BT4012 Project Final Report

Detecting and Predicting Vehicle Insurance Fraud

Github Repo

Student Name	Student ID
Bian Tong	A0234470L
Kang Xuan Hui Victoria	A0241262R
Ke Jiayi	A0241155N

Academic Year 2024/2025 Semester 1

1 Motivation

Auto insurance fraud is a significant global issue with far-reaching consequences. In the United States alone, it is estimated that car insurance fraud results in approximately $29 billion in annual premium leakage (Insurance Information Institute, 2022). Similarly, in Europe, total losses from insurance fraud, including car insurance, amount to tens of billions of euros. In Singapore, data from the General Insurance Association (GIA) reveals that around 20% of motor insurance claims carry elements of fraud, with a single scam potentially causing losses of up to SG$1.6 million (Araullo, 2023).

Imagine the transformative potential of effectively detecting fraudulent insurance claims. It could lead to lower car insurance premiums, easing the financial burden of car ownership for consumers. Moreover, robust fraud detection would deter criminal activities, such as staged accidents, that pose dangers to innocent bystanders.

2 The Problem

Detecting Vehicle Insurance Fraud: Challenges and Analytical Solutions

Vehicle insurance fraud is a significant issue for insurance companies, leading to financial losses and increased premiums for honest customers. Fraudulent claims can take various forms, such as staged accidents, exaggerated damages, vehicle theft fraud, ghost claims, and identity fraud. These diverse tactics, combined with high data volumes and nuanced indicators of fraud, make detection a complex and multifaceted challenge.

Key challenges include:

1. **Complexity of Fraudulent Behaviours**
    - Fraudsters employ diverse and evolving schemes, often involving collusion with external parties like repair shops or even law enforcement, which complicates detection.
2. **Data Overload**
    - Vast amounts of data must be processed, with fraudulent claims often mimicking legitimate ones, making traditional rule-based detection insufficient.
3. **Subtle Fraud Indicators**
    - Fraud indicators, such as inconsistencies in testimonies or recurring claim patterns, are often subtle and context-dependent. Behavioural data like reporting speed or claim history, while potentially insightful, are challenging to interpret consistently.
4. **Human Subjectivity**
    - Claims adjusters may introduce cognitive biases or overlook suspicious details due to time pressures, leading to errors in judgement.
5. **Legal and Ethical Constraints**
    - Techniques like social media monitoring or location tracking raise concerns about privacy and compliance with ethical standards.

Given these challenges, there is a need for advanced and efficient analytics approaches to enhance detection of fraudulent claims amidst a sea of legitimate claims. This report aims to explore how data-driven solutions can identify fraudulent claims while balancing operational efficiency and ethical considerations. The analysis will focus on leveraging structured data, developing predictive models, and addressing key challenges to improve fraud detection outcomes effectively.

3 Datasets

To better understand how we can understand and predict vehicle insurance fraud, we will be studying this dataset of vehicle insurance claims data from the United States between the years of 1994 and 1996, from Kaggle. As this dataset is rather comprehensive, and comprise data within a reasonable time-frame where technology did not undergo rapid development, it is a suitable starting point in understanding fraud patterns.

The dataset, Vehicle Insurance Claim Fraud Detection Dataset by Oracle (Vehicle Insurance Claim Fraud Detection, 2021), has 33 features indicating policy information, policyholder information, claim details and the class variable 'FraudFound_P' that contains the label – 1 stands for "true" and 0 stands for "false". It contains 15,420 records of policy claims with only 6% (923 records) are fraudulent.

Column Name	**Description**	**Type**
Month	The month during which the insurance claim was submitted.	Categorical
WeekOfMonth	The specific week of the month when the claim was filed.	Numerical
DayOfWeek	The day of the week on which the claim submission occurred.	Categorical
Make	The vehicle manufacturer associated with the insurance claim.	Categorical
AccidentArea	The location type where the accident happened (e.g., urban or rural).	Categorical
DayOfWeekClaimed	The day of the week when the insurance claim was processed.	Categorical
MonthClaimed	The month during which the claim was processed.	Categorical
WeekOfMonthClaimed	The specific week of the month when the claim was processed.	Numerical
Sex	The gender of the policyholder.	Categorical
MaritalStatus	The marital status of the policyholder.	Categorical
Age	The age of the policyholder.	Numerical
Fault	Specifies if the policyholder was at fault in the accident.	Categorical
PolicyType	The category of the insurance policy (e.g., comprehensive or third-party).	Categorical
VehicleCategory	The classification of the vehicle (e.g., sedan, SUV).	Categorical
VehiclePrice	The value of the vehicle in question.	Categorical
FraudFound_P	Indicates whether fraud was identified in the claim.	Numerical
PolicyNumber	A unique number assigned to the insurance policy.	Numerical
RepNumber	A unique identifier for the insurance representative managing the claim.	Numerical

Deductible	The out-of-pocket cost the policyholder must cover before insurance takes over.	Numerical
DriverRating	An evaluation of the driver's record, often based on driving history.	Numerical
Days_Policy_Accident	The time elapsed (in days) between the policy's start date and the accident.	Categorical
Days_Policy_Claim	The number of days between the policy's start date and the claim submission.	Categorical
PastNumberOfClaims	The total number of previous claims by the policyholder.	Categorical
AgeOfVehicle	The age of the vehicle involved in the claim.	Categorical
AgeOfPolicyHolder	The age of the policyholder.	Categorical
PoliceReportFiled	Indicates whether a police report was submitted for the incident.	Boolean
WitnessPresent	Specifies if a witness was present at the accident scene.	Boolean
AgentType	The category of the insurance agent handling the policy (e.g., internal or external).	Categorical
NumberOfSuppliments	The count of additional documents or related claims associated with the main claim, grouped into ranges.	Categorical
AddressChange_Claim	Indicates whether the policyholder's address changed at the time of the claim, grouped into ranges.	Categorical
NumberOfCars	The total number of cars insured under the policy, grouped into ranges.	Categorical
Year	The calendar year during which the claim was submitted or processed.	Numerical
BasePolicy	The primary type of insurance coverage (e.g., Liability, Collision, or All Perils).	Categorical

Table 1. Dataset columns and description

4 Data Preprocessing and Modelling

Data Preprocessing

The dataset was first analysed for missing values. It was found that all columns were fully populated, indicating no missing data, which eliminated the need for imputation. Next, the dataset was examined for data types and unique value counts. Most variables had a limited number of unique values, except for 'PolicyNumber' (a unique identifier) and 'Age' (66 unique values). We had converted those variables with few unique values to object data type for better categorical analysis. This is critical because treating categorical variables as numerical without proper transformation could lead to incorrect assumptions in models.

Exploratory Data Analysis (EDA) was conducted to better understand the data distribution and relationships between variables. A significant problem identified during this step was the highly imbalanced nature of the target variable, 'FraudFound_P'. Fraudulent cases constituted only 6% of the dataset (923 observations), with the majority being non-fraudulent (14,497 observations). Class imbalance could result in models biassed toward the majority class, predicting non-fraudulent cases accurately while failing to detect fraud. To address this issue, **Synthetic Minority Oversampling Technique (SMOTE)** was applied. SMOTE generates synthetic samples for the minority class by interpolating between existing samples, thereby balancing the dataset and improving the model's sensitivity to fraud.

Moreover, histogram analysis revealed that 'Age' contained unrealistic values - 0, which were treated as outliers and dropped. 'PolicyNumber', being a unique identifier with no meaningful pattern for predicting fraud, was deemed uninformative and dropped too. Additionally, a **Cramér's V correlation heatmap** was used to analyse the relationships between categorical variables. This helped to identify highly correlated variables ('Month', 'VehicleCategory' and 'BasePolicy'), which were removed to avoid redundancy and multicollinearity that might lead to overfitting.

After cleaning, transformations were applied based on the nature of each variable. Binary variables, such as 'AccidentArea', 'Sex' and 'PoliceReportFiled', were label-encoded, converting them into numerical values of 0 and 1. For ordinal variables like 'Days_Policy_Accident' and 'AgeOfVehicle', logical mappings were created to preserve the inherent order. Nominal variables, such as 'Make', 'MaritalStatus' and 'PolicyType', were label-encoded for model compatibility. Finally, the dataset was split into training, validation and test sets, allowing for robust model training, tuning, and evaluation. StandardScaler was applied to standardise feature magnitudes. It is a critical step for models sensitive to scale, such as Logistic Regression.

Feature Selection

Initial feature selection was performed using LASSO (Least Absolute Shrinkage and Selection Operator) regression, which uses an L1 penalty to shrink less important coefficients to zero. LASSO was applied with a regularisation parameter (α=0.01) and identified only three features with non-zero coefficients ('Fault', 'PastNumberOfClaims' and 'PolicyType'). While LASSO effectively eliminates irrelevant features, the limited number of retained predictors suggested it might be too aggressive, potentially missing meaningful variables. This could lead to underfitting, where the model lacks sufficient complexity to capture important patterns.

Therefore, to address this limitation, **ElasticNet** regression was employed, combining L1 (LASSO) and L2 (Ridge) penalties. ElasticNet with α=0.01 and L1 ratio=0.5 struck a balance between feature sparsity and robustness to multicollinearity. This approach identified 15 features with non-zero coefficients, providing a more comprehensive set of predictors for model training, while minimising the risk of overfitting.

Modelling

For this fraud detection problem, we had implemented 6 types of machine learning model. Each model was chosen for its unique strengths in addressing the challenges posed by this dataset, such as the non-linear relationships between features, the imbalanced class distribution, and the need for robust predictive performance. To evaluate the impact of addressing class imbalance, all models were trained separately on the original (non-SMOTE) dataset and the SMOTE-augmented dataset. This allowed a direct comparison to determine whether SMOTE effectively improved performance on imbalanced data.

Logistic Regression was chosen as a baseline model due to its simplicity. It is a linear classifier that predicts the probability of fraud based on the weighted combination of features. It is particularly suitable for datasets where features exhibit linear relationships with the target variable. However, its limitations in capturing non-linear patterns makes it a good baseline model for comparison with more sophisticated models.

Random Forest was included for its ability to handle non-linear relationships and interactions between features. As an ensemble model, it combines the outputs of multiple decision trees, reducing the risk of overfitting and improving generalizability. Random Forest is also robust to noise and multicollinearity, making it highly suitable for fraud detection.

XGBoost is a gradient boosting algorithm. It was employed for its ability to capture complex, non-linear relationships in the data. Its regularisation techniques help prevent overfitting, and its scalability makes it efficient for large datasets. XGBoost natively handles missing data and supports custom evaluation metrics, making it highly adaptable to the challenges of fraud detection.

CatBoost was selected for its ability to handle categorical variables natively without requiring extensive preprocessing, such as one-hot encoding. This characteristic is particularly useful for datasets like this one, where categorical features play a critical role. CatBoost is robust to overfitting and performs well on imbalanced datasets.

A **Neural Network** was implemented using TensorFlow to capture complex, non-linear relationships in the data. The network consisted of multiple dense layers with ReLU activations, and dropout layers were added to prevent overfitting. Neural Networks are particularly suitable for datasets with intricate patterns, but their performance can be sensitive to imbalanced data.

K-Nearest Neighbors (KNN) is a non-parametric model. It was included for its intuitive approach to classification based on the majority class of neighbouring data points. While simple and easy to implement, KNN can struggle with high-dimensional data and imbalanced class distributions.

Evaluation Metrics

To comprehensively evaluate the models, we used several metrics. Each metric was chosen for its ability to capture different aspects of model performance, particularly in the context of imbalanced datasets.

Accuracy measures the proportion of correctly classified cases but is insufficient alone for imbalanced datasets as it may be dominated by the majority class.

Precision evaluates the proportion of correctly predicted fraud cases out of all predicted fraud cases. High precision ensures fewer false positives, reducing unnecessary investigations.

Recall (a.k.a Sensitivity) measures the proportion of actual fraud cases that were correctly identified. High recall is critical for minimising missed fraud cases (false negatives), which could lead to significant financial losses for the insurance company.

ROC-AUC quantifies the model's ability to distinguish between fraudulent and non-fraudulent cases across various thresholds. It is particularly valuable for imbalanced datasets as it reflects the model's overall discrimination capability, irrespective of the decision threshold.

By comparing the models across these metrics on both the original and SMOTE datasets, we aimed to identify the most effective model and determine the extent to which SMOTE improved performance. This

evaluation provided a nuanced understanding of each model's strengths and weaknesses in addressing the challenges of fraud detection.

Modelling Insights and Observations

Comparison between Original and SMOTE Data

Models	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression (Baseline)	74.15%	11.35%	48.84%	73.92%
KNN	93.98%	47.62%	7.75%	64.02%
Random Forest	93.28%	28.95%	8.53%	79.15%
XGBoost	93.38%	20.83%	3.88%	6.54%
CatBoost	82.91%	19.70%	60.47%	83.99%
Neural Network	94.03%	0.00%	0.00%	80.60%

Table 2. Model Results on Original Data

Models	Accuracy	Precision	Recall	ROC-AUC
Logistic Regression (Baseline)	67.67%	11.61%	66.67%	72.58%
KNN	80.92%	14.18%	43.41%	69.54%
Random Forest	90.23%	25%	31.78%	79.45%
XGBoost	89.02%	22.73%	34.88%	82.61%
CatBoost	88.74%	21.78%	34.11%	82.86%
Neural Network	80.13%	15.91%	54.26%	82.47%

Table 3. Model Results on SMOTE Data

After training the models on both the original dataset and the SMOTE-augmented dataset, we observed the following performance trends across key metrics:

Accuracy generally dropped when models were trained on the SMOTE dataset, with the exception of **CatBoost**, which slightly improved accuracy. This drop is expected, as SMOTE balances the dataset by introducing synthetic minority samples, which may reduce the dominance of the majority class and shift focus toward correctly predicting minority (fraudulent) cases. However, accuracy is not a reliable metric in imbalanced datasets, so this drop is not a significant concern.

Precision dropped significantly for **KNN**, likely due to its sensitivity to noisy or synthetic data introduced by SMOTE. Conversely, **Neural Networks** exhibited a significant increase in precision on the SMOTE dataset, indicating that the model became better at correctly predicting fraud cases while reducing false positives. For other models, precision remained relatively stable, showing that SMOTE did not drastically affect their ability to identify true fraud cases.

Recall generally increased on the SMOTE dataset for most models, with **KNN**, **Random Forest**, **XGBoost**, and **Neural Networks** showing significant improvements. This demonstrates SMOTE's effectiveness in enhancing sensitivity to minority (fraudulent) cases by exposing models to a more balanced representation of classes. Interestingly, **CatBoost** did not exhibit significant recall improvement on the SMOTE dataset. We plan to assess whether tuning hyperparameters further improves this metric.

ROC-AUC, which measures the model's ability to distinguish between classes, remained largely similar across datasets, with minor changes for most models. However, **XGBoost** demonstrated a significant increase in ROC-AUC when trained on the SMOTE dataset, indicating that its gradient boosting mechanism benefited substantially from the balanced data. This makes XGBoost a strong candidate for further tuning.

The results highlight the trade-offs introduced by SMOTE. While accuracy and precision may decrease in certain models due to the altered data distribution, **Recall** - a critical metric for fraud detection - improved significantly in most cases. Since our primary focus is on **Recall** and **ROC-AUC**, as these metrics directly impact the model's ability to detect fraudulent cases and its overall discrimination power, we conclude that training models on the SMOTE dataset is preferable.

Moving forward, we will prioritise further hyperparameter tuning of the models trained on the SMOTE dataset to enhance their performance. This approach aligns with our goal of minimising false negatives (missed fraud cases) while maintaining strong overall classification capabilities.

5 Results

Models	**Accuracy**	**Precision**	**Recall**	**ROC-AUC**
Logistic Regression (Baseline)	66.33%	10.97%	65.12%	72.67%
KNN	79.16%	15.03%	53.49%	75.04%
Random Forest	89.12%	22.40%	33.33%	81.74%
XGBoost	88.78%	21.19%	32.13%	82.59%
CatBoost	70.73%	13.71%	73.64%	80.60%
Neural Network	-	-	-	-

Table 4. Model Results after Hyperparameter Tuning

During hyperparameter tuning, we chose to prioritise recall as the primary metric for optimization, as it reflects the model's ability to correctly identify positive cases—a critical requirement for our use case. In the context of vehicle insurance fraud detection, emphasising recall ensures the system flags as many

potentially fraudulent claims as possible, even at the cost of some false positives. This approach aligns with business priorities, as the cost of investigating false positives is significantly lower than the financial and reputational risks of failing to detect fraudulent claims. In Table 4, the Random Forest model achieved the highest accuracy and precision. However, since our focus is on recall, the CatBoost model performed the best on the validation set. Therefore, we selected the tuned CatBoost model for testing on the test set.

The tuned CatBoost model outperformed others in terms of recall, achieving 79.42% on the test set, a significant improvement from its initial results and the best among all models. This prioritisation of recall ensures the model effectively identifies potential fraud cases, making it more suitable for the imbalanced nature of vehicle insurance fraud detection. While models with high accuracy, such as Random Forest and XGBoost, struggled to achieve high recall, indicating a challenge in detecting minority class instances. These models were better at detecting legitimate claims but frequently missed fraudulent ones. The trade-off between recall and precision highlights the inherent complexity of fraud detection. While the CatBoost model flags more potential fraud cases, a higher proportion of these are false positives.

Integration of the model into the business process

1. **Claim Pre-screening**
   The model can be used to automatically analyse incoming insurance claims and assign a fraud risk score to each submission. Claims that are identified as high risk can be flagged for further investigation by specialised fraud detection teams. This pre-screening process not only reduces claims adjusters' workload, but also ensures that investigative resources are allocated more efficiently, with a focus on claims that are most likely to be fraudulent.

2. **Enhanced Fraud Detection Workflow**
   Integrating the model into existing fraud detection systems enables insurers to supplement traditional rule-based approaches with advanced, data-driven insights. The model's ability to detect subtle patterns and anomalies, such as repeated claims or inconsistencies in reporting timelines, significantly improves detection accuracy. This integration allows for a more comprehensive and robust fraud detection workflow, which enhances the system's ability to detect sophisticated or evolving fraudulent schemes.

3. **Real-Time Decision Support**
   The model can also be integrated into a real-time decision-making tool to help claims adjusters with initial claim evaluations. By providing instant fraud risk assessments, the model helps to identify potentially fraudulent claims early in the process. This approach increases operational efficiency and reduces delays in processing legitimate claims, which benefits both insurers and honest policyholders.

Additional Insights from Experiments

In an effort to boost the model's capacity to detect false claims, we also tried reducing the decision thresholds to enhance recall. Recall significantly increased as a result of this modification, suggesting that the model may be able to identify more possible fraud situations. However, there were significant trade-offs associated with this advancement. Precision dropped considerably, meaning a higher proportion of the flagged claims were false positives. Similarly, overall accuracy and ROC-AUC declined, reflecting a reduced ability of the model to correctly classify both legitimate and fraudulent claims and to distinguish between the two classes effectively.

Moving forward, we aim to refine the decision threshold further and explore tailored cost-sensitive strategies to optimise recall while maintaining acceptable precision levels. This approach ensures the fraud detection system aligns with the business context, effectively addressing operational inefficiencies and mitigating financial risks.

6 Limitations and Conclusion

Limitations

There are several limitations to consider in this study. A key limitation is that the dataset used for analysis is nearly 30 years old. Over this period, advancements in technology, particularly in closed-circuit television (CCTV) and the widespread use of dashboard cameras (dash cams), have transformed fraud detection and monitoring methods. These technologies provide real-time visual evidence of accidents, which can significantly simplify the process of identifying fraudulent claims. As a result, the insights gained from this study may not fully account for these modern developments and their impact on fraud detection practices.

Additionally, the data is sourced exclusively from the United States, which limits the generalisability of the findings to other regions. Driving regulations, practices, and the make and model of vehicles can vary greatly between countries, influencing the nature of fraudulent claims. For instance, countries with stricter driving laws or different insurance structures may experience fraud patterns that are not captured by the dataset used in this study. This geographical and contextual specificity must be considered when applying the results to other regions or adapting the findings to global insurance markets.

Lastly, while the model provides valuable insights, human oversight still remains critical. Fraudulent claims often involve complex and nuanced contexts that automated systems may not fully capture. Adjusters and fraud teams must validate the model’s outputs to ensure accurate and fair claim assessments.

Conclusion

The experiments demonstrate how machine learning, and specifically the CatBoost model, can identify fraudulent insurance claims with a high recall score, highlighting its ability to correctly identify a significant proportion of fraudulent cases. The technology can significantly improve fraud detection by utilising sophisticated modelling approaches, giving insurers a powerful tool for reducing losses and improving operational efficiency. The challenges posed by class imbalance, as well as the trade-offs between recall and precision, highlight the problem's complexities and the need to prioritise detecting fraudulent claims over overall accuracy.

Nonetheless, the study provided valuable insights into which types of predictive models might be most effective for detecting vehicle insurance fraud. Despite the evolution of technology, the nature of data collected for insurance claims has remained largely consistent from the 1990s to the modern day. This means that the findings from this study are still applicable to contemporary insurance claims, enabling insurers to identify similar patterns when detecting fraudulent activities.

To further improve the system, future efforts should focus on exploring more sophisticated resampling methods, cost-sensitive learning, and feature engineering to address the limitations in detecting minority class instances. Additionally, integrating external data sources and employing ensemble models could enhance the model's ability to identify subtle fraud patterns.

Ultimately, this fraud detection system holds significant promise for reducing the economic and social impact of insurance fraud. By identifying fraudulent claims more effectively, insurers can minimise losses, lower premiums for honest policyholders, and contribute to a fairer and more efficient insurance ecosystem.

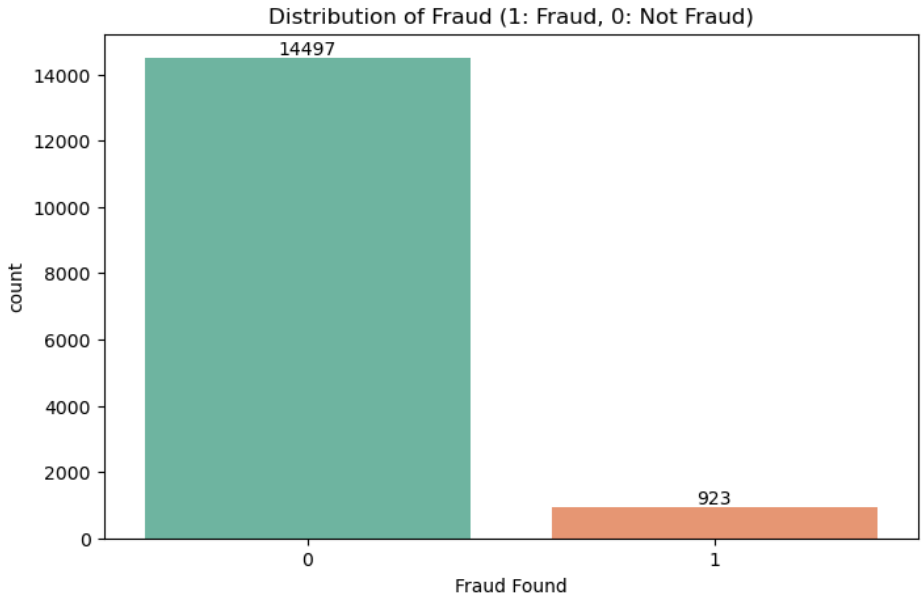
7 References

Araullo, K. (2023, November 7). One in five motor claims in Singapore fraudulent – GIA. Insurance Business Asia. https://www.insurancebusinessmag.com/asia/news/auto-motor/one-in-five-motor-claims-in-singapore-fraudulent--gia-465901.aspx

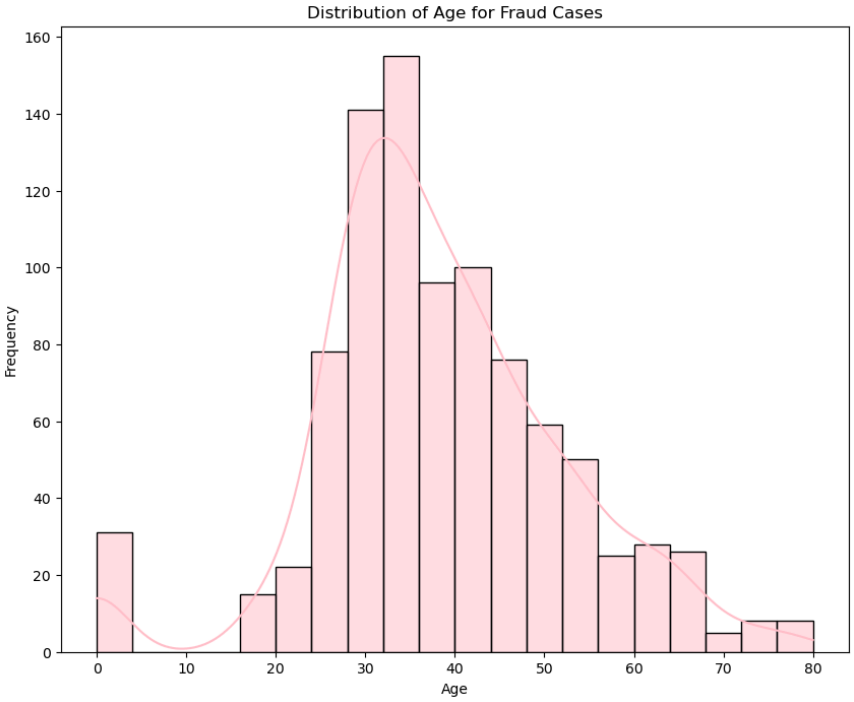
Insurance Information Institute. (2022, August 1). Background on: Insurance fraud. https://www.iii.org/article/background-on-insurance-fraud

Vehicle insurance claim Fraud Detection. (2021, December 20). Kaggle. https://www.kaggle.com/datasets/shivamb/vehicle-claim-fraud-detection

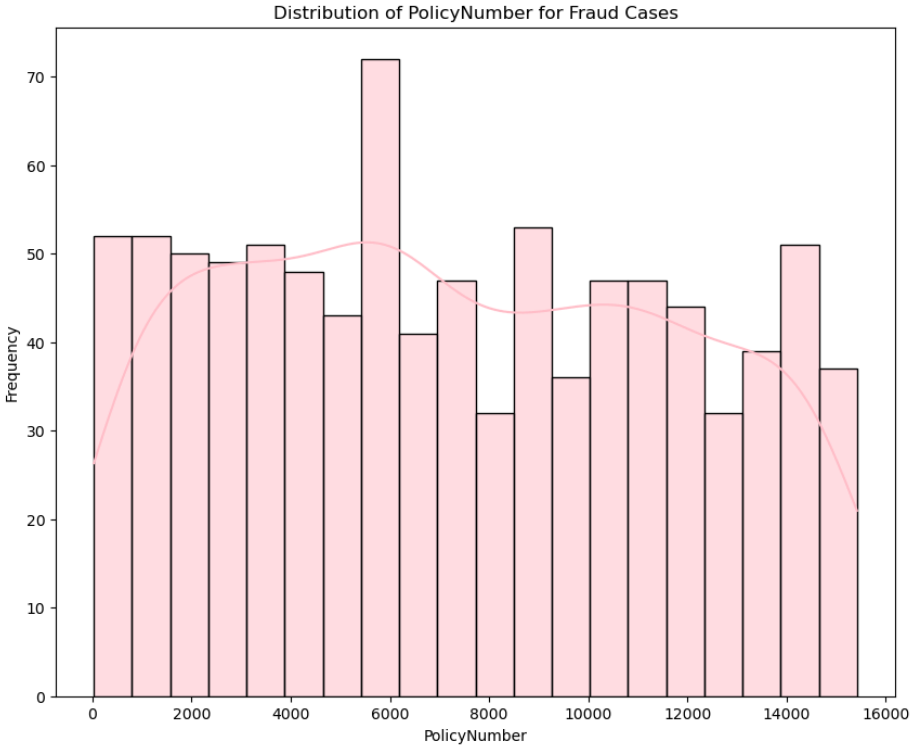
8 Appendix



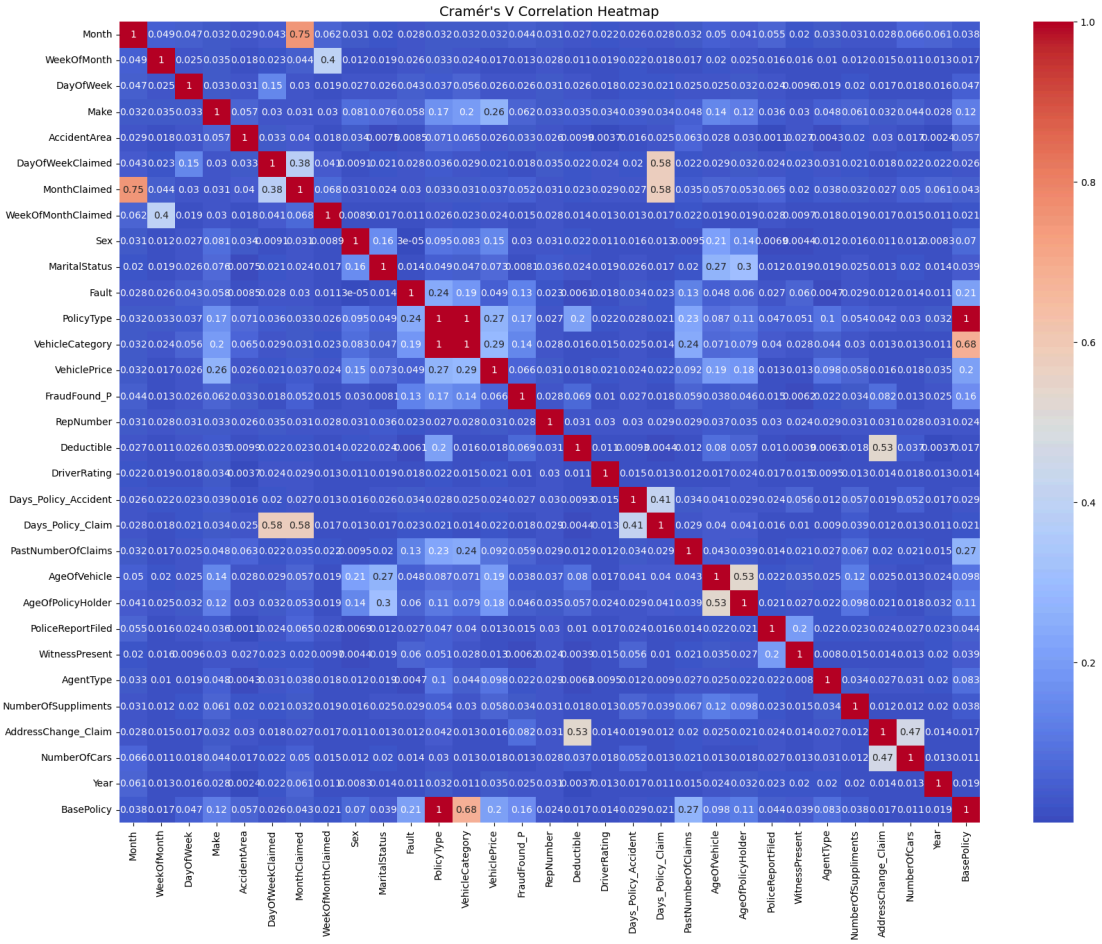
Graph 1. Distribution of Fraud Classes



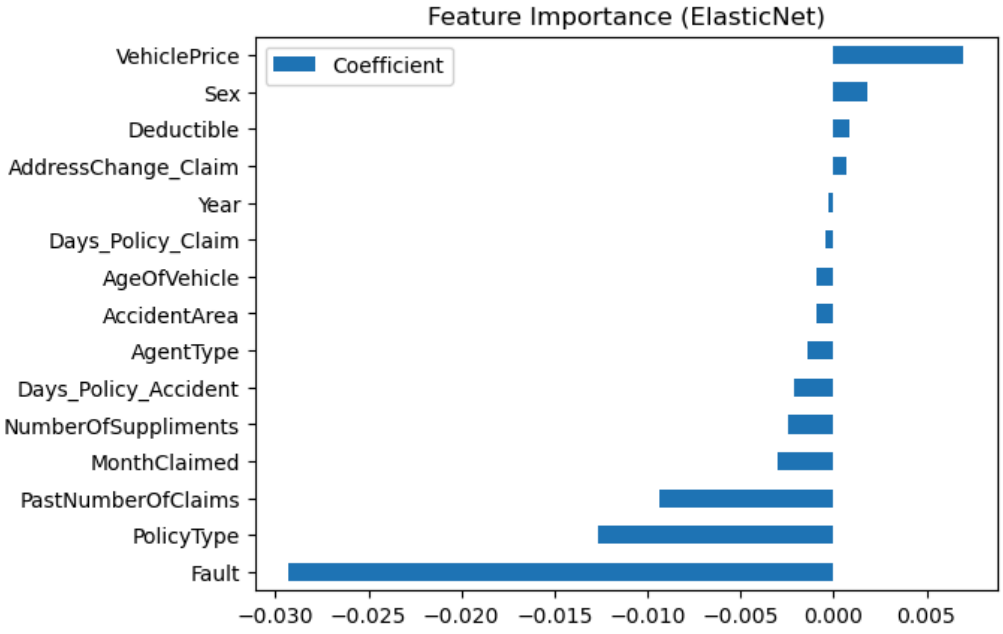
Graph 2. Distribution of Age for Fraud Cases



Graph 3. Distribution of PolicyNumber for Fraud Cases



Graph 4. Cramér's V Heatmap: Correlation Analysis Between Categorical Variables



Graph 5. Feature Importance Plot: Coefficients from ElasticNet Model