Classifying U.S. Flights: Regional vs. Mainline Operations

Packages needed

```
library(tidyverse)
library(tidymodels)
library(tidygraph)
library(usmap)
library(sf)
```

Dataset

The `flights_raw` data is a subset of flight records in 2024 from eight airline companies: “AA”, “DL”, “UA”, “WN”, “9E”, “YX”, “OH”, “OO”. We have been using a similar dataset to visualize the flight routes on US map (Week 6 Friday) and look at the flight arrival departure pattern (Week 7 Friday).

```
flights_raw <- read_csv("data/flights_2024.csv.gz")
```

Your goal

From a business perspective, mainstream airlines (AA, DL, UA, WN) may want to identify routes that can be served by regional airlines (9E, YX, OH, OO) to reduce operational costs. **You task is to build a classification model to predict whether a flight is operated by a regional airline or not.**

In the following code, we create a new variable `regional` that indicates whether the flight is operated by a regional airline or not. The regional airlines in our dataset are: “9E”, “YX”, “OH”, “OO”. If the flight is operated by one of these airlines, the `regional` variable should be 1, otherwise it should be 0. Once you have created the `regional` variable, we will drop the `reporting_airline` variable from the dataset, i.e. **you can not use it as a predictor in your classification model.**

To simplify this problem, we only consider the route specific information. That is, we will only keep one record for each unique route (origin-destination pair) and airline combination and we will leave time-specific information (e.g., flight date, flight hour, etc) for more advanced analysis.

```
regionals <- c("9E", "YX", "OH", "OO")
mainlines <- c("AA", "DL", "UA", "WN")
flights_dt <- flights_raw |>
  mutate(regional = as.factor(ifelse(reporting_airline %in% regionals, 1, 0))) |>
  select(-reporting_airline) |>
  select(regional, origin_airport_id:dest_wac, distance, distance_group) |>
  distinct()
```

We will also use the `airports_raw` dataset to identify airports located in the continental US.

```
airports_raw <- read_csv("data/airports.csv")
us_map_sf <-  us_map(regions = 'states') |> filter(!abbr %in% c("AK", "HI"))
```

```
airports_trans <- airports_raw |>
  st_as_sf(coords = c("x", "y"), crs = 4326) |>
  st_transform(st_crs(usmap::us_map())) |>
  mutate(x = st_coordinates(geometry)[,1], y = st_coordinates(geometry)[,2]) |>
  as_tibble() |>
  filter(between(x, -3000000, 3000000), y < 700000) # exclude HI and AK

flight_cont <- flights_dt |>
  filter(origin %in% airports_trans$ident & dest %in% airports_trans$ident)
```

The `flight_cont` dataset will be your entry point for building the classification model.

```
glimpse(flight_cont)
```

Exploratory data analysis

Part 1

The variable `distance` indicates the distance of the flight in miles. Would it be a useful predictor for whether a flight is operated by a regional airline or not? Create a visualization to support your answer.

```
# write your code here
```

Part 2

You may realize that some routes are bidirectional, i.e., there are flights from airport A to airport B and also from airport B to airport A. Run the following code to see an example of such routes.

```
flight_cont |> select(origin, dest) |> head(4)
```

To simplify the analysis, we will only keep one record for each unique route (origin-destination pair). Write code to achieve this and name the resulting dataset `flights_dt_unique_routes`.

```
# write your code here
flights_dt_unique_routes <- flight_cont |> ...
```

Part 3

Before building the classification model, it is a good practice to consider which variables to include as predictors. Some variables may not be useful or may introduce noise into the model. For example, the variables `origin_airport_id`, `dest_airport_id`, `origin_city_market_id`, and `dest_city_market_id` are unique identifiers for airports and cities, and they may not provide useful information for predicting whether a flight is operated by a regional airline or not.

Consider the proposal by student A and B. Are the variables they suggest likely to be useful predictors of whether a flight is operated by a regional airline or not, given the information available in the `flights_dt_unique_routes` dataset? Explain your reasoning.

- Student A: I would like to add a collection of binary variables of the main airport hubs for each airline.
- Student B: I would like to include the geographic location of the origin and destination airports (e.g., latitude and longitude).

Your Answer: explain your answer here.

Part 4

One of the characteristics of regional airlines is that they often serve smaller airports that may not be served by mainstream airlines. Therefore, the importance of the origin and destination airports in the flight network could be a useful predictor for whether a flight is operated by a regional airline or not. Here we will consider using the concept of centrality from network analysis to identify important airports in the flight network. There are various centrality measures, such as degree centrality, closeness centrality, and betweenness centrality. For simplicity, we will use degree centrality, which is defined as the number of direct connections an airport has to other airports.

In a simple example, for three airports A, B, and C, if there are flights from A to B and from A to C, then airport A has a degree centrality of 2, while airports B and C each have a degree centrality of 1. Graphically, we can represent this flight network as B - A - C.

Here is the script to compute the degree centrality of each airport in the flight network and save the results to a CSV file named `airport_centrality.csv` in the `data` folder.

```r
library(tidygraph)
route_count_df <- flights_dt |>
  select(origin, dest) |>
  count(origin, dest, sort = TRUE) |>
  rename(from = origin, to = dest)

route_nodes <- tibble(airport = c(route_count_df$from, route_count_df$to)) |> distinct()

route_graph <- tbl_graph(nodes = route_nodes, edges = route_count_df) |>
  activate(nodes) |>
  mutate(degree = centrality_degree())

centrality_df <- as_tibble(route_graph) |> arrange(-degree)
write_csv(centrality_df, "data/airport_centrality.csv")
```

The centrality of each airport has been pre-computed and saved in the `data/airport_centrality.csv` file. Load this dataset and write the code to include degree centrality of each airport in the `flights_dt_unique_routes` dataset and name the resulting dataset `flight_mod`.

```r
# write your code here
```

Part 5

Other than using centrality measures, can you think of other ways to identify important airports in this data? Propose your ideas below and briefly explain how you would obtain the data.

Your answer: explain your answer here.

Model building and evaluation

Part 6

Now we are ready to build our classification model. Specify the model formula you would use to predict whether a flight is operated by a regional airline or not. Explain your choice of predictors.

```r
# write your code here
# example with the penguins dataset: sex ~ bill_length_mm + bill_depth_mm
```

Your answer: explain your answer here.

Part 7

Assuming you will build three different classification models: logistic regression, decision tree, and random forest. Explain how you would compare the performance of these models for this classification task. Which evaluation metrics would you use and why?

Your answer: explain your answer here.

Part 8

Implement what you described in Part 7 with `tidymodels` for the three models (logistic regression, decision tree, and random forest).

```
# write your code here

# fit a logistic regression


# fit a decision tree


# fit a random forest
```

Part 9

Compare the performance of the three models using

1) the confusion matrix from the three models and
2) the ROC curve of all three models in one plot.

Which model performs the best?

```
# write your code here
```

Part 10

If you work for one of the mainstream airlines (AA, DL, UA, WN), you may be interested in identifying routes that are currently operated by mainstream airlines but can be served by regional airlines. Suggest such routes based on the predictions from your best classification model.

```
# write your code here
```

What are the features of these routes?

Your answer: explain your answer here.