Project 2 – Walmart Store Sales Forecasting

Fall 2022 STAT 542

Authored By:

Mohammad Hassanpour	MH57
Jason Young	JasonY6
Unnati Narang	Unnati

Abstract

Linear Regression, data pre-processing, a sound prediction process, and the application of principal components were used to produce predictions of Walmart weekly sales data for 10 2-month periods using historical data resulting in the accuracy on test date in Table I, below.

Table I

Model Performance, measured by WMAE, with average and total run time.

Fold	1	2	3	4	5	6	7	8	9	10	**Average**	**Run Time:**
WMAE	1941.58	1363.46	1382.50	1527.28	2056.66	1635.78	1682.75	1399.60	1418.08	1426.26	**1583.40**	**1.82**

Introduction

We built a model to predict Walmart Store Sales using sales data for 45 Walmart stores located in different regions that contains 421,570 observations of store, department, date, weekly sales, and a holiday week indicator. We will predict sales for each department within each store.

First, we will use a training data set with February 2010 to February 2011 data to predict sales in a testing data set with March 2011 to October 2012 data that is divided into 10 two months folds. We initially forecast the first fold using our February 2010 to February 2011 dataset. Then we will add the data from the first fold to the training data set, retrain the model, and predict the next fold and continue until all 10 folds are predicted.

Model performance will be evaluated using weighted mean absolute error (WMAE):

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

where *n* is the number of data points, y_i denotes actual SALES, and $\hat{y}_i$ is predicted weekly sales, w_i are the weights. The holiday weeks, Super Bowl, Labor Day, Thanksgiving, and Christmas are 5 times those of other days.

Methods

Data Pre-processing

The training and testing datasets were preprocessed in the same way except that only the training data was transformed into a dataset of principal components.

1. We extract the response variable, historical weekly sales for each department in each store.
2. We extract a set of weeks from the training dataset that corresponds to the set of weeks that must be predicted in each of the 10 folds to train a model specific to that fold. That is if we are predicting weeks 7 through 14 of 2021, we will use a data set containing weeks 7 through 14 of 2020 to train a model.
3. Next, we extract the data for the pairs of stores and departments that appear in the training and testing datasets we created above.

This results in 10 training and 10 testing datasets each of which has a unique range of dates and store-department combinations.

4. *Principal components-*

We use principal component analysis to retain only a subset of linear combinations of predictors called principal components responsible for most of the variation in the training dataset because we believe the others represent noise.

Procedure:

A. Arrange data from a department as matrix X with stores as rows and weeks as columns.

B. Center: Subtract the predictor's means from their corresponding value. We subtracted the mean of each week's weekly sales from each store's weekly sales for that week.

C. SVD: Decompose the training data matrices into three matrices so that we can extract the singular values of the matrix.
Decompose the centered data matrix

$$X_{m\times p} \text{ as } U_{m\times m} S_{m\times p} V^{\top}_{p\times p} \tag{2}$$

D. Reconstruct a transformed data matrix using the eight largest of these singular values from the S matrix in (2) as our principal components. Use U and V from (2) to create a transformed centered data matrix.

E. Add the predictor's means to their corresponding value. We added each week's mean weekly sales to each store's weekly sales for that week to create the data matrix used to forecast.

.

Fit model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{3}$$

, where X_1 is a week factor and X_2 is a year factor. We constructed the design matrices for both training and testing data, fit the model on training data, replaced "na" values in the testing data with zero then predicted the response values in the test data.

Results

The computer system used for prediction:

MacBook Pro (13-inch, 2020, Four Thunderbolt 3 ports)
Processor: 2.3 GHz Quad-Core Intel Core i7
Memory 16 GB 3733 MHz LPDDR4X

Table I

Model Performance, measured by WMAE, with average and total run time.

Fold	1	2	3	4	5	6	7	8	9	10	**Average**	**Run Time:**
WMAE	1941.58	1363.46	1382.50	1527.28	2056.66	1635.78	1682.75	1399.60	1418.08	1426.26	**1583.40**	**1.82**

Table I reports the accuracy of the model on test data.

Discussion

Understanding the data and making some reasonable guesses suggested data pre-processing to match training data to prediction data, in terms of date range, department and store pairs for each fold. It also suggested fitting a model for each store and department pair. Technical knowledge suggested that with the number of levels we had in each factor Principal Components for would likely help with smoothing or noise reduction.

References

Liang, Feng. "Project FAQ.", *Campuswire.com,* 14 November 2022, https://campuswire.com/c/G3D46BBBA/feed/19.