COTIVITI DATA SCIENTIST ASSESSMENT TEST
Blake Hamm
2021-08-16

Assessment Write Up

For this assessment, I wanted to explore the data before fitting a model. This helps me understand the distribution of the two variables as well as checking for correlation. With only one independent and dependent variable and only 500 observations, I was limited on how much I could explore.

First, I checked descriptive statistics and visualized the individual fields. This involved verifying the shape (500, 2), describing the data and plotting a boxplot and histogram. I found the independent variable (X) had a very uniform distribution with a subtle skew to the right; I did not find any outliers. The dependent variable (y) had some sort of log normal, geometric or exponential distribution; there were outliers on the higher end (> ~20 units). With this information in mind, I considered removing outliers before fitting a model.

In addition to checking the individual fields, I wanted to check some correlation plots for further exploratory data analysis. I plotted a scatter plot, 2d histogram and built a correlation matrix. I noticed a strong correlation at smaller values of X. As X increased, the covariance increased. I also found what appeared to be a positive trend with a correlation coefficient of 73%. Because of the increase in covariance over X and a correlation of 73%, this was not the strongest dataset, but could give some directional information as there is a clear, positive trend.

Next, I built a function in Python to fit a linear model, describe and visualize the results; this way, I could easily test a model with and without outliers. I found the R-squared did not improve when outliers were removed; I decided to leave the outliers in my final results. Finally, I used sklearn for cross validation to obtain an accuracy score; I chose a test size of 25%. I found the model accuracy was ~53% with the training data and ~49% with the test data. Depending on our KPI and accuracy goals, this may not be a very strong model and I suggest we need more data or a more robust model for a better fit.

My final results indicate a slope of 3.16 and an intercept of -2.33. Both the constant and independent variable were statistically significant with P-value close to zero. Also, the R-squared was .523 indicating that only 52% of the data fit the regression. Further results can be found in my Jupyter Notebook available here: https://github.com/blake-hamm/cotiviti_test/blob/main/Cotiviti_Assesment.ipynb .

I used the following Python packages for this assessment:
numpy, pandas, seaborn, matplotlib, statsmodels, sklearn, scipy