IMT 573: Problem Set 8

Prediction

Tommy Huynh

Due: August 15, 2021

Collaborators:

Instructions: Before beginning this assignment, please ensure you have access to R and RStudio; this can be on your own personal computer or on the IMT 573 R Studio Cloud.

1. Download the `08_ps_prediction.Rmd` file from Canvas or save a copy to your local directory on RStudio Cloud. Supply your solutions to the assignment by editing `08_ps_prediction.Rmd`.
2. Replace the "YOUR NAME HERE" text in the `author:` field with your own full name. Any collaborators must be listed on the top of your assignment.
3. Be sure to include well-documented (e.g. commented) code chucks, figures, and clearly written text chunk explanations as necessary. Any figures should be clearly labeled and appropriately referenced within the text. Be sure that each visualization adds value to your written explanation; avoid redundancy – you do no need four different visualizations of the same pattern.
4. Collaboration on problem sets is fun and useful, and we encourage it, but each student must turn in an individual write-up in their own words as well as code/work that is their own. Regardless of whether you work with others, what you turn in must be your own work; this includes code and interpretation of results. The names of all collaborators must be listed on each assignment. Do not copy-and-paste from other students' responses or code.
5. All materials and resources that you use (with the exception of lecture slides) must be appropriately referenced within your assignment.
6. Remember partial credit will be awarded for each question for which a serious attempt at finding an answer has been shown. Students are *strongly* encouraged to attempt each question and to document their reasoning process even if they cannot find the correct answer. If you would like to include R code to show this process, but it does not run without errors you can do so with the `eval=FALSE` option as follows:

```
a + b # these object don't exist
# if you run this on its own it with give an error
```

7. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, download and rename the knitted PDF file to `ps7_YourLastName_YourFirstName.pdf`, and submit the PDF file on Canvas.

Setup: In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
library(gridExtra)
```

```
library(MASS)
library(pROC)
library(dplyr)
library(arm)
library(randomForest)
library(Metrics)
library(knitr) # this will keep code on the page!
opts_chunk$set(tidy.opts=list(width.cutoff=60),tidy=TRUE)
```

Data: In this problem set we will use the `flights` dataset used previously in class. The flights dataset (via the the *nycflights13* library) contains information on flight delays and weather.

As part of this assignment, we will evaluate the performance of several statistical learning methods. We will fit our learning models using a set of *training* observations and measure its performance on a set of *test* observations.

Problem 1: Discuss the advantages of using a training/test split when evaluating statistical models.

The advantages of using a training/test split when evaluation statistics models is when you have a large data set, you can separate them to have two seperate models and compare them to ensure accuracy.

Problem 2: Predictions with a continuous output variable

```
library(nycflights13)
data(flights)
data(weather)

nyc_flights <- merge(flights, weather)
```

(a) Load in the flights dataset. Join the flights data to the weather data based on the departure location, date, and hour of the flight. Exclude data entries which cannot be joined to weather data. Copy the joined data so we can refer to it later.

```
nyc_flights <- nyc_flights[, c(9, 4, 7, 20, 24, 26, 28)]
nyc_flights <- nyc_flights[complete.cases(nyc_flights),]
```

(b) From the joined data, keep only the following columns as we build our first model: departure delay, origin, departure time, temperature, wind speed, precipitation, and visibility. Omit observations that do not have all of these variables present.

```
first <- nyc_flights[1:65380,]
second <- nyc_flights[65380:326898,]
```

(c) Split your data into a *training* and *test* set based on an 80-20 split. In other words, 80% of the observations will be in the training set and 20% will be in the test set. Remember to set the random seed.

```
m1 <- lm(dep_delay ~ origin + dep_time + temp + wind_speed + precip + visib, data = first)
summary(m1)
```

(d) Build a linear regression model to predict departure delay using the subset of variables indicated in (3.). What is the RMSE on the training set? What is the RMSE on the test set? Which is higher and is this expected?

```
##
## Call:
## lm(formula = dep_delay ~ origin + dep_time + temp + wind_speed +
##     precip + visib, data = first)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -54.60  -14.25   -6.46    2.00 1308.34
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.789e-01  7.554e-01   0.369    0.712
## originJFK   -6.395e+00  2.985e-01 -21.423   <2e-16 ***
## originLGA   -5.338e+00  2.997e-01 -17.810   <2e-16 ***
## dep_time     1.637e-02  2.589e-04  63.230   <2e-16 ***
## temp        -8.860e-02  8.242e-03 -10.749   <2e-16 ***
## wind_speed   3.890e-01  2.164e-02  17.974   <2e-16 ***
## precip       1.123e+02  1.155e+01   9.720   <2e-16 ***
## visib       -1.191e+00  5.833e-02 -20.422   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.16 on 65372 degrees of freedom
## Multiple R-squared:  0.07538,    Adjusted R-squared:  0.07528
## F-statistic: 761.4 on 7 and 65372 DF,  p-value: < 2.2e-16
```

```
m2 <- lm(dep_delay ~ origin + dep_time + temp + wind_speed + precip + visib, data = second)
summary(m2)
```

```
##
## Call:
## lm(formula = dep_delay ~ origin + dep_time + temp + wind_speed +
##     precip + visib, data = second)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -102.60  -19.73   -8.41    3.13 1124.37
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.2361877  0.5243945  -2.357   0.0184 *
## originJFK   -4.3232905  0.1904896 -22.696   <2e-16 ***
## originLGA   -3.8963529  0.1923260 -20.259   <2e-16 ***
## dep_time     0.0227259  0.0001619 140.368   <2e-16 ***
## temp         0.1262778  0.0044170  28.589   <2e-16 ***
```

```
## wind_speed   0.2073733  0.0148886  13.928   <2e-16 ***
## precip      58.0797412  2.5671789  22.624   <2e-16 ***
## visib       -2.4613865  0.0438279 -56.160   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.86 on 261511 degrees of freedom
## Multiple R-squared:  0.09352,    Adjusted R-squared:  0.09349
## F-statistic:  3854 on 7 and 261511 DF,  p-value: < 2.2e-16
```

```
first_mse <- mse(first$dep_delay, predict(m1, first))
second_mse <- mse(second$dep_delay, predict(m2, second))
```

The RMSE in the test set is 970 whereas the RMSE in the training set is 1589. The training set is higher than expected.

(e) Now, improve upon these prediction results by including additional variables in your model. Make sure you keep at least 95% of original data (i.e. about 320K observations across both the training and test datasets). Do not include the arrival time, scheduled arrival time, or the arrival delay in your model. Use the same observations as above for the training and test sets (i.e. keep the same rows but add different variables/columns at your discretion). Can you improve upon the training RMSE? Once you have a model that you feel adequately improves the training RMSE, does your model improve the test RMSE? Which variables did you include in your model?

EXTRA CREDIT: Predictions with a categorical output (classification)

In this problem our goal is to predict the survival of passengers. First, let's train a logistic regression model for survival that controls for the socioeconomic status of the passenger.

(a) Load in the `titanic` data. Split your data into a *training* and *test* set based on an 80-20 split. In other words, 80% of the observations will be in the training set and 20% will be in the test set. Remember to set the random seed.

(b) Fit the model described above (i.e. one that only takes into account socioeconomic status) using the `glm` function in R.

(c) What might you conclude based on this model about the probability of survival for lower class passengers?

(d) Predict the survival of passengers for each observation in your test set using the model fit in Problem 2. Save these predictions as `yhat`.

(e) Use a threshold of 0.5 to classify predictions. What is the number of false positives on the test data? Interpret this in your own words.

(f) Using the `roc` function, plot the ROC curve for this model. Discuss what you find.

(g) Suppose we use the data to construct a new predictor variable based on a passenger's listed title (i.e. Mr., Mrs., Miss., Master). Why might this be an interesting variable to help predict passenger survival?

(h) Fit a second logistic regression model including this new feature. Use the `summary` function to look at the model. Did this new feature improve the model?

(i) Comment on the overall fit of this model. For example, you might consider exploring when misclassification occurs.