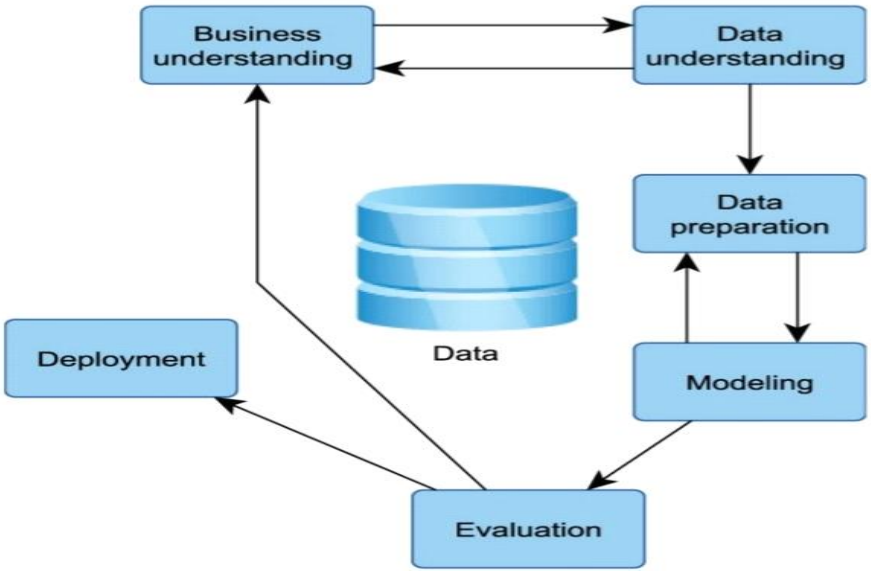
FLIGHT DELAY PREDICTION SOLUTION ARCHITECTURE

Introduction:

The purpose of this project is to look at the approaches used to build models for predicting flight delays that occur due to bad weather conditions.In this project, we look at using Python based Logistic Regression along with Support Vector Machine and then plugging the dataset into our classifier for results.



Predictive Models For Flight Delay:

We would like to build an analysis dataset by choosing the threshold of 15 minutes, beyond which we consider the class change to "delayed" flight. This is a standard threshold in the aviation industry, with indicators on delayed flights commonly based on 15 minutes of delay.

For flight departure delay prediction, the following features are potential candidates for the model:

1. Month

2. Day of Week (weekday vs. weekend)

3. Departure Hours (convert from (CRSDepTime))

4. Arrival Hours (convert from (CRSArrTime))

5. Departure Airport

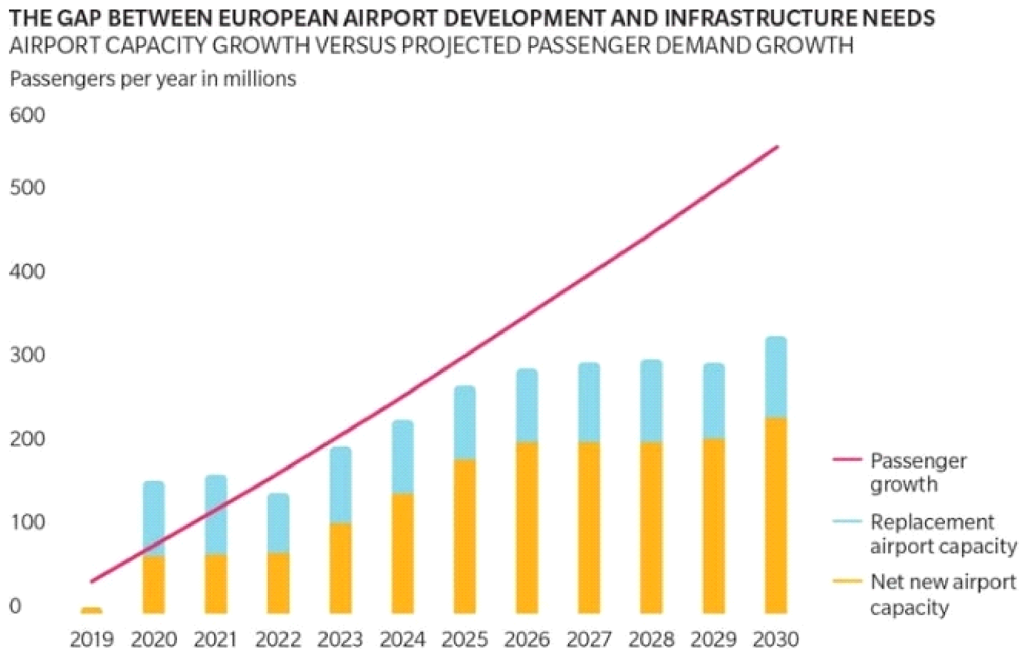
6. Arrival Airport

7. CRSElapsedTime (total time for a flight)

8. Flight Distance

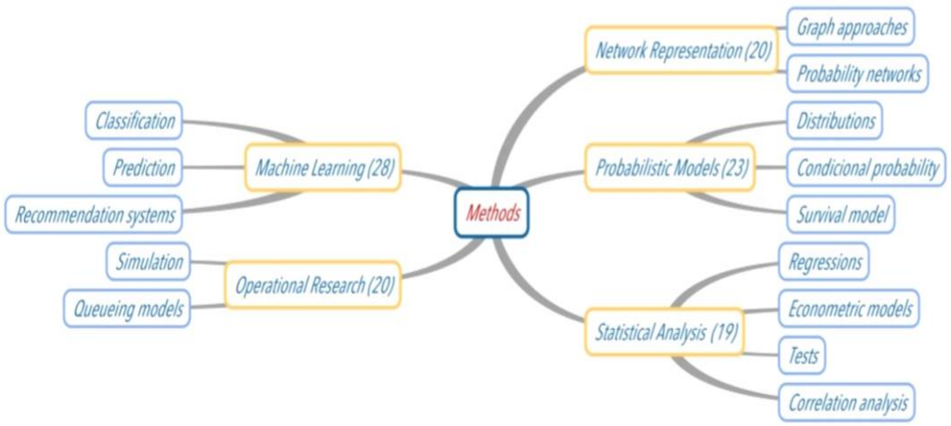
9. Carrier Name

we choose the direction to consider - that is, whether Chicago ORD is the origin or the destination. Additionally, we decide what type of delay to consider (YColumns), either Departure delay measured by DepDelay, or the Arrival delay measured by ArrDelay. We can come back to this section to change the choices. We select a subset of columns of most interest (XColumns). CRSArrTime will be dropped since the hour is sufficient. Also, the TaxiIn, TaxiOut and Diverted variables will be dropped. TailNum is the tail number of the plane and could be interesting to match with plane information like the number of seats. Since the dataset is so large, we have to constrain it so that Python can perform analysis on our system in a reasonable amount of time. To do so, the dataset is sampled randomly for 20k rows. Further, these observations have be randomly split into training and test sets, so that 10k observation is used for analysis.



Method:

The flight delay prediction problem may be modeled in many ways, depending on the objectives of the research.Methods were divided into five groups. The numbers next to each category represent the number of related papers.



Statistical Analysis:

Statistical analysis usually encompasses the use of regression models, correlation analysis, econometric models, parametric tests, non-parametric tests, and multivariate analysis (MVA). When it comes to regression models, both delay multiplier and recursive models can help airlines to understand delay propagation effects through the network and to estimate the costs of delays [16, 115, 84, 124, 127]. Many econometric models are also build to evaluate the efficiency flight systems, such as the analysis of the investments done by a governmental agency [88] or to evaluate the equilibrium point considering the relationship between delays and passenger demand, fares, frequency and size of the aircrafts [131]. Xiong et al. [122] built an econometric model based on pre-existing delays, potential delay savings, distance, characteristics of the destination airport and airline, frequency, aircraft size, occupancy rate and fare to understand which reasons lead airlines to cancel their flights.

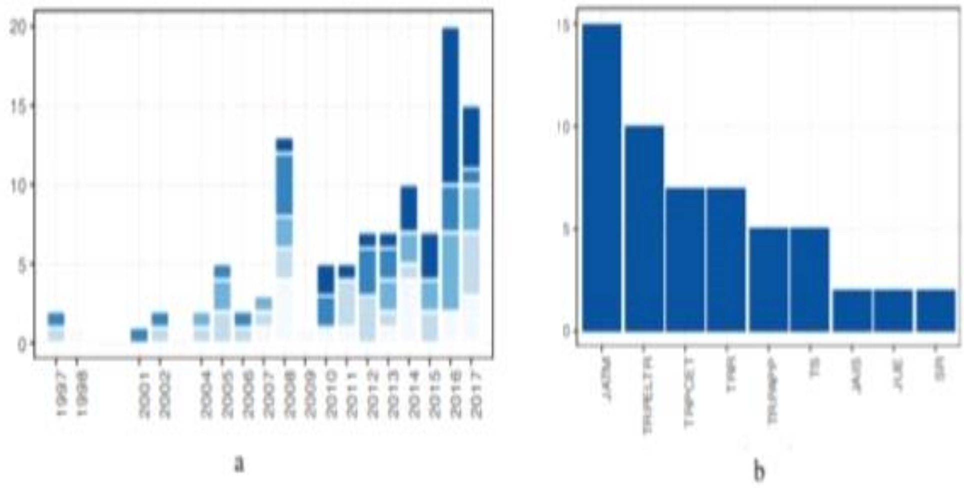
Qin et al. [101] studied the periodicity of flight delay rate, whereas Mofokeng et al. [87] studied the impact of aircraft turnaround time during maintenance check. Finally, Hao et al. [61] built a model to quantify how delays originated at New York are propagated to other airports. Some works focus on statistical inference. Pathomsiri et al. [94] used a non-parametric function to evaluate the efficiency of airports of the United States regarding delays. Reynolds et al. [103] computed the correlation between levels of delays and capacities of the European airports. They also suggested different approaches to deal with the congestion problem, describing their advantages and disadvantages. Finally, Abdel-Aty et al. [1] calculated daily average of delays to detect correlations to understand the principal causes of delays at Orlando International Airport.

Network Representation:

Network representation encompasses the study of flight systems according to a graph theory. Abdelghany et al. [2] built direct acyclic graphs to model the schedule of an airline (including flight times and resources availability) to detect disruptions and their impacts on the rest of the network. They used the classical shortest path algorithm to evaluate propagation effects. Ahmadbeygi et al. [3] built propagation trees to compare two different airlines, one operating in a conventional huband-spoke scheme and the other in a low-cost point-to-point system. Xu et al. [123] and Wu et al. [120] built a Bayesian network to model delay propagation. Baspinar [14] built a network-epidemic process using historical flighttrack data of Europe to create a novel delay propagation model**.**

Result And Discussion:

Since flight delays cause economic consequences to passengers and airlines, recognizing them through prediction may improve marketing decisions. Due to that, several forecast models have been built over the last twenty years. These models have sought to understand how delays propagate through the network of flights or airports, to predict root delay in the system or to comprehend the cancellation process. Beyond these three points of view for treating the flight delay prediction problem, models could also differ by their scope of application, data issues and methods.



Conclusion:

Flight delays are an important subject in the literature due to their economic and environmental impacts. They may increase costs to customers and operational costs to airlines. Apart from outcomes directly related to passengers, delay prediction is crucial during the decision-making process for every player in the air transportation system. In this context, researchers created flight delay models for delay prediction over the last years, and this work contributes with an analysis of these models from a Data Science perspective. We developed a taxonomy scheme and classified models in respect of detailed components.