CS598 Project Report- Walmart Store Sales Forecasting (Netid: VKK2)

1. Introduction & Goal

For this project we are provided historical data for Walmart stores located in different regions. Each store contains many departments. The goal is to predict the future weekly sales for each department in each store based on the historical data. Evaluation of model is per Kaggle using weighted mean absolute error and goal is to get below 1630.

2. Exploratory data analysis and Pre-processing

The data in "train_ini.csv" contains 5 columns ("Store","Dept","Date",Weekly_Sales", "IsHoliday) and ranges from 2010-02 to 2011-02. This data was be used to predict sales from 2011-03 to 2012-10. The file test.csv contains same columns as train except for "Weekly_Sales" which was predicted. Prediction was made in 2 month step sizes, for example since training data is provided from 2010-02 to 2011-02 which will be used to predict from 2011-03 and 2011-04 and then original fold_1.csv containing actual data from these months will be appended to original training data from 2010-02 to 2011-02 and will be used to predict from 2011-05 and 2011-06. This process of predicting and appending is done 10 time for total of 20month prediction until 2012-10. As a part of pre-processing this fold_x.csv where x represents fold number between 1 and 10 is created.

The training data from 2012-02 to 2012-10 consists of data from 45 stores and 99 different departments. To understand how Weekly sales changes with respect to store and department, data plot was done and shown in Figure 1. From Weekly Sales by Store plot we can observe the highest peak occurs close to end of the year and not all stores are subject to the same peaks. From Weekly Sales by Dept. plot we can see huge variance between weekly sales difference between departments and each department may exhibit a peak at a different time of the year. There was some missing data, but it was quite insignificant, and no attempt was made to interpolate, instead N/A was replaced with 0's.

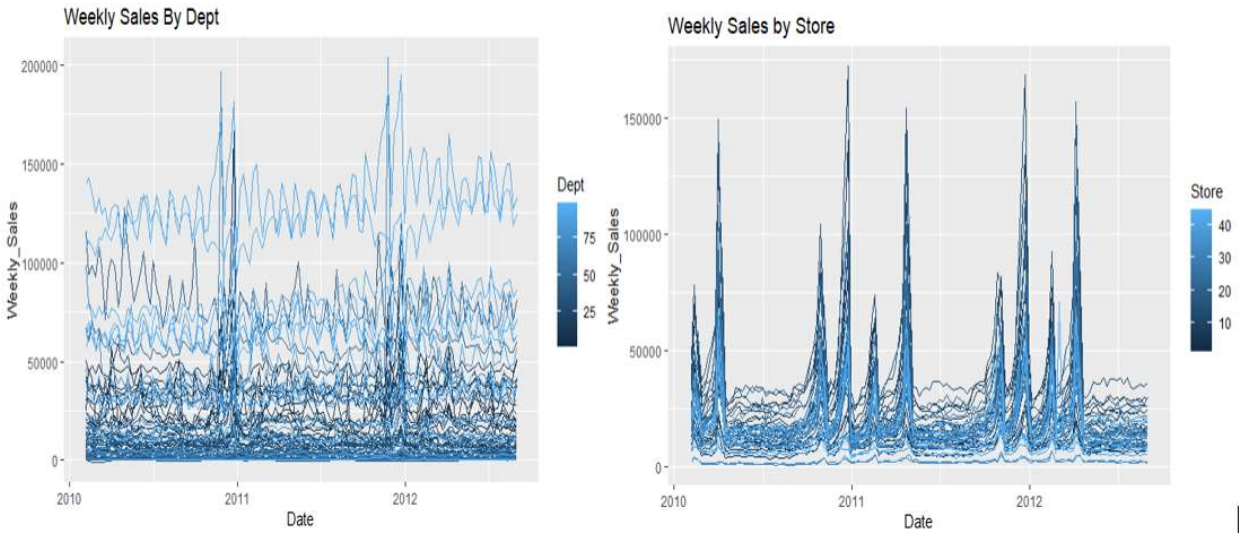

Figure 1: Plot of Weekly Sales by Dept & Store

3. Naïve Model

This is the simplest model used in which we predict all future weekly sales using data from the most recent week. From plot in figure 2, it can be seen with this model predicated sale remains the same for 2 months. This runs the fastest but is highly inaccurate, its mean **WAE is 2812.67**.

4. Seasonal Naïve Model

In this model we use the past data and use the same data from that week from previous year for the current year. Since using the week function between 2010 and 2011 gives 52 and 53 weeks respectively we subtract 1 after week is computed so all holidays (Super Bowl, Labor Day, Thanksgiving & Christmas) line up by week. Even though the weeks lineup but sale shift still exists as days prior to Christmas is one of the peak shopping periods. Refer to Figure 2 for summary of results, mean **WAE is 1858.85**.

6. Trend & Seasonal linear Model

Modeling was performed on every combination of store and department. Corresponding sales data from each combination of store and department was extracted and using variable of "Wk" and "Yr" a linear regression model was fit. Design matrix used for this model combined both training and test to avoid errors during lm and manual matrix multiplication was done in lieu of built-in predict function. Additionally, of the NA coefficients were replaced with 0. Also, note for model fitting "Wk" was used as a categorical feature. Refer to Figure 2 for summary of results, without circular shift (discussed next) mean **WAE is 1659.71**.

5. Post-processing via circular shift

The winner on Kaggle competition did shifting a portion of departments sale occurring 3 weeks prior Christmas to the weak of Christmas. Instead, I used approach of shifting 1/7 of sales from 48th week to week 49 and repeating this shift for week 49 to week 50 until week 42. The reason for this shift is Christmas falls on Sunday (Week 53) in year 2011 instead of week 52. This shift was only applied after Weekly sales was determined by the model and for week 49-52. This circular shift was done to fold 5 only as it applies to this peak Christmas shopping period (Thanksgiving to Christmas). The reason for this shift is as our evaluation metric weighs heavily on getting holiday period accurately. This circular shift reduced for linear model above mean WAE of fold5 improved from 2324.5 to 2028.68 and overall mean **WAE to 1630.12**, which is slightly above our target of 1630.

6. Final Model Submitted

Final model submitted is combination of snaive and linear trend & seasonal model with applied circular shift on fold5. Since there are holidays in fold 4, 5, 6, and 10 and our WAE is weighted highly on these periods. Prediction of these periods was plotted, and it was observed actual value was close to somewhere in between the 2 models on some of the holidays. So, to reduce mean WAE slightly further, average of the predictions between the two models was done for the period when there was a holiday. Doing this average, reduced mean **WAE to 1625.09.**

4. Results & Conclusion

For this analysis hand coded approach of using different forecasting methods were used. This report outlines multiple approaches to a solution each provide a tradeoff of accuracy vs complexity/runtime. Figure 2 below shows example of prediction from store and Dept#1. It can be observed Naïve model has various step up/down type response showing lack of inaccuracies. From the table in figure 2 it can be observed using a model that average linear regression & season naïve model when there is a holiday period provides best accuracy for forecasting for the whole year. From this project not only, I learned data wrangling techniques, but how it can help give basic forecasting prediction fast. Additionally, this project has peaked by interest more in learning more about forecasting and plan to explore more by using advanced time series analysis techniques using "forecast" package in R in the coming future.

Fold#	Naïve	Snaive	TSLM	TSLM(with Shift)	SubmittedModel(If Holiday: Do avg of LM & snaive shift)
1	2078.72	2262.42	2045.243	2045.243	2045.24
2	2589.34	1787.08	1466.91	1466.91	1466.91
3	2253.93	1779.05	1449.85	1449.85	1463.33
4	2823.09	1716.12	1593.99	1593.99	1542.56
5	5156.012	2108.91	2324.49	2028.68	1980.08
6	4218.35	1696.9	1677.48	1677.48	1547.38
7	2269.9	2086.97	1722.27	1722.27	1776.34
8	2143.84	1750.28	1428.21	1428.21	1467.01
9	2221.14	1719.88	1443.96	1443.96	1478.5
10	2372.42	1680.96	1444.65	1444.65	1483.52
Mean	2812.67	1858.85	1659.71	1630.12	1625.09

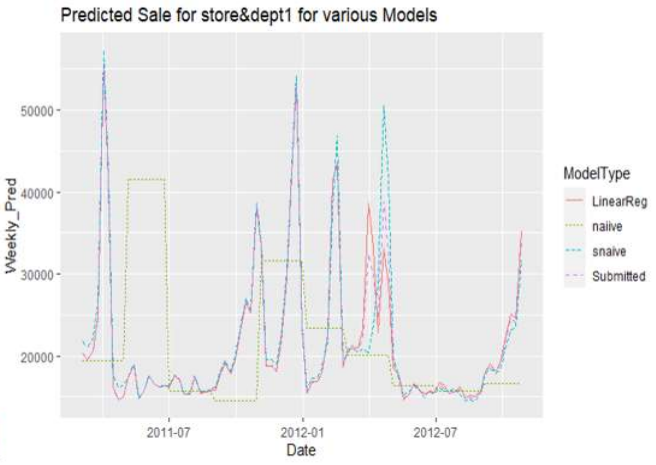

Figure 2:Summary of WAE of each model & sample plot of prediction

4. Acknowledgement

CS598 Professor Liang's Hints provided on piazza on what has been tried was extremely helpful

Kaggle Competition here .

David Thaler- Winner of Kaggle Competition. His idea of post processing suggested here.

5. Appendix- Model Runtime

All models were trained on a Windows PC using R(version 3.6.0) with library of lubridate and tidyverse. Detail computer configuration are as follows: Windows 8 64bit OS, Processor: Intel Core i5-4460 CPU @3.20GHz, 8GB RAM. Below is runtime for each of the models. The runtime includes time to train and predict for all 10 folds

Fold#	Naïve	Snaive	TSLM	TSLM(with Shift)	SubmittedModel(If Holiday: Do avg of LM & snaive shift)
Runtime(sec)	1.9	2.5	792	714	714

Figure 3: (Left) Runtime for different models