Insurance Claim Fraud Detection

By : Deepak Singh

Project on auto insurance fraud claim detection using a number of different classifiers and ensembles.

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 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.
People who commit insurance fraud include:

- organized criminals who steal large sums through fraudulent business activities,
- professionals and technicians who inflate service costs or charge for services not rendered, and
- Ordinary people who want to cover their deductible or view filing a claim as an opportunity to make a little money.

Auto insurance fraud

Auto insurers lose at least $29 billion a year, according to a 2017 study by Verisk, to premium leakage, the "omitted or misstated underwriting information that leads to inaccurate rates." A number of information failures and fraudulent practices drive costs up, such as unrecognized drivers ($10.3 billion); underestimated mileage ($5.4 billion); violations/accidents ($3.4 billion); and, false garaging to lower premiums ($2.9 billion). While not always a result of malicious or even conscious actions, premium leakage creates problems for consumers, too—as much as 14 percent of all personal auto premiums can be attributed to the cost of covering premium leakage.
No-fault auto insurance is a system that lets policyholders recover financial losses from their own insurance company, regardless of who was at fault in a motor vehicle accident. However, in many no-fault states, unscrupulous medical providers, attorneys, and others pad costs associated with legitimate claims – for example, by billing an insurer for a medical procedure that was not performed.

Another attempt to solve the problem of title washing is the National Motor Vehicle Title Information System (NMVTIS), a database that requires junk and salvage yard operators and insurance companies to file monthly reports on vehicles declared total losses.

After the hurricanes of 2005, the National Insurance Crime Bureau (NICB) created a database in which vehicle identification numbers (VINs) and boat hull identification numbers (HINs) from flooded vehicles and boats are stored and made available to law enforcers, state fraud bureaus, insurers and state departments of motor vehicles. The database (VINcheck) is online and can be accessed by the general public.

Problem Statement:

Business case:

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, 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.

Dataset link: https://github.com/dsrscientist/Data-Science-ML-Capstone-rojects/blob/master/Automobile_insurance_fraud.csv

Criteria for success:

The model should be able to classify if a claim is a fraud or not on a data set that it has not seen, accurately. The area under curve of the ROC (ROC AUC) will also be taken into consideration in model selection as a secondary criterion as it is important to distinguish between fraud and legit claims. This is because investigations into frauds can be time consuming and expensive and may even negatively affect customer experience. As a compulsory criterion, the ROC AUC must be above 0.55. On top of that, I aim to have a ROC AUC of at least 0.70.

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.
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.

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.
Prior to modeling, the data was clean and exploratory data analysis was conducted, pre-processed for the modeling. After that, models were evaluated,and best fitted model using the ROC AUC score.

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. 24.7% of the data were frauds while 75.3% were non-fraudulent claims.

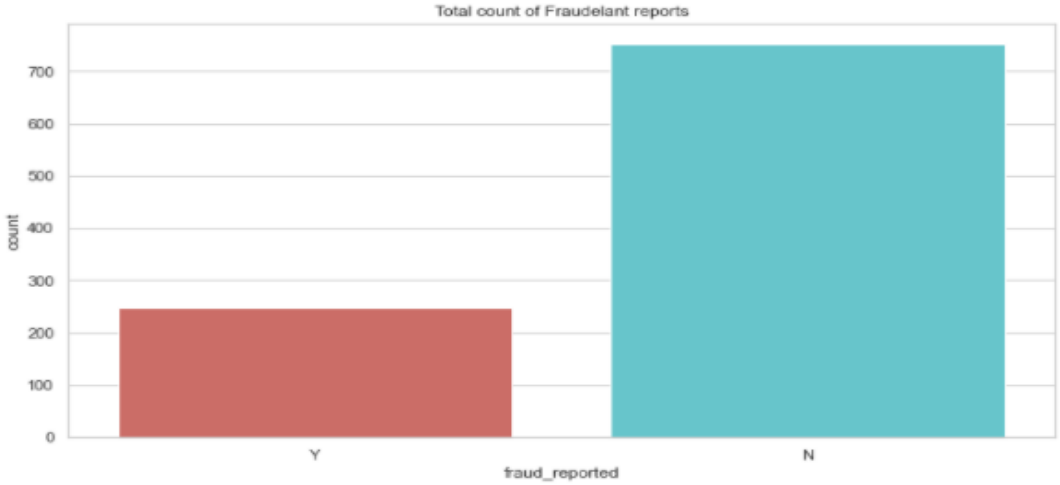
Percentage of people who reported fraud cases:

```
fraud_percent = (df["fraud_reported"].value_counts()/df.shape[0])*100
print(fraud_percent)
```

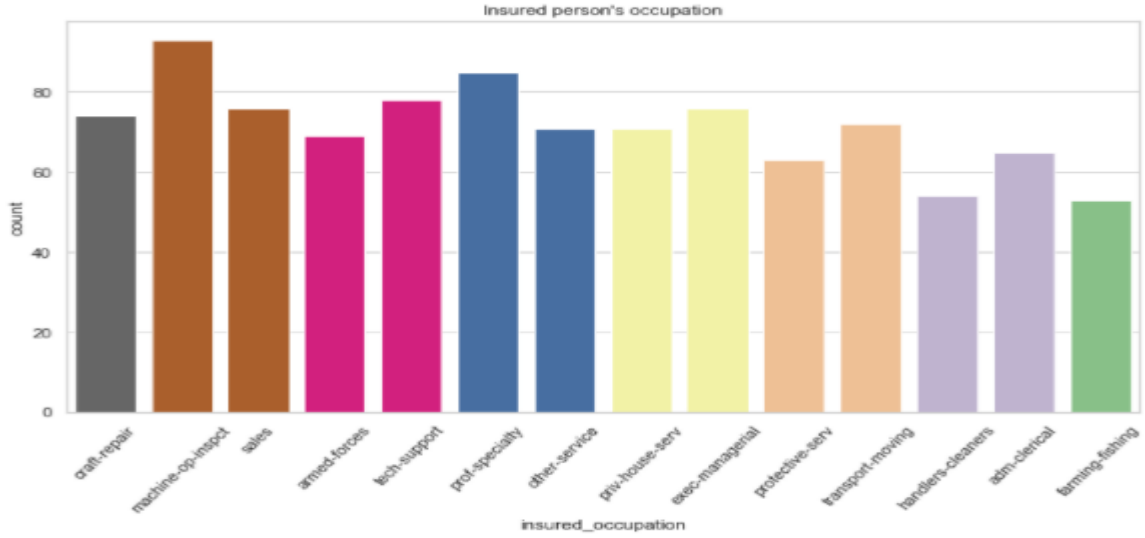
```
N    75.3
Y    24.7
Name: fraud_reported, dtype: float64
```

As per the data approximately 75.3 % of cases are authentic and about 24.7 % of cases are reported as fraud.

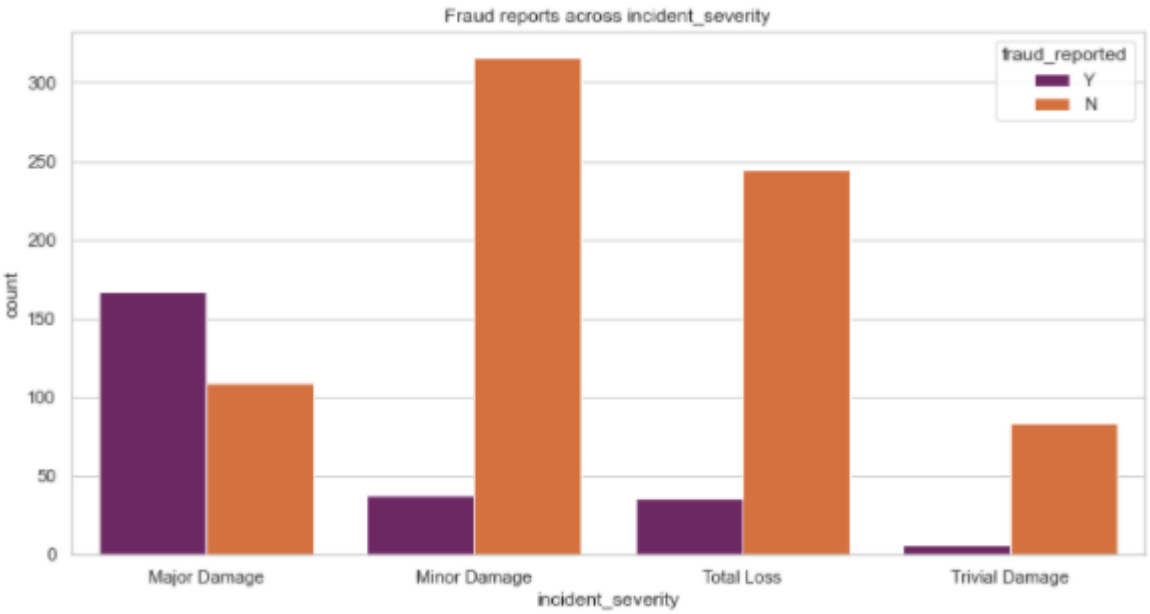
```
# Fraud Percentage:
plt.figure(figsize=(12, 6))
sns.set_theme(style="whitegrid")
ax = sns.countplot(x="fraud_reported", data=df, palette="hls").set(title='Total count of Fraudelant reports')
```



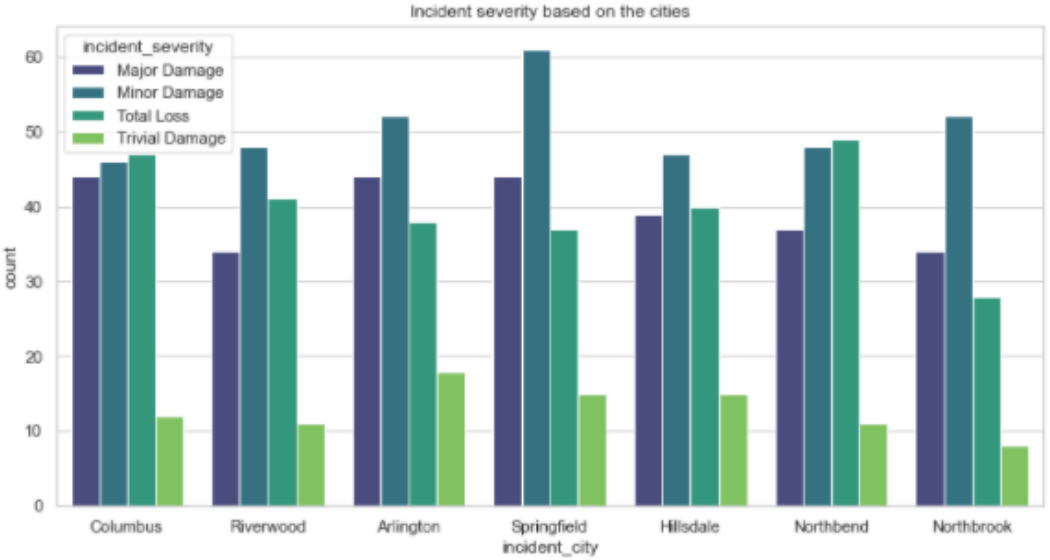
Insured person's occupation:



Fraud reported over the severity of incidents and the Incident severity based on the cities:



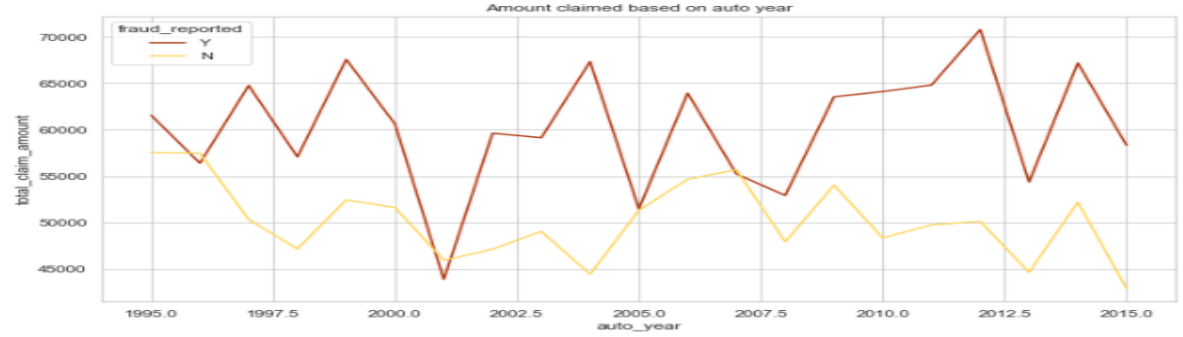
auto year and total claimed amount



'Columbus" city has highest number of "Major accidents" and Springfield has majority of "Minor accidents"

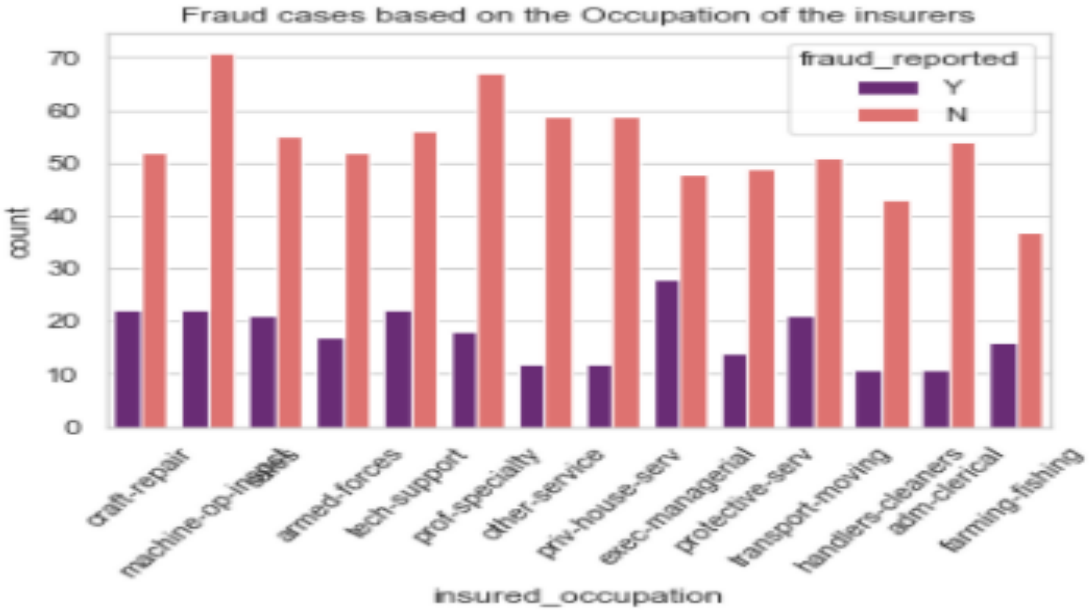
Auto year and total claimed amount

```
[Text(0.5, 1.0, 'Amount claimed based on auto year')]
```

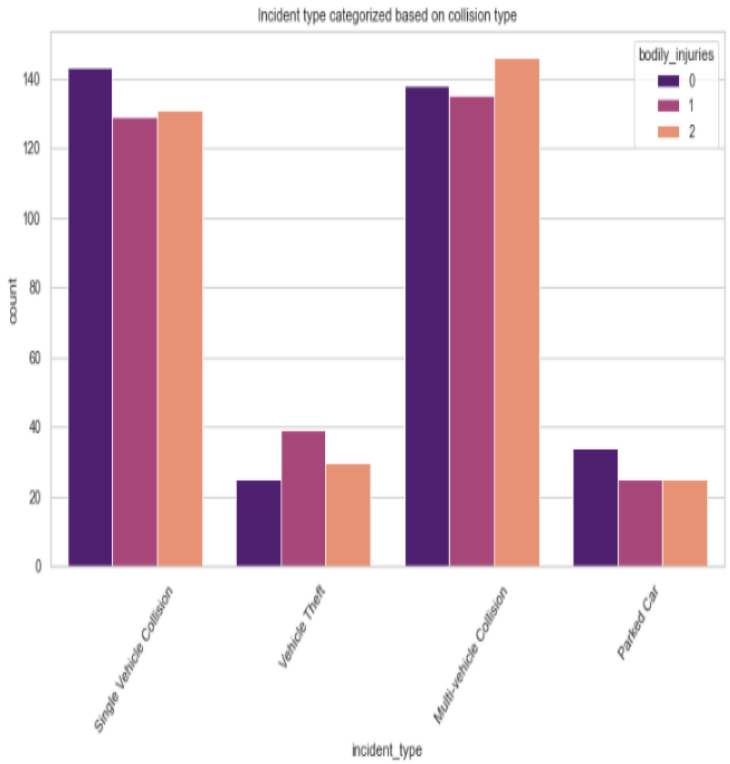


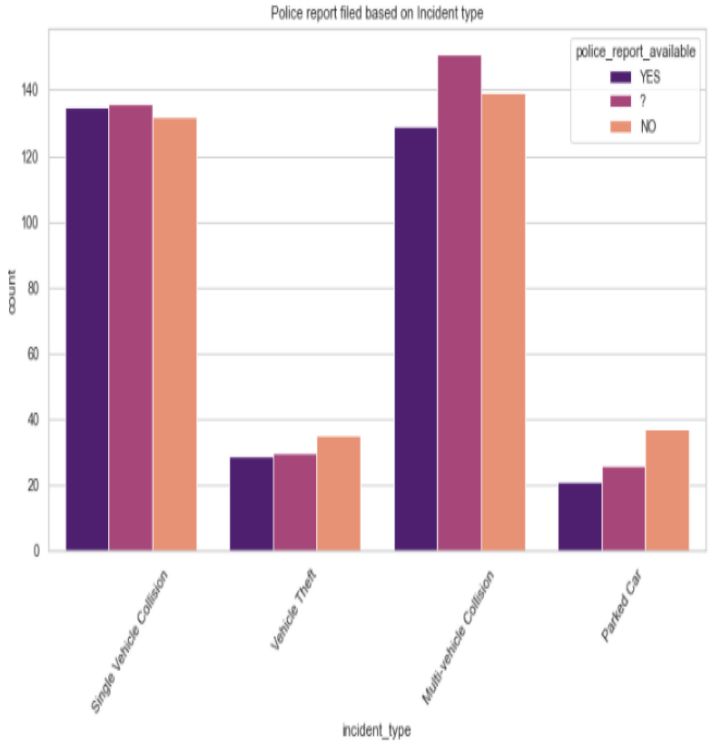
Suspected that fraud differed across. It seems like players or cross-fitters have tendencies of fraud. There seem to be more frauds than non-fraud claims along the mean of total claims.

Majority of people haven't reported false fraud; people fall under occupation "exec-managerial". any fake fraud claimed:

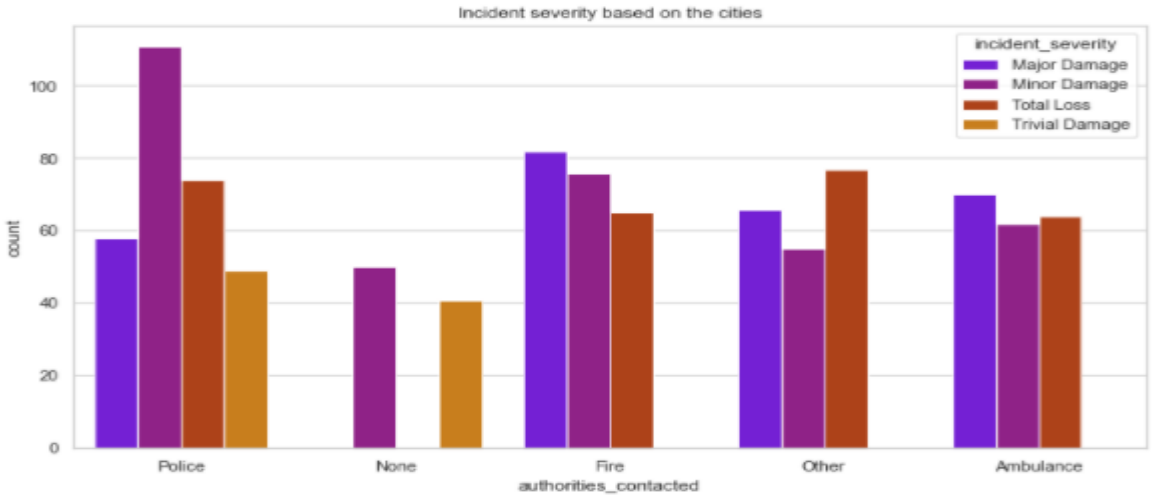


Incident types, bodily injuries in case of "vehicle theft" and "parked car" and Police report filed or not:





Incident severity and authorities contacted:



Cases of "Major accident", Fire department was called first. Maybe the vehicle was compromised in such a way causing fire hence fire departments were approached then most contact authorty is "ambulance"

Losses due to 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.

The insurance company lost $8,198,060.09 through fraudulent claims in the year 2015. The average lost for fraud is $43,752.03*(M = 43752.03, SD = 21812.68)*, which is $10,383.35 more *(p < .001)*. than average lost through legit claims *(M = 33368.68, SD = 29690.41)*.

The mean of total claim amount : $114920.00000 in year 2015.

Pre-processing :

Nominal variables were Label Encoder, and the data set was split into 75% train and 25% test set, stratified on fraud reported.
Using Z-Score : (Z-scores are measures of an observation's variability and can be used by traders to help determine market volatility.)
Apply Z-score gives clear prediction of fraud.

Models:

Different classifiers were used in this project:
- Decision Tree
- Extra Trees
- Random forest

After that apply cross-validation:

Cross-validation

```
from sklearn.model_selection import cross_val_score

scr = cross_val_score(dt, x, y, cv=5)
print("Cross Validation score of DecisionTree model is:", scr.mean())

scr = cross_val_score(rf, x, y, cv=5)
print("Cross Validation score of RandomForestRegressor model is:", scr.mean())


scr = cross_val_score(ex_reg, x, y, cv=5)
print("Cross Validation score of ExtraTreesRegressor model is:", scr.mean())
```

```
Cross Validation score of DecisionTree model is: 0.7762353672433442
Cross Validation score of RandomForestRegressor model is: 0.7802755620014503
Cross Validation score of ExtraTreesRegressor model is: 0.7640111882316379
```

Now Getting the ROC and AOC Curve :

The ROC curves for models for the test data are superimposed to obtain a visual comparison of the 'classifiers' performance on new data points.

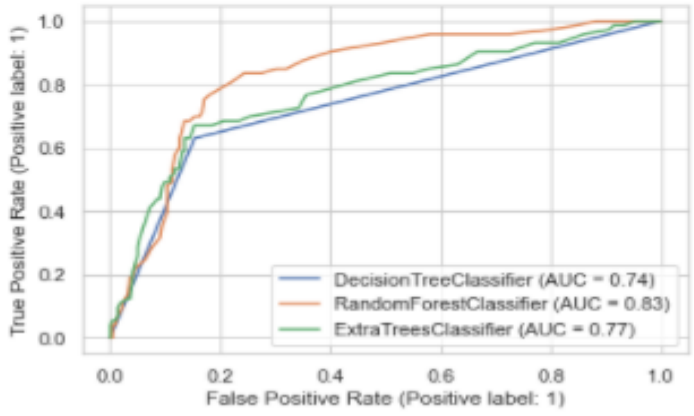
```
disp = plot_roc_curve(dt, x_test, y_test)

plot_roc_curve(rf, x_test, y_test, ax=disp.ax_)

plot_roc_curve(ex_reg, x_test, y_test, ax=disp.ax_)

plt.legend(prop={"size":11}, loc="lower right")

plt.show()
```



Using Hyper-parameter:
Hyperparameter tuning and selection was done for all the models using Randomized Search. Due to the number of parameters and models that were ran, Randomized Search is a faster more efficient choice as compared to gridsearch. the model with its selected hyperparameters were fitted on the training set.
Cross validation accuracy scores, accuracy scores on training set, accuracy scores on test set, sensitivity, specificity, precision, F1score and ROC AUC was computed and printed as shown in the table below.

Below summarizes the top 20 ranked by feature importance, based on weight. Incident severity is once again amongst the most important.

```
from sklearn.model_selection import GridSearchCV
from pprint import pprint
pprint(rf.get_params())
```

```
{'bootstrap': True,
 'ccp_alpha': 0.0,
 'class_weight': None,
 'criterion': 'gini',
 'max_depth': None,
 'max_features': 'auto',
 'max_leaf_nodes': None,
 'max_samples': None,
 'min_impurity_decrease': 0.0,
 'min_impurity_split': None,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'n_estimators': 100,
 'n_jobs': None,
 'oob_score': False,
 'random_state': None,
 'verbose': 0,
 'warm_start': False}
```

Models had much lower F1 and AUC scores and thus may have hindered performance of the ensemble.
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 86%.

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

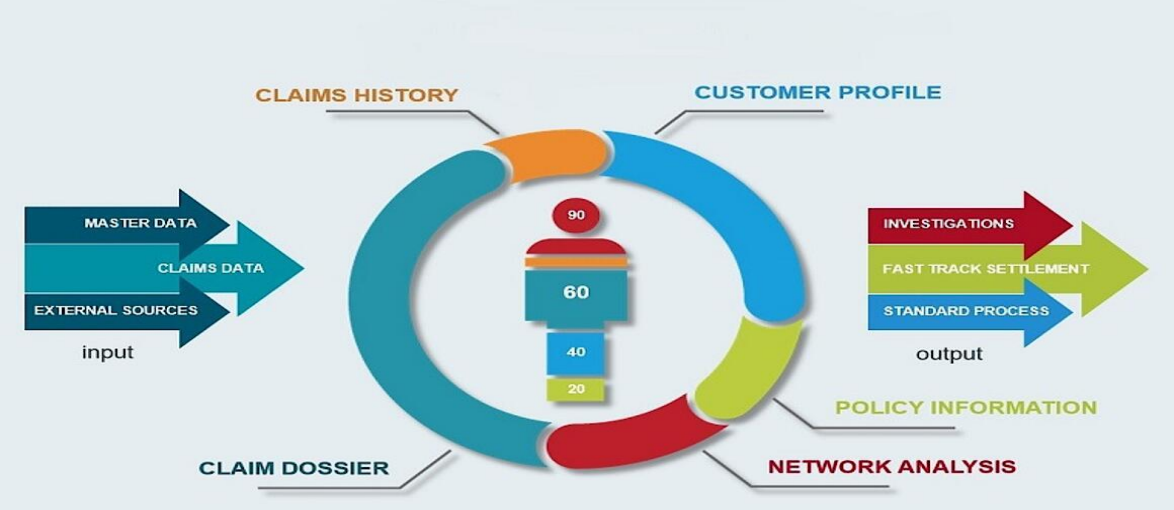
```
0.7932203389830509
[[188  34]
 [ 27  46]]
              precision    recall  f1-score   support

           0       0.87      0.85      0.86       222
           1       0.57      0.63      0.60        73

    accuracy                           0.79       295
   macro avg       0.72      0.74      0.73       295
weighted avg       0.80      0.79      0.80       295
```

Final Model:

The model had a training accuracy score of 0.797 and a test accuracy of 0.864. The high accuracy score hint of a low bias (only a hint as accuracy is not a good measure bias in imbalance problems). An accuracy score difference between train and test is relatively small. Thus, this model can be said to have low variance and is generalizable on unseen data.



Conclusion:

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 2012, compared with a range of $4.3 billion to $5.8 billion in 2002.
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.

```
# Saving the model

import joblib
joblib.dump(mod_dt_class,"Insurance_Claim_Fraud.pkl")
```

```
['Insurance_Claim_Fraud.pkl']
```

Findings:

Firstly, this study is restricted by its small sample size. Statistical models are more stable when data sets are larger. It also generalizes better as it takes a bigger proportion of the actual population. Furthermore, the data only capture incident claims of 3 states from 01 January 2015 to 01 March 2015. This means that we do not know the proportion of auto insurance policy holder who had no incidents compared to those who had incidents. We are also restricted to incidents between 2 months which may not be an accurate picture of the year. This is important as certain time of the year may correlate to higher incident rates such as St. Patrick's Day or other holidays.

Future studies may investigate acquiring a larger data set with multiple years. However, due to the sensitive nature of fraud and confidential information tagged to such data, this may remain a challenge.