Predicting Flight Delays and Cancellations - Report

Yen Tran - gb3386

05/14/2019

1. Introduction

Flight delays is a serious problem, which costs airlines, passengers, and U.S. economy. A 2007 study by National Center of Excellence for Aviation Operations Research (NEXTOR) estimates that air transportation delays costs total of $32.9 billion (Ball, 2010). The same report mentions over 25 percent of flights delayed (15+ minutes) and cancelled. 43.6% of all flight delays is caused by weather-related conditions (BTS, 2019). A better understanding of how weather affects flights can help mitigate the uncertainty of flight delays and flight cancellations.

This research is motivated by the following questions. Which factors have the most influence on flight on-time performance? Among those, which ones are measurable and have available data? Does weather data improve prediction of flight delays?

2. Data Description

Data

The data used in this paper come from the *nycflight13* R package. In the package, I use 3 datasets: flights, weather, and airlines. The flights dataset contains data for all flights that departed New York City airports in 2013, including on-time performance. The weather dataset contains hourly meteorological data for LaGuadia Airport, John F. Kennedy International Airport, and Newark International Airport. The airlines dataset maps airline names to their carrier codes in the flights dataset.

To answer my research questions, I joined flights with weather on the origin airport and time at the hour. After removing missing values and irrelevant columns, the dimension for the data table is 291,220 rows and 15 columns. Since my goal is to predict whether a given flight is delayed/cancelled or on-time at take-off, I categorized my response variable, **delay**, as 1 when a flight is either dep_delay of more than 15 minutes or cancelled (dep_delay is NA) and 0 otherwise. Potential predictors are the flight information (month, day, day_of_week, carrier, origin, distance, hour) and associated weather features (temp, dewp, humid, wind_speed, precip, pressure, visib). Among of all predictors, day_of_week, carrier, and origin are categorical variables; while the rest are quantitative variables.

Data Exploration & Visualization

To understand about the relationship of predictors and response variable, I perform an exploratory analysis. There is a correlation of 0.91 between arrival delays and departure delays, which means a flight that is delayed at take-off is more likely to arrive late than scheduled. However, the correlation is not exactly 1 because a flight could pick up more speed on air to avoid late arriving as much as possible or other factors (such as weather, busy traffic) at the destination might affect the scheduled arrival. Since departure delay is associated to cause arrival delay, my model will predict the probability of flight delayed at take-off. Before building the model, I did an exploratory analysis of the proportion of delayed and cancelled flights in my dataset. There are approximately 24% delayed and cancelled flights (81,169 out of 291,220 flights) in my dataset. Among those delayed flights, the months (July, June, and December) contributed the highest proportion compared to other months in the data (figure 1). Thus, I used the date of the flights (month, day, day of the week) as predictors to account for the variability of flight delays. Besides that, the number of delayed flights are highest between 3 - 7 pm. The explanation for that could be from an accumulation of earlier flight delays.

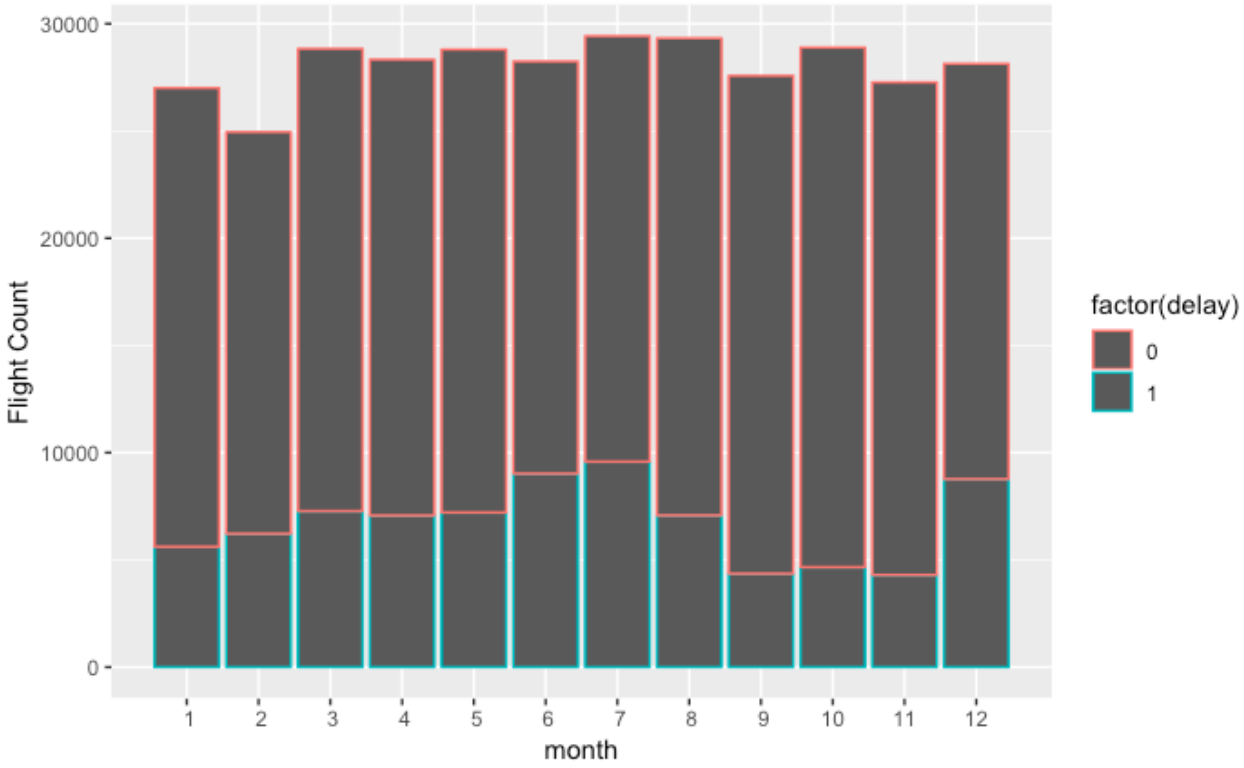

Figure 1: Distribution of flight delays and cancellation group by month.

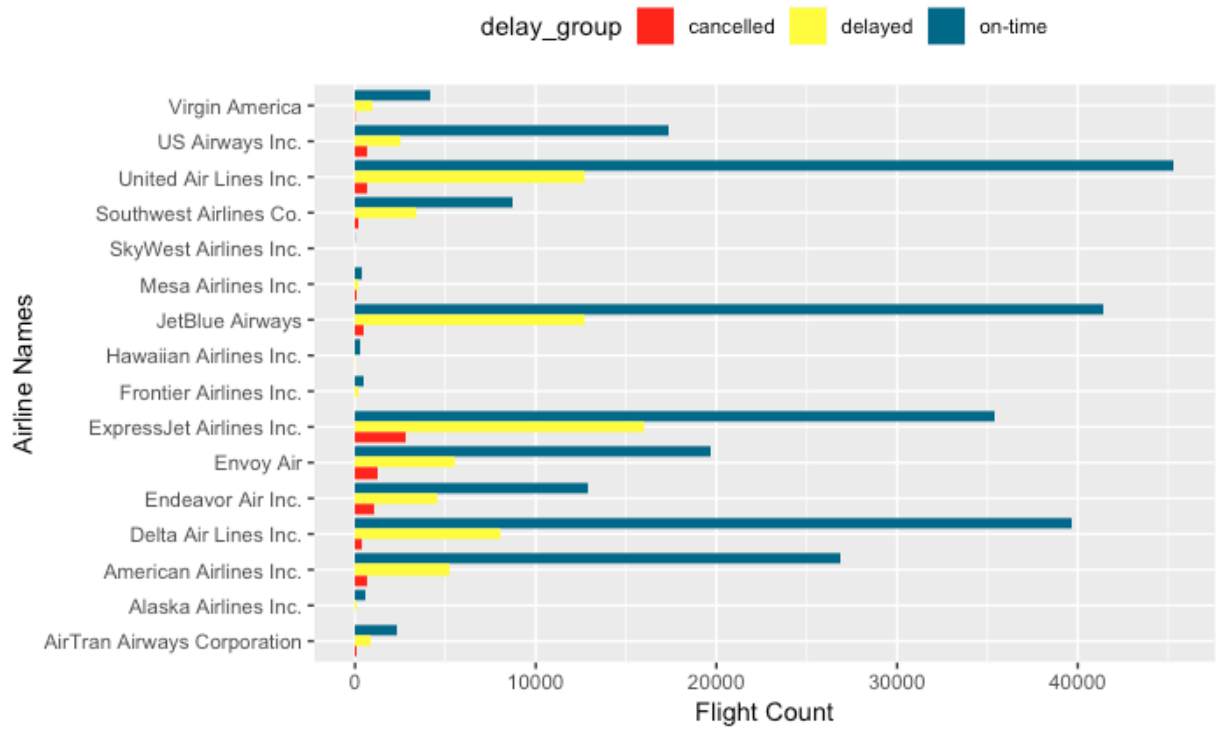

Figure 2: Number of flights categorized by different airlines.

	delay	month	day	distance	hour	temp	dewp	humid	wind_speed	precip	pressure	visib
delay	1.000	-0.024	0.006	-0.043	0.226	0.056	0.080	0.079	0.053	0.058	-0.126	-0.067
month	-0.024	1.000	0.008	0.022	0.001	0.264	0.268	0.075	-0.133	0.000	0.090	0.039
day	0.006	0.008	1.000	0.003	-0.010	0.013	-0.005	-0.037	-0.012	-0.003	0.009	0.030
distance	-0.043	0.022	0.003	1.000	-0.019	0.008	0.022	0.036	0.014	0.000	0.011	-0.005
hour	0.226	0.001	-0.010	-0.019	1.000	0.098	0.008	-0.162	0.120	0.010	-0.078	0.058
temp	0.056	0.264	0.013	0.008	0.098	1.000	0.892	0.060	-0.150	-0.026	-0.246	0.057
dewp	0.080	0.268	-0.005	0.022	0.008	0.892	1.000	0.496	-0.245	0.051	-0.278	-0.132
humid	0.079	0.075	-0.037	0.036	-0.162	0.060	0.496	1.000	-0.261	0.199	-0.160	-0.481
wind_speed	0.053	-0.133	-0.012	0.014	0.120	-0.150	-0.245	-0.261	1.000	0.005	-0.215	0.105
precip	0.058	0.000	-0.003	0.000	0.010	-0.026	0.051	0.199	0.005	1.000	-0.088	-0.358
pressure	-0.126	0.090	0.009	0.011	-0.078	-0.246	-0.278	-0.160	-0.215	-0.088	1.000	0.108
visib	-0.067	0.039	0.030	-0.005	0.058	0.057	-0.132	-0.481	0.105	-0.358	0.108	1.000

Figure 3: Correlation matrix.

```
temp          2.658e-02  3.365e-03   7.899 2.81e-15 ***
dewp         -2.287e-02  3.615e-03  -6.326 2.51e-10 ***
```

Figure 4: Collinearity affecting coefficient estimates.

I also looked into the distribution of categorized flights for all airlines (figure 2). My statistics show Mesa Airline with the highest proportion of flight delays (35.94% of its total flights), but the graph shows the number of flights that Mesa Airline operated is very small compared to other airlines. I looked up and found that Mesa Airline is a regional airline. Moving down to the second highest rank in the list, ExpressJet Airline (34.71% of its total flights) is confirmed to have the highest count of flight delays and cancellations. ExpressJet operates scheduled United Express flights to destinations in the U.S. (East, Midwest, and South regions), Canada and Mexico. Thus, the delays could be caused by bad winter storms in winter month and storms in summer months contributing to 24% of total flight delays.

Figure 3 below shows the correlation matrix of the quantitative predictors in my dataset. There is a strong correlation between temperature and dewpoint (0.893). A consequence of highly correlated predictor variables is that the coefficients in the regression model are of the opposite sign than expected. From the full model output (figure 4), I suspect that temp and dewp have opposite signs even they have a positive correlation. I decide to remove dewp from the model because I want to see the effect of temperature on the probability of flight delays. From figure 1, I suspect that either low temperature or high temperature will increase the probability of flight delays, suggesting to add a quadratic term in **temp** to my model. Not only the quadratic term is significant in the summary output (figure 5), but the prediction rate results also increase comparing to the one, not including the quadratic term.

3. Methods

Because the response variable is a binary categorical variable, I applied Logistic Regression model to predict the likelihood of a given flight is delayed. For model validation, I used cross-validation technique to split 70:30 ratio of my dataset to training and validation sets. The fitted model in probability form is given by:

$$\hat{p}(\boldsymbol{x}) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X_1+...+\hat{\beta}_{34} X_{34}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X_1+...+\hat{\beta}_{34} X_{34}}}$$

where $\mathbf{x} = (X_1, X_2, ..., X_{34})$ are 34 predictors in my model.

We wish to test:
H_0 : logistic regression model is appropriate
H_A : logistic model is inappropriate so a saturated model is needed

Since p-value is 1, we fail to reject null hypothesis. Thus, the deviance goodness-of-fit test finds that the logistic regression model is an adequate fit overall for the training data.

4. Results

The results from the summary output in figure 5 show all weather features are significant to the likelihood of flight delays since all the p-values are below 0.001. Also, looking only to the coefficient estimates of associated with temperature while holding other predictors constant, we can see the relationship between logit function of the probability of flight delays and the temperature is quadratic $y = -0.0664 * temp + 0.0006 * temp^2 + c$ (a u-shapes curve). Graphing this equation, I see that y is at a minimum when X is 55. We can interpret as when the temperature is around 55*F degree, a given flight is estimated with the lowest chance of being delayed or cancelled; while extreme temperature (low or high) will have more chance of being delayed.

The cross-validation results have a high accuracy rate (78.74%) and sensitivity (97.58%), but low specificity (12.87%) because my response variable is imbalanced. Thus, I evaluate my model using AUC (Area under the ROC curve) metric with the result of 0.7144.

5. Discussion

While fixing other sources of variability in my model, weather features have a significant impact on the likelihood of flight delays and cancellations. Even though the chi-square goodness-of-fit test confirms my model overall is appropriated, perform model diagnostics on residuals of multiple logistic regression for more than 3 predictors is limited. For future work, I want to analyze the significant impact of each weather feature invidually and/ their combination factors to the likelihood of flight delays. Besides that, I also want to find different machine learning algorithms such as XGBoost, Random Forest, or Neural Network to improve the prediction rates comparing to the model in this project.

6. References

Ball, Michael, et al. “A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States.” (2010, October). Retrieved May 12, 2019, from https://isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf

Route Maps. https://www.expressjet.com/about/route-maps/

Sheather, S. J. (n.d.). Logistic Regression. In A Modern Approach to Regression with R.

Understanding the Reporting of Causes of Flight Delays and Cancellations. (2019, March 29). Retrieved May 12, 2019, from https://www.bts.gov/topics/airlines-and-airports/understanding-reporting-causes-flight-delays-and-cancellation

7. Code Appendix

Please refer to the Code file.

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    2.344e+01  8.903e-01  26.327  < 2e-16 ***
month         -1.475e-02  1.804e-03  -8.179 2.87e-16 ***
day            3.351e-03  6.364e-04   5.266 1.40e-07 ***
day_of_week.L  2.448e-02  1.552e-02   1.578 0.114637
day_of_week.Q -1.529e-01  1.556e-02  -9.827  < 2e-16 ***
day_of_week.C -1.943e-01  1.499e-02 -12.961  < 2e-16 ***
day_of_week^4 -2.347e-01  1.450e-02 -16.186  < 2e-16 ***
day_of_week^5  1.267e-01  1.440e-02   8.794  < 2e-16 ***
day_of_week^6  4.657e-02  1.449e-02   3.214 0.001308 **
carrierAA     -5.052e-01  2.994e-02 -16.875  < 2e-16 ***
carrierAS     -1.006e+00  1.500e-01  -6.710 1.95e-11 ***
carrierB6     -2.414e-01  2.579e-02  -9.361  < 2e-16 ***
carrierDL     -6.303e-01  2.790e-02 -22.593  < 2e-16 ***
carrierEV      2.649e-01  2.902e-02   9.128  < 2e-16 ***
carrierF9      2.867e-02  1.175e-01   0.244 0.807161
carrierFL      4.531e-02  5.845e-02   0.775 0.438238
carrierHA     -1.035e+00  2.893e-01  -3.579 0.000345 ***
carrierMQ     -1.901e-01  3.048e-02  -6.237 4.46e-10 ***
carrierOO      1.749e-01  4.987e-01   0.351 0.725781
carrierUA     -3.101e-01  2.929e-02 -10.586  < 2e-16 ***
carrierUS     -7.377e-01  3.620e-02 -20.378  < 2e-16 ***
carrierVX     -3.836e-01  5.330e-02  -7.197 6.14e-13 ***
carrierWN      1.361e-01  3.737e-02   3.642 0.000270 ***
carrierYV     -2.606e-02  1.205e-01  -0.216 0.828811
originJFK     -2.175e-01  1.937e-02 -11.229  < 2e-16 ***
originLGA     -8.005e-02  1.713e-02  -4.672 2.98e-06 ***
hour           1.286e-01  1.274e-03 100.921  < 2e-16 ***
temp          -6.643e-02  2.002e-03 -33.183  < 2e-16 ***
I(temp^2)      6.154e-04  1.694e-05  36.323  < 2e-16 ***
humid          1.592e-02  3.869e-04  41.162  < 2e-16 ***
wind_speed     2.533e-02  1.143e-03  22.156  < 2e-16 ***
precip         2.266e+00  4.043e-01   5.605 2.09e-08 ***
pressure      -2.497e-02  8.580e-04 -29.104  < 2e-16 ***
visib         -3.661e-02  4.190e-03  -8.739  < 2e-16 ***
---
```

Figure 5: Regression summary output for final model.