Project 2: Walmart Store Sales Forcasting

XXX
Email

XXX
Email

November 8, 2021

Abstract

This report is for the second project in CS-542 (Practical Statistical Learning) for the Masters in Comptuer Science program at the University of Illinois Urbana-Champaign. The goal of this project was to predict future weekly sales for each department in each store based on historical data of 45 Walmart stores.

1 Contributions

Both authors of this paper contributed evenly to this project. We each experimented with the different models before deciding together to pick one. XXX organized the paper and XXX organized the code.

2 Introduction

The methodology for our implementation of a model began with implementing a series of suggestions from Dr. Liang. We eventually settled on an improved linear model with SVD as a processing step. We looked at two naive models and three linear models.

3 Model Approaches

3.1 Naive Model

The first approach we attempted was simple naive model. We simply predicted all future sales with a linear model by using previous week sales without any adjustments. The average WAE for this approach was **2813**. We didn't expect this model to be great as intuitively there are not any trends detectable from only using the previous week. The model does not take into account factors such as seasonality or sales periods which have more predictive power.

3.2 Better Naive - Matching Week from Previous Year Model

With the considerations for improvements from the naive model, we considered a second model that predicted sales by using the same week from the previous year instead of just the previous week. This helps address the issues from the naive model with some type of predictability. Intuitively, one would think seasonality and sales would be reflected in the previous year. For example, weeks of holidays will have more sales than regular weeks of the year. The average WAE for this approach was **1888**. However, this model is still naive and we can do better as seasonality and sales periods often last longer than a week.

3.3 Linear Model

We decide to use a linear model using the weeks and years as categorical and numerical variables respectively. This model was also suggested from Dr. Liang. By doing this, we are able to achieve the same intuition from matching the weeks from a previous year, but also potentially capture longer lasting effects and trends of an overall year. The average WAE for this approach was **1660**. This was the bare minimum of a linear model and we added some basic improvements which is discussed in the next subsection.

3.4 Better Linear Model

This approach is an extension of the previous model described. For this approach we only trained the model using the data from the same store/dept. We predicted future weekly future sales using historical weekly sales data and the store/dept. We also applied a square root transformation for the weekly sales response variable, we did this reduce skewness in data. For missing data, we replaced those values with zero; we also replaced any negative values with zero. The average WAE for this approach was **1622**.

4 Including Pre-Processing

The main pre-processing of data we did was to use SVD. The intuition for this is to reduce noise. With 45 stores, it's very likely that the "signals" that we care about can be reduced to much smaller dimensional that will work better for a generalized predictive model. We used Dr. Liang's suggestion of 8 principal components. We also extracted year and week to predictor "Yr" and "Wk" from Date. And to make the weeks line up for Year 2010 we subtracted 1 from Wk.

4.1 Final Model - SVD Better Linear Model

The final model we used was our better linear model with SVD pre-processing. The performance of this model was an average WAE of **1580.**. The basic implementation was to call SVD pre-processing on the training sets before the data was extracted and prepped for predictions by our better linear model.

5 Hardware/Run Time/Results

System Model	Processor	Memory
ASUS ZenBook	AMD Ryzen 5 4500 CPU @ 2.38 GHz, 2375 Mhz	8 GB

Model	Prediction Time	Avg. WMAE
Naive	1.583 secs	2812.677
Better Naive	1.543 secs	1888.006
Linear	10.005 mins	1659.709
Better Linear	10.153 mins	1622.649
Final	10.875 mins	1580.127

6 What We Learned / Challenges / Conclusion

It was an interesting step by step process as we proceeded between models. The naive model with no intuitive predictability served as a good baseline. Matching the weeks up from the previous year provided the biggest jump in improvement between the models. Moving to a linear model with provided another significant jump in improvement, but took considerably longer to run. The linear model added another level of predictability that got us closer to our goal. However there still needed to be fine tuning as we ended up with a linear regression model that trained sqrt(Y) Yr + Wk for each Store/Dept. Finishing by taking the square root of response variable Y in order to reduce skewness of data. SVD provided the final improvement that we needed. Generalizing the model so that we focused on just the signals that likely had predictive power filtered out the noise that may have been varying some of the predictions. For the future, we can also look at other pre-processing techniques and compare.

One challenge that we noticed was that Fold 5 seemed higher than others in terms of performance. This was discovered because of some variation of holiday placement within varying weeks. Although we didn't adjust for this in our model, it brought up the dealing with imperfect matches by week in the training data. This is something to look into for further improvement; one suggestion that was recommended by the official Kaggle winner was to make a post-prediction adjustment. Another adjustment that could be considered would be to add weights to the training data for those respective weeks that are affected that smooths them out to be more consistent with the rest of the folds.

In conclusion, we found that a combination of intuitive predictability and generalization through SVD provided a suitable model for this project.