Python Assignment Report

Devansh Bansal

April 2024

1 Methodology

1.1 Data Preprocessing Steps

Code Link
In this section, the outline of steps taken for preprocessing the data before model training are presented. This includes:

1. **Dropping Irrelevant Columns**: Columns such as ID were dropped as they were not an indicator of the person's education level. This also helped in dimensionality reduction and helped in increasing the efficiency of the model.

2. **Encoding Name Titles and Constituency Caste**: Encoding for things such as name titles of Dr. and Adv. along with Caste such as SC and ST was done. The reason for the above is that a Dr. or an Adv. is sure to be well educated and the caste can impact one's education level due to various societal reasons and problems.

3. **Converting Monetary Values**: Monetary values in the dataset were converted to numeric format using a custom function. This conversion allows for proper model training as it is able to use the mathematical values to gain insights into the data.

4. **Encoding Categorical Data**: 'Party' which is categorical data was encoded into numbers to allow for model training using One-Hot Encoding provided by the pandas library. This allows for numerical data for adequate training of the model.

2 Experiment Details

In this section, I provide details of the models used for training, along with their hyperparameters and other relevant information. The final model I went with was a Random Forest.

Random Forest is an ensemble learning technique that builds multiple decision trees during training. It constructs each tree in the forest based on a random subset of the training data and a random subset of features. Each tree in the forest independently predicts the output, and the final prediction is determined by averaging or taking a vote across all trees (for regression and classification tasks, respectively).

Here are the reasons why I chose it:
High Accuracy: Random Forest generally provides higher accuracy compared to single decision trees, especially for complex problems with non-linear relationships.
Reduced Overfitting: By using random subsets of the data and features, Random Forest reduces the risk of overfitting, which is common in individual decision trees.
Handle Large Datasets: Random Forest can efficiently handle large datasets with many features and instances.
Feature Importance: It can provide insights into feature importance, allowing users to identify which features contribute the most to the prediction.

Now the next most important thing was hyperparameter tuning and so I used Randomized Search to iterate through a bunch of hyperparameters and select the best one using cross validation.

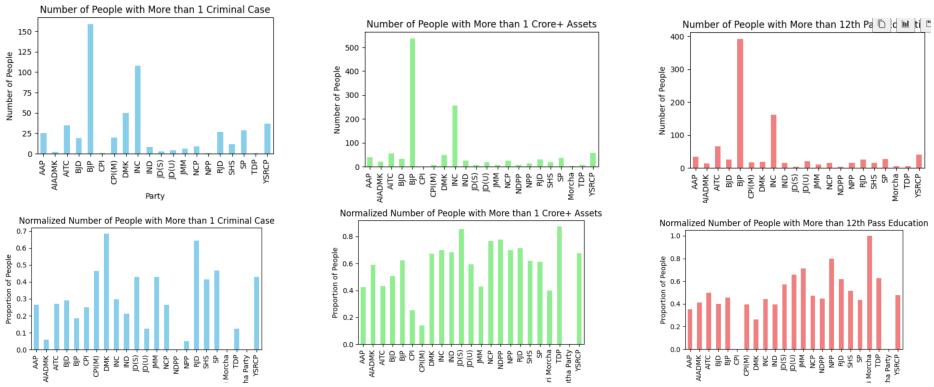
Random Search CV is an optimization technique used to find the best set of hyperparameters for a machine learning model. It randomly samples a predefined number of hyperparameter combinations from the specified search space. Each hyperparameter combination is evaluated using cross-validation on the training data, typically using k-fold cross-validation. The hyperparameter combination that results in the best performance metric (e.g., accuracy, F1 score) on the validation set is selected as the optimal set of hyperparameters.

Some advantages of using Random Search:
Exploration of Hyperparameter Space: It explores a wide range of hyperparameter values by randomly sampling from the search space, increasing the likelihood of finding good hyperparameter combinations.
Better Performance: Random Search CV often leads to better performance compared to using default hyperparameters or manual tuning, as it systematically searches the hyperparameter space.

Model	Hyperparameters	Details
SVM	Kernel: RBF, C: 1.0	Support Vector Machine classifier with Radial Basis Function kernel.
Random Forest	n_estimators: 100	Random Forest classifier with 100 trees and unlimited depth.
Random Forest	Randomized Search with Parameter Dictionary 'n_estimators': randint (10,1000), 'random_state': [42, 25, 47]	Random Forest classifier using Randomized Search Cross Validation for hyperparameter tuning.

Table 1: Summary of Models Used

3 Data Insights



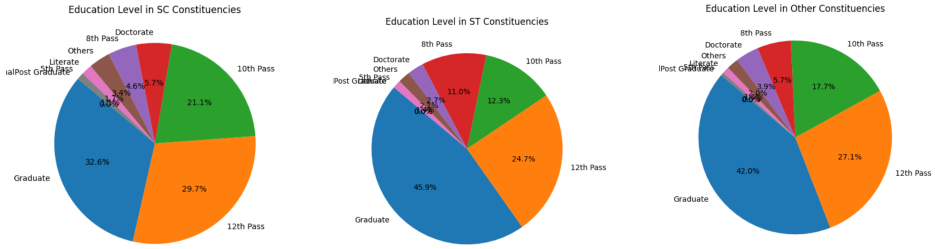
The first row of graphs shows the absolute number of people with more than 1 criminal case, with some parties having significantly higher numbers than others. For example, BJP has around 150 such people, much higher than most other parties.

The second row of graphs shows the same data but normalized by the total number of candidates for each party. This provides a more meaningful comparison, showing that parties like DMK and RJD have a relatively high proportion of candidates with more than 1 criminal case compared to their overall candidate pool.

The normalized graph reveals that some smaller parties like YSRCP, CPI(M) also have a high proportion of candidates with multiple criminal cases, perhaps indicating issues with candidate selection in those parties.

The contrast between the two graphs highlights the importance of considering both the absolute and relative numbers when analyzing data on criminal cases involving political candidates. Looking at just the raw numbers can be misleading.

ing without accounting for the overall size of the candidate pool for each party.



In SC constituencies, the education level is skewed more towards lower levels with Graduate level education (32.6%), followed by 12th Pass (29.7%). The proportion of candidates with a Doctorate degree is at 5.7%.
In ST constituencies, the majority of candidates have a Graduate level education (45.9%) followed by 12th Pass (24.7%). The percentage of candidates with a Doctorate degree is only 2.0%.
In Other constituencies, the education profile is more evenly distributed, with a significant proportion of candidates having a Graduate degree (42.0%) and 12th Pass (27.1%) education. The representation of candidates with higher degrees like Doctorate (3.8%) is still relatively low compared to the Graduate and 12th Pass levels.
Overall, the data suggests that the education levels of candidates vary across different types of constituencies, with SC and ST constituencies having a relatively lower proportion of candidates with higher degrees compared to Other constituencies.

4 Results

F1 Score: 0.23799
Public Leaderboard Rank: 106
Private Leaderboard Rank: Not Released

5 References

Intro-to-ML-and-DL Link
This was a project I had done in my first year which provided me with majority of the insights I needed for this assignment. Apart from this, whenever I got stuck, I looked up the error on Google and mostly StackOverflow came to the rescue.