**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

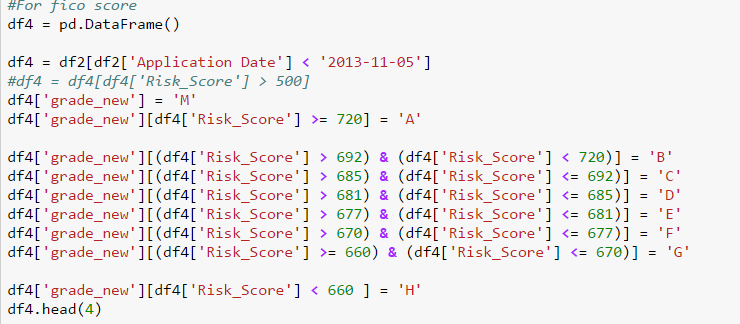
CHOSE BEST MODEL

**Data Modeling before running models:**

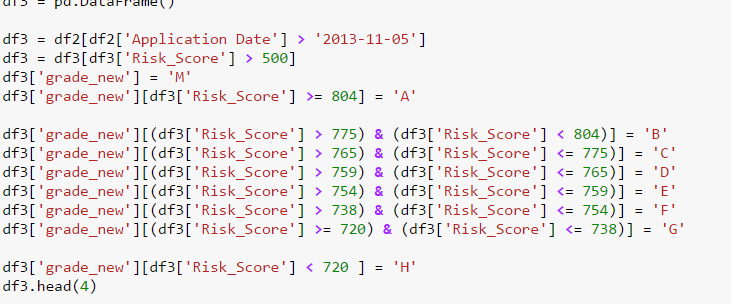
1. As Risk Score in declined loan data set is FICO score for date ranging before November 2013 and Vantage score for date later, we handled both the scenario differenly
2. Removed all those rows where risk score is zero (As zero risk score will incease the deviation of model towards declined data set)
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)**

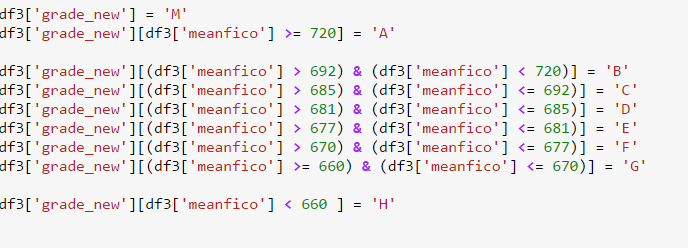
For less than 11th November (FICO Score)



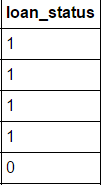
For more than 11th November 2013 (Vantage Score) (Grades scaled according to Fico score Grades)



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

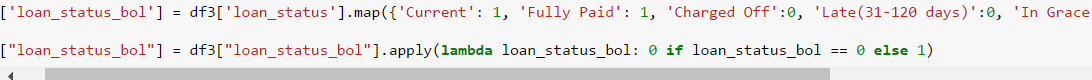


Created column for Loan Declined and Not Declined for classification

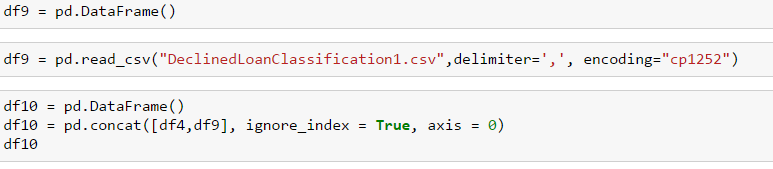


**Note : To make the classification model even better we included all the loans in Loan data set which are not Paid or current as Failed or Declined Loan**

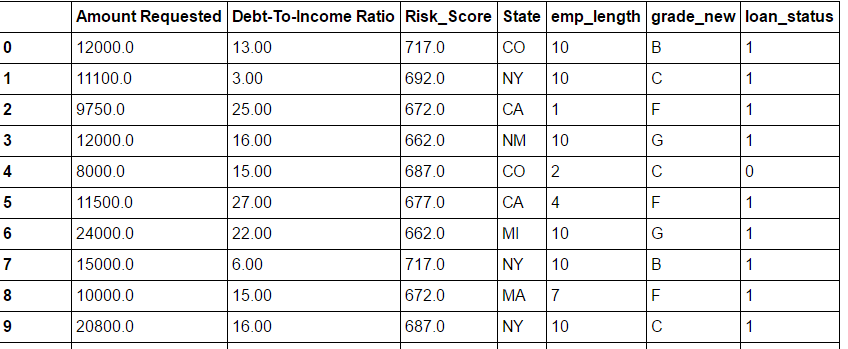
**Reason for doing this : As all the non -paid loans are actually deliquent loans so putting them in declined status makes for a better prediction**



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

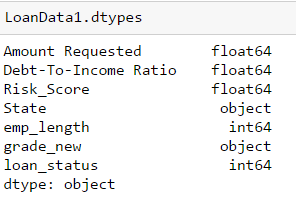


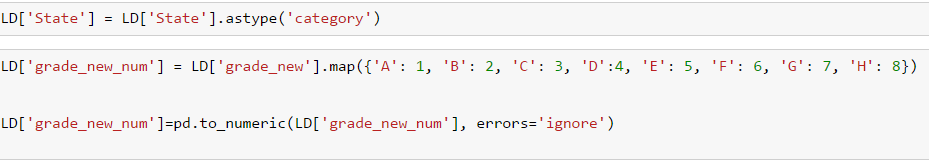
1. Values/Columns chosen for merged dataset:



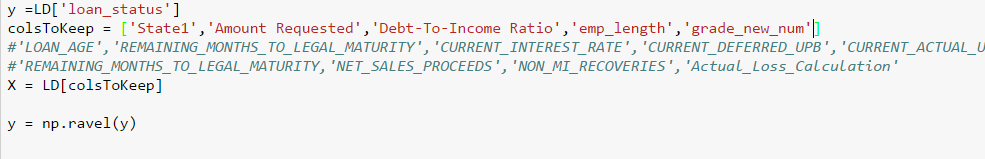
**Running Classification Models on Python Notebook:**

1. Changed data types as object to categorical:





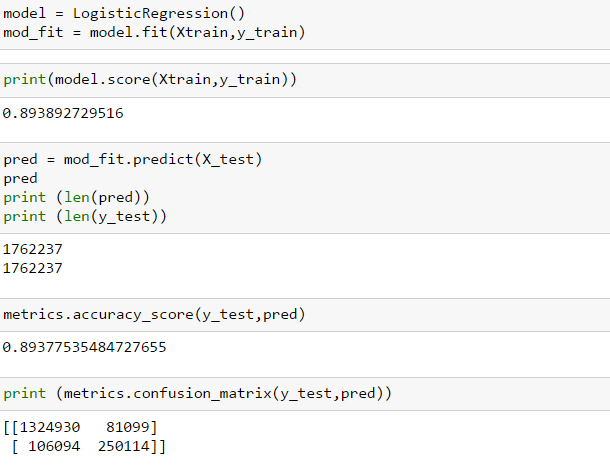
1. Chose the columns for data division:



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

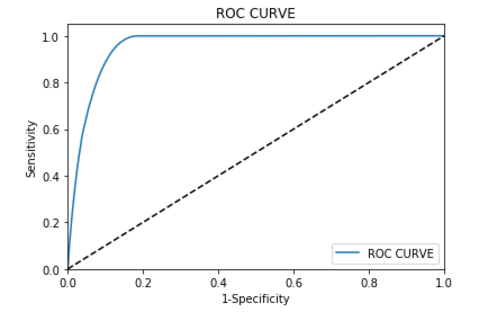


**Logistic Regression:**

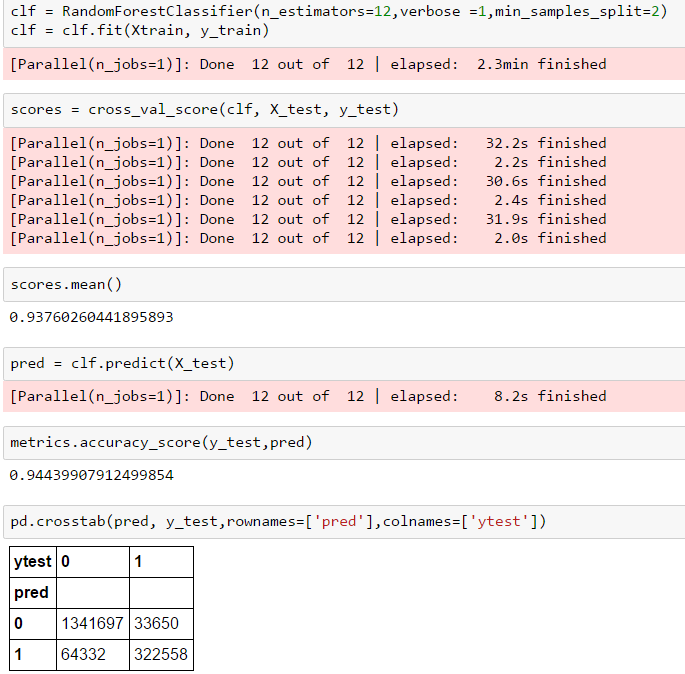


ACCURACY

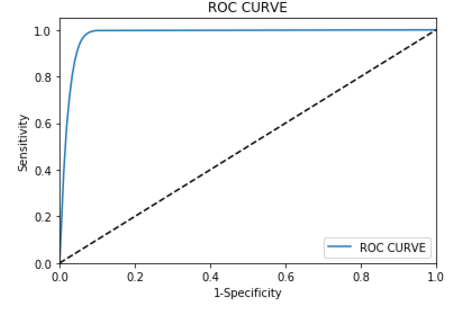
ROC curve for Logistic Regression:



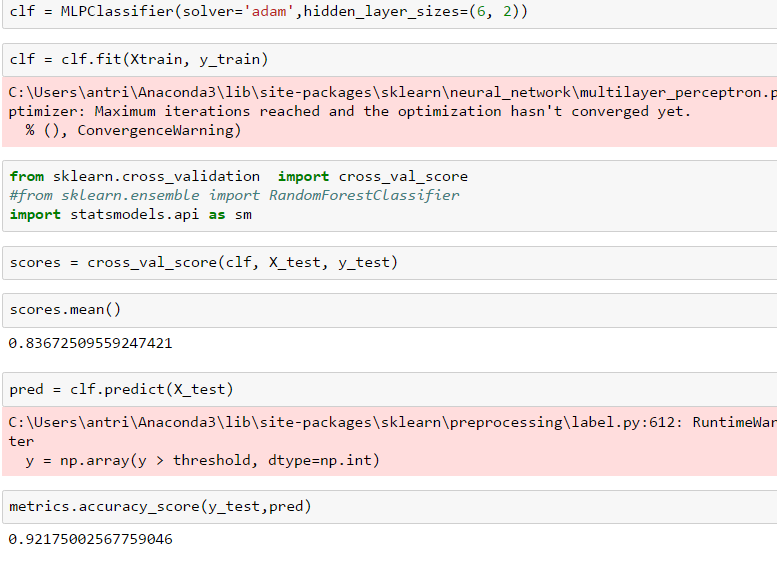
**Random Forest**

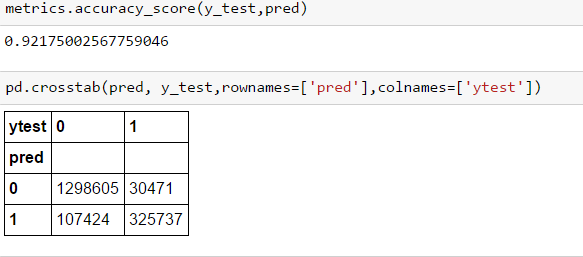


**ROC Curve for Random Forest:**

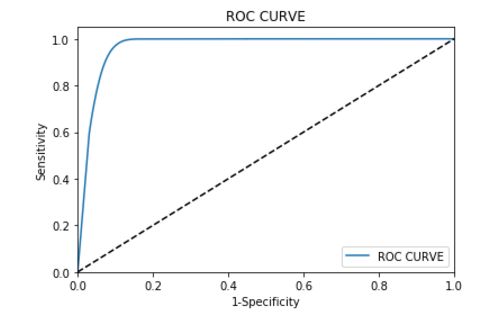


**Neural Network:**



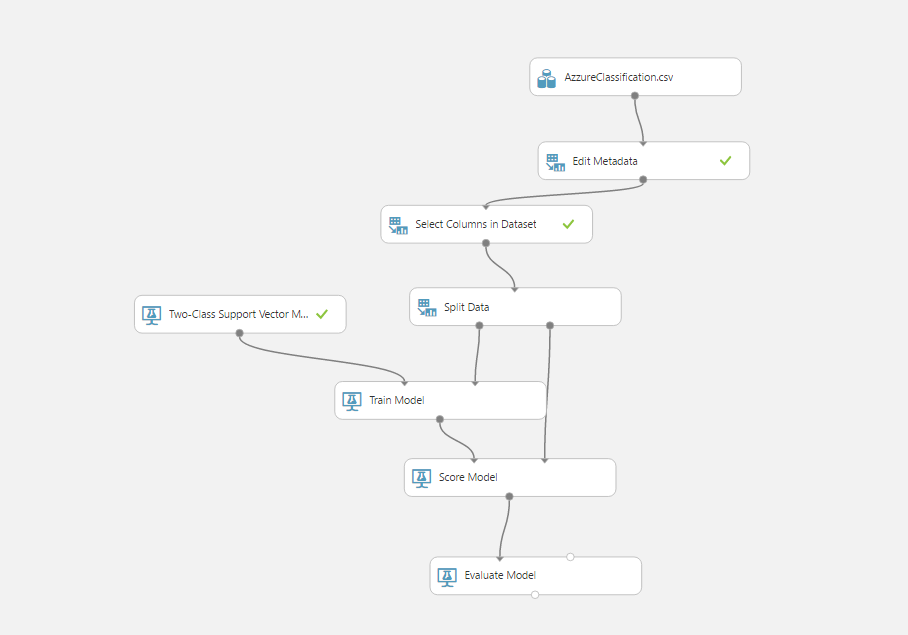


**ROC Curve for Neural Network :**

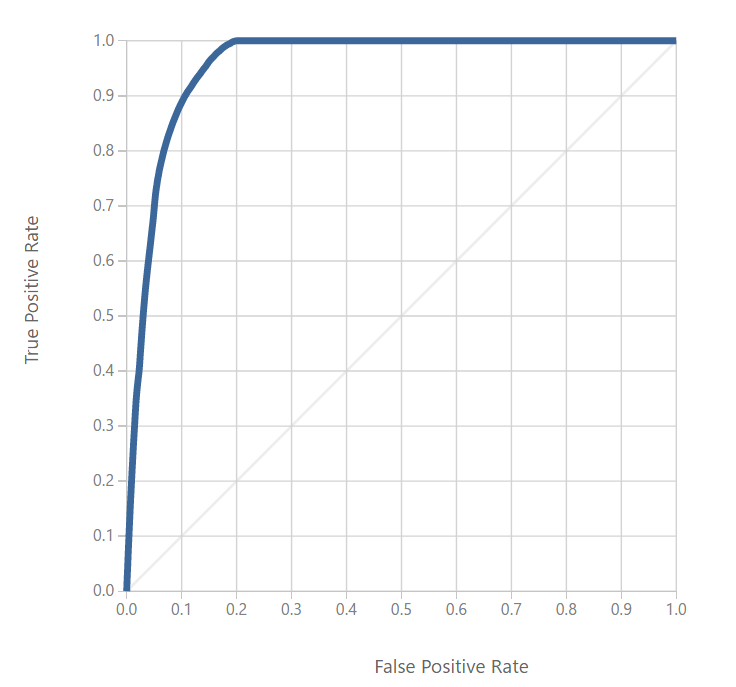


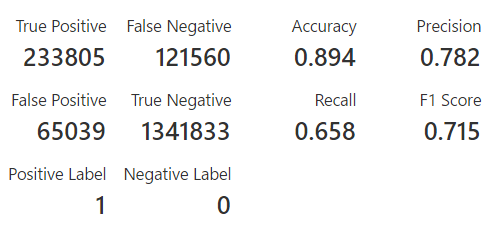
**Support Vector Machine:**

**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:





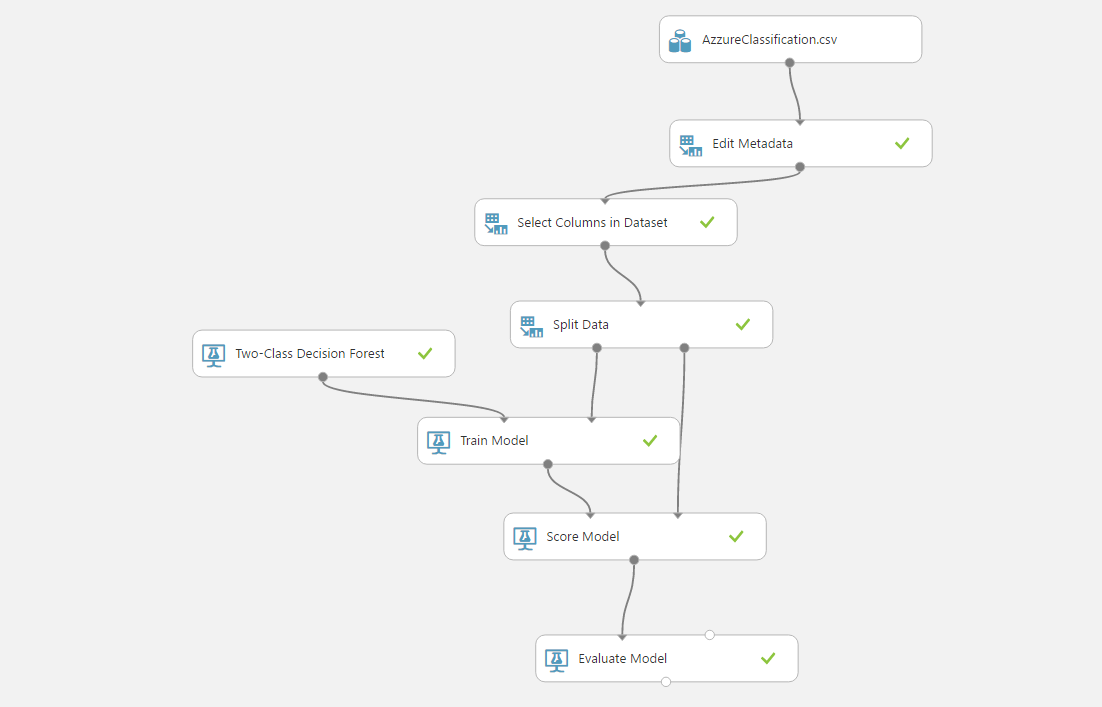
**Choosing Best Model:**

We looked at 4 values to make our choice :

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

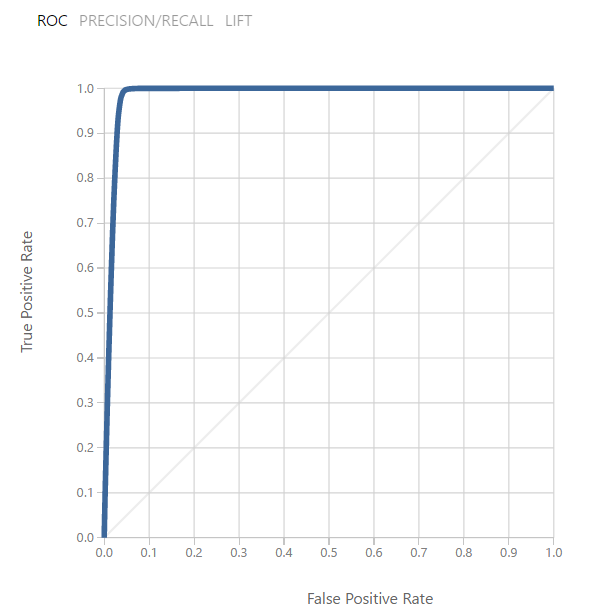
Based on the values above and looking at the ROC Curve , we chose Random Forest as our best model.

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

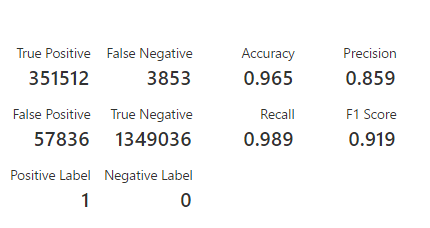


1. Used the input dataset created as a merge of Loan and Declined Dataset
2. We did not take Risk score as a parameter as Vantage score needs to be scaled and there is no fixed formulae and changed it using a formulae is not a good practice
3. As earlier stated , we used grade\_new based on Fico and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

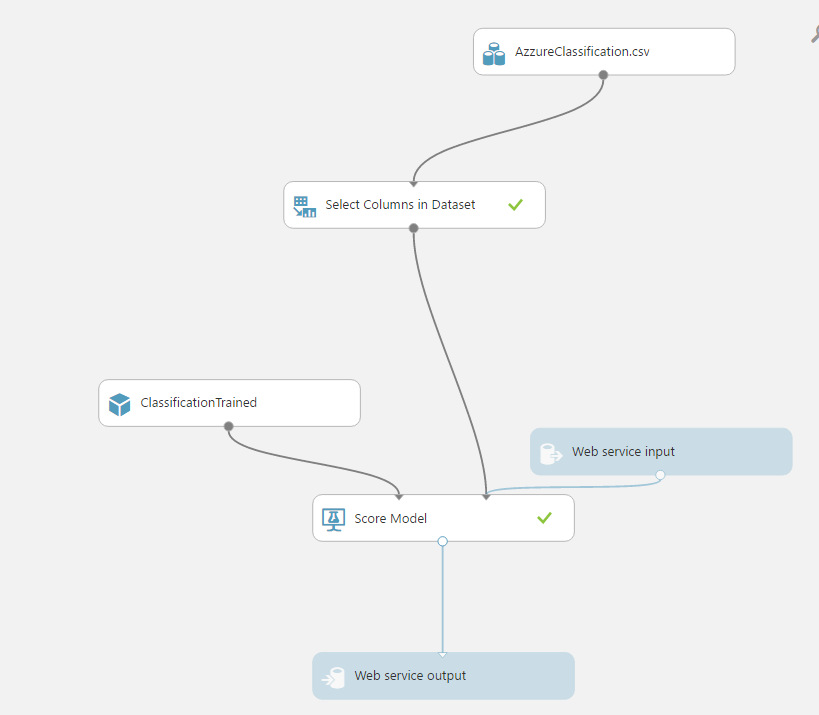
ROC Curve :



Stats on Azzure:



**Generating Rest API for Classification:**



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

