CSE587 : Data Intensive Computing (Spring 2015)

Project 2: R

Naveen Narayan

UB ID : nnarayan

UB Person # : 50134647

Aim :

The aim of this project is to do time-series forecast of stock price for given data using R language/tools in CCR. 3 techniques, namely, Linear Regression Model, Holt-Winters Model, and ARIMA model are implemented and analyzed.

Introduction :

R is a programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis.

R is a GNU project. The source code for the R software environment is written primarily in C, Fortran, and R. R is freely available under the GNU, General Public License, and pre-compiled binary versions are provided for various operating systems. R uses a command line interface; there are also several graphical front-ends, like RStudio, for it.

The capabilities of R are extended through user-created packages, which allow specialized statistical techniques, graphical devices (ggplot2), import/export capabilities, reporting tools (knitr, Sweave), etc. These packages are developed primarily in R, and sometimes in Java, C, C++ and Fortran. A core set of packages is included with the installation of R, with more than 5,800 additional packages and 120,000 functions (as of June 2014) available at the Comprehensive R Archive Network (CRAN), Bioconductor, Omegahat, GitHub and other repositories.

Time series data have the property of being temporally ordered. Each individual observation have a date and these dates are organised sequentially. The basic function in R that de_nes time series is ts().The main point with the function ts() is that it de_nes a time series object which consists of the data and a time line (including frequency). There is no need to de_ne time as a speci_c variable or use speci_c time variables in existing data.

forecast is a generic function for forecasting from time series or time series models. The function invokes particular methods which depend on the class of the first argument.
For example, the function forecast.Arima makes forecasts based on the results produced by

arima.

forecast.Arima Forecasting using ARIMA or ARFIMA models
Returns one-step forecasts for the data used in fitting the ARIMA model.

forecast.HoltWinters Forecasting using Holt-Winters objects
Returns forecasts and other information for univariate Holt-Winters time series models.

forecast.lm Forecast a linear model with possible time series components
forecast.lm is used to predict linear models, especially those involving trend and seasonality components.

Implementation:

- Stock data is provided in .csv files(<stockName.csv>). Each line in file corresponds to a day's trading information for that stock.
- All '.csv' files in the given data directory are read, one at a time.
- In R, each file is read into a variable.
- If the number of rows read (corresponds to number of days for which data is available) is less than 754 (less than 36 months), ignore.
- Convert the text data read from the file to time-series data (starting from January 1st, 2012) with a frequency of 365(daily frequency).
- Split the time series data into training data (744 instances) and testing data (10 instances).
- Fit training data by applying Arima (or Holt-Winters or Linear Regression) model.
- Forecast using the 'forecast' function.
- Calculate Mean Absolute Error (MAE) for each stock using the below formulae.

$$\mathrm{MAE}_i(\text{each day}) = \left|\text{forecastData}_{\mathrm{i}} - \text{testData}_{\mathrm{i}}\right|$$

$$\text{sum of MAE} = \sum_{i=1}^{10} \mathrm{MAE}_i$$

- Store the stock name (name of the file read) and MAE into the first and second columns of a matrix respectively.

- After all files are read, sort the matrix on MAE (2nd column of the matrix) and extract the top 10 stocks.

- Plot the graph displaying MAE for the top 10 stocks.

Observations :

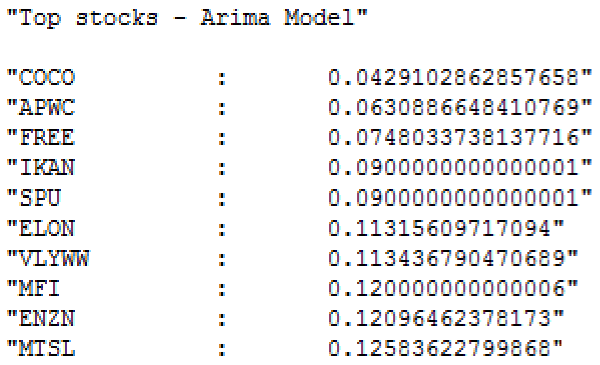
Of the 3 models, Arima, Holt-Winters and Linear Regression, Holt-Winters takes the least time while Linear Regression runs the longest.

Model	**Time taken**
Arima	0:12:05
Holt-Winters	0:02:17
Linear Regression	0:18:15

Table 1 – Time taken for each model

Running all the 3 models takes a total time of 30 minutes and 1 second.

Below are the top 10 stocks derived from the Arima model and the corresponding graph plot.



```
"Top stocks - Arima Model"

"COCO          :       0.0429102862857658"
"APWC          :       0.0630886648410769"
"FREE          :       0.0748033738137716"
"IKAN          :       0.0900000000000001"
"SPU           :       0.0900000000000001"
"ELON          :       0.11315609717094"
"VLYWW         :       0.113436790470689"
"MFI           :       0.120000000000006"
"ENZN          :       0.12096462378173"
"MTSL          :       0.12583622799868"
```

Fig 1 – Top 10 stocks based on MAE (Mean Absolute Error) – Arima Model

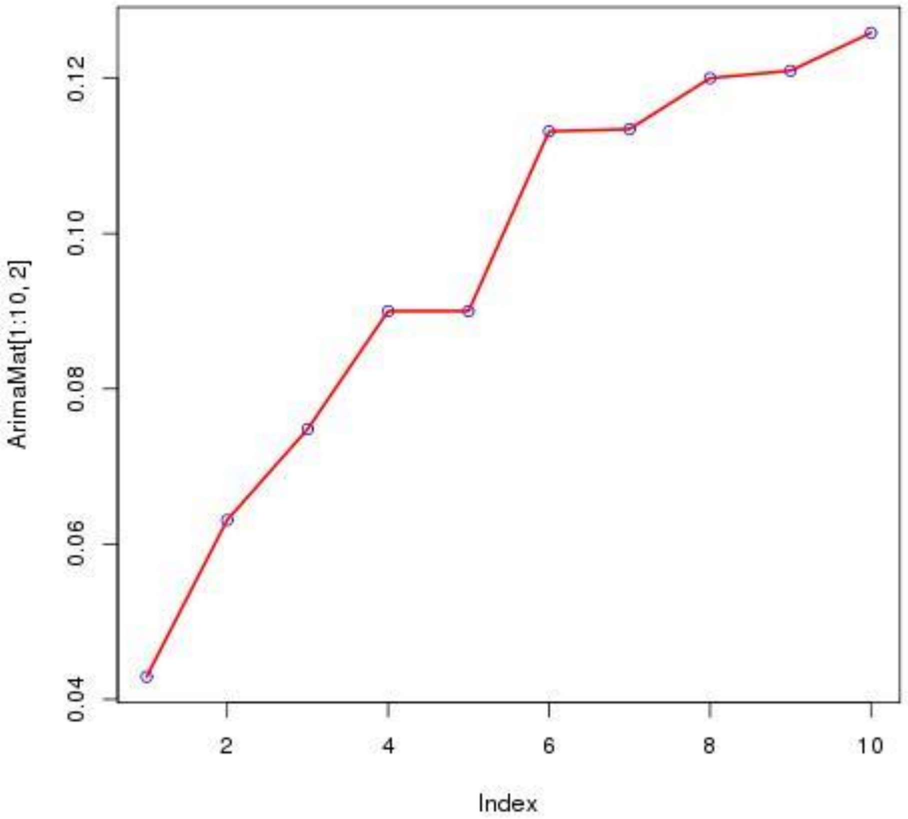
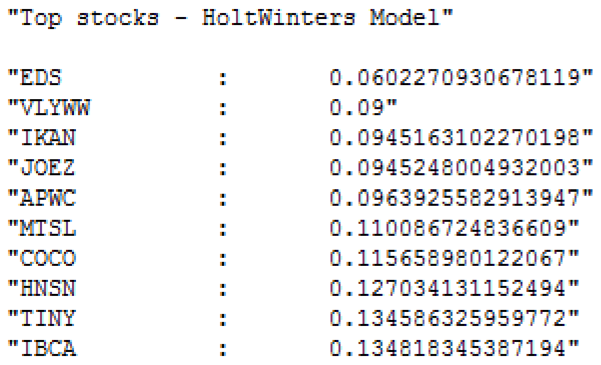


Fig 2 – Plot of MAE (Mean Absolute Error) for the top 10 stocks - Arima Model

Below are the top 10 stocks derived from the Holt-Winters model and the corresponding graph plot.



```
"Top stocks - HoltWinters Model"

"EDS            :       0.0602270930678119"
"VLYWW          :       0.09"
"IKAN           :       0.0945163102270198"
"JOEZ           :       0.0945248004932003"
"APWC           :       0.0963925582913947"
"MTSL           :       0.110086724836609"
"COCO           :       0.115658980122067"
"HNSN           :       0.127034131152494"
"TINY           :       0.134586325959772"
"IBCA           :       0.134818345387194"
```

Fig 3 – Top 10 stocks based on MAE (Mean Absolute Error) - Holt-Winters Model

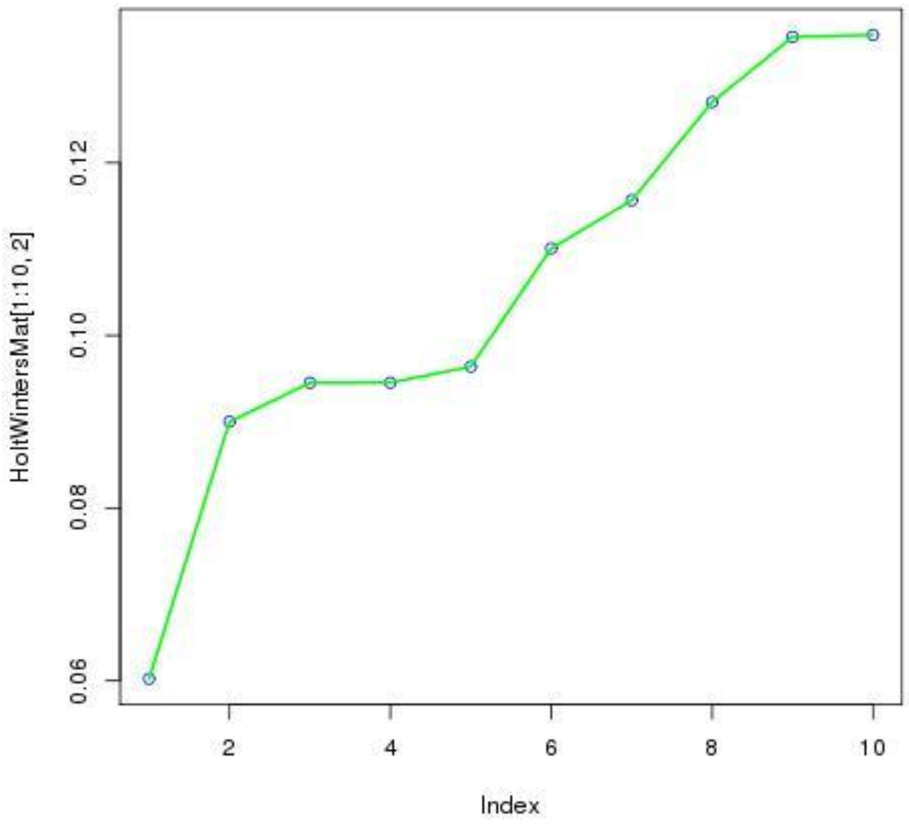
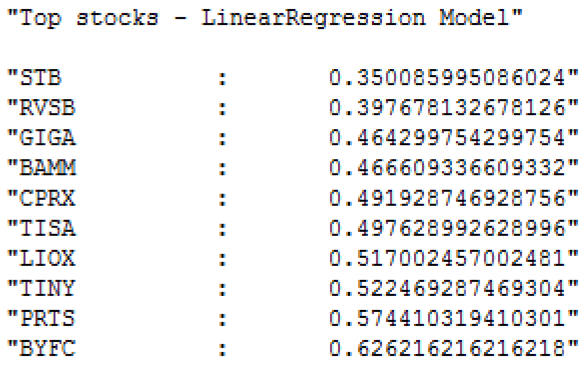


Fig 4 – Plot of MAE (Mean Absolute Error) for the top 10 stocks – Holt-Winters Model

Below are the top 10 stocks derived from the Linear Regression model and the corresponding graph plot.



```
"Top stocks - LinearRegression Model"

"STB            :       0.350085995086024"
"RVSB           :       0.397678132678126"
"GIGA           :       0.464299754299754"
"BAMM           :       0.466609336609332"
"CPRX           :       0.491928746928756"
"TISA           :       0.497628992628996"
"LIOX           :       0.517002457002481"
"TINY           :       0.522469287469304"
"PRTS           :       0.574410319410301"
"BYFC           :       0.626216216216218"
```

Fig 5 – Top 10 stocks based on MAE (Mean Absolute Error) - Linear Regression Model

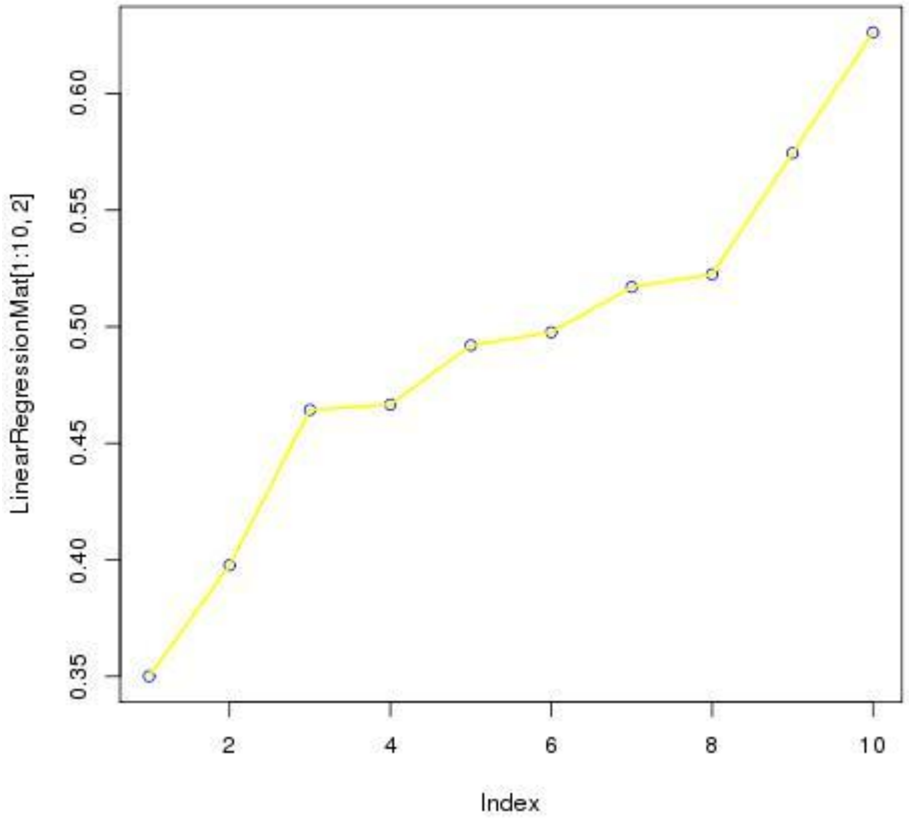


Fig 6 – Plot of MAE (Mean Absolute Error) for the top 10 stocks – Linear Regression Model

- 4 stocks are common between the top 10 list derived from Arima and Holt-Winters model.
- Only 1 stock is common between the top 10 list derived from Holt-Winters and Linear Regression model.
- However, ther are no common entries in the lists derived from Arims and Linear Regression model.

As we do not have any standard list to compare these lists against, we cannot determine which model is optimal.

Arima and Holt-Winters seem to yield most common entries but using Arima model takes more than 5 times longer than the time required with Holt-Winters model.

References :

http://en.wikipedia.org/wiki/R_%28programming_language%29

Introduction to R's time series facilities by Michael Lundholm

Package 'forecast' - Forecasting Functions for Time Series and Linear Models by Rob J Hyndman