Insurance Claims- Fraud Detection

Insurance fraud occurs when someone deceives an auto insurance company in order to benefit financially. Some car insurance fraud cases are more severe than others, but fraud is by no means a victimless crime.

The total cost of insurance fraud is approximately $40 billion annually, according to the FBI. This means the average family pays between $400 and $700 extra per year on insurance premiums to make up fraud costs

Fraud is causing billions of $$ in loss for insurance industry. This project has attempted to develop a ML algorithm to detect. The project has used the historical transaction data including normal transactions and fraud ones to obtain normal/fraud behavior features based on machine learning techniques, and utilized these features to check if a transaction is fraud or not. A cmparative study has been conducted to decide which classifier is best for this project to train the behavior features of normal and abnormal transactions.

Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem

Plan

The data that I have is from Automobile Insurance. I will be creating a predictive model that predicts if an insurance claim is fraudulent or not. The answere between YES/NO, is a Binary Classification task. This report deals with classification algorithm models to detect fraud transaction in Python. Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.

In this project, we 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, we 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.

Problem Statement

The goal of this project is to build a model that can detect auto insurance fraud. The challenge behind fraud detection in machine learning is that frauds are far less common as compared to legit insurance claims. This type of problems is known as imbalanced class classification .Frauds are unethical and are losses to the company. By building a model that can classify auto insurance fraud, I am able to cut losses for the insurance company. Less losses equates to more earning.

About the Dataset

Data source: <https://github.com/nidhi169/datatrained/blob/main/Automobile_insurance_fraud.ipynb>

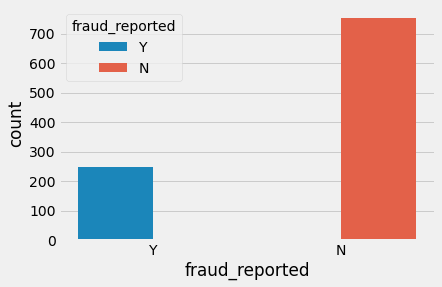
The inspiration for this project was to perform classification on imbalance class data sets, in particular fraud. Fraud data sets are very hard to come by and often unlabeled due to its sensitive nature

The current data set was labelled with n=1000 samples. Before any cleaning or feature engineering, the data set has a total of 40 variables. It is not stated if this data is from multiple insurance companies or just one company. However, throughout the report, “the insurance company” will be used to refer to the origin of this data.

Exploratory Data Analysis

Dependent variable

Exploratory data analysis was conducted started with the dependent variable, Fraud\_reported. There were 247 frauds and 753 non-frauds. 24.7% of the data were frauds while 75.3% were non-fraudulent claims.

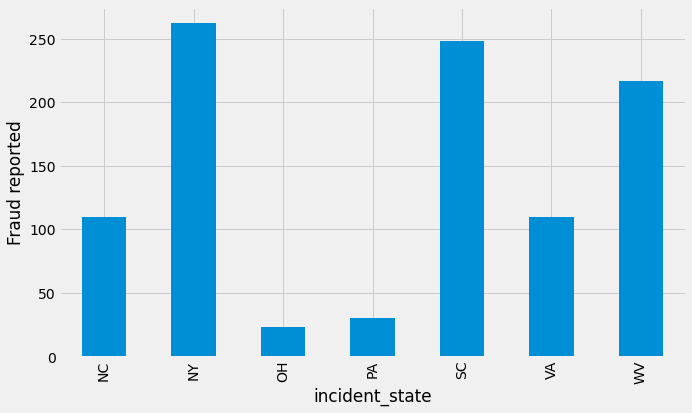


Correlations among variables

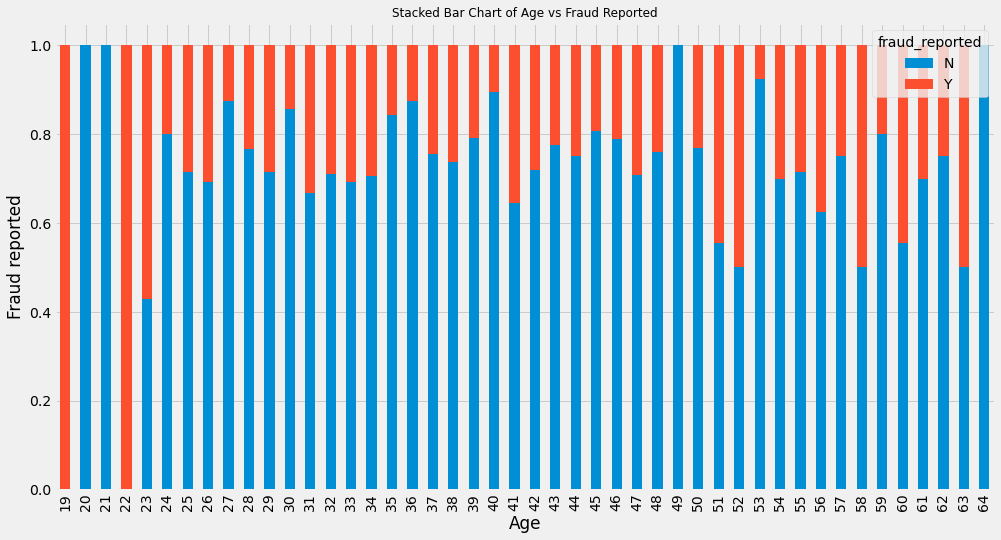
Next, correlations amongst continuous variables (ordinal, interval/ratio variables) were inspected. Heatmap was plotted for variables with at least 0.3 Pearson’s correlation coefficient , including the DV. Month as customer and age had a correlation of 0.92. Probably because drivers buy auto insurance when they own a car and this time measure only increases with age. Incident severity and different types of claims have a clear correlation .Apart from that, there don’t seem to be much correlations in the data. There don’t seem to be multicollinearity problem except maybe that all the claims are all correlated, and somehow total claims have accounted for them.

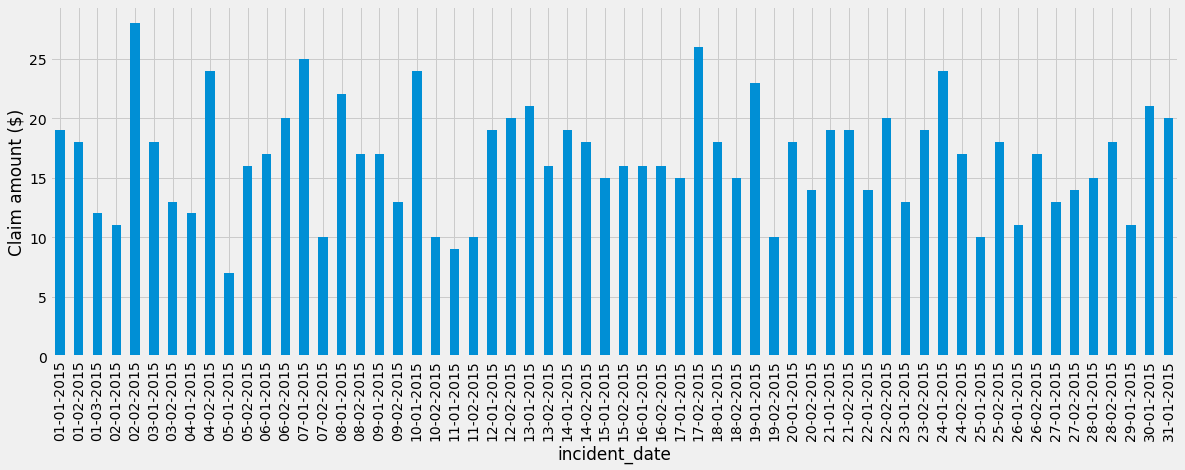
However, the other claims provide some granularity that will not otherwise be captured by total claims. Thus, these variables were kept.

**Visualizing variables against the DV**  
 Below are a few notable plots. Little have I suspected that fraud differed inicident state . It seems like NY and SC have higher tendencies of fraud. I stopped to think if the OH and PA develop one to be inclined to fraud or that people with tendencies are drawn to it

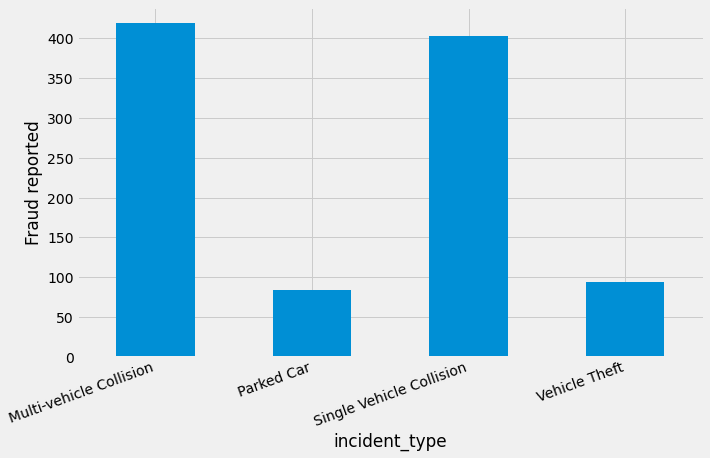


From below plot, it is obvious that, age is an important predictor for fraud reported. Age between 19-64 shows substantial number fraud report.

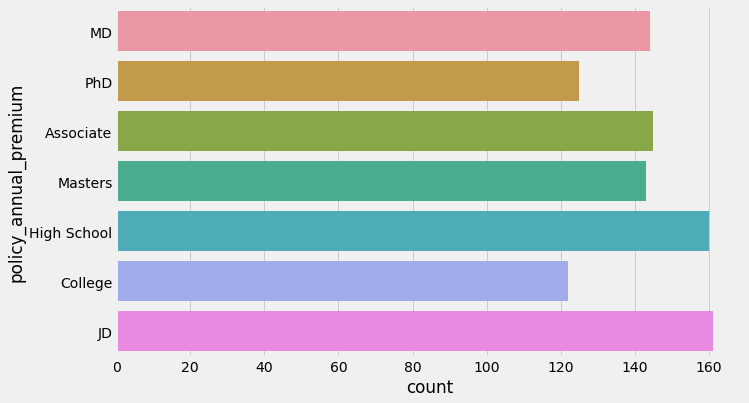


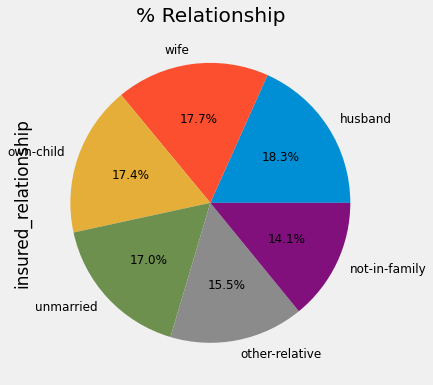
In below plot,We see that, all the cases in above plot are for the months of January and February 2015

In below plot, this data shows the fraud repoted at particular incident type which is the most in multi vehicle collision and least in parked car and vehicle theft.

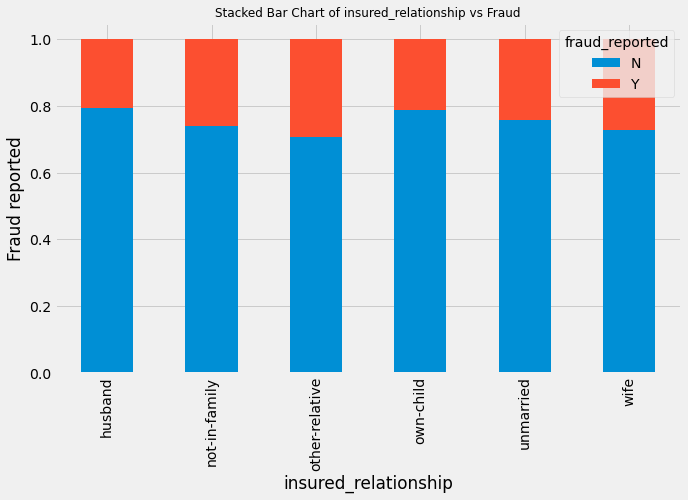


Below graph gives Breakdown of Average Vehicle claim by insured's education level, grouped by fraud reported which is greatest in jd and least in ph.d and college.



Below graph shows the fraud report on the basis of relationship percentage

this below data show the relation between fraud repoted vs insured relationship , which is the most in other relative and least in husbands and own child.

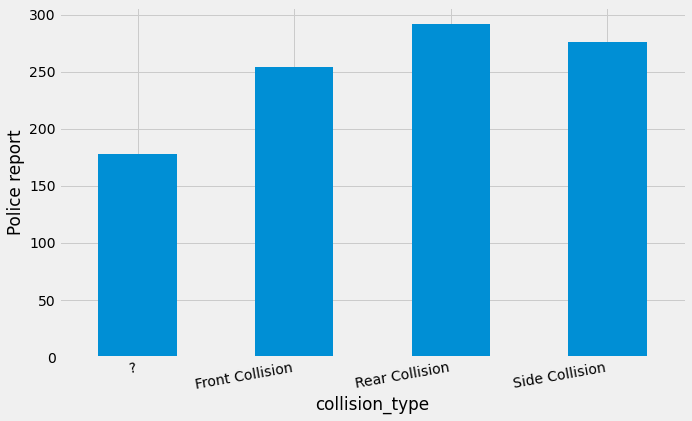


Preprocessing

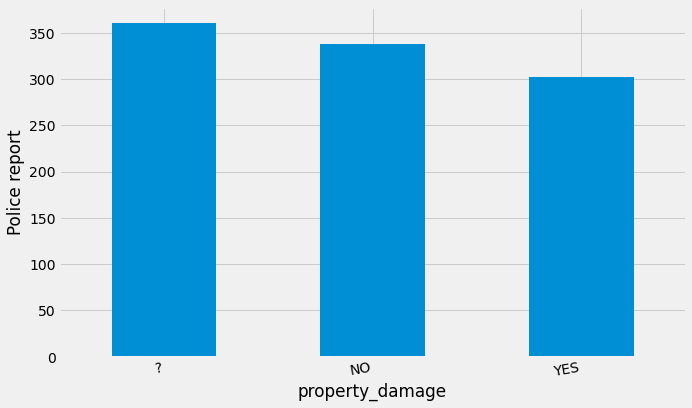
The DV, fraud\_reported was coded 1 for fraud and 0 for non-fraud.Six interaction terms were created. Interaction between property claim amount and incident severity, vehicle claim amount and incident severity, injury claim amount and incident severity, total claim amount and incident severity, policy annual premium and total claim amount, umbrella limit and total claim amount. Nominal variables were one-hot encoded, and the data set was split into 75% train and 25% test set, stratified on fraud reported. Here, I define loss as simply money going out from the insurance company. Source of money coming in, on the other hand, are premiums. Although we know premiums and claims are not the only source of money going in or out of an insurance company, these 2 variables are used since they are the only information, we have from this data set.

In line 43 we have checked the categorical values of our dataset and then we have dropped all unimportant colimns i.e policy number, policy csl, insured zip ,policy bind date, incident date, incident location, c39, Auto year, incident hour of the day.

In below graph, we have seen the relationship between police report and collision type , which shows that the most police reports are done of rear collisions, in comparision of front collision side collision , side collision and missecelanios.



In below graph , we see the relation between police report and property damage.



In line no. 54 we have applied one-hot encoding to convert all categorical variables except out target variables. And then have applied dummies on the columns i.e policy state, insured sex, insured education level, insured occupation, insured hobbies, insured relationship, incident type, incident severity, authorities contacted, incident state, incident city ,auto make, auto model, csl per person, Csl per accident ,incident period of day , and these are joint with columns i.e collision type, property damage, police report available, fraud reported.

Than we have defined our predictor variables and target variable for encoding of our data.

As our dataset is imbalance, accuracy is not a good measure of success. A high accuracy can be achieved by a poor model that only selects the majority class, hence, not detecting and measuring the accuracy of classifying the class of interest. In fact, predicting only the majority class will give an accuracy of 75%, specificity of 100% but a sensitivity of 0%.

# Modeling

Five different classifiers were used in this project:  
- train test split  
- Decision Tree Classifier  
- Random Forest Classifier

-Support Vector Classifier

-cross validation score

Hyperparameter tuning and selection was done for all the models using RandomizedSearch. Due to the number of parameters and models that were ran, RandomizedSearch is a faster more efficient choice as compared to gridsearch.

After a 2-fold RandomizedSearchCV, the model with its selected hyperparameters were fitted on the training set. Parameters we used in hypertuning was max depth and criterion. We find that Random Forest Classifier with (criterion='entropy', max depth=7)was the best model . Mean accuracy scores for the best estimators of the Randomized Search CV, accuracy scores on the training set and accuracy scores on the test set was computed

The three classifiers as per stated above were ran with hyperparameter tuning. Models that had option for class weighting had class weighting as one of the hyperparameters in this block of models. That is, the RandomizedSearch will test a weighted and non-weighted model to see which performed better. A best practice for using the class weighting is to use the inverse of the class distribution present in the training dataset

# Evaluation

Cross validation accuracy scores, accuracy scores on training set, accuracy scores on test set, sensitivity, specificity, precision , f1 score was computed and printed as shown in the table below. Our best model was reandom forest classifier with 73% accuracy with f1 score 0.83%.