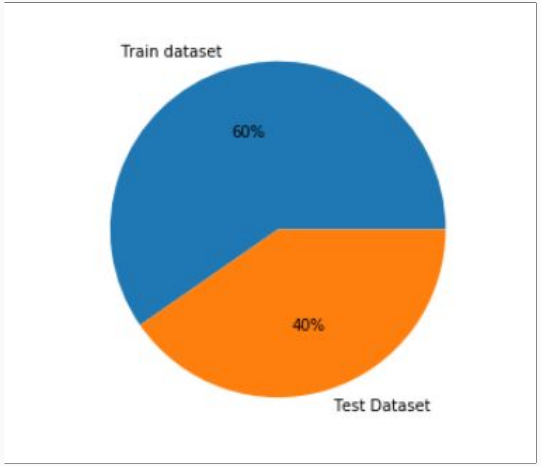
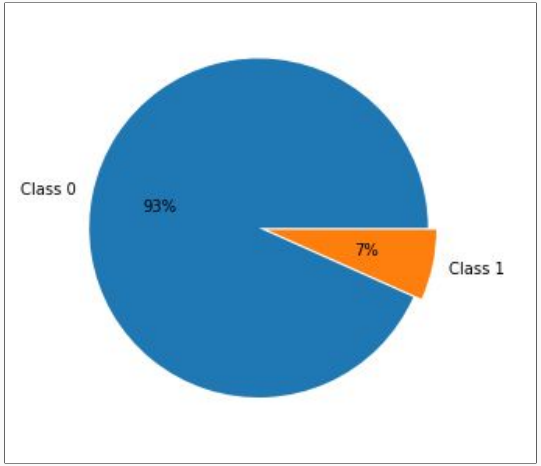
Kaggle : GiveMeSomeCredit Data Analysis

Sandheep Gopinath

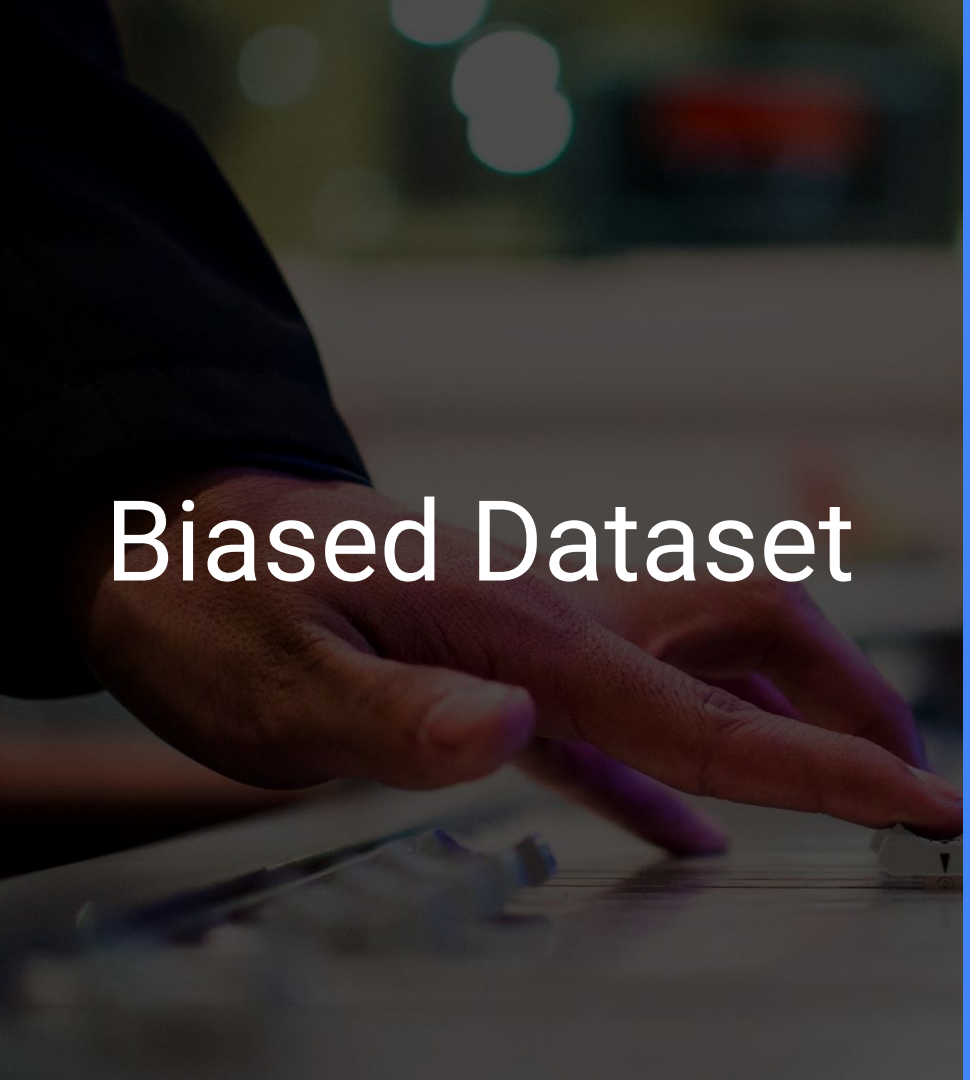
The Dataset



60% Training data and 40% Testing data



Baseline Accuracy : 93%

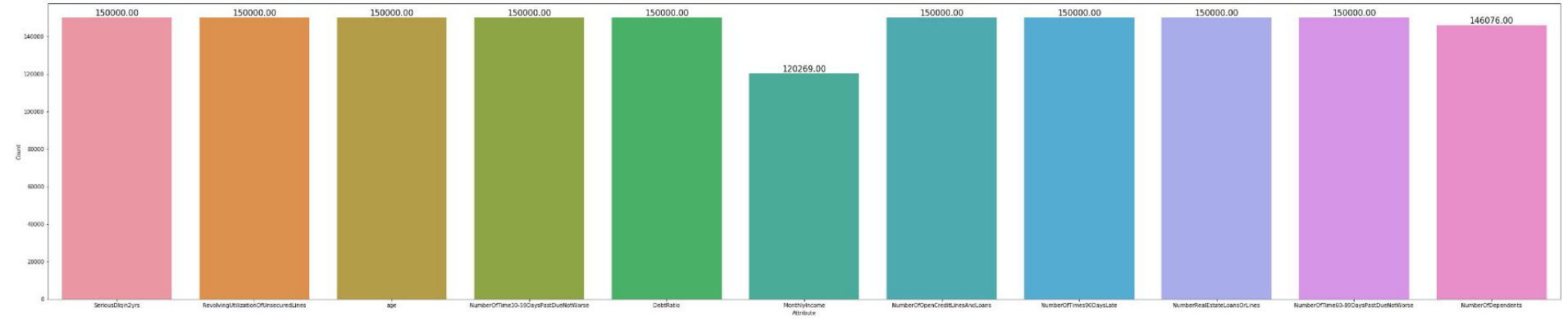


Biased Dataset

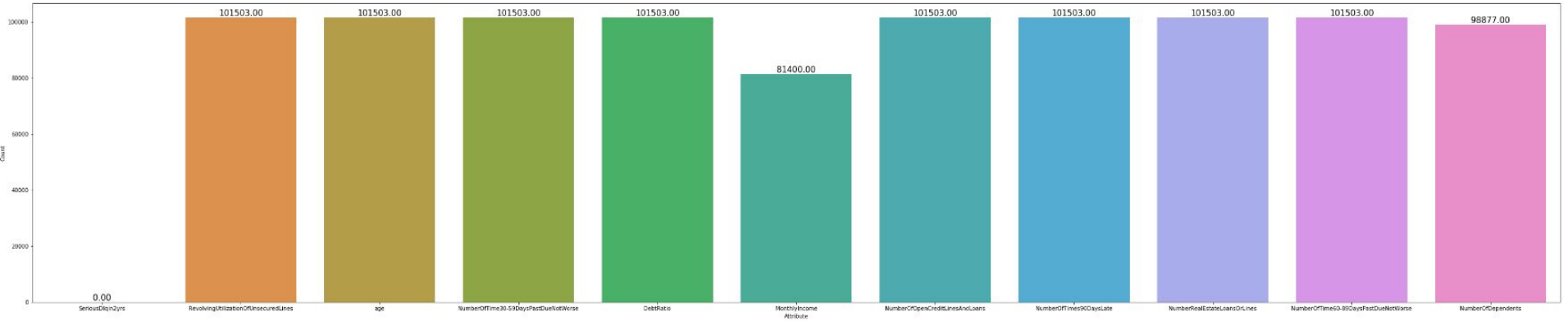
Since the dataset is biased towards one class there is a high chance that the model will predict more of the outputs as Class 1. The models will also show accuracy around 93% with least training.

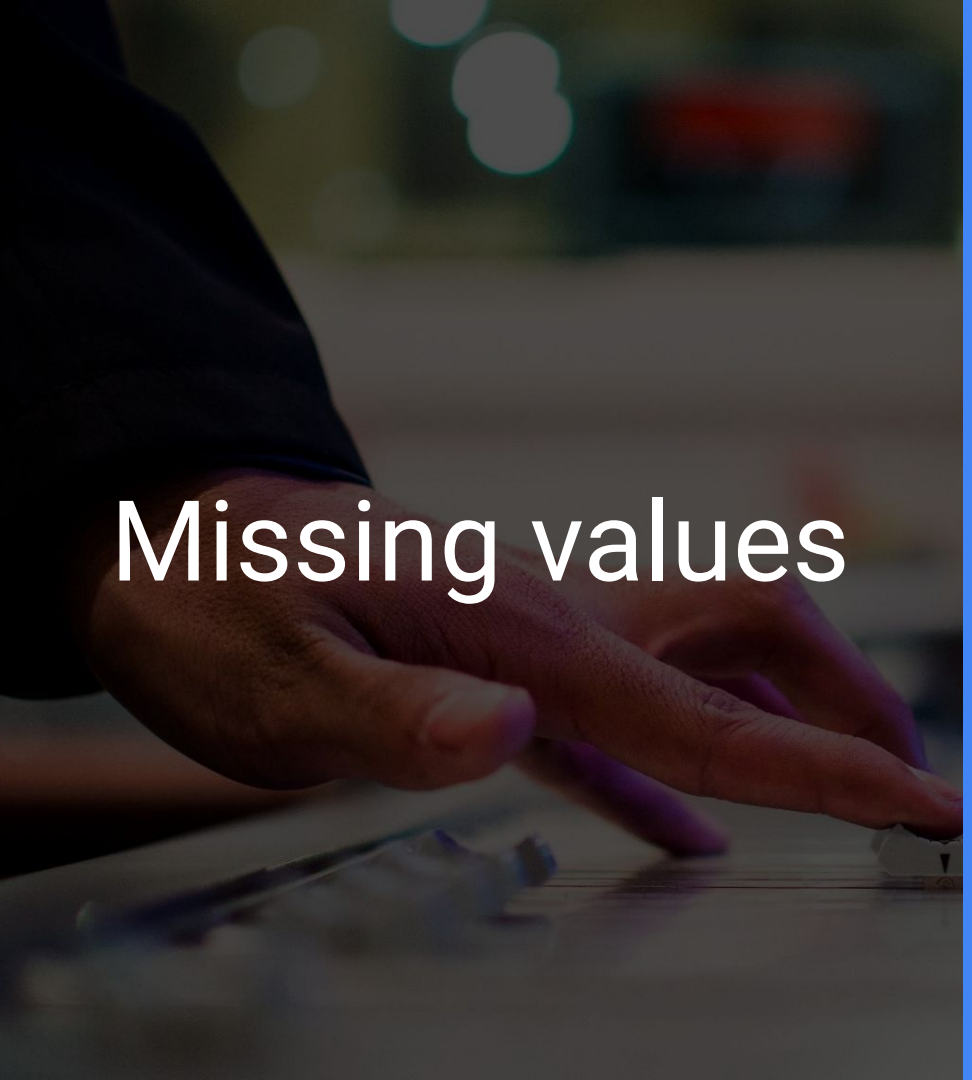
Missing Values

Train Data Info



Test Data Info

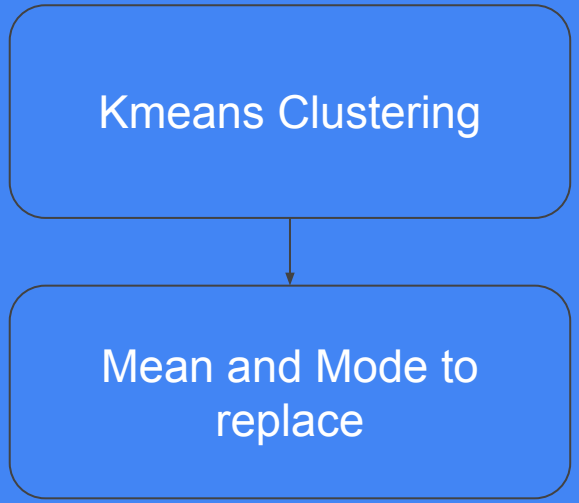




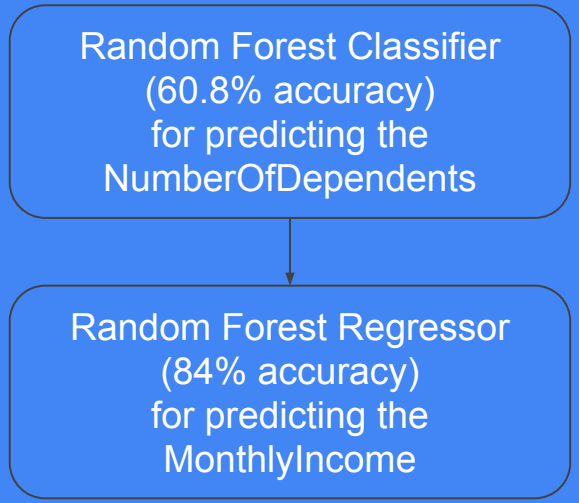
Missing values

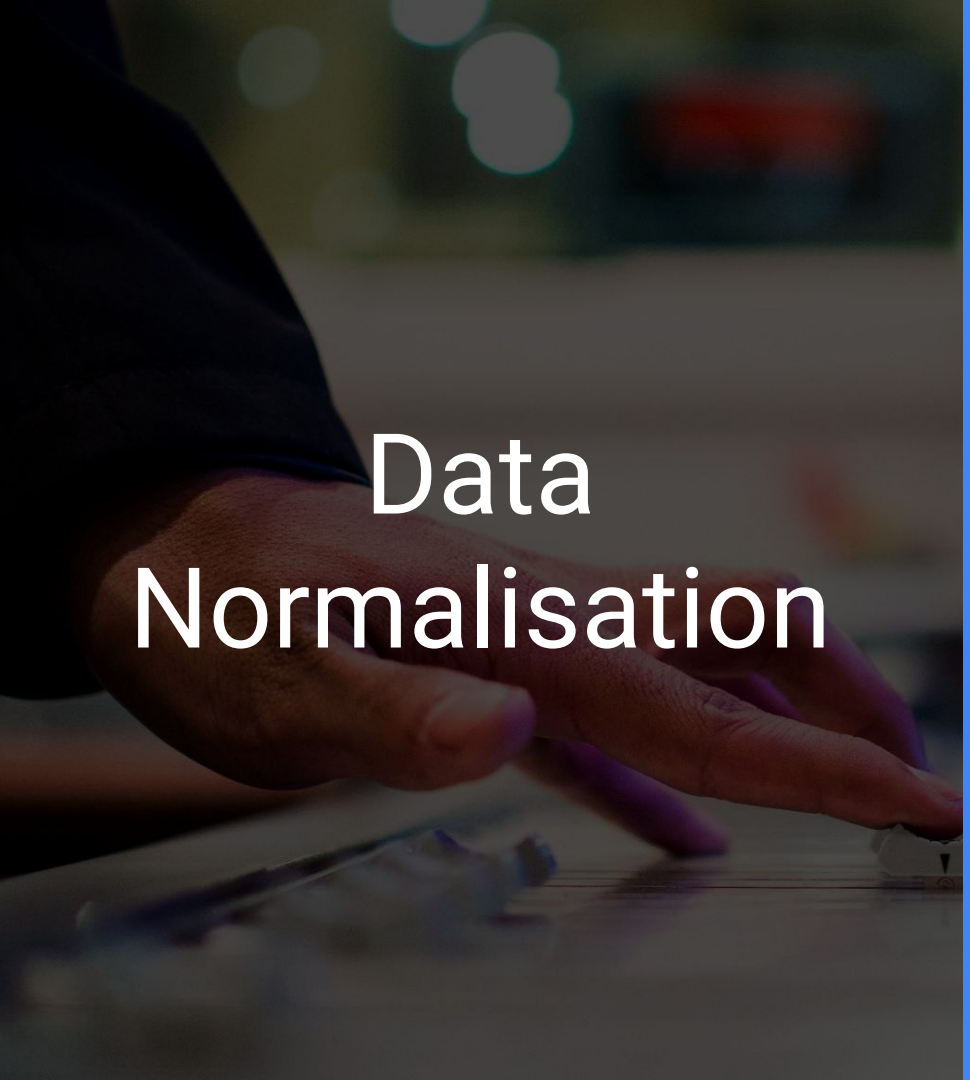
The dataset has missing values in Number Of dependents and Income in both training and test dataset.

Handling Missing Values



Handling Missing Values

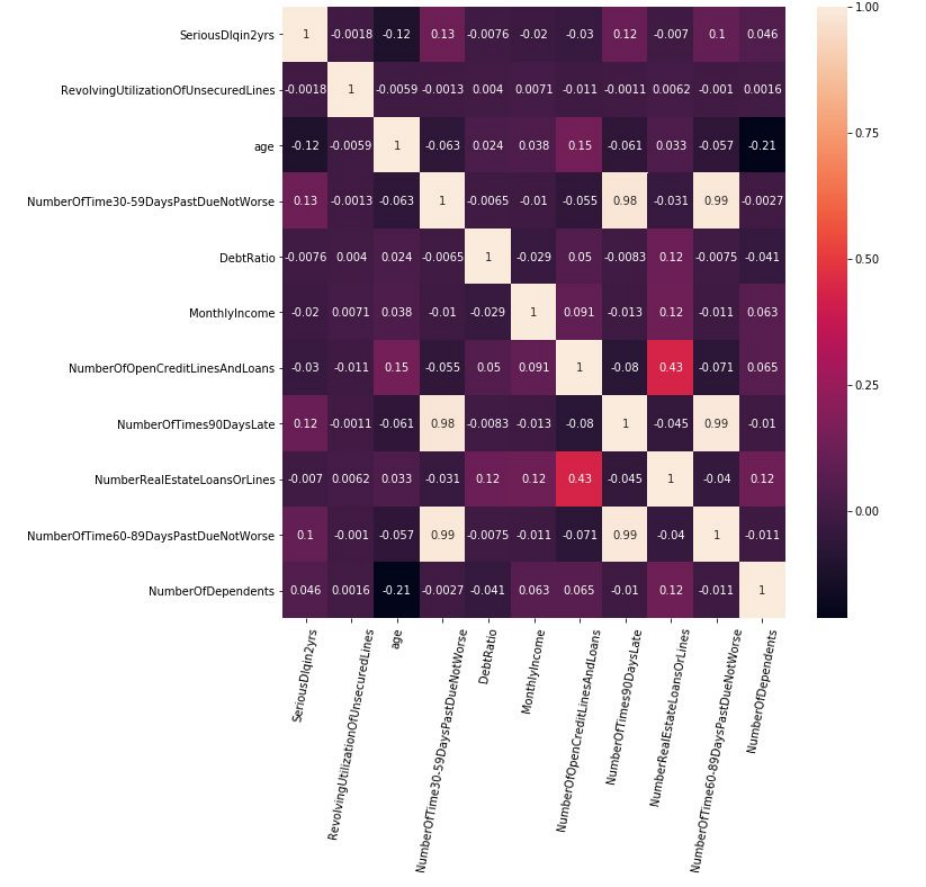




Data Normalisation

Using StandardScaler to bring all attributes to a common range.

Correlation



Two attributes which had high correlation more than 0.90 were dropped

NumberOfTime60-89DaysPastDueNotWorse

and

NumberOfTimes90DaysLate

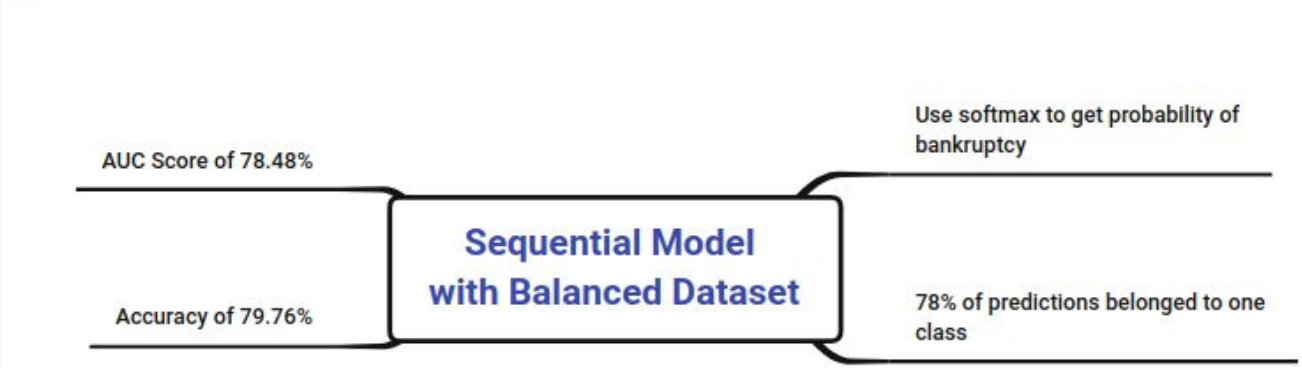
Model Building with Logistic Regression

Model Building with Neural Networks

```
Model: "sequential_3"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_18 (Dense)             (None, 1, 16)             144

 dense_19 (Dense)             (None, 1, 32)             544

 dense_20 (Dense)             (None, 1, 16)             528

 dense_21 (Dense)             (None, 1, 8)              136

 dense_22 (Dense)             (None, 1, 4)              36

 dense_23 (Dense)             (None, 1, 1)              5

=================================================================
Total params: 1,393
Trainable params: 1,393
Non-trainable params: 0
```

The data has to be converted to np array and the dimension had to be changed by using np.expand_dims so as to fit the model

Model Building with Neural Networks



Summary

- For the unbalanced dataset, the Accuracy score was higher but the AUC score was low
- The high accuracy in unbalanced dataset was because of the model doing predictions in the same class always and that class consisted of 93% of the total data
- For the balanced dataset, where we had to drop a large number of data points, the Accuracy was low. However, the AUC had increased
- The neural network models were slightly better than Logistic regression.
- Kaggle Score : 77.286%