678 Midterm Project

Flights Arrival Delay Prediction

Xinyi Wang

11/17/2018

1 Abstract

The inconveniences resulted from flight delays have been a long-time challenge for passengers, airports and airlines. This report establishes two models under the assumptions of logistic regression, discussing the "Airline on-time performance" data and applying The R Programming Language to predict flight arrival delays. By applying Boruta feature selection method and checking correlation bewteen all varibales, the models finally used 10 variables among 33 variables. After comparing with Multilevel Logistic Regression Model, the Logistic Regression which is no pooling model has higher accuracy with 67.32%. Some of model checking method such as Binned Residual Plot are also support a fair performance of the model.

2 Introduction

2.1 Background

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. According to the study conducted by the U.S. Federal Aviation Administration (FAA) in 2010, 32.9 billion US$ was borne by the American passengers and airlines as a result of flight delays. Airlines won't tell you if your flight is likely to be delayed or not. In this analysis I try to develop a model that aims to predict if a flight arrival will be delayed by 15 minutes or more which departed from Atlanta.

3 Method

3.1 Data source

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/

3.1.1 Read & Clean Data

```r
str(train)
```

```
## 'data.frame':    29417 obs. of  23 variables:
##  $ DAY_OF_MONTH     : num  1 2 3 4 5 6 8 9 10 11 ...
##  $ DAY_OF_WEEK      : Factor w/ 7 levels "1","2","3","4",..: 7 1 2 3 4 5 7 1 2 3 ...
##  $ OP_UNIQUE_CARRIER: chr  "AA" "AA" "AA" "AA" ...
##  $ ORIGIN_STATE_ABR : Factor w/ 52 levels "AK","AL","AR",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ DEST_STATE_ABR   : chr  "FL" "FL" "FL" "FL" ...
##  $ DEP_TIME         : int  820 820 802 804 801 808 813 754 753 751 ...
##  $ DEP_DELAY        : num  10 10 -8 -6 -9 -2 3 -6 -7 -9 ...
##  $ DEP_DEL15        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ARR_DELAY        : num  7 2 -11 19 9 -4 4 -24 -20 -7 ...
##  $ ARR_DEL15        : num  0 0 0 1 0 0 0 0 0 0 ...
##  $ AIR_TIME         : num  91 85 80 94 87 94 87 76 84 86 ...
##  $ DISTANCE         : num  594 594 594 594 594 594 594 594 594 594 ...
##  $ WT01             : num  1 1 1 0 0 1 0 0 0 0 ...
##  $ WT02             : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ WT03             : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ WT05             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ WT08             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SNOW             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ AWND             : num  7.83 7.61 8.5 12.97 6.04 ...
##  $ TAVG             : int  47 52 59 52 42 40 24 33 40 59 ...
##  $ PRCP             : num  0.84 2.26 0 0 0 0.47 0 0 0 0 ...
##  $ Description      : Factor w/ 1652 levels "21 Air LLC","40-Mile Air",..: 281 281 281 281 281 281 28
##  $ total_delay      : num  17 12 -19 13 0 -6 7 -30 -27 -16 ...
##  - attr(*, "na.action")= 'omit' Named int  7 38 65 145 220 261 1146 1737 2307 2404 ...
##   ..- attr(*, "names")= chr  "7" "38" "65" "145" ...
```

The raw flights dataset contains 450,017 observations and 28 variables for US flights in January 2017. The raw weather dataset has 729 observations and 22 variables. For future analysis, I will join flights and weather datasets by date. The airlines.csv contains translation between two letter rrier code and names of US airlines. I will aldo join this with flights dataset. Some of the columns are useless to our data analysis, so we NULL 'em to get my train dataset which is flights history from Jan. 2017. I applied same method for test data, which is flights history from Jan. 2018. Above displays the basic structure for my train dataset after remove NAs.

3.1.2 EDA

Fig 1. Distribution of total delay

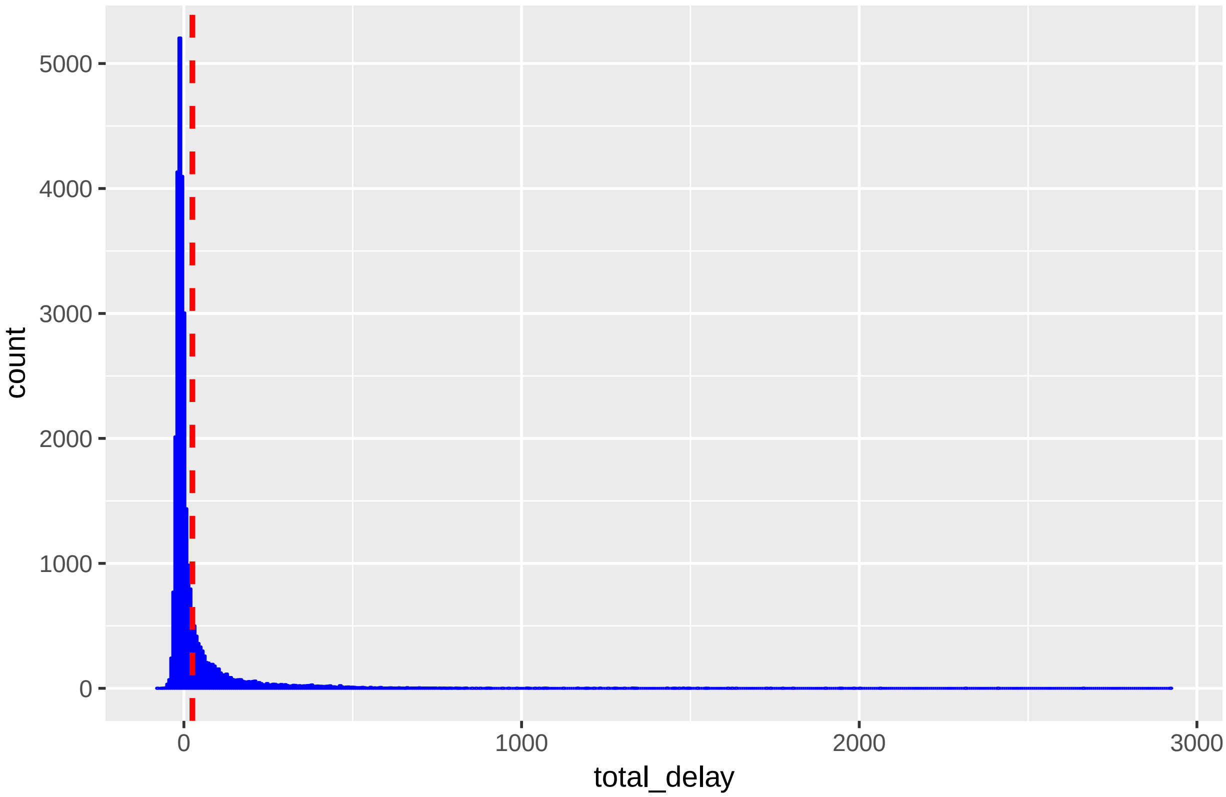


Fig 1 shows that the distribution of the outcome is right skewed, it has long tail in the high values. And we can see from the plot that most of flights are delay less than 25min but there are some outliers(flight delay alomost 3000min). Under this situation, I choose "ARR_DEL15" (a binary number indicates weather a flight delay 15 min or not) as my outcome instead of using total delay time.

Fig 2. Probability of Delay across Airlines

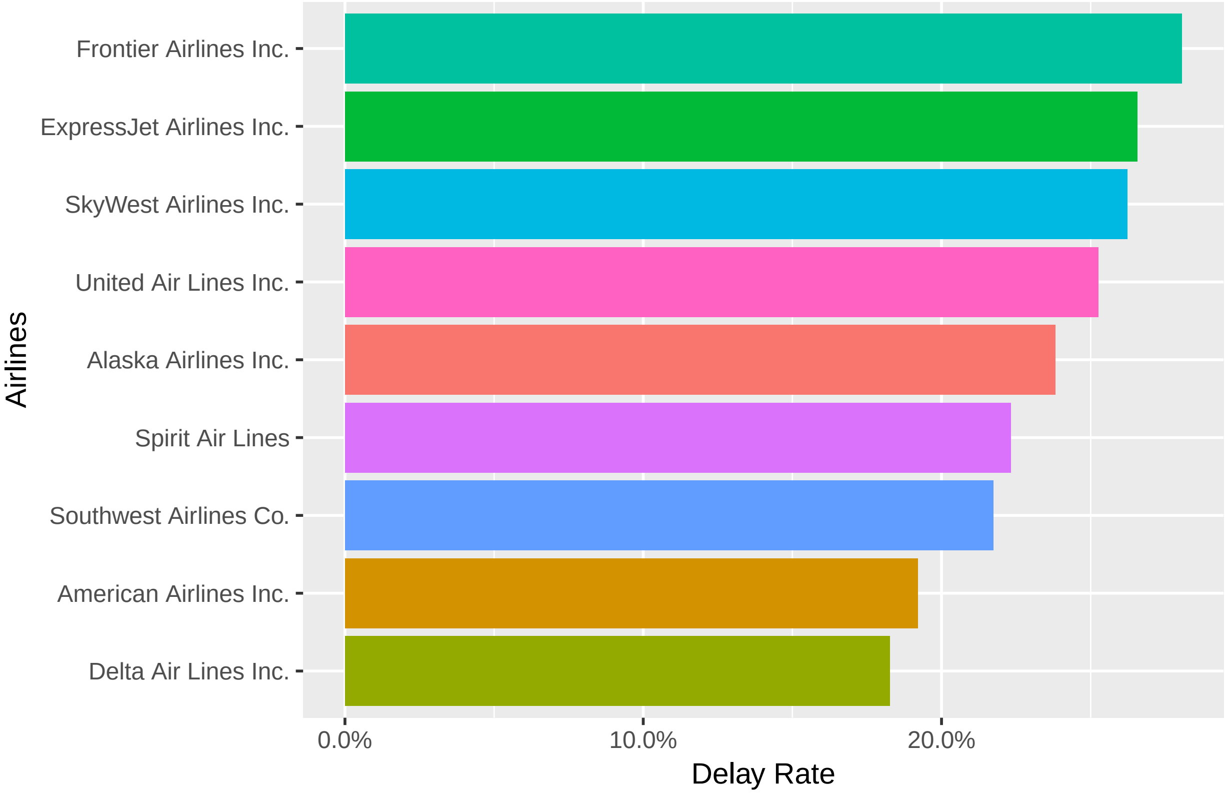


Fig 2 shows the probability of delay across different airlines. The delay rate is calculate by number of delay($>$15min) / total flights of specific airline. It shows that Frontier Airline has the highest delay rate among those 9 airlines. On the contrary, Delta Airline has the lowest delay rate.

Fig 3.Pie Chart of Delay Distribution

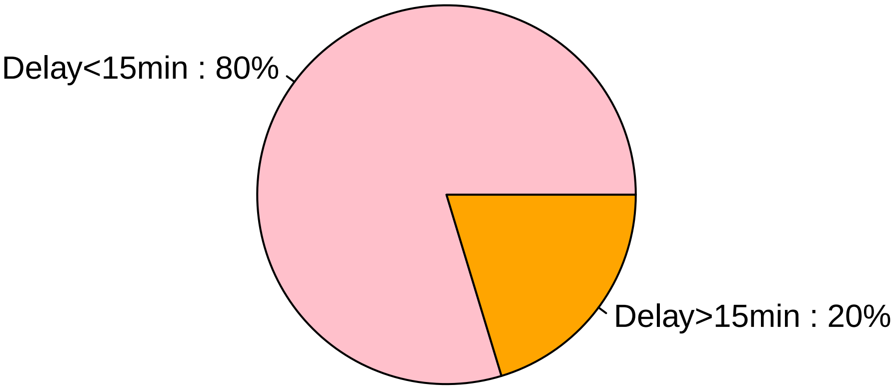


Fig 3 is the Pie chart of Delay distribution. It is clear that most of flights were delayed less than 15 minutes. Delay $<$ 15min flights is about 4 times as much as Delay $<$ 15min flights. To improve performance of the model, I resample the train dataset to creates possibly balanced sample by using the following method:

```
library(ROSE)
set.seed(321)
train<-ovun.sample(ARR_DEL15~.,data = train,method = "under")$data
```

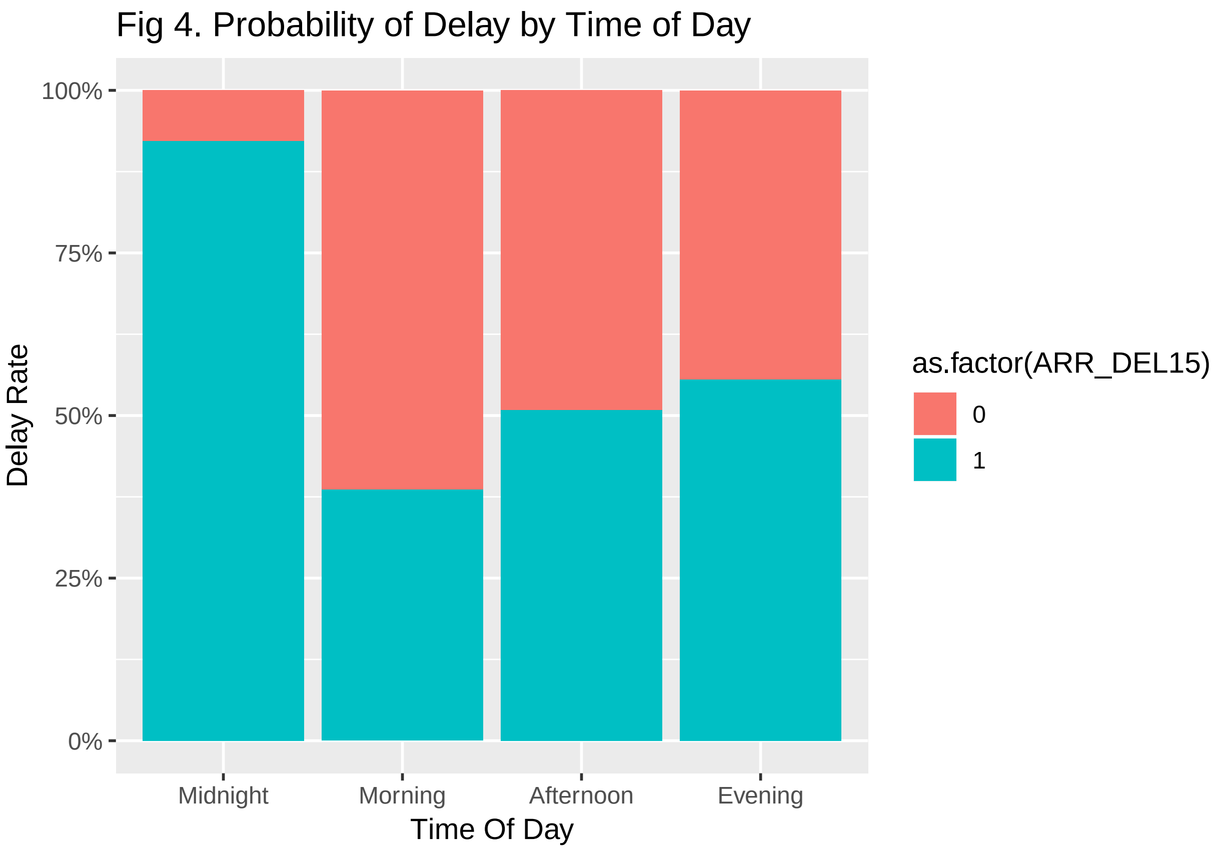


Fig 4. Probability of Delay by Time of Day

Fig 4 shows the probability of delay by Time of Day. I created a new variable "TimeOfDay" by splitting of hour variable into six hour segments:

1. **0am-6am: Midnight**

2. **6am-12am: Morning**

3. **12am-6pm: Afternoon**

4. **6pm-0am: Evening**

As we can see from the plot, flights departured during midnight are more likely delayed.

Fig 5. Probability of Delay by Day of Week

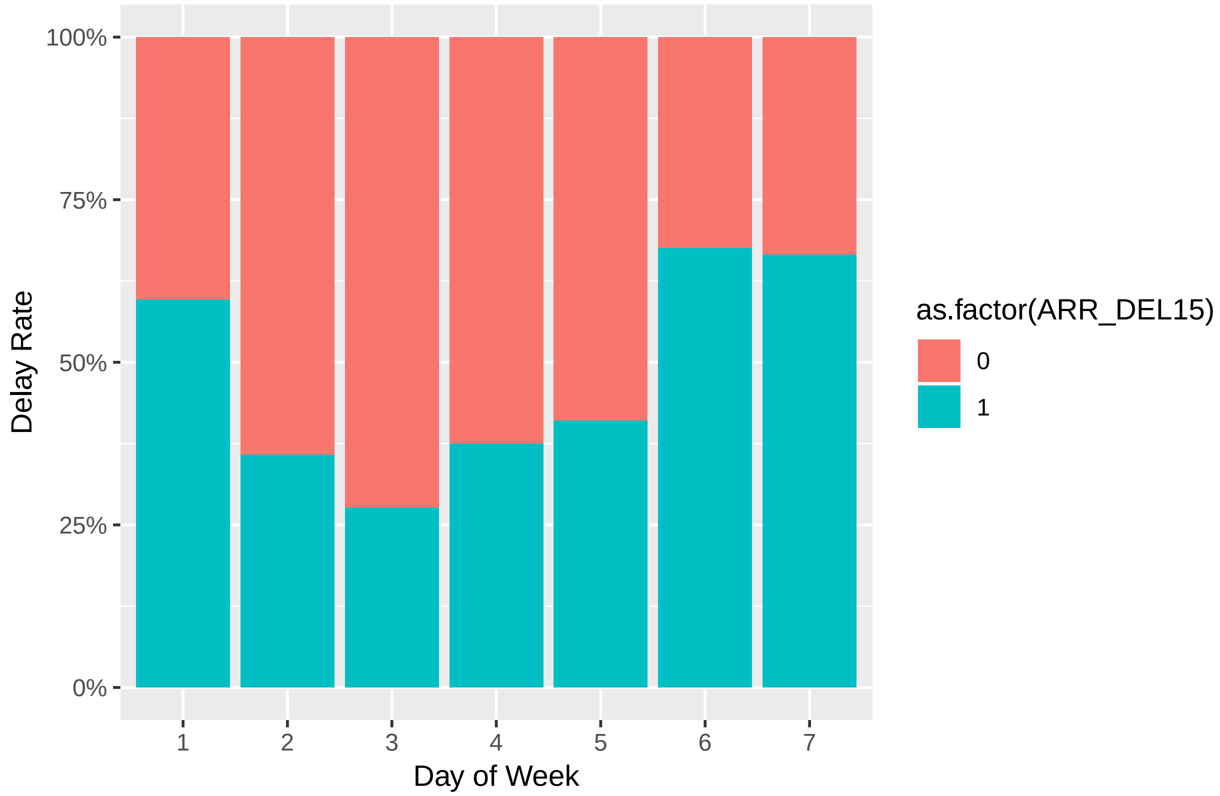


Fig 5 shows the probability of delay by Day of Week. It shows that flights are more likely delayed on Monday, Saturday and Sunday.

Fig 6. Probability of Delay by Destination

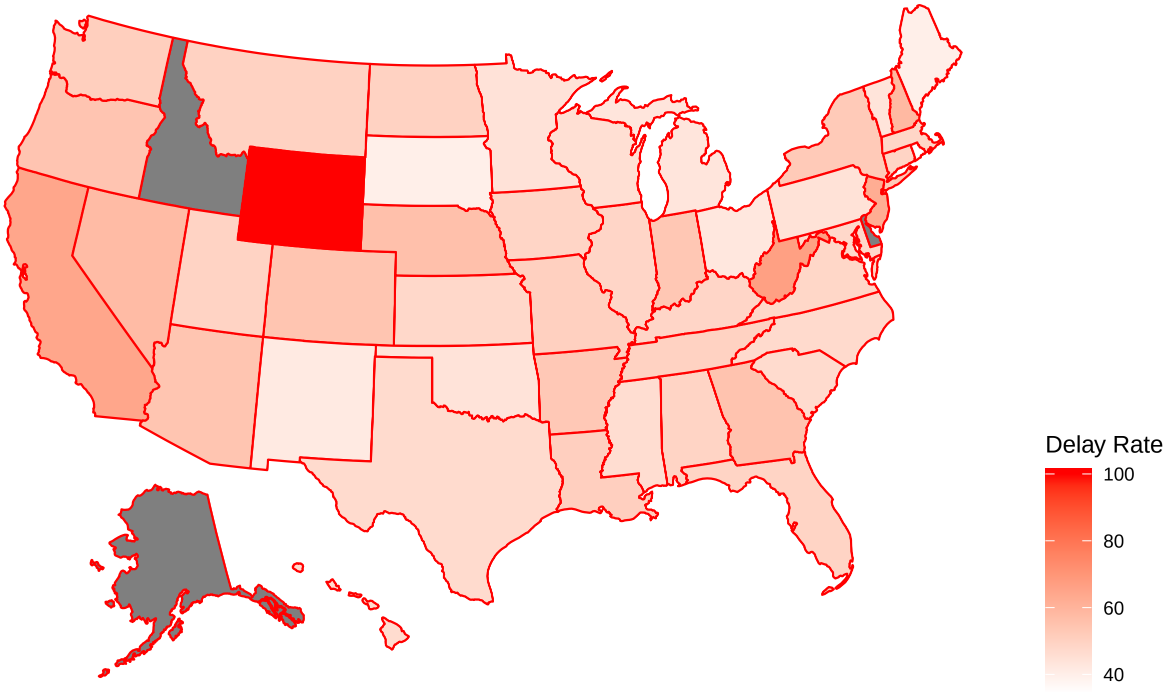


Fig 6 shows probability of delay by destination in U.S. It is clear that Wyoming has higher delay rate, and we don't have data for flights depature from Atlanta to Alaska and Idaho.

Fig 7.Correlation Plot

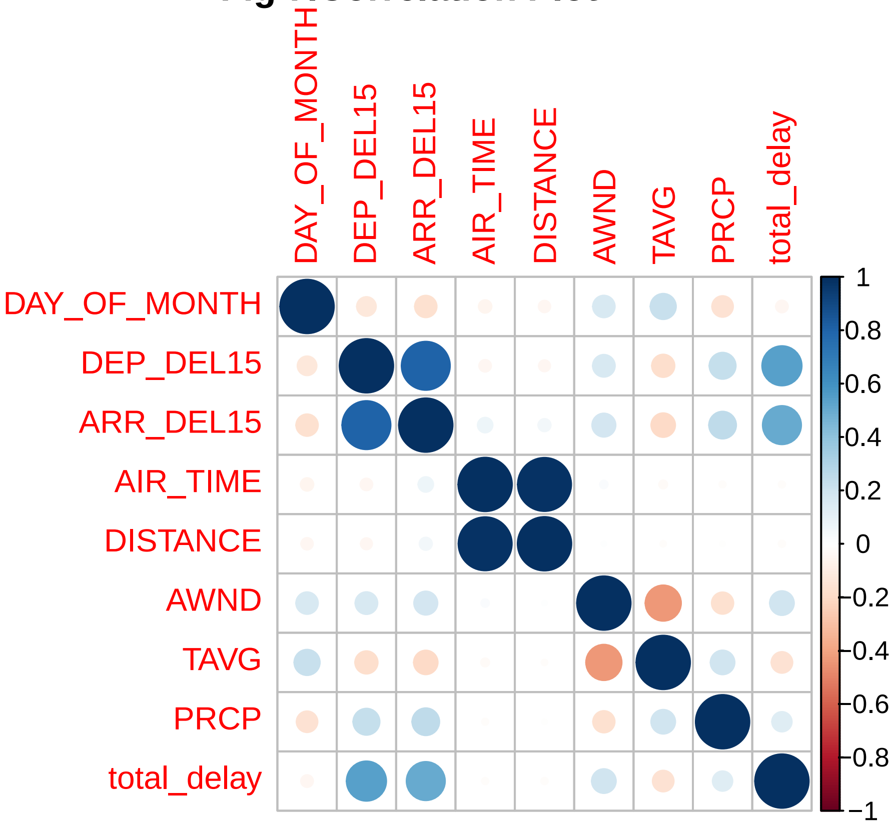


Fig 7 visually examine the between-predictor of the data.The variables "DEP_DEL15" and "AIR_TIME" are highly correlated with others predictors. So we remove them from potential predictors.

3.2 Model used

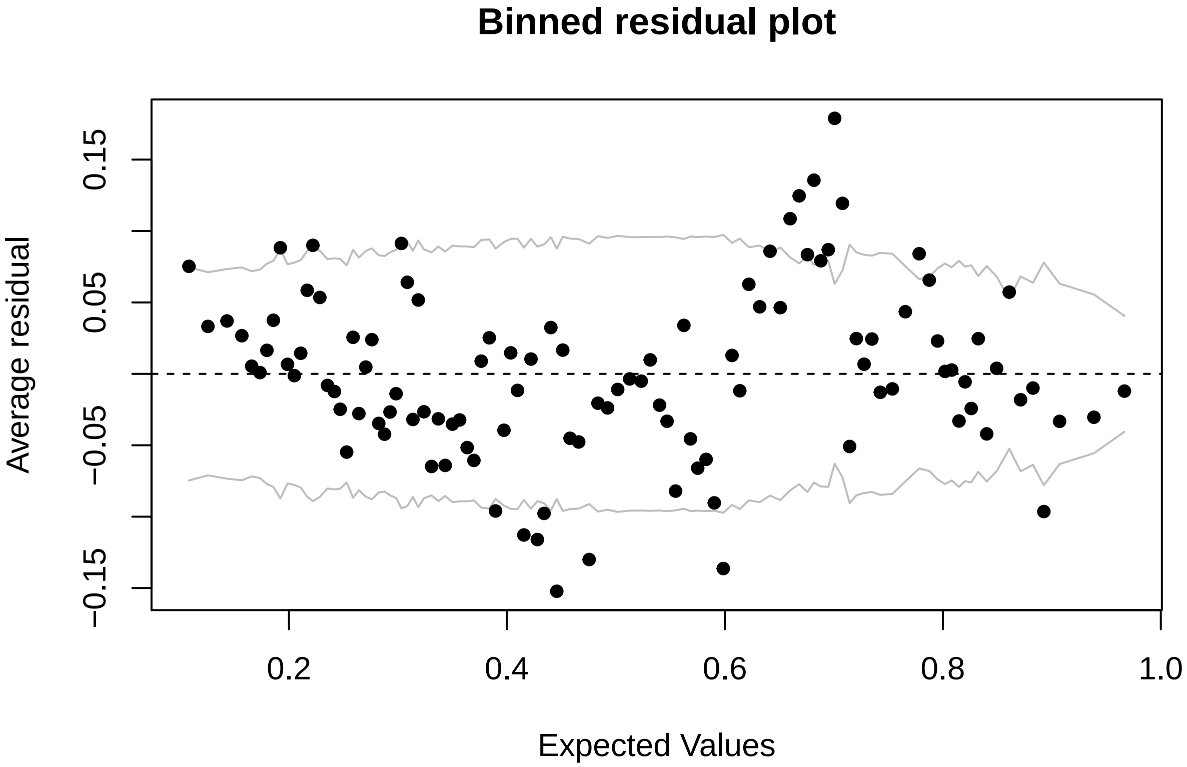
```
#Multilevel logistic model
library(car)
library(lme4)
train$OP_UNIQUE_CARRIER <- as.factor(train$OP_UNIQUE_CARRIER)
m3 <- glmer(ARR_DEL15 ~  (1|OP_UNIQUE_CARRIER)  + scale(DISTANCE) + DAY_OF_MONTH + WT01 + WT02 + DAY_OF_
summary(m3)
```

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## ARR_DEL15 ~ (1 | OP_UNIQUE_CARRIER) + scale(DISTANCE) + DAY_OF_MONTH +
##     WT01 + WT02 + DAY_OF_WEEK + TAVG + (1 | TimeOfDay)
##    Data: train
##
##      AIC      BIC   logLik deviance df.resid
##  13897.2  14000.6  -6934.6  13869.2    11915
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -8.9431 -0.7443  0.1493  0.7356  3.0314
##
## Random effects:
##  Groups            Name        Variance Std.Dev.
##  OP_UNIQUE_CARRIER (Intercept) 0.1156   0.3399
##  TimeOfDay         (Intercept) 1.1603   1.0772
## Number of obs: 11929, groups:  OP_UNIQUE_CARRIER, 9; TimeOfDay, 4
##
## Fixed effects:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      2.955157   0.567006   5.212 1.87e-07 ***
## scale(DISTANCE)  0.210394   0.021882   9.615  < 2e-16 ***
## DAY_OF_MONTH    -0.008027   0.002701  -2.971 0.002966 **
## WT011            0.839547   0.057261  14.662  < 2e-16 ***
## WT021            0.916443   0.075636  12.117  < 2e-16 ***
## DAY_OF_WEEK2    -0.770123   0.071593 -10.757  < 2e-16 ***
## DAY_OF_WEEK3    -0.340197   0.091073  -3.735 0.000187 ***
## DAY_OF_WEEK4    -0.362002   0.078584  -4.607 4.09e-06 ***
## DAY_OF_WEEK5    -0.484549   0.077037  -6.290 3.18e-10 ***
## DAY_OF_WEEK6     0.010498   0.077385   0.136 0.892091
## DAY_OF_WEEK7    -0.015421   0.070172  -0.220 0.826063
## TAVG            -0.046602   0.002428 -19.195  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##              (Intr) s(DIST DAY_OF_M WT011  WT021  DAY_OF_WEEK2
## s(DISTANCE)  -0.004
## DAY_OF_MONT  -0.007  0.033
## WT011         0.018  0.018  0.143
## WT021         0.033  0.011  0.367   -0.371
## DAY_OF_WEEK2 -0.030 -0.037  0.117    0.113 -0.004
## DAY_OF_WEEK3  0.004 -0.035  0.286    0.328  0.166  0.406
## DAY_OF_WEEK4 -0.008 -0.035  0.201   -0.011  0.290  0.412
## DAY_OF_WEEK5 -0.030 -0.028  0.100   -0.079  0.242  0.388
## DAY_OF_WEEK6 -0.044 -0.012 -0.109   -0.162 -0.069  0.348
## DAY_OF_WEEK7 -0.081 -0.009 -0.105    0.126 -0.252  0.396
## TAVG         -0.149 -0.014 -0.440   -0.319 -0.340 -0.186
##              DAY_OF_WEEK3 DAY_OF_WEEK4 DAY_OF_WEEK5 DAY_OF_WEEK6
## s(DISTANCE)
## DAY_OF_MONT
## WT011
## WT021
## DAY_OF_WEEK2
## DAY_OF_WEEK3
## DAY_OF_WEEK4  0.424
## DAY_OF_WEEK5  0.357        0.428
## DAY_OF_WEEK6  0.201        0.302        0.326
## DAY_OF_WEEK7  0.273        0.283        0.312        0.383
## TAVG         -0.432       -0.304       -0.155        0.079
##              DAY_OF_WEEK7
## s(DISTANCE)
## DAY_OF_MONT
## WT011
## WT021
## DAY_OF_WEEK2
## DAY_OF_WEEK3
## DAY_OF_WEEK4
## DAY_OF_WEEK5
## DAY_OF_WEEK6
## DAY_OF_WEEK7
## TAVG          0.144
```

```
## 1 2 3 4 5 6
## 0 1 0 1 1 1
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction     0     1
##          0 11594   874
##          1 11155  3503
## 
##                Accuracy : 0.5566
##                  95% CI : (0.5506, 0.5625)
##     No Information Rate : 0.8386
##     P-Value [Acc > NIR] : 1
## 
##                   Kappa : 0.1591
##  Mcnemar's Test P-Value : <2e-16
## 
##             Sensitivity : 0.5096
##             Specificity : 0.8003
##          Pos Pred Value : 0.9299
##          Neg Pred Value : 0.2390
##              Prevalence : 0.8386
##          Detection Rate : 0.4274
##    Detection Prevalence : 0.4596
##       Balanced Accuracy : 0.6550
## 
##        'Positive' Class : 0
## 
```

Fig 8.Binned Residual Plot for Multilevel Logistic Regression

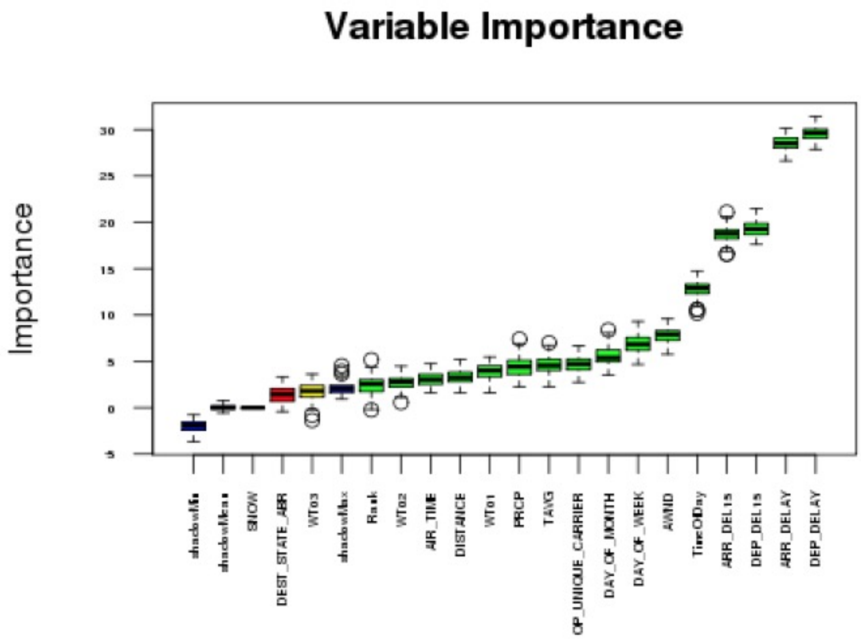


I applied multilevel logistic regression model by put different carrier and time of flight departures into groups, however, unfortunately the model gives lower accuracy with 56%. And binned residual plot also shows a wired trend(Fig 8).

4 Result

4.1 Model choice

Fig 9. Boruta Feature Selection Result



To start building the model, I start with feature selection using Boruta to decide if a variable is important or not. Fig 8 is the box plot of variable importance that Boruta produced.To predict the Arrival delay, the following attributes are considered in feature selection out of which "DEST__STATE__ABR" and "SNOW" happen to have the least influence on the delays, therefore we exclude them from the model.

```r
#Logistic Model
library(car)
m4 <- glm(ARR_DEL15 ~ DAY_OF_MONTH  + OP_UNIQUE_CARRIER  + WT01 + WT02 + poly(AWND,3) + TAVG + poly(PRC
```

Fig 9.Marginal Model Plot for Logistic Regression

```r
marginalModelPlots(m4)
```

```
## Warning in mmps(...): Splines and/or polynomials replaced by a fitted
## linear combination

## Warning in mmps(...): Interactions and/or factors skipped
```

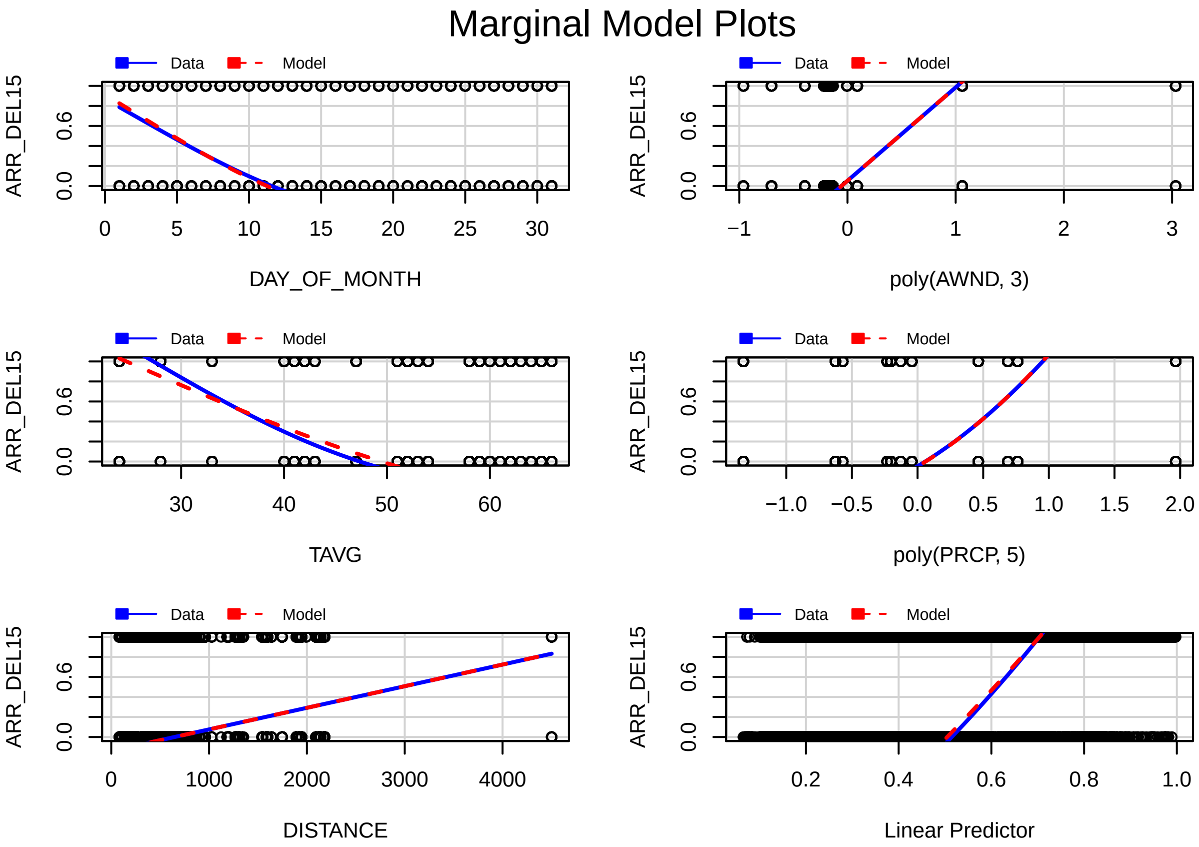


Fig 9 is Marginal Model Plots drawing response variable against each predictors and linear predictor. Fitted values(red lines) are compared to observed data(blue line). I also transformed some of my predictors since they are not having linear relationship with the outcome, such as AWND and PRCP.

4.2 Interpretation

```
summary(m4)
```

```
##
## Call:
## glm(formula = ARR_DEL15 ~ DAY_OF_MONTH + OP_UNIQUE_CARRIER +
##     WT01 + WT02 + poly(AWND, 3) + TAVG + poly(PRCP, 5) + DAY_OF_WEEK +
##     TimeOfDay + DISTANCE, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0078  -0.8824   0.1123   0.8749   2.2879
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             3.258e+00  2.962e-01  11.000  < 2e-16 ***
## DAY_OF_MONTH           -2.125e-02  4.089e-03  -5.196 2.04e-07 ***
## OP_UNIQUE_CARRIERAS -2.367e-01  4.747e-01  -0.499 0.618009
## OP_UNIQUE_CARRIERDL -2.302e-01  1.158e-01  -1.988 0.046867 *  
## OP_UNIQUE_CARRIEREV  6.090e-01  1.249e-01   4.877 1.08e-06 ***
## OP_UNIQUE_CARRIERF9  8.146e-01  2.576e-01   3.162 0.001565 ** 
## OP_UNIQUE_CARRIERNK  2.050e-01  1.854e-01   1.105 0.268958    
## OP_UNIQUE_CARRIEROO  6.331e-01  1.835e-01   3.450 0.000561 ***
## OP_UNIQUE_CARRIERUA  6.574e-01  2.292e-01   2.868 0.004131 ** 
## OP_UNIQUE_CARRIERWN  1.334e-01  1.273e-01   1.048 0.294524    
## WT011                2.744e-01  1.345e-01   2.040 0.041346 *  
## WT021                5.787e-01  1.423e-01   4.066 4.77e-05 ***
## poly(AWND, 3)1       5.725e+01  4.391e+00  13.036  < 2e-16 ***
## poly(AWND, 3)2       4.038e+01  4.604e+00   8.771  < 2e-16 ***
## poly(AWND, 3)3       3.327e+01  3.740e+00   8.894  < 2e-16 ***
## TAVG                -2.236e-02  4.132e-03  -5.412 6.23e-08 ***
## poly(PRCP, 5)1       4.653e+01  5.013e+00   9.282  < 2e-16 ***
## poly(PRCP, 5)2       1.621e+01  4.912e+00   3.300 0.000966 ***
## poly(PRCP, 5)3       5.587e+00  3.543e+00   1.577 0.114819    
## poly(PRCP, 5)4      -1.080e+01  5.309e+00  -2.035 0.041840 *  
## poly(PRCP, 5)5      -3.218e+01  2.820e+00 -11.413  < 2e-16 ***
## DAY_OF_WEEK2        -3.732e-01  1.003e-01  -3.721 0.000198 ***
## DAY_OF_WEEK3        -4.621e-01  1.213e-01  -3.810 0.000139 ***
## DAY_OF_WEEK4        -4.762e-02  9.413e-02  -0.506 0.612943    
## DAY_OF_WEEK5        -1.310e-01  9.299e-02  -1.409 0.158791    
## DAY_OF_WEEK6        -5.374e-01  1.602e-01  -3.354 0.000797 ***
## DAY_OF_WEEK7         7.488e-01  9.494e-02   7.887 3.10e-15 ***
## TimeOfDayMorning    -2.670e+00  2.277e-01 -11.730  < 2e-16 ***
## TimeOfDayAfternoon  -2.204e+00  2.261e-01  -9.749  < 2e-16 ***
## TimeOfDayEvening    -2.043e+00  2.265e-01  -9.019  < 2e-16 ***
## DISTANCE             4.490e-04  4.590e-05   9.782  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 16537  on 11928  degrees of freedom
## Residual deviance: 13203  on 11898  degrees of freedom
## AIC: 13265
## 
## Number of Fisher Scoring iterations: 5
```

In the output above, the first thing we see is the call, there are 10 predictors finally used in model which are: DAY_OF_MONTH,OP_UNIQUE_CARRIER,WT01,WT02,AWND,TAVG,PRCP, DAY_OF_WEEK,TimeOfDay, and DISTANCE.

The next part of the output shows the coefficients, their standard errors, the z-statistic (sometimes called a Wald z-statistic), and the associated p-values. In this model, most of predictor variables are statistically significant, except two terms of airline carrier. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

Therefore, for example, for every one unit change in day of month(DAY_OF_MONTH), the log odds of delay over 15 min (versus under 15 min) decreases by 0.0196.

For a one unit increase in average temperature(TAVG), the log odds of delay increases by 0.000439.

The catagorical indicator variables have a slightly different interpretation. For example, flights departs in the morning, versus departs in the the midnight, change the log odds of delay by -3.086.

Below the table of coefficients are fit indices, including the null and deviance residuals and the AIC. In our model, the residual deviance decreases from null deviance which means the predictor variables made the model performs better.

4.3 Model checking

```
## 1 2 3 4 5 6
## 0 0 0 0 1 1
```

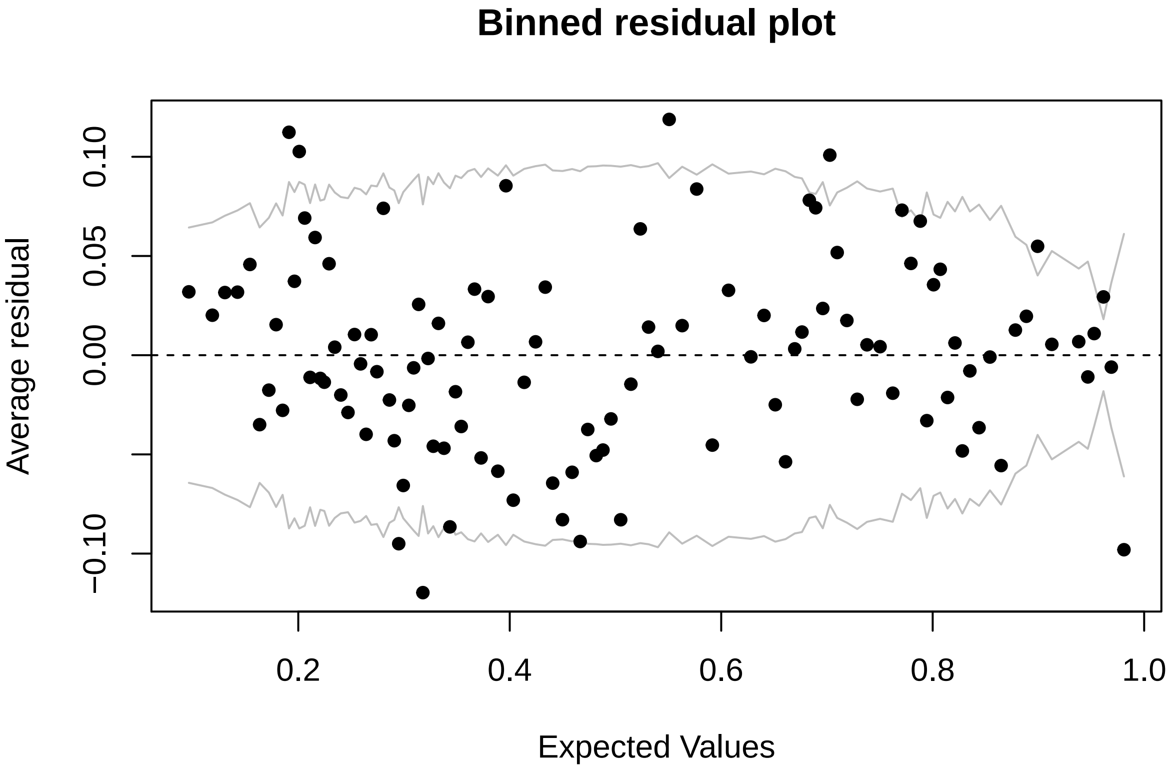
Before model checking, firstly we need to preapre test dataset. I use Jan.2018 flights history as testing data and select all predictors are selected above.

```
#Confusion Matrix
library(caret)
confusionMatrix(m4.predict, test$ARR_DEL15)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction     0     1
##          0 16076  1734
##          1  6673  2643
##
##                Accuracy : 0.6901
##                  95% CI : (0.6845, 0.6956)
##     No Information Rate : 0.8386
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.2133
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7067
##             Specificity : 0.6038
##          Pos Pred Value : 0.9026
##          Neg Pred Value : 0.2837
##              Prevalence : 0.8386
##          Detection Rate : 0.5926
##    Detection Prevalence : 0.6566
##       Balanced Accuracy : 0.6553
##
##        'Positive' Class : 0
##
```

The first method I used to test my model is Confusion Matrix. The number of correct predictions is along the diagonal– where correct was 0 and the prediction was 0 and where correct was 1 and the prediction was 1. Based on the result, most of the data points are classified as class "0", and the overall accuracy is 68.72% which is the sum of the diagonal divided by the sum of the whole matrix.

Fig 10. Binned Residual Plot for Logistic Regression



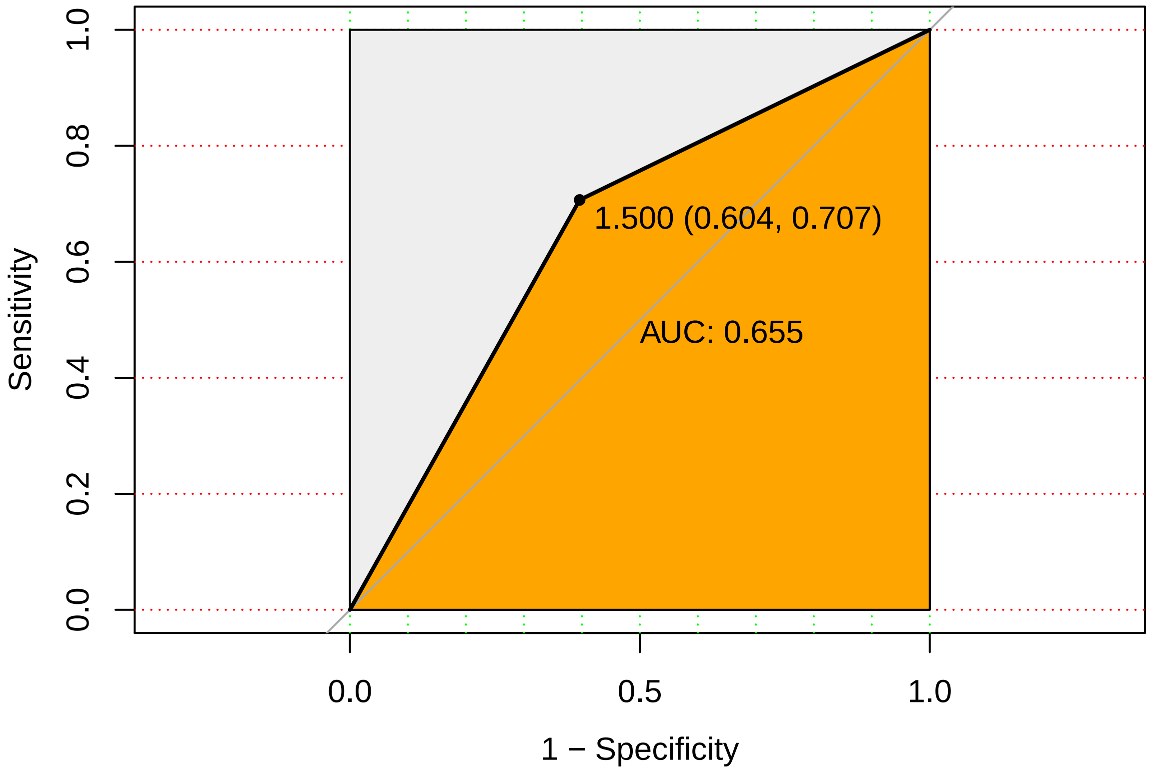
The other method is Binned Residual Plot. Fig 10 shows many points fall inside the confidence bands and there is not a distinctive pattern to the residuals.

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

Fig 11. ROC Curve Plot of Logistic Regression



The last method I used is ROC curve. The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Fig 11 shows the model gives curve closer to the top-left corner indicate it has better performance.

5 Discussion

5.1 Implication

The results of this study have implications for potential positive social change on the individual level. Passengers especially those who travels for a living might need a model to predict how likely their flights will delay and prepare for another plan early.

5.2 Limitation

The limitation of the model is that I only use January's flights history as training data and only picked ATL as origin departure location since I don't have weather data for all locations in U.S. The limited data used for training this model could lead some problem when applying other location's flights and weather information to it.

5.3 Future direction

A future analysis using machine learning methods may be conducted to carry out the estimation of delays flights departing from ATL, for instance SVM, random forests those like machine learning method can be used in this matter.

6 Acknowledgement

I would like to acknowledge with much appreciation the crucial role of Dr.Yajima, who gave me a lot of suggestions when I confronted with difficulties. A special thanks goes to Bureau of Transportation that provides me the dataset.

7 Reference

Fahad Alsehami. Medium. Weblog. [Online] Available from: https://medium.com/@famsfsu/predicting-flight-delay-u-s-airports-886748c617d5 [Accessed 11th November 2018].

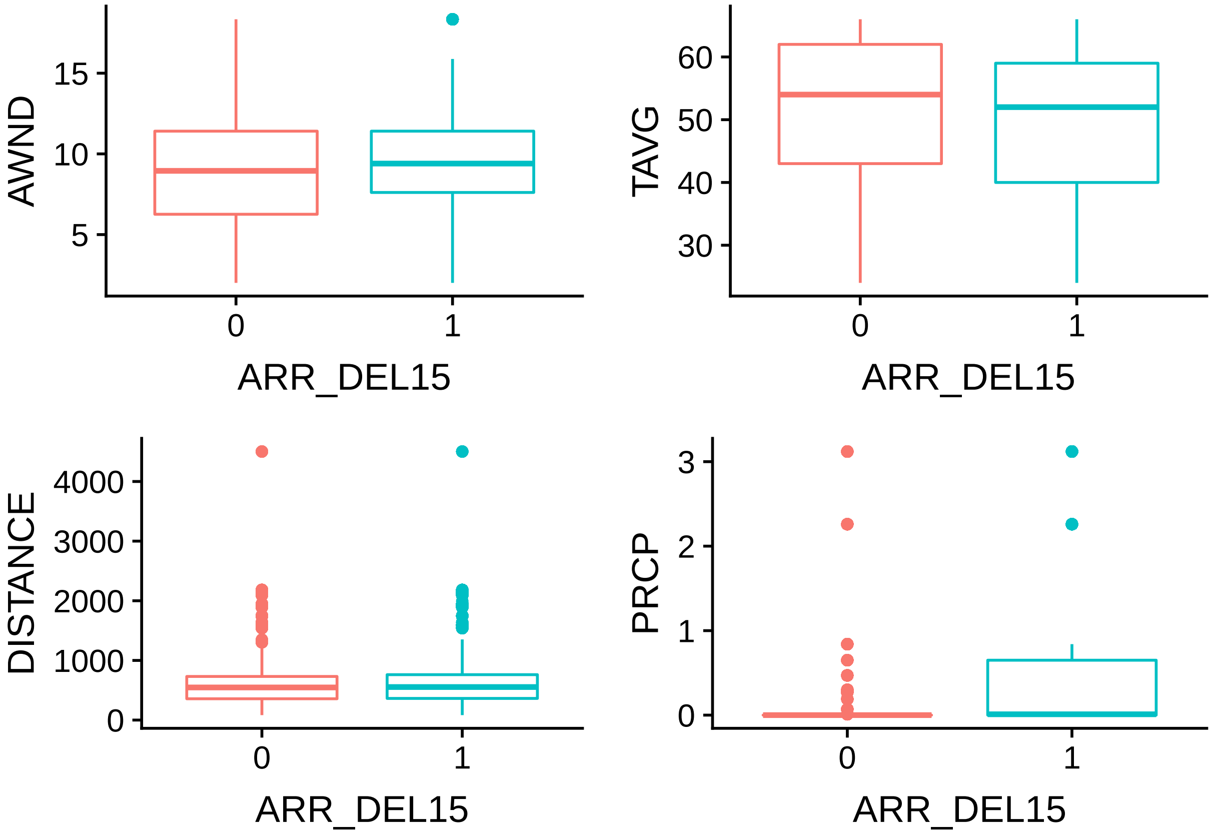
Ayman Siraj. RPubs. Weblog. [Online] Available from: https://rpubs.com/aymansir/usflightdelay [Accessed 11th November 2018].

8 Appendix

Fig 12. Box Plot for each numeric predictor(AWND,TAVG,DISTANCE,PRCP) vs. response variable(ARR_DEL15)

```
##
## Attaching package: 'cowplot'

## The following object is masked from 'package:ggplot2':
##
##     ggsave
```


AWND
15
10
5
0
1
ARR_DEL15
TAVG
60
50
40
30
0
1
ARR_DEL15
DISTANCE
4000
3000
2000
1000
0
0
1
ARR_DEL15
PRCP
3
2
1
0
0
1
ARR_DEL15