Predictive Analytics Modeling Exercise

You have been provided with data scattered across 5 files. The dataset consists of 12073 samples in the *learning set* and 5366 samples in the *prediction set*. There are 79 variables drawn from different databases, which is why they are split between 'Group A' and 'Group B' files. The *code* variable is a categorical variable. The *zip* variable contains zip codes. Remaining variables are a mix of continuous and categorical variables. Some features include missing values. A classification response *y* is provided for learning set samples, with the positive class labeled *1* and the negative class labeled *0*. The *id* column is a unique sample identifier and hence not a variable.

No additional data may be added or used.

Most modeling efforts at Nationwide are written in Python or R, so we prefer if you do this exercise using either language, but if you're more comfortable using another language (such as Matlab), please let us know upfront.

Your primary task is to build 2 different machine learning or statistical learning models on learning set samples and generate predictions on prediction set samples. The assessment metric for these models is *area under the receiver operating characteristic curve*.

You may wish to use these 2 models to contrast linear versus nonlinear performance, or models from different model families, or different hyperparameter combinations within the same modeling framework, or any other distinguishing modeling choice that you think is important. The best score of the two models submitted will be used to assess your performance in this modeling exercise.

Your deliverables are:

1. A written report, in PDF format. The filename should take the form *report_FirstnameLastName.pdf*.
2. Your code. We may examine it and/or run it locally. While we understand that, in the limited time assigned for this project, your code may not be deployment-grade, please ensure that it's reasonably readable and commented.
3. For each model, a CSV file with the first column being the samples 'id' and the second column being the predictions. These predicted values should be real numbers (e.g. propensities or probabilities of belonging to the positive class) and not binarized so that we can assess performance. The filenames should take the form *predictions_FirstnameLastname_model#.csv*.

We don't impose a length range for your report. We strive for clarity and concision. The length of your report will depend on the extent of the relevant insights into the data and models that you wish to share. Minimally, it should include the following:

- A description of your modeling process and design choices, including any data preparation, exploratory data analysis, feature manipulations, model tuning and assessment. State any assumptions you made along the way.

- For each model, an estimate of its generalizability (expected performance on previously-unseen data).

Some questions you may wish to address in your report are:

- How did you handle the class imbalance?
- Had a metric not been imposed, what metric(s) would you have considered using for this model? Why?
- What modeling choices were impacted by scale? Put otherwise, if the learning set's sample size had been (i) 20 times greater or (ii) 20 times smaller, briefly describe high-level changes in approach you might have considered.
- If data augmentation had been permitted for this project, what data might you have considered incorporating?
- Given another week to work on this modeling exercise, what would you want to try next towards better understanding this dataset and improving predictive performance?
- Compare and contrast both submitted models with regards to strengths, weaknesses, and performance.