Sample Questions

True/False

For supervised models the value of the target variable is known when the model is deployed. False

We can build unsupervised models when we lack labels for the target variable in the training data. True

Finding the characteristics that differentiate my most profitable customers from my less profitable customers is an example of an unsupervised learning task. False

Data Science applications are only useful when automated in a large technology system. False

As long as data is available, Data Scientists can always create value for their organization. False

It is sometimes necessary to engage in a sub-optimal decision policy to collect the necessary data required for modeling. True

Data Scientists should not invest resources learning about the problem domains they are modeling. False

Data Science problems are best solved by breaking the problem down into well defined sub-problems True

A “target variable” and a “label” commonly refer to the same thing in the Data Science lexicon True

A data “leak” is a situation where good predictors can be identified in a data set False

Research Data Scientists should not spend time working with engineers to implement prototypes – their time is better spent experimenting with new methods. False

Pruning is a technique for reducing complexity in Decision Trees. True

In a classification tree induction, the next attribute added is the one with the largest information gain. True

Support-Vector Machines (SVMs) approach classification problems by finding the widest possible bar that fits between points of two different classes. True

Cross-validation is used to estimate generalization performance. True

**Multiple Choice**

In order for data points to be taken as input to most data mining programs, they must be represented as:

* + a)  text documents
  + b)  feature vectors
  + c)  dependent variables
  + d)  targets

Regression is distinguished from classification by:

* + a)  class probability estimation
  + b)  numerical attributes
  + c)  numerical target variable
  + d)  hypothesis testing

Entropy

* + a)  is a measure of information gain
  + b)  is used to calculate information gain
  + c)  is a measure of correlation between numeric variables
  + d)  has a strong odor

Which of the following does not describe SVM (support vector machine)?

* + a)  SVM can estimate class membership probability
  + b)  SVMs are based on supervised learning
  + c)  SVM chooses the line to minimize the margin between two classes
  + d)  SVM can be applied when the data are not linearly separable

Which of the following is not true about logistic regression:

* a)  Logistic regression predicts probability of membership in a certain class.
* b)  Logistic regression takes a categorical target variable in training data.
* c)  A logistic regression represents the odds of class membership as a linear function of the attributes.
* d)  Logistic regression requires numeric attributes and categorical attributes should be converted to numeric attributes.

A learning curve aims to test:

* a)  True positive rate vs. false positive rate
* b)  True positive rate vs. false negative rate
* c)  Generalization performance vs. size of training set
* d)  Generalization performance vs. alternate feature transformations

The area under the ROC curve is not

* + a)  equal to the Mann-Whitney-Wilcoxon statistic
  + b)  a measure of the quality of a model’s probability estimates
  + c)  likely to be at least 0.5
  + d)  larger when false positive errors cost more

The points on a model’s ROC curve

* + a)  represent the performance of different thresholds
  + b)  represent different rankings of examples
  + c)  represent the cost of different classifications

I want to rank credit applicants by their estimated likelihood of default. Which technique would be least helpful in assessing the quality of a ranking model mined from data?

* + a)  holdout testing
  + b)  calculate area under the ROC curve
  + c)  calculate percent correctly classified instances
  + d)  cross-validation
  + e)  domain knowledge validation

Which is not a technique for reducing/avoiding overfitting in tree induction?

* + a)  choose largest improvement in information gain
  + b)  prunetree
  + c)  select tree size based on validation data
  + d)  reduce tree size based on statistical test

Unsupervised data mining

* + a)  Requires less effort early in the data mining process
  + b)  Is easier to evaluate than supervised data mining
  + c)  Cannot be applied if we have a well-defined target variable
  + d)  Needs minimal domain knowledge

Similarity measures are most essential for

* + a)  Naïve Bayes
  + b)  Tree Induction
  + c)  Hierarchical Clustering
  + d)  Logistic Regression

Which is not true of clustering?

* + a)  Centroid-based clustering is the procedure that all observations start in one cluster, and  splits are performed recursively as one moves down the hierarchy
  + b)  Domain knowledge can be incorporated
  + c)  The “K-means” algorithm is an iterative center-based algorithm
  + d)  It is used to group objects represented by multiple attributes

Which is not true of k-Nearest Neighbor (k-NN)?

* + a)  It can incorporate domain knowledge
  + b)  It builds a simple induction model
  + c)  It is robust to noisy data
  + d)  It is easy to explain how it works

One key part of the data mining process is creating attributes to describe examples. In order to represent documents (such as emails) as examples, we create term (word) based attributes to describe the documents. Which of the following is not a common approach?

* a)  whether or not the term appears in the document (binary attribute)
* b)  term frequency (number of times term appears in document)
* c)  term frequency/total number of terms in document
* d)  term frequency times the term’s frequency in the document corpus

Data Science for Business Analytics

Which technique divides the population into disjoint segments described by combinations of features?

* + a)  Naïve Bayes
  + b)  Treeinduction
  + c)  Logistic Regression
  + d)  k-NN

Which data mining technology would be most useful in answering the following business question? “Of all my accounts, which are the most likely to exhibit fraud, based on my experience with prior cases of accounts that have and have not been defrauded?”

* a)  Classification tree induction
* b)  Hierarchicalclustering
* c)  k-Means
* d)  Linear regression
* e)  Association finding

Which business intelligence technology would be most useful in answering the following business question? “If this customer responds to my offer, how much will she spend?”

* + a)  Classification tree induction
  + b)  Hierarchicalclustering
  + c)  k-Means
  + d)  Linear regression
  + e)  Association finding

Return to the Target/pregnancy case from class. What was the actual target variable that Target used to build their “pregnancy prediction” models?

* + a)  Which customers are most likely to sign up for Target’s baby registry?
  + b)  Which customers are most likely to actually be pregnant?
  + c)  What are the characteristics of customers who switch to Target after major life changes?
  + d)  Which customers are most likely to buy lots of unscented lotion and vitamins?
* 25)  Which of these organizations would have the most challenge in applying supervised predictive modeling?
  + a)  A business school that wants to start a new Master’s degree program in Business Analytics, and would like to estimate the likely number of applicants.
  + b)  A grocery store that is trying to identify which of its loyalty-card-carrying customers will spend more than $100 next month.
  + c)  A city government that is trying to predict which neighborhoods will see the most new businesses open up next quarter.
  + d)  An online marketing company that wants to estimate the number of clicks that the ads it serves will receive when shown to a particular population.

Data Science for Business Analytics

9

Matching

Find the best matching of the items on the right with the items on the left, where each item on the right appears in only one blank on the left.

Data Science for Business Analytics



|  |  |
| --- | --- |
| \_\_\_\_\_ entropy \_\_\_\_\_ logistic \_\_\_\_\_ information gain \_\_\_\_\_ regression \_\_\_\_\_ SVMs | 1. log odds 2. numeric target 3. how mixed up classes are 4. difference between parents and children 5. maximum margin |
| \_\_\_\_\_ model \_\_\_\_\_ induction \_\_\_\_\_ attribute \_\_\_\_\_ example | a. feature vector b. learning c. simplified representation d. feature |
| \_\_\_\_\_ cross-validation \_\_\_\_\_ domain-knowledge validation \_\_\_\_\_ ROC curve \_\_\_\_\_ overfitting avoidance | a. Ranking b. Comprehensibility c. Generalization performance d. Complexity control |

10

Data Science for Business Analytics

|  |  |
| --- | --- |
| \_\_\_\_\_ holdout evaluation \_\_\_\_\_ domain-knowledge validation \_\_\_\_\_ learning curve \_\_\_\_\_ overfitting \_\_\_\_\_ increasing comprehensibility | 1. Sanity checking 2. Pruning 3. Increasing data 4. Cross-validation 5. Divergence between training and testing accuracies |
| \_\_\_\_\_ learning curves \_\_\_\_\_ fitting curves for kNN \_\_\_\_\_ information gain \_\_\_\_\_ cumulative response curves \_\_\_\_\_ increasing Mann-Whitney-Wilcoxon | 1. increasing proportion targeted 2. increasing tree size 3. increasing training data 4. increasing complexity 5. increasing AUC |
| \_\_b\_\_ Classification task \_\_c\_\_ Scoring/Ranking task \_\_a\_\_ Regression task \_\_d\_\_ Unsupervised learning | 1. How many cell phone minutes will each customer use next month? 2. Which customers will leave within 90 days of their current contract expiration? 3. Which 500 customers should I target with a special offer? 4. Are there any interesting natural groupings of my customers? |

11

Short Answer

* 1)  You would like to build a model for predicting defaults on student loans. You are given a large number of categorical attributes of each loan such as the type of the school that a student is going to attend, the state where it is located etc., as well as numerical attributes such as outstanding loan amount, student’s age, loan interest rate and so on. Your client asks that your model must provide a clear explanation of the reason for its predictions, since the final judgment on whether to give a loan or not will be made by a human agent. What data mining technique would you suggest using? Explain why briefly (one or two sentences).
* 2)  You came up with the pregnancy prediction idea for Target. Explain precisely to your tech team how to formulate the data for supervised mining. In particular, (1) give a precise statement of the target variable to use; (2) explain how to create the features, and (3) briefly describe the use phase for your model.
* 3)  Briefly give an example of a task where you’d use each type of model, being clear about why that model is appropriate:
  + a)  Linear regression
  + b)  Logistic regression
  + c)  Decision tree
* 4)  I want to build a logistic regression, but I have a very large number of possible attributes describing my instances. Why might this be a problem? How should I deal with the problem? (Two sentences for each question)
* 5)  What exactly does the area under the ROC curve represent? Be as precise as possible.

Data Science for Business Analytics

12

* 6)  Give two different reasons why using ROC curves can be more effective for assessing model quality than the percent of classifications that are correct (a.k.a. “vanilla” accuracy).
* 7)  Last month your boss sent a mailing to 20,000 of your existing customers with a special offer on a Hoosfoos credeen. The response was exciting: 1% of them responded, which brought in $200,000 in revenue. She has now delegated to you the task of continuing the program, and has given you a budget of $10,000, which will allow you to target another 20,000 customers (out of your customer base of 100,000). You don’t want to just target them randomly, as your boss did. You build a tree model and a logistic regression. Describe how to evaluate them as follows. Describe (a) the confusion matrix and (b) how you will fill it out for one of the models. Describe (c) the cost/benefit matrix for this problem, including the costs and benefits for this case. (d) Show the evaluation function you will use to compare your systems. (e) How do (a) and (c) come into play in this evaluation function?
* 8)  Explain the meaning of each of the different terms in Bayes Rule. Describe one way that this rule is used for data mining. What is naïve about Naïve Bayes?
* 9)  You have been asked to oversee the design of a new system for targeting advertisements to web browsers. You realize that the designers have not considered building a predictive model based on the text of web pages. Assume that you can associate a precise target variable of interest to each web page. Describe how you would suggest that they represent the web pages as examples in order that they can be taken as input to standard supervised predictive modeling techniques. Be as complete as you can.
* 10)  You are a manager for two data miners A and B that work on a project for preliminary screening of a population of people for the early detection of Provost’s Quizinoma. Although very rare, this disease is deadly for the person bearing it if not identified in time, so your task is quite important. After preliminary screening, a $750 blood test can determine the presence of the disease with almost perfect accuracy. You decided to motivate your analysts by

Data Science for Business Analytics

13

structuring their work as a competition: both data miners A and B have to work independently on the problem and then present their results separately. After the competition period is over, on the test data, data miner A reports 99.9% percent correctly classified instances from her model, while data miner B reports only 86.3% percent correctly classified instances from his model. Describe carefully how you would determine which algorithm is preferable? Illustrate with some hypothetical example numbers.

* 11)  Distance is a key notion underlying many data mining algorithms, such as k-nearest neighbor (k-NN). What problem is there with comparing consumers using regular Euclidean distance, for example when they are described by age (in years), income (in dollars), and number of credit cards? How can this problem be fixed?
* 12)  In order to apply data mining to text documents, we need to represent text documents by some term (word) based features/attributes. Please describe three common approaches to create attributes for text documents. Please give a clear explanation of all your terminologies.
* 13)  Similarity is a key notion underlying many data mining techniques. Assume that you are employed by Pandora to make music recommendations. Give Pandora an artist or song, and it will find similar music in terms of melody, harmony, lyrics, orchestration, vocal character and so on. If you plan to use k-NN algorithm to finish the music recommendation task, please carefully describe how you would proceed.
* 14)  Similarity is a key notion underlying many data mining techniques. If you use Euclidean distance to find similar examples, how can you deal with categorical attributes? The k- nearest-neighbor technique estimates the target variable based on the k most similar examples. How exactly would you estimate the target variable for a regression problem? Explain the pros and cons of using different values for k, for example k=1 and k=N, where N is the total number of training examples. How would you choose k?

Data Science for Business Analytics

14

* 15)  My retail customers buy lots of products over the course of a year. I want to analyze the products purchased by my customers to determine whether there are patterns of co-purchases that I’m not aware of. What technique would I use? How could I include customer demographics in those patterns?
* 16)  Evaluation for clustering can be challenging; briefly discuss two different ways to understand the meaning of the clusters found by k-means clustering.
* 17)  A key notion underlying k-means clustering and k-nearest neighbor methods is the same-- what is it? Considering this, what should one watch out for during problem formulation for k- means and for k-nearest neighbor? (One sentence or phrase each.)
* 18)  Tree induction and clustering both can be used to segment customers. Contrast the process for using these two different types of modeling for customer segmentation. For what sort of problems would you use each? What would you have to do differently?
* 19)  You sell IT products and are using k-NN to build an IT wallet estimation predictor. Recall that wallet share is the percentage of a company’s total budget that is spent with your company. You have information on the total IT budgets of a large set of companies, which will constitute your database of potential neighbors. You already have defined a distance measure, and have chosen k. Now you want to estimate your wallet share for Acme Corp., one of your current customers for whom you do not know the IT budget.
  + a)  Explain precisely how you will estimate your wallet share for Acme with this technique.
  + b)  If you choose k=N, the total number of training examples, what is the effect?

Data Science for Business Analytics

15

* 20)  How do classification trees estimate the likelihood of class membership for a test example? In analogy, how might a regression tree estimate the target variable for a test example?  Answer:  Decision trees would deal with a test example first by assigning it to a leaf in the tree based on its attributes. For classification trees, that leaf will have been assigned a certain number of positive and negative examples during the training phase; the percent of positive training examples associated with the leaf will be the likelihood of class membership. For a regression tree, the estimate of the target variable will be the average of the training target variable values for the leaf that the test example is assigned to.
* 21)  Assume that you work for a credit card company. You’ve noticed that there has been a recent increase in the number of late payments on bills and want to predict which customers will be late in future time periods. In one sentence, formulate a useful target variable. In another sentence, describe precisely how you would formulate the feature vector. Finally, briefly describe the use phase for your model.  Answer:  A useful target variable for this problem would be whether a customer will make their next payment more than 30 days after their next bill is due. I would formulate the feature vector by looking at some of the data we have available for each consumer; in particular, demographic data such as age, income, zip code, etc., as well as any history with the company that they have such as amount of time they’ve had a credit card, average monthly payment, average balance, etc. In the use phase, I would apply my model to all of the customers to estimate which ones are the most likely to make a late payment on their next bill, and inform my boss so that they could perhaps make some intervention to prevent the late payments.

Data Science for Business Analytics

16

Proposal Evaluation

There will be one “Critique this proposal” question. To practice, see the example on p. 325 of the book, the example in Appendix B (p 351), and the corresponding HW.

Answer:

Answers to the book examples are in the book. Some possible answers to the HW are:

* Unclear that there are suitable training data. It is not clear that there are any negative examples -- it was a word-of-mouth campaign, so it is unlikely there are negative examples. Also, even forgetting that, the examples were not sampled from Blue Moon's database; it is unclear what selection bias is incorporated in the word-of-mouth campaign.  Remedy: invest in targeting a sample of consumers first, and then (re)model.
* Why choose LR beforehand? Maybe some other technique will have better  generalization performance. Remedy: try several predictive modeling techniques and compare the models' generalization performances.
* Expanding on the attributes question: even besides usage, what is the guarantee that the attributes we have correspond to those that BM has. Remedy: again, invest in gathering response data using BM's consumer base
* Statistical significance and interpretability of the coefficients do not guarantee accuracy of predictions.  Remedy: better to look at some measure of generalization performance if you want to  gauge the accuracy of your model in predicting service uptake
* You don't necessarily want to target the most likely adopters since some of those people  would adopt anyway. Therefore the ranking model won't necessarily rank by expected  profit of targeting.
* If the model is built with usage of the service as a feature, you won't be able to use that to  predict adoption for current non-users
* There's no notion of cost of marketing or budget. So we don't know how many people we  would/should target.

Data Science for Business Analytics

17

Other Comprehensive Essay Questions

1) Your company has hired a data mining consulting firm to help with your fraud problem. You are present in the meeting where the consulting firm presents the results of their pilot study, showing that the model has a very low error rate (percent incorrectly classified instances). They argue to your boss that based on this great performance, she should hire them to build the system. You need to explain to your boss the notions of false positive and false negative errors, and how the system should be evaluated. You may assume that the only relevant decision is (binary): if the system predicts fraud, block the account; if the system predicts no fraud, do nothing. Show me the key notions of

* a)  confusion matrix and
* b)  cost matrix.
* c)  Explain briefly why these are important for this problem (1-2 sentences).
* d)  Show the proper evaluation function (equation) for the consultant’s system.
* e)  How do the confusion and cost matrices come into play in this function?

Answer: The first part is straight from the book:

* a)  See Ch. 7 p. 189-190
* b)  See Ch 7 p. 193 & Figure 7-2. You should give some example costs benefits. Remember the “double counting” warning on p. 203. See (d) below.
* c)  These are important because in fraud detection the costs/benefits of the different errors/correct decisions are not equal. Thus, it does not make sense to simply add up the errors and report the error rate. Rather, one should determine the corresponding costs and benefits of using the system.
* d)  The equation is essentially the expected profit equation on pg. 200 – except here we might want to cast it as the expected loss – and we’d want it to be as close to zero as we can get. Let’s say “Y” means that we caught the fraud with the new system, and “N”

Data Science for Business Analytics

18

means that it would go uncaught until noticed by some other means (e.g., the customer contacts us). Let’s say that F means it’s really fraud and NF means that it’s not.

EL = p(Y,F)\*b(Y,F) + p(N,F)\*b(N,F) + p(N,NF) \* b(N,NF) + P(Y,NF) \* b(Y,NF)

we here assume that we will estimate the ELS by using an estimated average loss per account if the fraud were not detected. (You might notice that depending on the variance in the losses, we might do better with a more elaborate estimation, for example considering each test case individually. Not necessary for this question due to the problem statement.)

So, for the benefit (loss saved) matrix we might have:

Data Science for Business Analytics

|  |  |  |
| --- | --- | --- |
|  | F | NF |
| Y | 0 | -$50 |
| N | -$1000 | 0 |

This matrix codifies that on average a case of fraud costs us $1000 more if we don’t catch it (it may already have cost us something that we now can’t avoid), and falsely bothering the customer costs us $50 in customer aggravation.

[NB: in reality for fraud detection we might want to create a more complex expected cost setting where we actually look at the actual individual cases, rather than assuming a fixed cost. But not necessary for the answer to the question. Often one would use a formulation like this early on, in order to get moving on other things, and then improve the cost/benefit formulation later—for example, if results are looking good.]

e) These come into play because they specify the terms in the ELS equations: the “p” terms from the confusion matrix (matrix cell counts divided by total test examples) and the “b” terms from the cost/benefit matrix. [NB: I notice here that between the time I wrote this

19

question and the writing of the book, I standardized on the cost/benefit matrix representing benefits, with costs as negative benefits. Generally the term “cost matrix” is used throughout data mining practice and writings, but it ends up being a little misleading because it seems to miss the benefits side of the equation. On the other hand, casting costs as negative benefits can seem awkward in a true cost-minization application. So, the upshot is to be flexible and try to figure out what makes the most sense and will be most sensible to any stakeholder with whom one needs to communicate the results.]

2) Consider our telco churn example from class. We would like to target some subset of our customers with a brand new special offer prior to the expiration of their contracts, in order to minimize our losses due to churn. The special offer has a fixed cost of C irrespective of whether the customer accepts it. As we mentioned in class, in a real business application you want to design your solution carefully and specific to the application. You can assume that the value of a particular customer, if she were to stay, can be estimated directly using this particular customer’s past value and knowledge of the business going forward and the incentive, so you don’t need to use data mining to estimate the value of the customer. You can ignore any value of the customer if she leaves the company.

* a)  How would you approach this problem from the point of view of expected value in order to select the subset of customers who should be targeted? Describe carefully your procedure.
* b)  How would you decide which customers are legitimate candidates for targeting? Assume you have a budget of $5,000,000 for your campaign. How would you determine how many of these customers you will target?
* c)  Specify what model or models you would suggest building, including specifying the target variable(s).
* d)  Do you expect to have the data necessary to build these models? If so, from where, if not, what do you propose to do about it?

Data Science for Business Analytics

20

Answer: This is essentially the running example in the book, culminating in the expected value decomposition in Chapter 11.

* a)  The main idea is that the expected value framework can decompose this complicated business problems into subproblems that can be addressed by classification and regression modeling. For this problem we want to compute (i) the expected value if we do not target the customer, and (2) the expected value if we do target the customer. Then we will target those customers with the highest expected value; this could either be those with overall expected value > 0, or the top K such customers such that the total cost of targeting is less than some budget limit B. (Technically, not just the top K highest expected value customers, because there might be fewer than K with positive expected value. Maybe reserve one point for this if everyone looks to be doing very well, but it’s not a critical detail for this exam.) So, for a customer described by feature vector X and an incentive I with fixed cost C:  EV(X, notI) = p(stay|X)\*value(X,notI) , where value is the “lifetime” value (e.g., suitably discounted future value) of the customer presuming that she stays. p can be estimated using a classification model. One might formulate a regression modeling for value, but that is not necessary given the problem statement.  EV(X,I) = p(stay|X,I)\*value(X,I) – C , where p and value can be estimated similarly.  So the expected value to the company of targeting X with a special offer is EV(X,I) – EV(X,notI), i.e., how much added value we get by targeting the customer over not targeting the customer.
* b)  I would build classification (class probability estimation) models separately for p(stay|X,notI) and p(stay|X,I). Let’s call those Pni and Pi. The target variables would be:  Pni: whether or not a customer left/will leave within 90 days of contract expiration. Pi: whether or not a customer who has been given the incentive stays/will stay (e.g., enter into a new contract)  Any reasonable supervised classification model will suffice here (say, logistic regression).
* c)  We should have plenty of data available to compute Pni. Lots of customers’ contracts expire all the time, we just have to look back more than 90 days. We should have or be

Data Science for Business Analytics

21

able to obtain all these attributes. We do not have the data for Pi. This is a brand new special offer. We will need to somehow get training data. Possibilities include investing in a pilot study, where we make the offer to some customers and then wait to see their reactions (and maybe to customers with values “high enough” to care), or use a proxy target variable as discussed in class (such as a similar offer—but do note that there could be a sampling bias issue if the prior offer were targeted).

3) I would like to see whether my investment customers tend to cluster in understandable groups.

* a)  What precisely is the difference between segmenting my customers using clustering and segmenting my customers using tree induction, and when would I use one rather than the other? What is the practical difference— i.e., in the data mining process where and what would the main differences be?
* b)  Describe what is the important concept that I need in order to apply clustering to my customer data.
* c)  Considering that you answered part (b), what might still be very unclear for you before running k-means algorithm?
* d)  Describe another type of clustering method that I should consider. How is it different from k-means?
* e)  Once I get clusters, it is important for me to see if I can understand the meaning of the clusters. Describe three ways to help understand the resultant clustering (as discussed in the book).

Answer:

a) Segmenting using tree induction is supervised. That means there is a particular target variable that guides the segmentation, such as which customers use on-line brokerage vs. phone brokerage. For supervised modeling, we need to have values for the target variable for the training data. Clustering is unsupervised, meaning the clustering is not driven by a particular target variable. An example problem for unsupervised clustering is “Do we have different characteristic groups of high-net-worth customers?” You’d use

Data Science for Business Analytics

22

supervised segmentation if you want to segment your customers based on some particular target, and you have values for it.

The practical difference is that with supervised modeling one spends more time formulating the problem as a classification/regression problem; with clustering one spends much more effort on evaluation of the results. (And messing with the distance function, but I don’t think we discussed that explicitly in class.)

* b)  Define a distance function; maybe normalize values.
* c)  How to choose k.
* d)  Hierarchical. Agglomerating/doesn’t need k in advance. A.k.a. divisive, or  “dendrogram-based”. As of class 4/28 we also have soft clustering by fitting the  parameters of a numeric function/distribution – such as the mixture of Gaussians.
* e)  Three ways to understand the clusters:
  + i)  look at the cluster centers (as if they were instances)
  + ii)  look at exemplars from the clusters (like Glenmorangie)
* iii) label the examples with the cluster ids, and then learn a comprehensible  classification model to discriminate them!

4) You are on an interview where they notice that you’ve taken a data mining class.

* a)  They ask you about what you learned there, and besides talking about nitty-gritty modeling stuff, you want to give a bigger picture. Explain why it is important to think about data mining project strategically, with respect to making internal investments. What sort of investments might you have to make?
* b)  Now they’re interested and ask you if you believe a firm can achieve sustained competitive advantage from data mining. Of course you start with “Well, that depends...”, but then you want to go on. Give 5 reasons why data mining may indeed give sustained competitive advantage, even though the basic data mining technologies are easily acquired/replicated. Be as precise as possible.

Answer:

a) It’s important to think about data mining projects strategically because making the investments can require higher-level decision-making than that of a particular project

Data Science for Business Analytics

23

manager, and the investments may involve incurring losses in the short terms, in expectation of a (potential) payoff in the longer run. This was the lesson from Signet Bank/Capital One. Also, data mining requires cross-functional cooperation, which is greatly facilitated by explicit strategic focus.

The investments involve people (data mining projects need a broad spectrum of expertise + top-notch data scientists can be expensive to acquire), software and systems, and perhaps most overlooked, possibly data (cf., Capital One or the end to our churn case).

You may want to invest in patents on your data mining process/techniques, or to

keep trade secrets. (See Chapter 13.) b) Some reasons (Ch. 13):

* i)  How to get data mining projects to work involves getting lots of little things to work simultaneously.
* ii)  You may have complementary assets that are not mobile, such as particular data. (You may have acquired these, for example, through particularly favorable historical circumstances.)
* iii)  You may have a particularly suitable corporate culture: cooperative, experimental
* iv)  Your firm may be particularly attractive to the analytical or technical workforce, and  so you may be able to attract the best for less.
* v)  You may have one or more managers who have proven themselves to be very effective  at producing successful data mining projects.
* vi)  You may have patents on your data mining process/techniques, or trade secrets.

5) We would like to target some subset of the huge number of visitors to our main retail web page with a new special offer. Instead of the normal early May special offer of a discounted flower bouquet for Mom, we’ve decided to offer select customers a 30% discount on any electric razor purchase from our stock.

* a)  Show how computing expected value provides a framework for thinking about what models need to be built for this problem.
* b)  Specify what models you would build.

Data Science for Business Analytics

24

c) Do you expect to have the data necessary to build these models? If so, from where, if not, what do you propose to do about it?

Answer:

* a)  We should compute the expected profit for a visitor from each of the offers, and offer each visitor V the offer that promises the highest expected value. Call the offers B(ouquet) and R(azor).  EP(V,B) = p(B|V)\*v(B,V) EP(V,R) = p(R|V)\*v(R,V) Here the costs of the offers can be built into the v functions, since we won’t incur the costs unless the offers are taken.
* b)  The p functions actually are different models, one for B and one for R, that estimate a visitor’s probability of accepting the offer. These could be any of the models we have discussed for classification (that produce decent prob estimates, so not regular NB for example). Trees or LR for example. The v functions would be our estimated value from a customer responding to each offer. We’d probably want to build models for these, because we’d need to estimate how much a razor buyer would spend, and similarly for the bouquet buyer, who may be upsold or cross-sold in the process. Again, tree models, linear models, etc. Compare this to the discussion of the donation case in Chapter 11.
* c)  We ought to have data for both models on the normal May special offer. We would need to gather data for the razor offer. We could conduct a small pilot study to randomly selected customers, offering them the razor discount and recording response vs. not, plus ultimate revenue. These would give us our training labels for both models (probability and value). As to attributes, we’d have to look to see what information we have on visitors. Since we have a huge number of them, we probably have prior browsing history, and hopefully even know many of them as prior customers, in which case we can get potentially a lot of features (e.g., by registration, prior activity on the site, or linking to third-party data via credit card or even billing zipcode).

Data Science for Business Analytics

25

6) You scored a cool data-mining-oriented support job with Right Media, owned by Yahoo!. You’ve been working with engineers from Yahoo Research to design new data-mining-based techniques to better target display advertisements. Right Media is one of the top two ad exchanges, and delivers over 5 billion ad impressions each day. Your new techniques describe a browser-site-ad combination by various features, and then predict the probabiity that a user will click on the ad. (You have similar models for other sorts of conversions, too.) You are considering offering a premium targeting service. Based on a small pilot study, your preliminary evaluations show promise for the new techniques. You have a meeting scheduled with the Chief Marketing Officer (CMO) to discuss your project, and you would like to prepare for strategic-level questions regarding scaling up your project.

* a)  In framing the problem of targeting on-line advertisements as a supervised data mining problem, what will be an instance? What will be the target variable? What will be the features?
* b)  What investment do you see as being most important to the success of this venture? Why? Be precise. What implications might this have that the CMO should be aware of?
* c)  BONUS: You would like to be prepared to field questions about whether this capability has the promise to yield sustained competitive advantage for Yahoo over Ad Networks who are also designing methods for targeting display ads. List the top three reasons that it might, for this specific case. Be specific.

Answer:

a) An instance will be an impression opportunity: a “slot” on a page, plus a browser visiting the page, plus any information about the moment of the impression (time of day, day of week, etc.). The target variable will be some quantity of interest to the advertiser, such as clicking on an ad or buying a product. The features will be feature of the page, such as its topic, and features of the browser, such at what can be gleaned from the IP address (such as geographic location), what can be seen in the http request (they might not know that), and data we have gathered on the pages that this browser has visited in the past. We can get this information because we are a huge ad exchange: the publishers are our customers, and we can have them characterized. We see the browser, so we know the IP address (etc.). We see massive numbers of browsers visiting massive numbers of

Data Science for Business Analytics

26

pages, so we can gather a good deal of data on which pages browsers visit (and if we have the publishers categorized, we know the categories of pages that browsers visit, so we can create variables based on this). We observe browsers clicking on ads, so we can get that target variable. The one tricky thing is where we might get the values for conversions, but we did discuss the notion of pixeling sites to get conversion actions (and other actions). However, if they just use clicks as the target variable, that’s fine.

* b)  We’llneedtoinvestindatatotrainandtoevaluatethenewtechniquesatascalelarger than the pilot study. This will be an investment, because we can’t just show the ad to those who we think are most likely to respond. To train and evaluate the model, we have to target more-or-less randomly. The implications are that the ads served for the data gathering purpose may not be as profitable as if we use our current best practices.
* c)  BONUS: Some possible reasons (see Ch. 13):
  + i)  We have a much larger volume of data than the ad networks, across many many  advertisers and many many publishers. This is a complementary resource that will be hard to replicate, and we believe that more data will lead to more accurate targeting models. Furthermore, this can create a virtuous cycle, wherein we create better targeting, which brings us more customers, thereby increasing the amount of data we have.
  + ii)  We are Yahoo!, and have tremendous analytical/technical capability via Yahoo! Research. We probably have much more sophisticated data mining/machine learning expertise. These experts value being at the research lab, and the freedom and prestige that it affords them. This will be very costly/impossible for most ad networks to replicate. What would it take to get the very best machine learning expertise to choose an ad network over Yahoo! Research?
  + iii)  Research labs like Y!R are patenting machines. We presumably can create an intellectual property fortress that can sustain competitive advantage.
  + iv)  Competitors will not be able to tell by looking at our results exactly what has led to our success (causal ambiguity, plus trade secrets). Data mining-based products are attractive in this respect, because how the work exactly is not outwardly apparent.
* Lesser reasons for this particular case:

Data Science for Business Analytics

27

* i)  Yahoo already has managers who can succeed at data mining projects. (They may ... do we know that?) and/or has a management culture of cooperation (evidenced by this pilot project) and experimentation (ditto).
* ii)  Beyond technical trade secrets, competitors will not know exactly how we’ve achieve what we have. (Causal ambiguity -- a lesser answer here.)