Insurance Claim Fraud Detection

Author Kiruthika N

Date 13-Sep-2024

Batch No DS2403

Abstract

Insurance Company working as commercial enterprise from last few years have been experiencing fraud cases for all type of claims. Amount claimed by fraudulent is significantly huge that may causes serious problems, hence along with government, different organization also working to detect and reduce such activities. Such frauds occurred in all areas of insurance claim with high severity such as insurance claimed towards auto sector is fraud that widely claimed and prominent type, which can be done by fake accident claim. So, we aim to develop a project that work on insurance claim data set to detect fraud and fake claims amount. The project implement machine learning algorithms to build model to label and classify claim. Also, to study comparative study of all machine learning algorithms used for classification using confusion matrix in term soft accuracy, precision, recall etc. For fraudulent transaction validation, machine learning model is built using Python Library.

Introduction

With the growth of the population of automobiles, auto insurance has progressively become an essential sector linked to global economic growth and people's lives. It is primarily responsible for covering the cost of damages due to natural catastrophes and vehicle accidents, which includes auto insurance and third-party motor car liability insurance. With more trust in the positive growth of the insurance sector, more capital will join the insurance market, thus making the competition between insurance firms very intense. Therefore, insurance firms concentrate on lowering costs and keeping a lead over competitors. But insurance fraud accounts for a significant proportion of insurance expenses. Insurance fraud not only lowers earnings in the insurance business, leading to substantial losses but also impacts the price strategy and social-economic advantages of the insurance firm in the long run.

The main objective of this project is to build a predictive machine learning model for the detection of fraudulent activity in the insurance company using the most influential features from the chosen dataset. We detect the insurance fraud claim by using ten robust machine learning models, which include logistic regression, decision tree classification, random forest classification, Kneighbors classification, support vector machine, naïve Bayes, ada boost classification, gradient boosting classification, bagging classification and extra tree classification models. Finally, each algorithm's performance is examined based on the six evaluation measurements (accuracy, recall, precision, specificity, sensitivity, and F-1 score) using the confusion matrix. The Gradient Boosting Classification model performs better in prediction of Insurance fraud. The results of this study are helpful for the insurance industry, investors, and policyholders to obtain a better understanding of fraudulent claims and for detecting it using a predictive model. The model is effective since it is less time-consuming and demonstrates a higher accuracy rate, which is always desired by the three key stakeholders of an insurance claim.

# Problem Definition

In this project, you are provided a dataset which has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made.

In this example, you will be working with some auto insurance data to demonstrate how you can create a predictive model that predicts if an insurance claim is fraudulent or not.

## Independent Variables

1. months\_as\_customer: Number of months of patronage

2. age: the length of time a customer has lived or a thing has existed

3. policy\_number: It is a unique id given to the customer, to track the subscription status and other details of customer

4. policy\_bind\_date:date which document that is given to customer after we accept your proposal for insurance

5. policy\_state: This identifies who is the insured, what risks or property are covered, the policy limits, and the policy period

6. policy\_csl: is basically Combined Single Limit

7. policy\_deductable: the amount of money that a customer is responsible for paying toward an insured loss

8. policy\_annual\_premium: This means the amount of Regular Premium payable by the Policyholder in a Policy Year

9. umbrella\_limit: This means extra insurance that provides protection beyond existing limits and coverages of other policies

10. insured\_zip: It is the zip code where the insurance was made

11. insured\_sex: This refers to either of the two main categories (male and female) into which customer are divided on the basis of their reproductive functions

12. insured\_education\_level: This refers to the Level of education of the customer

13. insured\_occupation: This refers Occupation of the customer

14. insured\_hobbies: This refers to an activity done regularly by customer in his/her leisure time for pleasure.

15. insured\_relationship: This whether customer is: single; or. married; or. in a de facto relationship (that is, living together but not married); or. in a civil partnership

16. capital-gains: This refers to profit accrued due to insurance premium

17. capital-loss: This refers to the losses incurred due to insurance claims

18. incident\_date: This refers to the date which claims where made by customers

19. incident\_type: This refers to the type of claim/vehicle damage made by customer

20. collision\_type: This refers to the area of damage on the vehicle

21. incident\_severity: This refers to the extent/level of damage

22. authorities\_contacted: This refers to the government agencies that were contacted after damage

23. incident\_state: This refers to the state at which the accident happened

24. incident\_city: This refers to the city at which the accident happened

25. 1ncident\_location: This refers to the location at which the accident happened

26. incident\_hour\_of\_the\_day: The period of the day which accident took place

27. number\_of\_vehicles\_involved: This refers to number of vehicles involved the accident

28. property\_damage: This refers to whether property was damaged or not

29. bodily\_injuries: This refers to injuries sustained

30. witnesses: This refers to the number of witnesses involved

31. police\_report\_available: This refers to whether the report on damage was documented or not

32. total\_claim\_amount: This refers to the financial implications involved in claims

33. injury\_claim: This refers to physical injuries sustained

34. property\_claim: This refers to property damages during incident

35. vehicle\_claim: This refers to property damages during incident

36. auto\_make: This refers to the make of the vehicle

37. auto\_model: This refers to the model of the vehicle

38. auto\_year: This refers to the year which the vehicle was manufactured

39. \_c39:

40. fraud\_reported

***Dataset Link***

https://github.com/dsrscientist/Data-Science-ML-Capstone-Projects/blob/master/Automobile\_insurance\_fraud.csv

### Data Analysis

The first step is to import the libraries for data preprocessing, ML model building, training and evaluation. The Insurance claim fraud detection dataset consist of 1000 rows and 40 columns. The dataset contains punctuation and null records are present in the dataset. I used regular expression to remove the punctuations in the dataset. I applied the Scikit learn library SimpleImputer function to remove the null records present in the dataset. There are no duplicate records present in the dataset. The ML model can only understand the number type data for training and prediction so I used the scikit learn LabelEncoding technique to encode the categorical column data with value between 0 and n\_classes-1. It can be used to transform non-numerical labels to numerical labels. The summary Statistics includes mean, median, mode, standard deviation using pandas describe method to understand the distribution of data.

***Feature Extraction***

Feature extraction is a crucial step in the machine learning process. It involves transforming raw data into numerical features that can be used by machine learning algorithms to make predictions or identify patterns. The main goal of the feature extraction is to reduce the dimensionality of the data while retaining the most important information. This helps in improving the efficiency and performance of machine learning models. I applied statistical Chi2 test for feature extraction using hypothesis testing to extract the important features for prediction of target or label data.

***Exploratory Data Analysis***

Exploratory Data Analysis (EDA) is a crucial step in detecting insurance claim fraud. It helps uncover patterns, correlations, and anomalies in the data that might indicate fraudulent activity. I used different EDA technique for univariate and multivariate analysis using matplotlib and seaborn library. The seaborn box Plot is used to Identify the outliers and understand the distribution of numerical variables. The bar charts is used to visualize the distribution of categorical variables. Then I used heatmaps to identify correlations between variables. The seaborn histplot is used to identify the skewness present in the dataset.

Pre-Processing Pipeline

1, Handling Missing Data

Missing data in machine learning is a type of data that contains “None” or “NaN” type of values. Missing data can be filled using basic python programming, pandas library, and a sci-kit learn library named SimpleImputer. Handling missing values using the sci-kit learns library SimpleImputer is the easiest and most convenient method of all the other missing data handling methods. The analysis on the dataset shows that there are missing data present for the columns “collision\_type”, “property\_damage”, “police\_report\_available”, “authorities\_contacted” and “\_c39” in the dataset. I used Scikit learns library Simple Imputer to remove the null records in the dataset.

2, Data Transformation

The output from the univariate and bivariate analysis shows that there are outliers and skewness present in the dataset. The outliers are removed using statistical z-score test and the skewness are removed using power Transformer method in scikit learn library. SMOTE is the over sampling technique specifically designed to tackle imbalanced datasets by generating synthetic samples for the minority class. I used the SMOTE oversampling technique to balance the imbalanced dataset. Then the standardization technique using scikit-learns library StandardScaler method is applied to the input data (x) that ensures the data is on a consistent scale. This process transforms the data such that it has a mean of 0 and a standard deviation of 1.

***Building Machine Learning Models***

The dataset is split into training and test data with 30 percent of data is used for predicting the insurance claim is fraud or not and remaining 70 percent of data is used for training the classification models. The ten robust machine learning models, which include logistic regression, decision tree classification, random forest classification, Kneighbors classification, support vector machine, naïve Bayes, ada boost classification, gradient boosting classification, bagging classification and extra tree classification models are trained on the training data and tested on the test data. On the basis of the initial model performance, different features of the model are engineered and tested again. A predictive model is created that predicts if an insurance claim is fraudulent or not. Binary Classification task takes place which gives answer between YES or NO. The cross validation technique is performed for each model with the cross validation value of 5 i.e cv=5 to evaluate the performance of a model on unseen data. It helps ensure that the model generalizes well and isn’t overfitting to the training data. For each algorithm performance is examined based on the four evaluation measurements that includes accuracy score, precision score, recall score, R2 score. The ROC-Curve performance metrics on test data shows that the gradient boosting classification model result in best performance and the model is fine tuned using scikit learns library GridSearchCV method to optimize the model. This method involves specifying a grid of hyperparameter values and evaluating every combination to find the best one. The Gradient boosting classification model achieved 92 percent accuracy in prediction for the insurance claim fraudulent or not.

**Table data shows the comparison between different machine learning algorithm**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model name** | **R2\_score** | **Accuracy Score** | **Precision Score** | **Recall Score** |
| AdaBoostClassifier | 0.686411 | 0.922222 | 0.916667 | 0.942857 |
| GradientBoostingClassifier | 0.686411 | 0.922222 | 0.913386 | 0.946939 |
| RandomForestClassifier | 0.560976 | 0.891111 | 0.904959 | 0.893878 |
| BaggingClassifier | 0.507218 | 0.877778 | 0.892562 | 0.881633 |
| ExtraTreesClassifier | 0.42658 | 0.857778 | 0.878661 | 0.857143 |
| DecisionTreeClassifier | 0.390742 | 0.848889 | 0.876596 | 0.840816 |
| SVC | 0.336984 | 0.835556 | 0.905213 | 0.779592 |
| KNeighborsClassifier | 0.283225 | 0.822222 | 0.818533 | 0.865306 |
| Logistic Regression | 0.256346 | 0.815556 | 0.852174 | 0.8 |
| GaussianNB | 0.112992 | 0.78 | 0.787402 | 0.816327 |

***Concluding Remarks***

The gradient boosting classification algorithm performed better in prediction of insurance claim fraud when compared to other classification model.