Stat 652 Project

Vincent Chiu

2019-10-16

Loading Libraries

```
library(tidyverse)
```

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

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     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()
```

Loading the data

```
library(nycflights13)
set.seed(42)
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.
```

Methods

Preprocessing

dropping the columns “dep_time”, “arr_time”, “air_time”, “arr_delay”, because that leaks the response variable. dropping column “year.x” because it contains redundant information dropping tailnum because it produces too many dummy variable columns for one hot encoding.

```
class(fltrain)
```

```
## [1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

```
fltrain['sched_air_time'] <- fltrain['sched_arr_time']-fltrain['sched_dep_time']
drops <- c("dep_time", "arr_time", "air_time", "arr_delay", "year.x", 'tailnum')
fltrain <- fltrain[ , !(names(fltrain) %in% drops)]
fltrain
```

```
## # A tibble: 200,000 x 38
##    month   day sched_dep_time dep_delay sched_arr_time carrier flight origin
##    <dbl> <dbl>          <dbl>     <dbl>          <dbl> <chr>    <dbl> <chr>
##  1    11     7            600         0            825 WN        1716 LGA
##  2    10    30           1250         2           1400 AA         178 JFK
##  3    12    18           1715         8           2020 DL        1585 LGA
##  4    11    20           2030        -1           2205 WN        3494 EWR
##  5    10    21           1625        -5           1831 DL        2231 LGA
##  6    11     7            900        -8           1157 B6          27 EWR
##  7     9    29           1529       -10           1649 EV        4580 EWR
##  8    12    21           1530        -4           1710 EV        5207 LGA
##  9    11     7           1650         0           1906 9E        2910 JFK
## 10     3    31           1700        -8           1821 US        2183 LGA
## # ... with 199,990 more rows, and 30 more variables: dest <chr>,
## #   distance <dbl>, hour <dbl>, minute <dbl>, time_hour <dttm>, temp <dbl>,
## #   dewp <dbl>, humid <dbl>, wind_dir <dbl>, wind_speed <dbl>, wind_gust <dbl>,
## #   precip <dbl>, pressure <dbl>, visib <dbl>, name <chr>, lat <dbl>,
## #   lon <dbl>, alt <dbl>, tz <dbl>, dst <chr>, tzone <chr>, year.y <dbl>,
## #   type <chr>, manufacturer <chr>, model <chr>, engines <dbl>, seats <dbl>,
## #   speed <dbl>, engine <chr>, sched_air_time <dbl>
```