CS598 - Project 2: Walmart Store Sales Forecasting

Author: Gunther Correia Bacellar, NetID: gunther 6 Date: 11/05/2020

Introduction

In this project I analyzed historical sales data of 45 Walmart stores located in different regions, with up to 99 departments per store to produce two months of weekly forecast from 2010-02 to 2011-02 and calculated the weighted mean absolute error (wmea) of all models in order to define the best one.

Exploratory Data Analysis

I started analyzing the training time series (figure 1). I clearly observed the seasonality aspect of the data as well as peaks in some specific weeks around holidays for both the consolidated data and by store and department data.

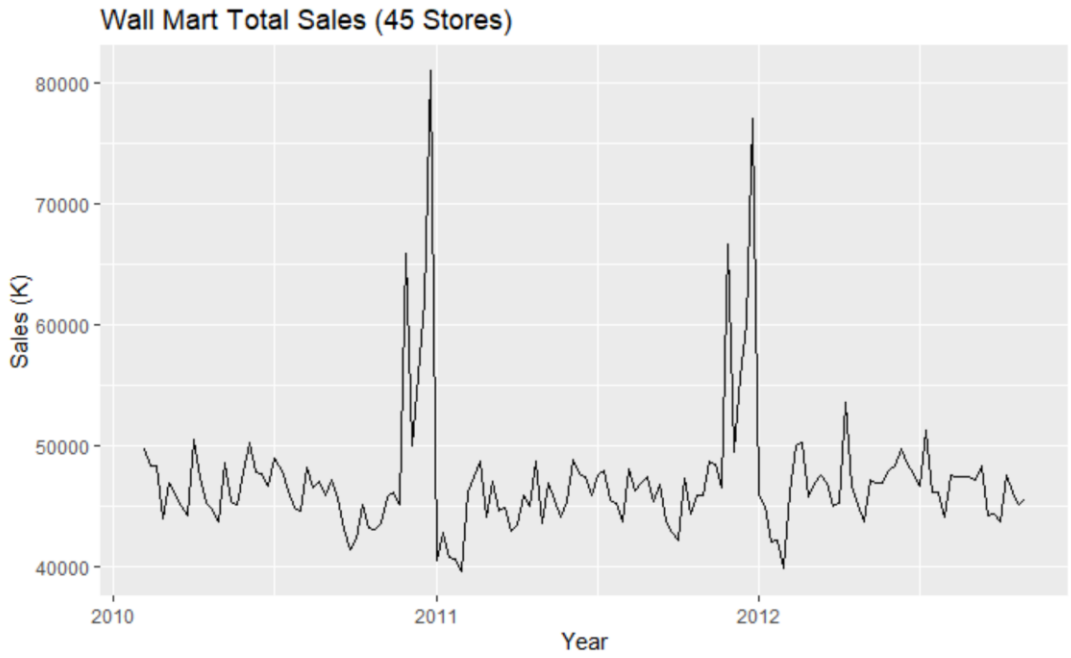

Figure 1: Graph with consolidated weekly sales of 45 stores from 2010 to 2012

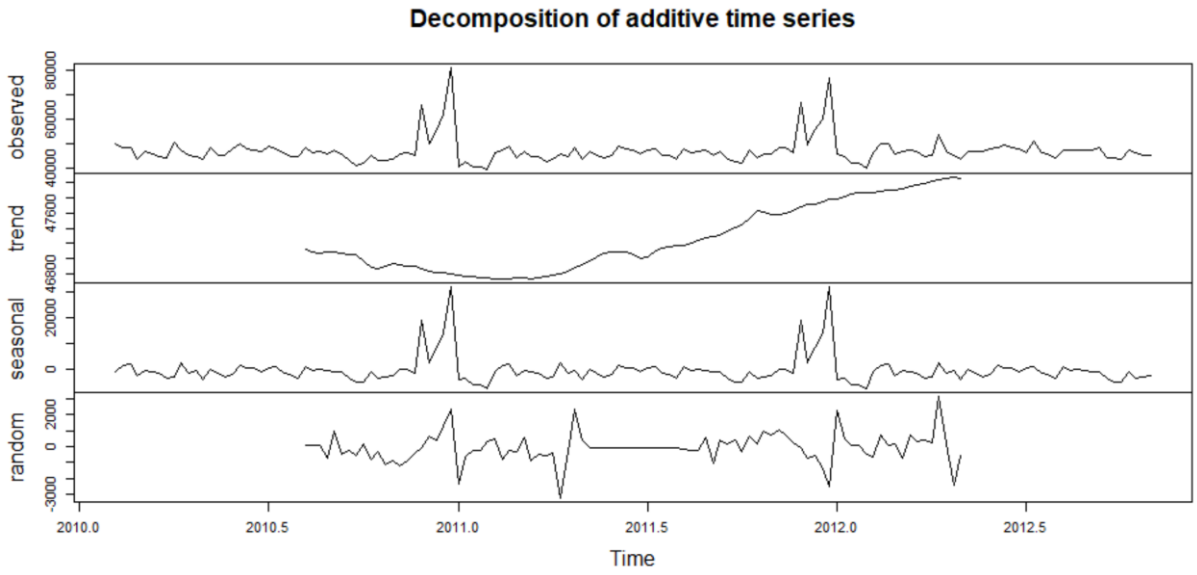

Figure 2: Decomposition of consolidated weekly sales of 45 stores from 2010 to 2012 using the decompose function in the time series

After that I run the ***decompose*** function of training time series (figure 2), separating the trend, seasonal and random components of observed graph. Figure 2 also shows a declining trend until the beginning of 2011, and an increasing trend after that. Also the random graph looks not totally random, as it is influenced by holiday sales.

Between 2010 and 2011 there is a mismatch in the week numbers. 2010 has 53 weeks and 2011 has 52 weeks. As 52 and 53 weeks are considered holiday weeks, I created an ***Adjust_Calendar*** function to shift the 2010 weeks as well as to calculate Year and Week per date to support the modeling process. After that, all 5 holidays in 2010 (weeks 6, 36, 47 and 52) matches between 2010 and 2011. I also addressed the missing values when analyzing each model individually.

Model 1: Season Naïve with and without shift on 2011 weeks 49 to 52

Due to the seasonality of data (figure 2), the modeling of Naïve or average models didn't provide significant WMAE, so I didn't consider them for further analysis. So, I started with season Naïve model, predicting future weekly sales based on information from 52 weeks before using the Week and Year variables. This model provided a mean WMAE of **1,888.0** from 10 different test folds. Out of all 10 WMAE fold results, two had very larger values (larger than 2100): fold 1 and fold 5. Fold 1 is impacted by the size of training data, while fold 5 is impacted by having two holiday weeks in the end of year, with slight shift of the Christmas shopping season from 2010 to 2011, generating higher weights in WMAE.

To address this issue with fold 5 I created a ***Shift_Calendar*** function, which shifts the weekly sales in **n** days the period between two weeks (**Wk_start** and **Wk_end**) in a specific year (**y**). I applied this transformation to sales prediction of fold 5, for year 2011, weeks 49 and 52 (after adjusting 2011 in 1 week) and n=1. This shift transformation produced a lower WMAE for fold 5, dropping from **2400.4** to **2108.7**. A similar drop of around 300 points on WMAE was observed in all models tested. This shift resulted in a new mean WMAE of **1,858.8**.

Model 2: Season Naïve with a growth multiplier, with and without shift

The trend graph of figure 2 shows a clear growth influence year by year in the overall weekly sales. Something similar was observed by store and department. Based on this fact, I tested a new Season Naïve model combined with a mean growth multiplier value by department and store. My assumption is that departments and stores with some reasonable level of yearly increase or decrease of weekly sales would keep this trend in the following year.

To calculate the mean growth multiplier, I considered the yearly growth for each combination of stores and departments. I generated a list of growth rates by dates, and I limited it only to values in the range of -30% to + 30% yearly growth (multiplier: 0.7 to 1.3) to avoid that extreme growth values had a big influence in the prediction. Finally, I calculated the mean of this list of growth rates to have an unique multiplier by each combination of department and store, and multiplied it by the Season Naïve forecast values. This resulted in a model with mean WMAE of **1,686.2** without shifting 2011 forecast (fold 5), and WMAE of **1,656.5** after shifting forecast of fold 5.

Model 3: Linear regression, with and without shift

The third type of model I analyzed was the linear regression using weekly sales as response variable, and Year and Week as feature variables. I developed this model using as reference the Piazza example provided by professor (What we have tried (III)) and modified it. I made the Week variable as categorical, while keeping Year numerical. To avoid error running the linear regression, I used the design matrices for the training and test data before fitting the regression, replacing NA coefficients with zero and finding the prediction using the design matrices. This model resulted in average WMAE of **1,659.7** without shifting forecast in fold 5 and **1,629.4** after shifting forecast in fold 5.

Although this model produced an acceptable mean WMAE, below **1,630**, for the first 6 folds the linear regression with shift didn't performed well, with an average WMAE (6 folds) of 1,790.1, much larger than the Seasonal Naïve with growth multiplier with shift model, with average WMAE (6 folds) of **1,628**. On the other hand, for the folds between 7 to 10, the linear regression showed the best results. One explanation for such difference is that the small

size of training dataset for the first 6 folds (1 to 6) created impacted the accuracy of the linear model. Also, longer periods (fold > 6) created a multiplier index (average growth for all dates) that influence less the overall forecast as longer periods used to provide higher variance in the growth rate.

Model 4: Combining linear regression and season Naïve with a growth multiplier

As the size of period and training dataset influence the WAE of model 2 (Seasonal Naïve with growth multiplier) and model 3 (linear regression), I decided to combined then using the model 2 for smaller training datasets (folds 1 to 6) and model 3 for larger training datasets (folds 7 to 10). For fold 5, I also applied the 2011 shift in the predicted values for weeks 49 to 52. This model resulted in an average WMAE of **1581.2**, the best of all models tested. Figure 3 summarizes the WMAE results of all models analyzed, with the numbers and graph in red for the best model.

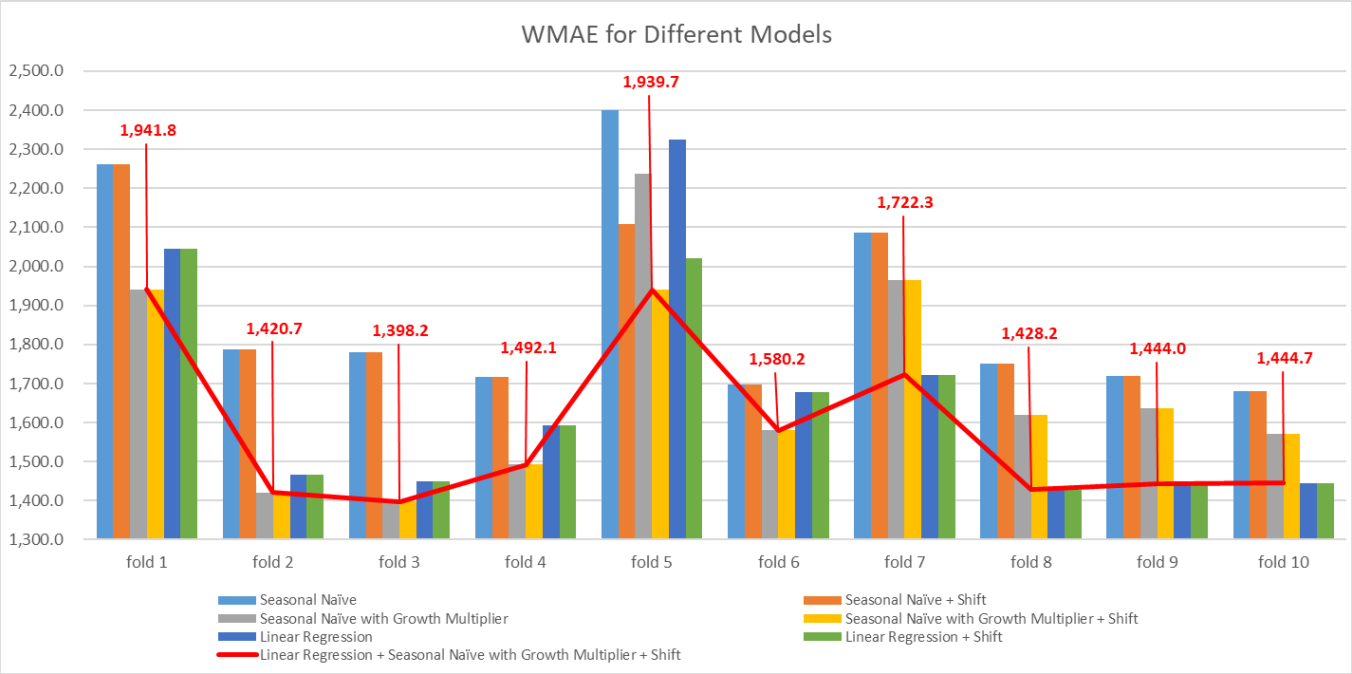

Figure 3: WMAE of all models analyzed with and without shifts to address fold 5. In red the graph and numbers for the best model, mixing linear regression and seasonal Naïve with growth multiplier and shift

Conclusion

Working with time series to predict weekly sales requires some adjustment to capture the right influence of holidays in the training and test data as well as the mismatches between week periods for different years. Linear regression produces good results, especially when the training data is large enough. A modified version of season naïve with growth multiplier produced better results for smaller training data. Table 1 shows a summary of average WMAE of all models tested in this analysis as well as the computer time to train and test the 10 folds combined. The processing time for training and testing all 10 folds are based on a Surface Book 3 machine with Intel Quad-Core 10th Gen i7-1065G7 CPU @ 1.50GHz, 32GB running Windows 10 64-bit.

Model	Mean WMAE	Time (s)
Seasonal Naïve	1,888.0	9.95
Seasonal Naïve + Shift (fold 5)	1,858.8	10.07
Seasonal Naïve with Growth Multiplier	1,686.2	10.16
Seasonal Naïve with Growth Multiplier + Shift (fold 5)	1,656.5	11.85
Linear Regression	1,659.7	50.91
Linear Regression + Shift (fold 5)	1,629.4	52.26
Linear Regression + Seasonal Naïve with Growth Multiplier + Shift (fold 5)	1,581.2	27.02

Table 4: Mean WMAE and computer time for the different models analyzed