Predicting flight delays

Objective of this project:

Predicting flight delays with sufficient lead time. Imagine I am working for the Department of Transportation in US. If accurate estimations of flight delays can be predicted once we received the flight schedule of each airlines, we can advise them to alert their flight schedules. This can reduce to average flight delays and thus reducing the cost caused by flight delays.

1. Introduction

The NEXTOR study by the University of California, Berkeley's Institute of Transportation Studies has revealed that in 2007, flight delays across the US airspace system cost the US economy USD32.9 billion. While the year was the worst on record in terms of flight delays – with 24% of all domestic airline flights more than 15 minutes late and 2% cancelled – it is also indicative of the cost of flight delays not just for airlines, but for the broader US economy.

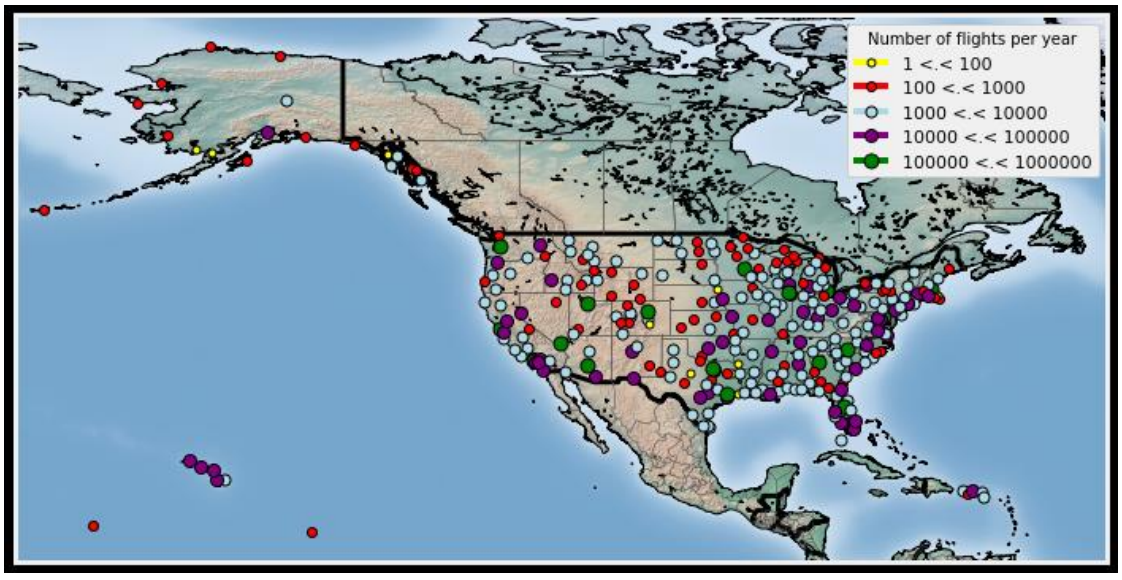
This project aims to create an accurate model of predicting flight delays with sufficient lead time and understand the most influential factor on flight on-time performance. Thus, the cost caused by flight delays can be reduced by proper and timely responses.

2. Data Description

Among the provided datasets, the table "flight_traffic.csv" is the major one to use in this project. One of the crucial challenges of using this table is that the datetime of fields "actual_departure" and "actual_arrival" are not given, making the happening time of these events ambiguous. For example, when actual_departure equals to 0030 and scheduled_departure equals to 2330, we cannot decide whether that flight has taken off 1 hour earlier or 23 hours later compared to scheduled time.

Therefore, the original table, which "departure_delay" and "arrival_delay" are available, is downloaded from the website of US Department of Transportation. Moreover, tables of flight traffic of 2015, 2016, 2017 are downloaded to facilitate the modeling process which requires more data.

3. Exploratory Data Analysis (EDA)



All the airports location covered, and the number of flights recorded during the year of 2017.

3.1 Missing Values

	column	missing values	missing percentage (%)
0	airline_id	0	0.000000
1	origin_airport	0	0.000000
2	destination_airport	0	0.000000
3	scheduled_departure	0	0.000000
4	scheduled_arrival	0	0.000000
5	scheduled_elapsed	7	0.000123
6	actual_departure	80308	1.415213
7	departure_delay	80343	1.415830
8	actual_arrival	84674	1.492153
9	arrival_delay	95211	1.677839
10	actual_elapsed	95211	1.677839

After filtering the data with only interested columns, we can see that less than 2% of data have missing values. And all the 95211 entries with missing data are due to cancellations or diversions, which mostly caused by unpredictable factors like extreme weathers.

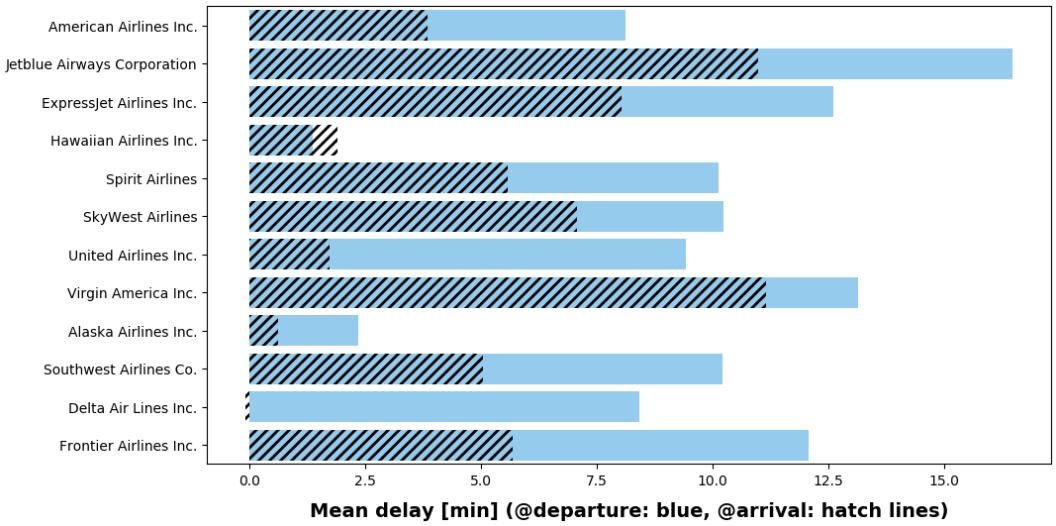
	date	counts
0	2017-03-14	4036
1	2017-09-11	3286
2	2017-02-09	2372
3	2017-09-10	2179
4	2017-04-06	1937

airline_id	min	max	count	mean
Virgin America Inc.	-91.0	712.0	69682.0	11.149536
Hawaiian Airlines Inc.	-78.0	1315.0	79861.0	1.905736
Frontier Airlines Inc.	-76.0	995.0	101917.0	5.676904
Spirit Airlines	-74.0	1619.0	151468.0	5.582539
Alaska Airlines Inc.	-89.0	861.0	183056.0	0.630687
Jetblue Airways Corporation	-98.0	1094.0	289627.0	10.977184
ExpressJet Airlines Inc.	-66.0	1810.0	328574.0	8.027348
United Airlines Inc.	-82.0	1539.0	577212.0	1.737166
SkyWest Airlines	-88.0	1792.0	694014.0	7.082484
American Airlines Inc.	-84.0	2189.0	882218.0	3.850768
Delta Air Lines Inc.	-238.0	1240.0	912792.0	-0.074738
Southwest Airlines Co.	-151.0	760.0	1308989.0	5.046553

Top 5 dates with cancelled/diverted flights. //// Descriptions of punctuality of airlines (“departure_delay”)

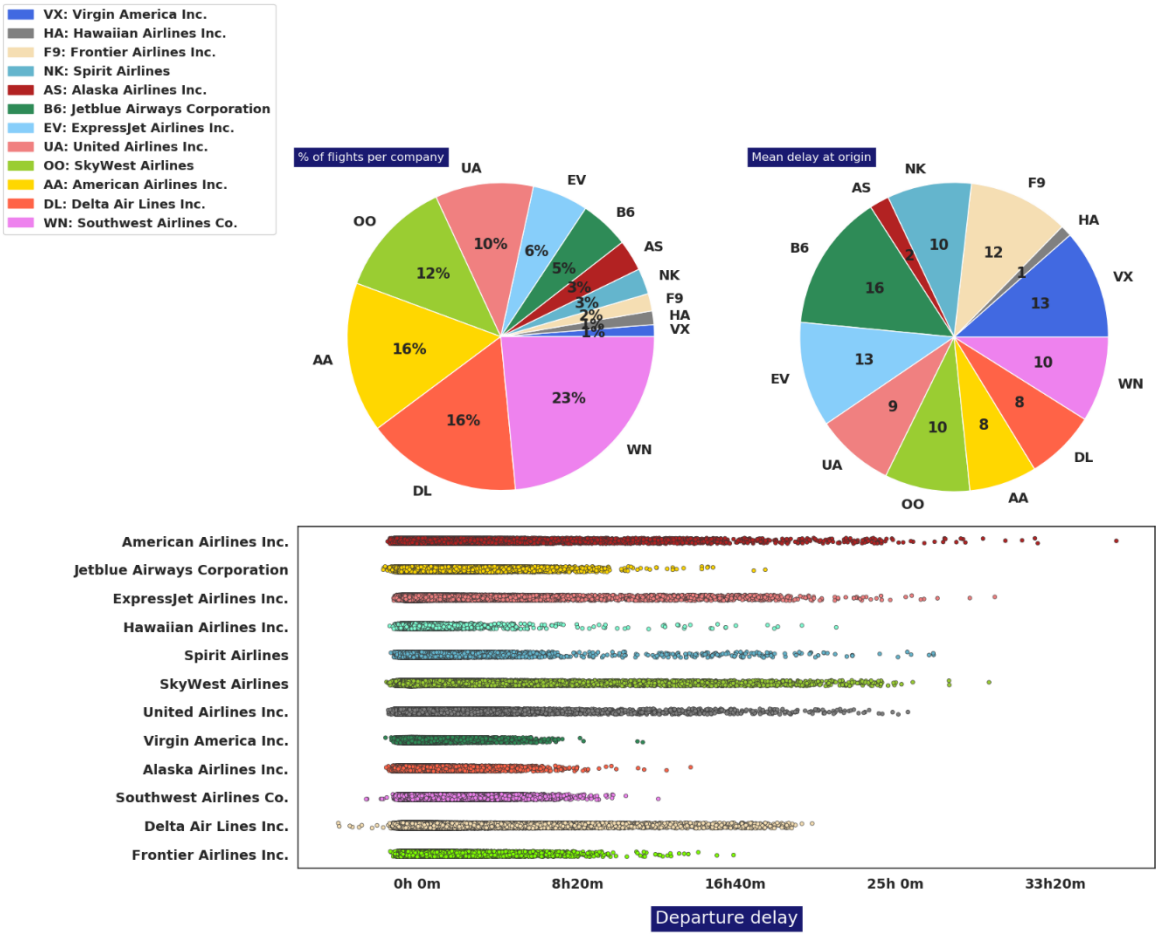
All these flights are affected by extreme weather which is not predictable with sufficient lead time (for example: there were snowstorms on 2017-03-14 and Hurricane Irma on 2017-09-11). I would like to exclude these events based on the hypothesis so that I simply remove the entries with missing values.

3.2 Comparing departure delays and arrival delays



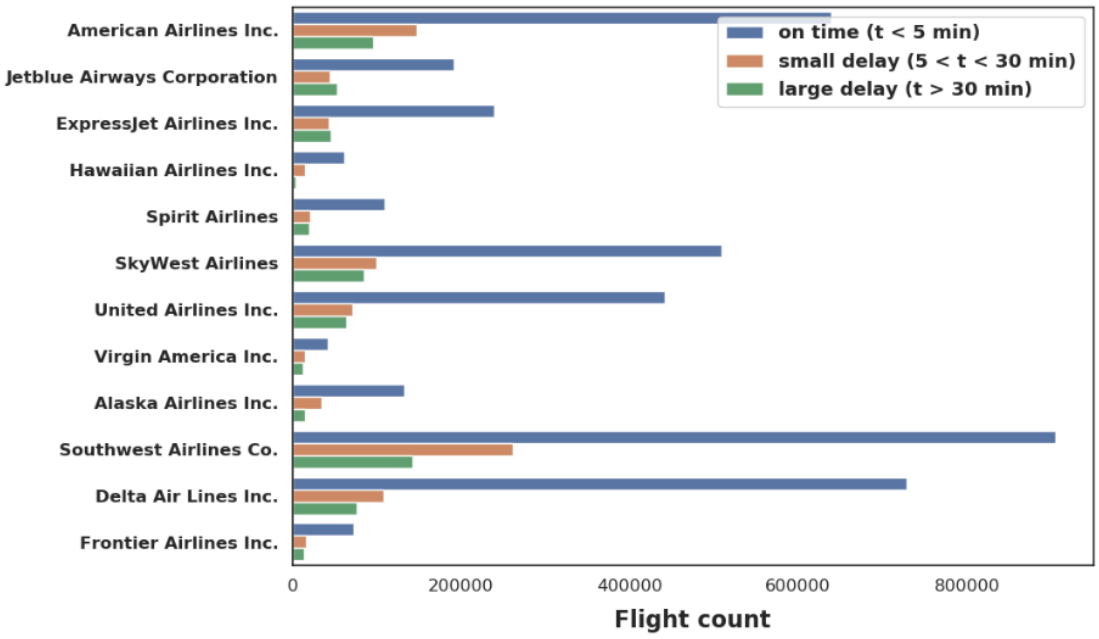
On this figure, we can see that delays at arrival are generally lower than at departure. This indicates that airlines care more about reducing delays at arrival and they may adjust their flight speed in order to reduce the delays at arrival. In what follows, only delays at arrivals would be discussed because the public usually care whether would they arrive their destinations on time more.

3.3 Punctuality of airlines



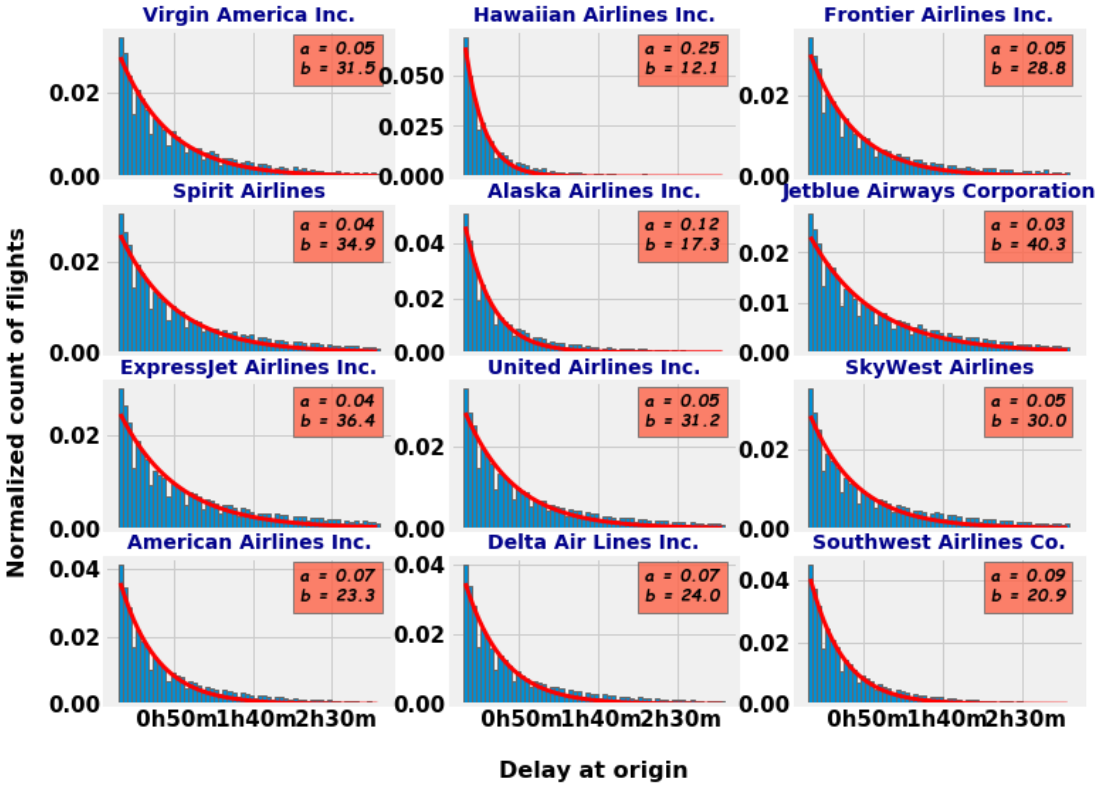
Considering the first pie chart that gives the percentage of flights per airline, we see that there is some disparity between the carriers. For example, Southwest Airlines accounts for 23% of the flights which is similar to the number of flights chartered by the 7 tiniest airlines. However, if we have a look at the second pie chart, we see that here, on the contrary, the differences among airlines are less pronounced. Excluding Hawaiian Airlines and Alaska Airlines that report extremely low mean delays, we obtain that a value of **∼12±4** minutes would correctly represent all mean delays. Note that this value is quite low which mean that the standard for every airline is to respect the schedule!

Finally, the figure at the bottom makes a census of all the delays that were measured in 2017. This representation gives a feeling on the dispersion of data and put in perspective the relative homogeneity that appeared in the second pie chart. Indeed, we see that while all mean delays are around 10 minutes, this low value is a consequence of the fact that most flights take off on time. However, we see that occasionally, we can face really large delays that can reach a few tens of hours!

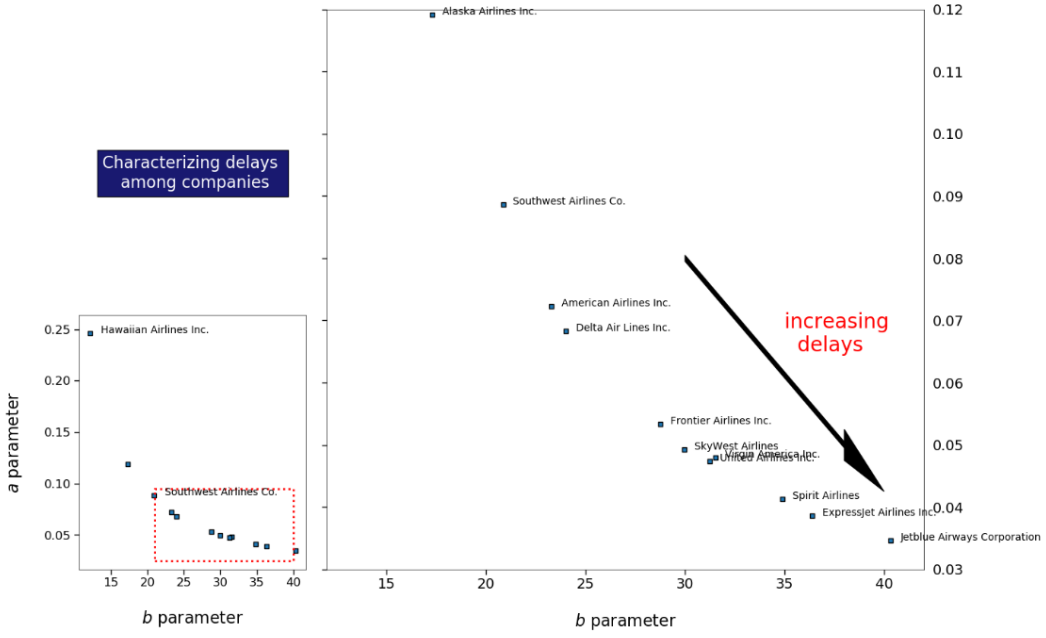


This figure gives a count of the delays of less than 5 minutes, those in the range $5 < t < 30$ min and finally, the delays greater than 30 minutes. Hence, we are aware that independently of the airline, delays greater than 30 minutes only account for a few percentages. However, the proportion of delays in these three groups depends on the airline: as an example, in the case of *SkyWest Airlines*, the delays greater than 30 minutes are only lower by ∼5% with respect to delays in the range $5 < t < 30$ min. Things are better for *SouthWest Airlines* since delays greater than 30 minutes are 2 times less frequent than delays in the range $5 < t < 30$ min.

3.3 Ranking of airlines



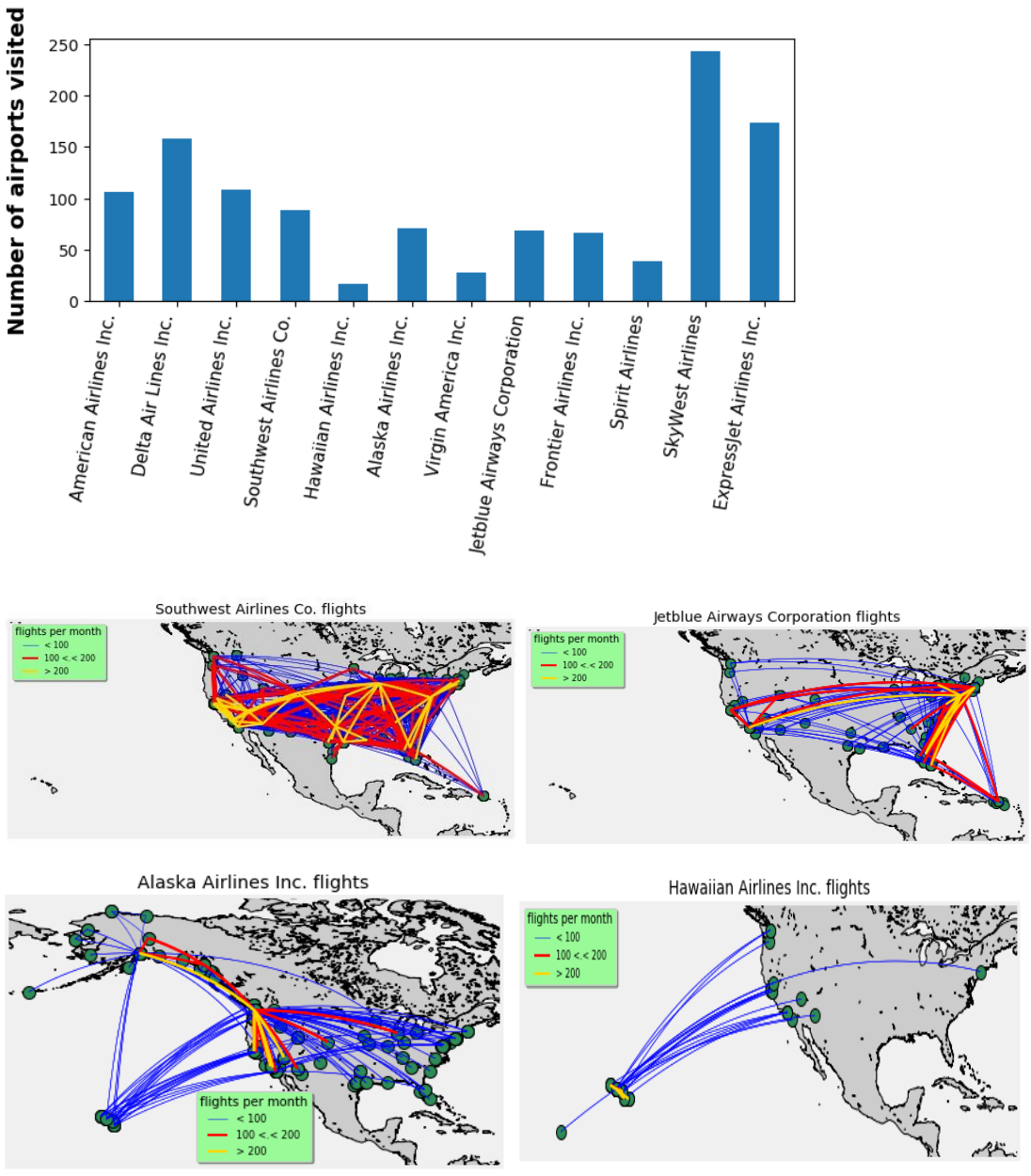
This figure shows the normalized distribution of delays that modelized with an exponential distribution $f(x)=a\exp(-x/b)$. The a and b parameters obtained to describe each airline are given in the upper right corner of each panel. Note that the normalization of the distribution implies that $\int f(x)dx \sim 1$. According to the value of either a or b, it is possible to establish a ranking of the companies: the low values of a will correspond to airlines with a large proportion of important delays and, on the contrary, airlines that shine from their punctuality will admit high a values:



The left panel of this figure gives an overview of the a

and b coefficients of the 14 airlines showing that *Hawaiian Airlines* and *Delta Airlines* occupy the first two places. The right panel zooms on 12 other airlines. We can see that *SouthWest Airlines*, which represent 23% of the total number of flights is well ranked and occupy the third position. According to this ranking, *JetBlue Airlines* is the worst carrier.

3.4 Geographical area covered by airlines



<Flights routes of indiviual airline during december 2017>

From the above graphs, we can observe that different airlines have their own focused market and business model. For *Alaska Airlines* and *Hawaiian Airlines*, they are obviously specialized airlines, which means that they have few flight routes and these

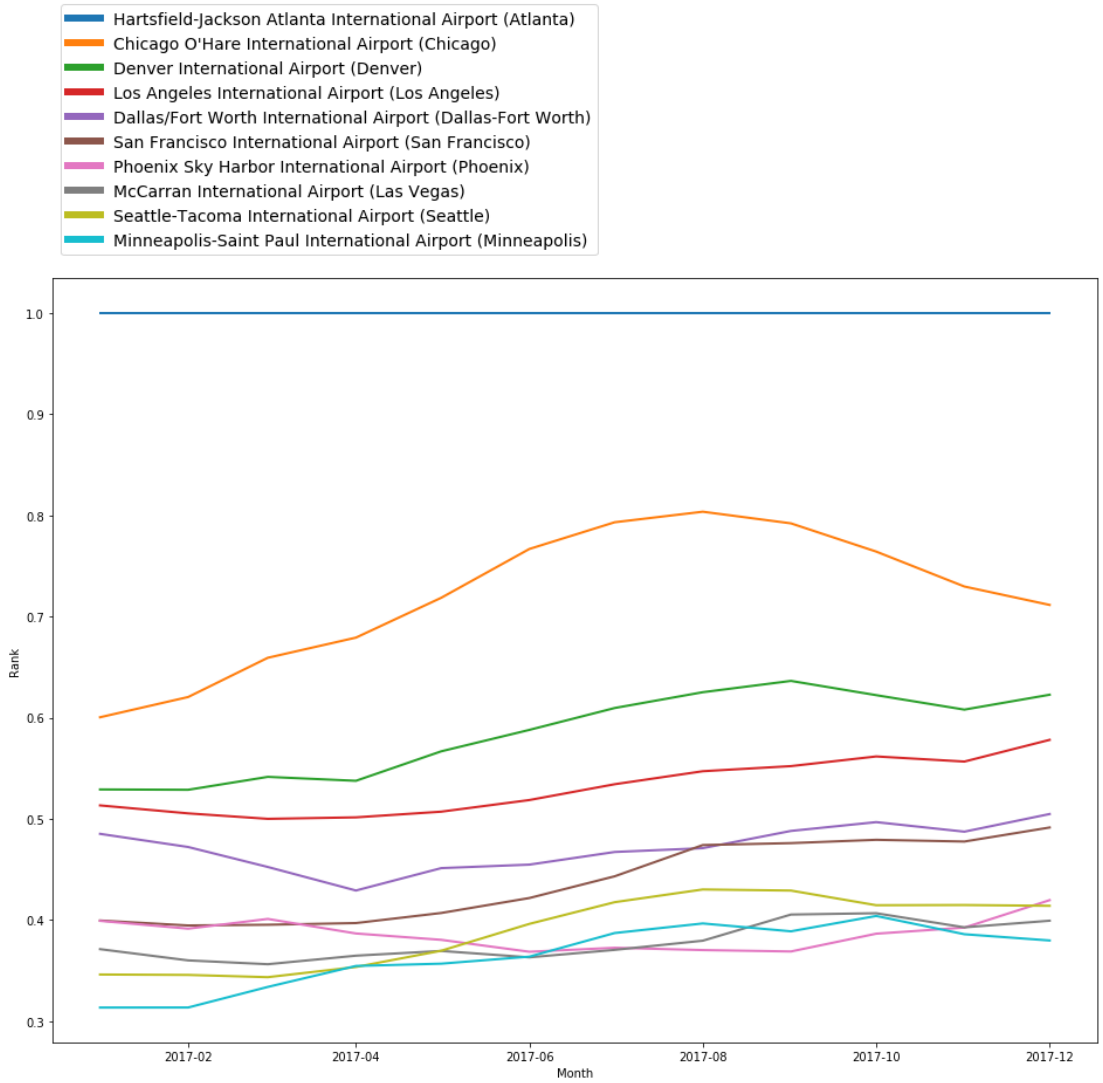
flight routes are not overlapping with other airlines. This could explain why they often have fewer delays. Comparing *SouthWest Airlines* and *JetBlue Airways*, almost all of the JetBlue's flight routes overlap with those of SouthWest's and there are fewer of them. However, the delay rate of JetBlue is much higher. So, flight routes should be one of the factors affecting delay rates, but not the only one.

3.5 PageRank of airports

	id	name	city	pagerank
0	ATL	Hartsfield-Jackson Atlanta International Airport	Atlanta	20.679329
1	ORD	Chicago O'Hare International Airport	Chicago	15.094673
2	DEN	Denver International Airport	Denver	12.607411
3	LAX	Los Angeles International Airport	Los Angeles	11.973382
4	DFW	Dallas/Fort Worth International Airport	Dallas-Fort Worth	10.324691
5	SFO	San Francisco International Airport	San Francisco	9.767376
6	PHX	Phoenix Sky Harbor International Airport	Phoenix	8.619960
7	LAS	McCarran International Airport	Las Vegas	8.465340
8	SEA	Seattle-Tacoma International Airport	Seattle	7.801273
9	MSP	Minneapolis-Saint Paul International Airport	Minneapolis	7.679868
10	MCO	Orlando International Airport	Orlando	7.390900
11	IAH	George Bush Intercontinental Airport	Houston	7.324005
12	DTW	Detroit Metropolitan Airport	Detroit	7.218357
13	BOS	Gen. Edward Lawrence Logan International Airport	Boston	6.995195
14	EWR	Newark Liberty International Airport	Newark	6.492210
15	SLC	Salt Lake City International Airport	Salt Lake City	6.398184
16	CLT	Charlotte Douglas International Airport	Charlotte	6.313920
17	BWI	Baltimore-Washington International Airport	Baltimore	5.639308
18	JFK	John F. Kennedy International Airportξ(New Yor...	New York	5.298627
19	LGA	LaGuardia Airport (Marine Air Terminal)	New York	5.218091

<PageRank of top 20 busiest airports in 2017, higher pagerank implies busier>

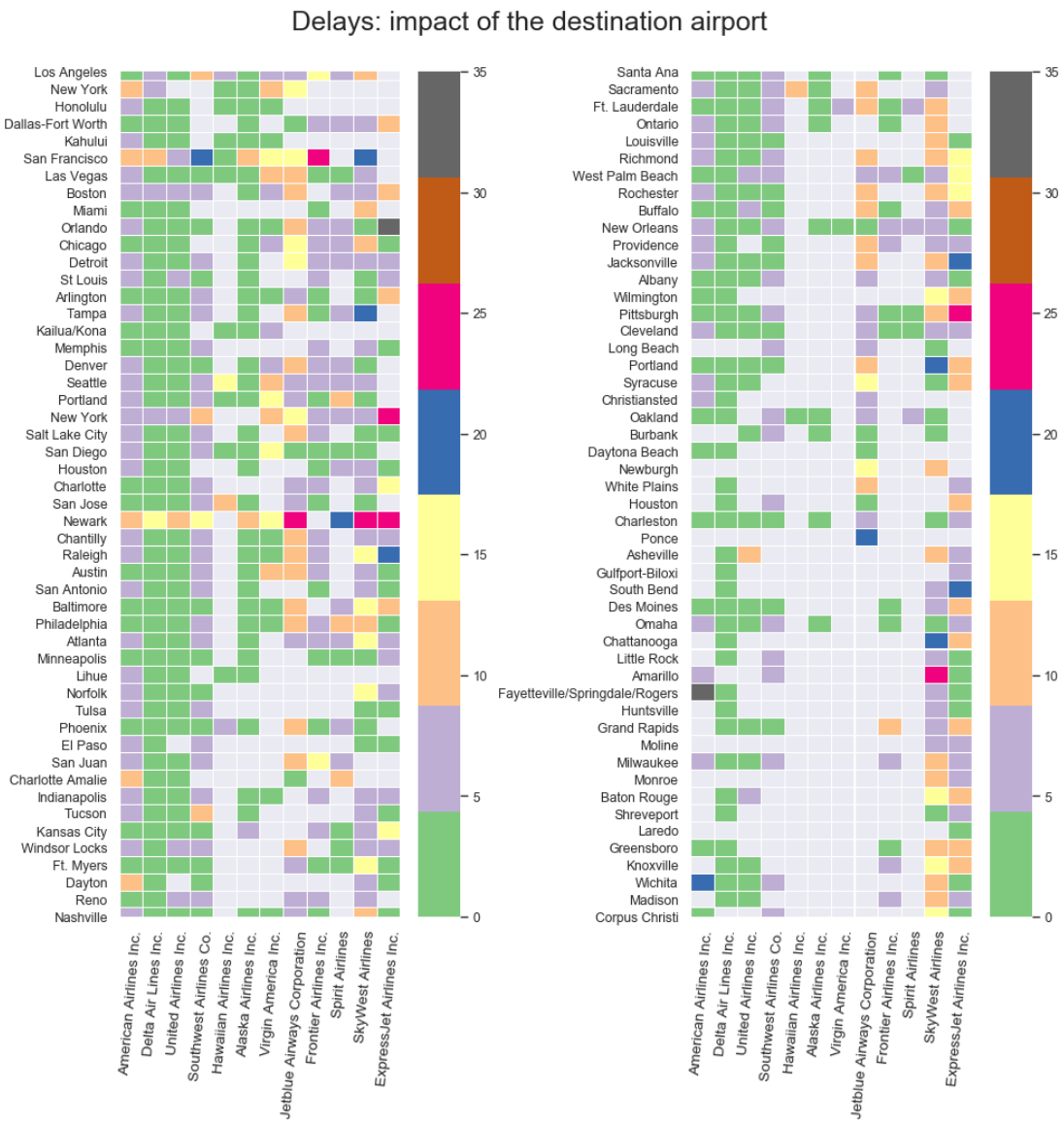
From the last section, we can observe that there is a huge flight network if we connect the airports (vertices) with flights (edges). And the PageRank algorithm, which originally developed by Google for ranking search results, can be useful for us to evaluate the busyness of each airports.



<Airports PageRank Trends in 2017>

Looking at the singular most important airport in 2017 is interesting but identifying the trend in an airport's PageRank over time is even more interesting. To do that, we must first calculate the PageRank of each airport month-by-month over time. We can see that there are few airports (especially *Hartsfield-Jackson Atlanta International*) are always the busiest airports. However, the busyness of other airports does vary with different months with a year.

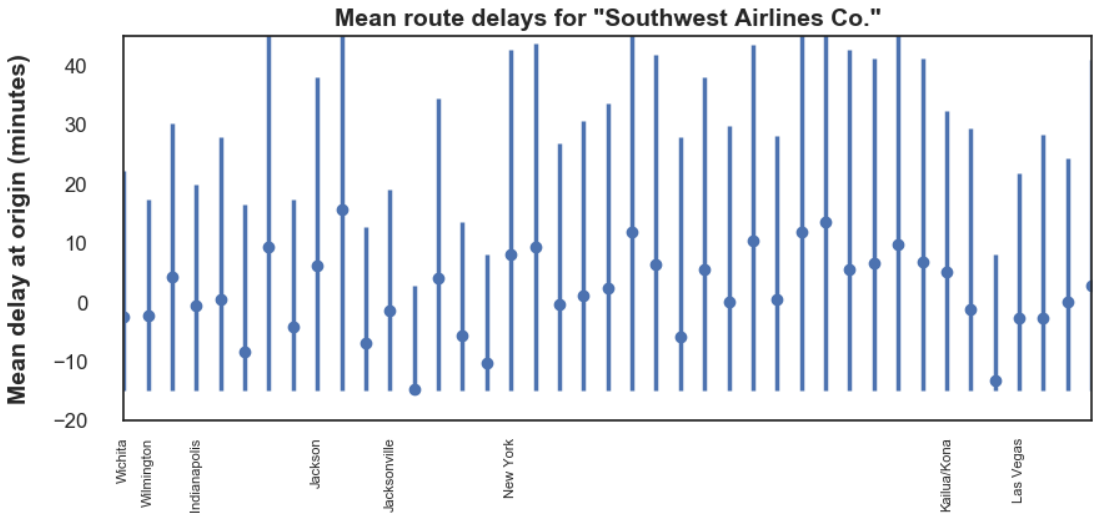
3.6 How the destination airport impact delays



This figure allows to draw some conclusions. First, by looking at the data associated with the different airlines, we find the behavior we previously observed: for example, if we consider the right panel, it will be seen that the column associated with *ExpressJet Airlines and JetBlue airways* mostly reports large delays, while the column associated with *Delta Airlines* is mainly associated with delays of less than 5 minutes. If we now look at the airports of origin, we will see that some airports favor late arrival: see e.g. Newark and San Francisco. Conversely, other airports will mainly know on time departures such as Las Vegas and San Jose.

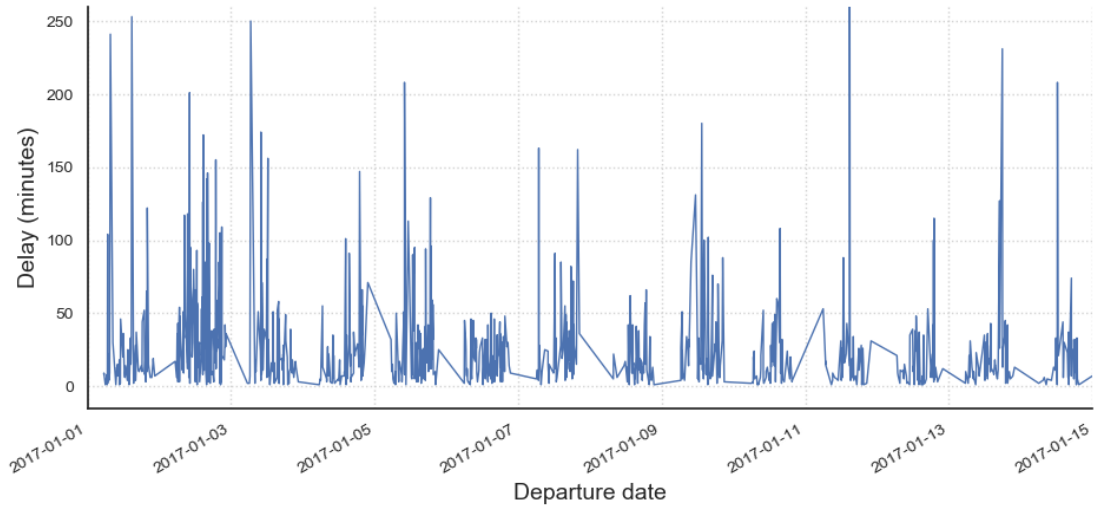
Finally, we can deduct from these observations that there is a high variability in average delays, both between the different airports but also between the different airlines. This is important because it implies that in order to accurately model the

delays, it will be necessary to adopt a model that is specific to the company and the destination airport.



This figure gives the average delays for *SouthWest Airlines*, according to the city of origin and the destination (note that on the abscissa axis, only the origin is indicated for the sake of clarity). The error bars associated with the different paths correspond to the standard deviations. In this example, for a given airport of origin, delays will fluctuate depending on the destination. We see, for example, that here the greatest variations are obtained for New York or Jackson where the initial average delays vary between 0 and ∼30 minutes.

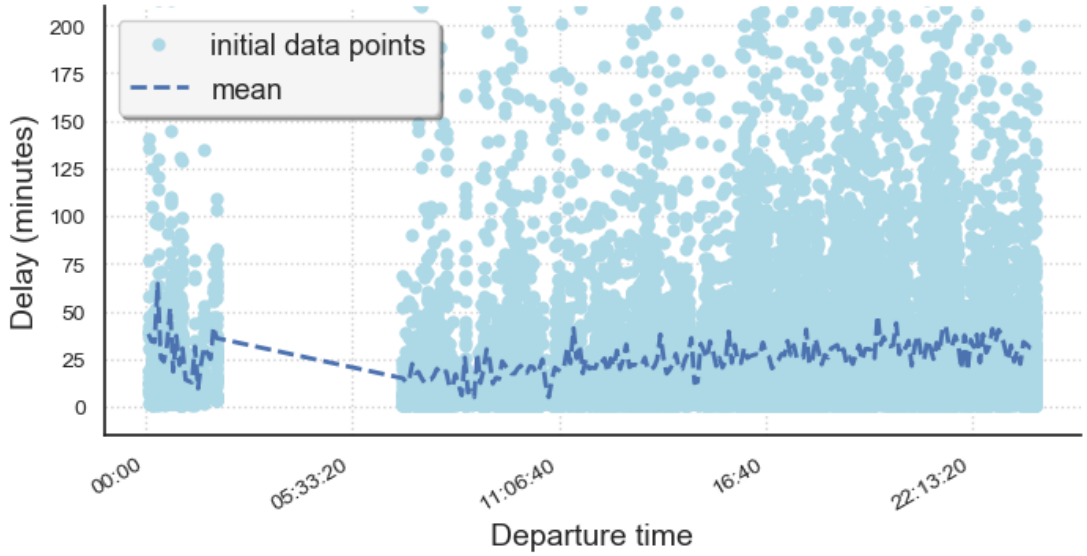
3.7 How delays vary with time



This figure shows the existence of cycles, both in the frequency of the delays but also in their magnitude. In fact, intuitively, it seems quite logical to observe such cycles since they will be a consequence of the day-night alternation and the fact that the airport activity will be greatly reduced (if not inexistent) during the night. This suggests that an important variable in the modeling of delays will be take-off time and

arrival time. Also, on 2017-01-03, we can observe higher delays in both day and night, which could be due to the end of holiday.

To check this hypothesis, I look at the behavior of the mean delay as a function of departure time, aggregating the data of the current month:



Here, we can see that the average delay tends to increase with the departure time of day: flights arrive on time in the morning and the delay grows almost monotonously up to 30 minutes at the end of the day. During midnight, the delay could spike up to 60 minutes. In fact, this behavior is quite general and looking at other airports or companies, we would find similar trends.

4. Modelling arrival delays

In the Exploratory Data Analysis stage, we have seen that the effect of seasonality and special dates is affecting the delay rate of flights a lot. However, the original provided dataset contains only flights data in 2017 which I believe it is far from enough for us to model these effects. So, I again downloaded the dataset published by U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics from 2015 to 2017. The train/test/validation set splitting is as follow:

Training set: 2015

Testing set: 2016

Validation set: 2017

For this project, I built a regression model to predict how long a flight would be delayed. Less negative effect due to class imbalance and importance of predicting the level of delay are the main reasons of selecting the objective of regression but not objective of classification.

4.1 Features and Model Selection

In order to achieve predictions with sufficient lead time, many features (for example: weather data) that can only be obtained within shorter period before the flights' departure are not allowed to use. Most of the features left behind are categorical features which is hard to deal with typical machine learning algorithm. Therefore, I have chosen to use the CatBoost model developed by Yandex for my modelling.

Advantages of CatBoost Library:

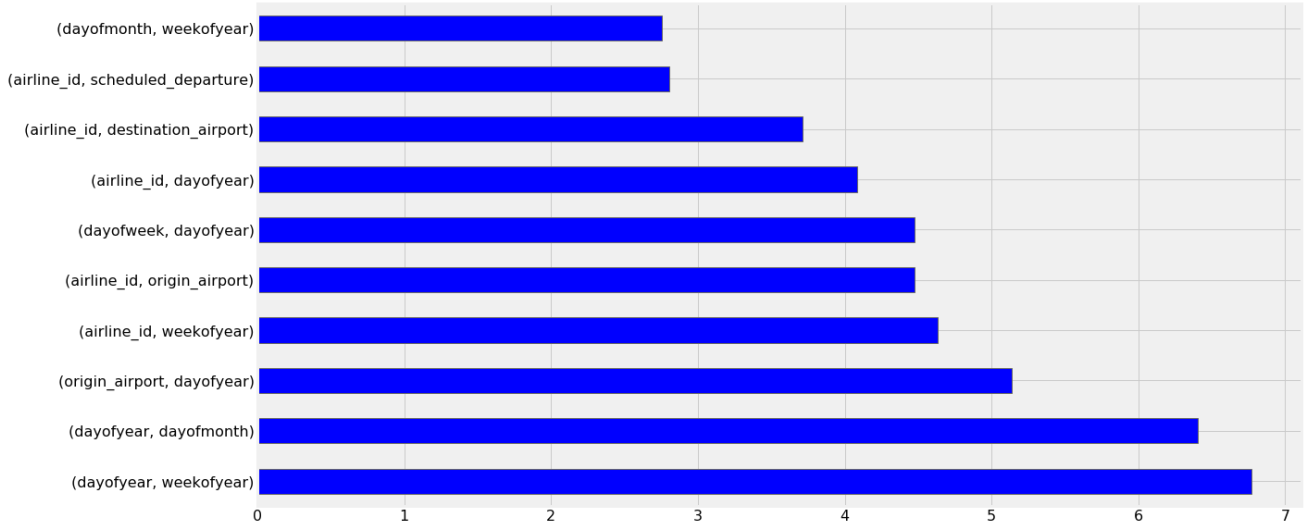
Performance: CatBoost provides state of the art results and it is competitive with any leading machine learning algorithm on the performance front.

Handling Categorical features automatically: We can use CatBoost without any explicit pre-processing to convert categories into numbers. CatBoost converts categorical values into numbers using various statistics on combinations of categorical features and combinations of categorical and numerical features.

Robust: It reduces the need for extensive hyper-parameter tuning and lower the chances of overfitting also which leads to more generalized models. Although, CatBoost has multiple parameters to tune and it contains parameters like the number of trees, learning rate, regularization, tree depth, fold size, bagging temperature and others.

Interpretability: CatBoost is also a tree-based algorithm which can be interpreted easier. This property is essential in making business decisions because it is hard to convince the management based on some black-box models. Also, the availability of feature importance makes the process of feature selection more reliable.

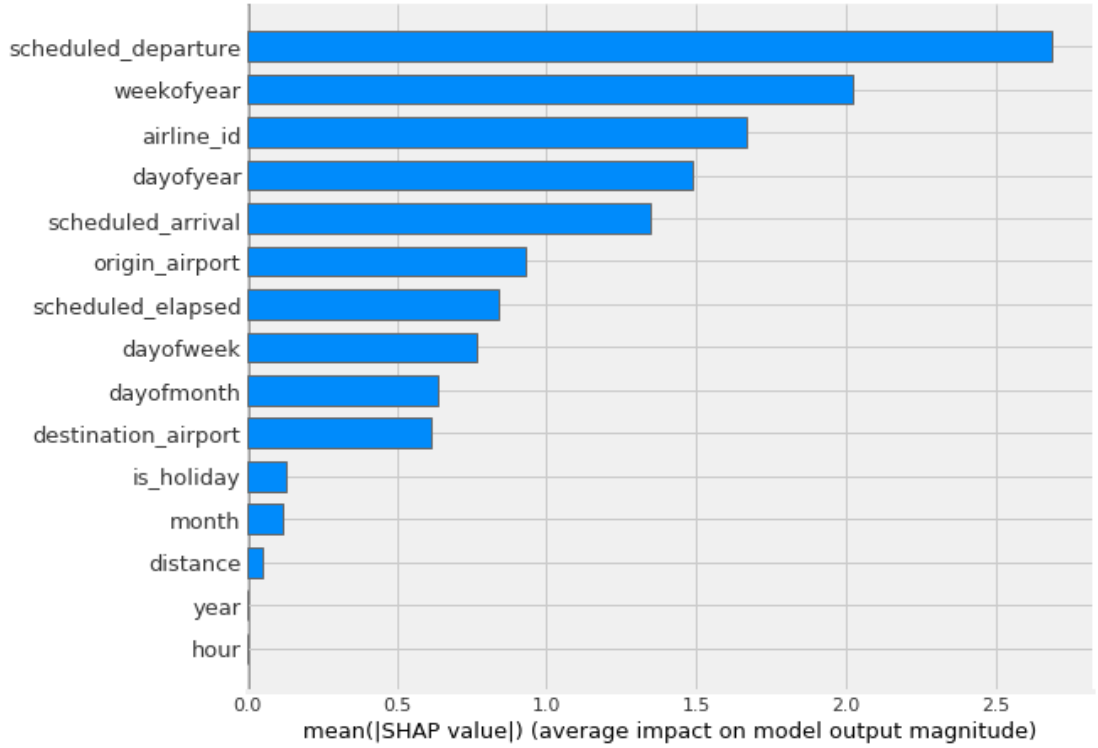
4.2 Features Importance



<levels of interaction between different features>

We can see that features of days have higher degree of interactions. This makes sense because day, week and month solely do not contain a lot information but when they combine with each other, they can reveal some special days or event (for example, Christmas Holiday).

The above graph also implies that some airports or some airlines have higher possibility of delay in certain days, weeks or months.



Considering individual features importance only, scheduled departure is the most important feature which align with our conclusions in exploratory data analysis (flights arrive at night often delay more).

5. Results, Conclusions and Limitations

The median of flight delays in 2015 are used to calculate the baseline for our model in 2017. Assume all the predictions of flight delays in 2017 are the median of flight delays in 2015, the root mean squared error is 46.8.

The root mean squared error of the predictions from CatBoost is 40.1 which beats the baseline and we can conclude that the CatBoost model built on these features can help us figure out some flights with higher possibility of delay. Although this model is certainly not good enough, it shows that this direction of research is possible to reduce the possibility of flight delays in USA.

Another valuable insight is that some of our common reasonings of why a flight would delay are proved to be true. Flights of some airlines, some airports, some timings have higher chances of delaying. These factors can actually be quantified and modelled.

Limitations

1. Lack of time and data

Most of my time are used in EDA and more can be done in the modelling part to improve the model performance. Especially grid search for hyper-parameter tuning and cross-validation on several years of data if more data is available.

2. Lack of computational power

In EDA stage, I have calculated the rolling monthly PageRank of each airport in 2017 but it is not used in modelling as there was no enough time and computational power to compute those. I have already used disturbed computing with spark on google cloud platform to speed up the computation, but it is still not fast enough to get the results before the deadline.

3. Alternative data

Probably most of the flight delays are still caused by extreme weathers which our model cannot interpret this factor. Ideally, we can collect the dates of extreme weathers and remove the delayed flights due to extreme weather. So, we can assume the remaining flights are delayed because of bad schedules.

Also, the US event data could be useful too because we can obtain this information with sufficient lead time.