K579 Team Project: Predictive Modeling Competition

Task Description

Website XYZ, 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 Canvas site) from the previous marketing campaign which targeted a number of non-subscribers.

Specifically, the **labeled dataset** contains 58,077 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 the outcome variable!**)
- *user_id* / integer / unique user id (*obviously, this is not a predictive feature, just a unique identifier*)
- *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 market-ing 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).

The general 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 technique, methodology, and approach that you have learned in class or anywhere else (though I won't assist in the implementation of algorithms we haven't learnt in the class). Here are some basic guidelines:

- Before sending out the predictions, make sure to follow the standard practice in building and evaluating predictive machine learning models, i.e., training-validation split or cross-validation;
- Explore different modeling techniques. For each modeling technique, explore different hyperparameter combinations;
- Consider feature selection;
- Note that this is an imbalanced dataset, i.e., the adopters only account for less than 2% of the population. Therefore, you may also consider sampling techniques to deal with imbalanced data;
- A starter R Script is also provided, which contains a very simple decision tree model that makes all the predictions as zero. However, after using oversampling, the tree model achieved an F-measure of 0.0707 for the "adopter = 1" class. This baseline should be fairly easy to beat.

Model Performance Assessment

To assess the out-of-sample performance of your model, you will also be provided with an **unlabeled dataset** of another set of 28,605 users with the same attributes as described above, except this dataset does not have the out-come labels. **No more than thrice per week**, you can use your current best model to predict the outcomes in this dataset and email the predictions to the instructor. The week restarts at Monday midnight. Your predictions should come in a .csv file named "Team-X-Submission.csv" (replace X with your group number) with two columns: *user_id* and *prediction*, where the ***user_id*** **column must match the user IDs in the unlabeled dataset, and** ***prediction*** **column contains binary (0/1) predictions for each user. Prediction files that do not follow this format will not be scored.**

Upon receiving your group's predictions, I will score your prediction against actual outcome labels and will email you back your model's performance. The best-to-date performance of your team will be continuously updated on the leaderboard.

The scoring metric used for this project is the F-measure for the "adopter = 1" class. Make sure not all the predictions are 0, as the F-measure will be NA. Do not submit if that is the case and try to improve your model. The team predictions will be accepted by the instructor from March 20 (first weekly submission) until **end-of-day on May 1 (completing seven weekly submissions)**.

Evaluation: 25 points

- **Performance**: 20 points. This is based on two aspects: (1) the final performance achieved by your best reported model, and (2) the diversity of techniques, methodologies, and approaches you try (i.e., there will be a penalty if you only tried a few models). You will be required to submit your R scripts (no knitted files are required) in which you attempted/ran all the models.
- **Presentation**: 5 points. On May 2 during class, each team will make a brief presentation (no more than 5 minutes + up to 1 minute for Q&A) about their experience with the project (various modeling explorations performed, what worked, what did not work, etc.) and describe their best performing model.