# Exploring the Predictive Power of Machine Learning for Identifying Flight Cancellations in the U.S. Domestic Airline Industry

Paul Seward
*Mathematical and Computational Sciences*
*University of Prince Edward Island*
Charlottetown, Canada
pseward@upei.ca

***Abstract*—This paper investigates the predictability of flight cancellations using classification models. Two models, logistic regression and random forests, were trained on a large dataset of American domestic flights in 2008. The results demonstrate that neither model performed well in predicting cancellations, with the logistic regression model slightly outperforming the random forests model. Important features were identified, such as DepDelay, ActualElapsedTime, TaxiOut, and ArrDelay which could contribute to a cancelled flight. The heavily biased dataset composition towards non-cancelled flights was identified as a significant limitation to this study. Future research should consider oversampling or undersampling techniques and incorporating real-time data such as weather forecasts and airport maintenance data.**

***Index Terms*—classification, dimension reduction, feature selection, flight cancellation**

## I. Introduction

In an increasingly connected and mobile world, air travel is a vital contribution to the world's efficiency. Individuals and businesses depend on a seamless experience. Delays can cause significant financial consequences due the interconnected nature of airport networks. We have chosen this data set to provide valuable insights to the contributing reasons for flight disruptions, uncover areas of optimization for airlines, and reveal information to enhance a customer's maximum chance of a smooth flight experience.

The objective of this study is to develop accurate classification models that can predict whether a flight will be cancelled or not, based on various features such as the security delay, specific airports or airlines, and distance of the flight. Accurately predicting cancellations could allow airlines to proactively manage their schedules and minimize the impact of cancellations. By developing accurate classification models, we can improve the efficiency and reliability of air travel, benefitting both airlines and passengers.

This paper begins with a background on the dataset and an exploration on the effects of a large data size. Different machine learning techniques are then explained leading to the presentation of our two classification models, logistic regression and random forests. We will analyze the performance of these models, including evaluating their accuracy and examining which features are most important in predicting cancellations. The paper then concludes with a discussion of the limitations of our study and suggest areas for future work.

## II. Background

### *A. Dataset*

While performing initial explorations and preprocessing, we encountered certain trade offs of size and performance. The size and depth of information this dataset provides was an attractive quality at first. This led to the decision to select this dataset for further data exploration. However, we quickly discovered that exact quality was also the biggest hindrance from extracting any useful information as the model became too computationally expensive to evaluate.

This dataset is comprised of 1,936,758 flights in the US in 2008. The data comes from the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics(BTS). There are 29 feature variables, and while most are self explanatory, the variables pertaining to a delay type require further elaboration. There are six different delay types, all recorded in minutes delayed. The feature explanation [1]. is as follows:

- CarrierDelay: Carrier delay is within the control of the air carrier. Examples of occurrences that may determine carrier delay are: aircraft cleaning, aircraft damage, awaiting the arrival of connecting passengers or crew, baggage, bird strike, cargo loading, catering, computer, outage-carrier equipment, crew legality (pilot or attendant rest), damage by hazardous goods, engineering inspection, fueling, handling disabled passengers, late crew, lavatory servicing, maintenance, oversales, potable water servicing, removal of unruly passenger, slow boarding or seating, stowing carry-on baggage, weight and balance delays.
- WeatherDelay: Weather delay is caused by extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival.
- NASDelay: Delay that is within the control of the National Airspace System (NAS) may include: non-extreme

weather conditions, airport operations, heavy traffic volume, air traffic control, etc.

- SecurityDelay: Security delay is caused by 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.
- LateAircraftDelay: Arrival delay at an airport due to the late arrival of the same aircraft at a previous airport. The ripple effect of an earlier delay at downstream airports is referred to as delay propagation.
- ArrDelay: A flight is counted as ”on time” if it operated less than 15 minutes later the scheduled time shown in the carriers' Computerized Reservations Systems (CRS). Early arrivals show negative numbers, in minutes

One-hot encoding was applied to letter-valued feature variables, Origin and Unique Carrier. The variable we tried to predict was, Cancelled, which has only 0 and 1 values, representing cancelled and non-cancelled respectively. The dataset is heavily biased towards non-cancelled flights. There are only 633 cancelled flights compared to 1,936,125 non-cancelled flights.

Due to the large data size and multiple categorical variables with numerous categories, we found we needed to reduce the data size by sampling down to one tenth of original size to perform several machine learning techniques in our exploratory analysis stage of this study. The constraints of a platform like Google Colab prevented further exploration in polynomial regression to predict delay times. Moreover, we experienced difficulties from large data size in creating a logistic regression classifying model. The model failed to converge when training this model originally. It was necessary to sample the dataset and increase the maximum iterations parameter by a factor of 10, from the default to 5000, for our model to converge.

In the exploratory phase of understanding this dataset, we also applied univariate and multivariate linear regression to predict the total delay time. However, these models performed poorly and the polynomial regression was inconclusive as stated above. The learning curve for linear regression is included in this study to showcase its results.

### B. Machine Learning Techniques

In our research, we applied Principle Component Analysis (PCA) to reduce the dimensionality of our dataset. Specifically, we used PCA to transform our data into a lower-dimensional space while preserving as much of the original information as possible. PCA works by finding new variables, the principal components, which are linear combinations of those in the original dataset, that successively maximize variance and that are uncorrelated with each other [2]. These principal components can then be used to represent the data in a lower-dimensional space. In our study, we discovered using PCA led to a relatively small reduction in the number of features needed to accurately classify our data

The first classification technique we deployed on this data is logistic regression. Logistic regression works by applying a nonlinear logistic function (1) to the input data with weights, $\theta^T x$ which maps the input features to a probability value between 0 and 1 as seen in the hypothesis function (2) [3]. The logistic function is used to model the relationship between the input features and the target variable, the likelihood of a flight being canceled.

$$g(z) = \frac{1}{1+e^{-z}} \tag{1}$$

$$h_\theta(x) = \frac{1}{1+e^{-\theta^T x}} \tag{2}$$

Since this hypothesis function is not linear, the loss function for logistic regression incorporates logarithmic terms (3). As a result, a subsequent gradient descent algorithm applied on the entire cost function J($\theta$) (4), produces a convex shape that can be optimized .

$$\mathcal{L}(h_\theta(x), y) = -ylog(h_\theta(x)) - (1-y)log(1-h_\theta(x)) \tag{3}$$

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} -y^{(i)}log(h_\theta(x^{(i)})) - (1-y^{(i)})log(1-h_\theta(x^{(i)}))\right] \tag{4}$$

Following this, we implemented a random forests classifying model. The main idea behind the random forest algorithm is to create an ensemble of decision trees, where each decision tree is trained on a random subset of the features and a random subset of the data. The output of the ensemble is then obtained by aggregating the outputs of individual trees.

The advantage of using an multiple decision trees is that it reduces the risk of overfitting and improves the generalization performance of the model. Injecting randomness into the model not only reduces the correlation and variance, but also makes it more robust to outliers. [4]

The Random Forest algorithm also produces a measurement pertaining to the importance of each feature in the dataset. This information can be used to gain insights into variable importance, which can be useful for feature selection and data analysis.

## III. Data Analysis

### A. Feature Selection

During the feature selection process we prioritized removing logically irrelevant features and duplicates in efforts to improve the accuracy and efficiency of the model by reducing the complexity of the dataset. Another focus in this process is to weigh the importance of certain categorical features against the cost of one hot encoding. Without careful consideration, this dataset will become too large for the computing power of platforms such as Google Colab. Therefore, this study aims to identify the most important categorical variables that can be used to create an efficient and accurate model while keeping the dataset size manageable.

Foremost, logically irrelevant features were dropped such as 'TailNum' since the plane's number would have no logical reason to delay a flight. Afterwards, any variables with logical duplicates which represented similar values but measured with

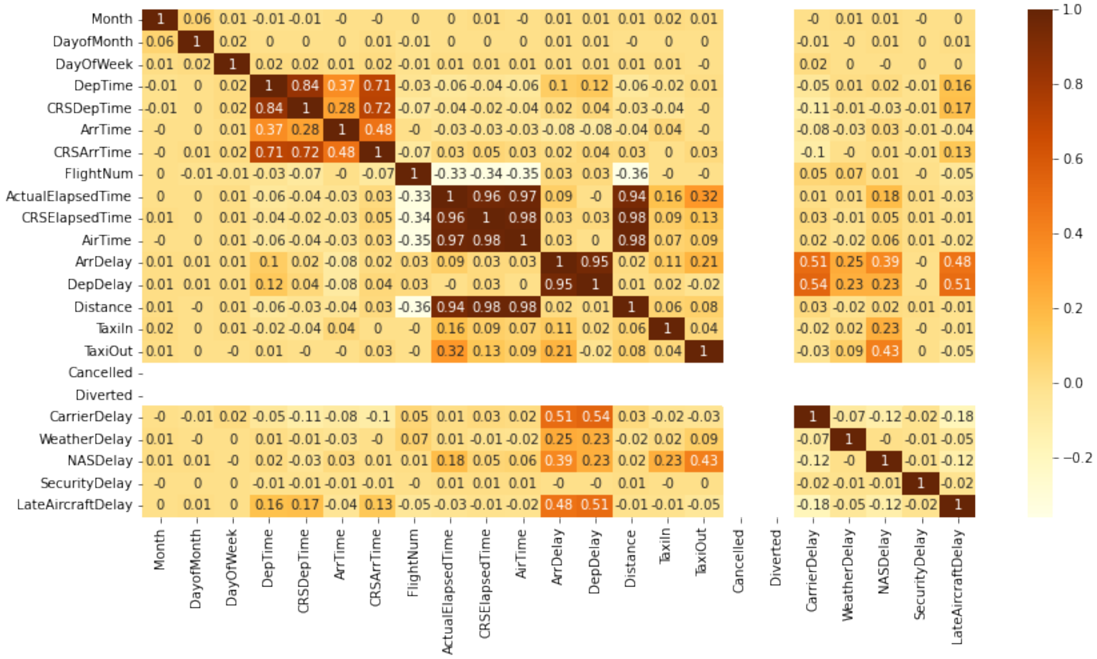

Fig. 1. Correlation Matrix

different formalities were also removed. An example of this is AirTime, CRSElapsedTime, ActualElapsedTime where we kept the latter feature and removed the former duplicates. The process of removing the logical counterparts can be confirmed in the correlation matrix in Figure 1. Amongst the three features, we can see a near perfect correlation implying they are measuring the same idea.

The categorical variable, Destination, is removed seeing that the trade-off for adding another three-hundred one-hot encoded variables will be too computationally expensive. We prioritized including the Origin categorical variable over Destination since the information derived from airport maintenance operations is more logically connected to the airport from where an airplane departs. Furthermore, the information provided by the destination airport variable is not necessarily lost since the distance between airports remains in the model.

## B. Learning Curve

The learning curve of a model may illustrate if the model is underfit to the data with high bias and high variance or overfit with low bias and high variance in the validation error. In Fig. 2, the learning curve does not show over-fitting since it does converge towards the training error. Although there appears to be high bias, I would not say this model is underfit to the data as more complex models do not lower the training error significantly.

The learning curves of logistic regression and random forests are not visually interpretable because metrics such as accuracy and precision remain around 100%. However, that error measurement is misleading in this specific context.

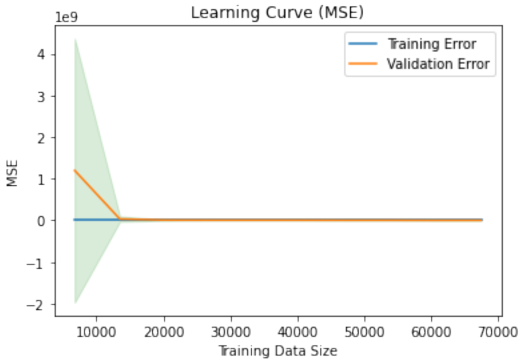

Fig. 2. Multivariate Regression Learning Curve

## C. Dimension Reduction

Principle Component Analysis is applied in this study to reduce the dimensionality of the data with a target of maintaining 90% of the total variance. The results of PCA demonstrate that no large benchmark variance levels are explained by the first several principle components. Table I illustrates that no specific principle component explains significantly more variance than other components seen lower in the table. We found the first 10 components explain just less than 6% of the total variance.

We also discovered to achieve 90% of the variance it will require using 269 principal components. These findings indicate there is not a small amount of features that maintain a significant percentage of variance, but rather, many features collectively explain the variance. In spite of that, the threshold of 90% variance explained was still deemed appropriate for the analysis.

TABLE I
PRINCIPLE COMPONENTS ANALYSIS VARIANCE

| **Number of Principle Components** | ***Cumulative Variance Explained by Principle Components*** |
|---|---|
| 1 | 0.00869284 |
| 2 | 0.01581579 |
| 3 | 0.02226346 |
| 4 | 0.02785745 |
| 5 | 0.03328908 |
| 6 | 0.03870814 |
| 7 | 0.0439704 |
| 8 | 0.0491875 |
| 9 | 0.05440036 |
| 10 | 0.05959877 |
| . . . | . . . |
| 267 | 0.89574934 |
| 268 | 0.89837591 |
| 269 | 0.90099462 |

The cumulative sum of variance by principle components

### D. Classification Results

I) Logistic Regression Classification

TABLE II
LOGISTIC REGRESSION CLASSIFIER

| | **Confusion Matrix** | |
|---|---|---|
| | ***Condition Positive*** | ***Condition Negative*** |
| Predicted Positive | 26 | 1 |
| Predicted Negative | 128 | 484035 |

| Metric | Value |
|---|---|
| Recall | 0.169 |
| Precision | 0.963 |
| Accuracy | 1.000 |

Results of Logistic Regression.

Table II displays the summary statistics of a logistic regression classifying model. The recall score of the model is low, at 0.169, indicating only 16.9% of cancelled flights (positive cases) were correctly identified. Since True Positive Rate (TPR) is equal to 1 - False Negative Rate (FNR), the FNR is very large meaning the model missed on most of the cancelled flights.

The precision score is very high at 0.963 or 96.3%. Although the model predicted 27 flights as cancelled, the model performed very well noting that 26 out of 27 predictions were correct. The accuracy score of the model is 1.0, indicating the model correctly classified all the test cases (with rounding). Again however, this is misleading because the dataset is heavily imbalanced towards non-cancelled flights

II) Random Forests Classification

Table III displays the summary statistics of a random forests classifying model. The recall score of the model is very low, at 0.026, indicating only 2.6% of cancelled flights were correctly identified. Again, this implies the FNR is very large meaning the model missed on most of the cancelled flights. The precision score is very high at 1.0, although it is misleading since the model only predicted a total of four predicted cancelled flights even if it was correct about each prediction.

TABLE III
RANDOM FORESTS CLASSIFIER

| | **Confusion Matrix** | |
|---|---|---|
| | ***Condition Positive*** | ***Condition Negative*** |
| Predicted Positive | 4 | 0 |
| Predicted Negative | 150 | 484036 |

| Metric | Value |
|---|---|
| Recall | 0.026 |
| Precision | 1.000 |
| Accuracy | 1.000 |

Results of Logistic Regression.

The accuracy score of the model is 1.0 (rounded), indicating that the model correctly classified all the test cases. Again however, this is misleading because the dataset is heavily imbalanced towards non-cancelled flights. It appears the best summary statistic of this model is recall. Since the model only predicted 4 out of 154 cancelled flights, it demonstrates poor performance in predicting cancelled flights.

### E. Feature Importance

Table IV highlights the most important features to their respective model. We noticed that DepDelay is the most important feature to predicting if a flight is cancelled or not given it is the second most important feature to both models. Other noteworthy features with high importance to both models are ActualElapsedTime and TaxiOut. While ArrDelay is mentioned as most important by a random forests model, the logistic regression model does not include this feature in a top ten list. These measure were extracted from the models without dimension reduction applied. Here the results are much more interpretable than attempting to analyze PCA principle components.

Top 10 Feature Importance Lists from Random Forests and Logistic Regression

The measure of feature importance for Random Forests is Mean Decrease Impurity (MDI). MDI scores are calculated by measuring how much each feature reduces the impurity of the decision tree or forest. Features that reduce impurity the most are considered more important because they provide the most information gained by the tree and would therefore be placed higher in the tree hierarchy for decision making [5].

Feature importance in Logistic Regression Classifiers can be measured by their coefficient magnitude. This importance score is calculated by simply obtaining each feature's coefficient. The coefficient magnitude is representative of how important a feature is to its model because small variations in the variable carry more weight in adjusting the prediction output. Therefore, features with the largest coefficient magnitude are identified as most important.

## IV. CONCLUSIONS

This paper presented different classification techniques to examine the predictability of cancellations in a large dataset

TABLE IV
Feature Importance

| Rank | Feature | Importance |
|---|---|---|
| | **Random Forests** | (MDI) |
| 1 | ArrDelay | 0.124 |
| 2 | DepDelay | 0.089 |
| 3 | ArrTime | 0.082 |
| 4 | ActualElapsedTime | 0.072 |
| 5 | TaxiOut | 0.060 |
| 6 | CRSArrTime | 0.060 |
| 7 | Distance | 0.057 |
| 8 | LateAircraftDelay | 0.057 |
| 9 | CarrierDelay | 0.056 |
| 10 | CRSDepTime | 0.055 |
| | **Logistic Regression** | (coefficients) |
| 1 | Month | 3.509 |
| 2 | DepDelay | 2.920 |
| 3 | ActualElapsedTime | 0.995 |
| 4 | TaxiOut | 0.592 |
| 5 | UniqueCarrier_MQ | 0.299 |
| 6 | UniqueCarrier_OO | 0.284 |
| 7 | TaxiIn | 0.216 |
| 8 | UniqueCarrier_YV | 0.182 |
| 9 | UniqueCarrier_9E | 0.181 |
| 10 | CRSDepTime | 0.133 |

of American domestic flights in 2008. The poor performance of both models in predicting cancellations, as indicated by the recall statistic, is highlighted by summary statistic tables III and IV. The logistic regression model performed slightly better than the random forests model in areas of recall and False Negative Rate. This demonstrates the logistic model is preferable when predicting whether a flight will be cancelled.

Although the random forests model displays a higher precision score it only predicts 4 flights as being cancelled when the logistic model predicts 27 flights as cancelled but had one incorrect prediction. Since the goal of this classification problem is to identify flights as being cancelled, it would be important to identify more than 4 out of 154 flights as likely to be cancelled. This suggests the optimal classification model is the logistic regression model, however neither model produce feasible results.

The foremost important features derived from each classifier are DepDelay, ActualElapsedTime, and TaxiOut. The feature importance scores imply no single feature stands out as most critical to classification. Nonetheless, these results can produce inferences for which factors contribute most to a cancelled flight. DepDelay, the total delay feature, makes logical sense as the most important feature because if plane is significantly delayed, the underlying reasons would also contribute towards an entire cancellation of the flight.

ActualElapsedTime, pertaining to the time duration of the flight, is labelled as an important feature which suggests longer flight lengths also affect the cancellation of a flight. TaxiOut relates to the time it takes for the plane to be escorted to its runway for takeoff. This indicates once a plane is boarded, the longer it takes to take off, the higher chance it is to be cancelled. ArrDelay pertains to the arrival delay at an airport due to the late arrival of the same aircraft at a previous airport. This is another important feature for cancelled flights and is noteworthy because one delayed flight can initiate several other disruptions due to the interconnected nature of airport traffic. The ripple effect of an earlier delay at downstream airports is referred to as delay propagation [6].

Neither of our classification models presented evidence that a flight can be confidently predicted as cancelled. Despite the large size of the dataset with 29 initial features, each model suffered from high bias. The amount of data did not seem to be a contributing reason because as models became more complex, such as random forests, the error did not decrease. A deep neural network could be applied and analyzed whether a significantly more complex model could help improve accuracy. Other potential reasons that explain the current models' poor performance are a heavily biased dataset composition and an absence of crucial information to each flight such as weather and airport maintenance data.

The largest reoccurring problem to this study is the heavily biased dataset composition towards non-cancelled flights. Future research into this dataset should include an exploratory analysis into whether using bootstrapping sampling techniques with oversampling or undersampling could help improve the model's predictions.

Future work to this study should also investigate the application of neural networks. A standard fully connect neural network will be the initial step. Following this, will be attempting to access more real time data such as weather forecasts and any publicly accessible airport maintenance data. The inclusion of real time data would provide access to opportunities of applying a time-series approach using a recurrent neural network. This also will have the potential to capture the effects of delay propagation.

## References

[1] L. Findsen et al, ”Airline 2008 Dataset.” Department of Statistics, Purdue University, Airline2008 Dataset Variable Definition, 2009.

[2] J. T. Jolliffe and J. Cadima, ”Principal component analysis: a review and recent developments,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, pp. 1-16.

[3] J. M. Hilbe, “Logistic Regression,” *International Encyclopedia of Statistical Science*, pp. 755-758, 2011.

[4] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[5] ChatGPT. (2023). Explain quantitative feature importance scores and how each are calculated. [Online conversation]. OpenAI. Available: 2023, March.

[6] J.T. Wong and S.C. Tsai, “ A survival model for flight delay propagation”, *Journal of Air Transport Management*, vol. 23, pp. 5-11, 2012.