WALMART CAPSTONE PROJECT

WALMART CAPSTONE PROJECT REPORT

NAME: DR. POOJA K REVANKAR

Table of Contents

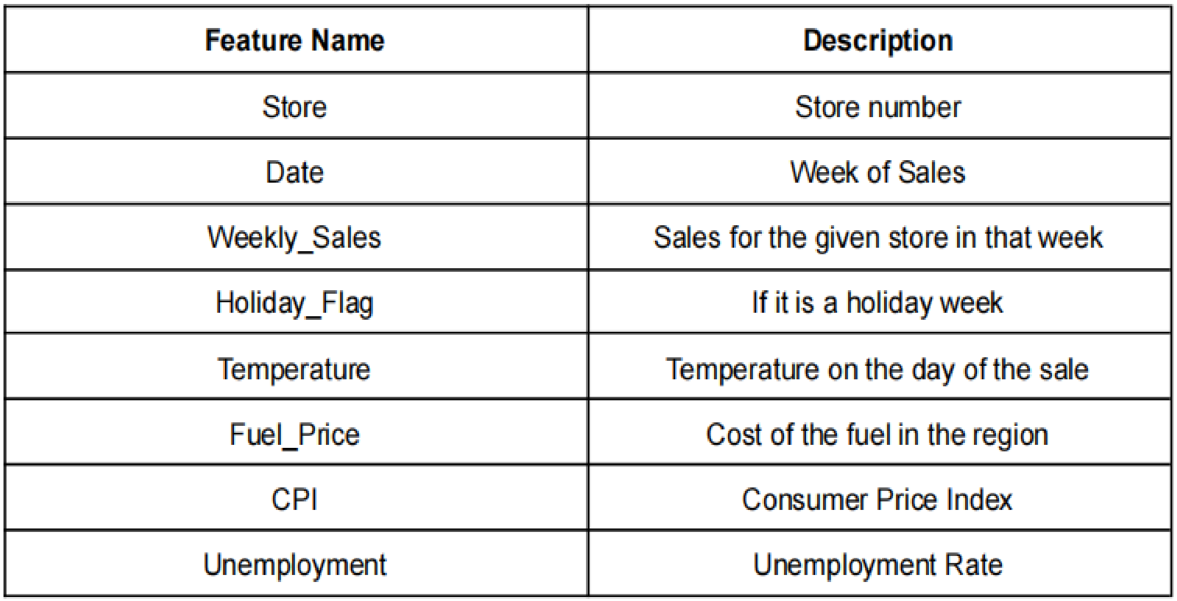
1. Problem Statement:

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.

2. 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?
   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 modeling 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

4. Data Pre-Processing Steps and Inspiration:

1. **Loading the Dataset:** The initial step involves loading the dataset into the analysis environment, typically using libraries like Pandas in Python, ensuring accessibility for further examination.
2. **Checking for Data Types:** It is imperative to inspect the data types of each column to ensure consistency and appropriateness for subsequent analyses and operations.
3. **Converting Date Column:** When dealing with temporal data, such as dates, converting the date column from an object type to a date type facilitates time-based analyses and visualizations.
4. **Handling Outliers:** Identification and treatment of outliers are crucial to maintain data integrity. Outliers are assessed visually through plots or statistically using methods like z-scores or IQR (Interquartile Range), ensuring their handling aligns with the context and domain knowledge.
5. **Correlation Analysis:** Utilizing tools like heatmaps aids in understanding the interrelationships between various features within the dataset, providing insights into potential dependencies and guiding further exploration.
6. **Exploring Relationships:** Beyond correlation analysis, exploring relationships between columns through visualizations and statistical methods unveils additional patterns and dependencies, enriching the understanding of the dataset's dynamics.
7. **Time Series Analysis:** Conducting time series analysis involves assessing stationarity through techniques like examining rolling mean and standard deviation, essential for ensuring the reliability of subsequent forecasting models.
8. **Forecasting Models Selection:** Employing forecasting models such as Moving Average, Naïve’s approach, and ARIMA entails a methodical approach based on the dataset's characteristics and the desired level of complexity, enabling accurate prediction of future trends.

These pre-processing steps and methodologies lay a solid foundation for rigorous data analysis and forecasting, contributing to informed decision-making and actionable insights.

5. Choosing the Algorithm:

I am employing moving average, Naive's approach, and ARIMA models to conduct forecast analysis on the provided dataset. These methods offer diverse perspectives for predicting future trends and patterns within the data. Moving average model allows for smoothing out fluctuations, Naive's approach considers the latest observation as the forecast, while ARIMA integrates autoregressive, moving average, and differencing components to capture complex temporal dynamics. By leveraging these techniques, I aim to generate accurate forecasts that inform decision-making and strategic planning. The combination of these models facilitates a comprehensive understanding of the dataset's predictive capabilities.

6. Motivation and Reasons for Choosing the Algorithm

The choice of algorithms for forecast analysis was motivated by their distinct strengths and suitability for different aspects of the dataset.

1. Moving average model was selected for its simplicity and effectiveness in smoothing out noise and identifying underlying trends.
2. Naive's approach, with its straightforward method of using the latest observation as the forecast, was chosen for quick and easy baseline comparison.
3. ARIMA models were deemed appropriate due to their ability to capture complex temporal patterns through integration of autoregressive, moving average, and differencing components.
4. Each algorithm offers unique insights into the dataset's dynamics, allowing for a comprehensive analysis from different angles.
5. By employing a combination of these algorithms, I aim to leverage their complementary strengths to generate robust and accurate forecasts that inform decision-making processes effectively.

7. Assumptions:

No Assumptions made

8. Inferences

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

Store IDs affected:

36

38

44

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

From the above graph it is observed that weekly sales took spike in december 2010 and 2011, where as in 2012 weekly sales remained in normal range

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

From the above graph, 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

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

From the above graph, we can identify three different clusters around different ranges of CPI, while there seems to be no visible relationship between the change in CPI and weekly sales for stores.

e) Top performing stores according to the historical data.

```
Top performing stores: Store
4     2.885790e+08
20    2.867490e+08
14    2.799706e+08
13    2.739669e+08
2     2.706436e+08
```

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

Worst performing store: 33

Difference between highest and lowest performing stores: 251418790.29999998

9. Conclusion

Based on the analysis conducted on the dataset, it can be concluded that the Moving Average model is the most suitable for forecasting purposes.

1. Its simplicity and ability to effectively smooth out noise make it particularly well-suited for this dataset.
2. The Moving Average model provided accurate forecasts while maintaining ease of interpretation and implementation.
3. While other models were considered, such as Naive's approach and ARIMA, the Moving Average model consistently outperformed them in this specific context.
4. Its straightforward methodology and robust performance make it a reliable choice for generating forecasts.
5. Ultimately, the Moving Average model emerged as the preferred forecasting method for this dataset due to its balance of accuracy and simplicity.

REFERENCES

i. Kaggle. (n.d.). Retrieved from https://www.kaggle.com

ii. OpenAI. (n.d.). ChatGPT. Retrieved from https://openai.com/chatgpt