CS 598 PSL – PROJECT 2

1.INTRODUCTION

For this project we were provided with historical sales data for 45 Walmart stores located in different regions 2010-02 (February 2010) to 2012-10 (October 2012). This dataset contains 421570 observations and 5 variables including the store and department number, date, the weekly sales for each of the stores by department and an indicator variable that shows if each of the given week is a special holiday week or not. The goal for our project is to forecast the future weekly sales for each department in each store based on the historical data given to us.

For this project, we tried several approaches including the solution approach provided by Dr. Liang, Time Series Linear Model (*tslm*), Seasonal and Trend Decomposition using Loess Forecasting Model (*stlf.svd*) all of which will be outlined below. Let us look at the implementation steps that were done for this project.

2.DATA PROCESSING

The train.csv file that was provided to us was split into the following files:

1. Train_ini.csv – This file contains the same weekly sales data as the train.csv file but from the time frame 2010-02 to 2011-02.
2. Test.csv – This file contains 4 variables Store, Dept, Date and IsHoliday for the test time frame 2011-03 to 2012-10. The Weekly_Sales column was removed from this dataset.
3. Fold1.csv... Fold10.csv – The train.csv file provided for the test time frame, 2011-03 to 2012-10, was further split into 10 folds each with each fold containing 2 months' worth of weekly sales data.

For each of the 10 folds the following initial data processing steps were performed:

1. Since the dates are not same across each fold, we computed the unique test dates for each fold that needed prediction.
2. Similarly, we computed the unique stores and unique department within each store that need prediction and stored them all together in a dataset. Same steps were performed for the train dataset as well.
3. We then create a time series view for each store's department for the train dataset.
4. For the test dataset, we created a place holder dataset with a time series view.

3.IMPLEMENTATION

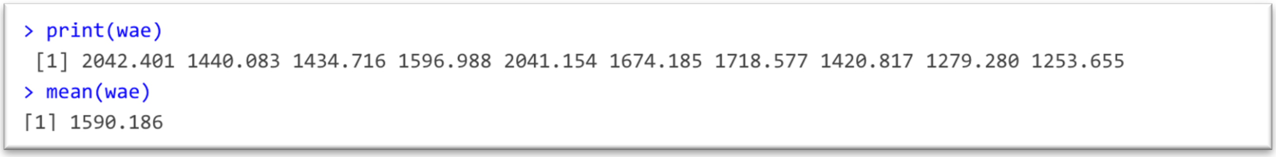
Approach 1: We started by following a combination of approaches suggested by Dr. Liang. We created a new categorical variable called "Wk", which enabled us to use each week's data from the previous year to help predict the corresponding week's data for the following year. In order to make the week data match accordingly, we subtracted 1 from the weeks in 2010. We also created a numerical variable "Yr" for Year. "Wk" and "Yr" served as the explanatory variable in the linear regression model. We then looped through the relevant departments and stores as mentioned in the Data Processing steps 2 and 3, obtained the sales data, created the design matrices for the training and test sets, and fit the linear regression model. Coefficients with NA were replaced with 0. The predictions were made using ordinary matrix computation. Using this approach for all the folds resulted in an average WAE score of 1659.709.

Approach 2: Using the Time Series Linear Model (tslm) approach for all the folds resulted in an improved average WAE score of 1653.324.

Approach 3: We first tried putting Approaches 1 and 2 together, using Approach 1 for folds 4, 5, and 10, and Approach 2 (TSLM model) for folds 1, 2, 3, 6, 7, 8, and 9. This combination of approaches resulted in an average WAE score of 1652.083.

Approach 4: We tried combining approaches 2 with stlf.svd. Approach 2 (TSLM) was used for folds 1, 2, 3, 4, 5, 6, 7, and 8. Approach 3 (stlf.svd) was used for folds 9 and 10. The resulting average WAE score from this strategy was our lowest, at 1638.101.

Approach 5: As the next step, we focused on 'Fold 5' for which we were getting the highest WAE of 2327.638 in Approach 4. The high WAE in 'Fold 5' is caused due to a slight shift in the Christmas shopping season from 2011 to 2012. To address this issue and to bring the WAE lower, we reimplemented David Thaler's shift function that won the Kaggle competition. This resulted in a an average WAE score of 1609.452 which meets the desired threshold as set by the Professor. We further fine-tuned by varying parameters for tslm that resulted in an average WAE score of 1590.186.



```
> print(wae)
 [1] 2042.401 1440.083 1434.716 1596.988 2041.154 1674.185 1718.577 1420.817 1279.280 1253.655
> mean(wae)
[1] 1590.186
```

4. TIME SERIES LINEAR MODEL (TSLM)

Time Series Linear Model or tslm() is a wrapper for the lm() which includes the trend and seasonality components for time series. Here 'trend' indicates the time trend and 'season' is the factor indicating the season in this case. This approach was applied to Folds 1-9 for all departments.

5. SEASONAL AND TREND DECOMPOSITION LOESS FORECASTING MODEL(STLF.SVD)

The *stlf* method uses forecasting and decomposition. For this project we applied *stlf* method to folds 9-10. The ARIMA model was used for forecasting the seasonally adjusted time series, the final model parameters were selected using ‘BIC’. ‘s.window = 7’ and ‘t.window = 21’ parameters were also utilized to arrive at the desired WAE scores.

7.CHALLENGES

There were a lot of resources available for this project, including Piazza Notes, Office Hours, and code from the Kaggle competition. It was challenging to adjust the code to work with newer packages or more recent versions of R. We also ran into issues when troubleshooting data formatting, as well as figuring out an error-free way to use different models depending on the fold in question.

8.CONTRIBUTIONS

Our group strategy involved all of us giving different approaches a try to achieve the lowest average WAE score. We began by approaching the project each independently, to ensure we learned by doing. We then had check ins where we discussed what strategies we tried, and shared findings.

1. XXX worked on the report, researched different models, worked on the shift(), and got the combination of 1 + 2 working at WAE score of 1659.709.
2. XXX got the tslm and stlf.svd approaches 2 + 3 and the shift() working, leading to a WAE score of 1609.452.
3. XXX worked on the report, researched potential techniques, code troubleshooting, worked on shift() reimplementation and fine tunings for sltf.svd leading to a final WAE score of 1590.186

9.CONCLUSION

By using a combination of both *tslm* and *stlf.svd* models and employing post processing steps and averaging across the 10 different folds we were able to meet the required benchmarks and arrive at a final WAE score of 1590.186. Working together to try several methods/models, brainstorming and trying to improve the WAE scores even more was a great learning experience for our team.

7.TECHNICAL SPECIFICATIONS

The final model combinations used in this project were tested on the following systems:

1. Macbook Pro, 2.3 GHz Quad-Core Intel Core i7 and 16GB RAM. *Run time - 13.59846 mins*
2. MacBook Air, 1.6GHz, 8GB RAM. *Run time - 8.00142 mins*
3. OS Catalina with a i5 1.4 GHz Quad Core processor, 8 GB RAM, *Runtime - 8.196028 mins*

8.ACKNOWLEDGEMENT

1. Dr. Liang’s solution approach provided on Piazza, Piazza posts and Office Hours
2. https://github.com/davidthaler/Walmart_competition_code/blob/master/grouped.forecast.R
3. https://liangfgithub.github.io/Example_Code_Project2_Josh.html