Logistic Regression Assignment

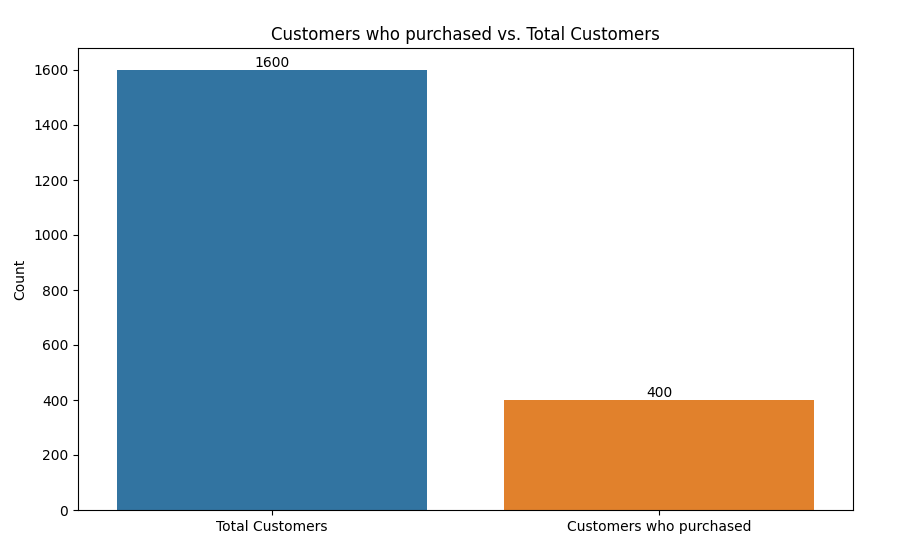
Cade Muntz

**Overview**

For this project, we have been asked to create a model which will allow a local book company called BookNerds to predict which customers will benefit more from having mailers sent to them. The benefit will come due to the customers responding to said mailers by making a purchase. To make this model, we used data given by BookNerds and used different predictor variables to perform a logistic regression. Logistic regression is by far the most practical model to use due to the outcome variable of the customers choice being binary in nature, as they either make a choice to purchase or do not.

**Initial Insights:**

To begin the process, we first started by looking at the data given to us by BookNerds. We first wanted to see how successful they were without making a model. With their current data, BookNerds has a 25% success rate (400 Customers, out of 1600 total) in terms of customers responding to a mailer with a purchase. You can see this visualized in the graph below.



Based on this graph we are able to conclude that BookNerds is only producing a sale with 25% of the mailers they send out.

**Predictive Insights:**

When making a model for BookNerds we started by looking at how different predictor variables might be able to influence / predict a customer's choice. Using these predictor variables we will be able to tell BookNerds what customer data they should keep an eye on when it comes to a customer purchasing from a mailer, so they know who the mailers should be sent out to..

**Initial Model:**

We first considered that a customers choice to purchase a book upon receiving a mailer (Choice 0/1) could be seen as a function of a variety of predictor variables. These variables include : Gender, Frequency, First purchase, last purchase, and Amount purchased which should all be self explanatory variables. While the variables P\_History, P\_Science, P\_Educ, and P\_Health are books under specific categories with Educ standing for education.

Upon running this model we were immediately able to conclude that all of the variables were fantastic predictors aside from First Purchase which had an insignificant predicting value (p = 0.197.) Therefore, First Purchase was taken out of any future models as is was insignificant.

**Model Fit:**

Based on the set of predictors used for the model upon taking out First Purchase our model was able to whether a customer would purchase or not purchase with slightly higher than 80% accuracy (Accuracy = 80.25%) However, this model seems to be much better at predicting whether someone will not purchase rather than whether they will purchase. This model was able to predict that a customer would not purchase with 93.50% accuracy while it was only able to predict if a customer would purchase with 40.50% accuracy. There is clearly still some work to be done with this specific model, however 40.50% is still much better than BookNerds original 25%.

**Significant predictors:**

In this model, every variable was extremely good at predicting with every variable other than Amount Purchased having a p value of less than .001. Amount Purchased was still a significant predictor having a p value of 0.018. This is good news for BookNerds as it means the majority of the data they have been collecting on their customers is valuable for drawing solid conclusions through predictive modeling. \

**Insignificant Predictors:**

The only insignificant predictor in the initial model was First Purchase. As mentioned earlier First Purchase had a p-value of 0.197 meaning that it did not meet our significance threshold of .05. In terms of what this means, it likely was insignificant due to other variables that were similar in nature taking up most of the significance. These variables would likely be Frequency, and Last Purchase. This is because, intuitively speaking, a customer's first purchase is much less valuable to BookNerds than when they purchased last and how frequently they purchased when it comes to predicting whether or not they will purchase again.

**Interpretations of Predictors:**

Knowing that all of these variables are predictive is good, but understanding what each variable actually means in terms of predictive power is much more important. In this section, we will go over what exactly each predictor variable means, and what it means in terms of influencing a customer's purchasing choice.

* Gender (0 = Male, 1= Female) - This variable has a coefficient of 0.867. Since this coefficient is positive, it means that Females are more likely to make a purchase when receiving a mailer, while Males are less likely to do the same.
* Amount Purchased - This variable has a coefficient of -0.002. Because this variable has a negative coefficient, it means that for each additional purchase the customer makes, the least likely they will be to purchase after receiving a mailer. However, since this coefficient is so small and almost positive, this variable is much harder to interpret, and it could likely swing both ways.
* Frequency - This variable has a coefficient of 0.090. Since this coefficient is positive, it means that the more frequently a customer purchases a book, the more likely they will be to purchase a book upon receiving a mailer.
* Last Purchase - This variable has a coefficient of -0.563. Because this variable is negative, that means that the more time that has passed since the last purchase of a customer, the less likely they are to purchase a book upon receiving a mailer.
* P\_History - This variable has a coefficient of 0.832. Because this variable is positive, this means that customer who typically purchase History books are more likely to purchase books upon receiving a mailer. This could be because customers interested in history books are more interested in getting their hands on even more books, leading them to purchase more books to gain more knowledge.
* P\_Science - This variable has a coefficient of 0.678. Because this variable is positive, this means that customers who typically purchase science books are more likely to purchase a book upon receiving a mailer. Just like History books, this could be because it is a studious customer who enjoys learning about science and would like to learn more by reading upon receiving a mailer.
* P\_Business - This variable has a coefficient of 0.939. Because this variable is positive, this means that customers who typically buy business books will likely purchase a book upon receiving a mailer. Just like the previous 2 variables, they likely want to learn more and are more apt to purchase a book upon receiving a mailer.
* P\_Educ - This variable has a coefficient of 0.915. Because this variable is positive, this means that customers who typically buy Educational books will likely purchase a book upon receiving a mailer. And again, like the previous book variables, these customers are likely interested in more so are more likely to continue purchasing books upon receiving a mailer.
* P\_Health - This variable has a coefficient of -0.652. Because this variable is negative, this means that customers who typically buy Health books will likely not purchase a book upon receiving a mailer. This could be because these customers purchase health books for specific health related problems, and do not need more health books after the first.

**Testing for Model Accuracy Using Split Data:**

Splitting the data is a useful tool when modeling predictive data, as you will be able to see the model predict data versus the actual data to get a better sense of the accuracy of the model. In this case, we split the model into 70% of our data into traIning data, which is the actual data, and the other 30% will be split into testing data, which is data we are testing to see how accurate our predictive model is.

Upon running this model, the results are shown in the confusion matrix below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | True 1 | True 0 | Class Precision |
| Predicted 1 | 338 | 76 | 0.39 |
| Predicted 0 | 17 | 49 | 0.95 |
| Class Recall | 0.74 | 0.82 |  |

The results show that this model is able to predict a customer's choice with an accuracy of 81%. It is also noted that it was able to predict a customer which won't purchase with an accuracy of 88% and was able to predict a customer who will purchase with an accuracy of 51%. Essentially, this means that this model is much better at predicting which customers will likely not respond to a mailer, which is just as valuable as we better understand what customers to not send mailers too. While this is still better than the 25% that BookNerds originally had, it is still nowhere near a perfect model.

Gathering more data and potentially getting more predictor variables could greatly impact the accuracy of predicting customers who will respond to the mailer.

**Testing Model Accuracy Using Untested Data:**

Instead of splitting the data, we then used untested data to apply to the model. This untested data is a predictive set of data with new customers from a different geographical region to the original data. This will give us the opportunity to use our model for practical purposes given BookNerds a chance to test the model on this new data. We will go over what this new model has in store for BookNerds in the upcoming recommendation section

|  |  |  |  |
| --- | --- | --- | --- |
|  | True 1 | True 0 | Class Precision |
| Predict 1 | 221 | 10 | 0.79 |
| Predict 0 | 57 | 32 | 0.76 |
| Class Recall | 0.36 | 0.96 |  |

**Recommendations:**

When applying the untested data to the predictive model, the results tell us that the recommended number mailers to be sent out is approximately 221. Considering our model does have errors when predicting customers who purchase more than customers who don't purchase this should be adjusted to a higher value. Since the accuracy of this model specifically on people who will receive mailers is only 36%, by taking 36% of 221 we get 79, adding that we can recommend that at least 300 mailers should be sent out for best results. Therefore, this number should give BookNerds the highest proportion of mailers sent out that turn into purchases. This should save BookNerds money in terms of costs for the mailers, and returns on the purchases

When using the model, BookNerds should focus on the predictive variables mentioned above, while not worrying about the First Purchase variable. I believe that this model still has work to be done to it, and playing around with the model further would likely result in an even more accurate model.