PROBLEMS

13.1 Acceptance of Consumer Loan Universal Bank has begun a program to encourage its existing customers to borrow via a consumer loan program. The bank has promoted the loan to 5000 customers, of whom 480 accepted the offer. The data are available in file *UniversalBank.csv*. The bank now wants to develop a model to predict which customers have the greatest probability of accepting the loan, to reduce promotion costs and send the offer only to a subset of its customers.

We will develop several models, then combine them in an ensemble. The models we will use are (1) logistic regression, (2) k-nearest neighbors with $k = 3$, and (3) classification trees. Preprocess the data as follows:

- Zip code can be ignored.
- Partition the data: 60% training, 40% validation.

a. Fit models to the data for (1) logistic regression, (2) k-nearest neighbors with $k = 3$, and (3) classification trees. Use Personal Loan as the outcome variable. Report the validation confusion matrix for each of the three models.

b. Create a data frame with the actual outcome, predicted outcome, and each of the three models. Report the first 10 rows of this data frame.

c. Add two columns to this data frame for (1) a majority vote of predicted outcomes, and (2) the average of the predicted probabilities. Using the classifications generated by these two methods derive a confusion matrix for each method and report the overall accuracy.

d. Compare the error rates for the three individual methods and the two ensemble methods.

13.2 eBay Auctions—Boosting and Bagging Using the eBay auction data (file *eBayAuctions.csv*) with variable Competitive as the outcome variable, partition the data into training (60%) and validation (40%).

a. Run a classification tree, using the default controls of *rpart()*. Looking at the validation set, what is the overall accuracy? What is the lift on the first decile?

b. Run a boosted tree with the same predictors (use function *boosting()* in the `adabag` package). For the validation set, what is the overall accuracy? What is the lift on the first decile?

c. Run a bagged tree with the same predictors (use function *bagging()* in the `adabag` package). For the validation set, what is the overall accuracy? What is the lift on the first decile?

d. Run a random forest (use function *randomForest()* in package `randomForest` with argument *mtry* = 4). Compare the bagged tree to the random forest in terms of validation accuracy and lift on first decile. How are the two methods conceptually different?

13.3 Predicting Delayed Flights (Boosting). The file *FlightDelays.csv* contains information on all commercial flights departing the Washington, DC area and arriving at New York during January 2004. For each flight there is information on the departure and arrival airports, the distance of the route, the scheduled time and date of the flight, and so on. The variable that we are trying to predict is whether or not a flight is delayed. A delay is defined as an arrival that is at least 15 minutes later than scheduled.