CS598 - Project 1

Xiaoming Ji

Computer System

Hardware

- Dell Precision Tower 5810
- CPU: Intel Xeon E5-1607 @ 3.10GHz
- Memory: 32GB

Software

- OS: Windows 10 Professional 64bit
- R: 3.5.1
- R Packages:
  - forecast_8.4
  - tidyverse_1.2.1
  - lubridate_1.7.4

Models

3 approaches (total 4 models) are used to generate the prediction:

- Naive model
- Seasonal naive model
- Dynamic model: for fold 1 to 6, regression model (tslm) is used. Start from fold 7, since the training data has more than 2 years of data, STL+ARIMA (method='arima', ic='bic') model is built to make the prediction.

Pre-processing

- Run SVD (first 12 components) on each by-department sales data and then transform it back to the original matrix size.
- Missing value handling
  - Weekly_Sales: replace missing value with 0.
  - IsHoliday: search through the training data to find the IsHoliday of same date. See function: `fill_missing_holiday`

Note: my testing show more sophisticated imputation approach won't improve the performance.

Test results

Fold	Naive	SNaive	Dynamic
1	2043.412	15282.78	15282.78
2	2551.222	15776.52	15776.52
3	2223.819	15861.54	15861.54
4	2772.357	15390.24	15390.24
5	5147.755	18588.14	18588.14
6	4190.999	15670.65	15670.65
7	2225.886	15723.94	15723.94
8	2103.463	16157.46	16157.46
9	2194.096	15954.42	15954.42
10	2320.815	15686.73	15686.73
Average	2777.382	16009.24	16009.24

Computation time: 194.26 seconds