Data Science for Online Customer Analytics
UIC Fall 2017
Assignment I

Name: Laavanya Ganesh
UIN: 654324917
e-mail: lganes2@uic.edu

Multiple Choices / Matching (10 points)

Select one answer for each of the following questions:

1. The points on a model's ROC curve (2 points)
    a. represent the performance of different thresholds
    b. represent different rankings of examples
    c. represent the cost of different classifications

2. You have built several predictive models 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? (3 points)
    a. holdout testing
    b. calculate area under the ROC curve
    c. calculate percent of instances correctly classified
    d. cross-validation
    e. domain knowledge validation

3. Which of these organizations would have the most challenge in applying supervised predictive modeling? (3 points)
    a. A grocery store that is trying to identify which of its loyalty-card-carrying customers will spend more than $100 next month
    b. 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
    c. A city government that is trying to predict which neighborhoods will see the most new business 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

Match tasks with the appropriate task type (2 points):

Task Type	Example Task
_____Classification task _____Scoring/ranking task _____Regression task _____Unsupervised learning	a) Are there any interesting natural groupings of my customers? ----- unsupervised b) Which 500 customers should I target with a special offer? ---- scoring/ranking c) Which customers will leave within 90 days of their current contract expiration? ----- classification d) How many cell phone minutes will each customer use next month? ----- regression

Short Answer (10 points)

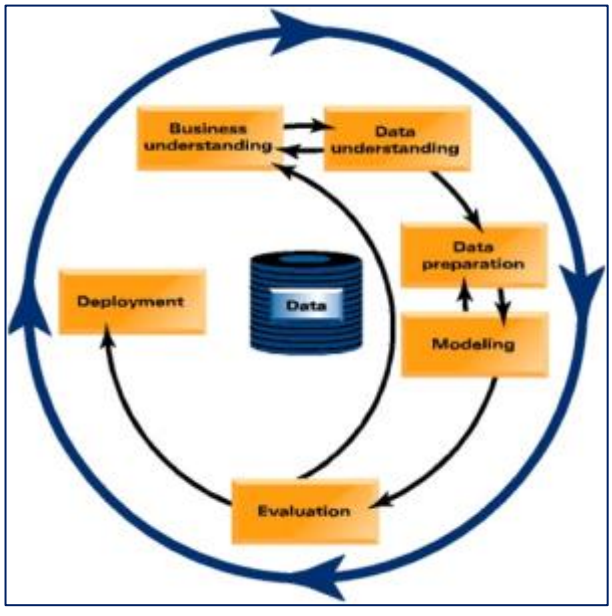
Plumbing Inc (5 points)

Plumbing Inc. has been selling plumbing supplies for the last 20 years. The owner, Joe, decides that next year it is time to diversify by adding gardening tools to the products. Having had success using customer data to build predictive models to guide direct mail campaigns for special plumbing offers, he considers that data mining could help him to identify a subset of customers who should be good prospects for his new set of products. Is Joe ready to solve this a supervised learning problem? (Write a few sentences to explain your answer.)

ANSWER:

In my opinion, Joe is not ready to solve this as a supervised learning problem since he has to address several problems before he thinks of solving it in the supervised learning way. Even with prior success using customer data to build predictive models to guide direct mail campaigns for special plumbing offers, he should not be using the same model for the new set of products which are gardening tools. One major menace in this dealing would be that the plumbing supplies data used to train the model does not bear resemblance to the garden supplies data that is needed for prediction.

The workaround for this loophole is that Joe has to redo the entire Data Mining Cycle for the garden supplies data.



The solution involves the following steps:

- **Business understanding:** Define the target variable which I would suggest could be whether or not a person would be a possible customer for Joe's new products i.e gardening supplies. The possible values this target variable can take are 1 and 0.
- **Data understanding:** Joe needs to understand and analyze what data he currently has in hand, what data he needs, what data he has to fetch and from where could he retrieve the data required to predict the target variable he defined. He would require historic purchases of gardening supplies and demographics of people who purchased them in the past. First and foremost, source of this data

would be survey data possessed by various third party sources featuring in marketing, sales and advertising. Consolidating and integrating various survey data may help building a robust model covering a lot of customer behavior attributes. Data for gardening supplies can be gathered in a variety of ways like focusing on customer purchasing behavior, conducting market basket analysis, tracking likes/dislikes of customers, focusing on past orders placed for gardening tools etc.

The further steps of the Data Mining cycle that are to be carried out by Joe may be similar to his prior campaigns for plumbing supplies. An overview of the further steps are as follows:

- **Data Preparation:** Involves a lot of data pre-processing like imputing missing values; data transformations like converting to categorical or numeric data; removing data on the basis of domain knowledge. Acquiring data from various sources would increase our information base.
- **Data Modeling:** Involves splitting the available data into training and validation data to improve its performance on unseen test data; Fitting different models suitable to training and test data and obtaining accuracies to evaluate model performance.
- **Data Evaluation:** Involves measuring the model effectiveness by comparing the actual results with the predicted results through various model performance metrics like precision, recall, ROC-AUC curve, lift analysis etc.

The best model out of all could be then used to predict the possible customers. By following the given approach, the problem could be solved easily.

Netflix (5 points)

Assume that you work for the data science team of Netflix. Assume that Netflix pays fees in royalties' every time a user requests to watch a movie/series. Even if a user has watched the first five minutes of a movie, Netflix has to pay the full fee. In the past two years there has been a worrying increase in the number of customers that play a movie/series on Netflix without actually watching (e.g. they fall asleep). You want to predict which users indulge in this costly (for your company) habit. In one sentence, formulate a useful target variable. In another sentence, describe precisely how you would formulate the feature vector. Finally, briefly describe how can Netflix use your model.

ANSWER

Target Variable: Whether customer has actually watched the movie/series or no (possible values are 1 and 0)

Feature vectors:

- User Watch history
- User Rating History
- Genre of the movie being watched
- Time of the day
- Volume (increase, decrease)
- Seek times (forward, pause, rewind)
- Movie/series duration

The dataset has to be divided into training, validation and test using the best split obtained from k-fold validation. With the target variable in hand, a series of supervised learning methods can be run (For example, Gradient Boosted, Decision Trees, Random Forests etc.) on the training, validation and test data. The model with the best accuracy can be used to accurately predict whether the customer has actually watched a movie/series. With the predictions acquired, recommender systems can be built using the feature vector to give Netflix the movies/series that are least popular. Using this result, Netflix can increase its revenue by placing the movies/series that are least watched at the bottom of the Home UI. Removing the movies that are least watched and adding the movies from the genre and language that are more often watched, may prove to be a better option for increasing revenue.

Problem Analysis (20 points)

Mail Marketing (10 points)

Last month your boss sent mailing to 20,000 of your existing customers with a special offer on a Hoosfoos Credeen (some cool product). 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, so you decide to build predictive models for targeting. Describe how to evaluate them as follows:

1. What is the cost and benefit of sending a mail?

ANSWER:

- **Cost of sending mails:** 10000$/20000 users mailed = 0.5$ per customer
- **Response:** 1% of 20000 = 200
- **Revenue Per Response:** 200000$/200 = 1000$ on an average

Using this cost of sending a mail and benefit of response we would put weights to the components of confusion matrix (TP, TN,FP,FN), which would enable us to find a model which could generate high profits.

2. What does your model predict (i.e., what's the target variable)?

ANSWER: Response of a customer (1 or 0) i.e. whether a customer will respond to a particular marketing campaign or not.

3. How would you use the data for training and validation?

ANSWER: Cross validation (k-fold concept) would be ideal to come up with the best split for training and validation. Training data will be used to fit the model and validation data will be used to improve the performance of the chosen model on unseen test data. Training data will be used to build the model and test data will be used as unseen data to calculate the model performance.

4. Describe the evaluation function you will use to compare your models.

ANSWER

ROC, AUC, Confusion Matrix and Lift Curve Analysis can be used as evaluation functions to compare the models.

i. The ROC and AUC curves will give us the True positive vs false positive rates, which should be high.

ii. The confusion matrix gives us the measure of opportunity cost lost and false positive identification penalties of a given model. The following performance metrics can be calculated from the confusion matrix:

- **Accuracy:** the proportion of the total number of predictions that were correct.
- **Positive Predictive Value or Precision:** the proportion of positive cases that were correctly identified.
- **Negative Predictive Value:** the proportion of negative cases that were correctly identified.
- **Sensitivity or Recall:** the proportion of actual positive cases which are correctly identified.
- **Specificity:** the proportion of actual negative cases which are correctly identified.

iii. The lift is the difference between the baseline and model performance. It is the advantage of using the model over random selection, meaning that it describes the increase in the number of responses for every depth in the mail file as opposed to mailing randomly.

These features could be used to effectively differentiate between models based on their performance on the data and the given costs and profits. Using these parameters, we can efficiently differentiate between models and select the best out of them.

GloboBank (10 points)

You're working for one of the world's largest financial institutions (GloboBank). They're building a system to monitor salespeople's electronic communications with the company's customers. The goal is to help reduce bad behavior among the company's salespeople, such as overpromising, understating risks, and so on. The company is unhappy with their current surveillance system, because it creates tons of false alarms, which wastes the time of the analysts, and also they know that it misses a lot of important cases. Below is a proposal they have received from a vendor, for a better system, that will address their issues. Specifically, it will monitor each outgoing email from a salesperson and flag those that are suspicious. The suspicious emails would be examined by an analyst, who would decide which ones ought to be escalated for further investigation.

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 use machine learning techniques to build a model to classify emails as "suspicious" or not. Those classified as suspicious will be "flagged"; our compliance analysis system will rank them and provide the most suspicious ones to the analysts. The system will maximize the lift at the top of the ranking, and minimize the number of missed cases (false negatives).*
>
> *The flagging model will take as input a feature-vector representation of the email, where each word is a feature and the feature value represents whether the word is present in the email; more sophisticated representations will be added later. We will leverage the existing system to provide training labels. Specifically, if the existing system flags the emails as being suspicious, we will give them a label of yes. Otherwise we will give a label of no. As we archive all salesperson emails specifically for compliance purposes, this will allow us almost unlimited training data.*
>
> *We will evaluate the system based on its generalization accuracy and the area under the ROC curve (AUC) on holdout data. The system should be able to achieve accuracies greater than 90% as well as high AUCs. We also will show the flagged emails to compliance experts for domain knowledge validation.*

ANSWER:

Flaw 1: Since, the existing system misses important cases and gives a lot of false alarms, leveraging the existing system to provide training labels is not a good idea.

Fix 1: Instead, data from past breaches or malicious e-mails can be taken into consideration for deciding the attributes required to make the dataset on which the model will be run.

Flaw 2: Since, the existing system gives a lot of false alarms, we do not know accurately whether the e-mail is suspicious or not. The existing system does not randomly sample out the e-mails. it's not clear that they will be useful data points from which to build a model to apply.

Fix 2: Conduct a domain validation to gather data to build the machine learning models.

Flaw 3: Each word is used as a feature vector which overlooks compound words like e-mail, website, web page etc.

Fix 3: n-grams (bi-grams and tri-grams) can be used apart from uni-grams. In other words, phrases can also be used to identify the suspicious emails. These phrases require machine learning algorithms like word2vec or word2phrase to be identified. Once identified they can be again treated like a token and their tf- idf values in the emails could be used to identify suspicious e-mails.

Flaw 4: Not all words will be present in the dictionary

Fix 4: In more sophisticated representations that will be added later, the following ideas

can be incorporated:

- Latent Semantic Indexing, which consists of applying Singular Value Decomposition to the DocumentXTerm matrix in order to identify relevant (concept) components, or in other words, aims to group words into classes that represent concepts or semantic fields.
- Using a lexical database like WordNet or BabelNet concepts in order to index the documents, allowing semantic-level comparison of documents.

Flaw 5: Presence of a word in an email might not be sufficient to make it suspicious.

Fix 5: Use Natural Language Processing to find out the context in which the word is being used and based of tf-idf values of certain words in the emails classify them as suspicious.

Flaw 6: We don't need almost infinite corpus of documents to acquire enough data to identify suspicious emails.

Fix 6: This could be done just by hand-picking a few suspicious emails and analyzing the text in these emails using Natural Language Processing tools, to come up with term and document frequency of certain words and use this information to identify other similar emails. This would reduce the size of data to be processed and increase the efficiency.