2024

MOTOR INSURANCE CLAIMS PREDICTION: COMPARATIVE STUDY USING MACHINE LEARNING

Sankhadip Roy
University of Rhode Island, sankhadiproy@gmail.com

MOTOR INSURANCE CLAIMS PREDICTION: COMPARATIVE STUDY USING MACHINE LEARNING

BY

SANKHADIP ROY

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
STATISTICS

UNIVERSITY OF RHODE ISLAND

2024

MASTER OF SCIENCE THESIS

OF

SANKHADIP ROY

APPROVED:

Thesis Committee:

Major Professor Guangyu Zhu

Haihan Yu

Xiaowei Xu

Brenton DeBoef

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2024

ABSTRACT

The report by the Insurance Information Institute underscores a notable increase in both claims frequency and severity, particularly highlighting a significant surge in accident claim severity within US car insurance from 2010 to 2019. Simultaneously, there has been a marked rise in the average expenditure on US car insurance during this timeframe. These shifts emphasize the critical need for accurate predictions to fine-tune premium adjustments and enhance the accessibility of car insurance coverage for a broader demographic of drivers. Consequently, numerous insurance companies are transitioning from traditional methodologies to incorporate machine learning (ML) techniques, providing a more sophisticated and reliable framework for generating outcomes. Nonetheless, the challenge persists in selecting the most optimal ML predictive model to effectively identify probable claims or potential premium defaulters. This study tackles these complexities by employing diverse classification methods and proposing specific techniques for feature selection and data resampling, with the overarching goal of constructing comprehensive classification models tailored for in-depth claim analysis.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. Guangyu Zhu, for his unwavering support, guidance, and encouragement throughout the entire thesis process. I am also immensely thankful to the inside committee member, Dr. Haihan Yu, and the outside committee member, Dr. Xiaowei Xu, for their timely support. Special thanks are also due to the Defense Chair, Dr. Weiwei Jia, for presiding over my defense. Their assistance has been instrumental in completing this thesis, and I am truly grateful for their contributions.

TABLE OF CONTENTS

CHAPTER

Page

LIST OF FIGURES

Figure **Page**

LIST OF TABLES

CHAPTER 1

Introduction

According to the insights conveyed in the Insurance Information Institute report (III, 2020), there exists a discernible upward trajectory in claims frequency, the number of claims per car, and claim severity. Noteworthy is the 35% surge in the severity of accident claims for US car insurance observed between 2010 and 2019. The average expenditure on US car insurance also witnessed an increase from USD $78,665 in 2009 to USD $100,458 in 2017 (III, 2020). The importance of precise predictions in the insurance domain cannot be overstated, as they enable the industry to refine premium adjustments, fostering increased affordability of car insurance coverage for a broader spectrum of drivers.

A palpable shift in the industry's operational approach is apparent, with numerous insurance companies transitioning from conventional methods to embracing machine learning (ML) techniques. This transition is pivotal, providing a more nuanced and robust framework for generating outcomes that are both dependable and representative. A recent study by McKinsey & Company (Columbus, 2017), focused on the intersection of artificial intelligence and business profitability, revealed compelling insights. Businesses wholeheartedly adopting artificial intelligence projects experienced a substantial boost in profit margins, ranging from 3% to 15%. However, despite these advancements, the challenge of selecting an optimal ML predictive model remains an aspect that demands comprehensive consideration and attention.

In the context of claim prediction challenges, classification models function as decision-making tools, employing techniques like feature selection, feature discretization, and data resampling. These models are pivotal for effective risk assessment. Optimal feature subset selection not only reduces computational costs but also enhances the model's efficiency and interpretability, as underscored by (Rawat et al., 2021). Additionally, the imbalanced nature of datasets, where positive and negative cases are unevenly distributed, necessitates strategies such as data resampling to enhance overall performance. Surprisingly, despite the recognized importance of integrating feature selection, feature discretization, resampling, and classification techniques, there is a paucity of literature that amalgamates all these strategies into a unified processing approach for constructing a comprehensive classification model in the realm of claim analysis.

In this study, three datasets were employed to analyze claims using various classification methods, namely Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, Extreme Gradient Boosting, and Naive Bayes. The Logistic Lasso, Random Forest, and knockoffs methods were suggested for feature selection to reduce dimensionality, and various resampling strategies, including under sampling, over sampling, and SMOTE, were employed to address the data's imbalance problem.

1.1 Literature Review

In the higly competitive landscape of the insurance sector, the imperative to secure and maintain customer loyalty has risen to the forefront. A profound comprehension of customers, encompassing their purchasing behaviors and potential losses, is now a critical necessity. Consequently, the classification of customers and the predictive capacity to assess their potential losses assume paramount importance. This predictive prowess not only enhances the profitability of insurance companies but also facilitates strategic interventions to mitigate loss ratios. At the heart of this process is underwriting, a pivotal procedure that assesses the risk associated with the insured. The determination of premium rates and insurance contract terms is intricately tied to this risk assessment (Briys and Varenne, 2001; Gay et al., 1998).

As each policyholder introduces a unique level of risk to the insurance company, ensuring fair and equitable premiums mandates the categorization of policyholders into distinct risk classes, where higher risks correspond to higher premiums. This rationale underscores the critical need for insurers to precisely evaluate the risks associated with their customers. Developing a highly effective model for classifying customers into distinct risk groups has perennially posed a fundamental and formidable challenge in the insurance industry.

In their pursuit of reducing financial losses for insurance companies, (Dhieb et al., 2020) introduced a system that minimizes human intervention, enhances process security, issues alerts about risky customers, and identifies fraudulent claims. Leveraging the XGBoost algorithm for insurance services, their study systematically compared its performance with other established algorithms such as Decision Trees (DT), K-Nearest Neighbors (k-NN), and Support Vector Machines (SVM). To fortify the foundation for secure transactions and data exchange within the insurance network, the authors advocated for the implementation of a blockchain-based infrastructure. The application of this solution to a dataset comprising vehicle insurance claims underscored the superiority of the XGBoost algorithm over its counterparts.

Exploring predictive models in the insurance domain, (Kate A. Smith and Brooks, 2000) conducted a comprehensive assessment of machine learning models, favoring neural networks over decision trees. Similarly, (Jing et al., 2018) opted for a Bayesian network as the exclusive tool for classifying the occurrence of insurance claims. In a unique approach, (Pesantez-Narvaez et al., 2019) conducted a comparative analysis using XGBoost and logistic regression to predict the frequency of motor insurance claims, with XGBoost demonstrating marginally superior performance. (Shady et al., 2020) contributed by developing four classifiers, including XGBoost, for predicting the occurrence of insurance claims.

While these studies focused on insurance-related predictive modeling, the works of (Hanafy and Ming, 2021a; Hanafy and Ming, 2021d; Hanafy and Ming, 2021c; Hanafy and Ming, 2021b) stand out for leveraging substantial large data, providing insights into the performance of various machine learning algorithms. Despite XGBoost's prominence, their findings revealed the superiority of random forest and decision tree (C50), while naive Bayes emerged as the least effective model. Noteworthy is the use of a real-world database from Porto Seguro, adding authenticity to their results.

Moreover, Hanafy and Ming addressed the challenges of large dataset volumes and missing values, employing diverse resampling strategies and incorporating a range of machine learning algorithms. In their recent work (Hanafy and Ming, 2022), they integrated feature selection, feature discretization, resampling, and classification techniques, presenting a comprehensive approach.

1.2 Background tools to find an effective predictive model

1.2.1 Data collection

Data collection for machine learning (ML) models is a critical process that lays the foundation for effective model training and performance. It involves gathering relevant and representative datasets from various sources such as databases, APIs, web scraping, or sensor networks. The collected data should encompass a diverse range of examples that reflect real-world scenarios and variations. Careful attention must be paid to data quality, ensuring accuracy, consistency, and completeness. Preprocessing steps may include cleaning, normalization, and feature engineering to refine the dataset for optimal model training. Ethical considerations such as privacy and consent are also paramount throughout the data collection process. Ultimately, the success of an ML model hinges on the quality and relevance of the data it is trained on, making meticulous data collection a crucial precursor to meaningful insights and predictive capabilities.

1.2.2 Data Preparation

Data Preparation involves the essential transformation of data to render it suitable for utilization by machine learning algorithms, thereby influencing the model's performance significantly. It encompasses a series of crucial steps including data cleaning, exploratory data analysis (EDA), normalization, encoding, addressing imbalanced data issues, and dimensionality reduction. Each of these facets plays a pivotal role in refining the dataset to ensure its compatibility with ML algorithms, ultimately enhancing the model's efficacy and predictive capabilities.

1.2.3 Data Cleaning

Data cleaning is an essential step in the data preprocessing pipeline, crucial for ensuring the accuracy and reliability of datasets before analysis. It involves identifying and rectifying errors, inconsistencies, and missing values within the data. This process may include tasks such as removing duplicates, correcting typos, standardizing formats, and imputing missing values using statistical methods or domain knowledge. Effective data cleaning not only improves the quality of the dataset but also enhances the robustness and validity of subsequent analyses and machine learning models.

1.2.4 Exploratory data analysis (EDA)

Exploratory Data Analysis (EDA) serves as a crucial step in comprehending data prior to implementing any machine learning models. It involves visualizing data through various graphs and charts to uncover diverse characteristics that may not be apparent merely by examining the dataset table. By leveraging EDA techniques, analysts gain insights into hidden relationships among different features. These relationships might be elusive when solely inspecting the dataset in tabular form. EDA empowers data scientists to grasp the intricate dynamics within the data, enabling them to make informed decisions regarding feature engineering, model selection, and preprocessing strategies.

1.2.5 Feature engineering

Feature engineering stands as a critical pillar in the realm of data preparation, paving the way for subsequent model training and evaluation. It involves the strategic creation of new features, drawn from insights garnered through Exploratory Data Analysis (EDA) and domain expertise, all in pursuit of bolstering model performance. Despite its demanding nature, feature engineering wields a profound impact on the accuracy of models.

These new features are crafted through an array of calculations including ratios, transformations, and statistical formulas. This transformative process extends beyond the confines of linear regression or text classification, proving beneficial across a spectrum of algorithms such as support vector machines, random forests, neural networks, and gradient boosting machines.

Furthermore, encoding categorical variables assumes paramount importance, given that the majority of machine learning algorithms operate on numerical data. Nominal encoding disregards the order of data, while ordinal encoding takes it into account. One-hot encoding, a method we utilize, circumvents multicollinearity by removing one column. Additionally, normalization plays a pivotal role by scaling data within a predefined range, typically from 0 to 1, thereby augmenting numerical scalability.

1.2.6 Feature selection

Feature selection is a vital process in machine learning where the most relevant and informative features are chosen from the dataset for model training. By selecting the most significant features and discarding irrelevant or redundant ones, feature selection helps in improving model performance, reducing overfitting, and enhancing interpretability. Various techniques, including filter, wrapper, and embedded methods, are employed to identify and select the optimal subset of features that contribute the most to predictive accuracy. Efficient feature selection not only streamlines computational resources but also aids in understanding the underlying patterns within the data, leading to more effective decision-making. The feature selection methods employed in this thesis are mentioned below.

Lasso Feature selection

Logistic Lasso feature selection is a powerful technique used in the context of logistic regression models to enhance predictive accuracy and interpretability. It combines the strengths of both L1 regularization (Lasso) and logistic regression, effectively shrinking coefficients towards zero while simultaneously performing variable selection. By imposing a penalty on the absolute size of regression coefficients, Logistic Lasso encourages sparsity in the model, automatically excluding irrelevant or redundant features. This attribute is particularly beneficial in high-dimensional datasets where the number of predictors exceeds the number of observations, thereby mitigating the risk of overfitting and improving the generalization capability of the model. Furthermore, Logistic Lasso facilitates variable selection by assigning zero coefficients to irrelevant features, thus simplifying the model and enhancing its interpretability, making it a valuable tool in fields such as biostatistics, epidemiology, and machine learning.

Random Forest Feature selection

Random Forest feature selection is a versatile and effective method for identifying important variables within a dataset. It operates within the framework of ensemble learning, constructing numerous decision trees and aggregating their predictions to generate robust and accurate results. In the process of building each tree, Random Forest randomly selects subsets of features, thereby promoting diversity among the trees and reducing the risk of overfitting. Feature importance is determined by measuring the decrease in node impurity, typically using metrics like Gini impurity or information gain, across all trees in the forest. Features with higher impurity decrease values are deemed more important in explaining the variance within the data. This method not only provides insight into the relative importance of different predictors but also offers a natural mechanism for feature selection, as less informative features tend to have lower importance scores.

Knockoffs' Feature selection

Knockoffs' feature selection (Barber and Candès, 2015; Kormaksson et al., 2020) is an innovative technique tailored to tackle the complexities of high-dimensional data analysis, particularly prevalent in fields such as genomics, economics, and social sciences. The essence of this method lies in the creation of a set of "knockoff" variables, which are crafted to mirror the statistical characteristics of the original features while deliberately avoiding redundancy. Through a meticulous comparison between the original features and their knockoff counterparts, researchers can discern genuinely significant variables while simultaneously exerting control over the false discovery rate. This strategy proves instrumental in mitigating the challenges posed by multiple testing, ultimately facilitating the revelation of meaningful patterns within intricate datasets. By employing knockoffs feature selection, analysts can navigate the intricacies of high-dimensional data analysis with greater precision and confidence, thereby unlocking valuable insights that might otherwise remain obscured.

1.2.7 Machine Learning Classifiers

Logistic Regression (LR)

Linear regression serves as a reliable tool for approximating the (linear) relationship between a continuous response variable and a set of predictor variables. However, when the response variable is binary, such as ”Yes” or ”No,” linear regression isn’t appropriate. Thankfully, analysts can turn to an alternative method, analogous to linear regression in various aspects, known as logistic regression .

Decision Tree (DT)

A decision tree is a graphical representation or model that resembles a tree structure, with its root positioned at the top and branches extending downward, akin to an inverted tree. This visual representation of data offers a clear and straightforward interpretation compared to other methods. Each input attribute corresponds to an internal node within the tree. The number of branches stemming from a hypothetical internal node equals the number of potential input attribute values. As data traverses from the root to the leaf nodes, each leaf node signifies a particular value of the label attribute.

In algorithms like Simple Cart, decision trees are constructed by recursively partitioning each decision node into two distinct branches based on various separation criteria. This iterative process enables decision trees to effectively capture relationships and patterns within the data, facilitating predictive modeling and classification tasks.

Random Forest (RF)

Random Forest (RF) stands as a widely utilized machine-learning model rooted in the decision theory pioneered by (Breiman et al., 1984). Leveraging the Classification and Regression Tree (CART) algorithm, RF constructs trees within its framework. Whether the response variable is categorical or continuous, RF adeptly handles classification and regression tasks respectively.

Within the RF model, CART initially grows an extensive tree, later subject to pruning. (Grömping, 2009) suggests that trimming an expansive tree, as opposed to limiting the number of trees grown, enhances RF's predictive accuracy. This strategy highlights RF's adaptability and efficacy in handling complex datasets, making it a favored choice in diverse analytical scenarios.

K-nearest Neighbor (KNN)

The K-nearest neighbor (KNN) algorithm operates on a fundamental principle: predicting each observation's outcome by assessing its similarity to neighboring observations. KNN is characterized as a memory-based algorithm, which implies that it relies on training samples during runtime, crafting predictions grounded on sample associations. Hence, KNN models are often referred to as ”lazy learners” , underscoring their reliance on stored data for decision-making rather than extensive upfront computation.

Extreme Gradient Boosting (XGB)

XGBoost, or eXtreme Gradient Boosting, stands as a pinnacle in the realm of ensemble learning techniques, particularly gradient boosting machines. Renowned for its exceptional performance and scalability, XGBoost operates by sequentially building an ensemble of decision trees during training, with each subsequent tree rectifying errors made by its predecessors. What sets XGBoost apart is its meticulous optimization strategies, including parallelization, tree-pruning, and regularization, aimed at enhancing both training speed and model accuracy. Moreover, XGBoost offers robust mechanisms for handling missing values and controlling overfitting through a suite of regularization parameters. With its innate ability to extract valuable insights on feature importance and support cross-validation for hyperparameter tuning, XGBoost has become the de facto choice for a myriad of machine learning tasks, from classification and regression to ranking and recommendation systems. Its widespread adoption and proven track record in data science competitions underscore its status as a cornerstone algorithm in the machine learning landscape.

Gaussian Naïve Bayes (NB)

The Gaussian Naïve Bayes classifier is a fundamental and efficient machine learning algorithm based on Bayes' theorem and the assumption of feature independence. Despite its simplicity, it remains a powerful tool for classification tasks, particularly in domains with continuous feature variables. This classifier assumes that the features follow a Gaussian (normal) distribution, making it well-suited for numerical data. Through the process of calculating probabilities using Bayes' theorem, the Gaussian Naïve Bayes classifier determines the likelihood of a particular class given the observed features. Despite its ”naïve” assumption of feature independence, this classifier often performs admirably well in practice, especially with datasets that exhibit reasonably independent features. Its computational efficiency and ability to handle high-dimensional data make it a popular choice in various applications, including text classification, medical diagnosis, and spam filtering. While it may not capture complex relationships between features, the Gaussian Naïve Bayes classifier serves as a reliable and interpretable baseline model for classification tasks in machine learning.

1.2.8 Resampling Methods

In scenarios where the distribution of classes in the training set is uneven, machine learning classifiers tend to favor categorizing all instances as belonging to the majority class to optimize overall accuracy. However, this approach often results in a significant disparity in accuracy for the minority class, which is underrepresented in the training set. Despite its lesser prevalence, the minority class can hold crucial significance in real-world applications. Consequently, overlooking the minority class can lead to inadequate performance of the classifier, particularly when accurate identification of these instances is essential. This imbalance underscores the importance of employing strategies that address class imbalance, ensuring that classifiers effectively capture patterns and nuances across all classes, regardless of their frequency in the dataset.

In this context, three resampling methods are employed, namely Random Over Sampling, Random Under Sampling, and SMOTE (Synthetic Minority Oversampling Technique).

Random Over Sampling

Random over sampling is a technique used to address class imbalance in machine learning datasets. In this method, instances from the minority class are randomly replicated and added to the training dataset until the class distribution is balanced or reaches a desired ratio between the minority and majority classes. By increasing the number of instances in the minority class, random over sampling helps prevent classifiers from being biased towards the majority class and improves their ability to learn patterns from the data. However, random over sampling may lead to overfitting, especially when the minority class is already well represented in the dataset.

Random Under Sampling

It involves randomly removing instances from the majority class to balance the class distribution with the minority class. By reducing the number of instances in the majority class, the dataset becomes more balanced, which can help classifiers better learn from the data and improve their performance, especially on the minority class. However, random under sampling may lead to information loss since it removes instances from the majority class without considering their importance or relevance to the classification task. Therefore, it is essential to carefully evaluate the trade-offs and potential impact on model performance when employing random under sampling.

Synthetic Minority Over-sampling Technique or SMOTE

The Synthetic Minority Over-sampling Technique (SMOTE) represents a pivotal approach in mitigating class imbalance within machine learning datasets. By strategically synthesizing new instances for the minority class, SMOTE effectively addresses the disparity in class distribution. Through a process of feature space interpolation between existing minority class instances and their nearest neighbors, SMOTE generates synthetic samples that accurately reflect the underlying characteristics of the minority class. This technique not only rebalances the dataset but also mitigates the risk of classifier bias towards the majority class. While SMOTE offers a powerful solution to class imbalance, its efficacy is contingent upon factors such as dataset structure, feature space dimensionality, and clustering of minority class instances. Therefore, careful consideration and evaluation of SMOTE's impact on model performance are essential to its successful implementation in machine learning tasks.

1.2.9 Discretization Methods

Feature discretization enhances the performance of certain classification algorithms by transforming continuous attributes into categorical ones. This process involves segmenting continuous features into distinct ranges or intervals, effectively converting numerical data into nominal data. The challenge in feature discretization lies in selecting suitable cut points as continuous data can be discretized in numerous ways. The optimal discretization method strives to identify a minimal number of cut points that effectively partition the data into meaningful bins. Thus, the key lies in locating cut points that facilitate accurate representation of the underlying patterns within the data while minimizing computational complexity and maximizing predictive performance.

1.2.10 Evaluation of Models (Prediction Performance)

In the quest to determine the most suitable model, assessing classifier performance is paramount. Various evaluation metrics are employed to gauge the effectiveness of machine learning algorithms. This study employs a diverse range of evaluation techniques, encompassing measures such as prediction accuracy, sensitivity, specificity, and the area under the curve (AUC). By considering multiple metrics, researchers gain comprehensive insights into classifier performance, enabling informed decisions regarding model selection and optimization strategies.

Confusion Matrix

The terms TP, TN, FN, and FP serve as fundamental components in describing Sensitivity, Specificity, and classification Accuracy. Sensitivity, represented by the formula Sensitivity = TP / (TP + FN), measures the accuracy of correctly identifying positive examples (actual events). On the other hand, Specificity, denoted by Specificity = TN / (TN + FP), quantifies the proportion of correctly identified negative examples (non-actual events). Accuracy, computed as Accuracy = (TN + TP) / (TN + TP + FN + FP), provides an overall measure of correct classifications. An effective classifier must yield highly accurate results for both Sensitivity and Specificity simultaneously, as they are crucial indicators of a model's ability to accurately identify positive and negative instances.

Area Under Receiver Operating Characteristic Curve (AUROC)

The Area Under the Receiver Operating Characteristic (AUROC) curve serves as a crucial metric for assessing the quality of classification. The Receiver Operating Characteristic (ROC) curve provides a graphical representation of a predictive model's performance, illustrating the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) across a spectrum of cutoff points. TPR and FPR are computed using the equations TPR = TP / (TP + FN) and FPR = FP / (FP + TN) respectively. AUROC quantifies the accuracy of the classifier by estimating the probability thresholds for the next event, whether it is positive or negative. It represents the area beneath the ROC curve geometrically. A higher AUROC value corresponds to better classification outcomes, while a value less than 0.5 suggests an ineffective classifier, one that performs worse than random chance. An AUROC of 0.5 indicates a random classifier, while an AUROC of 1 signifies an ideal classifier.

1.2.11 Hyperparameter Tuning

To mitigate the risks of overfitting and underfitting, it's imperative to fine-tune model parameters within stable zones where training and validation scores exhibit minimal fluctuations. The grid search technique, a prominent tool in the realm of insurance analytics, serves as a crucial mechanism for optimizing model parameters. In pursuit of achieving optimal ROC values, GridSearchCV was employed. This method systematically explores a range of parameter combinations to identify the configuration that yields the highest ROC values, ensuring that the model's predictive capabilities are maximized while maintaining robustness and generalizability across different datasets and scenarios.

List of References

(2020). Facts + statistics: Auto insurance. [Online]. Available: https://www.iii.org/fact-statistic/facts-statistics-auto-insurance.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and regression trees. *Wadsworth International Group*, 37.

Briys, E. and Varenne, F. (2001). *Insurance: From Underwriting to Derivatives.*

Columbus, L. (2017). Mckinsey's state of machine learning and ai. Technical report, Forbes.

Dhieb, N., Ghazzai, H., Besbes, H., and Massoud, Y. (2020). A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement.

Gay, F. H., Lai, G. C., Patterson, G. A., and Witt, R. C. (1998). Underwriting cycles in property and liability insurance: An empirical analysis of industry and by-line data. *The Journal of Risk and Insurance*, 65(4):539–61.

Grömping, U. (2009). Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63(4):308–319.

Hanafy, M. and Ming, R. (2021a). Comparing smote family techniques in predicting insurance premium defaulting using machine learning models. *International Journal of Advanced Computer Science and Applications*, 12(9).

Hanafy, M. and Ming, R. (2021b). Improving imbalanced data classification in auto insurance by the data level approaches. *International Journal of Advanced Computer Science and Applications*, 12(6).

Hanafy, M. and Ming, R. (2021c). Machine learning approaches for auto insurance big data. *Risks*, 9(2).

Hanafy, M. and Ming, R. (2021d). Using machine learning models to compare various resampling methods in predicting insurance fraud. *Journal of Theoretical and Applied Information Technology*, 99:2819–2833.

Hanafy, M. and Ming, R. (2022). Classification of the insureds using integrated machine learning algorithms: A comparative study. *Applied Artificial Intelligence*, 36(1):2020489.

Jing, L., Zhao, W., Sharma, K., and Feng, R. (2018). Research on probability-based learning application on car insurance data. In *Proceedings of the 2017 4th International Conference on Machinery, Materials and Computer (MACMC 2017)*, pages 59–63. Atlantis Press.

Kate A. Smith, R. J. W. and Brooks, M. (2000). An analysis of customer retention and insurance claim patterns using data mining: a case study. *Journal of the Operational Research Society*, 51(5):532–541.

Kormaksson, M., Kelly, L. J., Zhu, X., Haemmerle, S., Pricop, L., and Ohlssen, D. (2020). Sequential knockoffs for continuous and categorical predictors: With application to a large psoriatic arthritis clinical trial pool. *Statistics in Medicine*, 40:3313–3328.

Pesantez-Narvaez, J., Guillen, M., and Manuela, A. (2019). Predicting motor insurance claims using telematics data -xgboost versus logistic regression. *Risks*, 70(7).

Rawat, S., Rawat, A., Kumar, D., and Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector.

Shady, A., Khaled, E., and Mohamed, A. (2020). A proposed model to predict auto insurance claims using machine learning techniques. *Journal of Theoretical and Applied Information Technology*, 98:3428–3437.

CHAPTER 2

Data Achieving and Preprocessing

2.1 The Data

In our investigative analysis of insurance claims, we harnessed three distinct datasets. The initial two datasets pertain to the occurrences of claims in the realm of car insurance, while the third dataset specifically focuses on defaulters within the domain of car insurance. Detailed information about these datasets, along with a comprehensive description of their variables, can be found in Table 1. Dataset 1 was originally featured in the Dataverse Hackathon organized by Analytics Vidhya on November 14, 2022. Dataset 2, was acquired from Kaggle.com. Comprising a total of 10000 rows and 19 columns. Dataset 3, utilized in this study, was also sourced from Kaggle.com. It has 79,853 rows and 17 columns. This particular dataset was previously employed by Hanafi et al. in their research. These datasets are meticulously maintained with a steadfast commitment to safety and confidentiality. Client personal information is encrypted to uphold stringent privacy standards.

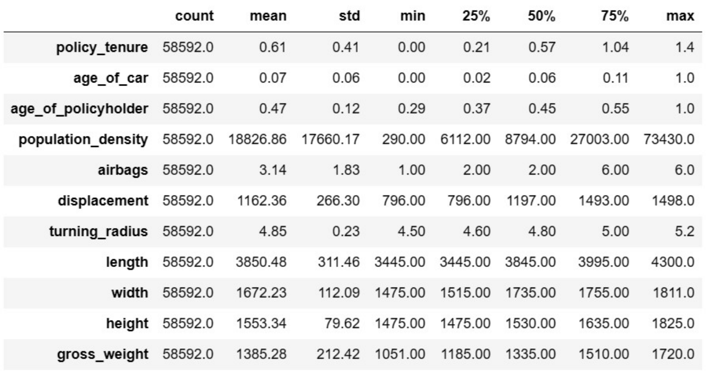
Dataset 1	Dataset 2	Dataset 3
The **dataset 1** used in this case study was sourced from Kaggle.com. There are 58592 rows and 44 features in total. Dataset was a part of dataverse hack - hackathon by analytics vidhya on the 14th of November 2022. The details of all the columns in the dataset are as follows: (1) policy_id- Unique identifier of the policyholder. (2) policy_tenure- Time period of the policy. (3) age_of_car- Normalized age of the car in years. (4) age_of_policyholder- Normalized age of policyholder in years. (5) area_cluster- Area cluster of the policyholder. (6) population density- Population density of the city (Policyholder City). (7) make- Encoded Manufacturer/company of the car. (8) segment- Segment of the car (A/ B1/ B2/ C1/ C2). (9) model- Encoded name of the car. (10) fuel_type- Type of fuel used by the car. (11) max_torque- Maximum Torque generated by the car (Nm@rpm). (12) max_power- Maximum Power generated by the car (bhp@rpm). (13) engine_type- Type of engine used in the car. (14) airbags- Number of airbags installed in the car. (15) is_esc- Boolean flag indicating whether Electronic Stability Control (ESC) is present in the car or not. (16) is_adjustable_steering- Boolean flag indicating whether the steering wheel of the car is adjustable or not. (17) is_tpms- Boolean flag indicating whether Tyre Pressure Monitoring System (TPMS) is present in the car or not. (18) is_parking_sensors- Boolean flag indicating whether parking sensors are present in the car or not. (19) is_parking_camera- Boolean flag indicating whether the parking camera is present in the car or not. (20) rear_brakes_type- Type of brakes used in the rear of the car. (21) displacement- Engine displacement of the car (cc). (22) cylinder- Number of cylinders present in the engine of the car. (23) transmission_type- Transmission type of the car. (24) gear_box- Number of gears in the car. (25) steering_type- Type of the power steering present in the car. (26) turning_radius- The space a vehicle needs to make a certain turn (Meters). (27) length- Length of the car (Millimetre). (28) width- Width of the car (Millimetre). (29) height- Height of the car (Millimetre). (30) gross_weight- The maximum allowable weight of the fully-loaded car. (31) including passengers, cargo and equipment (Kg). (32) is_front_fog_lights- Boolean flag indicating whether front fog lights are available in the car or not. (33) is_rear_window_wiper- Boolean flag indicating whether the rear window wiper is available in the car or not. (34) is_rear_window_washer- Boolean flag indicating whether the rear window washer is available in the car or not. (35) is_rear_window_defogger- Boolean flag indicating whether rear window defogger is available in the car or not. (36) is_brake_assist- Boolean flag indicating whether the brake assistance feature is available in the car or not. (37) is_power_door_lock- Boolean flag indicating whether a power door lock is available in the car or not. (38) is_central_locking- Boolean flag indicating whether the central locking feature is available in the car or not. (39) is_power_steering- Boolean flag indicating whether power steering is available in the car or not. (40) is_driver_seat_height_adjustable- Boolean flag indicating whether the height of the driver seat is adjustable or not. (41) is_day_night_rear_view_mirror- Boolean flag indicating whether day & night rearview mirror is present in the car or not. (42) is_ecw- Boolean flag indicating whether Engine Check Warning (ECW) is available in the car or not. (43) is_speed_alert- Boolean flag indicating whether the speed alert system is available in the car or not. (44) ncap_rating- Safety rating given by NCAP (out of 5). (45) is_claim- Outcome: Boolean flag indicating whether the policyholder filed a claim or not.	The **dataset 2** under examination in this case study, was procured from Kaggle.com. The dataset consists of 10000 rows and 19 columns. The details are as follows: (1) ID:Unique Customer Id (2) AGE: Age of driver(Categorical) (3) GENDER (4) RACE (5) DRIVING_EXPERIENCE Years(Categorical) (6) EDUCATION (7) INCOME (Categorical) (8) CREDIT_SCORE: Credit Score of the driver. (9) VEHICLE_OWNERSHIP (10) VEHICLE_YEAR: Manufacturing year of the vehicle (Categorical). (11) MARRIED (12) Children(Categorical) (13) POSTAL_CODE. (14) ANNUAL_MILEAGE (Numerical) (15) VEHICLE_TYPE (16) SPEEDING_VIOLATIONS: Number of speeding violations. (17) DUIS: Number of DUI. (18) PAST_ACCIDENTS: Number of past accidents. (19) OUTCOME: The customer has claimed or not.	The **dataset 3** used in this study was also collected from Kaggle.com. 79,853 rows and 17 columns make up the dataset's total. The details of all the columns in the dataset: (1) id: Unique customer ID. (2) percent of the premium paid by cash credit. (3) age in days: age of the customer in days. (4) Income: Income of the customer. (5) Count_3-6_months_late: Number of times premium was paid 3–6 months late. (6) Count_6-12_months_late: Number of times premium was paid 6–12 months late. (7) Count_more_than_12_months_late: Number of times premium was paid more than 12 months late. (8) Application_underwriting_score: Risk score of customers. (9) number of premiums paid: Number of premiums paid till date. (10) sourcing channel:Channel through which customer was sourced. (11) residence area type:Residence type of the customer. (12) premium: Total premium amount paid till now. (13) default: 0 indicates that customer has defaulted the premium and 1 indicates that customer has not defaulted. (14) Marital Status: Married/Unmarried. (15) Number of vehicles. (16) Number of dependents. (17) Accommodation: Owned /Rented.

Table 1: Data Review

2.2 Exploratory Data Analysis(EDA)

Exploratory data analysis (EDA) serves as the foundational step preceding the training of predictive models. It involves a comprehensive investigation of the dataset's characteristics, aiming to unveil insights crucial for subsequent modeling decisions. Through EDA, one can delve into summary statistics, distributions, and relationships among variables, identifying potential outliers and missing values. This process aids in feature selection by pinpointing the most influential variables for predictive accuracy while illuminating any redundancies or irrelevancies. Furthermore, EDA facilitates the detection of patterns and trends, guiding appropriate model selection and preprocessing strategies. By addressing data quality issues and assessing model assumptions, EDA ensures the reliability and validity of the ensuing predictive models.

This section endeavors to extract meaningful insights from each dataset, starting with an analysis of the continuous variables within Dataset 1. We explore the summary statistics and correlations for numerical variables. Additionally, we juxtapose the distributions of categorical variables for overall and claimed instances. However, we only present variables exhibiting significant differences between the distributions of the overall dataset and the claimed subset.



	count	mean	std	min	25%	50%	75%	max
policy_tenure	58592.0	0.61	0.41	0.00	0.21	0.57	1.04	1.4
age_of_car	58592.0	0.07	0.06	0.00	0.02	0.06	0.11	1.0
age_of_policyholder	58592.0	0.47	0.12	0.29	0.37	0.45	0.55	1.0
population_density	58592.0	18826.86	17660.17	290.00	6112.00	8794.00	27003.00	73430.0
airbags	58592.0	3.14	1.83	1.00	2.00	2.00	6.00	6.0
displacement	58592.0	1162.36	266.30	796.00	796.00	1197.00	1493.00	1498.0
turning_radius	58592.0	4.85	0.23	4.50	4.60	4.80	5.00	5.2
length	58592.0	3850.48	311.46	3445.00	3445.00	3845.00	3995.00	4300.0
width	58592.0	1672.23	112.09	1475.00	1515.00	1735.00	1755.00	1811.0
height	58592.0	1553.34	79.62	1475.00	1475.00	1530.00	1635.00	1825.0
gross_weight	58592.0	1385.28	212.42	1051.00	1185.00	1335.00	1510.00	1720.0

Figure 1: Summary of Numerical Variables of Dataset 1.

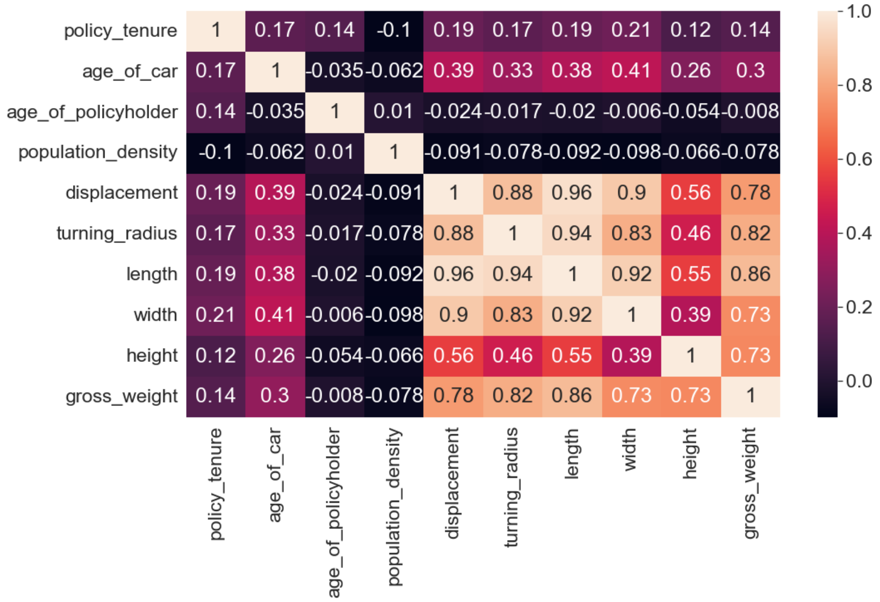


Figure 2: Correlatin Matrix of Numerical Variables of Dataset 1

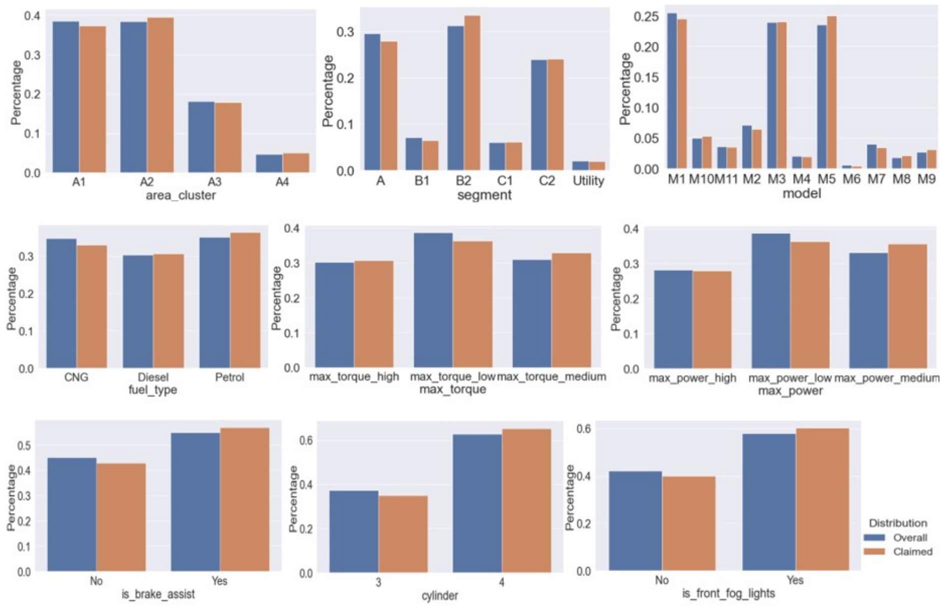
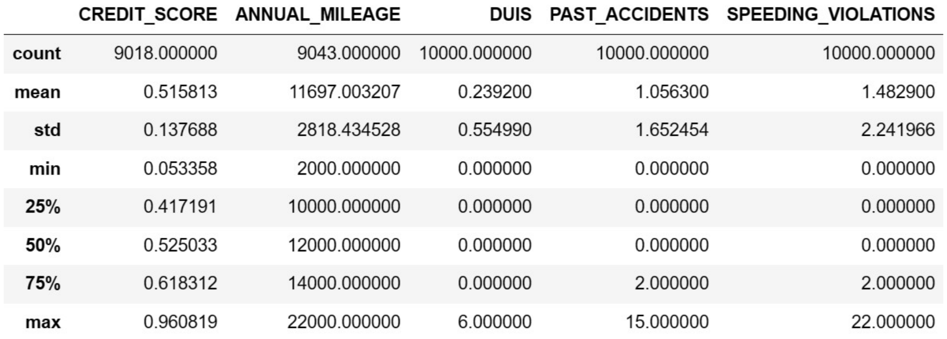


Figure 3: The distribution of Categorical Variables of Dataset 1.

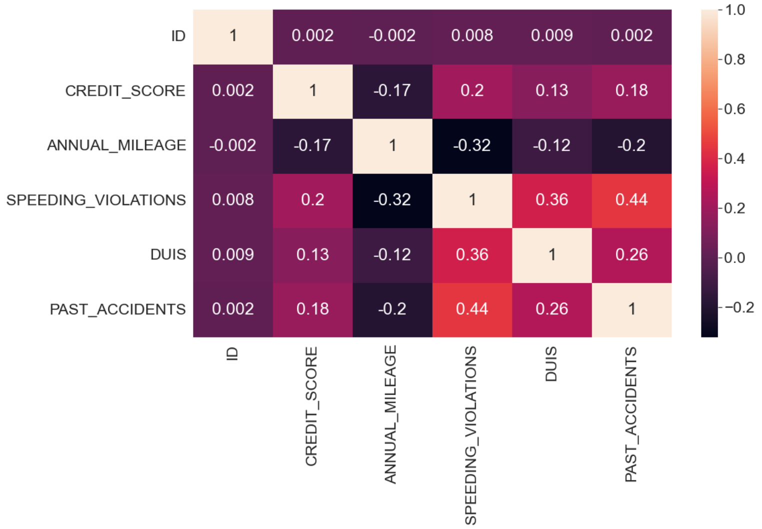
Analyzing Figure 1 and 2 sheds light on the characteristics of continuous variables in Dataset 1. It is recommended to normalize these variables before proceeding with subsequent analyses. Additionally, the correlation matrix unveils interesting patterns, indicating a significant correlation among various car-related features, while the correlation between policy holder attributes and car characteristics remains relatively low.

Another focal point of our inquiry involves a comparative scrutiny of the overall distribution and distribution under a claim for categorical variables in Dataset 1. Figure 3 exclusively highlights variables that exhibit substantial differences in both overall distribution and distribution under a claim. This selective representation aims to underscore and clarify significant variations within these categorical variables. Similar exploratory data analyses were conducted for Dataset 2 and Dataset 3.



	CREDIT_SCORE	ANNUAL_MILEAGE	DUIS	PAST_ACCIDENTS	SPEEDING_VIOLATIONS
count	9018.000000	9043.000000	10000.000000	10000.000000	10000.000000
mean	0.515813	11697.003207	0.239200	1.056300	1.482900
std	0.137688	2818.434528	0.554990	1.652454	2.241966
min	0.053358	2000.000000	0.000000	0.000000	0.000000
25%	0.417191	10000.000000	0.000000	0.000000	0.000000
50%	0.525033	12000.000000	0.000000	0.000000	0.000000
75%	0.618312	14000.000000	0.000000	2.000000	2.000000
max	0.960819	22000.000000	6.000000	15.000000	22.000000

Figure 4: Summary of Numerical Variables of Dataset 2.



	ID	CREDIT_SCORE	ANNUAL_MILEAGE	SPEEDING_VIOLATIONS	DUIS	PAST_ACCIDENTS
ID	1	0.002	-0.002	0.008	0.009	0.002
CREDIT_SCORE	0.002	1	-0.17	0.2	0.13	0.18
ANNUAL_MILEAGE	-0.002	-0.17	1	-0.32	-0.12	-0.2
SPEEDING_VIOLATIONS	0.008	0.2	-0.32	1	0.36	0.44
DUIS	0.009	0.13	-0.12	0.36	1	0.26
PAST_ACCIDENTS	0.002	0.18	-0.2	0.44	0.26	1

Figure 5: Correlatin Matrix of Numerical Variables of Dataset 2.

We have also identified noticeable skewness and potential outliers in variables such as DUIS, PAST_ACCIDENTS, and SPEEDING_VIOLATIONS. To address this, we've chosen to discretize these variables into distinct groups. Additionally, CREDIT_SCORE has already been standardized, while ANNUAL_MILEAGE has undergone standardization to mitigate the influence of high-magnitude values. The distribution of categorical variables, alongside the newly discretized ones, is presented below for further examination.

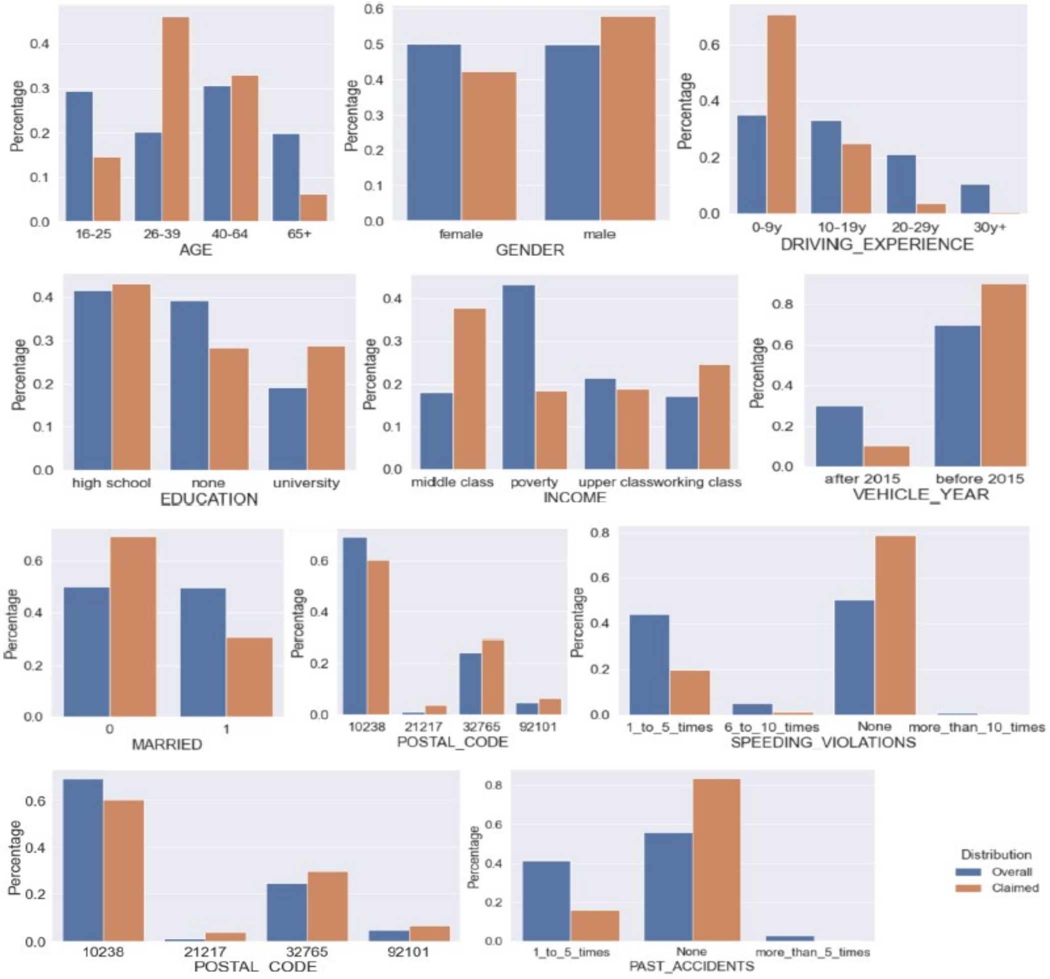
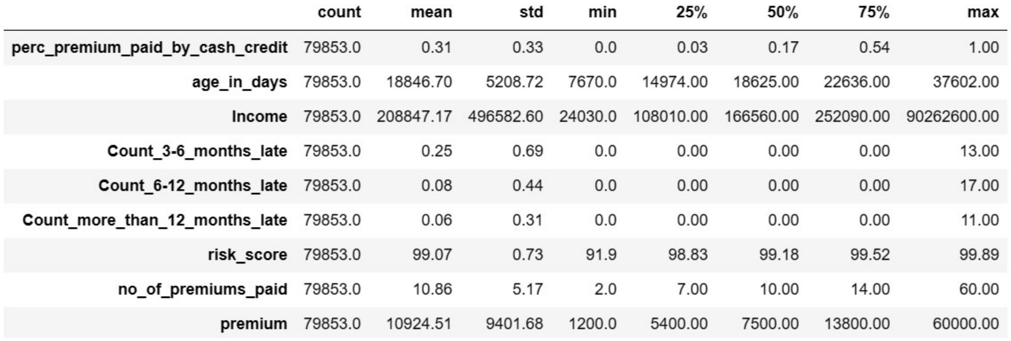


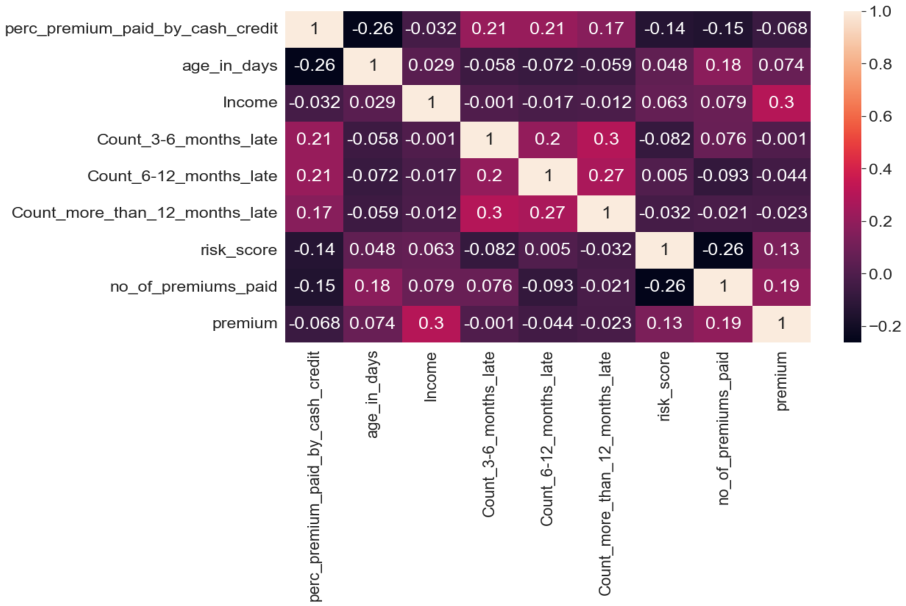
Figure 6: The distribution of Categorical Variables of Dataset 2.

Figure 6 illustrates that variables such as AGE, DRIVING EXPERIENCE, VEHICLE YEAR, and others can exert a notable influence on insurance claims. We can now move forward with the exploratory data analysis of the predictors within data set 3.



	count	mean	std	min	25%	50%	75%	max
perc_premium_paid_by_cash_credit	79853.0	0.31	0.33	0.0	0.03	0.17	0.54	1.00
age_in_days	79853.0	18846.70	5208.72	7670.0	14974.00	18625.00	22636.00	37602.00
Income	79853.0	208847.17	496582.60	24030.0	108010.00	166560.00	252090.00	90262600.00
Count_3-6_months_late	79853.0	0.25	0.69	0.0	0.00	0.00	0.00	13.00
Count_6-12_months_late	79853.0	0.08	0.44	0.0	0.00	0.00	0.00	17.00
Count_more_than_12_months_late	79853.0	0.06	0.31	0.0	0.00	0.00	0.00	11.00
risk_score	79853.0	99.07	0.73	91.9	98.83	99.18	99.52	99.89
no_of_premiums_paid	79853.0	10.86	5.17	2.0	7.00	10.00	14.00	60.00
premium	79853.0	10924.51	9401.68	1200.0	5400.00	7500.00	13800.00	60000.00

Figure 7: Summary of Numerical Variables of Dataset 3



	perc_premium_paid_by_cash_credit	age_in_days	Income	Count_3-6_months_late	Count_6-12_months_late	Count_more_than_12_months_late	risk_score	no_of_premiums_paid	premium
perc_premium_paid_by_cash_credit	1	-0.26	-0.032	0.21	0.21	0.17	-0.14	-0.15	-0.068
age_in_days	-0.26	1	0.029	-0.058	-0.072	-0.059	0.048	0.18	0.074
Income	-0.032	0.029	1	-0.001	-0.017	-0.012	0.063	0.079	0.3
Count_3-6_months_late	0.21	-0.058	-0.001	1	0.2	0.3	-0.082	0.076	-0.001
Count_6-12_months_late	0.21	-0.072	-0.017	0.2	1	0.27	0.005	-0.093	-0.044
Count_more_than_12_months_late	0.17	-0.059	-0.012	0.3	0.27	1	-0.032	-0.021	-0.023
risk_score	-0.14	0.048	0.063	-0.082	0.005	-0.032	1	-0.26	0.13
no_of_premiums_paid	-0.15	0.18	0.079	0.076	-0.093	-0.021	-0.26	1	0.19
premium	-0.068	0.074	0.3	-0.001	-0.044	-0.023	0.13	0.19	1

Figure 8: Correlatin Matrix of Numerical Variables of Dataset 3.

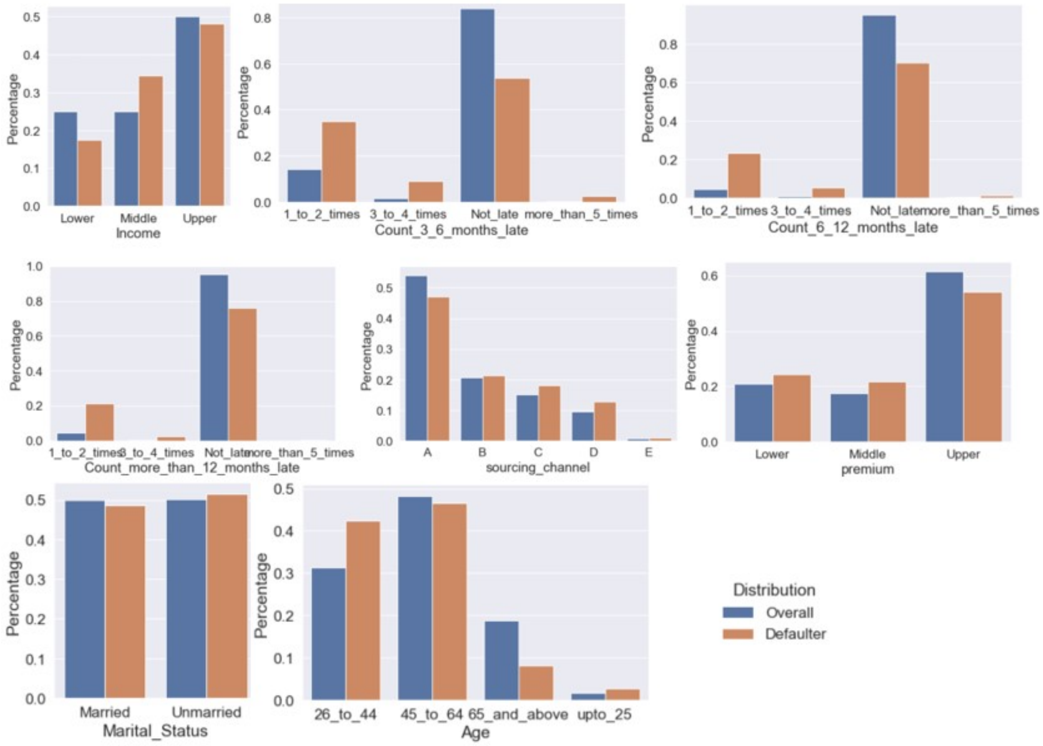


Figure 9: The distribution of Categorical Variables of Dataset 3.

For Dataset 3, the numerical variables like age_in_days, Income and premium are discretized. Other numerical variables are normalized. Besides, the categorical variables like Count_3-6_months_late,Count_6-12_months_late and Count_more_than_12_months_late are merged into smaller number of categories to improve the performance of the predicltive models.

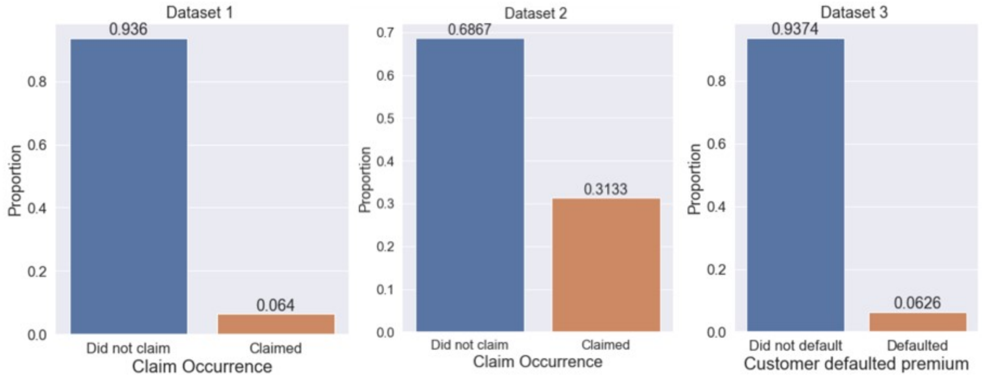


Figure 10: Distribution of response variable.

Moving on to Figure 10, it illustrates the distribution of claim occurrences or defaulters across all three datasets. As anticipated, the distributions appear highly imbalanced, prompting consideration of oversampling or undersampling techniques before delving into the fitting of various classification models.

2.3 Data Source

Dataset 1: https://www.kaggle.com/datasets/avikumart/analytics-vidhya-nov22-insurance-claims-dataset

Dataset 2: https://www.kaggle.com/datasets/sagnik1511/car-insurance-data

Dataset 3: https://www.kaggle.com/prakharrathi25/premium-default-prediction/data

CHAPTER 3

Methodology

Our study unfolds through a systematic progression comprising 10 key steps. Commencing with the initial stride, we embark on data collection for subsequent analysis. The three datasets integral to this study have been comprehensively detailed. The procedural roadmap, elucidated as a visual representation in Figure 11, delineates the multifarious stages of our analytical journey.

The initial phase involves data cleaning, primarily addressing missing values. Subsequently, categorical variables undergo transformation, while continuous variables undergo normalization, rendering them apt for machine learning algorithms. In our analysis of Dataset 1 , we opted for Lasso and Random Forest algorithms to pinpoint significant features. Conversely, for Dataset 2 and 3, feature selection was achieved through the application of the Lasso, Random Forest and Knockoffs algorithm. The datasets are then partitioned into training and testing subsets, adhering to a 70:30 split. Recognizing the high imbalance in the target variable, as illustrated in Figure 13, resampling techniques—namely regular oversampling, undersampling, and Synthetic Minority Over-sampling Technique (SMOTE)—are employed to rectify the imbalance within the training datasets.

The subsequent steps encompass the fitting of machine learning models, including K-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, and Extreme Gradient Boosting onto the designated training datasets. Evaluation of these models is executed using the corresponding test datasets. Ultimately, a comparative analysis unfolds, scrutinizing the accuracy, sensitivity, specificity,and AUC-ROC score across diverse model-performance metrics for each strategy combination. This comprehensive approach serves to discern and articulate the efficacy of the models under various strategic combinations.

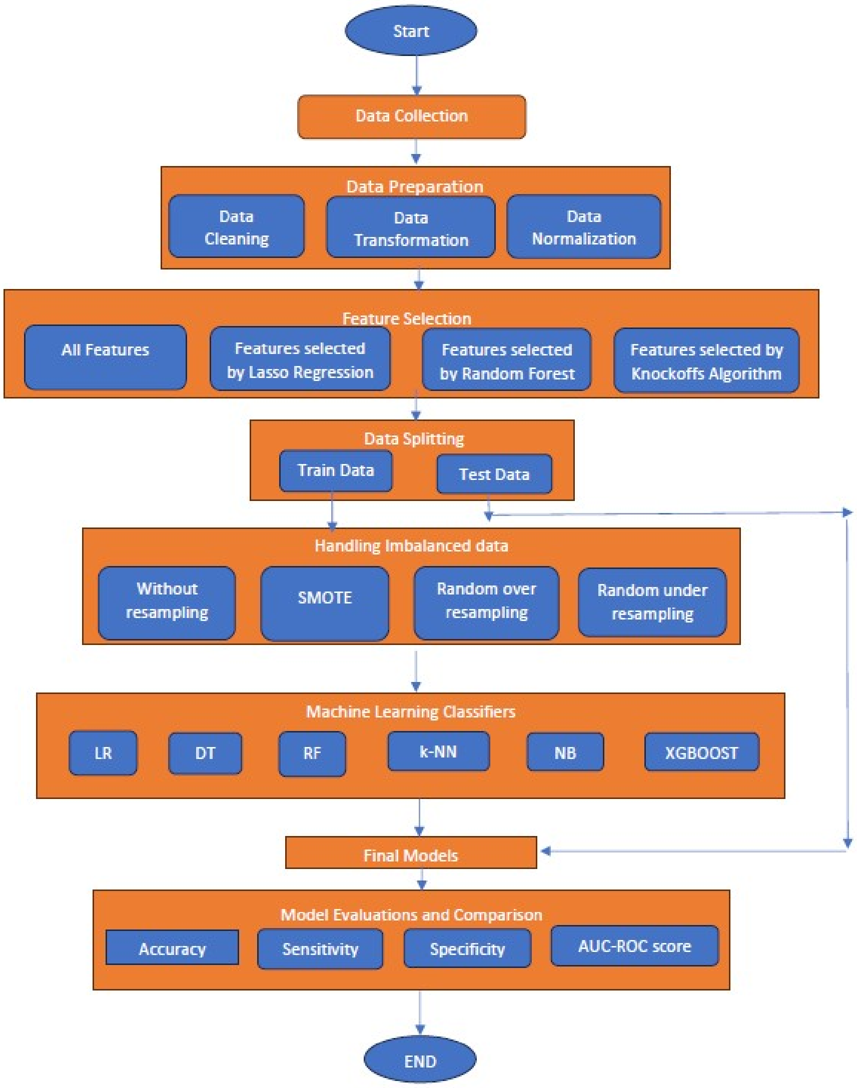


Figure 11: Working Diagram.

CHAPTER 4

Results

Following the outlined methodology, the feature analyis of Dataset 1 with Lasso and Random Forest algorithms respectively, along with Dataset 2 and 3 using the Lasso, Random Forest and Knockoffs algorithm, provided insightful results. After partitioning the datasets into training and testing subsets, adhering to a 70:30 split, the high imbalance in the target variable was addressed using resampling techniques such as regular oversampling, undersampling, and Synthetic Minority Over-sampling Technique (SMOTE). We have used a set of abbreviations in our representation. US- Random Under Sampling, OS- Random Over Sampling, SM- SMOTE, FSL- Feature Selection by Lasso, FSR- Feature Selection by Random Forest, and FSK- Feature Selection by Knockoffs.

Subsequently, a variety of machine learning models including K-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, Extreme Gradient Boosting, and Naive Bayes were fitted onto the designated training datasets. Through evaluation using the corresponding test datasets, a comparative analysis unfolded. This analysis scrutinized key performance metrics including accuracy, sensitivity, specificity, and AUC-ROC score across diverse model-performance metrics for each strategy combination.

Upon completing exhaustive computations, we have compiled the outcomes for the various datasets under consideration. Across these datasets, we have calculated metrics such as accuracy, sensitivity, specificity, and AUC-ROC score for all conceivable combinations of predictive models and accompanying methods. This meticulous analysis provides a comprehensive overview of the performance of different models and techniques, facilitating informed decisions regarding the most suitable approaches for predictive modeling in each scenario.

4.1 Results from Dataset 1

Knockoffs' Algorithm was not very successful for dataset 1. The reason maybe the existence of several categorical types in some variables. We have used random forest and logistic Lasso regularization to identify important features from dataset 1.

LASSO Feature Selection (23)	**Random Forest Feature Selection (25)**
policy_tenure,age_of_car,	policy_tenure,age_of_car,
age_of_policyholder,population_density,	age_of_policyholder, population_density
area_clusterA2,area_clusterA3,	area_clusterA2,area_clusterA3,
area_clusterA4,	area_clusterA4,
fuel_typePetrol,height	fuel_typePetrol,height,
max_torquemax_torque_low	max_torquemax_torque_low,
max_powermax_power_low	max_powermax_power_low,
max_powermax_power_medium	max_powermax_power_medium,
modelM11,modelM2,	max_torquemax_torque_medium,
	modelM8 ,
is_adjustable_steeringYes ,	is_adjustable_steeringYes,
is_parking_cameraYes,	is_parking_cameraYes,
steering_typePower,	steering_typePower,
is_power_door_locksYes,	is_power_door_locksYes,
is_ecwYes,	is_ecwYes,
rear_brakes_typeDrum,	displacement,ncap_rating,
transmission_typeManua,l	is_front_fog_lightsYes,
gear_box6,	is_brake_assistYes,
is_central_lockingYes	steering_typeManual,
	is_day_night_rear_view_mirrorYes

Table 2: Important features of Dataset 1.

The results obtained from the respective models for all possible combinations different methods are as follows.

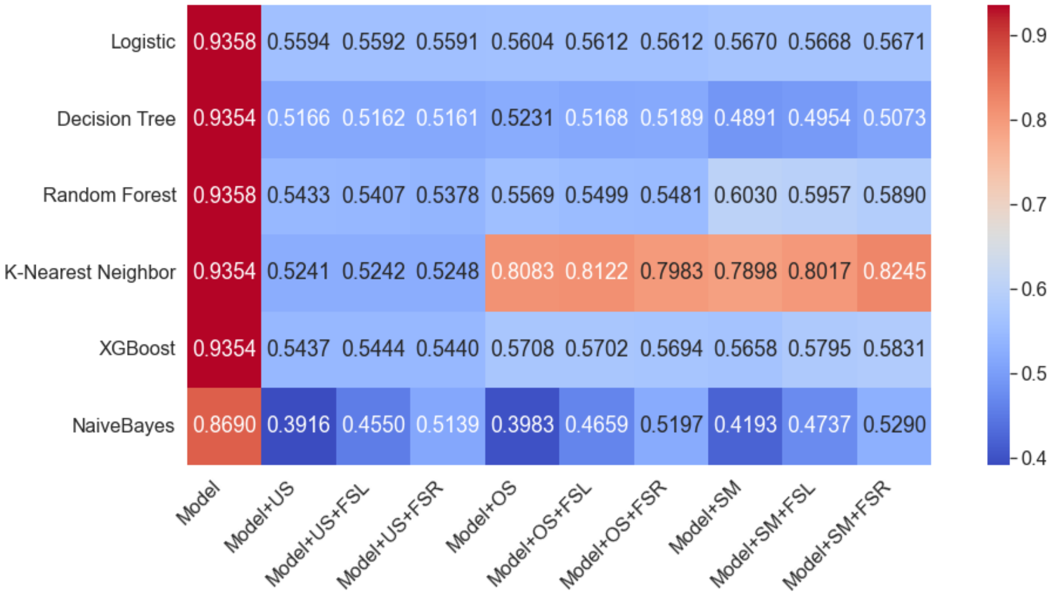


Figure 12: Accuracy Scores for Dataset 1.

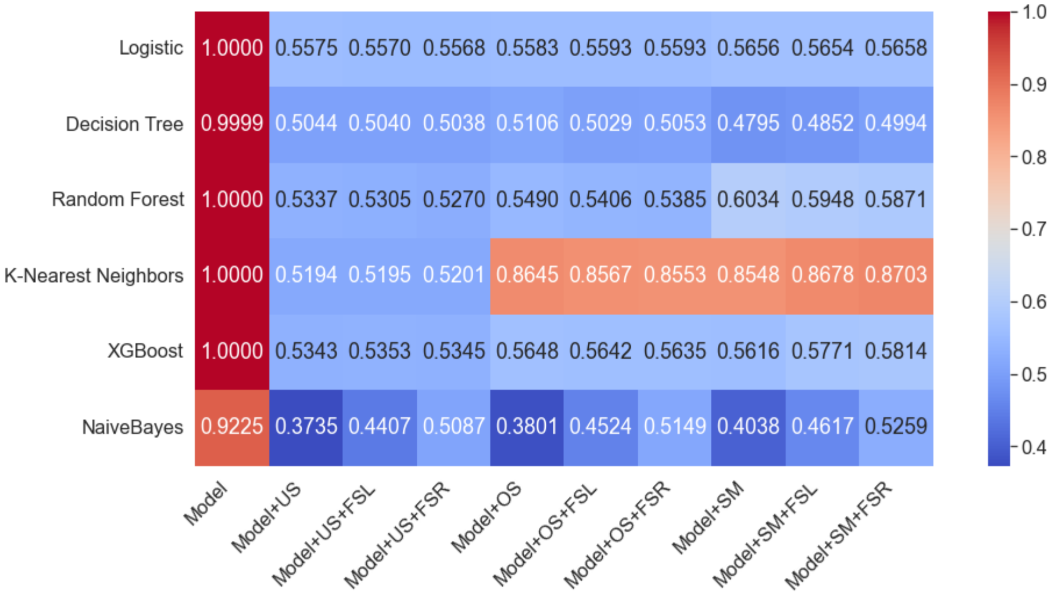


Figure 13: Sensitivity Scores for Dataset 1.

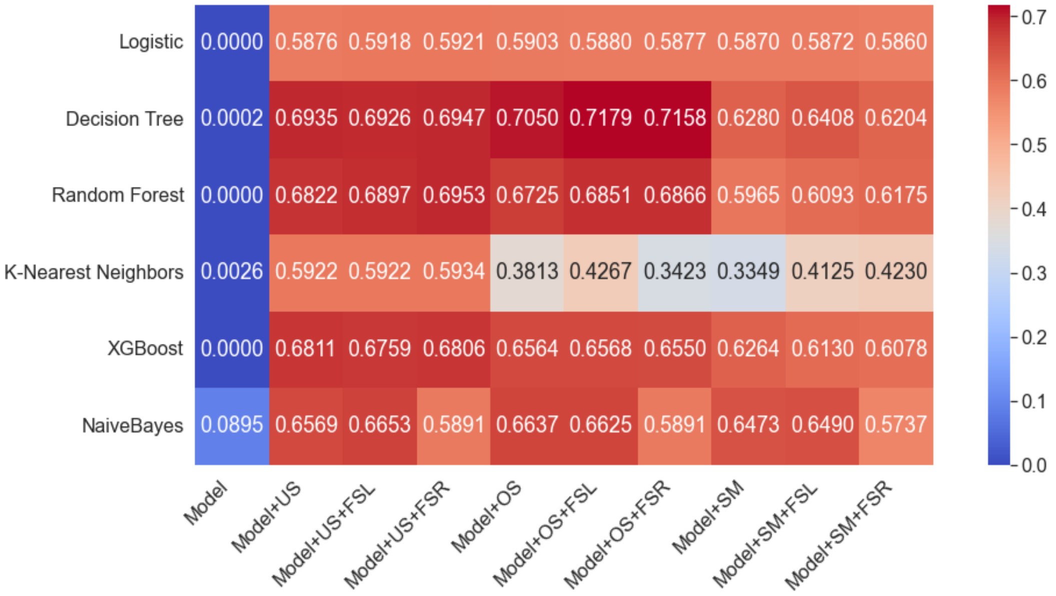


Figure 14: Specificity Scores for Dataset 1.

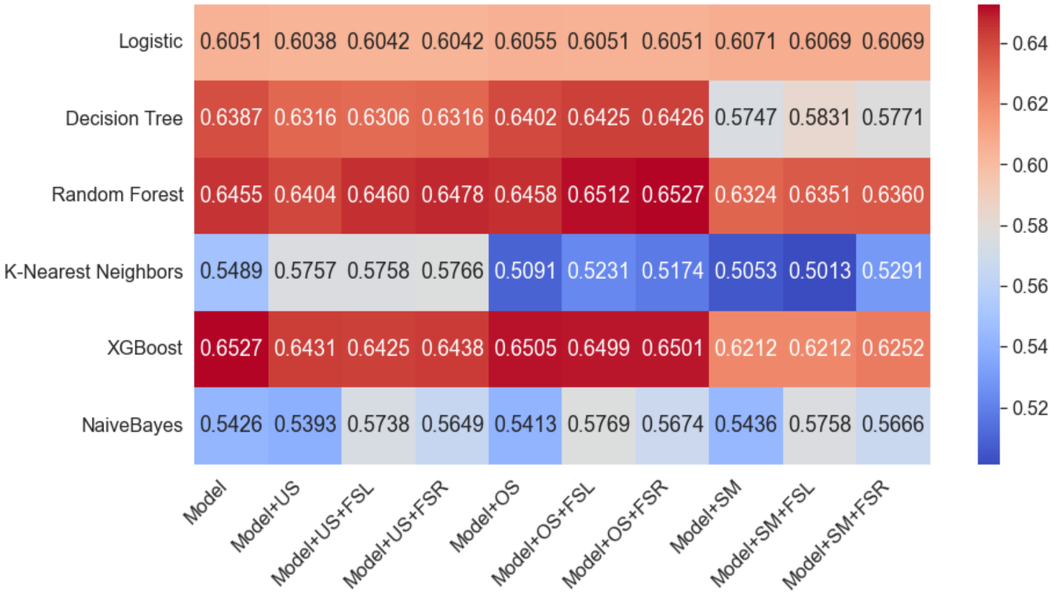


Figure 15: ROC AUC Scores for Dataset 1.

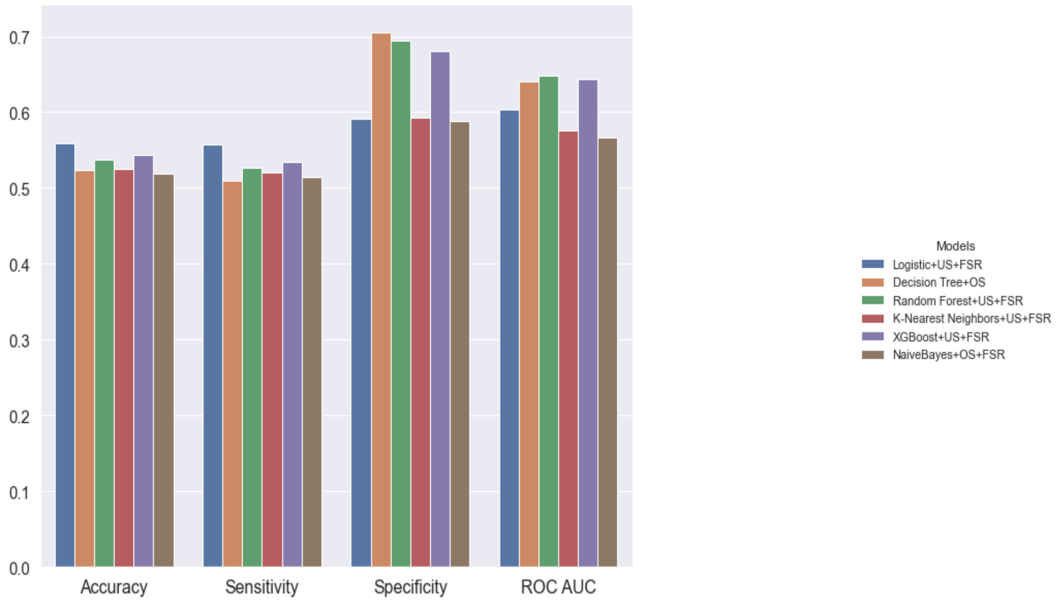


Figure 16: Comparison of best performing combinations from each model for Dataset 1.

In this context, our primary focus lies on achieving high specificity while maintaining reasonable sensitivity and ROC AUC score. Specificity plays a crucial role as it indicates the proportion of actually claimed customers correctly predicted as claimed. It is evident that the best-performing models for dataset 1 are Decision Tree+OS, Random Forest+US+FSR and XGboost+US+FSR.

4.2 Results from Dataset 2

We have applied the Knockoffs' Algorithm, random forest and logistic Lasso Regularization to dataset 2 to identify important features. Features that appeared in more than 75% of the trials were identified as important features. The important features selected by all three methods respectively are mentioned below.

LASSO Feature Selection (18)	**Random Forest Feature Selection (20)**	**Knockoffs' Feature Selection (16)**
AGE26.39, AGE65., GENDERmale, RACEminority, DRIVING_EXPERIENCE10.19y, DRIVING_EXPERIENCE20.29y, DRIVING_EXPERIENCE30y., INCOMEpoverty, INCOMEupper.class, VEHICLE_OWNERSHIP1, VEHICLE_YEARbefore.2015, MARRIED1, POSTAL_CODE21217, POSTAL_CODE32765, POSTAL_CODE92101, ANNUAL_MILEAGE, SPEEDING_VIOLATIONSNone, PAST_ACCIDENTSNone	PAST_ACCIDENTSNone, VEHICLE_OWNERSHIP1, SPEEDING_VIOLATIONSNone, VEHICLE_YEARbefore.2015, DRIVING_EXPERIENCE20.29y, DRIVING_EXPERIENCE10.19y, INCOMEupper.class, CREDIT_SCORE, INCOMEpoverty, POSTAL_CODE21217, DRIVING_EXPERIENCE30y., MARRIED1, AGE26.39, GENDERmale, AGE65., DUISNone, AGE40.64, ANNUAL_MILEAGE, POSTAL_CODE32765, CHILDREN1	VEHICLE_YEARbefore.2015, VEHICLE_OWNERSHIP1, POSTAL_CODE92101, POSTAL_CODE32765, POSTAL_CODE21217, MARRIED1, GENDERmale, DRIVING_EXPERIENCE10.19y, DRIVING_EXPERIENCE20.29y, DRIVING_EXPERIENCE30y., ANNUAL_MILEAGE, PAST_ACCIDENTSNone, DUISmore_than_3_times, AGE26.39, INCOMEpoverty, RACEminority,CHILDREN1

Table 3: Important features of Dataset 2.

The results obtained from the respective models for all possible combinations of different methods are as follows.

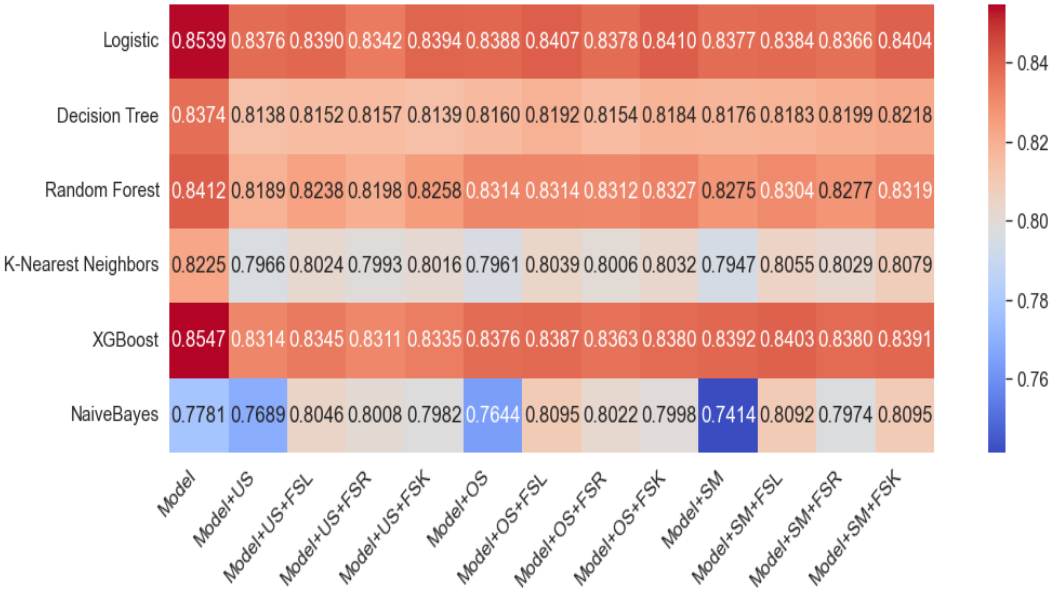


Figure 17: Accuracy Scores for Dataset 2.

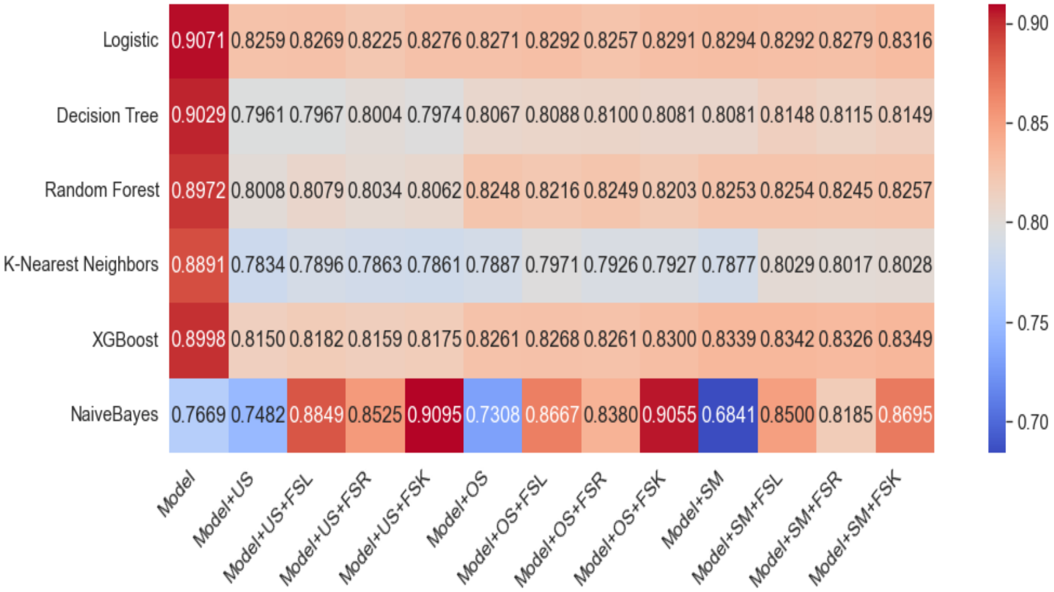


Figure 18: Sensitivity Scores for Dataset 2.

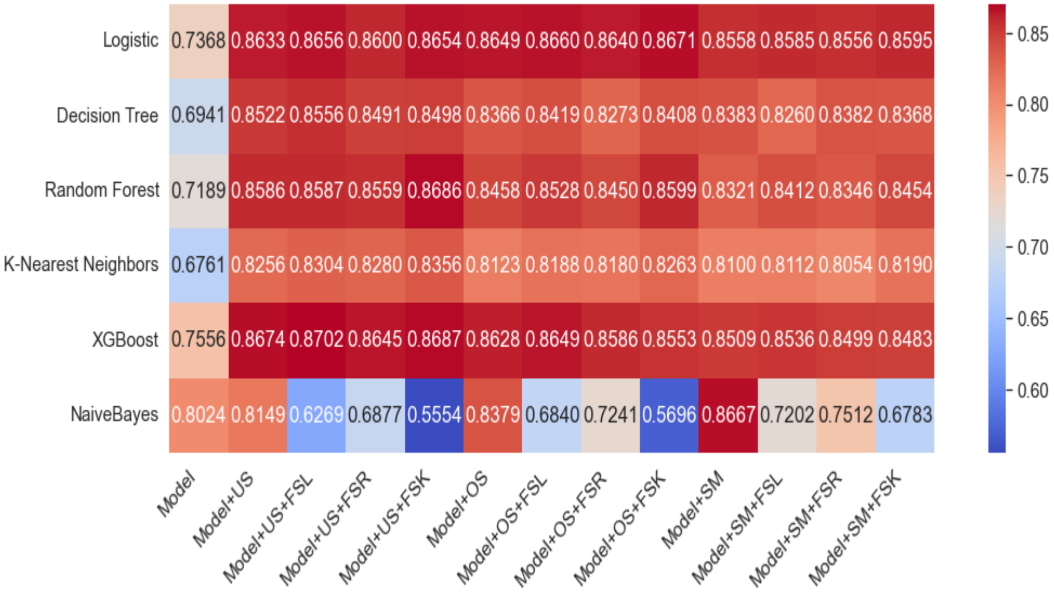


Figure 19: Specificity Scores for Dataset 2.

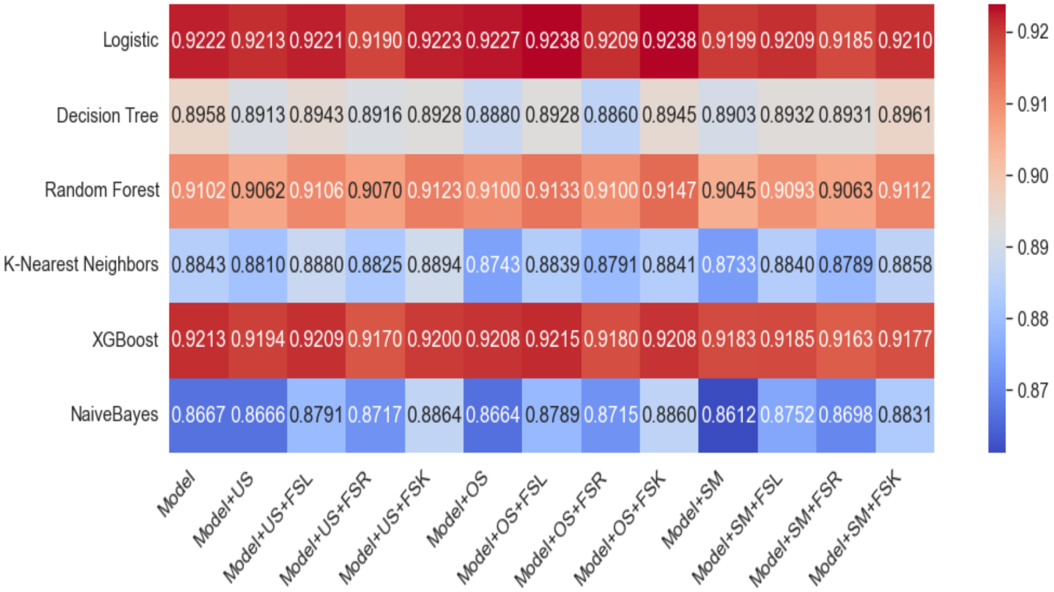


Figure 20: ROC AUC Scores for Dataset 2.

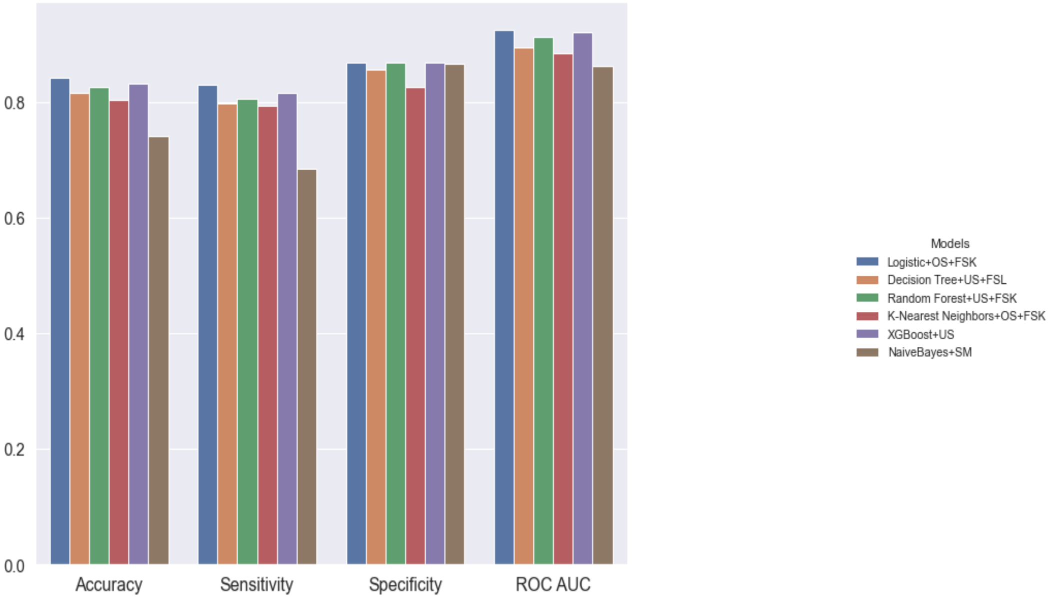


Figure 21: Comparison of best performing combinations from each model for Dataset 2.

In Dataset 2, the majority of models demonstrated strong performance when combined with either under-sampling (US), over-sampling (OS), or SMOTE (SM) alongside various feature selection methods. The top-performing models in this scenario are Logistic+OS+FSK, Random Forest +US + FSK and XGBoost+ US. Notably, the Knockoffs Algorithm exhibited a significant impact in identifying important features, highlighting its efficacy in enhancing model performance.

4.3 Results from Dataset 3

Just like Dataset 2, we have applied the Knockoffs' Algorithm, random forest and logistic Lasso Regularization to dataset 3 to identify important features. The important features selected by all three methods respectively are mentioned below.

LASSO Feature Selection (18)	**Random Forest Feature Selection (20)**	**Knockoffs' Feature Selection (15)**
perc_premium_paid_by_cash_credit,	perc_premium_paid_by_cash_credit,	perc_premium_paid_by_cash_credit,
IncomeMiddle, premiumMiddle,	IncomeMiddle, premiumMiddle,	IncomeMiddle, premiumMiddle,
premiumUpper,	premiumUpper, IncomeUpper,	IncomeUpper,
no_of_premiums_paid ,	no_of_premiums_paid ,	no_of_premiums_paid ,
Age45_to_64,Age65_and_above,	Age45_to_64,Age65_and_above,	Age45_to_64,Age65_and_above,
	Ageupto_25,	Ageupto_25,
No_of_dep,risk_score,	No_of_dep,risk_score,	risk_score,
Marital_StatusUnmarried,		
sourcing_channelC,sourcing_channelD,	sourcing_channelD,	
Count_3_6_months_late3_to_4_times,	Count_3_6_months_late3_to_4_times,	Count_3_6_months_late3_to_4_times,
Count_3_6_months_lateNot_late,	Count_3_6_months_lateNot_late,	Count_3_6_months_lateNot_late,
	Count_3_6_months_latemore_than_5_times,	Count_3_6_months_latemore_than_5_times,
Count_6_12_months_late3_to_4_times,	Count_6_12_months_late3_to_4_times,	Count_6_12_months_late3_to_4_times,
Count_6_12_months_lateNot_late,	Count_6_12_months_lateNot_late,	Count_6_12_months_lateNot_late,
	Count_6_12_months_latemore_than_5_times,	Count_6_12_months_latemore_than_5_times
Count_more_than_12_months_lateNot_late,	Count_more_than_12_months_lateNot_late,	
	Count_more_than_12_months_late3_to_4_times	
residence_area_typeUrban		

Table 4: Important features of Dataset 3.

The results obtained from the respective models for all possible combinations different methods are as follows.

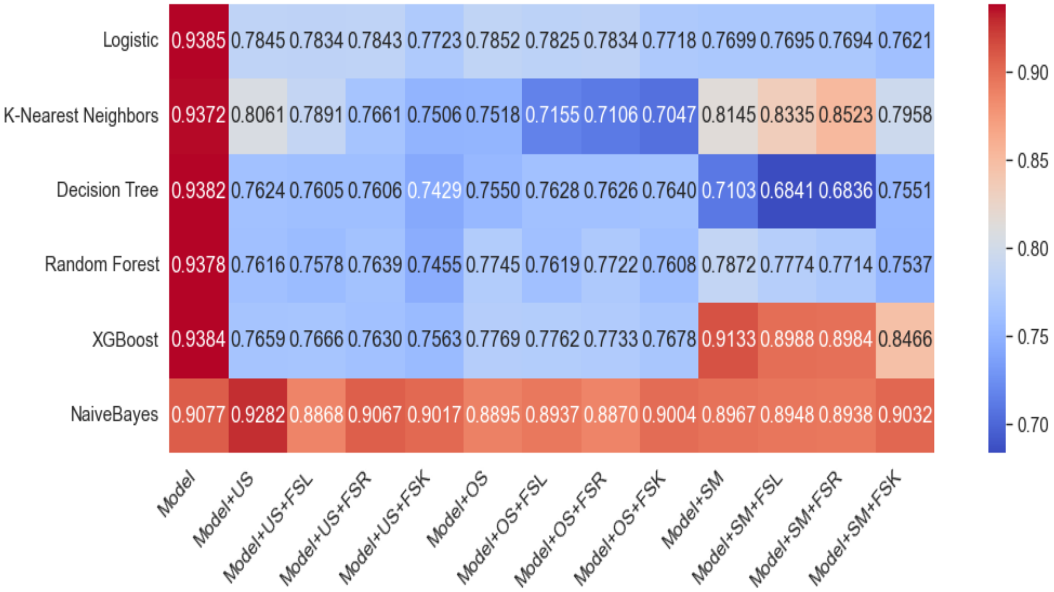


Figure 22: Accuracy Scores for Dataset 3.

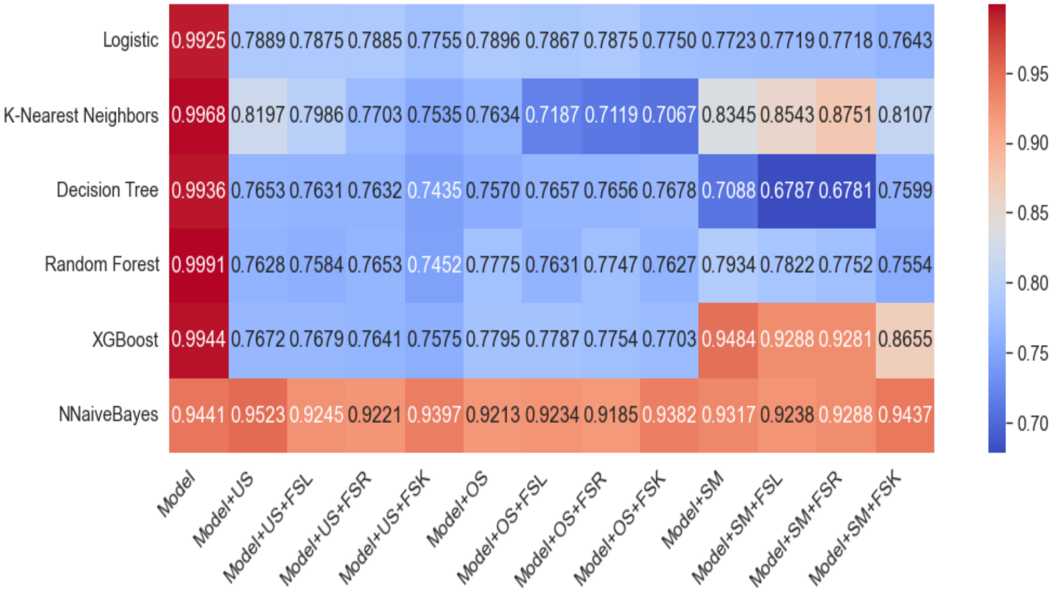


Figure 23: Sensitivity Scores for Dataset 3.

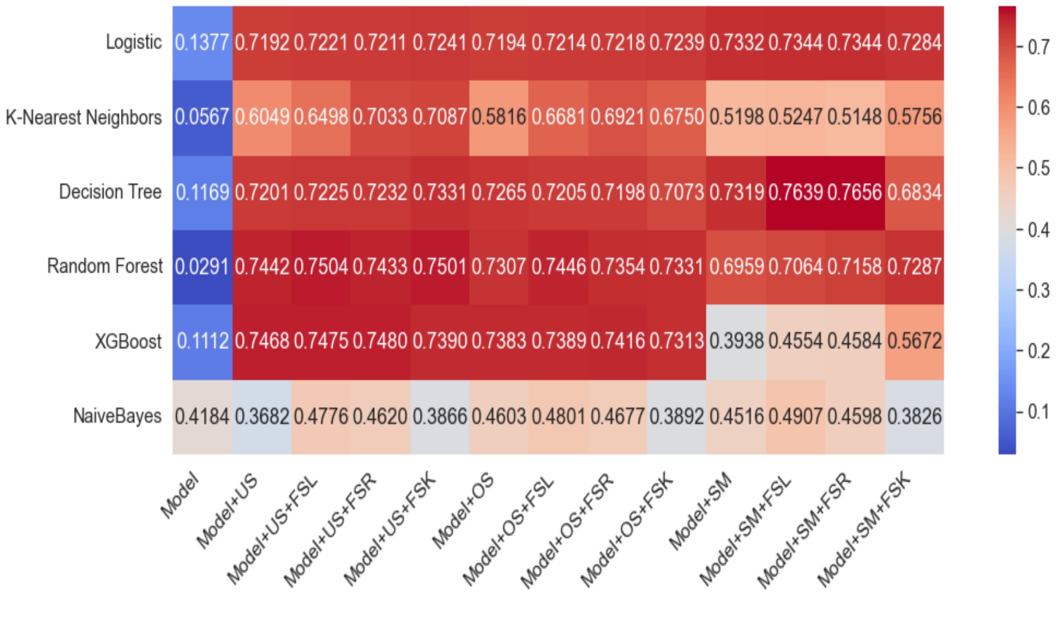


Figure 24: Specificity Scores for Dataset 3.

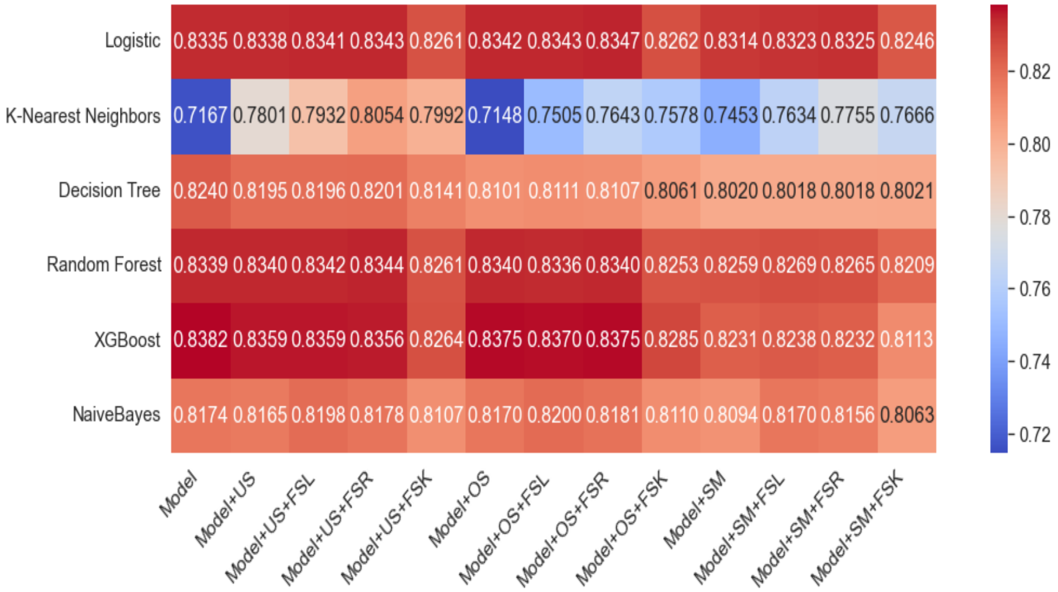


Figure 25: ROC AUC Scores for Dataset 3.

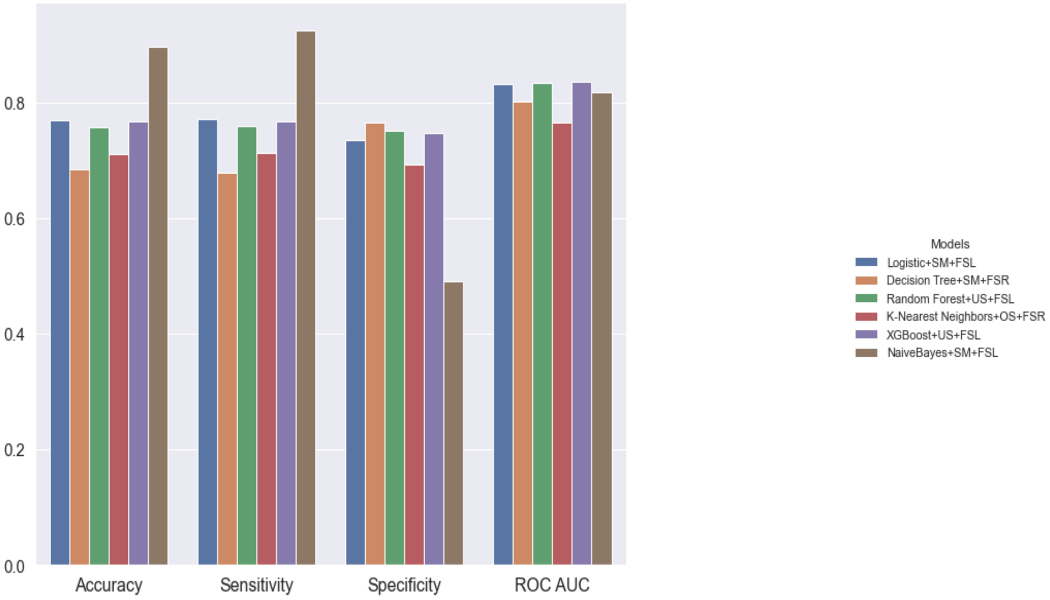


Figure 26: Comparison of best performing combinations from each model for Dataset 3.

Once more, a significant number of models showcased strong performance when paired with either under-sampling (US) or over-sampling (OS), coupled with diverse feature selection methods. In this case, Random Forest +US+FSL, XGBoost+US+FSL and Logistic +SM+FSL emerged as the top performer. Surprisingly, the performance of NB (Naive Bayes) did not meet expectations across various combinations. However, the Random Forest Algorithm notably demonstrated a remarkable ability to identify crucial features, underscoring its effectiveness in bolstering overall model performance.

CHAPTER 5

Conclusion

In this thesis, we embarked on a comprehensive exploration of predictive modeling techniques, leveraging a diverse array of machine learning algorithms and methodologies to analyze three distinct datasets. Through meticulous adherence to a structured methodology, encompassing the application of Lasso, Random Forest, and Knockoffs algorithms, along with strategic dataset partitioning and resampling techniques, we endeavored to unravel valuable insights into predictive modeling across varied scenarios.

The initial phase of our study involved preprocessing the datasets and addressing the challenge of class imbalance through resampling techniques such as oversampling, undersampling, and SMOTE. Subsequently, a suite of machine learning models spanning K-Nearest Neighbor, Logistic Regression, Decision Tree, Random Forest, Extreme Gradient Boosting, and Naive Bayes were trained and evaluated on designated training and test datasets. This rigorous evaluation was characterized by a comprehensive analysis of key performance metrics, including accuracy, sensitivity, specificity, and AUC-ROC score, enabling a nuanced understanding of model efficacy under different conditions.

The culmination of our efforts yielded a rich tapestry of findings across the three datasets. In Dataset 1, while the Knockoffs Algorithm exhibited limited success, likely due to the presence of categorical variables with diverse types, Random Forest and logistic Lasso regularization proved instrumental in identifying important features. Notably, the Decision Tree+OS, Random Forest+US+FSR and XGboost+US+FSR combinations emerged as the top performer, underscoring the importance of feature selection and resampling strategies in enhancing model performance.

Transitioning to Dataset 2, the incorporation of Knockoffs' Algorithm, alongside

Random Forest and logistic Lasso Regularization, facilitated the identification of crucial features. Here, the Logistic+OS+FSK, Random Forest +US + FSK and XG-Boost+ US combinations demonstrated superior performance, underscoring the efficacy of advanced feature selection methodologies in optimizing predictive modeling outcomes.

Dataset 3, subjected to rigorous analysis employing the Knockoffs' Algorithm, showcased promising results, with Random Forest +US+FSL, XGBoost+US+FSL and Logistic +SM+FSL emerging as the optimal combination. Notwithstanding, the exploration of feature selection methods underscored the pivotal role of innovative techniques such as Knockoffs in bolstering model performance, even in scenarios where it may not emerge as the top performer.

Reflecting on our findings, it is evident that the choice of feature selection method and resampling strategy significantly impacts model efficacy. The successful integration of Knockoffs' Algorithm, particularly in datasets with diverse data types, highlights its potential for advancing predictive modeling capabilities. However, challenges persist, as evidenced by the limitations encountered in applying Knockoffs to Dataset 1, necessitating further refinement and adaptation to accommodate categorical variables with multiple types.

Looking ahead, the identified areas for improvement, including enhancing the applicability of Knockoffs Algorithm to datasets with heterogeneous data types, pave the way for future research endeavors. Moreover, our study underscores the importance of methodological rigor and innovation in navigating the complexities of predictive modeling, offering valuable insights for practitioners and researchers alike.

In essence, this thesis represents a significant contribution to the field of predictive analytics, offering a nuanced understanding of model performance under diverse conditions and charting a course for future advancements in predictive modeling

In light of our findings, there exists a clear avenue for future improvement concerning the application of the Knockoffs feature selection method, particularly in Dataset 1. Unfortunately, we encountered limitations in utilizing this method due to its inability to effectively evaluate knockoffs for predictors with mixed data types. It appears that the presence of numerous categorical variables with diverse category types posed a significant challenge for the algorithm and the associated R package.

To address this issue and enhance the applicability of the Knockoffs Algorithm, modifications are necessary. By refining the algorithm to better handle datasets with mixed data types, particularly those containing a multitude of categorical variables, its effectiveness can be significantly enhanced. Without such improvements, leveraging the Knockoffs Algorithm for large datasets characterized by a plethora of categorical variables will remain a daunting task.

Therefore, future research efforts should focus on refining and adapting the Knockoffs Algorithm to accommodate the complexities inherent in datasets with diverse data types. By overcoming these challenges, researchers can unlock the full potential of the Knockoffs feature selection method, thereby expanding its utility and applicability in predictive modeling tasks across various domains.

BIBLIOGRAPHY

Barber, R. F. and Candès, E. J., “Controlling the false discovery rate via knockoffs,” *The Annals of Statistics*, vol. 43, no. 5, p. 2055–2085, 2015.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., “Classification and regression trees,” *Wadsworth International Group*, vol. 37, 1984.

Briys, E. and Varenne, F., *Insurance: From Underwriting to Derivatives*, 01 2001.

Columbus, L., “Mckinsey’s state of machine learning and ai,” Forbes, Tech. Rep., jul 2017. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2017/07/09/mckinseys-state-of-machine-learning-and-ai-2017

Dhieb, N., Ghazzai, H., Besbes, H., and Massoud, Y., “A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement,” march 2020.

Gay, F. H., Lai, G. C., Patterson, G. A., and Witt, R. C., “Underwriting cycles in property and liability insurance: An empirical analysis of industry and by-line data,” *The Journal of Risk and Insurance*, vol. 65, no. 4, p. 539–61, 1998.

Grömping, U., “Variable importance assessment in regression: linear regression versus random forest,” *The American Statistician*, vol. 63, no. 4, pp. 308–319, 2009.

Hanafy, M. and Ming, R., “Comparing smote family techniques in predicting insurance premium defaulting using machine learning models,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 9, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0120970

Hanafy, M. and Ming, R., “Improving imbalanced data classification in auto insurance by the data level approaches,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0120656

Hanafy, M. and Ming, R., “Machine learning approaches for auto insurance big data,” *Risks*, vol. 9, no. 2, 2021. [Online]. Available: https://www.mdpi.com/2227-9091/9/2/42

Hanafy, M. and Ming, R., “Using machine learning models to compare various resampling methods in predicting insurance fraud,” *Journal of Theoretical and Applied Information Technology*, vol. 99, pp. 2819–2833, 07 2021.

Hanafy, M. and Ming, R., "Classification of the insureds using integrated machine learning algorithms: A comparative study," *Applied Artificial Intelligence*, vol. 36, no. 1, p. 2020489, 2022. [Online]. Available: http://dx.doi.org/10.1080/08839514.2021.2020489

Insurance Information Institute. "Facts + statistics: Auto insurance." dec 2020. [Online]. Available: https://www.iii.org/fact-statistic/facts-statistics-auto-insurance

Jing, L., Zhao, W., Sharma, K., and Feng, R., "Research on probability-based learning application on car insurance data," in *Proceedings of the 2017 4th International Conference on Machinery, Materials and Computer (MACMC 2017).* Atlantis Press, 01 2018, pp. 59–63. [Online]. Available: https://doi.org/10.2991/macmc-17.2018.14

Kate A. Smith, R. J. W. and Brooks, M., "An analysis of customer retention and insurance claim patterns using data mining: a case study," *Journal of the Operational Research Society*, vol. 51, no. 5, pp. 532–541, 2000. [Online]. Available: https://doi.org/10.1057/palgrave.jors.2600941

Kormaksson, M., Kelly, L. J., Zhu, X., Haemmerle, S., Pricop, L., and Ohlssen, D., "Sequential knockoffs for continuous and categorical predictors: With application to a large psoriatic arthritis clinical trial pool," *Statistics in Medicine*, vol. 40, pp. 3313–3328, 2020.

Pesantez-Narvaez, J., Guillen, M., and Manuela, A., "Predicting motor insurance claims using telematics data -xgboost versus logistic regression," *Risks*, vol. 70, no. 7, 2019.

Rawat, S., Rawat, A., Kumar, D., and Sabitha, A. S., "Application of machine learning and data visualization techniques for decision support in the insurance sector," 2021.

Shady, A., Khaled, E., and Mohamed, A., "A proposed model to predict auto insurance claims using machine learning techniques," *Journal of Theoretical and Applied Information Technology*, vol. 98, pp. 3428–3437, 2020.