7390–Machine Learning and Data Sciences

Northeastern University, Fall 2017

Final, Saturday Dec 16, 2017

- The exam is open book.
- You are given 90 minutes for this exam.
- **Show your work.** Partial credit will be given.

Name: ____

Problem Description: We are given a dataset containing information on ten thousand customers. We want to predict which customers will default on their credit card debt.

The dataset consists of 10000 records including the following fields:

- **default**: A factor with levels No and Yes indicating whether the customer defaulted on their debt
- **student**: A factor with levels No and Yes indicating whether the customer is a student
- **balance**: The average balance that the customer has remaining on their credit card after making their monthly payment
- **income**: Income of customer

Questions:

1. Use logistic regression to analyze this data.
   (a) Use 100 bootstrap samples to estimate the the standard errors of the coefficients from the logistic regression fit AUC. Provide a 68% and 95% confidence interval for the model parameters of the logistic regression.
   (b) Calculate p-values for each feature. Do any of the features appear to be statistically significant? If so, which ones?
   (c) Use 5-fold cross-validation to estimate test AUC. What test AUC do you obtain? Provide a 68% and 95% confidence interval for test AUC.
   (d) Use 5-fold cross-validation and *Forward Stepwise Selection* approach for feature selection. In each iteration, choose the best model using the highest AUC. Report the features and AUC that appears to provide the best results.
2. Use random forests to analyze this data.
   (a) Use 100 trees for a range of values of max features consider. Describe the effect of the number of features considered at each split on the test AUC. What is the optimal value for max features?
   (b) What are most important features?
   (c) What test AUC do you obtain using 5-fold cross-validation?
3. Use boosting to analyze this data.
   (a) Perform boosting with 100 trees for a range of values of the shrinkage parameter λ. Produce a plot with different shrinkage values on the x-axis and the corresponding 5-fold cross-validation AUC on the y-axis. What is the optimal value for λ?
   (b) What are most important features?
   (c) What test AUC do you obtain using 5-fold cross-validation?
4. Compare the results between logistic regression, random forests, and boosting.
   (a) Plot ROC curves for the best logistic regression model, the best random forest model, and the best boosting model.
   (b) Which of these approaches yields the best performance?
5. Submit a Python or R script to show your work and evaluate the results on a test dataset.
   (a) For a given test dataset, it should provide the AUC and plot ROC curves for all three models
   (b) For a given test dataset and a given threshold, it should generate the confusion matrix and provide accuracy and FP and TP rates for all three models.