INSURANCE CLAIM FRAUD DETECTION

**Introduction :**

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. Submitting claims for injuries or damage that never occurred and staging accidents.

**Problem Statement :**

The goal of this project is to build a model that can detect car insurance fraud claims .

**About Dataset :**

The data has 15420 insurance claims information in it . The challenge behind fraud detection in machine learning is that frauds are far less common as compared to legit insurance claims. Only 5.9% of the claims are fraudulent . This type of problems is known as imbalanced class classification.

**Hypothesis :**

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. This is measured by the F1 score .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. As a compulsory criterion, the ROC AUC must be above 0.50.

**Insights :**

* There are 5.9% of fraudulent claims in the data , the data is highly imbalanced .
* Most of the customers own Toyota , Honda , Pontiac , Mazda cars .
* Majority of the claims are received from the urban area , there are just 10% of claims from the rural area .
* Most claims were received on the weekdays , very few on weekends .
* Month doesn’t show any pattern , the claims throughout all the months in an year are the same .
* From the data it is also observed that , most of the claims are from Male customers , only 15% of the claims are from Female claimants .
* Majority of the claimants are married and a few are single , it seems to be that widow / divorced people are least involved in claiming the insurance or driving a car .
* majority of the times , damage to the car is caused by the policy holder, although 27% of claims have entered that the damage was caused due to third party .
* It can be observed that , Sedan owners are much likely to claim for insurance . However , most of the sports car owners tend to claim for collision policy type and most of the utility car owners tend to claim for All Perils policy type .
* Most of the claims own the cars of the price ranging from 20k to 29k , very few people own cars above 40k price .
* It is also observed that the Claims are received from the claimants after minimum of 30days of signing in for the policy . If not it is suspicious that they are fraudulent claims .
* Most of the claimants have either claimed none or up to 4 claims in the past .
* Many of the customers have vehicles of age 7 years .
* Most of the policy holders are 31 to 40 years old .
* Most of the claimants preferred not to file a police report , just 2.7% of the claimants have reported to the police about the accident .
* Not even 1% of the claimants had witness present in the accident .
* 99% of the claimants opted for external agent for claiming the insurance.
* Most of them did not raise address change claim , though they have changed the address it does not effect Fraud\_found variable .
* Maximum customers had owned only 1 car .
* The Deductible amount has been 400 for 99% of the claimants
* We have equal no.of drivers of all different ratings in the data .

**Model :**

* Target Variable is Fraud Found .
* Month and week have no significant relation with the target variable . Hence they can be dropped for training the model .
* Sedan – All Perils has the highest percentage of fraudulent claims in it.
* Policy holder IDs are all unique , there is no correlation of this column with the target variable .
* Age of vehicle effects the target variable in a significant way , the cars with age greater than 7 years are suspicious of a fraudulent claim .
* Witness present does not effect the target variable , since in maximum claims , witness is not present .
* No.of suppliments effects the target variable , different proportions of fraud claims can be seen in different categories .
* Address change claim there is no effect , since almost all of the claims haven’t claimed address change .
* Similarly no.of cars has no significant effect on the target variable , hence can be dropped from the target variable .

The project compares the results of different Machine learning techniques:

• Logistic Regression

• SVC

• Random Forest

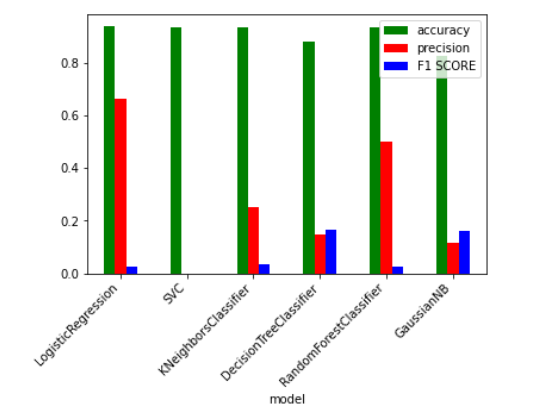
• KNN

• Naive Bayes

• Decision Tree

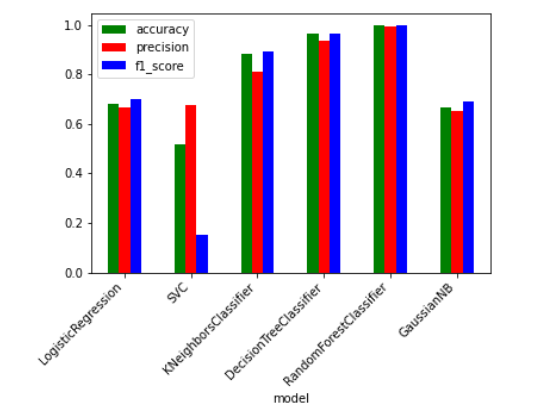
Results before upsampling :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.no | Model name | Accuracy | Precision score | F1 score |
| 1 | Logistic regression | 93.96% | 66.66% | 91% |
| 2 | SVC | 93.92% | 0.0.% | 91% |
| 3 | KNN | 93.69% | 25% | 91% |
| 4 | Decision Tree | 88.54% | 16.08% | 89% |
| 5 | Random Forest | 93.88% | 37.5% | 91% |
| 6 | Gaussian NB | 83.02% | 11.61% | 86% |



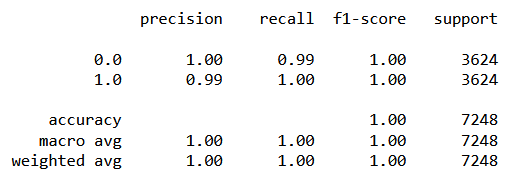
Results after upsampling :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.no | Model name | Accuracy | Precision score | F1 score |
| 1 | Logistic regression | 68% | 67 % | 68% |
| 2 | SVC | 52% | 68.% | 41% |
| 3 | KNN | 88% | 81% | 88% |
| 4 | Decision Tree | 96% | 93% | 96% |
| 5 | Random Forest | 99% | 99% | 99% |
| 6 | Gaussian NB | 67% | 67% | 67% |



**CONCLUSION :**

* Among all the models , random forest classifier was giving optimum results .
* For handling the class imbalance ,we did upsampling of the data .
* We obtained



* Confusion matrix :

