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.

A screenshot of a cell phone

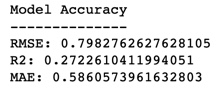
Description automatically generated

These 4 independent variables were chosen because month makes sense, given that a person needs to replace certain items throughout the year. Also, the decision to purchase will rely on how much the person has spent to date on items since the data has been collected. A person is also more likely purchase something if either a catalog or coupon were sent to the customer’s home. Given the variables that were chosen for the model, the probability of accuracy of the model is 0.697, this is a better indicator of model performance than a pseudo-R2 value because this lets us see the probability of the predicted model actually producing a value that isn’t just random. To determine if the model was overfit, we look at the confusion matrix and see that the model had 630 correct predictions and 273 incorrect predictions.

* 1. The linear regression analysis performed in the second part returned the following:

A close up of text on a white background

Description automatically generated



Each variable was chosen based off a developed correlation matrix to ensure no multi-collinearity. Using supervised learning techniques with a training set the model was able to do incrementally better when it came to predictions on the testing set of data, reaching an R2 value of 0.272 with a Mean Absolute Error (MAE) of 0.586. The MAE indicates that the values from the training set and the predicted values from the testing set were on average 0.586 units separate from each other either along the X or Y axis. This indicates strong model performance and given the low R2 value, overfitting was not observed. All but one of the variables chosen had significant effects on the dependent variable (spend), but even the non-significant variable (store) was still significant with P < 0.10. It is also interesting to note that all but two independent variables had a positive effect on spend except for catalog\_sent and web. This is most likely due to the idea that if a person sees it in a catalog before they see it in store then they can take the time to make a more thoughtful decision on if they need it or not and the same is true for seeing an item on the web. Giving the customer more time to make a decision leads to a negative effect on spending.

* 1. The final regression analysis performed was another logistic regression on what contributes to a person returning an item. The results below indicate that a person is most likely to return an item when they bought it through the catalog. It’s also interesting to note that customers were less likely to return an item if a discount was given to them. Indicating that it wouldn’t be worth their time to return an item if they didn’t pay full price. This model’s accuracy was 0.945, indicating that this model was a little easier to predict values as there were more variables that could explain the outcome of the predicted dependent variable. To ensure overfitting was not a concern the confusion matrix showed 609 correct predictions and 35 incorrect predictions. This is close to an overfit model, but it is expected given the increase in independent variables.

A close up of text on a white background

Description automatically generated

1. Mr. Shoemaker should target the customers who purchase goods in store over other channels and usually purchase a lot of goods at once. These two variables would generate about $0.26 per customer, so over the 100 customers he would expect to receive $26 in profit per trip to the store. Combining this with other factors, he would be expecting almost $50 per hundred customers each shopping trip. He should avoid sending customers catalogs and sending deals to web shoppers because he would be losing a combined $0.70 per customer who shops in these ways.
2. In order to improve the analytical approach Mr. Shoemaker should actually note down the months instead indexing them. Then a seasonality analysis can be performed, and forecasting models can be applied to determine the best times to send out catalogs and coupons to entice customers to purchase. Also Mr. Shoemaker should note down what individuals click on the website and track what people might be thinking about buying. Doing so could shed some light on the percentage of goods bought vs browsed to determine growth potential for his market.