A REPORT ON LOAN DEFAULT PREDICTION

PRESENTED BY:
MADUABUCHI ANAMELECHI
maduabuchianamelechi@gmail.com

INTRODUCTION

Any financial organization suffers greatly as a result of bad loans. The effort to be proactive in order to stop loans from defaulting via the conventional ways has been time-consuming and fruitless.

NEXT STEP: MACHINE LEARNING

- Therefore, the ability of machine learning algorithms to forecast whether a borrower will fail on a loan or not must be leveraged immediately.

- Moving further, we'll examine a machine learning technique called **Logistics Regression** that uses historical data to forecast whether a borrower would fail on a loan.

WHY LOGISTIC REGRESSION?

LOGISTIC REGRESSION IS WELL KNOWN FOR ITS ABILITY TO PREDICT BINARY OUTCOMES (0 / 1).

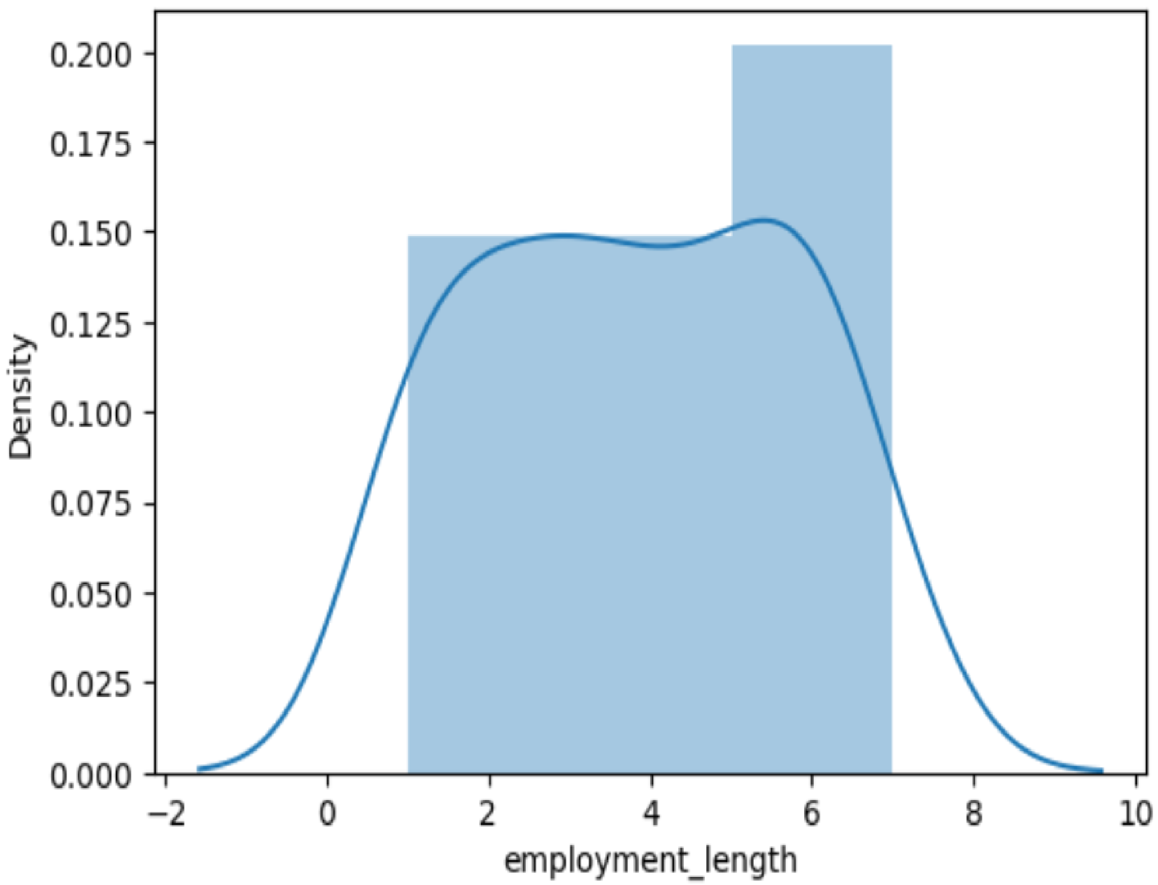
IT BECOMES A PREFERRED ALGORITHM FOR US BECAUSE WE ARE TRYING TO CLASSIFY WHETHER A PERSON WILL DEFAULT ON A LOAN.

ALSO, ITS ABILITY TO SHOW THE PROBABILITY OF ITS PREDICTION HELPS EXPLAIN TO WHAT DEGREE THE LIKELY EVENT WILL OCCUR.

DATA WRANGLING

While exploring the dataset, I discovered the following issues

- The column headers was named inappropriate.
- The Annual income column contains currency symbols and a comma
- The Employment length column contains more than one-character types.
- The Employment length, Debt-to-income ratio, Loan default columns have missing row of data.
- The dataset was imbalance and had inappropriate data types.



Steps taken to clean them

- The column headers was renamed following the best naming convention
- The columns with mixed data types was cleaned using the pandas replace and split function.
- The missing data in the Loan default column was filled using the mode function because it has only two unique values (0 and 1).
- The missing data in other columns was filled using the mean function because the distribution of those columns was not skewed.
- The columns was converted to its appropriate data types using the pandas astype function

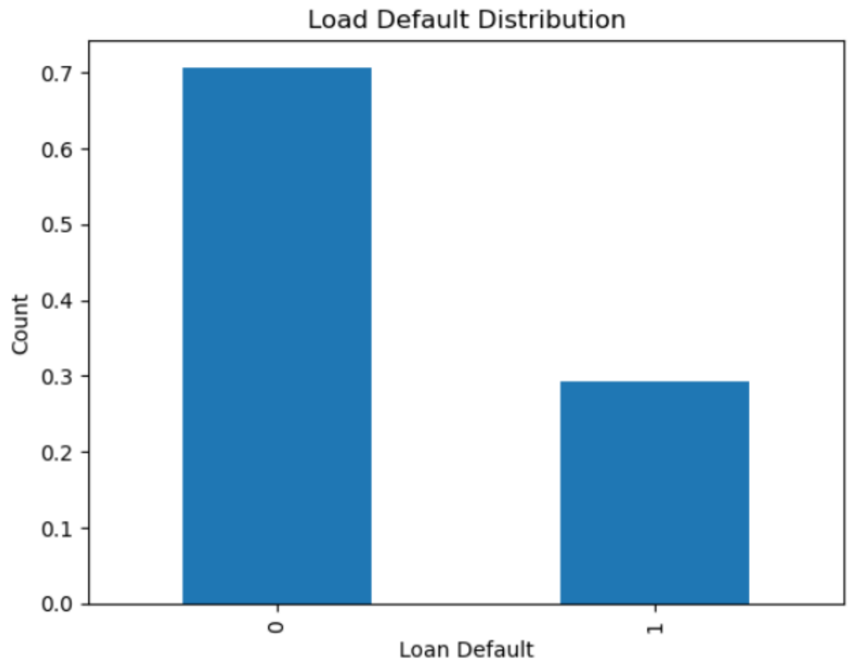
HANDLING MULTI-DIMENSIONALITY AND IMBALANCED DATASET

From the result of the pandas nunique() function, it is seen that the customer_id column has a high dimensionality and was therefore dropped as it wont feed anything to our model.

Check the Multi-dimensionalism of the dataset

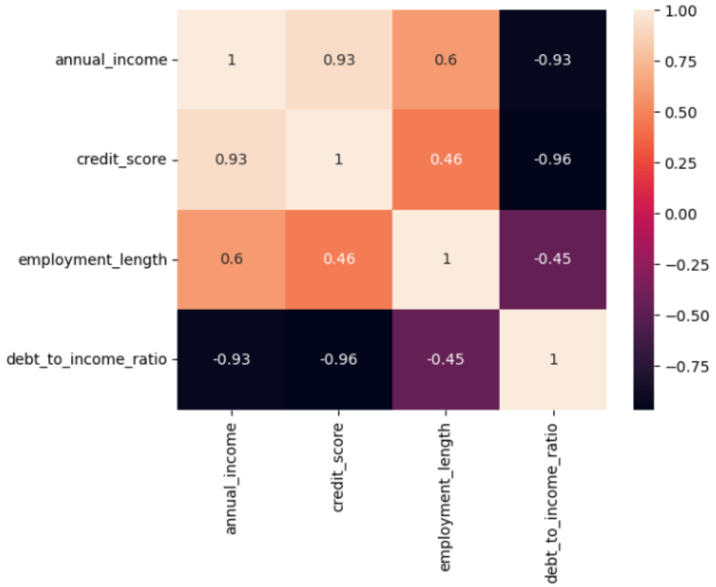
In [23]:
```
df.nunique()
```

Out[23]:
```
customer_id             58
annual_income           24
credit_score            25
employment_length        7
debt_to_income_ratio    32
loan_default             2
dtype: int64
```



Visualizing the distribution of our target column (loan_default) shows that the distribution is imbalanced as non defaulters makes up about 70.7% of the dataset. This was fixed by doing over sampling on the training dataset using the RandomOverSampler function from the Imblearn library.

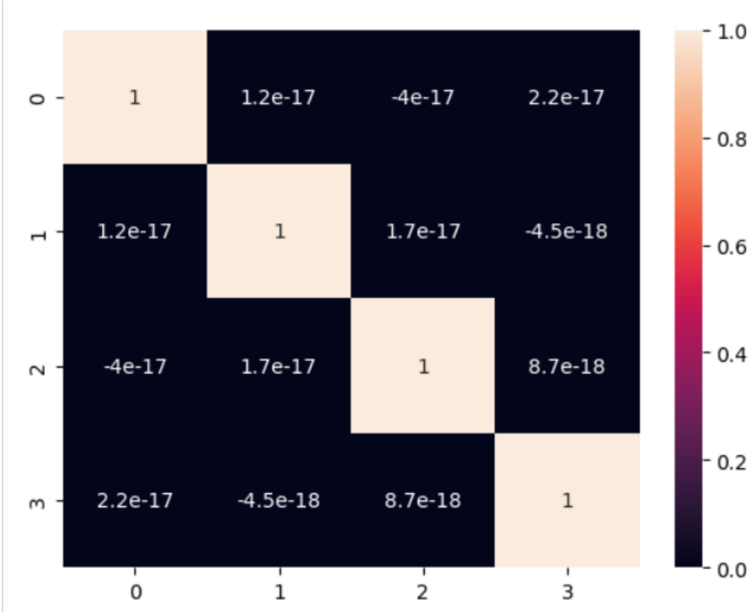
HANDLING MULTICULLINEARITY



Using the seaborn heatmap() function to visualize the correlation plot of the dataset shows that there are high correlation between the following columns:

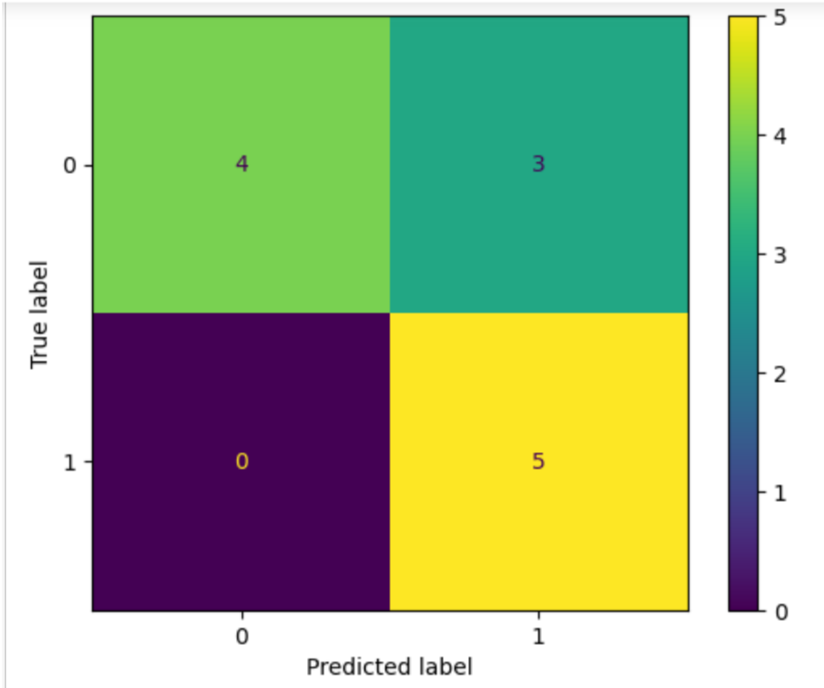
- annual_income and credit_score
- annual_income and employment_length
- credit_score and employment_length

This was fixed using the dimensionality reduction technique PCA



MODELING AND PERFORMANCE

After the **LogisticRegression()** was trained, prediction was obtained. When evaluated, the model had an AUC-ROC score of 79%.



Check the AUC-ROC score

```
In [43]: pred_prob = model.predict_proba(X_test)

In [44]: roc_auc_score(y_test, pred_prob[:, 1])

Out[44]: 0.7857142857142857
```

Taking a quick look at the Confusion Matrix, we can see that the model did a great job in precision than recall.

RECOMMENDATIONS

- The company should give out more loans to those who have spent more that 2 years in the organization as they are more likely to perform.

- To increase the efficiency of the model, More features should be added. Like spending pattern, age, average account balance, debit and credit inflows, location data, etc.

- Also, the volume of the data should be increased as this will help the model to learn better during training and further generalizes on test data.


Thank You