Final exam due on December 7

Start Assignment

Due Wednesday by 11:59pm **Points** 50 **Submitting** a file upload **Attempts** 0
Allowed Attempts 3 **Available** Nov 30 at 11:59pm - Dec 7 at 11:59pm

STA 4364 Final Exam

Due Wednesday December 7 by 11:59PM on Webcourses

Instructions: Choose ONE of the problems below as your final exam problem. You can complete both problems for at most 5 points of extra credit.

Format: Please submit your final as both 1) a HTML or pdf document, and 2) also submit the source file in either R Markdown or Jupyter notebook format (at most one of each type of file).

Problem 1: (50 points) This problems will examine logistic regression for the Bank Marketing dataset in the file bank_data.csv (https://webcourses.ucf.edu/courses/1415323/files/96623605?wrap=1) (https://webcourses.ucf.edu/courses/1415323/files/96623605/download?download_frd=1) . Descriptions of the features are available here (https://archive.ics.uci.edu/ml/datasets/Bank+Marketing) . The variable of interest is the y feature, a binary outcome that indicates if a customer has made a deposit. The original dataset has been edited so that there are an equal number of positive and negative outcomes. All string-valued features are categorical and all number-valued features are numerical. Although the dataset description says that the duration feature might not be appropriate for true predictions, you will use it as a feature for this problem. Split your dataset into a training set with 80% of observations and a validation set with 20% of observations.

(a) Learn a logistic regression model using the full set of variables. From your full set of variables, select the most relevant features and create a logistic regression model using the reduced set of features. Plot an ROC curve and report the AUC for the full and reduced model on both the training and validation sets (4 curves in all). Comment on the degree of overfitting that you observe. Compare the performance of the full and reduced model on the validation set.

(b) Learn a LASSO logistic regression model (the R command model.matrix() might be useful for formatting your dataframe to use with glmnet, see the ridge and lasso regression (https://webcourses.ucf.edu/courses/1415323/files/94536192?wrap=1) (https://webcourses.ucf.edu/courses/1415323/files/94536192/download?download_frd=1) file (https://webcourses.ucf.edu/courses/1415323/files/94536192?wrap=1) (https://webcourses.ucf.edu/courses/1415323/files/94536192/download?download_frd=1)). Tune the value of λ using 10-fold cross-validation. Visualize the cross-validation error across different values of lambda, and report the value of λ that minimizes cross-validation error. Report the features that your LASSO model selects at the optimal value of λ, and compare these features to the features you selected in part a). Make an ROC curve and calculate the AUC for the training and validation data for your LASSO model.

Problem 2: (50 points) This problem will examine the data set **online_shoppers_intention.csv (https://webcourses.ucf.edu/courses/1415323/files/96623634?wrap=1)** **(https://webcourses.ucf.edu/courses/1415323/files/96623634/download?download_frd=1)** . Each row gives a variety of information about a single individual's online habits. Information about the features in this dataset can be found **here (https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset)** . We are interested in the Revenue response, which is a binary variable indicating whether an individual made a purchase while browsing. This dataset features class imbalance, so we will investigate some tools for dealing with this. Code related to oversampling and Precision-Recall curves for unbalanced data can be found here **class_imbalance_example.Rmd (https://webcourses.ucf.edu/courses/1415323/files/96623691?wrap=1)** **(https://webcourses.ucf.edu/courses/1415323/files/96623691/download?download_frd=1)** . Analyze this dataset by following the steps below:

(a) Load the data. Make sure categorical variables are formatted correctly. Merge rare categories for categorical features (40 or fewer observation) into a single feature. Split the data into a train and test set.

(b) The response in this dataset provides an example of class imbalance. In particular, the proportion of shopppers who do make a purchase is much smaller than the proportion of shoppers who do not make a purchase. Models might not learn anything in this case, because high accuracy can be achieved by always predicting the majority class. To deal with class imbalance, you can oversample the minority class, which involves sampling rows of the minority class with replacement. From your training dataset only (not the validation dataset), create a new training dataset by oversampling the minority class so there are an equal number of responses in both categories. You will use this training dataset when learning your model. You should use the original validation set when evaluating your model.

(c) Use the training data to create 2 different models to predict Revenue given all other features: an LDA model, and a QDA model. Your models can use all features as predictors, for this problem you don't have to investigate feature selection.

(d) Using the validation data, make an ROC curve for each model and calculate the AUC for the ROC curves. Comment on the differences that you observe between the different methods.