Impact of COVID-19 Pandemic on US Flight Traffic

Mithila Sivakumar
msivakum@yorku.ca
York University
Toronto, Canada

Tasneem Naheyan
tnaheyan@yorku.ca
York University
Toronto, Canada

Armin Gholampoor
arminir@yorku.ca
York University
Toronto, Canada

ABSTRACT

In this project, we examine the impact of the Covid-19 pandemic on domestic flight traffic in the United States. The Covid-19 pandemic has brought unprecedented negative impact on the aviation industry, leading to a sharp decline in the number of flights and passengers. To assess the extent and duration of the pandemic's impact on US flight traffic, we analyze the data from Transtats [1] library. In addition to highlighting annual trends in flight traffic, our findings provide insights into the effects of the pandemic on the aviation industry and can inform policy decisions aimed at mitigating the impact of future pandemics on air travel. All project code and related materials are available here: https://github.com/tasn19/DataAnalyticsFlightPatterns.

KEYWORDS

massive datasets, flight traffic analysis, visualization, tableau, machine learning models, COVID-19 pandemic

1 INTRODUCTION

The United States of America's commercial airline industry is one of the busiest, fastest evolving, and most diverse airline industries in the world. Over the last decade, with the advancement of aircraft technology there has been a steady increase in revenue of the airline industry. This growth halted in 2019 with the onset of the COVID-19 pandemic. The pandemic disrupted many businesses globally and the airline industry was no exception as travel was restricted or banned everywhere. In this project, we have analyzed air traffic data over a five year period, with a focus on flight delays and airline statistics before, during, and after the peak of the pandemic.

Figure 1: Travelers navigate a security checkpoint at Denver International Airport on November 22, 2022 in Denver, Colorado. Source - Bloomberg [4]

Historical flight data is useful because it enables airlines and governments to identify patterns in flight demand and plan to allocate resources accordingly. Awareness of peak flight times and popular routes can help airports to prepare their staff for busier times to prevent crowding and long line-ups at screening and security checkpoints, as seen in [1]. Data analysis in this domain can help airlines decide whether to increase flights for popular routes or during peak times. Commercial airlines incurred major losses due to the pandemic. Identifying trends in flight demand in pre-pandemic years and projecting this to future post-pandemic years can help them decide the best route for recovery.

2 PROBLEM DEFINITION

In our project, we analyze patterns in our selected US flight records dataset and perform preliminary experiments to forecast the future number of flights per day given historical data. Our dataset, obtained from Kaggle [3], contains detailed records of US flights between 2018 and 2022, including information on airlines, delays, cause of delays, origins, and arrival destinations. This dataset has been extracted from the Marketing Carrier On-Time Performance (Beginning January 2018) data table of the "On-Time" database from the TranStats data library [1]. Flight records are available from before and after the start of the COVID-19 pandemic, making the dataset suitable to observe the impact of the pandemic on the airline industry, in addition to finding general trends in commercial flights across years. Some questions we answered through our data analysis are as follows:

- *What are the peak times/months for flights?, Which airports and routes are busiest?*
- *Which airlines are the most popular?, Which airlines experience the most delays?*
- *How has the amount of flights, flight delays, and cancellations changed over the years, and before and after the pandemic?*
- *Unusual trends/patterns uncovered from the dataset*

3 RELATED WORK

AI Czerny et al [8], have analyzed the impact of Covid-19 in the Chinese aviation industry in the paper "Post pandemic aviation market recovery". They review the recovery pattern influenced by the Chinese government's aviation policy in the hope that their findings would help improve the aviation policies elsewhere. In contrast to our analysis, this study focuses only on the recovery pattern used by the Chinese government rather than the impact of the pandemic.

The study performed by Lap Hang Chung [7] on the controlling measures adopted by various airports (specifically 12 selected airports all over the world) during outbreaks over the last decade and not just on the Covid-19 pandemic concludes that more efficient airport pandemic control plans cause less severe economic

impact on airports during pandemic and recommends a streamlined approach that improves overall effect of pandemic control while minimizing economic impacts on airport businesses. Our study is focused only on the United States domestic air traffic and the impact of the Covid-19 pandemic.

Another case study performed by Sandro Nižetić [9] analyzes the air transport mobility in Europe (EU), based on available data from the relevant sources associated with the airline industry. Compared to our study, data were analyzed in specific periods from January to April of 2020, which corresponded with the start of the pandemic in the EU and later in its full development. Also this study focuses on only two selected airports in Croatia.

In kaggle, the source of our selected dataset, there are only five solutions posted for this dataset. All of the five solutions provide visualizations in the form of bar charts, heat maps, histograms etc. depicting delays and cancellations. In our solution, we have used Tableau [2] for visualizations and also performed predictive analysis using ensemble techniques.

4 METHODOLOGY

We used Python, PySpark, Pandas on the Google Colab IDE to load and preprocess the dataset, and to run predictive analysis algorithms. Tableau was used for data visualization. The stages of our methodology are described in the subsections.

4.1 Data Mining

We have performed the following data mining and data pre-processing techniques to prepare the data for predictions and visualizations

Data Cleaning: In this initial stage, we removed unwanted columns in our dataset. These columns were either redundant or did not contain any useful information. Moreover, we had to separate some of the columns into multiple other columns to make the data suitable for processing. For example, we divided the time columns into hours and minutes to make the data more interpretable for processing and for the models.

Encoding Techniques : For processing the categorical data, we used different encoding methods. Since we had many unique values in the string columns, we decided to encode them with frequency encoding method instead of one-hot encoding. However, for our small categorical columns (like GeoOrientation and Pandemic) we used one-hot encoding.

Data Scaling : After trying a few scaling methods (such as MinMax, Robust, and Standard) we concluded that the Standard scaling method best suits our purposes and used this method to scale our data.

Feature Engineering : We engineered new columns from the existing columns to provide more useful information to our models. For example, we added the 'Pandemic' column which says whether a flight was before, during, or after the peak pandemic period. As another example, we created the GeoOrientation column to account for the geographical orientation of each flight e.g. flights originating in states located in the Northeast region of the United States were labelled as 'Northeast' in the Geoorientation column.

Challenges : Due to the large amount of data, we could not load the data with Pandas, the most common data processing package in Python. Hence, we had to use PySpark to load our data. However, after defining preprocessing functions over our data, the loading time was still too much. So we had to divide the dataset into smaller pieces and preprocess each of them individually.

While the dataset contains some information on the cause of delays, this information was missing for a large portion of the flights, so we were unable to include it in our analysis.

It is worth mentioning that handling missing values is tricky in this dataset. This is due to the fact that some of the columns are highly related to each other and null values do not mean a missing value in all cases but in fact, it implies some hidden information about the flight. For example, we had a few null values in the 'DepDelay' column, the column containing the number of minutes the flight's departure was delayed. Upon closer observation, we realized that these null values are for the flights that were canceled. Consequently, if we wanted to naively remove all flights with null values, we would lose all of our canceled flight data which is not optimal. Instead we replaced these null values with a special value not already existing in the 'DepDelay' column. The column contained negative values for flights that departed early. So we set these negative values to zero because another dataset column already captures the early departure information and set the null values for cancelled flights to -1.

4.2 Data Visualizations

To visualize our preprocessed data and identify trends, we opted to use Tableau as it is the market-leading choice for modern business intelligence. The Tableau platform is also known for taking any kind of data from almost any system, and turning it into actionable insights with speed and ease. Since our dataset is massive (29M rows) we identified Tableau as our best resource for visualizations. Figure 2, created with Tableau, shows domestic flight arrivals at US airports in the year 2018.

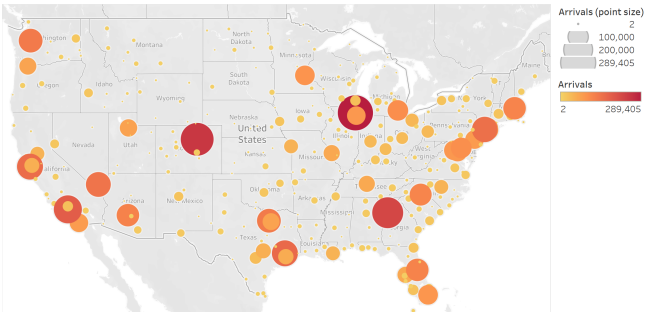

Figure 2: US map showing domestic flight arrivals in 2018.

Even though we applied multiple preprocessing techniques to the data as explained in the previous sections, we had to create a few calculated columns in Tableau for the purpose of effective visualizations. For example, we created a calculated column to rank the airlines (top 10) and applied it to the filter in Tableau so the chart will display only the top 10 airlines. Similarly, we created a column to calculate the percentage of cancellations and delay over all flights for a particular airline as this facilitates fair comparisons.

4.3 Predictive Analysis

We performed preliminary experiments in predictive analysis using our preprocessed data and select features to forecast the number of flights per day in future years. Due to limited resources and the time constraint, we ran experiments on a smaller condensed set of data generated by combining flight records for each day. We used the date, average delay of flights in minutes, total cancellations in a day, total diversions in a day, whether the day is pre-, during or post-pandemic, and number of flights in a day as input features to our model. Number of flights in a day is the target variable.

We tested the performance of two models: XGBoost [5] regressor and random forest regressor. We used these models and related functions from the scikit-learn Python library. XGBoost regressor is an ensemble method that uses gradient boosting to iteratively train decision trees on residuals, or the difference between predicted and actual values, of previous trees. Random Forest regressor is also an ensemble method; it is a simpler model than XGBoost regressor in which each tree is built independently using a random subset of features.

To use XGBoost or Random Forest regressors for time series forecasting, the data is first converted into a supervised learning problem with past values of the target variable as input features (and other relevant variables) and predicted values of the target variable as output.

We used walk forward validation to evaluate the model's performance and obtain the best model. Walk forward validation involves iteratively training the model on a rolling basis using a training set that includes historical data up to a certain point and evaluating it's performance on a test set consisting of subsequent time steps. It is more suited to evaluating time series forecasting models compared to other methods such as cross validation because it assesses the model's performance in a progressive manner and accounts for changes in underlying patterns in the data over time. We split our dataset into a training set, and a test set containing data in the next time interval, train the model, predict and validate on the test set. In the next step the training set is updated to include data from the subsequent time step and the process is repeated.

We applied standard scaling to the data to normalize it to have a mean of zero and standard deviation of one. Normalization helps reduce the non-stationarity in data, which is important for time series forecasting. Models will produce more accurate predictions if trained on stationary data - data for which statistical properties (mean, variance, covariance) are constant over time.

5 RESULTS

5.1 Data Analysis and Visualizations

We created two dashboards in Tableau. The first dashboard illustrates historic flight data trends for the number of flights, cancellations and delays. Figure 3 presents the top 10 airlines ranked by the most number of flights for all 5 years from 2018-2022. It can be inferred that Southwest airlines is in the first position with almost 5M flights. Similarly figures 4 and 5 represent the top 10 cancellations and top 10 delays over the 5 year period. Note that we have expressed this in percentage rather than real numbers to accurately reflect the value and make comparisons among airlines possible. The dashboard has an option to filter the data in these three charts by year. The viewer can then interact with the charts and see how the performance of the airlines varied over the years 2018-2022.

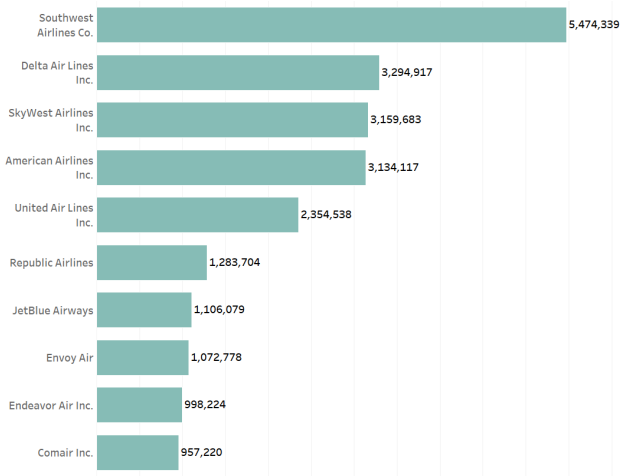

Figure 3: Bar chart showing Top 10 airlines with the highest numbers of flights for the period 2018 - 2022.

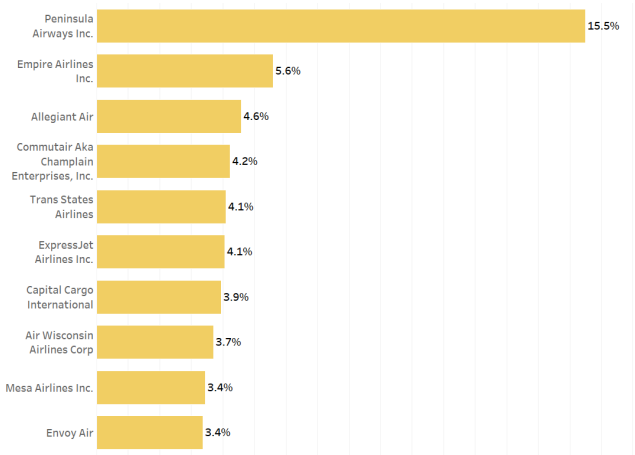

Figure 4: Bar chart showing Top 10 airlines with most cancellations for the period 2018 - 2022, where cancellations is the percent cancelled flights out of all flights for a particular airline.

The tree map in Figure 6 visualizes the number of flights per month in each year. The busiest months of each year have a darker blue color. It can be seen that the year 2019 was the boom of airline industry with all months having a high number of flights. This trend continued until March of 2020 after which travel in the US was limited due to the Covid-19 pandemic. Again, the number of flights started to increase from March 2021 onwards (after the peak pandemic time when travel started to pick up) and it continues to do so till July 2022.

We created an interactive map of the United States to show the busiest routes. For example, in the figure 7, we can see that the

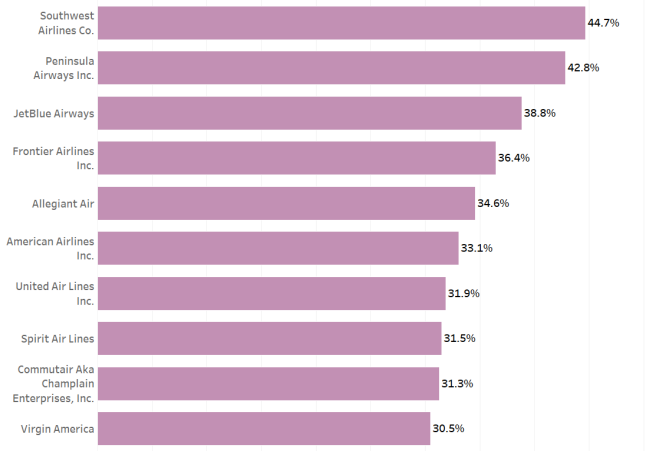

Figure 5: Bar chart showing Top 10 airlines that experienced the most delayed flights for the period 2018 - 2022, where delays is the percent delayed flights out of all flights for a particular airline.

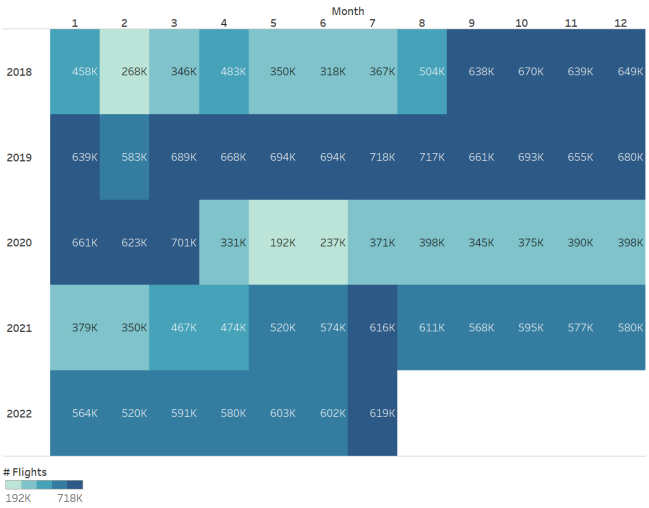

Figure 6: Tree map showing the number of flights per month per year. Light color indicate low value and darker color indicates high value.

route from Lexington, Kentucky to Chicago, Illinois is one of the busiest routes with the highest number of flights. The viewer can select different origin airports and see air traffic out of it using this map.

After visualizing the historic flight data for the 5 year period, we did an in-depth analysis of the impact of the pandemic on flight traffic. To do this, we labelled flights as 'On time', 'Delayed' or 'Cancelled' according to flight status. A total of ***65.9 percent of flights were on time, 31.4 percent delayed and 2.61 percent cancelled*** over the 5-year period.

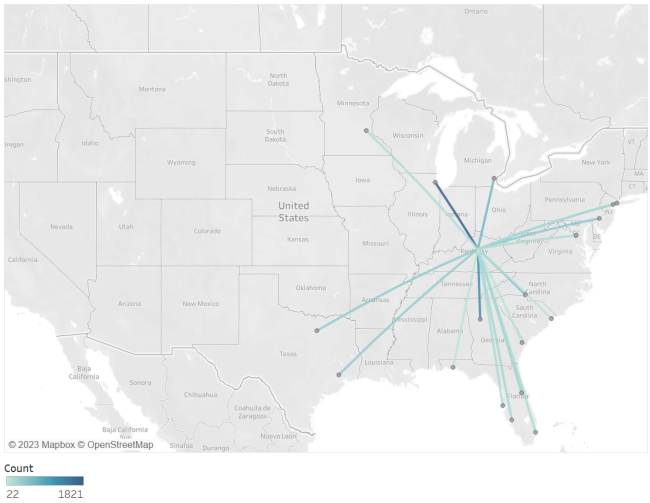

Figure 7: Image of Tableau interactive map of the United States showing the routes out of LEX airport in Kentucky. Darker colors indicate greater number of flights on that route. On the Tableau dashboard, different origin airports can be selected to see different routes.

To analyze the performance of airlines before, during and after the peak of the Covid-19 pandemic, we split the 5-year period (2018-2022) into three time intervals: ***"Pre-pandemic"*** corresponding to the time before March 2020, ***"Pandemic"*** corresponding to the period between March 2020 and March 2021 and ***"Post-pandemic"*** which is from April 2021 to July 2022. Figure 8 shows how the proportions of all flights that were on-time, delayed or cancelled varied in the three time intervals.

We spotted an unusual pattern in the ***"Pandemic"*** period. We would assume that during the peak of the pandemic the number of flights that were on-time would be less and more flights would be delayed e.g. due to increased pre-boarding procedures introduced due to COVID-19 and confusion due to changing travel restrictions. However, as we can see from the stacked chart, the data says otherwise. There was an increase in the on-time flights percentage (almost 12 percent) when compared to the pre-pandemic period which resulted in a lower percentage of delayed flights. This unusual pattern could be due to many factors such as drastic reduction in the number of passengers owing to travel measures preventing people showing symptoms from boarding flights or people choosing not to travel because of health concerns. Lower passenger traffic may have made it easier for the airports and airlines to manage departures and arrivals, It can also be seen from figure 10 that the number of cancellations are also 4 percent higher than the average. The proportion of cancellations is higher, likely as a result of some routes being cancelled during that period.

We created another stacked bar chart that can be used as an indicator of airline performance. Figure 9 reflects the performance of airlines during the pandemic period. For example, from this chart ***TranStates Airlines*** has almost the same percent of cancellations as the on-time flights; this indicates that their performance during this period was suboptimal.

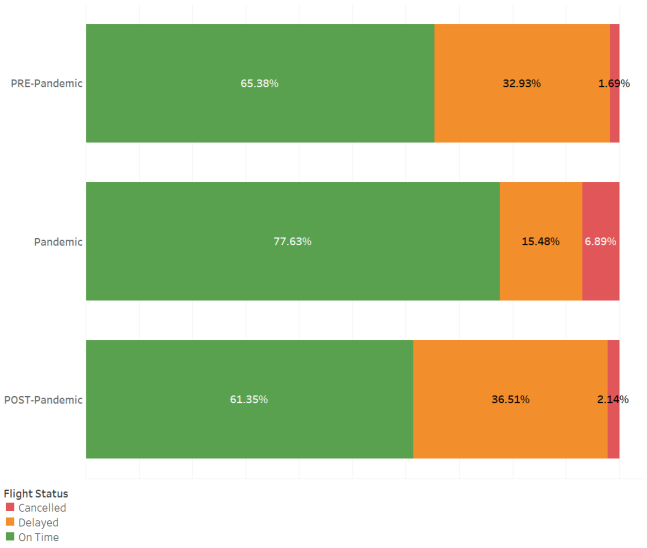

Figure 8: 100 percent stacked bar chart showcasing the 3 flight statuses (On time, Delayed and Cancelled) percentage during the three time periods Pre-Pandemic, Pandemic and Post-Pandemic.

Figure 9: 100 percent stacked bar chart indicating the airlines performance during pandemic period. This chart displays airlines whose average delay and average cancellations are higher than the 5-year average.

Death of Airlines: From our dataset, we inferred another unexpected pattern: some of the airlines went out of business after the pandemic. In our data exploration stage we noticed some airlines appeared in earlier years but were missing from the records in later years. Upon investigation, we discovered these airlines had actually gone out of business or merged with a larger airline. If we take the same example as above ***TranStates Airlines***, from figure 10, we see that there were around 466 flights cancelled per 1000 flights during the pandemic after which they went out of business, indicated by the blank space post-pandemic. From figure 11, which shows average delay per flight in minutes, it can be seen that for the same ***TranStates Airlines***, the average delay was actually reduced during the pandemic period, which might project a wrong impression that the airline is doing well if considered independently without taking into account other factors e.g. cancellation statistics, when in fact, the airline was not performing well. ***Compass Airlines*** followed the same trend. Therefore, delay times alone cannot be used as an effective attribute for determining airlines going out of business. There are other factors such as revenue, fuel consumption etc that play a role in an airline ceasing operations, which we need to incorporate in order to arrive at a strong conclusive result.

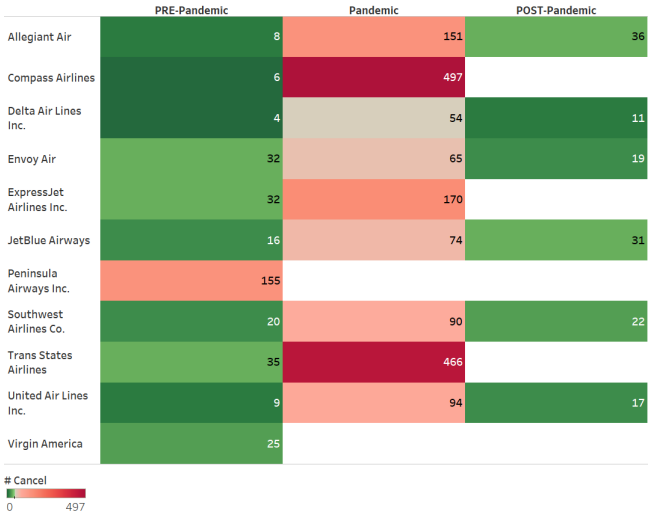

Figure 10: Number of flights cancelled per 1000 flights for selected airlines (best and worst performance) for each of the three periods pre-pandemic, pandemic and post-pandemic.

5.2 Predictive Analysis

The mean squared errors in our predictions of number of flights calculated on validation sets of our best XGBoost regression model and random forest regression model were 0.050 ad 0.038 respectively. Qualitatively the models performed similarly (Figures 12 and 13). The figure shows historical number of flights from 2018 to August 2022 and forecasted values from August 2022 to May 2023. Both models underpredict the number of flights and loosely follow the pattern seen in early 2018 and early 2021. However since a very small training set is used here (total dataset used for predictive analysis contained flight information for 1672 days), the results are expected to be inaccurate. Further work to improve forecasting is discussed in section 7.

6 CONCLUSIONS

Through our data analysis, we were able to identify trends in flight patterns across airlines and years. With our Tableau dashboard, we compared number of flights, percentage delays and cancellations across years to identify peak travel months and the best year for airlines in terms of number of flights. We explored flight patterns before, during and after the peak of the COVID-19 pandemic. A

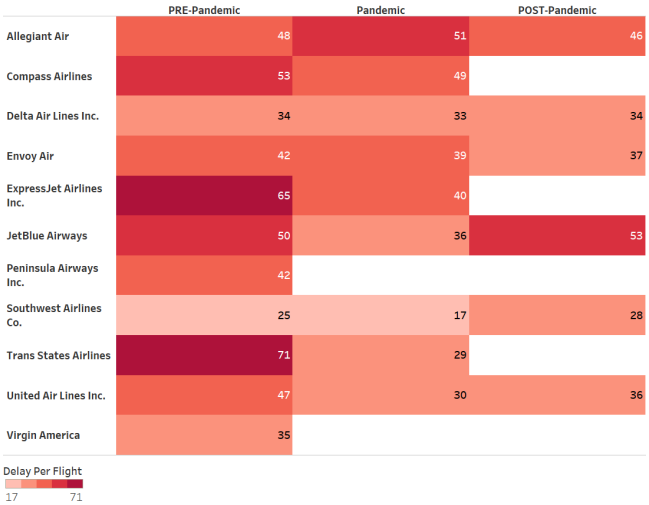

Figure 11: Average delay per flight in minutes for selected airlines (best and worst performance) for each of the three periods pre-pandemic, pandemic and post-pandemic.

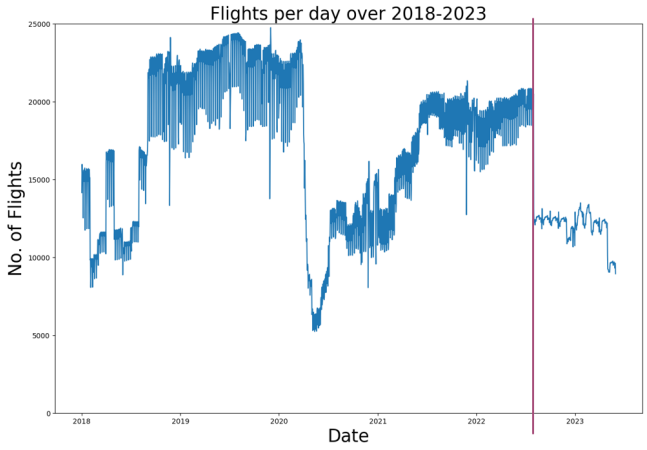

Figure 12: Results of forecasting number of flights using XGBoost regressor from August 2022 to May 2023 given historical data from January 2018 to July 2022. Plot after red line shows predictions.

key finding was some airlines went out of business during the pandemic period despite having less delays, and this highlights the importance of analysing several flight attributes before drawing conclusions. We created a chart (figure 9) which depicts which airlines have percentage cancellations and delayed flights higher than respective thresholds. In our analysis we found that some airlines (TranStates, Peninsula, Compass, ExpressJet) that crossed these thresholds went out of business in consecutive years. Thus, this chart could be a useful indicator to airlines, alerting them when they should make improvements or changes in their strategy to maintain performance.

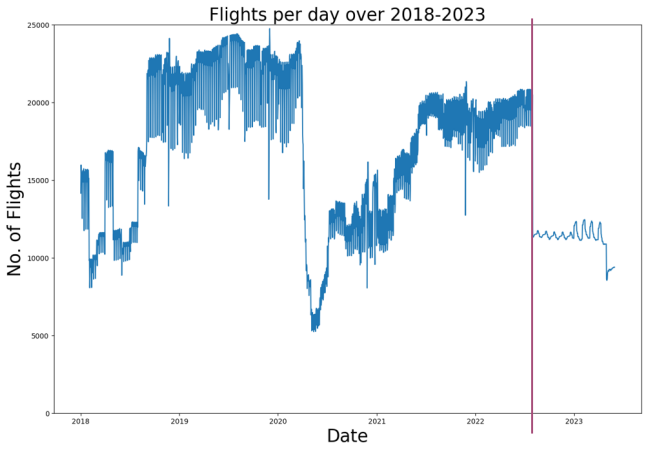

Figure 13: Results of forecasting number of flights using random forest regressor from August 2022 to May 2023 given historical data from January 2018 to July 2022. Plot after red line shows predictions.

However factors such as delay time, number of cancellations and diversion are not enough to understand all trends in observed flight patterns. Several other factors such as revenue, operating costs, cause of delays, ticket prices, airline market share, etc influence airline performance and more data needs to be considered for conclusive results. To summarize, we preprocessed and prepared our dataset for predictive analysis and ran initial experiments to forecast the number of flights in future years. We implemented XGBoost and random forest regression models, but were limited by data in generating accurate predictions. In addition to training the models on larger datasets, there are other modifications we can make to improve our results.

7 FUTURE WORK

One option to improve forecasting results is to use a model that is more suited for time series analysis than XGBoost and random forest regression, such as the Autoregressive Integrated Moving Average (ARIMA) model [6]. This model is useful for forecasting data that shows seasonality. It is a combination of three models:

- Autoregression (AR): this model predicts future values of a variable from past values,
- Integrated (I): this model differences the time series model to make it stationary. Differencing is a method to remove seasonality in data by subtracting the current values from previous values,
- Moving average (MA): this model uses past prediction errors to predict future values.

We attempted to implement an ARIMA model to predict future number of daily flights, however to generate accurate predictions we would need to fine-tune the model by adjusting the parameters. To determine the best set of parameters, we should apply different statistical techniques, such as autocorrelation plots, which show the degree of similarity of a time series and a lagged version of it, and domain expertise and knowledge on potential causes of

patterns in the data. Time constraints prevented us from performing this analysis so we leave fine-tuning the ARIMA model for flight forecasting as future work.

Another direction of future work would be to increase our dataset to include more historical data beyond 2018. This data can be extracted from the Transtats [1] library. To get a better insight on flight patterns, in addition to the features available in the dataset used in this project, we also need to find sources of other relevant input features e.g. revenue, operating costs, airline market share, as these all impact the performance of the commercial airline industry. With this additional data, it could also be possible to forecast trends other than number of flights, such as the likelihood a flight from a certain airline will be delayed, whether an airline might go out of business, and flight traffic patterns in the event of another pandemic.

REFERENCES

[1] Bureau of transportation statistics.
[2] Data analytics. https://www.tableau.com/.
[3] Flight status prediction. https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022?select=Combined_Flights_2020.csv.
[4] Thanksgiving travel is now busier than it was before the pandemic. https://www.bloomberg.com/news/articles/2022-11-23/thanksgiving-travel-2022-tops-pre-pandemic-levels-busiest-us-airports?leadSource=uverifywall.
[5] Xgboost documentation. https://xgboost.readthedocs.io/en/stable/.
[6] Jenkins G. M. Box, G. E. P. Time series analysis: Forecasting and control. *San Francisco: Holden-Day*, 1976.
[7] Lap Hang Chung. Impact of pandemic control over airport economics: Reconciling public health with airport business through a streamlined approach in pandemic control. *Journal of Air Transport Management*, 44:42–53, 2015.
[8] Achim I Czerny, Xiaowen Fu, Zheng Lei, and Tae H Oum. Post pandemic aviation market recovery: Experience and lessons from china. *Journal of Air Transport Management*, 90:101971, 2021.
[9] Sandro Nižetić. Impact of coronavirus (covid-19) pandemic on air transport mobility, energy, and environment: A case study. *International Journal of Energy Research*, 44(13):10953–10961, 2020.

A COLUMN DETAILS

In this section we provide information regrading the preprocessing of each column in the dataset.

Table 1: Pre-processed columns

Column Name	Explanation
CRSDepTime	Convert to CRSDepHour and CRSDepMinute
DepTime	Convert to DepHour and DepMinute
DepDelay	Convert to IsDepEarly and IsDepDelay
CRSArrTime	Convert to CRSArrHour and CRSArrMinute
ArrTime	Convert to ArrHour and ArrMinute
ArrDelay	Convert to IsArrEarly and IsArrDelay
Distance	Convert to ShortOrLong
DayOfMonth	Convert to TimeOfMonth
GeoOrientation	Derived from OriginState
Pandemic	Derived from Year and Month
Cancelled	Changed to binary 0 or 1
Diverted	Changed to binary 0 or 1

Table 2: Removed columns

Column Name
DOT ID Marketing Airline
IATA Code Marketing Airline
Flight Number Marketing Airline
DOT ID Operating$_{A}irline$
IATA Code Operating Airline
Tail Number
Flight Number Operating Airline
OriginAirportID
OriginAirportSeqID
OriginCityMarketID
DestAirportID
DestAirportSeqID
DestCityMarketID
DivAirportLandings
OriginWac
OriginStateFips
DestStateFips
DestWac
Removed since not helpful