Usecase 2:

Lending Club Loan Dataset :

This dataset contains complete loan data for all loans issued through the 2007-2015, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file containing loan data through the "present" contains complete loan data for all loans issued through the previous completed calendar quarter.

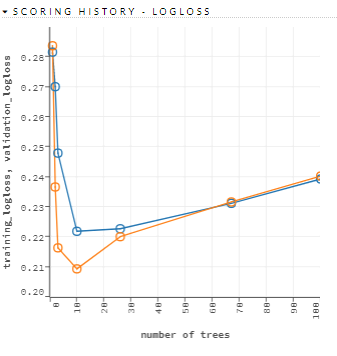
Use case is to predict the default and the fully paid loan status in the dataset.

To predict the bad loans or the default status, I have used Distributed Random Forest.(We can also do it using gradient boosting model, but for more accurate performance, I have used DRF)

Definition of DRF : Distributed Random Forest (DRF) is a powerful classification and regression tool. When given a set of data, DRF generates a forest of classification (or regression) trees, rather than a single classification (or regression) tree. Each of these trees is a weak learner built on a subset of rows and columns. More trees will reduce the variance.

The dataset was split into 70,30(train and test set). I have increased the number of trees to 100 and given nbins=10 as the loan status is of 10 categories for modelling the data.

SCORING HISTORY :



<https://sriyamsoft-my.sharepoint.com/personal/gayatrip_sriyamsoft_com/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fgayatrip_sriyamsoft_com%2FDocuments%2Fscore2%2EPNG&parent=%2Fpersonal%2Fgayatrip_sriyamsoft_com%2FDocuments>

We can observe that in the above scoring history the logloss(error rate) has decreased from 28.16% to 23.93%,I.e., our model has improved by approx. 4%.

The above given link has the data regarding the logloss at each point.

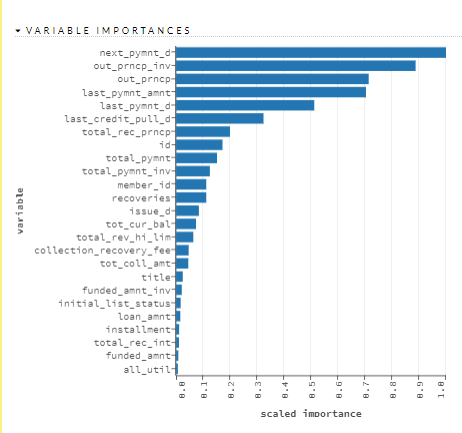
OUTPUT : CONFUSION MATRIX :

<https://sriyamsoft-my.sharepoint.com/personal/gayatrip_sriyamsoft_com/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fgayatrip_sriyamsoft_com%2FDocuments%2Fcm1%2EPNG&parent=%2Fpersonal%2Fgayatrip_sriyamsoft_com%2FDocuments>

In the above given link, it represents the confusion matrix, i.e., actual vs predicted loan status.

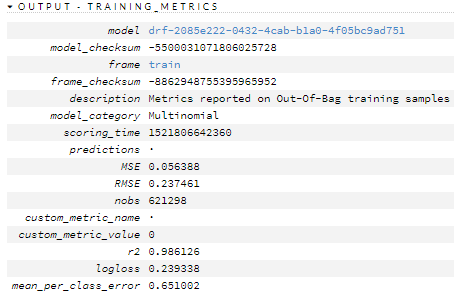
We can be able to know how many people are in the default category and work on improvising the number.

VARIABLE IMPORTANCE :



The above given figure represents variable importance of the model which illustrates which all predictor columns have dependencies or are related to the response columns. This is used for implementing feature engineering, taking into consideration the more important variables.

SUMMARY :



The training metrics summary shows the r2 value is 98.6% which explains that the model is a great fit. The mean square error(MSE) is also very low , I.e., 0.05%, that says the prediction is almost correct.