Clothing Store Customer Data

Abhi Gudimella

Introduction

In this project, I attempted to use customer data from a clothing store to analyze the customer base and predict how much they would annually spend. The variables used to determine that are based on the customer's profile and behaviors. They are gender, age, height, waist size, inseam, membership test group status, self-reported salary, months active in the rewards program, number of purchases, and the year of data collection. The models here can help estimate revenue and offer valuable insights for the clothing store, letting them understand their customers better.

Methods

To begin, I dropped missing values in the dataset to be as precise as possible. On one hand, all the continuous variables (age, height, waist size, inseam, self-reported salary, months active in the rewards program, number of purchases, and the year of data collection) were z-scored, which is a way of standardizing all values despite their different ranges. On the other hand, the categorical variables (gender and test group membership status) were one-hot encoded, which is a way of converting non-continuous variables into numerical values. The project observed two different models–linear and logistic regression. The linear model looked at a linear relationship between the inputs and output. The logistic model looks at the polynomial, more complicated, relationship. These models use a 80/20 Test/Train Split. Therefore, 80% of the available data was used to train whereas the remaining 20% was used to test the effectiveness and accuracy of the model. We tested the accuracy of the model by looking at its Mean Squared Error (MSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R-squared values.

Results

First, we looked at a Linear Regression model to represent the data and tried to linearly predict the amount of money a customer annually spends. This model produced the following results.

```
Train MSE :  12943.090588067958
Train MAE :  89.84199252272599
Train MAPE:  0.1275533278620406
Train R2  :  0.5255533158348613

Test MSE  :  13124.379634459647
Test MAE  :  90.2184075540761
Test MAPE :  0.13046607058489756
Test R2  :  0.5070407781149123
```

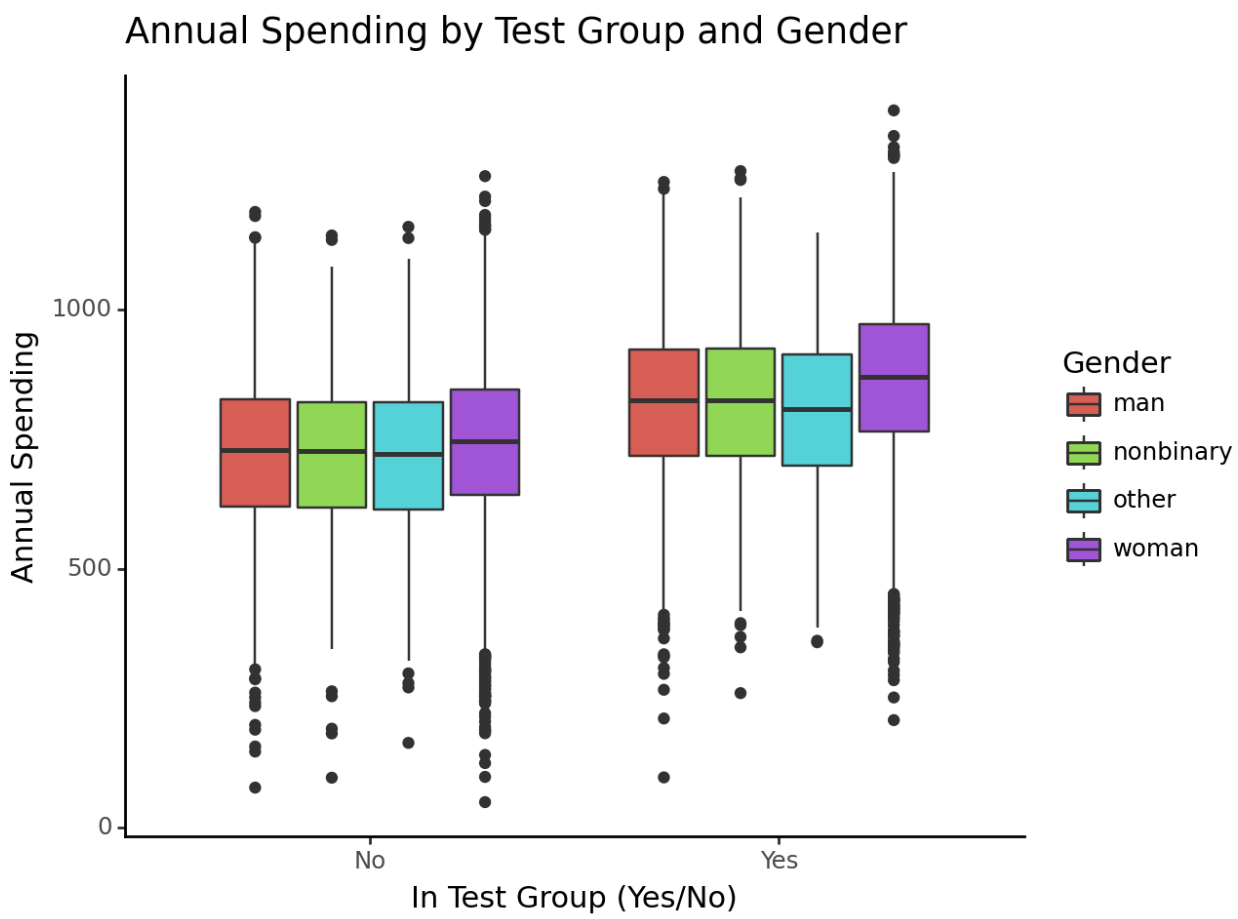
Here, I found that a linear model was clearly not effective. It only presented an R-squared of 0.53 in training and a 0.51 in testing. Moreover, the predictions made by the model were, on average, around 13% off the actual values. Therefore, a polynomial regression model was needed. First the model was trained as a degree 2 polynomial. This produced the following results.

```
Train MSE :  3100.8744139977534
Train MAE :  44.478693156008674
Train MAPE:  0.05940932374047139
Train R2  :  0.8863332081527708

Test MSE  :  3063.7689572930335
Test MAE  :  44.25702033710133
Test MAPE :  0.059961383727182856
Test R2  :  0.8849230818302947
```

Here, the model performed a lot better. Other than the obviously highly improved R-squared values, the model was also able to minimize errors. For example, this time the model predicted values within 6% of the actual values. Even the MSE dropped by a lot, meaning the larger errors were significantly minimized.

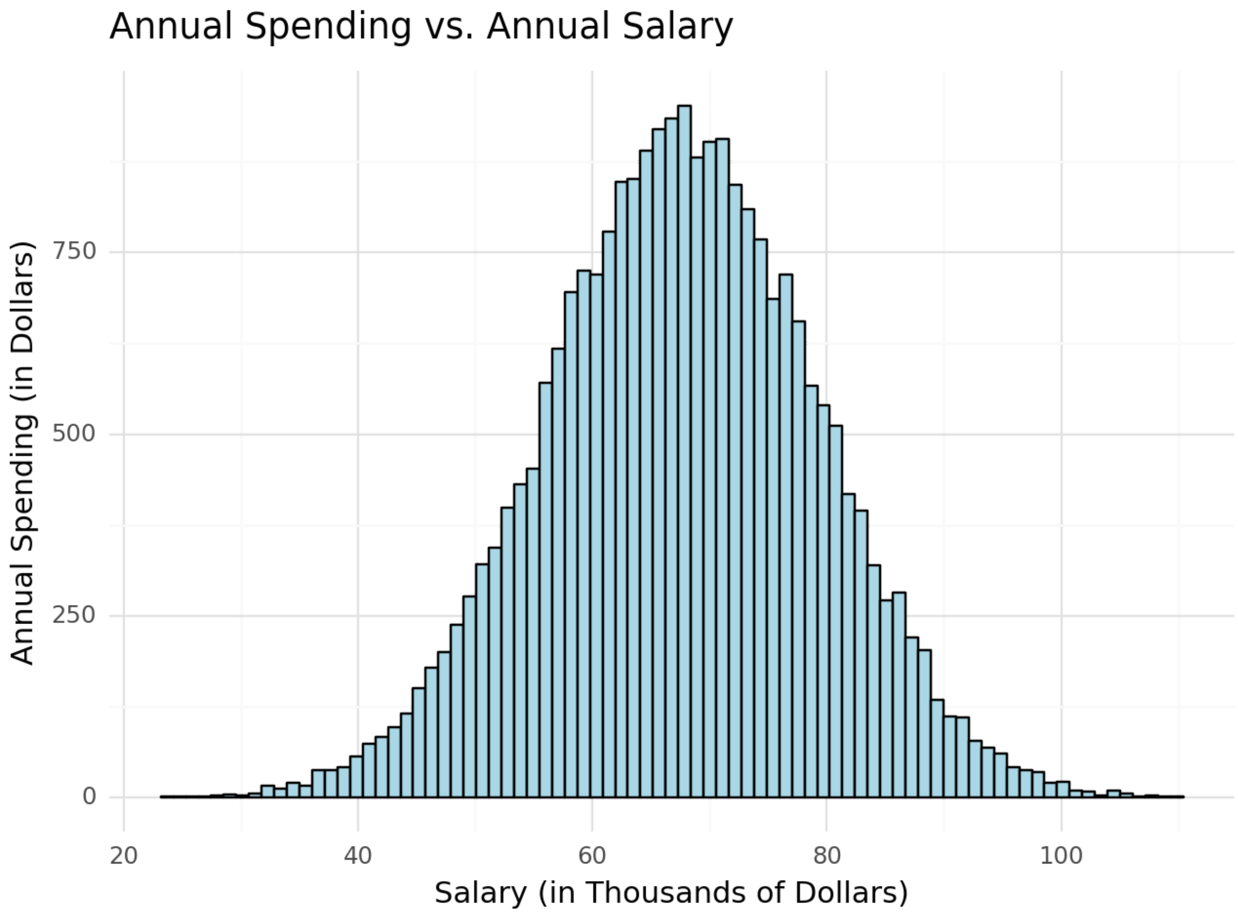
Does being in the experimental test_group actually increase the amount a customer spends at the store? Is this relationship different for the different genders?



Caption 1: This graph shows 4 pairs of graphs. The former elements of the pair make up the people not in the clothing stores' test group while the latter represents the other. Each pair represents a different gender. On the y-axis is the amount annually spent by each customer at the store. Those not in the test group hover around a median of 700 dollars of annual spending whereas those in the group hover around a median of 850.

Clearly, the members in the test group spend more money across all genders. The median values are always higher for those in the test group. It is interesting to see that those who identify their gender as “other” spend lower than those identifying as other genders across those in the test group and those not in it.

Does making more money (salary) tend to increase the number of purchases someone makes? Does it increase the total amount spent?



Caption 2: This graph is a histogram where the x-axis represents the salary earned by each customer at the store in thousands of dollars–ranging from 20 to 100. The y-axis represents each customer's corresponding annual spending in dollars ranging from 0 to 1000. The distribution is approximately a normal distribution showing a bell curve.

The graph very clearly shows that customers making more money do not spend more money at this clothing store. In fact, the graph shows a bell curve relationship between the annual salary and the annual spending. Therefore, those in the middle class of earning are much more likely to spend more money at this store.

Discussion/Reflection

From performing these analyses, I've learned several key insights about customer spending behavior at the clothing store. With this report, the company should be able to identify certain spending habits and change their business model to apply them.

Firstly, there's a clear trend indicating that customers enrolled in the test group tend to spend more annually compared to those not in the test group, across all genders. This suggests that the special rewards offered to the test group may be effective in driving higher spending. Secondly, the histogram reveals an interesting relationship between customer salary and annual spending. Contrary to the expectation that higher earners would spend more, the data shows a bell curve distribution, indicating that customers with moderate salaries are more likely to spend higher amounts annually at the store. This suggests that the store's pricing strategy or product offerings may resonate more strongly with middle-income customers.