**Data Mining for Stock Prediction**Predicting the next day Stock Price

WQD7005 – Data Mining Assignment Project  
University of Malaya

Name - Vikas Mann  
Matic Number – WQD180051  
01-Jun-2019

<https://github.com/vikas-mann/WQD7005>

Contents

[1 Introduction, Research Goal & Objective 3](#_Toc10410937)

[1.1 Introduction 3](#_Toc10410938)

[1.2 Research Goal 3](#_Toc10410939)

[1.3 Objective 3](#_Toc10410940)

[2 DataSet preparation 4](#_Toc10410941)

[2.1 Data Collection 4](#_Toc10410942)

[2.2 Data Management 4](#_Toc10410943)

[2.3 Database Schema 5](#_Toc10410944)

[2.4 Data Analytics Software 6](#_Toc10410945)

[2.5 Load Data 7](#_Toc10410946)

[3 Descriptive Data Analysis 8](#_Toc10410947)

[3.1 Analysis using SAS Enterprise Miner 9](#_Toc10410948)

[3.2 Analysis using Tableau 10](#_Toc10410949)

[4 Prediction Models 13](#_Toc10410950)

[4.1 Feature Selection 13](#_Toc10410951)

[4.2 Linear Regression (Model 1) 13](#_Toc10410952)

[4.3 Neural Networks (Model 2) 14](#_Toc10410953)

[4.4 Decision Tree (Model 3) 15](#_Toc10410954)

[4.5 Gradient Boosting (Model 4) 16](#_Toc10410955)

[5 Model Comparisons and Validation 17](#_Toc10410956)

[6 Analysis & Recommendations 18](#_Toc10410957)

[6.1 Observations 18](#_Toc10410958)

[6.2 Publication 18](#_Toc10410959)

# Introduction, Research Goal & Objective

## Introduction

The stock market prediction is the process of predicting future value of a company listed on an exchange. The success of this prediction can result in achieving significant profits for the market investments.

## Research Goal

The purpose of the research is to understand the impact of different data sets on a stock price and perform different data analysis e.g. descriptive data analysis, exploratory data analysis etc. to find correlation between different variables and seeking patterns. Ultimately a prediction model will be developed for prediction of Stock Price.

During the process of achieving the ultimate goal of building the most accurate prediction model, different alternative models will be built and compared.

The prediction model will be developed using the training data set and will be validated against the testing data set.

## Objective

The ultimate objective of this assignment is to predict a Stock Price, based on the historical stock data. The historical stock data, along with the supporting data e.g. Commodities price, major forex, major stock indices and social data e.g. Twitter is scrapped from internet on daily basis.

An end to end Data Mining process is developed for Stock Price prediction, by capturing the data using web crawling methods, performing data analysis and building Machine Learning model for prediction.

# DataSet preparation

## Data Collection

Data collection is done using a Python script using the web scraping methods. Below are the data sources –

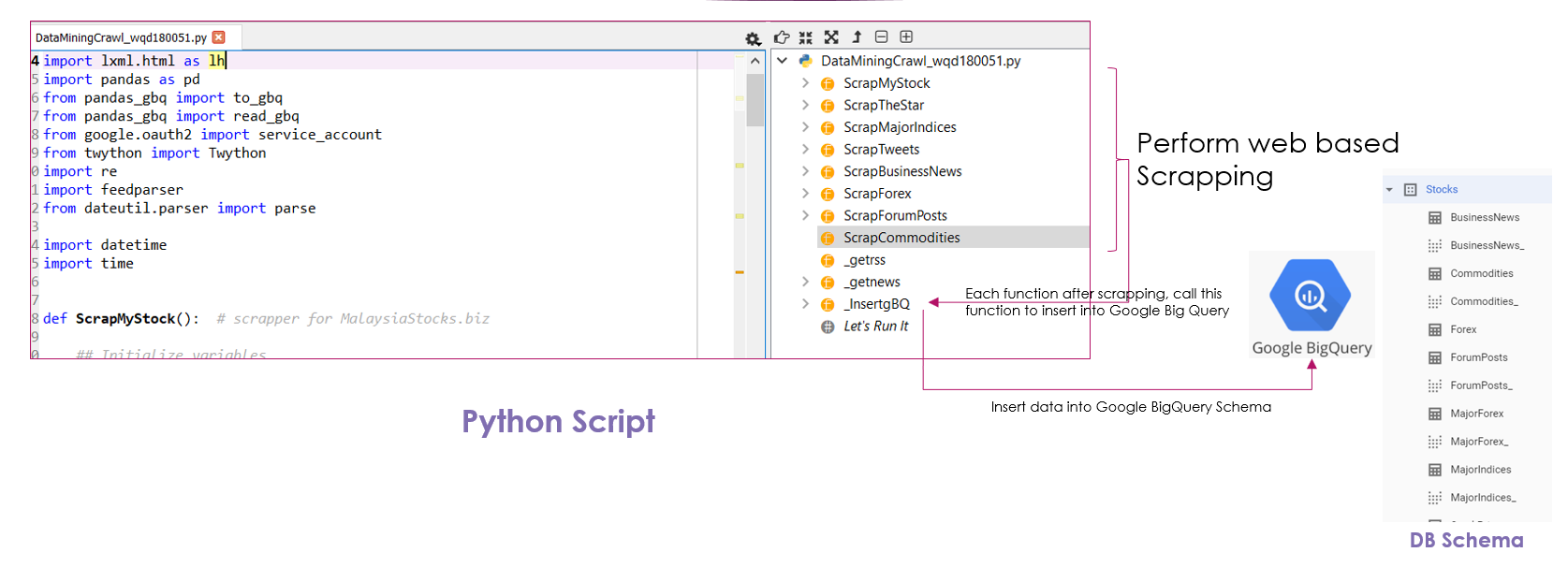
* Stock Price for all KLSE listed stocks – the data is collected using Web Crawler methods
  + Malaysiastock.biz
  + Star Online
  + Yahoo Finance
* Stock News – data collected from RSS feed
  + Malaysiakini
  + The Star Business
* Stock Tweets – Twitter API is used to scrap the tweets
  + Twitter
* Stock Forum Posts – Web Crawler methods
* Major Forex – Web Crawler methods
  + XE.com
* Commodities – Web Crawler methods
  + Gold, Crude Oil etc.

## Data Management

After scrapping the daily data, the data is stored in Google BigQuery platform. Google BigQuery is a fast, highly scalable, cost-effective, and fully managed cloud data warehouse for analytics, with built-in machine learning.

The data ingestion is performed directly from Python script using the Google BigQuery library in Python i.e. pandas\_gbq. Two important functions are used for ingestion and retrieval of data from Google BigQuery.

* Read\_gbq – load data from Google Big Query
* To\_gbq – write a dataframe to Google Big Query Table



*Figure 1: Python Script for data collection and storage.*

Below functions are developed for the scrapping of data from internet –

* ScrapMyStock
* ScrapTheStar
* ScrapMajorIndices
* ScrapTweets
* ScrapBusinessNews
* ScrapForex
* ScrapForumPosts
* ScrapCommodities

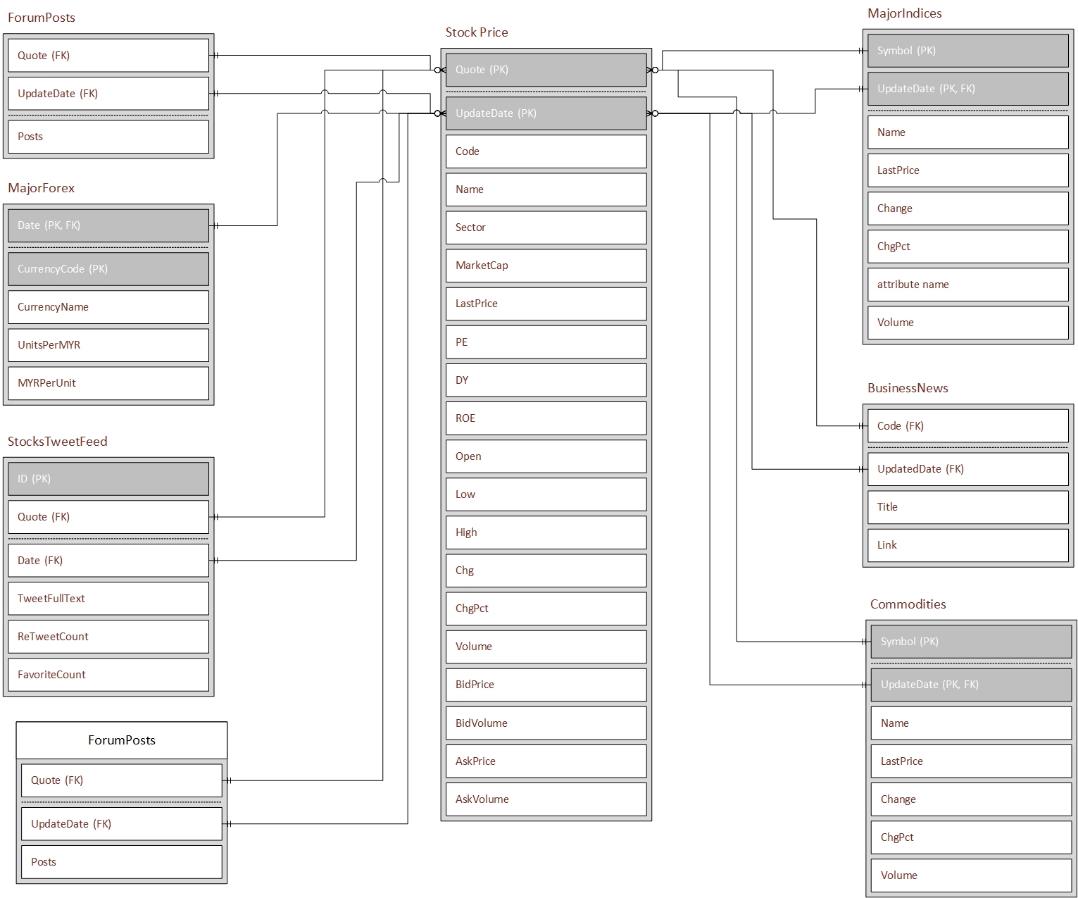
A function is developed for ingestion of scrapped data from these functions, directly into Google BigQuery defined Schema.

* \_InsertgBQ

## Database Schema

The database schema defined in Google BigQuery consists of 7 tables –

* StockPrice (Primary)
* BusinessNews
* ForumPosts
* StocksTweetFeed
* MajorForex
* MajorIndices
* Commodities



*Figure 2: Database Schema.*

## Data Analytics Software

Below software and frameworks are used to perform the Data Analysis

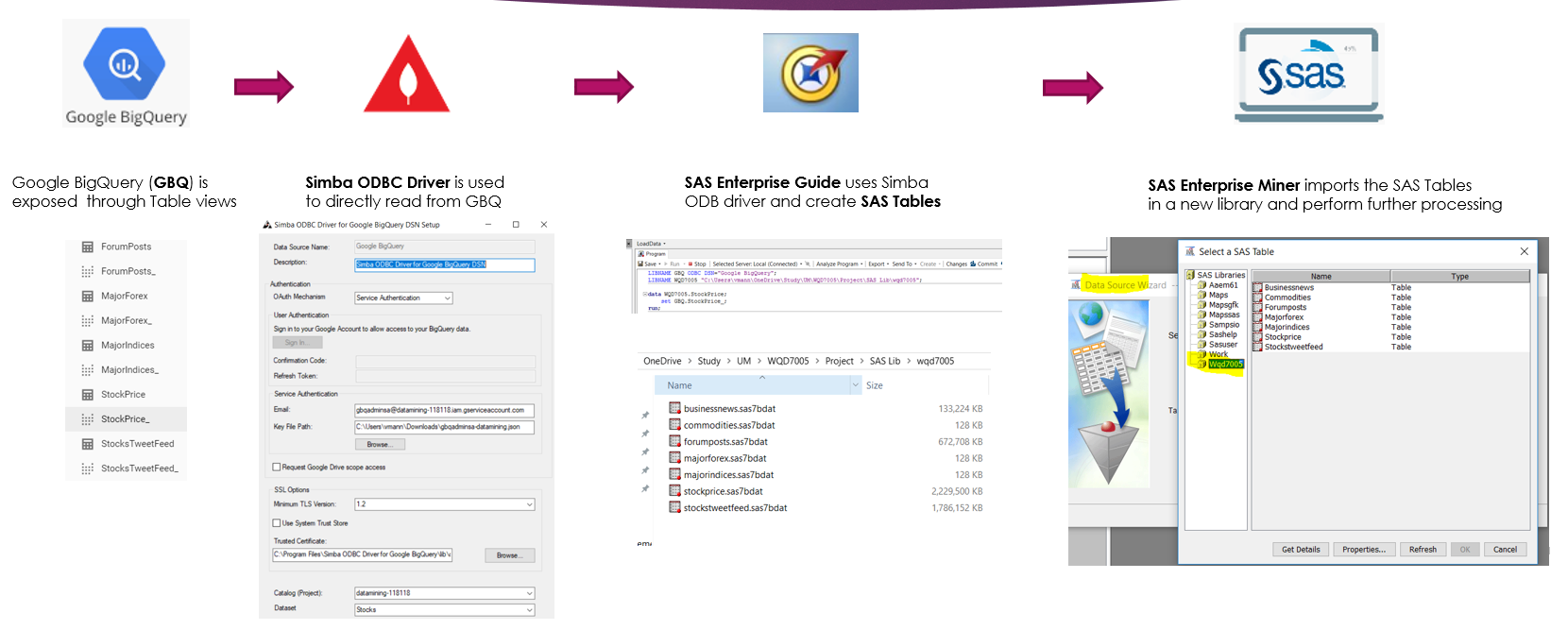
* SAS Enterprise Guide
* SAS Enterprise Miner
* Python
* Spyder IDE
* Anaconda
* Tableau

## Load Data

The data analysis and the prediction modelling are targeted to be done in SAS Enterprise Miner software. Hence, it is required to bring this data into SAS ecosystem.

The data stored in Google BigQuery is ingested into SAS Enterprise Miner as per following steps –

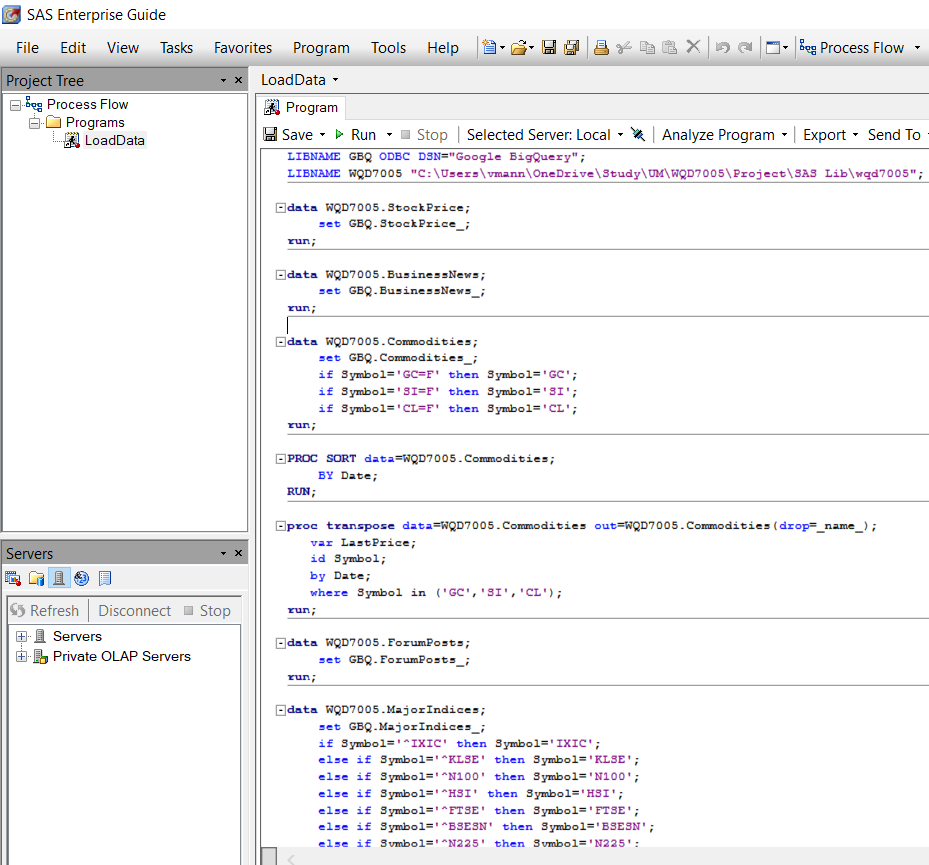
* Google BigQuery database tables are exposed through Table views to expose the relevant variables only.
* Simba ODBC driver is used to directly read data from Google BigQuery
* A program is developed in SAS Enterprise Guide to convert the relational tables in SAS tables. This program uses Simba ODBC driver for directly accessing the data
* The imported SAS tables are added into SAS enterprise miner software by creating a new SAS library.



*Figure 3: Data Ingestion process in SAS.*

**Steps in SAS Enterprise Guide to import and process Data**

* Create a new library for ODBC Source
* Create a new library for imported SAS Tables
* Import from ODBC Source library to new SAS Tables Library
* Perform Sorting of Data
* Perform Transposing of Data for below tables
  + Commodities
  + MajorIndices
  + MajorForex

*.*

*Figure 4: Data Ingestion program in SAS Enterprise Guide.*

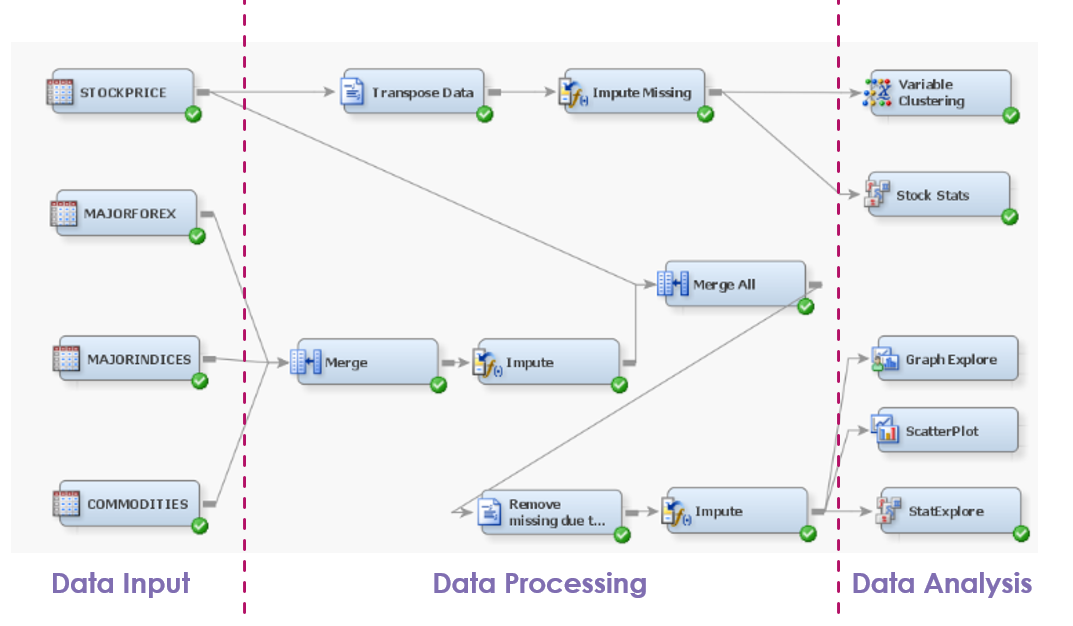
**Steps in SAS Enterprise Miner**

* Import the new Library created during SAS Enterprise Guide Processing
* Create new Data Sources from the new Imported library
  + COMMODITIES
  + MAJORFOREX
  + MAJORINDICES
  + STOCKPRICE
* Create new Diagram for Data Processing and Analysis

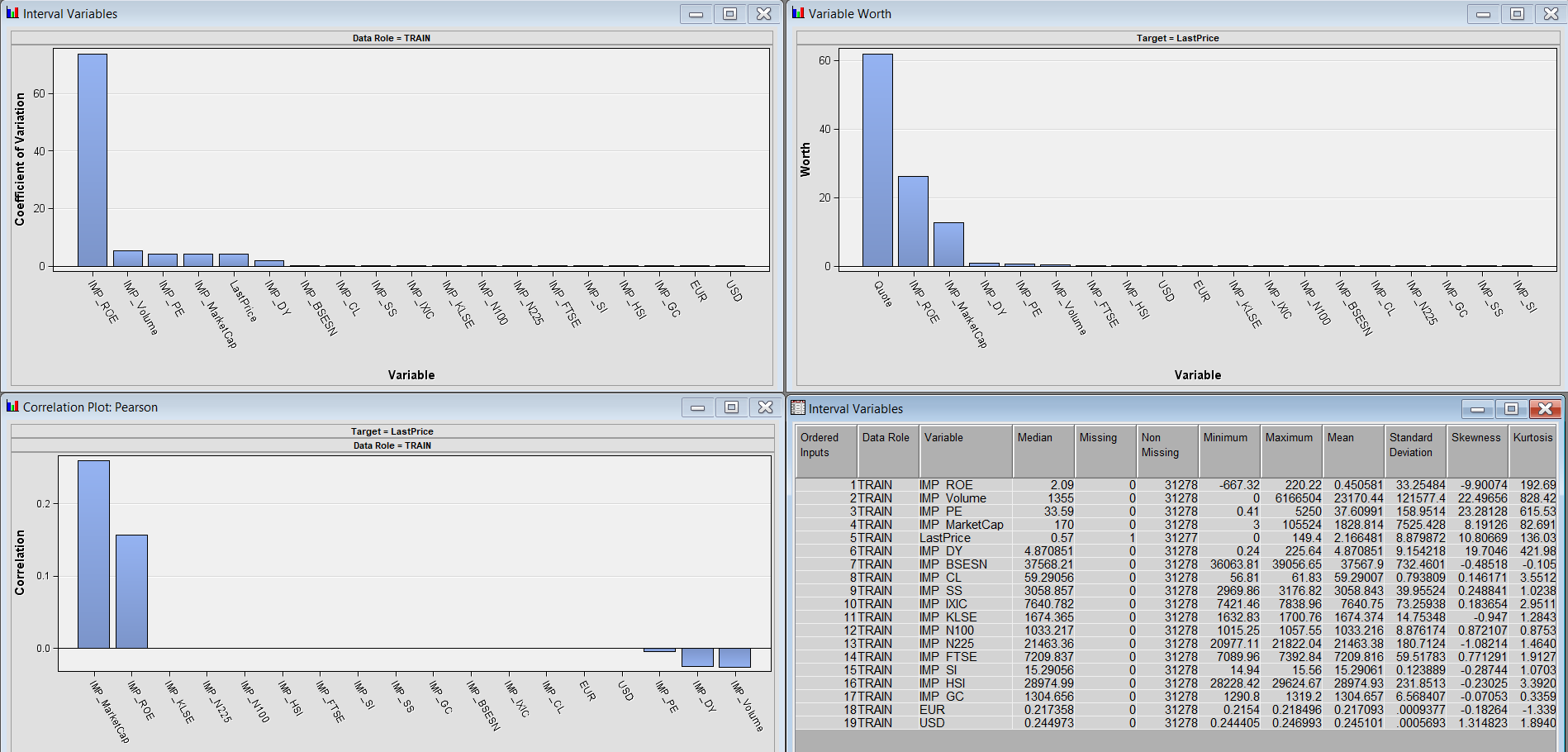
# Descriptive Data Analysis

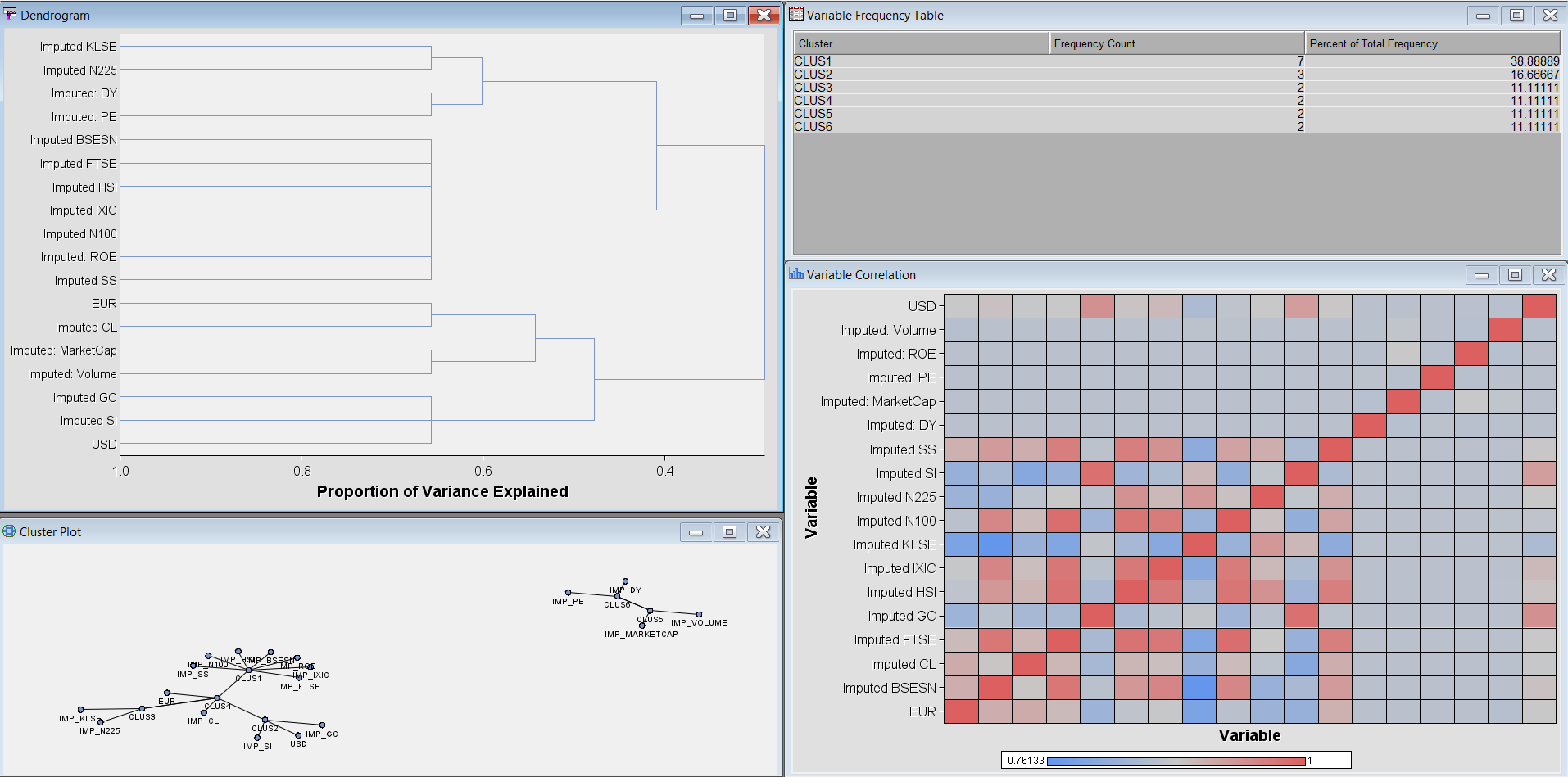
## Analysis using SAS Enterprise Miner

The data imported into SAS enterprise miner software are processed by defining a process workflow. This analysis ranges from Variable clustering, Covariation/Correlation analysis and variables statistical exploration.



*Figure 5: Analysis in SAS Enterprise Miner.*

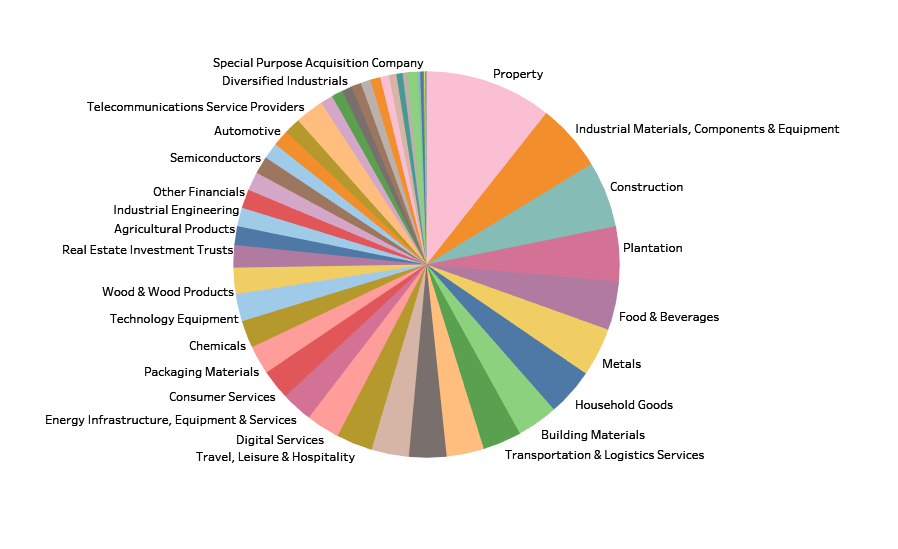


\

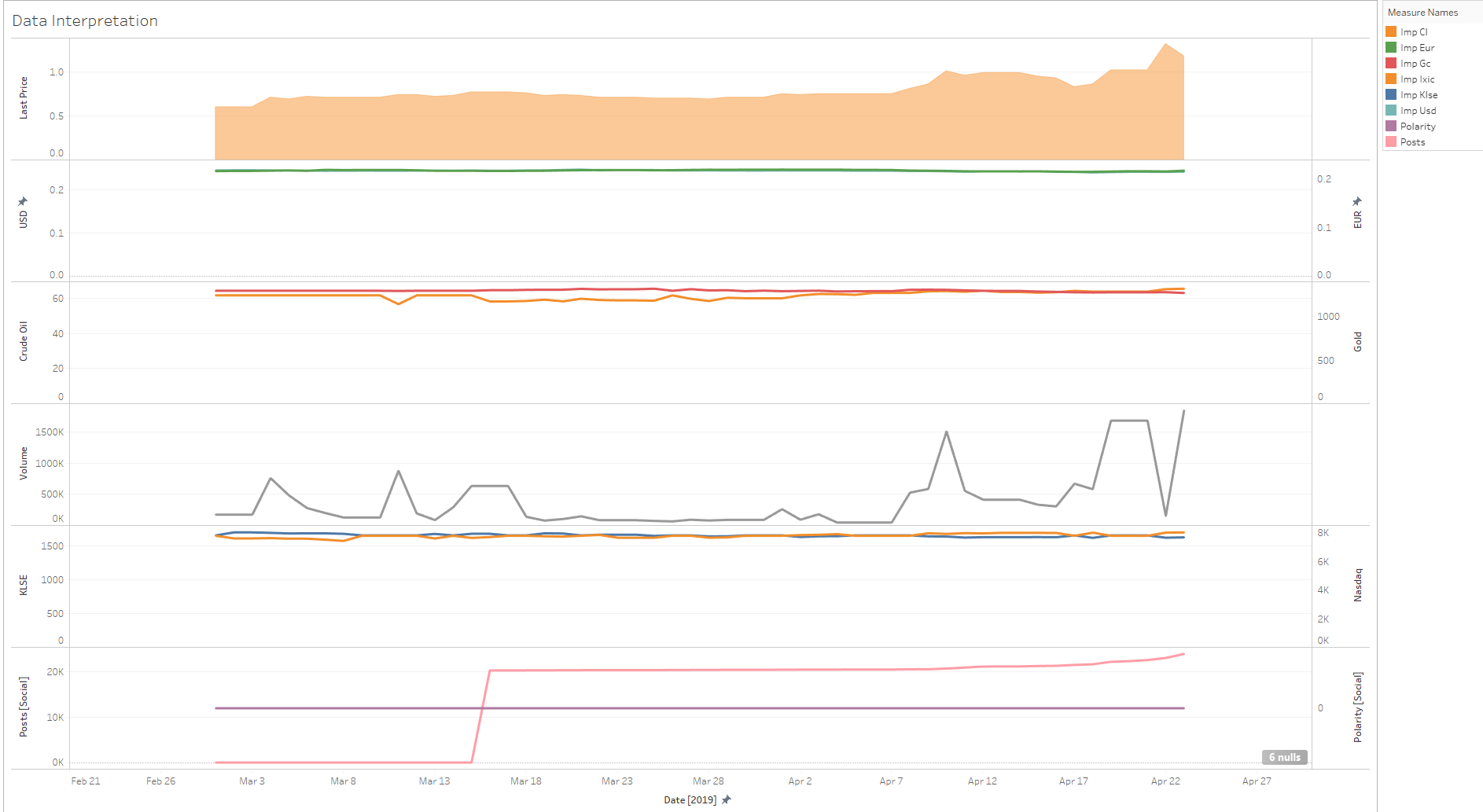
*Figure 6: Co-variation / Correlation analysis and Variable Clustering*

## Analysis using Tableau

Tableau is also used to perform detailed data analysis.



*Figure 7: Composition of different stocks listed in KL Stock Exchange.*

**

*Figure 8: Relation of Target Variable with other Variables.*

This chart displays relationship between the Target Variable and the other variable. This chart is drawn for one stock only ‘IWCITY’ and shows how the Target Variable i.e. Stock Last price, changes with other important variables i.e. major currencies, commodities, Stock volume, Stock exchange Indices, and Social Sentiments.

Crude oil price, stock trade volume, social posts count found to be having positive impact on Stock Last price.

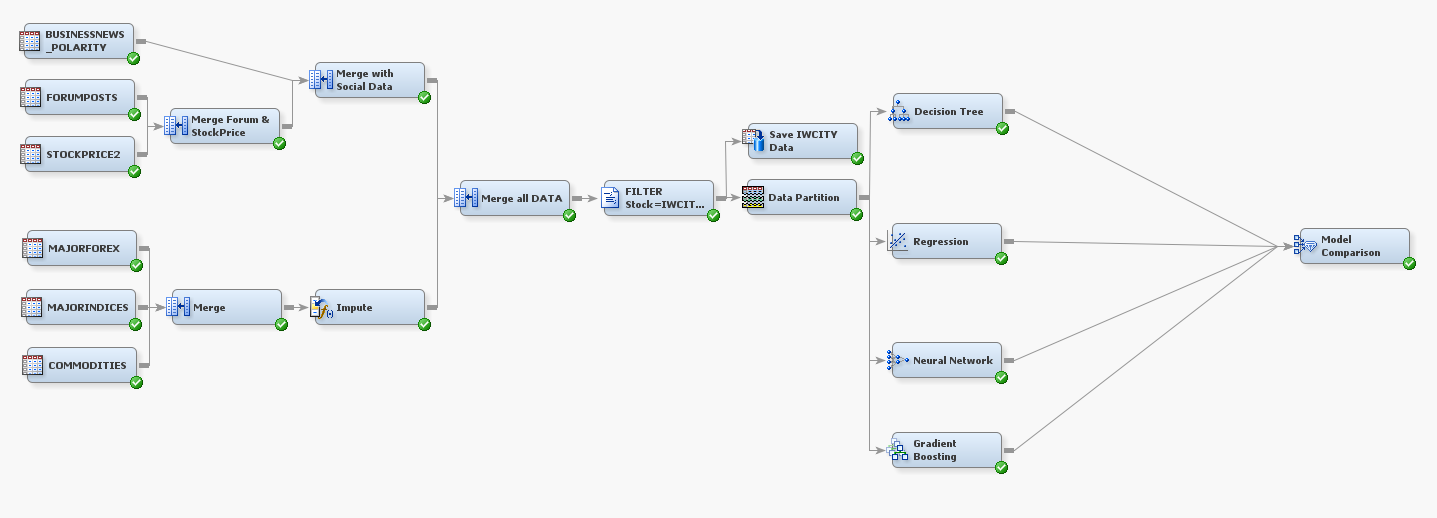


*Figure 9: Social Sentimental Analysis.*

This chart clearly shows impact of social sentiments on Stock Last Price. The stock price is moving up during the positive social sentiments e.g. when lot of people are commenting on social forums about the stock (shown by color intensity) and overall positive News on Business News sources (shown by size of the line).

# Prediction Models

There are total of four prediction models developed using Neural Networks, Decision Tree, Logistic Regression and Gradient Boosting algorithm.



*Figure 10: End to End process defined in SAS Enterprise Miner.*

## Feature Selection

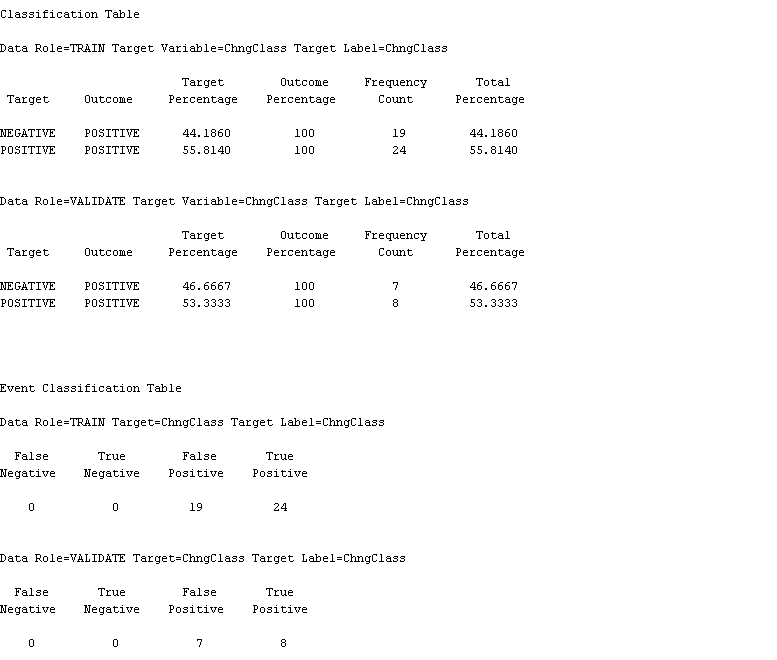
There are three different feature selection techniques are used, to select the most optimal features to develop the prediction model i.e. removing the non-optimal features doesn’t hugely impacting the overall results.

|  |  |  |
| --- | --- | --- |
| **Feature Selection Method** | **Total selected Features** | **Selected Features** |
| Info Gain Evaluation | 19 | IMP\_SI,MarketCap,IMP\_FTSE,IMP\_KLSE,Posts,Volume,IMP\_LB, IMP\_N100,IMP\_N225,IMP\_GC,IMP\_HSI,IMP\_PL,IMP\_IXIC, IMP\_HKD,IMP\_USD,IMP\_EUR,IMP\_BZ,IMP\_CL,IMP\_CNY |
| Correlation Evaluation | 14 | IMP\_EUR,IMP\_PL,IMP\_CNY,IMP\_AUD,Polarity,IMP\_THB, IMP\_GBP,IDR,IMP\_USD,IMP\_HKD,IMP\_SGD,Volume, MarketCap,IMP\_JPY |
| Logistic Regression P-value | 7 | IMP\_EUR,IMP\_GBP,Volume,IMP\_SI,IMP\_CL,IMP\_GC,IMP\_NG |

## Linear Regression (Model 1)

Logistic Regression is a basic and most commonly used Predictive Analysis technique.

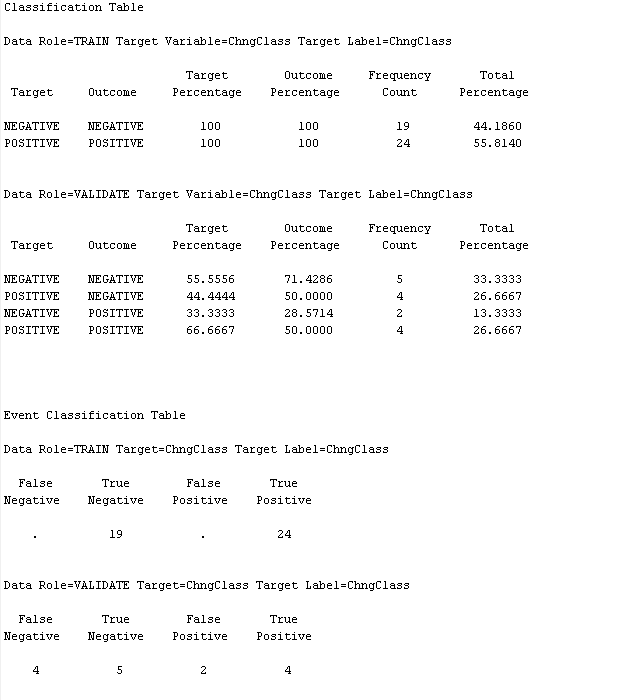
The Logistic Regression is applied on the training dataset by using Regression node in SAS Enterprise Miner.



*Figure 11: Logistic Regression results.*

## Neural Networks (Model 2)

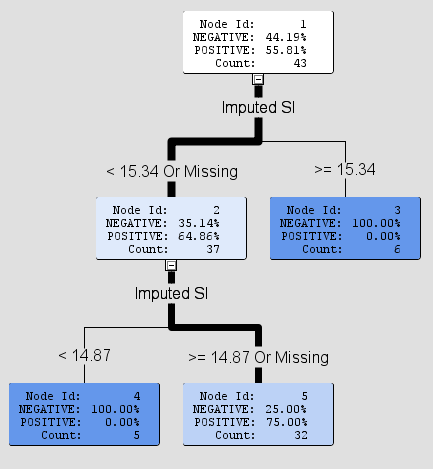
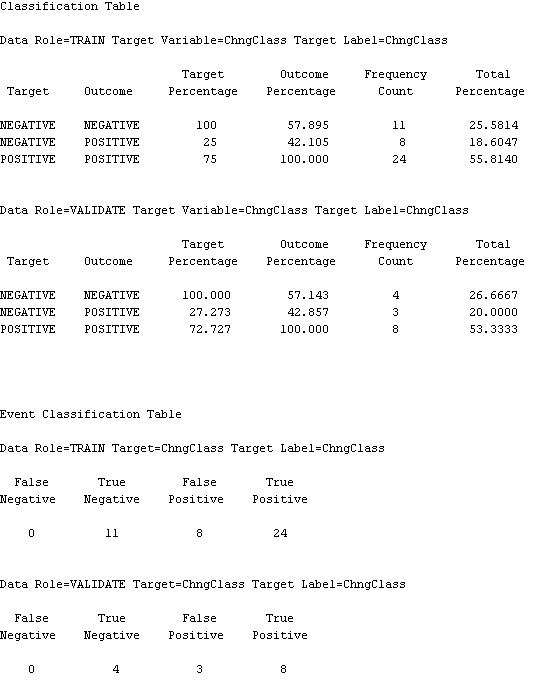
A simple neural networks model will be fit to the dataset using the neural network node in SAS.



*Figure 12: Neural Networks model results.*

## Decision Tree (Model 3)

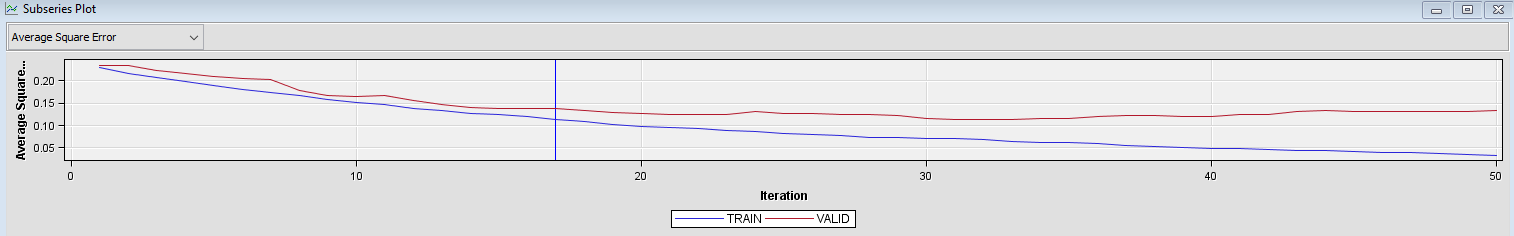
Decision tree is the most powerful and popular tool for classification and prediction. In SAS Enterprise Miner, Decision Tree node is available to fit the model.

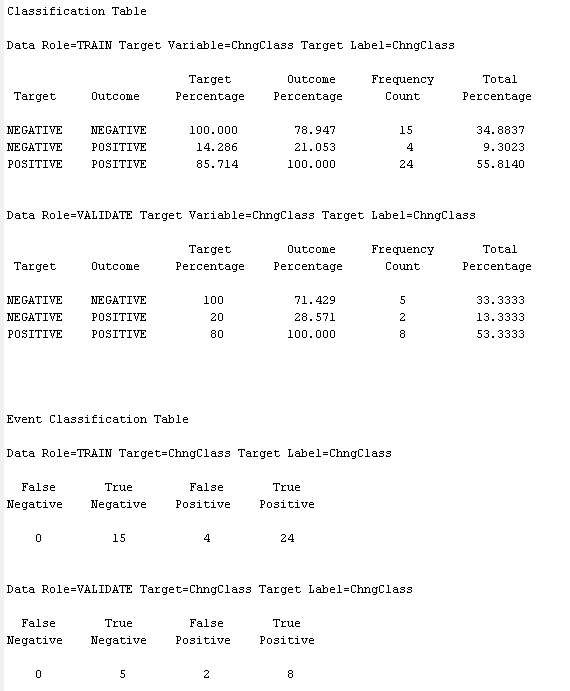
 

*Figure 13: Decision Tree model results.*

## Gradient Boosting (Model 4)

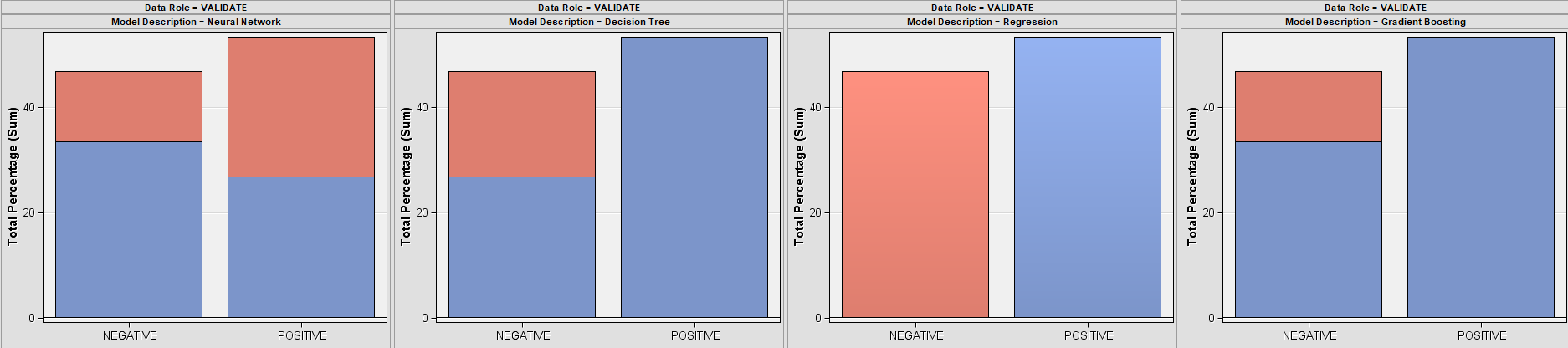
Gradient boosting is a machine learning technique produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. In SAS Enterprise Miner, Gradient Boosting node is available to develop the model.

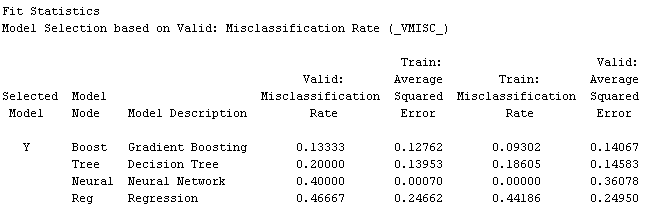
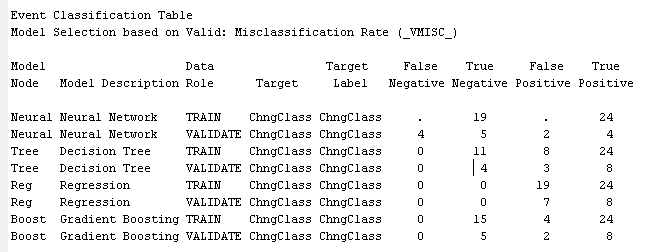




*Figure 14: Gradient Boosting model results.*

# Model Comparisons and Validation



*Figure 15: Model Comparison results.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Models** | **% Accuracy** | **% TP Rate** | **% F-Measure** | **% ROC AUC** |
| Decision Tree | 80 | 72.7 | 84.2 | 78.6 |
| Logistic Regression | 53 | 53.3 | 69.6 | 50 |
| Neural Network | 60 | 66.7 | 57.1 | 64.3 |
| Gradient Boosting | 87 | 80 | 88.9 | 96.4 |

# Analysis & Recommendations

## Observations

Based on the Model comparison results, Gradient Boosting is evaluated as the best model.

The below features impact the Stock Price (IWCITY) the most

* Euro Currency Exchange rate
* Great Britain Currency Exchange rate
* Stock Volume
* Silver Price
* Crude Price
* Gold Price
* Natural Gas price

## Publication

The GitHub publication consists of

1. Research Report
2. Presentation Slide,
3. Data for 1 stock,
4. SAS Data Import code,
5. Python Script for Scrapping,
6. Tableau Analytics - published workbook and pptx,
7. SAS enterprise miner - exported diagram

<https://github.com/vikas-mann/WQD7005>