Table of Contents

Introduction

Provide an overview of the task, explaining the objectives and the importance of accurate time series forecasting in sales data analysis.

Exploratory Data Analysis (EDA)

Data Overview: Describe the dataset, its columns, and their significance.

The dataset contains dummy sales data for various items sold by a well-known brand on Amazon. Each row in the dataset represents the sales data for a particular item on a specific date. The columns included in the dataset are:

date: The date on which the sales data was recorded. This column is essential for time series analysis and forecasting. The format is typically YYYY-MM-DD.

Item Id: A unique identifier for each item. This column helps in distinguishing between different products.

Item Name: The name of the item. While this column is not directly used in numerical analysis, it provides a human-readable reference to the items.

anarix_id: An internal identifier that might be used for tracking or categorizing items within the brand's system.

ad_spend: The amount of money spent on advertising for the item on the given date. This column can be useful in understanding the relationship between advertising spend and sales performance.

units: The number of units sold for the item on the given date. This is the target variable for forecasting in this task.

orderedrevenueamount: The total revenue generated from the units sold on the given date. This can be used to calculate metrics like average unit price and return on ad spend.

unit_price: The price per unit of the item. This column helps in understanding the pricing strategy and its impact on sales.

Data Cleaning: Steps taken to handle missing values, outliers, and any inconsistencies.
Visualization: Use plots to show trends, seasonality, and patterns in sales data over time.
Summary Statistics: Present key statistics like mean, median, standard deviation, etc.

Feature Engineering

Date Features: Extract features like day of the week, month, quarter, etc.
Lag Features: Create lagged versions of the target variable to capture autocorrelation.
Rolling Statistics: Calculate rolling means, standard deviations to capture trends.
Ad Spend Impact: Analyze and create features based on advertising spend.
Revenue-Based Features: Metrics like average unit price and return on ad spend.

Model Selection

For the time series forecasting task, several models were considered, each with its unique advantages and drawbacks. Here's an overview of the models explored:

ARIMA (AutoRegressive Integrated Moving Average)
SARIMA (Seasonal ARIMA)
Prophet
LSTM (Long Short-Term Memory) Neural Network

ARIMA (AutoRegressive Integrated Moving Average)

Description: ARIMA is a widely used statistical method for time series forecasting. It combines autoregression (AR), differencing (I), and moving averages (MA) to model time series data.
Advantages:
- Well-suited for non-seasonal data.
- Simplicity and interpretability.

Drawbacks:
- Requires data to be stationary.
- Limited in handling seasonality.

SARIMA (Seasonal ARIMA)

Description: SARIMA extends ARIMA to handle seasonality in time series data by incorporating seasonal components.
Advantages:
- Captures both trend and seasonality effectively.
- Flexible and interpretable.

- Drawbacks:
  - Computationally intensive due to multiple parameters.
  - Requires careful parameter tuning.

Prophet

- Description: Prophet is an open-source forecasting tool developed by Facebook. It is designed to handle time series data with strong seasonal effects and missing data.
- Advantages:
  - Handles seasonality well with intuitive parameters.
  - Robust to missing data and outliers.
- Drawbacks:
  - Less transparent and interpretable than ARIMA/SARIMA.
  - Requires domain knowledge for optimal tuning.

LSTM (Long Short-Term Memory)

- Description: LSTM is a type of recurrent neural network (RNN) that is capable of learning long-term dependencies in sequence data.
- Advantages:
  - Handles complex patterns and non-linear relationships.
  - Effective for large datasets with multiple features.
- Drawbacks:
  - Requires large amounts of data for training.
  - Computationally expensive and less interpretable.

Evaluation Metrics

Mean Squared Error (MSE) was chosen as the primary evaluation metric for the following reasons:

- Simplicity: MSE is straightforward to compute and understand, providing a clear measure of the average squared differences between predicted and actual values.
- Sensitivity to Large Errors: MSE penalizes larger errors more than smaller ones, making it effective for highlighting significant prediction inaccuracies.
- Continuity: Unlike some metrics, MSE is continuous and differentiable, which is advantageous for optimization algorithms during model training.

Model Choice Justification

Based on the research, SARIMA was chosen as the final model for several reasons:

Interpretability: SARIMA provides clear insights into the components of the time series (trend, seasonality, and noise), making it easier to understand and communicate results.

Performance: During model comparison, SARIMA demonstrated a good balance between capturing seasonal patterns and maintaining predictive accuracy. While LSTM showed promise, it required extensive data and computational resources, which were beyond the scope of this project.
Computational Efficiency: Compared to LSTM, SARIMA is computationally more efficient, which is essential for iterative model development and hyperparameter tuning.
Flexibility: SARIMA's ability to handle both trend and seasonal components made it suitable for the sales data, which exhibited clear seasonal patterns.

By following this detailed process, SARIMA was chosen for its balance of interpretability, performance, and computational efficiency, making it suitable for the sales data forecasting task.

Hyperparameter Tuning

Grid Search/Random Search: Explain the process of hyperparameter tuning used.
Cross-Validation: Describe how cross-validation was employed to ensure model robustness.
Final Model Parameters: Present the best-found parameters and the reasoning behind their selection.

Conclusion

Summarize the findings, the model's performance, and potential future improvements.