Team 6 – Jonathan Zaremba, Tyler Kramer, Alex Holland, Maura Glynn

April 31, 2020

Lab 4 Write-Up

Exercise 1

1. *Execute the code above. Based on the results, rank the models from "most underfit" to "most overfit".*

Having executed the code, we found that the most underfit model was fit_caps, followed by fit_selected, fit_additive, and fit_over — in that order.

2. *Re-run the code above with 100 folds and a different seed. Does your conclusion change?*

The code was re-run using 100 folds (K = 100 in the cv.glm function) and a different seed (seed set to 21). The conclusion does not change: the most underfit model was *fit_caps, followed by fit_selected, fit_additive, and fit_over — in that order.*

3. *Generate four confusion matrices for each of the four models fit in Part 1.*

Fit_caps:

```
          actual
predicted nonspam spam
  nonspam    2022 1066
  spam        162  351
>
```

Fit_selected:

```
          actual
predicted nonspam spam
  nonspam    2073  615
  spam        111  802
>
```

Fit_additive:

```
          actual
predicted nonspam spam
  nonspam    2057  157
  spam        127 1260
>
```

Fit_over:

```
          actual
predicted nonspam spam
  nonspam    1725  103
  spam        459 1314
>
```

4. *Which is the best model? Write 2 paragraphs justifying your decision. You must mention (a) the overall accuracy of each model; and (b) whether some errors are better or worse than others, and you must use the terms specificity and sensitivity. For (b) think carefully… misclassified email is a pain in the butt for users!*

The fit_caps model predicted 162 false positives (positive meaning it was flagged as spam) and 1066 false negatives. The fit_selected model predicted 111 false positives and 615 false negatives. The fit_additive model predicted 127 false positives and 157 false negatives. The fit_over model predicted 459 false positives and 103 false negatives. In this situation, we believe false positives (non-spam email misclassified as spam) is a worse error than false negatives — this is because it can be annoying and costly for important messages to get "lost" in a spam folder. Since this is the case, a model with high sensitivity is more important than a model with high specificity.

Considering that misclassified email is a pain in the butt for users

Exercise 2

In our fit_education model, the educationtertiary and educationsecondary are both positive meaning that a person having gotten a college degree or above made it more likely that they would deposit money with the bank, while a negative educationunkown means a person whose education was unknown is more likely to not have made a deposit.

In our fit_selected model, the education variables all have the same sign as in the previous model. The balance coefficient is near zero indicating people with small balances have roughly the same likelihood of making a deposit as people with large balances. The housingyes coeffient is negative to a larger degree indicating that people with housing were less likely to make a deposit. For the month coefficients, it appears that observations taking place in March, September, October, and December were most likely to yield a positive result.

Describing every single coefficient in our fit_over model would take several pages, but essentially they describe the interactions between each variable, each level of education, and y.