Predicting flight delays using Machine Learning: K-Nearest Neighbors, Naive Bayes Classifier and Random Forest algorithms.

Literature review

Flight delays result in significant costs for both airlines and passengers. While certain factors, such as adverse weather conditions, runway closures, or air traffic control issues, are beyond the airlines' control, recent statistics from the Bureau of Transportation Statistics (n.d) indicate that a considerable portion of flight delays from January 2022 to December 2023 were attributable to the airlines themselves. Specifically, Air Carrier Delay and Aircraft Arriving Late contributed the highest percentages to the delay factors, at 7.18% and 7.19% respectively, compared to uncontrollable factors such as Weather Delay (0.72%), Security Delay (0.06%), and National Aviation System Delay (5.08%). This suggests that many delays could have been prevented with better planning and operational management. It's important to note that according to the Bureau of Transportation Statistics, a delay is considered if airlines arrive 15 minutes later than scheduled. This threshold underscores the significance of timely arrivals and highlights the impact of even minor deviations from the planned schedule.

Costs Incurred due to Flight Delays

In 2010, a research study commissioned by the Federal Aviation Administration (FAA) conducted a comprehensive analysis of the costs attributed to flight delays, encompassing impacts on airlines, passengers, and broader economy. The findings revealed a staggering total cost of $32.9 billion. Remarkably, over half of this financial burden, amounting to $17.7 billion, was shouldered by passengers. This calculation factored in various variables, including lost passenger time, flight cancellations, missed connections, and expenses incurred for food and accommodations as a result of being away from home.

The remaining $8.3 billion represented expenses borne by airlines. These costs included expenditures for crew, fuel, maintenance, and expenses associated with forced rescheduling due to runway capacity limitations. Additionally, there were costs associated with the loss of demand from potential airline customers.

Moreover, beyond the direct impacts on passengers and airlines, evidence from the study highlighted broader economic repercussions. The research indicated that flight delays had led to a reduction in the U.S. gross domestic product by approximately $4 billion, primarily attributable to the loss of productivity stemming from delayed travel and disrupted schedules. These findings underscore the far-reaching implications of flight delays, not only on the aviation industry but also on the broader economy.

Purpose

The objective of this paper is to assist our client in enhancing their services and predicting flight delays more accurately. The primary goal is to develop a classification model capable of determining whether a flight will be delayed or not. Crucially, the focus will be on minimizing false positives, where the model incorrectly predicts a flight to be on time when it is actually delayed.

Given that many flight delays are avoidable through meticulous planning, the successful implementation of this model has the potential to significantly improve customer experience and scheduling accuracy for our client. By accurately identifying flights at risk of delay, proactive measures can be taken to mitigate potential disruptions, thereby enhancing overall operational efficiency and customer satisfaction. Through this endeavor, our client aims to optimize their services, minimize the impact of delays, and elevate the reliability of their flight schedules.

Data Overview

The dataset used for both training and testing models have 2201 data points with variables as follow:

- CRS_DEP_TIME = Scheduled Departure Time
- CARRIER = Airlines Code
- DEP_TIME = Actual Departure Time
- DEST = Airport Code of the Flight Destination
- DISTANCE = Distance of the flight
- FL_DATE = Flight Date
- FL_NUM = FLight Number
- ORIGIN = Airport Code of the Flight Origin
- Weather = Weather Code
- DAY_WEEK = Day of the Week in Number
- DAY_OF_MONTH = Date in Month
- TAIL_NUM = Tail Number of the Airlines
- Flight Status = Real Status of the Flight
- dep_delay_in_min = Delay in Minutes by Subtracting Actual Departure Time with Scheduled Departure Time. A positive results mean flights departed later than scheduled, and negative results mean flights depart earlier than scheduled

Descriptive Statistics

After conducting exploratory data analysis and data cleaning, a detailed examination of the dataset revealed interesting insights into the distribution of scheduled and actual departure times. The analysis in figure 1 indicates that the peak of both scheduled and actual departure times occurs between 14:00 and 15:00. However, notable disparities were observed between the scheduled and actual departure times during specific hours. For instance, while the scheduled departure times suggest approximately 250 flights were planned to depart at 16:00, the actual departure data reveals that over 400 flights departed during this hour. Similarly, during the peak hour of flight, where around 320 flights were scheduled to depart, the actual number of flights departing at 14:00 exceeded expectations, totaling roughly 450 flights.

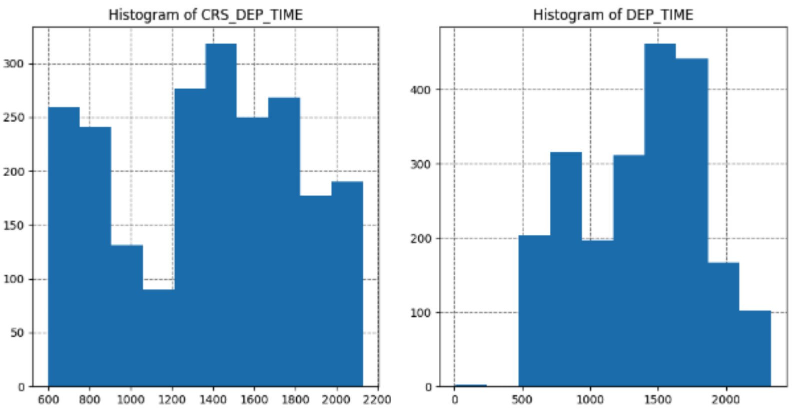


Figure 1. Scheduled Departure Time versus Actual Departure Time

Figure 2 reveals that the highest number of flights originate from DCA airport, with a predominant destination being LGA airport. This suggests a significant air traffic route between these two locations. Additionally, it was observed that approximately 63% of total flights cover a distance between 210 and 220, indicating a concentration of flights within a specific range.

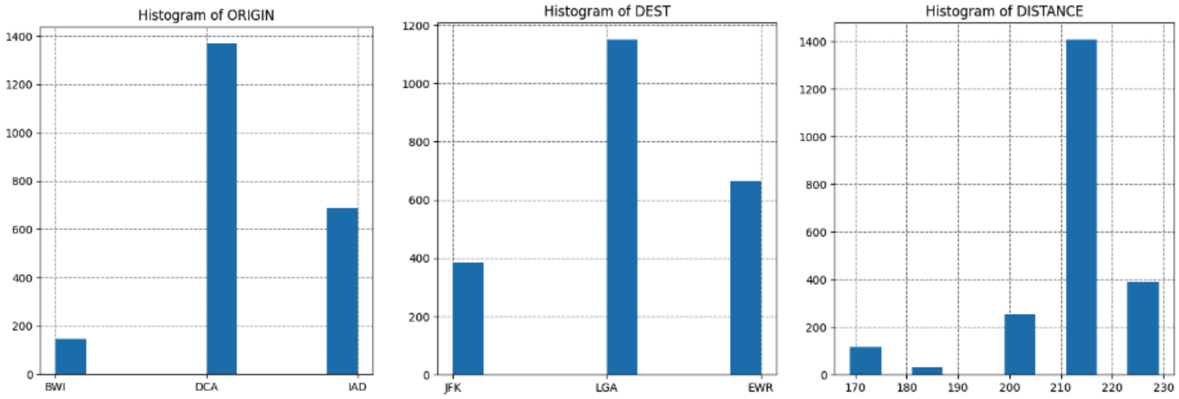


Figure 2. Distribution of Flight Origin, Destination and Distance Travelled

Moreover, the distribution of flight days appears to be relatively uniform, with peaks occurring on days 4 and 5, corresponding to Thursday and Friday. A closer examination of flight dates reveals distinct patterns in flight frequency throughout the month. Flights appear to peak during the early and late stages of the month, between dates 4 and 10, as well as 28 and 30.

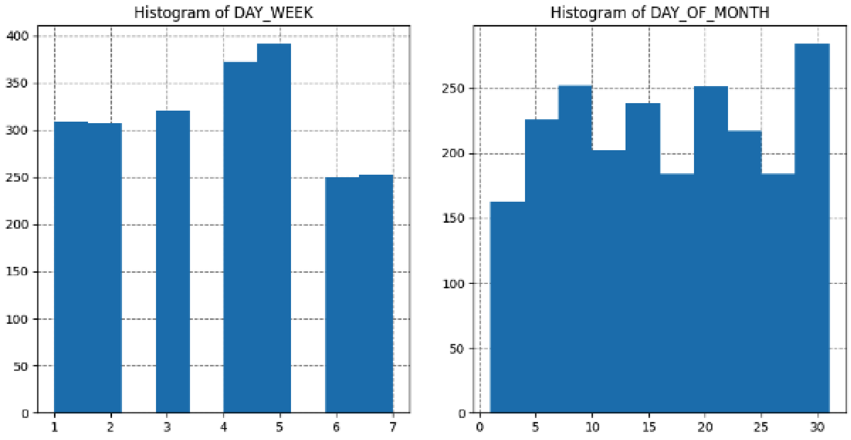


Figure 3. Distribution of Flights Based on Day in Week and Month

When examining flight status by carrier in figure 4, it was observed that carrier DH had the highest number of flights flown, followed by RU and US, respectively. The delayed proportion of flights appeared to closely align with the ratio of flights flown by each carrier, with carrier DH experiencing the highest number of delayed flights, followed by RU. This correlation suggests that carriers with a higher volume of flights may encounter a proportionate number of delays.

However, an interesting finding emerged when analyzing the delayed flights relative to the total number of flights flown by each carrier. Despite ranking 5th in terms of flights flown, carrier MQ unexpectedly ranked third in terms of delayed flights. This discrepancy highlights potential operational challenges or inefficiencies within carrier MQ's flight operations.

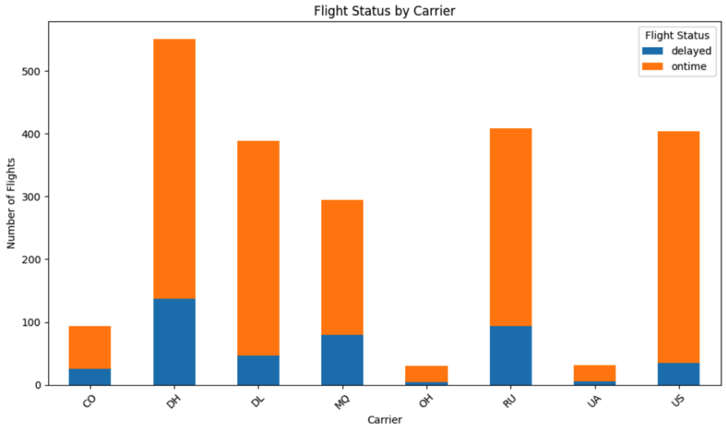


Figure 4 Ratio of Delayed and Ontime Flight by Carrier

In the final observation, Figure 5 indicates that, on average, flights experience a delay of approximately 61.41 minutes before taking off. To ensure a more accurate distribution, we've excluded any flights that departed more than 200 minutes earlier than their scheduled departure time yet still arrived late at their destinations due to their rarity, which could skew the data. Conversely, Figure 6 illustrates flights that arrived punctually. On average, these flights depart approximately 15.71 minutes earlier than their scheduled departure time. Most carriers depart around 50 minutes prior to the scheduled time, followed by departing 5 minutes earlier or later than the scheduled time.

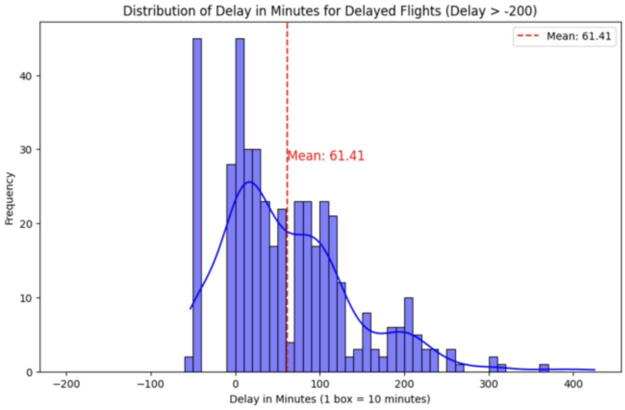


Figure 5. Distribution of Delay in Minutes for Flight Status = Delayed

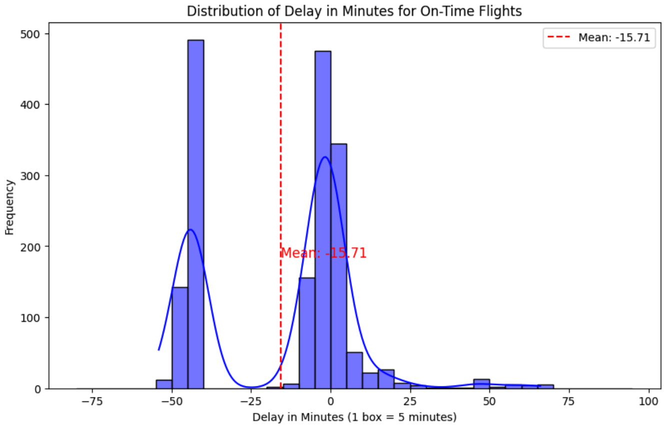


Figure 6. Distribution of Delay in Minutes for Flight Status = Ontime

Data Preparation and Model Setup:

Before implementing the machine learning algorithms, the cleaned dataset underwent preprocessing to ensure optimal performance. The dataset was split into training and testing sets using a rule of thumb of pareto principle (Joseph, V.R., 2022), 80:20 ratio, where 80% of the data was allocated for training the models and 20% for testing their performance.

As part of the preprocessing steps, the target variable, "FLIGHT_STATUS," was separated from the predictor variables and assigned to the variable "y," representing the outcome to be predicted. Subsequently, the "flight status" column was dropped from the dataset to prevent it from influencing the model predictions.

K-Nearest Neighbour (k-NN) Analysis

The K-Nearest Neighbor (k-NN) algorithm was the initial foray into machine learning. It works by identifying the k-nearest data points to a given point and classifying it based on the most prevalent class among those neighbors. In the analysis conducted, the accuracy of the algorithm was observed to reach its peak at k = 3, achieving a classification accuracy of 87.98%. Following closely behind was the accuracy at k = 6, with a slightly lower accuracy of 87.3%.

This implies that when considering the three nearest neighbors, the algorithm was most successful in accurately classifying the data points. However, the slight decrease in accuracy with k = 6 suggests that incorporating a few additional neighbors may introduce some noise or variability into the classification process, resulting in a slightly lower accuracy rate.

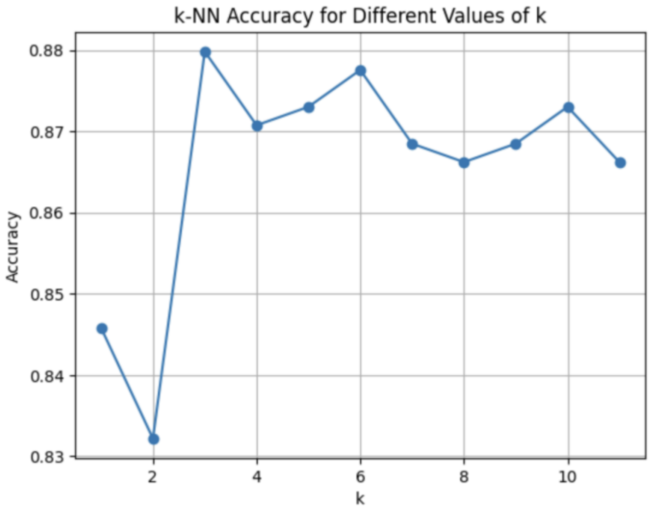
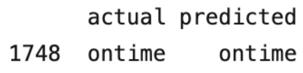


Figure 7. K-Nearest Neighbor Accuracy Plot

Naive Bayes Classifier Analysis

The second algorithm tested was the Naive Bayes Classifier. As its name suggests, the Naive Bayes approach treats each predictor independently, hence the "naive" designation. One of the initial steps in applying this algorithm is to segment the departure time data. This segmentation process reduces the 2201 unique values of departure times to a smaller set of categories based on rounding. For example, consider a flight scheduled to depart at 14:55. Instead of representing it as a unique value, it would be categorized into the bin representing 15. Similarly, a flight departing at 16:40 would fall into the bin representing 16. By categorizing departure times in this way, the data becomes simplified, allowing the Naive Bayes Classifier to focus on broader patterns within these categorized time bins.

After determining the outcome and predictors, the Naive Bayes Classifier was executed based on the specified criteria required by the predictors. In an attempt to test if the model was working properly, a test was conducted to predict the outcome for a carrier (DL) departing on Sunday within the 10AM range, with a destination of LGA airport and originating from DCA airport. Remarkably, the prediction accurately indicated an on-time departure.



```
     actual predicted
1748 ontime    ontime
```

Figure 8. Example of the Prediction in test case

A confusion matrix (Figure 9) was constructed to gain insights into the accuracy, precision, sensitivity, and specificity of the test model. According to the confusion matrix analysis, the test model demonstrates the following metrics:

- Accuracy: 78.2% (percentage of correct predictions out of the total sample)
- Precision: 79.17% (percentage of correctly predicted delayed flights out of all predicted delayed flights)
- Recall (Sensitivity): 98.56% (percentage of correctly predicted delayed flights out of all actual delayed flights)
- Specificity: 15.5% (percentage of correctly predicted on-time flights out of all actual on-time flights)

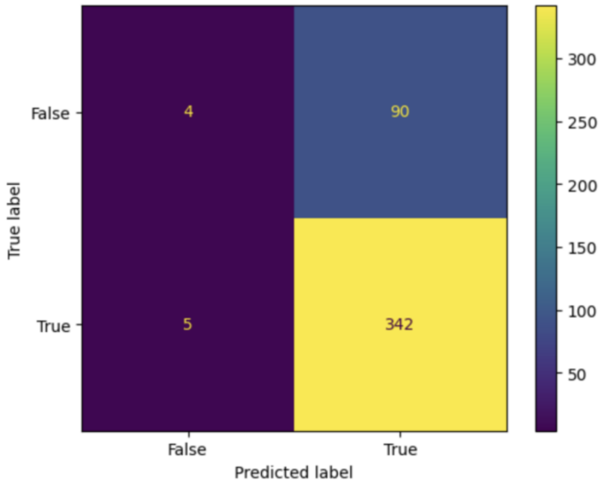


Figure 9. Confusion Matrix of Naive Bayes Test Model

Random Forest Analysis

In contrast to Naive Bayes, which treats each variable in isolation, Random Forest employs a sophisticated branching mechanism. This mechanism selects the lowest GINI value from each variable to determine the optimal split, constructing multiple decision trees simultaneously. In our analysis where identical predictors and outcomes were utilized, the Random Forest model demonstrated a remarkable accuracy of 91.84%. This notable improvement is attributed to its superior performance in accurately predicting delayed flights, as illustrated in Figure 10. Conversely, Naive Bayes often misclassified a significant portion of actual delayed flights as on-time, highlighting the comparative effectiveness of Random Forest in this context.

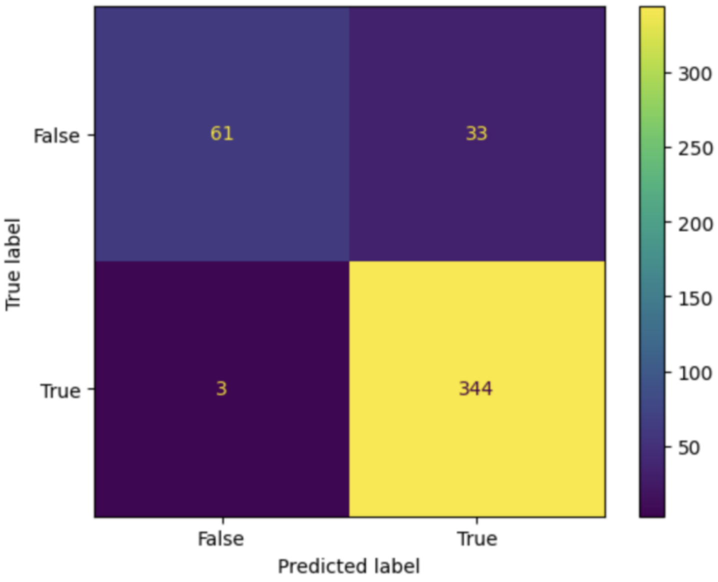


Figure 10. Confusion Matrix of Random Forest Test Model

Results and Conclusions

In our analysis, we examined three machine learning algorithms—K-Nearest Neighbor (k-NN), Naive Bayes, and Random Forest—for their effectiveness in predicting flight delays. While k-NN and Naive Bayes demonstrated reasonable accuracy, Random Forest emerged as the most effective, achieving an impressive 91.84% accuracy rate.

The superior performance of the Random Forest model can be attributed to its ability to analyze interactions between variables and optimize predictive accuracy by making optimal splits in the data. As the most sophisticated among the three models evaluated, Random Forest effectively addresses the limitations observed in k-NN and Naive Bayes, particularly in accurately forecasting flight delays, thus establishing its suitability for this dataset.

However, it's essential to acknowledge that the complexity of the Random Forest model also introduces the risk of overfitting. To mitigate this risk, careful consideration and appropriate regularization techniques are necessary during model training.

Furthermore, it's important to recognize that the ratio between training and test data can impact the results of the confusion matrix. Simply increasing the ratio of training to test data (e.g., from 90:10 to 60:40) does not guarantee more accurate results. Instead, it's crucial to ensure that the training set is both representative and sufficiently large to yield meaningful results without causing the model to overfit. Balancing these factors is essential for optimizing model performance and generalization to unseen data.

These findings carry significant implications for the aviation industry. By harnessing advanced machine learning techniques, airlines can refine their scheduling processes, reduce costs, and enhance customer satisfaction. The adoption of predictive models, notably Random Forest, holds the potential to optimize operational efficiency and deliver a smoother travel experience for passengers.

In conclusion, our analysis underscores the importance of proactive strategies in managing flight delays. Through the adoption of data-driven approaches, airlines can effectively mitigate disruptions, allocate resources more effectively, and ultimately elevate the quality of service provided to customers.

References

Bureau of Transportation Statistics. (n.d.). *Airline On-time Statistics and Delay Causes.* Retrieved from https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E

Guy, A.B. (2010). *Flight delays cost $32.9 billion, passengers foot half the bill.* Berkeley News. Retrieved from https://news.berkeley.edu/2010/10/18/flight_delays

Joseph, V.R. (2022). *Optimal ratio for data splitting*. Wiley Online Library. Retrieved from https://onlinelibrary.wiley.com/doi/10.1002/sam.11583