Project 2: Walmart Store Sales Forecasting

tianni2@illinois.edu
October 15, 2021

1. Overview

We are provided with historical sales data for 45 Walmart stores located in different regions ranging from 2010-02 to 2011-02, which was used in Kaggle competition. Each store contains many departments. The dataset contains 421570 observations with 5 variables:Store, Dept(Department), Date, Weekly_Sales, IsHolidays(whether the date is holiday or not). The goal of this analysis is to predict the future weekly sales for each department in each store based on the historical data.

For prediction, we need to first split the data as training and test data to test the accuracy of different model to get the most accurate one. Also, we tried several methods includes linear model, “naive time serious” etc. And we end up with the linear regression model with SVD(Singular Value Decomposition).

2. Data Preprocessing

2.1 Data overview

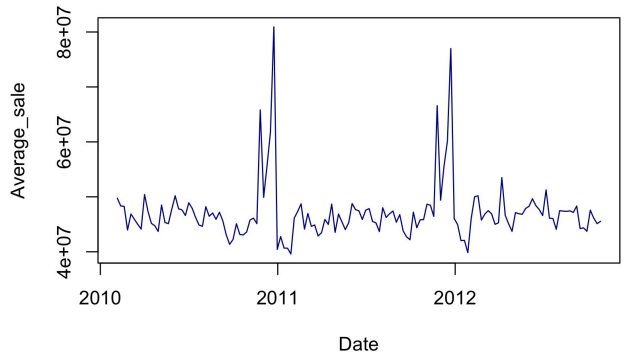

Figure1. Averaged Weekly Sale through 2010 to 2021

From figure 1 we could see that, there exists some exact pattern in the weekly sale. During the weeks of the end of each year, there would be peaks. This might related to the Thanksgiving Day and Christmas.

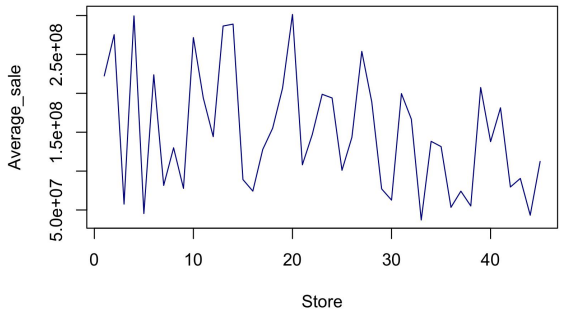

Figure2. Averaged Weekly Sale for different stores

Looking through figure 2, we could see that the sale difference between stores are huge and inconsistent. Location and store size might be the reason.

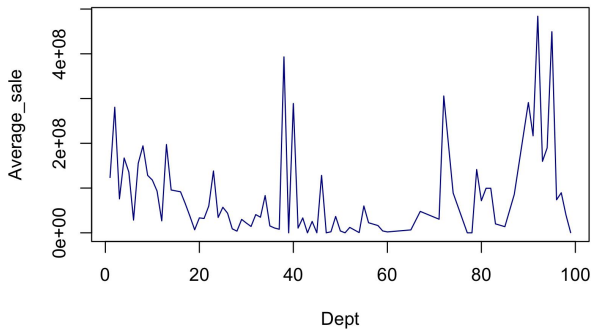

Figure3. Averaged Weekly Sale for different departments

We can see from figure 3 that, some specific department tend to have higher weekly sale. And the sale difference between departments are also huge.

2.2 Data pre-processing

We were given the zip file, train.csv.zip, and we need use necessary code to generate datasets I need for this project. After data splitting, we get 12 files:

1. train_ini.csv: 5 columns (“Store”, “Dept”, “Date”, “Weekly_Sales”, “IsHoliday”), while ranging from 2010-02 to 2011-02.
2. test.csv: 4 columns (“Store”, “Dept”, “Date”, “IsHoliday”), ranging from 2011-03 to 2012-10 with the “Weekly_Sales” column being removed.
3. fold_1.csv, …, fold_10.csv: contains 5 columns (“Store”, “Dept”, “Date”, “Weekly_Sales”, “IsHoliday”), one for every two months starting from 2011-03 to 2012-10, they are all splitting from the test.csv file.

Further preprocessing

The evaluation code would run my analysis in a loop with iteration 10, for each iteration:

4. Since the store and department of observations in training and test data are not exactly same, we need to find the unique pairs of store and department combination that appeared in both training and test sets.
5. Moreover, pick out the needed training samples and convert them to dummy codding and put them into a list. Do the same for the test set.

(Note: these two steps were included in my mypredict function while the 1-3 steps were stored in another file.)
After these preprocessing steps, my data were ready for analysis.

3. Building Model

Looking through these plots we find that the sales pattern of a department seems similar across stores. So we'll first apply SVD to reduce noise. To apply SVD, we first arrange data from a particular department as a m-by-n matrix, where m is the number of stores that have this particular department and n denoted the number of weeks. Then we could apply SVD and choose the top d components. Here we selected d=8 based on the accuracy. Then we get a reduced matrix and plug it into the following linear regression analysis.
With the SVD reduced matrix, we proceed the linear regression analysis and making prediction. Below is the wae table for each fold we get and the average wae is 1608.776, which is not bad.

Fold 1	1941.581
Fold 2	1363.462
Fold 3	1382.497
Fold 4	1527.280
Fold 5	2310.469
Fold 6	1635.783
Fold 7	1682.747
Fold 8	1399.604
Fold 9	1418.078
Fold 10	1426.258

4. Running Time

With 2019 MacBook Pro, 2.4 GHz Intel Core i5 CPU with 16GB memory. The running time for is around 2min and 20 seconds.

5. Conclusion

Proceeding with the combination of linear regression and SVD, we get a rather good prediction accuracy. Moreover, the pre-processing is also a necessary part of this analysis. At the end, the averaged WAE is 1608.776.

6. Acknowledgment

Professor Liang’s Campuswire posting for the Project 2(Fall 2021)
R document about the packages data.table,tibble, dplyr