AUTO INSURANCE FRAUD DETECTION USING MACHINE LEARNING

Wangari Kimani Ruth - 092833

Submitted in partial fulfillment of the requirements for the Master of Science Degree of
Data Science & Analytics at Strathmore University

ILAB
Strathmore University
Nairobi, Kenya

MARCH 2025

DECLARATION

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the Research Project contains no material previously published or written by another person except where due reference is made in the Research Project itself.

Wangari Kimani Ruth……..... [*Name of Candidate*]

……..... [*Signature*]

17/03/2025……..... [*Date*]

This Research Project has been submitted for examination with my approval as the Supervisor.

DR BW MUGANDA..... [*Name of Supervisor*]

……..... [*Signature*]

17/03/2025……..... [*Date*]

School of Finance and Applied Economics

Strathmore University

ABSTRACT

Rising vehicle insurance fraud significantly undermines the profitability of insurers and unfairly increases premiums for honest policyholders. To combat this growing threat, advanced machine learning (ML) techniques offer a promising solution for detecting fraudulent claims with greater accuracy and efficiency. This study develops and evaluates an ML-based fraud detection system using rich claim datasets that capture policyholder details, vehicle specifications, and claim attributes such as accident history and claim values. Four ML algorithms—Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), and XGBoost—were trained and assessed using key performance metrics, including accuracy, precision, recall, and F1-score.

The results indicate that while Random Forest and XGBoost achieved high accuracy, they exhibited lower recall, making them less effective in identifying fraudulent claims. In contrast, KNN and Logistic Regression demonstrated superior recall, essential for minimizing undetected fraud. Further optimization through hyperparameter tuning and ADASYN resampling improved KNN's recall to 0.52 and its AUC score to 0.71, while Logistic Regression maintained a recall of 0.60. Based on its balanced performance and interpretability, Logistic Regression was selected for deployment in a web-based fraud detection system.

The study concludes that implementing ML-driven fraud detection can significantly reduce fraudulent payouts, streamline claims processing, and enhance customer satisfaction. Future research should explore the use of more recent datasets, deep learning techniques, and alternative resampling methods to further refine fraud detection accuracy. Expanding the model to include other types of insurance, such as life and health, could enhance its applicability across the industry.

Table of Contents

Tables

Table of Figures

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

1.1.1 Auto Insurance Fraud in the Kenyan Context

According to AKI (2021), insurance fraud is when a person, business or organization knowingly provides false information to an insurance company in order to receive benefits that they do not deserve. Insurance fraud can take many forms such as falsifying or inflating an insurance claim, staging accidents, theft by submitting claims for existing losses or giving false information on an insurance application which is against the principle of utmost good faith in Insurance. Insurance fraud is a serious crime that carries imprisonment and hefty fines.

Insurance fraud is a big problem in Kenya that affects both the insurance industry and the general public. A survey by the Association of Kenya Insurers (AKI, 2021) estimates that 10% of all claims in Kenya are fraudulent, costing the sector billions of shillings in damages. Kenya has a high rate of insurance fraud for a number of reasons. One of the main causes is lack of awareness among the general public about the consequences of insurance fraud. Many individuals and organizations mistakenly believe that insurance fraud does not hurt anyone except themselves without realizing that it affects the entire industry and leads to increased premiums for everyone. Another factor is the lax regulatory environment that dominates Kenya's insurance industry. The body tasked with regulation of the country's insurance market is the Insurance Regulatory Authority (IRA, 2022). However, it is encumbered with lack of resources as well as personnel that can enable it to effectively handle instances of insurance fraud.

The economic issue of insurance fraud puts insurers' viability and financial stability in jeopardy (AKI, 2021). A class of insurance that is of particular importance is vehicle insurance, which accounts for 35.8% of non-life insurance business and has risen by 38.17% and 3.96% in motor private and commercial over the past five years, respectively. Fraud in this category includes reporting and claiming fictitious damage or loss, exaggerating damages covered by the insured, and making false statements about the truth.

Automobile insurance is one of the most difficult products for insurers to sell, according to Association of Kenya Insurers (AKI, 2021), local market actors in the underwriting sector. This is because of the significant technical loss, which accounts for 68.92% and 60.72% of total losses in motor private and motor commercial, respectively. In other words, of the Kes 100 in

premiums that the insurer receives, Kes 68.92 and Kes 60.72 are used to pay for the claims of the insured, respectively (AKI, 2021). The issue is made worse by the significant costs incurred, which account for 44.16% of each class' costs and are partially attributable to the investigation done to confirm the validity of the claim. As a result, for each 100 Kes earned in net premium, the insurer suffers losses of 13.08 and 4.88, respectively.

General insurers are faced with the challenge of balancing fraud risk and reputation risk because they must carefully evaluate claims to avoid exposure to fraudsters, which results in an increase in the cost of premiums for the good customers, while on the other hand, they must maintain an excellent reputation for customer service since they primarily write auto insurance (Karen et al., 2005). Through the provision of various routes for handling claims based on the amount of risk, statistical modeling offers a means to manage each case according to the information obtained and to give a better service in terms of both speed and safety (Irshad Hussain B, 2023).

1.1.2 Machine Learning Approach in Auto Insurance Fraud

For insurance companies to remain profitable and guarantee fair premiums for legitimate policyholders, it is essential to identify and stop such fraudulent activity. When it comes to spotting intricate fraud patterns and developing fraud tactics, traditional methods of fraud detection sometimes fall short because they are time consuming, have limited efficiency as well as accuracy (Karimi, 2019). Machine learning (ML) approaches have recently become more popular in fraud detection due to their capacity to evaluate enormous amounts of data, reveal hidden patterns, and generate precise predictions (Karimi, 2019). Indicators of fraud, abnormal claims, and suspicious activity that can be challenging to see manually, can all be found using machine learning (ML) models. This study focuses on using machine learning (ML) techniques to build a strong auto insurance fraud detection system that may greatly improve fraud detection skills.

Advanced machine learning algorithms have opened up new possibilities for detecting and stopping fraudulent activity involving cars. It is possible to create reliable fraud detection systems that can spot suspicious patterns, behaviors, and abnormalities in vehicle-related transactions by utilizing the power of machine learning (Ye, 2019). Large amounts of data, including financial transactions, insurance claims, vehicle registration information, and past fraud tendencies, can be trained into machine learning algorithms. Machine learning algorithms can learn to identify patterns that are suggestive of fraudulent behavior by extracting useful

features from this data, (Irshad Hussain B, 2023). In order to detect and stop automobile fraud early on, these models can then be utilized to automatically flag suspect activity for additional examination. Machine learning has numerous advantages for detecting automotive fraud (Bhasin, 2018). First of all, it enables real-time analysis of enormous amounts of data, facilitating the rapid detection of probable fraud situations. Second, as they are exposed to fresh data and developing fraud strategies, machine learning models can adapt and advance over time. This flexibility makes sure that the detection system keeps working even when fraudsters come up with new tactics.

Additionally, systems for detecting fraud based on machine learning can lessen false positives, which occur when a legitimate claim is mistakenly flagged as fraudulent, which will have a smaller negative effect on real customers (Rahamathunnisa, 2021). These algorithms can increase their accuracy in differentiating between legitimate and fraudulent transactions by continuously learning from prior data. This facilitates the investigation process and guarantees that resources are directed more effectively to look into actual fraud instances.

We will examine various machine learning methods and algorithms that can be used to identify auto insurance fraud in this study. We will go over how to preprocess data, how to create features, and how to choose the best algorithms to train fraud detection models. We will also explore the difficulties in detecting auto insurance fraud and how machine learning might assist in resolving these difficulties. We have the chance to greatly improve the effectiveness and efficiency of automobile fraud detection systems by leveraging the power of machine learning. The goal of this research project is to assist in the creation of solid and trustworthy solutions that can shield people, businesses, and the entire auto insurance sector from the negative effects of car fraud.

1.2 Problem Statement

It can be challenging to demonstrate fraud, and it can be tough to have a thorough grasp of the issue. Insurance fraud has historically been discovered through the expensive and ineffective manual inspection of claims (Cedervall, 2022). Additionally, since fraudulent activity is dynamic, it is impossible to recognize it using set features. Therefore, insurers must develop the ability to swiftly spot new and evolving fraud schemes. The growth of large-scale company operations has increased the amount of data that must be processed, which has made manual fraud detection challenging and unfeasible. Insurers' databases currently include enormous volumes of data, and the volume is increasing. Data mining is now acknowledged as a vital

activity to spot fraud when combined with various analytical methodologies, like machine learning techniques (Dwivedi, 2020; Gomes, 2021). Automatic techniques have the ability to discover suspicious cases in a scalable manner, which might greatly reduce economic losses, both to insurers and policyholders.

Traditional methods for detecting fraud, such as rule-based algorithms and manual investigation have their limitations when it comes to accuracy and effectiveness. However, the advent of machine learning and the availability of amounts of insurance data offer an opportunity to leverage analytical techniques for improving fraud detection capabilities. Insurance fraud, particularly when it comes to auto insurance, is a major problem in Kenya and all over the world, costing insurance firms a significant amount of money and increasing prices for genuine policyholders, (Karen M Gill, 2005). In order to detect sophisticated fraudulent activity, the current traditional methods of fraud detection are frequently inaccurate, insufficient and ineffective, (Bhasin, 2018). To solve this issue, machine learning (ML) techniques must be used to create a reliable and effective system for detecting vehicle insurance fraud. This system should be able to analyze massive amounts of data, spot suspicious patterns, and instantly flag possibly fraudulent claims, eventually enhancing the profitability and reputation of insurance companies and ensuring that policyholders pay reasonable premiums.

1.3 Research Objective

The main objective of this research project is to develop a system based on machine learning that can effectively identify instances of motor insurance fraud. The system aims to be strong, dependable and adaptable, in order to detect fraudulent claims at an early stage. This will help minimize losses for insurance companies and safeguard the rights of policyholders. It seeks to make use of machine learning capabilities to improve fraud detection abilities, reduce false positives, and successfully adapt to changing fraud techniques, ultimately helping to improve fraud detection systems in the auto insurance industry.

1.4 Specific Objectives

1. To investigate the most relevant factors that inform fraudulent and suspicious activity in the insurance sector.
2. To evaluate the performance of the machine learning models used in detecting fraudulent auto insurance claims.

3. To design and implement a machine learning-based fraud detection system capable of real-time identification of suspicious activities in insurance.

1.5 Research Question

Can a machine learning-based approach, as opposed to more conventional techniques, efficiently identify and prevent motor insurance fraud, resulting in lower financial losses for insurance firms and ensuring fairness for policyholders when it comes to premium pricing?

1.6 Justification of the Study

Automobile insurance fraud is an issue that costs the insurance industry billions of dollars each year. It involves activities, like staged accidents and false claims which pose challenges for fraud detection methods. This research aims to develop a machine learning driven system that can promptly identify auto insurance fraud. By doing so, it will help reduce losses for insurers and protect honest policyholders. The significance of this research lies in its potential to lower premiums for policyholders, advance our knowledge in machine learning based fraud detection and uphold the integrity of the insurance sector.

This research mainly dwells on the potential advantages of using machine learning to discover fraud patterns in the auto insurance industry. It aims to reduce losses caused by false claims and ensure profitability by detecting and stopping fraudulent activity. This could lead to more affordable premiums for policyholders, who can expect lower insurance premiums. The research will also benefit the academic community by providing a case study of practical application in real-world settings, enabling the next generation of data scientists and machine learning specialists to train.

Regulatory authorities, such as Kenya's Insurance Regulatory Authority (IRA), are interested in the results of this research, as efficient fraud detection systems support moral behavior and compliance with legal standards, ultimately improving the insurance industry's stability and standing. The research also has potential to influence technology and data privacy sectors by addressing the challenges of managing private and sensitive data in the insurance business. The research's findings could significantly impact international insurance fraud detection, potentially improving the efficacy and efficiency of fraud detection systems, enabling insurance firms to better resist fraudulent activity and leading to significant changes in industrial procedures and practices.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

According to Simha (2016), fraud is defined as an intentional deception committed through information concealment and misrepresentation with the goal of advancing one's own interests at the expense of another party's. The International Association of Insurance Supervisors (2011) defines fraud as any deceptive activity or inaction intended to unfairly benefit the person who is pulling off the hoax or another party. The term "insurance fraud" refers to a broad range of actions, such as willful fabrication of information, negligent handling of claims by the insurer, misuse of a trust account or its obligations, and hiding or destroying evidence related to communications, financial transactions, and insurance contracts (Young., 2011).

Insurance fraud, which includes numerous fraudulent actions such staged accidents, exaggerated claims, and misrepresentations on insurance applications, is a widespread problem with significant financial ramifications for the insurance business internationally (Viaene, 2015). Innovative strategies are required since traditional fraud detection techniques have failed to handle the complexity and adaptability of fraudulent schemes. Machine learning (ML) has become a promising technique for transforming fraud detection in the insurance industry in this context (Wang et al., 2018). With its capacity to examine big datasets and identify complex patterns, ML algorithms have the potential to improve the detection and avoidance of fraudulent claims.

This literature review's objective is to give a comprehensive overview of the traditional/manual fraud detection techniques as well as the current research and advancements in the field of ML-based auto insurance fraud detection. This review aims to shed light on the way toward more effective and adaptive fraud detection systems that are adapted to the particular requirements of the automotive insurance industry by reviewing the present body of knowledge, methodology, issues, and breakthroughs. We want to synthesize the state-of-the-art in ML-based fraud detection, identify relevant trends, identify research gaps, and highlight potential for improving fraud detection capabilities through critical examination of the literature.

2.2 Manual Fraud Detection Methods

According to Aisha Abdallah (2016), a fraud detection system is one that looks for unusual activity as it passes through the primary system. Previously, this approach involved manually detecting and identifying these behaviors by looking at a sample of actual fraud data. The

procedure has been laborious and prone to human mistake, misinterpretation, and omission of important details. As a result, fraud detection technologies were developed to automate the procedure and remove human intervention from the system's functional level. To provide better outcomes and conclusions for an efficient fraud detection system, data mining methods have improved significantly over the years. However, in previous iterations, many of these approaches were absent.

The insurance sector has long used manual fraud detection techniques when examining auto insurance claims. According to AKI (2021), skilled investigators and claims adjusters are employed by insurance firms to carefully evaluate and investigate suspicious claims in an effort to uncover fraudulent activity. Even though automated fraud detection systems have advanced, manual techniques are still essential for identifying and reducing fraudulent vehicle insurance claims.

Experts in the industry use a variety of tactics and strategies when doing manual fraud detection. These methods consist of:

1. Verification of Claim Documentation: Verification of data from sources like accident reports, repair bills, medical records, and witness testimony is part of this procedure. By closely scrutinizing these documents, investigators can spot errors or irregularities that might suggest fraud (Gill, 2005).
2. Surveillance and Investigation: Manual fraud detection frequently entails surveillance operations including conducting physical surveillance or reviewing surveillance film. In order to acquire more information, investigators may also visit accident scenes, evaluate damage, and work with law enforcement organizations.
3. Expert Opinion and Collaboration: To provide specialized insights and opinions regarding the validity of the claim, experts from other domains, such as automobile specialists, medical professionals, and forensic accountants, may be engaged.

Efforts have also been made to standardize investigation processes and establish best practices. According to the study by Subelj et al. (2011), a framework for manually detecting fraud in auto insurance claims has been developed. It describes a methodical process that includes data analysis, document verification, interviews, and cooperation with outside organizations. An advantage of standardized procedures is that they guarantee that important steps are not missed by offering an organized foundation for fraud detection. They can increase the handling of claims in terms of consistency and efficiency. However, standardized procedures could still

need a lot of resources, and their efficacy is dependent on the caliber and accessibility of outside data sources as well as collaboration with outside organizations.

Another prominent method in the field of vehicle insurance fraud detection is rule-based expert systems. The AKI (2021) states that these systems examine insurance claims using a predetermined set of rules, which are sometimes stated as "if-then" statements. Examples of rule-based rules that can be used to detect auto insurance fraud:

- If the policyholder has filed multiple claims in a short period of time, flag the claim for review.
- If the policyholder has a history of filing fraudulent claims, flag all of their claims for review.
- If the policyholder has made significant changes to their vehicle coverage shortly before filing a claim, flag the claim for review.
- If the policyholder is filing a claim for damage that is inconsistent with the type of accident that they are reporting, flag the claim for review.

Experts in the sector have considered and designed these criteria based on past fraud tendencies and business indications. The main goal is to compare every claim to this set of rules and, when certain conditions are met, to produce warnings that indicate possible fraudulent conduct. For example, when several claims from various regions are filed in a short period of time, the system initiates a manual review.

Over the years, these rule-based systems have improved thanks to decades of manual fraud data analysis and review. Baumann (2021) highlights that although these systems are intricate, their primary means of processing and assessing data, which helps identify possibly fraudulent insurance claims, comes from human-designed rules. Generally, each claim's legality is evaluated using a large number of rules, roughly 300 in all. These systems do, however, have certain drawbacks. Because of its relative simplicity, adding or changing restrictions when fraud techniques change will need manual involvement. They could also have trouble identifying implicit correlations in the data. Moreover, many of these rule-based systems are built with outdated technology that cannot handle the real-time streams of data that the digital world demands.

In conclusion, a tried-and-true but effective technique for identifying vehicle insurance fraud is the application of rule-based expert systems. These systems examine claims based on predetermined guidelines in an effort to quickly identify questionable behaviour. They may

need constant manual changes and have limitations when it comes to identifying intricate, dynamic fraud schemes in real-time digital environments, despite their speed and transparency.

2.2.1 Limitations of Manual Fraud Detection Methods

Although conventional/manual techniques for identifying vehicle insurance fraud have been effective in the past, a number of drawbacks and criticisms have surfaced:

1. Resource Intensiveness: Manual techniques need a lot of work and financial capital, especially when handling a large number of claims. Fraud detection processes may become less effective as a result of this resource strain.

2. Subjectivity: Human judgment can inject subjectivity and biases into the process of detecting. When evaluating the same claim, various investigators could come to different conclusions.

3. Scalability: In the digital age, when data flows are more vast and faster than ever, traditional approaches find it difficult to scale up to accommodate the growing number of claims.

4. Limited Adaptability: Because traditional approaches rely on established criteria and previous data, they may find it difficult to keep up with new fraud schemes.

5. Time Delays: The manual investigation process may cause delays in the claims settlement procedure that are incompatible with the requirement to provide policyholders with prompt responses.

In summary, conventional, manual techniques for identifying motor insurance fraud have significantly improved the insurance industry's ability to avoid fraud. They are very skilled at managing complicated and unusual issues. However, they are facing more and more difficulties in the face of rising data volumes, the need for rapid responses, and keeping up with the ever-evolving fraud schemes. In order to overcome these obstacles, a growing number of insurance companies are considering shifting to automated techniques such as data analytics and machine learning to improve their fraud detection performance, while keeping certain components of manual inquiry for more complex instances.

2.3 Automation of Fraud Detection Systems

The problem of vehicle insurance fraud has been a topic of interest for researchers in recent years. Effective and aggressive responses are necessary to mitigate the financial and reputational harm caused by fraudulent operations. Organizations are using automation as a potent tool in their toolbox as a response to the complexity and scope of fraudulent schemes that are becoming more numerous. In terms of how companies and industries battle fraudulent activity, the automation of fraud detection systems marks a paradigm change. Various approaches have been proposed to address this issue, including rule-based systems, statistical methods, neural networks, and machine learning-based techniques

Technology, particularly machine learning (ML), artificial intelligence (AI), and statistical modeling, lies at the heart of automated fraud detection systems. These advanced techniques empower organizations to create models that learn from historical data, identify patterns indicative of fraudulent behavior, and adapt to evolving tactics employed by fraudsters (Abdallah & Gaber, 2016). The integration of data from diverse sources, coupled with real-time monitoring and predictive analytics, equips organizations with the ability to stay one step ahead of fraudulent activities (Hassani et al., 2019).

2.3.1 Statistical Fraud Detection Methods

In this section we look at some statistical methods that have been used to detect fraud. We start with regression analysis which is a method that helps us understand the relationship, between factors (known as predictors or features) and a specific outcome variable in this case the likelihood of fraud. (Tse, 2018) logistic regression in auto insurance fraud detection is commonly used. It models the probability of an event happening (such as fraud) based on these predictors. Insurance companies gather amounts of data about their policyholders and claims including details like demographics, vehicle information and claim specifics. By using regression, we can create models that predict the likelihood of a claim being fraudulent based on these variables. The coefficients assigned to each predictor variable indicate how strongly they influence the probability of fraud. Logistic regression produces a probability score ranging from 0 to 1. Scores to 1 suggest a possibility of fraud. To categorize claims as fraudulent or legitimate we establish a predefined threshold.

Cluster analysis is a technique that groups similar data points into clusters or categories based on their similarities. In auto insurance fraud detection, it helps identify groups of claims, with shared characteristics or behaviors (Bilodeau, 2008). Insurance companies utilize cluster

analysis on claims data to identify patterns of behavior. For instance, they group together claims from policyholders that exhibit similarities in terms of the timing of the claim or the nature of the damage. The presence of clusters may suggest fraud rings or organized fraudulent activities. The output of cluster analysis is a set of clusters, where claims within the same cluster are more similar to each other than to claims in other clusters. Insurers can then investigate claims within suspicious clusters more closely.

Time-series analysis involves examining data collected over time to identify patterns, trends, and anomalies. In auto insurance fraud detection, time-series analysis is applied to claims data to uncover temporal irregularities (Huang, 2020). Insurance companies use time-series analysis to monitor the timing and frequency of claims. Sudden spikes or unusual patterns, such as a high volume of claims within a short timeframe, may indicate potential fraud. Detecting such anomalies can be critical in real-time fraud prevention. The output of time-series analysis includes visualizations, statistical summaries, and alerts when significant anomalies are detected. These alerts prompt further investigation into potentially fraudulent activities.

Bayesian networks are probabilistic graphical models that capture dependencies between variables using a directed acyclic graph. In the context of auto insurance fraud detection, Bayesian networks help model complex relationships among various factors related to claims and fraud likelihood (Dai, 2012). Insurance companies use Bayesian networks to assess the likelihood of fraud by considering multiple interconnected variables. These variables may include policyholder characteristics, claim details, vehicle information, and more. Bayesian networks enable insurers to account for the interplay of these factors in fraud detection. Bayesian networks provide a probabilistic assessment of fraud likelihood, taking into account the relationships among variables. This output helps insurers prioritize claims for further investigation based on their calculated fraud probability.

In as much as statistical methods are beneficial in detecting auto insurance fraud, they still have limitations that have to be considered when applying them. Some of the limitations include;

- Data Quality: For statistical fraud detection to be effective, data quality is essential. Incomplete or inaccurate data might impair the efficacy of models and result in inaccurate predictions.
- Interpretability: Although statistical models are generally comprehensible, communicating model results to stakeholders may be difficult when utilizing sophisticated statistical methodologies.

- Imbalanced Datasets: Models may be skewed when there is a large proportion of true claims compared to false ones in an imbalanced dataset. Preventing overfitting to the majority class requires addressing class imbalance.

In conclusion, an essential component in automating the identification of auto insurance fraud is statistical methodologies. Insurance companies can identify questionable claims, find patterns, and determine the possibility of fraud by using these techniques, which range from regression analysis to Bayesian networks. Even if they provide interpretability and transparency, problems with model scalability, class imbalance, and data quality still need to be addressed. Improved accuracy, real-time monitoring, and increased transparency are anticipated in the future of vehicle insurance fraud detection as insurers combine statistical techniques with machine learning and artificial intelligence.

2.3.2 Insurance Fraud Detection Using Machine Learning

Due to the challenges presented by the human element in the traditional approach and the limitations imposed by automated systems methods, insurers have begun to use machine learning techniques to anticipate and identify fraudulent motor insurance claims. (Alrais, 2022). Additionally, auto insurance fraud detection has grown even more automatic as a result of information technology advancements. A combination of data mining, analytical algorithms, and specialist knowledge is utilized to provide useful information from the mined data. Many scholars are primarily looking for useful and effective techniques that can be applied to predict and analyze the content of motor insurance claims using machine learning algorithms in order to distinguish between genuine and fraudulent claims. Machine learning approaches facilitate the enhancement of predicted accuracy, hence allowing loss control units to attain reduced false positive rates and increased coverage (Karimi, 2019).

Alrais (2022) claims that the goal of using machine learning techniques to predict auto insurance fraud is to develop a computerized system that can perform complicated analysis and not only replace human input, but also outperform it. Consequently, the system may learn from experience without requiring more programming thanks to machine learning. By analyzing large, labeled data sets, the system is able to do basic jobs and free up human resources to work on more complex claim analysis.

Markovskaia (2020) explains that adopting machine learning for auto insurance fraud prediction allows the system to effectively identify suspected fraudulent claims, process data rapidly, and infer and display situations where links between different parameters exist but are

imperceptible to the human eye. Machine learning algorithms make it possible to separate undetected fraud tendencies and eliminate human error by identifying exceptions. Algorithms are employed to build the model that predicts vehicle insurance fraud. A comparative evaluation of classification models for detecting insurance fraud was conducted by (Rukhsar et al., 2022). The algorithms included Support Vector Machine (SVM), Random-Forest (RF), Decision-Tree (DT), Adaboost, 15 K-Nearest Neighbor (KNN), Linear Regression (LR), Naive Bayes (NB), and Multi-Layer Perceptron (MLP). Precision, Recall, and Fl-Score are three performance indicators that were used to evaluate the algorithms' efficacy. When compared to other strategies for predicting insurance fraud, Decision Trees yielded the highest accuracy of 79%, according to comparative study results.

Many studies have focused on developing machine learning predictive models for the insurance industry. In Kenya, Njeru, (2022) investigated how machine learning algorithms could be used to detect fraudulent auto insurance claims using attributes extracted from vehicle insurance claim datasets. To ascertain the effectiveness and efficiency of detecting fraudulent auto insurance claims, the study examined several machine learning classification algorithms, such as Extreme Gradient Boosting (XGBoost), Adaptive Boosting (AdaBoost), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest Classifier (RF), Artificial Neural Networks (ANN), Decision Tree (DT), and Logistic Regression (LR). The performance of the proposed technique was evaluated using accuracy, precision, recall, and F1-score measures, and XGBoost outscored other machine learning methods considerably, with an accuracy of 99.25%.

Alrais (2022) also studied how machine learning may be used to identify phony insurance claims. The research developed a model that may identify potentially fraudulent claims, saving insurance companies money, time, and improving their effectiveness in handling false claims. To determine how effectively the generated model works, it is recommended to test it against additional datasets that include current data or against datasets of a similar kind. As an alternative, a predetermined set of parameters or random combinations may be examined. Supervised learning techniques in machine learning were used in the study. The study's dataset included exponentially good performance from KNN and Random Forest in addition to XGBoost, logistic regression, and Random Forest.

Fernando (2021) conducted research on machine learning methods for identifying fraudulent auto insurance claims. The primary objective of the research was to develop a model for

detecting auto insurance fraud by utilizing classification algorithms. The best model was suggested by applying a series of assessments. The study examined data from motor insurance claims made to Sri Lanka Insurance. In the dataset used for the investigation, 3,112 out of 30,098 vehicle claims were found to be fraudulent. The dataset was unbalanced since fraudulent claims, also known as positive occurrences, accounted for only 10% of all cases. Underwriting information was analyzed with past claims data. Three classifiers—Artificial Neural Network, Random Forest, and XGBoost—were employed to determine whether or not a claim was fraudulent. By splitting the dataset into training, validating, and testing sets, these algorithms were examined and assessed. Nevertheless, the machine learning model is biased towards the majority class when input data is given to it with an imbalanced class variable; in one example, the program incorrectly identified a fraudulent claim as a legitimate claim. To solve this issue, ensemble models and the oversampling technique known as Synthetic Minority Oversampling Technique (SMOTE) were used. To evaluate the model's performance, evaluation criteria such as receiver operating characteristics (ROC) curve, precision-recall (PR) curve, f1-score, recall, and precision were employed. Because the Random Forest and XGBoost classifier models have parameters that the researcher had to select, hyperparameter tuning was also employed and evaluated. In comparison to neural network models, Random Forest and XGBoost models were found to perform better. Performance differences between the random forest and XGBoost models were minimal, but the random forest model with modified hyperparameters performed marginally better than the other models. The study's conclusion, which highlights the significance of transforming weak learners into strong learners by ensembling approaches, is that ensemble models, such as the random forest model and the XGboost model, perform better in forecasting motor fraud claims.

The model developed by (Urunkar et al., 2022) was based on a variety of machine learning methods, and the dataset consisted of actual data from a well-known insurance business in Brazil. To develop a model that can assess if an insurance claim is fraudulent, they employed logistic regression, XGB, decision trees, random forests, and K nearest neighbor. They took characteristics from the profiles of known fraudsters that affect false claims. According to their model's results, ensemble techniques including random forest, gradient boosting, and deep neural networks outperformed logistic regression in terms of accuracy.

Because the logistic regression classifier can counterbalance the drawbacks of using other algorithms to construct the predictive model, the researcher used it to construct the model for

predicting fraudulent vehicle insurance claims, even though many of the aforementioned related research studies concluded that the logistic regression algorithm performed poorly than other algorithms in making predictions. Because of its straightforward probabilistic interpretation, the logistic regression technique uses less processing resources and less time for model training than alternative algorithms like Artificial Neural Networks (ANN). In addition to being incredibly quick in classifying unknown records, the logarithm is also incredibly effective in datasets with linearly separable variables. Finally, by using a stochastic gradient descent, the logistic regression algorithm makes it simple to update a prediction model to incorporate new data, in contrast to support vector machines and decision trees.

Irshad Hussain B, (2023), proposed a machine learning-based approach for detecting vehicle insurance fraud. They explored various machine learning techniques and algorithms, including XGBoost, K-NN, and decision tree, and discussed the data pre-processing steps and feature engineering techniques used in developing an effective fraud detection system. They also employed feature engineering techniques to extract relevant information, including policyholder demographics, vehicle details, accident reports, and claim amounts, which were used to identify patterns indicative of fraudulent behavior. Previous research has shown that machine learning-based techniques have promising results in detecting vehicle insurance fraud due to their adaptability and ability to learn from historical data.

The authors of this paper aimed to contribute to the development of robust and reliable solutions that can protect individuals, organizations, and the automotive industry as a whole from the detrimental effects of vehicle fraud. Overall, this study highlights the potential benefits of using machine learning for detecting vehicle insurance fraud and provides insights into the challenges associated with this problem. By harnessing the power of machine learning, researchers and practitioners can significantly enhance the effectiveness and efficiency of fraud detection systems in the vehicle insurance domain.

2.4 Research Gap

There are a number of significant gaps in the current body of research on machine learning-based auto insurance fraud detection. Although there has been significant success in creating predictive supervised machine learning models for fraud detection, there aren't many studies that make use of recent data, which could limit the models' capacity to adjust to changing fraud strategies. Furthermore, there are issues with data privacy and sensitivity that Kenyan insurance

businesses have. These issues restrict data access and make it challenging to create and evaluate machine learning algorithms that are specific to the peculiarities of the Kenyan auto insurance market. Transparency is another area where prior research fails, with some studies neglecting to reveal the elements that they used. Furthermore, the data sources needed to properly train machine learning models are not given enough attention, despite some research into privacy-preserving techniques for fraud detection. The insurance industry has focused mostly on centralized datasets, ignoring strategies that could help the insurance sector as a whole. In addition, the problem of imbalanced claim datasets has not gotten much attention, thus more research is necessary to create decentralized training models and evaluate cooperative model training methods. In conclusion, these research gaps highlight the need for studies that focus on the Kenyan auto insurance market, make use of up-to-date data, improve transparency, and address class imbalance. There is also a need to investigate decentralized and privacy-preserving training approaches in order to support the efficacy and quality of fraud detection systems in this field.

CHAPTER 3: METHODOLOGY

3.1 Introduction

The discovery of the dataset, data preparation, data exploration, testing and modeling, and results are some of the processes that this research will take. Car insurance data was acquired as part of the research process to better comprehend the data structure and extract the characteristics needed to train the machine learning classifiers. During the project's model training and testing phases, the chosen dataset will be generated. The project's first phase will include the preparatory phases of data cleaning and normalization, which deal with redundant and missing data and make sure the dataset's existing data conforms with integrity standards. Following a thorough analysis of the data, a variety of tools will be used to display the data in order to provide a variety of visuals that will aid in illustrating the relationships between the various variables. The models will then be trained and assessed. The most precise and effective machine learning classifier for identifying genuine vs fraudulent vehicle insurance claims will be found using the CRISP-DM technique in order to achieve the study's objectives.

3.2 CRISP-DM Methodology

The Cross Industry Standard Process for Data Mining (CRISP-DM) paradigm will be utilized in the study. The process model was chosen for this investigation due to its substantial backtracking, flexibility, and widespread application in data mining and analysis. A tried-and-true method for leading a data science or machine learning project is called CRISP-DM (Hotz., 2023). Furthermore, CRISP-DM offers a standard structure for organizing and overseeing a project involving machine learning. CRISP-DM is a 1996 invention that plans, coordinates, and carries out machine learning (data mining) initiatives, claims Rodrigues (2020). In light of this, the CRISP-DM serves as the process model for the purposes of this study. It provides an overview of the project's machine learning life cycle and describes the phases of the supervised machine learning project, the tasks associated with each phase, and the relationships between the tasks. The CRISP-DM Model breaks down the life cycle of a data mining project into six phases namely:

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modeling
5. Evaluation

6. Deployment

3.2.1 Business Understanding

The first stage of the cycle is centered on comprehending the project's goals from a business standpoint, applying that knowledge to define a data science problem, and creating a draft plan that aims to address the issue at hand as well as identify significant variables that may have an impact on the project's outcome (Hotz., 2023). We will focus on the study's objectives, which have to do with detecting fraud related to auto insurance claims. A system that can recognize and anticipate fraudulent automobile insurance claims in real time is required, as we found out that insurance firms are experiencing an increase in financial losses as a result of these claims.

3.2.2 Data Understanding

The second phase of the CRISP-DM process is data understanding. In this phase, researchers need to understand the data that they will be using to develop the machine learning model (Hotz., 2023). This includes performing exploratory data analysis (EDA) which helps in understanding the different features in the data, data structure, the relationships between the features, and the quality of the data. In this phase, statistical summaries and visualizations are used to identify trends and patterns in the data. To fully understand the problem of fraudulent auto insurance claims, I will use a secondary source of data at this stage of the CRISP-DM life cycle. After contacting a few Kenyan insurance providers, I was unable to get a dataset on Kenyan automobile insurance claims since the data is too sensitive and confidential.

The dataset that will be used in this research project was obtained from the Oracle Database and includes past insurance motor claims history which was originally collected by Angoss Knowledge Seeker software and published by Oracle between 1994 - 1996. The data contains 15,420 records of policy claims and 33 features, of which 14,497 is non-fraudulent records while 923 consists of fraudulent cases. The dataset contains historical auto insurance claims, policyholder information, accident details, claim amounts, car make and model, year of manufacture, sum insured, and other relevant contextual data from insurance firms. This data will be trained using different machine learning algorithms and later used to validate the best performing model.

The dataset contains the following features.

Feature Name	Description	Type
Month	Month of accident	Object
Week Of Month	The week in the month of the accident occurrence	Object
Day Of Week	Day of the accident	Object
Make	The car model	Object
Accident Area	General area of the accident	Object
Day Of Week Claimed	Day of the week the claim was filed	Object
Month Claimed	The month the claim was filed	Object
Week Of Month Claimed	Week of the month the claimed was field	int64
Sex	Gender of the person claiming	Object

Marital Status	Marital status of the person claiming	Object
Age	Age of the person making a claim	int64
Police Report Filed	Indicates whether a police report was filed for the accident	Object

Table 1 Dataset Features

Fault	Fault owner	Object
Base Policy	Type of insurance coverage	Object
Number Of Cars Year	Number of cars involved in the accident	Object
Address Change_Claim	Time from claim was filed to when the person moved.	Object
Number Of Suppliments	Additional coverage to the primary scope	Object
Witness Present	Presence of a witness	Object
Agent Type	Agent classification	Object
Age Of Policy Holder	Policy Holder age	Object

Vehicle Category	Categorization of vehicle	Object
Policy Type	1. Type of insurance 2. Category of vehicle	Object
Vehicle Price	Vehicle prices	Object
Deductible	The deductible amount	int64
Age Of Vehicle	Age of vehicle at the time of accident	int64
Past Number Of Claims	Previous number of claims	Object
Days_Policy_Claim	Number of days the purchased and claimed was filed	Object
Rep Number	Rep number	int64
Days_Policy_Accident	The number of days between the accident occured	Object
Driver Rating	Driver rating	Object
Fraud Found_P	Indicates whether a claim is fraudulent	int64
Policy Number	The policy number	int64

3.3.3 Data Preparation

To produce high-quality features that would be trained using the machine learning classifiers, data pre-processing is necessary to ensure the dataset is in a format that can be analyzed efficiently. An exploratory approach to data analysis will be used in this work to analyze and choose quality features. Data pre-processing is an essential step in order for machine learning to yield precise and insightful findings (Harrington, 2012). Data quality has an adverse relationship with the outcomes' dependability. Real-world datasets are by their very nature noisy, inconsistent, and flawed. Data pre-processing fills in the gaps in the data, lowers noise, and corrects inconsistencies to improve the quality of the data (Harrington, 2012). In order to remove redundant or unnecessary data and leave just the parts that offer important information to help in creating an effective and efficient categorization, Pandey (2019) defines data preparation as cleaning, integrating, transforming, and reducing data. Pandey (2019) outlines the steps involved in the process are as follows:

i. Data cleaning, the process of filling in missing values and removing anomalies from the dataset.

ii. Using data transformation methods, such as normalization and encoding. Normalization could, for example, improve the accuracy and efficiency of distance-based mining algorithms. Encoding is a process which involves converting categorical variables into integers or numerical format, which can be easily understood by the machine learning model (Harrington., 2012).

iii. Data integration, which creates a single data warehouse from data collected from several sources.

iv. Data reduction, which reduces the amount of data by eliminating features that are redundant. Methods for selecting and extracting features can be applied.

3.3.4 Modeling

Four models which include; XGBoost, Random Forest, KNN, and Logistic Regression will be trained and evaluated for this study in order to determine which technique worked best with the dataset. Based on earlier research that yielded encouraging results when accuracy tests were carried out, the four models were chosen. Twenty percent of the dataset will be used to test the classifier's prediction, while the remaining eighty percent will be used to train the classifiers. The accuracy of the outcomes for each classifier will be acquired in order to determine which classifier will perform the best.

Extreme Gradient Boosting, or XGBoost, is a powerful ensemble learning technique that combines the predictions of many decision trees in a sequential manner to perform very well in predictive modeling (Naseer, 2020). XGBoost is an iterative boosting strategy to generate models, wherein each new tree fixes mistakes in the combined model of its predecessors. A regularization term that discourages excessively complex trees and a loss function that measures prediction discrepancies make up the objective function that the method reduces during training.

$$\text{Objective} = \sum_{i=1}^{n} L(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_j)$$

Where; n = number of training samples, $\sum_{i=1}^{n} L(y_i, \hat{y_i})$ is the loss function which gets the difference between the actual label and the expected label, in our case the label being the fraud found column. $\Omega(f_j)$ is the regularization term used to penalize each of the individual trees and K represents the number of trees. By gradually adding new trees that fix the mistakes of the older ones, XGBoost employs gradient boosting to reduce the aforementioned loss function during training (Naseer, 2020).

By using gradient boosting, XGBoost is better able to handle challenging cases in the training set. XGBoost guarantees effectiveness, prevents overfitting, and enhances model performance via procedures like tree trimming and parallel training. Thanks to its scalability and versatility, XGBoost has gained popularity as an alternative to standard approaches for a wide range of machine learning applications (Chen, 2016). With its integrated feature significance score, XGBoost highlights the attributes that have the greatest influence on the model's predictions. This will assist in determining the most important variables linked to suspicious and questionable behavior. High accuracy in fraud detection tasks is a well-known attribute of XGBoost (Chen, 2016). The efficacy of the model in detecting true positives and reducing false positives may be evaluated by gauging its capacity to accurately categorize legitimate and fraudulent claims.

An ensemble learning technique called Random Forest builds a large number of decision trees during training and outputs a class that is the mean prediction (regression) or the mode of the classes (classification) of the individual trees (Liaw, 2002). There are two methods to add the randomness: one is to use a random subset of the training data for each tree, and the other is to randomly pick a subset of features for each tree. By adding up the forecasts of each individual tree, the Random Forest prediction is produced.

The Random Forest equation is written as: $F_{(x)} = \frac{1}{N}\sum_{i=1}^{N} h_i(x)$

Where *N* is the number of trees in the ensemble and $h_i(x)\ is\ the\ prediction\ from\ the\ i-th\ tree$. It is possible to determine which characteristics have the most influence on the Random Forest model's decision-making by examining the feature importance scores it produces after training (Liaw, 2002). Elevated significance implies a more robust impact on fraudulent activities.

For classification and regression, K-Nearest Neighbors is an easy-to-understand approach. The class membership that results from categorization is the majority class among the k closest neighbors. Regression analysis yields an output that is the mean of the values of the k closest neighbors (Weinberger, 2006). The training process does not directly teach the algorithm a model. The anticipated class or value, y, for a given query point x, may be expressed as follows:

$$y = \frac{1}{k}\sum_{i=1}^{k} y_i$$

Where, y is the value of the i-th closest neighbour. Based on the degree of similarity between features, KNN detects false patterns. According to Weinberger (2006),
patterns in the feature space that are suggestive of fraud may be found by analyzing the characteristics of the closest neighbors.

Machine learning algorithms such as logistic regression are often employed for tasks involving binary categorization. It uses a logistic function to simulate the likelihood of an instance falling into a certain class, as opposed to linear regression (Hosmer, 2004). A range between 0 and 1, which represents the likelihood of the positive class, is mapped by this function to the linear combination of the input characteristics. The logistic regression is shown through the following equation: $P(Y = 1) = \frac{1}{1+\epsilon^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_n X_n)}}$

Where; P (Y = 1) is the probability of the positive class label, $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients that will be trained in the model and will show how each factor affects the probability of fraud. $X_0, X_1, \ldots, X_n$ are the input features. Using the training data, the likelihood function is maximized to estimate the coefficients. Positive coefficients indicate a positive association while negative coefficients indicate a negative correlation. Through the process of training, Logistic Regression calculates the coefficients and produces a comprehensible model that provides information about how each variable affects the expected result, in our research the expected result being the presence of fraud (Hosmer, 2004). It is especially useful in situations when

figuring out the probability and comprehending the influence of each individual characteristic are crucial. Logistic Regression is a simple but powerful technique for binary classification issues, and despite its simplicity, it may be highly useful in many different applications (King, 2001).

The target feature in our dataset is imbalanced, so I will opt to apply class weights to that feature. According to George (2022), class imbalance occurs when a dataset has noticeably more samples of one class (the majority class) than the other (the minority class). In fields such as fraud detection or illness diagnosis, this might lead to models that disproportionately favor the majority class, resulting in poor performance on the minority class. During model training, it's important to give the minority class a higher weight than the other classes. This makes sure the model is more focused on accurately categorizing examples from the minority class, which enhances the model's generalization to that class (George, 2022). This will be done by using built-in mechanisms in particular, the class_weight='balanced' in scikit-learn, which assigns weights inversely proportional to class frequencies.

3.3.5 Evaluation

In order to effectively detect auto insurance fraud, this research will assess the effectiveness of four machine learning models. To evaluate accuracy, reduce missed fraudulent claims, and solve class imbalance, a variety of measures will be used, such as the confusion matrix, F1-score, and a classification report that emphasizes recall and precision. Following a review of the models using these standards, the research will determine which model best balances accuracy, F1-score, recall, interpretability, and efficiency to provide accurate and trustworthy fraud detection results. The study aims to achieve the optimal balance between reducing false positives and false negatives in this particular context by giving priority to recall and closely examining precision. In the end, it will be determined which model is best at spotting false vehicle insurance claims.

3.3.5.1 Confusion Matrix

Parab (2020) defines a confusion matrix as a performance classification metric that evaluates the effectiveness of a machine learning algorithm based on target classes. Using a confusion matrix, the following values will be calculated to generate the categorization metrics mentioned above:

- True Positives (TP) - The number of fraudulent auto insurance claims that will be found.
- True Negatives (TN) - The amount of legitimate auto insurance claims that will be reported as fraudulent.
- False Positives (FP) - The quantity of legitimate auto insurance claims that will be mistakenly labeled as fraudulent.
- False Negatives (FN) - The quantity of false auto insurance claims that will remain unreported.

3.3.5.2 Accuracy

Accuracy, according to Parab (2020), is the ratio of all the input data; the total of True Positives, False Positives, False Negatives, and True Negatives, to the accurately predicted observations (True Positives). Accuracy in auto insurance fraud detection will be useful in evaluating the model's overall correctness in distinguishing between fraudulent and legitimate claims.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

3.3.5.3 Precision

According to Parab (2020), precision can be defined as the ratio of correctly predicted positive samples, also referred to as True Positives, to the total number of expected positive samples, which is the sum of True Positives and False Positives. High recall is often preferred for auto insurance fraud detection since it might lead to more serious implications when a fraudulent claim is missed (false negative) than when a valid claim is investigated (false positive) (Brownlee, 2020).

$$Precision = \frac{TP}{(TP + FP)}$$

3.3.5.4 Recall

This represents the ratio of all samples in the actual class (the sum of True Positives and False Negatives) to the accurately predicted positive samples (True Positives) (Parab, 2020).

$$Recall = \frac{TP}{(TP + FN)}$$

A high recall rate means that a significant percentage of real fraudulent instances are successfully captured by the model. A low recall value indicates that a large percentage of fraud

instances are being missed by the model, which may be enabling fraudulent activity to continue unnoticed (Brownlee, 2020).

3.3.5.5 F-1 Score

The weighted average of recall and precision is known as the F1 Score. F1 Score is particularly very useful when there is an imbalance between fraudulent and non-fraudulent claims as is the case with our dataset. An improved balance between recall (reducing false negatives) and accuracy (minimizing false positives) is indicated by a higher F1 score (Parab, 2020).

$$F1\ score\ = 2 * \frac{(Precision\ *\ Recall)}{(Precision + Recall)}$$

3.3.6 Deployment

After training the model, I will validate the chosen model on the 20% dataset to assess its generalization capability on unseen data. Once the model is evaluated and found to be effective with the highest levels of prediction performance and classification accuracy, it can be deployed in a production environment to make real time detection of fraudulent auto insurance claims. This can be achieved by incorporating the model into a web design system or production environment where it will be used. This will involve linking the model into an application programming interface (API), web service, or other infrastructure. This will thereafter be used to flag suspicious claims for further investigation by human analysts.

3.4 Feature Engineering & Feature Selection

3.4.1 Feature Engineering

Feature engineering is the process of transforming raw data into features that are more informative and predictive for a machine learning model (Johnson, 2019). It involves creating new features, combining existing features, and transforming features into a format that is more compatible with the chosen machine learning algorithm. Feature engineering is an important step in any machine learning project, but it is especially important for projects involving complex and noisy data, such as auto insurance fraud detection (Gareth et al., 2021).

Claims datasets frequently include a variety of data kinds and formats. Furthermore, this data was gathered and combined throughout various time periods. Additionally, this information

might come from outside sources like demographics. An aggregation of the features seen in a dataset over a predetermined period of time can be one feature used in the prediction. The goal of feature engineering is to enhance the performance of the machine learning model by preparing datasets to meet the demands of a machine learning algorithm. Studies have demonstrated that for algorithms to function properly, a domain's features must depict particular qualities (Zhang et al., 2019). Various feature engineering techniques have been investigated and suggested for the purpose of detecting insurance fraud. Zhang et al. (2019) employed the HOBA technique, which enables claims to be homogenized while taking various heterogeneity factors into account.

In the context of auto insurance fraud detection, feature engineering can be used to create new features that represent the following:

- The policyholder's risk profile (e.g., age, gender, driving history, claims history).
- The vehicle involved (e.g., make, model, year, value).
- The circumstances of the claim (e.g., type of accident, date, time, location).
- The relationships between the different features.

Feature engineering can also be used to transform features into a format that is more compatible with the chosen machine learning algorithm. For example, some machine learning algorithms cannot handle categorical features. In these cases, the categorical features would need to be converted to numerical features using a technique such as one-hot or label encoding. Using feature engineering will lead to more informative and predictive features for the chosen machine learning models. This can lead to improved performance on the fraud detection task.

3.4.2 Feature Selection

Feature selection is the process of selecting a subset of features from a dataset that are most relevant to the target variable (Johnson, 2019). It is a critical step in machine learning, as it can help to improve the performance of the model and reduce the risk of overfitting.

According to Gareth et al. (2021), there are several reasons why feature selection is important:

- To improve model performance: Feature selection can help to improve the performance of a machine learning model by removing irrelevant and redundant features. Irrelevant features are those that do not contain any information about the target variable. Redundant features are those that contain the same information as other features. By

removing irrelevant and redundant features, feature selection can help to simplify the model and make it more efficient.

- To reduce the risk of overfitting: Overfitting occurs when a machine learning model learns the training data too well and is unable to generalize to new data. Feature selection can help to reduce the risk of overfitting by removing features that are specific to the training data and are not likely to be present in new data.
- To improve model interpretability: Feature selection can also help to improve the interpretability of a machine learning model. By selecting a smaller subset of features, it is easier to understand how the model is making predictions.

For this research project, we will use a variety of feature selection methods including:

- Univariate selection: Univariate selection methods, such as Pearson's correlation coefficient, which ranks features based on their individual correlation with the target variable (Gareth et al., 2021). Features with the highest correlation will be selected for the modeling.
- Multivariate selection: Multivariate selection methods take into account the relationships between features when ranking them. This allows them to identify features that are most informative when considered together (Morgan, 2022).

Once a feature selection method has been selected, it can be applied to the auto insurance claims dataset. The resulting subset of features can then be used to train a machine learning model for fraud detection. By using feature selection, we will be able to create a more robust and accurate machine learning model for auto insurance fraud detection.

3.5 Ethical Considerations and Fairness

I will also ensure that potential biases in the data and model predictions is addressed to ensure fairness and prevent discrimination. I will implement mechanisms to handle sensitive information and adhere to data privacy regulations. The data is highly anonymized, making it impossible to profile the individuals whose data we are using in this research project.

CHAPTER 4: RESULTS AND DISCUSSIONS

4.1 Introduction

The primary objective of this chapter is to showcase the research outcomes and explore the potential of machine learning algorithms in utilizing features extracted from vehicle insurance claim datasets to detect fraudulent claims. The effectiveness of fraudulent detection was assessed by comparing the performance of four classification models: XGBoost, Random Forest (RF), KNN, and Logistic Regression (LR).

4.2 Data Preparation

For the data preparation, we started by checking for null values and duplicate values, of which neither of them were present as seen below. Since missing data can distort the classification of a machine learning model, it is crucial to scrutinize and rectify them utilizing suitable imputation techniques.

```
#Check if there are any null values in the dataset
df.isnull().sum()
```

```
Month                   0
WeekOfMonth             0
DayOfWeek               0
Make                    0
AccidentArea            0
DayOfWeekClaimed        0
MonthClaimed            0
WeekOfMonthClaimed      0
Sex                     0
MaritalStatus           0
Age                     0
Fault                   0
PolicyType              0
VehicleCategory         0
VehiclePrice            0
FraudFound_P            0
PolicyNumber            0
RepNumber               0
Deductible              0
DriverRating            0
Days_Policy_Accident    0
Days_Policy_Claim       0
PastNumberOfClaims      0
AgeOfVehicle            0
AgeOfPolicyHolder       0
PoliceReportFiled       0
WitnessPresent          0
AgentType               0
NumberOfSuppliments     0
AddressChange_Claim     0
NumberOfCars            0
Year                    0
BasePolicy              0
```

Figure 1 Checking for missing values

```
df.drop_duplicates()
# Number of rows remain the same at 15420, no duplicates present and dropped
```

	Month	WeekOfMonth	DayOfWeek	Make	AccidentArea	DayOfWeekClaimed	MonthClaimed	WeekOfMonthClaimed	Sex	MaritalStatus	...	AgeOfVehicle	AgeOfPolicyHolder	PoliceRep
0	Dec	5	Wednesday	Honda	Urban	Tuesday	Jan	1	Female	Single	...	3 years	26 to 30	
1	Jan	3	Wednesday	Honda	Urban	Monday	Jan	4	Male	Single	...	6 years	31 to 35	
2	Oct	5	Friday	Honda	Urban	Thursday	Nov	2	Male	Married	...	7 years	41 to 50	
3	Jun	2	Saturday	Toyota	Rural	Friday	Jul	1	Male	Married	...	more than 7	51 to 65	
4	Jan	5	Monday	Honda	Urban	Tuesday	Feb	2	Female	Single	...	5 years	31 to 35	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
15415	Nov	4	Friday	Toyota	Urban	Tuesday	Nov	5	Male	Married	...	6 years	31 to 35	
15416	Nov	5	Thursday	Pontiac	Urban	Friday	Dec	1	Male	Married	...	6 years	31 to 35	
15417	Nov	5	Thursday	Toyota	Rural	Friday	Dec	1	Male	Single	...	5 years	26 to 30	
15418	Dec	1	Monday	Toyota	Urban	Thursday	Dec	2	Female	Married	...	2 years	31 to 35	
15419	Dec	2	Wednesday	Toyota	Urban	Thursday	Dec	3	Male	Single	...	5 years	26 to 30	

15420 rows × 33 columns

Figure 2 Checking for duplicates

From the above, we can see that the shape of the data remained the same as the original shape, indicating no presence of missing values and duplicate values.

4.3 Exploratory Data Analysis

Harrington (2012) states that Exploratory Data Analysis (EDA) is a procedure that yields a concise comprehension of the gathered data. EDA was used in the study to both explore the data and highlight the most important findings. We utilized EDA to look for trends, patterns, and outliers in the data that was gathered. We conducted descriptive statistics and presented the data in plot charts for better comprehension. The EDA completed is shown below;

0 (Fraud_Not_Found)	1 (Fraud_Found)
14497	**923**

Table 2 Count of Fraud Found

```
# Plotting the fraud counts
plt.figure(figsize=(6,5))
df["FraudFound_P"].value_counts().plot(kind="bar", color=['salmon', 'lightblue'])
```

```
<Axes: xlabel='FraudFound_P'>
```

Figure 3 Fraud Count Plot

The above plot shows us that the Fraud found cases are only 923 records, which is 6% of the whole data. This shows the data is highly imbalanced, of which will be dealt with later on during the analysis before the modeling.

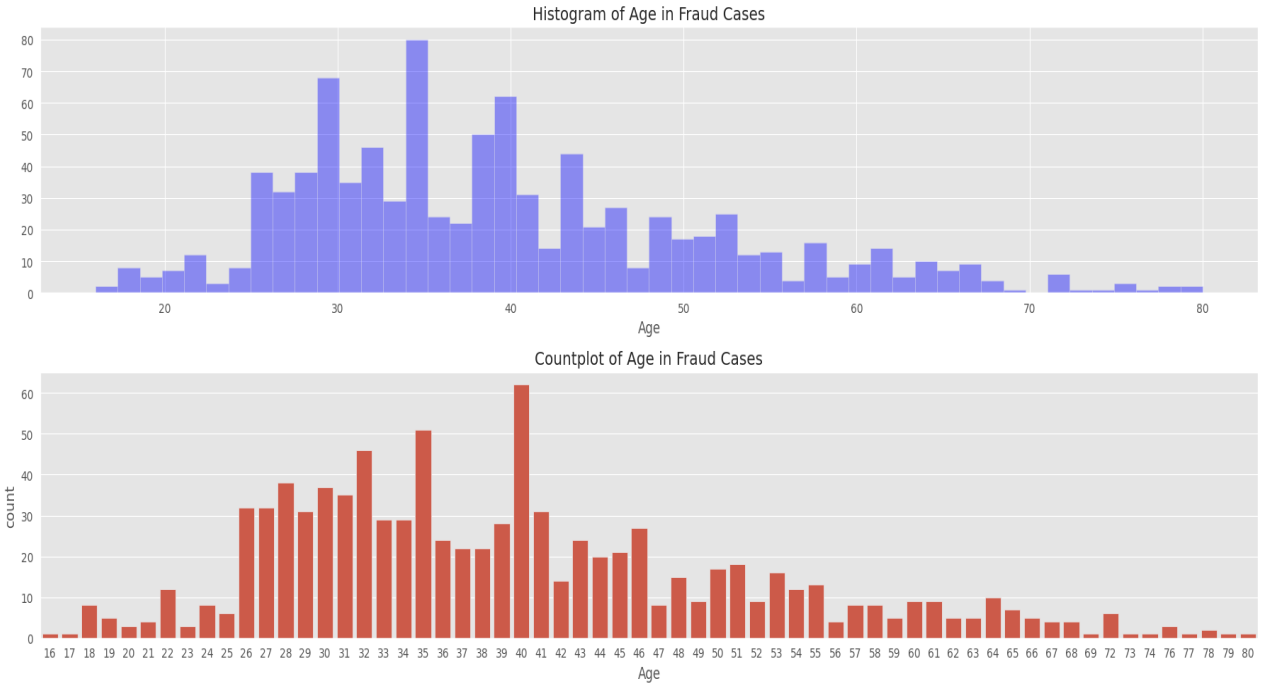


Figure 4 Distribution of age involved in fraud

The graph above shows us the age distribution of the fraud cases. The distribution of Age in Fraud dataset is also slightly right-skewed. In addition, most of the ages fall between 26 to 46 years old for fraud cases.

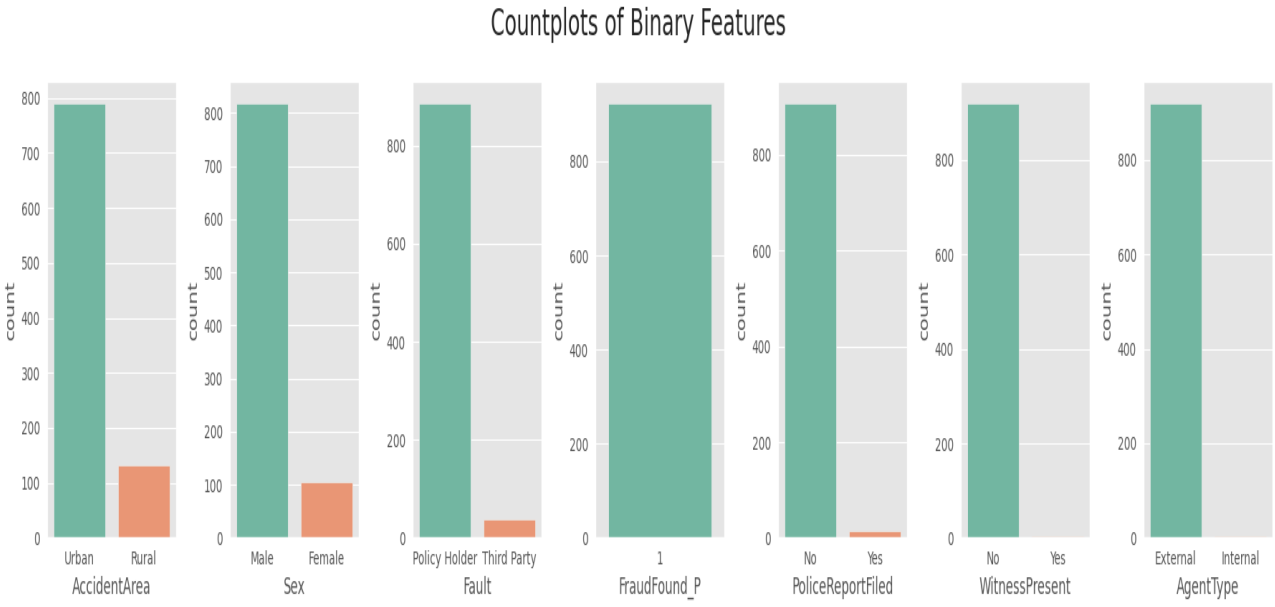


Figure 5 Countplots of Binary Features

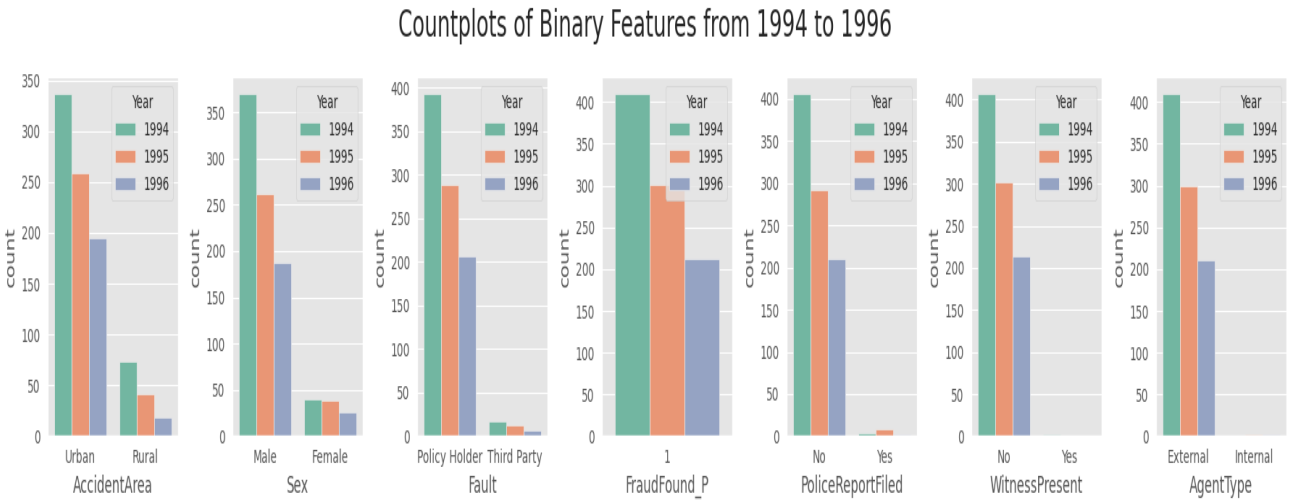


Figure 6 Countplots of Binary Features for the period

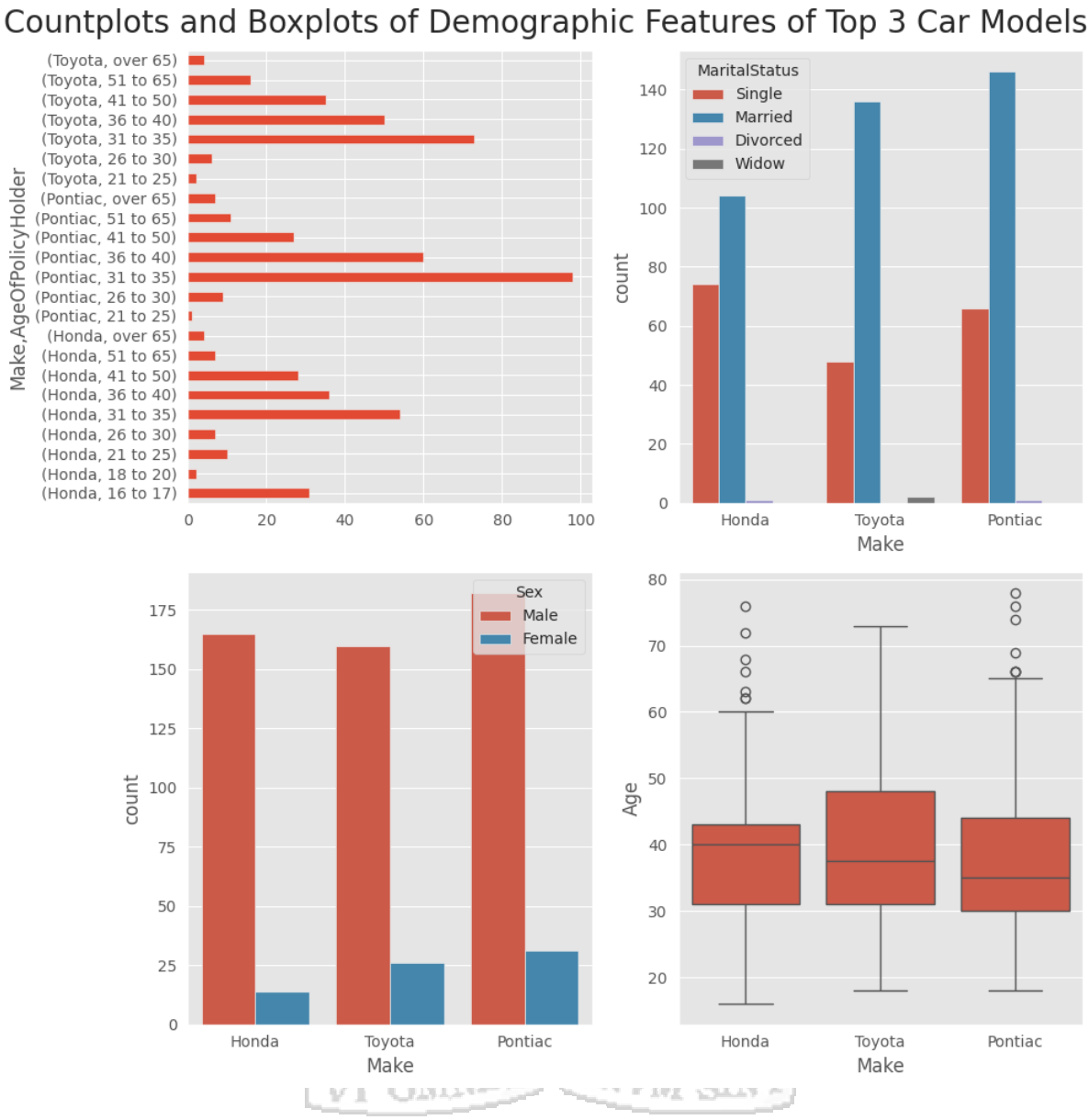


Figure 7 Demographic features of Top 3 Car Models

The graphs above show various demographic features of the owners of the top 3 car models which include; Toyota, Pontiac and Honda.

Beginning with the age, most popular age range for the 3 car models is 31 – 35 and most of them are men who are married.

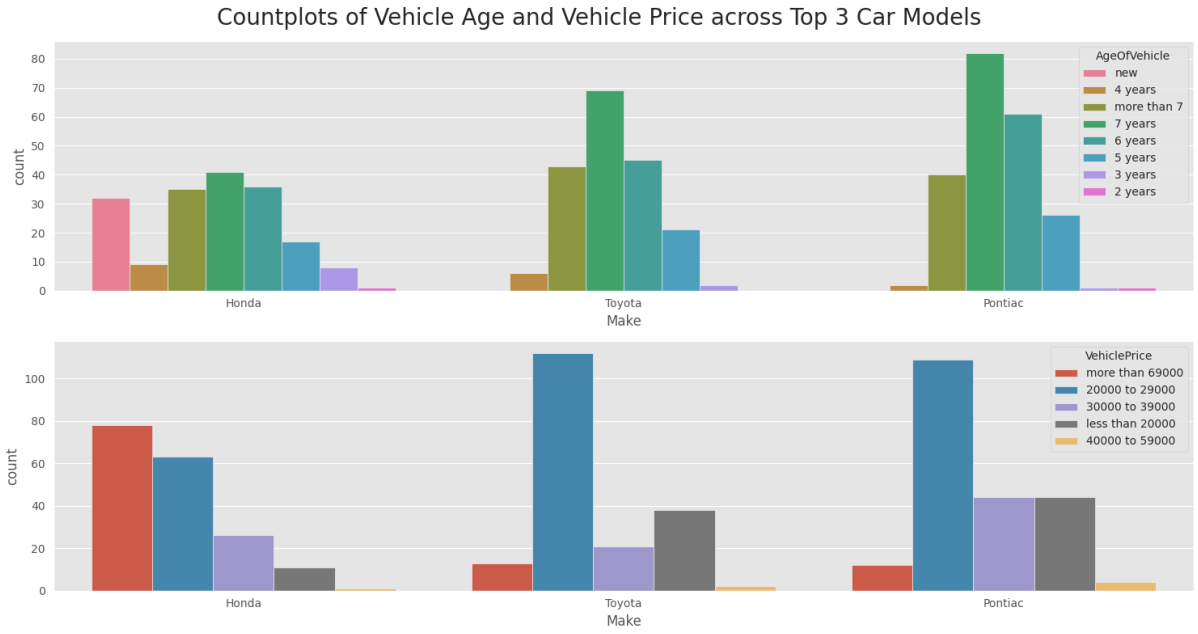


Figure 8 Vehicle Age and Price for the Top 3 Car Models

```
[ ]  fraud.loc[fraud['VehiclePrice'].isin(['more than 69000']) & fraud['AgeOfVehicle'].isin(['new']), 'Make'].value_counts()

Make
Honda    32
Name: count, dtype: int64
```

Figure 9 Checking for Car Model in positive fraud cases within more than 69000$ price

From the above plot, it is noted that all brand new cars priced over $69000 were Honda models. This indicates that more investigation is needed for Honda cars and its related insurance policies and regulations. For Toyota and Pontiac most cars were about 7 years old, and their prices ranged majorly between 20000$ - 29000$.

4.4 Data Transformation

The data was translated into forms that machine learning classifiers could understand. Text values must be transformed to integers since machine learning classifiers cannot read them. I encoded categorical data in integer format, allowing it to be used in various models. This enhanced the accuracy of our models. Verma (2021) describes categorical data encoding as the process of converting categorical variables into integers for use in models to improve predictions. He defined categorical data as information grouped into categories with restricted values. I used label encoding due to the nature of the features in the dataset. Label encoding

works well for ordinal data, where the order of categories is meaningful. The features transformed were 15 as shown below.

Label Encoding

```
from sklearn.preprocessing import LabelEncoder

# Keep a copy of the original DataFrame with the categorical columns
df_categorical = df[['AccidentArea', 'Sex', 'MaritalStatus', 'Fault', 'PolicyType', 'VehicleCategory',
                     'Days_Policy_Accident', 'Days_Policy_Claim', 'PoliceReportFiled', 'WitnessPresent',
                     'AgentType', 'AddressChange_Claim', 'NumberOfCars', 'Year', 'BasePolicy']].copy()

# Initialize LabelEncoder
label_encoder = LabelEncoder()


# Apply LabelEncoder to each categorical column
for column in df_categorical.columns:
    df_categorical[column] = label_encoder.fit_transform(df_categorical[column])

# Concatenate the label encoded DataFrame with the original numerical columns
df_final = pd.concat([df_categorical, df.select_dtypes(exclude=['object'])], axis=1)

# Final check to see if all the features are in integer or float datatypes after encoding
print(df_final.info())
```

Figure 10 Label Encoding Categorical Data

4.5 Data Modeling

In order to prepare the data for model training, we had to apply techniques to deal with the imbalanced data. Class imbalance in machine learning datasets can lead to biased models that favor the majority class. To address this, resampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling Approach) are employed. SMOTE generates synthetic data points for the minority class by leveraging existing data points and their k-nearest neighbors (Chawla, 2002). ADASYN builds upon SMOTE by assigning weights to minority class data points based on their local density distribution (He, 2008). This allows ADASYN to prioritize oversampling in areas where the classification boundary is less certain, potentially leading to improved model performance.

```
In [139]: # Define X and y dataset for Machine Learning
          X = df1.iloc[:, 1::]
          y = df1.iloc[:, 0]

In [141]: #Split the dataset into training and testing sets
          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

          #Apply SMOTE to balance the training data
          smote = SMOTE(random_state=42)
          X_train_resampled, y_train_resampled = smote.fit_resample(X_train, y_train)

In [142]: # Train classifiers
          classifiers = {
              'KNN': KNeighborsClassifier(),
              'Logistic Regression': LogisticRegression(),
              'Random Forest': RandomForestClassifier(),
              'XGBoost': XGBClassifier()
          }

          trained_models = {}
          for name, clf in classifiers.items():
              clf.fit(X_train_resampled, y_train_resampled)
              trained_models[name] = clf
```

Figure 11 Fitting the classifier models

The above figure shows the splitting of the train and test data, while applying the SMOTE resampler and fitting in our classification models; KNN, Logistic Regression, Random Forest and Xgboost.

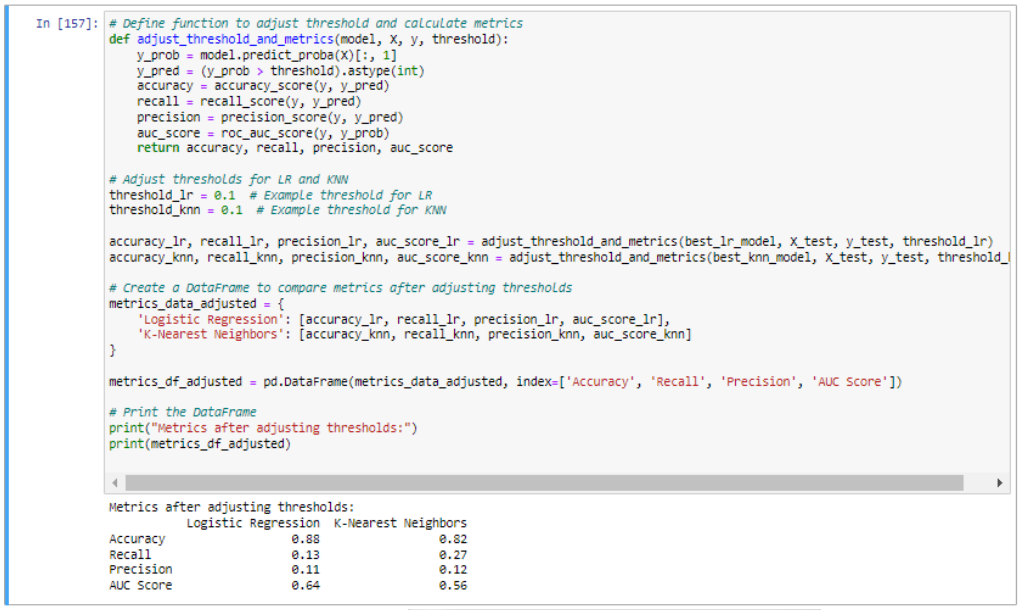
	Accuracy	Recall	Precision	AUC Score
KNN	0.76	0.47	0.13	0.68
Logistic Regression	0.75	0.60	0.15	0.79
Random Forest	0.93	0.04	0.26	0.80
XGBoost	0.93	0.06	0.38	0.83

Table 3 Evaluation Metrics for the 4 Classifiers

Table one gives us the evaluation metrics of the four classifiers. Since our data is imbalanced, we will focus on recall as the main evaluation criteria because recall gives the number of correctly identified fraudulent cases which is in alignment with our main research objective. Random Forest and Xgboost have the highest accuracy but have the lowest recall. If our data was balanced, we would have proceeded with the optimization of those models but in this case, we will go ahead to optimize the KNN and Logistic regression models.

I proceeded with the hyperparameter tuning on the 2 models as well as adjusting for probability thresholds, to see if we can improve the accuracy and recall scores. The accuracy improved,

however, the recall value decreased which was not the kind of results we were looking for. Hence I decided to apply the ADASYN resampler on both models.



```
In [157]: # Define function to adjust threshold and calculate metrics
def adjust_threshold_and_metrics(model, X, y, threshold):
    y_prob = model.predict_proba(X)[:, 1]
    y_pred = (y_prob > threshold).astype(int)
    accuracy = accuracy_score(y, y_pred)
    recall = recall_score(y, y_pred)
    precision = precision_score(y, y_pred)
    auc_score = roc_auc_score(y, y_prob)
    return accuracy, recall, precision, auc_score

# Adjust thresholds for LR and KNN
threshold_lr = 0.1  # Example threshold for LR
threshold_knn = 0.1  # Example threshold for KNN

accuracy_lr, recall_lr, precision_lr, auc_score_lr = adjust_threshold_and_metrics(best_lr_model, X_test, y_test, threshold_lr)
accuracy_knn, recall_knn, precision_knn, auc_score_knn = adjust_threshold_and_metrics(best_knn_model, X_test, y_test, threshold_

# Create a DataFrame to compare metrics after adjusting thresholds
metrics_data_adjusted = {
    'Logistic Regression': [accuracy_lr, recall_lr, precision_lr, auc_score_lr],
    'K-Nearest Neighbors': [accuracy_knn, recall_knn, precision_knn, auc_score_knn]
}

metrics_df_adjusted = pd.DataFrame(metrics_data_adjusted, index=['Accuracy', 'Recall', 'Precision', 'AUC Score'])

# Print the DataFrame
print("Metrics after adjusting thresholds:")
print(metrics_df_adjusted)

Metrics after adjusting thresholds:
           Logistic Regression  K-Nearest Neighbors
Accuracy                  0.88                 0.82
Recall                    0.13                 0.27
Precision                 0.11                 0.12
AUC Score                 0.64                 0.56
```

Figure 12 Adjusting probability thresholds of KN & LR

The ADASYN resampler had an improvement on the model's evaluation scores as seen below on table 4. We proceeded to do hyperparameter tuning for the models using the ADASYN resampler, and an improvement on KNN's recall was observed, while the LogisticRegression recall remained the same and observed in table 5.

	Logistic Regression	**K-Nearest Neighbors**
Accuracy	0.75	0.76
Recall	0.60	0.48
Precision	0.15	0.13
AUC Score	0.79	0.67

Table 4 Metrics after using ADASYN Resampler

	Logistic Regression	**K-Nearest Neighbors**
Accuracy	0.75	0.74
Recall	0.60	0.52
Precision	0.15	0.13
AUC Score	0.79	0.71

Table 5 Metrics after Hyperparameter Tuning

The recall for KNN Classifier has improved from 0.48 to 0.52 as well as the AUC score from 0.67 to 0.71, while the recall for LR has remained the same as before hyperparameter tuning at 0.60. Since the LR has performed better than KNN, we will save the LR Classifier model to be used in deployment.

4.6 Deployment

After considering the findings given above, the Logistic Regression classifier was selected as the model to be used in the web-based application for detecting fraudulent auto insurance claims. The Streamlit Framework was used in the development of our web-based application. Patil and Lokesha (2022) define Streamlit as a Python-based platform that is open-source and used for creating and implementing interactive data science dashboards and machine learning models on web applications. The figures below depict snapshots of a web application interface that allows the user to input details regarding the claim which will be used to predict the nature of the claim. The various input categories include; policy information, vehicle information, accident information and customer information.

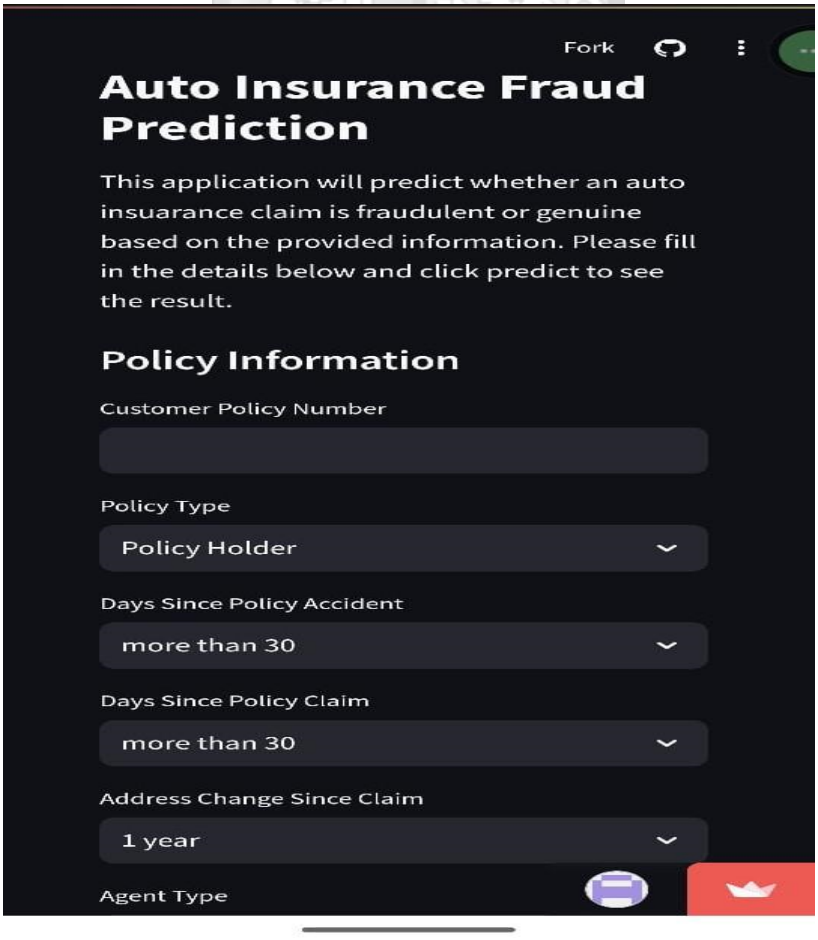


Figure 13 Streamlit web page - Policy Information

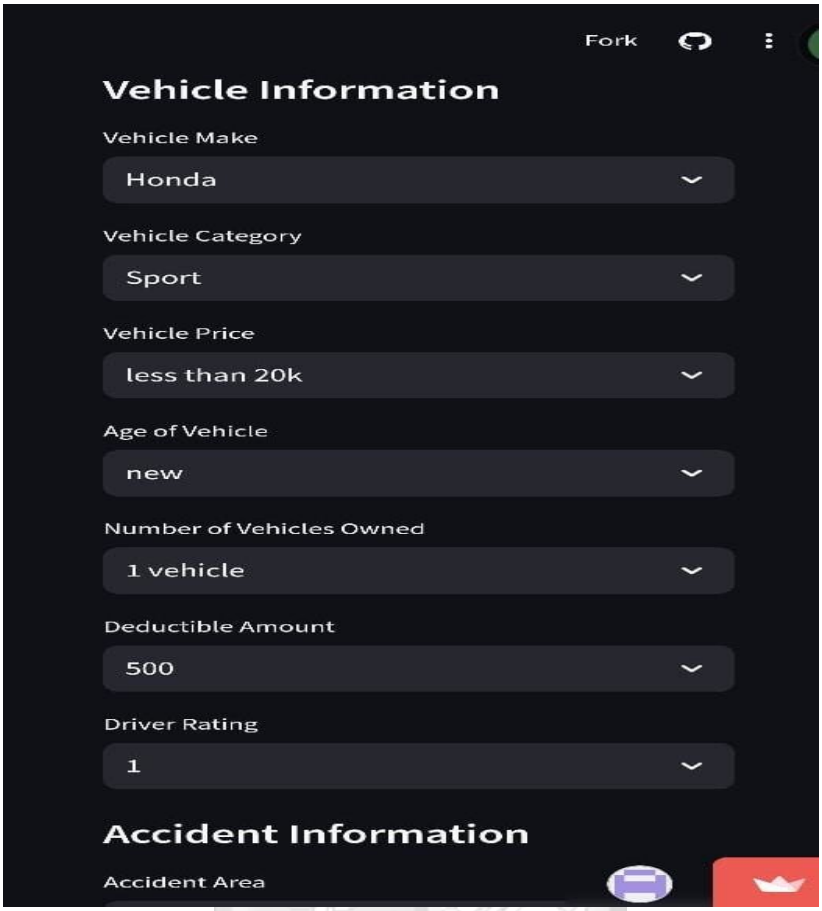


Figure 14 Streamlit web page - Vehicle Information

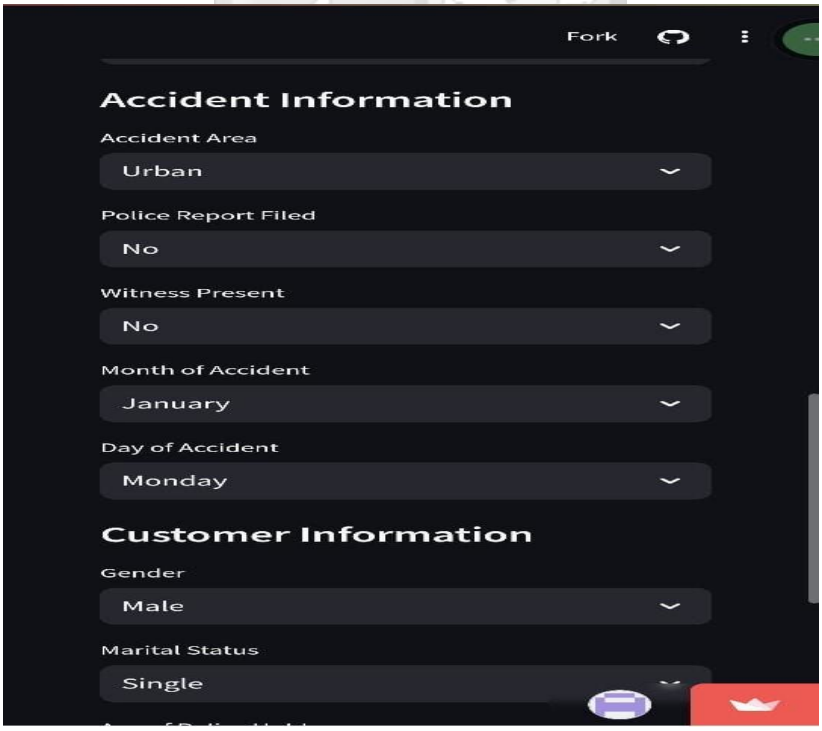


Figure 15 Streamlit web page - Accident Information

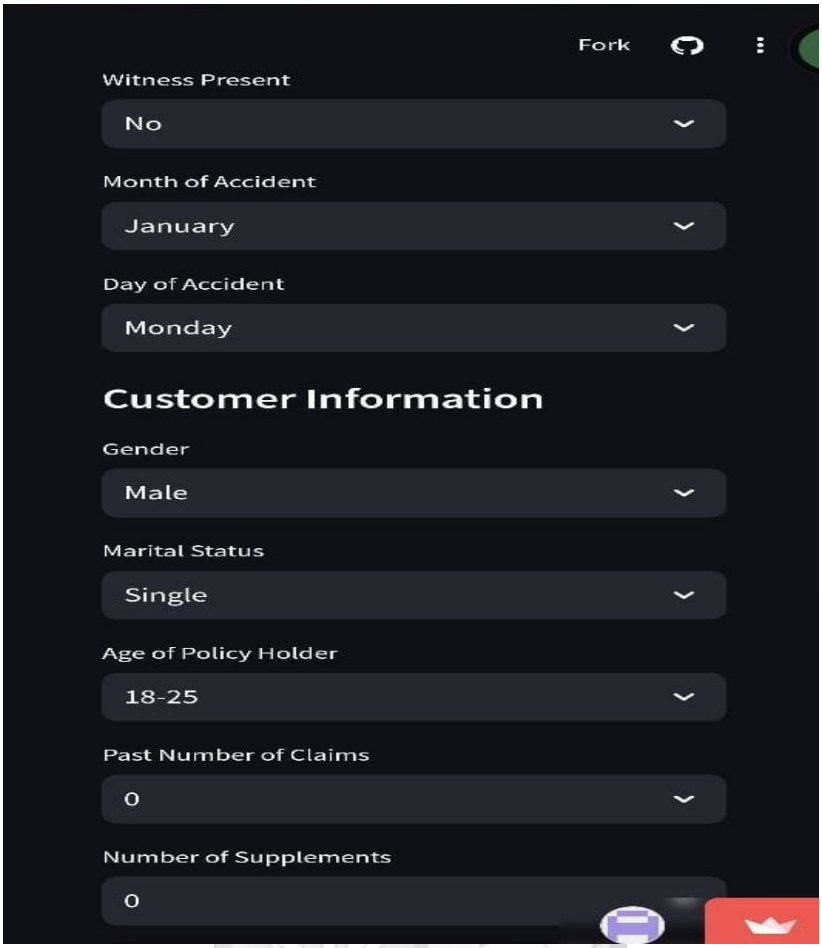


Figure 16 Streamlit web page - Customer Information

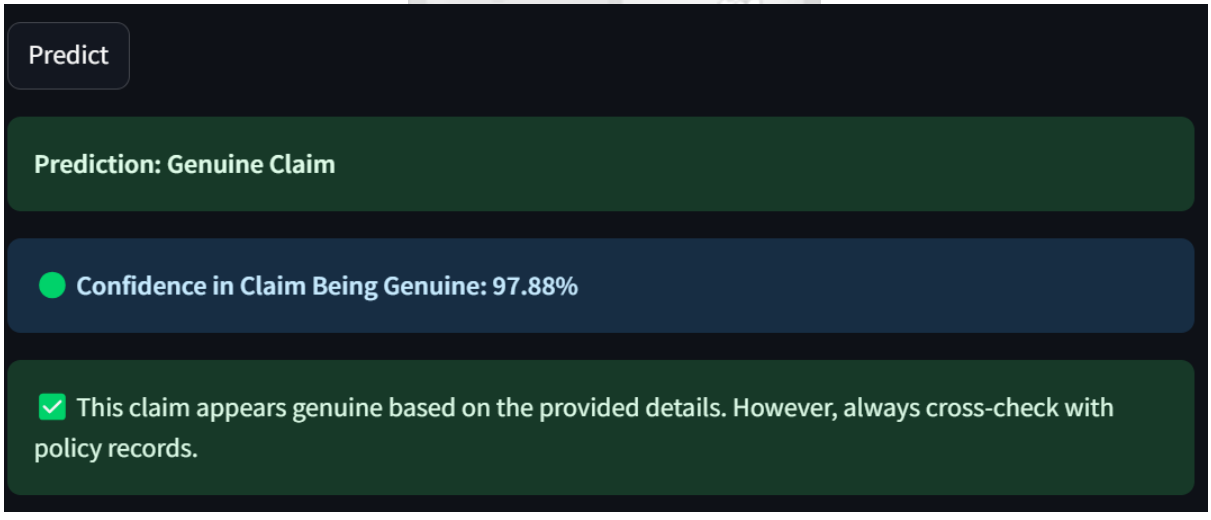


Figure 17 Streamlit Web Page - Prediction

CHAPTER 5: CONCLUSION

5.1 Conclusion

This study explored the use of supervised machine learning algorithms to detect fraudulent auto insurance claims. The models—Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, and XGBoost—were evaluated based on key performance metrics, including accuracy, recall, precision, and AUC scores. Given the class imbalance in the dataset, recall was prioritized as the main evaluation metric, as it determines the model's ability to correctly identify fraudulent claims.

Before deploying the final model, extensive training and parameter tuning was conducted. The initial results showed that Random Forest and XGBoost achieved the highest accuracy, but their recall was significantly lower, making them less effective for fraud detection. This was primarily due to the models' tendency to focus on the majority class (legitimate claims) while failing to capture fraudulent instances. In contrast, KNN and Logistic Regression demonstrated higher recall but required further optimization. To enhance performance, hyperparameter tuning was applied to KNN and Logistic Regression. Techniques such as adjusting probability thresholds and optimizing model hyperparameters were implemented. The ADASYN resampler was also applied to address the imbalanced dataset, improving recall and overall fraud detection capability. As a result, KNN's recall improved from 0.47 to 0.52, while Logistic Regression maintained a recall of 0.60, making it the most effective model for deployment since it was the most balanced model in terms of performance, interpretability, and fraud detection capability.

The comparison of model results before and after parameter tuning provided critical insights into model behavior. While Random Forest and XGBoost maintained high accuracy (~93%), they struggled to detect fraudulent claims, leading to a recall as low as 0.04 and 0.06, respectively. This suggested that, although these models performed well on overall classification, they were not ideal for fraud detection due to the minority class imbalance. KNN and Logistic Regression, on the other hand, demonstrated better fraud detection capabilities. KNN initially struggled with low precision, indicating some false positives, but parameter tuning and resampling significantly improved its performance. Logistic Regression, with a recall of 0.60, was ultimately chosen for deployment due to its balance between precision, recall, and interpretability.

The final deployed model was integrated into a web-based fraud detection system using the Streamlit framework. The prototype allowed users to input claim details and receive fraud predictions in real-time. The model demonstrated strong predictive performance, accurately identifying fraudulent claims and maintaining a recall above 0.55, indicating good generalization. This suggests that while the model is effective, continuous retraining with more recent and diverse data is necessary to maintain high detection accuracy.

The results underscore the importance of integrating machine learning in fraud detection systems to minimize financial losses for insurers and enhance operational efficiency. By leveraging machine learning and data analysis, insurers can detect suspicious claim patterns, reduce reliance on manual reviews, and improve the accuracy of fraud identification. This not only helps prevent fraudulent payouts but also streamlines legitimate claims processing, leading to reduced investigation time, lower administrative costs, and improved customer satisfaction.

5.2 Recommendation

The dataset that we used to construct this insurance prediction model covered the years 1994 through 1996. We would benefit by gathering the most recent dataset over the last two to five years. To evaluate how well the proposed solution performs, it's a good idea to test a mix of random and planned parameters. The model should also be compared against a similar dataset to see how adaptable it is. Finally, testing with other datasets that reflect different conditions from those used during data preparation can provide a clearer picture of its reliability. In order to further reduce computational expenses, it is recommended to reduce the number of attributes.

5.2 Future Work

Future work should try to employ a sophisticated or newly acquired dataset in order to evaluate the efficacy of machine learning and deep learning approaches. Furthermore, since the current dataset is imbalanced, it is recommended to use an alternative dataset. To develop a more universal feature selection and focus strategy, further evaluation is necessary to determine the significance of features across various datasets, regardless of whether they share similar properties.

A motor insurance claim fraud detection system has a lot of promise. It may be improved for more accuracy and efficiency with developments in machine learning and data analysis. While integration with social media and law enforcement databases may provide deeper insights into

fraud charges, using bigger datasets can aid in the detection of subtle fraud tendencies. Adding more categories for insurance claims, such as life, health, and property, to the system may provide insurers with thorough fraud detection, enhance corporate processes, and lower losses. Overall, the relevance of such a system is anticipated to increase in the future due to its critical role in preventing fraud.

Appendix I: Similarity Report

Auto Insurance Fraud Detection Using ML .pdf

ORIGINALITY REPORT

18%	15%	8%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	erepository.uonbi.ac.ke:8080 Internet Source	4%
2	Submitted to Midlands State University Student Paper	2%
3	www.jetir.org Internet Source	2%
4	scholarworks.rit.edu Internet Source	2%

Appendix II: Ethical Clearance Release Letter

Strathmore UNIVERSITY

8th May 2024

Wangari Kimani Ruth

092833

wangari.kimani@strathmore.edu

Dear Ruth,

RE: Auto Insurance Fraud Detection Using Machine Learning

This is to inform you that the Office of Graduate Studies in March/April 2024 received your request for intervention/assistance following your referral by the Strathmore University Institutional Scientific and Ethics Review Committee (SU-ISERC) to our Office due to the fact that you stated that you had already collected and/or analysed its data prior to seeking Ethical clearance. The ethics approval process is ONLY done before any collection of primary or secondary data.

We have taken note of your response that the information that you provided was misleading.

This is a letter for you to proceed with the next steps of your academic requirements.

Please be advised, that in future, all research proposals should be submitted to the SU-ISERC through the RHInnO Ethics platform: https://strathmoreuniversity.rhinno.net/login

Disclaimer: 1) This is not in any way an ethical approval letter. 2) Should there be any legal implications/actions emanating from the research in terms of any ethical violations, you will be personally liable.

Yours sincerely,

Dr. Bernard Shibwabo

Director of Graduate Studies

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000 Email admissions@strathmore.edu www.strathmore.edu

REFERENCES

Aisha Abdallah, M. A. (2016.). Fraud detection system: A Survey. *Journal of Network and Computer Applications*, 68:90–113.

AKI, A. o. (2021). *Insurance Industry Statistics Report.* Nairobi, Kenya.

Alrais, A. (2022). *Fraudulent Insurance Claims Detection Using Machine Learning.* Rochester Institute of Technology.

Automobile Insurance Fraud Detection using Artificial Neural Networks. (2022). From The Actuary: https://www.actuaries.or.ke/automobile-insurance-fraud-detection-using-artificial-neural-networks/

Bhasin, H. &. (2018). Fraud Detection in Automobile Insurance using Machine Learning Techniques. *International Journal of Computer Applications*, 181(5), 8-11.

Bilodeau, M. J. (2008). Cluster analysis in automobile insurance fraud detection: An introduction. *Journal of Data Science*, 429-446.

Brownlee, J. (2020). "Precision, Recall and F-Measure: Understanding the Metrics and When to Use Them".

Cedervall., A. H. (2022). Insurance fraud detection using unsupervised sequential anomaly detection.

Chamal Gomes, Z. J. (2021). Insurance fraud detection with unsupervised deep learning. *Journal of Risk and Insurance*, 88(3):591– 624.

Chen, T. &. (2016). *"XGBoost: A highly scalable gradient boosting machine".*

Dai, B. X. (2012). Application of Bayesian networks in insurance fraud detection. *IEEE 12th International Conference on Data Mining*, (pp. (pp. 139-148)).

Dwivedi., A. K. (2020). Fraud detection in credit card data using unsupervised machine learning based scheme. In 2020 international conference on electronics and sustainable communication systems (ICESC), . *IEEE*, pages 421–426.

Fernando, E. (2021). *Machine Learning Approaches On Motor Insurance Fraud Detection.*University of Colombo School of Computing.

Gareth James, D. W. (2021). *"The Elements of Machine Learning Algorithms".*

George, R. (2022). "Handling Class Imbalance in Fraud Detection Using Machine Learning Techniques".

Gill, P. L. (2005). *The Evolution of Fraud.* The Actuary.

Harrington., P. (2012). Preparing Your Data. In *"Machine Learning in Action".*

Hosmer, D. W. (2004). "Applied logistic regression".

Hotz., N. (2023, January 19). *What is CRISP DM?* From Data Science Process Alliance: https://www.datascience-pm.com/crisp-dm-2/

Huang, D. L. (2020). A time-series analysis method for insurance fraud detection. *PLoS ONE 15(1), e0227961*.

IRA, I. R. (2022). *Insurance Industry Annual Report.* Nairobi, Kenya.

Irshad Hussain B, P. K. (2023). Vehicle insurance fraud Detection using Machine Learning. *Journal of Emerging Technologies and Innovative Research (JETIR)* www.jstor.org.

Johnson, M. K. (2019). *"Feature Engineering and Selection: A Practical Approach for Predictive Models"* .

Karen M Gill, A. W. (2005). Insurance fraud: the business as a victim? *Crime At Work: Studies in Security and Crime Prevention Volume I*, 73–82.

Karimi, A. R. (2019). Machine learning in insurance fraud detection. *Expert Systems with Applications*, 105-120.

King, G. &. (2001). "Logistic regression in rare events data".

Liaw, A. &. (2002). *"Classification and regression by randomForest".*

Markovskaia, N. (2020, July 9). *Detecting Insurance Fraud with Machine Learning.* . From PlugandPlay: https://www.plugandplaytechcenter.com/resources/detecting-insurance-fraud-machine-learning/

Morgan, M. G. (2022). *"Applied Univariate, Bivariate, and Multivariate Statistics.*

Naseer, S. e. (2020). "XGBoost for financial fraud detection".

Njeru, A. (2022). *Detection of Fraudulent Vehicle Insurance Claims Using Machine Learning.* . Department of Computer Science & Informatics, University of Nairobi.

Parab, R. (2020). *Performance Evaluation Metrics for Machine Learning Models with Python Code*. From Medium: https://medium.com/swlh/performance-evaluation-metrics-for-machine-learningmodels-

Rahamathunnisa, S. &. (2021). Predicting Fraudulent Claims in Vehicle Insurance Using Machine Learning Techniques. *Proceedings of the 2nd International Conference on Innovative Computing and Communication (ICICC)*, (pp. 193-197.).

Rukhsar, L. H. (2022). Prediction of Insurance Fraud Detection using Machine Learning Algorithms. *Mehran University Research Journal of Engineering and Technology,vol. 41, no. 1,*, pp. 33–40,2022. doi:10.22581/muet1982.2201.04.

Simha, A. &. (2016). Straight From the Horse's Mouth: Auditors' on Fraud Detection and Prevention, Roles of Technology, and White-Collars Getting Splattered with Red! . *Journal of Accounting & Finance Vol. 16(1)*, pp. 26-44.

Subelj, L. F. (2011). Network-based Meta-analysis of Auto Insurance Fraud. *Expert Systems with Applications.*, 38(1), 384-392.

Supervisors, I. A. (2011, September). From Application Paper on Deterring, Preventing, Detecting, Reporting and Remedying Fraud In Insurance.: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.iaisweb.org/uploads/2022/01/Application_paper_on_fraud_in_insurance.pdf.pdf

Tse, T. H. (2018). Identifying auto insurance fraud with network analysis: A novel hybrid model. . In *Expert Systems with Applications.* (pp. 249-260).

Urunkar, A. K. (2022). Fraud Detection and Analysis for Insurance Claim using Machine Learning. *2022 IEEE International Conference on Signal Processing, Revue ame, Vol 4, No 4 (Octobre, 2022) 96-113 Page 113 Informatics,* (pp. 96 - 113). Communication and Energy Systems (SPICES).

Viaene, S. &. (2015). Insurance Fraud: Issues and Challenges. *Geneva Papers on Risk and Insurance-Issues and Practice.*, Vol. 29, No.2., pp. 313-333.

Weinberger, K. &. (2006). "Distance metric learning for large-scale nearest neighbor search" .

Ye, Y. F. (2019). A comparative study of classification algorithms. *Journal of Big Data*, 6(1), 1-21.

Young., E. &. (2011). Fraud Insurance on the Rise. *India Survey*.

Zhang, X. H. (2019). HOBA: A novel feature engineering methodology for credit card fraud detection with a deep learning architecture. . *Information Sciences*, 557, 302–316.