Sean Olson

Data Analytics

Homework 2

Regression Analysis

1. Regression Analysis was performed to examine three different patterns found within the supplied customer transaction data. Given the high variance within the data, the approach I took to perform this regression analysis was using a supervised learning library to train the model before running it on the data. Using this type of machine learning granted me the ability to customize the model on the fly without having to create separate data frames for each run. This also helped to create the best models possible while mitigating as much of overfitting as possible by splitting the data into testing and training sets and correlating that with the Mean Absolute Error (MAE for linear Regression) and the Logistic Score to determine how well the trained models performed on the test data when compared to the training data.
   1. The first regression performed was a logistic regression on determining what contributes to a customer’s decision to make a purchase or not. The chosen model was . These independent variables were chosen because they either ranked highly in a Recursive Feature Elimination (RFE) analysis or I felt they represented an individual’s purchasing habits.

Logistic Regression Results

-----------------------------------------------------------------------

Independent Variables Coefficients Log Odds Standard Errors \

0 Intercept -3.103669 0.044884 0.062561

1 month -0.110616 0.895282 0.005518

2 items 5.276033 195.592346 0.017942

3 catalog\_sent -0.303133 0.738501 0.055305

4 catalog 0.624634 1.867563 0.188493

5 store 2.679741 14.581321 0.068094

6 web 1.853758 6.383764 0.111337

z P > |z|

0 -49.609907 0.000000e+00

1 -20.047148 2.137522e-89

2 294.057229 0.000000e+00

3 -5.481071 4.227583e-08

4 3.313837 9.202511e-04

5 39.353576 0.000000e+00

6 16.649981 3.027005e-62

-----------------------------------------------------------------------

Accuracy of logistic regression on test set: 1.0

It’s no surprise that all the independent variables were significant, given the analysis