Week 8 Deliverables

Group Name: Data Forecasting Team

Team Member's Details

Name: Charles Ibitamuno

Email: charlesibitamuno@gmail.com

Country: United States of America

College/Company: Columbia University of General Studies

Specialization: Data Science: Retail Forecasting

Problem Description:

The large beverage company in Australia needs to forecast demand for each of their products at the item level, on a weekly basis. Their sales are influenced by various factors including promotions, holidays, and seasonality. The company currently uses an in-house software solution for forecasting, but it often produces unreliable results. They want to explore AI/ML-based forecasting to replace their current system.

Data Understanding:

The dataset we are analyzing is the "forecasting_case_study.xlsx" file, which contains 1218 rows and 12 features, comprising of "Product", "date", "Sales", "Price Discount (%) ", "In-Store Promo", "Catalogue Promo", "Store End Promo", "Google_Mobility", "Covid_Flag", "V_DAY", "EASTER" and "CHRISTMAS". Their features and datatype are listed below:

The dataset contains the following columns:

1. **Product**: SKU identifier (SKU1, SKU2, etc.) **Dtype: object**
2. **Date**: Weekly dates from 2017 to 2020 **Dtype: object**
3. **Sales**: Numeric sales figures **Dtype: int64**
4. **Price Discount (%):** Percentage of price discount offered **Dtype: object**
5. **In-Store Promo**: Binary (0 or 1) indicating in-store promotions **Dtype: int64**
6. **Catalogue Promo**: Binary (0 or 1) indicating catalogue promotions **Dtype: int64**
7. **Store End Promo**: Binary (0 or 1) indicating store end promotions **Dtype: int64**
8. **Google_Mobility**: Numeric values (possibly related to Google's mobility reports during COVID-19) **Dtype: float64**
9. **Covid_Flag**: Binary (0 or 1) indicating COVID-19 period **Dtype: int64**
10. **V_DAY**: Binary (0 or 1) possibly indicating Valentine's Day **Dtype: int64**
11. **EASTER**: Binary (0 or 1) indicating Easter holiday **Dtype: int64**
12. **CHRISTMAS**: Binary (0 or 1) indicating Christmas holiday **Dtype: int64**

Type of Data Analysis:

Categorical Data: Product, Binary promotion indicators and holiday flags

Numeric Data: Sales, Price Discount, Google_Mobility

For our Time Series Data, we use the Weekly Sales between the designated dates given in the file.

Problems in The Data:

1. There were no missing values in the file
2. There might be outliers in the Sales column, with some weeks showing significantly higher sales than others. This might be because of the seasonal holiday or promotions that occurred within that time.
3. Most of the Products are Right Skewed, for reasons I've not yet finalized
4. There's a lot of inconsistencies with the Data for example, when we compared the in-store, catalog, and store end promo and grouped them based on certain patterns, we saw there were 384 instances that there were no promotions, making them to have a value of 0. Then there were 141 instances when only the catalog had a promo and the rest were 0. Also, 118 instances when both catalog and store end had promotional sales, but in-store had 0. Then there were 268

instances where there was only an in-store promotion, with the other values as 0. And finally 307 instances where both in-store and store end had promotion sales, but 0 for the catalog.

For the Google_Mobility Feature, there is either no change in mobility, or negative mobility change throughout the data, showing us a lot of restricted periods for extended periods of time.

Approaches in Handling Missing Data:

Since there were no missing data on the csv file, I didn't have to make any changes on the data. However, I had to adjust the date and Percent to numerical values to enable me to analyze the data in future.

I also used "***winsorization***" to rectify the outlier data in my code.