ANALYZING OPERATIONAL EFFICIENCY AND PERFORMANCE METRICS OF THE USA'S AVIATION INDUSTRY: A CASE STUDY OF AMERICAN AIRLINES

By Emanuele Bossi

Data Science II: Foundations

College of Arts and Sciences

Department of Data Science

Professor: Dr. Das

29th April 2024

Table of Contents

Abstract

This research project utilizes Data Analysis and Machine Learning techniques to address the impacts of flight delays in the Aviation sector.

Analyzing extensive datasets, including flight schedules and delay records from the United States Department of Transportation (USDOT), the study identifies patterns and trends associated with delays and evaluates their financial ramifications. Various machine learning models such as Random Forest, Logistic Regression, K-Nearest Neighbor (KNN), and Extreme Gradient Boosting (XGBoost) are employed to predict and prevent flight delays effectively. A specific case study on American Airlines flights originating from Dallas Fort-Worth airport provides insights into airline-specific delay mitigation strategies.

The findings indicate substantial potential for cost savings for both airlines and passengers through proactive delay management strategies.

Introduction

Air transportation plays a vital role in driving global economic development by facilitating the exchange of goods, people, capital, technology, and ideas between countries and cities. The COVID-19 pandemic severely disrupted this connectivity in 2020, resulting in a significant decline in the number of city pair routes and routes eliminated. While there has been a gradual recovery since then, with travel restrictions being lifted unevenly across the globe, the restoration of service frequency and capacity lags behind the reestablishment of city-pair connections. *IATA*'s connectivity index indicates that international connectivity has rebounded to 89.5% of pre-pandemic levels, with domestic connectivity reaching 97.5%, marking a 12-percentage point increase compared to 2022 (“Global Outlook for Air Transport a Local Sweet Spot”).

Based on the annual economic report published by *Statista*, the estimated market size of the global airline industry in 2023 has been USD 841.4 billion, with operating profits of USD 41 billion. Passengers worldwide have unequivocally demonstrated the significance of air transportation by consistently choosing to utilize it, even amid periods of exceptionally high jet fuel prices relative to crude oil prices. According to a survey of 8,000 air travelers, passenger satisfaction stands at an impressive 82%, with 91% expressing the indispensability of air travel. This underscores the vital role air transportation plays in connecting our global community (“The Statistics Portal”).

Despite advancements in technology and meticulous scheduling, flight delays remain an enduring challenge for both airlines and passengers alike. Flight delays present a persistent and costly issue for airlines, airports, and travelers. In the United States alone, airlines face approximately $8.3 billion in financial losses due to delays, while passengers bear the brunt with losses totaling around $16.7 billion, equating to roughly $51 per person (Anupkumar).

The consequential impacts of delays extend beyond inconvenience, affecting operational costs, customer satisfaction, and overall industry performance. Consequently, there exists a pressing need for robust predictive models that can anticipate flight delays and empower airlines to proactively manage their operations.

This research paper focuses on the specific case study of American Airlines, one of the leading carriers in the United States aviation industry. Through the lens of data science methodologies, this study endeavors to dissect and analyze various factors influencing flight delays within the American Airlines network. By leveraging comprehensive datasets encompassing historical flight information, seasonal patterns, airport congestion, and other relevant variables, our aim is to construct predictive models capable of forecasting flight delays with a high degree of accuracy.

The significance of this research extends beyond its academic curiosity; it holds practical implications for airline operations, strategic planning, and passenger experience enhancement. By identifying key drivers of flight delays and developing predictive frameworks, airlines can implement targeted interventions to mitigate disruptions, optimize resource allocation, and enhance overall flight performance. Moreover, a deeper understanding of the complex interplay between different variables influencing flight delays can inform policy decisions and industry regulations aimed at fostering a more efficient and resilient aviation ecosystem.

Methodology

For this project, data sourced from the United States Department of Transportation (USDOT) via the United States Bureau of Transportation Statistics (BTS) was utilized. This dataset encompasses monthly flight information spanning from 1987 to the present day. Specifically, it comprises data on flights operated by U.S. certified air carriers representing at least one percent of domestic scheduled passenger revenues. The analysis focused on a subset of 6,847,899 flights, covering the period from January 1st, 2023, to December 31st, 2023.

Date	Month, Day of Month, Day of Week
Carrier	Name, Flight Number
Origin	Airport, State, City
Destination	Airport, State, City
Time	Departure/Arrival Scheduled/Actual, Air Time
Cancellation	Reason (if applicable)
Delay	Reason (if applicable)

Table 1: Major Information Collected

The database contains information on 15 carriers, listed below in descending order based on the number of domestic flights operated in 2023: Southwest Airlines (WN), Delta Airlines (DL), American Airlines (AA), United Airlines (UA), SkyWest Airlines (OO), Republic Airways (YX), JetBlue Airways (B6), Spirit Airlines (NK), Alaska Airlines (AS), Envoy Air (MQ), Endeavor Air (9E), PSA Airlines (OH), Frontier Airlines (F9), Allegiant Air (G4), and Hawaiian Airlines (HA).

Total Flights	6,847,899
Total Days	365
Total Carriers	15
Total Different Flights (by FL Number)	6,358
Total Origin Airports	350
Total Destination Airports	350
Total Flights Cancelled	87,943
Total Flights Delayed	2,472,530

Table 2: Data Summary

Following the Feature Engineering stage, during which certain variables were modified to suit the analysis objectives, an exhaustive Exploration Data Analysis (EDA) was conducted to gain deeper insights into the dataset and uncover potential patterns or significant insights.

A variety of Machine Learning techniques were employed to forecast flight delays, focusing specifically on American Airlines flights. Feature importance analysis was carried out using the Random Forest algorithm, while Random Forest, Logistic Regression, K-Nearest Neighbors, and XGBoost algorithms were utilized for predictive modeling. In the final phase of the project, to enhance the predictive capabilities, only data from flights operated by American Airlines departing from its central hub of Dallas Forth-Worth were considered for building the XGBoost model.

Multiple evaluation metrics, as detailed in the dedicated section, were utilized to determine the most effective Machine Learning algorithm for this project.

Results

Exploratory Data Analysis

The Exploratory Data Analysis (EDA) stage, which constitutes a pivotal aspect of this project, comprised various subparts designed to facilitate a thorough comprehension of the database and uncover potential patterns within the data.

1. *Dataset Overview*

   A foundational aspect of a well-executed Data Science project is gaining a concise understanding of the data distribution within the dataset. In this study, we conducted an analysis of General Performances (Figure 1), Reasons for Cancellation (Figure 3), and Reasons for Delays (Figure 4).

   As depicted in the graph below, approximately 60% of the total number of domestic flights operated in 2023 were on-time, while the remaining 40% experienced delays (around 35%) or cancellations (5%).

   Notably, even though the number of delayed flights is substantial, it's important to note a distinction: our primary objective is to minimize financial losses from delays and enhance passenger satisfaction, thus, we must consider that minor delays (i.e., delays under 15 minutes), which constitute the largest portion of delays as illustrated in Figure 3, do not significantly impact our analysis. Consequently, in our Machine Learning modeling, we will classify a flight as delayed only if the measured arrival delay is equal to or exceeds 15 minutes.

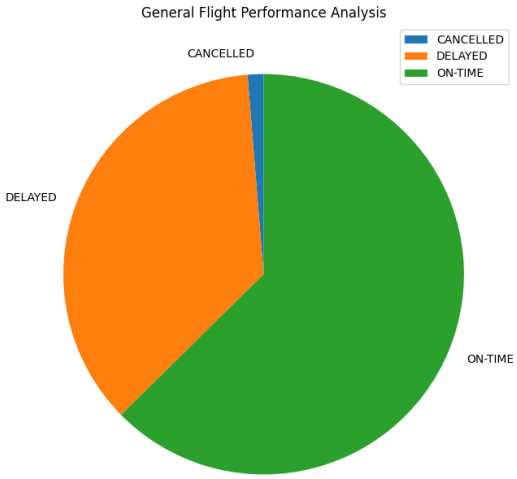

Figure 1: General Flight Performance Analysis

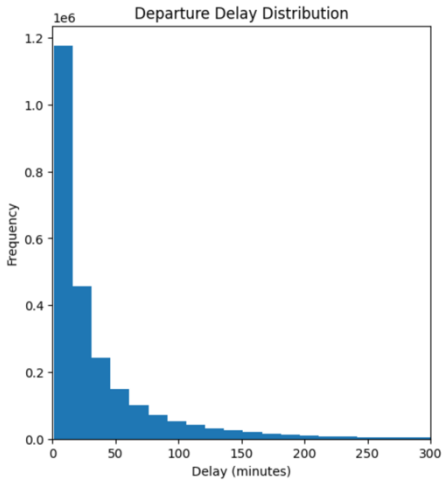

Figure 2: Departure Delay Distribution

From the Reasons of Cancellation plot and the Reasons of Delays plot, it's evident that the factors contributing to cancellations differ from those causing delays. Weather conditions emerge as the primary reason for cancellations, whereas they play a marginal role in delays. Therefore, it can be inferred that flights are much more likely to be cancelled during adverse weather conditions rather than just delayed.

Conversely, carriers' actions contribute to cancellations and delays to a similar extent (around 30-35%). Since our project operates under the assumption that we are working for the Technical Operations department of an airline, where we have direct control only over carrier decisions (while decisions based on weather forecasts would constitute an indirect form of control), our objective will be to minimize both cancellations and delays attributable to carrier-related factors.

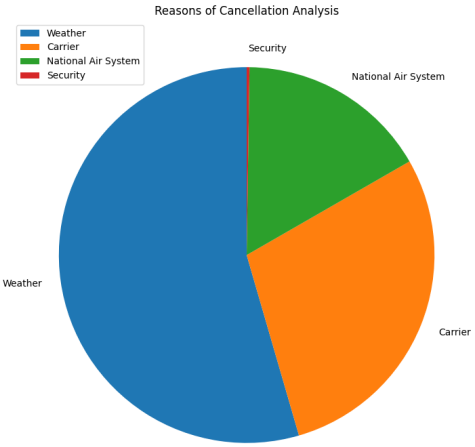

Figure 3: Reasons of Cancellation Analysis

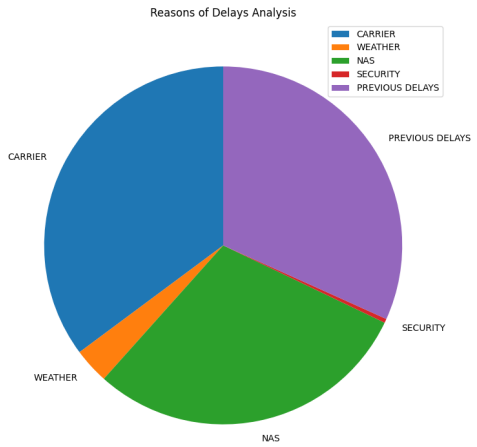

Figure 4: Reasons of Delays Analysis

2. *Time Analysis*

An outstanding aspect of Data Analysis is its ability to elucidate real-life phenomena through data. In the following time-series analysis, distinct maximum and minimum peaks can be discerned, each attributed to specific causes:

A. *Thanksgiving*: A national holiday during which individuals typically stay at home with their families, leading to a reduction in the number of flights on that day.

B. *Sunday after Thanksgiving*: Following several days spent at home for Thanksgiving, often enjoyed as a "long weekend," individuals return to their respective workplaces, resulting in a heightened demand for flights.

C. *2023 Winter Storm*: A significant winter storm, characterized by high accumulations of snow, swept across the USA, causing a substantial number of flight delays (National Weather Service).

D. *2023 FAA System Outage*: U.S. flights experienced a great number of delays on January 11, 2023, as the Federal Aviation Administration (FAA) endeavored to rectify a system outage (Josephs).

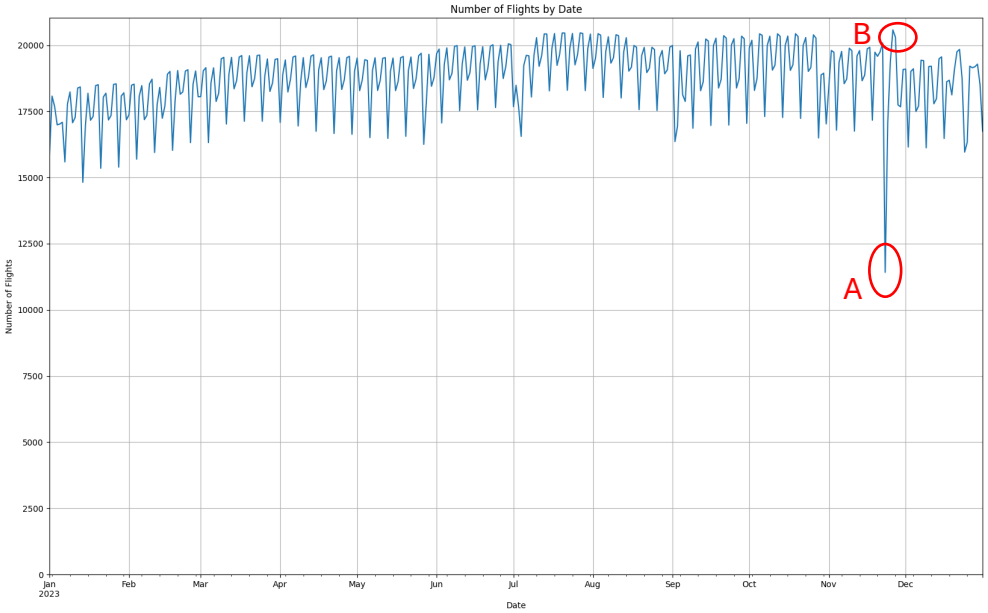

Figure 5: Number of Flights by Date

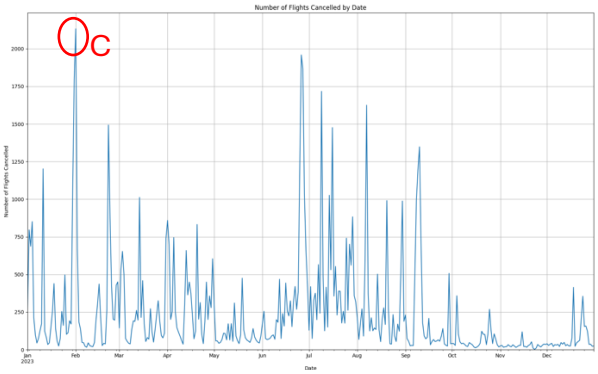

Figure 6: Number of Flights Cancelled by Date

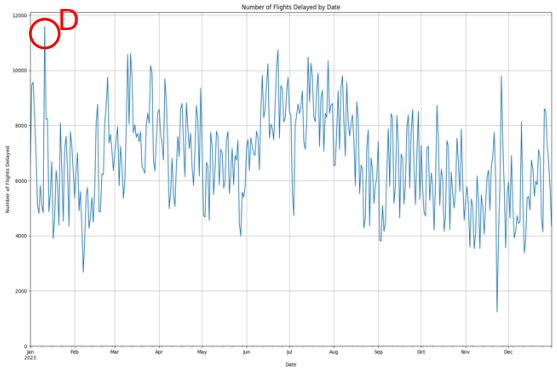

Figure 7: Number of Flights Delayed by Date

Another notable and significant pattern worth highlighting is the observation of distinct seasonality patterns evident in the plot depicting the Number of Flights Delayed by Date. Specifically, notable peaks can be observed in the following periods:

- March: A surge in the number of students and families can be observed as they flock to airports across the country during Spring Breaks.

- July-August: US airports experience a significant influx of American and international tourists during the summer vacation season, leading to a notable increase in the number of delays.
- December-January: The Christmas holidays prompt a substantial number of individuals to travel by air for vacations, resulting in heightened airport activity and a subsequent increase in delays.

It's crucial to emphasize that delays aren't solely attributed to the volume of domestic flights depicted in Figure 5, but both airports and international airlines, including the target airline for this research, American Airlines, are impacted by international flights. The number of passengers transported notably increases during vacation periods, affecting the overall operational dynamics and contributing to delays across the board.

3. *Number of Flights Analysis*

The examination of flight frequency yielded insightful findings regarding various factors including the days of the week, leading carriers, and prominent airports.

Figure 8 illustrates that Monday and Friday emerge as the peak days of flight activity, suggesting a trend where individuals likely commute for weekend getaways (Friday) or return to their workplaces at the start of the week (Monday). Conversely, Saturday registers as the least active day, potentially owing to reduced demand for work-related travel.

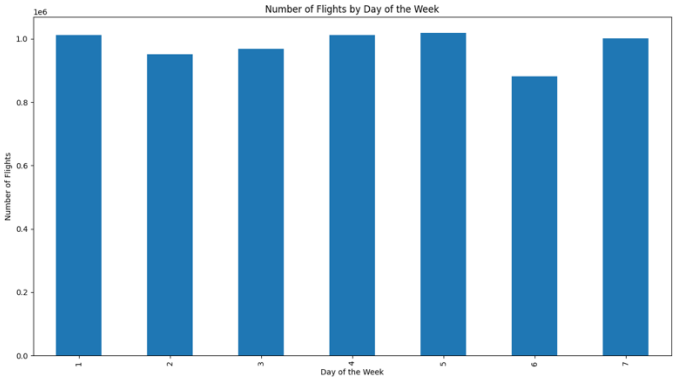

Figure 8: Number of Flights by Day of the Week

Figure 9 showcases the top three carriers in terms of domestic flight volume for the year 2023: Southwest Airlines, Delta Airlines, and American Airlines. The substantial margin between the leading carrier and the subsequent ones can be attributed to Southwest Airlines' exclusive focus on domestic operations, while Delta and American Airlines extend their services to international routes. Notably, American Airlines boasted the highest passenger enplanement figures in 2023, totaling approximately 210,692,000, compared to Southwest Airlines' 171,817,000 ("List of Largest Airlines in North America").

Upon scrutinizing Figure 9, a distinct pattern emerges delineating two clusters of carriers based on flight volume: the "big" airlines, encompassing WN, DL, AA, UA, and OO, and the "small" airlines, comprising YX, B6, NK, AS, MQ, 9E, OH, F9, G4, and HA. Notably, within the cohort of "big" airlines, all except SkyWest Airways (OO) operate on an international scale and hold positions among the top 12 worldwide airlines by market capitalization. Specifically, Delta Airlines holds the second position, followed by Southwest Airlines (3rd), United Airlines (4th), and American Airlines (12th).

An intriguing observation regarding the largest airlines by market capitalization is the positioning of Alaska Airlines. Despite ranking 9th in terms of the number of domestic flights performed in 2023, Alaska Airlines secures the 21st

position, surpassing renowned carriers such as Canada Air and Air France-KLM ("Largest airlines by market cap").

However, it's crucial to acknowledge that SkyWest Airlines, the largest regional airline in North America, operates in collaboration with 4 major carriers, such as American Airlines, Delta Airlines, United Airlines, and Alaska Airlines. Through these partnerships, SkyWest assumes responsibility for operating and maintaining aircraft utilized on flights scheduled, marketed, and sold by its four mainline partner airlines ("SkyWest Airlines").

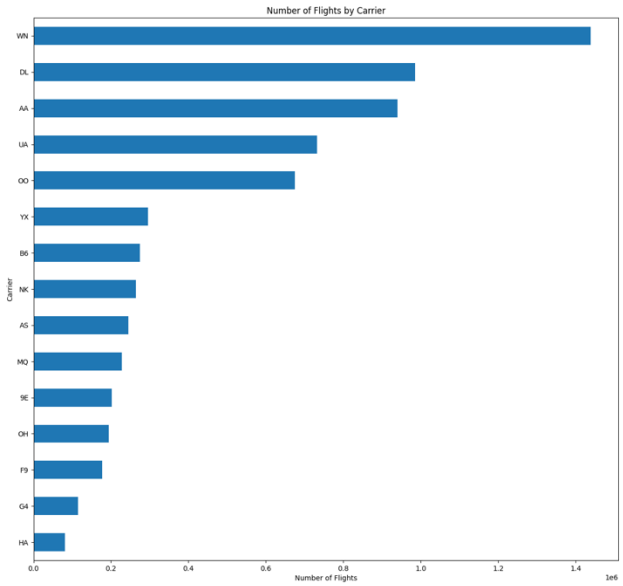

Figure 9: Number of Flights by Carrier

Figure 10 presents the leading 20 airports by domestic flight departures. Atlanta airport (ATL) secures the top rank, followed by Denver (DEN) and Dallas Fort-Worth (DFW). It is noteworthy that these airports serve as primary hubs for the respective top three carriers, with Denver being pivotal for Southwest Airlines, Atlanta for Delta Airlines, and Dallas Fort-Worth for American Airlines.

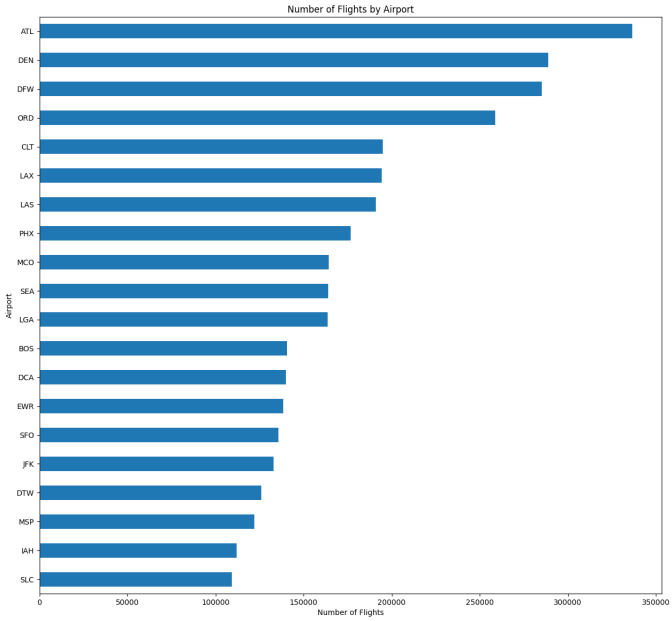

Figure 10: Number of Flights by Airport

4. *Cancellation Analysis*

In this section dedicated to the analysis of flight cancellations, our focus was directed towards assessing the cancellation ratio rather than simply the sheer number of cancelled flights across different categories.

Examining the cancellation ratio by month (Figure 11), a conspicuous pattern emerges, indicating a pronounced peak during the summer months of June, July, and August. This surge in cancellations can likely be attributed to the heightened demand for air travel during the peak vacation season. The resulting congestion at airports coupled with the strain on airlines' schedules contributes to the elevated cancellation ratio observed during this period. Furthermore, notable peaks in cancellation ratios are also evident during the winter months of January, February, and March. This trend aligns with our earlier observation (as depicted in Figure 3) highlighting adverse weather conditions as a primary cause for flight cancellations. While the precise factors contributing to these trends warrant further investigation within the aviation

industry, it is apparent that seasonal fluctuations and weather-related challenges play significant roles in shaping the cancellation landscape.

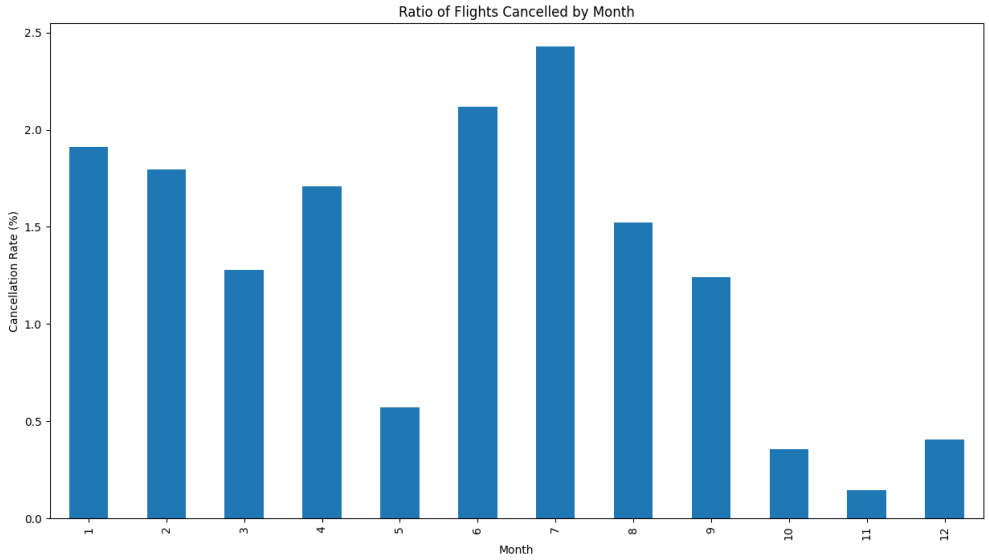

Figure 10: Ratio of Flights Cancelled by Month

Unlike the observed trend regarding months, where high traffic correlates with high cancellation ratios as depicted in Figure 11, an interesting deviation emerges when analyzing cancellations by day of the week. Surprisingly, Wednesday stands out as the day with the highest number of flight cancellations. Conversely, Thursday and Friday, which coincide with peak traffic days, exhibit fewer cancellations, along with Saturday, which experiences relatively lower flight activity.

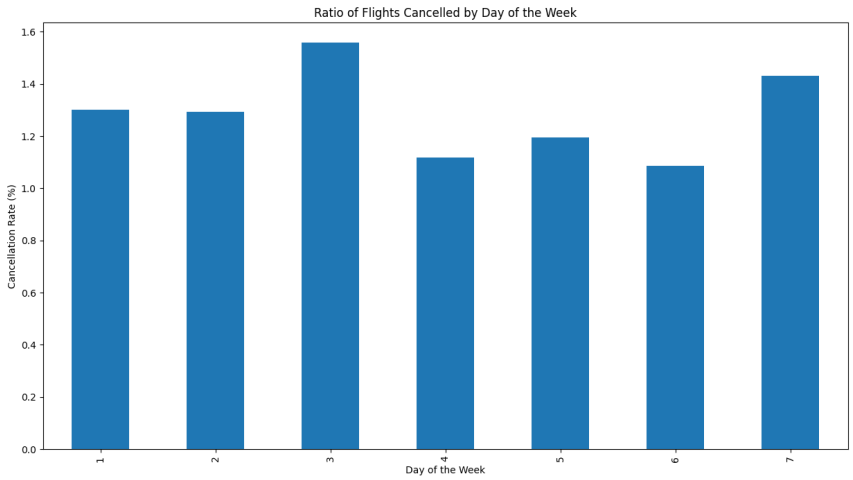

Figure 10: Ratio of Flights Cancelled by Day of the Week

The graph depicted in Figure 11, showcasing the Ratio of Flights Cancelled by Carrier, offers a compelling metric for evaluating the performance of individual carriers in domestic flight operations. Notably, Southwest Airlines, Delta Airlines, and American Airlines, occupying the top three positions in terms of total number of domestic flights in 2023, also emerge among the top six carriers with the lowest cancellation ratios, securing the 3rd (WN), 5th (DL), and 6th (AA) positions, respectively.

Conversely, certain carriers with comparatively lower flight volumes in 2023, such as Republic Airways, Frontier Airlines, and Endeavor Air, exhibit higher cancellation ratios, reflecting the potential impact of operational scale on cancellation performance within the aviation industry.

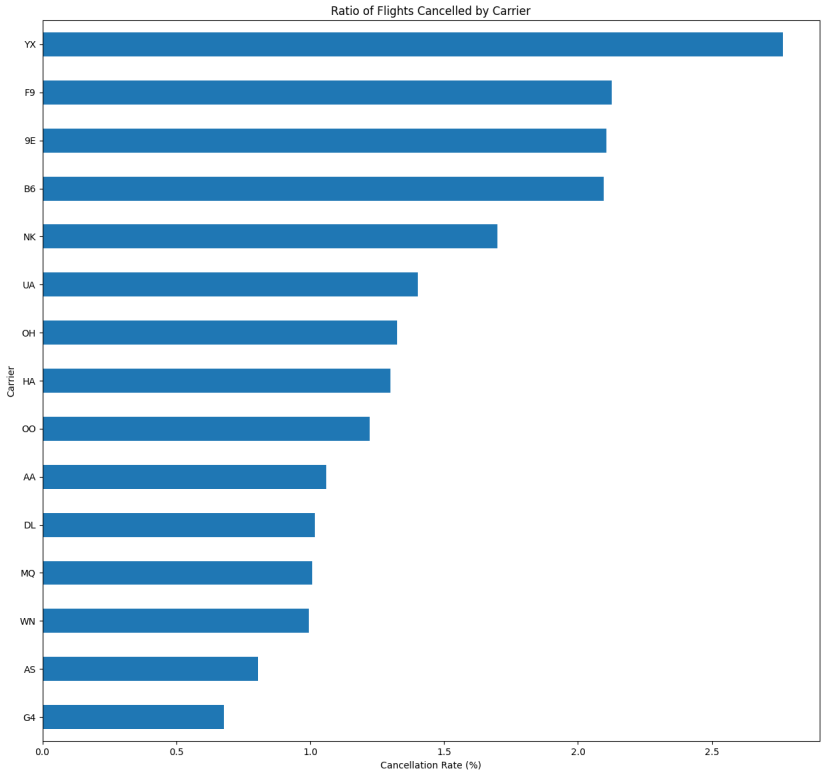

Figure 11: Ratio of Flights Cancelled by Carrier

5. *Delay Analysis*

Another crucial metric for evaluating the performance of various components within the aviation industry, such as airports and airlines, is delay analysis. It's essential to establish a clear definition of "delay" within the scope of this research:

"A flight is considered delayed if its actual arrival time exceeds the scheduled arrival time."

Def. 1: Arrival Delay Definition

In Figure 12, we examined the relationship between the number of flights originating from each airport and the delay factor. The scatter plots on the left suggest a predictable linear relationship between the number of flights operated and the total number of flights experiencing delays. However, akin to the cancellation analysis, what truly holds significance for our analysis is the ratio of delays, as illustrated in the scatter plot on the right.

As depicted, there is an evident increase in the ratio of delayed flights as the volume of flights operated escalates. Notably, this linear relationship holds true until approximately 200,000 flights operated. Interestingly, the top four airports in terms of flight volume, namely Atlanta, Denver, Dallas Fort-Worth, and Chicago O'Hare, do not exhibit a corresponding increase in the delay ratio. It's noteworthy that the mean ratio of delays remains relatively stable for airports handling more than 150,000 flights. However, smaller airports with fewer than 50,000 flights operated demonstrate substantially lower delay ratios, as indicated by the median line (purple line). This observation underscores the influence of airport size and operational scale on delay management.

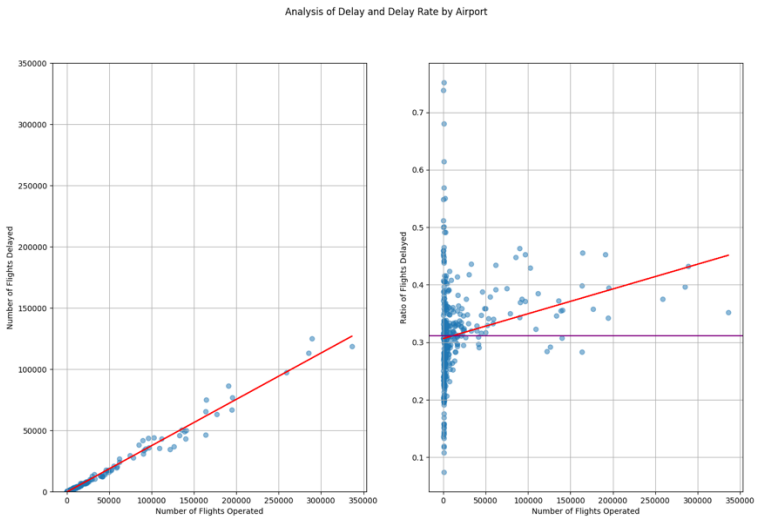

Figure 12: Analysis of Delay and Delay Rate by Airport

It's important to emphasize that Definition 1 still lacks precision for the specific objectives of our analysis. As outlined at the inception of this project, our overarching aim is to enhance both the economic performance and passenger satisfaction within the aviation industry. While minor delays can indeed present challenges, our primary focus is on mitigating substantial delays that significantly impact operations and customer experiences.

As depicted in Figure 13, our previous definition of delay, which considers a flight as delayed even with just a 1-minute deviation from the scheduled arrival time, encompasses a vast number of occurrences. However, these slight delays, while prevalent, hold lesser significance for our analytical purposes.

In light of this, we have introduced a refined definition of a delayed flight tailored specifically for data analysis and machine learning applications within the context of this project:

"A flight is considered delayed if its actual arrival time exceeds the scheduled arrival time by at least 15 minutes."

Def. 2: Arrival Delay Constraint Definition

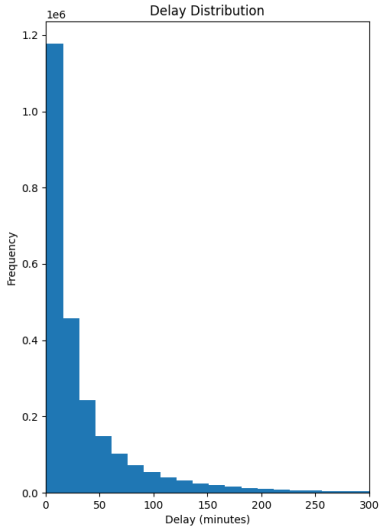

Figure 13: Delay Distribution

Based on the refined definition provided earlier, Figure 14 illustrates the distribution of flights delayed by carriers, with a clear differentiation between flights delayed by more than 15 minutes (highlighted in red) and the total number of delayed flights. Additionally, it's important to note that our analysis of carrier delays isolates instances where delays are attributable to the carriers themselves, thereby excluding external factors beyond their control. By doing so, we aim to assess the performance of airlines independent of external influences. For instance, Alaska Airlines, operating primarily in Alaska, might be disproportionately affected by adverse weather conditions, potentially inflating its delay ratio even if the delays are not directly attributable to the airline's operations.

An intriguing observation from the graph is that while YX emerges as the carrier with the highest ratio of flight cancellations, it also exhibits the lowest ratio of delays attributable to the carrier's fault.

The top airlines by the number of flights operated display comparable performance levels. However, notable discrepancies are observed for airlines such as F9, HA, and B6, which demonstrate poorer on-time performance metrics compared to

their counterparts, evidenced by both higher delay ratios and cancellation rates. This suggests a need for further examination into the operational practices and efficiency of these airlines to address performance shortcomings and enhance overall service reliability.

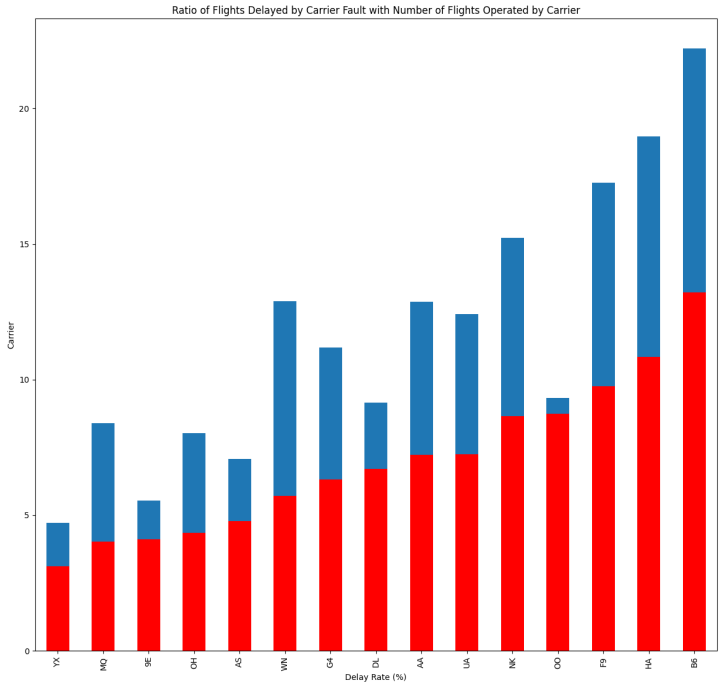

Figure 14: Delay Distribution

6. *Routes Analysis*

Another pivotal factor influencing flight delay performances is the volume of flights operated along specific routes. It stands to reason that a higher frequency of flights within a given route would lead to improved on-time performance, as all operational aspects, including check-in procedures, gate operations, air traffic control (ATC) coordination, and pilot schedules, are likely to be better managed. This hypothesis finds support in Figure 15, depicting the Ratio of Flight Delayed by Route.

The graph reveals a discernible negative linear correlation, indicating that as the number of flights operated on a route increases, the ratio of flight delays tends to decrease. This trend holds true with considerable accuracy up to a threshold of

approximately 3000 flights operated. Beyond this point, the ratio stabilizes, suggesting a plateau in the relationship between flight frequency and delay performance.

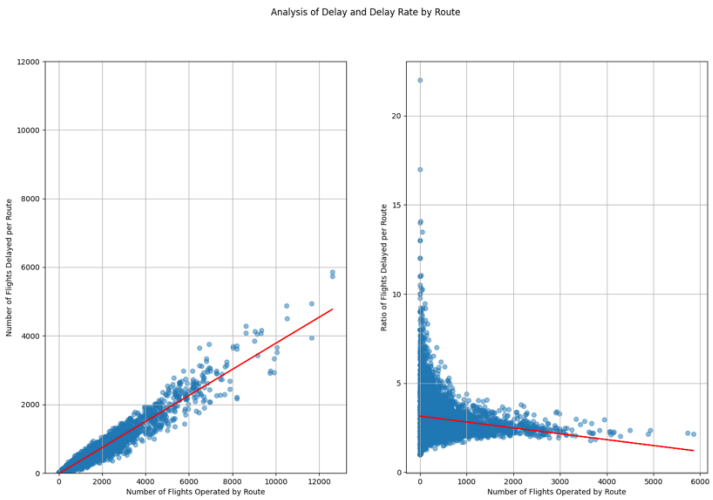

Figure 15: Analysis of Delay and Delay Rate by Route

Machine Learning

Introduction

The Machine Learning segment of this project stands out as its most crucial component. Predicting flight delays holds immense significance in shaping the future trajectory of the aviation industry, with the potential to drive substantial financial gains and enhance passenger satisfaction.

As highlighted in the introduction, the Machine Learning phase is centered on a case study of American Airlines, aiming to forecast delays for flights operated by this specific carrier. It's essential to clarify that the methodologies, techniques, and insights garnered from this study can be readily applied to analyze the performance of any airline.

While this research delves into the specifics of American Airlines' operations, it's imperative to recognize the broader implications and applicability of the findings across the aviation sector. American Airlines' prominence in the industry, reflected by its status as the

largest airline in the world by scheduled passengers carried, revenue passenger mile, and fleet size, underscores the relevance and significance of this focused investigation (“American Airlines”).

General Considerations

For this analysis, the dataset has been filtered to exclusively include American Airlines' scheduled flights, encompassing operated, cancelled, and delayed flights throughout the entire 2023. The predictive features utilized in this endeavor comprise the following:

- DAY: count of the number of day since the inception of the year;
- DEP_MIN: minute of scheduled departure since the commencement of the day;
- DAY_OF_WEEK: categorical variable spanning from 1 (Monday) to 7 (Sunday);
- ORIGIN_POP: number of flights scheduled in 2023 by the departure airport;
- DEST_POP: number of flights scheduled in 2023 by the destination airport.

To evaluate the efficacy of each predictive model, the dataset has been evenly divided into training and testing sets, with each accounting for 50% of the data. Additionally, a random seed of 0 has been set to ensure the reproducibility of the experiments, facilitating consistent and reliable results across iterations. Several Evaluation Metrics (Appendix A) has been adopted to assess model efficiency on the testing set*.

Following an examination of class representation within the dataset (Figure 16) and validating the Random Forest model's performance with the "pure" dataset (Table 1), it became imperative to tackle the imbalanced class issue using resampling techniques. The class distribution is notably skewed, with observations of class 0 outnumbering class 1

*All the models have been tested on the testing dataset. By definition, it is a portion of the original dataset reserved for evaluating the performance of a machine learning model after it has been trained on the training dataset, helping to assess the model's ability to generalize to new, *unseen* data .

observations by approximately 5 times. As depicted in Table 1, this imbalance poses a significant challenge, potentially resulting in subpar performance in identifying the positive class (i.e., delayed flights), with a resulting very low Recall score.

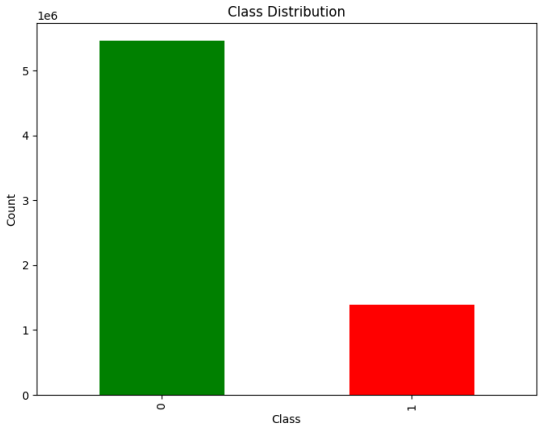

Figure 16: Class Distribution

Random Forest Evaluation Metrics	
Accuracy	77.2%
Recall	21.4%

Table 1: Initial Random Forest Evaluation Metrics

After implementing undersampling, a technique that randomly removes instances from the majority class until class distribution is balanced, with the *RandomUnderSampler* function from the *imblearn.under_sampling* library, and oversampling using the *SMOTE* function from the *imblearn.over_sampling* library, which generates synthetic samples for the minority class based on its nearest neighbors in the feature space, we evaluated the performance of each technique (see Table 2).

Resampling Techniques Evaluation Using RF		
	Accuracy	*Recall*
Undersampling	62.8%	61.9%
Oversampling	70.1%	43.3%

Table 2: Resampling Techniques Evaluation Using Random Forest

Given the better performances results highlighted above, we applied undersampling to the training dataset to ensure equal representation of both positive and negative observations. A possible reason of the outperformance of undersampling on oversampling is that oversampling techniques, like SMOTE, generate synthetic samples to increase the minority class representation that may not accurately represent the true underlying distribution of the minority class, leading to potential model bias or noise.

Assessing Feature Importance

The initial step in the Machine Learning process involved assessing the importance of features through training a Random Forest model. This analysis is pivotal as it identifies which features exert the most significant influence on the outcome variable. Understanding feature importance allows stakeholders within the Aviation sector to discern which variables hold the greatest sway over outcomes, thereby directing attention to areas that require improvement to enhance service delivery. Moreover, this understanding yields insights into the underlying dynamics of the system being modeled, thereby informing decision-making and strategy development. Lastly, analyzing variable importance enhances model interpretability, enabling stakeholders to grasp the factors driving model predictions.

As depicted in Figure 17, the departure minute and day emerge as the most critical features, collectively contributing to approximately 30% of the predictive power. The origin and destination each account for around 15% to 18%, while the day of the week represents approximately 7-8%.

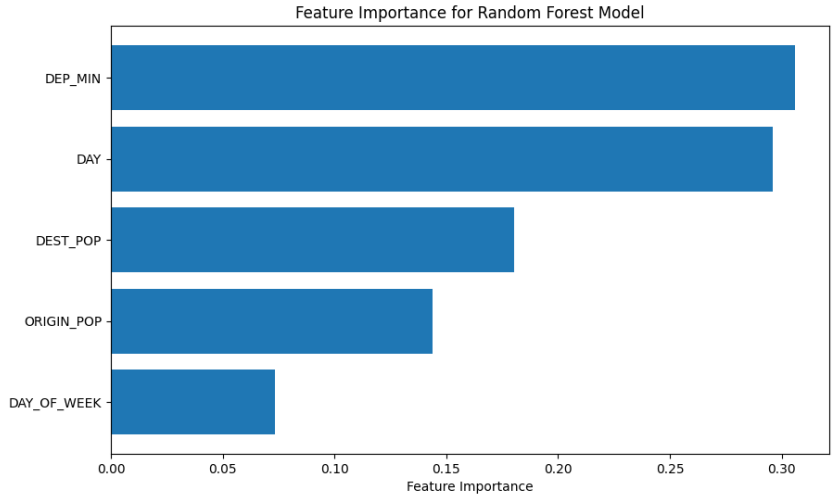

Figure 17: Feature Importance for Random Forest Model

While drawing definitive conclusions solely from this analysis may pose challenges, the variable importance plot aligns with the Time Analysis conducted in Step 2 of the Exploratory Data Analysis (EDA) section. This alignment underscores the imperative for airlines to prioritize management strategies during high-traffic seasons. Examples of such strategies include:

- Requiring passengers to arrive early to facilitate timely completion of check-in and gate procedures, thereby minimizing ground-time for aircraft, which is critical during busy flight schedules.
- Planning aircraft routes while considering potential delays or, alternatively, renting additional aircraft for periods of high demand.
- Minimizing vacation time for pilots, cabin crew, and ground personnel during peak periods, or augmenting the workforce with seasonal hires.
- Collaborating with ATC to proactively plan routes and taxiing procedures, thereby reducing queue wait times before take-offs and landings.

Models Performance Evaluation

For the purpose of this research project, four distinct Machine Learning models have been implemented and evaluated:

1. **Random Forest**: *Random Forest* is an ensemble learning method, considered an evolution of a "simple" *Decision Tree*, that builds multiple decision trees during training and combines their predictions through a voting mechanism or averaging to improve accuracy and reduce overfitting. The *Random Forest* model builds on the idea of bagging, which is the resampling of the observed dataset (and of equal size to the observed dataset), each of which is obtained by random sampling with replacement from the original dataset, but providing also an improvement since it decorrelates trees: each time a split in the tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors (usually, and often, $m = \sqrt{p}$). *Random Forest* is usually robust to noise and capable of capturing complex relationships within the data.

   To mitigate overfitting, it's also crucial to establish an optimal maximum depth, which refers to the maximum number of levels in a single decision tree. As illustrated in Figure 18 below, the optimal maximum depth for our case study is identified at 28. This value corresponds to the point where precision and recall intersect, given that in data science, the objective is to maximize both precision and recall simultaneously.

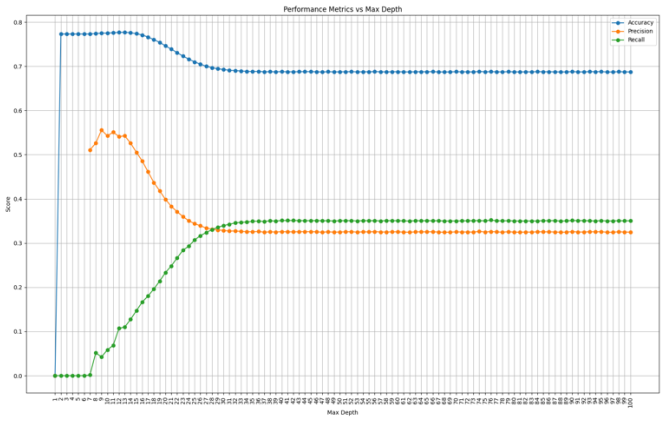

Figure 18: Performance Metric vs. Max Depth in Random Forest Model

2. **K-Nearest Neighbor (KNN)**: *KNN* is a non-parametric (which means that the scientist does not need to assume any functional form) and instance-based learning algorithm used for classification and regression tasks. It predicts the class of a data point by identifying the majority class among its K nearest neighbors in the feature space. *KNN* may suffer from computational inefficiency with large datasets and requires careful selection of the K parameter. The distance formula used by default by the Scikit-Learn library in Python is the Euclidean formula, namely: $d(P,Q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$, where q_i and p_i are the ith coordinates of points P and Q, respectively.

   To mitigate potential overfitting and ensure the robustness of our *KNN* model, it's important to establish an optimal number of neighbors to be considered. As illustrated in Figure 19 below, we determined that selecting K=15 achieved the highest classification accuracy-recall tradeoff on the testing set (we trained and tested KNN with K= 15, 50, 100, 150, 200, 300, 500, 1000).

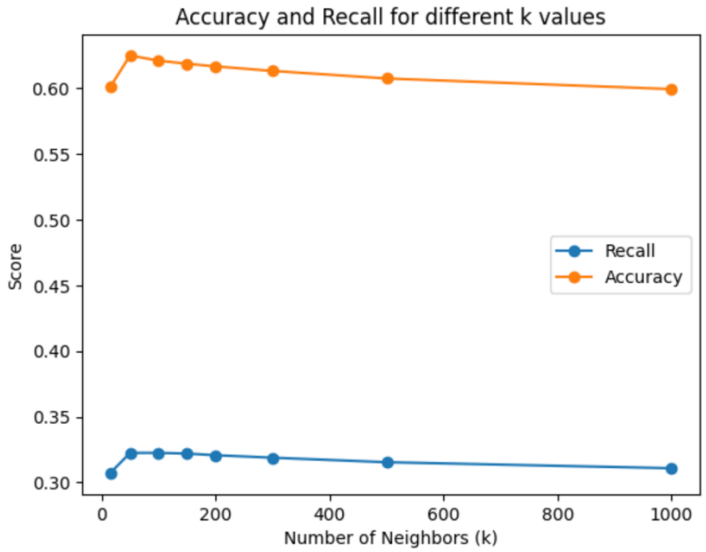

Figure 19: Accuracy and Recall for different k-values

3. **Logistic Regression**: *Logistic Regression* is a linear model used for binary classification tasks. It estimates the probability that a given input belongs to a certain class using a logistic function, which maps the input features to a probability between 0 and 1. Despite its name, *Logistic Regression* is a classification algorithm, not a regression algorithm. It is interpretable, computationally efficient, and robust to noise, but it assumes a linear relationship between the input features and the log-odds of the outcome variable.

   The logistic function for *Linear Regression* is the following: $y = \frac{e^{(\beta_0+\beta_1 X)}}{1+e^{(\beta_0+\beta_1 X)}}$

4. **Extreme Gradient Boosting (XGBoost)**: *XGBoost* is a scalable and efficient implementation of gradient boosting, a machine learning technique that builds an ensemble of weak learners (typically decision trees) sequentially to improve predictive performance. *XGBoost* employs a regularization term in the objective function to prevent overfitting and utilizes parallel and distributed computing techniques for faster training. It is highly flexible, interpretable, and often achieves state-of-the-art performance in a wide range of machine learning tasks.

Table 3 below shows the performance of each model trained for this research project.

Models Evaluation Metrics			
		Accuracy	*Recall*
Random Forest	*Training*	99.98%	99.99%
	Testing	62.8%	61.9%
KNN	*Training*	63.5%	61.3%
	Testing	62.5%	59.4%
Logistic Regression	*Training*	58.3%	63.4%
	Testing	55.4%	63.6%
XGBoost	*Training*	67.0%	65.5%
	Testing	65.3%	63.3%

Table 3: Models Evaluation Metrics

Focused Analysis

In many instances, the Technical Operations Department of airlines delves into highly specific domains. Thus, the latter segment of the Machine Learning phase centered on forecasting flight delays by narrowing the scope to a specific origin airport. Considering our focus on American Airlines, I opted to focus on their central hub: Dallas Fort-Worth airport. Presented below (Table 4) are the evaluation metrics results utilizing XGBoost as the Machine Learning model.

AA From DFW: XGBoost Performance	
Accuracy	66.2%
Recall	65.3%

Table 4: AA From DFW: XGBoost Performance

Discussion

This research project aimed to leverage Data Analysis and Machine Learning techniques to enhance the performance of the Aviation sector. As outlined in the *Introduction* section, delays in aviation operations incur significant financial repercussions. With the adoption of the techniques employed in this research, several key objectives can be achieved:

- Planning in advance fleet and personnel schedules by just considering the delays prediction techniques illustrated could mitigate the impact of delays by approximately 65%.
- Potentially saving airline companies a total of USD 5.5 billion.
- Potentially saving passengers a total of USD 11 billion.

Both airport management companies and airlines stand to benefit from the strategies elucidated in this research project. By conducting more focused analyses on specific aspects, akin to the examination of American Airlines flights originating from Dallas Fort-Worth, even greater improvements could be realized beyond those elucidated above.

Furthermore, with access to more detailed datasets than the one provided by USDOT, including tracking data pertinent to specific airlines for route optimization, and collaboration with other relevant units within the Aviation sector, such as the Meteorological division for weather condition analysis, the outcomes of this study could be further enhanced. This could involve not only reducing the percentage of delays but also devising more efficient schedules for aircraft and flights, thereby fostering improved operational efficiency and customer satisfaction within the Aviation industry.

Appendix A: Evaluation Metrics

Throughout the Machine Learning sub-section in the *Results* section, several common Evaluation Metrics have been employed to assessing the efficiency of each Machine Learning model used.

Before proceeding in the explanation of the metrics used, it's important to define the following:

- *True Positive (TP)*: Instances correctly identified as belonging to the positive class.
- *False Positive (FP)*: Instances wrongly identified as belonging to the positive class.
- *True Negative (TN)*: Instances correctly identified as belonging to the negative class.
- *False Negative (FN)*: Instances wrongly identified as belonging to the negative class.

The following is the explanation of the metrics that have been employed:

- *Accuracy*: important to assess the overall capacity of the model to identify correctly instances as belonging to the positive or negative class. It's sensitive to class imbalance issues.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}.$$

- *Recall (Sensitivity)*: important to assess the ability of the model to recognize observations belonging to the positive class.

$$Recall = \frac{TP}{TP+FN}.$$

- *Precision*: important to assess the quality of the model to recognize observations belonging to the positive class:

$$Precision = \frac{TP}{TP+FP}.$$

The goal of a Data Scientist is to find a model that maximize both recall and precision.

Works Cited

"American Airlines." *Wikipedia*, Wikimedia Foundation, en.wikipedia.org/wiki/American_Airlines. Accessed 22 Apr. 2024.

Anupkumar, Ashmith. "*INVESTIGATING THE COSTS AND ECONOMIC IMPACT OF FLIGHT DELAYS IN THE AVIATION INDUSTRY AND THE POTENTIAL STRATEGIES FOR REDUCTION*", 2023. Electronic Theses, Projects, and Dissertations. 1653. Accessed 22 Apr. 2024.

"Global Outlook for Air Transport a Local Sweet Spot." *IATA*, www.iata.org/en/iata-repository/publications/economic-reports/global-outlook-for-air-transport---december-2023---report/. Accessed 22 Apr. 2024.

Josephs, Leslie. "FAA System Outage Disrupts Thousands of Flights across U.S." *CNBC*, CNBC, 31 Jan. 2023, www.cnbc.com/2023/01/11/faa-orders-airlines-to-pause-departures-until-9-am-et-after-system-outage.html.

"Largest Airlines by Market Cap." *CompaniesMarketCap.Com - Companies Ranked by Market Capitalization*, companiesmarketcap.com/airlines/largest-airlines-by-market-cap/. Accessed 22 Apr. 2024.

"List of Largest Airlines in North America." *Wikipedia*, Wikimedia Foundation, en.wikipedia.org/wiki/List_of_largest_airlines_in_North_America. Accessed 22 Apr. 2024.

"The Statistics Portal." *Statista*, www.statista.com/markets/419/topic/490/aviation/#overview. Accessed 22 Apr. 2024.

"Winter Storm Summary for February 21-23, 2023." *National Weather Service*, NOAA's National Weather Service, 24 Feb. 2023, www.weather.gov/arx/feb2323.