INSURANCE CLAIM FRAUD DETCTION

Hey readers, this post is about a project I did on Insurance fraud claim detection using a number of different classifiers and ensembles.

**Problem Statement**

The goal of this project is to build a model that can detect 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.

**Relevance to businesses:**

Imbalance class problems are common in many industries. Many a times, we are interested in a minority class against another much bigger class or classes. For instance, classification of other types of frauds, classification of defective goods, classification of at-risk teenagers, identifying high potential employees, identifying people of interest such as terrorist, just to name a few.

**Background of insurance fraud:**

Insurance fraud is a deliberate deception perpetrated against or by an insurance company or agent for the purpose of financial gain. Fraud may be committed at different points in the transaction by applicants, policyholders, third-party claimants, or professionals who provide services to claimants. Insurance agents and company employees may also commit insurance fraud. Common frauds include “padding,” or inflating claims misrepresenting facts on an insurance application; submitting claims for injuries or damage that never occurred; and staging accidents.

The FBI estimates that the total cost of insurance fraud (excluding health insurance) is more than $40 billion per year.  
Auto insurance fraud ranges from misrepresenting facts on insurance applications and inflating insurance claims to staging accidents and submitting claim forms for injuries or damage that never occurred, to false reports of stolen vehicles.

Fraud accounted for between 15 percent and 17 percent of total claims payments for auto insurance bodily injury in 2012, according to an Insurance Research Council (IRC) study. The study estimated that between $5.6 billion and $7.7 billion was fraudulently added to paid claims for auto insurance bodily injury payments in 2022, compared with a range of $4.3 billion to $5.8 billion in 2002.

The current study aims to classify auto insurance fraud that arises from claims. The type of fraud is not disclosed in this data set and could be false reports, inflating claims, staging accidents or submitting claim forms for damages or injuries that never occurred.

**Executive Summary**

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.

Several models were tested with different methods of handling imbalance datasets. The top models were also fitted and tested with different ensembles.

The final fitted model is a weighted XGBoost which yielded an F1 score of 0.72 and a ROC AUC score of 0.84. The model performed far better than the baseline F1 score of 0.397 and ROC AUC target of 0.7. The model’s F1 score and ROC AUC scores were the highest amongst the other models. In conclusion, the model was able to correctly distinguish between fraud claims and legit claims with high accuracy.

Prior to modeling, the data was clean and exploratory data analysis was conducted. After which, the data was pre-processed for the modeling. After modeling, the models were evaluated, and the best fitted model was selected using the F1 score and the ROC AUC score. The performance of the final fitted model was discussed in further details and its top features were displayed. The project concluded by reiterated the importance of the research and what had been done and finally, with some limitations.

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.

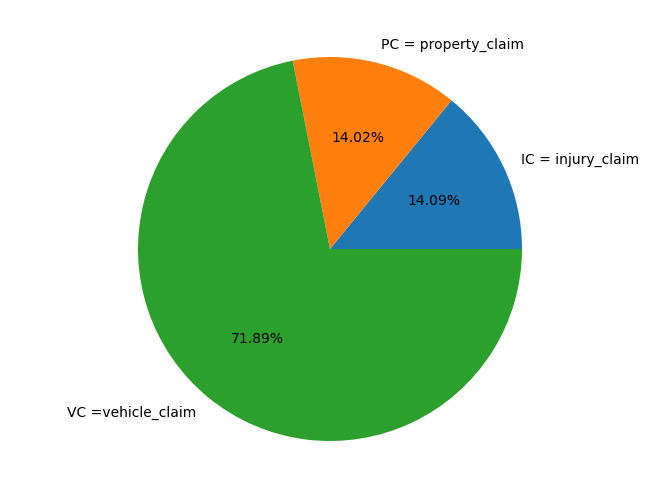
The current data set was labelled with n=1000 samples. Unlike many other data sets, this one was less popular with only the author and one other having a notebook of it on Kaggle, making this data set one that was rather novel in nature. The data set consist of 1000 auto incidents and auto insurance claims from Ohio, Illinois and Indiana from 01 January 2015 to 01 March 2015. Before any cleaning or feature engineering, the data set has a total of 39 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.

The obvious con of this data set is the small sample size. However, there are still many companies who do not have big data sets. The ability to work with what is available is crucial for any company looking to transition into leveraging data science. In the 2017 MIT tech review, EmTech presentation, Professor Andrew Ng penned a cyclical diagram on the white board and explained that many companies start off with some small data and develop a product which have users, which in turn leads to generation of more products. In similar vein, companies may start off with a small data set and build towards a bigger data set as time goes by. Compared to a company that waits for the day when it has a huge data set, the company that started with a small data set and worked on it will more likely succeed earlier in its data science journey and reap its rewards.

**Exploratory Data Analysis**

*Check out my GitHub for more detailed EDA*

**Dependent variable**  
Exploratory data analysis was conducted started with the dependent variable, Fraud\_reported. There were 247 frauds and 753 non-frauds. 14.7% of the data were frauds while 71.3% were non-fraudulent claims.

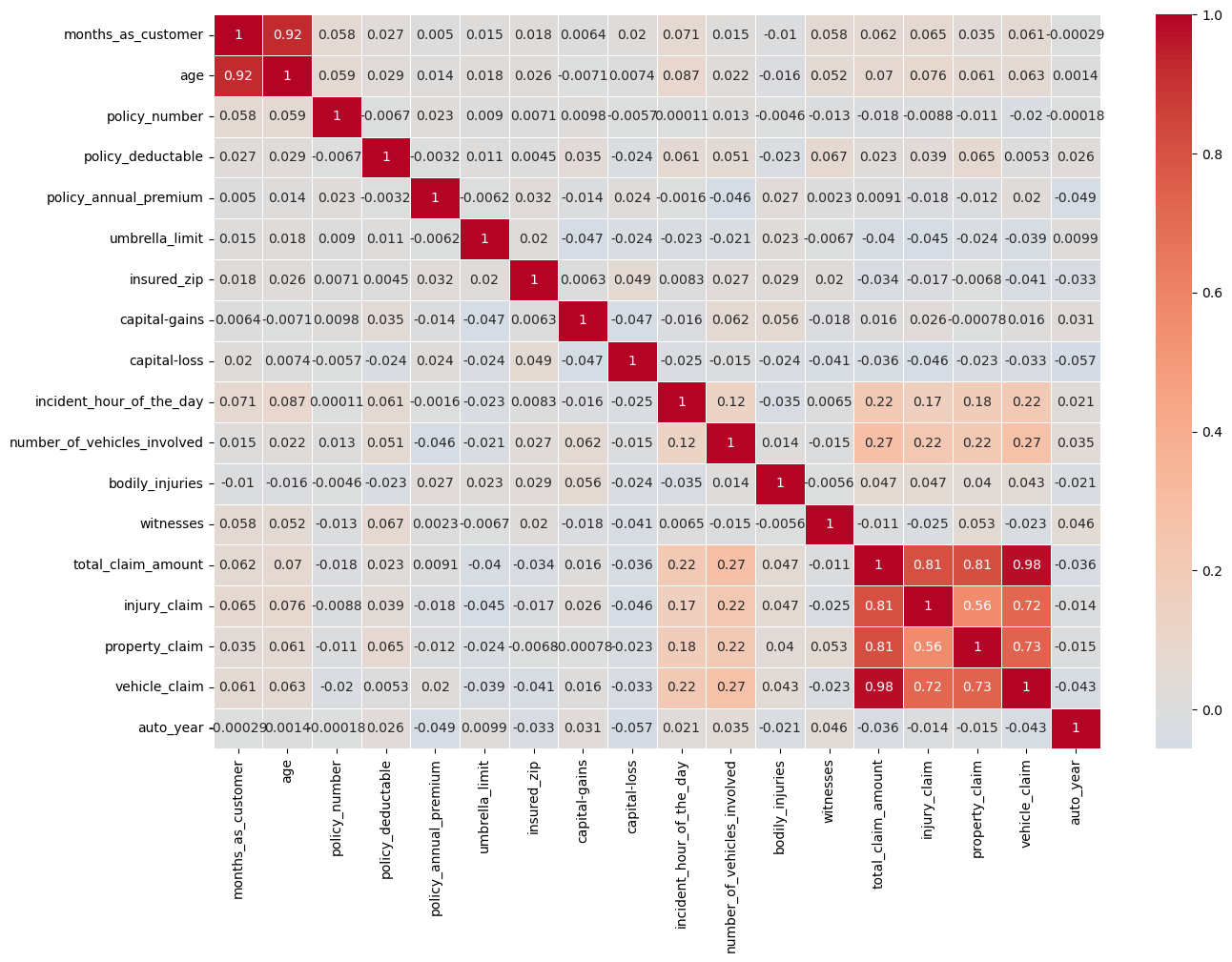


**Correlations among variables**  
Next, correlations amongst continuous variables (ordinal, interval/ratio variables) were inspected. Heatmap was plotted for variables with at least 0.3 Pearson’s correlation coefficient , including the DV.

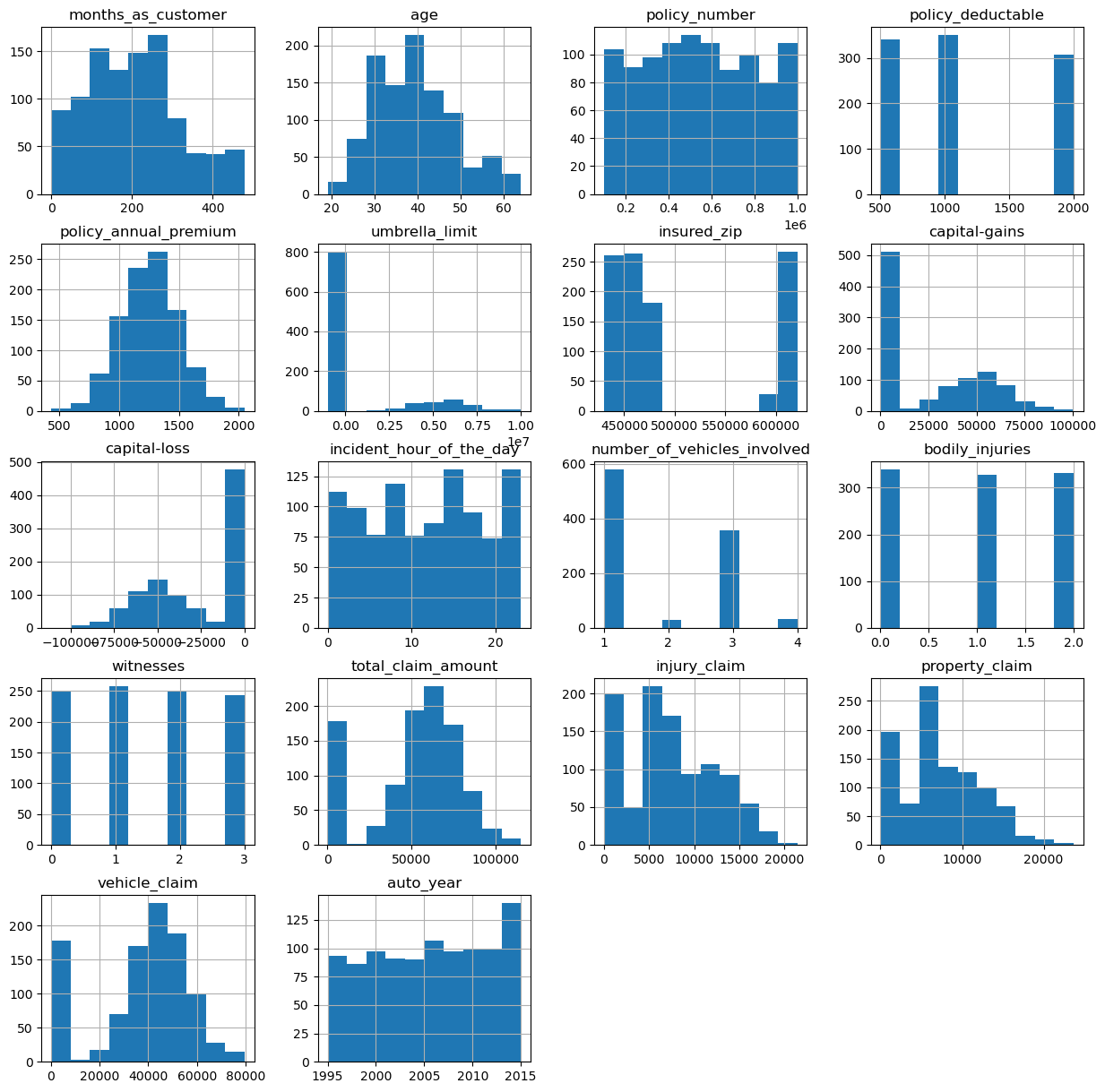
Month as customer and age had a correlation of 0.92. Probably because drivers buy auto insurance when they own a car and this time measure only increases with age.

Incident severity and different types of claims have a clear correlation (0.36–0.50)

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.

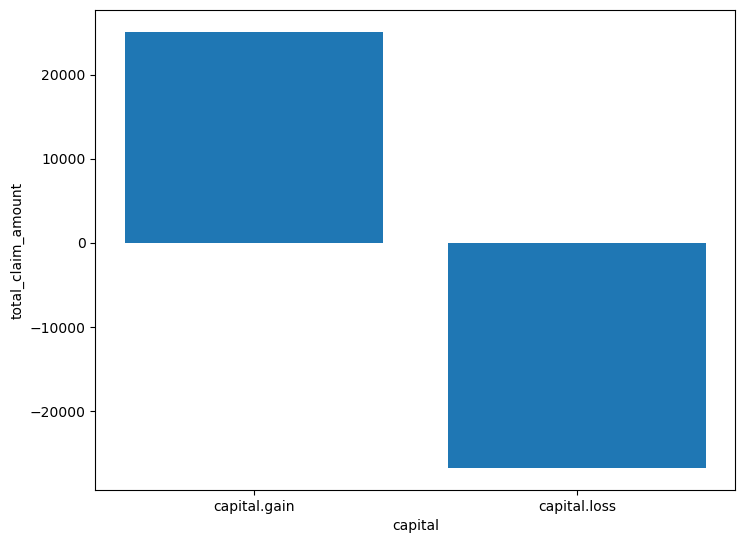


**Visualizing variables against the DV**  
Counts of every variable split by the DV was plotted. Below are a few notable plots.

Little have I suspected that fraud differed across hobbies. It seems like chess players and cross-fitters have higher tendencies of fraud. I stopped to think if the sports develop one to be inclined to fraud or that people with tendencies are drawn to it

Major incident severity seems to have highest fraud cases that exceed non fraud cases.

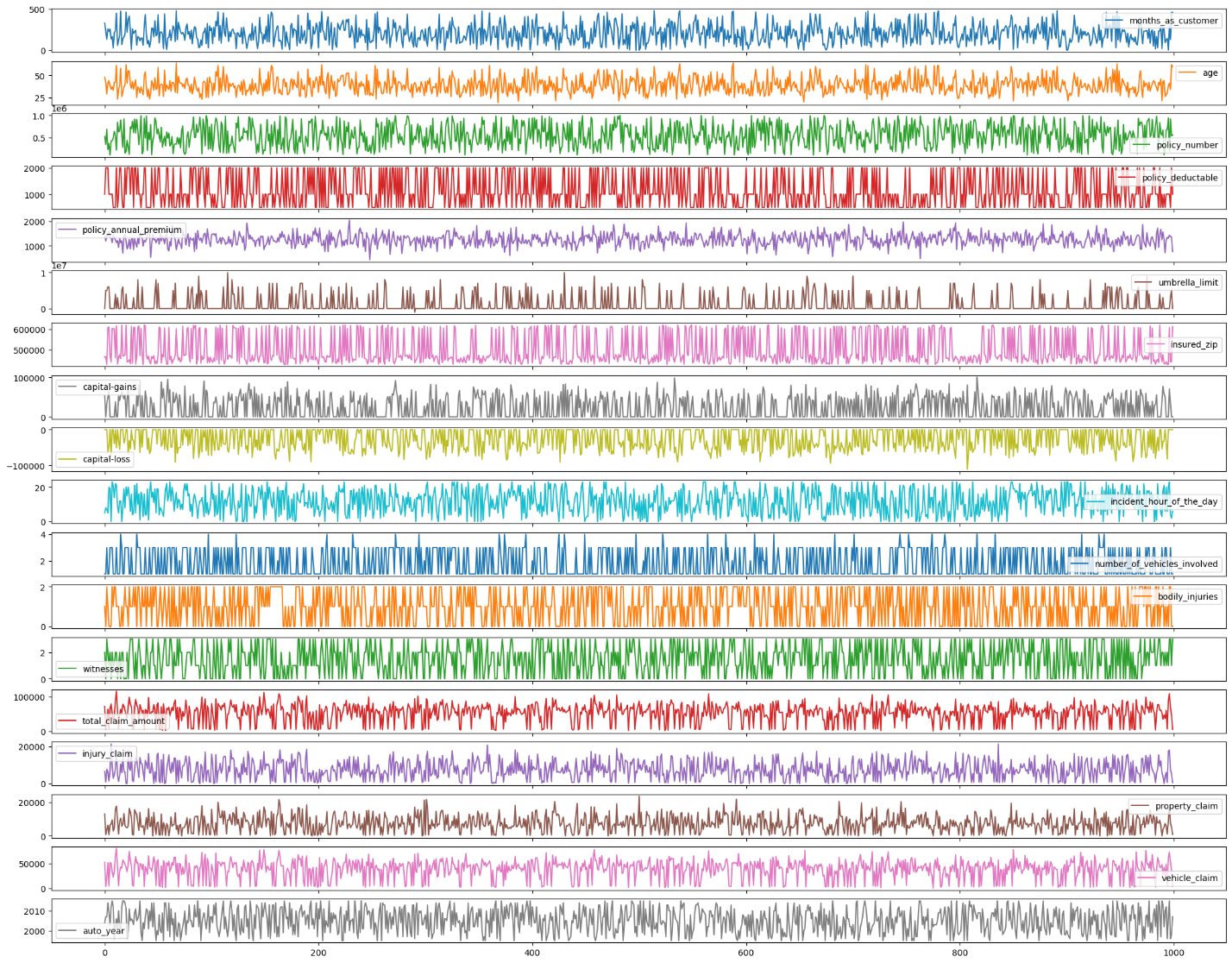
There seem to be more frauds than non-fraud claims along the mean of total claims.



**Losses by Claims**

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. Typically, other source of money movement maybe investments made by the insurance company, for instance.

I created a variable that measure how much claims minus how much premiums were paid by a client to indicate losses by claim. a positive will indicate a loss while a negative will be a profit. Every time a claim is more than the total premiums paid by a client; it is a loss for the insurance company.



**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.

**Baseline Score**

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%.

If we make a naive prediction that all claims are frauds, so that no frauds escape our watch, we will have a score as shown below:

- Sensitivity: 1.0  
- Specificity: 0.0  
- Precision: 0.248  
- F1 score: 0.397  
- ROC AUC Score: 0.50

As identifying as many frauds as possible is the goal, the F1 score of 0.397 was used as a baseline. However, investigations into frauds can be time consuming and expensive and may even affect customer experience. Thus, ROC AUC score will also be used to measure how well we distinguish between Fraud and legit claims. The baseline ROC AUC score is 0.50. I am to have a ROC AUC of at least 0.70.

**Modeling**

1. different classifiers were used in this project:  
   - logistic regression  
   - K-nearest neighbours  
   - Random forest

-DecisionTreeClassifier

-GaussianNB

-KNeighborsClassifier

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 10-fold RandomizedSearchCV, the model with its selected hyperparameters were fitted on the training set.

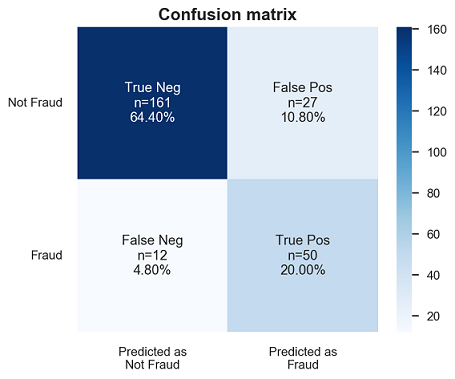
Mean accuracy scores for the best estimators of the RandomizedSearchCV, accuracy scores on the training set and accuracy scores on the test set was computed. Then, the sensitivity, specificity, precision, F1 score and ROC AUC scores were computed.

*Checkout the link for an experiment on RandomizedSearch Vs GridSearch (spoiler alert: RandomizedSearch won on 10, 50, 100, 500 trails by time and by cost function value BUT this is not always the case):*

This section discusses how different blocks of models were ran. Evaluation of the models will be the evaluation section.

**Modeling with Oversampling using bootstrapping**

Unlike SMOTE or ADASYN, bootstrap draws from the already existing distribution and does not create synthetic data. Thus, data are simply repeating of existing and are perceived by many to be less biased. Bootstrap oversampling was conducted by randomly drawing from the existing fraud dataset with replacement until both fraud and non-frauds had the same sample size of 565. Bootstrap was only done on the training set. The five classifiers were running on the bootstrapped data set, with hyperparameter tuning.



**The summary of the classification report is presented below.**

Sensitivity (recall of fraud cases) is derived from:

* True positive/(True positive + False negative)
* Sensitivity summarizes our true positive rate, which is how many we got correct out of all the positive cases.
* Sensitivity for the final model is 81%.

Specificity (recall of non-fraud cases) is derived from:

* True negative/(True negative + False positive)
* Specificity summarizes our true negative rate, which is how many we got correct out of all the negative cases.
* Specificity for the final model is 86%

Precision of fraud cases are derived from:

* True positive/(True positive + False positive)
* Precision of fraud cases summarize the accuracy of fraud cases detected. That is, out of all that I predicted as fraud, how many are correct.
* Precision of fraud detection is 65%.

Precision of non-fraud cases are derived from:

* True negative/(True negative + False negative)
* Precision of non-fraud cases summarize the accuracy of non-fraud cases detected. That is, out of all that I predicted as non-fraud, how many are correct.
* Precision of non-fraud detection is 93%.

F1 scores are the harmonic mean of recall and precision and is derived from

* (2 x recall x precision)\(recall + precision)
* As we are interested in fraud cases, only the F1 scores on fraud cases are reported.
* The F1 score of the model is 72%.

In sum, the model has outperformed the baseline F1 scores by a huge margin.

fraud\_reported total\_claim\_amount

1 1 60302.105263

0 0 50288.605578

**Conclusion and Limitations**

Fraud accounted for between 15 percent and 17 percent of total claims payments for auto insurance bodily injury in 2012, according to an Insurance Research Council (IRC) study.This project has built a model that can detect auto insurance fraud. In doing so, the model can reduces loses for insurance companies. The challenge behind fraud detection in machine learning is that frauds are far less common as compared to legit insurance claims.

RESOUCES:-

**Data source:** <https://www.kaggle.com/roshansharma/insurance-claim>

**Data source**: <https://www.iii.org/article/background-on-insurance-fraud>

**Data source** :[*https://towardsdatascience.com/random-search-vs-grid-search-for-hyperparameter-optimization-345e1422899d*](https://towardsdatascience.com/random-search-vs-grid-search-for-hyperparameter-optimization-345e1422899d)