Statistical learning based analysis and prediction of Airline On-Time Performance Data

Beverly Osborn, Md Ferdous Alam, Jiarong Zou, Shuqi Zhou, Yitong Zhou

April 28, 2019

1 Introduction

Airline travel is one of the most popular modes of transportation, and the economic costs associated with flight delays are enormous. Our research objective is to build prediction models for flight delays. We are interested in both arrival delays, which can negatively affect passengers, and departure delays, which can negatively affect operating costs for airlines (e.g., when pilots make up for late departures by accelerating more quickly and thereby increasing fuel consumption). We applied a wide variety of machine learning techniques to classify flights as delayed or on-time.

We use two main response variables: a binary response variable for whether a flight *departure* was at least 15 minutes behind schedule, and a binary response variable for whether a flight *arrival* was at least 15 minutes behind schedule. We build separate prediction models for the two response variables; we expect departure delay to be an important predictor of arrival delay. Otherwise, our prediction variables largely relate to the schedule, the origin airport, and the departure airport; variables will be discussed in more detail in later sections of this report.

An important complicating factor in this analysis is the fact that the delays we are measuring are a function of both environmental/external uncertainty, and the airline's own scheduling decisions. Unexpectedly long flights (i.e., delays) are more costly for airlines than expectedly long flights (i.e., a flight of the same duration, but for which schedule-buffer time has been allocated to absorb the delay), all else being equal. Therefore, it is rational for airlines to adjust their scheduling decisions such that predictable delays (e.g., caused by morning fog at some airports) do not occur. This makes our question a very difficult one. We will only be able to detect a pattern of delays against schedule when an airline: **(1)** is not behaving economically rationally **(2)** is facing especially high costs of schedule buffering **(3)** is facing especially low costs of unexpected delays **(4)** was not previously aware of the pattern of delays.

This report has been constructed in the following way: introduction of the data and sampling approach followed by description of our findings from exploratory data analysis, and then addressing the problem of dimensionality, performing analysis using support vector machines (SVM), logistic regression, and random forest, ans concluding remarks on the findings.

2 Data and Sampling

We use publicly available data on US domestic flights, published by the **?**, to develop and test our prediction models. This database currently contains more than 100 million flight-level observations. All airlines that carry a substantial portion of total US passenger traffic are required (by Federal Aviation Authority regulation) to report their flight-level data, and the data has been published on a monthly basis since 1987.

We randomly sampled 10,000 flights for the 60-month period of 2012 to 2017. This period of time allows us to observe seasonality and trends, while also providing a sufficient number of observations of infrequent events. After taking the random sample, we removed observations for cancelled flights (9233) and diverted flights (1378), resulting in a final data set of 589,389 observations. We then removed all variables related to diverted and cancelled flights, as well as variables redundant with other variables (e.g., the airline's unique ID and the airline's operating name). All random samples mentioned in subsequent sections of this report are randomly sampled from this parent-sample.

3 Exploratory Data Analysis (EDA)

Our exploratory data analysis focuses on three main areas: the influence of airlines, time, and geography.

3.1 Airline Analysis

Analysis of the 16 airlines included in the sample indicates a strong relationship between airlines and flight delays. Some airlines have much more total delay time than others. Most flights are on-time, however, a few flights have very long delays. Departure delays are typically greater than arrival delays; this is an expected result of airlines' ability to trade-off between flight speed and fuel consumption. A flight with a departure delay can sometimes increase its speed in order to arrive on-time. Figure 1 shows which airlines are responsible for most flights in the observation period (left pie chart). Note that Southwest Airlines is the largest carrier by this measure; it operated 20% of flights. The figure also shows the mean departure delay, in minutes, for each airline (right pie chart). Note that there is substantial variation in departure delays between carriers. Finally, we can also observe the long tail of flight delay durations, by carrier (bottom chart), and note that there is a small number of delays that exceed 20 hours. Additional figures on the relationship between airlines and flight delays can be found in the Appendix.

3.2 Time Related Analysis

We also observe relationships between month and year, and flight delays. As expected (recall our decision to sample from a five year period), we observe both seasonality and trend in flight delays. There are more delays in June, July, August and December (Fig.2). The number of delays is increasing from year 2012 to 2014 and decreasing after 2014.

Next, we consider the role of scheduled departure time and scheduled arrival time. Fig.3 shows that flights departing from 12pm to 10pm are delayed more often than flights departing at other time. Meanwhile, flights scheduled to arrive from 6am to 12am are delayed less often than flights scheduled to arrive at any other time.

Finally, we consider the relationship between airtime and delays. While Fig.4 does not indicate a linear relationship between distance (which is closely related to airtime) and delays, we can observe that flights with very long airtime rarely experience long arrival delays.

3.3 Map Analysis

In order to visualize the delay across the country, we plotted maps according to the number of flights, arrival delay, departure delay, and percentage of arrival delay for each state as a destination. Since the map we used does not contain the following 5 regions including Alaska, Hawaii, Puerto Rico, U.S. Pacific trust territories and possessions, and U.S. Virgin Islands, we excluded the 5 regions

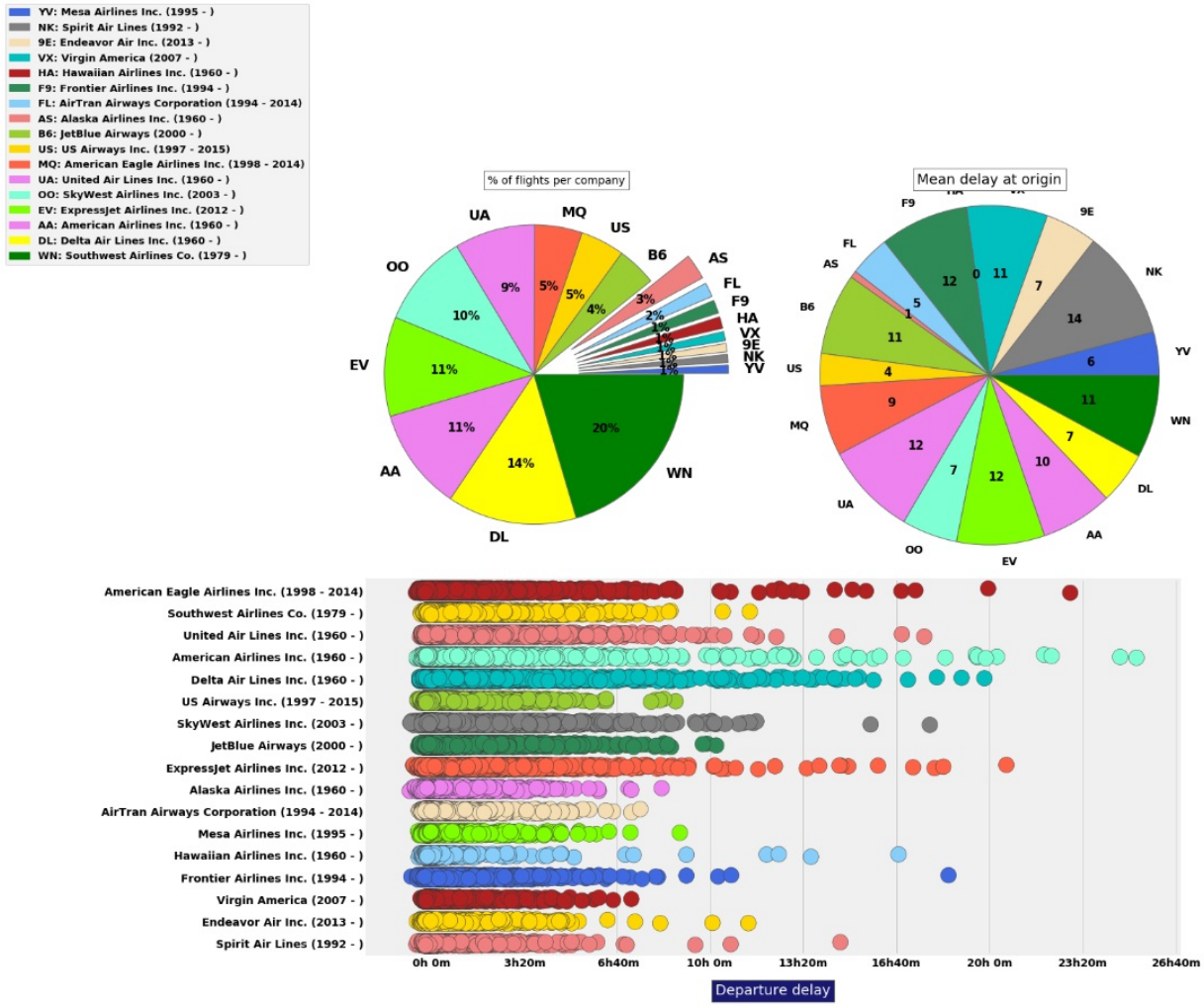


Figure 1: Delay performance by airline.

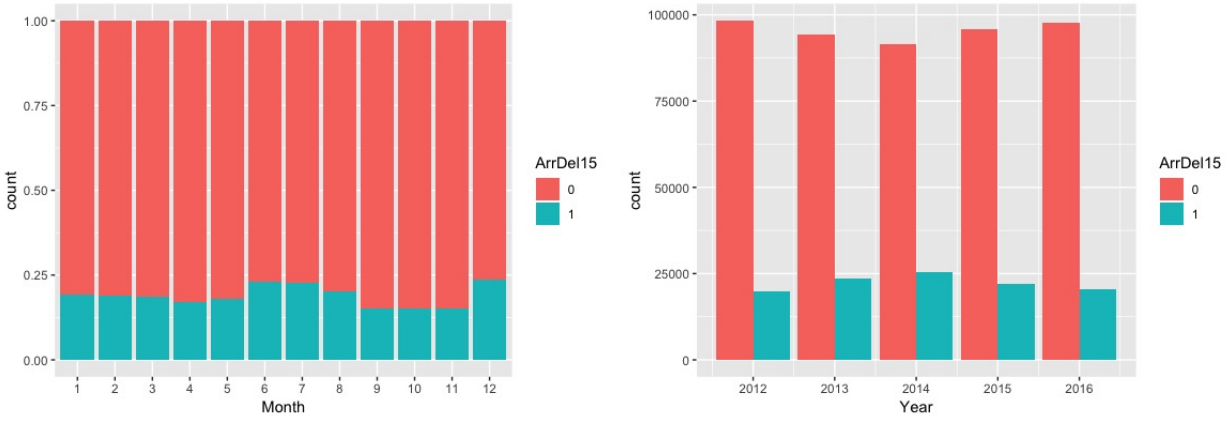


Figure 2: Flight Date.

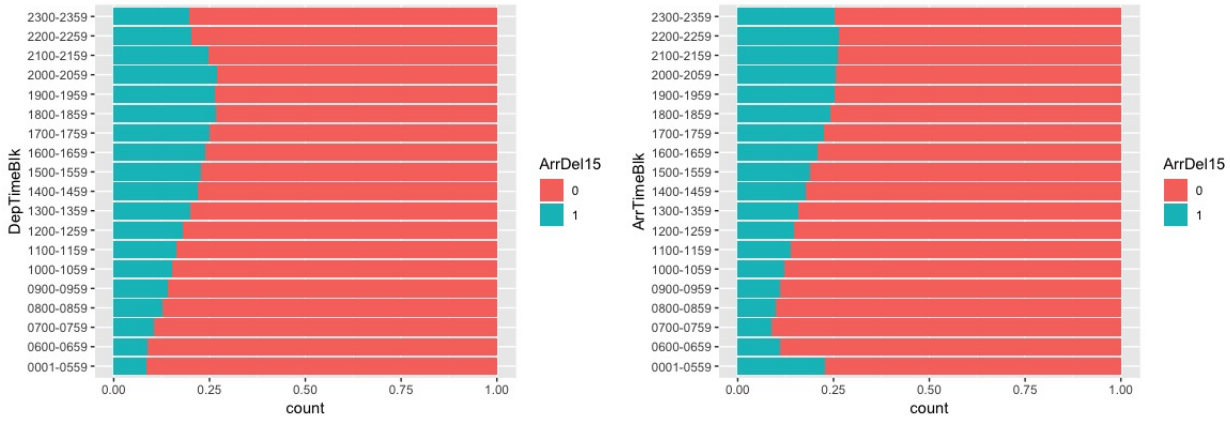


Figure 3: Departure Time and Arrival Time.

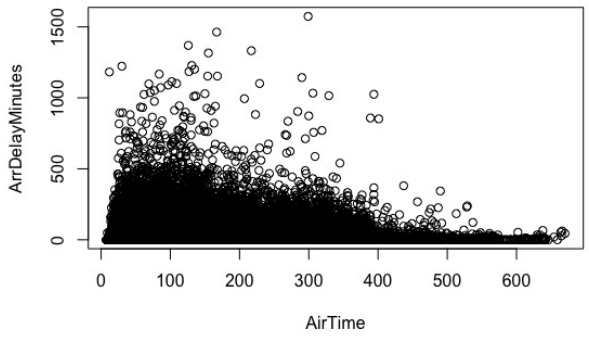


Figure 4: Flight Duration and Arrival Delay Time.

from the map. The first three maps in the appendix all show similar patterns: California, Texas, Florida, Illinois, New York, and Georgia are destinations with high counts of departure delays, high counts of arrival delays, and high total flights. This indicates that the busier the destination, the more delays happen.

The below figure shows the proportion of delayed arrivals for each state as a destination. The states with highest delay percentage are Delaware, New Jersey, Vermont, Oklahoma, and New York with delay percentages of 0.42, 0.26, 0.24, 0.23, and 0.23, respectively. This means that someone want to visit the above states should be prepared to arrive late.

Although Delaware has the lowest counts of arrival delay, due to its relative small total counts of flights, it also has the highest arrival delay rate. The states with lowest delay percentages are Utah, Montana, Minnesota, Georgia, and Michigan with values of 0.12, 0.13, 0.15, 0.15, and 0.15, respectively. Generally, arrival delay rate is higher in northeast and mid-west areas.

Similar maps showing the departure delay rates for different destination states are provided in the Appendix. Considering all of these maps as a group, we can observe a strong association between departure delays and arrival delays. Delaware, New Jersey, and Vermont are top states for departure and arrival delay no matter whether one flies in or out. We have also observed that delays are highly associated with locations. In the next stages of analysis, we expect locations to

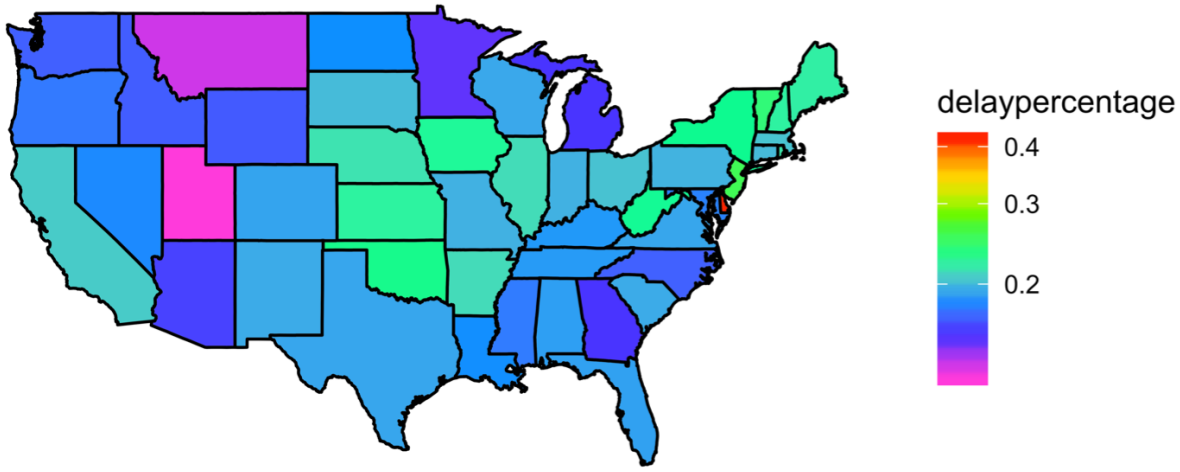


Figure 5: Arrival delay for destination states.

be important predictors of flight delays.

From the EDA, we conclude that airlines, months, years, departure and arrival times, flight durations, and geographic locations all have important associations with our response variables of interest.

4 Variable Selection and Dimensionality

Our EDA identified airlines and locations as important categorical predictors of flight delays. However, these categorical variables have many factor levels. If we conduct our analysis at the state-level of locations, we potentially face 2,500 combinations of origins and destinations. If we conduct analysis at the airport level instead, it gets even worse: 336 airports appear in our sampled data. Therefore, we conduct our analysis in two stages:

- Dimension Reduction
- Predictive Modelling

The outputs of our dimension reduction techniques will be inputs for some of the predictive models in the next stage of our analysis. First, we identify the variables that require dimension reduction by dividing the original variables into 4 types: **high dimensional variables to include, low dimensional variables to include, variables to exclude, and prediction variables**.

The high dimensional variables requiring dimension reduction are flight number, airline, origin airport, and destination airport. Low dimensional variables (including continuous variables) are CRS elapsed time (i.e., the scheduled duration of the flight), month, arr/dep time blk (i.e., the 'block,' or time interval, in which a flight is scheduled to arrrive or depart), and year. Our prediction variables, of course, are departure delay and arrival delay. We exclude all other variables, which either had little correlation with the prediction variables (based on the EDA), or were less granular measures of included variables (e.g., airports are nested within states). In the next subsection, we present one dimension reduction method. A second method (Multiple Correspondence Analysis) is described in the appendix.

4.1 K-Means Clustering

We apply k-means clustering to the high-dimensional variables variables identified in the previous section. By reorganizing our flight-level data, we are able to identify several potentially useful features airports, carriers, and flight numbers[1].

We cluster airports on the following measures:

- Flights: the sum of flights from that airport; this is a measure of the traffic, or volume, served by the location.
- Routes: the count of unique destinations served from that airport; this is a measure of the connectivity of the airport within the broader network of US airspace.
- Carriers: the count of unique carriers operating flights out of that airport; this is a measure of the competition for space and resources in the airport.
- DepDels: the sum of departing flights delayed by at least 15 minutes from that airport; this is a measure of the relationship between our response variable and the airport.
- ArrDels: the sum of departing flights delayed by at least 15 minutes from that airport; this is a measure of the relationship between our response variable and the airport.
- DepTimeBlks: the count of unique departure time blocks in which the airport operates; this is a measure of how long the airport is open each day.
- FlightNum: the count of unique flight numbers departing from that airport; this is a measure of the complexity of the airport's operations.

We cluster airlines similarly, but with the sums of flights operated by that airline, the count of unique routes served by that airline, etc. The Carriers variable is dropped, of course, and two others are added:

- Origins: the sum of unique airports from which flights operated by that airline depart.
- TailNum: the sum of unique aircraft (identified by tail number) operated by that airline.

We cluster flight numbers similarly, but with the sums of flights departing with that flight number, the count of unique Origin and Destination combinations included on that flight number, etc. Comparing this clustering approach to the one used for airports, the FlightNum variable is dropped, and the TailNum variable is included (i.e., as a measure of how many aircraft operate under that flight number).

We chose these variables for clustering the airports[2], airlines, and flight numbers for several reasons. For many variables, there is a strong theoretical association between the variables and the probability of delay. The availability of data also informed these choices; all variables were easily calculated from our flight-level data. After some consideration, we chose to also include variables

[1]For clarity, note that flight numbers can be multi-city route-carrier combinations, and the information contained in a flight number therefore has substantial, but imperfect, overlap with the information contained in a combination of origin, destination, and carrier.

[2]The k-means clustering of airports was conducted based on the origin airports in our sample. For consistency, we use the same clusters for both origin and destination airports. E.g., if JFK belongs to Cluster 1 when JFK is the origin airport, then JFK also belongs to Cluster 1 when JFK is the destination airport. This results in our number of sample observations being reduced by one, since one of the sampled observations had a destination airport which had not assigned to a cluster.

directly related to our main response variables (i.e., DepDels and ArrDels) as input variables for our cluster analysis.

Ultimately, the aim of this cluster analysis is to create clusters with predictive value for future delays. The theoretically optimal clustering approach in an unchanging system would be to use only the proportions of delayed departures and arrivals. However, the real-world system does change over time, and we therefore concluded that it would be more interesting to cluster on a combination of theory-driven and data-driven variables.

Our variables for clustering have different scales, suggesting that rescaling them (to have a mean zero and a standard deviation of one) may be appropriate. However, these clusters are intended for input into the next stage of our analysis, and enabling standalone-interpretation is not our main objective. We therefore try both approaches.

We chose the number of clusters for each of airports, airlines, and flight numbers by using scree plots. We find that 4 clusters perform well for airports and airlines, while 6 clusters performs better for flight numbers. In each scree plot, it is apparent that no significant additional explanatory power can be gained from increasing the number of clusters.

The clustering stage of analysis results in the following reduction in dimensionality:

- Airports: the original set of 336 locations is reduced to 4 clusters; since these clusters are used for both origins and destinations, this reduces the number of origin-destination combinations[3] from 5206 to 16.
- Airlines: the original set of 17 airlines is reduced to 4 clusters.
- Flight Numbers: the original set of 6809 flight numbers is reduced to 6 clusters.

Although this approach to reducing the dimensionality of our data is statistically imperfect in some ways (e.g., it uses informed, but not impartial, judgment for variable selection), it successfully reduced the dimensionality of our data, and yielded categorical predictors that can be used in the second stage of our analysis.

5 Predictive Modelling

Our main statistical analysis uses support vector machines (SVM), logistic regression, and random forest algorithms to classify flights as delayed or on-time. We find interesting differences between the prediction models for departure and arrival delays, as well as interesting differences between prediction models constructed with different statistical methods.

5.1 Support Vector Machines

In this section, we describe the performance of support vector machine (SVM) models with radial, linear, and polynomial kernels. Based on the EDA and dimension reduction stage, we include the following predictors in our SVM model:

- Flight: CarrierCluster, FlightNumCluster
- Location: OriginCluster, DestCluster
- Time: Month, Year, ArrTimeBlk, DepTimeBlk

[3]Note that the real-world connectivity is much lower than the 336 x 335 = 112,560 that would be observed if every US airport connected to every other US airport.

- Distance: AirTime

Rather than using the full random sample, we randomly select a subsample of 1000 observations and split it into a training (701 obs) and testing (299 obs) set. This sample is used for all SVM algorithms, with two response variables: a) Departure delay b) Arrival delay.

5.1.1 SVM with radial kernel

We define the following tuning parameter list: **a)** cost list $=10^{seq(-1,1,length.out=20)}$ **b)** gamma list $=10^{seq(-1,1,length.out=20)}$

After tuning both cost and gamma parameter, the best SVM model with radial kernel results in 700 support vectors in which 559 support vectors from non-delayed flight and 141 from delayed flight. The following is the confusion matrix for prediction of testing set:

		Reference	
		Positive	Negative
Prediction	Positive	238	60
	Negative	1	0

Here, we achieve accuracy of 79.6% and P-value of 0.5911. The no information rate is 0.7993. Note that the model did not successfully identify any delayed flights, and yielded accuracy worse than the no information rate.

5.1.2 SVM with linear kernel

Using the same cost list for tuning parameter, the best SVM model with linear kernel results in 361 support vectors in which 220 support vectors from non-delayed flight and 141 from delayed flight.

		Reference	
		Positive	Negative
Prediction	Positive	239	60
	Negative	0	0

Here, we achieve accuracy of 79.93% and P-value of 0.5345. The no information rate is 0.7993. Again, the model did not successfully identify any delayed flights, and the accuracy is the same as the no information rate (i.e., the model predicted on-time flights in all cases).

5.1.3 SVM with polynomial kernel

We tune the cost parameter for different degrees of polynomial and compare the accuracy of best model indexed by degree. For degree = 2, the best SVM model with polynomial kernel results in 379 support vectors in which 238 support vectors from non-delayed flight and 141 from delayed flight.

		Reference	
		Positive	Negative
Prediction	Positive	237	59
	Negative	2	1

Here, we achieve accuracy of 79.6% and P-value of 0.5911. The no information rate is 0.7993, so our model underperforms in terms of accuracy. However, this model did successfully identify one delayed flight.

For degree = 3, the best SVM model with polynomial kernel results in 483 support vectors in which 344 support vectors from non-delayed flight and 139 from delayed flight.

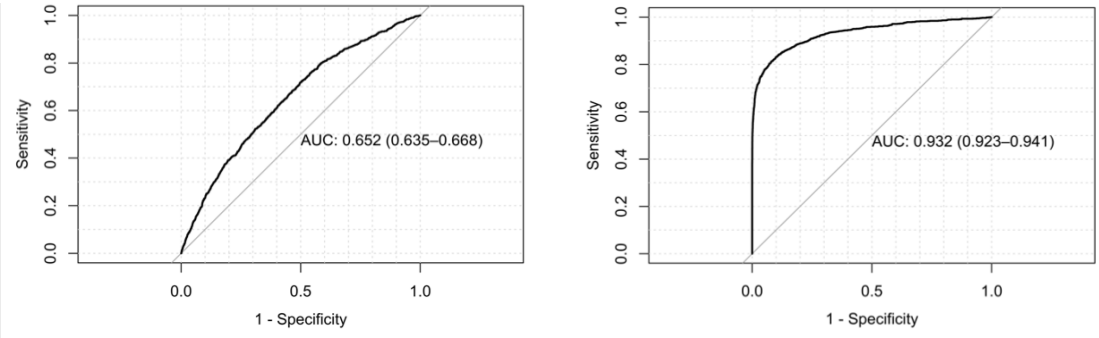
		Reference	
		Positive	Negative
Prediction	Positive	207	48
	Negative	32	12

Here, we achieve accuracy of 73.24% and P-value of 0.99791. The no information rate is 0.7993, so this model also underperforms in terms of accuracy. However, it was able to successfully identify 12 delayed flights. Recalling that airlines commonly use schedule buffering to adjust for known delay patterns, and that losses associated with false positives and false negatives are generally not equal in this context, we consider this an interesting and potentially useful prediction model.

All results reported above concern departure delays. For brevity, we do not present our models for arrival delays. For arrival delays, the no information rate is 0.82, and our prediction models generally achieve accuracy of 82% (by predicting all flights on-time) when departure delays are not included as predictors.

5.2 Logistic regression for arrival delay

We perform logistic regression on arrival delay using the same predictors described in the previous section, both with and without including departure delay minutes as a predictor. ArrDel15 (arrival delay for equal to or more than 15 min) is our response variable. We use 10,000 randomly selected observations and a training/test ratio 7/3.



(a) (b)

Figure 6: (a) ROC for arrival delay without using departure delay . (b) ROC for arrival delay with using departure delay.

We plot ROC curves for using and not using departure delay as a predictor in Figure 6. We observe that without using departure delay as a predictor, the AUC is only 0.65, meaning the logistic regression model can barely detect an arrival delay. However, when departure delay is included as a predictor of arrival delay, as shown in Figure 6 (b), the AUC is 0.93, indicating a good fit. Since we want higher sensitivity to detect arrival delay, we choose sensitivity as 0.8, which results in a threshold of 0.19 and specificity of 0.93.

Next, we use 0.19 as the threshold for test data set, resulting in a confusion matrix as below. Sensitivity, specificity, and accuracy are found to be 0.8, 0.92, and 0.89 respectively. This result is highly promising in arrival delay detection.

		Reference	
		Positive	Negative
Prediction	Positive	2205	117
	Negative	203	475

We conclude that most variables are poorly associated with arrival delay, indicating arrival delay is a complex phenomena. In addition, departure delay is a pivotal variable for predicting arrival delay, meaning that we are only able to consistently detect arrival delay after a flight takes off.

5.3 Random Forest:

Finally, we also apply a random forest algorithm to our classification problem for departure delays. The confusion matrix below relates to a randomly selected training set of 700 observations, with a tree number of 500 and using 3 variables at each splitting node. Here, we achieved an out-of-bag (OOB) error rate of 20.86% and correctly identified 2 of 146 delayed departures.

		Reference	
		Positive	Negative
Prediction	Positive	554	144
	Negative	0	2

The confusion matrix below is for a random forest using 6 variables per node, this time for a smaller random sample of 300 observations. Here, we achieved a slightly better OOB error rate of 20%, and correctly identified 3 of 57 delayed departures. Consistently with the earlier models results, these results support our conclusion that flight delays are a complex phenomenon that is difficult to predict (partially due to the role of airlines' scheduling practices).

		Reference	
		Positive	Negative
Prediction	Positive	233	54
	Negative	10	3

6 Discussion and Conclusions

Our research objective was to predict flight delays. As highlighted in the introduction, this is a harder problem than is apparent at first glance. Airlines' scheduling practices already account for most easily predictable delays. Furthermore, the high dimensionality of our key variables in our dataset complicated our analysis. Despite the difficulty of our problem, we were able to create prediction models that correctly identified up to 20 percent of departure delays. Furthermore, by including departure delays in our prediction model for arrival delays, we were able to correctly predict 80 percent of arrival delays.

A key finding from our analysis is that departure delays are a very strong predictor of arrival delays, and this relationship has not yet been exploited in airlines' scheduling and operating decisions. Although this relationship is not surprising, our findings provide useful direction for airlines interested in further improving their on-time performance. Airlines can best reduce unplanned arrival delays by implementing in-air schedule adjustments, automated notifications to subscribed individuals, and improving their abilities to cost-effectively adjust airtime (i.e., by adjusting flight speed or route) to compensate for departure delays. These techniques can lead to a better customer experience.