Sales Forecasting and Optimization

Full Documentation

This document provides full documentation for a sales forecasting and optimization project, covering each step from data understanding to Deployment.

This project was developed as part of the Digital Egypt Pioneers Initiative, an initiative affiliated with the Ministry of Communications and Information Technology (MCIT). It was carried out under the academic supervision of AMIT Learning, with ongoing mentorship and follow-up provided by Eng. Abdullah Wagih.

Table of Contents

1. Team Members

This project was successfully completed through the collaboration and dedication of the following team members. Each member contributed based on their specialization, mapped to specific milestones:

Zeyad Tarek – Team Leader – (Milestones 1, 2 & 5)

Directed the overall project vision, coordinated the team's workflow, and ensured timely delivery across all phases. He led the initial data exploration and preprocessing, oversaw advanced data analysis and feature engineering, and compiled the final documentation and presentation materials.

Mostafa Basheer - (Milestones 1, 2 & 3)

Contributed to the effort by collecting, cleaning, and exploring historical sales data, ensuring a strong foundation for the analysis. Afterward, worked on building and developing time-series forecasting models, using various approaches to ensure the models delivered reliable and precise predictions.

Ahmed Khaled - (Milestone 1, 2 & 3)

Explored and applied various data cleaning methods to ensure accuracy and consistency. Contributed to analysis by framing key questions and deriving insights, and helped develop and test diverse forecasting models to maximize performance and predictive power.

Amr Gaber – (Milestone 2)

Designed and implemented a comprehensive dashboard with MS Power BI, transforming raw sales data into an accessible and interactive visual format. This dashboard highlighted trends and patterns, offering key insights for further analysis.

Ahmed Gamal – (Milestone 4)

Developed a seamless deployment solution using Streamlit, creating an interactive interface that provided real-time sales predictions and enabled stakeholders to easily explore and interact with the forecasting model.

Nour Sameh – (Milestone 5)

Took charge of crafting the final presentation, summarizing the project’s methodologies, results, and business implications in an engaging format. Her presentation highlighted the model’s capabilities and provided actionable insights for stakeholders.

2. Define the Project and Objective

The purpose of this project is to forecast daily sales for Rossmann stores using historical data and store-specific features. We aim to:

- Predict future sales for each store up to six weeks ahead.
- Identify key factors influencing sales, such as promotions, holidays, and store types.
- Support better decision-making in inventory planning, staffing, and marketing strategies.

3. Data Presentation

3.1 Data overview

- **train.csv**: Contains historical sales records, including the target variable Sales. ***(1,017,209 entries)***
- **test.csv**: Contains future data points where sales need to be predicted. ***(41,088 entries)***
- **store.csv**: Provides store-specific metadata (e.g., store type, competition, promotions).

(1115 entries)

3.2 Feature Descriptions

A. A. Features from **" train.csv "** and **" test.csv "**

Feature	Description
Id	Unique ID for each row in the test set. *(Only in test.csv)*
Store	Unique identifier for each store.
Date	Date of the record (daily granularity).
Sales	Target variable — total sales for the store on that day. *(Only in train.csv)*
Customers	Number of customers that visited the store. *(Only in train.csv)*
Open	Indicates whether the store was open: 1 = yes, 0 = no.
Promo	Indicates whether a store ran a promo on that day: 1 = yes, 0 = no.
StateHoliday	Type of state holiday: a = public, b = Easter, c = Christmas, 0 = None.
SchoolHoliday	Indicates if the store was affected by school closure: 1 = yes, 0 = no.

B. Feature From **" store.csv "**

Feature	Description
StoreType	Type/category of store: a, b, c, or d.
Assortment	Assortment level: a = basic, b = extra, c = extended.
CompetitionDistance	Distance to the nearest competing store (in meters).
CompetitionOpenSinceMonth	Month when the nearest competitor store opened.
CompetitionOpenSinceYear	Year when the nearest competitor store opened.
Promo2	Indicates if the store is participating in Promo2: 1 = yes, 0 = no.
Promo2SinceWeek	Week number when Promo2 started for the store.
Promo2SinceYear	Year when Promo2 started for the store.
PromoInterval	Months when Promo2 is active (e.g., "Feb,May,Aug,Nov").

3.3 Merging Strategy

The **store.csv** file is merged with both **train.csv** and **test.csv** using the Store column. This ensures that each data point includes both temporal (daily) and static (store-level) features to enrich the forecasting process.

4. Data Wrangling

4.1 Data Cleaning

- **Train Data**
  - All train data features have no missing values.
  - All features' data types are numeric except " **StateHoliday** "has object data type.
  - Closed stores and days on which they didn't have any sales won't be counted in the forecasts.

- **Test Data**
  - There is (11 – Null values) at `**Open**' column,
  - As the store is expected to be open in the test dataset, we replace missing values in the `**Open**` column with `**1**`
- **Store Data**
  - Many Features have missing values as:

```
- CompetitionDistance        >>  3
- CompetitionOpenSinceMonth >> 354
- CompetitionOpenSinceMonth >> 354
- Promo2SinceWeek           >> 544
- Promo2SinceYear           >> 544
- PromoInterval             >> 544
```

  - Apparently this information is simply missing from the data. No particular pattern observed. In this case, it makes a complete sense to replace NaN with the **ZERO**.

4.2 Handling Duplicates

- No duplicate records found.

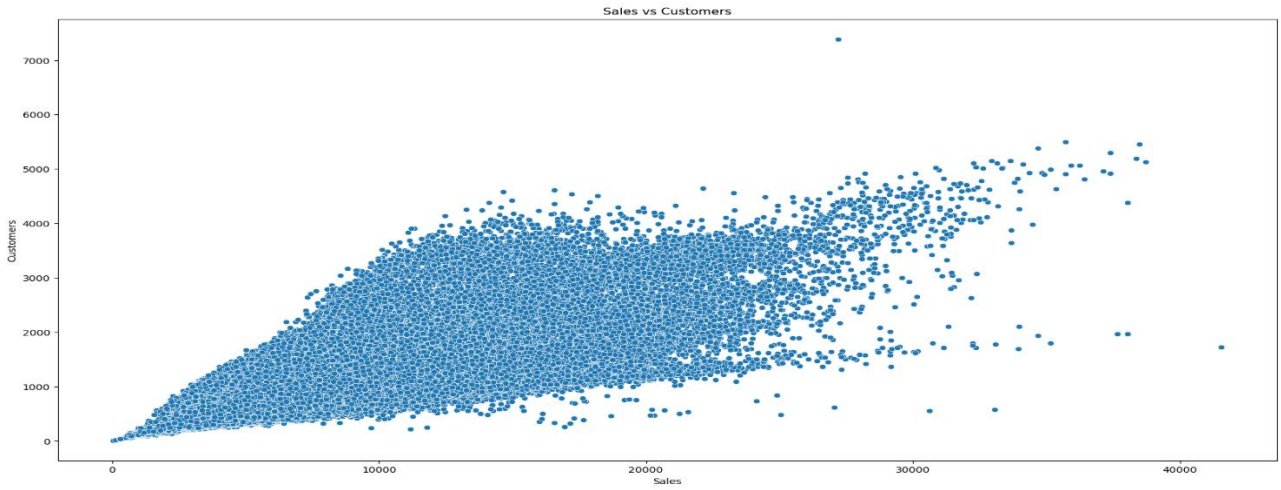
5. Exploratory Data Analysis (EDA)

5.1 Univariate Analysis Summary

- Customers: There is a strong positive correlation between **`Sales`** and **`Customers`**
- Sales: The Sales variable is **right-skewed**, with most transactions concentrated around lower to mid-range values.
- StoreType: The most common store type was StoreType A, followed by B, C, and D and StoreType A had both high frequency and typically higher average sales.
- Promotions:
  - Univariate plots showed that `Promo = 1` days had clearly higher sales and customer volumes.
  - `Promo2 = 1` (continuous promotions) didn't always improve sales, with some evidence suggesting a **decreasing effect** over time (possibly due to diminishing returns).
- Holiday Impact: StateHoliday and SchoolHoliday both showed distinct effects on store traffic and sales. Christmas (c) caused the most noticeable sales drops.
- Stores: We have data of 1115 different stores
- Date: Our data hold stores data in range between **[2013-01-01] => [2015-07-31]**

5.2 Bivariate Analysis Summary

Analyzed how each feature correlates with attrition:



✓ **Sales vs Customers**: Strong positive correlation — more customers consistently led to higher sales.

✓ **Sales vs Promo**: Days with active promotions (Promo = 1) saw significantly higher average sales than non-promo days.

✓ **Sales vs StoreType**: StoreType A stores had the highest average sales, while Type D stores trailed behind.

✓ **Sales vs Assortment**: Stores with extended assortments (Assortment = c) outperformed basic ones (a) in daily sales.

✓ **Sales vs StateHoliday**: Public holidays, especially Christmas (c), led to notable drops in sales.

✓ **Sales vs SchoolHoliday**: Some reduction in sales was observed during school holidays, though the effect was moderate.

✓ **Sales vs CompetitionDistance**: Weak negative correlation — stores closer to competitors had slightly lower sales.

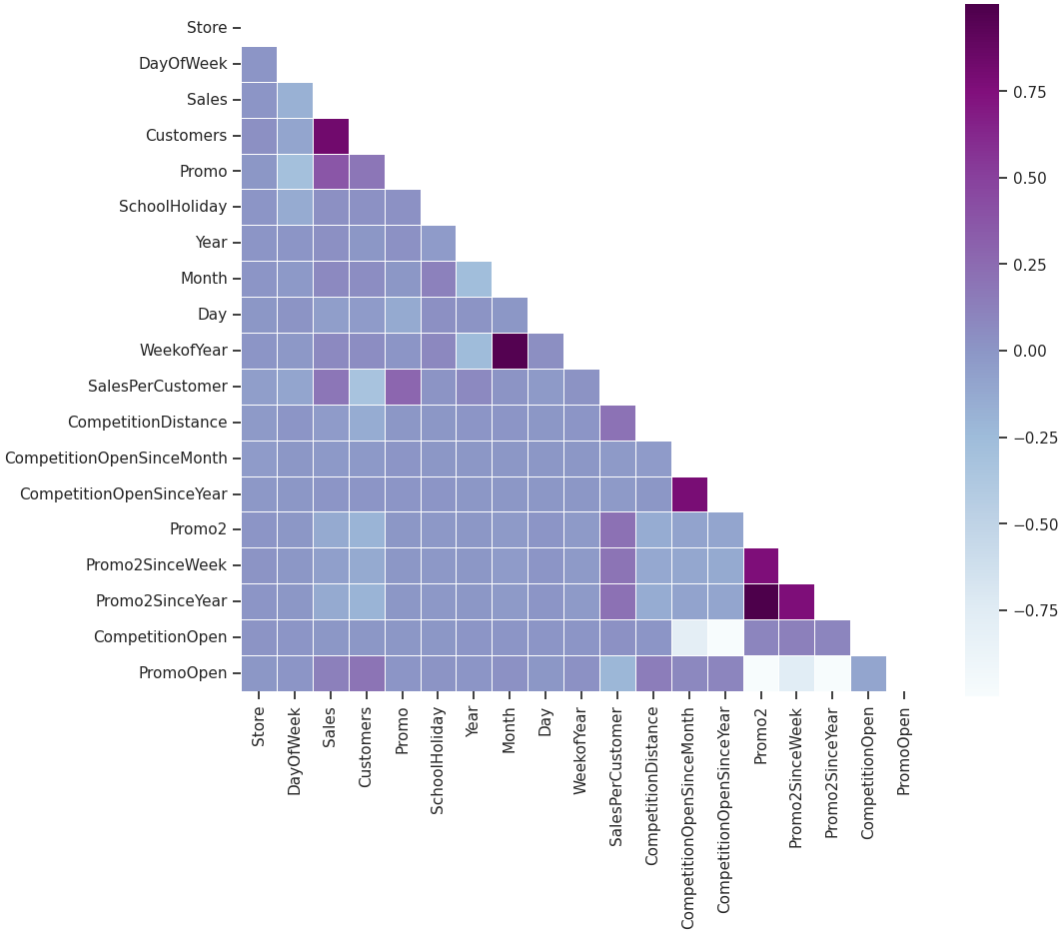
✓ **Sales vs Promo2**: Mixed impact — some stores benefited, but overall long-term promo participation didn't guarantee higher sales.

✓ **Sales vs Day of Week**: Weekends and Mondays typically showed lower sales, while mid-week (especially Fridays) performed better.

✓ **Sales vs Store Open Status**: Open = 0 always resulted in zero sales. Some anomalies were noted where Open = 1 but sales were zero.

5.3 Multivariate Analysis Summary

HEATMAP:



- **Sales vs Customers**: Strong positive correlation — more customers result in higher sales.
- **SalesPerCustomer vs WeekOfYear**: Moderate correlation — suggests seasonality in spending per customer.
- **Promo vs Customers**: Promotions bring more customers into stores.
- **Promo vs Sales**: Sales increase significantly during promotions.
- **CompetitionDistance vs Sales**: Slight negative correlation — stores nearer to competitors tend to earn slightly less.
- **Promo2SinceYear vs Promo2SinceWeek**: Moderate negative correlation — newer Promo2 stores tend to start later in the year.
- **Day, Month, Year vs Sales**: Weak correlations — temporal features may not directly influence sales but are still important for seasonal trends.
- **SchoolHoliday, StateHoliday, Promo2 vs Sales**: Very weak direct correlation — likely due to inconsistent application across stores or time.

6. Machine Learning Model Development

This section outlines the development, evaluation, and optimization of machine learning models to predict employee attrition.

6.1 Introduction

Retail sales forecasting plays a critical role in optimizing inventory, staffing, and marketing strategies within organizations. This report details the process of building a time series forecasting model using Prophet to predict future sales at Rossmann stores based on historical sales data. By accurately predicting sales trends, this model enables businesses to make data-driven decisions, improve operational efficiency, and better plan for fluctuations due to seasonal changes, promotions, and holidays.

6.2 Model of Time Series Forecasting

Selecting the appropriate model for time series forecasting is a critical step in ensuring accurate and reliable predictions. For the task of forecasting sales at Rossmann stores, several models can be evaluated and compared based on their ability to handle key aspects of the time series data, such as seasonality, trend, and external factors like holidays and promotions.

6.2.1 Time Series Characteristics

Time series data often exhibit certain patterns, such as:

- **Trend**: The long-term increase or decrease in the data (e.g., sales growing over time).
- **Seasonality**: Regular, repeating patterns (e.g., weekly, monthly, or yearly fluctuations in sales).
- **Noise**: Random fluctuations that do not follow any discernible pattern.
- **Holidays/Events**: Sales can be influenced by external factors like holidays or promotions.

6.2.2 Used Time Series Models

For the task of forecasting future sales at Rossmann stores, several time series models were evaluated to identify the most accurate method for predicting sales trends. The following time series models were tested:

6.2.2.1 Prophet Model

Prophet was chosen for its ability to model seasonality, holidays, and trends. This model is particularly effective in retail sales forecasting, as it handles multiple seasonalities and external factors (such as promotions and public holidays). Prophet is also user-friendly, allowing for quick adjustments to seasonal effects and holidays, making it a powerful choice for this analysis.

```
MAE: 674.1702896702294
RMSE: 832.6118394962039
MAPE: 14.747619716275343 %
Accuracy: 85.25238028372466 %
```

6.2.2.2 ARIMA Model

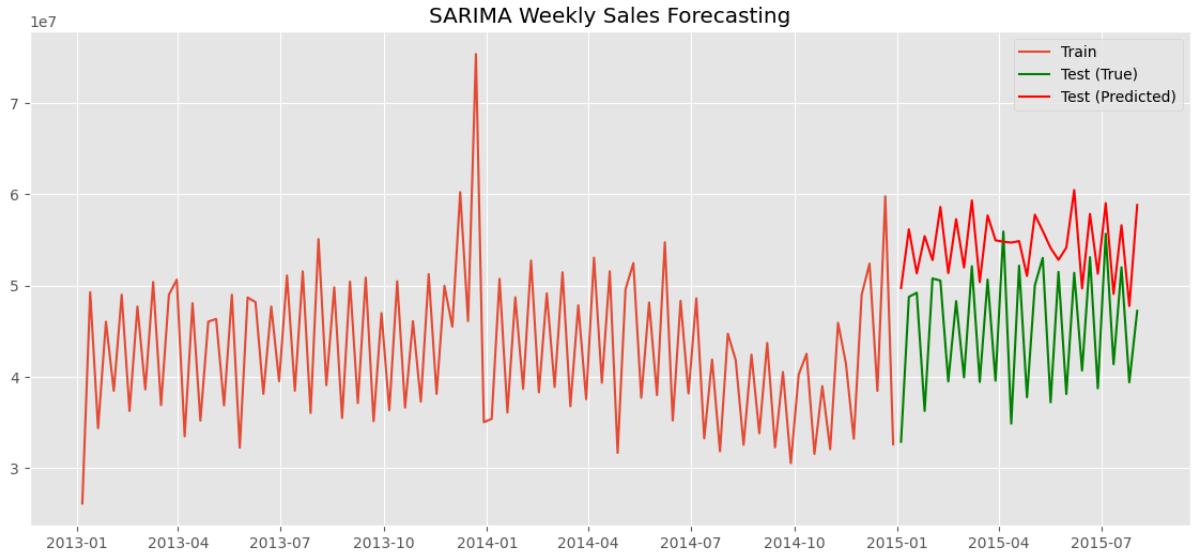
ARIMA is a classic time series forecasting model that captures trends and patterns in the data. However, ARIMA requires the data to be stationary, so it may require additional transformations (like differencing) before fitting. While ARIMA was tested, it was less effective than Prophet in capturing seasonal variations and handling external events like holidays.

```
RMSE: 3069.3578564951713
MAPE: 36.76%
Accuracy: 63.23789580938461 %
```

6.2.2.3 SARIMA Model

SARIMA extends ARIMA by explicitly modeling seasonal components in the data. It is particularly useful when the time series exhibits strong seasonality. SARIMA was evaluated to account for both trend and seasonal components in the sales data, but like ARIMA, it struggled to integrate external factors such as holidays or promotional events.

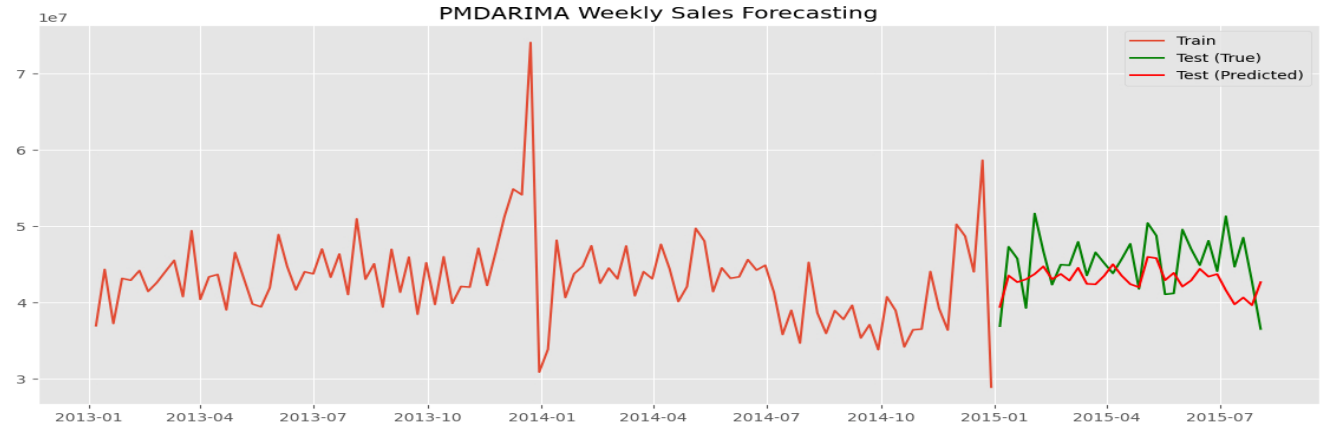
```
RMSE: 10528563.832442906
MAPE: 22.17%
Accuracy: 77.83%
```



6.2.2.4 PMDARIMA Model

PMDARIMA, an extension of ARIMA, is designed for automatic model selection and hyperparameter tuning. It was tested to see if it could improve the ARIMA model's performance by automating the process of selecting the best seasonal parameters and model orders. While it provided useful automatic model suggestions, Prophet outperformed PMDARIMA due to its better handling of complex seasonalities and external influences.

```
Mean Squared Error: 17641129373034.617
MAPE: 7.51%
Accuracy: 92.49%
```



6.2.3 Model Evaluation Criteria

Once the models are trained, they need to be evaluated based on the following metrics:

- **Mean Absolute Error (MAE)**: Measures the average magnitude of errors between predicted and actual values.
- **Root Mean Squared Error (RMSE)**: Highlights larger errors by squaring the differences, giving more weight to larger deviations.
- **Mean Absolute Percentage Error (MAPE)**: Calculates the percentage difference between predicted and actual values, useful for understanding forecast accuracy.
- **R-squared**: Indicates the proportion of variance explained by the model.

6.2.4. Model Selection Process

The model selection process typically involves the following steps:

1. **Exploring the Data**: Identify trends, seasonality, and external factors such as holidays or promotions.
2. **Choosing a Model**: Based on the identified patterns, select an initial set of models. For example, if your data has strong seasonality and holidays, Prophet might be a good candidate.
3. **Feature Engineering**: Create features that represent time-dependent patterns (e.g., lagged values, day of the week, month, etc.) and external factors (e.g., promotions).
4. **Model Training**: Train each model using the historical sales data.
5. **Model Evaluation**: Evaluate the models using appropriate performance metrics (e.g., MAE, RMSE, MAPE).
6. **Model Tuning**: Tune hyperparameters to optimize the model's performance.
7. **Comparison and Final Selection**: Compare the results from different models and select the one that provides the best accuracy and generalization on unseen data.

8. Rossmann Forecasting MLOps & Deployment

This section describes the MLOps pipeline for automating, deploying, and monitoring the employee attrition prediction model, ensuring scalability and reliability.

8.1 Objective

The MLOps implementation aimed to create a reproducible, maintainable, and production-ready machine learning pipeline for predicting Rossman Sales Forecasting. It supports both technical and non-technical users while ensuring continuous performance monitoring.

8.2 Model Deployment (With Streamlit)

8.3 Conclusion

The MLOps pipeline ensures a reliable, scalable, and maintainable system for predicting employee attrition. By integrating tracking, deployment, monitoring, and automated retraining, it delivers long- term value to both developers and business stakeholders.

9. Final Business Recommendations

Based on the insights derived from the data analysis and model performance, the following business actions are recommended to improve sales forecasting accuracy and optimize retail operations:

- **Focus on Seasonal Trends**: The model identified strong seasonal fluctuations in sales. It is crucial to plan marketing, promotions, and stock management around these seasonal peaks and troughs to optimize inventory levels and staffing.
- **Optimize Promotions**: Sales spikes often coincide with promotional periods. By using the sales forecast, businesses can better align promotional activities with expected sales, ensuring inventory is available and staffing is sufficient during high-demand periods.
- **Plan for Public Holidays and Events**: The forecast model reveals that sales are highly sensitive to public holidays and special events. Retailers can leverage this insight by planning for holiday-specific promotions and adjusting stock levels ahead of time.
- **Refine Inventory Management**: Accurate sales predictions enable better inventory management. By forecasting future sales, stores can reduce stockouts and overstock situations, improving product availability and minimizing waste.
- **Enhance Staff Scheduling**: Sales forecasts can be used to optimize staffing levels during peak sales periods. By anticipating high sales days, businesses can ensure they have adequate staff to provide quality customer service and manage the increased traffic.
- **Expand Remote Store Operations**: If the model identifies sales patterns related to location or store types, retailers can optimize the allocation of resources to stores showing high growth potential. Moreover, stores in remote areas could benefit from tailored marketing strategies and stock management.

10. Conclusion and Future Work

This documentation outlines the full cycle of retail sales forecasting — from data preprocessing and model development through to evaluation and deployment. The findings demonstrate that sales are influenced by various factors, including seasonality, promotions, holidays, and store-specific attributes.

The deployed Prophet model, supported by a robust forecasting pipeline, provides valuable insights for decision-making in retail operations. This model helps store managers and business teams predict sales trends more accurately and proactively plan for future periods.

10.1 Future Enhancements:

- **Incorporate Advanced Explainability Tools: Integrating tools like SHAP to explain the model's predictions and improve transparency for decision-makers.**
- **Integration with Retail Systems: Feeding forecasted sales into inventory and staffing management systems (e.g., SAP, Oracle) to automate real-time decision-making.**
- **Include More External Factors: Incorporate additional data such as customer behavior, weather conditions, or competitor activity to enhance prediction accuracy.**
- **Extend the Forecast to Multiple Products: Expanding the model to forecast sales at the product level rather than at the store level could provide even more granular insights.**
- **Predict Engagement and Customer Sentiment: By analyzing customer feedback and engagement data, future models could predict customer sentiment and its potential impact on sales.**

This project lays the foundation for a strategic sales forecasting framework, combining business insights with scalable, data-driven solutions to optimize retail operations and improve profitability.