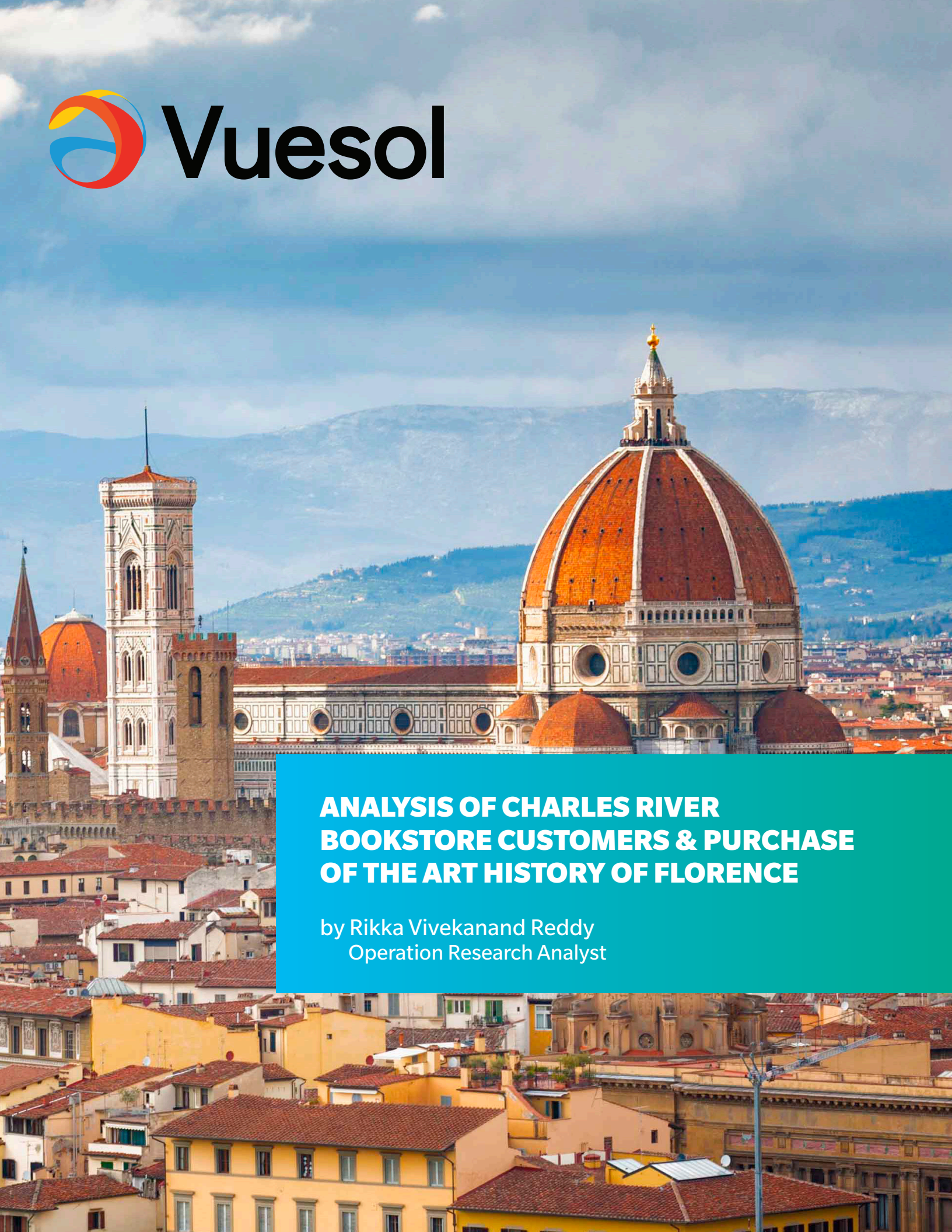


Vuesol

# ANALYSIS OF CHARLES RIVER BOOKSTORE CUSTOMERS & PURCHASE OF THE ART HISTORY OF FLORENCE

by Rikka Vivekanand Reddy
Operation Research Analyst

# Executive Summary

The Charles River Bookstore (CRB) is an independent bookstore struggling to stay afloat in a competitive, rapidly evolving market. It contends with large brick and mortar and online competitors, and the rise of e-books, all of which offer customers rock-bottom prices. Many of the books CRB sells offer thin profit margins. In spite of this intense competition, CRB has a loyal customer base, particularly among book clubs that buy these lower margin books, and has access to a large database of customers interested in purchasing from independent bookstores. The company learned of a coffee table book entitled The Art History of Florence that has potential to bring significant profit. The store must be creative and strategic in how it uses its limited resources to market and sell books that will increase profit margins, but it is weary of offering a new kind of book that could alienate their loyal customer base, many of whom turn to it to purchase their book club readings.

**Objectives:** The objective of this analysis is to help CRB understand the preferences and buying patterns of its customers, and identify opportunities to offer more diverse products like the Florence book to deliver greater profitability to the company. The data mining goal is to identify a model that can predict the probability of purchasing the Florence book and develop a profile of customers who will purchase the book. The company can utilize this information to develop a more targeted marketing approach that efficiently uses resources to increase sales of the book while maintaining their loyal book club customer base. The results will be decreased sunk costs, reduction of wasted resources, increased efficiency and greater profitability.

**Summary of Results:**

Who the customers are: Only 13% of CRB customers are book club members. This suggests that fears of alienating a large portion of customers are not accurate. Of those customers, 75% are women and almost half bought the Florence book. Overall, 70% of their customers are women. We also saw that book club members tended to buy more books across all book categories than non-book club members.

Profile of Florence book buyers: Men bought the book more than women. Customers who purchased art books were very likely to purchase the Florence book. There was also high correlation between customers that purchased the Italian cooking book, as well as recent shoppers, frequent shoppers and those shoppers who had been customers the longest. The analysis also revealed who is more likely to not buy the book: those customers who had purchased more than one children's book or youth book were the least likely to buy the Florence book. Those who had purchased DIY books or who had not purchased from CRB in several months were also less likely to buy the Florence book.

Profit potential: After completing our analyses, we identified a model that result in a net profit of $1,168,750. This model would also reduce the number of incorrect purchasing predictions, which will help accurately predict incoming revenues and reduce sunk costs, inventory costs and shipping costs.

**Conclusions:** The analysis revealed that there are many opportunities for CRB to maximize its profitability by offering The Art History of Florence to its customer database while maintaining a positive relationship with its important book club member customers.

# Index

## BA Lifecycle:

The BA Lifecycle process was important to clarify different aspects of the Charles River Bookstore and to develop our analysis plans. We have outlined how we applied all six phases of the BA Lifecycle methodology and have directly integrated our data visualizations, logistic regression and decision tree analyses, and model comparisons.

## Discovery:

Business problem: The Charles River Bookstore (CRB) operates in a market of large competitors and an on-line environment that have whittled profit margins and increased consumer buying power. These conditions have made it difficult for independent book sellers like CRB to compete and offer books that will deliver higher profitability. CRB must identify ways to remain profitable even as larger companies slash prices for consumers looking for the best deal. CRB also needs to understand the preferences of its customers in order to offer the right product mix, market to them appropriately and drive sales.

The situation: The company learned of a coffee table book entitled The Art History of Florence that has potential to bring significant profit. CRB has a loyal customer base, particularly among book clubs, and has access to a large database of customers interested in purchasing from independent bookstores. It is weary of offering a new kind of book that could alienate their loyal customer base, many of whom turn to it to purchase their book club readings.

Business objectives and data mining goals: The goal of this project is to develop a model that will best predict if a customer will buy or not buy the Florence book. We also want to build a profile of customers that will buy the book in order to efficiently market to those individuals in order to maximize our profitability.

Success/failure criteria: Our goal is to minimize "false positives"; that is the number of people we predict will buy the book but actually do not. This incurs a cost to the company and disrupts revenue predictions. We want to maximize correct predictions of purchasing the book and maximize net profit.

Key risks: The greatest risk to CRB is alienating what it perceives to be its key customers: book club members who are avid readers and loyal customers. The analysis may reveal that customers might not respond well to offering such a book. If it does not get a clear understanding of their customers, CRB may also risk wasting marketing resources. Another risk is that this analysis is particular to the Florence book. The model that the analysis reveals may not apply to all coffee table books of a different topic, such as nature or architecture.

Key stakeholders: The Charles River Bookstore, loyal customers, frequent customers, book club members, and new customers that CRB might attract through the sale of the new book.

## Data Preparation:

Before building our models, we wanted to ensure that our data was high-quality. We re-categorized ID, Gender and Florence from continuous to nominal data. There were no missing values or duplicate records. We also reviewed the summary statistics and box plots and found no unusual outliers. In some categories, there were higher maximum values, such as Monetary - Total money spent on books ($477) and First Purchase - number of months since first purchase (99 months, which equates to just over 8 years). We felt these were reasonable values. We also recoded the gender and Florence categories to make them easier to understand and analyze. See the summary statistics in the Appendix.

**Descriptive statistics:** More than 70% of the Charles River Bookstore customers are women. The total monetary value of money spent on books was fairly normally distributed: the average amount and median value of this variable was $207, suggesting that there are only a small proportion of customers spending large sums of money at the store. The average number of months since the last purchase from the store was 13.5 months. This data was skewed significantly to the right, meaning that the mean is greater than the median (12 months), and there are many larger values on the right side of the scale. This indicates that a number of customers have not purchased books in well over a year. It suggests that while many customers do not shop frequently, there are some that do. Number of months since first purchase was also skewed to the right.

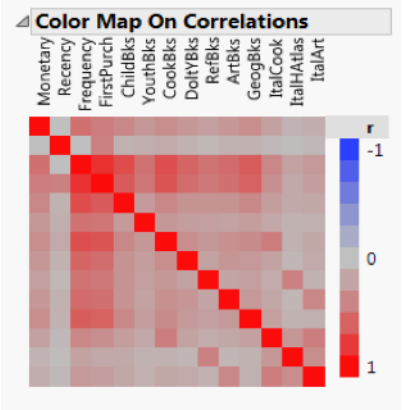
The average customer had not purchased anything in 26.5 months, and the median value was 20 months. This suggests that they have a number of newer customers who have bought for the first time ever within the last two years. These customers may not be as affected by a change in inventory, such as the inclusion of the Florence book. It also shows that there are a number of customers who made their first purchase several years ago.

Frequency and total number of purchases of the different book categories are all distributed to the right, with the average number of purchases being relatively small, and a smaller portion of customers making more frequent purchases, and purchases of greater numbers of books in the different book categories. See JMP file for all box plots.

**Data visualization:** These visualizations helped us to get a preliminary understanding of who the CRB customers were and their purchasing habits.

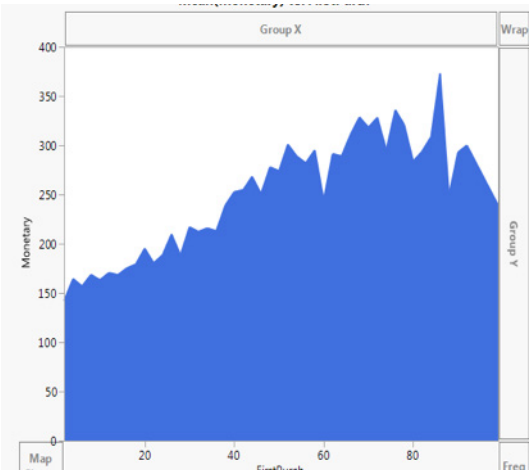
**Correlation Heat Map:**

This correlation heat map will be discussed in more detail later in model planning. It shows which variables are correlated with each other. Most variables do not have a strong correlation, though some are strongly positively correlated, which may affect how they interact with each other in our models.



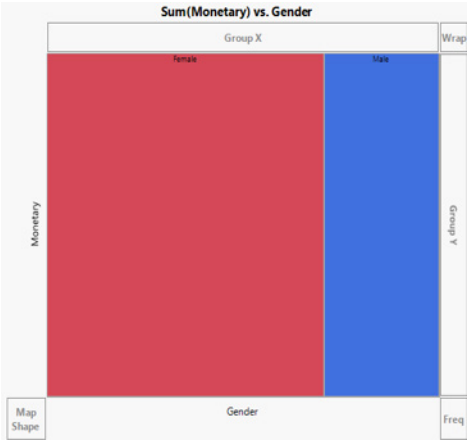

**First Purchase and Money Spent Over Time:**

This graph shows the relationship between total money spent and number of months since first purchase. The amount spent rises as the amount of time since first purchase increases. This suggests that the store has very loyal customers who continue to buy over time.



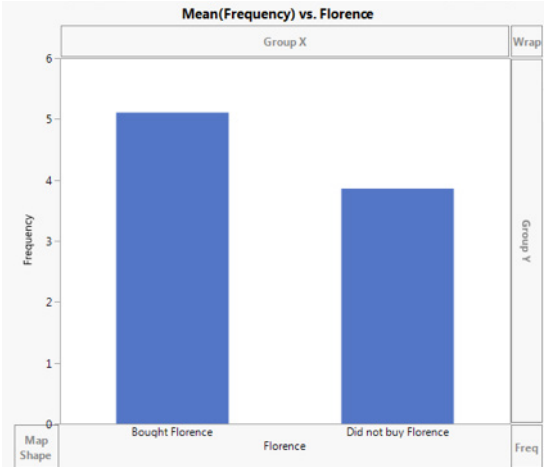

**Money Spent, By Gender:**

This chart shows that women spend more money than men. This fact may be interesting to explore as the company looks to find who will purchase the book and what types of books to offer.



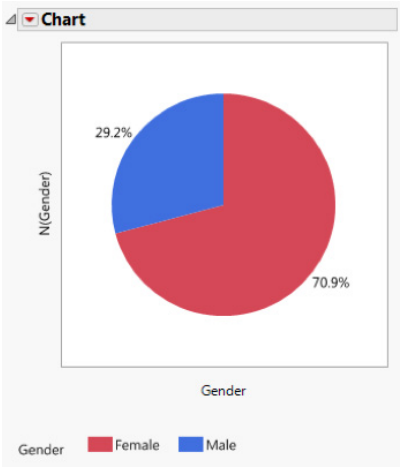
**Frequency and purchase of Florence book:**

This chart shows that frequent CRB customers also bought the Florence book more on average. This finding may improve marketing efforts of the Florence book to repeat customers.



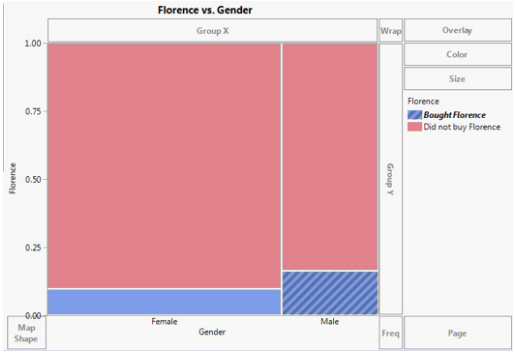

**Customers, by gender:**

Most of their current customers are women:



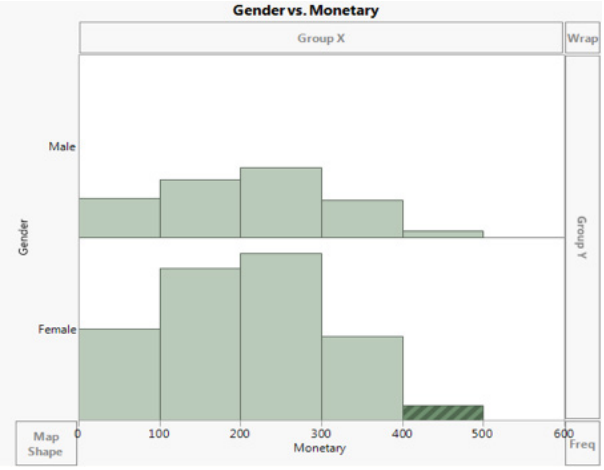

**Florence purchase by gender:**

This chart shows that more men bought the Florence book than women. This is an interesting finding, given that, on average, women spend more money at the bookstore than men. This may indicate that men are more likely to buy this type of book than women.



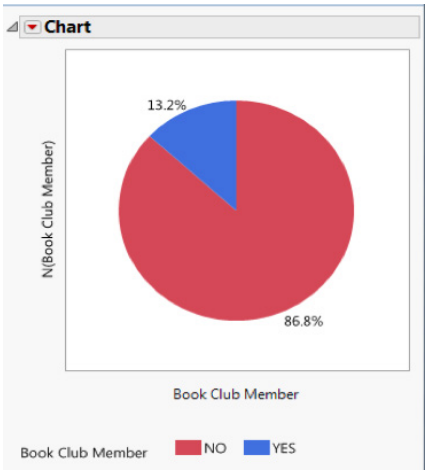
**Overall spending, by gender:**

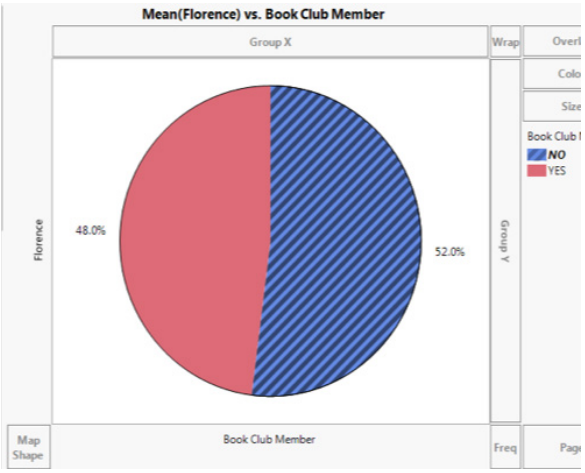
This chart shows that spending habits of men and women are similar. Though there are more women customers (indicated by the taller bars), the majority of customers spend between $100 and $300 regardless of gender.




**Book club members & Percentage of book club members that bought Florence:**

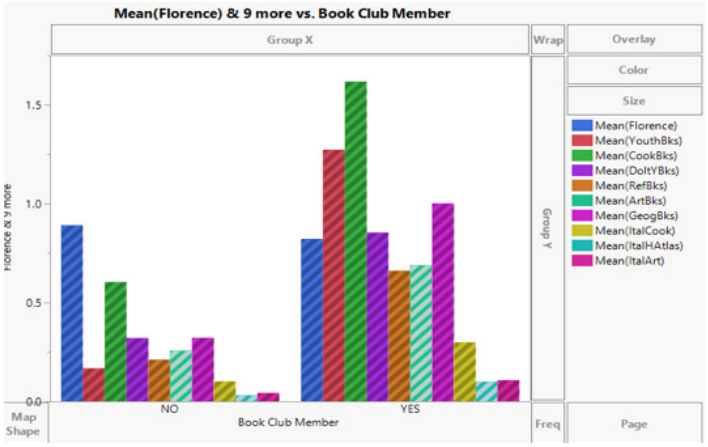
The chart on the left shows that only 13% of total customers are members of a book club. The chart on the right shows that almost half of the book club customers bought the Florence book.




**Purchases, book club vs. non-book club:**

We can observe from this graph that if someone is a book club member they tend to buy more books when compared to someone who is not a book club member.




## Model Planning:

We plan to use JMP to perform a Logistic Regression and a Decision Tree Analysis as methods to understand the buying patterns of customers to predict which customers will buy the Florence book. We are interested in identifying a profile of the type of customer that would buy the Florence book. By creating this profile, the Charles River Bookstore would be able to better target certain customers with the appropriate marketing campaigns. This model will be supervised learning, as there is the predefined outcome group of did/did not buy the Florence book. We will be searching for a pattern among the variables provided to use in order to understand customer buying habits.

We analyzed the relationships between variables in order to understand if more than one of the independent variables affects each other. As seen in the correlation chart below, we found that there is a strong correlation between the F variable and FirstPurch, ChildBks, and CookBks variables. These variables had a strong positive correlation with the frequency of the total number of purchases. This makes sense because the other variables are the number of purchases from the category and this directly affects the total number of purchases. What is interesting is that the other types of books do not have a strong correlation with the frequency of the total number of purchases.

**Correlations**

| | Monetary | Recency | Frequency | FirstPurch | ChildBks | YouthBks | CookBks | DoItYBks | RefBks | ArtBks | GeogBks | ItalCook | ItalHAtlas | ItalArt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Monetary | 1.0000 | 0.0119 | 0.5220 | 0.4523 | 0.3896 | 0.2887 | 0.3521 | 0.2960 | 0.2447 | 0.2930 | 0.3484 | 0.2319 | 0.1020 | 0.1428 |
| Recency | 0.0119 | 1.0000 | -0.0068 | 0.4289 | 0.1082 | 0.1367 | 0.1643 | 0.1059 | 0.0723 | 0.1054 | 0.1173 | 0.0605 | 0.0310 | 0.0614 |
| Frequency | 0.5220 | -0.0068 | 1.0000 | 0.8372 | 0.7080 | 0.5133 | 0.6971 | 0.5855 | 0.5083 | 0.5481 | 0.6225 | 0.3950 | 0.2236 | 0.2801 |
| FirstPurch | 0.4523 | 0.4289 | 0.8372 | 1.0000 | 0.6397 | 0.4945 | 0.6683 | 0.5388 | 0.4735 | 0.5229 | 0.5968 | 0.3454 | 0.1800 | 0.2480 |
| ChildBks | 0.3896 | 0.1082 | 0.7080 | 0.6397 | 1.0000 | 0.2739 | 0.3949 | 0.3175 | 0.3211 | 0.3217 | 0.3868 | 0.2559 | 0.1762 | 0.1868 |
| YouthBks | 0.2887 | 0.1367 | 0.5133 | 0.4945 | 0.2739 | 1.0000 | 0.2853 | 0.2582 | 0.2395 | 0.2198 | 0.2561 | 0.1799 | 0.1068 | 0.1247 |
| CookBks | 0.3521 | 0.1643 | 0.6971 | 0.6683 | 0.3949 | 0.2853 | 1.0000 | 0.3559 | 0.2850 | 0.3314 | 0.3671 | 0.4430 | 0.1133 | 0.1879 |
| DoItYBks | 0.2960 | 0.1059 | 0.5855 | 0.5388 | 0.3175 | 0.2582 | 0.3559 | 1.0000 | 0.2243 | 0.3012 | 0.2797 | 0.2211 | 0.0771 | 0.1671 |
| RefBks | 0.2447 | 0.0723 | 0.5083 | 0.4735 | 0.3211 | 0.2395 | 0.2850 | 0.2243 | 1.0000 | 0.2037 | 0.2398 | 0.1558 | 0.4267 | 0.1168 |
| ArtBks | 0.2930 | 0.1054 | 0.5481 | 0.5229 | 0.3217 | 0.2198 | 0.3314 | 0.3012 | 0.2037 | 1.0000 | 0.2805 | 0.1853 | 0.0885 | 0.3948 |
| GeogBks | 0.3484 | 0.1173 | 0.6225 | 0.5968 | 0.3868 | 0.2561 | 0.3671 | 0.2797 | 0.2398 | 0.2805 | 1.0000 | 0.2370 | 0.1075 | 0.1231 |
| ItalCook | 0.2319 | 0.0605 | 0.3950 | 0.3454 | 0.2559 | 0.1799 | 0.4430 | 0.2211 | 0.1558 | 0.1853 | 0.2370 | 1.0000 | 0.2936 | 0.4454 |
| ItalHAtlas | 0.1020 | 0.0310 | 0.2236 | 0.1800 | 0.1762 | 0.1068 | 0.1133 | 0.0771 | 0.4267 | 0.0885 | 0.1075 | 0.2936 | 1.0000 | 0.3548 |
| ItalArt | 0.1428 | 0.0614 | 0.2801 | 0.2480 | 0.1868 | 0.1247 | 0.1879 | 0.1671 | 0.1168 | 0.3948 | 0.1231 | 0.4454 | 0.3548 | 1.0000 |

Florence will be used as our dependent variable. We are interested in predicting whether a customer will buy The Art History of Florence (Florence = 1) or not buy the book (Florence =0). Our independent variables will

be selected through running a logistic regression. We will then remove independent variables that are not significant from the model using the backward elimination method.
We will run a logistic regression and create a decision tree analysis on this data.

## Model Building:

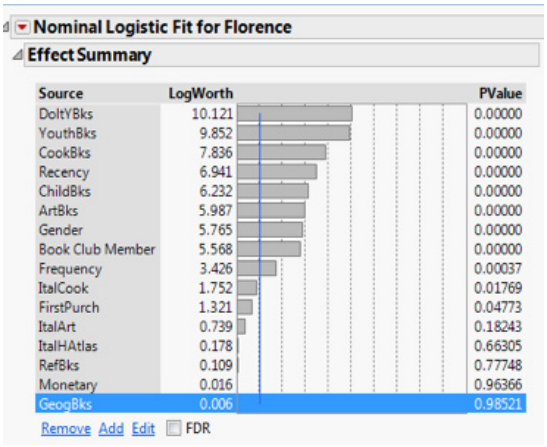
### Logistic Regression Analysis

Our initial hypothesis is that the betas of all the independent variables is equal to zero meaning that the factors do not have any effect on the probability of buying. Our alternative hypothesis is that at least one of the betas of the independent variables will not be zero.

| | |
|---|---|
| Null Hypothesis | Betai = 0 |
| Alternative Hypothesis | At least one Betai /= 0 |

### Steps involved in developing the LR analysis:

To begin with our logistic regression, we created a purely random validation column. We separated the data into training (70%) and validation (30%) datasets. We ran the logistic regression, using Florence as the dependent variable. We examined the p-values of the variables to determine the most significant values for the model:




### Model Significance:

After running the initial logistic regression, we identified the variables that were statistically insignificant and removed them one at a time, starting with the least significant variable, GeogBks (p value = .98521).
We were left with the following variables:

**Column Contributions**

| Term | Number of Splits | G^2 | Portion |
|---|---|---|---|
| Recency | 3 | 103.775651 | 0.2178 |
| YouthBks | 2 | 86.1608508 | 0.1808 |
| ArtBks | 1 | 67.827489 | 0.1424 |
| Book Club Member | 4 | 50.8793201 | 0.1068 |
| FirstPurch | 3 | 48.7563506 | 0.1023 |
| DoItYBks | 4 | 33.7686402 | 0.0709 |
| Gender | 3 | 24.830053 | 0.0521 |
| GeogBks | 1 | 15.2956247 | 0.0321 |
| CookBks | 1 | 14.0417524 | 0.0295 |
| ItalHAtlas | 1 | 11.096353 | 0.0233 |
| Monetary | 1 | 10.7971592 | 0.0227 |
| ChildBks | 1 | 9.24763879 | 0.0194 |
| Frequency | 0 | 0 | 0.0000 |
| RefBks | 0 | 0 | 0.0000 |
| ItalCook | 0 | 0 | 0.0000 |
| ItalArt | 0 | 0 | 0.0000 |

**Interpreting model significance:**

For the whole model test, the p-value was <.0001, which is less than alpha = .05. Therefore, we reject the null and can conclude that at least one beta is not equal to zero. This means that some of the variables in our model will affect whether or not someone will purchase the Florence book. The pseudo R^2 value is 0.1631, which implies that 16.31% of total uncertainty can be explained by this model. The Lack of Fit value is not significant at 0.1091, which implies that the most significant variables are included in this model.

**Odds Ratios:**

**Unit Odds Ratios**

Per unit change in regressor

| Term | Odds Ratio | Lower 95% | Upper 95% | Reciprocal |
|---|---|---|---|---|
| Recency | 0.931704 | 0.907088 | 0.956371 | 1.0733021 |
| Frequency | 1.315567 | 1.177094 | 1.469253 | 0.7601285 |
| FirstPurch | 1.018512 | 1.000537 | 1.037095 | 0.9818241 |
| ChildBks | 0.581802 | 0.476618 | 0.706349 | 1.7187989 |
| YouthBks | 0.325685 | 0.222548 | 0.463879 | 3.0704498 |
| CookBks | 0.513151 | 0.416498 | 0.627118 | 1.9487459 |
| DoltYBks | 0.408088 | 0.312012 | 0.526775 | 2.450451 |
| ArtBks | 1.971239 | 1.577158 | 2.471256 | 0.5072951 |
| ItalCook | 1.853198 | 1.315701 | 2.605273 | 0.5396077 |

Unit odds ratios show the increase in the odds of buying the Florence book as customers make one more purchase of a type of book or make one more purchase from the store. For example, the odds of buying the Florence book increases by 1.97 with every additional purchase of an Italian Art book, which was our highest units odds ratio. A person who makes one more purchase from the store (Frequency) has a 50-50 chance of buying the Florence book due to the odds ratio of 1.018512. This is not a good predictor of purchasing the Florence book as every random purchasing decision is a 50/50 chance. A person who buys one more Children's Book has odds of only .325685 of buying the Florence book, which means they have a higher probability of not buying this book.

**Range Odds Ratios**

Per change in regressor over entire range

| Term | Odds Ratio | Lower 95% | Upper 95% | Reciprocal |
|---|---|---|---|---|
| Recency | 0.090251 | 0.036315 | 0.219428 | 11.080201 |
| Frequency | 20.42879 | 6.01066 | 68.87534 | 0.0489505 |
| FirstPurch | 5.925583 | 1.053448 | 34.22956 | 0.1687598 |
| ChildBks | 0.038784 | 0.011723 | 0.124198 | 25.784009 |
| YouthBks | 0.003664 | 0.000546 | 0.021479 | 272.90406 |
| CookBks | 0.004808 | 0.000906 | 0.023922 | 207.98964 |
| DoItYBks | 0.011318 | 0.002957 | 0.040562 | 88.354792 |
| ArtBks | 29.76436 | 9.758354 | 92.16986 | 0.0335972 |
| ItalCook | 3.434344 | 1.731069 | 6.787445 | 0.2911764 |

Range odds ratios show how much the odds of purchasing the Florence book increases from the lowest value of the variable to the highest value. In this case, as the number of Art Books purchased increases from the minimum value of zero to the maximum value of five, the odds of buying the Florence book are 29.76436.

**Gender Odds:**

The odds of buying the Florence book is 1.86 for males, compared to only 0.537 for females, which means that males would have a higher likelihood of purchasing the Florence book than females. Among book club members, the odds of buying the Florence book are 3.1739, compared to only 0.315 for non-book club members. This indicates that the company should focus on who their book club members are and how they market to them. They should try to recruit more book club members, otherwise they may not sell many books because the percentage of customers that are book club members is actually quite small (13.2%), as shown in our visualization. This can also help to target marketing strategies because they know that people in book clubs are more likely to buy this type of book.

**Odds Ratios for Gender**

| Level1 | /Level2 | Odds Ratio | Prob>Chisq | Lower 95% | Upper 95% |
|---|---|---|---|---|---|
| Male | Female | 1.8606585 | <.0001* | 1.4426312 | 2.3953792 |
| Female | Male | 0.5374441 | <.0001* | 0.4174704 | 0.6931778 |

**Odds Ratios for Book Club Member**

| Level1 | /Level2 | Odds Ratio | Prob>Chisq | Lower 95% | Upper 95% |
|---|---|---|---|---|---|
| YES | NO | 3.1739028 | <.0001* | 1.9516665 | 5.1804139 |
| NO | YES | 0.3150695 | <.0001* | 0.1930348 | 0.5123826 |

**Confusion Matrix:**

The misclassification rate decreased between the training and validation model. This is a good sign meaning that our model becomes more accurate. The number of false positives, which is the number of people CRB thinks will buy the Florence book but actually do not, also decreases. This is good, because it will give the company a better prediction of response from customers as well as revenues earned from the sale of the Florence book.

**Confusion Matrix**

Training

| Actual Florence | Predicted Bought Florence | Predicted Did not buy Florence |
|---|---|---|
| Bought Florence | 47 | 302 |
| Did not buy Florence | 33 | 2402 |

Validation

| Actual Florence | Predicted Bought Florence | Predicted Did not buy Florence |
|---|---|---|
| Bought Florence | 16 | 108 |
| Did not buy Florence | 15 | 1077 |

The table below summarizes our interpretation of the LR confusion matrix:

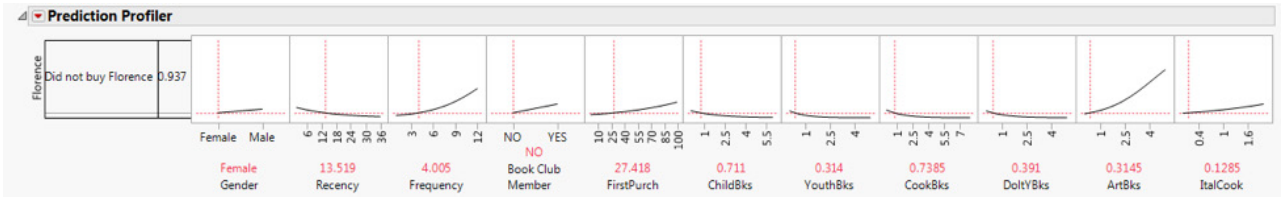
| | Training | Validation |
|---|---|---|
| Predictive Power: Accuracy | 87.97% | 89.88% |
| Misclassification | 12.03% | 10.12% |
| Sensitivity | TP / (TP+FN) = | Place |
| 47 / (47+302) = 13.47% | TP / (TP+FN) = | Promotion |
| 16 / (16+108) = 12.9% | Measurement (inspection) | People/Personnel |
| Specificity | TN / (TN+FP) = | Process |
| 2402 / (22+2402) = 99.09% | TN / (TN+FP) = | Physical Evidence |
| 1077 / (15+1077) = 98.6% | | |
| Number of False Positives | 22 | 15 |
| Number of False Negatives | 302 | 108 |

**Economic Costs:**

The net profit associated with the Linear Regression is depicted below. As seen in the chart below, the cost of having 15 false positives is $75. This may not seem like a large cost, but when analyzing compared to more customers, this cost can greatly increase. This unnecessary cost cuts into the profits of the company. The Book Cost is unavoidable and a cost of doing business. The higher this cost, the more books were sold, which is a positive outcome.

| | |
|---|---|
| Error Cost | $5 x 15 = $75 |
| Book Cost | $10 X (16+108) =$1,240 |
| **TOTAL COST:** | **$1315** |
| Gross Profit | $30 X (16+108) = $3,720 |
| **Net Profit (1,200 customers)** | $3,720 - $1,315 = **$2,405** |
| **Net Profit (500,000)** | **$1,001,750** |

**Profiler:**




By looking at the profiler, we can see by the steeper slope that purchasing Art Books has a very high effect on the probability of purchasing the Florence book. When moving the ArtBooks crosshairs to the right, the probability of buying the Florence book increases rapidly, while the probability for not buying it decreases. Frequency follows a similar pattern. When the crosshairs move from not being a book club member to being a book club member, the likelihood of buying the book also increases. We can see that the purchase of Children's, Youth, Cookbook, and DIY books have a negative slope. As the crosshairs in these categories move to the right, the likelihood of buying the book decreases. The profiler reaffirms the variables that are most predictive of purchasing the Florence book.

## Decision Tree Analysis:

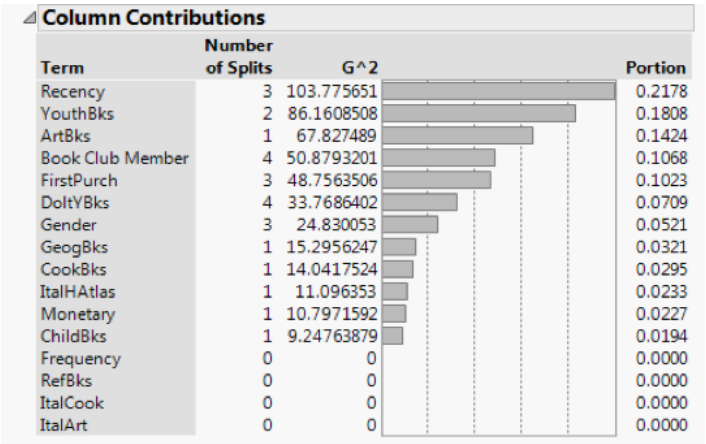
**Steps involved:**

We created a new validation column that put 70% into the training and 30% into the validation for the Decision Tree Analysis. We used stratified random with Florence as our dependent variable. Using Florence as our dependent variable, we added all independent variable to the decision tree. After hitting the "Go" button to find the maximum number of splits to make the most accurate model, the decision tree was split 25 times. The R^2 for the training set was 0.232. The R^2 for the validation set was 0.114. **See Appendix for the Decision Tree Analysis Fit Details**

The top variables in order of importance for understanding whether or not a customer will purchase the Florence book are as follows:

| Term | G^2 | Portion |
|---|---|---|
| Recency | 103.7757 | 0.2178 |
| YouthBks | 86.1609 | 0.1808 |
| ArtBks | 67.8275 | 0.1424 |
| Book Club Member | 50.8793 | 0.1068 |
| FirstPurch | 48.7564 | 0.1023 |

The column contributions shows that the most important variables in identifying whether or not someone will buy the Florence book are first Recency, YouthBks, ArtBks, then Book Club Members.



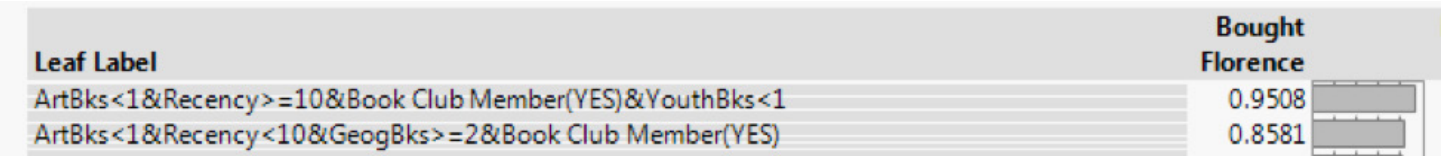
**Column Contributions**

| Term | Number of Splits | G^2 | Portion |
|---|---|---|---|
| Recency | 3 | 103.775651 | 0.2178 |
| YouthBks | 2 | 86.1608508 | 0.1808 |
| ArtBks | 1 | 67.827489 | 0.1424 |
| Book Club Member | 4 | 50.8793201 | 0.1068 |
| FirstPurch | 3 | 48.7563506 | 0.1023 |
| DoltYBks | 4 | 33.7686402 | 0.0709 |
| Gender | 3 | 24.830053 | 0.0521 |
| GeogBks | 1 | 15.2956247 | 0.0321 |
| CookBks | 1 | 14.0417524 | 0.0295 |
| ItalHAtlas | 1 | 11.096353 | 0.0233 |
| Monetary | 1 | 10.7971592 | 0.0227 |
| ChildBks | 1 | 9.24763879 | 0.0194 |
| Frequency | 0 | 0 | 0.0000 |
| RefBks | 0 | 0 | 0.0000 |
| ItalCook | 0 | 0 | 0.0000 |
| ItalArt | 0 | 0 | 0.0000 |

The most important class in this case is to identify the person who would buy the Florence book because by more accurately identifying who will purchase the books, CRB can market to specific customers more efficiently and effectively. This is also the important class because it will reduce costs of excess inventory and revenue prediction shortfalls. False positives of assuming someone will buy the book and then actually not buying the book is the most costly because of these reasons.

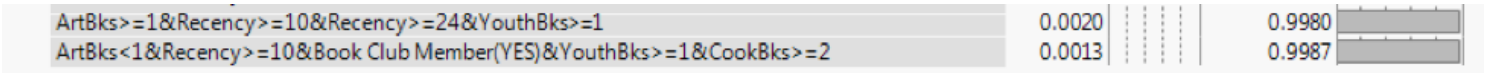
**Interpretation of Rules:**

Top two rules for Bought Florence:



| Leaf Label | Bought Florence |
|---|---|
| ArtBks<1&Recency>=10&Book Club Member(YES)&YouthBks<1 | 0.9508 |
| ArtBks<1&Recency<10&GeogBks>=2&Book Club Member(YES) | 0.8581 |

- If ArtBooks less than one and Recency is greater than or equal to 10 and the person is a book club member and Youth books less than one, the high probability of buying the Florence book is .9508.
- If ArtBooks less than one and Recency less than 10 and GeogBooks is greater than or equal to 2 and the person is a book club member, the high probability of buying the Florence book is .8581.




**Top two rules for Did Not Buy Florence:**

- If ArtBooks greater than or equal to 1 and Regency is greater than or equal to 10 and Recency is greater than or equal to 24 and YouthBooks is greater than or equal to one, the probability of not buying the Florence book is .9980.
- If ArtBooks is less than 1 and Recency is greater than or equal to 10 and the person is a book club member and YouthBooks is greater than or equal to 1 and Cookbooks is greater than or equal 2, the probability of not buying the Florence book is .9977.

**Interpret Confusion Matrices:** All of the models have similar total accuracy but, as seen in the table below, the model that we found to be the best is the Probability Cutoff of 0.8. We found that this model lowered the amount of false positives, which is key to lowering unnecessary costs and maximizing net profits. Each false positive, in which the company predicts that a customer will buy the Florence book, but actually does not, incurs a cost of $5. For every book the company sells, whether they predicted it as a true positive or it was a surprise as a false negative, the company will net $20 off of each book ($30 selling price minus $10 cost). Using the .8 cutoff model, we estimate net profits of $2,805, about $400 more than the worst model. The model also has the highest accuracy in detecting the customers that will not buy the book, which will further help to market and sell more efficiently. The chart further details accuracy, sensitivity, specificity, total profits among the 1,200 validation sample, and extrapolated to the 500,000 customers in their database. **See the Appendix for the JMP outputs and matrices for all three cutoffs.**

**The following chart is a comparison between all three Decision Tree cutoff models:**

| Validation | Prob Cutoff (0.2) | Prob Cutoff (0.5) | Prob Cutoff (0.8) |
|---|---|---|---|
| Total Accuracy | 86.89% | 88.25% | 88.33% |
| Accuracy for Florence = 1 (Bought) Sensitivity | 32.39% | 7.7% | 6.34% |
| Accuracy for Florence = 0 (Not Bought) Specificity | 91.11% | 99% | 99.34% |
| Total Net Profits (1,200 customers) - Validation | Cost : (94 * $5) = **$470**<br>Revenue : ($30 - $10 * (46+96)) = $2,840<br>**Net Profit** = 2,840 - 470 = **$2,370** | Cost : (10 * $5) = **$50**<br>Revenue : ($30 - $10 * (11+131)) = $2,840<br>**Net Profit** = 2,840 - 50 = **$2,790** | Cost : (7 * $5) = **$35**<br>Revenue : ($30 - $10*(9+133)) = $2,840<br>**Net Profit** = 2,840 - 50 = **$2,805** |
| Total Net Profits (50,000 customers) | Cost : (5 * $39,150) = **$195,750**<br>Revenue : $20*(19,150 + 40,000) = $1,183,000<br>**Net Profit** = $1,183,000 - $195,750 = **$987,250** | Cost : (5 * $4,167) = **$20,833**<br>Revenue : $20*(4583+54583) = $1,183,327<br>**Net Profit** = $1,183,327 - $20,833 = **$1,162,494** | Cost : (5 * $2,197) = **$14,583**<br>Revenue : $20*(3750+55417) = $1,183,333<br>**Net Profit** = $1,183,333 - $14,583 = **$1,168,750** |
| False Positive | 94 | 10 | 7 |
| False Negative | 96 | 131 | 133 |

## Model Comparison:

When we compare the results of the Logistic Regression and the three Decision Tree models, we see that the Decision Tree - .8 Probability Cutoff model results in stronger outcomes:

| Validation Set | LR | DT Cutoff = .2 | DT Cutoff = .5 | DT Cutoff = .8 |
|---|---|---|---|---|
| Accuracy | 89.88% | 84.17% | 88.25% | 88.33% |
| Misclassification | 10.12% | 15.83% | 11.75% | 11.67% |
| Sensitivity | 12.9% | 32.39% | 7.7% | 6.34% |
| Specificity | 98.6% | 91.11% | 99% | 99.34% |
| True positives | 16 | 46 | 11 | 9 |
| False positives | 15 | 96 | 131 | 7 |
| False negatives | 108 | 94 | 10 | 133 |
| Total net profits - 1,200 Validation customers | $2,405 | $2,370 | $2,790 | $2,850 |
| Total net profits - 500,000 total customers | **$1,001,750** | **$987,250** | **$ 1,162,494** | **$1,168,750** |

While the Logistic Regression is slightly more accurate and had a lower misclassification rate, it was not as effective in reducing false positives as the .8 DT Cutoff model, which we believe is one of the most important requirements of the model. This key difference is reflected in the much higher net profit of the .8 DT Cutoff model.

The .8 Cutoff model also resulted in the most False Negatives. While this classification does result in revenue for the company (the model predicts a customer will not purchase the book but actually does), methods to ultimately reduce this number and increase true positives may be warranted in order to more effectively predict revenues, but the model will help CRB anticipate it.

When comparing the four potential models, we were troubled by the low sensitivity rates. We realize that our validation column splits the data 70% to 30%, which resulted in a much smaller validation sample. If we were to change our validation column to a 50/50 split, we predict that sensitivity percentages would have been higher.

## Communication:

When communicating the results of our analysis to CRB, we would first review the primary business and data mining objectives of: 1) developing a model to help predict who would purchase the Florence book, while

reducing the costs associated with false positives and maximize revenues, and 2) developing a profile of customers who would buy the book.

We turn to our **Model Comparison chart**, which is an effective tool to quantify the business value of each model and clearly communicate key outcomes, followed by an explanation of what they mean to the business and our overall objectives. We would explain that we identified one particular model (the DT .8 Cutoff Model) as the best at achieving our goals and objectives defined at the beginning of the process.

We would also communicate that while the model was the most superior at eliminating costs of False Positives and maximizing profits, making further adjustments to reduce False Negatives, increase True Positives, and reduce misclassification could be done, if the company is willing to commit further resources. However, we believe that the model will help them realize significant, immediate gains.

We would also present the key profile characteristics of who would and would not purchase the book based upon the analyses. We would stress that while there were differences in accuracy-related percentages across the models, the customer profile remained constant:

| Customers most likely to purchase Florence | Customers most likely to NOT purchase Florence |
|---|---|
| People who purchased art books | People who purchased children's books |
| People who purchased the Italian cook book | People who purchased youth books |
| Frequent buyers | People who purchase infrequently |
| Loyal, long-time customers | People who have not purchased in over two years |
| Recently acquired customers within last two years | |
| Males | |
| To some degree, Book Club members | |

Lastly, we would present our recommendations for marketing tactics, which are detailed in the Optimization – Recommendations section below.

## Operationalization – Recommendations:

We recommend first segmenting customers by using the most predictive variables from our analyses to market the Florence book, including:

- People who have bought the Italian cookbook
- People who have purchased art books
- People who have made many purchases from the store over time (Frequency)

- People who have been buying from the store for several years (First Purchase)
- People who have purchased from the store recently
- Males

The company is concerned about alienating its book club customers. We do not feel that offering these higher margin books will discourage these customers from buying from the store for several reasons: 1) While book club members purchased art books and Italian cooking books less often than other categories, they did buy them at greater levels than non-book club members, 2) Book club members represent a small percentage of total customers (13.2%), and 3) Just under half of the book club customers also purchased the Florence book. Therefore, it would not be inappropriate to include some marketing of the Florence book to these customers as well, in addition to their book club purchases.

**Marketing tactics:**

We recommend that the bookstore implement the following steps:

1. Develop targeted emails, mailings, mobile content and special offers or loyalty programs to cus-tomers in the identified customer segments, particularly frequent and loyal customers.
2. Include a free, lower-margin book related to the Florence book, such as travel writing or a non-fiction book, to frequent customers to further promote it.
3. Encourage people to join the book club, as book club members tend purchase all book catego-ries at higher levels.
4. Purchase digital advertising on blogs, websites related to Italian cooking and art, and through search engines when people search for terms like Italian cooking, art, and other related terms.

These programs can be implemented one by one in order to determine the effectiveness of each one, but these ideas give them a list of options that they can pilot to effectively reach current and new customers.

**Conclusion:**

All four of the models we analyzed will give the Charles River Bookstore a strong idea of who will likely purchase the Florence book. Furthermore, we were able to identify the model that will accurately pinpoint who will purchase while minimizing errors, maximizing profits, and mapping out where they should devote the greatest resources to effectively market it. It is important to remember that this model is particular to the Florence book only. If the Charles River Bookstore intends to sell other coffee table books of other topics, such as nature or history, the model may not apply. The analysis demonstrates the importance of walking through the entire Business Analytics Lifecycle to ensure that the root business problem is addressed and that all possible solutions are evaluated.

# Appendix

## Summary Statistics

**Summary Statistics**

17 Columns | Clear Select | Distribution

| Columns | N Categories | Min | Max | Mean | Std Dev |
|---|---|---|---|---|---|
| Gender | 2 | . | . | . | . |
| Monetary | . | 15 | 477 | 206.789 | 101.32426229947 |
| Recency | . | 2 | 36 | 13.519 | 8.2386654199558 |
| Frequency | . | 1 | 12 | 4.0055 | 3.5467752471313 |
| Book Club Member | 2 | . | . | . | . |
| FirstPurch | . | 2 | 99 | 27.4175 | 18.738329034195 |
| ChildBks | . | 0 | 6 | 0.711 | 1.0386282718913 |
| YouthBks | . | 0 | 5 | 0.314 | 0.6144089636339 |
| CookBks | . | 0 | 8 | 0.7385 | 1.0790777785744 |
| DoItYBks | . | 0 | 5 | 0.391 | 0.7233600849184 |
| RefBks | . | 0 | 4 | 0.2705 | 0.5748150599004 |
| ArtBks | . | 0 | 5 | 0.3145 | 0.6314185793557 |
| GeogBks | . | 0 | 5 | 0.4115 | 0.774156043311 |
| ItalCook | . | 0 | 2 | 0.1285 | 0.3768597296262 |
| ItalHAtlas | . | 0 | 2 | 0.0395 | 0.2072449942023 |
| ItalArt | . | 0 | 2 | 0.052 | 0.2372974452441 |
| Florence | 2 | . | . | . | . |

Decision Tree Analysis Fit Details

**Fit Details**

| Measure | Training | Validation | Definition |
|---|---|---|---|
| Entropy RSquare | 0.2316 | 0.1143 | 1-Loglike(model)/Loglike(0) |
| Generalized RSquare | 0.2999 | 0.1544 | (1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n)) |
| Mean -Log p | 0.2792 | 0.3220 | $\sum$ -Log(p[j])/n |
| RMSE | 0.2849 | 0.3086 | $\sqrt{\sum(y[j]-p[j])^2/n}$ |
| Mean Abs Dev | 0.1634 | 0.1780 | $\sum \lvert y[j]-p[j] \rvert /n$ |
| Misclassification Rate | 0.1000 | 0.1175 | $\sum$ (p[j]≠pMax)/n |
| N | 2800 | 1200 | n |

**Confusion Matrix**

Training

| Actual Florence | Predicted Bought Florence | Predicted Did not buy Florence |
|---|---|---|
| Bought Florence | 59 | 272 |
| Did not buy Florence | 8 | 2461 |

Validation

| Actual Florence | Predicted Bought Florence | Predicted Did not buy Florence |
|---|---|---|
| Bought Florence | 11 | 131 |
| Did not buy Florence | 10 | 1048 |

| | RSquare | N | Number of Splits |
|---|---|---|---|
| Training | 0.232 | 2800 | 25 |
| Validation | 0.114 | 1200 | |

Decision Tree Confusion Matrices

**Confusion Matrix**

Training

| Actual Florence | Predicted Bought Florence | Predicted Did not buy Florence |
|---|---|---|
| Bought Florence | 59 | 272 |
| Did not buy Florence | 8 | 2461 |

Validation

| Actual Florence | Predicted Bought Florence | Predicted Did not buy Florence |
|---|---|---|
| Bought Florence | 11 | 131 |
| Did not buy Florence | 10 | 1048 |

| Cut-off = 0.5 | Training | Validation |
|---|---|---|
| Predictive Power: Accuracy | 90 % | 88.25% |
| Misclassification | 10 % | 11.75% |
| Sensitivity | TP / (TP+FN) = 17.8 % | 7.7% |
| Specificity | TN / (TN+FP) = 99.6 % | 99% |
| Number of False Positives | 8 | 10 |
| Number of False Negatives | 272 | 131 |

Alternative Cut-offs:

**Confusion Matrix for Cut-off = 0.2**

**Target: Florence**
**Predictor: Partition**

| | Modeling Data | | | |
|---|---|---|---|---|
| | Training | | Validation | |
| | Predicted | | Predicted | |
| Florence | Bought Florence | Did not buy Florence | Bought Florence | Did not buy Florence |
| Bought Florence | 142 | 189 | 46 | 96 |
| Did not buy Florence | 178 | 2291 | 94 | 964 |

**Confusion Rates** **Confusion Rates**

| Cut-Off = 0.2 | Training | Validation |
|---|---|---|
| Predictive Power: Accuracy | 1-13.1017% = 86.89% | 1-15.83% = 84.17% |
| Misclassification | (178+189)/(178+189+142+2291) = 13.107% | (94+96) / (94+96+46+964) = 15.83% |
| Sensitivity | TP / (TP+FN) = 142 / (142+189) = 42.9% | TP / (TP+FN) = 46 / (46+96) = 32.39% |
| Specificity | TN / (TN+FP) = 2291 / (2291+178) = 92.79% | TN / (TN+FP) = 964 / (964+94) = 91.11% |
| Number of False Positives | 178 | 94 |
| Number of False Negatives | 189 | 96 |

**Confusion Matrix for Cut-off = 0.8**

**Target: Florence**
**Predictor: Partition**

| | Modeling Data | | | |
|---|---|---|---|---|
| | Training | | Validation | |
| | Predicted | | Predicted | |
| Florence | Bought Florence | Did not buy Florence | Bought Florence | Did not buy Florence |
| Bought Florence | 51 | 280 | 9 | 133 |
| Did not buy Florence | 5 | 2464 | 7 | 1051 |

| Cut-Off = 0.8 | Training | Validation |
|---|---|---|
| Predictive Power: Accuracy | 90% | 88.33% |
| Misclassification | 10% | 11.67% |
| Sensitivity | TP / (TP+FN) = 15.41% | 6.34% |
| Specificity | TN / (TN+FP) = 99.80% | 99.34% |
| Number of False Positives | 5 | 7 |
| Number of False Negatives | 280 | 133 |

**Confusion Matrix**

Training

| | Predicted | |
|---|---|---|
| Actual Florence | Bought Florence | Did not buy Florence |
| Bought Florence | 59 | 272 |
| Did not buy Florence | 8 | 2461 |

Validation

| | Predicted | |
|---|---|---|
| Actual Florence | Bought Florence | Did not buy Florence |
| Bought Florence | 11 | 131 |
| Did not buy Florence | 10 | 1048 |

NOTE: All charts used for this analysis, including those posted in this report, can be found in the JMP file.


**Headquarters**

4625 Alexander Drive Suite
#245,Alpharetta, GA-30022, USA
+1-678-862-0550

**Sales office**

2033 Gateway Place,
Suite 605, San Jose,
CA 95110
+1-678-862-0550

**Development Centre**

Thinkspace Building, Plot no 82/11, Patrika
Nagar, HITEC City,
Hyderabad, Telangana 500081
040-29884477

**E-mail**

info@vuesol.com

To see how we empower your business
visit **www.vuesol.com**

/vuesol @vuesol /company/vuesol