Data Science Project Report (Week 9)

Number of team members: 1
Member Name: Yan-Ping Yu
Email: yy3407@columbia.edu
Nationality: Taiwan
College: Columbia University
Specialization: Data Science

1. Problem description:

Our client operates within the beverage industry in Australia. The primary challenge is to develop multivariate forecasting models, utilizing machine learning or deep learning techniques, to accurately predict weekly demand for their products. This initiative is driven by the need to analyze historical time series data, incorporating various factors that influence demand, to forecast the quantity of items required by customers each week.

2. Github Repo link

https://github.com/YanPing0227/yanping-my-repo

3. Data cleansing and transformation done on the data.

I began by loading the dataset into a pandas DataFrame to conduct an initial assessment. I checked for missing values and found none, which meant I could proceed without having to handle NA values.

Next, I converted the 'date' column to a datetime format for proper time series analysis. For the 'Price Discount (%)' column, I stripped the percentage signs and converted the strings to float values for numerical operations.

To ensure data integrity, I checked for duplicate rows but didn't find any. This allowed me to proceed without needing to remove duplicates, which is an important step in data cleansing to avoid skewed results.

Afterward, I conducted a descriptive analysis by dropping irrelevant columns like 'Product' and 'date' for this specific statistical summary. This gave me insights into the distribution of the numerical variables.

I defined a function to identify outliers using the z-score method, checking if any data points were beyond three standard deviations from the mean. I found outliers, particularly in the 'Google_Mobility' data during April 2020, which correlated with the onset of COVID-19 lockdowns. Since these outliers were due to an external event and not random anomalies, I decided to keep them, acknowledging their impact on mobility during the pandemic.

Moving on to the skewness, I found that the 'Sales' and 'Google_Mobility' data were significantly skewed. I applied log transformations to these columns to reduce skewness and normalize their distribution, which is beneficial for many statistical and machine learning models.

For the 'Sales' data, I added 1 before taking the logarithm to handle zero sales values. For 'Google_Mobility', I reflected the data (multiplied by -1) to handle negative values before the log transformation. This reflection-log technique is a common approach to handle negative skewness.

Finally, I standardized the log-transformed 'Sales' and 'Google_Mobility' data to have a mean of 0 and a standard deviation of 1, which is necessary for models that assume normally distributed input. I also normalized the 'Price Discount (%)' to bring it into the range of [0, 1], which is suitable for models that are sensitive to the scale of the input data.

Throughout this process, I took care to only apply transformations that made sense given the context of the data and the analysis goals. I documented each step carefully, ensuring transparency and reproducibility of the data cleansing and transformation process.

I visualized the distributions of the transformed variables using histograms and density plots, confirming the success of the transformations and the preparedness of the data for further analysis or modeling. This methodical approach ensured that the data was clean, relevant, and formatted correctly for the next stages of the data science project.