PSL (F20) Project 2

Vijayakumar Sitha Mohan (VS24), Waitong Matthew Leung (wmleung2)

Introduction

- The goal is to predict the future weekly sales for each department in each store based on the historical data for 45 Walmart Stores located in different regions. With given train_ini.csv, the data till 2011-02, we need to predict the weekly sales for 2011-03 and 2011-04. After that provided with the weekly sales data for 2011-03 and 2011-04 (fold_1.csv), and we need to predict the weekly sales for 2011-05 and 2011-06, and so on.
- t = 1, predict 2011-03 to 2011-04 based on data from 2010-02 to 2011-02 (train_ini.csv);
- t = 2, predict 2011-05 to 2011-06 based on data from 2010-02 to 2011-04 (train_ini.csv, fold_1.csv);
- t = 3, predict 2011-07 to 2011-08 based on data from 2010-02 to 2011-06 (train_ini.csv, fold_1.csv, fold_2.csv);
- ……
- t = 10, predict 2011-09 to 2011-08 baesd on data from 2010-02 to 2011-08 (train_ini.csv, fold_1.csv, fold_2.csv, …, fold_9.csv)

Overall Approach

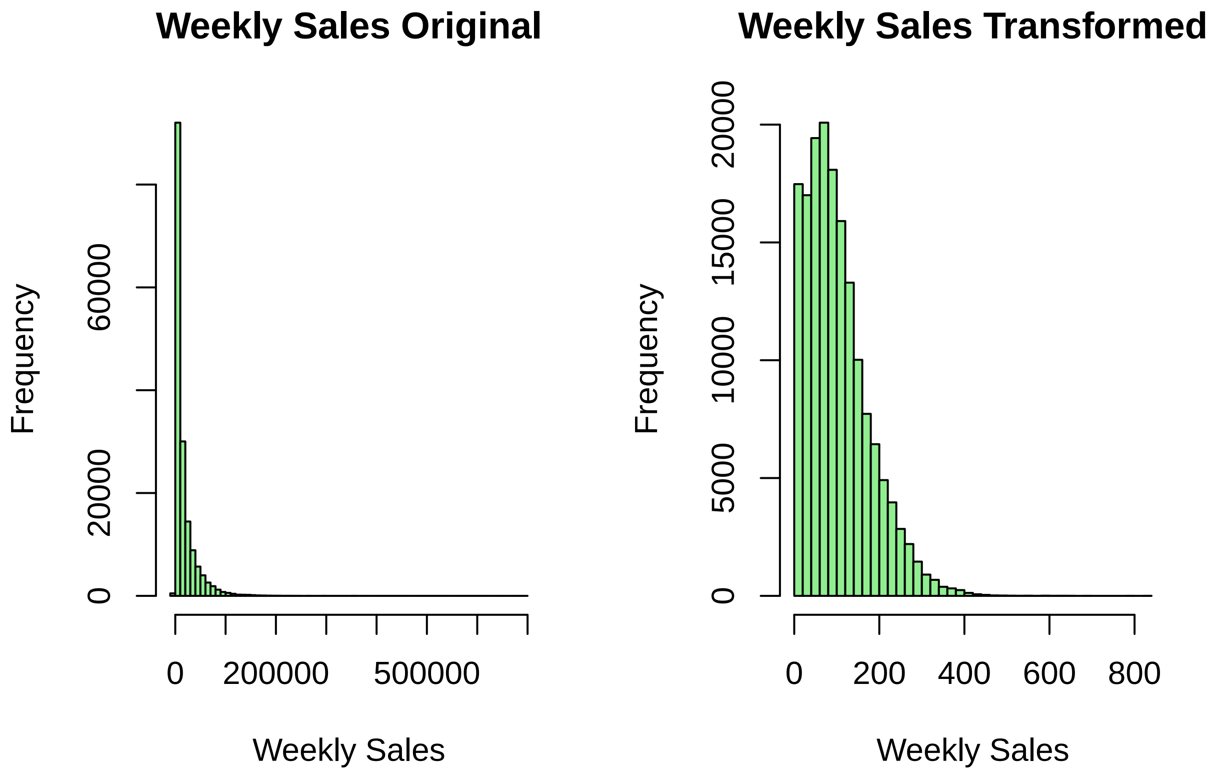
- Our approach was an extension to the solution(III) provided by the professor which trains the linear model for each Store, Dept combination. We use the historical weekly sales data to predict future Weekly_Sales (response variable Y) by Store/Dept combination.
- The idea is to predict each week by the same week from last year for each combination of Store/Dept. Since we train each linear model with only data from the same Store and Dept, we don't need to include Store/Dept in the predictor. Since we predict each week by the same week from last year, if this week is holiday week, this week last year was also holiday week. Thus, the linear model doesn't need to include IsHoliday as predictor.
- We derive two predictors, namely "Yr" and "Wk", from original data column Date. "Wk" is the number of the week in each year starting from 1 to 52(or 53). "Yr" is the year from data column Date. We train a linear regression model: sqrt(Y) ~ Yr + Wk for each combination of Store/Dept. We apply square root transformation to response variable Y to reduce skewness of data. It inflates smaller Y but stabilizes bigger Y.

Data PreProcessing

- The following preprocessing applied by the professor in the given solution(iii)
  - Make sure each year to have exactly 52 weeks
  - Not all stores and departments needed prediction

- extract year to predictor “Yr” from Date
- extract week to predictor “Wk” from Date
- check if “Yr” == 2010, subtract 1 from “Wk” since there is one more week in 2010. This is to line up the week from last year. The most important thing is to line up holiday weekend as a result.
- set “Wk” as factor variable with 52 levels, i.e. one per week.
- since there is some missing data, we use model.matrix() to build design matrix and replace any NA with zero before feeding to lm() to prevent erroring out lm().

- In addition, we applied
  - Replace any negative response variable Weekly_Sales with zero.
  - Apply Square root tranformation for response variable Weekly Sales.
- Scatter plot to show Outlier & Skewness in Weekly Sales of training dataset



Results

- 10-Fold Weighed Mean Absolute Error (WMAE)

Fold	WeightedMeanError
1	1981
2	1450
3	1423
4	1547
5	2276
6	1632
7	1680
8	1404
9	1426
10	1407

- Mean WMAE

Mean_WAE
1623

- Running time: System: MSI laptop, Intel i5, 2.0GHz, 8GB, Win 10

Run.Time
8.388 mins

Discussion

1. Log Transformation of Weekly Sales didn't yield desired result instead WMAE increased.
2. Since the given dataset didn't have many features as opposed to original dataset, it wouldn't be possible to predict negative weekly sales. So replaced negative weekly sales with zero to reduce WMAE.
3. Applied square root transformation yielded the desired with mean WMAE 1623 which was desirable.