Lab 7 - Modeling II

Introduction

In this lab you'll continue your practice of statistical modeling and then venture on to quantifying uncertainty.

Packages

In this lab we will work with the **tidyverse** and **tidymodels** packages.

```
library(tidyverse)
library(tidymodels)
```

Guidelines

Your plots should include an informative title, axes should be labeled, and careful consideration should be given to aesthetic choices.

In addition, the code should all the code should be be able to be read (not run off the page) when you render to PDF. Make sure that is the case, and add line breaks where the code is running off the page.

> **!** Important
>
> Render your document. If your code is running off the page for a given question such that we can't see your entire code, we will not evaluate any of the code for that question. The question will automatically receive a 0. This is something you can and should verify before you turn in your work.

You should have at least 3 commits with meaningful commit messages by the end of the assignment.

Additionally, if you're using functions that are not introduced in the course materials, you must cite your sources.

> **!** Important
>
> Failure to cite outside resources used, including Large Language Models like Chat GPT, is a violation of the Duke Community Standard and will be treated as such.

Part 1 - Building a spam filter

The data come from incoming emails in David Diez's (one of the authors of OpenIntro textbooks) Gmail account for the first three months of 2012. All personally identifiable information has been removed. The dataset is called `email` and it's in the **openintro** package.

The outcome variable is `spam`, which takes the value `1` if the email is spam, `0` otherwise.

Question 1

a. Fit a logistic regression model predicting `spam` from `exclaim_mess` (the number of exclamation points in the email message). Then, display the tidy output of the model.

b. Using this model and an R function like `predict()` or `augment()`, predict the probability the email is spam if it contains 10 exclamation points.

Question 2

a. Fit another logistic regression model predicting `spam` from `exclaim_mess`, `winner` (indicating whether "winner" appeared in the email), and `urgent_subj` (whether the word "urgent" is in the subject of the email). Then, display the tidy output of the model.

b. Using this model, predict spam / not spam for all emails in the `email` dataset with `augment()`. Store the resulting data frame with an appropriate name.

c. Using your data frame from the previous part, determine, in a single pipeline, and using `count()`, the numbers of emails:

- that are labelled as spam that are actually spam
- that are not labelled as spam that are actually spam
- that are labelled as spam that are actually not spam
- that are not labelled as spam that are actually not spam

Store the resulting data frame with an appropriate name.

d. In a single pipeline, and using `mutate()`, calculate the false positive and false negative rates. In addition to these numbers showing in your R output, you must write a sentence that explicitly states and identified the two rates.

Question 3

a. Fit another logistic regression model predicting `spam` from `exclaim_mess` and another variable you think would be a good predictor. Provide a 1-sentence justification for why you chose this variable. Display the tidy output of the model.

b. Using this model, predict spam / not spam for all emails in the `email` dataset with `augment()`. Store the resulting data frame with an appropriate name.

c. Using your data frame from the previous part, determine, in a single pipeline, and using `count()`, the numbers of emails:

- that are labelled as spam that are actually spam
- that are not labelled as spam that are actually spam
- that are labelled as spam that are actually not spam
- that are not labelled as spam that are actually not spam

Store the resulting data frame with an appropriate name.

d. In a single pipeline, and using `mutate()`, calculate the false positive and false negative rates. In addition to these numbers showing in your R output, you must write a sentence that explicitly states and identified the two rates.

e. Based on the false positive and false negatives rates of this model, comment, in 1-2 sentences, on which model is preferable and why.

Part 2 - Hotel cancellations

For this exercise, we will work with hotel cancellations. The data describe the demand of two different types of hotels. Each observation represents a hotel booking between July 1, 2015 and August 31, 2017. Some bookings were cancelled (`is_canceled = 1`) and others were kept, i.e., the guests checked into the hotel (`is_canceled = 0`). You can view the code book for all variables here.

The data can be found in the `data` folder: `hotels.csv`.

Question 4

Read the data and then explore attributes of bookings and summarize your findings in 5 bullet points. You must provide a visualization or summary supporting each finding.

> **Note**
>
> This is not meant to be an exhaustive exploration. We anticipate a wide variety of answers to this question.

Question 5

Using these data, we will try to answer the following question:

> *Do we expect reservations earlier in the month or later in the month to be cancelled?*

(a) **Exploration:** In a single pipeline, calculate the mean arrival date (`arrival_date_day_of_month`) for both booking that were cancelled and that were not cancelled.

(b) **Justification:** In your own words, explain why we can not fit a linear model to model the relationship between if a hotel reservation was cancelled and the day of month for the booking.

(c) **Model fitting and interpretation:**

- Fit the appropriate model and display a tidy summary of the model output.
- Interpret the slope coefficient in context of the data and the research question.

(d) **Predicted:** Calculate the probability that the hotel reservation is cancelled if it the arrival date date is on the 26th of the month. Based on this probability, would you predict this booking would be cancelled or not cancelled. Explain your reasoning for your classification.

Part 3 - Statistics experience

Question 6

You have two options for this exercise. Clearly indicate which option you choose. Then, summarize your experience in no more than 10 bullet points.

Include the following on your summary:

- Name and brief description of what you did.

- Something you found new, interesting, or unexpected
- How the talk/podcast/interview/etc. connects to something we've done in class.
- Citation or link to web page for what you watched or who you interviewed.

Option 1: Listen to a podcast or watch a video about statistics and data science. The podcast or video must be *at least 30 minutes* to count towards the statistics experience. A few suggestions are below:

- posit::conf 2023 talks
- rstudio::conf 2022 talks
- rstudio::global 2021 talks
- rstudio::conf 2020 talks
- Stats + Stories Podcast
- Casual Inference Podcast
- Not So Standard Deviations

This list is not exhaustive. You may listen to other podcasts or watch other statistics/data science videos not included on this list. Ask your professor if you are unsure whether a particular podcast or video will count towards the statistics experience.

Option 2: Talk with someone who uses statistics in their daily work. This could include a professor, professional in industry, graduate student, etc.

Part 4 - Get creative

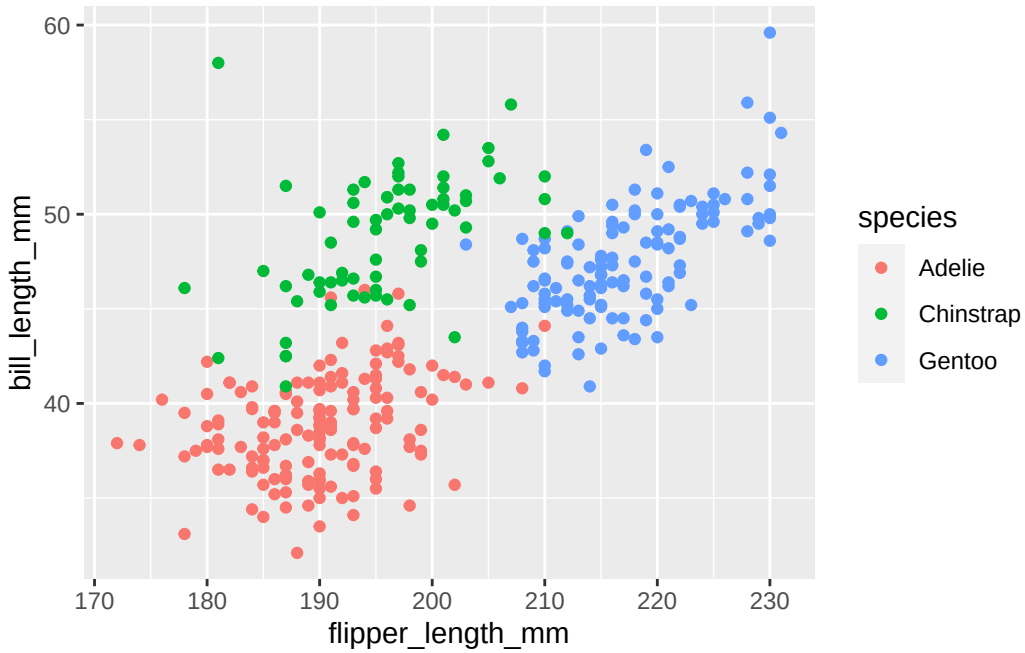
Question 7

Your task is to make the following plot as ugly and as ineffective as possible. Change colors, axes, fonts, theme, or anything else you can think of in the code chunk below. You can also search online for other themes, fonts, etc. that you want to tweak. Try to make it as ugly as possible, the sky is the limit!

In 2-3 sentences, explain why the plot you created is *ugly* (to you, at least) and ineffective.

```
ggplot(
  penguins,
  aes(x = flipper_length_mm, y = bill_length_mm, color = species)
  ) +
  geom_point()
```

```
Warning: Removed 2 rows containing missing values (`geom_point()`).
```



Part 5 - Harassment at work

The General Social Survey (GSS) gathers data on contemporary American society in order to monitor and explain trends and constants in attitudes, behaviors, and attributes. Hundreds of trends have been tracked since 1972. In addition, since the GSS adopted questions from earlier surveys, trends can be followed for up to 70 years. The GSS contains a standard core of demographic, behavioral, and attitudinal questions, plus topics of special interest. Among the topics covered are civil liberties, crime and violence, inter-group tolerance, morality, national spending priorities, psychological well-being, social mobility, and stress and traumatic events.

The data for this part comes from the `gss16` data frame (containing data from the 2016 GSS) from the **dsbox** package.

Question 8

In 2016, the GSS added a new question on harassment at work. The question is phrased as the following.

Over the past five years, have you been harassed by your superiors or co-workers at your job, for example, have you experienced any bullying, physical or psychological abuse?

Answers to this question are stored in the `harass5` variable in our data set.

a. Create a subset of the data that only contains `Yes` and `No` answers for the harassment question. How many respondents chose each of these answers?

b. Describe how bootstrapping can be used to estimate the proportion of all Americans who have been harassed by their superiors or co-workers at their job.

c. Calculate a 95% bootstrap confidence interval for the proportion of Americans who have been harassed by their superiors or co-workers at their job. Use 1000 iterations when creating your bootstrap distribution. Interpret this interval in context of the data.

Part 6 - Data science assessment

Question 9

> **!** Important
>
> This question is graded out of 10 points.

Take a data science assessment at https://duke.qualtrics.com/jfe/form/SV_bPZS3LIKoYA0fL8. Your assessment will be graded out of 10 points, and those points will be added to your score for the remaining points on this lab. There are no additional penalties for wrong answers, so please answer all questions.

Optionally, you can participate in the research study this assessment is a part of. I will not find out whether you participate in the study or not, your scores will be passed on to me by the research administrator regardless of your study participation.