Product Performance Analysis

Overview

In this assignment we analyzed supermarket product sales data using machine learning techniques. We implemented K-means clustering to discover product groupings and used regression models to predict product performance. This project emphasizes understanding machine learning concepts, proper data preprocessing, model comparison, and clear visualization of results.

Data Preprocessing

For missing values we use ChatGPT to make a function that counted the missing values and computed the fraction of the missing data per each feature.

Handling missing values Any rows with 50% missing values or mor were dropped in order to avoid unreliable imputations We used the median as an input for the numerical features, since it is resistant to outliers The mode was used for any categorical features which kept category consistency. Outliers were found using the IQR method - Lower bound = Q1 - 1.5 * IQR Upper bound Q3 + 1.5 * IQR. Instead of removing them we just capped them at the IQR boundaries. We used Z-score standardization on all the numerical features since k-means relies on Euclidean distance, this makes it so features like price and units_sold all contribute equally.

K means Clustering

ChatGPT helped a lot with the implementation of k-means clustering. We knew to randomly assign K centroids and then find the distance from the centroid to each point. From there we would assign the points to a centroid and then recalculate the centroid. This would

repeat until the centroids can no longer change values. For optimal K selection we used WCSS to see which K would be the most viable since its how you find the best fit. For the cluster analysis we computed different stats such as the average pricing, average profit, number of items, etc and then categorized each cluster into a different preference based on the results from the previous stats.

Regression Analysis

For the models chosen we used Linear Regression and Polynomial Regression as those were the 2 discusses in our powerpoints. To train our models we used certain stats like costs, price, units, and compared their profits to see how each model works. To compare the models we used the values of both Mean Absolute Error and Mean Squared Error and realized that the polynomial regression model works best as it can better capture the relationships. Linear Regression seemed to show signs of underfitting while Polynomial Regression seemed like the best fit. If you see through the websites, you will be able to notice the various visuals we added, whether it be charts, graphs or tables.

Conclusion

Some things we noticed were that data preprocessing improved the model stability. K-means clustering showed clear groups that were different in significant ways. The polynomial regression outperformed the linear regression for profit prediction. Some limitations that we had when working was, the k-means used predefined k and assumed spherical clusters. The polynomial degree was also manually selected. Some improvements we could do is add features to the dataset like dates, this could improve the prediction accuracy. We could also improve the

dashboard, like adding filters for the dataset, having live plots that change with inputs, and a dropdown to choose the k for clustering. We used AI in this project to help us out with some machine learning concepts and how to implement them into or code, debugging our code when we ran into errors, improving the way our code was organized, and designing the dashboard for users.