Research and Development on Time Series Analysis for Sales Forecasting

Objective

Develop a time series forecasting model to predict the number of units sold per item ID using Amazon's dummy sales data.

Step-by-Step Process

1. Understanding the Data

The initial phase involved loading the dataset and performing exploratory data analysis (EDA) to grasp the data's structure and characteristics. Key activities included:

- Checking for missing values.
- Analyzing the distribution and range of various features.
- Converting the 'date' column to a datetime format to extract useful time-based features like year, month, day, and week of the year.

Understanding the data helps identify anomalies and prepares it for subsequent analysis, ensuring the temporal structure is correctly managed for time series forecasting.

2. Data Visualization

Visualization techniques were employed to uncover patterns and trends, such as:

- Plotting monthly and weekly sales distributions.
- Identifying seasonal trends and periodicity.

Visualization helps reveal underlying patterns and seasonal effects, crucial for accurate time series forecasting. It shows how sales vary over time, highlighting any cyclical behaviors.

3. Feature Engineering

Creation of lag and rolling window features:

- Lag Features: Sales from previous days, weeks, or months to help the model learn from past values.
- Rolling Mean Features: 7-day and 30-day rolling averages to smooth out short-term fluctuations and highlight long-term trends.

These features capture temporal dependencies and trends essential for time series forecasting. Lag features allow the model to use historical sales data to predict future sales, while rolling means provide a smoothed version of the sales data.

4. Model Selection

Various models were considered for the forecasting task:

- **Linear Regression:** Assumes a linear relationship between input features and the target variable.
- **Ridge Regression:** Adds a regularization term to prevent overfitting.
- **Lasso Regression**: Performs feature selection by driving some coefficients to zero.
- **Gradient Boosting Machines (GBM):** An ensemble technique that builds multiple decision trees sequentially to enhance predictive performance.

Linear Regression serves as a baseline to understand linear relationships in the data. Ridge and Lasso Regression introduce regularization to handle multicollinearity and overfitting. Gradient Boosting Machines can capture complex relationships and interactions between features, making it highly accurate for various machine learning tasks.

5. Training and Evaluation

Models were trained on the training dataset and evaluated using the validation dataset, with the primary evaluation metric being the Mean Squared Error (MSE).

Results:

- Linear Regression MSE: 967.20
- Ridge Regression MSE: 967.15
- Lasso Regression MSE: 942.93
- Gradient Boosting Machines MSE: 516.63

6. Model Comparison

MSE scores of all models were compared after training. Gradient Boosting Machines significantly outperformed the others.

The superior performance of Gradient Boosting Machines is due to its ability to capture complex patterns and interactions in the data, which simpler models might miss, making it a robust choice for time series forecasting.

7. Hyperparameter Tuning

Hyperparameter tuning was performed for the best-performing model (GBM) to optimize its performance. This involved adjusting parameters such as the number of estimators, learning rate, and maximum depth of trees.

Results:

- Best Parameters for GBM:
- Learning rate: 0.1
- Maximum depth: 3
- Number of estimators: 200

Optimizing these parameters ensures that the model generalizes well to unseen data, thereby improving its predictive accuracy.

8. Final Model Training

Using the best hyperparameters identified, the GBM model was retrained on the entire training dataset.

Retraining the model with optimized parameters on the full dataset ensures it leverages all available data for learning, leading to better generalization and accuracy in predictions.

9. Prediction on Test Data

The final model was used to make predictions on the test dataset, with results saved for submission.

This step ensures that the model's performance is evaluated on unseen data, providing a realistic measure of its predictive capabilities. The predictions are then formatted for submission as required.

Conclusion

Through systematic EDA, feature engineering, and rigorous model selection and tuning, Gradient Boosting Machines were identified as the best model for the time series forecasting task. Its ability to capture complex relationships and robustness against overfitting made it the ideal choice for predicting sales data. The structured approach ensured that the model was optimized for accuracy and reliability in forecasting future sales trends. The final model, tuned with optimal hyperparameters, demonstrated superior performance and predictive accuracy, making it highly suitable for real-world forecasting applications.