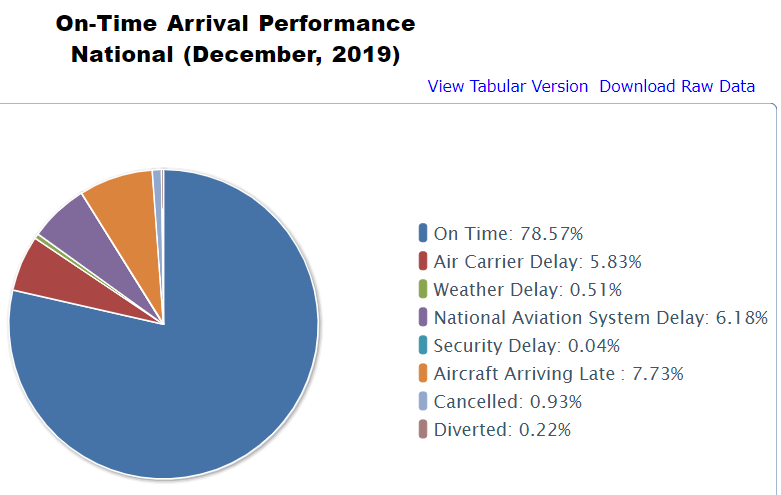
**Flight delay prediction**

1 Introduction

Over the last few decades, air transport is increasing popularity because of speed and comfort which eventually increase the traffic in the airspace. With the great increase in air traffic comes a large increase in the demand for airport capacity. However, airspace and airport capacity cannot keep increasing at a rate necessary to match the rising demand. When an airport's capacity is reduced during “peak hours", the demand for an airport's resources exceeds the capacity that the airport can afford. Many major airports around the world have significant delay problems as a result of an imbalance between capacity and demand. Flight delays are obviously frustrating to air travellers and costing billions to airlines and travellers. The study found that airlines with high rates of delay also have higher operating costs overall. The direct cost to airlines included increased expenses for crew, fuel, and maintenance, among others. Flight delays are complex to explain and subdivided into many types of delays like weather delays, Air Carrier delays, security delay, National Aviation system delay.



Related Work:

There are several works in the literature that focus on airport delays, air-traffic management and optimization. Airline operations are highly complex processes that are intended to regulate many expensive, tightly constrained, and interdependent resources, such as the crew, aircraft, airports, and maintenance facilities. Many studies have been carried out on airline planning problems, but only a few have been performed on the characteristics of airline delays and the prediction of delay statistics. Delays occur when an event takes place later than the time at which it is planned, scheduled, or expected to happen [12]. Delays in departure can occur due to bad weather conditions, seasonal and holiday demands, airline policies, technical issues such as the problems in airport facilities, luggage handling and mechanical apparatus, and accumulation of delays from preceding flights.

Dieterich Lawson and William Castillo, in their project” Predicting Flight Delays” in 2012 (Lawson & Castillo, 2012), used dataset of flights but included several years, which resulted in an impressive number of data points (135 million of flights). However, they limited their features to weather data only (33 features only in the end) obtaining 40% recall only. They used algorithms like support vector machines, Naïve Bayes and Random forest.

In another research article by Nathalie Kuhn and Navaneeth in their project “Application of Machine Learning Algorithms to predict flight delays “used machine Learning algorithms like decision trees, Neural Network and Logistic regression algorithms and concluded that Departure delay is main factor in aircraft delay.

Another report published by Neil and Samir report titled “Predicting flight delays and cancellations using weather as a feature” compared algorithms like Random forest, XGboost(Extreme Gradient), Linear regression, Lasso, and SVM with AUC(Area Under Curve) 0.81 highest among all algorithms.

**Dataset and Features:**

To train and test our models, we used a publicly available dataset for United States domestic air-traffic. The original source of our dataset is the Bureau of Transportation Statistics. The data set is for the year 2019 and consists of well over 1.5 GB data with 29 variables (But we use that is required for project) categorized as follows:

• Information about flight (day, day of the week, airline, flight number, tail number)

• Information about origin and destination (origin airport, originstate, destination airport, departurestate)

• Information about the departure (depdelay, depdelayMinutes, depdel15)

• Cause of Delay (Carrier Delay, Weather Delay, NASDelay, SecurityDelay, LateAircraftDelay)