PSL (F20) Project 2

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

Introduction The goal is to predict Walmart Store Sales Price for the given train_ini.csv, the data till 2011-02, you need to predict the weekly sales for 2011-03 and 2011-04. Then you'll be provided with the weekly sales data for 2011-03 and 2011-04 (fold_1.csv), and you need to predict the weekly sales for 2011-05 and 2011-06, and so on.

Overall Approach Our approach is to train a 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 We preform the following data preprocesssing before model training:

- replace any negative response variable Weekly_Sales with zero
- 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()

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
8.322 mins

Discussion We made couple of changes from Professor suggested approach in order to meet the bench mark WMAE, namely to handle negative Weekly_Sales and square root response variable transformation.

We tried to replace negative Weekly_Sales with 1 but WMAE got worse. We settled with replacing by 0 which improve WMAE.

We tried log response variable transformation but WMAE got worse. We settled with square root response variable transformation.