Predictive Modeling for Anticipating Flight Delays

Andres Ramirez, Gizelle Lao, Muhammad Ali Mirza, Rafael Frederico Muniz Albuquerque, and Trung Le

Business Insights & Analytics, Humber College Institute of Technology and Advanced Learning, Toronto, Canada

***Abstract* – An important application of machine learning models is predicting flight delays to assist businesses to diagnose operational problems and improve customer service. Airline businesses have various independent factors to consider, such as flight destination, flight origin, weather, day of the week, etc. In this paper, three algorithms – namely K-NN, Naïve Bayes, and Linear Discriminant Analysis – have been tested to model relationships between independent factors and flight delays. Test results show the best performance by the K-NN classifier which has an overall accuracy of 71% but correctly predicts 86% of actual flight delays.**

***Index Terms*—Flight delays, K-NN, Naïve Bayes, Linear Discriminant Analysis.**

I. INTRODUCTION

Accurately predicting flight delays can help airlines to reduce the cost of doing business by identifying operational inefficiencies and allocating appropriate resources to prevent or manage flight delays. The dataset contains records for all flights from the Washington, DC area into the New Your City area during January 2004.

The remaining content of the paper is structured in the following way. Section II describes the data pre-processing that includes data exploration and oversampling to balance classes before training the models. Section III discusses model training and performance, and Section IV concludes the results.

II. DATA PRE-PROCESSING

The dataset comprises flight data for the month of January 2004. There are 2,201 records, 13 variables, and no null values. The variables' descriptions can be seen in Table 1.

Table 1: Data Attributes

Attributes	Description	Data type
CRS_DEP_TIME	Expected flight time departure	Integer
CARRIER	Code of the carrier	Object
DEP_TIME	Actual flight time departure	Integer
DEST	Flight destination	Object
DISTANCE	Flight distance in miles	Integer
FL_DATE	Flight date	Object
FL_NUM	Flight number	Integer
ORIGIN	Flight Origin	Object
Weather	0: Good, 1: Bad	Integer
DAY_WEEK	Day of the week	Integer
DAY_OF_MONTH	Day of the month	Integer
TAIL_NUM	Airplane tail number	Object
Flight Status	Ontime or delayed flight	Object

We made some data exploration and preprocessing to understand the behavior of the flights. Figure 1 shows that Thursdays and Fridays are the days with the highest number of flights on time. On the other hand, the days with the highest number of delayed flights are Mondays, Fridays, and Sundays.

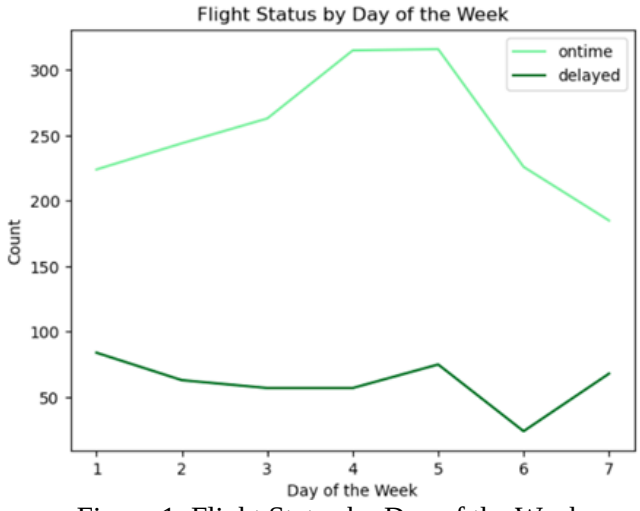

Figure 1: Flight Status by Day of the Week

Figure 2 shows that the days of the month with the most delayed flights are days 5, 16, 18, and 26.

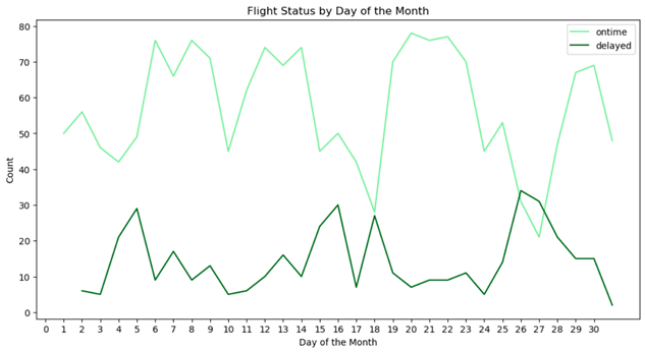

Figure 2: Flight Status by Day of the Month

As shown in Figure 3, Ronald Reagan Washington National Airport (also known as DCA) stands out not only for having the highest number of on-time flights among all origin airports but also for having the highest number of delayed flights.

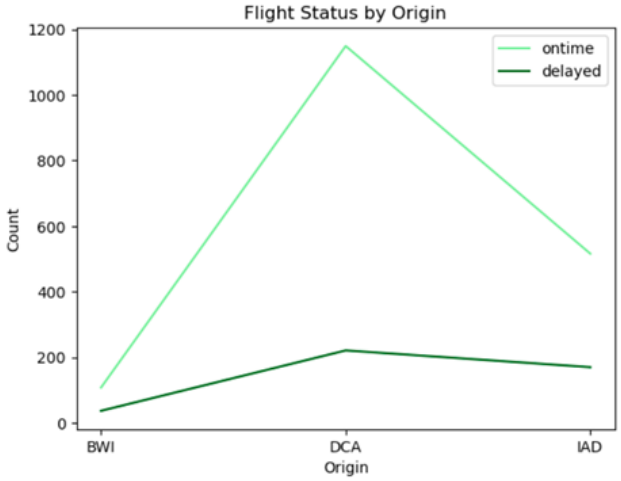

Figure 3: Flight Status by Origin

Figure 4 illustrates that LaGuardia Airport (also known as LGA) not only has the highest number of on-time arrivals among all destination airports but also the highest number of delayed flights.

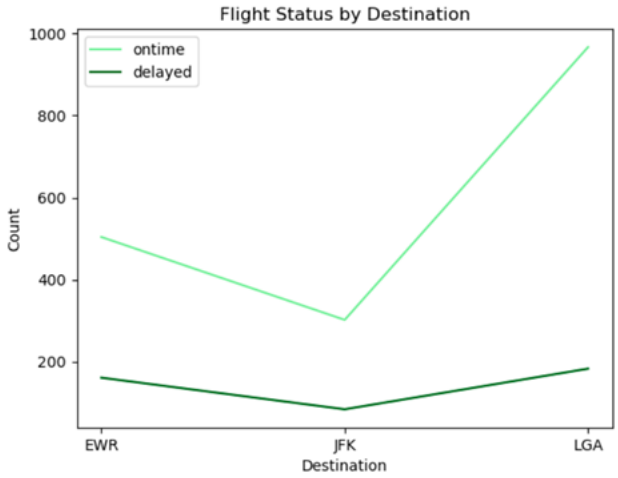

Figure 4: Flight Status by Destination

Figure 5 shows that out of the total 2,201 flights, 1,773 arrived on time while 428 experienced delays. We also observe that the dataset has an imbalanced class distribution where on-time flights account for 80% of the data and delayed flights account for 20%.

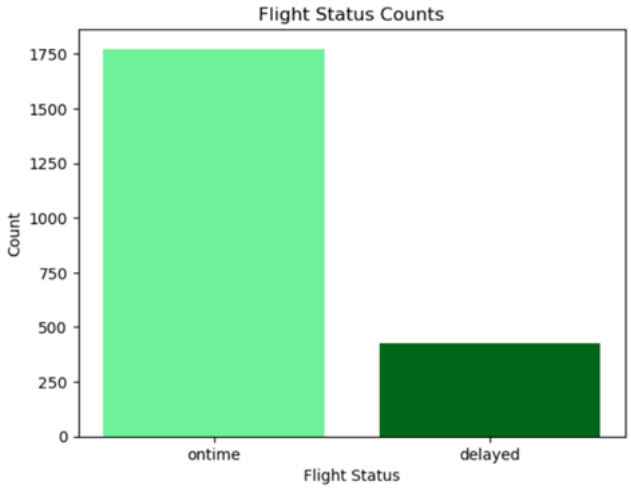

Figure 5: Flight Status Class Distribution

Due to the imbalanced nature of the target variables, we oversampled the delayed attribute by randomly replicating its data points to match the length of the on-time attribute. This balancing process ensures that both attributes are equally represented for modeling purposes. Figure 6 shows that after oversampling both classes have an equal count of 1773 records.

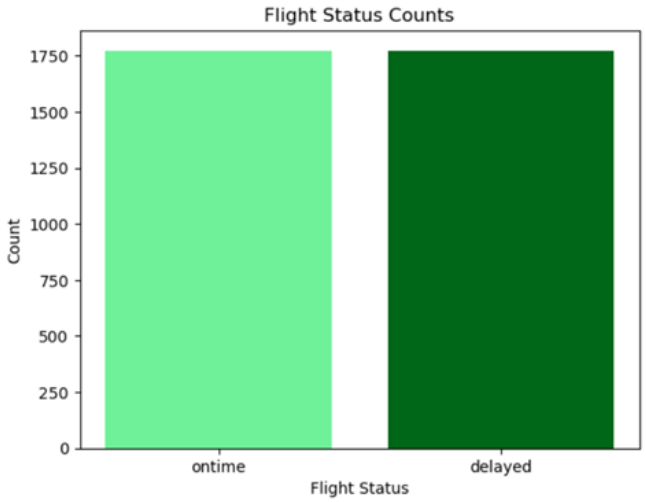

Figure 6: Balanced Flight Status Class Distribution

III. MODEL TRAINING AND PERFORMANCE

K-Nearest Neighbors (K-NN)

The main objective of this model is to classify the specific data according to the nearest data. The model will consider the classification of the majority around it to classify the analyzed data. The key point in this model is to use a good K value, which is the number of nearest values around the analyzed data to determine its classification.

For the dataset analyzed in this case, different values of K were tested (from 1 to 14) using the training data to check the best accuracy result. Table 2 below shows that k=5 has the best accuracy result, so that is the K value that will be used in the definitive model training:

Table 2: Finding Optimal K

k	accuracy
1	0.722
2	0.717
3	0.718
4	0.698
5	0.714
6	0.702
7	0.724
8	0.706
9	0.711
10	0.701
11	0.698
12	0.697
13	0.689
14	0.698

After knowing the best K value for this case, the model was trained, and the validation part of the dataset was used in the model to get the classification predictions as result. The results are shown in Tables 3 and 4 below:

Table 3: Confusion Matrix K-NN

Accuracy 0.71		**Prediction**	
		delayed	on-time
Actual	delayed	605	98
	on-time	308	408

Table 4: Classification Report K-NN

	precision	**recall**	**f1-score**	**support**
delayed	0.66	0.86	0.75	703
on-time	0.81	0.57	0.67	716

The output of the K-NN trained model was compared with the real classification values, and the result of this comparison is shown in the table above (table 3), in which it is demonstrated that the model has a 71% accuracy in the classification task. About the confusion matrix (table 3 above), it has 605 true positives: this is the number of correct predictions that the flights will be delayed. 308 false positives reflecting the number of incorrect predictions that the flights will be delayed. Also, there are 408 true negatives which show the number of correct predictions that the flights will be on time, and 98 false negatives [1] which are the number of flights predicted to be on time but were delayed.

Naïve Bayes (NB)

Naïve Bayes is a categorical classification model that assigns variables into specific categories based on their characteristics. It requires categorical variables and numerical variables need to be binned into categories. The model calculates the probability of each class for the values of predictor variables, and it is used to determine the likelihood of a variable falling into a specific category. Naïve Bayes is used as a benchmark against other classification algorithms for accuracy determination [2]. However, it tends to overfit or underfit the data [3], which can be seen through the predicted probabilities of the training and validation datasets. The choice of alpha can impact the model's output.

Table 5: Training Data Prediction Probabilities

	Delayed	On-Time
0	0.559	0.441
1	0.580	0.420
2	0.256	0.744

Table 6: Validation Data Prediction Probabilities

	Delayed	On-Time
0	0.308	0.692
1	0.798	0.202
2	0.482	0.518

After running the Naïve Bayes model, the confusion matrix (table 7) and the classification report (table 8) were used to assess that the Naïve Bayes algorithm was conducted correctly, confirming the accuracy of the algorithm. According to table 6 below, the confusion matrix has 475 true positives: this is the number of correct predictions that the flights will be delayed. 320 false positives reflecting the number of incorrect predictions that the flights will be delayed. Also, there are 396 true negatives which show the number of correct predictions that the flights will be on time, and 228 false negatives which are the number of flights predicted to be on time but were delayed. With these values, the model can determine the accuracy of the model, which is 61%.

Table 7: Confusion Matrix NB

Accuracy 0.61		**Prediction**	
		delayed	on-time
Actual	delayed	475	228
	on-time	320	396

Table 8: Classification Report NB

	precision	recall	f1-score	support
delayed	0.60	0.68	0.63	703
on-time	0.63	0.55	0.59	716

Linear Discriminant Analysis (LDA)

According to Balakrishnama et al. (1998), Linear Discriminant Analysis (LDA) is an approach for both data classification and dimensionality reduction. It effectively addresses scenarios where there is uneven distribution among classes and has been evaluated using randomly generated test data. This technique increases the ratio of between-class variance to within-class variance, which leads to a high level of separability and ensures optimal performance on any given dataset [4].

In this scenario, we apply LDA to classify flight status of "delay" or "ontime" and use a confusion matrix to illustrate in Table 9 and the classification report in Table 10.

Table 9: Confusion Matrix LDA

Accuracy 0.66		**Prediction**	
		delayed	on-time
Actual	delayed	488	215
	on-time	261	455

Table 10: Classification Report LDA

	precision	recall	f1-score	support
delayed	0.65	0.69	0.67	703
on-time	0.68	0.64	0.66	716

The f1-score represents the accuracy of the model by combining the precision and recall scores of a model. The overall accuracy of the Linear Discriminant Analysis model is 0.66, (or 66%) for the total samples in the validation data.

According to table 9 above, the confusion matrix illustrated 448 true positives: this is the number of correct predictions that the flights will be delayed. There are 261 false positives reflecting the number of incorrect predictions that the flights will be delayed. Also, there are 455 true negatives which show the number of correct predictions that the flights will be on time, and 215 false negatives [3] which are the number of flights predicted to be on time but were delayed. With these values, the model can determine the accuracy of the model, which is 66%

ROC Curve Summary

In addition to the classification report, we can evaluate the performance of Linear Discriminant Analysis for classification by plotting a ROC curve. The ROC curve, which stands for the Receiver Operating Characteristic curve, is a graph that illustrates the performance of a classification model at all classification thresholds. This curve plots two parameters: the True Positive Rate (TPR) and the False Positive Rate (FPR) [5].

A good classifier will have a ROC curve that is closer to the top left corner of the plot, indicating a high true positive rate and low false positive rate.

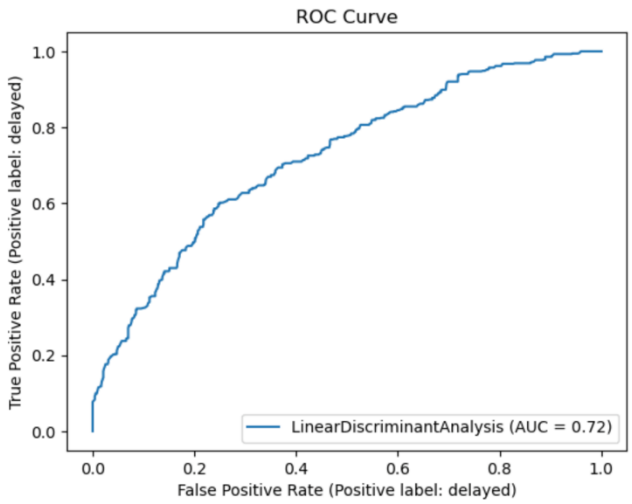

Figure 7. ROC Curve

Taking the TPR of 0.6 means that the model has correctly classified 60% of the positive outcomes, which is "delayed". However, at this threshold, the respective FPR is 0.2 which means it falsely classifies 20% of the "delayed" outcome; or in other words, it classified the 20% “on-time" as “delayed”. However, if we adjust the threshold higher to 0.8 (80%) of true possible, the false positive increase to 0.6 (60%), respectively.

In addition to this, the Area Under the Curve (AUC) measures the level of separability between classes and indicates the model's ability to distinguish between them. A higher AUC score indicates better performance in predicting 0 as 0 and 1 as 1. In this case, an AUC of 0.72 indicates that there is a 72% chance the model will be able to distinguish between on-time and delay correctly [6].

Based on the ROC Curve and AUC metric, the model performance is not considered as an effective classification at a higher threshold.

IV. CONCLUSION

The performance of all three models is summarized in Table 11 below:

Table 11: Model Performance Summary

Model	**Overall Accuracy**	**Recall of ‘delayed’ class**
K-NN	0.71	0.86
NB	0.61	0.68
LDA	0.66	0.69

Based on our analysis, we conclude that K-NN is the best model amongst the three as it has the highest overall accuracy and because it correctly predicts actual flight delays 86% of the time. It also performs significantly better than the baseline model, Naïve Bayes. Therefore, it is adding considerable predictive value.

We acknowledge that these results were achieved on a small dataset comprising only 2,201 flight records. Therefore, predictive performance may be improved by collecting more data as well as exploring other modeling techniques, such as decision tree classifiers.

REFERENCES

[1] Nugroho, A., & Fahmi, R. A. (2017). On-Time Flight Departure Prediction System Using Naïve Bayes Classification Method (Case Study: XYZ Airline). International Journal of Computer Trends and Technology, 54(1), 4-10.

[2] Menezes, R. (2019, August 28). Introduction to Naïve Bayes Classifier. Towards Data Science. https://towardsdatascience.com/introduction-to-na%C3%AFve-bayes-classifier-fa59e3e24aaf

[3] Parekh, R. (2021, January 23). Understanding Machine Learning Algorithms: Naïve Bayes. Analytics Vidhya. https://medium.com/analytics-vidhya/understanding-machine-learning-algorithms-Naïve-bayes-808ed649c1ec

[4] Balakrishnama, S., & Ganapathiraju, A. (1998). Linear discriminant analysis-a brief tutorial. Institute for Signal and Information Processing, 18(1998), 1-8.

[5] Google. (n.d.). ROC and AUC. Machine Learning Crash Course. Retrieved March 26, 2023, from https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

[6] Narkhede, S (2018, May 21). Understanding AUC-ROC curve. Towards Data Science. Retrieved March 26, 2023, from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5