FORECASTING AIRLINE DELAYS

On any given day, more than 87,000 flights take place in the United States alone. About one-third of these flights are commercial flights, operated by companies like United, American Airlines, and JetBlue. While about 80% of commercial flights take-off and land as scheduled, the other 20% suffer from delays due to various reasons. A certain number of delays are unavoidable, due to unexpected events, but some delays could hopefully be avoided if the factors causing delays were better understood and addressed.

In this problem, we'll use a dataset of 9,381 flights that occured in June through August of 2014 between the three busiest US airports -- Atlanta (ATL), Los Angeles (LAX), and Chicago (ORD) -- to predict flight delays. The dataset AirlineDelay.csv includes the following 23 variables:

- **Flight** = the origin-destination pair (LAX-ORD, ATL-LAX, etc.)
- **Carrier** = the carrier operating the flight (American Airlines, Delta Air Lines, etc.)
- **Month** = the month of the flight (June, July, or August)
- **DayOfWeek** = the day of the week of the flight (Monday, Tuesday, etc.)
- **NumPrevFlights** = the number of previous flights taken by this aircraft in the same day
- **PrevFlightGap** = the amount of time between when this flight's aircraft is scheduled to arrive at the airport and when it's scheduled to depart for this flight
- **HistoricallyLate** = the proportion of time this flight has been late historically
- **InsufficientHistory** = whether or not we have enough data to determine the historical record of the flight (equal to 1 if we don't have at least 3 records, equal to 0 if we do)
- **OriginInVolume** = the amount of incoming traffic volume at the origin airport, normalized by the typical volume during the flight's time and day of the week
- **OriginOutVolume** = the amount of outgoing traffic volume at the origin airport, normalized by the typical volume during the flight's time and day of the week
- **DestInVolume** = the amount of incoming traffic volume at the destination airport, normalized by the typical volume during the flight's time and day of the week
- **DestOutVolume** = the amount of outgoing traffic volume at the destination airport, normalized by the typical volume during the flight's time and day of the week
- **OriginPrecip** = the amount of rain at the origin over the course of the day, in tenths of millimeters
- **OriginAvgWind** = average daily wind speed at the origin, in miles per hour
- **OriginWindGust** = fastest wind speed during the day at the origin, in miles per hour
- **OriginFog** = whether or not there was fog at some point during the day at the origin (1 if there was, 0 if there wasn't)
- **OriginThunder** = whether or not there was thunder at some point during the day at the origin (1 if there was, 0 if there wasn't)
- **DestPrecip** = the amount of rain at the destination over the course of the day, in tenths of millimeters
- **DestAvgWind** = average daily wind speed at the destination, in miles per hour
- **DestWindGust** = fastest wind speed during the day at the destination, in miles per hour

- **DestFog** = whether or not there was fog at some point during the day at the destination (1 if there was, 0 if there wasn't)
- **DestThunder** = whether or not there was thunder at some point during the day at the destination (1 if there was, 0 if there wasn't)
- **TotalDelay** = the amount of time the aircraft was delayed, in minutes (this is our dependent variable)

PROBLEM 1 - LOADING THE DATA (2/2 points)

Load the dataset AirlineDelay.csv into R and call it "Airlines". Randomly split it into a training set (70% of the data) and testing set (30% of the data) by running the following lines in your R console:

```
set.seed(15071)

spl = sample(nrow(Airlines), 0.7*nrow(Airlines))

AirlinesTrain = Airlines[spl,]

AirlinesTest = Airlines[-spl,]
```

How many observations are in the training set AirlinesTrain?

6566 **Answer:** 6566

How many observations are in the testing set AirlinesTest?

2815 **Answer:** 2815

EXPLANATION

The dataset can be loaded into R using the read.csv function. After following the lines given to split the data, the number of observations can be found with the nrow or str functions.

Check | **Save** | Hide Answer *You have used 1 of 3 submissions*

PROBLEM 2 - METHOD OF SPLITTING THE DATA (1/1 point)

In this class, we have frequently used the sample.split function to randomly split our data. Why do we use a different approach here?

○ We don't want to randomly split our data here, so the sample.split function is not appropriate.

● The sample.split function is typically used to split data with a categorical dependent variable, and we have a continuous dependent variable. ✓

○ The sample.split function is typically used to split data with a continuous dependent variable, and we have a categorical dependent variable.

○ The sample.split function will give an error message if our dependent variable is not a factor variable.

○ We have too many observations for the sample.split function to work.

EXPLANATION

The sample.split function is used to split data for a classification problem (a categorical dependent variable) and we have a continuous dependent variable here.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 3 - A LINEAR REGRESSION MODEL (1/1 point)

Build a linear regression model to predict "TotalDelay" using all of the other variables as independent variables. Use the training set to build the model.

What is the model's R-squared? (Please report the "Multiple R-squared" value in the output.)

0.09475268 **Answer:** 0.09475

EXPLANATION

You can build the linear regression model using the lm function, and find the R-squared by looking at the summary of the resulting model.

Final Check **Save** Hide Answer *You have used 1 of 2 submissions*

PROBLEM 4 - CHECKING FOR SIGNIFICANCE (1/1 point)

In your linear regression model, which of the independent variables are significant at the p=0.05 level (at least one star)? For factor variables, consider the variable significant if at least one level is significant. Select all that apply.

- ☑ Flight ✓
- □ Carrier
- ☑ Month ✓
- ☑ DayOfWeek ✓
- ☑ NumPrevFlights ✓
- ☑ PrevFlightGap ✓
- ☑ HistoricallyLate ✓
- ☑ InsufficientHistory ✓
- □ OriginInVolume
- □ OriginOutVolume
- ☑ DestInVolume ✓
- □ DestOutVolume
- ☑ OriginPrecip ✓
- ☑ OriginAvgWind ✓
- ☑ OriginWindGust ✓
- □ OriginFog
- □ OriginThunder
- ☑ DestPrecip ✓
- □ DestAvgWind

☑ DestWindGust ✓

□ DestFog

□ DestThunder

EXPLANATION

The significance can be found by looking at the summary output.

Final Check | **Save** | **Hide Answer** *You have used 1 of 2 submissions*

PROBLEM 5 - CORRELATIONS (2/2 points)

What is the correlation between NumPrevFlights and PrevFlightGap in the training set?

-0.6520532 **Answer:** -0.652053189

What is the correlation between OriginAvgWind and OriginWindGust in the training set?

0.5099535 **Answer:** 0.509953488

EXPLANATION

The correlations can be computed using the cor function.

Final Check | **Save** | **Hide Answer** *You have used 2 of 3 submissions*

PROBLEM 6 - IMPORTANCE OF CORRELATIONS (1 point possible)

Why is it imporant to check for correlations between independent variables? Select all that apply.

☑ Having highly correlated independent variables in a regression model can affect the interpretation of the coefficients. ✓

☑ Having highly correlated independent variables in a regression model can affect the quality of the resulting predictions.

EXPLANATION

Highly correlated independent variables can affect the interpretation of the coefficients. We won't worry about dealing with correlated independent variables here, but if interpreting the coefficients is important, multicollinearity should be addressed.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 7 - COEFFICIENTS (1/1 point)

In the model with all of the available independent variables, what is the coefficient for HistoricallyLate?

47.913638 **Answer:** 47.913638

EXPLANATION

The coefficient for HistoricallyLate can be found in the summary output of the model.

Final Check | **Save** | Hide Answer | *You have used 1 of 2 submissions*

PROBLEM 8 - UNDERSTANDING THE COEFFICIENTS (1/1 point)

The coefficient for NumPrevFlights is 1.56. What is the interpretation of this coefficient?

○ For an increase of 1 in the predicted total delay, the number of previous flights increases by approximately 1.56.

● For an increase of 1 in the number of previous flights, the prediction of the total delay increases by approximately 1.56. ✓

○ If the number of previous flights increases by 1, then the total delay will definitely increase by approximately 1.56; the number of previous flights should be minimized if airlines want to decrease the amount of delay.

EXPLANATION

The second choice is the correct answer; the coefficient is defined as the change in the prediction of the dependent variable per unit change in the independent variable in question. The first choice is not correct because it flips the relationship.

The third choice is not correct because the coefficient indicates how the prediction changes, not how the actual value changes, and this option asserts that actual delay changes, i.e., there is a causal effect.

Hide Answer | *You have used 1 of 1 submissions*

PROBLEM 9 - UNDERSTANDING THE MODEL (2/2 points)

Let us try to understand our model.

In the linear regression model, given two flights that are otherwise identical, what is the absolute difference in predicted total delay given that one flight is on Thursday and the other is on Sunday?

6.989857 **Answer:** 6.989857

In the linear regression model, given two flights that are otherwise identical, what is the absolute difference in predicted total delay given that one flight is on Saturday and the other is on Sunday?

0.911413 **Answer:** 0.911413

EXPLANATION

The coefficient for DayOfWeekSunday is -5.418356 (look at the summary output of the model). For the first question, the coefficient of DayOfWeekThursday is 1.571501, so the absolute difference in the prediction is abs(-5.418356 - 1.571501) = 6.989857

For the second question, the coefficient of DayOfWeekSaturday is -4.506943, so the absolute difference in the prediction is

abs(-5.418356 - (-4.506943)) = 0.911413

Final Check | Save | Hide Answer | *You have used 2 of 3 submissions*

PROBLEM 10 - PREDICTIONS ON THE TEST SET (3/3 points)

Make predictions on the test set using your linear regression model. What is the Sum of Squared Errors (SSE) on the test set?

4744764 **Answer:** 4744764

What is the Total Sum of Squares (SST) on the test set? Remember to use the mean total delay on the training set as the "baseline model".

5234023 **Answer:** 5234023

What is the R-squared on the test set?

0.09347674 **Answer:** 0.09347674

EXPLANATION

Predictions on the test set can be computed with the predict function. Then the SSE can be computed by taking the sum of the squared differences between the predictions and the actual values, the SST can be computed by taking the sum of the squared differences between the mean on the training set and the actual values, and the R-squared can be computed as 1 - SSE/SST.

Check | Save | Hide Answer | *You have used 1 of 3 submissions*

PROBLEM 11 - EVALUATING THE MODEL (1/1 point)

Given what you have seen about this model (the R-squared on the training and test sets, the significance of the coefficients, etc.), which of the following are true? Select all that apply.

- ☑ Since our R-squared values are low, we can conclude that our independent variables only explain a small amount of the variation in the dependent variable. ✓
- ☐ We can't learn anything about what correlates with higher flight delays from this model.
- ☐ Since our R-squared value is much higher on the training set than the test set, we probably overfit our model to the training set by using all of the available independent variables.

EXPLANATION

We can conclude that this is a challenging problem, and our independent variables only explain a small amount of the variation in delays. However, since we do have some significant independent variables, we can get some insights from factors that correlate with delays. Additionally, it looks like we did not overfit our model because the R-squared on the training set and the R-squared on the test set are similar.

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 12 - A CLASSIFICATION PROBLEM (3/3 points)

Let's turn this problem into a multi-class classification problem by creating a new dependent variable. Our new dependent variable will take three different values: "No Delay", "Minor Delay", and "Major Delay". Create this variable, called "DelayClass", in your dataset Airlines by running the following line in your R console:

```
Airlines$DelayClass = factor(ifelse(Airlines$TotalDelay == 0, "No Delay", ifelse(Airlines$TotalDelay >= 30, "Major Delay", "Minor Delay")))
```

Note that a minor delay is a delay less than 30 minutes long, and a major delay is a delay at least 30 minutes long.

How many flights in the dataset Airlines had no delay?

4688 **Answer:** 4688

How many flights in the dataset Airlines had a minor delay?

3096 **Answer:** 3096

How many flights in the dataset Airlines had a major delay?

1597 **Answer:** 1597

EXPLANATION

After running the given lines in your R console, you can find these answers with the table or summary function.

Now, remove the original dependent variable "TotalDelay" from your dataset with the command:

```
Airlines$TotalDelay = NULL
```

Then randomly split Airlines into a training set, containing 70% of the observations, and a testing set, containing 30% of the observations. This time, you should use the sample.split function with the dependent variable "DelayClass". Right before calling the sample.split function, set the seed to 15071. Remember that you should set SplitRatio to 0.7. The TRUE observations from the split should go into the training set, and the FALSE observations should go into the testing set.

Check **Save** Hide Answer *You have used 1 of 3 submissions*

PROBLEM 13 - A CART MODEL (1 point possible)

Build a CART model to predict "DelayClass" using all of the other variables as independent variables and the training set to build the model. Remember that to predict a multi-class dependent variable, you can use the rpart function in the same way as for a binary classification problem. Just use the default parameter settings (don't set a value for minbucket or cp).

How many split are in the resulting tree?

2

Show Answer *You have used 2 of 2 submissions*

PROBLEM 14 - UNDERSTANDING THE MODEL (1/1 point)

The CART model you just built never predicts one of the three outcomes. Which one?

○ No Delay

○ Minor Delay

● Major Delay ✓

EXPLANATION

The CART model never predicts "Major Delay".

Hide Answer *You have used 1 of 1 submissions*

PROBLEM 15 - TRAINING SET ACCURACY (1/1 point)

Make predictions on the training set, and then create a confusion matrix. What is the overall accuracy of the model?

0.5261154 **Answer:** 0.5261154

EXPLANATION

Predictions can be computed with the predict function (remembering to add the argument type="class"), and then the confusion matrix can be created with the table function.

Final Check **Save** Hide Answer *You have used 1 of 2 submissions*

PROBLEM 16 - A BASELINE MODEL (1/1 point)

What is the accuracy on the training set of a baseline model that predicts the most frequent outcome (No Delay) for all observations?

0.4997716 **Answer:** 0.4997716

EXPLANATION

The accuracy of the baseline model can be computed by using the table function to see that 3282 observations have no delay, out of the 6567 total observations.

Final Check **Save** Hide Answer *You have used 1 of 2 submissions*

PROBLEM 17 - TESTING SET ACCURACY (1/1 point)

Make predictions on the testing set, and then create a confusion matrix. What is the overall accuracy of the model on the testing set?

0.5167022

Answer: 0.5167022

EXPLANATION

Predictions can be computed with the predict function, and then the confusion matrix can be created with the table function.

Final Check | **Save** | **Hide Answer** | *You have used 1 of 2 submissions*

PROBLEM 18 - UNDERSTANDING THE MODEL (1 point possible)

What can you conclude from the CART model? Select all that apply.

☑ Turning this problem into a classification problem significantly improved our ability to predict delays.

☑ Out of the independent variables in our dataset, the best predictor of future delays is historical delays. ✓

☑ While delays are hard to predict, using historical data can be helpful. ✓

☑ The CART model was able to find significant variables that the linear regression model didn't find.

EXPLANATION

According to the CART model, the best predictor of future delays is historical delays, which helps improve our predictive ability. Turning this problem into a classification problem did not really improve our ability to predict delays, and the CART model didn't really give us any insights that we didn't see in the linear regression model. However, the CART model is very simple, and could be easier to interpret.

Hide Answer | *You have used 1 of 1 submissions*

‹ ›

EdX offers interactive online classes and MOOCs from the world's best universities. Online courses from MITx, HarvardX, BerkeleyX, UTx and many other universities. Topics include biology, business, chemistry, computer science, economics, finance, electronics, engineering, food and nutrition, history, humanities, law, literature,

About edX

About

News

Contact

FAQ

Follow Us

Facebook

Twitter

LinkedIn

math, medicine, music, philosophy, physics, science, statistics and more. EdX is a non-profit online initiative created by founding partners Harvard and MIT.

Terms of Service and Honor Code

Privacy Policy (Revised 10/22/2014)

POWERED BY OPEN edX

edX Blog

Donate to edX

Jobs at edX

Google+

Tumblr

Meetup

Reddit

Youtube