**CIND820 Capstone Project: Walmart Sales Forecasting**

**Literature Review, Data Description and Approach**

**By: Hasib Chowdhury (500439135)**

**Supervisor: Dr. Uzair Ahmad**

**Date of Submission: June 6, 2022**

**Revised Abstract: (modification from the original abstract)**

For this project, I will use the Walmart Sales Forecasting dataset from the Kaggle website (<https://www.kaggle.com/code/aslanahmedov/walmart-sales-forecasting/data>) to conduct the research to fulfill the “Data Analytics Project (CIND820)” requirements. The data that will be used are the clean\_data.csv file from the site, which takes into account many different open data sources from Kaggle involving Walmart sales forecast.

This dataset has 420,213 rows with 23 columns (attributes). In total, there are 45 different Walmart stores with about 100 departments in each of 3 different types (A,B,C) and different sizes which will be analyzed throughout the project.

The data is provided on a weekly basis from February 2010 to October 2012, over a 2.5-year period, and each week has its own set of features such as different CPI (consumer price index), unemployment rate as well as whether a holiday was a factor in its performance.

The main themes for this project will be regression and classification (whether the Walmart was successful in sales or not or any other factor), as well time-series in terms of predictive analytics.

The 3 main questions that will be investigated for this project are whether the seasons are a factor in terms of the sales for this dataset. How the sales have been in that week if there was a specific holiday during that week such as Christmas, Thanksgiving, New Years, etc. and the final question would be what would be the method (for example either regression or classification) that we can use to improve the sales for a better outcome for Walmart in terms of sales forecasting.

One other thing that will be examined is to predict the store’s sales at a particular week and if there are any impacts on it. Lastly, also whether the temperature or fuel price for that week had influenced the sales in addition to other things such as the CPI (which varies depending on the specific store number in the region).

The regression methods will be employed to estimate the future sales of the store, and time series forecasting to see whether there is a trend in the data from February 2010 to October 2012 or specific time periods in between the years.

The software tools that will be used are Python & R.

**Introduction:**

For this project, I decided to work on the Walmart Kaggle dataset for sales forecasting throughout its many stores in certain regions as there are 45 different stores and about 90-100 departments in each specific location. In addition, this dataset takes into account several different variables including weekly sales (which is the target variable) within the time frame from February 2010 to October 2012 and whether certain holidays (such as Christmas, Thanksgiving, etc.) have a factor in terms of sales forecasting for the big box corporation. In this project, I am going to analyze the dataset that is given in further detail and predict how the sales forecast will be like in the future years and also to see if the weekly sales at particular locations are impacted by either time-based (during different seasons such as spring, summer, winter) or space-based factors (how big a particular location is) affect the sales.

**Literature Review:**

The literature review for the Walmart sales forecasting dataset will consist of regression and sales forecasting to predict the outcomes for Walmart in the near future in terms of its sales. There are several seasons that affect sales in a given time of year that are significantly more or less than averages depending on a variety of factors. In this scenario, if Walmart is unaware about the seasons (eg Spring, Summer, Fall, Winter), it can lose too much money which in effect can cause low profit for the company. Predicting future sales is one of the most crucial plans for a big company such as Walmart since sales forecasting gives us an idea to the company for arranging stocks, calculating revenue, and deciding to make a new investment.

In an article by Mentzer, J. T., & Cox, J. E.(1984), it states that “Sales Forecasting is quickly becoming one of the most crucial aspects of planning for companies and that a survey of 175 midwestern businesspeople indicated that 65% thought sales forecasting was very important to their company’s success and an additional 28% said forecasting was important, although not critical” (Mentzer et al, p. 27). This article from the 1980’s tells us that Walmart’s (being a big corporation) sales forecasting is one of the most important aspects in terms of sales, especially in the United States since there are more than 5,000+ locations in this country and over 6,000 locations worldwide.

“Forecasts have traditionally served as the basis for planning and executing supply chain  
activities. Forecasts drive supply chain decisions, and they have become critically important  
due to increasing customer expectations, shortening lead times, and the need to manage  
scarce resource” (Boone, T. et. al, p. 170)

In terms of this data, Walmart is a big corporation that sells several different products throughout North America and Worldwide. It also operates a chain of hypermarkets, in this scenario Walmart has provided data (in csv format) with a combination of 45 different stores (ranging from 3 different types – A,B,C as well as different sizes varying from 34,000 square feet all the way up to 200,000+ square feet) including store information and monthly sales. This data is provided on weekly basis, going back from February 2010 all the way to July 2013. Walmart tries to find the impact of holidays on the sales of store depending on the different time of year (eg. Spring, summer, fall, winter) eg. holidays such as Christmas, Thanksgiving, Super bowl, Labor Day, Family day, etc.

“Firms use product sales forecasting as a foundation to estimate sales revenue and make decisions regarding production, operation and marketing strategies” (Fan, Z. et. al, p. 90)

“Through product sales forecasting, firms can create a plan for marketing, sales management, production, procurement, logistics and so on to improve their economic benefits and reduce losses caused by weaknesses in the production plan” (Fan, Z. et al, p. 90).

In this scenario, Walmart can maintain its products in a yearly fashion and make good marginalized profit when it comes to their certain operations and marketing strategies in certain locations. They make good money off customers in things like electronics, furniture, etc. especially during the end of year during the Christmas holidays and black Friday/boxing day.

In August 2019 article study, it showed that “Walmart Inc. second-quarter sales came in ahead of market forecasts however, the corporation saw income fall by a third amid shrinking margins and rising costs. The world’s biggest retailer posted revenues of USD $138.6 billion for the quarter, 0.6% growth year-on-year and above analysts predictions of around USD $136.9 billion.” (Proactiveinvestors, Aug. 2021). This most likely occurred to inflation happening within the US economy as certain states have higher inflation than others but it still maintained to have over $100 billion in revenue over 3 months of the year in 2019. The sales more or less may have fluctuated afterwards in 2020 due to the COVID-19 pandemic as certain states were under restrictions.

In a challenge done on Kaggle in 2020, “The challenge was to predict future sales of Walmart products based on past sales. The competition was organised in two parallel challenges, one was a 28-day challenge and the other one was series of quantile estimates for the same period” (de Rezende, R., et al., p. 1) This study shows how the company did in that specific month in terms of the sales that it received from the consumers. Their “approach was conceived mainly to model product-store sales, as these are the most relevant for supply-chain decisions” (de Rezende et al, p. 1). This tells us how different Walmarts in different regions across the country do in terms of sales when it comes to supply-chain inquiries, the Walmart Kaggle dataset is consistent with this relation, based on the 3 years of data that is given.

**Data Description:**

There are a total of 23 attributes in this dataset. The attributes in the Walmart sales forecast are as follows:

1. **Store:** This is the first column of the dataset and is numbered between 1-45, indicating the 45 different Walmart stores in the region (ranging from types A,B,C)
2. **Department:** This is the second column of the dataset and is numbered between 1-98, indicating the 98 different departments in each store.
3. **Date:** This is the third column of the dataset and is in the time span between February 2010 to October 2012 (Friday of every week).
4. **Weekly Sales:** This is the fourth column of the dataset, this is the target variable of the dataset since it tells us how much each store (along with the department made) in that particular week.
5. **IsHoliday:** This is the fifth column of the dataset and is based on True/False classification value based on the specific week in which there was a holiday (eg. Christmas, New Year’s).
6. **Temperature:** This is the sixth column of the dataset and represents the temperature on that specific day, the values range from -2 degrees to 100 degrees.
7. **Fuel Price:** This is the seventh column of the dataset and represents the fuel price on that specific day, the values range from 2.472 to 4.468.
8. **Markdowns 1-5:** These are columns 8 to 12 in the dataset, these 5 variables represents the promotional markdowns that Walmart is running during that specific week.
9. **CPI (Consumer Price Index):** This is column 13 of the dataset and values range from 126 to 227.
10. **Unemployment:** This is column 14 of the dataset and tells us the unemployment rate in that particular Walmart store, and the values range from 3.879 to 14.313
11. **Type:** This is column 15 of the dataset and tells us whether it’s a type A, B or C Walmart in that particular region.
12. **Size:** This is column 16 of the dataset and tells us how big that particular store is, values range from 34,875 to 219,622 .
13. The next 4 columns are the 4 different holidays or the biggest season for sales, being Super Bowl weekend, Labor Day, Thanksgiving, Christmas
14. **Week:** This is column 21 of the dataset and tell us the specific week of that year the Walmart made sales in, values start from early January (week 1) to late December (week 52)
15. **Month:** This is column 22 of the dataset and tells us the specific month, varying from month1 (January) to month 12(December) of that particular year.
16. **Year:** This is the last column(column 23) and tells us the year of the sales in the Walmart, this dataset spans 3 years (2010, 2011, and 2012)

The Summary of the dataset is as follows:

A close-up of a document

Description automatically generated with medium confidence

The following is a boxplot of temperature:

Chart, box and whisker chart

Description automatically generated

The minimum value for Temperature in this dataset is 5.54 and the maximum value is 100 within the region. The median is 62 with outliers ranging being a negative value of -2.06.

The following is a boxplot of the fuel price:

Chart, box and whisker chart

Description automatically generated

The minimum value for Fuel Price in this dataset is 2.472 and the maximum value is 4.468 within the region. The median is 3.452 with no outliers, this tells us that the fuel price is more consistent than the temperature.

The following is a boxplot for the Consumer Price Index (CPI):

Chart, box and whisker chart

Description automatically generated

The minimum value for CPI in this dataset is 126.06 and the maximum value is 227.23 within the region. The median is 182.35 with no outliers.

The following gives us a pie chart of how the sales were during Walmart’s 3 years of operation based on the weekly sales of the given years:

Chart, pie chart

Description automatically generated

The above piechart tells us the sales during the 3 years, which are consistent as they range from 30-36% each, indicating the sales growth each year while in operation.

**Approach:**

The approach that I made when I extracted this dataset was to first import the dataset onto R and use the data from Kaggle website (clean\_data.csv file). I then explored the data in more detail by identifying the different types of variables in it (23 total, some categorical and some numerical). The target variable is the Weekly Sales (4th column) which is what we are trying to predict in terms of future sales forecasting for Walmart depending on the region it is in (45 different stores and approx. 100 different departments). The values range from $0 to ~$700k for the 3 years (2010-2012) that its given and it gives us a fairly rough indication (using regression) to predict how much the market may make in profit in the succeeding years (2013 onwards) and what the company can do to improve its sales especially during the holidays and during its peak seasons throughout the year. In terms of the steps for approach, I did the following:

* Step 1: Import the dataset from Kaggle
* In this step, I took the data for Walmart Sales Forecasting and imported it onto RStudio.
* Step 2: Data Processing/Extraction
* In this step, I took the data and observed the different types of variables it had and sorted it accordingly (i.e. numeric or categorical)
* Step 3: Data Cleaning
* In this step, it was to see whether different type of attributes needed to be added/removed, in this case I kept most of it as is.
* Step 4: Test/Train Model
* This step was used to split the data, I used the 70% training and 30% testing criteria to evaluate the dataset, this narrowed the dataset from over 400,000 rows into 294,000 rows for training and 126,000 rows for testing for better accuracy of the model.
* Step 5: Build a multiple linear regression model
* I used the weekly sales as the target variable and used it as dependent variable and the other attributes such as temperature, fuel price and CPI as the independent variable
* Step 6: Result and Discussion
* This is the part where I analyze the different types of results within the data such as boxplots, scatter plot, pie chart, etc. along with the summary table to see different types of results (ie max, median, min, etc.) and see if any outliers are there and if they have an impact of the sales forecasting
* Step 7: Conclusion
* This is the final part of the approach and it’s based on the evidence from the data and the results obtained through the regression model and other factors.

**References:**

Proactiveinvestors. (2021). *Proactive Investors: Walmart sales beat forecasts but profits drop by a third*. Newstex.

Boone, T., Ganeshan, R., Jain, A., & Sanders, N. R. (01/2019). *Forecasting sales in the supply chain: Consumer analytics in the big data era* Elsevier. doi:10.1016/j.ijforecast.2018.09.003

Fan, Z., Che, Y., & Chen, Z. (05/2017). *Product sales forecasting using online reviews and historical sales data: A method combining the bass model and sentiment analysis* Elsevier. doi:10.1016/j.jbusres.2017.01.010

Mentzer, J. T., & Cox, J. E. (01/1984). *Familiarity, application, and performance of sales forecasting techniques* Elsevier. doi:10.1002/for.3980030104

de Rezende, R., Egert, K., Marin, I., & Thompson, G. (12/2021). *A white-boxed ISSM approach to estimate uncertainty distributions of walmart sales* Elsevier. doi:10.1016/j.ijforecast.2021.11.006