|  |  |
| --- | --- |
|  | **Data Mining for Biz Analytics**  **info-ub.57**  **Spring 2019** |
| **Homework #2** | **Due: turned in on NYU Classes by 11:59pm Mon 4/3/2019** |

Ashley Tsoi

(put your name above)

ast418

(put your Net ID above)

Total grade: \_\_92\_\_ out of \_\_100\_\_ points

**Part 1:** 15 **/ 15**

**Part 2:** 25 **/ 25**

**Part 3:** 28 **/ 30**

**Part 4:** 15 **/ 15**

**Part 5:** 9 **/ 15**

***Please answer all questions/follow all directions. Put your name above, and include your last name in the filename of your homework submission. Include all of the requested material in a compressed file (.zip, .tar.gz, .rar, etc)***

**1) Label each case as describing either data mining process (DM), or the use of the results of data mining (Use). To help you answer this question, revisit the definition of data mining.**

a) Use --- Choose customers who are most likely to respond to an on-line ad.

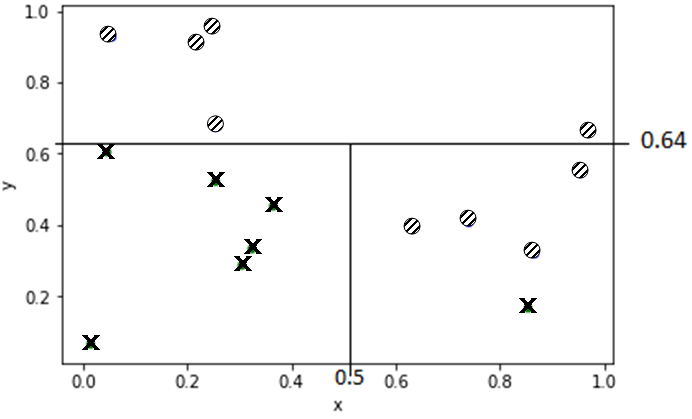
b) DM --- Discover rules that indicate when an account has been defrauded.

c) DM --- Find patterns indicating what customer behavior is more likely to lead to response to an on-line ad.

d) Use --- Estimate probability of default for a new credit application.

e) Use --- Predict whether a customer is pregnant.

**2) You are given a 2D plot, with instances that belong to 2 classes: circles and x’s. After learning a Decision Tree classifier on that dataset, you get the decision boundaries shown in the image below.**



**a) Present the resulting Decision Tree model. You can show it as a tree structure *or* as the rules that you’ll get out of them.**

**For the tree structure, include an image (here or w/ your submission) that you drew by hand, via MS / Mac paint, or any other tool you like. For each node in your tree, include the split condition (if any). Report the class value only in the leaf nodes. You do not need to report other values for any of the nodes.**

***Alternatively*, write the *rules* that correspond to the model. For each rule, the conditions need to be in the order as you would follow them in the tree model, i.e., the 1st condition you write must be at the root, the 2nd condition must be at layer 1, and so on.**



**b) Briefly (2-3 sentences) explain why the first feature was selected.**

The first feature (whether y is greater or less than 0.64) because it gives us the most information gain.

**c) What is the probability that a new instance with (x,y) = (0.7, 0.2) belongs to class “x” ? What is the probability for the same instance to belong to class “x” with Laplace correction?**

without Laplace: n/(n+m) = 1/5

with Laplace: (n+1)/(n+m+2) = 2/7

**3) Soc-Marketeer Inc. is a social media marketing agency that can help you “place your ad-content in front of 100K accounts” on a popular social network. They have a varied clientele who deal in *sports, fashion, tourism, entertainment, education, domestic retail, technology* and other domains. They charge (effectively) a flat rate for their 100K outreach, but you’ve heard that their customers very often see a large discrepancy between the advertised (100K) and real traffic that they receive from the social network, leaving them thinking they’ve overpaid.**

***Assume that these 100K are all real people (no bots) with verified accounts (no duplicates). Soc-Marketeer really has access to them, and all users see all ads. The users are knowingly targeted with ads and all legal paperwork is in order. Focus on data science-related problems.***

**You are confident you can (help them) do better! In half to one page (max), using concepts we’ve introduced in class – in discussions and Jupyter Modules - and the corresponding chapters in the textbook, write a brief, *yet precise*, business proposal. Your answer should (at least) include the following:**

1. Explain what the root cause of the traffic discrepancy is.
2. What can be done to reduce that discrepancy?
3. What data science *task* would you use to reduce the discrepancy?
4. What information would you need for your task?
5. Who would provide the data that you’d need for your task? Should you involve others?
6. Present ***at least 5*** different features that you think i) you can get and ii) will help you to address the problem. Be (very!) precise (e.g. “frequency of” is not precise).
7. If the task is supervised, give a (precise) definition of your target variable.
8. Give an example of an instance of your data.
9. How would you evaluate your suggestion?

The root cause of the traffic discrepancy may lie in Soc-Marketeer’s ability / accuracy of its client-targeting model. To reduce that discrepancy, Soc-Marketeer should better learn its models with which it uses to target the social media accounts. There are many possible explanations to why Soc-Marketeer does not have an accurate enough model. It may be due to a lack of training data, over-fitting which causes the models to “memorize” rather than “generalize”, failure to identify the target variable, etc. The key is to find the “sweet spot” between the accuracy of their model and its complexity.

The data science task would be building the model using cross-validation, tree-induction, or logistic regression. To use the tree induction method, for instance, our target variable would be if the customer would click on the link provided with the ad, with Boolean results. To estimate whether a user will respond to our ads, we need information datasets from the social networks and other data collection service companies to get more data.

The features Soc-Marketeer can use to better identify the target variable include: average number of hours spent on social networks each week, age of the user, average number of posts posted on social networks every week, educational background of the user, gender of the user, geographic background of the user, marital status, etc. On top of the boolean outcome, we can also estimate the probability of each features happening, and the probability of the user to respond using this method. For instance, one data can be a 20-year-old college female who spends an average 20 hours on social networks per week and posts an average 3 posts on social networks every week.

To evaluate this model, we can use nested cross-validation or cross-validation to test the model. We can split the datasets into different folds, let’s say 10, each consists 10% of the whole datasets. We can use the first dataset as the holdout dataset and the other nine as the training set and get one model, then we iterate them and test 9 more times. The accuracy is the average accuracy of the ten models. In this way use the original labeled data efficiently to estimate the performance of a modeling procedure.

**4) Due to time pressure, your firm is thinking about retaining the services of a consultant for your personalized advertising campaign, instead of building it in-house. You are asked to attend the meeting given your data science expertise, to help with the screening process.**

**During their presentation, they mention that they achieved an accuracy score of *91%* on a public, well-known dataset that you have worked on extensively. You find this result surprising, because following a rigorous evaluation methodology, you had only achieved accuracy around 74%, using the same modeling techniques that they presented: Decision Trees, Logistic Regression, SVMs.**

**Write below 2-3 different questions that you should ask them, to assess whether the reported 91% is truly reflective of their model’s effectiveness. Your questions must be specific, in the sense that their respective answer should be short (few words, “yes” / “no” answers). For example, asking “*what was your evaluation methodology?*” (or similar) is not a vaswlid question.**

1. Has any methods used to recognize overfitting, i.e. by testing with a holdout set by using

2. If yes, what methods? (e.g. Cross-validation,

3.

**5) Hands-on section: Log in to** [**JupyterHub**](https://jupyterhub.stern.nyu.edu/) **and start your server, if needed. Using the Jupyter Notebooks’ controls, navigate to your class material folder ( /notes/gv760/data\_mining\_spring2019/ ). In there, you should find a folder called “assignments” containing a Python notebook for the hands-on part. Follow the instructions in the notebook. Once you’re done, download the notebook and include the notebook – and any other needed code – in your submission file.**

**Reminder: *You will need to be connected to an NYU network to access JupyterHub, either by being on campus and connecting directly to one of NYU’s networks, or via a VPN connection. Instructions for setting up a VPN conection, depending on your platform:***

**For Mac Users:** [link](https://nyu.service-now.com/servicelink/search_results.do?sysparm_document_key=kb_knowledge,a6be768b1c8dd504bbcf4dc2835ec355&sysparm_search=vpn)

**For Windows Users:** [link](https://nyu.service-now.com/servicelink/search_results.do?sysparm_document_key=kb_knowledge,6177d7031c811904bbcf4dc2835ec340&sysparm_search=vpn)