Group Name: Solo Science Team

Name: Asher Chok

Email: asherchok@outlook.com

Country: United States

Specialization: Data Science

Problem Description

A ML model would yield better predictions if there are good features engineered from the existing dataset. The current dataset needs to have several new features to be extracted.

Data Cleansing

There are insufficient data for the variable *Google_Mobility* before the date 2022-02-09. If any model was fitted on this entire dataset, then it would not be accurate. Therefore, the dataset has been split into 3 groups to address the issue:

- Data set with *Google_Mobility* column dropped.
- Data set with zero *Google_Mobility* values before the first *Google_Mobility* non-zero value. *Google_Mobility* column is also dropped.
- Data set with *Google_Mobility* values after the first *Google_Mobility* non-zero value. The Google_Mobility column is kept.

Data Transformation

Few features were made from original columns:

- Categorizing *Product* into numerical values via label encoding method.
- A new column called *Promotions* where it is equal to 1 for when there's a promo for a product.
- A *Sales After Discount* column where it considers the actual sale from the product after discount.
- A *Sales Mobility Ratio* column that takes each *Sales After Discount* value and divide by absolute value of *Google_Mobility* value (might need to figure out if the absolute value is necessary later).
- A moving average that considers the previous 90 days of sales data, *MA_3months*.
- *Quarter* and *Season* from the date column.
- A label encoded seasonal column from *Quarter* called *Encoded_Season*.

There will be more features that will be needed to be engineered before performing model fitting.

Version Control and Source Code Management

GitHub Repository Link: asherchok/retail-forecasting: Data Glacier internship forecasting project repository (github.com)