Homework 1

Tiffany Le

Introduction

Companies, especially in the clothing industry, actively pursue strategies to boost revenue, enhance customer retention and experience, and offer products aligned with customer trends and consumer profiles. This data analysis aims to predict how much the clothing store's customers will spend with the company per year. To achieve this goal, a dataset, provided by the Clothing Store's company, on their customers provides key independent variables to build two different predictive models. These key independent variables are the customers':

- **gender** Self-disclosed gender identity (Male, Female, Nonbinary, or other).
- **age** The current age of the customer.
- **height_cm** Self-reported height in centimeters
- **waist_size_cm** Self-reported waist size in centimeters
- **inseam_cm** Self-reported inseam (measurement from the crotch of the pants to the floor) in centimeters
- **test_group** Whether or not the customer is in an experimental test group that receives special coupons once a month. (0 for no. 1 for yes)
- **salary_self_report_in_k** Self-reported salary of the customer, in thousands
- **months_active** The number of months the customer has been a part of the clothing store's preferred rewards program
- **num_purchases** The number of purchases the customer has made
  - A purchase is defined as a single transaction that could include multiple items.

Introduction cont.

These predictive models can be incredibly valuable if successful. By accurately forecasting customer annual spending based on consumer profiles, the clothing store can optimize the impacts of its marketing strategies, promotions, and consumer shopping experience. Furthermore, this model enables the store to allocate and offer products that better fit their customers' preferences. This can lead to not only a higher percentage of customer retention and satisfaction, but it can lead to higher revenues and lower costs.

Methods

The data analysis of the clothing dataset was independently completed by two different models: the linear regression model and the polynomial regression (to the degree of two [2]) model.

Data Preparation

Before fitting the data to the models, the data was prepared through multiple steps. First, the "null" or empty values were dropped from the dataset and its indices were reset. Afterward, the data was split into two lists: `x` (independent variables) and `y` (customer's annual spending). Following this, Train-Test-Split Model Validation with an 80/20 split was utilized and outputted the four splits: X-train, X_test, y_train, and y_test. Regarding the split, 80% of the data was split into the training dataset that will be used to build the model and the remaining 20% will be used as the testing dataset to assess the model's performance on unfamiliar data. After splitting the dataset 80/20, another list was created to contain only independent variables defined as continuous or numerical variables. Following this step, the continuous variables' z-score was computed, and for the one categorical variable (gender) was applied with One Hot Encoding.

Linear Regression Model

After preparing the data, an empty linear regression model was created. Subsequently, the linear regression pipeline was constructed using the dataset's z-score and the empty linear regression model. The model was then trained using X_train and y_train. Finally, the linear regression model was utilized to make two predictions of the customer's annual spending: one by inputting X_train and the other by inputting X_test into the model.

Methods cont.

Polynomial Regression Model

The model creation was also repeated for the polynomial regression model, but one change was made to the pipeline. In the polynomial pipeline, an additional step was added: `PolynomialFeatures()`, which creates new features to the regression by raising the independent variables to a certain degree, which in this model's case was to the degree of two (2).

Assessing the Models' Performances

After the completion of the four predictions, the two model's performance is conducted through the calculation of both the training and testing sets' MSE (Mean Squared Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and R^2.

Results

After conducting the training and performance analysis on the two models, the polynomial regression model's success and accuracy surpassed that of the linear regression. Displayed below in Table are the performance metrics of each model's training and testing sets.

	MSE	MAE	MAPE	R^2
Linear: Train	~15,217.678	~97.371	~0.1381	~0.431
Linear: Test	~16,032.59	~100.00	~0.145	~0.441
Polynomial: Train	~5463.356	~59.374	~0.079	~0.800
Polynomial: Test	~5746.88	~61.024	~0.083	~0.800

Model Comparison

This model's performance was tested through four (4) different metrics: MSE, MAE, MAPE, and R^2. The MSE signifies the sensitivity of a model to prediction errors and outliers. The MAE and MAPE are key measures of a model's prediction accuracy. Finally, the R^2 of a model explains the percentage of the dependent variable's (customer annual spending) variance is explained by the independent variables.

Results cont.

Fit of the Model

While there is a difference between the linear and polynomial regression training and testing sets' MSE, MAE, and MAPE, the difference is not significant enough to show signs that either model is overfitting the data. Additionally, the difference between the training and testing sets for the polynomial model is less than the linear model.

Interpretation of MSE - Mean Squared Error

As shown above, the polynomial regression model achieved approximately a 65% lower MSE on the testing dataset compared to the linear regression model. A 65% difference between the two models is significant towards identifying the sensitivity of the model to the prediction errors and outliers.

Interpretation of MAE - Mean Absolute Error

The polynomial regression model achieved a 40% lower MAE on the testing dataset compared to the linear model. The linear regression's testing dataset scored a MAE of 100.000, which means that, on average, the model's predictions are off by about 100 units. In more applicable terms, this error implies that if the model predicts a customer's annual spending of $1,000, the expectation of the actual spending is about $900 or $1,100 ($100 less or more). The polynomial regression's testing set scored a MAE of 61.024, so the model's prediction is, on average, about $61.024 off from the actual value.

Interpretation of MAPE - Mean Absolute Percentage Error

The polynomial regression model achieved a 6.2% lower MAE on the testing dataset compared to the linear model. The linear regression's testing data scored a 14.5% MAPE which means that, on average, the linear predictions are approximately 14% different from the actual value. The polynomial regression's testing data scored a 8.3% MAPE which means that, on average, the linear predictions are approximately 8.3% different from the actual value.

Interpretation of R^2

The polynomial regression model achieved a 35.9% higher R^2 value on the testing dataset compared to the linear model. With that being said, 80% of the customer annual spending's variance can be explained by the independent variables from the polynomial regression. 44.1% of the customer annual spending's variance can be explained by the independent variable from the linear regression.

Results cont.

Model Trust and its Caveats

The statistical comparison between the two models reveals that the polynomial regression model does a better job with predicting the customer annual spendings. While it is far from the perfect model, the `PolynomialFeatures`, which includes both the degrees and interactions, was needed in this prediction process in order to better fit to the trends and fluctuations of the training and testing data. With that being said, I trust the results of the polynomial model more than that of the linear model's, but I would advise the company to take caution with completely being dependent on the data. The reasoning behind this suggestion is that the fashion economy is a sensitive economy that is prone to substantial fluctuations based on different variables such as the rise of "fast fashion", ever changing fashion trends, beauty standards, and cost of the products. To combat this issue, I highly recommend the clothing store to continue conducting surveys and legally collecting data from their customers to constantly update the model with new data and potentially significant data trends. Additionally, I want to emphasize the importance of continuous monitoring and evaluation of the model.

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?

In Figure 1, we observe that there is a difference in the amount customers spend based on their membership in the experimental test group. Overall, being in the experimental test group does increase the amount a customer spends at the store, regardless of the gender. However, the increase is a more distinct increase for women, with a difference of approximately +150,000. Male customers also exhibit a noticeable difference in the spending between the men in the test and non-test groups, though the difference is less distinct.

For nonbinary and other gender categories, the distinction in annual spending between the test and non-test group is less pronounced than the female and male category. It is also important to note that these gender categories represent a much smaller portion of the clothing store dataset compared to the male and female population.

Regardless of the magnitude of the distinction between test-group membership and annual spending, there is still a noticeable difference when examining all of the graphs up close, which indicates that there is a relationship between the experimental test group status and the customer's annual spending, regardless of gender.

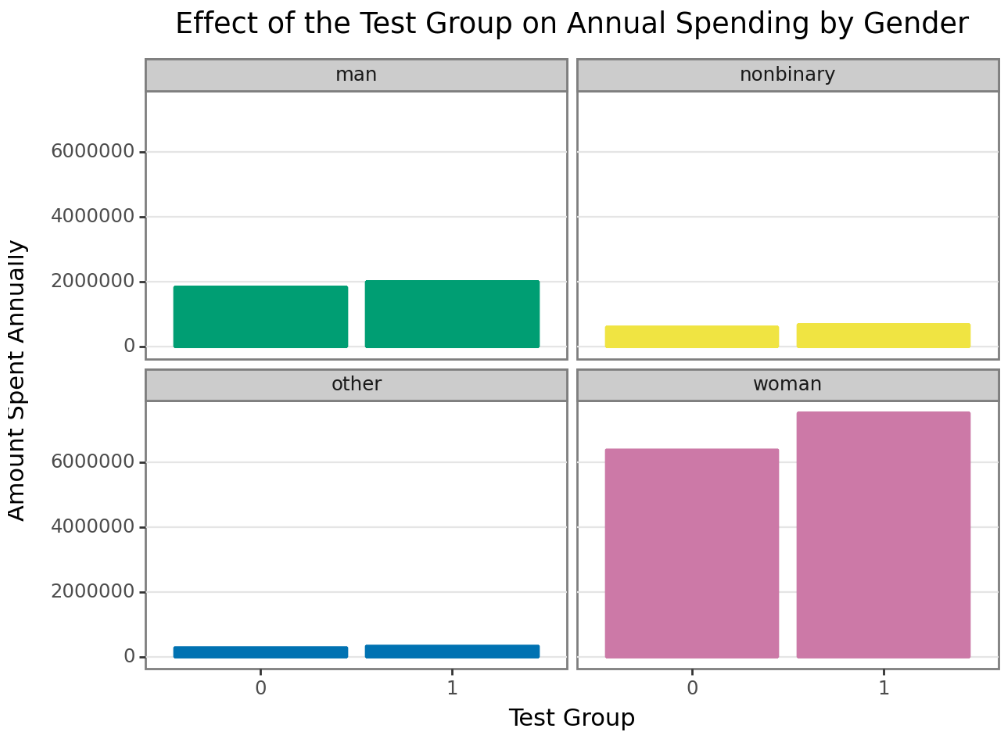

Figure 1: A group of four (4) bar graphs, split by the customers' self-disclosed gender, showing the relationship between customers' test group membership and their annual spending at the clothing store.

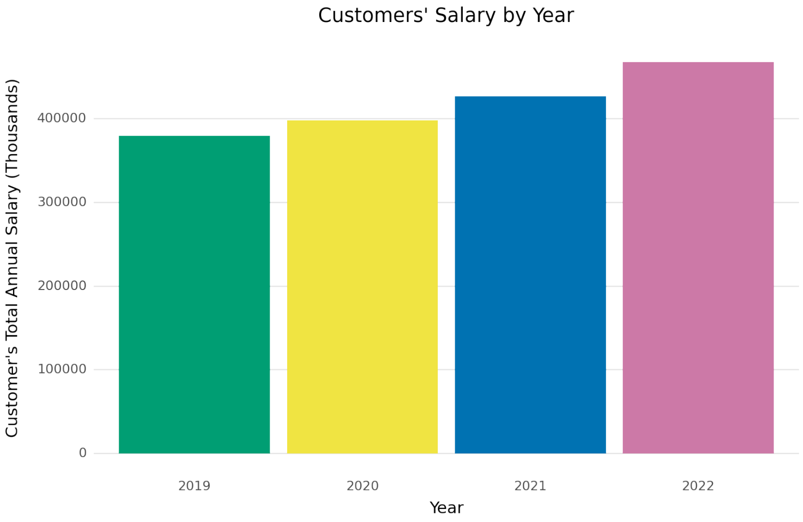

Figure 2: A bar graph showing the customer's total annual salary (y-axis) during the years of 2019 to 2022.

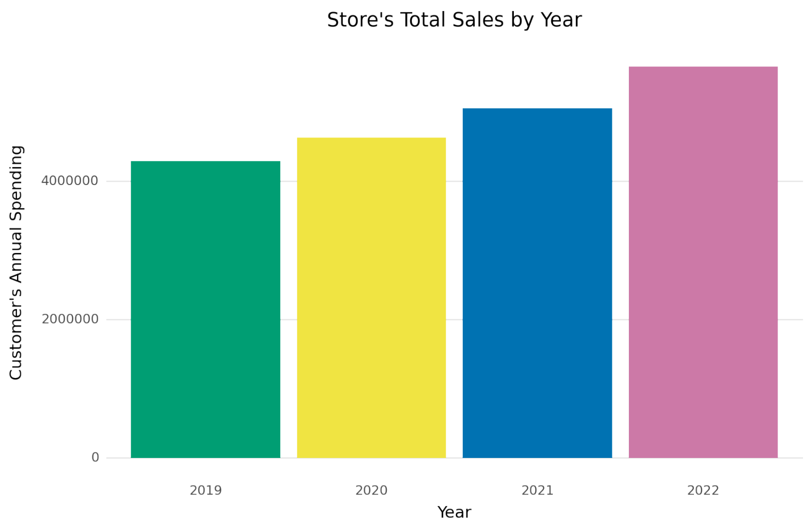

Figure 3: A bar graph showing the store's total sales, represented by the customer's annual spending, during the years of 2019 to 2022.

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

In Figure 2, we see that there is an increase in the clothing store's customer's annual spending each year leading up to 2022. In Figure 3, we see that there is also an increase in the clothing store's customer's annual spending (store's total sales) each year leading up to 2022. Overall, 2022 was the year that customers made the most money and spent the most. With that being said, there is a positive correlation between the customers' annual salary and their annual spending.

Discussion/Reflection

Through performing these analyses, we are able to see the impacts different variables have on the clothing store's customer's annual spending. While some variables such as the experimental test group membership have a less significant positive impact on a customer's spending, variables such as the customer's gender and annual salary have a more significant positive impact on their spendings. Furthermore, running these analyses, visualizes the importance of fitting the data to the correct model in order to reduce the potential errors in forecasting and predicting. We were able to see this through the comparison of the linear regression and polynomial regression's training and testing performance metrics.

If this analysis was to be performed again, I would not use the Train Test Split method in order to validate the models. Rather, I would recommend using K-Fold Cross Validation to validate the models that were run in the analyses. While K-Fold Cross Validation is computationally expensive, it is able to generalize well across the different subsets of the data. In order to mitigate the computational cost, we can subset the data or do parallel processing to handle the separate parts of the validation. Additionally, this validation method is able to capture seasonality and trend-related information, which can be important factors in predicting a customers' annual spending for a store in the fashion industry.