EMORY UNIVERSITY

Machine Learning II

Homework #4 **Due: turned in by Monday 04/20/2020**

(put your name above)

Total grade: _____ out of ___100___ points

Please complete the following task and submit your assignment on https://inclass.kaggle.com/c/emory-isom-676-hw4.

Task description: Website XYZ, which is a music-listening social networking website, follows the "freemium" business model. The website offers basic services for free, and provides a number of additional premium capabilities for a monthly subscription fee. We are interested in predicting which people would be likely to convert from free users to premium subscribers in the next 6-month period, if they are targeted by our promotional campaign. You have a dataset (provided on the course Moodle site) from the previous marketing campaign which targeted a number of non-subscribers.

Specifically, the training dataset contains approx. 86,700 records, each record representing a different user of the XYZ website who was targeted in the previous marketing campaign. Each record is described with 25 attributes. Here is a brief description of the attributes (attribute name/type/explanation):

- *adopter* / binominal (0 or 1) / whether a user became a subscriber within the 6-month period after the marketing campaign (this is an outcome variable!)
- *user_id* / integer / unique user id
- *age* / integer / age in years
- *male* / integer (0 or 1) / 1 – male, 0 – female
- *friend_cnt* / integer / numbers of friends that the current user has
- *avg_friend_age* / real / average age of friends (in years)
- *avg_friend_male* / real (between 0 and 1) / percentage of males among friends
- *friend_country_cnt* / integer / number of different countries among friends of the current user
- *subscriber_friend_cnt* / integer / number of friends who are subscribers of the premium service
- *songsListened* / integer / total number of tracks this user listened (or reported as listened)
- *lovedTracks* / integer / total number of different songs that the user "liked"
- *posts* / integer / number of forum or discussion board posts made by the user
- *playlists* / integer / number of playlists created by the user
- *shouts* / integer / number of wall posts received by the user
- *good_country* / integer (0 or 1) / country type of the user: 0 – countries where free usage is more limited, 1 – less limited.
- *tenure* / integer / number of months since the user has registered on the website.
- There are also a number of attributes with the following names: *delta_*<attr-name>, where <attr-name> is one of the attributes mentioned in the above list. Such attributes refer not to the overall number, but the change to the corresponding number over the 3- month period before the marketing campaign. For example, consider attribute *delta_friend_cnt*. If, for some user, *friend_cnt* = 50, and *delta_friend_cnt* = –5, it means that the user had 50 friends at the time of the previous marketing campaign, but this number reduced by 5 during the 3 months before the campaign (i.e., user had 55 friends 3 months ago).

Task: The general overall task is to build the best predictive model for the next marketing campaign, i.e., for predicting likely adopters (that is, which current non-subscribers are likely to respond to the marketing campaign and sign up for the premium service within 6 months after the campaign). Feel free to use any techniques, methodologies, and approaches that you learned in the class.

Model performance assessment: You will also be provided with the scoring dataset of different 86,700 users with the same attributes as described above, except this dataset does not have the outcome labels. No more than once per calendar day, you can use your current best model to predict outcomes in this dataset and submit the predicted data to Kaggle (***https://inclass.kaggle.com/c/emory-isom-676-hw4***), which will automatically score your prediction against actual outcome labels and will post your result on the group project leaderboard. Submissions are limited to only few per day – this may be an incentive to start working on the project early and not wait to try things out until the last days of the semester. The scoring metric used for this assignment is the F-measure for the "adopter" class. The leaderboard will be located on the Kaggle webpage for this assignment. You can use the provided RapidMiner process (or any other software you prefer) in order to create a submission file with the same format as the provided sample.

Evaluation: Evaluation is based on the final performance achieved by your best reported model(s) at the end of the competition.

Finally, you should also submit your final code/solution here on Canvas (i.e., the code/solution with which you achieved your highest score); *this is an individual assignment and your code/solution should be your own work.*