XXX NetID: xxx

Fall 2021 STAT 542 — Project 2 Report

1 Overview

1.1 Goals

In project 2, we work with the sales data for 45 Walmart stores located in different regions. In more details, the data set contains 421570 sample data points and 5 variables. These variables are: the store and the department number, the date and corresponding weekly sales (response variable) for the given store and department, and a categorical variable indicating whether the given week is a special holiday week. Our goal is to build a model to predict future weekly sales for each department in each store given the historical data.

To simulate the usual practical settings of machine learning models, we will be training a model where we continuously add in data to improve the accuracy of our prediction of the future weekly sales. In particular, we will start out with a training set containing data from February 2010 to February 2011. Our testing data set will contain observations from March 2011 to October 2012. The testing set will be further divided into 10 folds, each of which contains data of two months starting from March 2011. For example, fold 1 will contain data points of March 2011 and April 2011. We will start out by using only training data set to train our model, and then perform inference on test data (associated with fold 1) with the trained model. Next, we will add the data from fold 1 to the training data, retrain the model, and predict fold 2, which has data of May and June 2011. We iterate this process until the 10-th fold.

The model performance will be evaluated using weighted mean absolute error (WMAE), which is defined as below:

$$\text{WMAE} = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i |y_i - \hat{y}_i|$$

Here n is the number of data points, y_i is the true sales and $\hat{y}_i$ is the predicted weekly sales. Moreover, w_i are the weights, and we assign higher weights for holiday weeks including: Super Bowl, Labor Day, Thanksgiving, and Christmas. To be count as qualified, our model needs to achieve an overall average of WMAE below 1610 taken over the 10 folds.

1.2 Contribution

I work on project 2 individually and there is no other team member. My NetID is **xxx**.

2 Technical Details

2.1 Data details

In this section, we provide more details into the testing and training data sets the we briefly mention above. In particular, we have:

- Training data: contains 5 variables “Store”, “Dept”, “Date”, “Weekly Sales”, “Is Holiday”. The data points’ time frame ranges from 2010-02 to 2011-02.

- Test data: contains 4 variables “Store”, “Dept”, “Date”, “IsHoliday”. The data points’ time frame ranges from 2011-03 to 2012-10.
- Folds data (10 folds from Fold 1 to Fold 10): contain the same 5 variables as training data. Each fold (starting from Fold 1) contains data for every two months starting from 2011-03 to 2012-10.

2.2 Common Data Pre-processing

First of all, we note that the train and test data sets share similar processing steps except for the SVD part, which is for train data only. In this section, we describe the general processing steps and will go into more details (including SVD) in the methodology part (2.3). Moreover, we note that the test data for each iteration is associated with each fold described above, which consists of observation of two month each from the original testing data.

- For each (department, store) combo, we can extract the historical weekly sales data, which is the response variable Y.
- Since the dates are not same across each fold, we find the associated (test) dates for each fold that needed prediction.
- Next, we find the unique pairs of (Store, Dept) combo that appeared in both training and test sets, i.e the one that needed prediction. We create the training data set and testing data set using these unique pairs and the next steps in 2.3.

2.3 Methodology

SVD step: This step is applied to training data to remove noise and retrieve all the top principle components. So we can work with a smaller, less noisy data set while being able to keep the essential information. In particular, we will be performing SVD step on the original training data, and the result would be passed onto the processing steps above as actual training data. To perform SVD, we go through the following steps:

- Arrange data from a particular department as matrix $X_{m\times n}$, where m is the number of stores having this departments, and n is the number of weeks.
- Apply SVD on data where observations are stores and features are weeks.
- Choose the top $d = 8$ principle components and multiply with the original data X to obtain the reduced data set $\tilde{X}$.

Create variables: Once we are done with performing SVD and data processing, we move to creating variables to fit our model. We will be creating two variables **‘Wk’** and **‘Yr’**.

- Variable Wk: The intuition behind this categorical variable is that given a certain week we want to predict (of the current year), we can use the data of the same week from previous year to predicting our sales. So we create a new variable called ‘Wk’, which numbers our weeks in one year starting from 1 to 52 (or 53). Note that since there

migh be mismatch between the last week in 2010 (numbered 53) versus last week of 2011 (numbered 52) while both are Christmas week, we subtract 1 from weeks in 2010.

- Variable Yr: Besides the "Wk" feature, we also create another feature "Yr" that is a numerical feature to fit in our regression model.

Fit model: With categorical variable 'Wk' and numerical variable 'Yr', we fit a linear regression with **Weekly Sales ~ Yr + Wk**. Note that since it's possible that the training data doesn't contain all 52 weeks, we construct the design matrices for both training and testing data, and then fit the model on training data. Then to perform prediction, we replace any NA coefficients with zero and apply the model on testing data.

3 Results

In this section, we include the testing results of WMAE (rounded to 2 decimal places) for the 10 folds in Table 3.1. Moreover, in 3.2, we include the average of WMAE taken over 10 folds and the total running time. The results meet the benchmark of less than 1610 for the average WMAE take over all the folds.

Fold	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
WMAE	1941.58	1363.46	1382.50	1527.28	2310.47	1635.78	1682.75	1399.60	1418.08	1426.26

Table 3.1: WMAE results for 10 folds

Average WMAE	**Total Running Time (in seconds)**
1608.78	74

Table 3.2: Average WMAE result and total running time for 10 folds

To get this result, the computer system we used is Macbook pro with 2.6 GHz 6-Core I7, 32GB memory.

4 Conclusion and Reference

In this report, we follow the approach that is suggested by Prof. Feng Liang from her Campuswire post. We actually started out by testing other approaches that were also suggested by the Professor such as Naive, or Seasonal Naive, and fitting linear regression directly without the use of SVD. However, none of these showed performance as good as the above approach. Through going over all the approaches suggested, we gain two important learning points:

- It can be useful to start with simple model first and learn about the data through these models. Then we can add simple modifications to the model and improve the accuracy as we move from basic to more complex models.
- It's beneficial to understand the data well rather than just trying to apply complex models. Manipulation with the data itself can help improve the model performance even without complicated models.