

Acknowledgement:

This Corporate Project was developed as part of my MSBA program at Arizona State University. Team members of this project were Raul Contreras, Kyla Stewart, and Hannah Torrey. We all were involved during all the different stages of the project and the tools and software used were Tableau, Python, SPSS, and Azure ML.

Introduction

Purpose of project

The purpose of this project is recommending changes that will reduce delays and cancellations through descriptive, predictive, and prescriptive data analytics to Delta Airlines. Flight delays not only cause inconvenience to passengers, but also cost carriers billions of dollars. These delays and cancellations tarnish the airlines' reputation, often resulting in loss of demand by passengers. The scope of this project includes descriptive analytics, predictive analytics, and prescriptive analytics.

There were 6 important milestones during the project: Project kickoff, descriptive analytics, interim presentation, predictive analytics, prescriptive analytics, and final project presentation and report. There are no costs associated with the project, but there are resources needed to develop it. Mainly, we will need tools such as Tableau, SPSS, Azure ML, and Google Suite. Fortunately, they are accessible for free by having Citrix accounts from ASU and Microsoft - Google accounts.

Team

The project team consist of the following members:

Project lead: Kyla Stewart. Responsibilities: Lead project, delay data visualization, multiple correlations, business recommendations, and market research.

Team Member: Raul Contreras. Responsibilities: Data prep, overall data visualization, multiple correlations, ML modeling, and ML algorithm outputs, and Machine Learning recommendations.

Team Member: Hannah Torrey. Responsibilities: Cancelled data visualization, data summary, multiple correlations, ML training, ML modeling, ML algorithm outputs, and decision analysis.

Data source

Data was collected from the Bureau of Transportation Statistics ("On-Time : Reporting Carrier On-Time Performance (1987-present)") which holds 109 features of historical data related to time period, airline, origin, destination, departure performance, arrival performance, cancellations and diversions, flight summaries, cause of delay, gate return information origin airport, and diverted airport information.

Not all data was used during the machine learning model, so the most descriptive 29 features for our objective were selected. In addition, dummy variables or feature engineering were created from departure delay and arrival delay to create departure delay FAA and arrival delay FAA increasing the features from 29 to 31. A flight delay is when an airline flight takes off and/or lands later than its scheduled time. The Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes later than its scheduled time. So, the value of 1 represents those flights delayed with the latter description, and a value of 0 for those less than 15 minutes.

Data was collected for all Delta Airlines flights from 2019 with 991, 986 observations and a size of almost 100 MB as a CSV file. Below there is a link and screenshot (Figure 1) of the Bureau of Transportation Statistics website: https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time.

Methods used

Data visualization. Visualization was plotted with the help of Tableau. Total flights, delays, cancellations by month, day of week, airport, causes in minutes were analyzed.

Predictive Analytics. Binary Classification Model developed in Azure Machine Learning Studio. Smote, Hyperparameter Tuning and Feature Selection approaches were implemented.

Background

Delta's History

Delta's 95 year-old history exhibits its dedication to innovation, competition, safety and customer experience.

In 1925, Delta's predecessor and the first commercial agricultural flying company, Huff Daland Dusters, was founded in Macon, Georgia. At the time, the company's 18 planes were the largest privately owned fleet on earth, and were mainly used for crop-dusting. In 1928, C.E. Woolman purchased Huff Daland Dusters. He renamed the company after its service area, the Mississippi Delta region. A year later in 1929, Delta conducted its first passenger flight from Dallas, Texas, to Jackson, Mississippi. Quickly following, Delta made stops in Louisiana and offered services in Alabama. In 1941, Delta headquarters moved its headquarters to Atlanta, where it has remained and dominated ever since. Twenty years later, Delta operated its first nonstop flight from Atlanta, Georgia to Los Angeles, California. That year, the airline became the first to connect commercial jet passengers from California to the Caribbean. In 1961, Delta reached flying over 11 billion passenger miles without a fatality, and was given a National Safety Award to recognize the accomplishment. In 1970, Delta began flying the Boeing 747. Boeing 747 service allowed Delta to carry over 375 passengers per flight, increasing capacity and decreasing cost for the airline (Gay).

In 1991, Delta purchased Pan Am's trans-Atlantic routes and the Pan Am Shuttle, making Delta a global carrier, continuing to transport passengers in its jets domestically, but internationally as well. Delta's acquisition of Pan Am is the largest acquisition of flight routes to this day (Gay).

At the start of the new decade, Delta launched SkyTeam, a global alliance between Delta, Aeromexico, Air France and Korean Air, which now includes 19 airlines. That year, Delta ordered the largest purchase of regional jets in history, at a staggering 500 regional jets. These jets were used to help secure Delta's business model of dominating business travel and unique markets through expensive short-distance travel. By 2000, Delta had carried more than 120 million passengers on their planes (Gay).

Delta began struggling financially in the start of the new millennium, partly due to high fuel and labor prices. Though the company remained competitive, like most airlines, they were not making much of a profit. In an attempt to combat financial woes, in 2005, Delta carried out Operation Clockwork, resulting in an increasing amount of on-time departures, contributed to reducing congestion in airports and made more aircrafts available for scheduling. To this day, Operation Clockwork is the "largest single-day redesign in aviation history" (Gay).

Although Operation Clockwork was successful in achieving Delta's initial goals, it didn't improve the airline's financial situation enough. Due to high fuel prices, cost of labor, competition, Hurricane Katrina and other economic and political factors, in September of the same year, Delta filed for bankruptcy, after a number of other passenger airlines (Ramos). Following this, the company made numerous changes that allowed them to remain competitive and profitable. Despite filing for bankruptcy, two months later, Delta began one of its largest expansions in its history-- service to seven new Latin American and Caribbean routes to meet demand in the area (Isidore).

In 2008, Delta acquired Northwest Airlines. With this acquisition came the ability for Delta to fly in all areas of the world, completing their goal of becoming a global competitor. This year, Delta also became the first airline in the United States to offer in-flight Wi-Fi on domestic routes, the first of many major web-related and technology-based innovations (Gay).

A year later, Delta announced a partnership with Air France-KLM, vastly expanding their reach across Europe. In 2009, they also began offering nonstop flights between Los Angeles, California and Sydney, Australia, becoming the only airline in the United States to operate on six continents. Delta's quick and wide expansion allowed them to reach more markets, remain competitive and determine their most profitable routes. That year, Delta also integrated all of its frequent flyers into SkyMiles, becoming the world's largest loyalty program, boasting over 74 million members. SkyMiles points can be earned by travelers through shopping, dining and flying, and are redeemable on Delta flights, as well as partner routes (Gay).

In 2010, Delta announced a \$3.2 billion upgrade to its services, with the goal of improving customer experience. Product upgrade, the first in over ten years, included bed and video player installations in most aircrafts, additional first class seats and renovated terminal lounges. This upgrade also included a \$1 billion expansion of their JFK facilities. Much of this budget was focused on improving and innovating current technology so customers could experience a seamless travel experience from booking to landing in their destination. A year later, the airline became the first to provide a mobile bag tracker to travelers, in efforts to reduce stress and time associated with losing, locating and retrieving luggage. In 2017, Delta placed 63rd on Fortune's "Best Companies To Work For" list, based on employee ratings of workplace, culture and job satisfaction. Notably, Delta was the only airline included. That year, Delta also invested 49% in Aeromexico, creating a trust between Mexico and the United States, and improving travelers' schedules and overall flying experience through integration. In 2018, Delta partnered with Korean Air, creating a hub in Seoul, and launched a daily nonstop service from Atlanta to Shanghai. These improvements contributed to alleviating travel unknowns that are common when changing airlines across countries (Gay).

Three themes remain constant throughout Delta's history -- growth, innovation and customer service. Throughout its history, Delta has instilled these three values at the forefront of their business decisions, which are showcased through their business model and competitive advantages.

Business model

The airline industry, as mentioned in Delta's history, is known for competition, low profit margins and impact of the current market. Delta strives for an overall flexible business model to combat these challenges through its cost structure, target market and control of their fuel (Sam).

The airline aims to allow their cost structure to shift from fixed to variable whenever need be. This model allows Delta to react quickly to demand, scaling up or down when needed. One way Delta reacts quickly to demand is through purchasing used aircrafts. Used aircrafts are lower in fixed costs, but often higher in maintenance, or variable, costs. Based on demand, Delta can decide what type of costs they need to be inquiring about, for example, purchasing or fixing up aircrafts. Having the choice of where to allocate funds gives Delta options based on the market and its own needs. In case there is a sharp change in demand, Delta can quickly adjust its cost model to adapt. This enables the airline to maintain larger profit margins than fellow US carriers, resulting in a sustainable competitive advantage and reduced risk when changes in demand occur (Bachman).

Delta's business model also includes owning a fuel refinery. Fuel hedging and futures trading is very standard among the airline industry because of how much the fuel industry impacts them. High fuel prices was one of the main reasons Delta filed for bankruptcy in 2005. However, in 2012, Delta bought a fuel refinery for \$150 million in efforts to manage their costs and business processes. Now, when fuel prices increase, Delta does not have to worry as much about the future of their business because they are in control of their fuel. Furthermore, this helps them maintain competitive price advantage. As of 2020, Delta is the only U.S. airline that owns a fuel refinery -- a decision that showcases the company's innovation, investment into the future and forward thinking (Sam).

Another huge source of revenue is Delta's corporate travelers. Delta focuses on attracting and retaining this segment because they generally result in higher margins due to their lower price sensitivity. Furthermore, business trips must occur regardless of the price. Because of this, when Delta can retain business travelers, they are almost guaranteed revenue from the market because of the necessity of the trip, contrary to vacationers looking for a cheap time to travel and route (Bachman). Delta manages their corporate travelers through a multitude of business programs that are tailored to different types of workers, including contracts with companies, hubs that can be used for meetings and group travel discounts. Throughout each of these programs, Delta reiterates the high level of customer service a businessperson can expect and the benefits that come with partnering with Delta, including personal travel perks, airport lounges and discounts.

Competitive advantages

According to airline analysts, three business decisions separate Delta from other airlines, allowing them to hold a sustainable competitive advantage.

Delta relies heavily on their regional operations. These operations are more costly, which include price of a ticket, resulting in higher profit margins. Delta holds about a 60% share on its regional market, 7% higher than American Airlines, which has the second largest share. Though Delta conducts many cross-continent routes, their regional fares are what keeps their business competitive. Many of Delta's regional flyers are its corporate travelers that the company heavily invests in and cultivates. By attracting and retaining regional corporate travelers, who are less price sensitive, and holding a majority share, Delta maintains a sustainable competitive advantage (Bachman).

Another one of Delta's competitive advantages is its abundance of extra legroom premium economy seats on its regional jets. Not only do regional jet seats cost more, but seats with more leg room cost an additional fee, making these routes extremely profitable. Usually, these comfortable seats target business travelers, as the cost is more expensive than an economy seat, but don't price out the segment. Analysts have noted that Delta has more extra leg room seats than other airlines offering regional routes. By cashing in on these minor changes, Delta is able to offer luxury to their traveling customers and their profit margin increases (Bachman).

Lastly, and seemingly most importantly, Delta holds a market share of 76% in Atlanta, its major hub. This is crucial to Delta's success and competitive advantage because Atlanta is home to over 30% of its domestic capacity. Furthermore, at its other four largest hubs--Minneapolis, Detroit, New York, and Salt Lake City—Delta does not have much exposure to its largest low-cost competitor-- Southwest. By dominating their home market and avoiding their biggest price competitor, Delta can ensure customer loyalty without sacrificing price and profitability. Another contributing factor to Delta's competitive advantage is that its other competitors, American and United, have far more capacity overlap with Southwest in their home hubs. Delta's competitors have to worry more about their prices because of their route similarity with Southwest. This gives Delta a major price advantage and a long-term sustainable advantage (Bachman).

Through regional operations, extra legroom and unique hubs, Delta holds a sustainable competitive advantage in their domestic and international routes.

Customer service

Exquisite customer service is embedded in Delta's core values as a company. Three unique ways the airline provides this customer service is through an industry-leading credit card, a top-of-the-line main cabin experience and a new facial recognition service.

Delta's credit card partnership with American Express is tailored to customers' spending habits. Credit card owners earn miles faster when purchasing groceries, spending money on entertainment and buying Delta products and services. Through credit card purchases, customers earn SkyMiles that accumulate after paying off the card. Delta also partners with travel-related companies including AirBnB and Lyft to give customers more SkyMiles and special deals and coupons. Additionally, each customer receives a welcome bonus offer with their new card that allows them to accrue points faster if they spend a certain amount of money in the first three months of owning the card. Delta offers four different tiers of credit cards that include offers relative to a customer's travel and spending habits, ranging from an infrequent traveler to a weekly corporate traveler (Steele). Customer service bonuses related to the credit card include free checked bags and lounge access. Furthermore, credit card owners are able to use their SkyMiles on Delta partner flights, which dramatically opens up route paths and time availability and options. Delta's American Express credit card is an extension of their customer service that escalates as customer purchase loyalty increases.

Another way Delta prioritizes customer service is through their "industry-leading main cabin experience." Rather than only treating first class as unique, Delta ensures all members of their flights feel special by including perks throughout the journey. These perks include a welcome cocktail, bistro-style dining and on-demand snacks("Delta debuts industry-leading international Main Cabin experience"). Customers can feel comforted that they will be taken care of when on a Delta flight regardless of the price of their seat. In turn, Delta hopes to retain happy customers who refer their network to the airline.

Delta's customer service is based on a seamless travel experience. A major way this is possible is through the airline's facial recognition program. Their facial recognition aims to solve the rigid flight process, from arriving for a flight to getting off at a destination. Facial recognition is currently being rolled out all over the United States, and is offered during baggage handling, check-in, TSA checkpoints, boarding, bag tracking and flying ("INFOGRAPHIC: Rolling out optional facial recognition technology to improve the travel experience"). Though the program is new, each airline equipped with facial recognition is already saving over nine minutes when boarding and deplaning. Passengers are always given the option to opt out, and have seen an average of 2% opt out thus far . Delta hopes to roll out facial recognition at all touchpoints throughout the traveling journey and to all their terminals worldwide as soon as possible (Steele).

Innovation

Delta believes in continuous innovation, especially technology-related innovation that can improve customer experience. Three ways the airline is currently experimenting and debuting innovative solutions is through their continual bluetooth tracking improvements, pet travel solutions and quantum computing.

Delta was the first passenger airline to incorporate bluetooth tracking into their fleet, and they are continuing to make progress with the technology through providing real-time tracking for unit load devices. Unit load devices move cargo shipments, baggage and mail all over the world, and are what are used to track checked bags. Delta Cargo ships products weighing over 500 kilograms from organs to flowers to over 300 destinations on six continents worldwide. Previously, tracking was conducted manually. With this innovation comes a completely automated solution that can track all cargo and luggage via GPS, allowing handlers and owners transparency and security in knowing where their goods are (Egerton).

Delta has recently debuted CarePod, a travel carrier that provides pet owners real-time updates through GPS tracking. The most important features surrounding the new CarePod are safety-related, as the airline wants to provide a secure and comfortable travel experience for all passengers, including pets. CarePod offers strong and insulated walls, so that regardless of the external temperature, each pet does not feel a major change inside their carrier. Additionally, multi-layer blinds and windows block out visual distractions and worries for pets without leaving them in darkness. A brand new hydration system included in the CarePod that can hold a liter of water in a spill-proof bowl that is automatically refilled. Delta is monitoring CarePods at all times to ensure pets are safe and their GPS is providing owners the most updated information that can be accessed on a mobile device. All of these features make for a safe and innovative pet carrier that pushes the limits on technology and engineering to ensure customers' peace of mind (Egerton).

Recently, Delta partnered with IBM to explore how quantum computing can improve travel processes. With the idea of bettering the travel of customers and employees, Delta plans to work with IBM's Q Hub to expand and improve upon and discover previously undeveloped and unknown touchpoints and experiences throughout the travel process. As in line with Delta's core beliefs, the end goal of this partnership is to provide a seamless and safe travel experience for everyone on Delta planes ("Delta partners with IBM to explore quantum computing – an airline industry first").

Airline industry overview

The United States airline industry is massive, with 18 passenger airlines flying 926 million passengers (Smallen, 2020) and generating \$4.6 billion in net profit in Q3 2019 (Smallen, 2019). This is a 21.5% increase in profit over Q3 prior year. With more passengers than ever, airlines must positively differentiate themselves to increase their market share. As with any process improvement, it's imperative to understand the current situation before identifying where improvements can be made.

Customer satisfaction & complaints

Since customers are the source of market share, their satisfaction is integral to the success of the company. While Delta routinely surveys its current customers, the results are limited, as they do not cover all airlines and can be biased.

Instead, we will rely on three independent sources: the JD Power 2019 North America Airline Satisfaction survey, Wall Street Journal's "Best and Worst Airlines of 2019", and Business Travel News' 2019 corporate travel survey. JD Power reports that customers were more satisfied with the quality of their air travel experiences than in any of the previous 14 years. They attribute this increase to newer planes, better ticket value from traditional airlines, and improved customer touchpoints. They note that the ratings of low cost and traditional carriers are converging, indicating an improvement on the part of traditional carriers like Delta (JD Power).

The company scores well on all three surveys: second in JD Power’s traditional carrier ranking (JD Power), “Overall Best Airline” from the Wall Street Journal (with Delta winning the on-time arrival rate, cancelled flight rate, and passenger bumping rate categories) (McCartney), and first in Business Travel News’ best airlines for corporate travel buyers (Baker). However, with competitors continually making improvements to their service, Delta must do the same to remain a top performer. For this project, we have decided to focus on flight delays and cancellations as a lever for improvement.

Delays and Cancellations

The Bureau of Transportation Statistics categorizes delays and cancellations into five possible causes (FAA).

Weather	Extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival
National Airspace System (NAS)	Non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc.
Security	Evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas
Carrier	Within the control of the air carrier (cleaning, repair, crew legality, slow passenger boarding, catering, etc.)
Late Aircraft	Arrival delay at an airport due to the late arrival of the same aircraft at a previous airport

For the purposes of this report, we will concentrate our efforts on the two types within Delta’s control: late aircraft and carrier delays. NAS delays are determined by the National Airspace System and are out of Delta’s control. Similarly, security delays are controlled by the Transportation Security Agency (TSA) and law enforcement. Weather is beholden only to itself. These three types generally affect all airlines equally, so will not be considered.

Impact of delays and cancellations

Delays and cancellations, whatever their cause, are impactful in operations, customer satisfaction, and to the airline’s financial position. Operationally, the largest impact is a phenomenon called delay propagation (FAA). Delay propagation is a domino effect that occurs when a single delayed flight “causes flight disruptions for hundreds of other flights” (Centives). In 2014, Mashable reported, “[i]n the airline world, delays build as the day wears on. This summer, for example, airlines were on-time around 85% or better until mid-morning. By mid-afternoon, the rate dropped into the low 70s, then plunged into the 60s by dinner time. Delays early in the day are particularly problematic” (Mashable). From a customer satisfaction standpoint, customers that experience problems with a business’ product or service are less likely to repatronize that business. This lost revenue can have a significant impact on Delta’s profits. Delays and cancellations also cost the airline directly. Delayed flights, depending on whether the delay occurs in the air or on the ground, can incur additional fuel, personnel, and airport costs. Delays cost the airline industry \$8 billion in 2010 (Mashable). In 2014, a cancelled flight was estimated to cost \$5770. Assuming this average holds true five years later, Delta’s cancellation costs in 2019 were approximately \$10.6 million (Mashable).

Data Overview

All Delta flights

Analyzing data from the Bureau of Transportation Statistics, there are 3 flight status: Cancelled Flights, Delayed Flights, and On Time Flights. With 991,986 flights occurred in 2019 by Delta Airline, we can see in the next graph that June, July, and August are the busier month for our client having a maximum point in August with 91,278 flights from this 76,059 were on time, 15,005 delayed, and 214 cancelled. The months with fewer flights were January and February having a minimum point in February with 67,337 flights from which 55,629 on time, 11,645 delayed, and 63 cancelled flights. The busier months match with summer which are the months with more recreational activities due to the end of classes and the nice weather in several parts around the globe.

All Flights by Month

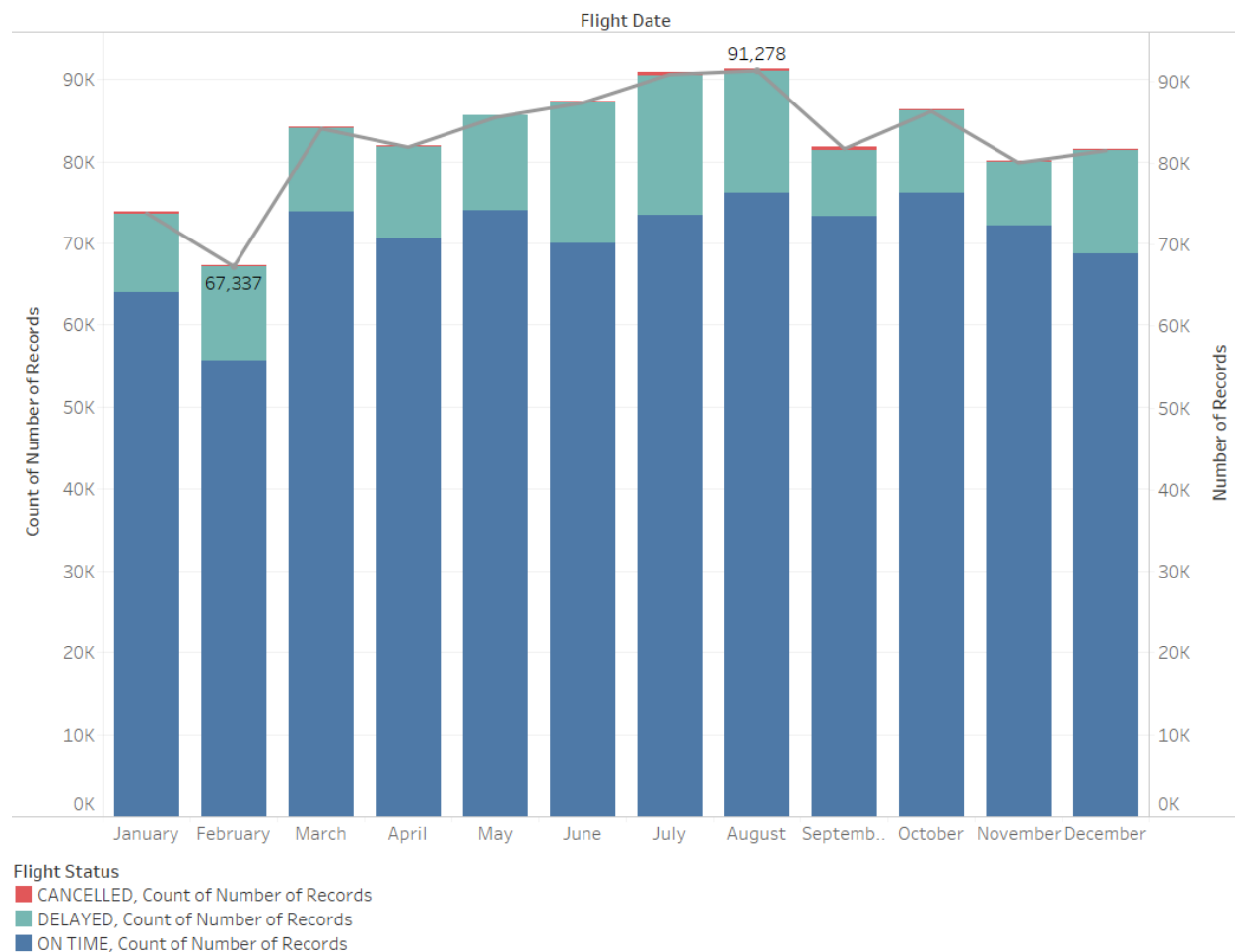


Figure 1

If we drill this down to the day of the week, the busier days are Monday to Friday identifying Monday as the busiest day in 2019 with 149,560 flights divided as 118,871 on time, 20,285 delayed, and 68 cancelled flights. Contrary to that, we have Saturday as the slowest day of the week with 114,035 from this 98,996 are on time, 14,771 are delayed, and 268 are cancelled flights. The graph looks as following:

All Flights by Day of Week

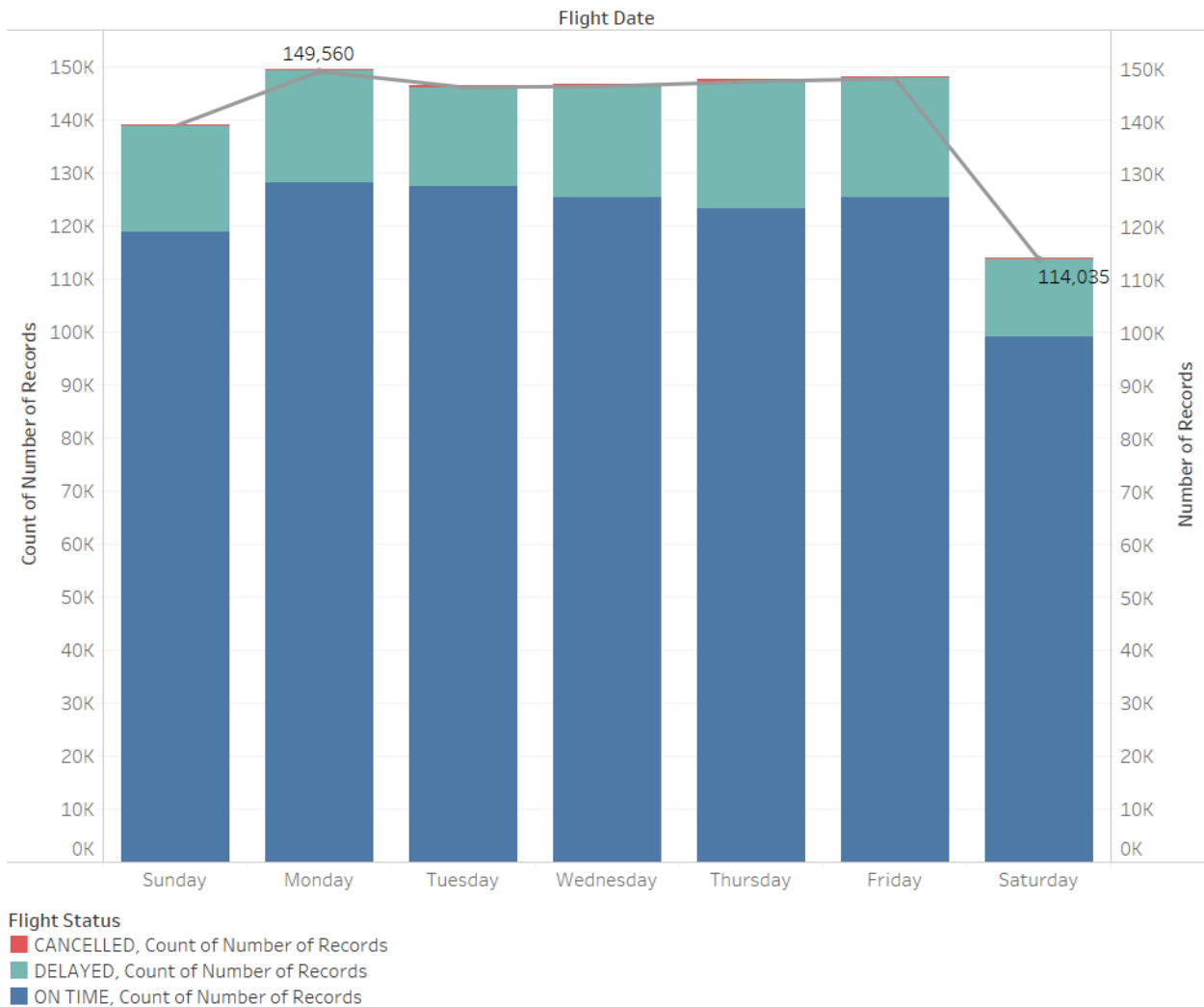


Figure 2

We can see a bar chart with a high percentage of flights that departed on time. The second highest percentage comes from delayed flights, and some few cases are cancelled flights. Flights are broken down by month being June, July, and February the months with less flights on time and with a higher concentration of delayed flights. This represents around 17% to almost 20% of delayed flights.

Similarly, the days of the week with less flights on time were Thursday and Friday and with a higher concentration of delayed flights. The delayed flights portray almost 15% to 17% of the flights during these days. Here are some displays of these numbers.

Percentage Flights by Month

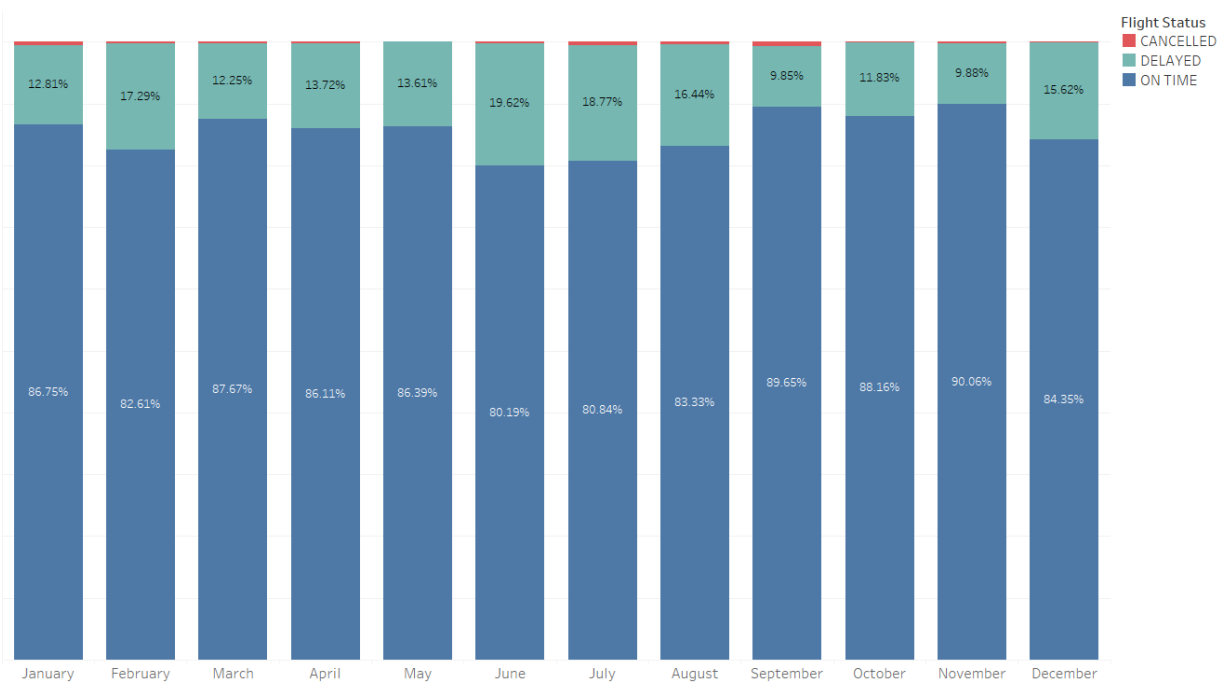


Figure 3

Percentage Flights by Day of Week

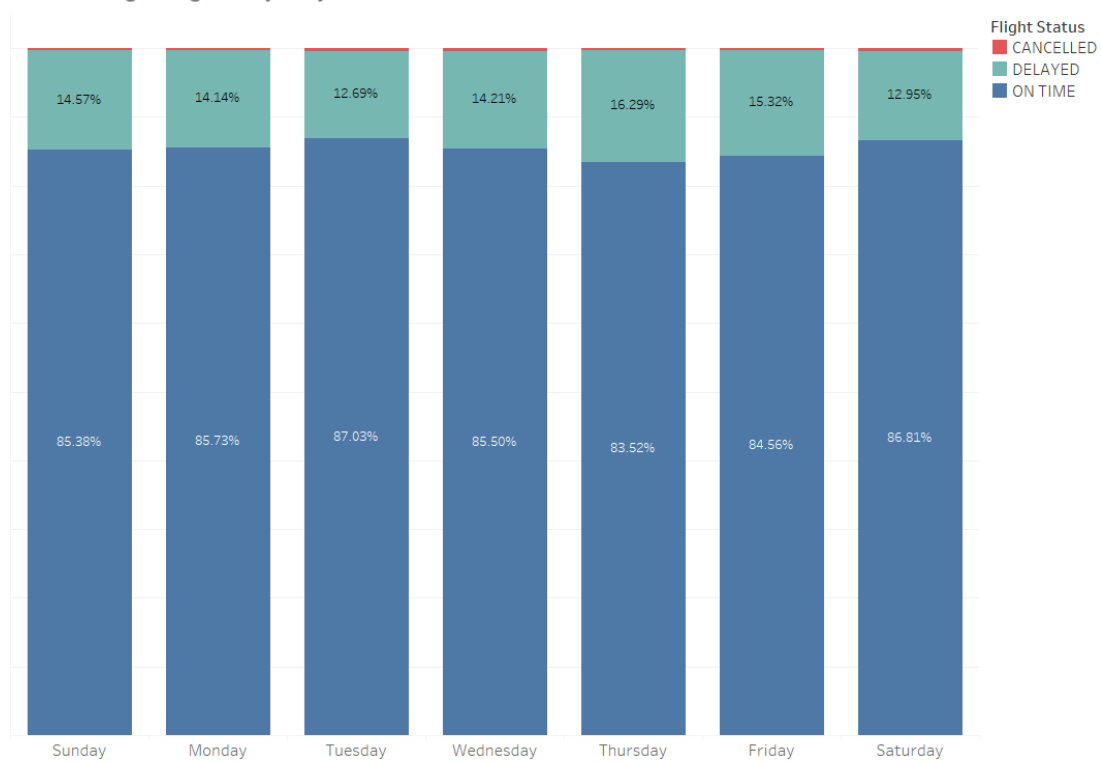


Figure 4

Delay Deep-dive

Delays are most likely to occur in June, July and August and least likely to occur in January, September, and November. Carrier, late aircraft, and NAS delays are most common, with comparatively very few weather and security delays (Figure 5). On a weekly basis, Thursdays and Fridays have the highest rate of delays, while Saturday has the lowest (Figure 6). Delta's hub, Atlanta, is either the origin or destination of most of the delayed flights in 2019. Los Angeles, Salt Lake City, Minneapolis, Detroit, and New York City are also notable in their rates of delayed incoming and outgoing flights (Figure 7).

Delays by Month



Figure 5

Delays by Weekday

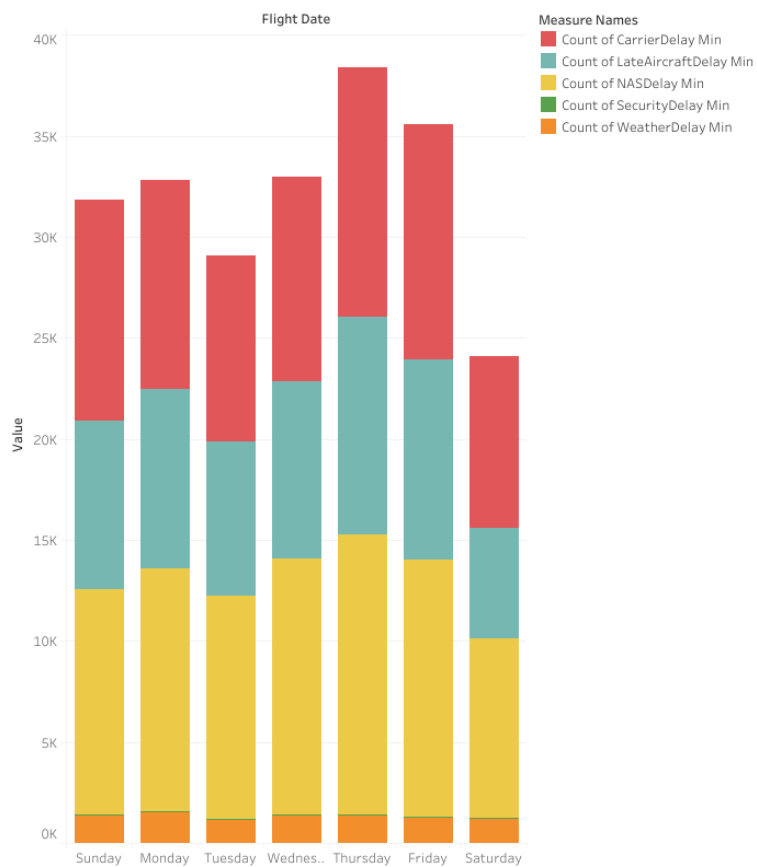


Figure 6

Delays by Origin

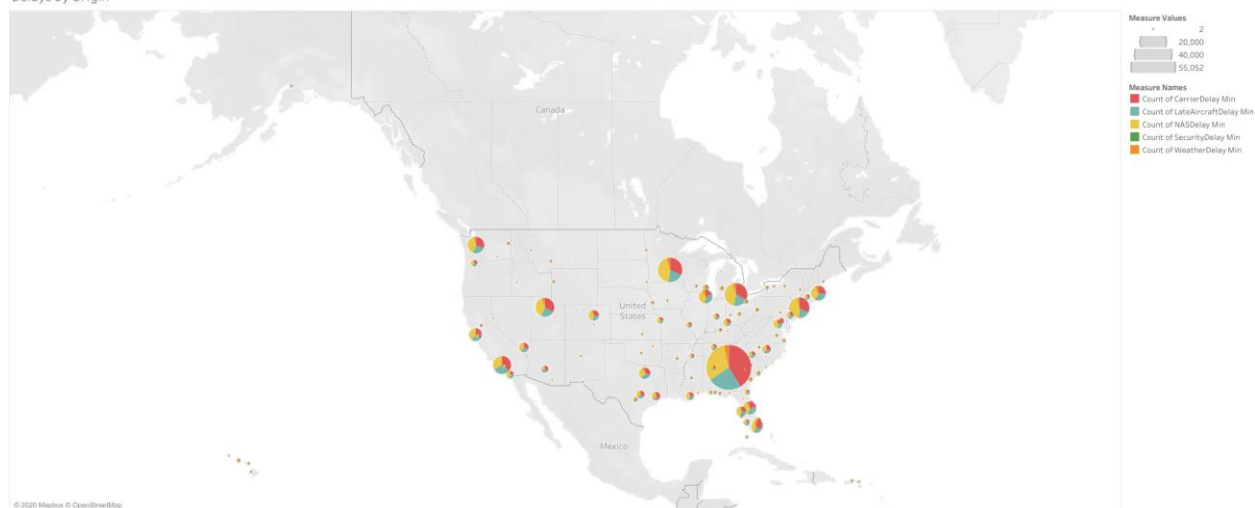


Figure 7



Figure 8

In 2019, Delta flights experienced 9.7 million minutes of delays, the equivalent of nearly 6,800 days. Each minute of delay incurs crew, fuel, airport, and other costs, making delays extremely expensive for the airline. Carrier and late aircraft delays, both under Delta's control, account for 65% of the total delay minutes, indicating that reducing instances in these two categories will have a significant impact on Delta's overall On Time Performance.

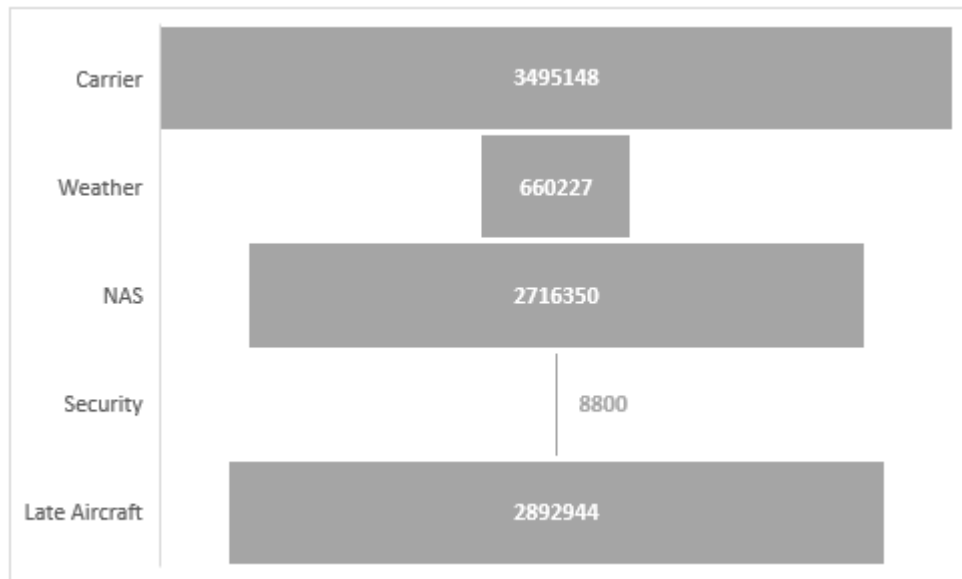


Figure 9

The top 10 origin airports for delay minutes account for 61% of Delta's total delay minutes in 2019. These airports are: Atlanta, Minneapolis, Detroit, La Guardia, JFK, Salt Lake City, Los Angeles, Seattle, Chicago O'Hare, and Boston.

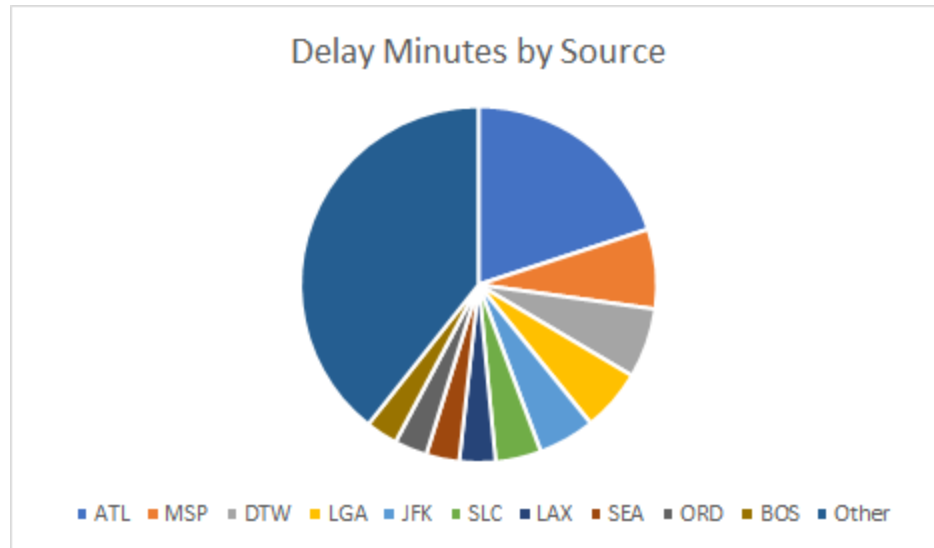


Figure 10

Cancellation Deep-dive

Of 991,986 Delta flights in 2019, 1,842, or .19% were cancelled. As mentioned earlier, the four reasons for flight cancellations are:

- Airline/carrier
- Weather
- National Air System
- Security

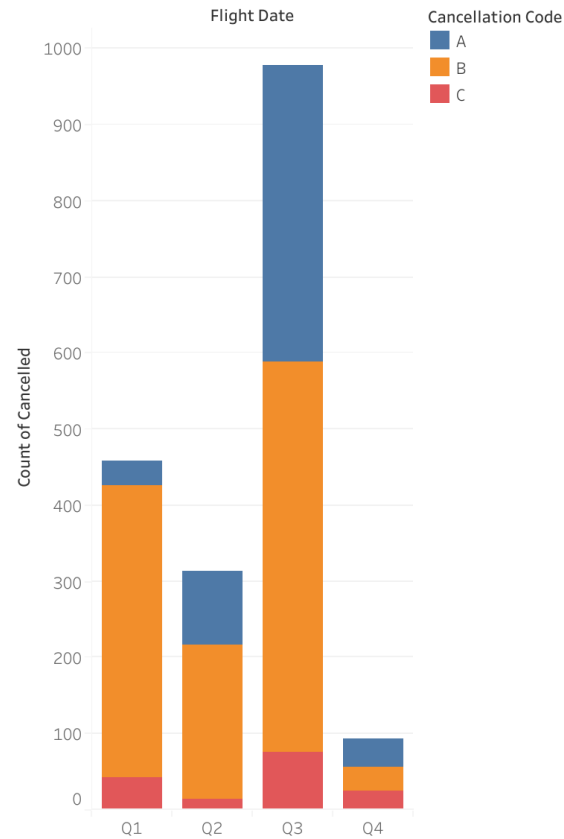
In the dataset, most flights were canceled due to weather, secondly by the airline, and lastly by the National Air System. No flights in 2019 were canceled for security reasons. Below are charts that display visualizations of cancellations.

Flights canceled by quarter show that the most cancellations occurred in the third quarter of 2019, at nearly 1,000 flights canceled. Of these canceled flights, about half were due to harsh weather, followed by airline cancellation and the National Air System. Quarter three of 2019 was the quarter with the most extreme weather in Atlanta, Georgia -- Delta's hub ("Monthly weather forecast in Atlanta, GA"). These weather conditions correspond with Delta's flights canceled due to weather.

The second highest cancellation quarter in 2019 was quarter one, with around 450 cancellations. Of these cancellations, approximately 400 were due to weather, and about 25 each were because of airline and National Air Systems. Quarter one had the second worst weather in Delta's main hub city of Atlanta.

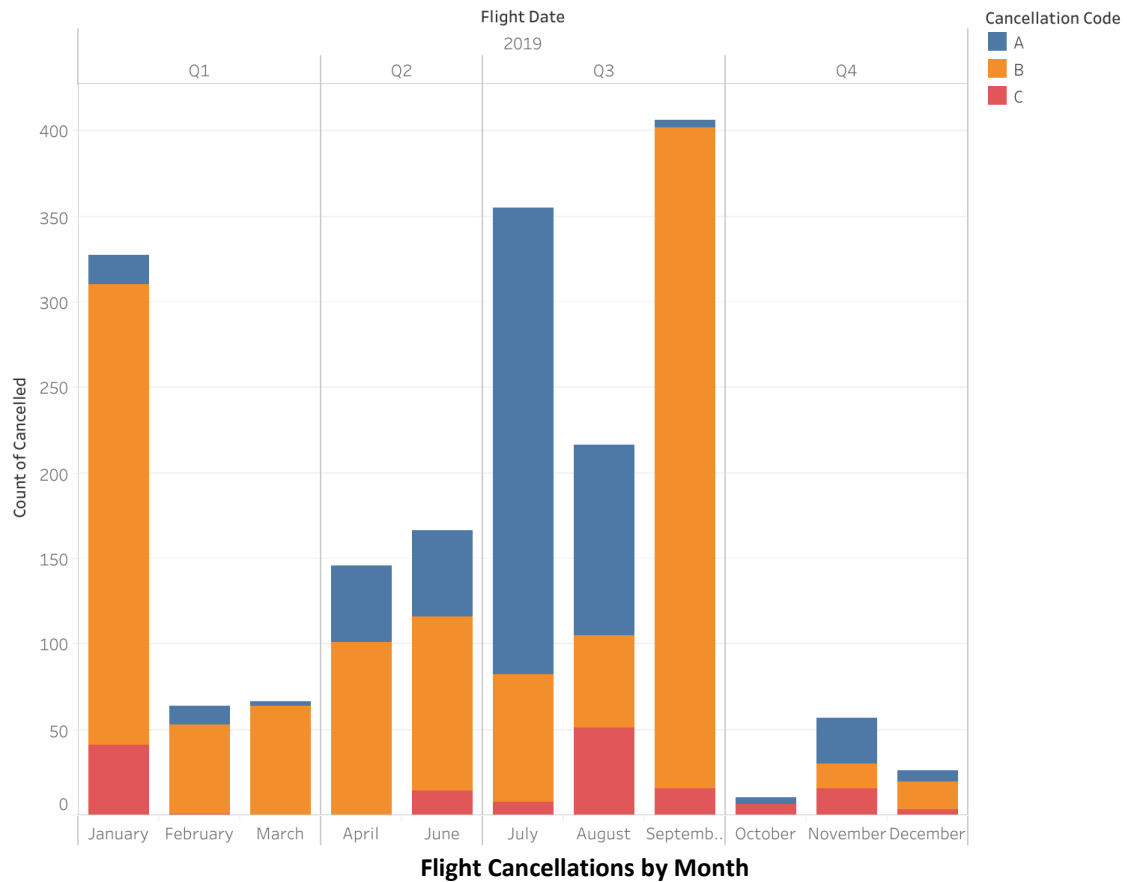
Quarter three saw the third lowest number of cancellations -- just over 300. Approximately 200 of these cancellations were weather related, 75 due to the airline and 25 by the National Air System.

The fourth quarter of 2019 saw the fewest cancellations, at under 100. Each cancellation was split fairly evenly across the quarter.

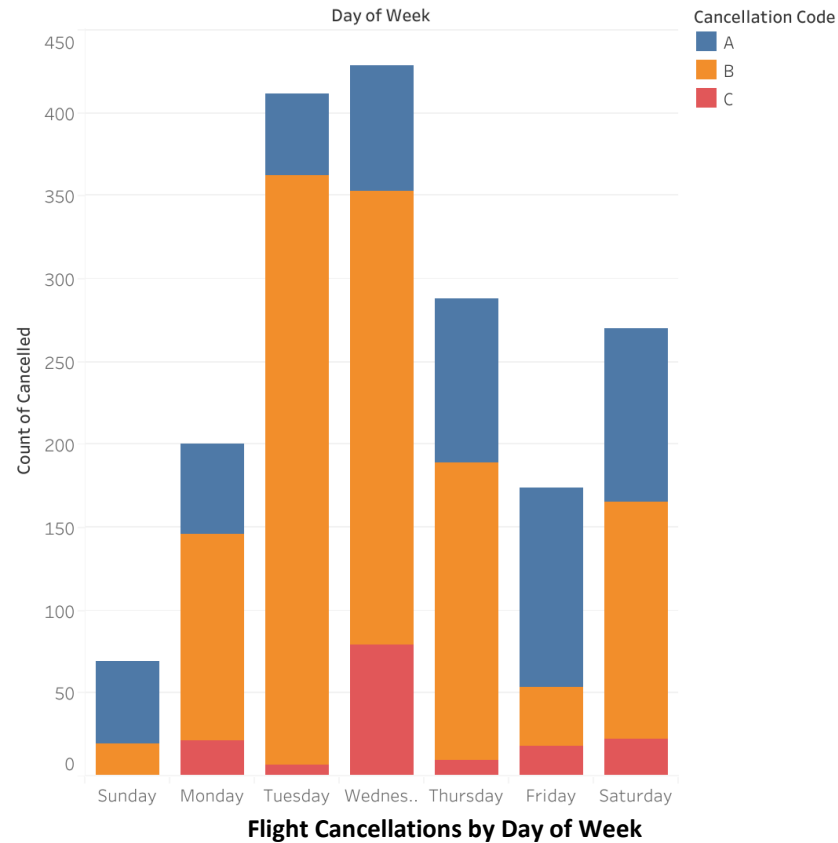


Flight Cancellations by Quarter

Flight cancellations by month provide deeper insight to how months performed against each other. The four highest cancellation months were September, July, January and August. July, August and September comprise quarter three, which was the quarter with the highest number of cancellations. Furthermore, January belongs to quarter one - the quarter with the second highest number of cancellations. October saw the lowest cancellation, followed closely by December and November, which are the months in quarter four.



Flight cancellations by weekday offer understanding of patterns and expectations Delta should have and recognize in the future. Most flights were canceled on Wednesday, followed closely by Tuesday. Wednesday (Tuesday is a close second) is the busiest travel day in the United States because it is the most popular day to travel for business. Delta has the highest volume of flights operating this day, especially on their regional planes. Therefore, most flights being canceled these days is logical. Monday is not a popular day for travel because it is right after weekend travel and before business travel begins, as reflected in cancellations as well.



Flight cancellations by origin and destination looked nearly identical. Most flight cancellations occurred in Delta's major hubs across the United States -- Atlanta, Chicago, Minneapolis, New York and Detroit.

Notable differences between origin and destination cancellations include:

- More cancellations by origin than destination in
 - Salt Lake City
 - Las Vegas
 - Kansas City
 - New York



Flight Cancellations by Origin



Flight Cancellations by Destination

Analytics

Correlations

Correlations between variables were identified using Tableau, SPSS, and Excel.

Time

Delays are highly correlated with June and July. Interestingly, August has the most flights, but not the highest number of delays. Similarly, November has the fewest of every delay type, but the third fewest flights. It may be worth exploring why these mismatches exist. When isolating the two delay types Delta can control (late aircraft and carrier delays), June and July again have the most delays, while November has the fewest.

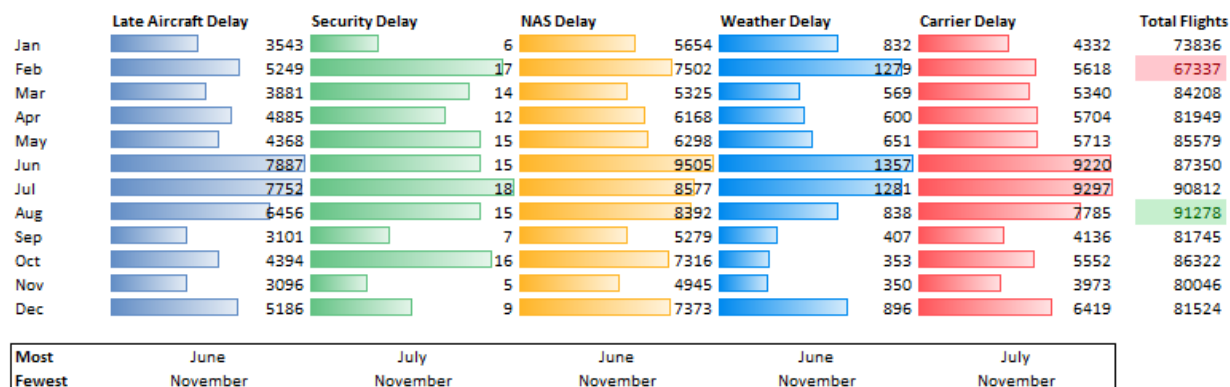


Figure 16

Cancellations are highly correlated with January, July, and September. Again, concentrating on the delay type under Delta's control, carrier delays are by far the most frequent in July and August. This is in line with expectations, because these months have the highest flight volume. May is an outlier, as there were no cancellations for any reason that month, despite it having the fifth most total flights. While this could be a reporting error, this outlier should be examined more deeply to determine whether the reasons for the lack of carrier delays can be replicated.

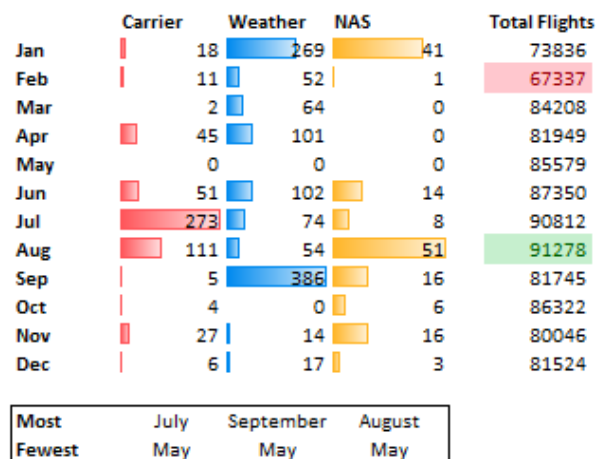


Figure 17

Delays are most highly correlated with Thursdays, but delays are fairly evenly distributed among the days of the week. Thursday has the highest rate of the two controllable delay types, while Saturday has the lowest.

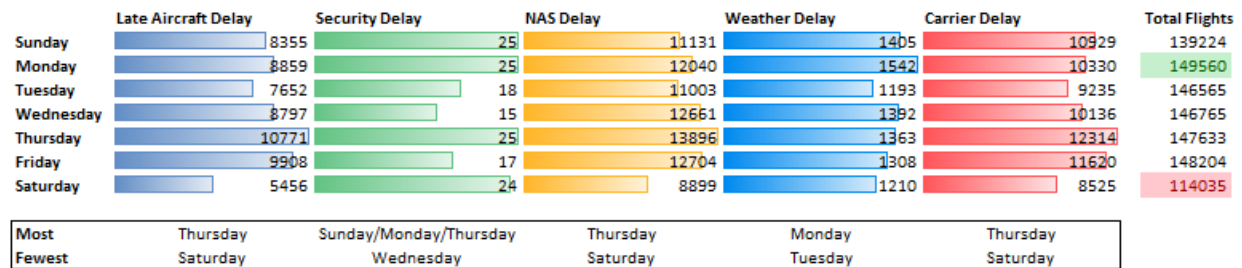


Figure 18

Overall, cancellations are highly correlated with Tuesdays and Wednesdays. However, Fridays and Saturdays see the greatest number of carrier cancellations. Fridays have the second highest flight volume, so a higher number of cancellations is expected. However, Saturday has the lowest flight volume, so should not account for so many carrier delays. The reasons for these cancellations should be explored further.

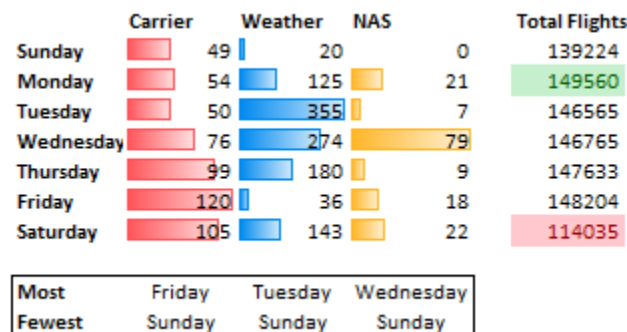


Figure 19

For this purpose, flights were divided into four buckets based on scheduled departure time. The buckets are as follows: Early AM (12am - 7am), Morning (7am - 12pm), Afternoon (12pm - 6pm), and Evening (6pm - 12pm). With respect to delay association to time of day, it is important to remember the effect of delay propagation. Delays can have a cumulative effect as the day progresses. It is also important to note that one type of delay can lead to another type; for example, a security delay at the flight's origin may result in a late aircraft delay at the flight's destination.

All delay types occur most frequently in afternoon flights (12pm - 6pm), with evening flights (6pm - 12am) following closely in the controllable delay types. Early AM flights are least likely to experience delays because they are the first flights of the day and are generally not dependent on preceding flights. This backs up our understanding that delay propagation effects are reset at the end of each day.

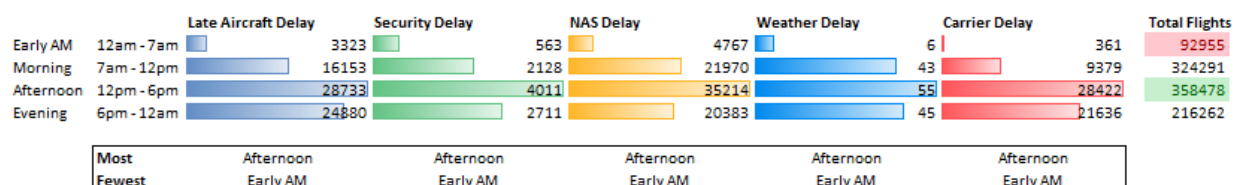


Figure 20

Cancellations are overall most likely to occur in the Afternoon, when flight volume is at its highest. However, when focusing only on carrier cancellations, Evening flights are the most likely to be cancelled. This is potentially due to the effects of delay propagation, with a flight being so delayed that the airline chooses to cancel rather than continue the route.

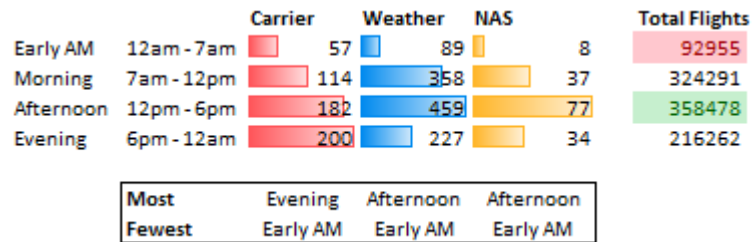


Figure 21

Location

Delays are most associated with the following airports: Atlanta, Minneapolis, Detroit, La Guardia (New York City) and JFK (New York City). They account for 44% of delays in 2019.

Cancellations are most associated with the following airports: Atlanta, Detroit, La Guardia, Minneapolis, and Chicago O'Hare. They account for 48% of cancellations in 2019.

There is significant overlap between the top delay and cancellation airports. That the lists overlap and that these are the top airports is not surprising. All of the airports are very large, extremely busy international airports in large cities and are in areas with extreme weather.

Multivariate

Multifactorial analysis using various combinations of the above variables identified areas where the risk of delay or cancellation is the highest. These measures can both guide further exploration of proprietary data we did not have access to and can be used as KPIs to track improvement progress.

This analysis is intended to show very specifically which flights are most at risk of delays. The first step was to identify the day of the week and range of flight numbers that have the most delays (Figure 22). Next, this group of 4952 was further broken down into times of day (Figure 23). Finally, a filter was applied to determine how many of each type of controllable delay occurs within this window (Figure 24).

	Flight Number						
	1-499	500-999	1000-1499	1500-1999	2000-2499	2500-2999	3000+
Sunday	1172	3020	3687	3935	4162	3766	191
Monday	1234	3029	3873	4188	4263	4081	206
Tuesday	1035	2737	3476	3872	4077	3585	173
Wednesday	1201	3022	3889	4342	4521	4152	179
Thursday	1405	3510	4421	4892	4952	4616	217
Friday	1247	3190	4134	4532	4769	4334	171
Saturday	935	2301	3200	3284	3052	2659	132

Figure 22

Thursday Delays, Flt Nos. 2000-2499

12am - 7am	264
7am - 12pm	1180
12pm - 6pm	1982
6pm - 12am	1526

Figure 23

Thursday, Flt No. 2000-2499, Afternoon flights

LateAircraft Delay	1011	10%
Carrier Delay	992	9%
Total Flights	10625	

Figure 24

Using this process, we determined that 19% of the total Thursday afternoon flights with flight numbers from 2000-2499 are delayed, much higher than Delta's 6 year average of 14% delayed flights. Because it is so specific, this metric is a good KPI. While we know that Thursdays and afternoons are correlated with delays due to high flight volume, we were not able to determine why flight number is significant. Airlines traditionally had fairly strict numbering conventions. For example, 1 and 2 digit flight numbers were for international flights and even flight numbers would be associated with north- and east-bound routes. Today, aside from a few legacy flight numbers, most airlines use software to automatically assign flight numbers each day. We were not able to access flight number assignment information for Delta, so cannot correlate this group of flight numbers to any particular geography or other attribute that would explain the high rate of delays.

The next multivariate analysis delves into delays by month and time of day. June, July, and August have the highest number of delays, all on flights within the Afternoon scheduled departure bucket (Figure 25). Diving deeper into this segment and looking at only the controllable delay types, there are some surprising observations. In (Figure 26), June, July, and August flights are combined. We see the number of minutes of delay, the number of delays, and the average duration of the delay for each scheduled departure time bucket. The highest average delays, in orange, and the lowest, in yellow, are highlighted. We can see that most carrier delays over the summer months occur in the evening, which is expected due to delay propagation. Most late aircraft delays occur on afternoon flights, likely for similar reasons. Interestingly, both items are associated with the lowest average delay duration. Conversely, early morning flights experience the fewest delays, but the longest delay durations, with late aircraft delays being an average of 2 hours long. Undoubtedly, these delays are causing delay propagation and are likely responsible for many of the delays in subsequent departure time buckets.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
12am - 7am	624	774	579	568	414	569	546	488	410	638	649	783
7am - 12pm	2768	3284	2603	2504	2501	3531	3154	3114	2069	2962	2546	3518
12pm - 6pm	3754	5053	3724	4824	4676	7190	7248	6303	3476	4744	3092	4744
6pm - 12am	2257	3072	2838	3129	3602	5788	5744	5002	2752	3084	1930	3399

Figure 25

Jun, Jul, Aug

	CarrierDelay_Min	CarrierDelay	Avg Delay	LateAircraftDelay_Min	LateAircraftDelay	Avg Delay
12am - 7am	85716	995	86	16360	134	122
7am - 12pm	286750	5287	54	144741	2910	50
12pm - 6pm	433382	9905	44	489212	9925	49
6pm - 12am	484458	10115	48	487767	9126	53

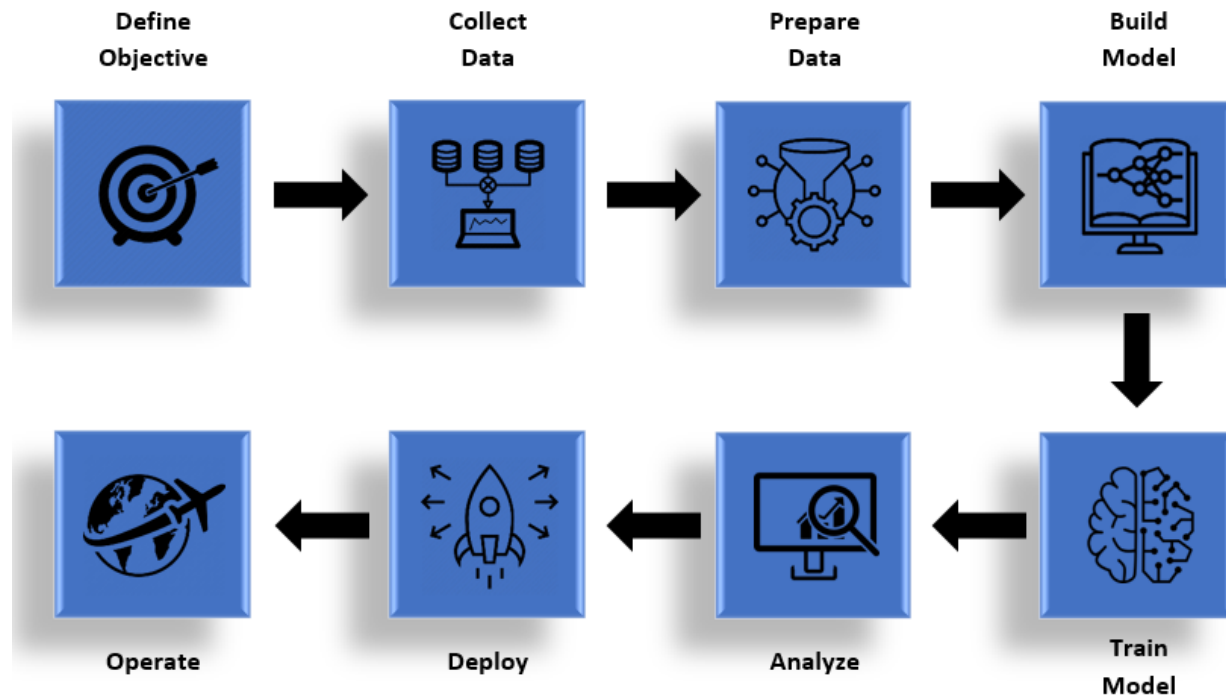
Figure 26

Predictive analytics

To predict if a flight will be delayed or canceled, we have created a Machine Learning Model with data collected from the Bureau of Transportation Statistics which holds information such as flight date, departure time, arrival time,

destination, origin as well as other flight information. The model is a type of supervised learning which uses historical flight data for prediction.

The framework to develop the Machine Learning Model follows the next steps (Linthicum):



To reduce the size of the predictive analytics section, we explain this process for the Arrival Delay Model Prediction; however, other models were created for Departure Delay and Cancellations, but those are summarized in section IX - "Other Models Summary." It is important to highlight that these models followed the same approach.

The objective is to create and deploy a Machine Learning model that predicts if a flight will be cancelled or delayed by using Azure Machine Learning to build and train the model; additionally, Python to prepare our data.

As it was mentioned on Data Source section, data was collected from the Bureau of Transportation Statistics which holds 109 features of historical data related to time period, airline, origin, destination, departure performance, arrival performance, cancellations and diversions, flight summaries, cause of delay, gate return information origin airport, and diverted airport information.

Not all data was used during the machine learning model, so the most descriptive 29 features for our objective were selected. In addition, dummy variables or feature engineering were created from departure delay and arrival delay to create departure delay FAA and arrival delay FAA increasing the features from 29 to 31. A flight delay is when an airline flight takes off and/or lands later than its scheduled time. The Federal Aviation Administration (FAA) considers a flight to be delayed when it is 15 minutes later than its scheduled time. So, the value of 1 represents those flights delayed with the latter description, and a value of 0 for those less than 15 minutes. Data was collected for all Delta Airlines flights from 2019 with 991, 986 observations and a size of almost 100 MB as a CSV file.

The screenshot shows the Bureau of Transportation Statistics website. The header includes the U.S. Department of Transportation logo and navigation links: Topics and Geography, Statistical Products and Data, National Transportation Library, and Newsroom. A search bar is present on the right. Below the header, the 'OST-R > BTS' breadcrumb is visible. The main content area features the 'TranStats' logo and a search bar. To the left of the search bar are links for 'Resources' (Database Directory, Glossary, Upcoming Releases, Data Release History) and 'Data Tools' (Analysis, Table Profile, Table Contents). The central panel is titled 'On-Time : Reporting Carrier On-Time Performance (1987-present)' and includes a 'Download Instructions' link, a 'Latest Available Data: February 2020' notice, and filters for Geography (All), Year (2020), and Period (January). Below these are checkboxes for 'Prezippped File', '% Missing', 'Documentation', and 'Terms'. A 'Download' button is on the right. A table lists available fields with their descriptions and links to 'Get Lookup Table'.

Field Name	Description	Support Table
<input type="checkbox"/> Time Period		
<input type="checkbox"/> Year	Year	
<input type="checkbox"/> Quarter	Quarter (1-4)	Get Lookup Table
<input type="checkbox"/> Month	Month	Get Lookup Table
<input type="checkbox"/> DayOfMonth	Day of Month	
<input type="checkbox"/> DayOfWeek	Day of Week	Get Lookup Table
<input type="checkbox"/> FlightDate	Flight Date (yyyymmdd)	
<input type="checkbox"/> Airline		
<input type="checkbox"/> Reporting_Airline	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.	Get Lookup Table
<input type="checkbox"/> DOT_ID_Reporting_Airline	An identification number assigned by US DOT to identify a unique airline. (Source: DOT's carrier database)	Get Lookup Table

The Bureau of Transportation Statistics data offers a very clean and complete dataset, yet data preparation was done in Python and Azure Machine Learning. First, data exploratory analysis was developed in Python to look into the shape and data types, missing values, and useful features for our model (Lianne & Justin). Then, in Azure Machine Learning we selected the best features to build and train our model.

We started by importing all the libraries and defining some parameters.

```
# Importing all the libraries needed
import seaborn as sns
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from heatmap import heatmap, corrplot
import matplotlib
plt.style.use('ggplot')
from matplotlib.pyplot import figure

%matplotlib inline
matplotlib.rcParams['figure.figsize'] = (18,14)

pd.options.mode.chained_assignment = None
```

We read the dataset from a cloud service where all data is stored.

```
# Reading the dataset
data = pd.read_csv('C:/Users/RC_x/OneDrive/ASU MSBA/CIS-SCM 593 Capstone Project/airline-delay-and-cancellation-data-2009-201
```

Then, it was important to check the data shape and type.

```
# Shape and data types of the data
print(data.shape)
print(data.dtypes)
```



```
(991986, 31)
FlightDate          object
Reporting_Airline    object
Flight_Number_Reporting_Airline  int64
Origin              object
Dest                object
CRSDepTime          int64
DepTime             float64
DepDelay            float64
DepDelayFAA         int64
TaxiOut             float64
WheelsOff           float64
WheelsOn            float64
TaxiIn              float64
CRSArrTime          int64
ArrTime             float64
ArrDelay            float64
ArrDelayFAA         int64
Cancelled            int64
CancellationCode     object
Diverted            int64
CRSElapsedTime      int64
ActualElapsedTime    float64
AirTime             float64
Distance            int64
CarrierDelay         float64
WeatherDelay         float64
NASDelay            float64
SecurityDelay        float64
LateAircraftDelay    float64
Month               int64
DayOfMonth           int64
dtype: object
```

From this output, we learned that the dataset has 991,986 observations and 31 features. We have numeric types (integer, float) and categorical (object). We verified in Azure ML that these were set correctly as booleans, numeric, float, or string types. We analyzed which features were numeric and which were not numeric.

```
# Selecting numeric columns
data_numeric = data.select_dtypes(include=[np.number])
numeric_cols = data_numeric.columns.values
print(numeric_cols)

['Flight_Number_Reporting_Airline' 'CRSDepTime' 'DepTime' 'DepDelay'
 'DepDelayFAA' 'TaxiOut' 'WheelsOff' 'WheelsOn' 'TaxiIn' 'CRSArrTime'
 'ArrTime' 'ArrDelay' 'ArrDelayFAA' 'Cancelled' 'Diverted'
 'CRSElapsedTime' 'ActualElapsedTime' 'AirTime' 'Distance' 'CarrierDelay'
 'WeatherDelay' 'NASDelay' 'SecurityDelay' 'LateAircraftDelay' 'Month'
 'DayOfMonth']
```

```
# Selecting non numeric columns
data_non_numeric = data.select_dtypes(exclude=[np.number])
non_numeric_cols = data_non_numeric.columns.values
print(non_numeric_cols)

['FlightDate' 'Reporting_Airline' 'Origin' 'Dest' 'CancellationCode']
```

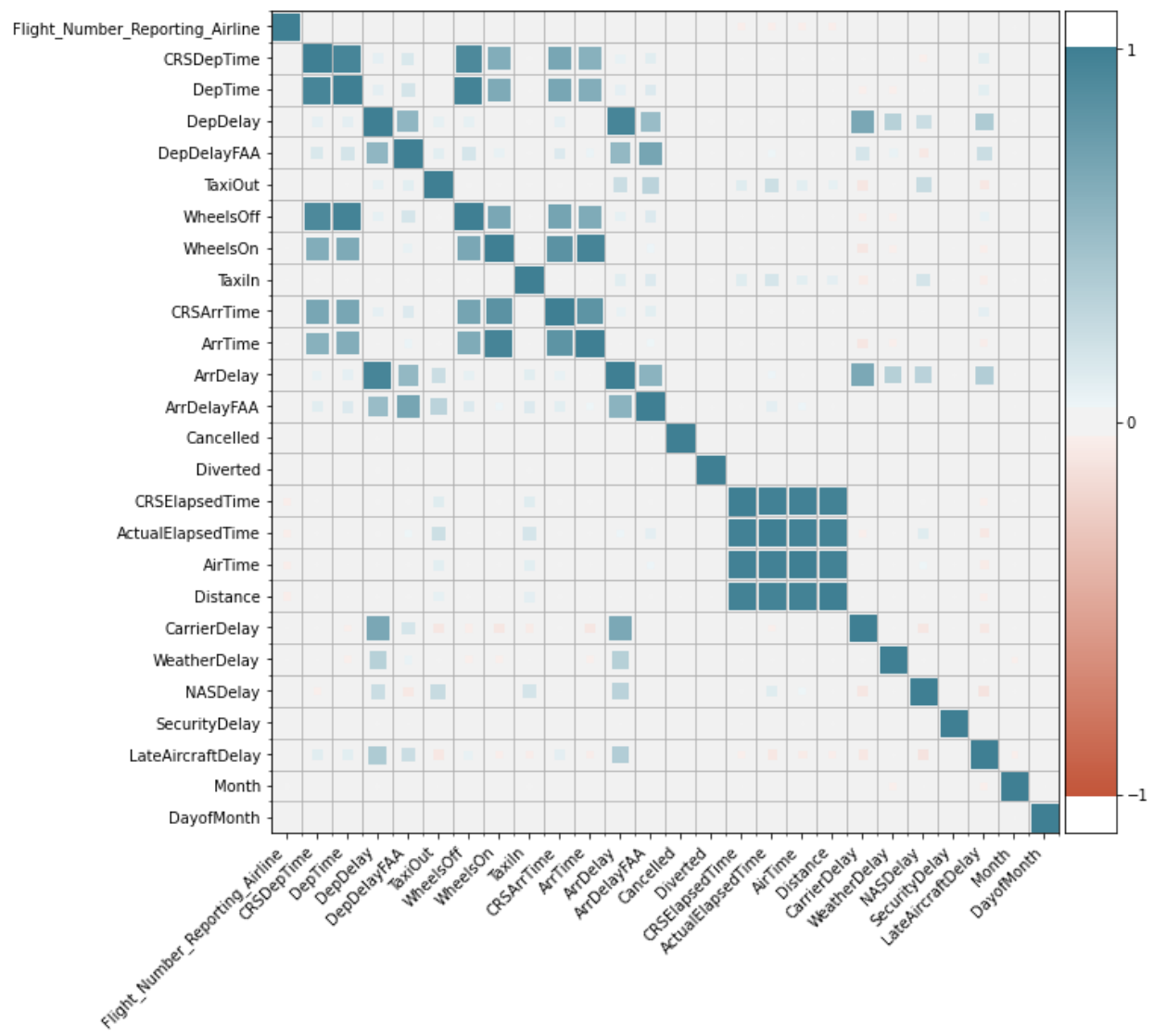
From these results, we identified we have 26 numeric and 5 categorical variables. This was useful information when we started building our model into Azure Machine Learning. Next, we saw how our data looks.

```
# Showing how data looks
data.head(5)
```

	FlightDate	Reporting_Airline	Flight_Number_Reporting_Airline	Origin	Dest	CRSDepTime	DepTime	DepDelay	DepDelayFAA	TaxiOut	...	ActualElaps
0	1/1/2019	DL	14	TPA	ATL	1445	1439.0	-6.0	0	14.0	...	
1	1/1/2019	DL	15	ATL	TPA	1951	1948.0	-3.0	0	13.0	...	
2	1/1/2019	DL	31	ATL	DFW	1914	1917.0	3.0	0	14.0	...	
3	1/1/2019	DL	32	DFW	ATL	1610	1606.0	-4.0	0	13.0	...	
4	1/1/2019	DL	40	LAX	JFK	920	915.0	-5.0	0	16.0	...	

As it was mentioned before, there are some features/variables that are more important than others for Machine Learning. In the correlation section, we explained some of those variables, so it is useful plotting a correlation matrix heatmap to see which features we will be selecting.

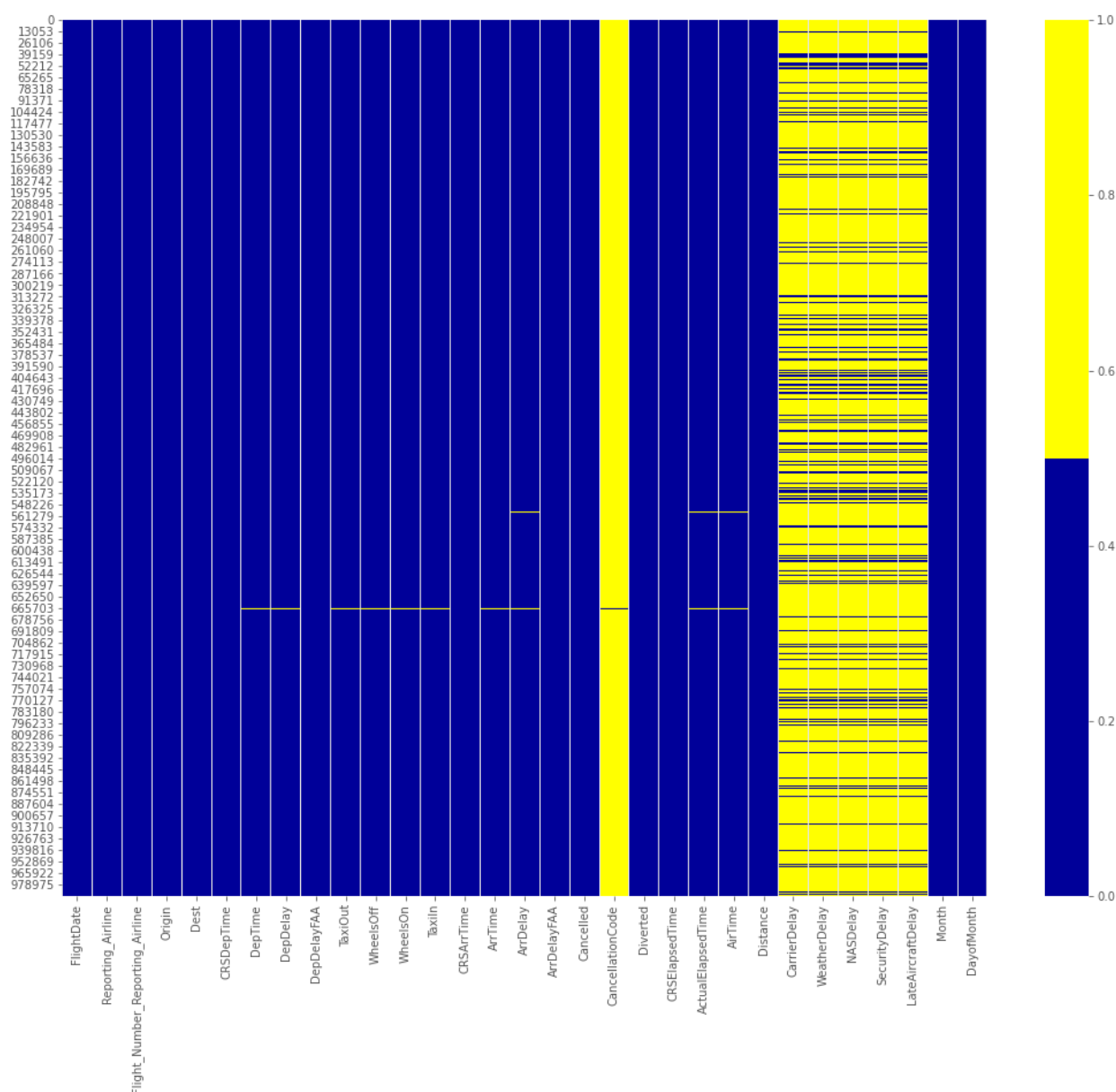
```
# Plotting a correlation matrix heatmap
plt.figure(figsize=(10, 10))
corrplot(data.corr(), size_scale=300, marker = 's');
```



Green means positive and red means negative correlation. The bigger the square and the more emphasized the color, the stronger the correlation between the variables (Zaric). So, if we wanted to predict flights which will be arriving delayed, what are the three variables more correlated with ArrDelayFAA?. As it is illustrated in the correlation matrix heatmap, the three top variables are DepDelay, DepDelayFAA, and ArrDelay, so DepDelay is a great variable to predict a flight delay because it happens before than ArrDelayFAA. If a flight arrives delayed; consequently, it will depart delayed, and so on.

Despite the data being clean and complete in the Bureau of Transportation Statistics website, it was a good idea to take a look for missing values. To see this graphically, we displayed a missing values heatmap (Lianne & Justin).

```
# Specifying the colours - yellow is missing. blue is not missing.
colours = ['#000099', '#ffff00']
# Plotting the heatmap with missing values
sns.heatmap(data.isnull(), cmap=sns.color_palette(colours))
```



The chart above represents missing data patterns in the data set. The X axis represents the feature name whereas the Y axis represents the number of observations (Lianne & Justin). We observed there are 6 features with plenty of missing values, but why are there so many missing values for those features even if the dataset was supposedly clean and complete? The answer is simple, Cancellationcode has values only when a flight is cancelled and CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, and LateAircraftDelay have values when a flight is delayed. Having more than 85% of flights on time, we did not find values for those features. Also, these features were not useful for our model because they happen when a flight is already delayed or cancelled.

Saying that, how would the data look if we drop those features and others that might not help our model?

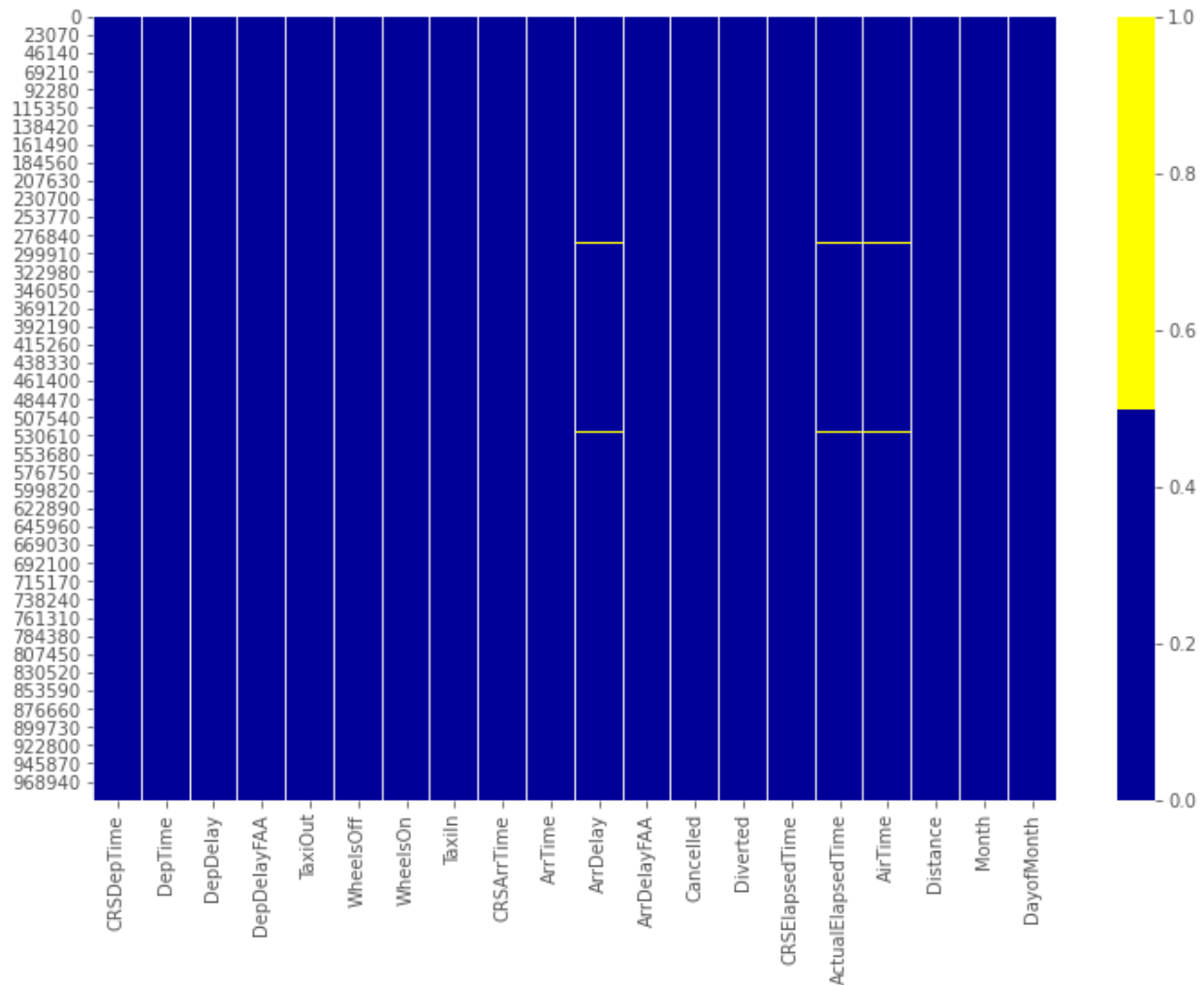
```
# Dropping some features that might not be needed in our Machine Learning model
cols = data.drop(columns=['FlightDate', 'Reporting_Airline', 'Flight_Number_Reporting_Airline', 'Origin', 'Dest',
                          'CancellationCode', 'CarrierDelay', 'WeatherDelay', 'NASDelay', 'SecurityDelay',
                          'LateAircraftDelay'])

cols.head(3)
```

	CRSDepTime	DepTime	DepDelay	DepDelayFAA	TaxiOut	WheelsOff	WheelsOn	TaxiIn	CRSArrTime	ArrTime	ArrDelay	ArrDelayFAA	Cancelled	Div
0	1445	1439.0	-6.0	0	14.0	1453.0	1600.0	4.0	1621	1604.0	-17.0	0	0	
1	1951	1948.0	-3.0	0	13.0	2001.0	2105.0	3.0	2119	2108.0	-11.0	0	0	
2	1914	1917.0	3.0	0	14.0	1931.0	2023.0	13.0	2042	2036.0	-6.0	0	0	

Next, our missing values dataset heatmap looks like this:

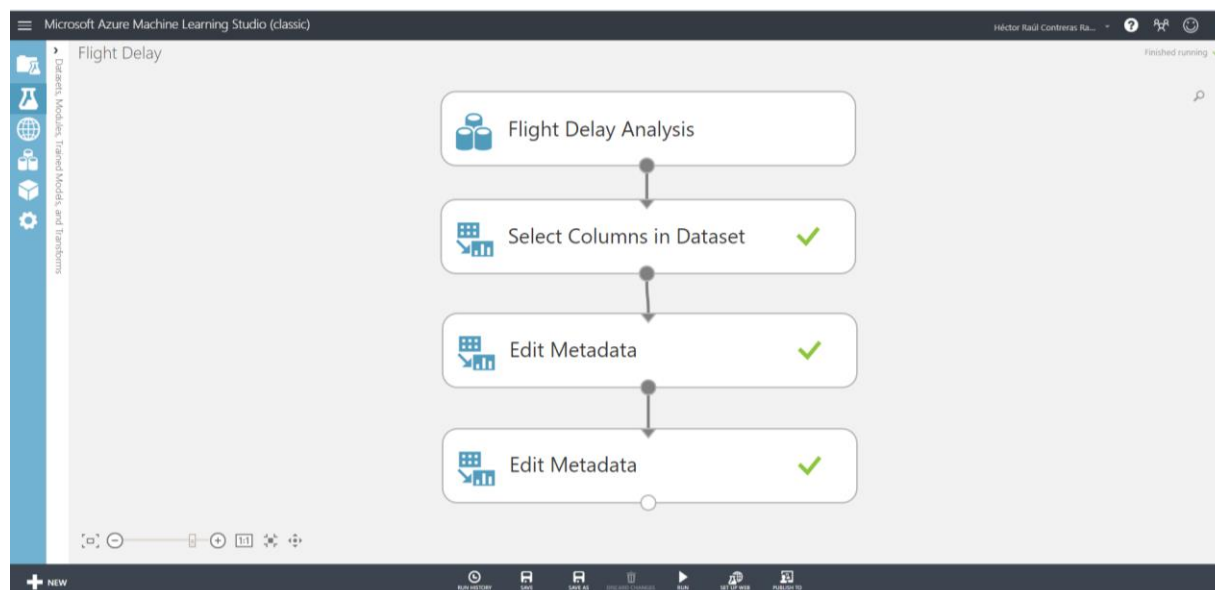
```
# Specifying the colours - yellow is missing. blue is not missing.
colours = ['#000099', '#ffff00']
# Plotting the heatmap with missing values
sns.heatmap(cols.isnull(), cmap=sns.color_palette(colours))
```



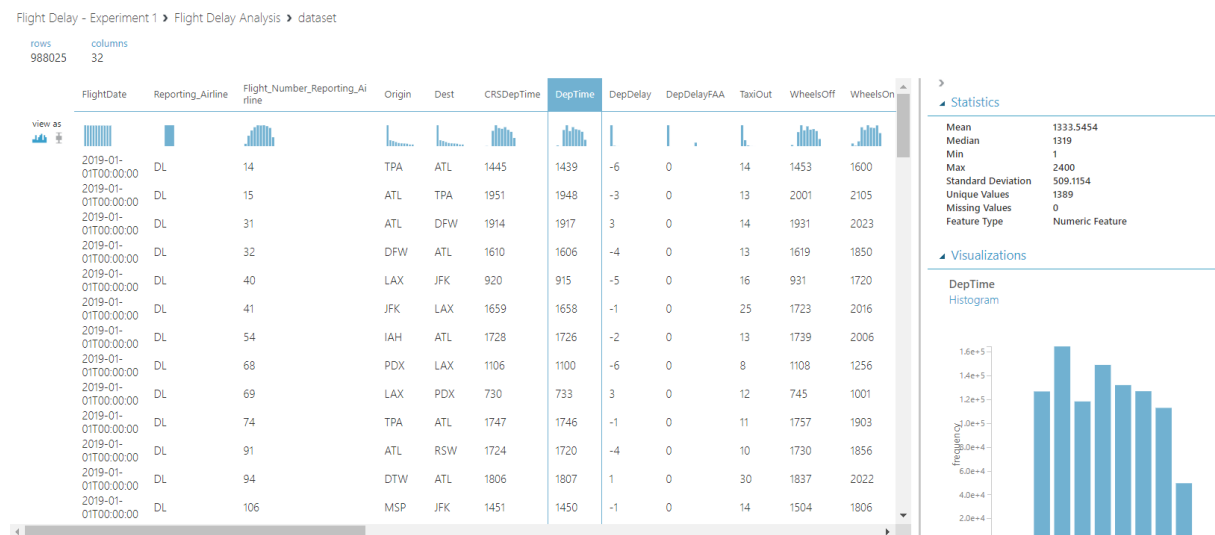
This heatmap shows a cleaner dataset, but there are still some observations with missing values. We can execute listwise deletion technique (dropping the observations), replace the missing values, or impute the missing values; however, this is not needed because those missing values are related to delay and cancellation. For instance, if a flight is cancelled, there are no values in some features such as ArrDelay, ActualElapsedTime, or AirTime.

We performed our data exploratory analysis to understand more of our data, so we could see the shape and data types, detected missing values, and identified some useful features for our model. We did not convert the final dataset because we will select our useful features in Azure Machine Learning.

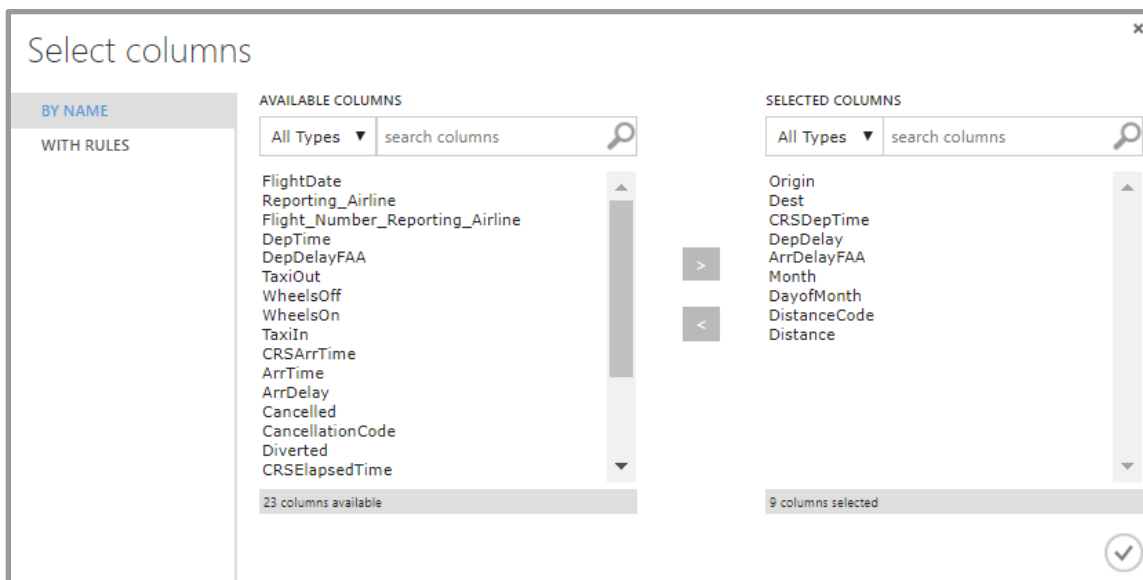
We created two experiments; one for flight delay and other for cancelled. In arrival flight delay, our data was prepared as following with the dataset uploaded, one select column in dataset module, and two edit metadata modules:



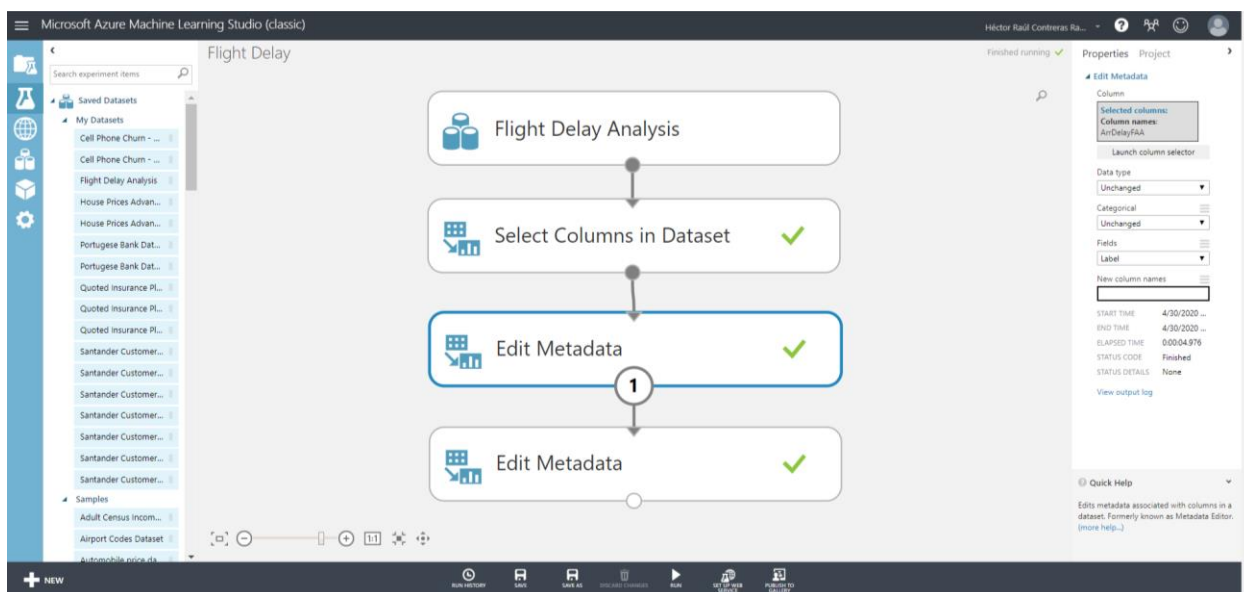
During our data exploratory analysis, we observed there were 5 categorical and 26 numeric features, but we added one more feature here, DistanceCode, which we thought is an important feature for our model. Additionally, observations related to cancelled flights were dropped. Let's visualize our dataset in Azure Machine Learning.

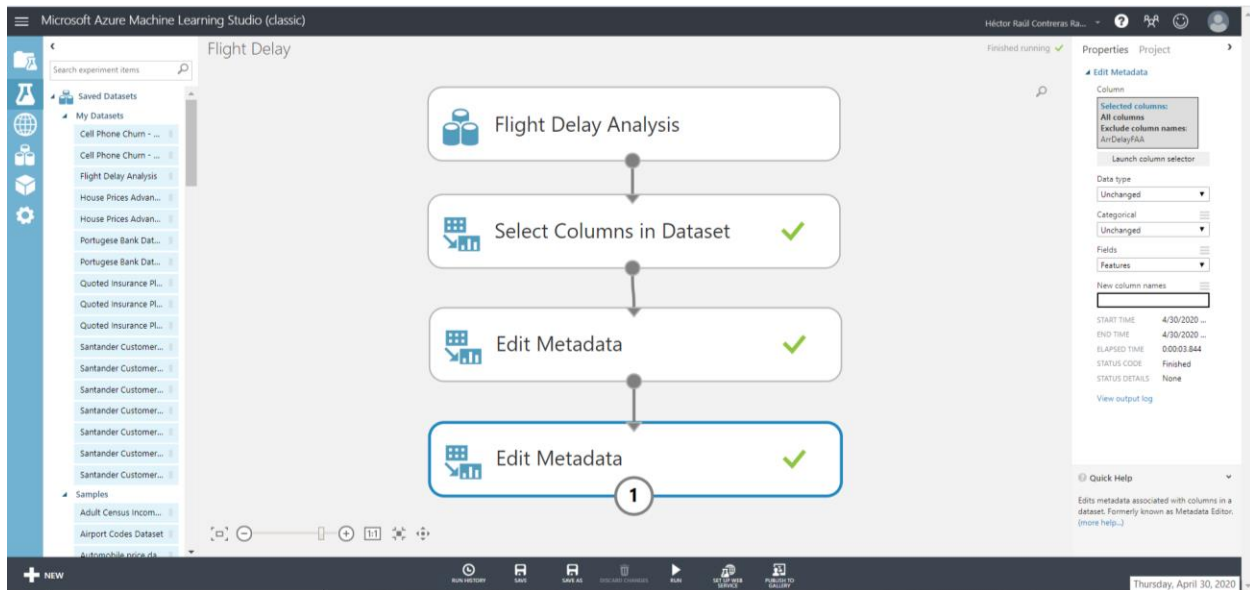


The features that were selected in our model were Origin, Dest, CRSDepTime, DepDelay, ArrDelayFAA, Distance, DistanceCode, Month, and DayMonth which are critical for a flight to arrive delayed and have a positive correlation with the variable ARRDelayFAA which is our label or target.



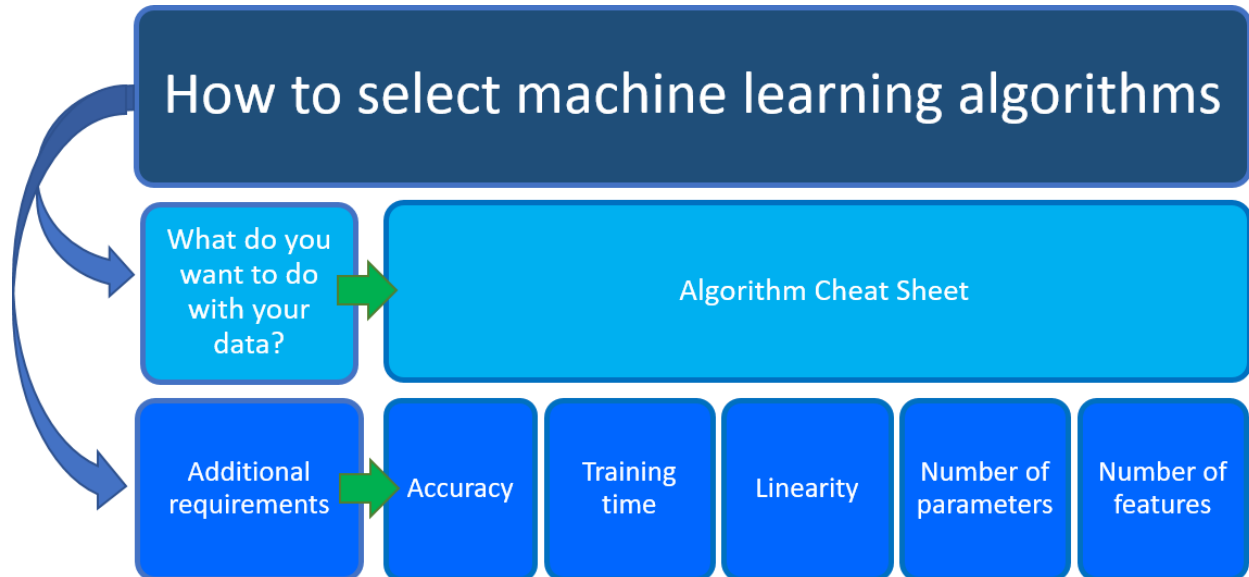
Finally, in our Metadata Editor modules, we defined ArrDelayFAA as our label or target and the other variables as features. Data preparation takes time, and it was one of the most important steps for our Predictive Analytics stage in order to offer a prediction to Delta Airlines.



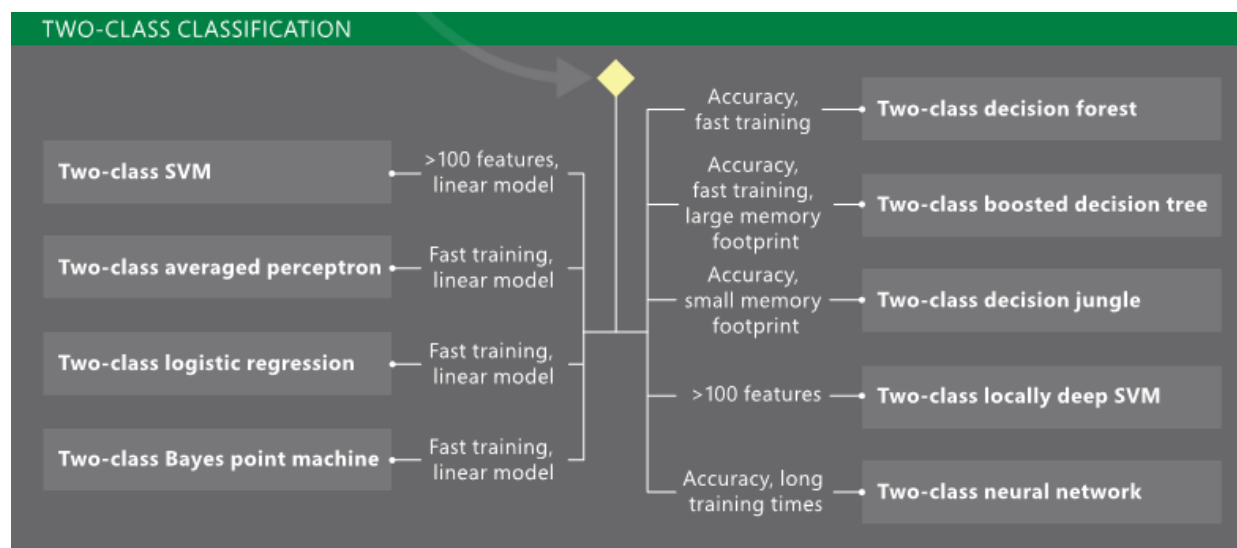


Building the model

From this point, we prepared our data and selected our label and features to train our model. As it was mentioned, the model is a type of supervised learning which uses historical flight data for prediction. In order to select the best algorithm for the model, we analyzed our business scenario and decided based on this flow by Microsoft (“How to select algorithms for Azure Machine Learning”):



Therefore, what do we want to do with our data? Well, we want to predict between two categories, so the model we used is a two-class classification algorithm to predict if a flight is delayed or not. There are different types of two-class classification algorithms, but these are the options (“Machine Learning Algorithm Cheat Sheet for Azure Machine Learning designer”):



The algorithms we selected are Two-class logistic regression, Two-class decision forest, Two-class SVM, Two-class boosted decision tree, and Two-class decision jungle.

For the additional requirements (“How to select algorithms for Azure Machine Learning”), we have the following brief explanation of selection which will be fully covered in their respective sections.

Training Time

To measure the effectiveness of the model we used accuracy which is a good metric when evaluating classification problems. Accuracy is the proportion of true results to total cases, and here is the formula:

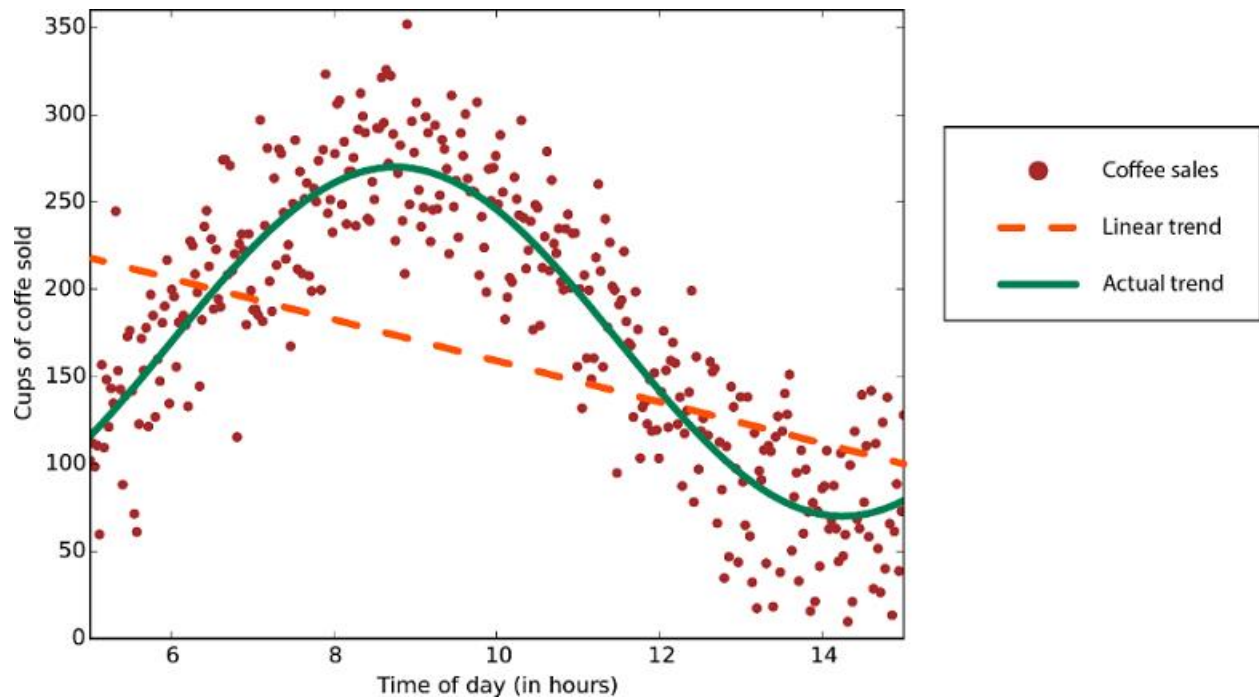
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Training Time

We selected 4 fast training algorithms and a Two-class SVM during our experiment. The number of hours or minutes varies from one algorithm to another.

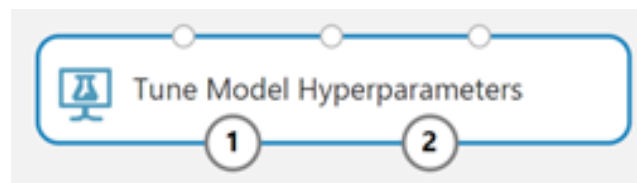
Linearity

Linearity means that there is a linear relationship between a variable and a constant, so we selected two algorithms with linear model as a strategy: Two-class SVM and Two-class logistic regression. Nonetheless, relying on a linear classification algorithm would result in low accuracy. This is an example of a nonlinear data points:



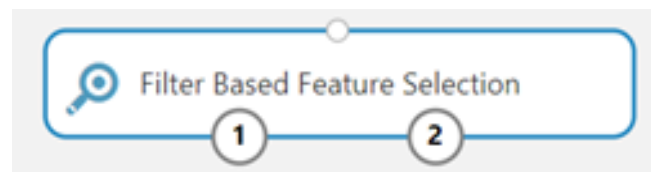
Number of parameters

Parameters for the best first strategy algorithms were changed as well as some tune model hyperparameter modules were used as another strategy.



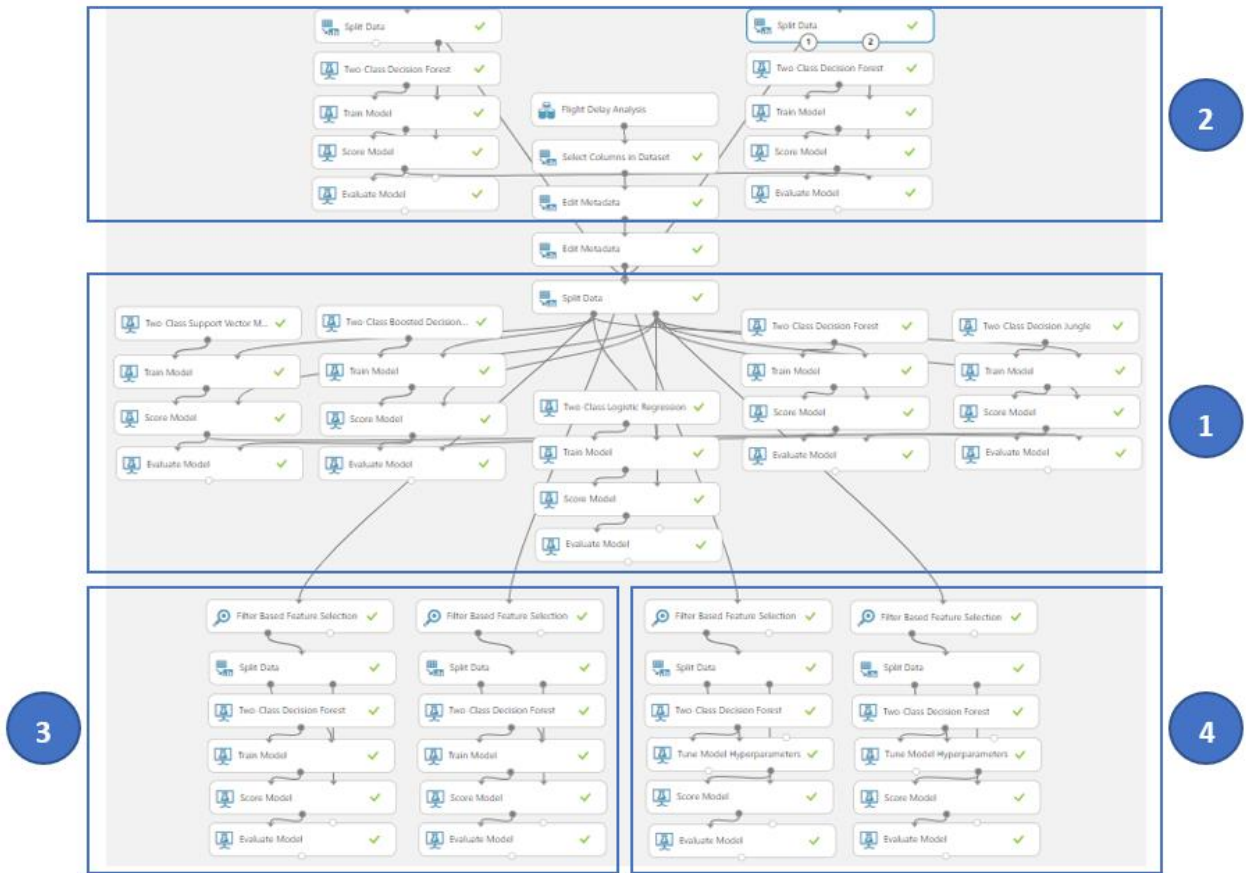
Number of features

Feature selection was implemented in some strategies even if there were not enough features. The module was implemented to find a better performance and improvements during our experiment.



Delay Experiments and strategies

The model was built in two different experiments and 4 strategies, this is the Machine Learning Model experiment 1:



This is the Machine Learning Model experiment 2:



Strategy 1: Training different algorithms and selection of the two best algorithms. From here, Experiment 1 and Experiment 2 were built from these two best algorithms. The next strategies are

Strategy 2: Best algorithm, splitting the data with 75% and 80%, and parameters modifications.

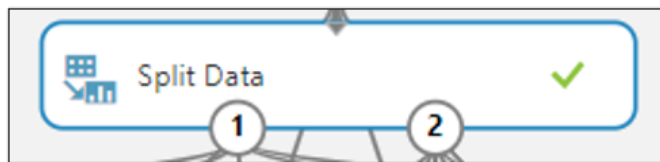
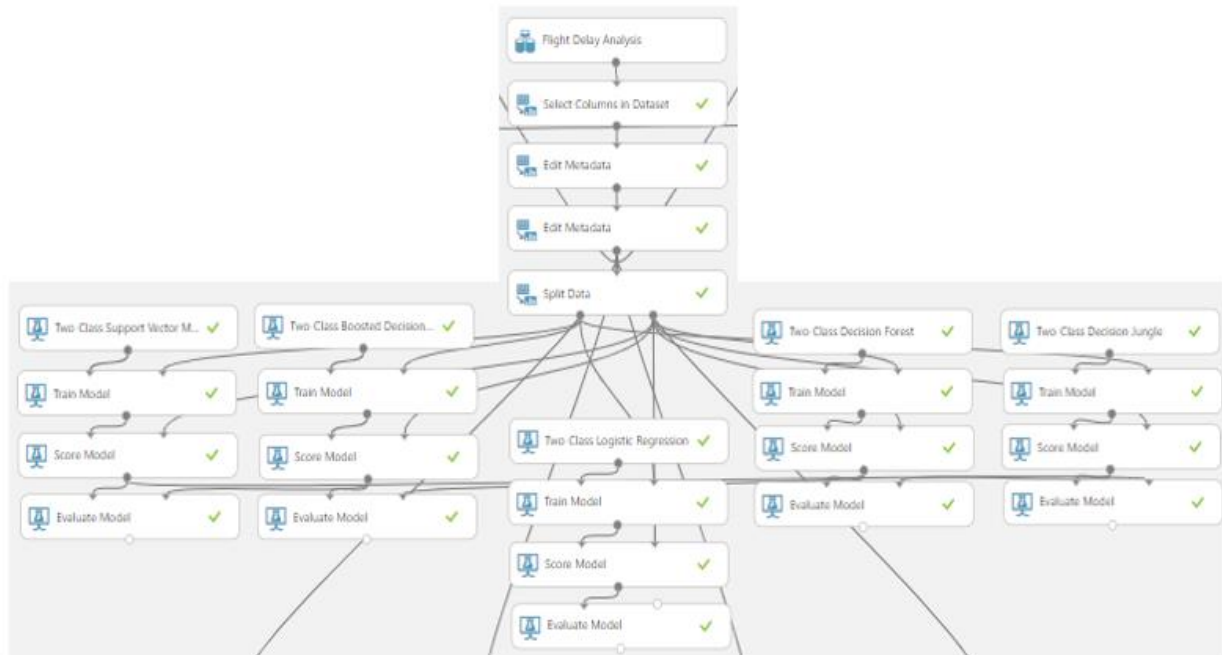
Strategy 3: Best algorithm, best data split, best parameters modifications, and feature selection (6 and 7 features).

Strategy 4: Best algorithm, best data split, best parameters modifications, best feature selection, and hyperparameter tuning with 10 and 20 number of runs on random sweep.

Training the model

Strategy 1 – Experiment 1 and 2

Five algorithms were trained with their predefined parameters, data was split on 75% for training data and 25% for testing data for validation, and label was stratified.



Properties Project

Split Data

Splitting mode

Split Rows

Fraction of rows in the first...

0.75

☒ Randomized split

Random seed

0

Stratified split

True

Stratification key column

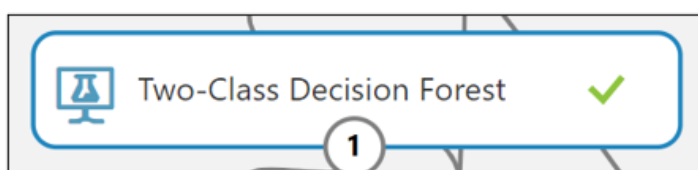
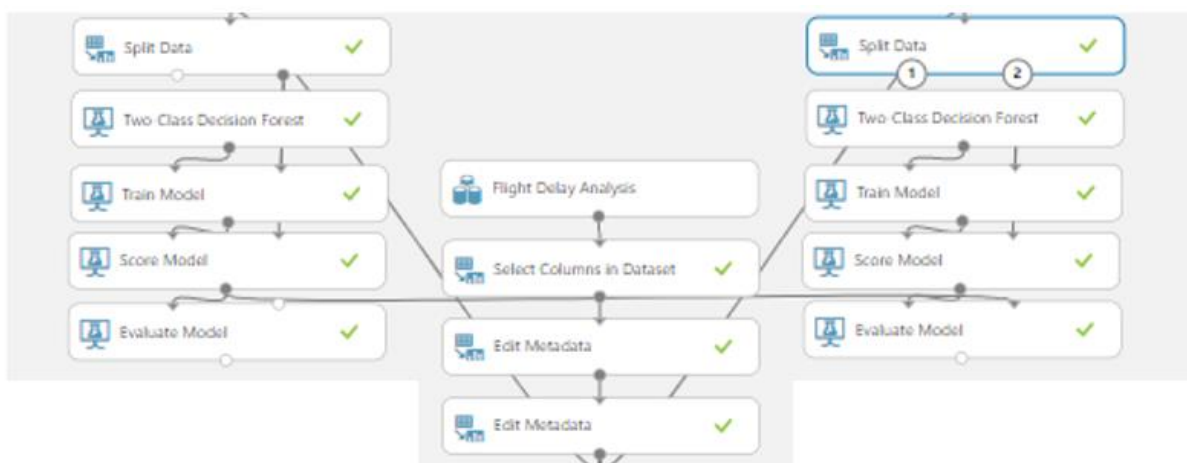
Selected columns:

All labels

Launch column selector

Strategy 2 – Experiment 1

One of the two best algorithms was split with 75% and 80% for training data. The parameters were increased in the number of decision trees from 8 to 16 and the maximum depth from 32 to 64.



Properties Project

Two-Class Decision Forest

Resampling method

Bagging

Create trainer mode

Single Parameter

Number of decision trees

16

Maximum depth of the de...

64

Number of random splits ...

128

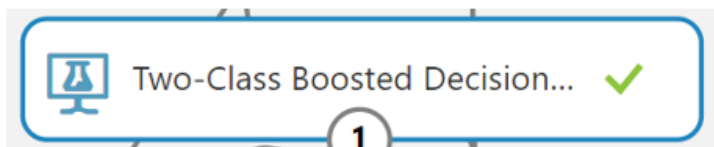
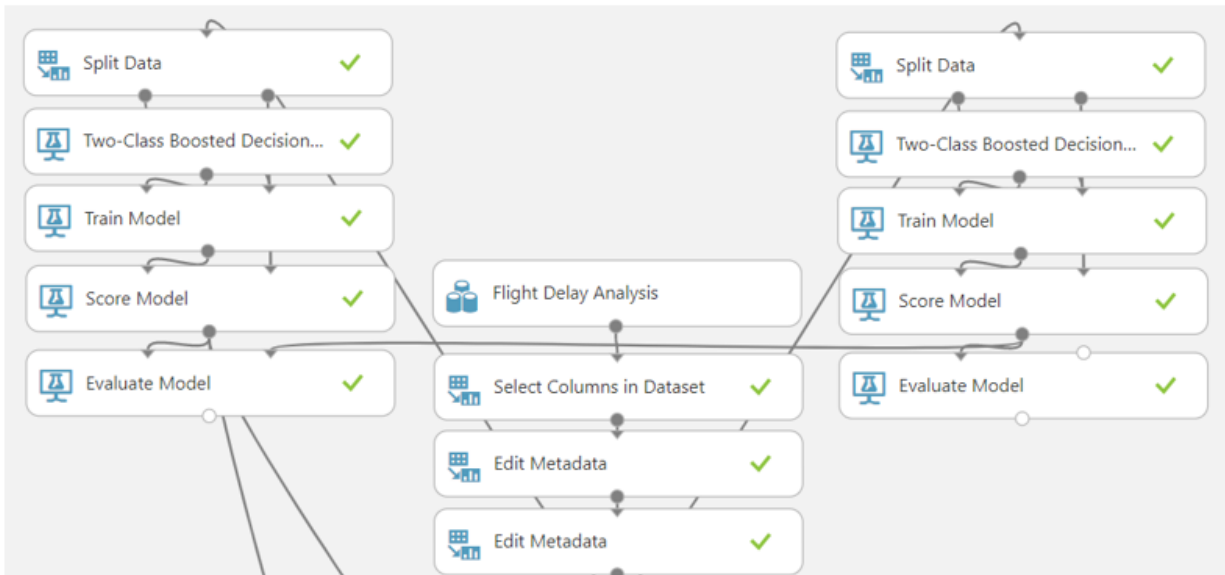
Minimum number of sam...

1

☒ Allow unknown values...

Strategy 2 – Experiment 2

One of the two best algorithms was split with 75% and 80% for training data. The parameters were increased in the maximum number of leaves per tree from 20 to 40 and number of trees constructed from 100 to 200.



Properties Project

Two-Class Boosted Decision Tree

Create trainer mode

Single Parameter ▼

Maximum number of leav...

40

Minimum number of sam...

10

Learning rate

0.2

Number of trees construct...

200

Random number seed

☒ Allow unknown categ...

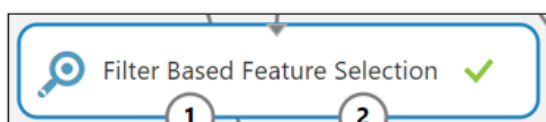
Strategy 3 – Experiment 1 and 2

The best model from strategy 1 and 2 was combined with feature selection of 6 and 7 desired features through the Pearson Correlation method.

Experiment 1



Experiment 2



Properties Project

Filter Based Feature Selection

Feature scoring method

Pearson Correlation

☒ Operate on feature co...

Target column

Selected columns:

Column names:

ArrDelayFAA

Launch column selector

Number of desired features

6

START TIME 4/27/2020 ...

END TIME 4/27/2020 ...

ELAPSED TIME 0:00:00.000

STATUS CODE Finished

STATUS DETAILS Task output was present in output cache

Properties Project

Filter Based Feature Selection

Feature scoring method

Pearson Correlation

☒ Operate on feature co...

Target column

Selected columns:

Column names:

ArrDelayFAA

Launch column selector

Number of desired features

7

START TIME 4/27/2020 ...

END TIME 4/27/2020 ...

ELAPSED TIME 0:00:00.000

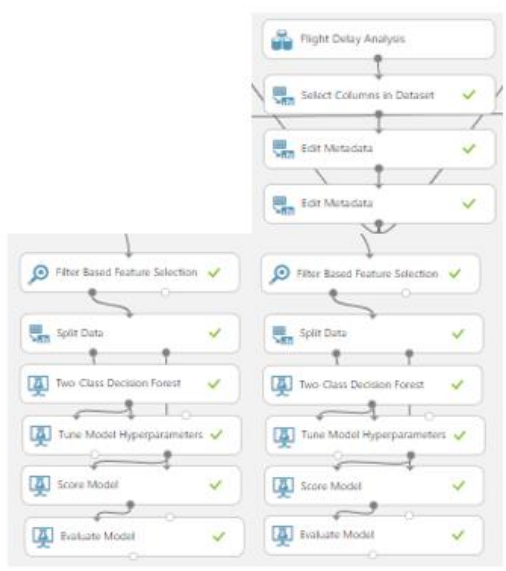
STATUS CODE Finished

STATUS DETAILS Task output was present in output cache

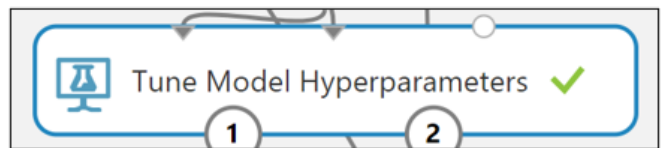
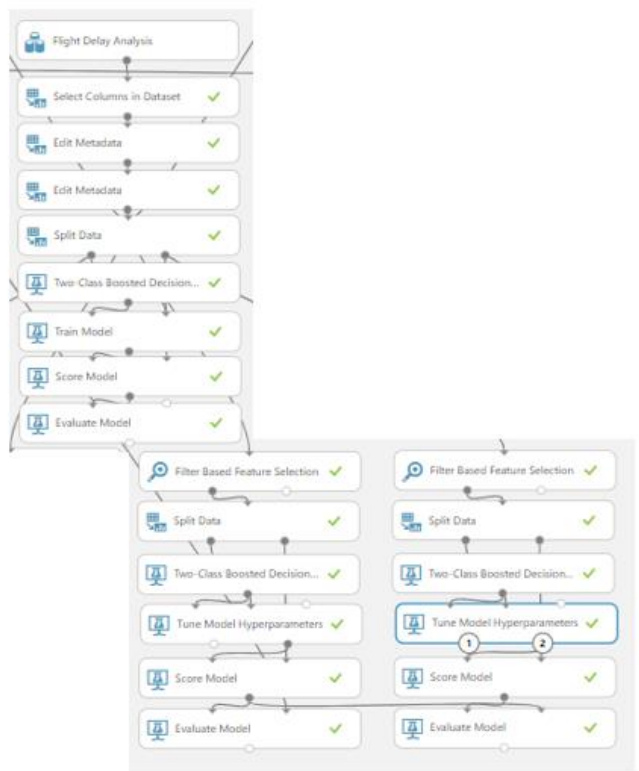
Strategy 4 – Experiment 1 and 2

The best model from strategy 1, 2, and 3 was combined with hyperparameter tuning with a maximum number of runs of 10 and 20.

Experiment 1



Experiment 2



Properties	Project	Properties	Project
Tune Model Hyperparameters		Tune Model Hyperparameters	
Specify parameter sweepin...		Specify parameter sweepin...	
Random sweep		Random sweep	
Maximum number of r...		Maximum number of r...	
10		20	
Random seed		Random seed	
0		0	
Label column		Label column	
Selected columns: Column names: ArrDelayFAA		Selected columns: Column names: ArrDelayFAA	
Launch column selector		Launch column selector	
Metric for measuring ...		Metric for measuring ...	
Accuracy		Accuracy	
Metric for measuring ...		Metric for measuring ...	
Mean absolute error		Mean absolute error	

Analyzing


Once the Machine Learning algorithms were trained, scores were generated and compared with the testing set. Some scores were compared with other models using the same set of data. The best models from each strategy

were compared between themselves keeping the best one out of 4 and the rest were discarded. This approach was used for experiment 1 and 2. The best scores for each strategy and experiment are displayed below:


Experiment 1

Experiment 1 Best Model – Two-Class Decision Forest with standard parameters, 75% Training Data and 25% Testing Data, no feature selection, and 20 max number of runs on Random Sweep. Strategy 4.


Strategy 1

True Positive	False Negative	Accuracy	Precision	Threshold		AUC
23383	12372	0.939	0.893	0.5		0.922
False Positive	True Negative	Recall	F1 Score			
2804	208447	0.654	0.755			
Positive Label	Negative Label					
1	0					


Strategy 2

True Positive	False Negative	Accuracy	Precision	Threshold		AUC
18903	9701	0.939	0.886	0.5		0.919
False Positive	True Negative	Recall	F1 Score			
2438	166563	0.661	0.757			
Positive Label	Negative Label					
1	0					

Strategy 3

True Positive	False Negative	Accuracy	Precision	Threshold		AUC
23358	12397	0.938	0.886	0.5		0.915
False Positive	True Negative	Recall	F1 Score			
3012	208239	0.653	0.752			
Positive Label	Negative Label					
1	0					


Strategy 4

True Positive	False Negative	Accuracy	Precision	Threshold		AUC
23473	12282	0.939	0.896	0.5		0.924
False Positive	True Negative	Recall	F1 Score			
2714	208537	0.656	0.758			
Positive Label	Negative Label					
1	0					


Experiment 2

Experiment 2 Best Model – Two-Class Boosted Decision Tree with 40 Maximum number of leaves per tree and 200 Number of trees constructed, stratified 75% Training Data and 25% Testing Data, no feature selection, and 10 max number of runs on Random Sweep. Strategy 4.


Strategy 1

True Positive	False Negative	Accuracy	Precision	Threshold		AUC
23893	11862	0.940	0.890	0.5		0.929
False Positive	True Negative	Recall	F1 Score			
2953	208298	0.668	0.763			
Positive Label	Negative Label					
1	0					


Strategy 2

True Positive	False Negative	Accuracy	Precision	Threshold		AUC
24354	11401	0.941	0.884	0.5		0.934
False Positive	True Negative	Recall	F1 Score			
3209	208042	0.681	0.769			
Positive Label	Negative Label					
1	0					

Strategy 3

True Positive	False Negative	Accuracy	Precision	Threshold		AUC
24308	11447	0.941	0.884	0.5		0.932
False Positive	True Negative	Recall	F1 Score			
3181	208070	0.680	0.769			
Positive Label	Negative Label					
1	0					

Strategy 4


True Positive	False Negative	Accuracy	Precision	Threshold		AUC
24706	11049	0.941	0.875	0.5		0.936
False Positive	True Negative	Recall	F1 Score			
3543	207708	0.691	0.772			
Positive Label	Negative Label					
1	0					

Best Machine Learning Model

As it was stated previously, the metric we used for scoring, evaluating, and comparing was Accuracy. However, there were several ties, so we used AUC and false negatives as secondary metrics leading us to our final decision. The reason for analyzing false negatives is because we do not want a false prediction when in reality there is a delay; it is better a false positive where a flight is predicted as delay, and business action can be executed to avoid the delay.

Finally, we compared both experiments finding the best machine learning model on Experiment 2 – Strategy 4.

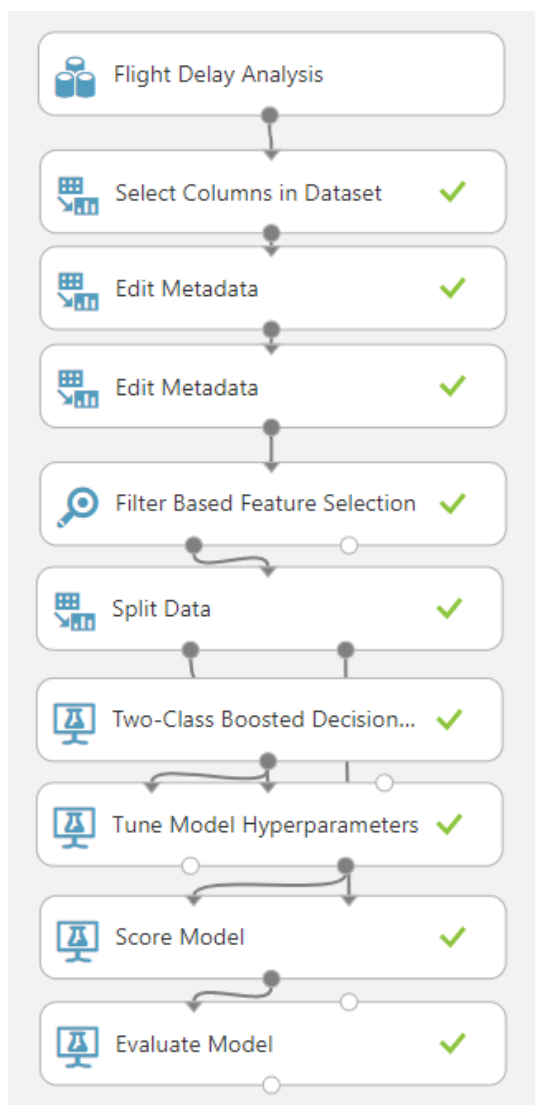
Best model from experiment 1.

True Positive	False Negative	Accuracy	Precision	Threshold		AUC
23473	12282	0.939	0.896	0.5		0.924
False Positive	True Negative	Recall	F1 Score			
2714	208537	0.656	0.758			
Positive Label	Negative Label					
1	0					

Best model from experiment 2.

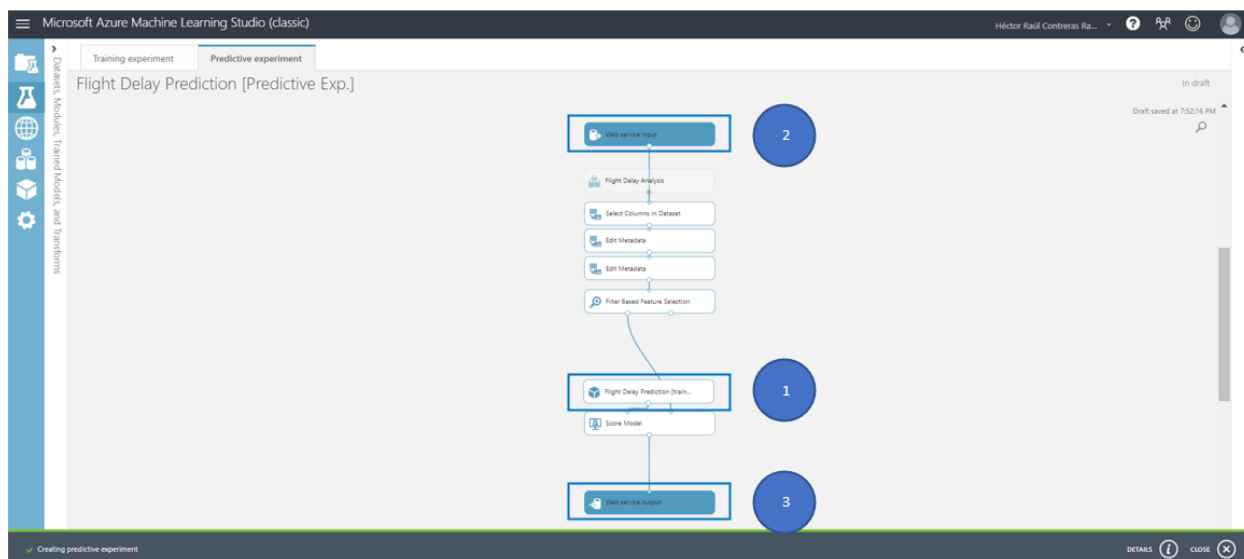
True Positive	False Negative	Accuracy	Precision	Threshold	AUC
24706	11049	0.941	0.875	0.5	
False Positive	True Negative	Recall	F1 Score		
3543	207708	0.691	0.772		
Positive Label	Negative Label				
1	0				

The best model of our training experiments contains the dataset and 9 modules. This is how the model looks like before deploying it as a web service, creating the predictive experiment, and making it operationalized.

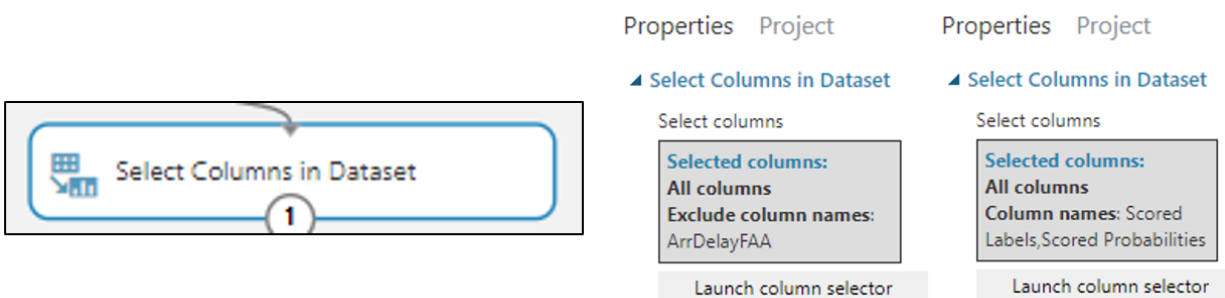


Deployment

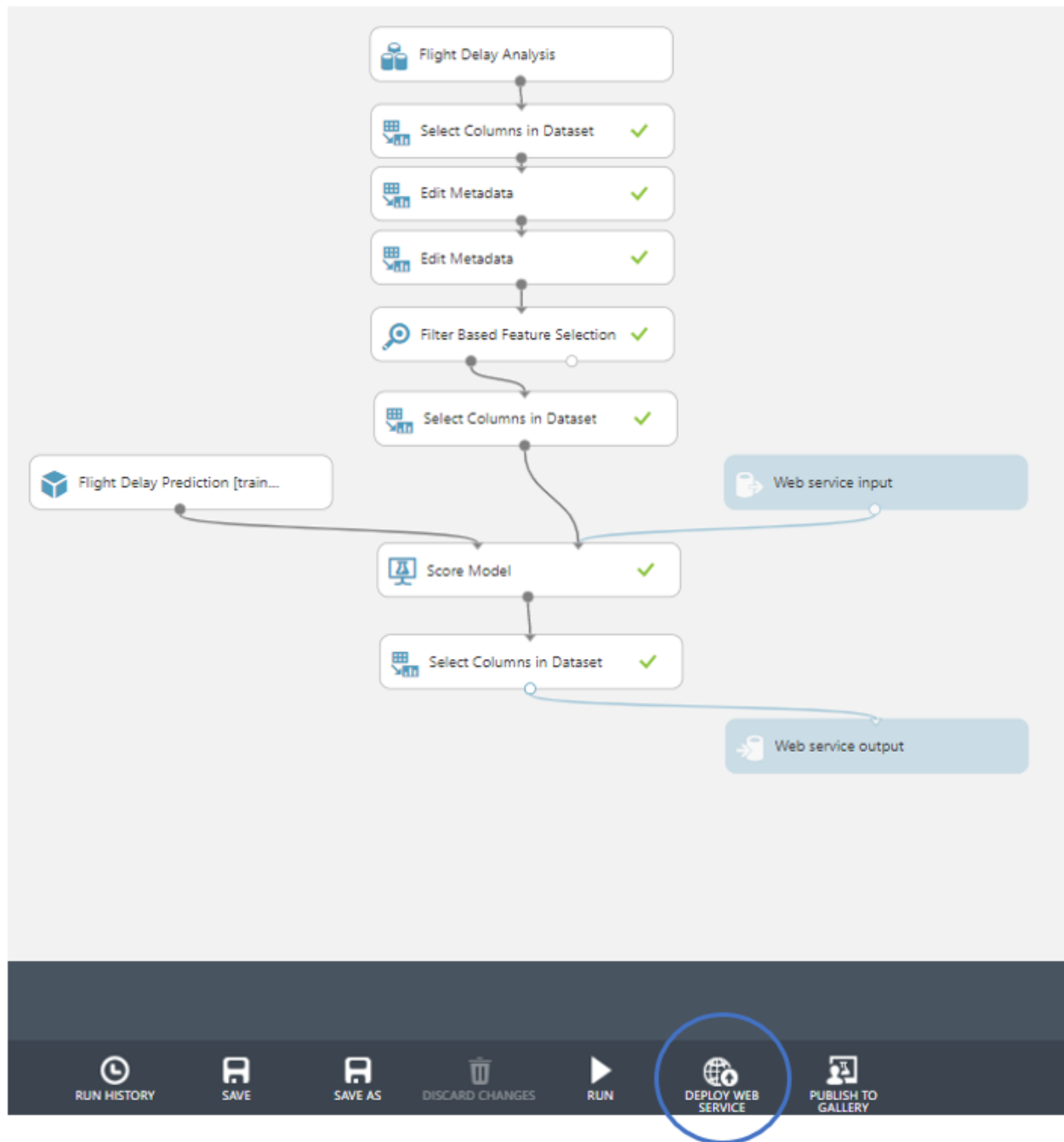
To deploy our model, we had to create our predictive experiment from our training experiment (best model). We set up the web service by selecting the hyperparameter tuning module (Tune Model Hyperparameter), clicking on set up web service, and this is the result:



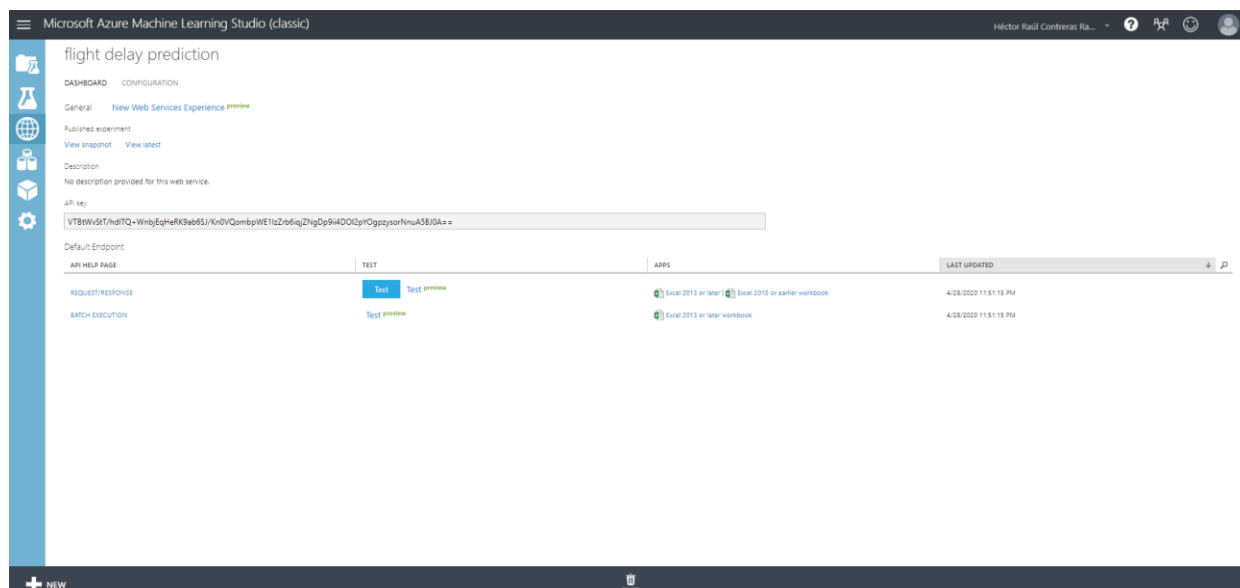
We ended up with a training experiment (previously called our best model) and a predictive experiment. In the predictive experiment the algorithm or trained model was compacted as a new module (1). In addition, “web service input” (2) and “web service output” (3) modules were created automatically (by the wizard), but we had to add two more “select columns” modules. One to specify ArrDelayFAA to be excluded because this is not an input from our model, and the other to select the “scored labels” (the prediction)” and “scored probabilities” (the probability of the prediction) as the output of our model.



After running the predictive experiment, we got our final model as following, then, we deployed it as web service:



The deployment as a web service created a dashboard that contains the API key which is needed for authorization purposes. Also, it contains help pages links for Request/Response and Batch Execution endpoints, workbooks with Azure Machine Learning Add-ins, and a testing section.



Operating the model

We can make our Machine Learning operationalized by different ways, but we tested the following:

1. Web APP through an Azure Subscription
2. Excel Add-ins with connection to the API

Web APP through an Azure Subscription

We can use Azure ML Request/Response web application or Azure ML Batch Execution web application. To do that we would need the API key and API URL for the application we want to execute. This how it looked during its testing:

Request/Response Web application

Request-Response
Batch

Flight_Inputs

DepDelay
27

CRSDepTime
1016

Distance
422

DistanceCode
2

Month
1

DayofMonth
1

Origin
RNO

Dest
SLC

Test Request-Response

Flight_Prediction

Scored Labels
0

Scored Probabilities
0.1185707077384

After submitting the inputs, we obtained the scored labels and scored probabilities as output. This results are obtained because we selected them during the predictive algorithm after the Machine Learning model was scored. This is how the output looks:

The screenshot shows a web interface for testing a machine learning model. It has two tabs: 'Request-Response' (selected) and 'Batch'. Under 'Request-Response', there are two sections: 'Flight_Inputs' and 'Flight_Prediction'. The 'Flight_Inputs' section contains eight input fields: DepDelay (27), CRSDepTime (1016), Distance (422), DistanceCode (2), Month (1), DayofMonth (1), Origin (RNO), and Dest (SLC). The 'Flight_Prediction' section shows the output: 'Scored Labels' as 1 and 'Scored Probabilities' as 0.907850623130798. A 'Test Request-Response' button is located at the bottom left.

The scored label indicates that the flight will not be delayed with a 0.4% of probability that it will be delayed. This makes sense because there is no delay when it departs.

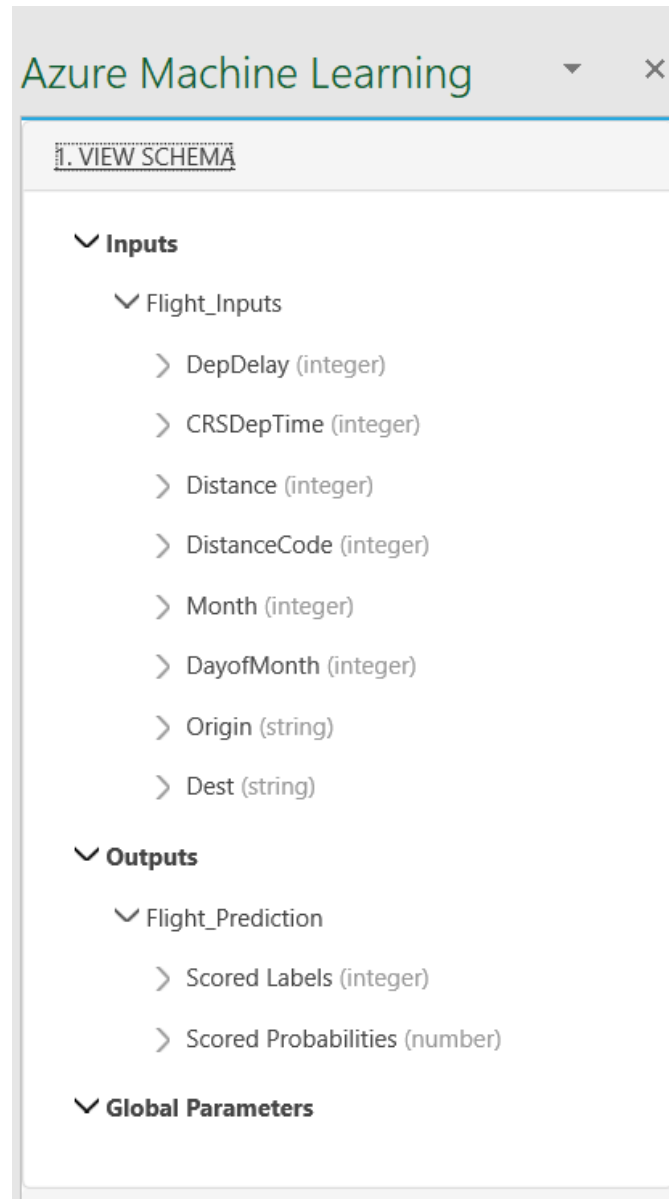
Batch Execution web application

To run the web application, we would need to upload the file to get the scored labels and scored probabilities outputs.

The screenshot shows the 'Test' tab of the Microsoft Azure Machine Learning Studio (classic) Web Services interface. The breadcrumb is 'Flight Delay Prediction > default'. There are two tabs: 'Request-Response' and 'Batch' (selected). Under 'Batch', there is a 'Flight_Inputs' section with a file upload area labeled 'Browse your machine for local files...' and a 'Browse...' button. Below this is a 'Storage account' dropdown menu with the text 'Select an item'. There is also a 'Test Batch Jobs' section. At the bottom, there is a note: 'Note: We will enable CORS on your storage account to upload this file' and a 'Test' button.

Excel Add-ins with connection to the API

The Excel Add-in follows the next schema which is linked to our predictive experiment. We selected our input and output variables with the web service input and web service output modules which were created by our wizard when we set the web service up:



The prediction was tested, and it was accurate after we set up our input and output cells. Below is how the Azure Machine Learning Excel Add-in looks like.

The screenshot displays a Microsoft Excel spreadsheet with a flight delay prediction dataset. The data is organized into columns: DepDelay, CRSDepT, Distance, Distance, Month, DayofMo, Origin, Dest, Scored L, and Scored P. The first six rows of data are visible, showing various flight details and predicted scores. To the right of the spreadsheet, the Azure Machine Learning interface is open, showing the 'Flight Delay Prediction' project. The 'Input' section is set to 'Flight_Inputs' with the data source 'Sheet1!A1:H15'. The 'Output' section is set to 'Flight_Prediction' with the data source 'Sheet1!I1'. The 'Predict' button is highlighted.

DepDelay	CRSDepT	Distance	Distance	Month	DayofMo	Origin	Dest	Scored L	Scored P
-6	1445	406	2	1	1	TPA	ATL	0	0.004464
-3	1951	406	2	1	1	ATL	TPA	0	0.0061055
3	1914	731	3	1	1	ATL	DFW	0	0.0409091
-4	1610	731	3	1	1	DFW	ATL	0	0.0107288
-5	920	2475	10	1	1	LAX	JFK	0	0.0110169

Other models summary

Flight Departure Delay Model

Using Azure ML, a model was created to predict flight delays. The image below exhibits the best model out of every model tested, with an accuracy of 73.8%.

To create this model, many different measures were testing, including:

- Two-Class Decision Forest
- Two-Class Decision Tree
 - Decision trees, ranging from 8 to 54
- Two-Class Logistic Regression
- SMOTE percentage, ranging from 20% to 1000%
- Pearson correlation
- Tuning Hyperparameters
- Splitting data, ranging from 50% to 70%

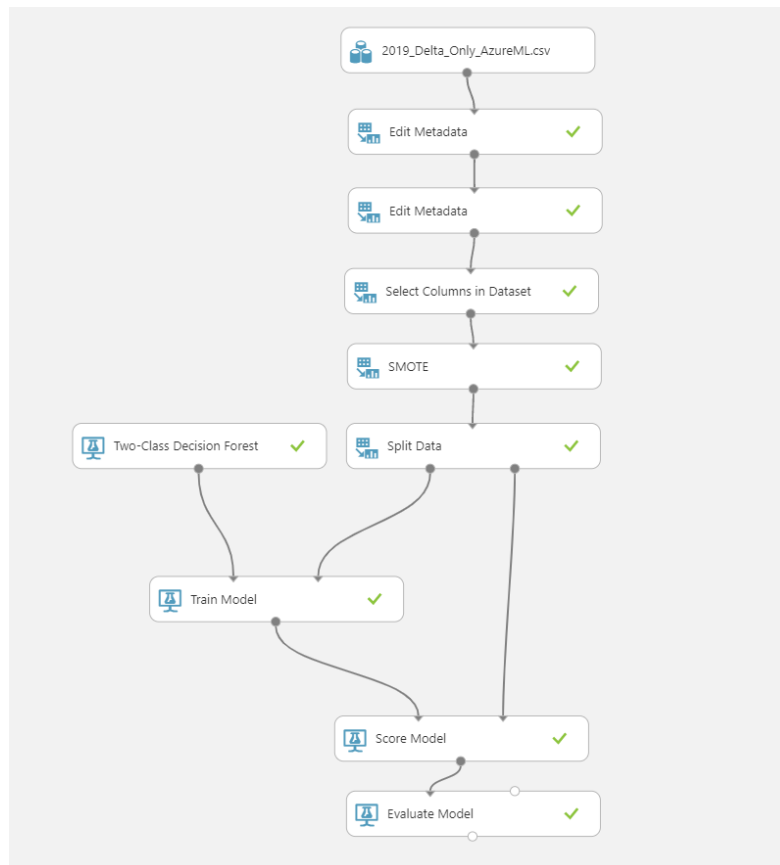
The final model first loaded the prepared data into AzureML. The next step was selecting the label column, "Delay," which was originally a numeric type of 0 (no delay) or 1 (delay). Through the "Edit Metadata" module, "Delay" was selected, then changed to the label and a boolean. Another "Edit Metadata" module was then selected to edit all features mentioned above. These features were also changed to boolean as selected because they were originally numeric, and only contained a 0 or 1.

Next, all the features used in the model were selected using "Select Columns in Dataset." The SMOTE module was selected to increase the label class because of the uneven distribution. The final SMOTE percentage was 200. The data was split 70/30, with 30% reserved for training. A stratified split was chosen when splitting the data to ensure randomness and a wide distribution of data.

The Two-Class Decision Forest produced the most accurate results. All of the module's settings were best at default, besides the number of trees, which increased from eight to 16. The model then trained the training data and scored it against the original, untouched data.

After the data was scored, it was evaluated. At 73.% accuracy and 66.4% precision, the model cannot confidently be used moving forward to predict delays. From these results, it is concluded that the features used in the prediction did not have obvious correlations, causations or patterns among them. Contrarily, according to the Azure ML model, delays are more due to randomness. Therefore, it would not be advised for Delta to use this model in the future to predict whether or not a flight will be delayed.

Flight Delays Model



Flight Cancellations Model

Azure ML was used to create another model to predict flight cancellations. The image below shows the winning model, which had an accuracy of 98.8%.

Similar measures to the delay model were used to test this one, including:

- Two-Class Decision Forest
- Two-Class Decision Tree
- Two-Class Logistic Regression
- SMOTE percentage, ranging from 100% to 2000%
- Tuning Hyperparameters
- Splitting data, ranging from 50% to 70%

The methodology surrounding the winning model first included loading the prepared data into AzureML and selecting the label, "Cancelled." The label was selected through "Edit Metadata" and changed it to a boolean, since

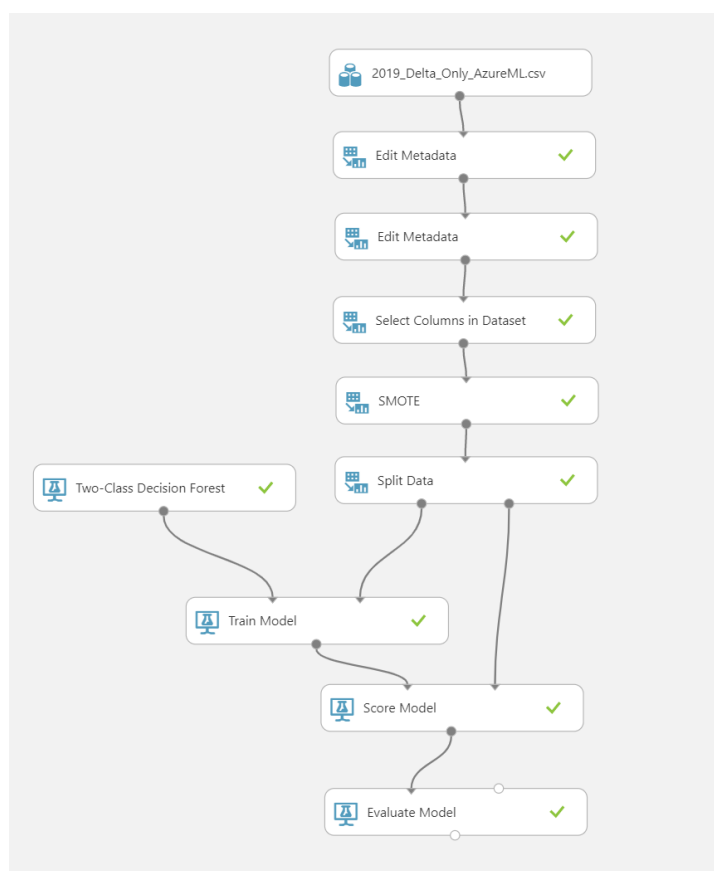
in the dataset, zero marks no cancellation and one marks a cancellation. A second “Edit Metadata” module was also used to edit all features that were used in this machine learning project (previously mentioned), and converted to a boolean since they are zeros and ones in the original dataset.

After, “Select Columns in Dataset” was used to choose the correct features for the model from the prepared dataset. Then, The SMOTE module was selected. The goal of the SMOTE module was to increase the label, “Cancellation,” because of the uneven distribution among the dataset -- a very important step because there were only about 1,000 cancellations out of nearly a million flights. The final SMOTE percentage was 2000, which allowed the data to be more evenly distributed among true and false labels. The original data was split 70/30, where 70% was saved for scoring and 30% was used for training. A stratified split was used when splitting the data to ensure randomness and a wide distribution of data.

The Two-Class Decision Forest method resulted in the highest accuracy, with all settings at default. This tree was used to train the data and then score it against the reserved data from the split.

The winning module’s accuracy was 98.8%, with precision of 96.8%. This model should confidently be used to predict whether or not Delta flights will be cancelled because of the large dataset and high reliability of results.

Flight Cancellations Model



Delta’s Competition

For the following analyses, data from all US passenger airlines operating as of 2019, over a 6 year span, from 2014-2019, was used. Over the 6 year period and across all airlines, on average, 19% of flights were delayed, while 1.99% were cancelled.

Figure 87 shows the cancellation landscape in the industry. Southwest and Skywest were both significantly affected by the grounding of Boeing 737 MAX aircraft. The top 3 airlines, those with the fewest cancellations, are Delta, Allegiant, and Hawaiian (Figure 88). Delta has the most cancellations of the three, with a large spike in 2017 due to several days of severe thunderstorms in Atlanta from which they took time to recover. However, it's important to contextualize these numbers.

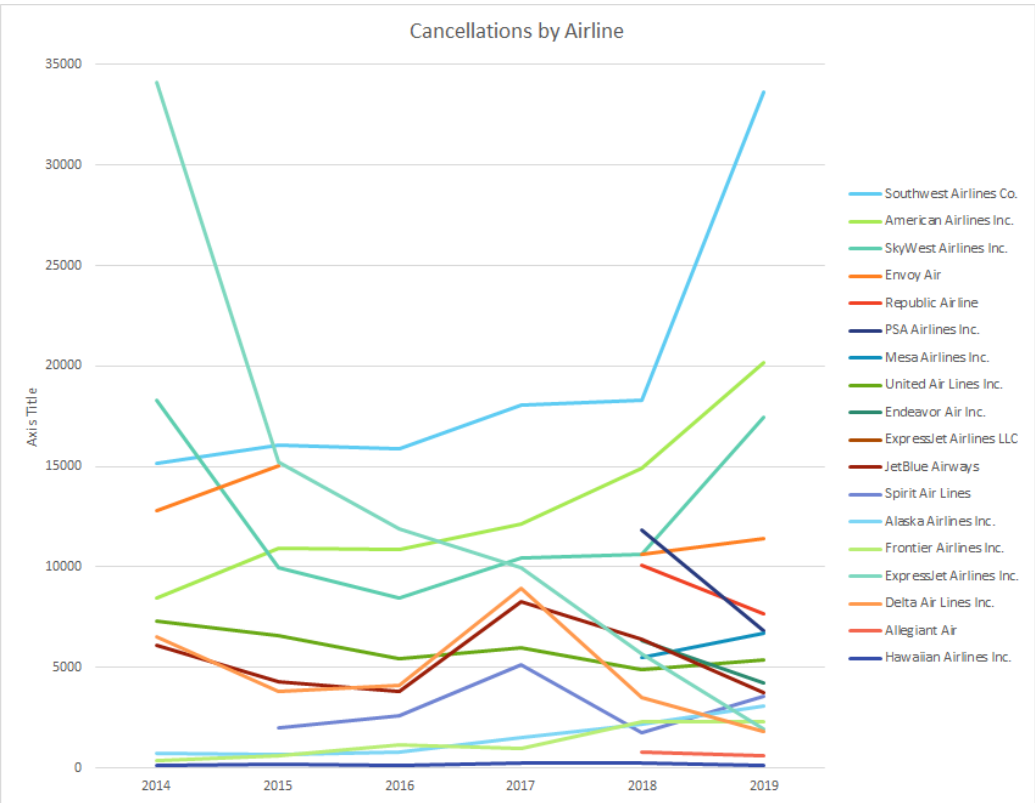


Figure 87

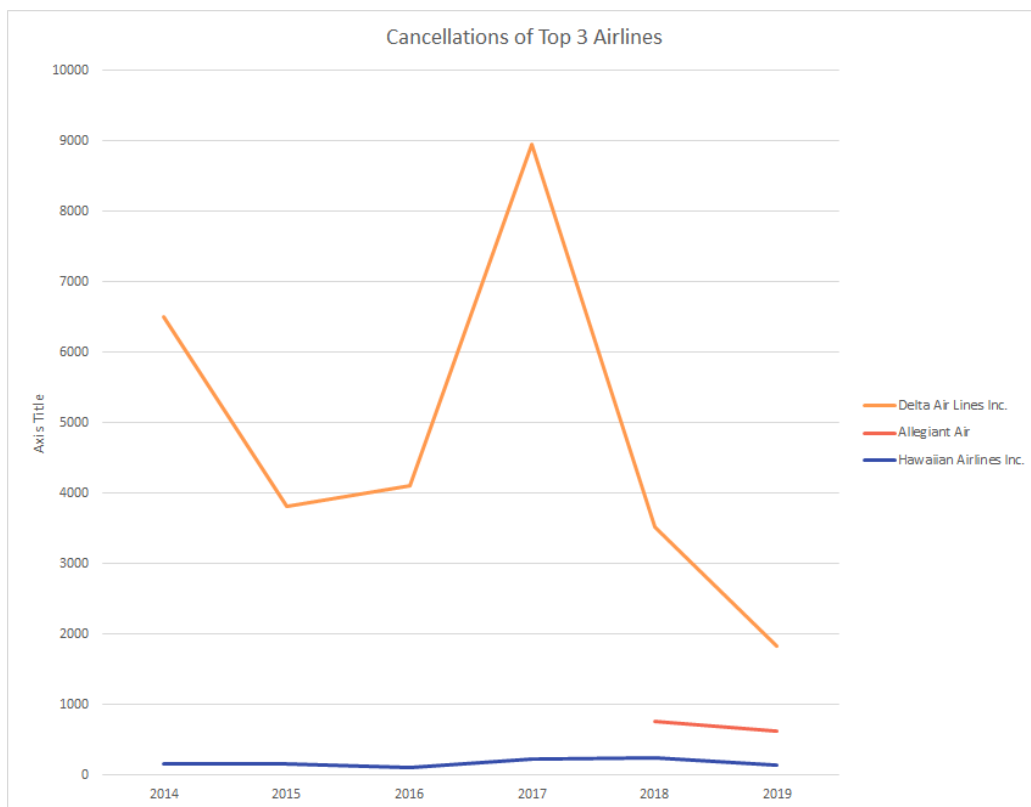


Figure 88

Looking at cancellation rates as a percentage of the total number of flights by each airline, the proportion of cancelled flights is extremely low for Delta and higher for Allegiant and Hawaiian (Figure 89).

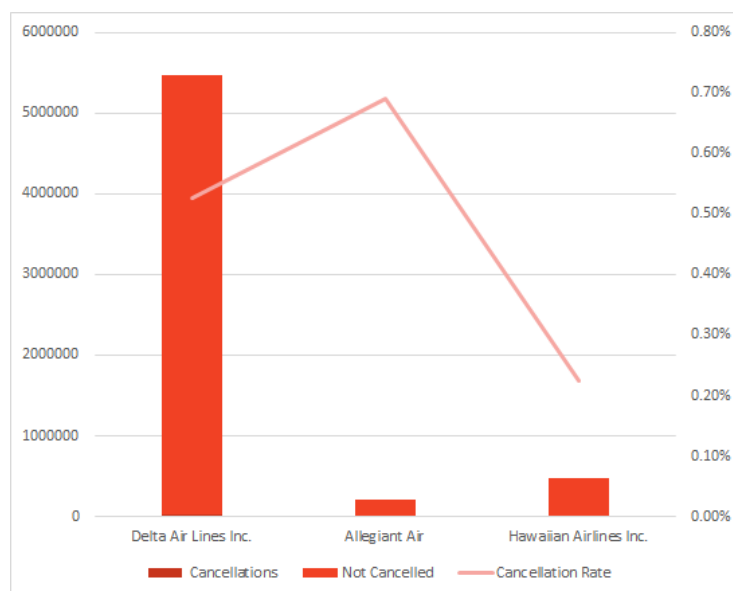


Figure 89

Moving on to delays, Delta is in the top 3 airlines with the lowest delay rate in each of the 6 years analyzed. Hawaiian had the lowest rate in all 6 years. Delta’s 6 year average is 14%, well under the 19% industry average, indicated by the orange line. Over the analyzed period, Hawaiian has the lowest rate of delay, at 10%. It is emerging as the airline to beat

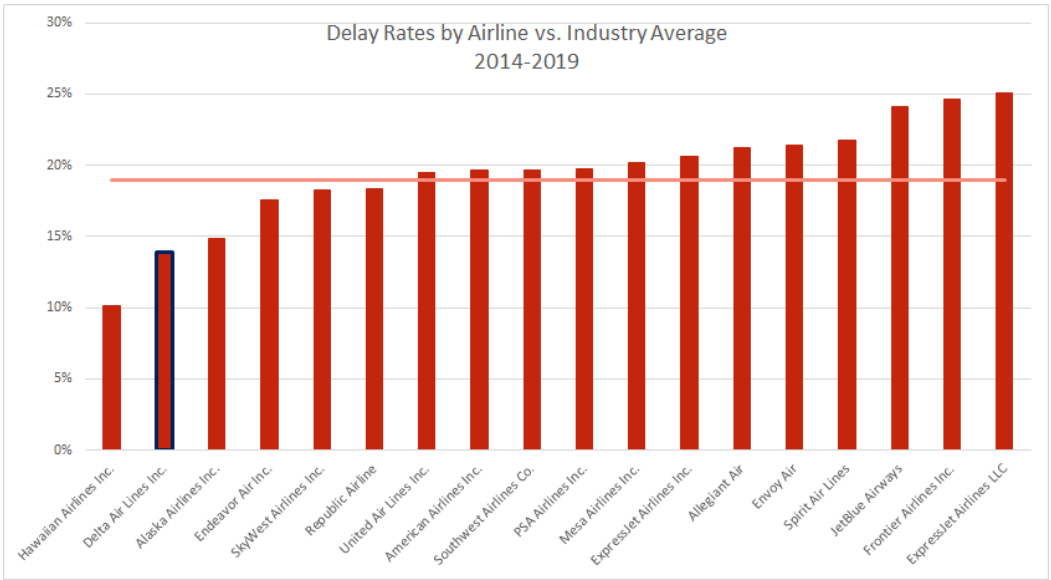


Figure 90

As shown in Figure 91, Delta operates below the industry average for all years studied, a positive sign. However, Figure 92 shows that the number of delayed Delta flights is increasing year over year.

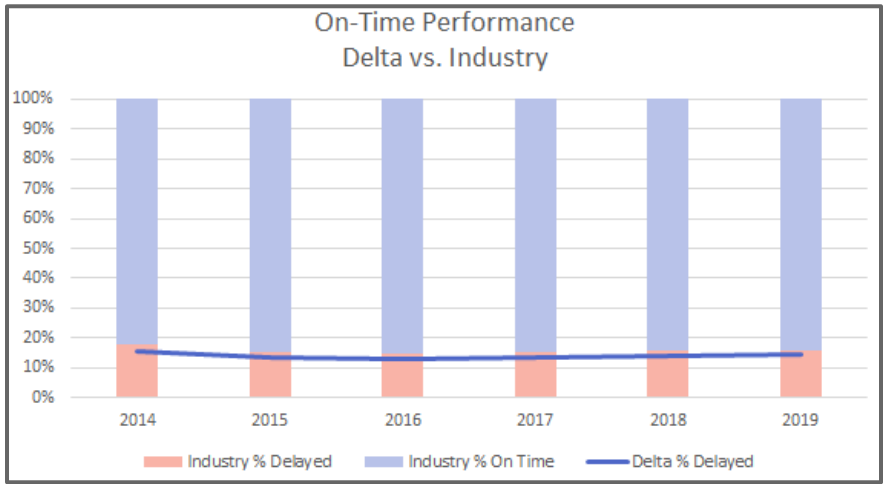


Figure 91

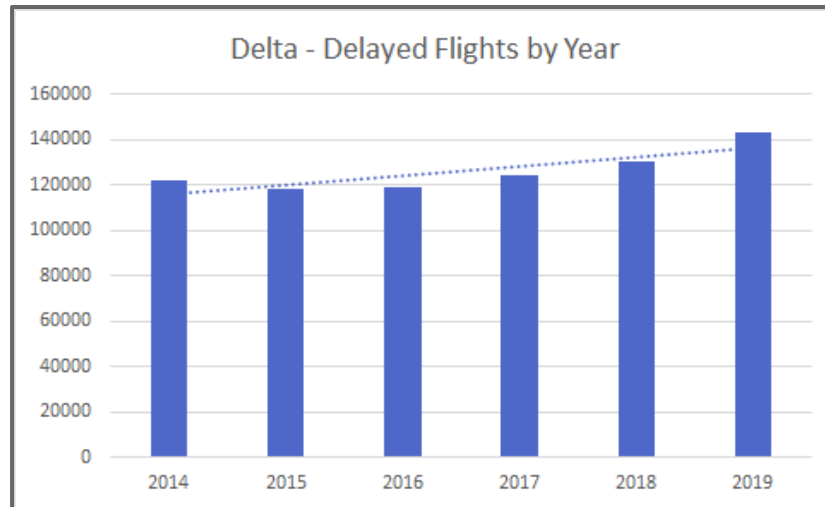


Figure 92

To determine whether this increase is unique to Delta, Figure 93 depicts the two airlines' number of delayed flights by year. Because both trend line slopes are positive, Hawaiian and Delta are both experiencing year-over-year increases in delays, although Hawaiian's trend is flatter than Delta's.

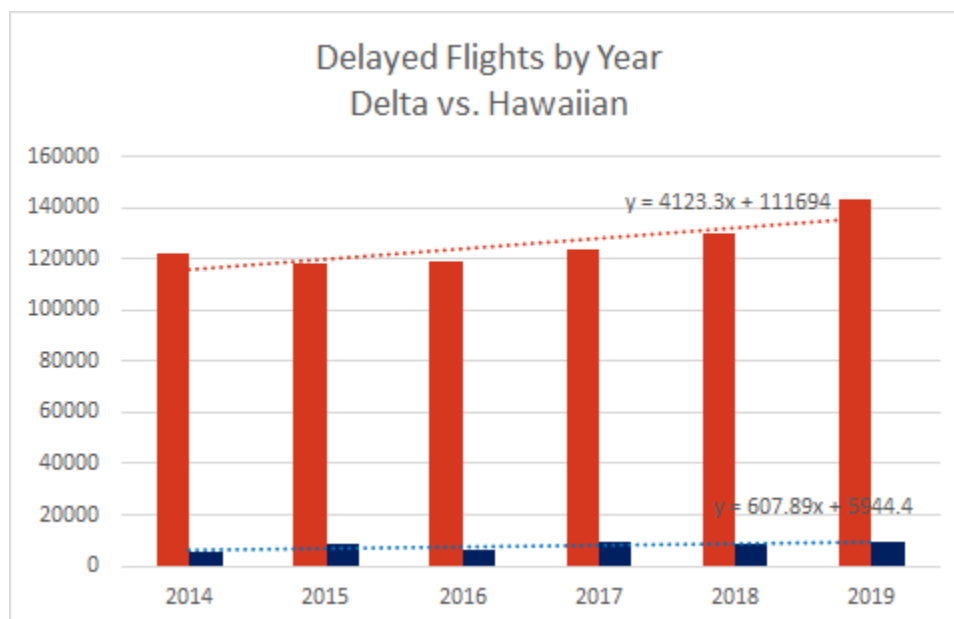


Figure 93

There are a number of possible reasons for this difference. First is the relative size of the two airlines. Hawaiian is significantly smaller than Delta, with Hawaiian flying about 8% of the annual flights Delta flies (Figure 94). When comparing the delay rates side by side, there is a significant difference between the numbers of flights by each airline every year; the difference in delay rates is much smaller (Figure 95). To put this in perspective, in 2019, Delta and Hawaiian had 14% and 11% delay rates respectively. To get to a 14% delay rate, Hawaiian would need to delay an additional 2168 flights. To match Hawaiian's 11% delay rate, Delta would need 29,776 fewer delayed flights over the year.

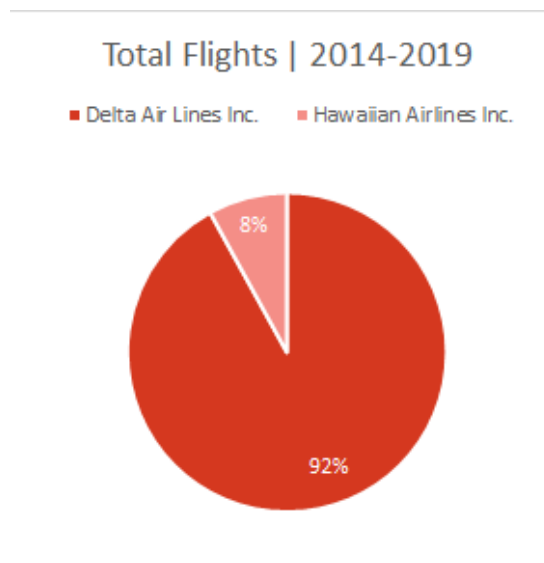


Figure 94

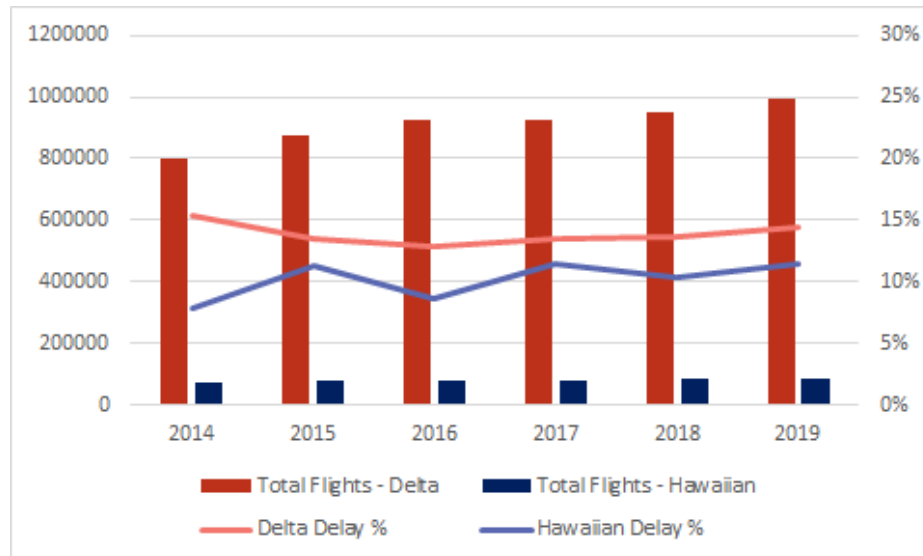


Figure 95

Another attributable difference is the difference in hubs. A hub is an airline's home airport, where most of their flights originate and return to. Hawaiian's hub is in Honolulu, Hawaii, while Delta is based out of Atlanta, Georgia. Hartsfield-Jackson airport in Atlanta was not only the busiest airport in the country in 2018 (FAA, 2020), it was also the busiest airport in the world that year (Bednarz, 2018). 52 million passengers flew through it, more than 5 times the number through Honolulu's airport, which was the 30th busiest in the US (FAA, 2020). Figure 96 shows that flights through Atlanta are significantly more likely to be delayed by weather. Similar numbers of security delays in the two airports indicate that they are equally efficient.

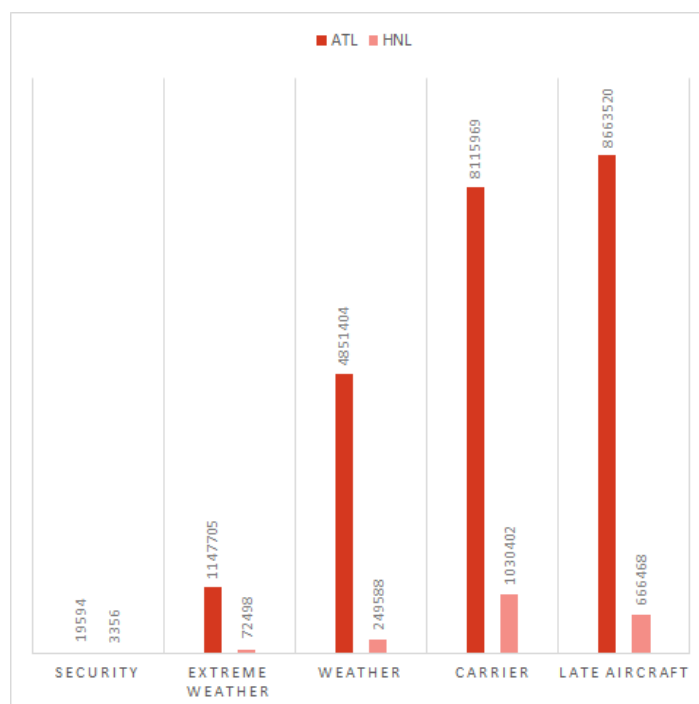


Figure 96

We've now established that Delta is more effective at keeping flight delays and cancellations low than similarly sized airlines. Smaller airlines, particularly Hawaiian, do have lower rates of delays and cancellations, but cannot be accurately used as a benchmark for Delta's business. However, Delta can use Hawaiian's 90% on-time percentage and 0.26% cancellation rate as goals. To reach those goals, the company will need to find a new lever that will help to improve their rates.

Recommendations

Delta is accustomed to making sweeping, effective operational changes to improve their on-time performance rate. Prior to 2011, Delta had the worst OTP of all US commercial airlines. Then, in 2011, it moved from worst to best. "Dave Holtz, senior vice president of operations, credited schedule changes, monthly bonuses of up to \$100 per employee for hitting on-time and other goals, and the rote, repeatable task of sticking to a minute-by-minute pre-departure checklist" (Mashable, 2014).

The following are recommendations for further research that will help Delta make another round of impactful changes.

First, we recommend that Delta focus their efforts on the two controllable delay causes, carrier and late aircraft delays, as they account for 65% of Delta's delays. "Michael Baiada, a recently retired United captain whose company, ATH Group, calculates arrival times for Delta, thinks airlines could eliminate most delays themselves if they had more consistent procedures, left early if they knew there were headwinds or other delaying factors looming, and worked harder to catch up when they fall behind schedule" (Mashable, 2014).

Second, we suggest concentrating on top ten origin airports with most delay minutes to make the biggest impact. Since these airports are the sites of 61% of Delta's controllable delays, this is an ideal test group. On a small scale of ten locations, changes will be easier to implement and track. With the majority of delays experienced here, reducing them will likely have a significant reverse delay propagation effect.

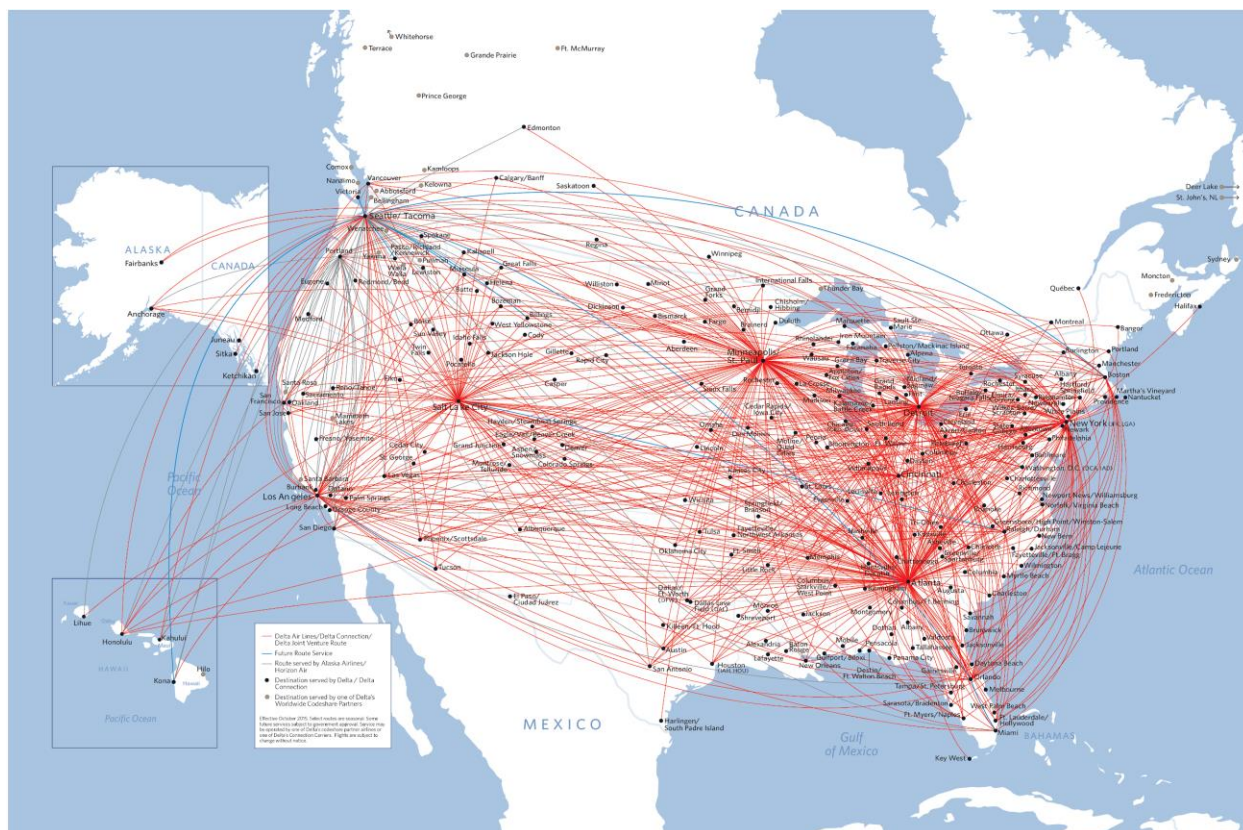
Third, Delta should explore the reasons behind mismatches between the number of delays and cancellations and month. The data shows that more flights do not necessarily equate to more delays, as evidenced by the month of November. The airline should determine whether there are factors during that month that allow for this phenomenon.

Fourth, Delta should determine the reason so many carrier cancellations occur on Saturdays. Saturdays have the fewest flights; one would assume that they also have the fewest cancellations. However, Saturday experiences the second-most cancellations. The cause of this discrepancy may be fixable, leading to fewer cancellations on this day and overall.

Fifth, Delta should determine why there are very long delays on early morning flights during summer. Logic would dictate that the first flights of the day should not experience carrier or late aircraft delays, as these flights are not dependent on equipment or crew from prior flights. They also theoretically have more preparation time before takeoff. If Delta can reduce the number or duration of these delays, there is a good chance that delays later in the day will be reduced by virtue of reverse delay propagation.

Conclusion

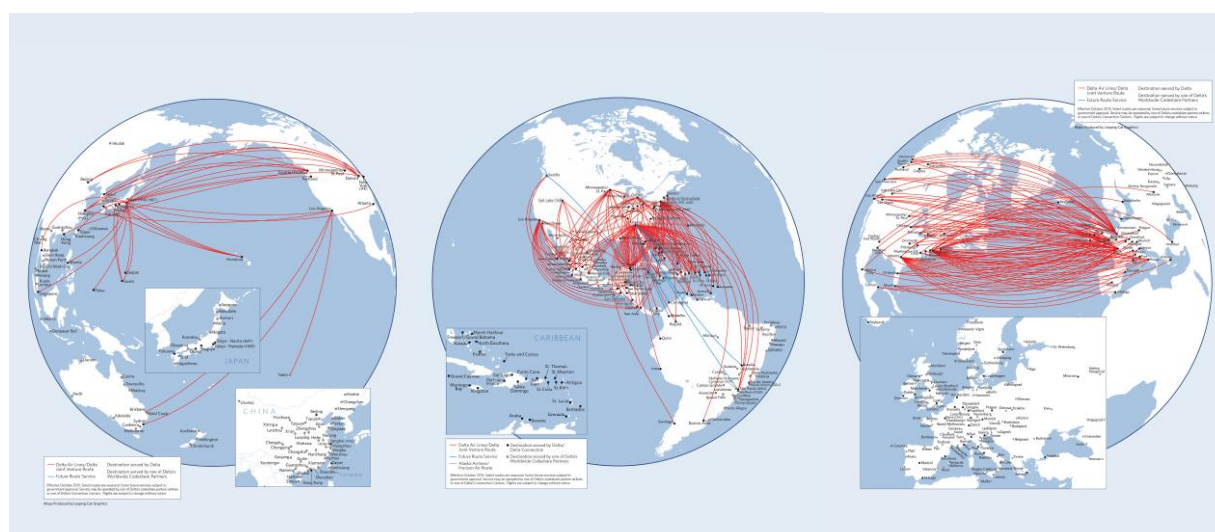
Because Delta is a competitive domestic and international airline, their market share looks different depending on the type of travel. In 2019, Delta ranked second in domestic market share, at 17.5%. American Airlines placed first, with 0.1% of the market share higher than them. Southwest Airline ranked third, at 16.8%, or 0.7% behind Delta. It is clear that Delta is very competitive domestically. Delta's hub in Atlanta received the most foot traffic in the United States in 2019, which exemplifies the strong and strategic hold Delta has on their market (Mazareanu, E.). The image below displays Delta's diverse flight paths across the United States and Canada, with more flights departing and arriving at their hub. In the image, the largest hub, Atlanta, is clearly marked by the most flight routes ("Route map for U.S., Canada").



Delta's main international United States-based competitors are American Airlines and United Airlines. Although Delta has the smallest fleet of the three, they were the most profitable airline in 2019 (Zhang, Benjamin). One of the main reasons Delta took the lead was because of their low fuel costs (Ciesluk, Karol). As shown in the images below, Delta has many routes world wide, and any savings on fuel is a huge advantage ("Route map between U.S., Asia", "Route map between U.S., Europe", "Route map between U.S., Latin America"). These low fuel prices can be attributed to the decision Delta made in owning a fuel refinery -- the only airline in the United States to do so.

Additionally, in 2019, American Airlines and other competitors were forced to cancel thousands of flights because of the grounding of the 737 MAX (Ciesluk, Karol). Because Delta does not own any of these planes in their fleets, they did not see the repercussions of that decision.

Another factor to Delta's success in 2019 was its growth in their key metrics. Firstly, Delta was able to notably increase the number of passengers carried. Although American Airlines carried 215 million passengers (compared to Delta's 204 million), they grew 5.6%, while Delta grew over 6%. Secondly, Delta's cost per available seat mile was lower than both American Airlines and United Airlines, at a mere 10.52 cents per seat. This low cost dramatically increases their profitability, considering they flew over 200 million passengers throughout the year. Delta also made the most revenue per seat. American Airlines brought in 17.41 cents per seat mile, while United Airlines took 16.55 per seat mile. Meanwhile, Delta profited 17.79 cents per seat mile. By investing in strategies that impact their bottom line, Delta is able to be the most profitable, even though their fleet is not the largest (Ciesluk, Karol).



In Addition to being the most profitable airline in 2019, Delta was also named the best carrier based on its on-time flight departures and arrivals, multitude of routes and customer service. Out of all the passengers it flew in 2019, only 32 were forced off their flight due to capacity reasons. This factor contributed to their high customer service ranking. By prioritizing customer service and loyalty, Delta is able to hold a sustainable competitive advantage (Talty, Alexandra).

Through its strategic positioning, flight routes, hubs, service strategy and employee treatment, Delta continues to show growth, competition and innovation that outlasts trends. With long term goals in place, they will continue to grow and progress as an airline and a culture that customers want to join.

Summary of recommendations

Predictive algorithm

Flight delays and cancellations tend to lead to a bad reputation for airlines, to lose reliability, and to generate financial losses to passengers, airlines, and airports. Additionally, they impact the environment because of the

increasing gas emissions and fuel consumption which cause higher costs for the company. In this project, we developed three Machine Learning Models from which two of them are very accurate. It is recommended to use these predictive analytics tools to take actions and prevent those operational issues. The science behind the algorithms and the analysis to develop the Machine Learning model are very useful for Decision Makers who can lead a data driven business and create competitive advantage which will result in a solid and profitable business.

The model can be implemented on a singular platform or in a private or non-private Delta's web application. When it was deployed as a web service, a dashboard was set up and APIs key / URL were generated; The API can be added to another product if it is needed. Also, a different approach can be implemented by using Azure Machine Learning Excel Add-in. The innovation and leverage that our predictive analysis offers would spotlight Delta Airlines against its competitors.

Recommendations for controllable factors

Delta Airlines is in complete control of carrier and late aircraft delays, which make up 61% of the airline's total 2019 delays. With delays and cancellations costing millions of dollars annually, as well as the constant fight for market share, it is in Delta's best interest to delve into the data they have collected and create actionable plans to reduce their risk factors.

References

“2019 North America Airline Satisfaction Study.” *2019 North America Airline Satisfaction Study*, J.D. Power, 29 May 2019, www.jdpower.com/business/press-releases/2019-north-america-airline-satisfaction-study.

“Airline On-Time Statistics and Delay Causes.” *Bureau of Transportation Statistics*, US Department of Transportation, 2020, www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp.

The Associated Press. “Flight Delays Are Costing Airlines Serious Money.” *Mashable*, Mashable, 10 Dec. 2014, mashable.com/2014/12/10/cost-of-delayed-flights/.

Bachman, Justin. “Delta Holds an Edge Over Competitors by Dominating Less-Competitive Markets.” *Skift*, Skift, 14 Jul. 2017, <https://skift.com/2017/07/14/delta-holds-an-edge-over-competitors-by-dominating-less-competitive-markets/>.

Baker, Michael B. “Delta Nabs Ninth Win as United Shows Stark Improvement.” *Business Travel News*, Business Travel News, 18 Nov. 2019, www.businesstravelnews.com/Research/Delta-Nabs-Ninth-Win-as-United-Shows-Stark-Improvement.

Bednarz, Christine. “The 10 Busiest Airports in the World.” *The 10 Busiest Airports in the World*, National Geographic, 4 Oct. 2018, www.nationalgeographic.com/travel/lists/transportation/worlds-busiest-airports-things-to-do-layover/.

Ciesluk, Karol. “Is Delta The Leading US Airline? These Stats Would Say So....” *Is Delta The Leading US Airline? These Stats Would Say So...*, Simple Flying, 9 Feb. 2020, <https://simpleflying.com/2019-us-airlines-performance/>.

David Linthicum “Leveraging Cloud-Based Machine Learning on Azure: Real-World Applications.” LinkedIn, 05 Dec. 2019, <https://www.linkedin.com/learning/leveraging-cloud-based-machine-learning-on-azure-real-world-applications/law-enforcement>

“Delta partners with IBM to explore quantum computing – an airline industry first.” *Delta partners with IBM to explore quantum computing – an airline industry first*, Delta News Hub, 8 Jan. 2020, <https://news.delta.com/delta-partners-ibm-explore-quantum-computing-airline-industry-first>.

“Delta debuts industry-leading international Main Cabin experience.” *Delta debuts industry-leading international Main Cabin experience*, Delta News Hub, 8 Dec. 2019, <https://news.delta.com/delta-debuts-industry-leading-international-main-cabin-experience>.

Egerton, Debbie. “An industry first: Delta launches innovative solution for pet travel.” *An industry first: Delta launches innovative solution for pet travel*, Delta News Hub, 22 Jan. 2020, <https://news.delta.com/industry-first-delta-launches-innovative-solution-pet-travel>.

Egerton, Debbie. “Delta becomes first major carrier to offer real-time Bluetooth tracking on container shipments.” *Delta becomes first major carrier to offer real-time Bluetooth tracking on container shipments*, Delta News Hub, 2020, <https://news.delta.com/delta-becomes-first-major-carrier-offer-real-time-bluetooth-tracking-container-shipments>.

Gay, Charles. “Delta’s history: From dusting crops to connecting the world.” *Delta’s history: From dusting crops to connecting the world*, Delta News Hub, 23 Apr. 2016, <https://news.delta.com/deltas-history-dusting-crops-connecting-world>.

"How Flight Delays Set Off A Domino Effect." *Centives*, 11 Jan. 2013, www.centives.net/S/2013/how-flight-delays-set-off-a-domino-effect/.

"How to select algorithms for Azure Machine Learning." Microsoft, 05 Mar. 2020, <https://docs.microsoft.com/en-us/azure/machine-learning/how-to-select-algorithms>

"INFOGRAPHIC: Rolling out optional facial recognition technology to improve the travel experience." *INFOGRAPHIC: Rolling out optional facial recognition technology to improve the travel experience*, Delta News Hub, 6 Sep. 2019, <https://news.delta.com/infographic-rolling-out-optional-facial-recognition-technology-improve-travel-experience>.

Isidore, Chris. "Delta Air Lines files for bankruptcy." *Fortune 500*, CNN Money, 15 Sep. 2005, <https://money.cnn.com/2005/09/14/news/fortune500/delta/>.

Lianne & Justin. "Data cleaning in Python: The Ultimate Guide (2020)." Medium, Towards data science, 03 Feb. 2020, <https://towardsdatascience.com/data-cleaning-in-python-the-ultimate-guide-2020-c63b88bf0a0d>

"Machine Learning Algorithm Cheat Sheet for Azure Machine Learning designer." Microsoft, 05 Mar. 2020, <https://docs.microsoft.com/en-us/azure/machine-learning/algorithm-cheat-sheet>

Mazareanu, E. "Domestic market share of leading U.S. airlines from February 2019 to January 2020." *Domestic market share of leading U.S. airlines from February 2019 to January 2020*, Statista, 20 Apr. 2020, <https://www.statista.com/statistics/250577/domestic-market-share-of-leading-us-airlines/>.

McCartney, Scott. "The Best and Worst U.S. Airlines of 2019." *The Wall Street Journal*, Dow Jones & Company, 15 Jan. 2020, www.wsj.com/articles/the-best-and-worst-u-s-airlines-of-2019-11579097301.

"Monthly weather forecast in Atlanta, GA." *Monthly weather forecast in Atlanta, GA*, Weather Atlas, n.d., <https://www.weather-us.com/en/georgia-usa/atlanta-climate>.

"On-Time : Reporting Carrier On-Time Performance (1987-present)." *Bureau of Transportation Statistics*, US Department of Transportation, 2020, https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time

"Passenger Boarding (Enplanement) and All-Cargo Data for U.S. Airports." *Federal Aviation Administration*, US Department of Transportation, 9 Jan. 2020, www.faa.gov/airports/planning_capacity/passenger_allcargo_stats/passenger/.

"Prevent overfitting and imbalanced data with automated machine learning." Microsoft, 09 Apr. 2020, <https://docs.microsoft.com/en-us/azure/machine-learning/concept-manage-ml-pitfalls>

Ramos, Rachel. "Delta files for Chapter 11 bankruptcy." *Atlanta Business Chronicle*, Atlanta Business Chronicle, 14 Sep. 2005, <https://www.bizjournals.com/atlanta/stories/2005/09/12/daily17.html>.

"Route map between U.S., Asia." *Route map between U.S., Asia*, Delta News Hub, 15 Oct. 2015, <https://news.delta.com/route-map-between-us-asia>.

- "Route map between U.S., Europe." *Route map between U.S., Europe*, Delta News Hub, 28 Oct. 2015, <https://news.delta.com/route-map-between-us-europe>.
- "Route map between U.S., Latin America." *Route map between U.S., Latin America*, Delta News Hub, 28 Oct. 2015, <https://news.delta.com/route-map-between-us-latin-america>.
- "Route map for U.S., Canada." *Route Map for U.S., Canada*, Delta News Hub, 28 Oct. 2015, <https://news.delta.com/route-map-us-canada>.
- Sam. "Delta Airlines: Flying High in a Competitive Industry." *Technology and Operations Management*, Technology and Operations Management, 8 Dec. 2019, <https://digital.hbs.edu/platform-rctom/submission/delta-airlines-flying-high-in-a-competitive-industry/>.
- Smallen, Dave. "Preliminary Estimated Full Year 2019 and December 2019 U.S. Airline Traffic Data." *Bureau of Transportation Statistics*, US Department of Transportation, 17 Jan. 2020, www.bts.gov/newsroom/estimated-full-year-2019-and-december-2019-us-airline-traffic-data.
- Smallen, Dave. "Third Quarter 2019 U.S. Airline Financial Data." *Bureau of Transportation Statistics*, US Department of Transportation, 6 Dec. 2019, www.bts.gov/newsroom/third-quarter-2019-us-airline-financial-data.
- Steele, Kathryn. "Delta expands optional facial recognition boarding to new airports, more customers." *Delta expands optional facial recognition boarding to new airports, more customers*, Delta News Hub, 8 Dec. 2020, <https://news.delta.com/delta-expands-optional-facial-recognition-boarding-new-airports-more-customers>.
- Steele, Kathryn. "Relaunched Delta SkyMiles American Express Cards – now with new Card designs – debut with even more benefits for travelers." *Relaunched Delta SkyMiles American Express Cards – now with new Card designs – debut with even more benefits for travelers*, Delta News Hub, 30 Jan. 2020, <https://news.delta.com/relaunched-delta-skymiles-american-express-cards-now-new-card-designs-debut-even-more-benefits>.
- Talty, Alexandra. "Delta Named Best Airline In 2019." *Delta Named Best Airline In 2019*, Forbes, 19 Mar. 2019, <https://www.forbes.com/sites/alexandratalty/2019/03/19/delta-named-best-airline-in-2019/#4d8b92fd6f91>.
- "Train models with Azure Machine Learning." Microsoft, 05 Mar. 2020, <https://docs.microsoft.com/en-us/azure/machine-learning/concept-manage-ml-pitfalls>
- "Types of Delay." *Aviation System Performance Metrics (ASPM)*, Federal Aviation Administration, 7 Mar. 2019, aspmhelp.faa.gov/index.php/Types_of_Delay.
- Drazen Zaric, "Better Heatmaps and Correlation Matrix Plots in Python." Medium, Towards data science, 15 Apr. 2020, <https://towardsdatascience.com/better-heatmaps-and-correlation-matrix-plots-in-python-41445d0f2bec>
- Zhang, Benjamin. "The 20 biggest airlines in the world, ranked." *The 20 biggest airlines in the world, ranked*, Business Insider, 6 Mar. 2019, <https://www.businessinsider.com/biggest-airlines-world-oag-2019-3>.

