Lab 4:
Classification and Logistic Regression

Team: Uno

Exercise 1

1. Execute the code above. Based on the results, rank the models from "most underfit" to "most overfit".
    a. 1.(Most Underfit) cv.glm(spam_trn, fit_caps, K = 5)$delta[1]
    b. 2. cv.glm(spam_trn, fit_selected, K = 5)$delta[1]
    c. 3. cv.glm(spam_trn, fit_over, K = 5)$delta[1]
    d. 4. (Most Overfit) cv.glm(spam_trn, fit_additive, K = 5)$delta[1]
2. Re-run the code above with 100 folds and a different seed. Does your conclusion change?
    a. When running the code with 100 folds and a different seed, the same conclusions about order can be made.
3. Generate four confusion matrices for each of the four models fit in Part 1.
    a.
    ```
    > conf_mat_1
             actual
    predicted nonspam spam
      nonspam    2004 1016
      spam        183  398
    ```
    b.
    ```
    > conf_mat_2
             actual
    predicted nonspam spam
      nonspam    2050  599
      spam        137  815
    ```
    c.
    ```
    > conf_mat_3
             actual
    predicted nonspam spam
      nonspam    2050  161
      spam        137 1253
    ```
    d.
    ```
    > conf_mat_4
             actual
    predicted nonspam spam
      nonspam    1979  153
      spam        208 1261
    ```
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!

Going in depth on each of the four models we explored - all generally outperform the classifier of simply classifying all observations to the majority class, as stated in the lab to be effective. As a standard, accuracy should be defined as the ability of the model to predict both specificity and sensitivity. When the number of false negatives are low, the sensitivity is high and when the number of false positives are low, the specificity is high. Model two (fit_selected) and three (fit_additive) both have the lowest number of false negatives at 137 each. As well, model three (fit_additive at 161) and four (fit_over at 153) have the lowest number of false positives.

With that being said, in terms of predicting the test data, the best model would be model model three as it is both specific and sensitive, making it the most accurate overall.The least accurate model would be model one (fit_caps). With model two four coming in second and model four (fit_over) coming in third. All assessed based on total accuracy (number of false positives and false negatives). In terms of some errors being better and/or worse than others is relative. Often receiving a false positive is worse than receiving a false negative. Better safe than sorry when predicting.

Exercise 2

1. Use the bank data and create a train / test split.
    a. Visible on R script.
2. Run any logistic regression you like with 10-fold cross-validation in order to predict the yes/no variable (y).
    a.
```
> fit_caps

Call:  glm(formula = loan ~ balance, family = binomial, data = bank_trn)

Coefficients:
(Intercept)      balance
 -1.5965122   -0.0001298

Degrees of Freedom: 999 Total (i.e. Null);  998 Residual
Null Deviance:      845.4
Residual Deviance: 837.5        AIC: 841.5
```
    b.
```
> fit_selected

Call:  glm(formula = loan ~ balance + duration + marital + age, family = binomial,
    data = bank_trn)

Coefficients:
   (Intercept)        balance       duration  maritalmarried   maritalsingle            age
    -2.0889307     -0.0001328      0.0005634       0.3571958       0.0768115      0.0020354

Degrees of Freedom: 999 Total (i.e. Null);  994 Residual
Null Deviance:      845.4
Residual Deviance: 831.9        AIC: 843.9
```

c.

```
> fit_additive

Call:  glm(formula = loan ~ ., family = binomial, data = bank_trn)

Coefficients:
       (Intercept)                age     jobblue-collar    jobentrepreneur       jobhousemaid
         -2.336194           0.007783          -0.101632          -0.178046          -0.631756
     jobmanagement         jobretired   jobself-employed        jobservices         jobstudent
         -0.568343          -0.316494           0.055188          -0.017650         -14.394983
     jobtechnician      jobunemployed         jobunknown     maritalmarried      maritalsingle
         -0.344574          -0.629785          -0.514692           0.383129           0.328527
educationsecondary  educationtertiary   educationunknown         defaultyes            balance
          0.406921           0.019857          -1.897590           1.569121          -0.000095
        housingyes   contacttelephone     contactunknown                day           monthaug
         -0.133863          -0.212262           0.418650          -0.028422          -0.151131
          monthdec           monthfeb           monthjan           monthjul           monthjun
          1.495865          -0.695605           0.032815           0.999117          -0.622996
          monthmar           monthmay           monthnov           monthoct           monthsep
          0.480030          -0.202381           0.758782           0.717785           0.047280
          duration           campaign           previous               yyes
          0.001167           0.033971           0.073684          -1.055121

Degrees of Freedom: 999 Total (i.e. Null);  961 Residual
Null Deviance:      845.4
Residual Deviance: 755  AIC: 833
```

d.

```
> fit_over

Call:  glm(formula = loan ~ balance * (.), family = binomial, data = bank_trn,
    maxit = 50)

Coefficients:
              (Intercept)                    balance                        age             jobblue-collar
               -2.236e+00                 -3.917e-04                  1.697e-02                 -3.546e-01
          jobentrepreneur               jobhousemaid              jobmanagement                 jobretired
               -7.167e-01                 -8.272e-01                 -9.575e-01                  5.687e-02
         jobself-employed                jobservices                 jobstudent              jobtechnician
               -3.878e-01                 -1.987e-01                 -9.240e+02                 -3.909e-01
            jobunemployed                 jobunknown             maritalmarried              maritalsingle
               -1.136e+00                 -5.600e-01                  5.911e-01                  6.266e-01
       educationsecondary          educationtertiary           educationunknown                 defaultyes
                1.708e-01                  5.031e-02                 -2.173e+00                  1.802e+00
               housingyes           contacttelephone             contactunknown                        day
               -3.764e-01                  5.420e-02                  6.351e-01                 -3.539e-02
                 monthaug                   monthdec                   monthfeb                   monthjan
               -3.883e-01                  2.395e+00                 -7.563e-01                 -6.981e-01
                 monthjul                   monthjun                   monthmar                   monthmay
                5.955e-01                 -7.162e-01                 -2.386e+02                 -3.204e-01
                 monthnov                   monthoct                   monthsep                   duration
                4.498e-01                  1.100e+00                  2.264e+00                  1.341e-03
                 campaign                   previous                       yyes                balance:age
                5.833e-02                  5.226e-02                 -1.285e+00                 -8.180e-06
   balance:jobblue-collar    balance:jobentrepreneur       balance:jobhousemaid      balance:jobmanagement
                3.273e-04                  7.759e-04                 -2.301e-04                  7.431e-04
       balance:jobretired    balance:jobself-employed        balance:jobservices         balance:jobstudent
               -8.845e-04                  6.426e-04                  3.070e-04                  1.707e-01
    balance:jobtechnician      balance:jobunemployed         balance:jobunknown     balance:maritalmarried
                1.347e-04                  1.363e-03                  1.297e-04                 -1.066e-04
    balance:maritalsingle balance:educationsecondary  balance:educationtertiary   balance:educationunknown
               -1.682e-04                  4.176e-04                 -1.390e-04                  1.228e-04
       balance:defaultyes         balance:housingyes   balance:contacttelephone     balance:contactunknown
               -5.498e-04                  1.121e-04                 -4.371e-04                 -2.277e-04
              balance:day           balance:monthaug           balance:monthdec           balance:monthfeb
                6.535e-06                  9.961e-05                 -9.678e-04                 -3.840e-04
         balance:monthjan           balance:monthjul           balance:monthjun           balance:monthmar
                4.862e-04                  5.013e-04                 -2.828e-04                  4.444e-02
         balance:monthmay           balance:monthnov           balance:monthoct           balance:monthsep
               -1.023e-04                  1.409e-04                 -2.974e-03                 -6.245e+00
         balance:duration           balance:campaign           balance:previous               balance:yyes
               -8.824e-09                 -5.311e-05                  7.958e-05                 -3.585e-05

Degrees of Freedom: 999 Total (i.e. Null);  924 Residual
Null Deviance:      845.4
Residual Deviance: 709.3        AIC: 861.3
```

3. Discuss the interpretation of the coefficients in your model. That is, you must write at least one sentence for each of the coefficients which describes how it is

related to the response. You may use transformations of variables if you like. FAKE EXAMPLE: age has a positive coefficient, which means that older individuals are more likely to have y = yes.

   a. Balance which has a negative coefficient saying if you have balance in you account the less likely you are to have a loan
   b. Marital status which are positive for both single and married making it more likely you have a loan
   c. Duration which is positive meaning the longer you have an account the more likely you have a loan
   d. Age which has a positive coefficient meaning the older you are the more likelihood you would have a loan

4. Create a confusion matrix of your preferred model, evaluated against your test data.
   a.
```
> (conf_mat = make_conf_mat(predicted = bank_tst_pred, actual = bank_tst$loan))
         actual
predicted   no  yes
      no  2948  516
      yes   32   25
```