Connor Adams

12/7/20

Charles Book Club Data Report

Part 1

Overview:

Our report will answer our assumption of the more recent the last purchase, the more products bought from the company in the past, and the more money spent in the past buying the company's products, the more likely the customer is to purchase the product offered. The way our report will answer this question is through the variables: Total money spent on books over a period of time (M), the total number of previous purchases (F), and the months since last purchase (R). We will also be going over how the Charles Book Club data set helps us make decisions on which books to put on the shelves in the bookstore and how well the new book, "*The Art History of Florence"* is performing in the bookstore.

Background:

We are using the data set named Charles Book Club. To give a brief overview on the data set, so when reading the report, the reader will be able to understand what each data field means, we have attached all variable descriptions below.

Variable name	Description
Seq#	Sequence number in the partition
ID#	Identification number in the full (unpartitioned) market test dataset
Gender	0 = Male, 1 = Female
M	Monetary—Total money spent on books
R	Recency—Months since last purchase
F	Frequency—Total number of purchases
FirstPurch	Months since first purchase
ChildBks	Number of purchases from the category child books
YouthBks	Number of purchases from the category youth books
CookBks	Number of purchases from the category cookbooks
DoItYBks	Number of purchases from the category do-it-yourself books
RefBks	Number of purchases from the category reference books (atlases, encyclopedias, dictionaries)
ArtBks	Number of purchases from the category art books
GeoBks	Number of purchases from the category geography books
ItalCook	Number of purchases of book title *Secrets of Italian Cooking*
ItalAtlas	Number of purchases of book title *Historical Atlas of Italy*
ItalArt	Number of purchases of book title *Italian Art*
Florence	= 1 if *The Art History of Florence* was bought; = 0 if not
Related Purchase	Number of related books purchased

Data Preprocessing

When we brought in the data, we were able to see that there were missing variables for some of the variables (First Purchase = 1.98%, Do It Yourself Books = 1.43%, etc.), which could make our report and calculations inaccurate. In order to stop this from happening we used an impute function for all missing variables to have a value that was the median (rounded the median to a whole number) of that variable. Once we used the impute function, we had an output that, showed a 0.0 (0%) for all variables which means there are no missing variables, so every observation had an entry for all variables. The next thing that we wanted to accomplish was finding the summary statistics for the M (Total money spent on books over a period of time), R (Months since last purchase), and F (Total number of previous purchases) variables. Once we calculated this, we got the Table shown in Figure 1 as a result. Based off of the Figure shown, if we look at variable M, we can see that the maximum money spent on books over a period of time was $479 with a minimum of $15, a median of $208, a standard deviation of $100.52, and the average amount a customer spent over a period of time on books was $208.13. Based off of the high standard deviation we can see that most people are spending around $108-308 on books over a period of time. From looking at the M variable we can see that people are spending a lot of money in the bookstore on average which means that customers tend to spend a lot once we get them into the store. Also, from Figure 1, if we look at variable F, we can see that the maximum number of purchases from the company over a period was 12 books, the minimum was 1 book, the median was 2 books, the standard deviation was 3.44 books, and the average number of books purchased over a period from the company was 3.81 books. Based off the standard deviation most people are purchasing 1-7 books over a period of time. This is a good sign for the bookstore because it is showing that customers have bought on average multiple books from the bookstore in the past. Lastly, if we look at variable R, we can see that the maximum time since a customer's last purchase was 36 months (3 years), the minimum was 2 months, the median was 12 months (1 year), the standard deviation was 8.07 months, and the average months since a customer's last purchase was 13.37 months. Based off of the standard deviation we can conclude that most of the customers have spent 5-21 months since their last purchase at the bookstore, which is a bad sign because we would like to see the customer back in the store earlier than what the data is showing. After finding the summary statistics for these variables we want to group the data for each variable, so the data is not so spread out for each variable. For the R variable we created dummy variables so that a 1 means that the customers time since the last purchase was between 0-2 months, a 2 means that the customers last purchase was between 3-6 months, a 3 means that the customers last purchase was between 7-12 months, and a 4 means that the customers last purchase was 13 months and above. For the F variable we created dummy variables so that a 1 means that the customer purchased 1 book from the company over a period, a 2 means that the customer purchased 2 books from

	Minimum	Mean	Median	Standard Deviation	Maximum
M	15.0	208.12825	208.0	100.519664	479.0
F	1.0	3.80550	2.0	3.438841	12.0
R	2.0	13.37450	12.0	8.067963	36.0

Figure 1

the company over a period, and a 3 means that the customer purchased 3 or more books from the company over a period. For the M variable we created dummy variables so that a 1 means that the customer has spent between 0-25 dollars on the company's products over a period, a 2 means that the customer has spent between 26-50 dollars on the company's products over a period, a 3 means that the customer has spent between 51-100 dollars on the company's products over a period, a 4 means that the customer has spent between 101-200 dollars on the company's products over a period, and a 5 means that the customer has spent above 201 dollars on the company's products over a period. Once we created these dummy variables, we found that we added 12 new columns to our data set (Figure 2).

```
Number of rows BEFORE merging: 4000
Number of columns BEFORE merging: 24

Number of rows AFTER merging: 4000
Number of columns AFTER merging: 36
```

Figure 2

Data Visualization

One of the main variables that we want to look at is gender. We want to know what gender is coming into the bookstore and making more purchases. This will help us pick how we decorate our store and which type of advertisements to use if we have more of one gender than the other. One way that we can find this is through creating a Bar graph on the gender variable in the data set. We don't want to look at the count necessarily but we want to look at the percentage of males and females in the data set. We can see this by figure 3. If you look at the two bars in the graph, the blue bar which is named 1.0, is equal to females, and the red bar which is names 0.0, is equal to males. Based off of the graph we can see that there is over double the number of women that responded to the test mailing than men. Women at 70.45% responded more than men at 29.55% to the test mailing. This can be concluded that more women are coming into the bookstore and buying books than men are. We would want to set up the store in a way that women are more likely to come into the store and feel comfortable to buy a book, and we would want to set up advertising in the store that speaks to women more than men.

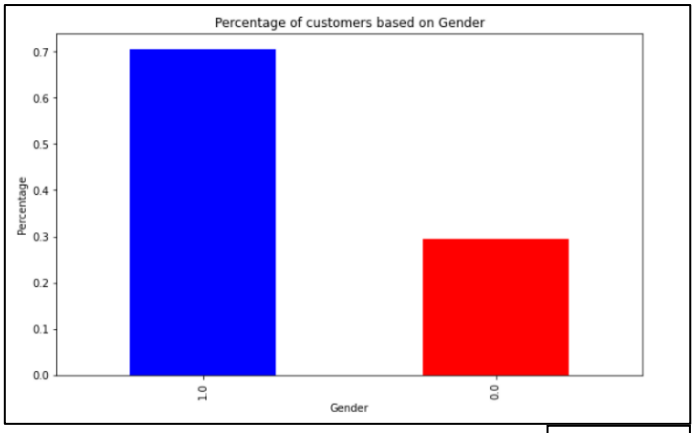

Figure 3

The next variable that is important to us, is the Florence variable. This variable lets us know if customers in the random sample are buying the new book "*The Art History of Florence*". Based off of this information, we want to create a bar chart to see the percentage of people in the data set that have bought the new book, compared to a percentage of people who have not

bought the new book. We can see our results in figure 4. The blue bar that is named 0.0 is the percentage of people who did not buy the new book, and the red bar that is named 1.0 is the percentage of people who did buy "*The Art History of Florence".* Based off of this graph we can see that only 8.45% of people in the random sample bought the "*The Art History of Florence"* and 91.55% of people in the sample did not buy the book. This is telling us that many people in the data set either did not know about the book, or are not interested in reading that book. The bookstore can either stop selling the book or can spend more on advertising the book in the store.

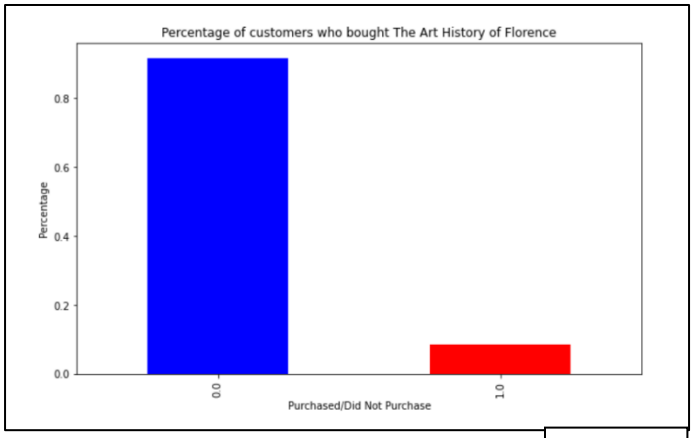

Figure 4

We want to be able to see if there is a difference in the three most used variables (M, R, and F) when grouped by if someone bought "*The Art History of Florence"* or not. This will be able to tell us what audience to target when selling the book and what is the most likely buyer of the book or the most likely not buyer. We will be able to use this information to target people that may be interested in buying the book in the future. The way we can do this is through a Bar Graph that displays two bars with one showing if someone bought the book and the other showing if someone didn't buy the book. The y axis will show the average variable (M, R, or F) that we are looking at. The average will be the average for each group. The first graph that we are looking at is the average time since a customer's last purchase (R variable) grouped by if they bought *"The Art History of Florence"* or not. Our result can be seen in figure 5. From the graph we see that when someone buys the book (1.0, red bar), they on average have bought a book more recent (12 months), then someone who hasn't bought the book (0.0, blue bar) (13.5 months). Although both of these numbers are still high, we can see that people that have bought a book recently are the ones on average that are buying the *"The Art history of Florence"*. The next graph that we took a look at was the average

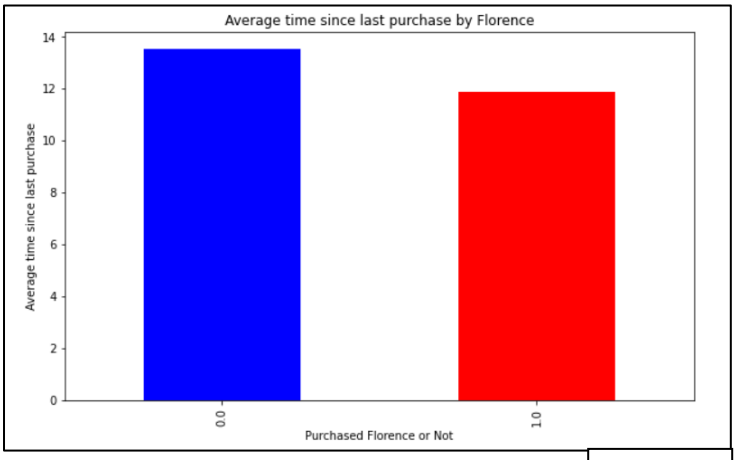

Figure 5

money spent on a company's product over a period (M variable) based on if they bought the "*The Art History of Florence"* or not. We can see our results in figure 6. Based on these results we can see that people who bought the book (1.0, red bar) on average spend more money on the company's product over a period than people who didn't buy the book (0.0, blue bar). Although both groups spent over $200 on the company's product the people who bought the book tended to spend around on average $50 more on the company's product. This tells us that people who bought "*The Art History of Florence"* tend to spend more money on the company's product. We want to target customers that have spent more money on the company's product over a period of time to sell the new book to. The last variable that we want to look at is the average number of previous purchases from the company over a period of time (F variable) grouped by if the customer bought "*The Art History of Florence"* or not. We can see our results in figure 7. Based on the results we can see that the red graph (people who bought the book, 1.0) has a greater average number of previous purchases than the blue graph (people who did not buy the book, 0.0). This means that if someone bought "*The Art History of Florence"* they on average had more previous purchases (5) than people who didn't buy the book (3.5). This shows us that we want to target people who have more previous purchases when selling the book. After looking at all three of these scenarios we can tell that people that bought "*The Art History of Florence"* have had on average more previous purchases, a purchase more recently, and more money spent on the company's product. These are the customers that seem to be the most devoted to the bookstore which is what we could expect and the one's we want to target when selling a new book.

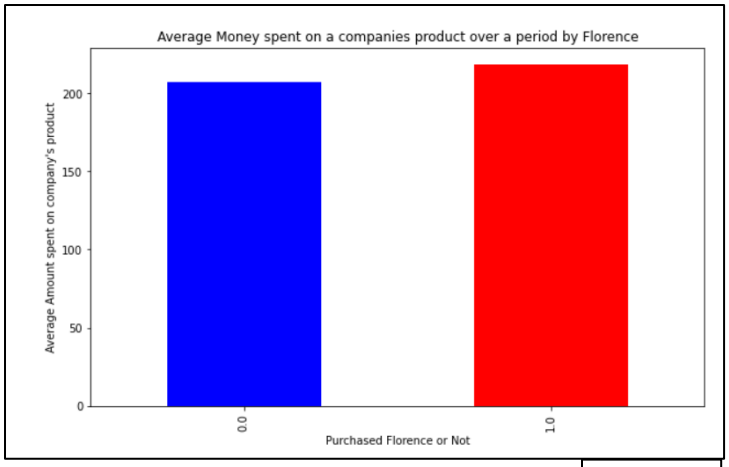

Figure 6

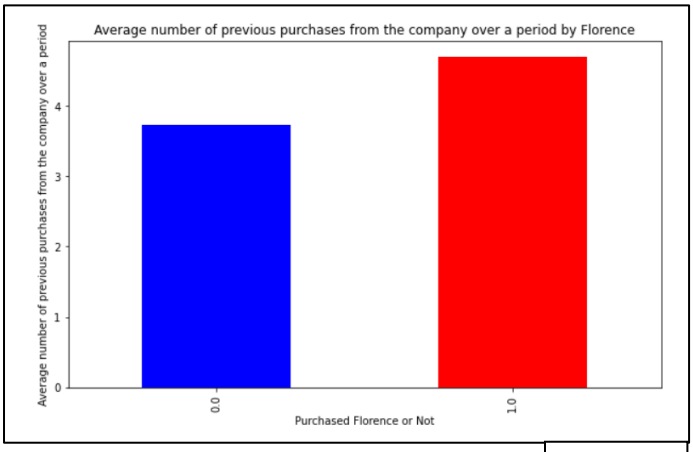

Figure 7

The next thing we want to look at is the relationship between certain variables. This can help us see if there are variables that are related/correlated to each other that will help us predict the other variable. The variables that we want to look at are First Purchase (FirstPurch), Child Books (ChildBks), Youth Books (YouthBks), Cook Books (CookBks), Do It Yourself Books (DoItYBks), Reference Books (RefBks), Art Books (ArtBks), Geography Books (GeogBks), Secrets

of Italian Cooking (ItalCook), History Atlas of Italy (ItalAtlas), Italian Art (ItalArt), and Related Purchases (Related Purchase). The way we can see how all of these are related is through a scatter plot matrix. The scatter plot matrix can be seen in Figure 8. When looking at the scatter plot matrix we want to single out three variables and comment on their scatter plot. These three variables are ones that we think will have importance in determining if there is correlation between any of the variables. The first is Months since first purchase (FirstPurch) and Number of Cook Books purchased (CookBks), the second is Months since first purchase (FirstPurch) and Number of Art Books purchased (ArtBks), and the third is Number of Geography Book purchased (GeogBks) and Number of related books purchased (Related Purchase). We can see the months since first purchase and number of cook books purchased scatter plot in figure 9. Based off of the scatter plot we can see that although the variables seem to not be directly correlated, we can see there seems to be a trend that when the months since the first purchase gets larger the number of cook books purchased also tends to go up. They seem to have a positive correlation between the number of months since the first purchase and the number of cook books purchased. We can expect this however because as more time goes on that a customer has purchased something at the store, they are more likely to buy more books, in this case more cook books. The next scatter plot we are going to look at is between months since the first purchase and the number of art books purchased. We can see the results of this scatter plot in figure 10. Based on the results of the scatter plot we can see that there seems to be a negative correlation between months since the first purchase and the

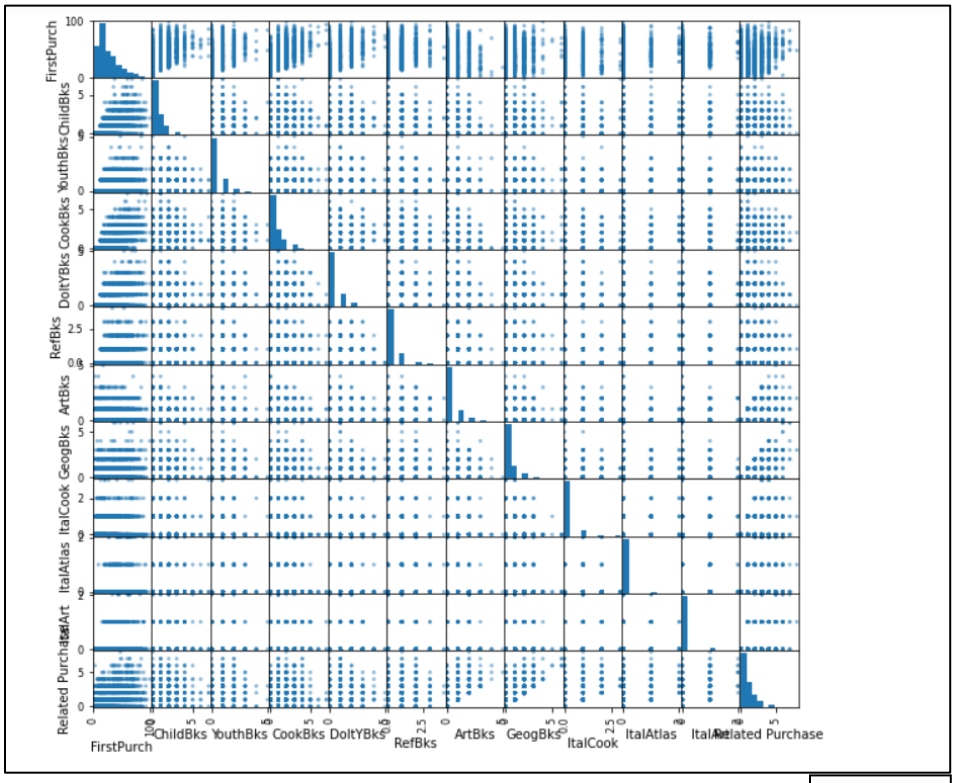

Figure 8

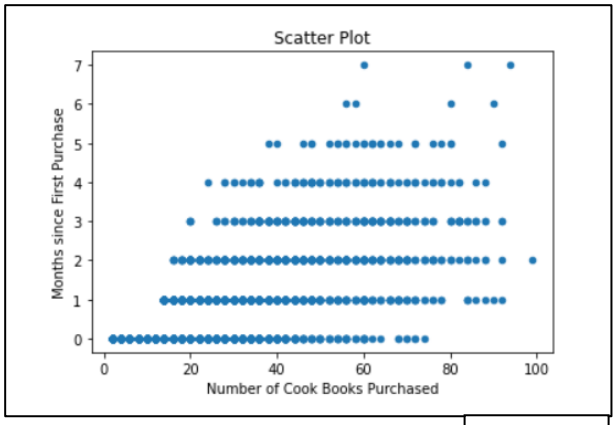

Figure 9

number of art books purchased. This is very shocking because when more time passes since the first purchase, one would think that the more art books would be sold. This does not seem to be the case. It seems as most people buy art books around the time of their first purchase and they don't buy many more after that. The last scatter plot that we want to take a look at is between the number of geography books purchased and the number of related books purchased. This scatter plot can be seen in figure 11. Based on the results of the scatter plot we can predict that there is going to be a positive correlation between the number of related books purchased and the number of geography books purchased. This means that when a customer is making a related purchase it seems that the purchase that is related will be a geography book. This may show that customers that have a related purchase may be buying more geography books.

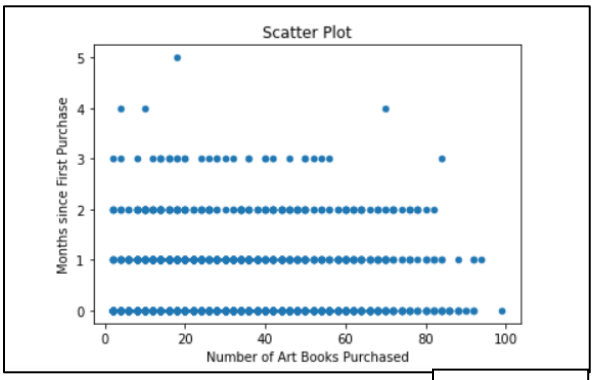

Figure 10

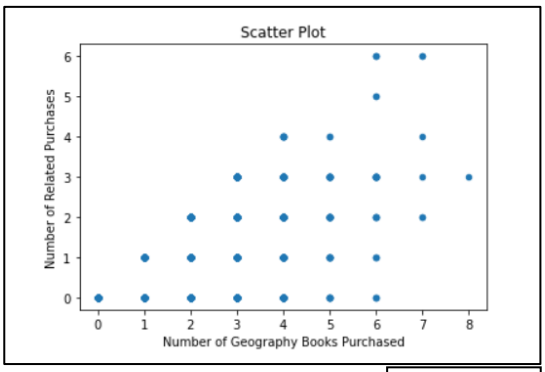

Figure 11

After looking at the scatter plots of all of the variables we wanted to look at, we want to figure out if any of the scatter plots show that there is a strong enough correlation between two variables that will help us predict one variable based off of the other variables value. The way we can see this is through a heatmap. A heat map will show a table between variables that show the correlation to the 2 decimal places, with a dark blue meaning the correlation is near one and light blue meaning the correlation between 2 variables is close to 0. We also want to use the heat map to see if our observations that we made about the scatter plots are true. The heatmap can be seen in figure 12. Based off the heatmap we can see that the observation between Number of Cook Books Purchased (CookBks) and Months since the First Purchase (FirstPurch) confirms the suspicion that they are positively correlated at 0.68. Another observation between Number of

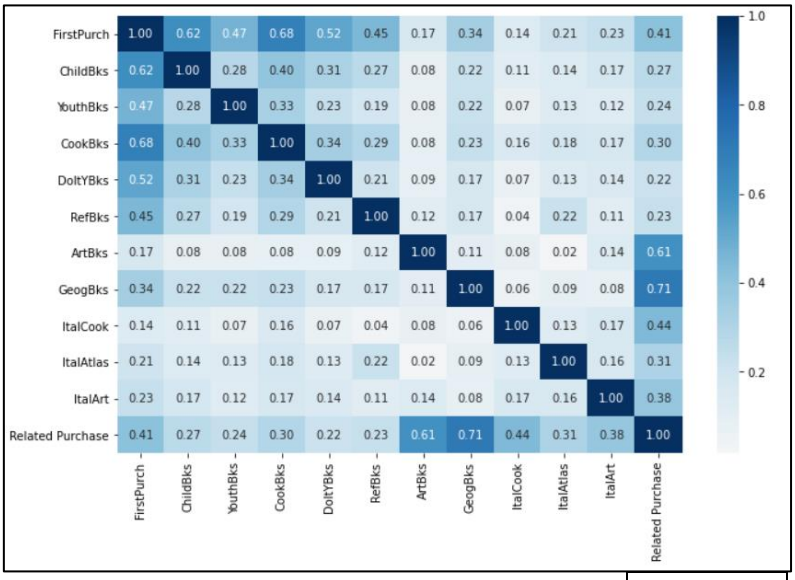

Figure 12

Geography Books purchased (GeogBks) and Number of Related Book purchases (Related Purchase) confirms our suspicion that they are positively correlated at 0.71. However, our suspicion that Number of Art Books Purchased (ArtBks) and Months since First Purchase (FirstPurch) are negatively correlated was wrong. The heatmap shows that these two variables are weakly positively correlated (0.17). After looking at the heatmap what is surprising is how positively correlated Number of Art Books Purchased (ArtBks) and Number of Related Books purchased (Related Purchase) are to each other. Although the graph does look positively correlated, it is surprising that they have a correlation of 0.61.

The last thing that we want to take a look at is the box plots of the variables, total money spent on books (M), months since last purchase (R), and total number of purchases (F) grouped by if the customer bought "*The Art History of Florence"*. We want to see this and compare if there is truly a difference between the sample of people who bought the book and the sample of people who didn't buy the book by these three most used variables. In order to see this, we have created a side-by-side boxplot of the variables M, R, and F grouped by the Florence variables, seen in figure 13. Based off of the side-by-side boxplot we can see that the for the total money spent on books variable (M) there isn't much of a difference in the data between if someone has bought the book "*The Art History of Florence"* (1.0) and if they didn't buy the book (0.0). In fact, if we look at the boxplot, the maximum total money spent is higher for people who didn't buy the book. The IQR for both people who bought the book and didn't buy the book is relatively the same, even though the median is a little bit higher for people who bought the book. From this information we can conclude that there isn't sufficient evidence that the variable of total money spent on books (M) plays a role in if a customer will buy the book "*The Art History of Florence".* By taking a look at the months since the last purchase (R) grouped by if the customer bought the book or not, we can see that although the median since last purchase is higher for customers who didn't buy the book (boxplot named 0.0), the IQR and maximum is bigger for the customers who did buy the book (boxplot named 1.0). Therefore, there isn't sufficient evidence to say that the number of months since last purchase (R) plays a role in if a customer will buy the book "*The Art History of Florence".* Finally, looking at the total number of purchases (F) grouped by if a customer purchased the book (1.0) or not (0.0) we can see that customers who purchased the book have the same median as customers who did not purchase the book. However, we can see that the IQR is larger for customers who did buy the book, than customers who didn't buy the book by 2 books. Although the IQR for customers who bought the book is larger and greater than customers who didn't buy the book, there isn't sufficient

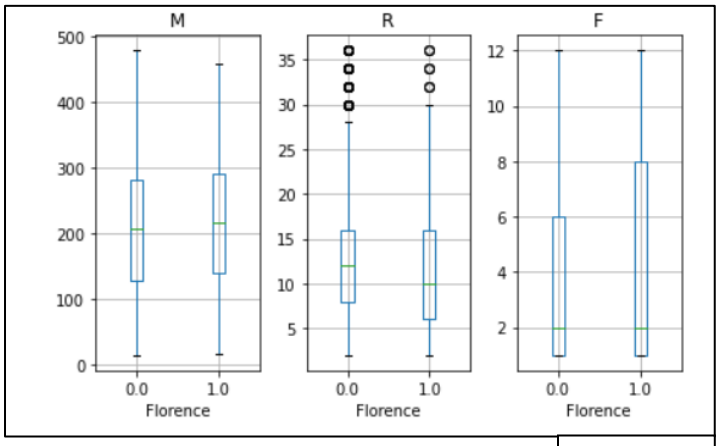

Figure 13

evidence that total number of purchases plays an important role on a customer buying the book.

Conclusion

Based on our calculations we can conclude that there isn't sufficient evidence that the variables M, R, and F play an important role in a customer purchasing "*The Art History of Florence"*. By looking at the bar graphs showing the average M, R, and F variable grouped by if a customer bought the book or not, it made it seem that the more total money spent on books, the fewer months since the last purchase, and the more total number of purchases meant that the customer was more likely to purchase the book. However, after taking a look at the boxplots at the end of our data we could get a more accurate picture of the role the M, R, and F variables play on if a customer has bought the book or not. The picture painted by the boxplots points out that by taking just the average of each variable grouped by if the customer purchased *"The Art History of Florence",* it gives us false information. Once we are able to see more information (Maximum, IQR, Median, Minimum, and Outliers) we were able to conclude that there isn't sufficient evidence that the M, R, and F variables play a role on if a customer is to buy *"The Art History of Florence*" or not.

Part 2:

Background:

In the next part of the Charles Book Club report, we will be going over using different types of classification models to accurately predict if the new book, *"The Art History of Florence*" will be bought or not. In order to do this, we have chosen 5 different classification models that we think have a good chance in correctly predicting this. Those 5 models are, Neural Networks, Decision Trees, Logistic Regression, Linear Discriminant Analysis, and K Nearest Neighbor. To give a background of these models, Neural Networks are a set of algorithms that try to recognize the relationships in a data set through a process that is similar to a human. The Decision Tree model is a sequence of branching the data set based on similarities in the data that help show predictions for a variable. The Logistic Regression model uses the logistic function to predict a binary variable (in this case if the book was bought or not). The Linear Discriminant Analysis model that finds linear combination of the variables to separate the data into two or more classes (if the data was bought or not). The K Nearest Neighbor model uses pattern recognition to find a pattern in the data set to predict a variable in the data set.

Modeling

In order to see which model can accurately predict if *"The Art History of Florence*" was bought or not we need to be able to have all of the right data fields in our data set. From part one of our report we did all of the data preprocessing we needed to do but there are still some variables that could potentially cause issues within the models causing them not to be accurate. In order for this not to happen we need to get rid of these variables. The variables that need to be removed are the Sequence number (Seq#), ID number (ID#), Didn't buy the book (No_Florence), and whether they bought the book or not (Florence). The reason we don't want the Sequence number or the ID number is that they don't mean anything in the data set. Those two variables do not predict anything, they just give us an explanation of that entry. The No_Florence and Florence data field needs to be deleted because it will tell us if the book was bought or not which would potentially cause all of the models to have 100% accuracy which would be inaccurate.

Once these variables are deleted from the data set, we need to specify a dependent variable in the data set and then delete that variable from the data set. In this case, we want to predict the Yes_Florence variable because this will tell us if the book, *"The Art History of Florence*" was bought or not. Once we identify that this variable is the dependent variable and drop that column, we have all the other variables being independent variables. The next step to allow us to run our 5 models is to standardize the data. This allows us to have the mean to be 0 for all of the variables and give us a number that is the unit variance. This helps the data read and process more accurately within the models. After the data is standardized, we need to split the data into training data and testing data. We split the data up 80% training data and 20% testing data to use the testing data to see how accurate our models are compared to the actual classification (bought the book or did not buy the book). The reason we split the data up is because we also want to model the classification models with the training data and then do our model evaluation with the testing data to see how accurate our classification model truly is.

After the data is split 80% and 20%, we wanted to scale the data for the Neural Network classification model. What this does is helps the Neural Network read the data better and perform at a higher rate. When we scale the standardized data, this makes all values to be within 0 and 1. When this step is complete the classification models can now have modeling performed on them. When doing modeling we want to use the training data on the classification models, so that the testing data can be used on them later to see how accurate the models are. Once we have all of the classification models all modeled using the training data, we wanted to see how accurate the modeling made each classification model, which we made confusion metrix to see how many times the modeled data would predict the book was or was not bought compared to if the book was or was not bought. When looking at the data modeled for Neural Networks, we can see the confusion metrix in figure 14. What this tells us is that the Neural Network classification

Confusion Matrix (Accuracy 0.9234)

Actual	Prediction 0	1
0	2913	11
1	234	42

Figure 14

model using the training data accurately predicted if *"The Art History of Florence"* was not bought 2913 times and correctly predicted if the *"The Art History of Florence"* was bought 42 times. The model predicted 11 times that the book was bought but it wasn't bought and the model also predicted the book wasn't bought 234 times when it in fact was bought. In the next classification model, Logistic Regression, we can see the confusion metrix in figure 15. What this tells us is that the Logistic Regression classification model using the training data accurately predicted if *"The Art History of Florence"* was not bought 2923 times and correctly predicted if the *"The Art History of Florence"* was bought 8 times. The model predicted 1 time that the book was bought but it wasn't bought and the model also predicted the book wasn't bought 268 times when it in fact was bought. This is an interesting confusion matrix because it is showing that the model only predicted the book was bought 9 times but in reality, the book was bought 276 times. The next classification model's confusion matrix to look at is K Nearest Neighbor's in figure 16. This tells us is that the K Nearest Neighbor classification model using the training data accurately predicted if *"The Art History of Florence"* was not bought 2912 times and correctly predicted if the *"The Art History of Florence"* was bought 20 times. The model predicted 12 times that the book was bought but it wasn't bought and the model also predicted the book wasn't bought 256 times when it in fact was bought. The Linear Discriminant Analysis classification model confusion matrix can be seen in figure 17. This tells us is that the Linear Discriminant Analysis classification model using the training data accurately predicted if *"The Art History of Florence"* was not bought 2912 times and correctly predicted if the *"The Art History of Florence"* was bought 17 times. The model predicted 12 times that the book was bought but it wasn't bought and the model also predicted the book wasn't bought 259 times when it in fact was bought. This is very interesting because we can see that when the K Nearest Neighbor and Linear Discriminant Analysis models are used using the training data, they had the same number of accurately predicted books not bought and the same number of books predicted to be bought when they weren't bought. The only difference between the two models is that the K Nearest Neighbor accurately predicted 3 more books bought than the Linear Discriminant Analysis model. The last classification model that we want to model using the training data is the Decision Tree model. In order to do this, it is a little different than the other models. In order to model the Decision Tree model, the best we want to grab the best Decision Tree. To find the

Confusion Matrix (Accuracy 0.9159)

Actual	Prediction 0	1
0	2923	1
1	268	8

Figure 15

Confusion Matrix (Accuracy 0.9163)

Actual	Prediction 0	1
0	2912	12
1	256	20

Figure 16

Confusion Matrix (Accuracy 0.9153)

Actual	Prediction 0	1
0	2912	12
1	259	17

Figure 17

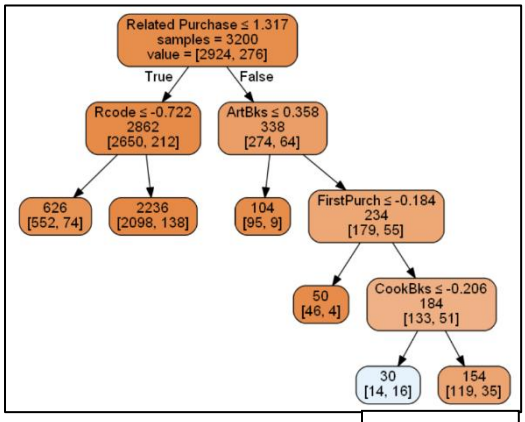

Figure 18

best Decision Tree we have to use the grid search function and set parameters and the cross-validation to find it. The parameters we used for max depth was the range between 2 and 10, the minimum samples split was a range between 10 and 20, and the minimum impurity decrease was 0.0009, 0.001, and 0.0011. The cross-validation we used to find the best tree was set to 10. The best decision tree plot is shown in figure 18. Once we found the best decision tree, we could then model the classification model using the training data. The results for the Decision Tree confusion matrix can be seen in figure 19. This tells us is that the Decision Tree classification model using the training data accurately predicted if *"The Art History of Florence*" was not bought 2910 times and correctly predicted if the *"The Art History of Florence*" was bought 16 times. The model predicted 14 times that the book was bought but it wasn't bought and the model also predicted the book wasn't bought 260 times when it in fact was bought. The confusion matrix results for the Decision Tree model are very similar to the results from the Linear Discriminant Analysis model and the K Nearest Neighbor model. This brings us to the end of our data modeling and it is time for us to move to our model evaluation to see which model is the most accurate and best to use.

Confusion Matrix (Accuracy 0.9144)

	Prediction	
Actual	0	1
0	2910	14
1	260	16

Figure 19

Model Evaluation

The first way we are going to look into model evaluation process we want to look at is the 10-fold cross validation. This will partition the data into the training data and then test that model using the testing data (we use the testing data for all of the model evaluation). When looking at the 10-fold cross validation we want to find the model with the highest average accuracy. Below we can see the 10-fold cross validation scores for each of the 5 models.

Decision Tree

```
[0.8625 0.9    0.875  0.9    0.8875 0.8625 0.85   0.875  0.8125 0.85  ]
Average accuracy score: 0.868
```

Neural Network

```
[0.825  0.925  0.875  0.8625 0.875  0.9    0.925  0.925  0.8625 0.8875]
Average accuracy score: 0.886
```

Logistic Regression

```
[0.925  0.925  0.925  0.925  0.925  0.925  0.925  0.925  0.9125 0.9125]
Average accuracy score: 0.922
```

Linear Discriminant Analysis

```
[0.925  0.925  0.925  0.925  0.9    0.9125 0.925  0.925  0.9125 0.9125]
Average accuracy score: 0.919
```

K Nearest Neighbor (Knn)

```
[0.925  0.925  0.925  0.925  0.9125 0.925  0.925  0.925  0.8875 0.875 ]
Average accuracy score: 0.915
```

From the results shown above we can see that the Logistic Regression model is the best model to use using the 10-fold cross validation. The reasoning behind this answer is because of the average accuracy score. Using the 10-fold cross validation we want to find the average accuracy score not just the highest. This will show us which model on average will have the highest accuracy and chance of being right when predicting if the book *"The Art History of Florence"* was bought or not. The Logistic Regression model is best because it has an average accuracy of 0.922 which means that the model on average is correct 92.2% of the time. The Linear Discriminant Analysis is at 91.9% correct on average and the K Nearest Neighbor model is 91.5% correct on average. The worst model is the Decision Tree model that is correct 86.8% of the time on average. Although none of these results are bad based off of the confusion matrix, they probably all have high accuracies because of them predicting more books to not be bought than there really are (way more *"The Art History of Florence"* books not being bought than being bought). One way that we can see how much each of the classification models are misclassifying the data, which shows at what percent are the models showing that the *"The Art History of Florence"* was not being bought when it actually is being bought and when the book is being bought when it is actually not being bought. If we find this information it is the misclassification rate for the model. In order to find this using, the 10-fold cross validation we need to take 1 – Average Accuracy rate using the 10-fold cross validation from above. We were able to calculate the data and once we did, we got figure 20 as a result. The misclassification rate that is the best based off of the figure shown to the right is the Logistic Regression model which is what we would have expected based off of our results for the 10-fold cross validation average accuracy. This shows us that the Logistic Regression on average misclassifies whether the book *"The Art History of Florence"* was bought or not. The misclassification rate using the 10-fold cross validation for the Logistic Regression model is 0.078 which is 7.8% of the time. This means that 7.8% of the time the Logistic Regression model will misclassify either the book *"The Art History of Florence"* as bought when it isn't bought or it will misclassify the book as not bought when it is bought. The next best model is the Linear Discriminant Analysis model at 8.1% and the worst model is the Decision Tree model at 13.2%. We would like to get this under 5% but the Logistic Regression model is the best based off of the 10-fold cross validation.

```
Misclassification rate for MLP: 0.114
Misclassification rate for decision tree: 0.132
Misclassification rate for kNN: 0.085
Misclassification rate for linear discriminant analysis: 0.081
Misclassification rate for logistic regression: 0.078
```

Figure 20

To build off our model evaluation of the 10-fold cross validation the ROC curve for each model serves as another way to see which model is best when predicting if the book *"The Art History of Florence"* was bought or not. The ROC curve shows us the curve between true positives and false positives in the data set. The curve will show us on the graph where the classification models accurately give true positives and when there starts to be more false positives than the increase in true positives. We can see our results from running the ROC curve using the testing data in figure 21. The results that we got from the ROC curve are very shocking. From looking at all of the past information from model evaluation we had a fairly good indication of what the ROC curve would look like but when running the code for the ROC curve it shows us that no classification model has a score over 0.530. This is very bad and shows that the model only predicts a true positive 53% of the time. However, based off of the ROC curve shown we can see that the Neural Network (MLP) ROC curve is the most accurate at a curve of 0.526, which shows that the model will show a true positive 52.6% of the time. At the beginning every model besides the K Nearest Neighbor (kNN) model show that they show all true positives for the first 0.05 of the data then after that is about even in a true positive or a false positive. One big thing that we can conclude from the ROC curve is that there could be a problem with the data set. There is no reason for the ROC curve to have there low of scores based off of the 10-fold cross validation and the confusion matrix, which shows that the Charles Book Club data may be wrong or have issues with it.

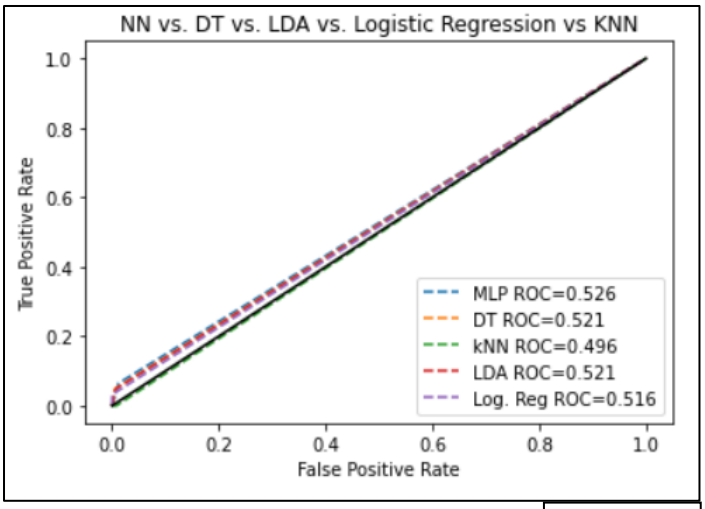

Figure 21

The last model evaluation that we are going to look at is the lift chart for each of the classification models. This will help us see if the model is doing a good job at predicting the variable of "Yes_Florence" (If *"The Art History of Florence"* was bought or not). We want to see the lift chart have a high bar at the 10 percentile in the model with it diminishing to almost 0 by the 100 percentile. This will show us if the classification model is working properly. If the bars are all even for the percentiles, this shows us that the model is not working at all and the model isn't binning the data any better than if you randomly binned the data. If the bars on the lift chart are out of order than it is showing that the classification model isn't doing a good job at predicting if someone bought the book *"The Art History of Florence"* or not. The first lift chart that we looked at is the Neural Network Lift Chart.

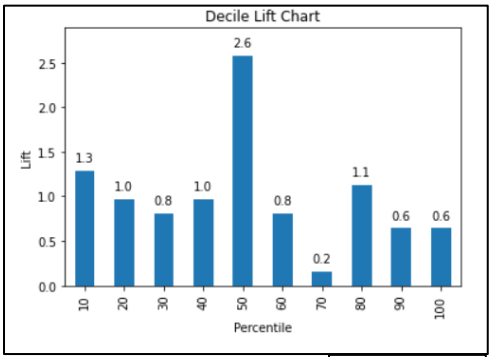

Figure 22

We can see what the lift chart looks like in figure 22. From this lift chart for Neural Networks we can see that the most of the bars have close to the same height with the 50-percentile shooting up to 2.6. This shows us that the Neural Network classification model isn't working very well and binning the data very well. This shows us that there needs to be an improvement to the Neural Network classification model. The next lift chart that we will look at is the Logistic Regression Lift Chart. This lift chart can be seen in figure 23. The Logistic Regression Lift Chart shows us that the bars seem to be in a different order and mixed up a little bit with varying height on each of the bars. This tells us that the model is not doing a good job at predicting the actual response in the data set, which could cause the ROC curve to not have a score that is high. The next lift chart that we will look at is the K Nearest Neighbor (kNN) model. You can see the results of the Lift Chart in figure 24. The lift chart for the kNN model shows us the same thing as the Logistic Regression Lift Chart. It shows us that the kNN model is not doing a good job at predicting the actual response in the data set which in this data set is if someone bought the book *"The Art History of Florence"* or not. This is not good because we would like our chart to have a high value for the first 10 percentile and then diminish to around 0 by the time it is at the 100 percentile. With the varying heights of the ROC curve it shows us that the kNN model is not doing a good job at predicting the dependent variable (Yes_Florence). The next lift chart that we will look at is the Linear Discriminant Analysis Lift Chart. The results from this chart are found in figure 25. The Lift Chart for the Linear Discriminant Analysis shows us that the model is in between not being a good model at all not being much better than guessing (flat graph) and not predicting if the book *"The Art History of Florence"* was bought or not accurately (varying bar heights). Either way this reflects the low ROC curve score, which could mean that the data that is being used may not be good data given to us from the Charles Book Club. The last lift chart that we are going to look at is the lift chart for the Decision Tree model. The lift chart can be seen in Figure 26. The Decision Tree Lift Chart shows us again that it is in the middle of being not much better than guessing and not predicting if the book *"The Art History of Florence"* was bought or not. Again, this trend seems to be continuing throughout all of the

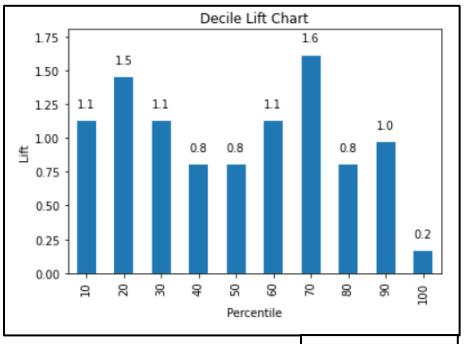

Figure 23

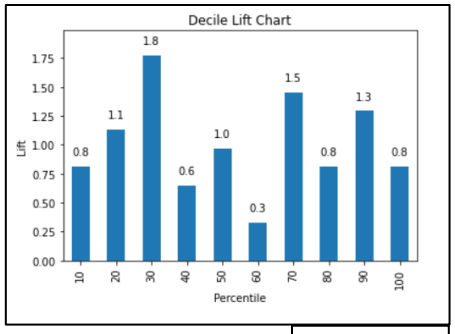

Figure 24

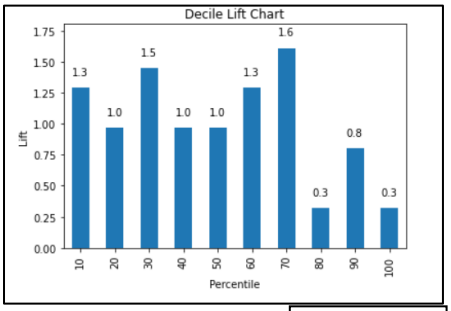

Figure 25

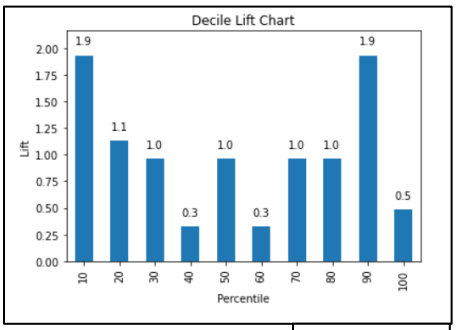

Figure 26

model where they are not doing a good job at predicting the independent variable (Yes_Florence) very well. From all of the lift charts for the classification models, we can conclude that there could be a problem in the data because of the results we found in the Lift Charts.

Conclusion

In this part of the Charles Book Club report we have found that predicting if the book *"The Art History of Florence"* will be bought is very tough given the data set that we have been given. If we had to use a classification model to predict if the book will be bought or not, we would recommend the Logistic Regression model because of the highest average accuracy rate using the 10-fold cross validation. From the ROC curve and the Lift Charts we can see that there could be a problem with the Charles Book Club data because the ROC curve should have a higher score than it did for all of the classification models. The Lift charts should all have a high bar for the first 10 percentile and then diminish to around 0 for the 100 percentile. From our results this is not the case and the charts show that the models are not good at predicting if the book *"The Art History of Florence"* was bought or not, which calls for the data to possibly be bad. We would recommend that the data given is to be looked at because the results show us that there could be a potential flaw in the data set.

Part 3

Background:

In this part of the project the main focus is to find a classification model that will be able to give an accurate measure for the "M" (total money spent) variable when the "Yes_Florence" variable is equal to 1, which means *The Art History of Florence* was bought. The given classifications that are going to be used in this section are Forward Selection, Backward Elimination, Stepwise Regression, Decision Trees, Random Forest, and Boosted Trees. We were going to use the Exhaustive Search Approach but decided to not use it when it caused our code to not run properly. To give a background on all of the classification models, the Forward Selection model tests the variables one at a time and sees if that variable is significant to improving the model. Once the model doesn't get any better than the process stops. The Backward Elimination model deletes any independent that do not contribute to the equation. The Stepwise Equation the model will throw the best predictor variable for determining the variable we want to find at each step and so on and so on. In the Decision Tree model, the model takes a series of queries that cause the outcome of the previous tests to influence the test performed next. The Random Forest model makes a bunch of decision trees and then outputs the mode of all the trees created. The Boosted Trees model uses a learning algorithm to build a stronger tree to give the best output. In this part, we will use all of these classification

models to see which model most accurately predicts the "M" variable if someone bought the book "*The Art History of Florence*".

Classification Modeling

In order to achieve our goal of finding the best classification model for predicting the "M" variable, there are a couple things we needed to accomplish first. The first thing that we needed to do was take out a couple of variables that did not mean much to us in the data set. Those variables were Seq#, ID#, No_Florence (when people didn't buy the book), Florence (tells us if the books was purchased or not), Mcode, Rcode, and Fcode. Although the M, R, and F codes are useful information for these models, we already created dummy variables for these variables and no longer need them. The next step was for us to only use cases where the Yes_Florence variable was equal to 1. The reason we needed to do that is we only wanted to see instances where people bought the book "*The Art History of Florence"* and the Yes_Florence variable tells us this. Once we filtered down the data and deleted the variables that we did not need we then needed to set the dependent variable as "M" and split the data 80% training and 20% testing. The reason we needed to set the dependent variable as "M" is because that is the variable that we are going to try and predict using the models. Once we set the dependent variable and split the data, we then were ready to start our analysis.

The first thing we needed to do was predict "M" using all of the variables. This is an important step because we needed each model to predict what "M" would be using the data we had in the training data. This would then be able to be used later and see how the classification models performed against the testing data. The next step that we needed to figure out was the top 5 most important variables in the Forward Selection, Backward Elimination, and Stepwise Regression model. When we ran our code, we found that the Forward Selection and Stepwise Regression models had the same 5 most important variables and in the same order. The order that they went were Mcode_5.0, Mcode_4.0, F, GeogBks, and Mcode_3.0. This is expected because the Mcode variables show how much money was being spent in that order, so it is likely those variables will be important since they are the money spent by a customer when we are trying to predict money spent by a customer. The F variable being important is also not much of a shocker because that shows the number of purchases made by a customer, so likely the higher number of purchases you have, the higher amount you spend. Also, the GeogBks (Geographic Books) was an important variable which is somewhat surprising because it didn't see conclusive that someone who bought a geographic book spent a lot of money or a lot less money. For backward elimination, the variables were very similar in an order of F (Frequency), ChildBks (Children Books), GeogBks (Geographic Books), Mcode_3.0, and Mcode_4.0. I thought these results were a little more surprising although the backward elimination model had 4 of the same variables. The reason for this is that Children Books and Geographic Books are both meaningful in determining "M".

The next step for us was to plot the top most meaningful variables for the Decision Tree, Boosted Tree, and Random Forest mode. To start off we will talk about the Decision Tree model. You can take a look at Figure 27 to show you the plot we had. These results tell us that the five most important variables that are in the Decision tree model is CookBks, GeogBks, F, Mcode_4.0, and Mcode_5.0. The most important variable out of the 5 is for sure mcode_5.0, with an importance of 0.8. This is significant because we can see that Mcode variables are meaningful in determining what the M variable will be. These results are similar to the Forward Selection and Stepwise Regression models. The next plot that we will be taking a look at is the Random Forest model. The results that we got from our code can be represented in Figure 28. As we can see to our right, the order and variables are the exact same as the Decision Tree model, but the only difference is that the last variable is FirstPurch for the Random Forest model and CookBks for the Decision Tree. These results are again what we expected. The last and final model that we had to plot was the Boost Tree model. We can see the results that we got in figure 29. The Boost Tree model is also very similar to the Stepwise Regression model and especially the Random Forest model. For this model the results are almost identical to the Random Forest variables with Mcode_5.0 being the most important, Mcode_4.0, F, FirstPurch, and GeogBks then followed in that order.

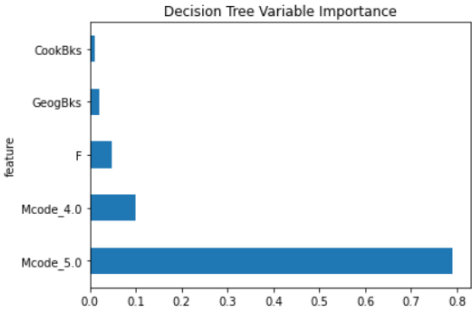

Figure 27

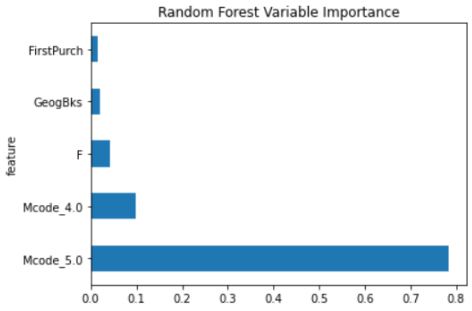

Figure 28

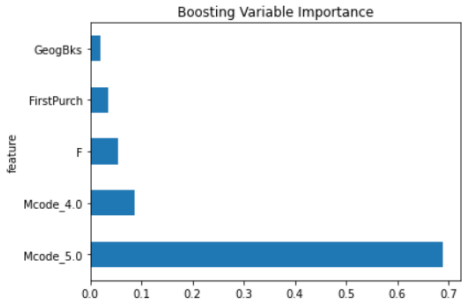

Figure 29

When looking at all of the plots that we made and the outcomes we got, it is reasonable to say that there was a significant overlap between important variables within the models. The most important variable to almost all of the models is the Mcode_5.0 variable. This is very predicted because this variable shows that the customer bought a lot of items from the store. The next important in Mcode_4.0 (second biggest), followed by the F variable, and mainly GeogBks (Geography). This is surprising because you wouldn't think that someone to buy Geography Books and buy *The Art History of Florence*. Also, the Mcode_3.0 does have some significance but for the most part the universal top 5 importance variables throughout all of the classification models. We would say the most important variable is the Mcode_5.0, followed by Mcode_4.0, F, GeogBks, and Mcode_3.0. Since many of the models have the top 4 variables

listed, we can conclude that it is likely these variables are truly the most important when predicting if the "*The Art History of Florence*".

The next step in the process was to use the top 5 variables on the testing data to predict "M". There are a couple steps that we needed to do to ensure that we could do this. The first is that for each of the models we had to fit a regression with the training data with the top 5 variables. Next, we had to select the columns that we wanted to use in the testing dataset, which were the top 5 variables we found from above. The next step was to perform the model prediction. Once we did all of those steps, the last step was to do the model evaluation for all of the models. We will start by talking about the Random Forest model. Looking at Figure 30, we can see the results that we get. From the model evaluation we can see that the mean error is -1.69, the root mean squared error is 45.02, the mean absolute error is 35.08, the mean percentage error is -4.46, and the mean absolute percentage error is 20.56. To take a look at the next model evaluation, we will be looking at the Boosted Trees Model. From the model evaluation in Figure 31, we can see that the mean error is -5.40, the root mean squared error is 48.70, the mean absolute error is 36.77, the mean percentage error is -5.15, and the mean absolute percentage error is 20.55. The next model that we will evaluate is the Decision Tree model. From the model evaluation in Figure 32, we can see that the mean error is 0.69, the root mean squared error is 46.31, the mean absolute error is 36.46, the mean percentage error is -3.74, and the mean absolute percentage error is 21.27. The next model that we will look at is the Forward Selection model. From the model evaluation in Figure 33, we can see that the mean error is -3.24, the root mean squared error is 41.08, the mean absolute error is 32.56, the mean percentage error is -5.86, and the mean absolute percentage error is 17.90. The next model we will look at is the Backward Elimination model. From the model evaluation in Figure 34, we can see that the mean error is 9.66, the root mean squared error is 56.94, the mean absolute error is 44.65, the

```
**********Random Forest**********

Regression statistics

                    Mean Error (ME) : -1.6887
     Root Mean Squared Error (RMSE) : 45.0243
          Mean Absolute Error (MAE) : 35.0762
        Mean Percentage Error (MPE) : -4.4639
Mean Absolute Percentage Error (MAPE) : 20.5599
```

Figure 30

```
**********Boosted Trees**********

Regression statistics

                    Mean Error (ME) : -5.3961
     Root Mean Squared Error (RMSE) : 48.6952
          Mean Absolute Error (MAE) : 36.7697
        Mean Percentage Error (MPE) : -5.1534
Mean Absolute Percentage Error (MAPE) : 20.5495
```

Figure 31

```
**********Decision Trees**********

Regression statistics

                    Mean Error (ME) : 0.6941
     Root Mean Squared Error (RMSE) : 46.3075
          Mean Absolute Error (MAE) : 36.4616
        Mean Percentage Error (MPE) : -3.7421
Mean Absolute Percentage Error (MAPE) : 21.2746
```

Figure 32

```
**********Forward Selection**********

Regression statistics

                    Mean Error (ME) : -3.2420
     Root Mean Squared Error (RMSE) : 41.0761
          Mean Absolute Error (MAE) : 32.5651
        Mean Percentage Error (MPE) : -5.8550
Mean Absolute Percentage Error (MAPE) : 17.8953
```

Figure 33

```
**********Backward Elimination**********

Regression statistics

                    Mean Error (ME) : 9.6634
     Root Mean Squared Error (RMSE) : 56.9392
          Mean Absolute Error (MAE) : 44.6498
        Mean Percentage Error (MPE) : -14.7286
Mean Absolute Percentage Error (MAPE) : 36.4134
```

Figure 34

mean percentage error is -14.73, and the mean absolute percentage error is 36.41. The last model that we will look at is the Stepwise Regression model. From the model evaluation in Figure 35, we can see that the mean error is -3.24, the root mean squared error is 41.08, the mean absolute error is 32.57, the mean percentage error is -5.86, and the mean absolute percentage error is 17.90. After this step we are finished with our job. The next and final step is to figure out which model is best.

```
************************Stepwise Regression************************

Regression statistics

                    Mean Error (ME) : -3.2420
      Root Mean Squared Error (RMSE) : 41.0761
           Mean Absolute Error (MAE) : 32.5651
         Mean Percentage Error (MPE) : -5.8550
Mean Absolute Percentage Error (MAPE) : 17.8953
```

Figure 35

Conclusion

There are only two models that we want to take a look at when deciding which model is the best at predicting the "M" variable for when someone bought the book, *The Art History of Florence*". Those two models are the Random Forest model and the Decision Tree model. The reason that we have chosen those two models is that they have the lowest mean error or average error when determining the "M" variable. Because we want to help the book store, we do not want to over estimate or under estimate the amount someone will spend in the store, we want it as close as possible. The Random Forest model has a mean error -1.69, compared to the Decision Tree model which has a mean error of 0.69. This shows that the Decision Tree model has the lowest average error out of all the variables when predicting the "M" variable. This means that it is the closest model to actually predicting the correct amount for "M". We would want to use this model because of its accuracy to the mean and it is the best model for this situation. An interesting side note is that the Stepwise Regression model and Forward Selection model had the same model evaluation. This may be in part to them both having the same important variables.