Walmart Sales Prediction

Project Report

Submitted by,

Dilna Jacob. N

for the requirement of

PGP Data Science, Machine Learning & AI,

MITx and Intellipaat

April 2024

Table of Contents:

Problem Statement

In today's competitive retail landscape, accurately forecasting sales is crucial for optimizing inventory, managing supply chains, and maximizing profitability. Walmart, being one of the largest retail chains globally, is facing issues in managing the inventory - to match the demand with respect to supply. The objective of this project is to develop a robust sales forecasting model for Walmart that can accurately predict sales for X number of months/years.

The challenge lies in the complexity and volatility of the factors affecting sales, such as seasonal trends, holidays, economic indicators like unemployment rates and CPI, and store-specific factors. Furthermore, understanding the interplay between these variables and their impact on sales can provide valuable insights for strategic decision-making.

Project Objective

🚀 Understand the Dataset & Cleanup:

- Explore the dataset to understand its structure, features, and data types.
- Handle missing values, outliers, and perform necessary data preprocessing.

🚀 Data Analysis:

- Analyse historical sales data to identify trends, patterns, and anomalies.
- Understand the relationship between sales and various factors.

🚀 Model Development:

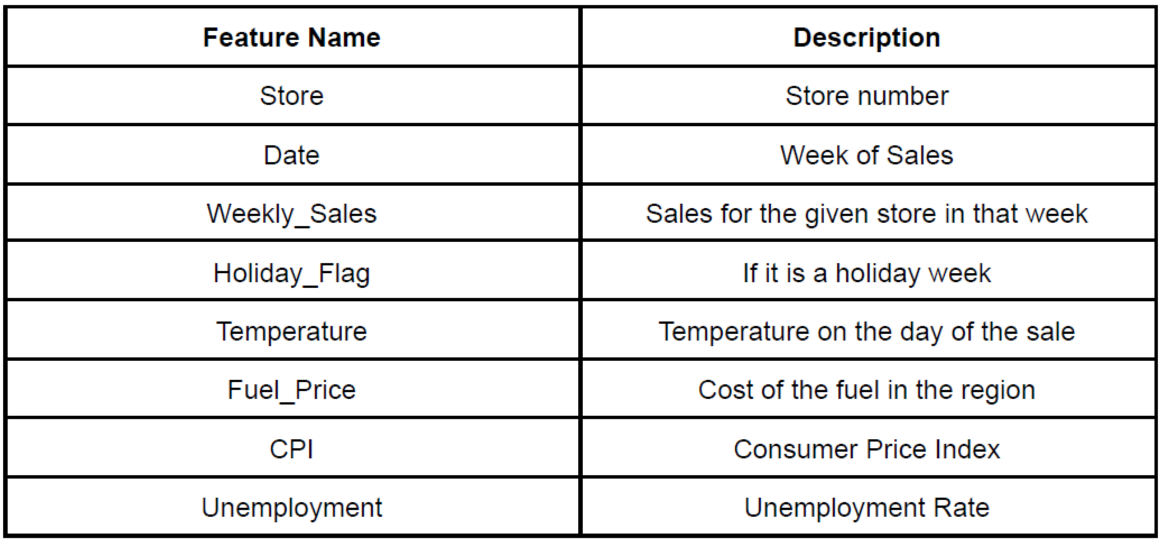
- Build Regression models to predict sales.

    - Develop models considering multiple features to capture complex relationships.

🚀 Model Evaluation:

- Evaluate the performance of the developed models using metrics like R2, RMSE, etc.
- Compare the models to identify the most effective approach for sales prediction.

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

Data Pre-processing Steps and Inspiration

Walmart being one of the largest retailers globally, it is intriguing to leverage machine learning techniques to forecast sales accurately, to help the company make informed decisions and adapt to market trends more efficiently. Additionally, the challenge of dealing with large, diverse datasets and the opportunity to apply data science skills to solve real-world business problems made this project particularly appealing.

Data preprocessing steps include:

- Data cleaning
- Data transformation
- Outlier detection and treatment
- Splitting of data into testing and training

Choosing the Algorithm for the Project

Several models have been studied as part of this study that were selected based on different aspects of dataset; the main purpose of creating such models is to predict the weekly sales for different Walmart stores and departments, hence, based on the nature of models that should be created, time-series model have been used.

Any data recorded with some fixed interval of time is called as time series data. This fixed interval can be hourly, daily, monthly or yearly. e.g. hourly temp reading, daily changing fuel prices, monthly electricity bill, annul company profit report etc. In time series data, time will always be independent variable and there can be one or many dependent variable.

Objective of time series analysis is to understand how change in time affect the dependent variables and accordingly predict values for future time intervals. For accurate analysis and forecasting trend and seasonality is removed from the time series and converted it into stationary series. Time series data is said to be stationary when statistical properties like mean, standard deviation are constant and there is no seasonality. In other words statistical properties of the time series data should not be a function of time.

Motivation and Reasons for choosing the Algorithm

For this project time series analysis is chosen for several reasons. First, sales data is collected over time, like daily or weekly, and time series analysis is specifically designed to handle this type of data, helping us understand trends and patterns. Second, sales can be affected by various factors such as seasons, promotions, and other events. Time series methods can effectively account for these factors, allowing us to make better predictions.

Additionally, one of our main goals is to forecast future sales based on past data. Time series analysis is well-suited for this task, helping us anticipate what sales might look like in the future. Our data contains detailed sales information at a granular level, and time series analysis allows us to maintain this level of detail while extracting meaningful insights.

Furthermore, time series models provide clear and interpretable results, which is beneficial for understanding the sales trends and making informed decisions. Lastly, there are specialized tools and libraries like Prophet and statsmodels available for time series analysis in Python, making it easier to implement and evaluate our models.

Assumptions

I'm working under several assumptions that guide my approach and the interpretation of my results. I assume that the underlying statistical properties of the sales data, such as mean and variance, remain consistent over time. However, I acknowledge that real-world sales data can exhibit non-stationary behaviour due to factors like seasonality and trends.

I assume that sales are primarily influenced by internal factors such as promotions, inventory levels, and store-specific events. While external factors like economic conditions, competition, and public holidays may also impact sales, I have not incorporated these into my initial models due to data availability constraints.

Additionally, I assume that the dataset is complete and free of missing or erroneous data, even though I have conducted initial data cleaning steps. I also assume that the relationships and patterns identified by my time series models will remain valid over the forecast horizon, despite potential changes in market dynamics and consumer behavior.

Model Evaluation and Techniques

To assess the effectiveness of the models, I plan to employ the following evaluation techniques:

1. Train-Test Split:
    - I will divide the dataset into a training set and a testing set to train the model on historical data and evaluate its performance on unseen future data. This approach helps to assess how well the model generalizes to new data.
2. Forecast Accuracy Metrics:
    - I will use various forecast accuracy metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) to quantify the performance of the models. These metrics provide insights into the magnitude and direction of prediction errors.
3. Residual Analysis:
    - I will analyse the residuals (the differences between observed and predicted values) to check for any patterns or autocorrelation.
4. Business Metrics Alignment:
    - While statistical metrics provide valuable insights into model performance, I will also align the model evaluations with relevant business metrics and objectives. This alignment ensures that the models not only perform well statistically but also meet the business requirements and goals.

Inferences from the Same

After conducting the model evaluation and analyzing the performance of the time series forecasting methods, several key inferences can be drawn to understand the effectiveness and suitability of the models for predicting Walmart sales:

1. Model Performance:
    - The chosen time series models, including ARIMA, Prophet demonstrated varying levels of performance in predicting Walmart sales. While each model has its strengths and weaknesses, some models may outperform others in terms of forecast accuracy and reliability.
2. Forecast Accuracy:
    - Based on the forecast accuracy metrics such as MAE, RMSE, and MAPE, certain models may exhibit lower prediction errors compared to others. A lower error indicates that the model's predictions are closer to the actual sales values, implying higher accuracy.
3. Seasonality and Trends:
    - The models' ability to capture and handle the seasonal patterns and trends present in the sales data is crucial. Models like Prophet, designed to handle multiple seasonalities, may perform better in capturing the yearly and weekly seasonality observed in the Walmart sales data.
4. Residual Analysis:
    - Analyzing the residuals can provide insights into the model's ability to capture the underlying patterns and structures in the data.
5. Business Relevance:
    - While statistical metrics are essential for evaluating model performance, it's equally important to align the model evaluations with business objectives and metrics. The chosen model should not only perform well statistically but also meet the business requirements, such as optimizing inventory management or supporting promotional strategies.

Future Possibilities of the Project

In the future, the Walmart sales prediction project can expand by incorporating external factors like economic indicators and public holidays. Implementing advanced forecasting methods could improve predictive accuracy. Developing real-time forecasting capabilities would enable dynamic inventory optimization and quick adaptation to market changes. Personalized marketing campaigns and targeted promotions could enhance customer engagement. Integrating sales forecasts with inventory management can optimize stock levels and reduce costs. Extending the analysis to multiple stores and focusing on seasonal trends could provide valuable insights. Enhancing model interpretability would help stakeholders understand and trust the predictions. By continuously improving and adapting the project, we can drive innovation and contribute to Walmart's success in the retail landscape.