Homework 7

GSBA 524

Fall 2022, Term 3

Please submit *two* files for this homework:

1. A report document (with extension `.nb.html`) that is created in Radiant (as we saw in the MLR lab, this is available from Report > Rmd.
2. The state file (with extension `.state.rda`) that is created when you click "Save Radiant state file."

Background

In this problem we will be working with `oj.csv`, which contains 1070 purchases where a customer either purchased Citrus Hill or Minute Maid orange juice. We will use various pieces of information of the customer and the product to predict which juice is purchased. Here is information about each variable in the data set:

```
Purchase
A factor with levels CH and MM indicating whether the customer purchased Citrus
Hill or Minute Maid Orange Juice

WeekofPurchase
Week of purchase

StoreID
Store ID

PriceCH
Price charged for CH

PriceMM
Price charged for MM

DiscCH
Discount offered for CH

DiscMM
Discount offered for MM

SpecialCH
Indicator of special on CH

SpecialMM
Indicator of special on MM

LoyalCH
Customer brand loyalty for CH
```

```
SalePriceMM
Sale price for MM

SalePriceCH
Sale price for CH

PriceDiff
Sale price of MM less sale price of CH

PctDiscMM
Percentage discount for MM

PctDiscCH
Percentage discount for CH

ListPriceDiff
List price of MM less list price of CH
```

Problems

1. Make a scatter plot with the difference in sales price on the y-axis, customer loyalty to CH on the x-axis, and with points colored by the brand of orange juice they purchased.
2. What are some observations from the plot about consumer purchase behavior? Include in your description comparisons of the bottom right to the bottom left and of the bottom right to the top right. Are your observations in line with what you would expect?
3. Create a new column in the dataset called `training` that randomly assigns each purchase to either a training or test set[1]. After creating this column, press the autopaste button so that it will include the code for inserting this column. (Note: You will not see any output show up in the notebook, but you'll see a code chunk on the left side. Having this code there ensures that the notebook records this step.) Set the filter to be `training==1` so that all models fit below will be trained specifically on the observations in which `training==1`. Now fit a logistic regression model where you predict whether a customer purchases CH based on loyalty to the brand. Provide the output (by autopasting).
4. Is there evidence at the 0.05 level that consumer loyalty to CH is a statistically significant feature in this model? Justify your answer.
5. According to this model, how many times larger is the odds of buying CH for a customer with loyalty 1 for CH compared to a customer with loyalty 0 for CH? Justify your answer.
6. According to this model, what is the probability that a customer with loyalty 0 for CH purchases CH rather than MM? Hint: Go to "Predict" tab and then select "Command" as the input type and enter `LoyalCH=0`. Then autopaste the result in support of your answer to the question. Hint: To avoid the repetitive output, you can delete the `summary(result)` line in the code chunk.
7. Using the output from the "Summary" tab, write out an expression for the probability in the previous part, and verify (using a calculator) that the expression gives the same value as in the previous part. Hint: To write out an expression in nice looking math, you can use some special syntax. For example, if you write `$\frac{e^{0}}{1+e^{0}}=0.5$` it will render as $\frac{e^0}{1+e^0} = 0.5$.
8. What is the odds that a customer with loyalty 0 for CH purchases CH rather than MM? Show the expression for how you calculate this and then use a calculator to get the value.
9. Repeat Problems 6 and 8 for a customer of loyalty 1 for CH.

[1] Please keep the defaults how they are, with `Seed` set to 1234 and `Size` set to 0.7.

10. Using your answers to Problems 8 and 9, calculate how many times larger the odds is of purchasing CH when you compare the most CH-loyal customers to the least CH-loyal customers. Does this (approximately) agree with your answer to 5?

11. Add a column called `pred_loyal` to the data set that has the predicted probabilities using the model in Problem 3. Include in this answer the Radiant output you get after clicking "Store predictions." For the first purchase in the dataset **that was not used for training**,[2] give the CH-loyalty level of the customer, the predicted probability of purchasing CH, and the brand of orange juice that was actually purchased by this customer.

12. Consider the classifier that predicts that a customer will purchase CH whenever our estimated probability is at least 0.5. What is the accuracy of this classifier on the test set? Hint: Go to Model > Evaluate classification, select `pred_loyal` as your stored predictions and leave `Cost` at 1 and `Margin` at 2 (which will make it so that the probability cutoff used is 0.5). Choose "Both" for the "Show results for:" option, so that we can see both training and test set evaluations. Go to the "Confusion" tab to find the accuracy. (Autopaste to get the full table from the "Confusion" tab included in your report.)

13. Now fit a logistic regression model with both CH-loyalty and the difference in sales price. Calculate the predictions on the dataset and save this to a column called `pred_loyal_pricediff`. Provide the output.

14. Suppose the price of MM were increased by 10 cents while the price of CH is held fixed. How many times larger would the odds of a customer buying CH be?

15. Suppose Customer A's CH loyalty is higher than Customer B's, with a difference in loyalty scores of 0.1. According to the model, how much of a discount would you have to give Customer B so that the two customers would have the same odds of purchasing CH? Justify your answer and show your work.

16. What is the test-set accuracy of this model? Does including the price difference appear to help in terms of test-set accuracy? Justify your answer including any Radiant output you generated to answer this question.

[2] To answer this, you'll need to briefly uncheck the `training==1` filter. But be sure to put a check in that box again before proceeding with later parts.