Capstone Project – Walmart

Submitted by: Amit Jagtap

Email ID: amitjgtp2@gmail.com

I. Problem Statement

A retail store that has multiple outlets across the country is facing issues in managing the inventory. To match the demand with respect to supply, they would like to predict the sales and demand accurately. Here historical sales data for 45 stores located in different regions is available. The business is facing a challenge due to unforeseen demands and runs out of stock some times, due to the unsuitable machine learning algorithm and forecasting model. An ideal ML algorithm and forecasting model will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc.

II. Data Description

This is the historical data that covers sales from 2010-02-05 to 2012-11-01 in the CSV file Walmart Dataset. The 'Walmart DataSet.csv' contains 6435 rows and 8 columns. Within this file you will find the following fields:

Feature Name	Description
Store	Store number
Date	Week of Sales
Weekly_Sales	Sales for the given store in that week
Holiday_Flag	If it is a holiday week
Temperature	Temperature on the day of the sale
Fuel_Price	Cost of the fuel in the region
CPI	Consumer Price Index
Unemployment	Unemployment Rate

III. Data Pre-processing Steps and Inspiration

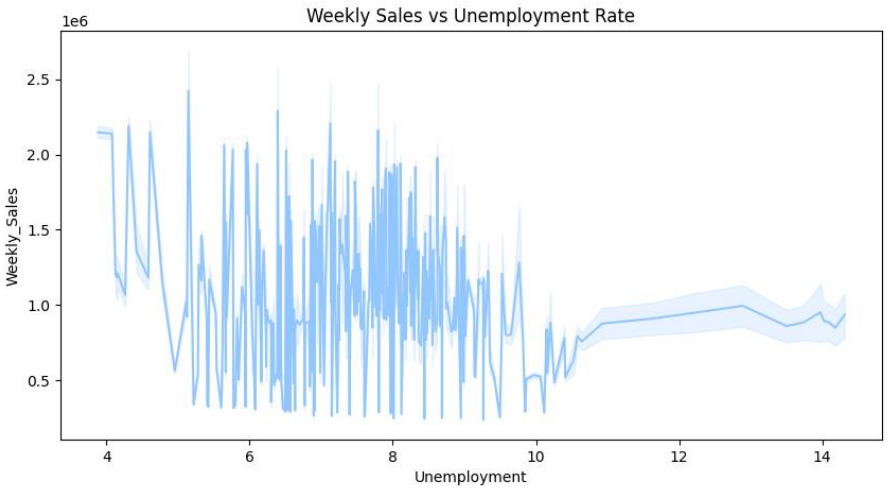
The Data Pre-processing of the data includes the following steps:

- Data Cleaning: Extracting and cleaning the data by removing missing values, duplicates, outliers and other inconsistencies.
- Data Exploration: Exploring the data to gain insights about the shape, dimension of dataset and understanding the data. Checking for uniformity in data based on categorical and numerical data types
- Data Visualization: Visualizing the data for better understanding.

IV. Project Objective

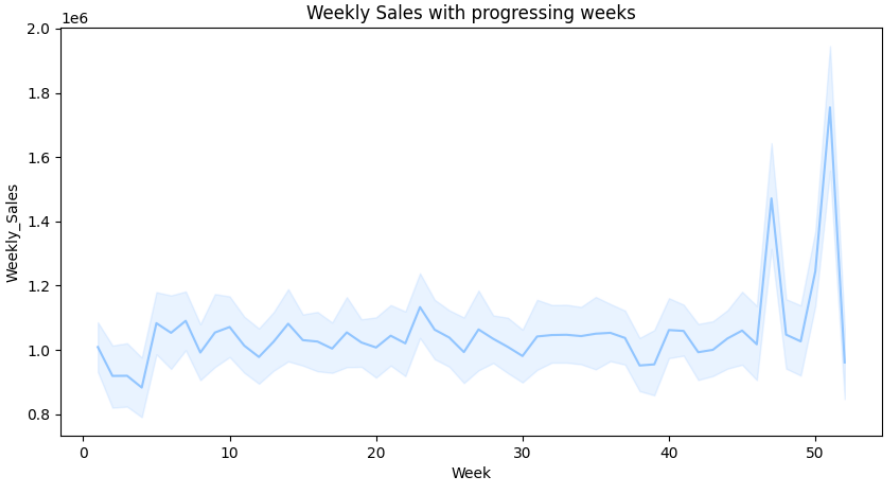
1. You are provided with the weekly sales data for their various outlets. Use statistical analysis, EDA, outlier analysis, and handle the missing values to come up with various insights that can give them a clear perspective on the following:

a. If the weekly sales are affected by the unemployment rate, if yes - which stores are suffering the most?



Weekly Sales are weakly correlated to Unemployment. Moreover from above graph, as the unemployment increases, there is decline in weekly sales as opposed to average unemployment rate (6-10) which contributes to notable weekly sales

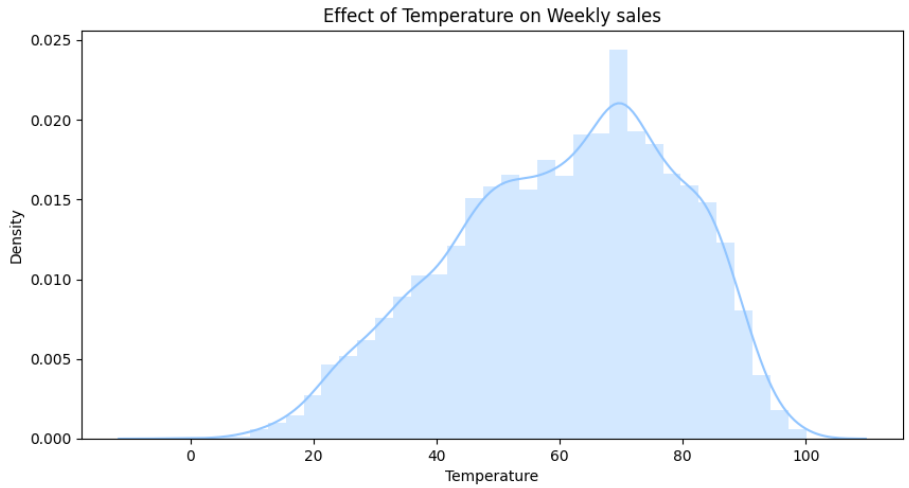
b. If the weekly sales show a seasonal trend, when and what could be the reason?



It can be understood that during seasons there is an expected surge in sales as there would be more requirement of goods at houses in general. There is slight rise in sales during first

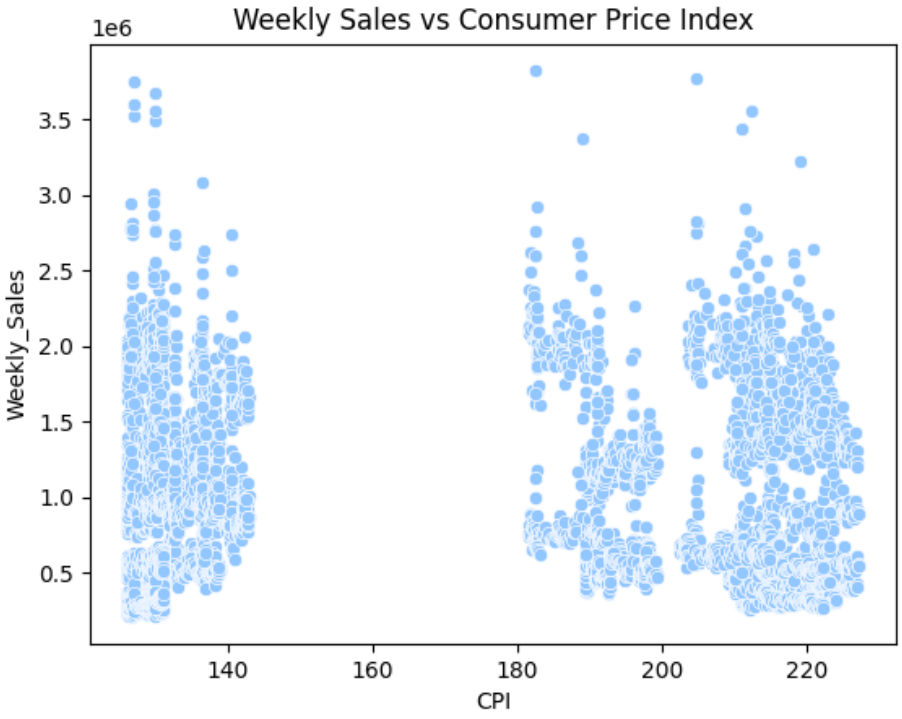
quarter and noticeable increase at year-end during last 6 weeks of year most likely due to Christmas season

c. Does temperature affect the weekly sales in any manner?



Weather has a profound effect on sales. Extreme temperatures can result in drop in number of incoming customers as it is a factor to affect commute to the store.

d. How is the Consumer Price index affecting the weekly sales of various stores?

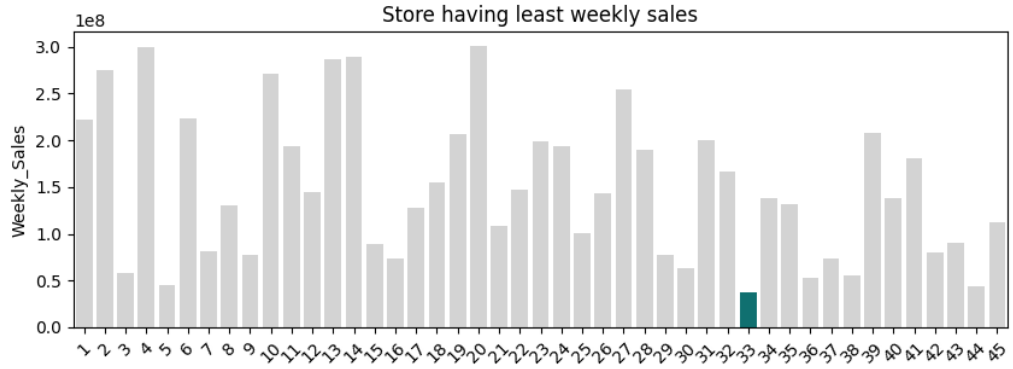


In scatter plot above, there are briefly three different clusters around different ranges of CPI. While there seems no visible relationship between the change in CPI and weekly sales for Walmart stores (sales still occur at high CPI rates), the only negligible observation that can be made is that very high amount of sales for store when CPI is at a low rate of 140. Moreover moderate to average CPI (150 to 180) accounts for very low to no contribution for weekly sales.

e. Top performing stores according to the historical data.

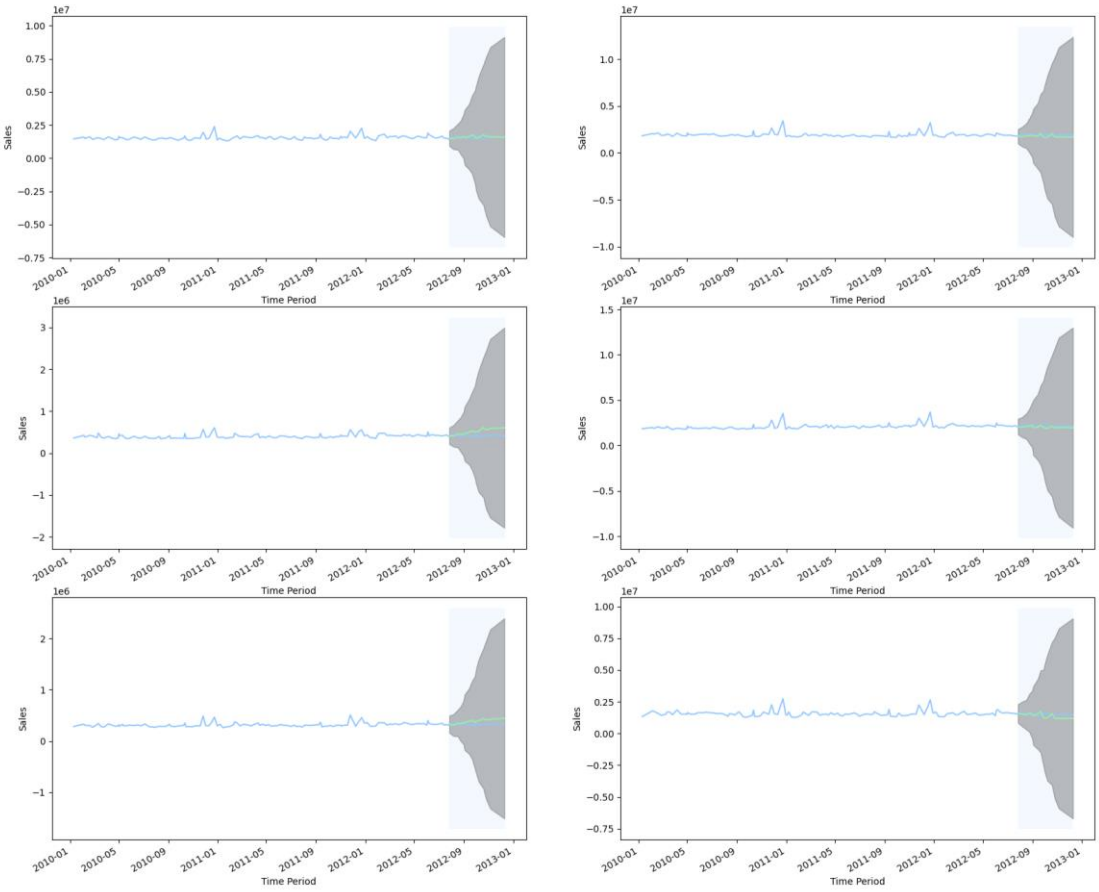
It is visible that the top performing store among all others is store ID 20 having max sales, the sales of this particular store have surpassed sales of all other stores

f. The worst performing store, and how significant is the difference between the highest and lowest performing stores.



The difference between to top and least performing store in terms of weekly sales is: 264237570.49. This is confirmed even from above graph that store 33 generated very less revenue and is worst performing stored. This store has the least amount of sales compared to other stores

2. Use predictive modelling techniques to forecast the sales for each store for the next 12 weeks.



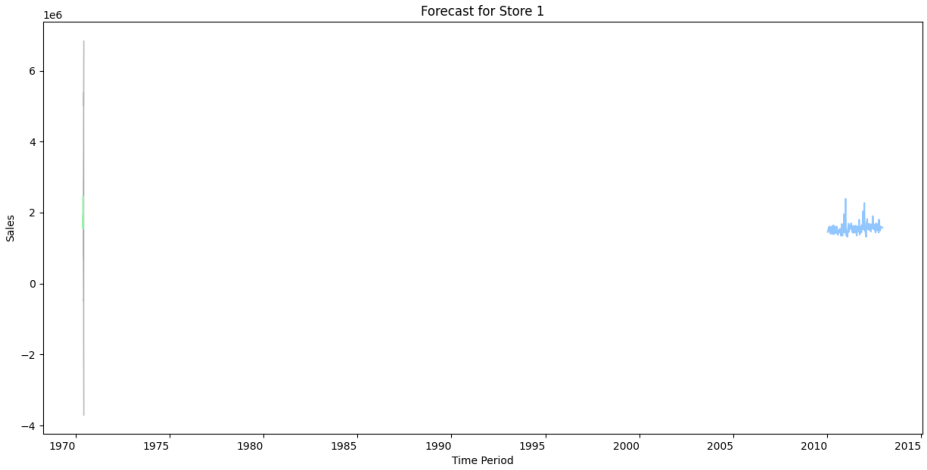
Sales Forecasting using Predictive Modelling for store 1 to store 6 for next 12 weeks is displayed above. It can be observed that forecast for some of the stores [store 1, 2, 4] (lime green coloured line) matches very well with existing sales trend (blue coloured line). This indicates that the pattern of sales is bound to remain pretty much same for upcoming 12 weeks. Whereas for some other stores [store 3, 5, 6], the dynamic forecast rises or drops for upcoming period suggesting that sales might improve or fall respectively

V. Choosing the Algorithm for the Project

The algorithm for predictive modelling and time series forecasting models like ARIMA, SARIMA is chosen depending upon the provided dataset and type of problem to solve. Generally, it depends on number of factors affecting the output variable i.e. either one-univariate or multiple-multivariate. Likewise for a machine learning project supervised and unsupervised learning algorithms are available. Supervised are used for classification and regression problems, while unsupervised are used for clustering and dimensionality reduction tasks. Some of the most popular algorithms used in machine learning include Random Forests, Decision Trees, k-Nearest Neighbours (kNN), etc.

VI. Motivation and Reasons For Choosing the Algorithm

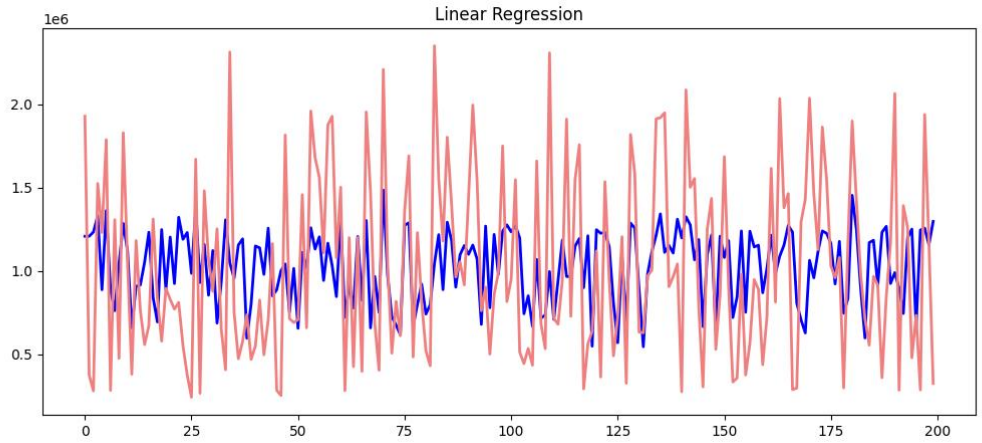
SARIMA time series forecasting algorithm is used here to forecast sales for upcoming 12 weeks. These SARIMA models have some advantages over ARIMA models when the data exhibits strong seasonal patterns. For instance as forecasting of weekly sales data is required, one may encounter higher sales in certain months due to holidays or seasonal demand. SARIMA models can capture this effect correctly and adjust the forecasts accordingly. SARIMA models can also handle multiple seasonal cycles apart from weekly, such as monthly and yearly patterns too. This is useful for data that has complex seasonality and multivariate data, which makes selecting this model an appropriate choice in this scenario.

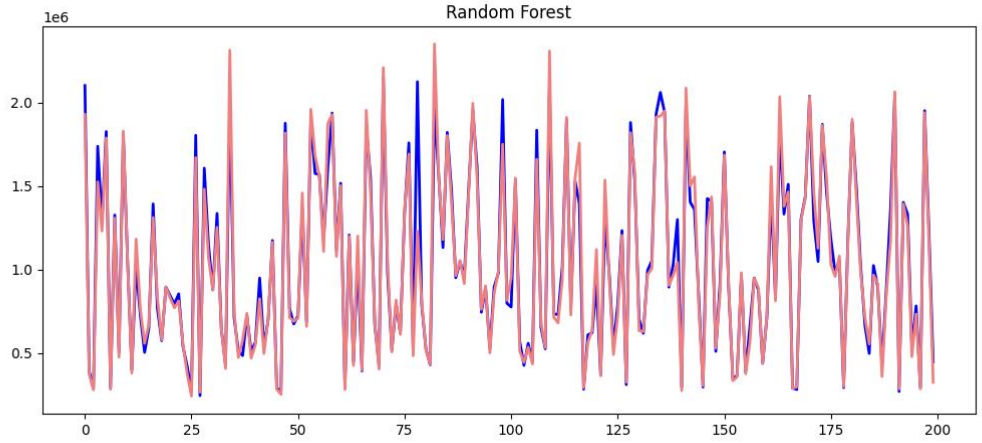


Above plot is a representation of SARIMA model for store 1 showing forecast for next 12 weeks

In terms of machine learning, algorithms Linear Regression & Random Forest are applied here. As observed, Linear Regression model is not quite fitting for this dataset but Random Forest is working well for this dataset.

Random forest is an efficient and powerful algorithm that is well-suited for Walmart's needs. It is a powerful algorithm that can handle large datasets and is capable of handling complex non-linear relationships. It is also highly scalable and can be used to make predictions on large datasets. It is also known for its high accuracy. Additionally, it is easy to use and can be implemented quickly.





VII. Assumptions

It is safe to assume that constraints like Fuel Price, Unemployment rate will not be a major influence on weekly sales even before visualizing heat map with the help of correlation matrix. Although there will be difference due to change in these parameters but can be considered insignificant compared to other important factors. This is because every person does not necessary travel with a vehicle to the store and even can commute by public transport or can walk to the store. Hence increase in Fuel price will not majorly impact nature of sales. Likewise even if unemployment rises, it can be assumed that the sales will not drop drastically and instead might rise as one cannot compromise on basic needs that can be obtained from stores. As store supplies are a required necessity and not an optional luxury

Remaining primary factors determining movement of sales are Temperature, Holiday flag CPI. From these, at least certain conclusions can be directly drawn from Temperature and Holiday Flag that temperature will be inversely related to sales and sales will be more on holidays. Assumption can be made regarding CPI that as CPI rises, the sales may rise with rising purchasing power.

VIII. Model Evaluation and Techniques

Before actual modelling, firstly data needs to be identified whether it is stationary or non-stationary. This can be verified using Ad Fuller test. Seasonal decomposition graphs can be plotted to understand Seasonality, Trend, and Residuals

Various metrics like Mean Square Error, Root Mean Square Error, and Residual can be used for model evaluation. These values are calculated mathematically. Ultimately predicting and forecasting techniques can be applied to the data to obtain desired results and can be visualized to get a better comprehension

IX. Inferences from the Same

Inferences can be derived from studying data and completing the objectives. By leveraging historical sales data, Walmart can use predictive analytics to identify patterns and trends in sales and use them to make accurate predictions and forecast about future sales for each store. Walmart can also use machine learning to identify factors that influence sales, such as weather, seasonality, and customer demographics. By incorporating these factors into their forecasting models, Walmart can make more accurate predictions about future sales. This can help them to correct their mistakes, analyse their shortfalls and implement necessary changes to maximize sales and improve customer retention

X. Future Possibilities of the Project

This project could be developed by implementing another machine learning algorithm and could be built with another time series forecasting model just to check if more desirable outcomes are achieved with advantage of less computational expense.

The sales for each store can be forecasted using machine learning algorithms such as regression, decision trees, and neural networks. These algorithms can be used to predict the sales for each store based on historical data, such as sales figures from previous years, customer demographics, and other factors. The predictions can then be used to inform decisions about inventory management, pricing, and marketing strategies. Additionally, the predictions can be used to identify trends and opportunities for growth.

Additionally these concepts can be modified and used for other related fields in the market to solve similar real-world problems and help in prediction of trends based on historical data for financial instruments like stocks or even for superstores by combining with market basket analysis

XI. Conclusion

Thus, this capstone project proved to be a great source which gives idea of forecasting, prediction based on existing data and available historical records. Certain tips which can be suggested to Walmart to enhance the sale performance if implemented are:

- Offer occasional promotional coupons, seasonal discount to encourage regular as well as new customers
- Appoint a customer relationship manager to maintain loyal customer base
- Follow-up with old customers which have reduced frequency of visit to store and find out reason for them doing so
- Don't stock up perishable items before extreme weathers as sales during those period decline/drop
- Collect feedback through modes of survey to gain insight on needs of customers
- Offer option of pick-up and home delivery at subsidized cost during extreme temperatures
- Use reasonable pricing for frequently bought items
- Organize sales aligned with holidays and weekends to attract public
- Use online marketing to invite potential buyers
- Focus on shortcomings of certain stores and develop strategy to revive sales.

XII. References

1. Provided documents on LMS and reference attachment file ‘walmart(letsee) (1).ipynb’
2. Statistical analysis for superstore sales – Kaggle