flight delay analysis

Aileen Gutierrez

Using the "flight.csv" from the dataset "2015 Flight Delays and Cancellations". The question is: which airline has the least risk of canceling. Looking at the columns, they vary on telling you about the status of a given flight. To start with a simpler model, only use the factors most related to this would be those related to date of flight, time delay, and distance for flight.

Because the answer is binary (cancel, not cancel), use a logistic regression.

Start by importing the data. The column "Canceled", shows 1 for canceled flight, and time variables as integers.

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Rows: 5819079 Columns: 31

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): AIRLINE, TAIL_NUMBER, ORIGIN_AIRPORT, DESTINATION_AIRPORT, SCHEDUL...
## dbl (20): YEAR, MONTH, DAY, DAY_OF_WEEK, FLIGHT_NUMBER, DEPARTURE_DELAY, TAX...

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Summary of data to check for NAs, outliers, and other issues. And there is imbalance with the dependent varaible. 0 is over 98% of the total. Also, based on count of rows per airline, (excluding WN which has significantly more than the rest) filter down to DL, AA, OO and EV. Also, "DEPARTURE_DELAY" cannot have negative values, its minimum is 0, since that means a flight departed at the time it was supposed to.

```
flight_delay <- flight_delay %>% filter(AIRLINE == 'DL' | AIRLINE == 'AA' | AIRLINE == 'OO' | AIRLINE ==
flight_delay <- flight_delay %>% filter(DEPARTURE_DELAY >= 0)
summary(flight_delay)
```

```
##       YEAR          MONTH             DAY          DAY_OF_WEEK
##  Min.   :2015   Min.   : 1.000   Min.   : 1.00   Min.   :1.000
##  1st Qu.:2015   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.:2.000
##  Median :2015   Median : 7.000   Median :16.00   Median :4.000
##  Mean   :2015   Mean   : 6.567   Mean   :15.68   Mean   :3.918
##  3rd Qu.:2015   3rd Qu.: 9.000   3rd Qu.:23.00   3rd Qu.:6.000
##  Max.   :2015   Max.   :12.000   Max.   :31.00   Max.   :7.000
##
```

```
##    AIRLINE          FLIGHT_NUMBER  TAIL_NUMBER        ORIGIN_AIRPORT
##  Length:1019572     Min.   :   1   Length:1019572     Length:1019572
##  Class :character   1st Qu.:1302   Class :character   Class :character
##  Mode  :character   Median :2237   Mode  :character   Mode  :character
##                     Mean   :2802
##                     3rd Qu.:4670
##                     Max.   :9794
##
##  DESTINATION_AIRPORT SCHEDULED_DEPARTURE DEPARTURE_TIME     DEPARTURE_DELAY
##  Length:1019572      Length:1019572      Length:1019572     Min.   :   0.00
##  Class :character    Class :character    Class :character   1st Qu.:   2.00
##  Mode  :character    Mode  :character    Mode  :character   Median :  11.00
##                                                             Mean   :  29.92
##                                                             3rd Qu.:  33.00
##                                                             Max.   :1988.00
##
##     TAXI_OUT      WHEELS_OFF        SCHEDULED_TIME   ELAPSED_TIME
##  Min.   :  1.0   Length:1019572     Min.   : 20.0   Min.   : 16.0
##  1st Qu.: 13.0   Class :character   1st Qu.: 88.0   1st Qu.: 85.0
##  Median : 16.0   Mode  :character   Median :122.0   Median :118.0
##  Mean   : 18.6                      Mean   :141.1   Mean   :136.6
##  3rd Qu.: 21.0                      3rd Qu.:174.0   3rd Qu.:169.0
##  Max.   :225.0                      Max.   :718.0   Max.   :711.0
##  NA's   :990                                        NA's   :5147
##     AIR_TIME        DISTANCE       WHEELS_ON           TAXI_IN
##  Min.   :  8.0   Min.   :  21.0   Length:1019572     Min.   :  1.000
##  1st Qu.: 60.0   1st Qu.: 373.0   Class :character   1st Qu.:  4.000
##  Median : 91.0   Median : 630.0   Mode  :character   Median :  6.000
##  Mean   :110.4   Mean   : 796.9                      Mean   :  7.576
##  3rd Qu.:141.0   3rd Qu.:1045.0                      3rd Qu.:  9.000
##  Max.   :669.0   Max.   :4983.0                      Max.   :202.000
##  NA's   :5147                                        NA's   :1950
##  SCHEDULED_ARRIVAL  ARRIVAL_TIME       ARRIVAL_DELAY        DIVERTED
##  Length:1019572     Length:1019572     Min.   : -79.00   Min.   :0.000000
##  Class :character   Class :character   1st Qu.:  -4.00   1st Qu.:0.000000
##  Mode  :character   Mode  :character   Median :   8.00   Median :0.000000
##                                        Mean   :  25.31   Mean   :0.003804
##                                        3rd Qu.:  32.00   3rd Qu.:0.000000
##                                        Max.   :1971.00   Max.   :1.000000
##                                        NA's   :5147
##    CANCELLED        CANCELLATION_REASON AIR_SYSTEM_DELAY SECURITY_DELAY
##  Min.   :0.000000   Length:1019572      Min.   :   0.0   Min.   :  0.0
##  1st Qu.:0.000000   Class :character    1st Qu.:   0.0   1st Qu.:  0.0
##  Median :0.000000   Mode  :character    Median :   0.0   Median :  0.0
##  Mean   :0.001245                       Mean   :  12.5   Mean   :  0.1
##  3rd Qu.:0.000000                       3rd Qu.:  15.0   3rd Qu.:  0.0
##  Max.   :1.000000                       Max.   :1049.0   Max.   :573.0
##                                         NA's   :607177   NA's   :607177
##  AIRLINE_DELAY    LATE_AIRCRAFT_DELAY WEATHER_DELAY
##  Min.   :   0.0   Min.   :   0.0      Min.   :   0.0
##  1st Qu.:   0.0   1st Qu.:   0.0      1st Qu.:   0.0
##  Median :   3.0   Median :   6.0      Median :   0.0
##  Mean   :  23.9   Mean   :  25.4      Mean   :   3.6
##  3rd Qu.:  23.0   3rd Qu.:  32.0      3rd Qu.:   0.0
```

```
##  Max.   :1971.0   Max.   :1331.0      Max.   :1211.0
##  NA's   :607177   NA's   :607177      NA's   :607177
```

```
prop.table(table(flight_delay$CANCELLED))
```

```
##
##          0          1
## 0.99875536 0.00124464
```

Filter to the columns that will predict whether or not a flight is delayed (as stated above). The rows that have missing information seem to not have DEPARTURE_TIME,DEPARTURE_DELAY and were cancelled. These rows can be deleted

```
flight_clean <- flight_delay %>% select(c('MONTH','AIRLINE','ELAPSED_TIME','DEPARTURE_DELAY','DISTANCE'
flight_clean[!complete.cases(flight_clean),]
```

```
## # A tibble: 5,147 x 6
##    MONTH AIRLINE ELAPSED_TIME DEPARTURE_DELAY DISTANCE CANCELLED
##    <dbl> <chr>          <dbl>           <dbl>    <dbl>     <dbl>
##  1     1 EV                NA              41      295         0
##  2     1 EV                NA             206      107         0
##  3     1 EV                NA             125      667         1
##  4     1 OO                NA               9      913         0
##  5     1 EV                NA               7      295         1
##  6     1 OO                NA              29      524         0
##  7     1 OO                NA              24      737         0
##  8     1 EV                NA              81     1215         0
##  9     1 EV                NA              49      563         1
## 10     1 EV                NA               2      984         0
## # ... with 5,137 more rows
```

```
flight_clean %>% filter(is.na(DEPARTURE_DELAY) & is.na(ELAPSED_TIME)) %>% nrow()
```

```
## [1] 0
```

```
flight_fl <- flight_clean %>% filter(!(is.na(DEPARTURE_DELAY) & is.na(ELAPSED_TIME)))
summary(flight_fl)
```

```
##      MONTH          AIRLINE           ELAPSED_TIME   DEPARTURE_DELAY
##  Min.   : 1.000   Length:1019572     Min.   : 16.0   Min.   :   0.00
##  1st Qu.: 4.000   Class :character   1st Qu.: 85.0   1st Qu.:   2.00
##  Median : 7.000   Mode  :character   Median :118.0   Median :  11.00
##  Mean   : 6.567                      Mean   :136.6   Mean   :  29.92
##  3rd Qu.: 9.000                      3rd Qu.:169.0   3rd Qu.:  33.00
##  Max.   :12.000                      Max.   :711.0   Max.   :1988.00
##                                      NA's   :5147
##     DISTANCE        CANCELLED
##  Min.   :  21.0   Min.   :0.000000
##  1st Qu.: 373.0   1st Qu.:0.000000
##  Median : 630.0   Median :0.000000
##  Mean   : 796.9   Mean   :0.001245
##  3rd Qu.:1045.0   3rd Qu.:0.000000
##  Max.   :4983.0   Max.   :1.000000
##
```

Import libraries, and get correlations for numeric columns. Correlations are very high between ELAPSED_TIME and DISTANCE.

```
library(dplyr)
library(tidyr)
all_num <- flight_fl %>% select(c('ELAPSED_TIME','DEPARTURE_DELAY','DISTANCE'))
round(cor(all_num, use="complete.obs"), digits=2)
```

```
##                 ELAPSED_TIME DEPARTURE_DELAY DISTANCE
## ELAPSED_TIME            1.00           -0.02     0.97
## DEPARTURE_DELAY        -0.02            1.00    -0.03
## DISTANCE                0.97           -0.03     1.00
```

First turn the columns into factors:

```
flight_fl$CANCELLED = as.factor(flight_fl$CANCELLED)
flight_fl$AIRLINE = as.factor(flight_fl$AIRLINE)
flight_fl$MONTH = as.factor(flight_fl$MONTH)
```

Filter to DISTANCE having positive values, and not having outliers. Delete ELAPSED_TIME since it has high correlation with DISTANCE.

```
flight_delay_new <- flight_fl  #na.omit(flight_fl)
flight_delay_new <- flight_delay_new %>% filter(DISTANCE >= 0 , DISTANCE < 10170)
flight_delay_new <- flight_delay_new %>% select(-c('ELAPSED_TIME'))
```

Check nulls again

```
flight_delay_new[!complete.cases(flight_delay_new),]
```

```
## # A tibble: 0 x 5
## # ... with 5 variables: MONTH <fct>, AIRLINE <fct>, DEPARTURE_DELAY <dbl>,
## #   DISTANCE <dbl>, CANCELLED <fct>
```

```
summary(flight_delay_new)
```

```
##      MONTH        AIRLINE     DEPARTURE_DELAY      DISTANCE      CANCELLED
##  7      :104493   AA:279559   Min.   :   0.00   Min.   :  21.0   0:1018303
##  8      : 99076   DL:349596   1st Qu.:   2.00   1st Qu.: 373.0   1:   1269
##  12     : 95059   EV:191701   Median :  11.00   Median : 630.0
##  6      : 94987   OO:198716   Mean   :  29.92   Mean   : 796.9
##  3      : 88206               3rd Qu.:  33.00   3rd Qu.:1045.0
##  1      : 79090               Max.   :1988.00   Max.   :4983.0
##  (Other):458661
```

Test for multicollinearity using VIF. All are less than 5, so no issues there. As for the coeficients, several of the months are not significant.

```
flight_delay_new$AIRLINE <- factor(flight_delay_new$AIRLINE, levels=c('DL','AA','OO', 'EV'))

model1 <- glm(flight_delay_new$CANCELLED  ~., family = binomial(link = 'logit'), data = flight_delay_ne
summary(model1)
```

```
##
## Call:
## glm(formula = flight_delay_new$CANCELLED ~ ., family = binomial(link = "logit"),
##     data = flight_delay_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5987  -0.0604  -0.0464  -0.0232   4.2818
```

```
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.312e+00  1.568e-01 -53.021  < 2e-16 ***
## MONTH2           5.707e-01  1.294e-01   4.412 1.02e-05 ***
## MONTH3           1.406e-01  1.391e-01   1.010 0.312288
## MONTH4           1.859e-01  1.428e-01   1.302 0.192962
## MONTH5           2.955e-01  1.369e-01   2.159 0.030877 *
## MONTH6           7.159e-02  1.379e-01   0.519 0.603629
## MONTH7          -4.688e-02  1.401e-01  -0.335 0.737844
## MONTH8           1.054e-01  1.371e-01   0.769 0.441998
## MONTH9          -7.002e-01  1.919e-01  -3.649 0.000263 ***
## MONTH10         -9.820e-01  2.067e-01  -4.750 2.03e-06 ***
## MONTH11         -1.656e-02  1.487e-01  -0.111 0.911332
## MONTH12          5.118e-01  1.264e-01   4.050 5.12e-05 ***
## AIRLINEAA        1.717e+00  1.216e-01  14.124  < 2e-16 ***
## AIRLINEOO        1.996e+00  1.211e-01  16.480  < 2e-16 ***
## AIRLINEEV        2.011e+00  1.215e-01  16.549  < 2e-16 ***
## DEPARTURE_DELAY  3.890e-03  1.474e-04  26.386  < 2e-16 ***
## DISTANCE        -3.123e-04  6.752e-05  -4.625 3.74e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 19513  on 1019571  degrees of freedom
## Residual deviance: 18375  on 1019555  degrees of freedom
## AIC: 18409
##
## Number of Fisher Scoring iterations: 11
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
car::vif(model1)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## MONTH           1.021518 11        1.000968
## AIRLINE         1.336117  3        1.049480
## DEPARTURE_DELAY 1.024236  1        1.012045
## DISTANCE        1.305402  1        1.142542
```

For a model with only AIRLINE as independent variable, with DL as reference group. AA has an odds of being cancelled of 5.4493 times that of being canceled in DL, holding everything else the same. For OO, it's 8.4390, for EV 8.748527. This makes airline EV as highest odds of getting cancelled.

```r
model1_2 <- glm(flight_delay_new$CANCELLED  ~ AIRLINE, family = binomial(link = 'logit'), data = flight_
summary(model1_2)
```

```
##
## Call:
## glm(formula = flight_delay_new$CANCELLED ~ AIRLINE, family = binomial(link = "logit"),
##     data = flight_delay_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.0652  -0.0640  -0.0515  -0.0221   4.0797
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -8.3216     0.1085  -76.71   <2e-16 ***
## AIRLINEAA     1.6955     0.1203   14.09   <2e-16 ***
## AIRLINEOO     2.1329     0.1193   17.88   <2e-16 ***
## AIRLINEEV     2.1689     0.1193   18.18   <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: 19513  on 1019571  degrees of freedom
## Residual deviance: 18904  on 1019568  degrees of freedom
## AIC: 18912
##
## Number of Fisher Scoring iterations: 11
```

```r
### Exponentiated coefficients.
exp(model1_2$coefficients)
```

```
##  (Intercept)    AIRLINEAA    AIRLINEOO    AIRLINEEV
## 0.0002431969 5.4493580462 8.4390567542 8.7485279510
```

Subset into test and train, and run a logistic regression. Looking at the results of confusion matrix, the prediction does not do well for Cancel = 1, despite high accuracy overall. Looking at the confusion matrix, the Specifity is 0, which means when CANCELLED=1, it was not predicted correctly at all. While for Sensitivity it is 1,meaning that CANCELLED=0 was correctly predicted all the time. To address this problem, balance the classes (through upsmpling), and test the model again.

```r
library(caTools)
set.seed(123)
split = sample.split(flight_delay_new$CANCELLED, SplitRatio = 0.75)
training_set = subset(flight_delay_new, split == TRUE)
test_set = subset(flight_delay_new, split == FALSE)

prop.table(table(training_set$CANCELLED))
```

```
##
##           0           1
## 0.998755033 0.001244967
```

```r
model_train <- glm(training_set$CANCELLED ~., family = binomial(link = 'logit'), data = training_set)
summary(model_train)
```

```
##
## Call:
## glm(formula = training_set$CANCELLED ~ ., family = binomial(link = "logit"),
##     data = training_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6170  -0.0600  -0.0465  -0.0236   4.2428
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.207e+00  1.781e-01 -46.074  < 2e-16 ***
## MONTH2           5.344e-01  1.480e-01   3.611 0.000305 ***
## MONTH3           1.398e-01  1.584e-01   0.883 0.377498
## MONTH4           4.450e-02  1.684e-01   0.264 0.791539
## MONTH5           3.254e-01  1.544e-01   2.107 0.035080 *
## MONTH6           2.555e-02  1.583e-01   0.161 0.871768
## MONTH7          -1.836e-02  1.581e-01  -0.116 0.907506
## MONTH8          -2.871e-02  1.604e-01  -0.179 0.857944
## MONTH9          -8.283e-01  2.281e-01  -3.631 0.000282 ***
## MONTH10         -1.034e+00  2.388e-01  -4.332 1.48e-05 ***
## MONTH11         -4.703e-02  1.701e-01  -0.277 0.782143
## MONTH12          5.047e-01  1.436e-01   3.514 0.000441 ***
## AIRLINEAA        1.713e+00  1.382e-01  12.398  < 2e-16 ***
## AIRLINEOO        1.936e+00  1.379e-01  14.033  < 2e-16 ***
## AIRLINEEV        1.942e+00  1.384e-01  14.024  < 2e-16 ***
## DEPARTURE_DELAY  3.872e-03  1.707e-04  22.687  < 2e-16 ***
## DISTANCE        -3.562e-04  7.836e-05  -4.545 5.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 14638  on 764678  degrees of freedom
## Residual deviance: 13787  on 764662  degrees of freedom
## AIC: 13821
##
## Number of Fisher Scoring iterations: 11
```

```
exp(model_train$coefficients)
```

```
##     (Intercept)          MONTH2          MONTH3          MONTH4          MONTH5
##    0.0002728501    1.7064898356    1.1500165732    1.0455095099    1.3845859491
##          MONTH6          MONTH7          MONTH8          MONTH9         MONTH10
##    1.0258754760    0.9818044064    0.9716966077    0.4368014487    0.3554130517
##         MONTH11         MONTH12       AIRLINEAA       AIRLINEOO       AIRLINEEV
##    0.9540613513    1.6565117312    5.5464182499    6.9293670257    6.9697839371
## DEPARTURE_DELAY        DISTANCE
##    1.0038797129    0.9996439063
```

```
pred_test = predict(model_train, test_set[-5], type = 'response')
pred = ifelse(pred_test > 0.5, 1, 0)
pred = as.factor(pred)

library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
## The following object is masked from 'package:purrr':
##
##     lift
```

```
confusionMatrix(pred,test_set$CANCELLED)
```

```
## Warning in confusionMatrix.default(pred, test_set$CANCELLED): Levels are not in
## the same order for reference and data. Refactoring data to match.
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 254576    317
##          1      0      0
##
##                Accuracy : 0.9988
##                  95% CI : (0.9986, 0.9989)
##     No Information Rate : 0.9988
##     P-Value [Acc > NIR] : 0.5149
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.0000
##          Pos Pred Value : 0.9988
##          Neg Pred Value :    NaN
##              Prevalence : 0.9988
##          Detection Rate : 0.9988
##    Detection Prevalence : 1.0000
##       Balanced Accuracy : 0.5000
##
##        'Positive' Class : 0
##
```

The next step would be to make a model with all the independent variables, with a training and test set. Before fitting model, upsample so that there is a better prediction for Cancel = 1 (which is the minority in the data set).

```
set.seed(456)
split = sample.split(flight_delay_new$CANCELLED, SplitRatio = 0.7)
training = subset(flight_delay_new, split == TRUE)
test = subset(flight_delay_new, split == FALSE)
```

Upsample step:

```
set.seed(1234)
up_samp <- upSample(x = training[,-5], y=training$CANCELLED)
table(up_samp$Class)
```

```
##
```

```
##      0      1
## 712812 712812
```

Apply logistic regression to the upsamled training set, lastly get confusion matrix. This is an improvement, all the varaibles are significant, and correct predictions are between 67% and 75%.

The odds of a Canceled flight are for AA 6.5678 and for OO its 10.0045, for EV its 9.1567 compared to odds for DL (holding all other variables at fixed values). In this model, airline OO has the highest odds of getting canceled.

```
model_upsample <- glm(up_samp$Class ~., family = binomial(link = 'logit'), data = up_samp)
summary(model_upsample)
```

```
##
## Call:
## glm(formula = up_samp$Class ~ ., family = binomial(link = "logit"),
##     data = up_samp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.6814  -0.9952  -0.0715   0.9696   2.4284
##
## Coefficients:
##                   Estimate Std. Error  z value Pr(>|z|)
## (Intercept)     -2.270e+00  9.758e-03 -232.646  < 2e-16 ***
## MONTH2           7.023e-01  9.310e-03   75.430  < 2e-16 ***
## MONTH3           2.471e-01  9.502e-03   26.005  < 2e-16 ***
## MONTH4          -2.629e-02  1.000e-02   -2.629  0.00858 **
## MONTH5           2.435e-01  9.490e-03   25.658  < 2e-16 ***
## MONTH6          -7.478e-02  9.353e-03   -7.996 1.29e-15 ***
## MONTH7           5.394e-02  9.104e-03    5.925 3.13e-09 ***
## MONTH8           1.534e-01  9.187e-03   16.701  < 2e-16 ***
## MONTH9          -8.631e-01  1.165e-02  -74.113  < 2e-16 ***
## MONTH10         -1.083e+00  1.224e-02  -88.439  < 2e-16 ***
## MONTH11         -5.325e-02  9.834e-03   -5.415 6.13e-08 ***
## MONTH12          3.531e-01  8.911e-03   39.630  < 2e-16 ***
## AIRLINEAA        1.882e+00  6.937e-03  271.319  < 2e-16 ***
## AIRLINEOO        2.303e+00  7.107e-03  324.066  < 2e-16 ***
## AIRLINEEV        2.214e+00  7.109e-03  311.492  < 2e-16 ***
## DEPARTURE_DELAY  1.217e-02  3.667e-05  331.949  < 2e-16 ***
## DISTANCE        -2.624e-04  4.141e-06  -63.373  < 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: 1976335  on 1425623  degrees of freedom
## Residual deviance: 1556343  on 1425607  degrees of freedom
## AIC: 1556377
##
## Number of Fisher Scoring iterations: 5
```

```
exp(model_upsample$coefficients)
```

```
##     (Intercept)          MONTH2          MONTH3          MONTH4          MONTH5
##       0.1033014       2.0183276       1.2802932       0.9740505       1.2757156
```

```
##          MONTH6          MONTH7          MONTH8          MONTH9         MONTH10
##       0.9279477       1.0554206       1.1658263       0.4218336       0.3386732
##         MONTH11         MONTH12       AIRLINEAA       AIRLINEOO       AIRLINEEV
##       0.9481402       1.4235156       6.5678480      10.0045599       9.1567086
## DEPARTURE_DELAY        DISTANCE
##       1.0122463       0.9997376
```

```
pred_upsample = predict(model_upsample, test[-5], type = 'response')
pred_sample = ifelse(pred_upsample > 0.5, 1, 0)
pred_fl = as.factor(pred_sample)

library(caret)
confusionMatrix(pred_fl,test$CANCELLED)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 206154     94
##          1  99337    287
##
##                Accuracy : 0.6749
##                  95% CI : (0.6733, 0.6766)
##     No Information Rate : 0.9988
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0033
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.674828
##             Specificity : 0.753281
##          Pos Pred Value : 0.999544
##          Neg Pred Value : 0.002881
##              Prevalence : 0.998754
##          Detection Rate : 0.673988
##    Detection Prevalence : 0.674295
##       Balanced Accuracy : 0.714055
##
##        'Positive' Class : 0
##
```

Now try without month since that could be affecting the prediction of canceling. For this model, all variables are signficant. For airlines, compared to DL, AA has an odds of 5.9567992,OO 9.5136029 and EV 8.8455421 times greater. The airline with highest odds of getting cancelled is OO.

```
samp_nomonth <- up_samp %>% select(-c('MONTH'))
test_nomonth <- test %>% select(-c('MONTH'))

model_nomonth <- glm(samp_nomonth$Class ~., family = binomial(link = 'logit'), data = samp_nomonth)
summary(model_nomonth)
```

```
##
## Call:
## glm(formula = samp_nomonth$Class ~ ., family = binomial(link = "logit"),
##     data = samp_nomonth)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7481  -1.0215  -0.1309   1.0188   2.3802
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.173e+00  7.006e-03 -310.16   <2e-16 ***
## AIRLINEAA        1.785e+00  6.731e-03  265.10   <2e-16 ***
## AIRLINEOO        2.253e+00  6.948e-03  324.24   <2e-16 ***
## AIRLINEEV        2.180e+00  6.975e-03  312.55   <2e-16 ***
## DEPARTURE_DELAY  1.247e-02  3.642e-05  342.46   <2e-16 ***
## DISTANCE        -2.317e-04  4.061e-06  -57.05   <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: 1976335  on 1425623  degrees of freedom
## Residual deviance: 1597395  on 1425618  degrees of freedom
## AIC: 1597407
##
## Number of Fisher Scoring iterations: 5
```

```
exp(model_nomonth$coefficients)
```

```
##     (Intercept)       AIRLINEAA       AIRLINEOO       AIRLINEEV DEPARTURE_DELAY
##       0.1138483       5.9567992       9.5136029       8.8455421       1.0125516
##        DISTANCE
##       0.9997683
```

```
pred_nomonth = predict(model_nomonth, test_nomonth[-4], type = 'response')
pred_sample_n = ifelse(pred_nomonth > 0.5, 1, 0)
pred_n = as.factor(pred_sample_n)

library(caret)
confusionMatrix(pred_n,test$CANCELLED)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 205174    106
##          1 100317    275
##
##                Accuracy : 0.6717
##                  95% CI : (0.67, 0.6733)
##     No Information Rate : 0.9988
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.003
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.671620
```

```
##             Specificity : 0.721785
##          Pos Pred Value : 0.999484
##          Neg Pred Value : 0.002734
##              Prevalence : 0.998754
##          Detection Rate : 0.670784
##    Detection Prevalence : 0.671130
##       Balanced Accuracy : 0.696703
##
##        'Positive' Class : 0
##
```

Another item to check is to build model with seasonal grouping. In this case, AA odds are 6.37794813 , OO are 9.66681983, and EV would be 8.93214550,times that of canceling for DL airline. Autumn has the least odds of cancelling. The airline with highest odds of getting canceled is OO.

```
ver3 <- flight_delay_new %>% mutate(season = case_when(MONTH ==12|MONTH==1|MONTH==2 ~ "winter",
                                                       MONTH == 3|MONTH==4|MONTH==5 ~ "spring",
                                                       MONTH == 6|MONTH==7|MONTH==8 ~ "summer",
                                                       MONTH == 9|MONTH==10|MONTH==11 ~ "autumn")) %>%
                                    select(-c("MONTH"))

ver3$season = as.factor(ver3$season)

set.seed(456)
split3 = sample.split(ver3$CANCELLED, SplitRatio = 0.7)
training3 = subset(ver3, split3 == TRUE)
test3 = subset(ver3, split3 == FALSE)
## upsample because of lack of Cancel = 1
set.seed(1234)
up_samp3 <- upSample(x = training3[,-4], y=training3$CANCELLED)
table(up_samp3$Class)
```

```
##
##      0      1
## 712812 712812
```

```
model_season <- glm(up_samp3$Class ~., family = binomial(link = 'logit'), data = up_samp3)
summary(model_season)
```

```
##
## Call:
## glm(formula = up_samp3$Class ~ ., family = binomial(link = "logit"),
##     data = up_samp3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.6676  -0.9926  -0.0971   0.9794   2.4123
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.793e+00  8.436e-03 -331.09   <2e-16 ***
## AIRLINEAA        1.853e+00  6.830e-03  271.29   <2e-16 ***
## AIRLINEOO        2.269e+00  7.016e-03  323.35   <2e-16 ***
## AIRLINEEV        2.190e+00  7.035e-03  311.25   <2e-16 ***
## DEPARTURE_DELAY  1.218e-02  3.646e-05  334.04   <2e-16 ***
## DISTANCE        -2.514e-04  4.104e-06  -61.24   <2e-16 ***
```

```
## seasonspring     7.084e-01  6.231e-03  113.69   <2e-16 ***
## seasonsummer     5.896e-01  5.987e-03   98.48   <2e-16 ***
## seasonwinter     9.160e-01  6.000e-03  152.67   <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: 1976335  on 1425623  degrees of freedom
## Residual deviance: 1572813  on 1425615  degrees of freedom
## AIC: 1572831
##
## Number of Fisher Scoring iterations: 5
```

```
exp(model_season$coefficients)
```

```
##     (Intercept)       AIRLINEAA       AIRLINEOO       AIRLINEEV DEPARTURE_DELAY
##      0.06123122      6.37794813      9.66681983      8.93214550      1.01225231
##        DISTANCE    seasonspring    seasonsummer    seasonwinter
##      0.99974867      2.03064071      1.80327591      2.49934594
```

```
pred_season = predict(model_season, test3[-4], type = 'response')
pred_season_res = ifelse(pred_season > 0.5, 1, 0)
pred_f_season = as.factor(pred_season_res)

library(caret)
confusionMatrix(pred_f_season,test3$CANCELLED)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction      0      1
##          0 203269     93
##          1 102222    288
##
##                Accuracy : 0.6655
##                  95% CI : (0.6638, 0.6672)
##     No Information Rate : 0.9988
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0031
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.665385
##             Specificity : 0.755906
##          Pos Pred Value : 0.999543
##          Neg Pred Value : 0.002809
##              Prevalence : 0.998754
##          Detection Rate : 0.664556
##    Detection Prevalence : 0.664860
##       Balanced Accuracy : 0.710645
##
##        'Positive' Class : 0
##
```

After looking at all the different models, after upsampling, the airline with highest odds of getting cancelld is OO.