Problem Statement – To produce a trained model which can be used to order a set of transactions by likelihood of matching a receipt image.

Accuracy of the model – 94.8%

Reason for the choice – This model was chosen as it showcased evidence that suggested that the model would generalise better. However, class imbalance makes this evaluation a very tricky scenario. For accounting for this, 5 other models have also been included in the assessment.

\*\*\* A more detailed analysis can be found within the Jupyter notebook where every step has been explained to great detail.

Approach –

1. EDA was performed

Based on the EDA performed, it was found that the datatset did not have any missing values, and from a basic understanding – it seemed to contain a few ID columns and certain metric columns which was a probability of match

1. Each of the columns giving out the probability was on a different scale, so the immediate inference was that it needed to be standardized – This was done by binning and creating buckets of probabilities such as [0-20%] [20-40%] etc.
2. Conversion of these probability buckets to scores so that these scores could be added to form a weightage
3. There was an obvious class imbalance and to manage this, we tried different configurations of models
4. 6 different models were fit, with different configurations and their comparison was provided, something called an “Area under the curve” “Receiver operating characteristic curve” was plotted to form the comparison.
5. The best model was picked and a function was created which would take in input data in the form of a data frame and it would find predictions which suggests if a data point is infact the right receipt id or not. The auc-roc score was at 95%
6. Al alternate method to improve the model was also evaluated which makes use of scoring datasets – This however increased the number of classes and there was little evidene to suggest that it was a better outcome
7. Combining both these methods would probably yield better results, however that was ot performed as it takes quite a bit of time.