**Data Science with Python Course-end Project 1**

**Name: Ryne Mascarenhas**

**Email id:** [**rynem8@gmail.com**](mailto:rynem8@gmail.com)

**Retail Analysis with Walmart Data**

1. Introduction

This report presents an analysis of sales data for Walmart stores, with a focus on identifying trends and building predictive models to forecast future demand. The analysis utilizes various statistical techniques and machine learning algorithms to gain insights into the factors that influence sales and to develop models that can accurately predict weekly sales.

Description

One of the leading retail stores in the US, Walmart, would like to predict the sales and demand accurately. There are certain events and holidays which impact sales on each day. There are sales data available for 45 stores of Walmart. The business is facing a challenge due to unforeseen demands and runs out of stock some times, due to the inappropriate machine learning algorithm. An ideal ML algorithm will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc.

Walmart runs several promotional markdown events throughout the year. These markdowns precede prominent holidays, the four largest of all, which are the Super Bowl, Labour Day, Thanksgiving, and Christmas. The weeks including these holidays are weighted five times higher in the evaluation than non-holiday weeks. Part of the challenge presented by this competition is modelling the effects of markdowns on these holiday weeks in the absence of complete/ideal historical data. Historical sales data for 45 Walmart stores located in different regions are available.

**Dataset Description**

This is the historical data that covers sales from 2010-02-05 to 2012-11-01, in the file Walmart\_Store\_sales. Within this file you will find the following fields:

* Store - the store number.
* Date - the week of sales.
* Weekly\_Sales -  sales for the given store.
* Holiday\_Flag - whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week.
* Temperature - Temperature on the day of sale
* Fuel\_Price - Cost of fuel in the region.
* CPI – Prevailing consumer price index.
* Unemployment - Prevailing unemployment rate.

**Holiday Events**

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13  
Labour Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13  
Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13  
Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

**Analysis Tasks**

**Basic Statistics tasks**

* Which store has maximum sales.
* Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation.
* Which store/s has good quarterly growth rate in Q3’2012
* Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together.
* Provide a monthly and semester view of sales in units and give insights.

**Statistical Model**

For Store 1 – Build prediction models to forecast demand.

* Linear Regression – Utilize variables like date and restructure dates as 1 for 5 Feb 2010 (starting from the earliest date in order). Hypothesize if CPI, unemployment, and fuel price have any impact on sales.
* Change dates into days by creating new variable.

Select the model which gives best accuracy.

1. Results and observations

The following results are of the code provided separately as a pdf file named Sales\_forcast.

The analysis begins with an exploration of the sales data, which includes information on weekly sales, store location, date, and various economic indicators such as CPI, unemployment, and fuel price. We perform basic statistics task for understanding the dataset with respect to weekly sales. Furthermore, the data is cleaned and pre-processed to remove outliers and ensure data quality.

A screen shot of a computer

Description automatically generatedA computer screen shot of a cross

Description automatically generated A screen shot of a computer

Description automatically generated A screen shot of a computer

Description automatically generated A computer screen shot of a black screen

Description automatically generated A screenshot of a computer

Description automatically generated

The above images show the results for store 14 with maximum mean and standard deviation weekly sales. Moreover, we can also see the 2012 3rd quarter growth of each store, with store 7 having the maximum growth for weekly sales. Next, comparisons between different weekly sales for holidays and non-holidays is done to find if holidays impact weekly sales as show above.

A screen shot of a computer screen

Description automatically generated

A group of blue and white vertical bars

Description automatically generated

A graph with blue lines

Description automatically generated

The above images show the weekly sales for each month for each year and the entire dataset. This shows the month 12 having highest sales for each year and the entire dataset. Additionally, we also we show a barplot with and without the outliers and also check for correlation between different variables shown below.

A collage of blue graphs

Description automatically generated

The above image shows correlation between different variables and weekly sales cannot be modelled with a linear line.

A diagram of a graph

Description automatically generated with medium confidence A row of blue rectangular bars

Description automatically generated with medium confidence

Finally, we create 2 models to predict the weekly sales for store 1 using linear regression and random forest. The dataset is scaled to standard values before training the models.

A screenshot of a computer

Description automatically generated A screenshot of a computer program

Description automatically generatedA screenshot of a computer screen

Description automatically generated

The accuracy score shows that linear regression model performs very poorly giving only 15% accuracy, while Random Forrest gives 94% accuracy, indicating its superior predictive capabilities for the entire dataset. However, for predicting store 1 the Linear regression model scores negative accuracy while Random Forest algorithm gives a slightly lower but useful score with 71% accuracy.

1. Conclusion

The analysis provides valuable insights into the sales performance of Walmart stores and identifies key factors that influence demand. The predictive models developed can be used to forecast future sales, optimize inventory levels, and make informed decisions regarding sales strategies. By leveraging the Random Forest models, Walmart can improve its operational efficiency and maximize its revenue potential.