Technical Career Education Private Limited

5th floor, Sahyadri Campus, Adyar, Mangalore 575007

Innoventure Internship

PROJECT REPORT

2022 - 23

Project Title: Protecting systems and data from internal or external threats

Submitted by:

Md Mainuddin	4SF21IS057
Keerthan S Suvarna	4SF21CS070
Shreenidhi D K	4SF21CS152
Mahaan N Bhat	4SF21CD015
Samarth S Rao	4SF21CS137
Rahul Rohith Kottary	4SF21IS070

Institution:

Sahyadri College of Engineering and Management

Adyar Mangalore 575007

CONTENTS

Project Overview

Problem Statement	The problem statement is to develop a predictive models for vehicle insurance and fraud detection			
Solution Proposed *(video Link)*	The proposed solution involves data-centric approach by utilising the dataset, risk mitigation by understanding the attributes and predicting modeling for early fraud detection and intervention.			
Link to the final Challenge presentation	cyds.pptx			
Link to photos/ videos drive				
Github Link				
Team Name				
Team Members	Name	USN	Class/Section	College Internship Report
	Md Mainuddin	4SF21IS057	ISE/B	
	Keerthan S Suvarna	4SF21CS070	CSE/B	
	Shreenidhi D K	4SF21CS152	CSE/C	
	Mahaan N Bhat	4SF21CD015	CSE-DS	
	Samarth S Rao	4SF21CS137	CSE/A	
	Rahul Rohith Kottary	4SF21IS070	ISE/B	

1. Introduction

In recent years, the insurance industry has witnessed a significant rise in fraudulent activities, particularly in the realm of vehicle insurance. Fraudulent claims not only incur substantial financial losses for insurance companies but also disrupt the overall market stability, leading to higher premiums for honest policyholders. As a result, the need for robust fraud detection mechanisms has become paramount to safeguard the integrity of the insurance sector.

The advent of advanced data analysis techniques, coupled with the proliferation of digital technologies, has empowered insurance companies to tackle fraud more effectively. In this context, the focus of this study is on vehicle insurance fraud detection, employing a comprehensive data analysis approach. By leveraging sophisticated analytical tools and machine learning algorithms, insurers can sift through vast datasets to identify patterns, anomalies, and trends indicative of fraudulent activities.

Significance of Vehicle Insurance Fraud Detection:

Vehicle insurance fraud comes in various forms, ranging from staged accidents and false injury claims to fictitious policies and identity theft. These fraudulent activities not only lead to financial losses but also pose risks to public safety. Staged accidents, for instance, can result in severe injuries or even fatalities, making it crucial for insurers to identify and prevent such incidents promptly. Additionally, insurance fraud raises premiums for law-abiding citizens, creating an additional burden on honest policyholders.

2. Problem Statement

Development of predictive models for vehicle insurance and fraud detection

Data-Centric Approach: Utilizing a dataset of vehicle insurance claims allows for a data-driven analysis, enabling a comprehensive examination of the diverse characteristics and factors associated with both legitimate and fraudulent claims.

Risk Mitigation through Understanding: Gaining insights into these attributes is pivotal as it aids in recognizing patterns and behaviors indicative of potential fraudulent activities. This comprehension serves as a foundation to proactively mitigate risks associated with fraudulent claims.

Predictive Modeling for Fraud Detection: Leveraging the knowledge acquired from the dataset facilitates the development of sophisticated predictive models. These models play a pivotal role in identifying potentially fraudulent claims, offering insurance companies a powerful tool for early detection and intervention, ultimately reducing financial losses and maintaining the integrity of the insurance system.

3. Solution

1. Data Preparation

- Data Loading and Exploration: Using Pandas to load the dataset and understand its structure and unique values within each column.
- Data Cleaning: Dropping irrelevant columns, handling missing values, and filtering based on specific conditions.
- Balancing Classes: Oversampling the minority class to match the majority class instances.
- Encoding Categorical Variables: Transformation of categorical variables into a numeric format using label encoding.

2. Model Training and Evaluation

- Splitting the Dataset: Division into training and testing sets for model training and evaluation.
- KNeighborsClassifier: Utilizing the KNeighborsClassifier for predictive modeling, calculating accuracy, AUC score, confusion matrix, and classification report.
- Feature Importance: Evaluating feature importance using entropy, information gain, and other metrics for classification.

3. Implementing XGBoost Classifier

- XGBoost Training: Training an XGBoost classifier on the dataset.
- Model Evaluation: Calculating accuracy and generating a classification report for the XGBoost model.

4. Interpreting and Analyzing Results

- Metrics Evaluation: Analyzing accuracy, precision, recall, F1-score, and other metrics for the trained models.
- Feature Importance Analysis: Utilizing information gain to understand feature importance in the classification process.

5. Error Handling and Recommendations

- Addressing error handling strategies and recommending further analysis, model enhancement, and continuous monitoring of model performance.
- Suggesting the potential advantages of utilizing XGBoost over simpler models like KNeighborsClassifier.
- Emphasizing the importance of ongoing assessment of model performance considering the problem nature and business requirements.

Key Data Insights and Preprocessing Analysis

1. Dependent Variable - Sampling the dependent variable of interest in this analysis is 'FraudFound_P,' which represents the presence or absence of fraud in insurance claims. To investigate this variable, we can perform counts and sampling to understand its distribution and characteristics.

2. Value Counts - Each Variable Value counts have been performed for each attribute in the dataset to gain insights into the frequency of different categories within each attribute. This analysis reveals the distribution of data across attributes.

3. Data Types - Data types have been examined for all attributes, and it was found that the dataset contains a combination of categorical and numeric data.

4. Unique Values for Each Attribute - The unique values for each attribute have been analyzed. In this analysis, some attributes contain '0' or 'None' values, such as '0' for 'DayOfWeekClaimed' and 'MonthClaimed,' and 'None' for 'Days_Policy_Accident' and 'Days_Policy_Claim.' Further investigation is required to understand the nature of these values, and appropriate actions, such as imputation or data cleaning, should be taken as necessary.

5. Unique Attributes - The dataset comprises various unique attributes, including 'Month,' 'Make,' 'AccidentArea,' 'Sex,' 'MaritalStatus,' 'Age,' 'Fault,' 'PolicyType,' 'VehicleCategory,' 'VehiclePrice,' 'PolicyNumber,' 'RepNumber,' 'Deductible,' 'DriverRating,' 'Days_Policy_Accident,' 'Days_Policy_Claim,' 'PastNumberOfClaims,' 'AgeOfVehicle,' 'AgeOfPolicyHolder,' 'PoliceReportFiled,' 'WitnessPresent,' 'AgentType,' 'NumberOfSuppliments,' 'AddressChange_Claim,' 'NumberOfCars,' 'Year,' and 'BasePolicy.'

6. Independent Variables - There are a total of 32 independent variables in the dataset, including both categorical and numeric attributes. These variables are used to predict the dependent variable 'FraudFound_P.'

7. Filter Based on None Values Attributes with 'none' values – None values such as 'Days_Policy_Accident,' 'PastNumberOfClaims,' and 'NumberOfSuppliments,' may require filtering or special handling during analysis to ensure that these values do not interfere with modeling and analysis.

8. Outliers - Age The 'Age' attribute contains numeric data, and potential outliers were observed in the data. It is essential to consider how these outliers might affect the analysis and whether outlier treatment or transformation is needed.

9.Categorical Data - Conversion Categorical data transformation may be required for several attributes, including one-hot encoding, label encoding, or other methods, to prepare the data for modeling and analysis. Eg: Month, DayOfWeek, Make, NumberOfCars.

10.Categorical Total Data - 25 Out of the 32 independent variables, 25 are categorical, requiring specific techniques for handling categorical data, such as encoding, to make them suitable for modeling.

In summary, the analysis of the 'vehicle' dataset revealed several key aspects related to data types, unique values, independent variables, handling of '0' or 'None' values, and categorical data transformation. These insights serve as a foundation for further exploratory data analysis (EDA) and the development of predictive models aimed at addressing the problem statement of identifying fraudulent insurance claims.

Correlation and Bivariate Analysis

1. Correlation (Corr):

• In the provided analysis, correlation is not applicable for categorical variables. Correlation typically measures the strength and direction of the linear relationship between two continuous variables. However, the analysis focuses on categorical variables, which do not have a linear relationship.

2. Bivariate Analysis:

• The bivariate analysis involves examining how each categorical independent variable relates to the binary dependent variable 'FraudFound_P.'

a. Boxplot:

• Boxplots provide a visual representation of how the distribution of each category within a categorical variable differs for fraud (1) and non-fraud (0) cases.

• Significant differences in boxplots could indicate a relationship between the categorical variable and fraud occurrence.

b. t-Test:

• The t-test assesses whether there is a significant difference in the means of each category within a categorical variable between fraud and non-fraud cases.

• Non-significant t-tests suggest that means are not significantly different, while significant t-tests indicate a potential relationship.

c. Crosstab:

• Cross tabulations provide counts and percentages of cases for each combination of a categorical variable and 'FraudFound_P.'

• The chi-square test assesses the association between a categorical variable and the binary 'FraudFound_P.'

• Significant p-values in chi-square tests indicate a relationship between the categorical variable and fraud occurrence.

d. Chi-Square Test:

• Chi-square tests specifically evaluate the association between categorical independent variables and the binary dependent variable 'FraudFound_P.'

• Significant chi-square test results suggest that a categorical variable is related to the occurrence of fraud.

Data Analysis and Model Performance

Chi-Square Analysis and Entropy Calculation

- Chi-Square Analysis: Identifying significant relationships between variables and 'FraudFound_P'.
- Entropy and Information Gain: Calculating the importance of categorical variables in predicting fraud.

Model Performance:

- K-Nearest Neighbors (KNN):
- Accuracy: 81.79%

XGBoost Model:

- Accuracy: 82.86%
- Sensitivity (True Positive Rate): 0.96
- Specificity (True Negative Rate): 0.73
- Precision: 0.78

Support Vector Machine (SVM):

- Accuracy: 73.36%
- Sensitivity (True Positive Rate): 0.86
- Specificity (True Negative Rate): 0.61
- Precision: 0.69

4. Conclusion/Outcome

In conclusion, the comprehensive analysis of the vehicle insurance fraud detection dataset provided valuable insights into the characteristics and factors associated with insurance claims and fraud presence. Through extensive data preparation, including class balancing and categorical variable encoding, we effectively prepared the data for modeling. The utilization of machine learning models, such as K-Nearest Neighbors (KNN), XGBoost, and Support Vector Machine (SVM), enabled us to predict fraud occurrences with varying degrees of accuracy and precision.

The results demonstrated that the XGBoost model outperformed other models, achieving an accuracy of 82.86% and a high sensitivity of 0.96, indicating its capability to correctly identify most fraudulent claims. Additionally, the chi-square analysis highlighted significant relationships between specific variables and fraud occurrence, providing valuable insights into potential fraud indicators.

However, it's crucial to acknowledge that further refinement and continuous monitoring are necessary to enhance model performance. Recommendations include exploring hyperparameter tuning, experimenting with additional models like Random Forest or Gradient Boosting, and considering business-specific features for a more precise fraud detection system.

This analysis lays the groundwork for an effective fraud detection framework, emphasizing the importance of ongoing evaluation and adaptation to evolving fraud patterns. By leveraging advanced analytics and machine learning techniques, insurance companies can proactively identify fraudulent claims, thereby minimizing financial losses and ensuring a more secure insurance environment for all stakeholders.

5. References

Websites:

https://www.javatpoint.com/machine-learning-models

https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/