Python Assignment Report

S.Sai Nikhil

April 13, 2024

1 Methodology

1.1 Data Preprocessing Steps

- The strings in the Assets and Liabilities (20 Crore+, 10 Lac+, 15 Thou+) are converted into their respective numerical values (20,00,00,000; 10,00,000; 15,000 resp)

- The categorial variables such as Party and State are encoded using the Labelencoder function.

- Splitting the data into features (X) and the target variable (y).

1.2 Feature Engineering

- Transformation of string based numerical values in "Total Assets" and "Liabilities" columns to numerical values using a function (get_num).

1.3 Normalization, standardization, or transformation

- Imputation of missing values using SimpleImputer with median strategy.

- Standardization of features using StandardScaler in the pipeline.

1.4 Others

- GridSearchCV for hyperparameter tuning.

2 Experiment Details

- n_neighbors: Number of neighbors to consider (values: 3, 5, 10, 15)

- weights: Weight function used in prediction (values: 'uniform', 'distance')

- p: Power parameter for Minkowski distance (values: 1, 2)

- Data preprocessing steps include imputation of missing values using median strategy and standardization using StandardScaler.
- The best model selected based on the F1 weighted score on the validation set.
- Predictions are made on the testing data, and the results are output to a CSV file named 'KNNimproved.csv'.

2.1 Data Insights

1. **Distribution of Education level across total assets**

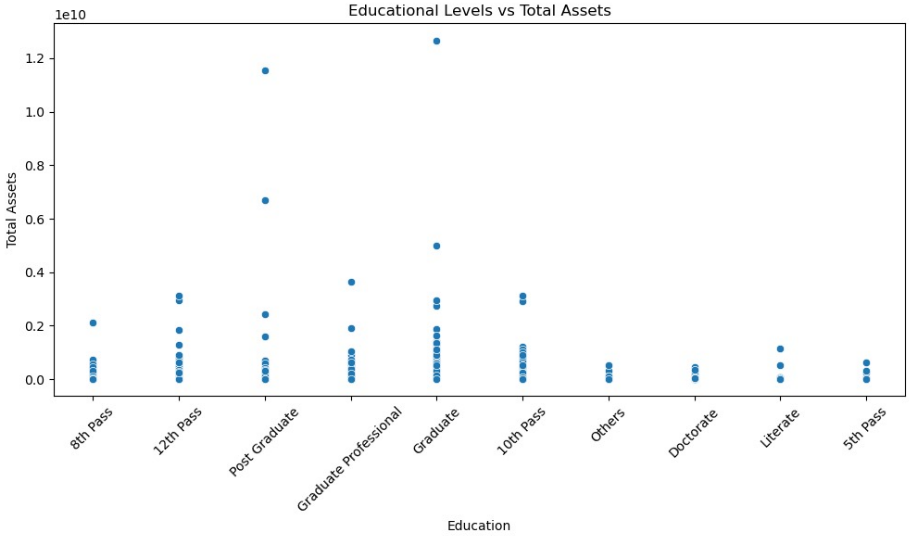

Figure 1: Education vs Assets

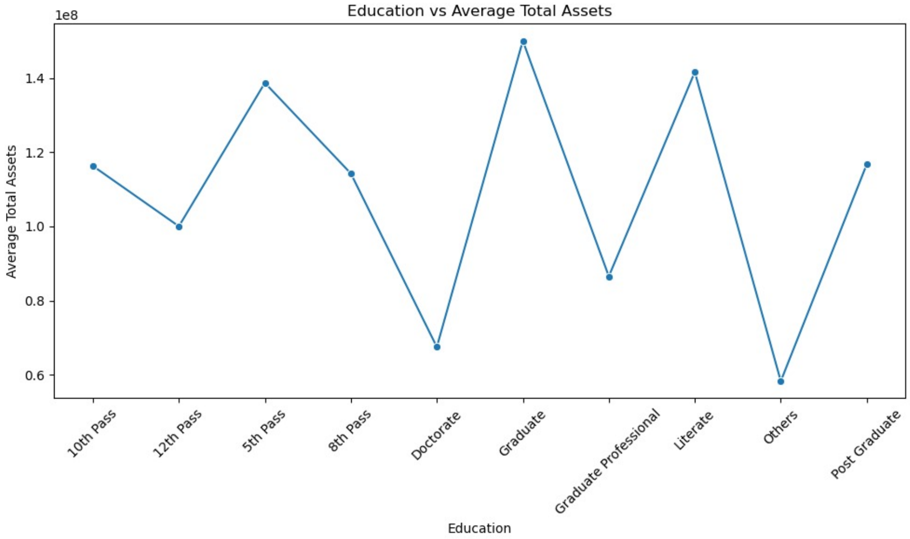

Figure 2: Education vs Assets

2. **Distribution of Education level across Liabilities**

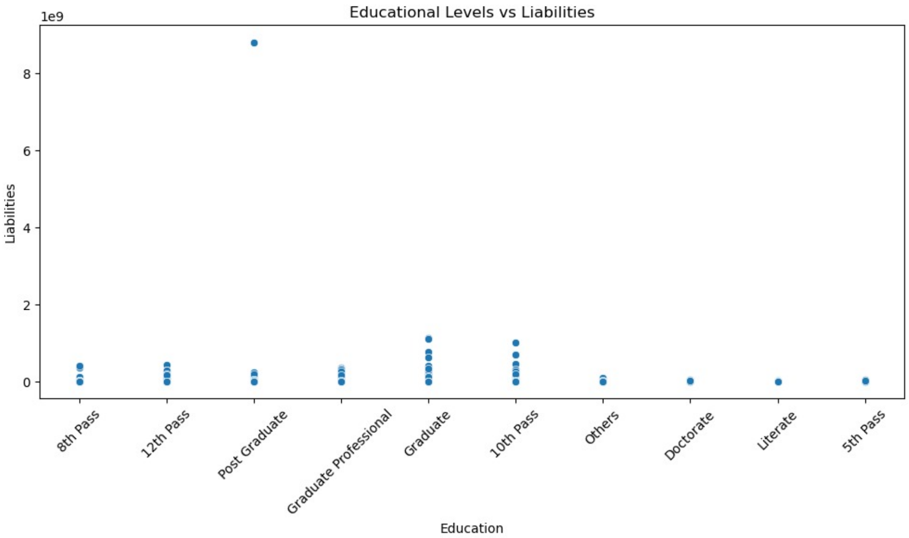

Figure 3: Education vs Liabilities

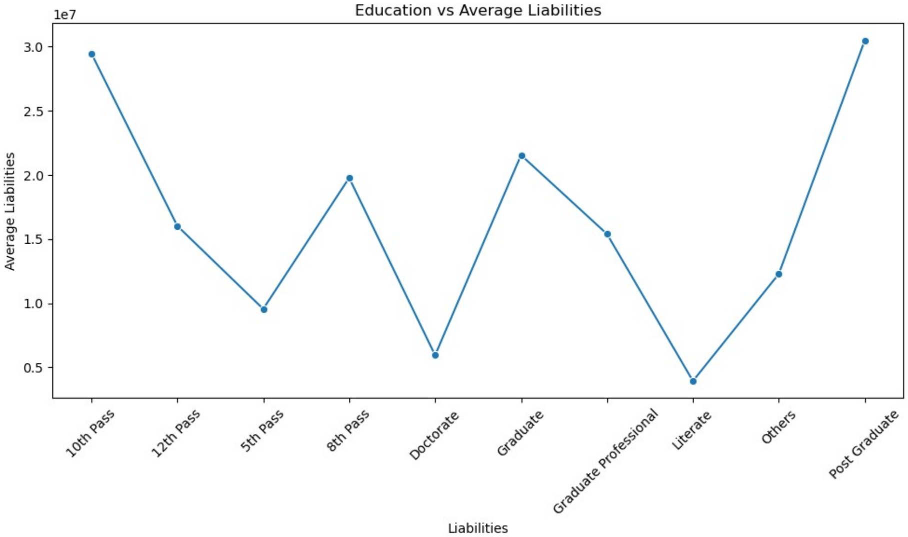

Figure 4: Education vs Liabilities

3. **Distribution of Education level across Parties**

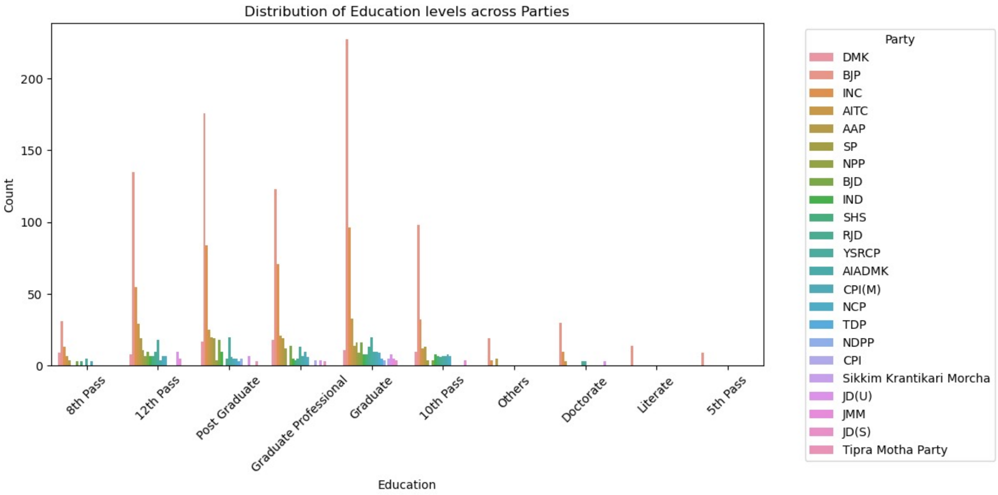

Figure 5: Education across Parties

4. **Distribution of Education level across States**

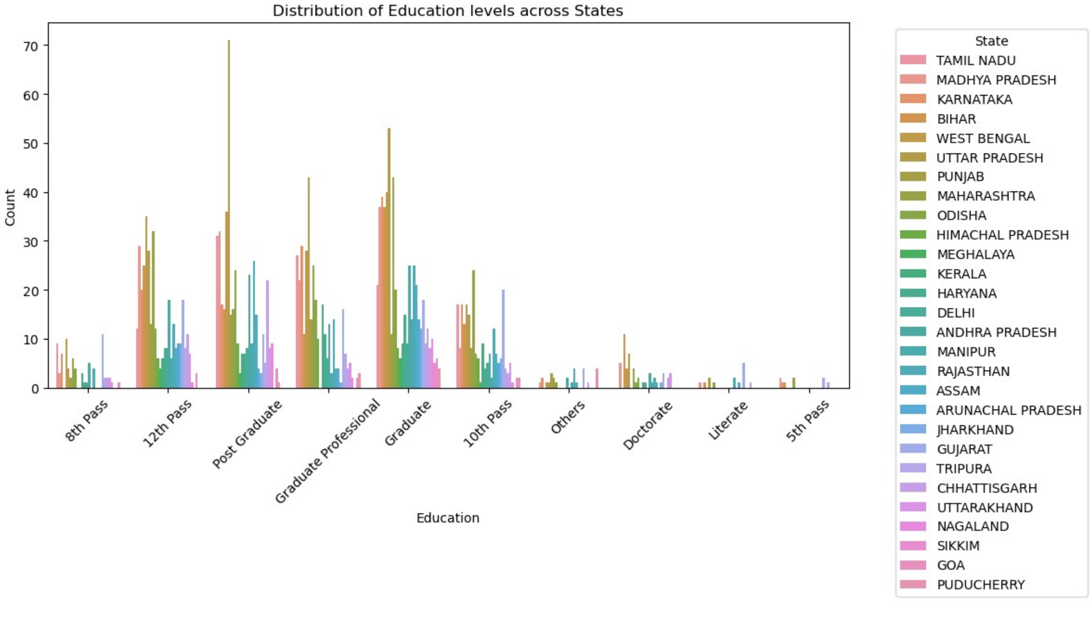

Figure 6: Education across states

3 Results

- F1 Score: 0.23433
- Rank in Public Board: 119
- Rank in Private Board: 108

4 References

- Link to the Code: https://github.com/Sainikhilsudhamandu/CS253_Python_Assignment/tree/main
- To get Introduced to ML: https://www.youtube.com/watch?v=i_LwzRVP7bg&ab_channel=freeCodeCamp.org
- For further understanding of scikit Library: https://scikit-learn.org/stable/
- For doubts regarding the scikit Library: https://stackoverflow.com/questions/tagged/scikit-learn
- For Data Visualization referred to: https://github.com/mwaskom/seaborn and https://matplotlib.org/stable/users/installing/index.html