BDS 21

Devika S Nair

BS21DMU010

SP JAIN SCHOOL OF GLOBAL MANAGEMENT

Analysis and prediction of Walmart sales

Project report

Table of Contents

[1. Introduction 2](#_Toc103531286)

[1.1 Data Description 2](#_Toc103531287)

[2 Understanding the data 3](#_Toc103531288)

[3](#_Toc103531289)

[2.1 Key statistics of the data 4](#_Toc103531290)

[4](#_Toc103531291)

[2.1.1 observations: 4](#_Toc103531292)

[3 Scatterplots and boxplots 5](#_Toc103531293)

[3.1 scatterplots 5](#_Toc103531294)

[3.2 Boxplots and finding outliers 6](#_Toc103531295)

[4 Correlation Matrix 7](#_Toc103531296)

[5 Data analysis 8](#_Toc103531297)

[5.1 Linear regression and random forest analysis 8](#_Toc103531298)

[5.1.1 Linear Regression 8](#_Toc103531299)

[5.1.2 Random forest analysis 9](#_Toc103531300)

[5.2 Multiple regression analysis 9](#_Toc103531301)

[5.2.1 Multiple regression 10](#_Toc103531302)

[5.2.2 Outputs of this model 10](#_Toc103531303)

# Introduction

Walmart is one of the leading retail stores in the US. It was founded by a man named Sam Walton in 1962. The aim of this project is to predict and analyse the sales demand. The dataset has events like sales and holidays which impact the sales output each day. The most prominent holidays that can be observed include the Super Bowl, Labour Day, Christmas, and Thanksgiving.

This is the historical sales data of 45 Walmart stores in different regions. The main aim of this study is to analyse and predict sales and find the best fit model for this data.

## Data Description

Store – The store number

Date – The date

Weekly\_Sales – sales of a given week

Holiday\_Flag - whether the week is a special holiday week 1 – Holiday week 0 – non-holiday

Temperature - Temperature on the day of sale

Fuel\_Price - Cost of fuel in the region

CPI – Prevailing consumer price index

Unemployment – The prevailing unemployment rate

Holiday Events:

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

# Understanding the data

The table shows the first few stores of the data i.e., it is a sample of the dataset used for this project.

Note: the outputs shown below are the results from python R and JMP

### Table Description automatically generated

Table 1

The data information is shown in figure 1

Text

Description automatically generated

Figure 1

NULL value test was performed on the dataset. No NULL values were present in the dataset

## Key statistics of the data

Table 2

Graphical user interface, text, application

Description automatically generated

### 

Table 3

Text

Description automatically generated

Table 2 and table 3 show the key parameter of the data’s various attributes.

### observations:

From the tables output some key observations can be made:

1. There are 45 stores each having 143 observations
2. Extreme values greatly influence the mean weekly sales.

# Scatterplots and boxplots

## scatterplots

A picture containing diagram

Description automatically generated Scatterplots for visualising each attribute with weekly sales. This is to primarily visualise the data.

Figure 2

## Boxplots and finding outliers

Chart, box and whisker chart

Description automatically generated

Figure 3

Figure 3 shows the various outliers in each of the attributes. It is clear to see that unemployment and temperature show outliers. Once these are cleaned from the data, the model building is done.

# Correlation Matrix

Correlation coefficient between two random variables X and Y, usually denoted by

𝑟(𝑋, 𝑌) 𝑜𝑟 𝑟𝑋𝑌 is a numerical measure of linear relationship between them and is defined as:

𝑟𝑋𝑌 =

A screenshot of a computer

Description automatically generated with medium confidenceA picture containing graphical user interface

Description automatically generated 𝑟𝑋𝑌 provided a measure of linear relationship between X and Y.It is a measure of degree of relationship

Figure

Figure 5

Figure 4 and 5 shows the correlation matrix,

# Data analysis

## Linear regression and random forest analysis

### Linear Regression

Doing a 70;30 split for the train and test data, a linear regression analysis is done the output obtained is however not optimum.

Chart, scatter chart

Description automatically generated

Text

Description automatically generatedFigure 6

As seen in figure 6 it is not a feasible model

### Random forest analysis

The data was analysed using a random forest analysis and fig 7 shows that is more feasible than the linear regression model.

Text, letter

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 7

## Multiple regression analysis

Removing all the variables with a high correlation, the considered parameters in this model is *Weekly Sales, Fuel Price, Week number, Unemployment, Event(categorical), month(categorical)*

Text, letter

Description automatically generatedCreating a train test split of 70:30

Figure 8

### Multiple regression

Now that the data is split into training and testing part. We will take the training data for our regression model.

Chart

Description automatically generated A multiple linear regression model takes the form

### Outputs of this model

Text, letter

Description automatically generated

Figure 9

Table

Description automatically generated

Figure 10

To see how the model performs we do a test and train vs prediction graph Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Figure 12

Figure 11

# References

<https://www.kaggle.com/datasets/yasserh/walmart-dataset>

<https://corporate.walmart.com/about/history>