Project2 - Walmart Sales Forecast – CS513 -FA20

Saurav Chetry
schetry2@illinois.edu

Dataset, Exploration and Goals:

The dataset for this project is the popular Walmart Forecasting which was introduced in 2012 by Walmart for their recruiting challenge. The dataset consists of sales history for their 45 stores which contained 65 departments. The historical data ranges from February 2010 until February 2011.
I explored the data for NAs and assigned 0 appropriately. The goal was to predict the sales from the stores for 2 months starting from March 2011 & April 2011. In the subsequent iterations the train data was appended with actual Weekly_Sales from March 2011 & April 2011, and the models are to be trained to forecast Weekly_Sales for May 2011 and June 2011. This is done iteratively using Weekly_Sales data provided as 10 folds for dates until 2012-09 to 2012-10.
Weighted Mean Average Error was uses as the metric to evaluate the models prediction performance.

Implementation:

Summary: I used two implementations from Piazza post "What we have tried (II)" and "What we have tried (III)" (the TA's modified version) for all weeks and all folds. Shifted the weekly predictions for fold_5 weeks with Holidays and finally averaged the forecasts for all folds to achieve a **mean WAE of 1612**.

Model details:

What we have tried (II) model is an extension of the naïve model where for each week we need to predict, we use the "same" week from last year. We create a new variable called "Wk" to store values 1 to 52 by using the function "week" in package "lubridate". These values are used in predicting sales in the future, corresponding to the same week number in the previous year. However, as 2010 and 2011 had 53 and 52 weeks respectively, this model substracts 1 week in 2010 to align the weeks in both years. As historical data starts only from Feb 2010, substracting 1 week is acceptable. To avoid any rounding issues due to the way weeks are counted, a slightly higher than 365 days was subtracted for finding the matching starting week from previous year; similarly, a slightly lower than 365 days was subtracted for finding the matching end week from previous year.

What we have tried (III) model uses Y response as weekly Sales data, Yr as numerical feature for year and Wk as categorical feature for Week to compute the linear model lm(Y = Yr + Wk)", to predict future sales for the department, store combination. The design matrices for both training and test were created to avoid any errors corresponding to folds not having 2 weeks. NAs were replaced with zeros.

Shifting:

The instructor's provided models did not account for shifting the forecasts to accurately match the dates from previous year. Holidays occur in the weeks 49 to 52 and for those week's predicted Weekly Sales, I used the circular shifting of the weekly predictions to adjust the future predictions in sync to the exact Holiday Sales from previous year's weeks.

Shifting was done to after the forecasts were calculated from both models. As week numbers in 2011 and 2012 contain different dates, it was required to make Holiday week predictions for precisely the same dates that corresponded to the dates from the previous year.
Fold 5 had dates which had the holidays. Weeks 49,50,51,52 have column IsHoliday as True.
I used $1/7*(previous\ week\ pred) + 6/7\ *(current\ week\ pred)$ to calculate the matching prediction for the new holiday week.

Averaging: After shifting the forecasts for fold 5, the forecasts from both models were averaged by doing a simple means calculation.

Results & Statistics:

RunTime:	1.261687 mins
platform	x86_64-w64-mingw32
year	2020
R version	4.0.2 (2020-06-22)
System:	
Manufacturer:	Windows 10 – SAP Release 9.0 in-place upgrade
Processor:	Intel(R) Core(TM) i5-5300U CPU @ 2.30GHz 2.29 GHz
Installed memory (RAM):	8.00 GB (7.88 GB usable)
System type:	64-bit Operating System, x64-based processor

Accuracy on each of the 10 folds:

Fold 1	1985.681
Fold 2	1479.666
Fold 3	1484.015
Fold 4	1534.974
Fold 5	1972.567
Fold 6	1538.227
Fold 7	1743.826
Fold 8	1463.208
Fold 9	1472.048
Fold 10	1446.211
Average All Folds	**1612.042**

Acknowledgement: Instructors for sharing the sample started codes, previous sample reports, OH discussions and Piazza discussions from everyone.