

Predicting Walmart Sales, Exploratory Data Analysis, and Walmart Sales Dashboard

By

Siva Raj K

Dedication:

To my Family and Guvi mentors who have supported me endlessly throughout this Data science journey.

Predicting Walmart Sales, Exploratory Data Analysis, and Walmart Sales Dashboard:

Data About:

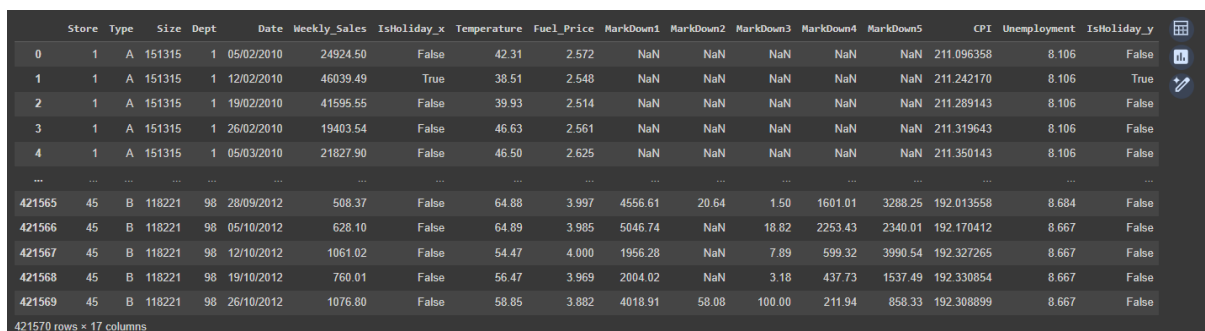
Data is one of the most essential commodities for any organization in the 21st century. Harnessing data and utilizing it to create effective marketing strategies and making better decisions is extremely essential for organizations. For a conglomerate as big as Walmart, it is necessary to organize and analyze the large volumes of data generated to make sense of existing performance and identify growth potential. The main goal of this project is to understand how different factors affect the sales for this conglomerate and how these findings could be used to create more efficient plans and strategies directed at increasing revenue.

This paper explores the performance of a subset of Walmart stores and forecasts future weekly sales for these stores based on several models including linear and lasso regression, random forest, and gradient boosting. An exploratory data analysis has been performed on the dataset to explore the effects of different factors like holidays, fuel price, and temperature on Walmart's weekly sales.

About the data-set:

It contains historic weekly sales information about 45 Walmart stores across different regions in the country along with department-wide information for these stores.

The main goal of this study is going to be to predict the department-wide weekly sales for each of these stores. 17 columns and 421570 rows in total.



	Store	Type	Size	Dept	Date	Weekly_Sales	IsHoliday_x	Temperature	Fuel_Price	Markdown1	Markdown2	Markdown3	Markdown4	Markdown5	CPI	Unemployment	IsHoliday_y
0	1	A	151315	1	05/02/2010	24924.50	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False
1	1	A	151315	1	12/02/2010	46039.49	True	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
2	1	A	151315	1	19/02/2010	41595.55	False	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	False
3	1	A	151315	1	26/02/2010	19403.54	False	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	False
4	1	A	151315	1	05/03/2010	21827.90	False	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	False
...
421565	45	B	118221	98	28/09/2012	508.37	False	64.88	3.997	4556.61	20.64	1.50	1601.01	3288.25	192.013558	8.684	False
421566	45	B	118221	98	05/10/2012	628.10	False	64.89	3.985	5046.74	NaN	18.82	2253.43	2340.01	192.170412	8.667	False
421567	45	B	118221	98	12/10/2012	1061.02	False	54.47	4.000	1956.28	NaN	7.89	599.32	3990.54	192.327265	8.667	False
421568	45	B	118221	98	19/10/2012	760.01	False	56.47	3.969	2004.02	NaN	3.18	437.73	1537.49	192.330854	8.667	False
421569	45	B	118221	98	26/10/2012	1076.80	False	58.85	3.882	4018.91	58.08	100.00	211.94	858.33	192.308899	8.667	False

421570 rows × 17 columns

Another big aspect of this study is to determine whether there is an increase in the weekly store sales because of changes in temperature, fuel prices, holidays, markdowns, unemployment rate, and fluctuations in consumer price indexes, the above sample screen-shot contains all necessary factors along with weekly sales.

Here, is the Holidays which is True in our dataset:

It consists of Christmas, New-year(30th dec). Other than Christmas and New year, there is no consistent Holidays in our dataset.

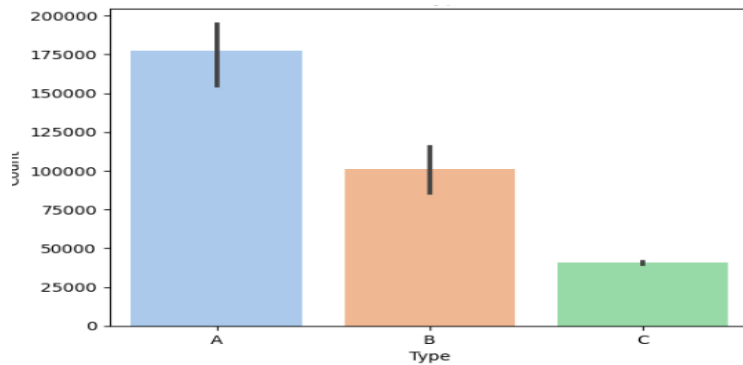
```
date_list
[ '2011-11-25',
  '2011-12-30',
  '2012-10-02',
  '2012-07-09',
  '2011-09-09',
  '2010-12-02',
  '2010-12-31',
  '2010-11-26',
  '2011-11-02',
  '2010-10-09' ]
```

Exploratory Data Analysis:

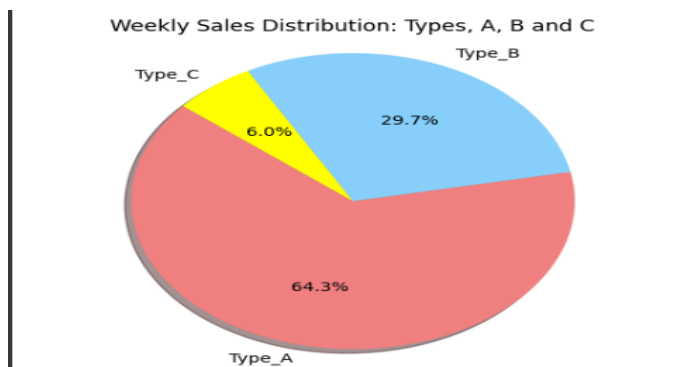
It is crucial to have an in-depth understanding of the dataset that is used in this analysis to understand the models that would give the most accurate prediction. Several times there are underlying patterns or trends in the data that would not be identified as easily, hence the need for an extensive exploratory data analysis. This thorough examination is necessary to understand the underlying structure of the dataset and to draw conclusions or insight about the validity of our analysis. The study is going to begin with a brief analysis of the available dataset to get a sense of the main characteristics and components that are relevant to the research. An exploratory data analysis is crucial to this study considering the numerous attributes that are a part of the dataset that will be essential when trying to draw insights and making predictions. As part of the exploratory data analysis, several visualizations have been created that will help us understand what it is that we are trying to achieve and to keep in mind the various attributes that we can use to improve results.

Analysis high number of stores: Here we have only 3 types of stores('A', 'B', 'C'). In which, it clearly seen that A – type store had high numbers, B in 2nd place and C in 3rd place.

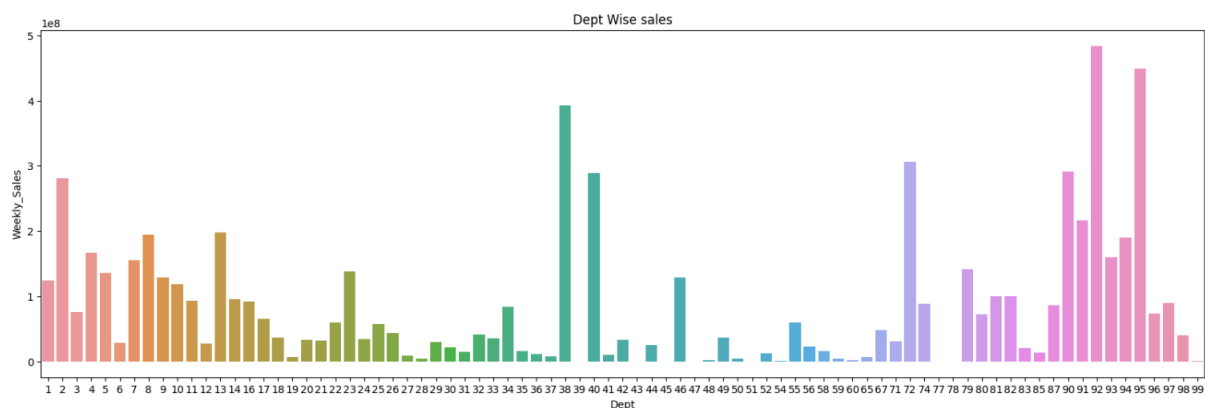
Fig for Highest number of stores:



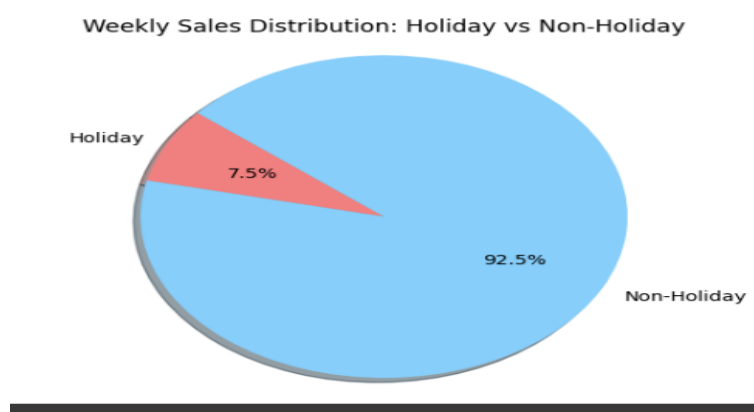
- *Pie-Graph for Weekly sales distribution of Types A, B & C: In which, we can easily find that Type A had high number of weekly sales while comparing to other Types.*



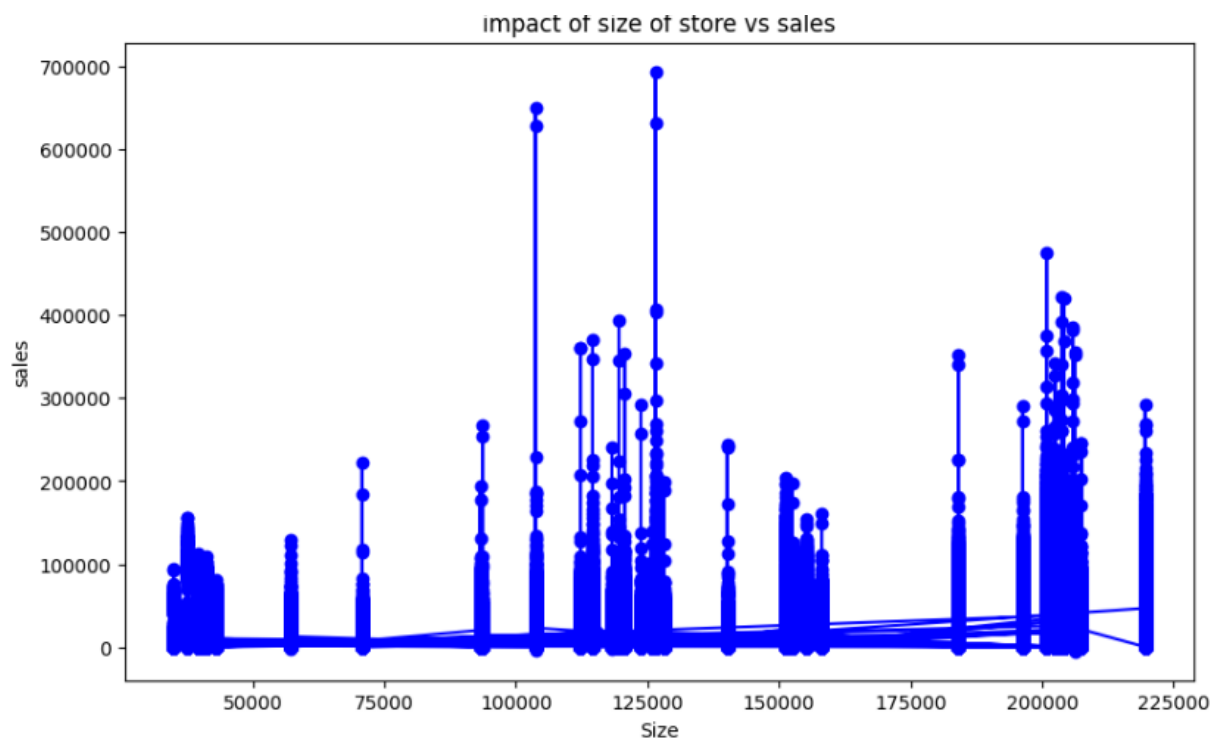
- *Yes, from this, we need to find the departmental wise sales. We have nearly 98 department. In this, dept = [38, 40, 72, 90, 92, 95] had more number of sales.*



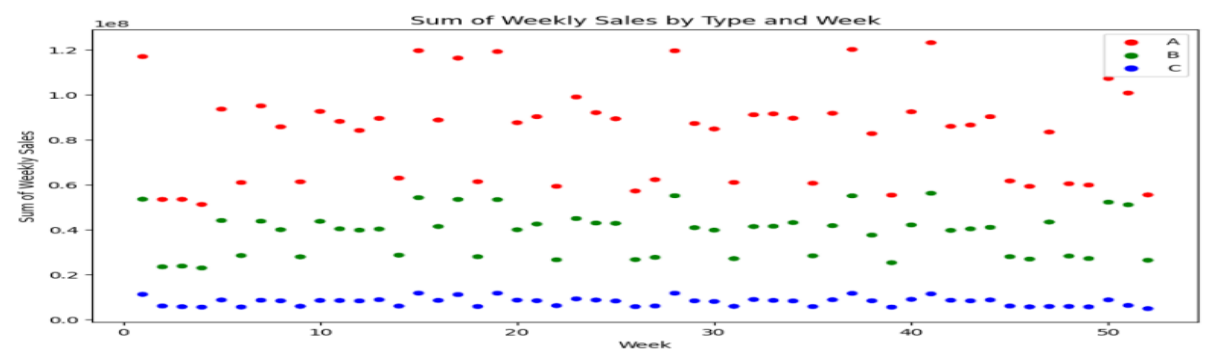
- *Next, while comparing weekly sales with Holidays and non – Holidays, we can seen that Non-holidays days had high number of sales.*
- *There are many reasons can affect the sales on Holidays list, as most of the holidays are non consistent.*



- Finding the impact of store size and weekly_sales. It been observed that, the size would not affect the sales of store. Even less size store can attain decent amount of profit in weekly sales.

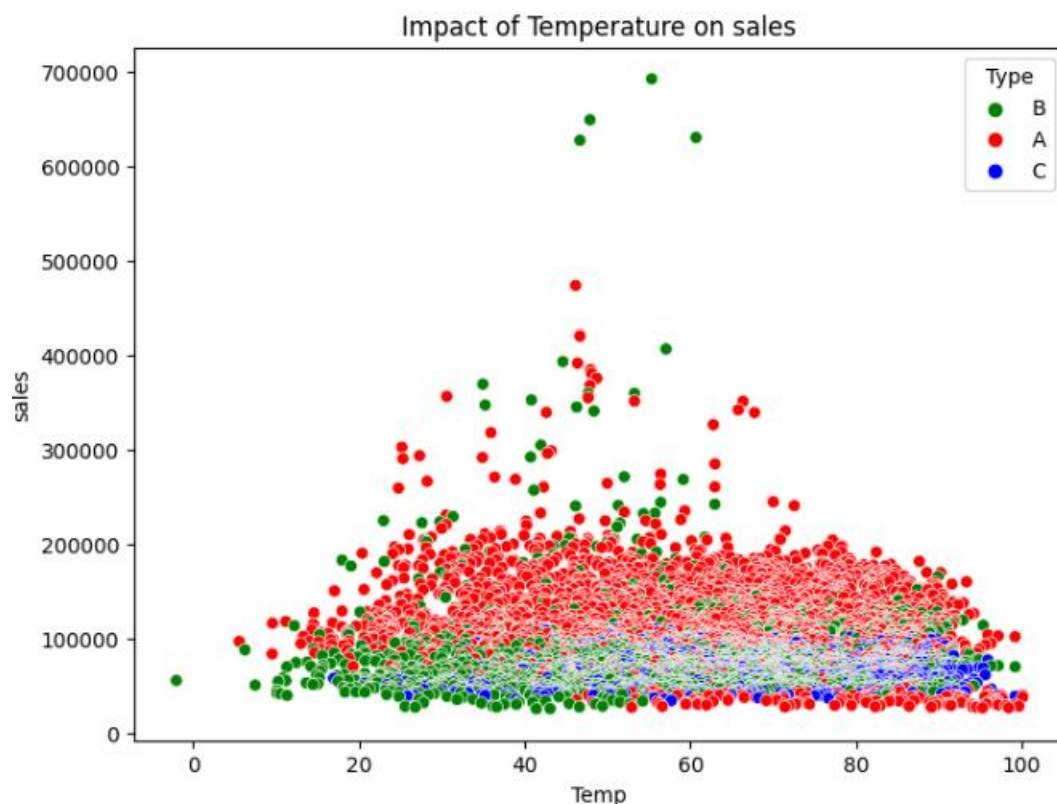


- Yes, we already knew from above analysis, Type A had high number of sales. Lets see the trend of sales from week over week for better understanding.



Impact of Temperature on Sales:

It has widely been known in the retail sector that weather has a profound effect on sales. While warmer weather promotes sales, cold/harsh or extremely hot weather is generally not a great encouragement for shoppers to get outdoors and spend money. Generally speaking, temperatures between 40 to 70 degrees Fahrenheit are considered as favorable for humans to live in considering they are not as hot or cold. As seen below, the highest sales occur for most store types between the range of 40 to 80 degrees Fahrenheit, thus proving the idea that pleasant weather encourages higher sales. Sales are relatively lower for very low and very high temperatures but seem to be adequately high for favorable climate conditions.

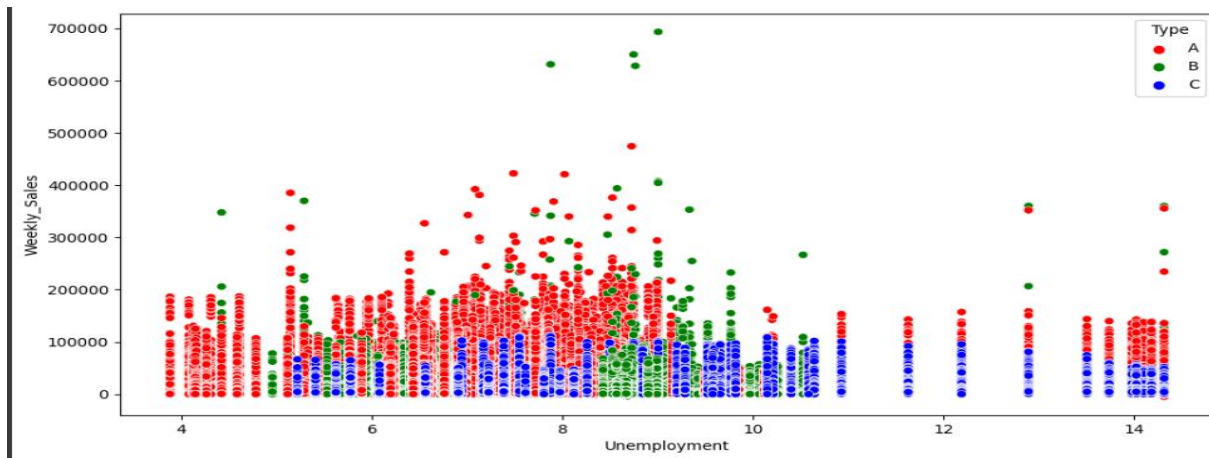


Impact of Unemployment with Sales:

Spending sharply drops on the onset of unemployment; a higher unemployment index would generally result in a dip in sales as individuals tend to decrease overall spending. In our dataset, unemployment is presented through an index of the unemployment rate during that week in the region of the store. From our scatter plot, it is easier to gather the following information:

- For the given store types, there seems to be a visible decrease in sales when the unemployment index is higher than 11
- Even when the unemployment index is higher than 11, there is no significant change in the average sales for Type C stores when compared to the overall sales

- There seems to be a significant drop in sales for store types A and B when the unemployment index increases
- Highest recorded sales for store types A and B occur around the unemployment index of 8 to 10; this gives ambiguous ideas about the impact of unemployment on sales for each of the stores



Model Selection and Implementation:

Model selection in Machine learning is the process of selecting best model and algorithm for a specific job or dataset.

Train Test split: We need to split the data into training set and testing set, where training set is used to train the model and test set is used to test the unseen data with the help of trained model, to determine the accuracy of the model.

```
X = df_1.drop('Weekly_Sales', axis = 1)
y = df_1['Weekly_Sales']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42)
```

Used Algorithms:

```
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error

from sklearn.tree import DecisionTreeRegressor
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.neighbors import KNeighborsRegressor
from lightgbm import LGBMRegressor
from xgboost import XGBRegressor
```

Used, R2 score, mean squared error & mean absolute error for validation metrics.

R² Scores:

	Model	R ² Score	MSE Score	MAE Score
0	DecisionTreeRegressor	0.939117	3.174849e+07	2202.153941
1	LinearRegression	0.089494	4.748024e+08	14551.873248
2	GradientBoostingRegressor	0.738415	1.364091e+08	6918.654114
3	KNeighborsRegressor	0.344759	3.416893e+08	11384.471601
4	LGBMRegressor	0.903140	5.050964e+07	4214.225788
5	XGBRegressor	0.932767	3.506001e+07	3245.224909

Upon checking, DecisionTreeRegressor had high R² score.

Lets see how Decision Tree algorithm working....

DecisionTreeRegressor:

What is a Decision Tree?

A decision tree is **a non-parametric supervised learning algorithm for classification and regression tasks**. It has a hierarchical tree structure consisting of a root node, branches, internal nodes, and leaf nodes. Decision trees are used for classification and regression tasks, providing easy-to-understand models.

How decision tree algorithms work?

Decision Tree algorithm works in simpler steps

- 1. Starting at the Root:** The algorithm begins at the top, called the “root node,” representing the entire dataset.
- 2. Asking the Best Questions:** It looks for the most important feature or question that splits the data into the most distinct groups. This is like asking a question at a fork in the tree.
- 3. Branching Out:** Based on the answer to that question, it divides the data into smaller subsets, creating new branches. Each branch represents a possible route through the tree.

4. **Repeating the Process:** The algorithm continues asking questions and splitting the data at each branch until it reaches the final “leaf nodes,” representing the predicted outcomes or classifications.

Decision Tree Assumptions

Several assumptions are made to build effective models when creating decision trees. These assumptions help guide the tree's construction and impact its performance. Here are some common assumptions and considerations when creating decision trees:

Binary Splits:

Decision trees typically make binary splits, meaning each node divides the data into two subsets based on a single feature or condition. This assumes that each decision can be represented as a binary choice.

Recursive Partitioning:

Decision trees use a recursive partitioning process, where each node is divided into child nodes, and this process continues until a stopping criterion is met. This assumes that data can be effectively subdivided into smaller, more manageable subsets.

Feature Independence:

Decision trees often assume that the features used for splitting nodes are independent. In practice, feature independence may not hold, but decision trees can still perform well if features are correlated.

Homogeneity:

Decision trees aim to create homogeneous subgroups in each node, meaning that the samples within a node are as similar as possible regarding the target variable. This assumption helps in achieving clear decision boundaries.

Top-Down Greedy Approach

Decision trees are constructed using a top-down, greedy approach, where each split is chosen to maximize information gain or minimize impurity at the current node. This may not always result in the globally optimal tree.

Categorical and Numerical Features

Decision trees can handle both categorical and numerical features. However, they may require different splitting strategies for each type.

Overfitting

Decision trees are prone to overfitting when they capture noise in the data. Pruning and setting appropriate stopping criteria are used to address this assumption.

Impurity Measures

Decision trees use impurity measures such as Gini impurity or entropy to evaluate how well a split separates classes. The choice of impurity measure can impact tree construction.

No Missing Values

Decision trees assume that there are no missing values in the dataset or that missing values have been appropriately handled through imputation or other methods.

Equal Importance of Features

Decision trees may assume equal importance for all features unless feature scaling or weighting is applied to emphasize certain features.

No Outliers

Decision trees are sensitive to outliers, and extreme values can influence their construction. Preprocessing or robust methods may be needed to handle outliers effectively.

Sensitivity to Sample Size:

Small datasets may lead to overfitting, and large datasets may result in overly complex trees. The sample size and tree depth should be balanced.

Conclusion:

The main purpose of this study was to predict Walmart's sales based on the available historic data and identify whether factors like temperature, unemployment, fuel prices, etc affect the weekly sales of particular stores under study. This study also aims to understand whether sales are relatively higher during holidays like Christmas and Thanksgiving than normal days so that stores can work on creating promotional offers that increase sales and generate higher revenue. As observed through the exploratory data analysis, store size and holidays have a direct relationship with high Walmart sales. It was also observed that out of all the store types, Type A stores gathered the most sales for Walmart. Additionally, departments 92, 95, 38, and 72 accumulate the most sales for Walmart stores across all three store types; for all of the 45 stores, the presence of these departments in a store ensures higher sales. Pertaining to the specific factors provided in the study (temperature, unemployment, CPI, and fuel price), it was observed that sales do tend to go up slightly during favorable climate conditions as well as when the prices of fuel are adequate. However, it is difficult to make a strong claim about this assumption considering the limited scope of the training dataset provided as part of this study. By the observations in the exploratory data analysis, sales also tend to be relatively higher when the unemployment level is lower. Additionally, with the dataset provided for this study, there does not seem to be a relationship between sales and the CPI index. Again, it is hard to make a substantial claim about these findings without the presence of a larger training dataset with additional information available.