Flight Delay Prediction

Shivanirudh S G

September 17, 2020

1 Introduction

Air transport, right from when the first airliner took off, has been steadily gaining importance, and has claimed a top spot in the transport sector on a global scale. The aviation industry is the only transport industry that has a worldwide network, making it a necessity for international business and tourism. With a headcount of close to 2 billion passengers and 40% of the total worth of all transported cargo annually, this industry affects the lives of 29 million people via direct, indirect, induced and catalytic means to provide employment. The global economy has a contribution of USD 2960 billion from this sector, translated to approximately 8% of the world's Gross Domestic Product (GDP) [1].

Such tremendous growth naturally has a side-effect, namely an increase in the number of flights, causing delays. In the United States alone, it was found that about 25% of passengers experienced a delay in their flights, with an average amount of 1 hour and 54 minutes in the year 2007 [2]. In the years 2016 and 2017, 32% of all airline operations in the United States have been delayed, with an average of 1.7 million delayed minutes each month [3]. 10 airports in the USA are part of the top 30 busiest airports in the world, with approximately a third of the flights being delayed.

It might be unfair, however, to put all the blame on just traffic and not the medium. The ever-changing weather, predictable only to a bare minimum extent, influences where and when any flight can take off or land.

This paper begins by describing the premises and the problem statement. Next, the dataset under scrutiny is explained, and further analysed in detail. Following this, the models used to make the predictions are dealt with, and the observations compared. The results and conclusions derived constitute the final section of the paper.

2 Problem Statement

As mentioned previously, it becomes very important to determine delays beforehand as each incident questions the reliability and fragility of the industry.
This paper attempts to make a prediction on whether a given flight is delayed upon its arrival at the destination and if so, by how long. This paper considers a flight as delayed on arrival only if it arrives at least 15 minutes after its scheduled arrival time. Also, all issues that could have arose for a flight during its journey are ignored and only the conditions at the end points of the route are considered.

3 Dataset

The dataset focuses on 15 of the busiest airports in the USA, and every flight scheduled to take off from or land in these airports in 2016 and 2017. With about 1.8 million records, this dataset consists of details regarding the routes, and the weather conditions at the time and location of departure. Each record has the details of the route the flight is on, including origin and destination airport, the scheduled and actual times of departure and arrival, and the prevailing weather conditions on the scheduled date and time at the origin airport. The attributes of the dataset are given below.

FlightDate	Quarter	Year	Month	DayofMonth
DepTime	DepDel15	CRSDepTime	DepDelayMinutes	OriginAirportID
DestAirportID	ArrTime	CRSArrTime	ArrDel15	ArrDelayMinutes
windspeedKmph	winddirDegree	weatherCode	precipMM	visibility
pressure	cloudcover	DewPointF	WindGustKmph	tempF
WindChillF	humidity	date	time	airport

Table 1: Attributes of dataset

4 Exploratory Data Analysis

The dataset under consideration consists of multiple features that could have determined the arrival time of a flight, with a few interesting observations and trends with respect to the delayed flights. With delayed flights having a low percentage of approximately 21%, the dataset is skewed heavily on flights that have been on time. Some of the broad categories under which the trends are observed are dealt with here.

4.1 Origin and Destination Airport

In the matter of origin and destination airports of a flight, there has been a significant amount of debate on whether the delays are fault of the airports themselves, or something beyond their control, with the balance tipping toward the unprecedented nature of weather. However the following prominent trends have been observed.

- Contributing the most, the Los Angeles International Airport holds the top place as both an origin and a destination site to flights that were delayed.
- The George Bush Intercontinental Airport in Houston functions closer to the schedule, having the least number of flights that take off from there being delayed.
- Flights scheduled to land at the Charlotte Douglas International Airport seemed to be more on time than other airports.

4.2 Weather Conditions

As mentioned in Section 1, weather plays an important role in the aviation industry. It is expected that a greater number of delays are seen as the weather gets worse.
However, an analysis of this dataset gave the result that most of the flights were delayed under the weather conditions being ”CLEAR”, as illustrated in Figure 1, plotted using the top 8 most frequent weather conditions.
To support this observation two main aspects of weather are considered in this analysis.

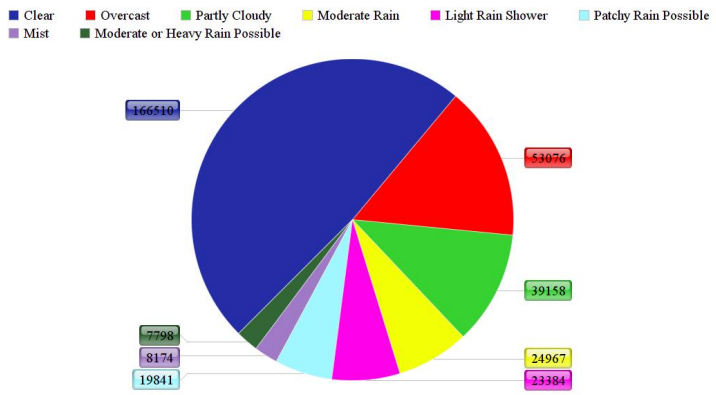

Figure 1: Weather conditions during delay

4.2.1 Visibility

Visibility is defined as the greatest distance through the atmosphere toward the horizon that prominent objects can be identified with the naked eye. Flight visibility is the average forward horizontal distance, from the flight deck, at which prominent unlighted objects can be seen and identified. A minimum technical legal requirement for landing is that visibility be at least 3 miles [4].
Atmospheric phenomena that impact visibility greatly are fog, rain, snow, volcanic ash, haze, airborne dust and sand [4]. As the presence of these factors change across the year, they seem to have a notable grip over the take off and landing aspects at the airports.
Despite this, as illustrated in Figure 2, maximum number of flights were found to be delayed at a visibility level much higher than the benchmark, at 10 miles.

4.2.2 Cloud Cover

Cloud cover or ceiling is a measure of the height of the cloud base relative to the ground. It is measured in terms of one-eighth of the total sky cover ranging from SKY CLEAR (SKC) at $\frac{0}{8}$ to OVERCAST (OVC) at $\frac{8}{8}$ [5].
Cloud cover and visibility together play a vital role in determining the landing of a flight. If permitted for a "visual" approach during landing, or controlling the aircraft by what can be seen, the pilots are given control to choose their separation from other aircrafts. This means that when ceiling gets very low and the visibility

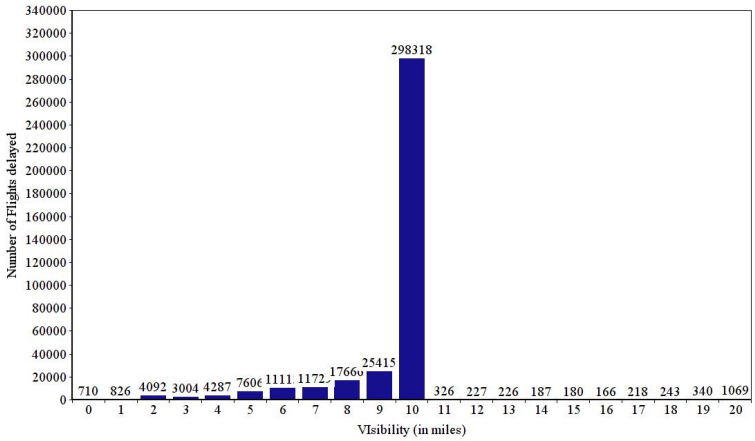

Figure 2: Effect of visibility on flight delay

drops below 3 miles, most aircrafts will be operating on an "instrument" based approach for landing, that is, relying on the internal devices of the aircraft to take readings and make decisions [6]. In such situations, the number of flights that could land at or take off from the airports become lower than usual.
For example, airports like SFO and LAX operating in close parallel arrivals, handling about 60 landings an hour in good weather conditions, have the numbers reduced to just 25 per hour[6]. Not only would a departing flight not be able to take off, but also an arriving flight may not be able to land, and end up getting diverted to an alternative airport.
As proved by the depiction in Figure 3, the maximum number of flights have been delayed when the cloud cover is 0, the least value, rather than at 100.

5 Prediction Model

5.1 Features considered

Out of the 30 features in the dataset mentioned in Section 2, the following fields may be omitted.

- **DepTime, CRSDepTime and DepDelayMinutes**: To check if the arrival delay is simply a translation of the departure time, a s mention of whether it departed on time is enough, since there is a buffer zone of 15 minutes

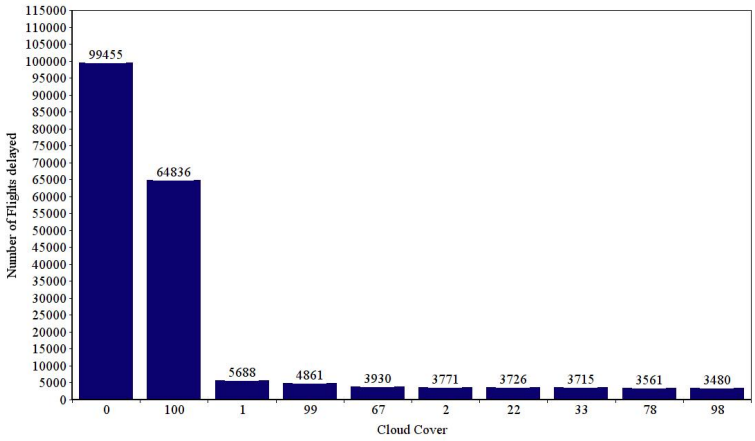

Figure 3: Effect of cloud cover on flight delay

for each flight and all issues that may arise during the flight time are not considered.

- **date, FlightDate, Quarter and airport**: Since the dataset is already matched according to the departure time and the prevailing weather conditions, these features are redundant.
- **humidity, WindChillF, tempF, DewPointF**: These factors directly or indirectly affect or derived from visibility and cloud cover, and hence, in a way, already taken into account.
- **windGustKmph** : This value is represented by another field "**windspeedKmph**" and therefore, redundant.
- **weatherCode** : This field is simply influenced by all other weather factors, and does not play an independent role in the prediction.

This results in the dataset having only the below mentioned columns.

Year	Month	DayofMonth	DepDel15	OriginAirportID
DestAirportID	ArrTime	CRSArrTime	ArrDel15	ArrDelayMinutes
windspeedKmph	winddirDegree	precipMM	visibility	pressure
cloudcover	time			

Table 2: Features under consideration

5.2 Preprocessing

- **Scaling**: Since almost every field that falls under the weather category of the dataset has a different range, the values are brought to an approximate central value. This also ensures that the model is not biased over fields that have a higher value range.
- **Sampling** : As mentioned in Section 4, this dataset is heavily skewed with flights that have arrived on time at almost 80%. To balance both classes of data, sampling is performed. To avoid loss of data points that would arise from under sampling, the dataset is processed through over sampling, where the records of the minority class are duplicated until the number of records in each class becomes equal.

5.3 Performance Metrics

5.3.1 Classification

- **Precision:** $Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} = \frac{True\ Positive}{Total\ Predicted\ Positive}$
- **Recall:** $Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} = \frac{True\ Positive}{Total\ Actual\ Positive}$
- **F1 Score:** $F1Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

Precision and recall are superior to accuracy in classification owing to the fact that mis-classification of data is difficult to trace with just accuracy, and hence, the former two metrics are taken under consideration. F1 score helps in maintaining a balance of the two metrics, as it is their Harmonic Mean.

5.3.2 Regression

- **Mean Absolute Error (MAE) :** $MAE = \frac{1}{n} \times \sum |y_{actual} - y_{predicted}|$
- **Mean Squared Error (MSE)** : $MSE = \frac{1}{n} \times \sum (y_{actual} - y_{predicted})^2$
- $R^2 : R^2 = 1 - \frac{MSE(model)}{MSE(baseline}$

All three of these metrics are considered in evaluating the regression models. However, MSE is given a priority over MAE because of the latter's inability to penalize large errors, which is necessary in this prediction problem.

5.4 Classification

The information in Table 3 compares the results of various models used for classification.

Classifer	Class	Precision	Recall	F1-Score
Logistic Regression	0	0.94	0.93	0.94
	1	0.76	0.77	0.76
Decision Tree Classifier	0	0.95	0.16	0.27
	1	0.23	0.97	0.38
Extra Trees Classifier	0	0.94	0.94	0.94
	1	0.78	0.77	0.77
Gradient Boosting Classifier	0	0.96	0.54	0.69
	1	0.34	0.91	0.50

Table 3: Results of Classification

As can be clearly seen, the Extra Trees Classifier gave the best outcome for both the classes, amongst the four classifiers.

5.5 Regression

The original dataset is put through various regression models to predict arrival delay in minutes. The results are tabulated in Table 4. It is clear that the Extra Trees Regressor performed better than the others for the original dataset.

Regressor	MAE	MSE	R^2
Linear Regression	12.28	309.10	0.939
Extra Trees Regressor	11.91	282.98	0.944
Gradient Boosting Regressor	11.71	283.43	0.944

Table 4: Results of Regression

5.6 Pipeline

The dataset is put through the best performing classifier, filtered and then through the best performing regressor to evaluate its performance sequentially. Figure 4 illustrates the structure of the pipeline.

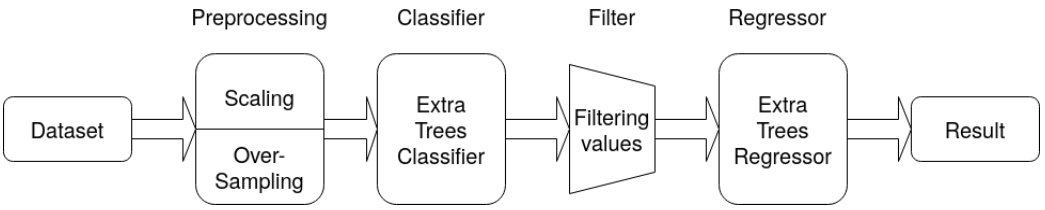

Figure 4: Structure of pipeline

Table 5 shows the results of the pipeline.

Component	Name	Metrics		
Classifier	Extra Trees Classifier	Precision	Recall	F1-Score
		0.77	0.77	0.77
Regressor	Extra Trees Regressor	MAE	MSE	R^2
		22.6	601.8	0.662

Table 5: Results of Pipeline

5.7 Regression Testing

In order to determine the performance of the model under various ranges of arrival delays, the test set is split into multiple ranges on the basis of the amount of arrival

delay and passed through the model. Table 6 tabulates the results.

Range	Data points	MAE	MSE	R^2
15 to 100	64588	10.73	199.8	0.59
100 to 200	9797	18.6	717.4	0.013
200 to 500	2943	20.03	920.2	0.784
500 to 1000	237	22.73	1084.7	0.94
Above 1000	47	17.09	1039.8	0.92

Table 6: Results of Regression Testing

The model performs better at a lower range between 15 and 100, and at a much higher range of above 1000. As illustrated in Figure 5, it proves to be more effective for a vast range of daily occurrences.

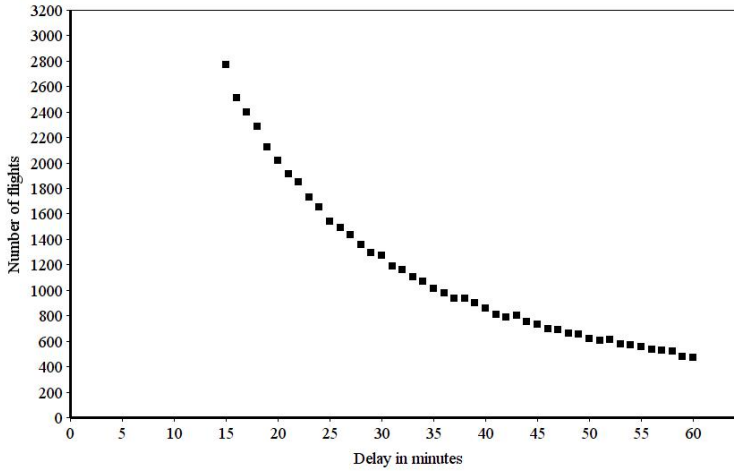

Figure 5: Flights delayed by less than an hour

6 Results and Conclusion

The various models that were implemented to predict the arrival delay of flights, performed as well individually on the original dataset, as when the classifiers and regressors were in tandem.
The results of regression testing also showed that the model is more accurate to be used in regular scenarios, where delays generally do not exceed 100 minutes.
Majority of the weather related fields in the feature set are interrelated more closely than expected, allowing elimination of that many redundancies.

References

[1] Air Transport Action Group. *The importance of the industry - facts and figures*. 2005. URL: https://www.icao.int/meetings/wrdss2011/documents/jointworkshop2005/atag_socialbenefitsairtransport.pdf.

[2] *Delays in flights in the USA*. URL: https://en.wikipedia.org/wiki/Air_transportation_in_the_United_States#Delays.

[3] *Weather's share of delays*. URL: https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?pn=1.

[4] *Visibility*. URL: https://www.universalweather.com/blog/aviation-weather-tips-visibility/.

[5] *Cloud Cover*. URL: https://www.universalweather.com/blog/aviation-weather-tips-all-you-need-to-know-about-ceilings/.

[6] *Why does weather cause air traffic delays?* URL: https://gizmodo.com/why-does-weather-cause-air-traffic-delays-1706282676.