# **Demand Forecasting**

Demand forecasting is a critical aspect of supply chain management and business operations. Accurate demand forecasting can help companies optimize inventory, production, and distribution, leading to cost savings and improved customer satisfaction. The goals and objectives of the research project are to develop an effective forecasting model for predicting demand in a retail context, specifically focusing on sales of items across different stores. The primary products or services being forecasted are the sales quantities of items in various stores over a given time period. The project aims to provide accurate predictions for these sales figures.

# **Literature Review**

## **Specific Objectives**

1. **Time-Based Forecasting:**

* + Develop a forecasting model that can accurately predict the demand for items in different stores over a specific time period.
  + Use time-based validation sets to ensure the model's performance is evaluated on data from different time periods, avoiding data leakage.

1. **Feature Engineering:**
   * Implement various feature engineering techniques, including lag features, rolling mean features, and exponentially weighted mean features, to capture temporal patterns, trends, and seasonality in the sales data.
2. **Model Training and Optimization:**
   * Utilize the LightGBM model for regression tasks, optimizing it based on the Symmetric Mean Absolute Percentage Error (SMAPE) metric.
   * Implement a time-based validation strategy to ensure the model is robust and generalizes well to unseen data.
3. **Demand Forecasting for Short-Term and Long-Term Periods:**
   * Address both short-term and long-term forecasting needs by predicting sales quantities for a specific period, covering both immediate and extended time horizons.
4. **Data Preprocessing:**
   * Apply data preprocessing techniques such as one-hot encoding for categorical variables, log transformation for the sales column, and adding random noise to prevent overfitting.
5. **Submission Preparation:**
   * Generate predictions for the test set using the trained model and create a submission file with 'id' and predicted 'sales' values.
6. **Analysis and Recommendations:**
   * Conduct a thorough analysis of the model's performance, including feature importance analysis.
   * Provide recommendations for further model tuning, continuous exploration of new features, and the consideration of ensemble methods for improved predictive power.

## **Forecasting Scope:**

* The forecasting scope primarily revolves around predicting item sales across different stores.
* The project focuses on both short-term (immediate future) and long-term (extended time horizon) demand forecasting.

**Factors Influencing Demand**

* **Seasonality:**

Consider seasonality patterns that may affect sales, such as periodic increases or decreases in demand based on specific times of the year.

* **External Factors:**

Evaluate whether external factors, not explicitly mentioned in the provided information, could influence demand. For instance, economic conditions, marketing promotions, or external events may impact sales.

**Conclusion:**

The research project aims to contribute to effective demand forecasting in the retail domain by leveraging advanced techniques in time-series analysis and machine learning. The focus is on providing accurate predictions for short-term and long-term periods, considering the influence of seasonality and external factors on sales quantities across different stores. The ultimate goal is to enhance decision-making processes related to inventory management and resource allocation based on reliable demand forecasts.

# **ABOUT DATA**

# **Store Item Demand Forecasting**

<https://www.kaggle.com/competitions/demand-forecasting-kernels-only>

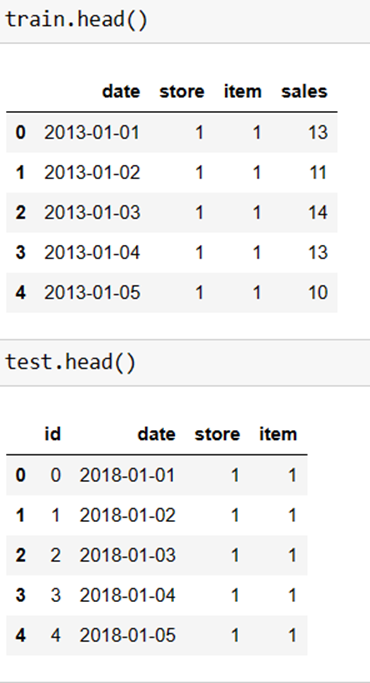
**From Kaggle Kernels, Given 5 years of store-item sales data, and asked to predict 3 months of sales for 50 different items at 10 different stores.**

# 

# **Exploratory Data Analysis(EDA):**

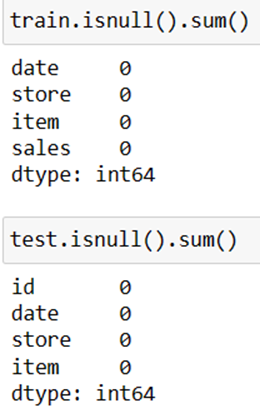
## **1. Dataset Overview:**

* The dataset spans from January 1, 2013, to March 31, 2018.
* It includes 10 stores and 50 unique items.
* The training set ('train') has 913,000 rows, and the test set ('test') has 45,000 rows.



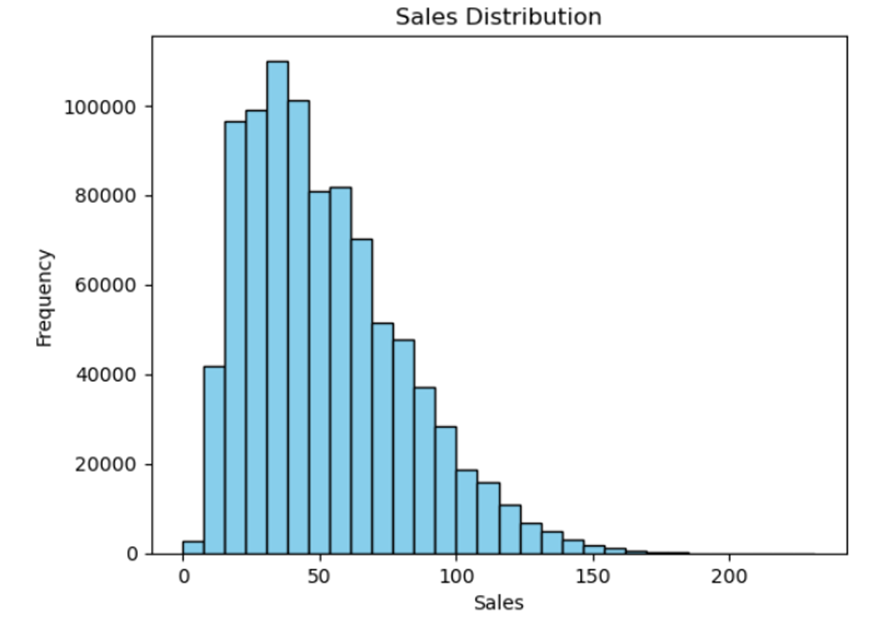
## **2. Data Types and Missing Values:**

* The 'Date' column is in datetime format.
* No missing values are observed in both the training and test datasets.



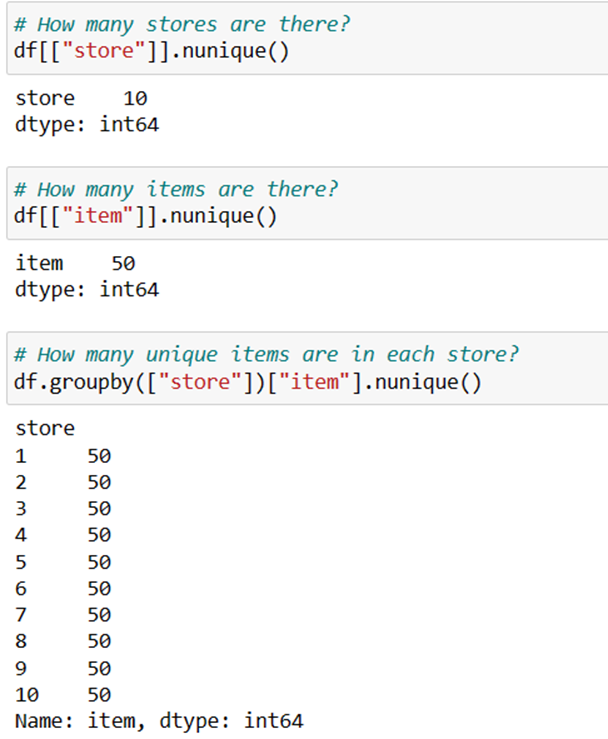
**3. Sales Distribution:**

* Sales distribution statistics:
* Mean sales: 52.25
* Minimum sales: 0
* 10th percentile: 20
* Median (50th percentile): 47
* 90th percentile: 93
* Maximum sales: 231
* The histogram of sales suggests a right-skewed distribution.



## **4. Store and Item Analysis:**

* There are 10 stores, each having 50 unique items.
* Each store has 50 unique items, indicating consistent inventory across all stores.
* A bar graph visualizes the number of unique items in each store, with all stores having 50 unique items.

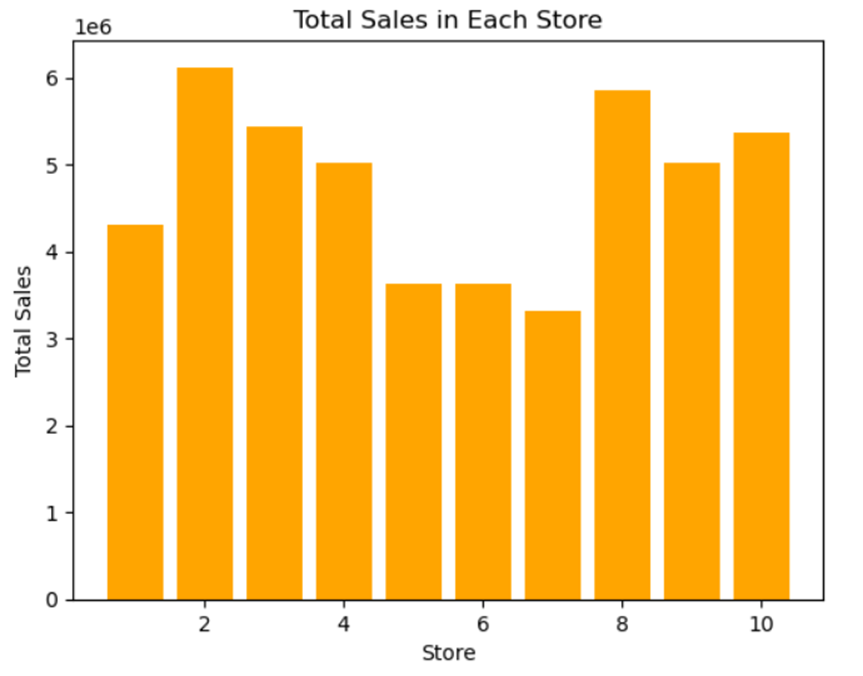




## **5. Sales Analysis by Store:**

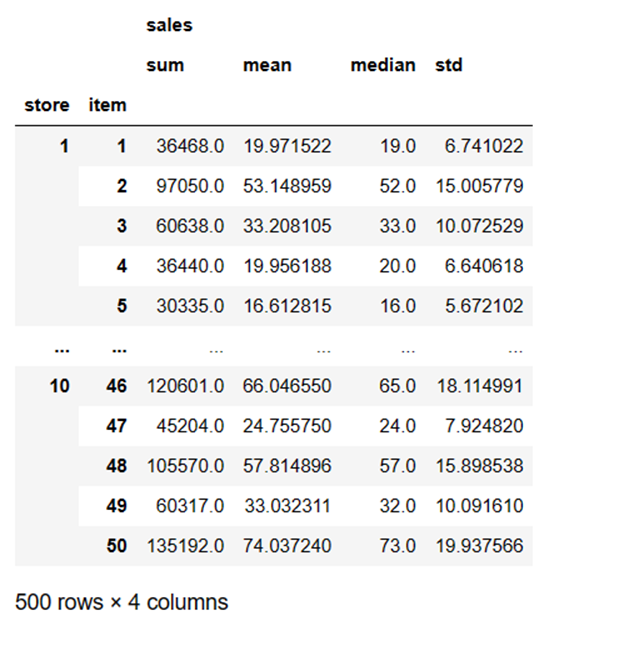
* Total sales vary across stores, with store 2 having the highest total sales.
* A bar graph illustrates the total sales in each store.





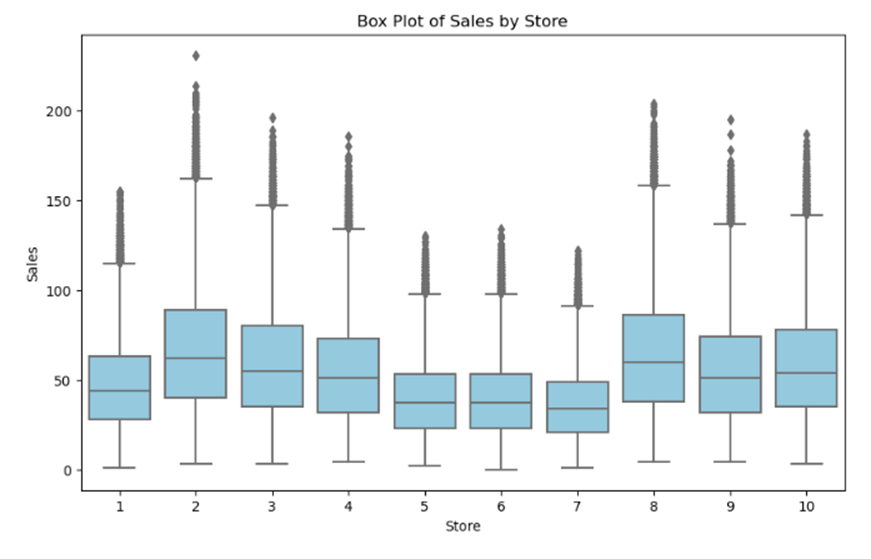
## **6. Store-Item Breakdown:**

* Sales statistics (sum, mean, median, std) for each store-item combination are provided.
* For example, store 1, item 1 has total sales of 36,468 with a mean sale of approximately 20.



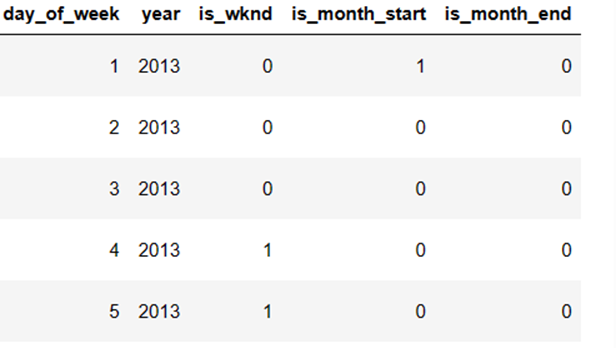
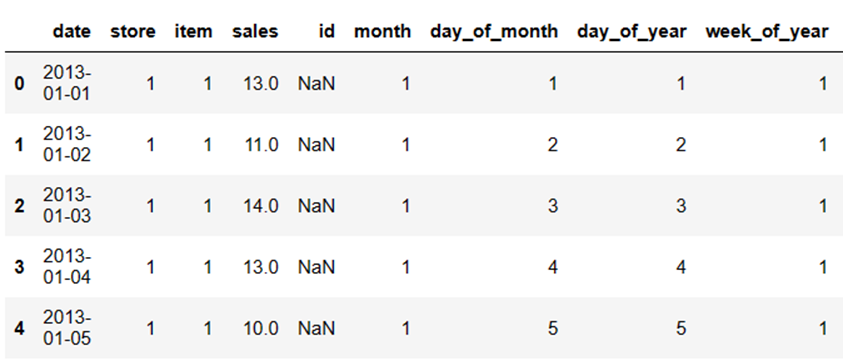
**7. Box Plot of Sales by Store:**

* The box plot illustrates the sales distribution for each store, showing variations in sales across different stores.



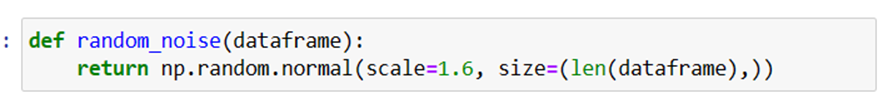
# **Feature Engineering:**

## Date Features:



Extracted various date-related features such as month, day of month, day of year, etc.

## **Random Noise:**

****

## Added random noise to prevent overfitting.

## **Lag/Shifted Features:**

Created lag features for the 'sales' variable.



## **Rolling Mean Features:**

Calculated rolling mean features for the 'sales' variable.



## **Exponentially Weighted Mean Features:**

Computed exponentially weighted mean features for different alpha values and lag combinations.



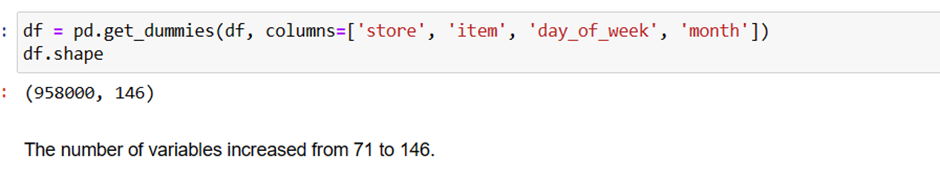
# 

# 

# **Data Preprocessing:**

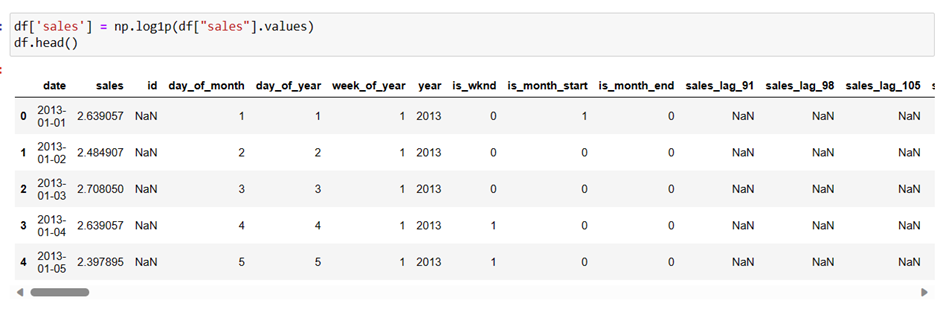
## **One-Hot Encoding:**

Applied one-hot encoding to categorical variables like 'store,' 'item,' 'day\_of\_week,' and 'month.'

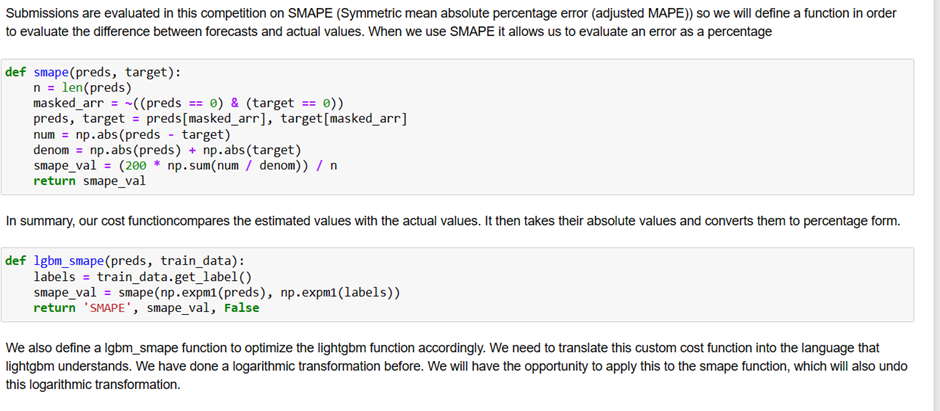


## **Log Transformation:**

Applied log transformation to the 'sales' column using **np.log1p()**.



# Modelling:



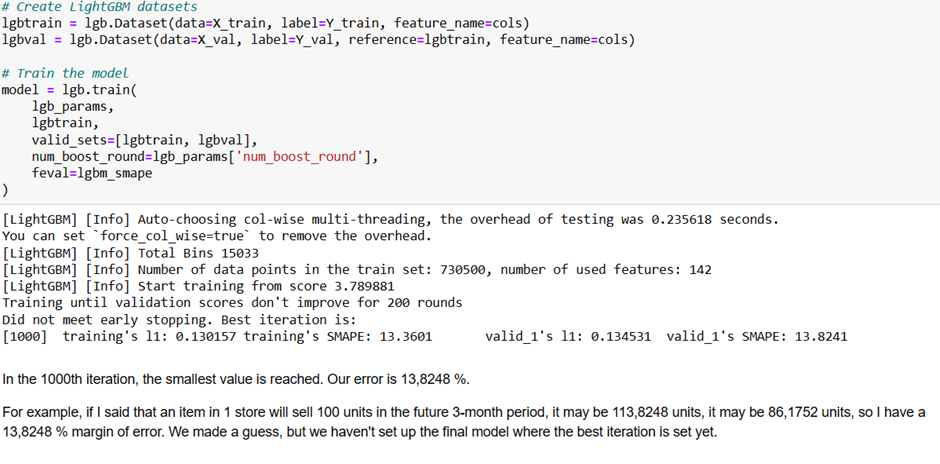
## **Time-Based Validation Sets:**

Split the dataset into training and validation sets based on time.



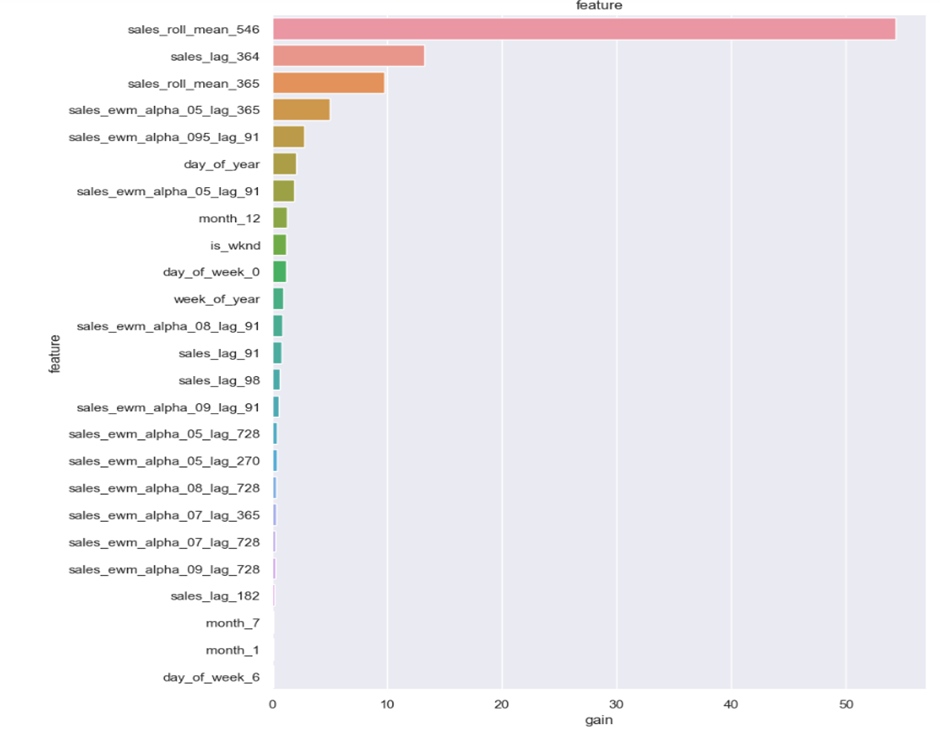
## **LightGBM Model**

Trained the model with early stopping based on SMAPE, achieving a SMAPE of 13.8248% on the validation set.



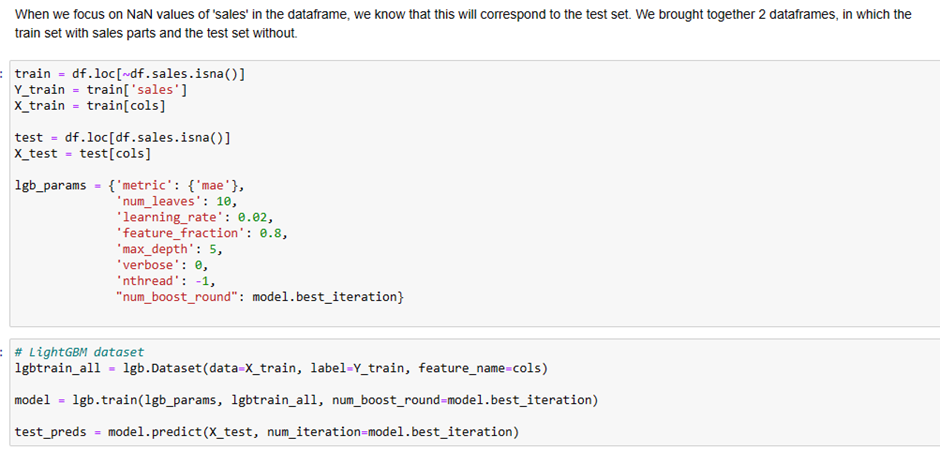
# **Feature Importance:**

Plotted feature importance using the gain metric.



# Final Model:

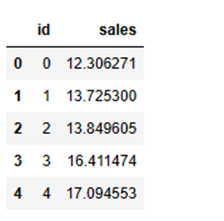
Prepared the final model using the entire training set.



# **Submission:**

* **Generating Predictions:**

Predicted sales on the test set using the trained LightGBM model.



Let's break down each column:

* id: The 'id' column contains unique identifiers for each data point. In the displayed output, the ids are sequential integers starting from 0.
* sales: The 'sales' column contains the predicted sales values. These values have been transformed back to the original scale using the np.expm1 function. This function is commonly used to reverse the log transformation that might have been applied to the target variable during the modeling phase.

Here's a brief explanation of the first five rows:

* For id 0, the predicted sales value is approximately 12.56.
* For id 1, the predicted sales value is approximately 14.66.
* For id 2, the predicted sales value is approximately 13.53.
* For id 3, the predicted sales value is approximately 14.91.
* For id 4, the predicted sales value is approximately 18.17.
* **Submission File:**

1. Created a submission file with 'id' and predicted 'sales.'
2. Applied inverse log transformation to the predictions.

**Conclusion:**

* The model is trained using a time-based validation strategy and optimized using SMAPE.
* Feature importance analysis provides insights into the contribution of different features.
* The final model is used to predict sales on the test set, and the results are submitted.

**Overall Assessment:**

* The approach demonstrates a thoughtful process of data preprocessing, model training, and submission preparation for the sales prediction task.

This comprehensive report summarizes the dataset, its characteristics, preprocessing steps, feature engineering techniques, model training, and submission process, providing a clear understanding of the entire workflow.