

COMP/INDR 421/521 ELEC 443/543
HW07: Modeling Outbound Reach Rate of a Call Center
Deadline: January 7, 2019, 11:59 PM

In this homework, you will develop a machine learning solution in R, Matlab, or Python for a real-life binary classification problem from finance industry. Your machine learning algorithm needs to predict whether a customer will answer the phone call initiated by the outbound call center of a bank using the information given about each particular customer and call time. Here are the steps you need to follow:

1. You are given three input data files, namely, `training_data.csv`, `training_labels.csv`, and `test_data.csv`, in a single zip file, which you can download from the Assignments section of course web site in Blackboard. The training set contains 300,000 labeled data instances, where each training data point has 142 features. The categorical features (F03, F05, F06, and F63) were converted into binary features using one-hot encoding. For example, the fifth feature F05 has three levels (i.e., A, B, and C), and these levels are converted into three binary features (i.e., F05_A, F05_B, and F05_C). You are also given a very simple solution strategy using a decision tree classifier in the file named `quick_and_dirty_solution.R`.
2. Develop your own machine learning solution for this problem. You are free to use any publicly available packages in R, Matlab, or Python. If you need to pick some parameters for your machine learning algorithm, you should perform a cross-validation strategy on the training data set. The predictive quality of your solution will be evaluated in terms of its AUROC value on the test set.
3. Use the trained classifier from the previous step to perform predictions for the test data set, which contains 186,226 data points. You are not given the class labels of these instances. You need to generate either probability or score estimates for the positive class and to write these estimates into a file. For example, the decision tree strategy implemented in `quick_and_dirty_solution.R` file generates probability estimates for the test set and writes these values into a file named `test_predictions.csv`.

What to submit: You need to submit your source code in a single file (.R file if you are using R, .m file if you are using Matlab, or .py file if you are using Python), the estimated probabilities/scores that you calculated for the test set (`test_predictions.csv`), and a detailed report explaining your approach (.doc, .docx, or .pdf file). You will put these three files in a single zip file named as ***STUDENTID.zip***, where ***STUDENTID*** should be replaced with your 7-digit student number.

How to submit: E-mail the zip file you created to both arahimi14@ku.edu.tr and jsayilgan14@ku.edu.tr with the subject line ***Intro2MachineLearningHW07***. Please follow the exact style mentioned for the subject line and do not send a zip file named as ***STUDENTID.zip***. Submissions that do not follow these guidelines will not be graded.

Late submission policy: Late submissions will not be graded.

Cheating policy: Very similar submissions will not be graded.