Data Science : Retail Forecasting Group Project

Group Name: Data Solver

Member 1:
Name: Yiu Hung Leong
Email: ernest9773@gmail.com
Country: Malaysia
College/Company: Monash University
Specialization: Data Scientist

Member 2:
Name: Oladimeji Adaramoye
Email: adaramoyeoladimeji@gmail.com
Country: United Kingdom
College/Company: Roehampton University
Specialization: Data Science

Member 3:
Name: Roshan Gurung
Email: gurung999roshan@gmail.com
Country: United Kingdom
College/Company: Solent University
Specialization: Data Science

Github repo link: Data-Glacier-Internship/Group Project/Week 8 at main · ErnestCodeHub/Data-Glacier-Internship · GitHub

Problem description: Forecasting Beverage Demand

A large company in the beverage industry in Australia sells their products through various supermarkets and heavily invests in promotions throughout the year. They face the challenge of accurately forecasting the demand for each product at the item level on a weekly basis. This forecast is crucial for effective inventory management and production planning.

The company's demand patterns exhibit various complexities, including trends, seasonality, and non-linear patterns. As a result, we will build 4-5 multivariate forecasting models, including ML or deep learning-based models, using PySpark and parallel computing techniques and demonstrate best-in-class forecast accuracy by achieving a low weighted Mean Absolute Percentage Error (MAPE). Lastly, we will provide explainability of the models by showcasing the contribution of each variable in the forecasting process.

Data Understanding

The provided data is part of a retail forecasting group project focused on predicting beverage demand. This data comprises historical demand for beverage products sold by a company in Australia through various supermarkets. The primary goal is to build accurate forecasting models that can predict future demand. This predictive capability is crucial for the company’s effective inventory management and production planning. Understanding historical sales trends and external influencing factors is vital for developing reliable models that can help the company optimize its operations and ensure timely supply of products to meet consumer demand.

Type of Data for Analysis

The dataset encompasses various types of data essential for comprehensive analysis. It includes historical sales data, which provides insights into past demand patterns and helps in identifying trends and seasonality. Additionally, the dataset incorporates external factors such as promotional activities, holidays, and weather data. These factors are crucial as they significantly impact consumer purchasing behavior and can cause fluctuations in demand. By integrating these external variables, the forecasting models can account for these influences and improve their predictive accuracy.

Problems in the Data

Upon examining the dataset, it was found that there were no duplicates or missing values that could have directly affected the accuracy of the models. However, the dataset still presented some significant challenges. One of the main issues was the presence of outliers, which are extreme values that do not conform to the expected range. These outliers can skew the results and lead to inaccurate predictions if not properly managed. Additionally, the data exhibited skewness, indicating that the data distribution was not normal. Skewed data can introduce bias into the models, making it difficult to achieve accurate forecasts.

Approaches to Overcome Data Problems

To address these data issues, several methods can be applied. For managing outliers, the first step is detection, which can be done using visual tools such as box plots. Once identified, outliers can be treated either by removing them from the dataset or by transforming them to minimize their impact. Addressing skewed data involves applying transformations to normalize the data distribution. Common techniques include log transformation, square root transformation, or Box-Cox transformation. These transformations help to stabilize variance and make the data more suitable for modeling, thereby improving the overall accuracy and reliability of the forecasting models.