Capstone Project – Walmart Sales Forecasting

T SAI KISHORE

Table of Contents	Page Number

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.

Dataset Information: The walmart.csv contains 6435 rows and 8 columns

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

2.Project Objective:

1. 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?

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

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

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

e. Top performing stores according to the historical data.

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

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

3.Data Description

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

The dataset contains features as mentioned in the above table as feature names. All the features have numeric data.

4. Data Pre-processing Steps and Inspiration

4.1 Data Preprocessing Steps

Data preprocessing is the process of generating raw data for machine learning models. This is the first step in creating a machine-learning model.

Steps in data preprocessing

4.1.1 Check for Missing Values

Assess the loaded data and check for missing values. If missing values have been found, there are particularly two ways to resolve this issue:

- Either remove the entire row that contains a missing value. However, removing the entire row can generate a possibility of losing some important data. This approach is useful if the dataset is very large
- Or estimate the value by taking the mean, median or mode.

4.1.2 Outlier Treatment

Depending on the specific characteristics of the data, there are several ways to handle outliers in a dataset. Some of the most common approaches to handle outliers are:

a. Remove outliers*:*

In some cases, it may be appropriate to simply remove the observations that contain outliers. This can be particularly useful if you have a large number of observations and the outliers are not true representatives of the underlying population.

b. Transform outliers:

The impact of outliers can be reduced or eliminated by transforming the feature. For example, a log transformation of a feature can reduce the skewness in the data, reducing the impact of outliers.

c. Impute outliers:

In this case, outliers are simply considered as missing values. You can employ various imputation techniques for missing values, such as mean, median, mode, nearest neighbour, etc., to impute the values for outliers.

d. Use robust statistical methods:

Some of the statistical methods are less sensitive to outliers and can provide more reliable results when outliers are present in the data. For example, we can use median and IQR for the statistical analysis as they are not affected by the outlier's presence. This way we can minimize the impact of outliers in statistical analysis.

4.1.3 Arrange the Data

Machine learning modules cannot understand non-numeric data. It is important to arrange the data in a numerical form in order to prevent any problems at later stages.

4.1.4 Do Scaling

Scaling is a technique that can convert data values into shorter ranges. Rescaling and Standardization can be used for scaling the data.

4.1.5 Distribute Data into Training, Evaluation and Validation Sets

The final step is to distribute data in three different sets, namely

- Training
- Validation
- Evaluation

4.2 Walmart Dataset Preprocessing Steps

shape( ) command

shape command will help us to find out how many rows and columns are there in the dataset. Walmart dataset contains 6435 rows and 8 columns.

Isnull( ) command

isnull command will help us to find if there are any null values. There are no null values in the dataset.

Info( ) Command

The dataset information is obtained using info() command which will help us print a concise summary of a DataFrame. This method prints information about a DataFrame including the index dtype and columns, non-null values and memory usage.

```
#dataset information
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6435 entries, 0 to 6434
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   Store         6435 non-null   int64  
 1   Date          6435 non-null   object 
 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
dtypes: float64(5), int64(2), object(1)
memory usage: 402.3+ KB
```

describe( ) Command

The describe command will help us to generate descriptive statistics. Descriptive statistics include those that summarize the central tendency, dispersion and shape of a dataset's distribution, excluding ``NaN`` values. Analyzes both numeric and object series, as well as ``DataFrame`` column sets of mixed data types.

duplicated( ) command

To check if there are any duplicate rows we use duplicated() command. There are no duplicate rows in the dataset.

4.3 Outliers

Outliers, or data points that are significantly different from the rest of the data, can affect the accuracy and reliability of statistical measures such as the mean and standard deviation. To ensure that these measures accurately represent the data, it is necessary to identify and properly handle outliers. To address this issue, we will develop two functions: one to detect outliers and another to count them. These functions will help us identify and understand the impact of outliers on our data, and allow us to make informed decisions about how to handle them.

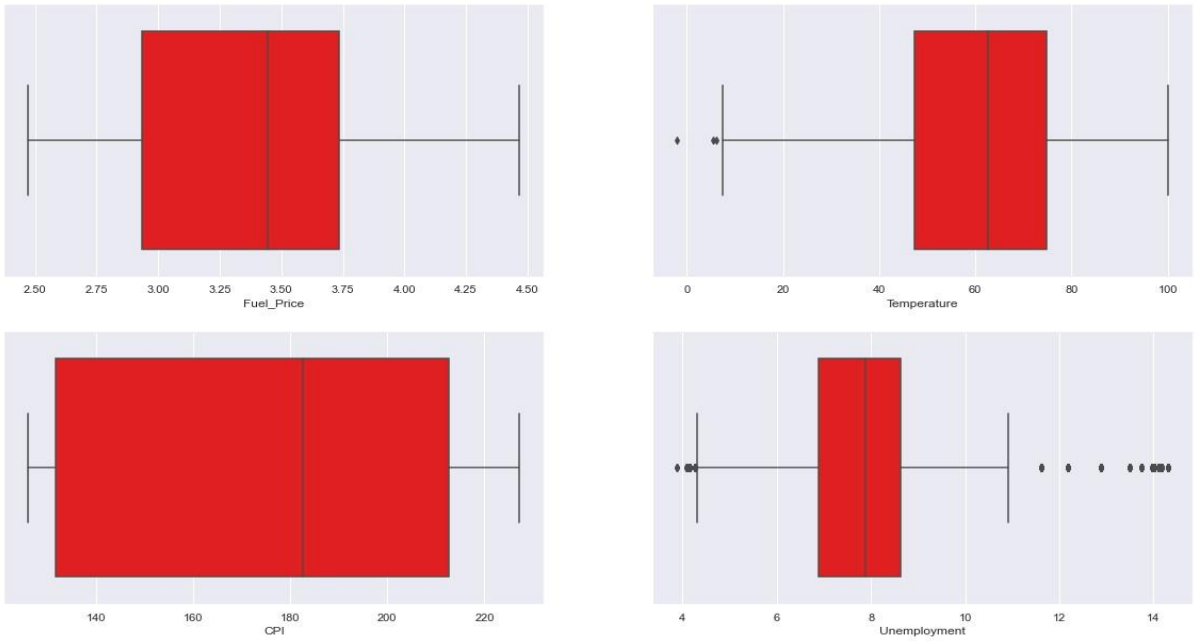

Fig 4.1 Outliers in the dataset

4.3.1 Outliers Treatment:

The interquartile range (IQR) is a measure of the spread of the middle 50% of the data. The IQR can be calculated as the difference between the 75th percentile and the 25th percentile of the dataset. Any data point outside the range of 1.5 times the IQR below the 25th percentile or above the 75th percentile can be considered an outlier.

To identify outliers using the IQR, we can use the quantile() function in pandas to calculate the 25th and 75th percentiles of the dataset. We can then calculate the IQR and use it to identify outliers.

Once we have identified the outliers in our dataset, we can either exclude them from our analysis or replace them with more accurate values.

4.4 Distribution of the data

Fig 4.2 Data Distribution

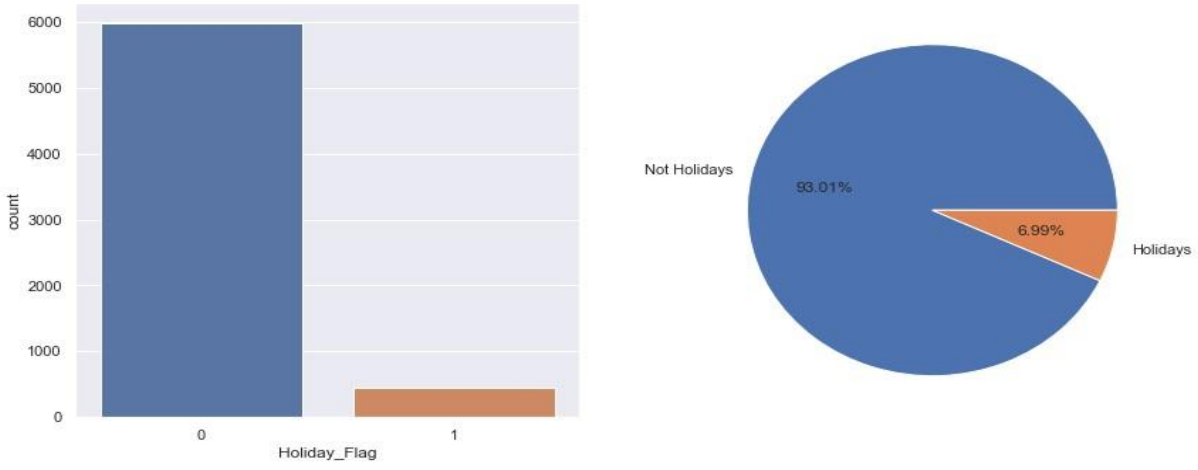

Fig 4.3 categorical Data Distribution

Days of no holiday are the most frequent than days of holiday in the dataset with a percentage of 93 % .

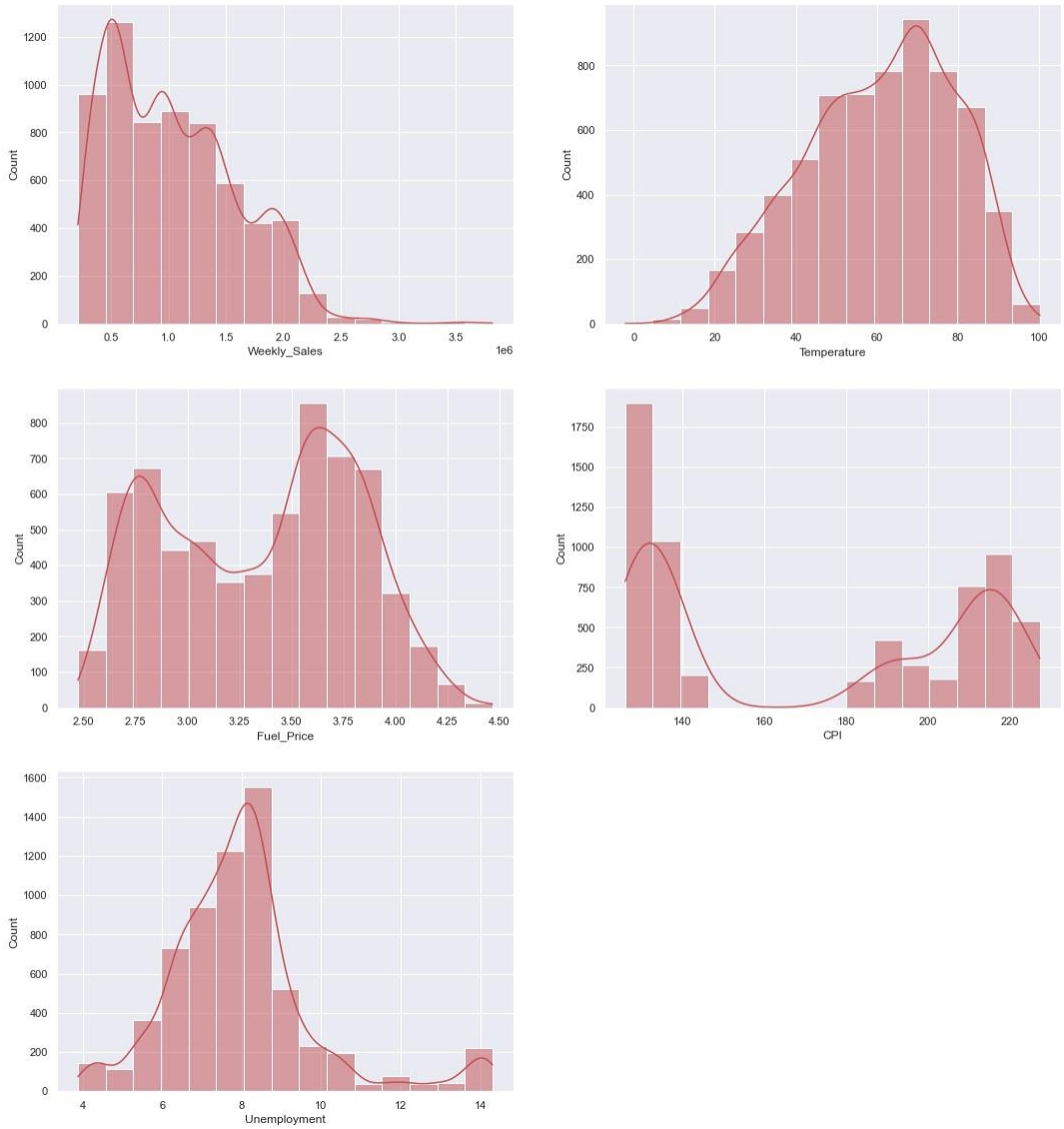

Fig 4.4 Numerical Data Distribution

The data distribution helps us to understand various statistical parameters.

From the above histograms, we can understand that:

- the number of transactions occurred almost evenly across various stores and years.
- The distribution of weekly_sales right-skewed. Only a few of the weekly sales are above 2 million USD.
- The distribution of temperature is approximately normal.
- The distribution of fuel_price is bi-modal.
- CPI formed two clusters.
- unemployment rate is near normally distributed.
- Four consecutive months November-February recorded the highest sales.

4.5 Correlation among the features

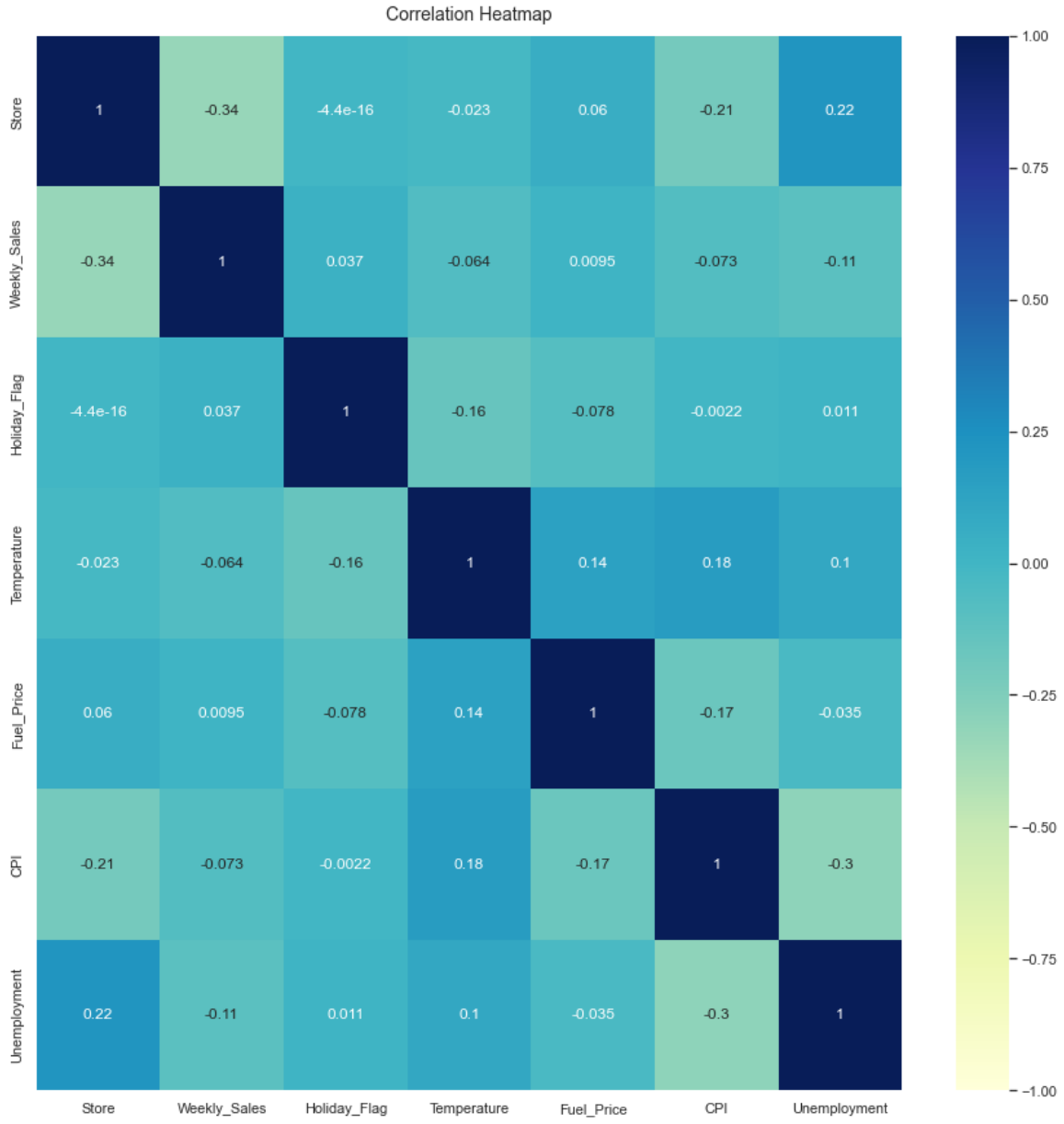

Fig 4.5 Correlation Heatmap

Correlation matrix gives us correlation of each variable with each of other variables present in the dataframe. To calculate correlation, we first calculate the covariance between two variables and then covariance is divided by the product of standard deviation of same two variables. Correlation has no units so it is easy to compare correlation coeffient.

In pandas, we don't need to calculate co-variance and standard deviations separately. It has corr() method which can calculate the correlation matrix for us.

4.6 Analysis on the basis of date

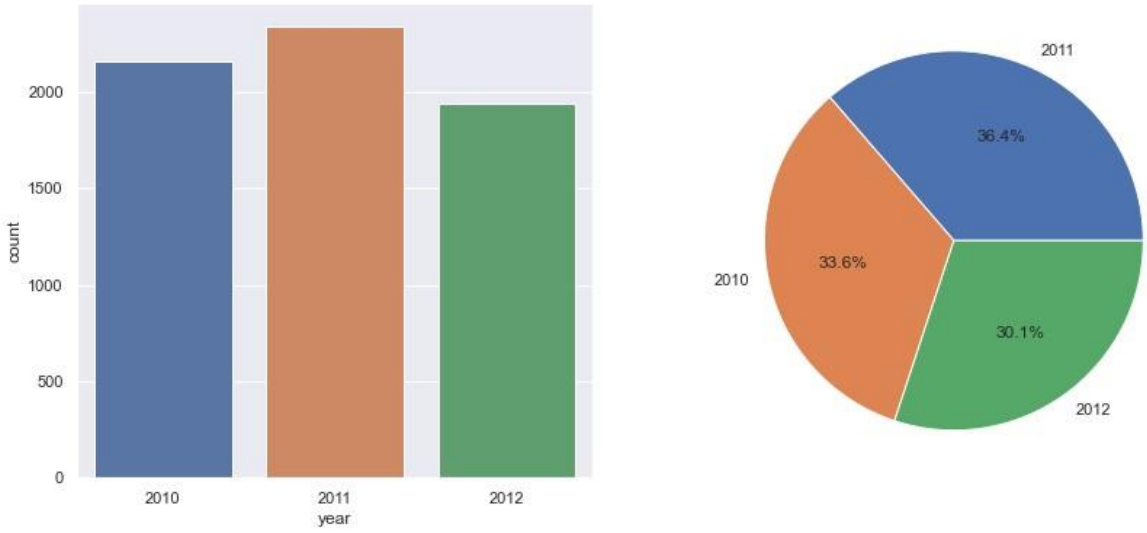

Fig 4.4.1 Year Wise Analysis

2011 is the most frequent in the dataset because most of the weekly sales were recorded during this year.

Fig 4.4.2 Monthly Wise Analysis

April and July are the most frequent in the dataset because most of the weekly sales were recorded in these months.

5. Choosing the Algorithm for the Project

The target values of the dataset are continuous value and is a regression problem. The following algorithms are implemented:

1. Linear Regression
2. KNN Regressor
3. Decision Tree Regressor
4. Random Forest Regressor
5. Time Series Forecasting using SARIMAX

Regression is the technique which will be used to train the model since the target variable is continuous. After the model is trained, we need to predict how good the model has performed. We evaluate this using metrics like RMSE, MAE etc.

Some of the most commonly used metrics are:

Model Evaluation and Technique

To evaluate a regression problem, we need to analyse and evaluate the error difference between the actual value and the predicted value. Some of the metrics used to evaluate the model performance is:

R2 Score

The R2 score (pronounced R-Squared Score) is a statistical measure that tells us how well our model is making all its predictions on a scale of zero to one. we can use the R2 score to determine the accuracy of our model in terms of distance or residual.

Mean Absolute Error (MAE)

The MAE is simply defined as the sum of all the distances/residuals (the difference between the actual and predicted value) divided by the total number of points in the dataset. It is the absolute average distance of our model prediction.

If you want to know the model's average absolute distance when making a prediction, you can use MAE. In other words, you want to know how close the predictions are to the actual model on average.

Root Mean Squared Error (RMSE)

Another commonly used metric is the root mean squared error, which is the square root of the average squared distance (difference between actual and predicted value). RMSE is defined as the square root of all the squares of the distance divided by the total number of points.

5.1 Linear Regression

The reason Linear Regression algorithm was chosen because it is a regression problem. We will have to predict the weekly sales based on certain parameters.

Linear regression algorithm shows a linear relationship between a dependent (y variable) and one or more independent (x variable) variables hence it is known as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

There are four assumptions associated with a linear regression model:

1. **Linearity**: The relationship between X and the mean of Y is linear.
2. **Homoscedasticity**: The variance of residual is the same for any value of X.
3. **Independence**: Observations are independent of each other.
4. **Normality**: For any fixed value of X, Y is normally distributed.

The model was trained and evaluated.

- Root Mean Squared Error: 525536.16
- R-Square score Training: 14.99 %

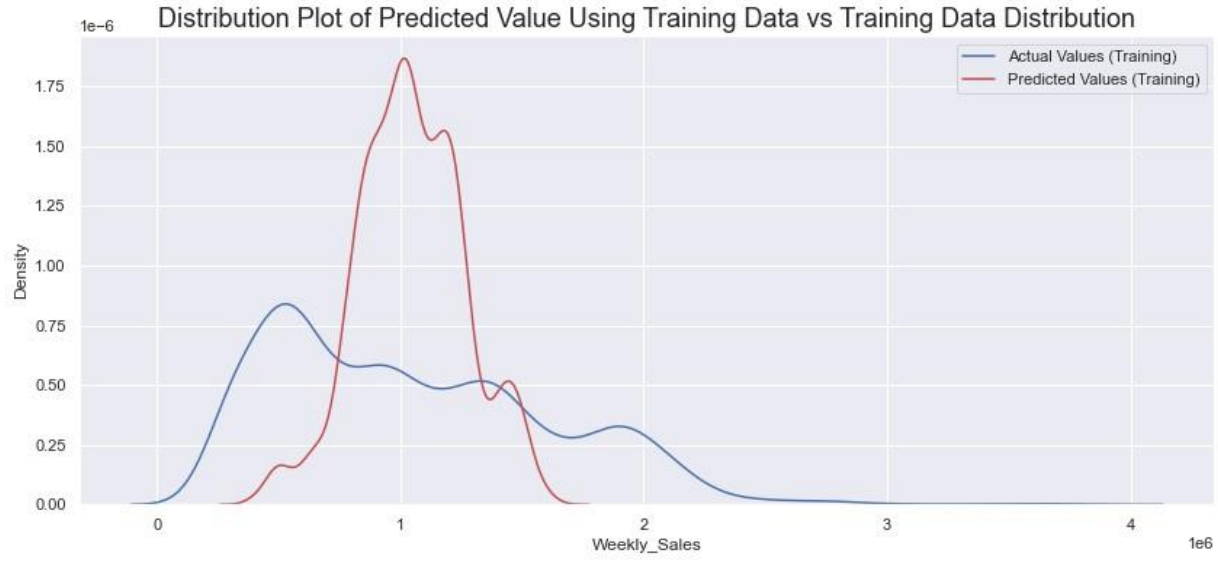

Fig 5.1 Linear Regression Distribution Plot

Since there are no linear correlations between variables and targets, The model seems to be not doing well in learning from the training dataset, so we need to increase the complexity of this model. let's do Polynomial Features for the data before modelling.

Since the model has not performed well, we will try implementing Polynomial regression. Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.

Training accuracy is :

- Root Mean Squared Error: 67608.93
- R-Square score Training: 98.59 %

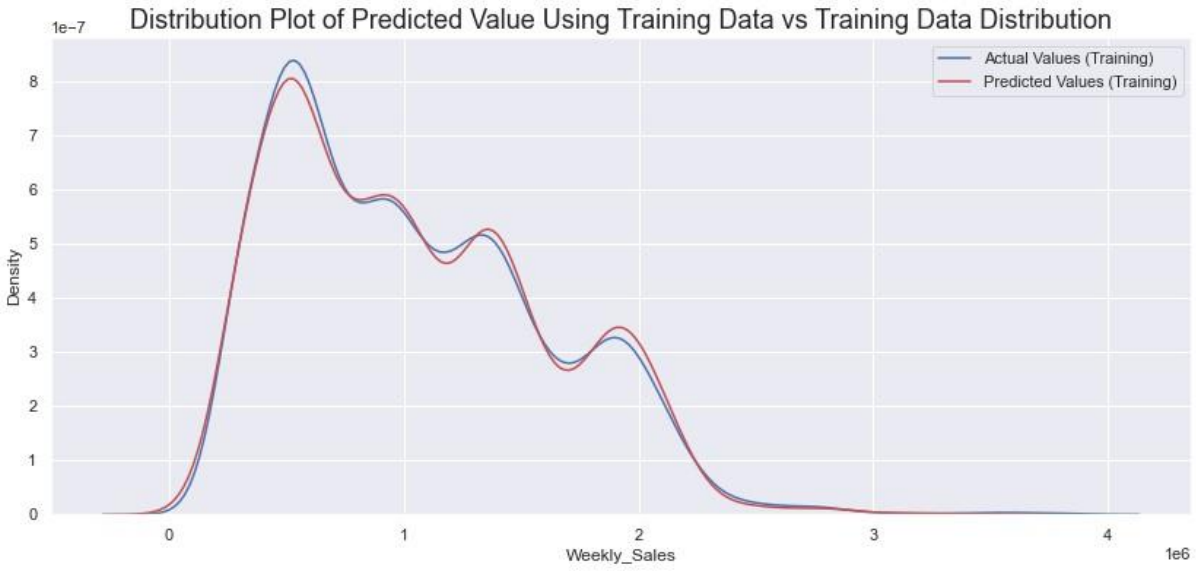

Fig 5.2 Polynomial Distribution Actual vs Predicted

After applying Polynomial Features to the training data, the model seems to have improved more in learning from the training dataset.

Cross Validation of Linear Regression

Cross-validation (CV) is an essentially simple and intuitively reasonable approach to estimating the predictive accuracy of regression models. CV is developed in many standard sources on regression modelling and "machine learning"

Upon implementing cross validation the accuracy is

- Root Mean Squared Error: 115051.15
- R-Square score Training: 95.95 %

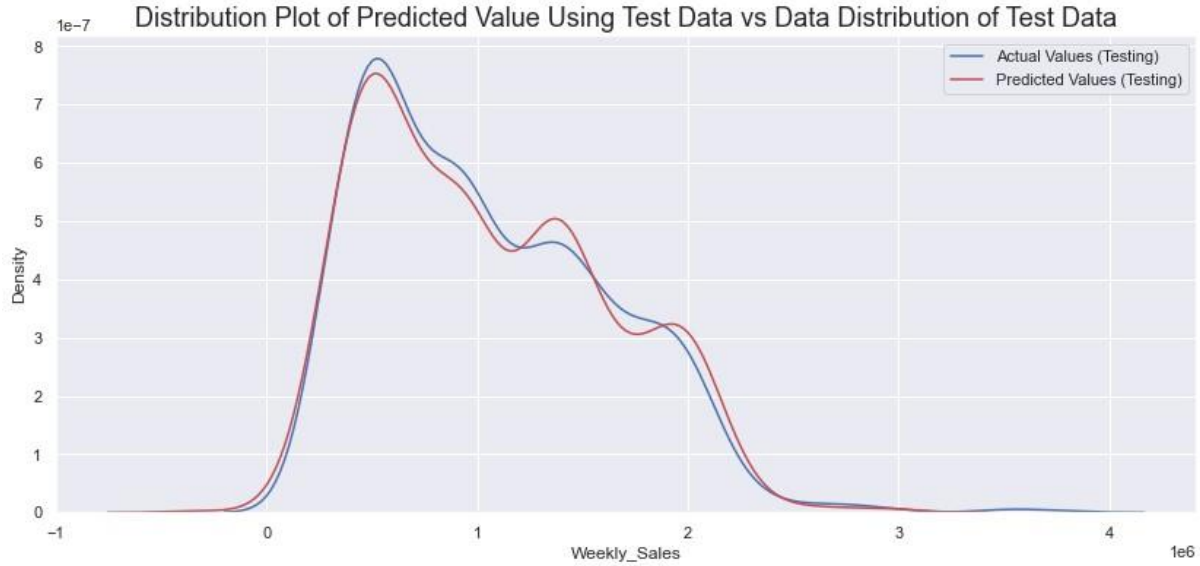

Fig 5.3 Cross Validation Actual vs Predicted

After testing the polynomial regression model, it became clear that the model had learned sufficiently, as its accuracy was 95.95 %.

5.2 KNN Regressor

K-Nearest Neighbours (KNN) is a non-parametric machine learning algorithm that can be used for both classification and regression tasks. In the context of regression, KNN is often referred to as "K-Nearest Neighbours Regression" or "KNN Regression." It's a simple and intuitive algorithm that makes predictions by finding the K nearest data points to a given input and averaging their target values (for numerical regression) or selecting the majority class (for classification).

The primary assumption that a KNN model makes is that data points/instances which exist in close proximity to each other are highly similar, while if a data point is far away from another group it's dissimilar to those data points.

After the model the trained and tested, the metrics values are:

- Root Mean Squared Error: 355737.84
- R-Square score Training: 61.05 %

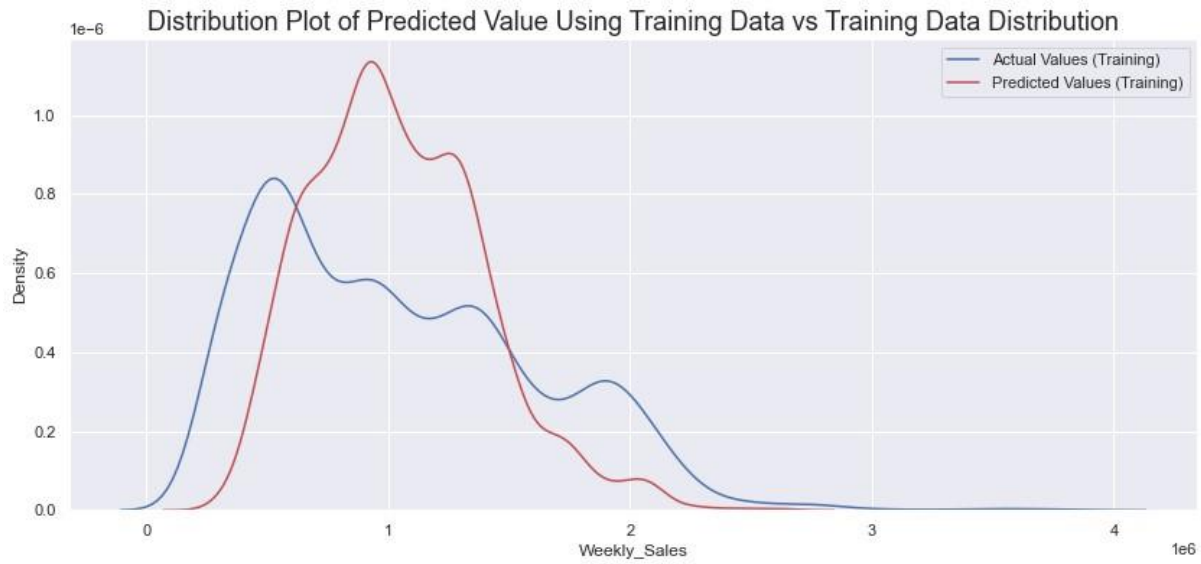

Fig 5.4 KNN Actual vs Predicted

The model seems to be doing not bad in learning from the training dataset.

Upon tuning the parameters and cross validation, the model performance is average.

5.3 Decision Tree Regressor

A decision tree is one of the most frequently used Machine Learning algorithms for solving regression as well as classification problems. As the name suggests, the algorithm uses a tree-like model of decisions to either predict the target value (regression) or predict the target class (classification).

The process of splitting starts at the root node and is followed by a branched tree that finally leads to a leaf node (terminal node) that contains the prediction or the final outcome of the algorithm. Construction of decision trees usually works top-down, by choosing a variable at each step that best splits the set of items. Each sub-tree of the decision tree model can be represented as a binary tree where a decision node splits into two nodes based on the conditions.

Some of the assumptions we make while using Decision tree:

- In the beginning, the whole training set is considered as the root.
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- Records are distributed recursively on the basis of attribute values.
- Order to placing attributes as root or internal node of the tree is done by using some statistical approach.

While training the model, there was overfitting and hence the model had to be tuned using hyperparameters.

Results after hyperparameter tuning:

- Root Mean Squared Error: 142796.43
- R-Squere score Training: 93.72 %

After implementing cross validation, the results are

- Root Mean Squared Error: 160666.31
- R-Squere score Training: 92.11 %

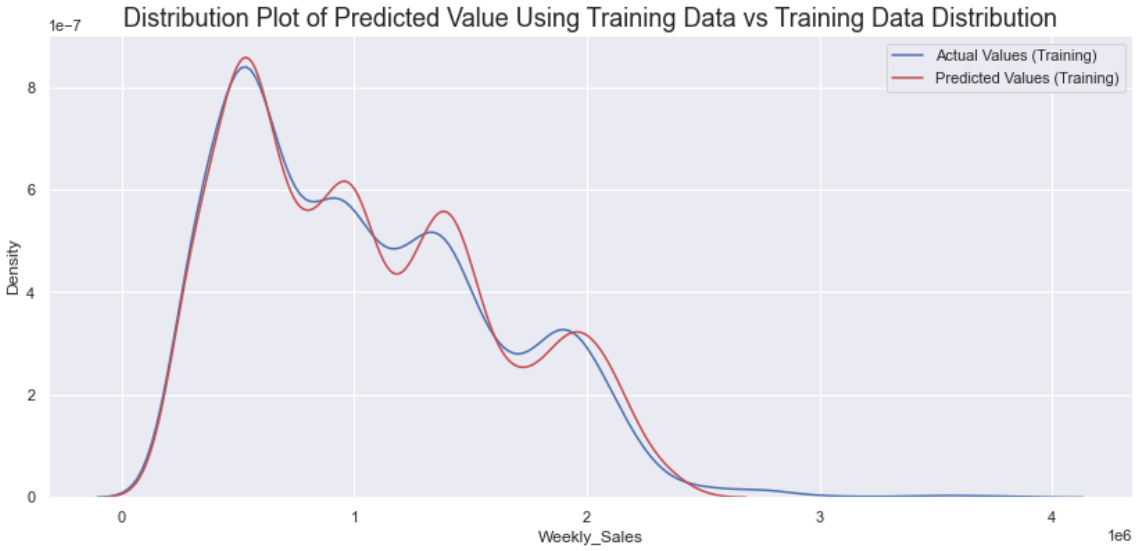

Fig 5.5 Decision Tree Actual vs Predicted after hyperparameter tuning

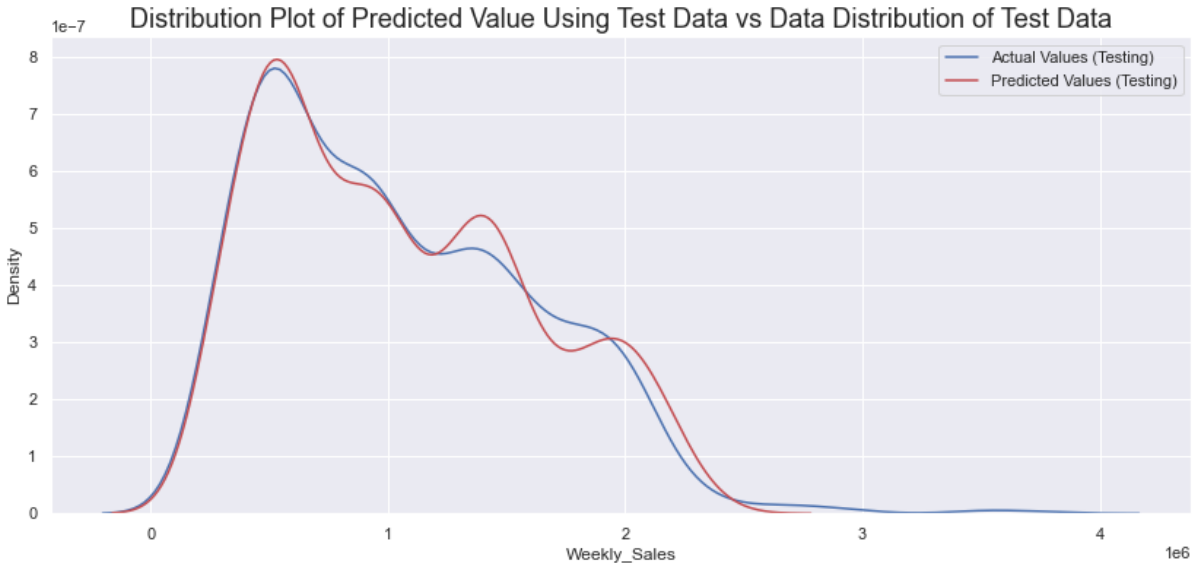

Fig 5.6 Decision Tree Actual vs Predicted after cross validation

5.4 Random Forest Regressor (Bagging)

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees.

After the model is trained and tested, the results are

- Root Mean Squared Error: 55054.2
- R-Square score Training: 99.07 %

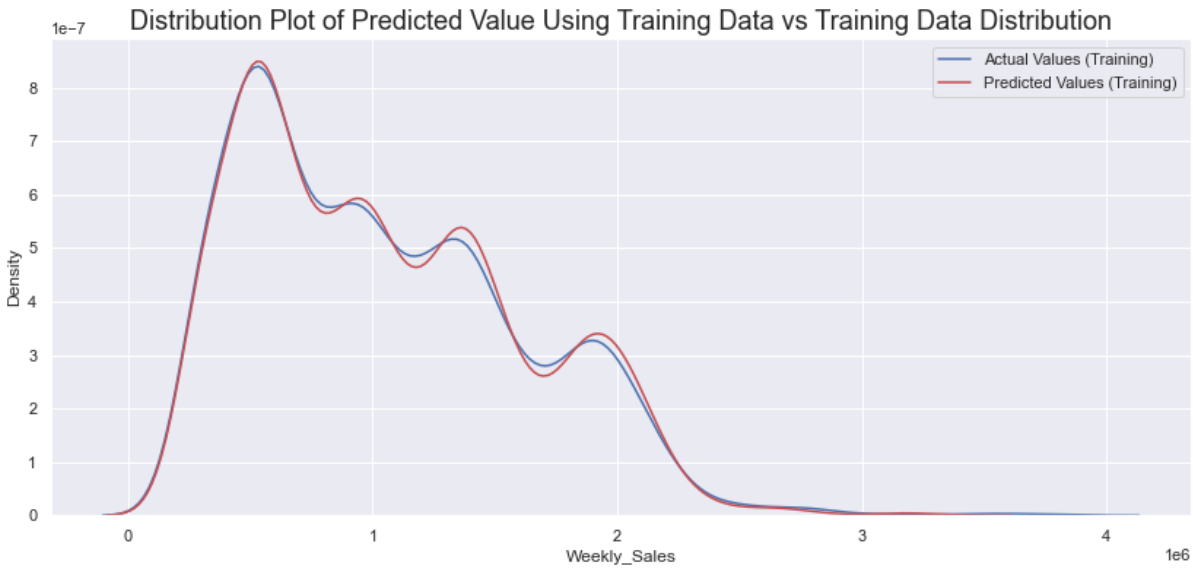

Fig 5.7 Random Forest Actual vs Predicted

To check if the model has overfitted we implement cross validation and the results are

Cross Validation Scores: [0.93690105 0.94290209 0.95282048 0.89407298 0.91357314 0.93925958 0.95934304 0.93600437 0.91999997 0.93803849]

Mean of Scores: 93.33 %

Standard Deviation of Scores: 0.018260718614136556

5.5 Time Series Analysis

The reason for choosing time series analysis is the weekly sales data is time-dependent variables

SARIMA

If there is seasonality a SARIMA (Seasonal ARIMA) model should be used. When applying an ARIMA model, we are ignoring seasonality and using only part of the information in the data. As a consequence, we are not making the best predictions possible.

SARIMA models include extra parameters related to the seasonal part. Indeed, we can see a SARIMA model as two ARIMA models combined: one dealing with non-seasonal part and another dealing with the seasonal part.

Therefore, a SARIMA(p,d,q)(P,D,Q,S) model have all the parameters described above (non-seasonal parameters) and P,D,Q,S that are the seasonal parameters, i.e.,

Non-seasonal orders

p: Autoregressive order

d: Differencing order

q: Moving average order

Seasonal orders

P: Seasonal autoregressive order

D: Seasonal differencing order

Q: Seasonal moving average order

S: Length of the seasonal cycle

In order to predict the future weekly sales value, we use time series.

There are 45 different stores and each store has to be trained and predicted independently.

Automated Model Selection

pmdarima allows us to automate the search of model orders. Automated Model Selection can speed up the process of choosing model orders, but needs to be done with care. Automation can make mistakes since the input data can be imperfect and affect the test scores in non-predictable ways.

The only non-optional parameter in auto_arima is data. However, using our knowledge to specify other parameters can help finding the best model.

Upon implementing auto_arima to find best parameters we found order (2,1,3), and seasonal_order (0, 1, 1, 52) to be best performing.

5.6 Decision Making: Determining a Good Model Fit

Now that we have visualized the different models, and generated the R-squared and MSE values for the fits, how do we determine a good model fit?

What is a good R-squared value?

When comparing models, the model with the higher R-squared value is a better fit for the data.

What is a good MSE?

When comparing models, the model with the smallest MSE value is a better fit for the data.

Comparing these four models, we conclude that The Linear Regression Model is the best model with an accuracy of 95.95 % to be able to predict weekly sales from our dataset.

6. Inferences from the Project

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

Pearson Correlation Coefficient will help us to determine if sales are affected by unemployment rate.

Pearson Correlation

The Pearson Correlation measures the linear dependence between two variables X and Y.

The resulting coefficient is a value between -1 and 1 inclusive, where:

1: Perfect positive linear correlation.

0: No linear correlation, the two variables most likely do not affect each other.

-1: Perfect negative linear correlation.

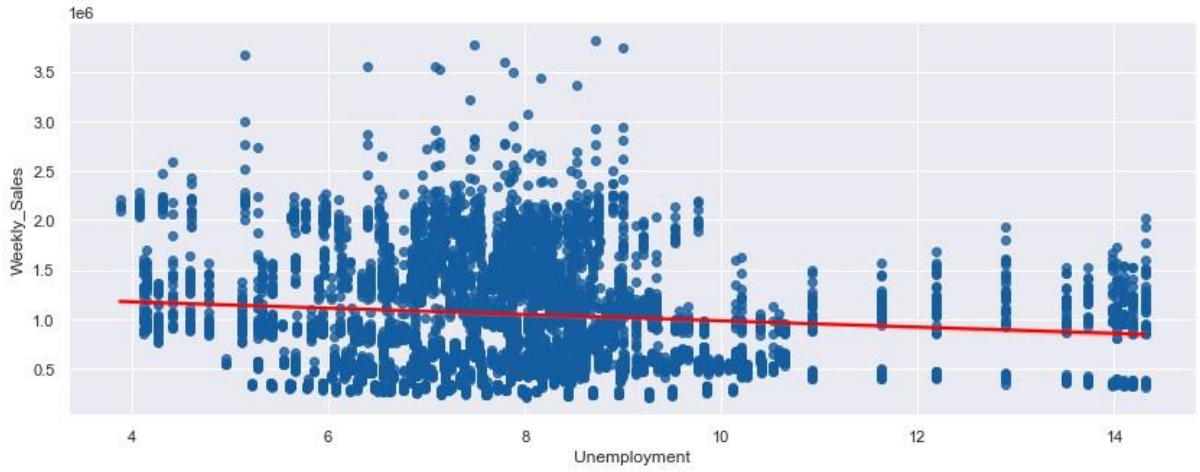

Fig 6.1 Pearson Correlation between weekly sales and unemployment

Since the p-value is < 0.001, the correlation between unemployment and weekly sales is strong evidence that the correlation is significant.

Unemployment seems like a good predictor of the weekly sales, The higher the unemployment rate, the lower the weekly sales.

Store 12,28 and 38 are affected the most.

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

Seasonality trends analysis can be extremely valuable for businesses, as it allows us to better forecast future sales, make more informed decisions about inventory and staffing, and understand the drivers of customer demand leading to improved efficiency and profitability.

We will create a pivot table to group the data by month and year and calculate the average sales for each period. We will then plot the average sales of the table using line chart for the three years. This will allow us to see if there are any patterns in the data that repeat at regular intervals.

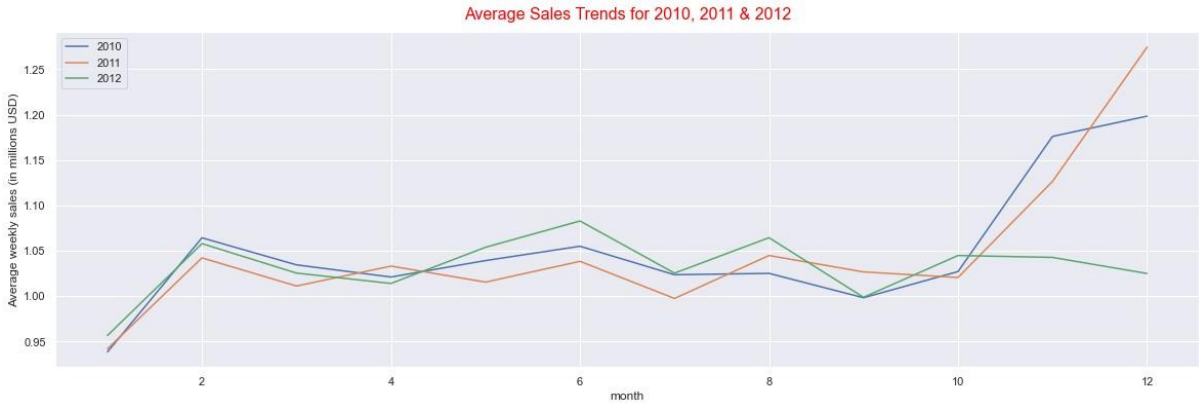

Fig 6.2 Average Sales Trend

We can observe that the line charts for the three years for the month of January to October simultaneously follow a sawtooth shape with big rises experienced in November and December due to holidays. This indicates seasonality trends as months do have consistencies in bigger or smaller sales for the three years. We can also observe that although 2011 performed worse than 2010 in terms of average sales for Walmart, the trend was reversed for the year 2012 which performed better than 2010. However, the data for 2012 ends in October, which may explain the significant drop in sales for November."

6.3 Does temperature affect the weekly sales in any manner?

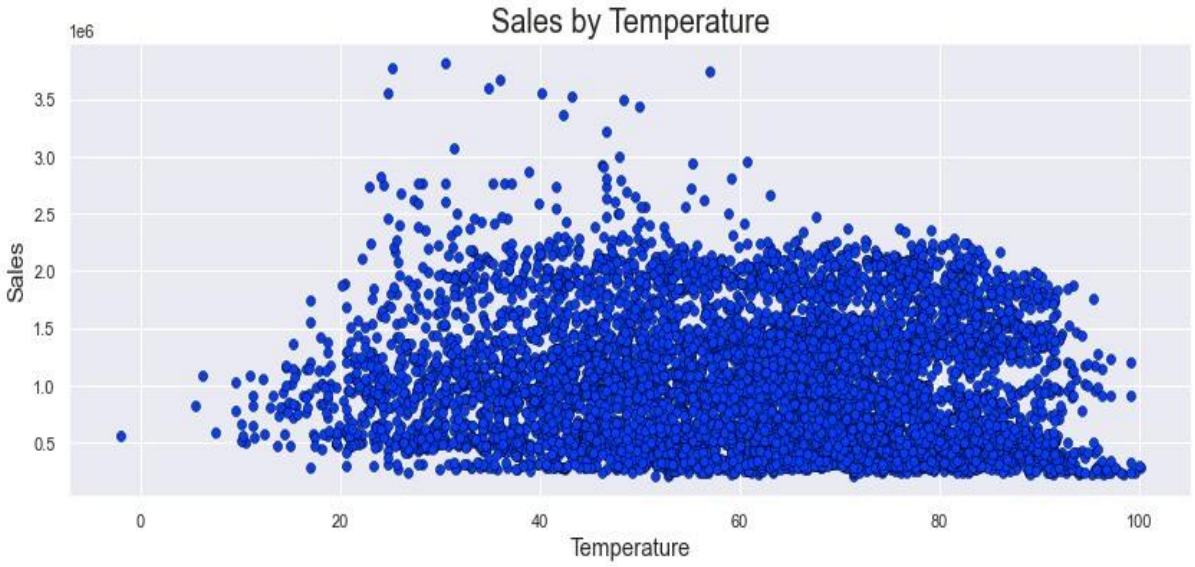

Fig 6.3 Sales vs Temperature

Sales are not affected by changes in temperature.

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

Fig 6.4 Sales vs CPI

Consumer Price Index (CPI) does not affect sales. And based on the distribution of average consumer prices in the above figure, customers can be divided into two categories:

customers who pay from 120 to 150 (Middle-class customers). customers who pay from 180 to 230 (High-class customers).

6.5 Top performing stores according to the historical data

Fig 6.5 Store Performance

There is a high variance in weekly sales from one store to another.

Store No. 20 has the highest sales from any store with 301,397,792 followed by Store No. 4 with 299,543,953 and Store No. 33 comes last with 37,160,222$.

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

Worst performing store is store number 33 with 37,160,222. The significant difference between best performing store with sales 301,397,792 is 264,237,570

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

To predict the next 12 weeks sales, Time series SARIMAX trained model was chosen as it is best suited to predict on weekly basis. The predicted results are saved as pickle file.

7. Future Possibilities

With growing technology and increasing consumer demand, Walmart can shift its focus on e-commerce aspect of the business. Walmart can grow its retail business with already established warehouses and using predictive modelling in predicting the future sales.

The Walmart dataset contained past data of certain stores which helped us to build model to predict the future sales. The study also helped us to analyse the buying behaviour and trend of the products over the seasons. Customer segmentation can help the organization in reaching out to customers by targeted messages of particular product based on regions thus establishing a better customer relationships.

8. Conclusion

The main purpose of the study was to predict Walmart sales based on available historic data and identify whether factors like weather, unemployment, fuel etc would affect the sales. The study also aims to analyse the relationship between holiday season and sales over the period.

As observed through exploratory data analysis, various findings were conclude which helped in building a suitable model using the most relevant features. Relationships between independent and target variables were also observed.

Finally certain suitable algorithms were implemented and metrics analysis was performed. Random Forest Regressor turned out to be the best model to train and predict. Random forest avoided overfitting and provided best results. To predict the future sales Time Series using SARIMAX was best suited.

9. References

- Intellipaat recorded lectures and study material.
- https://towardsdatascience.com/exploratory-data-analysis-using-spermarket-sales-data-in-python-e99d329a07fc
- https://hdsc.medium.com/a-beginners-guide-to-eda-tutorial-country-sales-data-cf0eb56ca213
- https://scikit-learn.org/stable/
- https://matplotlib.org/
- https://seaborn.pydata.org/