Flight Delay Classification and Prediction Models based on Naïve Bayes, Regression Tree, and LogisticRegression Algorithms

Divyansh Shah
Faculty of Business
Humber Institute of Technology and Learning
Toronto, Canada
n01472284@humbermail.ca

Subhanjan Das
Faculty of Business
Humber Institute of Technology and Learning
Toronto, Canada
n014314743@humbermail.ca

Yuvraj Shand
Faculty of Business
Humber Institute of Technology and Learning
Toronto, Canada
n01479401@humbermail.ca

Abhi Tyagi
Faculty of Business
Humber Institute of Technology and Learning
Toronto, Canada
n01474042@humbermail.ca

Hardi Patel
Faculty of Business
Humber Institute of Technology and Learning
Toronto, Canada
n01480409@humbermail.ca

***Abstract*— The aviation or airline industry is undoubtedly one of the most rapidly growing industries currently. This has made aviation to be one of the most regulated industries as global economic dependence grew. A delay in the arrival of flights can have a massive financial and reputational impact on the airline but the losses are not only limited to the airless, but air passengers also have to the bear majority of the losses. Hence, it is critically significant to predict delayed flights to minimize the losses faced by consumers and airlines. In this study, we examine the key variables that have an impact on the flight schedule. This study aims to build models which analyzed and classify the occurrence of flight delays with the help of Machine Learning models based on Naïve Bayes (NB), Classification and Regression Tree (CART) and Logistic Regression algorithms. We have used Chi Squared feature selection, which is a categorical feature selection. We also evaluate the performances of our models and compare them based on the accuracy metrics. With the presented dataset, all three models performed well, with an average accuracy of ~84%. The model that forecasts flight arrival status using logistic regression performs best (85.14%). This study will support and aid airlines, authorities, and passengers by relaying accurate flight information, which will decrease economic losses and increase passenger confidence.**

***Keywords*— Naïve Bayes, Classification and Regression Trees (CART), Linear Regression, Machine Learning.**

I. INTRODUCTION

The aviation or airline industry experienced a rapid and massive rise in the growing travel, transportation, and logistics sectors since the dawn of its adoption. As the airline industry became more regulated and widespread, the number of air passengers increased a staggering 534% from 310,441,392 passengers in 1970 to 1.97 billion passengers in 2005 [1]. This increased our dependency on air travel, which gave birth to the increasing problem of flight delays.

Flight delays have become widespread in every part of the world, especially in the major metro cities of North America because of the globalization movement. With the expansion of airlines and growing connectivity of destinations, we now routinely face scenarios that lead to congestion of jets and complexities in the Air Traffic Control systems which breeds major delays in the aviation industry [2]. The impact that flight delays have are generally underestimated as a layman does not possess an outlooking perspective. These delays can cause significant economic damage to the passengers as well as the airline directly. In many cases, the airline's reputation gets a huge blow as an impact of the delays [3]. According to a recent study led by researchers at the University of California, Berkeley, domestic flight delays cost the United States economy a staggering $32.9 billion, with airline customers bearing nearly half of that expense and $8.2 billion directly affecting the airlines [4][5]. According to the Air Travel Consumer Report by the Department of Transportation, 28.8% of the total complaints were received concerning flight delays and deviations [6]. It is therefore indispensable to conduct substantial studies by adopting newer methodologies and approaches to help minimize the losses and support the airline authorities in decision-making related to flight delays. Delays in flight departure and arrival occur due to a multitude of factors with the most relevant being the weather. Predicting flight delays is hence an extremely challenging research avenueas the delay is induced by several mutually correlated factors.

In this paper, we attempt to build realistic Machine Learning models based on Naïve Bayes (NB), Classification and Regression Tree (CART) and Logistic Regression by extracting and analyzing the flight delays dataset. All commercial flights that left the Washington, D.C., region and arrived in New York in January 2004 are included in this dataset. We extend our research by evaluating theperformances and efficiencies of each model, to accurately classify the status of a flight.

II. DATA

The dataset includes statistics on each commercial aircraft that left Washington, D.C. and arrived in New York in January 2004.

Table 1. Data Description

Variable	Description	Data Type
CRS_DEP_TIME	The scheduled departure time of the flight	Integer
CARRIER	Name of flight	Ordinal
DEP_TIME	The departure time of the flight	Integer
DEST	Destination of the flight	Nominal
DISTANCE	Geographical location difference between the two airports in terms of kilometres	Integer
FL_DATE	The departure date of the flight	Date
FL_NUM	The route number is followed by the flight	Nominal
ORIGIN	The geographical location of flight departure	Nominal
Weather	The climate condition	Nominal
DAY_WEEK	The day on which the flight departures	Ordinal
DAY_OF_MONTH	The month of flight departures	Ordinal
TAIL_NUM	Flight identification number	Nominal
Flight Status	The current position and situation of the flight	Ordinal

There are 3 integer-type columns the scheduleddeparture time of the flight, departure time, and destination of the flight, along with that there are 8 variables which have unique descriptive identification for the flight dataset such as geographical location between the airports, the climate condition, the flight identification number, and the current position [7].

Along with that, there are 3 columns with values or observations that are ranked such as the name of the flight, the departure day of the flight, the month of flight departure, and the status of the flight [7].

To note the dataset, have a column name weather that has a distinct type which differentiates 0 to be good weather and 1 to be bad. We do not have any NULL values.

III. DATA EXPLORATION

The flight data in the CSV is from 1st January to 31st January 2004. There are a total of 2201 flights out of which 19.45% of the flights were delayed and 80.55% of the flights were on-time. Nearly all flights originated from Washington National (DCA) airport.

62.24% of flights originated from Washington (DCA) they had a mind-boggling 10.04% delay and most of the flight delays were not due to bad weather. The second greatest number of flights were from Dulles Airport (IAD) with 31.16% of the total flights, 7.72% of the flight delays were from Dulles (IAD) airport and for most delays the weather condition was good and only 6.59% of the flights originating from Washington International (BWI) airport the delays were just under 2%.

The dataset contained details of 2201 flights that flew in January. We looked at a few insights from the datato analyze major parameters of the data.

Table 2. Flight Status vs Weather

Weather	**0**	**1**
Flight Status		
Delayed	396	32
Ontime	1773	0

Table 2 illustrates that, during a given bad weather (1) condition, none of the flights were on time and 32 flights were delayed. However, under favorable conditions (0), 396 flights were delayed and 1773 flights to their destination on time.

Table 3. Count of flights from different origins on days of the week

Origin	**EWR**	**DCA**	**IAD**
DAY _WEEK			
1	24	237	111
2	25	246	120
3	24	137	89
4	19	152	82
5	17	196	95
6	17	198	92
7	19	204	97

Table 3 breaks down the number of flights originating from different airports by days of the week. It is evident that DCA Airport is the busiest airport while EWR being the quietest. Also, maximum flights are scheduled to fly on the second day of the week.

Table 4. Count of flights of different carries vs flight status

Flight Status	Delayed	Ontime	Average Delay
CARRIER			
CO	26	68	0.28
DH	137	414	0.25
DL	47	341	0.12
MQ	80	215	0.27
OH	4	26	0.13
RU	94	314	0.23
UA	5	26	0.16
US	35	369	0.09

Table 4 helps us analyze the number of flights that have been delayed or arrived on time from different carriers. CO airline has the greatest number of average delays at 26 delays of 94 flights, closely followed by MQ airline with 80 delays of 295 flights. The least delays were with US airlines with only 35 delays out of 404 flights.

Table 5. Count of flights with different destinations on days of the week

DEST	EWR	JFK	LGA
DAY_WEEK			
1	113	63	196
2	122	64	205
3	86	59	105
4	72	55	126
5	90	47	171
6	86	48	173
7	96	50	174

From table 5 we can see that the maximum number of flightsarrive at LGA airport with Mondays being the busiest. And the last flights arrive at JFK with Thursday the slowest.

IV. WORKING

This work aims to create 3 classification models using Naives Bayes Classifier, CART Algorithm and Logistic Regression to predict if a flight will arrive on time or bedelayed on the different parameters of the dataset.

We created a new column from the scheduled departure time column and actual departure time column named 'delayed_departure' that says if the flight took off ontime or was delayed where 1 is ontime and 0 is delayed. Out of the 2201 flights recorded 2 flights that were scheduled for a later night took off after midnight who's delayed_departure was manually changed to 0. Along with that the Flight Status column was also changed to

To synchronize the way people would count the day of the week we replaced Sundays with 1 Saturday to 7. Furthermore, the scheduled departure time was changed to just having the hour number in the cells to analyze between hours. All the nominal columns had their datatype change from 'int' to 'category' as these algorithms work better with categorical columns.

Table 6. Reason for removal of columns

Removed Variable	Reasons
DEP_TIME	A new variable using the same was created and keeping this would lead to redundancy.
Distance	As all the data in the set are categorical
FL_DATE	As Day of week and day of month already exist, reducing redundancy.
FL_NUM	Too many categories that do not contribute to the result
TAIL_NUM	Too many categories that do not contribute to the result

Along with this we used the Chi-Squared Feature selection method [8] to find if any of the remaining columns can be removed from the dataset that wouldn't work well with the model. Though there were a few feature dummies that had low fs scores, but a few dummies of that feature had good scores. We decided to keep all the other features to train the model on.

The above dataset was split with 67% training data and 33% testing data. All the variables except Flight status had their dummies created for the model to fit smoothly.

V. RESULT

All the 3 models performed well with the dataset provided and all of them average approx. ~84% accuracy.

Table 7. Predictions of models on split test data

True Label	Predicted Label	Naïve Bayes	Decision Tree	Logistic Regression
Ontime	Ontime	569	515	543
Ontime	Delayed	20	74	46
Delayed	Ontime	95	54	62
Delayed	Delayed	43	84	76

Table 8. Accuracy of models

Model	Accuracy
Naïve Bayes Classifier	84.18%
Decision Tree	82.39%
Logistic Regression	85.14%

From table 7 and table 8 we understand that the model that uses Logistic Regression to predict values works the best with 85.14%.

We randomly selected 5 rows from the dataset to predict what the status of their arrivals could be and all three models predicted the same outcome for each flight.

Table 9. Predicting 5 random flights

Flight	**Naïve Bayes Classifier**	**Decision Tree**	**Logistic Regression**
1	ontime	ontime	ontime
2	delayed	delayed	delayed
3	ontime	ontime	ontime
4	ontime	ontime	ontime
5	ontime	ontime	ontime

ACKNOWLEDGMENT

We extend our sincere gratitude to the Humber Institute of Technology and Learning for guiding us throughout the process of this work.

REFERENCES

[1] *Air Transport, passengers carried*. Data. (n.d.). Retrieved December 3, 2022, from https://data.worldbank.org/indicator/IS.AIR.PSGR?end=2020&start=1970&view=chart

[2] Shadare, W. (2022, February 15). *Flight delays cost more than just time, airlines' reputation at stake*. Aviation metric. Retrieved December 3, 2022, from https://aviationmetric.com/flight-delays-cost-more-than-just-time-airlines-reputation-at-stake.

[3] N. L. Kalyani, G. Jeshmitha, B. S. Sai U., M. Samanvitha, J. Mahesh and B. V. Kiranmayee, "Machine Learning Model - based Prediction of Flight Delay," 2020 Fourth International Conference on I-SMAC (IoT inSocial, Mobile, Analytics and Cloud) (I-SMAC), 2020, pp. 577-581, doi: 10.1109/I-SMAC49090.2020.9243339.

[4] Ball, M.O., Barnhart, C., Dresner, M.E., Hansen, M., Neels, K., Odoni, A.R., Peterson, E.B., Sherry, L., Trani, A.A., & Zou, B. (2010). *Total Delay Impact Study: A Comprehensive Assessment of the Costs and Impacts of Flight Delay in the United States.*

[5] Guy, B. A. B. (2010, October 18). *Flight delays cost $32.9 billion, passengers foot half the bill*. Flight delays cost $32.9 billion, passengers foot half the bill | Research UC Berkeley. Retrieved December 3, 2022, from https://vcresearch.berkeley.edu/news/flight-delays-cost-329-billion-passengers-foot-half-bill

[6] *Air Travel Consumer Report: Consumer Complaints up from May,nearly 270 percent above pre-pandemic levels.* U.S. Department of Transportation. (2022, August 26). Retrieved December 3, 2022, from https://www.transportation.gov/briefing-room/air-travel-consumer-report-consumer-complaints-may-nearly-270-percent-above-pre

[7] Team, G. L. (2022, November 16). *4 types of data - nominal, ordinal, discrete and continuous*. Great Learning Blog: Free Resources what Matters to shape your Career! Retrieved December 3, 2022, from https://www.mygreatlearning.com/blog/types-of-data/

[8] Brownlee, J. (2020, August 18). How to perform feature selection with Categorical Data. MachineLearningMastery.com. Retrieved December 5, 2022, from https://machinelearningmastery.com/feature-selection-with-categorical-data/