BANA 7042: Assignment 3

Logistic regression (part II)

Note: Please be sure to properly label all figures and include a caption for each!

For this assignment, you'll be working with data from the direct marketing campaigns (phone calls) of a Portuguese banking institution. The goal is to predict the likelihood that a client will subscribe a term deposit (i.e., the binary variable labeled `y`). The data are available from the UC Irvine Machine Learning Repository.

Direct URL: https://archive.ics.uci.edu/dataset/222/bank+marketing

I have already downloaded the data and split them into two samples, which are available on the corresponding `Assignments` tab in our course's Canvas page:

- `bank.csv` - these data will be used to build/train your logistic regression models;
- `bank_new.csv` - these data will be used for testing your final logistic regression model.

Question 1: Use the `bank.csv` data to build an initial logistic regression model using all available predictors.

a) Does there appear to be any issues with the model? If so, please elaborate on what you see.
b) Use the methods described in class to explore and potentially reduce the model further and explain your process (e.g., look for potential multicollinearity and redundant features, etc.).
c) Draw an ROC curve and calibration plot for your final model. Which plot do you think is more useful given the task at hand?
d) Read the background documentation for these data here. Does there appear to be any *leakage* variables in the data? If so, describe which ones and refit your model without those predictors. How do the ROC and calibration curves compare to the previous model (e.g., better or worst)? Does this make sense?

Question 2: Try exploring forward selection (BS) and backward elimination (BE). For FS, start with an intercept-only model. For BE, you can start with your final model from **Question 1**. How do the results compare between FS and BE? How do they compare to your final model from **Question 1**?

Question 3: Using the Brier score metric (see function below), compute the LOCO-based variable importance for all of the variables in your final model and display the results using a simple dotchart (do this using the Brier score). Do the results make sense? (You can use and modify the R code we walked through in class.)

```
# Brier score function to compute MSE between binary y and probability p
#
# Args:
#   y - the binary 0/1 outcome
#   p - the predicted probability that y=1
brier_score <- function(y, p) {
  mean((y - p) ^ 2)
}
```

Question 4: Apply your final model to the `bank_new.csv` data to obtain the predicted probability that each customer will subscribe to a term deposit if offered. Using these results, construct a cumulative gains chart (like we did in class). If we used your model to identify the 1000 customers most likely to subscribe, how many successes can we expect? Be sure to include your plot as well.