Understanding Flight Delays

***Abstract*—This comprehensive flight delay prediction analysis dives into the complex world of aviation disruptions, meticulously examining diverse datasets comprising flight details, weather records, and operational metrics. Spanning sources like Bureau of Transportation Statistics, FAA, and Weather Underground, the study dissects various factors contributing to delays, encompassing flight specifics, weather attributes, and airport operations. Leveraging techniques including ridge regression, XGBoost, and LSTM, the report aims to pre-empt and manage disruptions by delving into extensive data analysis and modeling. The project's integrated approach scrutinizes vast datasets to unravel insights into mitigating disruptions and optimizing aviation operations, offering a roadmap to proactively address flight delays and enhance operational efficiency within the aviation industry. Through detailed analysis and predictive modeling, this study navigates through multifaceted data to unveil strategies for reducing disruptions and improving overall aviation performance.**

I. Introduction

Flight delays pose a pervasive challenge in contemporary aviation, impacting passengers, airlines, and airports significantly. In 2013, approximately 31.1% of flights encountered delays surpassing 15 minutes, underlining the severity of this issue. This project takes a deep dive into extensive flight delay data, encompassing both domestic and international flights. Its overarching objectives revolve around dissecting the complex array of factors contributing to flight delays and, more importantly, constructing predictive models geared towards anticipating and mitigating these disruptions. The initiative involves a comprehensive examination of various variables such as weather conditions, airline operations, airport congestion, aircraft statuses, and external influences. By scrutinizing patterns, correlations, and causative factors from vast data repositories, predictive models empower stakeholders to make informed decisions, manage operational challenges, and preemptively address potential disruptions. Moreover, the project endeavors to develop predictive models using advanced machine learning algorithms. These models will harness the insights gleaned from data analysis to forecast potential delays, enabling stakeholders to take proactive measures and make informed decisions to preempt, manage, and reduce the impact of flight disruptions on the aviation ecosystem.

II. Dataset Description

This research delves into the realm of domestic commercial flights within the United States, analyzing data collected between January 2015 and August 2023. The information derives from reputable sources, namely the Bureau of Transportation Statistics and the Federal Aviation Administration. Focused on 45 key airports, the dataset encompasses comprehensive monthly reports from various airlines. It encompasses an array of critical details including flight times, airline specifics, origin and destination data, departure and arrival performance, cancellations, diversions, flight summaries, delay causes, available seat miles, daily flight operations, and staffing metrics.

Supplementing this aviation data, weather information was sourced from Weather Underground, spanning a nine-year hourly record. This weather dataset includes significant attributes such as wind speed, humidity levels, weather conditions, atmospheric pressure, and dew point readings.

A. Data Collection

To gather this wealth of data, a combination of scraping tools—Selenium and BeautifulSoup—was deployed. These tools were instrumental in extracting information from three primary sources: https://www.transtats.bts.gov, https://www.aspm.faa.gov, and https://www.wunderground.com. Through these efforts, a comprehensive dataset was compiled, enabling an in-depth analysis of the intricate interplay between flight operations and meteorological conditions over an extended time span.

To address the problem, our strategy involved creation of new features for each flight and tried to handle all possible types of delays by collecting data from multiple sources.

Type of Delay	***Data used***
Weather Delay	Historic weather data
Carrier Delay	Aircraft, Staff details of the carrier
Security Delay	No of Passengers travel through airport
Late Aircraft Arrival	Previous flight delay

TABLE I
Types of Delays

B. Feature Extraction

From each dataset,we employed feature engineering to generate the most predictive and informative features for the intended task. we were able to extract some relevent features for each flight. Below are some examples of features we selected:

Feature	*Description*
Day of Month	The day of the month
Day of Week	The weekday of flight
Operating Airline	Airline company operated flight
Origin Airport	flight departure airport
Destination Airport	flight arrival airport
Scheduled Departure Time	Planned departure time
Actual Departure Time	Time of flight departed
Scheduled Elapsed Time	Planned flight duration
Actual Elapsed Time	Actual flight duration in minutes
Wheels-Off Time	The time a flight takes off
Temperature	Temperature(inp F) at origin airport
Wind Speed	WindSpeed(mph) at origin airport
Wind Direction	Wind Direction at origin airport
Condition	Weather condition at origin airport
Is Holiday	1 if holiday in +/-2days, else 0
Age	Age of the aircraft
Number of Seats	Total seats in aircraft
Model	Model type of Aircraft
TailNumber	Used for cascading effect

TABLE II
SOME OF THE FEATURES EXTRACTED FOR THE ANALYSIS

C. Data Cleaning and Preparation

We leveraged distributed framework like Dask to perform pre-processing like parsing dates and text on the gathered large corpus of raw-data. We added several features related to aircrafts(age, model, number of seats), holiday season. We handled missing features by replacing with average/median/best possible approximation based on the feature type. We converted categorical variables into a numerical format suitable for analysis and modeling using label encoding. We normalised the data to maintain uniformity across each feature ranges. Once the feature engineering is completed and final set of data is decided, we shifted to Python to leverage Pandas, Matplotlib libraries for exploratory analysis and visualisations.

D. Data Integration

Data integration is the process of merging information from diverse sources to create a unified and coherent dataset. This amalgamation is fundamental in generating comprehensive data that facilitates deeper insights, better decision-making, and more robust analysis. We collected data from various websites and employed scraping techniques to compile and consolidate it into a singular file. Specifically, we merged flight-related data with available weather information using departure times and the originating airport's location details as key parameters. This combined dataset enables a more holistic analysis, intertwining flight-specific details with pertinent weather conditions for a more comprehensive understanding of their relationship and impact.

E. Data Analysis and Visualization

Gathering and preparing a dataset through web scraping encompasses multiple stages, starting with data extraction and culminating in cleaning and formatting for analytical purposes. This process entails retrieving weather-related information, such as temperature, humidity, and precipitation, for each day within a specified date range from relevant websites. Initially spanning 16 gigabytes, the collected data underwent extensive analysis, extraction, and transformation, resulting in a streamlined dataset of 4 gigabytes. Multiple procedures were employed to refine and condense the information, ensuring its suitability for analysis and facilitating a more manageable and efficient dataset for further examination and utilization.

III. Literature Review

Analyzing existing research on flight delay prediction encompasses methodologies, data sources, and challenges in aviation. This review investigates predictive models, data features, and their impact, aiming to improve operational efficiency, passenger experience, and safety within the aviation industry.

IV. Exploratory Data Analysis

In order to get a qualitative understanding of the different factors of delays, we did some exploratory data analysis

A. Weather factors effect on flight delays

Individually, specific weather factors such as temperature, wind speed, humidity, pressure, precipitation, and dew point don't typically lead to significant flight delays unless they reach extreme levels. However, when multiple weather elements coincide or interact, a noticeable increase in flight delays becomes evident. depicted in fig.1

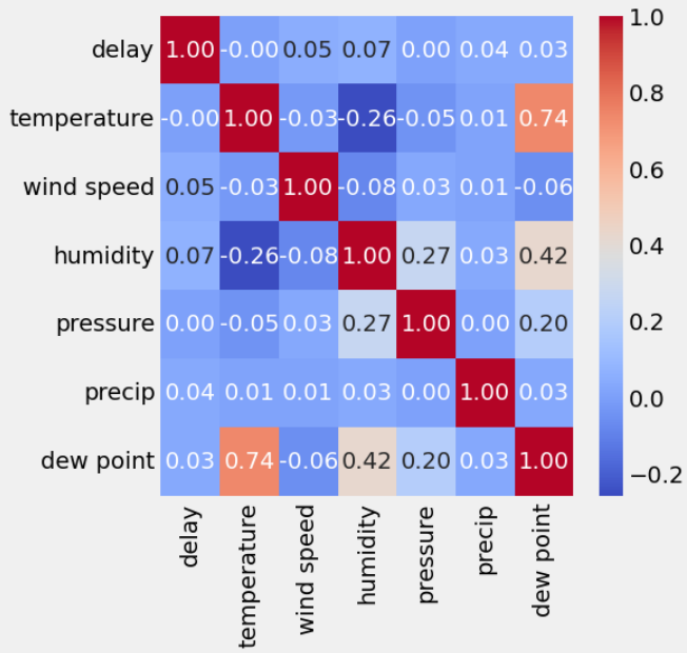

Fig. 1. Correlation matrix of weather factors and delay.

B. Average Delays in the US States

An interesting question that comes up during analysis is if certain states experience delays to a higher extent that other states. Post-analysis, we've come to the conclusion that Ohio usually leads the country in frequency of delays while Puerto Rico is overwhelmingly the best performer at minimizing delays experienced by flight travelers. More interesting questions could follow as to if there is a disparity of delays between airports on the east coast and the west coast, as well as the Midwest, and if there's a significant difference in delay rate between blue and red states.

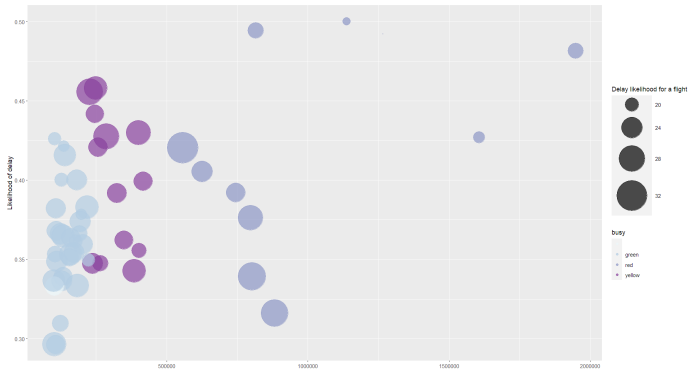

Fig. 2. The above bubble plot shows the relation between how busy an airport is and the degree of delay experienced. Bubble size corresponds to average delay duration. Busiest airports to the right.

C. Do busier airports experience a higher rate of delay than less busier airports?

While the extremes in Figure.3 do seem to answer in positive to the above question. We can not quite assert this statement due to the relation between busy-ness of an airport and rate of delay being quite nebulous. While JFK in particular experiences close to the highest degree of delay, other similarly busier airports report delay rates close to the median. The overall correlation between these two axes is close to 0.22 showing a shaky correlation at best between busy-ness of an airport and degree of delay experienced. This points to a greater influence of other factors like weather conditions, geographical location dominating over airport size.

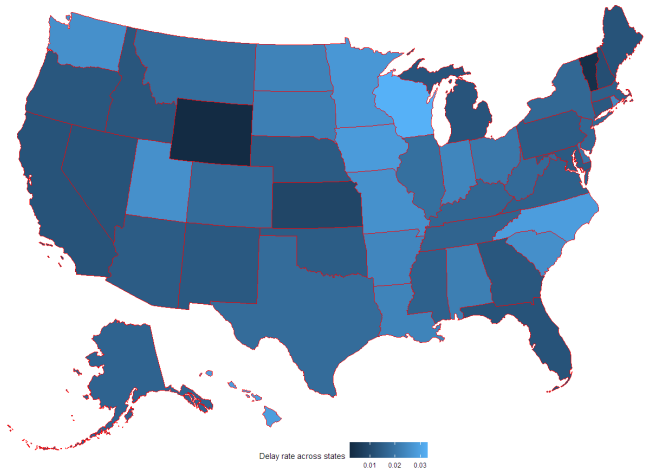

Fig. 3. The above US map shows the regularity of delays by each state

D. Weather factors by airport location

Airports across the United States can be divided into four different clusters based on certain weather events afflicting these areas. This data of weather events has been gathered from https://smoosavi.org/datasets/lstw. Airports were then clustered based on frequency and severity of extreme weather events like snow, hail etc, geographical location (based on latitude and longitude) for the airports in each state to be clustered as shown in figure 4.

E. Weather factors by airport location

When comparing rate of delays in extreme conditions as opposed to typical weather conditions, airlines usually face a much higher rate of delay in extreme weather conditions. The bubble plot in Figure 2 shows sand/dust whirlwinds causing a near certain delay in airline disruption whenever they occur with freezing rain and

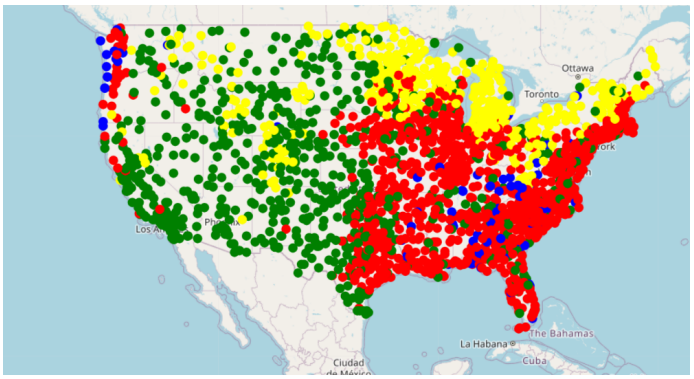

Fig. 4. Airports across the United States can be divided into 4 clusters based on weather conditions.

thunderstorms also disrupting regular airline functioning to a high degree of certainty. On the other hand, typical weather conditions like fogginess and cloudiness cause a much lower degree of delay, with even the delays that are actually caused turning out to be of a much lower duration.

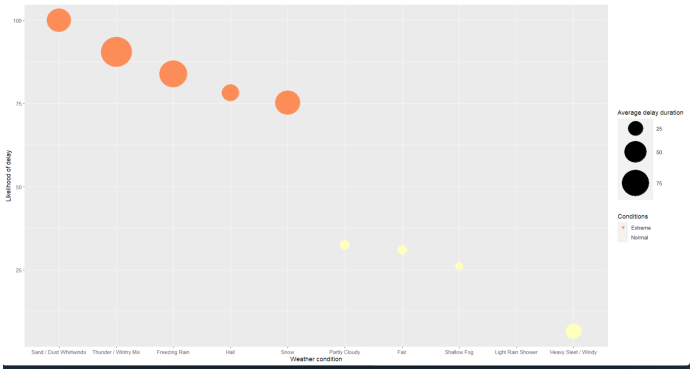

Fig. 5. Bubble plot of weather conditions and rate of delay caused by each. Size corresponds to average delay duration. Extreme weather conditions on the left and regular weather conditions on the right.

F. Delay rate and geographic location

When analysing the relation between delay and degree of flight activity at airports, we realised that, airports concentrated towards the center of the United States seemingly had a greater degree of delay. The above plot confirms a part of the hypothesis with centralized airports having a bigger bubble on average than those towards the West and East Coasts, and the airports on the coast having almost definitely smaller degrees of delay compared to those on the inside. Just inductively a hypothesis that could be worth exploring is if rate of delay increases the further we go inland. After analysis, we were able to confirm a degree of correlation of 0.55 of rate of delay with the distance of an airport from the oceans. With most of the airports concentrated heavily on the East Coast or towards the eastern part of the United States, some interesting insights could possibly be found there.

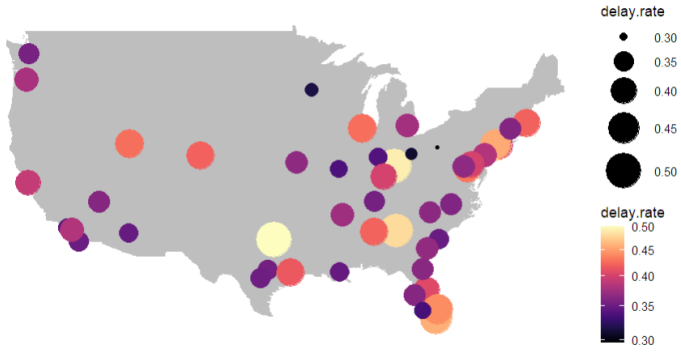

Fig. 6. Plot of airports across map. Size and color of bubble for each airport corresponds to degree of rate of delay.

G. Cascading effect on flight delays

The primary objective of this section is to understand how delay propagates through airlines and airports. Delay propagation occurs when a delay at a flight stage causes a ripple effect in the subsequent stages of a flight. Delays propagate into and out of an airport. Arrival delays are tracked at the end of each flight leg traveled by the same aircraft identified by a tail number. In order to compute delay propagation, the mean and variance for the late aircraft delay values are calculated. In fact we observed that there has been a delay propagation in the following conditions:

- A flight arrives late at an airport.
- A flight departs late in subsequent stages.
- A flight arrives late at the next destination.

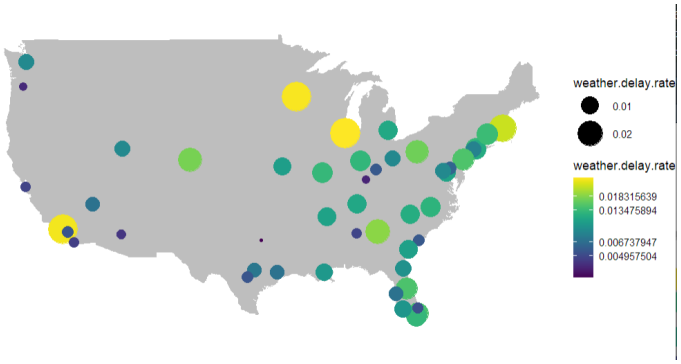

Fig. 7. Plot of airports across a map of the United States. Size and color of bubble corresponds to the ratio of weather-effected delays as opposed to non-weather related delays.

H. Geographic weather factors and delays

Figure 7 shows the top 50 most busiest airports on a map of the United States with biggest and palest bubbles corresponding to airports with a greater ratio of weather-affected delays. Crucially, the biggest and lightest bubbles lie in the 'Blizzard Valley' region of the United States which refers to the the three upper midwestern states - North Dakota, South Dakota and Minnesota - states known for experiencing a much higher range of winter storms when compared with the rest of the country.

I. Delay rate by airline

Preliminary analysis of the correlation between delay rate and the date of founding of an airline revealed some interesting results as delay rates for 2022 were in the exact same order as the ages of the 10 major airlines. The oldest airlines including Delta and Hawaiian Airlines flyers suffered the least amount of delays, with the newer airlines experiencing a much higher rate of delay on the other hand. Expanding the dataset to include all flights over the past decade from the current big 10 of commercial airlines showed a less consistent pattern, but the correlation between its position on the list when ordered by age of airline and its position on the list when ordered by rate of delay proved to be very high at around 0.78. The stacked area chart above shows the contribution of each airline in the overall delay rate of airlines over the past decade.

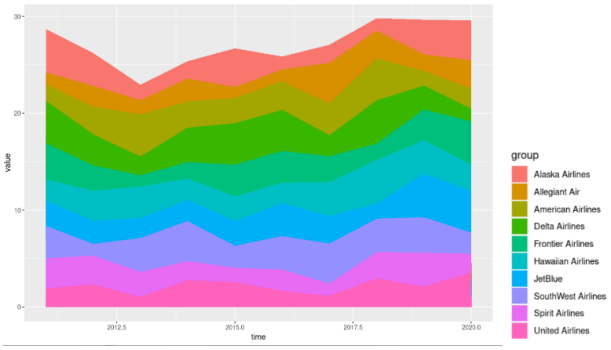

Fig. 8. The above stack chart shows each airline's contribution to overall delay from 2010-2020.

J. Delay rate of popular American airlines

Following analysis of the current top 10 airlines in the United States, flights by Hawaiian Airlines experienced the least amount of delay, with SouthWest Airlines and Delta Airlines being delayed the most. The barplot shows Hawaiian Airlines, Alaska Airlines and United Airlines as the three best performing airlines with respect to delay rate, which brings out an interesting insight. These three airlines are substantially smaller than the other seven in all regards ranging from annual revenue to staff size to fleet size, showing a link between scale of operations and rate of delay. However, when comparing the relation between scale of operations of an airline and the delay rate for each airline. We had to take into consideration several parameters, predominantly those that measured ASM(available seat miles), staff size and gross annual revenue of an airline. Post calculation of a parameter that took into account all three of these factors, we found out that scale of operations was correlated with the delay rate of an airline at a value of around 0.42. Notorious 'budget' airlines like Spirit Airlines and Allegiant Air performed as expected, netting a higher delay rate as compared to other airlines with a similar staff size but at a lower ASM. Hawaiian Airlines however outperformed airlines with a similar staff size and annual revenue by a large margin.

Fig. 9. The above stack chart shows each airline's contribution to overall delay from 2010-2020.

V. Modelling Techniques

We have explored several techniques for modelling this regression problem of predicting flight delays. We selected algorithms that are more suitable for the flight delays problem and created the predictive models.

A. Regression Technique

1) Linear regression: It is a supervised machine learning algorithm used for predicting a continuous tar-

get variable based on one or more independent features. It models the relationship between the variables by fitting a linear equation to the observed data, aiming to minimize the sum of squared differences between predicted and actual values.

2) Ridge regression: It is a regularization technique applied to linear regression. It introduces a regularization term (L2 penalty) to the linear regression objective function, preventing overfitting by penalizing large coefficients. This regularization term is controlled by a hyperparameter (alpha), and as alpha increases, the impact of regularization on the model increases, leading to a more robust and stable model, enhances model generalization, and improves prediction accuracy by mitigating the impact of noisy or irrelevant variables often present in complex datasets related to air travel.

B. XGBoost Technique

The exploratory data analysis showed that our target variable delay is non-linear, which requires us to use more sophisticated techniques to achieve better results. Once such technique is XGBoost Regressor, it is an ensemble learning algorithm that combines the predictions of multiple weak learners, typically decision trees, to produce a robust and accurate regression model. It is suitable for flight delay prediction due to its ability to handle complex relationships in data, handle missing values, and resist overfitting, making it well-suited for capturing the intricate patterns and factors affecting flight delays in a dynamic and diverse dataset.

C. Long Short-Term Memory(LSTM)

Recurrent neural networks (RNN) are well designed for predicting sequences over time, and are a natural fit for this problem. In particular the LSTM variant of RNNs, which have have gained traction in recent years, nicely fit the characteristics of the problem. Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture designed to overcome the vanishing gradient problem in traditional RNNs, allowing it to effectively capture and remember long-term dependencies in sequential data.

LSTM is suitable for flight delay prediction because it can model complex temporal relationships in historical flight data, considering factors such as weather conditions, air traffic, and airport congestion. Its ability to retain and update information over extended periods makes it well-suited for handling the intricate patterns and dependencies inherent in the domain of flight delays.

VI. Results and Discussions

A. Metrics Used

To gauge the quality and effectiveness of the models, we used standard metrics like Mean Squared Error(MSE), Mean Absolute Error(MSE), R-Squared Score. F1 Score, AUC Metrics are not considered since they are more suitable for classification problems.

1) Mean Squared Error(MSE): The Mean Squared Error measures how close a regression line is to a set of data points. It is a risk function corresponding to the expected value of the squared error loss. Mean square error is calculated by taking the average, specifically the mean, of errors squared from data as it relates to a function.

2) Mean Absolute Error(MSE): The magnitude of the difference between the individual measurement and the true value of the quantity is called the absolute error of the measurement. The arithmetic mean of all the absolute error is taken as the mean absolute error of the value of the physical quantity.

3) R-Squared Score: R-Squared (R^2 or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

B. Discussion

In this study, we conducted a comprehensive evaluation of above discussed predictive models to assess their performance in the context of flight delay prediction. To ensure a rigorous evaluation, we utilized a dataset spanning three months (September to November 2023), which was intentionally excluded from the training process to provide an unbiased assessment of model generalization.

Model	***MSE***	***MAE***	***R2-Squared***
Linear Regression	154.62	4.26	0.72
Ridge Regression	104.01	3.39	0.88
XGBoost	76.69	2.65	0.98
LSTM	72.46	2.51	0.98

TABLE III
PERFORMANCE OF EACH MODEL

The results, as summarized in Table III, reveal distinctive performance characteristics among the models. Notably, the LSTM model stands out as the top performer across all metrics, exhibiting an MSE of 72.46, MAE of 2.31, and an impressive R2-Squared value of 0.98. This

superior performance can be attributed to the LSTM's capability to effectively learn and generalize complex temporal dependencies within the flight delay data. The model's proficiency in capturing and understanding sequential patterns positions it as a promising choice for applications demanding a nuanced understanding of time-series data.

The observed performance hierarchy showcases the limitations of traditional regression models like Linear and Ridge Regression in capturing the intricate dynamics inherent in flight delay prediction. Meanwhile, the XGBoost algorithm, renowned for its efficacy in handling structured data, demonstrated commendable performance, but it fell short of surpassing the LSTM in this specific context. The findings underscore the importance of selecting models tailored to the inherent characteristics of the data, with a preference for specialized architectures like LSTM when dealing with temporal dependencies in predictive modeling scenarios.

C. Feature Importance

In our pursuit of comprehending the primary contributors to flight delays, we explored the significance of various features using an XGBoost model. Notably, 'Carrier Code', and 'Condition', 'Scheduled Elapsed Time (Minutes)' emerged as the most influential factors in the model's training.

The prominence of 'Carrier Code' (representing the Airline Company) as a top feature raises interesting considerations. It could potentially be attributed to a bias in the model training stemming from the substantial variability in market shares among different airline companies in our dataset. Notably, Southwest (19.4%), Delta (19.4%), American (15.7%), and United (10.3%) airlines collectively account for approximately 65% of all flights(see Fig. 10) and these major airlines exhibit lower delay rates compared to their counterparts(See Fig. 11).

The 'Condition' feature, reflecting the weather conditions at the airport, also emerged as a significant influencing factor. Days characterized by adverse weather conditions such as Sand/Dust Whirlwinds (100%), Thunder (90%), Wintry Mix, Freezing Rain (84%), and Snow (76%) exhibited higher delay rates. In contrast, normal weather conditions, such as Fair, Light Rain, and Shallow Lower, were associated with comparatively lower delay rates, which has been illustrated in Fig. 5 in exploratory analysis section. As weather condition being one of the top features, we also explored whether flight delays can be predicted solely based on weather factors. But, the performance of the models dropped significantly.

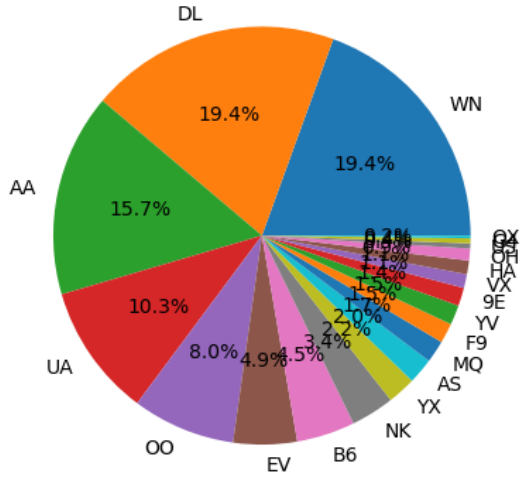

Fig. 10. Market share of airlines

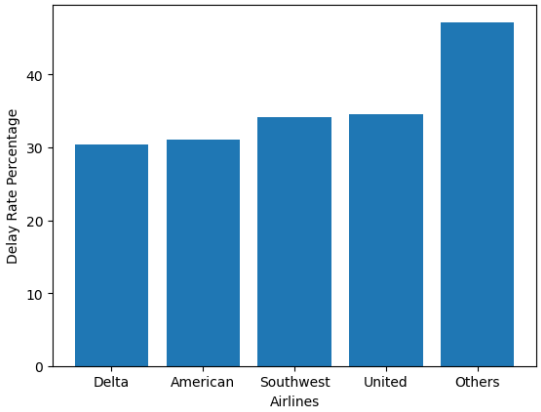

Fig. 11. Comparison of delay rates by top airlines and remaining other airlines

'Scheduled Elapsed Time (Minutes)' is identified as a crucial factor, as flights with longer durations tend to experience higher delays. This association can be attributed to the increased likelihood of encountering multiple obstacles during extended flight durations.

The observed patterns emphasize the need for nuanced interpretations, especially in the context of biases related to airline market shares and the influence of adverse weather conditions on flight punctuality.

VII. Conclusions

In conclusion, this study has addressed the pervasive challenge of flight delays within the aviation sector, providing a comprehensive analysis of diverse factors influencing disruptions. Through the integration of extensive datasets from reputable sources such as the Bureau of Transportation Statistics, Federal Aviation

Administration, and Weather Underground, we meticulously examined the interplay between flight delay and flight operations, Temporal (timeof-day, day-of-week, holidays), local (airport delay, carrier delay) and weather (temparature, wind speed, humidity) factors. Employing Selenium and BeautifulSoup scrapers facilitated the creation of a unified dataset, enabling in-depth analysis and modeling.

The research encompassed a thorough exploration of various modeling techniques, including linear regression, ridge regression, XGBoost, and LSTM, to predict flight delays. The comparative performance evaluation revealed that the LSTM model outperformed other algorithms, showcasing its ability to learn and generalize strong temporal dependencies within the data. This superior performance positions LSTM as a promising choice for flight delay prediction, particularly when dealing with intricate temporal patterns.

The study contributes valuable insights into the complexities of flight delays and provides a foundation for optimizing aviation operations. By leveraging advanced machine learning algorithms and considering diverse factors such as weather conditions, airport congestion, and aircraft statuses, stakeholders can make informed decisions to preempt, manage, and mitigate the impact of flight disruptions. The findings underscore the importance of adopting sophisticated modeling techniques tailored to the unique characteristics of the aviation domain.

As the aviation industry continues to grapple with the challenges of timely and accurate predictions, the outcomes of this research offer a roadmap for the development and deployment of effective predictive models. Future work in this domain may involve exploring additional features, refining existing models, and considering real-time data integration to enhance the accuracy and applicability of flight delay predictions.

REFERENCES

[1] Evangelos Mitsokapas, Benjamin Schäfer, Rosemary J. Harris & Christian Beck, “Statistical characterization of airplane delays“ Scientific Reports volume 11, Article number: 7855 (2021)

[2] 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.

[3] Martina Zamkova 1ORCID,Stanislav Rojik 2, ORCID,Martin Prokop 1 andRadek Stolin 1 “Factors Affecting the International Flight Delays and Their Impact on Airline Operation and Management and Passenger Compensations Fees in Air Transport Industry: Case Study of a Selected Airlines in Europe“

[4] Fen Zhou,1Guosong Jiang,1Zhengwu Lu,1and Qingdong Wang ”Evaluation and Analysis of the Impact of Airport Delays”, Volume 2022 — Article ID 7102267 — https://doi.org/10.1155/2022/7102267

[5] Kerim KiliçORCID andJose M. Sallan, ”Study of Delay Prediction in the US Airport Network”, Aerospace 2023, 10(4), 342; https://doi.org/10.3390/aerospace10040342

[6] ”A survey of flight delay models in air traffic management” by Lai, C., Zhuang, X., & Li, D. (2017).

[7] ”An empirical study on the impacts of weather conditions on flight delays” by Zhang, J., Fu, X., & Lei, Z. (2013).

[8] ”Predicting air travel delays: A comparison of classification and regression methods” by Grigas, J., Potts, C., & Camerer, C. F. (2013).

[9] ”Airport and airline choice by travelers” by Basso, L. J., & Zhang, Y. (2012).

[10] ”Statistical modeling of air traffic delays in the United States” by Haghani, A., & D’Apuzzo, K. (2005).

[11] ”Impact of airline flight schedule changes on passengers’ air travel decisions” by Wang, D., & Zheng, Q. (2005).

[12] ”A multi-objective model for aircraft arrival, departure, and taxi-out scheduling” by Ball, M. O., et al. (2007).

[13] ”A survey of air transportation and air traffic control” by Barnhart, C., et al. (2003).

[14] ”Managing delays in air traffic” by Ball, M. O., & Barnhart, C. (2002).

[15] ”Airport and airline choice in a multiple airport region: An empirical analysis for the San Francisco Bay Area” by Zou, B., & Hansen, M. (2007).