Contents

[Problem Statement 3](#_Toc498906270)

[Data Wrangling and Exploratory Data Analysis 3](#_Toc498906271)

[Data Download 3](#_Toc498906272)

[LUIGI Pipe Line 5](#_Toc498906273)

[Variable Selection 8](#_Toc498906274)

[Exploratory Data Analysis 10](#_Toc498906275)

[Exploratory Data Analysis for Loan Data 10](#_Toc498906276)

[Exploratory Data Analysis for Declined Loan Data 14](#_Toc498906277)

[Analysis of Approval of Loans 15](#_Toc498906278)

[Classification 18](#_Toc498906279)

[Logistic Regression 19](#_Toc498906280)

[Neural Networks 20](#_Toc498906281)

[Random Forest 22](#_Toc498906282)

[Deployment on Azure 24](#_Toc498906283)

[Summarization of Results for Classification 25](#_Toc498906284)

[CLUSTERING 26](#_Toc498906285)

[Manual Clustering 26](#_Toc498906286)

[Clustering using K Means Algorithm 27](#_Toc498906287)

[TSNE 28](#_Toc498906288)

[PREDICTION 29](#_Toc498906289)

[Manual Clustering 29](#_Toc498906290)

[Cluster 1 29](#_Toc498906291)

[Cluster 2 30](#_Toc498906292)

[Cluster 3 31](#_Toc498906293)

[Cluster 4 32](#_Toc498906294)

[Manual Clustering Prediction and Web Service Deployment on Azure 33](#_Toc498906295)

[Consuming Manual Clustering WebService 38](#_Toc498906296)

[Final Results for Manual Clustering 39](#_Toc498906297)

[Prediction using Clustering Algorithm K Means 40](#_Toc498906298)

[Cluster 0 40](#_Toc498906299)

[Cluster 1 41](#_Toc498906300)

[Cluster 2 42](#_Toc498906301)

[Cluster 3 43](#_Toc498906302)

[Prediction and Deployment using KMeans Clustering in Azure 44](#_Toc498906303)

[Consuming the web service using Azure 48](#_Toc498906304)

[Summary of Results using KMeans Clustering 49](#_Toc498906305)

[Prediction with No Clustering 50](#_Toc498906306)

[Model Deployment using Azure 52](#_Toc498906307)

[Consuming web service using Azure 53](#_Toc498906308)

[Summary of Results for data with No Clusters 54](#_Toc498906309)

[DEPLOYMENT 55](#_Toc498906310)

# Problem Statement

You are working at a bank and you are considering investing in Lending club. Since there are no standard models, you are expected to build prediction models that will help you predict the interest rates based on various parameters users would input.

# Data Wrangling and Exploratory Data Analysis

## Data Download

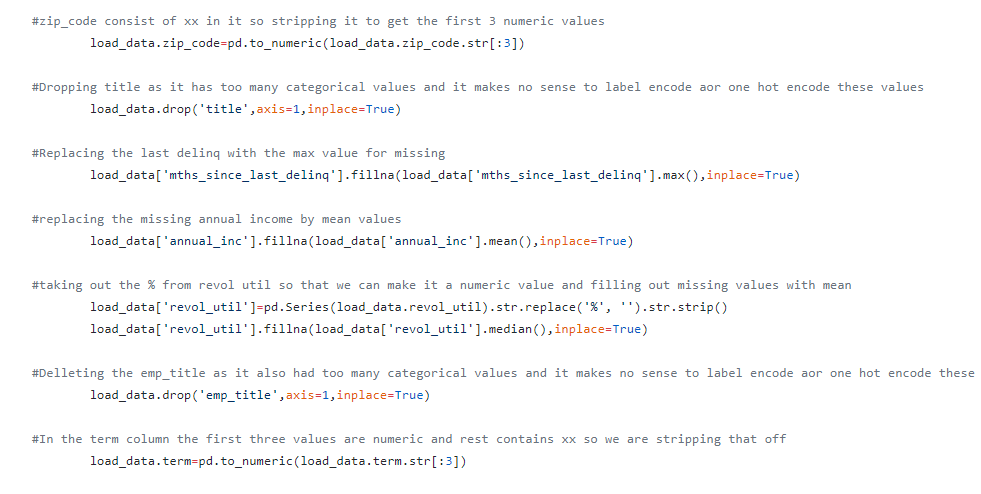
1. We first signed in to Lending club website to collect the full data.

<https://www.lendingclub.com/info/download-data.action>

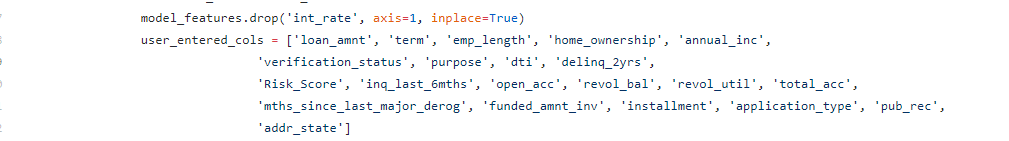
We programmatically downloaded the data as follows



1. Followed the same procedure for Declined Loan Data
2. We did Missing Data Analysis



1. Followed by Feature Engineering. Our aim was to keep all fields that could be defined by the user



1. Next we used LUIGI to automate this entire process.

## LUIGI Pipe Line

**Commands to run**

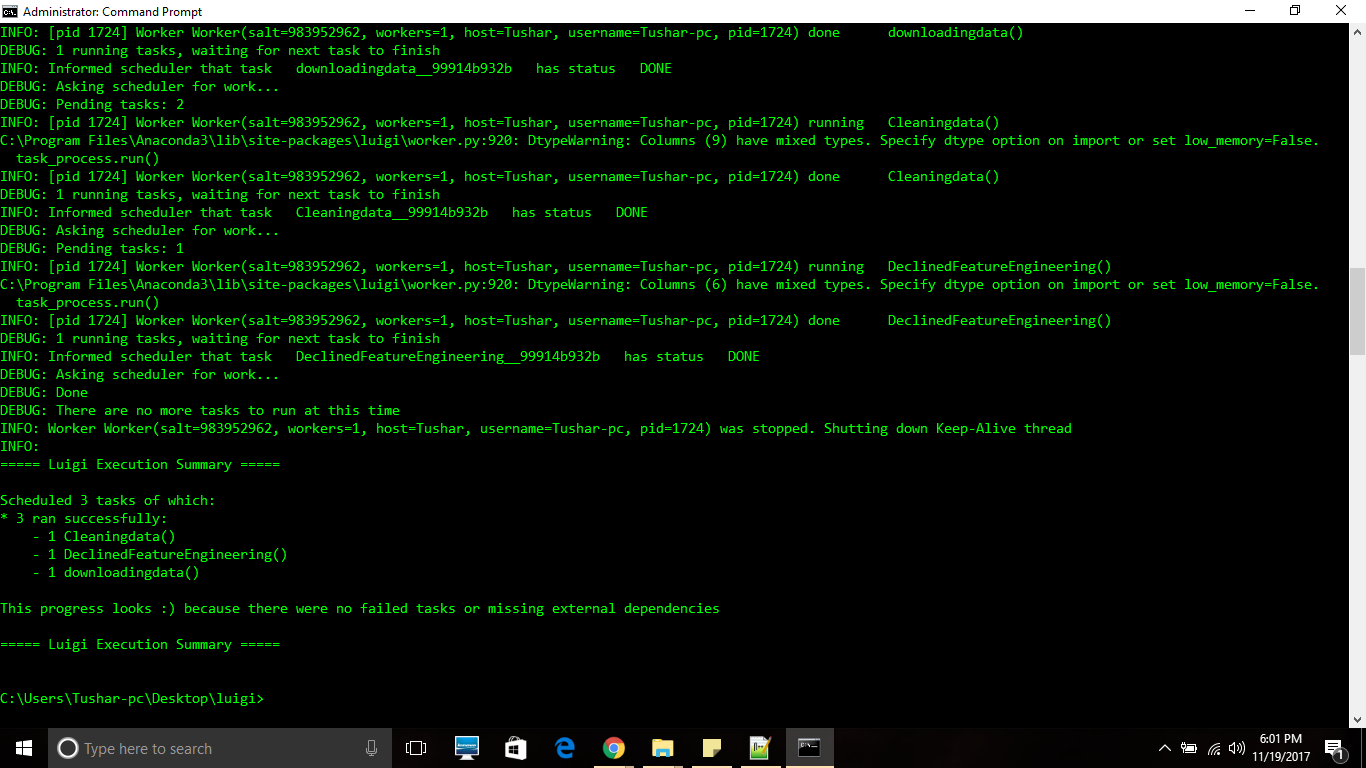
#### Command to run the LUIGI pipeline for Accepted Loan Data

python loandata.py --local-scheduler FeatureEngineering

**Command to run the LUIGI pipeline for Declined Loan Data**

python declined.py --local-scheduler DeclinedFeatureEngineering

**Luigi Pipeline for Declined Loan data Set**



**Luigi Details**

We have 3 Classes

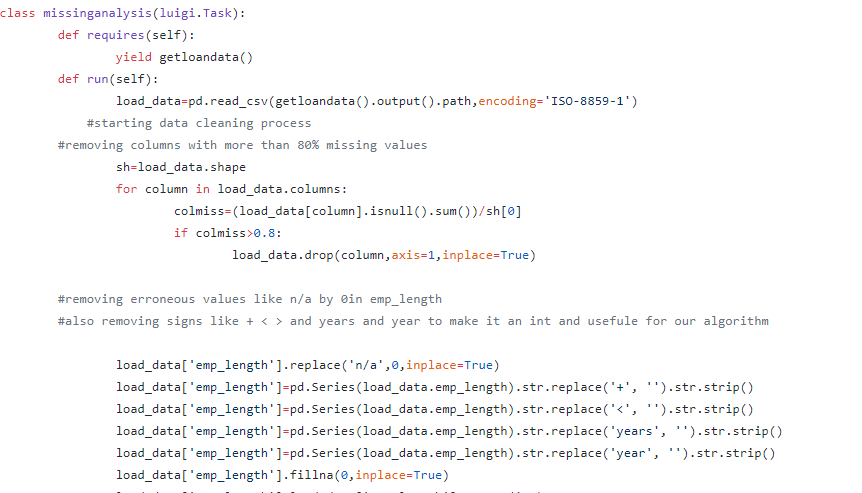
1. Fetching the Loan Data
2. Missing data Analysis
3. Feature Engineering

Fetching Loan Data : Here we use beautiful soup to extract the urls and then download the files and merge them



Similarly we did handled missing values and anomalies.

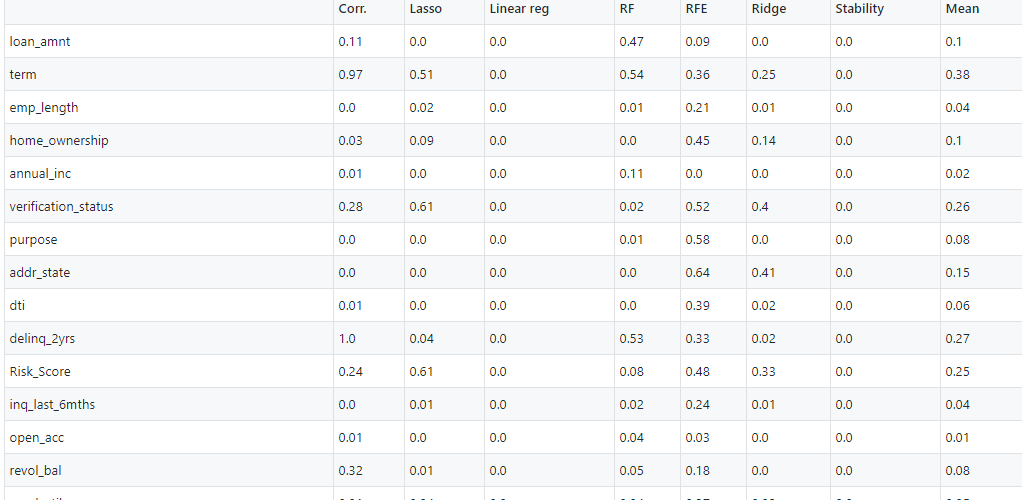
Our cutoff was to drop the columns with more than 70% missing data in them and rest we used different imputing strategies

****

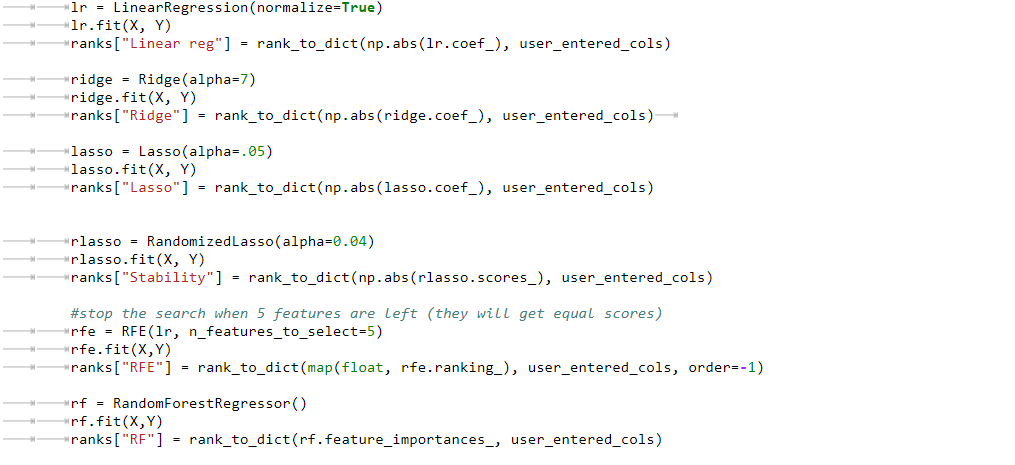
After that we did Feature Engineering and stability of the features and generate a csv file with the metrics



## Variable Selection

****

**Code Snippet**



The Figure above shows the Featureselection.csv generated after feature engineering

The First column named corr defines the linear correlation between feature and the label

The second and third columns represent the ranking and coeff of lasso and linear regression

The RF column represents the importance given to the feature by randomforest regressor

The RFE column represents the results after doing recursive feature engineering

The Ridge column specify the importance given after doing ridge regression

The Stability column is the result of applying RandomizedLasso Regression and checks how many times a feature appeared in every model and then ranks it

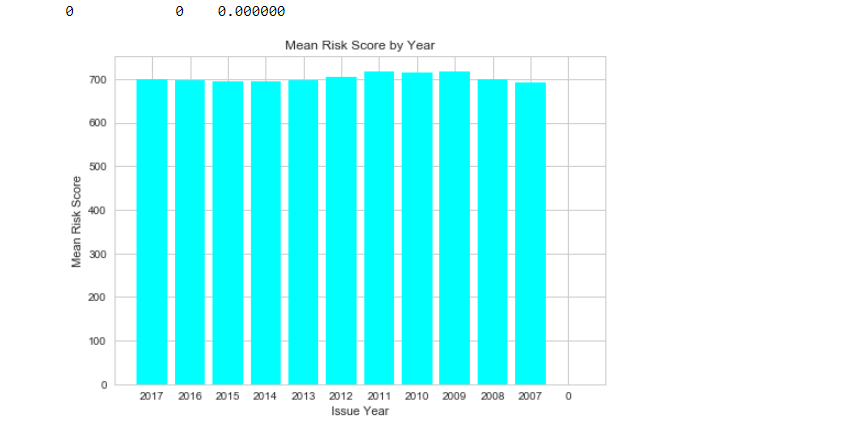
# Exploratory Data Analysis

The Loan Data has 145 features and Declined Loan Data had 8 Features. We have also combined the data for Loan Data and the Declined Data and summarized Approval Statistics for the same. The Summarization is done in length and is present in SummaryMetrics.ipynb. A few of the snapshots are present below.

## Exploratory Data Analysis for Loan Data

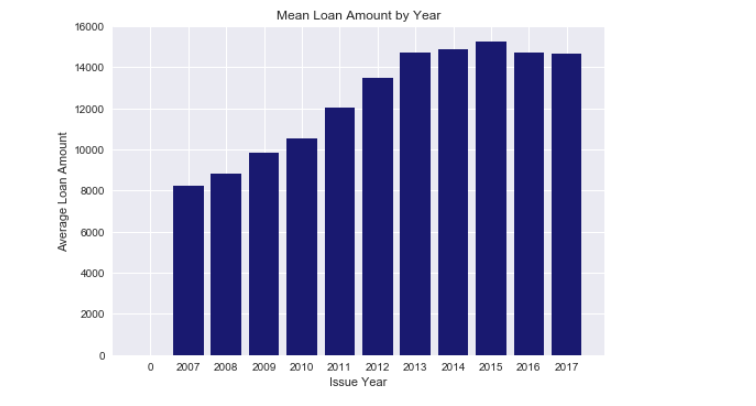
Mean Risk Score by Interest Rate

We note that the mean risk score does not change significantly



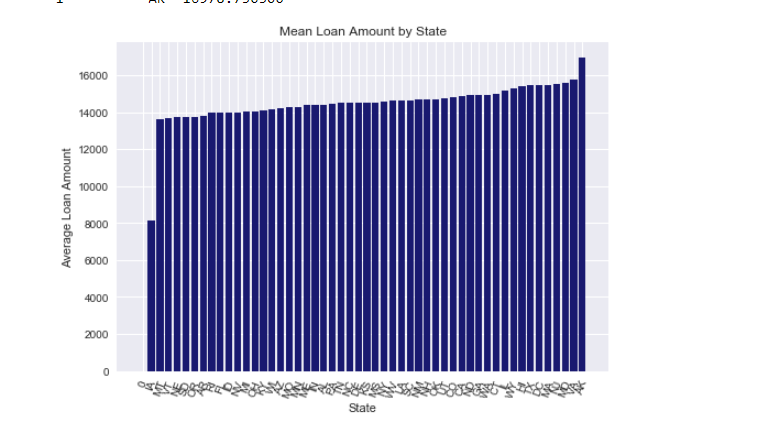
**Loan Amount by Issue Year**

We observed that the loan amounts grew steadily from 2007 to 2017



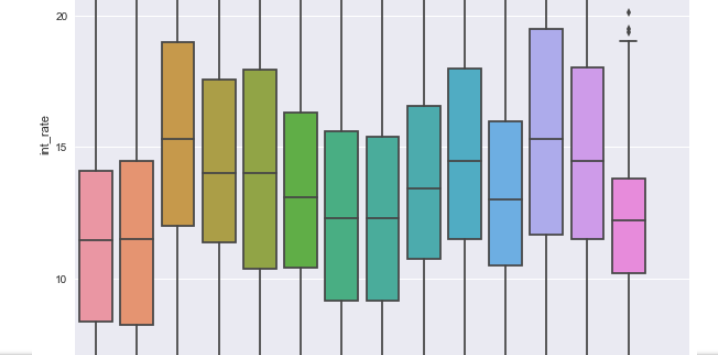
**Mean Loan Amount by State**

The Maximum mean loan amount was provided to Alaska but remains consostent for the other stat

We guage that home\_improvement and major purchases have the same range of interest rates provided. Additionally the interest rates are go to the highest for purpose House.

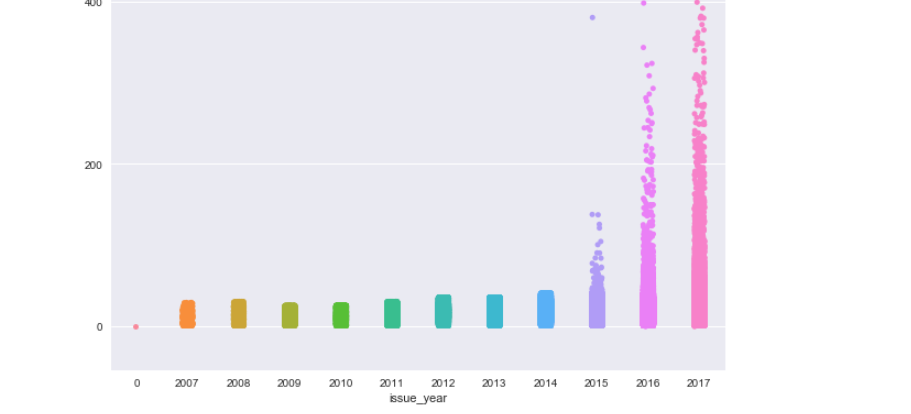
**Purpose by Interest Rate**

We guage that home\_improvement and major purchases have the same range of interest rates provided. Additionally the interest rates are go to the highest for purpose House.

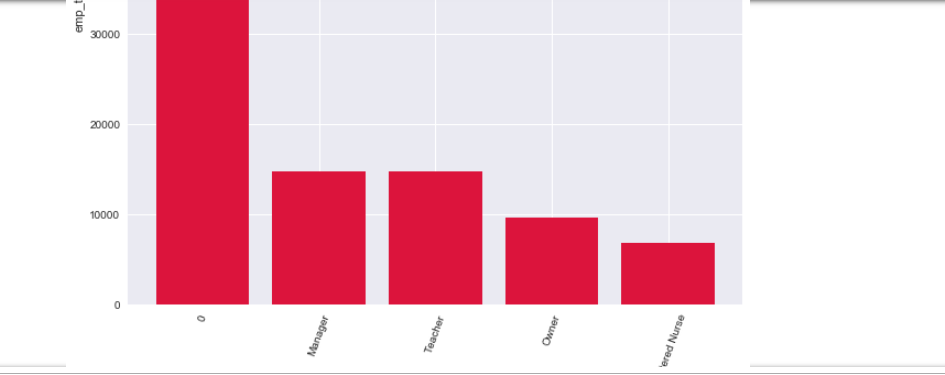


**Analysis of DTI**

Next, we tried to understand the ratio of debt payment by the total debt obligation of all employees by understanding distributionby issue year. Since, this is a ratio it cannot be aggregated further



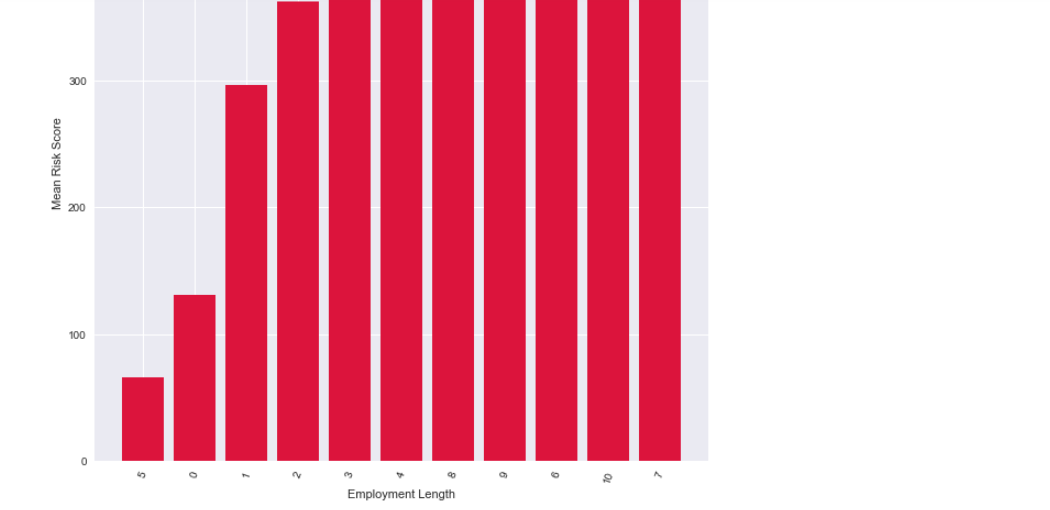
**Employee Title by Interest Rate**



## Exploratory Data Analysis for Declined Loan Data

**Analysis of Risk Score and Employee Year Length**

As per the rejected loan data set the mean risk score is the highest for a borrower who has a an employment lenght of 7 years. Hence there are other factors involved in rejecting a loan apart from risk score

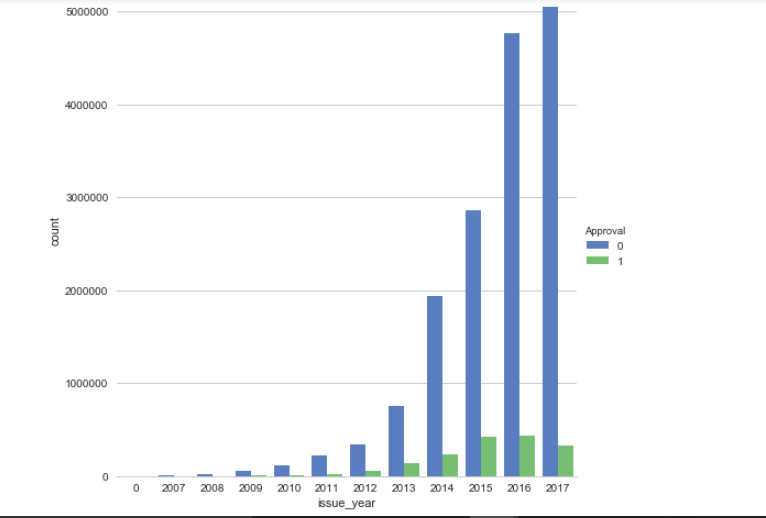


## Analysis of Approval of Loans

Next we analyse Approval by various features. The loans that have been approved have Approval=1 and loans that are not approved have Approval=0. We first combine the data of Approved and Rejected Loans. First we established all the columns common to both data sets.

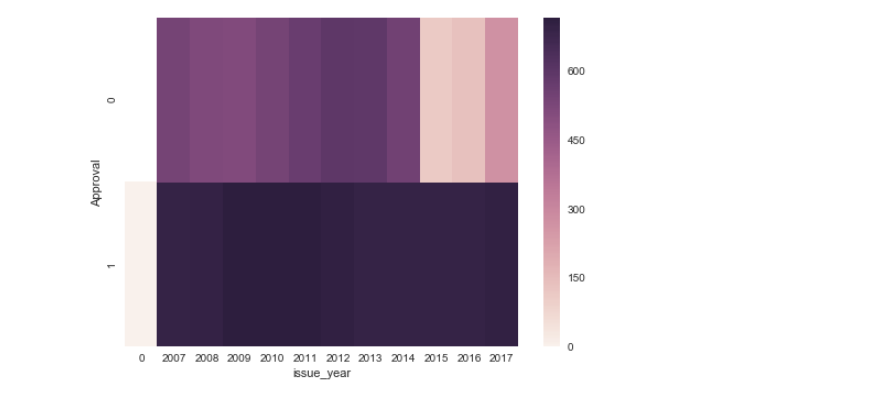
**Analysis of Approval Counts versus Issue Year**

Analysed the comparison Approvals and Rejects by the Issue Year. Highest number of rejects were for the year 2017 and the highest approvals were for the year 2016



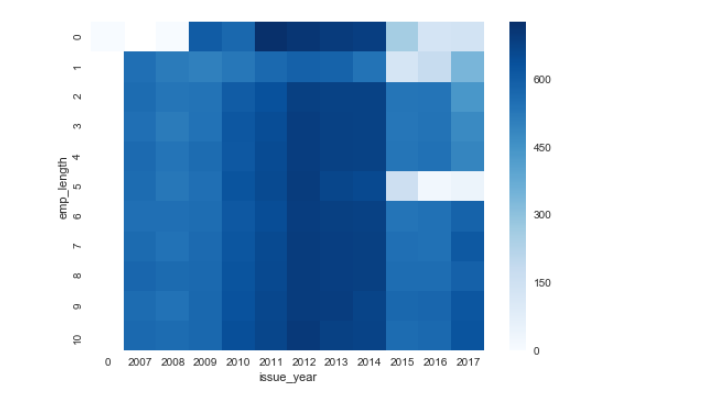
#### Analysis of Risk Score with Approval and Issue Year

#### From the heat map it is clearly visible that Approval score of 0 have varied risk scores of less than 150 and greater than 450 for the years of 2017 and 2012 respectively. Loans that were approved have conssitent risk scores of above 450.



#### Analysis of Mean Risk Score with emplyement length and Issue Year

We observe that the mean risk score is the highest for employment length 9 years followed by 10 years for years 2013 and 2014. Another noted observation is mean risk score is the highest for employment lenght 10 but is consistenty high for all employment years for the year 2012.

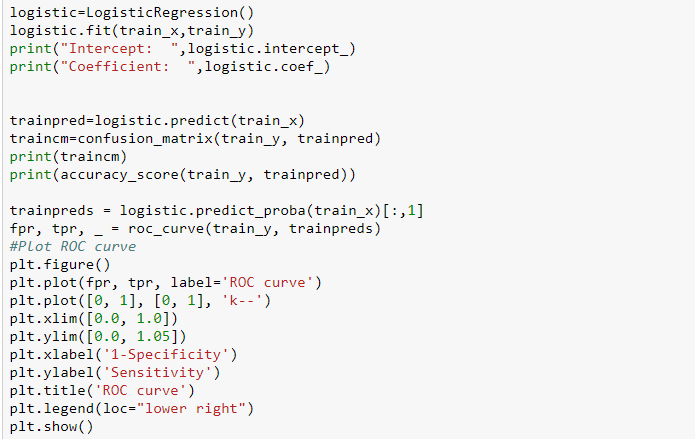


**PART 2**

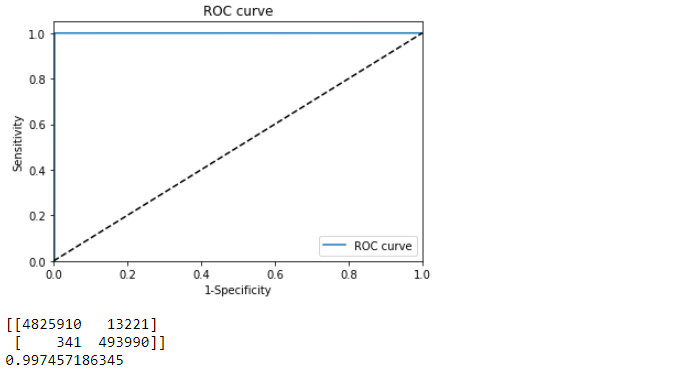
# Classification

In order to predict if a user is eligible to get a loan from Lending Club we used Logistic Regression, Neural Network and Random Forest Algorithms. Please find the screenshots attached below.

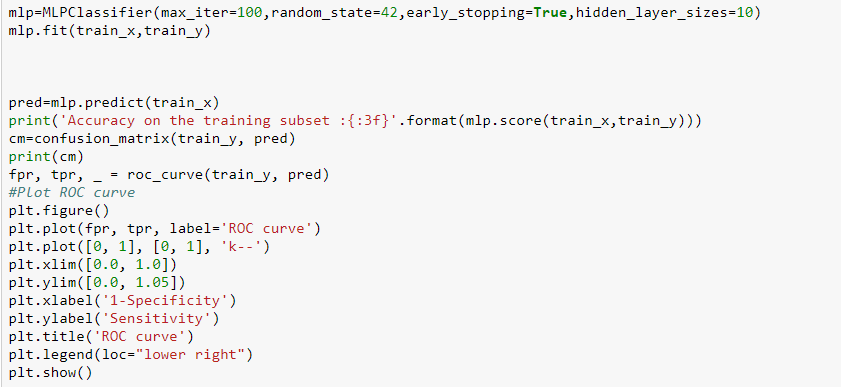
## Logistic Regression



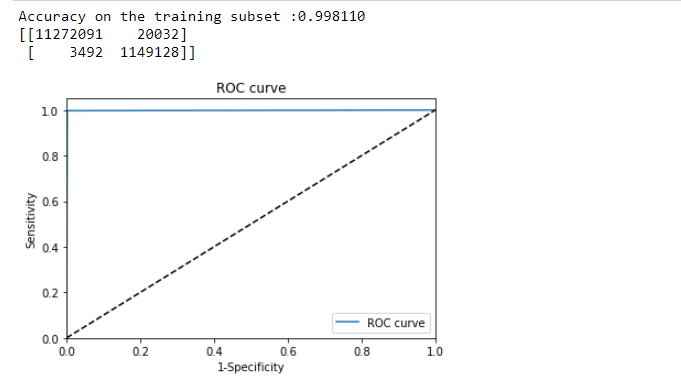
**Result Set**



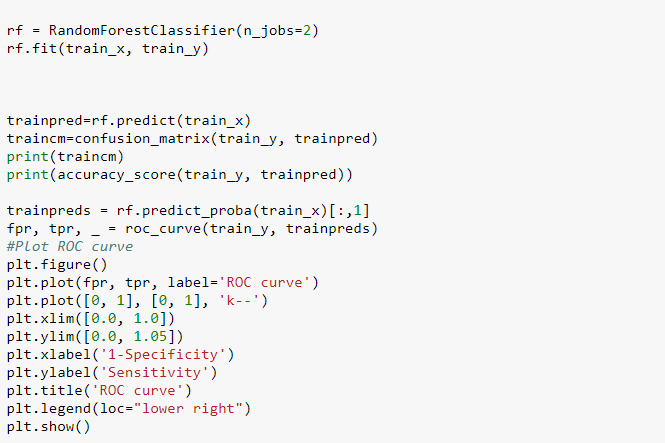
## Neural Networks



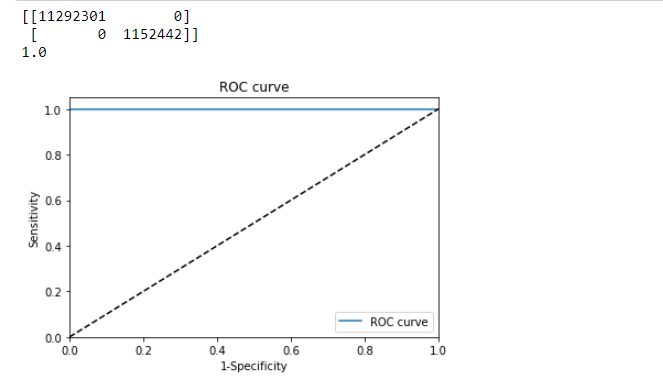
**Result Set**



## Random Forest

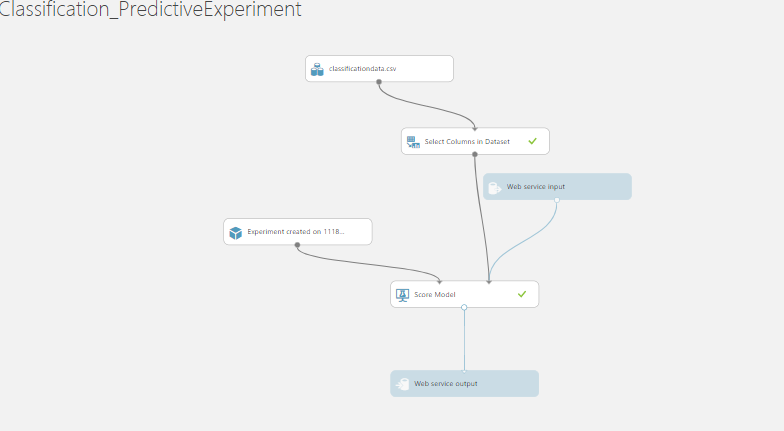


**Result Set**

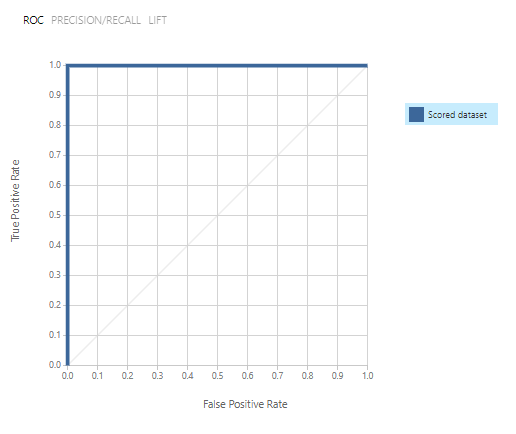


## Deployment on Azure

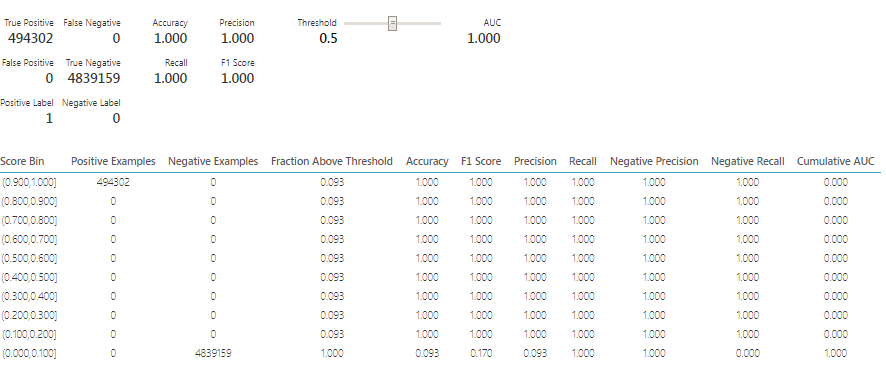
**Random Forest Classifier**



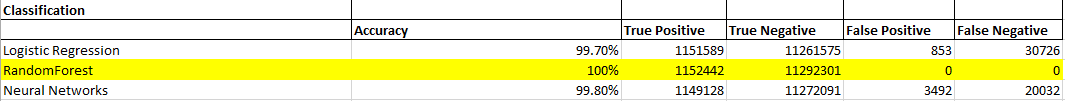
**Resulting ROC Curve**



**Confusion Matrix**



## Summarization of Results for Classification



# CLUSTERING

After deciding on the eligibility of a user to get a loan, next we predict the interest rate that would be assigned to the user based on clustering. In order to predict interest rate we clustered the data in the following ways:

1. Manual Clustering
2. Clustering Algorithm – Kmeans
3. No Clusters only Data

The best model for each cluster was then deployed on Azure to create a REST API

## Manual Clustering

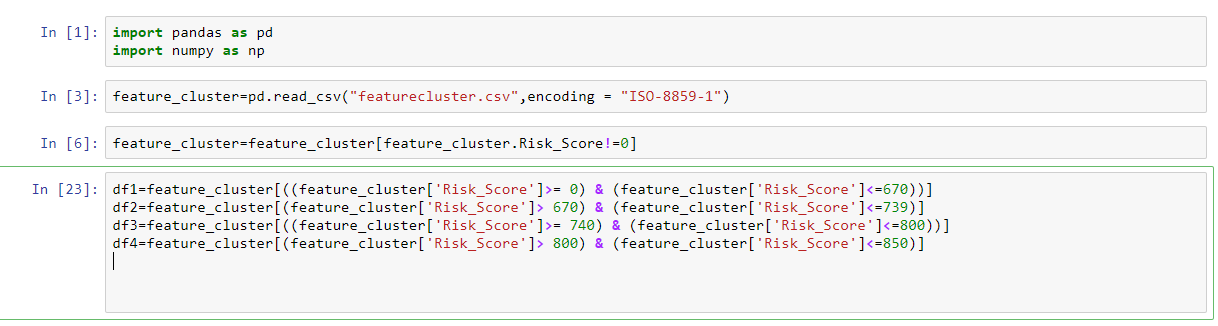
We created clusters on the basis of Risk Scores. Attached is the screenshot of the clusters

Cluster 1: Risk Score – 0-670

Cluster 2: Risk Score – 670-739

Cluster 3: Risk Score – 740-800

Cluster 4: Risk Score – 800-850

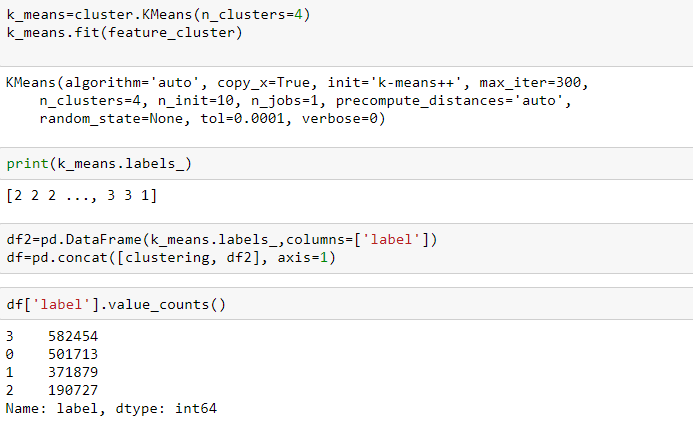


## Clustering using K Means Algorithm

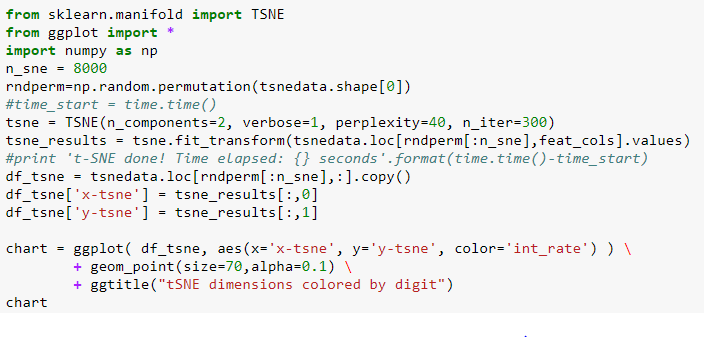
**Features Used**



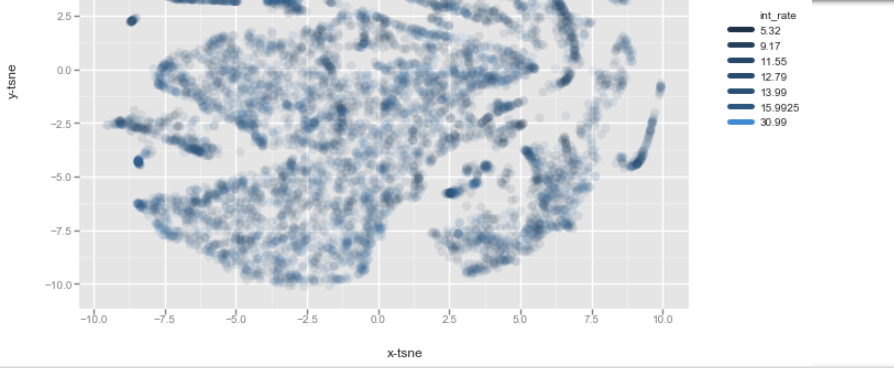
**Clusters Created**



## TSNE



**Result Set**



# PREDICTION

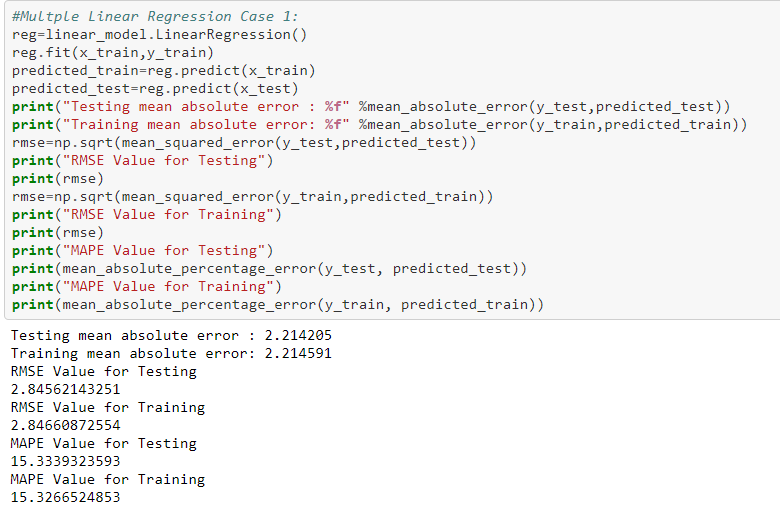
After clustering the data, next we predicted the interest rate. We first ran Neural Networks, KNN and Linear Regression on each of the clusters. Next, deployed the best model on Azure and created a REST API for each cluster. In all clusters the best algorithm was Random Forest Regressor.

## Manual Clustering

Manual Clustering resulted in 4 clusters. Below is the attached screenshot for cluster 1. The example screenshot attached is for Linear Regression

### Cluster 1

Example screenshot for Linear Regression. We also ran Random Forest Regression and KNN on Cluster 1. Please refer to ManualCluster1Regression.ipynb for all algorithms



### Cluster 2

Example screenshot for KNN Regression. We also ran Random Forest Regression and Linear Regression on Cluster 2. Please refer to ManualCluster2Regression.ipynb for all algorithms

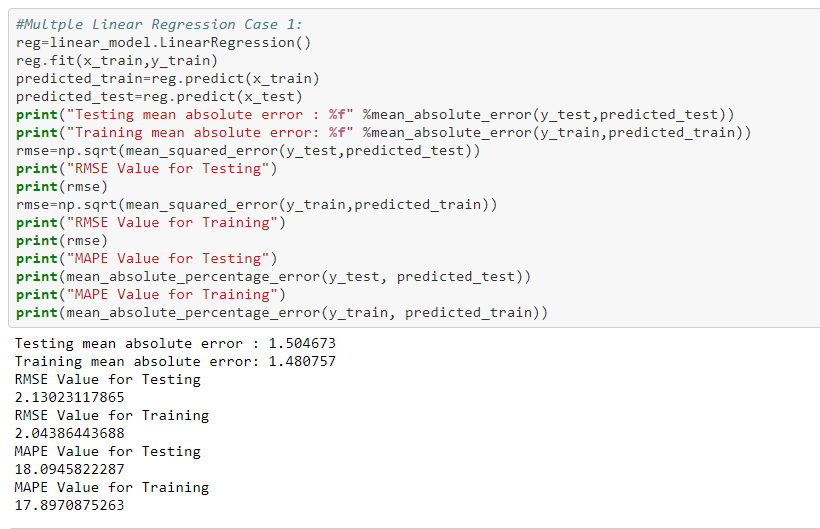


### Cluster 3

Example screenshot for Random Forest Regressor. We also ran KNN and Linear Regression on Cluster 3. Please refer to ManualCluster2Regression.ipynb for all algorithms



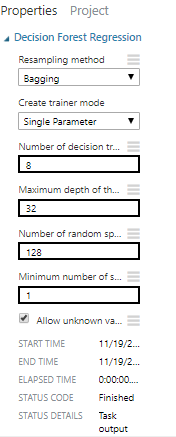
### Cluster 4



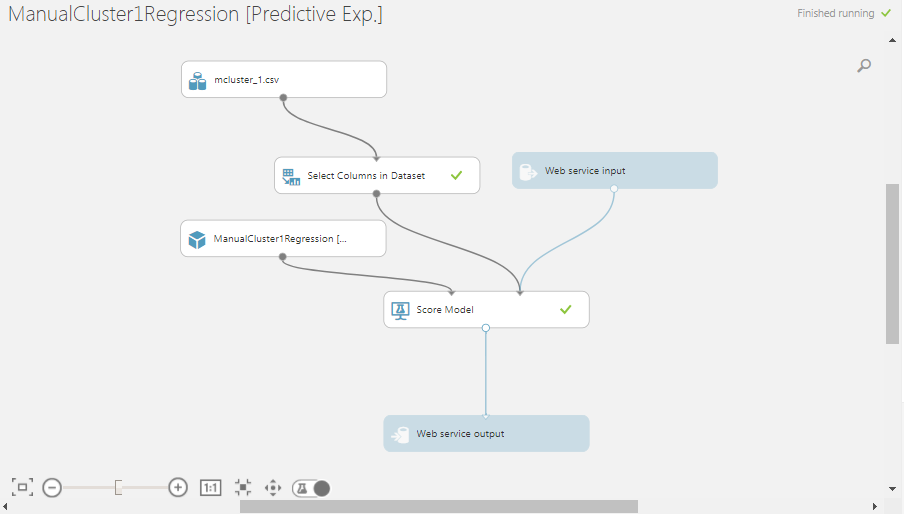
### Manual Clustering Prediction and Web Service Deployment on Azure

We used the following parameters. Though number of trees did play a major role in reducing MAE. Due to computational time overhead we reduced the number of trees. The MAE came down by 0.002 when number of trees was increased.

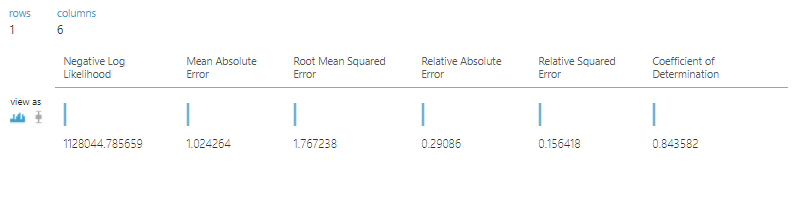
**Parameters**



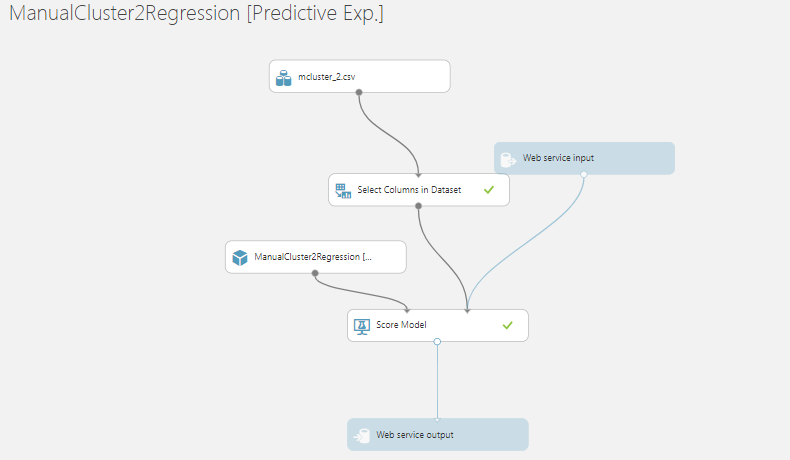
**Manual Clustering 1**



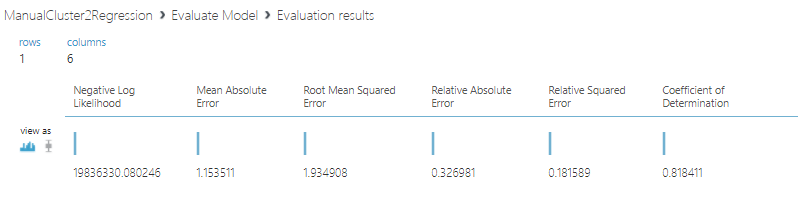
Results



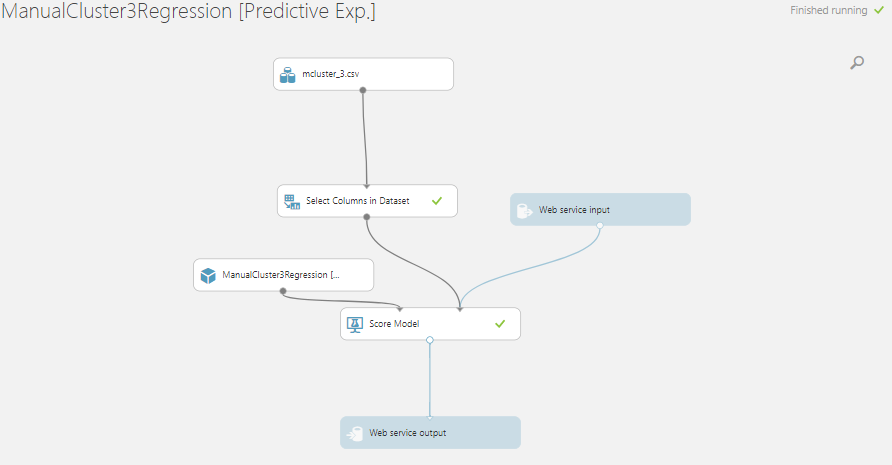
**Manual Clustering 2**



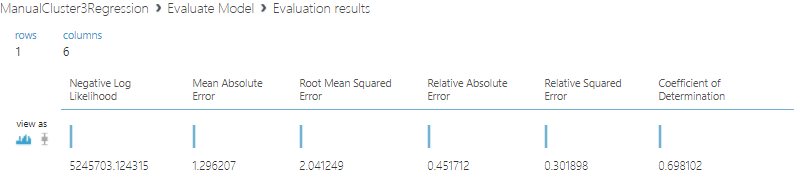
Results



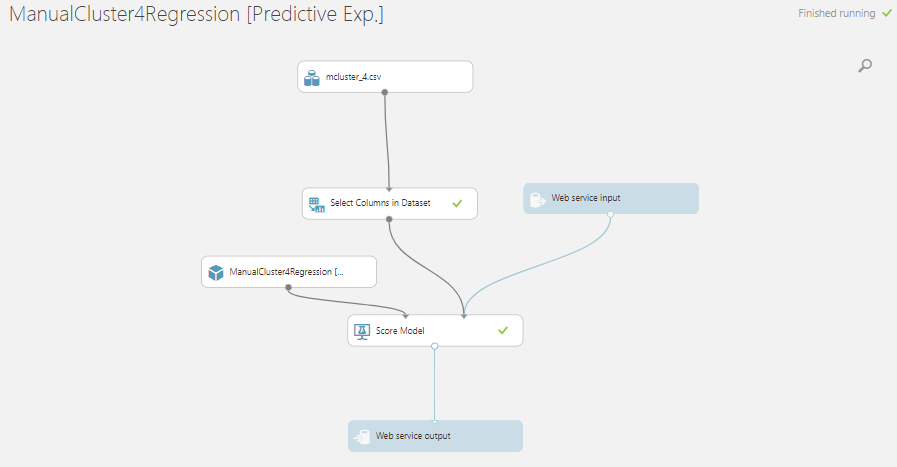
**Manual Clustering 3**



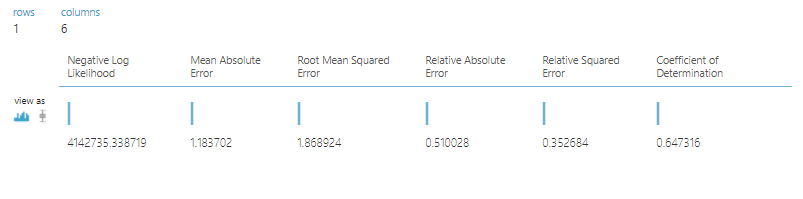
**Results**



**Manual Clustering 4**

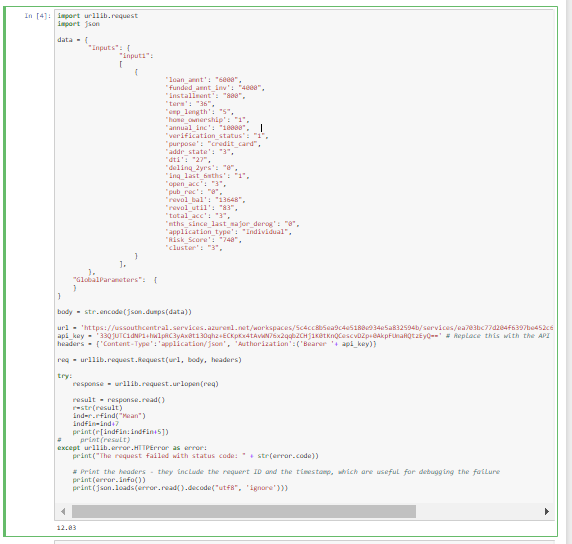


**Results**

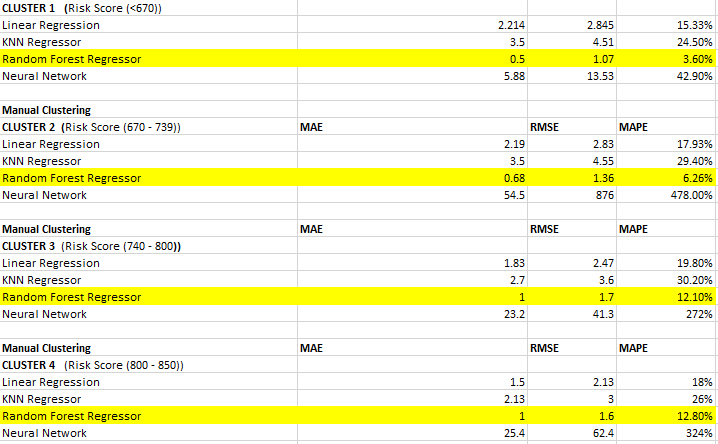


### Consuming Manual Clustering WebService

**We consumed all Manual Clustering Web Services using Jupyter Notebook**



### Final Results for Manual Clustering



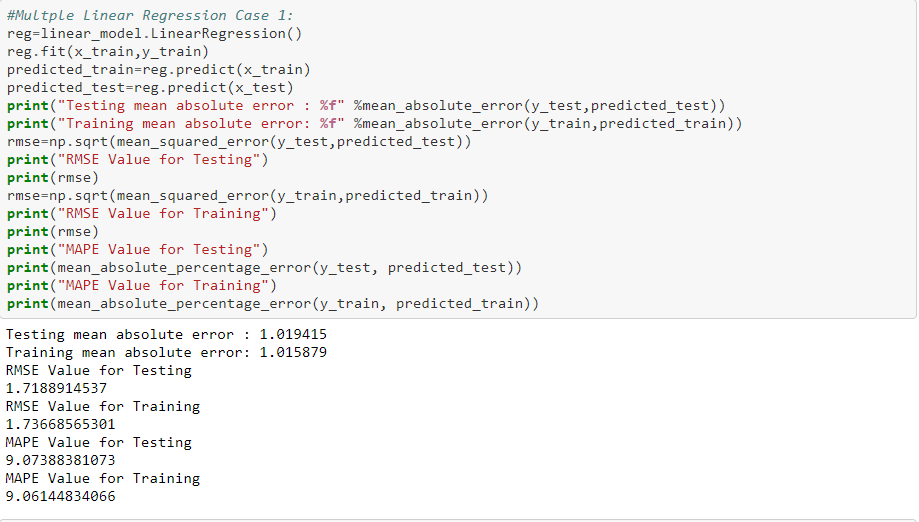
## Prediction using Clustering Algorithm K Means

Using K Means we created the 4 clusters namely

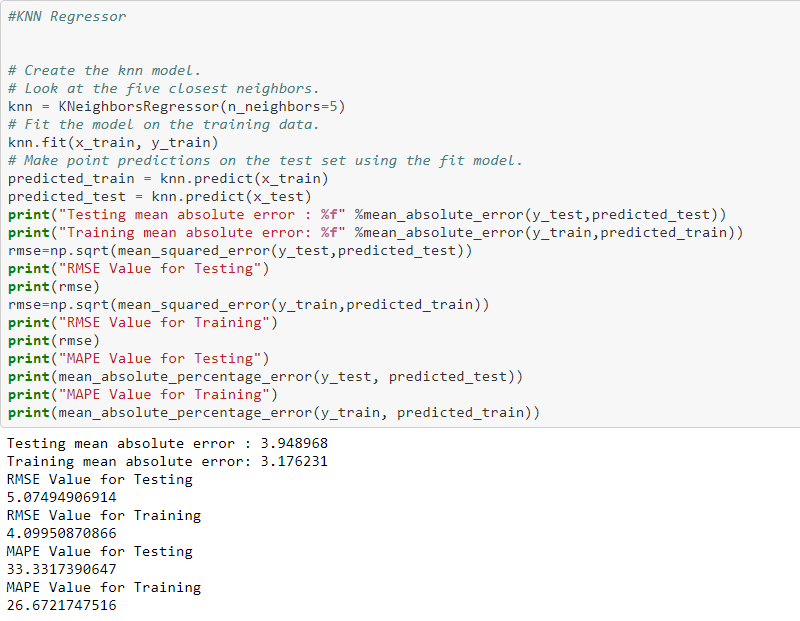
Cluster 0, Cluster 1, Cluster 2, Cluster 3. We used Linear Regression, Random Forest Regressor and KNN Algorithms for each of the cluster. We then deployed the best algorithm for each cluster on Azure. Have attached sample snapshots for each cluster. Please refer to the KMeansCluser.ipynb for detailed analysis

### Cluster 0

Example screenshot for Linear Regression. We also ran Random Forest Regression and KNN on Cluster 1. Please refer to KMeansCluster0 Regression.ipynb for all algorithms



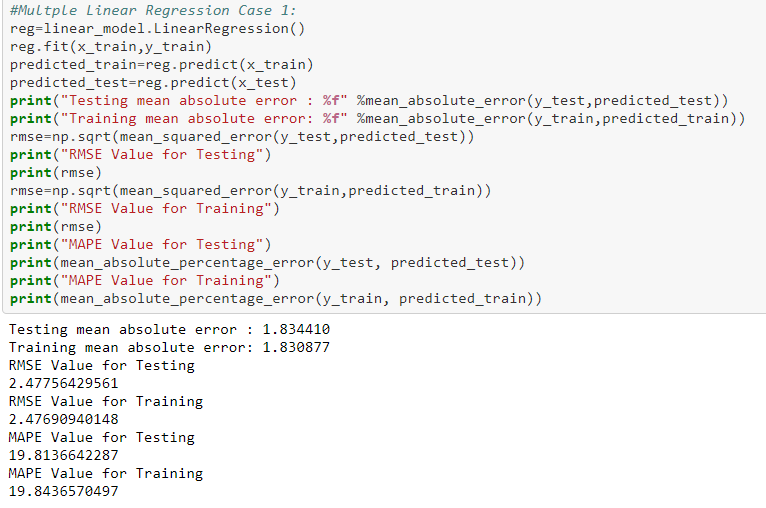
### Cluster 1



### Cluster 2

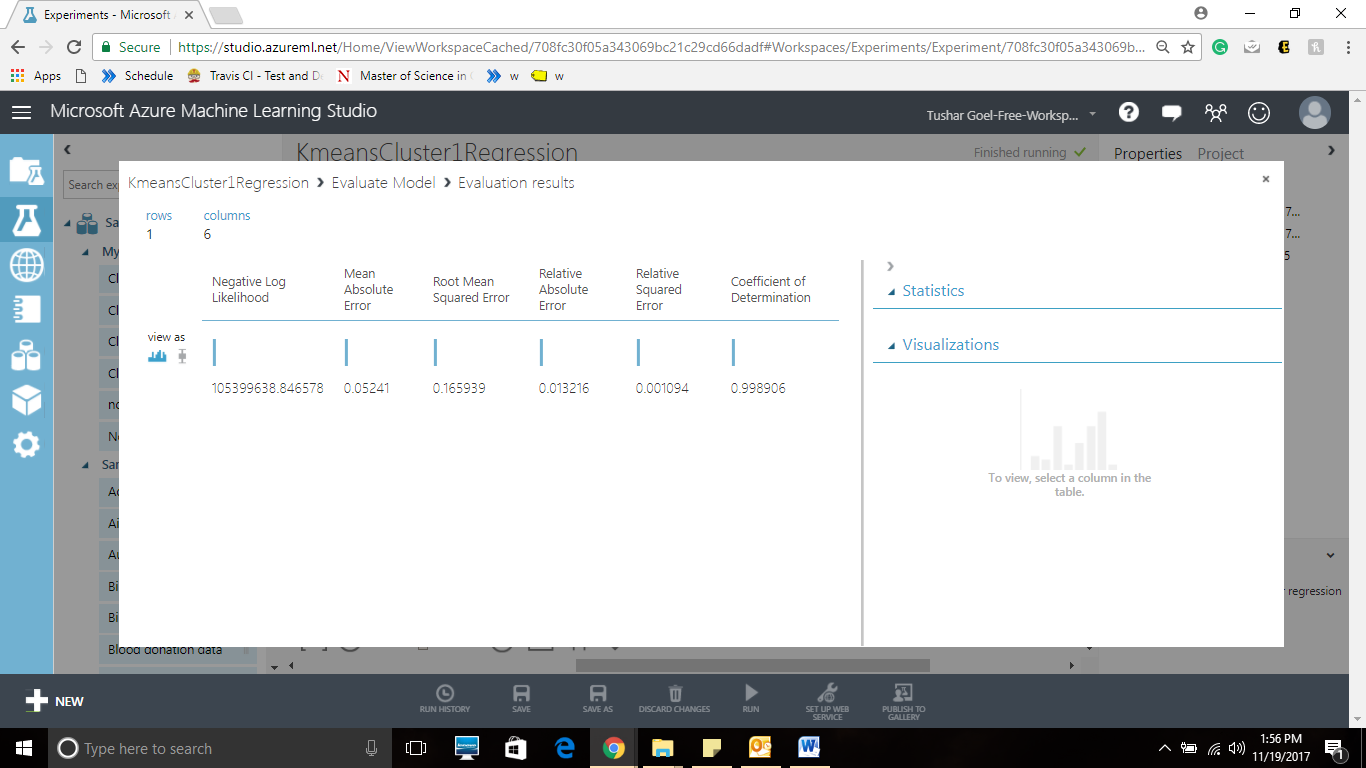


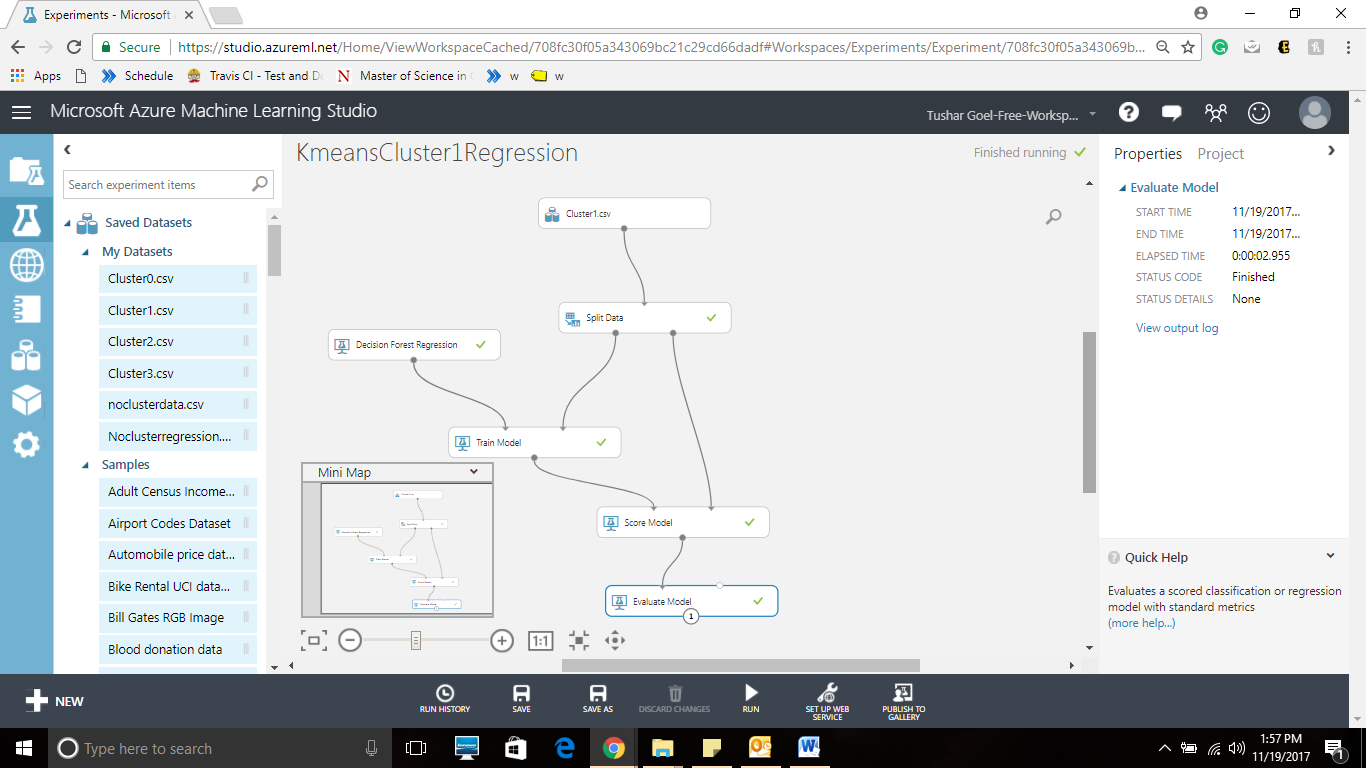
### Cluster 3



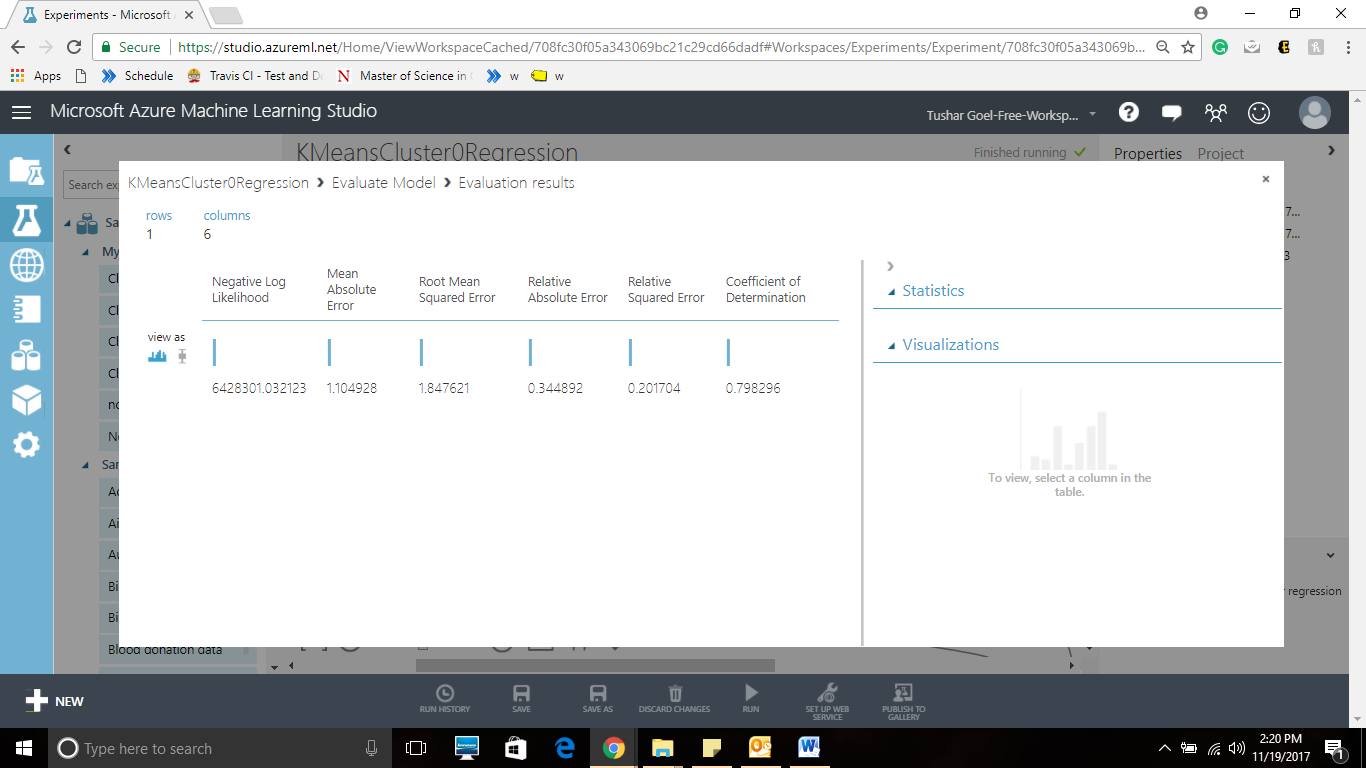
### Prediction and Deployment using KMeans Clustering in Azure

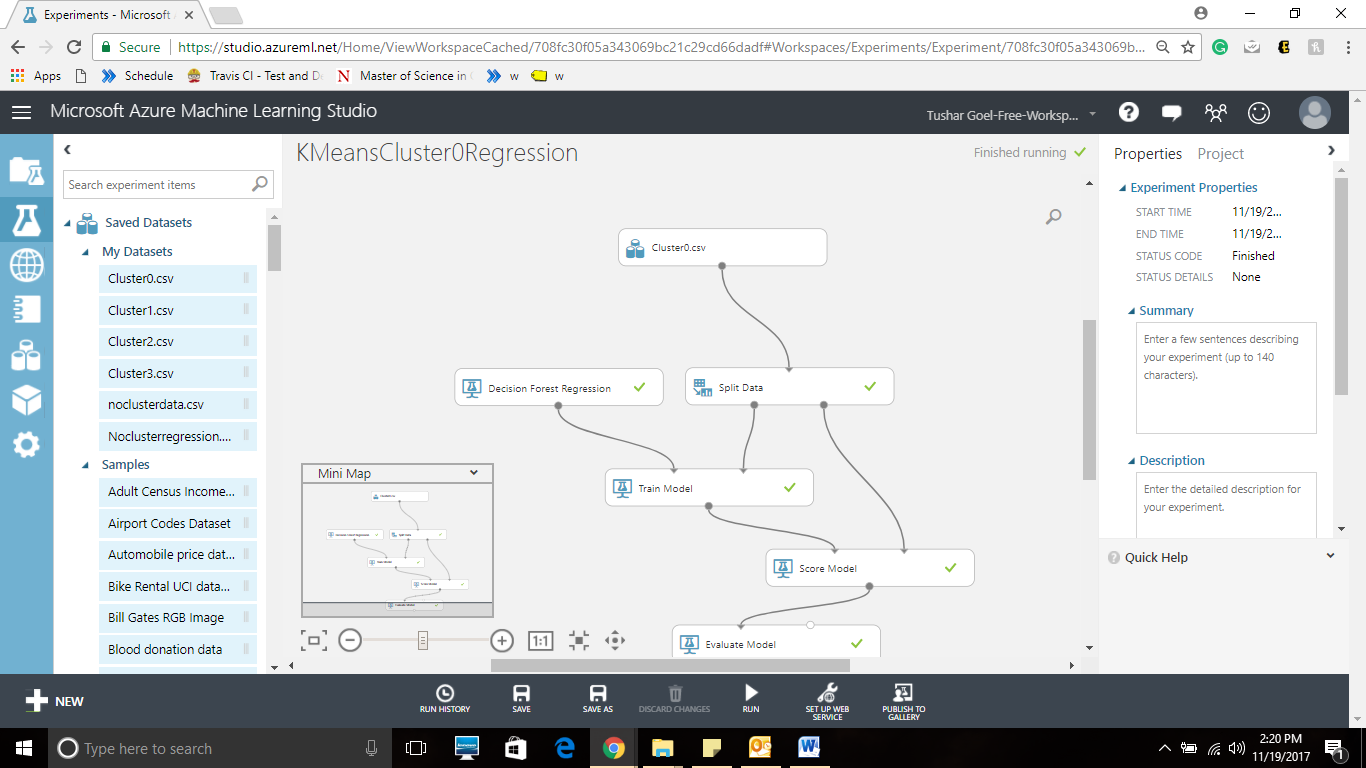
**Cluster 1**



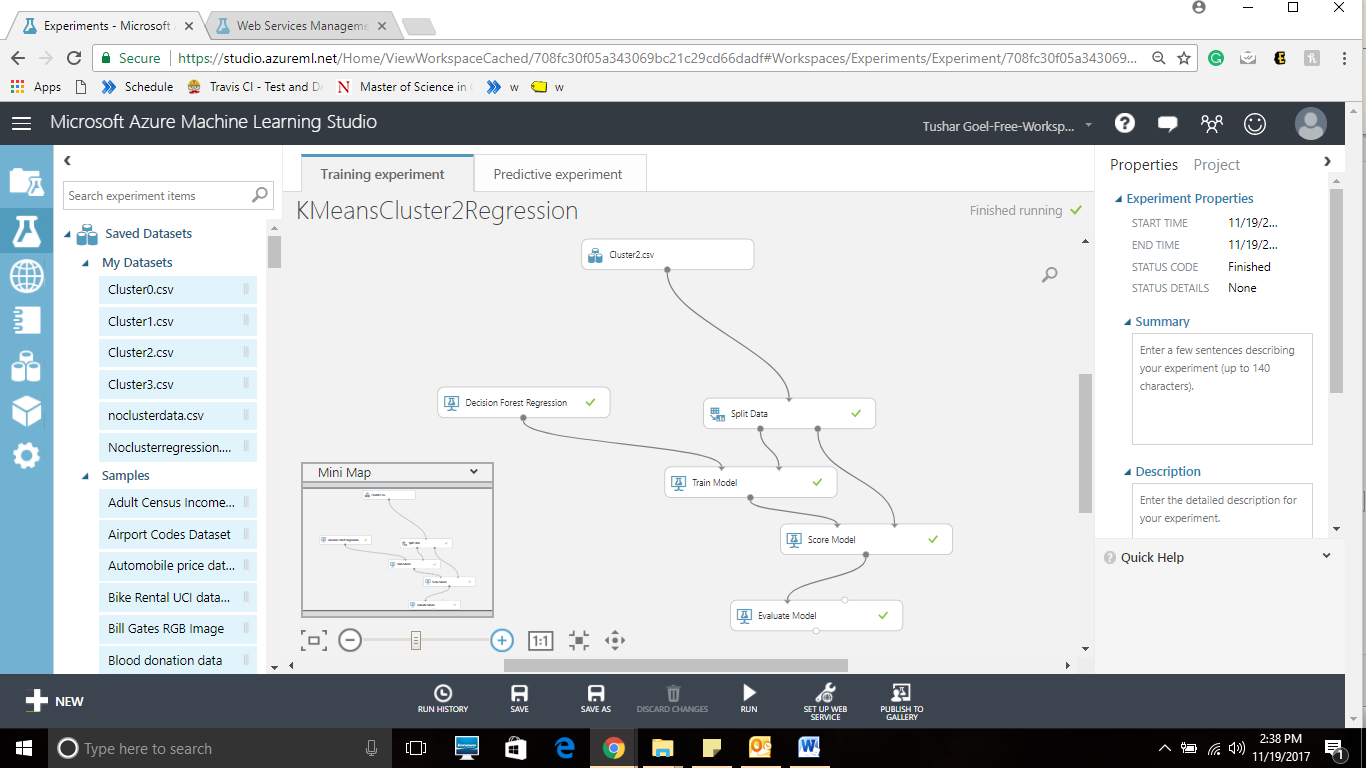


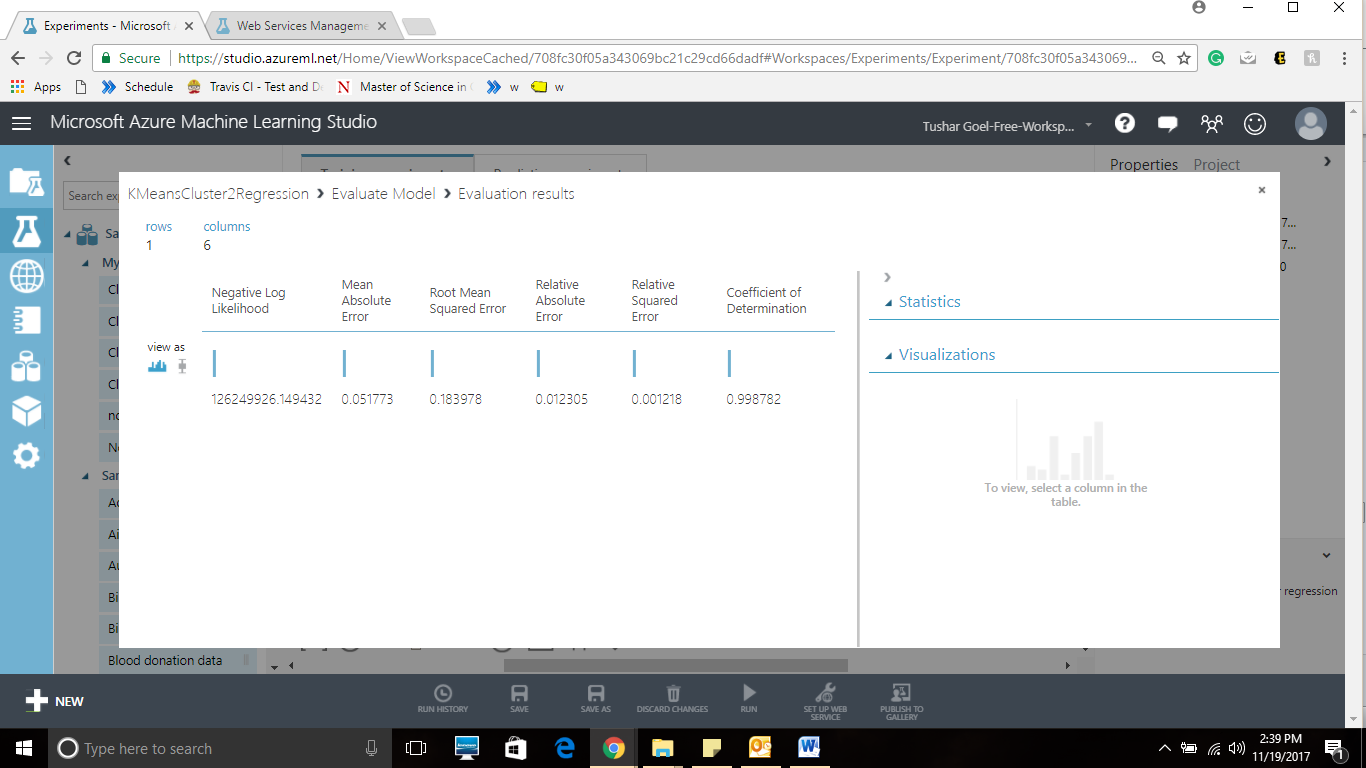
**Cluster 0**



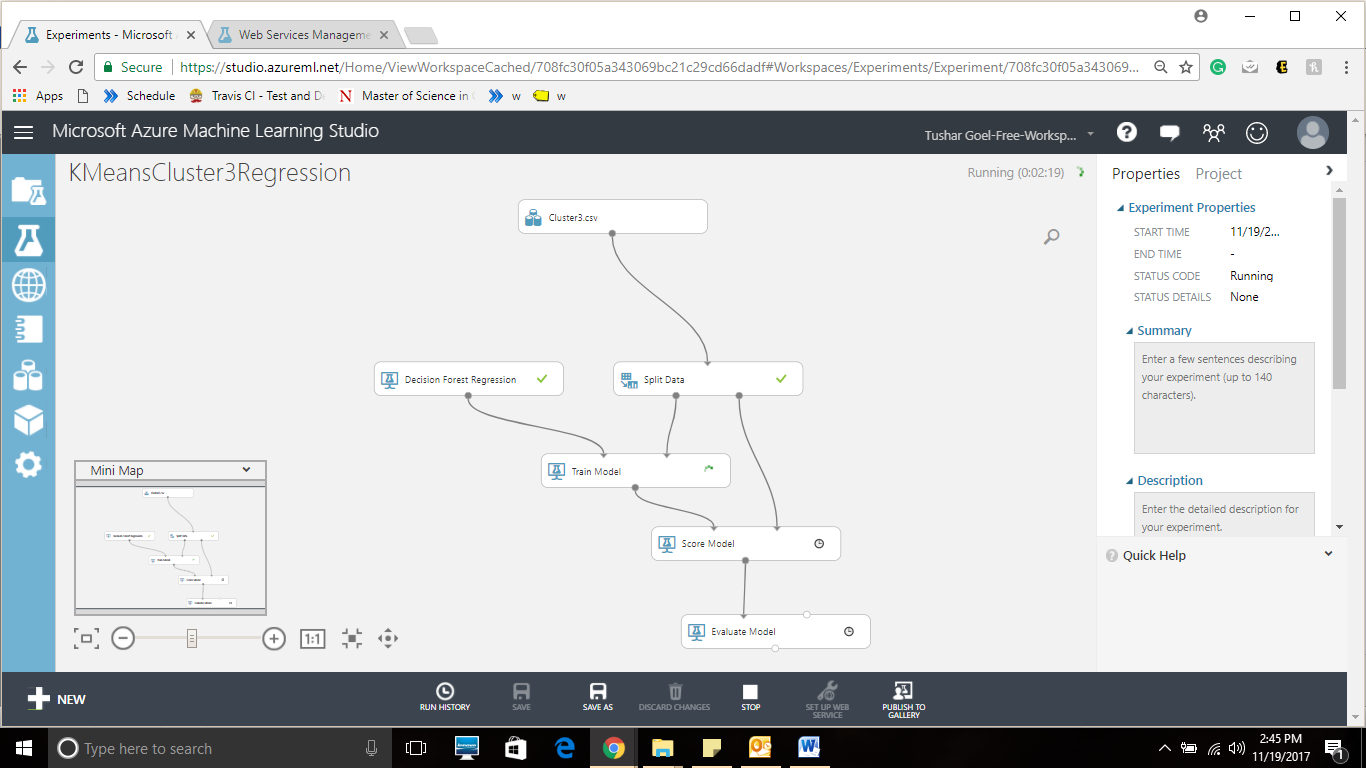


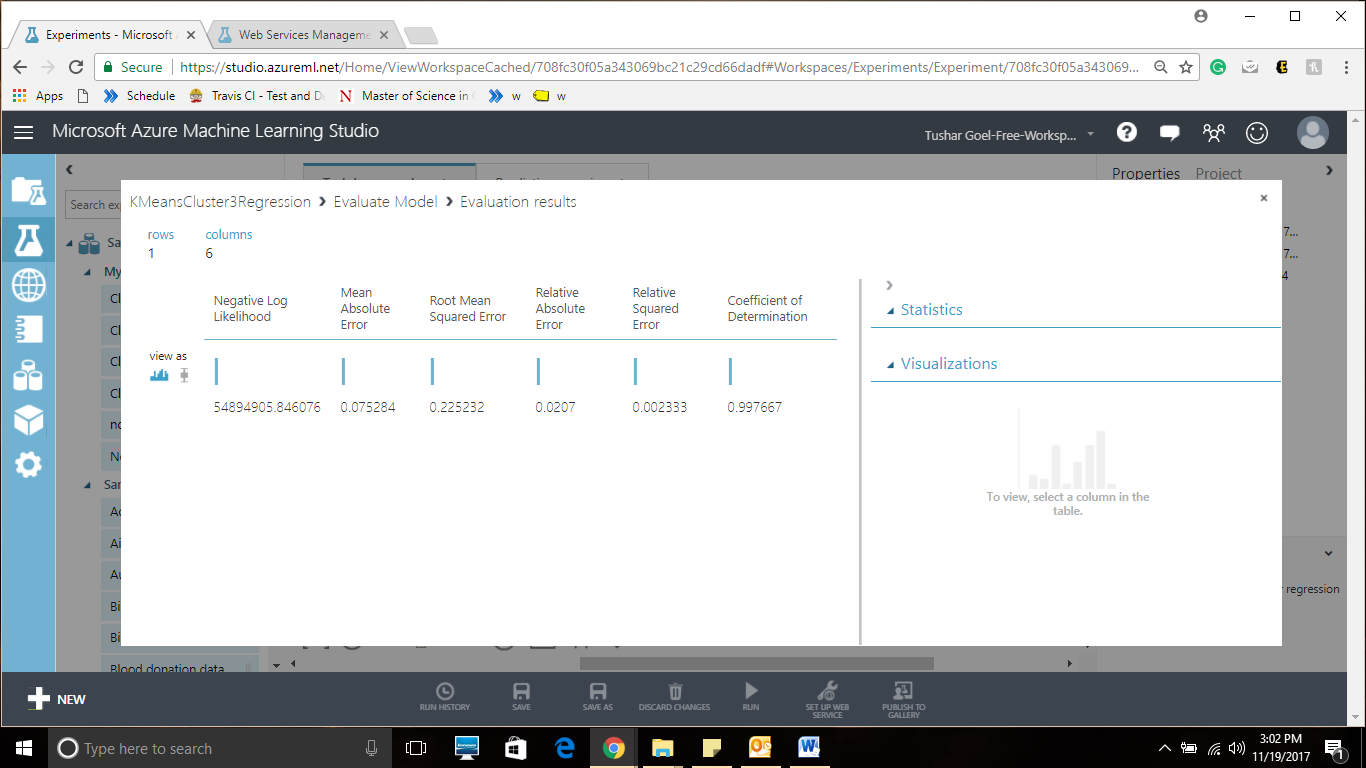
**Cluster 2**



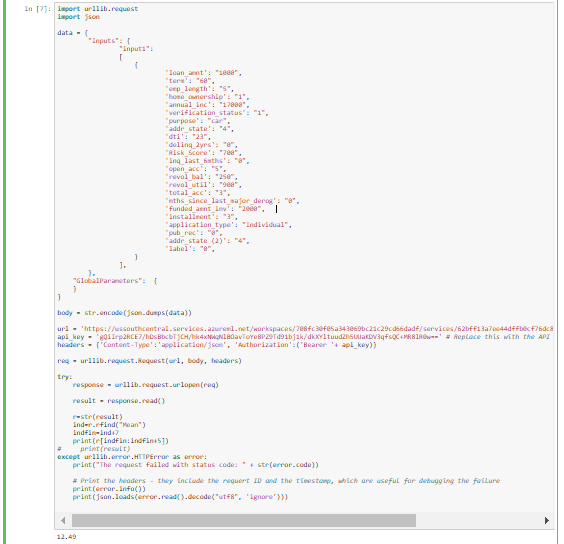


**Cluster 3**





### Consuming the web service using Azure

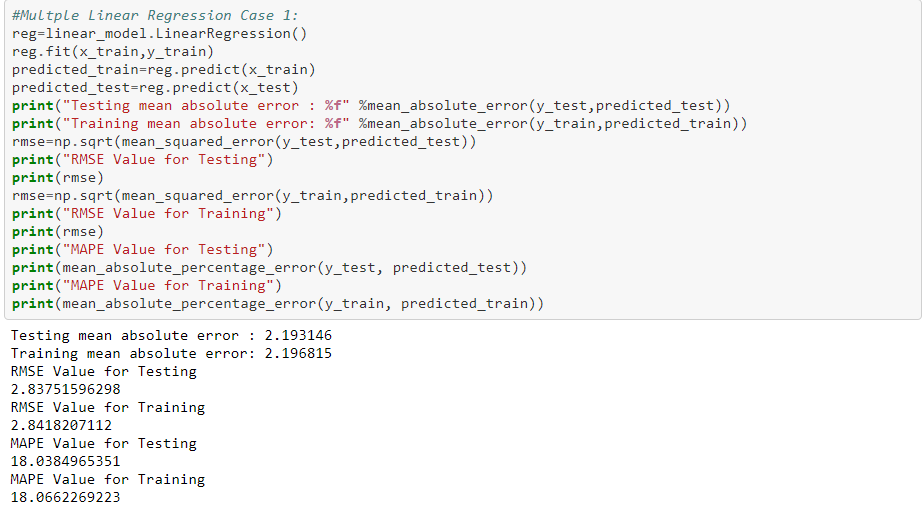


### Summary of Results using KMeans Clustering

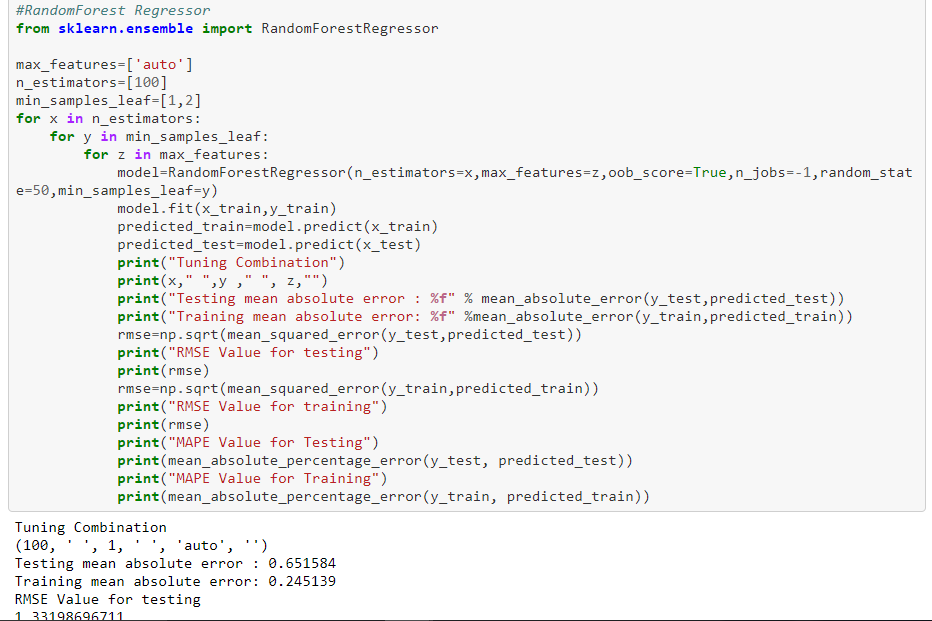


## Prediction with No Clustering

We then ran the data with no clustering for all three logarithms ie Linear Regression, Random forest and KNN Algorithms

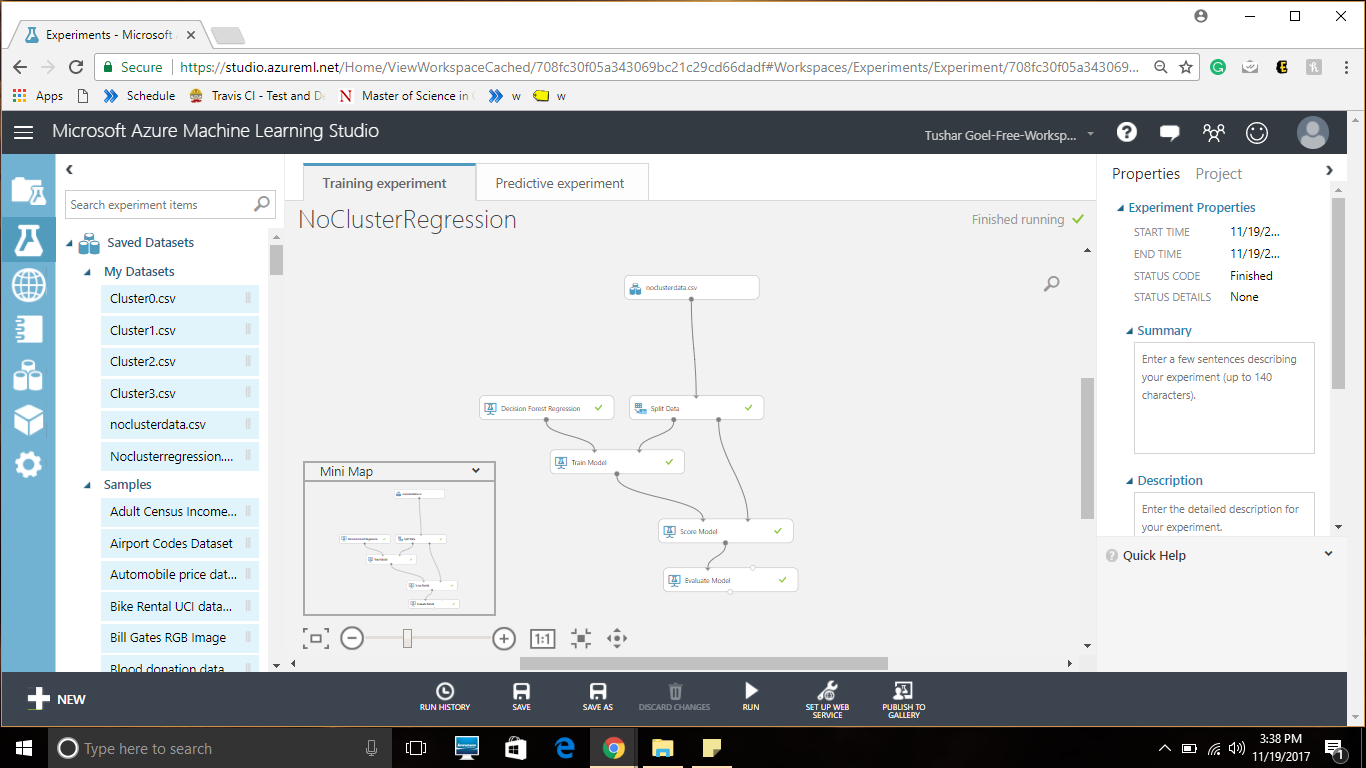


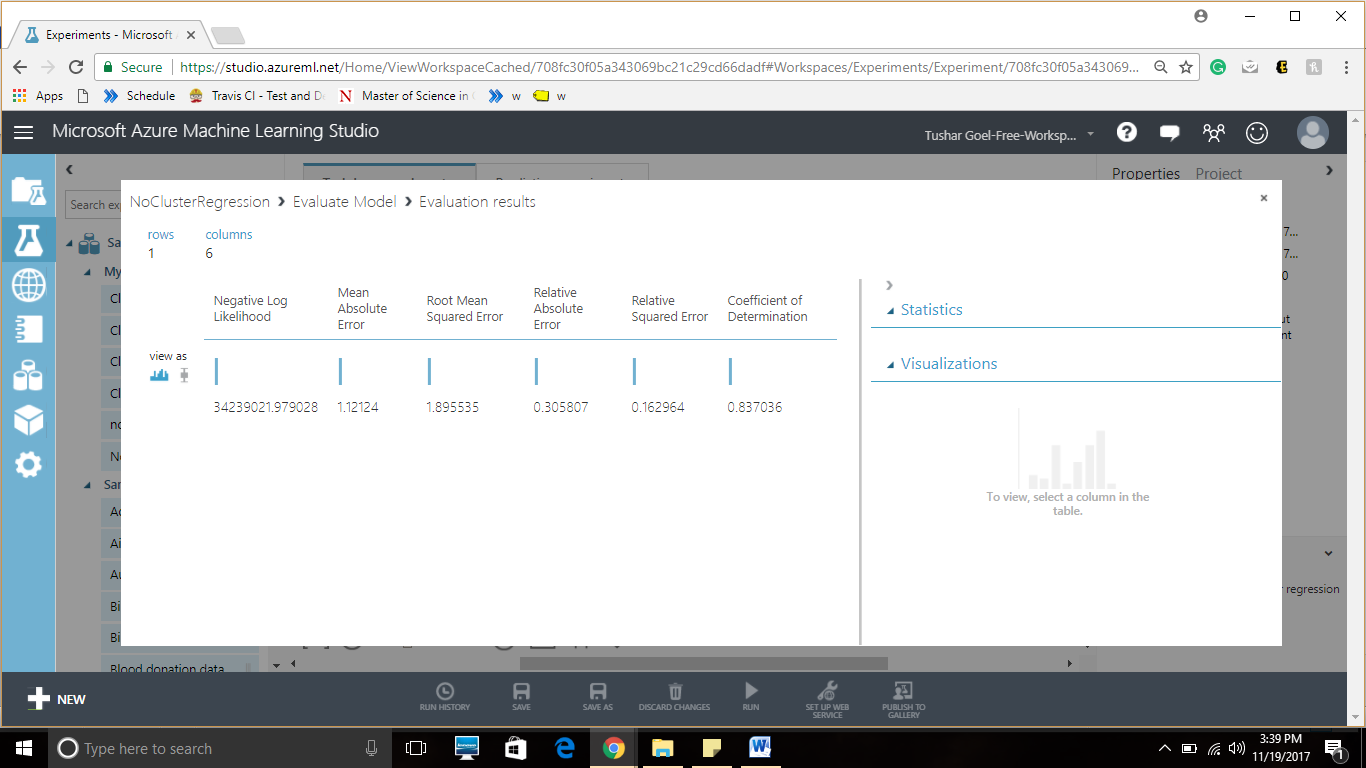




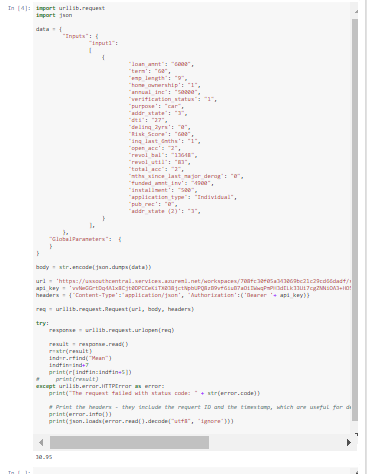
## Model Deployment using Azure

No Clustering Regression

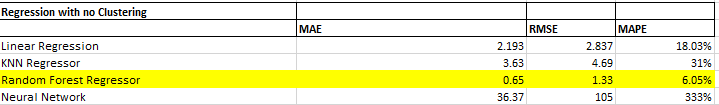




## Consuming web service using Azure



### Summary of Results for data with No Clusters



# DEPLOYMENT

We integrated this entire workflow. The deployment has been done in Jupyter Notebook where the user enters the user defined parameters . Next it is identified if the user is eligible for the loan. If eligible models are run for the clusters in Manual Clustering, Kmeans and No clusters . Next, after fetching interest rates for each of these, the highest interest rate is provided to the user. Below is the screenshot



