Programming Assignment #2

Mimanshu Shisodia
mimanshu@buffalo.edu
Person Number: 50133318

Introduction

We were needed to use R to implement the time-series forecast of stocks in NASDAQ of which the data was provided. The execution of R code has to be done in CCR. The approaches used are:

- Linear Regression Model
- Holt-Winters Model
- ARIMA model

Package Usages

- **Forecast**: Provides the necessary functionality of converting and evaluating the data for time series analysis. The models that are implemented in the code use the library function provided by this package
- **Fpp**: This library has functionality that builds over the Forecast package.
- **Plotly**: This is the Graphics library used.

Evaluation Metric

The MAE (Mean Absolute Error) is a common measure to evaluate error in time series analysis. For each stock, the MAE is calculated as below:

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

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

Goal of the Project

- Based on the above metric we needed to find stocks with best-forecasted performance, i.e. stocks with the minimum value of Sum MAE value. Stock which are best fitting the model learned using the Models mentioned above.
- Also we were required to compare three techniques, namely, Linear Regression Model, Holt-Winters Model, and ARIMA model.

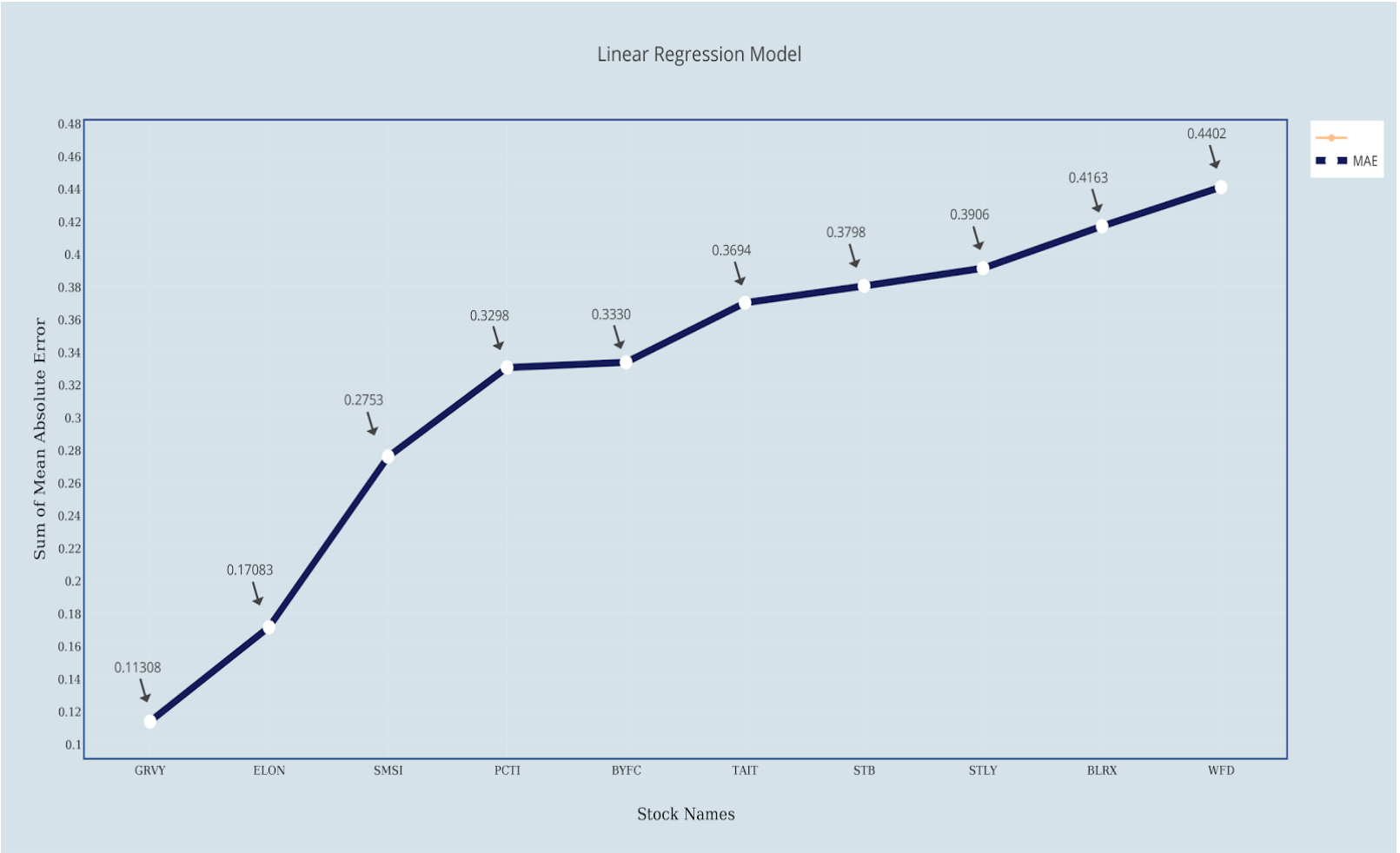
Implementation and Result of Linear Regression Model

Function used for calculation:

- tslm(trainData ~ trend)
    - tslm = Name of the function
    - trainData = Adjusted close price time series data fetched from the csv file for the first 744 days.
    - trend = Seasonality used in the formula. Uses “dummy variables” for seasons. In R, tslm automatically generates seasonal dummies for a ts object
- Result:
    - **------RESULTS OF LINEAR REGRESSION FORECASTING------**

    - **1 GRVY 0.113081527029937**
    - **2 ELON 0.170833278456325**
    - **3 SMSI 0.275305529339237**
    - **4 PCTI 0.329864586468727**
    - **5 BYFC 0.333033614789245**
    - **6 TAIT 0.36946510507102**
    - **7 STB 0.379810373230695**
    - **8 STLY 0.390640313562363**
    - **9 BLRX 0.416331393406878**
    - **10 WFD 0.440248351480182**

- Plot:

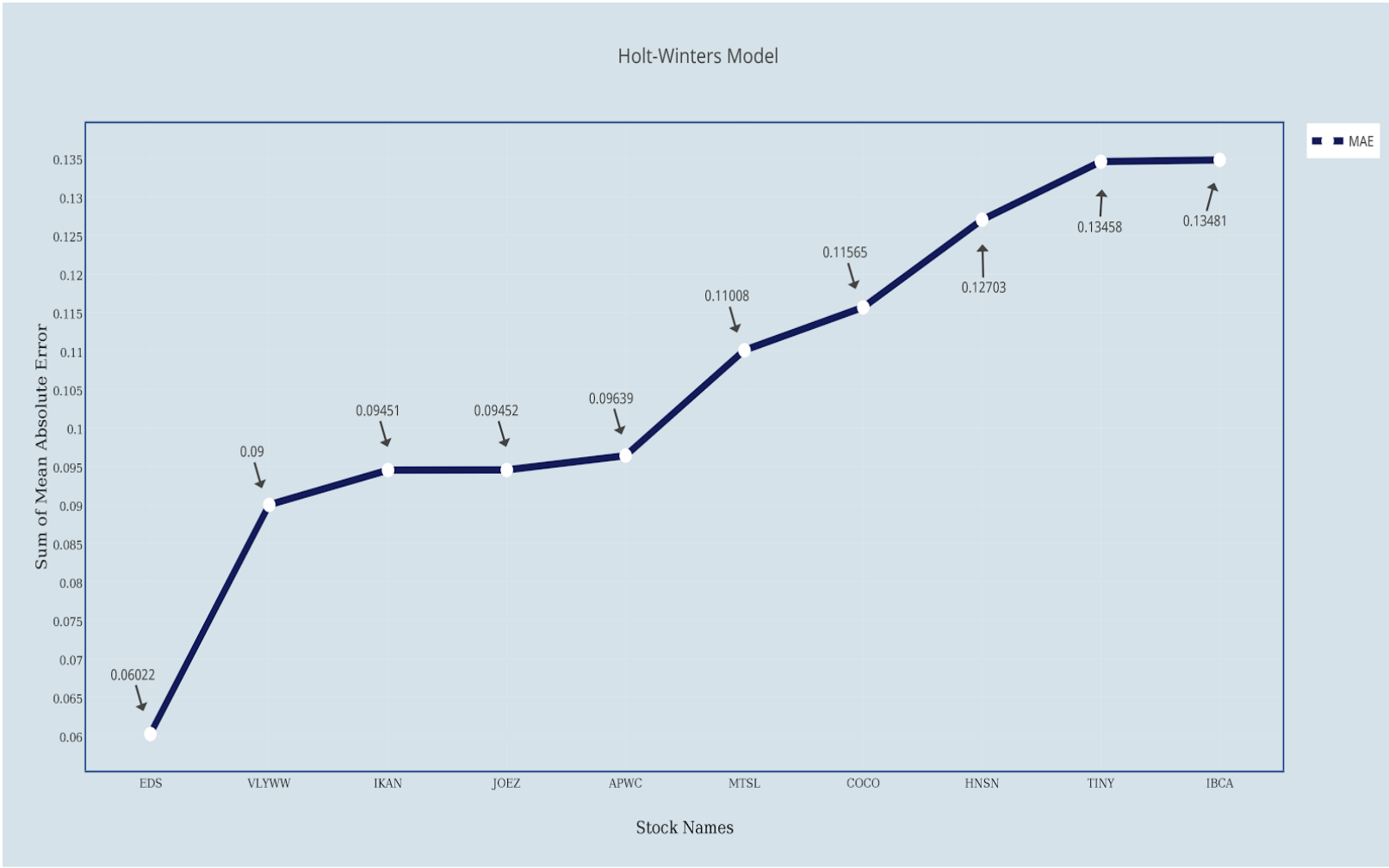


Implementation and Result of Holt-Winters Model

Function used for calculation:

- HoltWinters(trainData, gamma = FALSE)
    - HoltWinters = An object of class "HoltWinters"
    - trainData = Adjusted close price time series data fetched from the csv file for the first 744 days.
    - gamma = This parameter will do the exponential smoothing

- Result:
  - **------RESULTS OF HOLT-WINTERS FORECASTING------**
  - **1 EDS 0.0602270930678119**
  - **2 VLYWW 0.09**
  - **3 IKAN 0.0945163102270198**
  - **4 JOEZ 0.0945248004932003**
  - **5 APWC 0.0963925582913947**
  - **6 MTSL 0.110086724836609**
  - **7 COCO 0.115658980122067**
  - **8 HNSN 0.127034131152494**
  - **9 TINY 0.134586325959772**
  - **10 IBCA 0.134818345387194**
- Plot:

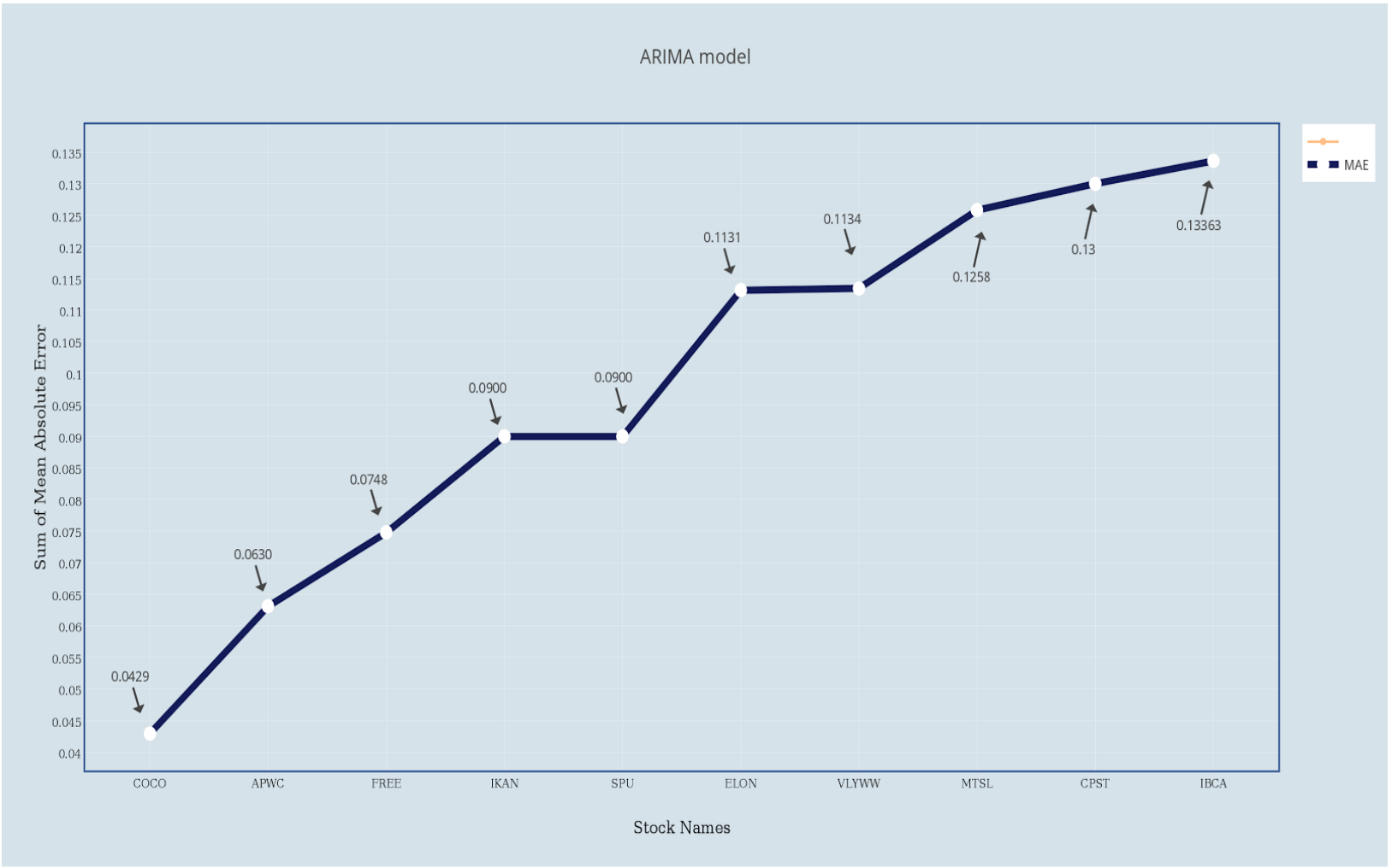


Implementation and Result of ARIMA Model

Function used for calculation:

- auto.arima(trainData)
    - auto.arima = Fit best ARIMA model to univariate time series (from documentation)
    - trainData = Adjusted close price time series data fetched from the csv file for the first 744 days.
- Result:
    - **------RESULTS OF ARIMA FORECASTING------**
    - **1 COCO 0.0429102862857658**
    - **2 APWC 0.0630886648410769**
    - **3 FREE 0.0748033738137716**
    - **4 IKAN 0.0900000000000001**
    - **5 SPU 0.0900000000000001**
    - **6 ELON 0.11315609717094**
    - **7 VLYWW 0.113436790470689**
    - **8 MTSL 0.12583622799868**
    - **9 CPST 0.13**
    - **10 IBCA 0.133633916554524**

- **Plot**


ARIMA model
Sum of Mean Absolute Error
Stock Names
MAE
0.135
0.13
0.125
0.12
0.115
0.11
0.105
0.1
0.095
0.09
0.085
0.08
0.075
0.07
0.065
0.06
0.055
0.05
0.045
0.04
COCO
APWC
FREE
IKAN
SPU
ELON
VLYWW
MTSL
CPST
IBCA
0.0429
0.0630
0.0748
0.0900
0.0900
0.1131
0.1134
0.1258
0.13
0.13363

Problems Faced

- While executing the code, the values were changing as the parameter of various mode functions were changing. Hence the experimentation process followed was focussed on getting the actual trends with the minimum error.
- At the CCR, the Single Node 8 core machine was selected in the SLURM script but this was leading to printing the value 8 times, each time for each core.
- Also, care needed when dealing with the non existing files or files which are empty since the execution may get halted in between and will be failed job.

Comparison Between the Models

Different trends were observed for different models. The reason could be different fitting of the training data. Since the models are trying to learn a generalized representation of data which results in a Mathematical equation (Linear or Nonlinear, depends upon the model and parameters used). When the test data is given to this equation, it will generate the values and we will calculate the value. Now the deviation of this calculated value with respect to the actual values is the Error. Since we are getting dissimilar trends for different models we can rightly say that the Equation differs for each case.

Note: For plots, Plotly was used instead of R plot.