2016

INFO 7390 ADS Final Project Report

team 6

hina gandhi

nikita anand

Shivam Goel

# Introduction:

The project focuses on economic analysis of two countries India and USA. We have predicted certain indicators that are required for determining economic growth of a country. Based on the values of economic indicators, we have predicted whether we will face any recession in upcoming 10 years for both countries. Also, we have compared two economies.

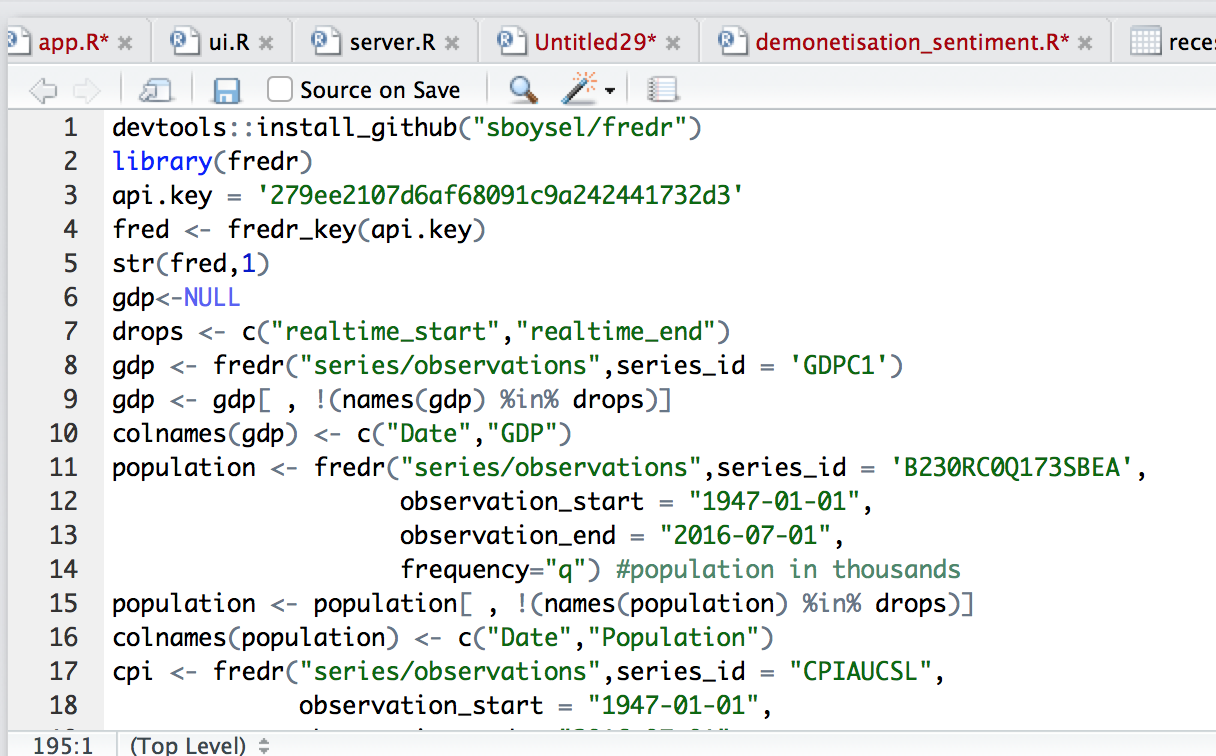
# STEPS:

* Data Scraping and wrangling
* Missing Values
* Parameter Selection
* Building Models
* Deploying Web Service
* Deploy application on cloud

# Data Scraping and Wrangling:

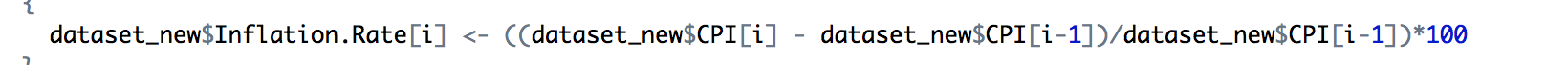
Firstly, we did research on the economy indicators. We determined what all indicators can impact an economy of a nation. We scraped the data from <https://fred.stlouisfed.org/>.

We have used library fredr and api key is created on fred.stlouis website. Fredr method is used to scrape data from website where we have mentioned the start date, end date and frequency = ”quarter”. We have taken data quarterly for all the selected indicators.



# Missing Values:

Some of the data was not available for inflation rate and unemployment rate. To fill in the missing inflation rates values, we have taken percent change of cpi between two quarters to determine inflation rate for the present quarter.



The missing unemployment rate was taken as mean of the column:

# C:\Users\hinag\Downloads\Screen Shot 2016-12-16 at 9.40.26 PM.png

# Creating new classification column “Recession”:

To have one classification column, we have added new column %change in gdp. The column is calculated by %change of gdp between two quarters.

There is another column created “Recession” based on the percent change of GDP. If %change in GDP is less than -1.1 then recession is “YES” for that year. This formula is determined after going through various research papers of economy.

# Parameter Selection:

We founded one new parameter selection package “Boruta” package. Boruta is a feature selection algorithm. Precisely, it works as a wrapper algorithm around Random Forest.

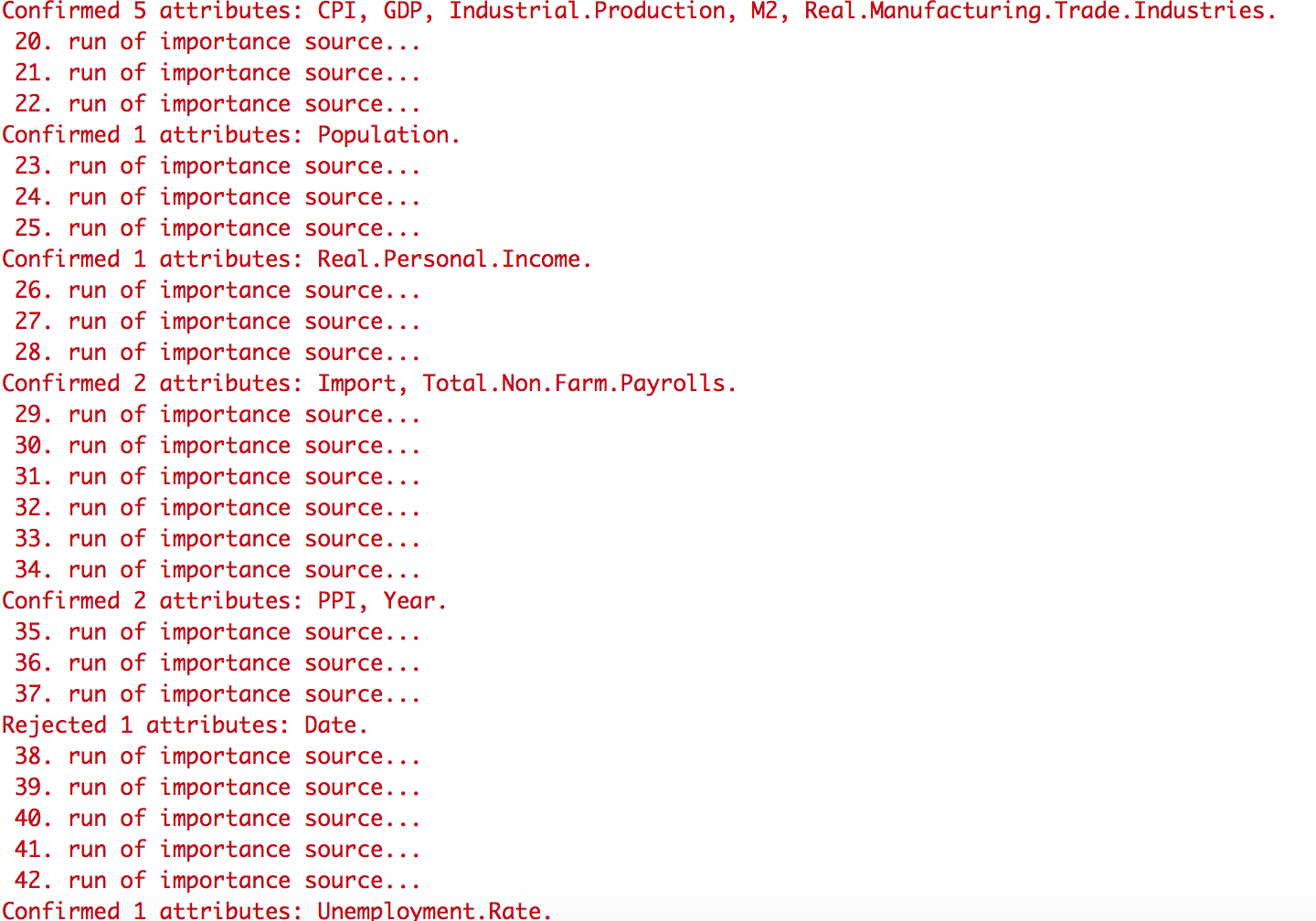
**How it works?**

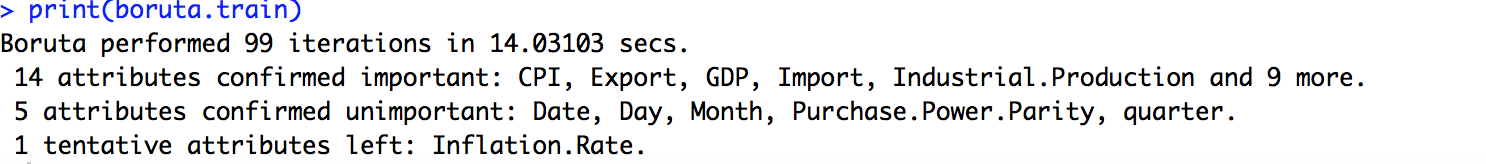
1. Firstly, it adds randomness to the given data set by creating shuffled copies of all features (which are called shadow features).
2. Then, it trains a random forest classifier on the extended data set and applies a feature importance measure (the default is Mean Decrease Accuracy) to evaluate the importance of each feature where higher means more important.
3. At every iteration, it checks whether a real feature has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z score than the maximum Z score of its shadow features) and constantly removes features which are deemed highly unimportant.
4. Finally, the algorithm stops either when all features gets confirmed or rejected or it reaches a specified limit of random forest runs.

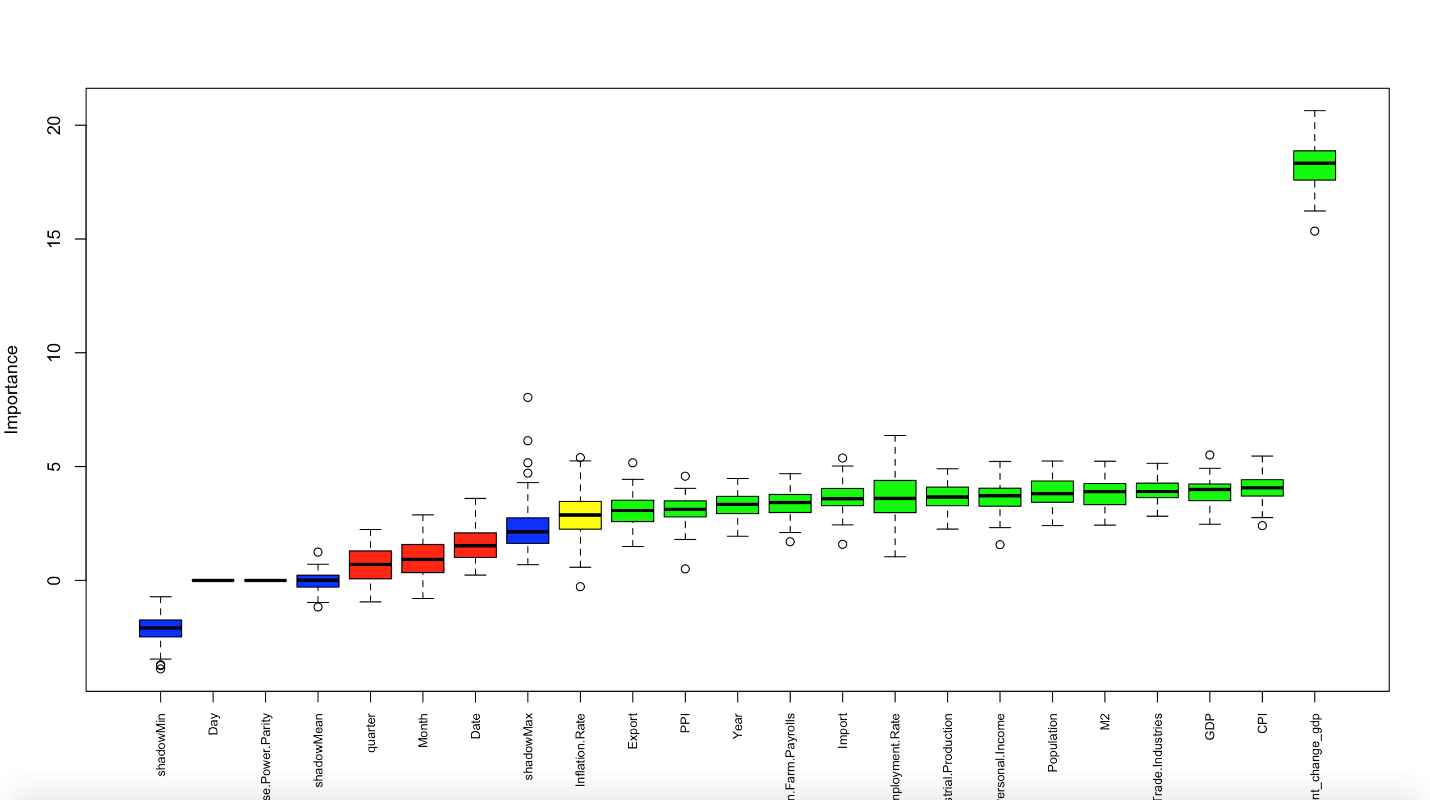
**How this package is different from other R packages?**

This package goes through iterations and in each iteration, it removes one non important variable.

This way going through 50-100 iterations, it selects most important variables from list of variables.







The graph is plotted where green boxes indicate the important parameters, yellow represents average indicators and blue, red shows those indicators that are do not have any importance in predicting recession.

Based on the above graph, we have selected those parameters that are green boxes as they have more significance.

# Predicting the economic indicators for 10 years:

As our dataset is time series.

# Azure ML Studio

Read the consolidated csv file.

Select the desired column GDP, Population, CPI, PPI Import,M2,Unemployment rate, Real Personal Income, Industrial Production, Recession, Export predicted/Classification values

**Producer Price index (PPI)**

The Producer Price index (PPI) is a family of indexes that measures the average change in selling prices received by domestic producers of goods and services over time. PPIs measure [price change](http://terms/p/price-change.asp) from the perspective of the seller. The PPI looks at three areas of production: industry-based, commodity-based and commodity-based final demand-intermediate demand.

**[Consumer Price Index](http://www.investopedia.com/terms/c/consumerpriceindex.asp" \t "_blank)(CPI)**

The [Consumer Price Index](http://www.investopedia.com/terms/c/consumerpriceindex.asp), that measure price change from the purchaser's perspective

The Consumer Price Index (CPI) is a measure that examines the [weighted average](http://www.investopedia.com/terms/w/weightedaverage.asp) of prices of a basket of [consumer goods](http://www.investopedia.com/terms/c/consumer-goods.asp) and services, such as transportation, food and medical care. It is calculated by taking [price changes](http://www.investopedia.com/terms/p/price-change.asp) for each item in the predetermined [basket of goods](http://www.investopedia.com/terms/b/basket_of_goods.asp) and averaging them. Changes in the CPI are used to assess price changes associated with the [cost of living](http://www.investopedia.com/terms/c/cost-of-living.asp); the CPI is one of the most frequently used statistics for identifying periods of [inflation](http://www.investopedia.com/terms/i/inflation.asp) or [deflation](http://www.investopedia.com/terms/d/deflation.asp).

**Gross domestic product (GDP)**

Gross domestic product (GDP) is the [monetary value](http://www.investopedia.com/terms/v/value.asp) of all the finished goods and services produced within a country's borders in a specific time period. Though GDP is usually calculated on an annual basis, it can be calculated on a [quarterly](http://www.investopedia.com/terms/q/quarter.asp) basis as well. GDP includes all private and public consumption, government outlays, investments and [exports](http://www.investopedia.com/terms/e/export.asp) minus [imports](http://www.investopedia.com/terms/i/import.asp) that occur within a defined territory. Put simply, GDP is a broad [measurement of a nation’s overall economic activity](http://stats.oecd.org/glossary/detail.asp?ID=1163)

**Money Supply (M2)**

M2 is a measure of the [money supply](http://www.investopedia.com/terms/m/moneysupply.asp) that includes all elements of M1 as well as "near money." M1 includes cash and checking [deposits](http://www.investopedia.com/terms/d/deposit.asp), while [near money](http://www.investopedia.com/terms/n/near-money.asp) refers to savings deposits, [money market](http://www.investopedia.com/terms/m/moneymarket.asp) securities, [mutual funds](http://www.investopedia.com/terms/m/mutualfund.asp) and other [time deposits](http://www.investopedia.com/terms/t/timedeposit.asp). These assets are less liquid than M1 and not as suitable as exchange mediums, but they can be quickly converted into cash or checking deposits.

M2 is a broader money classification than [M1](http://www.investopedia.com/terms/m/m1.asp), because it includes assets that are highly liquid but are not cash. A consumer or business typically doesn't use savings deposits and other non-M1 components of M2 when making purchases or paying bills, but it could convert them to cash in relatively short order. [M1](http://www.investopedia.com/video/play/m1/) and M2 are closely related, and economists like to include the more broadly defined definition for M2 when discussing the money supply, because modern economies often involve transfers between different account types. For example, a business may transfer $10,000 from a [money market account](http://www.investopedia.com/terms/m/moneymarketaccount.asp) to its [checking account](http://www.investopedia.com/terms/c/checkingaccount.asp). This transfer would increase M1, which doesn’t include money market funds, while keeping M2 stable, since M2 contains money market accounts.

**Purchasing Power Parity (PPP)**

Purchasing Power Parity (PPP) is an economic theory that compares different countries' currencies through a market "basket of goods" approach. According to this concept, two currencies are in equilibrium or at par when a market basket of goods (taking into account the exchange rate) is priced the same in both countries.

This is how the relative version of PPP is calculated:

urchasing Power Parity (PPP)

Where:

"S" represents exchange rate of [currency](http://www.investopedia.com/terms/c/currency.asp) 1 to currency 2

"P1" represents the cost of good "x" in currency 1

"P2" represents the cost of good "x" in currency 2

**Inflation**

Inflation is the rate at which the general level of prices for goods and services is rising and, consequently, the [purchasing power](http://www.investopedia.com/terms/p/purchasingpower.asp) of currency is falling. [Central banks](http://www.investopedia.com/terms/c/centralbank.asp) attempt to limit inflation, and avoid [deflation](http://www.investopedia.com/terms/d/deflation.asp), in order to keep the [economy](http://www.investopedia.com/terms/e/economy.asp) running smoothly.

**Real Person Income**

Real income refers to the income of an individual or group after taking into consideration the effects of [inflation](http://www.investopedia.com/terms/i/inflation.asp) on [purchasing power](http://www.investopedia.com/terms/p/purchasingpower.asp). For example, if you receive a 2% salary increase over the previous year and inflation for the year is 1%, then your real income only increases by 1%. Conversely, if you receive a 2% raise in salary and inflation is at 3%, then your real income shrinks by 1%.

**Unemployment rate**

Unemployment rate is defined most basically as the percentage of the total labor force that is unemployed but actively seeking employment and willing to work.  
Unemployment is measured in a number of ways in the U.S.: labor force sample surveys, Social Insurance statistics, which are numbers drawn from the number of people drawing unemployment benefits and other welfare programs, and employment office statistics which measure the number of people who come into unemployment offices looking for work. When calculating the official unemployment rate, the Bureau of Labor Statistics consider all the information gathered by these methods to provide a holistic picture of the labor market and the challenges facing workers. 

**Industrial Production**

Industrial production figures are based on the monthly raw volume of goods produced by industrial firms such as factories, mines and electric utilities in the United States. Also included in the industrial production figures are the businesses of newspaper, periodical and book publishing, traditionally labeled as manufacturing.

**Total Nonfarm Payroll**

All Employees: Total Nonfarm, commonly known as Total Nonfarm Payroll, is a measure of the number of U.S. workers in the economy that excludes proprietors, private household employees, unpaid volunteers, farm employees, and the unincorporated self-employed. This measure accounts for approximately 80 percent of the workers who contribute to Gross Domestic Product (GDP).  
This measure provides useful insights into the current economic situation because it can represent the number of jobs added or lost in an economy. Increases in employment might indicate that businesses are hiring which might also suggest that businesses are growing. Additionally, those who are newly employed have increased their personal incomes, which means (all else constant) their disposable incomes have also increased, thus fostering further economic expansion.

# Azure Modules

**Split Data -** To split a dataset into two equal parts, just addthe [Split Data](https://msdn.microsoft.com/en-us/library/azure/dn905969) module after the dataset without no other changes. By default, the module splits the dataset in two equal parts. For data with an odd number of rows, the second output gets the remainder.

**Train Model** - Training a classification or regression model is a kind ofsupervisedmachine learning. That means you must provide a dataset that contains historical data from which to learn patterns. The data should contain both the outcome you are trying to predict, and related factors (variables). The machine learning model uses the data to extract statistical patterns and build a model.

When you configure Train Model**,** you must also connect an already configured model, such as a regression algorithm, decision tree model, or another machine learning module.

**Score Model -** Score Model is used to generate predictions using a trainedclassification or regression model. The predicted value can be in many different formats, depending on the model and your input data: If you are using a classification model to create the scores, Score Model outputs a predicted value for the class, as well as the probability of the predicted value. For regression models, Score Model generates just the predicted numeric value.

* + **Evaluate Model** - Evaluate Model is used to measure the accuracy of a trainedclassification model or regression model. You provide a dataset containing scores generated from a trained model, and the Evaluate Model module computes a set of industry-standard evaluation metrics. The metrics returned by Evaluate Model depend on the type of model that you are evaluating

# Classification Models

We build three classification model to classify Predicted Values into Yes (recession) and No (not a recession) and compare those models among them to find the best classification algorithm as per our model

# Two class Decision Forest

The decision forest algorithm is an ensemble learning method for classification. The algorithm works by building multiple decision trees and then voting on the most popular output class. Voting is a form of aggregation, in which each tree in a classification decision forest outputs a non-normalized frequency histogram of labels.

# Two class Neural network

A neural network is a set of interconnected layers, in which the inputs lead to outputs by a series of weighted edges and nodes. The weights on the edges are learned when training the neural network on the input data. The direction of the graph proceeds from the inputs through the hidden layer, with all nodes of the graph connected by the weighted edges to nodes in the next layer. Most predictive tasks can be accomplished easily with only one or a few hidden layers.

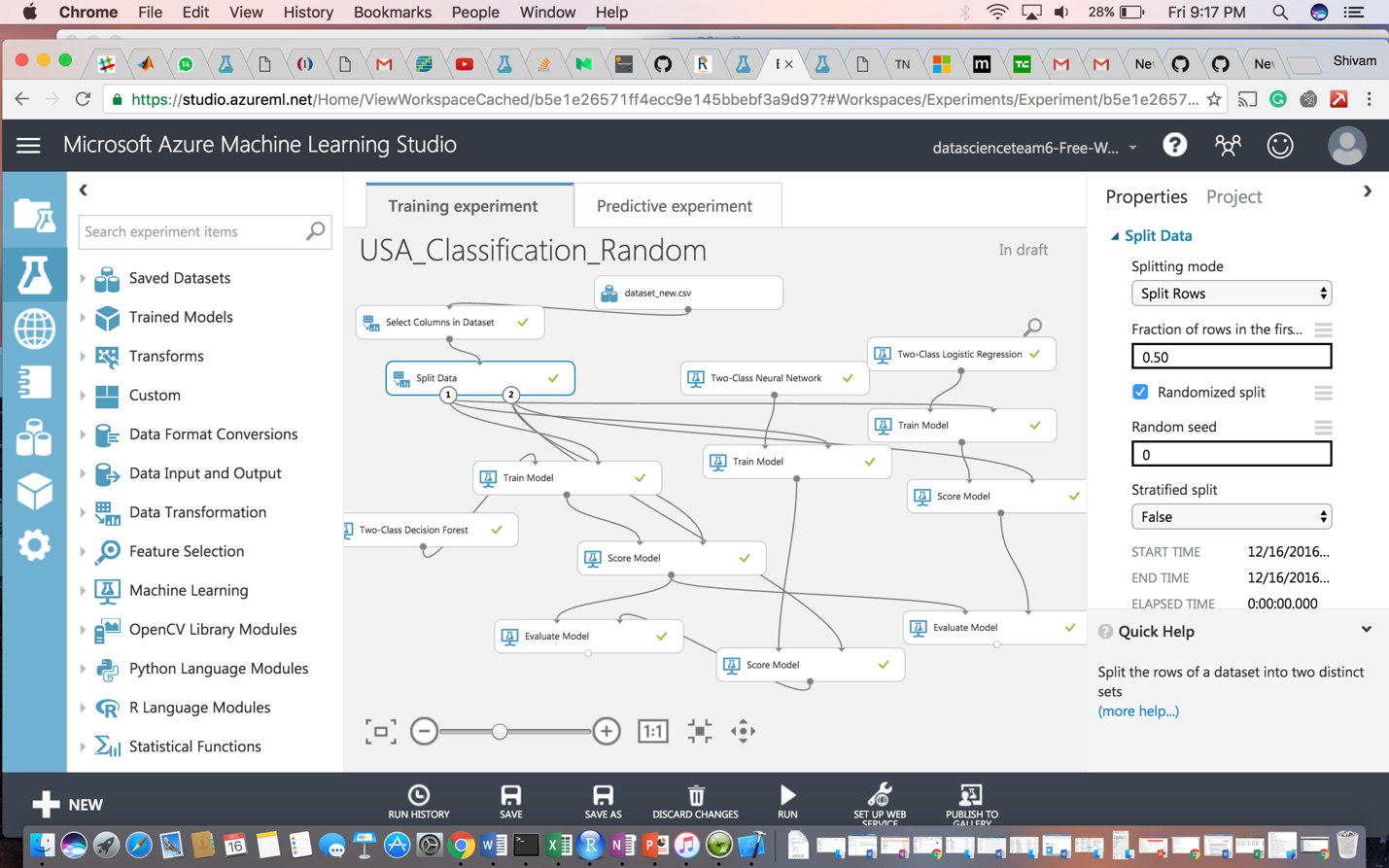
Recent research has shown that deep neural networks (DNN) can be very effective in complex tasks such as image or speech recognition, in which successive layers are used to model increasing levels of semantic depth. To compute the output of the network for any given input, a value is calculated for each node in the hidden layers and in the output layer. For each node, the value is set by calculating the weighted sum of the values of the nodes in the previous layer and applying an activation function to that weighted sum.

# Two class Logistic Regression

Logistic regression is a well-known method in statistics that is used to predict the probability of an outcome, and is especially popular for classification tasks. The algorithm predicts the probability of occurrence of an event by fitting data to a logistic function.

Logistic regression requires numeric variables. Therefore, when you use categorical columns as variable, Azure Machine Learning converts the values to an indicator array internally.

**Azure Model:**

****

**ROC curve**

The ROC curve plots the pairs {sensitivity, 1-specificity} as the cutoff value increases from 0 and 1

* **Sensitivity** (also called the **true positive rate**, or the **recall** in some fields) measures theproportion of positives that are correctly identified (e.g., the percentage of sick people who are correctly identified as having the condition).
* **Specificity** (also called the **true negative rate**) measures the proportion of negatives thatare correctly identified

as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

• Better performance is reflected by curves that are closer to the top left corner

**Confusion Matrix and other performance metrics:**

The following statistics are shown for our model:

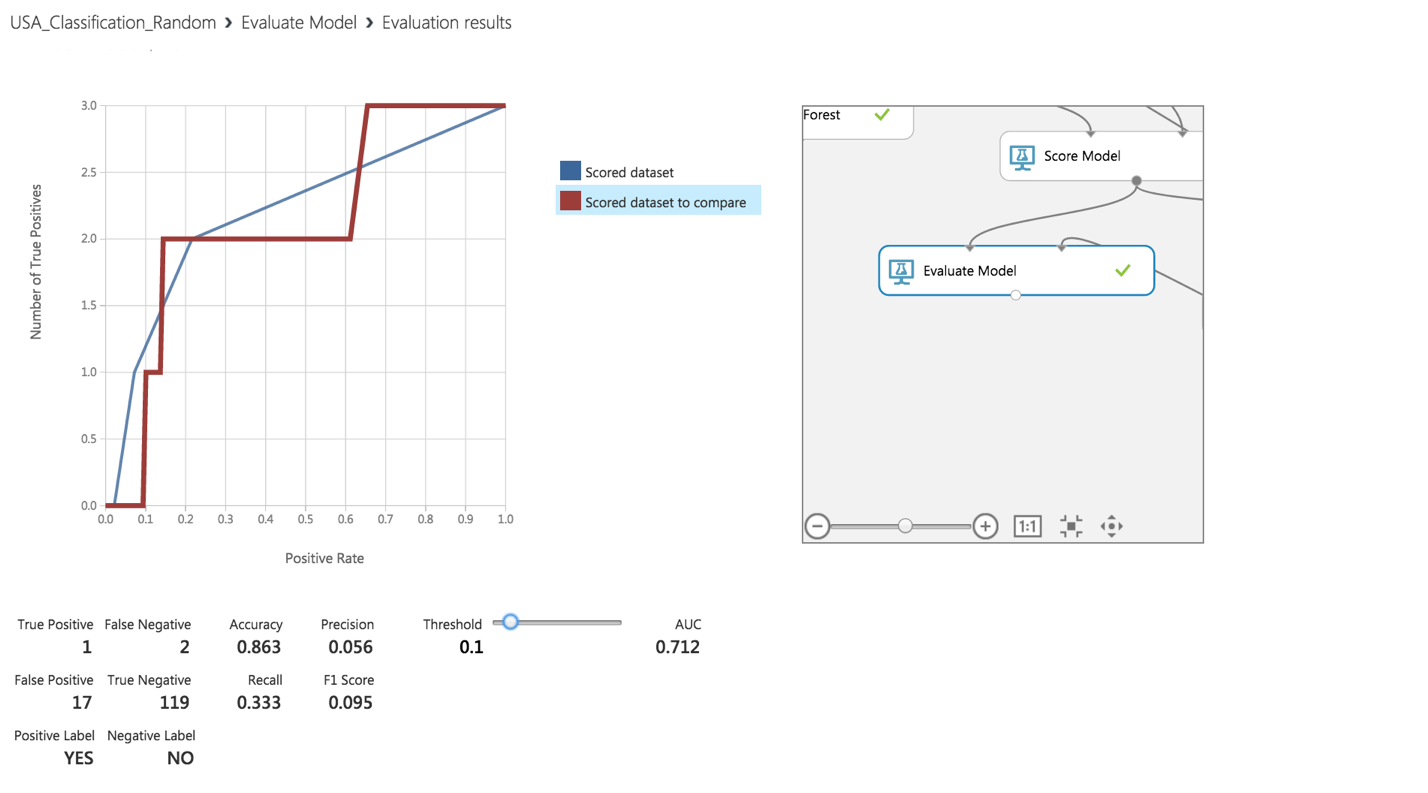
* **Mean Absolute Error** (MAE): The average of absolute errors (an*error*is thedifference between the predicted value and the actual value).
* **Root Mean Squared Error** (RMSE): The square root of the average of squared errorsof predictions made on the test dataset.
* **Relative Absolute Error**: The average of absolute errors relative to the absolutedifference between actual values and the average of all actual values.
* **Relative Squared Error**: The average of squared errors relative to the squareddifference between the actual values and the average of all actual values.
* **Coefficient of Determination**: Also known as the **R squared value**, this is a statisticalmetric indicating how well a model fits the data.

For each of the error statistics, smaller is better. A smaller value indicates that the predictions more closely match the actual values. For **Coefficient of Determination**, the closer its value is to one (1.0), the better the predictions.

Comparing between Random forest and Neural Network classification

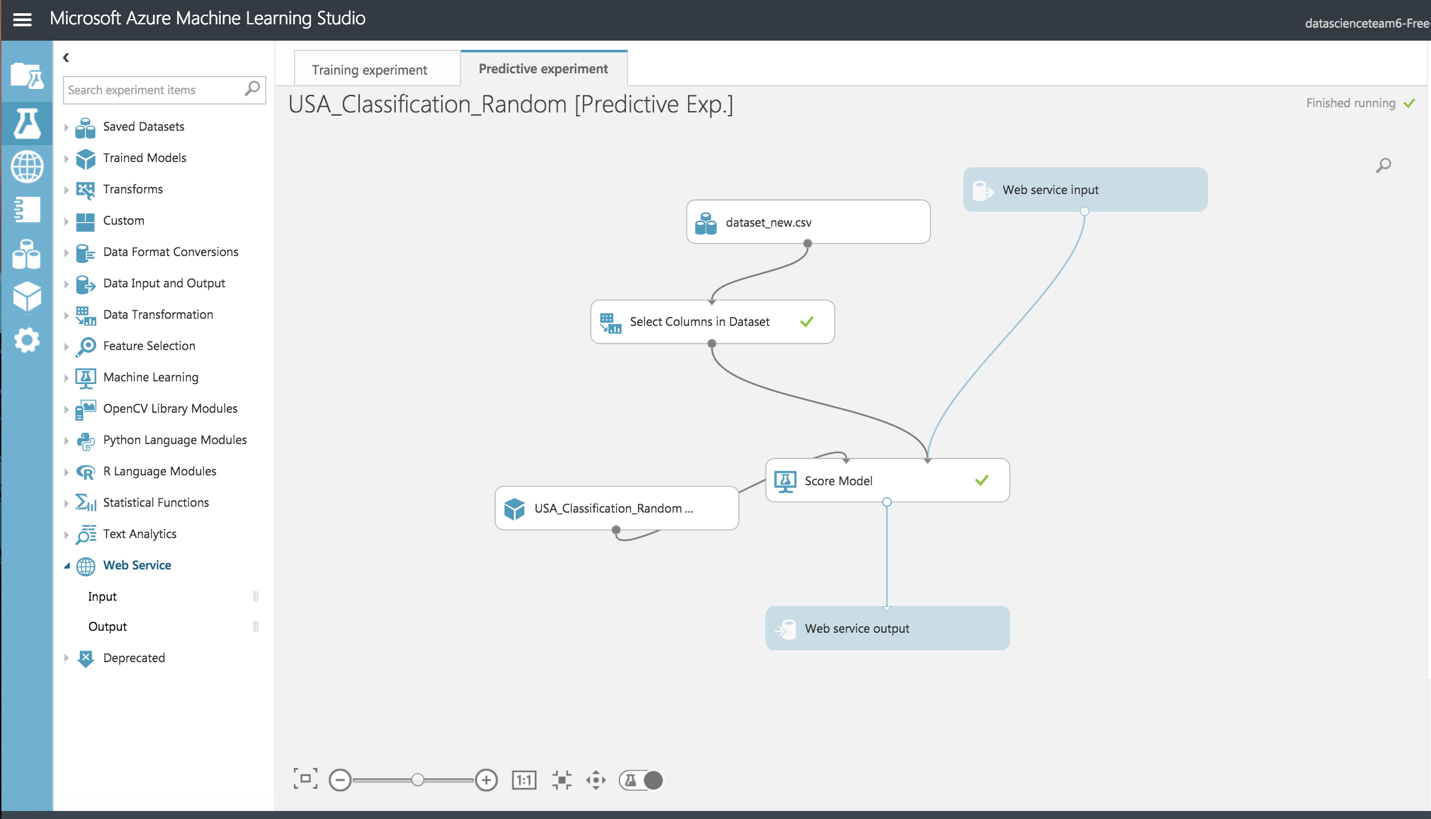


Comparing between Random forest and Logistic Regression classification

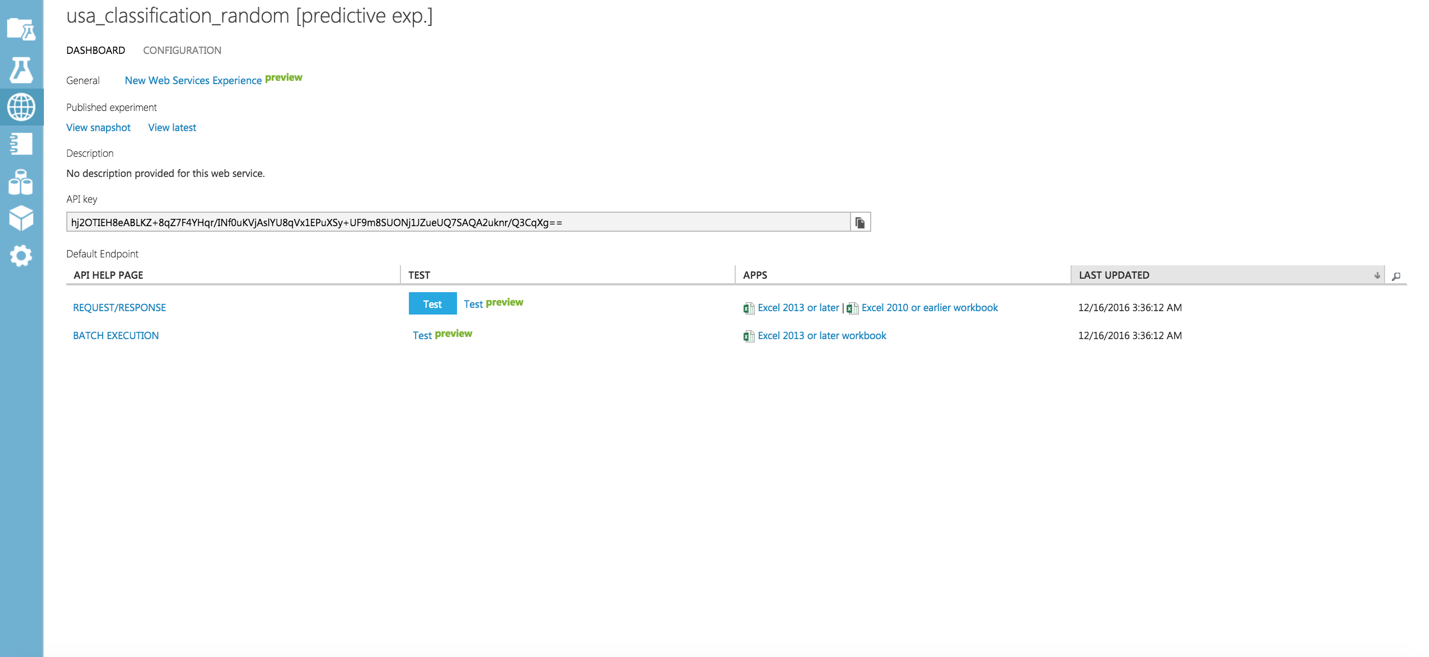


**Web Service**

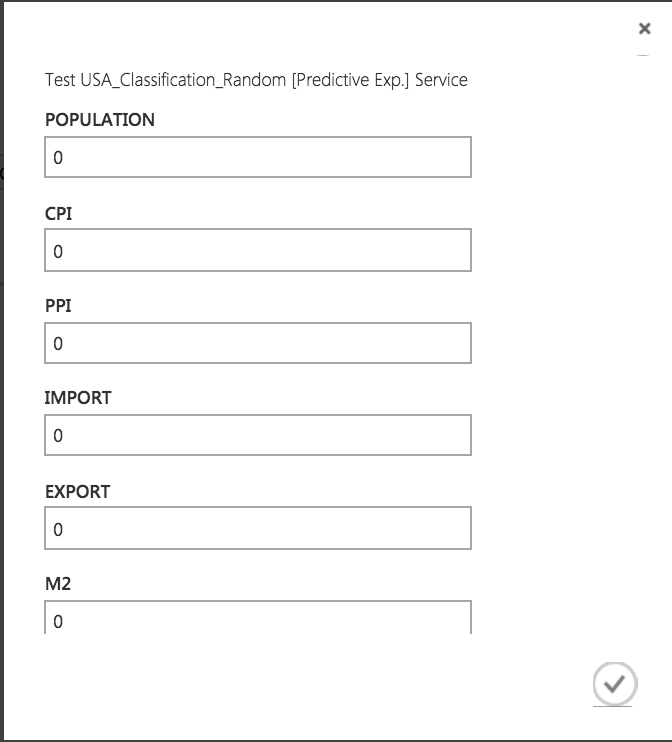
* Once the classification model is ready, we set up **Web Service**.
* The model we trained is saved as a single **Trained Model** module into the module palette to the left of the experiment canvas (you can find it under **Trained Models**)
* Then we added the saved trained model back into the experiment.
* **Web service input** and **Web service output** modules are added.



* Now run the model and publish the web service



* On running the web service, we get the following form which can be used to invoke the web service and do prediction.

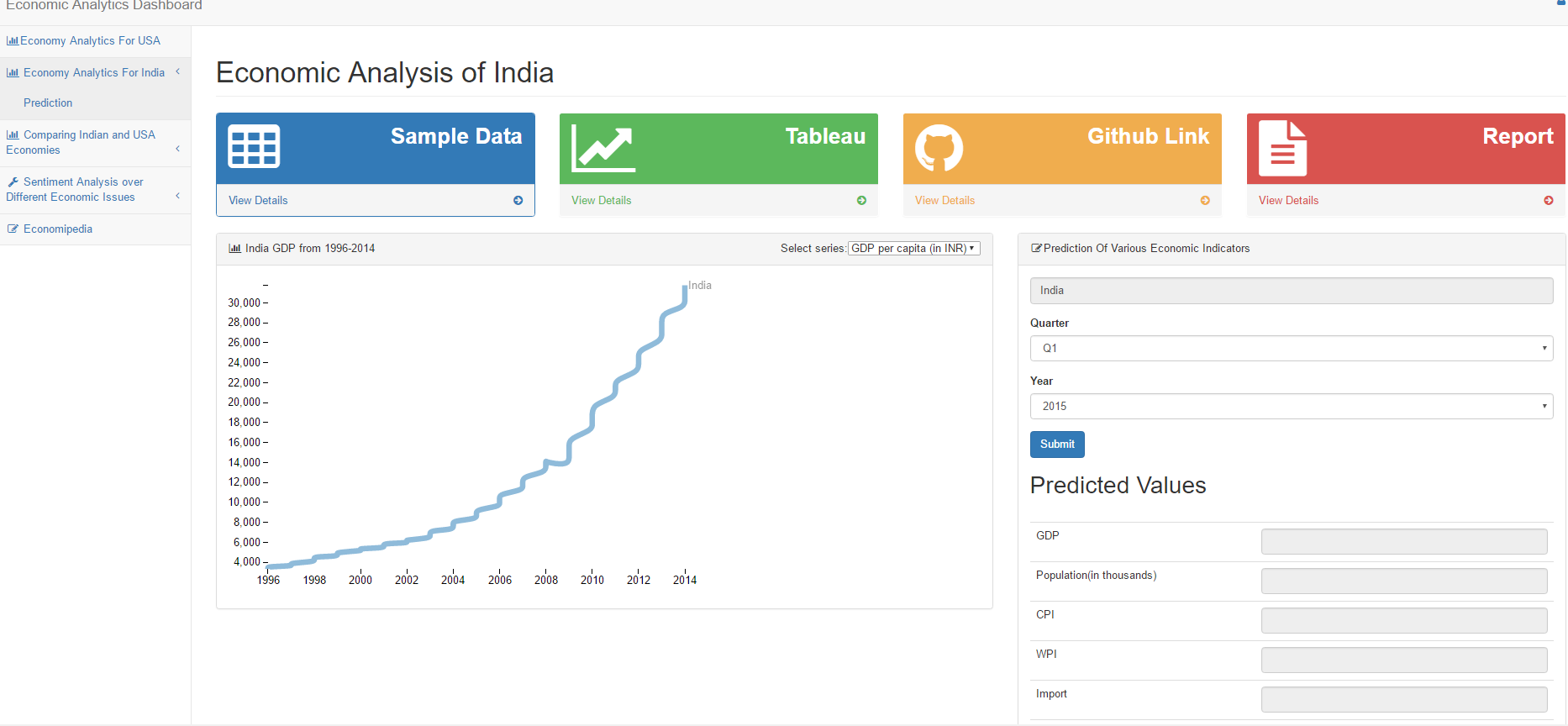


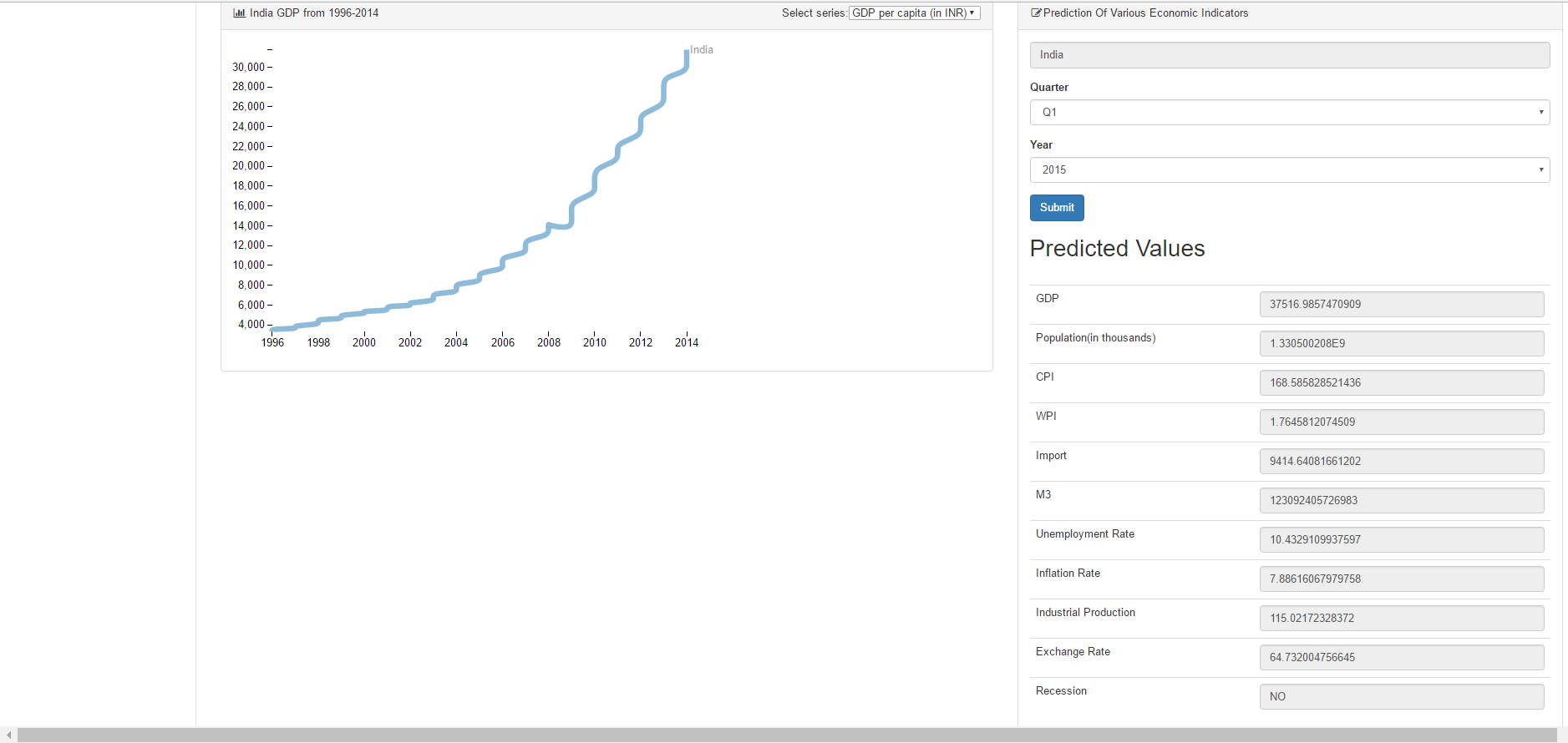
**Conclusion:** Thus from above it’s clear that the best model amongthe classificationmodels is Two class Decision Tree as it has high accuracy rate at 86.3%. Also, the Area Under Curve (AUC) is highest in Two class Decision Tree.

# Web Application

AWS Link: http://sample-env-2.zq3wjanf63.us-west-2.elasticbeanstalk.com/

We created a MVC Java application on Netbeans and deployed the application on Amazon AWS

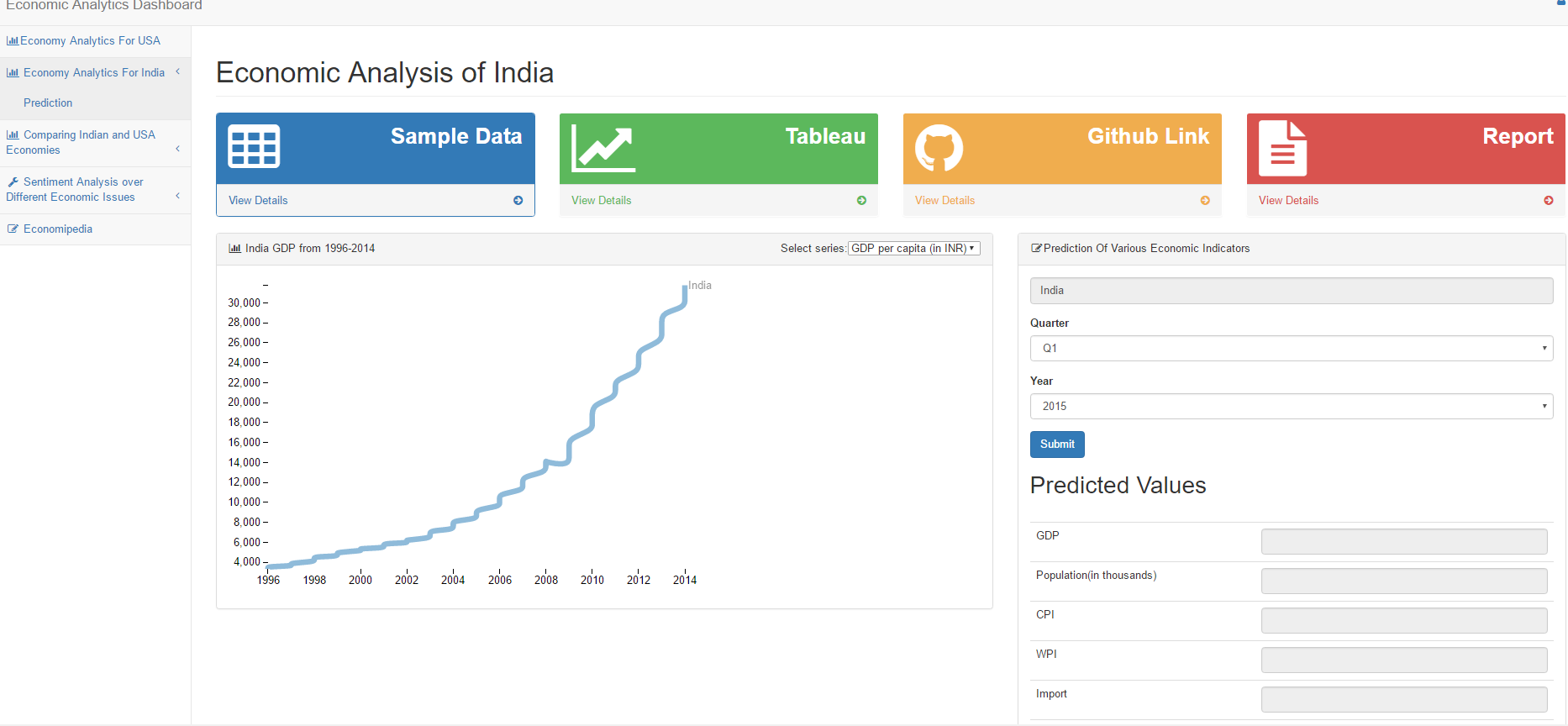




**We have created Tableau Dashboard on the Links at the heading of this website**

**This link will take you to tableau public dashboard**

**You can also find various relevant tabs as below for Report, Presentation, Github Link, Sample Data**

****