CS253 ML report

Rishikesh Sahil 220892

April 13, 2024

1 Methodology

In this section, we outline the methodology used in the assignment.

1.1 Data Preprocessing

loading the data from CSV file and encoding categorical features while defining both the features and the target variable.

1.2 Feature engineering

we defines a list of features (features) and the target variable (target). It then converts categorical variables into numeric format using LabelEncoder from scikit-learn.

1.3 Random Forest Model

initialising the RF model, conducted hyperparameter tuning via grid search, and assess its performance on the validation set.

1.4 Gradient Boosting Model

initialising the GB model, perform hyperparameter tuning, and evaluate its performance on the validation set.

1.5 Model Selection

selecting the model with the higher F1 score on the validation set as the final model.

1.6 Prediction and Submission

making predictions on the test set, converting them back to original labels, and saving the results as a CSV file for submission.

1.7 Data Visualisation

Feature importance plots for Random Forest and Gradient Boosting models. Confusion matrices for both models on the validation set. We also delved deep in the candidates wealth and their criminal record.

2 Experiment Details

In this section, we provide details of the experiments conducted, including the models used and their hyperparameters.

2.1 Model Selection

Table summarizes the models used along with their hyperparameters.

Table 1: Hyperparameters Used in the Code

Model	**Hyperparameters**
Random Forest	`n_estimators`: [200, 500] `max_depth`: [None, 10] `min_samples_split`: [2, 5] `min_samples_leaf`: [1, 2]
Gradient Boosting	`n_estimators`: [200, 500] `learning_rate`: [0.01, 0.1] `max_depth`: [3, 5]

2.2 Data Insights

We generated various graphs and plots to gain insights into the data distribution, relationships between variables, and class distributions.

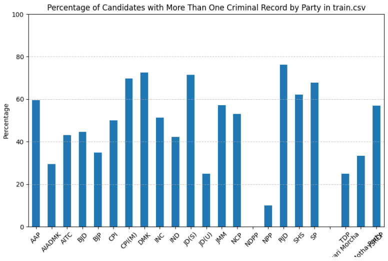

Figure 1: Percentage of criminal candidates of each party in training set.

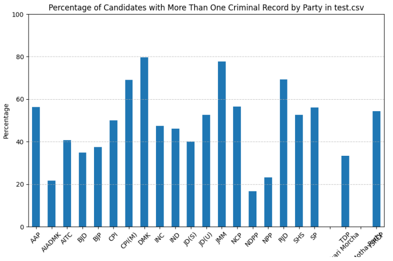

Figure 2: Percentage of criminal candidates of each party in testing set.

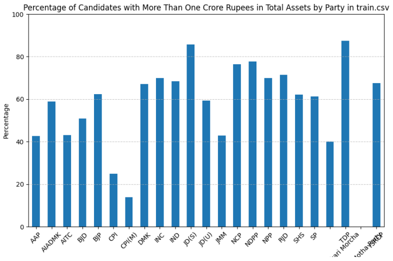

Figure 3: Percentage of candidates with more than one crore asset of each party in training set.

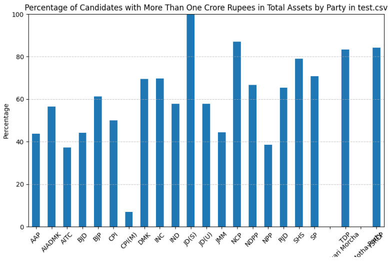

Figure 4: Percentage of candidates with more than one crore asset of each party in testing set.

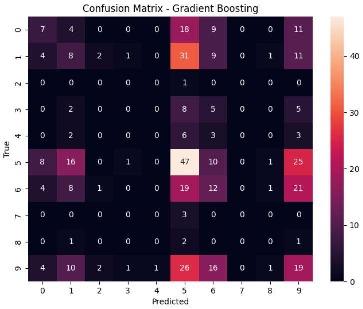

Figure 5: Confusion matrix of gradient boosting method

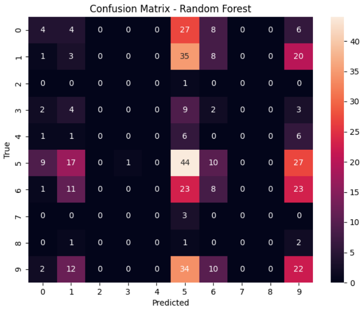

Figure 6: Confusion matrix of random forest method

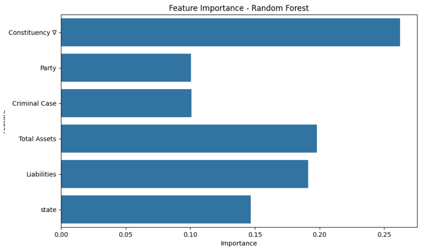

Figure 7: feature importace on random forest model

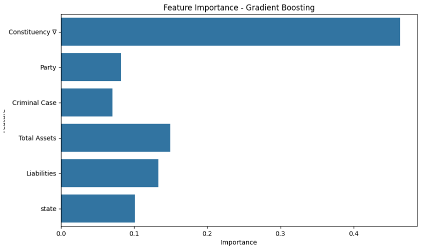

Figure 8: feature importace on gradient boosting model

2.3 Results

F1 score (public) = 0.23
Public Rank Leaderboard = 113

3 References

1. Link of RandomForestClassifier
2. Link of GradientBoostingCLassifier
3. Link of GridSearch

4 Github Repository

Link of Github repository