Flight Delay Prediction

Abstract - Flight delays can cause airline carriers millions of dollars as they bear the costs of alternate flight options or compensations for the passengers. It also inconveniences passengers, resulting in a loss of demand. Various issues like bad weather conditions, air traffic, unavailability of runways, maintenance, etc. cause flight delays. These delays could even lead to cancellations which could jeopardize the reputation of the carrier and lead to a tremendous loss in revenue. Moreover, rectifying these inefficiencies could require an increase in effort and resources which would impact the cost of doing business. In this report, we will analyze the Flight delay data provided to predict which flights have a higher probability of getting delayed by applying machine learning algorithms like the Naïve Bayes model, Classification and Regression Tree, and Logistic Regression. This will help the carriers increase precautionary measures to avoid delays and smooth running of their flight operations. We observe that the Logistic regression model is appropriate to predict flight delays giving a test accuracy of 89%.

Keywords – Flight Delay, Delay Prediction, Naïve Bias, Logistic Regression, CART

I. INTRODUCTION

Commercial flights have become the most convenient form of transport as it is the fastest mode of transport and has brought the world closer to people. However, there have been many calamities like 9/11 and the most recent covid pandemic which led to a huge drop in revenue and loss of passengers and employees. After the pandemic, the re-entry of people into the skies has made for a bumpy ride. In the first half of 2022 alone, one in five flights were recorded to have delays (Adams, 2022). Flight delays also indicated that passenger baggage on these flights was most likely to be mishandled i.e. reported as lost, damaged, or stolen. Predicting delays can help in building a good strategy and precautionary and mitigation measures to avoid and control them, which will benefit the air carriers by minimizing loss, streamlining, and continuously improving flight operations which would boost customer service levels. In this paper, we will test out and compare various predictive machine-learning algorithms to produce the best model to predict flight delays.

II. DATA EXPLORATION AND FEATURES

To predict the flight status, we analyse the trends in flight delays dataset which features more than 28000 sample data with 13 variables for different carriers and their flight status in the month of January 2004. The data provides the details about flight origin, its destination, and the distance it will cover, planned departure time and the actual departure time of the flight, day, date and weather conditions on the given day and their flight status whether it was on-time or delayed. To assess the dependency, we plotted graphs for different variables and examined for any trend. Following is the analysis for different variables-

- Weather – This is an essential factor in predicting the flight status as bad weather conditions cause delays in flight. From the graph plotted, it is evident there is always a delay in flight when the weather conditions are not normal (Weather = 1).

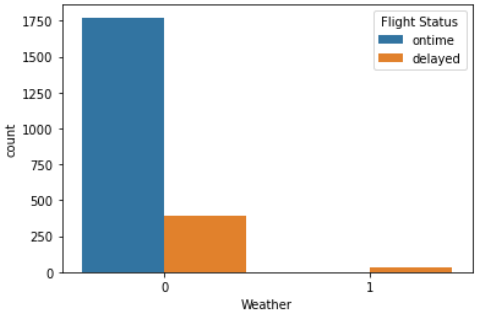

Figure 1. Flight status by Weather

- Carrier – Certain Airline companies are prone to delays due to various reasons such as frequent technical issues, staffing problem, weather constraints for the place of origin etc. Therefore, it helps in predicting the delay for a flight.

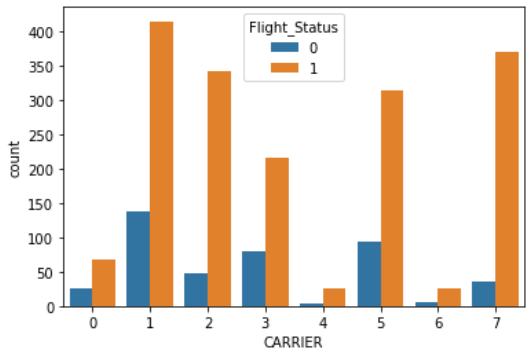

Figure 2. Flight status by Carrier

- Day of the Month – The delays can also be predicted based on the day and month, as holiday season is expected to be extremely busy, causing delays. Some months are colder than the others with bad weather conditions.

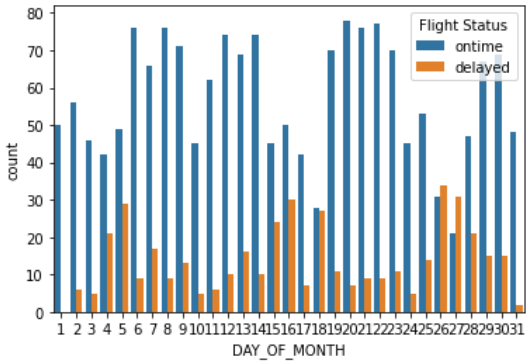

Figure 3. Flight status by Day of Month

III. EXPERIMENT DISCUSSIONs

A. Naïve Bias

Running on the basic concept of Bayes theorem, Naive Bayes is a classification technique and it developed by assuming that if a feature is present in a dataset, it is independent of the other features present in the same dataset. This algorithm is generally known to outperform even the highly complex machine learning models. For complex dataset such as ours for flight delay Naive Bayes makes it easier to predict the class of test datasets. In cases where the assumption of independence is true, the Naive Bayes algorithm performs much better than other algorithms. Ny providing the parameters of flight delays in this project, the Naive Bayes algorithm calculates the probability of a delayed flight. Once the model has been trained, it is applied to the testing dataset to get the accuracy

Our goal is to correctly predict whether a flight's departure will be delayed, where delayed is operationally defined as the actual departure is at least 15 minutes later than scheduled. Therefore, classification models were applied to discriminate negative examples (flight not delayed) from positive examples (flight delayed by at least 15 minutes). It is important to note that this is an unbalanced dataset. Therefore, a raw accuracy measure may be misleading because a naive classifier that solely predicts that every flight is not delayed achieves an accuracy of about 82%. Though this accuracy is higher than a coin flip, this model is uninformative. Therefore, we chose to evaluate other models.

B. Logistic Regression

By using regression model, we are trying to reach the highest accuracy to predict the flight delays. Though you can look at individual results and compare the predicted flight delay to the actual value you typically need to evaluate the success of the regression model to compute the overall prediction.

A mean squared error (MSE) of zero means that the model predicts the dependent variable with perfect accuracy. This is the ideal but is typically not possible. Likewise, an R-squared value of 1 indicates that all of the variances in the dependent variable can be explained by the feature variables. Typically, you compare the MSE and R-squared values from multiple regression models to find the best balance or fit for your data.

With regression we can also use the multiclass confusion matrix, it provides a summary of the performance of the classification analysis. It contains the number of occurrences where the analysis classified data points correctly with their actual class as well as the number of occurrences where it misclassified them.

C. CART

A Classification and Regression Tree (CART) is a predictive model, which explains how an outcome variable's values can be predicted based on other values. A CART output is a decision tree where each fork is split in a predictor variable and each end node contains a prediction for the outcome variable. For flight delays keeping in mind the 15 minutes window the tree will classify it into a delay or on time. Classification and Regression Trees: In CART, the decision trees will lay input factor such as Dep_Time, Destination, Weather, Origin etc. to identify values in the leaves. The branches of the trees then split the input values based on the observation values. This process is repeated till the predictions are reached. Decision trees: Decision trees are a supervised type of learning algorithm in which the data is continuously split according to certain parameters. A decision tree can be best explained with the help of two

terms, decision nodes and leaf nodes, the leaf nodes act as the outcomes, whereas the decision nodes are where the data is split. Decision trees are of two types, namely, Classification and Regression. While making a decision tree, different types of questions are asked, at each node of the tree. Based on the questions asked, the information gained can be calculated corresponding to it. The entropy of the dataset after a transformation is compared with that of the same dataset before the transformation, this allows one to calculate the information gain. It is then used to choose a feature which would then be split to further the tree.

IV. MODEL COMPARISION

Naïve Bayes, CART, and Logistic regression algorithms are applied on the dataset using Python programming language, and their accuracies were determined based on characteristics such as Dep_Time, Destination, Weather, Origin etc.

Naïve Bayes model	Cart algorithm	Logistic Regression Model
88.83%	87.30%	89.05%

Table 1. Model accuracy Comparison

From the above table we can see that Logistic Regression is observed to yield the highest accuracy in the prediction of flight delay.

Confusion matrix: Confusion matrix in machine learning is used to understand the performance of classification model. There are four terminology which are related to confusion matrix. They are as follows.

	Predicted Values	
Actual Values	True Positive (TP)	False Negative (FN)
	False Positive (FP)	True Negative (TN)

Table 2. Confusion Matrix

- True positive - This represents that situation where the actual value and the expected value are true. For example, the flight has been identified as delayed, and the model also expected that the flight was delayed.
- Fales Negative - This denotes that situation where the actual value is true, but the expected value is false. For example, the flight has been identified as delayed, and the model expected that the flight was not delayed.
- Fales positive - This characterizes that situation where the actual value is false, but the expected value is true. For example, the flight has not been identified as delayed, and the model expected that the flight was delayed.
- True Negative - This denotes that situation where the actual value is false, and the expected value is false as well. For example, actually the flight has not been delayed, and the model also anticipated that the flight was not delayed but it was delayed.

From the confusion matrix we got using the models, following are the values for various performance matrix:

	Naïve Bayes	**CART**	**Logistic Regression**
Accuracy	0.8934	0.839	0.873
Precision	0.907	0.625	0.9762
Sensitivity	0.4756	0.5914	0.4271
Specificity	0.9889	0.9052	0.9971
F1-Score	0.624	0.6077	0.5942

Table 3. Performance Matrix

V. CONCLUSION

By data exploration for trend analysis and running the correlation heatmap of all the variables with flight status, we were able to predict the flight delays effectively by using the factors – carrier, destination, weather, scheduled departure, actual departure, and day of the month. We used three models to get the prediction accuracies and found that the logistic regression model yielded the highest accuracy of prediction at 89.05%. Furthermore, we can increase the accuracy and granularity of the prediction by getting more detailed data on delays due to maintenance, shortage of crew, air traffic, etc.

VI. FUTURE SCOPE

Based on the classification and regression model, we can suggest using KNN model to predict the flight status as KNN is compatible with both classification and regression problems and can provide great accuracy for predicted results.

REFERENCES

[1] Sternberg, A., Soares, J., Carvalho, D., & Ogasawara, E. (2017). A review on flight delay prediction. arXiv preprint arXiv:1703.06118.

[2] Understanding the reporting of causes of flight delays and cancellations. Understanding the Reporting of Causes of Flight Delays and Cancellations | Bureau of Transportation Statistics. (n.d.). Retrieved December 9, 2022.

[3] Adams, K. (2022, August 22). 2022 has brought more air travel delays and cancellations - and nearly double the risk of having a bag mishandled. ValuePenguin. Retrieved December 9, 2022.

[4] Understanding the reporting of causes of flight delays and cancellations. Understanding the Reporting of Causes of Flight Delays and Cancellations | Bureau of Transportation Statistics. (n.d.). Retrieved December 9, 2022.

[4]