# PYTHON ASSIGNMENT REPORT

Aditi Khandelia

April 2024

This is the report submitted for the python assignment of the course CS253: Software Development and Operations, completed by Aditi Khandelia, under the guidance of Prof. Indranil Saha.

## Model Trained

Many experimentation techniques were applied for solving the PS:

1. Training with synthetic data obtained from CTGAN[1] [2]
2. Training with variable features:
   (a) Creating and using the 'Dr.' column
   (b) Dropping columns 'Total Assets' and 'Liabilities'
   (c) Dropping the 'Constituency' column

The best results were obtained with the approach of dropping 'Total Assets', 'Liabilities', and 'Constituency', followed by label encoding 'state' and 'Party'. 'Education' was also label encoded.
The model used was **Decision Tree Classifier** with hyper-parameters:

1. max_depth : 14
2. min_samples_leaf : 1
3. min_samples_split : 2

The final metrics were the following on the test split (20% of 'train.csv' ):

1. Accuracy: 0.2524271844660194
2. Precision: 0.22970790300788921
3. Recall: 0.2524271844660194
4. F1: 0.2401183976703066

On the same approach, different models gave different F1 scores on the test split:

| **Model** | **F1 Score** |
|---|---|
| Linear SVC | 0.2136 |
| Random Forest Generator | 0.2084 |
| K-Nearest Neighbours | 0.1902 |

Table 1: F1 scores of different models on the test split.

# Data Analysis

The following Data Analysis was done on "train.csv", which contains the following columns:

1. ID : Unique ID of each candidate
2. Candidate : Candidate's Name
3. Constituency : Name of the Candidate's Constituency
4. Party : Name of the Candidate's political party
5. state : Name of the Candidate's state
6. Total Assets : Value of the candidate's total assets
7. Liabilities : Value of the candidate's liabilities
8. Education : The level of Education of the candidate

The dataset has 2059 datapoints.

## Distribution of Labels

The following is the distribution of 'Education' Label in the train dataset.

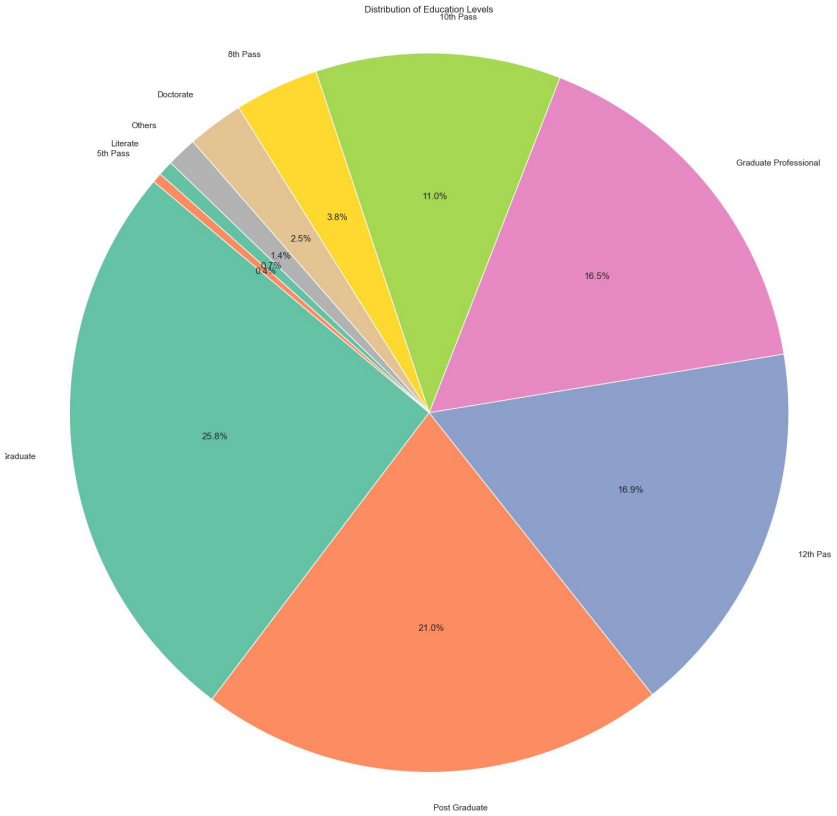


Figure 1: Labels in the Train Dataset

## State-wise Distribution of Labels

The following is the distribution of 'Education' label in the train dataset further classified under unique 'States'.

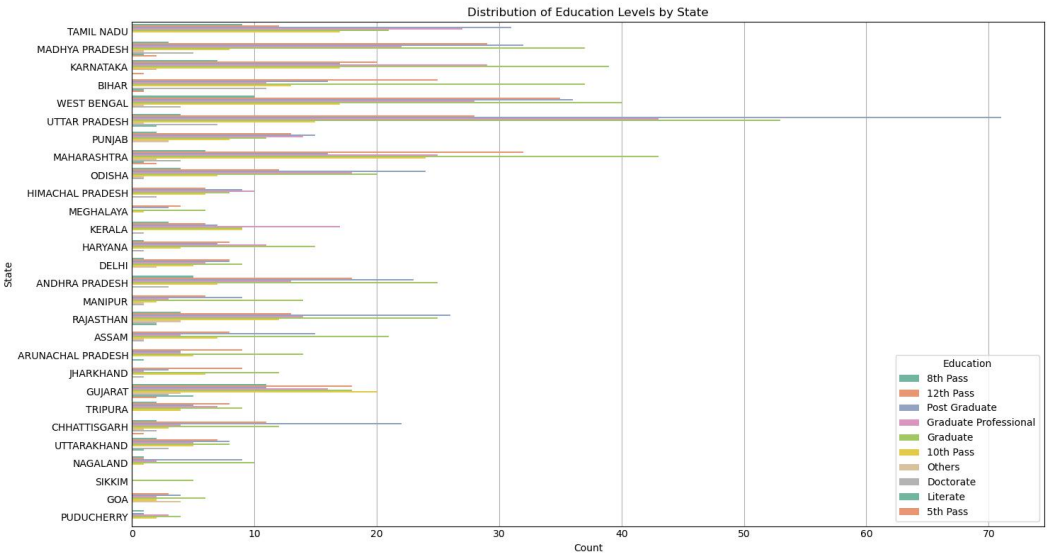


Figure 2: State-wise Distribution of Labels

## Party-wise Distribution of Labels

The following is the distribution of 'Education' label in the train dataset further classified under unique 'Party'.

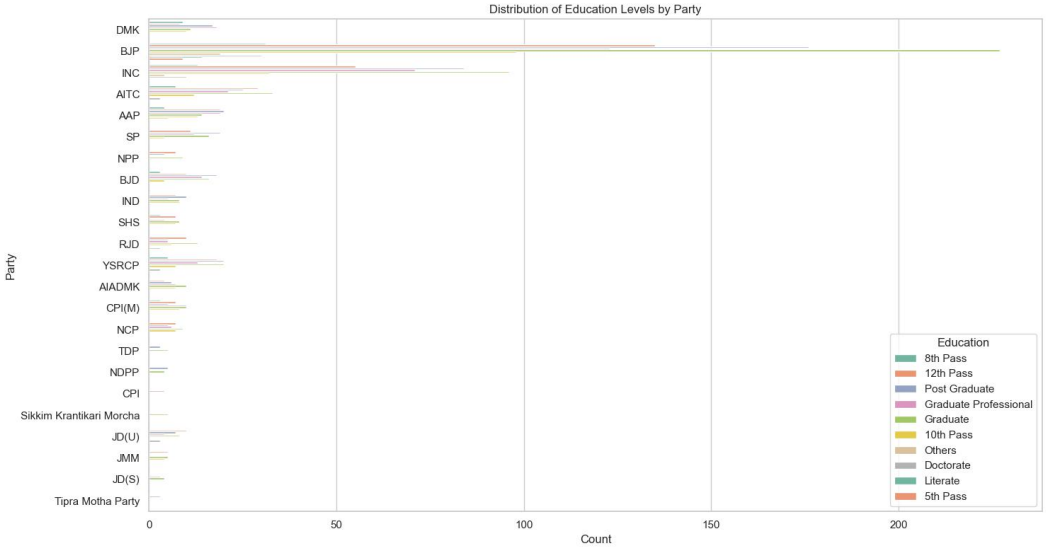


Figure 3: Party-wise Distribution of Labels

## Total Assets-wise Distribution of Labels

The following is the distribution of 'Education' label in the train dataset further classified under distribution of 'Total Assets'.

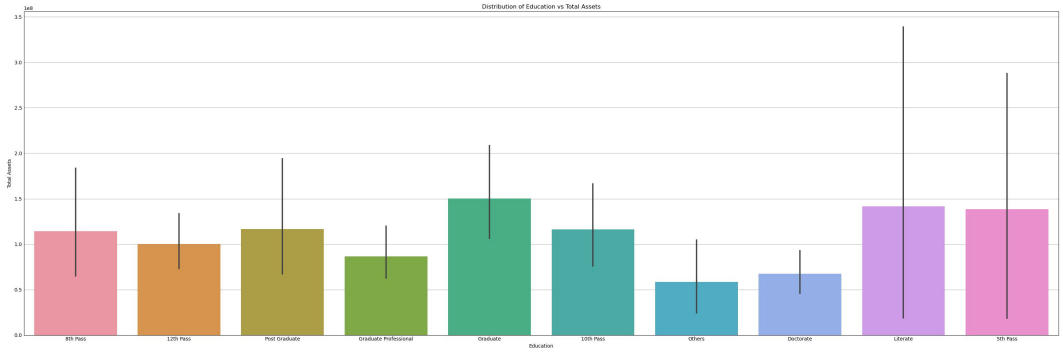


Figure 4: Total Assets-wise Distribution of Labels

## Liabilities-wise Distribution of Labels

The following is the distribution of 'Education' label in the train dataset further classified under distribution of 'Liabilities'.

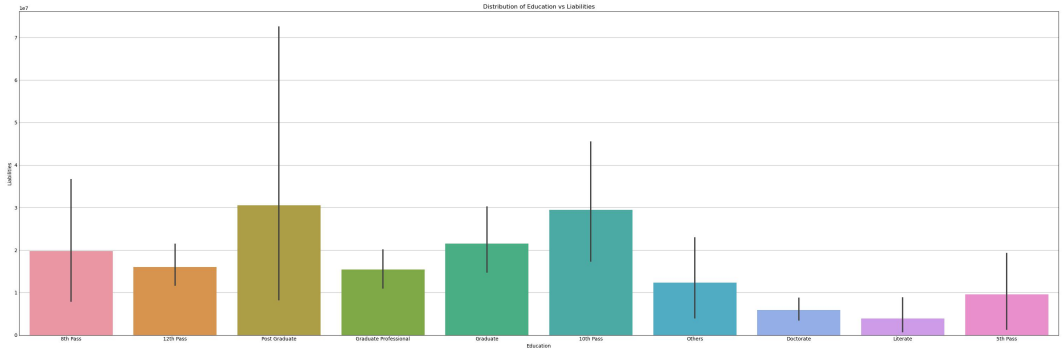


Figure 5: Liabilities-wise Distribution of Labels

## Net Income-wise Distribution of Labels

The following is the distribution of 'Education' label in the train dataset further classified under distribution of 'Net Income'.

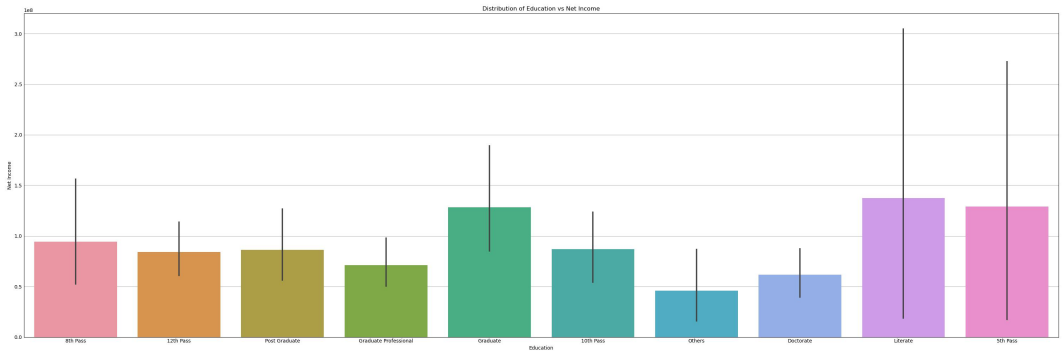


Figure 6: Net Income-wise Distribution of Labels

## Criminal Cases-wise Distribution of Labels

The following is the distribution of 'Education' label in the train dataset further classified under distribution of 'Criminal Cases'.

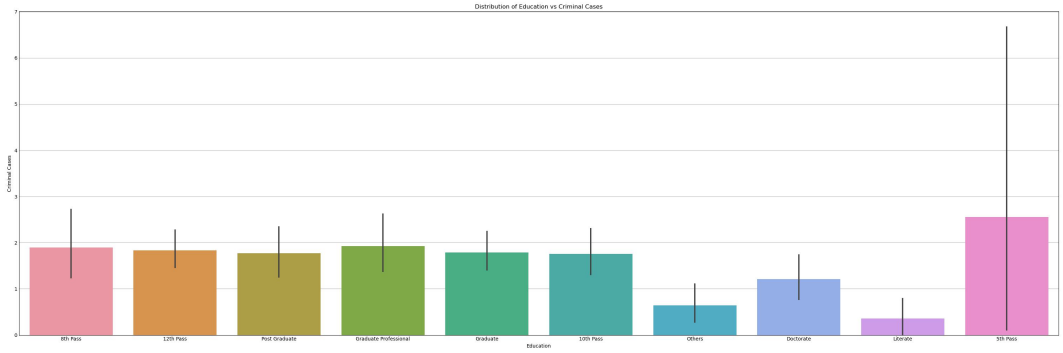


Figure 7: Criminal Cases-wise Distribution of Labels

## Distribution of Criminal Cases Across Political Parties

### Percentage of Members with High criminal Case Record w.r.t. to their Own Political Party

The following plot showcases the percentage of members of a particular political party that have criminal case record higher than the mean no. of criminal cases, which is 1.7775619232637203

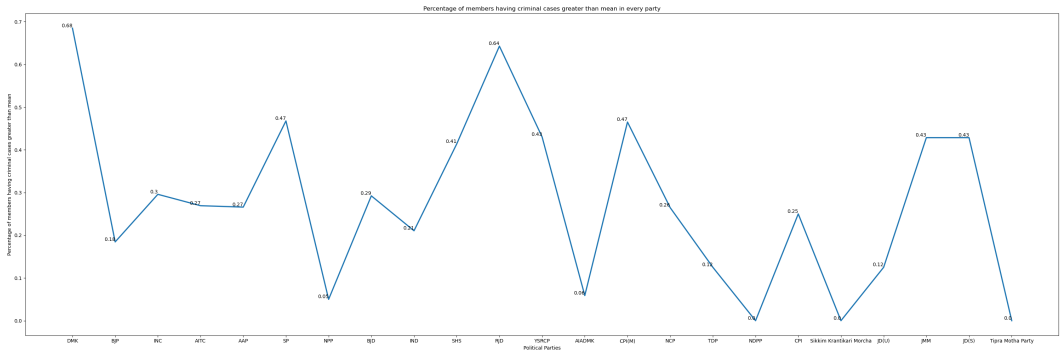


Figure 8: Percentage of Members with High criminal Case Record

### Percentage of Members with High criminal Case Record w.r.t. all the Candidates having Criminal Record greater than the Mean

The following plot showcases the percentage of members of a particular political party that have criminal case record higher than the mean no. of criminal cases w.r.t. all the candidates with criminal records greater than the mean, which is 1.7775619232637203

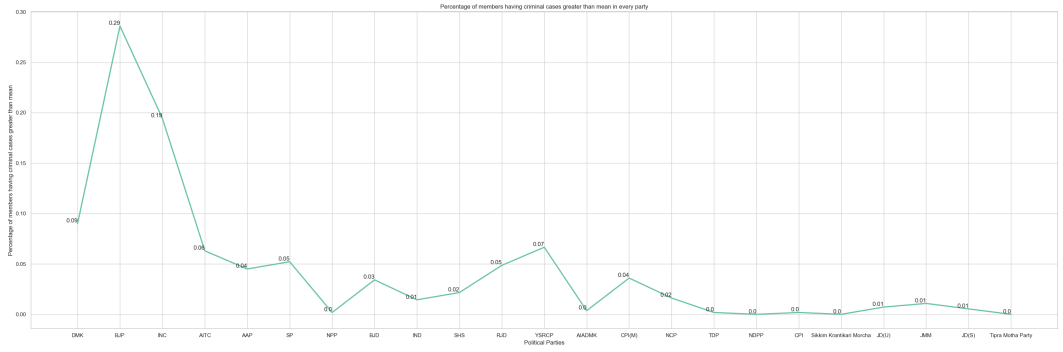


Figure 9: Percentage of Members with High criminal Case Record w.r.t. all the Candidates having Criminal Record greater than the Mean

### Individual Candidates

The following plot showcases the distribution of criminal cases for different political parties. It separately showcases the criminal case records for individual candidates.

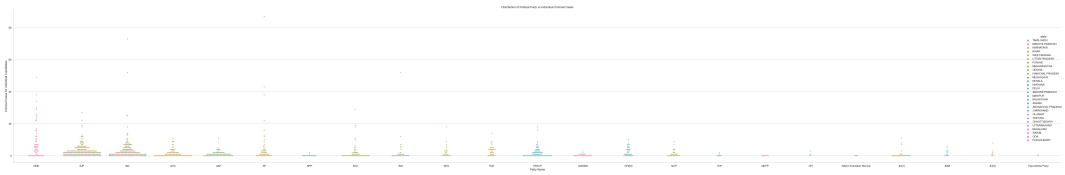

Figure 10: Distribution of Criminal Cases Across Political Parties - Individual Candidates

### Sum of All Candidates

The following plot showcases the distribution of criminal cases for different political parties. It showcases the sum of criminal cases for all candidates.

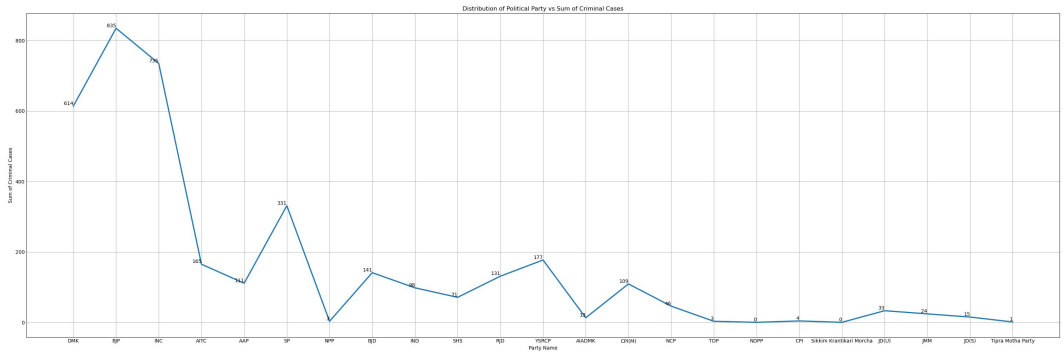


Figure 11: Distribution of Criminal Cases Across Political Parties - Sum of All Candidates

### Mean of All Candidates

The following plot showcases the distribution of criminal cases for different political parties. It showcases the mean of criminal cases for all candidates belonging to a particular political party. The error margins represent the min and max no of criminal cases for that particular political party.

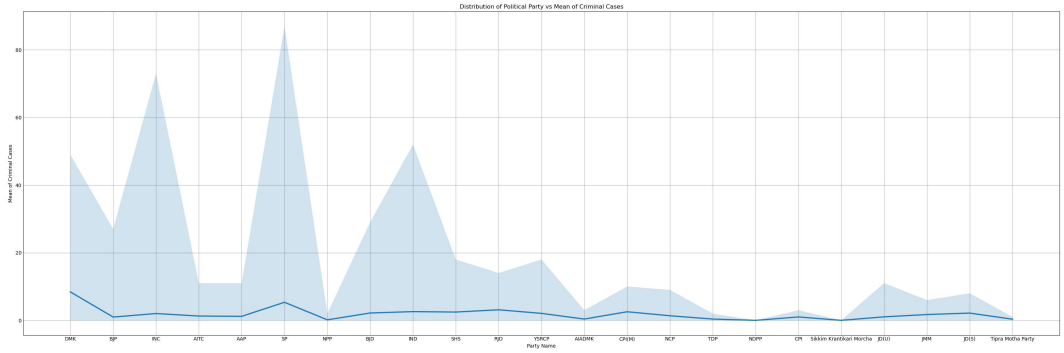


Figure 12: Distribution of Criminal Cases Across Political Parties - Mean of All Candidates

## Distribution of Total Assets Across Political Parties

### Percentage of Members with High Total Assets w.r.t. to their Own Political Party

The following plot showcases the percentage of members of a particular political party that have total assets higher than the mean total assets, which is 115599132.1029626

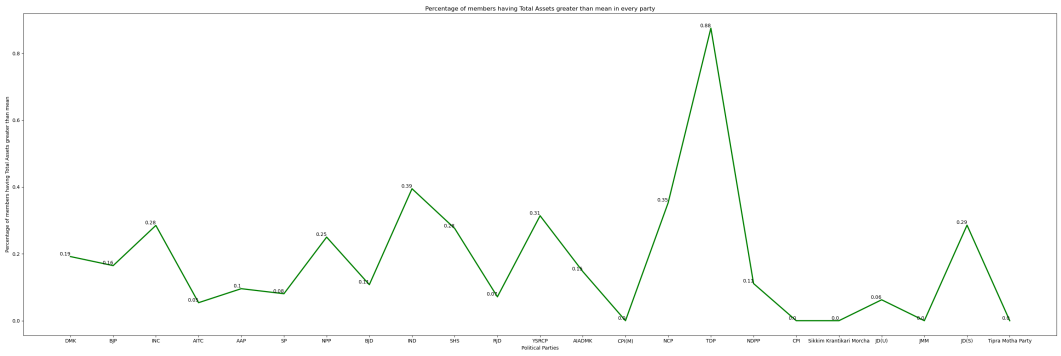


Figure 13: Percentage of Members with High Total Assets

### Percentage of Members with High Total Assets w.r.t. all the Candidates having Total Assets greater than the Mean

The following plot showcases the percentage of members of a particular political party that have total assets higher than the mean total assets, w.r.t. all such candidates across all political parties.

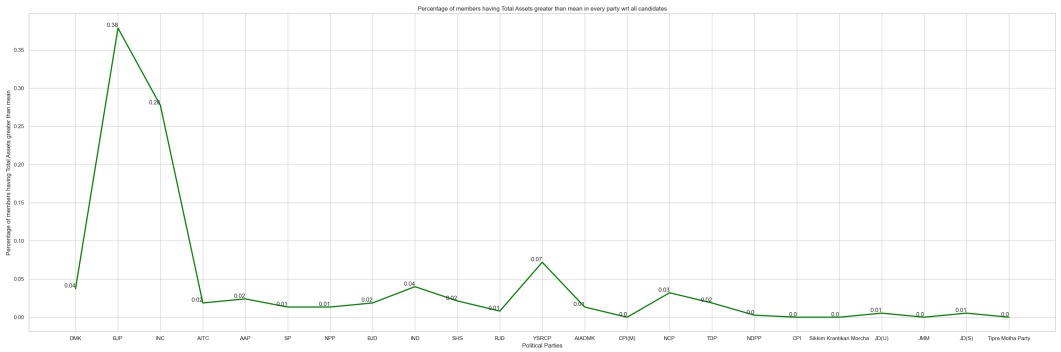


Figure 14: Percentage of Members with High Total Assets w.r.t. all the Candidates having Total Assets greater than the Mean

### Individual Candidates

The following plot showcases the distribution of Total Assets for different political parties. It separately showcases the Total Assets for individual candidates.

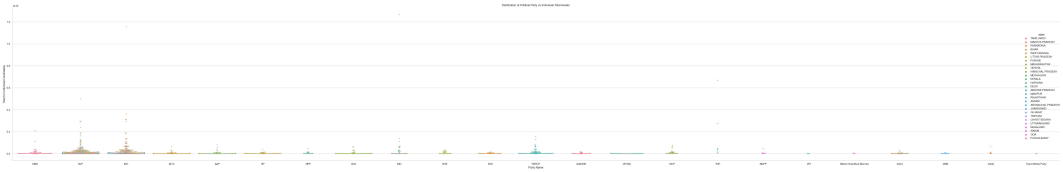

Figure 15: Distribution of Total Assets Across Political Parties - Individual Candidates

### Sum of All Candidates

The following plot showcases the distribution of Total Assets for different political parties. It showcases the sum of Total Assets for all candidates.

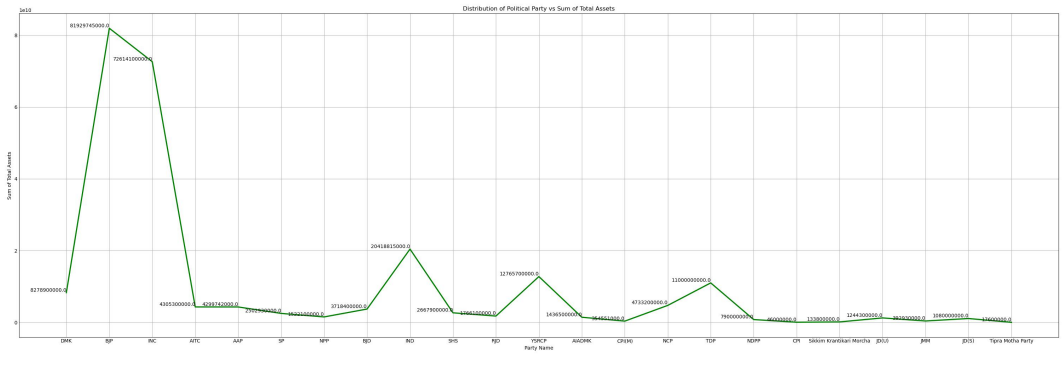


Figure 16: Distribution of Total Assets Across Political Parties - Sum of All Candidates

### Mean of All Candidates

The following plot showcases the distribution of Total Assets for different political parties. It showcases the mean of total assets for all candidates belonging to a particular political party. The error margins represent the min and max value of total assets for that particular political party.

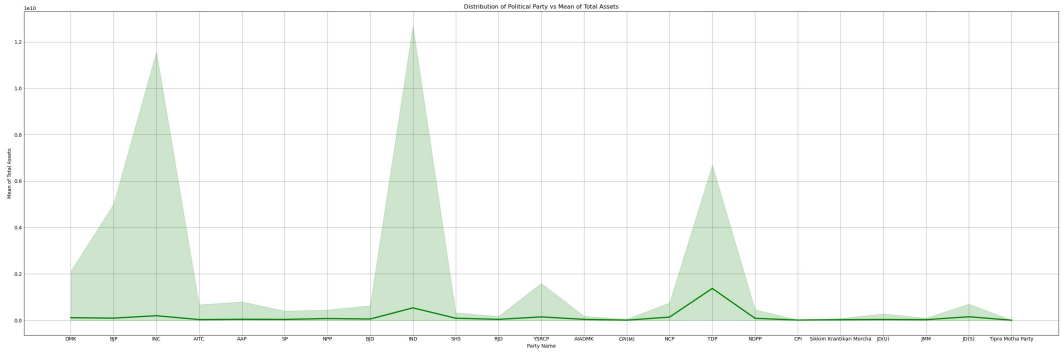


Figure 17: Distribution of Total Assets Across Political Parties - Mean of All Candidates

## Distribution of Liabilities Across Political Parties w.r.t. to their Own Political Party

### Percentage of Members with High Liabilities

The following plot showcases the percentage of members of a particular political party that have liabilities higher than the mean liabilities, which is 21590683.389995143

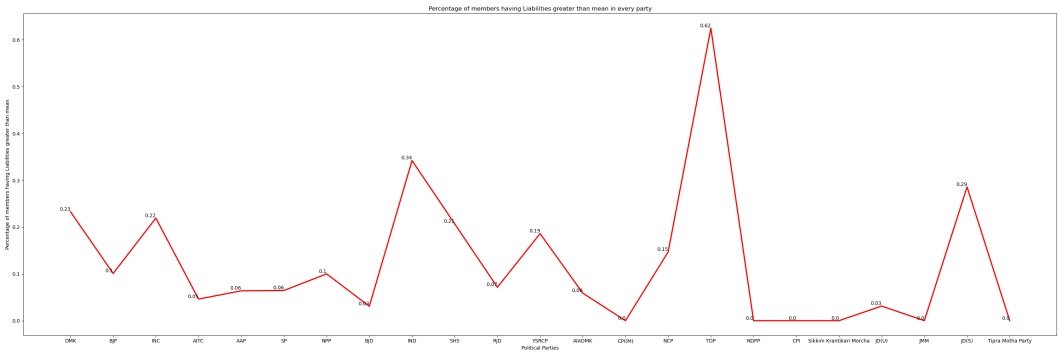


Figure 18: Percentage of Members with High Liabilities

### Percentage of Members with High Liabilities w.r.t. all the Candidates having Liabilities greater than the Mean

The following plot showcases the percentage of members of a particular political party that have liabilities higher than the mean liabilities wrt to all such candidates across all political parties.

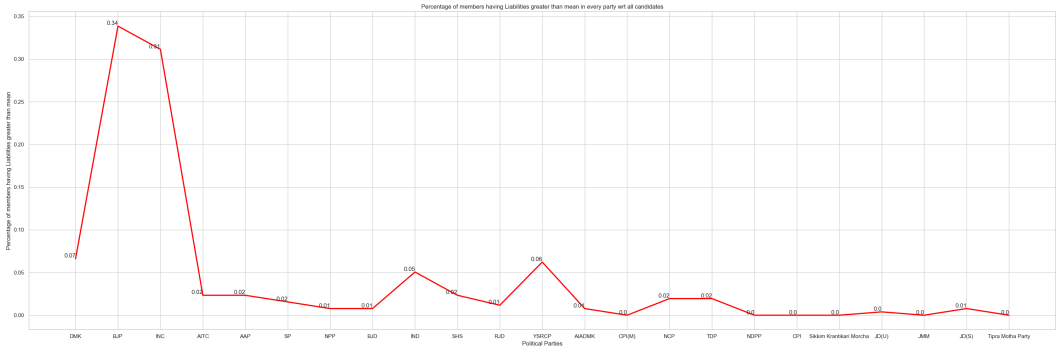


Figure 19: Percentage of Members with High Liabilities w.r.t. all the Candidates having Liabilities greater than the Mean

### Individual Candidates

The following plot showcases the distribution of Liabilities for different political parties. It separately showcases the Liabilities for individual candidates.

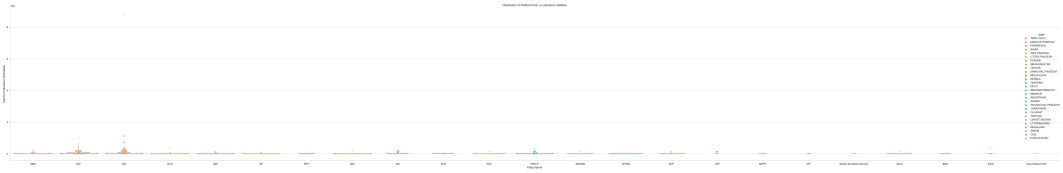

Figure 20: Distribution of Liabilities Across Political Parties - Individual Candidates

### Sum of All Candidates

The following plot showcases the distribution of Liabilities for different political parties. It showcases the sum of Liabilities for all candidates.

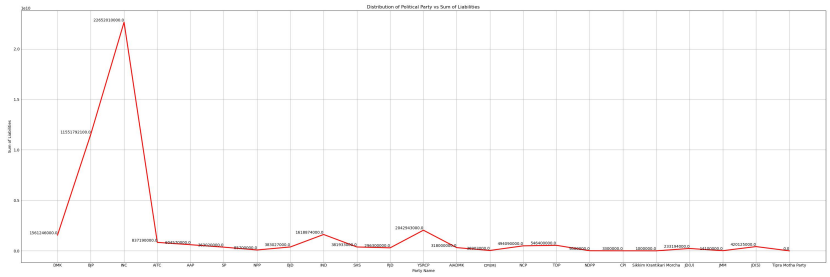


Figure 21: Distribution of Liabilities Across Political Parties - Sum of All Candidates

### Mean of All Candidates

The following plot showcases the distribution of Liabilities for different political parties. It showcases the mean of Liabilities for all candidates belonging to a particular political party. The error margins represent the min and max value of Liabilities for that particular political party.

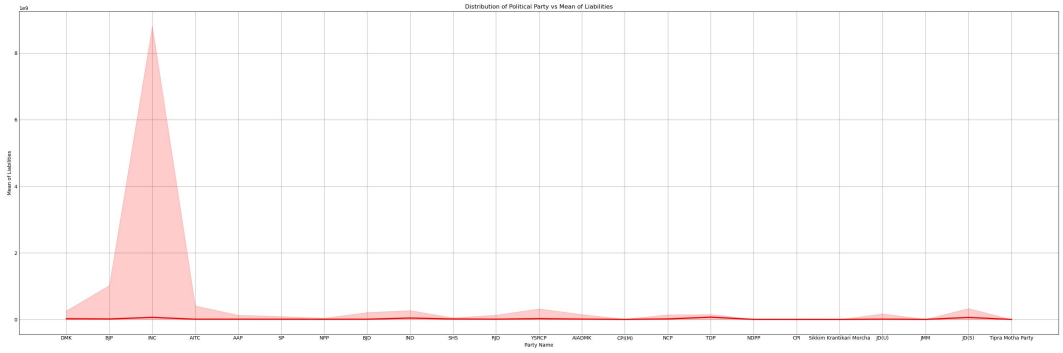


Figure 22: Distribution of Liabilities Across Political Parties - Mean of All Candidates

## Distribution of Net Income Across Political Parties

### Percentage of Members with High Net Incomes w.r.t. to their Own Political Party

The following plot showcases the percentage of members of a particular political party that have net income higher than the mean net income, which is 94008448.71296746

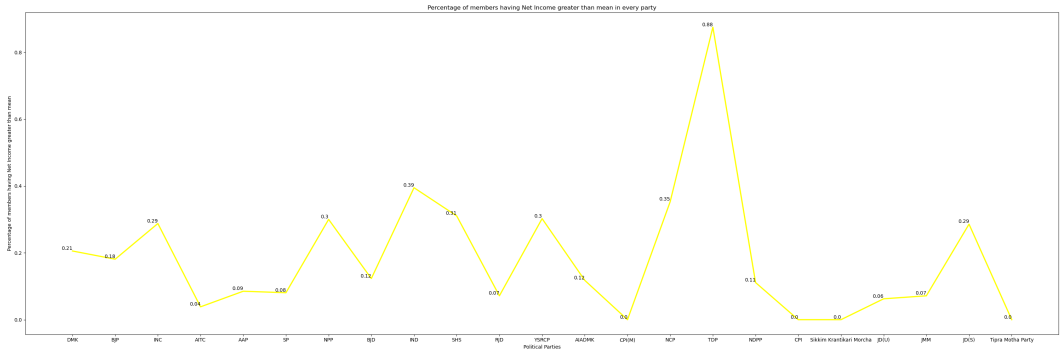


Figure 23: Percentage of Members with High Net Incomes

### Percentage of Members with High Net Incomes wrt all such Candidates

The following plot showcases the percentage of members of a particular political party that have net income higher than the mean net income w.r.t. all such candidates that have net income higher than the mean net income, which is 94008448.71296746

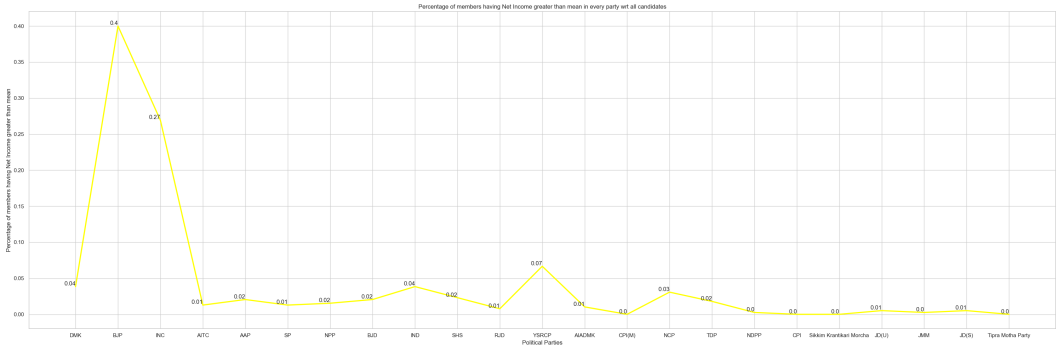


Figure 24: Percentage of Members with High Net Incomes wrt all such Candidates

### Individual Candidates

The following plot showcases the distribution of Net Income for different political parties. It separately showcases the Net Income for individual candidates.

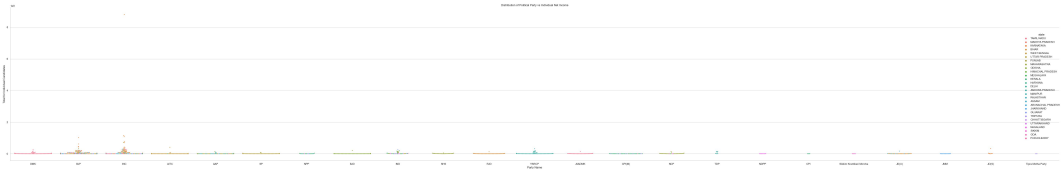

Figure 25: Distribution of Net Income Across Political Parties - Individual Candidates

### Sum of All Candidates

The following plot showcases the distribution of Net Income for different political parties. It showcases the sum of Net Income for all candidates.

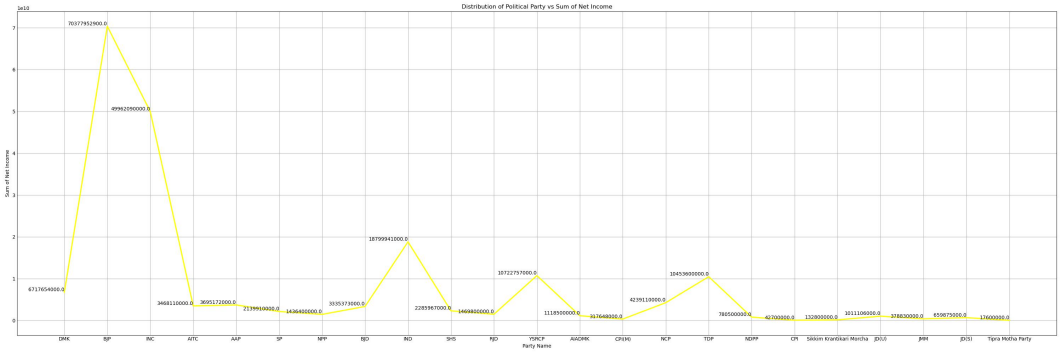


Figure 26: Distribution of Net Income Across Political Parties - Sum of All Candidates

### Mean of All Candidates

The following plot showcases the distribution of Net Income for different political parties. It showcases the mean of Net Income for all candidates belonging to a particular political party. The error margins represent the min and max value of Net Income for that particular political party.

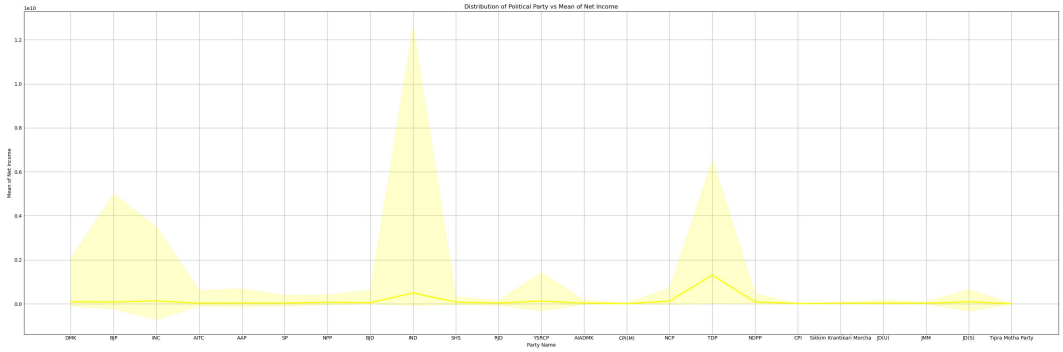


Figure 27: Distribution of Net Income Across Political Parties - Mean of All Candidates

# Final Submission

**Github Repository link :** https://github.com/AditiKhandelia/CS253-PYTHON-ASSIGNMENT
**Final F1 score for the public submission :** 0.24045
**Final F1 score for the private submission :** 0.23498
**Final public leaderboard position :** 100
**Final private leaderboard position :** 98

# References


[1] Xu, L., Cuesta-Infante, A., Veeramachaneni, K., & Skoularidou, M. (2019). Modeling Tabular Data using Conditional GAN. In *Advances in Neural Information Processing Systems 32* (pp. 8896–8906). Curran Associates, Inc. Retrieved from `https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf`

[2] SDV-Dev. (n.d.). CTGAN. GitHub repository. Retrieved from `https://github.com/sdv-dev/CTGAN`