Project 2
Wallmart Store Sales
CS598 Fall 2020
oorona2 - Oswaldo Orona

Prediction Models

After experimenting with multiple models using the forecast package I focus on using the following two models

1. Naive Seasonal. This method is very simple but with weighted average with lm I got decent results.
2. Linear Regression. The main Model for this prediction and used in combination with Naive Season to get some default values and a weighted average

Data Manipulation

The main challenge of this project was able to process the data in a way what would make the use of the forecast efficient as well as allow multiple test and faster development cycles. Since there are lot of stores and Dept this becomes a significant problem for a efficient algorithm.

All the processing was done using Tidyverse and data processing similar to Database Data transformations techniques such as joins, distinct and pivot operations.

- Joins. These operations are mainly used to make sure that the data is complete. Some training data is does not have data for all the weeks and test data does not predicts every single week. The forecast packages requires to have data to have the proper frequency of 52 week for the prediction. This was achieved my a join with a list dates for the data period.
    - Full Joints (cross join) were used to get a full table with Store,Dept and Date.
    - Left joins were used to get the empty(zero) for the missing Weekly Sales.
    - Inner joins were used filter the data back to only the values int the test set.
- Distinct. These operations are mainly used to select a unique set of data such the prediction weeks or the combination of Store and Dept.
- Pivot. These operations to avoid having to do small operation to select specific data for a combination of Store and Dept.
    - Pivot wider. This was used to get all the associated values for a Store, Dept combinations in a single column allowing a faster operation during the selection of the data for the processing of the models.
    - Pivot longer. Once the data was been process and predicted values are created we need to return the data back to the original format for the evaluation of the code.

Finding models

The first success model was a single lm model using the tslm from the forecast package. These model was using season and trend ***data_ts ~ season + trend.*** After testing other single models without much improvement I started experimenting with combination of models.

The proper model combinations were found using a cartessian search of the weights for each model

Interesting Findings

I found then that a average of prediction for linear and Seasonal naive produces acceptable results but the biggest surprise was that the best results for my project were encountered with a weighted average for the combination of Seasonal and linear.

Some extra manipulations of data such as avoiding negative values from lm models add a small improvement to prediction but add around 20% processing time for each split. This was caused by the large number of operations over relative small number of predictions per Store and Dept.

Results

Below is a table with all the results of this project.

Split	WMAE	Time(s)
1	1962.687	19.821
2	1404.143	19.676
3	1412.367	19.718
4	1539.074	20.224
5	2274.898	20.470
6	1561.262	20.425
7	1685.636	20.376
8	1400.858	20.404
9	1414.671	20.535
10	1411.977	20.509
Mean	1606.757	20.216

Equipment

These results were obtained on a desktop computer running Ubuntu 20.04 using a processor AMD 3950x 3.8Ghz Base clock and boost of 4.6Ghz with 16cores 32 threads, 128Gb of ram.