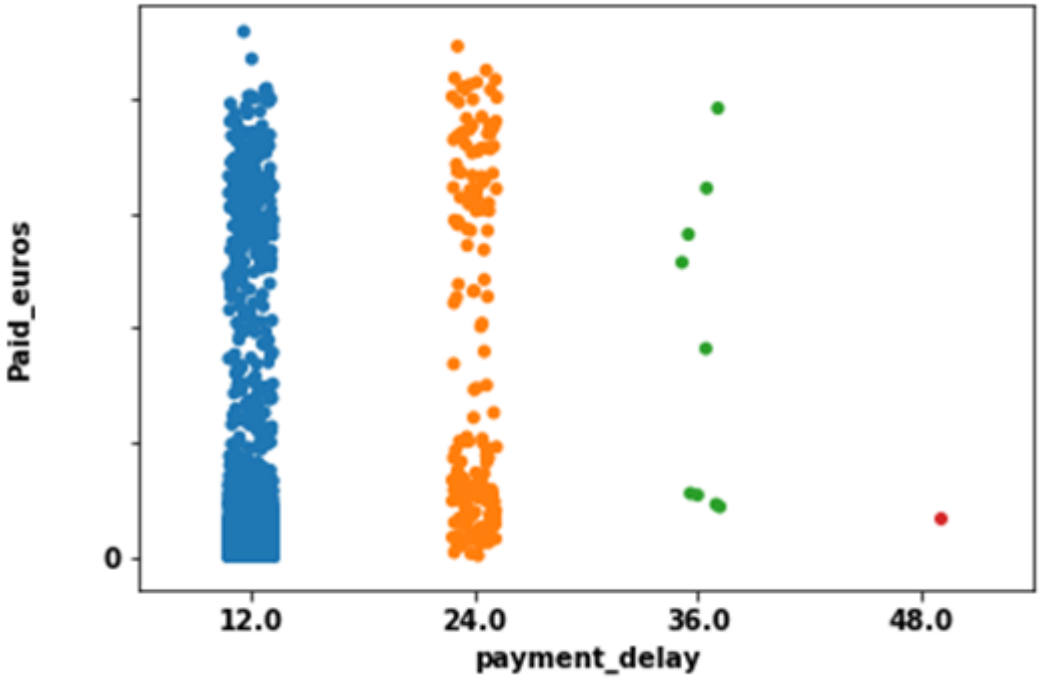
Figure 9. Scatter plot of paid euros vs. reporting delay 2016-2019 distribution



In figure 9 the reporting_delay_days variable was grouped into quarters as well. Most of the claims are reported within three months after the accident has happened. It seems larger claims are reported more promptly than smaller ones. It could be affected by the harmfulness of the accident that occurred to the insured.

Figure 10 visualises the payment_delay which was divided in to 12 months periods.

Figure 10. Distribution of paid euros vs. payment delay 2016-2019



Based on the visualisations from figures 8, 9 and 10 it can be concluded most of these types of claims are paid soon after they have been reported, in general only few claims seem to get paid after years after the accident happened.

3.5.1 Issues with the data

Common problem in the insurance industry is the lack of adequate amount of good quality data to draw solid conclusions. The reasons for this are depending on the nature of the insurance policy, the accident frequency and the accident monetary amount. For example, there might be few incidents per year and the claim amount was substantial or in contrary, there was many accidents and claims but the amounts are small in monetary value.

The dataset was reduced significantly by 30% during the data cleansing and feature engineering phases. Major part of the data was not useful. Immense amount of time was spent

on transforming the data due to the incomplete and dirty data. In addition, there exists heavy data bias with the payment_delay variable. The month 12 is more heavily presented than the other months. The OSLR variable in the dataset that was assumed to be able to use as the output label in the model was missing almost all the information since the values were 0. This could mean there is not enough experience for deeper analysis, or that the data is not recorded properly to make predictive conclusions. A biased dataset means it does not accurately represent the target group to build the model, it results in skewed outcomes, low accuracy and analytical errors (Lim, H., 2020). There exists association bias in the data, even though not in the scope of this thesis the gender distribution is imbalanced too (Lim, H., 2020). If the same data is used in further research, it might not give the adequate results to say what kind or how severe accident occur to the other gender.

The bias in the data was noticed in the beginning of the research and it was tried to be avoided by labelling the data as accurately as possible, talking to the industry experts how to deal with the missing information and what does the missing data mean. It was not wanted to exclude other variables from the dataset, as those exist in the data in real world setting as well and it was not wanted to include more bias that were based on the authors decisions.

3.5.2 Conclusion

Based on the descriptive analysis it was noticed the dataset is not equally distributed. Therefore, logistic regression and random forest methods were the appropriate options to use for the predictions of the development year. The linear ridge regression was considered to be the most appropriate option to predict the number of claims in euros to be paid.

The dataset was divided into subgroups based on the accident year. The model was trained with the 2016 data since it had full actual paid claim values and there were no outstanding loss reserves left (according to the theory).

4 RESULTS

4.1 Predicting loss reserves with chain-ladder method

The run-off triangle with chain-ladder method was created using Excel spreadsheet. The accident_year variable was created from the loss_date column which is the date when the accident had occurred to the claimant. In addition, the payment_delay variable was calculated based on the difference between the reported_date and the paid_date.

The calculation was followed by instructions in Merz & Wüthrich (2008) and Amin et.al. (2020). The accident year, payment_delay and paid_euros variables from the data were used in the triangle. The actual total paid claims in the tringles have been masked with dummy figures in table 8.

Table 8. Total claims paid until 11/2020

	Development year/payment delay			
	0	**1**	**2**	**3**
Accident year	**12**	**24**	**36**	**48**
2016	100,0 €	200,0 €	5,0 €	3,0 €
2017	91,0 €	181,9 €	20,0 €	
2018	91,5 €	182,9 €	10,0 €	
2019	90,4 €	180,7 €		
2020	90,5 €			

According to the theory of chain-ladder, the first accident year should be fully consolidated. However, in this dataset when the numbers are illustrated in triangular format, it can be seen it does not follow the same step-by-step format precisely as in the examples presented before. The development year 3 does not have any actual claims for year 2017, additionally the 2016 does not extend to development year 4. Therefore, this creates uncertainty with the 2016 and 2017 predictions and the cumulative amount if they are correctly reported in the data. Overall, the data quality has raised concerns at the different analytical steps. Many years have passed therefore, both years should be fully consolidated and remain unchanged.

After this step, the cumulative claims were calculated by summing up the development year 0 with development year 1, the same method was consolidated for all the years filling up the triangle. In table 9 below is the filled in cumulative triangle of the paid claims. The numbers are dummy numbers used to mask the actual data.

Table 9. Cumulative claims paid until 11/2020

	Development year/payment delay			
	0	**1**	**2**	**3**
Accident year	**12**	**24**	**36**	**48**
2016	100,0 €	300,0 €	305,0 €	308,0 €
2017	91,0 €	272,9 €	292,9 €	292,9 €
2018	91,5 €	274,4 €	284,4 €	
2019	90,4 €	271,1 €		
2020	90,5 €			

The next step was to calculate the development factors. The development factors were calculated from the development year 1 factors divided by the development year 0 matching factors; the rest were calculated using this same method. After this the run-off tringle was completed with the occurred claim amount divided by the matching development factor. The predictions are written in table 10 with italic and highlighted with blue. The numbers are dummy numbers used to mask the actual data.

Table 10. Completed run-off triangle Chain-Ladder method

	Development year/payment delay			
	0	**1**	**2**	**3**
Accident year	**12**	**24**	**36**	**48**
2016	100,0 €	100,0 €	100,0 €	100,0 €
2017	91,0 €	91,0 €	91,0 €	91,0 €
2018	91,5 €	91,5 €	91,5 €	*91,5 €*
2019	90,4 €	90,4 €	*91,2 €*	*91,2 €*
2020	90,5 €	*108,2 €*	*109,2 €*	*109,3 €*

Next, the outstanding claims and predictions of the ultimate claims were combined in a table. The results of the November 2020 data were compared with the March 2021 updated data. Due to the nature of the insurance business, there exists no fixed figures or estimates where to compare. As noted earlier in this chapter when analysing table 8, the accident year 2016 should have fully consolidated paid claims and remain unchanged.

When the results were compared with the March 2021 data, it was noticed there was more paid claims reported to years 2016 and 2017.

In table 11 below are the predictions of outstanding claims amount as of November 2020, in the column Actual 3/2021 are the outstanding claims from March 2021. The accuracy and precision of the November 2020 result is calculated against the March result. In the table dummy figures have been used (the ratios are the same as in the triangles with actual figures), however the accuracy scores are the actual results.

Table 11. Prediction of outstanding claims and measure of accuracy

Accident year	Prediction 11/2020	Actual 3/2021	Accuracy
2016	- €	- €	100,0%
2017	- €	- €	100,0%
2018	0,0 €	0,0 €	98,7%
2019	0,8 €	0,8 €	97,1%
2020	18,8 €	**18,7 €**	99,7%
Total	**19,6 €**	**19,6 €**	**99,6%**

As, stated by Merz & Wüthrich (2008), the method is quite straightforward and simple, in addition the method is highly accurate as seen from the table 11 results. The original predictions from November did not change much compared to March actuals.

The prediction of the ultimate claim amount did not have much variance either, the accuracy scores varied between 100% to 97%. The accident year 2019 had the most changes between November 2020 to March 2021, which is still expected to have changes in the figures. The method and the calculations were validated by a professional who is used to work with the loss reserving calculations in the company, and confirmed the method was calculated correctly.

4.2 Predicting the claims amount with Ridge regression

The claims dataset was divided into four different subsets by the accident year, in order to be able to predict the claims amount for each year. The 2016 subset was used as the training set, as it is supposed to have the fully consolidated ultimate claims and it had no

outstanding reserves left. The predictions were tested against the test split as well as with the subsets to see how well the model can predict to unseen data.

The documentation and theory of the linear regressions models for multiclass regression stated the model predicts well if the input and output variables have linear relationship and Gaussian distribution. For example, from the figure 4 it was noticed the Paid_euros variable has the Gaussian distribution however, the top of the bell curve is skewed to the left and the dataset was imbalanced. The model was first trained with all the variables available in the dataset, however they were over leaking information to the model. The variable that was passing too much information was the Paid_Not_Paid variable, which was removed in order to avoid overfitting and over leakage of information. The final input features were: 'Insured_gender', 'Policy_period', 'Reserves', 'Paid_1_0','accident_year', 'Manual_reserve', 'OSLR', 'reporting_delay_days', 'settlement_delay_days', 'payment_delay'. The output feature was 'Paid_euros'.

The linear ridge regression training set score was 0.38 and test set score 0.45. The model was tuned by changing the alpha parameter. The alpha values given were decreased as close to zero as possible and increased up to 100 in order to find the optimal setting. However, neither of the alpha values had any influence on the train test set scores. Lasso regression was trained as comparison to see if there was any variance with the results, but the same results were achieved as with ridge regression. The ridge regression model was not able to predict better on unseen data of the different subsets as demonstrated in table 12. The model performance deteriorates when predicted to the 2017, 2018 and 2019 subsets.

Table 12. Linear ridge regression test set results

Test_score	**subset_2017**	**subset_2018**	**subset_2019**
0.45	0.4367	0.3963	0.3960

The comparison of the coefficient magnitudes for ridge regression with different values of alpha was investigated too, however the chart give not any additional insight. The ridge has probably penalized the other features so much that they were all close to zero or were off the charts.

Figure 10. Learning curve of ridge regression on the 2016 claims dataset

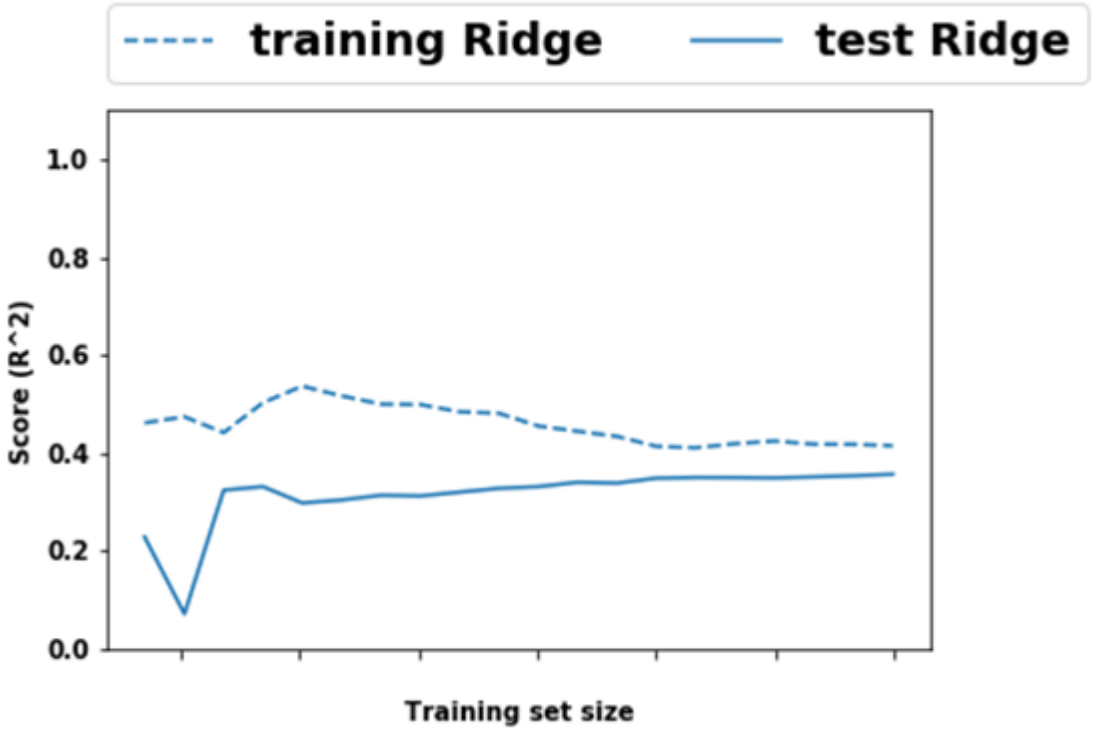


Figure 10 examines how many datapoints the model used for training and for testing. For the smaller datasets the training set is better, however the more datapoints are given the closer the curves get with the test set, and the training performance decreases. The more data there will be available, the risk for overfitting will decreases but for it to memorize the data decreases (Guido & Müller, 2016).

4.2.1 Model evaluation and improvement

Cross validation with five-fold cross validation was done to evaluate the generalisation performance. The method is more stable and systematic, than using the train-test split as it splits the data into five parts that are the same size, after that it trains multiple models (Guido & Müller, 2016).

The default of five-fold cross-validation was used to evaluate the ridge model, table 13 shows the results. The mean crossvalidation score received was 0.40, which means it is expected this model to be 40% accurate. The cross validation scores range from 34% to

51% accuracy. Due to small dataset, it might be dependent on the fold used for training. The other fold can have more datapoints than the other fold.

Table 13. Cross-validation results

Cross-validation scores				
	fit_time	**score_time**	**test_score**	**train_score**
0	0.006000	0.004001	0.422030	0.411379
1	0.008001	0.004999	0.510461	0.383905
2	0.007000	0.005000	0.308568	0.435736
3	0.007001	0.004000	0.343662	0.430171
4	0.005999	0.005000	0.393376	0.416674

Mean times and scores	
fit_time	0.006800
score_time	0.004600
test_score	0.395619
train_score	0.415573

After performing the cross-validation, it can be seen the model can predict on new data at its best with 51% accuracy, which is not very reliable result to base future decisions. The range of the accuracy is vast, there is 17 percentage points difference between the best and the worst result. It might be a result caused by the small dataset, therefore in the future larger dataset could provide more accurate result. It implies the dataset is not very good with predicting the amount of the claims with these variables available in the dataset. However, an answer to the research question that was set in the beginning was obtained. There can exists databias which was caused when the variable paid_not_paid was removed to avoid over leakage of information. On the positive side the cross-validation was able to improve the test set score from 0.45 to 0.51.

4.2.2 Comparison of chain-ladder vs. ridge regression results

The ridge regression and the chain- ladder methods are difficult to compare together since the chain-ladder calculates ultimate claim amount; ridge regression predicts the single amount of claims. The chain-ladder method seems to be easier to understand under this

context of reserving amount of money for the full accident year. Linear ridge regression could be better to be used on predicting individual claims amounts and perhaps attached to different type of information, for example type of insurance products or benefit the claimant has had or the type of injury that has occurred. In the Harej et al. (2017) paper they found the chain-ladder estimates the reserves accurately but badly with individual claims. The dataset includes individual claims.

For predicting and estimating the loss reserves the traditional way seems in this research to be the better option. The issues faced with the dataset throughout the analytical steps of the research raised doubts of the data adequacy and accuracy. The ridge regression results should not be overlooked since it takes more variables into count than the chain-ladder and as stated by the literature the individual claims is not accurately predicted on this type of data. In addition, more data would need to be collected in order to provide more reliable predictions.

4.3 Logistic regression to predict development month

Logistic regression was used in order to predict the development year or development month the claims will be paid. This information is needed in the process of estimating the loss reserves. The problem is handled as multiclassification problem since there are four classes of development months in the data: 12, 24, 36 and 48.

In the logistic regression model, the 2016 data was used as training for the model, which was then predicted to test set 2016 and the subset dataset 2018-2019. The data is highly imbalanced regarding the output value which in this task is the payment_delay feature as was seen the descriptive analysis section. The class imbalance between the accident years of the payment_delay can be seen from table 14:

Table 14. Payment delay value counts 2016-2019

payment_delay	count_2016	count_2017	count_2018	count_2019
12	5717	1377	1440	1481
24	178	44	56	56
36	9	5	3	3
48	1	0	0	0

The tables 15 and 16 present the logistic regression training and test set scores.

Table 15 and 16. Training and train-test set scores

2016 data	Logreg_C=1	Logreg_C=100	Logreg_C=10	Logreg_C=001
Training set score	0.967	0.968	0.968	0.967
Test set score	0.935	0.935	0.935	0.932

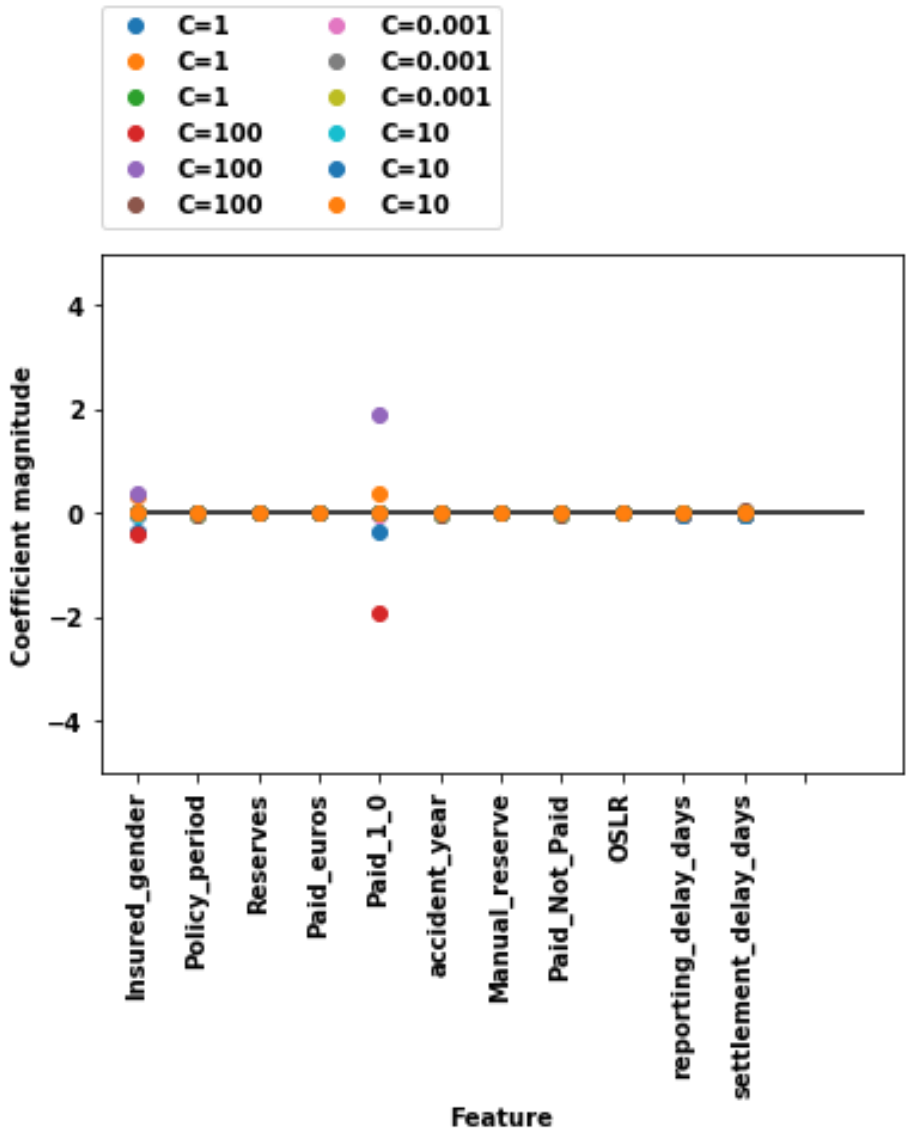
Logistic regression	**2016 test data**	**2018-2019 combined**
Training set score	0.967	
Test set score	0.935	0.972

The logistic regression gives good result, however the performance decreases when introduced predicted to test set with 2016 data. The test set result improved when predicted on the 2018-2019 combined subset. The different values of C do not provide much difference.

Since the 2016 dataset had only 9 datapoints of class 36, it was also trained on combined 2016 and 2017 dataset as comparison to have more datapoints in that group to see the behaviour of the model when more datapoints have been added. The model performance improved when the 2016 and 2017 data was combined and trained. The score was 0.972 in training set and 0.96 with the test set score. The results trained with the 2016 dataset are used here after to retain consistency of the results, and as it was the official dataset which was decided to be used in the beginning for the research problem.

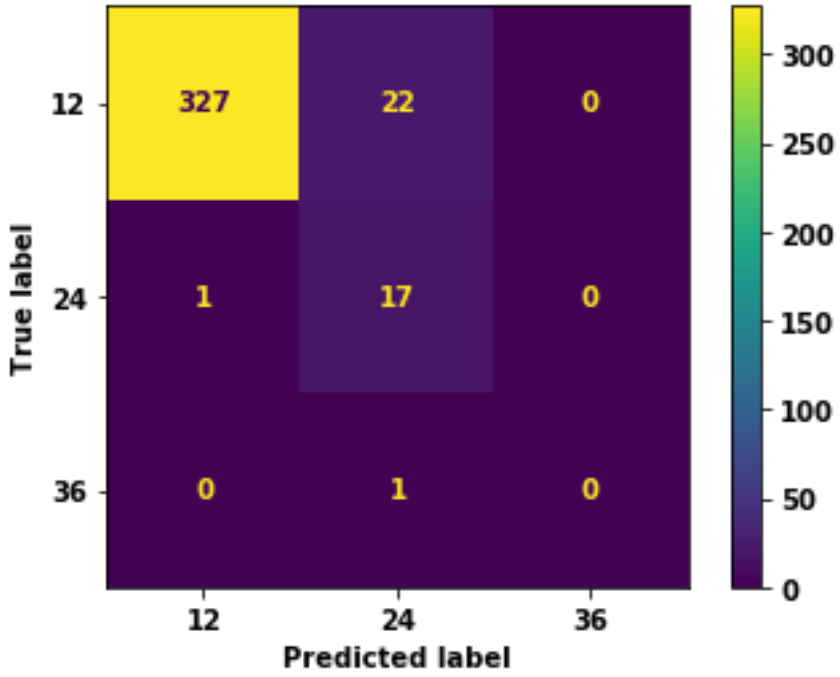
Below it is visualized the relationship of the coefficients of the logistic regression different values of C. Most of the feature coefficients are closer to zero, therefore it can be concluded the prediction is correct. The features Insured_gender and Paid_0_1 have the models with the highest values of C. Depending on the class, there is a shift between positive 2 and negative 2. Downside of the scatter plot in figure 13 is that the class information is not shown, therefore it cannot be concluded which classes cannot be predicted accurately.

Figure 11. Coefficients learned by the logistic regression model



Confusion matrix presents the predictions between true labels and predicted labels in multiclassification problem with true positives, true negatives, false positives and false negatives. The classification performance was analysed with confusion matrix and classification report. The model accuracy is 93,5 percent, which is reasonably good result. With the confusion matrix it was investigated how the model predicts the different classes in figure 12.

Figure 12. Confusion matrix predicted vs. actual results



The confusion matrix shows the model predicted for class 12 true positives result 327, for class 24 true positives result 17, and for the class 36 there was no true positives. However, the model predicted no false positives for class 36 which is a good result. The model classified 23 false negatives belonging to class 24, therefore the classifier predicted the positive class as negative. The biggest errors in the performance and accuracy are coming from this class.

The model accuracy classification report results are shown in table 17, where the precision, recall and f1 scores are reported for each class. The precision score 1.00 for class 12 shows the model can predict with certainty the payment will be paid within 12 months. The recall for this class is 0.94 which means how many actual positives are correctly classified. For class 24 the precision drops significantly to 0.42, therefore the model is not very good predicting if the payment is made in 24 months. The recall is the same as with class 12, therefore the model can predict the same rate of positives for both classes. For class 36 the model cannot make predictions based on the classification report.

Table 17. Classification report logistic regression

Classification report

Output	precision	recall	f1-score	support
12	1.00	0.94	0.97	349
24	0.42	0.94	0.59	18
36	0.00	0.00	0.00	1
accuracy			0.93	368
macro avg	0.47	0.63	0.52	368
weighted avg	0.97	0.93	0.94	368

In addition, the multiclass version of the f-score was calculated as it is the mostly used metric for imbalanced datasets. The micro average f1 score is 0.935 and the macro average f1 score is 0.517. There is quite significant difference between these two results. If the macro average is considered, it would mean the payment delay is unreliable to predict since all the classes are considered equally important.

Random forest classifier was trained as a comparison. Both models' results will be evaluated with the AUC score to select the best model to answer the research question.

4.4 Random forest classifier results

The same logic was used to train the random forest classifier as with the logistic regression model. The 2016 dataset was used to train the model and the predictions were made on the 2018-2019 subsets.

Table 18. Random forest classifier train-test scores

Random forest	2016 test data	2018-2019 combined
Training set score	1.00	
Test set score	0.989	0.997

From table 18 it can be seen the model gives a very good score for the test set and it can predict well on the 2018-2019 combined dataset too.

Figure 13. Random forest classifier feature importance

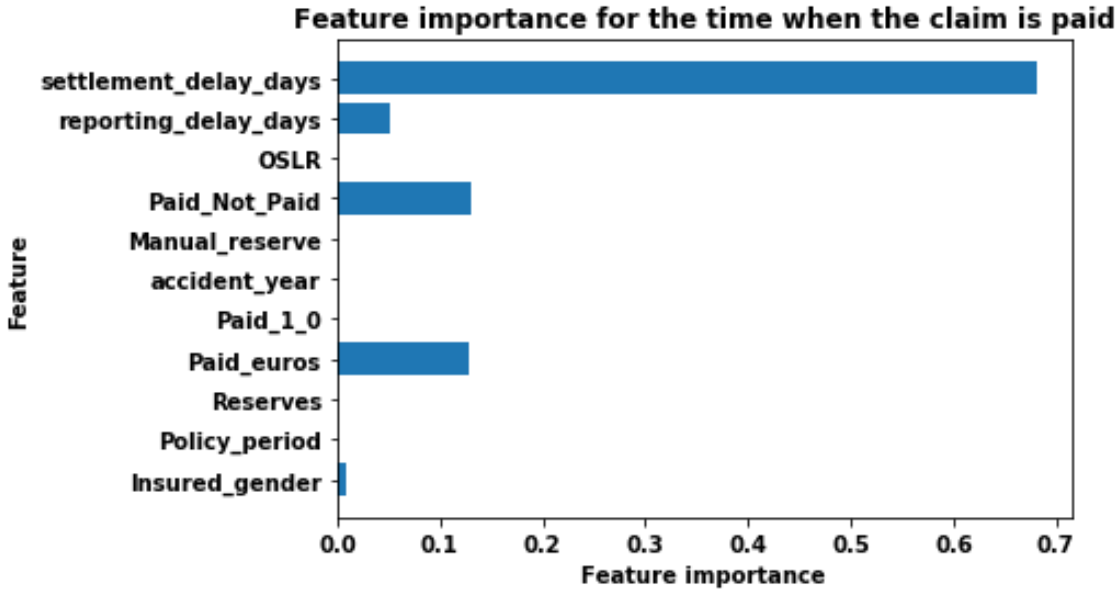


Figure 13 shows the feature importance, the settlement delay is the most important feature in the dataset, which is important information. However, it could also mean this variable is over leaking information to the model. Compared to a single decision tree, the random forest gives more importance's to the other features than just the settlement and reporting delays. In addition, the Paid_euros and Paid_Not_Paid features hold importance. The single decision tree gave results only for settlement delay and payment delays. Decision tree feature importance's can be found from the appendix 2. The benefit of the random forest takes into consideration broader picture of the data compared to a single tree (Guido & Müller, 2016). Therefore, with this type of dataset the random forest provides better view of the results and features affecting them.

4.4.1 Model evaluation and selection

Grid search with cross validation were done in order to improve the model performance. Since the accuracy is not very good estimator on imbalanced data, the final evaluation and selection of the model was analysed using the AUC score.

In order to receive a better generalization performance, the cross-validation was used to evaluate the performance of the different parameter combinations. The grid search with cross-validation selects the parameters with the highest mean validation accuracy. The following values for gamma and C that were given:

- for gamma in [0.001, 0.01, 0.1, 1, 10, 100]
- for C in [0.001, 0.01, 0.1, 1, 10, 100]

In addition, five-fold stratified cross-validation strategy was used as it is in generally done by default.

Table 19. Grid search with cross validation results

Grid search with cross validation results

Test set score:	0.97
Best parameters:	{'C' 0.001, 'gamma' 0.001}
Best cross-validation score:	0.96
Best estimator:	SVC (C=0.001, gamma=0.001)

The evaluation of the model was done on the test set which gave 97 percent accuracy. In table 19 it can be found which are the best parameters and best estimators as well as the cross-validation score. The grid search result of the test set is better compared to the logistic regression and it gave slightly lower score than what was received form the random forest. However, the random forest performed almost perfectly on the other subset, therefore it is possible the model was overfitting on the 2016 test set. Analysis of the grid search results table with the first five rows can be found from appendix 3. The accuracy measurements are not adequate with this problem and it justifies the conclusion that the accuracy measures are not enough to evaluate the model performance.

The AUC (the area under the ROC curve) measures how well the predictions are ranked instead of their total values, it measures the model's predictions regardless of the chosen classification threshold (Agarwal, 2019). The AUC scores for each model were calculated using the OvR scheme for grid search, logistic regression and random forest classifier. The results were:

AUC for Logistic regression: 0.983
AUC for Random Forest: 0.980
AUC for Grid search: 0.969

Based on the AUC score the logistic regression model trained gave the best score, however the random forest classifier performed also very well, but is most likely overfitting. The logistic regression model based on the confusion matrix was able to predict with high accuracy also on the subsets and when trained with the combined dataset. Based on the AUC score the logistic regression model should be used to predict within which development month the claims are being paid.

As a conclusion of this chapter, the model can predict well based on the training data that the claims are most likely paid within 12 months of the claim reported date. Based on the AUC score and the confusion matrix it can be concluded the logistic regression model is accurate enough to base decisions in the future.

5 DISCUSSION

In the discussion part it will be reviewed the different machine learning methods trained, their usability into this type of dataset and if the research questions set in the beginning were able to be answered. Issues and further recommendations how the company should proceed with this data in order to automate and implement the predictions process in their systems are proposed.

The chain-ladder method was easy to use, and it is a simple way to calculate the estimates of the loss reserves. In this thesis the simple chain-ladder method was used. In order to make more advanced calculations there is vast actuarial literature to go through more advanced calculations. As this is stochastic prosses, the more advanced calculations require more mathematical statistics applied. After the extremely in-depth analysis process of this kind of highly imbalanced, sparse and incomplete data, it is no wonder that in general the research around the topic will recommend using the chain-ladder method in the estimations. As seen from the results and comparison with the actuals, the predictions done with chain-ladder method are accurate. There needs to be a set refence point in historical data to which the evaluation can be compared.

From the linear ridge regression results it is quite difficult to see the full value of paid claims similar way as from the chain-ladder methods result. The latter predicts full value of ultimate claims, while the other predicted the financial value of individual claims paid to the claimant. The ridge regression solution therefore did not really answer the research question set in the beginning of the thesis. The ridge regression model score was not accurate enough to base future decisions. It was concluded the ridge regression could be used in other type of research problem for example, concerning the characteristics of the individual claims. The answer to the first research question, with the variables available in the dataset it can be predicted outstanding loss reserves and ultimate claims with the chain-ladder method. The machine learning methods in this context cannot the way it was described in the setting.

The logistic regression problem which development year or month the claims are paid was multiclass prediction problem. The prediction with the dataset was challenging since

the training set needed to be set as balanced for it to work with this dataset. Even when the balanced scaler was used, the model balanced the dataset in a way that it provided binary predictions. There was very few datapoints for classes 36 and 48. In addition, it was difficult to visualise and improve the results as multiclass problem. The scikit learn library usually expected more than one value for each class in order compute the calculations or improve the results. Therefore, it is considered the full potential and correct answer to this prediction problem was not reached the way it was supposed to be achieved. In order to solve this problem properly, the author would have needed more knowledge of how to apply this type of imbalanced data in machine learning setting. The dataset gave many challenges throughout each step of the model training processes regardless of the algorithm that was applied. Based on the logistic regression and random forest classifier findings, it can be concluded machine learning method can be applied with this type of data to predict accurately the development months or years when the claims will be paid out to the claimants. If this is compared to the chain-ladder method, the chain-ladder does not predict the month when the claim will be paid which is disadvantage of the traditional method due to the stochastic nature. The machine learning method's advantage therefore is that it can provide new information and new insight from unused features that the chain-ladder cannot take into consideration.

In order to combine the two machine learning methods to properly compare the findings with the chain-ladder method, algorithm chaining and pipelines should be built. The literature review discussed the usage of artificial neural networks which were not decided to be used in this thesis. The training process of the simpler models showed true what the other researchers on the field have found as well, due to the nature of the data available in the industry and the various challenges.

5.1 Recommendations for future research

For the future is recommended that more data is collected to make accurate prediction. In addition, the company should decide what variables and data is needed to be collected and how it is stored, in order to receive all the needed information and enhance the ability to make prediction for future purposes. If it is wanted to analyse the performance in a fixed point in time, they should agree on at what point in time the data is fixed in order make better estimations and comparisons about the performance. For future reference and in order to automate the claims reserving prediction process, the Python's own library for chain-ladder method is recommended to be tested. In addition, it is recommended the data should be delivered in a format that can be uploaded without much feature engineering and data cleaning. With a few alterations the dataset can be changed in away the preparation of dataset does not become overwhelming process. The data collection could be handled in the same way as it is being delivered to date however, instead of storing the data on various separate files instead, they can be combined and delivered in one file. It would help to automate the processes and reduce manual work for data analysis for the business. In order to find out more insight from this type of challenging data, it would be interesting to see how the ANN's perform.

The value of this research and the outcome of it to the business is, that a method for simple calculation for business professionals was found, without the need of actuarial science knowledge. The descriptive analysis has provided valuable knowledge of the account's performance to the business underwriters. The machine learning methods used can help with evaluation and analysis of the claims in the future with other predictive problems once the data structure is changed and more data has been collected. The methods calculated can be generalised into other type of datasets as well.

6 CONCLUSION

The research investigated if real world claims reserving problem can be predicted with machine learning methods compared to statistical reserving method. The simplest and most used algorithms were trained in order to answer the problem, as actuarial mathematics can be overwhelming to understand and heavy to implement for an individual not familiar with advanced mathematics and statistics. With machine learning methods this kind of predictive process can be easier to understand, automated, and it can reduce manual work and reduce errors in predicting and evaluating profitability of the business.

The thesis reviewed and discussed each method in detail and provided in depth data analysis with statistical and predictive results. It was found out the traditional statistical method in its simplest version is accurate enough and easily implemented to create the prediction in accurate manner. The machine learning algorithms can reach almost the same accuracy, however with the data provided from individual claims point of view. In addition, machine learning methods was noticed to provide new insight from the variables used in the analysis. The predictions were done on small dataset, which was analysed as imbalanced and biased, therefore for future it is recommended that more data should be gathered in order to prepare accurate predictions.

REFERENCES

Agarwal, R., 2019. The 5 classification evaluation metrics every Data Scientist must know. *Towards Data Science, Medium*. Available at https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226. Accessed 13.5.2021.

Amin, Z., Antonio,K., Beirlant, J., Charpentier, A., Dean, C.G., Frees, E. D., Gao, L., Garrido, J., Hua. L., Ismali, N., Kim, J.H.T., Okine, N-A., Saridaş, E.S., Shi, P., Shyamalkumar, N.D., Su, J., Verdonck, T., Viswanathan, K., 2020. Loss Data Analytics. *The Actuarial Community*. Available at Github. Accessed 12.2.2020.

Asif, J., 2018. Granular Reserving Dialogistic in Machine Learning. *Institute and Facility of Actuaries*. Available at: https://www.actuaries.org.uk/system/files/field/document/Reserving%20ML%20Presentation%2020Jun18%20v1.2_2.pdf. Accessed 24.01.2021.

Brownlee, Jason., 2016. *Linear Regression for Machine Learning*. *Machine Learning Mastery.com*. Available at: https://machinelearningmastery.com/linear-regression-for-machine-learning/. Accessed 19.4.2021.

Bryman, A., 2015. *Social Research Methods*. Oxford UniversityPress, Incorporated. Available at: Ebook Central. Accessed: 21.3.2021.

De Alba, E., 2002. Bayesian Estimation of Outstanding Claim Reserves, *North American Actuarial Journal,* Vol.6(4), pp.1-20.

Forfar,D., D.Raymont, D,. 2002. General insurance Definitions.doc. *Actuaries.org*. Available at: https://www.actuaries.org.uk/system/files/documents/pdf/GeneralInsuranceDefinitions.pdf. Accessed 21.4.2021.

FormPlus, 2020. *Correlational Research Designs: Types, Examples & Methods*. Available at: https://www.formpl.us/blog/correlational-research. Accessed 21.03.2021.

Friedland, J., 2010. Estimating Unpaid Claims Using Basic Techniques. *Casualty Actuarial Society*. Available at: https://www.casact.org/sites/default/files/database/study-notes_friedland_estimating.pdf .Accessed 22.4.2021.

Golfin, D., Kuo, K., 2016. *A Machine Learning Framework For Loss Reserving*. KPMG. Available at: https://hubb.blob.core.windows.net/78275ba5-1abe-42e2-9f97-b79ae754dd3c-published/765fbf7b-2223-4e32-941e-f243b61da3e8/AR-1%20-%20A_Machine_Learning_Framework_for_Loss_Reserving_CLRS_final.pdf?sv=2017-04-17&sr=c&sig=2sR6msQ1fyVGfCfjYHonXE7bAsSCuLf-PJmcQ8EsN8lc%3D&se=2021-05-19T03%3A01%3A30Z&sp=r. Accessed 24.9.2020.

Guido, S., Müller, A., 2016. *Introduction to Machine Learning with Python*. O'Reilly Media, Inc. USA.

Harej, B., Gächter, R., Jamal, S., 2017. Individual Claim Development with Machine Learning, *Astin Bulletin - Journal of the IAA*. Available at: actuaries.org. Accessed 3.6.2020.

Joseph, 2020. Bootstrapping Statistics. What it is and why it's used? *Towards Data Science, Medium*. Available at: https://towardsdatascience.com/bootstrapping-statistics-what-it-is-and-why-its-used-e2fa29577307. Accessed 24.9.2020.

Kagan, J. 2020a. Chain Ladder Method – CLM. *Investopedia*. Available at: https://www.investopedia.com/terms/c/chain-ladder-method-clm.asp. Accessed 3.6.2020.

Kagan, J. 2020b. Incurred But Not Reported. *Investopedia*. Available at : https://www.investopedia.com/terms/i/incurredbutnotreported.asp. Accessed 29.11.2020.

Kenton, W., 2020. Stochastic Modeling. *Investopedia.* Available at: https://www.investopedia.com/terms/s/stochastic-modeling.asp. Accessed 21.09.2020.

Kuang, D., B., Nielsen, B., 2020. Generalized log-normal chain-ladder. *Scandinavian Actuarial Journal*, vol. 2020:6, pp. 553-576. Available at: Taylor and Francis Online. Accessed 22.09.2020.

Kuo K., 2019. DeepTriangle: A Deep Learning Approach to Loss Reserving. *Risks*. 7(3). Available at: MDPI. Accessed 22.11.2020.

Lim, H., 2020. 7 types of Data Bias in Machine Learning. *Lionbridge AI Lionbridge Technologies, Inc.* Available at : https://lionbridge.ai/articles/7-types-of-data-bias-in-machine-learning/. Accessed 25.4.20201.

Mack, T., 1994. Which Stochastic Model is Underlying the Chain Ladder Method? *Insurance: mathematics and economics*, vol 15, issues 2-3, pp. 133-138. Available at: Science Direct. Accessed 22.11.2020.

Merz, M., Wüthrich, M.V., 2008. *Stochastic Claims Reserving Methods in Insurance: Reserving Methods in Insurance*. John Wiley & Sons, Incorporated. Available at: Ebook Central. Accessed 20.9.2020.

Merz, M., Wüthrich, M., 2006. Stochastic Claims Reserving Methods in Non-Life Insurance, *ETH Zürich, Switzerland. University Tübingen, Germany*. Available at: www.actuaries.ch. Accessed 20.9.2020.

Mohajon, J., 2020. Confusion Matrix for Your Multi-Class Machine Learning Model. *Towards Data Science, Medium*. Available at: https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826. Accessed 18.4.2021.

Pedregosa, F.,Varoquaux, G., Gramfort, A., Michel, V.,Thirion, B.,Grisel, O., Blondel, M., Prettenhofer, P.,Weiss, R. Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830.

Qiu, D., 2019. *Individual Claims Reserving: Using Machine Learning Methods*, thesis Department of Mathematics and Statistics. Concordia University, Montreal.

Singh, S., 2018. Statistics: Descriptive and Inferential. *Towards Data Science, Medium*. Available at: https://towardsdatascience.com/statistics-descriptive-and-inferential-63661eb13bb5. Accessed 18.4.2021.

Stojiljković. M., 2021. Logistic Regression in Python. *RealPython.com*. Available at : https://realpython.com/logistic-regression-python/. Accessed 29.4.2021.

Uria-Recio, P., 2018. 5 Principles for Big Data Ethics. *Towards Data Science, Medium* Available at: https://towardsdatascience.com/5-principles-for-big-data-ethics-b5df1d105cd3. Accessed 2.5.2021.

Williams, C., 2007. Research Methods, *Journal of Business & Economics Research (JBER)*, vol. 5 (3). Available at: Clute Institute Journals. Accessed 21.3.2021.

Wüthrich, M., 2018. Machine learning in individual claims reserving, *Scandinavian Actuarial Journal*, vol. 6, pp. 465-480. Available at: Taylor & Francis Online. Accessed 21.3.2021.

Bayesian statistics, 2020. Wikipedia. Available at: https://en.wikipedia.org/wiki/Bayesian_statistics. Accessed 3.6.2020.

Hyvä tieteellinen käytäntö. Tutkimuseettisen neuvottelukunnan ohje 2012 (PDF). Tutkimustieteellinen neuvottelukunta (TENK). Available at: https://tenk.fi/fi/tiedevilppi/hyva-tieteellinen-kaytanto-htk . Accessed 9.5.2021.

Outstanding Loss Reserves (OSLR), 2021. International Risk Management Institute, IRMI. Available at: https://www.irmi.com/term/insurance-definitions/outstanding-loss-reserves. Accessed 3.6.2020.

The chain ladder method the most common reserving-method, 2021. The Actuarial Club. Available at: https://theactuarialclub.com/2019/05/17/chain-ladder-method-the-most-common-reserving-method/. Accessed 22.09.2020.

Yleinen tietosuoja-asetus, 2021. Europa.eu. Available at : https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/index_fi.html. Accessed 9.5.202.

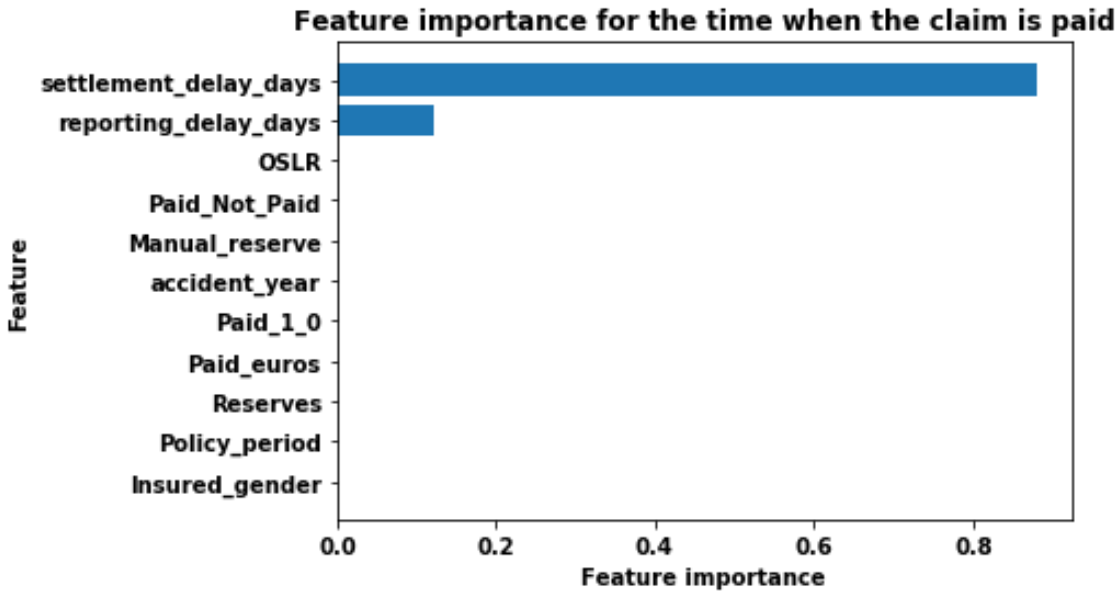
APPENDICES

Appendix 1. Descriptive vs. Inferential statistics

S. No	Descriptive Statistics	Inferential Statistics
1	Concerned with the describing the target population	Make inferences from the sample and generalize them to the population.
2	Organize, analyze and present the data in a meaningful manner	Compares, test and predicts future outcomes.
3	Final results are shown in form of charts, tables and Graphs	Final result is the probability scores.
4	Describes the data which is already known	Tries to make conclusions about the population that is beyond the data available.
5	Tools- Measures of central tendency (mean/median/ mode), Spread of data (range, standard deviation etc.)	Tools- hypothesis tests, Analysis of variance etc.

Source: Singh, S., 2018.

Appendix 2. Decision tree classifier feature importance's



Appendix 3. Analysis of the grid search results

	mean_fit_time	std_fit_time	mean_score_time	std_score_time	param_C
0	0.007010	0.000631	0.002607	0.000498	0.001
1	0.006401	0.000490	0.002985	0.000020	0.001
2	0.007800	0.000761	0.003401	0.000489	0.001
3	0.007598	0.000780	0.003199	0.000401	0.001
4	0.007399	0.000491	0.003399	0.000490	0.001

	param_gamma	params	split0_test_score
0	0.001	{'C': 0.001, 'gamma': 0.001}	0.963801
1	0.01	{'C': 0.001, 'gamma': 0.01}	0.963801
2	0.1	{'C': 0.001, 'gamma': 0.1}	0.963801
3	1	{'C': 0.001, 'gamma': 1}	0.963801
4	10	{'C': 0.001, 'gamma': 10}	0.963801

	split1_test_score	split2_test_score	split3_test_score	split4_test_score
0	0.963801	0.959276	0.959276	0.963636
1	0.963801	0.959276	0.959276	0.963636
2	0.963801	0.959276	0.959276	0.963636
3	0.963801	0.959276	0.959276	0.963636
4	0.963801	0.959276	0.959276	0.963636

	mean_test_score	std_test_score	rank_test_score

0	0.961958	0.002191	1
1	0.961958	0.002191	1
2	0.961958	0.002191	1
3	0.961958	0.002191	1
4	0.961958	0.002191	1

Appendix 4. Explanation of confusion matrix

True Positive (TP): It refers to the number of predictions where the classifier correctly predicts the positive class as positive.

True Negative (TN): It refers to the number of predictions where the classifier correctly predicts the negative class as negative.

False Positive (FP): It refers to the number of predictions where the classifier incorrectly predicts the negative class as positive.

False Negative (FN): It refers to the number of predictions where the classifier incorrectly predicts the positive class as negative.

(Mohajon, 2020)

Appendix 5. CL prediction of when the claim is paid

	Development year/payment delay			
Accident year	**12**	**24**	**36**	**48**
2016	79,8%	20,1%	0,1%	0,1%
2017	81,9%	16,4%	1,7%	0,0%
2018	78,9%	20,0%	1,1%	*0,0%*
2019	92,3%	6,7%	*0,0%*	*0,0%*
2020	82,8%	*0,0%*	*0,0%*	*0,0%*