Title : Sales Forecasting using Time Series and Neural Networks		
Class/Section :	Register Number	Name
Big Data and Data Analytics	171046001	Carol Steffi Dcunha
2nd Semester	171046037	Sushmitha Shetty

Guide: B.V Ravindra

Contents

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 analyse the behaviour 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

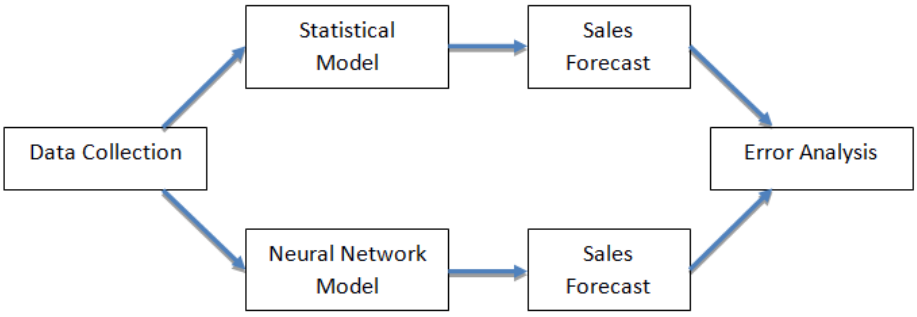
Sales forecasting is crucial for many retail operations. Both the models presented in the project can be used for forecasting sales. However, the error comparison analysis depicts which model can be preferred over the other.

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.

2. General Description

2.1 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.

2.2 Specifications of the Project

2.2.1 Rossmann Dataset

Sales forcasting 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

2.2.2 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 it's lagged series to extract trends from the data e.g. November's sales values are subtracted with October's values to produce trend-less 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.

2.2.3 RMSE

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 RMSD represents the sample standard deviation of the differences between predicted values and observed values. RMSD 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.

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

2.3 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.

3. Specific Requirements

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

4. References

4.1 Bibliography

[1] Angela P. Ansuj, M. E.Camargo, R. Radharamanan and D. G. Petty, “Sales Forecasting using Time Series 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

4.2 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