Data Science for Online Customer Analytics
UIC Fall 2017
Final Exam

Name: LAAVANYA GANESH
UIN: 654324917

Multiple Choices / Matching (15 points)

Select one answer for each of the following questions:

1. Which is **not** a reason why data mining technologies are attracting significant attention nowadays? (3 points)
   a. There is too much data for manual analysis
   b. Data are difficult to transfer from databases
   c. Data can be a resource for competitive advantage
   d. Machine learning algorithms are easily available

2. More complex models (3 points)
   a. have better predictive performance
   b. tend to overfit more
   c. are easier to train than simpler models
   d. are very interpretable

3. A binary classifier achieves 95% accuracy on a test set consisting of 95% positive and 5% negative instances. If we use the same classifier on a test set comprised of 50% positive and 50% negative instances, we expect to get: (3 points)
   a. higher accuracy
   b. lower accuracy
   c. the same accuracy
   d. cannot be determined

Match tasks with the appropriate task type (4 points):

Task Type	Example Task
____________Regression c) ____________Classification d) ____________Database query a) ____________Hypothesis testing b)	a) How many students over 26 years old are in the MSBA program? b) Is there an academic performance difference between MSBA students over and under 26 years old? c) What GPA will a new MSBA student achieve? d) Will a new MSBA student fail the program?

Match concepts with the appropriate definition (2 points):

Task Type	Definition
_______True Positive Rate (y-axis of ROC curve, also known as "Recall") ----- a) _______False Positive Rate (x-axis of ROC curve) ------- b)	a) TP / (TP + FN) b) FP / (FP + TN) c) TP / (TP + FP) d) FP / (FP + TP) TP=True Positive, FP=False Positive TN=True Negative, FN=False Negative

Short Answer (15 points)

ANSWER

Vaccine Cost-Benefit Matrix (5 points)

In a classification application we are asked to predict whether kids are going to be infected with the flu virus during 2017 or not, and if yes vaccinate them against it. The vaccine costs $10. If a child is vaccinated, there is only a 10% chance that she will be infected. If a kid gets infected, the cost of treatment is about $1000. Write down the cost-benefit matrix for the problem.

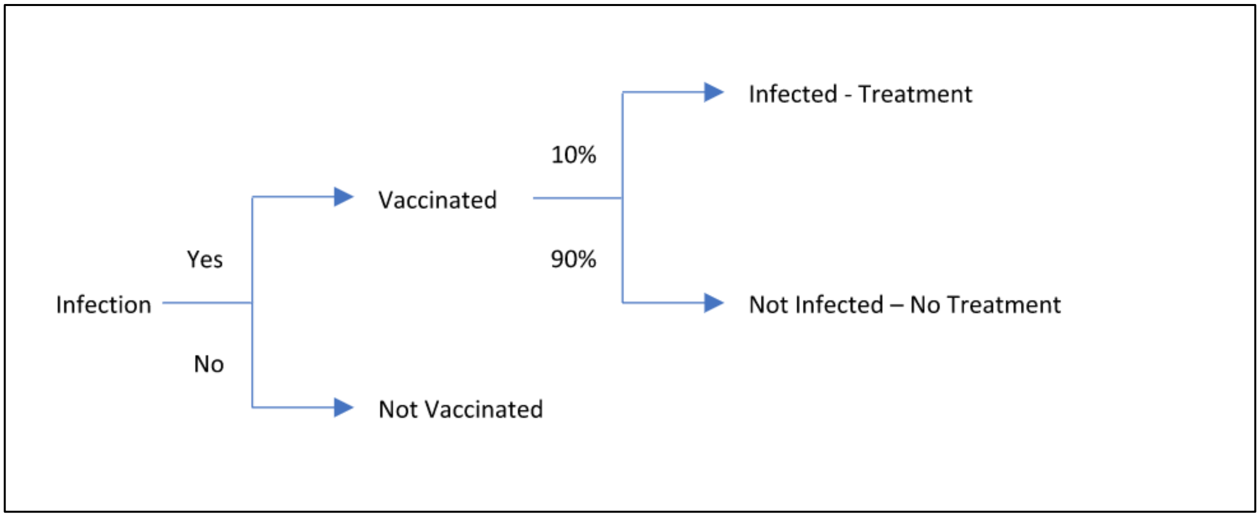
	Actual Infected	
Predicted as Infected	**TP** Cost = $10+(0.1*$1000)=$110	**FP** Cost = $10
	FN Cost = $1000	**TN** No cost is associated

Cost of Vaccination:

- For every person predicted to be infected by the disease the vaccination costs would be $10, whereas for those predicted to not have any infection, the vaccination cost would be $0.

Cost of Treatment:

- For every vaccinated person, there is a 10% chance of infection. Hence the predicted cost of treatment for every vaccinated person would be 0.1*1000=$100.
- For every predicted to get infected, but is actually infected i.e FN, the cost of treatment would be $1000.



Provost's Quizinoma (10 points)

Two of your data scientists A and B are working 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 structuring their work as a competition: both data scientists 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 scientist A reports 95% percent correctly classified instances from her model, while data scientist B reports only 86% percent correctly classified instances from his model. Describe carefully **how you would determine which algorithm is preferable**?

ANSWER:

The first means of determining which model is the best would be the analyzing of ROC curve. A good model should have a very high true positive rate vs positive rate, on the unseen data set. This gives us how well a model distinguishes between the positives and the negatives.

This brings us to the second metric for evaluation which is the Area under the ROC curve (AUC), Higher Area Under the ROC Curve denotes a higher model accuracy. Hence one should look for both True positive rate in the ROC curve and the Area under the ROC curve to compare between models.

We can also assess the misclassification of the model, that is estimate the precision and recall of the model, because a lot of times accuracy is skewed by the majority category present in the data. That is if there are 1000 data-points available, out of which 100 are positive and 900 are negative, then predicting everything to be negative would still have a 90% accuracy. Hence, its better to look at accuracy along with the precision and recall.

Problem Analysis (30 points)

Alumni Donation (15 points)

The UIC Graduate School of Business has a large alumni base, but only recently has been working to engage them in a lifelong relationship with the school. It is not a good idea to bombard alumni with every possible opportunity. There are various different sorts of opportunities and engagements, and the school wants to match them with the alumni for whom they seem to be best aligned. There are many positive advantages to such relationships; right now the school is interested specifically in increasing alumni giving. The school administration has learned that you studied data science for business analytics, and has asked you to help them assess a proposal from Blue Moon Consulting, to help increase alumni giving. As a trial, Blue Moon has been asked to consider one fundraising engagement: the Undergraduate Scholarship Drive (the USD).

Assess the proposal and provide constructive criticism: identify what you assess to be the three most important potential flaws and suggest ways to fix each of them.

> We will mine the data from the prior USD campaign that was delivered to a random sample of 10,000 alumni. We propose to build a model to predict how much each alumnus will give, and then target those who will give the most. The school has collected various data points on each alumnus, including demographic, geographic, major, year, interest, and first-job data, and stored it in the Alumni Database. We will use the amount donated as the target variable, and the data from the Alumni Database as the features. We will build a classification tree and a logistic regression from the data of the random campaign to estimate the amount donated. We will compare the models built based on the area under the ROC curve. The School administration has told us that they would like to target another 5000 alumni in the next test. The 5000 alumni with the highest area under the ROC curve will be targeted.

ANSWER:

Potential Flaw 1: Trying to predict the amount of donation directly, follows the presumption that each of the alumnus will donate some amount, which is not right. Logistic regression is also a kind of classification modelling which cannot be used to predict the amount of donations.

Solution 1: We need to first predict if a person would donate or not using a classification modelling, and then predict the amount of donation for the predicted donors. Logistic regression is also a kind of classification modelling technique, which cannot be used to predict the amount of donation by the alumni. We can apply classification trees or logistic regression to predict if an alumnus will donate or not (1 or 0) and then apply a linear regression model to predict the amount of donation.

Potential Flaw 2: ROC cannot be the sole evaluation technique to evaluate the model.

Solution 2: ROC curve should be used alongside of LIFT curve analysis to evaluate the accuracy of the model. Area under the ROC Curve would only give the accuracy of the classification model, and not the regression model, which could be measured by calculating the value of adjusted R 2 , and root mean square error.

Potential Flaw 3: Area under the curve cannot be used to pick the top donors, it is only a model evaluation parameter.

Solution 3: The area under the curve is not a right measure to pick the donors, instead cumulative LIFT curve should be used to select the data points, by sorting the probabilities of donation in descending order and selecting the top n decile of alumni to send the promotions.

Potential Flaw 4: The feature set could have been a bit more elaborate.

Solution 4: The school should have collected a few more data points about each alumnus, like the company where they currently work, current Salary, Domain of work, Personal Interests, Other Donation works, etc. This could have helped the model to be more accurate.

Facebook Targeting (15 points)

UNICEF is leading a project that aims to collect enough signatures on a petition on the U.S. Government website by the end of May, in order to petition the White House for additional support for early education of disadvantaged children in eastern Africa. An important factor in the success of past campaigns was the participation of younger people, who are notoriously hard to reach via traditional channels (mail, phone). However, it is clear that young people are interested in this cause, as they tend to "Like" the petition on Facebook. Roughly 500,000 UNICEF donors have elected to allow the UNICEF Facebook application to share content on their behalf, and choose which of their 50 million unique friends that the UNICEF content will reach.

You are in charge of building an analytics solution to prioritize which of the 50 million friends will be targeted with a "sign the petition" message. Selective targeting is important because over-sharing leads Facebook algorithms to de-prioritize the content in the friends' streams, making it less likely to be seen. Explain your design by answering the following questions. One or two sentences will suffice to answer each question.

ANSWER

a) **What precisely is your problem formulation? What general category of data mining task does this correspond to?**

The problem is to identify to which 50 million unique friends of the 500,000 million UNICEF donors the UNICEF content will reach. The general category of the data mining task is an unsupervised learning problem.

b) **What features will you use? Describe a few elements of your feature vector precisely.**

The features that can be used for this problem are:

- **Demographics:** Select an audience based on age, gender, education, relationship status, job title and more.

- **Location:** Whether next door or across the world, reach people in the cities, communities and countries where you want to do business.

- **Interests:** Choose the interests and hobbies of the people you want your ad to reach, from organic food to action movies.

- **Behaviors:** Select people based on their prior purchase behaviors, device usage and other activities. For example, if you're a shoe shop you can target people who've recently purchased shoes.

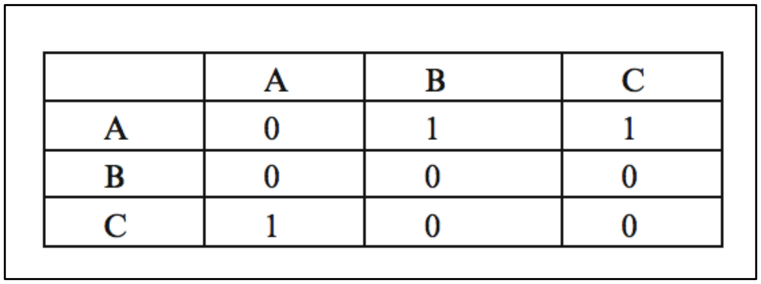
- **Connections:** Reach people who are connected to your Facebook Page, app or event, or exclude them to find new audiences. For example, if you want new Page likes, you can exclude people who already like your Page.

c) What method do you propose to use? Why?

Network Analysis collaborated with k-cluster analysis & Tabu search can be used to tackle this target marketing problem. A network consisting of the roughly selected 500,000 donors along with their friends can form the nodes of the network while the connection between two people on Facebook can form the edge. The feature vectors can become node and edge properties. Property Maps, a network graph with attributes can be used for this purpose. The following questions need to be taken into consideration while building the network:

- who knows whom and how well do they know each other?
- how does the information or resources flow within a network?
- how members of a network know each other?

Although these are not the only factors to consider they provide a good starting point. The nodes and edges and the link between two nodes is represented by a matrix where O's represent no connection and 1's signify a link. The below figure illustrates the same. It is a binary matrix. A signed matrix assigns a nodes relationship with another with +1(positive relationship), 0(no relationship) and -1 (negative relationship) However, an ordinal matrix allows the researcher to assign a strength rating to each connection. In other words, if a strong connection exists, instead of placing a 1 in the matrix, a number would be placed which would signify the strength like 5 or a 10.



	A	B	C
A	0	1	1
B	0	0	0
C	1	0	0

Once the network graph has been made, various clustering and community detection algorithms can be run on the graph to identify the people who are closely connected to each other and people who are sparsely connected. Local bridges connecting the sparsely connected communities can be identified for the purpose of viral marketing. Shortest paths between two people in the network can be observed for the purpose of quick propagation of the shared content. The clusters with highest modularity need to be targeted to identify the 50 million unique friends. Using Social network clustering analysis, the network nodes can be divided into classes based on their links as well as their attributes. Tabu search is a numerical method for finding the best division of nodes into a given number of partitions on the basis of approximate automorphic equivalence. Having selected a number of partitions, it is useful to re-run the algorithm a number of times to insure that a global, rather than a local minimum has been found. The method begins by randomly allocating nodes to partitions. A measure of badness of fit is constructed by calculating the sums of squares for each row and each column within each block, and calculating the variance of these sums of squares. These variances are then summed across the blocks to construct a measure of badness of fit.
Search continues to find an allocation of actors to partitions that minimizes this badness of fit statistic.

d) How will you evaluate whether your model has captured any useful knowledge? What metric would you propose to employ?

The following measures (Metrics) can be used to evaluate the model explained above in order to determine whether it has captured useful knowledge or not:

- **Centrality:** This measure gives a rough indication of the social power of a node based on how well they “connect” to the network. “Betweenness”, “Closeness”, and “Degree” are considered to fall under the measures of centrality.

- **Betweenness:** It is defined as the extent to which a node lies between other nodes in the network. Here, the connectivity of the node’s neighbours is taken into account in order to provide a higher value for nodes which bridge clusters. This metrics reflects the number of people who are connecting indirectly through direct links.

- **Closeness:** This refers to the degree with which an individual is nearer to all others in a network either directly or indirectly. Further, it reflects the ability to access information through the “grapevine” of network members. In this way, the closeness is considered to be the inverse of the sum of the shortest distance (sometimes called as geodesic distance) between each individual and all other available in the network.

- **Degree:** It is the count of the number of ties to other actors in the network.

- **Clustering coefficient:** This provides the likelihood that two associates of a node are associates with themselves. A higher clustering coefficient indicates a greater “cliquishness”.

- **Centralization:** It is calculated as the ratio between the numbers of links for each node divided by maximum possible sum of differences. While a centralized network will have many of its links dispersed around one or a few nodes, the decentralized network is one in which there is little variation between the number of links each node possesses.

- **Density:** It is the degree that measures the respondent’s ties to know one another. The density may be sparse or dense network depends upon the proportion of ties in a network relative to the total number of possibilities.

- **Modularity:** Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules.

- **Assortativity Co-efficient:** This statistic measures the ability of higher degree nodes connecting with other higher degree nodes.