Airline Ontime Model Report

Mark R. Locatelli

April 8, 2015

1 Introduction

In this post, we'll use logistic regression to predict delayed flights. This analysis is conducted using a public data set that can be obtained here:

- https://catalog.data.gov/dataset/airline-on-time-performance-and-causes-of-flight-dela
- http://stat-computing.org/dataexpo/2009/the-data.html

Other websites of interest:

- http://users.stat.umn.edu/~geyer/Sweave/
- http://www.datasciencecentral.com/profiles/blogs/predicting-flights-delay-using-supervi

Note: This is a common data set in the machine learning community to test out algorithms and models given it's publicly available and have sizable data.

In this report, we will look at small sample snapsot(flights in and out of HI December 2014).

Let's look at a summary of the data:

```
> summary(AirlineDataSummary)

 Day of the Week    Carrier       Origin City     Origin State    Destination City
 Min.   :1.00    Min.   :1.000   Min.   : 1.00   Min.   : 1.000   Min.   : 1.00
 1st Qu.:2.00    1st Qu.:3.000   1st Qu.: 9.00   1st Qu.: 6.000   1st Qu.: 9.00
 Median :4.00    Median :4.000   Median :11.00   Median : 6.000   Median :11.00
 Mean   :3.85    Mean   :3.689   Mean   :12.46   Mean   : 5.986   Mean   :12.46
 3rd Qu.:6.00    3rd Qu.:4.000   3rd Qu.:14.00   3rd Qu.: 6.000   3rd Qu.:14.00
 Max.   :7.00    Max.   :6.000   Max.   :28.00   Max.   :16.000   Max.   :28.00
```

```
 Destination State  Delay (min)       Delay 15 min?    Arrival Dealy (min)
 Min.   : 1.000    Min.   : -30.000   Min.   :0.0000   Min.   : -71.000
 1st Qu.: 6.000    1st Qu.:  -6.000   1st Qu.:0.0000   1st Qu.:  -9.000
 Median : 6.000    Median :  -2.000   Median :0.0000   Median :  -2.000
 Mean   : 5.985    Mean   :   6.662   Mean   :0.1305   Mean   :   3.767
 3rd Qu.: 6.000    3rd Qu.:   4.000   3rd Qu.:0.0000   3rd Qu.:   8.000
 Max.   :16.000    Max.   :1403.000   Max.   :1.0000   Max.   :1423.000
                   NA's   :33         NA's   :33       NA's   :76
 Arrival Delay 15 min? Distance (mi)
 Min.   :0.0000       Min.   :  84
 1st Qu.:0.0000       1st Qu.: 121
 Median :0.0000       Median :2384
 Mean   :0.1643       Mean   :1590
 3rd Qu.:0.0000       3rd Qu.:2603
 Max.   :1.0000       Max.   :4983
 NA's   :76
```

Data available includes the following elements:

- Departure Time
- Carrier
- Destination
- Distance
- Flight Number
- Day of the Week
- Day of the Month
- Tail Number
- Flight Status

The goal here is to identify flights that are likely to be delayed. In the machine learning literature this is called a binary classification using supervised learning. We are bucketing flights into delayed or ontime(hence binary classification). (Note: Prediction and classification are two main big goals of data mining and data science. On a deeper philosophical level, they are two sides of the same coin. To classify things is predicting as well if you think about it.)

Logistic regression provides us with a probability of belonging to one or the two cases (delayed or ontime). Since probability ranges from 0 to 1, we will use the 0.5 cutoff to determine which bucket to put our probability estimates in. If the probability estimate from the logistic regression is equal to or greater tha 0.5

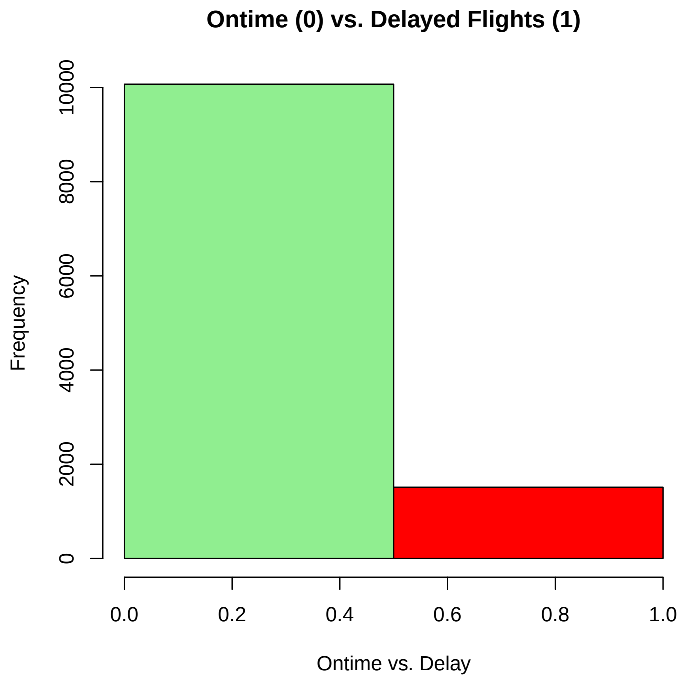

Figure 1: Simple histogram of delays vs. ontime flights

then we assign it to be ontime else it's delayed. We'll explain the theory behind logistic regression in another meeting.

But before we start our modeling exercise, it's good to take a visual look at what we are trying to predict to see what it looks like. Since we are trying to predict delayed flights with historical data, let's do a simple histogram plot to see the distribution of flights delayed vs. ontime (See Figure 1).

We see that most flights are ontime (86.95%, as expected). But we need to have delayed flights in our dataset in order to train the machine to learn from this delayed subset to predict if future flights will be delayed.

Table 1: Carriers Distribution in the Data Set

	0	1
AA	628	191
AS	1242	194
DL	704	76
HA	5944	465
UA	1248	539
US	307	47

Table 2: Day of the Week in the Data Set

	0	1
1	1623	240
2	1445	355
3	1579	209
4	1342	152
5	1362	216
6	1347	200
7	1375	140

2 Exploratory Data Analysis (EDA)

The next step in predictive analytics is to explore our underlying data. Let's look at a few tables of our explantory variables to see how they look against Delayed Flights.

3 Data Transformations And Pre-Processing

One of the main steps in predictive analytics is data transformation. Data is never in the format you want. Transformations of the data are requirede to get it the way we need them (either because the data is dirty, not of the type we want, out of bounds, and a host of other reasons).

This first transformation we'll need to do is to convert the categorical variables into dummy variables.

The categorical variables of interests are: 1) Destination (state) 2) Origin (state) 3) Day of the Week. For simplicity of model building, we'll NOT use Day of the Month, because of the combinatorial explosion in number of dummy variables. The reader is free to do this as an exercise on his/her own. :)

Table 3: Flight Destinations in the Data Set

	0	1
Anchorage, AK	38	7
Atlanta, GA	30	1
Bellingham, WA	31	8
Chicago, IL	33	12
Dallas/Fort Worth, TX	76	31
Denver, CO	74	5
Guam, TT	20	10
Hilo, HI	475	42
Honolulu, HI	3431	547
Houston, TX	25	6
Kahului, HI	1703	224
Kona, HI	858	108
Las Vegas, NV	68	7
Lihue, HI	887	99
Los Angeles, CA	899	146
New York, NY	39	1
Newark, NJ	23	8
Oakland, CA	117	21
Pago Pago, TT	9	2
Phoenix, AZ	193	15
Portland, OR	143	27
Sacramento, CA	55	7
Salt Lake City, UT	30	1
San Diego, CA	88	19
San Francisco, CA	289	91
San Jose, CA	120	21
Seattle, WA	306	44
Washington, DC	13	2

Table 4: Flight Origins in the Data Set

	0	1
Anchorage, AK	38	7
Atlanta, GA	31	0
Bellingham, WA	34	5
Chicago, IL	29	16
Dallas/Fort Worth, TX	57	51
Denver, CO	36	43
Guam, TT	22	8
Hilo, HI	478	37
Honolulu, HI	3621	362
Houston, TX	18	13
Kahului, HI	1671	249
Kona, HI	868	97
Las Vegas, NV	70	5
Lihue, HI	908	78
Los Angeles, CA	825	223
New York, NY	37	3
Newark, NJ	18	13
Oakland, CA	130	7
Pago Pago, TT	8	3
Phoenix, AZ	172	36
Portland, OR	160	11
Sacramento, CA	57	5
Salt Lake City, UT	29	2
San Diego, CA	90	16
San Francisco, CA	227	153
San Jose, CA	122	20
Seattle, WA	310	40
Washington, DC	7	9

This is done usinge code like:

```
M <- cbind(AirlineData, rep(0, nrow(AirlineData)))
names(M)[ncol(M)] <- "origin.AK"
M$origin.AK[M$ORIGIN_STATE_NM == "Alaska"] <- 1
```

4 Logistic Regression

The commands for running the logistic regressions look like:

```
day.model <- glm(formula = M$DEP_DEL15 ~ 1 +  M$day.2 + M$day.3 + M$day.4 + M$day.5 + M$day.
```

Three seperate regressions were run for each of the three categorical vairables.

```
> summary(day.model)

Call:
glm(formula = M$DEP_DEL15 ~ 1 + M$day.2 + M$day.3 + M$day.4 +
    M$day.5 + M$day.6 + M$day.7, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6628  -0.5262  -0.4986  -0.4632   2.1824

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.911393   0.069158 -27.638  < 2e-16 ***
M$day.2      0.507646   0.091059   5.575 2.48e-08 ***
M$day.3     -0.110820   0.100998  -1.097 0.272533
M$day.4     -0.266643   0.110027  -2.423 0.015374 *
M$day.5      0.069962   0.100730   0.695 0.487342
M$day.6      0.004075   0.102592   0.040 0.968317
M$day.7     -0.373174   0.112473  -3.318 0.000907 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8975.2  on 11584  degrees of freedom
Residual deviance: 8876.2  on 11578  degrees of freedom
  (33 observations deleted due to missingness)
AIC: 8890.2

Number of Fisher Scoring iterations: 4
```

```
> summary(dest.model)

Call:
glm(formula = M$DEP_DEL15 ~ 1 + M$dest.AK + M$dest.AZ + M$dest.CA +
    M$dest.CO + M$dest.GA + M$dest.IL + M$dest.NV + M$dest.NJ +
    M$dest.NY + M$dest.OR + M$dest.TX + M$dest.TER + M$dest.UT +
    M$dest.VA + M$dest.WA, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8322  -0.5097  -0.5097  -0.5097   2.7162

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.97544    0.03341 -59.123  < 2e-16 ***
M$dest.AK    0.28377    0.41266   0.688  0.49167
M$dest.AZ   -0.57920    0.27012  -2.144  0.03201 *
M$dest.CA    0.33820    0.07094   4.767 1.87e-06 ***
M$dest.CO   -0.71919    0.46328  -1.552  0.12057
M$dest.GA   -1.42575    1.01612  -1.403  0.16057
M$dest.IL    0.96384    0.33875   2.845  0.00444 **
M$dest.NV   -0.29816    0.39835  -0.748  0.45417
M$dest.NJ    0.91939    0.41182   2.233  0.02558 *
M$dest.NY   -1.68810    1.01001  -1.671  0.09465 .
M$dest.OR    0.30843    0.21248   1.452  0.14661
M$dest.TX    0.97124    0.19505   4.979 6.38e-07 ***
M$dest.TER   1.09305    0.34487   3.170  0.00153 **
M$dest.UT   -1.42575    1.01612  -1.403  0.16057
M$dest.VA    0.10364    0.76029   0.136  0.89157
M$dest.WA    0.10660    0.15269   0.698  0.48508
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8975.2  on 11584  degrees of freedom
Residual deviance: 8891.5  on 11569  degrees of freedom
  (33 observations deleted due to missingness)
AIC: 8923.5

Number of Fisher Scoring iterations: 5

> summary(origin.model)

Call:
glm(formula = M$DEP_DEL15 ~ 1 + M$origin.AK + M$origin.AZ + M$origin.CA +
    M$origin.CO + M$origin.GA + M$origin.IL + M$origin.NV + M$origin.NJ +
```

```
    M$origin.NY + M$origin.OR + M$origin.TX + M$origin.TER +
    M$origin.UT + M$origin.VA + M$origin.WA, family = binomial)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.286  -0.455  -0.455  -0.455   2.343

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.21582    0.03671 -60.361  < 2e-16 ***
M$origin.AK    0.52414    0.41294   1.269 0.204339
M$origin.AZ    0.65184    0.18692   3.487 0.000488 ***
M$origin.CA    0.98554    0.06630  14.866  < 2e-16 ***
M$origin.CO    2.39350    0.22887  10.458  < 2e-16 ***
M$origin.GA  -12.35025  158.54539  -0.078 0.937910
M$origin.IL    1.62111    0.31358   5.170 2.34e-07 ***
M$origin.NV   -0.42324    0.46436  -0.911 0.362062
M$origin.NJ    1.89039    0.36582   5.168 2.37e-07 ***
M$origin.NY   -0.29649    0.60142  -0.493 0.622027
M$origin.OR   -0.46146    0.31386  -1.470 0.141483
M$origin.TX    2.05721    0.17409  11.817  < 2e-16 ***
M$origin.TER   1.21251    0.35439   3.421 0.000623 ***
M$origin.UT   -0.45833    0.73200  -0.626 0.531228
M$origin.VA    2.46713    0.50529   4.883 1.05e-06 ***
M$origin.WA    0.18184    0.16272   1.118 0.263777
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8975.2  on 11584  degrees of freedom
Residual deviance: 8519.4  on 11569  degrees of freedom
  (33 observations deleted due to missingness)
AIC: 8551.4

Number of Fisher Scoring iterations: 13
```

Based on this, a selection of significant variables were selected to craft a final version of this model.

```
> summary(sigvar.model)

Call:
glm(formula = DEP_DEL15 ~ 1 + day.2 + day.4 + day.7 + dest.CA +
    dest.IL + dest.NJ + dest.TX + dest.TER + origin.AZ + origin.CA +
    origin.CO + origin.IL + origin.NJ + origin.TX + origin.TER +
    origin.VA, family = binomial, data = M)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5074  -0.5885  -0.3860  -0.3393   2.4479

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.55958    0.05208 -49.147  < 2e-16 ***
day.2        0.56275    0.07240   7.772 7.70e-15 ***
day.4       -0.26657    0.09584  -2.781 0.005412 **
day.7       -0.38521    0.09880  -3.899 9.66e-05 ***
dest.CA      0.89382    0.07779  11.491  < 2e-16 ***
dest.IL      1.52424    0.34293   4.445 8.80e-06 ***
dest.NJ      1.47838    0.41639   3.550 0.000385 ***
dest.TX      1.53314    0.19912   7.700 1.37e-14 ***
dest.TER     1.63015    0.34910   4.670 3.02e-06 ***
origin.AZ    0.99257    0.19000   5.224 1.75e-07 ***
origin.CA    1.30460    0.07201  18.118  < 2e-16 ***
origin.CO    2.74580    0.23251  11.809  < 2e-16 ***
origin.IL    1.94843    0.31761   6.135 8.53e-10 ***
origin.NJ    2.22253    0.37040   6.000 1.97e-09 ***
origin.TX    2.39465    0.17789  13.461  < 2e-16 ***
origin.TER   1.49885    0.35808   4.186 2.84e-05 ***
origin.VA    2.87676    0.51111   5.629 1.82e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8975.2  on 11584  degrees of freedom
Residual deviance: 8244.2  on 11568  degrees of freedom
  (33 observations deleted due to missingness)
AIC: 8278.2

Number of Fisher Scoring iterations: 5
```

5 Conclusions

In this example, we've looked at a publicly available data source, run some simple analyses, and used the `Sweave` tools in R to create a report of the results we've generated. Looking at data flying in and out of Hawaii in December: flights on Tuesdays are more likely to be delayed ($Coef = 0.5627$), and flights on Sundays are more likely to be on time ($Coef = -0.3852$); flights to the outlying territories (Guam, American Samoa, etc.) are most likely to be delayed ($Coef = 1.6301$); as are flights from Virginia (2.8768).

Table 5: Flights by Probability of Delay

	ProbDelay	N
1	0.05	919
2	0.06	908
3	0.07	4101
4	0.11	244
5	0.12	1095
6	0.12	32
7	0.13	242
8	0.14	27
9	0.16	1098
10	0.16	243
11	0.17	122
12	0.18	243
13	0.19	4
14	0.19	3
15	0.19	6
16	0.20	18
17	0.21	4
18	0.21	4
19	0.21	4
20	0.21	5
21	0.22	18
22	0.22	1095
23	0.23	4
24	0.25	289
25	0.25	18
26	0.26	27
27	0.26	27
28	0.26	80
29	0.27	27
30	0.27	6
31	0.28	26
32	0.29	5
33	0.33	4
34	0.33	294
35	0.35	27
36	0.35	4
37	0.37	18
38	0.37	5
39	0.38	7
40	0.38	7
41	0.39	22
42	0.39	18
43	0.41	7
44	0.42	18
45	0.45	10
46	0.46	81
47	0.48	10
48	0.48	4
49	0.49	7
50	0.51	2
51	0.55	48
52	0.56	5
53	0.58	8

Table 6: Actual Delays vs Predicted Delays

	ProbDelay	Actual Percent Delayed
1	0.05	0.04
2	0.06	0.03
3	0.07	0.07
4	0.11	0.12
5	0.12	0.15
6	0.12	0.12
7	0.13	0.15
8	0.14	0.11
9	0.16	0.16
10	0.16	0.17
11	0.17	0.17
12	0.18	0.24
13	0.19	0.00
14	0.19	0.00
15	0.19	0.50
16	0.20	0.17
17	0.21	0.75
18	0.21	0.00
19	0.21	0.00
20	0.21	0.40
21	0.22	0.17
22	0.22	0.22
23	0.23	0.00
24	0.25	0.22
25	0.25	0.17
26	0.26	0.30
27	0.26	0.19
28	0.26	0.32
29	0.27	0.30
30	0.27	0.33
31	0.28	0.31
32	0.29	0.20
33	0.33	0.25
34	0.33	0.27
35	0.35	0.41
36	0.35	0.50
37	0.37	0.39
38	0.37	0.40
39	0.38	0.43
40	0.38	0.29
41	0.39	0.23
42	0.39	0.39
43	0.41	0.57
44	0.42	0.44
45	0.45	0.80
46	0.46	0.47
47	0.48	0.50
48	0.48	0.50
49	0.49	0.29
50	0.51	0.50
51	0.55	0.50
52	0.56	0.40
53	0.58	0.50

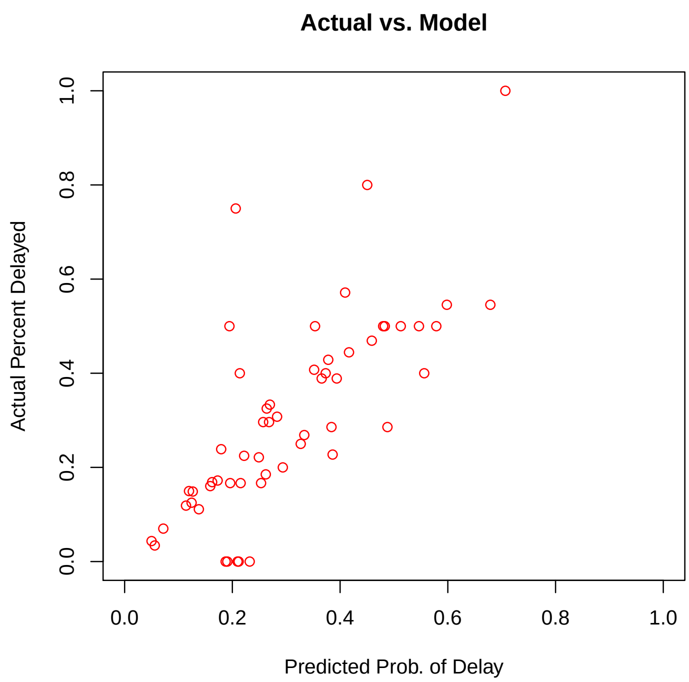

Figure 2: Model's predicted probability of delay vs. actual percentage delayed.