CS253 Python Assignment Report

Sankalp Mittal | 220963

April 2024

1 Introduction

This is the report for the CS253 Python Assignment. I have used various Multi-Class Classification methods to determine which gives the best f1-score. The code that is used has been uploaded to this repository. It contains four different models, and a train-validation split of 20%, that I used to test the best model, which was later removed so that the best model for testing could be trained.

2 Methodology

The dataset given had to be modified before it was fit for training, the following steps were taken

- Since the **name and constituency** features were unique for each entry they had to be *dropped*
- Replaced the *string* values in the **assets and liabilities** columns with the corresponding numerical values
- After that the **party and state** columns had object type values so *one-hot encoding* was done
- The education level values also had object type so, I *label encoded* them as **XGBoost** requires numerical values

3 Models Trained

3.1 KNN

The best parameters found out via *Grid Search* were

- **metric**: manhattan
- **n_neighbors**: 11
- **weights**: uniform

3.2 Random Forest

The best parameters found out via *Grid Search* were

- **bootstrap**: True
- **max_depth**: None
- **max_features**: log2
- **min_samples_leaf**: 1
- **min_samples_split**: 5
- **n_estimators**: 100

3.3 Decision Tree

The best parameters found out via *Grid Search* were

- **criterion**: gini
- **max_depth**: None
- **min_samples_leaf**: 4
- **min_samples_split**: 10

3.4 XGBoost

The best parameters found out via *Trails and Error* were

- **eta**: 0.3
- **max_depth**: 15
- **objective**: multi:softmax
- **num_class**: 10
- **eval_metric**: merror

4 Data Analysis

4.1 Party versus Criminal Record

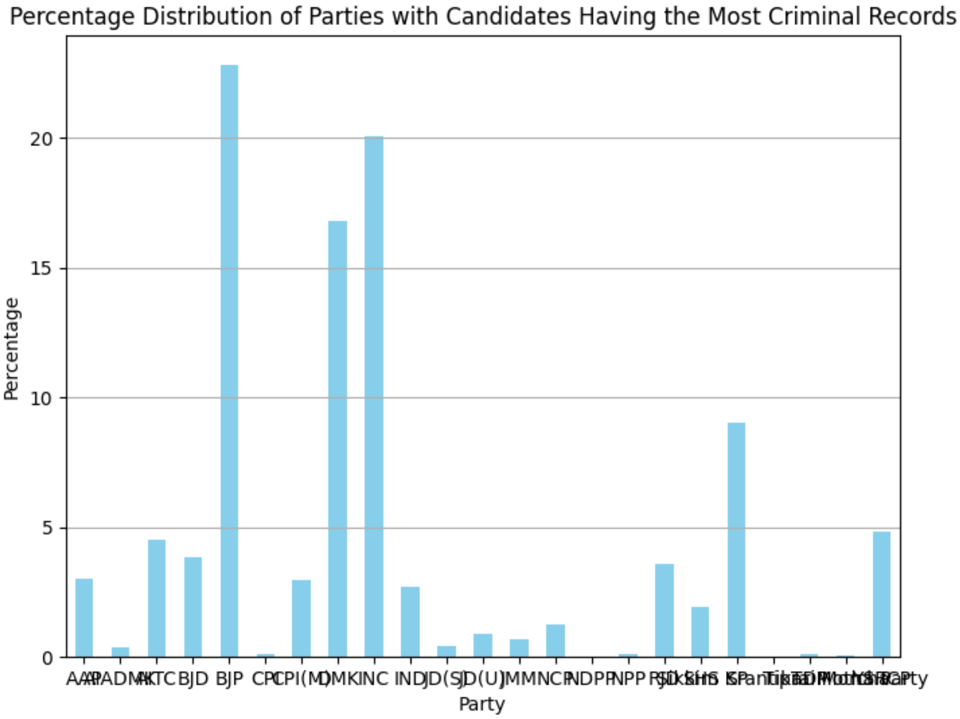

Figure 1: Party versus Criminal Records distribution

4.2 Party versus Wealth of Candidates

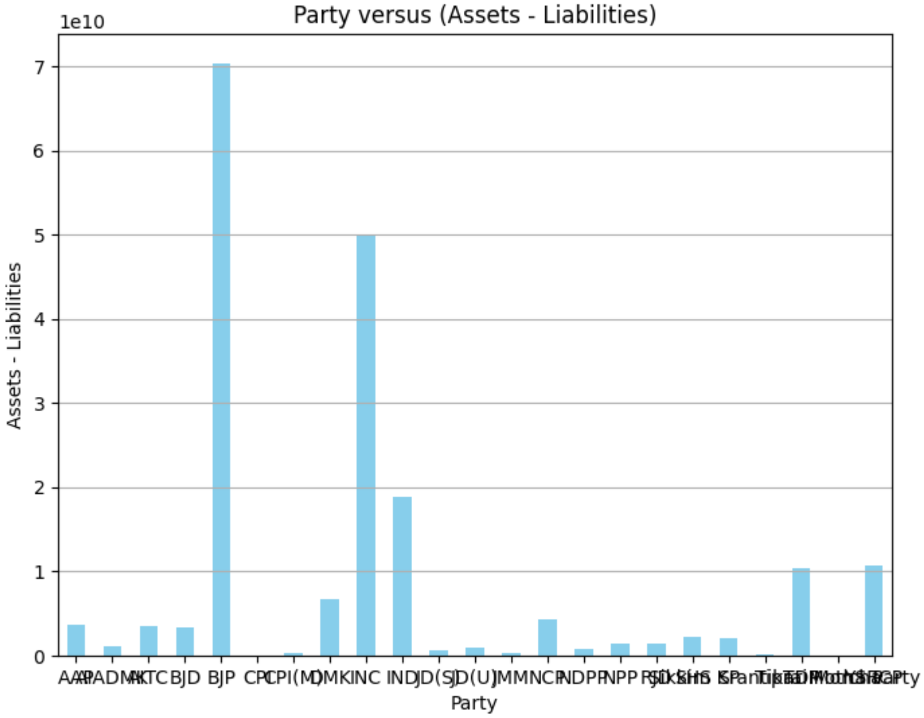

Figure 2: Party versus Criminal Records distribution

5 Results

5.1 Best F1 score

The best F1 scores were as follows

- **Public:** 0.23150
- **Private:** 0.25035

5.2 Ranking

There is a problem as the system did not consider my final(and best) submission, I had submitted on 23:58 which Kaggle took as late Otherwise with the other submissions, rankings were as follows

- **Public:** 179
- **Private:** 93

6 GitHub

The link to the repository is Github_Link

7 References

The following sources were used

1. KNN Documentation
2. Decision Tree Classifier Documentation
3. Random Forest Documentation
4. XGBoost Documentation