FoML hackathon report

Sri Tejaswini Challa

November 2024

1 Data Preprocessing

I observed the dataset description and there are a few attributes that are duplicates and some are irrelevant to the target prediction. So, i dropped those attributes.

Then I removed the columns that have most of the values NAN and the remaining columns, I replaced the missing values with median.
Observing the dataset, the target attribute has imbalance of classes. Hence the model might be biased, so I used XGBoost because it helps in distinguishing different classes better than other models.

I tried other resampling methods like SMOTE, but it didn't improve f1 score much as it is just replicating the minority samples and not finding underlying trends better.

2 Model Performance

Performance of XGBoost model on splitting train set into train and validation set

	precision	recall	f1-score	support
low	0.33	0.15	0.21	4526
medium	0.35	0.17	0.23	4462
high	0.63	0.86	0.73	13526
accuracy			0.58	22514
macro avg	0.44	0.39	0.39	22514
weighted avg	0.52	0.58	0.52	22514

Figure 1: XGBoost

Performance of Random forest model on splitting train set into train and validation set

	precision	recall	f1-score	support
low	0.40	0.10	0.16	4526
medium	0.39	0.09	0.15	4462
high	0.62	0.94	0.75	13526
accuracy			0.60	22514
macro avg	0.47	0.38	0.35	22514
weighted avg	0.53	0.60	0.51	22514

Figure 2: Random forest

I ensembled these models together and the final f1 score on the validation set provided on kaggle was 0.405.