Title: Leveraging Machine Learning to Combat Insurance Fraud in the Auto Insurance Industry

Name: Adeola Oluwagbenga Odunewu

Project: Third Phase Evalution Project (Insurance Claim Fraud Detection)

Internship N0: DS2311

Introduction

Insurance fraud poses a significant challenge for the auto insurance industry, leading to substantial financial losses and increased premiums for policyholders. Traditional methods of detecting fraudulent claims often fall short due to their reliance on manual investigation and subjective judgment. However, advancements in machine learning offer promising solutions to this problem by enabling the development of predictive models that can automatically identify suspicious claims with high accuracy.

Problem Definition

The primary objective of this project is to develop a predictive model that can classify insurance claims as either fraudulent or legitimate based on the provided data. This involves creating a binary classification system where the model assigns a label of "fraudulent" or "non-fraudulent" to each insurance claim. The scope of the problem includes analyzing historical data of insurance policies, customer information, and accident details to train the predictive model. The model will then be deployed to predict the likelihood of fraud for new insurance claims as they are submitted.

Data

The dataset provided for this project contains details of insurance policies, customer demographics, accident information, and a binary target variable indicating whether the claim is fraudulent or not. Features such as policy duration, claim amount, number of previous claims, customer demographics, and type of accident are included in the dataset.

Methodology

Data Collection and Preprocessing

The dataset is download from FlipRobo Github repository and preprocessed to handle missing values, encode categorical variables, and scale numerical features if necessary. It is then split into training and testing sets.

Feature Engineering: New features are created, and feature selection techniques are applied to identify the most relevant features for detecting fraud.

Model Selection: Different machine learning algorithms such as Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machines are experimented with, and the model that provides the best performance is selected.

Model Training: The selected model is trained on the training data.
Model Evaluation: The performance of the model is evaluated using appropriate classification metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).

Hyperparameter Tuning: The hyperparameters of the selected model are fine-tuned to optimize its performance further.
Model Deployment: Once satisfied with the model's performance, it is deployed to production to automatically flag potentially fraudulent claims.

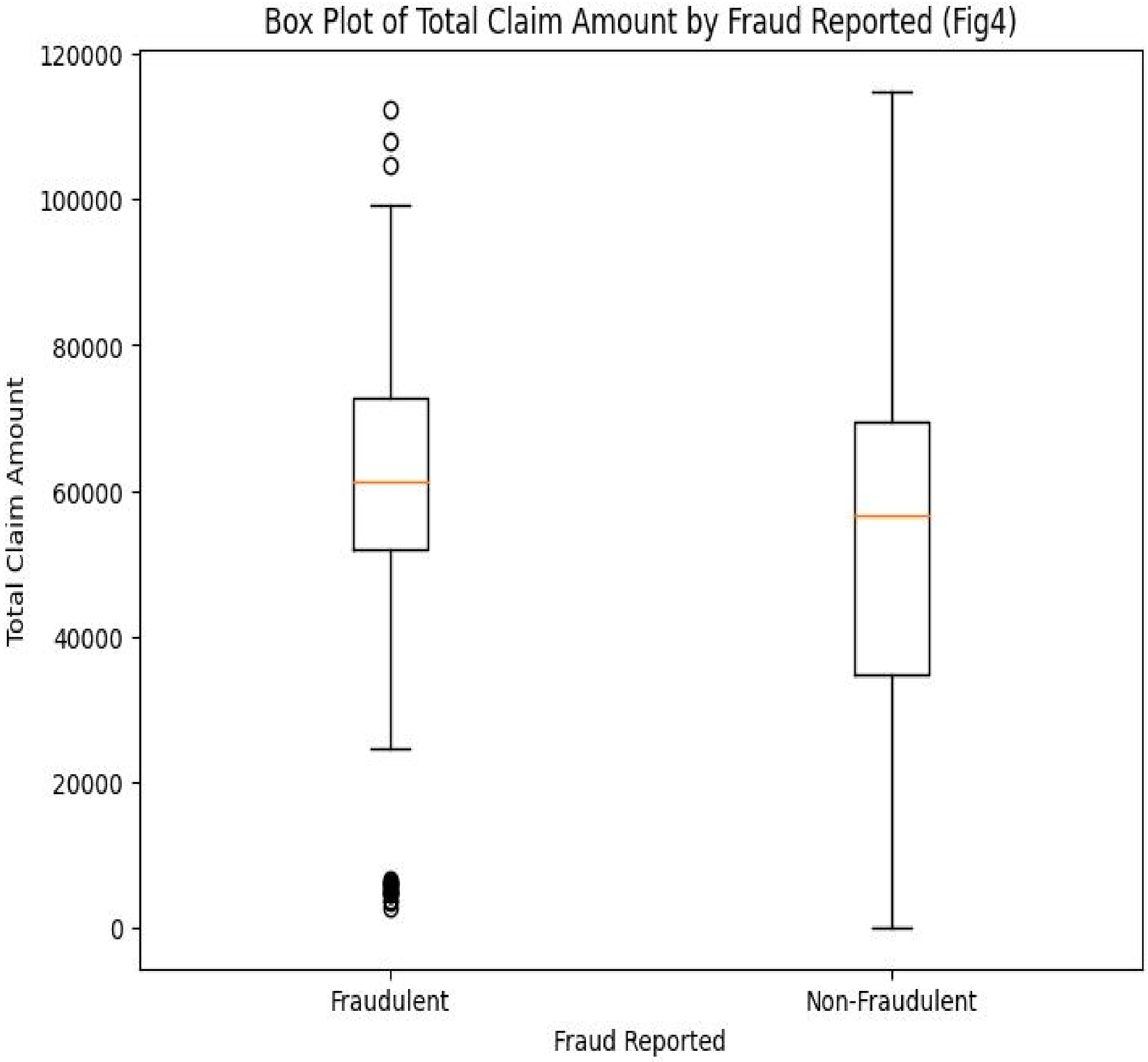
Data Analysis

Uncovering Insights to Combat Insurance Fraud

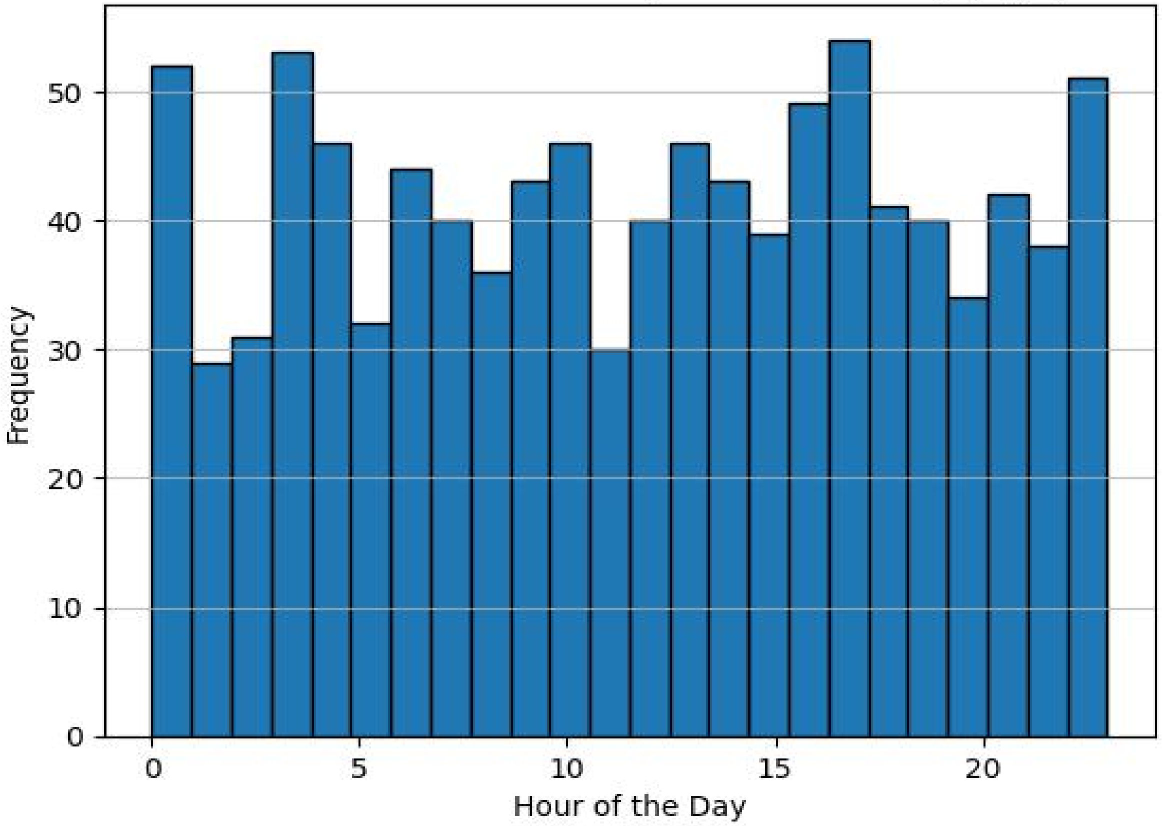
In our quest to combat insurance fraud within the auto insurance industry, thorough data analysis serves as the cornerstone. Through a multifaceted approach, we delve deep into the dataset, unraveling patterns, addressing missing values, and identifying outliers to lay the groundwork for an effective predictive model.

Understanding the Landscape

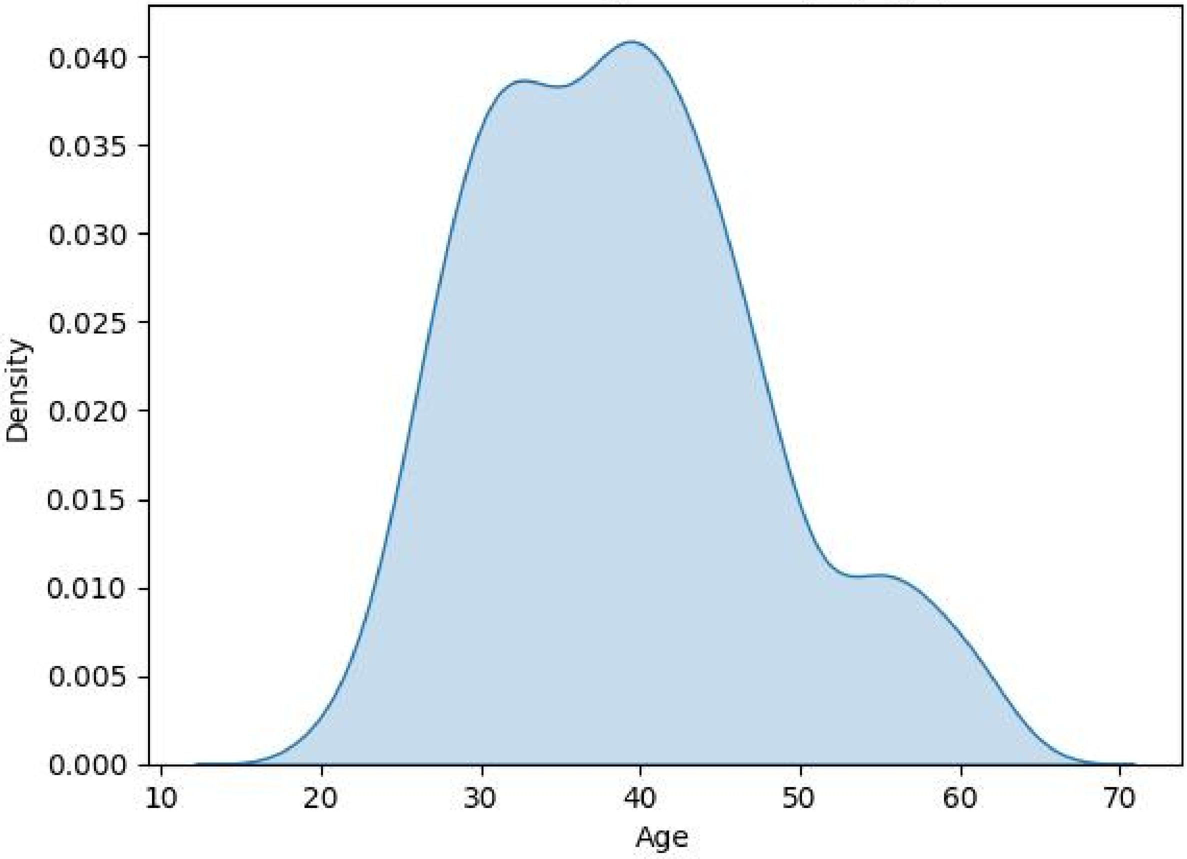
This analysis begins by dissecting individual variables. Through count plots and histograms, we gain insights into categorical features such as Total Claim Amount and Incident by Hours of the Days, shedding light on their distributions and potential significance in fraud detection. Meanwhile, kernel density plots unveil the nuanced distribution of key variables like Age, guiding our understanding of demographic trend.



Distribution of Incidents by Hour of the Day (Fig2)



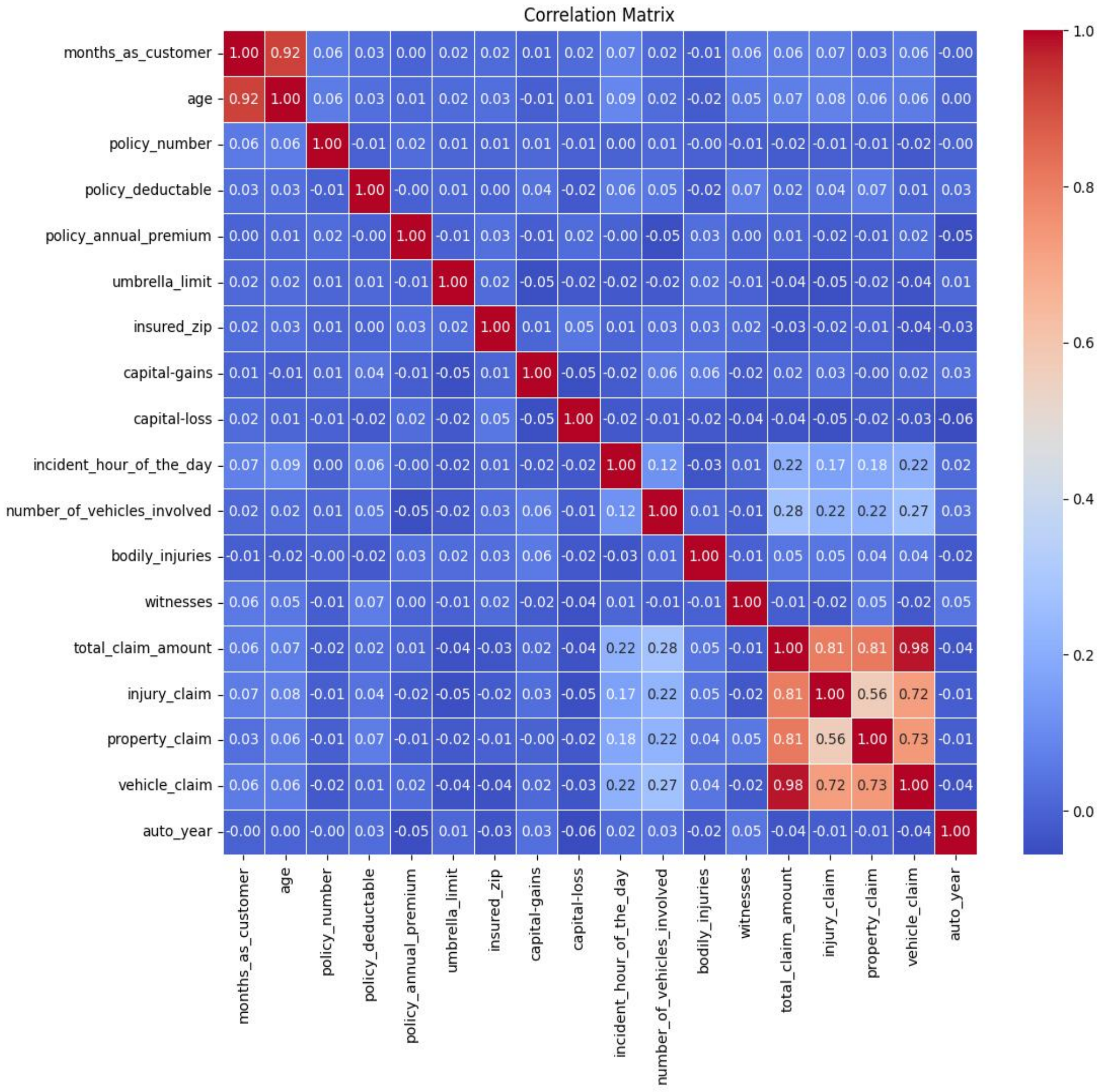
Kernel Density Plot for Age (Fig3)



Ensuring Data Integrity
Missing values pose a threat to the integrity of our analysis. By employing a pragmatic approach, we replace missing values with the mode for categorical features, ensuring that our dataset remains robust and representative of the underlying population.

Optimizing Analysis Efficiency
Efficiency is paramount in this endeavor. By converting object columns to categorical type, the study streamline memory usage and facilitate subsequent analysis. Furthermore, the transformation of the fraud_reported column to numeric values empowers our predictive model with actionable insights.

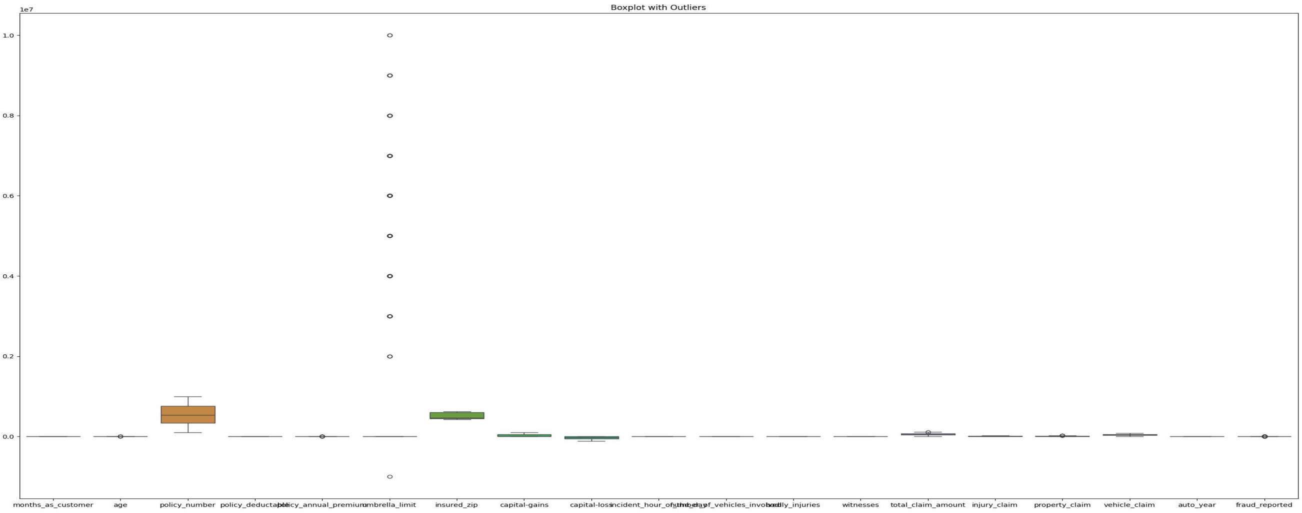
Unveiling Relationships
The interplay between variables holds valuable clues. Through correlation matrices and heatmap visualization, the study uncover nuanced relationships between numerical features, offering glimpses into potential predictors of insurance fraud. This also lays the groundwork for feature selection and model refinement.



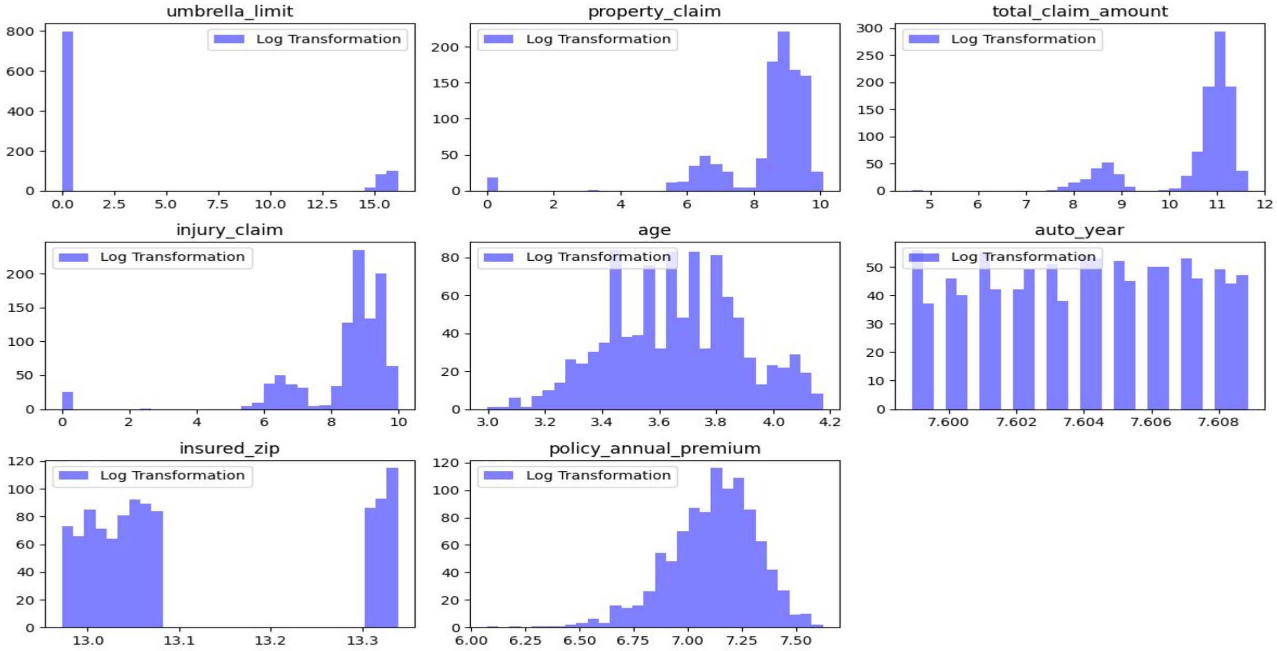
Mitigating Anomalies

Outliers, while informative, can skew our analysis. Leveraging z-scores, we identify and scrutinize data points that deviate significantly from the norm. Visualization through boxplots provides clarity, allowing us to discern patterns and anomalies within the dataset.

In the context of insurance claim fraud detection, log transformation is good particularly useful for features, and other variables that exhibit skewed distributions with outliers. By transforming these variables using the natural logarithm the resulting distributions become more symmetrical, making them more suitable for modeling purposes.



Box plot indicating Outliers



Plot Distributions After Log Transformation

In essence, this data analysis journey serves as a compass, guiding the efforts to develop a robust predictive model capable of discerning fraudulent insurance claims amidst the complexities of the auto insurance landscape.

EDA Concluding Remarks

The study exploration of the dataset, unveiled a complexity of insights that offer valuable guidance for detecting fraudulent activity within the auto insurance sector.

Total Claim Amount Distribution Analysis:
The distribution of total claim amounts presents a nuanced picture, with a slight positive skew indicating a prevalence of lower claim amounts. However, outliers, identified through box plots, beckon further investigation, as they may signify anomalous or extreme cases deserving of scrutiny.

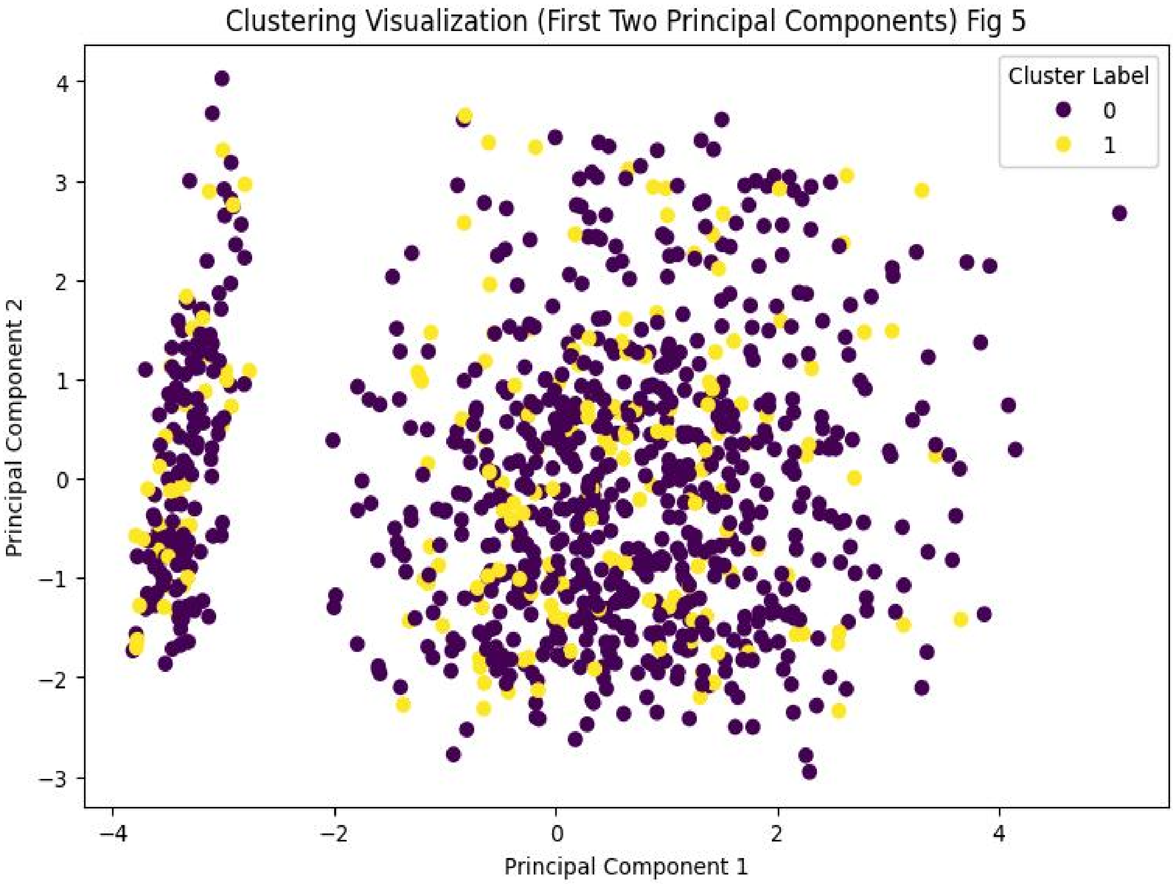
Average Incident Hour of the Day:
Notably, incidents peak around 11 hours of the day, suggesting a potential temporal window for heightened claim activity. This insight could inform resource allocation and operational strategies to manage surges in claim processing during peak hours.

Age Distribution Analysis:
The age distribution paints a revealing portrait of policyholder demographics, exhibiting a symmetric pattern with a discernible right tail. This demographic insight may offer valuable context for understanding customer behavior and tailoring insurance products and services accordingly.

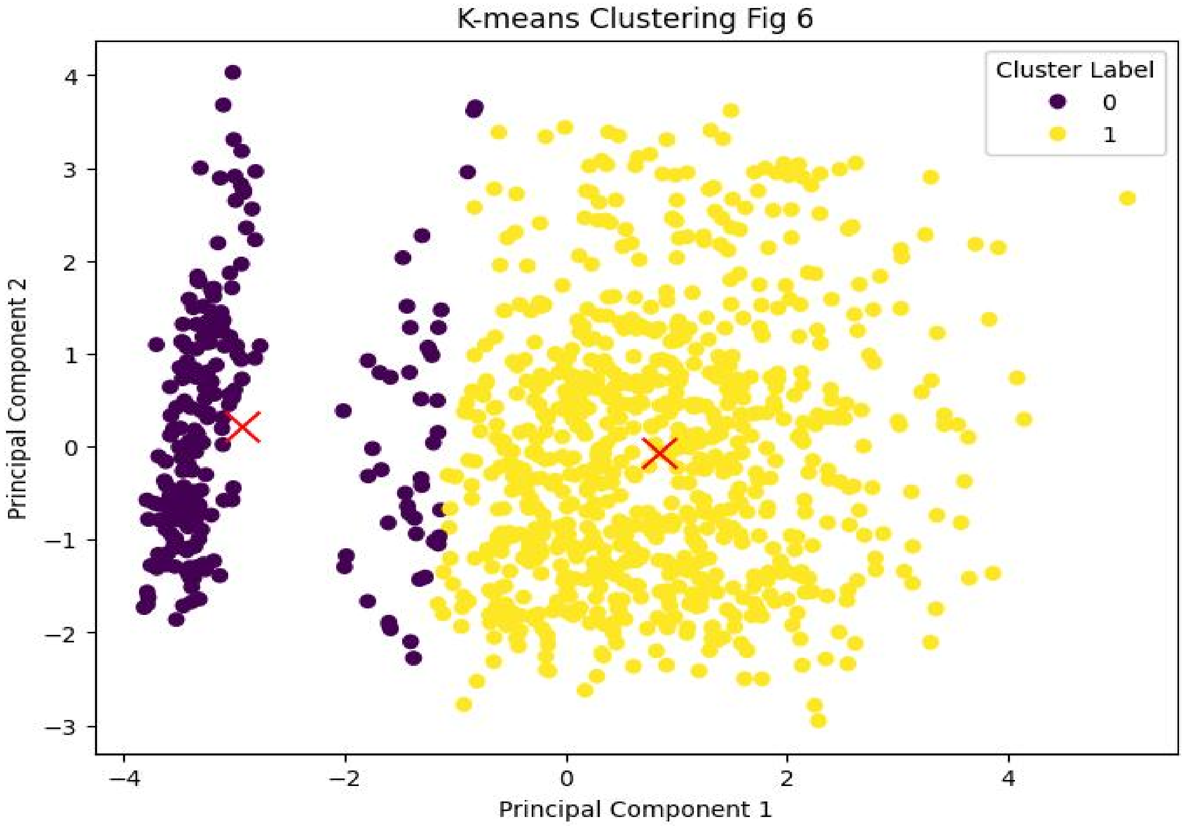
Comparison of Total Claim Amounts for Fraudulent and Non-Fraudulent Claims:
A striking finding emerges from the comparison of total claim amounts between fraudulent and non-fraudulent claims. Fraudulent claims, on average, boast higher median total claim amounts, suggesting a potential correlation between claim amount and the likelihood of fraudulence. The presence of outliers, particularly among fraudulent claims, underscores the need for vigilant scrutiny and anomaly detection to root out potential instances of fraud.

Correlation Analysis Insights:
Delving into correlations between key variables uncovers intricate relationships within the dataset. Strong positive correlations between Total Claim Amount (TCA) and various claim types hint at underlying patterns driving insurance claims. Meanwhile, the correlation between Month as Customer and age unveils intriguing dynamics surrounding customer loyalty and tenure, offering fertile ground for further exploration and hypothesis generation.

Cluster Analysis:
Cluster analysis reveals a rich tapestry of policyholder groups with diverse characteristics and behaviors. These distinct clusters, with their unique profiles and tendencies, provide valuable fodder for customer segmentation, fraud detection, and risk management strategies. By leveraging insights gleaned from cluster analysis, insurance companies can tailor their approaches to better meet the needs of different customer segments while simultaneously identifying and mitigating potential instances of fraud.



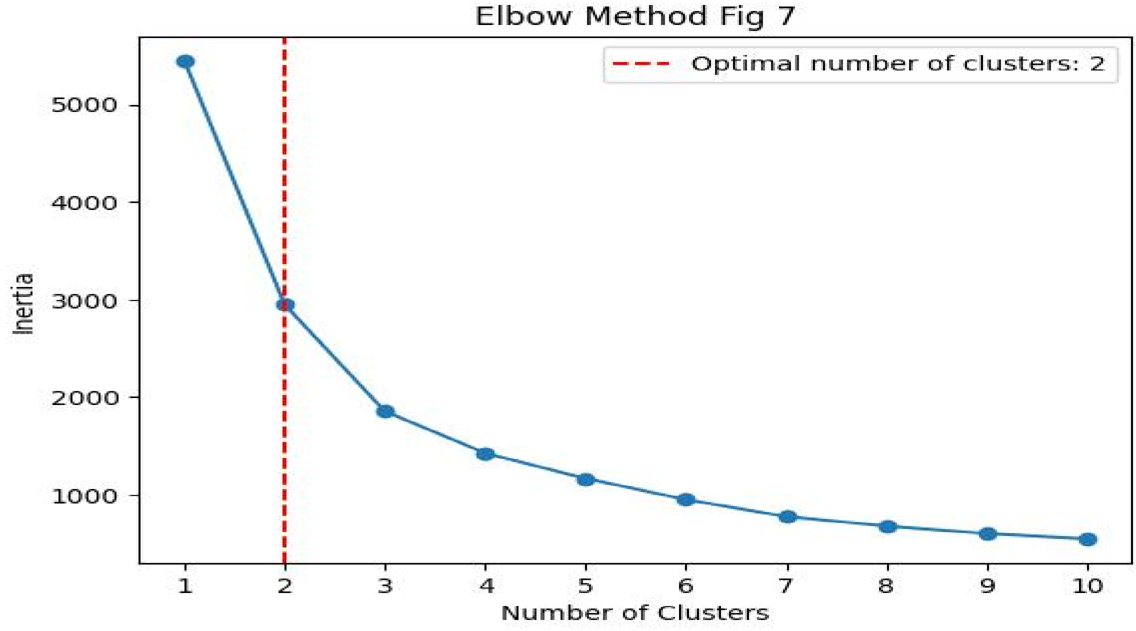
Cluster scatterplot of PCA 1&2

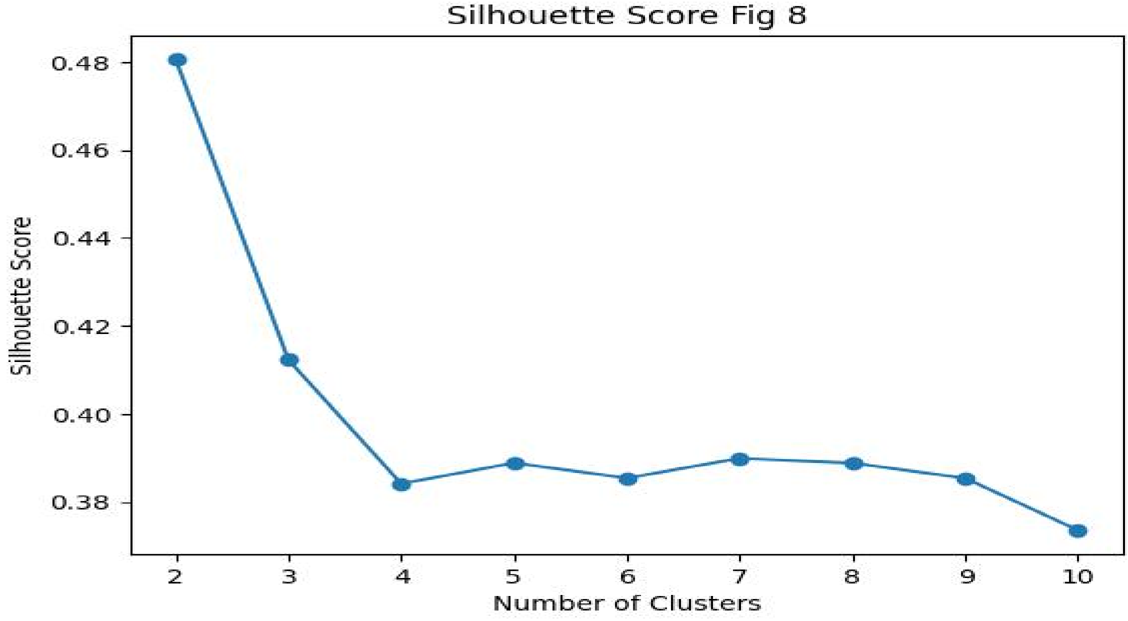


Cluster Centriod

Elbow Method and Silhouette Score:

The convergence of the elbow method and silhouette score underscores the presence of two distinct clusters within the data. This clustering validation offers a robust foundation for subsequent analysis and interpretation, empowering stakeholders to make informed decisions grounded in data-driven insights.





The journey through exploratory data analysis has illuminated the intricate landscape of insurance claims within the auto insurance industry. By leveraging these insights, insurance companies can fortify their defenses against fraudulent activity, bolster customer satisfaction, and drive sustainable business growth in an increasingly complex and dynamic marketplace.

Pre-processing Pipeline

The pre-processing pipeline is a critical step in preparing the insurance dataset for effective fraud detection. It involves addressing missing values and outliers, ensuring data integrity and reliability. New features, such as age groups and claim amount ratios, are engineered to capture nuanced patterns identified during exploratory data analysis (EDA), enhancing the predictive power of the model. Numerical features may be scaled to maintain consistency and facilitate model convergence, while categorical variables are encoded to transform them into a format suitable for machine learning algorithms.

Additionally, dimensionality reduction techniques like Principal Component Analysis (PCA) were applied to reduce the complexity of the dataset while preserving essential information. The pre-processed dataset is then split into training and testing subsets, enabling the model to learn from the training data and evaluate its performance on unseen data. This systematic approach ensures that the model is trained on clean, transformed data, ready to detect fraudulent insurance claims accurately and efficiently in the auto insurance industry.

Building Machine Learning Models

In building machine learning models claim fraud detection, the selection of variables is informed by a comprehensive analysis encompassing correlation, Variance Inflation Factor (VIF), Principal Component Analysis (PCA), and RandomForestRegressor (RFR) feature importance. These methods collectively provide insights into variable relationships, dimensionality reduction, and predictive power. Variables deemed relevant and impactful are selected for inclusion in the modeling process, enhancing the model's ability to generalize and make accurate predictions.

The dataset is split into features and the target variable, with features including numerical and one-hot encoded categorical variables. Standardization is applied to the features, and the dataset is further split into training and testing sets. To address class imbalance, Synthetic Minority Over-sampling Technique (SMOTE) is employed. Classifiers including RandomForestClassifier, XGBClassifier, GradientBoostingClassifier, LogisticRegression, and Support Vector Machine (SVC) are initialized for model training. K-fold cross-validation is performed for each classifier to evaluate accuracy.

Result:

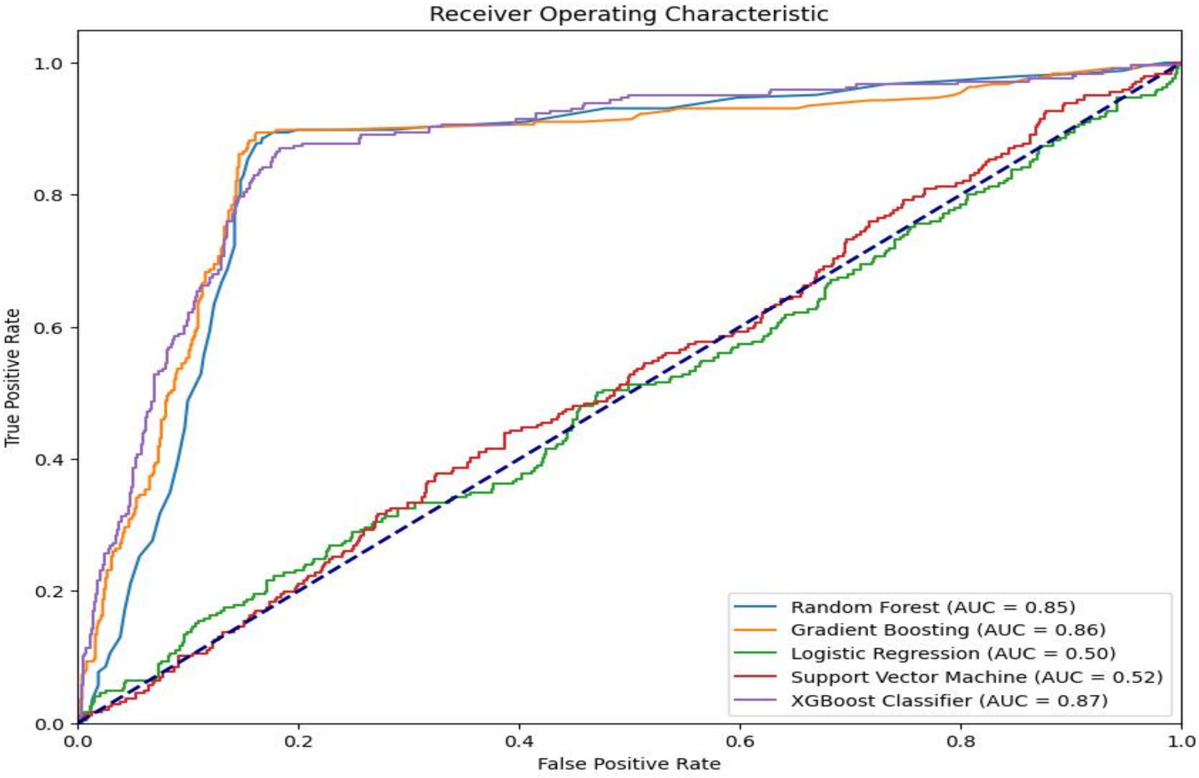
Classifier: RandomForestClassifier
Accuracy: 0.7577 +/- 0.0175

Classifier: XGBClassifier
Accuracy: 0.8319 +/- 0.0408

Classifier: GradientBoostingClassifier
Accuracy: 0.8409 +/- 0.0304

Classifier: LogisticRegression
Accuracy: 0.7538 +/- 0.0046

Classifier: SVC
Accuracy: 0.7538 +/- 0.0046



Additionally, a pipeline incorporating imputation and gradient boosting is defined, and hyperparameter tuning using GridSearchCV is conducted to optimize model performance. The best model is evaluated on the test set, and its accuracy is computed. Finally, the best model is saved using joblib for future use.

Result:
Best parameters found:
{'gb__learning_rate': 0.05, 'gb__max_bins': 150, 'gb__max_depth': 4, 'gb__min_samples_leaf': 1}

Accuracy on Test Set: 0.8350

In conclusion, a comprehensive examination of various insurance-related factors uncovers crucial insights essential for informed decision-making within the insurance sector. Through detailed analysis of claim amounts, incident timing, age demographics, and fraud indicators, the study develop a nuanced understanding of customer behavior and risk patterns. These insights, combined with advanced analytical techniques like PCA and cluster analysis, enable us to optimize customer segmentation, improve fraud detection mechanisms, and refine risk management strategies. By integrating these findings with a finely-tuned

HistGradientBoostingClassifier, we enhance predictive accuracy and facilitate data-driven excellence in insurance operations, ensuring sustainable growth and robust risk mitigation in the ever-evolving insurance landscape. Additionally, the study save the best model using joblib.dump.

Reference

Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. Computers & Electrical Engineering, 89, 106903. https://doi.org/10.1016/j.compeleceng.2020.106903

Khan, A., Ghosh, S. K., Ghosh, D., & Chattopadhyay, S. (2021). Random wheel: An algorithm for early classification of student performance with confidence. Engineering Applications of Artificial Intelligence, 102, 104270. https://doi.org/10.1016/j.engappai.2021.104270

Aslam, F., Hunjra, A. I., Ftiti, Z., Louhichi, W., & Shams, T. (2022). Insurance fraud detection: Evidence from artificial intelligence and machine learning. Research in International Business and Finance, 62, 101744. https://doi.org/10.1016/j.ribaf.2022.101744

Datatrained/Flip Robo. (2024). LCC Online tutorial material [Online tutorial]. Retrieved from https://learning.datatrained.com/myaccount/#/course/65839/lesson/965649?lesson=965649&lesson_type=material&section=249729&subject=190542