DETECTION OF FRAUDULENT INSURANCE CLAIM – Using machine learning

Problem Definition:

The insurance industry is usually hit with high volumes of claims with the majority being fraudulent. Claims of fraudulent nature has severe impact on most of the insurance companies and the industries at large. This has a negative consequence for calculating premiums. It is therefore obvious that huge number of insurance claims could result in increasing insurance premiums. The two most common fraud perpetrated by customers of insurance companies are misrepresentation of material facts (when the policy is being incepted, usually motivated by a desire to secure a cheaper premium) and dishonest claim (where the policy holder has genuinely suffered a loss but has then represented the claim dishonestly).

Several research regarding insurance claims has been done to cover other areas using different statistical means. The works of Ranjodh Singh, and et al.in 2019, built a system that was used to take images of damaged cars as inputs and produces other relevant information such the repairing cost in deciding the insurance claim amount and damage localization. This research did not consider several aspects of insurance claim such as the sum assured prior to purchasing the insurance cover. The sum assured is a key factor in determining the premium of cover thus it becomes inaccurate for a model to consider using the damaged vehicle to determine the cost of insurance claim. This model will therefore not be applicable when there is a complete damage to the vehicle and thus needs replacement.

On the other hand, the project by Dan Huangfu, 2015 the aimed at comparing the performances of various statistical models and methods to predict the bodily injury liability insurance claim payments based on the characteristics of the insured's vehicles in a particular data set. However, the data set includes a large number of missing values for the categorical variables which were not appropriately handled.

Another work done by G. Kowshalya, M. Nandhini, 2018 considered data mining techniques to predict fraudulent claims and to calculate insurance premium amount for different customers based on their personal and financial details. Three classifiers namely, Random forest, J48, and Naïve Bayes were built to predict fraudulent claims and percentage of the premium amount. Depended on the synthetic dataset, the results show Random forest outperforms the remaining techniques. The data used for this study had a very small dataset and thus likely to have a downside if applied on a much bigger dataset. This current study seeks to correct this downside and the application of other alternative research method.

It could be seen from the previous research reviewed above that there were issues relating to both big volumes of data and missing values in their works, but they depended on common and traditional methods which studies have proven incorrect.

consequently, we focus on advanced statistical methods and machine learning algorithms that are the most suitable method for the problem of claim prediction with many missing values.

This current article however focuses on advanced statistical methods and machine learning algorithms that are the most suitable method for the problem of fraudulent claim prediction with no missing values. This is made possible by looking at how machine learning could be used to analyse a set of historical data to aid in predicting whether an insurance claim is fraudulent or otherwise. The auto insurance dataset has the details of the insurance policy along with the customer details. It also has the details of the accident based on which the claims have been made. A machine learning model is thus expected to be built from the dataset upon which an accurate prediction of whether the claim is genuine or fraudulent could be determined.

Data Analysis:

The auto insurance dataset used for this analysis consist of one thousand (1,000) customer information shared under forty (40) unique columns describing the customer information. The training data was used to build the model for this study in other to predict the likelihood of a claim being fraudulent or otherwise. The testing data on the other hand is used to check the accuracy of the model.

The customer information collected include the following: Months as customer, age, policy number, policy bind date, policy state, policy csl, policy deductible, policy annual premium, umbrella limit, insured zip, insured gender, insured educational level, insured occupation, insured hobbies, insured relationship status, capital gains, capital loss, incident loss, incident date, incident type, collision type, incident severity, authorities contacted, incident state, incident city, incident location, incident hour of the day, number of vehicles involved, total claim amount, insured claim, property damaged, auto made and model, and fraud reported.

The dependent variable used for building the model is “Fraud Reported”.

The following observations were made in the collected data;

1. There were no null values in the data collected, hence there was no need to provide assumed information using either the mode or the mean in the respective column to fill it up.
2. The data type consisted of integers, floats and objects. It was therefore necessary to convert the object values into integers to aid in the analysis.
3. A huge gab was observed between the maximum amount and the 75th percentile for the following columns; months as customers, policy annual premium, capital gains, total claim amount. This signifies the existence of outliers in the data set.
4. The data set revealed the existence of a high inverse correlation of about 40% between Fraud Reported and Incident Severity.
5. It is also observed that, Incident Location and Auto Model has an almost no correlation with Fraud report.
6. Vehicle Claim recorded the highest positive correlation with Fraud Reported at a value of 17%.

Data Visualization:

Chart, scatter chart

Description automatically generated

*Fig. 1 Scatter plot for months as customer and total claim amount*

Figure 1 above shows an even distribution of data between months as customer and total claim amount. It is obvious that from the figure that as the number of months increases, the total claim amount reduces.

Chart, line chart

Description automatically generated

*Fig. 2. Scatter plot of number of vehicles involved and total claim*

Figure 2 above also shows an unclear correlation between the number of cars and the claim amount. It is seen that, at some point when the number of cars are less the total claim is high and vice versa.

Chart, line chart

Description automatically generated

*Fig. 3 Scatter plot of number of property damage and property claim*

Figure 3 indicates the trend of distribution between property damage and property claim. There is an even distribution of data between the two variables.

There seem to be property claims which has not been indicated as being damaged or not damaged.

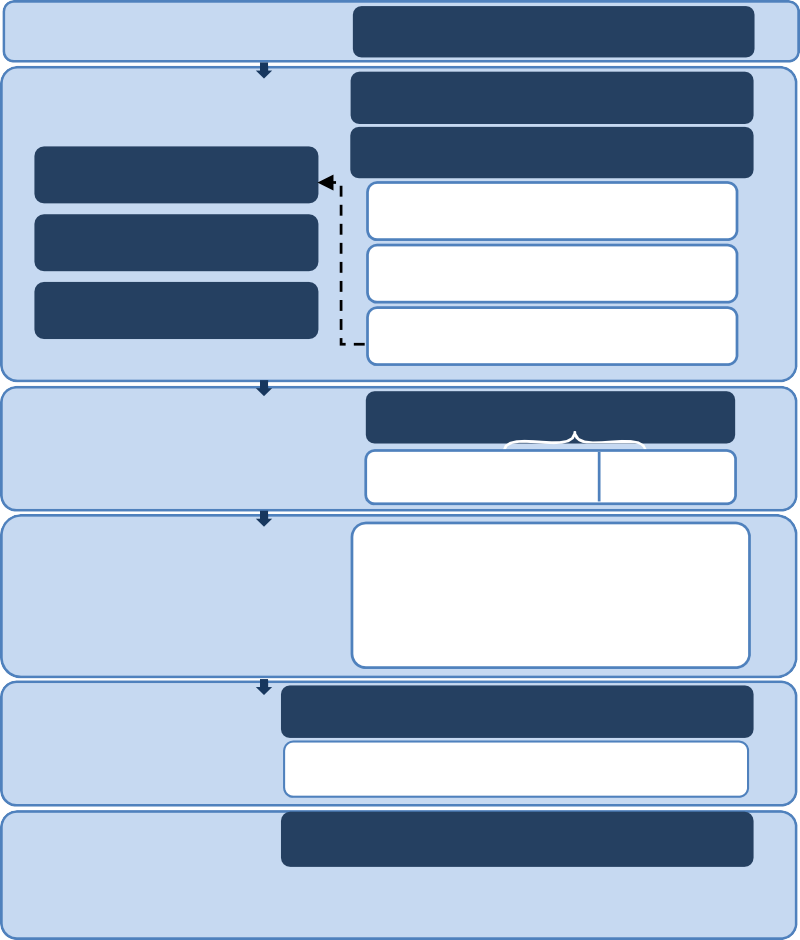
Chart, line chart

Description automatically generated

*Fig. 4 Scatter plot of police report and vehicle claim.*

The above figure shows an even distribution of the data. There is however some claims made which do not have response to whether there is a police report available or not.

*Model Breakdown*



Collection of Data

Data Set

Data preprocessing

Variables Discretization

Discarding the irrelevant Handle Missing Data

Missing Data Analysis

Variables Encoding

Detection Missing Data

Standard Scaler

Impute missing data

Splitting Data

Cross Validation

Train Data

Test Data

Machine Learning Techniques

* Neural Network
* XGboost
* Decision tree
* naïve Bayes

Evaluation Model

Evaluation Model Using Test Data

Confusion Matrix | Accuracy | ROC

Results

Comparison Between Results of

Neural Network**,** XGboost**,** Decision tree**,** naïve Bayes

In this analysis, predicting the auto insurance claims using machine learning from the data collected from customers was through the above designed model. The processes involved are;

1. Data Collection phase
2. Data preprocessing phase
3. Data splitting into training and testing
4. Selection of classification models
5. Evaluation of the accuracy of the built model

Data Preprocessing:

To improve the predictive effect of our proposed model, the raw data needed to be cleaned or corrected from any missing values, duplicated values, inconsistent values etc. It is therefore imperative to preprocess the data before the predictive model. The following processing were followed to enhance the accuracy of the model.

*Handling missing data*

There were no missing data in the data set. Hence no tool was employed to fill up the null values. This could have been done using the Simple Imputer imported from Sklearn.impute

*Label Encoding*

The algorithm used will not work with string dataset such as gender, months as customer, age etc., hence needed to be converted into integers.

*Outliers*

There was the possibility of the existence of outliers in the dataset which needed to be taken out. The values of zscore imported from scipy.stats was used to identify the outliers and were eventually filtered out of the independent dataset.

*Standardization*

The difference that existed between the qualities of the features, standard scaler processing generally plays a crucial role in transforming raw data into a dimensionless index which was applied on all attributes in the data set to ensure each index value is at the same scale level. to standardize training set. The MinMaxScaler method was employed to ensure this is achieved.

Modeling:

The data set has been divided into training and testing parts, with 70% being training and the rest being for the testing set. The model was fitted with the training dataset and the testing data set was used to calculate the accuracy of the prediction model. This article utilized four of the widely used classification models namely Decision Tree, Naïve Bayes, KNeighbors, and SVC.

*Models Evaluation and Experimental Results*

The performance of the classifier was evaluated using confusion matrix and accuracy score. The confusion matrix known as the contingency table is a specific table layout that displays the performance of a model. For binary classification, it contains two rows and two columns that report the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN).

|  |  |  |
| --- | --- | --- |
|  | Model Prediction Positive | Model Prediction Negative |
| Truth:  Positive | 215 | 1 |
| Truth:  Negative | 78 | 0 |

|  |  |  |
| --- | --- | --- |
|  | Model Prediction Positive | Model Prediction Negative |
| Truth:  Positive | TP | FN |
| Truth:  Negative | FP | TN |

Tab 1. Confusion Matrix Tab. 2 *Confusion matrix for SVC*

|  |  |  |
| --- | --- | --- |
|  | Model Prediction Positive | Model Prediction Negative |
| Truth:  Positive | 183 | 33 |
| Truth:  Negative | 35 | 43 |

|  |  |  |
| --- | --- | --- |
|  | Model Prediction Positive | Model Prediction Negative |
| Truth:  Positive | 216 | 0 |
| Truth:  Negative | 78 | 0 |

*Tab. 3 Confusion matrix for Decision Tree Tab. 4 Confusion matrix for MultinomialNB*

|  |  |  |
| --- | --- | --- |
|  | Model Prediction Positive | Model Prediction Negative |
| Truth:  Positive | 202 | 14 |
| Truth:  Negative | 70 | 8 |

*Table 5. Confusion matrix for KNeighbors Classifier*

From the above tables, MultinomialNB recorded the highest number of True Positive of 216 followed by the SVC model. The lowest recorded True Positive is with the Decision Tree model. The Decision Tree recorded the highest True Negative of 43 and highest False Negative of 33. The number of true prediction made by the Decision Tree model resulted in the higher accuracy as recorded in table 6 below.

*Table 6. Accuracy Score*

|  |  |
| --- | --- |
| Technique | Accuracy |
| Multinomial NB | 73.4% |
| KNN | 71.4% |
| DTC | 76.8% |
| SVC | 73.1% |

The above results indicates that, the decision tree model is better able to make a lot more true predictions than all the other models used for this analysis.

Conclusion:

Fraudulent claims pose persistent challenges for the insurance industry to deal with. Insurance companies are therefore advantaged if there is a readily available model that could test details of a claims in anticipation of the claim being genuine or fraudulent. This comes in handy as numerous insurance companies are burdened with the use of traditional methods to analyze clients’ details. However, the historical claim data are usually big data filled with several missing information for many attributes of the data. Therefore, the application of advanced statistical methods and machine learning algorithms can help resolve these problems. This work constructed a model to predict fraudulent insurance claims, with the use of four classifiers which were built to predict the claims. The algorithms KNeighbors, Decision Tree, SVC, and MultinomialNB, were the choice for this classification. The Decision Tree model performed best among the four models. With an accuracy score of 76.8%, the DTC model is well able to predict the extent to which a claim could be deem as fraudulent and that which is not.