A Study on Domestic Air Travel in the US in 2008 - Guides on R codes

Thu Nguyen

Introduction

This is a guide on running the `R` codes, which were used to generate the results reported. The codes can be found on GitHub here, and were last updated on September 28, 2019.

Getting data

As mentioned in the Data section of every chapters in the main report, the main data - the collection of all domestic air flights in the US in 2008 - can be obtain from Statistical Computing.

Supplementary data such as the Census data of all US cities in 2010 - collection of information such as the city demographic and physical area, among other - can be obtained from 2010 US Census Survey.

Additionally, for convenience, the collection of all airlines (their IATA codes and full names), and that of all airports (IATA codes and full names) can be obtain from Supplemental data. Please note that not all airlines in the 2008 data set were available in the data set. This will be taken care of later in the Data Preparation process.

Running `R` codes

Given the raw data from above, the first step is to prepare data, from the files in the `0 - Data Preparation` folder. Once the data have been prepared, further analysis - Exploratory Data Analysis, Machine Learning: Classification, and Network Analysis - can be run independently.

0. Data Preparation

There are 3 files, and they should be run in this order:

```
[Done] --- Air traffic 2008 - Data Preparation - Airport Info.R
```

- Preparing census information per cities having airports featuring in the 2008 data set, and merge with the airports' geographic information.

```
[Done] --- Air traffic 2008 - Data Preparation - 2008 Data.R
```

- Creating dummy variables and selecting relevant variables to be used for chapters 1: Exploratory Data Analysis and 2: Machine Learning: Classification.

```
[Done] --- Air traffic 2008 - Data Preparation - 2008 Data - Network Analysis Data.R
```

- Preparing data for chapter 3: Network Analysis.

1. Exploratory Data Analysis

There is only 1 file, from which all results in the report can be obtained.

```
[Done] --- Air traffic 2008 - EDA - Statistics and Delays and Cancels.R
```

2. Machine Learning: Classification

There are 5 files, all numbered, and ideally should be run in that order, although 3 files following the 3 approaches *Alternate Cur-offs, Down-Sampling*, and *Cost-sensitive Training* can be run in any order.

```
[Done] --- Air traffic 2008 - ML - 1 - GOF Subset Finding.R
```

- Preparing 10 representative subsets of sizes $10,000$ and $100,000$ each, checked by the *Chi-squared Goodness-of-Fit* test.

```
[Done] --- Air traffic 2008 - ML - 2 - Alternatve Cut-offs.R
```

- Building, evaluating, and selecting the best performing algorithms under the *Alternate Cut-offs* approach.

```
[Done] --- Air traffic 2008 - ML - 3 - Down-Sampling.R
```

- Similar but under the *Down-Sampling* approach.

```
[Done] --- Air traffic 2008 - ML - 4 - Cost-sensitive Training.R
```

- Similar but under the *Cost-sensitive Training* approach.

```
[Done] --- Air traffic 2008 - ML - 5 - Test Set Testing.R
```

- Testing of the best performing algorithms on the common test set.

3. Network Analysis

There are 4 files, all numbered. They can be run independently, but should be run in the order to match that results in the report.

```
[Done] --- Air traffic 2008 - Network Analysis - 1 - US Air Network.R
```

- Descriptive statistics of the network structure of all flights in the US.

```
[Done] --- Air traffic 2008 - Network Analysis - 2 - Top 4 Airlines.R
```

- Descriptive statistics among airlines' networks.

```
[Done] --- Air traffic 2008 - Network Analysis - 3 - Southwest - Link Prediction.R
```

- Predicting if there would be a direct flight between airports offered by Southwest Airlines.

```
[Done] --- Air traffic 2008 - Network Analysis - 4 - Southwest - Graph Modeling.R
```

- Modeling the Southwest Airlines air route network via random graph modeling.