

---

# **A MACHINE LEARNING APPROACH FOR MODELING FREQUENCY AND SEVERITY OF VEHICLE INSURANCE**

---

**Frits Tuininga**

s1739409

University of Twente

Department of Behavioural, Management and Social Sciences  
Master Industrial Engineering & Management (IEM)

November 4, 2022

**Acknowledgements**

For the creation of this thesis project, I would like to thank my IEM supervisors: Wouter van Heeswijk and Lucas Meertens. Without their extensive feedback on both the direction of the project and the academic value, this thesis could not have taken the shape it has today. In addition, I would like to thank my AM supervisor: Hans Hangyuan. His expertise in the area of machine learning is highly appreciated. My company X supervisor, Malou Smink, had an extremely important role in determining the direction of this project. Malou was always available and ready to share her expertise when required. I am very grateful for all her help when it mattered most. Lastly, I would like to thank Ruben Lucas and Bas van Tintelen for supporting me whenever computer/model related problems occurred.

### Executive Summary

Gradient Boosting Regressor (GBR), eXtreme Gradient Booster (XGB), Random Forest (RF) and Neural Network (NN) with specified parameters do not improve or outperform Generalised Linear Model (GLM) when following the frequency-severity method on vehicle insurance data. The four machine learning models are selected due to their explainable results and outstanding performances in related areas (see Section 2). Feature importance and partial dependence plots are made for GBR, XGB and RF to gain more insight into prediction explainability.

Furthermore, permutation importance and partial dependence plots are created for NN to acquire a better understanding of prediction explainability. In a nutshell, this research consists of two experiments. The first experiment is divided into four phases: pre-processing, training & testing, importance plot creation and evaluation of risk premium predictions. The second research experiment is concerned with the generalisability of the pre-processing. The generalisability of these phases is demonstrated by running the program on another data set (California Housing Data [1]). By generalising these phases, the same machine learning models can be applied to a range of other data sets within the working environment of company X. In conclusion, our study found that when trained on vehicle insurance data, GBR, XGB, RF, and NN cannot outperform GLM. Nonetheless, when trained on different data sets this approach has the potential of improving or replacing other models. Training the models on new data is relatively easy due to the generalisability of the pre-processing and training & testing phase. Therefore, it is strongly recommended to apply the program on different data sets.

## TABLE OF CONTENTS

<b>List of Figures</b>	<b>v</b>
<b>List of Tables</b>	<b>vi</b>
<b>1 Introduction: Problem identification and the potential of improving risk premium prediction model for Company X</b>	<b>1</b>
<b>2 Related Literature and Background: Identification of suitable machine learning models and techniques</b>	<b>3</b>
2.1 Data . . . . .	3
2.2 Generalised Linear Model . . . . .	4
2.3 Suitable Machine Learning Models . . . . .	4
2.4 Overfitting and Underfitting . . . . .	5
2.5 Pre-processing . . . . .	5
2.6 Parameter Selection . . . . .	6
2.7 K-Fold Cross-Validation . . . . .	6
2.8 Partial Dependence Plots . . . . .	7
2.9 Importance Plots . . . . .	7
2.10 Evaluation Metric . . . . .	8
<b>3 Research Design: Machine learning experiment and generalisability</b>	<b>10</b>
3.1 First Experiment . . . . .	10
3.2 Second Experiment . . . . .	13
<b>4 Results: A visualisation of parameter search results, model comparisons, distribution of model predictions and importance plots</b>	<b>14</b>
<b>5 Discussion: The interpretation of experiment outcomes</b>	<b>21</b>
<b>6 Conclusion: Interpretation of results and opportunities of future research</b>	<b>25</b>
<b>References</b>	<b>26</b>
<b>A Generalised Linear Model (GLM)</b>	<b>28</b>
<b>B Variable Histograms</b>	<b>30</b>
<b>C Pie Charts</b>	<b>36</b>
<b>D Regression Tree</b>	<b>41</b>
<b>E Gradient Boosting Regressor (GBR)</b>	<b>44</b>
<b>F eXtreme Gradient Booster (XGB)</b>	<b>45</b>

<b>G Random Forest (RF)</b>	<b>47</b>
<b>H Multi-Layer Perceptron (MLP)</b>	<b>48</b>
<b>I Grid Search Tables</b>	<b>51</b>
<b>J Partial Dependence Plots</b>	<b>53</b>

## LIST OF FIGURES

1	Feature $X_1$ types . . . . .	4
2	Feature $X_2$ types . . . . .	4
3	Example: one-hot encoding [16] . . . . .	6
4	Example: K-fold cross-valuation (k=5) . . . . .	6
5	Example: Feature impact on prediction [21] . . . . .	7
6	Example: Decision Tree [23] . . . . .	8
7	Example: Permutation Importance [24] . . . . .	8
8	First Experiment Setup . . . . .	10
9	Pre-Processing Phase . . . . .	11
10	Training & Testing Phase . . . . .	12
11	Performance Plots Phase . . . . .	12
12	Evaluation Phase . . . . .	13
13	GBR Grid Search (Frequency) . . . . .	15
14	XGB Grid Search (Frequency) . . . . .	15
15	RF Grid Search (Frequency) . . . . .	15
16	NN Grid Search (Frequency) . . . . .	15
17	GBR Grid Search (Severity) . . . . .	15
18	XGB Grid Search (Severity) . . . . .	15
19	RF Grid Search (Severity) . . . . .	15
20	NN Grid Search (Severity) . . . . .	15
21	MAE scores (Frequency) . . . . .	16
22	MAE scores (Severity) . . . . .	16
23	GBR predictions (Frequency) . . . . .	16
24	XGB predictions (Frequency) . . . . .	16
25	RF predictions (Frequency) . . . . .	16
26	NN predictions (Frequency) . . . . .	16
27	GLM predictions (Frequency) . . . . .	16
28	GBR predictions (Severity) . . . . .	16
29	XGB predictions (Severity) . . . . .	17
30	RF predictions (Severity) . . . . .	17
31	NN predictions (Severity) . . . . .	17
32	GLM predictions (Severity) . . . . .	17
33	GBR Feature Importance (Frequency) . . . . .	17
34	XGB Feature Importance (Frequency) . . . . .	17
35	RF Feature Importance (Frequency) . . . . .	17
36	NN Permutation Importance (Frequency) . . . . .	17
37	GBR Feature Importance (Severity) . . . . .	18
38	XGB Feature Importance (Severity) . . . . .	18
39	RF Feature Importance (Severity) . . . . .	18
40	NN Permutation Importance (Severity) . . . . .	18

41	GBR $X_6$ (Frequency)	18
42	XGB $X_6$ (Frequency)	18
43	RF $X_6$ (Frequency)	18
44	NN $X_6$ (Frequency)	18
45	GBR $X_6$ (Severity)	19
46	RF $X_6$ (Severity)	19
47	NN $X_6$ (Severity)	19
48	GBR Grid Search (California Housing)	20
49	XGB Grid Search (California Housing)	20
50	RF Grid Search (California Housing)	20
51	NN Grid Search (California Housing)	20
52	MAE scores (California Housing)	20
53	Example: Ensemble Algorithm	22
100	Regression Tree: one split example	42
101	Regression Tree: multiple splits example	42
102	MLP Structure	48
103	MLP Connections	49

## LIST OF TABLES

1	Example frequency data	3
2	Example severity data	3
3	Average Risk Premiums	19
4	Parameter combinations selected by Grid Search	21
5	Average Risk Premiums	22
6	Parameter combinations California housing	24
7	Feature Table Example	42
8	Iris Data Set Example	47
9	Iris Data Set Bootstrap Example	47
10	GBR, XGB and RF Grid Search (Frequency)	51
11	NN Grid Search (Frequency)	51
12	GBR, XGB and RF Grid Search (Severity)	53
13	NN Grid Search (Severity)	53

# 1 INTRODUCTION:

## PROBLEM IDENTIFICATION AND THE POTENTIAL OF IMPROVING RISK PREMIUM PREDICTION MODEL FOR COMPANY X

*In this chapter, the case is introduced, research problem is identified and primary objectives of this research are elaborated upon. Moreover, research questions are stated and the framework in which this research takes place is described.*

This study focuses on vehicle insurances and is conducted at department A of Company X. Company X is a large Dutch corporation which specialises in financial services. Company X was born after a large series of mergers. These mergers made Company X a large insurer. As of today, key operations of Company X include insurance, asset management, and providing financial services.

Company X issues a variety of insurance products, one of which is vehicle insurance. If a customer has insured his/her vehicle against damage, then risk and (in the case of incidents) associated costs are transferred to Company X. In return, the customer pays Company X a unique risk premium based on customer characteristics. Consequently, it is critical for Company X to determine suitable risk premiums. A suitable risk premium equals expected payout plus a profit margin. If profit margins are too high, customers switch to competitors. If profit margins are too low, Company X generates losses when actual payouts exceed expected payouts. Hence, Company X benefits greatly from a model which accurately predicts cost of damages. This allows Company X to determine risk premiums more accurately and, as a result, improve its competitive position. At this moment, Company X determines risk premiums based on historical data. In the past, Company X has compared this approach with simply setting all risk premiums equal to a constant (i.e. average claim size of all customers plus a small profit margin). The result was that the model based on historical data functioned best (i.e. mean absolute error was smallest). These findings suggest a relation between

independent and dependent variables. The historical data consists of customer characteristics, occurrence of damages and cost of damages. Company X uses the so-called frequency-severity method [2] to generate suitable risk premiums. Frequency-severity method is an actuarial method for calculating expected number of claims received by an insurer over a certain time period and calculating average claim cost. Accordingly, Company X splits its data set in a frequency data set which describes average number of annual claims per customer and a severity data set which describes average cost of damage per customer per year. Company X applies a Generalised Linear Model [3] (Appendix A) to both historical data sets in order to make predictions regarding occurrence (frequency) and cost (severity) of damages. Predictions of occurrence and cost of damages are multiplied and a small profit margin is added to obtain risk premiums. A GLM fits probability distributions on frequency and severity data to generate predictions. These predictions are made based on customer characteristics. In this study, customer characteristics are referred to as features. Features are selected by applying Akaike's Information Criterion [4].

Company X believes their current risk models (GLM) could be improved or replaced by a machine learning models. Machine learning is the study of computer algorithms that improve automatically through experience [5]. These computer algorithms appear in various forms and apply statistical models to optimise performance [6]. If a machine learning model is on average able to outperform GLM, then Company X could apply this model to improve its competitive position. A better risk model enables Company X to determine risk premiums more accurately, which provides Company X with an competitive edge.

The primary research aim is to explore and evaluate a series of machine learning models that have the potential to outperform GLM. Note that it is possible to identify the current GLM as best predictive model.

GLM has the ability to generate explainable predictions. From the perspective of Company X, explainability of model predictions is extremely important. Afterall, if



Company X cannot explain to its customers why their risk premiums changed, then this leads to incomprehensibility on side of the customer. As a result, customers might leave or sue Company X. Company X does not stand alone in the pursuit of explainable models. For instance, the Dutch Council of State identified a series of risks associated with machine learning models including the issue of explainable results. In their report "*Digitalisering Wetgeving en Bestuursrechtspraak*", the Council of State pleads for protection of Dutch citizens against algorithmic decision making in public domain [7]. In addition, the Dutch government issued a series of guidelines for application of algorithms by governmental organisations [8]. Even though Company X is not a governmental organisation, these examples do illustrate developments and importance of explainable results. Moreover, they highlight the relevance of this research. Results are regarded as explainable if variables can be identified that contribute most to prediction. Methods used to identify most contributing variables are elaborated upon in Section 2. Hence, the secondary research aim is to achieve explainable results. In addition, Company X has identified a maximum risk premium its customers are willing to pay. This premium equals €3500 per year. So, models which generate risk premiums that occasionally exceed €3500 must not be chosen to improve or replace GLM.

In recent years, Company X has collected a large number of data points which are used to calculate risk premiums for a variety of different insurances. So, if a selection of machine learning models has the potential to outperform GLM on severity and frequency data sets, then these models might outperform similar models when trained on different data sets. Hence, Company X has a lot to gain from applying machine learning models on all available data sets. For this reason, the tertiary research aim is to automate preparation of data and to automate training and testing machine learning models.

In summary, the three research aims are:

1. **Model exploration:** Explore and evaluate machine learning models which hold the potential to improve or replace the current GLM.

2. **Explainability:** Predictions made by machine learning models must be explainable to the customer (i.e. variables which contribute most to prediction must be identified).
3. **Generalisability:** Pre-processing data, training and testing machine learning models must be automated.

The primary and secondary research aim are combined in the main research question:

*To which extent can explainable machine learning algorithms improve or outperform a Generalised Linear Model when following the frequency-severity method on vehicle insurance data?*

The tertiary research aim is incorporated in the sub-research question:

*To which extent can pre-processing, training and testing be automated for a set of machine learning models?*

In Section 2, relevant papers and contribution to literature is described in more detail. In *Data Description*, an in-depth analysis is given of data as provided by Company X. In *Background*, machine learning techniques are discussed and most suitable machine learning models (given data type) are discussed. In *Research Design*, the research experiment is described and the case for generalisation is made. In *Results*, a visualisation and experiment outcomes are presented. In *Discussion*, an interpretation of results is given and evaluated. In *Conclusion*, the research question is answered and conclusions are drawn based on results. Lastly, recommendations for Company X and opportunities for future research are stated.

## 2 RELATED LITERATURE AND BACKGROUND:

### IDENTIFICATION OF SUITABLE MACHINE LEARNING MODELS AND TECHNIQUES

*In this chapter, a description is given of data as provided by Company X. Furthermore, suitable machine learning models are identified and important machine learning techniques are discussed. In addition, the evaluation metric is elaborated upon.*

### 2.1 Data

As stated in Introduction, Company X must generate risk premiums that are profitable and acceptable for customers. To achieve this, Company X applies the 'frequency-severity method' [2]. This method implies that two data sets are formed: a frequency and severity data set. Both data sets consist of the same independent variables, but data sets differ in number of rows and dependent variable. All insured vehicles appear in the frequency data set (i.e. each row in frequency data refers to an insured vehicle). In contrast, the severity data set consists only of vehicles that suffered damage and for which a claim was issued. As a result, the frequency data set is far larger than the severity data set. Another difference is that the dependent variable for frequency data is average annual number of claims, whereas the dependent variable for severity data is average annual claim size (cost of damage). An example of frequency and severity data is given in Tables 1 and 2 (Var. stands for Variable).

Row	Var. 1	Var. 2	...	Dependent Var.
1	Manual	10	...	0.01
2	Manual	8	...	0
...	...	...	...	...
901169	Automatic	11	...	0

**Table 1:** Example frequency data

Row	Var. 1	Var. 2	...	Dependent Var.
1	Manual	10	...	1000
2	Manual	8	...	300
...	...	...	...	...
48536	Automatic	2100	0	

**Table 2:** Example severity data

For frequency some dependent variables display invalid values. An example of such an invalid value is 365, which would imply that a specific vehicle had 365 annual accidents. Company X explained that this is due to calculation errors. Hence, it was decided to remove all outliers which are clearly invalid. The cutoff point was set at 50 accidents per year. Severity and frequency data sets used in this research were created by collecting data starting from 2011 up and until 2019. The impact of the Corona pandemic was not taken into account on request of Company X.

Since the vast majority of vehicles have not suffered any damages during insurance period, frequency data is very unbalanced. Roughly 95% of all vehicles have not suffered any damage and have therefore an average annual occurrence of damage equal to zero. If a machine learning model consistently estimates zero regardless of vehicle insurance characteristics, then it would be right in 95% of all cases. In that case, however, vehicle characteristics are completely ignored which is exactly what Company X tries to prevent. This problem is discussed in more detail in Chapter 3.

To obtain more insight into feature distributions, for each feature a histogram of unique observations is created. If a unique observation forms 2% or less of all observations, then this observation is grouped under label *Other*. Two histograms are highlighted in this section (Figures 1 and 2). These figures show the way in which observation types could be distributed for a given feature. All remaining histograms and pie charts of the same variables can be found in Appendices B and C respectively.

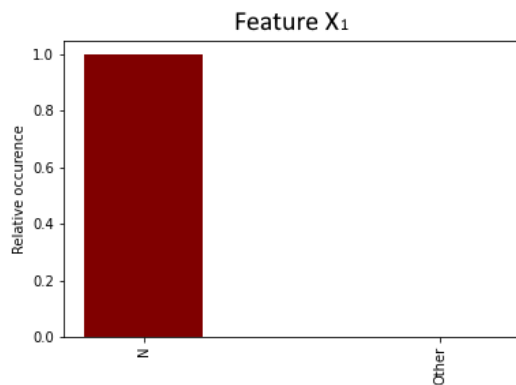


Figure 1: Feature  $X_1$  types

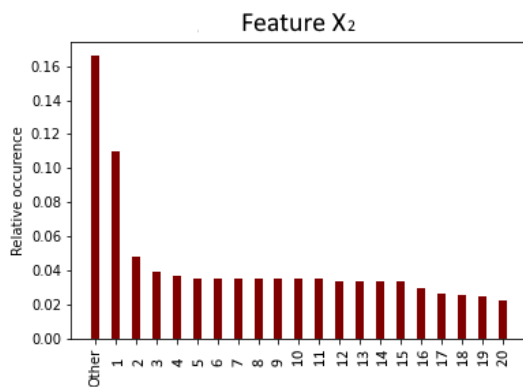


Figure 2: Feature  $X_2$  types

Figure 1 shows two observation types for feature  $X_1$ . As can be seen, observation type 'O' has a relative occurrence of 99%. This illustrates that the vast majority of insured vehicles is quite unbalanced in feature  $X_1$ . In Figure 2 it can be seen that feature  $X_2$  has a variety of unique observations of which many have a relative occurrence of more than 3%. These two features are highlighted to illustrate different ways in which observations could be distributed per feature.

## 2.2 Generalised Linear Model

Company X uses a Generalised Linear Model (GLM) to predict average annual number of claims based on frequency data and average annual claim size (cost of damage) based on severity data. To generate the best possible predictions, GLM bases predictions on probability distributions. Hence, it is key to fit the most suitable

probability distribution on frequency and severity data to generate predictions. For mathematical background of GLM see Appendix A. Omari [9] states that a Gamma distribution is often used as a predictive model for claim severity. In addition, Erdemir shows that a Poisson distribution is often used for modelling claim frequency [10]. The idea that severity data often follows a Gamma distribution and that frequency data often follows a Poisson distribution is substantiated by Ohlsson and Johnson in their book *Non-Life Insurance Pricing with Generalized Linear Models* [11]. Company X follows this approach by assuming that frequency and severity data follow a Poisson and Gamma distribution respectively. Frequency and severity predictions made by GLM are multiplied to obtain suitable risk premiums.

## 2.3 Suitable Machine Learning Models

It is important to determine which type of machine learning models are applicable given frequency and severity data. As stated previously, both frequency and severity data include dependent variables. Since supervised models use target/dependent observations to evaluate and adjust predictions, these type of machine learning models must be used to model frequency and severity data.

Another important aspect of machine learning is the type of target variable. If target variables can be divided in classes, then classification algorithms must be used. In contrast, if target variables are numerical, then regression algorithms must be applied. In both frequency and severity data, target variables are numerical. This implies that regression models must be applied to the data. Note that GLM as used by Company X is a regression model. In summary, only supervised regression models are considered in this research.

In this chapter, an argumentation is given as to which machine learning models are most applicable given the data. Random Forest (RF) makes decisions based on a 'forest' of regression trees. Neural Network (NN) is a collection of nodes called neurons, which model neurons in a biological brain. Staudt and Wagner [12] studied pricing of car insurance contracts by applying

RF, NN and GLM. The aim of this study was to compare these machine learning models to more traditional frequency-severity regression models for car insurance pricing. The study found that RF and NN generated results similar to GLM and could therefore be considered as alternatives to GLM. Furthermore, Gradient Boosting (GB) can be used to make valuable predictions. Similar to RF, GB makes predictions based on a 'forest' of regression trees. The difference lies in how these regression trees are constructed. For more information about construction of regression trees within a GBR or RF model, see Appendices E and G respectively. Guelman [13] presents the theory of Gradient Boosting (GB) and its application to the problem of predicting auto 'at-fault' accident loss cost using data from a major Canadian insurer. The predictive accuracy of GB model is compared to conventional Generalised Linear Model (GLM) approach. Guelman found that GB significantly outperformed GLM in terms of accuracy. In addition, GB produced interpretable results. GB is an umbrella term for a variety of models, but in his research Guelman applied standard GB. Another GB model which must be discussed is eXtreme Gradient Boosting [14]. As described in the book *Gradient Boosting with XGBoost and scikit-learn* [15], this type of Gradient Boosting has outperformed many models in a variety of machine learning competitions. This study applies both eXtreme Gradient Boosting (XGB) and standard Gradient Boosting (GBR) due to impressive results generated by both models.

In the scientific body of knowledge, GBR, XGB, RF and NN are not yet trained and tested on vehicle insurance data (actual name of this data set is censored), and subsequently compared to GLM. Furthermore, this approach is not yet generalised for GBR, XGB, RF and NN. This study aims to extend the scientific body of knowledge by applying GBR, XGB, RF and NN on vehicle insurance data and comparing results to GLM, and standardising this approach to prevent duplication of work. A more in-depth explanation of each machine learning model is given in Appendices E, F, G and H.

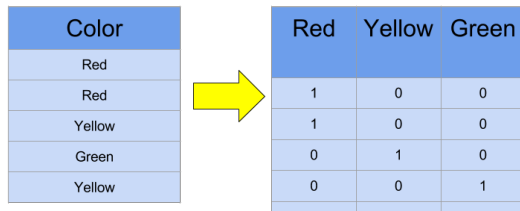
## 2.4 Overfitting and Underfitting

The bias and variance are important tools to evaluate model performance. The bias is an error which refers to average absolute difference between predictions and targets. A model with high bias performs quite poorly on both training and test data. The variance is an error which describes prediction variability [6]. A model with high variance tends to predict training targets very well, but performs extremely poor on test data. Overfitting is the phenomenon where a model performs excellent on train data, but extremely poor on test data (low bias, high variance). Underfitting refers to a situation in which a model performs equally poor on both train and test data (high bias, low variance). Both bias and variance can be influenced by selecting model parameters, which implies a sweet spot that minimises combined error. Furthermore, parameters can be adjusted to prevent over- and underfitting.

## 2.5 Pre-processing

Data must be pre-processed before being fed to machine learning models. In this study, three pre-processing techniques are implemented: removal of non-contributing features, data type transformation and data normalisation. Regarding non-contributing features, it is fairly possible that certain features consist of only one observation type (each row has the same value). Since no correlation between these features and target variable exists, these features increase computation time but do not contribute to predictions. Hence, these features must be removed from the data set. Furthermore, some features have a categorical data type. Machine learning models cannot interpret categorical data directly. Hence, categorical data must be transformed to another data type which can be understood by machine learning models. In this research, one-hot encoding transformation is applied.

As shown by Figure 3, observation types are transformed from a one-column structure to a multi-column structure. Note that by one-hot encoding data, categorical data is transformed to binary data which makes it readable for machine learning models. // Normalisation improves both convergence and generalisation in most tasks for Neural



Color	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	1	0
Yellow	0	0	1

**Figure 3:** Example: one-hot encoding [16]

Networks [17]. Hence, normalisation is an important part of training Neural Networks (Equation 1).

$$X_i^{norm} = \frac{X_i - X_{min}}{X_{max} - X_{min}} \quad (1)$$

In here,  $X_i$  represents the original observation.  $X_{min}$  and  $X_{max}$  refer to minimum and maximum observation value. By applying normalisation, relative differences between observations are preserved but the problem of large observations exists no longer.

## 2.6 Parameter Selection

Grid search is a method to identify best parameters for a machine learning model [18]. Machine learning model performance is highly dependent on its parameters. Since no straight-forward exact method to determine optimal parameters exists, a search method must be applied. Grid search is an exhaustive search method which evaluates a pre-determined number of combinations of parameters for a given machine learning model. The parameter combination which minimises total error is identified as best (given the set of parameters used).

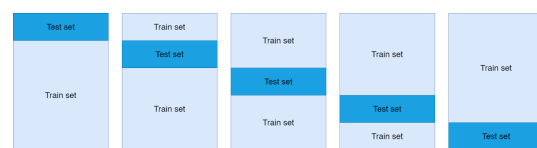
The Neural Network, Random Forest, Gradient Boosting Regressor and eXtreme Gradient Booster are dependent on a series of parameters. For a Neural Network three parameters are important. These are learning rate, number of neurons per layer and number of layers. As shown by Yu and Lui [19], an adaptive learning rate with momentum generates highly accurate results while reducing computational time. For these reasons, an adaptive learning rate is applied for the Neural Network. The best combination of number of neurons per layer and number of layers is derived by grid search.

For Random Forest maximum number of trees and maximum tree depth are the most important parameters for prediction. A larger maximum tree depth, increases both complexity and computational cost. To balance both, often a maximum tree depth of 8 is used [20]. For this reason, this research also applies a maximum tree depth of 8. The best maximum number of trees is determined by grid search.

For Gradient Boosting Regressor and eXtreme Gradient Booster, maximum number of trees, maximum tree depth and learning rate are the most important parameters. As stated previously, to balance complexity and computational cost a maximum tree depth of 8 is most suitable. The best combination of maximum tree depth and learning rate is determined by grid search.

## 2.7 K-Fold Cross-Validation

In machine learning research, data is split into a train and test set. A machine learning model uses train set to establish a mathematical relationship between features and targets. The test set is used for evaluation purposes. Generally, the train set is larger than the test set. A problem that arises is that of an unrepresentative test set (i.e. models perform too well or too poor due to test set structure). To counter this, K-fold cross-validation can be applied. This method utilises the same data set several times for training and testing (Figure 4).



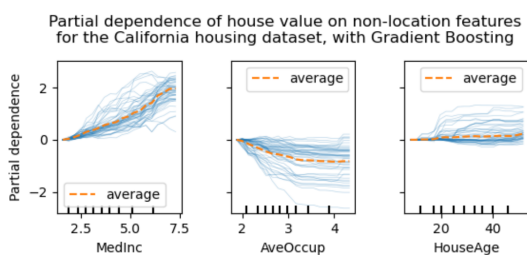
**Figure 4:** Example: K-fold cross-valuation ( $k=5$ )

In the figure above an example is given of standard K-fold cross-validation with  $k = 5$ . In here,  $k$  represents number of times the same data set is used for training and testing. In addition,  $k$  describes test set size. For instance, consider a data set with 100 observations. If  $k = 2$ , then the test set consists of  $\frac{100\%}{k} = \frac{100\%}{2} = 50\%$  of data or 50 observations. If  $k = 4$ , then the test set consists of  $\frac{100\%}{k} = \frac{100\%}{4} = 25\%$  of data or 25 observations. Standard K-fold cross-validation works well for

severity data, but not for frequency data. The reason is that frequency data is unbalanced. In total 95% of targets equal 0, whereas 5% of targets are strictly positive. This was expected since most customers of Company X have not suffered any damages. As a result, average number of damages equals 0 for most customers. An unbalanced data set can result in an unrepresentative test set. After all, it is quite possible that one test set consists of a lot of strictly positive targets whereas another test set has little strictly positive targets. This paints a distorted picture of reality and could result in an ineffective model evaluation. This problem can be solved by applying stratified K-fold cross-validation. In stratified K-fold cross-validation, it is ensured that each test set for each fold consists of an equal number of strictly positive observations. Hence, for severity data standard K-fold cross-validation is used, whereas for frequency data stratified K-fold cross-validation is applied.

## 2.8 Partial Dependence Plots

Partial dependence plots show dependence between one feature and predictions generated by a given model. On x-axis of a partial dependence plot, all unique observations of a particular feature are displayed. On y-axis corresponding change in average prediction is displayed (Figure 5).



**Figure 5:** Example: Feature impact on prediction [21]

In the Figure above three examples of partial dependence plots are given. Consider a data set which includes a feature called MedInc (Figure 5 left). Suppose this data set is fed to a machine learning algorithm and the goal is to construct a partial dependence plot. In that case, all unique observations of MedInc must be gathered and minimum unique observation must be identified. In case

of MedInc, this minimum is 2 (see Figure 5). Next, an adjusted data set is created by setting all observations of feature MedInc equal to this minimum unique observation (i.e. all observations are set equal to 2). Lastly, the data set is fed to the machine learning algorithm and average prediction is registered. This average is the starting point for the partial dependence plot. In other words, all changes in prediction are measured in comparison to this average. Suppose this average equals 100. Let the second lowest unique observation of MedInc equal 2.5. Now, a data set is constructed in which all observations for MedInc are set equal to 2.5. This adjusted data set is fed to the machine learning algorithm and a set of predictions are generated. The average of these predictions is taken and suppose that this average prediction equals 100.5. The difference between average predictions is 0.5 which is shown in the plot (i.e.  $100.5 - 100$ ). In this manner, the partial dependence plot is constructed for all possible unique observations. The dashed line in Figure 5 represents this average prediction and the blue lines represent individual predictions. By constructing partial dependence plots, the impact a feature has on prediction can be measured. This makes it easier to interpret (machine learning) models.

## 2.9 Importance Plots

Importance plots show which features are most important for a given model. Roughly, there are two types of importance plots: feature importance and permutation importance plots. Feature importance plots are solely used for tree based models [22], whereas permutation importance plots are applicable to all models. Feature importance plots rely on GINI importance. To illustrate this idea, consider a decision tree (Figure 6).

In here, there are three parent nodes: 'Has features?', 'Can fly?' and 'Has fins?'. The remaining nodes are referred to as child nodes. For each parent node GINI impurity is calculated (Equation 2).

$$GINI(A) = 1 - \sum_{i=1}^k p_i^2 \quad (2)$$

Here,  $A$  refers to a parent node and  $k$  refers to total num-

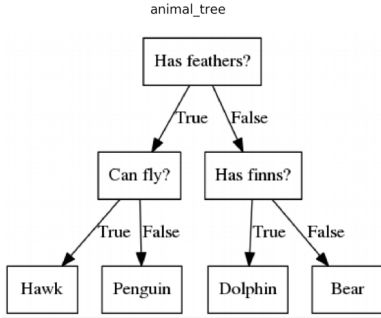


Figure 6: Example: Decision Tree [23]

ber of child nodes.  $p_i$  refers to number of observations in child node  $i$  divided by total number of observations in all child nodes. In this manner, GINI impurity is calculated for each parent node in the decision tree. A feature could appear multiple times in the same decision tree. Hence, this should be taken into account when calculating feature importance (3).

$$f_x = \frac{\sum_{j \in S_x} GINI(j)}{\sum_{j \in S_x} GINI(j)} \quad (3)$$

$f_x$  refers to feature importance of feature  $x$ .  $S_x$  is the set of all features and  $GINI(j)$  is GINI impurity of parent node  $j$ . Hence,  $\sum_{j \in S_x} GINI(j)$  refers to summed GINI importance of all parent nodes in a decision tree corresponding to feature  $x$ . If multiple decision trees are used in a model, then average feature importance of all decision trees in model is taken. As can be seen, feature importance shows relative importance of each feature given a tree based model.

Suppose a data set is fed to a machine learning model. For each observation in the data set a prediction is generated. These predictions are compared to actual target values. The average difference is taken between all predictions and target values. This difference is referred to as accuracy score. To create a permutation importance plot for a feature  $x$ , a new data set must be constructed. To achieve this, the original data set is slightly adjusted. To be more precise, all features in the data set remain the same with exception of feature  $x$ . This feature is randomly shuffled (observations are swapped at random). This newly created data set is fed to the same machine

learning model. Once again average difference is taken between all predictions and target values. As a result, in total two accuracy scores are obtained: an accuracy score based on the original data set and an accuracy score based on slightly adjusted data set. To understand the impact of feature  $x$  on predictive capabilities of the machine learning model, the difference between two accuracy scores is taken. This procedure is executed a given number of times and each time the difference between original accuracy score and accuracy score based on a slightly adjusted data set is taken. Lastly, these differences are averaged and this final result is called permutation importance. Similarly to feature importance, permutation importance shows the impact of a feature on predictive capabilities of a model. Whereas the feature importance plot show relative feature importance, permutation plots show absolute impact of a feature on predictions made. An example of a permutation importance plot is given in Figure 7.

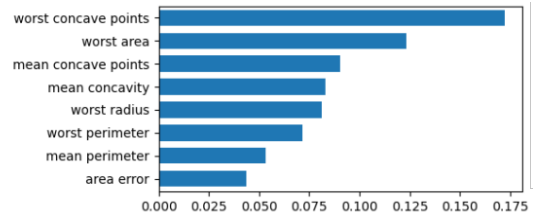


Figure 7: Example: Permutation Importance [24]

In here, the x-axis shows the impact of each feature on prediction accuracy and the y-axis shows each feature.

### 2.10 Evaluation Metric

Suitable performance metrics must be identified to evaluate which machine learning model and combination of parameters generates the best result. Since the problem at hand is a regression problem, three types of performance metrics could be used: Mean Squared Error (MSE), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Both MSE and RMSE emphasise larger errors. Consequently when using MSE or RMSE, models do not focus solely on reducing average error during training. Instead models are incentivised to reduce large errors partly at the expense of average er-

rors. In the context of this study, reducing large errors is irrelevant and the focus must lie on reducing average error. Hence, it was decided to apply Mean Absolute Error (MAE). The MAE refers to mean absolute difference between predictions and targets (Equation 4).

$$MAE = \frac{1}{|S_{test}|} \sum_{i \in S_{test}} |y_i - \hat{y}_i| \quad (4)$$

In here,  $S_{test}$  refers to the test set,  $y_i$  to the target corresponding to observation  $i$ , and  $\hat{y}_i$  refers to prediction. During grid search, MAE is applied to identify the best combination of parameters per machine learning model. Given the best combination of parameters, each model generates a set of predictions for frequency and severity data. Each set of frequency predictions is multiplied by each set of severity predictions. From this result, the average can be taken. This number is the average risk premium which is used to compare the best combination of predictive models. Therefore, MAE is an excellent performance metric to determine whether models have improved or outperformed GLM.



### 3 RESEARCH DESIGN: MACHINE LEARNING EXPERIMENT AND GENERALISABILITY

The research design consists of a research approach and an argumentation as to why certain steps are taken. The research approach is divided into two distinct experiments. The first experiment follows four phases: pre-processing, training & testing, creating performance plots, and evaluation. The second experiment focusses on generalisability of the pre-processing and training & testing phases.

The first experiment consists of four distinct phases: pre-processing, training & testing, creating performance plots, and evaluation. Figure 8 provides an overview of the (first) experiment setup.

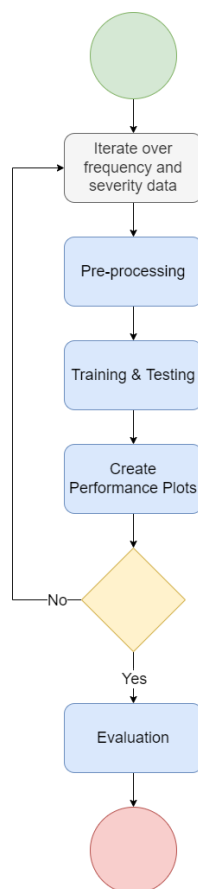


Figure 8: First Experiment Setup

As can be seen, the first three phases (pre-processing, training & testing, and creating performance plots) are executed for both the frequency and severity data set. The last phase (evaluation) is applied on both data sets simultaneously. In the future, Company X wishes to execute similar research for different data sets. In other words, Company X wants to evaluate and compare machine learning models with other models currently in use. Evaluation and comparisons are made based on Mean Absolute Error (MAE) scores. Company X wants to ensure that predictions are made based on customer characteristics. In other words, if models converge to simply predicting averages regardless of inputs, then these models are not useful for Company X. Hence, models are compared on MAE scores but models that learn to predict averages must be excluded. To make the process of evaluation and comparison easier, pre-processing and training & testing phases are standardised.

At the end of this chapter, the second research experiment is discussed in order to show that standardised approach works on another (unrelated) data set and is, hence, applicable to more data sets than solely frequency and severity vehicle insurance data sets. In this research, performance plots are applied to gain insight in importance of attributes. As shown by various studies [25], [26], [27], importance plots are an excellent way to gain insight in attribute importance.

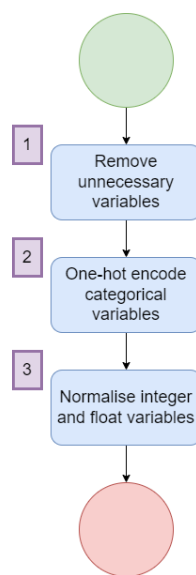
#### 3.1 First Experiment

As shown in Figure 8, the first phase is the pre-processing phase. This phase is visualised in Figure 9 and consists of three actions. First, a set of variables could be removed since these variables are not useful for generating predictions. These variables do not have a negative effect on predictions, but they slow down the training process. For instance, if all observations of a given variable are equal (e.g. only one unique observation exists), then this variable increases computational time but does not contribute to final prediction. Second, categorical variables must be transformed to numeric variables. The reason is that categorical variables are not understood by machine learning models. Frequency and severity data

include several categorical variables. Third, "unscaled input variables [for a Neural Network] can result in a slow or unstable learning process" [17] and, therefore, input variables must be normalised to ensure a stable learning process. Note that normalising data has no impact on tree-based model predictions (such as Gradient Boosting Regressor, eXtreme Gradient Booster or Random Forest).

In summary, the pre-processing phase removes unnecessary variables and adjusts the data set to make it understandable for machine learning models.

(Appendix A). Hence, grid search is unnecessary for this model. During testing performance is measured in MAE scores for GLM. For argumentation as to why MAE scores are used as measurement, see Section 2. The previously discussed steps are executed for both frequency and severity data. As a result of grid search the best performing parameters are identified (given the set of parameters used for grid search).



**Figure 9:** Pre-Processing Phase

As shown in Figure 10, the training & testing phase consists of four actions. First, K-fold cross validation is applied for both data sets. For frequency stratified K-fold cross validation is applied, whereas for severity normal K-fold cross validation is applied. Stratified K-fold cross validation prevents problems which are caused by unbalanced data. For more information on this matter see Section 2. Second, all four machine learning models are trained and tested. During testing, performance is measured with MAE scores. In addition, grid search is applied to generate an overview of the impact of model parameters on predictions made (measured in MAE). Third, GLM is trained and tested. Parameters for GLM are determined with the maximum likelihood estimator

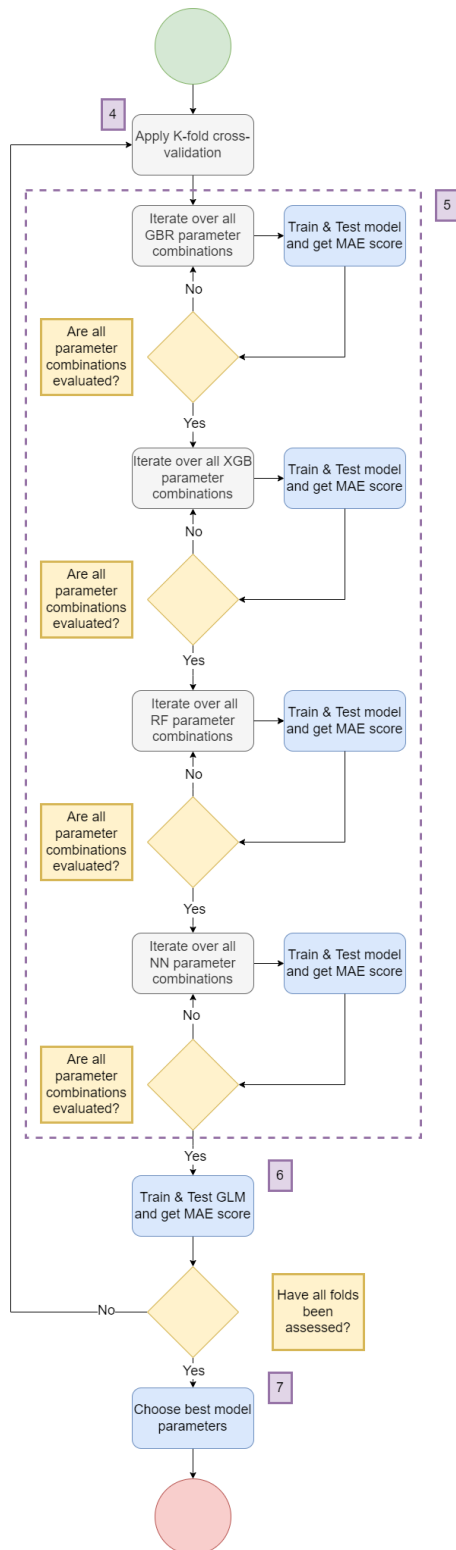


Figure 10: Training & Testing Phase

As shown in Figure 11, creation of performance plots consists of 5 actions.

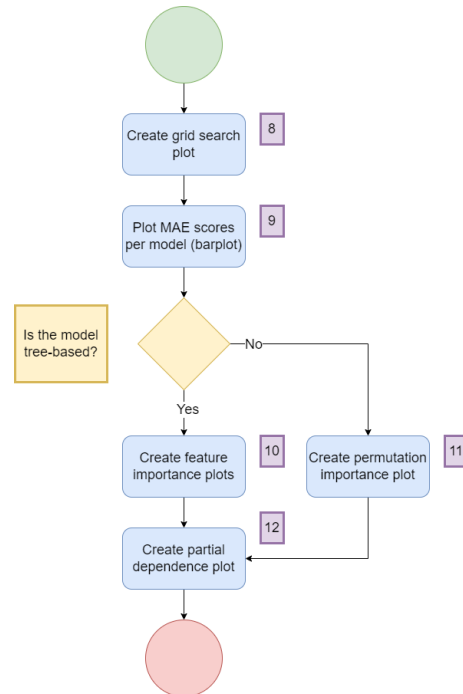
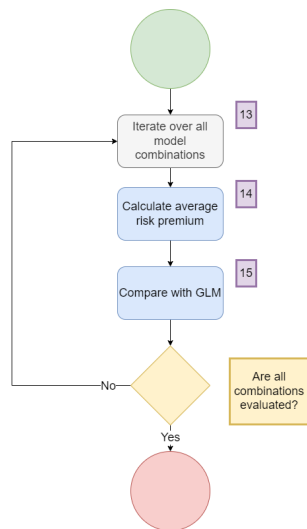


Figure 11: Performance Plots Phase

A grid search plot is created by visualising all possible parameter combinations per model. Predictive accuracy is expressed in MAE scores. These type of plots show in which area there is potential for improvement and which parameter combination generated lowest MAE score. Furthermore, given the best combination of parameters, for each model minimum MAE score is expressed in a bar plot.

For tree-based models (Gradient Boosting Regressor, eXtreme Gradient Booster and Random Forest), feature importance plots are generated. For Neural Network, a permutation importance plot is constructed. Both types of plots give a lot of insight into the impact of individual features on final predictions. Consequently, a ranking of most important to least important features can be made. Lastly, a partial dependence plot is constructed for each model. Similarly to feature and permutation importance plots, partial dependence plots show the impact of a feature on predictive abilities. In contrast to feature and permutation plots, partial dependence plots provide an

understanding of the impact of each unique observation on predictions. For more information regarding these plots, see Section 2. As shown in Figure 12, the evaluation phase consists of 3 actions.



**Figure 12:** Evaluation Phase

As stated previously, the evaluation phase uses both frequency and severity data simultaneously. This works as follows. From previous phases, four machine learning models and the current GLM generated two sets of predictions. The first set of predictions refers to frequency data and the second set of predictions refers to severity data. Hence, in total ten sets of predictions are generated (5 models \* 2 data sets = 10 sets of predictions). Currently, a GLM is used to generate predictions based on frequency and severity data. These predictions are combined into risk premium predictions. To discover whether or not machine learning models may outperform GLM in either frequency, severity or both types of prediction, all combinations of prediction sets must be evaluated. Hence, five prediction sets corresponding with severity data must be combined with five prediction sets corresponding with frequency data. This results in 25 sets of risk premium predictions (e.g. 5 frequency predictions \* 5 severity predictions = 25 sets of risk premiums). Since actual cost and frequency of damages is known, 25 risk premium prediction sets can be compared to the actual situation. In other

words, one can take the difference between the set of risk premiums in an ideal situation (all predictions are exactly correct) and the set of risk premium predictions generated by a combination of two given models. The best combination of two models has the lowest absolute error.

### 3.2 Second Experiment

To illustrate that the pre-processing and training & testing phase can be applied to other data sets, the same program used for Company X is applied on an unrelated data set. The data set selected for this purpose is California housing dataset (obtained from scikit-learn [1]). The data contains information from 1990 California census. In California housing dataset, a clear relation between independent and dependent variables. In addition, this data set is often used for machine learning purposes. For these reasons California housing data is selected as an unrelated data set to show generalisability of the pre-processing and training & testing phase. MAE scores obtained by applying this standardised approach are included in results.

## 4 RESULTS:

### A VISUALISATION OF PARAMETER SEARCH RESULTS, MODEL COMPARISONS, DISTRIBUTION OF MODEL PREDICTIONS AND IMPORTANCE PLOTS

*In this chapter, an overview of all experiment outcomes is given. This implies 8 grid search plots, 2 visualisations of MAE scores (one for frequency and one for severity), 10 histograms which show the distribution of predictions made, 8 importance plots and 7 partial dependence plots are highlighted.*

In following Figures and Tables, Gradient Boosting Regressor, eXtreme Gradient Booster, Random Forest, Neural Network and Generalised Linear Model are referred to as GBR, XGB, RF, NN and GLM respectively. Grid search plots are shown in Figures 13-20. These plots show in which area best results are obtained given the set of parameter combinations. For instance, in Figure 14 it becomes clear that the best combination of parameters for XGB, when trained on frequency data, is achieved when a learning rate of 0.100 is selected in combination with 150 estimators (trees). This combination of parameters results in minimum MAE score given the search space. In Appendix I-13 give a detailed overview of grid search outcomes for frequency and severity data. Given the best combination of parameters, models can be compared based on MAE scores. These results are shown in Figures 21 and 22. Note that the aim is to minimise MAE scores, meaning that GBR and GLM appear to outperform remaining models for frequency and severity data respectively. In addition, based on the best combination of parameters, a histogram is constructed for each model and data set. This histogram shows the prediction distribution per model (Figures 23-32). These histograms show on the x-axis the prediction made by the model and on the y-axis number of times this prediction was made. Hence, these plots give a clear overview of prediction distribution per model. Feature importance plots are displayed in Figures 33-39, whereas permutation importance plots are shown in Figures 36 and

40. Feature importance and permutation plots are constructed to show which features are most important for generating predictions. On the x-axis five most important features are shown and on the y-axis corresponding relative importance is displayed. Figure 33, shows that this feature has a relative importance of 20%. This implies that this feature is the most important feature in 20% of all frequency predictions made by GBR.

Highlights of partial dependence plots are expressed in Figures 41- 47 (for all partial dependence plots see Appendix J). Dependence plots show the impact of observation types on predictions made. For example, in Figure 42 it is shown that changing this feature from 0 to 20 translates to a reduction in frequency predictions by roughly 0.06. Table 3, shows all combinations of frequency and severity models, their average risk premium and error measured. In this table, low errors indicate strong performing models. Lastly, to illustrate generalisability of the pre-processing and training & testing phase, the four machine learning models were fitted to California housing data set [1]. In Figures 48-51, grid search outcomes given California housing data set is given. In Figure 52, accuracy measured in MAE scores of the four machine learning models is displayed.

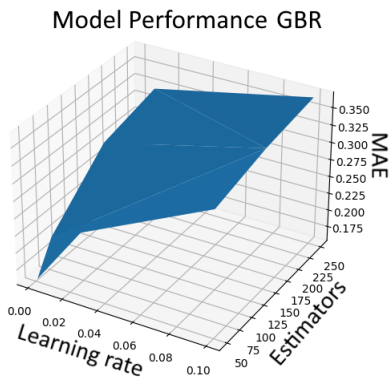


Figure 13: GBR Grid Search (Frequency)

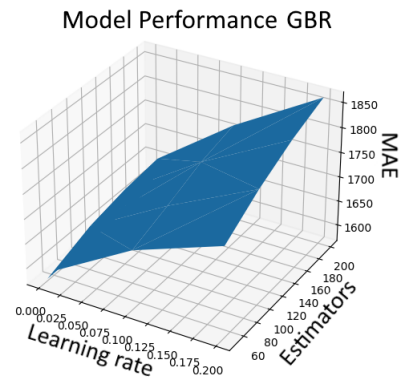


Figure 17: GBR Grid Search (Severity)

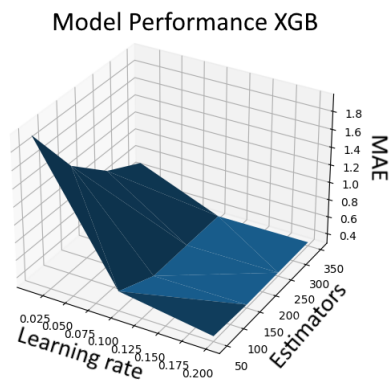


Figure 14: XGB Grid Search (Frequency)



Figure 18: XGB Grid Search (Severity)

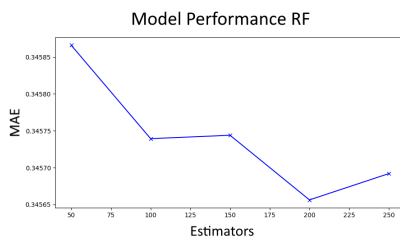


Figure 15: RF Grid Search (Frequency)

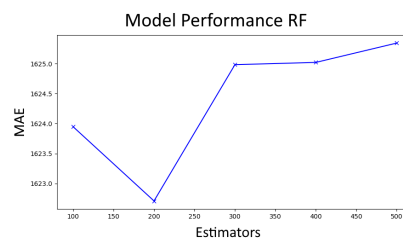


Figure 19: RF Grid Search (Severity)

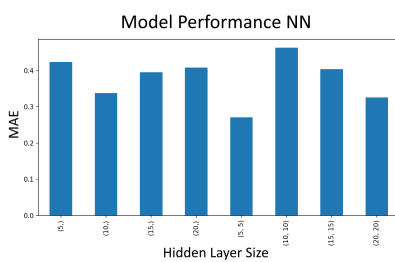


Figure 16: NN Grid Search (Frequency)

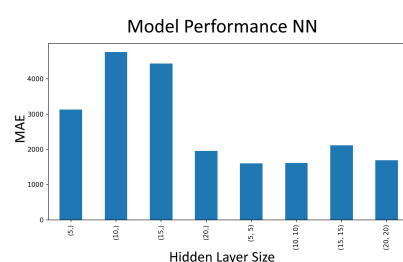


Figure 20: NN Grid Search (Severity)

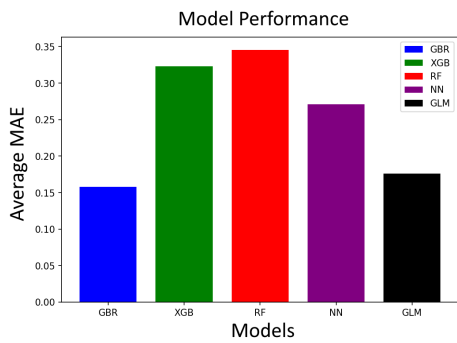


Figure 21: MAE scores (Frequency)

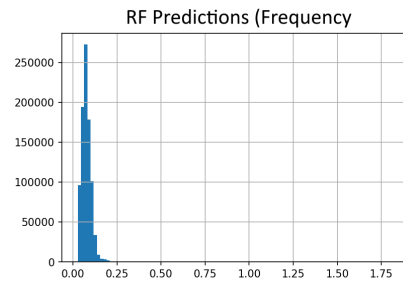


Figure 25: RF predictions (Frequency)

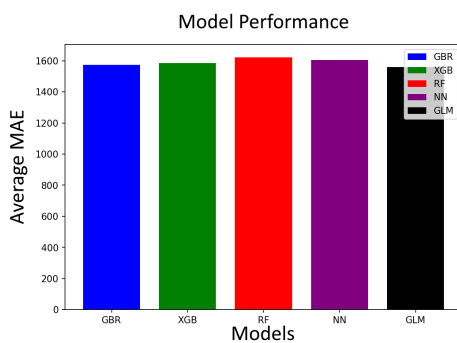


Figure 22: MAE scores (Severity)

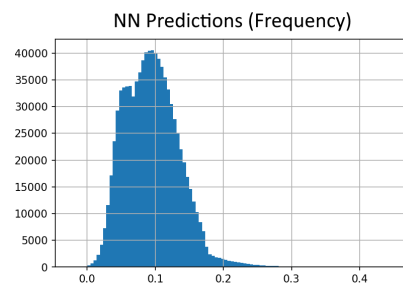


Figure 26: NN predictions (Frequency)

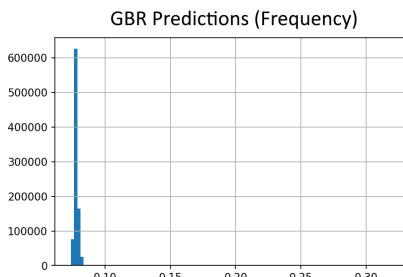


Figure 23: GBR predictions (Frequency)

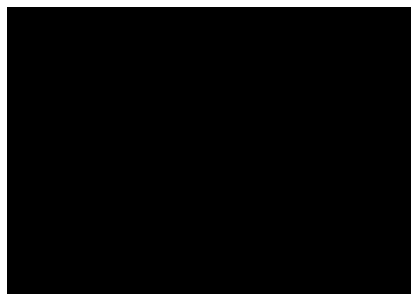


Figure 27: GLM predictions (Frequency)

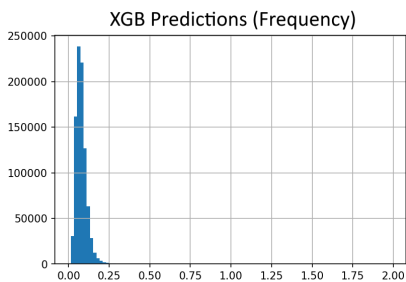


Figure 24: XGB predictions (Frequency)

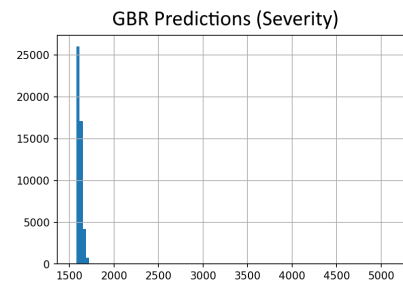


Figure 28: GBR predictions (Severity)

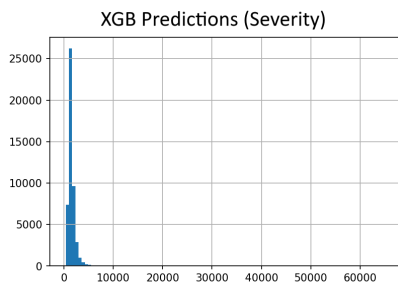


Figure 29: XGB predictions (Severity)

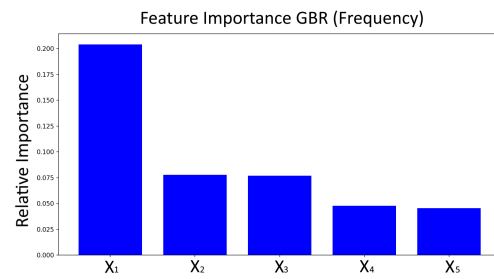


Figure 33: GBR Feature Importance (Frequency)

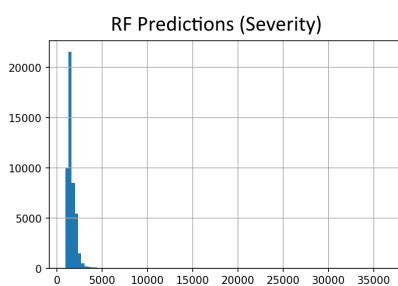


Figure 30: RF predictions (Severity)

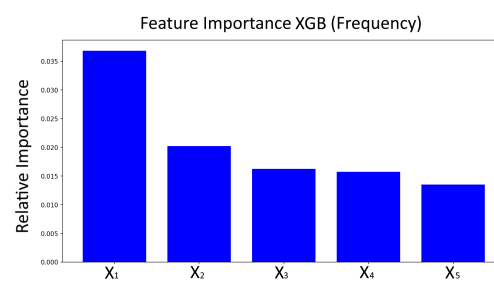


Figure 34: XGB Feature Importance (Frequency)

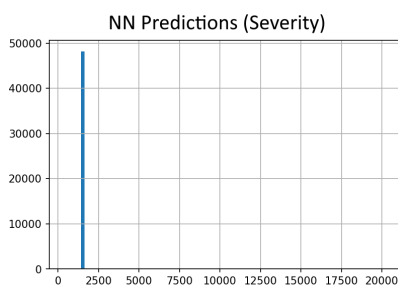


Figure 31: NN predictions (Severity)

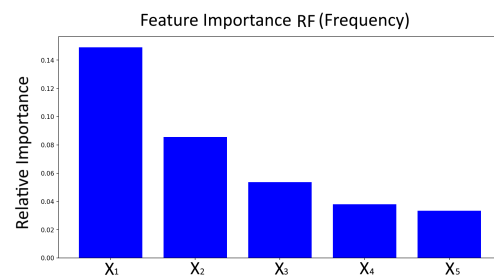


Figure 35: RF Feature Importance (Frequency)

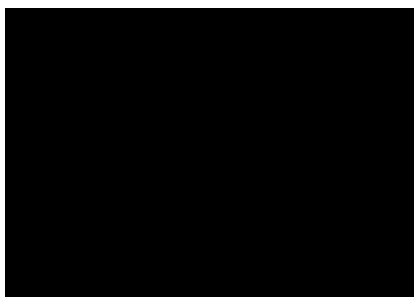


Figure 32: GLM predictions (Severity)

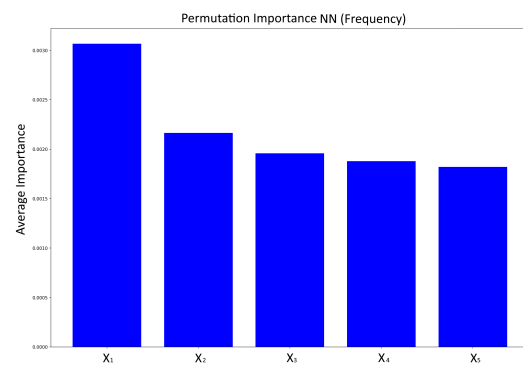


Figure 36: NN Permutation Importance (Frequency)



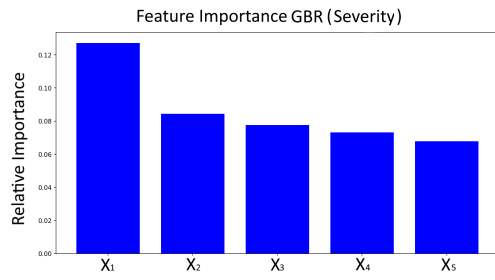


Figure 37: GBR Feature Importance (Severity)

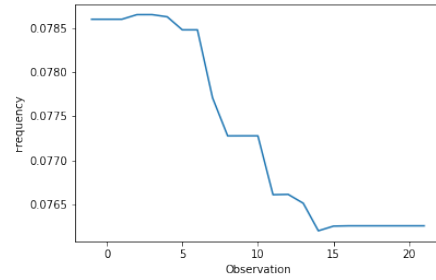


Figure 41: GBR X<sub>6</sub> (Frequency)

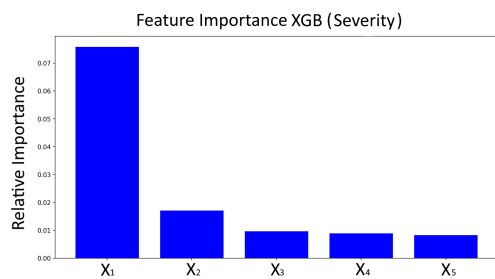


Figure 38: XGB Feature Importance (Severity)

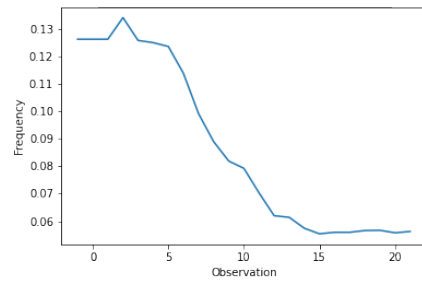


Figure 42: XGB X<sub>6</sub> (Frequency)

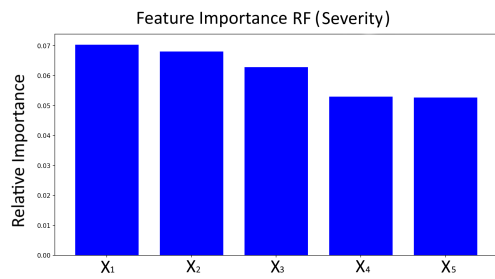


Figure 39: RF Feature Importance (Severity)

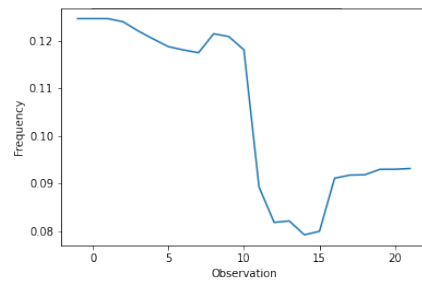


Figure 43: RF X<sub>6</sub> (Frequency)

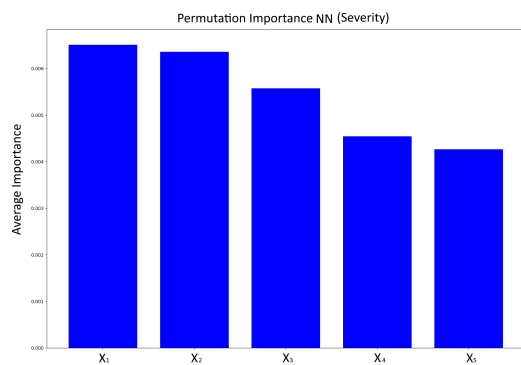


Figure 40: NN Permutation Importance (Severity)

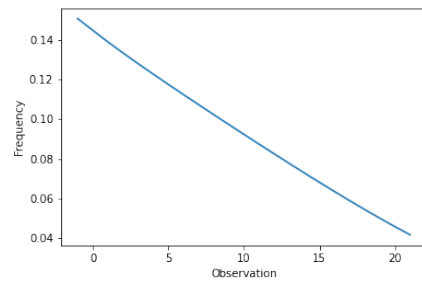


Figure 44: NN X<sub>6</sub> (Frequency)

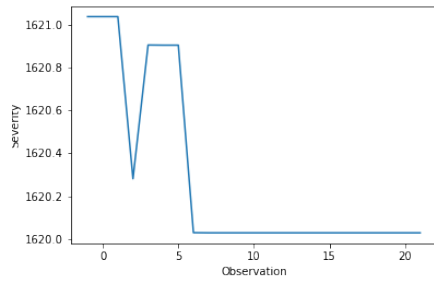


Figure 45: GBR  $X_6$  (Severity)

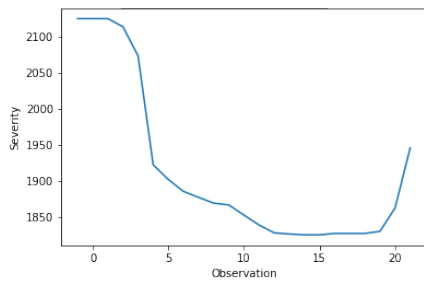


Figure 46: RF  $X_6$  (Severity)

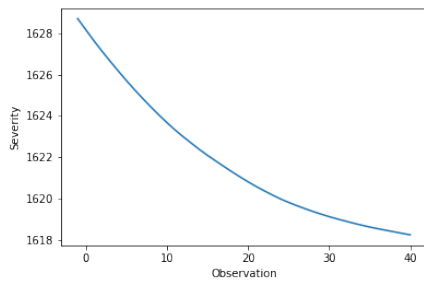


Figure 47: NN  $X_6$  (Severity)

Frequency	Severity	Min. RP.	Max. RP.	Avg. RP.	Error
GBR	GBR	0.33	408.64	126.16	31.19
GBR	XGB	0.11	4374.53	120.59	36.75
GBR	RF	0.21	2359.01	124.88	32.47
GBR	NN	0.14	1714.39	127.35	30.00
GBR	GLM	0.14	448.77	124.85	32.50
XGB	GBR	0.07	1928.17	129.87	27.48
XGB	XGB	0.03	6466.76	125.76	31.58
XGB	RF	0.05	4671.85	129.07	28.27
XGB	NN	0.05	2773.46	131.01	26.33
XGB	GLM	0.04	1784.78	129.06	28.28
RF	GBR	0.12	1399.30	127.40	29.95
RF	XGB	0.04	6847.95	123.71	33.63
RF	RF	0.08	4433.98	127.02	30.32
RF	NN	0.06	2192.84	128.65	28.69
RF	GLM	0.06	1996.16	127.20	30.14
NN	GBR	0.00	789.85	151.81	5.53
NN	XGB	0.00	7068.14	148.58	8.76
NN	RF	0.00	4511.51	152.22	5.13
NN	NN	0.00	3548.98	153.12	4.22
NN	GLM	0.00	1115.16	152.63	4.72
GLM	GBR	0.06	1175.01	151.47	5.87
GLM	XGB	0.04	7545.33	147.36	9.98
GLM	RF	0.07	7243.77	151.74	5.61
GLM	NN	0.06	3677.12	152.86	4.48
GLM	GLM	█	█	█	█

Table 3: Average Risk Premiums

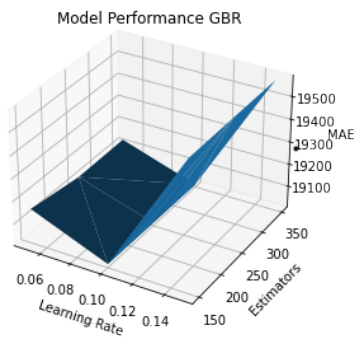


Figure 48: GBR Grid Search (California Housing)

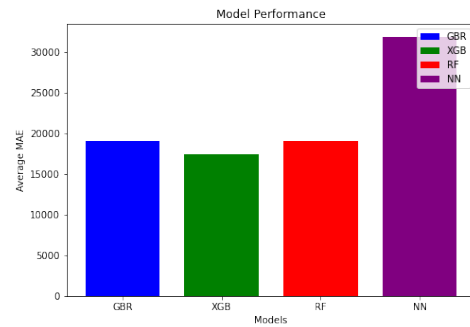


Figure 52: MAE scores (California Housing)

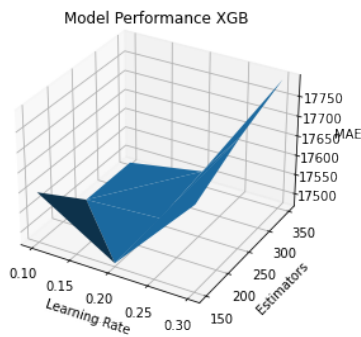


Figure 49: XGB Grid Search (California Housing)

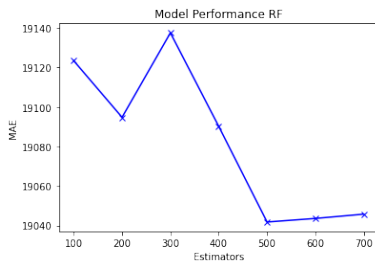


Figure 50: RF Grid Search (California Housing)

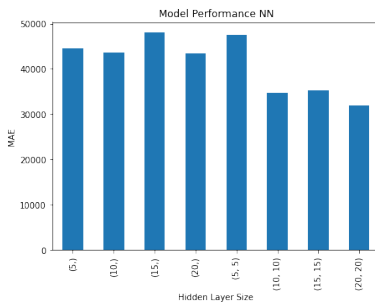


Figure 51: NN Grid Search (California Housing)

## 5 DISCUSSION:

### THE INTERPRETATION OF EXPERIMENT OUTCOMES

*In this chapter, an interpretation of all experiment outcomes is given. Best parameter combinations, best frequency and severity models, and best risk premium predictors are identified. Furthermore, certain model flaws are discussed.*

The four machine learning models cannot improve or outperform GLM. In Figures 13-20, grid search results are visualised per model and data set. These results are summarised in Table 4, which provides an overview of best parameter combinations. These parameter combinations were evaluated based on Mean Absolute Error (MAE) scores.

Name	L.Rate	Trees	Layers	MAE
GBR (Freq.)	0.001	50	x	0.158
GBR (Sev.)	0.001	50	x	1573
XGB (Freq.)	0.100	150	x	0.323
XGB (Sev.)	0.200	300	x	1586
RF (Freq.)	x	200	x	0.346
RF (Sev.)	x	200	x	1622
NN (Freq.)	x	x	(5,5)	0.271
NN (Sev.)	x	x	(5,5)	1604

**Table 4:** Parameter combinations selected by Grid Search

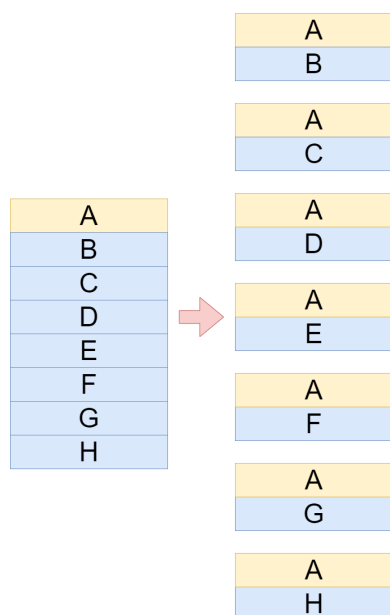
Given these parameters, models were evaluated and compared based on MAE scores (Figure 21 and 22). For frequency data, GBR and GLM outperform other models, whereas for severity data GLM slightly outperforms all other models.

In Figures 23-32, histograms are displayed which show prediction distribution per model. In here, x-axis represents prediction and y-axis represents number of times a prediction was generated. Most models have a prediction distribution which roughly represents a (skewed) normal distribution. Two models must, however, be highlighted. The first is GBR model trained on frequency data. In contrast to other models, this model consistently predicts low target values. As shown by Figure 21, this strat-

egy allows GBR to outperform all other models when trained on frequency data. Recall that the frequency data set is unbalanced. In other words, 95% of all target observations equal zero. Hence, a model which predicts values close to zero yields a very impressive MAE score. As shown by Figure 23, GBR predictions do not vary a lot which implies that features have little impact on predictions made. This is substantiated by the best combination of parameters found by grid search (Figure 13). This combination of parameters leads to most basic GBR model (low number of trees and a low learning rate). Consequently, the model predicts an average and deviates little from this prediction. This is exactly what Company X wishes to prevent. Risk premiums should be calculated based on customer characteristics and should not be an average that has no relation to customer characteristics. Hence, GBR model trained on frequency data might generate impressive MAE scores, it does go against the wishes of Company X and should therefore not be used.

In addition to this model, NN trained on severity data (Figure 31) shows similar behaviour. As stated in Section 2, models that converge to predicting average regardless of inputs are not deemed useful by Company X. For this reason, NN must not be used.

The problem of estimating averages regardless of input is caused by unbalanced data. Liu et al [28] showed that this problem can be solved by applying the ensemble algorithm.



**Figure 53:** Example: Ensemble Algorithm

In Figure 53, the left side represents the entire data set. Subset A represents non-zero target rows and subsets B – H represent zero target rows. As demonstrated, subset A is combined with each zero target row subset. Hence, subset A is used multiple times to construct, in this example, seven balanced subsets (AB, AC, ..., AH). Each balanced subset, has approximately an equal number of non-zero and zero target rows. A machine learning model can now be trained on each balanced subset. Predictions based on each balanced subset are averaged into one prediction. This prediction does not suffer from the consequences of unbalanced data. This is what the ensemble method implies and it is recommended to Company X utilise this algorithm in future research.

From Figures 33-40, it becomes apparent that feature  $X_6$  is seen by most models as an important feature. This feature refers to [REDACTED]. As stated in Chapter 1, Company X applied Akaike’s Information Criterion to select suitable features for GLM. The most important feature was  $X_6$  which is reflected by feature and permutation importance plots.

In Figures 41- 47, one feature is highlighted for each model. A problem that arises here is that of model ex-

plainability. For instance, consider Figure 34. In this Figure, feature  $X_6$  is used. [REDACTED] These sudden price hikes and falls cannot be explained to the customer and appear quite often for GBR, XGB and RF. NN does not suffer from this problem. The reason is that this particular NN has only two layers. This implies that it resembles a polynomial of degree 2. As a result, sudden price hikes/falls cannot occur in this structure. GLM also applies a second order polynomial which does not capture sudden price hikes or falls.

In Table 3, an overview is given of maximum risk premium (Max. RP.), average risk premium (Avg. RP.), and difference between predicted average risk premium and desired risk premium as indicated by Company X (Error). As shown, only model combinations in which an NN or a GLM is used for frequency prediction generate a relatively accurate average risk premiums (low error). Note that if XGB, RF or NN is used for severity prediction, then their maximum risk premium is greater than price ceiling of €3500 set by Company X (Chapter 1). This is unacceptable for individual customers and therefore Company X has indicated not to use XGB, RF or NN for severity prediction. Hence, solely based on this table, combinations in Table 5 seem viable.

Frequency	Severity	Max. RP.	Avg. RP.	Error
NN	GBR	789.85	151.81	5.53
NN	GLM	1115.16	152.63	4.72
GLM	GBR	1175.01	151.47	5.87
GLM	GLM	[REDACTED]	[REDACTED]	[REDACTED]

**Table 5:** Average Risk Premiums

In Table 5, model combinations are shown which generate acceptable maximum and average risk premiums. As discussed previously, GBR suffers from sudden prices hikes and falls which are unexplainable to customers. Furthermore, NN is vastly outperformed by GLM when trained on frequency data and measured by MAE. Consequently, the best combination of models is using GLM for both frequency and severity. In conclusion, given the best combination of parameters the four machine learning models cannot be used to improve or replace GLM. Hence, it is recommended to keep using

GLM for both frequency and severity prediction. In Figures 48, 49, 50 and 51 results of parameter search are shown when models are trained and tested on California housing data set. As can be seen, the program has executed the pre-processing and training & testing phase when applied on an unrelated data set. The accuracy measured in MAE can be found in Figure 52. As demonstrated, the program is able to apply pre-processing on California housing data set, to identify the best combination of parameters and to compare models based on MAE scores. Hence, it is shown that this program can be applied to other data sets than solely data provided by Company X.

The primary research question is:

*To which extent can explainable machine learning algorithms improve or outperform a Generalised Linear Model when following the frequency-severity method on vehicle insurance data?*

Four machine learning models were identified as being explainable and having the potential to improve or replace the model used by Company X. These are: GBR, XGB, RF and NN. The models can be fitted, evaluated and compared with GLM by the four step approach as shown in Figure 8. The result is that these four models with specified parameters cannot improve or outperform GLM when following the frequency-severity method on vehicle insurance data. To be more specific GBR, XGB and RF are unexplainable and NN is less accurate than GLM.

The secondary research question is:

*To which extent can pre-processing, training and testing be automated for a set of machine learning models?*

The pre-processing, training and testing phases can be automated completely. For Company X, a program is developed to standardise the pre-processing and training & testing phase. This program can be applied on different data sets as illustrated by California housing data

set. In this example, XGB has generated lowest MAE score. Best parameter combinations (given the set of parameters) for GBR, XGB, RF and NN are shown in Table 6.

---

---

Name	L.Rate	Trees	Layers	MAE
GBR	0.10	300	x	19067
XGB	0.20	300	x	17477
RF	x	500	x	19041
NN	x	x	(20,20)	30120

---

---

**Table 6:** Parameter combinations California housing

## 6 CONCLUSION:

### INTERPRETATION OF RESULTS AND OPPORTUNITIES OF FUTURE RESEARCH

*In this chapter, a brief summary of results and findings is given. Both research questions are answered, recommendations regarding the future strategy of Company X are made and the opportunities of future research are discussed.*

The four machine learning models with specified parameters cannot improve or outperform Generalised Linear Model (GLM) when following the frequency-severity method on vehicle insurance data. This study has three research aims:

1. **Model exploration:** Explore and evaluate machine learning models which hold the potential to improve or replace the current GLM.
2. **Explainability:** Predictions made by machine learning models must be explainable to the customer (i.e. variables which contribute most to prediction must be identified).
3. **Generalisability:** Pre-processing data, training and testing machine learning models must be automated.

From literature search, it became apparent that four machine learning models held the potential to outperform GLM in context of vehicle insurance data: Gradient Boosting Regressor (GBR), eXtreme Gradient Booster (XGB), Random Forest (RF) and Neural Network (NN). Feature importance plots are used to evaluate explainability of GBR, XGB and RF, whereas permutation importance plots are used to evaluate explainability of NN. The plots display most important features for each model. Partial dependence plots are used to assess the impact of most important features on predictions. Customers perceive the machine learning model as unexplainable if sudden price increases/falls occur. GBR, XGB, and RF are subject to such sudden price fluctuations. As a result, these models must not be used in place of GLM. Hence,

only combinations of GLM and NN are viable solutions. NN is, however, outperformed by GLM when Mean Absolute Error (MAE) scores are registered. Hence, GLM must be used to predict both frequency and severity. As a result, it is advised to continue using GLM for risk premium calculation.

To illustrate generalisability, the program used for pre-processing and training & testing phases is applied to another dataset: California housing. Generalisability is shown by applying the same program on different data sets. The program is able to apply pre-processing on California housing data, to identify the best combination of parameters and to compare models based on MAE scores. Hence, it is shown that this program can be applied to other data sets than solely data provided by Company X. Hence, to save time, it is recommended to apply this program to other data sets within Company X when similar research is conducted.

As stated in Section 2, frequency data set is unbalanced. To mitigate problems associated with an unbalanced data set, stratified K-fold cross validation is applied. As illustrated by GBR when trained on frequency data, it is still possible for a machine learning model to systematically estimate 0 regardless of the set of features. Hence, the problem is not completely resolved. As stated in Section 2, stratified K-fold cross validation solves the problem of unreliable train and test sets, but it does not guarantee that models do not fall in the trap of estimating constants regardless of inputs. Hence, for future research it is advised to solve this problem by implementing methods to handle unbalanced data.

Another area in which results can be improved is parameter search. Once again due to time constraints and acceptable results, the best parameters generated by grid search were used. However, actual best combination of parameters could lie around the point chosen. For instance, instead of using a learning rate of 0.100 and 150 number of trees by XGB (severity) it could be that a learning rate of 0.094 and 133 number of trees is optimal. Hence, more in-depth grid search is advised for future research. To be more specific, it is advised to apply a



more fine-grained grid around the current minimum. As discussed by Wu et al. [29], Deep Neural Networks can be constructed and tuned in such a way that they mirror GLMs. As stated by the author '*Neural Networks are nothing more than recursive Canonical GLMs*'. Hence, applying a neural network on frequency and severity data might generate more success when different activation functions and layer structures are considered. Thus, it is recommended to explore Deep Neural Networks in more detail when applied to frequency and severity data. The model can still be explainable by applying permutation importance plots. Finally, predictions were made using historical data, but there is no guarantee that the past can always be used to predict the future accurately. For example, if the Dutch government decides to invest significantly less in infrastructure, accidents may become more common. As a result, even though the model used by Company X did not anticipate this, more customers may file claims. This fact must not be overlooked. As stated previously, the program developed is able to apply the pre-processing and training & testing phase on multiple data sets. To truly demonstrate generalisability, it is recommended to apply the program to a multitude of different regression data sets. Furthermore, the program could be expanded by including performance plots.

## REFERENCES

- [1] "California housing." (2022), [Online]. Available: [https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch\\_california\\_housing.html](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.fetch_california_housing.html) (visited on 2022).
- [2] E. Z. Fisman, I. Shapira, M. Motro, A. Pines, and A. Tenenbaum, "The combined cough frequency/severity scoring: A new approach to cough evaluation in clinical settings.," *Journal of medicine*, vol. 32, no. 3-4, pp. 181–187, 2001.
- [3] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society*, 1972.
- [4] T. W. ARNOLD, "Uninformative parameters and model selection using akaike's information criterion," *The Journal of Wildlife Management*, vol. 74, no. 6, pp. 1175–1178, 2010. DOI: <https://doi.org/10.1111/j.1937-2817.2010.tb01236.x>. eprint: <https://wildlife.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1937-2817.2010.tb01236.x>. [Online]. Available: <https://wildlife.onlinelibrary.wiley.com/doi/abs/10.1111/j.1937-2817.2010.tb01236.x>.
- [5] T. M. Mitchell and M. Learning, "Mcgraw-hill science," *Engineering/Math*, vol. 1, p. 27, 1997.
- [6] M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [7] Council of State. "Digitalisering, wetgeving en bestuursrechtspraak." (2021), [Online]. Available: <https://www.raadvanstate.nl/@125918/publicatie-digitalisering/>.
- [8] Dutch Government. "Richtlijnen voor het toepassen van algoritmen door overheden en publieksvoorlichting over data-analyses." (2021), [Online]. Available: <https://www.rijksoverheid.nl/documenten/richtlijnen/2021/09/24/richtlijnen-voor-het-toepassen-van-algoritmen-door-overheden-en-publieksvoorlichting-over-data-analyses>.
- [9] C. Omari, S. Nyambura, and J. Mwangi, "Modeling the frequency and severity of auto insurance claims using statistical distributions," 2018.
- [10] Ö. K. ERDEMİR and Ö. KARADAĞ, "On comparison of models for count data with excessive zeros in non-life insurance," *Sigma Journal of Engineering and Natural Sciences*, vol. 38, no. 3, pp. 1543–1553, 2020.
- [11] E. Ohlsson and B. Johansson, *Non-Life Insurance Pricing with Generalized Linear Models*. Springer, 2010.

- [12] F. Y. Wu and K. K. Yen, "Applications of neural network in regression analysis," *Computers & industrial engineering*, vol. 23, no. 1-4, pp. 93–95, 1992.
- [13] L. Guelman, "Gradient boosting trees for auto insurance loss cost modeling and prediction," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3659–3667, 2012.
- [14] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.
- [15] C. Wade, *Gradient Boosting with XGBoost and scikit-learn*. Packt Publishing Ltd., 2022.
- [16] P. Rodríguez, M. A. Bautista, J. González, and S. Escalera, "Beyond one-hot encoding: Lower dimensional target embedding," *Image and Vision Computing*, vol. 75, pp. 21–31, 2018, ISSN: 0262-8856. DOI: <https://doi.org/10.1016/j.imavis.2018.04.004>. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0262885618300623>.
- [17] J. Shao, K. Hu, C. Wang, X. Xue, and B. Raj, "Is normalization indispensable for training deep neural network?" *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 434–13 444, 2020.
- [18] P. Liashchynskiy and P. Liashchynskiy, "Grid search, random search, genetic algorithm: A big comparison for nas," *arXiv preprint arXiv:1912.06059*, 2019.
- [19] C.-C. Yu and B.-D. Liu, "A backpropagation algorithm with adaptive learning rate and momentum coefficient," in *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No. 02CH37290)*, IEEE, vol. 2, 2002, pp. 1218–1223.
- [20] M. Segal, "Machine learning benchmarks and random forest regression," *University of California*, 2003.
- [21] *scikit learn*. "Partial dependence and individual conditional expectation plots." (2022), [Online]. Available: [https://scikit-learn.org/stable/modules/partial%5C\\_dependence.html](https://scikit-learn.org/stable/modules/partial%5C_dependence.html) (visited on 2022).
- [22] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "Evaluating feature importance estimates," 2018.
- [23] C. Kingsford and S. L. Salzberg, "What are decision trees?" *Nature biotechnology*, vol. 26, no. 9, pp. 1011–1013, 2008.
- [24] Scikit-learn. "Permutation importance with multicollinear or correlated features." (), [Online]. Available: [https://scikit-learn.org/stable/auto%5C\\_examples/inspection/plot%5C\\_permutation%5C\\_importance%5C\\_multicollinear.html](https://scikit-learn.org/stable/auto%5C_examples/inspection/plot%5C_permutation%5C_importance%5C_multicollinear.html) (visited on 2022).
- [25] H. Marcos-Pasero, G. Colmenarejo, E. Aguilar-Aguilar, A. R. de Molina, G. Reglero, and V. Loria-Kohen. "Ranking of a wide multidomain set of predictor variables of children obesity by machine learning variable importance techniques." (2021), [Online]. Available: <https://www.nature.com/articles/s41598-021-81205-8#Sec2> (visited on 2021).
- [26] Y. Zhao, J. Zhang, T. Yuan, *et al.* "Study on rapid identification of medicinal plants of paris polyphylla from different origin areas by nir spectroscopy." (2014), [Online]. Available: <https://europepmc.org/article/med/25269290> (visited on 2021).
- [27] K. Maheswari and A. Valarmathi. "A novel approach for gene selection based on random forest-variable importance." (2019), [Online]. Available: [https://www.ripublication.com/ijaersp12019/ijaerv14n15spl\\_19.pdf](https://www.ripublication.com/ijaersp12019/ijaerv14n15spl_19.pdf) (visited on 2021).
- [28] L. Liu, X. Wu, S. Li, Y. Li, S. Tan, and Y. Bai, "Solving the class imbalance problem using ensemble algorithm: Application of screening for aortic dissection," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–16, 2022.
- [29] F. Y. Wu and K. K. Yen, "Applications of neural network in regression analysis," *Computers & industrial engineering*, vol. 23, no. 1-4, pp. 93–95, 1992.
- [30] L. Williams, "A modification to the half-interval search (binary search) method," *Association for Computing Machinery*, 1976.
- [31] K. Dwyer and R. Holte, "Decision tree instability and active learning," *Springer*, 2007.
- [32] A. Papanicolaou, "Taylor approximation and the delta method," *April*, vol. 28, p. 2009, 2009.
- [33] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [35] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.