STAT 542 Project 2: Walmart Store Sales Forcasting

Wenbo Fu (679744457), Bingyan Liu (668046518)

1 Data and Objective

The data is historical sales data from 45 Walmart stores spread across different regions. The goal is to predict the future weekly sales for every department in each store. The evaluation is based on the weighted mean absolute error (WMAE):

$$\text{WMAE} = \frac{1}{\sum \omega_i} \sum_{i=1}^{n} \omega_i |y_i - \widehat{y}_i| \tag{1}$$

where n is the number of rows, $\widehat{y}_i$ are predicted sales, y_i are actual satles, ω_i are weights such that $\omega = 5$ if the week is a holiday week or 1 otherwise.

Details can be found in `https://liangfgithub.github.io/Proj/F23_Proj2.pdf`.

2 Method

Our algorithm contains two parts: data preprocessing and model fit.

The data preprocessing part takes the training data and testing data as inputs, which contains the following steps:

- SVD on the training data for the same department, i.e. $X_{m,n}$ where m is the index for stores and n is the index for weekly sales. Fill NAN entries with 0. Keep the largest 8 eigenvalues. This step is used to smooth out noise.
- Use the testing data (Dept, Store) as keys to partite the training data. Specify week as categorical feaures, year as an numerical feaure and include square of year as a new feature.
- Remove constant or colinear columns from the training data, drop the same columns for the testing data.

The model fit part takes the processed traning data and the testing data and the testing data with label as inputs and take the OLS model to make predictions.

3 Result and Discussion

The WMAE on the 10 folds:

$$1947.7, 1391.5, 1393.8, 1524.6, 2318.2, 1637.7, 1616.0, 1362.8, 1351.4, 1332.6 \tag{2}$$

average of the WMAE over the 10 folds is 1587.6. The The processing time on the 10 folds:

$$25.38, 26.35, 30.03, 32.42, 36.35, 38.61, 39.99, 41.4, 42.91, 45.75 \tag{3}$$

All the time are in the unit of seconds. The code is excuted on a Macbook Air, 8GB memory, Apple M1 chip.

To summarize, we use methods described in the instructor's posts: `https://campuswire.com/c/G06C55090/feed/363`, `https://campuswire.com/c/G06C55090/feed/364`. We have also tried other SVD cutoff dimensions or use weighted linear regression, but the WMAE does not improve. One possible way to improve is to adjust the holiday information in fold 5, as discussed in `https://campuswire.com/c/G06C55090/feed/366`, but we did not implement this post-preprocessing here.