**TEAM 8**

**ADVANCED DATA SCIENCE**

**LENDING CLUB - ANALYSIS**

**TOTAL FLOW :**

TAKE REQUIRED VARIABLES TO CONFIRM LOAN GIVEN OR NOT

TAKE ALL THE INPUTS FROM USER ON R SHINY WEB

PREDICT INTEREST RATE WITHOUT CLUSTER

PREDICT INTEREST RATE USING MANUAL CLUSTER

PREDICT INTEREST RATE USING CLUSTER

**Classification Models**:

Flow of the steps Performed:

DATA MODELING FOR DECLINED LOAN DATA(CHECK FICO SCORE)

CHOSE COLUMNS AND CREATE DERIVED COLUMNS

RUN MODELS ON PYTHON NOTEBOOK

MERGE DERIVED AND LOAN FILES

DEPLOY THE REST API TO R WEBAPP

CREATE BEST MODEL ON AZURE

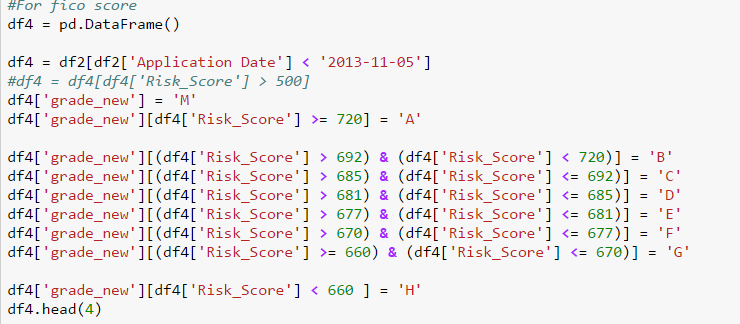
CHOSE BEST MODEL

**Data Modeling before running models:**

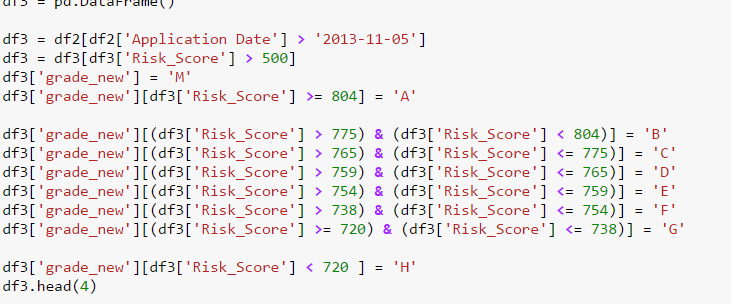
1. As Risk Score in declined loan data set is FICO score for date ranging before November 2013 and Vantage score for date later, we handled both the scenario differenly
2. Removed all those rows where risk score is zero (As zero risk score will incease the deviation of model towards declined data set)
3. Delete the rows where risk score is less than 500 ( For Vantage Score) as both Vantage 2.0 and 3.0 were there for both of them (Assumption)
4. Created a new derived column, Grade new wherein we mapped the risk/fico score according to a grade value chosen by us
5. For declined loan data this was chosen seperately for FICO and Vantage score
6. Vantage Score grade scale was scaled according to FICO score range.

**Derived Column : grade\_new (custom created grade values based on fico/vantage score)**

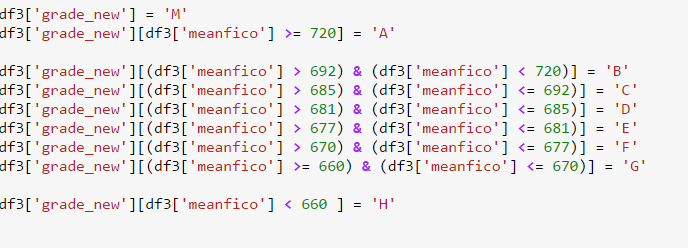
For less than 11th November (FICO Score)



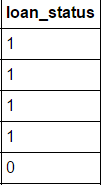
For more than 11th November 2013 (Vantage Score) (Grades scaled according to Fico score Grades)



Same steps were performed for Loan data set to create custom Grade values:

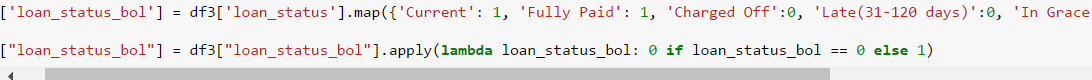


Created column for Loan Declined and Not Declined for classification

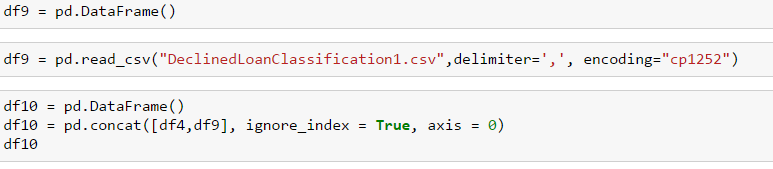


**Note : To make the classification model even better we included all the loans in Loan data set which are not Paid or current as Failed or Declined Loan**

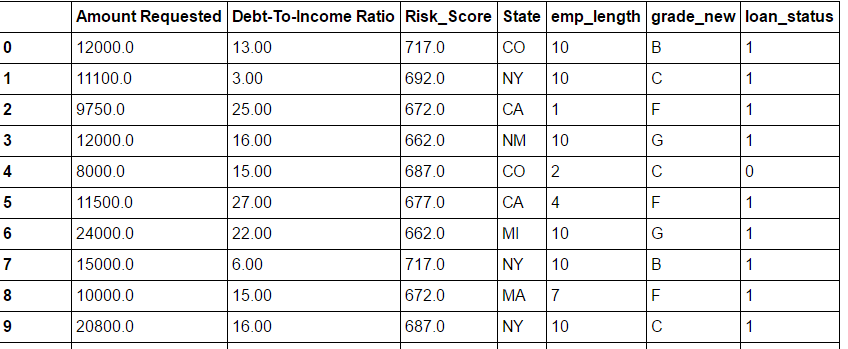
**Reason for doing this : As all the non -paid loans are actually deliquent loans so putting them in declined status makes for a better prediction**



1. The two data sets were merged together (Declined and Loan) :

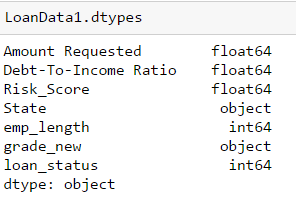


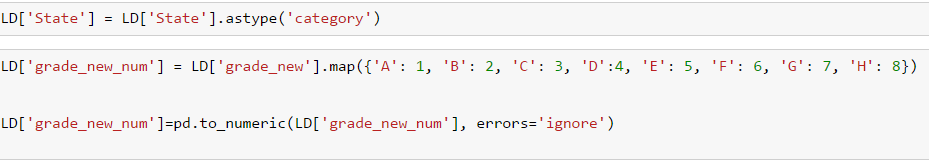
1. Values/Columns chosen for merged dataset:



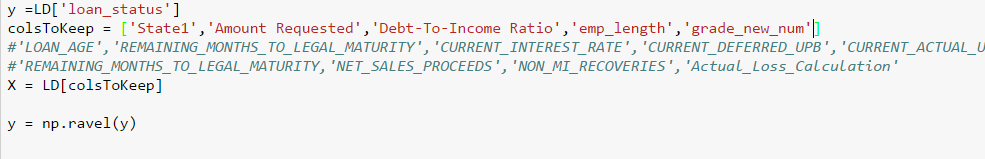
**Running Classification Models on Python Notebook:**

1. Changed data types as object to categorical:





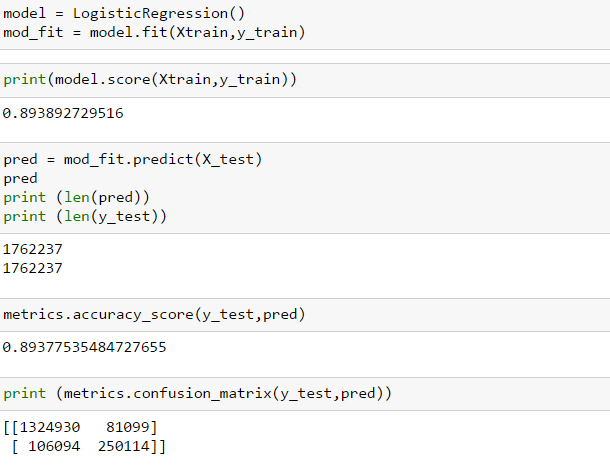
1. Chose the columns for data division:



1. Train /Test Data based on 70-30%

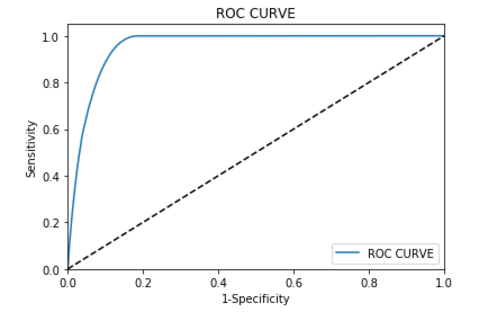


**Logistic Regression:**

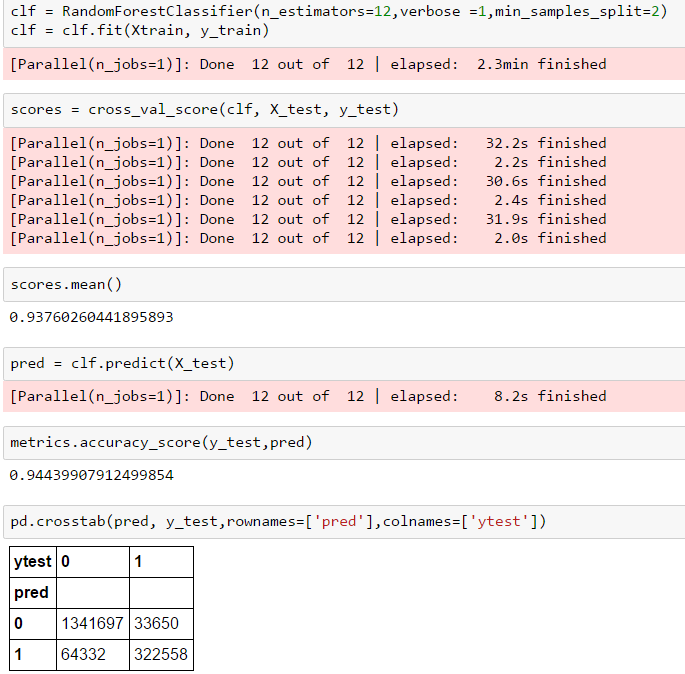


ACCURACY

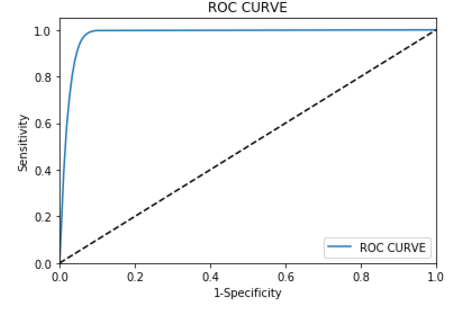
ROC curve for Logistic Regression:



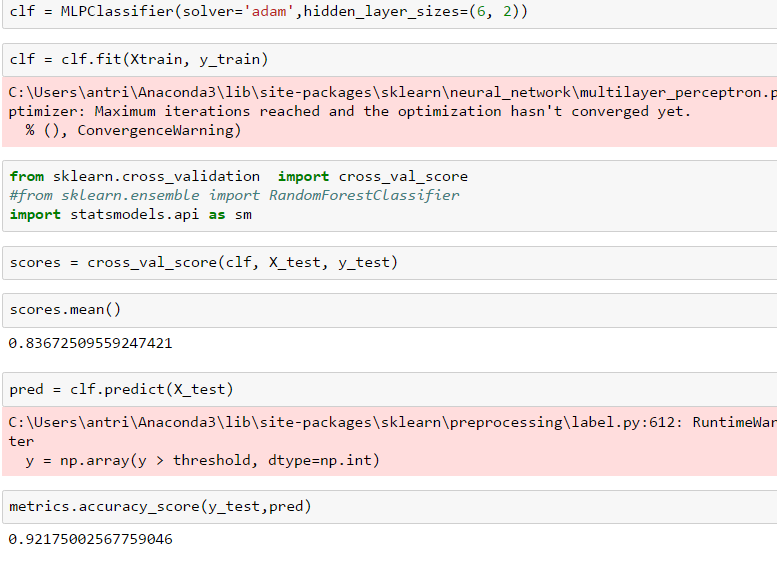
**Random Forest**

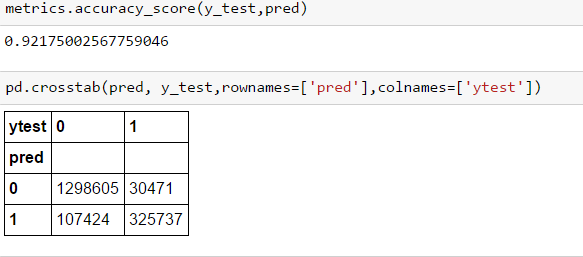


**ROC Curve for Random Forest:**

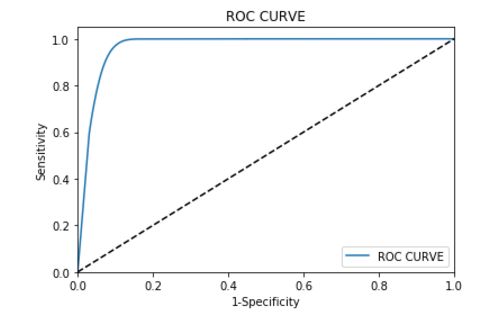


**Neural Network:**



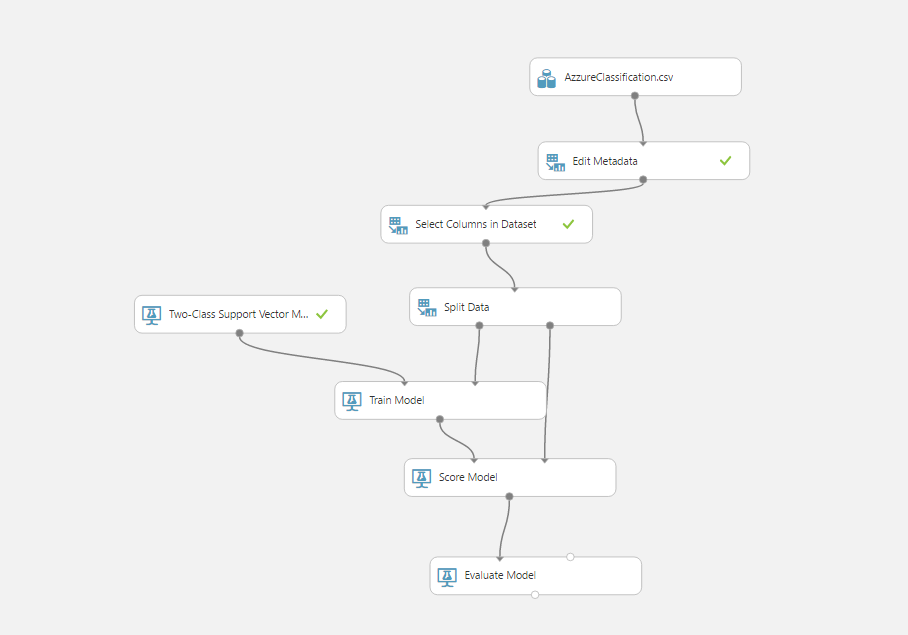


**ROC Curve for Neural Network :**

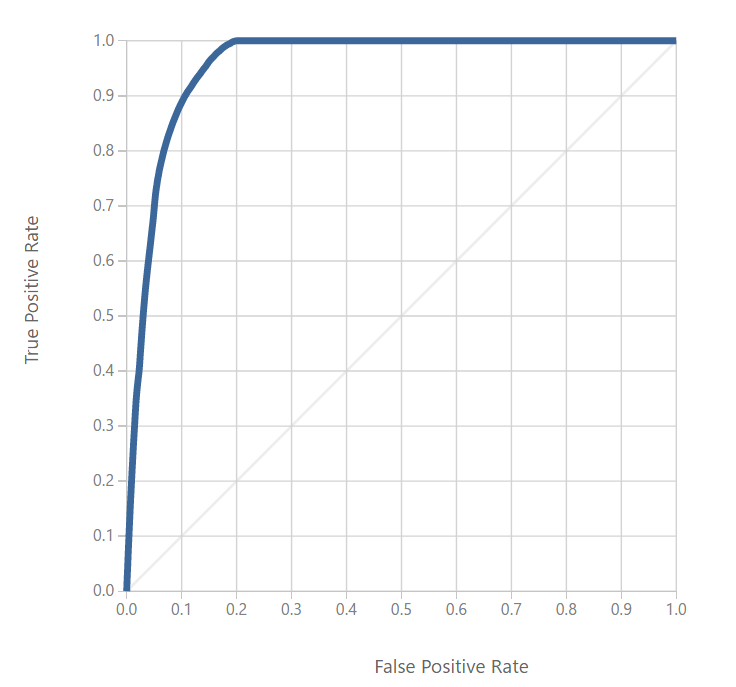


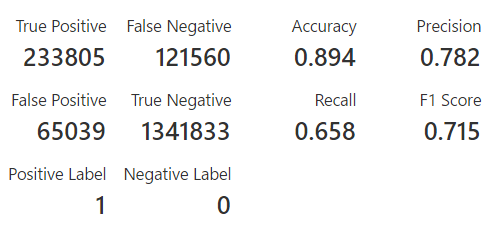
**Support Vector Machine:**

**Note : We tested SVM straight away on Azure as SVM was not running on our python machines (checked for more than 6 hours)**



ROC Curve for SVM:



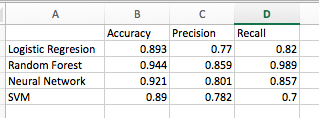


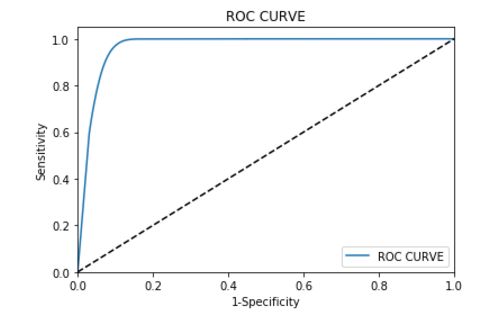
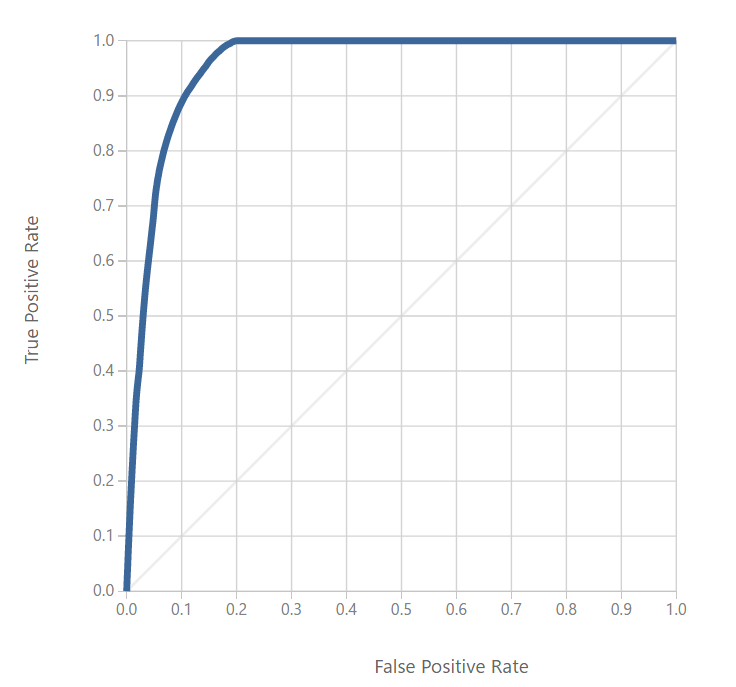
**Choosing Best Model:**

We looked at 4 values to make our choice :

1. Accuracy
2. Precision
3. Recall
4. ROC Curve

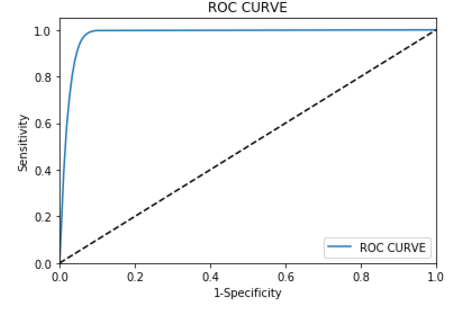
RANDOM FOREST HAS BEST ACCURACY, PRECISION



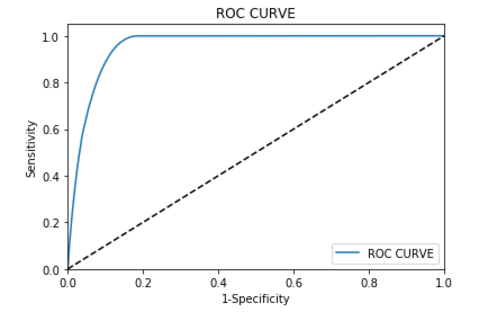


NEURAL  
NETWORK

SVM



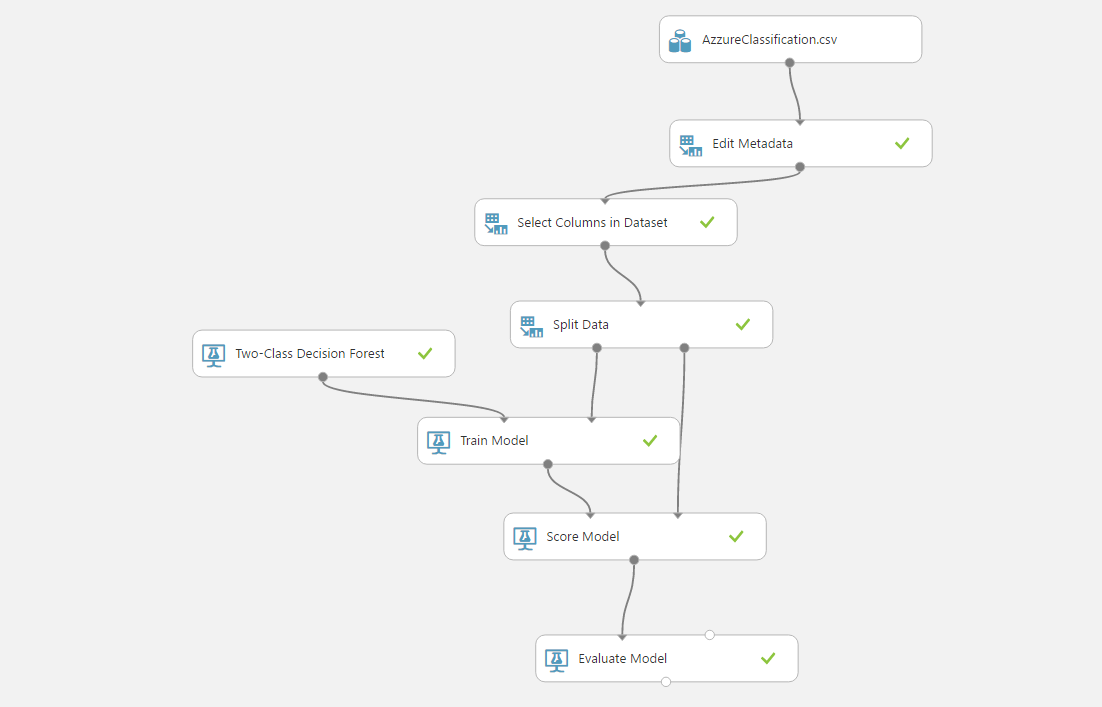
BEST AREA UNDER CURVE  
RANDOM FOREST



LOGISTIC REGRESSION

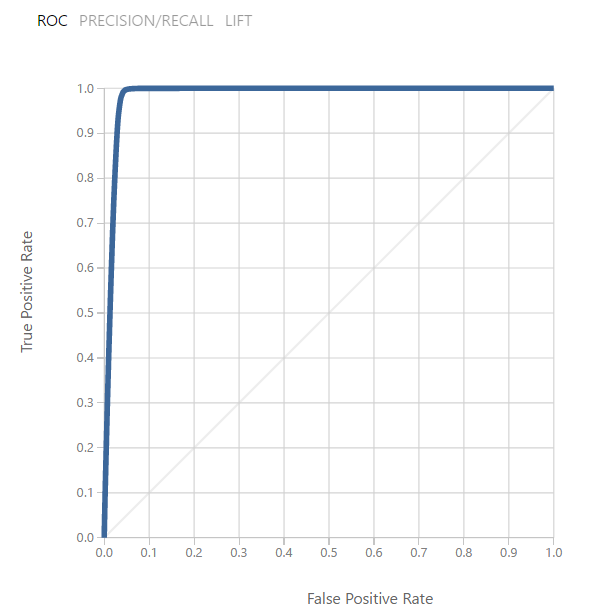
**Based on the values above and looking at the ROC Curve , we chose Random Forest as our best model.**

**Deploying the Best Model On Microsoft Azure:**

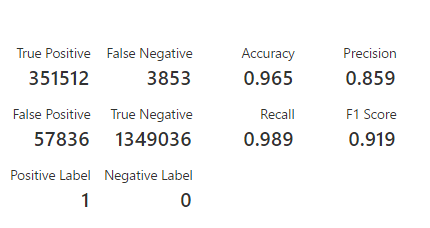


1. Used the input dataset created as a merge of Loan and Declined Dataset
2. We did not take Risk score as a parameter as Vantage score needs to be scaled and there is no fixed formulae and changed it using a formulae is not a good practice
3. As earlier stated , we used grade\_new based on Fico and vantage score

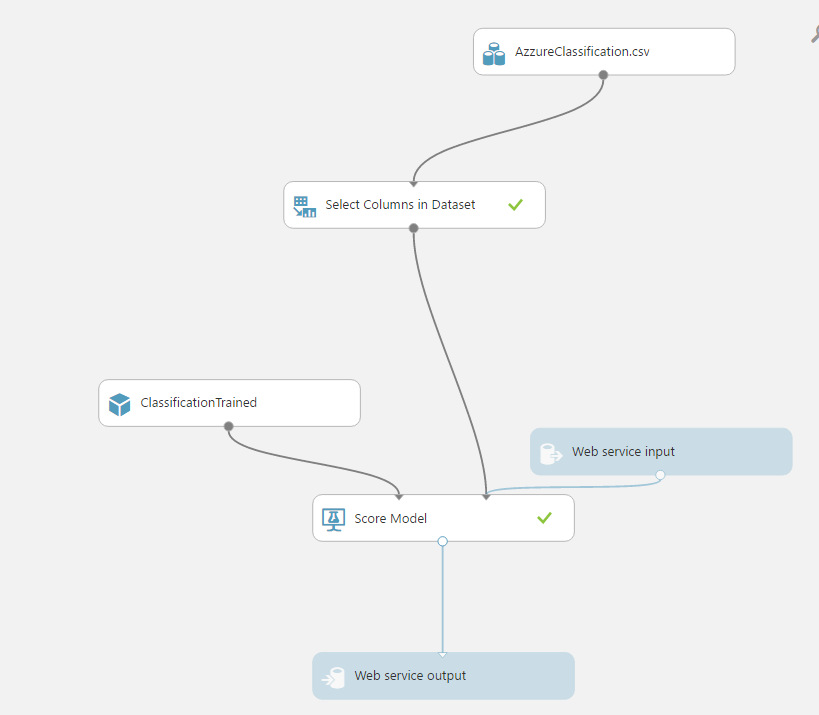
ROC Curve :



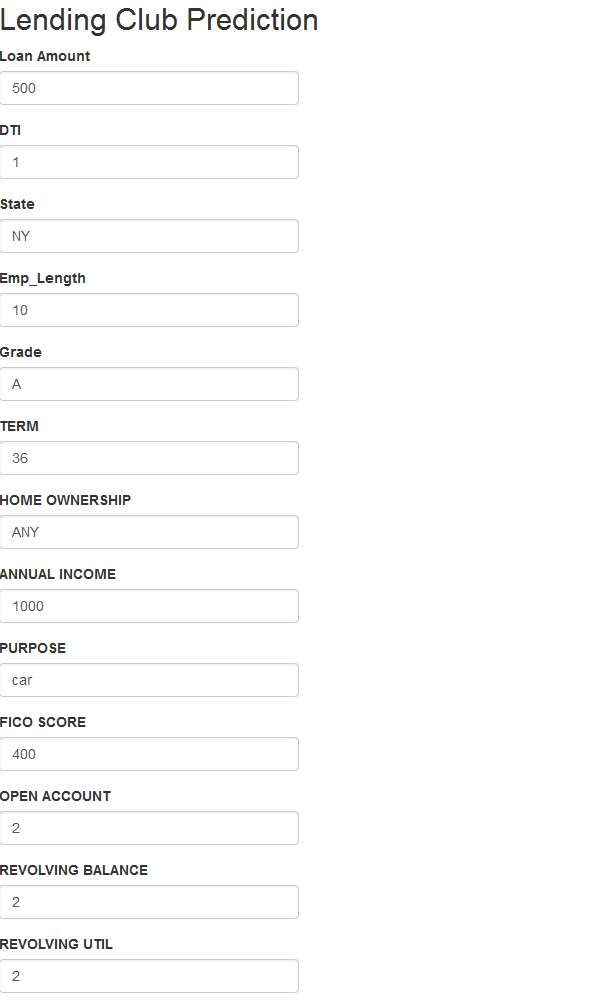
Stats on Azzure:



**Generating Rest API for Classification:**



**R Shiny was used for deployment of this API.**



**R-Shiny:**

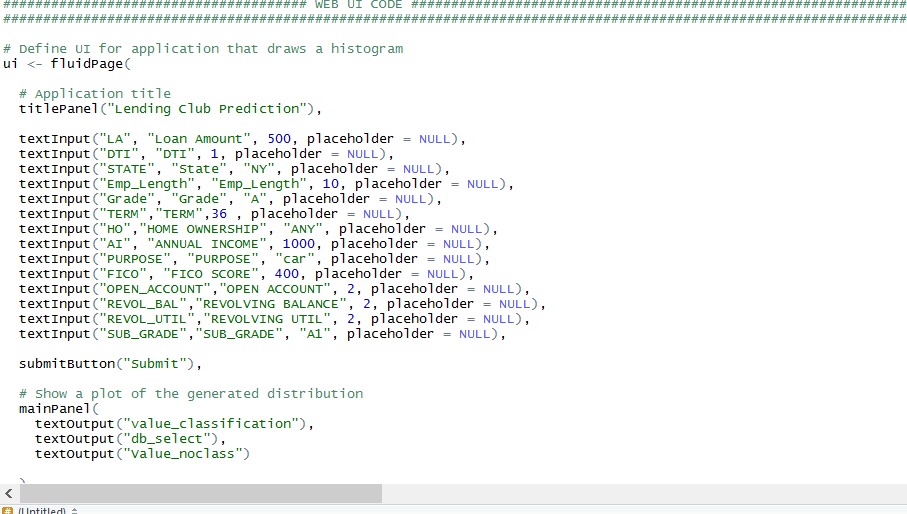
We chose R-Shiny for adhoc prediction of interest rate.

Shiny is a new package from RStudio that makes it incredibly easy to build interactive web applications with R.

**Features**

* Build useful web applications with only a few lines of code—no JavaScript required.
* Shiny applications are automatically “live” in the same way that spreadsheets are live. Outputs change instantly as users modify inputs, without requiring a reload of the browser.
* Shiny user interfaces can be built entirely using R, or can be written directly in HTML, CSS, and JavaScript for more flexibility.
* Pre-built output widgets for displaying plots, tables, and printed output of R objects.
* Fast bidirectional communication between the web browser and R using the websockets package.
* Uses a reactive programming model that eliminates messy event handling code, so you can focus on the code that really matters.

Sample Code in R-shiny

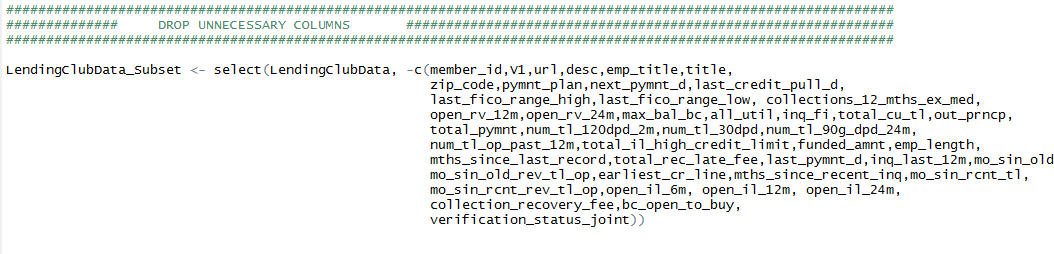


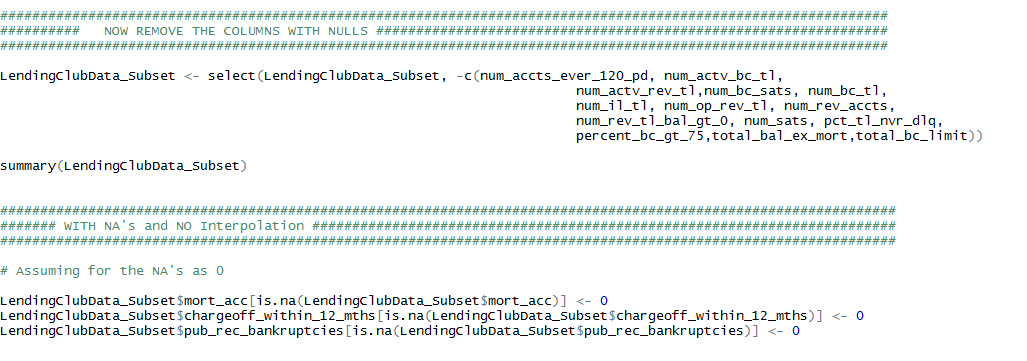
**Unsupervised Learning & Prediction**

Unsupervised Learning is performed when we are trying to predict labels for the unlabelled dataset.

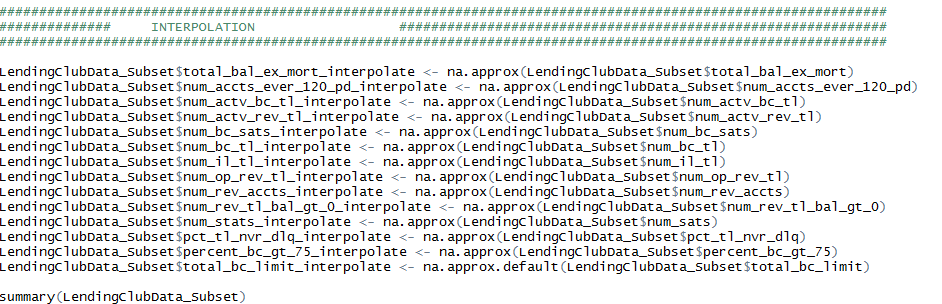
**Feature Engineering:**

1. In Unsupervised learning first we did feature engineering, in which we first validated the entire file and removed the missing values

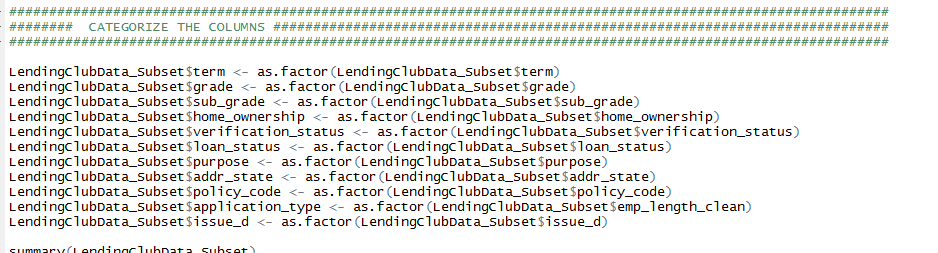




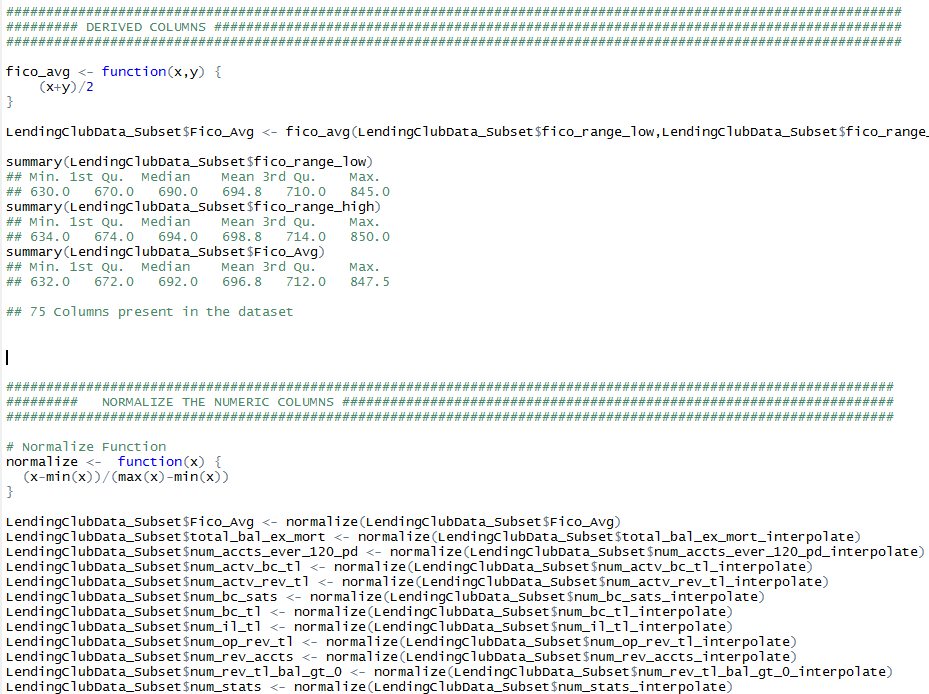
1. Some columns that seemed relevant to us, we interpolated those values



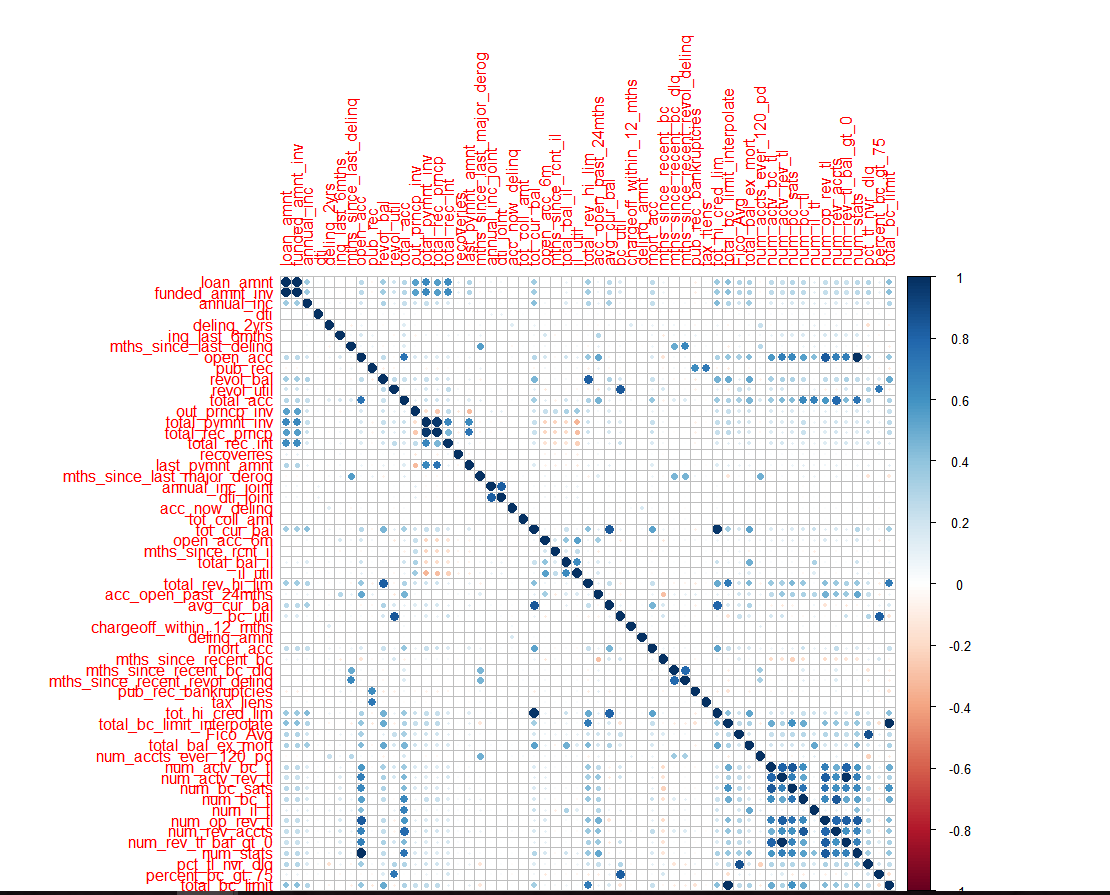
1. Then we chose to convert the categorical columns in the categorical values



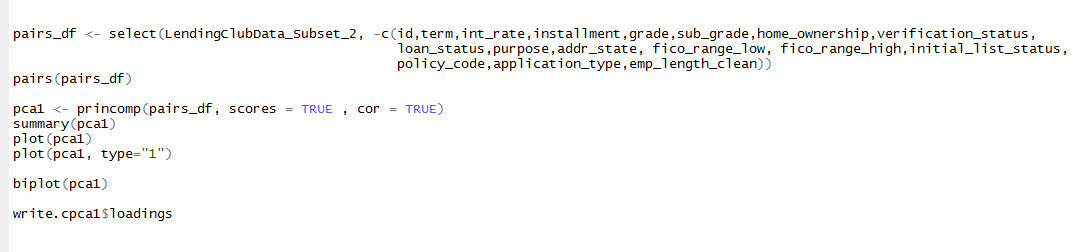
1. We, then chose to normalize the numeric column to put them one scale



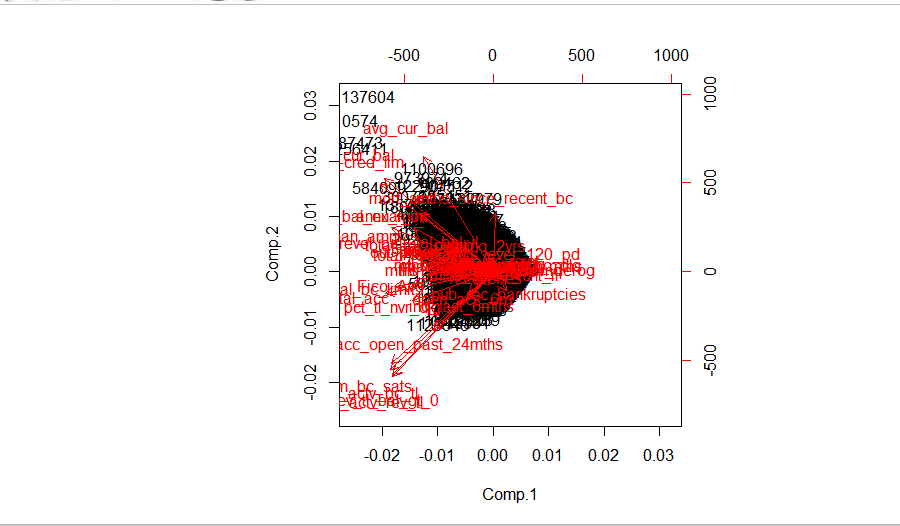
1. After which we created a correlation plot and removed the most correlated columns which weren’t relevant to us.



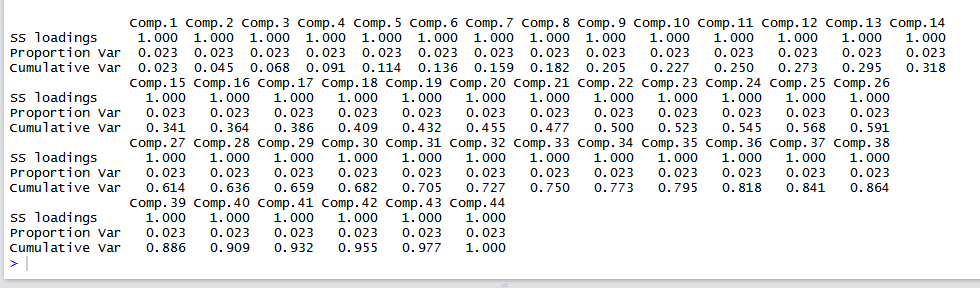
1. We then performed PCA – that is principal component analysis and analysed the loadings to check the variables that cause the maximum variance in the dataset and reduce the dimensionality, so that we get the best features



PCA Graph



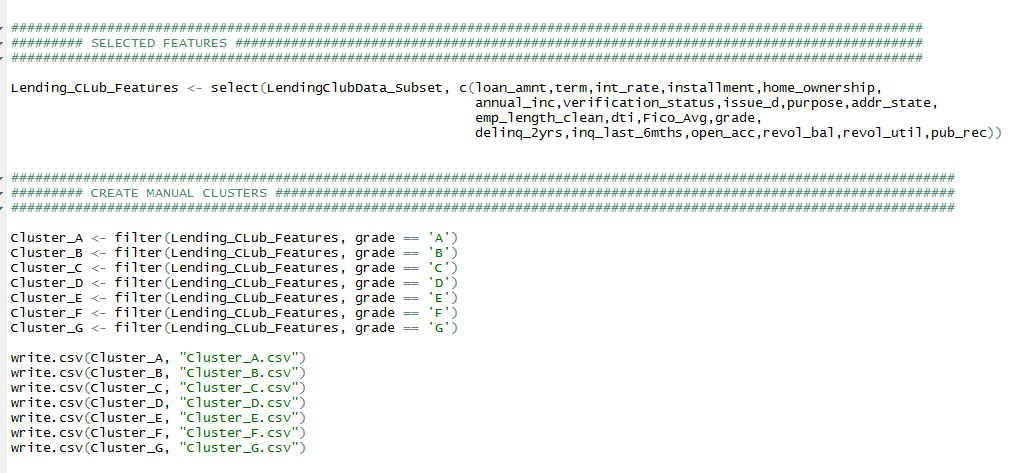
PCA Loadings



1. After dimensionality reduction, we then further chose our features based on RFE in python and exhaustive and forward backward searches and has to remove some columns that were not relevant for user input or were used even before the user got the loan.

**Manual Cluster Creation**

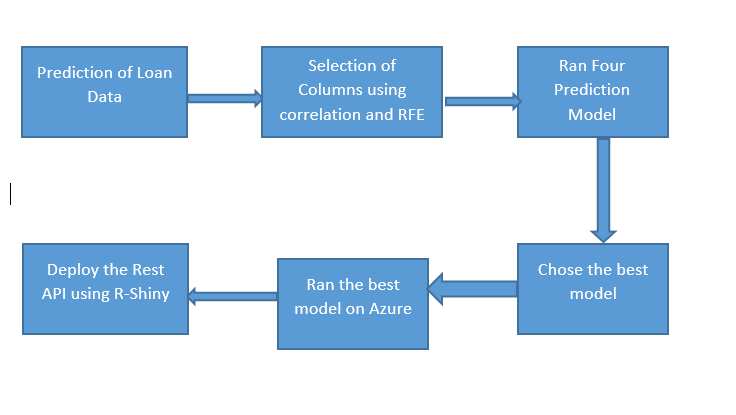
We, then created the Manual Clusters per the grades present in the dataset, there were around 7 Clusters formed per 7 grades namely A, B, C, D, E, F, G



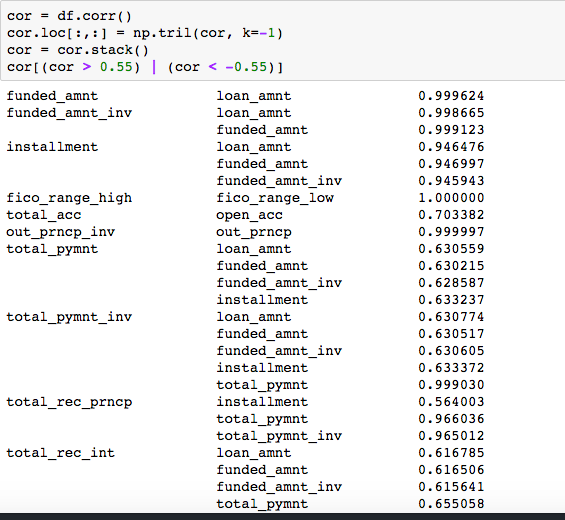
**Predictions on Clusters**

As Per the Clusters we then chose to perform Interest Rate predictions on the manually clustered data.

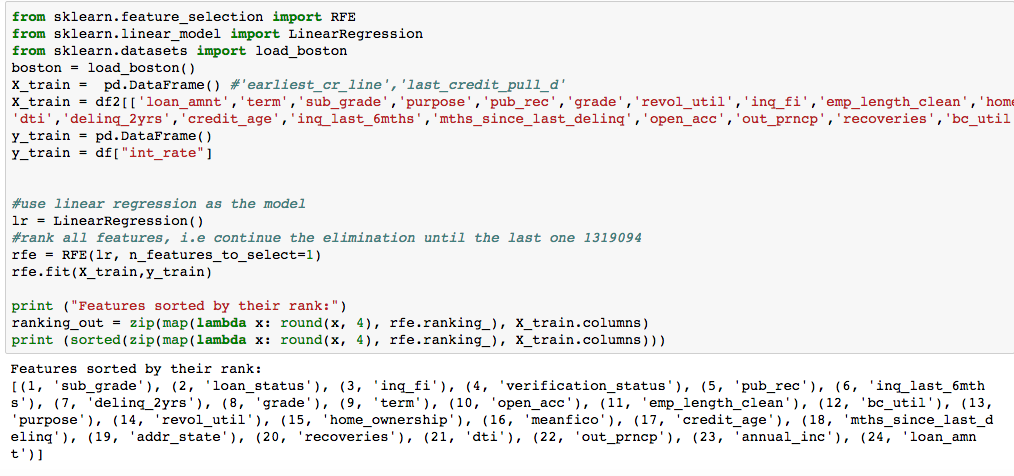
***Flow for Prediction:***



1. Once we have decided that we will give a loan and clusters are created both manually and using clustering algorithm, we will predict the loan’s interest rate. We have further done even more feature engineering and chosen 17 columns for the prediction
2. We found correlation between all the columns and columns with high correlations were dropped. We removed id, member id, title, desc and other irrelevant columns.

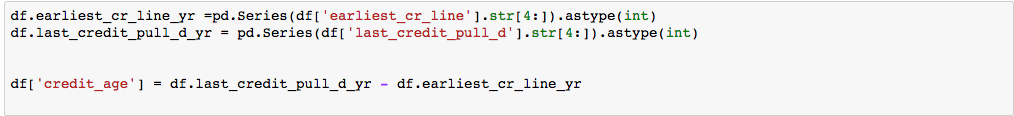


1. After correlation, we calculated ranking of each column with respect to interest rate using RFE. RFE helps in identifying which parameter affects the most in predicting the interest rate.

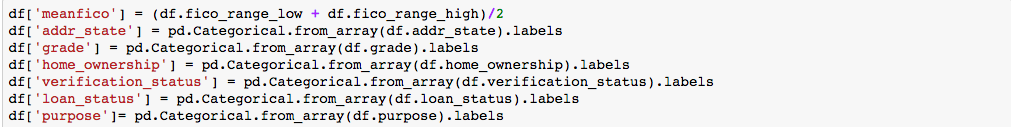


1. Derived Columns –

Age of credit history reflects the length of your experience with the credit system. Created Credit age by subtracting last credit pull and earliest credit line.



Created average of fico and labels for categorical columns as Regression models take only numeric data



Screen%20Shot%202017-04-14%20at%201.53.12%20PM.png

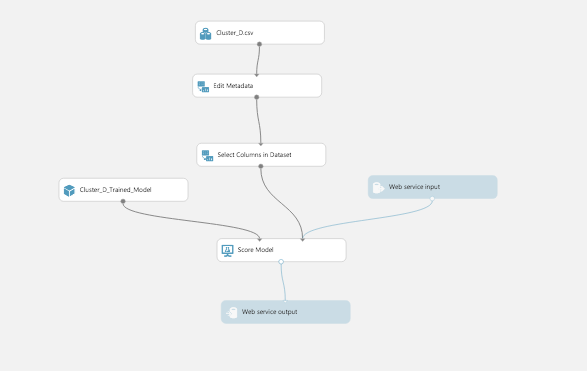
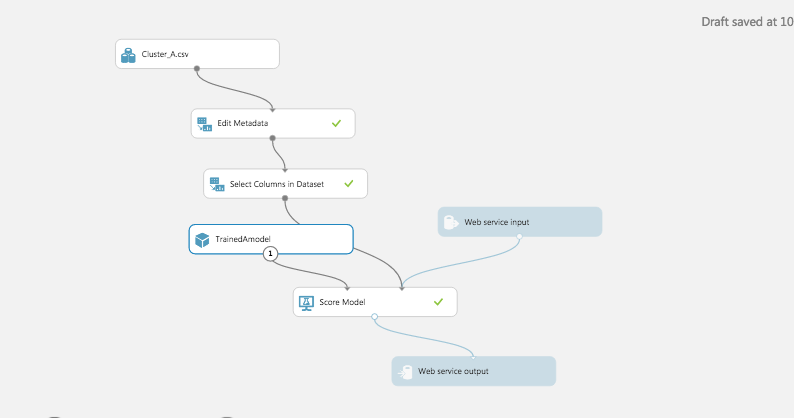
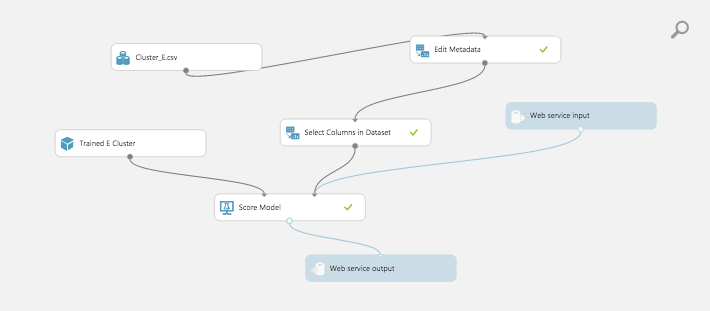
***Selected Columns for Prediction:***

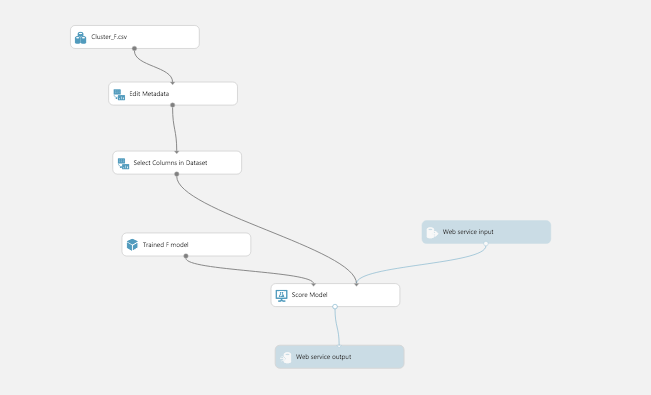
* ***loan\_amnt:*** The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
* Term: The number of payments on the loan. Values are in months and can be either 36 or 60.
* ***sub\_grade:*** The number of payments on the loan. Values are in months and can be either 36 or 60.
* ***pub\_rec***
* ***home\_ownership:*** The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
* ***annual\_inc:*** The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
* ***verification\_status***: Indicates if income was verified by LC, not verified, or if the income source was verified.
* Purpose: A category provided by the borrower for the loan request.
* ***addr\_state:*** The state provided by the borrower in the loan application
* ***dti:*** A ratio calculated using the borrower’s total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower’s self-reported monthly income.
* ***delinq\_2yrs:*** The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
* ***inq\_last\_6mths:*** The number of inquiries in past 6 months (excluding auto and mortgage inquiries)
* ***meanfico:*** average fico of fico\_range\_high and fico\_range\_low
* ***emp\_length:*** Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
* ***Open\_acc:*** The number of open credit lines in the borrower's credit file.
* ***Revol\_util:*** Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.

***Prediction Algorithms were performed to predict the best Interest Rate (Manual Clustering)***

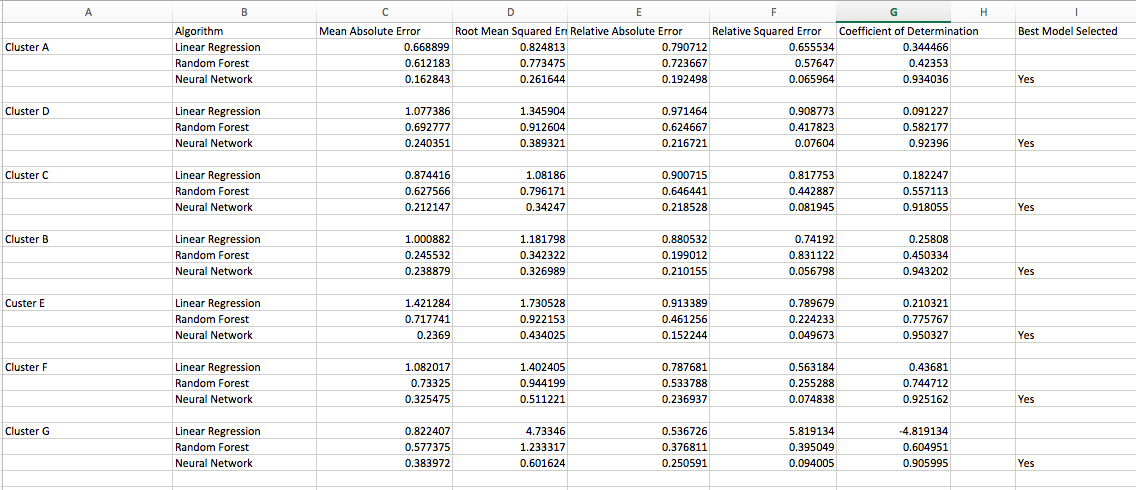
After feature selection, we saved the dataset for each cluster separately in 7 csv files.

We used this as the input data to our prediction model in Azure. The files are imported into the experiment as mentioned previously. Once that is done we took each cluster individually and created a trained model of it. Once the trained model is saved we ran the model with the test data for prediction and set up the web service. With the help of the API key and the R sample code we created the rest API.





***The matrix for each cluster is:***



**Based on Root Mean Squared Error and Co-efficient of Determination the best model for each cluster is Neural Network.**

**Algorithm Cluster Creation**

For programmatically downloading creating the clusters, we have used Rstudio with the help of some important functions like daisy and pam to calculate the medoids.

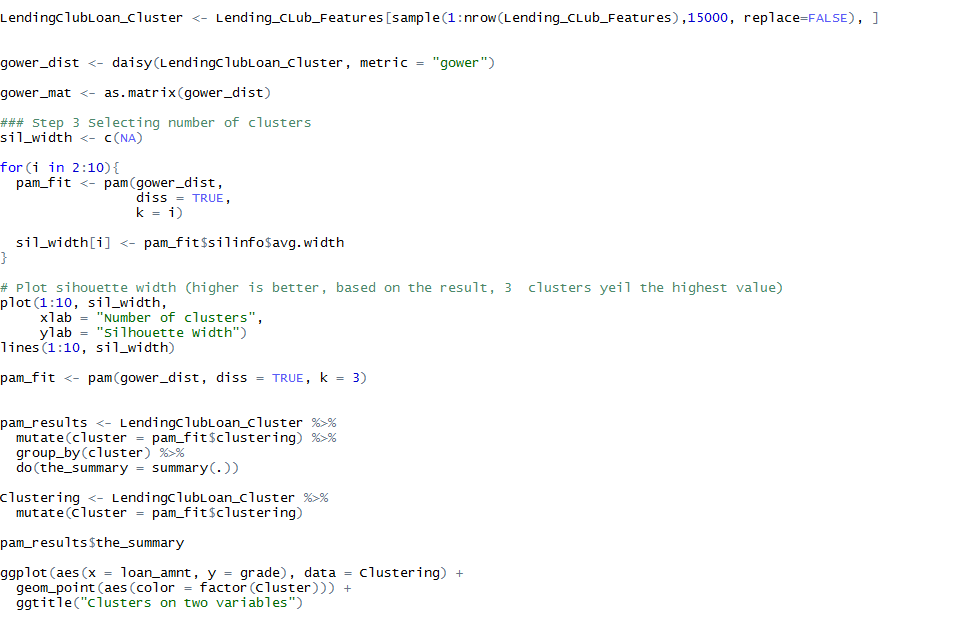
**Objective:**

* Our objective here is to first calculate the dissimilarity between the observations by using a distance formula
* Grouping/binning the same observations together into clusters and choosing the best number of clusters
* Then visualize those clusters

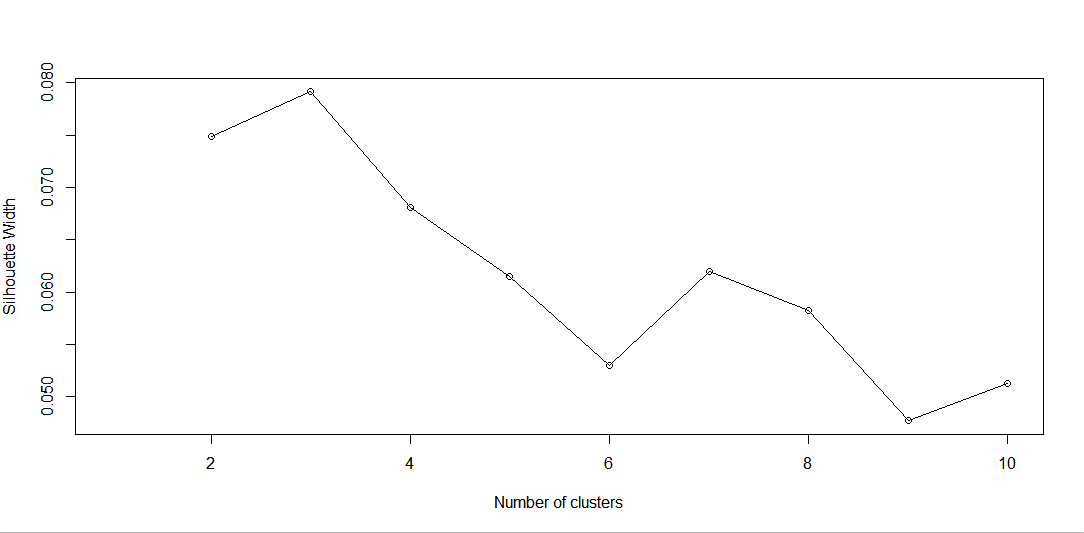
We have used **PAM** here to choose k random entities and group them into medoids through the distance matrix

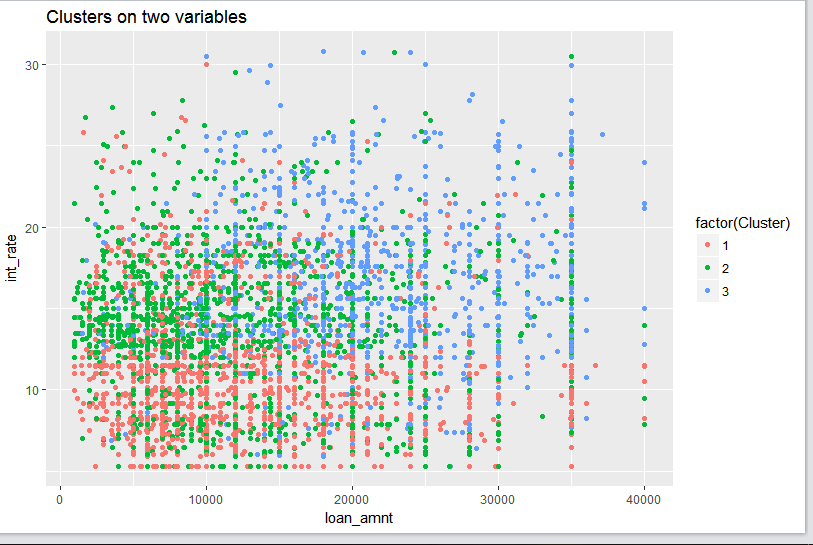
**Gower –** is the distance formula used to calculate the distance between the continuous and categorical features together. We use the daisy function to calculate the distance and convert it into a matrix

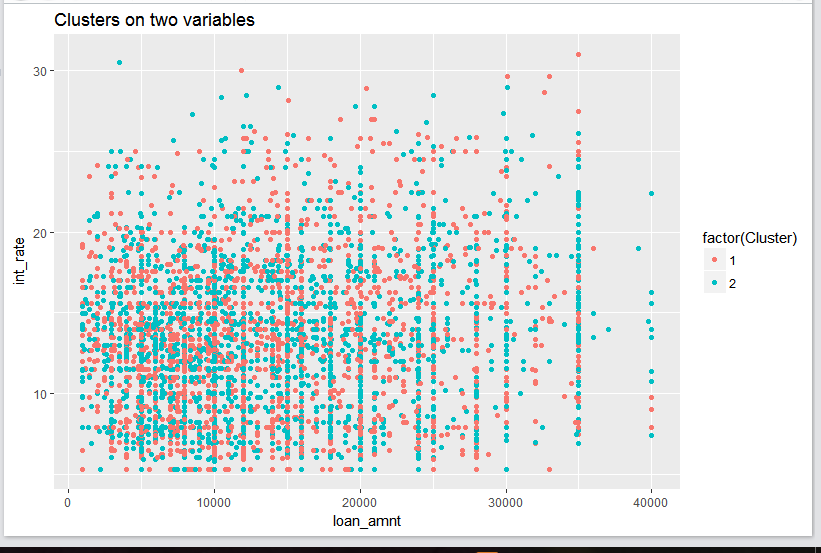
Once the matrix is created we use the PAM function to run iteratively on ***2 to 20 clusters*** and choose the best number of clusters through this PAM function and the ***Bend Graph***



According to this graph we know that we can use 3 clusters, we tried the predictions from 2 to 3 clusters and found that 2 clusters would be our best selection





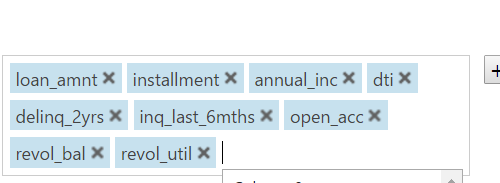


***Running K-Means on Azure***

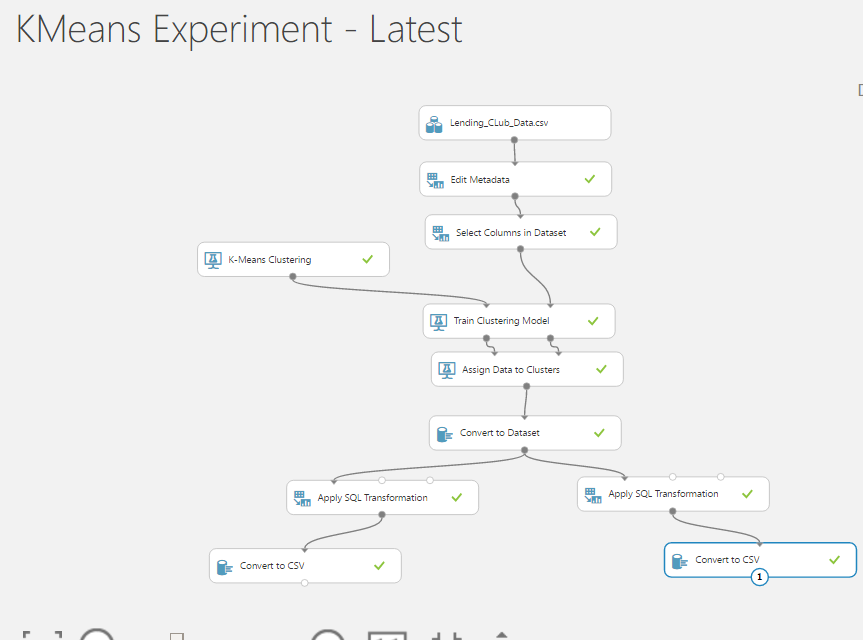
K-means is used for creating clusters of similar observations grouped together on the basis on the numeric columns that we provide by calculating the distance between the observations.

We implemented the K-Means clustering on Azure by creating two clusters

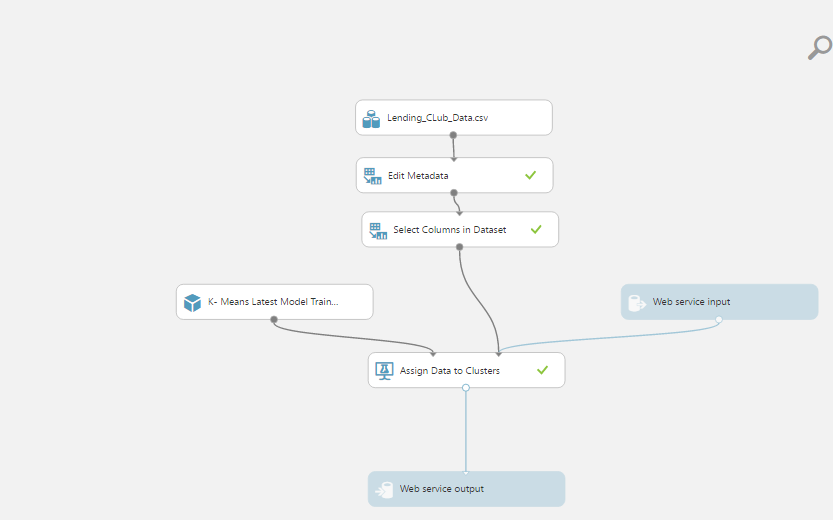
1. *Selected Columns*



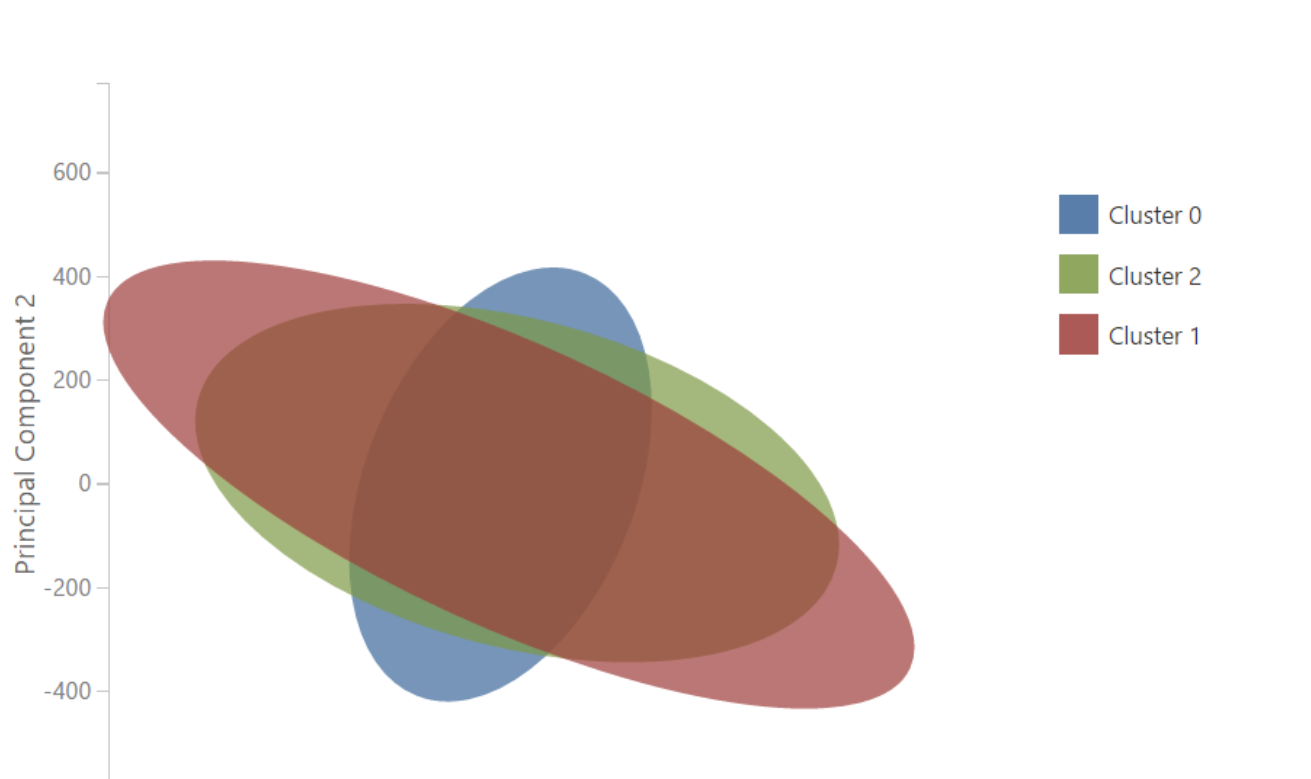
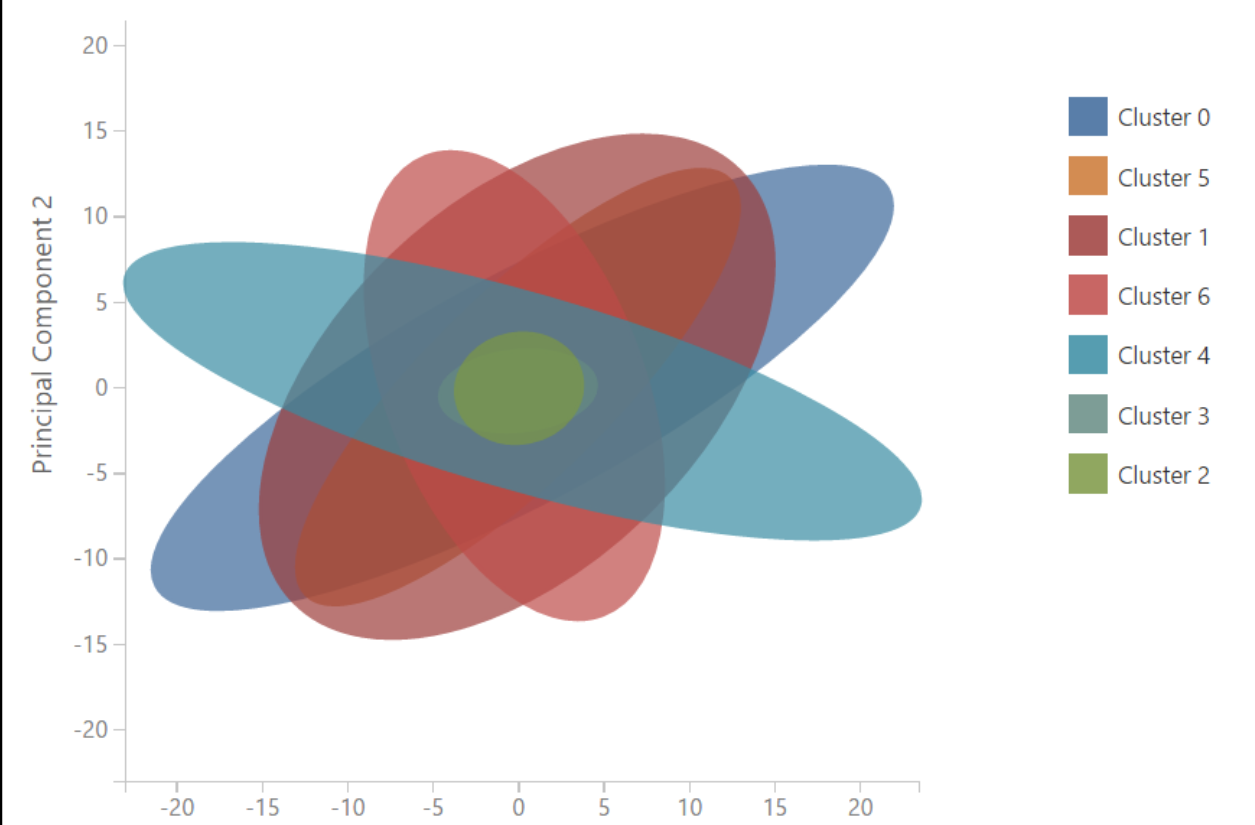
1. *KMeans Network on Azure*



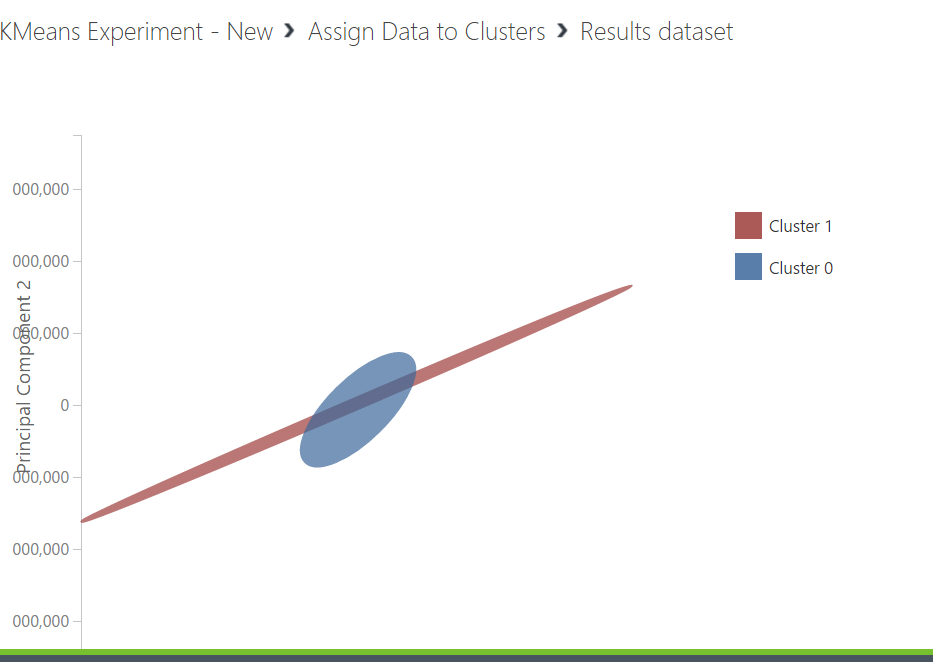
*Web Deployment*



**Experimental Clusters**



**Analysis Graph**

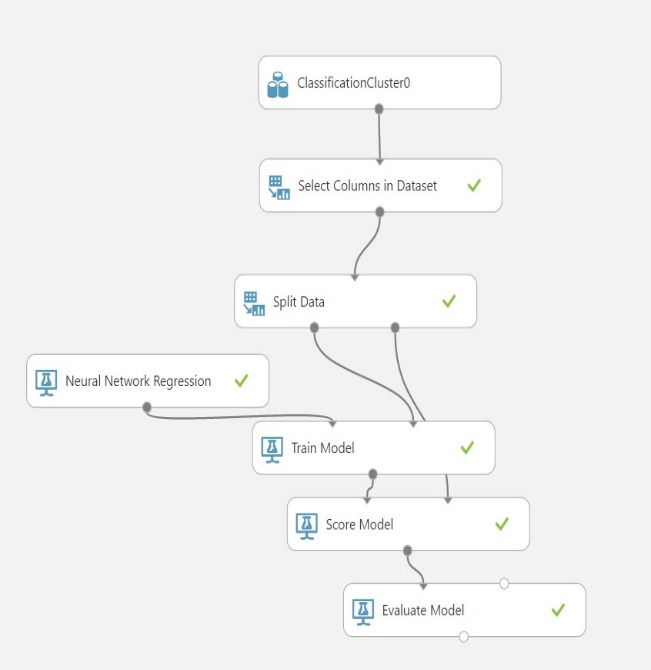
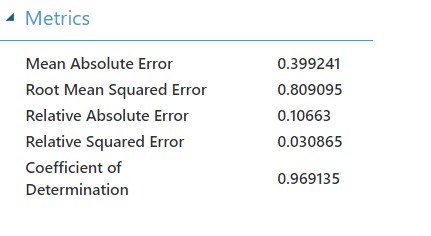


1. Saved the Cluster Datasets separately and used for Predictions to predict the interest rate

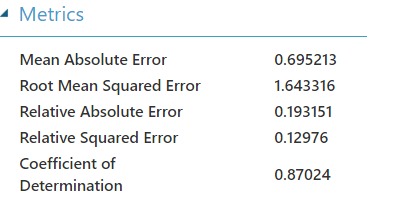
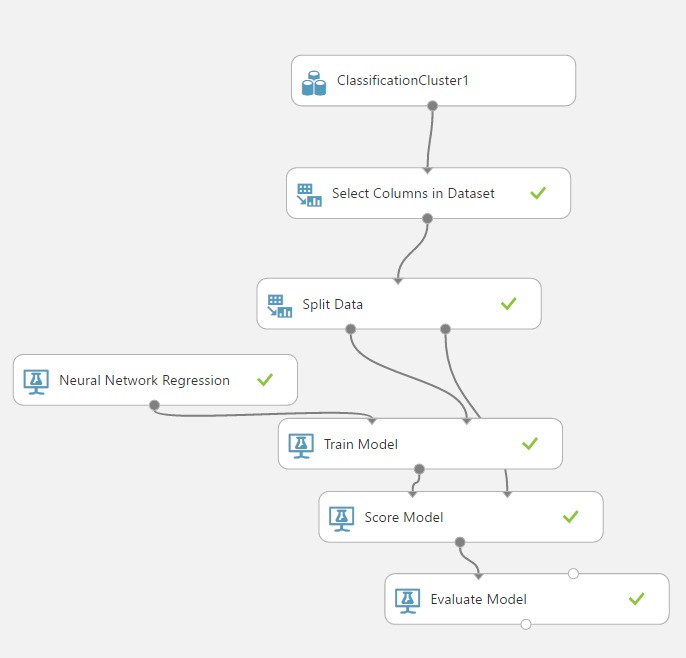
**Prediction on Algorithmic Clusters**

We Chose Neural Network as the best network to run on the Clustered Data

1. Cluster 0 Predictions



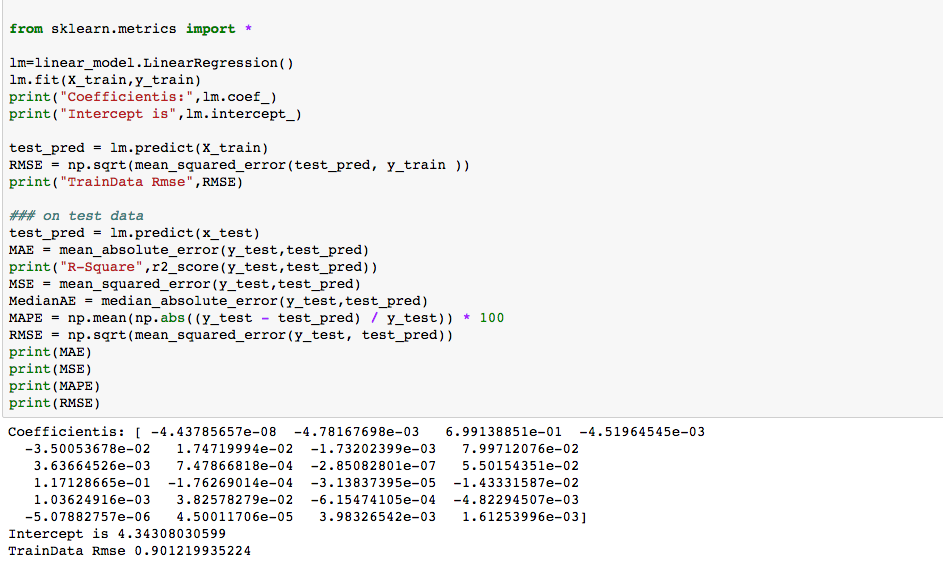
1. Cluster 1 Predictions



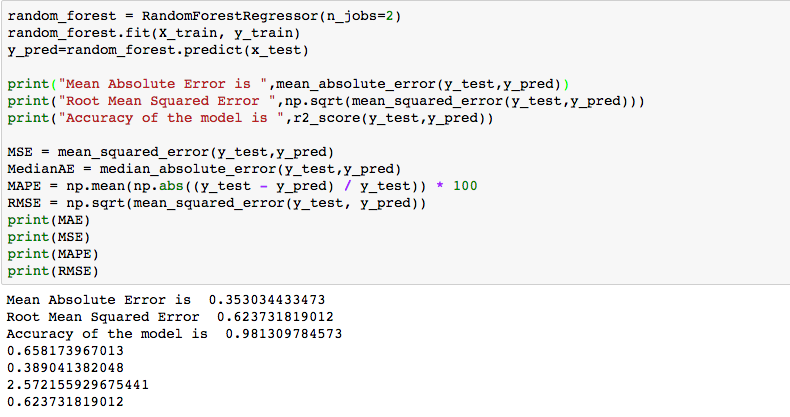
**Prediction on No Cluster Creation**

This is where we run the code without any clustering and run four prediction models and chose the highest predicted interest rate

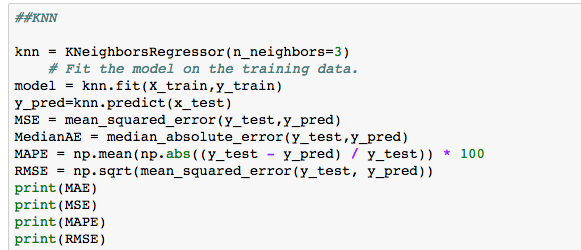
Linear Regression:



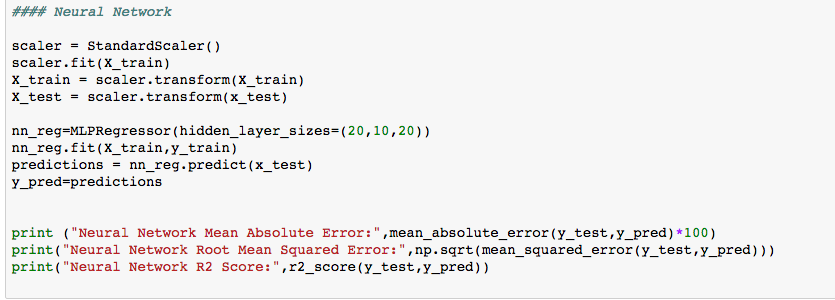
Random Forest:



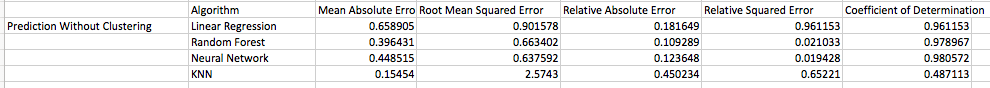
KNN



Neural Network:



Matrix for comparison:



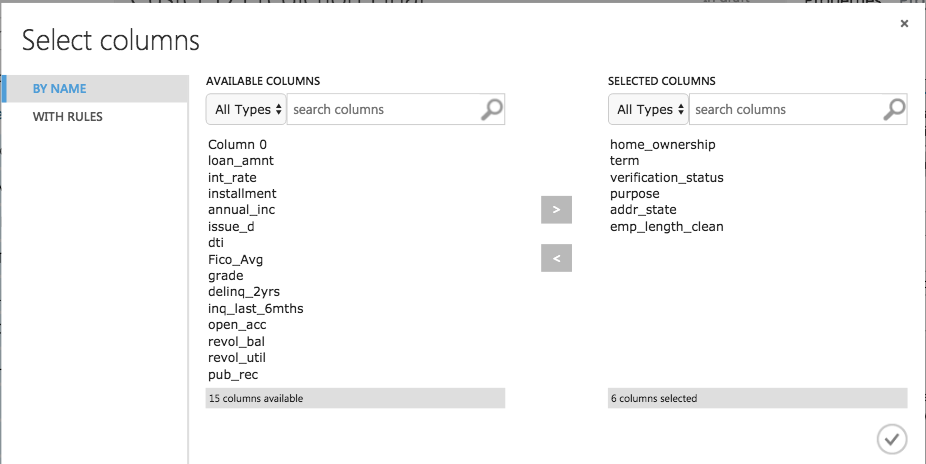
**Since Neural Network’s RMSE and Coefficient of Determination was less compared to others we chose Neural Network as the best model.**

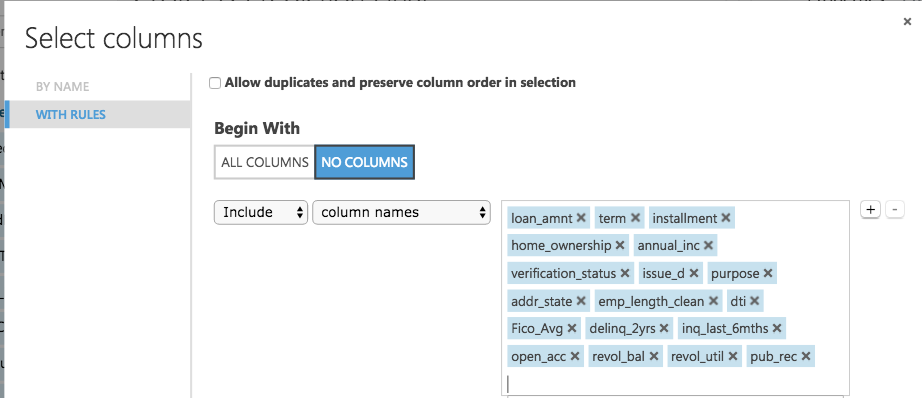
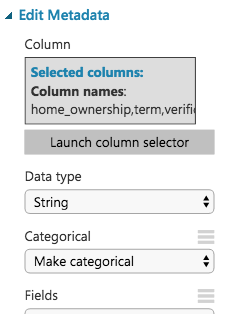
Note: we performed the tasks using Windows Azure also, to perform the same tasks

***Best Model Selection in Azure***

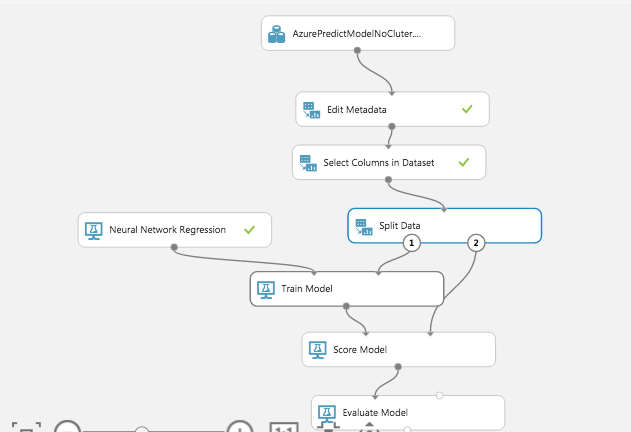
We saved the dataset with the columns that are required for the prediction in a csv file from Jupyter python after feature selection.

1. Imported the data from the local into the new experiment created.
2. Converted string type columns to category type using edit metadata module.

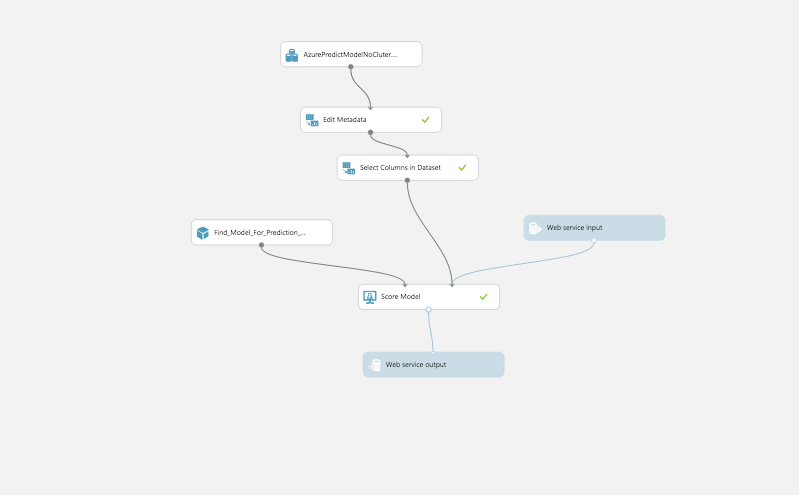




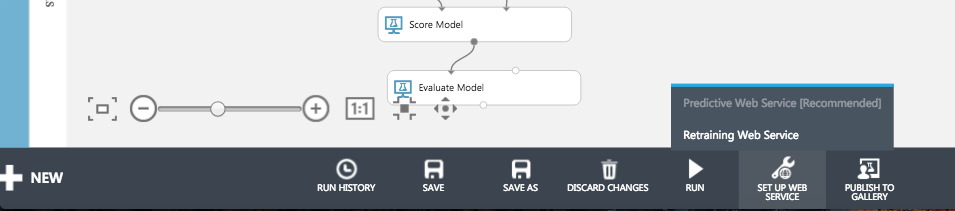
1. Then we split the data into 70-30 using the split module. The output of spilt goes to the train model module and to the score module. We chose Neural Network Regression module in the prediction and train model that the model as the input and trains the model. Once the model is trained its score and other parameters such as RMSE, RME can be viewed in score and evaluate modules.



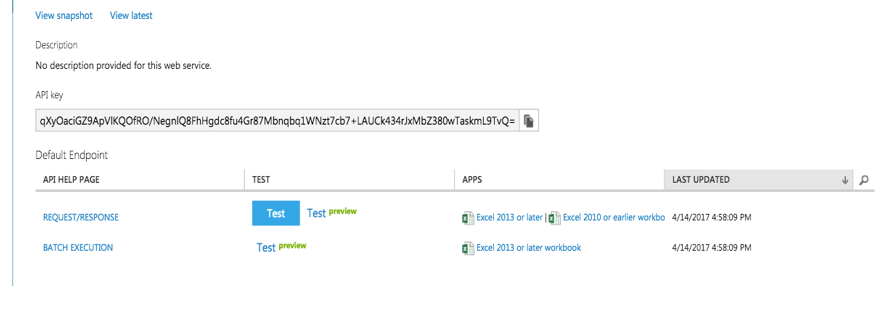
1. Once the model is trained we saved the model and then used that model on the test data to predict the interest rate as shown below.



1. Once, the model runs successfully, we set it up as a web service as shown below



1. Once it is deployed an API key will be generated.



1. We have used this API and the R code present in the Request/Repose link to create the rest API.

