Predictive Modelling for Insurance Claim Probability

Karthika Surendra
School of Computing and Mathematics
University of South Wales

A submission presented in
partial fulfilment of the requirements
of the University of South Wales/Prifysgol De Cymru
for the degree of MSc Data Science

September 30, 2024

UNIVERSITY OF SOUTH WALES

PRIFYSGOL DE CYMRU

FACULTY OF COMPUTING, ENGINEERING & SCIENCE
SCHOOL OF COMPUTING & MATHEMATICS

STATEMENT OF ORIGINALITY

This is to certify that, except where specific reference is made, the work described in this project is the result of the investigation carried out by the student, and that neither this project nor any part of it has been presented, or is currently being submitted in candidature for any award other than in part for the MSc Data Science degree of the University of South Wales.

Signed Karthika Surendra ..... (student)

Date 30 / 09/ 2024 .....

Acknowledgments

I would like to express my sincere gratitude to my parents, Navaretnarajah and Uthaya Tharki, as well as my husband, Surendra, for their unwavering love, support, and encouragement. They instilled in me the importance of pursuing my passions and relentlessly chasing my dreams, which has significantly contributed to my personal growth.

Through undertaking this postgraduate Master's project, I have acquired valuable insights and research experience in the field of Machine Learning within Data Science. I wish to extend my heartfelt appreciation to my project supervisor, Adeesha Gamage, for your steadfast support and guidance. Our weekly meetings have been immensely beneficial, and I am grateful for the constructive feedback that has greatly advanced my project. I will miss the collaborative nature of this work and wish we had more time to delve deeper into the subject matter. I would also like to thank my project sponsor, Jenni Whewell, for your assistance during the challenges I faced in selecting an appropriate dataset and identifying suitable techniques. Your support during that critical time was invaluable.

Lastly, I wish to express my gratitude to the Data Science academic staff at the School of Computing and Mathematics at the University of South Wales. I have thoroughly enjoyed my studies and look forward to applying the skills and knowledge I have gained in my future career.

Abstract

The application of machine learning to predict car insurance claims has rapidly emerged as a transformative strategy within the insurance industry, driven by advancements in data analytics and the availability of extensive datasets, such as the Kaggle.com dataset comprising 10,303 observations and 25 features. This study utilizes this dataset, applying various machine learning techniques and rigorous data preprocessing methods to improve predictive accuracy and address challenges, notably class imbalance. Essential preprocessing steps included data cleaning, feature encoding, feature scaling, and exploratory data analysis, which revealed a significant class imbalance. To rectify this, the Synthetic Minority Over-sampling Technique was employed to enhance model training. Several classification algorithms, including Random Forest, Support Vector Machine, and XGBoost, were implemented, with performance assessed using accuracy and ROC-AUC metrics. Distinct approaches were systematically applied during the model training phase to identify the most effective model for predicting car insurance claims. The results indicate that the Random Forest and SVM algorithms outperformed the others, with Support Vector Machine achieving the highest accuracy of 0.7858 in Approach Two, while Random Forest exhibited an accuracy of 0.7816 in Approach Five. Additionally, Chi-Square tests revealed significant associations between the target variable (claim flag) and features such as education, occupation, and vehicle type. These findings highlight the critical importance of feature selection and preprocessing in developing effective predictive models and provide valuable insights for improving decision-making processes in insurance claim management.

Table of Contents

List of Figures

List of Tables

Chapter 1 Introduction

In the insurance industry, accurate claim prediction is critical for fair pricing. Inaccuracies can lead to higher premiums for low-risk drivers and lower costs for high-risk drivers, with low-risk clients effectively subsidizing the others. Premiums are calculated based on demographics, vehicle specifications, and claims history. Notably, EY estimates that in 2022, UK motor insurers paid £1.11 per £1 received in premiums, a figure expected to rise, underscoring the need for precise predictions (Hooson and Mark, 2024). Accurately predicting insurance claims is vital for effective risk management and pricing. Traditional methods often lead to adverse selection, where low-risk customers subsidize high-risk ones (Abdelhadi, Elbahnasy and Abdelsalam, 2020). Inaccurate claim predictions, particularly in car and health insurance, result in unfair pricing and financial instability, highlighting the need for better predictive models (Pesantez-Narvaez ,Guillen and Alcaniz, 2019) (Christiansen, et al, 2016).

The desired result of this project is to develop a robust predictive model that leverages machine learning techniques to analyse historical data and accurately predict the likelihood of individuals filing a claim after purchasing car insurance. This model will empower car insurance companies to identify potential claimants at the point of purchase, enabling them to implement more precise risk assessments and pricing strategies. Ultimately, this will lead to better resource allocation, enhanced customer satisfaction through fairer pricing, and improved financial outcomes for car insurance providers (Fauzan and Murfi, 2018).

This project aims to develop a machine learning-based predictive model to accurately forecast car insurance claims, enhancing risk assessment and pricing strategies. The model's effectiveness will be measured using metrics like precision, recall, F1 score, and ROC-AUC. The project also seeks to reduce adverse selection by aligning premiums with actual risk levels, improving customer retention and satisfaction. Additionally, the project will optimize resource allocation within insurance companies, leading to cost savings and better financial performance. Success will be measured by comparing misclassified claims, analysing policyholder retention, and assessing customer satisfaction through surveys post-implementation (Hanafy and Ming, 2021).

Accordingly, this project going to address the research questions "How do various algorithms compare in the context of predictive modelling for insurance claim probability, and what is the significance of data preprocessing in improving model performance?"

The research was developed based on five key objectives, as outlined below

- Development of a Robust Predictive Model for Estimating the Probability of Car Insurance Claims Based on Historical Data.
- Evaluation of Various Classification Algorithms for Predicting Car Insurance Claim Probabilities Using Metrics Such as Accuracy and ROC-AUC.
- Addressing Class Imbalance Through the Synthetic Minority Over-sampling Technique.
- Comprehensive Data Preprocessing for Machine Learning Applications.
- Examination of Feature Importance Through Chi-Square Tests and Correlation Metrics.

The scope of this project encompasses applied research in machine learning to improve the prediction of car insurance claims. It addresses the challenge of inaccurate claim prediction, which leads to suboptimal risk management and unfair pricing. The project relies on historical car insurance data, focusing on customer demographics, policy details, and claims history. The primary beneficiaries are car insurance companies, which will gain enhanced risk assessment and pricing tools, while policyholders benefit from fairer pricing. If a company partner is involved, they will provide data and industry expertise, ensuring the model's practical applicability and alignment with industry needs (Abdelhadi, Elbahnasy and Abdelsalam, 2020).

This research builds upon the existing body of knowledge in predictive modelling by integrating advanced machine learning techniques with comprehensive data preprocessing methods specifically tailored for the insurance sector. By addressing the critical issue of class imbalance through the application of the SMOTE, this project enhances the model's predictive accuracy while also mitigating the risks associated with adverse selection. The incorporation of various classification algorithms facilitates a comparative analysis of their performance, thereby enabling the identification of the most effective strategies for predicting car insurance claims. Furthermore, the emphasis on feature importance through statistical testing, including Chi-Square tests, underscores the significance of selecting relevant variables that substantially influence claim outcomes. This multifaceted approach not only aims to provide actionable insights for insurance providers but also contributes to the ongoing discourse on improving fairness and accuracy in risk assessment and pricing within the insurance industry. Ultimately, the findings of this study are anticipated to pave the way for future research, enhancing the understanding and application of machine learning in the context of insurance claim prediction and fostering innovation in risk management practices.

The expected deliverables for this study include a fully developed and trained predictive model in Python, saved for future use. A Python script (Otarb, et al, 2024) will be provided to facilitate the integration of the model into Power BI, where it will be applied to the dataset to generate claim predictions. These predictions, along with the relevant input features, will be used to develop an interactive Power BI dashboard that visually represents the model's outcomes and highlights the factors influencing the claim predictions. Comprehensive documentation will also be provided, detailing the model development process, the integration into Power BI, and instructions for using the dashboard. This integrated approach ensures that stakeholders can efficiently monitor, analyse, and interpret claim predictions in real-time through an intuitive and accessible platform.

The project is constrained by several factors that must be meticulously managed to ensure its successful completion. The quality and availability of the dataset will significantly influence the accuracy and reliability of the predictive model, while the model’s performance is limited by its ability to generalize to new data and scale effectively within Power BI. Furthermore, Power BI's capacity to handle complex Python models and large datasets presents limitations that may impact the efficiency of prediction and visualization processes. Computational resources, including memory and processing power, may also create potential bottlenecks. The project must fulfil stakeholder expectations regarding accuracy and usability, while strictly adhering to data privacy and security regulations. Additionally, the timeline and budget impose limitations on the extent of model tuning and dashboard sophistication. Ensuring compatibility between Python, Power BI, and any associated tools is critical to avoid integration challenges. Effectively managing these constraints is essential to delivering a robust and successful solution.

Chapter 2 provides a comprehensive review of the literature relevant to this project, offering an overview of studies and papers that have utilized machine learning techniques within the realm of insurance claims prediction. Chapter 3 delineates the methodology employed in this research, elaborating on the data preprocessing procedures involved. The dataset utilized in this study is described, covering critical aspects such as data cleaning, feature encoding, feature scaling, and exploratory data analysis, all of which collectively ensure the dataset's readiness for analytical purposes. Furthermore, the implementation of the Synthetic Minority Over-sampling Technique (SMOTE) is discussed to mitigate class imbalance, along with methods for feature selection. Chapter 4 critically examines various machine learning algorithms applied in predictive modelling efforts, including XGBoost, Random Forest, Support Vector Machine, Multi-Layer Perceptron, Linear Regression, K-Nearest Neighbours, and Decision Tree classifiers. The performance of each algorithm is evaluated, assessing their efficacy in predicting insurance claims. This chapter also focuses on the evaluation of the predictive models, analysing their performance metrics, such as accuracy and ROC-AUC, while investigating feature importance associated with each model. Finally, Chapter 5 concludes the dissertation by synthesizing the key findings and proposing potential avenues for future research aimed at augmenting the contributions of this study.

Chapter 2 Literature Review

This project employed several techniques that were utilized both before and during its development to provide essential theoretical and practical support in the car insurance industry. The following is a brief overview of the related papers its benefits and limitation in implementing insurance car claim prediction model. be that offered explanations, examples, and inspiration for the successful completion of the project. A comprehensive understanding of car insurance was essential for this project, with insights derived from the "UK Car Insurance Statistics 2024." As a result, this source has been included in the related work section.

There are three primary types of car insurance policies: Third Party, which covers liability for injuries and damages caused to others; Third Party Fire and Theft, which provides additional protection against theft and fire damage to the vehicle; and Comprehensive, which offers broad coverage, including personal medical expenses and damage to the policyholder's own property. The average annual cost of car insurance varies significantly across different regions of the UK. The cost a driver pays for insurance is influenced by a range of factors, such as age, the make and model of the vehicle (including its value and engine size), years of driving experience, no-claims history, annual mileage, and the type of coverage selected. Additionally, young drivers tend to face higher insurance premiums due to their relative inexperience and a statistically higher likelihood of filing claims. However, they can manage these costs by exploring various premium options and employing effective strategies to reduce their insurance expenses. This information is sourced from the "UK Car Insurance Statistics 2024."(Hooson and Mark, 2024).

The research was conducted to investigate how modern technologies enhanced decision-making within the insurance industry, with a particular focus on advanced claims analysis (Rawat et al., 2021). The transformative role of InsurTech and Big Data in revolutionizing traditional insurance practices was underscored, particularly through the application of artificial intelligence (AI) and machine learning (ML) in areas such as underwriting, risk selection, pricing, and claims processing. Machine learning classification algorithms were employed on insurance datasets, utilizing data preprocessing and feature selection techniques to improve accuracy in distinguishing between fraudulent and legitimate claims. It was demonstrated that these methods significantly enhanced fraud detection, claim prioritization, and overall decision-making, resulting in more efficient operations and improved customer satisfaction. The paper concluded that considerable advantages were offered to insurers through the integration of ML and data visualization techniques, which streamlined processes and enhanced the accuracy of decision making.

The application of multivariate decision tree models for predicting insurance claims was examined in another research study, which demonstrated superior accuracy compared to traditional univariate approaches (Quan and Valdez, 2018). Decision tree models were recognized for their nonparametric nature, ability to manage missing data, detect non-linear interactions, and perform variable selection, making them highly effective for actuarial data analysis. The traditional decision tree method was extended by the authors to address multivariate response variables, which account for dependencies frequently observed in insurance data, such as multiple coverages held by a single policyholder. Additionally, advancements in decision tree methodologies, including random forests and gradient boosting, were reviewed, further enhancing predictive capabilities. By applying multivariate decision tree models to insurance claims data from the Wisconsin Local Government Property Insurance Fund (LGPIF), a marked improvement in predictive accuracy was demonstrated by the study. The findings suggested that these models offered a valuable tool for capturing complex relationships in insurance claims without requiring specific probability distributions.

A critical gap in insurance claim prediction research was addressed by incorporating advanced imputation techniques alongside machine learning models, specifically XGBoost and Decision Tree, to effectively manage

missing data (Abdelhadi, Elbahnasy, and Abdelsalam, 2020). High predictive accuracy was demonstrated by these models, with XGBoost achieving 92.53% and Decision Tree 92.22%, thereby improving the reliability of predictions used for setting insurance premiums. Additionally, a shift from traditional statistical methods to advanced machine learning models, particularly XGBoost, in insurance claim prediction was highlighted by another study (Fauzan and Murfi, 2018). XGBoost was noted for its superior capacity to handle large, complex datasets and missing values, outperforming models such as Random Forest. The effectiveness of XGBoost in enhancing both prediction accuracy and reliability was further validated by this study.

Furter study focused on developing a robust statistical framework to model the frequency and severity of insurance claims (Ng'elechei, et al, 2020). It was addressed the complexities of premium pricing, which are exacerbated by the inherent variability in claim data. The analysis was utilized historical motor insurance claim data, applying specific statistical distributions: the Negative Binomial distribution to model claim frequency and the Pareto distribution to model claim severity. Parameters for these distributions are estimated using Maximum Likelihood Estimation (MLE). The goodness-of-fit for the models is rigorously assessed through chi-square and Anderson-Darling tests. These validated models are then used to project future claim amounts. The research highlights the critical importance of selecting appropriate statistical models to ensure accurate prediction and effective insurance premium determination

A further study was presented in a formal review of customer churn prediction methods within the telecom industry, where churn was defined as the termination of service due to dissatisfaction or the availability of superior alternatives (Ahmed and Linen, 2017). The necessity of accurate churn prediction for retaining at-risk customers was emphasized. Techniques such as support vector machines (SVM), artificial neural networks (ANN), and meta-heuristic algorithms were employed to enhance prediction accuracy. The use of balanced datasets, achieved through methods such as SMOTE, along with effective feature selection, was identified as critical. Hybrid models, which integrated multiple algorithms, were generally found to outperform single models. It was suggested that future research should prioritize improvements in feature selection, address data imbalance, and explore advanced hybrid methods.

Another research study was conducted to compare the predictive performance of XGBoost and logistic regression in forecasting motor insurance claims using telematics data, which includes real-time driving behaviours such as distance driven and urban driving percentage (Pesantez-Narvaez, Guillen, and Alcaniz, 2019). Logistic regression was valued for its interpretability and simplicity, rendering it suitable for regulatory environments where transparency is essential. In contrast, XGBoost was found to offer slightly higher predictive accuracy but was noted to be more complex and less interpretable, requiring extensive tuning. The importance of telematics data in creating fairer insurance premiums through Pay-As-You-Drive (PAYD) schemes was highlighted by the research. While potential benefits of XGBoost were recognized, the study concluded that logistic regression remains a strong candidate for practical insurance applications due to its transparency. The authors recommended that a hybrid model should be explored to combine the strengths of both methods, balancing predictive accuracy with the interpretability needed for regulatory compliance.

The paper titled "Machine Learning in P&C Insurance: A Review for Pricing and Reserving" was presented to offer a detailed examination of the integration of machine learning (ML) in Property and Casualty (P&C) insurance, with a focus on pricing and reserving (Christopher et al., 2020). Historically, linear models such as Generalized Linear Models (GLMs) were relied upon in actuarial science. However, recent advancements in ML, including neural networks and enhanced computational resources, enabled the modelling of complex nonlinear relationships and interactions. The ability of ML to handle both structured and unstructured data, such as text and images, was highlighted without extensive manual feature engineering. A comprehensive review of literature up to August 2020 was conducted, encompassing peer-reviewed papers and industry reports. It was noted that a growing but still limited application of ML in actuarial science existed, particularly in pricing, where ML adoption was more established compared to reserving, which remained less common.

Challenges in ML adoption were identified, including the need for fostering a data-centric culture, providing ongoing education for actuaries, and addressing ethical issues. A call for continued research was made to refine ML models and integrate them with traditional actuarial practices, emphasizing the transformative potential of ML in enhancing pricing and reserving within the insurance sector.

The similar study was conducted to investigate the application of Back Propagation Neural Networks (BPNN) for forecasting motor insurance claims, addressing key challenges such as handling large, uncertain datasets and the limitations of traditional statistical methods in dealing with extreme values (Yunos et al., 2016). The research focused on issues of data quality, particularly the presence of noise and missing values, and emphasized the superior ability of Artificial Neural Networks (ANNs) to manage complex, non-linear relationships within the data. The development of the BPNN model was carried out through a rigorous process that involved the careful selection of input variables, data normalization, and extensive training using the backpropagation algorithm. The model was fine-tuned through a process of trial and error to optimize learning rates and momentum. During the experimental phase, nine different network structures were evaluated across two categories of claims: third-party property damage (TPPD) and own damage (OD). The results demonstrated that BPNN models were effective, with the best models achieving low error rates across various evaluation metrics. It was concluded that BPNN is a robust and reliable tool for predictive modelling in the motor insurance sector. The research was supported by the Ministry of Higher Education of Malaysia and the UTM Big Data Centre, providing a solid foundation for the study's findings.

The article titled "Machine Learning Approaches for Auto Insurance Big Data" was published to explore the application of machine learning (ML) models for enhancing risk assessment and premium pricing in the auto insurance industry (Hanafy and Ming, 2021). Six ML models, including Logistic Regression, XGBoost, Random Forest, Decision Trees, Naïve Bayes, and K-Nearest Neighbours, were employed using a dataset from Porto Seguro, a Brazilian insurance company, which comprised 1.5 million observations and 59 variables. The performance of the models was assessed using metrics such as accuracy, kappa statistics, and the Area Under the Curve (AUC). It was found that Random Forest (RF) demonstrated superior performance, while Naïve Bayes was identified as the least effective model. The study also reviewed previous research on ML applications in insurance, noting the varying success of models such as Neural Networks and Support Vector Machines. In conclusion, it was asserted that Random Forest is a highly effective tool for predicting insurance claims, enabling insurers to offer more accurate and tailored premiums, thereby improving customer service and accessibility in the industry. However, the paper could be improved by further analysing the Logistic Regression, XGBoost, Random Forest, Decision Trees, and K-Nearest Neighbours models, along with their ROC-AUC metrics, particularly in terms of computational efficiency and ease of implementation in the current study.

The paper was conducted to assess the effectiveness of five machine learning models—Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), and XGBoost—in predicting car insurance claims (Li, 2023). The dataset underwent thorough pre-processing, which included normalization, encoding, and the application of the Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance. Feature selection was rigorously performed using ANOVA and Chi-squared tests, ensuring that only the most relevant variables were retained. Among the models evaluated, the Random Forest model was found to be the most effective, offering a well-balanced performance across precision, recall, and overall accuracy, thereby making it particularly reliable for predicting both the occurrence and non-occurrence of claims. Although the highest accuracy was achieved by XGBoost, its lower recall indicated a higher rate of false negatives, which limited its overall effectiveness. The study concluded that the Random Forest model is the optimal tool for predicting car insurance claims, providing insurance companies with a powerful mechanism to improve risk management and premium setting. This research underscored the significant potential of machine learning, particularly the Random Forest model, in enhancing the accuracy and reliability of insurance claim predictions. This aspect is intriguing and could be effectively integrated into

the current study to enhance its scope. Incorporating techniques such as Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), Multi-layer Perceptron (MLP), and XGBoost, along with data preprocessing methods like the Synthetic Minority Oversampling Technique (SMOTE) and statistical assessments through Chi-squared tests, would provide a comprehensive understanding of the subject matter. This integration has the potential to uncover new dimensions in the analysis and improve the conclusions regarding insurance claim predictions. This is interesting that could also consider incorporate to the current study.

This project builds upon these insights by leveraging the strengths of the Random Forest, Linear Regression, and Support Vector Machine classifiers to develop a robust predictive framework for car insurance claims. By addressing the gaps identified in the literature, this research aims to contribute to improved risk assessment, enhanced pricing strategies, and optimized resource allocation within the insurance industry. Ultimately, these efforts seek to facilitate more informed decision-making and promote greater customer satisfaction.

Chapter 3 Methodology

3.1 Dataset Description

In this project, a machine learning model was developed using Python to predict car insurance claims. The dataset used in this study was sourced from Kaggle (Jha and Mukul Kumar, 2020).com and comprised 25 features across 10,303 observations. Each observation represented a distinct policyholder and their respective vehicle. The dataset included a mix of numerical, categorical, and binary variables. Numerical features included variables such as the Claim id, the number of children who drive, the policyholder's age, the number of children residing at home, years on the job, annual income, the market value of the policyholder's home, and policy tenure. Binary variables included indicators such as whether the policyholder is a single parent, marital status, gender, and whether the car is a red vehicle. Categorical variables included education level, job title, travel time, the purpose of the car, the vehicle's book value, car type, the amount of prior claims, claim frequency, policy revocation status, motor vehicle record points, claim amount, and the age of the car. The target variable, claim flag, indicated whether a claim had been made by the policyholder.

3.2 Ethical Considerations

The Kaggle dataset was utilized in accordance with its licensing agreements, ensuring compliance with ethical standards. Acknowledgment of limitations, including potential biases and gaps in the data, was made, as these factors may affect the analysis's generalizability.

3.3 Experimental set-up

The current project was developed using the Python programming language, along with its extensive libraries, within the Jupiter Notebook environment, which facilitated efficient data analysis and visualization. Additionally, Power BI was utilized to create an interactive dashboard, allowing stakeholders to explore and interpret the data intuitively. This combination of Python and Power BI provided a robust framework for analysing complex datasets and effectively communicating insights

3.4. Data Preprocessing

The methodology employed in this study was illustrated in Figure 1.

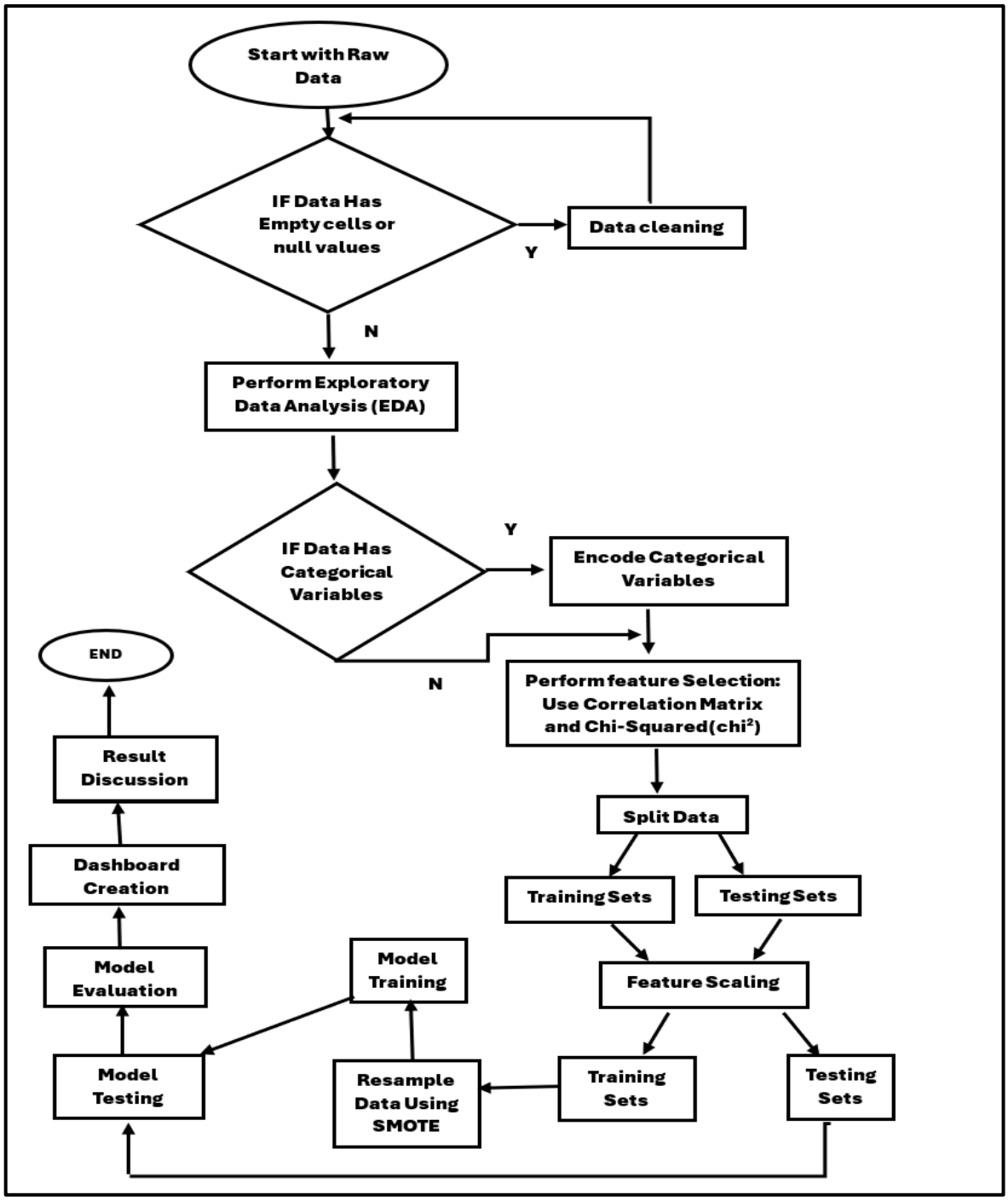

Figure 1:Flowchart of methodology used in developing machine learning models for car insurance

3.4.1. Data Cleaning

The initial step was to identify the number of missing entries in each column of the dataset. Out of the 10,302 instances (exoplanets), it was found that only five features had missing data. To ensure that Machine Learning tasks could be effectively performed, the issue was addressed by imputing the missing values in four numerical columns with the mean value of each respective column. For the categorical column with missing data, all instances containing at least one missing entry were removed. As a result, the dataset was reduced to 9,637 instances, reflecting a removal of just over 6% of the original data. Consequently, nearly 94% of the instances retained complete data for all numerical and categorical features, thus providing a robust dataset for further analysis (Rawat, et al, 2021).

3.4.2. Feature Encoding

In this study, label encoding was utilized to convert categorical and Boolean features into numerical labels for constructing the correlation matrix, which facilitated the quantification of relationships between features. For the classification model, one-hot encoding was applied to both categorical and Boolean features to avoid introducing unintended ordinal relationships. This approach improved the model's ability to process the data effectively. Accurate categorical feature encoding is vital in machine learning as it ensures that algorithms can properly interpret the data, maintains the integrity of information, reduces biases, and enhances overall model performance and interpretability. In summary, encoding is critical for developing accurate, interpretable, and efficient models, preventing overfitting, and ensuring the proper use of categorical data.

3.4.3. Feature Scaling

The Standard Scaler was employed in this study during classification tasks, particularly for algorithms such as Support Vector Machines (SVM) and Logistic Regression, which rely on distance calculations. By standardizing features through mean subtraction and scaling to unit variance, it ensures equal contribution from each feature. This process enhances model performance by preventing features wit larger ranges from dominating the analysis and accelerates convergence in gradient-based methods. Additionally, it normalizes feature importance, thereby improving the efficacy of regularization. Although not necessary for tree-based models, the Standard Scaler is essential when features have varying scales, ensuring balanced and accurate classification outcomes.

3.4.4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) (Rawat, et al, 2021) is a critical step in data science, especially in the context of insurance claim prediction. It involves the preliminary examination of a dataset to uncover patterns, identify anomalies, test hypotheses, and validate assumptions. EDA is crucial for understanding the dataset's underlying structure, which informs the prediction of insurance claims by summarizing key characteristics through statistical graphics and visualizations. Figure2 presents scatter plots that visualize pairwise relationships between numerical features in the dataset, revealing interactions that could impact claim predictions. Each point is color-coded according to the claim flag, allowing us to observe how the occurrence or absence of a claim is distributed across feature pairs. This visual thus playing a significant role in refining predictive models and enhancing decision-making in the insurance industry.

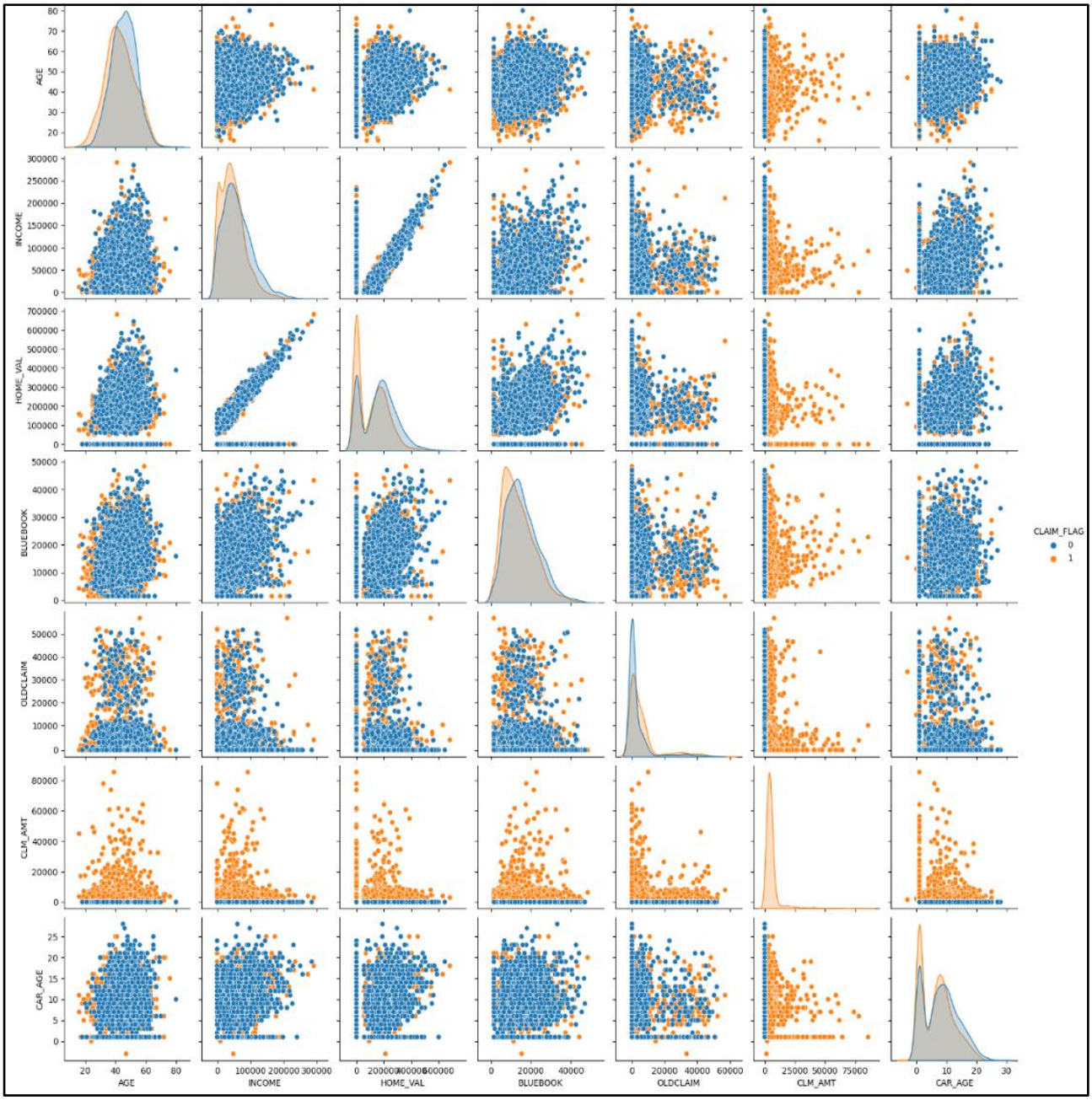

Figure 2:Scatter plots that visualize pairwise relationships between numerical features

Figure 3 presents histograms for the dataset's features, comprising 9 numerical features, with the remainder being categorical. The AGE feature is particularly notable, displaying an almost bell-shaped histogram that suggests an approximately normal (Gaussian) distribution, making it suitable for parametric methods. In contrast, the other numerical features demonstrate skewness or multiple peaks, akin to the categorical features, indicating a variety of distribution patterns. This variability underscores the need for distinct analytical approaches, particularly when addressing the non-normality present in the data.

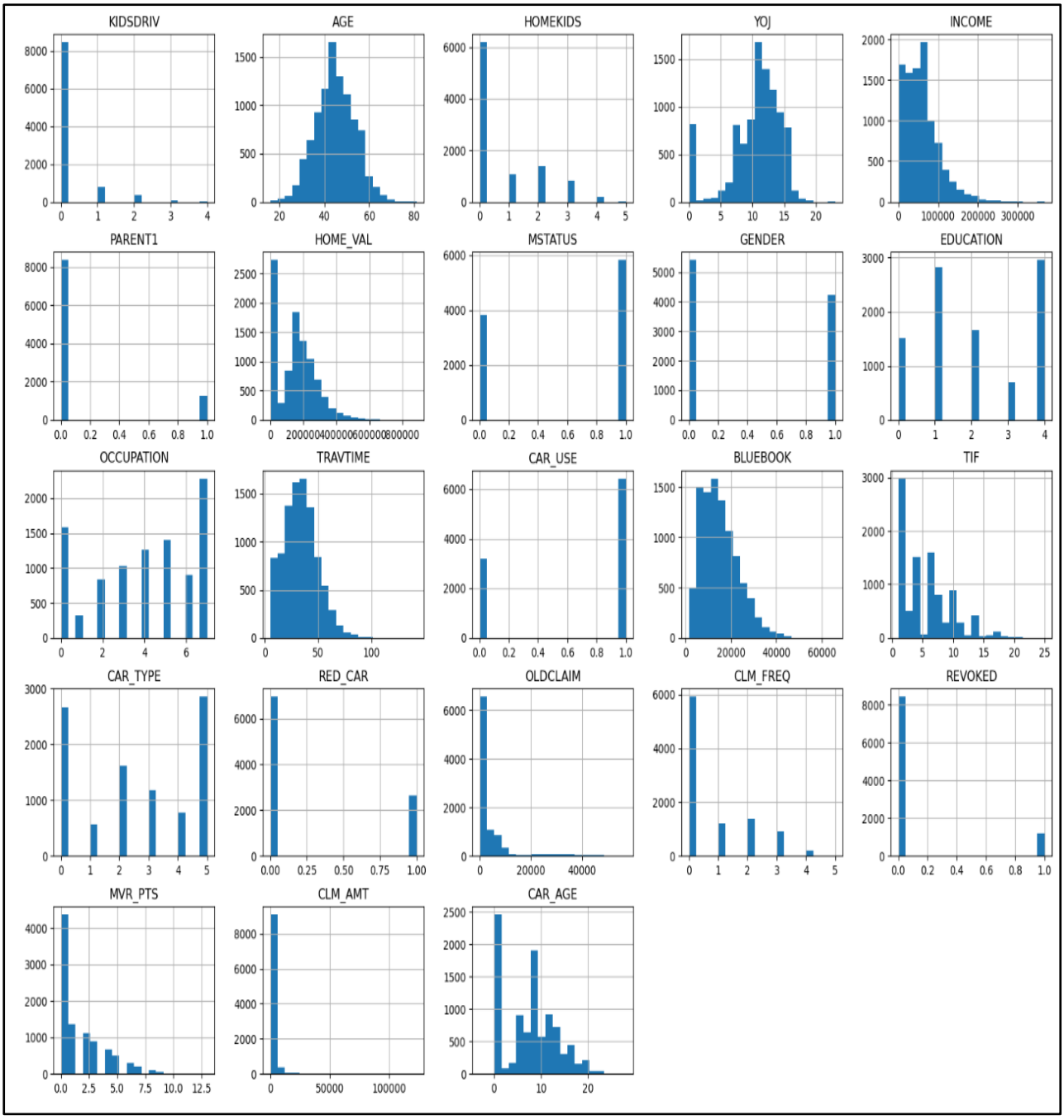

Figure 3:Histograms of 23 Data Features.

In Figure 4, a comparison of claim status across various features is presented, and several key patterns are highlighted. A higher claim percentage is observed among customers aged 16 to 27 compared to those with no claims. Certain car types are associated with higher claim percentages, with commercial vehicles particularly noted for their elevated rate of claims. Additionally, higher claim percentages are seen among individuals in occupations with more frequent car usage.

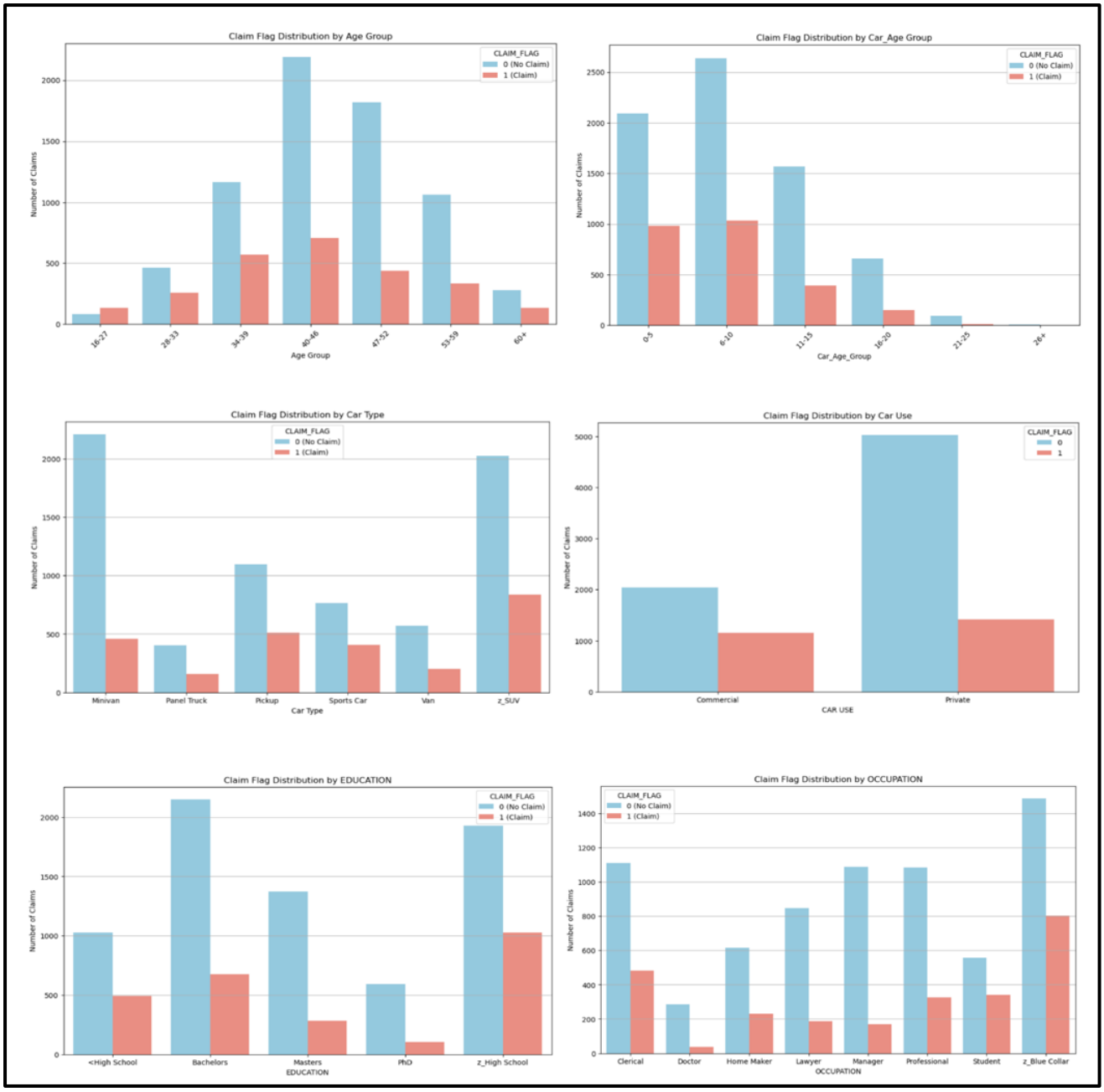

Figure 4:Comparison of Claim Status Across Different Feature

3.4.5. Synthetic Minority Over-sampling Technique (SMOTE)

Figure 5 illustrates a balancing plot for the claim flag, providing a visual assessment of the distribution and balance of claim outcomes within the dataset. The plot reveals a notable imbalance, with 73% of instances showing no claim made and only 27% involving a claim. This imbalance is critical as it can impact the accuracy and fairness of subsequent analyses and modelling, necessitating corrective measures. SMOTE offers an effective solution to this issue (Li, 2023), (Ahmed and Linen, 2017). It addresses class imbalance by generating synthetic samples for the minority class rather than duplicating existing data. By creating new instances based on the nearest neighbours of existing minority class samples, SMOTE enhances the balance of the dataset. This technique improves the model's ability to predict minority class instances accurately, reducing bias and overfitting. SMOTE is particularly beneficial in insurance claim prediction, where precise prediction of the minority class is crucial for effective decision-making and risk management. Figure 6 illustrates the balancing plot for the claim flag after applying SMOTE, demonstrating the improved balance in the dataset.

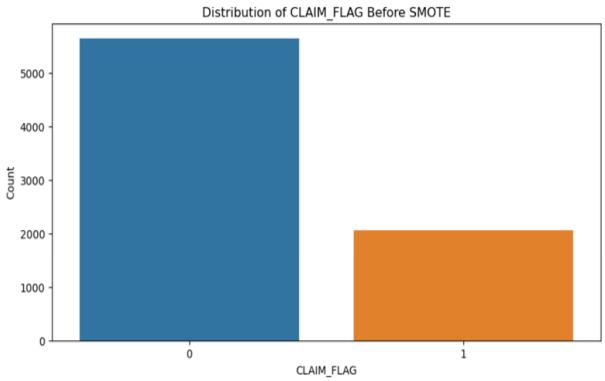

Figure 6:Distribution of Claim Flag Before SMOTE

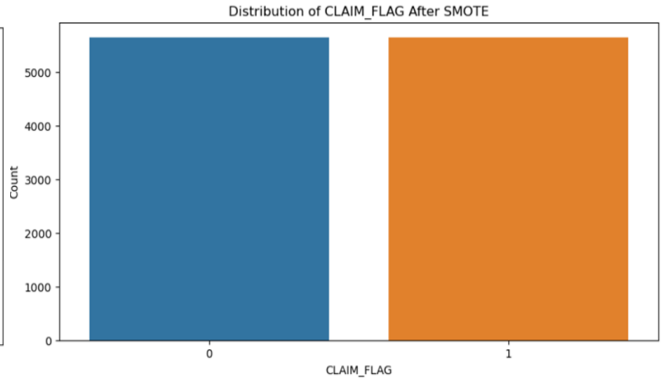

Figure 5:Distribution of Claim Flag After SMOTE

3.4.6. Feature selections

A correlation matrix was utilized to identify the relationships between numerical and Boolean variables and the target variable, *Claim Flag*. Figure 7 presents this matrix, depicting the strength and direction of these relationships. The colour scale ranges from red (indicating a positive correlation) to blue (indicating a negative correlation), with lighter shades representing weaker correlations. Most of the correlations appear in lighter shades, signifying generally weak relationships among the variables. Understanding these correlations is essential for feature selection in predictive modelling, as it reveals potential interactions between demographic factors, vehicle characteristics, and insurance claims.

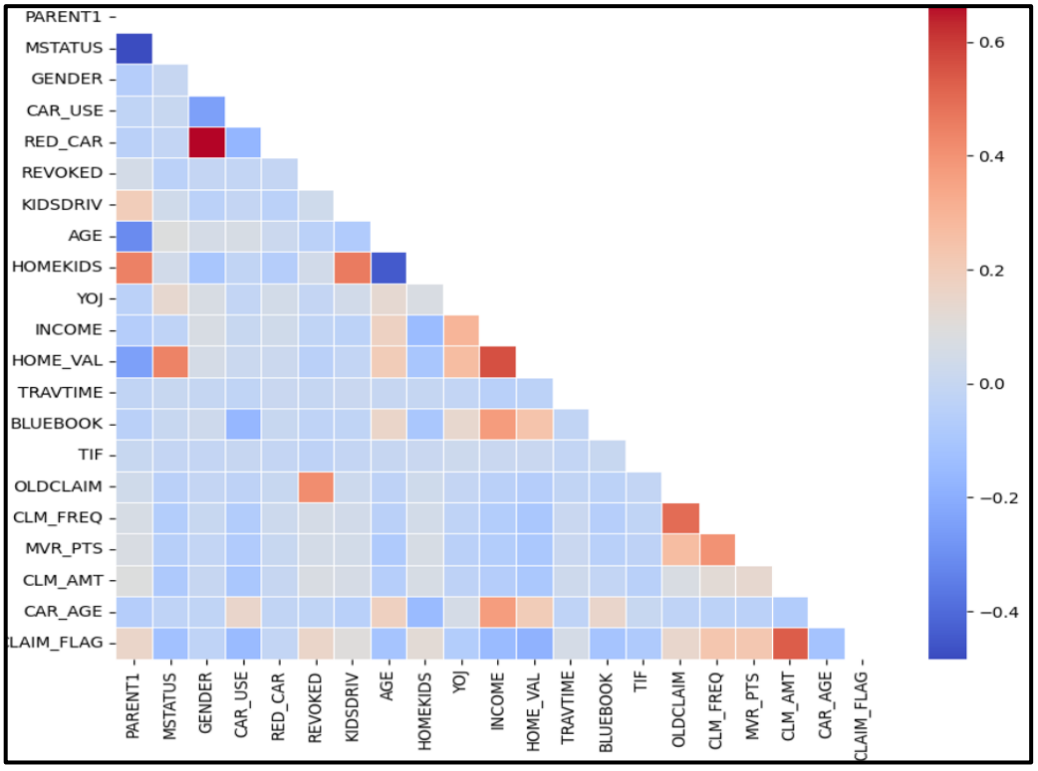

Figure 7:Correlation Matrices

Table 1 highlights pairs of variables with notably high correlation values, both positive and negative. Only those pairs with significant correlations were considered. A correlation threshold of 0.8 was discussed as a criterion for identifying highly correlated variables. However, as all correlation values were below 0.8, no decisions were made to remove any variables from the dataset at this stage.

Table 1: Highly Correlated Variable Pairs and Their Corresponding

Paired Relationships Between Variables		Values of correlation
PARENT1 with HOMEKIDS	Positive correlation	0.45
MSTATUS with HOME_VAL		0.44
RED_CAR with GENDER		0.67
OLDCLAIM with REVOKED		0.42
HOMEKIDS with KIDSDRIV		0.46
HOME_VAL with INCOME		0.56
BLUEBOOK with INCOME		0.38
CAR_AGE with INCOME		0.37
CLM_FREQ with OLDCLAIM		0.50
MVR_PTS with CLM_FREQ		0.40
CLM_AMT with CLAIM_FLAG		0.53
PARENT1 with MSTATUS	Negative correlation	-0.48
PARENT1 with AGE		-0.31
CAR_USE with GENDER		-0.25
HOMEKIDS with AGE		-0.45

```
CLAIM_FLAG     1.000000
CLM_AMT        0.532571
CLM_FREQ       0.230024
MVR_PTS        0.228257
PARENT1        0.156701
REVOKED        0.153644
OLDCLAIM       0.143939
HOMEKIDS       0.119880
KIDSDRIV       0.102568
CAR_TYPE       0.094852
OCCUPATION     0.069513
TRAVTIME       0.054102
EDUCATION      0.045958
RED_CAR       -0.011999
GENDER        -0.024079
YOJ           -0.067366
TIF           -0.080072
AGE           -0.110485
BLUEBOOK      -0.110849
CAR_AGE       -0.115493
MSTATUS       -0.127247
CAR_USE       -0.149994
INCOME        -0.152756
HOME_VAL      -0.184198
Name: CLAIM_FLAG, dtype: float64
```

Figure 8:Correlation Between Claim Flag and other Variables in the Dataset.

Significant categorical variables with CLAIM_FLAG:

	Variable	Chi-Square Stat	p-value
0	EDUCATION	262.913440	1.074350e-55
1	OCCUPATION	348.965337	2.051992e-71
2	CAR_TYPE	194.157017	5.047552e-40

	Variable	Chi-Square Stat	p-value
0	PARENT1	235.595610	3.590010e-53
1	MSTATUS	155.451568	1.115612e-35
2	GENDER	5.478147	1.925566e-02
3	CAR_USE	216.095643	6.427727e-49
4	RED_CAR	1.327491	2.492521e-01
5	REVOKED	226.439901	3.562597e-51

Figure 9: Significant Categorical and Boolean Variables and Their Statistical Association with Claim Flag (Chi-Square Test)

The Chi-Square Test is crucial for analysing categorical and Boolean variables, as it evaluates relationships and data distribution *(Ng'elechei, et al, 2020)*. For categorical variables, it compares observed and expected frequencies to identify significant associations. In Boolean scenarios, it assesses whether the occurrence of one outcome is independent of another factor. This test is advantageous because it does not assume a specific data distribution, offering a versatile and straightforward method for examining relationships within categorical datasets. Figure 9 provides a detailed summary of the Chi-Square test results used to assess the relationship between various categorical and Boolean variables with claim flag. It is evident that the categorical variables education, occupation, and car type exhibit extremely low p-values and high Chi-Square statistics, highlighting their strong statistical relationships with claim flag. Significant associations with claim flag are also observed for Boolean variables such as parent1, mstatus, car use, and revoked, as indicated by very low p-values and substantial Chi-Square statistics. In contrast, gender and red car demonstrate weaker associations,

with higher p-values reflecting reduced significance. These findings suggest that education, occupation, car type, parent1, mstatus, car use, and revoked are particularly influential in predicting claim flag, offering valuable insights for further analysis and model development.

Figure 8 illustrates the varying degrees of correlation with the claim flag variable, encompassing both strong positive and weak negative associations. It is observed that red car and gender exhibit very weak correlations with claim flag. Additionally, Chi-Square test results confirm that red car and gender show very weak associations with claim flag. Therefore, red car and gender have been excluded from the dataset for subsequent analysis and model development.

During the exploratory data analysis (EDA), Figure 10 presents the Box Plot and Scatter Plot of Claim Amount versus Claim Flag. These visualizations clearly illustrate a significant relationship between claim amounts and the occurrence of claims. Specifically, for instances where claim flag equals 0 (indicating no claim), the claim amount is consistently zero. Conversely, for claim flag equal to 1 (indicating a claim was made), a range of non-zero claim amounts is observed. The presence of outliers suggests that while most claim amounts are moderate, some are exceptionally high. This analysis indicates a perfect correlation between claim amount and claim flag: claim amounts are always zero when no claim is made and non-zero otherwise.

Including claim amount in a predictive model would yield 100% accuracy in predicting claim flag, as this feature directly determines the outcome. However, such inclusion would result in data leakage, undermining the model's validity for generalization (Krishna, 2024). To prevent artificially inflating the model's performance and to ensure meaningful predictions, claim amount was excluded from the dataset. This decision preserves the integrity of the model by avoiding the direct learning from a feature that is perfectly correlated with the target variable.

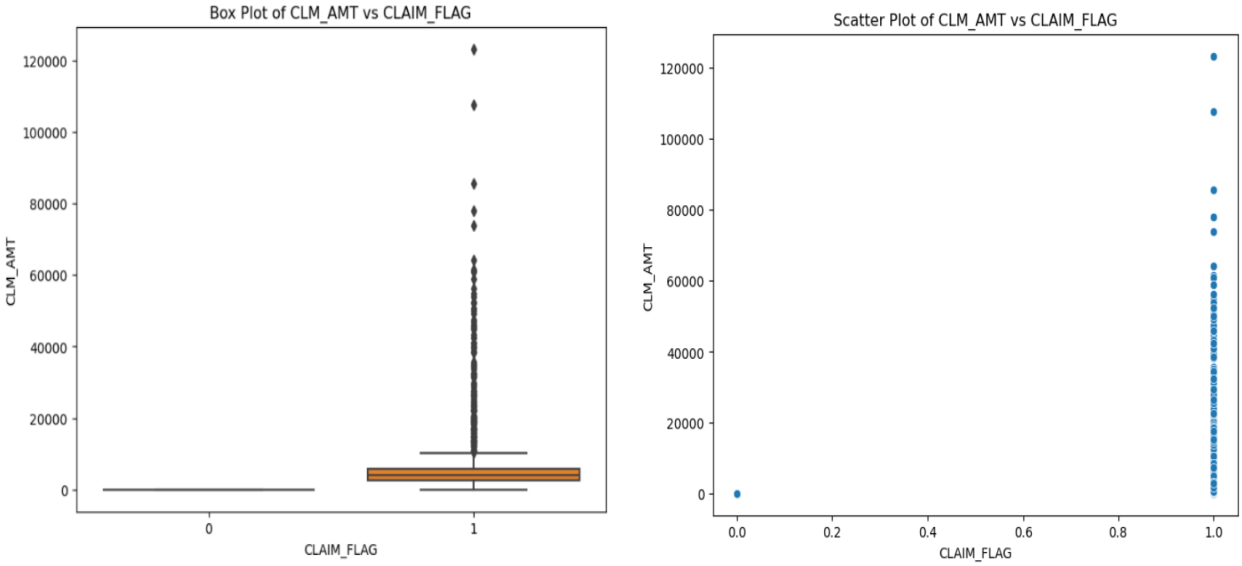

Figure 10:Box Plot and Scatter Plot of Claim Amount vs. Claim Flag

3.5. Building Predictive Models

Figure 11 illustrates the selected features and their corresponding data types for model development. The dataset was partitioned into 80% for training and 20% for testing. A Standard Scaler was applied to the numerical variables in both the training and testing sets to ensure consistent feature scaling. The model was trained using several machine learning algorithms, including Random Forest, XGBoost, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Linear Regression, K-Nearest Neighbours (KNN), and Decision Tree classifiers. Each algorithm was subsequently evaluated to determine the best-performing model for the dataset.

KIDSDRIV	int64
AGE	float64
HOMEKIDS	int64
YOJ	float64
INCOME	float64
PARENT1	bool
HOME_VAL	float64
MSTATUS	bool
GENDER	bool
EDUCATION	category
OCCUPATION	category
TRAVTIME	int64
CAR_USE	bool
BLUEBOOK	float64
TIF	int64
CAR_TYPE	category
RED_CAR	bool
OLDCLAIM	float64
CLM_FREQ	int64
REVOKED	bool
MVR_PTS	int64
CAR_AGE	float64

Figure 11:Selected Features and Corresponding Data Types for Model Development

3.5.1. XGBoost classifier

XGBoost (Extreme Gradient Boosting) is a highly effective machine learning algorithm used for classification and regression tasks. It builds an ensemble of decision trees sequentially, where each tree corrects the errors of its predecessors, thereby enhancing overall accuracy and reducing overfitting. XGBoost is adept at handling large datasets and supports both numerical and categorical data. It manages missing values internally, minimizing the need for extensive preprocessing. The algorithm incorporates regularization techniques to improve generalization and mitigate the risk of overfitting. Furthermore, XGBoost provides insights into feature importance, enabling users to identify the most significant variables affecting predictions. By fine-tuning various hyperparameters, such as learning rate and tree depth, XGBoost consistently delivers accurate and robust results across a wide range of predictive modelling applications. Its efficiency and flexibility make it a popular choice among data scientists and researchers.

Objective Function:

The objective function in XGBoost is constructed to minimize a combination of the loss function and a regularization term. The loss function L(θ) quantifies the error between predicted values and actual values, while the regularization term $\Omega(f_k)$ penalizes model complexity to prevent overfitting. The objective function is mathematically expressed as:

$$\text{Obj}(\theta) = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$ (Fauzan and Murfi, 2018)

Where:

- y_i denotes the true labels

- $\hat{y}_i$ denotes the predicted value,
- $\Omega(f_k)$ is the regularization term.

Gradient Boosting Process:

The gradient boosting process utilized by XGBoost involves the iterative construction of decision trees, each focusing on correcting residual errors from preceding trees. The residuals at iteration t, denoted as r_i^t are computed as:

$$r_i^t = y_i - \hat{y}_i^t \quad \text{(Chen and Guestrin, 2016)}$$

These residuals, representing the differences between true labels and predicted values at the current iteration, are used to train a new decision tree. The predictions are updated in each iteration, and this iterative process continues until the residual errors are minimized, thereby enhancing model accuracy.

Tree Pruning:

Model complexity in XGBoost is managed through tree pruning, which limits the growth of decision trees according to specific criteria. Pruning is controlled by setting a maximum depth for the trees and employing early stopping mechanisms. The decision to prune is based on the gain in loss reduction, calculated as:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right]\text{-}\gamma \quad \text{(Chen and Guestrin, 2016)}$$

Where:

- G_L and G_R are the gradients for the left and right splits,
- H_L and H_R are the Hessians (second derivatives),
- and λ and γ are regularization parameters.

Overall, the advanced features of XGBoost, such as regularization, missing data handling, parallel computations, and adjustments for class imbalance, make it exceptionally suited for predicting car insurance claims. The model can capture complex relationships between features and target variables, resulting in accurate and robust predictions.

3.5.2. Random Forest classifier

Random Forest is a powerful machine learning algorithm frequently employed for classification and regression tasks (Schott, 2019). It operates by constructing multiple decision trees during training and aggregating their predictions, which enhances accuracy and reduces overfitting. Each tree is built using a random subset of the training data and a selection of features, fostering diversity that improves model generalization. A notable advantage of Random Forest is its ability to handle both numerical and categorical data, making it suitable for various applications. The algorithm is resilient to noise and effectively manages missing values. Furthermore, it provides insights into feature importance, allowing users to identify the most significant predictors. Hyperparameters, such as the number of trees and the maximum depth of each tree, can be fine-tuned to optimize performance (Hanafy and Ming, 2021). Overall, Random Forest is recognized for its accuracy, flexibility, and ease of use, making it a favoured choice among data scientists and researchers.

Hyperparameters and Calculations

During the training process, several key hyperparameters were carefully fine-tuned to optimize the performance of the Random Forest model:

Number of Trees (n_estimators): The number of decision trees in the forest was set to 100, a common starting point in Random Forest models. While increasing the number of trees can generally enhance model performance, it also leads to a higher computational cost.

Maximum Depth (max_depth): The maximum depth of each tree was limited to 10, preventing overfitting by restricting the growth of the trees.

Minimum Samples Split (min_samples_split): This parameter was set to 5, ensuring that each node must contain at least 5 samples before it can be further split. This constraint helps avoid overly complex trees.

Criterion: The model used Gini Impurity as the criterion for evaluating the quality of splits. Gini Impurity is calculated as:

$$G = 1 - \sum_{i=1}^{C} P_i^2$$ (Schott, 2019)

where p_i represents the proportion of samples in the i^{th} class (CLAIM_FLAG). A lower Gini Impurity value indicates purer nodes within the decision trees.

3.5.3. Support Vector Machine classifier

Support Vector Machines (SVM) represent a robust and versatile machine learning algorithm, widely applied in classification, regression, and outlier detection tasks. The primary goal of SVM is to identify the optimal hyperplane that maximizes the margin between two classes. The data points closest to this hyperplane, referred to as support vectors, play a critical role in defining its position. For datasets that are not linearly separable, SVM utilizes the kernel trick, which projects the data into a higher-dimensional space to facilitate better separation. Moreover, SVM incorporates a soft margin to handle noisy or overlapping data, thereby improving its applicability in real-world settings.

SVM is particularly well-suited for high-dimensional data, with the ability to manage both linear and non-linear relationships through the use of different kernel functions (IBM, 2023). A linear kernel is applied in cases where the data is linearly separable, while non-linear kernels such as the radial basis function (RBF) are used to capture more complex patterns. In this study, the flexibility of SVM proved advantageous in modelling intricate relationships between demographic, vehicle-related, and claim history features.

The decision to employ SVM in this study was motivated by its capacity to efficiently process high-dimensional datasets, particularly in scenarios where the number of features exceeds the number of available data points. Additionally, SVM is less prone to overfitting, especially when proper regularization techniques are applied. This property makes it particularly effective in managing imbalanced datasets, such as those encountered in car insurance claim predictions, where claims are significantly fewer than non-claims.

SVM's decision function calculates the weighted sum of the kernel function applied to the support vectors, determining whether an input is classified as a claim or non-claim. This approach enabled the model to generate accurate and balanced predictions, thereby demonstrating the effectiveness of SVM in handling complex classification tasks such as insurance claim prediction.

3.4.4. Multi-Layer Perceptron classifier

A Multi-Layer Perceptron (MLP) classifier is a type of neural network used for classification tasks. It consists of an input layer, one or more hidden layers, and an output layer, where each neuron is fully connected to the next layer (Bento, 2021). MLP uses activation functions like ReLU or Sigmoid to introduce non-linearity and backpropagation to minimize errors by adjusting weights. It is effective for learning complex, non-linear

patterns but can be computationally expensive and prone to overfitting if not properly tuned. MLP is widely applied in tasks like image recognition and text classification.

3.5.5. Linear Regression classifier

Linear Regression is a basic algorithm that models the relationship between input features and a continuous target variable using a straight line (Fauzan and Murfi, 2018). It predicts the target by minimizing the error between the predicted and actual values, typically through the least squares method. Though primarily used for regression, it can be adapted for binary classification by applying a threshold. However, it is less effective for classification tasks compared to models like logistic regression, as it struggles with complex, non-linear data. Linear Regression is simple to implement but performs best when the data shows a clear linear relationship.

3.5.6. K-Nearest Neighbours classifier

K-Nearest Neighbours (KNN) is a simple classification algorithm that assigns a class to a data point based on the majority class of its nearest k neighbours (Hanafy and Ming, 2021). It uses distance metrics like Euclidean distance to measure proximity. As a non-parametric and lazy learning algorithm, KNN makes no assumptions about the data distribution and does not build a model, instead classifying points at runtime. While easy to implement and effective for small datasets, KNN can be computationally expensive for large datasets and sensitive to the choice of k and distance metrics.

3.5.7. Decision Tree classifier

A Decision Tree classifier is a supervised learning algorithm that classifies data by recursively splitting it based on feature values(Quan and Valdez, 2018). It builds a tree-like structure where internal nodes represent features, branches represent decisions, and leaf nodes represent class labels. Splitting is based on criteria like Gini impurity or information gain, aiming to separate the classes effectively. Decision Trees are easy to interpret and can handle both numerical and categorical data. However, they are prone to overfitting, especially with deep trees, and can be sensitive to small changes in the data. Proper tuning is required for optimal performance.

During the model training process, several distinct approaches were methodically applied to identify the most effective model for car insurance claim prediction. These approaches were designed to address challenges associated with categorical, numerical, and imbalanced data, ensuring comprehensive data preprocessing.

Approach 1: The OneHot Encoder was employed to encode both categorical and boolean features. After encoding, the model was trained on the transformed dataset.

Approach 2: The second approach also utilized OneHot Encoder to encode categorical and boolean features, while the Standard Scaler was applied to normalize numerical features, ensuring all variables were on a comparable scale before model training.

Approach 3: The third approach introduced SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance in the dataset. OneHot Encoder was used for categorical and boolean features, Standard Scaler for numerical features, and SMOTE for balancing the dataset, followed by model training.

Approach 4: In this approach, Label Encoder was used to encode categorical and boolean features. The encoded data was then used for model training.

Approach 5: In the fifth approach, Label Encoder was applied to categorical and boolean features, while the Standard Scaler was used for normalizing numerical features before the model was trained.

Approach 6: In the final approach, Label Encoder was employed for categorical and boolean features, Standard Scaler was applied for numerical scaling, and SMOTE was used to address data imbalance before the model training process.

Each of these approaches was systematically evaluated to assess the impact of different encoding techniques, scaling methods, and strategies for handling imbalanced data. The ultimate objective was to determine the most accurate and robust model for predicting car insurance claims.

3.6. Model Evaluation

In this study, the evaluation of the model utilized various metrics, including Accuracy, Precision, Recall (True Positive Rate), F1-Score, Cross-Validation Accuracy, and Receiver Operating Characteristic (ROC) analysis. These techniques collectively offer a thorough assessment of model performance, balancing true positives with false positives and supporting informed decision-making in classification tasks.

3.6.1. Confusion Matrix

The confusion matrix serves as a fundamental instrument in the assessment of classification models, offering a comprehensive breakdown of prediction outcomes. Through the analysis of this matrix and the associated metrics, practitioners can obtain critical insights into the performance of the model, pinpoint areas necessitating enhancement, and make informed decisions concerning model selection and threshold optimization. It offers valuable insights into the model's accuracy and aids in identifying the sources of errors. The matrix is generally organized in the following manner:

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

3.6.2. Accuracy

Accuracy is defined as the ratio of the number of correct predictions (both true positives and true negatives) to the total number of predictions made. It is calculated using the formula:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Predictions}$$

3.6.3. Precision

Precision measures the proportion of true positive predictions relative to all positive predictions made by the model. It is calculated using the formula:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

3.6.4. Recall (Sensitivity/True Positive Rate)

Recall measures the proportion of actual positive instances that the model correctly identifies. It is calculated using the formula:

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

3.6.5. F1-Score:

The F1 Score measures the harmonic mean of precision and recall, providing a single score that balances both metrics. It is calculated using the formula:

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

3.6.6. Cross-Validation Accuracy

Cross-validation accuracy is a fundamental metric in machine learning that quantifies the average accuracy of a model by training and testing it across various splits of a dataset. This methodology entails partitioning the dataset into several subsets, known as "folds," wherein the model is trained on a selection of these folds and subsequently evaluated on the remaining fold. By iterating this process across all folds, one obtains an aggregate accuracy score that reflects the model's performance. This technique is particularly advantageous for assessing a model's capacity to generalize to unseen data, as it effectively mitigates the risk of overfitting—wherein a model performs satisfactorily on training data yet poorly on novel data. In summary, cross-validation provides a more robust and reliable estimate of a model's performance, thereby enhancing the credibility of performance evaluation in machine learning.

3.6.7. Receiver Operating Characteristic (ROC)

The Receiver Operating Characteristic (ROC) curve is an essential instrument for assessing the performance of binary classification models by graphically representing the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) across various threshold settings (Hanafy and Ming, 2021). This methodology facilitates the comparison of multiple models, enabling analysts to identify the most effective model, which is typically characterized by a curve that approaches the top-left corner, indicating a higher sensitivity and a lower rate of false positives. The Area Under the Curve (AUC) serves as a summary measure of overall performance; an AUC of 1 denotes a perfect model, while an AUC of 0.5 indicates no discriminative capability. The ROC curve's advantages include its insensitivity to class distribution, rendering it particularly valuable for imbalanced datasets, and its intuitive graphical representation, which enhances the understanding of the trade-offs between sensitivity and specificity. Frequently utilized in fields such as medical diagnostics, machine learning, and finance, the ROC curve is a powerful tool for informed decision-making regarding model evaluation and threshold selection.

These metrics collectively enhance the understanding and evaluation of model performance, facilitating informed decision-making in the methodology.

3.7. Model Integration and Dashboard Creation

In the present study, a predictive model was developed using Python and subsequently saved for future use (Otarb, 2024). The model was later imported into Power BI through a Python script, where it was applied to the same pre-processed dataset, which consisted of training and testing subsets, maintaining consistent column names to generate predictions regarding insurance claims. The saved model was utilized to make predictions based on the input dataset, and these predictions were stored for further analysis to gain insights into claim occurrences and their potential impacts.

The predicted data was subsequently utilized to create an interactive dashboard within Power BI. This dashboard was designed to calculate the accuracy of the prediction model using DAX code and the confusion matrix. The predictions, along with the key features influencing the claims, were visually represented to provide a clearer understanding of the contributing factors. Various visualization tools, including graphs and charts, were incorporated into the Customer Overview Dashboard to enhance data interpretation and facilitate informed decision-making for stakeholders.

This integrated approach was designed to ensure the practical application of the predictive model while offering an accessible platform for stakeholders to monitor and analyse claim predictions. The combination of Python for model development and Power BI for data visualization was found to effectively streamline data processing and enhance the overall efficiency and clarity of the analysis.

Chapter 4 Result and Discussion

4.1. Performance Results and Evaluation

Seven classification algorithms were employed to train the model, with each algorithm selected for its unique strengths in addressing various classification tasks. A concise overview of these algorithms was presented in the methodology section, wherein their key characteristics, underlying assumptions, and suitability for the specific problem were detailed. The analysis yielded a comprehensive set of performance metrics, including accuracy, precision, recall, and F1-score, which facilitated a thorough comparison of the effectiveness of each algorithm. Additionally, the results were elaborated upon to evaluate not only the predictive capabilities of each algorithm but also their potential limitations. This multifaceted approach allowed for a nuanced understanding of the performance of each algorithm under different conditions, ultimately guiding the selection of the most appropriate algorithm for the predictive modelling task at hand.

4.1.1 Model Accuracy with Different Approaches and Algorithms

The accuracy metrics for Approaches One through Six, encompassing model accuracy, cross-validation mean accuracy, and cross-validation standard deviation, are detailed in the tables presented in Figures 12 to 17.

Model Comparison:

	Model	Accuracy	CV Mean Accuracy	CV Std Dev
0	Random Forest	0.735477	0.731659	0.005801
1	Decision Tree	0.731328	0.732282	0.006272
2	Logistic Regression	0.746888	0.747120	0.001232
3	SVM	0.747407	0.746809	0.003803
4	MLP	0.739108	0.740065	0.005472
5	XGBoost	0.727178	0.736120	0.006901
6	K-Nearest Neighbors	0.702801	0.709764	0.006947

Figure 12:Model Accuracy with OneHotEncoder Applied

Model Comparison:

	Model	Accuracy	CV Mean Accuracy	CV Std Dev
0	Random Forest	0.781120	0.771505	0.007286
1	Decision Tree	0.671680	0.684859	0.008915
2	Logistic Regression	0.770228	0.769223	0.009381
3	SVM	0.785788	0.773996	0.002646
4	MLP	0.691909	0.708516	0.016069
5	XGBoost	0.774896	0.768808	0.003171
6	K-Nearest Neighbors	0.747407	0.741413	0.004747

Figure 15:Model Accuracy with OneHotEncoder and StandardScaler Applied

Model Comparison:

	Model	Accuracy	CV Mean Accuracy	CV Std Dev
0	Random Forest	0.768672	0.759884	0.003316
1	Decision Tree	0.675830	0.680296	0.007350
2	Logistic Regression	0.693983	0.692747	0.010708
3	SVM	0.723029	0.730413	0.007852
4	MLP	0.696058	0.703747	0.005248
5	XGBoost	0.766079	0.764658	0.006763
6	K-Nearest Neighbors	0.630187	0.637022	0.012137

Figure 14:Model Accuracy with OneHotEncoder, StandardScaler and SMOTE Applied

Model Comparison:

	Model	Accuracy	CV Mean Accuracy	CV Std Dev
0	Random Forest	0.778008	0.773580	0.005704
1	Decision Tree	0.678942	0.682059	0.008213
2	Logistic Regression	0.764523	0.753139	0.004248
3	SVM	0.732884	0.733008	0.000215
4	MLP	0.740145	0.705615	0.029622
5	XGBoost	0.765041	0.767147	0.005502
6	K-Nearest Neighbors	0.710581	0.704369	0.012776

Figure 13:Model Accuracy with Label Encoder Applied

Model Comparison:

	Model	Accuracy	CV Mean Accuracy	CV Std Dev
0	Random Forest	0.781639	0.773581	0.003752
1	Decision Tree	0.678942	0.685482	0.004947
2	Logistic Regression	0.774896	0.767251	0.005306
3	SVM	0.778008	0.769430	0.005095
4	MLP	0.739627	0.734875	0.006128
5	XGBoost	0.772822	0.764864	0.005056
6	K-Nearest Neighbors	0.738071	0.735706	0.007056

Figure 16: Model Accuracy with Label Encoder and StandardScaler Applied

Model Comparison:

	Model	Accuracy	CV Mean Accuracy	CV Std Dev
0	Random Forest	0.761929	0.754383	0.007168
1	Decision Tree	0.658714	0.664004	0.010385
2	Logistic Regression	0.687241	0.691501	0.010518
3	SVM	0.721992	0.720867	0.013214
4	MLP	0.691390	0.686209	0.009383
5	XGBoost	0.776452	0.758431	0.006256
6	K-Nearest Neighbors	0.652490	0.636712	0.010199

Figure 17: Model Accuracy with Label Encoder, StandardScaler and SMOTE Applied

Table 2: Model Performance Comparison Across Different Approaches and Algorithms

Approach 1					
Algorithm	Accuracy	Macro Precision	Macro Recall	Macro F1-Score	Mean Cross-Validation Accuracy
Random Forest	0.7355	0.64	0.59	0.59	0.7317
Decision Tree	0.7313	0.63	0.57	0.57	0.7323
Logistic Regression	0.7469	0.67	0.57	0.56	0.7471
SVM	0.7474	0.68	0.56	0.55	0.7468
MLP	0.7391	0.65	0.58	0.58	0.7401
XGBoost	0.7272	0.62	0.56	0.56	0.7361
KNN	0.7028	0.59	0.57	0.57	0.7098
Approach 2					
Random Forest	0.7811	0.74	0.64	0.66	0.7715
Decision Tree	0.6717	0.59	0.59	0.59	0.6849
Logistic Regression	0.7702	0.71	0.63	0.64	0.7692
SVM	0.7858	0.75	0.64	0.66	0.7740
MLP	0.6919	0.61	0.61	0.61	0.7085
XGBoost	0.7749	0.71	0.66	0.67	0.7688
KNN	0.7474	0.67	0.61	0.62	0.7414
Approach 3					
Random Forest	0.7687	0.70	0.67	0.68	0.7599
Decision Tree	0.6758	0.60	0.61	0.60	0.6803
Logistic Regression	0.6940	0.66	0.70	0.66	0.6927
SVM	0.7230	0.66	0.69	0.67	0.73
MLP	0.6961	0.62	0.62	0.62	0.7037
XGBoost	0.7661	0.70	0.65	0.67	0.7647
KNN	0.6302	0.61	0.64	0.60	0.6370
Approach 4					
Random Forest	0.7780	0.73	0.64	0.66	0.7736
Decision Tree	0.6789	0.60	0.61	0.60	0.6821
Logistic Regression	0.7645	0.71	0.60	0.61	0.7531
SVM	0.7329	0..87	0.50	0.42	0.7330
MLP	0.7401	0.69	0.53	0.48	0.7056
XGBoost	0.7650	0.70	0.65	0.66	0.7671
KNN	0.7106	0.60	0.57	0.57	0.7044
Approach 5					
Random Forest	0.7816	0.74	0.65	0.66	0.7736
Decision Tree	0.6789	0.60	0.61	0.60	0.6855
Logistic Regression	0.7749	0.73	0.63	0.65	0.7673
SVM	0.7780	0.74	0.63	0.65	0.7694
MLP	0.7396	0.66	0.63	0.64	0.7349
XGBoost	0.7728	0.71	0.66	0.67	0.7649
KNN	0.7381	0.65	0.61	0.62	0.7357
Approach 6					
Random Forest	0.7619	0.69	0.67	0.68	0.7544
Decision Tree	0.6587	0.59	0.60	0.59	0.6640
Logistic Regression	0.6872	0.65	0.69	0.65	0.6915
SVM	0.7220	0.66	0.69	0.67	0.7209

MLP	0.6914	0.62	0.64	0.63	0.6862
XGBoost	0.7765	0.72	0.67	0.68	0.7584
KNN	0.6525	0.62	0.64	0.61	0.6367

Table 2 presented the results of all approaches; however, this study emphasizes the superior performance observed in Approaches Two and Five. The performance metrics for various algorithms in Approach Two are detailed as follows: The Random Forest model exhibited the highest accuracy at 0.7811, closely followed by the Support Vector Machine (SVM), which achieved an accuracy of 0.7858. The XGBoost model also demonstrated strong performance, attaining an accuracy of 0.7749. Regarding the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) values, XGBoost led with an AUC of 0.76, while Random Forest achieved an AUC of 0.77, and SVM recorded an AUC of 0.75. Other models, such as Logistic Regression and K-Nearest Neighbours (KNN), displayed lower performance metrics, with accuracies of 0.7702 and 0.7474, respectively. These results indicate that while Random Forest and SVM are robust contenders, XGBoost offers a well-balanced performance in terms of both accuracy and ROC AUC, rendering it a reliable choice for classification tasks.

Similarly, the performance metrics for various algorithms in Approach Five are summarized as follows, The Random Forest model demonstrated the highest accuracy at 0.7816, followed closely by the Support Vector Machine (SVM), which achieved an accuracy of 0.7780. Logistic Regression also performed admirably, attaining an accuracy of 0.7749. Regarding the Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) values, Random Forest led with an AUC of 0.76, while XGBoost recorded an AUC of 0.75, and SVM followed with an AUC of 0.74. Other models, such as the Decision Tree and K-Nearest Neighbours (KNN), exhibited lower performance metrics, with accuracies of 0.6789 and 0.7381, respectively. These results indicate that while Random Forest and SVM are strong performers, the overall results suggest that Random Forest provides a well-rounded performance across various metrics, making it a suitable choice for classification tasks. According to the results, Approaches Two and Five demonstrated the highest accuracy. Consequently, Figures 18 and 19 depict the confusion matrices for these high-accuracy models.

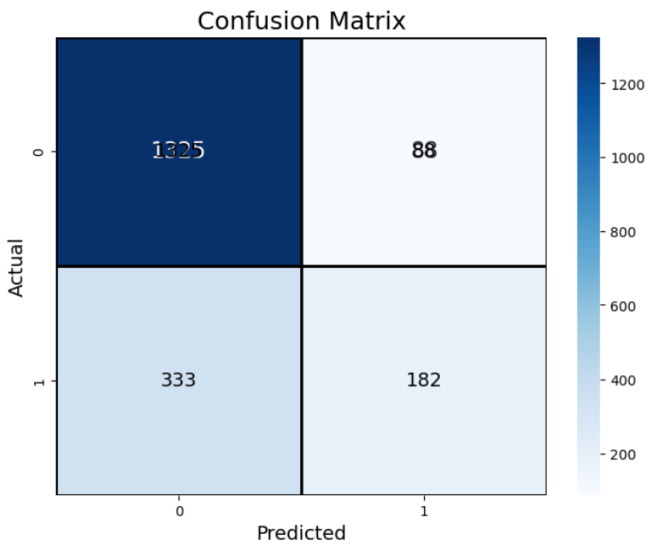

Figure18: Approach 5: Confusion Matrix for Random Forest

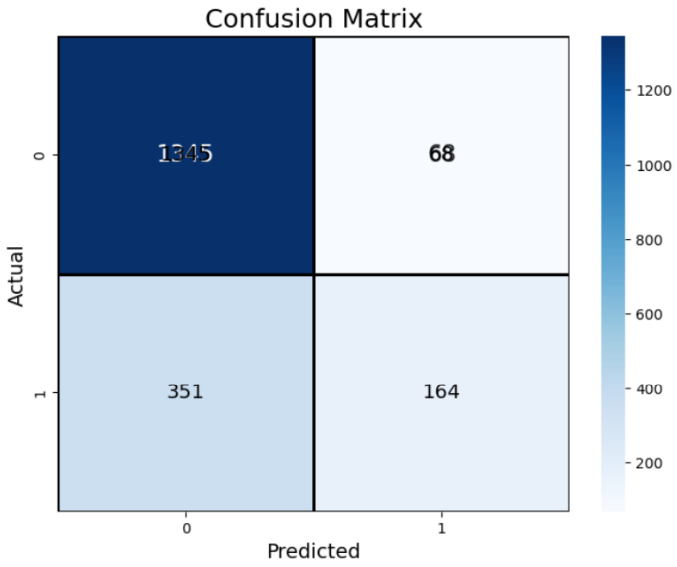

Figure19: Approach 2: Confusion Matrix for SVM

Figures 20 and 21 illustrate the ROC curves, which were used to compare the performance of various machine learning models across two different approaches (Approach 2 and Approach 5). In both figures, the algorithms evaluated included Random Forest, Decision Tree, Logistic Regression, SVM, MLP, XGBoost, and K-Nearest

Neighbours. The curves demonstrated the relationship between the true positive rate and the false positive rate for each model, with the Area Under the Curve (AUC) being utilized as the performance metric. Random Forest and Logistic Regression were consistently identified as the top-performing models in both approaches, with AUC values of approximately 0.77 in Approach 2 and 0.76 in Approach 5. SVM and XGBoost were also observed to perform well, with AUC scores exceeding 0.74. In contrast, the Decision Tree was determined to have performed poorly, with an AUC of 0.61 in both cases. A notable improvement in the MLP model was recorded in Approach 5, where the AUC increased from 0.69 to 0.71. Overall, the graphs provided a clear comparison of model performance, aiding in the identification of the most effective classifiers.

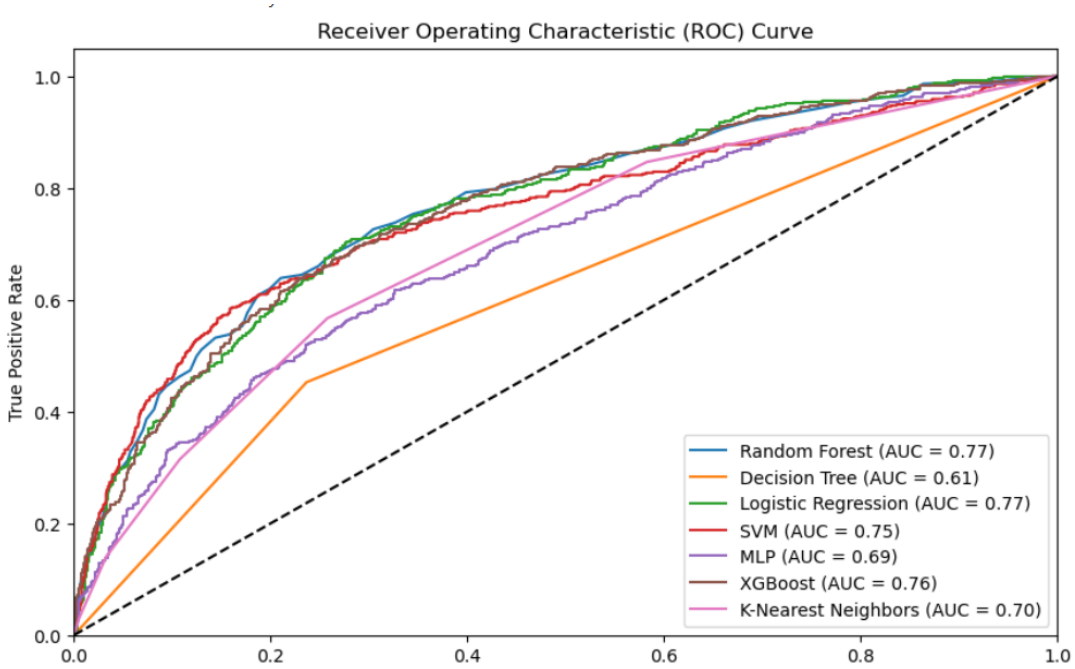

Figure 20: ROC Curves for All Algorithms in Approach 2

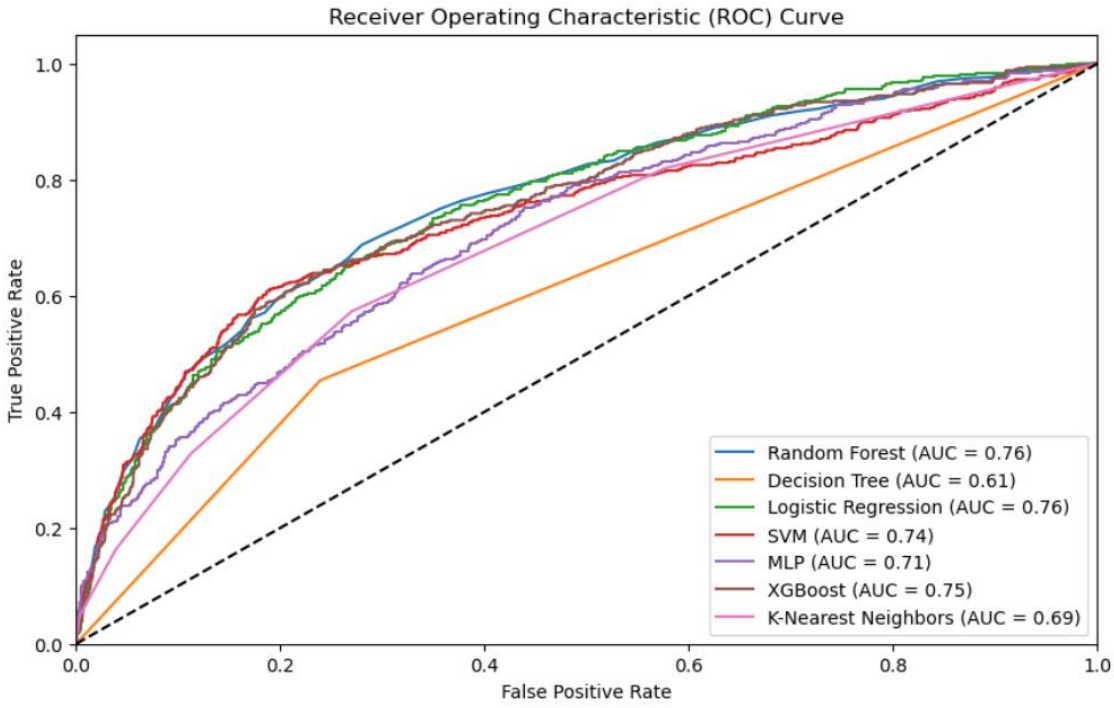

Figure 21: ROC Curves for All Algorithms in Approach 5

4.1.2. Development of Power BI Dashboard for Model Performance Evaluation

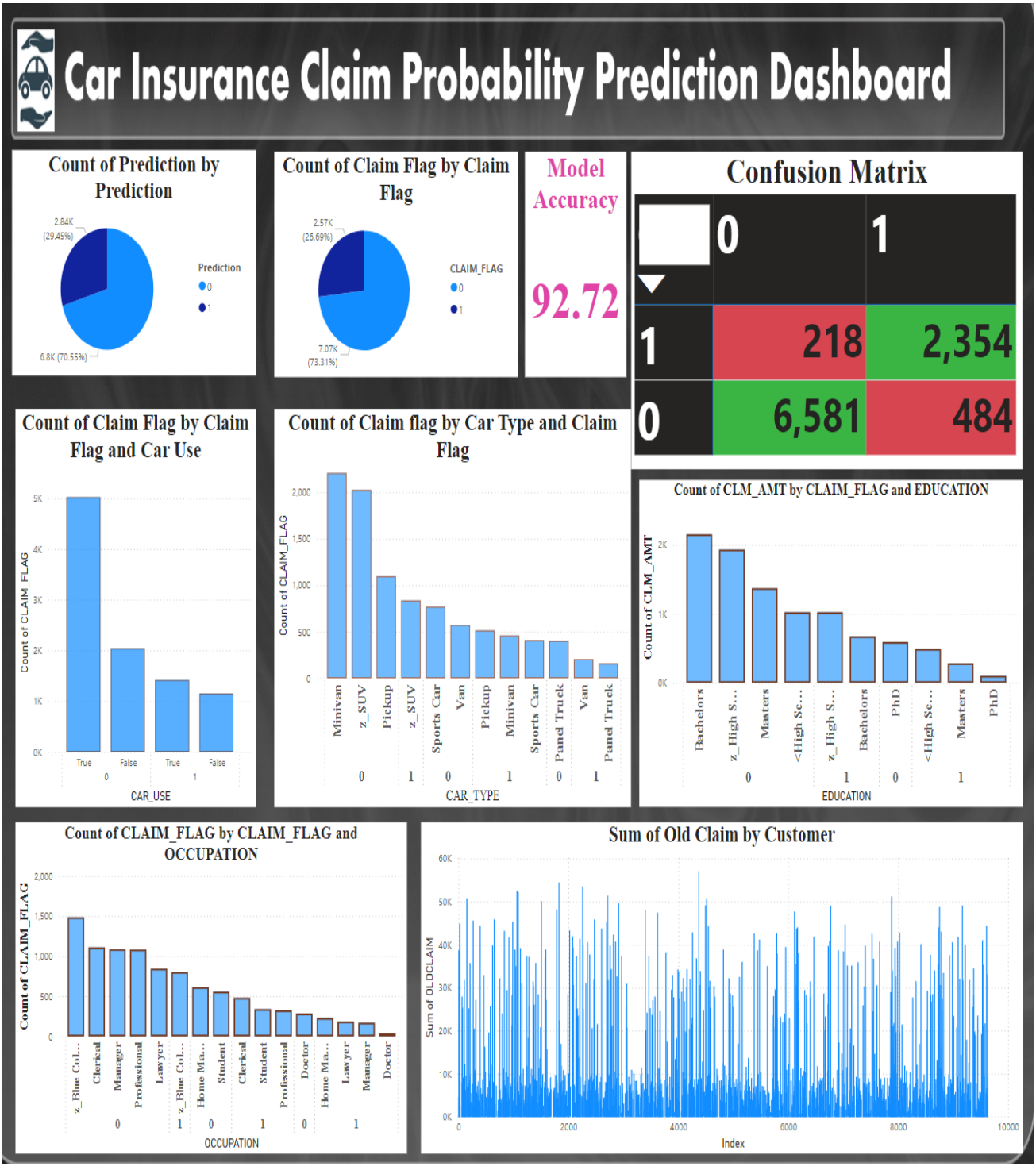

Figure 22: Car insurance claim probability prediction Dashboard

Figure 22 shows the Power BI dashboard, which was developed to provide a comprehensive overview of the Car Insurance Claim Probability Prediction model, highlighting key insights derived from the data. At the top, the Model Accuracy is prominently displayed, showing that a high predictive performance of 92.72% had been achieved. The Confusion Matrix was used to break down the model's classification results, revealing that 6,581 true negatives and 218 true positives were identified, while 2,354 false negatives and 484 false positives were detected. This breakdown provided a clear understanding of how effectively the model predicted both claim and non-claim outcomes.

In addition to performance metrics, multiple visualizations were incorporated into the dashboard to delve deeper into the dataset. Pie charts were used to represent the distribution of predictions and actual claim flags, with the majority classified as non-claims. Bar charts were generated to offer a detailed analysis of claims based on several factors, including Car Use, Car Type, Occupation, and Education. For instance, it was observed that SUVs and trucks had higher claim rates, while occupations such as clerical workers and high school graduates were associated with a higher likelihood of filing claims.

Lastly, a line graph at the bottom right was created to track the Sum of Old Claims by Customer, providing valuable insights into past claim behaviour and trends over time. Overall, this dashboard was designed to serve as an effective tool for visualizing and analysing key factors influencing car insurance claims, offering both high-level performance metrics and granular data insights.

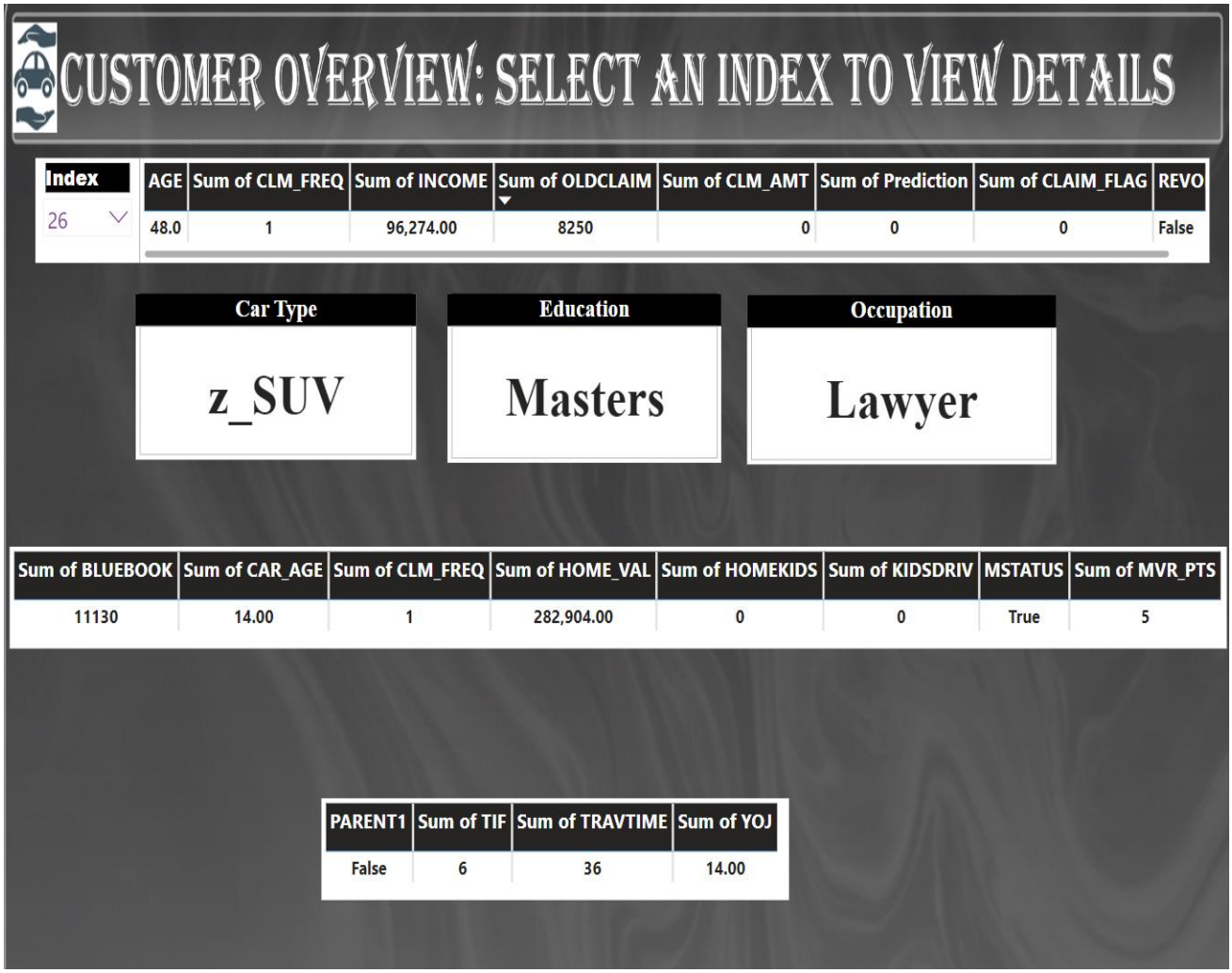

Figure 23: Customer overview dashboard

Figure 23 illustrates the Customer Overview Dashboard, developed in Power BI to provide a detailed view of individual customer data related to car insurance claims. Users are enabled to select a specific customer index, allowing various attributes associated with that customer to be accessed. For instance, it was observed that the selected customer is 48 years old, has an income of $96,274, and has made one claim, amounting to $8,250. The Sum of Prediction and Sum of Claim Flag indicate that no future claims are expected, as the claim flag is set to 0.Key factors such as car type (SUV), educational attainment (Master's degree), and occupation (lawyer) are central to the profile, offering insights into lifestyle and potential risks. Additional details, including a car value of $11,130, car age of 14 years, a home valued at $282,904, marital status (married), and 5 points on the Motor Vehicle Report, are provided. This dashboard is considered a valuable tool for assessing customer risk and tailoring services.

4.2. Discussion

This section provides a critical evaluation of the results derived from the various machine learning models, underscoring both the strengths and limitations of each approach, while also identifying potential areas for future improvement.

4.2.1 Model Performance and Strengths

The highest overall accuracy across various algorithms was exhibited by Approaches Two and Five, particularly with Random Forest and Support Vector Machines (SVM). Several key factors were attributed to the consistently superior performance of these models. Firstly, the handling of categorical and numerical features was identified as significant. In Approach Two, One-Hot Encoding was applied to categorical and Boolean variables, in conjunction with feature scaling for numerical variables, effectively enhancing model accuracy. Similarly, in Approach Five, Label Encoding was utilized alongside feature scaling to ensure that categorical variables did not possess unintended ordinal relationships, and that balance was maintained across numerical features.

Additionally, the implementation of Standard Scaling was observed to mitigate issues related to features with varying scales disproportionately influencing results. This normalization was especially beneficial for algorithms such as SVM and Logistic Regression, which rely on distance-based calculations, thus resulting in improved performance with normalized data. In addressing class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed in Approaches Three and Six. While SMOTE was effective in improving class balance, it was determined that simply balancing the dataset did not guarantee optimal performance; other factors, such as feature scaling and encoding, were found to exert a more significant influence on accuracy.

Ultimately, Random Forest and SVM were identified as the top-performing algorithms in this analysis. The strength of Random Forest was recognized for its ability to efficiently handle both categorical and numerical features while demonstrating robustness against overfitting due to its ensemble nature. Conversely, SVM was noted for its strong predictive power, effectively identifying optimal decision boundaries, particularly in high-dimensional feature spaces. The combination of advanced preprocessing techniques and robust algorithm selection was credited with underpinning the high accuracy achieved in these approaches.

4.2.2 Limitations and Model Shortcomings

Despite the notable performance of certain models, several limitations impacted the overall accuracy and generalizability of the results. Class imbalance was identified as a significant issue; although SMOTE was employed to help balance the class distribution, it did not fully eliminate the bias toward the majority class. Some models, such as Decision Trees, were found to underperform, potentially due to overfitting or an inability to capture the complex relationships among features. Furthermore, it was suggested that SMOTE may have introduced synthetic bias, which could alter the distribution of minority class features.

Feature engineering also presented challenges. Correlation analysis revealed weak relationships between most features and the target variable (Claim Flag), indicating that the dataset's features may not be sufficiently predictive of insurance claims. More advanced feature engineering techniques, such as interaction terms or polynomial features, were suggested for exploration to capture non-linear relationships. Additionally, the application of domain-specific knowledge could lead to the creation of more meaningful variables that align closely with insurance claim predictions.

The exclusion of key variables was necessary to avoid data leakage, particularly concerning the Claim Amount feature, but this decision resulted in the loss of potentially valuable information. While the exclusion of perfectly correlated features like Claim Amount was critical for preserving model integrity, future research could investigate alternative methods to leverage this information indirectly without compromising model generalization.

Concerns regarding overfitting were noted, particularly with Decision Trees and K-Nearest Neighbours (KNN), both of which exhibited signs of overfitting. Decision Trees, in particular, recorded the lowest accuracy across several approaches, suggesting that the model may have been overly complex for the training data and failed to generalize well on the test set. Techniques such as pruning and hyperparameter tuning could be employed to mitigate overfitting in these models.

The discrepancy in accuracy between the Random Forest model, which achieved 78.16% accuracy in Python and 92.72% in Power BI, can be attributed to the manner in which the model was applied and the dataset utilized for generating predictions. In Python, the accuracy was calculated strictly using a test set that was not part of the training data, thereby reflecting the model's ability to generalize to unseen data, resulting in a lower accuracy figure. Conversely, in Power BI, the model was applied to a dataset that included both training and testing subsets, which may have inadvertently led to a more favourable evaluation of its predictive capabilities. By utilizing the combined dataset, the model had access to more data points, potentially resulting in higher accuracy due to its familiarity with the data it was predicting. This highlights the importance of ensuring that model evaluation metrics accurately reflect the model's ability to generalize to new, unseen data rather than solely relying on performance measures derived from datasets it has previously encountered.

Finally, although Random Forest and SVM demonstrated strong performance, both models were often viewed as black-box models, which limited their interpretability. In high-stakes decision-making environments such as insurance, interpretability is critical for understanding the factors driving predictions. Future research could employ interpretability techniques, such as SHAP (Shapley Additive Explanations), to provide insights into the underlying influences on model predictions.

Chapter 5 Conclusions and Future work

5.1. Conclusions

the Power BI dashboards were designed to provide comprehensive insights into car insurance claim predictions and individual customer profiles. By integrating performance metrics, visualizations, and detailed customer data, a deeper understanding of the factors influencing claim behaviour and risk assessments was facilitated. The high predictive accuracy of the model was highlighted, demonstrating its reliability, while the granular data allowed for services to be tailored to meet individual customer needs. Ultimately, these dashboards were established as essential tools for informed decision-making, enhancing the capability to analyse trends and improve overall operational efficiency within the insurance industry.

The primary objective of this project was successfully achieved through the development of a robust predictive model that was designed to estimate the likelihood of car insurance claims based on historical data. Advanced machine learning techniques were employed alongside comprehensive data preprocessing methods, resulting in significant improvements in predictive accuracy and risk assessment capabilities for car insurance providers. Among the various approaches evaluated, Approaches Two and Five were found to demonstrate the highest accuracy rates, with the Random Forest model exhibiting particularly strong performance across multiple metrics. This result affirmed the effectiveness of a combination of OneHot Encoding and Standard Scaling in enhancing model performance.

The comparative analysis of classification algorithms revealed that both the Random Forest and Support Vector Machine (SVM) models were identified as robust contenders for this task, with Random Forest achieving an accuracy of 78.16% and SVM attaining 77.80% in Approach Five. This finding underscored their applicability in predictive modelling within the insurance sector. Although the Synthetic Minority Over-sampling Technique (SMOTE) was incorporated to address class imbalance, it was determined that this method did not significantly impact model accuracy in this study. This observation emphasized the importance of evaluating the relevance and efficacy of various preprocessing techniques in the context of specific datasets and models.

By accurately predicting the likelihood of claims, the developed model sought to mitigate adverse selection, enabling insurance companies to align premiums more closely with actual risk levels. Such alignment was expected to promote improved customer satisfaction through fairer pricing and to enhance customer retention and loyalty. Furthermore, the insights derived from the predictive model empowered car insurance companies to optimize resource allocation, resulting in cost savings and improved financial performance. By identifying potential claimants at the point of purchase, informed decisions were facilitated that positively influenced the bottom line. Overall, this project highlighted the potential of machine learning to transform the car insurance industry, providing both insurers and policyholders with tools to enhance fairness and accuracy in risk assessment and pricing.

5.2 Future work

The methods presented in this study successfully establish a framework for predicting insurance claims; however, several avenues for future research are recommended to enhance the robustness of model evaluation and performance. The observed discrepancy in accuracy between the Random Forest model—achieving 78.16% in Python and 92.72% in Power BI—underscores the need for a more nuanced approach to model validation. Future studies should prioritize the implementation of a consistent evaluation protocol that strictly separates training and test datasets to ensure a fair assessment of model generalization capabilities. Additionally, researchers could investigate the impact of cross-validation techniques to provide a more comprehensive understanding of model performance across varying data splits. Further exploration of ensemble methods or hybrid models may also prove beneficial in improving predictive accuracy. By focusing on these areas, future research may contribute to a deeper understanding of model behaviour in practical applications, ultimately leading to more reliable predictions in the insurance domain.

The methods presented in this study aim to establish a robust framework for accurately predicting insurance claims. While the findings are promising, several areas warrant further exploration and improvement. The focus of this research was narrowed to a binary classification problem; however, future studies may consider formulating the issue as a multiclass classification problem. Instead of merely predicting whether a claim will be filed, researchers could categorize claim values into distinct bins and predict the probability of an observation falling into each bin. This approach is likely to enhance individual prediction accuracy, although it may negatively affect the overall error associated with the total claim amount.

By tackling the classification problem in a multiclass manner, it may be possible to achieve more precise estimations for individual claims. Nonetheless, this focus on individual claims could detract from the accuracy of the total claim amount, as the multiclass problem centres more on individual outcomes rather than broader financial implications. A hybrid approach could be adopted, where the model forecasts the total expected claim amount through binary classification while simultaneously specializing in individual claims using multiclass classification techniques.

Furthermore, enhancements in translating claim probabilities into actual claim values could be achieved as more data becomes available. Currently, predictions are made by multiplying the claim probability by the average claim amount of the entire training dataset. Future methodologies could refine this by employing a conditional average claim amount, which would utilize the average claim amounts of similar records from historical databases. However, the effectiveness of this approach hinges on the availability of a larger dataset, as the current dataset's sparsity may lead to volatile results.

While this study employed the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, the findings indicated that it did not significantly impact model accuracy. Nonetheless, future research should explore advanced techniques for managing class imbalance, such as synthetic data generation, ensemble methods tailored for imbalanced data, or cost-sensitive learning. These strategies could potentially enhance the model's ability to predict minority classes effectively.

The framework developed in this study can be extended to other domains requiring binary decision-making. For instance, a practical application for Finaps lies in their employment screening service, CV-OK. This service is designed to assess candidates and employees through an online platform, determining whether an individual has provided false information on their resume. This scenario can similarly be distilled into a binary classification problem. By leveraging the framework constructed in this research, it is feasible to develop a model that predicts the probability of a candidate having misrepresented themselves.

Bibliography

1. Abdelhadi, Elbahnasy and Abdelsalam, 2020. *a proposed model to predict auto insurance claims using machine learning techniques,* s.l.: Ain Shams University and Helwan University.

2. Ahmed and Linen, 2017. A review and analysis of churn prediction methods for customer retention in telecom industries. *International conference on advanced computing and communication systems*.

3. Bento, 2021. *Medium.* [Online]
   Available at: https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141
   [Accessed September 2024].

4. Chen and Guestrin, 2016. *XGBoost: A Scalable Tree Boosting System,* s.l.: University of Washington.

5. Christiansen, et al, 2016. Whois changing health insurance coverage? empirical evidence on policyholder dynamics.. *Risk and Insurance.*

6. Christopher, et al, 2020. *Machine learning in P&C insurance: A review for pricing and reserving,* s.l.: Université Laval.

7. Fauzan and Murfi, 2018. *The accuracy of XGBoost for insurance claim prediction,* s.l.: Universitas Indonesia.

8. Hanafy and Ming, 2021. *Machine learning approaches for auto insurance big data,* s.l.: Zhejiang Gongshang University.

9. Hooson and Mark, 2024. *Forbes ADVISOR.* [Online]
   Available at: https://www.forbes.com/uk/advisor/car-insurance/car-insurance-statistics
   [Accessed August 2024].

10. IBM, 2023. *Support Vector Machine.* [Online]
    Available at: https://www.ibm.com/topics/support-vector-machine
    [Accessed September 2024].

11. Jha and Mukul Kumar, 2020. *www.kaggle.com.* [Online]
    Available at: https://www.kaggle.com/datasets/mukuljh2/car-insurance-claim

12. Krishna, 2024. *kaggle.com.* [Online]
    Available at: https://www.kaggle.com/discussions/general/513005
    [Accessed 26 September 2024].

13. Krishnamurthy, et al, 2021. *Machine learning prediction models for chronic kidney disease using national health insurance claim data in Taiwan,* s.l.: s.n.

14. Li, 2023. *Identifying the optimal machine learning model for predicting car insurance claims: a comparative study utilising advanced techniques,* Central University of Finance and Economics: s.n.

15. Ng'elechei, et al, 2020. Modeling frequency and severity of insurance claims in an insurance portfolio. *American journal of applied mathematics and statistics,,* Issue Science and Education Publishing.

16. Otarb, et al, 2024. *learn.microsoft.com.* [Online]
    Available at: https://learn.microsoft.com/en-us/power-bi/connect-data/desktop-python-scripts
    [Accessed Septembter 2024].

17. Otarb, 2024. *learn.microsoft.com.* [Online]
Available at: https://learn.microsoft.com/en-us/power-bi/connect-data/desktop-python-scripts
[Accessed September 2024].

18. Pesantez-Narvaez ,Guillen and Alcaniz, 2019. *Predicting motor insurance claims using telematics data—XGBoost versus logistic regression,* s.l.: Universitat de Barcelona.

19. Quan and Valdez, 2018. Predictive analytics of insurance claims using multivariate decision trees. *Depend. Model.*

20. Rawat, et al, 2021. *Application of machine learning and data visualization techniques for decision support in the insurance sector,* s.l.: Amity University Uttar Pradesh.

21. Schott, 2019. *Medium.* [Online]
Available at: https://medium.com/capital-one-tech/random-forest-algorithm-for-machine-learning-c4b2c8cc9feb
[Accessed September 2024].

22. Yunos, et al, 2016. *Predictive modelling for motor insurance claims using artificial neural networks,* s.l.: Universiti Teknologi Malaysia & Universiti Kebangsaan Malaysia.