CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES
Institute of Economic Studies

Machine Learning Methods in Motor Insurance Fraud Detection

Master's thesis

Author: Bc. Barbora Bajgarová

Study program: Economics and Finance

Supervisor: doc. PhDr. Jozef Baruník, Ph.D.

Year of defense: 2024

Declaration of Authorship

The author hereby declares that he or she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain any other academic title.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis in whole or in part and agrees with the thesis being used for study and scientific purposes.

Prague, July 30, 2024

Barbora Bajgarova

Abstract

This thesis explores the application of machine learning models for detecting fraudulent claims in motor insurance. It compares the effectiveness of several algorithms, including logistic regression, random forest, XGBoost, histogram-based gradient boosting, and multilayer perceptron (MLP). The study addresses the challenge of class imbalance in fraud detection, utilizing techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and class weighting to enhance model performance. Real-world data provided by UNIQA pojišťovna a.s., including detailed information on insurance contracts and claims, serve as the basis for the empirical analysis. Among the models tested, XGBoost with SMOTE resampling and class weighting achieved the highest recall rate, detecting over 90% of fraudulent claims, while maintaining a reasonable level of precision. The feature importance analysis highlighted key predictors of fraud, such as claim amount, type of coverage or vehicle age. The findings underscore the potential of advanced machine learning techniques to improve the efficiency of fraud detection systems in the insurance industry.

JEL Classification	C49, G22, K42,
Keywords	machine learning, fraud detection, insurance, unbalanced data
Title	Machine Learning Methods in Motor Insurance Fraud Detection

Author's e-mail	85952280@fsv.cuni.cz
Supervisor's e-mail	barunik@fsv.cuni.cz

Abstrakt

Tato diplomová práce zkoumá aplikaci modelů strojového učení pro detekci podvodných pojistných událostí v pojišťení motorových vozidel. Porovnává účinnost několika algoritmů, včetně logistické regrese, random forest, XGBoost, histogram-based gradient boosting machine a multilayer perceptron (MLP). Studie se zabývá výzvou nevyrovnaného datasetu v detekci podvodů, přičemž využívá techniky jako je Synthetic Minority Over-sampling Technique (SMOTE) a vážení efektu jednotlivých kategorií. Reálná data poskytnuta UNIQA pojišťovnou a.s., včetně podrobných informací o pojišťovacích smlouvách a nárocích, slouží jako základ pro empirickou analýzu. Mezi testovanými modely dosáhl model XGBoost s využitím SMOTE tranformace a vážení kategorií nejvyšší míry zachycení podvodů s více než 90% detekovaných podvodných nároků. Analýza duležitosti jednotlivých proměnných zdůraznila klíčové ukazatele podvodů, jako je výše nároku, typ krytí nebo stáří vozidla. Závěry této práce ukazují potenciál pokročilých technik strojového učení ke zvýšení efektivity systémů detekce podvodů v pojišťovnictví.

Klasifikace JEL	C49, G22, K42,
Klíčová slova	strojové učení, odhalování podvodů, pojišťovnictví, nevyvážená data
Název práce	Využití strojového učení při odhalovaní pojistných povodů v autopojištění
E-mail autora	85952280@fsv.cuni.cz
E-mail vedoucího práce	barunik@fsv.cuni.cz

Acknowledgments

The author is especially grateful to the thesis supervisor doc. PhDr. Jozef Baruník, Ph.D. for his valuable insights, guidance and support throughout the process of writing the thesis.

Moreover, a great thanks goes to the UNIQA pojišťovna a.s., who provided data for the empirical research. Without them, this thesis would not be possible. Cooperating with them on the dataset preparation has been a pleasure.

Last but not least, I would like to express my immense gratitude to my family and friends for their endless support and love.

Typeset in FSV LaTeX template with great thanks to prof. Zuzana Havrankova and prof. Tomas Havranek of Institute of Economic Studies, Faculty of Social Sciences, Charles University.

Bibliographic Record

Bajgarová, Barbora: *Machine Learning Methods in Motor Insurance Fraud Detection*. Master's thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2024, pages 73. Advisor: doc. PhDr. Jozef Baruník, Ph.D.

Contents

List of Tables viii

List of Figures ix

Acronyms x

Thesis Proposal xi

1 Introduction 1

2 Theoretical background 3

2.1 Vehical insurance in Czechia and Slovakia 3

2.2 Insurance Fraud 4

2.3 Fraud detection systems 5

3 Methodology 7

3.1 Evaluated models 8

3.1.1 Logistic regression 8

3.1.2 Decision tree 9

3.1.3 Random forest 9

3.1.4 Gradient boosting methods 10

3.1.5 Multilayer perceptron 11

3.2 Evaluation methods 12

3.3 Unbalanced dataset problem 14

3.4 Feature importance 15

3.4.1 Permutation importance 16

3.4.2 SHAP values 16

4 Data 18

4.1 Data source 18

4.2 Data pre-processing 20
4.3 Inherited bias in the data 22
4.4 Data description 23

5 Results **30**
5.1 Individual model results 31
5.1.1 Logistic regression 31
5.1.2 Random forest 34
5.1.3 Histogram-based gradient boosting 36
5.1.4 XGBoost 39
5.1.5 Multilayer perceptron 42
5.2 Performance comparison 44
5.3 Feature importance 46

6 Conclusion **49**

Bibliography **54**

A Data Description **I**

B Models **IV**

List of Tables

3.1 Confusion Matrix 14

4.1 Proportion of fraudulent claims by the type of insurance 19
4.2 Distribution of damaged vehicle type 23
4.3 Descriptive statistics of claim amount 23
4.4 Descriptive statistics for sum insured (in millions) 25
4.5 Representation of men, women and companies in the sample . . 28

5.1 Logistic regression results 32
5.2 Random forest results 34
5.3 Histogram-based gradient boosting results 37
5.4 XGBoost results 40
5.5 Multilayer perceptron results 43
5.6 Comparison of models' results 45

A.1 Variables description I
A.2 Descriptive statistics for numeric variables II
A.3 Distribution of other binary variables III

B.1 Model parameters IV
B.2 Full comparison of models' results VI

List of Figures

4.1 Distribution of claim amount 24
4.2 Distribution of car age 25
4.3 Distribution of sum insured of CASCO insurance 25
4.4 Distribution of policy age 26
4.5 Distribution of days it took to report the claim 27
4.6 Distribution of damage type 27
4.7 Distribution of policyholder age 28
4.8 Correlation matrix of numerical and binary variables 29

5.1 Permutation importance for logistic regression 32
5.2 SHAP values for logistic regression 33
5.3 Permutation importance for random forest 35
5.4 SHAP values for random forest 36
5.5 Permutation importance for histogram-based gradient boosting . 38
5.6 SHAP values for histogram-based gradient boosting 38
5.7 Permutation importance for XGBoost 40
5.8 SHAP values for XGBoost 41
5.9 Permutation importance for multilayer perceptron 43
5.10 SHAP values for multilayer perceptron 44
5.11 SHAP values for claim amount 47
5.12 SHAP values for vehicle's age 48

A.1 Numerical variables distributions and correlations plot II
A.2 Distribution of damage cause III

Acronyms

AUC Area Under ROC Curve

ČAP Česká associace pojišťoven (*Czech association of insurers*)

CASCO Casualty and Collision

GBM Gradient Boosting Machine

HGBM Histogram-based Gradient Boosting Machine

MLE Maximum Likelihood Estimation

MLP Multi Layer Perceptron

MTPL Motor Third Party Liability

ROC Receiver Operating Characteristic

SGD Stochastic Gradient Descent

SMOTE Synthetic Minority Oversampling Technique

SVIPO Systém pro výměnu informací o podezřelých okolnostech (*System for exchange of information about suspicious circumstances*)

UNIQA UNIQA pojišťovna a.s.

XGBoost Extreme Gradient Boosting

Master's Thesis Proposal

Author	Bc. Barbora Bajgarová
Supervisor	doc. PhDr. Jozef Baruník, Ph.D.
Proposed topic	Machine Learning Methods in Motor Insurance Fraud Detection

Motivation Insurance companies all over the world lose billions of dollars due to fraudulent insurance claims. In Czech Republic alone there were 13 820 frauds detected in the total amount of 1 422.7 million CZK in 2022 out of which 6 605 frauds in the total amount of 434.4 million CZK were in stemming from car insurance (ČAP 2023). Moreover, both number of frauds and their total amount has been increasing in recent years, making it a pressing issue for Czech insurance companies as well as insurance companies all over the world. Note that these are only fraud that get detected, one can only guess how much more goes undetected.

As of now, the fraud detection systems in insurance companies are usually not very sophisticated and system solution is often missing. The market standard is implementing a scenario-based risk fraud detection techniques rather than utilizing more advanced data scientific methods available. This thesis will provide a comparison of several machine learning techniques which might be used to fight insurance fraud by flagging potentially fraudulent claims. Moreover, the interpretability of given machine learning methods will be explored in an aim to provide an insight into determinant of insurance fraud, which might further help insurance companies to understand its clients and potential risk groups among them.

This thesis will focus on auto insurance - more specifically on MTPL and CASCO insurance. It will make use of real-world data on insurance contracts and claims provided by one UNIQA pojišťovna, a.s.. The data set consist of contract data (both on policy holder and insured vehicle) and claim data (on damaged vehicle, accident circumstances and claim value) since 2020. There is total 783 claims flagged as a fraud uncovered by the adjusters complemented by a sample of 21 688 non-fraudulent claims.

Hypotheses

Hypothesis #1: Machine learning methods outperform standard econometric methods such as logistic regression.

Hypothesis #2: Ensemble methods, such as Random forest or XGBoost, provide the best overall results.

Hypothesis #3: Data synthetization can improve model performance.

Methodology First of all, data regarding frauds are highly unbalanced as detected frauds usually amount to only a few percent of the total volume of insurance claims. In order to correct for this imbalance, SMOTE resampling (Synthetic Minority Oversampling Technique), over-sampling or under-sampling methods will be considered.

Furthermore, several different statistical models will be use with the purpose of comparing them and trying to find the best ones. These models will be mainly logistic regression, decision tree, random forest or XGBoost models and neural network. The goal of this thesis is to compare strengths and weakness of these methods in the context of insurance fraud detection and try to find the most suitable one for practical deployment. To evaluate the models, we will use standard statistical metrics such as area under curve (AUC), precision, recall and F1 score with employment of k-fold validation to avoid sampling bias.

Expected Contribution This thesis aims to improve understanding of determinants of car insurance fraud in order to help fight auto insurance frauds and to find a suitable model for fraud prediction by comparing several suggested machine learning techniques. This thesis hopes to show that machine learning techniques are a viable alternative to currently used fraud detection systems and that they might increase success rate in insurance fraud detection. Moreover, this thesis will hopefully provide further insight into determinants of insurance fraud and thus provide a better understanding of the client portfolio structure and its associated risk.

Outline

1. Introduction
2. Literature review
3. Methodology
4. Data description
5. Results
6. Conclusion

Core bibliography

- NALLURI, Venkateswarlu, et al. Building prediction models and discovering important factors of health insurance fraud using machine learning methods. Journal of Ambient Intelligence and Humanized Computing, 2023, 14.7: 9607-9619.

- HANAFY, MOHAMED; MING, Ruixing. Using machine learning models to compare various resampling methods in predicting insurance fraud. J. Theor. Appl. Inf. Technol, 2021, 99.12: 2819-2833.

- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

- ITRI, Bouzgarne, et al. Performance comparative study of machine learning algorithms for automobile insurance fraud detection. In: 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS). IEEE, 2019. p. 1-4.

- Kaur, H., H. S. Pannu, & A. K. Malhi (2019): 'A systematic review on imbalanced data challenges in machine learning: Applications and solutions.' ACM Computing Surveys (CSUR) 52(4): pp. 1-36.

- ASLAM, Faheem, et al. Insurance fraud detection: Evidence from artificial intelligence and machine learning. Research in International Business and Finance, 2022, 62: 101744.

- Rukhsar, L., W. H. Bangyal, K. Nisar, & S. Nisar (2022): 'Prediction of insurance fraud detection using machine learning algorithms.' Mehran University Research Journal of Engineering & Technology 41(1): pp. 33-40.

- ČAP (Česká associace pojišťoven), Pojistný obzor | Pojistný obzor 01/23 [online]. Available at: https://pojistnyobzor.cz/archiv/pojistny-obzor-01-23

Author

Supervisor

Chapter 1

Introduction

Insurance companies around the world lose billions of euros every year due to fraudulent insurance claims. In the Czech Republic alone, in 2023 6 242 frauds were confirmed in a total amount of 1 664 million CZK, of which 434.4 million CZK were originating from vehicle insurance (Česká associace pojišťoven (2024)). Moreover, both the number of frauds and their total amount have been increasing in recent years, making it a pressing issue for Czech insurance companies as well as insurance companies all over the world. It is important to note that these figures only represent detected fraud, leaving the extent of undetected fraud to speculation. Currently, the level of sophistication of fraud detection systems varies significantly among insurance companies and a system solution is often missing. The market standard, especially for smaller insurers that are not part of an international insurance group, is to implement scenario-based fraud detection techniques rather than to utilize more advanced data analytics or machine learning methods.

The purpose of this thesis is to analyze the suitability of a machine learning solution for detection of insurance fraud. It provides a comparison of several supervised machine learning techniques that could be used to fight insurance fraud by flagging potentially fraudulent claims. These include logistic regression, random forest, histogram-based gradient boosting, XGBoost and multilayer perceptron models. Moreover, fraud detection is a highly unbalanced problem as frauds usually amount to only a few percentage points of the total number of incurred claims. Therefore, two approaches are tested for each of the models to correct for this imbalance. These are the Synthetic Minority Oversampling Technique (SMOTE) and weighting the effect of each class. The goal is to develop a model that can be used to score or pre-flag claims before the liquidation

process.

Furthermore, the interpretability of given machine learning methods is explored in order to verify the decision-making process of the model and provide insight into the determinants of insurance fraud, which could further help insurance companies understand its clients and potential risk groups among them. The interpretability and understandability of the machine learning solution are often overlooked by modern empirical research. We use the permutation importance and the concept of SHAP values to evaluate the importance of individual variables and their effects on the model prediction. This enables us to provide an interpretable automated mechanism for detecting fraudulent claims.

This thesis focuses on motor insurance - more specifically on MTPL and CASCO insurance. The empirical analysis is performed on a unique real-world dataset assembled and provided for the purpose of this thesis. The data were provided by one of the leading insurers in the Czech market and contain information on real insurance contracts and claims. This enables us to derive information about the effects of individual features without biasing the results by using a fully synthesized dataset. Using a dataset on actual claims from the Czech Republic and Slovakia helps illustrate the potential benefit for Czech and Slovak insurers from applying the model.

The structure of this thesis is as follows. First of all, in Chapter 2 we provide a theoretical background on vehicle insurance in the Czech Republic and Slovakia and current fraud detection systems and research available. In Chapter 3 the proposed models are presented, and the class imbalance correction, model evaluation and feature importance extraction methods are discussed. Followed by Chapter 4 which provides a description of the data used for the modeling, their specifics and pre-processing performed. In Chapter 5 the results of the models are presented and discussed - both the performance of respective models and the importance of variables used for the estimation, and the main implication of the research are summarized. Finally, Chapter 6 summarizes the thesis and concludes on the most important findings.

Chapter 2

Theoretical background

2.1 Vehical insurance in Czechia and Slovakia

The **Motor Third Party Liability** insurance (in Czech *'povinné ručení'* or *'pojištění odpovědnosti z provozu vozidla'*) is a mandatory insurance in both Czech Republic and Slovakia, as well as in the whole European Union. It is required by act n.168/1999 Sb. - The vehicle liability insurance act[1] and act n. 381/2001 Z.z. - Act on compulsory contractual liability insurance for damage caused by the operation of a motor vehicle[2] respectively. These regulations are an incorporation of European regulations and determine that any motor vehicle that is to be operated on public communications needs to be insured. The MTPL insurance is designed to protect individuals and their property from damage caused by motor vehicles operated by third party. It covers bodily injury and property damage caused to third parties in the event of a road traffic accident up to an amount specified in the respective act. Therefore, the MTPL insurance covers the damage caused by the insured vehicle, not the damage to that vehicle.

On the other hand, the **Casualty and Collision** insurance (in Czech *havarijní pojištění*) is a voluntary insurance intended to cover damage to the insured vehicle and/or its passengers. Generally, this insurance covers accidents, natural disasters, theft or vandalism. Additionally, this insurance is often complement by supplementary insurances such as windshield insurance, health insurance for people in the vehicle, luggage insurance and costs of renting a replace-

[1] in Czech *Zákon o pojištění odpovědnosti z provozu vozidla*

[2] in Slovak *Zákon o povinnom zmluvnom poistení zodpovednosti za škodu spôsobenú prevádzkou motorového vozidla*

ment vehicle and assistance services coverage. As this is a voluntary insurance, the coverage differs by contract given the client and insurer's agreement.

2.2 Insurance Fraud

Insurance fraud is a significant and persistent challenge in the insurance industry. It is defined as any act committed with the intent to obtain a benefit to which the claimant is not entitled. It can manifest in various forms - from exaggerated claims or providing untruthful or incomplete information to entirely fabricated incidents. In the context of motor insurance, fraud can involve, for example, staged accidents, inflated repair costs, or false theft claims. These deceptive practices not only lead to substantial financial losses for insurance companies but also result in higher premiums for honest customers and a general mistrust in the insurance system.

In 2017, the value of fraudulent claims detected in Europe reached 2.6 billion euros and Insurance Europe (the European insurance and reinsurance federation) estimated that combined detected and undetected fraud in Europe reached up to 13 billion euros (InsuranceEurope (2019)). In the Czech Republic, the Czech Association of Insurers (ČAP) collect and annually publishes statistics on insurance fraud. In 2023, Czech insurance companies confirmed 6 242 fraudulent cases amounting to 1.7 billion CZK, which is an almost 17% increase compared to the previous year, arising mainly from property and liability insurance. The motor insurance amounted to 408 million CZK (approx. 25% of the total amount). Note that this is only the value of convicted frauds, leaving the total amount to speculation. The volume of fraud and its increasing tendency over the past years makes this significant challenge for insurance companies (Česká associace pojišťoven (2024)).

From a risk management perspective, insurance fraud is naturally also perceived as a potential risk. The non-insurance risks connected to the insurance business as defined by Solvency II are market risk, credit risk (also known as default risk), and operational risk. All of these have a direct impact on the calculation of the solvency capital requirement. Solvency II defines operation risk as *the risk of loss arising from inadequate or failed internal processes, personnel, or systems, or from external events* and divides the operational risk into internal and external, which also includes fraud in insurance claims. Having an efficient and reliable fraud detection system can significantly decrease the

level of risk faced by the insurance companies and, subsequently, decrease not only unnecessary cost to the business, but also the capital requirement.

2.3 Fraud detection systems

As of now, the level of sophistication of fraud detection systems in insurance companies varies significantly. Although there has been intensive development recently, the market standard, especially for smaller insurers, is still a scenario-based detection system rather than utilizing advanced machine learning algorithms. This means that there is a predetermined set of condition rules, usually set by the insurers' employees. If those rules are satisfied, the claim is flagged as suspicious. However, the main responsibility for fraud reporting still remains in the judgment of claim adjusters, appraisers, examiners, and investigators, which can be resource-consuming and inefficient.

In order to help the insurance companies fight the insurance fraud, Czech Association of Insurers (*Česká associace pojišťoven*) developed and maintains two information systems for automated exchange of information about suspicious insurance activities. This aims to help with both prevention and detection of insurance fraud. The SVIPO I system provides information on motor vehicle insurance activity and can flag suspicious insurance claims that might have been a purposeful or unlawful act. The SVIPO II is a very similar system focused on personal insurance, more specifically life and health insurance. The systems are not publicly accessible, but according to ČAP the majority of insurers on the Czech market have joined the initiative and actively use the systems.

In this thesis, we explore the feasibility of machine learning application in the detection of fraudulent claims in motor vehicle insurance. More specifically, the models are developed as a pre-flagging mechanism to indicate which claims are at risk of being fraudulent. Therefore, the models are designed to detect as many potentially fraudulent claims as possible while keeping prediction errors reasonable. Due to the very low availability and quality of the data, academic research in this area is not very extensive. To our knowledge, there is virtually no publicly available dataset for the Czech or Slovakian insurance market, and therefore there has not been any academic research yet. In international research, several studies have been conducted on insurance fraud. Hanafy & Ming (2021b) provides a comprehensive overview of existing research and a comparison of multiple machine learning models and sampling techniques for automobile insurance based on data from an Egyptian car insurance company.

Hanafy & Ming (2021a) and Singhal *et al.* (2023) show a comparison of several machine learning methods on data from Brazilian automotive company and publicly available data from kaggle.com respectively. Severino & Peng (2021) compared several machine learning techniques, including a deep neural network and GBM, on a Brazilian property insurance dataset. They also tested several techniques, such as the permutation importance and the SHAP values, to interpret variable impotence and their effects on the prediction of the model. In all of this research, machine learning models appear to provide interesting performance for application within insurance fraud detection, as the best models range from 0.5 to 0.9 in recall and from 0.6 to 0.9 in AUC scores. However, the quality of the underlying dataset can significantly influence the research results. Nalluri *et al.* (2023) compared several machine learning models for medical insurance fraud detection and tested them on two dataset, data obtained from kaggle.com and data from the CMS database, and the resulting performance of the models differed drastically. Therefore, it will be difficult to make a direct comparison with previous research and its results.

An area, where fraud detection research has been extensive in recent years and machine learning techniques provide satisfactory results, is the area of credit card fraud detection (e.g. Alarfaj *et al.* (2022), Trivedi *et al.* (2020), Khatri *et al.* (2020) or Sinčák (2023)). However, it is important to note a crucial difference between these two fraud detection applications. In credit card fraud, the fraud is usually committed by a third-party and the goal is to prevent the fraudulent transaction from happening. In case the fraudulent transaction goes undetected, it will usually be reported by the card holder. Thus, there is a feedback mechanism for the model to learn from. The situation of insurers is a little more complicated. In case of insurance fraud, the fraudster is usually the counterparty, i.e. the client. If the insurance fraud goes undetected by the insurance company, there is no natural feedback or back-testing mechanism to correct the model's prediction. Improving and correcting the fraud prediction model for insurance claims is thus a complex process and regular testing needs to be conducted to ensure correctness of training data.

Chapter 3

Methodology

There are two main approaches to machine learning - the supervised and unsupervised learning. Supervised learning involves training models on labeled datasets, where each training example is paired with an output label. In contrast, unsupervised learning deals with unlabeled data, focusing on identifying patterns and structures within the data without explicit guidance on what to predict. In this thesis, we will focus on comparing several models from the supervised learning field.

There are numerous models in supervised learning that are suitable for classification problems such as fraud detection. In this thesis, the logistic regression, random forest, XGBoost, Histogram-Based GBM and multilayer perceptron are compared. This chapter provides a theoretical background and discusses the advantages and disadvantages of the proposed models and the established literature that summarizes their applications. Moreover, to provide a comparison between the models, suitable metrics are chosen and explained. Since fraud detection is a highly imbalanced problem, its implications for modeling are discussed and several methods to correct the imbalance are introduced. Lastly, in addition to understanding the models and their performance, it is also important to understand the decision-making of the models and the role that the respective features in the dataset play in the prediction. Therefore, two methods are introduced for the evaluation of the importance of individual features and their effects on the predictions. These include the permutation importance and the SHAP values.

3.1 Evaluated models

3.1.1 Logistic regression

First of the proposed models is the logistic regression. The logistic regression is an extension of linear regression, where the estimated class is assumed to follow the Bernoulli distribution. It is used for binary classification problems; however, it can also be extended for multi-class classification. First, the linear part of the model is estimated and then the logistic transformation is applied. Linear regression is defined as $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$ or using the matrix notation as $y = xw+b$. Therefore, the output of the linear part of the regression is equal to $\bar{y}(x; w) = xw$. To ensure that the predicted probability is in the range $< 0, 1 >$, the sigmoid transformation is applied $\sigma(x) = \frac{1}{1+e^{-x}}$. Hence, the output of the logistic regression can be represented as:

$$y(x; w) = \sigma(\bar{y}(x; w)) = \sigma(xw).$$

The algorithm, as applied in this thesis, is estimated based on maximum likelihood estimation (MLE) and the weighs updates are performed using stochastic gradient descent (SGD) with L2 regularization.

Logistic regression has been researched for various areas of financial fraud detection, such as automobile insurance fraud (Rukhsar *et al.* (2022), health care insurance fraud (Obodoekwe & Van der Haar (2019)) or credit card fraud (Sinčák (2023)). Although it is usually outperformed by more complicated or ensembled models, some research shows that, for specific datasets, logistic regression combined with under-sampling can actually achieve competitive results (Itoo *et al.* (2021)).

The main advantage of this method is its interpretability. The logistic regression is based on a linear method, therefore the effects of individual variables on the probability of a claim being fraudulent can be easily extracted and examined. Another advantage of this model is the estimation time, as the training time of this model is significantly faster than the ensembled methods (Sinčák (2023)). However, since our dataset is rather small with limited number of features, the training time is not an issue and does not need to be taken into account. Logistic regression is used mainly as a baseline model for comparison of the performance of the other methods.

3.1.2 Decision tree

To understand more complex and ensembled methods, we need to first introduce their building blocks, also called base learners, which are decision trees. Decision trees are a versatile machine learning method that can be used for both classification and regression. It is a non-parametric method that splits the data into subsets based on feature values, effectively creating a tree-like structure. In each splitting node, a chosen criterion (such as Gini impurity, log loss or entropy) is evaluated, and feature and its values, which divide the data into the most homogeneous subsets, are chosen for the splitting. In the context of classification, a decision tree predicts the class of data entry by traversing from the root to a leaf, following the branches according to the feature values of the instance. The predicted class is then based on the majority class of training data belonging to a given leaf.

The main advantage of the decision tree model is its intepretability and comprehensibility. The decision making process is based on a series of if-based steps which can be easily explained and visualized. In addition, the importance of individual data features can be easily extracted. Moreover, it is a non-parametric model, hence no underlining distribution needs to be assumed. On the other hand, if left unrestricted, the decision tree model is prone to create an overly complex tree that overfits the training data and struggles to generalize on unseen data. They can also be unstable and sensitive to minor changes in data (Murphy (2012)).

Despite its limitations, decision three is a strong baseline model and is used as an underlying model for many ensembled methods. Its application to fraud detection has been investigated for applications within automobile insurance fraud, but it is usually outperformed by its ensembled versions such as the Random forest and XGBoost classifiers (Hanafy & Ming (2021b), Rukhsar *et al.* (2022)).

3.1.3 Random forest

The random forest is effectively an extension of the decision tree model. This method builds multiple decision trees and merges their results to improve performance and control overfitting. By combining the predictions of several base estimators (also called weak learners), it enhances the robustness and generalization capability of the model.

Each tree in the forest is trained on a random subset of the training data.

Additionally, at each node of the tree, a random subset of features is selected, and the best split is determined only from this subset. This introduces an additional layer of randomness, which helps reduce the correlation between trees and leads to a more diverse set of base models. The final prediction, as implemented in the *scikit-learn* library, is then obtained by averaging the probabilistic prediction of individual trees (Pedregosa *et al.* (2011)).

Since each tree in the random forest is trained using different data subsets and uses different features, they remain uncorrelated with one another reducing instability commonly seen in single decision trees. Additionally, the large number of trees significantly reduces the risk of overfitting. Studies have shown its effectiveness in identifying fraudulent claims in various fields - in credit card fraud (Udeze *et al.* (2022), Sinčák (2023)), car insurance (Hanafy & Ming (2021b)) or health insurance fraud (Nalluri *et al.* (2023)).

However, the increase in performance of the model comes with its cost, as the random forest sacrifices some of the interpretability compared to a single decision tree. Moreover, with a large number of features and a large number of base estimators (i.e. trees in the forest), the estimation time can increase significantly (Sinčák (2023)). However, considering the usual size and number of features of the insurance claims dataset, this is not of concern.

3.1.4 Gradient boosting methods

Gradient boosting is a generic framework that employs the gradient descent algorithm that can be used to minimize various loss functions (Friedman (2001)). The descent is done in subsequent steps, where each estimation is built to correct mistakes made by previous estimation by optimizing the model weights based on the errors of previous iterations. Finally, all the predictions obtained from each of the iterations are combined to form a final prediction. An extension of this algorithm is the gradient-boosted decision trees method (Friedman (2001)), where each iteration in the gradient descent is based on a decision three. All trees built during the estimation are then combined into a final prediction.

Extreme Gradient Boosting (XGBoost) is a relatively new implementation of gradient-boosted decision trees proposed by Chen & Guestrin (2016). It is an advanced implementation designed for speed and performance and has become one of the most popular machine learning algorithms for structured or tabular data due to its high accuracy, efficiency and flexibility. In addition to the other

gradient boosting methods, it adds a regularization form to the loss function during its training. This limits the complexity of the model and prevents overfitting. The binary log loss, also known as binary cross-entropy loss, was used as a loss function. Moreover, XGBoost leverages parallel and distributed computing to enhance speed. The suitability of the XGBoost algorithm for fraud detection has already been shown in several fields - in credit card fraud (Sinčák (2023)), health insurance (Gupta *et al.* (2021)) or motor insurance (Baran & Rola (2022)), and this model usually ranks among the highest performing within the compared methods.

Another extension of the gradient-boosted trees is the Histogram-Based Gradient Boosting method. This algorithm was inspired by LightGBM as proposed by Ke *et al.* (2017). It bins the input samples into integer-valued bins, which tremendously reduces the number of splitting points to consider, and allows the algorithm to leverage integer-based data structures (histograms) instead of relying on sorted continuous values when building trees (Pedregosa *et al.* (2011)). It is an efficient estimator developed for a larger dataset (with a number of samples over 10 000). Binary log loss is used as a loss function, and the implementation in the *scikit-learn* library also allows for the inclusion of an L2 regularization term, similar to the XGBoost algorithm. Histogram-Based Gradient Boosting Classifier is a rather new method developed to enhance speed, but has already been shown to provide competitive results in the field of fraud detection (Nhat-Duc & Van-Duc (2023), Mahmood *et al.* (2023)).

3.1.5 Multilayer perceptron

The multilayer perceptron (MLP) is esentially a feed-forward artificial neural network. The artificial neural network is a powerful model capable of learning complex patterns within the data. It was originally inspired by neurons in the human brain and can be used for both regression and classification.

The MLP consists of the input layer, one or more hidden layers, and the output layer. In the hidden layer, each neuron applies a weighted sum of its input (neurons from the previous layer) followed by an activation function. Some common examples of activation function are the ReLU (Rectified Linear Unit), sigmoid or tanh function. In our case, ReLU proved to provide the best results. In the output layer, a softmax activation function is applied for a classification task to transform the output to probabilities. The estimation process in an MLP involves adjusting the weights of the connections between neurons

to minimize the loss function, typically using backpropagation combined with an optimization algorithm such as stochastic gradient descent (SGD) or Adam optimizer.

The main advantage of neural network model is its capability to capture complex relationships and work even with unstructured data. However, it is considered a 'black-box' model, as it is very difficult to interpret. It also contains a lot of hyperparameters, which need careful fine-tuning. Moreover, with large dataset and complicated architectures, the network can quickly become computationally complex and time-consuming.

Nowadays, there are naturally many extensions of the original algorithm. Zakaryazad & Duman (2016) examined several neural networks for the purpose of detecting credit card fraud, also considering the savings provided by the different models. Their research showed that the original simple neural network provided the best accuracy. However, in recent years, there has been an extensive development in the field of deep learning and especially in neural networks based on transformers. Specifically, transformers and pre-trained transformers show promising results in the field of fraud detection (Yu *et al.* (2024), Yuan (2022)). However, we include MLP mainly to provide a comparison to tree-based methods and leave the development of the optimal advanced neural network for further research.

3.2 Evaluation methods

Choice of the right evaluation methods is important to ensure the correct setting of the model and its usability and efficiency for a given classification task. There are several performance evaluation methods suitable for a classification problem.

Accuracy: The accuracy gives the percentage of the right predictions out of all the predictions. Thanks to its intepretability, it is a common choice for classification problems. However, given the high class imbalance in the data, since fraudulent claims only form approximately 3.5% of the data, accuracy is not a suitable measure for the fraud detection problem.

Precision: The precision reports the proportion of actual frauds in the total number of frauds predicted. In terms of the confusion matrix shown in Figure 3.1, it is the True Positives (that is, fraudulent claims indeed marked by

the model as fraudulent) divided by True Positives and False Positives (that is, nonfraudulent claims marked by the model as fraudulent), i.e $\frac{TP}{TP+FP}$. Hence, this measure refers to how efficient a given model would be in fraud detection.

Recall: The recall (also called sensitivity) refers to the proportion of correctly predicted frauds to all fraudulent claims. In terms of the confusion matrix, this is the $\frac{TP}{TP+FN}$. Hence, this metric reports how much of the actual fraudulent claims is the given model capable of capturing.

Specificity: The specificity, also called the true negative rate, measures the proportion of correctly classified non-fraudulent claims to all non-fraudulent claims. In terms of the confusion matrix, this is $\frac{TN}{TN+FP}$. Since we are mainly concerned with fraudulent claims, this measure is not as informative.

F1 score: Obviously there is a trade-off between precision and recall. The F1 score attempts to capture this trade-off. It is calculated as follows:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Therefore, this metric penalizes the model for poor performance in either precision or recall and is often used to evaluate classification models.

ROC and AUC: Another metric used to evaluate classification models that captures the sensitivity-specificity trade-off is the receiver operating characteristic curve (ROC). This is a graphical representation of the true positive rate (sensitivity) and the false positive rate (1 - specificity) at various classification thresholds. The closer the curve is to the top-left corner of the plot, the higher is the specificity and the lower is the false positive rate across different decision threshold and hence the better the model performs.

The Area Under the Curve (AUC) is a quantification of the ROC curve and refers to the area under ROC. Its value falls in the range from 0 to 1, where the higher the value, the better the discrimination power of the model. A model that has no discrimination power (i.e., is similar to random guessing) has an AUC of 0.5. Therefore, an AUC is a good measure of model performance regardless of whether precision or recall is intended to be maximized, since those can be highly influence by threshold setting. Therefore, an AUC of at least 0.5 was used as a baseline condition for the models.

	Actual Positive	Actual Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

Table 3.1: Confusion Matrix

The goal of this thesis is to test algorithms for the purpose of pre-flagging potentially fraudulent claims. This assumes that the program will be used as input information for a subsequent process of liquidation of the insurance claim. Therefore, the models were designed with the purpose of detecting as many potentially fraudulent claims as possible while keeping the prediction errors reasonable. Hence, recall was maximized while keeping the F1 score reasonable. Furthermore, the AUC was used to assess the overall quality of the model.

3.3 Unbalanced dataset problem

The fraud detection problem is a high unbalanced dataset problem as detected frauds usually represent only up to a few percent of the total number of claims. Such an extreme under-representation of one class can cause the classification problem to be biased towards the majority class, which leads to worse performance of the classifier (Hanafy & Ming (2021b), Kaur *et al.* (2019)).

Several techniques have been developed for handling unbalanced datasets. Generally, there are two approaches to the problem of class imbalance. Either the effect of each class is adjusted by weighting, or the dataset itself is adjusted (resampled) so that the classes are represented more equally. For the purpose of this thesis, we have decided to employ two techniques - the weight balancing and SMOTE, which represent the two different approaches.

The weight balancing, also called the class weighting, is a simple method where observations are assigned weights based on the distribution in the training data, usually inversely proportional to their frequency in the data. Hence, the weights are set as follows:

$$weight_{class_i} = \frac{\text{Total number of observations}}{\text{Number of observations of } class_i}$$

The exception to this is the implementation of the XGBoost algorithm, where the class weighting is performed by parameter *scale_pos_weight* and the weight is set to *number of negative instances/number of positive instances*.

Hence, the fraudulent data, which are significantly under-represented in the dataset, will be assigned higher weights, whereas the non-fraudulent data will assigned lower weights, effectively imposing stronger regularization for updates based on non-fraudulent observations. The application of this method for fraud detection has been explored for example by Udeze *et al.* (2022) and shows promising results, especially for the XGBoost estimation.

The second method that will be applied is the Synthetic Minority Over-sampling Technique (SMOTE) as proposed by Chawla *et al.* (2002). This method creates a new synthetic observation by interpolating between minority class observations already existing in the dataset. For each of the under-represented observations, k-nearest neighbors which also belong to the minority class are selected (based on the kNN algorithm), and new observations are generated by interpolating between the selected observation and its randomly chosen k-nearest neighbor. This method has proven to outperform standard over-sampling and under-sampling techniques, where observations are simply randomly dropped or duplicated, for application in fraud detection in insurance (Hanafy & Ming (2021b)) and also in other finance frauds such as credit card frauds (Muaz *et al.* (2020)) and perform especially well with the Random Forest classifier (Muaz *et al.* (2020)). The SMOTE method also performs well in combination with under-sampling methods (Haixiang *et al.* (2017)). However, in our dataset the fraudulent claims have already been complemented by only a representative sample of non-fraudulent claims, therefore under-sampling in fact has already been applied, in this thesis only SMOTE is applied.

The SMOTE also enables to control the degree of over-sampling by controlling how many synthetic observations will be created. Therefore, we will also explore the possibility of a combination of both approaches - the class weighting and the SMOTE. Firstly, the dataset will be partially balanced by the synthetic over-sampling and then the observations will be weighted based on their inverse frequency in the resampled dataset to further limit the bias of the model towards the majority class.

3.4 Feature importance

With more complicated models being developed, it is ever more important to also develop mechanisms to understand the model processes and uncover what is happening within the 'black-boxes'. In order to understand and be able to verify the decision-making of the model, it is important to understand on

what the model bases its decisions. Since most of the compared methods are nonlinear, the effects of individual features on the prediction cannot be easily extracted. Therefore, two methods, which can be applied to all the compared models, were chosen to evaluate the importance and effect of individual features on models' predictions. These include the permutation importance, which assess solely the importance of individual features for the models' performance, and the SHAP values, which are more complex with longer estimation time, but provide also an evaluation of the effect of given feature on the prediction. It is important to note that both methods evaluate the importance of given features, but not the quality of the prediction itself. That is, some feature might not be important for a bad model, but might be important for a good one. Therefore, it is also important to always separately evaluate the quality of the model itself.

3.4.1 Permutation importance

Permutation feature importance is a model-agnostic technique that provides a measure of the importance of each feature. This method assesses the decrease in model performance when the values of a particular feature are randomly shuffled, thus causing the relationship between the feature and the target variable to be broken. Firstly, the model is estimated and the baseline performance is measured with respect to the chosen metric. In our case, the recall is chosen as the evaluation metric. Then, for each feature separately, its values are shuffled to break the relationship with the target variable. The decrease in performance is then measured and this corresponds to the importance of a given feature. The shuffling of feature values can be performed repeatedly to avoid selection bias and to provide a confidence interval on the estimated importance.

This approach is particularly useful due to its simplicity, ability to be applied to any model, and short estimation time. Moreover, compared to the impurity-based importance implemented for tree-based algorithms, it does not exhibit bias toward high cardinality (usually numeric) features (Pedregosa *et al.* (2011)).

3.4.2 SHAP values

The SHapley Additive exPlanations (SHAP) values, as introduced by Lundberg & Lee (2017), is a method based on cooperative game theory that aims to provide a model-agnostic evaluation of the importance of model features and

their effect on the model's prediction. It was derived from Shapley values that originate from cooperative game theory (Shapley *et al.* (1953)), but it was adapted for the interpretation of machine learning models. The effect of each feature on the final prediction is based on its marginal contribution to the prediction, and it is calculated by considering all possible combinations (subsets) of features. The key idea is to fairly distribute the difference between the prediction and the average prediction across the features based on their contributions. Formally, SHAP values are defined as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|! \cdot (|N| - |S| - 1)!}{|N|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

where N is the set of all features, S is a subset of N not containing feature i_i, $f_{S \cup \{i\}}$ is the prediction of the model including the examined feature and f_S is the model where the feature is withheld, $|S|$ is the number of features in subset and $|N|$ is the total number of features.

The SHAP values satisfy the assumption of local accuracy, missingness and consistency (Lundberg & Lee (2017)). They are calculated for each observation individually, which allows for local feature importance extraction and interpretation of how every individual prediction of the model has been made. Most importantly, they do not only indicate which features are important, but also how they contribute to the final prediction (both in terms of direction and magnitude), offering valuable insights into the model's decision-making process.

The main disadvantage of SHAP values is their computational complexity. As they are calculated for every observation, every feature and all possible subsets of features, the estimation time can be considerably high. Therefore, since the introduction of the original kernel algorithm, several algorithms for different models were developed to approximate the SHAP values to make the estimation more feasible, e.g. TreeSHAP (Lundberg *et al.* (2018)).

Chapter 4

Data

This chapter provides a description of the dataset used for empirical analysis. It discusses the source and specifics of the data, initial pre-processing needed, and feature engineering performed. Moreover, it presents summary statistics of the most important variables.

4.1 Data source

This thesis utilizes real-world data on insurance contracts and claims from the Czech Republic and Slovakia from the period of 01/2020-8/2023. The dataset was provided by UNIQA pojišťovna, a.s., who is one of a leading insurers in motor insurance in both Czech and Slovakian market.[1] The dataset provided by the insurance company was essentially obtained from two sources - the contract data and the accident information.

The contract data consist of both information about the policyholder (e.g. year of birth and municipality of residence) and the contract itself. Data about the contract contain information about the type of insurance coverage - whether it is MTPL or CASCO. Moreover, it includes information about the sum assured, frequency of premium payment, contract signing and effective dates. Furthermore, these were complemented by information on the number and amount of claims on a given contract in the last three years and the total number of claims in the contract history. Unfortunately, the information about premium is not available due to maintaining the trade secrets. The data are captured as of the moment the claim was reported (except for the premium frequency and

[1] As of year end 2022, UNIQA is fifth largest insurer in the MTPL in the Czech market both in terms of number of vehicles insured and gross written premium and fourth largest insurer in CASCO insurance in terms of gross written premium (Česká associace pojišťoven (2022)).

the number and amount of claims on a given contract in the last three years, which are captured at a date of data extraction).

Then the accident data is available. This includes data on both the damaged and insured vehicle, the accident circumstances (such as a location, number of vehicles involved, if a police was called, if there were witnesses or bodily injuries, etc.), the date of the accident and the date that the accident was reported, information about how many claims were filed with respect to given accident and given insurance contract (there can be damaged to the car, property, bodily injuries etc.) and most importantly the claimed amount.

The data provided by the insurance company were available from 01/2020 to 10/2023, when they were extracted from their internal systems. However, we have decided to exclude the last two months (in terms of a date when the claim has been reported) since there were no frauds detected in this period. This is clearly given by the fact that fraud detection requires a considerable amount of time and these claims were not fully resolved by the time of the data extraction.

Therefore, the dataset consists of 22 000 reported claims in the period of 01/2020 to 8/2023. Considering the features available, we are focusing solely on insurance claims regarding damage on a vehicle and the information about bodily injuries, property and other damages are used as additional feature. This leaves us with 19 661 observations, of which 760 have been marked by the insurance company as frauds. It is important to note that this is a complete list of frauds detected during the period, but not a complete list of claims incurred. The fraudulent data have been complemented by only a representative sample of the non-fraudulent claims. Nevertheless, we are dealing with a highly unbalanced dataset as for both MTPL and CASCO the proportion of fraudulent claims amounts to approximately 3.5% of the total number of observations available.

Table 4.1: Proportion of fraudulent claims by the type of insurance

	CASCO	**MTPL**
No Fraud	13486	5415
Fraud	486	274

4.2 Data pre-processing

In total, the obtained dataset contained 41 columns. However, some of these columns were identifiers, some were directly not suitable for machine learning tasks (such as, for example, the car brand or address), and many suffered from very low data quality and many missing values. For example, the car millage had almost half of the values missing or filled with zeros and therefore had to be excluded from the analysis. Since a significant part of the data is filled manually by the client when reporting the claim, the quality of the data and the number of missing values were the main issue when processing the dataset. Moreover, many columns contained time-related information (e.g., year of birth, year of manufacture of the car, or date of the accident and its reporting). Those were converted to information relative to the moment of the accident (e.g. age at the time of the accident, age of the car at the time of the accident, or number of days it took to report the accident). Finally, 18 variables with sufficient data quality and meaningful informational value were selected for analysis. This includes 10 numerical variables, 4 categorical variables and 4 binary variables. The complete list with a detailed description can be found in Appendix A Table A.1 and the most interesting variables are further discussed in Section 4.4.

In addition to choosing suitable variables, the data need to be pre-processed in order for the model to process them correctly. Firstly, numerical features need to be appropriately scaled, as machine learning models often perform better when the input variables are on a similar scale. This ensures that no single feature dominates the learning algorithm due to its larger magnitude. The most common scaling techniques include standardization (scaling features to have zero mean and unit variance) or normalization (scaling features to a range, typically between 0 and 1). Since our data clearly do not follow the normal distribution, as can be seen in Figure A.1 in Appendix A, normalization is applied. More specifically, MinMax scaling is applied to all numerical variables. The MinMax scaler is defined as follows:

$$X_{scaled} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

This transforms the data into an interval $< 0, 1 >$ while keeping the original distribution and relative relationship between the values.

Furthermore, some machine learning algorithms require a numerical input. Therefore, categorical variables need to be encoded as numbers and it needs to

be done in such a way that the model does not assume an ordinal relationship where one does not exist. A relatively straightforward approach for dealing with categorical variables encoding is the one-hot encoding. This method creates a separate column for each of the categories of a given variable, which then acquires binary values indicating whether the observation belongs to a given category or not. Since machine learning models process binary variables well, this is a commonly used method. However, in case of dataset with many categorical variables with a large number of categories, the resulting dataset can become high-dimensional, which may slow down the estimation. There are techniques to tackle working with categorical variables with a large number of categories. This includes, for example, the learned embeddings. This method was developed to represent words and their relative meanings in text analyses but is very flexible and also often used in other fields. It transforms categorical variables into one or more vectors of continuous numerical values. However, by using the learned embeddings, additional interpretability is lost as it is hard to map the original categories to the new numerical values. Therefore, we decided to manually adjust variables with an undesirably high number of categories. For example, in variable *'DAMAGE_CAUSE'* the various categories for flood, storm, fire and animal encounter causes were merge into a category 'nature' and various regulation violation and criminal activity causes were merged into a category 'lawbreak'. Similarly, for the variable *'DAMAGED_VEH_TYPE'* the various vehicle types in the dataset were divided into big (such as buses, truck, tractors, etc.), small (such as motorcycles or quad bikes), auto (cars) and other (marked as other or non-specified). Analogously, similar transformation and grouping were performed for other categorical variables and categories. Since there were no more than 21 values per variable, this approach was feasible and allowed us to retain as much interpretability as possible. Moreover, by being able to supervise the grouping process, we ensure meaningful encoding without having to rely on a 'black-box' embeddings. After the encoding of categorical variables is applied, the dataset contains a total of 32 numerical or binary variables.

Finally, the data are divided into training and validation sets. Data are randomly assigned to the sets in the proportion of 70% to the training set and 30% to the validation set. This split is performed before any methods to correct for the class imbalance are applied (as described in Section 3.3). The idea of splitting data into training and validation sets is that the validation data should not influence the model training in any way. Therefore, it is important

to divide the data before resampling so that the resampling does not bias the reported results. We are interested in the performance of the models on the real dataset, not the artificially adjusted balanced one. Otherwise we might end up with a model that performs well on balanced data but struggles on unbalanced data, which frauds clearly are.

4.3 Inherited bias in the data

We are working with a real-world dataset on insurance fraud, therefore it is important to keep in mind some specifics of such a dataset. The fraud flag in the dataset refers to the frauds detected by the insurance company as there is no way to account for the undetected fraud. Therefore, it is important to note that the controls and fraud detection mechanism in place can have a direct influence on which frauds are most likely to be detected by the model. For example (and note that this is just an illustrative example), due to its limited resources, the company might decide to concentrate on insurance claims with higher claim value and thus higher potential loss, which inevitably brings a bias into the input data and might cause the models to underestimate the true probability of fraud for nominally lower claims. Moreover, some types of fraud might be more difficult to detect and prove than others. Due to obvious reasons, the current processes in place to detect fraudulent claims cannot be, have not been and will not be shared - neither publicly, nor with the author. Nevertheless, current detection systems may significantly influence the variables' importance and models' predictions, and it is important to be aware of these limitations.

Additionally, as mentioned previously, we are dealing with a real-wold dataset where major part of the data is filled in by people (e.g. clients when reporting the claim) and it is not possible to verify the correctness of the input data. There are certainly some controls in place when the claim is being investigated, especially for more suspicious claims, but no backward check is possible. Considering the amount of missing or suspicious values in some features, which might have been useful but had to be excluded from the analysis, this is one of the main drawbacks in using real-world data. Although the quality of the incident report also contains interesting information about the claim (e.g., claim with a lot of missing information might also be a sign fraud), improving data quality and data controls would likely improve model performance significantly.

All of this brings bias into the data we are dealing with. This bias cannot be avoided (in case of detection systems bias) or is very difficult to avoid (in case

of data validation problem) and it belongs to such fraud detection problems, however, it is important to keep it in mind when interpreting the results and potential next steps to be taken.

4.4 Data description

As mentioned previously, the claims to be investigated are motor (MTPL or CASCO) insurance claims with damage to a vehicle. Bodily injuries, property damage and other damages are excluded and used as additional feature - the *'OTHER_CLAIMS'* as a binary flag indicating whether any additional claims were filled. The majority of incidents (93%) were a single claim on the damaged vehicle, the rest contained 1 to 5 additional claims. Figure 4.2 shows the representation of individual damaged vehicle types in the dataset. The vast majority of vehicles in the dataset are cars (98%). The other categories included are 'big' (such as trucks, buses or tractors), small (such as motorcycles and quad bikes) and other (non-specified).

Table 4.2: Distribution of damaged vehicle type

	car	**big**	**small**	**other**
No Fraud	18 514	252	92	43
Fraud	738	15	4	3

Probably the most interesting and important variable in our dataset is the *'CLAIM_AMOUNT'*. This corresponds to the value of the given claim reported in Czech crowns (CZK). Table 4.3 shows the summary statistics of the variable. This variable should be particularly useful in combination with some variables that refer to the value of the car or the extent of the damage. Together, they should help identify exaggerated claims.

Table 4.3: Descriptive statistics of claim amount

	mean	**std**	**min**	**25%**	**50%**	**75%**	**max**
CLAIM_AMOUNT	34 290	54 434	0.00	7 825	18 355	40 000	1 458 500

Moreover, Figure 4.1 represents the distribution of claim amount graphically. For better representation, the data in the figure are restricted to 250 000 CZK, which corresponds to approximately the 99^{th} percentile. On the left there is the distribution of claim amount and on the right the proportion of frauds in

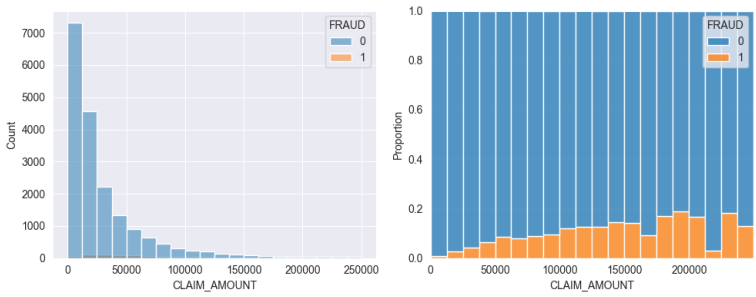

Figure 4.1: Distribution of claim amount

each of the bins is displayed (fraudulent claims are depicted in orange whereas non-fraudulent claims are depicted in blue). We can observe that majority of claims received by the insurance company have a lower nominal value as the data are strongly skewed to the right. There is a clear trend in the data where with a higher claim amount the proportion of frauds in the sample increases. This can be an example of inherited bias in the data caused by the detection mechanism in place, as mentioned in Section 4.3, but it can also be the true underlying distribution of frauds. Nevertheless, the claim amount is expected to be an important variable for fraud detection.

Regarding the variables that indicate the value of the damaged car, unfortunately the value itself is not available. However, there are two variables that can serve as a partial proxy - the *'CAR_AGE'* and the *'SUM_INSURED'*. Unfortunately, information on premium paid for the policy, which would indicate the value of the vehicle as well as the riskiness of the policyholder, is also not available due to maintaining trade secrets.

The variable *'CAR_AGE'* refers to the number of years between the manufacture of the vehicle and the incident, effectively the age of the vehicle. The vehicle to which it refers is the damaged vehicle, not the insured vehicle (which can be different, for example in case of the MTPL insurance). Figure 4.2 shows the distribution and proportion of frauds in the samples for this variable. Again we observe an increasing trend, as with the age of the vehicle the proportion of frauds increases. This is true at least for the first 20 years, where majority of the data lies. The damaged vehicles in the dataset are generally rather new, as the average age of the vehicles is just 6 years.

The *'SUM_INSURED'* variable corresponds to the coverage limit of a given insurance policy. That is the maximum amount that an insurance company will pay out in the event of a claim for damage or loss to the insured vehicle. For MTPL insurance, the minimum coverage is prescribed in the respective

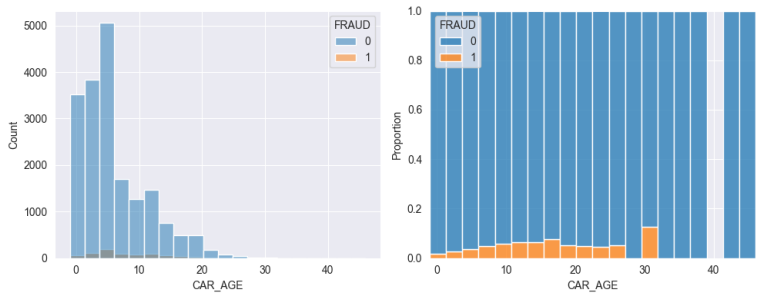

Figure 4.2: Distribution of car age

regulation. As of 2024, in the Czech Republic the mandatory coverage for cars is 50 million for bodily injuries for each of the harmed parties and 50 million for property damage and loss of profit (regardless of the number of parties). However, the client can choose to arrange a higher coverage if he wants. For CASCO insurance, the coverage for given policy is agreed upon by the client and the insurance company. Table 4.4 shows the overall summary statistics, as well as the summary statistics for each insurance product separately. We can see that the sum insured for the MTPL insurance is strongly determined by the legal requirements. For CASCO, the coverage varies significantly. Figure 4.3 presents the distribution of the sum insured for CASCO policies for more details. The data are again restricted to the 99^{th} percentile for better visualization.

Table 4.4: Descriptive statistics for sum insured (in millions)

	mean	**std**	**min**	**25%**	**50%**	**75%**	**max**
SUM_INSURED	40.82	72.33	0.00	0.33	5.78	53.00	300.00
SUM_INSURED (CASCO)	0.52	1.05	0.00	0.27	0.42	0.63	110.00
SUM_INSURED (MTPL)	139.79	65.51	0.00	110.00	110.00	110.00	300.00

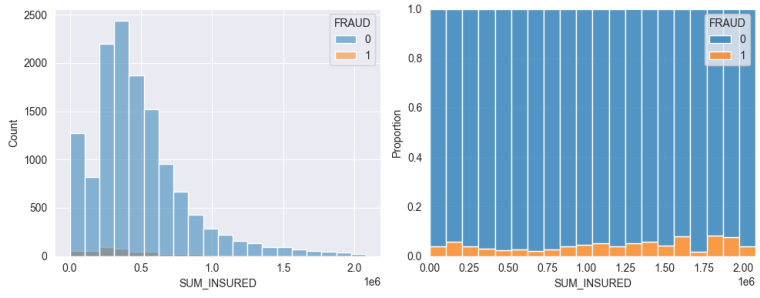

Figure 4.3: Distribution of sum insured of CASCO insurance

The next numerical variable, that is closely followed in the insurance industry, is the *'POLICY_AGE'*. This refers to the number of days since the signing

of the contract to the time of the accident. In Figure 4.4 the distribution and the proportion of frauds in the distribution can be observed. Again, the data are restricted to the 99^{th} percentile, which corresponds to approximately 11 years. We observe that most of the policies are rather young as the average policy is younger than 3 years at the moment of the accident. This is not surprising, as it is not uncommon to change the insurance provider and the average age of the vehicle in the dataset is only 6 years. The policy age variable also has negative values, where the accident happened before the contract was signed (total of 68 negative records, with a maximum of 56 days and a median of 2 days). Although counterintuitive, this is possible, but unsurprisingly the fraud rate for these claims is 16.2% (compared to 3.5% on the whole dataset).

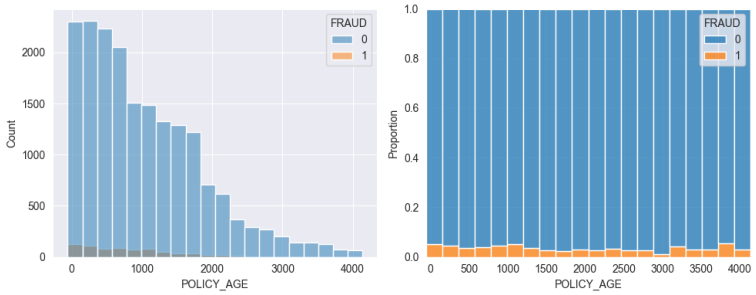

Figure 4.4: Distribution of policy age

Another interesting aspect of claim reporting can be its timing. Therefore, we have created a variable *'DAYS_TO_REPORT'*, which captures how many days it took to report the claim. Figure 4.5 shows the distribution of this variable and the proportion of frauds across the distribution. Once again, the values are restricted to the 99^{th} percentile, which corresponds to 188 days (approximately half year). The vast majority of claims are filled within the first days after the accident. We can observe a marginal increase in the proportion of frauds with increasing time to report the claim, but probably not a statistically significant trend.

From the categorical variables, probably the most interesting one is the *'DAMAGE_TYPE'*. This variable refers to the classification of damage to the vehicle, or more precisely to the type of coverage applicable. Various initial categories were grouped into seven categories: accidents (traffic accidents), parking, nature (nature forces, collision with an animal or inanimate object), vandalism (vandalism and theft), windshield damage, other (other damages or non-specified) and MTLP (where there is no further division). Figure 4.6 presents the distribution of the data between the categories and the proportion

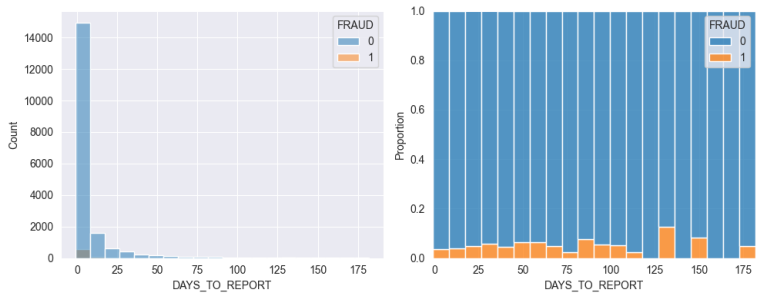

Figure 4.5: Distribution of days it took to report the claim

of frauds in individual categories. The most represented CASCO coverages are accident and windshield. For the windshield coverage, it is interesting to note that, although there are many entries, there is not a single fraud among them. This could be another example of the selection bias of the detection system, as mentioned in Section 4.3, as the average claim amount for the windshield coverage is significantly lower than for the other coverage categories (approximately half the next lowest category). In contrast, we observe a significantly higher proportion of frauds for vandalism coverage.

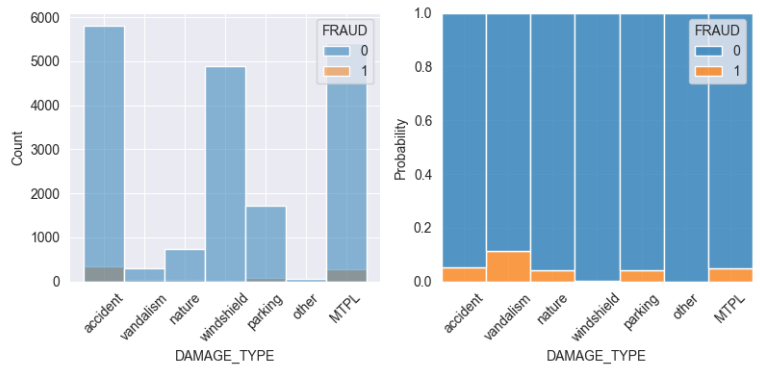

Figure 4.6: Distribution of damage type

Lastly, let us take a look at some demographic variables available in the dataset - the *'SEX'* and the *'PH_AGE'* variables. The *'SEX'* variable contains information on whether the policyholder is a female (F), male (M) or a company (C). Note that the variable does not refer to the driver or the owner of the vehicle but to the owner of an insurance policy. As can be seen in Table 4.5, the data in the dataset are fairly equally divided between the policy being owned by a natural person and a legal person, with men approximately two times more represented than women. However, neither of the groups displays a higher propensity to frauds as the proportion of fraudulent claims is similar.

The *'PH_AGE'* represents the age of the policyholder at the time of the accident (in years). For companies, the age (or the year of foundation) was not available, therefore the values were filled with '-1' and they are excluded from the following figure. Figure 4.7 presents the age distribution for women and men. We can observe that majority of the the policies held by natural persons are owned by people in the age between 40 to 60 years. It seems that younger people are slightly more prone to frauds, but no clear trend is observable.

Table 4.5: Representation of men, women and companies in the sample

	Company	**Female**	**Male**
No Fraud	9 784	3 030	6 087
Fraud	376	116	268

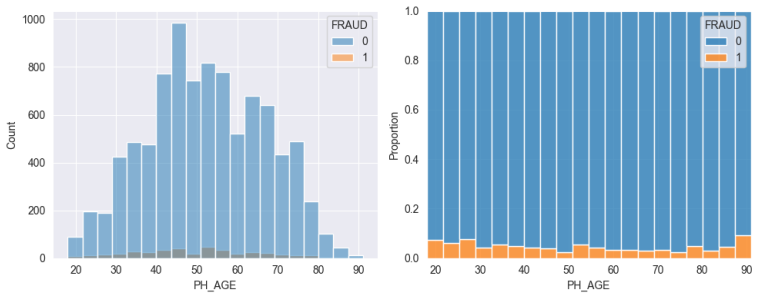

Figure 4.7: Distribution of policyholder age

This discussion only contains some of the most interesting variables from all the variables chosen for the modelling. More information on the other variables and additional summary statistics can be found in Appendix A.

Before building any model on the data, it is important to check the correlation between the variables. Therefore, we construct a correlation matrix, which is presented in Figure 4.8. The matrix shows the strength and direction of the correlation for all numerical and binary variables used.

We can observe that fraud detection in insurance claims is indeed very difficult, as there is no strong correlation between fraud and any of the variables in the dataset. The most significant is the slightly positive correlation with the *'CLAIM_AMOUNT'* variable. As for the other variables, there is a fairly strong correlation between variables *'N_CLAIMS_SINCE_2020'*, *'CLAIMS_VALUE_SINCE_2020'* an *'NT_CLAIM'*, which can be expected as all of them refer to the client's claim history. Another significant correlation

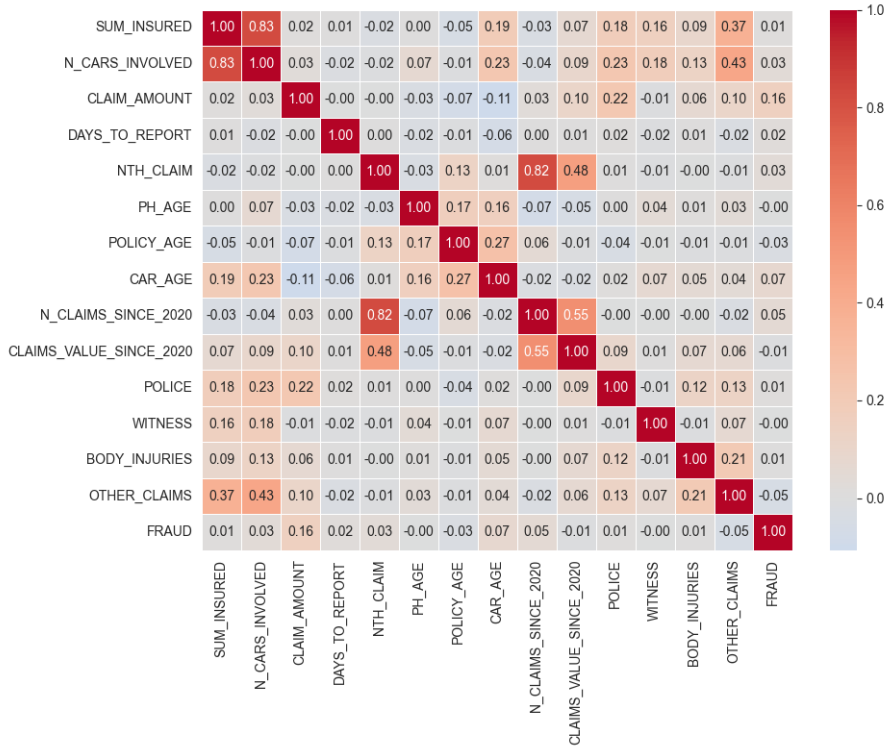

Figure 4.8: Correlation matrix of numerical and binary variables

can be seen among variables *'N_CAR_INVOLVED'*, *'SUM_INSURED'* and a binary variable *'OTHER_CLAIMS'*. This also makes sense to have more claims in accidents in which more vehicles were involved. Moreover, such accidents are likely stemming from MTPL insurance which generally has a significantly higher sum insured than CASCO policies. Besides that, the degree of correlation among our variables is fairly low.

Chapter 5

Results

This chapter presents empirical results of the proposed models on the real-world motor insurance dataset. Firstly, individual models' performance and results are described in detail and features important for individual models are discussed. Then we provide a comparison of the model and an overall discussion of the modelling result and the most important features and their implication on further development in insurance fraud detection. Lastly, we discuss possible extensions of our modelling approach.

Each of the proposed models has many hyperparameters that influence their performance and decision-making. Therefore, hyperparameter tuning was done for each model separately using the grid search of a grid of proposed parameters to find the most suitable ones. The size of the grid varies between the models as the number of parameters and their possible values also varies. The k-fold validation was applied to prevent sampling bias and, since the goal of the model is to pre-flag fraudulent claims, the recall was chosen as the metric to be optimized. In the following text, the results presented are the already optimized models on the validation set. An overview of used parameters can be found in Appendix B in Table B.1.

One of the parameters that were optimized is the degree of resampling applied. As described in Chapter 3, in a highly unbalanced classification problem such as fraud detection, it is important to account for the class imbalance. Therefore, both of the techniques proposed in that chapter, the class balancing and the SMOTE, were tested. Empirical research on real data showed a clear need to correct the class imbalance, as all the models performed better when some class balancing was applied. However, a simpler class weighting proved to outperform the more complex SMOTE and their combination for many of

the evaluated models. Even in cases where the SMOTE transformation improved performance, the improvement is not truly difference making. This is likely given by the fact that our data are fairly sparse as they contain a lot of categorical variables or numerical variables with limited number of values. The interpolation performed within the SMOTE rebalancing then does not create records that would add significant value to the model's learning ability. Nevertheless, in the following text, results for the models estimated both with and without SMOTE resampling are presented to provide a comparison of their performance and to evaluate the value added by the SMOTE transformation. The class weighting is applied in all the models presented, as it significantly improved performance in all of them[1].

5.1 Individual model results

5.1.1 Logistic regression

The first of the evaluated models is the logistic regression. This model was included mainly as a baseline comparison for the more complex methods. However, it seems to provide fairly competitive performance. In Table 5.1 the results of the logistic regression without and with SMOTE (with sampling strategy equal to 0.4) are presented.

As we are mainly interested in detecting frauds, the most interesting performance metric is the recall of the fraudulent data. Here, the model achieved a recall of 0.83 without SMOTE and 0.8 with SMOTE, which means that the model can detect more than 80% frauds. The precision and F1 scores are almost identical for the two models. However, the low precision on fraudulent data is one of the main drawbacks of the model, since only 9% of the claims marked as frauds by the model are actually fraudulent. On the other hand, if a claim is marked by the model as non-fraudulent, then we can be fairly sure it is not a fraud as 99% of 'no-frauds' predictions are made correctly. Lastly, the AUC is 0.75 for model without SMOTE and 0.74 with SMOTE, which means that even though there is still considerable room for improvement, the model is significantly better than random guessing. Overall, the results for with and without SMOTE are fairly similar, but in the two most important metrics, which are the recall of fraudulent records and the AUC, the model without SMOTE

[1]Except for the MLP model where the class weighting is not implemented.

slightly outperforms the one with SMOTE transformation applied. In any case, these results are quite good for a baseline model.

Table 5.1: Logistic regression results

		Precision	Recall	F1-score
no SMOTE	**No Fraud**	0.99	0.67	0.80
	Fraud	0.09	0.83	0.16
	Macro avg	0.54	0.75	0.48
	Weighted avg	0.95	0.68	0.78
	AUC		0.75	
SMOTE	**No Fraud**	0.99	0.69	0.81
	Fraud	0.09	0.80	0.17
	Macro avg	0.54	0.74	0.49
	Weighted avg	0.95	0.70	0.79
	AUC		0.74	

To better understand the model results, we perform an analysis of features importance. Even though exact coefficients can be extracted for logistic regression, the main goal of this thesis is to provide a comparative study of the models. Therefore, the same metrics and methods are applied and presented for all the models, and hence the permutation test and SHAP values are also presented for the evaluation of feature importance for logistic regression. The feature importance analysis is performed for the better model from the ones presented above, which is the one without SMOTE resampling.

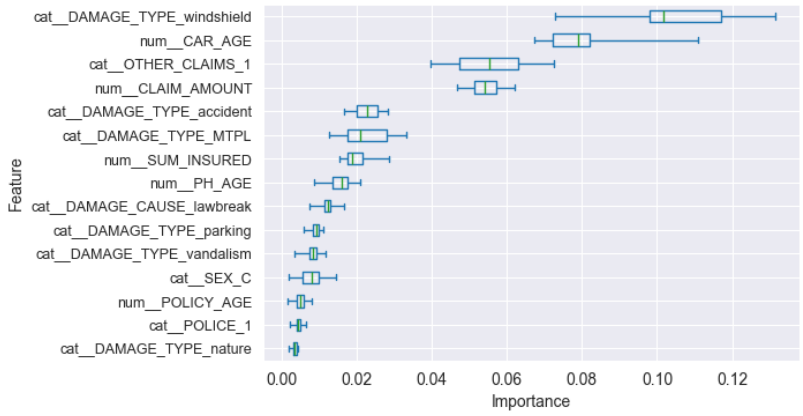

Figure 5.1: Permutation importance for logistic regression

Figure 5.1 shows the permutation importance of variables for logistic regression. The 'num' before the variable names denotes numerical variables and

the 'cat' denotes binary and categorical variables. For spatial reasons, only 15 of the most important features out of the 32 total features are shown in the figure. It is important to note that the permutation importance shows only the importance for the model performance, not the effect on the target variable (in our case the probability, or more specifically log-odds, of the claim being a fraud). We observe that windshield damage type, vehicle age, information on if other claims were filled and the claim amount are the most important variables in our dataset for fraud detection. This is in line with our expectations based on the analysis of the dataset described in Section 4.4.

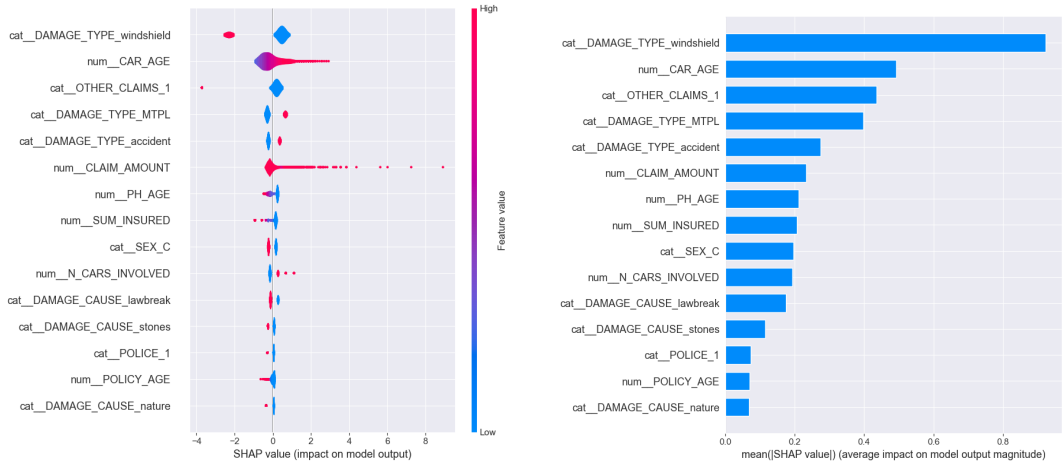

Figure 5.2: SHAP values for logistic regression

Figure 5.2 presents a summary plot of SHAP values on the left and plot of the mean of absolute SHAP values on the right. Again, only 15 most important variables are displayed. The most important variables based on SHAP values are again the damage type windshield, the vehicle age and the information about other claims during the incident, followed by whether the coverage is MTPL or CASCO and damage type accident. We can observe that with damage type windshield, the probability of the claim being fraudulent decreases. On the contrary, with damage type accident, the probability of fraud slightly increases. Moreover, with a higher vehicle age, the probability of fraud increases. Whereas with the claim being from the CASCO insurance (in other words not being MTPL) or if there are other claims, the probability of fraud decreases. All of this is in line with what we observed during the data analysis performed in the previous chapter. We can observe that the prediction is strongly driven by the windshield category, but then the magnitude of the effect is distributed across many variables. This is likely given by the linear nature of logistic regression.

5.1.2 Random forest

Next model whose suitability for fraud detection is evaluated is the random forest model. This model is based on combing the output of multiple decision trees to obtain a robust, accurate and stable estimate. The results of the random forest model can be found in Table 5.2 and are again presented for a model on the original data and on the data with SMOTE resampling applied (with a sampling strategy of 0.3).

The results of the two models are fairly similar. However, we are mostly interested in detecting frauds, and there the model without resampling outperforms the one with SMOTE resampling (recall of 0.83 compared to 0.77). On the other hand, the model with SMOTE performs better in detecting non-fraudulent cases and provides higher precision and F1-score. Nevertheless, the precision for fraudulent cases still remains the main drawback of the models, as it is still only 0.08 and 0.09 for models without and with SMOTE respectively, meaning that less than 10% of the claims marked as frauds actually prove to be fraudulent. In terms of AUC, the model with SMOTE transformation is slightly better (0.73 compared to 0.72), but the difference is only minor. Overall, the model with SMOTE seems to provide slightly better general results, but for the purpose of pre-flagging fraudulent claims, the model without SMOTE works better.

Table 5.2: Random forest results

		Precision	Recall	F1-score
no SMOTE	**No Fraud**	0.99	0.61	0.75
	Fraud	0.08	0.83	0.14
	Macro avg	0.53	0.72	0.45
	Weighted avg	0.95	0.62	0.73
	AUC		0.72	
SMOTE	**No Fraud**	0.99	0.70	0.81
	Fraud	0.09	0.76	0.16
	Macro avg	0.54	0.73	0.49
	Weighted avg	0.95	0.70	0.79
	AUC		0.73	

Interestingly, the performance of the random forest model is not better than the baseline logistic regression model as would be expected. On the contrary, the logistic regression actually outperforms the random forest model in both

precision and recall and also in the AUC, although the differences are rather minor.

To better understand the decision-making of the model, an analysis of the importance of individual features is performed. Again, the analysis is performed for the better of the two models described above, which, for the purpose of fraud pre-flagging, is the one without SMOTE transformation. Only 15 most important features are shown in the figures for a clearer presentation.

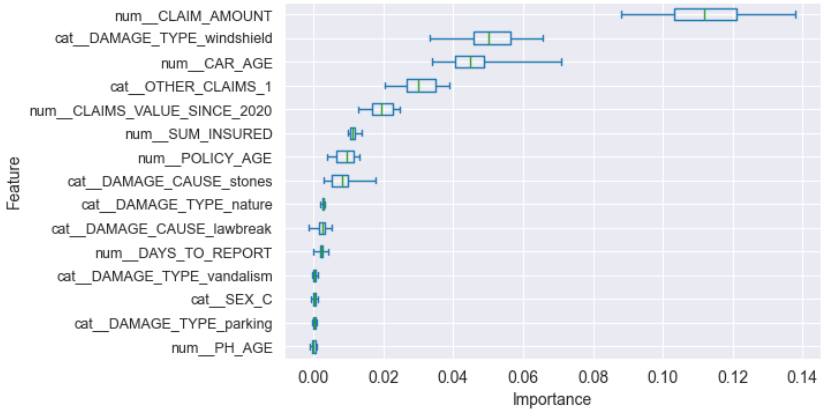

Figure 5.3: Permutation importance for random forest

Figure 5.3 displays the permutation importance of individual features for the random forest estimation. We can see that clearly the most important variables for model performance are the claim amount, followed by windshield damage type and vehicle's age and the information about other claims filled with respect to the incident and previous claims on the policy. We can see a shift in the distribution of the importance of individual features compared to logistic regression. This is probably given by the random forest model's ability to capture also non-linear relationships.

In Figure 5.4 the SHAP values for the random forest model are presented. On the left there is a summary plot of individual SHAP values, and on the right the mean of absolute SHAP values for each feature is illustrated. We can observe that the variable that influences the prediction the most is again the claim amount, where with a higher claim amount the probability of the claim being fraudulent also increases. It is followed by the damage type windshield, where if the claim is a claim on the windshield, the probability of being fraudulent significantly decreases. Similarly, for damages caused by stones, the probability of fraud is lower than for those not caused by stones. Next, with a higher vehicle's age, the probability of fraud increases, and the amount of

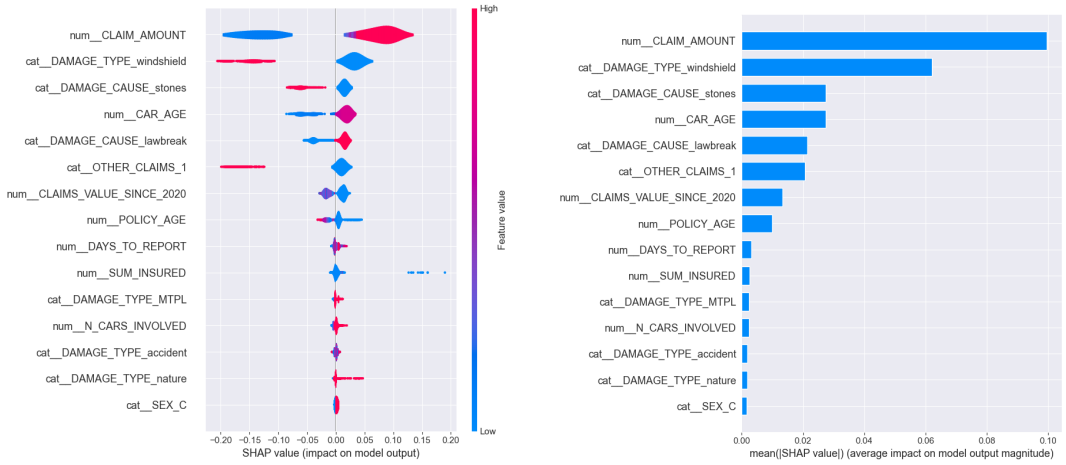

Figure 5.4: SHAP values for random forest

previous claims since 2020 seems to slightly decrease the probability of fraud, whereas if there were no other claims filled with respect to the incident, the probability of fraud slightly increases. However, we can observe that the prediction is strongly determined mainly by two variables - the claim amount and the damage type windshield. The directions of the influences are similar to what we observe for the logistic regression and correspond to the expectation based on the dataset analysis. The magnitude of effects and order of the variables differ for random forest and logistic regression as the models are based on different principles and treat variables significantly differently. Moreover, in the case of random forest the SHAP values refer to a change in probabilities, whereas for logistic regression the SHAP values refer to change log-odds.

5.1.3 Histogram-based gradient boosting

The next model considered for fraud detection purposes is the histogram-based gradient boosting. This is an advanced implementation of the gradient boosting algorithm provided by the *scikit-learn* library. Using histogram-based techniques, the algorithm is able to enhance efficiency and therefore is particularly useful for larger datasets. The results of the histogram-based gradient boosting model without and with SMOTE resampling (with a sampling strategy of 0.2) can be found in Table 5.3.

In terms of the ability to detect fraud, the model without SMOTE performs better than the one with SMOTE (with recall of 0.81 compared to 0.75). The issue of our model with low precision, and therefore rather low efficiency, per-

sists as the models still achieve a precision of only 0.9 without SMOTE and 0.10 with SMOTE, meaning that 90% of the claims marked by the model as frauds are actually not frauds. However, the main goal is to test suitable models for pre-flagging before further claim processing, therefore high recall is the main priority. In terms of the overall quality of the model, proxied by the AUC, the model without SMOTE seems to be slightly better, but the difference is rather minor. In conclusion, once again, the model with SMOTE resampling provides slightly better results in terms of precision and F1-score, but in the two most important metrics, which are recall for frauds and the AUC, the model without SMOTE resampling achieves higher results.

Table 5.3: Histogram-based gradient boosting results

		Precision	**Recall**	**F1-score**
no SMOTE	**No Fraud**	0.99	0.67	0.80
	Fraud	0.09	0.81	0.16
	Macro avg	0.54	0.74	0.48
	Weighted avg	0.95	0.67	0.77
	AUC		0.74	
SMOTE	**No Fraud**	0.99	0.71	0.83
	Fraud	0.10	0.75	0.17
	Macro avg	0.54	0.73	0.50
	Weighted avg	0.95	0.72	0.80
	AUC		0.73	

Interestingly, the histogram-based gradient boosting model also does not outperform the baseline logistic regression model. Logistic regression provides slightly better results both in terms of recall for the fraudulent class and the AUC, which are the two most important metrics in our use case.

An analysis of feature importance is again carried out for the better of the two models presented above, which, for the purpose of detecting insurance fraud, is the model without SMOTE. Figures 5.5 and 5.6 show the 15 most important variables based on permutation importance and SHAP values, respectively.

In Figure 5.5, we can observe that the predictions are strongly influenced by the variable claim amount, followed by damage type windshield, information on additional claims, vehicle's age and value of other claims since 2020. These variables essentially determine the model's decision process, as the importance

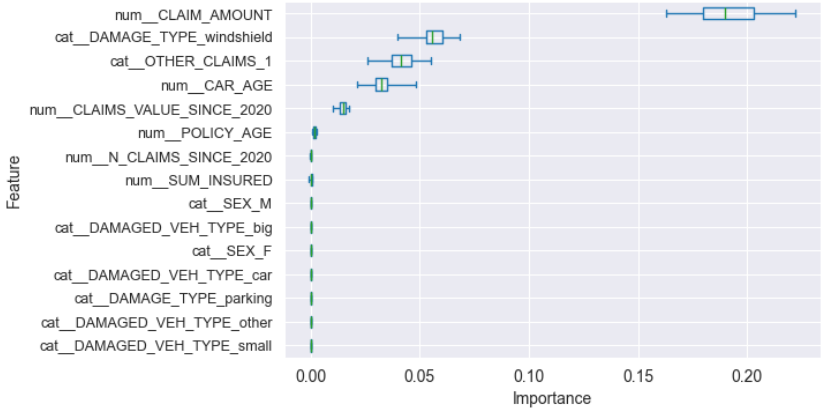

Figure 5.5: Permutation importance for histogram-based gradient boosting

of the other variables is negligible. In comparison to the previous models, here we clearly see the most influential feature and the four other important features, whereas for logistic regression and random forest models the importance was more distributed among different features. Nevertheless, the most important features remain the same, although their ranking slightly differs.

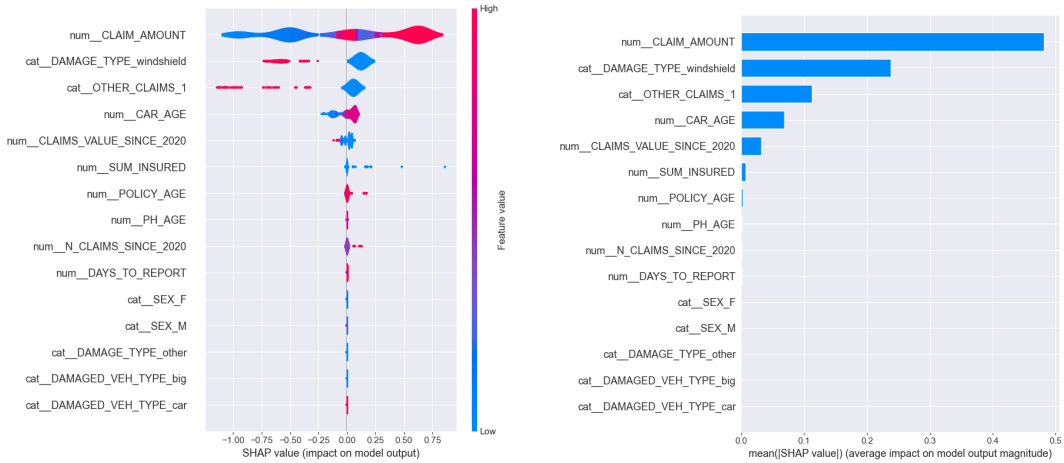

Figure 5.6: SHAP values for histogram-based gradient boosting

Figure 5.6 presents the individual SHAP values on the left and the mean of absolute SHAP values on the right. In line with the permutation importance and same as for random forest, we can observe that the claim amount is driving our prediction. However, the effect in the case of histogram-based gradient boosting model is not as simple as in the case of the random forest. It seems that with a higher claim value, the probability of fraud still increases, but there

is a cluster of high values also around SHAP values equal to zero. This is likely given by an integration with another variable within the decision trees. For example, the claim amount might increase the probability of fraud, except for cases when the car is new (i.e. vehicle age is low) and the high amount is justified. However, we would need to examine the relationship between the variables more closely to confirm this hypothesis. As for the damage type windshield and information about additional claims, they both significantly decrease the probability of the claim being fraudulent. On the other hand, with a higher vehicle age, the probability of fraud increases. Lastly, with a higher value of other claims since 2020, the probability of this claim being fraudulent seems to decrease, which might be a little counterintuitive, as we would expect a policy holder with a high number of claims to be more suspicious for fraud commitment, but this does not prove in the data. These five variables drive the model's predictions, as the SHAP values of the other variables are negligible.

5.1.4 XGBoost

The last tree-based model tested is the XGBoost algorithm. XGBoost is an advanced implementation of the gradient boosting framework, which utilizes decision trees as base learners and employs regularization techniques to improve model generalization. XGBoost package enables a parallel tree boosting to improve efficiency and decrease training time. Table 5.4 presents the results for the XGBoost estimation without SMOTE resampling and with the SMOTE resampling with a sampling strategy of 0.4.

The XGBoost model without SMOTE resampling achieves a recall of 0.74, precision of 0.12 and hence the F1-score of 0.21, and the AUC of 0.77. This is slightly lower performance in recall than we observed in the previous models, but a not insignificant improvement in precision and AUC compared to the previous models. Overall, the general results of the model are good, but in the recall, which is the most important metric for fraud detection, the XGBoost model without SMOTE resampling is falling behind the other models. On the contrary, for the XGBoost model with SMOTE resampling, we observe the best performance in recall so far with 0.92. This means that this model is able to capture more than 90% of fraudulent claims. The precision of 0.9 is similar to logistic regression, random forest and the histogram-based gradient boosting and remains the biggest weakness of our modelling approach. The AUC is 0.77 which is also the best result among all models considered. Overall, for

the purpose of fraud detection, the XGBoost model with SMOTE transformation seems to provide the best results with a recall of 0.92 and reasonable efficiency costs, in terms of claims incorrectly classified as frauds.

Table 5.4: XGBoost results

		Precision	Recall	F1-score
no SMOTE	**No Fraud**	0.99	0.79	0.88
	Fraud	0.12	0.74	0.21
	Macro avg	0.55	0.76	0.54
	Weighted avg	0.95	0.78	0.85
	AUC		0.76	
SMOTE	**No Fraud**	0.99	0.62	0.77
	Fraud	0.09	0.92	0.16
	Macro avg	0.54	0.77	0.46
	Weighted avg	0.96	0.63	0.74
	AUC		0.77	

To better understand the model and its decision making, a feature importance analysis is again performed for the better of the two models described above. This time, the best model for fraud detection is the one with SMOTE transformation applied.

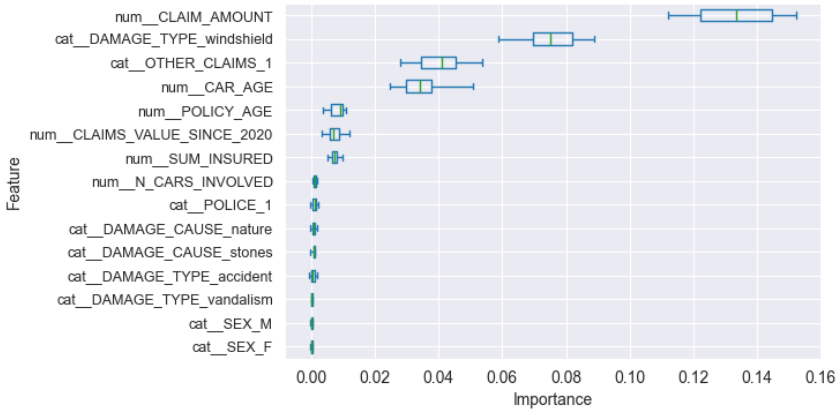

Figure 5.7: Permutation importance for XGBoost

Figure 5.7 illustrates the permutation feature importance for the 15 most important variables for the XGBoost model with SMOTE resampling. We observe that the performance of the model is again strongly determined by the claim amount variable, followed by the damage type windshield. The other significant variables are information about additional claims for the incident and

the vehicle's age. The importance of other variables is low or even negligible. This is similar to what we observed in the previous models. Compared to the other gradient boosting method, here the importance is slightly more evenly distributed, but the performance is still strongly driven by the claim amount.

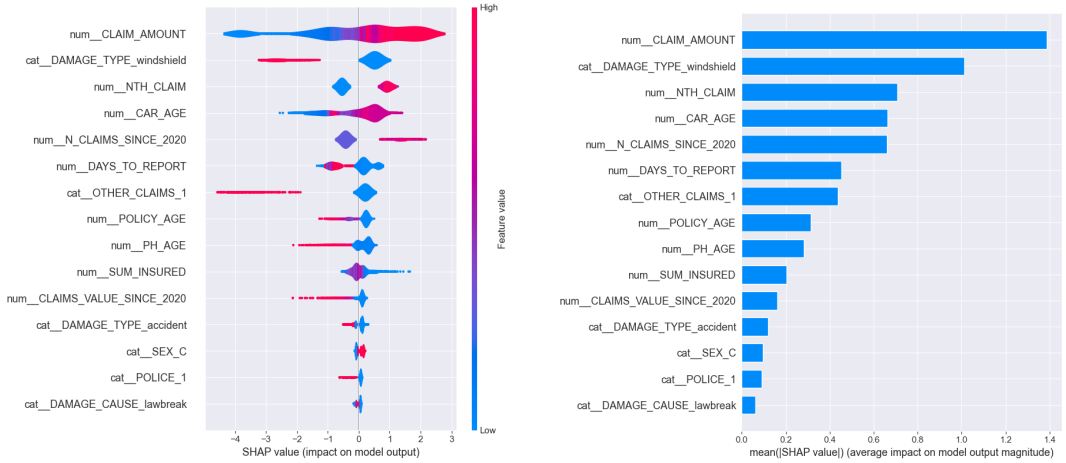

Figure 5.8: SHAP values for XGBoost

The SHAP values for the XGBoost model with SMOTE resampling are presented in Figure 5.8, where again the SHAP values for individual observations are displayed on the left and the mean absolute SHAP values are shown on the right. The feature that drives the prediction is, similarly to the other tree-based models, the claim amount, where with a higher claim amount, the probability of fraud increases. It is followed by the damage type windshield, which significantly decreases the probability of the claim being fraudulent. With increasing vehicle's age, the probability of the claim being fraudulent also increases, but the relationship is not as clear as we have seen in previous models. Generally, these variables and the direction of their effects are in line with what we have observed in the previous models and the data analysis performed in Section 4.4. Regarding the number of previous claims and the number of claims since 2020, a higher number of previous claims in the policy history and previous claims since 2020 tend to increase the probability of fraud. Compared to the other tree-based models, especially the other gradient boosting method, we observe that the SHAP values are more distributed among many variables.

5.1.5 Multilayer perceptron

Lastly, to provide a comparison to the tree-based methods, the multilayer perceptron (MLP) is trained. The MLP is an artificial neural network that consists of at least three layers of fully connected neurons (in our case 4 layers - an input, 2 hidden and a output layers) with a nonlinear activation function. Table 5.5 presents the results of two different MLP models, where both rely on the same underlying algorithm, but were trained on a different set of parameters. The parameters used can be referred to in Table B.1 in Appendix B. The SMOTE resampling has been applied for both models, since the class weighting is not implemented within the MLP framework and the class balancing proved to improve model performance significantly. The sampling strategy was set to 0.7 for the first model and to 0.8 for the second model.

The first MLP model achieves a very high recall for fraudulent cases (recall of 0.97), but the precision (0.06) for frauds and the AUC (0.65), as well as the recall for non-fraudulent claims (0.34), are very low compared to the other models. Basically, this model is predicting everything as a fraud unless it is absolutely sure that the claim is not a fraud. Therefore, even though this model performs the best in terms of recall on fraudulent claims, which is our main evaluation metric, it is not a good model, as it is not suitable for real-word application due to its very low efficiency (given by the low precision) and poor model quality (given by the low AUC). The second MLP model provides more balanced results. The precision is 0.11, the recall for 'no-frauds' is 0.76 and the AUC bounces back to 0.73, which is comparable to the previous models. However, the improvement in precision causes the recall for fraudulent claims to drop significantly to 0.7. Here we can clearly see the problem of precision-recall trade-off. The results of the second MLP model are generally worse than the previous models, including the baseline logistic regression model. Therefore, we would probably need a more complex architecture of the neural network to properly capture the relationships between the data. The future of application of neural networks for fraud detection might lie in the deep neural network based on transformers and pre-trained transformers. These have already been shown to outperform traditional machine learning techniques, as well as some other deep learning architectures, for the purpose of detecting credit card fraud (Yu *et al.* (2024)).

Nevertheless, we also perform a feature importance analysis for the MLP

Table 5.5: Multilayer perceptron results

		Precision	Recall	F1-score
MLP 1	**No Fraud**	1.00	0.34	0.50
	Fraud	0.06	0.97	0.10
	Macro avg	0.53	0.65	0.30
	Weighted avg	0.96	0.36	0.49
	AUC		0.65	
MLP 2	**No Fraud**	0.98	0.76	0.86
	Fraud	0.11	0.70	0.18
	Macro avg	0.55	0.73	0.52
	Weighted avg	0.95	0.76	0.83
	AUC		0.73	

model. The feature importance measures displayed are calculated for the second, more balanced, MLP model.

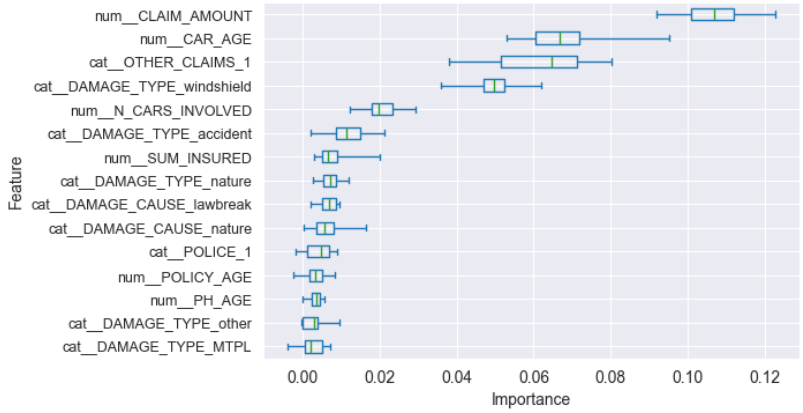

Figure 5.9: Permutation importance for multilayer perceptron

The permutation importance for 15 most important variables for the MLP is shown in Figure 5.9. We observe that the importance is more distributed among the variables than in the case of tree-based methods. Nevertheless, the performance of the model is still strongly determined by the claim amount, followed by the vehicle's age, information about additional claims filled with respect to the incident, and damage type windshield. This is similar to what is observed in the other models and it is in line with expectation based on the dataset exploration.

Figure 5.10 presents the SHAP values for the MLP model. Due to high computational demand, the SHAP values for MLP were calculated only for

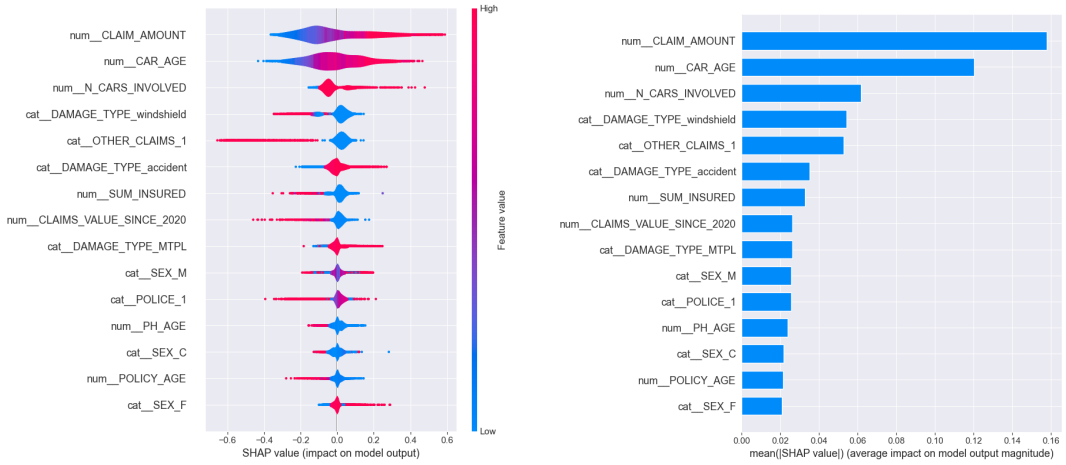

Figure 5.10: SHAP values for multilayer perceptron

a randomly sampled subsample, which should provide sufficient evidence to derive conclusions about the effects of individual features. We can observe that the prediction are again determined mainly by claim amount and vehicle's age, where with a higher value of these variables the probability of fraud increases. Followed by the number of cars involved, which is a new one compared to the previous models, but the effect from simple plotting of the SHAP values is rather inconclusive. Furthermore, with damage type windshield and if other claims were filled with respect to the incident, the probability of the claim being fraudulent decreases. These are the main drivers of the prediction, then the SHAP values are evenly distributed among many features. Except for the number of vehicles involved in the accident, these are similar to what we have observed in the previous models. Compared to previous models, especially the tree-based methods, the effects of individual feature values have a more continuous nature, but also exhibit less outliers than logistic regression. This is given by the significantly different and more complex structure of the MLP model.

5.2 Performance comparison

The goal of this thesis was to compare several machine learning methods and their suitability for application within the detection of motor insurance frauds. These include logistic regression, random forest, histogram-based gradient boosting, XGBoost and multilayer perceptron. For each of them, a version with and without the SMOTE transformation was tested. The only models

where the SMOTE provided a significant value to the model performance were the XGBoost and the MLP models, which does not support the class weighting option. Table 5.6 shows an overview of the results of the models. A complete overview of all 10 models can be found in Appendix B in Table B.2. The table presents the precision, recall and F1-score for fraudulent claims and the AUC.

Table 5.6: Comparison of models' results

Model	Precision	Recall	F1-score	AUC
Logistic Regression	0.09	0.83	0.16	0.75
Random Forest	0.08	0.83	0.14	0.72
Histogram-based GBM	0.09	0.81	0.16	0.74
XGBoost with SMOTE	0.09	0.92	0.16	0.77
MLP (2)	0.11	0.70	0.18	0.73

From the overview table, we can clearly see that the XGBoost outperformed the other models in number of fraud detected, as it is the only model that reached a recall of more than 0.9. The precision and F1-score are almost identical across the models, except for the MLP, but there the recall is significantly lower. In term of AUC, which is used to evaluate overall model quality, the XGBoost also provides the best result among all the tested approaches with AUC of 0.77. Hence, XGBoost as the best of the tested models can detect 92% of frauds, but at the cost of only 9% of the claims marked as frauds being an actual fraud. This precision-recall trade of can be adjusted to current needs by setting a different decision threshold, but the XGBoost should still provide the best results thanks to it higher AUC score.

In terms of comparison with existing literature, it is very difficult to make one, as the performance of the models is strongly determined by the characteristics of the underlying dataset. Since we are using a unique real-world dataset that was assembled for the purpose of this thesis, there is no research on the same data to compare our results with. However, XGBoost as the best model is not surprising, as it has been shown to provide the best results among many models for several fraud detection use cases (Maina *et al.* (2023), Sinčák (2023)).

The main tool to increase the performance of the models would be the input data - both in sense of size of the dataset, specifically the number of detected frauds, but also in terms of features available and feature engineering. There are several important characteristics of the insurance contract that were

not available in the current dataset. This includes the policy premium or information about the vehicle's value. Moreover, when the incident is reported, the report form is accompanied by a text description of the incident and photo documentation of the scene and the damage to the vehicle. These could also be processed and used as additional input information for the models. Some advanced text analytics for motor insurance claim has been explored by Wang & Xu (2018) and considerably improved the results. Lastly, many of the variables suffer from poor data quality. Therefore, introducing thorough controls for the information entered in the report form would be beneficial.

All the tested models can be adjusted for practical applications to provide a probability or scoring instead of 0/1 flags. Additionally, information on value at risk, in other words the potential loss due to the fraud, can also be added to the prediction process. This can further help insurance companies to efficiently allocate resources available for claim investigation.

5.3 Feature importance

Besides the evaluating the model themselves, an analysis of feature importance was also performed. This is important to better understand the complex machine learning models and to be able to verify that they base their prediction on reasonable inference, not a random coincidence in the sample (note that the data on frauds were complemented by only a representative sample of non-fraudulent claim).

The most important features were similar for all the tested models, even if their ranking differed slightly between the models. The main drivers of our models in terms of both permutation importance and SHAP values include the *'CLAIM_AMOUNT'*, *'DAMAGE_TYPE_WINDSHIELD'*, *'CAR_AGE'* and *'OTHER_CLAIMS'*. The XGBoost estimation also adds a *'NTH_CLAIM'*, *'N_CLAIMS_SINCE_2020'* and *'DAYS_TO_REPORT'* into the mix.

The claim amount is the variable that drives the prediction for majority of the models. With a higher claim amount, the probability that the claim is fraudulent increases. The relationship between the claim amount values and their SHAP values (hence, their effects on prediction) can be seen in Figure 5.11, which shows the SHAP values for individual observations in the validation set based on the XGBoost model with SMOTE transformation.[2] We can see that

[2]Note that the data in the figure are normalized and restricted of outliers for better illustration of the relationship.

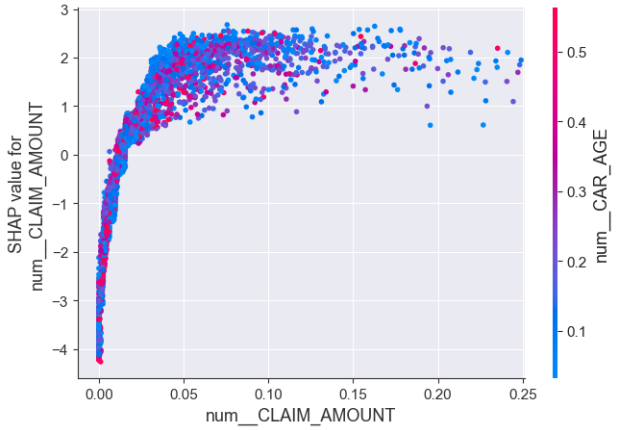

Figure 5.11: SHAP values for claim amount

indeed with increasing claim amount the SHAP values increase. However, the relationship is not linear, as from certain claim amount the SHAP values remain at a similar level. The color of the points represents the vehicle's age. However, we do not observe any distinct pattern between the claim amount and the vehicle's age. Overall, the effect of the claim amount on the probability of fraud is reasonable. Since the insurance company operates with limited resources towards the claim liquidation, they are likely to concentrate more on higher claims. Also, from the client point of view, committing a fraud for a nominally low damage might not be worth the reward, whereas for big damage, the clients might deem it worth the attempt for a fraud.

The second variable that usually drives the prediction is the damage type windshields. With a damage type windshield, the probability of fraud significantly increases. This might be due to several reasons. Firstly, this type of insurance usually incurs claims with a nominally lower amount (in our data the mean claim amount for windshield coverage is half the next lowest). As described for the claim amount variable, the negative effect on probability of fraud might be caused by this type of damage not being a priority for fraud detection due to low potential loss. In addition, windshield damage frauds are especially hard to prove.

The next variable that strongly influences the predictions for most of the models is the vehicle's age. Figure 5.12 shows the SHAP values of the vehicle's age for individual validation set observations. The SHAP values are again based on XGBoost model with SMOTE resampling, as the best performing model. For newer vehicles, we observe that with increasing age of the vehicle the probability of fraud also increases. But from a certain age the probability starts to slightly

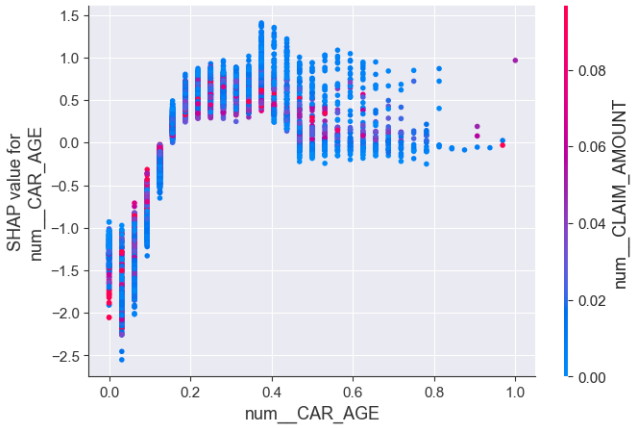

Figure 5.12: SHAP values for vehicle's age

decrease with increasing vehicle's age. Generally, it can be assumed that older cars might be more prone to breaking which the client might try to stage as an accident claim. Moreover, one of the most common frauds is inflated repair costs, which is more likely in an older car than in a brand new one. Additionally, this is the best proxy for the value of the damaged car available in the dataset. The color scale shows the value of the claim amount for the individual observations. Similarly to the claim amount plot, here we also do not see any clear pattern.

Chapter 6

Conclusion

Insurance fraud is a constantly present challenge for insurers around the world. Fraudulent activities in the insurance sector not only lead to substantial financial losses but also contribute to increased premiums for honest customers and erode trust in the insurance system. This thesis explored the suitability of advanced machine learning algorithms for application within insurance fraud detection. It provides a comparison of several supervised machine learning algorithms and modelling approaches. The selected algorithms include logistic regression, random forest, histogram-based gradient boosting machine, XGBoost and multilayer perceptron (MLP). In addition, the importance and effects of individual features were explored to improve the interpretability and understandability of the machine learning solution.

For the empirical analysis, a unique real-world dataset was provided by one of the leading MTPL and CASCO insurers in the Czech Republic. The fraud detection is a highly unbalanced problem as fraudulent claims amount to only about 3.5% of the observations in the dataset. Therefore, two approaches for correcting class imbalance were tested - the Synthetic Minority Over-sampling Technique (SMOTE) and the class weighting. Empirical research showed a clear need to address the class imbalance, as all models performed significantly better with the class weighting. The synthetization of additional fraudulent observations proved to improve the results only for the XGBoost and MLP models.

The performance of the real fraud detection system was not available, therefore, logistic regression was used as the baseline model. However, even the baseline logistic regression achieved satisfactory results in terms of recall and AUC, which were the main metrics used to evaluate the quality of the models. The only model that significantly outperformed logistic regression was the XGBoost

model with the SMOTE transformation. This model achieved an AUC of 0.77 and was able to detect 92% of frauds in the dataset. The biggest weakness of the modelling approach was precision as only 9% of the frauds predicted by the models were actual frauds.

Nevertheless, machine learning models offer a flexible and fast solution with simple implementation that can be easily updated and adjusted to current needs. However, many of them are considered 'black-box' models as the decision-making process of the model cannot be easily extracted. Therefore, permutation feature importance and the SHAP values were used to evaluate the effects of individual features. The most important features for fraud prediction were shown to be the nominal amount of the claim, type of insurance coverage (i.e., type of damage), the vehicle's age and information about other claims filled for the same incident. The analysis of feature effects within the models proved a reasonable interference of the machine learning algorithms as the effects are in line with the general logic behind the insurance fraud and the data analysis performed.

In conclusion, the integration of machine learning into the fraud detection systems of insurance companies could provide a significant advancement in the fight against fraud. The results of this thesis underscore the potential of these technologies to provide efficient, scalable and automated solution as an alternative to traditional scenario-based methods and people-dependent models.

Bibliography

Alarfaj, F. K., I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, & M. Ahmed (2022): "Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms." *IEEE Access* **10**: pp. 39700–39715.

Baran, S. & P. Rola (2022): "Prediction of motor insurance claims occurrence as an imbalanced machine learning problem." *arXiv preprint arXiv:2204.06109* .

Chawla, N. V., K. W. Bowyer, L. O. Hall, & W. P. Kegelmeyer (2002): "Smote: synthetic minority over-sampling technique." *Journal of artificial intelligence research* **16**: pp. 321–357.

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

Friedman, J. H. (2001): "Greedy function approximation: a gradient boosting machine." *Annals of statistics* pp. 1189–1232.

Gupta, R. Y., S. S. Mudigonda, & P. K. Baruah (2021): "A comparative study of using various machine learning and deep learning-based fraud detection models for universal health coverage schemes." *International Journal of Engineering Trends and Technology* **69(3)**: pp. 96–102.

Haixiang, G., L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, & G. Bing (2017): "Learning from class-imbalanced data: Review of methods and applications." *Expert systems with applications* **73**: pp. 220–239.

Hanafy, M. & R. Ming (2021a): "Machine learning approaches for auto insurance big data." *Risks* **9(2)**: p. 42.

Hanafy, M. & R. Ming (2021b): "Using machine learning models to compare various resampling methods in predicting insurance fraud." *J. Theor. Appl. Inf. Technol* **99(12)**: pp. 2819–2833.

InsuranceEurope (2019): "Insurance fraud: not a victimless crime." *IE report* .

Itoo, F., Meenakshi, & S. Singh (2021): "Comparison and analysis of logistic regression, naïve bayes and knn machine learning algorithms for credit card fraud detection." *International Journal of Information Technology* **13**: pp. 1503–1511.

Kaur, H., H. S. Pannu, & A. K. Malhi (2019): "A systematic review on imbalanced data challenges in machine learning: Applications and solutions." *ACM Computing Surveys (CSUR)* **52(4)**: pp. 1–36.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, & T.-Y. Liu (2017): "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in neural information processing systems* **30**.

Khatri, S., A. Arora, & A. P. Agrawal (2020): "Supervised machine learning algorithms for credit card fraud detection: a comparison." In "2020 10th international conference on cloud computing, data science & engineering (confluence)," pp. 680–683. IEEE.

Lundberg, S. M., G. G. Erion, & S.-I. Lee (2018): "Consistent individualized feature attribution for tree ensembles." *arXiv preprint arXiv:1802.03888* .

Lundberg, S. M. & S.-I. Lee (2017): "A unified approach to interpreting model predictions." *Advances in neural information processing systems* **30**.

Mahmood, T., S. K. Hashemi, S. L. Mirtaheri, & S. Greco (2023): "Machine learning techniques for detecting fraud in credit card transactions." In "SEBD," pp. 469–478.

Maina, D. G., J. C. Moso, & P. K. Gikunda (2023): "Detecting fraud in motor insurance claims using xgboost algorithm with smote." In "2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)," pp. 61–66. IEEE.

Muaz, A., M. Jayabalan, & V. Thiruchelvam (2020): "A comparison of data sampling techniques for credit card fraud detection." *International Journal of Advanced Computer Science and Applications* **11(6)**.

Murphy, K. P. (2012): *Machine learning: a probabilistic perspective*. MIT press.

Nalluri, V., J.-R. Chang, L.-S. Chen, & J.-C. Chen (2023): "Building prediction models and discovering important factors of health insurance fraud using machine learning methods." *Journal of Ambient Intelligence and Humanized Computing* **14(7)**: pp. 9607–9619.

Nhat-Duc, H. & T. Van-Duc (2023): "Comparison of histogram-based gradient boosting classification machine, random forest, and deep convolutional neural network for pavement raveling severity classification." *Automation in Construction* **148**: p. 104767.

Obodoekwe, N. & D. T. Van der Haar (2019): "A comparison of machine learning methods applicable to healthcare claims fraud detection." In "Information Technology and Systems: Proceedings of ICITS 2019," pp. 548–557. Springer.

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

Česká associace pojišťoven (2022): "Statistické údaje dle metodiky ČAP 1-12/2022." .

Česká associace pojišťoven (2024): "Pojistný podvod se Čechům příčí, počet podezřelých pojistných událostí přesto i v roce 2023 rostl." https://www.cap.cz/tiskove-centrum/tiskove-zpravy/, accessed: 16/07/2024.

Rukhsar, L., W. H. Bangyal, K. Nisar, & S. Nisar (2022): "Prediction of insurance fraud detection using machine learning algorithms." *Mehran University Research Journal of Engineering & Technology* **41(1)**: pp. 33–40.

Severino, M. K. & Y. Peng (2021): "Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata." *Machine Learning with Applications* **5**: p. 100074.

Shapley, L. S. *et al.* (1953): "A value for n-person games." .

Sinčák, J. (2023): "Machine learning methods in payment card fraud detection." Masters thesis. Charles University, Faculty of Social Sciences, Institute of Economic Studies, Prague. 2023, pages 71. Advisor: doc. PhDr. Jozef Baruník, Ph.D.

Singhal, A., N. Singhal, K. Sharma *et al.* (2023): "Machine learning methods for detecting car insurance fraud: Comparative analysis." In "2023 3rd International Conference on Intelligent Technologies (CONIT)," pp. 1–5. IEEE.

Trivedi, N. K., S. Simaiya, U. K. Lilhore, & S. K. Sharma (2020): "An efficient credit card fraud detection model based on machine learning methods." *International Journal of Advanced Science and Technology* **29(5)**: pp. 3414–3424.

Udeze, C. L., I. E. Eteng, & A. E. Ibor (2022): "Application of machine learning and resampling techniques to credit card fraud detection." *Journal of the Nigerian Society of Physical Sciences* pp. 769–769.

Wang, Y. & W. Xu (2018): "Leveraging deep learning with lda-based text analytics to detect automobile insurance fraud." *Decis. Support Syst.* **105**: pp. 87–95.

Yu, C., Y. Xu, J. Cao, Y. Zhang, Y. Jin, & M. Zhu (2024): "Credit card fraud detection using advanced transformer model." *arXiv preprint arXiv:2406.03733* .

Yuan, M. (2022): "A transformer-based model integrated with feature selection for credit card fraud detection." In "Proceedings of the 2022 7th International Conference on Machine Learning Technologies," pp. 185–190.

Zakaryazad, A. & E. Duman (2016): "A profit-driven artificial neural network (ann) with applications to fraud detection and direct marketing." *Neurocomputing* **175**: pp. 121–131.

Appendix A

Data Description

Variable	Description	Type
CAR_AGE	Age of the damaged car [years]	Numerical
CLAIM_AMOUNT	Value of the claim [CZK]	Numerical
CLAIMS_VALUE_SINCE_2020	Sum of claims on given policy since 2020 [CZK]	Numerical
DAYS_TO_REPORT	Number of days from the incident to the day the claim was reported	Numerical
N_CARS_INVOLVED	Number of cars involved in the incident	Numerical
N_CLAIMS_SINCE_2020	Number of claims on given policy since 2020	Numerical
NTH_CLAIM	Rank of the claim in the whole policy history (i.e. how many claim there are in the policy history)	Numerical
PH_AGE	Age of the policy holder [years]	Numerical
POLICY_AGE	Age of the policy [days]	Numerical
SUM_INSURED	The sum assured of the policy [CZK]	Numerical
DAMAGE_CAUSE	Cause of the damage [stones / nature (nature forces as floods, fires etc. and animals encounters) / traffic]	Categorical
DAMAGE_TYPE	Type of the damage incurred [accident / parking / theft / windshield /MTPL / nature (nature force, damage by both living and inanimate things)]	Categorical
DAMAGED_VEH_TYPE	Type of the damaged vehicle [big (e.g. bus, truck, work machinery) / auto / small (e.g. motorcycle) / other (other or unspecified)]	Categorical
SEX	Sex of the policy holder or indication of legal entity owner [M (Male) / F (Female) / C (the policy is held by a company)]	Categorical
BODY_INJURIES	If there were bodily injuries during the accident	Binary
POLICE	If police was called to the incident	Binary
OTHER_CLAIMS	If there were any other claims filled with respect to the incident (bodily injuries, other property damages etc.)	Binary
WITNESS	If any external witnesses were reported	Binary

Table A.1: Variables description

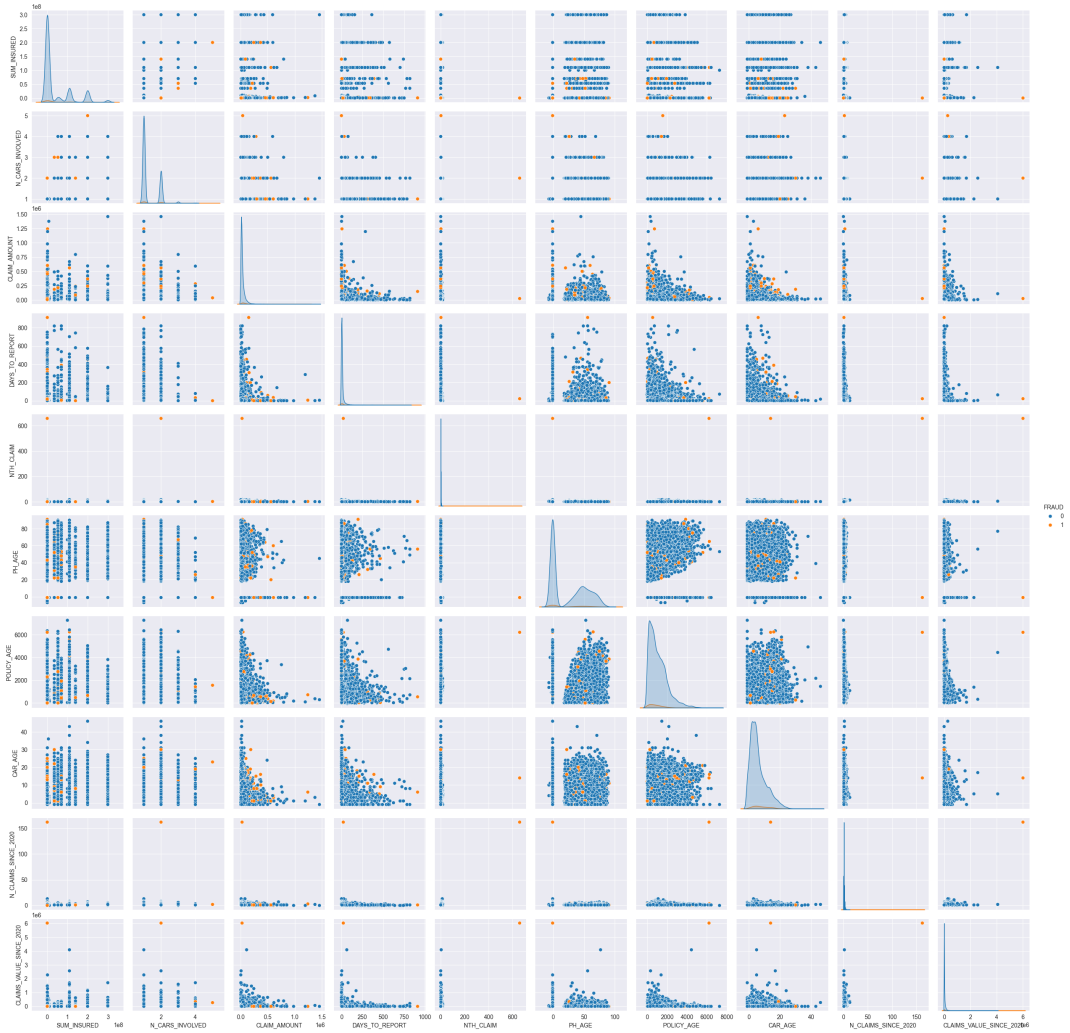

Figure A.1: Numerical variables distributions and correlations plot

Table A.2: Descriptive statistics for numeric variables

	mean	std	min	25%	50%	75%	max
SUM_INSURED	40 818 280	72 325 590	0	331 516	577 600	53 000 000	300 000 000
N_CARS_INVOLVED	1	0.4	1	1	1	2	5
CLAIM_AMOUNT	34 291	54 434	0	7 825	18 355	40 000	1 458 500
DAYS_TO_REPORT	12	42	-1	0	2	7	913
NTH_CLAIM	2	5	1	1	1	2	658
PH_AGE	25	28	-7	-1	-1	51	91
POLICY_AGE	1 070	910	-56	369	849	1 552	7 266
CAR_AGE	7	5	-1	2	5	9	46
N_CLAIMS_SINCE_2020	1	2	0	1	1	1	162
CLAIMS_VALUE_SINCE_2020	29 199	94 881	0	0	0	23 671	6 017 390

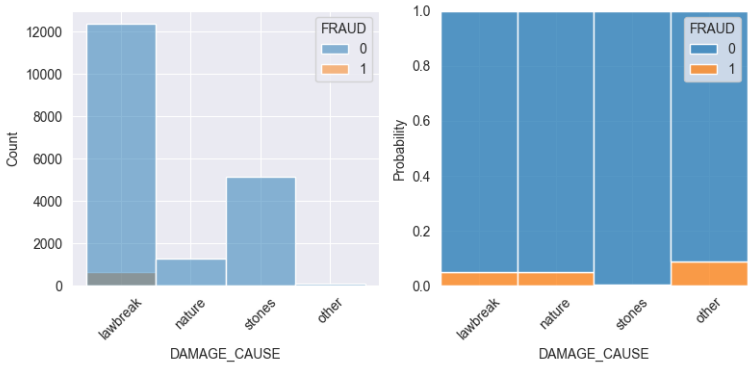

Figure A.2: Distribution of damage cause

Table A.3: Distribution of other binary variables

	POLICE		**WITNESS**		**BODY_INJURIES**	
	0	**1**	**0**	**1**	**0**	**1**
No Fraud	16 451	646	18 576	748	18 822	755
Fraud	2 450	114	325	12	79	5

Appendix B

Models

Table B.1: Model parameters

Model	Parameters
Logistic Regression	C: 0.3, penalty: 'l2', solver: 'liblinear'
Logistic Regression with SMOTE	C: 0.3, penalty: 'l2', solver: 'saga' , smote___sampling_strategy: 0.4
Random Forest	bootstrap: False, class_weight: 'balanced', criterion: 'entropy', max_depth: 3, min_samples_split: 2, n_estimators: 50
Random Forest with SMOTE	bootstrap: False, class_weight: 'balanced', criterion: 'entropy', max_depth: 3, min_samples_split: 2, n_estimators: 100, smote___sampling_strategy: 0.2

Continued on next page

Model	Parameters
Histogram-based Gradient Boosting	class_weight: 'balanced', min_samples_leaf: 50, l2_regularization: 0.5, learning_rate: 0.1, max_depth: 3
Histogram-based GBM with SMOTE	class_weight: 'balanced', min_samples_leaf: 30, l2_regularization: 0.5, learning_rate: 0.1, max_depth: 3, smote__sampling_strategy: 0.2
XGBoost	scale_pos_weight: 24.86, max_depth: 3, colsample_bynode: 0.7, eta: 0.1, gamma: 0.5, reg_lambda: 1.2
XGBoost with SMOTE	scale_pos_weight: 24.86, max_depth: 3, colsample_bynode: 0.7, eta: 0.1, gamma: 0.5, reg_lambda: 1.0, smote__sampling_strategy: 0.4
MLP 1	random_state: 420, hidden_layer_sizes: (32,16), learning_rate_init: 0.3, activation: 'tanh', batch_size: 'auto', early_stopping: True, learning_rate: 'adaptive', solver: 'adam', smote__sampling_strategy: 0.7

Continued on next page

Model	Parameters
MLP 2	random_state: 420, hidden_layer_sizes: (64,32), learning_rate_init: 0.1, activation: 'relu', batch_size: 'auto', early_stopping: True, learning_rate: 'adaptive', solver: 'adam', smote___sampling_strategy: 0.8

Other parameters are kept to their default values.

Table B.2: Full comparison of models' results

Model	Precision	Recall	F1-score	AUC
Logistic Regression	0.09	0.83	0.16	0.75
Logistic Regression with SMOTE	0.09	0.80	0.17	0.74
Random Forest	0.08	0.83	0.14	0.72
Random Forest with SMOTE	0.09	0.76	0.16	0.73
HGBM	0.09	0.81	0.16	0.74
HGBM with SMOTE	0.10	0.75	0.17	0.73
XGBoost	0.12	0.74	0.21	0.76
XGBoost with SMOTE	0.09	0.92	0.16	0.77
MLP (1)	0.06	0.97	0.10	0.65
MLP (2)	0.11	0.70	0.18	0.73