INFLUENTIAL FACTORS AFFECTING SALES AT WALMART AND PREDICTING SALES AT WALMART USING LINEAR REGRESSION.

Anupama Kadambi
December 8, 2022
ASU ID: 1222559637

Applied Regression Analysis
STP 530

Under the Guidance of Dr. Yi Zheng

CONTENTS

INTRODUCTION

A multinational retail corporation, operator of discounted stores which happens to be one of the world's biggest retail corporations is Walmart Stores Inc. Walmart was founded by Sam Walton in Arkansas in 1962, and primarily focused on small- scale stores in rural areas. It has now become a household name and has Supercenters across the North American subcontinent. As it grew, the company developed many retail formats such as Sam Club's discount warehouses and general stores. Throughout its expansion, various marketing strategies, cost control, customer attention, efficient distribution networks and promotions have played a big part in making it one of the largest grocers in the United States.

Such a retail giant has employed the use of customer metrics and data to improve their revenue and stay ahead of the competition. Metrics such as product reviews, seller rankings, special holiday discounts such as Black Friday Sale, are most widely used to analyze sales which involve predictive analysis also known as sales forecasting to adopt marketing and sales strategies accordingly. Lesser-known factors such as temperature, population density of the region, GDP (gross domestic product), inflation and other factors are usually used in predictive analysis to predict future sales.

Sales forecasting is the process of estimating future revenue by predicting the amount of product sold in a specific time period. It is a projected measure of how market will respond to the company's efforts to increase revenue. The finance department relies on the forecasts to prepare budgets for hiring, capacity planning, etc. Through forecasting, supply chain can be improved, and the company can be better prepared for unwarranted delays or surge in demand for goods. Data analysis includes a time series analysis. Time series analysis is the technique of analyzing a sequence of data points collected over a period of time. Time happens to be a crucial variable when forecasting is the goal. It helps us understand how the trend in data changes, or the data adjusts over the course of time. This is what sets the time series analysis apart from other predictive techniques. Use of machine learning algorithms has become increasingly popular over the past few years since the algorithms are capable of handling big data and produce accurate results. In this project, sales prediction is accomplished with the help of multiple linear regression.

Now, we might wonder how this data, internal and external factors impact the growth on sales in Walmart. This project focuses on determining the influential factors on Walmart sales, and the factors are discussed in depth in the following sections. Various internal and external factors are considered, which affect the weekly sales in Walmart Supercenters in forty-five different stores.

DATA SOURCE

Kaggle is a popular online resource for data scientists and machine learning enthusiasts. Kaggle has found a data driven community and allows millions of people to perform surveys, gather data and publish them, to solve common statistical problems, to perform predictive analysis and to mainly promote collaboration among the online statistical and data science community. Kaggle allows others to review the data set and verify its authenticity.

The 'Walmart' dataset was published by Mr. M Yasser H, on Kaggle. The data was collected by researching Walmart's annual sales for the time period: 5th February 2010 to 1st November 2012 through web scraping and using resources such as Google search engine. The sales were recorded in millions of dollars per week.

Weekly sales from forty-five different anonymized stores were compiled to generate this dataset. This dataset purely comprised of in-store sales, and not sales occurring through the Walmart app or website. It does not comprise of any sales due to online order of products either.

We need to take into account that this data was collected pre-Covid and the trend in online shopping has changed since then. Additionally, Walmart has introduced online shopping since August 2016.With more people shopping from home and shopping online, some of the factors considered in the sales forecast may no longer be relevant or might have little importance. This dataset needs to be updated with more variables including online activity of customers, and other factors related to online shopping. Further, more research needs to go into analyzing the extent of bias in the data set since the data was mainly compiled through web scraping and online data search engine tools. Bias can later be mitigated by introducing weights onto the different predictor variables to improve the accuracy of results.

DATA DESCRIPTION

The Walmart dataset contains weekly sales in millions of USD (United States dollars) recorded from February 2nd, 2010, till November 1st, 2012 in forty-five different Walmart stores. The data set contains eight predictor variables and 6435 observations. It comprises of six numerical variables and two categorical variables. They are discussed in depth in the following section.

The eight predictor variables are as follows:

Numerical predictors:

- Date (Date) - it gives us details about the different weeks during which the sales were recorded.
- Weekly sales (Week_sales) - it is the number of sales in a particular store, and the units of this variable is in millions of USD.
- Temperature (Temp) - it is the day temperature recorded on the day of the sale in the Fahrenheit scale.
- Fuel price (Fuel_pr) - it is the cost of the fuel per gallon on the day of a sale.
- Unemployment rate (Unemp_rt) - unemployment rate is a measure of the number of workers in the labor force who are not employed and are looking for work. The unemployment rate was recorded during the years 2010, 2011, and 2012 for each day of a sale.
- CPI (CPI)- it stands for the consumer price index. It is a measure of the average change in prices paid by consumers in a market basket over a period of time.

Categorical predictors:

- Holiday Flag (Holiday) - it takes the value of either 1 or 0. It takes a value of 1 if aa particular day is a holiday. This includes special holidays such as thanksgiving, Super bowl, Black Friday, Christmas and Labor Day. Else it takes the value of 0. There are 450 Holidays and 4985 normal days out of the total 6435 days for which the sales were recorded.
- Store Number (Str_num) - it is a number which represents the store from which the data was gathered. Number ranges from 1 to 45 depending on the store number.

DATA CLEANING AND DATA PREPROCESSING

Data cleaning is the process of incorrect data or data format, checking for missing values and taking necessary actions to fix these anomalies in the data set. Working with an incomplete dataset can produce incorrect results and leads to misinterpretation of data. Data cleaning also involves removal of irrelevant and duplicates from the dataset. Fixing structural errors is another important aspect of data cleaning. Fixing typos, incorrect capitalization, spelling errors is important to prevent mislabeling the data or categories. Further, an outlier test will help us understand if any of the observations in the dataset have a legitimate reason to appear so. If not, the outlier must be checked for an improper data entry. Improper entries must be removed. This will enhance the performance of the data.

In order to ensure more accurate results and improve interpretation from the plots, standardization of data is a must. Additionally, it simplifies the data visualization task so that people even with little to no statistical knowledge can interpret plots and graphs easily to understand trends in data. There is no one prescribed method for data cleaning.

Data preprocessing is the technique of transforming raw data into a format. That can be understood by the software, and the model being trained in regression analysis. Proper format of data can be used to make the code efficient and visualizations more meaningful and helps glean better insight on the data. It is recommended to check for any or all anomalies and rectify them. The data cleaning in this project consists of checking for missing values and data preprocessing involves changing the format of date, to extract days, months and years. These processes are explained in depth in the following sections.

Handling Missing Data:

The entire dataset and every column in the dataset was scrutinized to check if there were any missing fields. In R software, by employing the function 'which(is.na(variable_name))', we can determine if there are any missing values. If the function returns 'integer(0)', then it means that there are no missing values in that column. The same process is repeated for all the columns, and it is determined that the dataset does not contain any missing values.

Changing the format of date:

The data set comprised of the Date variable in 'day-month-year' format. For forecasting, and visualizing the data, time in terms of week, month or year makes more sense. Extracting the month and year for each day will aid in better time series analysis and is a better way to understand the data. In the R software, we use the following function to extract the month and year from the Date column.

- format(date, "%m") to extract the month and combine it to the existing data set. Every month is recorded in number. Example- February as 2.
- format(date, "%Y") to extract the year and combine it to the existing data set.

EXPLORATORY DATA ANALYSIS

It is crucial understand the data and perform initial investigation on the dataset before modeling and performing regression analysis. Doing so will aid in determining patterns existing in the dataset. It will also throw light upon any existing outliers and makes anomaly detection easier. With the help of summary statistics, and graphical representations, we can check our assumptions related to the data, the data analysis itself, and perform hypothesis testing. Exploratory Data Analysis(EDA) is the crux of preliminary investigation and making sense of the raw data. Below is a simple plot of the raw data, wherein all 6435 observations are plotted.

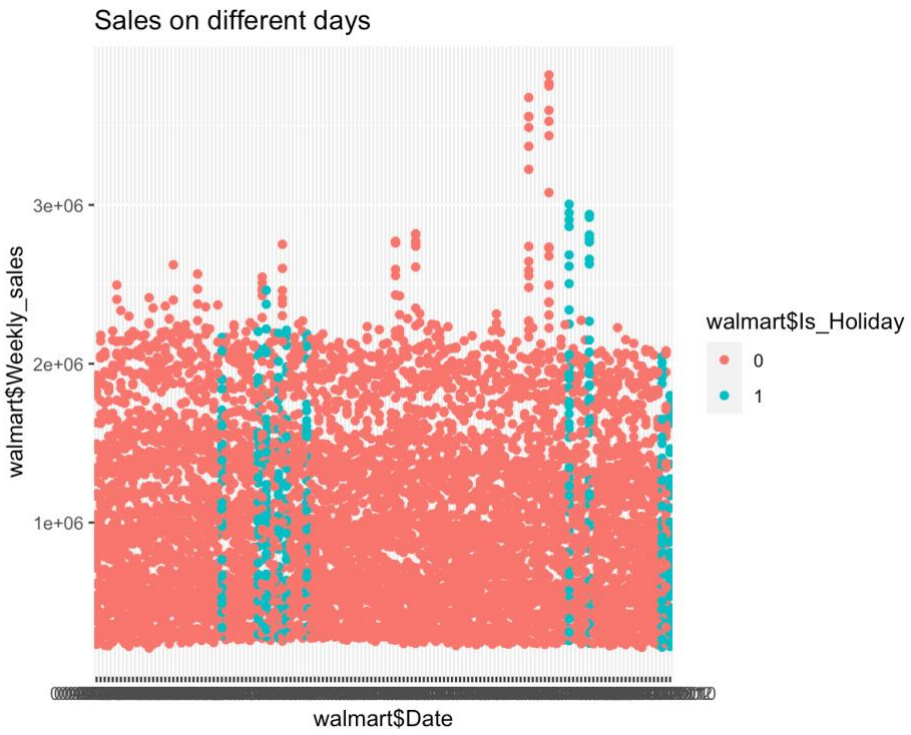


Fig. 1 Plot of sales on Holidays and Normal days.

The plot consists of all 6435 observations, and the color-coded points help us to understand the general trend in sales on normal days versus the sales during holidays. The lighter blue dots represent the sales on a holiday. The red dots represent the sales on normal days and the blue dots indicate sales on holidays. From this preliminary visual inspection, we observe that the maximum sales recorded usually occurred on holidays. But there are some outliers and the highest sales recorded were indeed on the normal days.

Since many data points are overlapping and it is just impossible for us to determine if the sales increased or decreased over time, we need to perform a structured data analysis. A linear regression model will help us obtain a better trend in the sales over the time period, and data transformation will make it easier to understand the plots as discussed in the following sections.

Generating histograms for every numerical predictor:

The histograms contain information regarding the distribution of data. By looking at various histograms, we can conclude if the data is normally distributed or not. We can even get an idea about the existence of outliers and if the data is right0-skewed, left-skewed, etc.

1. Histogram for Weekly sales recorded-

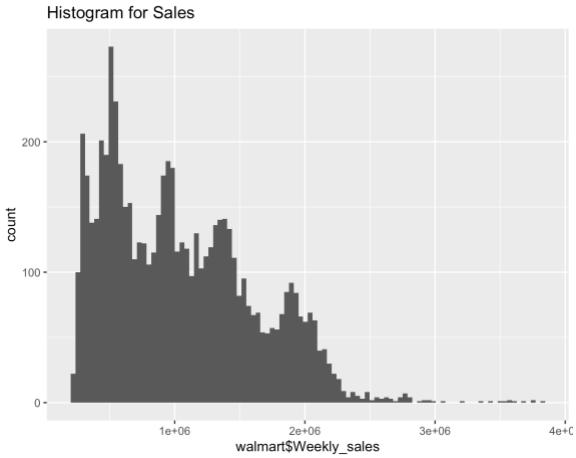


Fig 2. Histogram of Weekly sales in Walmart

From the above histogram it is evident that the Weekly sales at Walmart follows a normal, right-skewed distribution, also known to be positively skewed data. There are no evident outliers in the above distribution.

2. Histogram of Temperature-

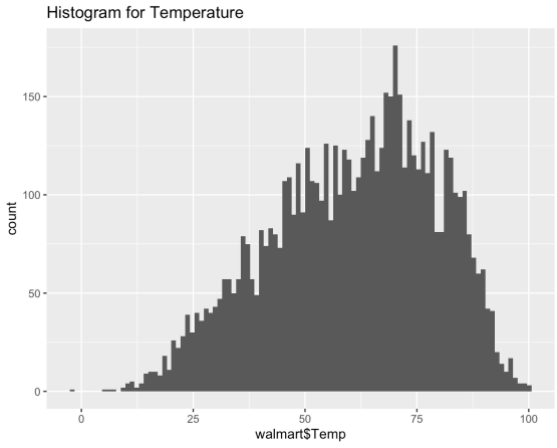


Fig 3. Histogram of Temperature

From the above histogram, it is evident that the temperature distribution is normal, and is left-skewed. It is also called as negative-skewed data and there are no evident outliers.

3. Histogram of Fuel prices-

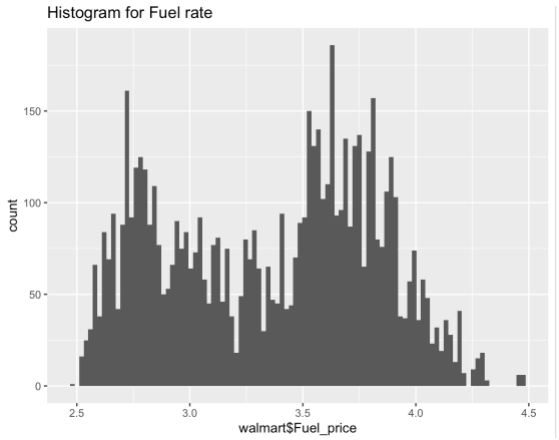


Fig 4. Histogram of Fuel prices

From the above histogram it is evident that the fuel prices follow a bimodal distribution. There's a peak for higher fuel rates and a separate peak for lower fuel rates. There are few outliers observed in the above histogram.

4. Histogram of Consumer price index - (CPI)

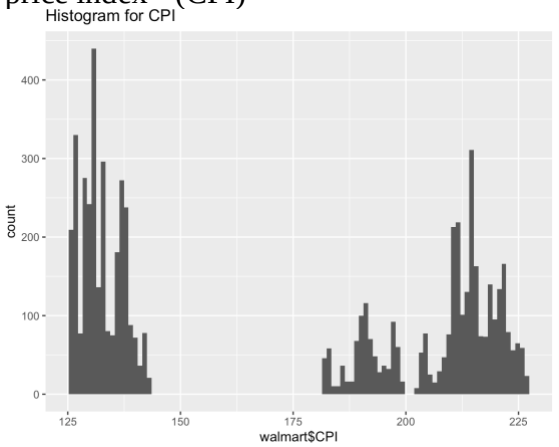


Fig. 5 Histogram of CPI

From the above histogram, we can infer that CPI follows a multimodal distribution. The data has three distinct peaks. There exists more than one normal distribution. Appropriate steps would be to the data with a mixture of two normal distributions. We can either use Least Squares or maximum Likelihood to fix the normal model. The general normal mixing model is

$$M = p\phi_1 + (1-p)\phi_2$$

Source: https://www.itl.nist.gov/div898/handbook/eda/section3/histogr5.htm

Where, p is the mixing proportion (between 0 and 1) and φ1, φ2 are the normal probability density functions.

Boxplots for numerical predictors:

In order to gain more numerical information apart from mean, median and mode, we can generate boxplots to determine the lower quartile, upper quartile, and also understand how the data is spread out. Additionally, boxplots represent the variability and dispersion in data as well. Boxplots are more compact and packed with data compared to histograms. All the information- mean, median, mode, lower quartile and upper quartile are all present in a single plot.

1. Boxplot for Sales in Walmart stores-

Fig 6. Boxplot for Sales

From the above boxplot, it is clear that Weekly sales has many outliers. Mean sales is about 104,696 USD. The median sales is about 960,746 USD, the lower quartile and upper quartiles are 553,350 USD and 1,420,158 USD respectively.

2. Boxplot for temperature-

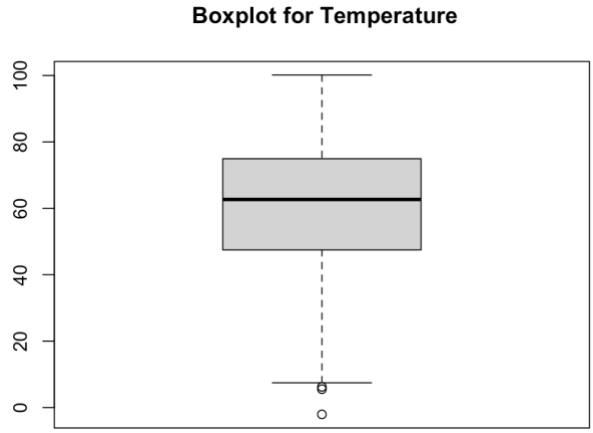


Fig 7. Boxplot for temperature

From the above boxplot, we find out that the mean temperature is about 60 degrees Fahrenheit. The median temperature is about 62 degrees Fahrenheit. The lower and upper quartiles are 47.46 degrees F and 74.94 degrees F. there are few outliers in this boxplot.

3. Boxplot for Fuel price-

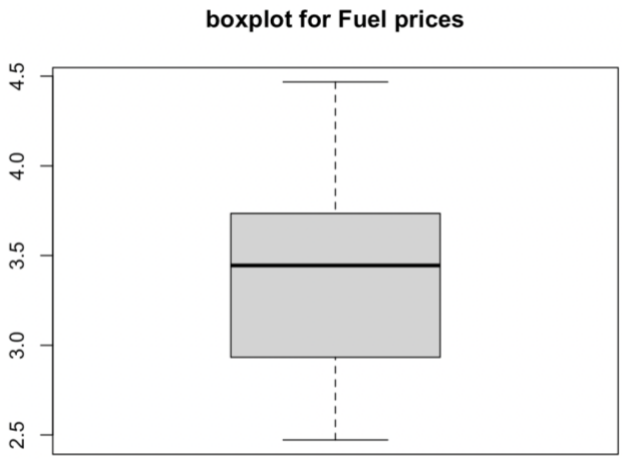


Fig 8. Boxplot for Fuel prices

From the above boxplot we can observe that median fuel price is 3.44 dollars per gallon. The upper and lower quartiles are 4.46 and 2.93 dollars per gallon respectively. There are no outliers in this boxplot. The mean fuel price is 3.35 dollars per gallon.

4. Boxplot for consumer price index (CPI)-

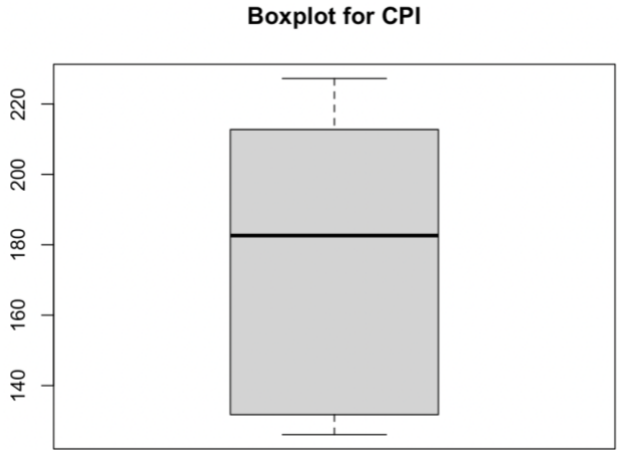


Fig 9. Boxplot for CPI

From the above plot, we observe that no outliers exist for CPI. The mean CPI is 171.57. The median CPI is about 182.6165. the lower and upper quartiles are 131.73 and 212.74 respectively. The data is pretty spread out.

Density plot:

The density plot is representation of the distribution of the data over a continuous period of time. They are used to observe the distribution of a variable in a dataset.
Also known as Kernel density plots, they are a variation of the histograms. The peaks obtained in a density plot display the values concentrated over a particular interval. Density plots are superior to histograms in the sense that the density curves are not affected by the number of bins. So, they also happen to be better at determining the distribution shape. A density plot is obtained for our raw data as follows. Density curves can be thought of as smoothed histograms.

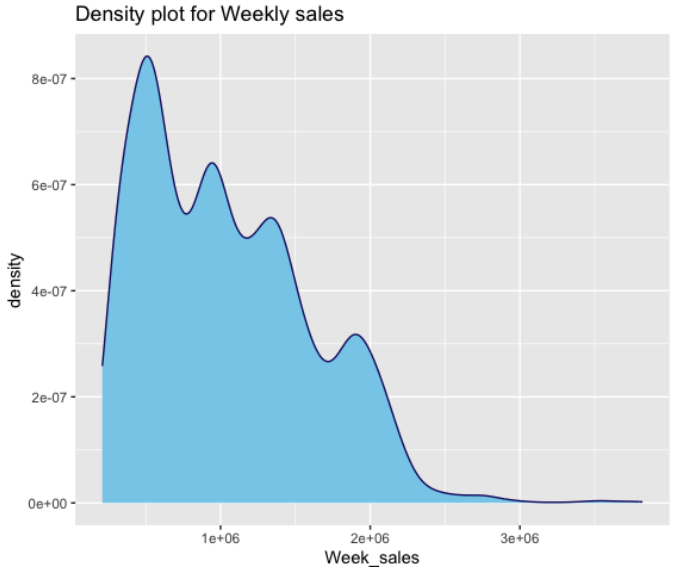


Fig 10. Density plot for Weekly sales at Walmart.

By comparing our density curve with the histogram for the weekly sales, we observe that there are 4 peaks. It is a right-skewed distribution.

Correlation:

Correlation is the measure of the extent to which two variables are linearly related to each other in a dataset. It helps us understand which variables are related to other variables in the dataset. When two or more independent variables in the dataset are moderately or highly linearly dependent, then we say that multicollinearity exists. It becomes increasingly difficult to detect the regression surface. We might encounter multiple standard errors and the estimates from our samples become less accurate and trustworthy. It also becomes difficult to determine the influence of one predictor variable on the response variable.

Use 'cor()' function in R software to obtain the correlation matrix. A positive value for two variable indicates that they are positively correlated. A negative value for two variables indicates that they are negatively correlated. The higher the value for correlation, the stronger is the relationship between the two variables. By using cor(Walmart) the following correlation matrix is obtained.

In addition to the correlation matrix, a correlation heatmap will help visualize the correlation among the variables much better. Darker colors represent weak to no correlation and lighter colors represent strong correlation among the variables.

```
> cor(walmart,method='pearson')
                Str_num   Week_sales       Holiday        Temp      Fuel_pr          CPI    Unemp_rt        Months       Years
Str_num     1.000000e+00 -0.335332015  6.250842e-20 -0.02265908  0.060022955 -0.209491930  0.22353127  8.402191e-19  0.00000000
Week_sales -3.353320e-01  1.000000000  3.689097e-02 -0.06381001  0.009463786 -0.072634162 -0.10617609  7.614332e-02 -0.01837754
Holiday     6.250842e-20  0.036890968  1.000000e+00 -0.15509133 -0.078346518 -0.002162091  0.01096028  1.229958e-01 -0.05678257
Temp       -2.265908e-02 -0.063810013 -1.550913e-01  1.00000000  0.144981806  0.176887676  0.10115786  2.358618e-01  0.06426923
Fuel_pr     6.002295e-02  0.009463786 -7.834652e-02  0.14498181  1.000000000 -0.170641795 -0.03468374 -4.215590e-02  0.77947030
CPI        -2.094919e-01 -0.072634162 -2.162091e-03  0.17688768 -0.170641795  1.000000000 -0.30202006  4.979672e-03  0.07479573
Unemp_rt    2.235313e-01 -0.106176090  1.096028e-02  0.10115786 -0.034683745 -0.302020064  1.00000000 -1.274559e-02 -0.24181349
Months      8.402191e-19  0.076143320  1.229958e-01  0.23586176 -0.042155900  0.004979672 -0.01274559  1.000000e+00 -0.19446452
Years       0.000000e+00 -0.018377543 -5.678257e-02  0.06426923  0.779470302  0.074795731 -0.24181349 -1.944645e-01  1.00000000
```

Fig 11. Pearson's Correlation matrix for Walmart dataset.

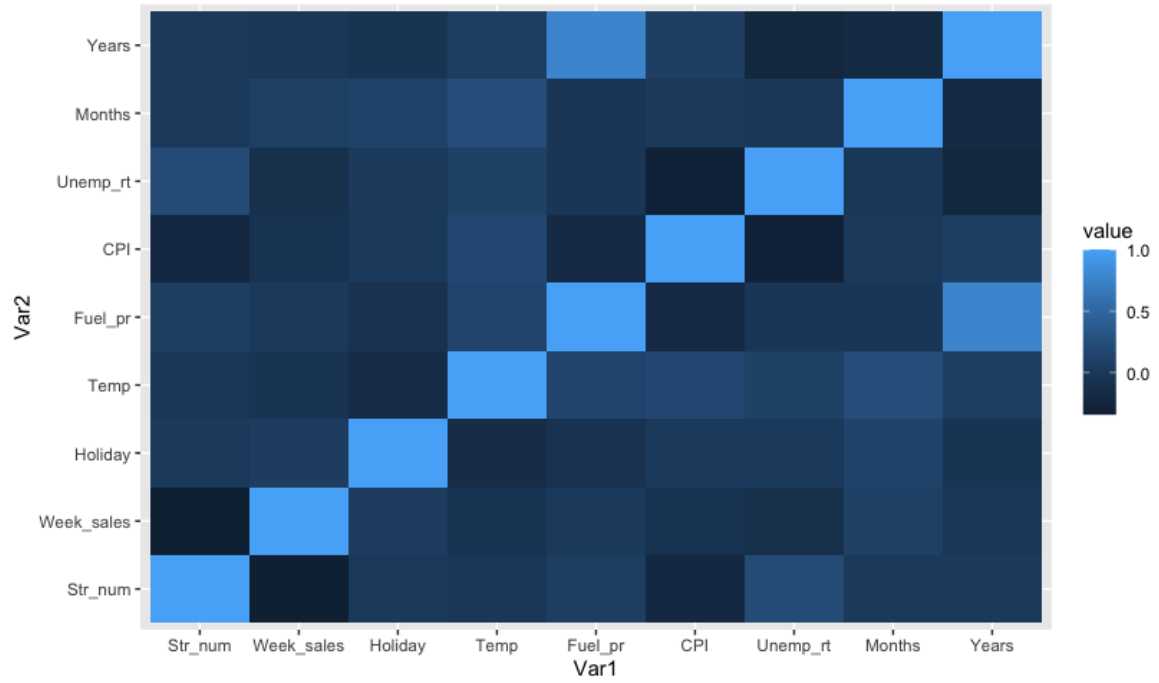


Fig 12. Correlation heatmap

The method chosen to quantify the relationship between the various predictor variables is Pearson's Correlation. Every correlation coefficient varies between -1 and 1. The closer the value is to 1, stronger is the linear relationship between the two variables. -1 indicates perfectly negative linear correlation between two variables. 0 indicates no linear correlation between the variables and 1 represents a strong positive linear correlation.

From the correlation heatmap and correlation matrix, some significant observations are pointed out-

- There is strong positive relationship between Years and Fuel Price. The fuel price increases linearly over the years.
- The consumer price index and the unemployment rate are negatively correlated.
- Weak positive correlation exists between temperature and the months.
- Not much correlation is detected between the other variables.

Thus, there is weak multicollinearity that exists between more than two of the independent variables. Thus, we can conclude that multicollinearity exists.

Feature importance for numerical variables:

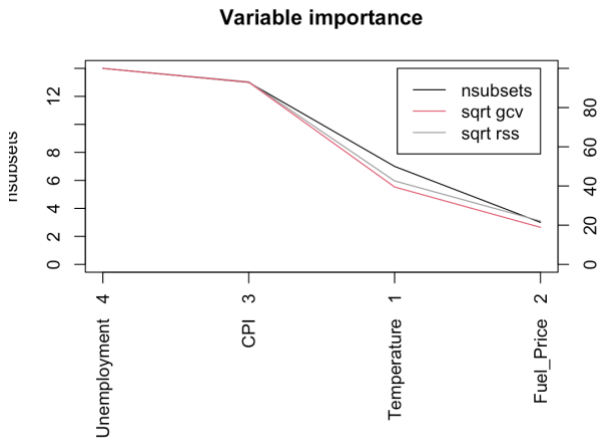


Fig 13 feature importance of numerical predictors

Using the earth() package in Rstudio, the most important numerical predictors are estimated. It happens to be Unemployment rate. Thus, the most important numerical predictor affecting the weekly sales at Walmart is the unemployment rate. The earth(0 package computes the most important feature based on the number of subsets a variable appears in. two more criteria used for calculating the important feature are general cross validation (gcv) and residual sum of squares (rss). Based on all three criteria, unemployment rate happens to be the most important feature affecting the sales.

MODELING

Before selecting a model, which fits the data well, the response variable Weekly sales was regressed over each predictor variable to understand the relationship between them. The following plots are obtained:

1. Weekly Sales ~ Store number:

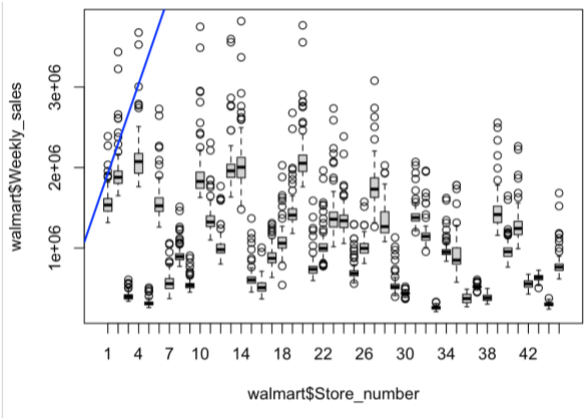


Fig 14 Weekly sales vs Store number

It is not possible to infer much from this plot since Store number is a categorical variable.

2. Weekly sales ~ Fuel rate:

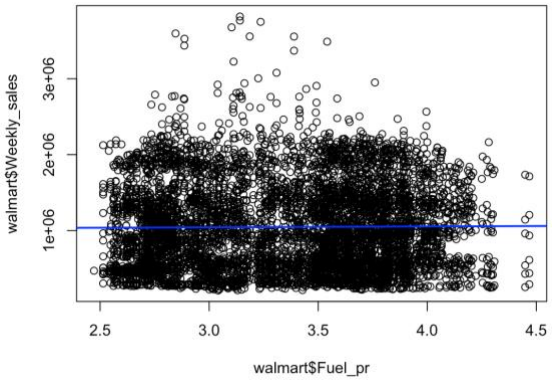


Fig 15 Weekly sales vs Fuel rate

There is little to no effect that fuel process has had on the sales at Walmart. The sales have remained same even though the fuel prices increased over time. However, there are a few outliers.

3. Weekly sales ~ Unemployment rate:

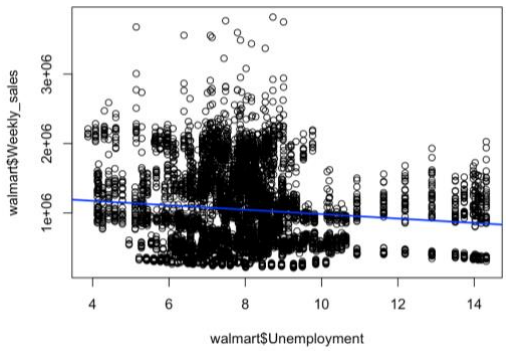


Fig 16 Weekly sales vs Unemployment rate

From the above plot, it is evident that there is a negative correlation between Unemployment rate and Weekly sales. As the unemployment rate increased over the years, the sales have decreased.

4. Weekly sales ~ CPI

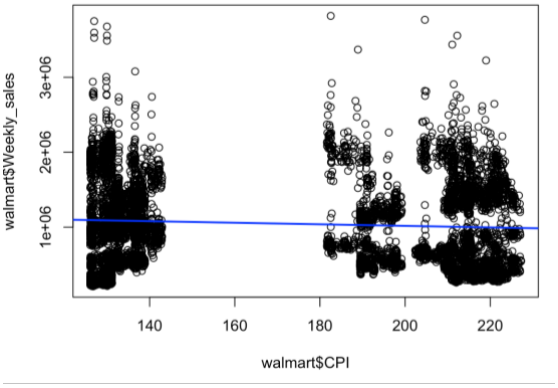


Fig 17 Weekly sales vs CPI

There is slight negative correlation between the consumer price index and sales.
There is very little decrease in sales with increase in CPI over the three years.

5. Weekly sales~ Holiday Flag

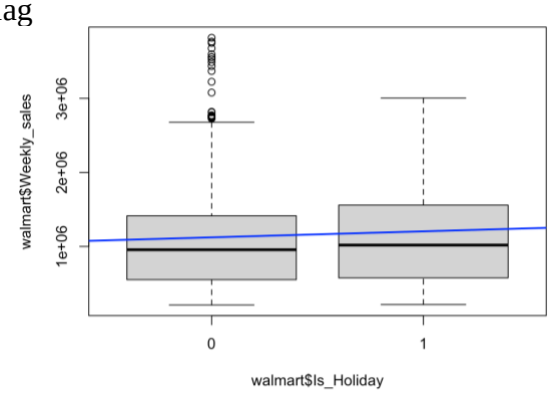


Fig 18 Weekly sales vs Holiday Flag

Since Holiday is a categorical variable, which takes the values of either 0 or 1, it is represented by boxplots. The trend line indicates that there is an increase in sales on Holidays. There are many

outliers in the boxplot at 0, which are normal days. There have been high sales which occurred on normal days as well.

Linear Regression Model:

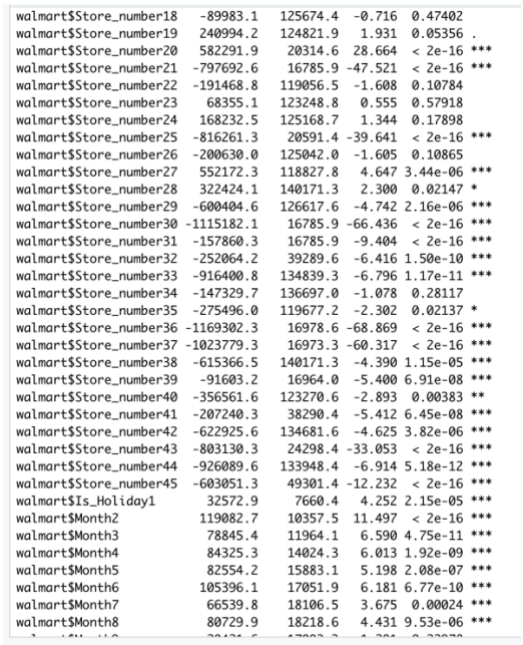
A linear regression model was fit for the data set, using the lm() function.

model<-
lm(walmart$Weekly_sales~walmart$Temp+walmart$Fuel_price+walmart$CPI+walmart$Unemployment+walmart$Store_number+walmart$Is_Holiday+walmart$Month+walmart$Year,data=walmart)

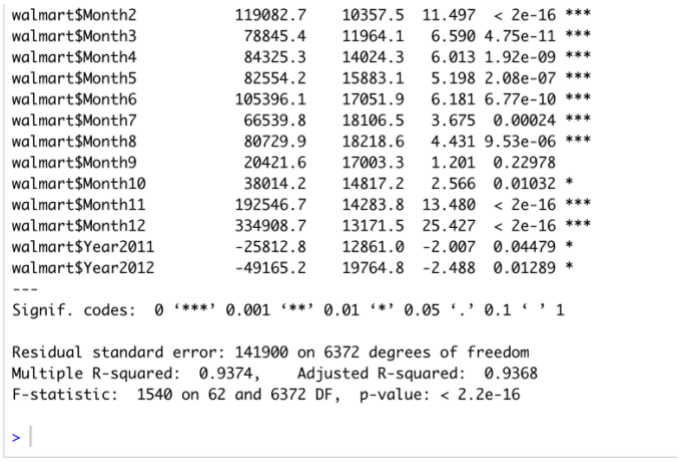
summary(model)

```
Call:
lm(formula = walmart$Weekly_sales ~ walmart$Temp + walmart$Fuel_price +
    walmart$CPI + walmart$Unemployment + walmart$Store_number +
    walmart$Is_Holiday + walmart$Month + walmart$Year, data = walmart)

Residuals:
    Min      1Q  Median      3Q     Max
-655452  -62616   -3827   44565 1626690

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            882763.2   322080.8   2.741  0.00615 **
walmart$Temp              725.1      337.2   2.150  0.03155 *
walmart$Fuel_price       -8696.3    12596.0  -0.690  0.48997
walmart$CPI               3982.5     1528.7   2.605  0.00920 **
walmart$Unemployment     -37137.1     4420.8  -8.400  < 2e-16 ***
walmart$Store_number2    372447.3    16785.0  22.189  < 2e-16 ***
walmart$Store_number3  -1184443.2    17795.6 -66.558  < 2e-16 ***
walmart$Store_number4    830439.0   133521.8   6.220 5.30e-10 ***
walmart$Store_number5  -1289153.2    17831.8 -72.295  < 2e-16 ***
walmart$Store_number6    -35012.1    17627.0  -1.986  0.04705 *
walmart$Store_number7   -838569.4    40144.3 -20.889  < 2e-16 ***
walmart$Store_number8   -712418.4    19211.6 -37.083  < 2e-16 ***
walmart$Store_number9  -1081451.1    19217.9 -56.273  < 2e-16 ***
walmart$Store_number10   720095.1   134681.6   5.347 9.27e-08 ***
walmart$Store_number11  -231523.3    17821.6 -12.991  < 2e-16 ***
walmart$Store_number12     7903.5   140171.3   0.056  0.95504
walmart$Store_number13   784667.8   134119.0   5.851 5.14e-09 ***
walmart$Store_number14   631945.7    49301.4  12.818  < 2e-16 ***
walmart$Store_number15  -580358.0   124829.7  -4.649 3.40e-06 ***
walmart$Store_number16  -972064.2    38388.8 -25.322  < 2e-16 ***
walmart$Store_number17  -336872.5   133956.4  -2.515  0.01193 *
walmart$Store_number18   -89983.1   125674.4  -0.716  0.47402
```

Fig 19 Summary of model



```
walmart$Store_number18   -89983.1   125674.4  -0.716  0.47402
walmart$Store_number19   240994.2   124821.9   1.931  0.05356 .
walmart$Store_number20   582291.9    20314.6  28.664  < 2e-16 ***
walmart$Store_number21  -797692.6    16785.9 -47.521  < 2e-16 ***
walmart$Store_number22  -191468.8   119056.5  -1.608  0.10784
walmart$Store_number23    68355.1   123248.8   0.555  0.57918
walmart$Store_number24   168232.5   125168.7   1.344  0.17898
walmart$Store_number25  -816261.3    20591.4 -39.641  < 2e-16 ***
walmart$Store_number26  -200630.0   125042.0  -1.605  0.10865
walmart$Store_number27   552172.3   118827.8   4.647 3.44e-06 ***
walmart$Store_number28   322424.1   140171.3   2.300  0.02147 *
walmart$Store_number29  -600404.6   126617.6  -4.742 2.16e-06 ***
walmart$Store_number30 -1115182.1    16785.9 -66.436  < 2e-16 ***
walmart$Store_number31  -157860.3    16785.9  -9.404  < 2e-16 ***
walmart$Store_number32  -252064.2    39289.6  -6.416 1.50e-10 ***
walmart$Store_number33  -916400.8   134839.3  -6.796 1.17e-11 ***
walmart$Store_number34  -147329.7   136697.0  -1.078  0.28117
walmart$Store_number35  -275496.0   119677.2  -2.302  0.02137 *
walmart$Store_number36 -1169302.3    16978.6 -68.869  < 2e-16 ***
walmart$Store_number37 -1023779.3    16973.3 -60.317  < 2e-16 ***
walmart$Store_number38  -615366.5   140171.3  -4.390 1.15e-05 ***
walmart$Store_number39   -91603.2    16964.0  -5.400 6.91e-08 ***
walmart$Store_number40  -356561.6   123270.6  -2.893  0.00383 **
walmart$Store_number41  -207240.3    38290.4  -5.412 6.45e-08 ***
walmart$Store_number42  -622925.6   134681.6  -4.625 3.82e-06 ***
walmart$Store_number43  -803130.3    24298.4 -33.053  < 2e-16 ***
walmart$Store_number44  -926089.6   133948.4  -6.914 5.18e-12 ***
walmart$Store_number45  -603051.3    49301.4 -12.232  < 2e-16 ***
walmart$Is_Holiday1       32572.9     7660.4   4.252 2.15e-05 ***
walmart$Month2           119082.7    10357.5  11.497  < 2e-16 ***
walmart$Month3            78845.4    11964.1   6.590 4.75e-11 ***
walmart$Month4            84325.3    14024.3   6.013 1.92e-09 ***
walmart$Month5            82554.2    15883.1   5.198 2.08e-07 ***
walmart$Month6           105396.1    17051.9   6.181 6.77e-10 ***
walmart$Month7            66539.8    18106.5   3.675  0.00024 ***
walmart$Month8            80729.9    18218.6   4.431 9.53e-06 ***
```



```
walmart$Month2           119082.7    10357.5  11.497  < 2e-16 ***
walmart$Month3            78845.4    11964.1   6.590 4.75e-11 ***
walmart$Month4            84325.3    14024.3   6.013 1.92e-09 ***
walmart$Month5            82554.2    15883.1   5.198 2.08e-07 ***
walmart$Month6           105396.1    17051.9   6.181 6.77e-10 ***
walmart$Month7            66539.8    18106.5   3.675  0.00024 ***
walmart$Month8            80729.9    18218.6   4.431 9.53e-06 ***
walmart$Month9            20421.6    17003.3   1.201  0.22978
walmart$Month10           38014.2    14817.2   2.566  0.01032 *
walmart$Month11          192546.7    14283.8  13.480  < 2e-16 ***
walmart$Month12          334908.7    13171.5  25.427  < 2e-16 ***
walmart$Year2011         -25812.8    12861.0  -2.007  0.04479 *
walmart$Year2012         -49165.2    19764.8  -2.488  0.01289 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 141900 on 6372 degrees of freedom
Multiple R-squared:  0.9374,    Adjusted R-squared:  0.9368
F-statistic:  1540 on 62 and 6372 DF,  p-value: < 2.2e-16

>
```

Fig 20, 21 Coefficeints of model, cntd.

Looking at the summary, the R-squared value is 0.9347, which means that this model explains for 93.74% of the variance. This seems to be a good fit for the data. Additionally, the p-value is 2.2e-16 which means the model is significant. Adjusted R sq is 0.9368.
A plot of the fitted values by the model for the data versus the actual values is shown below.

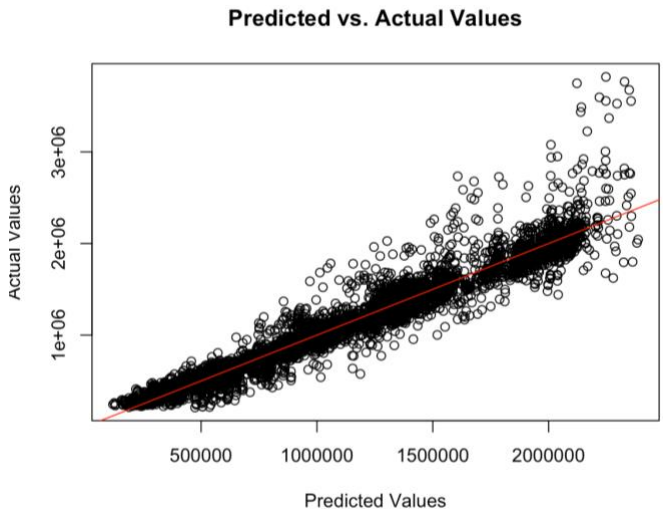


Fig 22 Plot of predicted values vs Actual values

The actual values versus the predicted values follow a linear trend. The red line represents the values fit by the model. Visually, the model is a good fit for the data. However, there are certain outliers visible in the top right part of the plot. Advanced diagnostics are required to analyze the outliers. It is important to deal with outliers because they can adversely affect the model and may not result in accurate predictions. Outlier is defined as an observation which has a very high residual.The mean, standard deviation and correlation are highly sensitive to the outliers. One way to eliminate outliers is to Use the subset() function, it is possible to simply extract the part of the dataset between the upper and lower ranges leaving out the outliers.

This model is further subjected to advanced diagnostics. These diagnostics are discussed in the following sections.

DIAGNOSTICS

Regression diagnostics are used to evaluate the model assumptions and investigate whether there are observations with a large, undue influence on the analysis. Some assumptions are taken into account while performing the regression analysis, they are discussed below.

- Linearity: it is assumed that the relationship between the response variable and predictor is linear.
- Homoscedasticity: it is assumed that the variance is constant for a residual for any value of a predictor.
- Independence: every observation is independent of each other.
- Normality: it is assumed that Y is normally distributed for any value X

Diagnostic plots are obtained for the linear regression model

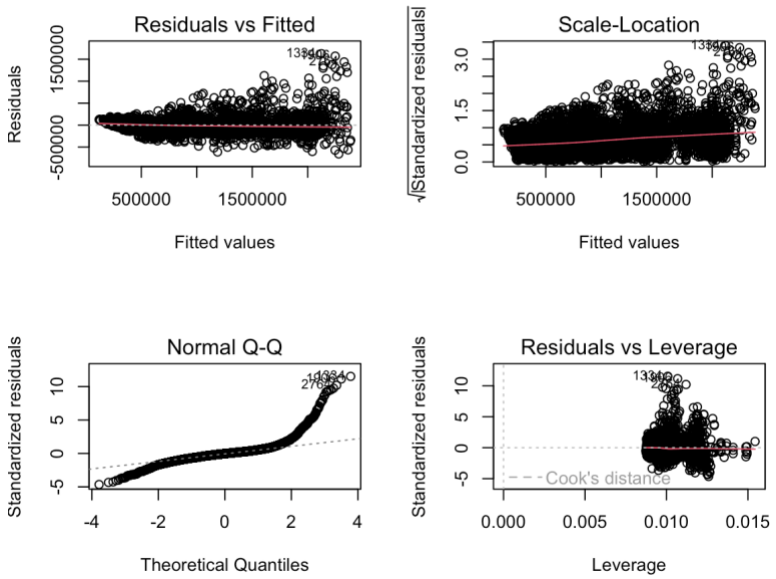


Fig 23 Diagnostic plots using plot(model) function

The first plot is that of residuals versus fitted values. This plot helps in checking if the linearity assumption is met. Our assumption is that there is homoscedasticity. Unfortunately, for the linear model, there is heteroscedasticity, and this violates the linearity assumption. Ideally, it is desired to obtain points randomly scattered about the 0 line, with no definite pattern. The residuals versus fitted values shows that there is a cone like structure, and thus heteroscedasticity is present.

The second plot is that of the Q-Q plot. It is desired that the points in the Q-Q plot follow the dotted line and stay within the line. Then the assumption that the response variable is normal, is true. But in the above Q-Q plot, though most of the points stay well along the dotted line, there are some points that deviate from the line. The assumption of normality is violated. In order to mitigate this situation, data transformation such as natural logarithm or squaring of data will produce better results.

The third plot is a scale-location plot. This plot again helps in checking the presence of homoscedasticity. There is slight upward trend here, indicating heteroscedasticity.

The final plot is of the Cooke's distance. It can also be called as an influence plot. The radius of the circles is proportional to the Cooke's distance. The closer the points are to the right of the plot, the more influence they tend to have on the regression curve. More points are concentrated on the right, which means that there are many influence points. The higher the points are from 0, the more residual they have. Below is a clearer picture of Q-Q plot and residual plot.

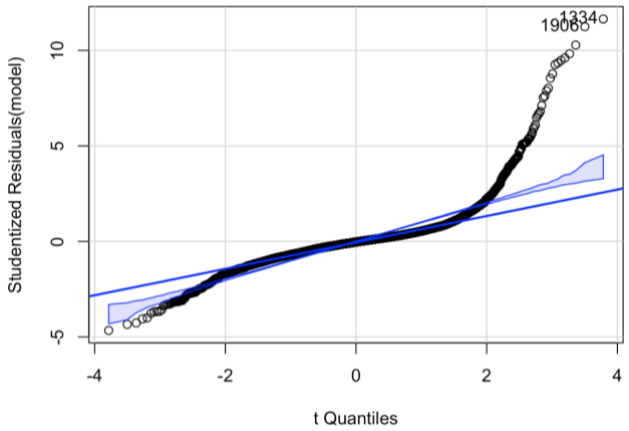


Fig 24 Q-Q Plot

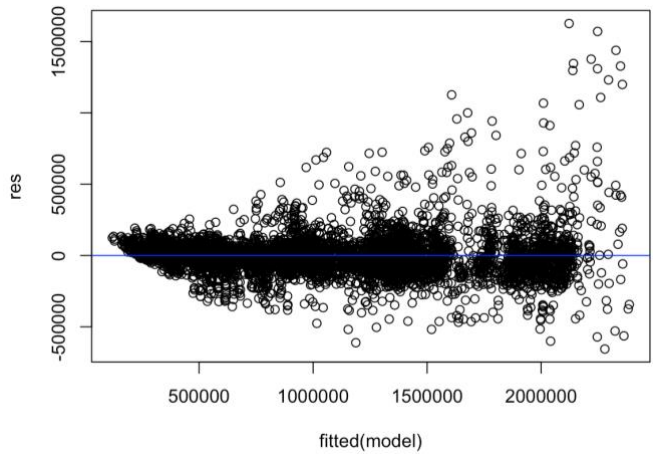


Fig 25 Residual plot

Marginal model plots:

Marginal model plot were proposed by Cook and Weisberg, to understand the marginal relationship that each predictor has on the response variable. The marginal plots have the response variable on the y axis and the independent variable/ predictor on the x axis. the smooth fit function is labeled 'data', and the predicted values are represented by a dotted red line.

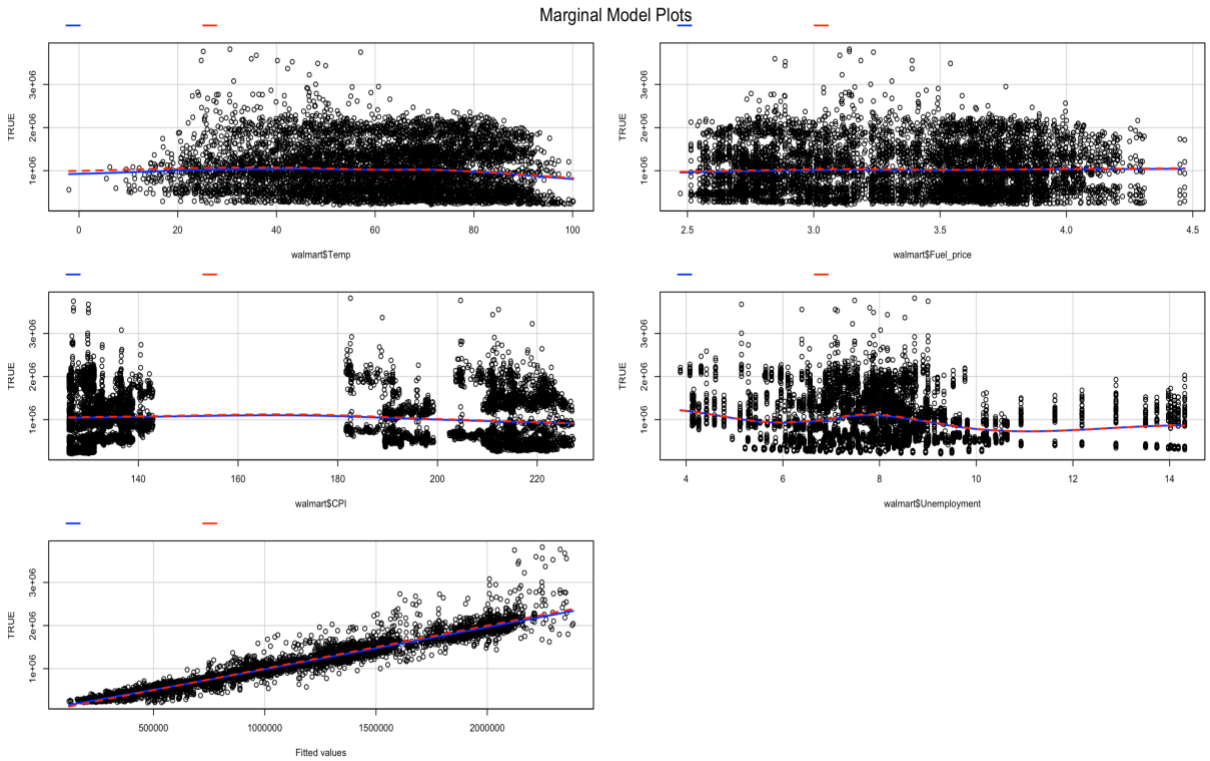


Fig 26 Marginal Model plots

The two functions in the above marginal model plots coincide, then it means that the model is a good fit. The model is a good fit for the Walmart dataset.

Influence plots:

In order to measure the extent of outliers, and to understand the extent of leverage they have on the regression line, influence plots come into picture. The influence plot obtained has the residuals on the y axis and the leverage on the x axis. several points have very high residuals, but have weak to moderate leverage on the regression line. The points 431 and points 4002 have the highest leverage. Points 1334 and 1906 have the highest residuals. These points had sales higher than average.

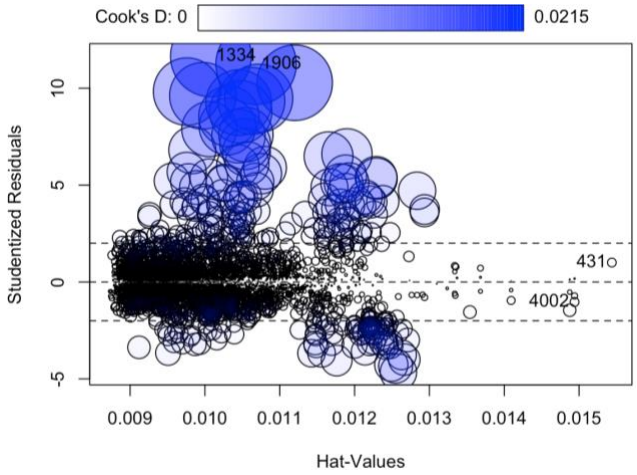


Fig 27 Influence plot

MODELING: RANDOM FOREST

Random Forest is a machine learning algorithm used in regression and classification problems.

This algorithm is based on decision trees, but the advantage of random Forest over decision trees is that the problem of over-fitting is mitigated. Rather than calling it a machine learning algorithm, it could be termed as an "ensemble" method and it means a combination of multiple models, which are collectively used for prediction. The algorithm is as follows:

- A certain number of records are selected from the dataset having I number of records.
- Every sample has an individual decision tree contracted for itself.
- An output is generated from every decision tree
- The output of all the decision trees for a particular feature are averaged out.
- This way, overfitting is prevented.

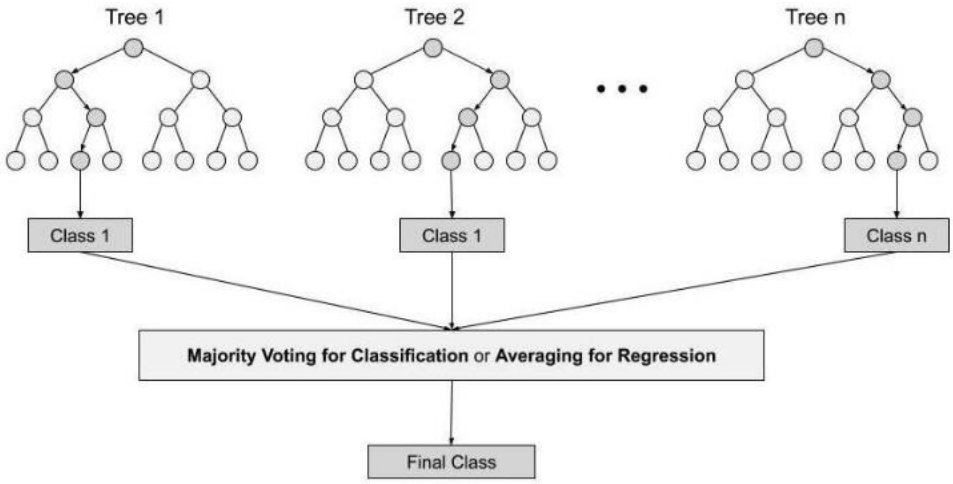


Fig 28 Random Forest model

In order to improve the accuracy of prediction for the Walmart dataset, a random Forest model was fit for the data using the 'randomForest' package in Rstudio. Additionally, feature importance for all predictors was determined.

rf.fit <- randomForest(Weekly_sales ~ ., data=walmart, ntree=170,keep.forest=FALSE, importance=TRUE)

rf.fit

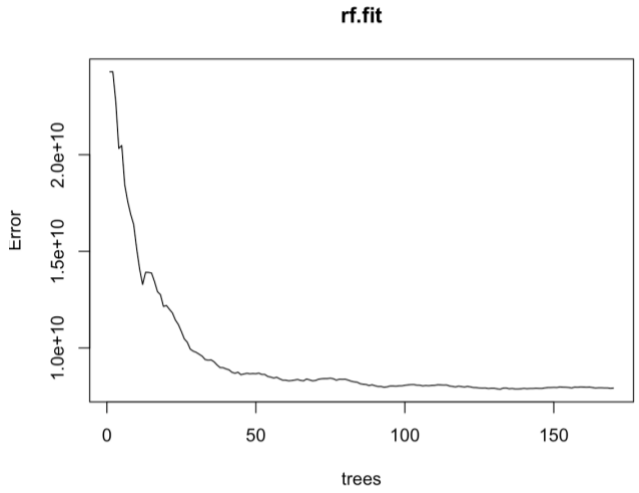


Fig 29 Random Forest Number of trees vs error

The above plot is generated for the random forest model. A total of 170 trees were involved in decreasing the error and producing accurate predictions. For the dataset, to minimize the error towards zero, utmost 90 trees are enough to accurately make predictions on the dataset i.e., to predict future sales.

```
Call:
 randomForest(formula = walmart$Weekly_sales ~ ., data = walmart,      ntree = 170, keep.forest = FALSE, impo
rtance = TRUE)
               Type of random forest: regression
                     Number of trees: 170
No. of variables tried at each split: 3

          Mean of squared residuals: 7919188474
                    % Var explained: 97.51
```

Fig 30 Summary for Random Forest model

The accuracy of predicitons by using the random forest model has increased to 97.7 percent. Thus RandomForest model is much better option than a simple linear regression model for the Walmart dataset. Since the Random Forest model used Bagging and Boosting for random sampling, the outliers are actually bagged and the output for these are averaged out. Thus, when dealing with datsets having high outliers, and the kind of outliers that cannot be deleted, RandomForest is a fantastic model to use, to make accurate predictions. Thus, I chose to use the Random Forest model and it has clearly performed much better than the linear regression model.

The feature importance for different variables is depicted next. The most important feature happens to be the Store number. In my opinion, the geographical location of the store must be a key contributor to in-store sales where people come in person to buy products. For example, in states receiving fewer snowfall such as Arizona, people may be more inclined to shop in person than a state receiving heavy snowfall, making it difficult to commute and reduces the number of in-person shopping.

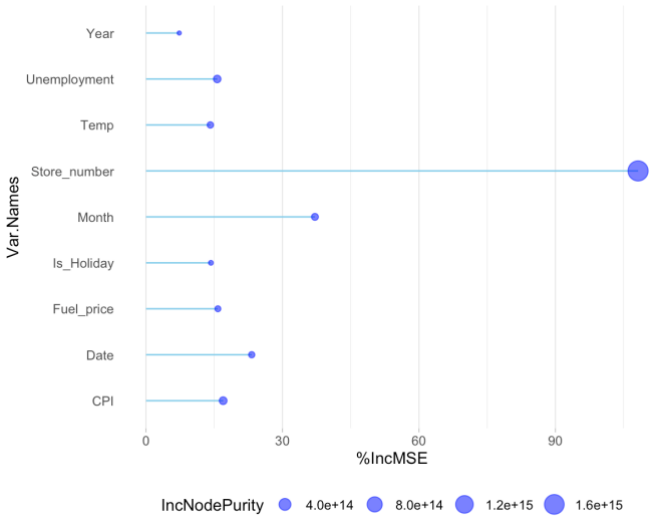


Figure 31 Feature importance using Random Forest

RESULTS AND DISCUSSION

In essence a linear regression model is fit for a Walmart sales dataset to make predictions on future revenue. Though the R squared value was satisfactorily high, the presence of unavoidable outliers reduced the confidence or accuracy of the predicted results. Further, the assumptions of linearity and normality were violated. The most important numerical predictor was found to be unemployment rate. The linear regression model produced an R square of 0.9374.(93.74% of total variance), and adjusted R squared of 0.9368. The model can explain 93.68% of total variance. To improve accuracy in linear regression models and to mitigate the violation of assumptions, log transforming and squaring of data is desirable. In future, Forward selection of variables, or backward selection of variable using VIF can help select the most important variables for prediction. Additional diagnostics and test for goodness of a fit can give more insight towards prediction with outliers. dealing with data sets having high outliers, and the kind of outliers that cannot be deleted, Random Forest is a fantastic model to use, to make accurate predictions. Thus, I chose to use the Random Forest model and it has clearly performed much better than the linear regression model. The accuracy of prediction with random forest was 97.5%. The most important feature in the entire dataset affecting the Weekly sales at Walmart is the Store number, i.e., the stores itself.

REFERENCES

1. Dr. Yi Zheng STP 530 lecture slides and assignments.
2. John Luke Gallup, Portland State University Portland, OR *Added-variable plots for panel-data estimation.*
3. R documentation: https://www.rdocumentation.org/
4. Julia Kho (2018, October 19) *Why Random Forest Is My Favorite Machine Learning Model* Medium. *https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706*
5. Sruthi E R (2021, June 17) *Understanding the Random Forest* Analytics Vidhya https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/