CS 598: PSL - Project 2 Report

Members: Amy Hwang (ahwang22), Christian Tam (cmtam2), Monil Kaneria (kaneria2)

Contributions: Amy Hwang worked on postprocessing, code refactoring, and report. Christian Tam worked on preprocessing, regression model, and result export. Monil Kaneria worked on preprocessing and report.

Section 1: Technical Details

Data Pre-processing

The data pre-processing pipeline began by reading in the test and train datasets.

Singular Value Decomposition (SVD) was applied to each department in the train dataset to help reduce data dimensionality, diminish noise, and help with the missing values (which were converted to zero).We retained the top 5 components for reconstruction.

Then both resulting datasets' Date column was converted to Week and Year columns for processing.

We then filtered each dataset for shared store-department pairs and filtered out pairs with zero occurrences. Those pairs were then merged with the post-SVD train dataset and test dataset on the Store and Dept columns.

Implementation Steps

The model used for prediction was built using the Ordinary Least Squares (OLS) approach from the Statsmodels library. First, the predictor variables were transformed into the appropriate format and adjusted to ensure that no unnecessary columns were retained. This involved removing columns with zero variance and centering the predictor variables based on their mean.

The training was conducted using the OLS function, and the model was fitted with the preprocessed data. After training, the parameters were accessed and missing values were replaced with zeroes to avoid issues during prediction. A loop was used to iterate over the features to refine the data structure, ensuring proper training. Department-level time-series data were also handled using pivot tables to create store-level panels for accurate sales predictions.

Prediction Post-processing

We implemented the post-prediction shift adjustment suggested by a Kaggle competition winner. This function adjusts weekly sales predictions for specific departments and weeks based on holiday sales surges in December. We looped through each department We shifted the prediction data by 1 week if there was a holiday sales surge where the average sales during the holiday weeks (weeks 2 through 4) was at least 1.1 times more than the baseline (weeks 1 and 5) December. Week 1's sales were preserved.

Section 2: Performance Metrics

Accuracy of Prediction

Fold	1	2	3	4	5
WMAE	1950.38	1375.48	1394.56	1540.03	2030.38

Fold	6	7	8	9	10
WMAE	1640.75	1692.46	1412.62	1423.20	1439.09

Overall average WMAE: 1589.894

Execution Time

Fold	1	2	3	4	5
Time (s)	38.57	41.06	42.66	43.63	105.15

Fold	6	7	8	9	10
Time (s)	44.50	44.59	46.11	47.02	46.65

System Details: Windows desktop, 3.6 GHz, 32GB memory