Exercise 3

Materials:

"dunnhumby - Breakfast at the Frat.xlsx" provided in "Handout 4"

Objective:

On the product (UPC) level predict future sales from the sales history

1 Time Series Analysis (5 points)

1. Choose 5 top products (with best total sales). Verify that these products have been sold during the last 4 weeks. Use the products-lookup sheet to find out information about the selected products.

2. Plot a graph of the total weekly number of units sold for the chosen 5 products. Are there products for which you can detect a trend or a seasonal behavior?

   Remark: be aware that data is originally organized on the granularity level: (product, household) and you are interested in total sales per product.

3. Investigate the influence of promotions on the number of products sold. There are 3 promotion types: FEATURE (*product was a part of in-store promotional display*), DISPLAY (*product was in in-store circular*), TPR_ONLY (*temporary price reduction only (i.e., shelf tag only, product was reduced in price but not on display or in an advertisement*). Which promotion types seems to have the most influence on the number of products sold? Support your answer by adding vertical lines on your plot for the dates when the chosen promotion happened.

4. Use the `fbprophet` library (available in python and R) to produce a weekly time-series forecast for the last 4 weeks of sales separately for the 5 selected products. Provide the values of your mean squared error for each product.

   Tips:

   - For each product split the time-series into train and test parts, where the test part corresponds to the last 4 weeks of sales.

   - Try to change hyper-parameters of the `fbprophet` model, for example seasonality. Did parameter change effect the quality of the prediction?

- Use the `fbprophet` functionality to take into account effect of special events and provide the dates of promotions as input to the parameter `holidays` of the model. How did this influence the performance of the model?

2 Machine Learning Approach (8 points)

Goal: For each product predict total sales for last week

1. Select products that have been in sale for the last month
2. For each product, produce base-line predictions:
   - sales in the previous week
   - average weekly sales
   - moving average weekly sales with a chosen window

   Split the data in train and test parts. Provide the score of these base-line predictions averaged over all products of the test set.

   Tips:

   - The number of observations in your data-set should correspond to the number of products
   - Be careful not to include information about the target in calculation of the base-line prediction

3. Select 2 different machine-learning algorithms and briefly describe them.
4. For each product use as features weekly sales for the previous 16 weeks to predict sales of the last week. Describe the performance of each algorithm and compare it to the performance of the best base-line prediction.
5. To increase the size of the data-set and make use of the seasonality we adopt a rolling-window strategy with a lag of 16 weeks. The idea is:
   - For each product use the values $X_{t+1}, \ldots, X_{t+16}$ as features and X_{t+17} as a target for all possible $t \geq 0$ (in this way the data-set will contain more multiple samples per product)
   - Add date related features like year, month, and week number of the target X_{t+17}
   - Add aggregated features like total, mean, median, standard deviation of the sales for each row

- Add product related features from the product-lookup sheet. Use category encoding if needed.

- If possible use other columns to generate new features: HHS (housholds that bought the product) or STORE_NUM

  How adding new features influenced the performance of your algorithms?

- Choose a feature selection algorithm and describe briefly what it does. Try to use it on the features that you engineered and describe if it influenced the performance of the algorithms.