Stat 652 Project Guidelines

Brad McNeney

2019-10-16

Data

The data are on flights from three New York City airports in 2013, from the `nycflights13` package. Data were combined from four datasets from this package:

- `flights`
- `weather`
- `airports`, and
- `planes`

Please read about the variables in each dataset by typing `help(datasetname)` from the R console.

```
library(tidyverse)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
## -- Attaching packages ---------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.1     v purrr   0.3.2
## v tibble  2.1.1     v dplyr   0.8.1
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(nycflights13)
#help(flights)
#help(weather)
#help(airports)
#help(planes)
fltrain <- read_csv("fltrain.csv.gz")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   carrier = col_character(),
##   tailnum = col_character(),
##   origin = col_character(),
##   dest = col_character(),
##   time_hour = col_datetime(format = ""),
##   name = col_character(),
##   dst = col_character(),
##   tzone = col_character(),
##   type = col_character(),
##   manufacturer = col_character(),
##   model = col_character(),
##   engine = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
# when the test data are made available:
# fltest <- read_csv("fltest.csv.gz")
```

Your task is to build a prediction model for departure delays. I have held out 1336776 observations as a test dataset that I will release a few days before the project is due.

Project Length and Scope

Your report should be no more than 5 pages long, plus references. You must also include an Appendix of R code that can be used to reproduce the analyses refered to in the report. There is no page limit for the Appendix, but please use judgement about what to include. Too long and it is not likely to be read. You are encouraged to try several prediction methods, and can compare these methods, but your report should focus on one method in particular. You **must** use methods discussed in class.

Grading Criteria

The criteria for the report are as follows.

Report (25 marks)

The report should have the following sections

1. Introduction (brief)
2. Data (brief)
3. Methods
4. Results
5. Conclusions and Discussion

The reports will be judged on the following criteria.

- Content (20): The content should be clear, accurate, complete and at the level of students in Stat 452. In the Methods you should provide a brief description of any statistical methods you use. Please restrict yourself to methods that were covered in class. You can mention methods not covered as areas of future work. Methods you considered but were not the focus of your report should be briefly mentioned here too. In Results you should summarize and interpret the fitted model. Though the primary goal is prediction, your insights into the data-generating process are important. Refer to the Appendix for the code that implements your prediction equation. In the Conclusions and Discussion present your conclusions, discuss short-comings of your approach, and, optionally, ideas for further work.
- Organization (3): Though the report is structured, you should present your ideas logically within each section.
- Grammar and spelling (2): Please proof-read your report.

Code (15 marks)

The code in your Appendix should look correct and be readable. Given the size of the dataset, you do not need to provide run-able code for all analyses. Use `{r, eval=FALSE}` in your computationally-intensive code chunks to prevent them from running. The Appendix will be judged on the following criteria.

- Software Details (2): List the version of R you are using and the names of all packages used in your analysis **at the beginning** of the Appendix. Please also provide an estimate of the time it will take to knit the code if more than about 2 minutes.
- Correctness (5): There should be no errors in data processing, function calls, etc.

- Readability (5): The steps of your analysis should be clearly layed out and it should be easy for the reader to find the final prediction equation/method.
- Efficiency (3): Please take steps to avoid computational inefficiencies, such as loops and excessive copying of large R objects.