IEE 598: DATA SCIENCE FOR SYSTEM INFORMATICS

FINAL PROJECT

Prediction of the Loan status using Machine Learning Classification Algorithms

Submitted under the guidance of

Professor Hao Yan

By,

Amrut Deshpande: 1211537636

Bharath Hebbar: 1211588024

Mithun Muralidhar: 1211309824

CONTENTS

Sl No	Contents	Pages
1	Motivation and Objective	2
2	Introduction to the problem setup	2
3	Initial Exploration and visualization	2
4	Data Challenge, Feature Engineering	3-4
5	Models, Solutions, and Interpretations	5
6	Future work	7
7	Appendix	8

IEE 598 PROJECT

MOTIVATION AND OBJECTIVE

The motivation is to utilize the techniques and concepts of the Data Science for System Informatics class with a focus on learning and building good Machine Learning algorithms for large-scale Prosper loan dataset to classify its multi class loan status variable.

The main objective is to build a classification model to successfully predict which loans will default by computing the key high-performance metrics for the classifiers and observing the learning curve behavior of training and testing sets. Further objective is to lay emphasis on data preprocessing stages and on feature engineering to extract important features that aids our analysis in making better predictions. Focus will be laid on exploratory data analysis in the beginning to understand the behavior of the variables and assess importance of predictors with response variable.

All three of us were interested in this project when we discovered this Udacity dataset, as this data will test Big Data challenge of a Financial Industry as an application of our data analysis skills, and felt that this would enable us to contribute in an important area of data science.

INTRODUCTION TO THE PROBLEM SETUP

As the objective suggests, the problem here is to identify the entries which gets defaulted out of the 100000 entries. We setup a master dataset by cleaning the initial unprocessed data. Initial analysis begins with identifying the significant or affecting variables for the loan status and this happens through intuition, visualization, and research about the credit and loan sector. This helped us decide the key variables which have impact on the decision making.

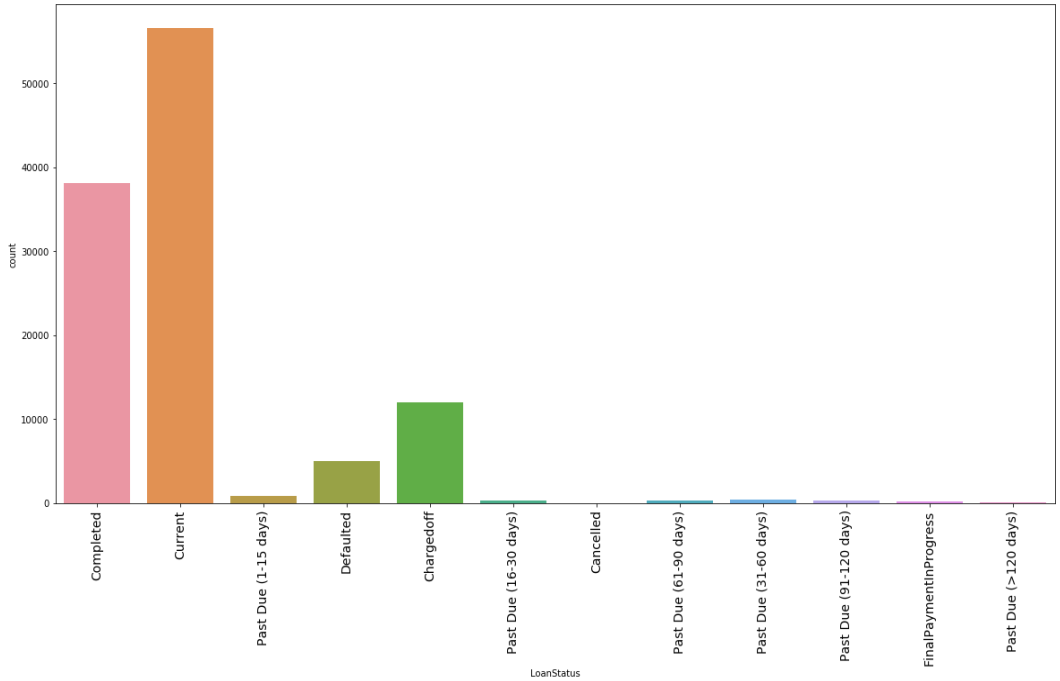
Once the significant variables are identified, then comes the feature engineering techniques like dummies, feature hashing, and transformations on categorical variables to obtain their numeric representation.

The prime problem here was the dimension and the big data. This problem had a setup of Principal Component Analysis to deal with the former and a SGD partial fit classifier for the latter. Principal component analysis (PCA) is a technique used to emphasize variation and bring out strong patterns in a dataset. The partial fit estimator implements regularized linear models with stochastic gradient descent (SGD) learning: the gradient of the loss is estimated each sample at a time and the model is updated along the way with a decreasing strength schedule (aka learning rate).

Finally, a few chosen Machine learning classifiers are partially fit through SGD Classifier and their performance is computed through relevant metrics such as AUC, F1 Score and Recall. Classifiers were subjected to regularization by introducing ‘L1’ and ‘L2’ penalties. Parameter tuning was handled manually as the model was implemented through SGD partial fit. A learning curve was plotted to visualize the trend in the metrics over several chunk iterations. Best Classifier is chosen considering the high metric value and the behavior of the learning curve to judge the complexity and good model fit.

INITIAL EXPLORATION AND VISUALIZATION

The visualization is not just representation of the variables. It plays a key role in exploring the characteristics of both the numerical and categorical variables and in inferring the inter-relationship with Loan status. Out of the 0.1 million data we remove half of the observations as it deals with the current status of loans which is out of our scope in the project. The visualizations involve some of the best plots from the seaborn and matplotlib packages such as pie chart, bar chart, histogram, distribution curve, joint plot, box plot, violin plot, etc.



The first stage involves the plots of the loan status. Pie chart and bar charts were used in estimating the number of all the classes present in the target variable loan status. Then, we progress to the numeric information of the data wherein we first obtain box plots of Employment status duration and Income metrics - Stated Monthly Income and Debt to Income Ratio with loan status which explains the relationship with the loans being defaulted.

We then proceed to Joint plot between Borrower Rate and Lender Yield which shows a perfect linear relationship. Then Bar plot is plotted between the Credit scores such as Prosper rating, Prosper Score, Credit Lower Range, and Credit Upper Range and the loan status which shows a linear relationship for both prosper rating and scores, whereas the upper range and lower range behave similarly against loan status.

Going ahead, Box plots are plotted for Credit history and Credit information data. Violin plot was used to explain the distribution of term and loan original amount with loan status. Distribution plot is plotted for Borrower Rate and LenderYield variables to show the behavior.

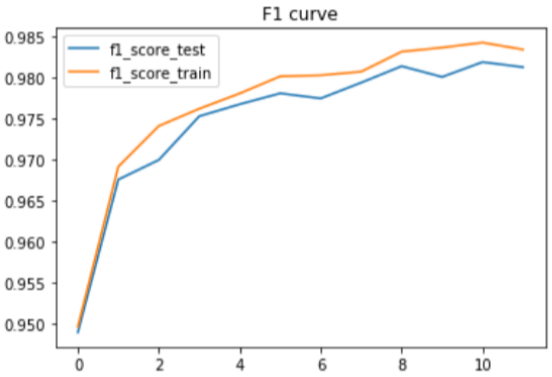
Secondly, we explore the categorical data mainly with the help of bar chart. The listing category shows the reason listed for the loan disbursements and the bar plot helps in estimating the counts of every list. Similarly, the Borrower state and occupation counts and the distribution is obtained through bar plot and a keen observation is made on which of the class of these three variables is close to getting defaulted.

The Boolean variable of Home owner too gets a bar plot to obtain a result if the customer is a home owner or not and finally we infer which employment status has less chances of getting defaulted and strangely part time workers win over full- time employees and it shows they are more regular in loan payments.

DATA CHALLENGE

Big Data Challenge: After completing the data preprocessing stage and feature engineering, we obtained a final master dataset with 375 variables and 57,361 observations. We used SGD classifier to deal with the big data challenge, by partially fitting the data with suitable chunk size. The SGD classifier was also subjected to penalization through ‘L1’ and ‘L2’ and the performance of the classifier was measured by plotting learning curve of each classifiers relevant metrics.

Learning curve for SVM model is as shown below,



High dimensionality challenge: After creating dummies for categorical variables, the size of the predictors increased greatly to 11830 columns. This high dimension challenge was addressed using feature hashing technique for one categorical variable that had almost 9000 unique values (date values). The final master data now contained only 375 variables. Further Principal Component Analysis was employed to reduce the dimension of the data by retaining features that explain most of the variance in the response variable.

Non-linearity challenge was addressed through non-linear classifiers such as MLP, Random Forest.

Lot of time was spent on data pre-processing stage, to combat the missing values, convert variables to relevant data-types and get a master dataset. Additionally, the dataset contained number of domain specific discrete and continuous numerical data along with categorical variables. We had to deal with the skewed distribution of few numerical data and extract a normally distributed variable which would enable in better representation and distribution of those variables.

FEATURE ENGINEERING

After cleaning the data, we have total of 63 features, out of which 56 are numerical and 7 categorical. We visualize the distribution of values of numerical data to see if any of the distributions are skewed and if it requires any transformation to stabilize the variance and make their distribution normal. For this cause we apply log transformations after which the values are somewhat distributed Gaussian or normal-like. Now the numerical data is prepared and ready to be used to train the model.
Coming to the categorical data, we notice “Income-Range” feature is of ordinal type and it is necessary to map and transform them into their numeric representation. Hence, we map Income-Range into numeric values from 0-6, and later create dummies since we cannot directly feed them to the algorithm although being numeric because the algorithm incorrectly interprets according to the magnitude of the values. For the remaining categorical variables except “FirstRecordedCreditLine” we create the dummy variables. Now on observing “FirstRecordedCreditLine, it has over 9000 unique values (obviously because it is the date the first credit line was opened). Hence it makes sense to use feature hashing to encode them rather use dummy variables to convert them into numerical values. Here we set a pre-defined vector of length 150, so that the hashed values of the features are used as indices in this pre-defined vectors and the values are updated accordingly. So now in the end we have a total of 322 columns. Now combining the numerical and the dummy variables we get a total of 376 features which can be used to train the model.

IEE 598 PROJECT

MODELS, SOLUTIONS, AND INTERPRETATIONS

PERFORMANCE METRICS OF ALL MODEL

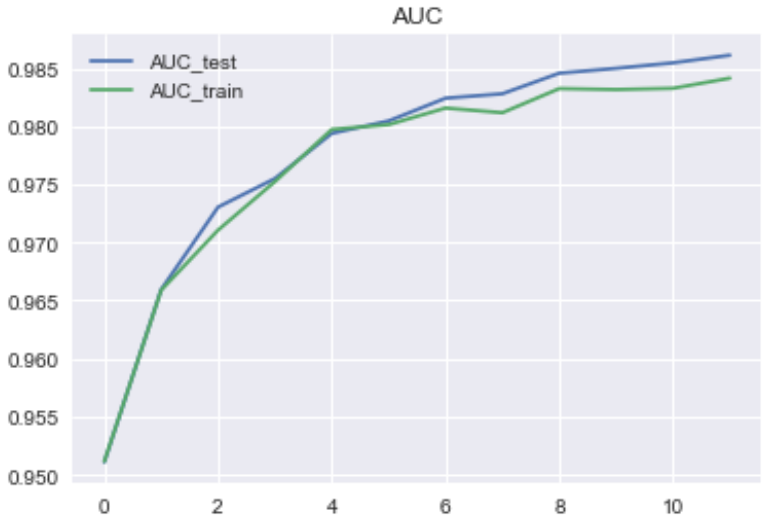
BEST MODEL: BINARY CLASSIFICATION

		Without PCA			PCA		
Classifiers	**Target Variable Type**	**AUC Score**	**Recall**	**F1 Score**	**AUC Score**	**Recall**	**F1 Score**
Support Vector Classification with L1 Regularization	Binary		0.97	0.98		0.97	0.97
Logistic Regression with L1 Regularization	Binary	0.99			0.98		
Support Vector with L2 Regularization	Binary		0.97	0.95		0.96	0.94
Logistic Regression with L2 Regularization	Binary	0.99			0.96		
Bernoulli Naïve Bayes	Binary	0.97			0.90		
Random Forest Classifier	Binary		0.99	0.99		0.73	0.83
Multi-Layer Perceptron	Binary		0.94	0.98		0.84	0.96
Support Vector Classification with L1 Regularization	Multi Class			0.93			0.88
Support Vector with L2 Regularization	Multi Class			0.92			0.86
Multi-Layer Perceptron	Multi Class			0.93			0.84

Binary – Logistic Regression with L1 penalty (LASSO)

Logistic regression model with L1 penalty is the best model performer to predict the classes for the binary type classification. The average Area under the curve score indicates that the model correctly predicts the classes with a probability of around 0.98. The curve is gradually increasing with increase in number of iterations, which suggests that the model is a good performer in predicting the classes accurately.

The AUC score of 0.98 is same before and after applying PCA which suggests that although the score being the same, PCA aids in extracting only the required features which best explains the model thereby reducing the complexity of the model.

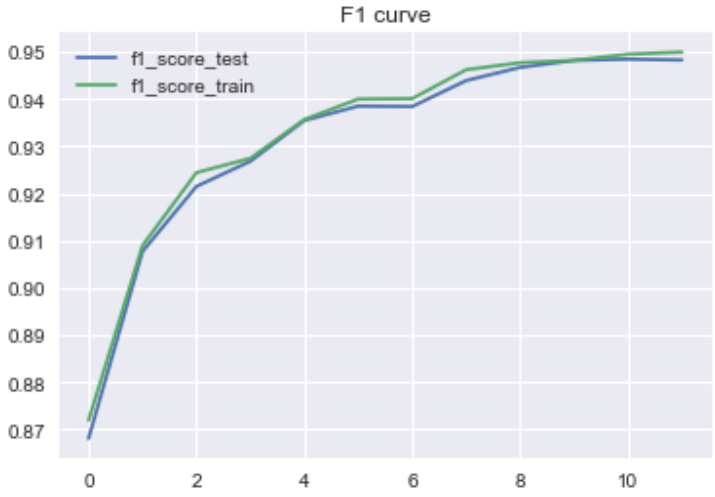


BEST MODEL: MULTI-CLASS CLASSIFICATION

For the multi-class classification, MLP Classifier has the highest f1 score among the classifiers. It is the weighted average of precision and recall score for each class in a multi class scenario. F1 score indicates to what extent the model correctly predicts all the classes (multi class case). Closer the score to 1, better is the model performance.

Although SVM with L1 penalty has a similar score, comparing the graphs of both the classifiers over many iterations we can infer that the MLP has a comparatively smoother curve, which indicates its robustness towards partial fit.

It was noted that the f1 score for MLP Classifier reduced when the features were trained into the model after applying PCA. This might be because, since the target variable has 11 unbalanced classes, it probably requires initial number of features (prior application of PCA) to have a good performance on the data.



MODEL CRITICIZATION

As noticed, the performance metrics for almost all the classifiers topped when it came to Binary Classification. The reason can be well noticed here as we clubbed several categories of loans that may be subject to default under default category. This binary reduction enabled the classifiers to easily categorize the loans as bad or good.

When it comes to Multi-Class Classification, the Loan Status target variable has 11 different categories. Here the performance of all the classifiers reduced to a certain extent when compared to binary classification.

Since, the model classification was handled by incremental learning through SGD Classifier, tuning the parameter was manually done because partial fit does not support implementation of cross-validation to fine tune the parameters. Also, as the specific classifiers are executed by defining the loss function in SGD Classifier, individual classifier's hyperparameters option is not available in this case.

SUMMARY

This project was a great learning platform to handle huge real-world datasets and perform extensive analysis to generate a good model. A lot of time was devoted to data pre-processing and was a great experience to confirm the fact that almost 80% of the time is spent in pre- processing and remaining 20% in analysis. Feature engineering is a vital step in Data Science as it extracts new features from existing features that makes more sense to the data. We deployed feature hashing technique to convert huge categorical variable into its numeric form.

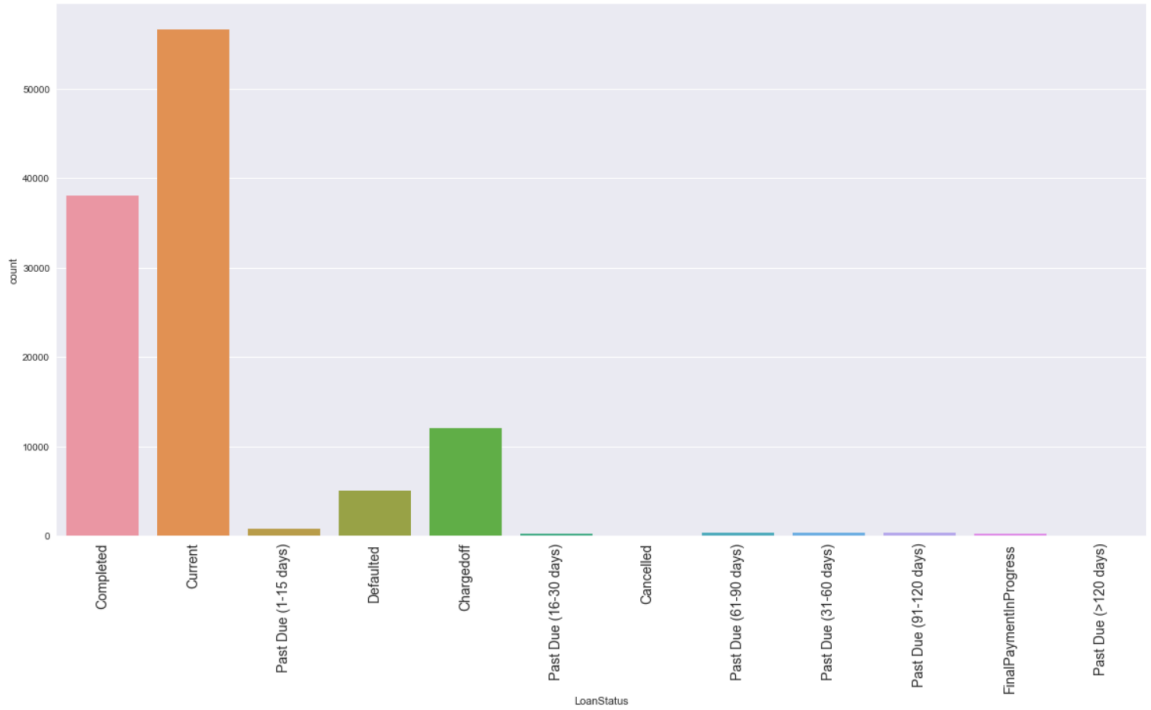
Having set the base for some Machine Learning classifiers, the big data challenge was addressed through partial fit. Dimensionality reduction was addressed through PCA. As the target variable is highly unbalanced, we chose specific performance metrics relevant to classifiers that are robust to imbalance. For Binary classification, the pos_label was set to '0' as our interest is focused more on predicting the default status of the loan. The best model for Binary classification is logistic regression with L1 penalty scoring a high AUC value of 98% and for multi-class classification, the best model is Multi-Layer Perceptron (MLP) scoring a good f-1 score of 93%.

FUTURE WORK

As the big data challenge was addressed through incremental learning, hyper-parameter tuning was not possible. We would like to explore Big Data frameworks such as Pytorch/ Tensor flow to combat the big data and also perform cross-validation to select the best tuning parameters for classifiers and see if this would result in increase in performance of classifiers in multi-class classification.

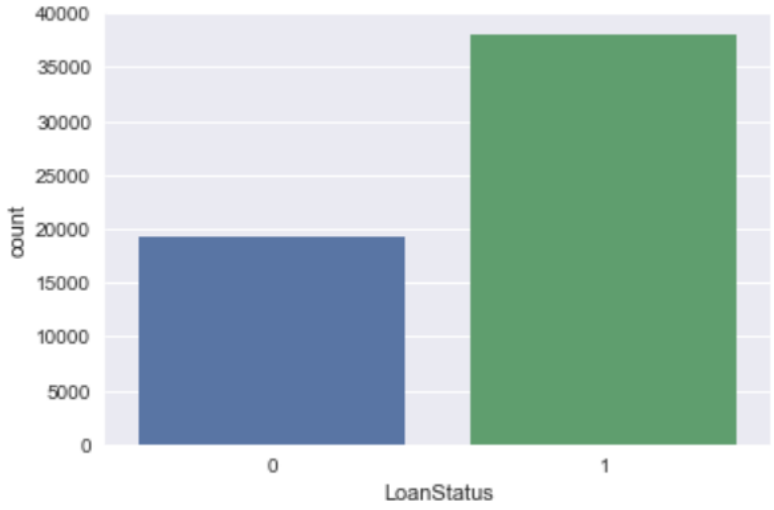
APPENDIX: Visualization Section

Target Variable : Loan Status



From the above visualization, Current loan status ranks the highest with almost nearly half of the dataset, followed by 'Completed' loan status with almost 39000 counts. Chargedoff and default follows the ranking list and the rest are fractionally distributed.

Binary transformation of Loan Status



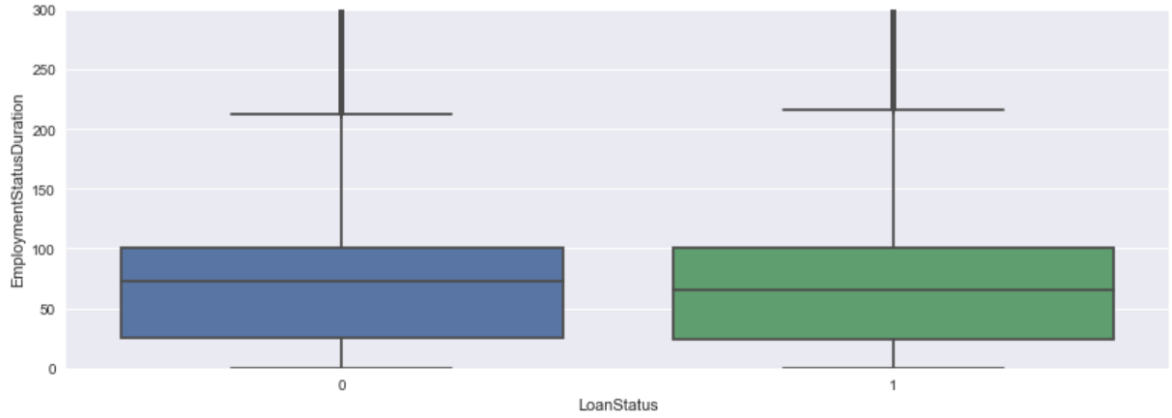
The current status of loan is removed from the vizualization as that is not part of the goals of the project. We only focus on historical data to build a predictive model.Further, all other classes of loan status are divided into binary format, categorizing them into good loans vs. bad loans. The one status is Completed which is a good loan and the other class has the mixture of all other sub classees

such as Chargedoff,Defaulted,Past Due (1-15 days), Past Due (31-60 days), Past Due (61-90 days), Past Due (91-120 days), Past Due (16-30 days), FinalPaymentInProgress, Past Due (>120 days), Cancelled and these are bad loans.

From the above bar chart, we can conclude that the majority class is of 1 i.e completed and the other class of 0 has about half the data of the completed class. Our sole focus is on the class '0' which is bad loans and is considered to be a true class for our analysis.

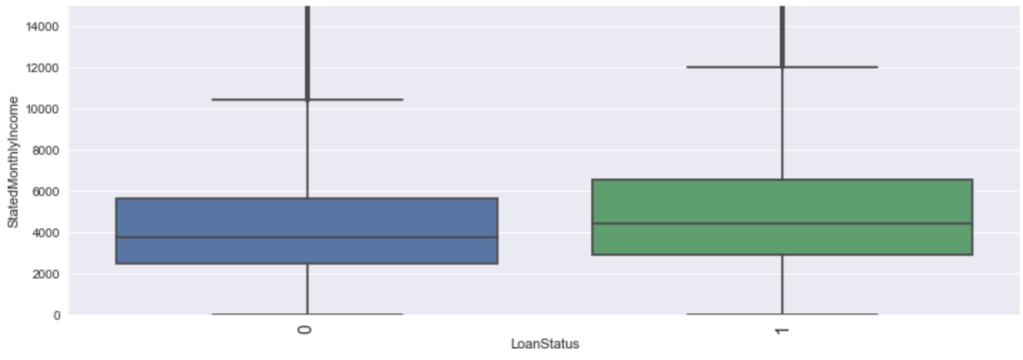
Exploring Numeric information of the data

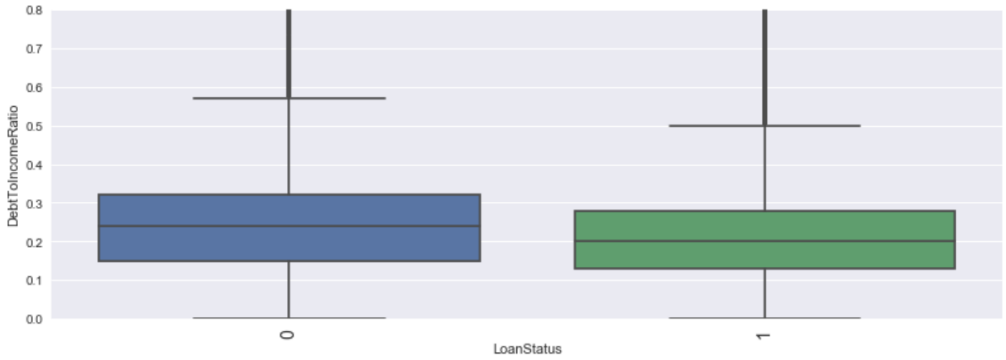
Employment status duration



The box plot shows the median status duration falls around 50 for both categories of loans. The lower and upper qaurtile including the range also dosen't really differ for Loan Status. It is quite evident from this plot that there is hardly a relationship between EmploymentStatusDuration and loan default.

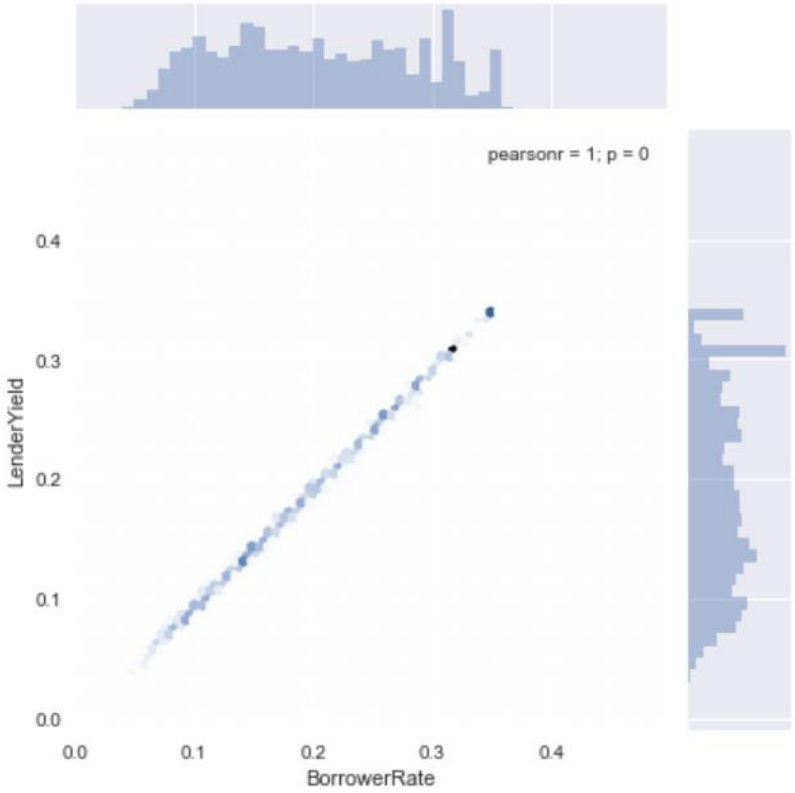
Income metrics - Stated Monthly Income and Debt to Income Ratio





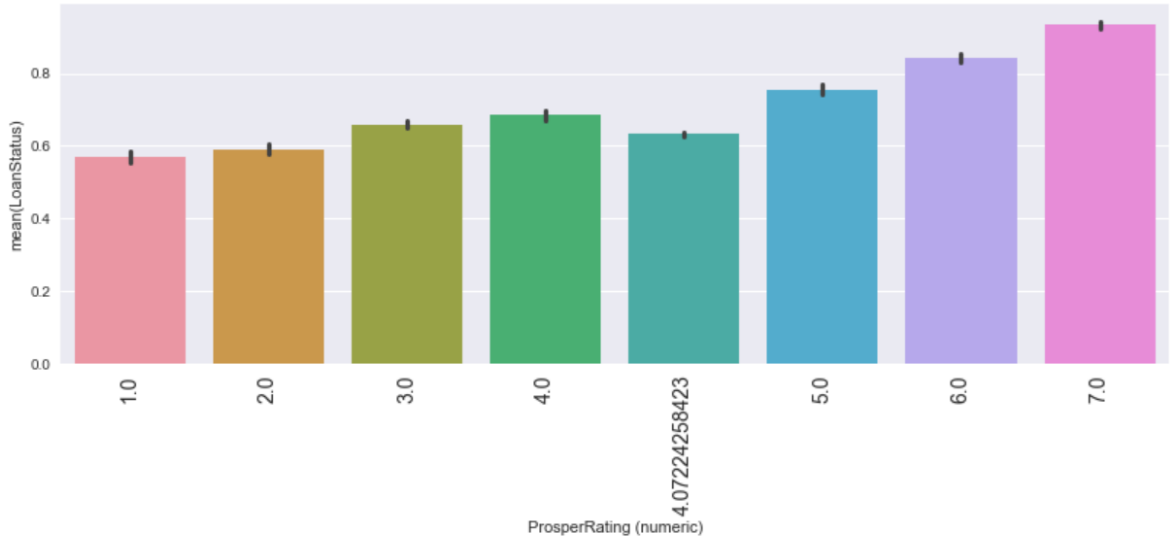
We observed the state monthly income and the debt to income ratio have a relationship with default i.e People with higher stated incomes defaulted less often than those with lower incomes. The range for monthly income for good loan status is also higher comparitively. People with higher debt to income ratio have more defaulted loans.

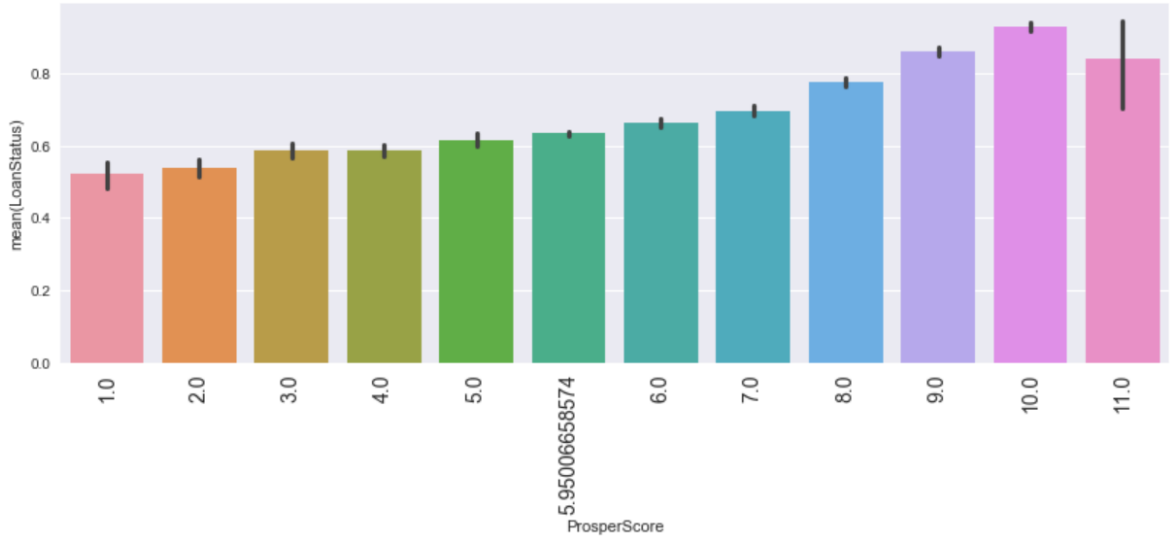
Borrower Rate and Lender Yield



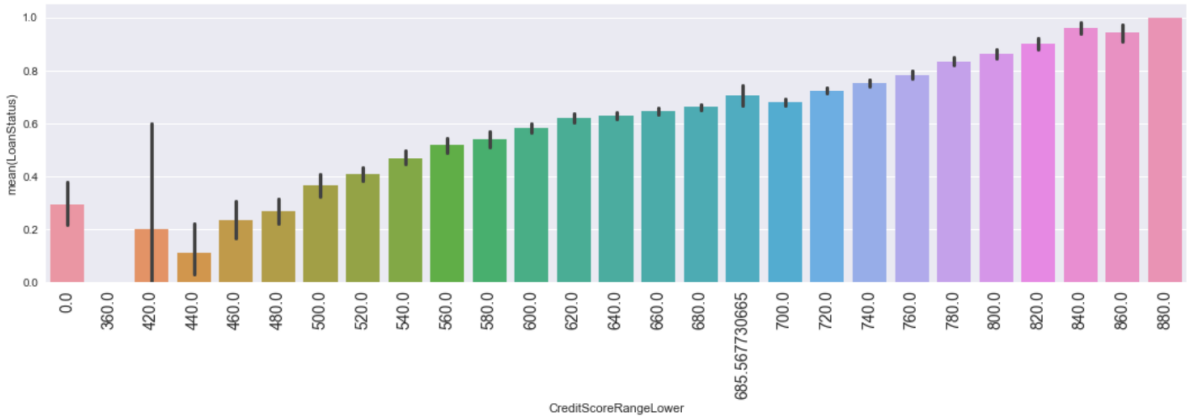
The above jointplot shows almost a perfect positive correlation between LenderYield and BorrowerRate. Higher the rate, higher is the lender yield. This makes sense to the data.

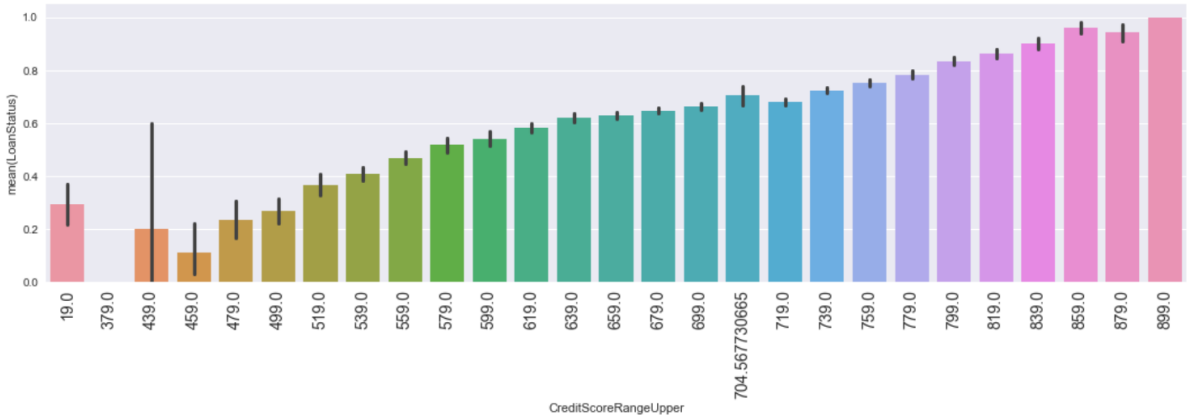
Credit scores: Prosper rating, Prosper Score, Credit Lower Range, Credit Upper Range





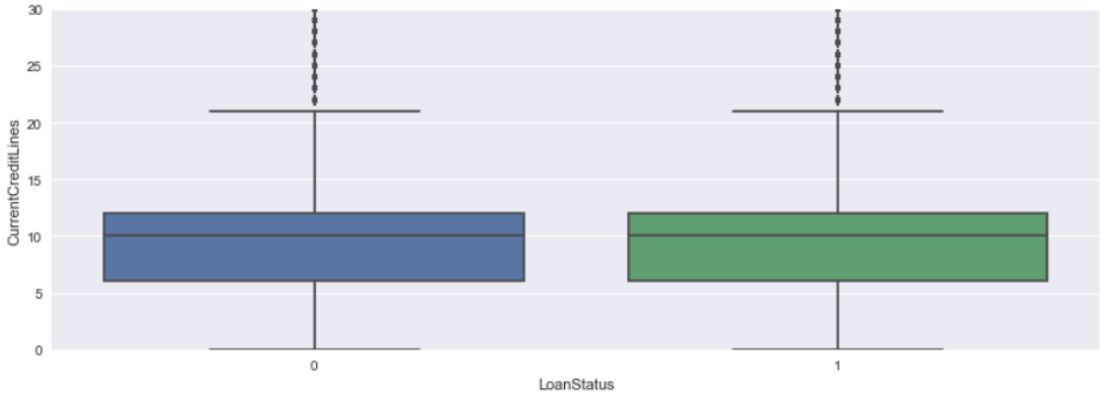
Higher the Prosper rating and ProsperScore, better the status of the loan being completed. Both the Prosper rating and Prosper score are linearly dependent on loan status and are doing pretty good in predicting the default. We can observe that as the rating number increases the probability of loan being defaulted increases. There is some unusual behavior noticed at ProsperScore valued 11, it seems to default more than its predecessors.

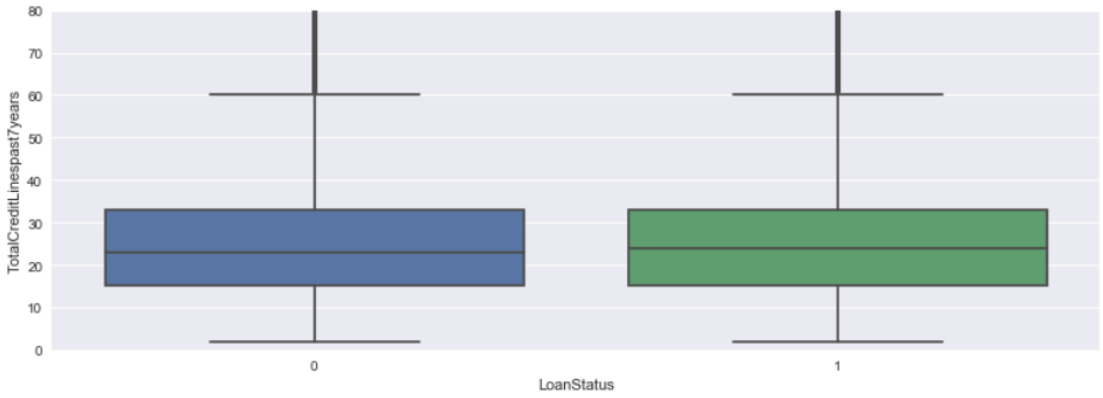


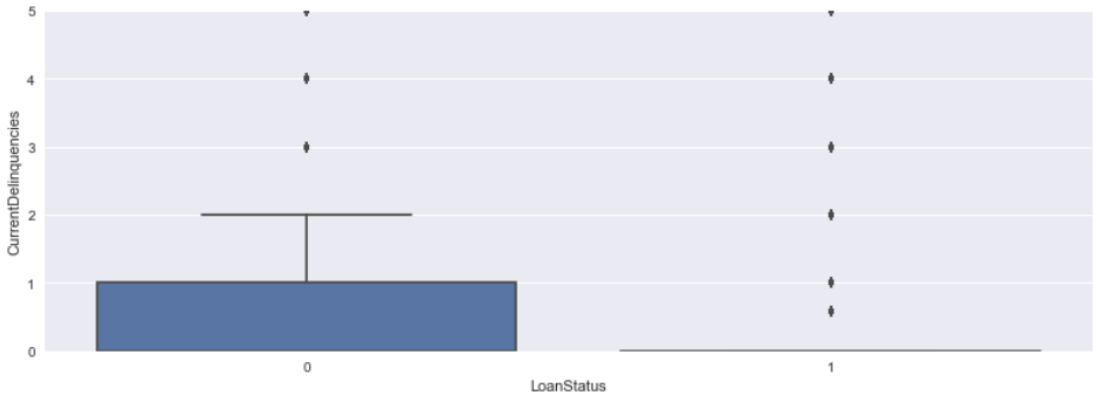


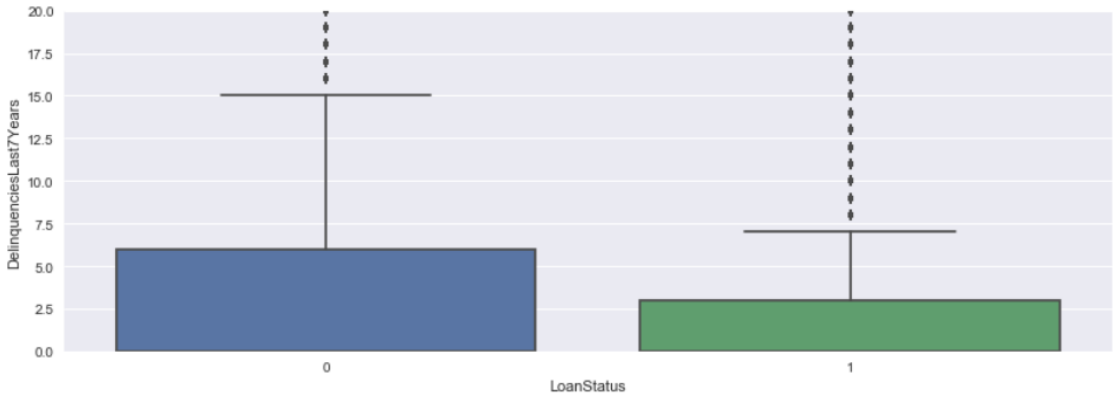
It should be noted here that the credit score "range" seems to be constant and behaves the same.

Credit History



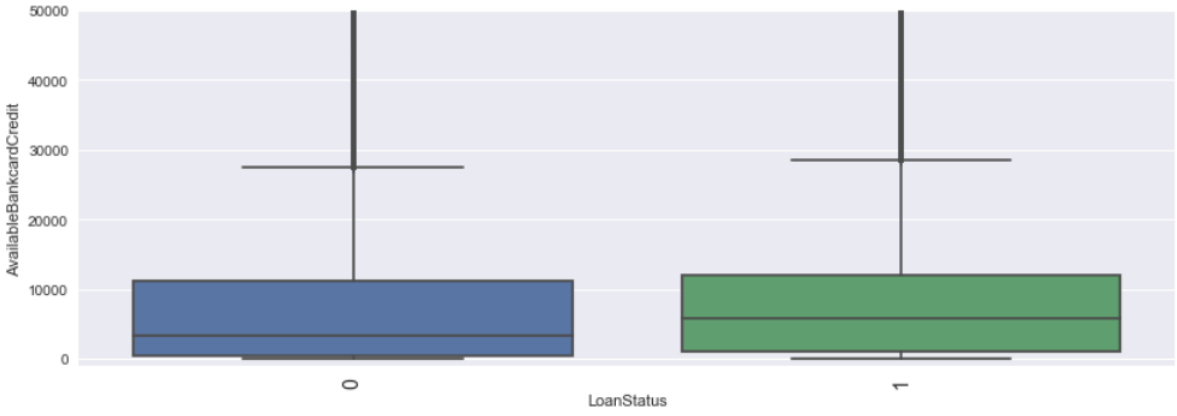


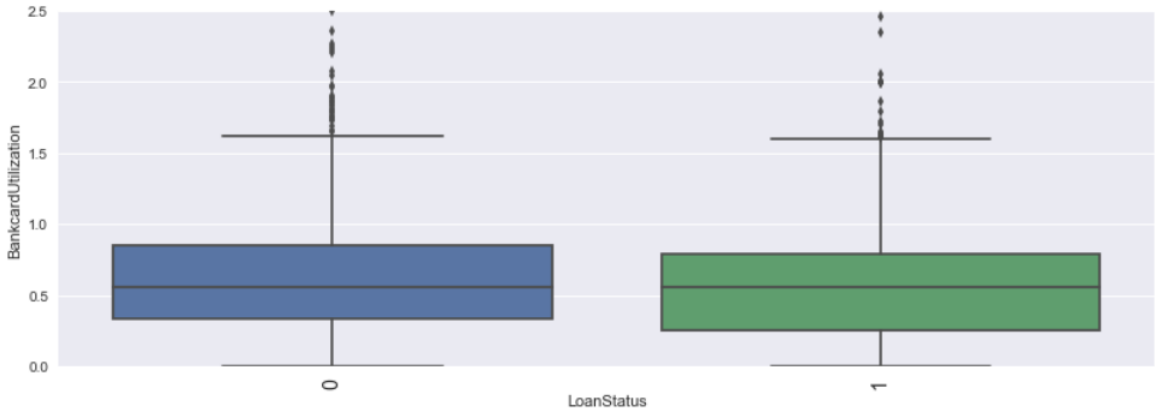




Credit Lines dosen't have significant relationship with Loan Status. However Delinquencies does seem to have some mild relationship.

Credit Information



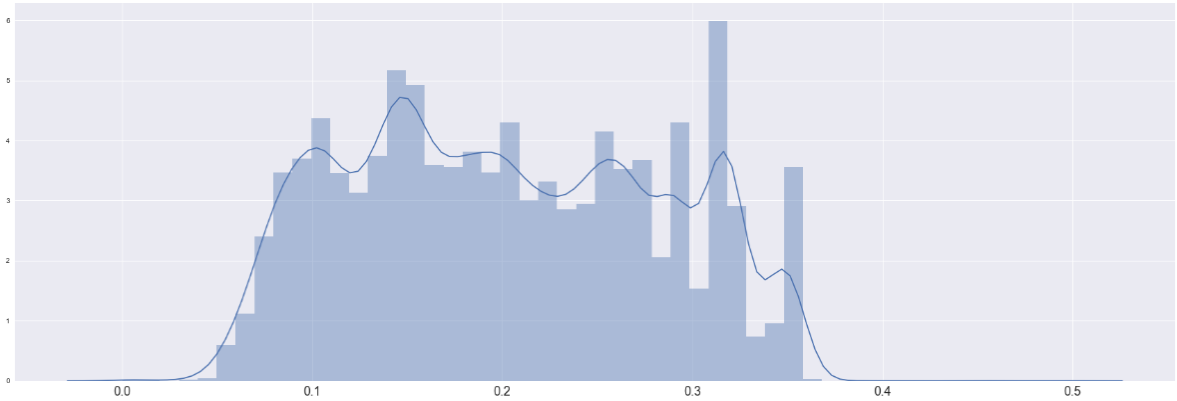


Loans that are subject to default has lower available bank card credit. Bank card utilization seem to be constant with Loan Status, however lower the proportion of bankcard utilization, lower the probaability of default.

Loan Characteristics

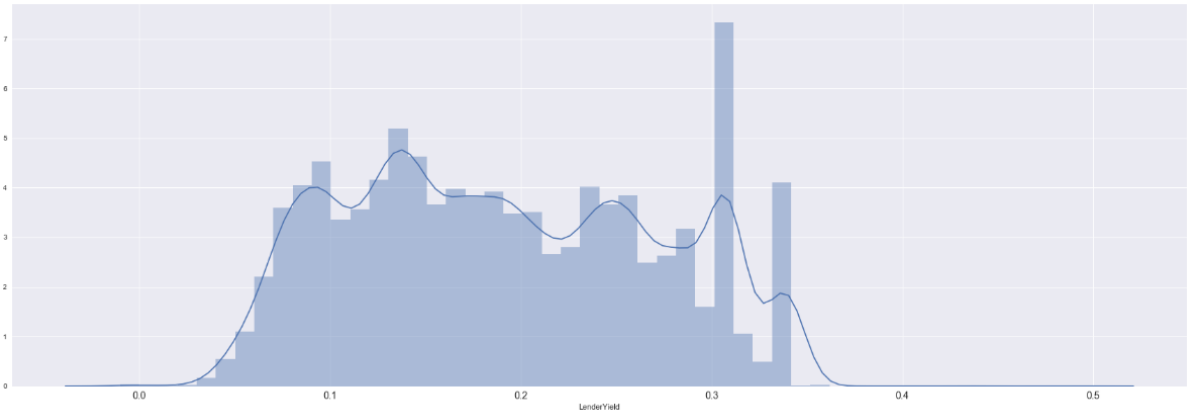
The violin plot helps in explaining the distribution of most of the points involved in the plot. In the above plots, the distribution appears to be similar for both variables against Loan Status.

Borrower Rate



Borrower Rate is normally distributed with a mean interest rate of 20.16%. Upper quartile is paying an interest upto 27%, while the lower quartile is paying upto 17% interest rate.

LenderYield

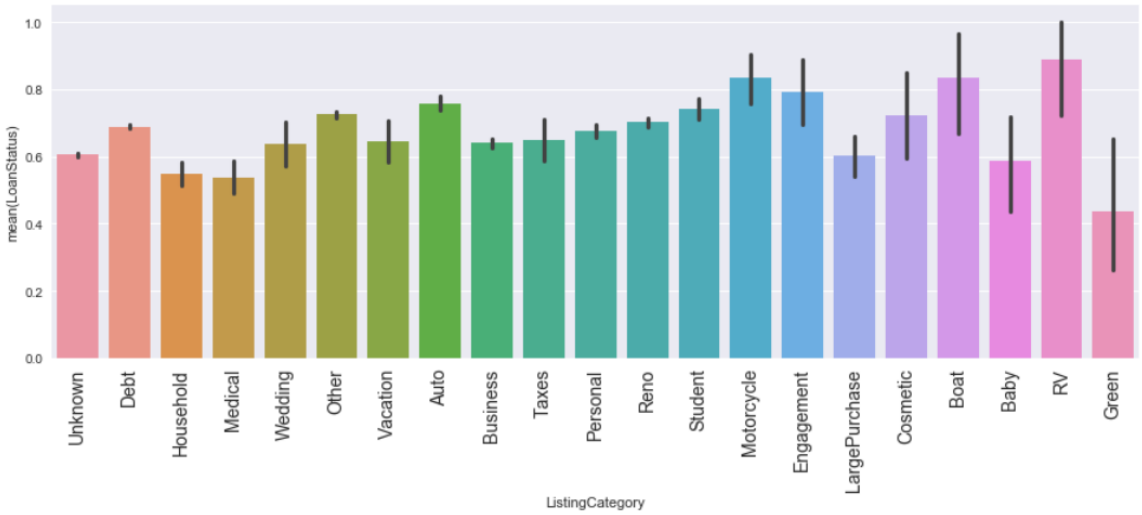


Lender yield distribution is normally distributed, mean of this distribution is around 0.19 and there is a yield with sharp peak ranging around 0.5, seems like this yield on the loan is with higher interest rate.

Exploring the categorical information of the data

Listing Category

The data type of this variable is numeric by default because of different loan types being distinguished by numbers. It is important to convert this to categorical type for better interpretability and avoiding false ordinality.



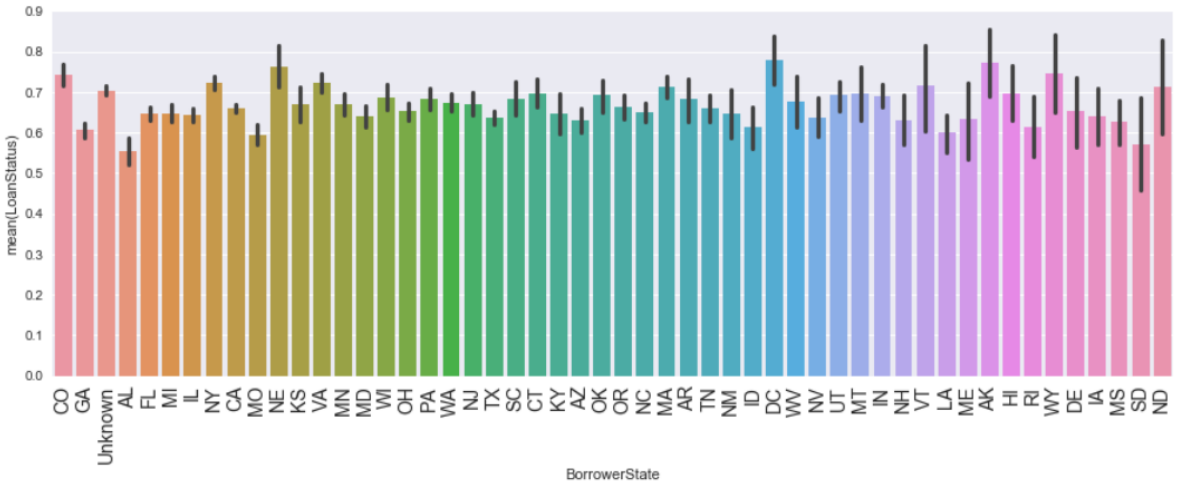
This visualization gives insights about probability of types of loans that are subject to good/ bad loans. Considering a threshold of 80% that makes excellent loans, loans such as 'Motorcycle','Engagement','Boat','RV' tend to complete their loans within the specified time.

It is evident to see 'Student' loans which is ususally a huge sum taken, to get completed almost 70% of their time.

Loans such as 'Household','Medical','LargerPurchase','Baby' complete their loans close to 60% of their time.These loans fall under common categories and usually have huge sums of loans taken, reason for which accounts to 60%.

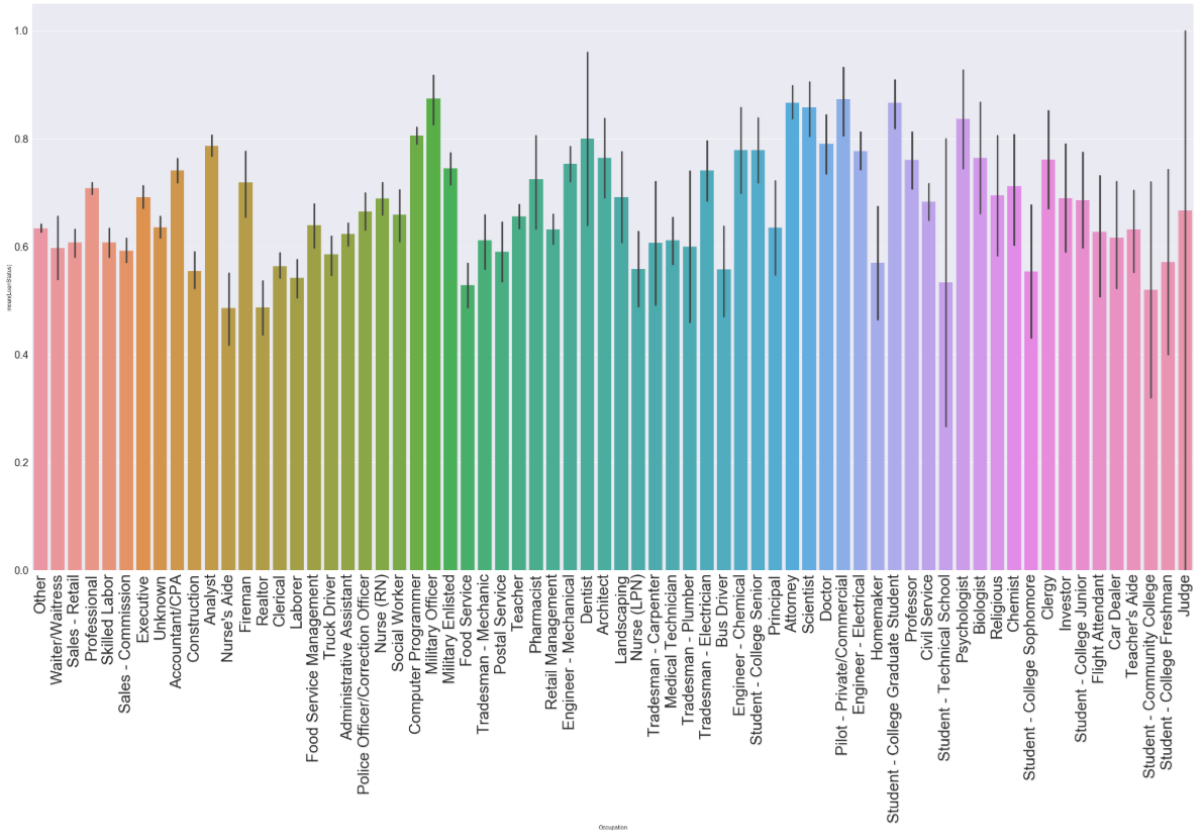
'Green' loans are subjected to default almost 60% of their time. These loans are probably deemed as bad loans.

Borrower State



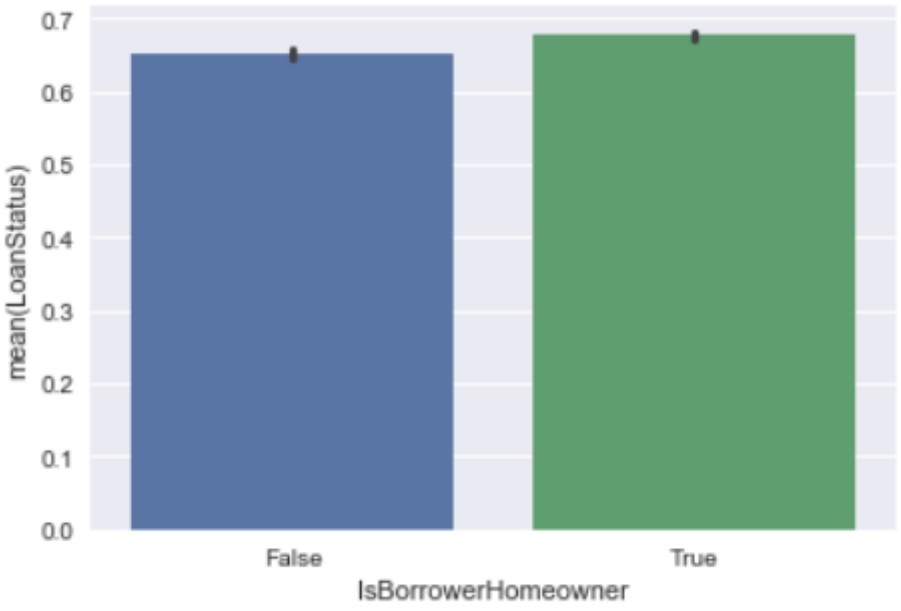
Fom the bar plot, we can see the non-defaulted values of loan statuses of different states of USA. Alabama, Missouri and San Diego states 'Completed' loan average strikes between 50% and 60%. The largest non-defaulted states are Nebraska and Washington DC followed by Arkansas, Wyoming and colorado ranging between 75% to 80% of their 'Completed loans'. This categorical data seems to give good insights.

Occupation



Setting a threshold of 80% and greater to be the top occupations that result in good return of the loans, some of these include 'Military Officer', 'Computer Programmer', 'Dentist', 'Attorney', 'Scientist', 'Pilot', 'Student-College Graduate Student' and 'Psychologist'. While occupations that have probability to complete the loans less than 50% include 'Nurse's Aide' and 'Realtor'.

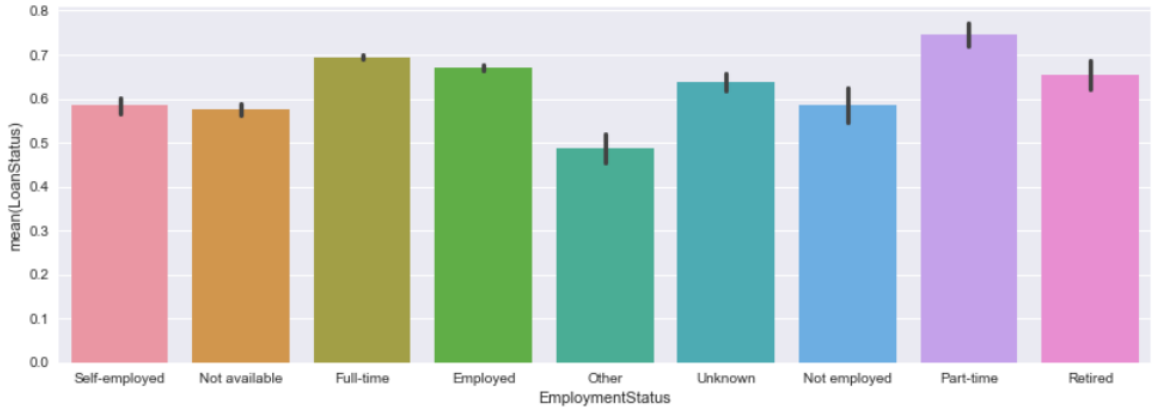
Is Borrower Home Owner



The correlation between IsBorrowerHomeowner and loan default is 0.026500826075958207, with a p-value of 2.1818058327590924e-10

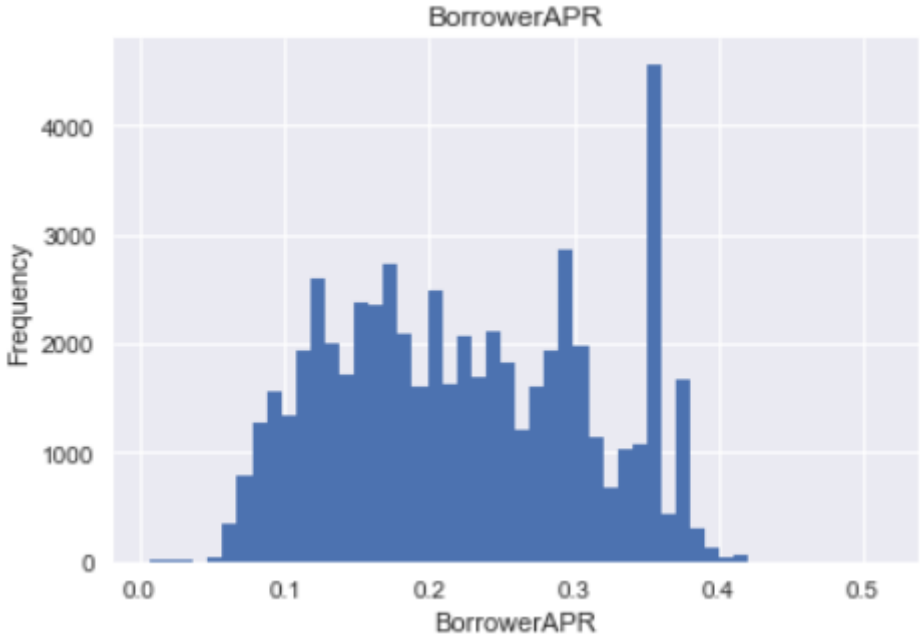
Looking at this plot, this variable has almost equal distribution in predicting loan status, however home owners tend to default less. Also the correlation is not so strong.So having information whether borrower is a home owner or not, dosen't really help in prediciting the status of loan payment.

Employment Status



EmploymentStatus has a relationship with default. We see a interesting observation here as part-time workers defaulted less often than the full-time workers and other employement status. It is great to see Retired people having successful loan payment percentage close to 70%.Also, unfortunately the people who listed their employment status as other defaulted even more often than those who were non employed.

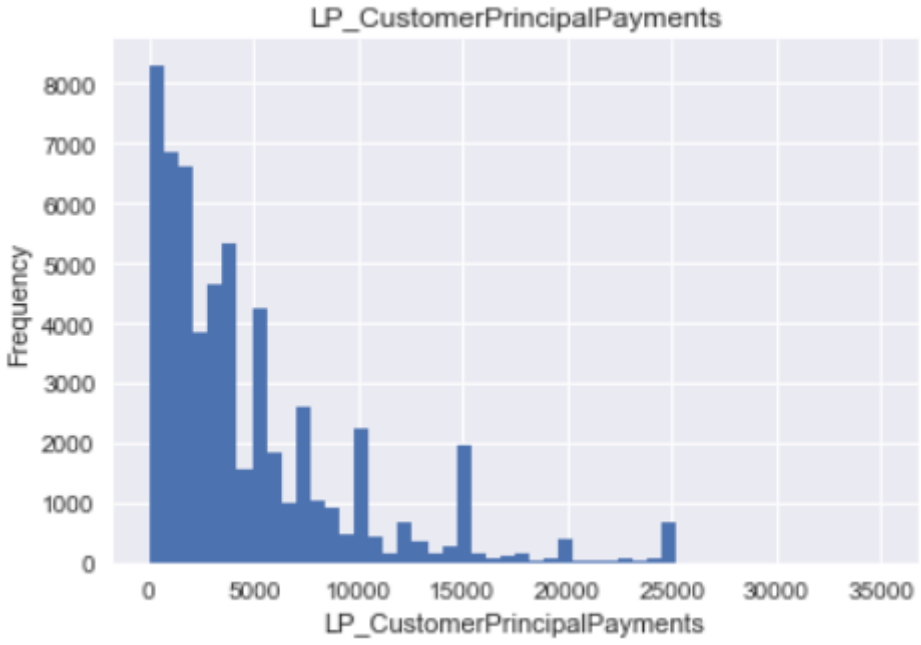
Feature Engineering: Numeric

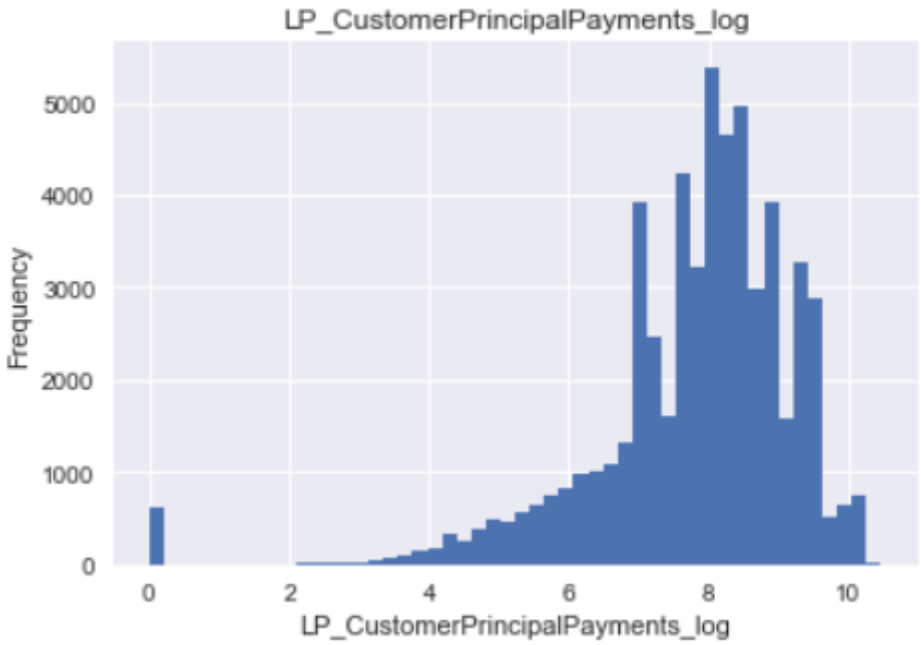


The distributions were obtained for all the numerical variables. Above is the distribution of one of the variable i.e. of BorrowerAPR.

Log transformations were applied on 3 variables listed below initially:

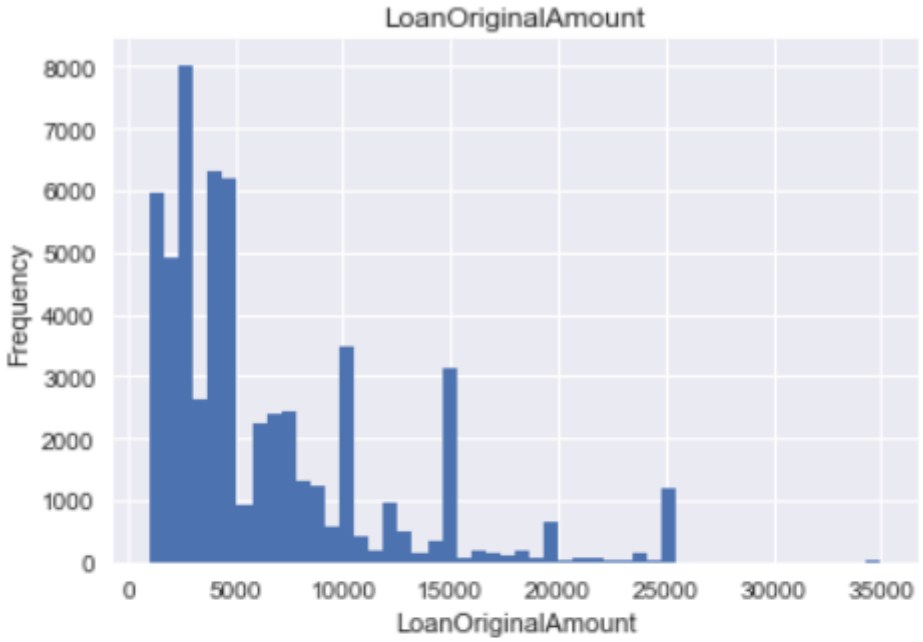
1. LP_CustomerPrincipalPayments
2. LP_CustomerPayments
3. LoanOriginalAmount

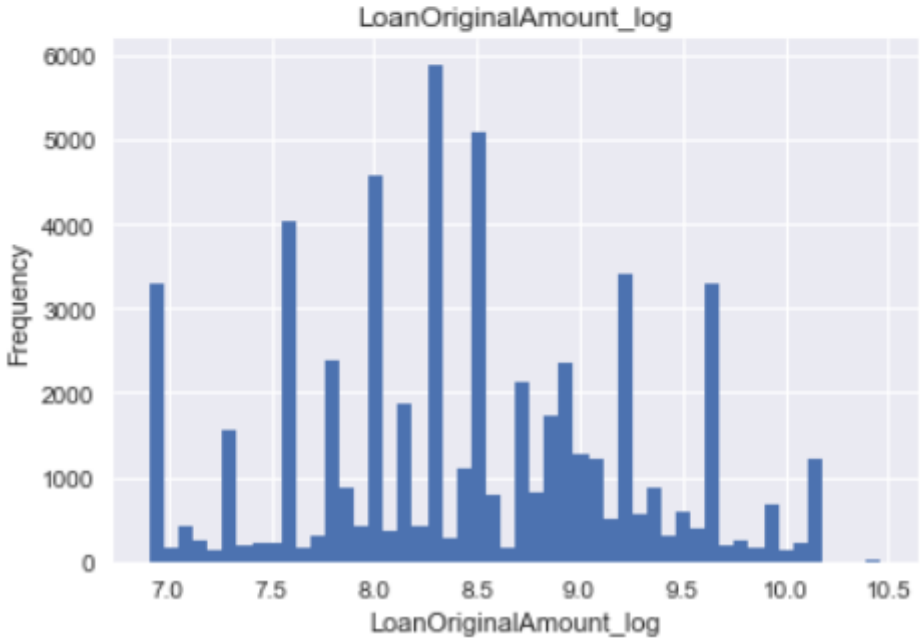




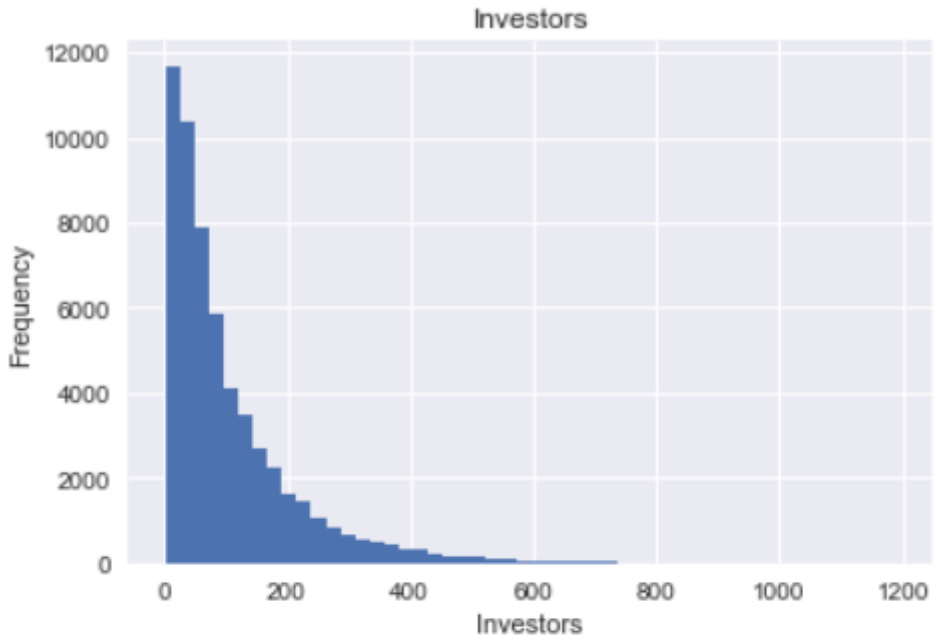

LP_CustomerPayments
Frequency
7000
6000
5000
4000
3000
2000
1000
0
0
5000
10000
15000
20000
25000
30000
35000
40000
LP_CustomerPayments

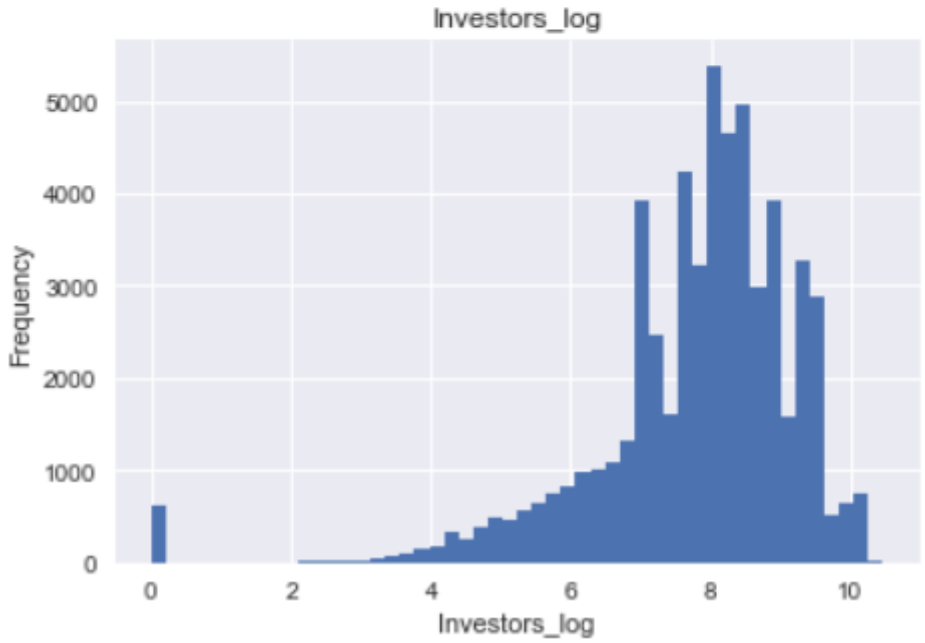

LP_CustomerPayments_log
Frequency
5000
4000
3000
2000
1000
0
0
2
4
6
8
10
LP_CustomerPayments_log


LoanOriginalAmount
Frequency
8000
7000
6000
5000
4000
3000
2000
1000
0
0
5000
10000
15000
20000
25000
30000
35000
LoanOriginalAmount

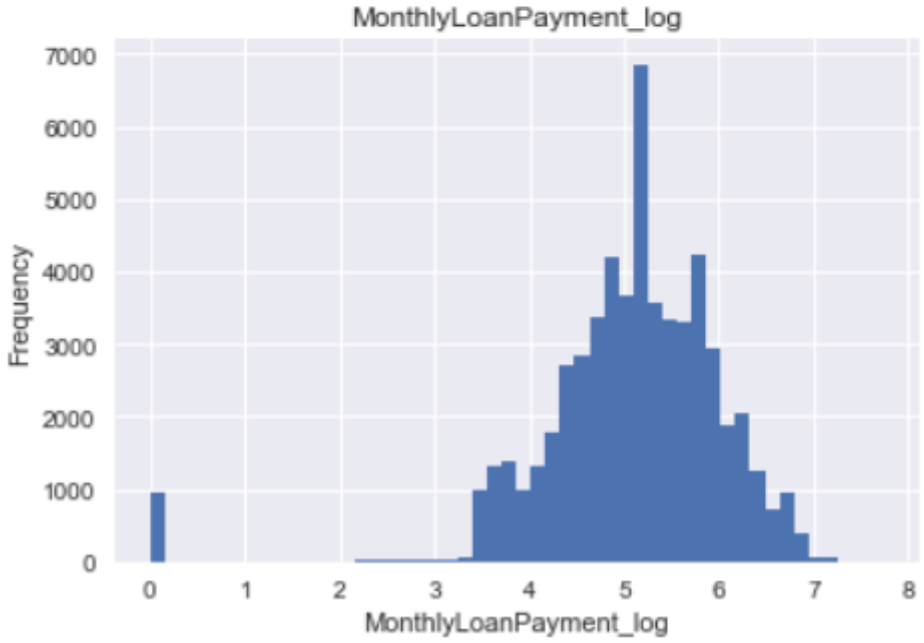

LoanOriginalAmount_log
Frequency
6000
5000
4000
3000
2000
1000
0
7.0
7.5
8.0
8.5
9.0
9.5
10.0
10.5
LoanOriginalAmount_log

Later, The Log transformation were applied on LP_CustomerPrincipalPayments, MonthlyLoanPayment shown below:



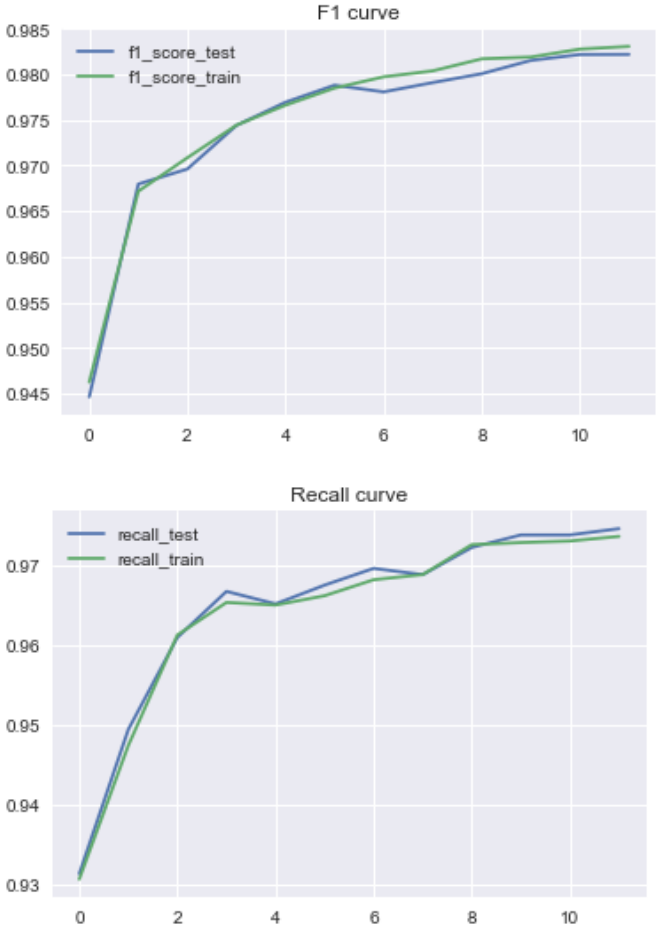



MonthlyLoanPayment
Frequency
10000
8000
6000
4000
2000
0
0
500
1000
1500
2000
MonthlyLoanPayment

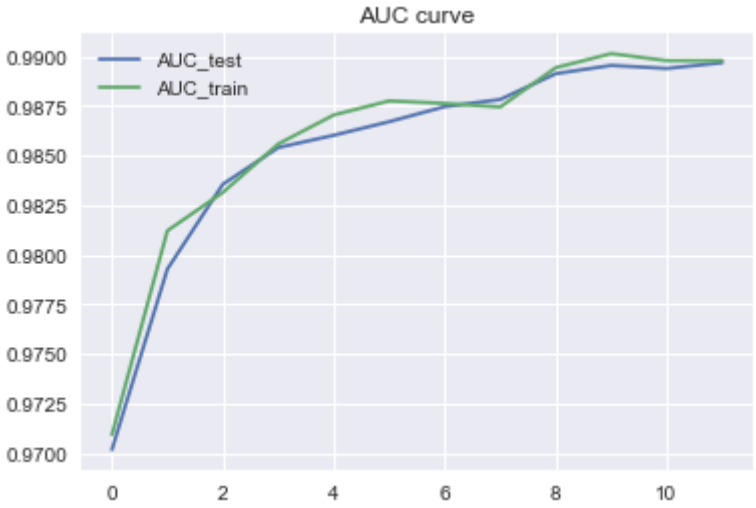

MonthlyLoanPayment_log
Frequency
7000
6000
5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
7
8
MonthlyLoanPayment_log

CLASSIFIERS

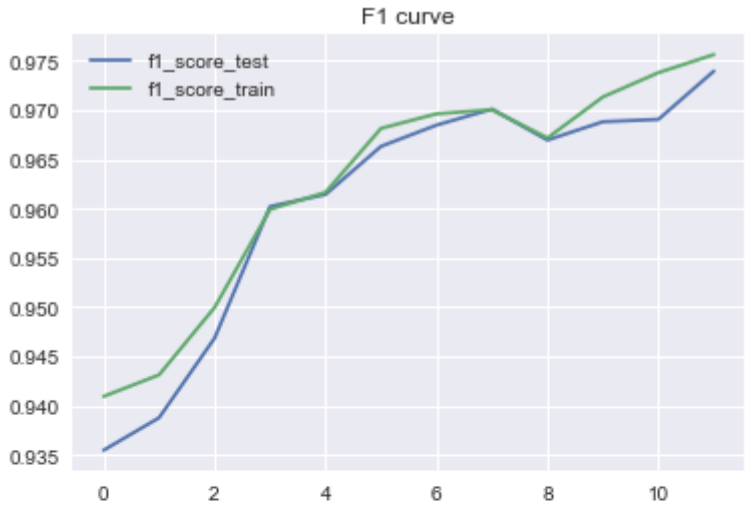
Model 1: SVM with L1(Binary/Linear)

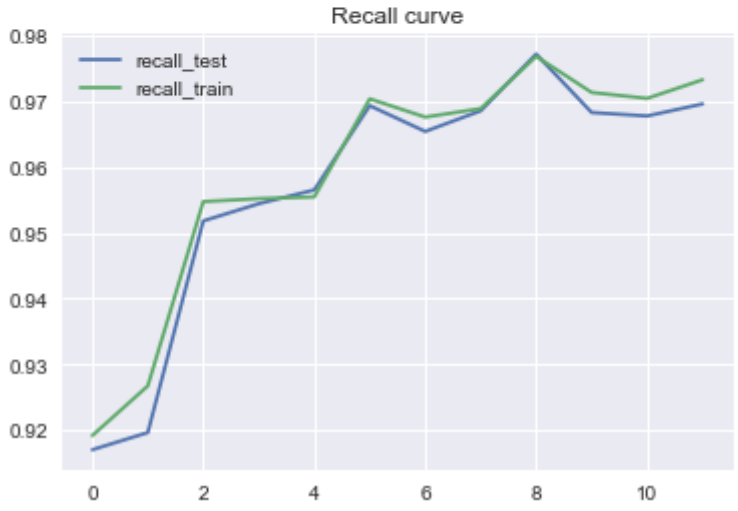


Model 2:Logistic Regression with L1(Binary/Linear)

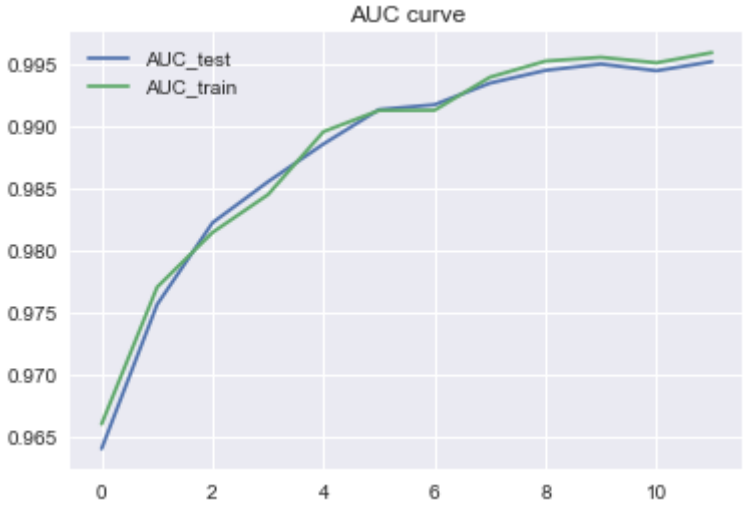


Model 3:SVM with L2(Binary/Linear)

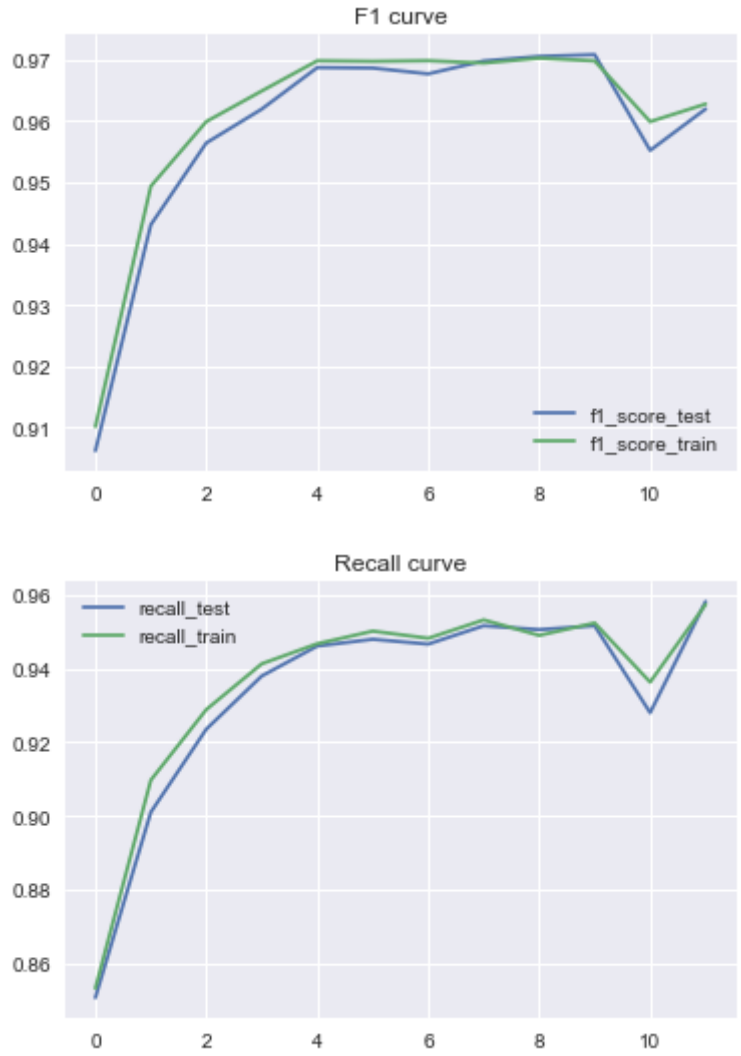




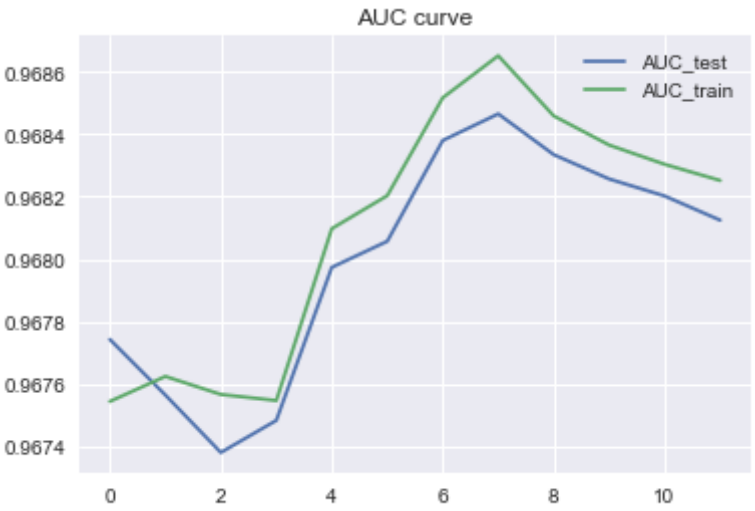
Model 4:Logistic Regression with L2(Binary/Linear)



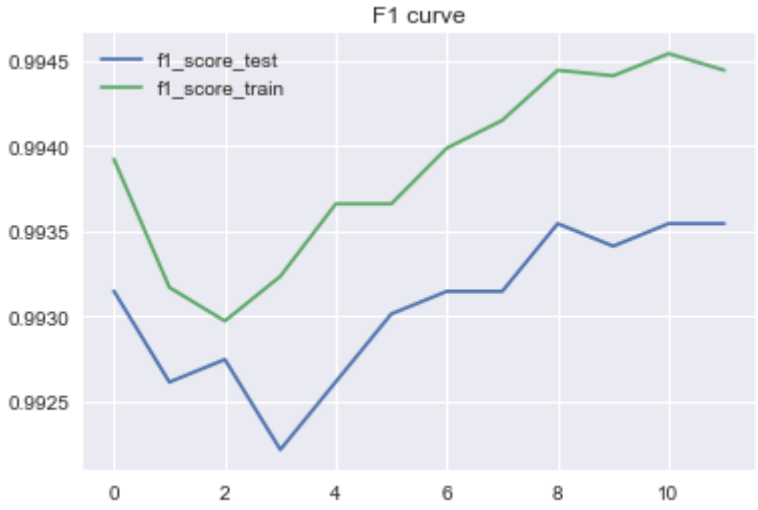
Model 5: Multi Layer Perceptron(Binary Classification/Non-Linear)



Model 6: BernoulliNB(Binary Classification/Non-Linear)

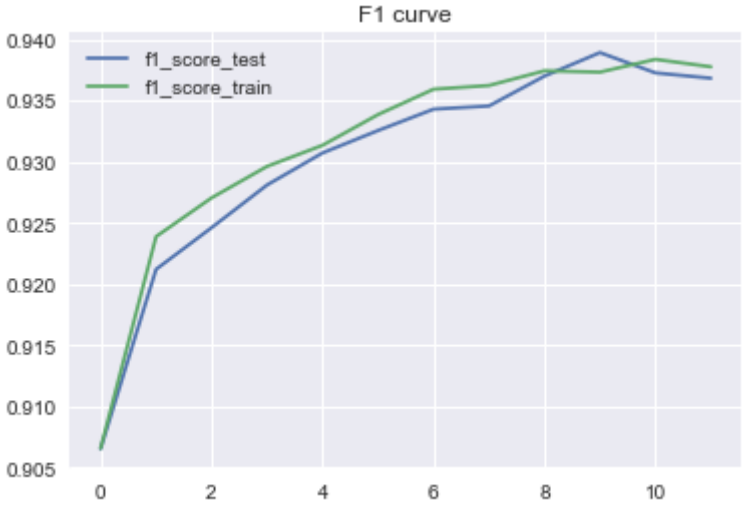


Model 7: Random Forest(Binary Classification/Non-Linear)

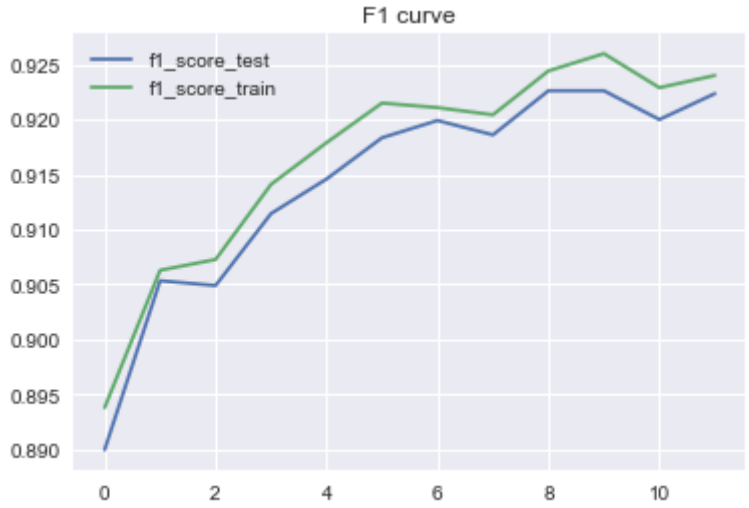


MULTI - CLASS CLASSIFICATION

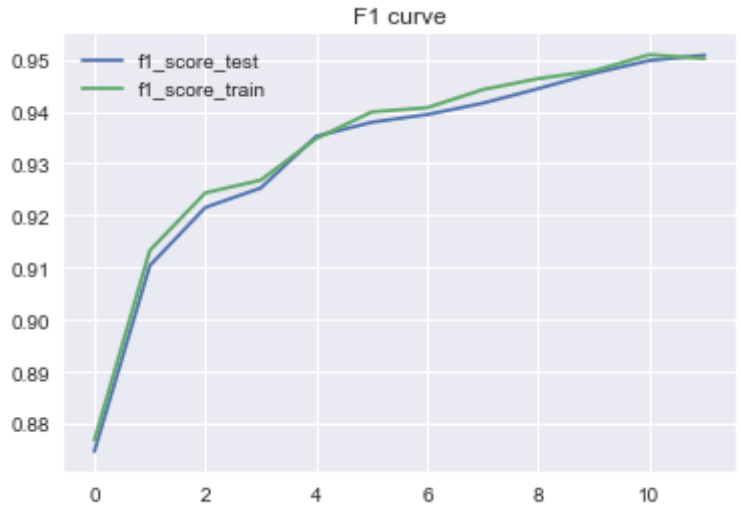
Model 8: SVM WITH L1(Multi Classification/Linear)



Model 9: SVM with L2(Multi Classification/Linear)



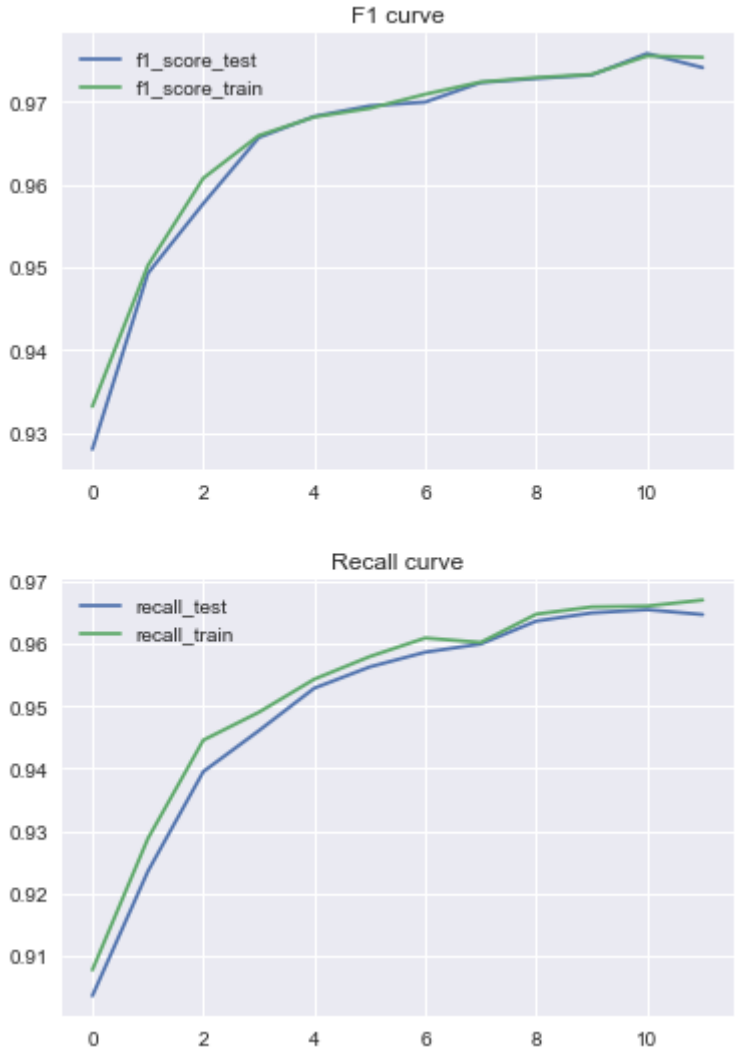
Model 10: Multi Layer Perceptron(Multi Classification/Non-Linear)



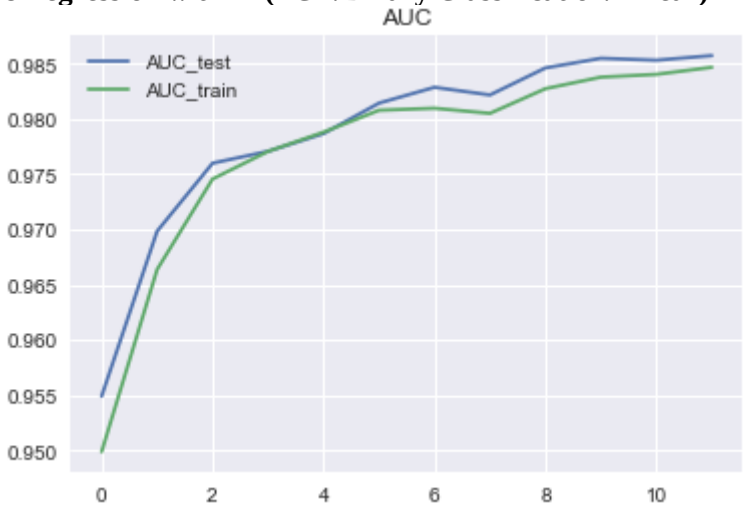
PRINCIPAL COMPONENT ANALYSIS(PCA)

BINARY CLASS CLASSIFICATION

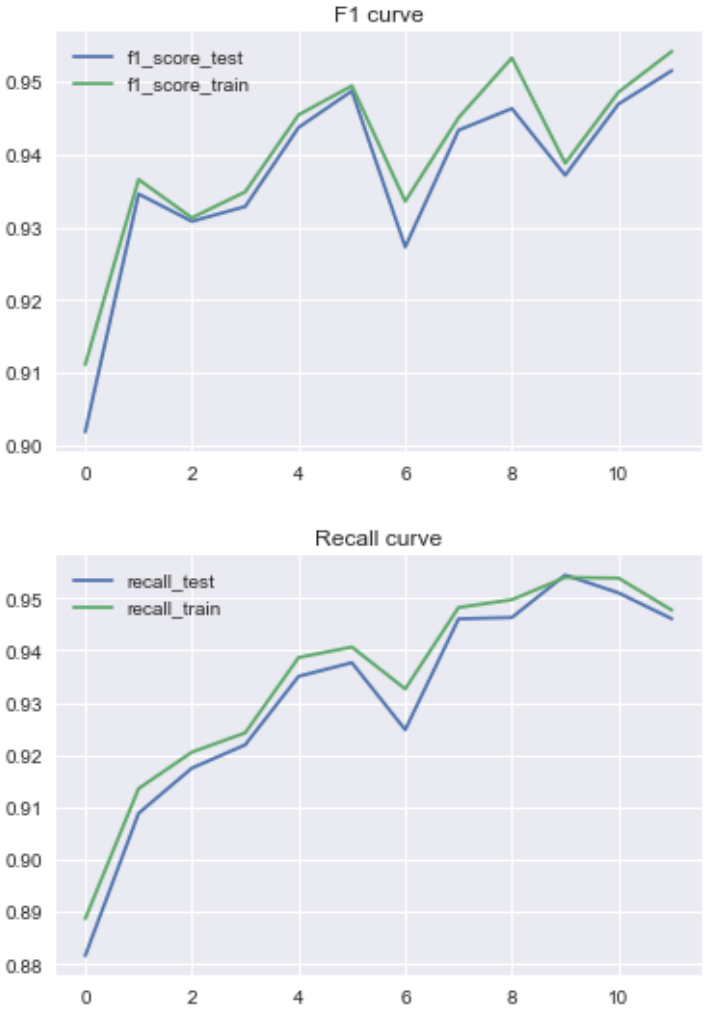
Model 11: SVM with L1(PCA/Binary Classification/Linear)



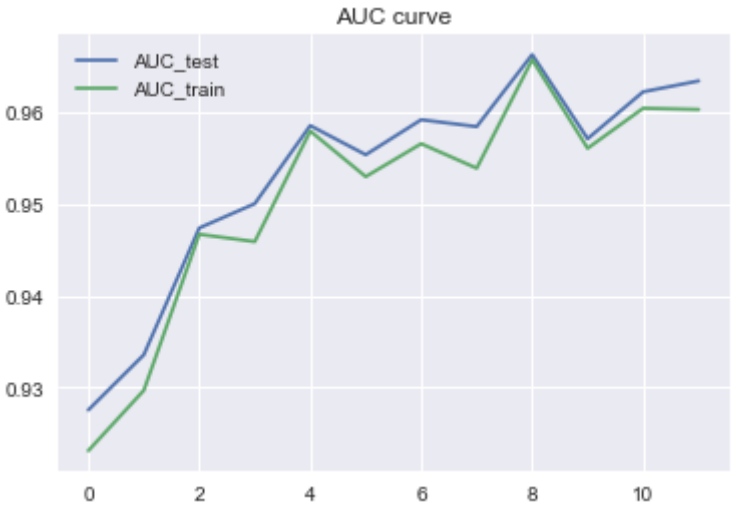
Model 12: Logistic Regression with L1(PCA/Binary Classification/Linear)



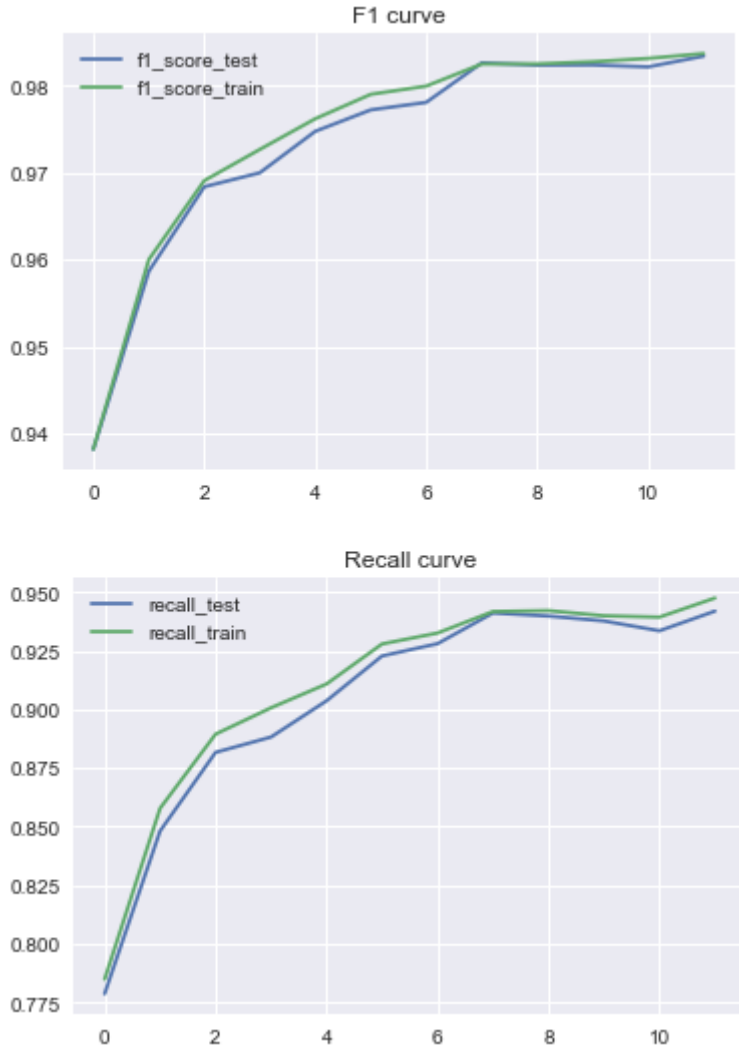
Model 13:SVM with L2(PCA/Binary Classification/Linear)



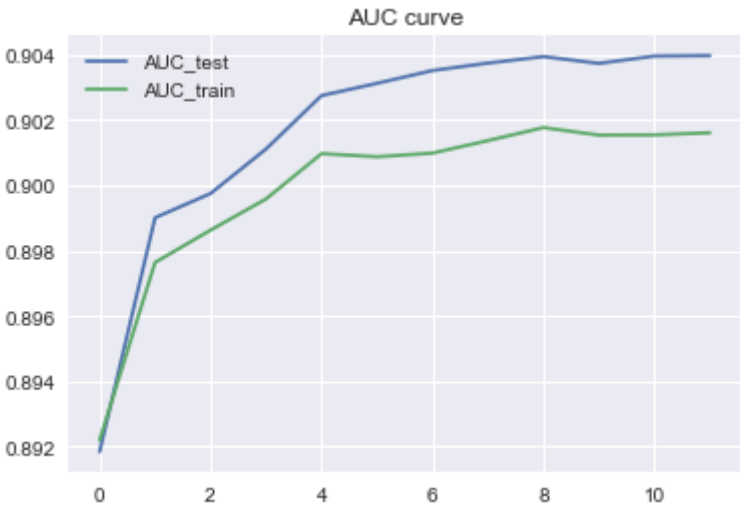
Model 14:Logistic Regression with L2 (PCA/Binary Classification/Linear)



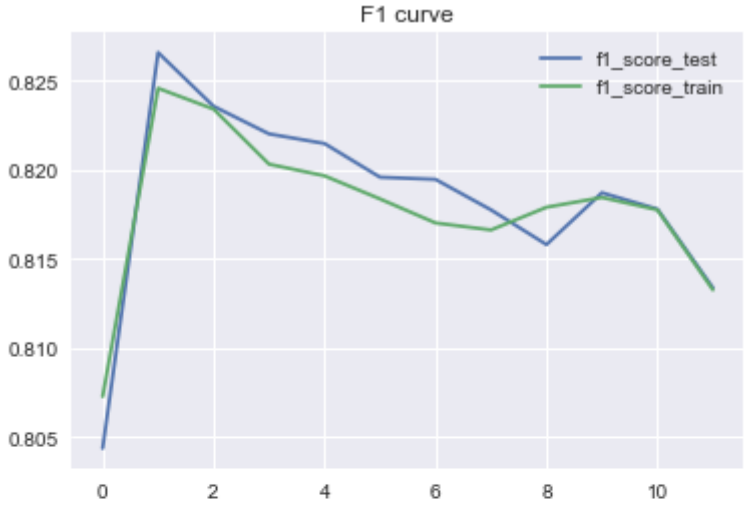
Model 15:Multi Layer Perceptron (PCA/Binary Classification/Non-Linear)

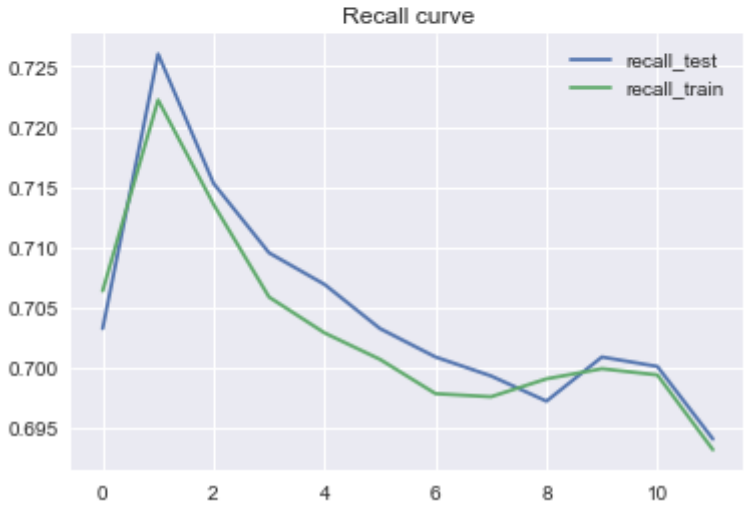


Model 16:BernoulliNB (PCA/Binary Classification/Non-Linear)



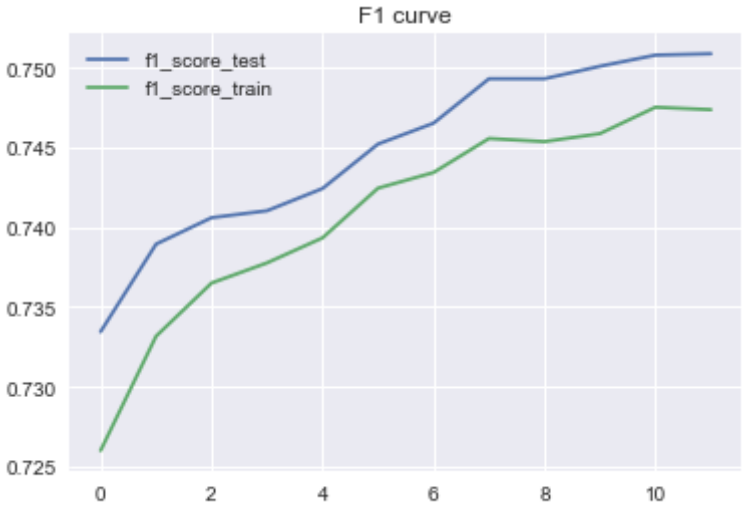
Model 17:Random Forest (PCA/Binary Classification/Non-Linear)



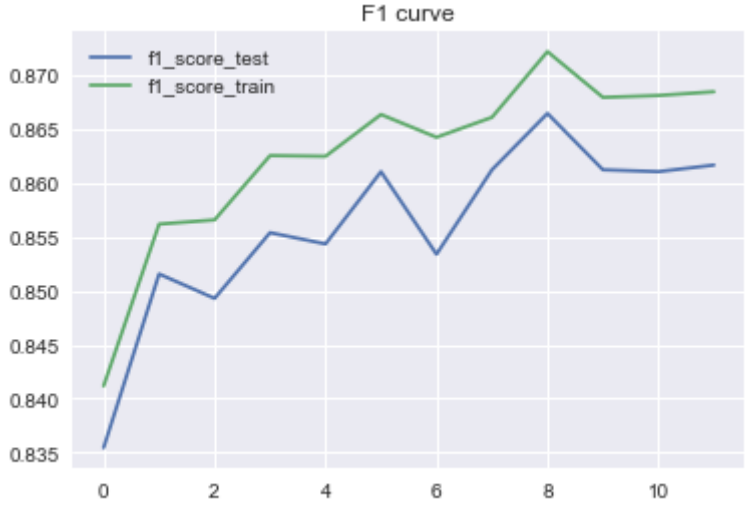


MULTI - CLASS CLASSIFICATION

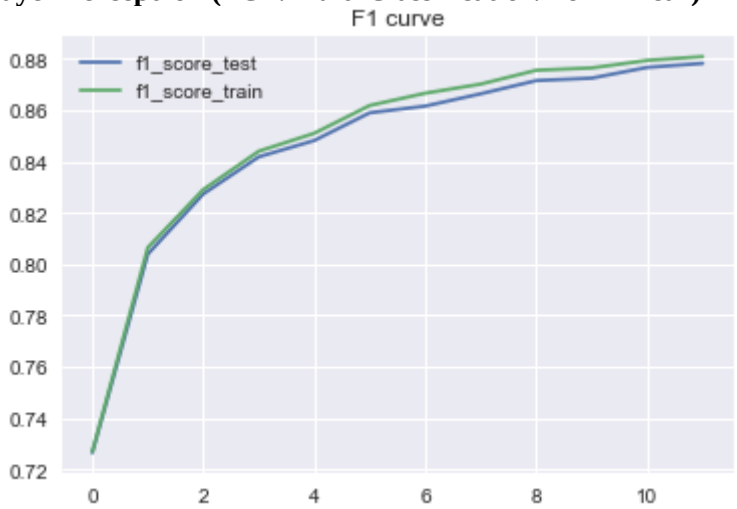
Model 18:SVM WITH L1 (PCA/Multi Classification/Linear)



Model 19:SVM with L2 (PCA/Multi Classification/Linear)



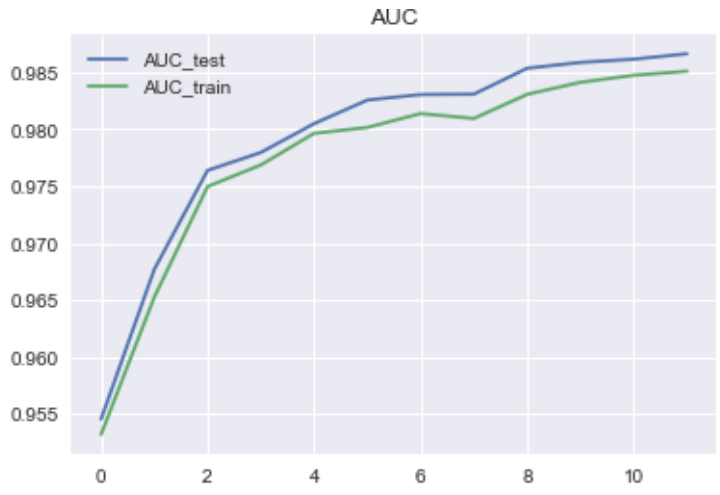
Model 20:Multi-Layer Perceptron (PCA/Multi Classification/Non-Linear)



RESULT AND CONCLUSION

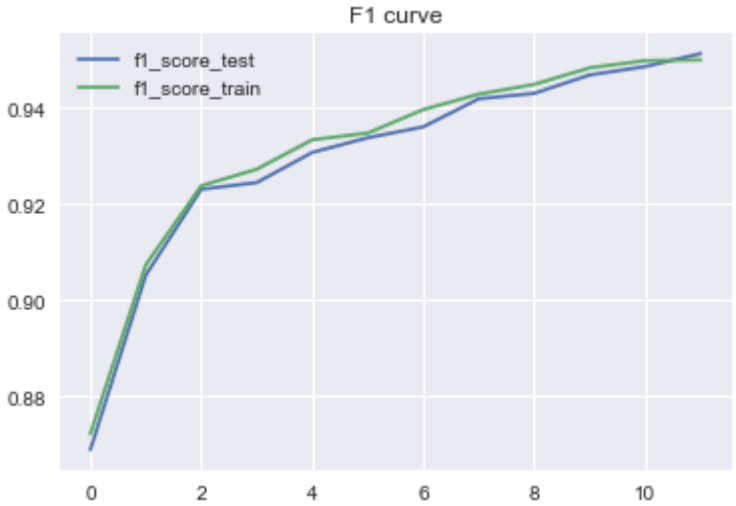
BEST MODEL FOR BINARY CLASSIFICATION

Model 12: Logistic Regression with L1(PCA/Binary Classification)



Logistic regression model with L1 is the best model performer to predict the classes for the binary type classification. The average Area under the curve score indicates that the model correctly predicts the classes with a probability of around 0.98. The curve is gradually increasing with increase in number of iterations, which suggests that the model is a good performer in predicting the classes accurately. The auc score of 0.98 is same prior and after applying pca which suggests that although the score being the same, pca aids in extracting only the required features which best explains the model thereby reducing the complexity of the model.

Model 10: Multi Layer Perceptron(Multi Classification)



For the multi-class classification, MLP Classifier has the highest f1 score among the classifiers. It is the weighted average of precision and recall score for each class in a multi class scenario. F1 score indicates to what extent the model correctly predicts all the classes (multi class case). Closer the score to 1, better is the model performance. Although SVM with L1 penalty has a similar score, comparing the graphs of both the classifiers over a number of iterations we can infer that the MLP has a comparatively smoother curve, which indicates its robustness towards partial fit.

It was noted that the f1 score for MLP Classifier reduced when the features where trained into the model after applying PCA. This might be for the reason that, since the target variable has 11 unbalanced classes, it probably requires initial number of features (prior application of pca) to have a good performance on the data.