Retail Task:

Introduction

The dataset contains retail sales records with different product and store features. The target variable is sales. The objective is to predict sales based on available predictors.

Methodology

Exploratory data analysis was conducted first. Scatterplots and histograms revealed that the variables were uniformly distributed, with no strong linear or non-linear relationships. A correlation heatmap confirmed the absence of meaningful correlation between predictors and sales.

For preprocessing, missing values were imputed using column means. Duplicates were also removed. Categorical variables were converted into numerical representations through one-hot encoding, though this had no significant effect due to the lack of correlation.

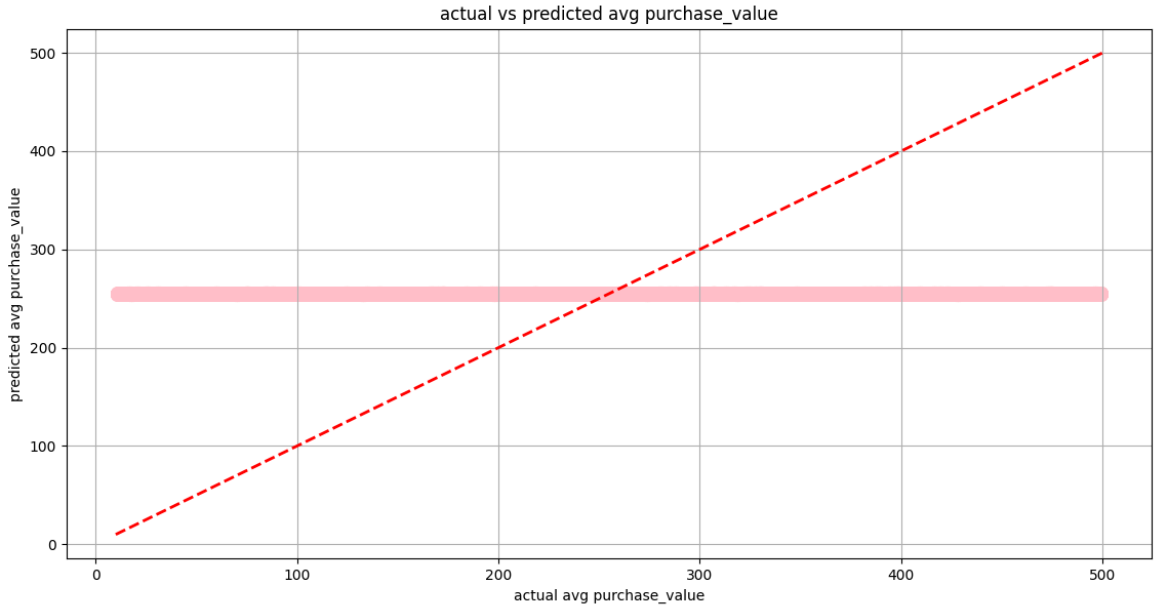
Experimentation

Multiple models were tested, including linear regression and regularized regressions. Feature selection and different combinations of variables were attempted. However, all approaches produced very similar outcomes, as the data carried no predictive signals.

Results

Performance was evaluated using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2. All models performed essentially the same, with R^2 close to 0, indicating that the prediction was no better than predicting the mean.

Model Name	MSE	RMSE	R^2
Model 1	20008	141.6	0
Model 2	20008	141.6	0
Model 3	20007	141.45	0



The best approach therefore was simply predicting the mean sales value for all cases.

Explanation

Since the data follows a near-uniform distribution and predictors are uncorrelated, no meaningful model could be learned. Regularization made no difference, as there were no dominant predictors to emphasize or suppress. Essentially, sales are random with respect to the features provided, and the mean is the most reliable estimate.

Challenges

The main difficulty was the lack of correlation or patterns in the dataset and the data being uniformly distributed. Along with this the sheer size of the dataset meant that the model would be slow. Traditional preprocessing and modeling steps did not help improve predictions. Unlike life expectancy data, where strong drivers like schooling or income exist, this dataset offered no clear signal.

Conclusion

The retail prediction task highlighted the limitations of modeling when the data contains no correlations or predictive structure. The best-performing strategy was to predict the mean sales value, as no algorithm could outperform this baseline. Possible improvements would require richer datasets with more meaningful features.

References

😭