**PSL Project 2 Report**

**Authors:**

* Qi Zhou - qizhou8 - MCS-DS
* Yogananth Mahalingam - ym24 - MCS-DS
* Derek Zhang - derekz3 - MCS-DS

**Contributions:**

Qi, Yoga and Derek worked on it independently. Brainstormed together & created the final version.

In this project, we use a linear model to forecast the future weekly sales for every department in each store based on historical sales data from 45 Walmart stores spread across different regions.

The dataset is from https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting

**Section1:**

**Data Pre-Processing**

We conducted the following preprocessing steps for the training and test datasets in order:

Smooth training data with PCA

* We implement PCA to each department, treating observations as stores and weeks as features. We retain the top 8 principal components, set the rest diagonal values to zero, then project it back into the original space. This process is employed to reduce the noise from consistent sales trends within a department across various stores.
* We exclude the “IsHoliday” column during PCA transformation and add it back in after PCA only for the purpose of later WMAE calculation.

Filter train data and test data:

* We retain only the “store” and “department” combinations that are present in both the training set and the test set.

Add new features for both train set and test set:

* We add a "Wk" column to assign numerical representations to each week of the year, ranging from 1 to 52, which is treated as a categorical variable and encoded to be binary variables using patsy.
* We introduce a "Yr" column, derived from the "date" column, and treat it as a categorical variable.
* We add the quadratic term of Yr under the “np.power(Yr, 2)” column using patsy.

Design matrix for model iterations on both train set and test set:

* We group the train set and test set by unique combinations of the "Store" and "Dept", and store the pairings as keys in a dictionary with the corresponding grouped subset of dataframe as values.

Finalize the input and output of the model:

* We drop three columns "Weekly\_Sales", "Store", "Dept" and use the remaining variables as input for the model.
* We fill in missing values of the model predictions with 0.

Adjustment post-prediction for fold 5

* We shift 1/9 of the sales from week 51 to week 52 to account for the inclusion of two holiday weeks in fold 5, mitigating its high WMAE.

**Model:**

We use Ordinary Least Squares from statsmodels as the training model to produce coefficients.

**Section 2: Performance Metrics**

* Accuracy: We evaluated our model performance for each fold based on the weighted mean absolute error, with weight set at 5 for a holiday week, 1 otherwise. We achieved an average WMAE of **1570.092** over the 10 folds. The results for each fold are as follows:

wae\_by\_fold\_1=1941.581

wae\_by\_fold\_2=1363.493

wae\_by\_fold\_3=1382.461

wae\_by\_fold\_4=1527.275

wae\_by\_fold\_5=2210.984

wae\_by\_fold\_6=1635.292

wae\_by\_fold\_7=1613.891

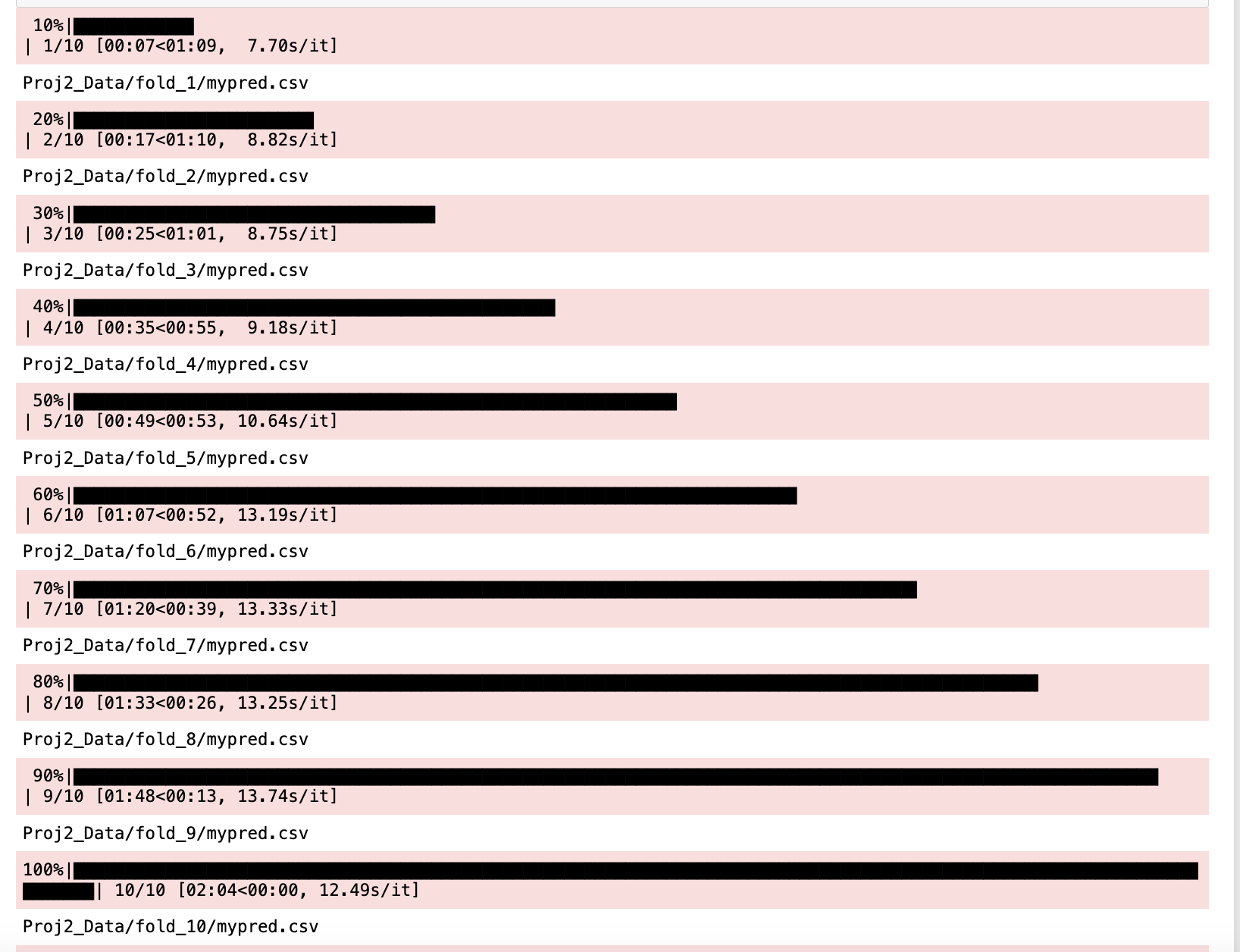
wae\_by\_fold\_8=1355.014

wae\_by\_fold\_9=1336.916

wae\_by\_fold\_10=1334.010

**overall wae=1570.092**

* Execution time: The time taken for each training/test split is displayed below.



* Computer system: The following outcome is derived from a computer system with the specifications: Macbook Pro, Apple M1 Pro, 32 GB memory