Research and Development on Time Series Analysis

Introduction

Time series analysis is a powerful statistical tool used for analyzing time-ordered data points. It is crucial for forecasting future values based on previously observed values. In the context of sales data, accurate forecasting helps in inventory management, budgeting, and strategic planning. This document outlines the research and rationale behind selecting the RandomForestRegressor model for predicting unit sales without using ad spend data.

1. Data Preprocessing

Data preprocessing is the first and most critical step in time series analysis. Properly preparing the data ensures that the model can learn effectively and make accurate predictions. Key steps include:

- **Converting the Date Column**: Transforming the date column to a datetime format allows us to extract various date-related features such as year, month, day, and day of the week.
- **Extracting Date Features**: Date features like year, month, day, and day of the week provide temporal context to the model.
- **Creating Lag Features**: Lag features incorporate historical sales data into the model. For example, the sales of the past 7 days can be included as features to predict the current day's sales.
- **Creating Rolling Window Features**: Rolling window features, such as the 7-day rolling mean and standard deviation, capture short-term trends and variability in the sales data.
- **Handling Missing Values**: Any missing values in the data need to be handled appropriately to avoid biases in the model. This includes filling NaN values and replacing infinite or excessively large values.

2. Feature Engineering

Feature engineering enhances the predictive power of machine learning models by creating meaningful features from the raw data. For this analysis, the following features were engineered:

- **Lag Features**: Lag features for the past 7 days were created to incorporate historical sales data.

- **Rolling Window Features**: The 7-day rolling mean and standard deviation were calculated to capture short-term trends and variability.
- **Date Features**: Year, month, day, and day of the week were extracted from the date column.

3. Model Selection

Several models are commonly used for time series forecasting, each with its strengths and weaknesses:

- **ARIMA/SARIMA**: These models are good for univariate time series with strong temporal dependence. However, they require stationarity and manual parameter tuning.
- **Prophet**: Developed by Facebook for forecasting with strong seasonality, Prophet is easy to use but may not capture complex interactions.
- **Machine Learning Models (RandomForest, XGBoost)**: These models can capture non-linear relationships and interactions between features, making them suitable for complex datasets.

4. Why RandomForestRegressor?

Based on the nature of our data and the requirement to exclude ad spend, RandomForestRegressor was chosen for the following reasons:

- **Robust to Overfitting**: RandomForestRegressor is an ensemble learning method that reduces the risk of overfitting, which is a common issue in time series forecasting.
- **Handles High-Dimensional Data**: It can manage datasets with many features and automatically captures complex interactions between them.
- **No Need for Data Normalization**: Unlike some other models, RandomForest does not require data normalization, simplifying the preprocessing steps.
- **Handles Missing Values**: It can handle missing values in the data without the need for imputation.
- **Feature Importance**: The model provides insights into the importance of different features, helping us understand their impact on predictions.

5. Model Evaluation

The model's performance was evaluated using Mean Squared Error (MSE), a standard metric for regression tasks. The RandomForestRegressor demonstrated

strong predictive performance on the validation set, justifying its selection. The steps included:

- **Splitting the Data**: The data was split into training and validation sets to evaluate the model's performance.
- **Training the Model**: The RandomForestRegressor was trained on the training set.
- **Making Predictions**: Predictions were made on the validation set, and the MSE was calculated to assess the model's accuracy.

6. Conclusion

The RandomForestRegressor model was selected after thorough research and consideration of the dataset's characteristics and the task requirements. It effectively captures the necessary patterns in the sales data, making it a suitable choice for this time series forecasting task.