ADS-505 Applied Data Science for Business

Module 1 Assignment 1.1: Use Case - Charles Book Club (Using Python) (80 points)

CharlesBookClub.csv is the dataset for this case study (linked in the assignment prompt in Blackboard).

For this assignment, you will use a Jupyter notebook to answer the questions in the "Assignment" section of this document. Be sure to include the number and question that you are answering in your notebook.

THE BOOK INDUSTRY

Approximately 50,000 new titles, including new editions, are published each year in the United States, giving rise to a $25 billion industry in 2001. In terms of percentage of sales, this industry may be segmented as follows:

16%	Textbooks
16%	Trade books sold in bookstores
21%	Technical, scientific, and professional books
10%	Book clubs and other mail-order books
17%	Mass-market paperbound books
20%	All other books

Book retailing in the United States in the 1970s was characterized by the growth of bookstore chains located in shopping malls. The 1980s saw increased purchases in bookstores stimulated through the widespread practice of discounting. By the 1990s, the superstore concept of book retailing gained acceptance and contributed to double-digit growth of the book industry. Conveniently situated near large shopping centers, superstores maintain large inventories of 30,000–80,000 titles and employ well-informed sales personnel. Book retailing changed fundamentally with the arrival of Amazon, which started out as an online bookseller and, as of 2015, was the world's largest online retailer of any kind. Amazon's margins were small and the convenience factor high, putting intense competitive pressure on all other book retailers. Borders, one of the two major superstore chains, discontinued operations in 2011.

Subscription-based book clubs offer an alternative model that has persisted, though it too has suffered from the dominance of Amazon.

Historically, book clubs offered their readers different types of membership programs. Two common membership programs are the continuity and negative option programs, which are both extended contractual relationships between the club and its members. Under a *continuity program*, a reader signs up by accepting an offer of several books for just a few dollars (plus shipping and handling) and an agreement to receive a shipment of one or two books each month thereafter at more-standard pricing. The continuity program is most common in the children's book market, where parents are willing to delegate the rights to the book club to make a selection, and much of the club's prestige depends on the quality of its selections.

In a *negative option program*, readers get to select how many and which additional books they would like to receive. However, the club's selection of the month is delivered to them automatically unless they specifically mark "no" on their order form by a deadline date. Negative option programs sometimes result in customer dissatisfaction and always give rise to significant mailing and processing costs.

In an attempt to combat these trends, some book clubs have begun to offer books on a *positive option basis*, but only to specific segments of their customer base that are likely to be receptive to specific offers. Rather than expanding the volume and coverage of mailings, some book clubs are beginning to use database-marketing techniques to target customers more accurately. Information contained in their databases is used to identify who is most likely to be interested in a specific offer. This information enables clubs to design special programs carefully tailored to meet their customer segments' varying needs.

DATABASE MARKETING AT CHARLES

The Club

The Charles Book Club (CBC) was established in December 1986 on the premise that a book club could differentiate itself through a deep understanding of its customer base and by delivering uniquely tailored offerings. CBC focused on selling specialty books by direct marketing through a variety of channels, including media advertising (TV, magazines, and newspapers) and mailing. CBC is strictly a distributor and does not publish any of the books that it sells. In line with its commitment to understanding its customer base, CBC built and maintained a detailed database about its club members. Upon enrollment, readers were required to fill out an insert and mail it to CBC. Through this process, CBC created an active database of 500,000 readers; most were acquired through advertising in specialty magazines.

The Problem

CBC sent mailings to its club members each month containing the latest offerings. On the surface, CBC appeared very successful: mailing volume was increasing, book selection was diversifying and growing, and their customer database was increasing. However, their bottom-line profits were falling. The decreasing profits led CBC to revisit their original plan of using database marketing to improve mailing yields and to stay profitable.

A Possible Solution

CBC embraced the idea of deriving intelligence from their data to allow them to know their customers better and enable multiple targeted campaigns where each target audience would receive appropriate mailings. CBC's management decided to focus its efforts on the most profitable customers and prospects, and to design targeted marketing strategies to best reach them. The two processes they had in place were:

1. Customer acquisition:
    - New members would be acquired by advertising in specialty magazines, newspapers, and on TV.
    - Direct mailing and telemarketing would contact existing club members.
    - Every new book would be offered to club members before general advertising.
2. Data collection:
    - All customer responses would be recorded and maintained in the database.
    - Any information not being collected that is critical would be requested from the customer.

For each new title, they decided to use a two-step approach:

1. Conduct a market test involving a random sample of 4000 customers from the database to enable analysis of customer responses. The analysis would create and calibrate response models for the current book offering.
2. Based on the response models, compute a score for each customer in the database. Use this score and a cutoff value to extract a target customer list for direct-mail promotion.

Targeting promotions was considered to be of prime importance. Other opportunities to create successful marketing campaigns based on customer behavior data (returns, inactivity, complaints, compliments, etc.) would be addressed by CBC at a later stage.

Art History of Florence

A new title, *The Art History of Florence*, is ready for release. CBC sent a test mailing to a random sample of 4000 customers from its customer base. The customer responses have been collated with past purchase data. The dataset was randomly partitioned into three parts: *Training Data* (1800 customers): initial data to be used to fit models, *Validation Data* (1400 customers): holdout data used to compare the performance of different models, and *Test Data* (800 customers): data to be used only after a final model has been selected to estimate the probable performance of the model when it is deployed. Each row (or case) in the spreadsheet (other than the header) corresponds to one market test customer. Each column is a variable, with the header row giving the name of the variable. The variable names and descriptions are given in Table 21.1.

DATA MINING TECHNIQUES

Various data mining techniques can be used to mine the data collected from the market test. No one technique is universally better than another. The particular context and the particular characteristics of the data are the major factors in determining which techniques perform better in an application. For this assignment, we focus on two fundamental techniques: *k*-nearest neighbors (*k*-NN) and logistic regression. We compare them with each other as well as with a standard industry practice known as *RFM (recency, frequency, monetary) segmentation*.

Table 21.1 List of Variables in Charles Book Club Dataset

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

(*Source*: Reproduced with permission of Jacob Zahavi)

RFM Segmentation

The segmentation process in database marketing aims to partition customers in a list of prospects into homogeneous groups (segments) that are similar with respect to buying behavior. The homogeneity criterion we need for segmentation is the propensity to purchase the offering. However, since we cannot measure this attribute, we use variables that are plausible indicators of this propensity.

In the direct marketing business, the most commonly used variables are the *RFM variables*:

R = *recency*, time since last purchase

F = *frequency*, number of previous purchases from the company over a period

M = *monetary*, amount of money spent on the company's products over a period

The assumption is that 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 1800 observations in the dataset were divided into recency, frequency, and monetary categories as follows:

Recency:
0–2 months (Rcode = 1)
3–6 months (Rcode = 2)
7–12 months (Rcode = 3)
13 months and up (Rcode = 4)

Frequency:
1 book (Fcode = l)
2 books (Fcode = 2)
3 books and up (Fcode = 3)

Monetary:
$0–$25 (Mcode = 1)
$26–$50 (Mcode = 2)
$51–$100 (Mcode = 3)
$101–$200 (Mcode = 4)
$201 and up (Mcode = 5)

ASSIGNMENT

1. Partition the data into training (60%) and validation (40%). Use seed = 1.

   1.1. What is the response rate for the training data customers taken as a whole? What is the response rate for each of the 4 × 5 × 3 = 60 combinations of RFM categories? Which combinations have response rates in the training data that are above the overall response in the training data? **(10 points)**

   1.2. Suppose that we decide to send promotional mail only to the "above-average" RFM combinations identified in part 1. Compute the response rate in the validation data using these combinations. **(10 points)**

2. The *k*-NN technique can be used to create segments based on product proximity to similar products of the products offered as well as the propensity to purchase (as measured by the RFM variables). For *The Art History of Florence*, a possible segmentation by product proximity could be created using the following variables:

   - *R*: recency—months since last purchase
   - *F*: frequency—total number of past purchases
   - *M*: monetary—total money (in dollars) spent on books
   - *FirstPurch*: months since first purchase

- *RelatedPurch*: total number of past purchases of related books (i.e., sum of purchases from the art and geography categories and of titles *Secrets of Italian Cooking*, *Historical Atlas of Italy*, and *Italian Art*)

2.1 Use the *k*-NN approach with **uniform** weights to classify cases with *k* = 1, 2, …, 11, using Florence as the outcome variable. Based on the validation set, find the best *k*. Remember to normalize all five variables. Create a lift curve for the best *k* model, and report the expected lift for an equal number of customers from the validation dataset. **(12 points)**
Hint: You can check sklearn.neighbors.KneighborsClassifier to assign different weights function.

2.2. The *k*-NN prediction algorithm gives a numerical value, which is a **weighted average** of the values of the Florence variable for the *k*-NN with **weights** that are inversely proportional to **distance**. Using the best *k* that you calculated above with *k*-NN classification, now run a model with *k*-NN prediction and compute a lift curve for the validation data. Use all 5 predictors and normalized data. What is the range within which a prediction will fall? How does this result compare to the output you get with the *k*-NN classification? **(10 points)**
Hint: You can check sklearn.neighbors.KneighborsClassifier to assign different weights function.

3. The logistic regression model offers a powerful method for modeling response because it yields well-defined purchase probabilities. The model is especially attractive in consumer-choice settings because it can be derived from the random utility theory of consumer behavior.
Use the training set data of 2400 records to construct two logistic regression models with Florence as the outcome variable and each of the following sets of predictors: **(10 points)**
   - The full set of **16 predictors** in the dataset
   - A subset of predictors that you judge to be the best

3.2. If the cutoff criterion for a campaign is a 30% likelihood of a purchase, find the customers in the validation data that would be targeted and count the number of buyers in this set. **(10 points)**

3.3. Briefly explain, in two to three paragraphs, the business objective, the data mining models used, why they were used, the model results, and your recommendations to your non-technical stakeholder team. **(18 points)**