**Homework 11 - Data Competition**

**Predicting the Usefulness of Online Questions**

Due May 9th, 11:59 PM

***Background:***

The dataset was collected from a community-based open Q&A website for user experience designers and professionals. In the community, users can ask questions related to the design of user interfaces and answer questions posted by other peers. After a user submits a question to the community, other users can vote up or vote down the usefulness of the question. Those questions with the highest net votes (i.e., positive votes – negative votes) are displayed on the top of the question list so that all community users can first view them when looking at the question list. Fig. 1 shows a sample question with usefulness votes.

Graphical user interface, text, application, email

Description automatically generated

**Fig. 1.** A sample question with usefulness votes.

The dataset contains 5000 questions posted from January 2010 to November 2021. The file “train.csv” contains a dataset of 4000 questions that is used to train classification models. Each row is an identical online question. There are 58 columns explained below.

* Id: unique identifier of an online question;
* IsUseful: a binary variable indicating if a question is useful or not (1= useful, 0=not useful);
* 709 Columns “TFIDF\_approach” through “TFIDF\_bmml”: frequency of words/characters in the online question, measured as TFIDF (term frequency–inverse document frequency). For example, “TFIDF\_approach” is the TFIDF score of the term “approach” in the content of the question.

The “test\_data.csv” contains a dataset of 1000 questions used to evaluate the performance of classification models. The test dataset has the same data structure as the training dataset.

The dataset is part of a research project. For more information about the dataset and its background, refer to the attached paper “Chen 2022 LNCS.pdf”. To simplify the predictive modeling, only the TFIDF features are included for the data competition. Other features used in the research project such as LIWC and Doc2voc features are not included.

***Instruction***:

The objective of this data competition is for students to apply various predictive modeling techniques to accurately predict whether an online question is useful or not. Students can either work in a pair of two students to finish the data competition or do the work individually. If you want to find a partner, you can submit a post on our discussion board. Students are also allowed to finish the work individually. Please work as early as possible so that you have sufficient time to refine your final model by the due time.

Student pairs or students should work individually to finish the data competition. Any form of collaboration or help from other pairs/persons/parties is NOT allowed. Please submit your work by the due time.

***Task:***

Apply various predictive modeling techniques to build a classification model that performs very well for predicting if an online question is useful. Below is a list of possible techniques that could help you tune your models:

1. Deal with imbalanced dataset

For more details, refer to <https://www.r-bloggers.com/2019/04/methods-for-dealing-with-imbalanced-data/>.

1. Assign class weights
2. Use regularization parameters such as C in SVMs.
3. Dimension reduction or feature selection
4. Tune hyperparamters
5. Use advanced models such as ensemble methods and deep learning
6. Other methods……

You can only use the training dataset to train the classification models without touching the test dataset. Using test dataset to as a validation set to refine your models is also not allowed. You can normalize/transform the variables in the test dataset, but any resampling applied to the test dataset is not allowed. The test dataset is only used to evaluate the performance of your model. Your models need to be evaluated by AUC score.

Make sure that your R Markdown can be reproducible. Re-running the Markdown will lead to the same results.

***Submit the Following Documents:***

1. Submit your R Markdown document with detailed explanations. Your R Markdown should at least include the following contents:

* Specify your pair information in the R Markdown document. Each pair only needs one submission.
* Specify contribution by each member. Students without significant contribution to the work will not receive the full regular points and extra credit.
* Data transformation and/or preprocessing.
* The final model trained from the training dataset.

***Note:*** I know you need to try multiple models and try many parameters. Please only submit the final model with the best parameters you find from your parameter tuning. Keep those time-consuming model optimization process in another Markdown document and submit it to the homework as a supplemental document. This ensures that your R Markdown with the final model can be re-run in a short time period.

* Test the performance of your final model using the test dataset. At least AUC score needs to be reported.

1. Submit an HTML report generated from your R Markdown for your final model.
2. If you have used hyperparameter tuning, please submit your R Markdown for the hyperparameter tuning as a supplemental document.
3. Submit a one-page Microsoft Word document that explains the logic of your predictive modeling.

***Grading:***

***Regular Points***

All students’ works will be graded based on the AUC score of the final model. Correct work submitted by the due time will earn 40 points as a regular homework. Errors in the submission will reduce the points.

***Extra Credit***

Students who finish the work by the due time can earn extra credit according to the following rule:

* Submissions ranked as top 3 in terms of the AUC score of the final model will receive additional 2% extra credit for final grade;
* Submissions ranked 4th to 5th best models in terms of the AUC score of the final model will receive additional 1% extra credit for final grade.