RCBC Case Study Data Analysis

Johnny Weng

Due to the decreasing profits of the Reader's Choice Book Club (RCBC), they believed that data intelligence would allow them to understand their customer better and segment a target market for upcoming book offerings. To incorporate data intelligence in their upcoming book offering, *"*The Art History of Florence*"*, a market test was conducted involving a random sample of 4000 customers to create an analysis to calibrate response models. Based on these response models, a score will be created for each customer in relation to the likelihood that they would buy the purchase. Through the use of data analysis, we will be able to identify a segmented target market to advertise *"*The Art History of Florence*"*.

Using the variables included in the market test on a training sample of 2400 customers, we conducted a logistic regression to identify a logit, odds, and probability that they would purchase *"*The Art History of Florence*"*. After our initial logistic regression, we identified that the variables to include Monetary, Recency, Months since first purchase, Youth books, *Secrets of Italian Cooking*, *Historical Atlas of Italy*, and *Italian Art* were not statistically significant at alpha =.10. We used backwards selection in our regression starting with the least significant variable to come up with our final model, which is shown in Table 1. Using this model, we could sufficiently use our training data and score values from the data set with the following model:

Logit (*Buyer = Yes*) = -2.591 (*Constant)* + .269 (*Frequency*) - .386 (*ChildBks*) - .507 (*CookBks*) - .225 (*DoItYBks*) - .260 (*RefBks*) + .436 (*ArtBks*) + .196 (*GeogBks*).

After creating our model in which we would base our odds and probability from, we wanted to identify if there was multicollinearity between our variables. This would allow us to see if any of the variables were redundant with one another, making it unnecessary in scoring our data. After running a multicollinearity test (seen in Table 2), we identified that the *Frequency* variable was a candidate for multicollinearity with a VIF of 5.604. We could address this issue of multi-collinearity by removing the variable, however we will move forward as it doesn't necessarily affect the fit of the model or create bad predictions (although it could make picking the right predictors more difficult).

Table 1

Variables in Reduced Logistic Regression Model

Variable	B	S.E.	Wald	df	Sig.	Exp (B)
Gender	-.532	.150	12.502	1	.000	.588
Frequency	.269	.043	39.487	1	.000	1.308
ChildBks	-.386	.101	14.754	1	.000	.680
CookBks	-.507	.101	25.336	1	.000	.602
DoItYBks	-.225	.120	3.513	1	.061	.799
RefBks	-.260	.139	3.486	1	.062	.771
ArtBks	.436	.101	18.610	1	.000	1.546
GeogBks	.196	.091	4.657	1	.031	1.217
Constant	-2.591	.149	300.388	1	.000	.075

Note. n = 2400 (Training Data).

Table 2

Collinearity Statistics For Reduced Logistic Regression Model

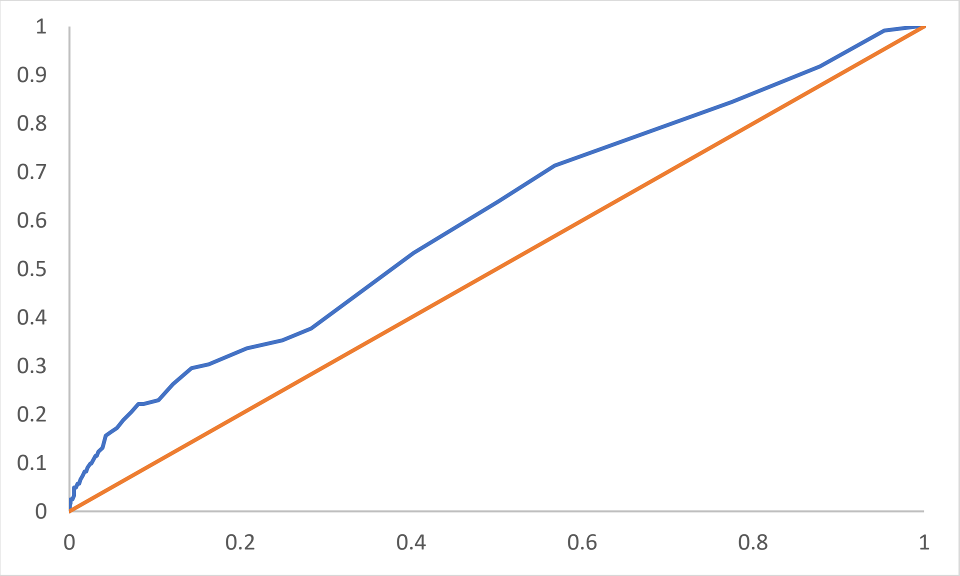
Variable	Tolerance	VIF
Gender	.996	1.004
Frequency	.178	5.604
ChildBks	.493	2.028
CookBks	.452	2.211
DoItYBks	.645	1.551
RefBks	.733	1.365
ArtBks	.956	1.047
GeogBks	.843	1.187

Note. Multicollinearity present with "F" (Frequency) variable. n = 2400 (Training Data).

In using our reduced model with each of the independent variables, it's clear to see that females and customers who have purchased child books, cookbooks, do it yourself books, and reference books reduce the probability of purchasing "The Art History of Florence". This is due to the coefficient of the variable being negative. Conversely, those who have purchased art books, geography books, and have made frequent purchases, increase the probability of purchasing "The Art History of Florence" because the value of the coefficient is positive. Holding all else constant, a one unit change in the negative variables would make a customer less likely to purchase *"*The Art History of Florence*"*, whereas a one unit change in the positive variables would make a customer more likely to purchase *"*The Art History of Florence*"*.

Once we identified our reduced model, we scored the values in both our training and validation sets. From the ROC Curve in our Validation dataset, we see the plot between the sensitivity, or True Positive Rate and the 1-specificty, or False Positive Rate. Looking at our ROC curve, we can tell that our data had a not so accurate test, resulting in many more False Positive results than True Positive. At a 0.20 cutoff level, we can infer that there were a large percentage of non-buyers who were classified as buyers. Comparing our classification matrices between our Validation and Training data, we can conclude that the Training model was much more accurate in defining both True Positives and False Positives. Since the dataset had a larger amount of observations, we see that the percentage of false identifications were actually lower in the Training than in the Validation dataset.

Chart 1
Validation Data ROC Curve



Classification Confusion Matrix Validation

		Predicted Class		
		1	0	Total
Actual	1	21	101	122
Class	0	82	1396	1478

Classification Confusion Matrix Training

		Predicted Class		
		1	0	Total
Actual	1	42	174	216
Class	0	92	2092	2184

Using the scoring data for our validation set, we sorted the probability of customers buying the “Art History of Florence” in descending order from highest probability to lowest probability. From the first 160 observations, we can see that 28 customers bought the book, or 17.5%. From the whole validation data set, we had 122 buyers out of 1600, which turns out to be 7.625%. The percentage of total buyers from the first 160 observations is higher than the average of the total data set, which makes sense due to the higher predicted probabilities. From the data below in Table 3, we can look at the number of buyers per every 160 observations.

Table 3

Buyers of Art History of Florence per every 160 observations

Observation	Buyers	Percentage
1-160	28	17.50%
161-320	12	7.50%
321-480	6	3.75%
481-640	17	10.63%
641-800	12	7.50%
801-960	13	8.13%
961-1120	7	4.38%
1121-1280	10	6.25%
1281-1440	10	6.25%
1441-1600	7	4.38%

Note. n = 1600 Observations are in increments of 160 up to 1600.

Based on the reduced regression model, we scored the following customers based on the logit, odds, and probability that they would respond to the offer in table 4.

Table 4

Logit, Odds, and Probability That Select Customers Will Respond to "The Art History of Florence" offer

ID#	Gender	Frequency	Child Bks	Cook Bks	DoIt YBks	Ref Bks	Art Bks	Geog Bks	Logit	Odds	Probability
A	0	2	0	3	0	0	0	0	-3.576	.028	.027
B	1	2	0	0	0	0	0	0	-2.585	.075	.070
C	1	1	8	0	0	0	0	0	-5.944	.003	.003
D	1	3	0	0	2	3	0	1	-3.351	.035	.034
E	0	2	0	0	0	0	3	0	-.747	.474	.322
F	1	3	0	1	0	0	0	0	-2.824	.059	.056

Note. Logit, Odds, and Probability based on Reduced Logistic Regression Model.

Next our task was to project our findings onto a larger population of 20,000 customers to forecast. To conduct this forecast based on the data that we had available, we first found the probability for each consumer to purchase "The Art History of Florence". After converting all our odds to probabilities, we reasoned that customers who are most likely to purchase the book "The Art History of Florence" are the ones who have a probability of 50% or higher. Coding all of those customers as a '1' if they did have a probability of purchase of 50% or higher and then all others as '0' we found that only 13 customers out of our 4000 sample would be projected to purchase the book. Forecasting this information to a population of 20,000, we estimate that 65 customers would purchase "The Art History of Florence".

After reducing our model to the significant independent variables, we found that gender was indeed significant in affecting which customer was more or less likely to purchase "The Art History of Florence". Within our data, the variable Gender was coded as '1' if the customer identified as a female, and a '0' if male. Since males are coded as '0' it means that they are the reference point in the model. Utilizing this data, we found our model ran a coefficient of -0.522 on Gender, which means that the log odds for women make them 0.522 times less likely to purchase “The Art History of Florence”, holding all else equal. This result also shows that men are the more likely purchasers of the book.

To determine the top quintiles in the categories Recency, Frequency and Monetary, a cutoff value for the top quintile in each category was determined and every observation outside of the cutoff value was eliminated. For Recency, the observations were sorted from smallest to largest to reflect the assumption that more recent a purchase, the more likely the consumer was to make another purchase. The cut off value for Recency was six months and while the top 800 reflected the top quintile in the 4,000 observations sample size, there were several ties at the cutoff value of six months which lead to a total of 852 observations sorted. The ties were not addressed at this point because of the assumption that after sorting for the top quintiles in the other categories, many of the observed ties would be eliminated. Frequency and Monetary were sorted from largest to smallest under the assumption that the more frequent a consumer is and the more money they have spent makes them a better candidate to target for purchases of “The Art History of Florence”. After sorting based on these conditions, the top 800 observations were used to determine cutoff values for each category, these values were seven for Frequency and $303 for Monetary. These cut off values were then used to eliminate observations that didn’t meet the criteria in the original 852 observations that were sorted for the category Recency. Using this method, the final list of observations was sorted to 96 out of the sample size of 4,000 and out of the 96 sorted observations there were 17 actual purchases which translated to a purchase rate of 17.7%. Based on this information, it is arguable that the method is sound in its ability to filter through the 4,000 total observations which had a purchase rate of 8.45% and optimize a target list which has a higher success rate and would also minimize targeting costs.

After running a logistic regression for purchases of "The Art History of Florence" (see Table 5), using only the variables Recency, Frequency and Monetary, it was determined that Monetary was insignificant in predicting a purchase and Recency and Frequency were the only significant predictors. Interpreting the coefficients tells us that Recency has a negative influence on predicting a purchase. Since Recency has an inverse relationship (the higher the number, the less frequent) it shows that the more recent a purchase, the more likely it is for a consumer to make the desired purchase. Frequency has a positive coefficient, indicating that consumers who more frequently made purchases from the "RCBC" were more likely to make the desired purchase. To verify if any multicollinearity exists, VIF scores were determined (see Table 6) for the variables Recency and Frequency which remained in the final model and elimination due to insignificance. The VIF scores for both variables were 1.000, indicating that there is no issue of multicollinearity.

To determine the relationship that exists between each individual independent variable remaining in the final logistic regression model and probability of purchase, scatter plots were created for each of the two variables (see Chart 1 & 2). In each chart, one variable was held constant at the median of the original 4,000 observations while determining changes in probability solely on the values of the other independent variable not being held at a constant. Based on the charts, it can be indicated that lower values for recency correlate with higher probabilities of purchase, this appears to be reasonable considering extended periods without a purchase could mean a consumer no longer has demand for the products offered by the RCBC. Examining the graph for probabilities based on frequency values, it can be indicated that higher values of frequency correlate with higher probabilities of purchase. This is evident with the highest probability recorded corresponding with the highest value of frequency and an inverse relationship existing for the lowest values in each category. This observation seems sound considering the more frequent a consumer makes purchases from the RCBC creates more opportunities to be introduced to other products such as "The Art History of Florence" and has a better chance of leading to purchase.

Table 5
Logistic Regression Model for Variables R, F, & M

Variable	B	S.E.	Wald	df	Sig.	Exp(B)
Recency	-.029	.008	14.229	1	.000	.971
Frequency	.075	.015	24.957	1	.000	1.078
Constant	-2.329	.129	327.547	1	.000	.097

Note. Based Off Recency, Frequency, and Monetary Variables. n = 4000.

Table 6
Collinearity Statistics for Reduced Logistic Regression Model

Variable	Tolerance	VIF
Recency	1.000	1.000
Frequency	1.000	1.000

Note. Based Off Recency, Frequency, and Monetary Variables. n = 4000.

Chart 2

Probability of Purchase based on Recency

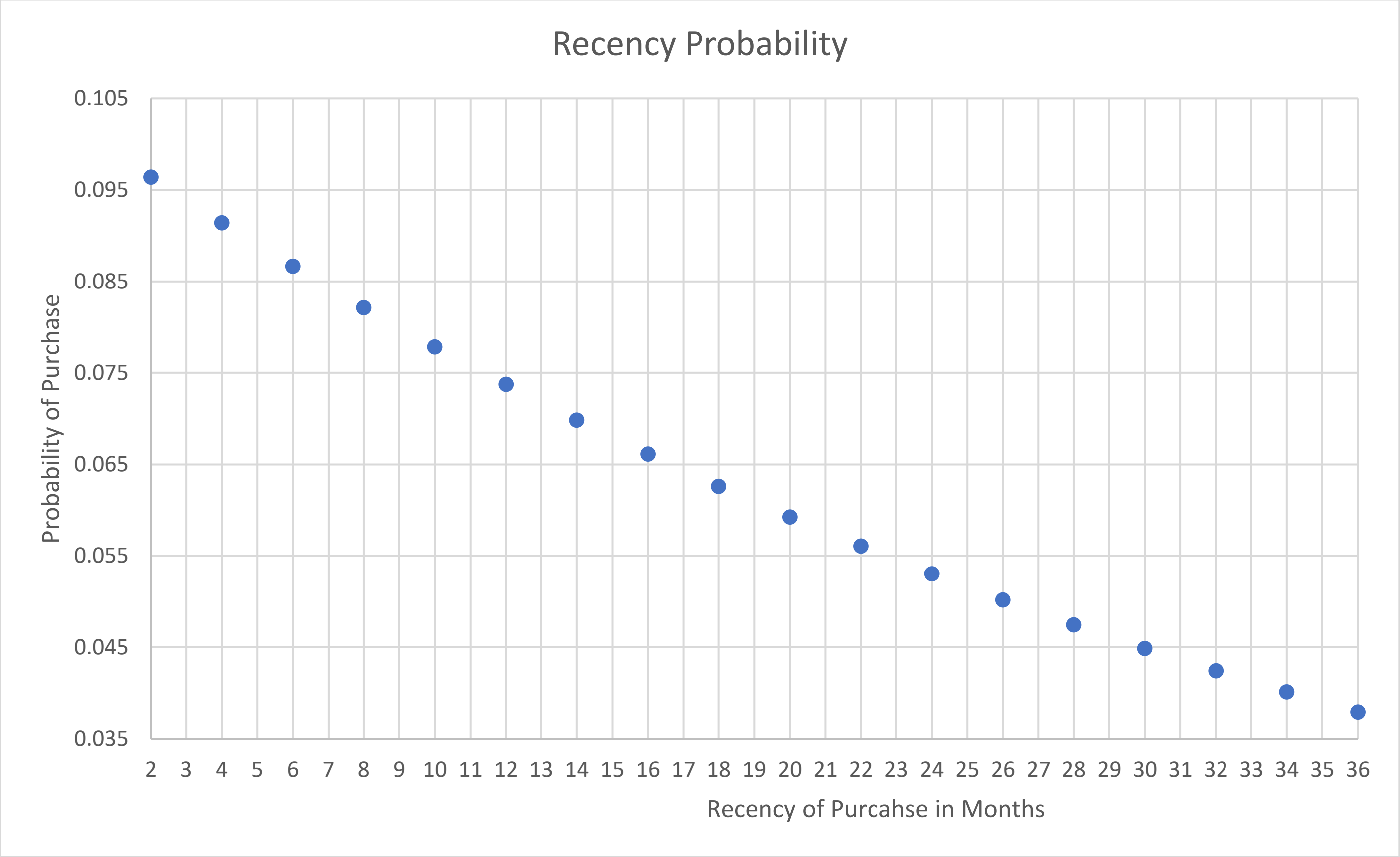
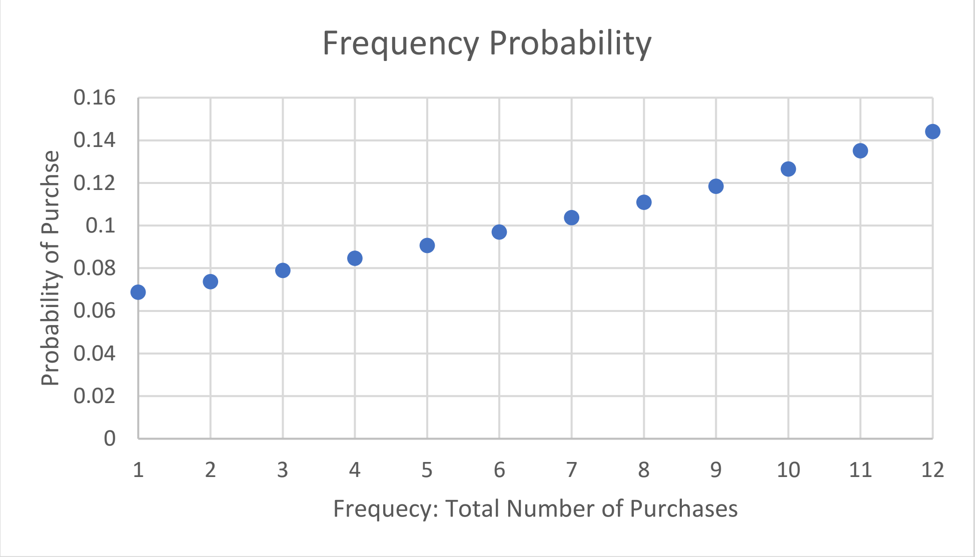


Chart 3

Probability of Purchase based on Frequency


Frequency Probability
Probability of Purchse
Frequecy: Total Number of Purchases