IE 582 Fall 2021 Project, Due final exam

Objective and Evaluation

Use the data provided with this project on Moodle.

One file is the **training data** that you should use for model building.

A second file contains **testing data** that will be used to evaluate your model.

The training and testing data were selected randomly from the original data set.

The aim is to build a classification model for the given data based on the methods described in the course. The classification task is related to gender prediction for the customers of an online retailer. This is a real world dataset. Provided data consists of actions of the users on the website. The fields in the data is as follows:

- time_stamp: Timestamp of the action
- contentid: Id of the product
- user_action: type of the action (i.e. visit, search, basket and etc.)
- sellingprice: price of the item
- product_name: name of the product
- brand_id: brand id of the product
- brand_name: brand of the product
- businessunit: business unit information for the product
- product_gender: gender of the product if defined
- category_id: caegory id in which product is belonging to.
- Level1_Category_Id: category id in the first hierarchy
- Level1_Category_Name: category name in the first hierarchy
- Level2_Category_Id: category id in the second hierarchy
- Level2_Category_Name: category name in the second hierarchy
- Level3_Category_Id: category id in the third hierarchy
- Level3_Category_Name: category name in the third hierarchy
- gender: this is the class information
- unique_id: id of the user
- type: type of the data (just for information purposes, train and test)

Unfortunately, this is an unstructured data set which consists of actions of the users. Hence, there is a need for feature extraction from the data to obtain a regular data matrix before

classification (i.e. feature engineering). You are expected to use the methods covered in the class and you are free to use extended versions of the approaches covered in the lectures.

Performance measure:

Balanced Error Rate: Your model will be scored as your error rate on class "a" plus your error rate on class "b". This is different from the overall (weighted) error rate commonly reported. Because the classes are unbalanced, you should work to minimize the error rate on each class.

Area under the ROC curve: Although correlated with balanced error rate, area under the ROC curve gives better idea about the performance for binary classification problems.

Your submission will be evaluated based on these two measures and you are expected to maximize both. These measures are going to be equally weighted as your final score.

Prediction and Report

This project is organized as a competition (like in platforms such as www.kaggle.com) in which you are expected to make submission everyday. For this purpose, we have built a system so that you will be able to make submissions via Google Form and monitor your progress through Google Sheets. We will be informing you about the submission system once we finish our tests.

Before **December 27th, 2021**, you will be given your passwords and usernames for access to the system. You will be able to make ten submissions every day and observe your performance on a predefined subset of the test data. The final evaluations will be performed based on your latest submission. You will be kept posted about the deadlines for final submissions (the one that will be used for final evaluation) towards the end of the project period. Note that 30% of your project grade will be determined by your final rank in this competition. First place will get full points (30 points) and this will decrease to a minimum of 15 points proportional to your deviation from the top performer (assuming that you did not miss any day during the second phase).

You are required to submit a written report with a brief description of your final method, how you evaluated your methods, and you choose the parameter settings. You are allowed to work as a group of at most 3 members.

Your report should have the following format:

1. *Introduction:* Problem description, summary of the proposed approach, descriptive analysis of the given data.
2. *Related literature:* Summarize relevant literature if there is any
3. *Approach:* Explain your approach to this problem.
4. *Results:* Provide your results and discussion.
5. *Conclusions and Future Work:* Summarize your findings and comments regarding your approach. What are possible extensions to have a better approach?
6. *Code:* Provide the Github link for your codes at the end of your report.