

# Flight Arrival Delays Inference and Prediction

Mary Silva<sup>1</sup>, Wyara Vanesa Moura e Silva<sup>1</sup>  
Graduate Student at UC Santa Cruz<sup>1</sup>

## Abstract

The United States Department of Transportation keeps records of departures, arrivals, destinations and various other flight details for most major airlines on an hourly basis. Many of the flights report a delay in arrival, which is the cause of many problems for travelers. The objective for this project is to determine which variables contribute to delays in flight arrival and to produce a model capable of predicting delays based on those variables. Two approaches are used to analyze the data. First, we produce a model in which the cities are separated. Second we attempt to model the data using cities as a categorical variable. In both cases, logistic regression was used. Initially we considered eight different variables, some of which were weather variables obtained from a separate source. Then, we attempt to modeling this data using different combinations of variables to see which of them best fit the data. To determine the best fitting model, we use several types of criterion: Receiver Operating Characteristic (ROC) areas under the curve (AUC), Akaike information criterion (AIC), the error rates of 10-fold cross validation, and the Hosmer-Lemeshow goodness of fit test. We observe that different combinations of variables including rainfall, distance of flight, visibility, wind speed, and temperature are statistically significant in determining the probability of having a flight delay.

**KEY WORDS:** Flight data, airport, weather, logistic regression.

## 1. Introduction

The U.S. Department of Transportation's *Airline On-Time Performance Data* (2016) contains over 5,000,000 flight records from 13 airline carriers and 311 origin cities for the year 2016. Among these records are the arrival delay in minutes and the distance of each flight. *The National Centers for Environmental Information* (2016) Local Climatological Data (LCD) data set is a source of hourly, daily, and monthly weather summaries for approximately 1,600 U.S. locations. From this data set, temperature, wind speed, visibility and precipitation can be found specific to each location.

For this project, logistic regression will be the primary method for modeling flight departure delays. The response will be treated as binary. Initially we separated the delays less than 30 minutes from the delays greater than or equal to 30 minutes. Later, we adjusted this cutoff. The explanatory variables will be weather, distance of flight, and date of flight. Models with fewer or

more predictors will be examined, as well as models with interaction terms added. Various criterion will be performed to choose the "best" fitting model. Ultimately, the goal is to produce a model that successfully predicts the probability of a flight delay occurring. All models and corresponding hypothesis tests will be obtained using R programming.

## 2. Exploring the Data Set

The airline data and the weather data must first be formatted so that they are compatible with each other. The problem here is that the weather data is given on an hourly interval, while the airline data contains hundreds of observations for each hour. For instance, one flight departing from San Francisco International Airport at 11:00 am may report a delay of 3 minutes, while another flight departing at the same time from the same airport may report a delay of 80 minutes. Similarly those two flights may have vastly different planned flight distances.

One solution to this problem was to take the mean of all the delay minutes as well as a mean of the distances of each flight for each hour. An alternate solution would be to assign a temperature corresponding to each hour for a specific airport location and specific day.

Another problem with our source of weather data is that the data must be downloaded separately by city. For simplicity, and to make the data from the airline delays compatible with the temporal and spacial locations, we have chosen to download only the weather data from high traffic airports.

## 3. Determining High Traffic Locations

We begin by separating the airline data by location and by the number of flights departing each location. Doing this allows us to separate the locations into categories of high, medium, and low traffic locations.

Using a dataset from Socrata (2017), *Airport Codes mapped to Latitude/Longitude in the United States*, we obtain the latitudinal and longitudinal coordinates of 13,400 airports across the country and match the Airport codes from this dataset to the Airline data set. We overlay these coordinates on to an image from Google maps. In Figure 1, the diameter of the circles correspond to the number of flight departures for that location relative to other locations. We see which airports correspond to low traffic (less than 18,000 flights per year), medium traffic (18,000 to 80,000 flights per year) and high traffic (more than 80,000 flights per year).

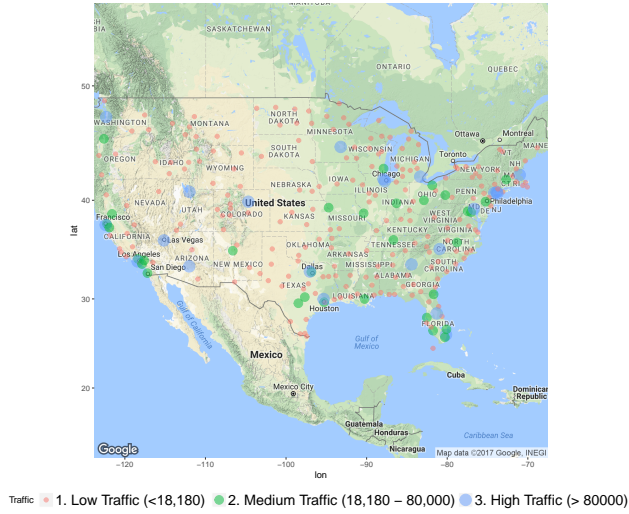


Figure 1: Airport locations

For simplicity, we focus on the 22 high traffic airports; of which, weather data can be obtained for 15 airports.

#### 4. Logistic Regression of Separate Cities

First, we attempt to model the delays of each of the 15 high traffic airports separately. After combining the datasets, we first separated the delay minutes into less than 30 minutes and greater than or equal to 30 minutes.

**Definition 1** *In logistic regression, we use the logistic function,*

$$p(X) = P(\text{Delay} = \text{Yes}|X) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

Where  $p(X)$  represents the response variable specific to each location and the list of possible predictors are

- $X_V$ : Visibility (a measure of the distance of visibility in feet),
- $X_{V_f}$ : Visibility as a factor (High/Low),
- $X_J$ : Day of the year represented by Julian calendar,
- $X_T$ : Temperature (degrees Fahrenheit),
- $X_H$ : Hour of the day (24-hr),
- $X_W$ : Wind Speed (miles per hour),
- $X_R$ : Rainfall (in hundredths of an inch),
- $X_D$ : Distance of Flight (Median distances for flights by hour)

The method of estimation used in this paper will be based in the maximum likelihood approach, for more details about the theory behind this method of estimation see Hosmer et al. (2000).

In order to compare the performance of one model over another, several criterion will be used: the Receiver Operating Characteristic (ROC) areas under the curve (AUC), Akaike information criterion (AIC), the error rates of 10-fold cross validation, and the Hosmer-Lemeshow goodness of fit.

The ROC curve is the sensitivity (true positive proportion) plotted against the false positive proportion for a range of threshold probabilities. A smooth curve is drawn through the points to derive the ROC curve. The areas under the ROC curve, which ranges from zero to one, provides a measure of the model's ability to discriminate between those subjects who experience the outcome of interest versus those who do not. (Hosmer et al., 2000)

As a general rule given by Hosmer et al. (2000):

- If  $\text{ROC}=0.5$ : this suggests no discrimination;
- If  $0.7 \leq \text{ROC} < 0.8$ : this is considered acceptable discrimination;
- If  $0.8 \leq \text{ROC} < 0.9$ : this is considered excellent discrimination;
- If  $\text{ROC} \geq 0.9$ : this is considered outstanding discrimination.

It should be mentioned that, in practice, it is extremely unusual to observe areas under the ROC curve greater than 0.9.

The AIC was first developed by Akaike (1973) as a criterion to compare different models on a given outcome. The AIC is calculated using the number of model parameters ( $k$ ) and the log-likelihood ( $\hat{\mathcal{L}}$ ), as we can see in Equation (1)

$$\text{AIC} = -2(\hat{\mathcal{L}}) + 2k \quad (1)$$

Thus, the AIC deal with the trade-off between the goodness of fit of the model and the complexity of the model. In addition, it penalizes the addition of parameters, and thus selects a model that fits well but has a minimum number of parameters. The model with the smaller AIC will be the one that fit best.

Cross validation is done by separating the data set into  $k$  equal folds where each fold is held as the validation set while the remainder is used to fit the model. For our data, we select  $k = 10$  folds. The validation set is then used to calculate the error rate of the model. This procedure is repeated ten times and the 10-fold cross validation estimate is computed by averaging these values (James et al., 2013). In other words,

$$CV(10) = \frac{1}{10} \sum_{i=1}^{10} MSE_i$$

We also apply the Hosmer-Lemeshow Goodness of Fit test to the logistic regression models. For the Hosmer-Lemeshow goodness of fit test, samples of the predicted probabilities are divided up based on the  $\beta$  estimates for each observation in the sample, and the probability that

Delay = Yes is calculated. The observations in the sample are then split up into 10 groups. Then the first group consists of the observations with the lowest 10% predicted probabilities. The second group consists of the 10% of the sample whose predicted probabilities are next smallest, and so on (Hosmer et al., 2000). The same procedure is done for the calculation of the probabilities that Delay = no. The Pearson goodness of fit test statistic is

$$\sum_{i=0}^1 \sum_{j=0}^{10} \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}}$$

Where the  $Obs_{0j}$  is the number of observed Delay = No observation in the  $j$ th group and  $Obs_{1j}$  denotes the number of Delay = yes observations in the  $j$ th group. The test statistic approximately follows Chi-squared with 10-2 degrees of freedom.

**Definition 2** *The hypothesis test for the Hosmer-Lemeshow goodness of fit is as follows*

$H_0$ : The model is a good fit for the data

$H_a$ : The model is a poor fit for the data

#### 4.1 Logistic Regression Model For Baltimore-Washington International Airport

Examining each location separately, we find that variables which are statistically significant for one city, are statistically insignificant for the model of another city. The model for Baltimore-Washington International (BWI) Airport, which had over 80,000 flight departures, showed that hour of the day and day of the year did not significantly affect the probability of a flight delay. For this airport there were a recorded 94,382 flights from 13 major airline carriers in the year 2016.

The plot of visibility against the arrival delays in minutes from Figure 2 indicate that visibility could be treated as a categorical variable—separating visibility into a factor of high and low visibility. The mean of the visibility recorded at BWI is used as cutoff for high and low visibility. Low is considered less than 9 feet visibility and high is greater than or equal to 9 feet of visibility.

Using 30 minutes as a cutoff for the delays, we produce separate models with statistically significant variables and compare the ROC areas under the curve (AUC), AIC, and the error rates of 10-fold cross validation. A summary of the results are shown in Table 1.

Table 1: Using 30 minutes as a cutoff delay for BWI Airport.

Predictors Included	AUC	AIC	CV
$X_R, \log(X_D), X_V, X_W$	0.666	1779.500	0.029
$X_R, X_D, X_V$	0.600	1806.200	0.029
$X_R, \log(X_D), X_V, X_W$	0.665	1785.200	0.029
$X_R, X_D, X_V, X_W$	0.665	1775.100	0.029
$X_R, X_D, X_{V_f}, X_W$	0.665	1780.800	0.029

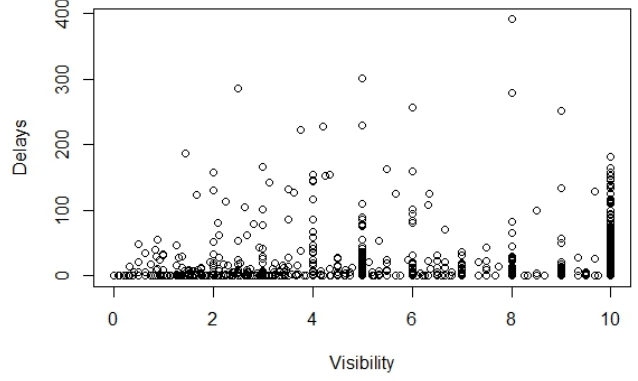


Figure 2: Scatter plot of Visibility versus Delays for BWI Airport.

The best area under the ROC curve we are able to obtain with 30 minutes as the cutoff is 67%, meaning that the true positive rate is not ideal. Raising this cutoff does not affect the significance of each variable, but it does reduce the cross validation error.

Using 60 minutes as a cutoff for delays, we are able to obtain a higher AUC and lower cross validation error rate. A summary of the results are shown in Table 2. Ultimately the model that produces the highest AUC and lowest AIC and cross validation error is the model that uses precipitation, distance of flight, visibility as a factor, and wind speed as predictors.

While the prediction accuracy of the model improves by increasing the cutoff for delays, the consequence is that a delay is now considered to be anything over one hour.

Table 2: Using 60 minutes as a cutoff delay for BWI Airport.

Predictors Included	AUC	AIC	CV
$X_R, \log(X_D), X_V$	0.622	955.000	0.014
$X_R, X_D, X_V$	0.663	949.100	0.014
$X_R, \log(X_D), X_V, X_W$	0.750	918.300	0.014
$X_R, X_D, X_V, X_W$	0.751	912.100	0.013
$X_R, X_D, X_{V_f}, X_W$	0.759	911.800	0.014

In Table 3 on the next page we have the estimates for the predictors using the models with rainfall, distance of flight, visibility (as factor) and wind speed. It can be seen that all predictors are significant. The intercept represents the log odds for a delay occurring when visibility is low. The odds ratio is represented by

$$\frac{\hat{p}(X)}{1 - \hat{p}(X)} = e^{\hat{\beta}_0 + \hat{\beta}_1 X_R + \hat{\beta}_2 X_D + \hat{\beta}_3 X_V + \hat{\beta}_4 X_W}$$

Examining the  $\beta$  estimates, we find that every one inch increase in rainfall increases the odds of having a delay

Table 3: Estimates of the  $\beta$ s for the model with  $X_R, X_D, X_{V_f}, X_W$  as predictors, using 60 minutes cutoff of delays for BWI Airport.

	<i>Estimate</i>	<i>Std. Error</i>	<i>t value</i>	<i>Pr(&gt;  t )</i>	
(Intercept) $\hat{\beta}_0$	-4.3722604	0.3934502	-11.113	< 2e-16	***
Rainfall $\hat{\beta}_1$	0.1205393	0.0175747	6.859	6.95e-12	***
Distance $\hat{\beta}_2$	0.0021077	0.0005092	4.140	3.48e-05	***
Visibility High $\hat{\beta}_3$	-0.8465841	0.2183840	-3.877	0.000106	***
Wind Speed $\hat{\beta}_4$	-0.1504343	0.0271698	-5.537	3.08e-08	***

by 13%. Furthermore, we expect to see a 57% decrease in the odds of having a delay when visibility is high (see Figure 3). While distance of flight is statistically significant, we notice that the  $\beta$  estimate is very small, meaning that for every one unit increase in distance we expect less than 1% increase in the odds of having a delay. Further examination of the original data we find that flights range in distance from 67 miles to 2585 miles. The average distance of flight is 855 miles, therefore a small estimate for  $\beta$  is reasonable.

When distance of flight and wind speed are the mean of both variables, Figure 3 represents the adjusted probabilities of having a delay for high and low visibility. The plot shows us that as the amount of rainfall increases, the probability of having a delay increases.

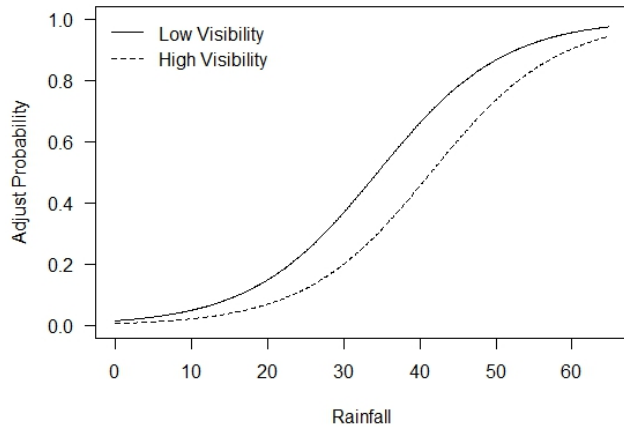


Figure 3: Adjust probability for delays less than 60 minutes, for BWI Airport.

Figure 4 is the ROC curve for the model with the highest AUC. The model was found to have an acceptable discrimination ability, having an area under the ROC curve of 0.759.

Next, we use the Hosmer-Lemeshow Goodness of Fit test (Hosmer et al., 2000) to determine if this model is indeed a good fit for the data using the hypothesis test defined in Definition 2. The test yields a p-value = 0.430. For an 0.05 alpha level of significance, we fail to reject to null hypothesis and conclude that this model is a good

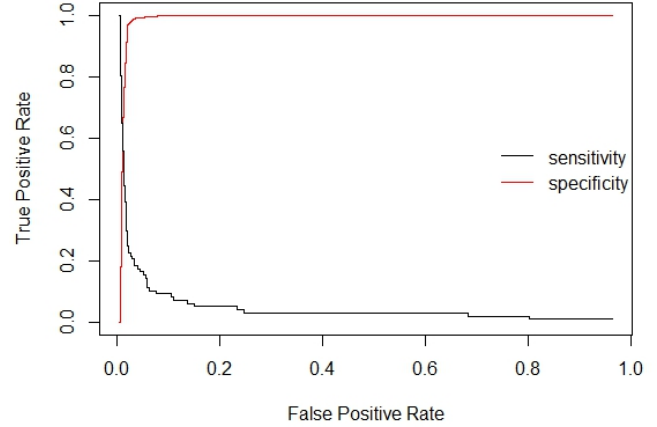


Figure 4: ROC curve for the model using  $X_R, X_D, X_{V_f}, X_W$  as predictors, for BWI Airport.

fit for the data.

## 4.2 Logistic Regression Model For Los Angeles International Airport

The model for Los Angeles International (LAX) Airport, which had 210,593 flights from 13 major airline carriers in 2016, had the same behavior as the models for BWI Airport in which the hour of the day and day of the year did not significantly affect the probability of a flight delay. Moreover, temperature did not affect the probability of the flight delay occurring.

In Figure 5, distance of flight for delays less than 30 minutes and delays greater than 30 minutes had almost the same median. There is greater variability in the distribution of distance of flight for delays less than 30 minutes than there is for delays greater than 30 minutes. For delays of less than 60 minutes, the distribution of distance of flight is not too different from the delays greater than 60 minutes; but median is higher for delays greater than 60 minutes, as we can see in Figure 6. However, we can not conclude anything from these two plots, whether or not this variable has significant influence on flight delays.

In Table 4 are the results for the estimates for five different models using 30 minutes as a cutoff for the delays.

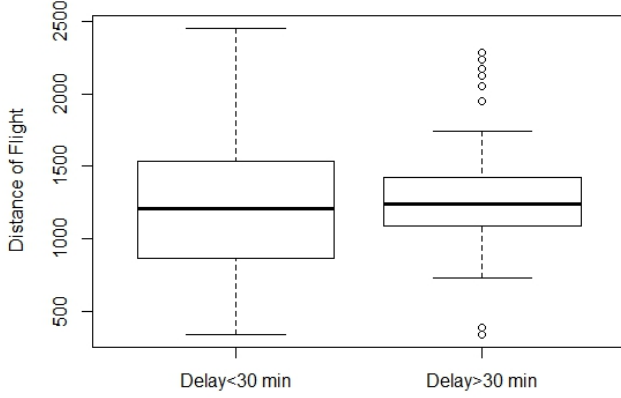


Figure 5: Boxplot for distance of flight and delays, using 30 minutes as a cutoff for delays, for LAX Airport.

The best area under the ROC curve was 65%, which is out of the range that is considered a acceptable discrimination.

Table 4: Using 30 minutes as a cutoff delay for LAX Airport.

Predictors Included	AUC	AIC	CV
$X_R, X_D, X_V$	0.633	632.200	0.008
$X_R, X_D, X_V, X_W$	0.647	633.200	0.008
$X_R, \log(X_D), X_{V_f}, X_W$	0.649	630.100	0.008
$X_R, \log(X_D), X_{V_f}$	0.634	629.200	0.008
$X_R, X_D$	0.614	631.200	0.008

Using 60 minutes as a cutoff for delays, we obtain a higher AUC and lower cross validation error rate. The results are summarized in Table 5. We were able to produce a model with an area under the curve of 84%, using rainfall, distance (with log transformation), visibility as a factor, and wind speed; however we found that visibility and wind speed are not statistically significant. Therefore, we determine that the model using only rainfall and distance of flight as predictors is the best.

Table 5: Using 60 minutes as a cutoff delay for LAX Airport.

Predictors Included	AUC	AIC	CV
$X_R, X_D, X_V$	0.830	130.100	0.001
$X_R, X_D, X_V, X_W$	0.833	132.000	0.001
$X_R, \log(X_D), X_{V_f}, X_W$	0.835	131.600	0.001
$X_R, \log(X_D), X_{V_f}$	0.834	129.700	0.001
$X_R, X_D$	0.831	128.100	0.001

In Table 6 are the estimates for the predictors using the model with only rainfall and distance of flight. The

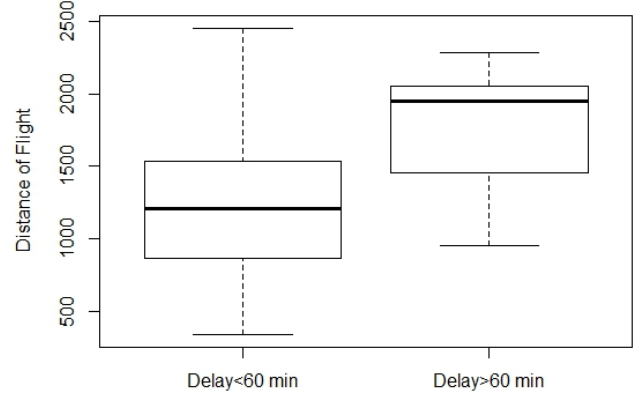


Figure 6: Boxplot for distance of flight and delays, using 60 minutes as a cutoff for delays, for LAX Airport.

odds ratio for this model is represented as follows

$$\frac{\hat{p}(X)}{1 - \hat{p}(X)} = e^{\hat{\beta}_0 + \hat{\beta}_1 X_R + \hat{\beta}_2 X_D}$$

The odds of having a delay increases by 23% for every one inch increase in rainfall. For distance of flight again we see that the  $\beta$  estimate is small, meaning for every one unit increase in distance, we expect less than 1% increase in the odds of having a delay. Again, the range of distance of flight is between 67 and 2585 miles, so this is reasonable.

Figure 7 shows the ROC curve for the model with the higher value of AUC, using 60 minutes as a cutoff for delays. The model was found to have a excellent discrimination ability, having an area under the ROC curve of 0.831.

Lastly, we evaluate the Hosmer-Lemeshow Goodness of Fit test (Hosmer et al., 2000) for the model using the hypothesis test defined in Definition 2. The test yields a p-value = 0.803. For an 0.05 alpha level of significance, we fail to reject the null hypothesis and conclude that this model is a good fit for the data.

## 5. Logistic Regression using Cities as a factor

Another logistic regression approach would be to treat origin airport locations as a factor of low, medium and high traffic locations.

**Definition 3** The logistic function is represented as

$$p(X) = P(\text{Delay} = \text{Yes} | X) = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

Where the response represents the probability of a delay occurring. The possible predictors are the same as those listed in Definition 1 but with additional categorical variables for low, medium and high traffic airports.



Table 6: Estimates of the  $\beta$ s for the model with  $X_R, X_D$  as predictors, using 60 minutes as the cutoff of delays for LAX Airport.

	<i>Estimate</i>	<i>Std. Error</i>	<i>t value</i>	<i>Pr(&gt;  t )</i>	
(Intercept) $\hat{\beta}_0$	-1.019e+01	1.396e+00	-7.301	2.85e-13	***
Rainfall $\hat{\beta}_1$	2.080e-01	5.735e-02	3.627	0.000287	***
Distance $\hat{\beta}_2$	2.282e-03	7.479e-04	3.051	0.002279	**

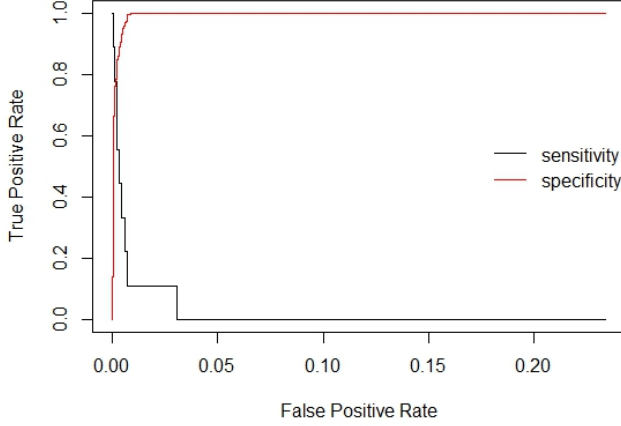


Figure 7: ROC curve for the model using  $X_R, X_D$  as predictors, for LAX Airport.

Ideally we would like to obtain the weather data for all 311 locations but due to time constraints and the manner in which the weather data for each location is downloaded, we only use the 15 high traffic airports described in section 3. Doing so, we separate these 15 airports into locations where the number of flight observations exceed 148,500 and locations where the number of flights do not exceed 148,500. The locations where traffic flights exceed 148,500 include are Hartsfield-Jackson Atlanta International Airport, Denver International Airport, Dallas/Fort Worth International Airport, McCarran International Airport in Las Vegas, Los Angeles International Airport, O'Hare International Airport, Phoenix Sky Harbor International Airport, and San Francisco International Airport. Flights that are still high traffic but do not exceed 148,500 include John F. Kennedy, Detroit Metropolitan Airport, and others.

We fit several models, with several combinations of predictors. First we notice that wind speed is not statistically significant. The model containing origin (as a factor of airport traffic), rainfall, visibility (as a factor of high and low), temperature and distance (seen in Table 7) produces the highest AUC. The odds ratio is given by

$$\frac{\hat{p}(X)}{1 - \hat{p}(X)} = e^{\hat{\beta}_0 + \hat{\beta}_1 X_{O_f} + \hat{\beta}_2 X_R + \hat{\beta}_3 X_{V_f} + \hat{\beta}_4 X_T + \hat{\beta}_5 X_D}$$

The interpretation of the estimate of the intercept  $\hat{\beta}_0$  is

the log odds of the probability of a delay occurring given a high traffic airport location and given visibility is high. One thing to note is that the estimate for the categorical variable representing "low" traffic airports is positive; in other words the odds of having a delay increases by 4% for locations in which flights do not exceed 148,500. For example, holding all other predictors constant, the odds of having a delay for John F. Kennedy International Airport is 4% higher than, say San Francisco International Airport. But again we are using only 15 of the 311 origin airports and we are splitting those 15 into high and low of the overall high traffic airport. Incorporating all 311 origin airports could change this estimate.

Furthermore, Table 7 using 60 minutes as the cutoff for flight delays, shows us that temperature is statistically significant. When logistic regression was applied to separate cities, temperature was never a statistically significant variable. For every one degree Fahrenheit increase in temperature, the odds of having a delay decrease by 1%. This seems reasonable because delays are more likely to occur in colder weather extremes. Also, adding interaction between visibility and temperature proves to be statistically significant. The interpretation of this is that for every one unit increase in temperature, the log odds for high visibility increases by 0.0059. So even though the interaction term is highly significant, the estimate is very small.

The AUC represented by the model containing the  $\beta$  estimates in Table 7 is 0.652 and the ROC curve is seen in Figure 8. This is the highest AUC we are able to achieve using location traffic as a factor; but again this may improve if we were to use all 311 airport locations.

As the Hosmer-Lemeshow Goodness of Fit test is influenced by sample size, it generally can not be applied to sample sizes larger than 25,000 (Yu et al., 2017). Therefore, we did not use this test to conclude whether this model is a good fit because the number of observations contained in the dataset using cities as a factor is 109,902.

## 6. Conclusion

Using 15 high traffic airports, we are able to produce a logistic regression model capable of predicting the probabilities of a delay occurring at a given airport. We find that the statistically significant variables are unique to each city. For Baltimore-Washington International, rainfall, distance of flight, high or low visibility and wind speed affect the probability of a delay occurring. Los Angeles International Airport, on the other hand, showed

Table 7: Estimates of the  $\beta$ s parameters using 60 minutes as the cutoff with airport traffic included as a categorical variable.

	<i>Estimate</i>	<i>Std. Error</i>	<i>t value</i>	<i>Pr(&gt;  t )</i>	
(Intercept) $\hat{\beta}_0$	-3.6379672	0.1440729	-25.251	< 2e-16	***
OriginLow $\hat{\beta}_1$	0.4303792	0.0614280	7.006	2.45e-12	***
Rainfall $\hat{\beta}_2$	5.6826286	0.5104060	11.134	< 2e-16	***
Visibility High $\hat{\beta}_3$	-2.4753090	0.1849836	-13.381	< 2e-16	***
Temperature (Fahrenheit) $\hat{\beta}_4$	-0.0284693	0.0023079	-12.336	< 2e-16	***
Distance $\hat{\beta}_5$	0.0007669	0.0000731	10.492	< 2e-16	***
Visibility High * Temperature $\hat{\beta}_6$	0.0344159	0.0031436	10.948	< 2e-16	***

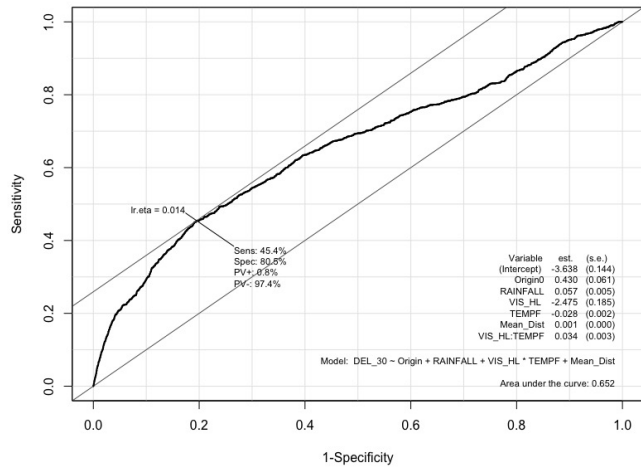


Figure 8: ROC curve for the model which has airport traffic as a categorical predictor

that visibility, rainfall and wind speed were not statistically significant. One possibility for this difference could be that Baltimore, Maryland witnesses weather extremes relating to rainfall, wind speed, and visibility while Los Angeles lacks these weather extremes. The  $\beta$  estimates for all of these variables provide reasonable and expected odds ratios. The models also meet all of the criterion for a well fitting model.

Separating the 15 high traffic airports into factors, we are also able to produce a model capable of predicting weather delays. One main difference in this approach is that temperature was statistically significant when it was not significant in any of the models of separate cities. However, we find that the criterion used to determine the accuracy of the model are not as strong as the approach of modeling separate cities. One reason for this is that we are focusing only on the 15 high traffic airports, when ideally all 311 origin airports may provide better models.

In both approaches, we found that time of day, and day of the year were not statistically significant. Something that would be interesting to add would be to consider another variable that indicates holidays and determine if this has an effect on the probability of a delay occurring.

Another factor that is not considered in any of the models is the effect of flight delays as well as the effect of weather variables at the textitdestination airport. However, the airline data is not formatted in a way that this could be calculated easily.

## References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281. Akadcmiai Kiado, Budapest, Hungary.
- Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2000). *Applied logistic regression*, volume 2nd. John Wiley & Sons.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.
- Oceanic, N. and Administration, A. (2016). National centers for environmental information. <https://www.ncdc.noaa.gov/cdo-web/datatools/lcd>.
- of Transportation, U. S. D. (2016). Airline on-time performance data. <https://www.transtats.bts.gov>.
- Socrata (2017). Airport codes mapped to latitude/longitudein the united states. <https://opendata.socrata.com/dataset/Airport-Codes-mapped-to-Latitude-Longitude-in-the-/rxrh-4cxm>.
- Yu, W., Xu, W., and Zhu, L. (2017). A modified hosmer-lemeshow test for large data sets. *Communications in Statistics - Theory and Methods*, 46(23):11813–11825.