Capstone Project - Project Walmart

Table of Contents

1.Problem Statement

A Retail store that has multiple outlets across the country are facing issues in managing the inventory - to match the demand with respect to supply. You are a data scientist, who has to come up with useful insights using the data and make prediction models to forecast the sales for X number of months/years

2.Project Objective

1.To know the certain events and holidays which impact sales on each day.

2.The business facing a challenge due to unforeseen demands and runs out of stock sometimes, due to inappropriate machine learning algorithm.

3.To predict the sales and demand accurately with the help of an ideal ML algorithm to predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index.

4. The objective is to determine the factors affecting the sales and to analyze the impact of markdowns around holidays on the sales.

3.Data Description

This is the historical data that covers sales from 2010-02-05 to 2012-11-01, in which 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

.

4.Data Preprocessing Steps And Inspiration

1. Viewing Data Information

```
# Convert date to datetime format
walmart_data['Date'] =  pd.to_datetime(walmart_data['Date'])
walmart_data.info()
```

2. Checking Null values

```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Store         6435 non-null   int64
 1   Date          6435 non-null   datetime64[ns]
 2   Weekly_Sales  6435 non-null   float64
 3   Holiday_Flag  6435 non-null   int64
 4   Temperature   6435 non-null   float64
 5   Fuel_Price    6435 non-null   float64
 6   CPI           6435 non-null   float64
 7   Unemployment  6435 non-null   float64
```

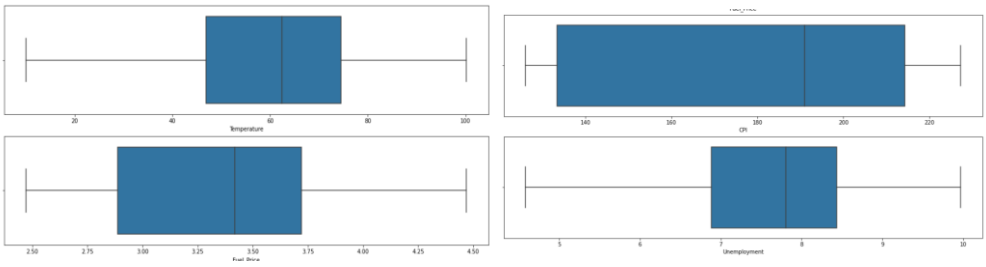
```
walmart_data.isnull().sum()
```

```
Store           0
Date            0
Weekly_Sales    0
Holiday_Flag    0
Temperature     0
Fuel_Price      0
CPI             0
Unemployment    0
dtype: int64
```

3. Converting non datetime to datetime

```
walmart_data["Day"]= pd.DatetimeIndex(walmart_data['Date']).day
walmart_data['Month'] = pd.DatetimeIndex(walmart_data['Date']).month
walmart_data['Year'] = pd.DatetimeIndex(walmart_data['Date']).year
walmart_data
```

4. Checking for outliers



5.Choosing the Algorithm for the Project

Random Forest regression can be a suitable algorithm for analyzing retail sales data. Random Forest is an ensemble learning method that combines multiple decision trees to make predictions.

I have chosen the Random Forest Regression algorithm for this project for the following reasons:

1. Non-linearity: Retail sales data often exhibits complex and non-linear relationships between the input features and the target variable (e.g., sales). Random Forest regression can capture non-linear patterns effectively by constructing an ensemble of decision trees.
2. Robustness to outliers and noise: Retail sales data may contain outliers or noisy observations due to various factors. Random Forest is robust to outliers and noise as it considers multiple trees and averages their predictions, reducing the impact of individual outliers or noisy data points.
3. Feature importance: Random Forest provides a measure of feature importance, indicating which features have the most significant influence on sales predictions. This information can be valuable in understanding the key factors driving sales and making data-driven decisions regarding product attributes, pricing, promotions, and more.
4. Handling categorical variables: Retail sales data often includes categorical variables such as product categories, store locations, or customer segments. Random Forest can handle categorical variables naturally without the need for explicit feature engineering, as decision trees can split on categorical variables directly.
5. Scalability: Random Forest can handle large-scale datasets efficiently. It can handle a high number of input features and handle a wide range of data sizes, making it suitable for analyzing large retail sales datasets.
6. Generalization: Random Forest is less prone to overfitting compared to individual decision trees, as it incorporates an ensemble of trees with reduced variance. This enables better generalization to new, unseen data, which is crucial for making accurate predictions on future sales data.

6.Assumptions

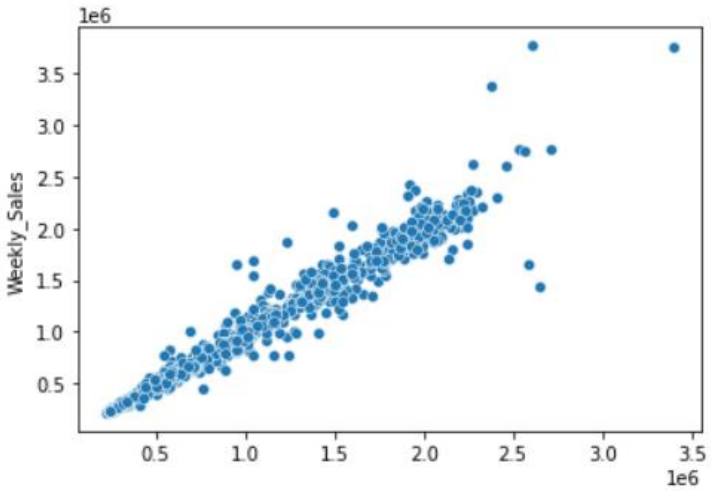
The following assumptions were made in order to create the model for Walmart project.

NO ASSUMPTIONS WERE MADE

7.Model Evaluation and Technique

The following techniques and steps were involved in the evaluation of the model

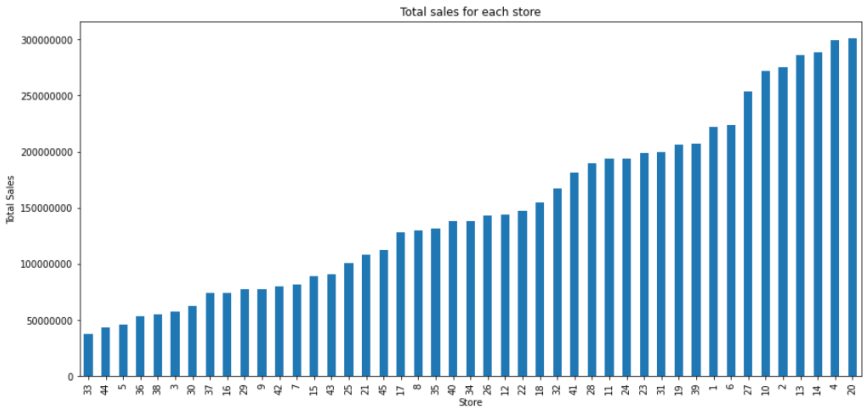
1. Accuracy: **95.91%**
2. Mean Absolute Error: **61601.64**
3. Mean Squared Error: **13632177059.04**
4. Root Mean Squared Error: **116756.91**



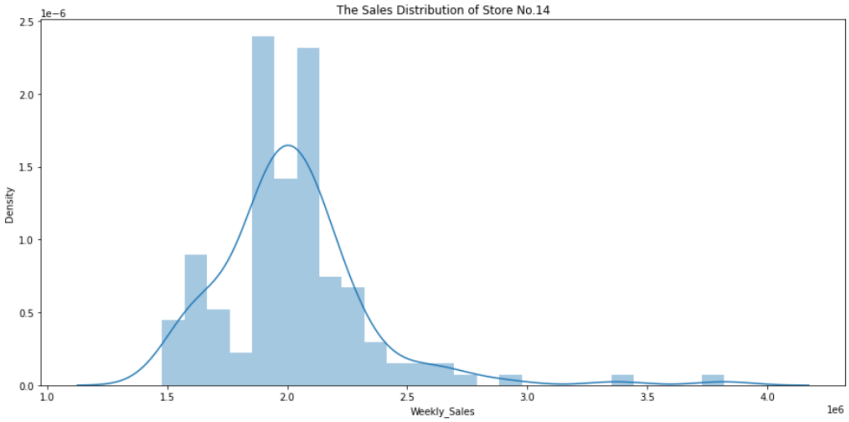
Predicted v/s Actual sales Values

8.Inferences from the Project

Clearly, from the below graph, it is visible that the store which has maximum sales is store number 20 and the store which has minimum sales is the store number 33.

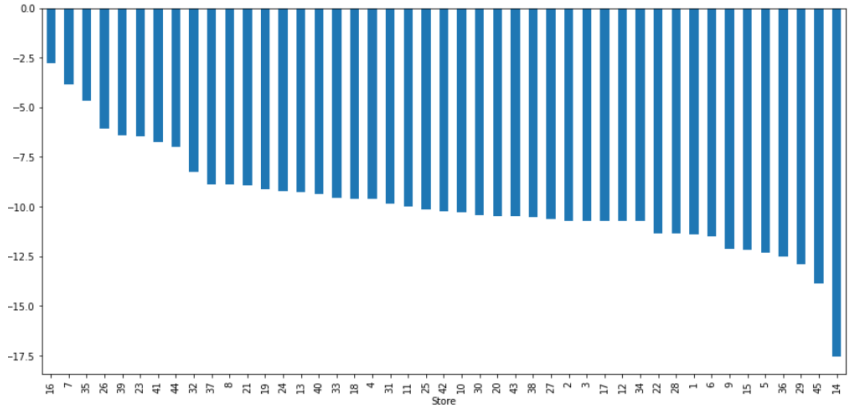


Total Sales for each store

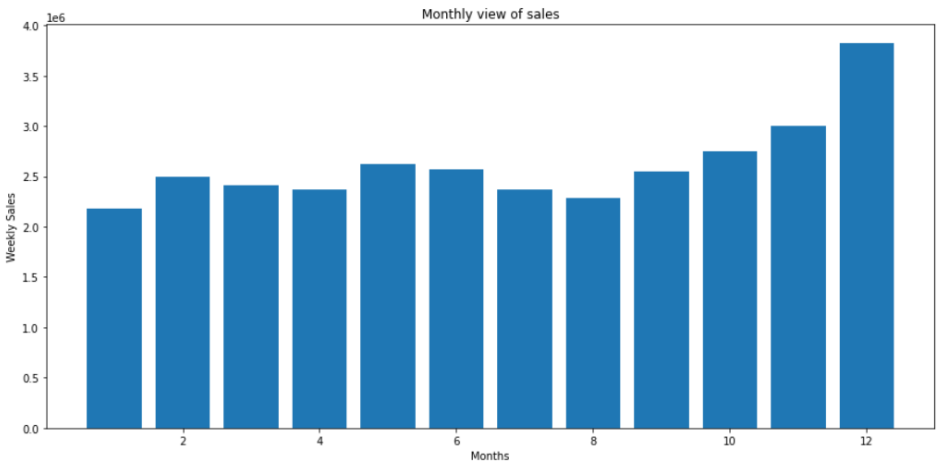


Distribution of weekly Sales for Store no,14

Here, there is no store which has performed better in the 3rd quarter as compared to the 2nd quarter.

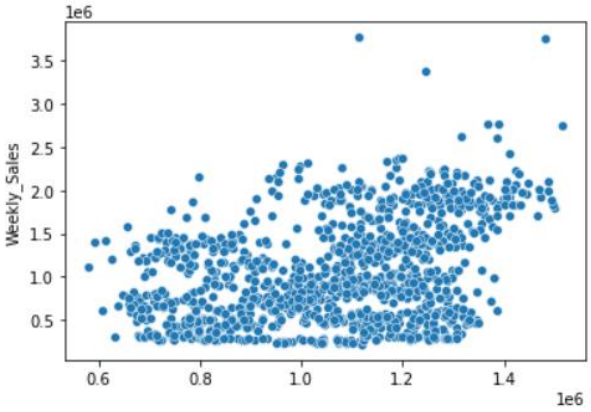


Performance in 3rd Quarter compared to 2nd Quarter



Monthly Views of Sales

Results from Linear Regression



Predicted v/s Actual Sales values

Accuracy: 12.73%

Mean Absolute Error: 447011.09

Mean Squared Error: 288871486677.95

Root Mean Squared Error: 537467.66

As we can see from the above chart and its results, Linear Regression is not the right Algorithm for our Analysis. One of the reasons could be our data follows a nonlinear relationship, which linear regression is not able to account for. Hence, I've chosen Random Forest regression, which accounts for non linear relationship in our data

9.Future Possibilities

The development of a good model that accurately predicts sales can have a significant impact on a retail business. Achieving a high level of accuracy in sales predictions enables several key benefits and opportunities:

1.Optimal Pricing and Promotions

2.Inventory Management

3.Resource Allocation

4.Marketing and Campaign Effectiveness

10.Conclusion

1.The store which has Maximum sales is store number 20 and the store which has minimum sales is store number 33.

2. The store which has maximum standard deviation is store number 14. i.e., the sales vary a lot

3. The Store which has good quarterly sales in Q3’2012 is store no. 4.

4. There is no store which has performed better in 3rd quarter as compared to the 2nd quarter.

5. Thanksgiving has higher sales than the mean sales on non-Holidays.

6.Overall monthly Sales are higher in the month of December while the yearly sales in the year 2011 are the highest.

7.Linear Regression is not an appropriate model to use which is clear from it's low accuracy. However, Random Forest Regression gives accuracy of over 95% , so, it is the best model to forecast demand

As the model is able to accurately predicts sales, it empowers retailers to make data-driven decisions, optimize business operations, and drive revenue growth. By leveraging insights derived from accurate sales predictions, retailers can enhance customer satisfaction, reduce costs, optimize inventory levels, and improve overall business performance in a highly competitive retail landscape.