Midterm Project Proposal

Flight Delays Prediction

Xinyi Wang

11/12/2018

1.Introduction

Every year in the United State of America, millions of passengers experience delays in flights, resulting in missing flights connections and distract the valuable time for people. Airlines won't tell you if your flight is likely to be delayed or not.

2.Objective

In this analysis, I try to develop a Multilevel regression model that aims to predict if a flight arrival will be delayed by 15 minutes or more which departed from Atlanta.

3.Data Preparation

The 3 datasets I will use can be found & downloaded by navigating to the following link: -flights.csv & airlines.csv https://www.transtats.bts.gov/DL_SelectFields.asp -weather.csv: https://www.ncdc.noaa.gov/cdo-web/

The raw flight dataset contains 621,462 observations and 29 variables for US flights in January 2018.

```
setwd("/Users/CindyWang/Desktop/678/midterm project")
flights <- read.csv("flights.csv")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
glimpse(flights)
```

```
## Observations: 621,461
## Variables: 29
## $ YEAR               <int> 2018, 2018, 2018, 2018, 2018, 2018, 2018, 2...
## $ MONTH              <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1...
## $ DAY_OF_MONTH       <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27,...
## $ DAY_OF_WEEK        <int> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6...
## $ MKT_CARRIER_FL_NUM <int> 369, 368, 367, 366, 365, 364, 363, 362, 361...
## $ TAIL_NUM           <fct> N26232, N477UA, N13720, N16217, N33714, N80...
## $ ORIGIN             <fct> FLL, SEA, DCA, LAX, JAX, IAH, EWR, HNL, LAS...
## $ ORIGIN_CITY_NAME   <fct> Fort Lauderdale, FL, Seattle, WA, Washingto...
```

```
## $ ORIGIN_STATE_NM    <fct> Florida, Washington, Virginia, California, ...
## $ DEST               <fct> IAH, SFO, IAH, ORD, EWR, PHX, HNL, EWR, SFO...
## $ DEST_CITY_NAME     <fct> Houston, TX, San Francisco, CA, Houston, TX...
## $ DEST_STATE_NM      <fct> Texas, California, Texas, Illinois, New Jer...
## $ DEP_TIME           <int> 602, 614, 828, 641, 1810, 1413, 842, 1623, ...
## $ DEP_DELAY          <dbl> -13, -4, -2, -9, -14, -7, 27, 8, -5, -7, -6...
## $ DEP_DEL15          <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0...
## $ DEP_DELAY_GROUP    <int> -1, -1, -1, -1, -1, -1, 1, 0, -1, -1, -1, -...
## $ DEP_TIME_BLK       <fct> 0600-0659, 0600-0659, 0800-0859, 0600-0659,...
## $ TAXI_OUT           <dbl> 19, 16, 17, 17, 11, 19, 21, 21, 11, 16, 22,...
## $ WHEELS_OFF         <int> 621, 630, 845, 658, 1821, 1432, 903, 1644, ...
## $ WHEELS_ON          <int> 749, 808, 1055, 1230, 2013, 1554, 1454, 618...
## $ TAXI_IN            <dbl> 7, 5, 13, 12, 8, 5, 3, 12, 7, 6, 6, 7, 4, 1...
## $ ARR_TIME           <int> 756, 813, 1108, 1242, 2021, 1559, 1457, 630...
## $ ARR_DELAY          <dbl> -12, -18, 1, -8, -24, -19, 19, -23, -22, -1...
## $ ARR_DELAY_NEW      <dbl> 0, 0, 1, 0, 0, 0, 19, 0, 0, 0, 0, 0, 0, 0, ...
## $ ARR_DEL15          <dbl> 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0...
## $ CANCELLED          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ AIR_TIME           <dbl> 148, 98, 190, 212, 112, 142, 651, 514, 68, ...
## $ DISTANCE           <dbl> 966, 679, 1208, 1744, 820, 1009, 4962, 4962...
## $ X                  <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
```

The raw weather dataset has 729 observations and 22 variables. For future analysis, I will join flight and weather datasets by date. Some of key variable will use are:

WT01-Fog,ice fog,or freezing fog(may include heavy fog)

WT02-Heavy fog or heavy freezing fog(not always distinguished from fog)

WT03-Thunder

WT05-Hail(may include small hail)

WT08-Smoke or haze

SNOW-snow fall

AWND-average wind speed

TAVG-average temperature

PRCP-preceptation

The airlines.csv contains translation between two letter rrier code and names of US airlines. I will aldo join this with flights dataset.

4.Next Steps

-Clean Data

-Exploratory Data Analysis

-Modeling and Prediction

-Model Check