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

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

Linear	Train	Test	Polynomial	Train	Test
MSE	13005.32355	12872.35241	MSE	3026.97443	3201.79839
MAE	90.10456	89.26341	MAE	43.96177	45.10156
MAPE	13005.32355	12872.35241	MAPE	3026.97443	3201.79839
R-Squared	0.52338	0.51605	R-Squared	0.88850	0.88209

How well did your model perform according to the various metrics?

The Polynomial Regression model outperforms the Linear Regression model in all metrics. It has lower MSE and MAE values for both the training and testing sets, indicating that it provides a better fit to the data. Additionally, the R-squared value is higher for the Polynomial Regression model, which means it explains more variance in the target variable.

Was the model overfit (how can you tell)? What do those performance metrics tell you about the model?

The Linear Regression model has lower training set error metrics (MSE and MAE) compared to its testing set error metrics. This indicates there is some level of overfitting, as the Linear Model performs better on the training data vs the test data. On the other hand, the Polynomial Regression model has slightly higher training set error metrics compared to its testing set error metrics, which suggests it is not overfitting.

In the Linear Regression Model, high MSE and MAE values indicate relatively large prediction errors. In addition, high MAPE shows significant percentage errors in predictions, and the low R-squared values suggest a poor fit to the data. In the Polynomial Regression Model, the lower MSE and MAE values indicate smaller prediction errors. The lower MAPE shows smaller percentage errors in predictions, and the higher R-squared values suggest a much better fit to the data.

In summary, the Polynomial Regression model provides a better fit to the data, with lower errors and a higher R-squared value, while the Linear Regression model performs less effectively, with higher errors and a lower R-squared value.

Did you need PolynomialFeatures (which includes both polynomial features and interactions)?

Yes as PolynomialFeatures was applied to the data when training the Polynomial Regression model. This is evident from the significantly improved performance of the Polynomial Regression model over the Linear Regression model, especially in terms of R-squared, indicating that polynomial features and interactions were useful in capturing the underlying patterns in the data.

How much do you trust the results of your model?

I would be much more confident using the Polynomial Regression model compared to the Linear Regression model because it provides a better fit to the data and explains a higher percentage of the variance in the target variable. However, there are still some caveats to consider:

- Overfitting: The Linear Regression Model exhibits some level of overfitting, which means its performance on unseen data may not be as good as it is on the training data. However, the Polynomial Regression model does not seem to be overfitting, which means it may be more robust to unseen data.
- Model Complexity: The Polynomial Regression model may be more complex and computationally intensive due to the inclusion of polynomial features. This could be a consideration if compute power and scalability are important.

In summary, I would generally recommend using the Polynomial Regression model, and keep in mind its strengths and limitations. While it does provide better predictions than the Linear Regression model, it should also be monitored for potential overfitting and used alongside using sufficient coputational resources.

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??

It does seem like for all four gender groups in the experimental test group end up spending more at the store vs those in the control group. In terms of gender, in both the test and control groups, women seem to be the gender that ends up spending the most, while the other three categories (men, nonbinary, other) seem to spend similar amounts regardless of whether they are in the control or test group. All in all, women spend the most, and the experimental test group does seem to spend more than the control group.

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

There seems to be a pretty linear relationship between the year and the self reported salaries of the customers. For example, as the years increased from 2019 to 2022, the customers reported had a consistent increase in salary. The store's sales were also highest in those years, as the store's sales increased as the years increased. This makes sense, as the customers are making more money, they are able to spend more at the store.

Discussion/Reflection

I learned that it is important to consider the strengths and limitations of different models, and to choose the model that best fits the data and the problem at hand. In this case, the Polynomial Regression model was a better fit to the data than the Linear Regression model, and it provided more accurate predictions. In addition, its important to consider what kind of data you have to choose an appropriate graph for said data. In this case, I used a boxplot to show the amount each group and gender spent, displaying the average, 1st and 3rd quartiles, and outliers. I also utilized a boxplot for the salary of customers throughout the years, and used another boxplot ot show the store sales throughout the same years. One suggestion for doing this analysis in the future would be to use data that encompasses more years, as the data only went from 2019 to 2022. This would allow for a more accurate representation of the data, and would allow for more accurate predictions.