US Flight Delay Predictions with ML

Fidel Garcia, Theodore Cox, Winnie Messa, Wendy Navarrete

April, 2023

1 Introduction

Delayed flights can have significant negative economic impacts on airlines, airports, and passengers which motivates the creation of more robust delay management programs like the existing ground delay program (GDP) [MS10]. Delays can also cause labor disruption by affecting personnel work periods, especially in the air transportation business, where there is strict legislation regarding flying times and off-times [DPS10]. The economic impact has motivated the previous analysis of flight delays. Airlines regulators will also benefit from a more reliable system for monitoring and enforcing compliance with regulations regarding flight schedules and delays. Our project consists of two parts: analyzing flight and weather datasets and predicting flight delays. One dataset has a year's worth of all US flight delay information [oT] and the other dataset has been gathered by scraping a weather website [wun]. After consolidating the flight delay and the weather data, we visualized the data distribution and identified periodic patterns. Using machine learning, we have predicted weather-induced airline delays and compared the accuracy of several classification algorithms, such as Decision Trees, KNN, and Random Forest.

2 Literature Survey

According to research commissioned by the Federal Aviation Administration (FAA) [FAA19] the cost of domestic flight delays puts a 33 billion dent into the US economy. [YKCK20] found that weather accounted for 70-75% of flight delays in the USA. Research such as [CSG+21], covers flight delay prediction topics like data processing, and several methods of analysis. [JS06] classifies the flight prediction analysis into five method groups including statistical, probability, network-based, operational, and machine learning. [ACL08] and [DPP12] provide insight into operational delays at an individual airport level. [CKBM16] employs weather-related features in their analysis to achieve better prediction model performance. In the context of conducting exploratory analysis [AALB+07] provides methods to detect seasonal patterns of arrival delay using average daily delay data. Whereas [TBJ08] identifies major factors that influence flight delays by applying a statistical approach to analyzing flight delay distributions. In recent years, an increasing number of papers have analyzed flight delays using machine learning approaches, among them [LMPM20] and [CKBM16] employ random forest and decision trees, and [Tan21] propose seven supervised algorithms finding that tree-based ensemble classifiers generally have better performance over other base classifiers. There are a variety of methods for flight delay prediction, which has provided insight into current prediction successes and limitations.

3 Proposed Method

We propose the following methods of innovation; consolidate two datasets (flight delay and weather) to provide a year's worth of intraday data, then perform various statistical analyses, fit machine learning models for flight delay prediction, and utilize an interactive web app using Streamlit to showcase our results. The original open-source flight delay dataset does not include significant features regarding weather. Including weather parameters in our analysis may help identify trends not visible with just flight, date, or location-related information. The open-source dataset "2015 flight delays and cancellations", is sourced from Kaggle [oT]. The weather dataset is compiled by scraping the wunderground.com website which gets its data from the NDFD (National Digital Forecast Database)

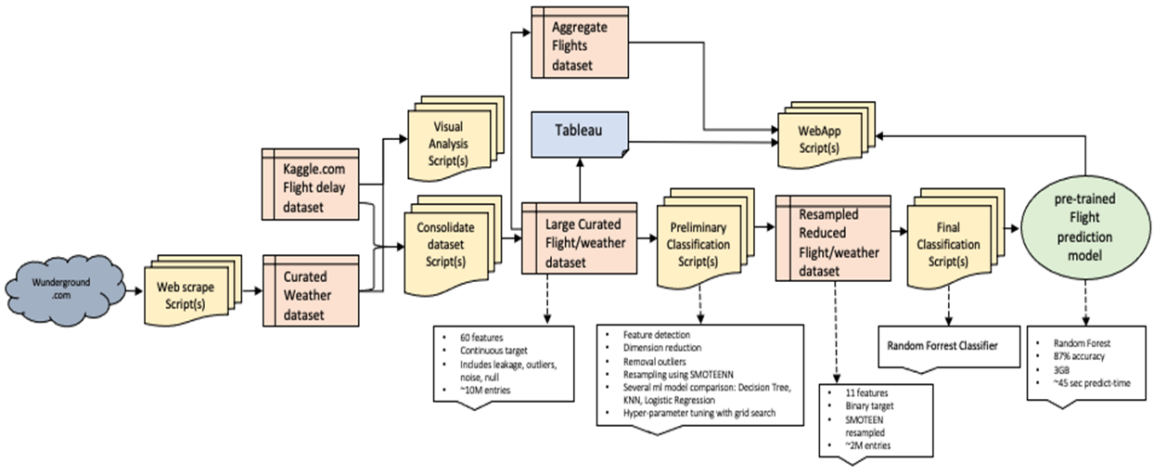

Figure 1: Process flow of sequential tasks performed to complete the analysis

provided by NOAA (National Oceanic and Atmospheric Administration) [wun]. This web-scraping approach allows a simpler process of data preparation than attempting to utilize the granular NDFD datasets for hundreds of locations and days. In classification-supervised machine learning, features are used as input variables to map to the target output. We will compare our best-performing features with Lambelho et al approach [LMPM20] where date, month, and season were used to predict flight delays. Feature abstraction and other types of input analysis will identify potential predictors of delays. A variety of classification models are used to predict whether a flight is delayed or not, followed by increasing the number of classes to predict more precise delays. Each model's performance will be measured using accuracy (a), precision (p), recall (r), and f1-score (f). The research from Lambelho et al. [LMPM20] provides insight into similar efforts to use classifiers to obtain a "prediction horizon of up to 1 day".

Further on, additional analytical steps taken include resampling, hyper-parameter tuning, and cross-validation. The use of the Streamlit app integrated with Tableau allows one to take advantage of some built-in features (filter, clean data, join, etc.) hence, less coding is needed which also contributes to a more seamless user experience. This method should provide a broader view of the entire commercial US flight delays. If we are successful in analyzing and visualizing the US airport traffic data for a full year, we can potentially gain several insights into the proximity of the driving causes of flight delays. For instance, we can identify the busiest airports and peak times, and can also uncover other patterns in flight delays, which can aid in improving flight scheduling and maintenance operations. We compare our flight delay prediction accuracy of the ML models with (a)(p)(r)(f) methods mentioned above with a flight-only dataset versus a flight-and-weather dataset. The diagram in Figure 1 shows the flow of sequential tasks performed to complete the analysis; from data collection and consolidation, statistical and visual analytics, machine learning prediction, and UI development linking everything together to successfully analyze flight delay patterns.

3.1 Data Gathering, Preprocessing, and Aggregating Datasets

As mentioned earlier, a key innovation of this project is consolidating two datasets (flight delay and local weather) into one tabular form to be fit in a supervised classification machine learning model. The flight delay data is open-source and provided by Kaggle, a community-driven dataset repository of various industries. The weather dataset was consolidated after using HTTP requests to scrape the wunderground weather history website. The wunderground website consolidates its information from NOAA and other pre-processing practices before displaying it to the public. The scraping effort was rather complex and took over 100,000 HTTP requests (data points for 365 days x 300 airports) to initially compile. This process was automated using Python and various libraries. The limitation of just using the open-source flight dataset is that there was only one feature that shows the relationship between the delay caused by the weather. The feature may not provide enough information to repeat-

edly predict delays when the weather changes. The scrapped weather dataset includes many other features for both the origin and destination locations including; temperature, dew point, heat index, relative humidity, pressure, visibility, wind chill, wind direction, wind speed, UV index, cloud cover, etc. Before narrowing the dataset down (using feature analysis, removing outliers, etc.) the dataset is about 2Gb large (0.5GB flight and 1.5GB weather).

We aggregated the dataset to reduce the size of the analysis. Aggregated datasets are a form of consolidating data by grouping it together based on some common characteristics. This process involves taking large amounts of data and combining it into smaller, more manageable datasets. Aggregate datasets may include metrics such as averages, sums, counts, percentages, and other statistical measures. Aggregating data has several advantages that can help us gain insights and make better decisions related to our project. Those advantages include; simplifying complex data sets; speeding up processing times, reducing the amount of storage needed, and removing noise and outliers. Our aggregate groups include; delays by month, delays by day of the month, and delays grouped with airline and month.

3.2 Machine Learning Models

The combined flights and weather dataset was fitted through a variety of supervised learning machine learning models to compare performance. The models are provided using the Python library scikit-learn. Feature analysis and hyper-parameter tuning with grid search cross-validation were used to further reduce the complexity and noise of the dataset. Additionally, there were several variations of supervised classification models available used in this analysis (i.e. Decision trees, KNN, Neural Networks, Random Forest). Each model was compared by their (a)(p)(r)(f) scores. The flight delay model prediction metrics were compared with the features and accuracy of the results from [CKBM16] which obtained 83.4% accuracy with 21 features encompassing data from 10 years (2005 to 2015) and 45 airports using SMOTE resampling. Their features included 3 flight-related, 6 time-and-date-related, and 12 weather-related items. Our analysis utilizes data from 1 year and 300 airports, initially starting with +50 features.

3.3 Stream-Lit Web-App

Streamlit is an open-source Python library that allows us to build powerful web apps that emphasize interactive data visualizations and machine learning models, all in one place [Tyl21]. The main advantages found using Streamlit in this project were rapid development, interactive visualizations, and ease to use. Streamlit provides a variety of tools and components that allow you to create custom visualizations, and integrate them with other libraries such as pandas, numpy, scikit-learn, matplotlib, plotly, and other common Python libraries. Additionally, the analysis conducted using Tableau was linked to the web app. Once the code was completed, it was easy to locally or remotely deploy the custom app.

4 Experiments and Evaluation

4.1 Initial Observations

Our raw dataset (i.e. flights and weather joined dataset) before preprocessing, resampling, and noise reduction had over 50 columns or features. To further analyze the dataset, histograms, and choropleth maps were used to observe trends and patterns. For the purpose of this study, a flight is considered delayed when the delta between the actual arrival and the scheduled arrival is greater than 0. The histogram in figure 2 describes the mean flight delay distribution by the airline. The top 5 airports with the least delays are all located in the West/Midwest region (see figure 3), and the top 5 airports with the most delays are located in the West region (mainly Alaska), the Samoan islands, and Delaware (see figure 4). These airports are smaller and most susceptible to flight disruptions because the carriers serving them have restricted capabilities in bad weather. The airlines accumulating the most delays are Spirit, Frontier, Hawaiian, and Jetblue. This is not very surprising, given that these are relatively low-cost carriers with tighter margins.

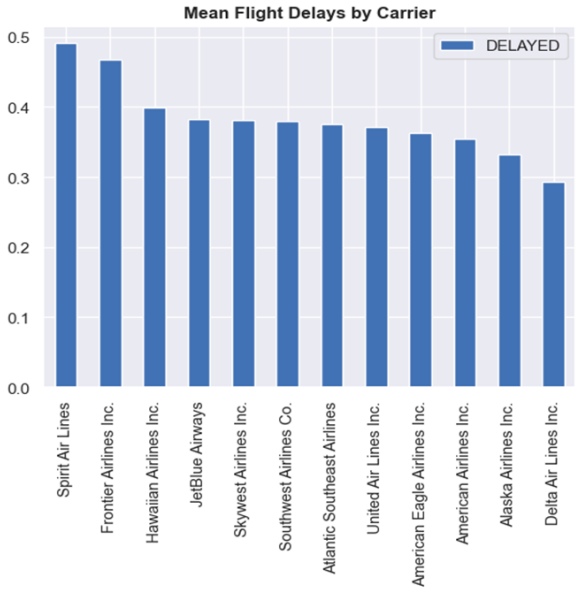

Figure 2: Mean flight arrival delays by airline.

Figure 5 shows an overview of the mean flight delay distribution per month, figure 6 shows the mean flight arrival delays by hour of the day, and figure 7 day of the week. There is not much variation when it comes to the month or day of the week, but we can observe that January, February, and June are the months with the highest amount of delays. This can probably be explained by the fact that January and February are busy travel months during the winter, due to cheaper costs, and lower temperatures can definitely affect flight departures. Not only is June a summer month, a time when there is heavy traffic volume, but it is also during peak tornado season (these are among the top causes of flight delays). Evening flights are also most likely to be delayed according to figure 6, this is correlated to heavy traffic volume because passengers prefer evening travels due to work schedules or school.

Figure 8 outlines the flight delay distribution according to origin-destination wind speed. Although there is no single maximum wind limit for planes, a crosswind above 40 mph and a tailwind above 10 mph can cause flight delays and cancellations, which explains the distribution of mean flight delays among wind speeds. Lower temperatures may cause the oil in the turbine engine to become so thick that it would be difficult to start the engine, hence causing flight delays [Wah].

4.2 Streamlit app with Tableau

In this project, the use of Tableau enabled us to allocate more time to data exploration and visualization and to reduce the time that could have been spent coding compared to the use of other visualization tools like D3.js. Figure 9 shows that in general, in 2015, Texas and California are the states with the highest cumulative sum of total arrival delayed flights. This can also be linked to adverse weather conditions and the presence of major airports in both states. Texas and California are among the top 40 states with the largest airports according to the List of Top 40 Airports in the US [Cod20]. Using Tableau, allows the user to select a month to visualize what states show the recorded highest number of arrival delayed flights. Another visualization created was a treemap, (see figure 10) which shows Southwest Airlines as the airline with the most arrival-delayed flights registered in 2015. Additionally, using Tableau provides the option for the user to interact with the chart. The user can select a state or a group of states to visualize the airlines with the most delays in that selection.

4.3 Machine Learning Models Performance

We reduced the number of predictors and improved classifier fit and prediction time by applying feature selection techniques to the original curated flight-delay and weather dataset, which resulted in dropping

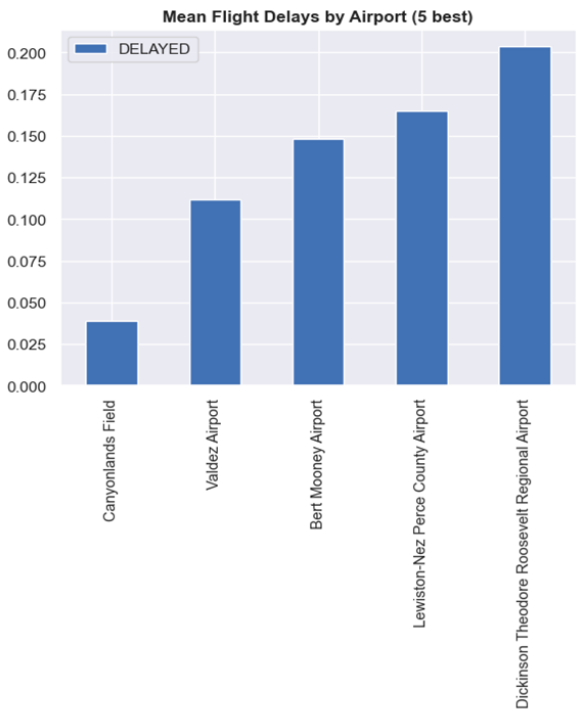

Figure 3: Top 5 best airports by mean flight arrival delays.

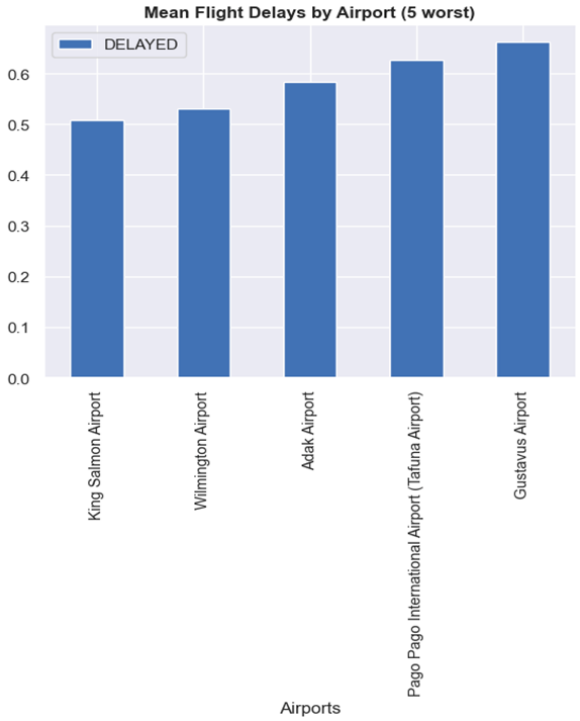

Figure 4: Top 5 worst airports by mean flight arrival delays.

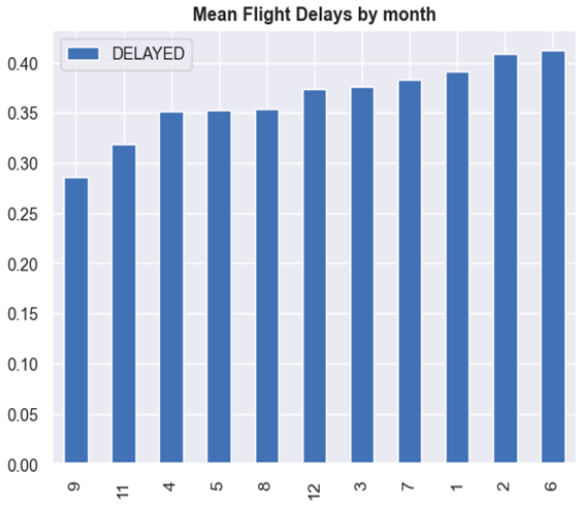

Figure 5: Mean flight delays by month.

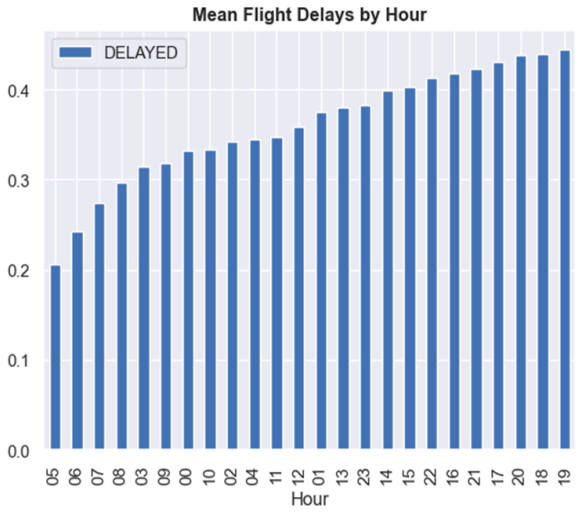

Figure 6: Mean flight delays by hour sorted in ascending order.

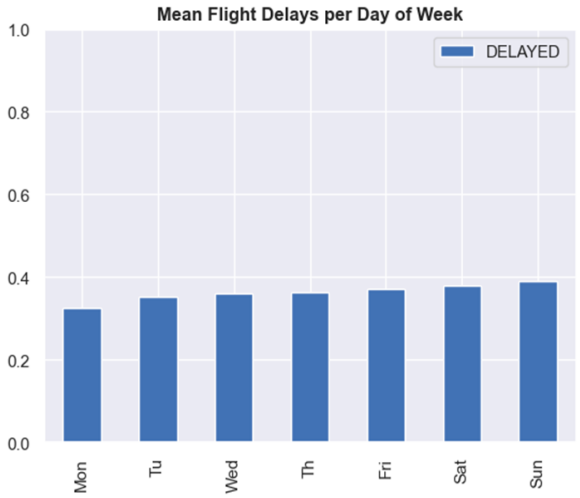

Figure 7: Mean flight delays by day of the week.

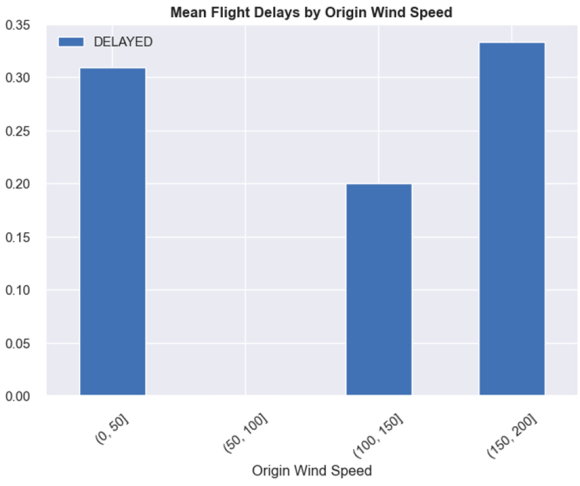

Figure 8: Mean flight delays by wind speed recorded at the origin

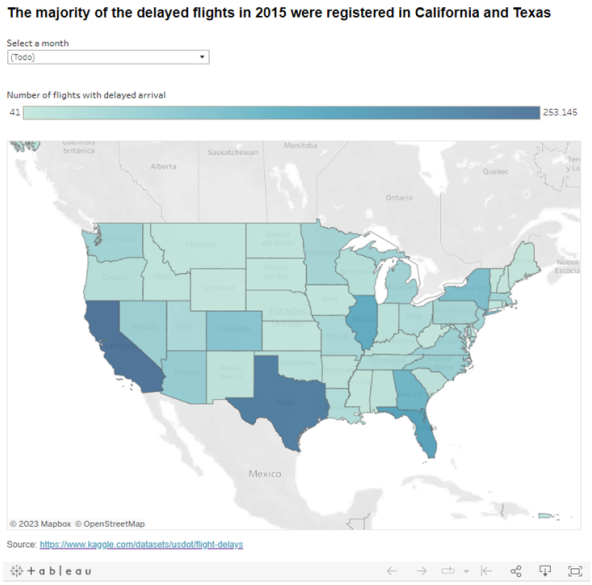

Figure 9: Interactive choropleth showing the total amount of delayed arrival flights per state. Users can filter data by month

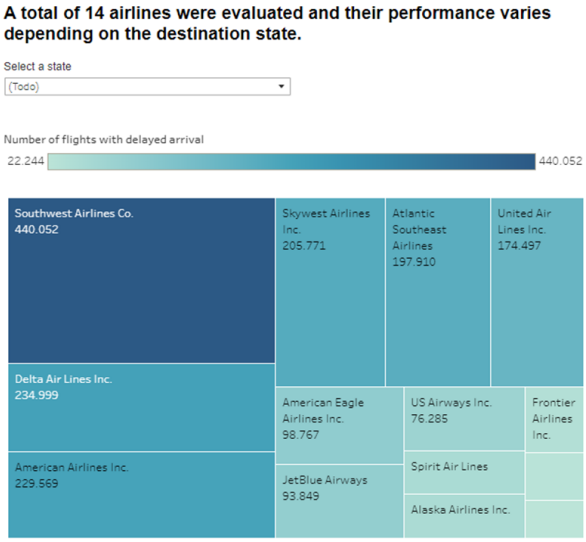

Figure 10: Interactive treemap showing the total amount of delayed arrival flights per airline. Users can filter data by state

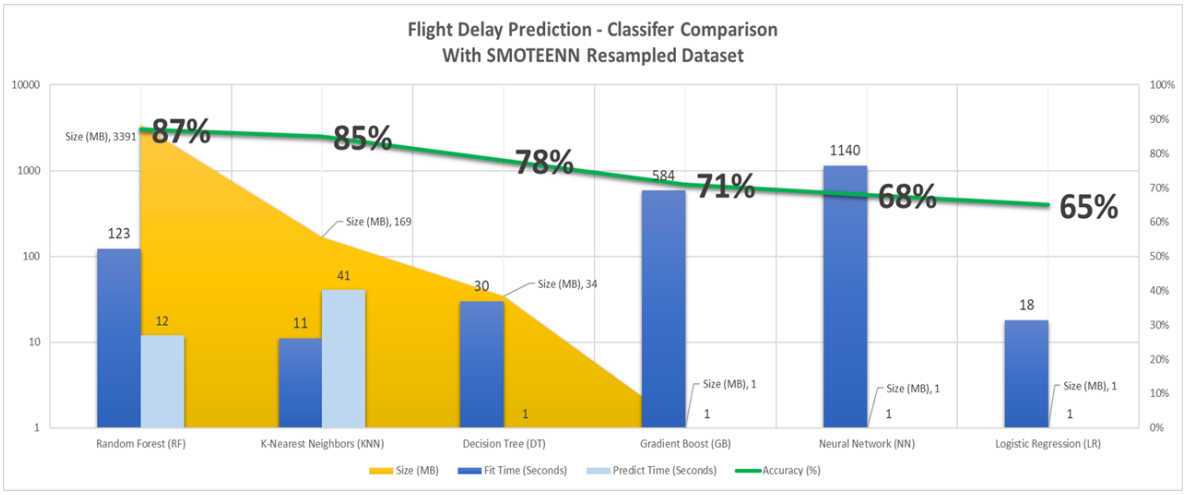

Figure 11: Classifier comparison with SMOTEEN resampled dataset.

the 60+ attributes down to the 11 most important features. We found that over half of them pertained to weather, while the remaining features were related to flight and date/time information. We also iterated fitting several classifiers and found that binary prediction for "late" vs "on-time" (where on-time included early arrivals), provided the highest level of accuracy as opposed to continuous data or a larger set of categories. We compared the results from several hyper-parameter tuned classifiers, including Decision Tree (DT), K-Nearest Neighbor(KNN), Neural Networks (NN), Logistic Regression (LR), Gradient Boosting (GB), and Random Forest (RF). To assess these models, we used three crucial metrics: Precision, Recall, and F1-Score. High precision indicates fewer false positives, while high recall indicates better identification of positive cases. F1-Score combines precision and recall by computing their harmonic mean, indicating a good balance between the two. To address imbalanced data in classification and improve prediction performance, we utilized the SMOTEENN (Synthetic Minority Over-sampling Technique Edited Nearest Neighbors) dataset resampling method, which combines SMOTE and ENN techniques (see figure 12). SMOTE generates synthetic samples to increase the representation of the minority class, while ENN removes misclassified majority samples. After performing hyper-parameter tuning with grid search cross-validation, the random forest classifier with 100 estimators and utilizing the gini criterion was the best-performing classifier (see figure 13), achieving 87% accuracy with a 70/30% split for train/test data, totaling over 2,300,000 rows of data. The other models performed closer to 55-80% accuracy (see figure 11). Final 11 features utilized:

1. Airline
2. Departure Hour
3. Scheduled Arrival Hour
4. Origin Airport
5. Destination Airport
6. Pressure at Origin
7. Pressure at Destination
8. Relative Humidity at Origin
9. Relative Humidity at Destination
10. Dew Point at Origin
11. Dew Point at Destination

Flight Delay Prediction Classifer Comparison (After SMOTEENN Resampling)							
performance metrics		RF	KNN	DT	GB	NN	LR
precision	on time	87%	87%	71%	57%	60%	57%
	delayed	86%	84%	83%	67%	72%	67%
recall	on time	76%	94%	69%	29%	49%	29%
	delayed	93%	70%	83%	87%	80%	87%
f1-score	on time	81%	77%	70%	38%	54%	38%
	delayed	90%	89%	82%	76%	76%	76%
support	on time	263450 (37% of test dataset)					
	delayed	438382 (67% of test dataset)					
total rows	test	701832 (30% of resampled dataset)					
	training	1637608 (70 % of resampled rdataset)					
accuracy (%)		87%	85%	78%	71%	68%	65%
time (seconds)	fit	10	12	30.36	584	1140	17.9
	predict	47	41	0.49	1.09	1.2	0.05
size of pkl (MB)		3391	169	34	0.12	0.04	0.001

Figure 12: Flight delay prediction classifier comparison after SMOTEENN resampling

Hyper Parameter Tuning with Grid Search Cross Validation of Random Forest Parameters			
	param 1	param 2	param 3
n_estimator	100	200	n/a
max_depth	10	None	n/a
min_samples_split	2	4	n/a
criterion	entropy	gini	n/a
min_samples_leaf	1	2	10

Figure 13: Hyperparameter tuning with grid search cross-validation of Random Forest parameters

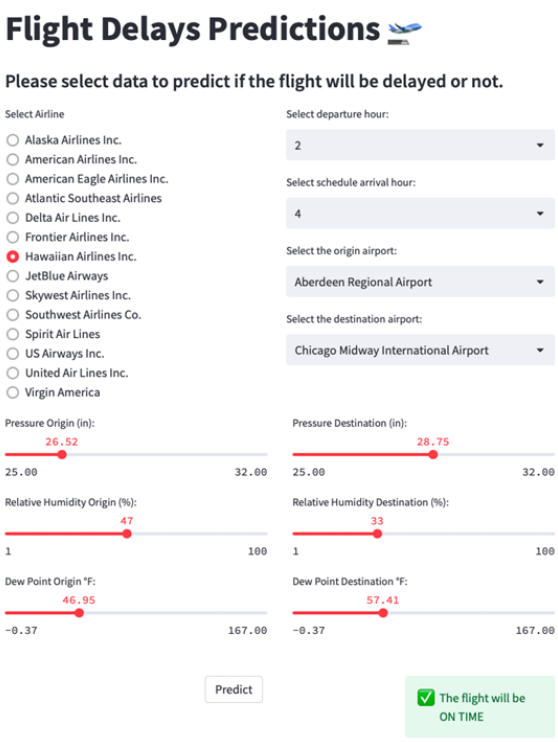

Figure 14: Interactive web app built in Streamlit. It allows the users to test different combinations of parameters in the Random Forest model

4.4 User Interface with Streamlit for Prediction

Since Streamlit provides an interactive user interface for Python code and integrates well with popular machine learning libraries, our team decided to use Streamlit to create a web application that allows users to input data and see the results of the model predictions in real-time (see figure 14). The file structure of the web app consists of 3 elements: data (CSV files), python scripts (for prediction web app and data visualization), and the prediction model pickle file (.pkl). The serialized pickle file allows leveraging the power of machine learning without the need for extensive training time and provides easy portability. It brought a significant reduction in execution time to our application allowing the users to quickly test various combinations of input parameters to determine the resulting effects on the flight delay.

5 Conclusion

The project provided valuable insights into the causes of flight delays, which are a widespread issue with national impact. Web scraping was necessary for obtaining weather data to train our machine learning models, but it proved to be a time-consuming process. Therefore, we learned that web scraping may not be the most efficient approach when working on projects with time constraints. It is important to note that since we are only using one year of data for our project, it was difficult to identify seasonality in the data (e.g. high-demand flight days could appear as outliers). We also learned that the SMOTTEEN class balancing technique gave us better accuracy overall, and without the weather predictors, our models would have had a worse performance. Our final model performed better than our goal of 83.4% [CKBM16] with only 11 features. Opportunities for improvement on the ML model include adding more data and deeper hyperparameter tuning to achieve the best result in accuracy, fit time, prediction time, and model size.

References

[AALB+07] Mohamed Abdel-Aty, Chris Lee, Yuqiong Bai, Xin Li, and Martin Michalak. Detecting periodic patterns of arrival delay. *Journal of Air Transport Management*, 13(6):355–361, 2007.

[ACL08] Shervin AhmadBeygi, Amy Cohn, and Marcial Lapp. Decreasing airline delay propagation by re-allocating scheduled slack. *Annual Conference. Boston*, 2008.

[CKBM16] Sun Choi, Young Jin Kim, Simon Briceno, and Dimitri N. Mavris. Prediction of weather-induced airline delays based on machine learning algorithms. *Digital Avionics Systems Conference (DASC)*, IEEE/AIAA(35th):Sacramento, CA, 2016.

[Cod20] World Airport Codes. List of top 40 airports in US. Retrieved April 2, 2023, from https://www.world-airport-codes.com/us-top-40-airports.html, 2020.

[CSG+21] Leonardo Carvalho, Alice Sternberg, Leandro Maia Gonçalves, Ana Beatriz Cruz, Jorge A. Soares, Diego Brandão, Diego Carvalho, and Eduardo Ogasawara. On the relevance of data science for flight delay research: a systematic review. *Transport Reviews*, 4(41):499–528. DOI:10.1080/01441647.2020.1861123, 2021.

[DPP12] Andrea D'Ariano, Marco Pistelli, and Dario Pacciarelli. Aircraft retiming and rerouting in vicinity of airports. *IET Intel Transp Syst*, 6(4):433–437, 2012.

[DPS10] Luis Delgado, Xavier Prats, and Banavar Sridhar. Cruise speed reduction for ground delay programs: A case study for san francisco international airport arrivals. *Transportation Research Part C: Emerging Technologies*, 36:83–96. DOI: 10.1016/j.trc.2013.07.011, 2010.

[FAA19] FAA. Cost of delay estimates. Retrieved February 24, 2023, from: https://www.faa.gov/, 2019.

[JS06] Liou JS. Delay prediction models for departure flights. 2006.

[LMPM20] Miguel Lambelho, Mihaela Mitici, Simon Pickup, and Alan Marsden. Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions. *Journal of Air Transport Management*, (82):DOI:10.1016/j.jairtraman.2019.101737, 2020.

[MS10] Bengi Manley and Lance Sherry. Analysis of performance and equity in ground delay programs. *Transportation Research Part C: Emerging Technologies*, 18(6):910–920. DOI: 10.1016/j.trc.2010.03.009/, 2010.

[oT] Department of Transportation. 2015 flight delays and cancellations. *Kaggle*, Retrieved February 15, 2023, from https://www.kaggle.com/datasets/usdot/flight-delays?select=flights.csv/.

[Tan21] Yuemin Tang. Airline flight delay prediction using machine learning models. *International Conference on E-Business and Internet*, DOI:10.1145/3497701.3497725, 2021.

[TBJ08] Yufeng Tu, Michael O. Ball, and Wolfgang S. Jank. Estimating flight departure delay distributions: A statistical approach with long-term trend and short-term pattern. *American Statistical Association*, 103(481):112–125. DOI: 10.1198/016214507000000257, 2008.

[Tyl21] Richards Tyler. Getting started with streamlit for data science. *Packt Publishing*, 2021.

[Wah] Madeline Wahl. This is exactly how cold it has to be to keep a plane from flying. *Reader's Digest*, Retrieved April 2, 2023, from https://www.rd.com/article/can-planes-fly-when-its-freezing/.

[wun] wunderground. Historical weather. *Weather Underground*, Retrieved March 1, 2023, from https://www.wunderground.com/history/.

[YKCK20] Maryam Farshchian Yazdi, Seyed Reza Kamel, Seyyed Javad Mahdavi Chabok, and Maryam Kheirabadi. Flight delay prediction based on deep learning and levenberg-marquart algorithm. *Journal of Big Data*, (106):DOI:10.1186/s40537–020–00380–z, 2020.