Homework 1

Spencer Au

Introduction

We are using both a Linear Regression and a Polynomial model to help predict the average amount a customer spends in a year given their gender, age, height, waist size, inseam length, whether they are in the test group, their salary, the months active, year, and the number of purchases. This model could be useful to the store if they want to know how much a customer will spend in a year given their information. This could help the store determine how much they should spend on advertising to a customer, or how much they should spend on a customer to get them to come back to the store.

Methods

Models:

- **Linear Regression:** Linear relationship between customer data and spending.
- **Polynomial Regression:** Captures non-linear spending patterns.

Data Preprocessing:

- Cleaned data by removing missing values and resetting indices.
- Split data into training (80%) and testing (20%) sets.
- Standardized continuous variables and used One-Hot Encoding for categorical variables.

Performance Assessment:

- **Mean Squared Error (MSE):** This metric quantifies the average squared difference between the predicted and actual spending values. A lower MSE indicates a better fit of the model to the data.
- **Mean Absolute Error (MAE):** It measures the average absolute difference between the predicted and actual spending values. Lower MAE signifies better accuracy.
- **Mean Absolute Percentage Error (MAPE):** MAPE calculates the percentage difference between predicted and actual values. It provides insight into the accuracy of our predictions in percentage terms.
- **R-squared (²):** This metric assesses the proportion of variance in the dependent variable explained by the independent variables. A higher ² value indicates a better-fitting model.

Results

How well did your model perform according to the various metrics, was the model overfit (how can you tell)? What do those performance metrics tell you about the model? Did you need `PolynomialFeatures` (which includes both ploynomial features and interactions)? How much do you trust the results of your model (in other words, would you be confident telling the store that they should use the model? Why or why not? Are there any caveats you'd give them?) Also answer the two questions you chose from part 2 above. Include the image, a caption as well as your written answer.

Linear Regression Model:

- **Training Set**
  - MSE: 13005.32355
  - MAE: 90.10456
  - MAPE: 13005.32355
  - R-squared: 0.52338
- **Testing Set**
  - MSE: 12872.35241
  - MAE: 89.26341
  - MAPE: 12872.35241
  - R-squared: 0.51605

Polynomial Regression Model:

- **Training Set:**
  - MSE: 3026.97443
  - MAE: 43.96177
  - MAPE: 3026.97443
  - R-squared: 0.88850
- **Testing Set:**
  - MSE: 3201.79839
  - MAE: 45.10156
  - MAPE: 3201.79839
  - R-squared: 0.88209

If you want a table you can make one with this website and paste the markdown table here. For example:

A	B	C	D	E
a	b	c	d	e
a	b	c	d	e
a	b	c	d	e

***Question 1*: Does being in the experimental test_group actually increase the amount a customer spends at the store? Is this relationship different for the different genders??**

Figure 1: This is a clothing store

***Question 2*: In which year did the store's customers make the most money? Were the store's sales highest in those years?**

Figure 2: These are clothes

Discussion/Reflection

A few sentences about what you learned from performing these analyses, and at least one suggestion for what you'd add or do differently if you were to perform this analysis again in the future.