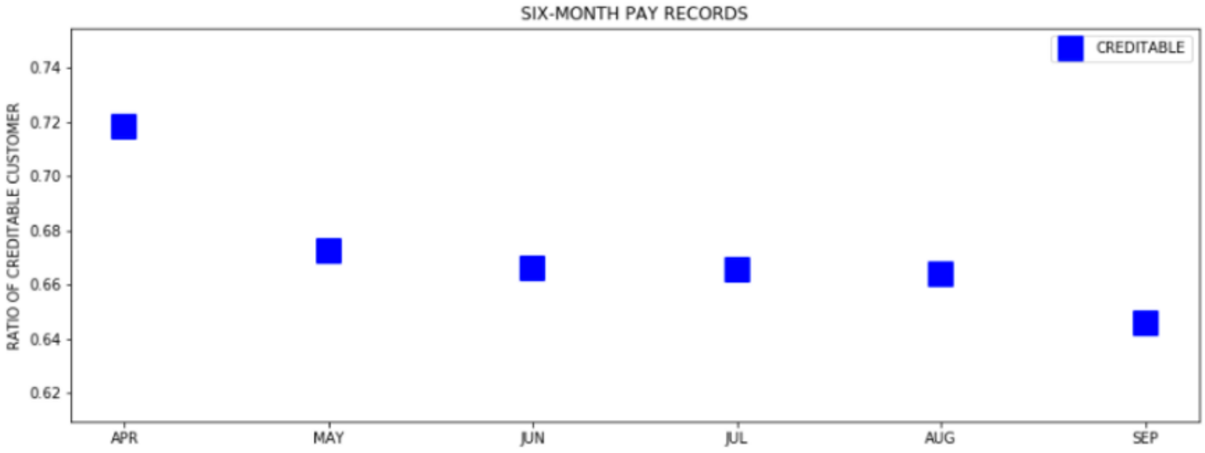
DEFAULT CREDIT CARD CLIENT

This project analyzed default credit card client data set to predict whether a customer will be default or not based on his/her past six months pay records as well as other attributes. Default credit card client data set stored in UCI Machine Learning repository system consists of 24 attributes and 30,000 instances. Various machine learning (ML) algorithm such as Logistic Regression, GaussianNB, Decision Tree Classifier, Extra Trees Classifier, have been applied to predict the default and prediction was validated statistically. Details analysis of the data shows that based on feature selection of the independent variables, the prediction turns out to be excellent with some of those ML models such as Extra Trees Classifier, Random Forest Classifier, and Ensemble Gradient Boosting Classifier. Also, it shows that the ratio of customer who pays on time drops down with time. Based on clients' attributes such as education and marriage pay records of certain groups are consistently better than other groups.

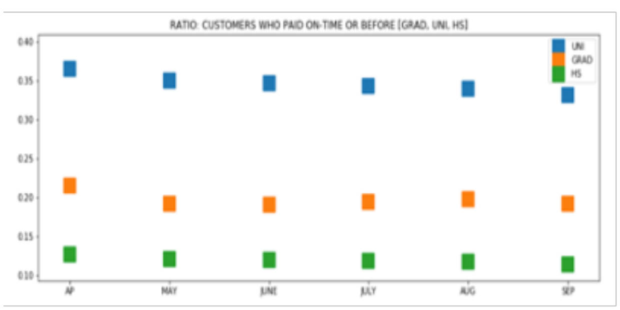
Data Visualization

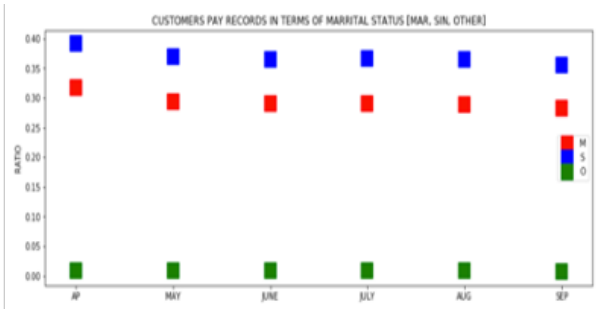
Among the following twenty four attributes such as 'ID', 'LIMIT_BAL', 'SEX', 'EDUCATION', 'MARRIAGE', 'AGE', 'PAY_0', 'PAY_2', 'PAY_3', 'PAY_4', 'PAY_5', 'PAY_6', 'BILL_AMT1', 'BILL_AMT2','BILL_AMT3', 'BILL_AMT4', 'BILL_AMT5', 'BILL_AMT6', 'PAY_AMT1','PAY_AMT2', 'PAY_AMT3', 'PAY_AMT4', 'PAY_AMT5', 'PAY_AMT6' , analysis was carried over based on overall features and selected features as well.

As we analyze the six pay cycles, we found the number of clients who paid on time or before the time decreases with pay cycles as shown here.



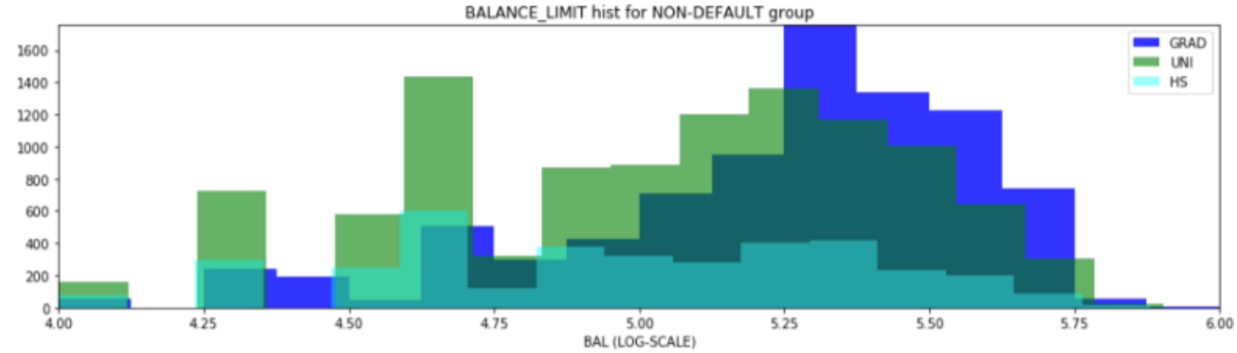
However, in terms of other attributes such as education, marital status we found that certain groups pay records are consistently better than other groups.





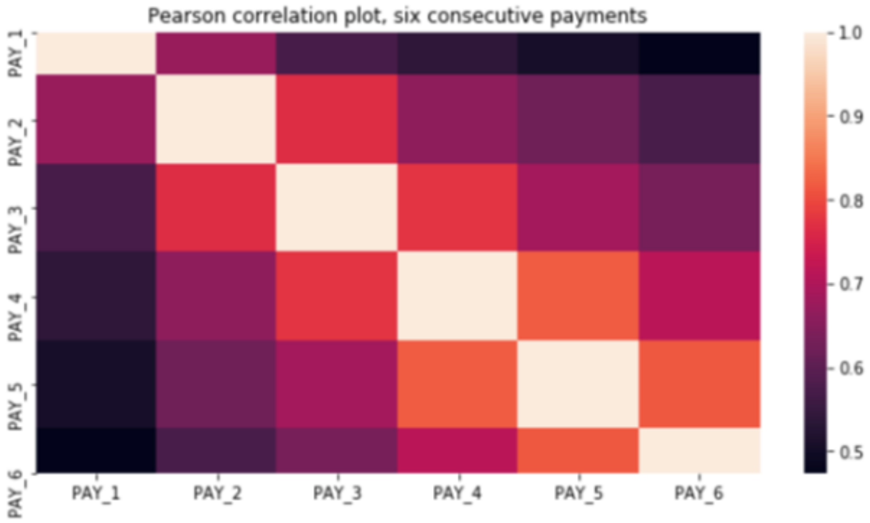
In terms of marital status single and other groups pay records are better than married group. In terms of age group, we found 27-32 years age group is the largest among clients irrespective of default or not default status.

In terms of credit limit, the largest group of customers credit ranges from a tenth of a million to three quarters of a million, but the density of grads mostly populates towards high credit limit end in the hist.

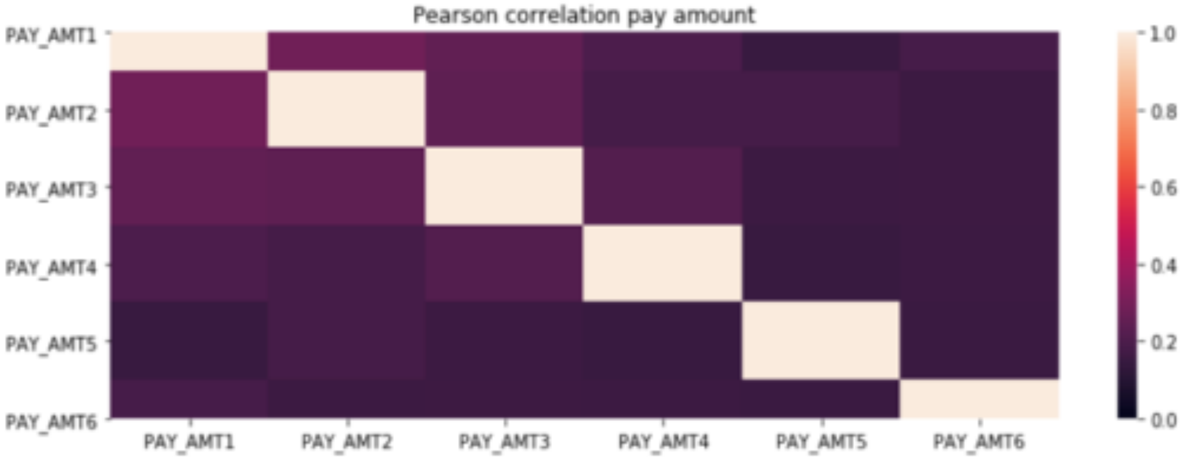


FEATURE CORRELATION:

We found consecutive pays (whether paid on time or not) are more correlated (items of correlation matrix ranges from 1 to 0.5)

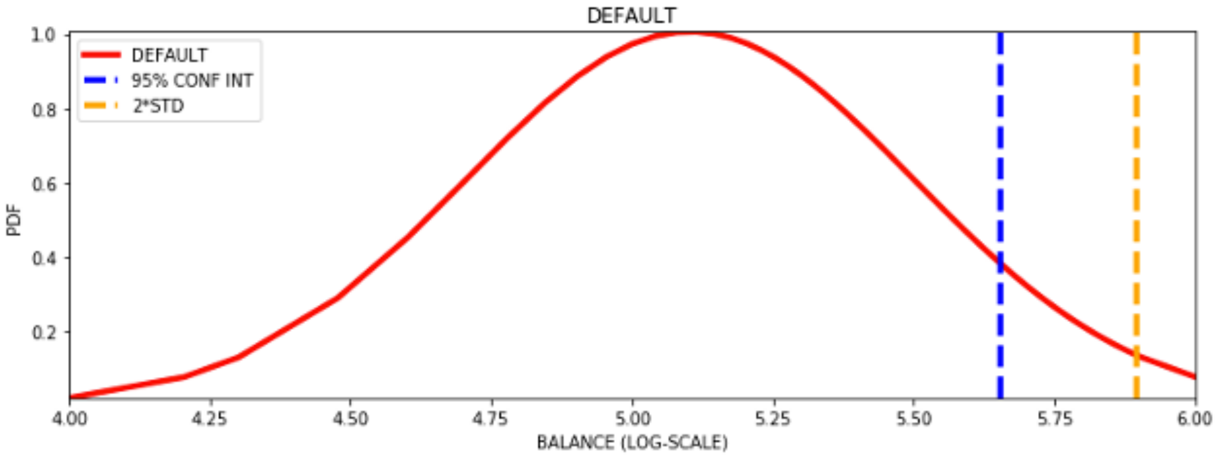


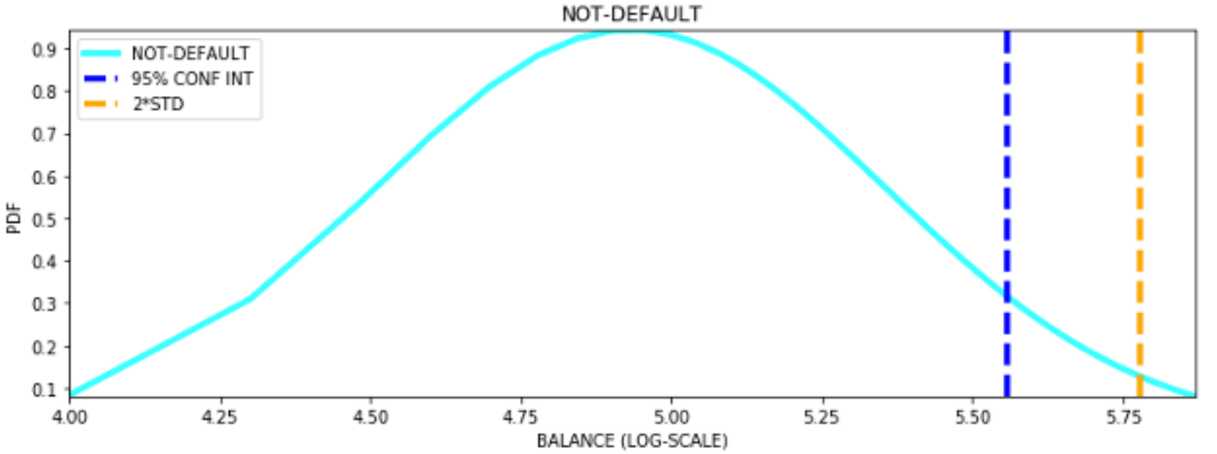
rather than consecutive pay amount or the ratio of pay amount to bill amount.



STATISTICAL ANALYSIS

We found over all clients' credit card balance or clients grouped in terms of various attributes such as default, not default, education and marriage and found the 95% confidence interval with respect to the mean is around the two times the STD of the distribution.

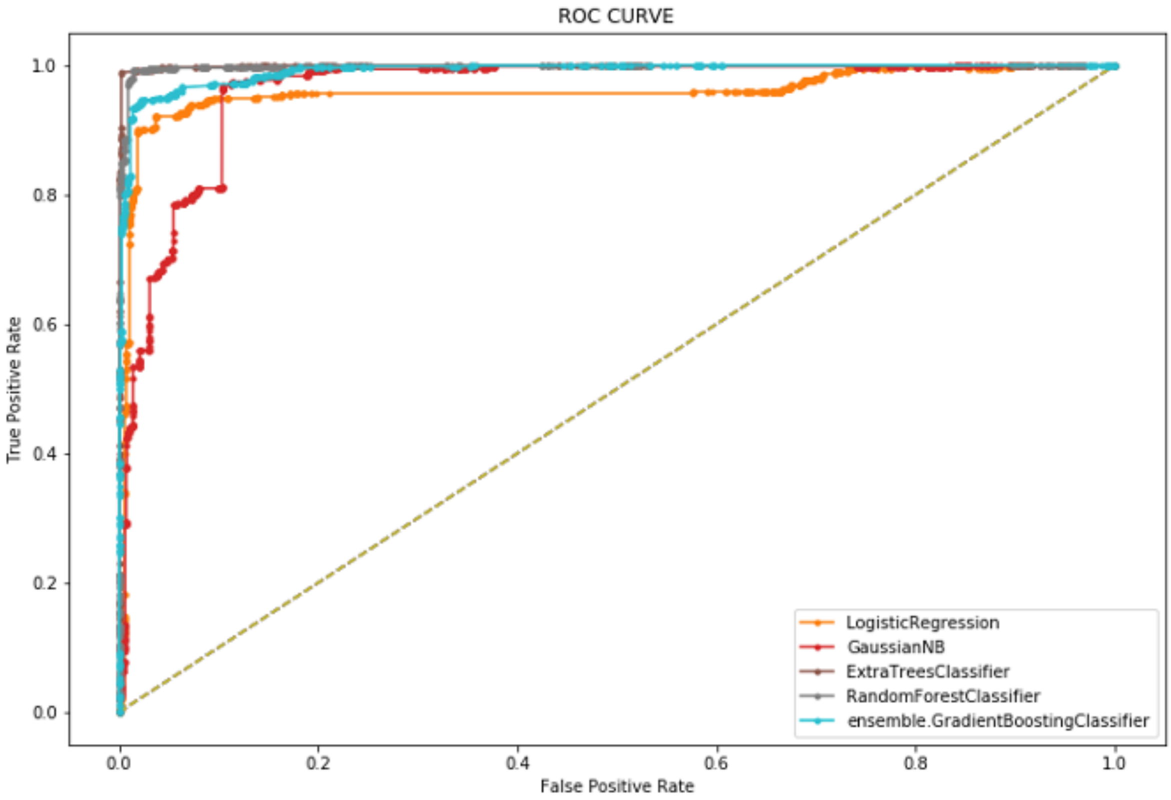




Though the distribution of the those groups are similar, however statistically those groups are not identical as we see from the ttest results, since the p values are less than 0.05.

MACHINE LEARNING PREDICTIONS:

We applied various ML algorithms to predict credit card default and prediction performance of the model was validated via Cross validation, grid search, ROC Curve. Among various ML model, Decision Tree, Extra Trees and Random Forest Classifier predict excellent ROC square based on feature selection.



BUSINESS VALUE

ML cross validations, grid search, feature selection is very important for finding the best algorithm for better prediction. Since total balance limit of all customer was around five billion, so based on prediction future customer selection is very important from business point of view.

RESULTS

Default credit card clients is a six-month data of customers pay records and other attributes as well. Based on analysis the following items can be highlighted

o Ratio of customer who was never late to pay the bill vs total number of customer decreases with time.

o In terms of attributes such as education, marriage, age we found certain group pay records are consistently better than other groups such as grad pay records are better than HS student, on the other hand university students pay records are better than grads.

o We also found that the 95% confidence interval of the PDF of the clients' balance distribution (log scale) with respect to the mean is close to two times the STD.

o We also found based on certain feature selection the overall prediction accuracy turns out to be excellent.

o Above claim was validated statistically. Also, we found ensemble gradient boosting classifier is the best classifier for ML prediction.