Project 2: Walmart Store Sales Forecasting

1. Introduction

In this project we have used historical sales data for 45 stores located in different regions and this data was used in Kaggle competition [1]. Each store contains many departments. Our goal is to predict the future weekly sales of each department in each store based on the historical data. The success criterion of the model is to achieve the WMAE (weighted mean absolute error) which is below the specified target (1630). The initial training sets starts from 2010-02 for 13 months and each test fold containing two months after that. The training set keeps incrementing by previous test fold data. The weight of the holiday week is five while that of a normal week is one. To achieve this, we have taken the following steps which would be discussed in detail in subsequent sections with overall structure as below.

- Analysis of input data
- Model determination and post-processing
- Result Analysis

2. Input Data Analysis

We performed some analysis of data using the hints provided by Professor Liang in Piazza [2] and were able to observe a few things which would be critical from modelling perspective.

i. This is time series data, so we analyse overall flow of weekly sales over years for all stores and depts. As a starting point, we look at average weekly sales [figure 1].

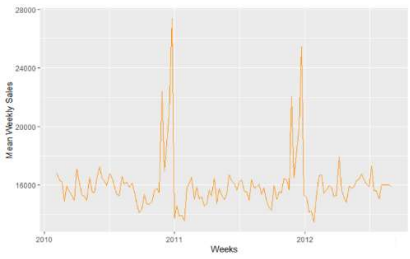

Figure 1: Mean Weekly Sales Over Time Horizon (Average at Walmart level)

ii. We clearly see the seasonality of the sales with sharp spikes near the end of each year for complete Walmart level by taking average of all stores and departments.

iii. We don't observe any clear trend from the overall average in figure 1 as the spikes have gone down a little but rest of the weeks are irregular.

iv. We dive deeper by randomly selecting some stores (#1, $10, and #25) and observing the trend and seasonality at the store level as shown in Figure 2. We can see that overall seasonality is same as the overall model, but we can see the trend getting marginally clearer at store level. At store #1 and #10, we see somewhat upward trend in sales except the holidays, while for store #25, the trend shows a slow downward trend but nothing conclusive. This encourages to dive deeper at department level to understand further.

v. Analysing at the most granular level of store and department, we see that each department shows the unique pattern with most showing seasonality but some being

even throughout the year such as department 40 in figure 3. The trend is also quite different for each department. In some cases, spike has got distributed over multiple weeks in 2011.

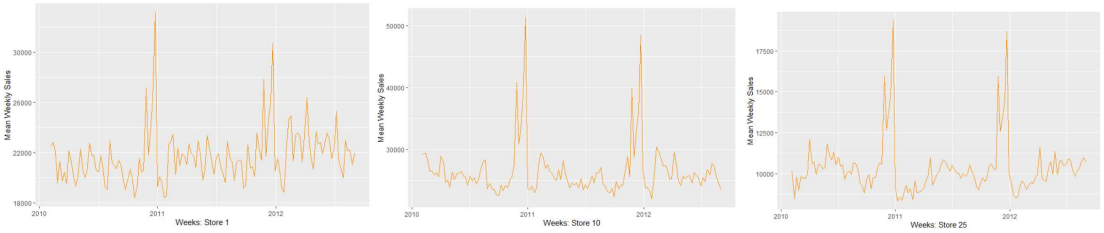

Figure 2: Mean Weekly Sales Over Time Horizon for Store 1, 10, 25

Figure 3: Weekly Sales Over Time Horizon for Store25, Dept 10, 30, 40

vi. All the previous analysis makes it clear that we should do the forecasting at each store and department separately.

vii. We replace all the missing value with 0 without any loss of generality.

viii. We saw that the last week of 2010 is 53rd, but the that of 2011 is 52nd, creating a mismatch as both weeks are regarded as the Christmas holiday week. However, this issue was easily fixed by subtracting 1 from 2010's weeks as it doesn't start from one anyway. We also observed the change in date included in week's which becomes the base of post-processing in our TSLM model discussed in subsequent sections.

3. Methodology and Analysis

We aimed to choose a strategy which would achieve target WMAE. We tried four strategies: Naïve, Seasonal-Naïve, Time-Series Linear Model, and TSLM with post-processing of prediction (TSLM*). The WAE for each model for each fold and its mean is given in Figure 4.

i. **Naïve**: In this model, we simply look at the latest forecast in training data and use that to forecast the weekly sales of each week of that department and store combination. By our data analysis, we knew that this wouldn't give a good result, but we are keeping it to start with this as benchmark.

ii. **Seasonal Naïve:** In this model, we try to use seasonality by taking the value from previous year's same week. This improves the WMAE significantly [figure 4].

iii. **TSLM:** Here, we try to use seasonality as well trend by introducing the time series linear model where *year* acts as trend, and *weeks* as seasonality factor. We change the weeks into categorical variables as we want to find seasonality factor in that group without looking at its actual value. The result has improved further but we observed that the holidays have moved from one week to another week as we move from 2010 to 2011. So, to take care of it, we decided to use a post processing as discussed in next model.

iv. **TSLM*:** We extend the TSLM model by adding a post processing. We observe that there is a difference of one day by looking at the dates included in each week. So, we use the 6/7 of the prediction of current week, and 1/7 of the next week, and combine them to predict the new value for holiday weeks 49-52 in fold 5 as that's where we observed the most error in TSLM model. As the score was below 1630, we stopped here but there are other methods which could be explored to reduce the error further.

Selected Model

Folds	TSLM*	Naïve	Seasonal-Naïve	TSLM
1	2045.243	2078.726	2262.422	2045.243
2	1466.912	2589.338	1787.081	1466.912
3	1449.852	2253.936	1779.052	1449.852
4	1593.998	2823.098	1716.117	1593.998
5	2021.182	5156.012	2400.395	2324.496
6	1677.483	4218.348	1696.900	1677.483
7	1722.274	2269.904	2086.967	1722.274
8	1428.212	2143.839	1750.283	1428.212
9	1443.960	2221.145	1719.887	1443.960
10	1444.656	2372.425	1680.956	1444.656
Mean	1629.377	2812.677	1888.006	1659.709

Figure 4: WAE of Models

4. Tech Specs

Operating System: Windows 10. **Processor**: Intel Core i5, 2.4GHz. **RAM**: 16GB

Overall, it took ~22 minutes to run the complete program (10 folds) with majority of folds taking ~2 minutes to run but fold five took ~3 minutes. One of the reasons is that we have added a post processing step to massage the forecast as discussed earlier only for fold five.

5. Conclusion

The project gave us deep insights into the time series models. We realized that how simple modifications in the model can improve the accuracy as we move from basic to complex models as discussed in previous sections. In time-series model, the linear model itself was powerful enough to improve the forecast significantly. We also found that some minute observations in datasets such as holidays being captured in different weeks, can help with improving accuracy without making models too complex.

Another major learning from this project was to use R for data wrangling by writing sql-like queries using mutations. Although we had to consult stack overflow [3] often, but it helped to write the post-processing in only a few lines rather than a complex loop.

6. Acknowledgement

[1] https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting.

[2] Professor and TAs: Amazing help were provided in the office hours to understand concepts of different elements of the projects.

[3] http://stackoverflow.com/: Several pages on this website helped in some nuances of coding.