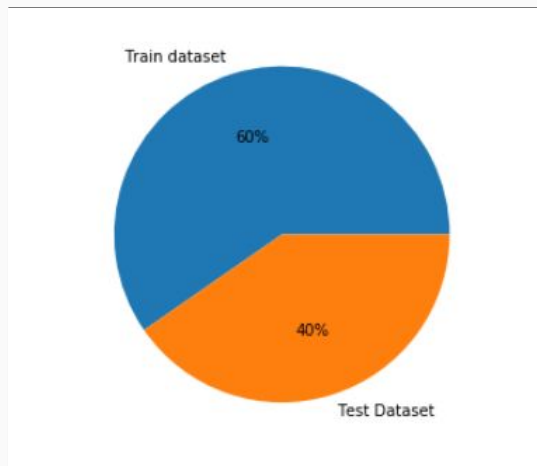


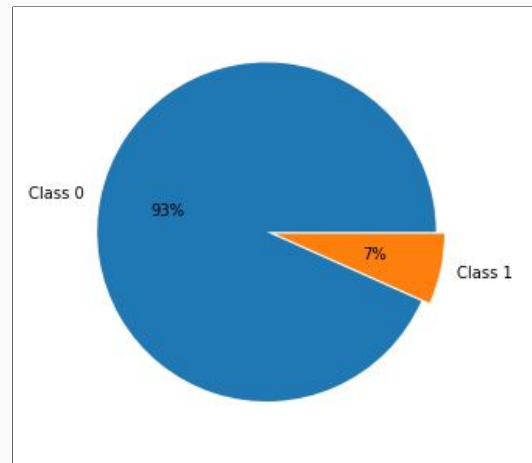
# Kaggle : GiveMeSomeCredit Data Analysis

Sandheep Gopinath

# The Dataset



**60% Training data and 40% Testing data**



**Baseline Accuracy : 93%**

A close-up photograph of a hand holding a pen, poised to write on a document. The background is blurred, showing a person's face and a red light source. The text 'Biased Dataset' is overlaid in white on the left side of the image.

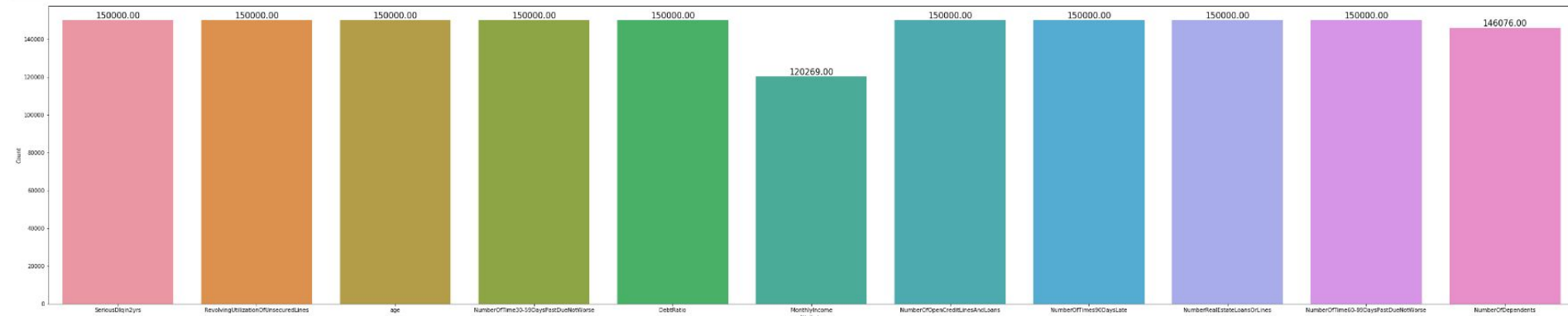
# Biased Dataset

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.

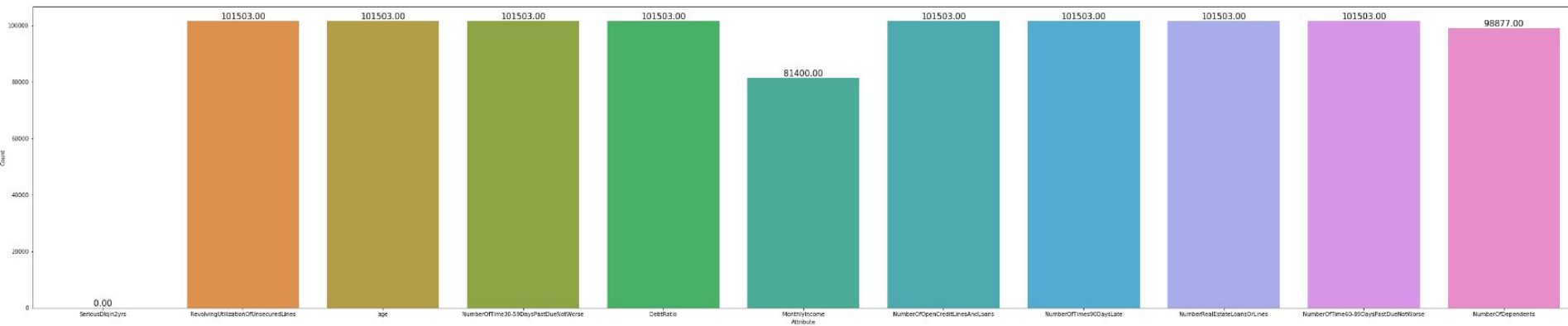
The models will also show accuracy around 93% with least training.

# Missing Values

## Train Data Info



## Test Data Info



A close-up photograph of a person's hand holding a pen, poised to write on a document. The background is out of focus, showing some bokeh light effects. The text 'Missing values' is overlaid in white on the left side of the image.

# Missing values

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

# Handling Missing Values

Kmeans Clustering



Mean and Mode to  
replace

# Handling Missing Values

Random Forest Classifier  
(60.8% accuracy)  
for predicting the  
NumberOfDependents



Random Forest Regressor  
(84% accuracy)  
for predicting the  
MonthlyIncome

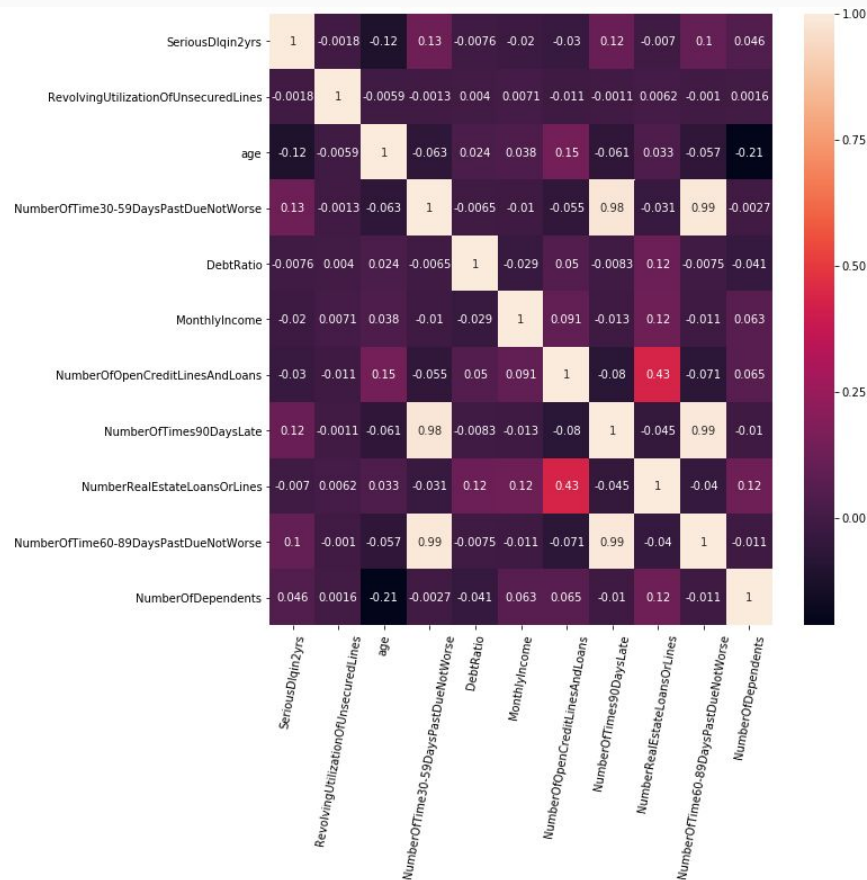


# Data Normalisation

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

**NumberOFTime90DaysLate**

# Model Building with Logistic Regression



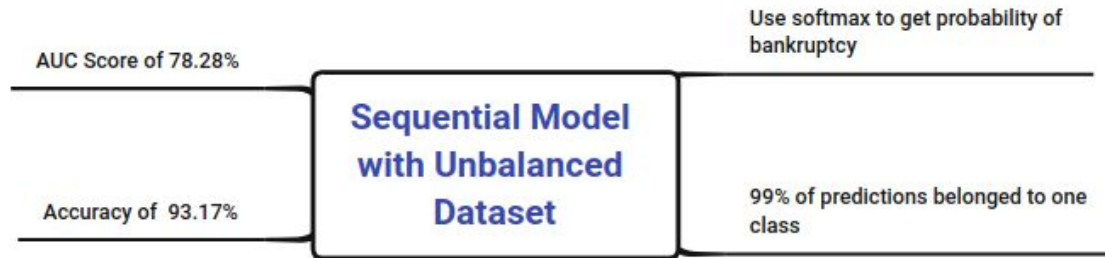
# Model Building with Neural Networks

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%