**Flight Research and Analytics**

Annie Chan, Azucena Faus, Mohammad Mahmoudighaznavi

Shiley-Marcos School of Engineering, University of San Diego

[Flight Research and Analytics Group GitHub](https://github.com/fausa/Flight-Research-and-Analytics-Group)

# Abstract

The airline industry plays an important and critical role in the world’s transportation sector. Various factors such as weather conditions, air traffic control, airport operations, etc., have the capability to affect airline services directly and indirectly, leading to delayed and canceled flights. Understanding the effects of such issues in advance can prepare customers and airline operators for potential reasons for flight cancellations.

*Keywords:* machine learning, predictive modeling, flight cancellations, weather, airlines, Sagemaker, AWS, airport weather

# Table of Contents

Contents

[Abstract 2](#_Toc132363977)

[Table of Contents 3](#_Toc132363978)

[Flight Research and Analytics Group 5](#_Toc132363979)

[Problem Statement 5](#_Toc132363980)

[Goals 5](#_Toc132363981)

[Non-Goals 6](#_Toc132363982)

[Data Sources 6](#_Toc132363983)

[Risks 7](#_Toc132363984)

[Data Ingestion and Exploration 7](#_Toc132363985)

[Data Ingestion 8](#_Toc132363986)

[Data Exploration 8](#_Toc132363987)

[Data Cleaning 10](#_Toc132363988)

[Measuring Impact 10](#_Toc132363989)

[Security Checklist, Privacy, and Other Risks 10](#_Toc132363990)

[Risks associated with Bias 11](#_Toc132363991)

[Ethical Considerations 12](#_Toc132363992)

[Data Preparation and Scrubbing 12](#_Toc132363993)

[Train, Test, and Validation Splits 15](#_Toc132363994)

[Balancing data 15](#_Toc132363995)

[Modeling, Methods, Validation, and Performance Metrics 15](#_Toc132363996)

[Sagemaker Autopilot 15](#_Toc132363997)

[Scikit-Learn Logistic Regression 16](#_Toc132363998)

[Modeling Results and Findings 17](#_Toc132363999)

[Final Model 18](#_Toc132364000)

[Future Enhancements and Next Steps 18](#_Toc132364001)

[Next Steps 20](#_Toc132364002)

[References 21](#_Toc132364003)

[Appendix 22](#_Toc132364004)

# Flight Research and Analytics Group

Flight Research and Analytics Group is a private contracting company made up of 3 employees that consult airlines, airports, and anyone in the Air travel industry with actionable insights derived from relevant data using predictive modeling and machine learning algorithms.

## Problem Statement

Although travelers know the propensity for canceled flights during the winter months due to weather, it is difficult to know when that likelihood is highest and what flight to schedule if travel is nonetheless required. Additionally, customer complaints at the Department of Transportation have increased dramatically when comparing pre and post pandemic numbers (Jacobsen, 2022). Customer dissatisfaction is not only painful for the customer, but the airlines, airport flow, and other customers. Having some indication that a flight might be canceled and then knowing you might get a lower airfare rate in case of a high likelihood of cancellation would be useful to set realistic expectations and hopefully increase customer satisfaction.

## Goals

The success of this project would contribute to the achievement of the following goals:

1. Combining weather and flight data to find connections and correlations between certain types and severity of weather conditions and canceled flights.
2. Training a model to predict cancellations on the flight dataset with weather condition data and predict whether a flight has a high probability of being canceled based on the weather for that date/time. We would like to attain an accuracy/F1 score/Sensitivity of: 0.8, 0.8, and 0.7, respectively.

Ultimately, the goal of this project is to provide airlines and airports with information regarding the likelihood of canceled flights due to weather. These entities can then notify customers within a more reasonable timeframe than when they arrive. This can then be extended to a long-term goal where airlines can provide this information at time of booking the flight - perhaps even lower the fare in case a cancellation is highly probable.

## Non-Goals

On the other hand, there are several non-goals that the team is not intentionally solving with this project:

1. Forecasting weather - The project is seeking to find if we can adequately predict canceled flights based on weather conditions. Once this concept is proven, a weather forecasting model can be used in conjunction with the canceled flights predictive model.
2. Predicting airline performance - This project intends to improve customer satisfaction relating to the added benefit of knowing in advance when a flight might be canceled. There is no desire to correlate these findings with airline performance.
3. Predicting delayed flights - The focus of this project is to predict canceled flights due to weather alone.

## Data Sources

One of the data sources comes from a relational database called Airline that contains US domestic flight information for January of 2016, along with information regarding canceled flights and whether those cancellations were due to weather. The second dataset was collected from Kaggle, which contains weather data for US airports for the years 2016-2021. This way, departure airport weather conditions can be added features for those flights that experienced cancellations due to weather. The final dataset contains more detailed daily weather conditions for singular airports of interest that we will be focusing on, such as JFK. Having more detailed weather data might help in predicting canceled flights.

[Flight Information Dataset](https://relational.fit.cvut.cz/dataset/Airline)

[US Airports Weather Dataset](https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events)

[JFK Airport Detailed Weather Dataset](https://www.weather.gov/wrh/Climate?wfo=okx)

## Risks

Some risks with the data are that it is only for one month, so any predictions would have to be for the following year, in the same month as weather conditions are seasonal. More data would have to be collected to further prove this concept and provide more accurate predictions, especially if weather trends change over time (climate change).

# Data Ingestion and Exploration

All datasets that used were downloaded from the source system and imported to the Amazon S3 bucket created for this project:

s3://sagemaker-us-east-1-993410942383/content-project/airline\_data/

The source data was stored in an S3 bucket by way of SQL extraction of the necessary tables and loading the comma-delimited files for weather.

## Data Ingestion

To ingest the data, Amazon Athena was employed to create a database and connection to the S3 bucket. After the connection is established, tables were created and stored to the table-to-S3 mapping columns accordingly.

## Data Exploration

During data exploration[[1]](#footnote-1), the goal was to find interesting insights on the relationships between cancellations, locations, airlines, flight dates and times. It was also important to ensure that the data is accurate and of high quality. Since the data source for the flights data came from the U.S. Department of Transportation and the weather data came from Kaggle as a result from a research paper on “Short and Long-term Pattern Discovery Over Large-Scale Geo-Spatiotemporal Data,” the data was confirmed as assuredly accurate (Moosavi et al., 2019). The final dataset came directly from the National Weather Service website where one can query the specific month, year, and location of desired data. Data quality for this final set was good with no null data for the fields of interest.

Data exploration revealed that the most common reason for flight cancellations was due to weather. Also uncovered was the seasonality in flights over the course of the month. When looking at all flights within the continental US, a curious finding was found where there was a spike in canceled flights between January 20-January 27. A deeper dive into what factors might have played a role in this spike revealed a major snowstorm forecast for those days in the Eastern part of the United States (Fountain, 2016). This might indicate a lack of seasonality in canceled flights, which would confirm that this problem must be addressed through a predictive model that uses weather information rather than depend on seasonal trends.

As would be expected, most flight cancellations happen over the weekend, but this is likely due to the increase in travel during those days. When looking at the proportion of cancellations with the total number of flights, unexpected airports came to the top of the list. This might be because the number of total flights is quite small, and the weather quite severe during that month, hence the higher proportion of canceled flights. It might also be that popular airports such as JFK and La Guardia are not so bad and mostly keep a good number of flights from getting canceled during the winter months. A deeper dive into popular, high traffic airports and their cancellation proportions might give further insights.

Within the first weather dataset, the most common weather type and severity combination was Rain/Light followed by Fog/Severe. When observing weather patterns for JFK airport, the top three weather Type/Severity conditions were Rain/Light, Rain/Moderate, and Snow/Light. This dataset also names airports with a prefix K which will have to be removed when matching up datasets for the flights and the weather is done. Also, since weather and flights have date and time attributes in addition to location, the flight time and date must fall within the start time and end time of the given weather condition for that airport found in the weather data.

The second weather dataset focused specifically on JFK airport revealed some pattern of rainfall during the month of January for 2016, while also revealing that on January 16, 2016, they had first snow. Snow depth grew from there to a maximum of 28 inches on January 24.

During the data exploration, Amazon Athena was also used since it can run ad hoc SQL queries on the athena tables created without having to move the data. Additionally, the pandas package was applied to generate data frames and helped with data cleaning and preparing for modeling. Matplotlib and seaborn packages facilitated data visualizations when SQL queries came up short. Finally, AWS Sagemaker Autopilot was employed to generate its own data exploration report that aided in gathering insights about potential issues with the data, like multicollinearity among variables included.

## Data Cleaning

The dataset contained all transactions relevant to the business, so data cleaning was needed to analyze real purchase transactions for the item of interest versus the rest of many transactions. Each transaction had 9 variables: ‘InvoiceNo,’ ‘StockCode,’ ‘Description’ (the item description), ‘Quantity’ (amount purchased), ‘InvoiceDate,’ ‘UnitPrice,’ ‘CustomerID,’ and ‘Country’ (from which order originated).

# Measuring Impact

In the long run, the metric this project seeks to change is to increase customer satisfaction. In the short term, having the ability to effectively predict canceled flights due to weather conditions will help airports and airlines mitigate and prepare for consequences that result from flight cancellations, perhaps letting passengers know in advance the likelihood of cancellation so they are not stuck at the airport, or prepare to provide alternate options.

# Security Checklist, Privacy, and Other Risks

* Since datasets are from a public source, no PHI, PII, user behavior, or credit card data will be tracked and stored.
* The application will read and write to following S3 bucket: s3://sagemaker-us-east-1-993410942383/content-project/airline\_data/

## Risks associated with Bias

One of the main goals in this study is to predict whether a flight will get canceled or not based on weather conditions and other variables present in the data. Hence for this purpose, it is important to consider bias when it comes to our target variable (Canceled). Bias is when the predictor of interest has the same value for a large majority of data that can result in a model that is skewed towards one type of classification. For example, in this case, most flights are not canceled. The flight data set for January 2016 includes less than 1% canceled flights from the total number of flights. This includes all airports and since there are an overwhelming number of airports that do not experience cancellations or have relatively few, it makes more sense to analyze canceled flights due to weather in terms of location since weather is location-based. The focus for this study will be JFK as it is a major airport with a great variance in weather throughout the year. Upon further investigation, it is found that 6% of flights coming out of JFK are canceled (Figure 1). This corroborates with the idea that it may be possible that popular airports have a better track record of non-canceled flights than perception may hold. Even then, the repercussions of canceled flights have real consequences for people and airports, and it is still imperative to mitigate this with predictive modeling.

**Figure 1**

*Flight Cancellations for JFK Airport*

Chart, bar chart

Description automatically generated

This imbalance can cause a predictive model accuracy of 94% to seem favorable, when, the minority class is simply not detected, and all records are predicted as non-canceled. Therefore, an attempt was made to balance out the data to about 20-30%. Since the data has a time component (date of flight FlightDate, scheduled departure time CRSDep, and weather Start and End datetimes), adding data points would falsify the time constraint of flights. As is customary when using up or down sampling data to balance a minority class, only training data was balanced while validation and test data remained in their original form.

It is also possible that despite this imbalance, a classifier with good minority class detection can achieve an accuracy higher than 94% without balancing the data. So models were generated first with the original data without balancing to see if a model can still detect the minority class without altering the original data.

## Ethical Considerations

The ethical concerns that should be addressed with this data are to make sure correct matching is done with respect to weather conditions within the timeframe and location of the intended flight records used in the final dataset. The model should be as explainable as possible where important features are indicated, and coefficients are accessible. Also, airlines should not be allowed to use this data to control the narrative in a way that makes another company look negative. Weather cancellations is the focus, and because of this, the airline id was removed from the dataset for model training and validation.

# Data Preparation and Scrubbing

The first task for data scrubbing was a data quality check. The flight records data set contained 84 columns where 24 of these columns had over 80% null values. These columns had mostly irrelevant information about flight cancellations and were more about flight delays and diverted flights, so all these columns, plus unnamed and eventID were removed.

Since the model will predict whether a flight gets canceled or not due to weather, a particular airport of focus was chosen: JFK. This airport was chosen because it is famously busy and central to many flights both domestic and abroad. It is also a good location to use given the city experiences a more seasonal trend in weather. The dataset focused on domestic flights but the natural progression of this project will be to focus on all outgoing flights. Flight records for uncancelled flights and flights that were canceled due to weather were extracted from the dataset. This meant the removal of canceled flights due to Carrier/Security reasons (about 10 records).

Since both flight and weather data have a time and location component, some of the main parameters for matching up flight and weather data are the date, time, and location of the flight. In the flight data, both were separate and not in the correct format. Data cleaning involved formatting the time data (CRSDepTime) into the proper time format and then appending this to the FlightDate column and converting this to a datetime type. Moreover, the weather dataset had a mismatch in airport names by adding a ‘K’ prefix to the names. This was removed to match airport codes from flights and weather when weather information and flight information for JFK airport was merged. Lastly, some column names were changed to avoid any problems with SQL commands.

Two features were combined (Origin\_Weather\_Type and Origin\_Weather\_Severity) as an engineered categorical feature called Origin\_Weather\_Comb. This feature was used instead of the individual weather type and weather severity features.

A correlation matrix of the features revealed there are strong correlations between the temperature columns in the data. So, all were dropped except average temperature to avoid multicollinearity in model training.

**Figure 2**

*Correlation matrix for data variables*

Timeline

Description automatically generated

Therefore, the final features used were: DayOfWeek, Origin\_Weather\_Comb, Origin\_Weather\_Precipitation, Origin\_Weather\_First\_Snow, Origin\_Weather\_Snow\_Depth, Origin\_Weather\_AvgTemp, Origin\_Weather\_HDD, Origin\_Weather\_CDD. The output label for the dataset was a binary type called Canceled that indicated whether or not a flight was canceled.

Categorical features like Origin\_Weather\_Comb and DayOfWeek were hot-encoded. In case the distributions for numerical variables were skewed, numerical data was scaled and normalized.

One of the lookup datasets for AirportCodes combined information of airport, city, and state. That column was split into 3 separate columns in case it was needed.

## Train, Test, and Validation Splits

To find the best model for our data, data was split into training, validation, and test sets. The training set was for model training, the validation set for validating the model training configuration (hyper-parameter tuning), and the test set was used to test the final model with proper hyper-parameter configuration and evaluate the model’s performance. The data split method used stratified random sampling on the label (canceled) to split the data into 90%, 5%, 5% train, validation, and test, respectively, using the scikit-learn’s train\_test\_split function.

## Balancing data

A couple balancing methods were programmed for worst-case scenarios where classifiers could not find the minority class. These methods were, in the end, not used. In one case, a combination of both oversampling and undersampling methods created new data for the underrepresented class (1: 'Canceled') and removed existing data for the overrepresented class (0: ‘Not Canceled;). The goal was to increase the representation of the minority class 'Canceled' to around 30% since doing a 50-50 balance would lead to a loss of too much data. During the test and validation steps, the data was not balanced. This was done mostly as a preparatory step, and model training moved forward with unbalanced data to see if classifiers could still adequately predict the minority class without the need for balanced data.

# Modeling, Methods, Validation, and Performance Metrics

## Sagemaker Autopilot

Sagemaker Autopilot was first used to build and train a baseline model and provide proof of concept that an unbalanced dataset can yield high accuracy for both classes, as well as determining the baseline set of hyper-parameters. Autopilot was able to split the data so the data entered into autopilot was without splitting. After training, Autopilot found that a weighted ensemble model was the best model across different algorithms. It was a binary model whose objective was to maximize the ‘accuracy’ metric. It achieved an accuracy of 0.99, recall of 0.93, and precision of 0.93. It also achieved an AUC of 0.99. All metrics were above and beyond initial goals.

**Figure 3**

*Autopilot Weighted Ensemble Results*

Chart, line chart

Description automatically generated

## Scikit-Learn Logistic Regression

To employ a simple and fast model that is also less complex than ensemble methods (hence employing the law of parsimony), a “bring your own script” method was subsequently used. The Scikit-Learn logistic regression model was selected. Logistic regression works well with binary outcomes and the label variable (‘Canceled’) is a binary categorical variable (‘Canceled’ or ‘Not Canceled’). This method is fast and will help with future data that would be large (all flights departing from JFK airport on a daily basis) with a requirement for fast results so that airline companies can make timely decisions based on predicted results. Another benefit to using logistic regression is that it can provide likelihood outputs, which will help in future renditions of this project that seek to provide customers with cancellation likelihood rather than a simple binary prediction of canceled or not. Prior to training the model, the numerical columns were scaled to make it easy for a model to learn and understand the problem. The categorical columns were encoded to make sure they get transformed and they can be properly consumed. The performance metrics such as accuracy, precision, recall, and f1 score are provided below. They are divided based on class to show that the scores for the minority class (1) is better than expected.

**Table 1**

*Logistic regression class-based metrics*

Table

Description automatically generated

**Table 2**

*Logistic regression summary statistics*

Table

Description automatically generated with medium confidence

# Modeling Results and Findings

Accuracy, precision, recall, and f1 score all together were considered as model evaluation metrics. Accuracy measures the total number of predictions the model gets right and is important to the business goal in getting the right number of flight cancellations. On the other hand, the cost of failing to predict a flight cancellation when it is happening, is much higher than incorrectly predicting a flight cancellation when it is not happening. Thus, precision was also an essential metric as it measures the accuracy of positive predictions (flights being canceled). Another crucial metric was recall (or sensitivity), to measure how complete the positive predictions are. Lastly, since the target variable ‘Canceled’ has a class imbalance problem, f1-score is a necessary metric as it takes into consideration data distribution.

When it came to the autopilot model, the permissions did not allow the team to access the code or parameters used. The Logistic Regression algorithm, however, was validated for hyperparameter tuning. The parameters tuned were solver, penalty and C (the variable that controls the strength of the penalty). In the end, the best results were found with no penalty, using the newton-cg solver, and with a C value of 0.0019. It was also decided that the entire dataset would be input to the model as one instance, since there is no significant time delay with this model, and the data size is not excessive for the model to work with. In total, the final JFK airport weather dataset was 7,977 rows before splitting, and so the total test dataset size was 399 records.

## Final Model

Even though the weighted ensemble model in the autopilot model performed well, it did not follow the law of parsimony. It also was less explainable given there was no access to the coefficients or parameters that were tuned. Therefore, the logistic model was selected as the final model to employ the simplest, most explainable, fastest model for this use-case. The results are far better than expected, and nothing is lost in the process of using this model.

## Future Enhancements

1. Add event-based triggers to the model pipeline

Adding triggers will allow the pipeline to automate execution without the intervention of users. Since the business goal for this project is to provide timely information about flight changes to both airlines and customers, it will be beneficial if the pipeline starts itself when new data arrives at the system. It can be achieved through utilizing ‘GitOps’ automation on SageMaker which triggers updates to the model whenever new data lands in S3 from an upstream application. This continuous update will ensure the prediction results align with the most current weather situations simultaneously, allowing more time for airlines to address the flight cancellation problem.

1. Implement an automated mechanism for continuous feedback:

It will also be beneficial to introduce a fast-paced feedback environment to continuously detect operational issues and recommend fixes. In the future, the team should consider leveraging the Amazon DevOps to inject this automated practice into the model pipeline. By optimizing the DevOps process, it will shorten the feedback loop and help developers make better-informed decisions and deliver updated features as quickly as possible. Not only does it help promote coordination between the technical teams, including the development team and the deployment team, it also helps improve the operations of other non-IT functions such as customer service support and airline staff working front line.

1. Introduce human-in-the-loop workflows:

Eventually, more and more features about different weather situations are going to be added to train the model to take into consideration all possible weather conditions that could happen. Thus, there might be times when the model is unable to make a high-confidence prediction due to the vast variety of weather conditions. If a low-confidence prediction wrongly classifies a to-be-cancelled flight as a not-canceled flight, it could cause economic losses and negatively impact the airline’s reputation. To avoid this, one future enhancement is to add human judgment into the workflow to enable manual evaluation of any low-confidence machine learning predictions. This can be achieved by using the Amazon A2I, which can incorporate human review processes into the pipeline. This will help improve the model accuracy and make certain airlines get high-confidence predictions for decision making.

## Next Steps

The next steps would be to gather more data for flights and weather over a larger timeframe and develop similar models for all US airports, as well as develop a weather forecast model so that predictions can be done with a lead time of at least a month. This will help not only airports and airlines prepare for worst-case scenarios, but also help them develop tools where customers can be made aware of such predictions either at the time of booking or as their flight date comes near. Feasibly, they can provide alternate solutions to mitigate the high likelihood of canceled flights.

Another step could be to develop similar models to predict flight delays due to weather. This might be less straightforward but may be worth doing in cases where delays are extreme and happen due to the weather at either the destination airports or wherever the aircraft scheduled for that flight is coming from. This can allow airlines to properly route planes in advance when those risks are present. Furthermore, maintenance records of all aircrafts can be used to provide likelihood of delays/cancellations due to aircraft malfunction or repair.

**Conclusion**

Air travel has been under fire as of late for many reasons, especially post-pandemic; with violence in airports and even within the aircraft themselves due to unhappy, unruly, and some might even say, entitled customers. Add to this the frustrations of delayed or canceled flights and therein lies the recipe for disaster. The ability to predict with high accuracy the likelihood of a canceled flight will not only help airlines and airports provide options, opportunities, and relief for tired travelers, but may also, in future renditions of this project, help set realistic expectations for clients at time of booking – thus easing the blow of unmet expectations from canceled flights.

# References

Fountain, T. (2016, January 21). Airlines waive fees to change flights leading up to snow storm. Fox61 News. <https://www.fox61.com/article/travel/bucket-list/airlines-waive-fees-to-change-flights-leading-up-to-snow-storm/520-5097a9e4-071d-49e6-8d0b-242dd7efdd6f>

Fregly, C. & Barth, A. (2021). Data Science on AWS Implementing End-to-End, Continuous AI and Machine Learning Pipelines. O'Reilly.

Jacobsen, P. (2022, June 9). Consumer Complaints Are Up 300% at Airlines. What’s Going On? FEE Stories. <https://fee.org/articles/consumer-complaints-are-up-300-at-airlines-what-s-going-on/>

Kaggle. (2020). [US Weather Events (2016 - 2021)] [WeatherEvents\_Jan2016-Dec2021.csv]. <https://relational.fit.cvut.cz/dataset/Airline>

Moosavi, S. (2019, July 25). Short and Long-term Pattern Discovery Over Large-Scale Geo-Spatiotemporal Data. ACM Digital Library. <https://dl.acm.org/doi/10.1145/3292500.3330755>

NOWData NOAA Online Weather Data. National Weather Service. <https://www.weather.gov/wrh/Climate?wfo=okx>

Relational Dataset Repository. [airline][U.S. Department of Transportation (DOT), Bureau of Transportation Statistics (BTS)]. <https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-event>

# Appendix

1. <https://github.com/fausa/Flight-Research-and-Analytics-Group/blob/main/FinalProject_Team5.ipynb> [↑](#footnote-ref-1)