CAPESTONE PROJECT : WALMART PROJECT

Table of Contents

1. Problem Statement
2. Project Objective
3. Data Description
4. Data Pre-processing Steps and Inspiration
5. Choosing the Algorithm for the Project
6. Motivation and Reasons For Choosing the Algorithm
7. Assumptions
8. Model Evaluation and Techniques
9. Inferences from the Same
10. Future Possibilities of the Project
11. Conclusion
12. References

Problem Statement: Financial Crisis and Inventory Management

Challenge : Our company is facing a financial crisis and needs to save money quickly.

Problem : The challenge is to find smart ways to manage our inventory and predict what we'll need to buy, so we can save money during this crisis.

Importance : Good inventory management can help us save money and ease our financial situation during tough times.

We'll focus on using data to make better decisions about what to keep in stock and what to buy.
We'll use our sales history, inventory records, and outside factors to help us make smarter choices about our inventory.

Project Objective

Our main goals are to:

1. Predict what we'll need to buy more accurately.
2. Spend money wisely on inventory.
3. Improve our financial situation during th e crisis.

Focus : This project is for our finance, inventory, and supply chain teams.

Challenges: Challenges include making the best financial decisions during the crisis, dealing with data quality issues, and being flexible as things change.

Assumptions: We'll assume that outside factors, like market conditions, won't change too drastically during the crisis.

Company Goals: Solving this problem will help us weather the financial crisis by managing our inventory better and spending wisely.

Data Description :

DataSet name : Walmart

Data source: The dataset used in this analysis was read from a CSV (Comma-Separated Values) file named [Walmart (1).csv].

Function - head (): shows all the columns and the top 5 rows ,
tail() : will shows all columns and the last 5 rows

Data size : data is comprise of 8 columns and 6435 rows

Data Structure : The dataset is structured with the following columns:

Columns name	Descriptions
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
Unemployment	Unemployment Rate

columns datatypes: what types of data is present in data set:

1 integer(numeric)
2 object(string)
3 float(deciaml numbers)

datatype covertion: converting Date object column to datetime format , for better perform .
null_value : there is nonull value in dataset
duplicates : no duplicate values is present in dataset.

Statistical analysis :

Store:

count of store =6435.000000
average store value=23.000000
standard deviation(how far data is from mean)=12.988182
minimum number of store =1
max number of store=45.000000

Quartiles are statistical measures that divide a dataset into four equal parts, each containing 25% of the data points. Quartiles are used to understand the distribution of a dataset

25% quartile=12.000000
50% quartile(median)=23.000000
75% quartile =34.000000

Statistical analysis :

Weekly Sales:

count of Weekly =6435.000000
average Weekly Sales value=1.046965e+06
standard deviation(how far data is from mean)=5.643666e+05
minimum number of sales in a week =2.099862e+05
max number of sales in a week=3.818686e+06

Quartiles are statistical measures that divide a dataset into four equal parts, each containing 25% of the data points. Quartiles are used to understand the distribution of a dataset

25% quartile=5.533501e+05
50% quartile(median)=9.607460e+05
75% quartile =1.420159e+06

Statistical analysis :

Holiday_Flag:

count =6435.000000
average Holiday=0.069930
standard deviation(how far data is from mean)=0.255049
minimum number of Holiday =0.000000
max number of Holiday=1.000000

Quartiles are statistical measures that divide a dataset into four equal parts, each containing 25% of the data points. Quartiles are used to understand the distribution of a dataset

25% quartile=0.000000
50% quartile(median)=0.000000
75% quartile =0.000000

Statistical analysis :

Temperature:

count =6435.000000
average Temperature=60.663782
standard deviation(how far data is from mean)=18.444933
minimum Temperature=-2.060000
max Temperature=100.140000

Quartiles are statistical measures that divide a dataset into four equal parts, each containing 25% of the data points. Quartiles are used to understand the distribution of a dataset

25% quartile=47.460000
50% quartile(median)=62.670000
75% quartile =74.940000

Statistical analysis :

Fuel_Price:

count =6435.000000
average fuel consumpition=3.358607
standard deviation(how far data is from mean)=0.459020
minimum fuel=2.472000
max fuel=4.468000

Quartiles are statistical measures that divide a dataset into four equal parts, each containing 25% of the data points. Quartiles are used to understand the distribution of a dataset

25% quartile=2.933000
50% quartile(median)=3.445000
75% quartile =3.735000

Statistical analysis :

CPI:

count =6435.000000
average =171.578394
standard deviation(how far data is from mean)=39.356712
minimum =126.064000
max =227.232807

Quartiles are statistical measures that divide a dataset into four equal parts, each containing 25% of the data points. Quartiles are used to understand the distribution of a dataset

25% quartile=131.735000
50% quartile(median)=182.616521
75% quartile =212.743293

Statistical analysis :

Unemployment:

count =6435.000000
average =7.999151
standard deviation(how far data is from mean)=1.875885
minimum =3.879000
max =14.313000

Quartiles are statistical measures that divide a dataset into four equal parts, each containing 25% of the data points. Quartiles are used to understand the distribution of a dataset

25% quartile=6.891000
50% quartile(median)=7.874000
75% quartile =8.622000

Data Pre-processing Steps

VISUALIZATION OF DATA

Transforming a date column into separate day, month, and year columns to facilitate sales data visualization by day, month, and year.

Through the power of data visualization, we uncover compelling insights:

1. The 31st day consistently registers lower sales compared to other days.
2. Monthly sales peaks in April and July, while January and November show lower sales.
3. Notably, 2011 outshines 2010 and 2012 in sales performance.
4. An intriguing correlation emerges: as sales rise, unemployment rates tend to decrease, highlighting the potential economic impact of higher sales.
5. Temperature spikes correspond with fuel price increases, possibly attributed to heightened vehicle usage in warmer weather—a revealing connection between climate and fuel consumption."

Detecting Outliers in the Dataset

Outliers in a dataset are data points that significantly differ from the rest of the data. They are rare and unusual, and they can have a substantial impact on statistical analysis and modeling. Outliers can occur due to errors in data collection or measurement errors.

One way to visualize and identify outliers is by using boxplots, which are available in the matplotlib.pyplot library. Boxplots provide a quick and intuitive way to visualize the spread and identify potential outliers in a dataset.

However, there are other methods for detecting outliers, such as using the z-score and the interquartile range (IQR). These methods can provide statistical measures of how far a data point is from the mean or the median.

In this project, i used boxplots to identify outliers. You found the presence of outliers in the "Weekly Sales," "Temperature," and "Unemployment" columns.

It's important to deal with outliers appropriately in your analysis, whether by removing them, transforming the data, or applying robust statistical techniques, depending on the nature of your data and the impact of outliers on your analysis.

"For this project, I have chosen not to address outliers because, in the context of time series analysis, removing outliers is not currently a part of my approach."

In the upcoming plot, we will visualize the weekly sales data for all the stores.

Highest selling stores:

store 2,
store 5,
store 10,
store 13,
store 14

Medium selling stores:

store 1
store 6,
store 11,
store 12,
store 19,
store 23
store 24
store 28
store 31
store 32
store 39
store 41

Other store have low selling.

Based on historical data, the store with the store number 33 had the worst performance with weekly sales of 37,160,221.96, while the store with the store number 20 was the best performer with weekly sales of 301,397,800

Analyzing Correlations:

Analyzing the correlation matrix allows us to understand the relationships between different columns in the dataset. Positive correlations are denoted as '+', zero correlations as '0', and negative correlations as '-'. This matrix helps identify the strength and direction of relationships between variables.

Time Series Forecasting:

In your next step, you're applying time series models to forecast 12 weeks of sales for all 45 stores. Time series forecasting involves using historical data to make predictions about future sales trends. This analysis will provide valuable insights into sales patterns and help with decision-making and resource allocation for the stores.

By utilizing the correlation matrix, we've identified the degree of correlation (positive, negative, or zero) between various columns, shedding light on which columns exhibit high, low, or no correlation with each other.

Motivation and Reasons for Choosing the Time Series Model:

In the context of time series analysis, the selection of an appropriate model is pivotal to the success of the project. Here, I outline the motivations and reasons behind opting for a time series model:

1. Problem Alignment: The project's primary objective revolves around forecasting future sales trends, a task inherently suited for time series analysis. The time-dependent nature of sales data demands a model capable of capturing sequential dependencies and seasonality.
2. Historical Precedence: Extensive research and practical applications have demonstrated the effectiveness of time series models for forecasting tasks in various domains. The time series model's historical success in similar scenarios provides a strong rationale for its selection.
3. Temporal Patterns: The dataset exhibits clear temporal patterns, with sales data varying over weeks or months. A time series model's inherent ability to account for these patterns was a compelling reason for its adoption.
4. Data Structure: The data's sequential nature, where each data point is influenced by prior observations, aligns well with the assumptions underlying time series models. This structural congruence was a critical factor in the decision-making .
5. Intrinsic Seasonality: Given that the sales data may exhibit seasonality (e.g., weekly or monthly sales patterns), the time series model's capability to handle and interpret seasonality was pivotal.

Assumptions:

Assumption 1: Stationarity of the Time Series

We assume that the time series data used in this project exhibits stationarity. Stationarity is a fundamental assumption in time series analysis, and it implies that the statistical properties of the data, such as mean and variance, remain constant over time. We have made this assumption to apply time series models effectively. If the data is not stationary, transformations or differencing may be necessary to achieve stationarity before modeling.

It's essential to be explicit about the assumptions you're making in your project, as they help readers understand the context and limitations of your analysis. You may have additional assumptions that you want to list and explain in a similar manner.

Assumption 2: Independence of Observations

We assume that the observations in the time series data are independent of each other. This means that the value of a data point at one time step is not influenced by the values at previous or future time steps. While independence of observations is a common assumption in time series modeling, it's important to acknowledge that real-world data may not always meet this assumption. If there is evidence of autocorrelation or temporal dependencies in the data, it may require more advanced modeling techniques to account for such dependencies.

By stating your assumptions clearly, you provide transparency regarding the conditions under which your analysis is valid and help readers understand the context of your work

Model Evaluation and Techniques:

Visual Inspection:

In this section, we visually inspected the model's performance by comparing the forecasted values to the actual sales data. Visualizations played a pivotal role in our evaluation process, offering a more intuitive understanding of how well the model captured the underlying patterns and trends in the data.

- Time Series Plots: We presented time series plots that display both the observed sales data and the model's forecasts. These plots allowed us to assess the alignment of predicted and actual values over time, making it easier to spot discrepancies or trends.
- Forecasted vs. Actual Sales Graphs: Utilizing forecasted vs. actual sales graphs, we visually compared the model's predictions against the true sales values. These graphs provided a straightforward means to evaluate the accuracy of our forecasts.

Inferences from the Project:

In this section, we present the main inferences and insights gleaned from our project. These findings are derived from our analysis and modeling efforts and shed light on the key takeaways:

1. Sales Trends:

- Our analysis revealed distinct sales trends among the 45 stores. Certain stores consistently outperformed others, indicating variations in sales performance across the dataset.

2. Seasonality and Patterns:

- We identified clear seasonal patterns in the sales data, with some stores experiencing increased sales during specific times of the year. This seasonality highlights the need for tailored forecasting strategies.

3. Time Series Model Effectiveness:

- The selected time series model effectively captured the temporal dependencies and trends in the sales data. It demonstrated its utility in forecasting sales trends for each store.

4. Outlier Identification:

- While we identified the presence of outliers in some columns, we opted not to address them due to their potential significance in time series analysis.

5. Store-Specific Insights:

- Our analysis provided a deeper understanding of the highest and lowest performing stores, enabling tailored strategies for each category.

6. Future Directions:

- We've identified areas for future work, including the refinement of forecasting techniques and the potential incorporation of additional features to improve the model's accuracy.

These inferences provide a comprehensive view of the project's outcomes, facilitating data-driven decisions and offering insights into improving sales forecasting for the 45 stores.

Conclusion:

In closing, this project encapsulates the insights and findings derived from our in-depth analysis of sales data for the 45 stores over a 12-week forecast period. It serves as a valuable resource for optimizing sales forecasting and decision-making, offering a data-driven approach to understanding and enhancing sales performance.

Visualizing the 12-Week Forecasts for 45 Stores: To enhance the project's overall impact and provide a compelling visual representation of our 12-week sales forecasts for all 45 stores, we've included a comprehensive plot that displays the forecasts in a single view. This visualization not only underscores the diversity in sales trends but also highlights the model's capacity to provide reliable and accurate forecasts for each store.

Significance: The insights gained from this project empower businesses to make informed decisions regarding inventory management, staffing, and other vital aspects of their operations. By leveraging the seasonality, trends, and reliable forecasting techniques, these stores are better equipped to address the unique challenges and opportunities within their sales data.

In conclusion, this project not only offers actionable insights into sales data but also provides a visually compelling overview of the 12-week sales forecasts for all 45 stores, further enhancing the accessibility and utility of the analysis.

References:

Kaggle. (Year). Title of the dataset or project. URL: https://www.kaggle.com

scikit-learn. (Year). Title of the library or specific page. URL: https://scikit-learn.org

TutorialsPoint Time Series Analysis Tutorial:

https://www.tutorialspoint.com/time_series_analysis/index.htm

YouTube Username.

https://www.youtube.com/username

THANK YOU