FRAUD INSURANCE CLAIM DETECTION

- Devendra Bupathi Rapolu (11639592)

Business Case

Insurance fraud detection is a significant challenge in the insurance industry. It involves identifying false claims made by policyholders or third parties. Effective fraud detection can save insurance companies considerable amounts of money and help in maintaining lower premiums for honest customers.

Problem Statement

The primary issue is the detection of fraudulent insurance claims. These claims can often be sophisticated and hard to distinguish from legitimate ones, leading to financial losses for insurance companies and higher premiums for customers.

Business Solution

Implementing an advanced fraud detection system that can accurately identify potential fraudulent claims. This system would reduce the financial burden on the insurance company and maintain trust in the insurance process. It would also streamline claim processing by quickly separating legitimate claims from fraudulent ones.

Technical Solution

Developing a machine learning model to detect fraudulent insurance claims. The solution involves preprocessing the dataset, handling missing values and outliers, transforming data, and encoding categorical features. Exploratory Data Analysis (EDA) is performed to understand the data better, followed by the application of machine learning models such as SVM, Random Forest, and a Voting Classifier for prediction.

Dataset Documentation: Insurance Claims Data

Overview

The dataset contains information regarding insurance claims, detailing aspects of policyholders' profiles, insurance policy details, claim specifics, and indications of whether a claim was fraudulent.

Data Source

This dataset is sourced from Mendeley Data, specifically from the dataset collection available at https://data.mendeley.com/datasets/992mh7dk9y/2.

Data Dictionary

The dataset consists of various columns which include:

- **months_as_customer:** The number of months the claimant has been a customer with the insurance company. This is a continuous variable indicating the length of the customer relationship.
- **age:** The age of the policyholder. This is a continuous variable representing the policyholder's age at the time of the claim.
- **policy_number:** A unique identifier for each insurance policy. This is a nominal variable used for identification purposes.
- **policy_bind_date:** The date when the insurance policy was issued/bound. It's a datetime variable that could be important for time-series analysis or understanding policy duration.
- **policy_state:** The U.S. state where the policy is registered. This is a categorical variable that can indicate regional patterns in insurance claims.
- **policy_csl:** Combined Single Limit (CSL) on the policy, indicating the maximum payout for a single claim. It's a categorical variable and can be important for understanding the coverage level.
- **policy_deductable:** The deductible amount of the policy. This is a continuous variable, typically impacting the claimant's decision to file a claim.
- **policy_annual_premium:** The annual premium amount for the policy. This is a continuous variable, reflecting the cost of the policy to the customer.
- **umbrella_limit:** The umbrella limit of the policy, if any. This is a continuous variable, usually indicating additional coverage beyond the basic policy.
- **insured_zip:** The ZIP code of the insured. This is a nominal variable, useful for geographical analysis of claims.
- **[Additional Columns]:** These may include details about the claim, the insured vehicle, and the fraud status, such as:
    1. **incident_type, collision_type, incident_severity:** Categorical variables describing the nature and severity of the incident.
    2. **authorities_contacted, incident_state, incident_city:** Categorical variables providing context about the incident's location and response.

3. **number_of_vehicles_involved, property_damage, bodily_injuries:** Continuous or ordinal variables detailing the extent of the incident.
4. **witnesses, police_report_available:** Variables indicating the availability of additional evidence or testimonies.
5. **total_claim_amount, injury_claim, property_claim, vehicle_claim:** Continuous variables detailing the financial aspects of the claim.
6. **auto_make, auto_model, auto_year**: Categorical and continuous variables providing details about the insured vehicle.
7. **fraud_reported:** A categorical variable indicating whether the claim was fraudulent.

Code Documentation

Data Preprocessing: Handling missing values, outliers, and transforming date columns.
EDA: Visualizations like pair plot, histograms, count plots, and heatmaps to understand data distributions and relationships.
Feature Engineering: One-hot encoding for categorical data.
Machine Learning Implementation

- **Models:** SVM, Random Forest, and Voting Classifier.
- **Class Imbalance Handling:** Using SMOTE.
- **Training and Testing:** Splitting data into training and test sets.
- **Evaluation:** Precision, recall, f1-score, ROC-AUC, and confusion matrices.

Conclusion

The implementation of the machine learning models demonstrates promising results in detecting fraudulent insurance claims. The Voting Classifier, combining SVM and Random Forest, provides the best performance, indicating the effectiveness of ensemble techniques in such scenarios.

Future Work

Feature Importance Analysis: To identify which features most significantly impact fraud detection.
Hyperparameter Tuning: To optimize model performance.
Cross-Validation: To ensure model robustness across various data subsets.
Model Interpretability: Implementing models that offer better interpretability for understanding the decision-making process.
Real-Time Integration: Consider deploying the model into a real-time system for immediate fraud detection as claims are processed.
Regular Updates and Training: Continuously updating the model with new data to maintain its accuracy over time.