Technical Report: Walmart Sales Prediction

John Zeiders

Section 1: Technical Details

1. **Data Processing and Dimensionality Reduction:**
   - Implemented a department-wise SVD (Singular Value Decomposition) approach using PCA
   - Set to maintain 8 principal components per department
   - Applied mean-centering by store before PCA transformation
   - Handled cases where PCA was not applicable (departments with insufficient samples) by retaining original data
   - Data was reconstructed back to original format after dimensionality reduction
2. **Feature Engineering:**
   - Created temporal features from the date column:
     - Week number (1-52) encoded as categorical variables
     - Binary indicators for years (2010, 2011, 2012)
   - Applied one-hot encoding to week numbers to capture seasonal patterns
   - Maintained store and department as categorical identifiers
3. **Model Implementation:**
   - Utilized Ridge Regression with the following specifications:
     - Alpha (regularization strength): 0.01
     - Random state: 42 for reproducibility
   - Applied separate models for each store-department combination
   - Implemented zero-floor constraint on predictions (no negative sales)
   - Built pipeline to handle training and prediction for each store-department pair independently
4. **Data Handling:**
   - Ensured prediction only for store-department pairs present in both training and test sets
   - Maintained original date and holiday flag information
   - Handled missing predictions through left merge with test data

Section 2: Performance Metrics

System Specifications: MacBook Pro, M2 Max, 32GB memory

Performance across 10 folds:

Fold	Training Samples	Test Samples	Weighted MAE	Runtime (s)
fold_1	164,115	26,559	1,918.42	4.62
fold_2	190,674	23,543	1,353.89	4.77

Fold	Training Samples	Test Samples	Weighted MAE	Runtime (s)
fold_3	214,217	26,386	1,377.22	4.96
fold_4	240,603	26,581	1,517.96	5.07
fold_5	267,184	26,948	2,290.46	5.65
fold_6	294,132	23,796	1,904.55	5.28
fold_7	317,928	26,739	1,608.17	5.58
fold_8	344,667	26,575	1,349.78	5.69
fold_9	371,242	26,599	1,334.48	5.54
fold_10	397,841	23,729	1,329.68	5.48

Summary Statistics: - Mean MAE: 1,598.46 - Standard Deviation: 332.34 - Minimum MAE: 1,329.68 - Maximum MAE: 2,290.46