Predicting Flight Delays & Cancellations

IST718–Big Data Analytics

Group 1: Abhiraj Singh, Lakshya Gupta, Surjit Singh, Sankalp Singh

Instructor: Prof. Williard Williamson

Contents

[1. Abstract 1](#_Toc39270028)

[2. Introduction 2](#_Toc39270029)

[2.1 Background 2](#_Toc39270030)

[2.2 Objective 2](#_Toc39270031)

[2.3 Data Set 2](#_Toc39270032)

[2.4 Business Problem 3](#_Toc39270033)

[3. Methodology - CRISP DM 4](#_Toc39270034)

[3.1 Data Preprocessing 4](#_Toc39270035)

[3.2 Data Wrangling & Feature Engineering 5](#_Toc39270036)

[3.3 Exploratory Data Analysis 6](#_Toc39270037)

[Correlation Matrix 6](#_Toc39270038)

[3.4 Predictive Modeling 13](#_Toc39270039)

[4. Results 16](#_Toc39270040)

[5.Conclusion 21](#_Toc39270041)

[6. Appendix 23](#_Toc39270042)

1. **Abstract:**

Commercial airlines are the backbone of USA transportation system, bringing significant socio-economic utility by enabling cheaper and easier long-distance travel. Growth in the aviation industry has resulted in air traffic congestion causing flight delays and cancellations. A lot of times, a modern passenger faces inconvenience by aircraft delays and cancellations. These delays and cancellations lead to negative impacts, mainly economical for commuters, airline industries, and airport authorities. Furthermore, in the domain of sustainability, it can even cause environmental harm by the rise in fuel consumption and gas emissions. These factors indicate how necessary and relevant it is to predict the delays.

The goal of this project is to predict airline delays, cancellations and analyze various factors causing inconvenience. A better understanding of these hindering and disrupting factors could improve airline scheduling and significantly reduce delays and cancellations. We aim to carry out descriptive analysis, predictive analysis, which encompasses a range of statistical techniques from supervised machine learning and data mining that studies current and historical data to make predictions or just analyze the future delays and cancellations. We have used binary and multiclass classification algorithms in order to predict the delays and cancellations. Using machine learning techniques, we were able to determine the various factors that have contributed most for cancellations and delays to occur. This report will provide the deep insights which can be helpful for the aviation industry in order to take actionable measures to avoid cancellations and delays.

# **2. Introduction**

## **2.1 Background**

The primary focus of this analysis is to predict the major factors leading to flight delays and cancellations. Major airlines suffer from the issue of flight delays due to increasing air traffic congestion, weather problems and several other reasons. These delays and cancellations result in a bad reputation for the air carrier and moreover leads to negative impact for airport authorities and airline industries.

## **2.2 Objective**

The main aim of this project is to predict the crucial factors which results in flight delays and cancellations. Flight delays and cancellations usually causes a lot of problems for the customers and mostly for the daily passengers travelling via flights. An analysis to determine the major factors leading to flight delays and cancellations will result in improving the performance of large air carriers and will lead to a robust aviation industry.

## **2.3 Data Set**

The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. Summary information on the number of on-time, delayed, canceled, and diverted flights is published in DOT's monthly Air Travel Consumer Report. The Dataset contains more than 5.5 million observations where each observation represents the information of a flight. There are 31 features in the dataset. We have used this dataset to find out the factors which lead to flight cancellations and delays. This dataset contains the following predictors:

|  |
| --- |
| YEAR - Year of the Flight Trip |
| MONTH - Month of the Flight Trip |
| DAY - Day of the Flight Trip |
| DAY\_OF\_WEEK - Day of week of the Flight Trip |
| AIRLINE - Airline Identifier |
| FLIGHT\_NUMBER - Flight Identifier |
| TAIL\_NUMBER - Aircraft Identifier |
| ORIGIN\_AIRPORT - Starting Airport |
| DESTINATION\_AIRPORT - Destination Airport |
| SCHEDULED\_DEPARTURE - Planned Departure Time |
| DEPARTURE\_TIME |
| DEPARTURE\_DELAY - Total Delay on Departure |
| TAXI\_OUT - The time duration elapsed between departure from the origin airport gate and wheels off |
| WHEELS\_OFF - The time point that the aircraft's wheels leave the ground |
| SCHEDULED\_TIME - Planned time amount needed for the flight trip |
| ELAPSED\_TIME - AIR\_TIME+TAXI\_IN+TAXI\_OUT |
| AIR\_TIME - The time duration between wheels\_off and wheels\_on time |
| DISTANCE - Distance between two airports |
| WHEELS\_ON - The time point that the aircraft's wheels touch on the ground |
| TAXI\_IN - The time duration elapsed between wheels-on and gate arrival at the destination airport |
| SCHEDULED\_ARRIVAL - Planned arrival time |
| ARRIVAL\_TIME |
| ARRIVAL\_DELAY |
| DIVERTED - Aircraft landed on airport that out of schedule |
| CANCELLED - Flight Cancelled (1 = cancelled) |
| CANCELLATION\_REASON - Reason for Cancellation of flight: |
| AIR\_SYSTEM\_DELAY - Delay caused by air system |
| SECURITY\_DELAY - Delay caused by security |
| AIRLINE\_DELAY - Delay caused by the airline |
| LATE\_AIRCRAFT\_DELAY - Delay caused by aircraft |
| WEATHER\_DELAY - Delay caused by weather |

## **2.4 Business Problem**

The business problem we have tackled here is to determine the prime factors leading to flight delay and cancellations. Flights delays and cancellations have a lot of negative impact on the airline industries and airport authorities. Our analysis aims to find factors resulting in flight delays and cancellations which can further be worked upon by airline carriers to reduce their delays and cancellations. A significant reduction in flight cancellations and delays from our analysis will help the airline industry build upon their reputation. This will eventually help the airlines achieve an improved amount of customer satisfaction. The other major drawback of the flight delay and cancellations include the economic loss encountered by the airline company. Such losses can be avoided, and more profitable business can be achieved if the airlines try to minimize the loss caused by the various factors.

# **3. Methodology - CRISP DM**

We have followed the Crisp DM – Cross industry standard procedure for data mining methodology. This is the most widely used analytics model. This model has five major phases –

* Business Understanding
* Data Understanding
* Data Preparation
* Modeling
* Evaluation

.

A close up of a logo

Description automatically generated

* 1. **Data Preprocessing:**

Initially, we identified the columns which comprises of the Null and NA values, such columns must be tackled with a different approach for different scenario. The columns which had null values were TAIL\_NUMBER, DEPARTURE\_TIME, DEPARTURE\_DELAY, TAXI\_OUT, WHEELS\_OFF, ELAPSED\_TIME, AIR\_TIME, WHEELS\_ON, TAXI\_IN, ARRIVAL\_TIME, ARRIVAL\_DELAY,CANCELLATION\_REASON,AIR\_SYSTEM\_DELAY,SECURITY\_DELAY, AIRLINE\_DELAY, LATE\_AIRCRAFT\_DELAY, WEATHER\_DELAY.

Now we try to find the statistically significant and non-significant features, based on the evaluation we were able to remove the features which do not contribute towards achieving the goal. We removed the CANCELLATION\_REASON, AIR\_SYSTEM\_DELAY, SECURITY\_DELAY, AIRLINE\_DELAY, LATE\_AIRCRAFT\_DELAY, WEATHER\_DELAY columns as they comprised of 98% of null values. Later, we removed the columns which do not share any trend with the dependent variable. These features included TAIL\_NUMBER, TAXI\_OUT, WHEELS\_OFF, WHEELS\_ON, TAXI\_IN, SCHEDULED\_TIME.

We were able to determine features like DEPARTURE\_TIME which would make our model less generalized for test cases therefore we removed this feature. We also removed ELAPSED\_TIME, AIR\_TIME as these features were redundant with features like Distance feature which captured the total elapsed time and airtime

* 1. **Data Wrangling & Feature Engineering**

Now moving to the next phase and making the data ready for the machine learning models. For this task we performed One Hot encoding for the categorical features- AIRLINE, ORIGIN\_AIRPORT and DESTINATION\_AIRPORT. The distance feature was transformed into 3 buckets as long distance, medium distance and short distance. For classification we transformed the target variable for delay prediction as-

* Binary Classification – Flight\_delayed – 1(Yes – Delay>15mins) / 0(No – Delay<15mins)
* Multiclass Classification – Delay\_Bucket – 0(No Delay), 1(Delay>15mins), 2(Delay>30mins)

# **3.3 Exploratory Data Analysis**

## **Correlation Matrix**

Variables containing numerical data are used to create the correlation matrix. Correlation matrix shows that the target feature ‘Flight\_Delayed’ has a small positive correlation with ‘SCHEDULED\_DEPARTURE’.

A close up of a piece of paper

Description generated with high confidence

**Month vs departure delay**

Below plot shows the departure delay variance according to the moths of a year. Mid-year months have the greatest delays.

A picture containing screenshot

Description automatically generated

**Month vs Departure Delay with Flight Duration**

Below plot shows the departure delays variance with the months along with duration of flights.

A close up of a pencil

Description automatically generated

**Day vs Departure delay**

The plot shows the departure delays for all the days of a month. Mid-month days show a peak in the delays.

A picture containing implement, pencil

Description automatically generated

**Day of Week vs Departure Delay**

The below plot shows departure delays for the days of a week. Highest departures are for Sunday.

A picture containing drawing

Description automatically generated

**Day of week vs Departure delay with Flight Duration**

The below plot shows departure delays for days of a week along with the flight duration.

A picture containing implement, stationary, pencil

Description automatically generated

**Airline vs Departure delay**

Below plot shows the departure delays for all the 14 airlines present in the dataset.

A screenshot of a cell phone

Description automatically generated

**Flight Distance vs Departure delay**

Below plot shows the departure delays for different duration of flights. Longer duration flights are more prone to delays.

A screenshot of a cell phone

Description automatically generated

**Origin Airport vs Percentage Departure Delay**

Below is bubble plot that shows the percentage delays for different origin airports. Size of the bubble represents the number of flights for that airport.

A close up of a logo

Description generated with high confidence

**Months vs Percentage Flight Cancelled**

Below plot shows the percentage of flights cancelled for different months of a year. February has the highest percentage of flights cancelled.

A screenshot of a cell phone

Description generated with very high confidence

**Airline Count Plot:**

Below plot shows the count of flights for all the different airlines.

A screenshot of a cell phone

Description automatically generated

**Top 10 Origin Airport Count:**

Below plot shows the Top 10 origin airport with their respective counts of flight

A screenshot of a cell phone

Description automatically generated

**Top 10 Departure Airport Count:**

Below plot shows the Top 10 origin airport with their respective counts of flight

A screenshot of a cell phone

Description automatically generated

# **Predictive Modeling**

We have approached the delay prediction and cancellation prediction in 3 ways:

1. **Delay Prediction – Binary Classification Problem**
2. **Delay Prediction – Multiclass Classification Problem**
3. **Cancellation Prediction**

Now we have the data which is ready to process by the machine learning techniques. We create the train and test sets. The train data comprises of the 90% of the total data, remaining 10% we consider as the test data which will be used for the evaluation of the models. Since the trained dataset was unbalanced because the class with label “0” is 80% and “1” had 20% of the training data. To overcome this problem, we performed the under-sampling technique. This helped us achieve the balanced dataset which we will use to train the model. This will help the models to learn about both the classes adequately and thus increases the performance of the overall model. However, we did not perform the under-sampling for the delay prediction for the multiclass classification because we wanted to perform an experiment with the pre-processed and featured data only. Therefore, the under-sampling techniques was thus used only for the delay prediction (Binary classification) and the cancellation prediction (Binary Classification)

**MODEL USED**:

1. Logistic Regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labelled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labelled "1" is a linear combination of one or more independent variables ("predictors"). The independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labelled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labelling. The function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a *logit*, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the *probit* model. The defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter. For a binary dependent variable this generalizes the odds ratio. In a binary logistic regression model, the dependent variable has two levels (categorical). Outputs with more than two values are modelled by multinomial logistic regression and, if the multiple categories are ordered, by ordinal logistic regression (for example the proportional odds ordinal logistic model). The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cut-off value and classifying inputs with probability greater than the cut-off as one class, below the cut-off as the other. This is a common way to make a binary classifier.

1. Random Forest

Random Forest is an ensemble learning technique for classification, regression and other tasks that operate by constructing a collection of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. They use the “bagging” technique and random selection of features to create the collection of decision trees. Given a training set X = x1, ..., xn with responses Y = y1, ..., yn, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For b = 1, ..., B:

1. Sample, with replacement, n training examples from X, Y; call these Xb, Yb.
2. Train a classification or regression tree fb on Xb, Yb.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

or by taking the majority vote in the case of classification trees. This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. Similarly, we used the grid search and cross validation to achieve the best parameters for the model and performed cross-validation.

1. Gradient Boosted Tree

Gradient-Boosted Trees (GBTs) are ensembles of decision trees. GBTs iteratively train decision trees in order to minimize a loss function. The spark.ml implementation supports GBTs for binary classification and for regression, using both continuous and categorical features.

1. Linear Support Vector Machine

In machine learning, support-vector machines are supervised learning models with associated learning algorithms that analyses data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces. This model is computationally very expensive.

**Delay Prediction** (**Binary Classification Problem**)

For this task, we used the above explained model with the following hyperparameters:

1)**Logistic Regression** – We implemented grid search and received the following parameters, regParam=0.2, elasticNetParam=0.1, threshold=0.45, thresholds = [0.55, 0.45]

2)**GBT** – We implemented the following parameter to achieve the best generalized model, maxIter=5, maxBins=15, stepSize=0.08, maxDepth=4

3)**Random Forest** – Using the cross validation and grid search we achieved the following parameters, impurity = gini, maxBins=15, maxDepth=4, num\_trees = 150

4)**SVM** – Due to lack of computation resources we implemented manually MaxIter = 20, RegParam= 0.08

**Delay Prediction** (**Multiclass Problem**)

For this task, we used the above explained model with the following hyperparameters:

1)**Random Forest** – num\_trees = 20 and all the other parameters were default.

2)**Logistic Regression** - maxIter=10, regParam=0.3, elasticNetParam=0.8, family='multinomial'

**Cancellation Prediction (Binary classification**)

For this task, we used the above explained model with the following hyperparameters:

1) **Logistic** **Regression**- maxIter=100, regParam=default, elasticNetParam=default, tol=1e-06, fitIntercept=True, threshold=0.5, thresholds=None, probabilityCol='probability', rawPredictionCol='rawPrediction', standardization=True, aggregationDepth=2, family='auto'

2) **GBT**- MaxIter = 20, rest other parameter were implemented by default model

3) **Random** **Forest**- The parameters were kept to default.

# **4. Results**

We have tackled 3 different problems by analyzing and performing predictive modeling on our dataset.

**Delay Prediction – Binary Classification**

|  |  |  |  |
| --- | --- | --- | --- |
| **PREDICTION MODEL** | **AUC SCORE** | **BALANCED ACCURACY** | **F1 SCORE** |
| Logistic Regression | 63 | 56 | 61 |
| Random Forest | 65 | 81 | 73 |
| Gradient Boosted Tree | 65 | 66 | 69 |
| Linear SVM | 64 | 68 | 71 |

**Delay Prediction – Multiclass Classification**

|  |  |  |
| --- | --- | --- |
| **PREDICTION MODEL** | **AUC** | **BALANCED ACCURACY** |
| Logistic Regression | 50 | 81.6 |
| Random Forest | 51 | 73.3 |

**Cancellation Prediction – Binary Classification**

|  |  |  |
| --- | --- | --- |
| **PREDICTION MODEL** | **AUC** | **BALANCED ACCURACY** |
| Logistic Regression | 75.9 | 93 |
| Random Forest | 69 | 90 |
| Gradient Boosted Tree | 80 | 94 |

The scoring criteria’s that we have used here for all the above models are:

* **AUC** – Area under the Receiver Operating Characteristic Curve
* **Balanced Accuracy** – Balanced accuracy is the weighted average accuracy of both the classes 0 and 1. We have calculated it by using the formula:
  + sensitivity <- TP / (TP + FN)
  + specificity <- TN / (TN + FP)
  + balanced accuracy <- (sensitivity + specificity)
* **F1 Score** – We have used the weighted F1 Score for evaluating our models

**Feature importance for delay prediction:**

* After analyzing feature importance from ensemble model – Random Forest, we found out the below top 5 features that are most important in predicting flight delays:
  + Origin Airport
  + Scheduled Arrival
  + Month
  + Day of Week
  + Day

A screenshot of a cell phone

Description generated with very high confidence

**Feature Importance for Cancellation Prediction:**

* After analyzing feature importance from ensemble model – Gradient Boosted Tree, we found out the below top 5 features that are most important in predicting flight cancellations:
  + Month
  + Day
  + Origin Airport
  + Day Of Week
  + Scheduled Departure

A screenshot of a cell phone

Description generated with very high confidence

**AUC Curve - Delay Prediction – Logistic Regression**

A screenshot of a map

Description generated with very high confidence

**AUC Curve – Cancellation Prediction – Logistic Regression**

A close up of a map

Description generated with very high confidence

# **5.Conclusion**

We analyzed 5.5 Million observations of delay and cancellation flight data. In our project our goal was to predict delay and cancellation of flights. For predicting delay of flights, we took two approaches, first is Binary classification where classes are 0 for non-delayed flights and 1 for delayed flights and second Multiclass Classification where classes are 0 for non-delayed, 1 for flights with delay greater than 15 minutes and 3 for delay more than 30 minutes. For predicting cancellation of flights, we worked on Binary classification solution where classes are 0 for non-cancelled flights and 1 as cancelled flight.

Below table shows us the best model for each problem and their evaluation metrics:

|  |  |  |  |
| --- | --- | --- | --- |
| Problem | Best Model | AUC | Balanced Accuracy |
| Delay Prediction (Binary Classification) | Random Forest | 65 | 81 |
| Delay Prediction (Multiclass Classification) | Random Forest | 51 | 73.3 |
| Cancellation Prediction (Binary Classification) | Gradient Boosted Tree | 80 | 94 |

Using our analysis, we can predict delay and cancellation of flights using flight data. Our analysis will be helpful for the large air carriers which will eventually lead to achieving higher customer satisfaction, better customer reputation and customer retention.

Air carriers can leverage our analysis and focus on important features from our results. Below are the main features affecting delay of flights and what aspects carriers can investigate for reducing delay in flights:

|  |  |
| --- | --- |
| Important Feature | Investigation |
| Origin Airport | Which airports cause more delay in flights. Reason can be inability of airport to handle more traffic. |
| Scheduled Arrival | Scheduled Arrival can be a factor which can be affected from airline fly time and inability of destination airport to give clearance for flight for landing. |
| Month | Colder Months can result in more delays in flights |
| Day of Week | Weekends can result in more delays in flights as there is more air traffic on weekends |
| Day | Similarly, to Day of Week, Day can be a factor for delays on Saturday and Sunday |

Similarly, carries can investigate below important features for reducing cancellation in flights:

|  |  |
| --- | --- |
| Important Feature | Investigation |
| Month | Colder Months can result in more cancellations in flights |
| Day | Weekends can result in more cancellation in flights as there is more air traffic on weekends |
| Origin Airport | Which airports cause more cancellation in flights. Reason can be inability of airport to handle more traffic. |
| Day of Week | Similarly, to Day, Day of Week can be a factor for cancellation on Saturday and Sunday |
| Scheduled Departure | Scheduled Departure can be a factor which can be affected from previous airline not arriving on time giving less time for maintenance checks and inability of airport to handle high takeoofs.to give clearance for flight for landing. |

# **6. Appendix**

Please find the IPYNB notebook files:

* EDA\_Delay
* BDA\_Multiclass
* SVM\_Delay
* BDA\_PredictCancellation