Financial Statistics & Risk Management

Auto Insurance Fraud Detection

Jayanth Dasamantharao

Abstract

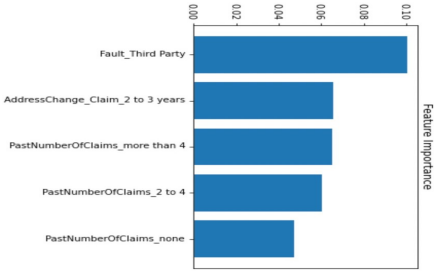
The project aims to develop a machine learning model for identifying instances of auto insurance fraud. This task is challenging as fraudulent claims are rare, but an accurate model can significantly reduce losses for insurance companies. The study focuses on classifying fraudulent claims without knowledge of the specific type of fraud. The classifiers used in this project are Random Forest, KNN, logistic regression, SVM, Decision Trees.

Problem

Insurance fraud is one of the major issues that causes significant financial losses for both insurance companies and the general public. This includes auto insurance fraud as well, which can result in higher premiums for honest drivers. Detecting and preventing this can result in insurance companies lowering their costs, while saving their customers money. Data analysis and machine learning techniques have proven effective in addressing such issues in automated systems.

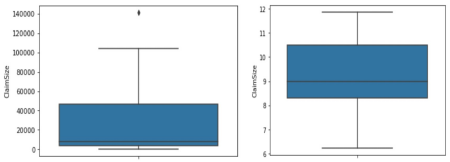
Data

The dataset consists of 33 attributes with more than 11,000 instances. The dataset is taken from Kaggle.com, which is an open-source website. There are 31 categorical and 2 numeric variables. The categorical variables are converted to binary type columns using One-Hot Encoding technique, resulting in 127 predictors. The target variable is FraudFound. The train and test data is split by 75% and 25% respectively.



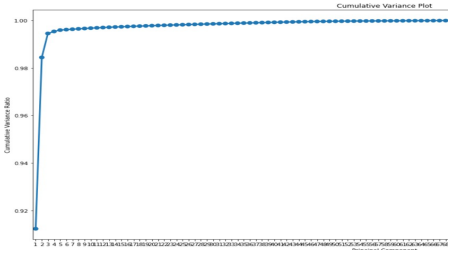
Data Exploration and Preprocessing:

1. Ten columns with a correlation coefficient greater than 0.8 were identified and removed from the dataset.
2. Log transformation was used to address outliers in the data.
3. Rows containing missing data were removed from the dataset, as they were a small fraction of the total number of rows.
4. Categorical variables were transformed into numerical variables using One-Hot Encoding.
5. Standardization was performed by subtracting the mean and scaling the data to have a unit variance, to ensure that all features had the same scale.



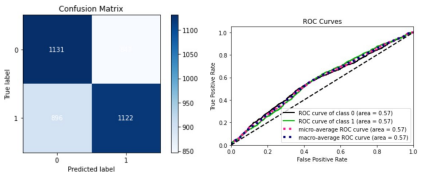
Findings and Evaluation:

1. The imbalanced data issue was addressed by applying the "SMOTE" re-sampling technique.
2. Principal Component Analysis (PCA) was utilized to decrease the number of features from 132 to 7.
3. A Grid Search approach was implemented to identify the optimal hyperparameters for each model.
4. Confusion matrices, ROC curves, and F1-scores were graphed for each model to evaluate their performance.



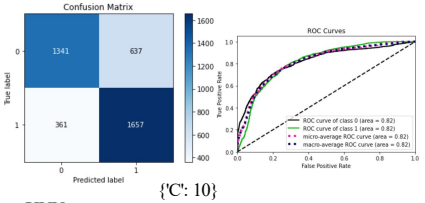
Methodology:

Logistic Regression



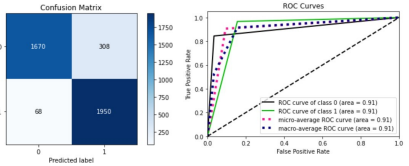
{'C': 0.01, 'penalty': 'l2'}

SVM



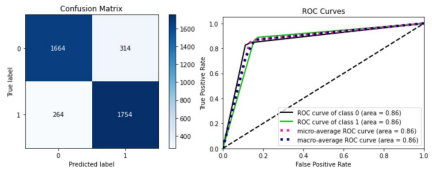
{'C': 10}

KNN



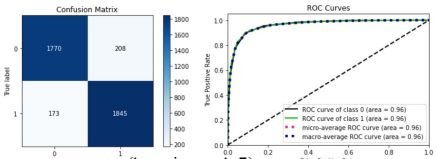
{'n_neighbors': 1}

Decision Tree



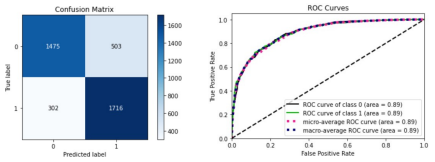
{'criterion': 'entropy', 'min_samples_leaf': 1, 'min_samples_split': 4, 'splitter': 'best'}

Random Forest



{'n_estimators': 7}

XG Boost



{'max_depth': 7, 'max_leaves': 1, 'n_estimators': 7}

Model	F1 - score (Target – 0)	F1 - score (Target – 1)
Logistic Regression	0.56	0.56
SVM	0.73	0.77
KNN	0.90	0.91
Decision Tree	0.85	0.86
Random Forest	0.90	0.91
XG Boost	0.79	0.81

Conclusion

The data was balanced to improve recall, which had the expected effect, but the resulting recall levels were still insufficient for real-world scenarios. To improve recall further, multiple models were tested on the balanced dataset, resulting in the Random Forest and KNN algorithms achieving an f1-score of 91%. KNN had a recall rate of 97% and an AUC of 91%, while Random Forest had a recall rate of 91% and an AUC of 96%. Since minimizing false negatives is critical in insurance fraud detection to avoid significant financial losses for the insurer, we have decided to select KNN as our final model, given its higher recall rate.