Detailed Report on the Code for Beverage Sales Forecasting

1. Introduction:

The code aims to address the challenge of forecasting beverage sales for a company operating in Australia. The dataset contains information on various products, sales, promotions, and external factors such as holidays and seasonality. The primary objectives include building accurate forecasting models using machine learning and demonstrating the effectiveness of the models through thorough evaluation.

2. Data Loading and Preliminary Analysis:

- The code begins by importing necessary libraries and loading the dataset from a CSV file.
- Preliminary analysis includes inspecting the head of the dataset and obtaining information about data types and non-null counts.

3. Data Cleaning and Transformation:

a. **Data Type Correction:**

- The 'Date' column is converted to the datetime format for consistency and ease of handling.

b. **Handling Missing Values:**

- Rows with missing values are dropped to ensure a complete dataset.

c. **Outlier Handling:**

- Sales data points above 100,000 are considered outliers and removed.

d. **Boolean Column Removal:**

- Boolean columns ('V_DAY,' 'EASTER,' 'CHRISTMAS') with limited occurrences are removed.

e. **Duplicate Entry Removal:**

- Duplicate entries based on a subset of columns are dropped to ensure data integrity.

f. **Product Code Transformation:**

- 'Product' codes are converted to integer format for numerical operations.

g. **Feature Engineering:**

- New time-related features (hour, dayofweek, quarter, month, year, dayofyear) are created.

4. Exploratory Data Analysis (EDA):

- The EDA process involves visual inspection of data, histogram and box plot analysis for 'Sales' distribution, correlation matrix visualization, boolean column analysis, and examination of product codes.

5. Data Separation for Individual Products:

- Recognizing potential differences in sales patterns among products, the code separates data for individual products. This step improves the granularity of analysis and model training.

6. Model Training and Testing:

- The code defines a set of features and a target variable for model training and testing.
- It splits the data into training and testing sets based on a specific timeframe.

7. Model Selection:

- Four machine learning algorithms (Linear Regression, Decision Trees, Random Forest, XGBoost) are chosen for forecasting.

8. Visualizations:

- The code includes visualizations to plot training and testing data, providing insights into sales trends.

9. Conclusion and Recommendations:

- The code concludes with an explanation of the challenges in the initial model performance and the solution found by separating products for individual training. It emphasizes the importance of understanding the context and characteristics of the data for effective forecasting.

10. Overall Assessment:

- The code demonstrates a systematic approach to data preprocessing, feature engineering, and model training for accurate beverage sales forecasting.
- EDA visualizations enhance the understanding of data distribution and relationships between variables.
- The inclusion of multiple machine learning algorithms allows for a comparative assessment of model performance.

Recommendations for Improvement:

- Further hyperparameter tuning and optimization of the selected machine learning models.
- Consideration of additional features or external data sources to enhance model accuracy.
- Evaluation of model performance on a broader timeframe for comprehensive insights.

This detailed report summarizes the code's key steps, from data loading and cleaning to model training and evaluation. It underscores the significance of each phase in achieving accurate and meaningful forecasts for beverage sales.