Homework 6

Christopher Williams

October 29, 2025

1 EDA

1.1 Shelve Location vs Sales Analysis

1.2 One-Hot Encoding

1.3 Train-Test Split and Standardization

Feature Selection & Prediction

2 Backward Stepwise Regression

2.1 Elimination Process Table

2.2 OLS Regression Summary

2.3 Final Regression Model

2.4 Prediction vs Test Set

2.5 Mean Squared Error

3 PCA

3.1 95% Variance Explained

3.2 Cumulative Variance Plot

3.3 95% Variance Threshold

3.4 What Does PCA Say?

PCA has told us that of the 11 original features, we need 8 principal components to explain 95% of the variance. This is indicating that our data has some redundancy/correlation, so after we transform the data into a new coordinate system, we can sufficiently use 8 dimensions instead of 11. However, since PCA gives us a weighted combination of all original features, and not the individual features, it doesn't tell us explicitly which features to remove, but has created new features for us to use. To learn which features we should remove, we should use backward/forward stepwise regression, or use a Random Forest model to find which features to split on.

4 Random Forest Analysis

4.1 Feature Importance Plot

4.2 Feature Selection Comparison

4.3 OLS Regression on Selected Features

4.4 Prediction vs Test Set

4.5 Mean Squared Error

5 Comparison of Feature Selection Methods

5.1 Metrics Comparison Table

5.2 Recommended Method and Features

6 Prediction Interval

6.1 95% Prediction Intervals

6.2 Prediction Interval Plot

7 Polynomial Regression and Grid Search

7.1 Grid Search with RMSE Minimization

7.2 Optimum Polynomial Order

7.3 RMSE vs Polynomial Order Plot

7.4 Training and Prediction

7.5 Mean Squared Error

8 Simple Linear Regression Proof

In a simple linear regression with n observations:

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{1}$$

Prove the following:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the sample mean.