Data Mining Group Projects

Spring 2016

Due Date: May 10, 2016

1. Tayko Software Cataloger
2. Segmenting Consumers of Bath Soap
3. Direct-mail Fundraising
4. Placement of Fordham Graduates
5. Charles Book Club

Descriptions of problems and datasets:

1. Tayko Software Cataloger

Background

Tayko is a software catalog firm that sells games and educational software. It started out as a software manufacturer and later added 3rd-party titles to its offerings. It has recently put together a revised collection of items in a new catalog, which it is preparing to roll out in a mailing.

In addition to its own software titles, Tayko's customer list is a key asset. In an attempt to expand its customer base, it has recently joined a consortium of catalog firms that specialize in computer and software products. The consortium affords members the opportunity to mail catalogs to names drawn from a pooled list of customers. Members supply their own customer lists to the pool, and can "withdraw" an equivalent number of names each quarter. Members are allowed to do data mining on the records in the pool so they can do a better job of selecting names from the pool.

The Mailing Experiment

Tayko has supplied its customer list of 200,000 names to the pool, which totals over 5,000,000 names, so it is now entitled to draw 200,000 names for a mailing. Tayko would like to select the names that have the best chance of performing well, so it conducts a test – it draws 20,000 names from the pool and does a test mailing of the new catalog.

This mailing yielded 1065 purchasers, a response rate of 0.053. Average spending was $103 for each of the purchasers, or $5.46 per catalog mailed. To optimize the performance of the data mining techniques, it was decided to work with a stratified sample that contained equal members of purchasers and non-purchasers. For ease of presentation, the dataset for this case includes just 1000 purchasers and 1000 non-purchasers, an apparent response rate of 0.5. Therefore, after using the dataset to predict who will be a purchaser, we must adjust the purchase rate back down by multiplying each case's "probability of purchase" by 0.053/0.5, or 0.107.

Data (Dataset: Tayko.xls, available at http://storm.cis.fordham.edu/~yli/data/Tayko.xls).

There are two response variables in this case. Purchase indicates whether or not a prospect responded to the test mailing and purchased something. Spending indicates, for those who made a purchase, how much they spent. The data has been partitioned into 2 sets (training and validation) randomly on the basis of the partition variable. You may like to do the partition by yourself. Validation data could be used to evaluate the performance of the models.

The overall procedure in this case will be to develop two models. One will be used to classify records as *purchase* or *no purchase*. The second will be used for cases that are classified as purchase and spending certain amount of money.

Codelist				
Var. #	Variable Name	Description	Variable Type	Code Description
1.	US	Is it a US address?	binary	1: yes 0: no
2 - 16	Source_*	Source catalog for the record (15 possible sources)	binary	1: yes 0: no
17.	Freq.	Number of transactions in last year at source catalog	numeric	
18.	last_update_days_ago	How many days ago was last update to cust. record	numeric	
19.	1st_update_days_ago	How many days ago was 1st update to cust. record	numeric	
20.	Web_order	Customer placed at least 1 order via web	binary	1: yes 0: no
21.	Gender=mal	Customer is male	binary	1: yes 0: no
22.	Address_is_res	Address is a residence	binary	1: yes 0: no
23.	Purchase	Person made purchase in test mailing	binary	1: yes 0: no
24.	Spending	Amount spent by customer in test mailing ($)	numeric	
25.	Partition	Variable indicating which partition the record will be assigned to	alpha	p1: training p2: validation

2. Consumers of Bath Soap

Background

CRISA is an Asian market research agency that specializes in tracking consumer purchase behavior in consumer goods (both durable and nondurable). In one major research project, CRISA tracks about 30 product categories (e.g., detergents), and within each category, about 60 to 70 brands. To track purchase behavior, CRISA constituted about 50,000 household panels in 105 cities and towns in India, covering about 80% of the Indian urban market. (In addition to this, there are 25, 000 sample households selected in rural areas, but we are working only with urban market data). The households are carefully selected using stratified sampling. The strata are defined on the basis of socioeconomic status and the market (a collection of cities).

CRISA has both transaction data (each row is a transaction) and household data (each row is a household), and for the household data, maintains the following information:

- Demographics of the households (updated annually).
- Possession of durable goods (car, washing machine, etc., updated annually; an "affluence index" is computed from this information)
- Purchase data of product categories and brands (updated monthly).

CRISA has two categories of clients: (1) advertising agencies that subscribe to the database services, obtain updated data every month, and use the data to advise their clients on advertising and promotion strategies; (2) and consumer goods manufacturers, which monitor their market share using the CRISA database.

Key Problems

CRISA has traditionally segmented markets on the basis of purchaser demographics. They would now like to segment the market based on two key sets of variables more directly related to the purchase process and to brand loyalty:

1. Purchase behavior (volume, frequency, susceptibility to discounts, and brand loyalty)
2. Basis of purchase (price, selling proposition)

Doing so would allow CRISA to gain information about what demographic attributes are associated with different purchase behaviors and degrees of brand loyalty, and thus deploy promotion budgets more effectively. More effective market segmentation would enable CRISA's clients to design more cost-effective promotions targeted at appropriate segments. Thus, multiple promotions could be launched, each targeted at different market segments at different times of the year. This would result in a more cost-effective allocation of the promotion budget to different market segments. It would also enable company to design more effective customer reward systems and thereby increase brand loyalty.

Data (Dataset: BathSoap.xls, available at http://storm.cis.fordham.edu/~yli/data/BathSoap.xls)

Measuring Brand Loyalty: Several variables in this case measure aspects of brand loyalty. The number of different brands purchased by the customer is one measure. However, a consumer who purchases one or two brands in quick succession, then settles on a third for a long streak, is

different from a consumer who constantly switches back and forth among three brands. How often customers switch from one brand to another is another measure of loyalty. Yet a third perspective on the same issue is the proportion of purchases that go to different brands – a consumer who spends 90% of his or her purchase money on one brand is more loyal than a consumer who spends more equally among several brands. All three of these components can be measured with the data in the purchase summary worksheet.

We could use clustering techniques to identify clusters of households based on:

- The variables that describe purchase behavior (including brand loyalty).
- The variables that describe the basis for purchase.
- The variables that describe both purchase behaviors and basis of purchase.

For k-means clustering method, the selection of K is important. Think about how the clusters would be used. It is likely that the marketing efforts would support two to five different promotional approaches.

How would the percentages of total purchases comprised by various brands be treated? Isn't a customer who buys all brand A just as loyal as a customer who buys all brand B? What will be the effect on any distance measure of using the brand share variables as is? Consider using a single derived variable.
After the customers are clustered, select what think is the best segmentation and comment on the characteristics (demographic, brand loyalty, and basis for purchase) of these clusters. This information would be used to guide the development of advertising and promotional campaigns.

Demographic Data Codelist	
MEM	Member ID
SEC	Socio economic class (1 = high, 4 = low)
FEH	**Food Eating Habits :** 1. Pure Vegetarian 2. Veg. But Serve Eggs 3. Non Vegetarian 0. Not Specified
MT	**Native Language (mother tongue) :** (0: not specified)
SEX	**Sex of homemaker:** 1. Male 2. Female
GE	**Age of homemaker** 1. Up to 24 2. 25-34 3. 35-44 4. 45+
EDU	**Education of homemaker** 1. Illiterate 2. Literate, but no formal schooling 3. Up to 4 years of school 4. 5-9 years of school 5. 10-12 years of school 6. Some college 7. College graduate 8. Some graduate school 9. Graduate or professional school degree 0. Not specified
HS	**Household size:** Number of people in the household
CHILD	**Presence of children in household** 1. Children up to age 6 present (only) 2. Children 7-14 present (only) 3. Both 4. None 5. Not specified

CS	**Television** 1. Cable or broadcast TV available 2. Unavailable
Affluence Index	Calculated from Durables sheet.
Purchase Summary Data Codelist	
Labels	**What they stand for**
No. Brands	*Number of brands purchased*
Brand Runs	*Number of runs (streaks) of purchasing the same brand*
Total volume	*Volume of product purchased (grams)*
No. of trans.	*Number of transactions*
Value	*Value in paise (100 paise = 1 rupee)*
Avg. Price	*Avg. price (rupees per 100 gram cake); computed from total volume and value*
Purch. Vol. no promo	*Percent of volume purchased not on promotion*
Purch Vol. promo 6	*Percent of volume purchased on promo code 6*
Purch. Vol other promo	*Percent of volume purchased on promo code other than 6*
Brand Codelist	(see spreedsheet)
Price Codelist	*1. ANY PREMIUM SOAPS 2. ANY POPULAR SOAP* *3. ANY ECONOMY/CARBOLIC 4. ANY SUB-POPULAR*
Promotion Codelist	1 *Price off* 2 *Exchange Offer* 3 *Coupons* 4 *Extra grammage* 5 *Value added Pack* 6 *Banded Offer* 7 *Free gift* 8 *Others*
Proposition Codelist	5 *ANY BEAUTY* *6 ANY HEALTH* *7 ANY HERBAL* *8 ANY FRESHNESS* *9 ANY HAIR* *10 ANY SKIN CARE* *11 ANY FAIRNESS* *12 ANY BABY* *13 ANY GLYCERINE* *14 ANY CARBOLIC* *15 ANY OTHERS*
Brand wise purchase	
Br. Cd. xxx	Percent of volume purchased of the brand xxx.
Price Category-wise purchase	
Price Cat 1 to 4	Percent of volume purchased under the price category
Selling Proposition-wise purchase	
Proposition Cat 5 – 15	Percent of volume purchased under the product proposition category.

3. Direct-mail Fundraising

Background

A national veterans' organization wishes to develop a data mining model to improve the cost-effectiveness of their direct marketing campaign. The organization, with its in-house database of over 13 million donors, is one of the largest direct-mail fundraisers in the United States. According to their recent mailing records, the overall response rate is 5.1%. Out of those who responded (donated), the average donation is $13.00. Each mailing, which includes a gift of personalized address labels and assortments of cards and envelopes, costs $0.68 to produce and send. Using these facts, we take a sample of this dataset to develop a classification model that can effectively capture donors so that the expected net profit is maximized. Weighted sampling is used, under-representing the non-responders so that the sample has equal numbers of donors and non-donors.

Data (Datasets: Fundraising.xls, FutureFundraising.xls, available at
http://storm.cis.fordham.edu/~yli/data/Fundraising.xls &
http://storm.cis.fordham.edu/~yli/data/FutureFundraising.xls)

The file Fundraising.xls contains 3120 data points with 50% donors (TARGET_B=1) and 50% non-donors (TARGET_B=0). The amount of donation (TARGET_D) is also included. The data in FutureFundraising.xls should be used as test data.

Codelist	
ZIP	Zip code group (zip codes were grouped into 5 groups). A 1 indicates that the potential donor belongs to this zip group. A 0 indicates that the potential donor does not belong to this zip group. 00000-19999 => 1 (not shown in the data set) 20000-39999 => zipconvert_2 40000-59999 => zipconvert_3 60000-79999 => zipconvert_4 80000-99999 => zipconvert_5
HOMEOWNER	1 = homeowner, 0 = not a homeowner
NUMCHLD	Number of children
INCOME	Household income
GENDER	Gender: 0 = Male 1 = Female
WEALTH	Wealth rating uses median family income and population statistics from each area to index relative wealth within each state. The segments are denoted 0-9, with 9 being the highest wealth group and zero being the lowest. Each rating has a different meaning within each state.
HV	Average Home Value in potential donor's neighborhood in hundreds of dollars.
ICmed	Median Family Income in potential donor's neighborhood in hundreds of dollars.

ICavg	Average Family Income in potential donor's neighborhood in hundreds of dollars.
IC15	Percent earning less than $15K in potential donor's neighborhood.
NUMPROM	Lifetime number of promotions received to date.
RAMNTALL	Dollar amount of lifetime gifts to date.
MAXRAMNT	Dollar amount of largest gift to date.
LASTGIFT	Dollar amount of most recent gift.
TOTALMONTHS	Number of months from last donation to July 1998 (the last time the case was updated).
TIMELAG	Number of months between first and second gift.
AVGGIFT	Average dollar amount of gifts to date.
TARGET_B	Target variable: binary indicator for response. 1 = Donor, 0 = Non-donor
TARGET_D	Target Variable: Donation Amount (in $).

4. Placement of Fordham Graduates

Background

To provide better services to Fordham students, the office of Career Service at Fordham (OCS) would like to investigate factors affecting the placement of Fordham graduates. Two categories of data are collected. The first part of data is from CareerInsights, which captures students' placement post-graduation, such as permanent job offers or internships. The second part of data is from CareerLink, which captures students activities related to a variety of career services offered by OCS, such as consulting/training, interview workshops, job fairs, etc. Regarding privacy issue, sensitive personal information was covered in both data sets.

Current data of students' placement from CareerInsights is of class 201X. Students of class 201X were active with OCS during a four-year period before/including 201X. Activities they participated in career service events were not completely collected by CareerLink due to the limitations of hardware and software during that period.

Data (Dataset: Class201X_outcome.csv, Class201X_StudentActivity.csv, available at http://storm.cis.fordham.edu/~yli/data/Class201X_outcome.csv & http://storm.cis.fordham.edu/~yli/data/Class201X_StudentActivity.csv)

Data mining techniques such as association rule, clustering and classification could be applied and tested on the given data sets. Questions such as "what kind of career services are most helpful for students to find a permanent job?", "Whether career service activities play an important role in students' job hunting process?", "Whether OCS is an important source for job searching?" should be studied.

The file ***Class201X_outcome.csv*** contains information about 1936 Fordham students. Each of them is identified with an ID number. The total instances are 2732. Some students reported their job placements more than once (permanent job and internship).

Codelist	
ID	1 – 1936
GraduationTerm	
UnderMajor	
UnderMinor	
UnderCollege	
UnderDegree	
ReportedPost	Status about post-graduation job *Multiple Job Intentions, Not Seeking, Seeking Employment (accepted offer), Seeking Employment (no received offer), Unreported*
ReportedIntern	Status about internship *Multiple Job Intentions, Not Seeking, Seeking Employment (accepted offer), Seeking Employment (no received offer), Unreported*
BeginLooking	Date to start looking for a job/intern
JobType	INTERN, POSTG

OfferType	JOB, INTERN, SEEKING_EDU (Still Seeking Education), SEEKING_JOB (Still Seeking Employment), CONTINUE_EDU(Continuing education), OTHER
OfferStatus	True – offer accepted, False
SalaryCurrency	USD, OTHER
Salary	Dollar amount of salary
PayPeriod	
OfferDate	
Industry	
JobFunction	
Intern	1 – Has intern experience; 0 – No intern experience
Location	Location of Job
LawOrHealth	0 – No, 1 – Yes, 2 - Unknown
LawRelated	0 – No, 1 – Yes.
FullTime	0 – No, 1 – Yes, 2 - unknown
SummerIntern	Whether the offer is for summer intern 0 – No, 1 – Yes, 2 - unknown
PostIntern	Whether the offer is for intern after graduation 0 – No, 1 – Yes, 2 - unknown
FreeLance	0 – No, 1 – Yes, 2 - unknown
Temporary	Whether the offer is temporary 0 – No, 1 – Yes, 2 - unknown
JobSource	How you get the information about the position

The file ***Class201X_StudentActivity.csv*** contains 1019 instances of Fordham students' activities related to services offered by OCS. Each of Fordham student is identified with an ID number.

Codelist	
ID	1 – 1936 The same id in both two datasets representing the same student.
Degree	
Major	
ClassLevel	Alumnus, Graduate Student, Junior, Senior, N/A
totalOnCamputRecruit	Number of On Campus Recruiting events this student has attended
totalInformationSession	Number of Information Sessions offered by OCS this student has attended.
totalConsutlingSession	Number of Consulting Sessions offered by OCS this student has attended.

5. Charles Book 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 mailing. CBC is strictly a distributor and does not publish any of the books that it sells. CBC built and maintained a detailed database about its club members.

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

1. Conduct a market test involving a random sample of customers from the database to enable analysis of customer response. The analysis would create and calibrate response models for the current book offerings.
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.

Data (Dataset: CBC.xls, available at http://storm.cis.fordham.edu/~yli/data/CBC.xls)

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. Each row corresponds to one market test customer. Each column is a variable, with the header row giving the name of the variable. The classification technique could be used to create segments based on product proximity to similar products of the products offered as well as the propensity to purchase.

CodeList	
Seq#	Sequence Position in training data
ID#	Customer ID# in market test database
Gender	Male=0, Female=1
M	monetary – total money spent on books.
R	recency – months since last purchase.
F	frequency – total number of past purchases.
FirstPurch	months since first purchase.
ChildBks	ChildrensBooks_purchased
YouthBks	YouthBooks_purchased
CookBks	CookBooks_purchased
DoitYBks	DoityourselfBooks_purchased
RefBks	Dict_Encycl_Atlases_purchased
ArtBks	ArtBooks_purchased
GeoBks	GeographyBooks_purchased
ItalCook	Num_items_purchased_of_Secrets_Italian_Cooking
ItalAtlas	num_items_purchased_of_Historical_Atlas_of_Italy

ItalArt	num_items_purchased_of_Italian_Art
Florence	bought_art_history_of_florence=1,else 0
RelatedPurchase	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.)
Mcode	0-25=1, 26-50=2, 51-100=3, 101-200=4, 201+=5
Rcode	0-2=1, 3-6=2, 7-12=3, 13+=4
Fcode	1=1, 2=2, 3+=3
Yes_Florence	bought_art_history_of_florence=1,else 0
No_Florence	not_bought_art_history_of_florence=0,else 1

Project Requirements:

Each team may select one case from the above 5 cases. Each group should perform data mining technique(s) on the given dataset and present the analysis report on (1) goal of project, (2) description of the quality/features of data (data processing) (3) data mining techniques adopted (4) performance evaluation and (5) final findings or conclusions. The team should submit the report in hard copy, and also present in class as the final exam of this course.

Grading Policy:

Each team should have at least 6 team members and have a clear definition of each team members' role in this project. The grade of this project for each student consists of two parts:

1. Quality of the project. The instructor and all students (other than team members themselves) will grade the project based on the final presentation and project report; (80%)
2. Performance of teamwork. Each team member will be graded by its team mates based on his/her performance and participation of the team project. (20%)

Timeline:

Time	**Activity**
April 1, 2016	Projects are assigned.
April 8, 2016	Lists of team members, team plans (dataset selection and roles of team members) are due.
April 19, 22, 26	Project discussion in class.
April 29th and May 3rd	Final Presentation (15 min./team)
1:30 p.m., May 10th	Final report due