Flight Status Prediction

RUI XU, MUTONG ZHANG, and YUZHE HE

ACM Reference Format:

Rui Xu, Mutong Zhang, and Yuzhe He. 2023. Flight Status Prediction. 1, 1 (December 2023), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The rapid evolution of technology has brought transformative changes to our world. Take the realm of transportation for example. Visionaries throughout history have ingeniously invented various modes of transportation to meet diverse needs arising from a myriad of circumstances. These inventions stand as important contributors to enhanced efficiency and mobility for humans, resulting in substantial savings on both time and resources. Moreover, these inventions have fostered connectivity among people and communities, significantly improving the overall quality of life. By facilitating economic activities and catalyzing economic development, these advancements in transportation not only benefit individuals but also contribute to the growth and prosperity of societies.

Nowadays, aviation technology, particularly airplanes, has emerged as a prominent element in the domain of long-distance journeys. The widespread availability of air travel has significantly improved global connectivity, fostering more convenience in transportation. However, it is also essential to acknowledge the existence of potential drawbacks. Beyond the environmental concerns tied to carbon emissions and noise, the susceptibility of flights to delays or cancellations due to adverse weather conditions is a long-standing challenge. This weather-dependency of flights introduces inconvenience for both individuals by disrupting travel plans and airlines by imposing unavoidable operational challenges.

For our project, the objective of our group is to harness flight status data spanning five years, from 2018 to 2022. Our focus is on utilizing this piece of comprehensive data to forecast the likelihood of delays and cancellations specific to every flight in the United States. Additionally, our project endeavors to go beyond predicting delays and cancellations by attempting to approximate the potential duration of delays in and early departures or arrivals. We would like to take advantage of our predictive model and analysis to contribute insights that can enhance the reliability and efficiency of travel by air.

2 DATASET

2.1 Dataset Information

The dataset we chose, named *Flight Status Prediction* , is sourced from *Kaggle* (https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022).

Authors' address: Rui Xu, r276xu@uwaterloo.ca; Mutong Zhang, m246zhan@uwaterloo.ca; Yuzhe He, y395he@uwaterloo.ca.

Manuscript submitted to ACM

The dataset, as noted in the introduction, presents a comprehensive repository of information encompassing all flight information that both departed from and arrived in the United States between January 2018 and July 2022. This dataset is remarkably sizable, approximately 25.78 gigabytes in total. The author partitioned the data into five distinct files based on year. Each file comprises of 61 columns that contains meticulously recorded attributes such as flight dates, airline information, and places of origin. One notable feature of this dataset is its intrinsic design for expansiveness for the anticipation of update with the inclusion of data for 2023.

2.2 Preprocessing Data

The dataset, once downloaded and extracted from a zip file, expanded to approximately 45 gigabytes, making it impractical for comprehensive analysis within the constraints of a single computer's RAM. With around 30,000,000 rows and 61 columns, the enormous size necessitates the utilization of tools like *PySpark* for efficient processing. Employing *PySpark* enables the distribution of computational tasks, overcoming the limitations imposed by the memory capacity of an individual machine. This significantly reduced the run time required for later data prediction and analysis.

We employed *Kaggle* API to first download data into Google Drive and then leveraged Spark DataFrame to load the entire dataset directly from *Kaggle*. After that, we prepared the data and developed the predictive model in *Google Colab*.

As mentioned, given our dataset contains approximately 30,000,000 rows, which is more than sufficient for subsequent predictive analysis, our first step for the data cleaning process was removing all rows (flights) containing null data. We proceeded to determine the necessity of including each column (attribute) to avoid overfitting of model. Some columns were eliminated due to reduplicated information. For example, we excluded 'year', 'month', 'day of month' and included 'flight date' column as this single column encapsulates the information conveyed by three columns together. As previously mentioned, all flight data is grouped by year, thus, we also removed the 'Year' column. Some columns were removed due to limitations in their compatibility with machine learning models. For instance, the 'OriginState' column was selected over 'OriginStateName', as 'OriginState' contains numerical data, a unique identification codes for departing states, and 'OriginStateName' contains categorical data. The selection of numerical data reduced the need to convert categorical variables into a format compatible with our machine learning algorithm.

After the first round of the selection process, we proceeded to adapt the chosen columns into formats suitable for machine learning. For example, we transformed 'Cancelled', originally a boolean variable, into a numerical variable by assigning 'False' to 0 and 'True' to 1. As mentioned in the previous paragraph, we attempted to avoid the inclusion of categorical variables by opting for their corresponding attributes in numerical formats. However, some categorical variables lack a direct numerical representation. For example, the 'Airline' column lacks a numerical representation. To make it suitable for later model implementation, we utilized *StringIndexer* to transform this string-based data into distinct indexes.

After completing above-mentioned process in Spark, our dataset now contains a total of 21 columns/attributes.

3 DATA VISUALIZATION

The link of code for the data visualizations is provided in the Appendix section at the end of this paper.

3.1 Line Graph Visualization

Before constructing a predictive model based on this dataset, it is important for us to understand the data. We attempted to do so through data visualization. Since each of the datasets contain approximately five million rows of data, plotting the data directly may be time-consuming and may lead to memory constraints. Thus, we used the Spark for selecting, transforming and aggregating the data. Once we get an aggregated dataframe, we collected these data as Pandas Dataframe.

For our first visualization, we generated line graphs of the average daily departure delay for each month for these five years. Firstly, we selected specific columns of interest, namely ’FlightDate’ and ’DepDelay’ (departure delay, negative values for early departure). In a for loop, we first filtered the data with selected columns by month. Then, we grouped this monthly flight data by day of the month, and aggregated the dataframe by the daily average delay. Finally, we converted this into a Pandas DataFrame and created twelve visualizations for each year. We wrote the above process in a function, and called this function for each of the five years in the dataset.

The visualizations are presented below and labeled as Figure 1 through 5.

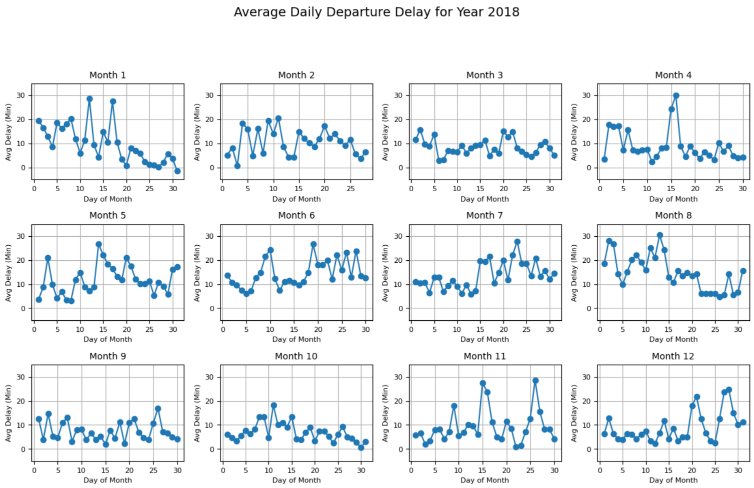

Fig. 1. Average Departure Delay from January to December 2018

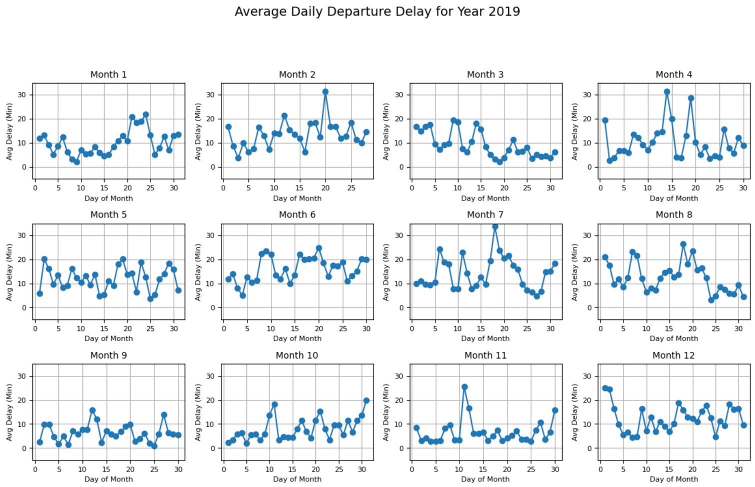

Fig. 2. Average Departure Delay from January to December 2019

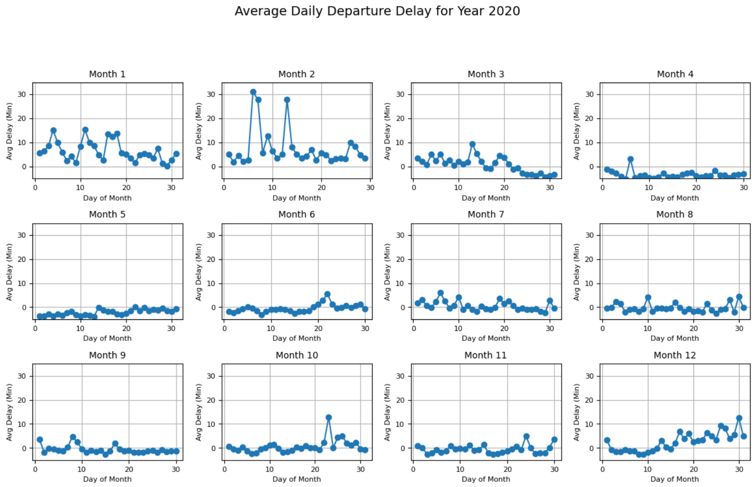

Fig. 3. Average Departure Delay from January to December 2020

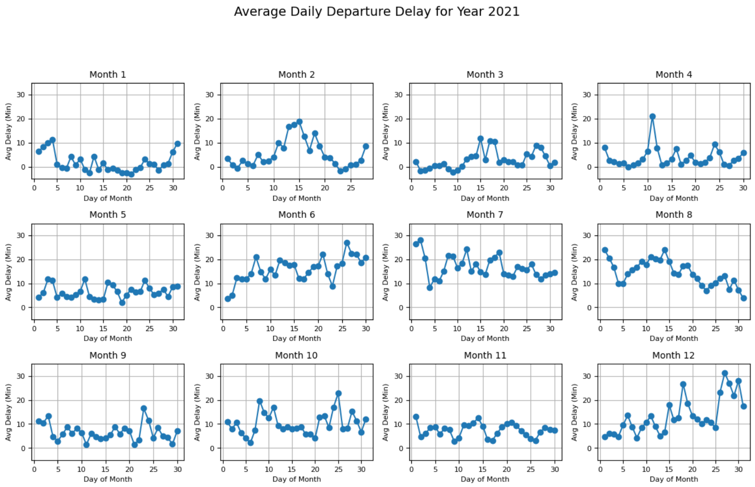

Fig. 4. Average Departure Delay from January to December 2021

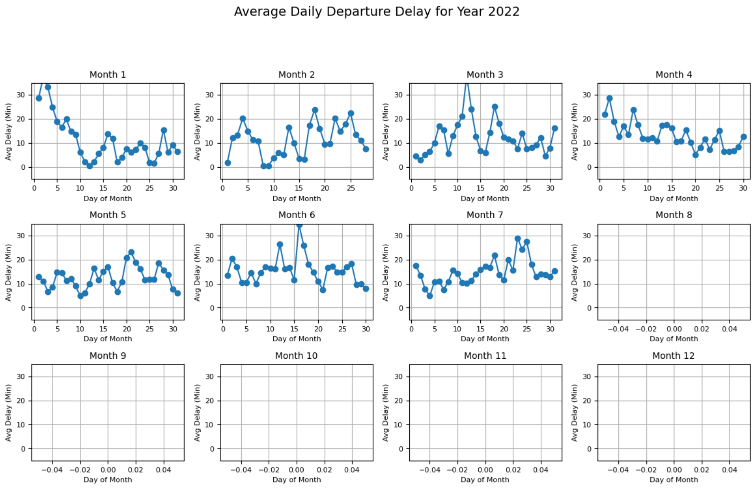

Fig. 5. Average Departure Delay from January to December 2022

3.2 Observations

At first sight, we did not discover any obvious trends in the average daily departure delays for each month. The fluctuations in delays seems to display random patterns for each plot. Therefore, we attempted to connect the identifiable characteristics of each visualization with real-world factors to understand the reasons behind the fluctuations. Through our research, we identified that these fluctuations were linked to factors such as seasonality, holidays, and extreme weather conditions. For example, in the April 2019 visualization, we observed a significant raise in average departure delay time over the Easter weekend. According to Yeager, this increase in delay time was due to severe storms across the United States, resulting in over 3,000 delayed flights (Yeager, n.d.). Our research led to the conclusion that the rise in delay times from December to February was primarily associated with blizzards during the winter months. From March to May, the delays were associated with thunderstorms, tornadoes, and heavy rainfall typical of the spring season. Delays during July to August were mainly due to thunderstorms and hurricanes.

In addition to the impact of weather on flight delays, it is important to point out that there were minimal fluctuations in average departure delays in 2020, after March.

3.3 Heatmap Visualization

For our second visualization, we generated heatmaps visualizing flight delays over a span of five years, focusing on the delays in relation to both days of the week and the time of day. The selected columns of interest included 'FlightDate,' 'CRSDepTime' (scheduled local departure time), and 'DepDelay.' We first transformed 'CRSDepTime' from 3 to 4 digits and extracted the (departure time) hour information. Then, we grouped the data by 'DayOfWeek' and the extracted hours, calculating the average delay for each group. The data was then converted into a Pandas DataFrame and we used Seaborn to create the heatmap visualization. We wrote the above process in a function, and called this function for each of the five years in the dataset.

The visualizations are presented below and labeled as Figure 6 through 10.

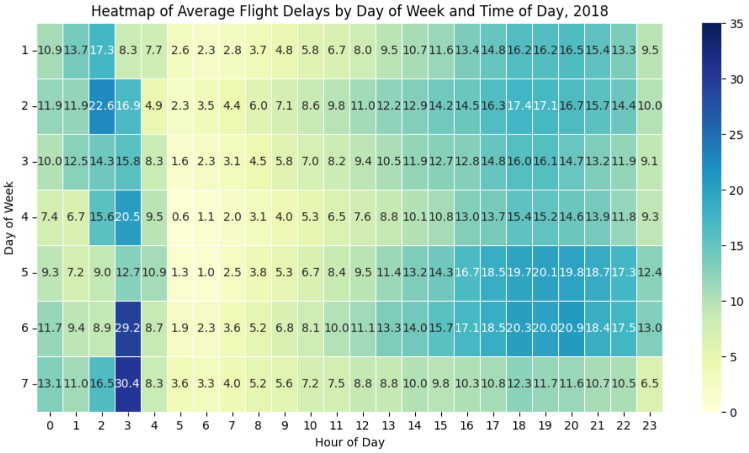

Fig. 6. Heatmap of Delays by Day of Week and Time of Day 2018

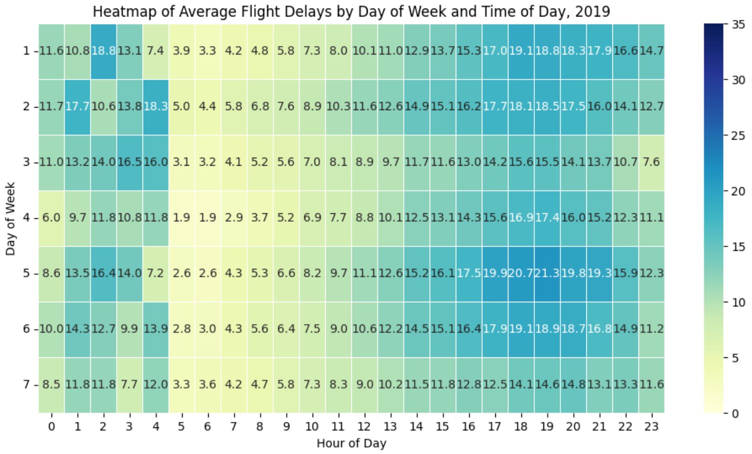

Fig. 7. Heatmap of Delays by Day of Week and Time of Day 2019

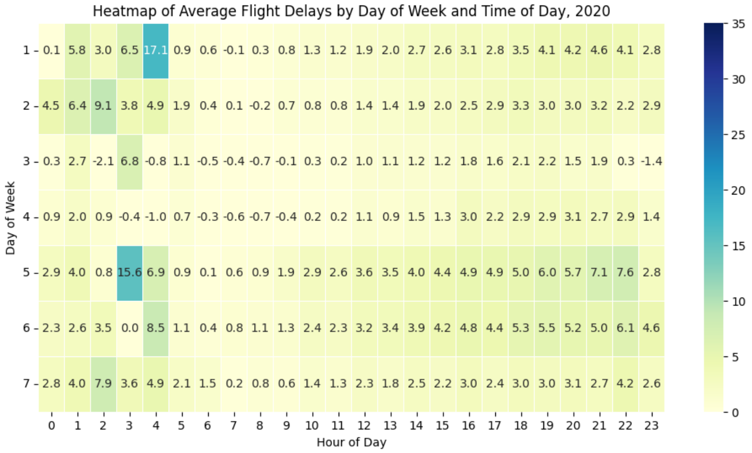

Fig. 8. Heatmap of Delays by Day of Week and Time of Day 2020

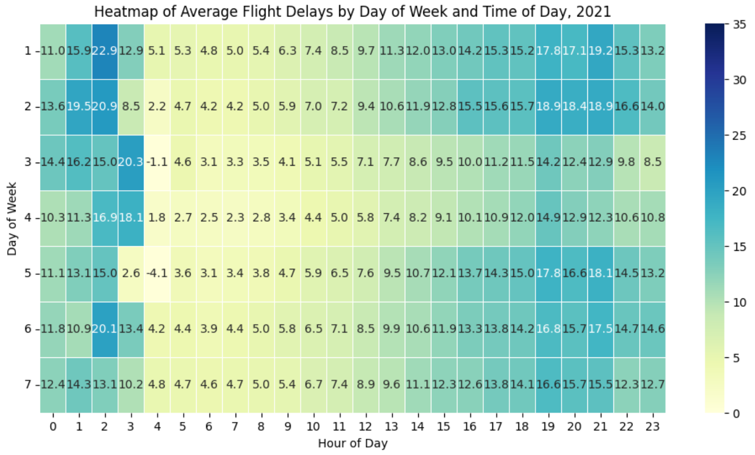

Fig. 9. Heatmap of Delays by Day of Week and Time of Day 2021

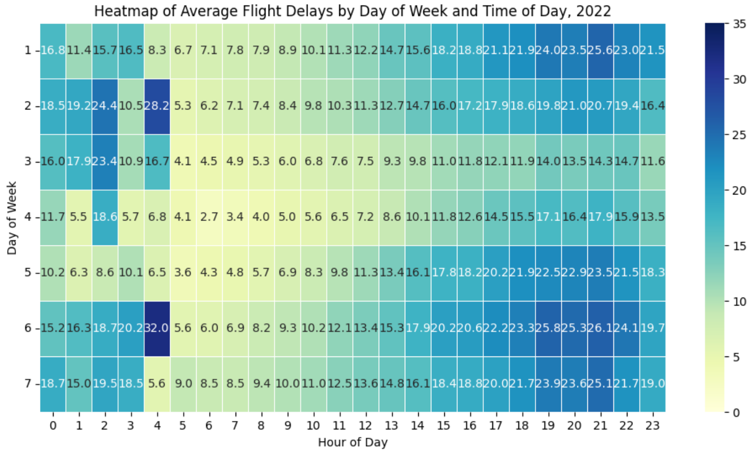

Fig. 10. Heatmap of Delays by Day of Week and Time of Day 2022

3.4 Observations

From the visualizations above, we discovered a discernible pattern: flight delays are prevalent during nighttime, particularly between 1 am and 4 am. This trend is visually represented by darker colors on our heatmaps. Delays are also quite common during weekends, spanning from Friday evening through Monday.

In addition to the observed patterns, it is important to point out the minimal flight delays in relation to both days of the week and the time of day in 2020, reflected visually by overall lighter colors on the heatmap.

4 PREDICTIVE MODEL

The link of code for the predictive model is provided in the Appendix section at the end of this paper.

4.1 Logistic Regression Model

For our predictive model, we chose to approach with logistic regression in Spark using *PySpark's RDD API*. This model demonstrates proficiency in binary classification problems. As mentioned in the Preprocessing Data section, our cleaned dataset contained boolean variables, making logistic regression a suitable choice for our model. The logistic regression outputs is more insightful than mere hard classification predictions, evidenced by the information conveyed through probabilities in output.

For our predictive model, we would like to forecast the probability of 'ArrDel15,' a binary variable that yields 'true' if the flight arrives more than 15 minutes behind schedule and 'false' if it arrives within the scheduled arrival time plus 15 minutes. Note that we have transformed the representation of this variable into a numerical format, where 'false' is represented as 0, and 'true' is represented as 1. For our logistic regression model, we first wrote the Sigmoid function, a function that transforms any real-valued number into a value within the range of 0 to 1, to map the output of linear regression onto a (0,1) scale, effectively representing the probability of 'ArrDel15'=1.0 based on the input attributes. Then, we wrote a function that estimates the model's coefficients by maximizing the likelihood function. It performed distributed gradient descent to optimize the logistic regression model parameters. This included the process of determining a set of coefficients that maximizes the probability of observing the given sample.

By using the method above, we can generate a diagram with Matplotlib to visualize the distribution of the product of weights (coefficients) and features (attributes selected in Proprocessing Data section excluding 'ArrDel15'). Given the substantial size of our dataset, we conducted model testing by randomly selecting a demo training dataset comprising 200 flight records from the combined dataset spanning five years' flight information. The scatterplot is presented below, labeled as Figure 11.

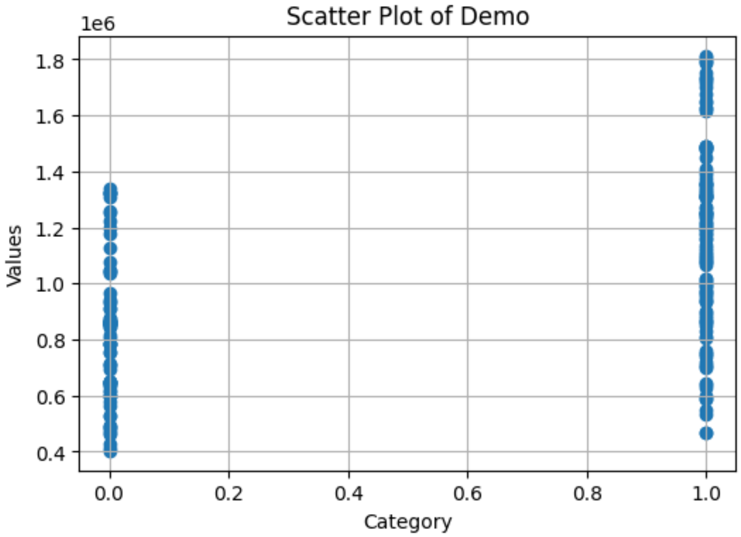

Fig. 11. Scatterplot of Distribution of weight*feature for Demo data

From this diagram, even though our algorithm is not fully optimized, we discovered our algorithms' potential to predict delay time. Therefore, we utilized a package from *pyspark.ml.classification* to conduct predictions based on our model written.

We then used *MLlib* to train the entire dataset.

4.2 Accuracy

We assessed the efficacy of our model's performance by using the *MulticlassClassificationEvaluator* in *MLlib*, calculating key metrics such as precision, recall, and F1 score. These values provide valuable insights into our model's ability to make accurate predictions on 'ArrDel15'.

The precision value of approximately 0.823 signifies that around 82.3% of the model's predictions in whether a flight would arrive more than 15 minutes late are accurate. A recall value of 1.0 indicates that the model correctly identified all flights that did arrive later than 15 minutes in reality. The F1 score, 0.74, indicates a balanced performance in terms of both precision and recall, providing a comprehensive assessment of the model's effectiveness. Based on the calculated values, we can see that our model is performing quite well, reflected through its high accuracy. The accuracy is especially high in predicting instances labeled as 'true' in 'ArrDel15', indicating our model excels in correctly identifying flights that actually arrived more than 15 minutes late, yet with some minor limitations in the accuracy of classifying flights that were labeled as 'false' in 'ArrDel15'.

5 LIMITATIONS

Our dataset has a few limitations. Firstly, in the five years of flight information, three of these years were marked by the disruptive influence of the Covid-19 pandemic, which significantly impacted flight schedules. This was reflected in both the line graph and heatmap visualizations. As mentioned in the observation of the data visualizations, there were minimal average departure delay regardless of grouped by day or by day of the week and time of day. This phenomenon is likely due to the widespread impacts of the Covid-19 pandemic. In contrast to the 2021 data, the dataset contains approximately 6.2 million rows, whereas there were only about 4.7 million scheduled flights in the 2020 data. This cancellation of numerous flights during this period led to a significant reduction in air traffic, which increased the likelihood of on-time departures for the remaining flights.

It is interesting to observe that, despite the prevalence of Covid-19 from 2019 to 2021, there appears to be no discernible difference in the visualizations of average departure delays in 2019 and 2021 when compared to years without the pandemic (2018 and 2022). This phenomenon may be due to the fact that, even though the onset of the Covid-19 pandemic was recognized in 2019, individuals and systems required time to respond to the pandemic. In 2020, their response to the pandemic is evident and can be reflected in flight operations. A significant number of scheduled flights experienced cancellations as individuals were reluctant to travel during the pandemic and governments' interference to protect their citizens from the adverse effects of the disease and to mitigate its spread. In 2021, as people developed immunity to the disease, in addition to a larger picture of the global economy recovering from substantial negative impacts brought by the pandemic, flights began to be rescheduled. In conclusion, the society and the people in the society gained adaptation in respond to Covid-19 pandemic.

Besides the Covid-19 pandemic, there were other unforeseen events like air crashes, terrorist threats, and interim governmental regulations can impact flight schedules. Yet, due to their unpredictable nature, we cannot include these factors in our model as it may induce inaccuracy to our predictive model.

For our future steps, we would like to go beyond our existing model by refining its architecture to achieve higher accuracy in predicting the occurrence of arrival delays. Furthermore, we would like to enhance our model by enabling it to predict the duration of arrival delays. This model would provide more detailed information about flight status, which is beneficial to the individuals whom may concern.

6 CITATIONS

Flight Status Prediction. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022

Yeager, M. (n.d.). Severe weather cancels, delays flights ahead of Easter travel weekend. The Arizona Republic. https://www.azcentral.com/story/travel/airlines/2019/04/18/travel-severe-weather-easter-weekend-flight-cancellations-delays/3508613002/

7 APPENDIX

https://colab.research.google.com/drive/13BfRriKF3U_cvQfH3ARXPW-Lyo8QmqRZ#scrollTo=NTE64RxqIKfw