Prediction and Analysis of Ground Stops with Machine Learning

Eugene Mangortey*, Tejas G. Puranik†, Olivia J. Pinon‡, and Dimitri N. Mavris§
Georgia Institute of Technology, Atlanta, GA, 30332

A flight is considered to be delayed when it arrives 15 or more minutes later than scheduled. Delays attributed to the National Airspace System are one of the most common type of delays. Such delays may be caused by Traffic Management Initiatives (TMI) such as Ground Stops (GS), issued at affected airports. Ground Stops are implemented to control air traffic volume to specific airports where the projected traffic demand is expected to exceed the airports' acceptance rate over a short period of time due to conditions such as inclement weather, volume constraints, closed runways, etc. Ground Stops can be considered to be the strictest Traffic Management Initiative (TMI), particularly because all flights destined to affected airports are grounded until conditions improve. Efforts have been made over the years to reduce the impact of Traffic Management Initiatives on airports and flight operations. However, these efforts have largely focused on other Traffic Management Initiatives such as Ground Delay Programs (GDP), due to their frequency and duration compared to Ground Stops. Limited work has also been carried out on Ground Stops because of the limited amount of time that traffic management personnel often have between planning and implementing Ground Stops and external factors that influence decisions of traffic management personnel. Consequently, this research primarily focuses on the prediction of weather-related Ground Stops at Newark Liberty International (EWR) and LaGuardia (LGA) airports, with the secondary goal of gaining insights into factors that influence their occurrence. It is expected that this research will provide stakeholders with further insights into factors that influence the occurrence of weather-related Ground Stops at both airports. This is achieved by benchmarking Machine Learning algorithms in order to identify the best suited algorithm(s) for the prediction models, and identifying and analyzing key factors that influence the occurrence of weather-related Ground Stops at both airports. This is achieved by 1) fusing data from the Traffic Flow Management System (TFMS) and Automated Surface Observing Systems (ASOS) datasets, and 2) leveraging supervised Machine Learning algorithms to predict the occurrence of weather-related Ground Stops. The performance of these algorithms is evaluated using balanced accuracy, and identifies the Boosting Ensemble algorithm as the best suited algorithm for predicting the occurrence of Ground Stops at EWR and LGA. Further analysis also revealed that model performance is significantly better when using balanced datasets compared to imbalanced datasets.

I. Nomenclature

AAR	=	Airport Arrival Rates
ASPM	=	Aviation System Performance Metrics
ASOS	=	Automated Surface Observing Systems
CASSIE	=	Computing Analytics and Shared Services Integrated Environment
CSV	=	Comma-Separated Value
EDCT	=	Expected Departure Clearance Times
EWR	=	Newark Liberty International Airport
FAA	=	Federal Aviation Administration
FIXM	=	Flight Information Exchange Model

*Graduate Research Assistant, Aerospace Systems Design Laboratory, School of Aerospace Engineering, AIAA Student Member
†Research Engineer II, Aerospace Systems Design Laboratory, School of Aerospace Engineering, AIAA Member
‡Senior Research Engineer, Aerospace Systems Design Laboratory, School of Aerospace Engineering, AIAA Member
§Regents Professor for Advanced Systems Analysis, School of Aerospace Engineering, AIAA Fellow

FN = False Negative
FP = False Positive
GDP = Ground Delay Program
GS = Ground Stop
LGA = New York La Guardia Airport
NAS = National Airspace System
$NTML$ = National Traffic Management Log
$SMOTE$ = Synthetic Minority Over-sampling Technique
TAF = Terminal Aerodrome Forecast
$TFMS$ = Traffic Flow Management System
TMI = Traffic Management Initiative
TN = True Negative
TP = True Positive

II. Introduction and Motivation

Traffic Management Initiatives (TMI) are implemented by the Federal Aviation Administration (FAA) to balance demand and capacity in a constrained area of the National Airspace System (NAS) [1–4]. These constraints can be attributed to inclement weather, volume constraints, equipment failures, runway-related incidents, etc. From Figure 1, it can be seen that in 2017, the National Airspace System was mainly constrained because of inclement weather, followed by volume and closed runways. A better understanding of the impacts of such constraints on airports and flight operations can be achieved by analyzing Traffic Management Initiatives (TMI) such as Ground Stops.

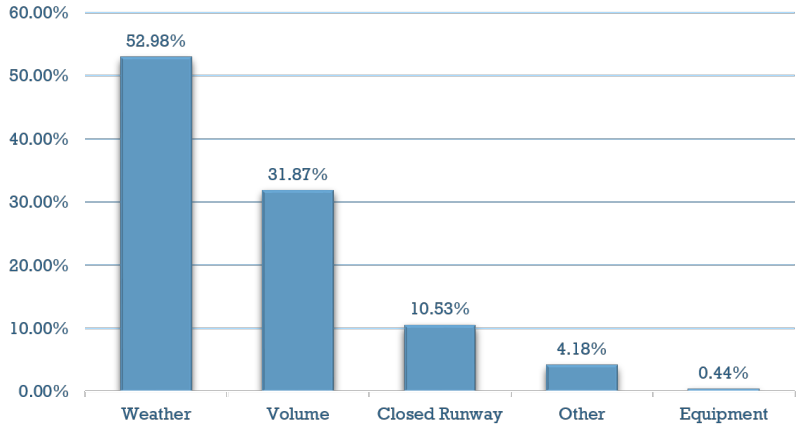


Fig. 1 Breakdown of causes of constraints in the National Airspace System (2017) [5]

A. Ground Stops

Ground Stops are TMI that are implemented to control air traffic volume to a specific airport where the projected traffic demand is expected to exceed the airport's acceptance rate for a short period of time due to conditions such as inclement weather, volume constraints, etc [6]. Whenever Ground Stops are issued, en-route flights are kept in airborne holding patterns or diverted, while flights that are yet to depart are grounded until the Ground Stop is terminated. This makes it difficult for airlines and passengers to determine departure times for affected flights because flights are not assigned runway release times, called Expected Departure Clearance Times (EDCT) [7]. Furthermore, if the duration of a Ground Stop is extended due to inaccurate predictions or a lack of improvement in airport conditions, airlines and passengers alike may incur additional delays. Figure 2 provides a breakdown of the incidence of Ground Stops at OPSNET 45 airports [8] in 2017. In particular, it shows that Ground Stops were predominantly caused by inclement weather conditions at majority of these airports.

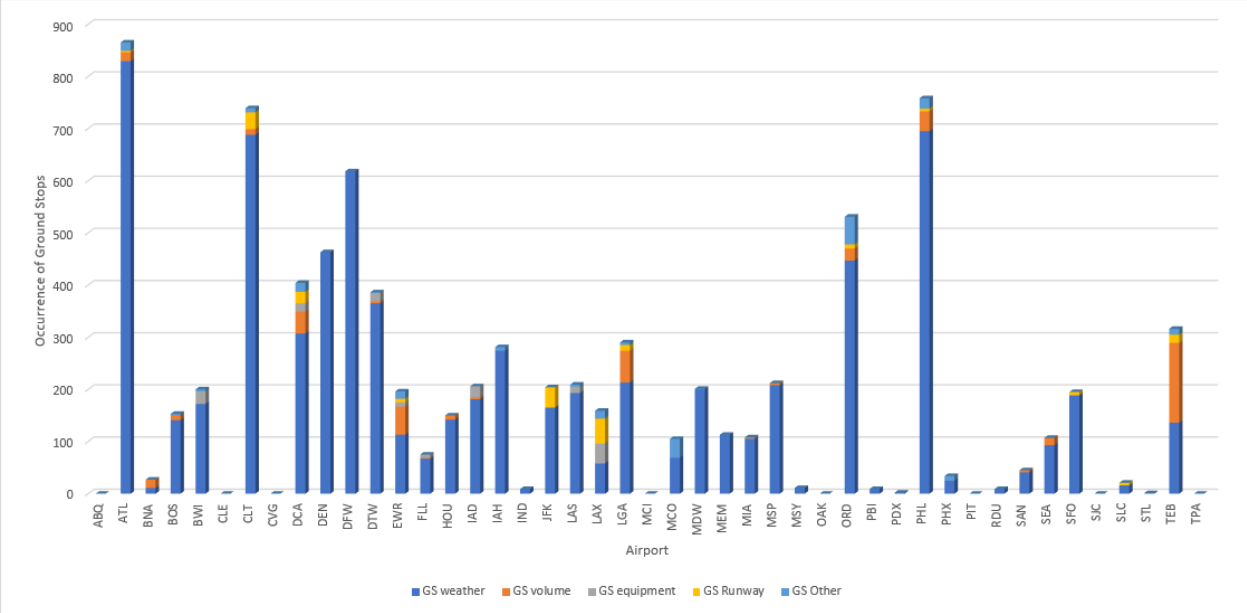


Fig. 2 Breakdown of Ground Stops by airport in 2017 [9]

Ground Stops can be considered to to be the strictest form of Traffic Management Initiatives, as they are only implemented as a last resort to ensure that the affected area of the NAS remains safe in spite of the constraint(s). This severely impacts airport and flight operations as all flights to the affected airport are not permitted to land until conditions improve. Furthermore, a Ground Stop at one airport may impact operations at other airports or the entire National Airspace System (NAS) due to the propagation of delays in the National Airspace System (NAS). The implementation of Ground Stops can also be considered to be more complicated compared to other Traffic Management Initiatives due to the limited amount of time that traffic management personnel often have between planning and implementing a Ground Stop. The complex nature of Ground Stops is further compounded by the severity of conditions at hand, as well as external factors such as airline-related delays, that influence decisions of traffic management personnel. Consequently, little work has been done to predict Ground Stops.

B. Review of prior research related to Ground Stops (GS)

Wang [10] used Ground Stop data from the National Traffic Management Log (NTML) database, airport data from the Aviation System Performance Metrics (ASPM) database, and weather data from the Rapid Update Cycle (RUC) database, to predict the occurrence of weather-related Ground Stops using the Ensemble Bagging Decision Tree Machine Learning algorithm. Even though the model had an accuracy of 85%, the model's performance could be improved by benchmarking Machine Learning algorithms to identify the best suited algorithm for the prediction model.

Although other efforts related to reducing the impact of Traffic Management Initiatives on airports and flight operations have largely focused on other TMI such as Ground Delay Programs (GDP) and Reroutes [2–4, 11–14], limited work has been carried out on Ground Stops. Consequently, this research focuses on benchmarking Machine Learning algorithms to identify a suitable algorithm for predicting the occurrence of weather-related Ground Stops.

Prior work related to Traffic Management Initiatives has also involved using heavily imbalanced datasets [3] which produced poorly performing prediction models. This research addresses this gap by reducing the imbalanced nature of data used. The need for utilizing balanced datasets for Machine Learning purposes is also highlighted by comparing the performance of prediction models developed with imbalanced and balanced datasets using metrics such as balanced accuracy.

Finally, Ground Stops can be considered to be the strictest Traffic Management Initiative. The implementation of Ground Stops in congested areas of the National Airspace System such as the airspace around New York City [15, 16] further constrains operations. Consequently, predicting and analyzing the occurrence of weather-related Ground Stops at two airports in the New York City airspace: Newark Liberty International (EWR) and LaGuardia (LGA) airports may provide a means for stakeholders to make more informed decisions to improve operations in that airspace. This research

will also lay out a framework for future studies at other airports as the same analysis can be replicated for other airports.

C. Research Objectives

As previously mentioned, Ground Stops can be considered to be the strictest Traffic Management Initiative, particularly because all flights to affected airports are grounded until the Ground Stop is terminated. This disrupts flight operations not only to the affected airport, but may also have an impact on operations in the National Airspace System (NAS). Several factors including weather conditions as well as external factors influencing the decisions of traffic management personnel are considered prior to and during the implementation of Ground Stops. Consequently, the overaching objective of this research is to predict and analyze the occurrence of Ground Stops using weather conditions, and to provide insights which may assist stakeholders to better understand, plan, and implement Ground Stops. This objective can be divided into four parts:

1) Predict the occurrence of weather-related Ground Stops at EWR and LGA by benchmarking Machine Learning algorithms
2) Highlight the benefit or need for using balanced datasets by evaluating the performance of prediction models developed with imbalanced and balanced datasets
3) Identify key factors that influence the occurrence of Ground Stops at EWR and LGA so as to help stakeholders make better decisions
4) Determine if the prediction models should be on an individual airport basis

The remainder of this paper discusses the methodology used for this research and obtained results. Sections III, IV, and V discuss the methodology used, obtained results, and concluding remarks and future work.

III. Methodology

Figure 3 provides an overview of the methodology used to address the aforementioned research objectives.The remainder of this section discusses each step of the methodology in detail.

Fig. 3 Overview of Methodology

A. Identify datasets

The following datasets containing information about Ground Stops and weather conditions are leveraged and used for this research:

- Traffic Flow Management System (TFMS)
- Automated Surface Observing System (ASOS)

1. Traffic Flow Management System (TFMS)

The Traffic Flow Management System (TFMS) is used to plan and execute traffic flow management initiatives to ensure that demand and capacity are balanced in the National Airspace System [17]. TFMS data is transmitted in two streams. The first stream, TFMS Flight, contains initial flight plan messages, amended flight plan messages, departure and arrival time notifications, flight cancellation messages, boundary crossing messages, and track position reports. The second stream, TFMS Flow, provides data on traffic flow management initiatives such as Ground Stops, Reroutes, Airspace Flow Programs, etc [17]. These datasets are obtained from the FAA's Computing Analytics and

Shared Services Integrated Environment (CASSIE), which brings FAA divisions, partners, and stakeholders together in a shared services environment consisting of Big Data, computing power, and analytical tools [18]. Ground Stop data provided by the TFMS dataset includes:

- Start and end dates and times of the Ground Stop
- Advisory Number
- Cause of Ground Stop (weather, volume, equipment, runway, other)
- Details of Ground Stop (wind, snow, runway construction, etc.)
- Probability of extending Ground Stop (low, medium, high)
- Affected airport

2. Automated Surface Observing Systems (ASOS)

The Automated Surface Observing Systems (ASOS) dataset provides weather data at airports which is widely used by meteorologists, climatologists, hydrologists, and aviation weather experts [19, 20]. The ASOS dataset contains actual weather data recorded at five minute intervals and provides the following weather parameters for each airport:

- Date and time
- Air temperature (Fahrenheit)
- Dew point temperature (Fahrenheit)
- Relative humidity (%)
- Wind direction (degrees)
- Wind speed (knots)
- Precipitation (inches)
- Pressure altimeter (inches)
- Sea level pressure (millibars)
- Visibility (statute miles)
- Wind gusts (knots)
- Cloud coverage type (overcast, scattered etc.) and altitude (feet)
- Ice accretion over 1, 3, and 6 hours (inches)
- Peak wind gust (knots)
- Peak wind direction (degrees)

The ASOS data used for this research was obtained from the ASOS database and downloaded in csv format [21].

B. Parse data

In order to utilize the data required for this research, the datasets should be converted from their raw formats into useful formats, and the necessary data extracted. This subsection covers the steps taken to parse the datasets.

1. Traffic Flow Management System (TFMS)

The Traffic Flow Mangement System (TFMS) datasets are stored in Flight Information Exchange Model (FIXM) [22] format, which is widely used for storing and transmitting aviation data. These datasets, which are stored as hourly files contain advisories generated during that hour, and need to be parsed from FIXM format to csv format. FIXM files have schema files which dictate the structure of the files. Consequently, FIXM files should be parsed using their respective schema to ensure that all required fields are extracted in their correct format. This is done using a Python [23] parser developed by Mangortey et al. [2], which follows the process highlighted in Figure 4 and is described below:

1) Since the datasets are comprised of advisories generated within the hour, there is no way to distinguish between the beginning of the file and the end of the file. Thus, it is important to enclose each file with a header and footer such as <root > and <\root >, respectively, to ensure that each file has unique starting and end points
2) Extract the schema location from the xsd file. The schema location is typically of the format "xlmns:....."
3) Parse the FIXM file using the ElementTree [24] Application Program Interface (API)
4) Extract "Active" weather and volume-related Ground Stop advisories for EWR and LGA airports

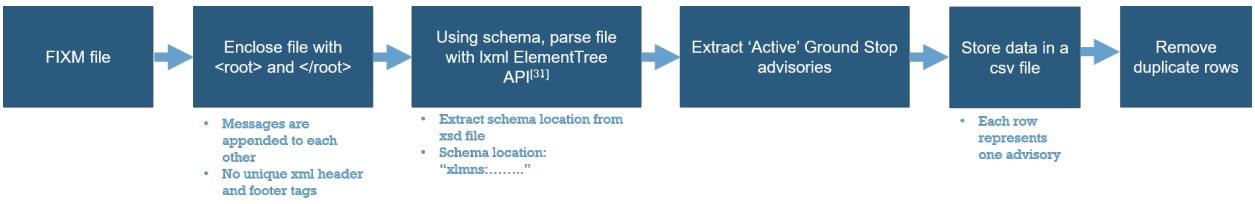


Fig. 4 FIXM to JSON conversion process

5) Store each Ground Stop advisory as a row in a csv file

Ground Stop advisories in the TFMS dataset can be of three forms:

1) Proposed: Ground Stop has been announced but has not been commenced
2) Active: Ground Stop is ongoing
3) Cancelled: Ground Stop has been terminated

Active Ground Stops advisories at the EWR and LGA airports from January 1 to August 13, 2017 were parsed into csv format and used for this research.

2. Automated Surface Observing Systems (ASOS)

Automated Surface Observing Systems (ASOS) data was extracted from the ASOS database [25] in csv format. The following parameters were extracted for EWR and LGA airports from January to August 2017:

- Date and time
- Air Temperature (Fahrenheit)
- Dew Point Temperature (Fahrenheit)
- Relative Humidity (%)
- Wind Direction (Degrees)
- Wind Speed (Knots)
- Precipitation Accumulation (Inches)
- Pressure Altimeter (Inches)
- Visibility (Miles)
- Wind Gusts (Knots)
- Cloud Coverage Type
- Cloud Altitude (Feet)

C. Clean datasets

The next step in the methodology focuses on identifying inconsistent and/or missing data and cleaning the datasets.

1. Traffic Flow Management System (TFMS)

1) The first step of the data cleaning process involves analyzing the data to ensure that fields are in their appropriate formats and do not contain any missing values or non-alphanumeric characters
2) The next step involves removing duplicate Ground Stop advisories. Duplicate advisories exist because TFMS occasionally stores the same Ground Stop advisory multiple times
3) The duration and scope of an ongoing Ground Stop may be modified whenever conditions change. This leads to overlapping Ground Stop advisories which is inaccurate. In order to address this inconsistency, the end time of the initial Ground Stop advisory is set as the start time of the new Ground Stop advisory as seen in Figure 5. In particular it shows that the start time of advisory number **0117** is prior to the end time of advisory number **0071**. Thus, the end time of advisory number **0071** is set as the start time of advisory number **0117**.

Advisory Number	Start Time	End Time
0071	2017-04-20T16:00:00Z	2017-04-21T03:00:00Z
0117	2017-04-20T18:15:00Z	2017-04-21T03:00:00Z

Advisory Number	Start Time	End Time
0071	2017-04-20T16:00:00Z	2017-04-20T18:15:00Z
0117	2017-04-20T18:15:00Z	2017-04-21T03:00:00Z

Fig. 5 Updating the end dates and times of updated active Ground Stop advisories

2. Automated Surface Observing Systems (ASOS)

The ASOS dataset was analyzed to ensure that fields were in their appropriate formats and to identify any fields with missing values. The ASOS datasets are recorded in five minute intervals. Thus, rows containing missing values were deleted to ensure uniformity. In addition, fields such as cloud coverage type and altitude, ice accretion, and peak wind gust and direction were not used for this research as over 80% of these fields contained missing values.

D. Fuse datasets

The next step in the methodology focuses on fusing the TFMS and ASOS datasets for EWR and LGA. Data Fusion is a method of data analysis that involves fusing data from different sources to produce more consistent and useful information than that obtained from a single data source [26]. The datasets are fused by date and time, and the occurrence of a Ground Stop serves as the target of the prediction models. Predictors include weather conditions, month, and hour of day.

E. Develop prediction models

1. Model generation, validation, and testing

Lantz [27] defines Machine Learning as "the field of study interested in the development of computer algorithms to transform data into intelligent actions". Machine Learning has been used in the aviation industry to identify parameters for flight risk identification [28], analyze the coincidence of Ground Delay Programs and Ground Stops [29], classify, predict, and analyze the daily operations of airports [30], detect flight anomalies during the approach phase [31], etc. This research focuses on using supervised learning algorithms to develop the prediction models. Supervised learning is the process of training a Machine Learning model to predict value(s) using other values in the dataset. In particular, supervised learning algorithms attempt to discover and model the relationship between the value(s) being predicted and other values (predictors). The supervised learning algorithms benchmarked to identify the best suited algorithm for predicting the occurrence of Ground Stops are Bagging Ensemble [27], Boosting Ensemble [27], and Random Forests [27].

Figure 6 provides an overview of the model generation, validation, and testing process. First, the fused dataset is randomly partitioned into three sets: training, validation and testing. Half of the data is assigned to the training set, which is used to train the model, one-fourth of the data is assigned to the validation set, which is used to iterate and refine the model, and one-fourth of the data is assigned to the test set, which is used to generate predictions for evaluations [27].

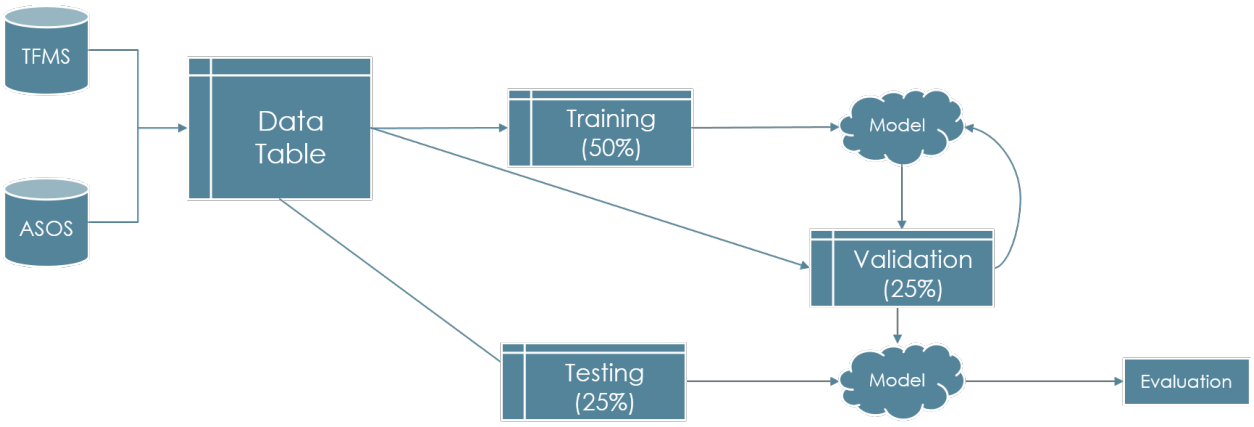


Fig. 6 Model Generation, Validation and Testing Process

As mentioned in Section II, a gap to be addressed by this research focuses on highlighting the need for using balanced datasets when leveraging Machine Learning techniques. This is achieved by developing prediction models using balanced and imbalanced datasets, and comparing their performance using evaluation metrics. The imbalanced nature of datasets can be addressed by implementing the Synthetic Minority Over-sampling Technique (SMOTE) algorithm [32, 33] on the imbalanced dataset. The SMOTE algorithm balances a dataset by increasing the minority class. This is achieved by randomly selecting a k-nearest-neighbor of each member of the minority class. Implementing the SMOTE algorithm instead of naively oversampling the minority class ensures that over-fitting is avoided. However, the SMOTE algorithm cannot be used for very large datasets. Finally, members of the majority class are also randomly under-sampled to further balance the data.

F. Evaluation

Evaluating the performance of prediction models is an important step as it informs as to how the model will perform on future data. In the context of a classification problem, such as this one, prediction models can be evaluated using results obtained from a confusion matrix, which categorizes predictions according to whether they match the actual value, as seen in Table 1. Performance metrics such as accuracy, sensitivity, specificity, and balanced accuracy are then computed to assess model performance [27].

Table 1 Confusion Matrix

	Actual: GS	**Actual: No GS**
Predicted: GS	True Positive (TP)	False Negative (FN)
Predicted: No GS	False Positive (FP)	True Negative (TN)

True Positive (TP) refers to the correct classification of the class of interest. True Negative (TN) refers to the correct classification of the class that is not of interest. False Positive (FP) refers to the incorrect classification of the class of interest. False Negative (FN) refers to the incorrect classification of the class that is not of interest [27].

1. Accuracy

This refers to the ratio of the number of true positives and negatives, to the total number of predictions. Accuracy varies from 0 to 1 and is specified as [27]:

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

2. Sensitivity

This refers to the proportion of true positives that were correctly classified. Sensitivity varies varies from 0 to 1 and is specified as [27]:

$$Sensitivity = \frac{TP}{TP + FN}$$

3. Specificity

This refers to the proportion of negative examples that were correctly classified. Specificity varies from 0 to 1 and is specified as [27]:

$$Specificity = \frac{TN}{FP + TN}$$

4. Balanced Accuracy

A model might have high accuracy because it correctly predicts the most frequent class, particularly when the dataset is imbalanced. Balanced accuracy adjusts accuracy by calculating the average of accurate predictions in each

class [34, 35] and is specified as:

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2}$$

IV. Results & Analysis

This section provides an overview and analysis of results from this research to achieve the following research objectives:

1) Predict the occurrence of weather-related Ground Stops at EWR and LGA by benchmarking Machine Learning algorithms
2) Highlight the benefit or need for using balanced datasets by evaluating the performance of prediction models developed with imbalanced and balanced datasets
3) Identify key factors that influence the occurrence of Ground Stops at EWR and LGA

The fused datasets for EWR and LGA were comprised of weather and Ground Stop data from January 1 to August 13, 2017. The fused EWR dataset was comprised of 5142 cases of No Ground Stops and 306 cases of Ground Stops. The fused LGA dataset on the other hand was comprised of 4765 cases of No Ground Stops and 385 cases of Ground Stops. The average training and prediction times for the algorithms were 2 minutes and 23 seconds on a desktop computer, respectively. The low prediction time indicates that this process is computationally inexpensive can be easily replicated and implemented.

A. Predicting the occurrence of weather-related Ground Stops

As mentioned previously, the fused data for EWR and LGA was randomly split into three sets: training, validation, and testing. The training, validation, and testing sets were used to train, tune, and evaluate model performance, respectively. The SMOTE algorithm was implemented to reduce the imbalanced nature of the EWR training set from 2566 cases of No Ground Stops and 158 cases of Ground Stops to 948 and 1106, respectively. The heavily imbalanced training set for LGA was comprised of 2381 cases of No Ground Stops and 194 cases of Ground Stops. The SMOTE algorithm was implemented to reduce the training set's imbalanced nature to 1164 cases of No Ground Stops and 1358 cases of Ground Stops. The performance of the Bagging Ensemble, Random Forests, and Boosting Ensembles algorithms with the balanced and imbalanced datasets were then compared with the same testing sets to identify the best suited algorithm for predicting the occurrence of Ground Stops at EWR and LGA.

1. Bagging Ensemble algorithm

Tables 2, 3, 4, and 5 show the confusion matrices for the Bagging Ensemble algorithm with the imbalanced and balanced EWR and LGA datasets, respectively. The confusion matrices show that the Bagging Ensemble algorithm predicted fewer cases of Ground Stops with the imbalanced dataset compared to the balanced dataset. This can be attributed to the fact that the models were trained with a significantly larger number of No Ground Stop cases. Thus, the models were more likely to predict No Ground Stop cases.

Table 2 Imbalanced EWR dataset

	Actual GS	**Actual No GS**
Predicted GS	26	4
Predicted No GS	50	1281

Table 3 Balanced EWR dataset

	Actual GS	**Actual No GS**
Predicted GS	55	338
Predicted No GS	21	947

Table 4 Imbalanced LGA dataset

	Actual GS	Actual No GS
Predicted GS	49	10
Predicted No GS	43	1186

Table 5 Balanced LGA dataset

	Actual GS	Actual No GS
Predicted GS	73	61
Predicted No GS	19	1135

Tables 6 and 7 provide a comparison of the performance of the algorithm with the imbalanced and balanced datasets using evaluation metrics. In particular, they show that the the models had higher accuracy with the imbalanced datasets. However, this assessment is not accurate as the models predicted majority of No Ground Stop cases and failed to predict majority of Ground Stop cases. Balanced accuracy thus serves as a more appropriate metric for assessing model performance as it averages the number of accurate predictions in each class. In addition, it can be seen that models had higher balanced accuracy with the balanced dataset compared to the imbalanced dataset. This highlights a benefit of using balanced datasets compared to imbalanced datatsets. Finally, it can be seen the model performance with the Bagging Ensemble algorithm for LGA was better compared to EWR.

Table 6 Comparison of the performance of the Bagging Ensemble algorithm with the imbalanced and balanced EWR datasets

Metric	Imbalanced dataset	Balanced dataset
Accuracy	0.9603	0.7362
Sensitivity	0.3421	0.7237
Specificity	0.9969	0.7367
Balanced Accuracy	0.6695	0.7303

Table 7 Comparison of the performance of the Bagging Ensemble algorithm with the imbalanced and balanced LGA datasets

Metric	Imbalanced dataset	Balanced dataset
Accuracy	0.9589	0.9379
Sensitivity	0.5326	0.7935
Specificity	0.9916	0.949
Balanced Accuracy	0.7621	0.8712

2. Random Forests algorithm

Tables 8, 9, 10, and 11 show the confusion matrices for the Random Forests algorithm with the imbalanced and balanced EWR and LGA datasets, respectively. Tables 12 and 13 also provide a comparison of the performance of the algorithm with the imbalanced and balanced datasets with evaluation metrics. Similarly to the Bagging Esemble algorithm, the prediction models better predicted the occurrence of Ground Stops and had higher balanced accuracy with the balanced datasets compared to the imbalanced datasets. Finally, model performance was also better for LGA compared to EWR.

Table 8 Imbalanced EWR dataset

	Actual GS	Actual No GS
Predicted GS	26	4
Predicted No GS	50	1281

Table 9 Balanced EWR dataset

	Actual GS	Actual No GS
Predicted GS	47	99
Predicted No GS	29	1186

Table 10 Imbalanced LGA dataset

	Actual GS	Actual No GS
Predicted GS	57	3
Predicted No GS	35	1193

Table 11 Balanced LGA dataset

	Actual GS	Actual No GS
Predicted GS	73	17
Predicted No GS	19	1179

Table 12 Comparison of the performance of the Random Forests algorithm with the imbalanced and balanced EWR datasets

Metric	Imbalanced dataset	Balanced dataset
Accuracy	0.9603	0.906
Sensitivity	0.3421	0.6184
Specificity	0.9969	0.9229
Balanced Accuracy	0.6695	0.7707

Table 13 Comparison of the performance of the Random Forests algorithm with the imbalanced and balanced LGA datasets

Metric	Imbalanced dataset	Balanced dataset
Accuracy	0.9705	0.972
Sensitivity	0.6196	0.7935
Specificity	0.9975	0.9858
Balanced Accuracy	0.8085	0.8896

3. Boosting Ensemble algorithm

Tables 14, 15, 16, and 17 show the confusion matrices for the Boosting Ensemble algorithm with the imbalanced and balanced EWR and LGA datasets, respectively. Tables 18 and 19 also provide a comparison of the performance of the algorithm with the imbalanced and balanced datasets. Similarly to the Bagging Ensemble and Random Forest algorithms, the prediction models better predicted the occurrence of Ground Stops and had higher balanced accuracy with the balanced datasets compared to the imbalanced datasets. Finally, model performance was also better for LGA compared to EWR.

Table 14 Imbalanced EWR dataset

	Actual GS	Actual No GS
Predicted GS	31	7
Predicted No GS	45	1278

Table 15 Balanced EWR dataset

	Actual GS	Actual No GS
Predicted GS	49	70
Predicted No GS	27	1215

Table 16 Imbalanced LGA dataset

	Actual GS	Actual No GS
Predicted GS	74	4
Predicted No GS	18	1192

Table 17 Balanced LGA dataset

	Actual GS	Actual No GS
Predicted GS	80	27
Predicted No GS	12	1169

Table 18 Comparison of the performance of the Boosting Ensemble algorithm with the imbalanced and balanced EWR datasets

Metric	Imbalanced dataset	Balanced dataset
Accuracy	0.9618	0.9287
Sensitivity	0.4079	0.6447
Specificity	0.9946	0.9455
Balanced Accuracy	0.7012	0.7951

Table 19 Comparison of the performance of the Boosting Ensemble algorithm with the imbalanced and balanced LGA datasets

Metric	Imbalanced dataset	Balanced dataset
Accuracy	0.9829	0.9697
Sensitivity	0.8044	0.8696
Specificity	0.9967	0.9774
Balanced Accuracy	0.9	0.9235

B. Identification of suitable algorithm for predicting the occurrence of Ground Stops

Figures 7 and 8 provide a comparison of the performance of Machine Learning algorithms in predicting the occurrence of Ground Stops at EWR and LGA, using Balanced Accuracy. The comparison shows that the algorithms performed better with the balanced datasets. It also shows that the **Boosting Ensemble algorithm** is the best suited for predicting the occurrence of Ground Stops at EWR and LGA.

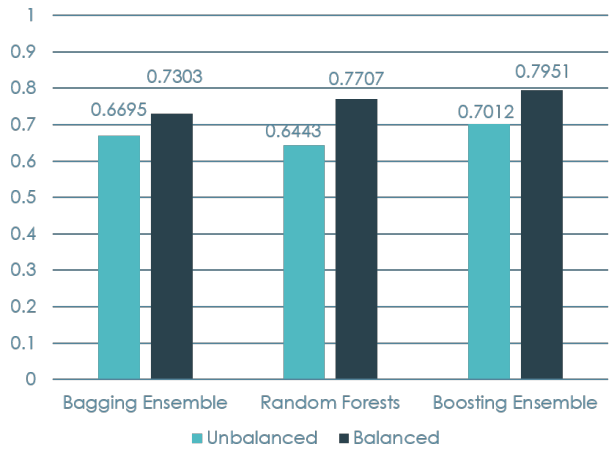


Fig. 7 Comparison of Machine Learning algorithms using Balanced Accuracy for EWR

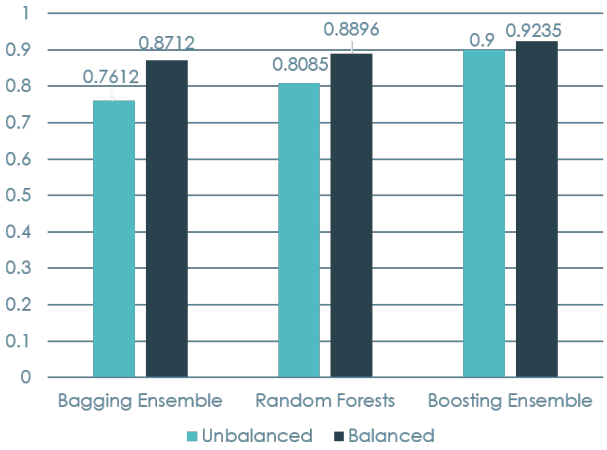


Fig. 8 Comparison of Machine Learning algorithms using Balanced Accuracy for LGA

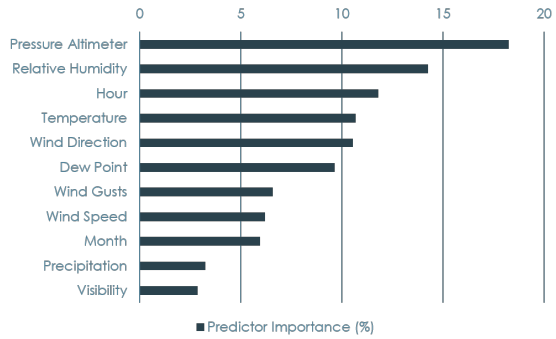
C. Ranking of predictors

Ranking predictors in order of their importance provides further insight into how predictors influence the performance of the prediction models. This subsection provides a comparison of predictor importance of the three Machine Learning algorithms with the balanced and imbalanced EWR and LGA datasets.

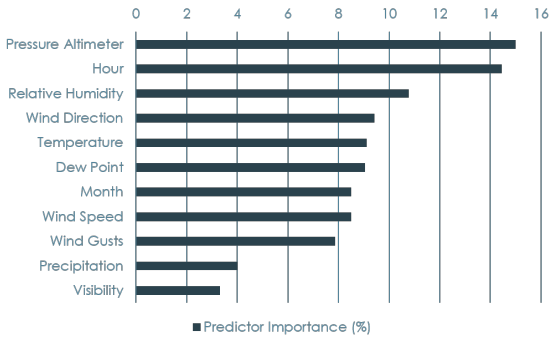
Figures 13 and 14 in the Appendix show the ranking of predictor importance for the prediction models developed with the Bagging Ensemble algorithm for EWR and LGA, respectively. Figure 13 shows that the ranking of predictor importance for EWR differed considerably with the balanced and imbalanced datasets. It also shows that model performance with the balanced dataset was primarily influenced by Hour, followed by Precipitation and Pressure Altimeter. However, predictor importance was distributed across parameters with the imbalanced dataset. Figure 14 shows that model performance with the balanced and imbalanced LGA datasets was largely influenced by three predictors: Hour, Month, and Pressure Altimeter. It also shows that the predictors were ranked almost similarly with the balanced and imbalanced LGA datasets.

Figures 15 and 16 in the Appendix show the ranking of predictor importance for the prediction models developed with the Random Forests algorithm for EWR and LGA, respectively. Figure 15 shows that the model developed with the balanced EWR dataset was primarily influenced by Hour. However, the model developed with the imbalanced EWR dataset was influenced by multiple predictors. It also shows that the predictors were ranked almost similarly with the balanced and imbalanced EWR datasets. On the other hand, Figure 16 shows that model performance with the balanced and imbalanced LGA datasets was largely influenced by three predictors in the same order: Pressure Altimeter, Hour, and Month.

Figures 9 and 10 show the ranking of predictor importance for the prediction models developed with the Boosting Ensemble algorithm for EWR and LGA, respectively. They show that the ranking of predictors of the models developed with the balanced and imbalanced was similar for both, EWR and LGA.

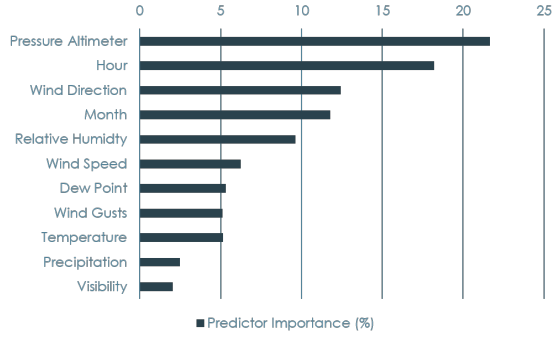


(a) Imbalanced EWR dataset

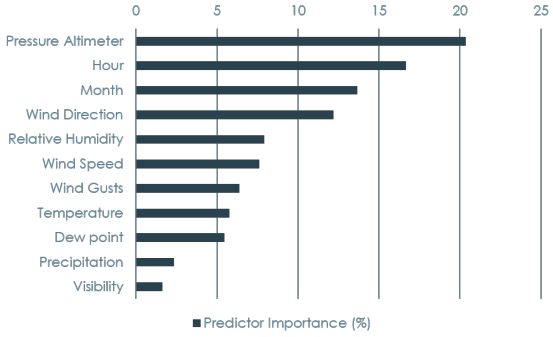


(b) Balanced EWR dataset

Fig. 9 EWR predictor importance with Boosting Ensemble algorithm



(a) Imbalanced LGA dataset



(b) Balanced LGA dataset

Fig. 10 LGA predictor importance with Boosting Ensemble algorithm

The ranking of predictors across the three Machine Learning algorithms with the balanced and imbalanced datasets for LGA and EWR were then compared to identify any similarities and/or differences. This comparison revealed that Hour, Month, and Pressure Altimeter were the key predictors across all algorithms with the balanced and imbalanced LGA datasets. Hour and Pressure Altimeter were also identified as the key predictors across all algorithms with the balanced and imbalanced EWR datasets. The comparison also identified Visibility as the most common, least influential predictor of the balanced EWR and LGA datasets, across all algorithms. Precipitation was identified as the most common, least influential predictor of the imbalanced EWR and LGA datasets, across all algorithms.

Tables 20 and 21 provide the average ranking of the key predictors in order of importance across the three Machine Learning algorithms with the balanced and imbalanced datasets for LGA and EWR, respectively. Table 20 shows that the average ranking of predictors was similar with the balanced and imbalanced LGA datasets. However, Table 21 shows that the average ranking of predictors differs between the balanced and imbalanced EWR datasets. Both tables also show that except for Hour and Pressure Altimeter, the average ranking of predictors differs at EWR and LGA. This may explain the difference in the performance of the Machine Learning algorithms with the EWR and LGA datasets. However, further analysis is required to better understand the differences in the average ranking of EWR and LGA predictors, and will be carried out as part of future work. This observation also indicates the prediction models should be on an individual airport basis.

This analysis provides an explanation for the variation in the performance of the Machine Learning algorithms. It also provides insight into predictors that influence the occurrence of Ground Stop at EWR and LGA.

Table 20 Ranking of key predictors for LGA

Balanced dataset	Imbalanced dataset
Pressure Altimeter	Hour & Pressure Altimeter
Hour	Month
Month	Wind Direction
Wind Direction	Relative Humidity
Wind Speed	Wind Speed
Relative Humidity	Dew Point & Visibility
Precipitation	Precipitation & Temperature
Temperature	Wind Gusts
Visibility & Dew Point	
Wind Gusts	

Table 21 Ranking of key predictors for EWR

Balanced dataset	Imbalanced dataset
Hour	Pressure Altimeter
Pressure Altimeter	Relative Humidity
Dew Point	Hour
Temperature	Dew Point
Relative Humidity	Temperature
Precipitation	Wind Direction
Wind Direction	Wind Gusts
Month & Wind Speed	Wind Speed
Wind Gusts	Month
Visibility	Precipitation
	Visibility

D. Analysis of the distribution of predictors

Analyzing the distribution of predictors can also be used to gain insights into, and validate the ranking of predictor importance. Figure 11 show the distribution of predictors of the balanced dataset for EWR using density plots. In particular, it shows that the distribution of No Ground Stop cases across all hours remains relatively constant while the occurrence of Ground Stops is much higher in the latter hours of the day. Figure 11 also shows that there are some variations in the distribution of Pressure Altimeter, Dew Point, Temperature, Relative Humidity, and Wind Direction between Ground Stop and Non-Ground Stop cases. Table 21 shows that the aforementioned parameters with variations in their distributions influenced the performance of the prediction models. However, the differences in the distributions of these parameters between Ground Stop and No Ground Stop cases is smaller, compared to that of Hour, and may explain why Hour is the highest, average ranked predictor. The distributions of No Ground Stop and Ground Stop cases of parameters such as Wind Gusts, Wind Speed and Visibility were very similar, validating their ranking as the least influential predictors.

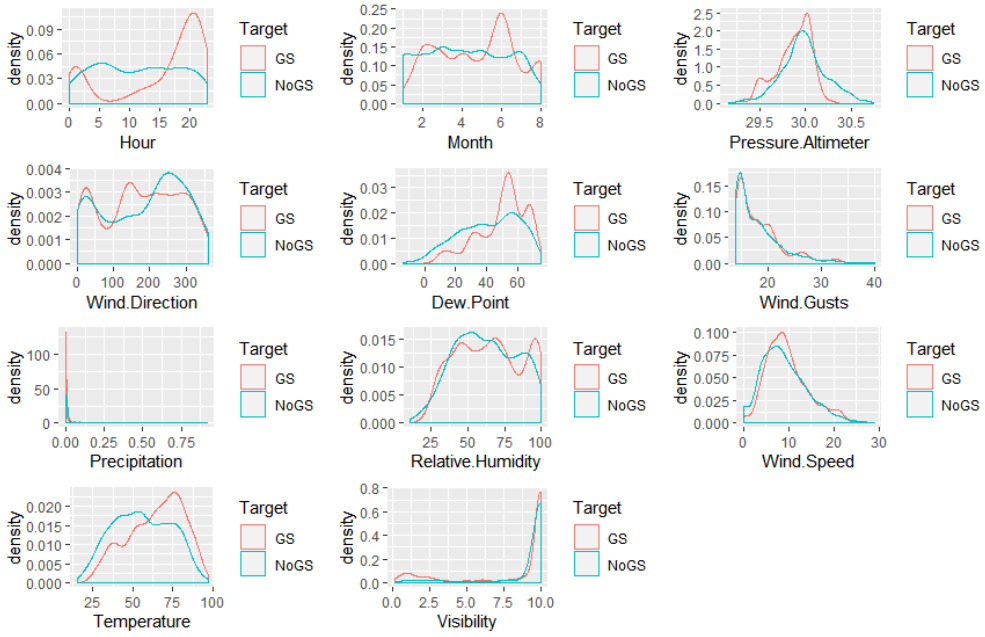


Fig. 11 Distribution of EWR predictors using a density plot

Figure 12 show the distribution of predictors of the balanced dataset for LGA using density plots. Similar to EWR, the occurrence of Ground Stops is much higher in the latter hours of the day. There are also significant variations in the distribution of Month, Pressure Altimeter, Wind Direction, and Relative Humidity between Ground Stop and No Ground Stop cases. These predictors with significant variations in their distributions are consistent with the top predictors identified from the average ranking of key predictors in Table 20. It can also be seen that the distributions of predictors is a lot smoother in the EWR dataset compared to the LGA. This variation provides an explanation for the differences in the performance of the algorithms with the EWR and LGA datasets, and the average ranking of predictors. The significant variations in the distributions of Ground Stop and No Ground Stop cases with the LGA dataset enables the algorithms to better predict the occurrence of Ground Stops.

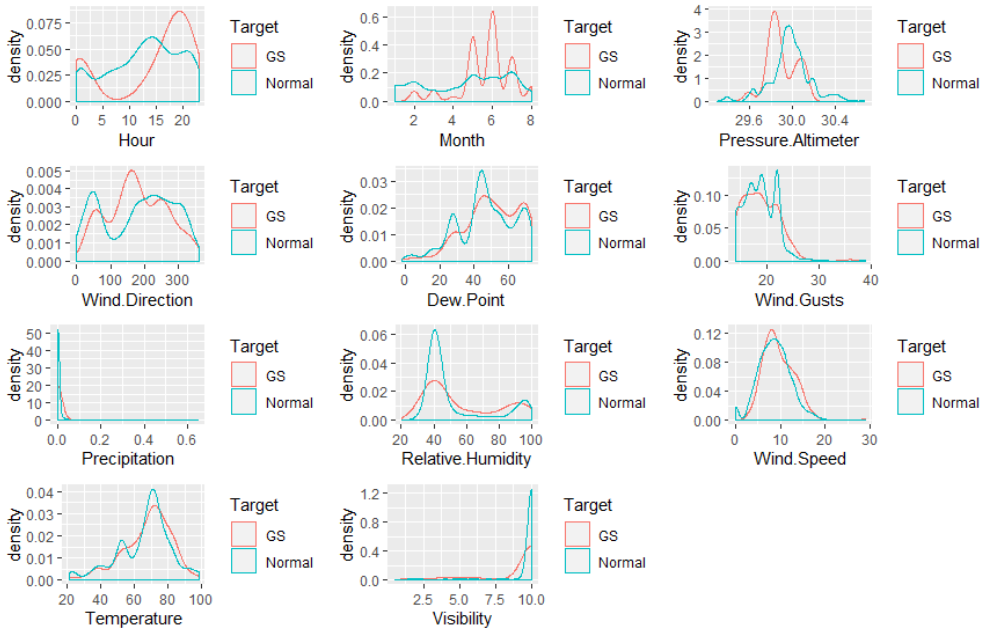


Fig. 12 Distribution of LGA predictors using a density plot

E. Summary of Findings

Below are a summary of findings obtained from this research:

1) The **Boosting Ensemble** algorithm was identified as the best suited algorithm for predicting the occurrence of weather-related Ground Stops at EWR and LGA
2) The prediction model for LGA performed better than the model for EWR because of differences in the distributions of parameters between Ground Stop and No Ground Stops cases
3) The Machine Learning algorithms performed better with the balanced datasets
4) The ranking of predictor importance revealed that different factors, except for Hour and Pressure Altimeter, influenced the prediction of the occurrence of Ground Stops at EWR and LGA. This is due to differences in the distributions of parameters between Ground Stop and No Ground Stops case
5) The prediction models should be on an individual airport basis due to the variations in predictor importance
6) Low training and prediction times indicate that this process is not computationally expensive and can be replicated with additional airports and over a wider range of dates to be deployed for real-time use by stakeholder

V. Conclusion and Future Work

Ground Stops are Traffic Management Initiatives that are implemented to control air traffic volume to specific airports where the projected traffic demand is expected to exceed the airports' acceptance rate over a short period of time due to conditions such as inclement weather, volume constraints, closed runways, etc. Ground Stops can be

considered to be the strictest Traffic Management Initiative as all flights to the affected airport are grounded. However, limited work has been carried out on Ground Stops because of the limited amount of time that traffic management personnel often have between planning and implementing Ground Stops and external factors that influence decisions of traffic management personnel. This research thus focuses on predicting weather-related Ground Stops at Newark Liberty International (EWR) and LaGuardia (LGA) airports, with the secondary goal of gaining insights into factors that influence their occurrence. The fusion of the Traffic Flow Management System (TFMS) and Automated Surface Observing Systems (ASOS) datasets, and the benchmarking of Machine Learning algorithms identified the Boosting Ensemble algorithm as the best suited algorithm for predicting the occurrence of Ground Stops at EWR and LGA. The SMOTE algorithm was also leveraged to determine that imbalanced datasets lead to poorly performing prediction models, compared to balanced datasets. Analysis of predictor importance also revealed that different factors influenced the occurrence of Ground Stops at EWR and LGA. Future work will focus on determining if a single or airport-type prediction model can be developed to predict and analyze the occurrence of Ground Stops at multiple airports. Future work will also focus on exploring the differences in the performances of the prediction models and the ranking of predictor importance, with the EWR and LGA datasets.

Acknowledgments

The authors would like to acknowledge the contribution of Chana Kim at Georgia Institute of Technology for her contribution to this work. The views and findings expressed in this document are those of the authors only, and do not represent those of the FAA.

References

[1] Xiong, J., "Revealed Preference of Airlines' Behavior under Air Traffic Management Initiative," Ph.D. thesis, University of California, Berkeley, 2010. URL https://www.semanticscholar.org/paper/Revealed-preference-of-airlines%27-behavior-under-air-Xiong/d905cb723ab247cee8dd7504745f01b8dd2cd926.

[2] Mangortey, E., Gilleron, J., Dard, G., Pinon-Fischer, O., and Mavris, D., "Development of a Data Fusion Framework to support the Analysis of Aviation Big Data," *AIAA Scitech 2019 Forum, AIAA SciTech Forum, (AIAA 2019-1538)*, 2019. URL https://doi.org/10.2514/6.2019-1538.

[3] Mangortey, E., Pinon, O., Puranik, T., and Mavris, D., "Predicting The Occurrence of Weather And Volume Related Ground Delay Programs," *AIAA AVIATION Forum*, 2019. URL https://doi.org/10.2514/6.2019-3188.

[4] Dard, G., Mangortey, E., Pinon, O., and Mavris, D., "Application Of Data Fusion And Machine Learning To The Analysis Of The Relevance Of Recommended Flight Reroutes," *AIAA AVIATION Forum*, 2019. URL https://doi.org/10.2514/6.2019-3189.

[5] Bureau of Transportation Statistics, "Causes of National Aviation System Delays National (May, 2018)," , 2018. https://www.transtats.bts.gov/OT_Delay/ot_delaycause1.asp?type=5&pn=1.

[6] Robyn, D., *Reforming the air traffic control system to promote efficiency and reduce delays*, The Brattle Group, 2007. URL https://brattlefiles.blob.core.windows.net/files/5677_reforming_the_air_traffic_control_system_to_promote_efficiency_and_reduce_delays.pdf.

[7] Federal Aviation Administration, "Expected Departure Clearance Times," , 2019. http://aspmhelp.faa.gov/index.php/Expect_Departure_Clearance_Times_(EDCT).

[8] Federal Aviation Administration, "OPSNET 45," , 2009. URL https://aspmhelp.faa.gov/index.php/OPSNET_45.

[9] Federal Aviation Administration, "OPSNET : Delays : EDCT/GS/TMI By Cause Report," , 2019. URL https://aspm.faa.gov/opsnet/sys/opsnet-server-x.asp.

[10] Wang, Y., "Analysis and prediction of weather impacted ground stop operations," *2014 IEEE/AIAA 33rd Digital Avionics Systems Conference (DASC)*, 2014, pp. 7A2–1–7A2–14. doi:10.1109/DASC.2014.6979510.

[11] Ball, M., and Guglielmo, L., "Ground Delay Programs: Optimizing over the Included Flight Set Based on Distance," *14th AIAA Aviation Technology, Integration, and Operations Conference*, 2014. URL https://doi.org/10.2514/atcq.12.1.1.

[12] Mukherjee, A., Grabbe, S., and Sridhar, B., "Predicting Ground Delay Program At An Airport Based On Meteorological Conditions," *Air Traffic Control Quarterly, vol. 12, no. 1*, 2004. URL https://www.aviationsystemsdivision.arc.nasa.gov/publications/2014/AIAA-2014-2713.pdf.

[13] Hansen, M., Mukherjee, A., and Grabbe, S., "Ground Delay Program Planning under Uncertainty in Airport Capacity," *Transportation Planning and Technology, vol. 35, no. 6*, 2012. URL https://www.semanticscholar.org/paper/Ground-delay-program-planning-under-uncertainty-in-Mukherjee-Hansen/218b536c384afb6b487b2b5b412641bef617aeb0.

[14] Smith, D., and Lance, S., "Decision Support Tool for Predicting Aircraft Arrival Rates from Weather Forecasts," *2008 Integrated Communications, Navigation and Surveillance Conference*, 2008. URL https://doi.org/10.1109/ICNSURV.2008.4559186.

[15] Nagle, G., Elliott, M., and Clarke, J.-P., *Hypothetical Redesign of the New York Metro Airspace (New Vehicle NRA)*, 2009. doi:10.2514/6.2009-7069, URL https://arc.aiaa.org/doi/abs/10.2514/6.2009-7069.

[16] National Business Aviation Association, "New York Metro Airspace," , 2019. URL https://nbaa.org/aircraft-operations/airspace/atc-issues-procedures/new-york-metro-airspace/.

[17] Federal Aviation Administration, *JAVA MESSAGING SERVICE DESCRIPTION DOCUMENT Traffic Flow Management Data Service (TFMData) Vol. 2.0.5*, Federal Aviation Administration, 2016.

[18] Mangortey, E., "Predicting The Occurrence OF Ground Delay Programs And Their Impact On Airport And Flight Operations," Ph.D. thesis, Georgia Institute of Technology, May 2019. URL http://hdl.handle.net/1853/61288.

[19] National Weather Service, "Automated Surface Observing Systems," , 2019. https://www.weather.gov/asos/asostech.

[20] Guttman, Nathaniel and Baker, Bruce, "Exploratory Analysis of the Difference between Temperature Observations Recorded by ASOS and Conventional Methods," *Bulletin of American Meteorological Society*, 1996. URL https://doi.org/10.1175/1520-0477%281996%29077%3C2865%3Aeaotdb%3E2.0.co%3B2.

[21] Iowa State University, "ASOS-AWOS-METAR Data Download," , 2019. https://mesonet.agron.iastate.edu/request/download.phtml.

[22] Lepori, Hubert, "Introduction to FIXM." , 2017. URL https://www.icao.int/MID/Documents/2017/SWIMInterregional/8.2IntroductiontoFIXM.pdf.

[23] Python.org, "Welcome to Python.org." , 2018. URL www.python.org/.

[24] Python Software Foundation, "19.7. Xml.etree.ElementTree - The ElementTree XML API," , 2018. URL docs.python.org/2/library/xml.etree.elementtree.html.

[25] Iowa State University, "ASOS Network," , 2019. URL https://mesonet.agron.iastate.edu/request/download.phtml?network=NJ_ASOS.

[26] Klein, L., *Sensor and Data Fusion: A Tool for Information Assessment and Decision Making*, Press Monographs, Society of Photo Optical, 2004. URL https://books.google.com/books?id=-782bo4u_ogC.

[27] Lantz, Brett, *Machine Learning with R: Discover How to Build Machine Learning Algorithms, Prepare Data, and Dig Deep into Data Prediction Techniques with R*, Packt Publishing, 2015. URL https://books.google.com/books?id=ZaJNCgAAQBAJ.

[28] Mangortey, E., Monteiro, D., Ackley, J., Gao, Z., Puranik, T., Kirby, M., Pinon, O., and Mavris, D., "Application of Machine Learning Techniques to Parameter Selection for Flight Risk Identification," *AIAA Science and Technology Forum (AIAA Scitech)*, 2020.

[29] Mangortey, E., Bleu-Laine, M., Puranik, T., Pinon, O., and Mavris, D., "Application of Machine Learning to the Analysis and Prediction of the Coincidence of Ground Delay Programs and Ground Stops," *AIAA Science and Technology Forum (AIAA Scitech)*, 2020.

[30] Mangortey, E., Puranik, T., Pinon, O., and Mavris, D., "Classification, Analysis, and Prediction of the Daily Operations of Airports Using Machine Learning," *AIAA Science and Technology Forum (AIAA Scitech)*, 2020.

[31] Sheridan, K., Puranik, T., Mangortey, E., Kirby, M., Pinon, O., and Mavris, D., "An Application of DBSCAN Clustering For Flight Anomaly Detection During The Approach Phase," *AIAA Science and Technology Forum (AIAA Scitech)*, 2020.

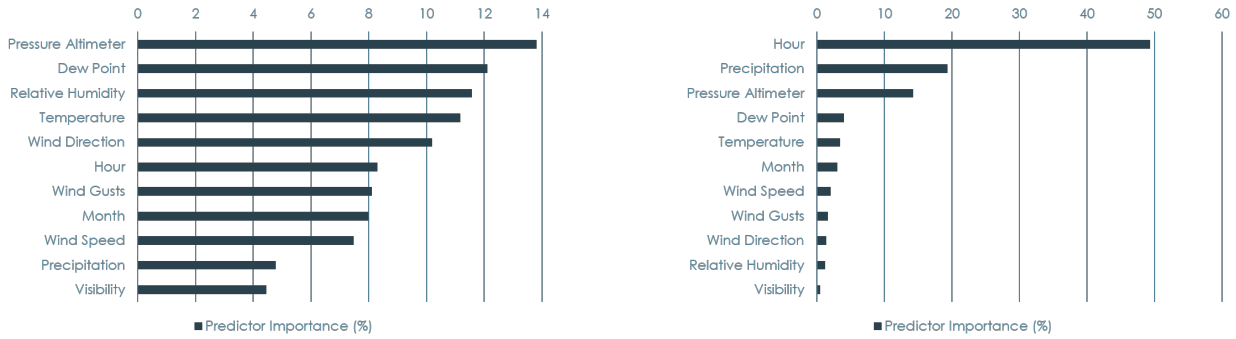
[32] Wang, J., Xu, M., Wang, H., and Zhang, J., "Classification of Imbalanced Data by Using the SMOTE Algorithm and Locally Linear Embedding," *2006 8th international Conference on Signal Processing*, Vol. 3, 2006. doi:10.1109/ICOSP.2006.345752.

[33] Bhagat, R. C., and Patil, S. S., "Enhanced SMOTE algorithm for classification of imbalanced big-data using Random Forest," *2015 IEEE International Advance Computing Conference (IACC)*, 2015, pp. 403–408. doi:10.1109/IADCC.2015.7154739.

[34] Brodersen, K. H., Ong, C. S., Stephan, K. E., and Buhmann, J. M., "The Balanced Accuracy and Its Posterior Distribution," *2010 20th International Conference on Pattern Recognition*, 2010, pp. 3121–3124. doi:10.1109/ICPR.2010.764.

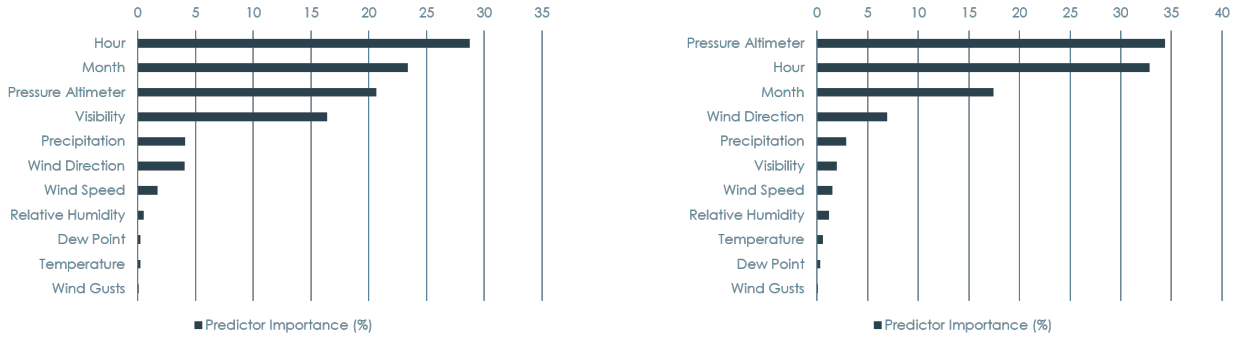
[35] García V. and Mollineda R.A. and Sánchez J.S., *Index of Balanced Accuracy: A Performance Measure for Skewed Class Distributions*, Springer, Berlin, Heidelberg, 2009. URL `https://doi.org/10.1007/978-3-642-02172-5_57`.

Appendix



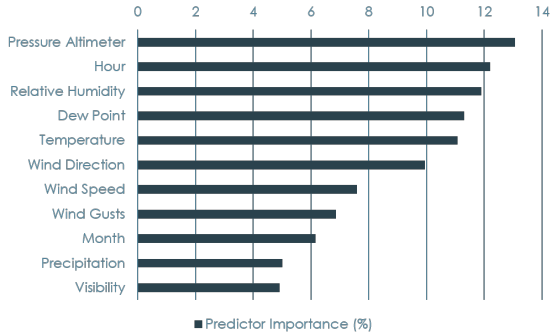
(a) Imbalanced EWR dataset **(b) Balanced EWR dataset**

Fig. 13 EWR predictor importance with Bagging Ensemble algorithm

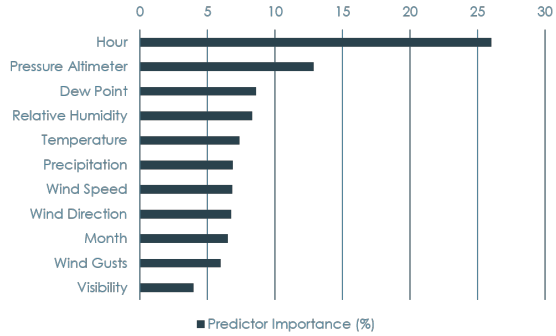


(a) Imbalanced LGA dataset **(b) Balanced LGA dataset**

Fig. 14 LGA predictor importance with Bagging Ensemble algorithm

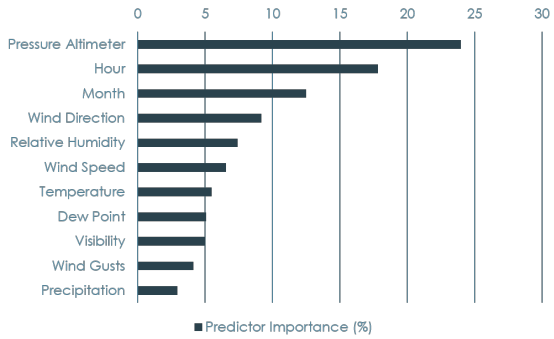


(a) Imbalanced EWR dataset

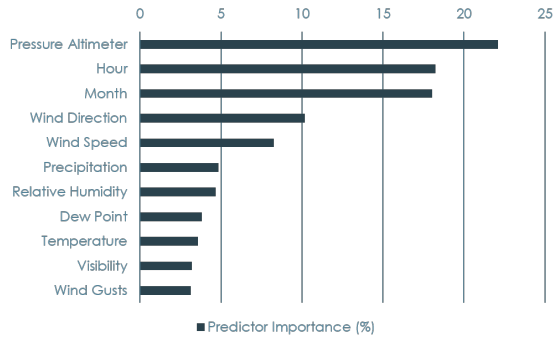


(b) Balanced EWR dataset

Fig. 15 EWR predictor importance with Random Forests algorithm



(a) Imbalanced LGA dataset



(b) Balanced LGA dataset

Fig. 16 LGA predictor importance with Random Forests algorithm