LENDING CLUB DATA ANALYSIS-ML

Applying machine learning models to the data and deploying best model in Azure



**Vasanti Mahajan**

**Nehal Bhanushali**

**Pranjal Jain**

# 

# 

# 

# INTRODUCTION

The dataset consists of borrower’s details on the basis of which LendingClub analyses if the loan should be approved or not. And what should be the interest rate on which the loan should be given to the borrower.

# BUILDING CLASSIFICATION MODELS TO DETERMINE LOAN TO BE APPROVED OR NOT:

# ACCEPTED LOAN DATASET AND DECLINED LOAN DATASET

The declined loan dataset has 9 features, while accepted loan dataset 111 features. Hence we select the common features.

**Columns selected:**  Amount\_Requested, Loan\_Title, Risk\_Score, Debt\_to\_Income\_Ratio, Zip Code (first three digits available which indicate state), Employment\_length

**Columns dropped :** Application\_Date , State ,Policy Code as it has direct relation with loan eligibility.

**Risk Score :** Converted as per scale to fit in limit 300 to 850 range. (Vantage 3.0)

Loan Status : Does not meet the credit policy.

Status:Fully Paid and Does not meet the credit policy.

Status:Charged Off- Considered as Decline

Finally the both the accepted and declined files are merged with selected columns.

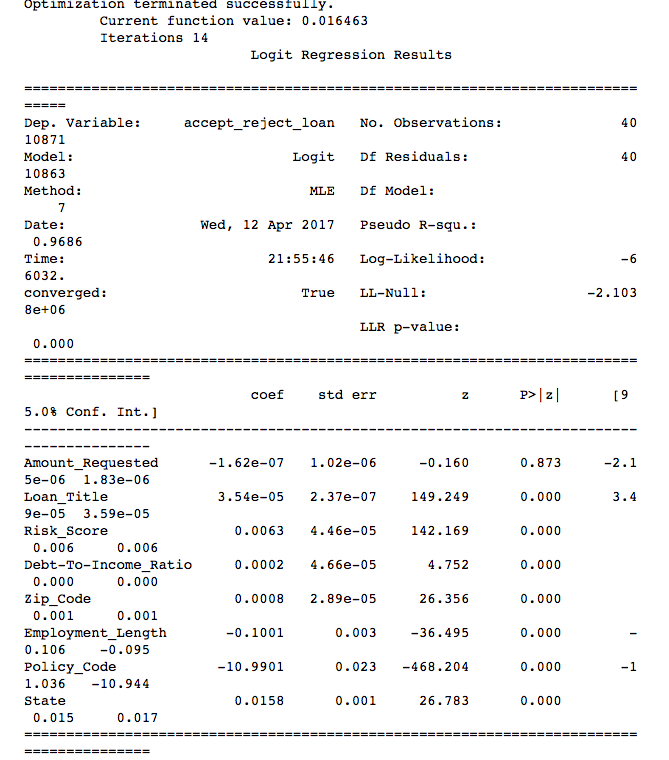
Number of loans declined in file 4679254

Number of loans accepted in file 1307121

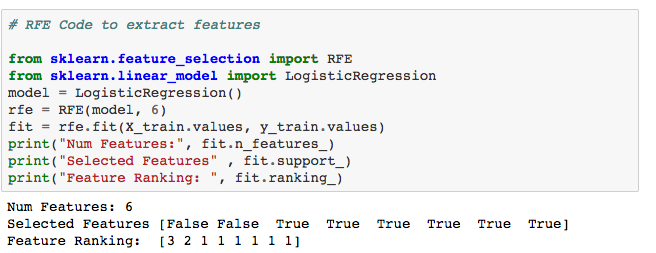
Ratio of accepted loans data to declined in file is 27.9%

# FEATURE SELECTION

**CALCULATING LOGIT VALUE FOR ALL FEATURES**



**RECURSIVE FEATURE ELIMINATION**



FEATURES SELECTED

'Amount\_Requested',

'Loan\_Title',

'Risk\_Score',

'Debt-To-Income\_Ratio',  
 ‘'Zip\_Code',

'Employment\_Length'

# CLASSIFICATION MODELS

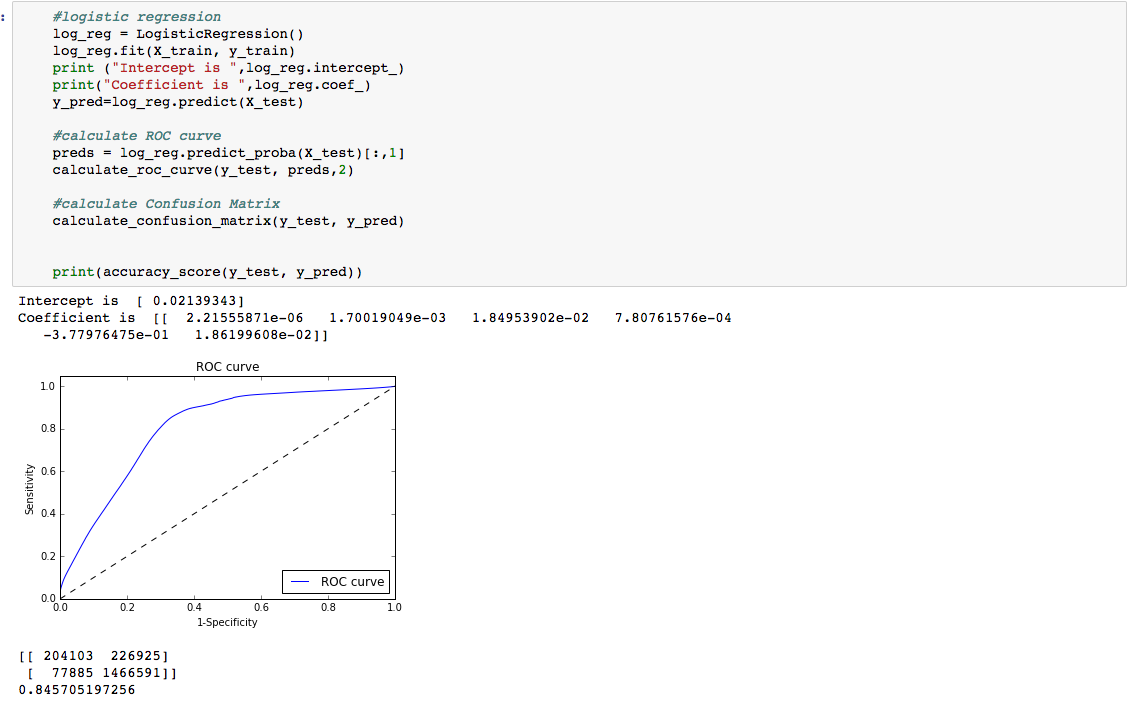
# LOGISTIC REGRESSION

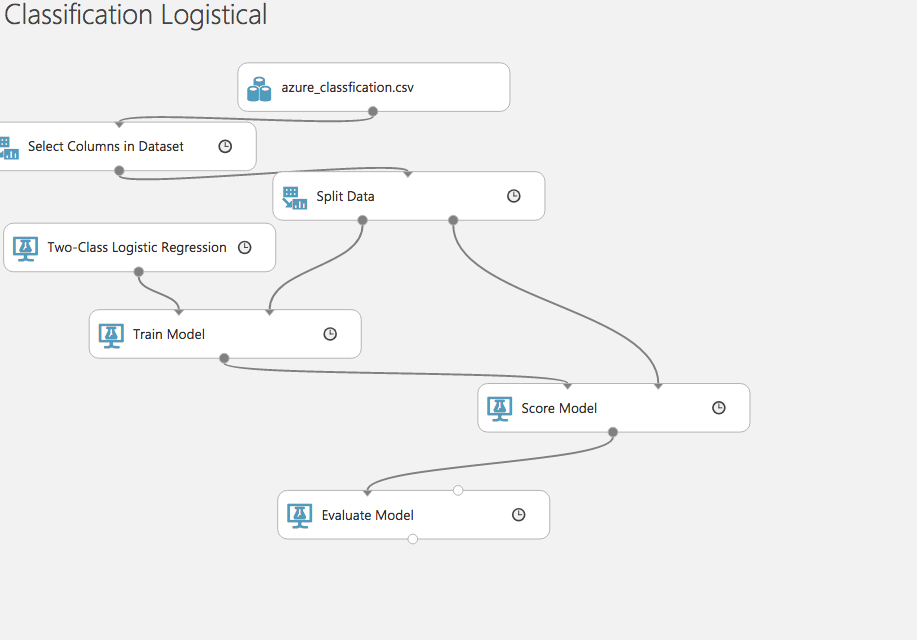
Confusion Matrix results

Number of loans classified in correctly as declined 1.6%

Number of loans classified in correctly as approved 17%

Accuracy 84%





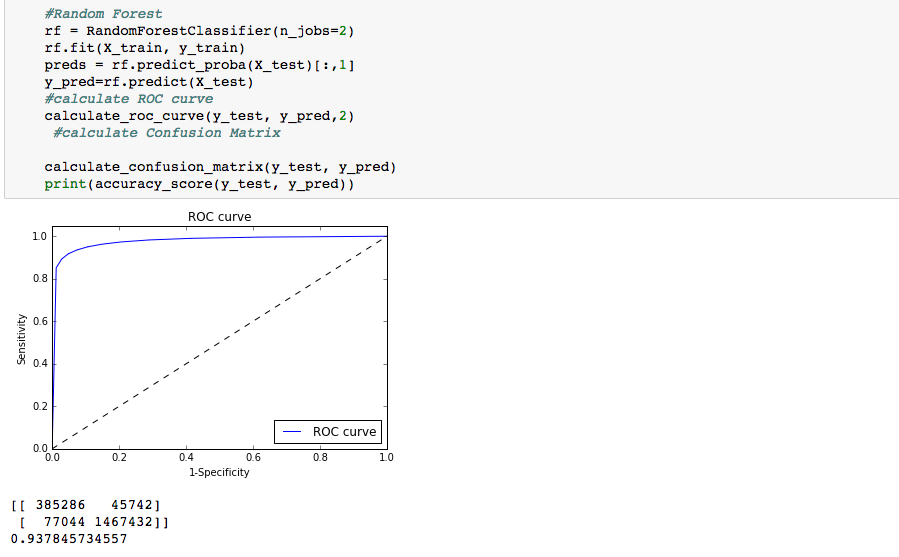
# RANDOM FOREST

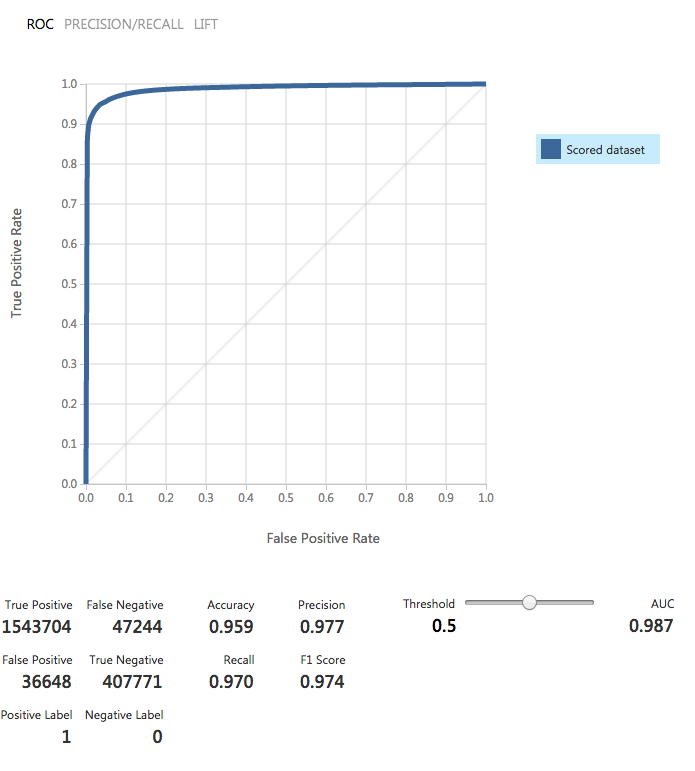
Confusion Matrix results

Number of loans classified in correctly as declined 1.64%

Number of loans classified in correctly as approved 0.9%

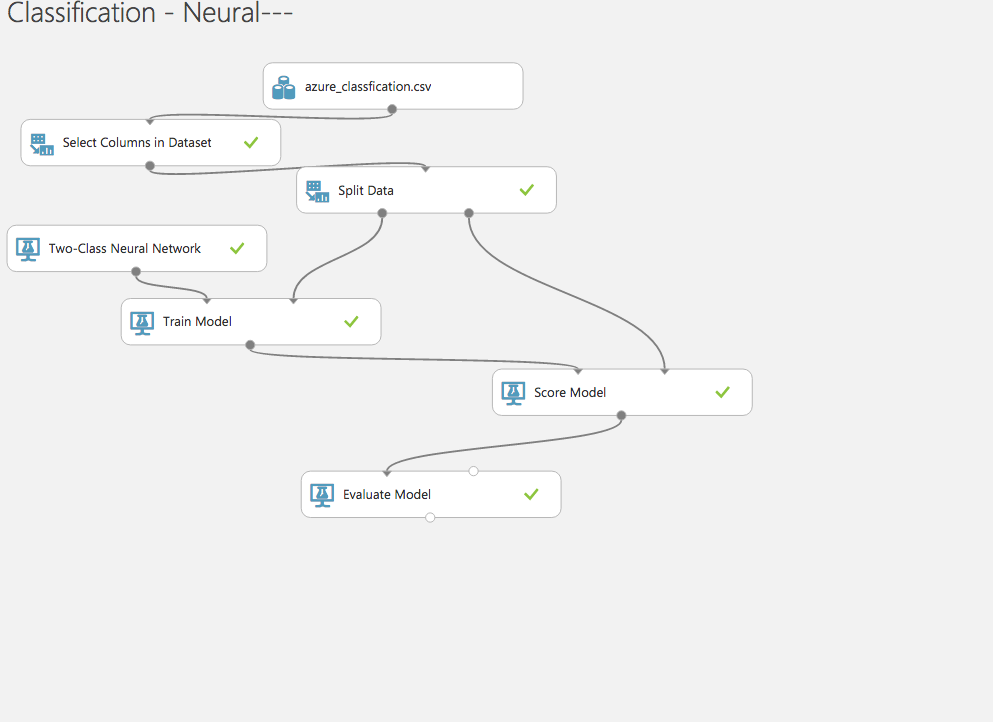
Accuracy 93%

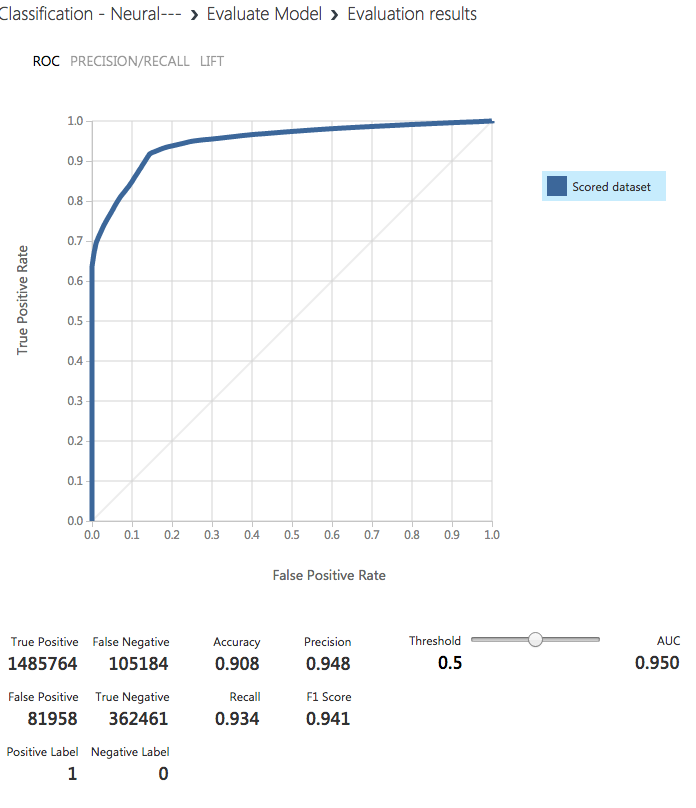




NEURAL NETWORK

Accuracy 90%





**Random Forest has less false positives for both accepted and declined.**

**Model Selected -Random Forest**

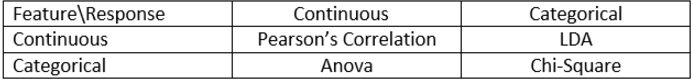
# ACCEPTED LOAN DATASET

The accepted loan dataset consists of 111 features. Feature engineering and selection is done to extract important features. Following four techniques are used for feature selection:

* Pearson correlation
* Sklearn’ s Select Percentile
* Randomized Lasso
* Recursive Feature Elimination

1. **FILTER METHODS:**





Since we are predicting a continuous label from a set of continuous features, we must compute Pearson Correlation between features.

The Pearson Correlation between int\_rate and all other features is shown below:



As we can see the features are not highly correlated, thus this is a poor technique for choosing the features.

Choosing the top most highly correlated features and running Linear Regression algorithm gives us the following scores:

**Parameters**:

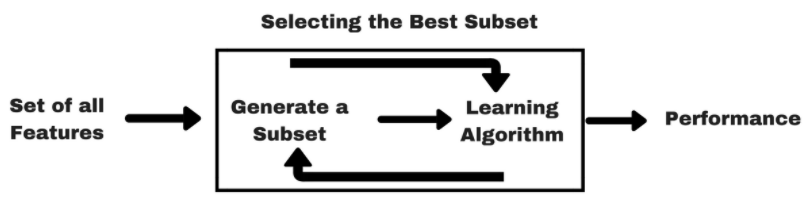
['acc\_open\_past\_24mths','num\_tl\_op\_past\_12m','percent\_bc\_gt\_75','term','revol\_util','total\_rec\_int','meanfico','int\_rate']

**Accuracy:**

Training score is **0.420957433573**  
Testing score is **0.418933610641**

The model did not train with great accuracy as the features were chosen on the basis of Pearson correlation.

2. **WRAPPER METHOD:**



2.1 **Select Percentile:**

Using Sklearn’s Select Percentile feature selection technique we can choose the top 20% percentile of features which contribute the maximum in predicting the

interest rate.

The top features obtained from this wrapper based technique and the accuracy of

the Linear Regression model is as follows:

**Features:**

'grade', 'total\_pymnt\_inv', 'revol\_util', 'loan\_status', 'fico\_range\_grade',

'total\_rec\_prncp', 'revol\_bal', 'grade\_based\_on\_inq\_last\_6mths',

'acc\_open\_past\_24mths', 'installment', 'last\_pymnt\_amnt', 'funded\_amnt\_inv',

'total\_acc', 'credit\_age', 'issue\_d', 'annual\_inc', 'meanfico','int\_rate']]

**Accuracy:**

Training score is **0.9169976353**  
 Testing score is **0.916618072069**

2.2 **RFE**

Using Sklearn’s RFE model, the following features were chosen and the Linear

Regression model was run for it:

**Features:**

'Grade','verification\_status','issue\_d','term','initial\_list\_status','grade\_based\_on\_inq

Last\_6mths','collections\_12\_mths\_ex\_med','fico\_range\_grade','chargeoff\_within\_12

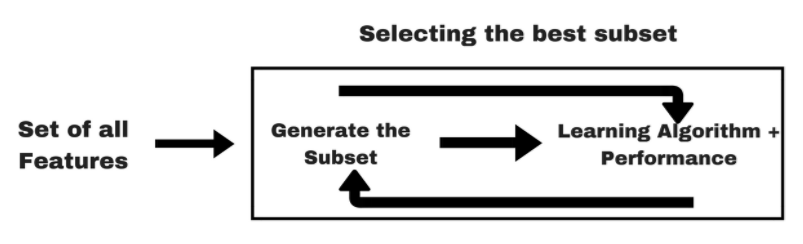
Mths','loan\_status','num\_tl\_op\_past\_12m','acc\_open\_past\_24mths','pub\_rec','pub\_re

C\_bankruptcies','purpose','meanfico','total\_acc','int\_rate'

**Accuracy:**

Training score is **0.916478380321**  
 Testing score is **0.916891270894**

3. **EMBEDDED METHOD**:



Implemented the L1 Regularization technique Lasso Regression for selecting the features and tested them on a Linear Regression model. The following is the accuracy of the model.

**Features:**

(1.0, 'verification\_status'), (1.0, 'total\_rec\_prncp'), (1.0, 'total\_rec\_int'), (1.0,

'total\_pymnt\_inv'), (1.0, 'total\_acc'), (1.0, 'term'), (1.0, 'revol\_util'), (1.0,

'revol\_bal'), (1.0, 'purpose'), (1.0, 'pub\_rec\_bankruptcies'), (1.0, 'pub\_rec'), (1.0,

'percent\_bc\_gt\_75')

Since all these features have a score of 1.0 thus this is not an accurate measure of

feature selection for our case study.

**Conclusion for Feature Selection:**

Comparing the results of the above methods, Select Percentile is the technique which gives the maximum accuracy and thus the final features for our model are:

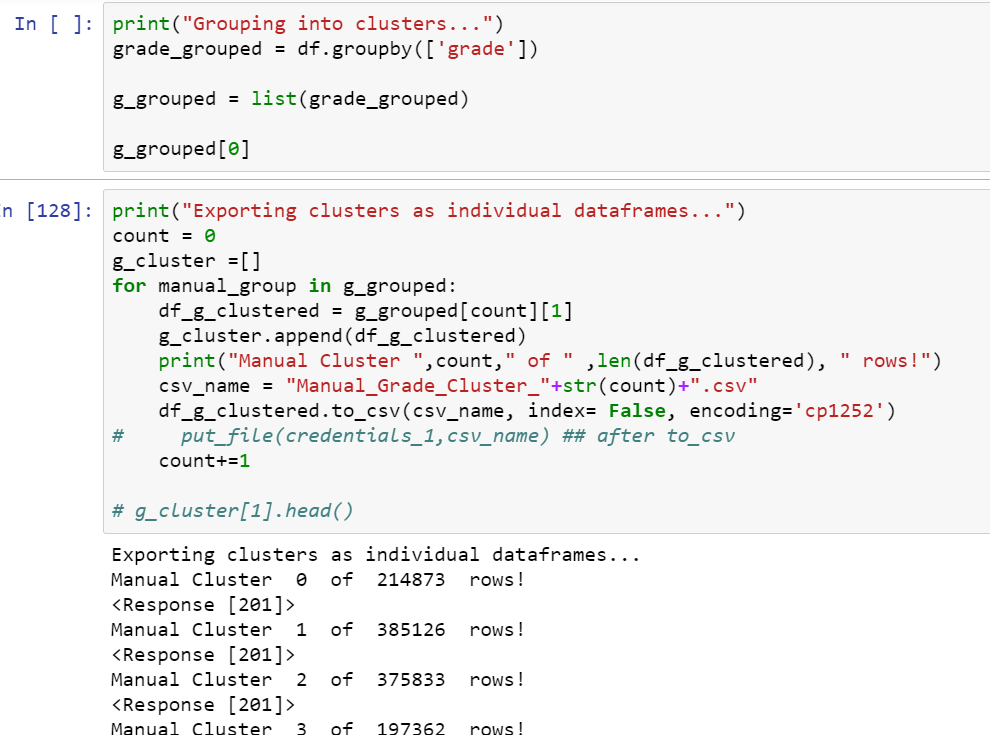
['grade','verification\_status','issue\_d','term','initial\_list\_status','grade\_based\_on\_inq\_last\_6mth','collections\_12\_mths\_ex\_med','fico\_range\_grade','chargeoff\_within\_12\_mths','loan\_status','num\_tl\_op\_past\_12m','acc\_open\_past\_24mths','pub\_rec','pub\_rec\_bankruptcies','purpose','meanfico','total\_acc','int\_rate']

# BUILDING REGRESSION MODELS FOR PREDICTING INTEREST RATE:

# There can either be a common prediction model for the entire data, or the data can be broken into clusters, for cluster specific prediction. The data is divided into three categories based on three different clustering techniques and four regression models are run on them.

CLUSTERING TECHNIQUES:

1. Segment data manually by a column or a combination of them. Clusters of grades have been made.

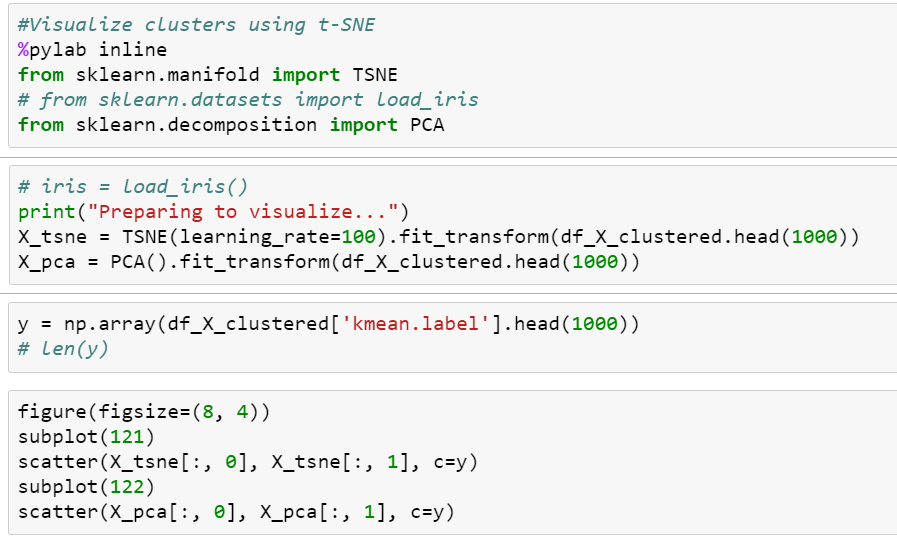


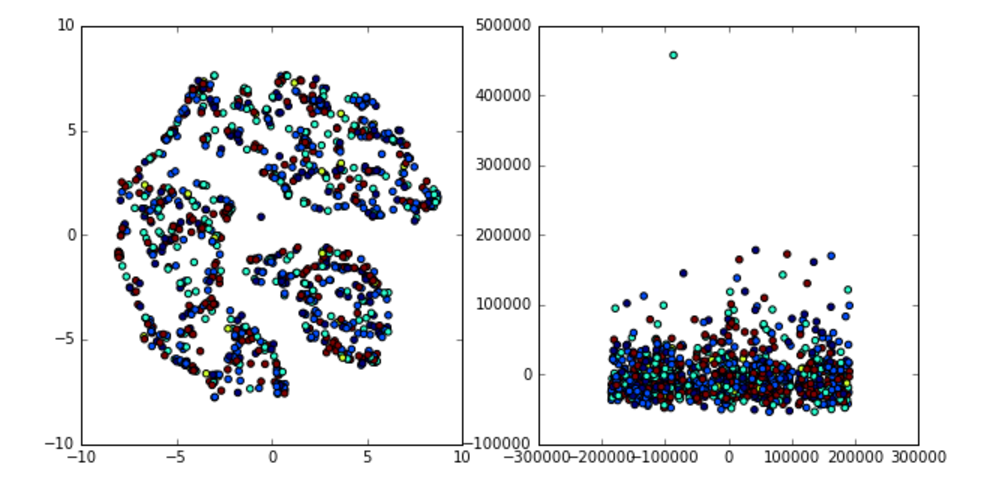
1. K-means clustering has been used to obtain 6 clusters as derived from the bend chart.



1. Use the same model on the entire dataset.

T-SNE VISUALIZATION v/s PCA:





REGRESSION MODELS:

**NORMALIZING DATA:**

The categorical data is transformed into numerical data using **Label Encoding**

Columns like **Fico Range** are grouped together to form buckets

**SPLITTING DATA:**

Use Random Split method for splitting data randomly on every iteration

**MODELS:**

On each of the clustered dataset the following four regression models are run. The following is the performance of each of them. The performance of all the models is evaluated on the basis of R2 score.

**CLUSTERING TECHNIQUE 1:**

The models are build on the entire dataset. The following is accuracy of the model.

LINEAR REGRESSION:

Training score is **0.916478380321**  
Testing score is  **0.916891270894**

NEURAL NETWORK: (MLP with 50 hidden layers, which optimizes the squared-loss using LBFGS)

Training score is  **0.952172955554**  
Testing score is **0.952300530304**

RANDOM FOREST: (n\_estimators= 8,n\_jobs=2, max\_depth=32, min\_samples\_leaf =1)

Accuracy of the model is **0.9884861333**

KNN: (k=11)

The most optimal accuracy is fetched using 11 nearest neighbours

Accuracy of the model is **0.346313043391**

|  |  |
| --- | --- |
| ALGORITHM | TEST ACCURACY |
| LINEAR REGRESSION | 0.916891270894 |
| NEURAL NETWORK | 0.952300530304 |
| **RANDOM FOREST** | **0.9884861333** |
| KNN | 0.346313043391 |

**CLUSTERING TECHNIQUE 2:**

The data has been split into clusters generated by the K-means algorithm. The four machine learning models are run on each of the clusters.

**LINEAR REGRESSION ON CLUSTER 1:**

**Accuracy:**

Training score is **0.920510496252**Testing score is **0.917243608695**

**LINEAR REGRESSION ON CLUSTER 2:**

**Accuracy:**

Training score is **0.917189266772**Testing score is **0.917465689397**

**LINEAR REGRESSION ON CLUSTER 3:**

**Accuracy:**

Training score is  **0.917078926071**Testing score is  **0.917293963913**

**LINEAR REGRESSION ON CLUSTER 4:**

**Accuracy:**

Training score is **0.916792230328**Testing score is **0.918187669818**

**LINEAR REGRESSION ON CLUSTER 5:**

**Accuracy:**

Training score is **0.955611624177**Testing score is **0.898463494407**

**LINEAR REGRESSION ON CLUSTER 6:**

**Accuracy:**

Training score is **0.91710948292**Testing score is **0.917061571887**

------------------------------------------------------------------------------------------------------------------------------------------------

**RANDOM FOREST ON CLUSTER 1:**

Accuracy of the model is **0.948861558115**

**RANDOM FOREST ON CLUSTER 2:**

Accuracy of the model is **0.976561777034**

**RANDOM FOREST ON CLUSTER 3:**

Accuracy of the model is **0.979234788392**

**RANDOM FOREST ON CLUSTER 4:**

Accuracy of the model is **0.977879647537**

**RANDOM FOREST ON CLUSTER 5:**

Accuracy of the model is  **0.869951505518**

**RANDOM FOREST ON CLUSTER 6:**

Accuracy of the model is **0.979173100573**

------------------------------------------------------------------------------------------------------------------------------------------------

**NEURAL NETWORK ON CLUSTER 1:**

Training score is **0.952292196395**  
Testing score is **0.948372053829**

**NEURAL NETWORK ON CLUSTER 2:**

Training score is **0.951950604013**  
Testing score is **0.951577445735**

**NEURAL NETWORK ON CLUSTER 3:**Training score is  **0.952644197997**Testing score is  **0.952930023223**

**NEURAL NETWORK ON CLUSTER 4:**  
Training score is **0.951856242041**  
Testing score is **0.952649989928**

**NEURAL NETWORK ON CLUSTER 5:**  
Training score is **0.999999947932**  
Testing score is **0.431639043406**

**NEURAL NETWORK ON CLUSTER 6:**  
Training score is **0.95241537991**  
Testing score is **0.95201616905**

------------------------------------------------------------------------------------------------------------------------------------------------

KNN ON CLUSTER 1:

Accuracy of the model is **0.402789185517**

KNN ON CLUSTER 2:

Accuracy of the model is **0.372754375617**

KNN ON CLUSTER 3:

Accuracy of the model is  **0.369376856041**

KNN ON CLUSTER 4:

Accuracy of the model is **0.369570681011**

KNN ON CLUSTER 5:

Accuracy of the model is  **-0.236425556647**

KNN ON CLUSTER 6:

Accuracy of the model is **0.366691042046**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ALGO | CLUSTER 1 | CLUSTER 2 | CLUSTER 3 | CLUSTER 4 | CLUSTER 5 | CLUSTER 6 |
| Linear Regression | 0.917243 | 0.917465 | 0.91729 | 0.918187 | 0.898463 | 0.917061 |
| Neural Network | 0.948372 | 0.951577 | 0.952930 | 0.952649 | 0.431639 | 0.952016 |
| Random Forest | 0.948861 | 0.976561 | 0.979234 | 0.977879 | 0.869951 | 0.979173 |
| KNN | 0.402789 | 0.372754 | 0.3693768 | 0.369570 | -0.2364255 | 0.366691 |

The highlighted algorithms are most optimal for each clusters.

**CLUSTERING TECHNIQUE 3:**

Manually building clusters based on grade. The accuracy of the models is:

**LINEAR REGRESSION ON CLUSTER 1:**

**Accuracy:**

Training score is **0.455652175482**Testing score is **0.456276853937**

**LINEAR REGRESSION ON CLUSTER 2:**

**Accuracy:**

Training score is **0.365314322198**Testing score is **0.36685141610**4

**LINEAR REGRESSION ON CLUSTER 3:**

**Accuracy:**

Training score is  **0.207024249752**

Testing score is  **0.212489386991**

**LINEAR REGRESSION ON CLUSTER 4:**

**Accuracy:**

Training score is **0.143422445692**

Testing score is **0.15064304605**

**LINEAR REGRESSION ON CLUSTER 5:**

**Accuracy:**

Training score is **0.273526433103**

Testing score is **0.273122977346**

**LINEAR REGRESSION ON CLUSTER 6:**

**Accuracy:**

Training score is **0.436481186121**

Testing score is **0.429011513207**

**LINEAR REGRESSION ON CLUSTER 7:**

**Accuracy:**

Training score is **0.81233138262**

Testing score is **0.831763844579**

------------------------------------------------------------------------------------------------------------------------------------------------

**RANDOM FOREST ON CLUSTER 1:**

Accuracy of the model is **0.814759971733**

**RANDOM FOREST ON CLUSTER 2:**

Accuracy of the model is **0.860161577495**

**RANDOM FOREST ON CLUSTER 3:**

Accuracy of the model is  **0.840512304021**

**RANDOM FOREST ON CLUSTER 4:**

Accuracy of the model is  **0.804460435558**

**RANDOM FOREST ON CLUSTER 5:**

Accuracy of the model is **0.840177974495**

**RANDOM FOREST ON CLUSTER 6:**

Accuracy of the model is **0.788567714919**

**RANDOM FOREST ON CLUSTER 7:**

Accuracy of the model is **0.853061120919**

------------------------------------------------------------------------------------------------------------------------------------------------

**NEURAL NETWORK ON CLUSTER 1:**

Training score is **0.745270388713**Testing score is **0.742271545532**

**NEURAL NETWORK ON CLUSTER 2:**

Training score is **0.704158689578**

Testing score is **0.70221552349**

**NEURAL NETWORK ON CLUSTER 3:**Training score is **0.593037872304**Testing score is  **0.594356349076**

**NEURAL NETWORK ON CLUSTER 4:**  
Training score is **0.590864093176**

Testing score is **0.585333874473**

**NEURAL NETWORK ON CLUSTER 5:**  
Training score is **0.76067320626**

Testing score is **0.756453002717**

**NEURAL NETWORK ON CLUSTER 6:**  
Training score is **0.765961706421**

Testing score is **0.73061791973**

**NEURAL NETWORK ON CLUSTER 7:**  
Training score is **0.889093988975**

Testing score is **0.841825362342**

------------------------------------------------------------------------------------------------------------------------------------------------

KNN ON CLUSTER 1:

Accuracy of the model is **0.634800543012**

KNN ON CLUSTER 2:

Accuracy of the model is **0.586986940027**

KNN ON CLUSTER 3:

Accuracy of the model is **0.431687984949**

KNN ON CLUSTER 4:

Accuracy of the model is **0.328095588573**

KNN ON CLUSTER 5:

Accuracy of the model is **0.475678694915**

KNN ON CLUSTER 6:

Accuracy of the model is **0.412201837081**

KNN ON CLUSTER 7:

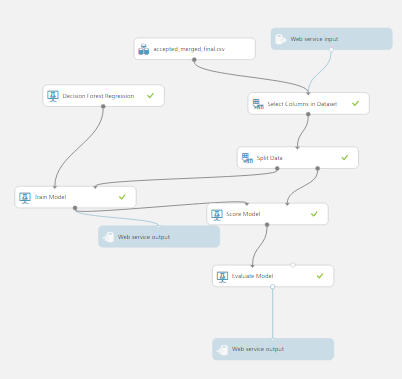
Accuracy of the model is  **0.542094748441**

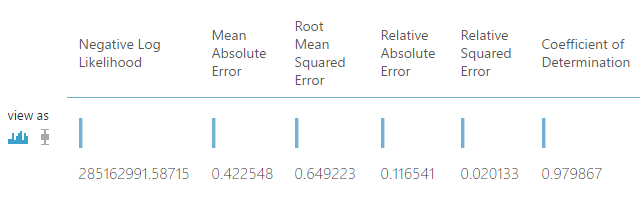
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ALGO | CLUSTER 1 | CLUSTER 2 | CLUSTER 3 | CLUSTER 4 | CLUSTER 5 | CLUSTER 6 | CLUSTER 7 |
| Linear Regression | 0.45627 | 0.36685 | 0.212489 | 0.15064 | 0.27312 | 0.42901 | 0.83176 |
| Neural Network | 0.745270 | 0.704158 | 0.594356 | 0.59086 | 0.76067 | 0.76596 | 0.88909 |
| Random Forest | 0.814759 | 0.86016 | 0.84051 | 0.84051 | 0.84017 | 0.78856 | 0.85306 |
| KNN | 0.634800 | 0.586986 | 0.431687 | 0.32809 | 0.47567 | 0.412201 | 0.54209 |

# AZURE ML IMPLEMENTATION:

1. **CLUSTERING TECHNIQUE 1:**

Applied Random Forest on the whole dataset in Azure ML

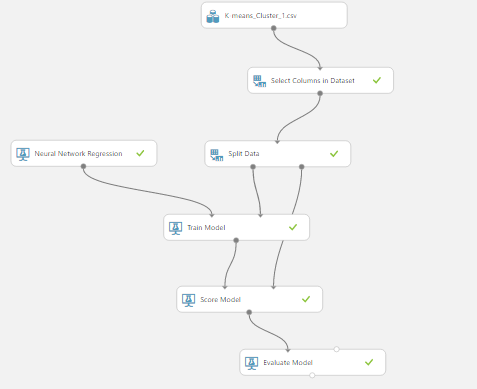


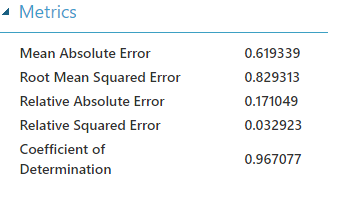


1. **CLUSTERING TECHNIQUE 2:**

Using K-means clustering algorithm for computing the interest rate using Neural

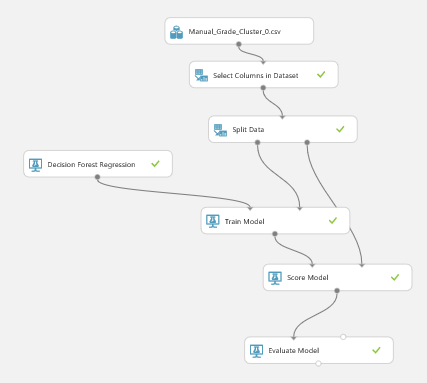
Network.



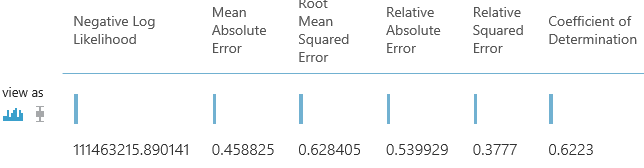


1. **CLUSTERING TECHNIQUE 3:**

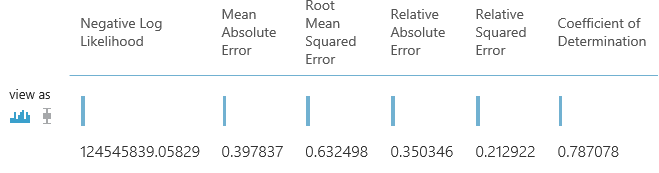
Using manual clustering technique for dividing data based on grade.



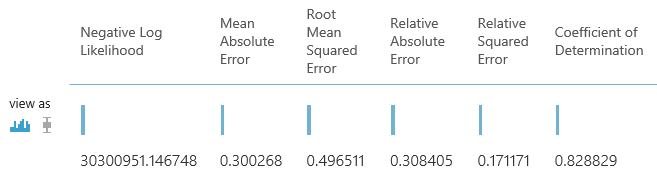
**Model 1-Grade A performance:**



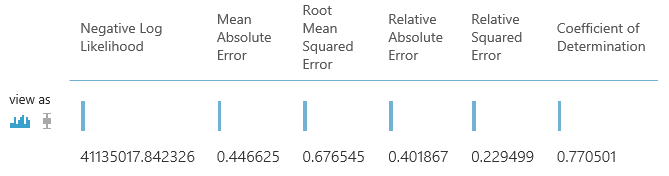
**Model 2-Grade B performance:**



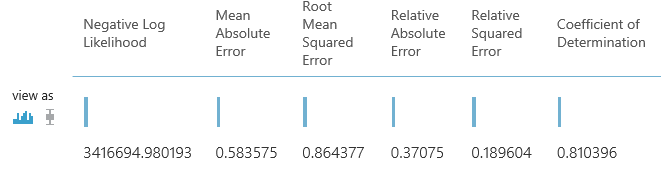
**Model 3-Grade C performance:**



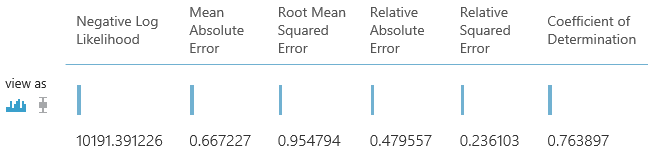
**Model 4- Grade D performance:**



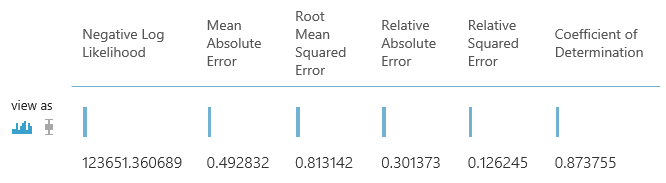
**Model 5-Grade E performance:**



**Model 6-Grade F performance:**



**Model 7-Grade G performance:**



# RESULTS

The pipeline works as follows:

1. The borrower feeds his/her data into the system which calls the Random Forest Classification trained model and returns a result 0/1 (loan should be given/ loan should not be given)
2. If the loan should be given to the user, additional information is fetched from user and three API calls are made to the trained models (trained on the three clustering techniques). The Random Forest algorithm is most optimal for clustering technique 1 and 2 and Neural Network is most optimal for clustering technique 3
3. The three models return three interest rates and the one with the highest interest rate is suggested for the user.