STAT 542 Project 2 Report

Tianhong Yin (tyin7)
Deli Yang (deliy2)
Rachel Banoff (banoff2; MCS-DS)

Contributions: Tianhong Yin, Deli Yang, and Rachel Banoff all wrote sections of the report. Tianhong Yin and Deli Yang wrote the code. Rachel Banoff reviewed the code and text for consistency with the project requirements.

1 INTRODUCTION

The goal of this project is to predict the future weekly sales for each department in each store based on the data from the Walmart Store Sales Forecasting Kaggle challenge. The historical data is from 45 Walmart stores located in different regions and is composed of weekly sales data for various stores and departments from February 2010 to February 2011. The data set has 164,115 observations and has five variables: "Store" (nominal), "Dept" (nominal), "Date" (discrete), "Weekly_Sales" (continuous), and "IsHoliday" (binary). The dataset includes selected holiday sales events, making it more challenging to predict sales.

The prediction model performance is based on the evaluation of the weighted mean absolute error (WMAE), which is defined as:

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

Where n is the number of data observations, y_i is the true weekly sales, $\hat{y}_i$ is the predicted weekly sales and w_i is the weights assigned. Higher weights are assigned to the following four holiday weeks: Super Bowl, Labor Day, Thanksgiving, and Christmas. To produce desired results, we need to get an average WMAE of the 10 folds smaller than 1580.

2 PRE-PROCESSING

2.1 Missing Values

Since missing data will cause NA coefficients in the model, we assign zeros to missing values.

2.2 Dimension reduction

Single value decomposition (SVD) was used to reduce the ranking of the training data set. The goal is to reduce noise in weekly variance per department in the training data, which, in turn, should help prevent overfitting.

The training data was first converted into wide formatting by Date and Weekly_Sales, and the missing values were replaced by 0. To implement SVD, we wrote a function that loops over all departments if the row number of a department data is larger than the number of components:

- Extracting data for each unique department and had the data matrix X_{m*n}, where m is the number of stores with a unique department id, and n is the number of weeks.
- Use X - store.mean to derive the updated matrix X.
- Apply the svd() function, where the number of left singular vectors to be computed and the number of right singular vectors to be computed are both the number of component, to get (1) a vector sorted decreasingly that contains the singular values of **X**, (2) a matrix **U** whose columns contain the left singular vectors of **X** and (3) a matrix **V** whose columns contain the right singular vectors of **X**.
- Derived the diagonal matrix $\boldsymbol{D_{r*r}}$, where r < min{m,n} and is the rank of **X**.
- Computing a singular-value decomposition of each matrix: $\mathbf{M} = \mathbf{UD}\boldsymbol{V^T}$+ store.mean. This will be repeated for all other departments and a transformed training dataset will be returned.

2.3 Training Set and Test Set

Similar pre-processing steps are used for both the training set and the test set. For each (store, dept) combo, we extract the historical weekly sales data, which is the response variable Y. The linear model was applied and features included in the model include "Wk", "Yr" and "Yr^2".

- "Wk" is a categorical variable that has 52 levels, which corresponds to the total number of weeks in a year. However, 2010 has 53 weeks, which is not reflected in R package lubricate. This means that the dataset does not extend back to the first week of 2010, and we subtract 1 from "Wk" if a date is in 2010.
- "Yr" is a numerical feature that corresponds to the year of a date.
- "Yr^2" is the quadratic term of "Yr". Adding this term helps improve the performance of the linear model.

2.4 Shift in Folder 5

Fold 5 has the highest WMAE since it contains two holiday weeks and therefore receives higher weights in WMAE. Following the post on Campuswire, we implemented a simple shift trick to define shift = 1/7. For fold 5, when predicting sales at Wk51, we only record 1 - shift = 6/7 of the predicted sales and shift 1/7 to the sales at Wk52. By using the shift trick, prediction accuracy can be further improved. To improve the generalization of the model, it's important to identify the discrepancies between the training and test data.

3 MODEL IMPLEMENTATION

We trained linear models to predict future weekly sales, and evaluated its accuracy using a 10-fold split of the test set. With the features selected and the preprocessing process, we fit a linear regression model: Y ~ Wk + Yr + Yr^2 for each (store, dept) combo. It is more efficient to first identify all unique pairs of (store, dept) in both training and test data and filter the relevant data based on this discovery. Next, we used the fitted model to make predictions on the corresponding pairs (store, dept) in the test data and were stored off in a pre-allocated list.

It is possible that the training data do not contain all 52 weeks. To avoid any errors from running lm(Y = Yr + Wk), we construct the design matrices for both training and test, fit the regression model, replace any NA coefficients with zero, and then compute prediction using ordinary matrix computation instead of the built-in predict function.

4 MODEL PREDICTION ACCURACY AND RUNNING TIME

Below are the WMAEs over 10 folds:

Fold	Test WMAE
1	1941.581
2	1363.462
3	1382.497
4	1527.280
5	2056.657
6	1635.783
7	1613.560
8	1354.611
9	1336.703
10	1333.700
Overall Average	1554.583

It is interesting to note that most of the folds have relatively low WMAE values. The one exception was fold 5 which happened to contain two holiday weeks.

We run the models on a MacBook Pro:

> macOS Monterey (version 12.1), MacBook Pro (M1, 2020), Apple M1 Chip, Total number of cores: 8 (4 performance and 4 efficiency), Memory 16GB.

The total running time over all 10 folds is: 137.3812.

Preprocessing data involves interpolating missing data, denoising, and encoding categorical variables. This is the most challenging part of this project. Based on historical data from 45 Walmart stores, a linear regression model was developed to predict future weekly sales data. To reduce variance and improve accuracy of predictions, Singular Value Decomposition was used with the training data along with the model that was developed. An average WMAE of 1554 was achieved across the 10 folds, which is less than the 1580 target value.