Machine Learning

Fraud Detection Models Report

Mobin Nesari
Dr. Hadi Farahani
May 8, 2023

Introduction:

Vehicle insurance fraud is a type of crime in which individuals collaborate to make deceitful or exaggerated claims regarding property damage or personal injuries that arise from an accident. This wrongful act can take many forms, such as staged accidents where criminals intentionally orchestrate a collision, phantom passengers who falsely claim to have suffered injuries despite not being present during the accident, and fraudulent personal injury claims where the extent of the injuries is exaggerated beyond reality.

General Description :

For this task, I've decided to create 10 different models to classify records to detect if they are fraud or not. Models which have been used in this task are XGBoost, Decision Tree, Random Forest, Balanced Random Forest, AdaBoost, AdaGradient, MLP, Support Vector Classification, K Nearest Neighbors and Logistic Regression. All these models have been fitted on normal, over and under sampling methods and then their performance are boosted by feature selection and hyper tuning.

General Results:

model	accuracy	recall	precision	f1	roc_auc
xgboost over	0.931907	0.086486	0.280702	0.132231	0.536172
dt over	0.889754	0.254054	0.188755	0.216590	0.592187
rf over	0.938067	0.059459	0.392857	0.103286	0.526798
blf over	0.920558	0.145946	0.236842	0.180602	0.557968
adaboost over	0.917964	0.091892	0.166667	0.118467	0.531286
gradientb over	0.927691	0.081081	0.220588	0.118577	0.531399

model	accuracy	recall	precision	f1	roc_auc
mlp over	0.940013	0.000000	0.000000	0.000000	0.500000
svc over	0.926719	0.086486	0.219178	0.124031	0.533412
KNN over	0.739300	0.270270	0.069541	0.110619	0.519751
LR over	0.837873	0.270270	0.120482	0.166667	0.572182
xgboost Under	0.824903	0.681081	0.207578	0.318182	0.757581
dt Under	0.773671	0.540541	0.140252	0.222717	0.664544
rf Under	0.834306	0.567568	0.195896	0.291262	0.709448
blf Under	0.683528	0.918919	0.150309	0.258359	0.793713
adaboost Under	0.805123	0.545946	0.163430	0.251557	0.683804
gradientb Under	0.820363	0.632432	0.194030	0.296954	0.732394
mlp Under	0.932879	0.048649	0.225000	0.080000	0.518978
svc Under	0.904345	0.113514	0.138158	0.124629	0.534163
KNN Under	0.831388	0.221622	0.098321	0.136213	0.545961
LR Under	0.829442	0.351351	0.138004	0.198171	0.605652
xgboost Normal	0.934825	0.070270	0.309524	0.114537	0.530133
dt Normal	0.898508	0.227027	0.198113	0.211587	0.584193
rf Normal	0.939689	0.000000	0.000000	0.000000	0.499828
blf Normal	0.662776	0.908108	0.141058	0.244186	0.777614
adaboost Normal	0.936446	0.010811	0.133333	0.020000	0.503163
gradientb Normal	0.940337	0.010811	0.666667	0.021277	0.505233
mlp Normal	0.940013	0.000000	0.000000	0.000000	0.500000
svc Normal	0.940013	0.000000	0.000000	0.000000	0.500000
KNN Normal	0.940013	0.005405	0.500000	0.010695	0.502530

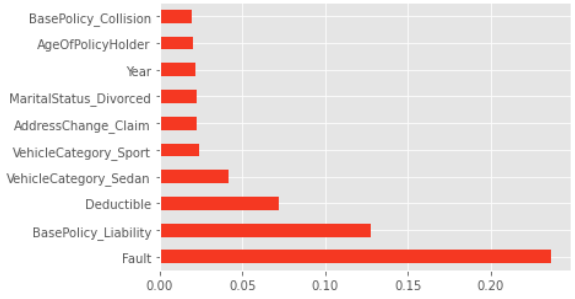
Table1 : Models General Report

As you can see in Table 1, all models have been trained on three different sampling methods and their accuracy, recall, precision, F1 and ROC / AUC have been reported. One of the novelties in this type of reporting is you can choose best model by your case. For instance, if you want to be sure that

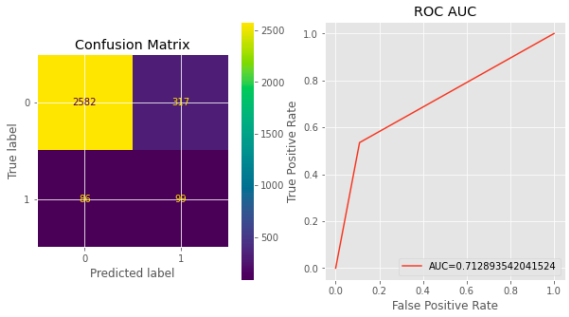
model can detect all frauds in cost of some false positive cases, you can choose the model with best precision and accuracy instead of best recall.

Hyper Tuning:

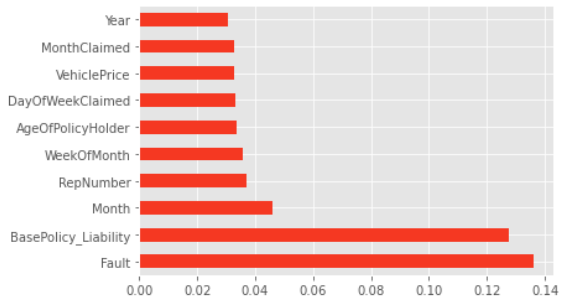
In this section, four tree based models, XGBoost, Decision Tree, Random Forest and Balanced Random Forest has been fine tuning with their number of estimators, max depth and learning rate as a given config. In the following, ROC/AUC and Confusion Matrix plots with feature importance for these four models has been shown:



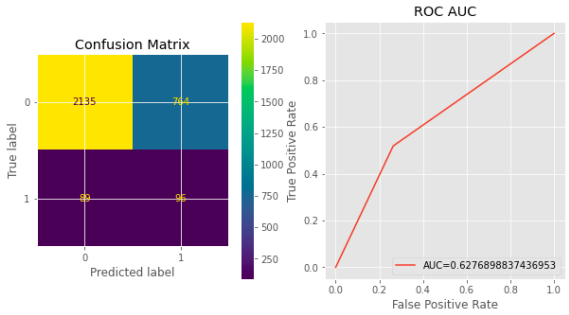
XGBoost Feature Importance



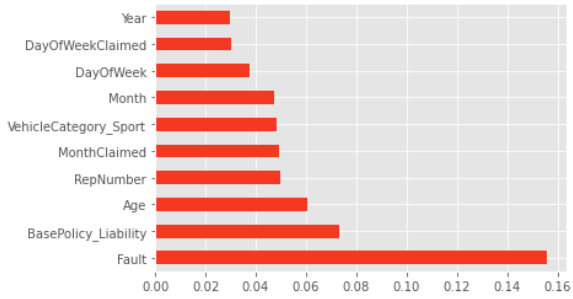
XGBoost Confusion Matrix and ROC/AUC Plot



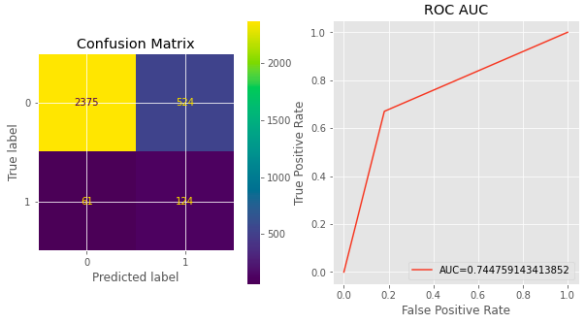
Decision Tree Feature Importance



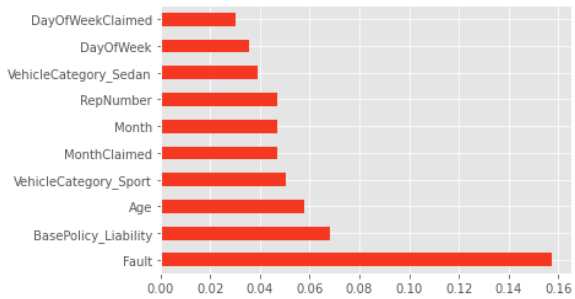
Decision Tree Confusion Matrix and ROC/AUC Plot



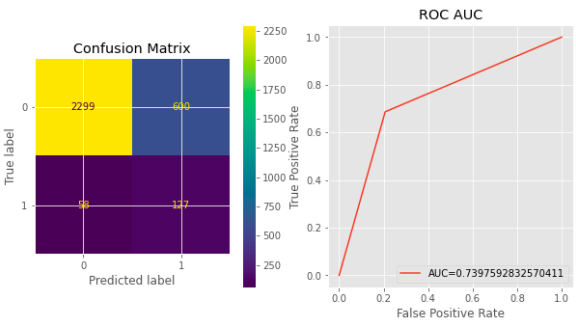
Random Forest Feature Importance



Random Forest Confusion Matrix and ROC/AUC Plot



Balanced Random Forest Feature Importance

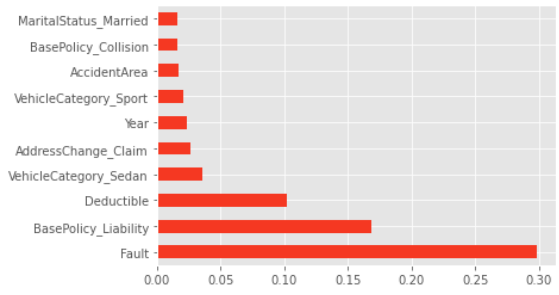


BRF Confusion Matrix and ROC/AUC Plot

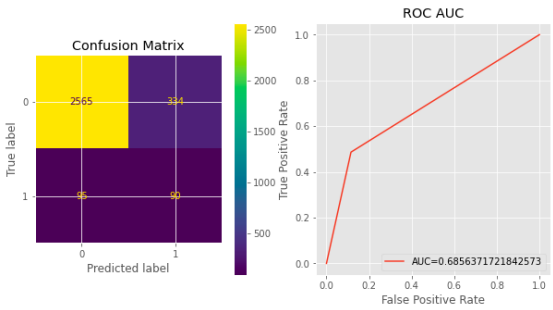
As you can see, all models has been decided that Fault is the most important feature in deciding about fraud cases which is logically true. The best Area Under Curvature is for Random Forest model with 0.7447.

Feature Selection and Hyper Tuning:

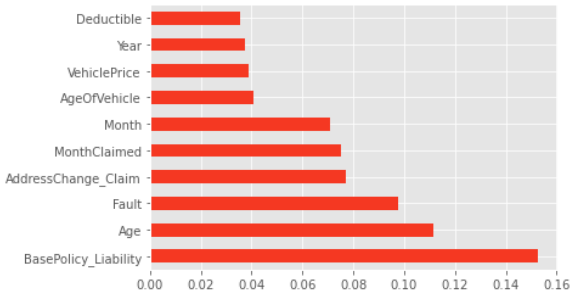
In this section, some obvious unimportant features has been removed from dataset to avoid deceiving model with distractions. ‘Days_Policy_Claim’, ’DayOfWeek’, ’WitnessPresent’, 'WeekOfMonthClaimed', 'DayOfWeekClaimed', 'DriverRating', 'WeekOfMonth', 'NumberOfCars' and ‘RepNumber' are unimportant features due to obvious facts and recent trained models’ importance features. In the following, feature importance, confusion matrix and ROC/AUC plot has been shown.



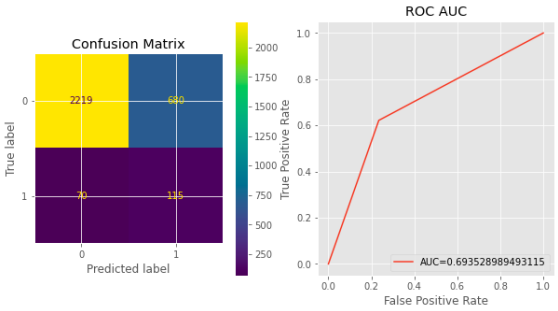
XGBoost Feature Importance



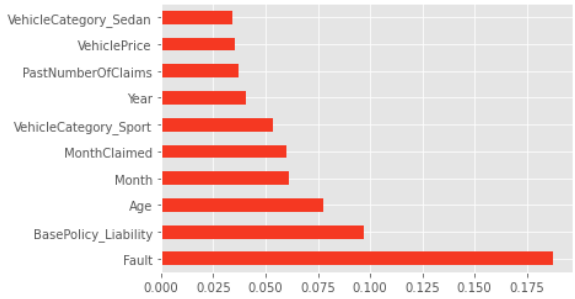
XGBoost Confusion Matrix and ROC/AUC Plot



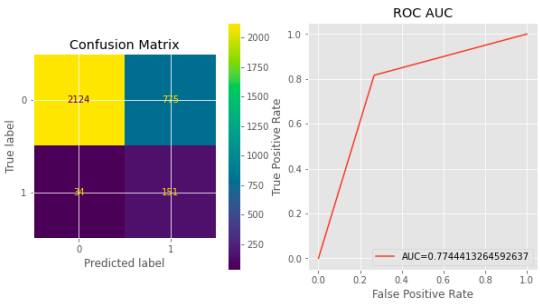
DT Feature Importance



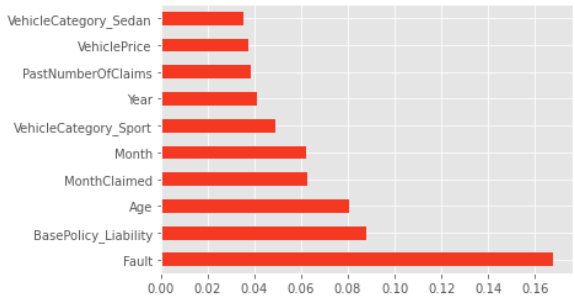
DT Confusion Matrix and ROC/AUC Plot



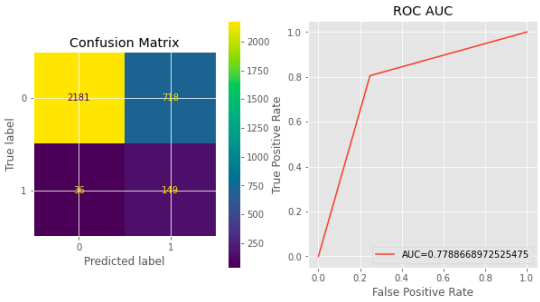
RF Feature Importance



RF Confusion Matrix and ROC/AUC Plot



BRF Feature Importance



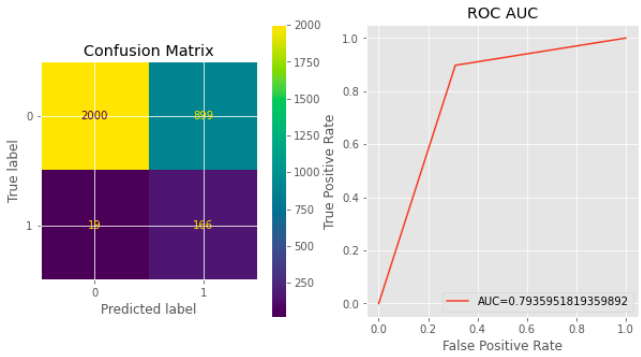
BRF Confusion Matrix and ROC/AUC Plot

As you can see, importance of features has been changed a little bit due to removing some unimportant features which take a small value for importance. Three out of four models agreed on second most important feature, BasePolicy_Liability. The highest AUC increased to 0.7788 and Balance Random Forest is the most reliable model after removing unimportant features.

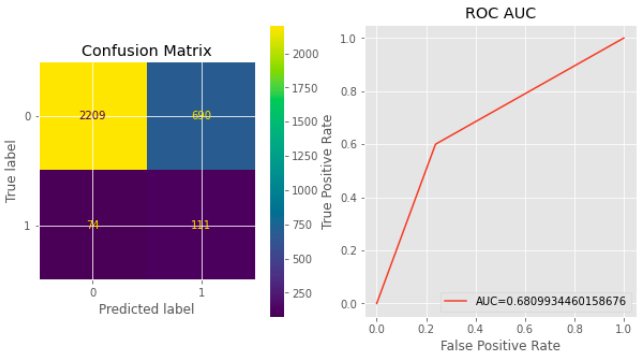
Predetermined Weights and Hyper Tuning:

In this section, due to unbalance dataset which we are working on, we tried XGBoost and Decision Tree models with predefined class weights and

thresholds. In the following, you can see confusion matrix and ROC/AUC plot of both of them.



XGBoost



Decision Tree

As you can see, XGBoost's AUC has been increased to 0.7935 which make it more reliable in this task.