**Classifier Selection:**

I trained the model on a number of classifiers starting with K-NN, Naïve Bayes, Decision Tree Classification and Random Forest Classification.

I recognised that since data was big so Naive Bayes and K-NN was not working properly.

Then I started using Boosting classifiers which are tree-based algorithms like light gb, GCD and xg\_boost which showed better results.

**Model Tuning:**

I tried using n\_estimators, max\_bin, learning\_rate, num\_iterations for tuning my model to reach to a model with was giving me highest accuracy. It was mostly a trial-and-error method with different parameters.

**Feature Selection:**

While training the model I tried a various feature technique. To list them:

1. I tried to create a new feature namely Debt – Multiplication of Income\*Debt\_Income. Including this new feature didn’t improve model performance.
2. I tried to remove a few features individually on the basis of correlation and feature importance. I personally feel feature importance as a better technique as it can interpret non linear data properly but removing features based on either of two didn’t help much.
3. I also tried to remove ‘-1’ in income section and with average of income and trained model but it too didn’t improve model’s efficiency. I also removed those entries but it didn’t help either.
4. In Length Employed also there were few entries as ‘-1’ which I removed but that also didn’t show much improvement.

**Features Understanding:**

From correlation matrix I observed that number of open accounts and total accounts are highly correlated.

All scatter plots are drawn to represent correlations between different features.