Predicting flight delays

Kathryn Falvo
Faculty of Business
Humber Institute of Technology & Advance Learning
Etobicoke, Canada
N01508439@humber.ca

Lorenzo Pepito
Faculty of Business
Humber Institute of Technology & Advance Learning
Etobicoke, Canada
N01468745@humber.ca

Rebecca Bennett
Faculty of Business
Humber Institute of Technology & Advance Learning
Etobicoke, Canada
N01516530@humber.ca

***Abstract*— In the United States there are "more than 45, 000 flights and 2.9 million airline passengers'' each day [1]. Flight delay variables were analyzed to build models to predict whether a flight departure will be delayed or on time. Features of the provided dataset were normalized and examined to select independent variables. The selected predictors: CARRIER, ORIGIN, DISTANCE, DEST, DEPTIME, WEATHER, and DAYWEEK. MINS_DIFF were used to validate the accuracy of the FLIGHT_STATUS classification. These variables were selected for training three machine learning algorithms - Naive Bayes (NB), Classification and Regression Trees (CART), and Logistic Regression - to determine the most accurate model for predicting whether a flight is delayed or not. After running each model numerous times to ensure consistency of results, it was found that the CART model had the greatest accuracy at 85.68%. This model may be used for future predictions and may help consumers choose flights based on the most optimal airline, airport, and time of day to increase the probability of being on time.**

Keywords—flight delay, Naïve Bayes, CART, Logistic Regression

I. Introduction

Each day, one out of five flights get delayed, inconveniencing thousands of travellers [2]. Of course, these delays result in mistrust and frustration from consumers, as flight delays may have significant impacts such as missing a connecting flight, arriving late to important events, having to extend time off at work, etc. Consequences are also present for airlines as they lose consumer trust, possibly resulting in a decline in bookings in the future.

Consumers want to do their best to avoid these delays, which can be made possible using machine learning algorithms. Predictive models can estimate which flights are likely to be delayed based on predictor variables. This analysis explores flight data from commercial flights that departed the Washington, D.C., area and arrived in New York in January of 2004. Variables in the dataset included information regarding departure and arrival airports, route distance, the scheduled time and date of the flight, and so on.

During the data pre-processing and cleaning stages, features were either selected as predictor variables or deemed irrelevant. The features that were maintained and used in model training include CARRIER, ORIGIN, DISTANCE, DEST, DEPTIME, WEATHER, and DAYWEEK. MINS_DIFF. TAIL_NUM, FL_NUM, and FL_DATE were deemed irrelevant and removed from the model, as they simply act as identifiers for the plane and or flight, which is unneeded or done so with other included variables. The selected predictors were used to train a model for classification and prediction to determine whether a flight will depart on time or be delayed (a flight is considered delayed if leaving ≥ 15 minutes later than originally scheduled).

II. Pre-processing The Data

Given that our models will require categorical inputs, the data was subject to three main tasks relating to pre-processing.

First, the data was cleaned. Rows were checked for non-conforming or anomalous values, and variable names were reformatted and renamed for consistency and simplicity. To guarantee the veracity of the dependent variable that the models will predict, the values for FlightStatus (STATUS) were checked if they complied with the condition of what constituted the delay. Specifically, this condition was that flights were "delayed" if the actual departure (DEPTIME) was late by 15 minutes or more compared to its scheduled time (CRSDEPTIME). Moreover, delays under 15 minutes or ahead of schedule were considered "on time." 146 records were identified to be misclassified and were overwritten with the appropriate status accordingly.

Table 1 Original data table prior to pre-processing.

	CRS_DEP_TIME	CARRIER	DEP_TIME	DEST	DISTANCE	FL_DATE	FL_NUM	ORIGIN	Weather	DAY_WEEK	DAY_OF_MONTH	TAIL_NUM	Flight Status
0	1455	OH	1455	JFK	184	37987	5935	BWI	0	4	1	N940CA	ontime
1	1640	DH	1640	JFK	213	2004-01-01	6155	DCA	0	4	1	N405FJ	ontime
2	1245	DH	1245	LGA	229	2004-01-01	7208	IAD	0	4	1	N695BR	ontime
3	1715	DH	1709	LGA	229	2004-01-01	7215	IAD	0	4	1	N662BR	ontime
4	1039	DH	1035	LGA	229	2004-01-01	7792	IAD	0	4	1	N698BR	ontime

Table 2 Data table after pre-processing and variable selection.

	CARRIER	ORIGIN	DISTANCE	DEST	DEPTIME	DELAYMINS	WEATHER	DAYWEEK	STATUS
1	DH	DCA	213	JFK	Afternoon	0	0	4	ontime
2	DH	IAD	229	LGA	Morning	0	0	4	ontime
3	DH	IAD	229	LGA	Afternoon	-6	0	4	ontime
4	DH	IAD	229	LGA	Morning	-4	0	4	ontime
5	DH	IAD	228	JFK	Morning	-1	0	4	ontime

Second, the data was binned, or separated into groups. Features with a continuous data type, particularly DEPTIME and DISTANCE were binned to segment values into categories that can more easily be managed by the models. Coercing numeric values as categories will prove unwieldy for the algorithms, especially with the case at hand containing similar yet different values. To illustrate, a departure at 1:05 A.M. can be uniquely different from a departure at 1:08 A.M.; albeit, from a less granular perspective, these flights can be considered similar enough to be considered as having taken place past midnight or early morning of an airport's local time. From this perspective, DEPTIME was initially segmented by hour and thereafter segmented into 4 bins corresponding to the different periods of the day: Morning, Afternoon, Evening, and Overnight). This binning logic was analogously applied to DISTANCE, which was cut into four bins. This was necessary as turning every observation for features of continuous data types would be too complex for the models and underrepresented in the context of records as it becomes less likely that records will be similar to one another.

Third, the independent variables, particularly the predictors to be discussed in the next section, were made into dummy variables such that additional features were added to express each observed category as a feature, which had values that were limited to either 0 or 1. This step is crucial in running classification algorithms such as those being built.

III. Determining Prediction Variables

After preprocessing was complete, an exploratory data analysis was conducted to identify patterns within the dataset. The proportion of the target value was assessed, illustrating that 19.45% of total flights in the dataset were delayed.

Table 3 Delayed and on time flight status by airline carrier.

STATUS	delayed	ontime	All
CARRIER			
CO	26	68	94
DH	137	414	551
DL	47	341	388
MQ	80	215	295
OH	4	25	29
RU	94	314	408
UA	5	26	31
US	35	369	404
All	428	1772	2200

Once the ratio of on time and delayed flights was identified, the flight status by airline was then observed (CARRIER). It was seen that American Airlines (US) had the highest percentage of flights on time at 91.34% (delayed = 8.66%) whereas Continental Airlines (CO) had the lowest percentage at 72.34% on time and therefore the highest percentage of delayed flights at 27.66% (Table 3). In addition to airline, variables such as airport origin, destination and the total distance of the flight will also have an impact. Because the dependability of airlines and airports varied greatly, they were selected as a predictor variable for the models.

Table 4 Flight status by airline carrier grouped by time of day.

	DEPTIME	Afternoon	Evening	Morning	Overnight	All
CARRIER	**STATUS**					
CO	**delayed**	16	7	3	0	26
	ontime	23	1	44	0	68
DH	**delayed**	73	38	21	5	137
	ontime	184	62	135	33	414
DL	**delayed**	27	9	11	0	47
	ontime	150	43	132	16	341
MQ	**delayed**	40	18	17	5	80
	ontime	104	5	75	31	215
OH	**delayed**	4	0	0	0	4
	ontime	25	0	0	0	25
RU	**delayed**	55	17	16	6	94
	ontime	165	18	77	54	314
UA	**delayed**	0	0	5	0	5
	ontime	0	0	26	0	26
US	**delayed**	10	10	10	5	35
	ontime	143	42	146	38	369
All		1019	270	718	193	2200

Next, the departure time was evaluated. Initially, the binned times alone were assessed, then the exploration was made more granular by adding the airline variable. This

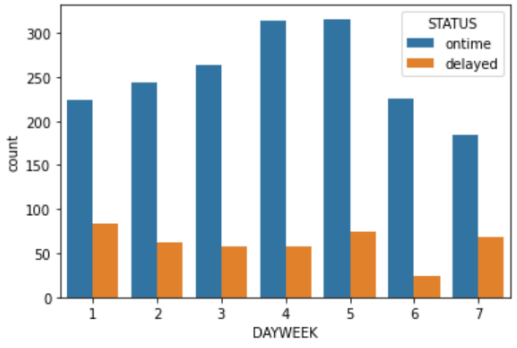

Fig. 1 Flight status based on the days of the week. With "1" representing Sunday, "2" representing Monday and so on.

provided insights as to the best time of day to fly, which was determined to be overnight with 89.12% of flights being on time, and the greatest number of delays occurring in the evening. To further evaluate the impact of time, day of the week was also analyzed with "1" representing Sunday, "2" representing Monday, and so on. From Fig. 1, it is seen that the least amount of flight delays occurs on Fridays, followed by Wednesdays, thus indicating that these are the best days to fly.

Based on the exploratory data analysis performed, the variables selected to be used in our models are as follows: CARRIER, ORIGIN, DISTANCE, DEPTIME, DELAYMINS, WEATHER, DAYWEEK, and STATUS (Table 2) [3].

IV. Model Comparison

Three models - Naïve Bayes (NB) model, Classification and Regression Tree (CART) and Logistic Regression - were then deployed to determine whether any given flight would be delayed or on time. The strengths and weaknesses of each model were investigated using clean training data and calculating the accuracy of each model. However, all three models are considered to be classification models, meaning that the test data is categorized based on the pre-categorized training dataset. The use case, advantages, and disadvantages of these models will be further explained and compared to determine the algorithm that should be used for future prediction.

A. Naïve Bayes (NB) Model

As a generative probability model, NB is used primarily for text classifications, sentiment analysis and spam filtering. This model is said to be naïve because it operates under the assumption that all features are independent [4]. This is one of the disadvantages of this model because rarely in real-life scenarios are features strictly independent from each other. Another disadvantage of NB is that data loss may occur when continuous variables are binned, required for this model specifically. Alternatively, NB will produce results with less training data, making it optimal for situations where limited training data is available. NB is also easy to implement and works quickly as probabilities can be directly computed [4].

B. Classification and Regression Tree (CART)

A CART is a tree-like diagram used to make non-linear decisions with a "simple linear decision surface" [5]. This is accomplished by setting decision rules from the root node to arrive at decision nodes and leaf nodes, creating sub-trees. The Gini index and entropy are used to create and select the next attribute [4]. Unfortunately, trees may grow to be incredibly complex and are sensitive to outliers.

To deal with the complexity, CART models must be pruned. Once the pruning is complete, CARTs can provide digestible insights into the predictors [4]. When compared to the NB model, CARTs are a discriminative model, whereas NB models are generative. Decision trees are also more flexible than NB, producing both categorical and numerical output types.

C. Logistic Regression (LR)

Logistic Regression is a discriminative classification algorithm that calculates the linear output and finds a relationship between features and the probability of a specific outcome where the response variable is categorical [6]. The logit function uses the Sigmoid function which creates an S-shaped curve resulting in a probability ranging from 0 to 1.

As a classification method, Logistic Regression may be used for multiclass classifications. This method provides insights regarding the directionality and significance of independent variables over the dependent variable; however, the insights are dependent on the features being appropriately selected [7]. Logistic Regression is unable to handle collinearity as well as CART, however it outperforms NB.

D. Best Classification Model

After running each model 15 times, it was determined that the CART model had the greatest accuracy and is the optimal predictive algorithm to use going forward (Fig. 2). The original CART had an accuracy of 82.12%, although it was incredibly intricate. To account for the complex nature of the CART model it was pruned so that it had only six sub-trees, which increased the accuracy to 85.68%.

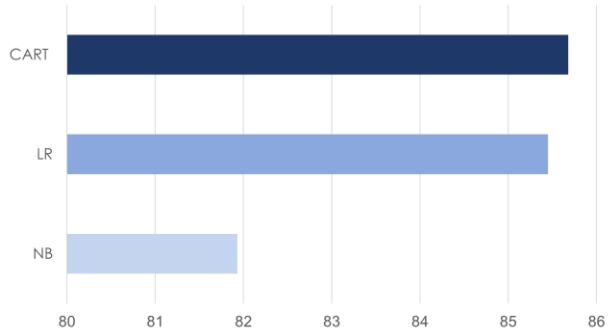

Fig. 2 Comparison of the accuracy of each of the three models (CART, LR, NB). CART was found to be the most accurate model at 85.68%.

Table 5 A random sample of five inputs predicted using the CART Model.

CARRIER	ORIGIN	DISTANCE	DEST	DEPTIME	WEATHER	DAYWEEK	STATUS
RU	IAD	213	EWR	Evening	0	2	**delayed**
US	DCA	214	LGA	Afternoon	0	3	**on time**
DH	IAD	228	JFK	Afternoon	0	2	**on time**
DL	DCA	214	LGA	Afternoon	0	1	**on time**
US	DCA	214	LGA	Morning	0	7	**on time**

We then decided to make 5 predictions based on randomly selected inputs, predicting either on time, or delayed departures. The predicted results of these randomly selected inputs are shown in Table 5. Please note that these predictions change with each run.

V. Conclusion

Based on the analysis performed, it has been determined that the CART model yielded the most accurate results (accuracy = 85.68%) for predicting flight delays. From the training data, it was observed that the airline carrier with the smallest percentage of delayed flights was American Airlines (US, = 91.34% of flights were on time). It was also observed that most flight delays occur in the evening, accounting for about 36% of delays. After the prediction and classification models were completed, it was found that the ideal flight for a traveller would be an overnight flight with American Airlines on either a Wednesday or Friday.

References

[1] "Air Traffic By The Numbers | Federal Aviation Administration," *Faa.gov*, Aug. 31, 2022. https://www.faa.gov/air_traffic/by_the_numbers

[2] H. Murphy, "Understanding the Summer Air Travel Mess," *The New York Times*, Jul. 01, 2022. Accessed: Dec. 09, 2022. [Online]. Available: https://www.nytimes.com/2022/07/01/travel/summer-travel-flight-delays-cancellations.html#:~:text=So%20far%20in%202022%2C%20an

[3] N. Kuhn and N. Jamadagni, "Application of Machine Learning Algorithms to Predict Flight Arrival Delays." [Online]. Available: http://cs229.stanford.edu/proj2017/final-reports/5243248.pdf

[4] D. Varghese, "Comparative Study on Classic Machine learning Algorithms," *Medium*, Dec. 06, 2018. https://medium.com/@dannymvarghese/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222 (accessed Dec. 08, 2022).

[5] T. Penumudy, "Decision Trees for Dummies," *Analytics Vidhya*, Jan. 29, 2021. https://medium.com/analytics-vidhya/decision-trees-for-dummies-a8e3c00c5e2e (accessed Dec. 08, 2022).

[6] A. Agrawal, "Logistic Regression. Simplified.," *Medium*, Mar. 31, 2017. https://medium.com/data-science-group-iitr/logistic-regression-simplified-9b4efe801389 (accessed Dec. 08, 2022).

[7] D. Varghese, "Comparative Study on Classic Machine learning Algorithms , Part-2," *Medium*, Dec. 11, 2018. https://medium.com/@dannymvarghese/comparative-study-on-classic-machine-learning-algorithms-part-2-5ab58b683ec0 (accessed Dec. 08, 2022).