**Advance Data Science/Architecture**

Assignment 2

Submitted By:

Karan Millan Thevar

Pratik Patre

Rahul Chandra

(Team 3)

14th April, 2017

Introduction

This project focuses on an entire cycle of whether to give loan to a customer or not and if yes then at what interest rate. The project is divided into three phases:

1. Classification
2. Clustering
3. Prediction

Classification: The project starts by first training the classification models based on Loan data from Lending Club Dataset. The classification models trained and tested are Logistic Regression, SVN, Neural Network and Random Forest. Based on the confusion matrix and ROC curve, the best classifying model is selected. This model is deployed on Microsoft Azure and a REST API is created for this model. Further, the REST API is linked to an interactive user input page. Based on the input from the user, the model will decide whether to provide the user with loan or no. The result will be displayed on the interactive page. If the user is approved for a loan amount, there would be further details required from the user.

Clustering: The Loan dataset is used to form three cluster segments.

1. Manual Cluster: We have made an algorithm where the data is segregated into clusters. We have described the algorithm in detail below. Each of the clusters will have a best prediction algorithm to predict the interest rate.
2. Clustering Algorithm: The Dataset is used to train and test two models, namely, Kmeans and Hierarchical model. The best model is deployed on the Microsoft Azure and a REST API is created for this model. This REST API will decide, the user falls in which cluster based on the information provided by the user.
3. No Cluster: The information provided by the user will directly be used for prediction.

Prediction: Based on the clusters, The Lending Club Loan dataset is used to train and test the prediction models namely Linear Regression, Neural Network, Knn and Random Forest. The Root mean square error (RMSE), Mean Absolute Error (MAE), Relative Absolute Error (RAE) and Relative Squared Error (Relative Squared Error) are computed for each of the prediction models. The best prediction models are derived from these matrices for each of the clusters and deployed on Azure. The user input details are further used by these best models to predict the interest rate. The highest interest rate is then shown as the result to the user.

The project has involved various technologies like Microsoft Azure, Rstudio, Jupyter, Spark, CSS, Visual Studio.

**Classification**

**Clustering**

**Prediction**

**Part 1: Classification:**

It is important to get data from both declined and accepted dataset so that classification models can work on a combined dataset which will be used to predict whether to give loan to customer or not.

We have merged data from Accepted loan dataset and Declined loan dataset along with a column indicating a whether the record is an accepted loan record or a declined loan record.

The Flag representing whether the loan was approved or reject is LOAN APPROVED. It has a flag of ‘1’ for approved loan and ‘0’ for rejected loan.

The following columns are included in the classification dataset.

1. Loan Amount
2. Fico Score
3. DTI
4. ZIP Code
5. Employee Length
6. State
7. Policy Code
8. Loan Approved

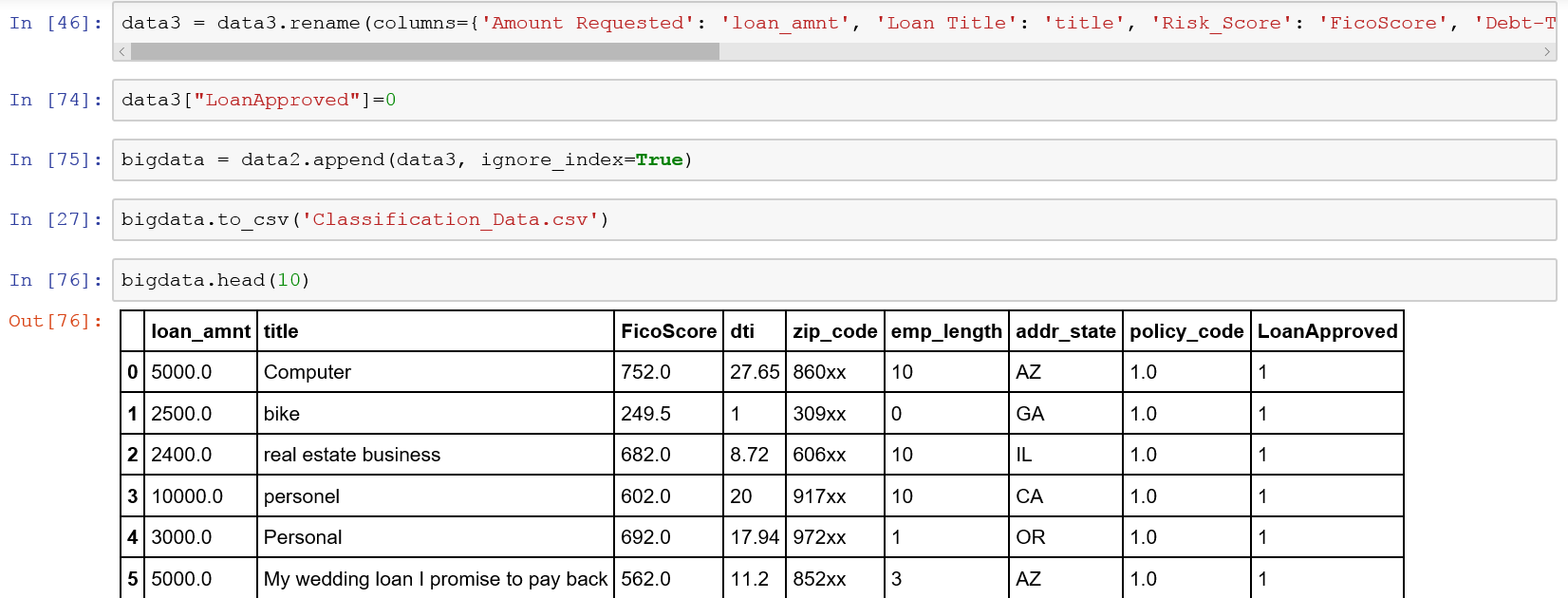
File:



Python Code:

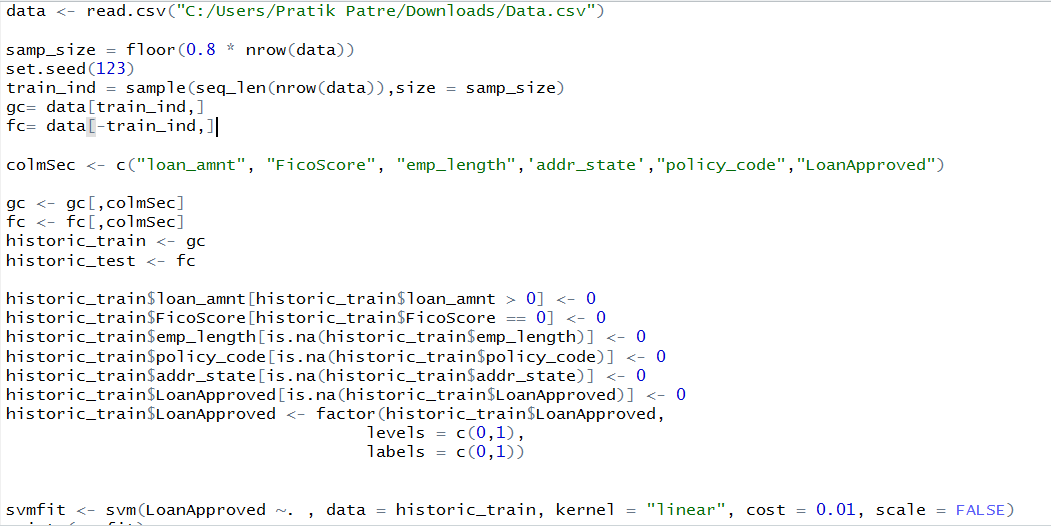


Python Code Snapshot:



We have run the following algorithms in R on this combined dataset.

1. SVN Algorithm



R Code:



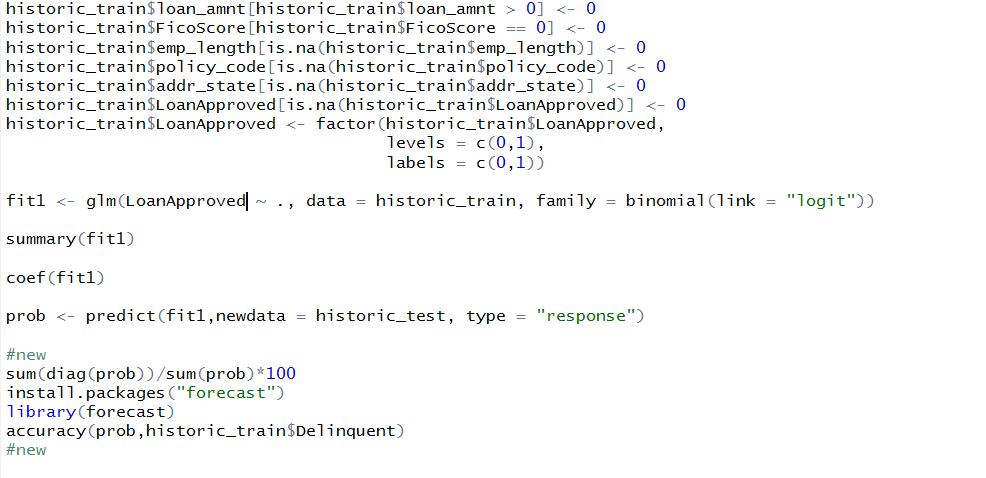
1. Neural Network Algorithm



R Code:



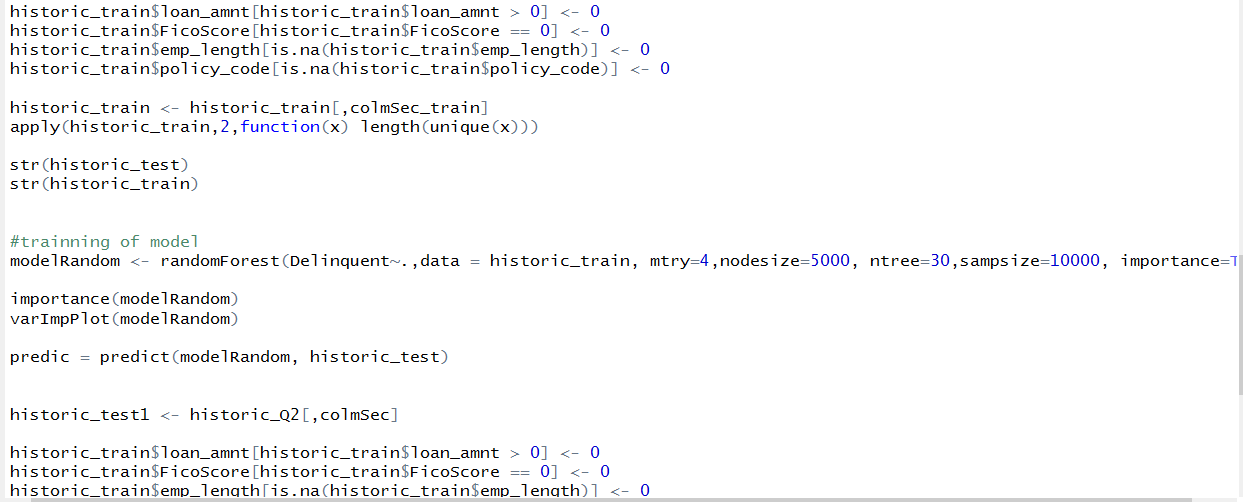
1. Logistic Regression



R Code:



1. Random Forest Algorithm



R Code:



We could not compute proper summary matrices for these models in R as the data was huge and we did not want our model to be trained on sample data.

For the prediction to be accurate approximately, the algorithm must work on the entire dataset.

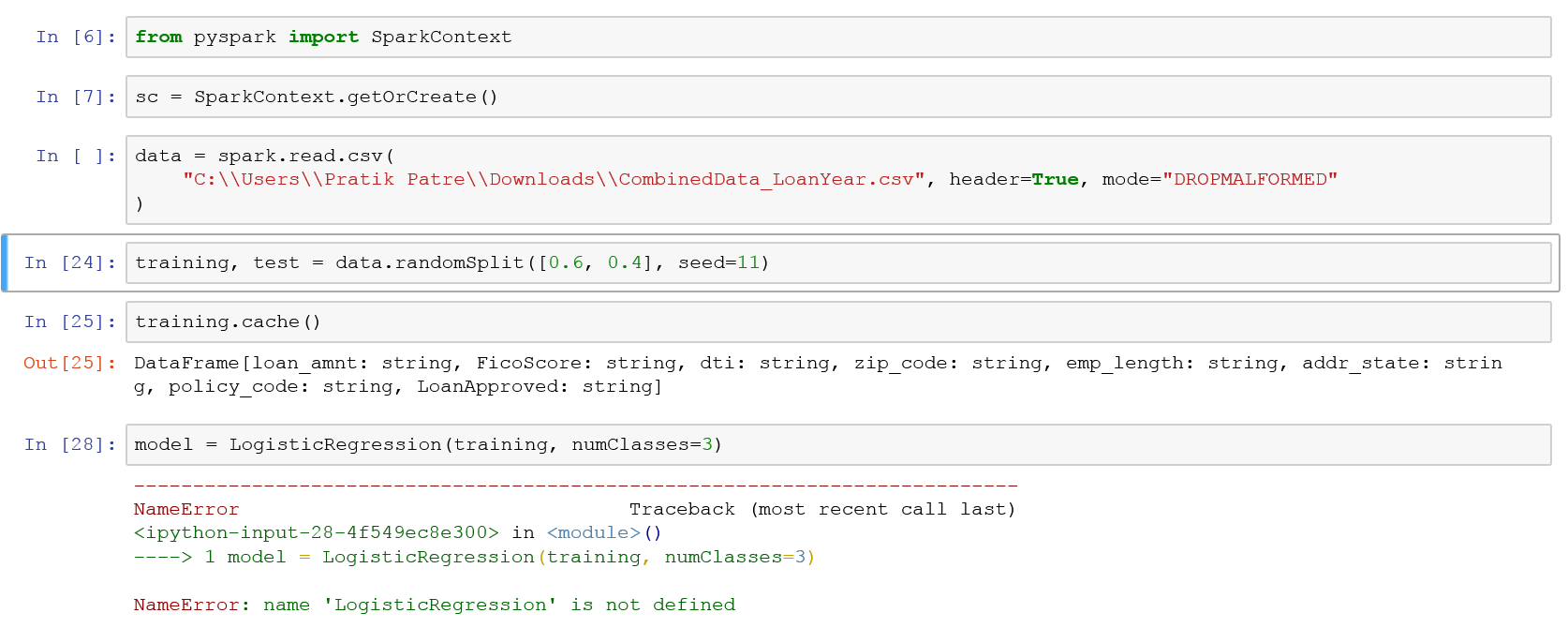
For this purpose, we have implemented spark. We have integrated spark with R using sparkR and have run the logistic regression algorithm. However, although we could run the model and retrieve predictions, we could not evaluate the model as per its performance.



R code:

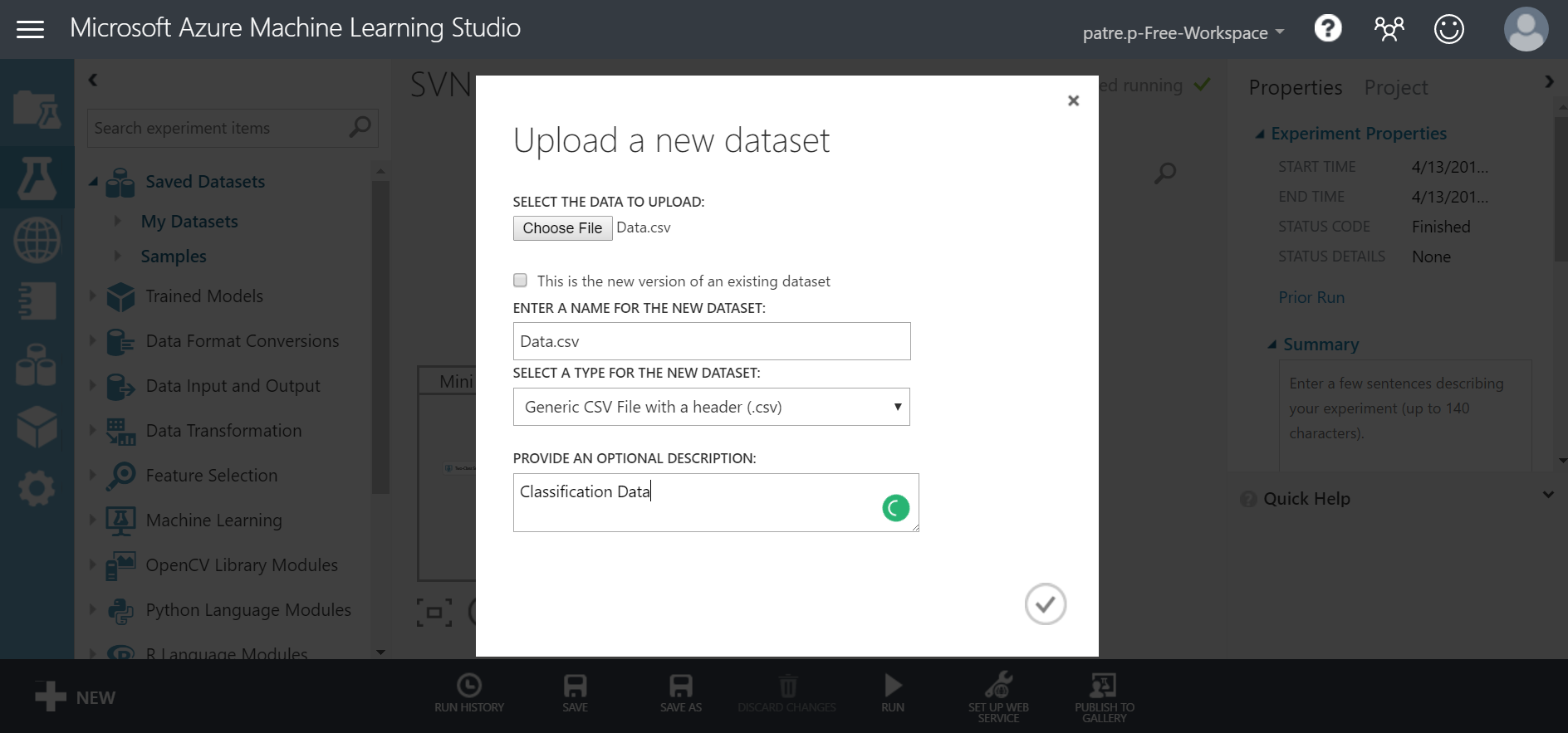


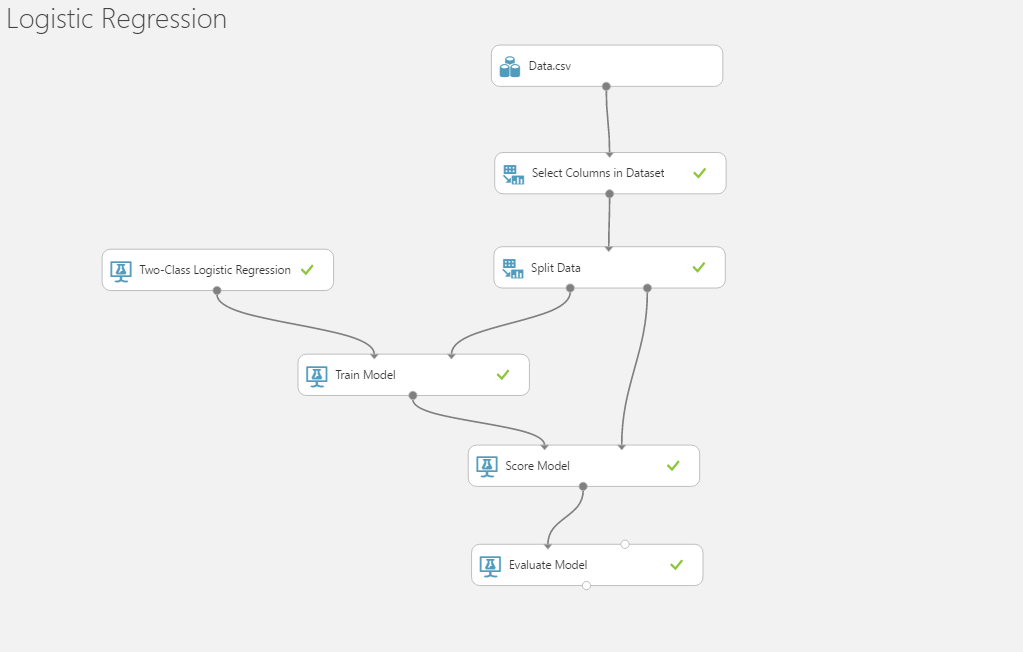
We have also implemented PySpark to speed up the process. Due to time constraints, it was not feasible for us to understand the functions in pyspark for classification algorithm.



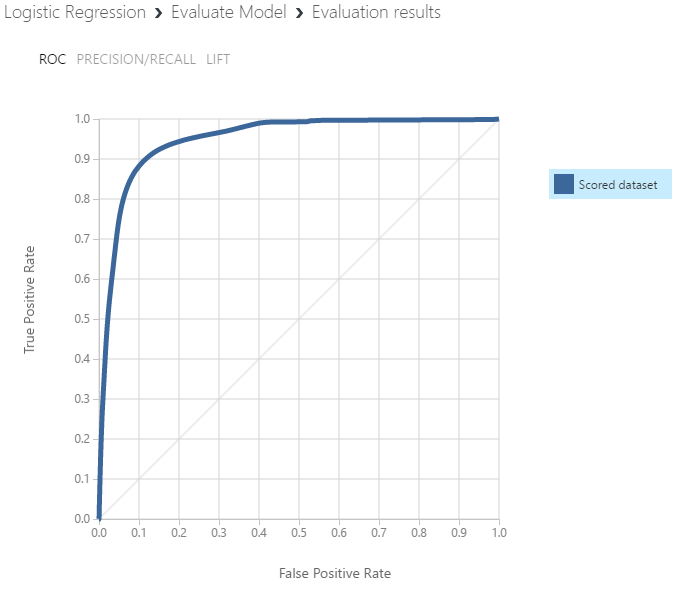
Finally, we decided to work on Azure and decide our best Classification model.

1. Logistic Regression

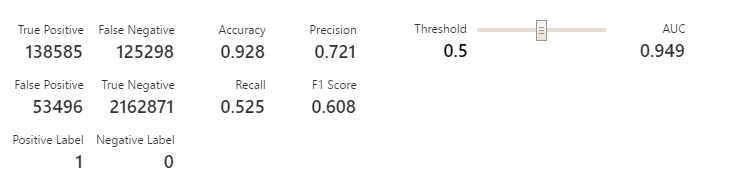




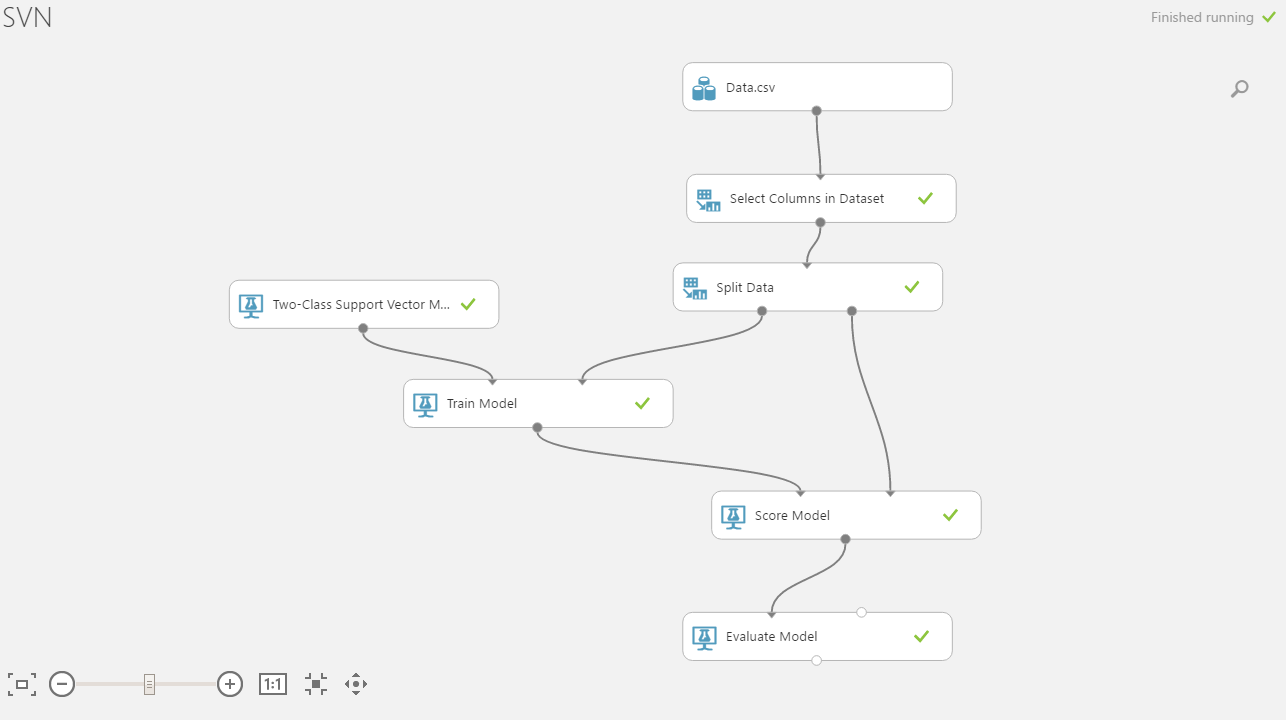
ROC Curve:



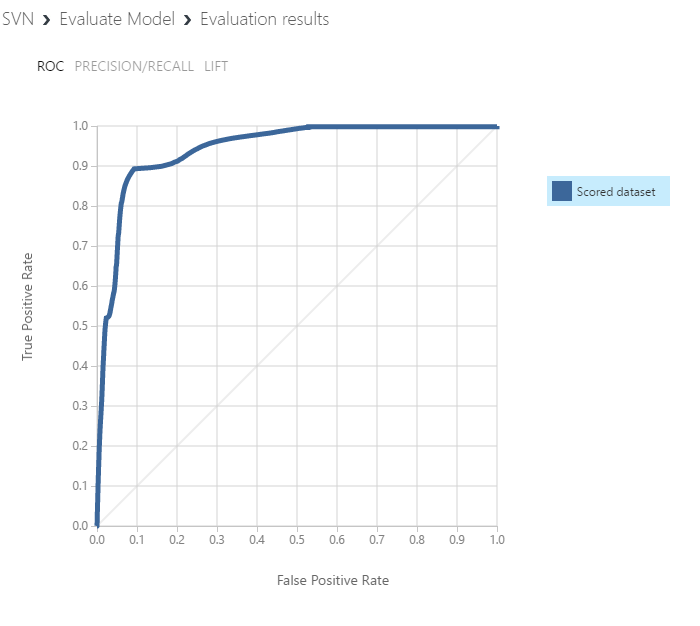
Confusion Matrix:



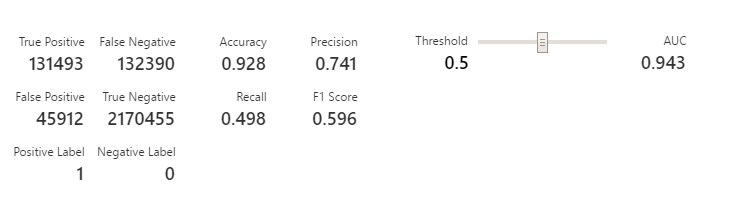
1. SVN Algorithm



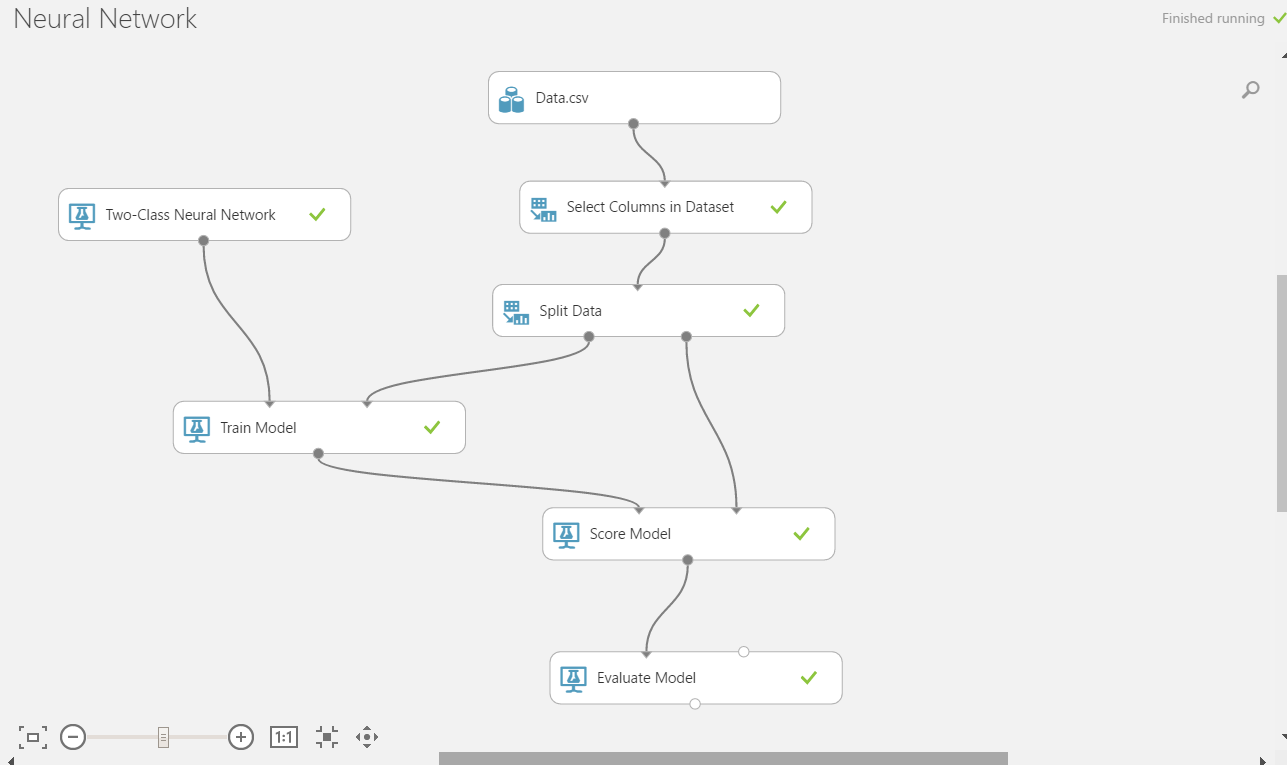
ROC Curve:



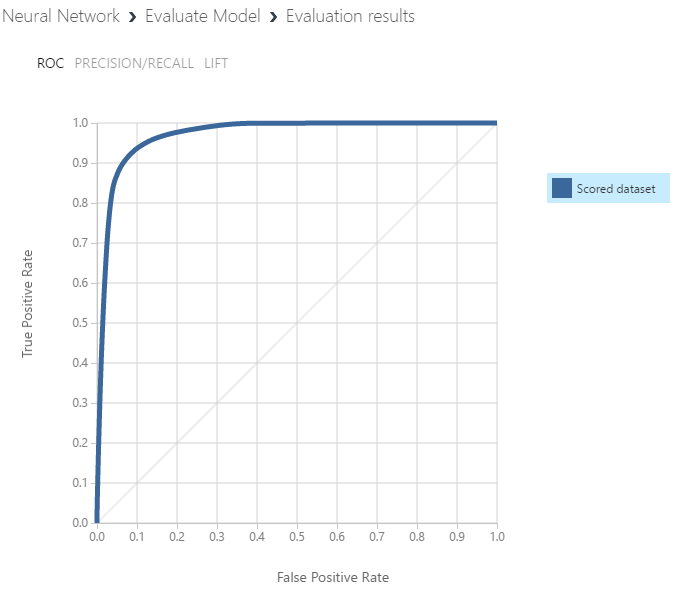
Confusion Matrix:



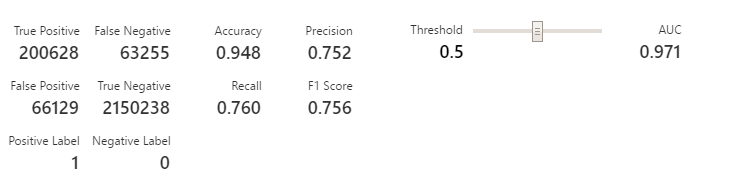
1. Neural Network



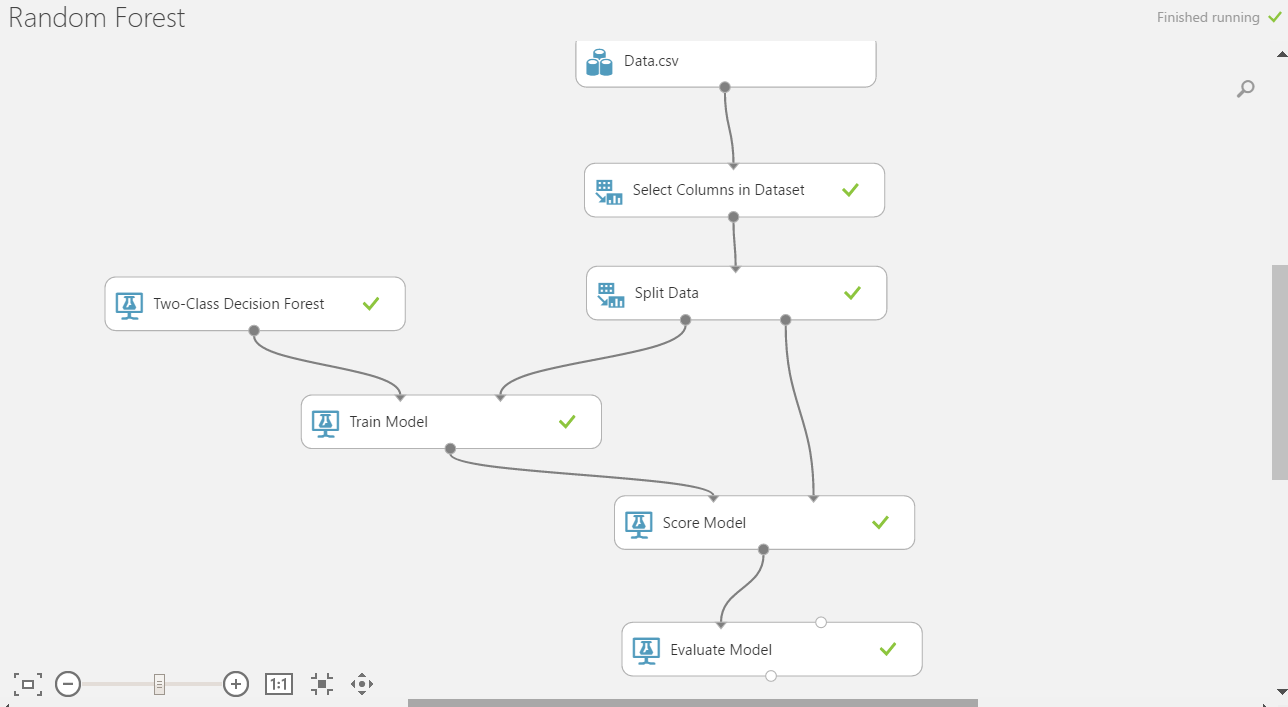
ROC Curve:



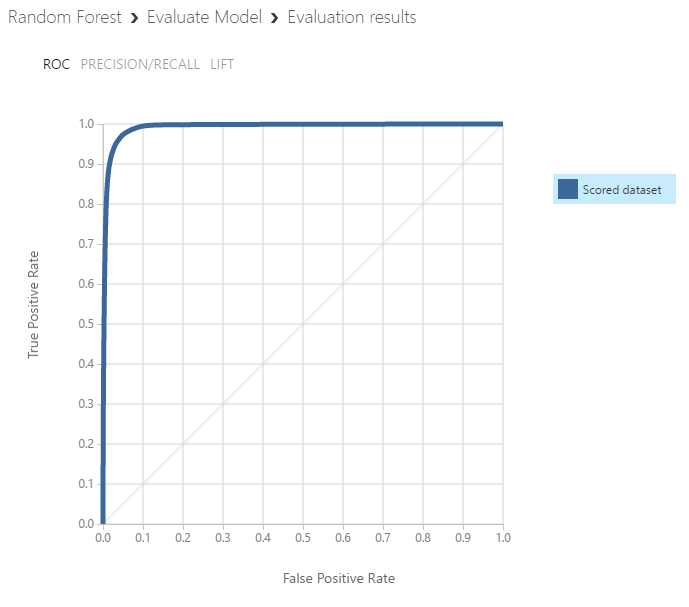
Confusion Matrix:



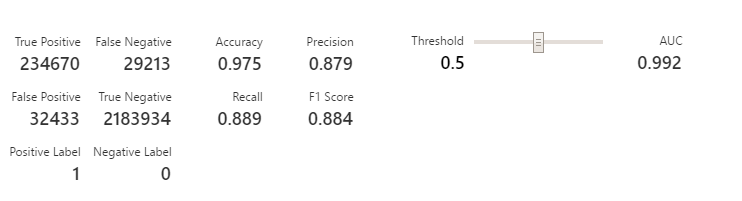
1. Random Forest:



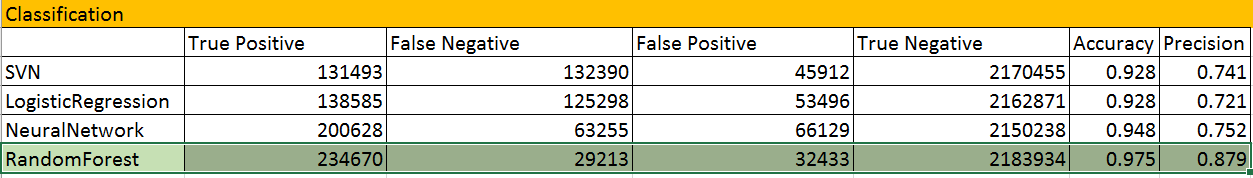
ROC Curve:



Confusion Matrix:



Based on different matrices we have found Random Forest to be Best model for Prediction:



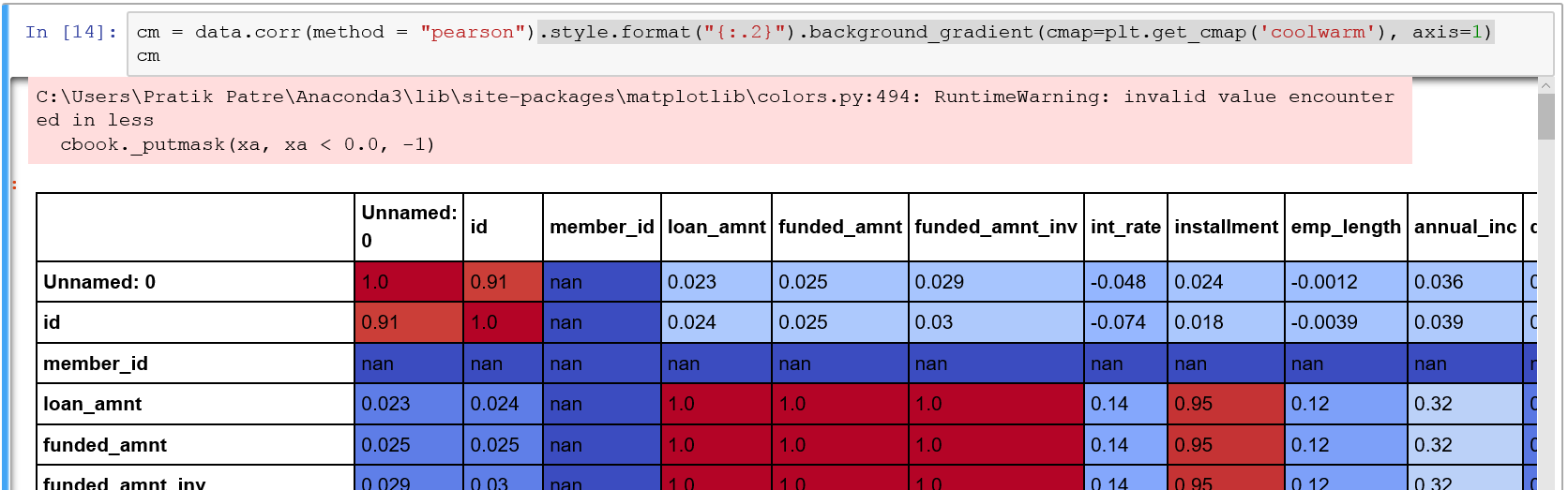
**Part 2: Clustering**

Now that the decision whether to give a loan or not is made, its time to decide the interest rate at which the loan is to be given.

**Feature Selection:**

Before we could proceed further it was important to understand which attributes are of our concern and which are not. We have used Pearson correlation, Spearson Correlation and Kendell correlation to determine various attributes which depend upon interest rate. It is important to determine which parameters change the value of interest rate. Although we could not find the perfect direct and indirect relationship through our correlations, we decided to go with last ten and first ten attributes. Below is the screenshot of the model.

**Pearson Correlation Snapshot:**



Python Code:



Feature Selection:



We have divided the clustering part into three segments:

***Manual Clustering:*** An algorithm was developed where the input of loan details is segregated into different clusters. There are 3 clusters in which the whole data set is divided. The clustering algorithm has 4 features that decide the cluster to be in, namely FICO score, Employment Length, State category, Annual Income. The conditions on which the algorithm is created is as follows:

First scores are given depending on the different brackets the value falls in.

1) FICO score

    - FICO score between 300 and 449 - 1

    - FICO score between 450 and 649 - 2

    - FICO score between 650 and above  -3

2) Employment Length

    - Employment Length between 0 and 4  - 1

    - Employment Length between 4 and 8  - 2

    - Employment Length between 8 and 10- 3

3) Annual Income

    - Annual income between 0 and 40000 -  1

    - Annual income between 40000 and 96000 -  2

    - Annual income between 96000 and above  - 3

4) State category it falls in

    - State category 1 - 1

    - State category 2 - 2

    - State category 3 - 3

Choosing the state category:

The economies of all 50 US states and Washington, DC, from worst to best

The US has an enormous economy, and that economy is the sum of the economies of 50 states and the District of Columbia.

We ranked the economies of all the states and DC on seven measures: unemployment rates; GDP per capita; average weekly wages; recent growth rates for nonfarm payroll jobs; GDP; house prices; and wages. For more on our methodology, click [here](http://www.businessinsider.com/state-economy-ranking-sources-and-methods-2016-1).

In addition to looking at some of the economic metrics that went into the ranking, we identified a [symbol for each state and DC, according to Wikipedia](https://en.wikipedia.org/wiki/Lists_of_United_States_state_symbols).

<http://www.businessinsider.com/state-economy-ranking-q4-2015-2016-1/#28-georgia-24>

The total score is calculated by adding all the scores. This score determines the cluster it will fall in. The way of clustering is determined in such a way that low score values are put in together and the high score value is in together. The range of the value is from 3 - 12. The clusters are formed as values {4,5,6} - cluster 1, {7,8,9} - cluster 2, {10,11,12} - cluster 3

Then upon this condition, the file is segregated into 3 parts.

The prediction algorithms are run on them and the best model is chosen for all the cluster.

ManualClustering Python Code:



ManualClustering Screenshot:

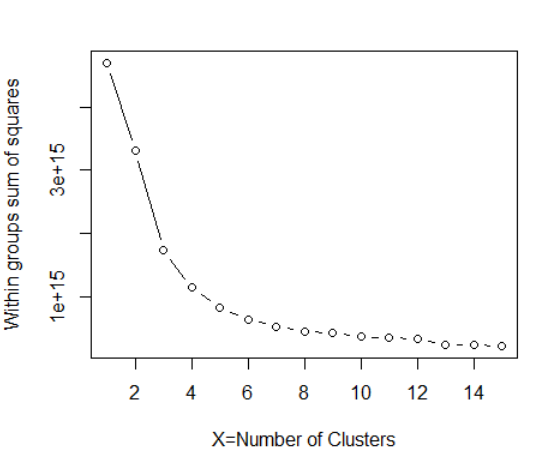


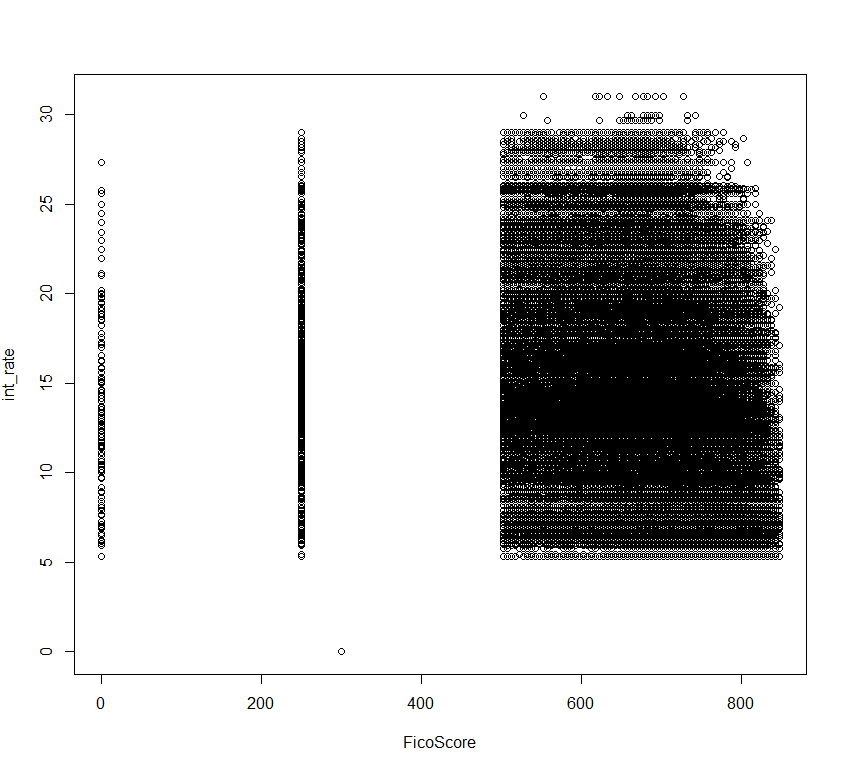
***Clustering Algorithm***: The data is clustered using two different clustering algorithm and based on the number of outliers, the best model is chosen. We have implemented Hierarchical clustering algorithm and K-means algorithm.

The best model which performed well is K means algorithm.

K-Means Algorithm:

Based on the bend graph we decided the value for K to be 5.

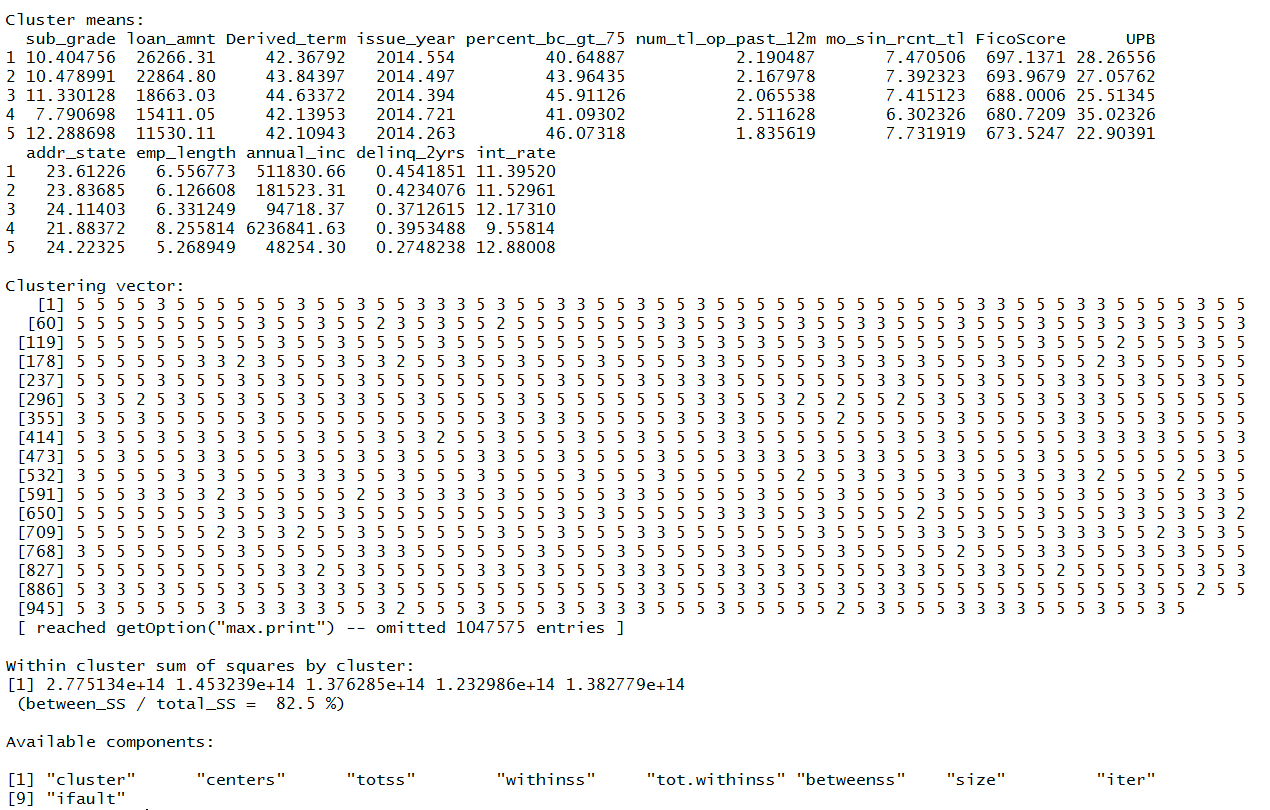




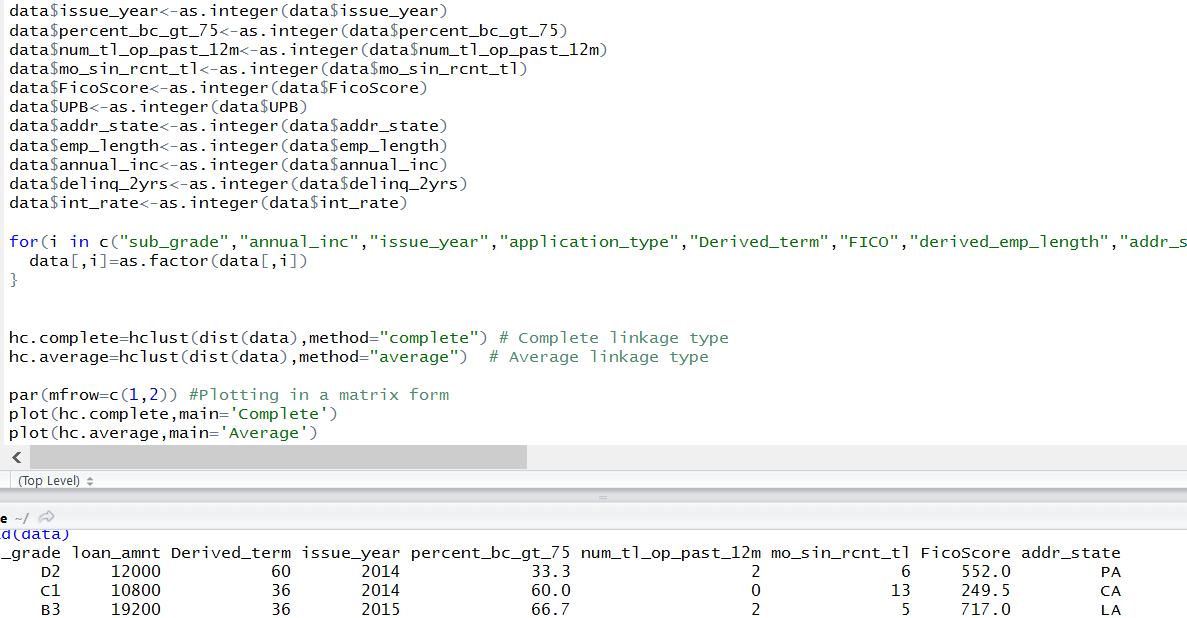
R code for K-Means Algorithm:



Kmeans Output:



Hierarchical Algorithm:

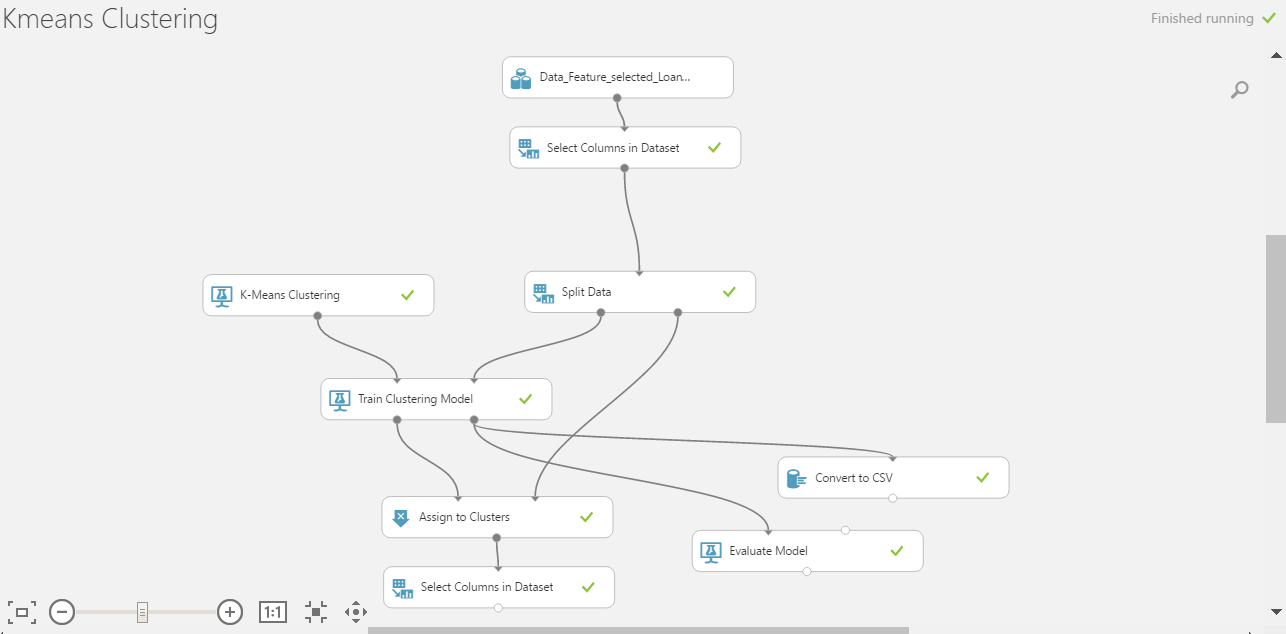


R code:

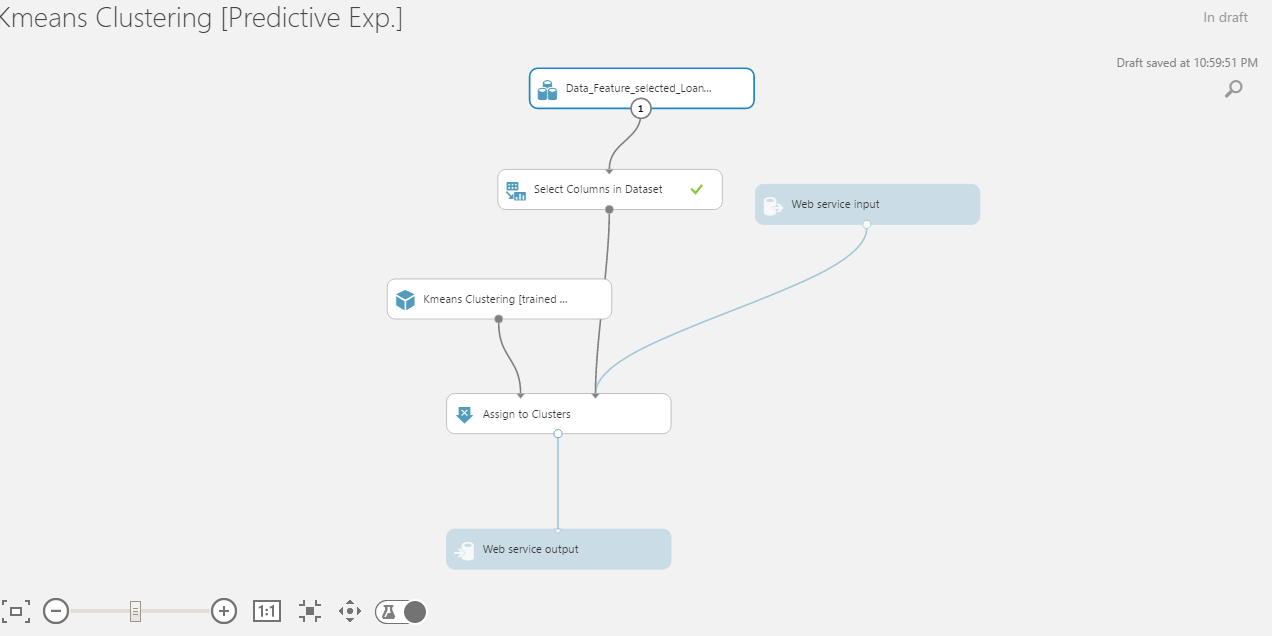


***No Cluster:*** The data is not segregated into any clusters and is fed directly to prediction model for predicting Interest rates.

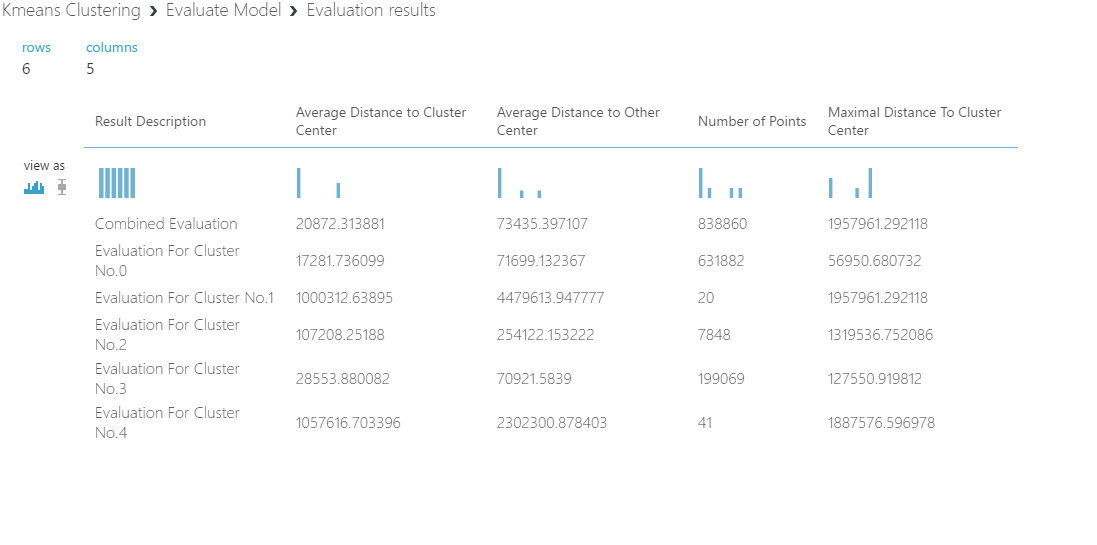
We deployed the Kmeans model in Azure and created a REST API.



Kmeans Web Deployed:



Kmeans Azure output:



**Part 3: Prediction**

Now that the data is segregated, it is time to decide the interest rates for each of the clusters. Prediction Algorithms are run on each of the clusters. We have 8 clusters which leads to running 24 prediction algorithms.

The algorithms which we implemented are:

1. Linear Regression
2. Random Forest
3. Knn
4. Neural Network

The algorithms took longer time to run in R so we planned to implement it in Azure for each of the clusters and compute the summary matrices.

Knn R Code:



Linear Regression R Code:



Neural Network R Code:



Random Forest R Code:



***We have run the prediction models in Azure for :***

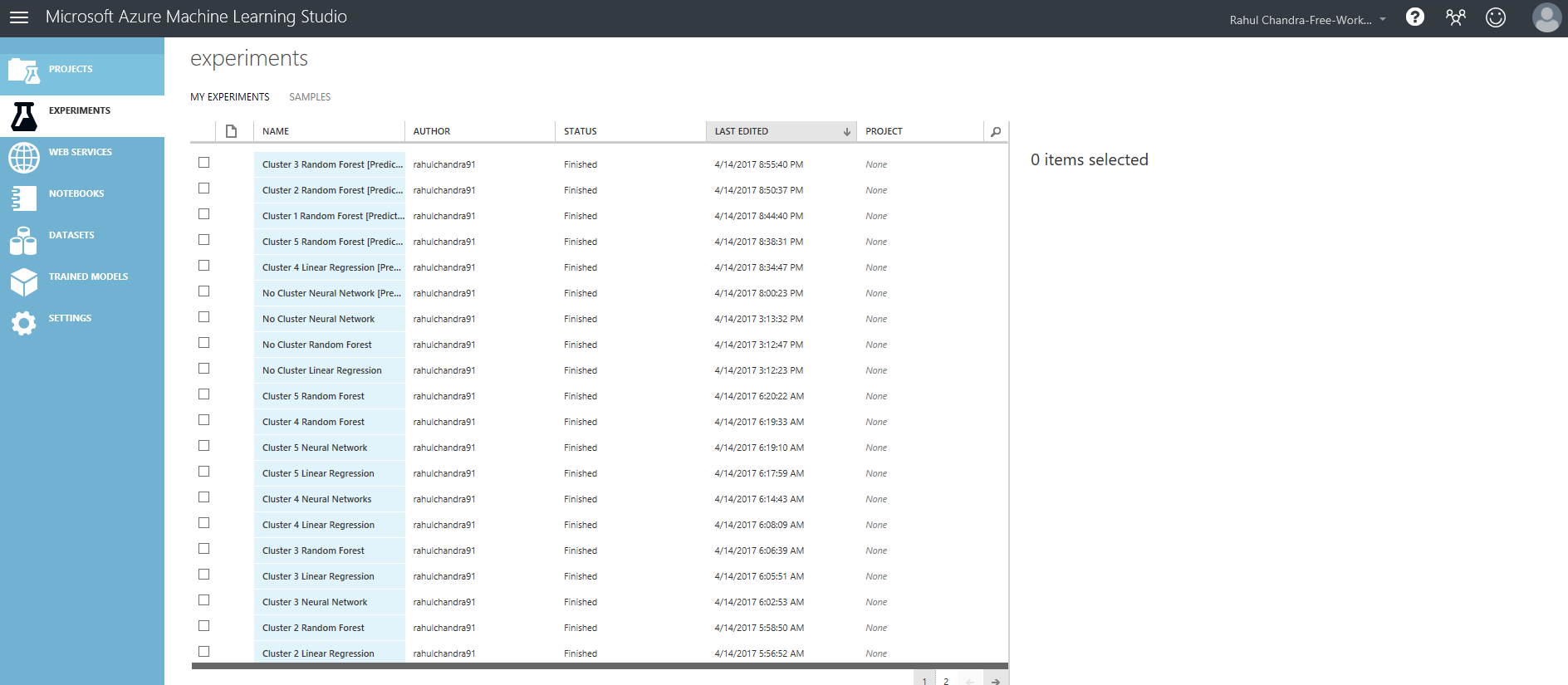
***9 for Manual Clusters***

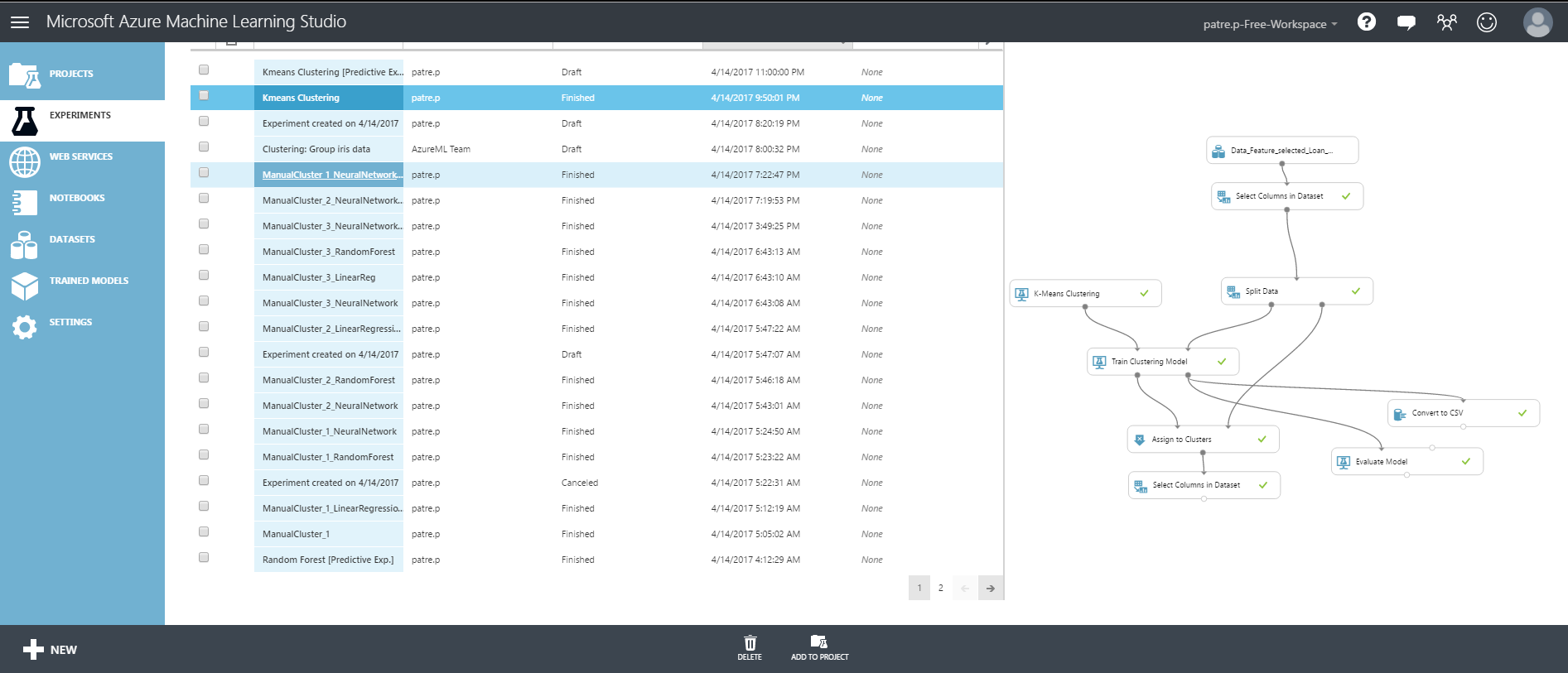
***15 for Clustering Algorithm***

***3 for No Cluster***

***Total of 27 Prediction Algorithms***

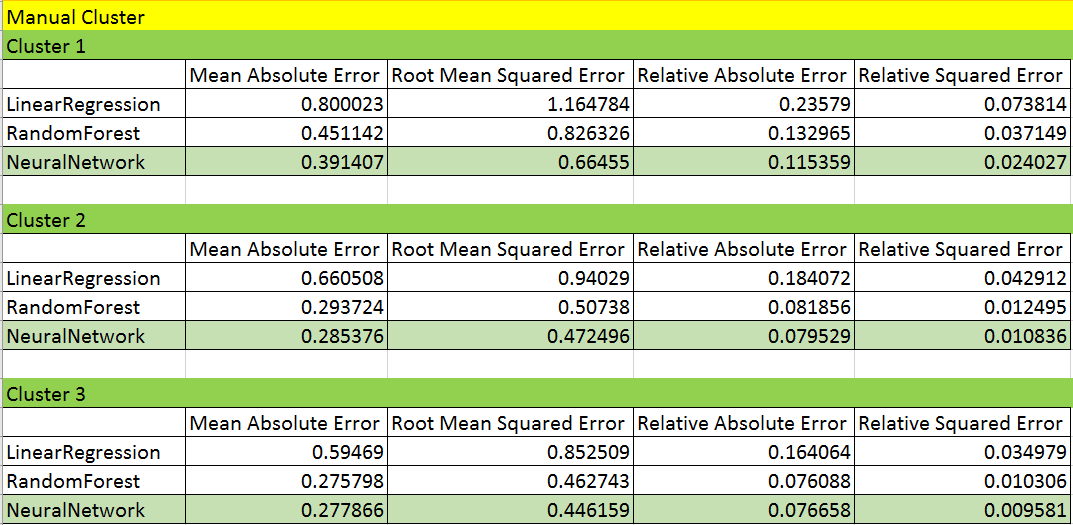




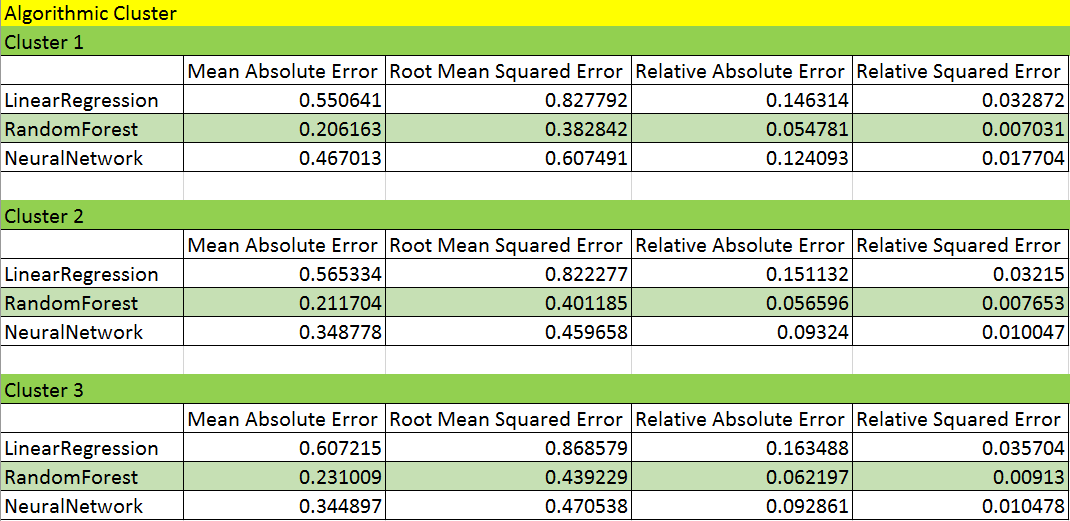


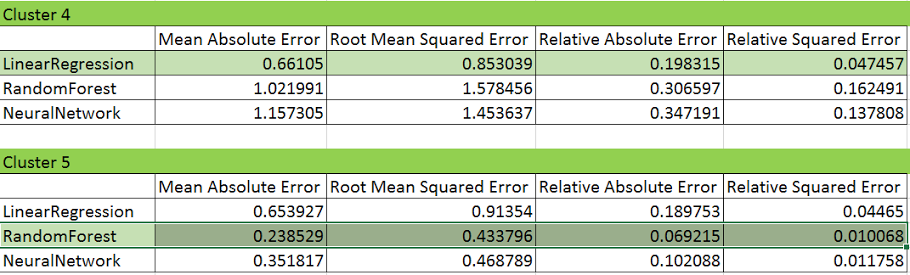


Summary Metrics for Manual Cluster:

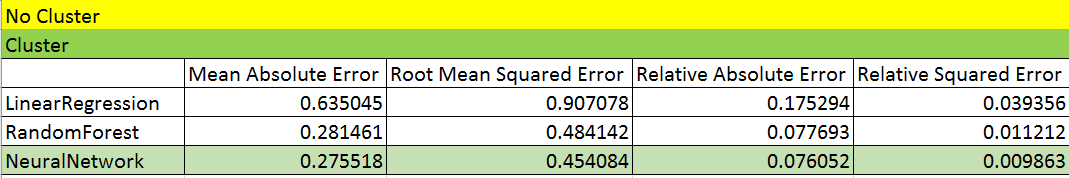


Summary Metrics for Clustering Algorithm:





Summary Metrics for No Cluster:



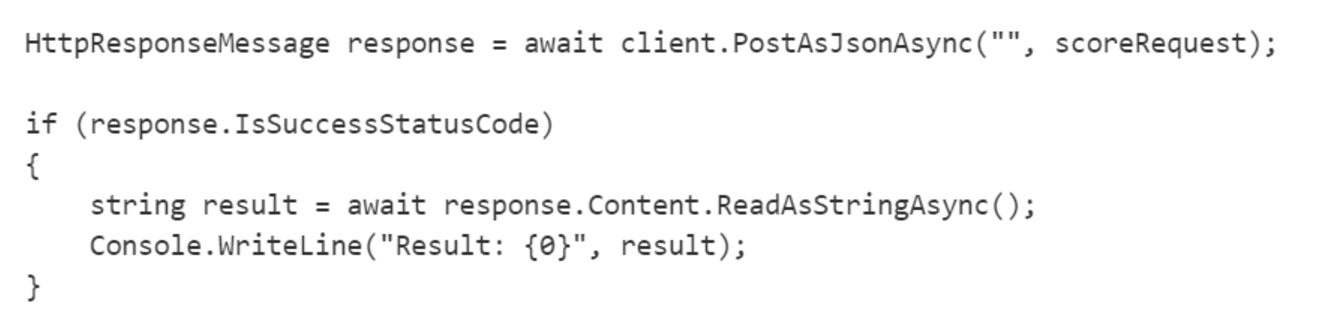
The models highlighted in light green are the best models for the respective clusters which are further deployed for creating a REST API.

Summary Metrics CSV:

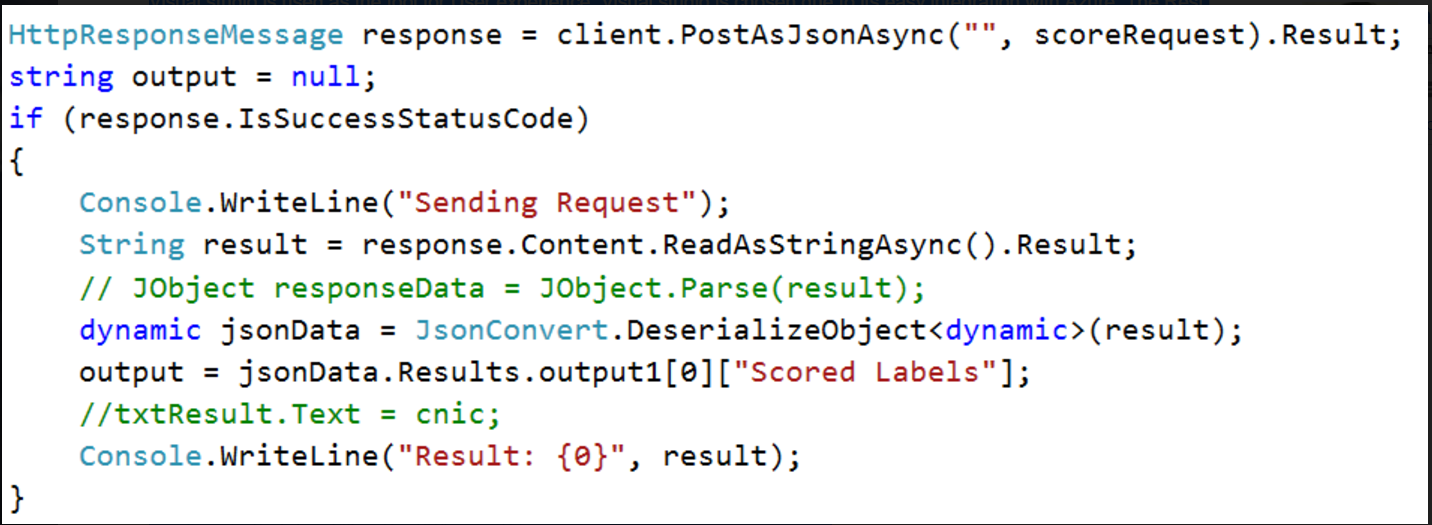


**Part 4: Deployement**

Visual studio is used as the tool for User experience. Visual studio is chosen due to its easy integration with Azure. The Rest API generated for the models where easy to include in the C# code. But there where some problem faced in the integration. The code snippet provided in the Azure was calling the function Asynchronously. Due to which the function was not waiting for response. Below is the screen shot of the code.



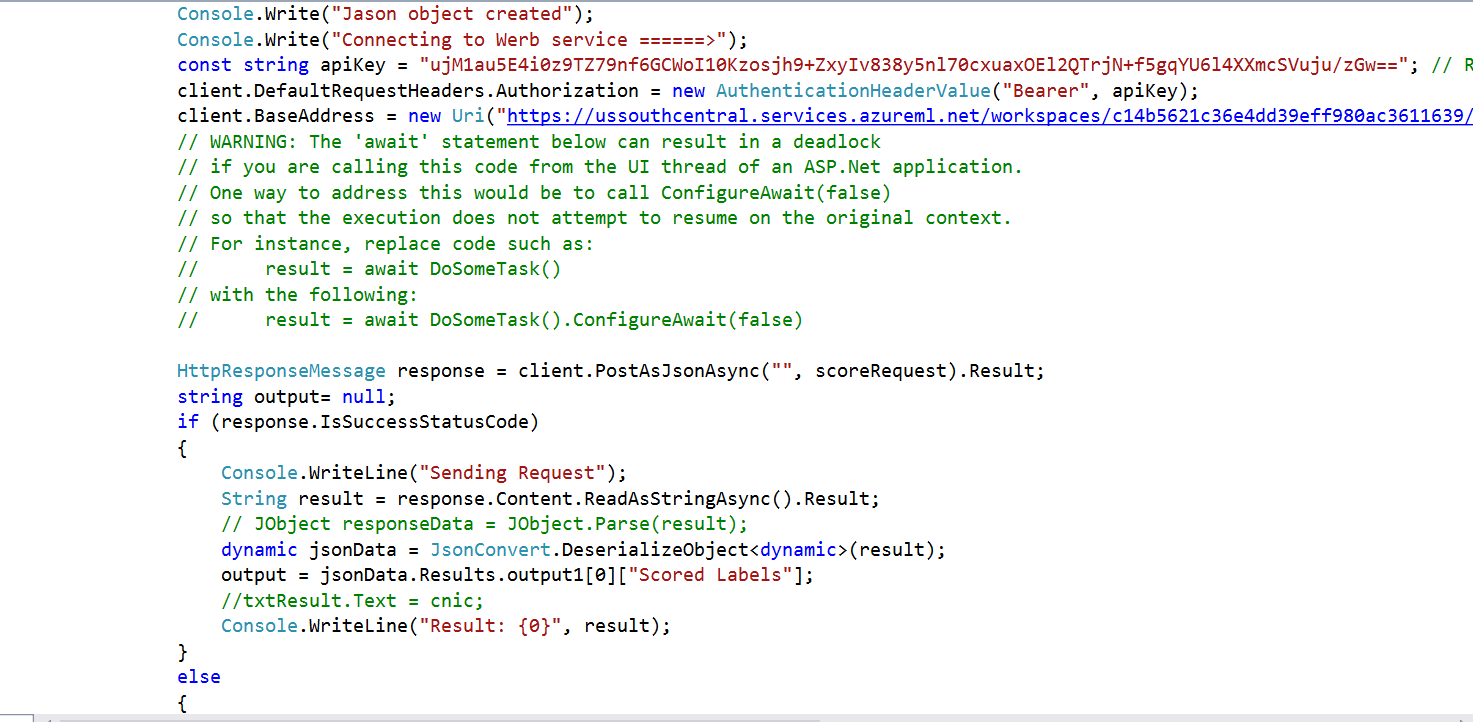
This issue was resolved by capturing the data in .result instead of ReadAsStringAsync().



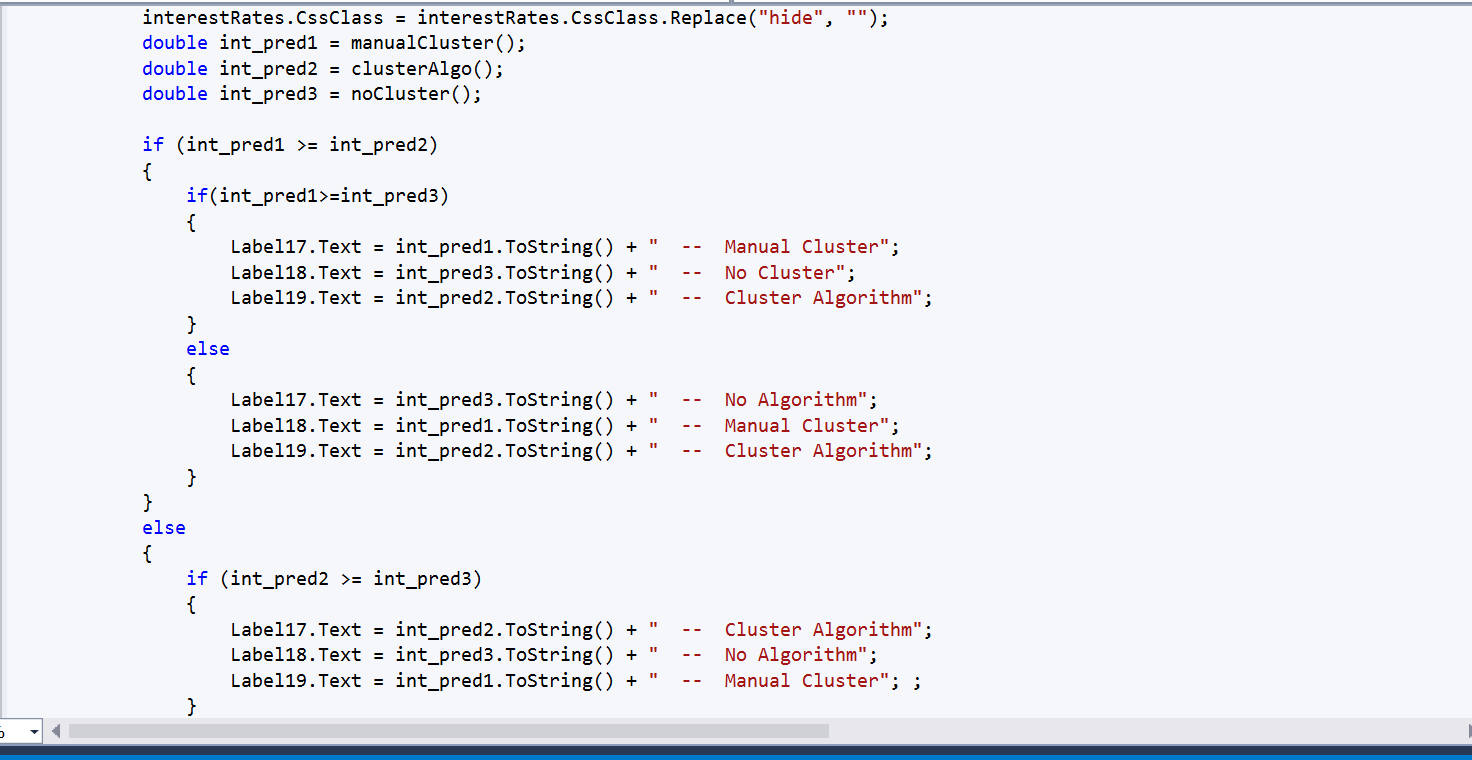
**Please find below the screenshots for Deployment:**

C# Code for Deployment:

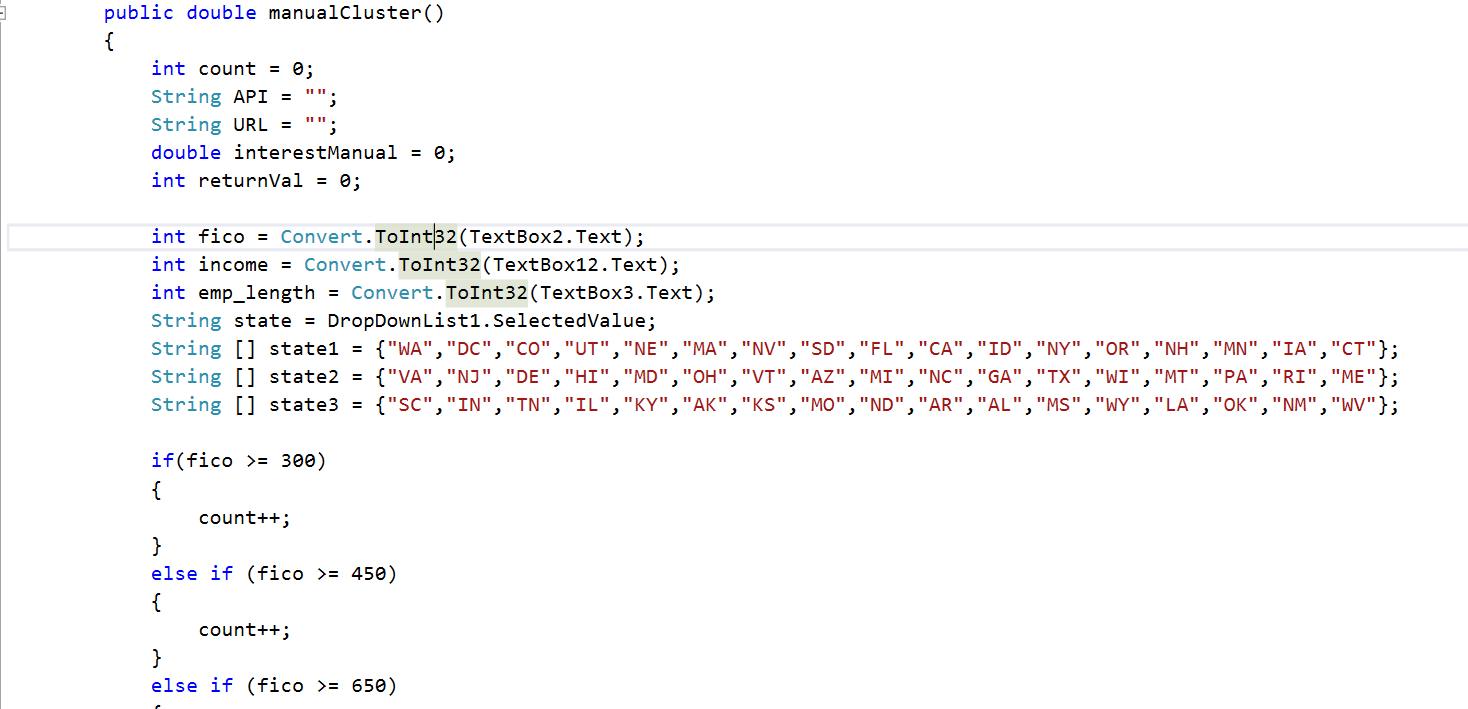


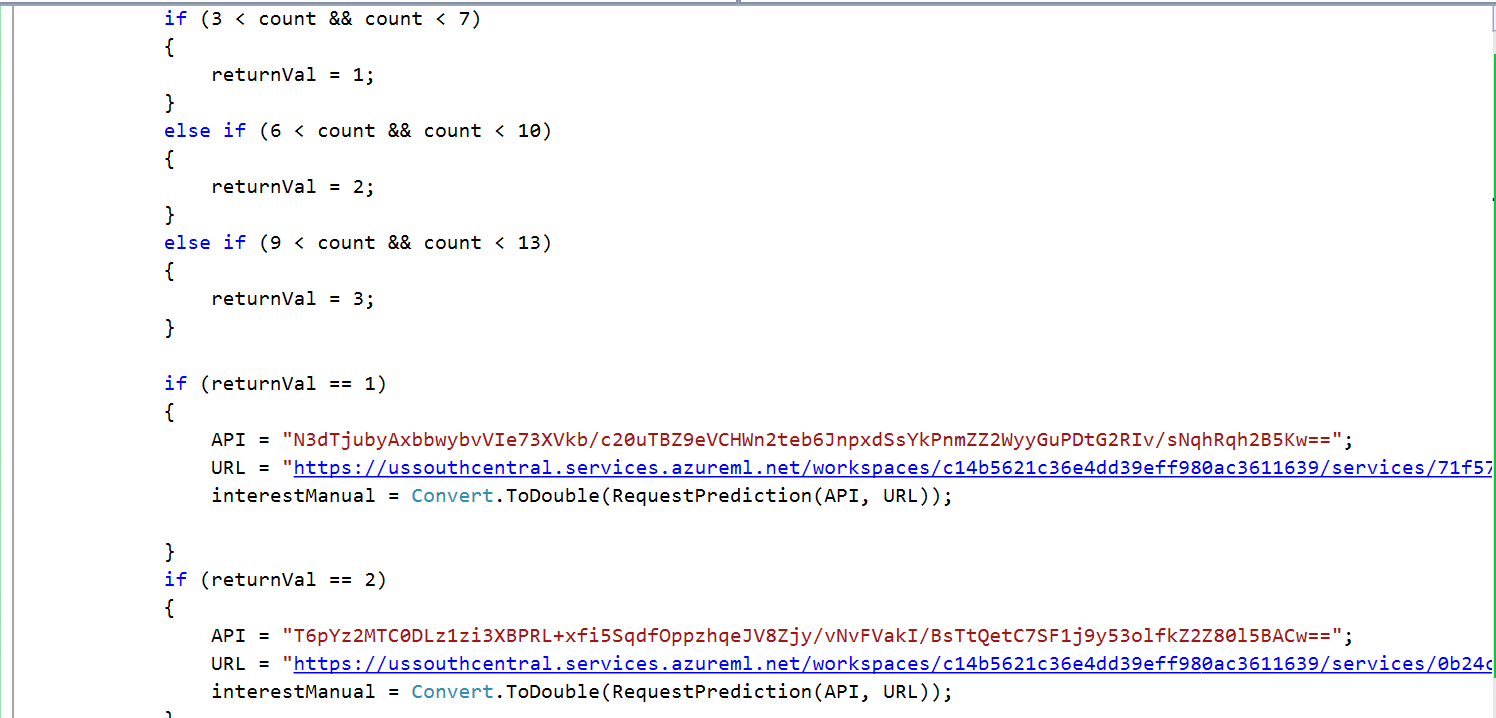


C# code for choosing the best Interest Rate:



C# code for Manual Clustering:

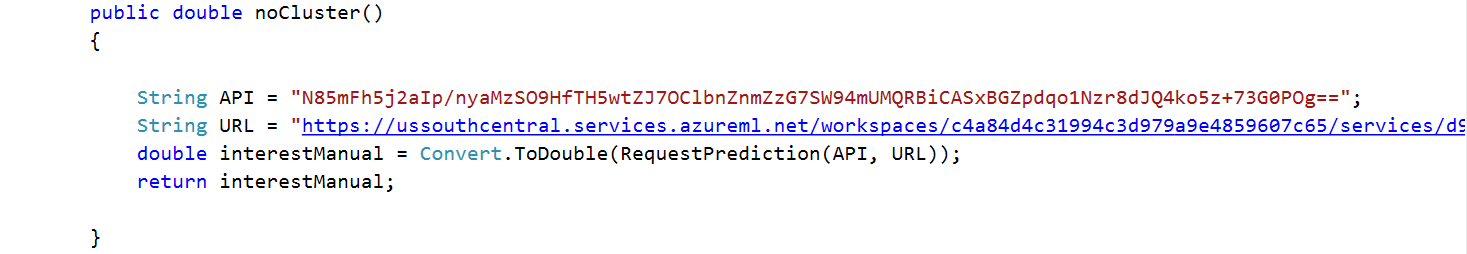




C# code for Clustering Algorithm:



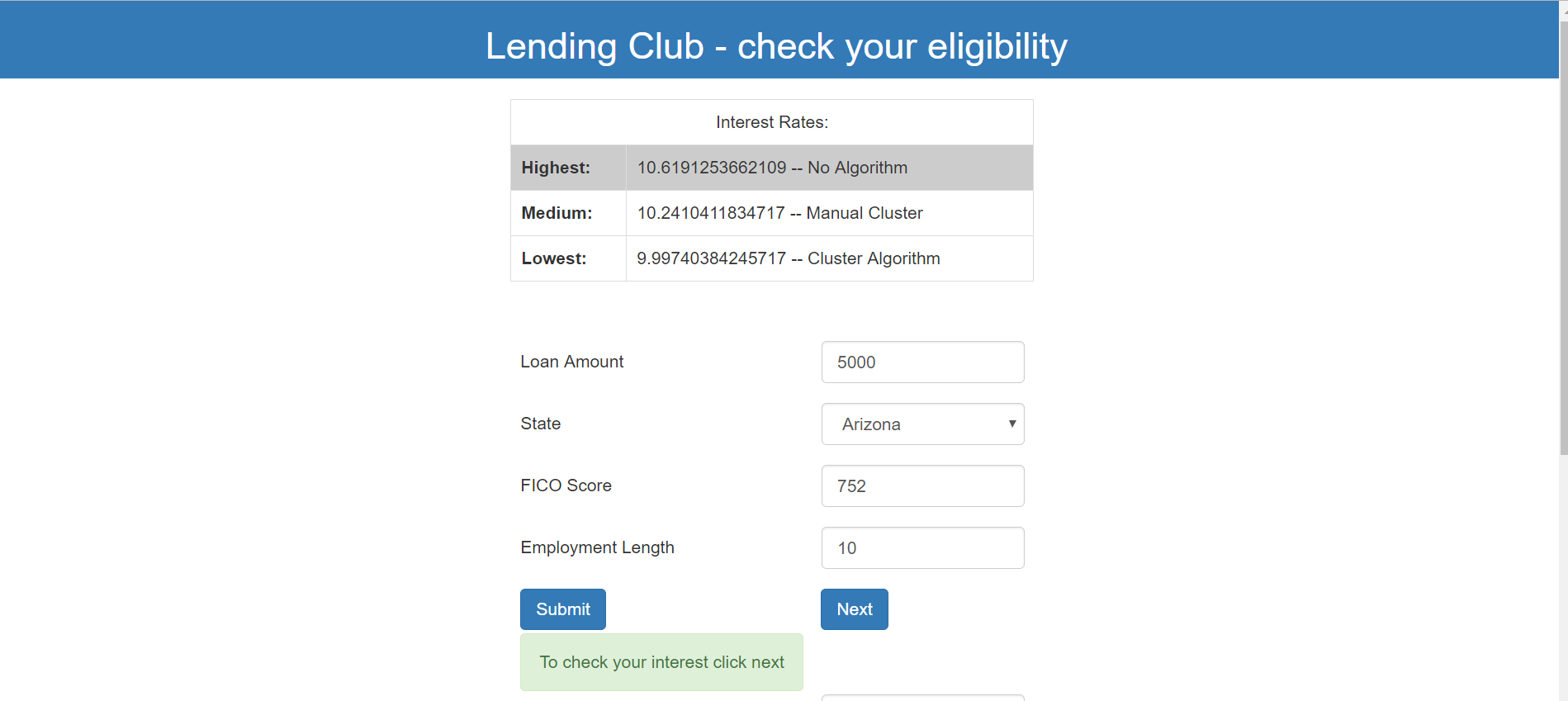
C# code for No clustering:



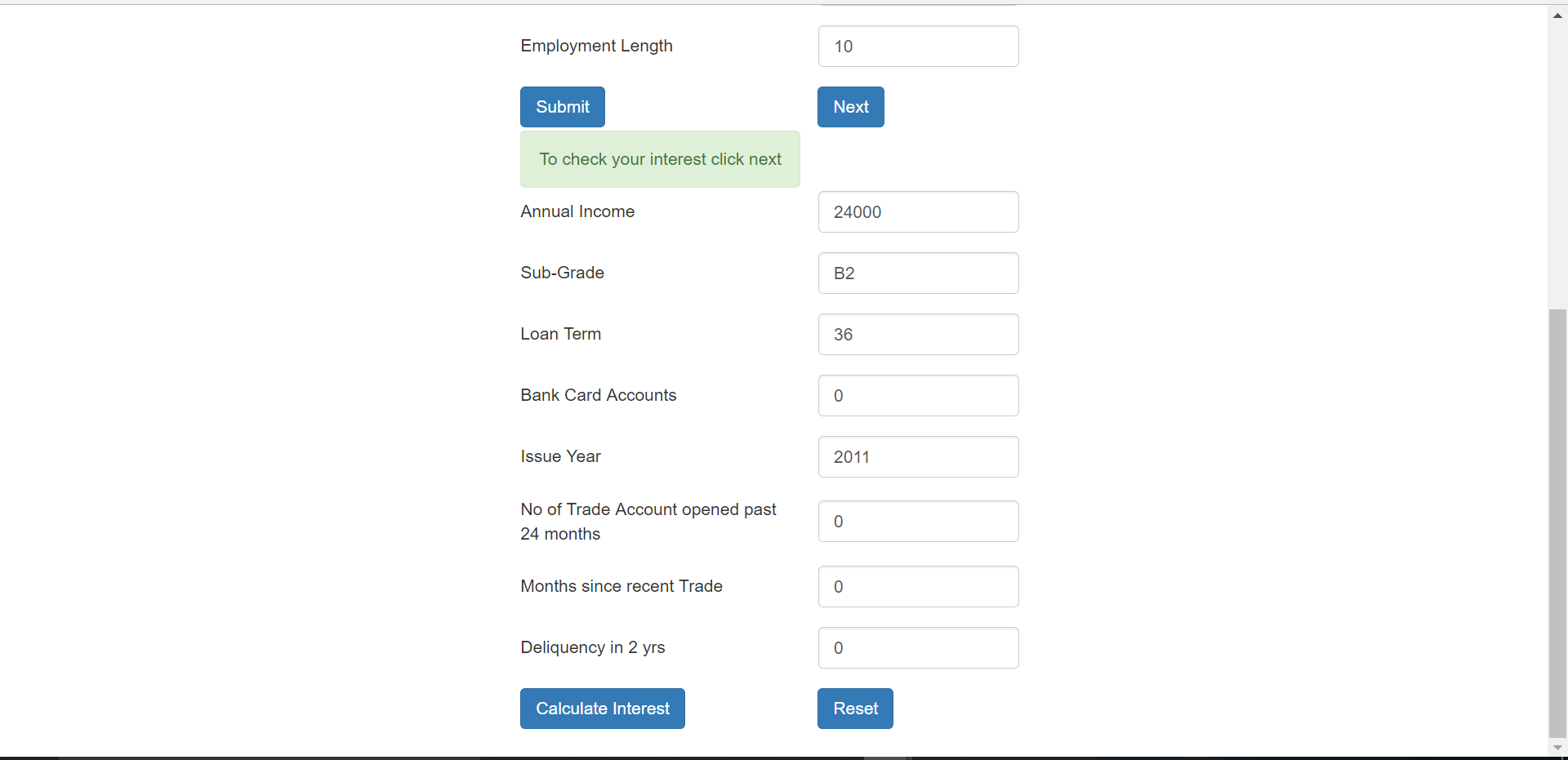
Backend HTML:

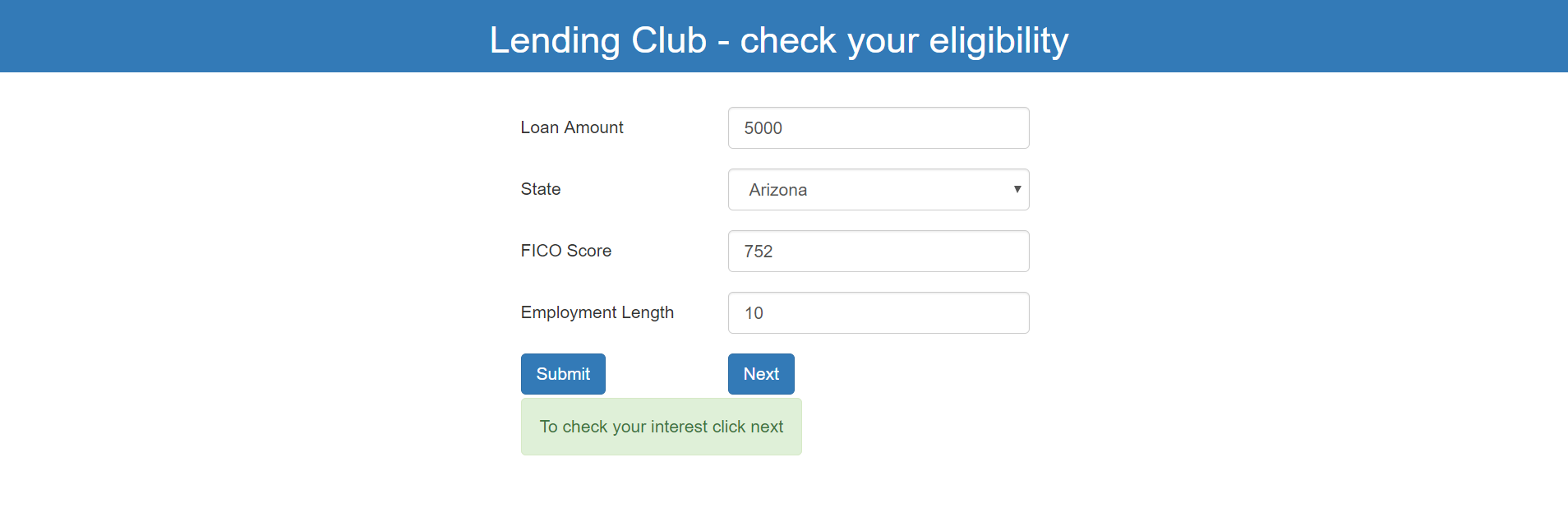


Result:



Approved Screenshot:





Rejected Screenshot:

