Project Proposal: Flight Status Prediction

Ardavasd Ardhaldjian, Andrew Fogarty, Calisto Betti, Michael Fawehinmi, Shreedhar Kodate, & River Yan

MIS587: Business Applications in Machine Learning

Group 6

10/27/22

Project Objectives

What is the business problem? Provide a detailed background on the problem domain including relevant research and other publications

In an increasingly global society, there are hundreds of thousands of flights and millions of people that fly every day. According to financeonline, Before the COVID pandemic, the number of flights around the world was over 100,000 and the number of people flying was in the tens of millions. Since the pandemic numbers have gone down, however, these numbers seem to be on a steady rise to pre-2019 numbers. People fly for various reasons including leisure, business, health, and family. The expectation is that flights will depart and arrive on time, however, In any given year around 20% of flights get delayed for various reasons. These could be due to punctuality mishaps, severe weather conditions, air traffic issues, and many more reasons. Usually, these reasons are hidden from any single person therefore it is difficult to predict whether a flight will be delayed.

Within our project, our business problem is to see whether the variables within our dataset can help predict whether a passenger's flight will be delayed or not. Having the ability to predict the status of a flight accurately can be extremely valuable for Air Traffic Control at airports.

Why is it important for the project sponsor (if you have one) or for a particular organization?

Predicting accurately whether a flight is going to be delayed could impact a variety of stakeholders. For starters, potential passengers might want to know if a flight is going to be delayed to increase their travel experience. Airlines might want to understand the correlations between delayed flights and various factors to better serve customers. The air traffic control at an airport would want to predict delays in order to better prepare for flights coming in and out of an airport and prevent traffic. Event shops, restaurants, and car rental companies at airports might want to know when flights are delayed to better predict when to expect customers. Through a model that can predict either departure or takeoff delays, the potential stakeholders mentioned before can save time, money, and effort.

Why is this problem suitable for machine learning approaches?

This business problem is fitting for machine learning methodologies. This is because there are over 60 columns to examine, providing a lot of potential feature variables that can have a strong correlation with the target variable. Through exploratory data analysis and feature engineering, we hope to find the best combination of features to have the most accurate model possible. This problem is also a machine learning problem because of the depth of the data. There are too many instances to manually analyze the data. Instead, machine learning should be

used to effectively analyze all of the data that includes millions of instances for each of the five datasets.

What is your target variable?

We have two potential target Variables. *ArrDelay*, which is the difference in minutes between scheduled and actual arrival time (Early arrivals show negative numbers), and *DepDelay,* which is the difference in minutes between scheduled and actual departure time (Early departures show negative numbers).

Who is going to use the model? Be specific about the use scenario.

Air Traffic Control (ATC) would benefit the most from our model. At high-traffic airports such as JFK, LAX, and Denver International Airport it is critical to keep the flow of flights going and to minimize hindrances.

How will the model generate value (bottom-line impact) for the project sponsor or the particular organization you have in mind?

Initially, we brainstormed several organizations that could benefit from this model including airlines, passengers, air traffic control, TSA, shops and restaurants, car rental services, and package delivery services. We narrowed our organization down to just air traffic control as we wanted to analyze how this model will generate value for ATC. This will be done through efficient planning and preparation of incoming and outgoing flights, leading to an increase in total number of flights that an airport can safely handle. This will also improve the passenger experience going through an airport, which will increase foot traffic and overall revenue. When traffic is reduced in an airport, it reduces the wait time for passengers when they are anxious to land or take off. The experience of being on an airplane for too long can irritate the passenger, leading to a negative overall passenger experience. If the experience is negative, then the passenger is less inclined to fly with that specific airline or airport. However, if the ATC handles traffic well, then the loss of passengers will be reduced and overall bring in more passengers and revenue.

Explore Data

Nature of Data

The Flight Delay data is a tabular data set containing 60 attributes describing flights from 2018 - 2022.

Data Source

The data is publicly available on Kaggle.com and is pulled from the Marketing Carrier On-Time Performance data table of the "On-Time" database in the TranStat Data library. The 60 attributes contain 50 categorical and 10 numerical variables. The following is a division of the attributes grouped by the general place or event the data describes.

Date: Year, Quarter, Month, DayofMonth, DayOfWeek, FlightDate

Airline: Marketing*Airline*Network, Operated*or*Branded*Code*Share_Partners, DOT*ID*Marketing_Airline, IATA*Code*Marketing_Airline

Flight: Flight*Number*Marketing_Airline, Originally*Scheduled*Code*Share*Airline, DOT*ID*Originally*Scheduled*Code*Share*Airline, IATA*Code*Originally*Scheduled*Code*Share*Airline, Flight*Num*Originally*Scheduled*Code*Share*Airline, Operating_Airline, DOT*ID*Operating_Airline, IATA*Code*Operating_Airline, Tail_Number, Flight*Number*Operating_Airline,Cancelled, CancellationCode, Flights, Diverted (Diverted Flight columns run 1 through 5), DivAirportLandings, DivReachedDest, Div1TailNum.

Time: DepTime, DepDelay, DepDelayMinutes, DepDel15, DepartureDelayGroups, DepTimeBlk, TaxiOut, WheelsOff, WheelsOn, TaxiIn, CRSArrTime, ArrTime, ArrDelay, ArrDelayMinutes, ArrDel15, ArrivalDelayGroups, ArrTimeBlk, CRSElapsedTime, AirTime, CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay, LongestAddGTime, TotalAddGTime, FirstDepTime, LateAircraftDelay, DivActualElapsedTime, DivArrDelay, Div1WheelsOn, Div1TotalGTime, Div1LongestGTime, Div1WheelsOff.

Origin: OriginAirportID, OriginAirportSeqID, OriginCityMarketID, Origin, OriginCityName, OriginStateOriginStateFips, OriginStateName, OriginWac.

Destination: DestAirportID, DestAirportSeqID, DestCityMarketID, Dest, DestCityName, DestState, DestStateFips, DestStateName, DestWac, CRSDepTime, Div1Airport, Div1AirportID, Div1AirportSeqID.

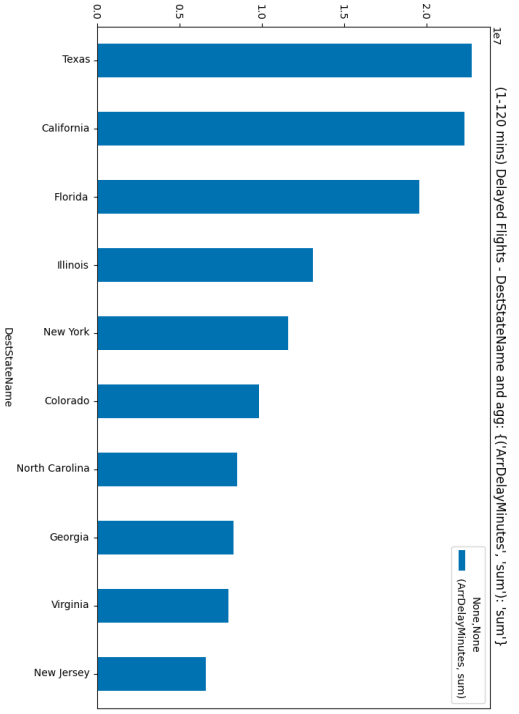
Distance: Distance, DistanceGroup, DivDistance,

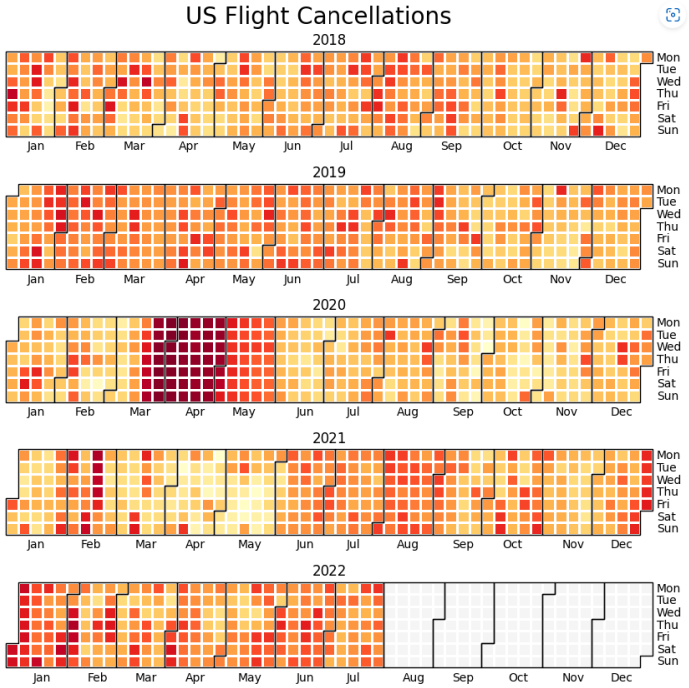
Additional Data Sources

The ability to accurately predict whether or not a flight will be delayed could benefit from information not contained within the Flight Data dataset. Detailed weather data for both the origin and destination would help provide additional context. This information can be pulled from the NOAA World Weather Records Data Set. The passenger traffic at the origin and destination airport is also an important factor to consider. Airports Council International provides data on passenger traffic at international airports around the world.

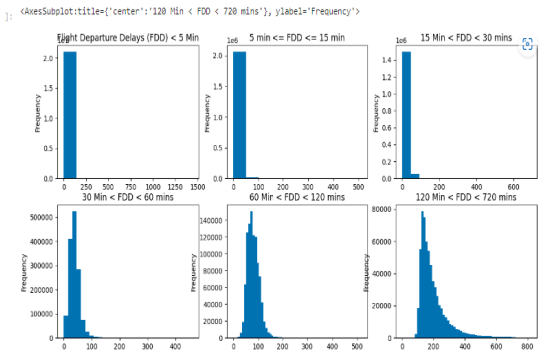
Exploratory Data Analysis

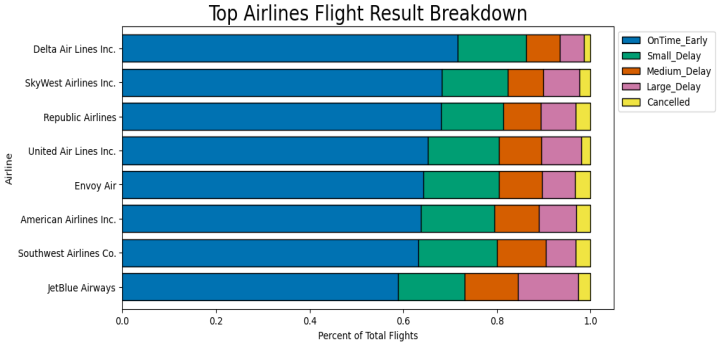
Initial understanding of the data requires understanding the impact of different factors on the cause of flight delays. In this case, the following table looks at the impact departure delay has on arrival delay. As the histograms shows, the more a flight's arrival is delayed the more unpredictable the departure delay becomes. This is especially true for delays of more than 30 minutes.





The heat map of cancelations from year to year shows the distribution of cancelations across days and weeks from year to year. This visualization also gives the team insight into real-world events that might impact the data. In this case, the effect of Covid-19 on flight delays





starting in March 2019. Descriptive information such as the number of flights per airline provides insight into the expected results of any prediction model. The following visual shows the number of flights.

Target Leakage

The raw data does contain target leakage that provides information about flight delays. This information is considered leakage because it is data that describes the flight after the point of analysis. In this case, adding data for predicting if a flight will be delayed has to come before the original departure time.

Feature Engineering

Our group will evaluate several feature engineering methods, both prior to importing our data into DataRobot or Jupyter Notebook and when our data is in the machine learning platform itself. Given the variability in both the *ArrDelay* and *DepDelay* target features, we will intend to use numeric column grouping to assist in classifying differences in flight delay times and determine categories of flight delay values. For example, a flight being delayed by five minutes is less concerning to customers and airlines, then a flight delay of one hour or more. Given the amount of numeric features, supervised learning for our model development seems the most accurate because our dataset can easily categorize our target feature. With the initial analysis of the dataset, dimensionality reduction techniques can be used to ensure proper feature representation is associated with our target variable.[1] Some techniques we are exploring include deleting certain categorical and numerical features which seem irrelevant to our target variable. We will also consider changing variable types as we begin our analysis in DataRobot or Jupyter Notebook, as certain dimensions may be misrepresented towards our target variable, while also adding potential features.

Risk Analysis

Our team assessed four risk categories which have potential impacts on our project. These include model variability, dataset outliers, model accuracy, and model misuse. The sections below discuss our assessments of the prospective risk towards our prediction.

Model Variance

- *Effect towards Prediction:* This can lead to inaccuracy for our prediction model if proper mitigation methods are not employed.

---

[1] Brownlee, J. (2020) *Introduction to dimensionality reduction for machine learning*, *Machine Learning Mastery*. Available at: https://machinelearningmastery.com/dimensionality-reduction-for-machine-learning/ (Accessed: October 26, 2022).

- *Assessment:* The dataset used for this project presents variability in the model prediction due to the different portions of flight delay times. Our team is aware of the concerns with overfitting, where our prediction model may try to match dataset features with non-existent delay times.[2] Given the variability in our target variables, decision tree algorithms may best suit our prediction models, however, the risk of high variance is always a possibility. The team is aware of this risk and intends to incorporate the appropriate techniques to find the right amount of variance for our model.
- *Mitigation Strategy:* Our team understands the importance of using the right sample size when training our prediction model. Several tactics our team will explore will include increasing our model size, reducing regulation, and implementing one-hot encoding to help increase model bias and develop more accuracy.[3]
- *Risk Level:* Medium

Dataset Outlier

- *Effects Towards Prediction:* Outlying variables can influence prediction outcomes by confusing different flight delay time categories.
- *Assessment:* Several outliers within this dataset can have profound influence towards our prediction model, such as weather and mechanical issues. Weather, which delays and cancels flights all the time, is a key influencer towards our prediction model because the variability in delay times cannot be controlled. Solving mechanical issues for an aircraft is also a key outlier, as our dataset does not contain features which identify when certain airline's kept planes grounded for repairs. These factors are difficult to predict, not just for our model, but for the end-user as well.
- *Mitigation Strategy:* Identify dataset outliers and perform exploratory data analysis.
- *Risk Level:* Medium

Model Accuracy

- *Effects Towards Prediction:* Accuracy is essential to correct model prediction and determines whether our model will be fit-for-use.
- *Assessment:* When generating any prediction model, accuracy is a risk that needs to be evaluated. Accuracy is the most important criterion for automated machine learning, and ensuring proper techniques are employed through the model development process is essential for our project employment.[4]

---

[2] Grant, P. (2019) *Introducing model bias and variance*, *Medium*. Towards Data Science. Available at: https://towardsdatascience.com/introducing-model-bias-and-variance-187c5c447793 (Accessed: October 26, 2022).

[3] Brownlee, J. (2021) *How to reduce variance in a final machine learning model*, *Machine Learning Mastery*. Available at: https://machinelearningmastery.com/how-to-reduce-model-variance/ (Accessed: October 26, 2022).

[4] Larsen, K.L.R. and Becker, D.S. (no date) "2.4 Eight Criteria for AutoML," in Automated Machine Learning for Business, pp. 40–41. Undergoing review with publishers.

- *Mitigation Strategy:* Ensure our team uses the proper methods when developing our model and validate our process.
- *Risk Level:* Low

Model Misuse

- *Effects Towards Prediction:* Our prediction model can be used for various consumers, and we need to ensure the reasons for consumption are direct for our target user.
- *Assessment:* Our project outcome is to be able to predict flight delay times. Passengers can use this to coordinate travel plans and accommodations more effectively. Airline companies can use this to reduce costs and increase profit through schedule management. Air Traffic Control (ATC) can use this to improve performance by providing efficiency tools for ATC employees when managing air traffic around airports. All consumers can gain great benefit from predicting flight delay times, however, depending upon who consumes the model, other users could misuse the information and not see the full potential benefits the model could provide. ACT control would not want to use a flight delay prediction model when the intent of the model is to generate positive effects on airline companies' bottomline. Critical information may be missing if the incorrect user is consuming our model.
- *Mitigation Strategy:* Ensure proper data governance is employed by the team and clearly identify who the target consumer of our model is.
- *Risk Level:* Low

Appendix

Essential Roles & Responsibilities:

All essential roles and responsibilities were agreed upon and confirmed by each member of MIS587: Business Applications in Machine Learning **Group 6**. Below you will find the fundamental job duties critical, imperative, and primary for the completion of the proposed project, Flight Status Prediction,

- **Recorder:** Takes notes summarizing team discussions and decisions and keeps all necessary records.
- **Reporter:** Serves as group spokesperson to the class or instructor
- **Timekeeper/ Checker:** Keeps the group aware of time constraints and deadlines and makes sure meetings start on time.
- **Business Analyst:** Defines project needs, assists in requirement gathering and hypothesis formulation, documents technical and qualitative.

- **Data Scientist:** Tests hypothesis, develops models, validates and develops code in collaboration with the team.
- **Project Facilitator:** Submits all project assignments and materials in accordance to project deadlines.

GitHub Repository Link:

https://github.com/krishnaShreedhar/mis587_project/blob/main/src/flight_delay_eda.ipynb

Team Rules:

Each team member is subject to abide by the following rules and act in accordance with the expectations agreed upon by MIS587: Business Applications in Machine Learning Group 6,

- Attend weekly meeting and in a timely manner
- Maintain communication throughout the week and update team in case of cancelations
- Complete assigned work in respect to due dates
- Be Attentive and collaborate during discussions and group meetings
- Be respectful to fellow teams during virtual discussions/ meetings
- Follow project guidelines and reference canvas guidance
- For decision making a majority vote will taken
- Weekly meetings will take place on Mondays at 6:00pm and Thursday at 6:00 pm(as needed)
- No weekend meetings
- Communications will take place via WhatsApp
- Meeting will be virtual via Zoom.

Project Work Schedule:

Task	Assigned to	Due Date
Business Memo (50 points):		
Business problem, project objectives, data-related questions and hypothesis.	**River, Ardavasd**	**30-Nov**
What are some data-related decisions (target variable, prediction type) you made.	**Shreedhar**	**30-Nov**

Project Report (100 points):		
Explain feature engineering techniques you used, justifications for changes, and outcomes for these changes.	**Andrew**	**7-Nov**
Explain model selection process, best models, and the model quality metrics.	**Everyone**	**7-Nov**
Explain areas where model struggles and areas for improvement(through more data, features, cases, etc..)	**Michael**	**14-Nov**
Explain most predictive features for modal building.	**Michael**	**14-Nov**
What feature types are most interesting to management(e.g., insights into the business problem and unknowns uncovered during the modeling process)	**Andrew**	**14-Nov**
Develop a concrete business recommendation supported by the best models from your work.	**River, Ardavasd**	**14-Nov**
Explain any business decisions to execute at various probability thresholds for the give target.	**Everyone**	**14-Nov**
Explain any actions the organization needs to take given these decisions and assertions on how implementation will change practice.(explain any baseline and profit matrices (payoff) related to the decisions and actions)	**Everyone**	**21-Nov**
Explain any assumptions you made due to uncertainty or lack of information.	**Andrew**	**21-Nov**
What is your final recommendation to implement the model or not?	**Everyone**	**21-Nov**
Project Presentation (50 points):		
PowerPoint	**Everyone**	**28-Nov**
Script/ Notes	**Everyone**	**28-Nov**
Final Project Report Deadline	**Everyone**	**11-Dec**