

MKT 568

Quiz 2 Spring 2020

Due: March 28, 11:59 pm

Scoring: 47 pts total

Fill in the blanks where applicable. Otherwise, paste in your answer below the question.

Note: Do this quiz on your own, with no help from your teammate or other classmate.

- A. (12 pts) Consider the following 6 rules and their characteristics, as found from an apriori analysis of the Churn dataset from an unnamed telecommunications company. See the appropriate module for the variable definitions. For each rule, consider its usefulness to the telecommunications company and fill in the last two rows below, saying what actions the company should consider ("None" if the rule is not useful or misleading) and why.

ID	Antecedent	Consequent	Support (%)	Confidence (%)	Lift	Business Action	Rationale
A	International Plan=YES	Churn	3%	41%	2.8		
B	Churn	International Plan_YES	3%	26%	1.2		
C	International Plan_YES	Voice Mail Plan_YES	2.50%	75%	2.5		
D	Refund	Churn	12%	95%	6.3		
E	Monthly Minutes>500	Churn	7%	25%	1.7		
F	Voice Mail Plan_YES	International Plan_YES	2.50%	6%	1.05		

Business Action	Rationale
Provide great customer service and immediately take care of issues or complaints	This information is interesting in that if a customer has an international plan, then they will likely churn. This may happen if the customer just wants to utilize the service for their great international plan for a short amount of time and then churn. Although the support level is low, the confidence and lift numbers, which suggests that a customer is more likely to churn given a customer is on an international plan, are on the higher side so it may make sense to keep these customers on the radar to both keep them from churning and also see whether this information is more likely to happen.
None	This is redundant and silly information because if a customer churned, then it doesn't matter if they may get an international plan because they've already stopped using the service. The support, confidence, and lift levels are also low so this information should not be accounted for.
Recommendations to customers with international plans to also get a voicemail plan	If customers have an international plan, then them being more likely to have a voicemail plan would be interesting information for the company to use in order to make more money from their plans. This could happen because if a customer with the international plan tries to

	make a call and the person they are trying to reach does not pick up because of the big time difference, the customer would be more likely to leave a voicemail so that the other person can receive their message. The company could have recommendations to promote their voicemail plans and tell customers about it to garner more interest. The confidence and lift levels are high so there could be a relationship between the two plans that the company can leverage to their advantage.
Recommend other products or services and/or a discount that could prevent the customer from churning	If a customer demands a refund, then they will most probably churn is important information for the company. If a customer asks for a refund, it means they are not satisfied with the products and may very well churn. In order to prevent this, the business could provide great customer service and resolve any issues immediately as well as recommend other services or provide a discount to keep the customer's interest. The support, confidence, and lift levels are the highest for this information out of any of the other rules so it would be in the business's best interest to act on it.
Misleading, but monitor	If a customer uses 500 minutes per month of the service, it doesn't make sense they will churn because they are using the service a lot. Instead, it makes sense that they most probably continue using the service so this information is misleading. In addition, the support and confidence levels are low so this information is not very useful. However, before totally discarding this rule, we can see if there is more supporting evidence since the support level is higher.
None	This is redundant information as the company already has a plan for whether the customer has an international plan and if they will get a voicemail plan and the same plan can be used vice versa. Moreover, this rule only has a confidence level of 6% and a lift of 1.05 which makes it less useful to use.

B. Adrian's Problem (8 pts)

You are working on a new, supplemental method to indicate fraud by doctors caring for Medicaid patients. Your colleague Adrian says: "Usually, a Medicaid patient will see several doctors each year

for their illnesses. Maybe we could somehow look for pairs (etc.) of doctors seen by some number of patients where the pair of doctors are too far apart for an individual patient to visit. For instance, some patient might say, in the Medicaid database, that they visited a doctor in Massachusetts in January, then a doctor in Texas in March, and so on. That would be suspicious, especially if those Massachusetts/Texas doctors also had other patients visiting them both too. The problem I see is that there are so many long distance doctor locations, and so many sequences of visits to a doctor. A patient might visit the Texas, Texas, Massachusetts doctors, while another might visit Massachusetts, Texas, Texas, so I don't know how to use a database query to find all those sequences and long distances. In a way, it reminds me of a patient going into a supermarket, and "buying" a Texas doctor and a Massachusetts doctor. Can you think of any way we can figure out how to discover these little patterns amongst such a huge database?"

1. Is there a target variable?

The max physician distance and day of usage

2. How is a patient's annual health history like a supermarket visit: what is a "transaction" and what is an "item"?

The transaction is the Medicaid patient annual physician visits and the purchases or items are the locations of physicians visited.

3. What methods might work?

Since there are so many sequences and long distances, we can use an automated method using association rules or collaborative filtering in order to find the patterns.

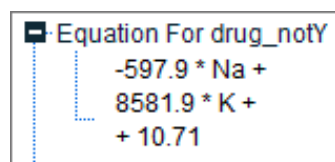
4. How would this solve Adrian's problem of huge numbers of patients and doctors?

With association rules (location-based), we can see what events happen with a lot of frequency. For example, we can count to see which pair of long distance doctor locations or which sequences of doctor visits are most frequent. So, using these pieces of information along with information generated from the Apriori algorithm such as support, confidence, and lift levels, we can find patterns that are most interesting and find our solution. With collaborative filtering (patient-based), we can find similar patients and their visited doctor locations that will be used to produce a ranked list. This could be used to look into the scenario mentioned in the problem of the suspiciousness of the MA and TX doctors having additional patients visit them.

C. Linear Classification Boundaries

- 1) The chart below shows the best drug (Y or not-Y) for treating someone with given levels of sodium (Na) and Potassium (K). It appears that the dividing line between the two best performers is a straight line. The data are contained in the csv file Drug2N in Module 9. Using logistic regression, calculate the formula that determines that line (3 pts)

Using the output from the logistic model:



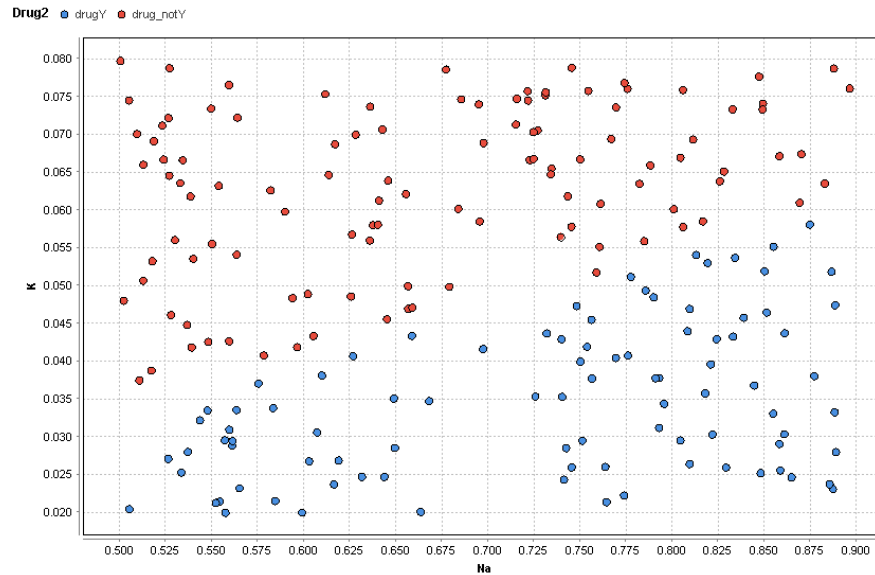
Equation For drug_notY

$$-597.9 * Na + 8581.9 * K + 10.71$$

The formula below was calculated from: $Y = e^{(a + b_1X_1 + b_2X_2)} / (1 + e^{(a + b_1X_1 + b_2X_2)})$

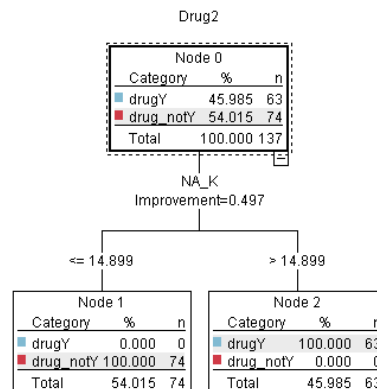
By plugging in the numbers, the equation is obtained:

$$Y = e^{(-597.9 \cdot Na + 8581.9 \cdot K + 10.71)} / (1 + e^{(-597.9 \cdot Na + 8581.9 \cdot K + 10.71)})$$



- 2) The variable Na_K is the ratio of Na and K, Na/K. Using your result in 1), or arguing from the scatterplot above, explain why this ratio would be a more useful input to a classification tree than the original inputs Na and K. (3 pts)

Inputting Na/K would be a more useful input to a classification tree than the original inputs Na and K in the logistic model because it will fit the data better using horizontal and vertical lines rather than in the logistic model that uses only a straight line to divide the data. For example, the model below was obtained which shows that 100% of the data was classified as drugY or drug_notY using the input Na/K.



- D. (6 pts) You are analyzing a customer satisfaction survey for clients in which (Overall) Satisfaction, Price Satisfaction and Reliability Satisfaction (and several other attributes) are measured on a 1-10 scale, 10 being the most favorable value for each variable. You fit a regression model of the form

$$Satisfaction = \alpha + \beta_1 Price + \beta_2 Reliability + \varepsilon$$

and find that in this model the estimated coefficient of *Price* is positive, but that of *Reliability*, is negative. You give a preliminary talk for your business clients, in which they declare this result must be false and cannot be presented to management.

- a) (2 pt) Why is this result unacceptable?

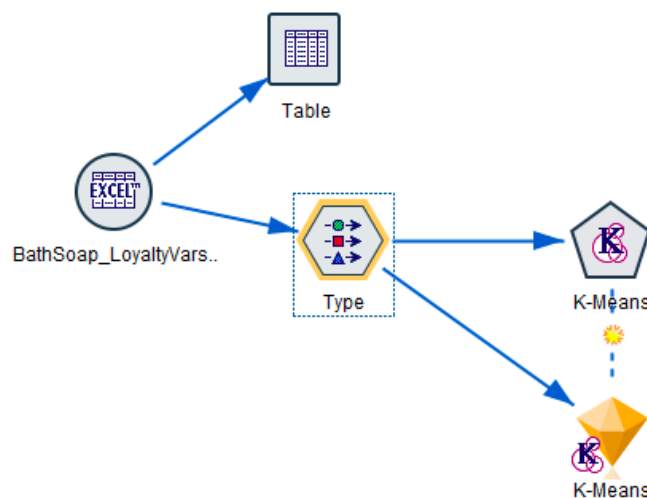
It is unacceptable because if the reliability satisfaction rate is high, it contributes to a more negative effect on the overall satisfaction rate and hence decreases it if the price satisfaction rate is held constant. So, it doesn't make sense that if a customer gives a high reliability rate then the overall score decreases.

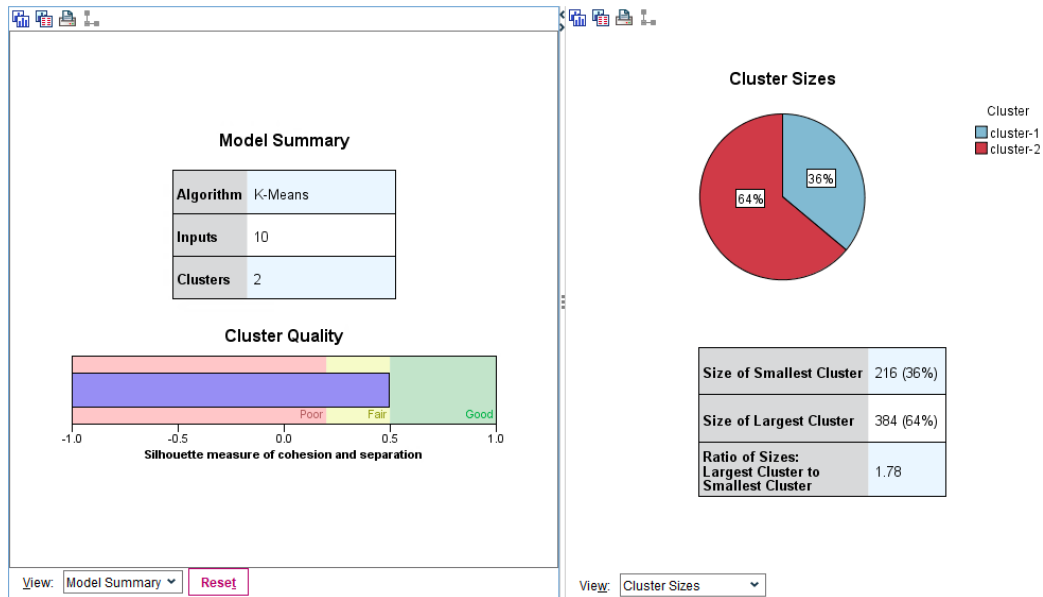
- b) (4 pts) Why might this result have occurred, and how could you modify the model to develop a more acceptable result?

This result could have occurred because the data could have been entered incorrectly or there could have not been enough of a sample size that could lead to underfitting or overfitting of the model. We could run a data audit in order to fill in any missing or null values. In addition, we could make sure to partition the data into training and testing. We could also include an interaction term in the model because the model could assume that the price and reliability are unrelated to each other, which may not be true.

E. Use the Bath Soap dataset on the Canvas site. Management's goal is to identify a small number of clusters of customers with similar patterns of brand loyalty, and target the best clusters for sales promotion campaigns. Note that the variable definitions are given in a sheet in this file.

1. Use the "brand loyalty" variables (No. of Brands – Brand Share Others) to construct k=2 clusters. Paste the stream below. (3 pts.)





2. What is the intuitive meaning of “Trans/Brand Runs”? (You may want to explain how someone with a large number is different from someone with a small number.) (2 pts.)

The meaning is how many of the transactions that a person made on each run to the store contains a purchase of the same brand. A number that is closer to 1 means that the customer has mostly bought that brand on every run to the store. A number that is larger, such as 5, means that the customer has bought different brands each time they are at the store. For example, if the total number of transactions of a customer is 100 and the brand run is also 100, this is equal to 1 (100/100) which means the customer bought this brand on every run to the store. However, if the transactions equal 100 and the brand run is 5, this is equal to 20 (100/5) so the customer has bought the same brand less amount of times.

In your cluster analysis above, paste the two cluster centroids below, and give a concise summary of the brand loyalty characteristics of those in each of the clusters. Note that in this table, cluster 2 is in the left-hand column. (3 pts)

Clusters		
Input (Predictor) Importance		
<div> <div></div> 1.0 <div></div> 0.8 <div></div> 0.6 <div></div> 0.4 <div></div> 0.2 <div></div> 0.0 </div>		
Cluster	cluster-2	cluster-1
Label		
Description		
Size	<div> <div></div> 64.0% (384) </div>	<div> <div></div> 36.0% (216) </div>
Inputs	Max Brand Share 0.19	Max Brand Share 0.70
	Brand Share Others 0.70	Brand Share Others 0.20
	Brand Runs 19.07	Brand Runs 9.85

Trans / Brand Runs 2.01	Trans / Brand Runs 3.69
No. of Trans 34.54	No. of Trans 25.13
Avg. Price 12.56	Avg. Price 10.55
Vol/Tran 368.34	Vol/Tran 498.09
No. of Brands 3.89	No. of Brands 3.19
Value 1,385.33	Value 1,252.15
Total Volume 11,569.48	Total Volume 12,528.61

The top three important inputs are max brand share, brand share others, and brand runs. These variables tell us important information about what kind of soaps each cluster tends to buy. Cluster 1 has a higher maximum percentage of purchase among all except “other” brands. But, cluster 2 has a higher brand share others, meaning they buy a higher percentage of products that are not among major brands. Their brand runs and trans/brand runs are also higher, meaning they buy the same brand at almost every run to the store and suggests they are more loyal to a brand. Cluster 2 also does a lot more transactions and spend a bit more than cluster 1 which means they may not be as constrained to price. The vol/Tran says that cluster 2 buys items that are less in volume per transaction than cluster 1. Cluster 2 also tend to buy a greater number of brands than cluster 1 which could suggest that they are willing to try new brands. The rest of the variables, value in paise and total volume of the product purchased, are not as important and have similar numbers among the two clusters.

3. Which cluster would be a better target for a promotional campaign, and why? You may want to compare the Affluence Index, Education Level and Household Size between the two clusters (try a Means node). (3 pts.)

Field	cluster-1*	cluster-2*	Importance ▾
Affluence Index	14.458	18.461	1.000 ★ Important
EDU	3.588	4.299	1.000 ★ Important
HS	4.264	4.151	0.436 □ Unimportant

The table from doing a Means Node analysis

The cluster that would be a better target for a promotional campaign would be cluster 2 because they seem to be more affluent and educated. This means that cluster 2 may not be as constrained to money and would be willing to try new brands since the number of brands they buy is higher than cluster 1. Cluster 2's brand runs and trans/brand runs are also higher, meaning they buy the same brand at almost every run to the store and suggests that they are more loyal to a brand. Also, the affluence and education factors seems to be more important measure than the household size so it would make sense to go with this information.

4. Although the Brand Loyalty variables are not readily available to identify those customer in the cluster you chose above, the demographic variables (SEC – Affluence Index) are, through a separate database kept by the marketing department. Choose a classification model method and state which variables appear to be most important in identifying your cluster? (4 pts.)

Using the logistic method, the following demographic variables were identified:

\$KM-K-Means ^a		B	Std. Error	Wald	df	Sig.	Exp(B)
cluster-2	Intercept	-.048	.882	.003	1	.956	
	SEC	-.061	.104	.349	1	.555	.940
	FEH	.055	.118	.216	1	.642	1.056
	MT	.057	.032	3.260	1	.071	1.059
	AGE	.226	.114	3.919	1	.048	1.254
	EDU	.142	.068	4.325	1	.038	1.153
	HS	-.135	.055	6.079	1	.014	.874
	CHILD	-.284	.100	8.111	1	.004	.753
	Affluence Index	.031	.012	6.273	1	.012	1.031
	[SEX=.000]	1.211	.723	2.805	1	.094	3.357
	[SEX=1.000]	-.090	.486	.035	1	.852	.914
	[SEX=2.000]	0 ^b	.	.	0	.	.
	[CS=.000]	.021	.482	.002	1	.965	1.021
	[CS=1.000]	-.314	.315	.993	1	.319	.730
	[CS=2.000]	0 ^b	.	.	0	.	.

From this table, Age, Education level of the housemaker, Household size, presence of children, gender, and presence of a television have the highest coefficients so these affect the output the most. So, these variables are important whereas the socioeconomic class, food eating habits, native language, and affluence index are less important according to the logistic classification model.