CS253: Software Development and Operations
Python ML Assignment:Who is the real Winner?

Submitted By:
Rohan Batra
210868

1 Data Visualisation and EDA:

The following points about the data are observed :

1. There are large number of unique values of ID , Candidate Name and the Constituency thus offering very low correlations with our target variable.

2. We can observe that some of the candidates have Dr. and Adv. in their names and thus this can be used to get strong idea about the Education Level.

3. We also have some reserved constituencies denoted by (SC) and (ST) in the name and this might unfortunately reflect the education level of the representative from there.

4. We see very few data points of some of the Categories in the Education Level and thus sampling data through synthetic means will be useful.

5. There are 28 different unique values for the state and the party categorical variables.

6. Total Data points are 2059 with largest number of points corresponding to "Graduate" Education Category of 531 points.

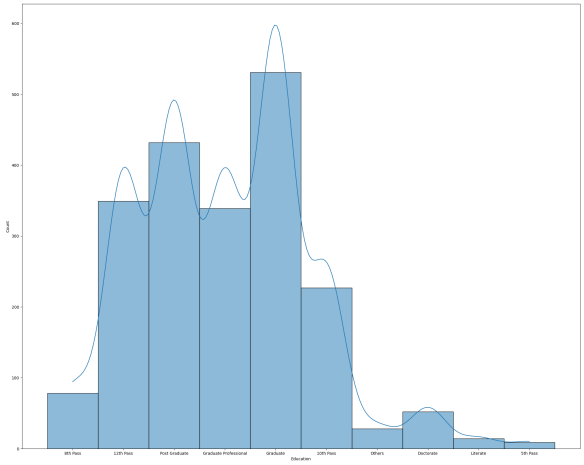

Figure 1: Different Values of Education Label and their counts

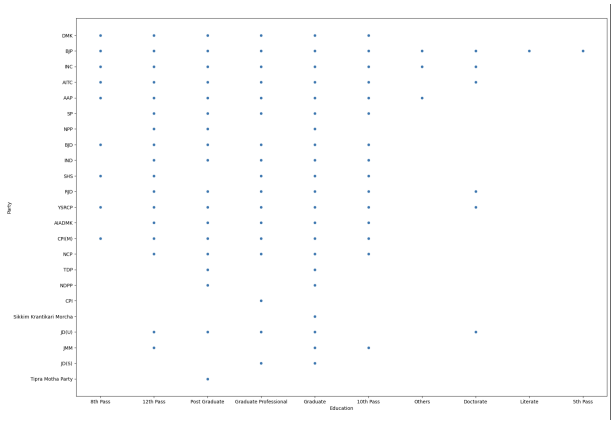

Figure 2: Relation between the Political Party and the Education Level in the data

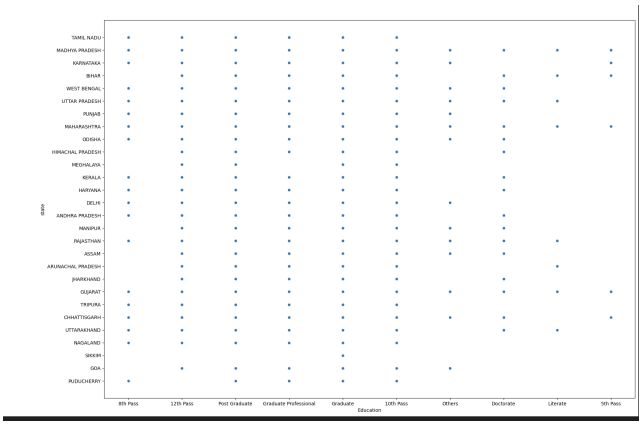

Figure 3: Relation between the State and the Education Level in the data

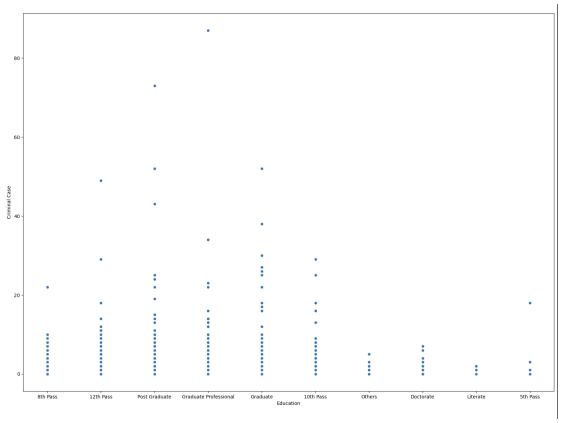

Figure 4: Relation between the Criminal Cases and the Education Level in the data

2 Feature Engineering:

The following steps were followed for feature engineering of the model data :

1. Extracting Boolean Features Dr and Adv based on the presence of "Dr." and"Adv." in the Candidate name strings, respectively.

2. Creating SC_reserv and ST_reserv if the constituency is an SC reserved or ST reserved constituency to be found from the presence of (SC) and (ST) in the Constituency name string.

3. The English value names in Total Assets and Liabilities are converted to numerical values.

4. Min Max Scaler is used to deal with the large values in the Total Assets and Liabilities columns.

5. Both the state and party categorical features are one hot encoded.

3 Synthetic Data Generation using CTGAN:

As discussed above some of the education categories have very less number of data points so generating synthetic data would be helpful. We use the CTGAN Synthesiser from the SDV library. It essentially uses Deep Learning to learn from the real data and synthesise artificial data.

First meta-data about the data is generated which is required by the CTGAN Synthesiser. The synthesiser is fitted and trained on real data and then 500 (around 25% of the dataset)points are sampled from the synthesiser.

Thus in total for the final dataset which has both the real data and the artificially synthesised data we have 2559 data points with the distribution of the different Categories of Education as shown in Figure 5.

One important observation that I made is: when I used CTGAN to specifically synthesise data points of the minority class to remove the class imbalance , the model didn't perform well (has less score). This allows to us to see that when the model resources are used up to capture the relationship of minority classes as well it doesn't do well , thus we can conclude that the test set has similiar distribution of categories of minority categories.
This makes perfect sense now as the complete population of India itself is similarly distributed with majority being graduates(lots of engineers lol), graduate professionals and post graduates and minority people being doctorates or 5th Pass. One

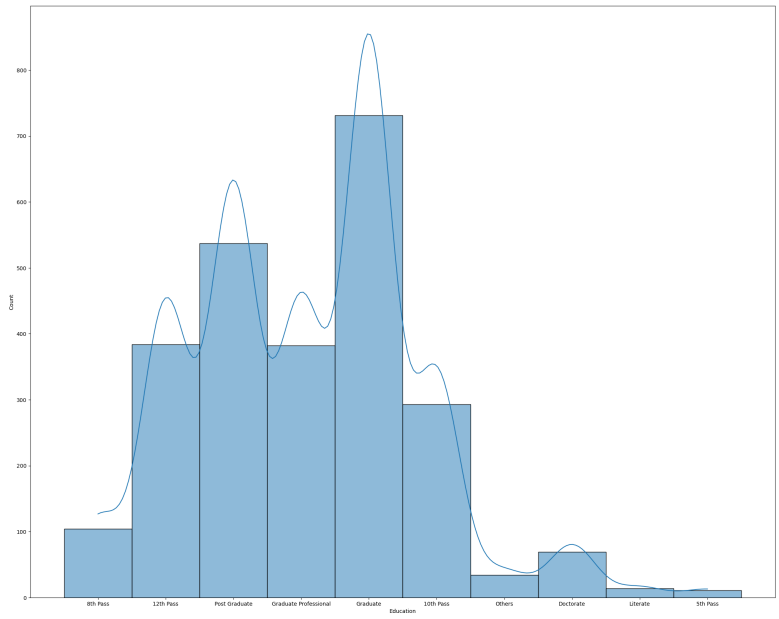

Figure 5: Final Distribution of different Education Categories

important lesson we get from this is mindlessly oversampling minority classes might not be a good idea if that doesn't reflect the true picture.(this mistake I made initially when I used SMOTE and over sampled the minority classes)

4 Machine Learning Model:

I tried multiple machine learning models during the competition: Logistic Regression, K nearest neighbour , SVC , Random Forest , Stacked Classifier of Random Forest with XG boosting and Bernoulli Naive Bayes.

The best performance was achieved using the **Bernoulli Naive Bayes** ML Model (public leader board score *0.25*).

The second best score (public leader board score *0.23*). was achieved using Random Forest with 800 classifiers on a SMOTE oversampled dataset (without CTGAN).

Please find the complete code at this link: Github Repository

5 Required Plots:

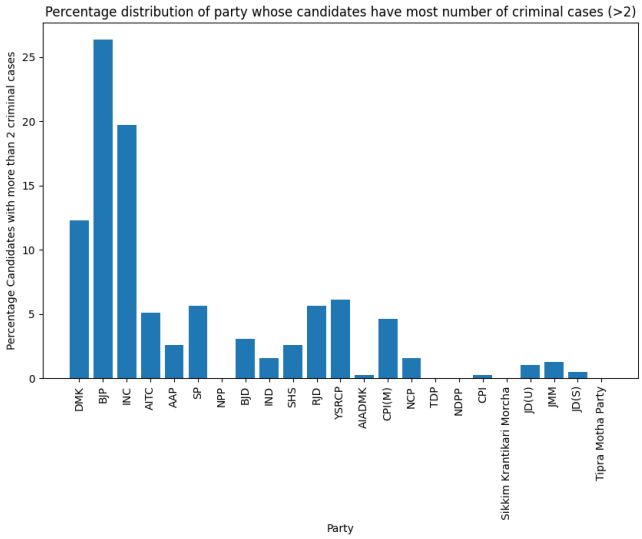

Figure 6: Different Values of Education Label and their counts

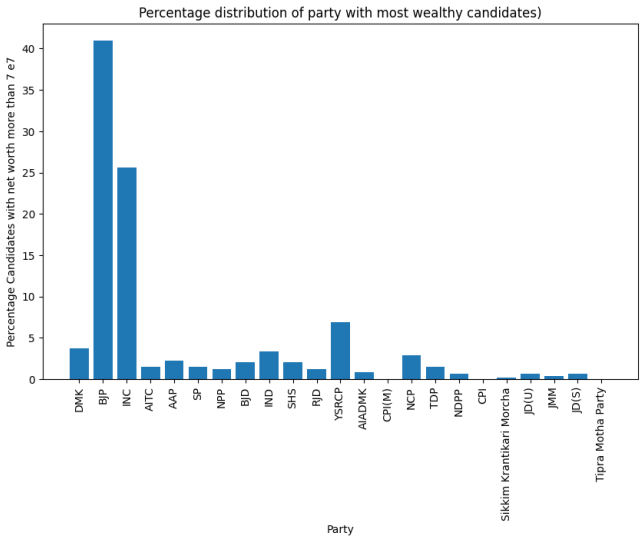

Figure 7: Relation between the Political Party and the Education Level in the data

The 75 percentile value for the number of criminal cases (2) and the weath(measured by net worth $>$ 7e7) to decide the people with most number of criminal cases and largest wealth.

6 Results and Evaluations:

Final Leaderboard Position: 15
Score : 0.26444