PREDICT RESPONSE TO MAILOUT CAMPAIGN

I. Domain Background

Predicting campaign response accurately will help companies target the most potential customers while saving marketing expense. To put it in numerical terms, if your overall response rate is 5% but you were able to predict the 10% most potential customers with a response rate of 80%, your return on investment would increase by 8 times compared to targeting everyone. Furthermore, targeting the wrong customer might backfire, making the product less appealing[1].

Targeting the right customers can be done with a supervised learning model, learning from the past responses/non-responses. This project aims to do that for a marketing campaign from a client of Arvato Financial Services. This type of modeling often face a big challenge of imbalanced class problem[2] as the response rate is very low. Hence, the models will bias towards predicting non-response since a naive guess of everything being the majority class would be a pretty good guess. Tackling imbalanced class problem could be used using a combination of sampling methods and boosting models[3].

II. Problem Statement

We will predict who will respond to a mailout campaign. Our model will be evaluated using area under the receiver operating characteristics curve (AUC) as this is the metrics for the Kaggle competition. We will also discuss the model's performance with respect to other metrics such as sensitivity, recall, and area under precision-recall curve so that one can evaluate the model with different objectives.

This is a binary classification problem with highly imbalanced class, with only about 1.23% of our customers in the data responds positively to the campaign.

III. Datasets and Inputs

The dataset is provided by Bertelsmann Arvato Analytics. The datasets include demographics data of Germany's general population and customers of a mail-order sales company. It also includes 2 datasets of targets of the company's marketing campaign, with one being the training dataset with an additional column indicating if the target responds to the campaign.

IV. Solution Statement

The goal is to optimize AUC, which measures the ability of the model to rank a random true positive higher than a random true negative.

Different sampling techniques and models will be employed.

V. Benchmark Model

The benchmark model is a Gradient Boosting machine with class weight balancing. I chose this algorithm because it has shown effectiveness in many problems and it doesn't require missing value handling. Since missing values might have different context and therefore different

meaning, I want to see how later I could beat the benchmark with no missing value handling. The 10-time-repeated 5-fold validation AUC is 76.4%

VI. Evaluation Metrics

Since the data is imbalanced, we should not focus on accuracy because of the model's inclination towards the majority class. Instead, we should focus on the model's ability to rank a random true positive as more probable than a random negative, which is what AUC represents. Therefore, the evaluation metric is AUC but other metrics will be reported.

VII. Project Design

The project will involve the following steps:

- *Run a baseline model with XGBoost with no tuning parameters and dataset as-is.*
  XGBoosts has proved itself in many problems so I want to see how much my extra effort in the following step will pay off compared to a default option
- *Handle missing data*
  Many algorithms requires non-missing data so this needs to be handle.
- *Feature engineering*
  The training dataset only includes nearly 43000 data points with over 300 features. Using all features would likely result in overfitting. In addition, some features needs preprocessing and new features could be created to increase predictive power for the model. We also run a sanity check on the dataset here
- *Model tuning*
  Define validation procedure to evaluate each modeling experiment. Each experiment concerns with algorithm family, hyperparameter-tuning, and model ensembling.
- *Compete on Kaggle*

References

1. *Response Modeling*, Big Data Analytics
   http://www.bigdatanalysis.com/response-modeling/
2. C.X. Ling, C. Li, "Data Mining for Direct Marketing—Specific Problems and Solutions", Proc. Fourth Int'l Conf. Knowledge Discovery and Data Mining (KDD '98), pp. 73-79, 1999
3. Sun, Yanmin, Andrew KC Wong, and Mohamed S. Kamel. "Classification of imbalanced data: A review." International Journal of Pattern Recognition and Artificial Intelligence 23.04 (2009): 687-719.