Project: Data Science : Retail Forecasting

Name: Siddharth Dudugu (siddharthdudugu@gmail.com)

Country: USA

College: Drexel University

Specialization: Data Science

Contents:

Problem description:

The beverage company in Australia faces a critical challenge in accurately forecasting the demand for its diverse range of products sold through supermarkets, compounded by frequent and substantial promotional activities throughout the year. The necessity for precise item-level forecasts on a weekly basis is further complicated by the influence of various factors such as holidays and seasonality on product demand. Currently utilizing an in-house forecasting solution, the company encounters issues with the sensibility of the predictions generated. In a strategic move, the company aims to explore the potential of AI and machine learning techniques to replace its existing software. The objective is to develop 4-5 multivariate forecasting models, incorporating advanced methodologies such as machine learning or deep learning. Accuracy is measured through the Weighted MAPE metric, with a focus on demonstrating forecast accuracy for the crucial time period of Q3-Q4 2020. The implementation emphasizes feature engineering to extract additional variables for enhanced accuracy, code optimization for speed, and the incorporation of explainability in the form of variable contributions to foster a deeper understanding of the forecasting models.

Business understanding:

Business understanding, in the context of the beverage company's challenge, involves a comprehensive grasp of the Australian beverage industry dynamics, the diverse product portfolio, and the impact of promotions on consumer behavior. Recognizing external factors like holidays and seasonality, understanding the limitations of the current forecasting system, and strategically focusing on Q3-Q4 2020 are crucial aspects. The aim is to align forecasting models with operational needs, considering inventory management and production planning as key components.

Data Intake Report

Name: Retail Forcasting

Report date: 19-12-2023

Internship Batch: LISUM 27

Version:<1.0>

Data intake by: Siddharth Dudugu

Data intake reviewer: -

Data storage location: GitHub:

Tabular data details:

Total number of observations	1217
Total number of files	1
Total number of features	12
Base format of the file	ipynb
Size of the data	-

Proposed Approach:

1. **Data Exploration and Cleaning:**
    - Correct data types and handle missing entries.
    - Visualize and analyze key variables.
2. **Feature Engineering:**
    - Create time-related features.
    - Separate data for individual products.
3. **Correlation Analysis:**
    - Identify and remove redundant features.

4. **Data Separation:**
    - Recognize diversity among products.
    - Separate data for each product.
5. **Model Selection and Training:**
    - Choose ML algorithms.
    - Split data into training and testing sets.
6. **Evaluation and Visualization:**
    - Assess model performance using visualizations.

Assumptions Made:

1. Products may have unique sales patterns.
2. Sales data is measured weekly.
3. Certain boolean columns can be removed.
4. Outliers in sales data above 100,000 can be discarded.
5. 'SKU' codes can be represented as integers.
6. Date column needs consistent formatting.
7. Evaluation focuses on Q3-Q4 2020.
8. Multiple ML algorithms will be used for forecasting.

Week 8

Data Understanding:

1. **Type of Data:**
    - The data appears to be related to beverage sales, with columns such as 'Product', 'Date', 'Sales', 'Price Discount (%)', and several promotional indicators. It is likely a time-series dataset capturing weekly sales data for different products.
2. **Problems in the Data:**
    - **Data Types:** The 'Date' and 'Price Discount (%)' columns have incorrect data types.
    - **Outliers:** There are potential outliers in the 'Sales' column.
    - **Skewed Distribution:** The 'Sales' distribution is highly skewed.
    - **Boolean Columns:** Certain boolean columns ('V_DAY', 'EASTER', 'CHRISTMAS') have limited occurrences.

Approaches to Address Data Issues:

- Data Types:

Approach: Convert the 'Date' column to datetime format for consistency.

Why: Ensures proper handling of date-related operations and analyses.

- Outliers and Skewed Distribution:

Approach: Identify and handle outliers, potentially by removing extreme values.

Why: Outliers can impact model performance and the skewed distribution may affect the accuracy of summary statistics and models.

- Boolean Columns:

Approach: Remove boolean columns ('V_DAY', 'EASTER', 'CHRISTMAS') with limited occurrences.

Why: These columns may not contribute significantly to the analysis due to their low frequency.

- Missing Values:

Approach: No explicit mention of missing values, but if present, addressing them using imputation or removal.

Why: Missing values can affect the accuracy of analyses and models.

- Product-Specific Challenges:

Approach: Separate data for individual products (e.g., SKU1) for more accurate modeling.

Why: Different products may exhibit unique sales patterns, and modeling them separately can improve accuracy.

- Feature Engineering:

Approach: Create additional time-related features (hour, dayofweek, quarter, month, year, dayofyear).

Why: Enhances the dataset with meaningful features that capture temporal patterns.

- Correlation Analysis:

Approach: Identify and remove redundant features based on correlation matrices.

Why: Redundant features may not contribute significantly to the analysis and can be removed for efficiency.

- Data Cleaning:

Approach: Handle duplicate entries based on a subset of columns.

Why: Duplicate entries can distort analyses, and removal ensures data integrity.

The overall approach involves a combination of data type correction, outlier handling, removal of less informative features, addressing skewed distributions, and creating additional features to enhance the dataset for accurate forecasting. The rationale behind these approaches is to ensure the quality and relevance of the data for subsequent analyses and model development.