School of Information Sciences

(A Constituent Institute of Manipal University)

Sales Forecasting using Time Series and Neural Networks

A Project Report Submitted By

Carol Steffi Dcunha	**171046001**
Sushmitha Shetty	**171046037**

Under the Guidance of

Mr. B.V. Ravindra

Assistant Professor

Second Semester Master of Engineering
in
Big Data and Data Analytics

Index

Abstract **i**

1 Introduction **1**

1.1 Scope 1

1.2 Definitions and Abbreviations 1

1.3 Objective 1

2 Specifications **2**

2.1 Specifications of the Project 2

2.1.1 Rossmann Dataset 2

2.1.2 ARIMA Model 2

2.1.3 Back-Propagation Neural Network Model 3

2.1.4 NRMSE 3

2.2 End User 4

3 Design **5**

3.1 Operating Environment 5

3.2 Software 5

3.3 Hardware 5

3.4 General Description 5

3.4.1 Functional Description 5

3.4.2 Assumptions 6

4 Results **7**

4.1 Cumulative Conclusion 11

5 Scope for future work **12**

6 References **13**

6.1 Bibiliography 13

6.2 Webiliography 13

Abstract

This project presents the use of time series ARIMA model and neural network back-propagation model in forecasting the sales in a medium size enterprise. The accuracy of the model is evaluated using normalized root mean square error (NRMSE); lower NRMSE indicates better predictive model. The forecasts obtained using the back-propagation model is found to be more accurate than those of ARIMA model.

Chapter 1

Introduction

Time series modelling, and forecasting has fundamental importance to various practical domains. The main aim of time series modelling is to carefully collect and rigorously study the past observations of a time series, to develop an appropriate model which describes the inherent structure of the series. This model is then used to generate future values for the series, i.e. to make forecasts. Time series forecasting thus can be termed as the act of predicting the future by understanding the past. Artificial neural networks (ANNs) have been extensively studied and used in time series forecasting. With ANNs, there is no need to specify a model form. Rather, the model is adaptively formed based on the features presented from the data. This project deals with statistical as well as neural network time series modelling to analyze the behavior of sales in a medium size enterprise. The main aim is to determine the appropriate model for sales forecasting by comparing the accuracies of the two models.

1.1 Scope

Sales forecasting is crucial for many retail operations. Both the models presented in the project can be used for forecasting sales. However, the error comparison analysis depicts which model can be preferred over the other.

1.2 Definitions and Abbreviations

- **Analytics:** The systematic computational analysis of data or statistics.
- **Time series:** A series of data points indexed (or listed or graphed) in time order.
- **Time series forecasting:** The use of a model to predict future values based on previously observed values.
- **Neural network:** A system of programs and data patterned on the operation of the human brain that learns from and adapts to initial rules and experience.

1.3 Objective

- Use of time series model and neural network back-propagation model in analyzing the behavior of sales in a medium size enterprise
- Compare the accuracy of sales forecast among the two models

Chapter 2

2.1 Specifications of the Project

2.1.1 Rossmann Dataset

Sales forecasting is done with historical sales data for 1,115 Rossmann stores. The data is collected from Kaggle. The data set is for the period from January 1st, 2013 to July 31st, 2015 and contains 1,017,209 records. The list of columns available in data is as follows:

- Store ID - a unique ID for each store
- Date - date of sales
- Day of Week - day of week
- Sales - predicting outcome
- Open - indicate if the store is open
- Promo - indicate if there is a promotion going on of Customers
- StateHoliday/SchoolHoliday - indicates a state or a school holiday
- StoreType/Assortment - different store models and assortment levels

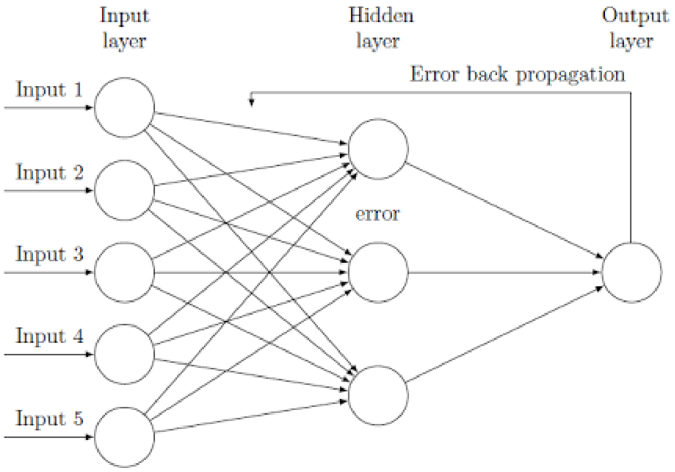
2.1.2 ARIMA Model

ARIMA is the abbreviation of AR (*AutoRegressive*), I (*Integrated*), and MA (*Moving Average*). A convenient notation for ARIMA model is ARIMA(p,d,q). Here p, d and q are the levels for each of the AR, I, and MA parts. In each step of ARIMA modeling, time series data is passed through these three passes as following:

- **Integrated (d)** - Here, the original series is subtracted with its lagged series to extract trends from the data e.g. November's sales values are subtracted with October's values to produce trendless residual series. When d=0, the series is trend-less and is called stationary on mean series.
- **AutoRegressive (p)** - Here, the influence of the values of previous periods is extracted on the current period e.g. the influence of the September and October's sales value on the November's sales.
- **Moving Average (q)** - Finally, the MA involves finding relationships between the previous periods' error terms on the current period's error term.

2.1.3 Back-Propagation Neural Network Model

A Neural Network is a machine learning approach inspired by the way in which the human brain performs a particular learning task. It is organized in layers. Layers are made up of a number of interconnected nodes which contain an activation function. Patterns are presented to the network via the input layer, which communicates to one or more hidden layers where the actual processing is done via a system of weighted connections. The hidden layers then link to an output later where the answer is the output.



The back-propagation neural network model is a paradigm commonly used in the areas of signal recognition, and principally in forecasting of time series. The learning is a supervised process that occurs with each epoch through a forward activation flow of outputs, and the backward error propagation of weight adjustments. Each hidden layer node is a sigmoidal activation function which polarizes network activity and helps it to stabilize.

2.1.4 NRMSE

The root-mean-square error (RMSE) is a measure of the differences between values (sample and population values) predicted by a model and the values actually observed. The RMSE represents the sample standard deviation of the differences between predicted values and observed values. RMSE is a measure of accuracy, to compare forecasting errors of different models for a particular data and not between datasets, as it is scale-dependent.

$$\mathrm{RMSE} = \sqrt{\sum \frac{(y_{pred} - y_{ref})^2}{N}}$$

NRMSE is the normalized RMSE given by,

$\mathrm{NRMSE} = \frac{RMSE}{S.D.}$, where S.D. is the standard deviation among the observations.

2.2 End User

Forecasting methods can be used by retailers to anticipate the future purchasing actions of consumers by evaluating past revenue and consumer behavior over the previous months or year to discern patterns and develop the upcoming months.

Chapter 3

Design

3.1 Operating Environment

- Ubuntu 16.04

3.2 Software

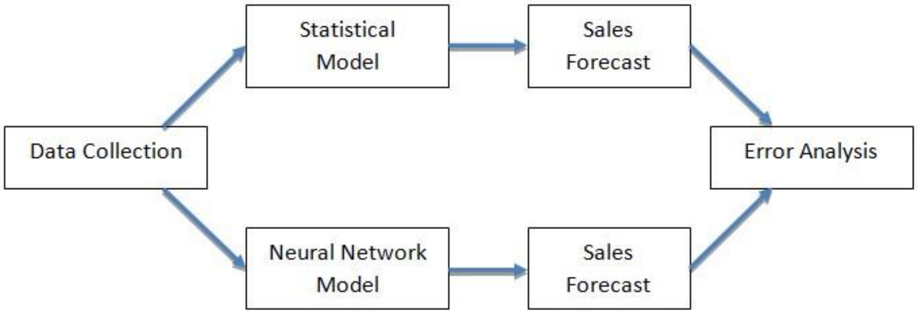
- R

3.3 Hardware

- 64-bit machine
- 2GB RAM or above
- Hard Disk Capacity: 1GB
- Processor Intel core i5/i7

3.4 General Description

3.4.1 Functional Description



The above diagram depicts the functional flow of the project. It includes:

- Data collection - This phase deals with collecting a sales dataset of certain store or supermarket. The project uses Rossmann sales dataset.

- Statistical Model - This involves fitting a mathematical model for the time series sales data that depends on many factors like customers, promotion days and holidays.
- Neural Network Model - This involves fitting a neural network model that uses back-propagation mechanism to learn the time series data.
- Sales Forecast - Here, the store sales is forecasted using the above fitted models with the test data.
- Error Analysis - This phase compares the RMSE values of the fitted models, thus concluding on one of the models that best forecasts the dataset.

3.4.2 Assumptions

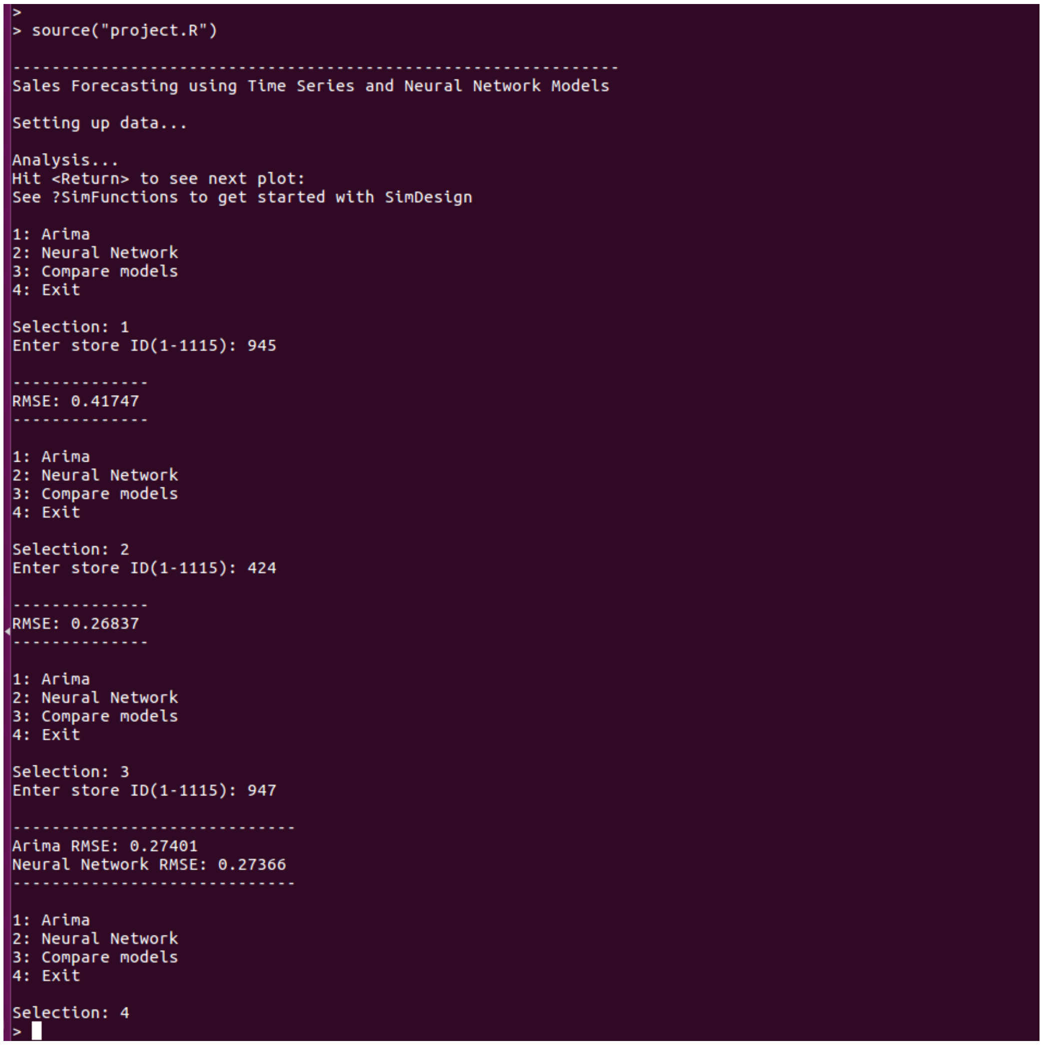
StateHoliday isn't included in the model because,

- Only 0.11% of stores are open on public holidays
- Model requires regressor matrix to have full rank which can't be achieved if StateHoliday is included.

Chapter 4

Results

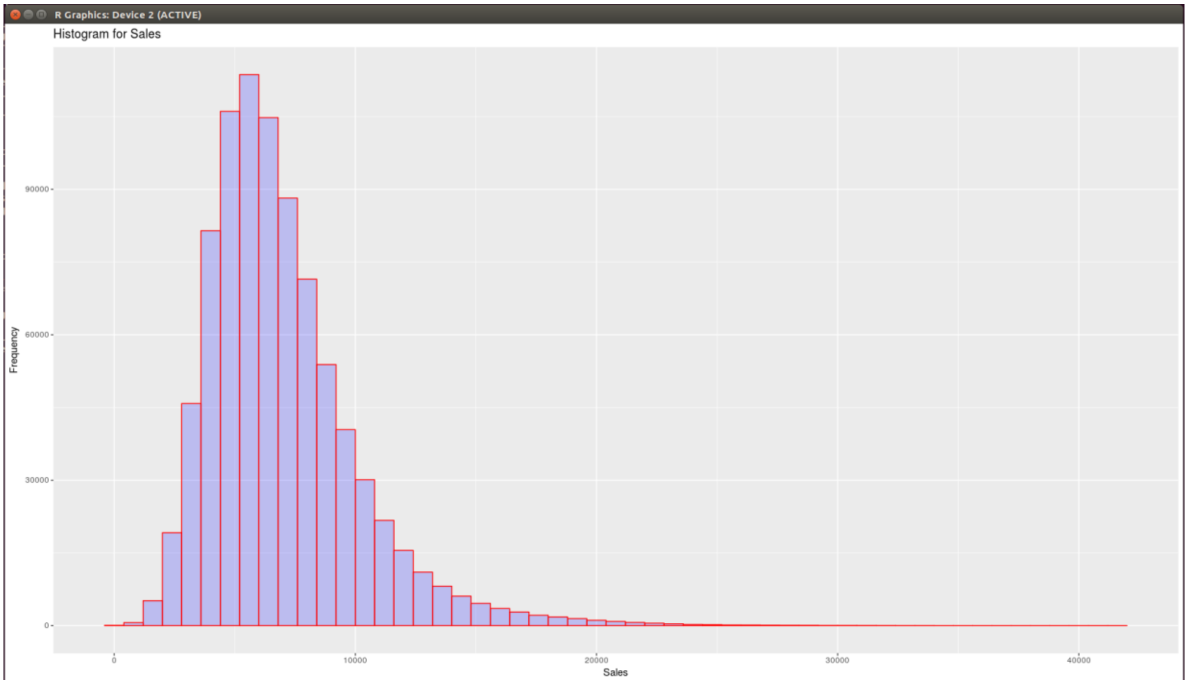
The following figure shows the user interface of the project with relevant options.



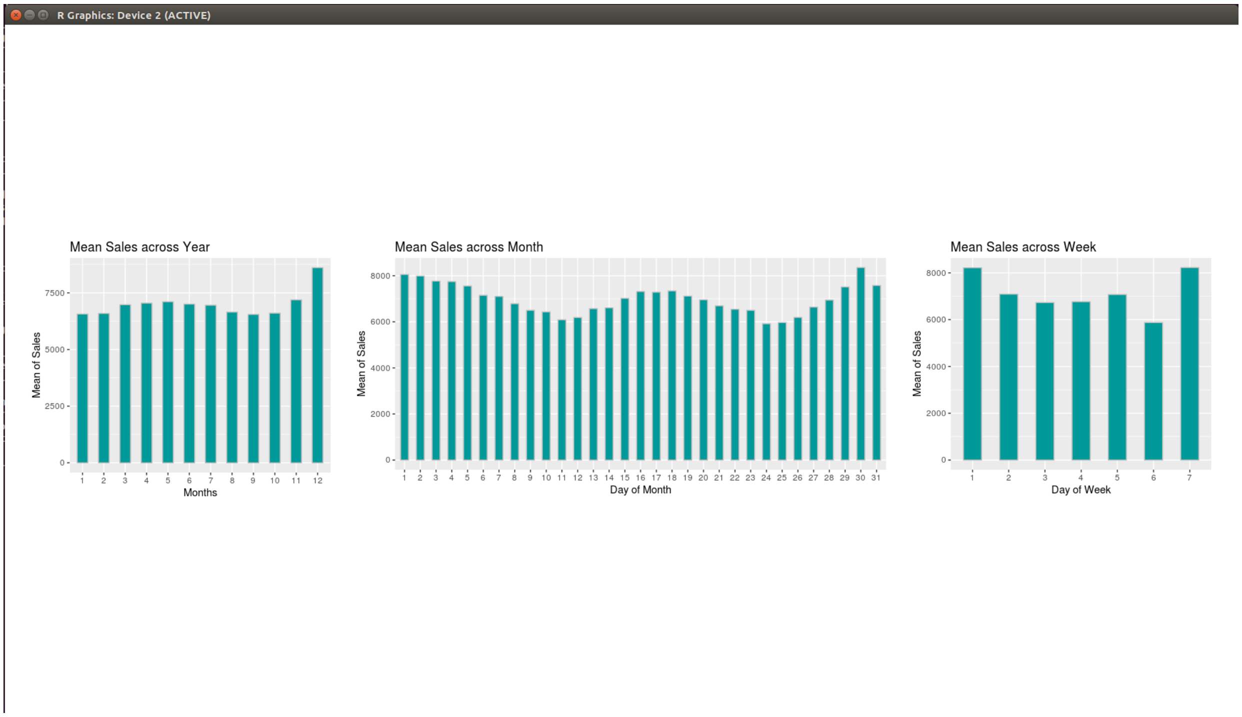
```
>
> source("project.R")

---------------------------------------------------------------
Sales Forecasting using Time Series and Neural Network Models

Setting up data...

Analysis...
Hit <Return> to see next plot:
See ?SimFunctions to get started with SimDesign

1: Arima
2: Neural Network
3: Compare models
4: Exit

Selection: 1
Enter store ID(1-1115): 945

--------------
RMSE: 0.41747
--------------

1: Arima
2: Neural Network
3: Compare models
4: Exit

Selection: 2
Enter store ID(1-1115): 424

--------------
RMSE: 0.26837
--------------

1: Arima
2: Neural Network
3: Compare models
4: Exit

Selection: 3
Enter store ID(1-1115): 947

-----------------------------
Arima RMSE: 0.27401
Neural Network RMSE: 0.27366
-----------------------------

1: Arima
2: Neural Network
3: Compare models
4: Exit

Selection: 4
> 
```

Analysis of Dataset

The histogram depicts that majority of the stores have sales within 10,000.

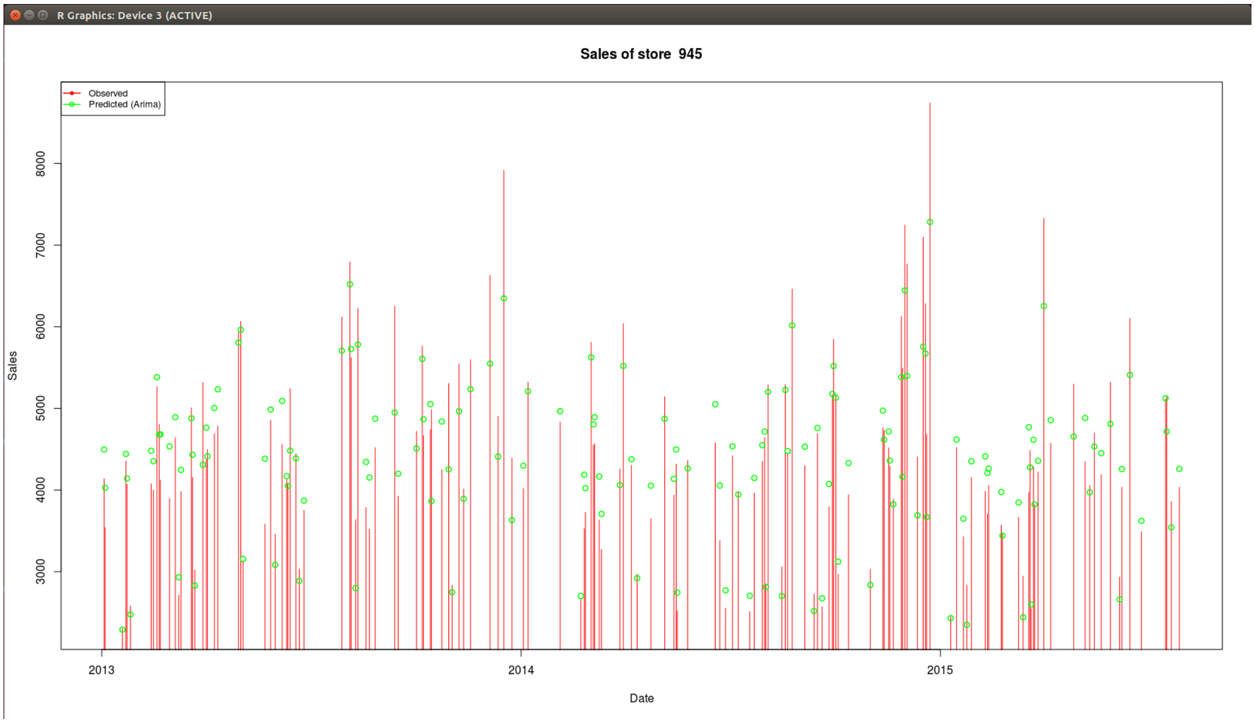


It is observed that sales are higher during December owing to festive season. Moreover, sales are peak during beginning and month end (monthly payroll) as well as during weekends.



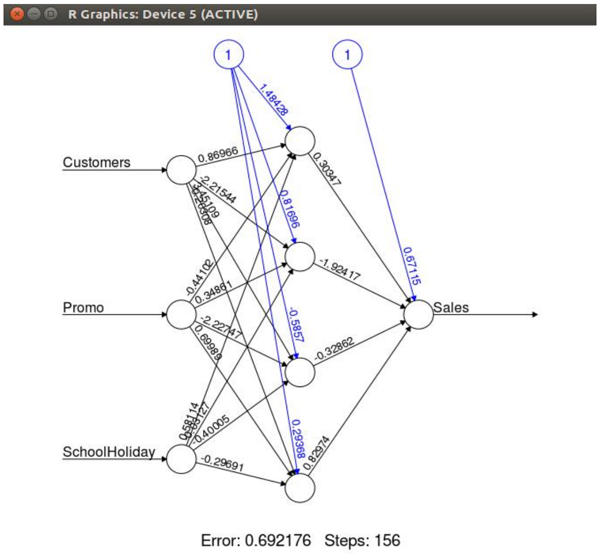
Choice 1: Arima

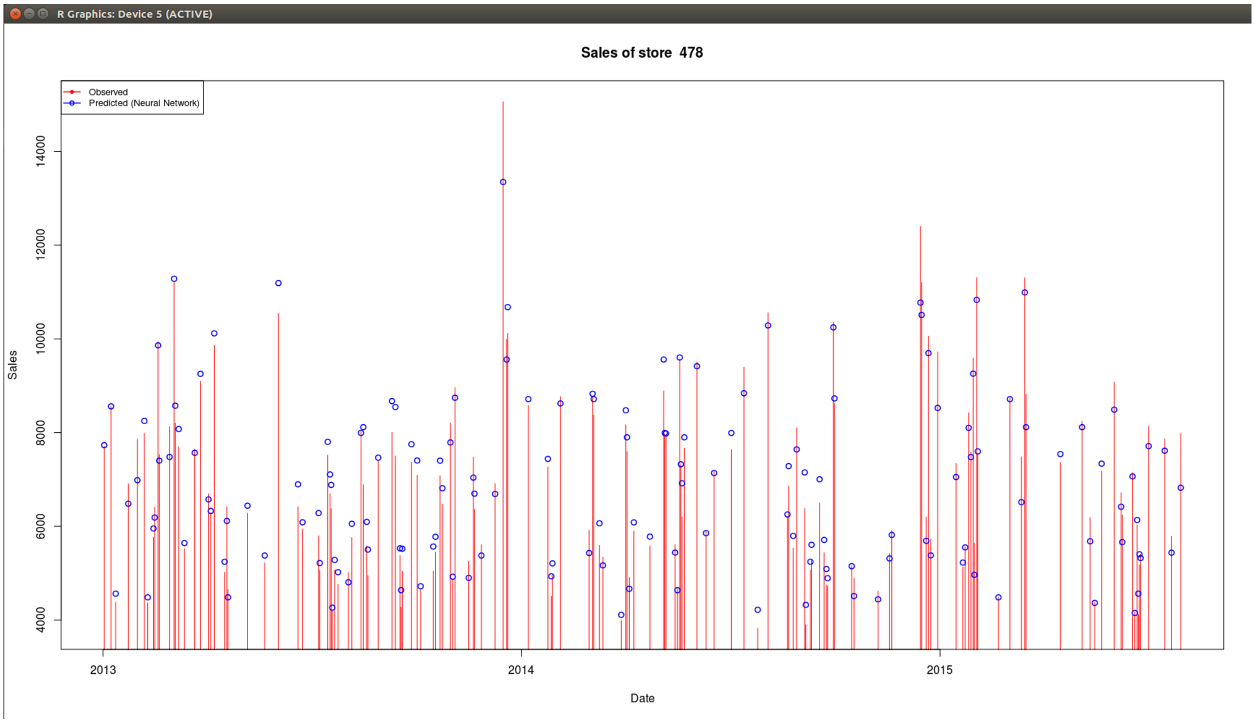
This ARIMA model is performed to forecast sales of store 945. A NRMSE of 0.41747 is obtained.



Choice 2: Neural Network

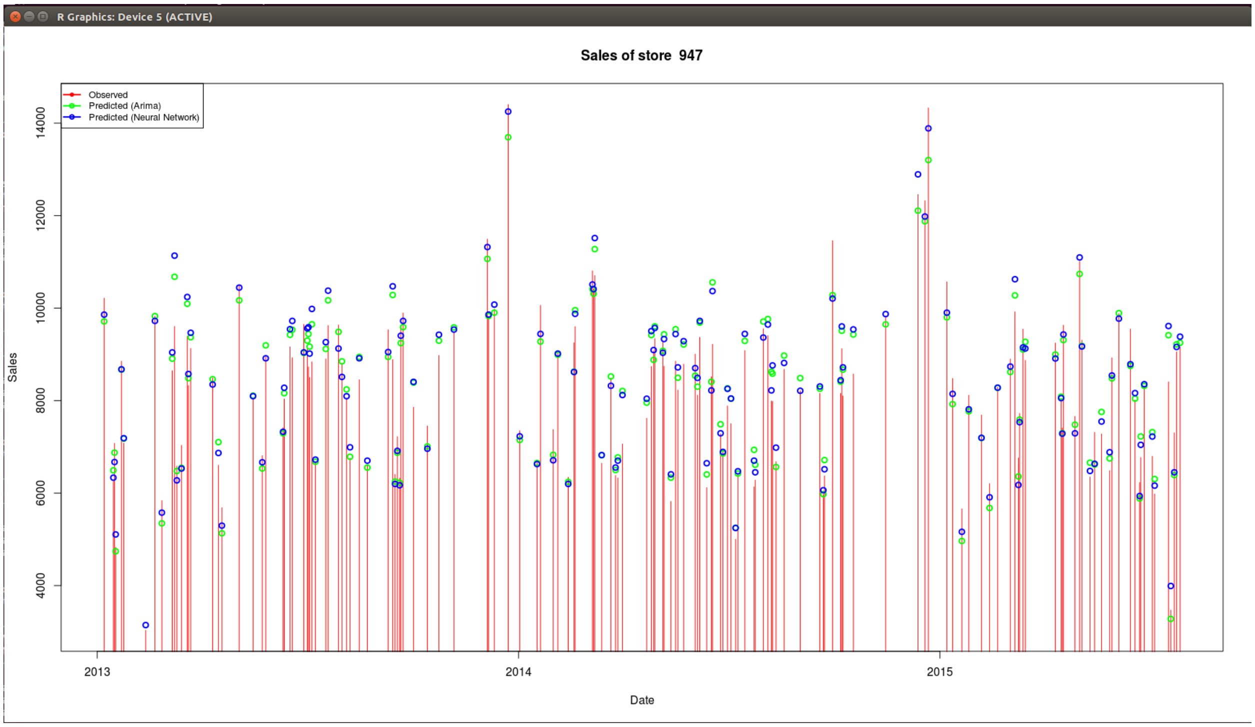
This neural network model is performed to forecast sales of store 478. A NRMSE of 0.19548 is obtained.





Choice 3: Compare models

A comparison is drawn between the two models for store 947. The NRMSE for ARIMA is 0.27401 and for Neural Network is 0.27366. This suggest that neural network shows slightly better forecast than ARIMA for store 947.



4.1 Cumulative Conclusion

The ARIMA model presented average NRMSE of 0.327348, whereas the neural network model presented an average NRMSE of 0.266387. Hence, the model obtained by the neural network was superior to ARIMA model, in adjustment as well as in forecasting for the data analyzed.

Chapter 5

Scope for future work

As observed, only three regressors - Customers, Promo and SchoolHoliday are considered in the models. As a part of future work, additional parameters such as store size, product quantity and weather details can be considered to develop a more robust and accurate forecasting model.

Chapter 6

References

6.1 Bibliography

[1] Angela P. Ansuj, M. E.Camargo, R. Radharamanan and D. G. Petty, "Sales Forecasting using Timeseries and Neural Networks", 19th International Conference on Computers and Industrial Engineering

[2] David Beam and Mark Schramm, "Rossmann Store Sales", December 2015

[3] G. Peter Zhang , "Time series forecasting using a hybrid ARIMA and neural network model", Neurocomputing 50 (2003) 159 – 175

6.2 Webliography

- https://www.kaggle.com/c/rossmann-store-sales/data
- http://rstudio-pubs-static.s3.amazonaws.com/137275_80c072e9458b48778c6629f2a2650592.html
- https://rpubs.com/gpetho/142772