Flight Delays and Propagation Effects - Machine Learning and Network Analysis Approach

INTRO

It all started back in March, when coming home from vacation. We'd been hearing about a bunch of delays, so we weren't completely surprised to hear the announcement that our flight was delayed. But, we then got another delay notification and then a third. Next we heard that our flight was "disrupted". Then the final announcement informing us that our flight had been canceled.

I had never had a flight canceled before. It's the oddest feeling. We were not in a place we're familiar with. We felt abandoned. My thought was "They're supposed to get us home. What do we do?" No flights for two days. We ended up finding another airport to fly home from. This meant getting in line for a refund, booking a flight that cost 4 times as much, a Lyft for an hour drive to the other airport, a hotel for the night, a couple more meals, besides the additional food we purchased throughout the day while waiting due to the delays. It cost over $1,000 to take care of everything we needed due to the delay and we were lucky to find the flight out the next day. Many ended up having to wait two days. My daughter missed work the next day.

In the end, they said it was due to weather conditions but I wasn't satisfied with that. You could delay our flight for a total of 4.5 hours, without telling us why and then cancel our flight 5 hours before the delayed scheduled time due to a weather forecast? The poor weather wasn't at the airport we were departing from or heading to, it was somewhere in between. This led to the search for the airline data.

Wouldn't it be great to predict a flight cancellation or delay?

This sets off our departure to investigate an ideal scenario with the goal of placing more information in the hands of travelers. Our aim was to utilize machine learning and network analysis to craft a predictive framework anticipating disruptions. By supplying travelers with timely alerts, alternative routes, and contingency strategies, our goal was to empower them to make well-informed decisions amidst unexpected challenges. We sought to decode the factors underlying delays and cancellations, ensuring travelers possess insights to navigate potential disruptions effectively.

We wanted to improve the travel experience by offering passengers not only the foresight of potential disruptions but also the tools to navigate them proactively.

We went on a bumpy ride exploring the airline world and found many insights into delays. Even with a ton of data, not everything had an explanation. But we can say confidently, weather snafus often kick off a chain reaction of delays. So, we zeroed in on that with our limited time and data stash. As we sifted through our analysis, one idea kept popping up: the hub and spoke model. You'll hear about this model trend often in this blog. Oh, and speaking of trends, delays love to hang out in summer and party big at year-end. Blame warm weather and moisture for most weather delays. And when it comes to predicting delays? Think of it as trying to forecast a paper plane's flight path – it's all up in the air!

METHODS

Literature Review

We had a high-level understanding of what we wanted to do, but we needed to see how the academic community approached the flight-delay problem and the potential for delays to cause further delays. We reviewed multiple research papers serving as a critical foundation to our approach. These papers focused on predictive models for flight delays and utilizing network analysis to research propagation.
Similar to how disease networks show infection and recovery rates, our network analysis can show propagation, subsequent flights affected following a weather delay, in the same manner an infection spreads to others. Recovery is when an airplane is back on its scheduled route, similar to the spread of infection ending. We leveraged insights from flight delay prediction models, Bayesian network techniques, weather data considerations, and systemic air traffic delay propagation modeling. With those, we could develop a comprehensive and effective machine learning model and a network analysis approach addressing complex phenomena in flight delays and propagation. *(Please see the appendix for more detailed information regarding our literature review. Appendix A - Literature Review)*

Qualitative Inquiry

We had the experiences, motivations and reasons to investigate flight delays. We had the literature reviews to provide direction in building models. We lacked a deeper understanding of the Airline industry. Certainly delays have existed ever since the birth of the 'Jet Age' around 1945. But, what are the factors and reasons for modern-day delays after almost 80 years of commercial aviation? In order to provide this level of understanding,

we decided to conduct qualitative research using semi-structured interviews and develop an affinity diagram to help guide our efforts. We modified the KJ Method of qualitative interviews to fit our goals.

For this we turned to two subject matter experts (SMEs) in the industry: a data analyst for a major airline who has experience analyzing flight data and a retired airline pilot with 40 years of experience flying both domestic and international commercial flights. We scheduled 1 hour interviews with both keeping a central question in mind: What trends or patterns exist regarding the reliability of domestic commercial flights? We structured the interview questions in ways that we believed complimented their different roles and perspectives while maintaining enough flexibility for the participants to elaborate beyond our questions and share personal stories about their experience with a specific topic or question.

Data extraction and qualitative data analysis was performed with the information collected. We extracted statements, and paraphrased comments and insights from these interviews creating digital sticky notes. These sticky notes were then clustered into small groups of concentrated themes. As themes emerged from the small cluster of notes, we then attempted to form broader themes combining clusters. We kept creating higher levels of clusters (pairing multiple clusters) until we identified central themes between the two interviewees resulting in an Affinity Wall.

Through the affinity wall, we consolidated the notes into 3 insights:

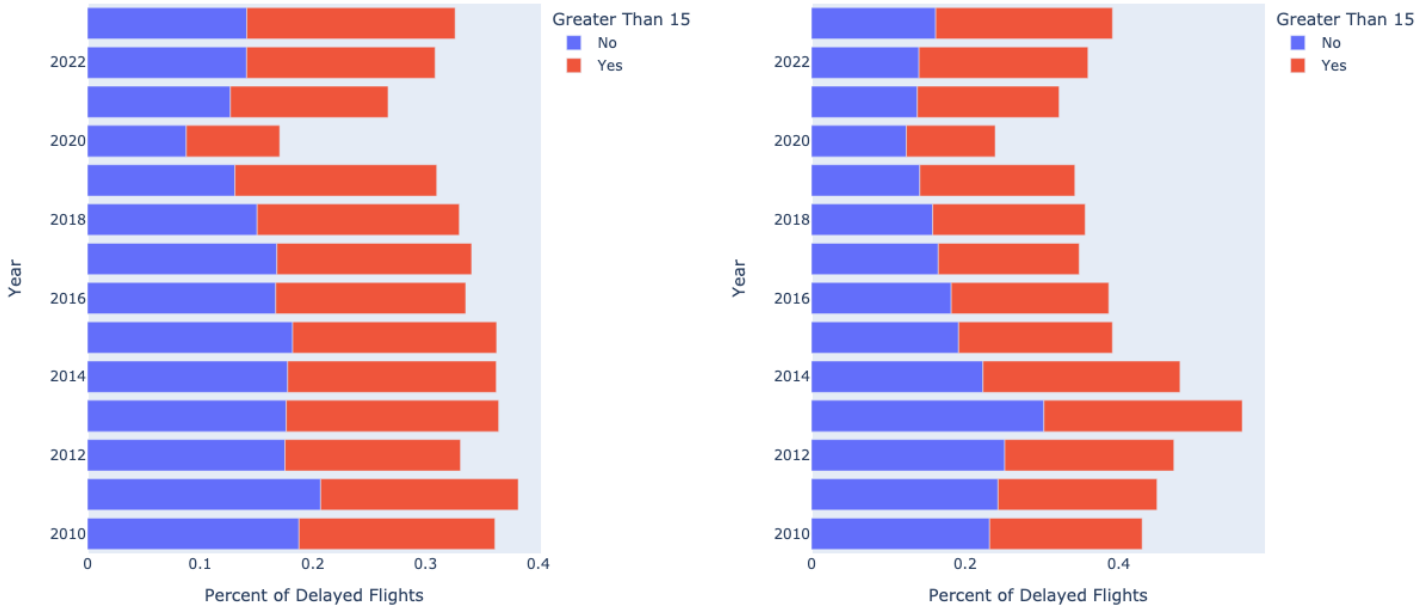
- **The complexities of the carrier-airport relationship**: Carriers and airports share a complex partnership that can potentially amplify delay issues. Carriers' decisions on aircraft fleet and storage locations influence delay potential and duration. Minimizing costs linked to grounded aircraft is a priority, driving investments in new planes and strategic plane placement. Partnerships between airlines help minimize ground time and maximize routes, impacting overall flight reliability and network effects.
- **The influence of the regulatory agencies (FAA/Air Control)**: Regulatory agencies emphasize safety, but their conservative risk approach can lead to delays due to weather and other factors. These regulatory frameworks, while enhancing safety, can sometimes introduce excessive risk avoidance and logistical complexity, resulting in delays. Air Traffic Control decisions, related to weather and air traffic, also contribute to delays.
- **Impact of weather**: Weather emerged as a significant challenge affecting both pilots' decision-making and air travel operations. Weather events, ranging from cyclical patterns to highly unpredictable conditions, can lead to delays, re-routings, or additional fuel consumption. The complexity of in-flight weather conditions alters the route a pilot may take, making accurate weather forecasting essential for proactive decision-making.

Data Collection and Exploration

Flight data

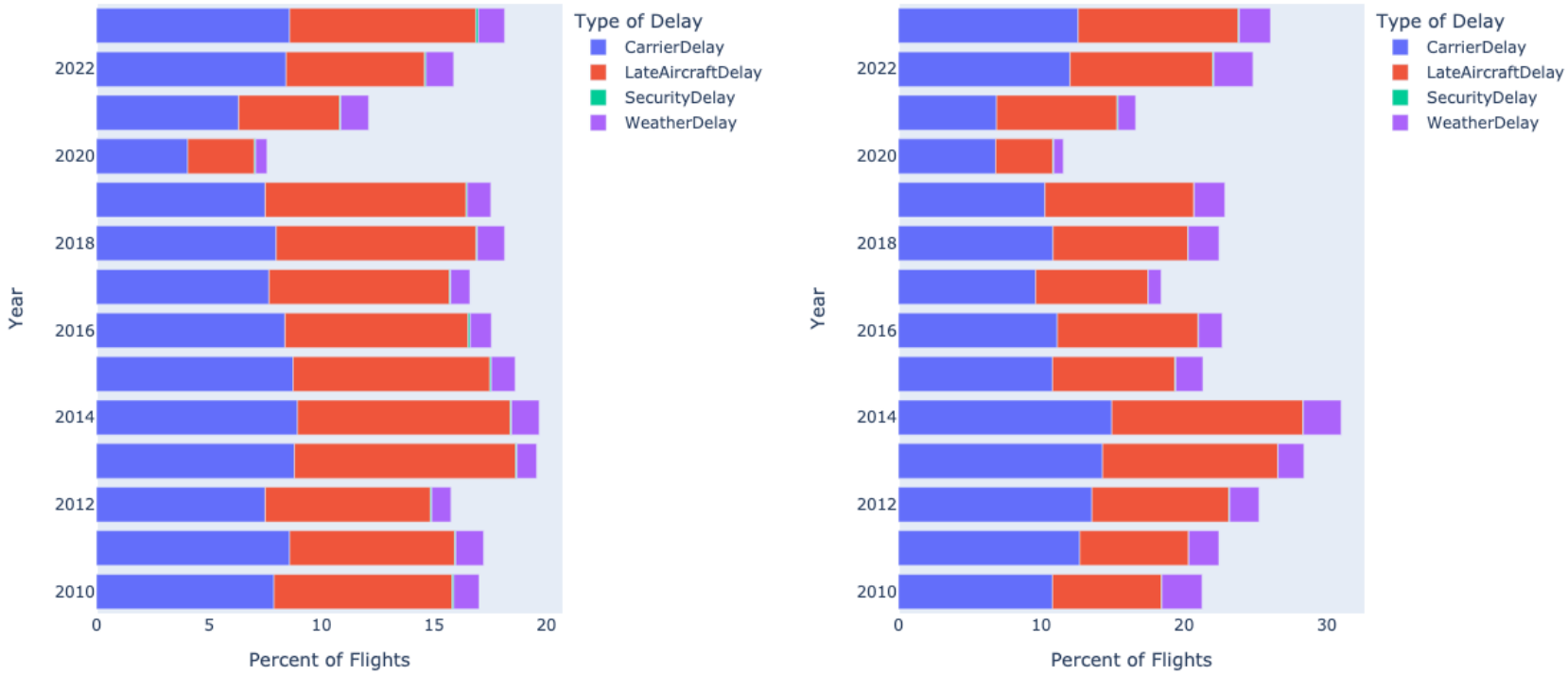
Having distilled our findings into these three illuminating insights, we moved from unraveling complexities to capturing real-world data. We conducted some early exploratory analysis with data downloaded from the Bureau of Transportation Statistics (BTS). They provide data called “OnTime” data. It can be downloaded on a monthly basis dating back to 1987. *(See Appendix C - Data Collection for further details.)* We downloaded data from 2010-2023. We looked at delays in general. About 30% of flights are delayed. We also knew we were going to want to narrow the data down to be able to provide more detail at an airport level. We decided to concentrate on one airport, encountering a representative amount of delays. We chose Dallas Fort Worth (DFW) after looking at frequencies for many airports. *(Appendix D - Delay Frequency By Airport)* In figure 1, you can see the frequencies of delays for all flights vs DFW only. Notice the bars are broken into two categories. A delayed flight is specifically designated as “delayed” if it is delayed for at least 15 minutes. In these bar charts you can see this specification narrows the data down further to approximately 15% of flights having delays 15 minutes or longer.

Figure 1: Left: Delays - 15 Minutes or More Vs Less than 15; Right: Delays - 15 Minutes or More Vs Less than 15



In figure 2, note a further breakdown of the delays by category. Here we see weather delays consist of approximately 1.5% of all flights. This is severely imbalanced data. Also note the late aircraft delays in these figures. We'll touch on these at a later point.

Figure 2: Percent of Flights with Delays of 15 Minutes or More by Type of Delay: Left is All Flights; Right is DFW Only



Aircraft

In addition to airline data, we gathered aircraft details via tail numbers from the Ontime data. In the US, the FAA assigns a unique registration number (tail number) to each plane. Most planes keep the same tail number throughout their service, though rarely it can change.

Studying aircraft characteristics helps understand their impact on delays. We used AeroBase Group, a government-approved aircraft supplier, to find model-level data using tail numbers. Though not detailed, this info aided our goals. With Beautiful Soup, we collected data including manufacturer, model, engines, type, weight class, and passenger capacity.

Merging datasets by tail number revealed an oversight: Ontime data had smaller aircraft (like Lear jets) and drones, noise for studying commercial flight delays. We filtered this by grouping flights based on manufacturer and average passenger capacity, focusing on flights with over 100 passengers. Our dataset included Boeing, Airbus, and defunct McDonnell Douglas planes (merged with Boeing in 1997).

Weather

Weather data was obtained from National Centers for Environmental Information (NCEI), available in tar.gz zipped files via the NOAA website. *(For complete weather details, refer to Appendix C - Data Collection / Weather.)* These files encompass global stations, including US stations beyond just those at airports. Unzipping all files would consume about 500GB. Each annual file, when unzipped, resulted in thousands of distinct files. File names used codes unrelated to flight data.

An NOAA Physical Scientist provided a cross reference to be able to link Station IDs to Country (*Appendix C*). By cross-referencing, US files were segregated to an annual US folder, others to a Non-US folder and then deleted to conserve space. The Station reference didn't offer an automated way to connect stations to airports. Later, the Master Location Identifier Database (MLID) from Weather Graphics (*Appendix C*) was found. This MLID reference associated WBAN numbers with a column matching to airport codes.

A code-driven approach was adopted for matching and renaming files. Code parsed file names, extracted WBAN numbers, and linked files to airport codes, adding this code as a column. Matched files were organized by year. Some airport codes lacked matches. Manual matching was attempted. However, some airport codes could not be matched. These flights were dropped from the OnTime data for further work. OnTime-weather files were merged creating annual OnTime-weather files. This narrowed down the number of files we had to work with but kept the current working file small enough to manage. We could finally move on to looking at the data inside the files.

Wow! There were over 170 columns in the weather data.

We realized this data contained more information than we needed for our machine learning models but we weren't weather experts and didn't know which columns to choose. The columns were named with codes as well. The "noaa global-hourly isd-format_document.pdf" (NOAA, 2018) provided the documentation needed to determine what each column contained and the meaning of the values. This still didn't dictate which columns to keep and which to drop. We used the 2019 weather data matched up to the flight data to create frequencies for each column. We created two sets of frequencies for each. One set was based on the frequencies for the data filtered on flights with delays and the other set was for data filtered to non-delays. Along with the frequency the percent matching to the filter was calculated. These two sets of frequencies were joined to be able to compare the frequencies of each value based on delay or not. Each column's frequencies were output to a tab in an Excel workbook for review. This is what the frequencies looked like for liquid precipitation quantity (column name: liq_precip_qty) and also for snow condition (column name: snow_cond) (Figure 3):

Figure 3: Weather features and Delay Frequencies

liq_precip_qty	No Delay	Delay	No Delay Pct	Delay Pct
0	2	0	0.00	0.00
1	64422	5354	76.12	76.79
3	7071	109	1.55	0.84
6	17429	197	2.80	2.08
12	1	0	0.00	0.00
24	9362	69	0.98	1.12
99	106	1	0.01	0.01

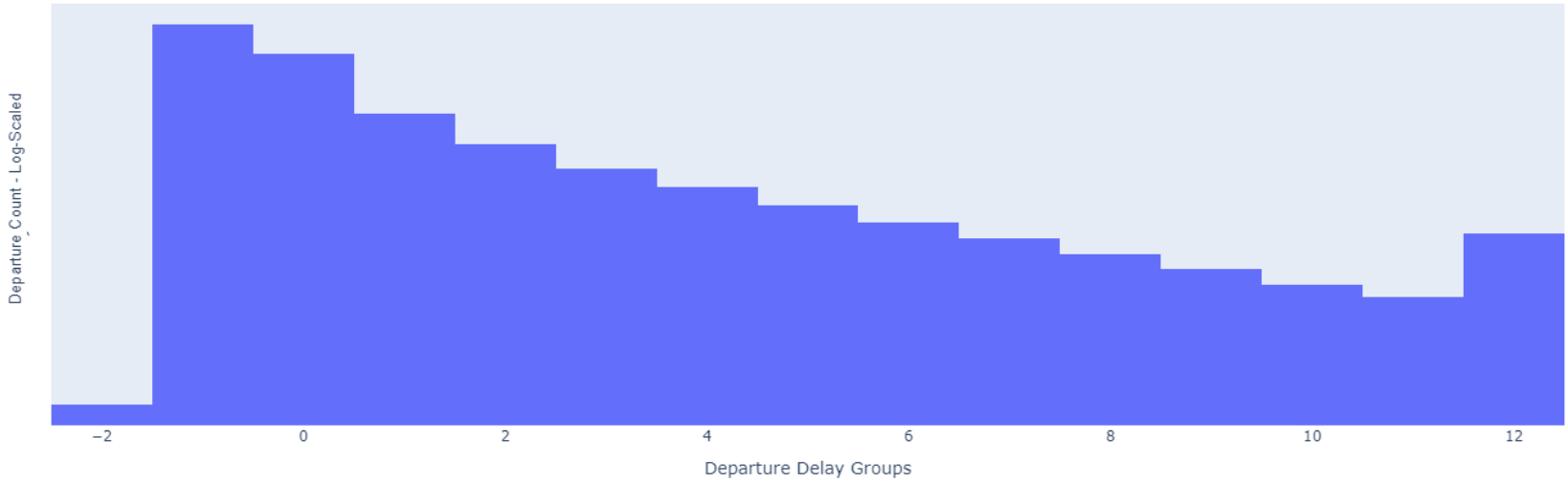
snow_cond	No Delay	Delay	No Delay Pct	Delay Pct
1	2	0	0.00	0.00
3	2859	54	0.77	0.34
9	2415	29	0.41	0.29

The frequencies do not contain missing data. The 99's and 9's in the frequencies represent missing values that were recorded as missing. There were many records with actual missing values. This did vary based on the column. For snow_cond you can see the data provided valid values for 0.77% of the non-delayed flights and only 0.34% of delays. Not only that, it's pretty much only one value, 3, and it doesn't provide any indication it would help predict a delay. Too sparsely populated. It was surprising to us the snow columns didn't provide anything valuable for our data. We dropped the snow columns. Notice the liq_precip_qty column provides values for close to 80% of the data. This is a well populated column. Looking at the percent of each value for delay vs non-delay still doesn't look like it provides any indication it would support a prediction. We reviewed all the columns in a similar manner keeping most columns that were well populated. None of the columns provided any indication they would be strong candidates for prediction based on the frequencies. We hoped the combination of the columns kept would be the key to the prediction.

Exploration

We conducted some exploratory data analysis on the combined dataset, providing valuable insights into the flight delay patterns. Among approximately 12.3 million flights, about 2.1 million experienced delays, with an average delay of 63 minutes and reaching up to 180 minutes, with outliers over 2 days. Interestingly, a significant portion of flights departed earlier than scheduled, with a gradual decline in frequency as the departure time gets closer to the scheduled time. A notable bump in frequency was observed in the 12th bin, which includes outliers with delays of 180 minutes or longer. (Figure 4).

Figure 4: The distribution of departure delay groups, ranging from flights departing earlier than scheduled to various degrees of delay.



When considering flights with delay information, nearly 17% of the total flights experienced departure delays, and approximately 18% had arrival delays. American Airlines Inc. (AA) had the highest percentage (29.4%) of total flights with delay information, followed by Delta Air Lines Inc. (DL) and JetBlue Airways (B6).

Moreover, focusing on frequently traveled flights, it was evident that Dallas Fort Worth International Airport served a disproportionately high number of flights with delay information, and an adjacency matrix highlighted several destinations, such as Miami, San Francisco, and Chicago, with a higher percentage of departure delays from Dallas Fort Worth.

Additionally, the analysis revealed that out of over 10.5 million flights originating from hubs, only 1.6 million were from non-hub airports. Departure delays affected around 15% of non-hub flights and over 18% of hub flights. Surprisingly, there was no significant increase in delays during November/December, as expected during the holiday travel season, but a noticeable spike was observed in Q1 travel.

Training Set Development

After integrating flight numbers and airport coordinates, we linked aircraft information smoothly. Amalgamating weather data involved assumptions, pinpointing weather measurements 45 minutes prior to departure to align with last-minute decisions. Training data spanned 2010-2019 (pre-COVID), and testing utilized 2023 Q1, excluding COVID years (2020-2022). Filtered for flights with at least 100 passengers, within the continental U.S., and matched weather stations.

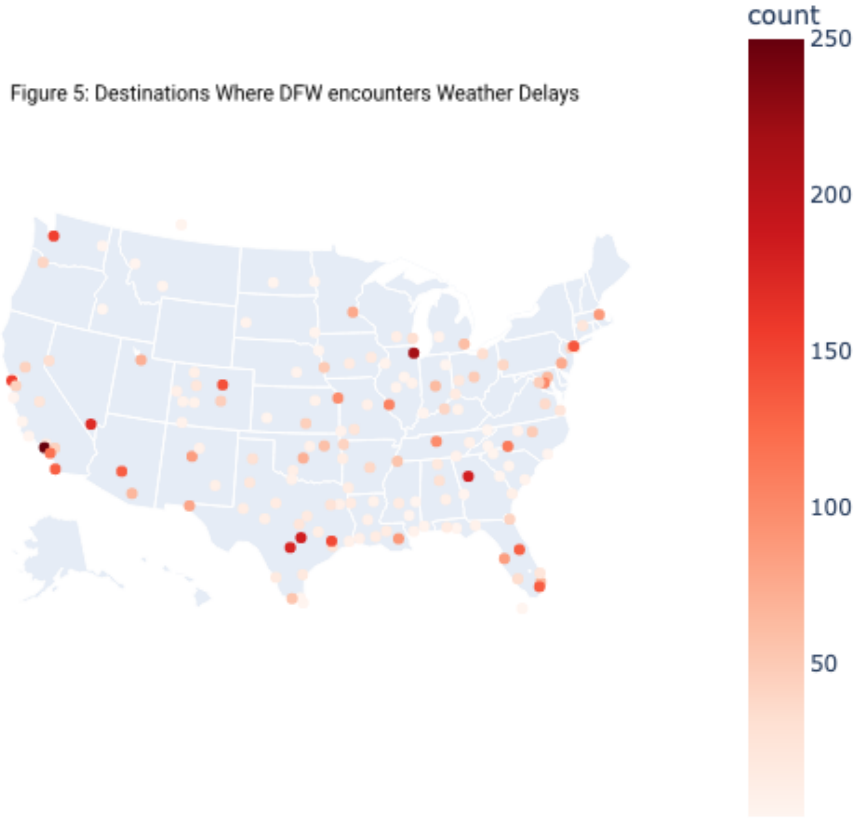
For machine learning models, training set columns were narrowed to prevent data leakage. Numeric data scaled for balanced impact. Categorical columns underwent one-hot encoding to avoid ordinal value assumptions. Additional model testing used a sample with up-sampled minority class (20%) and down-sampled majority class (80%).

Analysis and Modeling Approach

For our analysis we combined flight data, weather information and aircraft details. This combination of data allowed for modeling the carriers' operations, flight schedules, and airport partnerships we learned about up to this point. We'll provide details about each of these as we go.

Overall, qualitative interviews provided valuable insights that shaped our analysis and modeling process for predicting flight delays. Despite limited data sources, we focused on carriers' operations, flight schedules, and airport partnerships, understanding their impact on delays. Weather's significant influence on flight operations was also considered. Though lacking certain dimensions, we maintained realistic expectations and focused on crafting a predictive model to mutually benefit travelers and airlines. Figure 5 shows the pattern of weather delays at DFW based on their destinations. Figures 6 and 7 show the delay propagation densities to each subsequent connection.

Figure 5: Destinations Where DFW encounters Weather Delays



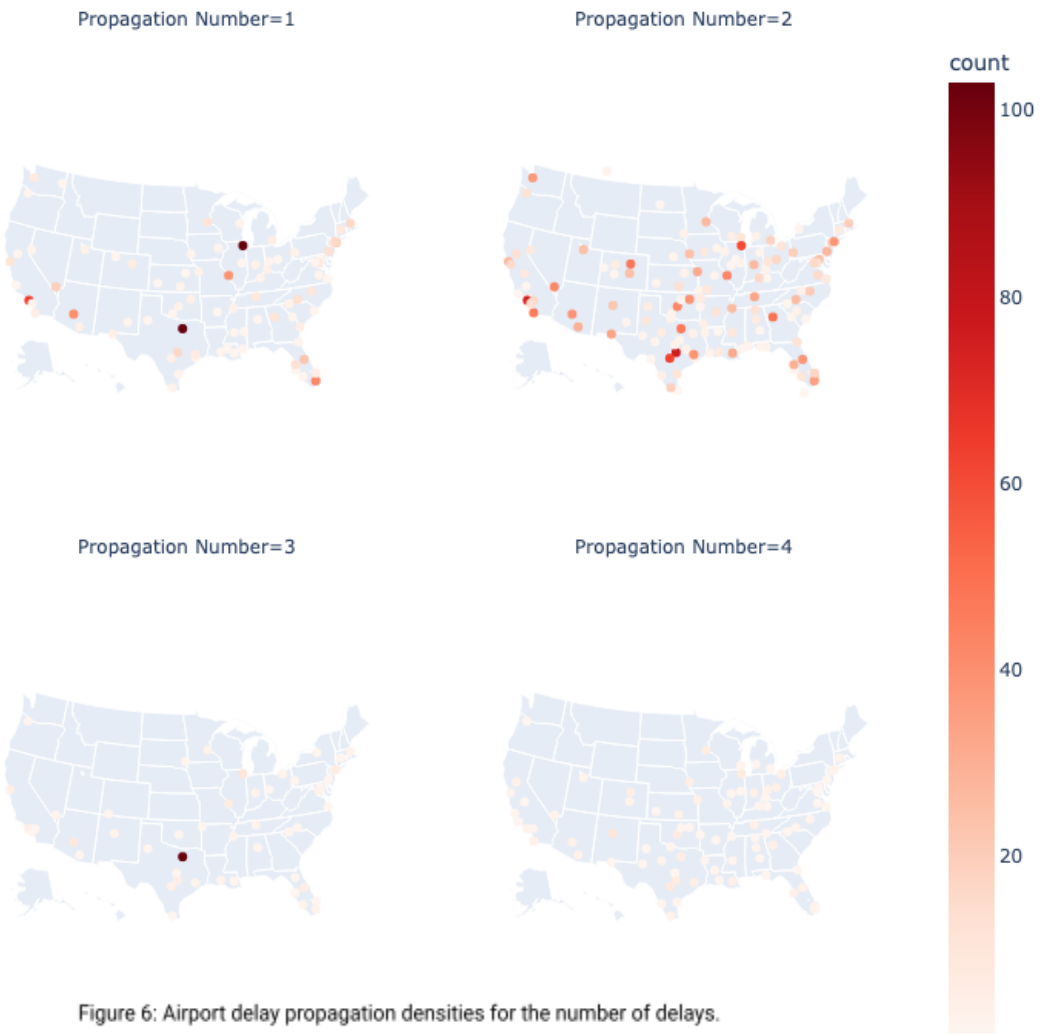


Figure 6: Airport delay propagation densities for the number of delays.

PREDICTIVE MODELING

The second leg of our journey entailed the creation of supervised machine learning classifiers to learn our dataset and predict which flights are weather-delayed. As our introductory story shared, delays can be a tremendous burden to travelers.

What if we could accurately predict weather delays based on our dataset? Could the model(s) benefit real-life travelers even if it was closer to the departure time?

To simplify our problem, we treated this task as a binary problem where a flight could either depart on schedule or delayed due to weather conditions,keeping in mind a delay is specified as at least 15 minutes behind schedule. This simplification also meant that we excluded other types of delays, such as aircraft issues, from our classification task. This was a necessary modification as we didn't obtain appropriate data to evaluate other types of delays.

Our earlier plans drew some inspiration from Li et al.'s 'A CNN-LSTM framework for flight delay prediction', where they incorporated a Long Short-Term Memory (LSTM) Deep Learning model using weather features to model relationships to flight delays. We had the idea that we could use an LSTM neural network to learn changes in weather over a period of time and provide the probability of inclement weather to a classifier, making the final determination of flight delay. However, we found the number of weather measurements we were able to extract prior to departure did not justify the complexity. This was later affirmed through a meeting with our project advisor. We decided it was best to simplify our approach and take a single set of measurements in a time window near the actual boarding of the plane and perform a comparison of linear, ensemble, and feed-forward neural network models.

We also had to think about the implications of our classification task. Despite our best intentions to help travelers, we could inadvertently introduce harm in ways we never intended. To help visualize this, we employ a commonly used tool in classification called a confusion matrix. We recreate the 2x2 matrix and relabel it as a "Harms Matrix" where the two purple quadrants correspond to potential harms falsely classifying either a weather delay (false positive) or on-time (false negative).

Figure 8: Harms matrix

The two green quadrants represent accurate predictions (non-delays and weather delays respectively) where no harm is identified. False positives, also referred to as Type 1 errors, are a major source of harm not only to the travelers but the airlines as well. Imagine the following scenario: A traveler uses this model and verifies it's flagged as a weather delay, cancels their flight and later finds out that the flight departed as scheduled. Not only did we harm our reputation, we also harmed the traveler, costing them time and money. We potentially cost the airline their profit margin from the ticket and created a vacant seat on the flight. Conversely, we can still cause harm if we falsely classify weather delays as being on-time. This creates frustration for travelers relying on our model to indicate delays as well as our trustworthiness as an information provider. This last scenario is akin to our opening story.

We needed to minimize the false positive rate to avoid the aforementioned scenario. Unfortunately, our data set was severely imbalanced as weather delays did not comprise a significant percentage of total flights (roughly 1.5% as seen in the visuals in the section about the OnTime data). We were concerned this imbalance would lead to a classifier only predicting non-weather delays (dominant class).

We couldn't rely on accuracy as a target metric. Choosing the dominant class for every sample would appear to be a very good classifier. As stated earlier, our motivation was to accurately predict positive weather delays to support travelers. Given the imbalance, recall was a possible choice to keep an eye on better scores for minimizing false negatives. However, we still had our earlier concern regarding harms to both travelers

and airlines. It seemed reasonable to leverage precision accounting for misclassification of the positive class but we could come at the expense of accurate nondelay predictions. Every metric seemed at best useful, none seemed ideal.

In the end, we found the weighted F1-score (aka harmonic mean) as a balance between precision (more quality predictions) and recall (the quantity of relevant predictions) an acceptable target metric. The F1 score takes on any value between 0 and 1 with 1 being optimal. As denoted in sklearn's F1 score documentation, the **weighted** variation of the F1 'calculate[s] metrics for each label, and find[s] their average weighted by support (the number of true instances for each label)'. We felt this was appropriate given the imbalance in labels. As we would later come to find, the weighted scores were much higher compared to their unweighted counterparts causing us to question whether this metric was creating a misperception of how well our models actually performed.

Logistic Regression

Logistic Regression (LR) is one of the most popular machine learning models for classification. Since we are working with a binary classification problem it makes sense to try logistic regression. It is easy to set up and is one of the easiest models to explain why it's producing the predictions it does. If you're familiar with the parameters of this algorithm it might interest you to know after many iterations with multiple variations of trial and error, our best model used the newton-cg solver with a balanced class weight. Additional parameters set were max iterations of 150 and C was set to 0.5. C is used to specify the strength of regularization, which is used to avoid overfitting. The training data providing the best results used the down-sampled majority with up-sampled minority dataframe.

Random Forest

Random Forest (RF) is also considered to be extremely popular. This one is also considered to be highly accurate since the idea behind it is to utilize several models and combine them providing the optimal result based on all, minimizing error. Feature importance scores can be obtained for these models making it fairly simple to analyze all columns comparing their importance. Again, for those interested in the best parameters determined, after many variations, our model used 50 trees, a maximum depth of 17, no bootstrapping, and a balanced class weight. Again, the training data providing the best results used the down-sampled majority with up-sampled minority dataframe.

Neural Network

We also developed a Neural Network (NN) Classifier using the Keras API. Initially, we built the model to be a deep learner with 3 hidden layers but found we achieved better performance and it was easier to train the model by simplifying the architecture to a single hidden layer. This layer contained 116 neurons, or 1 neuron per input feature with a relu activation function. The output layer utilized a sigmoid activation to produce the probability of weather/non-weather delay. All together, the model contained 13,456 parameters (ie the learned weights and bias). We tuned the model based on learning rate, drop out rate and batch size and leveraged an Adam optimizer which is a common and effective optimizer amongst deep learning and neural network models. We employed the KerasClassifier wrapper to provide some consistency with how we trained the other models.

The number of parameters combined with the various objects needed to support finding the optimal model made us feel like we were pilots in our own cockpit. Given our time constraints we were really just scratching the surface in terms of testing the model's levers and switches at our disposal.

Gradient Boosting Classifier

We developed a Gradient Boosting Classifier (GBC) model. It uses decision trees to efficiently explore complex delay patterns using decision trees, benefiting from their ensemble synergy for robust performance even with scarce data on rare delays. GBC's adaptability shines through targeted fine-tuning. Amplifying weights for delayed flights during training sharpens its focus on the minority class, heightening precision in spotting potential delays. Cross-validated grid searches refined GBC's performance, concentrating on precision and recall. Tuning parameters like learning rate, estimators, depth, features, split samples, and leaf samples tailored the model to our dataset's distinct nature. Sklearn's GBC implementation adeptly handles imbalanced data. With its ensemble approach, adaptability, and regularization, GBC proves a dependable model for predicting flight delays.

CLASSIFIER RESULTS

Our classifiers were trained on flight and weather data from 2010-2019, excluding pandemic years (2020-2022). We chose to avoid abnormal data and confirmed normalcy in early 2023. Our holdout set was the initial months of 2023 for model comparison.

The rarity of weather delays led to an imbalance during training, risking overfitting the dominant class. Initial training had low F1 scores due to this imbalance. To address this, we downsampled the non-delayed class by 20, boosting delayed class representation. Class weights were then applied, enhancing F1 scores, favoring correct predictions for weather delays, and adjusting for class imbalance.

$$\text{Class Weight} = (1 / N_s) * (N_{total} / N_c)$$

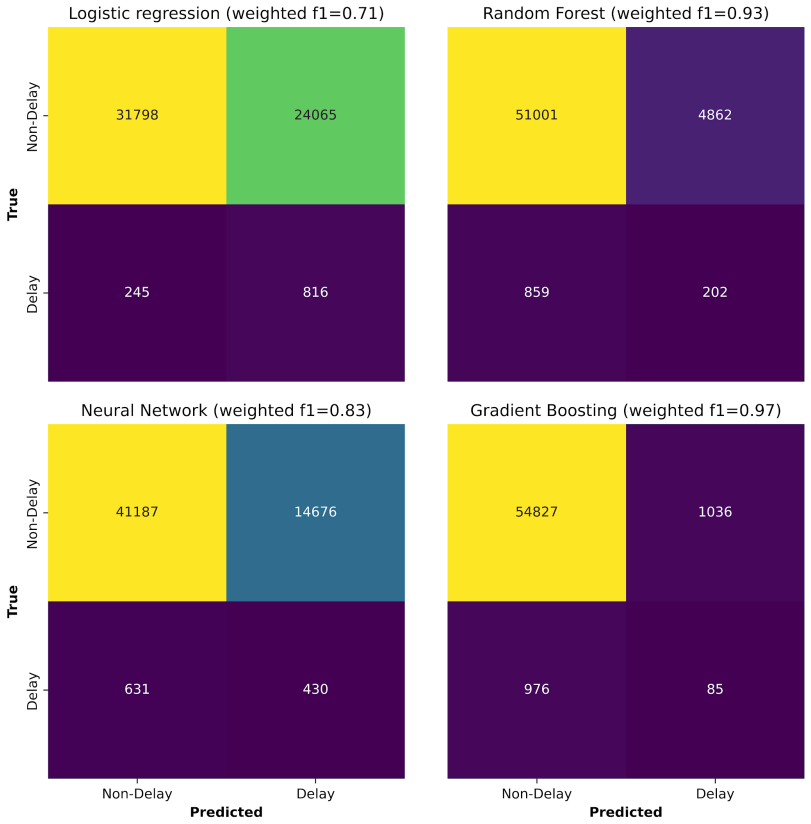
In the Confusion Matrix (Figure 9), the Gradient Boosting Classifier (GBC) achieved the highest weighted F1 score, accurately predicting the negative class but underestimating the positive class. Other models were affected by class weighting, aggressively predicting weather delays, yet sacrificing accuracy in opposite predictions. No model excelled in balanced predictions. Overall, the GBC achieved benchmark or near-benchmark scores amongst all of the classifiers (Figure 10).

We've learned growing up that two heads are better than one so why couldn't the same apply to machine learning? As another test, we attempted to ensemble the four models together, using the most frequent prediction across the models for each case, creating a 'majority rules' classifier. With these additional checks-and-balances, our ensemble of delay predictors were still not able to achieve the F1 bar established by the Gradient Boosting model, achieving a score of approximately .95 or .015 shy of the GBC model.

Weather-related Importance

Evaluating our model's performance is like flying blind – we need radar. So, we dived into the details to uncover the forces steering our models. The GBC assigned top priority to a variable tied to sky cover base height, as per NOAA's guidebook. This feature took center stage for GBC but seemed less important for other models. Across GBC, RF, and NN, dew point temperature and air temperature consistently topped the charts. Air temperature's impact on lift, key for achieving optimal altitude, raised concerns for higher temperatures, while lower temperatures might signal ice issues. Dew point temperature, a threshold for saturation, pointed to potential fog or ice complications.

Figure 9: Classifier performance confusion matrix



Figures 10: Performance Comparisons on Held-Out Test Set

models	accuracy	precision	recall	f1	auc
linear regression	0.572939	0.974	0.573	0.711	0.728
random forest	0.899498	0.966	0.899	0.930	0.509
neural network	0.731098	0.967	0.731	0.829	0.562
gradient boosting	0.964655	0.966	0.965	0.965	0.748

Interestingly, GBC's top 10 influential features consisted of 9 weather variables, compared to 8 for RF and LR, and a somewhat unexpected 4 for Neural Network (NN). Weather clearly weighed more than other categories like flight details. GBC's sole non-weather feature was scheduled departure hour, aligning with peak delay hours. Both GBC and RF shared several features, unsurprising given their shared ensemble lineage.

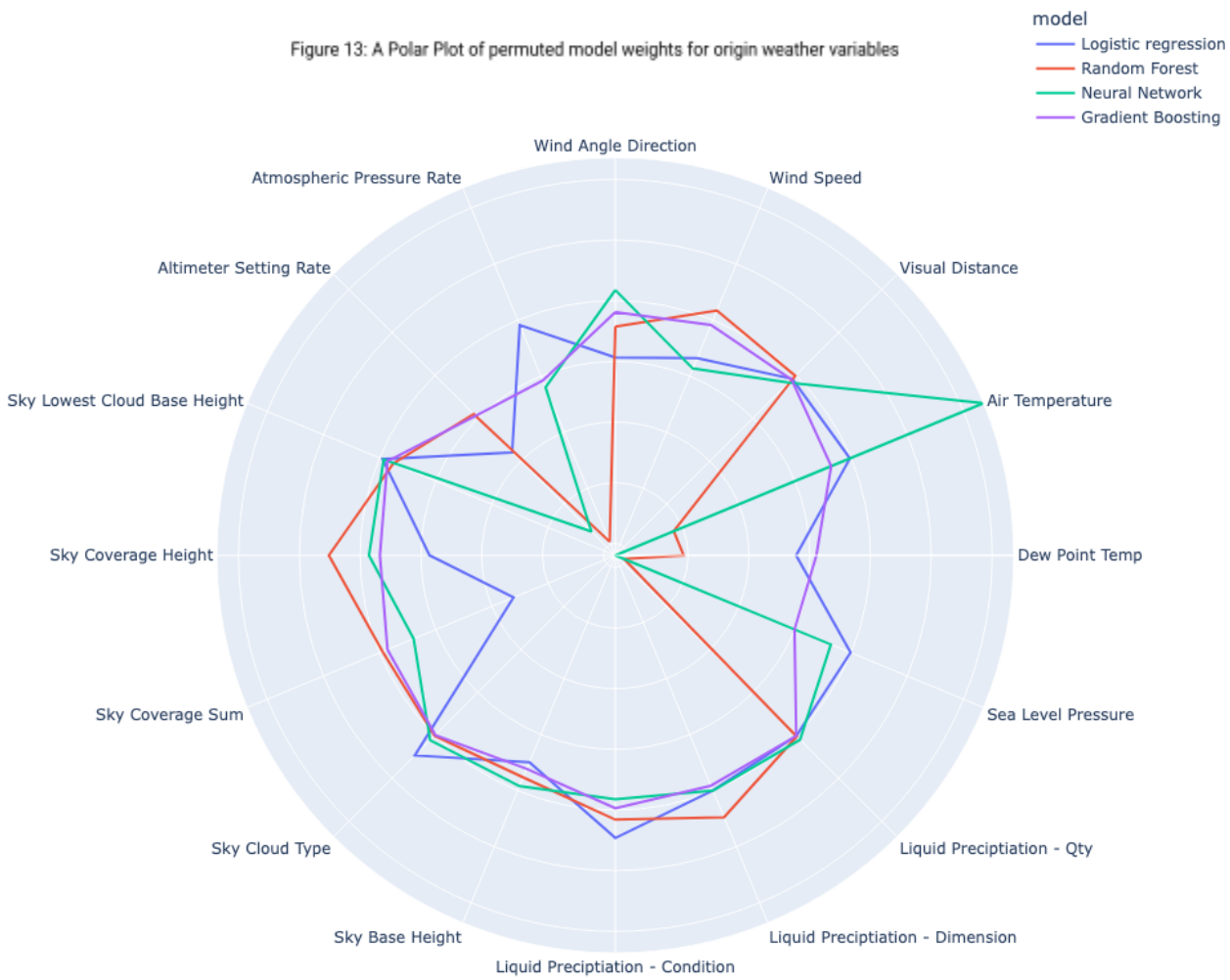
An aside on NN: its feature importance derived from shuffled values, highlighting time-based metrics (Day of Month, Month, Year, etc.) as critical contributors. Unlike other models, NN's weights stem from interactions with hidden layers' neurons.

Figure 11: A Comparison of Each Model's Feature Training Weights by Magnitude. Note: Neural Network's features were calculated using permutation importance. Negative weights indicate features associated with predicting the negative class.

Figure 12: Each Model's Most Important Features For Predicting on Held Out Dataset.

Next, we ran permutation importance for each model using the held-out test set. This measured what features were important to generalized classification. We then compared the results to the top training features. LR and RF shared 3 features in both top 10 lists, while NN shared 2. This suggested overfitting, leading to divergent training and test performance. A bar plot (Figure 11 and 12) shows the magnitude of weights for each model's top 10 permuted features. The Polar Plot (Figure 13) provides an easy comparison of how the models weighted the permuted features. The NN's larger weight on Air temperature raised red flags about overfitting. GBC didn't have agreement on top features for training and for

generalizing on the test set but its performance was more robust given its smaller, steadier weights compared to other models.



Finally, an error analysis of the GBC shed light on misclassifications. Pearson's correlation coefficient revealed the impact of liquid precipitation, primarily due to data flagging. Dew point temperature and atmospheric station pressure rate also showed notable correlations, despite being absent in permutation analysis. Misclassification cases are often correlated with high dew point values and related features, hinting at potential multicollinearity. This multicollinearity might contribute to models overfitting the training data. Addressing missing data and ensuring appropriate dew point sampling could help mitigate both multicollinearity and resulting misclassifications.

NETWORK ANALYSIS AND MODELING

For this model, we mimicked a model based on a disease network. We analyzed delays, restricted to the continental U.S, with weather delays beginning in Dallas Fort Worth. To capture delay propagation, we normalized local departure times and grouped data by tail numbers. We constructed a data processing loop covering 2010-2019 and 2023 for every single day in this timeframe, for each tail number. On average, delays propagated around 55.78 percent out of all subsequent flights on a given day for the affected tail numbers. This indicates frequent transmission to successive flight segments.

Centrality Analysis showcased major airport hubs, like Chicago O'Hare, Hartsfield-Jackson Atlanta, and Phoenix Sky Harbor, standing out due to high traffic, strategic connections, and network influence. These hubs significantly shaped air travel dynamics.

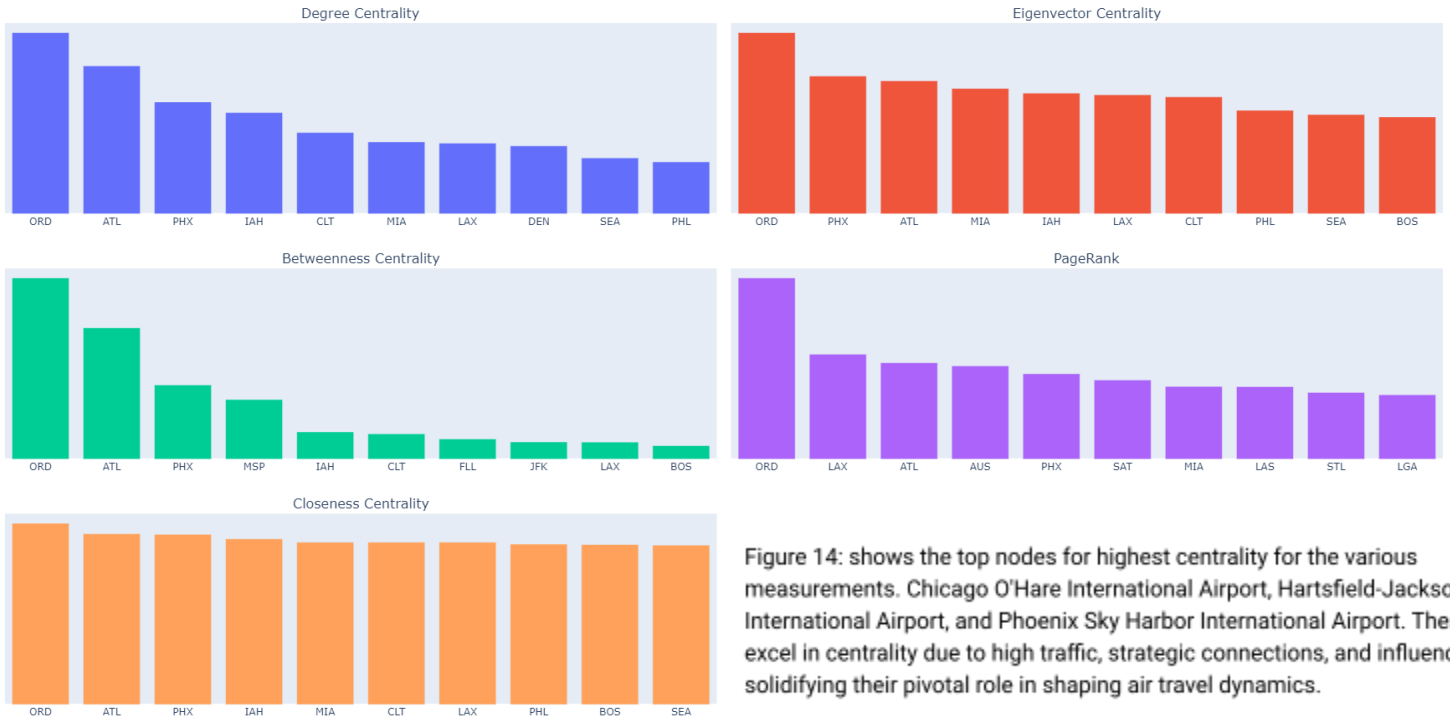


Figure 14: shows the top nodes for highest centrality for the various measurements. Chicago O'Hare International Airport, Hartsfield-Jackson Atlanta International Airport, and Phoenix Sky Harbor International Airport. These hubs excel in centrality due to high traffic, strategic connections, and influence, solidifying their pivotal role in shaping air travel dynamics.

Identifying delay bottlenecks was achieved using PageRank, another type of centrality measure, with delay-influenced edge weights. The edge weights were calculated based on the percentage of the route out of all flights. This revealed airports causing bottlenecks, influencing the entire network's delay propagation. Noteworthy bottlenecks included Phoenix Sky Harbor International Airport, Hartsfield-Jackson Atlanta International Airport, and Chicago O'Hare International Airport.

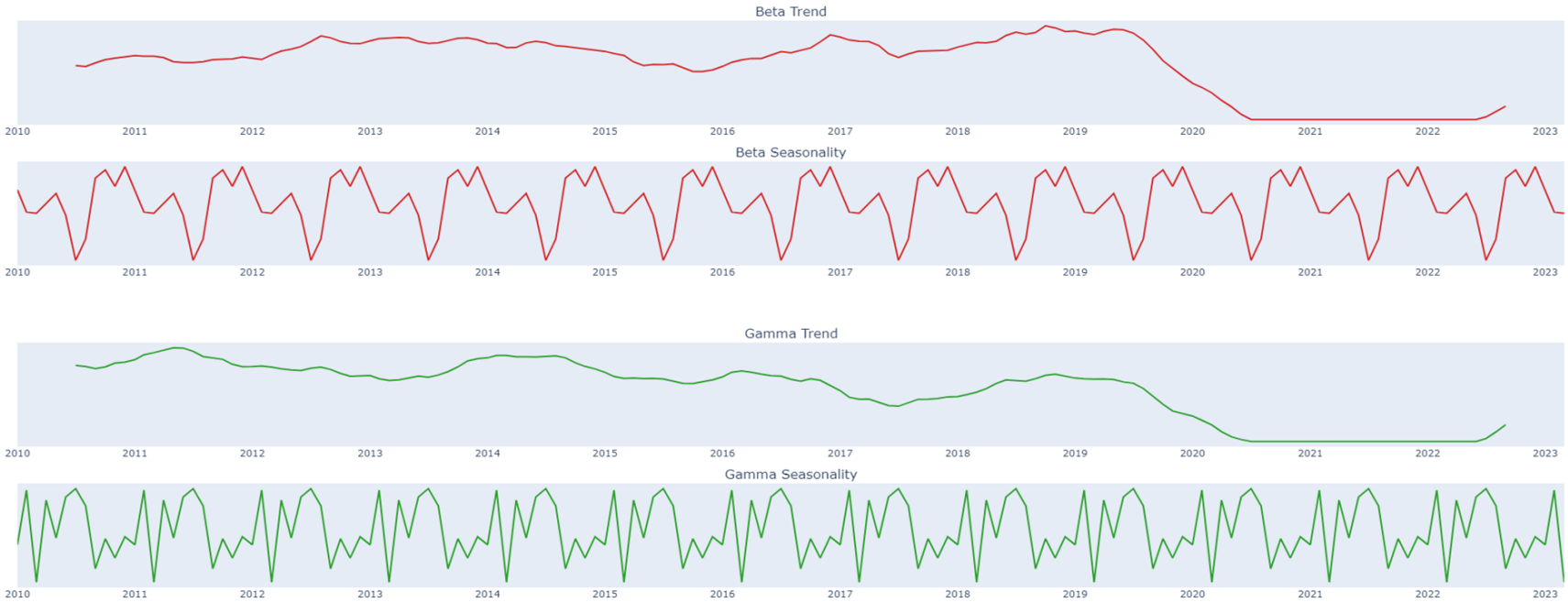


Figure 15 depicts the evolution of Beta and Gamma rates over time. Preceding the COVID pandemic, the Beta rate displays a relatively stable pattern, showing a slight upward trend. Concurrently, the seasonality of Beta maintains a predominantly consistent profile, with a subtle declining tendency. Notably, the Beta seasonality demonstrates an upward shift around the mid-year mark, reaching its zenith towards the conclusion of each year, and subsequently recurring. In contrast, the Gamma seasonality exhibits more pronounced fluctuations throughout the year. It becomes most pronounced around the midpoint of the year, subsequently diminishing during the latter half, and subsequently starting to ascend once again.Figure # depicts the evolution of Beta and Gamma rates over time. Preceding the COVID pandemic, the Beta rate displays a relatively stable pattern, showing a slight upward trend. Concurrently, the seasonality of Beta maintains a predominantly consistent profile, with a subtle declining tendency. Notably, the Beta seasonality demonstrates an upward shift around the mid-year mark, reaching its zenith towards the conclusion of each year, and subsequently recurring. In contrast, the Gamma seasonality exhibits more pronounced fluctuations throughout the year. It becomes most pronounced around the midpoint of the year, subsequently diminishing during the latter half, and subsequently starting to ascend once again.

We calculated beta (infected/delayed) and gamma (recovered) rates over time. This involved tallying delay types, optimizing parameters, and applying the SIRS model (*See Appendix F Network Modeling)*. The resultant rates of change for susceptible, infected, and recovered entities provided insights into infection and recovery rates, aligning with our objectives.

Seasonal decomposition provided insights into infection trends and seasonal patterns. Beta trend remained generally stable with a slow upward drift over time, interspersed with occasional spikes, e.g., in 2012 and 2018. Prolonged uptrends, e.g., 2015-2016 and 2020, indicated sustained higher infection rates. Consistent beta seasonality revealed cyclic variations, particularly around July-August and November-January, possibly due to weather and behavior influences.

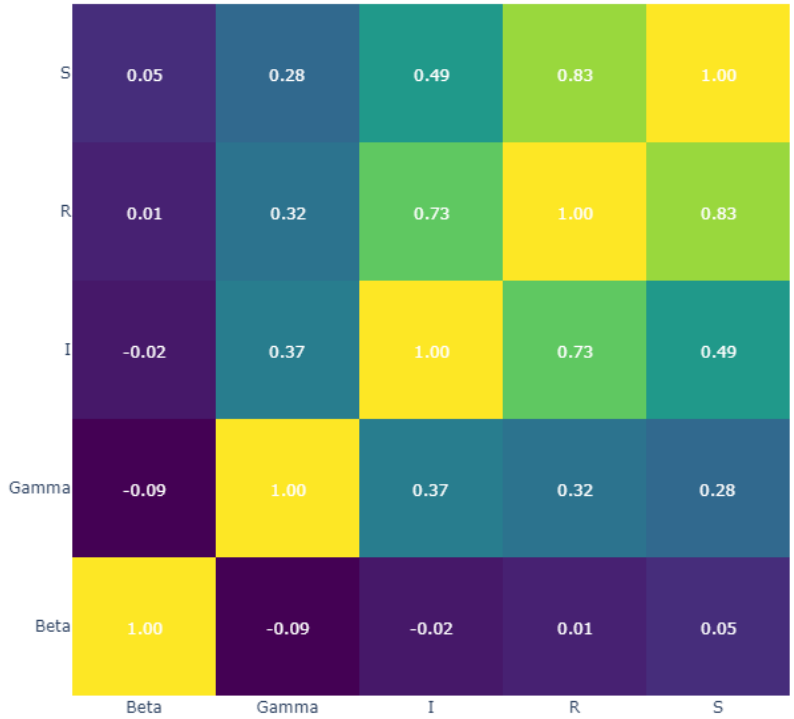
In contrast, gamma exhibited consistent recovery rates with a slight decline, implying gradual recovery deterioration. Notably, winter saw simultaneous increases in beta and gamma, suggesting higher delay propagation and better overall response to delays.

Understanding beta and gamma seasonal patterns benefits travelers as well. If a traveler feels a possible delay will be detrimental to their plans they can opt for a different flight still in their schedule. If the airlines utilized this information they could communicate the possibilities ahead of time allowing the travelers to modify their plans.

To validate our SIRS model, we conducted a correlation analysis between calculated beta and gamma, and the counts of infected (I), recovered (R), and susceptible (S) individuals. The coefficients reveal a mild linear relationship between beta and gamma rates and 'I', 'R', and 'S' counts. These correlations imply that fluctuations in these counts might not be heavily guided by infection and recovery rates. However, this doesn't undermine the beta/gamma significance in representing SIRS counts over time. Multiple factors could have influenced these rates:

- Simplification of Real-World Complexity: The model's analogy of flight delays as infections oversimplifies intricate air travel dynamics, potentially missing critical factors that affect delay propagation.
- Neglecting External Influences: The model doesn't incorporate external factors like weather, air traffic control decisions, and operational practices, which significantly impact delay cascades and may lead to deviations from real-world patterns.

Figure16: Correlation Analysis between Infection Rates (Beta) and Recovery Rate (Gamma) with SIR Counts.



- Limited Data Representation: The linear approach of the model may not effectively capture the multifaceted interactions and complexities involved in delay propagation, leading to potential inaccuracies in rate estimations.

While the correlation analysis suggests a subtler link between parameters and counts, it's essential to contextualize these findings within the model's specifications, data quality, and the multitude of factors influencing infection dynamics.

We adopted an infection model inspired by disease transmission to examine flight delay propagation within the continental U.S., centered around Dallas Fort Worth. Centrality Analysis highlights major airport hubs like Chicago O'Hare, Hartsfield-Jackson Atlanta, and Phoenix Sky Harbor as influential due to their extensive connections. Delay bottlenecks are identified using PageRank, unveiling key airports causing network-wide delays. By incorporating the SIRS model to analyze infection and recovery rates, offering insights into delay trends and seasonal patterns. The beta (infected/delayed) and gamma (recovery) rates provide valuable information for travelers to adapt plans, while airlines can enhance operational awareness and communication. Despite limitations like oversimplification of air travel dynamics and neglect of external influences, the findings offer a nuanced understanding of delay propagation dynamics in airline operations.

BROADER IMPACTS

INSIGHTS FROM THE ANALYSIS

In our analysis, we uncovered key insights that shed light on the intricate world of flight delays. Weather variables have emerged as powerful predictors, with sky cover base height, air temperature, and dew point temperature playing significant roles across different models. These findings underscore the impact of weather conditions on flight delays – adverse weather can set off a domino effect of disruptions.
What's fascinating is how different models view these weather features.

The Gradient Boosting Classifier (GBC) and Random Forest (RF) models, designed for ensemble learning, give more weight to weather factors. This alignment underscores the critical role that weather plays in understanding and anticipating flight delays.

Our analysis delves deeper into the flight network's intricate web, revealing a ripple effect. Using pagerank with delay-influenced edge weights, we pinpoint bottleneck airports—often hub airports—that trigger network-wide delays. Understanding this dynamic is key for airlines to manage disruptions, enhance operational resilience, and offer a smoother travel experience.

This insight uncovers the heartbeat of aviation, where delays act as signals rippling through the network's veins. Airlines can strategically address these delays at bottleneck hubs, minimizing disruptions and improving reliability for passengers. In a world of aviation complexities, our analysis underscores the importance of data-driven insights in navigating the skies with fewer interruptions.

PRACTICAL IMPLICATIONS

Given the timeframe for our project we weren’t able to create the app we intended. Being that our results weren’t as reliable as we would have liked, we worked hard and spent more time trying to improve our models the best we could with the time given as opposed to creating an app that would provide poor predictions. As mentioned earlier when starting to discuss our models, we thought about how our results could affect travelers, airlines, and ourselves. There *is* harm in providing poor predictions. Let's think back to where all of this started to get a feel for how this may have affected the flight plans back in March. Let's say I used the app when planning our spring break trip. I avoided taking a later flight on that same day because the app said it was likely to be delayed. In this scenario everyone is harmed. We got a bad reputation, we know from our story how I fared (time, money, anger, frustration). The airline was harmed due to the extreme circumstances here because I don’t want to fly with them again but what about the flight I avoided taking? Was it a different airline? The decision to take this flight may have cost a different airline the sale of the tickets and may have kept them from filling two seats on their plane.

If we felt good about the reliability of our predictions, the intended app could have helped the traveler make better plans. It could also help the airlines with scheduling. If the airlines could predict which flights are in jeopardy they could adjust their schedules accordingly. They don't want the delays and trouble just as much, if not more, than the passengers. They don't want to have to deal with all of the delay propagations and all of the unhappy customers from the original delay and the propagated delays. This costs them employees additional hours worked, as well as the extra expense for additional gate time at the airport. They know this is not good for their business too. The airlines could benefit by adjusting their schedules and the airports might be able to schedule gate availability for the airlines to better accommodate them.

LIMITATIONS

As we reflect on our journey, we recognize the potential for further advancements and future directions, especially in addressing the limitations of our data and approaches. Navigating the intricacies of weather data was a substantial challenge, demanding meticulous data preparation, thoughtful column selection, and handling of sparse data. To enhance our models' accuracy, a deeper dive into feature collinearity could mitigate overfitting, lending more robustness to our classifier.

Expanding the weather data sources beyond our initial scope could yield additional insights, bolstering the predictive power of our models. Additionally, exploring the impact of advanced technology-equipped aircraft capable of flying without visibility could open new avenues for enhancing predictions. By identifying and integrating these aircraft characteristics, we might unlock heightened model effectiveness. Furthermore, it's imperative to acknowledge potential constraints, such as the limited availability of flights for analysis within our dataset..

FUTURE DIRECTIONS

We could consider ideas beyond our current outlook. Some things may, or may not be possible. The weather turns out to be a very difficult thing to predict due to the many things the airlines, and pilots, can do to work around it but what about predicting the other types of delays? We learned about restrictions on airline staff working hours. This isn't limited to the pilots, other staff hours are limited as well. We could look at how much this affects flight schedules and how much delays actually affect their hours. What about the airport staff, beyond the airlines, those staff that keep the airport running. The staff out on the tarmac. The air traffic controllers. Can we obtain data that would assist in studying these areas? This would be something to look deeper into. We aren't aware of any of this data and not confident it's available. Would there be any kind of data that could help make inferences? Can we track congestion at the airport, would this provide anything beneficial? What about making what we have better?

REFERENCES

Wikipedia contributors. (2023). Post-war aviation. Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/wiki/Post-war_aviation

Kawakita, J. (2023, August 10). KJ Method. Lucid Meetings. Retrieved from https://www.lucidmeetings.com/glossary/kj-technique

Claiborne, Matt. "Airplane Tail Numbers Explained." *Aerocorner*, 29 Jan. 2022, aerocorner.com/blog/airplane-tail-numbers-explained/. Accessed 10 Aug. 2023.

Page, Charlie. "Can the Weather Get Too Hot to Fly a Plane Safely? A Pilot Explains." *The Points Guy*, 18 Jul. 2023, thepointsguy.com/news/can-it-get-too-hot-to-fly/. Accessed 10 Aug. 2023.

Martin, Swayne. "How Ice Affects Your Wings, And Why It Leads To An Early Stall." *Boldmethod*, 29 Jan. 2022, www.boldmethod.com/learn-to-fly/aerodynamics/how-aircraft-icing-affects-your-wing-and-leads-to-an-early-stall/. Accessed 10 Aug. 2023.

https://www.weather.gov/source/zhu/ZHU_Training_Page/thunderstorm_stuff/Thunderstorms/thunderstorms.htm#:~:text=Severe%20thunderstorms%20are%20more%20likely,as%20compared%20to%20relatively%20high

Li, Qiang, et al. "A CNN-LSTM Framework for Flight Delay Prediction." *Expert Systems with Applications*, vol. 227, 2023, https://doi.org/10.1016/j.eswa.2023.120287. Accessed 1 Jun. 2023.

NOAA - National Centers for Environmental Information, 'Federal Climate Complex Data Documentation for Integrated Surface Data (ISD)' Asheville, NC, 12 Jan 2018, https://www.ncei.noaa.gov/data/global-hourly/doc/isd-format-document.pdf Accessed 01 Jun 2023.

Shao, C., & Prabowo, R. (2019). Flight delay prediction using airport situational awareness map. arXiv preprint arXiv:1911.01605.

Rodriquez-Sanz, Alvaro, et al. "Assessment of Airport Arrival Congestion and Delay: Prediction and Reliability." *Transportation Research*, vol. 98, 2021, pp. 255-283, https://doi.org/10.1016/j.trc.2018.11.015. Accessed 1 Jun. 2023.

Sternberg, Alice. "A Review on Flight Delay Prediction." *Transportation Reviews*, vol. 41, no. 4, 2021, pp. 499-528, https://doi.org/10.1080/01441647.2020.1861123. Accessed 1 Jun. 2023.

“Sklearn.Metrics.F1_Score.” *Scikit-learn*, scikit-learn/stable/modules/generated/sklearn.metrics.f1_score.html.

Li, Shanmei. "Data-Driven Modeling of Systemic Air Traffic Delay Propagation: An Epidemic Model Approach." *Journal of Advanced Transportation*, vol. 2020, 2020, https://doi.org/10.1155/2020/8816615. Accessed 1 Jun. 2023.

Appendix

Appendix A - Literature review

In the pursuit of developing a comprehensive understanding of flight delays and propagation effects,we explored key research papers that offer valuable insights into flight delay prediction: Bayesian network approaches, weather data considerations, and systemic air traffic delay propagation.

Wei Shao et al. in "Flight Delay Prediction using Airport Situational Awareness Map" presented a real-time flight delay prediction model focused on delays within four hours. While their findings showed promising accuracy for short-term predictions, it was noted that weather data did not enhance the predictions and, in fact, led to decreased accuracy. As our project aims to predict flight delays further out, it is essential to acknowledge that the predictive accuracy may not match what was observed in this study.

Rodriguez-Sanz et al., in "Assessment of airport arrival congestion and delay: Prediction and reliability," utilized a Bayesian Network approach, considering multiple factors such as airport saturation, arrival rates, time of day, and weather conditions to showcase interdependencies influencing airport performance. Furthermore, they employed a Markov chain approach to assess system reliability. Drawing inspiration from this research, our project could explore network data and place greater emphasis on airport-related factors in predicting delays and cancellations.

Sternberg et al.'s "A Review on Flight Delay Prediction" provided a comprehensive overview of various studies, methodologies, data types, and sources related to flight delay prediction. The discussion surrounding weather data, its various components, and potential sources holds particular interest for our project. This review will guide us in selecting suitable weather variables and sources to align with our research goals.

In "Data-Driven Modeling of Systemic Air Traffic Delay Propagation: An Epidemic Model Approach" by Li, Xie, Zhang, and Bai, a novel modeling approach based on the epidemic model was introduced to understand air traffic delay propagation at a system level. The research highlights the significance of studying airport delay from the perspective of propagation and presents an integrated airport-based Susceptible-Infected-Recovered-Susceptible (ASIRS) epidemic model to simulate delay propagation in airport networks. The model offers promising predictive capabilities for both short-term and long-term temporal and spatial evolution of air traffic delay. Incorporating this research into our project will aid in creating a robust simulator to predict flight delay propagation effects within airport networks.

Appendix B - Data Dictionary (*Page 1 of 8*)

Descriptive Name	*Column Name	Description	Values
Altimeter Setting Rate	at_pres_altimeter_rate at_pres_altimeter_rate_d	The pressure value to which an aircraft altimeter is set so that it will indicate the altitude relative to mean sea level of an aircraft on the ground at the location for which the value was determined.	MIN: 08635 MAX: 10904 UNITS: Hectopascals SCALING FACTOR: 10 MISSING: Average of Values
Atmospheric Pressure Rate	at_pres_stn_rate at_pres_stn_rate_d	The atmospheric pressure at the observation point.	MIN: 04500 MAX: 10900 UNITS: Hectopascals SCALING FACTOR: 10 MISSING: Average of Values
Scheduled Departure Hour	CRSDepHour	The hour of the day the flight is scheduled to depart	MIN: 0 MAX: 23 No MISSING
Scheduled Elapsed Time	CRSElapsedTime	The scheduled elapsed time of the flight, in minutes	MIN: 45 MAX: 556 No MISSING
Day of Month	DayofMonth	The day of the month	MIN: 1 MAX: 31 No MISSING
Day of Week	DayOfWeek	The day of the week	MIN: 1 MAX: 7 No MISSING
Destination Airport	Dest	Destination airport.	MIN: 0 MAX: 999 No MISSING

Descriptive Name	*Column Name	Description	Values
Sequence Number (Dest)	DestAirportSeqID	Destination Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.	MIN: 1014001 MAX: 1591905 No MISSING
Destination State FIPS	DestStateFips	Destination Airport, State Fips - code identifying state	MIN: 1 MAX: 72 No MISSING
Flight Distance	Distance	Distance between airports (miles)	MIN: 175 MAX: 3847 No MISSING
Reciprocating Engine	engine_type_Reciprocating	**The model of aircraft uses a Reciprocating engine	1 = Reciprocating Engine
Turbofan Engine	engine_type_Turbofan	**The model of aircraft uses a Turbofan engine	1 = Turbofan Engine
Turbojet Engine	engine_type_Turbojet	**The model of aircraft uses a Turbojet engine	1 = Turbojet Engine
Liquid Precipitation - Condition	liq_precip_cond liq_precip_cond_d	The code that denotes whether a liquid precipitation depth dimension was a trace value.	1 = Measurement impossible or inaccurate 2 = Trace 3 = Begin accumulated period (precipitation amount missing until end of accumulated period) 4 = End accumulated period 5 = Begin deleted period (precipitation amount missing due to data problem) 6 = End deleted period 7 = Begin missing period 8 = End missing period 9 = Missing

Descriptive Name	*Column Name	Description	Values
Liquid Precipitation - Dimension	liq_precip_dim liq_precip_dim_d	The depth of liquid precipitation that is measured at the time of an observation.	MIN: 0000 MAX: 9998 UNITS: Millimeters SCALING FACTOR: 10 MISSING: Average of Values
Liquid Precipitation - Quantity	liq_precip_qty liq_precip_qty_d	The quantity of time over which the liquid precipitation was measured.	MIN: 00 MAX: 98 UNITS: Hours SCALING FACTOR: 1 MISSING: Average of Values
Airbus Aircraft	mfr_AIRBUS	**The model of aircaft was manufactured by Airbus	1 = Airbus Aircraft
Boeing Aircraft	mfr_BOEING	**The model of aircraft was manufactured by Boeing	1 = Boeing Aircraft
Number of Engines	no_engines	The number of engines	MIN: 2 MAX: 4 No MISSING
Sequence Number (Origin)	OriginAirportSeqID	Origin Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time.	MIN: 1129802 MAX: 1129806 No MISSING NOTE: JUST Origin Airport
Passenger Capacity	passengers	Capacity for passengers.	MIN: 102 MAX: 563 No MISSING
American Airlines	Reporting_Airline_AA	**American Airlines	1 = American Airlines
Alaska Airlines	Reporting_Airline_AS	**Alaska Airlines	1 = Alaska Airlines
JetBlue Airways	Reporting_Airline_B6	**JetBlue Airways	1 = JetBlue Airways

Descriptive Name	*Column Name	Description	Values
Continental Airlines	Reporting_Airline_CO	**Continental Airlines	1 = Continental Airlines
Delta Airlines	Reporting_Airline_DL	**Delta Airlines	1 = Delta Airlines
Frontier Airlines	Reporting_Airline_F9	**Frontier Airlines	1 = Frontier Airlines
Airtran Airways	Reporting_Airline_FL	**Airtran Airways	1 = Airtran Airways
Envoy Air	Reporting_Airline_MQ	**Envoy Air	1 = Envoy Air
Spirit Airlines	Reporting_Airline_NK	**Spirit Airlines	1 = Spirit Airlines
United Airlines	Reporting_Airline_UA	**United Airlines	1 = United Airlines
US Airways	Reporting_Airline_US	**US Airways	1 = US Airways
Virgin Airways	Reporting_Airline_VX	**Virgin Airways	1 = Virgin Airways
Sea Level Pressure	seal_lvl_p seal_lvl_d_p	The air pressure relative to Mean Sea Level (MSL).	MIN: 08600 MAX: 10900 UNITS: Hectopascals SCALING FACTOR: 10 MISSING: Average of Values
Sky Ceiling Determination Measured	sky_c_det_9 sky_c_det_d_9	** The code that denotes the method used to determine the ceiling.	1 = Missing
Sky Ceiling Determination Measured	sky_c_det_M sky_c_det_d_M	** The code that denotes the method used to determine the ceiling.	1 = Measured
Sky Ceiling Height	sky_c_hgt sky_c_hgt_d	The height above ground level (AGL) of the lowest cloud or obscuring phenomena layer aloft with 5/8 or more summation total sky cover, which may be predominantly opaque, or the vertical visibility into a surface-based obstruction.	MIN: 00000 MAX: 22000 UNITS: Meters SCALING FACTOR: 1 MISSING: Average of Values

Descriptive Name	*Column Name	Description	Values
Sky Coverage - 2 Oktas	sky_cov_02 sky_cov_d_02	** The fraction of the total celestial dome covered by the sky cover layer.	1 = Two oktas - 2/10 - 3/10, or FEW
Sky Coverage - 4 Oktas	sky_cov_04 sky_cov_d_04	** The fraction of the total celestial dome covered by the sky cover layer.	1 = Four oktas - 5/10, or SCT
Sky Coverage - 7 Oktas	sky_cov_07 sky_cov_d_07	** The fraction of the total celestial dome covered by the sky cover layer.	1 = Seven oktas - 9/10 or more but not 10/10, or BKN
Sky Coverage - 8 Oktas	sky_cov_08 sky_cov_d_08	** The fraction of the total celestial dome covered by the sky cover layer.	1 = Eight oktas - 10/10, or OVC
Sky Base Height	sky_cov_base_hgt sky_cov_base_hgt_d	The height relative to a vertical reference datum of the lowest surface of a cloud.	MIN: -00400 MAX: +35000 UNITS: Meters SCALING FACTOR: 1 MISSING: Average of Values

Descriptive Name	*Column Name	Description	Values
Sky Cloud Type	sky_cov_cld sky_cov_cld_d	*** The code that denotes the classification of the clouds that comprise a sky cover layer.	00 = Cirrus (Ci) 01 = Cirrocumulus (Cc) 02 = Cirrostratus (Cs) 03 = Altocumulus (Ac) 04 = Altostratus (As) 05 = Nimbostratus (Ns) 06 = Stratocumulus (Sc) 07 = Stratus (St) 08 = Cumulus (Cu) 09 = Cumulonimbus (Cb) 10 = Cloud not visible owing to darkness, fog, dust storm, sandstorm, or other analogous phenomena/sky obscured 12 = Towering Cumulus (Tcu) 13 = Stratus fractus (Stfra) 14 = Stratocumulus Lenticular (Scsl) 15 = Cumulus Fractus (Cufra) 16 = Cumulonimbus Mammatus (Cbmam) 17 = Altocumulus Lenticular (Acsl) 18 = Altocumulus Castellanus (Accas) 19 = Altocumulus Mammatus (Acmam) 20 = Cirrocumulus Lenticular (Ccsl) 21 = Cirrus and/or Cirrocumulus 22 = jenkins-content-114 Stratus and/or Fractostratus 23 = Cumulus and/or Fracto-cumulus
Sky Total Coverage - Clear	sky_obs_tot_cov_00 sky_obs_tot_cov_d_00	*** The code that denotes the fraction of the total celestial dome covered by clouds or other obscuring phenomena.	1 = None, SKC or CLR

Descriptive Name	*Column Name	Description	Values
Sky Total Coverage - 2 Oktas	sky_obs_tot_cov_02 sky_obs_tot_cov_d_02	*** The code that denotes the fraction of the total celestial dome covered by clouds or other obscuring phenomena.	1 = Two oktas - 2/10 - 3/10, or FEW
Sky Total Coverage - 4 Oktas	sky_obs_tot_cov_04 sky_obs_tot_cov_d_04	*** The code that denotes the fraction of the total celestial dome covered by clouds or other obscuring phenomena.	1 = Four oktas - 5/10, or SCT
Sky Total Coverage - 6 Oktas	sky_obs_tot_cov_06 sky_obs_tot_cov_d_06	*** The code that denotes the fraction of the total celestial dome covered by clouds or other obscuring phenomena.	1 = Six oktas - 7/10 - 8/10
Sky Coverage Sum	sky_sum_cov sky_sum_cov_d	The code that denotes the portion of the total celestial dome covered by all layers of clouds and other obscuring phenomena at or below a given height.	0 = Clear - No coverage 1 = FEW - 2/8 or less coverage (not including zero) 2 = SCATTERED - 3/8-4/8 coverage 3 = BROKEN - 5/8-7/8 coverage 4 = OVERCAST - 8/8 coverage 5 = OBSCURED 6 = PARTIALLY OBSCURED MISSING: Average of Values
Sky Coverage Height	sky_sum_hght sky_sum_hght_d	The height above ground level (AGL) of the base of the cloud layer or obscuring phenomena.	MIN: -00400 MAX: +35000 UNITS: Meters SCALING FACTOR: 1 MISSING: Average of Values
Air Temperature	tmp_air tmp_air_d	The temperature of the air.	MIN: -0932 MAX: +0618 UNITS: Degrees Celsius SCALING FACTOR: 10 MISSING: Average of Values

Descriptive Name	*Column Name	Description	Values
Dew Point Temperature	tmp_dew tmp_dew_d	The temperature to which a given parcel of air must be cooled at constant pressure and water vapor content in order for saturation to occur.	MIN: -0982 MAX: +0366 UNITS: Degrees Celsius SCALING FACTOR: 10 MISSING: Average of Values
Visual Distance	vis_dist vis_dist_d	The horizontal distance at which an object can be seen and identified.	MIN: 000000 MAX: 160000 UNITS: Meters SCALING: None NOTE: Values were capped at 160000. If higher it is recorded as 160000. MISSING: Average of Values
Wind Angle Direction	w_dir_angle w_dir_angle_d	The angle, measured in a clockwise direction, between true north and the direction from which the wind is blowing.	MIN: 001 MAX: 360 UNITS: Angular Degrees SCALING FACTOR: 1 MISSING: Average of Values
Wind Speed	w_speed_rate w_speed_rate_d	The rate of horizontal travel of air past a fixed point.	MIN: 0000 MAX: 0900 UNITS: Meters per second SCALING FACTOR: 10 MISSING: Average of Values

* Two column names are noted for each. The first one listed is the column containing data for the weather at the Origin. The second column listed corresponds to the weather at the Destination.

** One-Hot-Encoded Column - This column consisted of more than one value in the original data. It was split out to one column per category/value. A value of 1 in these columns indicates it is this value and 0 is not. This is done for efficient machine learning. This was discussed in the blog.

*** This column was treated as numeric. We realized late this should have been treated as categorical and one-hot-encoded.
Due to this missing is an average of values as opposed to a true missing.
The value of 11 was not used for this column.

Appendix C - Data Collection (*Page 1 of 2*)

In our quest to uncover flight delay insights, we identified several key variables needed to answer our questions, these include airlines, airports, aircrafts, and weather conditions

Airlines/Airports (https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FGJ):

To access the important aviation data needed for our analysis, we relied on the Bureau of Transportation Statistics (BTS), a U.S. government agency that collects, analyzes, and disseminates data and statistics related to all modes of transportation in the United States. This agency data respository is a treasure trove of information on airline on-time performance, which serves as the backbone of our research.

To streamline our data collection process, we relied on Selenium, a robust automation library. By leveraging Selenium, we efficiently retrieved the dataset through the web interface, ensuring utmost accuracy and effectiveness in acquiring the necessary information for our study.

The high level summaries of the BTS domestic on-time data:

- Airline Information: An abundant repository of vital details pertaining to the diverse airlines soaring the skies, complete with their names, codes, and various distinguishing attributes.
- Origin/Departure Performance: A comprehensive window into the world of flight departures, allowing us to understand the intricacies of take-offs and initial stages of each flight.
- Destination/Arrival Performance: A captivating section that sheds light on the final stages of flights as they gracefully land at their designated destinations.
- Flight Summaries: A succinct yet comprehensive roundup of individual flights, encapsulating a wealth of information vital to our understanding. This includes flight distance and elapsed times.
- Cancellations and Diversions: A critical compilation of flight disruptions, including delays, cancellations, and unforeseen diversions, presenting a valuable dimension to our analysis.

Appendix C - Data Collection (*Page 2 of 2*)

Weather (https://www.ncei.noaa.gov/data/global-hourly/archive/csv/)

Zip files available to download provided by the National Centers for Environmental Information (NCEI), part of the National Oceanic and Atmospheric Administration (NOAA), a U.S. government agency responsible for monitoring and predicting weather, climate, oceans, and coasts supports informed decision-making. The below weather measurements, enrich the aviation features bolstering our analysis and prediction of weather-related flight delays:

- Wind: Descriptive type of wind: calm, normal, or variable.
- Sky Condition, Coverage, and Observation: Provides several measures of the height above ground level of the lowest cloud, the fraction of the total celestial dome and the type of clouds.
- Visibility: Distance an object can be seen and identified.
- Temperature: Air temperature and dew point temperature.
- Sea Level Pressure
- Liquid Precipitation: Depth and dimension of precipitation, as well as length of time.
- Atmospheric Pressure Features: Altimeter setting rate and station pressure rate.

This data provided essential insights into visibility and potential impacts on flight operations, aiding us in exploring weather-related correlations with delays. The aim was to provide understanding of weather-related disruptions and contribute to making air travel safer and more efficient.

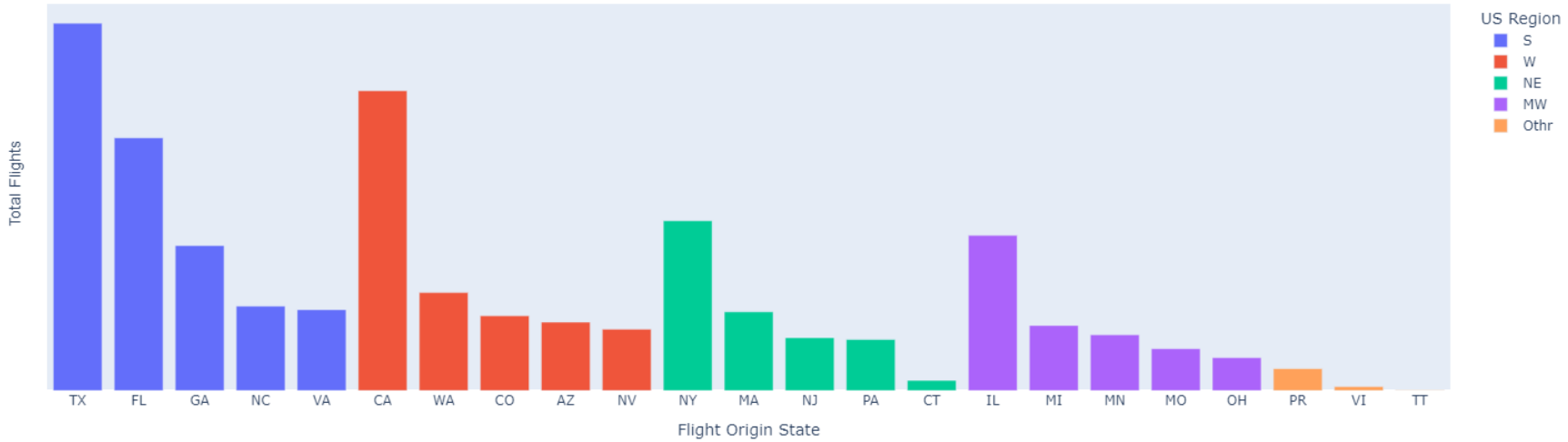
This data can be downloaded in tar.gz files for each individual year. Each file contains weather from all weather stations collected. This includes thousands of stations around the world. When narrowing down to US stations corresponding to airports, the files were narrowed down to approximately 375.

MLID - http://www.weathergraphics.com/identifiers/ master-location-identifier-database-20130801.csv Side note: I did not find the document at this location until writing this portion of our blog. Note the date on the file I used is 20130801.csv. It turns out there's a much more recent version and this is the official source and deserves all credit. Please refer to the document named: **master-location-identifier-database-202301_standard.xlsx** and note that our version is outdated. Our version can be found in our GitHub repository. They do provide this standard version for free but also have a professional version available for a fee. They put a lot of effort into creating the reference document and deserve payment if using this is a commercial manner.

LCD Station Numbers - An NOAA staff member provided an Excel spreadsheet, LCD Station Numbers.xlsx, used to cross reference the weather file names for each station to determine Country. This was used to initially narrow the files down to US weather stations only.

Appendix D - Delay Frequency By Airport

Figure 17: Top 5 total number of flights by state for each region.



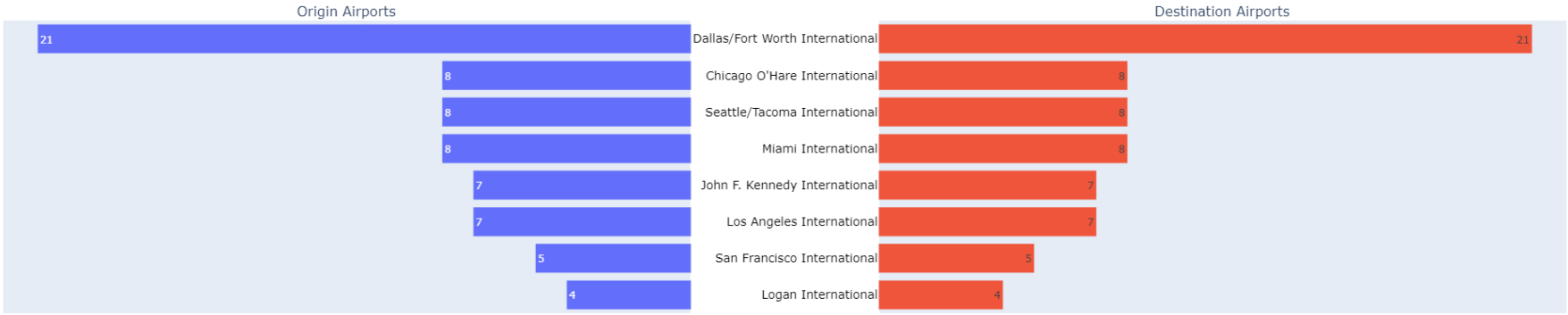


Figure 18: The count of the top 100 most frequently delayed airports

Appendix E - Model Tuning

Focused parameters for the gradient boosting classifier:

- Number of estimators: This controls the number of trees in the model. A higher number of estimators will generally lead to a better model, but it will also take longer to train
- Maximum depth: This controls the maximum number of levels in each tree. A higher maximum depth will allow the model to learn more complex relationships, but it may also lead to overfitting.
- Maximum features: This controls the number of features that are considered when splitting each node in a tree. A higher maximum feature will allow the model to consider more information, but it may also lead to overfitting.
- Minimum samples split: This controls the minimum number of samples that must be in a node before it can be split. A higher minimum sample split will prevent the model from overfitting to noise in the data.
- Minimum sample leaf: This controls the minimum number of samples that must be in a leaf node. A higher minimum sample leaf will prevent the model from underfitting the data.
- By tuning these parameters, you can find a model that is both accurate and generalizable. The specific values that you choose will depend on the dataset that you are using and the desired performance of the model.

Appendix F - Network Modeling

The SIRS model uses something called "ordinary differential equations" (ODEs) to represent how the number of susceptible, infectious, and recovered individuals change as the epidemic spreads.

- **Susceptible (S):** These are people who haven't been infected yet but could be.
- **Infectious (I):** These are people who are currently infected and can spread the disease to others.
- **Recovered (R):** These are people who were infected but have now recovered and gained immunity.

The model is characterized by two key parameters:

- **Beta (β):** This represents the rate at which infections spread. It shows how likely an infected person is to pass the disease to a susceptible person.
- **Gamma (γ):** This represents the rate of recovery. It indicates how quickly an infected person recovers and moves to the recovered category.

We created an objective function that looks for the best-fitting β and γ values to match real-world data of infections and recoveries. It does this by comparing the model's predictions with the actual data and adjusting the parameters to minimize the difference.

The process involves:

- **Initial Conditions:** The model starts with initial values for susceptible, infectious, and recovered individuals. These values are based on the available data for a particular day.
- **ODEs:** The ordinary differential equations describe how the number of people in each category changes over time. These equations use the β and γ values to simulate the spread and recovery of the disease.
- **Simulation:** The equations are solved numerically for each time step, creating a simulation of how the epidemic evolves over time.
- **Comparison:** The simulated values are compared to the actual recorded counts of infections and recoveries for that day.
- **Optimization:** The β and γ values are adjusted to minimize the difference between the simulated and actual counts. This process helps find the parameters that best match the real data.
- **Basic Reproduction Number (R0):** The ratio β/γ is calculated. It represents the potential for disease spread. If R0 is greater than 1, the epidemic can grow; if it's less than 1, the epidemic can die out.

This whole process is repeated for each day in the dataset, providing insight into how the epidemic's dynamics change over time and identifying the parameters that best explain the observed patterns.

Generated Network showing the Airport and their Clustering Coefficient with plotly:

- https://github.com/Team-Takeoff/Capstone/blob/main/assets/AirportCLusteringCoefficient.html
- https://github.com/Team-Takeoff/Capstone/blob/main/assets/DelayBottleNecksCLusteingCoef.html

Appendix G - Additional Classifier Comparisons (*Page 1 of 3*)

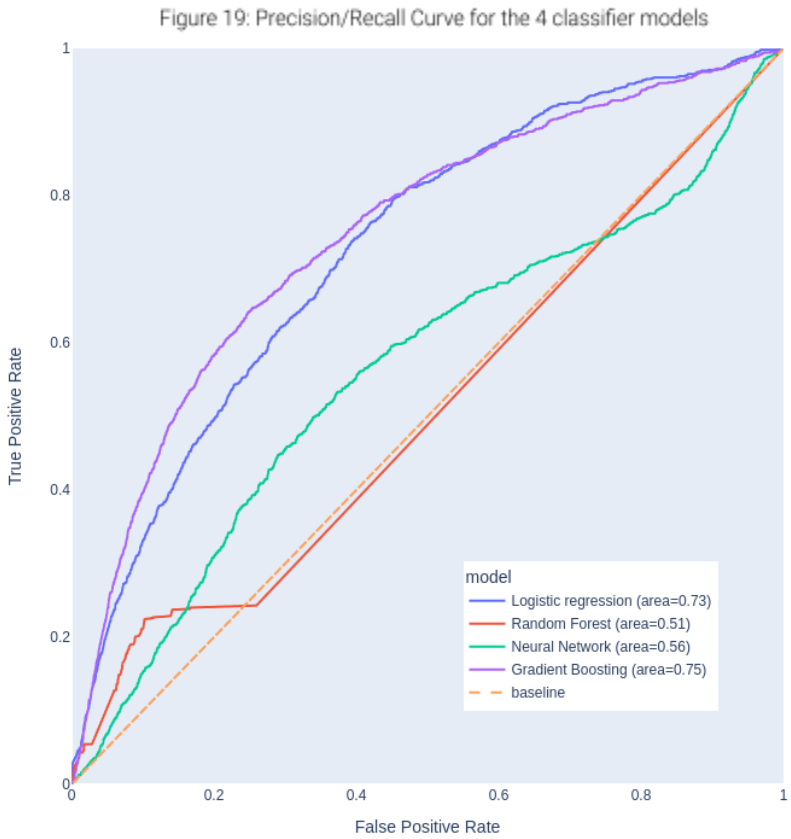


Figure 19: A comparison of the ROC and associated Area Under the Curve for all models. The ideal classifier would create a curve that has an area of 1, or the entire area of the graph. The GBC had the largest area primarily due to its ability to predict the dominant non-delay class, cementing its position as the most accurate classifier. By contrast the Random Forest proved to be the least effective classifier, hugging the diagonal baseline which represents a model that only slightly outperforms a dummy classifier that randomly guesses each prediction.

Appendix G - Additional Classifier Comparisons (*Page 2 of 3*)

Below, variables were categorized based and plotted based on the following: Airline, Aircraft, Origin and Destination weather variables. Regarding the Airlines and Aircraft variables, we recommend exercising extreme caution in the interpretation of these weights. Many of these are binary (one-hot encoded) variables and their weights do not suggest a propensity for delay or non-delay but merely how important these variables were to the overall prediction performance on unseen data. These weights can also be influenced by the imbalance in distributions of the data. For example, American Airlines will have higher representation of flights compared to Spirit as it is a larger Airline. The same can also be said of Airbus and Boeing.

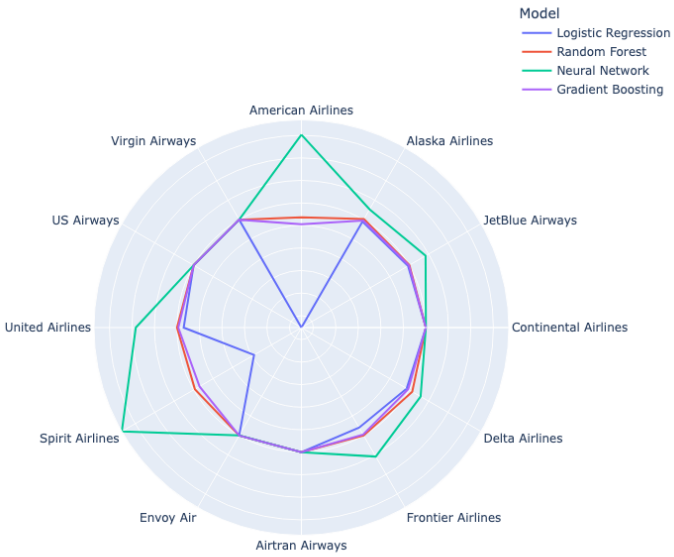


Figure 20: Permutation Importance on Airline Variables

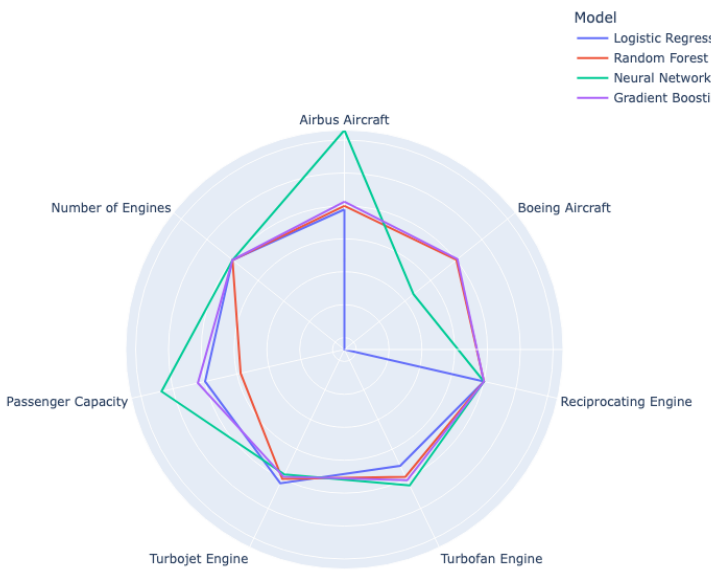


Figure 21: Permutation Importance on Aircraft Variables

Appendix G - Additional Classifier Comparisons (*Page 3 of 3*)

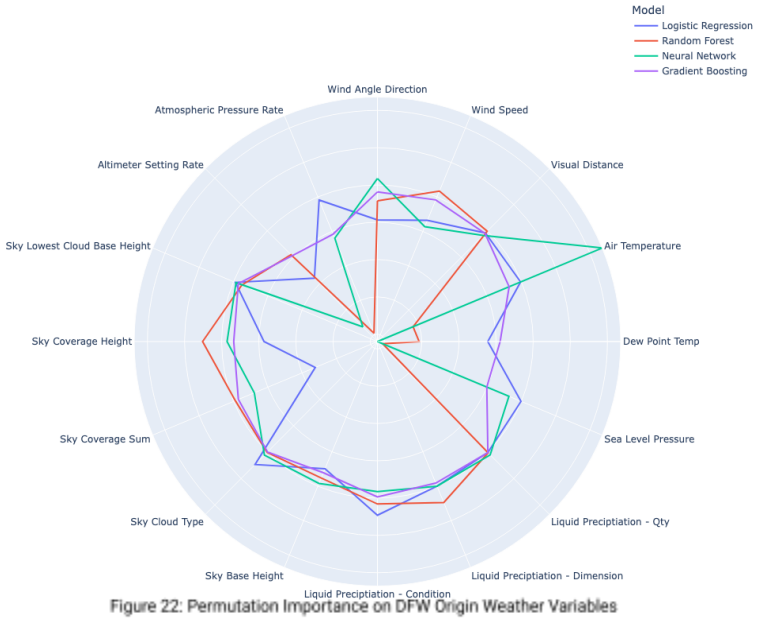


Figure 22: Permutation Importance on DFW Origin Weather Variables

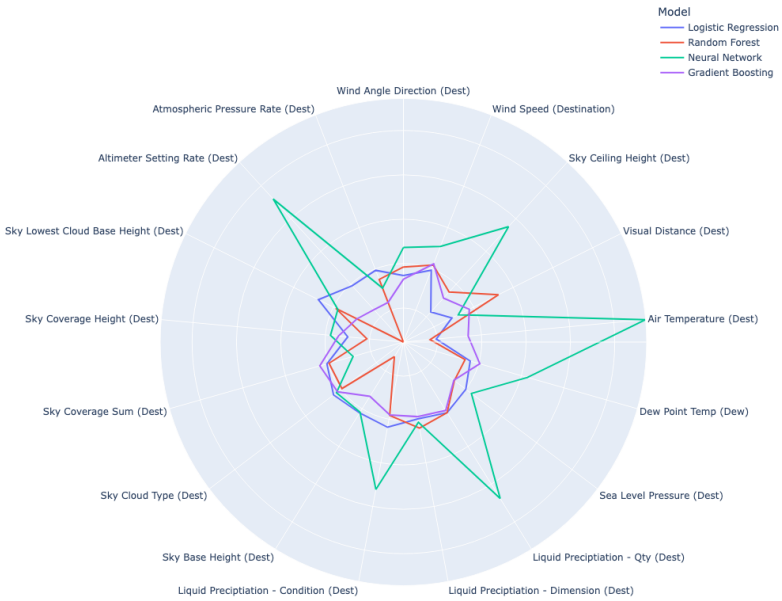


Figure 23: Permutation Importance on Destination Weather Variables

Appendix H - Limitation

We also found it to be extremely odd that the BTS data we downloaded contained exactly 1,200,000 records/flights per year from 2010-2019 and exactly one quarter of that, 30,000, for the first quarter of 2023. This, and when looking at frequencies of flights in various ways we felt we must be missing chunks of data.

Appendix I - Statement of Work *(Page 1 of 2)*

Project Work	Task	Assignee
Admin	Software tools write-up	Kuan Yu Chen
Admin	Standup 1 - Prep	Kuan Yu Chen
Admin	Standup 1	Kuan Yu Chen, Stacey Bruestle, David Boudia
Admin	Data leakage plan mini deliverable	Kuan Yu Chen, Stacey Bruestle, David Boudia
Admin	Report outline mini deliverable	Kuan Yu Chen, Stacey Bruestle, David Boudia
Admin	Standup 2 - Prep	Kuan Yu Chen
Admin	Standup 2	Kuan Yu Chen, Stacey Bruestle, David Boudia
Admin	Mentor check in	Kuan Yu Chen, Stacey Bruestle, David Boudia
Admin	Visuals mini deliverable	Kuan Yu Chen, Stacey Bruestle, David Boudia
Admin	Standup 3 - Prep	Kuan Yu Chen, Stacey Bruestle
Admin	Standup 3	Kuan Yu Chen, Stacey Bruestle, David Boudia
Admin	Revised outline mini deliverable	Kuan Yu Chen, Stacey Bruestle, David Boudia
Admin	Decide on medium for the write-up	Stacey Bruestle, David Boudia, Kuan Yu Chen
Admin	Outline the write-up - Used Outline Mini-Deliverable	Stacey Bruestle, David Boudia, Kuan Yu Chen
Admin	GitHub repository mini deliverable	David Boudia, Stacey Bruestle, Kuan Yu Chen
Admin	Pre-flight checklist mini deliverable	David Boudia, Stacey Bruestle, Kuan Yu Chen
Admin	Code Cleanup	Stacey Bruestle, David Boudia, Kuan Yu Chen
Gather Data	BTS data download	Kuan Yu Chen
Gather Data	FAA Data	David Boudia
Gather Data	Weather data download	Stacey Bruestle

Appendix Statement of Work *(Page 2 of 2)*

Project Work	Task	Assignee
Wrangling	Cleaning - Overall	David Boudia, Stacey Bruestle, Kuan Yu Chen
Wrangling	Cleaning - imputation	David Boudia, Stacey Bruestle, Kuan Yu Chen
Wrangling	Cleaning - filter unwanted entries	David Boudia, Stacey Bruestle, Kuan Yu Chen
Wrangling	Data Scaling	David Boudia, Stacey Bruestle
Wrangling	Categorical Encoding	Stacey Bruestle, David Boudia
Wrangling	Weather Data	Stacey Bruestle
QI	Obtain interviewees	Stacey Bruestle, David Boudia
QI	Interview Pilot	David Boudia, Stacey Bruestle, Kuan Yu Chen
QI	Interview Airlines Data Scientist	David Boudia, Stacey Bruestle, Kuan Yu Chen
QI	Cluster Notes	David Boudia
Modeling	Develop data input	David Boudia
Modeling	Neural Network Model	David Boudia
Modeling	Gradient Boosting Classifier	Kuan Yu Chen
Modeling	Random Forest Model	Stacey Bruestle
Modeling	Logistic Regression Model	Stacey Bruestle
Modeling	Network Analysis - Modeling	Kuan Yu Chen
Modeling	Classification Model Evaluation and Visualization	David Boudia
Final Report	Write-up	Stacey Bruestle, David Boudia, Kuan Yu Chen
Final Report	Visualizations	Stacey Bruestle, David Boudia, Kuan Yu Chen
Final Report	Pushed All Code to Git	David Boudia