MILestone 2

SC-22

* Bar Graphs.
* Chart, bar chart

  Description automatically generatedTrain -validation-test accuracy
* Chart, bar chart, waterfall chart

  Description automatically generatedTraining time
* Test time

Chart, bar chart

Description automatically generated

* Feature selection:
* Chart

  Description automatically generatedFirst, we make a Boxplot for all columns to detect outliers. We find that 3 columns at least have many outliers( 'StatedMonthlyIncome', 'AvailableBankcardCredit','RevolvingCreditBalance' , 'TotalTrades'), so we drop them.
* After dropping them and plot again we have found that other columns have outliers but if we drop outliers rows, data will be decreased too much so we let them as their was.

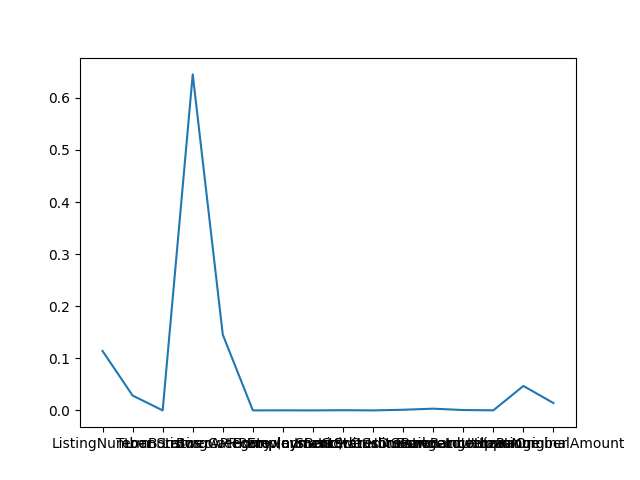
Chart, box and whisker chart

Description automatically generatedNote: after dropping columns (3d,9th) accuracy decreased

* Timeline

  Description automatically generated with medium confidenceThen we get the correlation between columns input and output and found that there is columns can be dropped-if decreasing time is wanted- such as (‘CreditScoreRangeLower’, ‘BorrowerRate’, ‘LoanNumber’)
* Then, after training models we have found that most of models depend on about just 5 columns which are

Chart

Description automatically generated[ListingNumber, Term, BorrowerAPR, BorrowerRate, LoanNumber] so we can drop other columns to reduce time but it decreases accuracy a bit

* Hyperparameter tuning

1. Tree:
2. Chart

   Description automatically generatedMax\_depth: try train with values from 1:30 in max depth and test with validation data to see overfitting point. We found that best accuracy in depth 10
3. Chart, line chart

   Description automatically generatedMin\_samples\_split: by trying values to 30 for min\_samples\_split and use max depth that predicted by the way before, we found that best min\_samples\_split with (maxdepth=10) equals to 23.
4. Decision tree:
5. N\_estimators: we calculate the best number by looping with values from 120:150 with stride 5 and found that best number is: 140

Chart, line chart

Description automatically generated

1. best\_min\_split : We best\_min\_split that calculated in tree which was 23
2. SVM:
3. C: regularization term: by trying values between 0.01 and 0.6, we found that 0.6 is good which can be noted in the following figure.

Chart, line chart

Description automatically generated

1. Kernel: by trying kernels (linear, sigmoid, polynominal, rbf ) we found that rbf is good which is the default for svc with accuracy about 96%

* Chart, line chart

  Description automatically generatedConclusion:

We have found that svm isn’t good because of long time and poor accuracy when it compared with the other models.

Data has outliers although we have tried to decrease it but still in data so, overfitting in this model is not good and we tried to reduce overfitting from tree in decision tree which is the best one of the three with test accuracy more than 98%