# Business Analytics: Marketing and Decisions

# HW4

Due date: **23:59PM June 2nd** (Tuesday), 2020

This is a group homework. Please upload your answers to homework problems (stored in **.pdf**) and the associated R code (stored in **.R**) to WM5. The file names must be BA_HW4_X, where X is your group ID. In addition, please make sure you specify your ID and name in the answer sheet. The above requirements and HW clarity are part of grading criteria. TAs have the rights to deduct points if you make the grading difficult, e.g. cannot tell the ownership of HW.

Q1. (24pts) Using *banks.csv*, run a logistic regression that models the status of a bank (strong or weak) as a function of two to financial measures: the ratio of total loans and leases to total assets (TotLns&Lses/Assets) and the ratio of total expenses to total assets (TotExp/Assets). Specify *the success class as weak*, and use the default cutoff value of 0.5.

(1) Write out the estimated equation that associates the financial condition of a bank with its two predictors in three formats: -12pts

(a) The logit as a function of the predictors

(b) The odds as a function of the predictors

(c) The probability as a function of the predictors

(2) Consider a new bank whose total loans and leases/asset ratio = 0.6 and total expense/asset ratio = 0.11. From your logistic regression model, estimate the following four quantities: the logit, the odds, the probability of being financially weak, and the classification of the bank. Hint: use R to do the calculation-7pts

(3) When a bank that is in poor financial condition is misclassified as financially strong, the misclassification cost is much higher than when a financially strong bank is misclassified as weak. To minimize the expected cost of misclassification, should the cutoff value for classification be increased or decreased? -5pts

Q2. (72pts) Please import *eBayAuction.csv* to R. The file contains information on 1972 auctions transacted on eBay.com during May-June, 2004. The goal is to use these data to build a model that will distinguish competitive auctions from non-competitive ones. A *competitive auction* is defined as an auction with at least two bids placed on the item being auctioned. The data include variables that describe the item (auction category), the seller (his or her eBay rating), and the auction terms that the seller selected (auction duration, opening price, currency, day of week of auction close). In addition, we have the price at which the auction closed. The goal is to predict whether or not an auction of interest will be competitive.

(1) Can you show competitive and non-competitive auctions in each product category? In addition, can you create a bar chart that shows the proportion of competitive auctions from Monday to Sunday?-8pts

(2) Can you create additional variables that can enrich the data analysis? -5pts

(3) Split the data into training (60%) and validation (40%) datasets. Run a logistic model with all predictors (including those from (2)) on the training data. Based on the result, what is the odds of bids in US dollars relative to that in EUR? Furthermore, what auction settings set by the seller (duration, opening price, ending day, currency) would you recommend as being most likely to lead to a competitive auction? Why is the question important? -14pts

(4) Build the confusion matrix on training data. Do we do better in identifying competitive auctions or non-competitive auctions?-9pts

(5) Following (3), can you compare prediction performance of the training with that of the validation? Why would we want to do so?-8pts

(6) Using the training data, what cutoff value should be used if the major objective is classification accuracy? How about validation data? Hint: the code below should help you answer the question (use in-class data as the example).-6pts

```
ROCpred=prediction(fitted(logit.reg3), loan3.df$Personal.Loan)
plot(performance(ROCpred, “acc”))
```

(7) Use Lasso regression on the training data. What are predicators with optimal penalty value equal to 0.005? Using validation data, would you recommend this model or the full model? Hint: the code below is the lasso logistic regression -10pts

```
glmnet(x, y, family=‘binomial’)
```

(8) “A model’s performance on validation data may be overly optimistic.” Can you explain the statement? -4pts

(9) If we want to *predict* at the beginning of an auction whether it will be competitive, we cannot use the information on the closing price. Run a logistic model with all predictors as above, excluding price. How does this model compare to the full model? Please show the ROC curve of the two models. Hint: think of what data should you apply-12pts

Q3. (50pts) Analyze the *UniversalBank.csv* (see lecture 5) and finish the following tasks.

(1) Drop the columns of ID and ZIP code. Create dummy variables Under (Education=1) and Grad (Education=2). After that drop Education. -3pts

(2) Use the *sample.split* ( ) function to create a training set with 70% observations and a test set with 30% observations based on the *Personal.Loan* variable. -3pts

(3) Fit a decision tree for **classification**, using the *information gain* as splitting criterion. Set the minbucket=25. Explain and interpret the fitted tree. -8pts

(4) Fit a random forest with 500 trees, in which the minbucket of each tree is 25. Also, each tree in the forest only randomly selects 4 variables. Show the variable importance plot and explain what you have found in terms of variables' effects. -8pts

(5) Fit a gradient boosting machine with 1500 trees, interaction depth=4, and shrinkage=0.05. Show the variable importance plot of *gbm*. Compare its variables' importance versus the random forest in part (5). -8pts

(6) Fit a logistic regression model using all available variables. Examine parameter estimates of the logistic regression model. What are the variables that are statistically significant as well as important in RF/GBM? -8pts

(7) Use the four models to generate predictions for the test set. Show the ROC curves of the four models. Explain what you observe. -8pts

(8) Report the AUC (area under curve) of the four models. Which model is the best? -4pts