Case Study

Jilly Hebert

11/15/2020

## Charles Book Club

The Charles Book Club (CBC) distributes books to their customer base after receiving their surveys in the mail. Even though the club was receiving increased mailing orders and diversifying their book selection, they were still loosing profits. CBC decided to revisit their original plan of using database marketing to improve mailing yields and to stay profitable.

CBC wants to start targeting their customers more appropriately, so to get to know them better, CBC management decided to focus on the most profitable customers and design targeted marketing strategies to best reach them.

Currently, CBC uses two processes. The first is “Customer Acquisition” which targets new members through magazine, newspaper, and television advertisements. Direct mailing and telemarketing would contact existing club members so they can be notified of new book arrivals before the general advertisements were released. The second process was “Data Collection.” This recorded all customer responses and collected any critical information. CBC workers would reach out to members if there was any missing information.

When new book titles are being released, CBC has a two-step approach. First, they conduct a test of 4,000 randomly sampled customers from the database to analyze their responses. This would create and calibrate CBC’s response models for the new book title. Second, they would compute a score for each customer based on the response model and use that score and a cutoff value to extract a target customer list for direct mail promotions.

## Variable Description

4,000 observations and 24 variables.

### Nominal Variables

* Sequence Number: Sequence number in the partition
* ID Number: Identification number
* Recency (R): Months since last book purchase
* Frequency (F): Total number of book purchases
* Monetary (M): Total money spent on books
* 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)
* GeoBks: Number of purchases from the category geography books
* ItalCook: Number of purchases of the book title *Secrets of Italian Cooking*
* ItalAtlas: Number of purchases of the book title *Historical Atlas of Italy*
* ItalArt: Number of purchases of the book title *Italian Art*

### Ordinal Variables

* Rcode: Recency breakdown
  + Rcode = 1: 0 - 2 months
  + Rcode = 2: 3 - 6 months
  + Rcode = 3: 7 - 12 months
  + Rcode = 4: 13+ months
* Fcode: Frequency breakdown
  + Fcode = 1: 1 book
  + Fcode = 2: 2 books
  + Fcode = 3: 3+ books
* Mcode: Monetary breakdown
  + Mcode = 1: $0 - $25
  + Mcode = 2: $26 - $50
  + Mcode = 3: $51 - $100
  + Mcode = 4: $101 - $200
  + Mcode = 5: $201+
* ArtBks: Number of purchases from the category art books
* Related Purchase: Number of related books purchased

### Binary Variables

* Gender
  + Gender = 0: Male
  + Gender = 1: Female
* Florence
  + Florence = 0: Did not buy *The Art History of Florence*
  + Florence = 1: Bought *The Art History of Florence*
* Yes\_Florence
  + Yes\_Florence = 0: Did not buy *The Art History of Florence*
  + Yes\_Florence = 1: Bought *The Art History of Florence*
* No\_Florence
  + No\_Florence = 0: Bought *The Art History of Florence*
  + No\_Florence = 1: Did not buy *The Art History of Florence*

## Analysis

A new book title, *The Art History of Florence*, is about to be released and a sample of 4,000 customers has been randomly selected. The customer base was split into a training (60%) data and validation (40%) data so the models could be created using the training data while the validation data was used to evaluate performance. There are 2401 customers in the training data set and 1599 customers in the validation data set. *k*-nearest neighbors and logistic regression models will be used to select targeted customers. Special attention will be paid to the RFM (recency, frequency, monetary) segmentation since that is standard industry practice.

### Response Rates

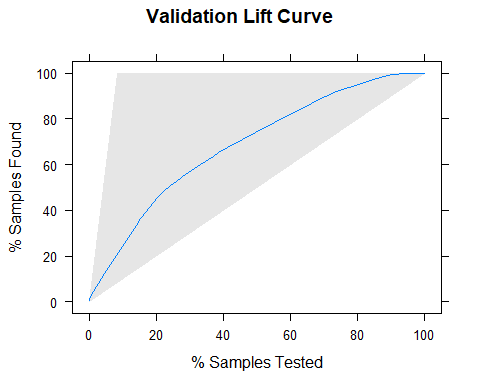
First, the response rates for the 4,000 customer sample were identified and the RFM combinations with higher response rates were identified.

The overall response rate for the training data set was 0.0845. There were 49 RFM combinations. Of those, there were 23 combinations that had higher response rates than the overall response rate.

The overall response rate for those with “above average” RFM combinations in the validation set was 0.1051.

### Segmentation

Segments were created using the response rates to identify the combinations that have higher rates. Those with response rates that exceeded twice the overall training response rate were placed in segment 1. Those with response rates that were greater than the overall training response rate but less than twice the overall training response rate were placed in segment 2. The remaining combinations were placed in segment 3.



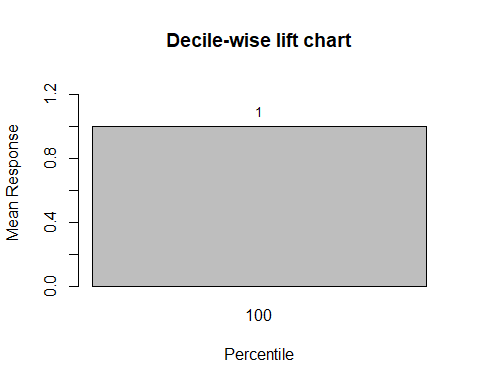
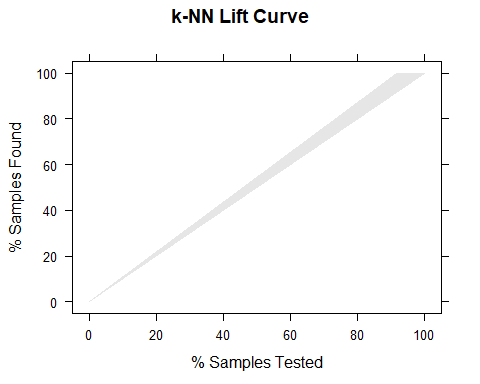
The validation lift curve shows the number of customers by the cumulative number of buyers in the validation data set. The gap between the shaded area and the blue line identifies the gains that the company sees from using the segmentation over randomly sampling.

### k-Nearest Neighbors Modeling

To more accurately identify customers that would purchase *The Art History of Florence*, two *k*-Nearest Neighbors models will be built. The first model will use a categorical response variable while the second model will use a numeric response variable. These models will be used to identify if targeting customers using the models is better than random sampling.

#### Categorical Response

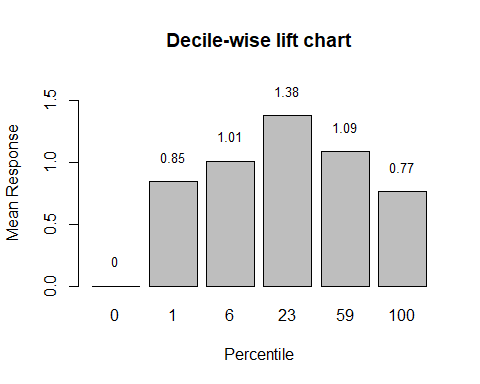
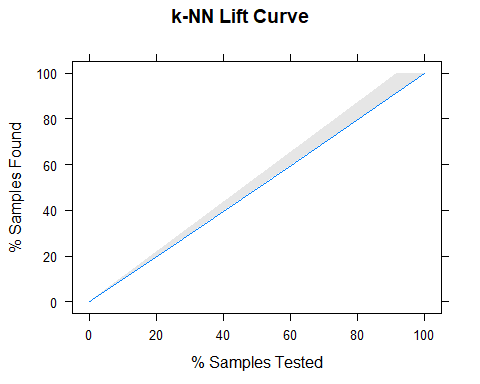
The first *k*-Nearest Neighbors model will use a categorical response variable.



The k value in the *k*-NN model indicates the count of the nearest neighbors. It is used to compute distances between test points and trained label points. Using the k value of 11 results in the highest accuracy of 91.56%. Because there are no predictions for purchasing the book, the lift curve is empty and the Decile-wise lift chart shows the same outcome as random sampling would produce. Since the first and only bar is equal to 1, it shows that using the *k*-NN model for targeted sampling is performing the same as random sampling. The lift curve shows that there would be no difference between using the *k*-NN model for targeted sampling and random sampling.

#### Numeric Response

The second *k*-Nearest Neighbors model will use a numeric response variable.



When Florence is treated as a numeric value, the model accuracy was 91.56%. The maximum predicted probability was 0.4545 and the minimum predicted probability was 0, giving a range of 0.45.

For the Decile-wise chart, since the bar probabilities are not in descending order, it can be determined that using the *k*-NN model for targeted sampling is not doing better than random sampling.

The lift curve further confirms this because it shows that there is no gap between the blue line and the shaded area, meaning that the *k*-NN model is not leading to more gains when using the *k*-NN model over randomly sampling.

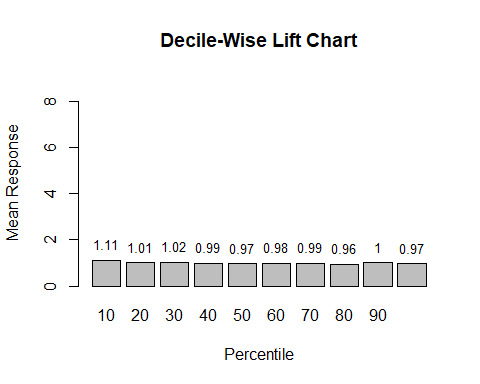
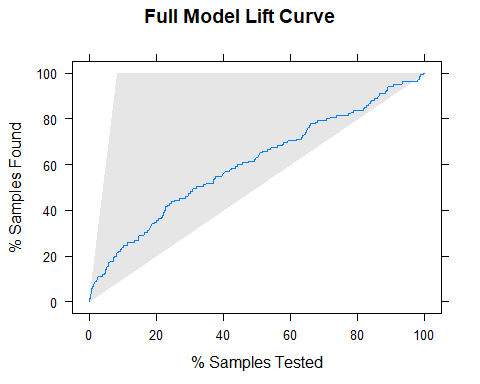
This model had the same accuracy than the *k*-NN model that treated the response as categorical, but did slightly better than random sampling, as seen by the Decile-wise lift chart.

### Logistic Regression Modeling

To more accurately identify customers that would purchase *The Art History of Florence*, three logistic regression models will be built. The first model will use all sixteen variables, the second model with use the best determined variables, and the last model will use only RFM variables. These models will be used to identify if targeting customers using these models are better than random sampling.

#### Full Model

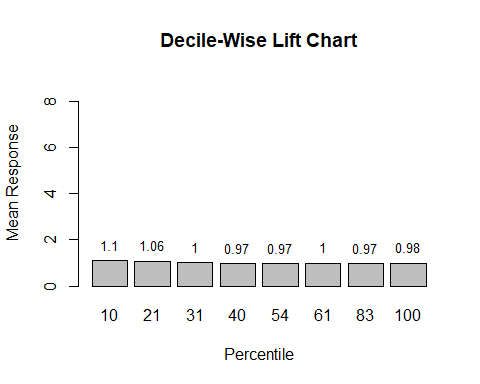
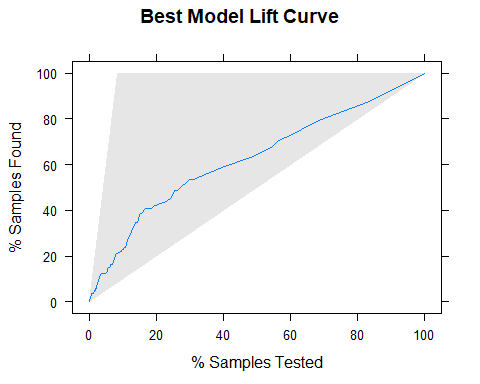
The full model will include all 16 variables.



When using the full model, the top 10% of customers who bought the book would lead to 1.11 times as many purchases than would selecting 10% of the customers at random. The full model has an accuracy of 91.56%. The gap between the shaded area and the blue line identifies the gains that the company sees from using the model over randomly sampling.

#### Best Predictors Model

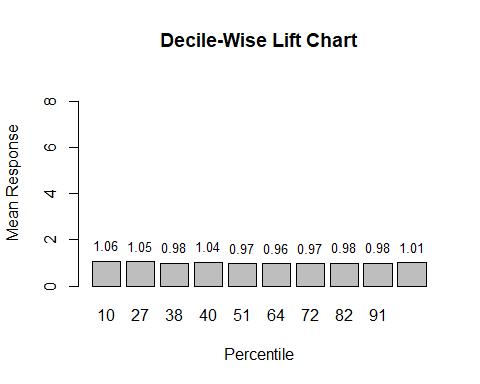
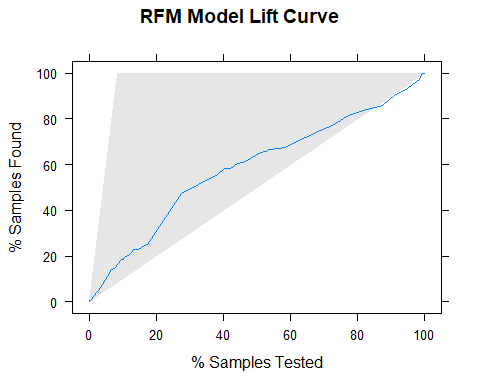
The best predictors model will be determined by using a backward stepping algorithm on the full model.



When using backwards stepping to reduce the full model to only the best predictors, the variables used in the best model were Gender, ArtBks, GeogBks, and Rcode. The best predictors model, the top 10% of customers who bought the book would lead to 1.10 times as many purchases than would selecting 10% of the customers at random. The best predictors model has an accuracy of 91.56%. The gap between the shaded area and the blue line identifies the gains that the company sees from using the model over randomly sampling.

#### RFM Variables Model

The RFM model will use the ordinal variables of recency, monetary, and frequency.



When using the RFM model, the top 10% of customers who bought the book would lead to 1.06 times as many purchases than would selecting 10% of the customers at random. The RFM model has an accuracy of 91.56%. The gap between the shaded area and the blue line identifies the gains that the company sees from using the model over randomly sampling.

### Model Comparisons

The Logistic Regression models all perform similarly, but the *k*-Nearest Neighbors models seemed to perform similarly to random sampling. The model that did the best job was the full logistic regression model with the best predictors logistic regression model as a close second. The models only performed slightly better than random sampling, so the more important issue to address is the low response rates. Because the response rates were so low, it would be suggested to increase the sample of customers for more accurate model predictions until more customers are purchasing new books.

### Mail Campaign

If CBC wants a 30% likelihood of purchase, then there are 4 people out of the 49 people in the validation data set that would be targeted. Of those 4 people, 3 people purchased *The Art History of Florence*.