**Predicting Commercial Airlines Delays & Cancellations**

Abhilash Bhuyan, Alicia Ringo, Julio Martínez, Sritha Darbha, and Van Nguyen

Department of Information Technology & Decision Sciences, University of North Texas

DSCI 5260

Dr. Sameh Shamroukh

April 18, 2023

**Table of Contents**

[**Acknowledgements** 5](#_Toc133417746)

[**Abstract** 6](#_Toc133417747)

[**Introduction** 7](#_Toc133417748)

[**Literature Review** 8](#_Toc133417749)

[**Research Method** 11](#_Toc133417750)

[**Data** 11](#_Toc133417751)

[**Participants** 11](#_Toc133417752)

[**Procedures** 11](#_Toc133417753)

[**Measures** 11](#_Toc133417754)

[**Analysis** 12](#_Toc133417755)

[**Data Analysis** 13](#_Toc133417756)

[**Factors that impact flight delays and cancellations** 13](#_Toc133417757)

[**Predicting Flight Delays** 25](#_Toc133417758)

[**XGBoost Model** 25](#_Toc133417759)

[**Random Forest Model** 26](#_Toc133417760)

[**CatBoost Model** 26](#_Toc133417761)

[**Evaluating Best Model: (New York)** 27](#_Toc133417762)

[**Predicting Flight Cancellations** 28](#_Toc133417763)

[**Naïve Bayes** 28](#_Toc133417764)

[**Decision Tree** 29](#_Toc133417765)

[**Random Forest Classifier** 29](#_Toc133417766)

[**Support Vector Machine** 30](#_Toc133417767)

[**K-Nearest Neighbors Classifier** 30](#_Toc133417768)

[**Model Evaluation** 31](#_Toc133417769)

[**Evaluating Best Model – New York** 31](#_Toc133417770)

[**Discussion** 33](#_Toc133417771)

[**Summary of Results** 33](#_Toc133417772)

[**Answer Research Question** 33](#_Toc133417773)

[**Exclusion of the Year 2020** 34](#_Toc133417774)

[**Limitations** 34](#_Toc133417775)

[**Future Directions** 36](#_Toc133417776)

[**Implications & Importance** 36](#_Toc133417777)

[**Conclusion** 38](#_Toc133417778)

[**References** 39](#_Toc133417779)

[**Author’s Contributions** 41](#_Toc133417780)

[**Appendix** 42](#_Toc133417781)

**Table of Figures**

[Figure 1: Linear Regression – Arrival Delays 12](#_Toc133011058)

[Figure 2: Linear Regression - Cancellations 13](#_Toc133011059)

[Figure 3: Delays - Time Series Plot 14](#_Toc133011060)

[Figure 4: Percentage of Flights with Arrival Delays be Seasons 14](#_Toc133011061)

[Figure 5: Texas Day of the Month 15](#_Toc133011062)

[Figure 6: Percent of Flights Delayed Per Weekday Total 16](#_Toc133011063)

[Figure 7: Total Flights with Arrival Delays by Time of the Day 16](#_Toc133011064)

[Figure 8: Delays per Hour of Day 17](#_Toc133011065)

[Figure 9: Percentage Contribution to Flight Delays 18](#_Toc133011066)

[Figure 10: Delay and On-Time Rates by Airline 19](#_Toc133011067)

[Figure 11: 2019-2022 Texas Flight Arrival Delays 20](#_Toc133011068)

[Figure 12: Distribution of Reasons of Flight Cancellation 21](#_Toc133011069)

[Figure 13: Cancellation Time Series 22](#_Toc133011070)

[Figure 14: Top 5 Cancellation Rates by Airlines 22](#_Toc133011071)

[Figure 15: 2019-2022 Texas Flight Cancellations 23](#_Toc133011072)

[Figure 16: XGBoost Model 24](#_Toc133011073)

[Figure 17: Random Forest Model 25](#_Toc133011074)

[Figure 18: CatBoost Model 25](#_Toc133011075)

[Figure 19: Evaluation 27](#_Toc133011076)

[Figure 20: New York Model 27](#_Toc133011077)

# **Acknowledgements**

We would like to extend our most sincere gratitude and appreciation to Dr. Sameh Shamroukh (University of North Texas) for patiently guiding us throughout our research and preparing us for industry. No email or request to meet was neglected regardless of day or time and he treated us with respect and compassion. Additionally, we would like to thank our family and friends for supporting us as we spent many nights and weekends working on this project. Their understanding and encouragement propelled us forward. Lastly, we would like to acknowledge our teamwork. We shared laughs, and frustrations, and supported one another all while learning and producing a final report we are proud of.

# **Abstract**

The complexity of airline flight logistics continues to be a relevant and ever-changing topic of research. In this paper we focused on determining which factors impacted flight delays and cancellations in the state of Texas. Additionally, we employed various models to predict flight delays and cancellations before they occurred with as much accuracy as possible. Our research methodology included reviewing literature from past similar cases and by utilizing data from The Bureau of Transportation Statistics. We were able to determine that factors that caused a flight to not adhere to its schedule, seasonality, day of flight, time, location and airline carrier all impacted flight delay and cancellations. Although this subject is complex, the implication of our research is that many controllable factors exist that leaders in aviation can improve upon to reduce flight delays and cancellations. Improvements will likely save money, improve passenger satisfaction, and reduce the number of subsequent flights being impacted by delays and cancellations.

**Keywords:** Texas flights,Airline Delays & Cancellations, CatBoost Classifier Model, Seasonality, Drill Down Causation

# **Introduction**

Coordinating air travel in the US is a complex and consistently evolving process that inevitably suffers from flight changes that include gate changes, departure & arrival delays, and flight cancellations. Historically, flight changes were most associated with weather or mechanical issues. Although those factors still impact air travel, long-term problems caused by the Covid-19 Pandemic have led to an increase in flight changes. Staff shortages & overbooked flights contribute heavily to the complexity of this problem (Matthews, 2023). Unexpected flight changes are an inconvenience to customers, leaving them frustrated and at times severely delayed to their destination or unable to make their trip. Decreasing customer satisfaction hurts the airline’s brand image and places a burden on already overworked staff. Using historical & real-time data to predict and mitigate impacts of flight changes is necessary for the future of air travel.

The research question to be answered is: Which factors impact flight delays and cancellations in Texas?

# **Literature Review**

With the spread of globalization, predicting factors that impact delays and cancellations in air travel is a widely researched subject throughout the world. A flight is considered delayed if it departs or arrives 15 or more minutes later than scheduled, whereas a flight cancellation occurs when the flight does not take place at all (Yi et al., 2021).

Past research has approached this problem from an economic standpoint, by considering Covid-19 factors, determining the causes and effects of flight delays and cancellations, and from a data analytics perspective. Less frequently discussed challenges are seasonality in relationship to delays and cancellations with impact on airport efficiency (Pagoni & Koumoutsidi, 2022) and the difficulty in predicting cancellations because of lack of flight cancellation data (Wang et al., 2022).

From an economic standpoint, delays and cancellations cost airlines and customers billions of dollars while leaving their customers frustrated by the inconveniences. The airlines' response to delays and cancellations may further dissatisfy their customers and cause them to consider a different airline in the future, impacting possible future revenue streams. The ability to accurately and seamlessly mitigate these issues is crucial to an airline’s reputation (Bombelli, 2023). Refunds are rare with some airlines only issuing reimbursements from travel delays when the cause of the delay is related to weather. On average, customers lose $500-$2,000 from unexpected expenses and loss of time (Venkatesh, 2017). Common sense seems to dictate that continuing to research and improve upon flight prediction models is critical for airlines and their customers' best financial interest.

COVID-19 effects on the U.S. airline industry have been particularly damaging, with projections that seem to indicate that the industry is unlikely to return to 2019 passenger volumes before 2023–2024 (Yimga, 2021). The pandemic was damaging to the airline's workforce. Some employees opted to quit or retire while thousands of others were laid off or furloughed (Gruenwald,2020). One implication of a decrease in the airlines’ workforce is an increase in flight delays and cancellations. During the pandemic, extra precautions that included social distancing, temperature screenings, and reduced flight capacities negatively affected delays and increased the departure standard deviation by 2 minutes and arrival standard deviation by 1 minute and 42 seconds (Yimga, 2021). As travel returns to normal, it is imperative to reevaluate delays and cancellations.

Extreme weather, a causal factor in delays and cancellations, is characterized by blizzards, hurricanes, severe winter weather and thunderstorms. In their recent work, Bombelli and Sallan found an interdependency between delays and cancellations caused by extreme weather. Meaning delays caused by extreme weather in one part of the country can have a cascading effect and cause delays or cancellations in another part of the country. Adopting clear policies based on data analyses for managing resources that include flight crews during these events was recommended (Bombelli, 2023). Additionally, a European study found that on average, half of their flights were delayed over a 6-year period and there was a correlation between aircraft capacity and length of delay. Aircrafts with a greater capacity had longer delay times than aircrafts with less capacity (Zámková et al.,2022). Increasing efficiency in boarding larger aircrafts could be a suitable solution for reducing delays.

Flight delays and cancellations are not only costly and dissatisfying to customers but may also result in disruptive behavior and safety concerns for airlines employees and other travelers. Major factors contributing to these types of disruptive behavior include length of delay, time of day, density of terminal, and customer service. Mitigating every delay or cancellation is an impossible goal but having ample time to prepare for the effects of delays and cancellations can prevent disturbances and safety concerns. Airlines can plan to deploy more security, customer service agents, and disperse large crowds in advance if there is a high likelihood of disruptive behavior (Gu et al., 2020).

High levels of flight cancellations cause airlines to overbook flights. Flights that are overbooked further increase the complexity of air travel and cause customer dissatisfaction. Revenue management is used to ensure airlines maximize revenues and are dependent upon demand. Cancellation uncertainty can affect this system further illustrating the need for accurate predictive models (Dewi, 2018)

From a data analytics perspective, a challenge that is not frequently discussed is the unobserved heterogeneity in flight data (Seyedmirsajad,2022). The standard way of thinking about unobserved heterogeneity is that there may be differences within the data that are not apparent but could be impactful. When optimizing future models this will be considered.

Common models used in past research include, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (CART), and Gaussian Naïve Bayes (GNB). A Turkish study was able to predict flight delays with all the previously mentioned models with a minimum F-Score of 0.8 (ATLIOĞLU et al., 2020). Using similar models, specifically SVM and decision tree models, a classification study was able to accurately predict cancellations 90% of the time (Yanying et al, 2019). Alternatively, Neural Networks have been proven to be a suitable choice in assessing airlines data when selecting models for accuracy. A study from Bangalore, India was able to accurately predict whether a flight would be delayed or not 92% of the time using their Neural Networks (Venkatesh, 2017). Another study pointing to the effectiveness of Neural Networks was conducted in China and found that using a graph conventional neural network allowed the impact of multiple airports to be studied versus a single airport. The results of predicting delays across multiple airports yielded more impactful findings than a single airport (Cai, 2021). Less commonly observed was a vector autoregression model for forecasting economic losses. (Xuan, 2021)

To continue to improve upon past research and aid future airlines decision making, data will be evaluated pre pandemic, during pandemic, and post pandemic to better understand what factors impact delays and cancellations in Texas. Seasonality will be considered as well as all statistically significant variables.

# **Research Method**

## **Data**

The utilized dataset was downloaded from The Bureau of Transportation Statistics website. It contains flight data of US flights from 2019 to 2022 and contains 4,649,078 entries & 86 rows. The dataset was selected because the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics tracks the performance of domestic flights in the US. We felt going with this source would ensure we were getting accurate and close to real time data.

## **Participants**

Our dataset contains flight data of US flights from 2019 to 2022 and contains 4,649,078 entries & 86 rows. The 86 rows are representative of the number of total variables in the dataset. No data was collected through interviews or surveys.

## **Procedures**

Exploratory data analysis such as feature selection, data transformation, data sampling, and data cleaning was performed to determine which factors affected flight delays and cancellations while addressing problems such as massive data, lack of computational space, and noisy data. R was used to clean data; Python was used to create models and Tableau was used to visualize patterns in delays and cancellations. AI machine learning models divided into training data (70%) and test data (30%) were used.

## **Measures**

The key independent variables identified were Year, Month, Dayofweek, Dayofmonth, Flight Date, origin, destination, Carrier Delay, Weather Delay, Nas Delay, Security Delay, Late Aircraft Delay, DepTime, DepDelay, ArrDelay, ArrTime,Taxi out, wheels On, wheels Off,Taxi In, CRS\_ELAPSED\_TIME, ACTUAL\_ELAPSED\_TIME, AirTime, Distance, Distance Group. The dependent variables used were ArrDelay 15 & Cancellation. The independent variables have a direct impact on the dependent variables as they impact a plane’s ability to stay on time and remain uncancelled.

## **Analysis**

To analyze and aid in the selection of the best model, the performance of models was evaluated using appropriate metrics that included accuracy, precision, recall, and F-1 scores. These metrics were selected because they are standard metrics used in evaluating classification models. The insights provided by the models were then analyzed and compared to information gained from the literature review in order to better understand the influence of specific independent variables on the dependent variables.

# **Data Analysis**

## **Factors that impact flight delays and cancellations**

The first step in determining which variables impact flight delays and cancellations was using linear regression. Using Tableau, it was determined that the following variables were correlated with arrival delays for years 2019-2022: Dep Delay, Distance, Wheels On, Wheels Off, and Taxi In. These results can be viewed below in Figure 1.

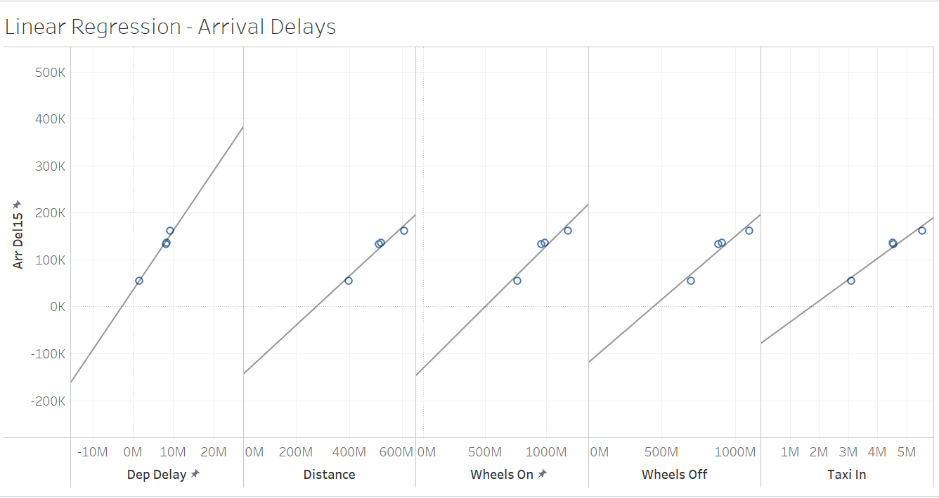


Figure 1: Linear Regression – Arrival Delays

Linear regression was also used to determine variables correlated with cancellation for the years 2019-2022 and determined to be Dep Delay, Arr Delay, Arr Delay Group, Arr Delay New, Dep Delay 15, Dep Delay Group, Distance, Wheels Off, and Taxi In. These results can be viewed below in Figure 2. Each of the listed variables for arrival delays and cancellations had p-scores less than 0.05 indicating statistical significance and r squared values greater than 0.85 indicating that the model fit was good.

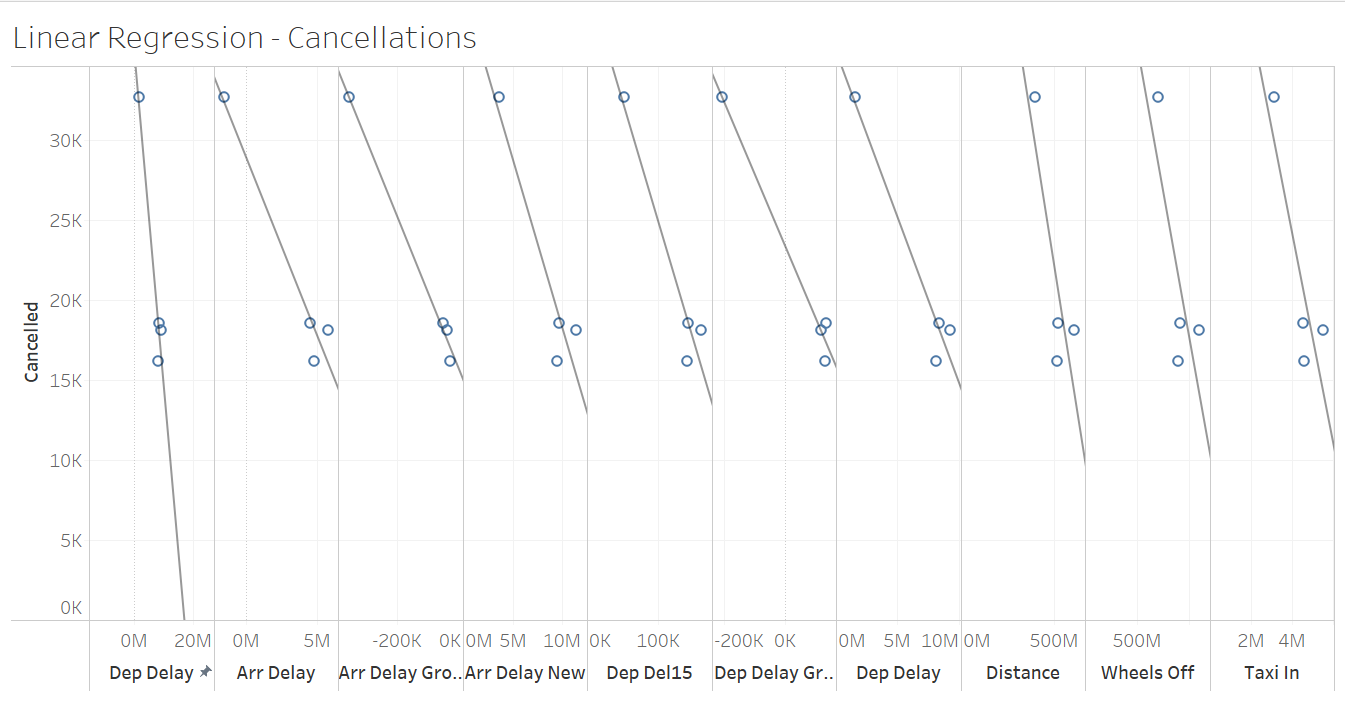


Figure 2: Linear Regression - Cancellations

The results indicate that a flight's ability to arrive on time and not become cancelled rely heavily on the previous flight’s arrival, the plane's ability to depart on time, the length of time the plane spent taxiing, the time the plane had wheels on or off the ground, and the distance of flight. It is unsurprising that most of the factors that impact delays and cancellations represent instances where the flight was unable to adhere to the original schedule. The distance of the flight having an impact is also unsurprising because logic supports the idea that further distance flights may encounter more unexpected events based on the additional time needed to make the flight or additional distance needed to be covered.

To further understand why flight delays and cancellations occur, a drill down approach was utilized to evaluate how seasonality, day of month, day of week, generalized time of day, and hour of day impact a flight's ability or inability to adhere to its schedule.

Chart, line chart

Description automatically generated

Figure 3: Delays - Time Series Plot

Starting with seasonality and referring to Figure 3, we can observe that flight delays were generally higher in 2019 and 2021, with 2021 experiencing the highest average delay time across all the years. The year 2020 is not displayed and this decision will be explained further in the discussion. From this time series plot, it is evident that the highest number of delays occur during the summer months.

Chart

Description automatically generated

Figure 4: Percentage of Flights with Arrival Delays be Seasons

Further supporting the time series plot, Figure 4 shows that flights during the summer season (IS\_SUMMER=1) had a higher percentage of delays at 51.51% compared to non-summer months (IS\_SUMMER=0) which had a percentage of 44.49%. This further suggests that summer months may have more challenges or factors that contribute to flight delays.

Next, the day of the month, city of origin, and individual airlines performances were evaluated for all flights that departed from Texas in 2019-2022. The factors were considered because we wanted to determine which days of the month and cities were most likely to be impacted by flight delays and cancellations.

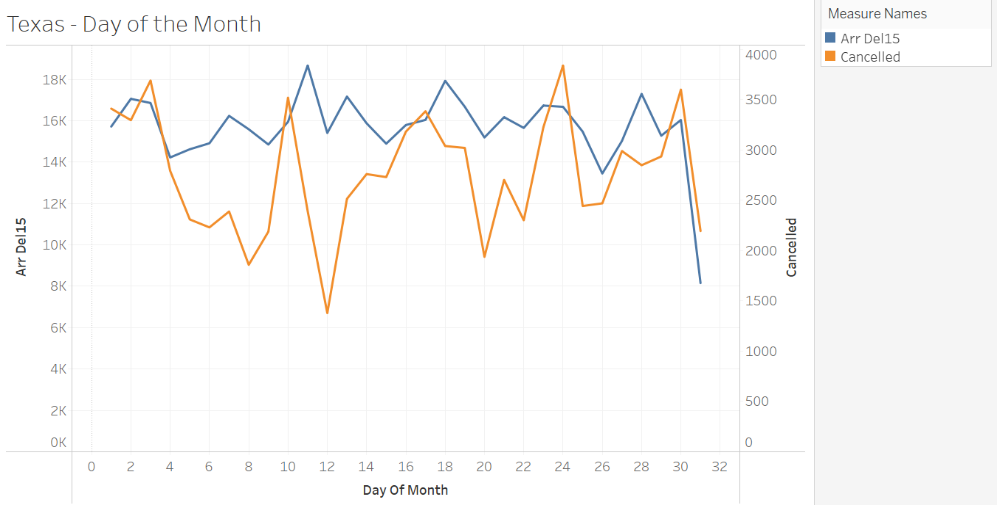


Figure 5: Texas Day of the Month

Referring to Figure 5, overall arrival delays, depicted in blue, did not have as much day-to-day variation as cancellations depicted in orange. On average, the 11th, 18th, and 28th day of each month experienced the most arrival delays whereas the 3rd, 10th, 17th, 24th, and 30th day of each month experienced the most cancellations. Interestingly, most delays and cancellation “highs” were spaced approximately one week apart. Another interesting finding was the closeness of dates between cancellation and delay highs that occurred mid-month. Cancellations on the 10th and 17th day of the month were followed by high rates of delays on the 11th and 18th days of the month.



Figure 6: Percent of Flights Delayed Per Weekday Total

Figure 6 shows an analysis of flight delays by day of the week shows that Fridays and Mondays have the highest number of flights, while Tuesdays and Saturdays have the fewest flights. Fridays have the highest percentage of delays at 19.87%, followed by Sundays at 19.21%, Thursdays at 19.09%, Mondays at 18.99%, Saturdays at 18.07%, Wednesdays at 17.57%, and Tuesdays with the lowest percentage of cancellations at 16.69% of total flights operated on those days.



Figure 7: Total Flights with Arrival Delays by Time of the Day

A generalized approach of evaluating delays based on time of day was evaluated using Python. Based on the above table Figure 7, we can see that the majority of flights experiencing departure delays occur in the evening with a count of 339,148. The afternoon and morning follow with 238,972 and 93,113 counts, respectively. The least number of flights experiencing departure delays occurs in the early morning with only 68,057 counts. These findings were expected as there are less flights in the earlier part of the day with afternoon and evening flights being more prevalent and impacted by the “snowball effect”, meaning that the average delay time builds upon itself.

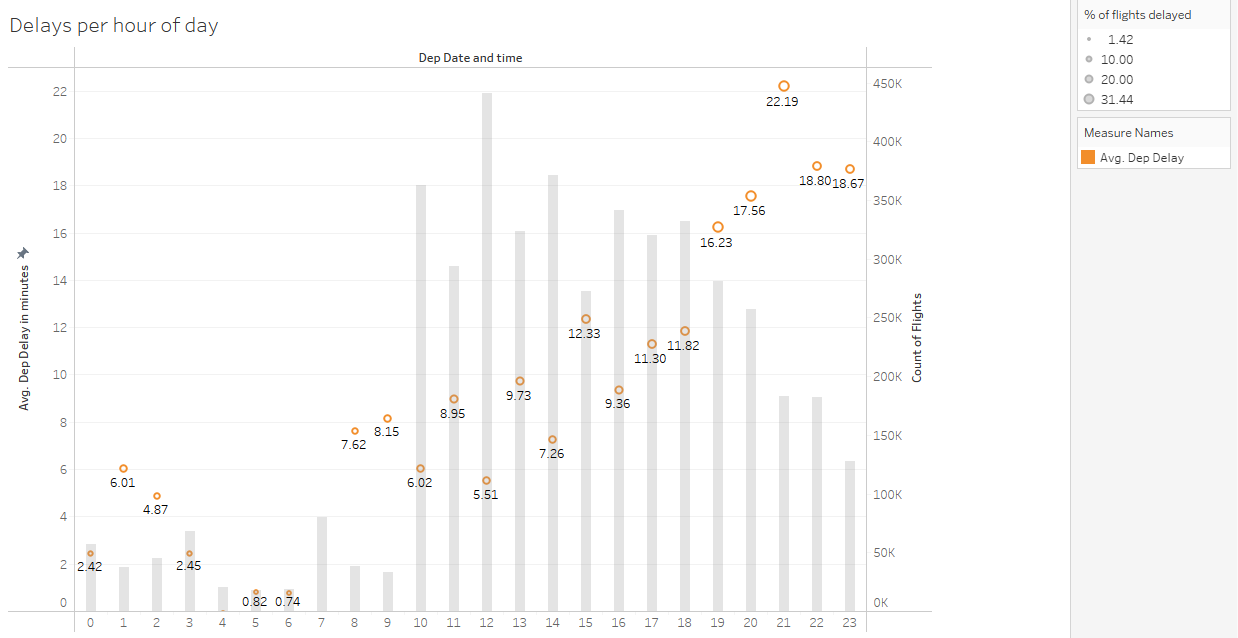


Figure 8: Delays per Hour of Day

Drilling deeper, Figure 8 shows an analysis of delayed flights throughout the day revealed that during the first 9 hours of the day, the average delay time ranged from 0.74 minutes to 8.15 minutes, this is expected considering that the total number of flights during these hours was below 100,000 flights. On the other hand, the rapid increase in the number of flights after 10:00 causes an increase from a minimum average of 6.02 minutes of delayed time to a maximum average of 22.19 minutes. One implication of these findings is that as the day progresses and the number of flights increases so does the average delay time. As mentioned previously, this is possibly due to a “snowball effect.”

After evaluating how time periods impacted delays and cancellations, causes of delays and cancellations were evaluated to provide deeper insights. Cities in Texas as well as airline carrier’s performance were considered to determine how these variables contributed to flight delays and cancellations. For this section of the analysis, delays will be discussed first and followed by cancellations.

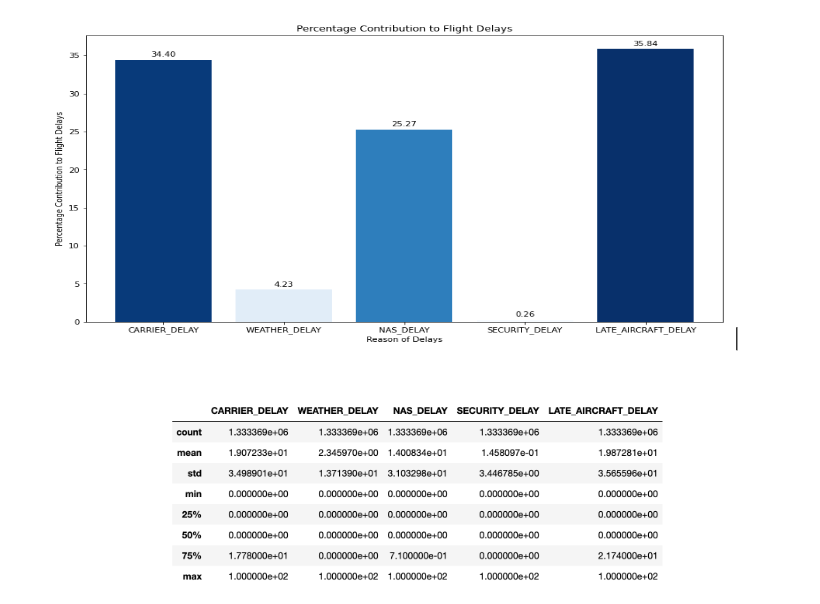


Figure 9: Percentage Contribution to Flight Delays

Figure 9 above clearly indicates that the total percentage contribution of flight delay times vary significantly by category, with the longest delays being caused by "LATE\_AIRCRAFT\_DELAY" (35.84 %) and "CARRIER\_DELAY" (34.40 %), while the shortest delays are due to "WEATHER\_DELAY" (4.3 %) and "SECURITY\_DELAY" (0.26 %).This suggests that carrier-related delays and delays due to late aircraft are more common and have a greater impact on overall flight delays than weather-related, National Air System or security-related delays.

The standard deviations for the percentage contribution of delay are quite large, indicating that there is a wide range of delay contribution for each category. The minimum delay contribution for each category is 0, which makes sense since some flights may not experience any delay. These findings are important because the most impactful causes of delays are variables that are somewhat controllable unlike delays caused by weather. Airlines can use this information to better mitigate delays.

The determination of delay causes was followed by an overview of how specific airlines and cities within Texas performed in terms of percentages of delays.

Chart, bar chart

Description automatically generated

Figure 10: Delay and On-Time Rates by Airline

Figure 10 shows a horizontal stacked bar chart that provides insights into the delay and on-time rates of the airlines in the dataset. Specifically, the graph showcases the delay and on-time rates of the top 5 airlines with the highest delay rates in comparison to the other 19 airlines. The chart allows us to compare the performance of these airlines in terms of their delay and on-time rates.

Upon analysing the graph, it is evident that Hawaiian Airlines has the highest delay rate, with around 50.1% of its flights getting delayed out of the total flights. This means that more than half of the Hawaiian Airlines flights are not arriving on time, which can cause inconvenience to passengers and affect their travel plans.

The table depicted in Figure 11 shows the total percentage of delays per city. This was done to visualize on average how locations in Texas ranked in terms of percentages of flight delays.

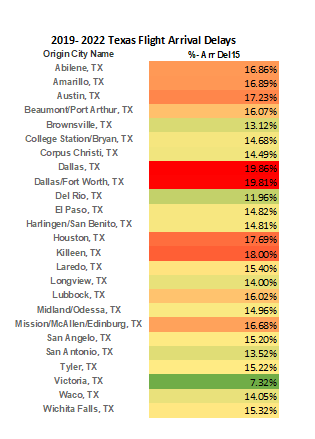


Figure 11: 2019-2022 Texas Flight Arrival Delays

Dallas, Dallas/Forth Worth, and Killeen had the highest percentages of arrival delays in Texas. It was unsurprising that delays were most common in Dallas & Dallas/Ft. Worth due to their central location and with both cities being hubs for various airlines with high numbers of flights. However, each city having on average nearly a 20% arrival delay percentage was surprising as the percentages seem high. This may indicate that highly trafficked locations struggle to mitigate delays. Another interesting finding was that Killeen had the third highest percentage of arrival delays.

The causes of flight cancellations were determined next. From the bar chart below, Figure 12, it is evident that the primary cause of flight cancellations is weather-related issues (CANCELLATION CODE B). The second most common reason for flight cancellations is carrier-caused disruptions (CANCELLATION CODE A), followed by issues within the National Aviation System (CANCELLATION CODE C) and security concerns (CANCELLATION CODE D).

Chart, bar chart

Description automatically generated

Figure 12: Distribution of Reasons of Flight Cancellation

Unlike flight delays, cancellations are more impacted by an uncontrollable variable, weather conditions. This is to be expected because severe weather events are more likely to prohibit a flight from taking place rather than delaying it because these events tend to be longer lasting and more widespread. Of note, carrier-related issues are a contributing factor in both delays and cancellations, and it can be deduced that these insights can be used to improve flight logistics.

Chart, line chart

Description automatically generated

Figure 13: Cancellation Time Series

The time series chart as seen above in Figure 13 reveals that the cancellation rate was notably higher in 2021 and 2022, with 2021 experiencing the highest average cancellation rate among all years. Furthermore, a seasonal pattern is evident, as cancellation rates increase from December to March.

The determination of cancellation causes was followed by an overview of how specific airlines and cities within Texas performed in terms of percentages of cancellations.

Chart, bar chart

Description automatically generated

Figure 14: Top 5 Cancellation Rates by Airlines

Figure 14 reveals the top 5 airlines with the highest cancellation rates among a total of 19 airlines in the dataset. Notably, Mesa Airlines had the highest cancellation rate among these top 5 airlines. Additionally, it is possible that Mesa Airlines has poor customer service, which could contribute to their high cancellation rate.

The table depicted in Figure 15 shows the total percentage of cancellations per city. This was done to visualize on average how locations in Texas ranked in terms of percentages of flight cancellations.

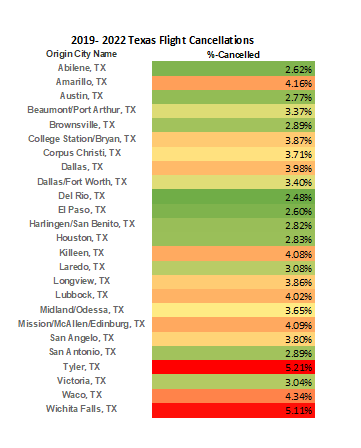


Figure 15: 2019-2022 Texas Flight Cancellations

Tyler, Wichita Falls, and Waco had the highest percentage of flight cancellations. Overall, the rates of cancellations were significantly lower than arrival delays and did not exceed 5.21%. However, it was unexpected that the cities with the highest percentages of arrival delays were not amongst the highest percentages of cancellations. In fact, Dallas, Dallas/Ft. Worth, and Kileen were on the middle to low end of the spectrum of cancellation percentages indicating although they face many issues with delays, they are average at mitigating cancellations. Likewise, Tyler, Wichita Falls, and Waco were on the middle end of the spectrum for arrival delays but ranked highest for cancellation percentages indicating they have a harder time mitigating cancellations than arrival delays.

## **Predicting Flight Delays**

After the evaluation of factors that impacted delays and cancellations was completed, prediction models were created for delays and cancellations. Then the best performing model's ability to replicate results was tested with New York’s data, another state with high air traffic. For this section of the analysis, delays will be discussed first and followed by cancellations.

Three tree-based AI machine learning models were employed to predict flight delays, including the XGBoost Classifier (eXtreme Gradient Boosting), Random Forest Classifier, and CatBoost Classifier. We opted for tree-based classifiers due to their effectiveness in handling large datasets, their ability to manage categorical and numerical values, as well as missing data. Furthermore, these classifiers demonstrate robustness against noise and overfitting. They are also less sensitive to outliers, as their decision-making process relies on data structure rather than actual values.

### **XGBoost Model**

Text

Description automatically generatedChart

Description automatically generated

Figure 16: XGBoost Model

The first model executed was the XGBoost model, utilizing parameters {'max\_depth': 2, 'eta': 1, 'objective': 'binary:logistic'}. Upon completion, the model achieved an accuracy of 93%, an F1 score of 81%, a precision of 89%, and a recall score of 75%. According to the confusion matrix, there were 2,160 instances of Type 1 error (False Positive) and 6,282 instances of Type 2 error (False Negative).

### **Random Forest Model**

Text

Description automatically generatedChart

Description automatically generated

Figure 17: Random Forest Model

The second model executed, the Random Forest Classifier, had with the parameter n\_estimators set to 40. Upon completion, the model yielded an accuracy of 97%, an F1 score of 91%, a precision of 95%, and a recall score of 88%. According to the confusion matrix, there were 1,118 instances of Type 1 error (False Positive) and 2,777 instances of Type 2 error (False Negative).

### **CatBoost Model**

Text

Description automatically generatedChart

Description automatically generated

Figure 18: CatBoost Model

The third model executed was the CatBoost Classifier, using default parameters. Upon completion, the model achieved an accuracy of 99%, an F1 score of 98%, a precision of 99%, and a recall score of 98%. According to the confusion matrix, there were 226 instances of Type 1 error (False Positive) and 444 instances of Type 2 error (False Negative).

**Model Evaluation**

Chart, bar chart, treemap chart

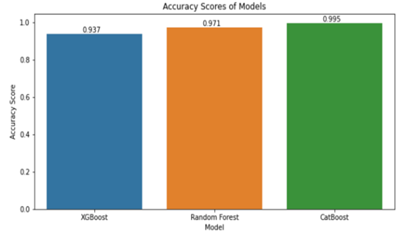
Description automatically generated

Figure 19: Evaluation

After thoroughly assessing the performance of the three models based on F1-Score, accuracy, and confusion matrix, we arrived at the conclusion that the CatBoost model outperformed the others in accurately predicting airline delays. Its superior results in terms of precision and recall, as well as the reduced number of errors in the confusion matrix, demonstrate its effectiveness in this specific task.

### **Evaluating Best Model: (New York)**

Text

Description automatically generatedChart

Description automatically generated with medium confidence

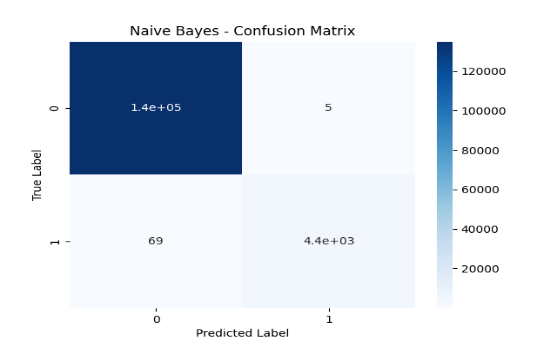
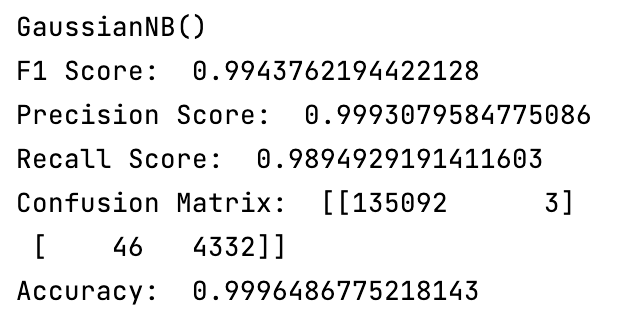
Figure 20: New York Model

After evaluating three models and identifying the CatBoost model as the best one, we used it to predict delays in the Texas dataset, achieving 99.5% accuracy, a 98.6% F-1 score, and 226 instances of Type 1 errors (False Positives) along with 444 instances of Type 2 errors (False Negatives). To further assess the effectiveness of our top-performing model, we applied it to the New York dataset, which had the same number of data points as the Texas dataset. The CatBoost model produced similar results, with an accuracy of 99.5%, an F1 score of 98.6%, 261 instances of Type 1 errors (False Positives), and 482 instances of Type 2 errors (False Negatives). The consistent performance across both datasets in terms of accuracy, F1 score, and confusion matrix indicates that the CatBoost classifier is an effective model for predicting flight delays.

## **Predicting Flight Cancellations**

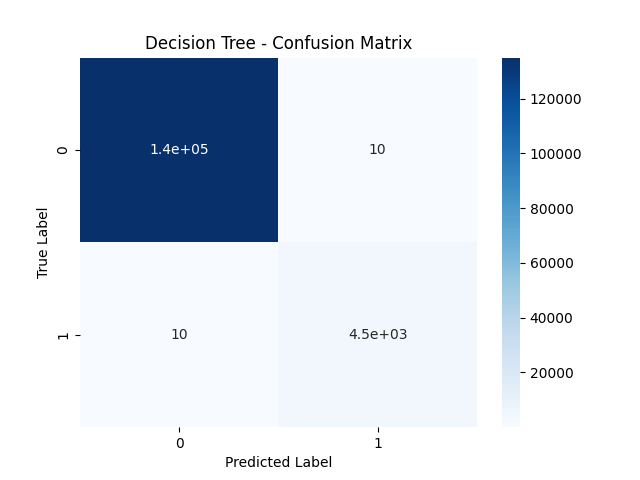
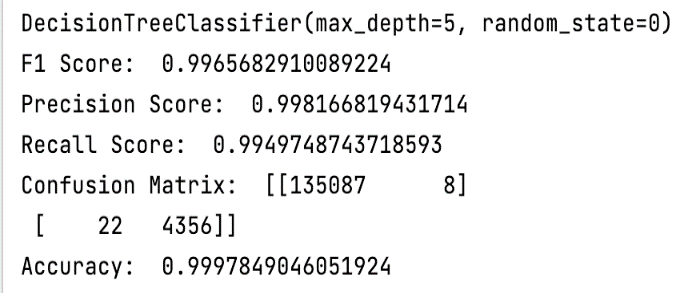
The next step was to predict flight cancellations for Texas by using Machine Learning Models Naïve Bayes Classifier, Decision Tree Classifier, Random Forest Classifier, Support Vector Machine and K-Nearest Neighbors Classifier. As we are predicting if the flight is cancelled or not cancelled with a large dataset, we used classification algorithms to compare their evaluation metrics and select the best model for predicting cancellations.

### **Naïve Bayes**



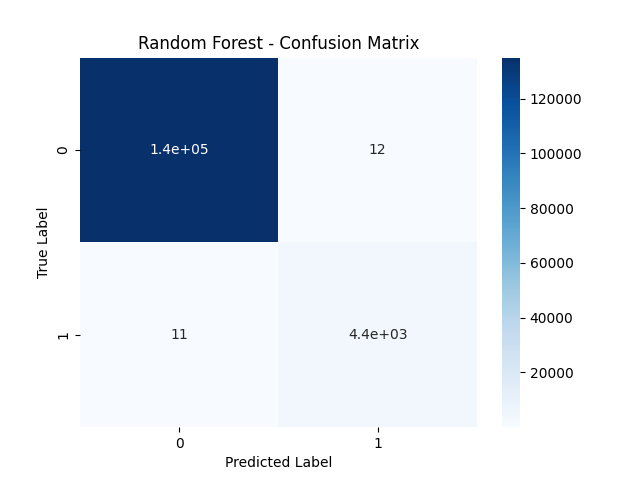
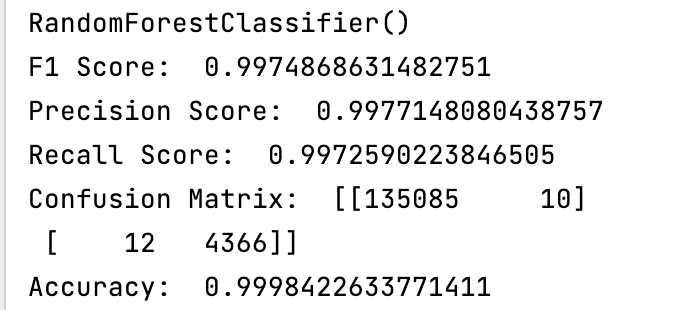
The first model executed was the Naïve Bayes model. Upon completion, the model achieved an accuracy of 99.96%, an F1 score of 99.4%, a precision of 99.9%, and a recall score of 98.9%. According to the confusion matrix, there were 3 instances of Type 1 error (False Positive) and 46 instances of Type 2 error (False Negative).

### **Decision Tree**



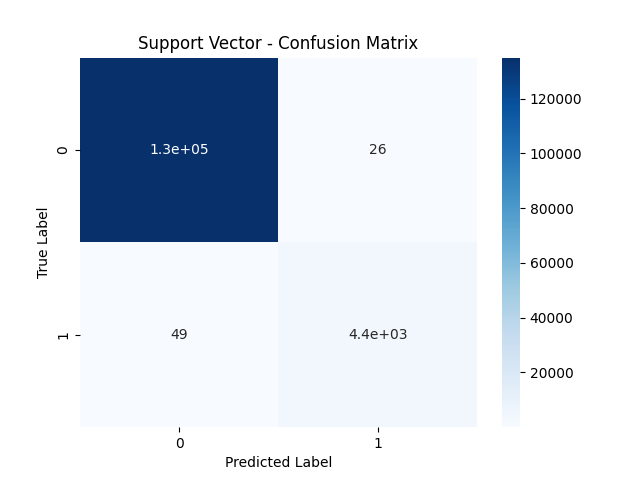
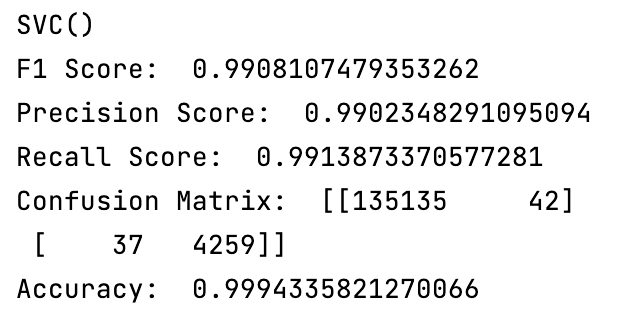
The second model executed was the Decision Tree model, with parameters maximum depth set to 5 and random state to 0. Upon completion, the model achieved an accuracy of 99.97%, an F1 score of 99.6%, a precision of 99.81%, and a recall score of 99.49%. According to the confusion matrix, there were 8 instances of Type 1 error (False Positive) and 22 instances of Type 2 error (False Negative).

### **Random Forest Classifier**



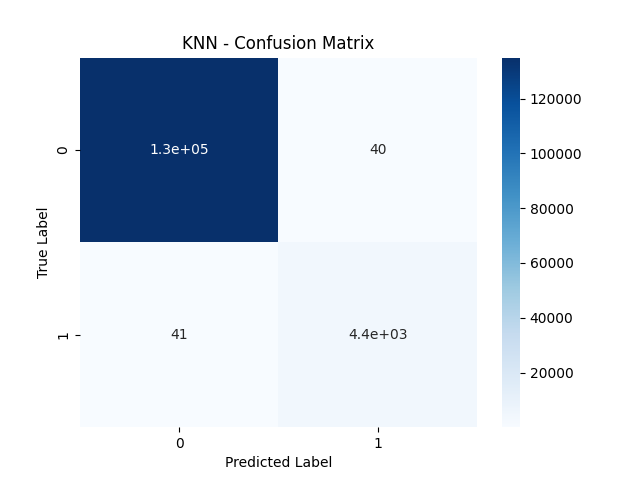
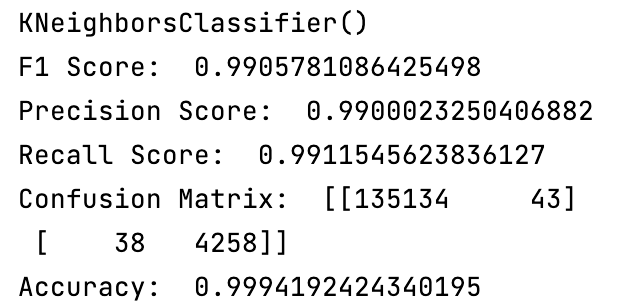
The third model executed was the Random Forest model. Upon completion, the model achieved an accuracy of 99.98%, an F1 score of 99.7%, a precision of 99.77%, and a recall score of 99.72%. According to the confusion matrix, there were 10 instances of Type 1 error (False Positive) and 12 instances of Type 2 error (False Negative).

### **Support Vector Machine**



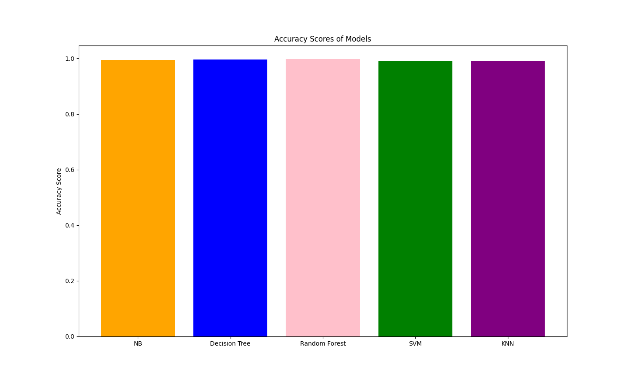
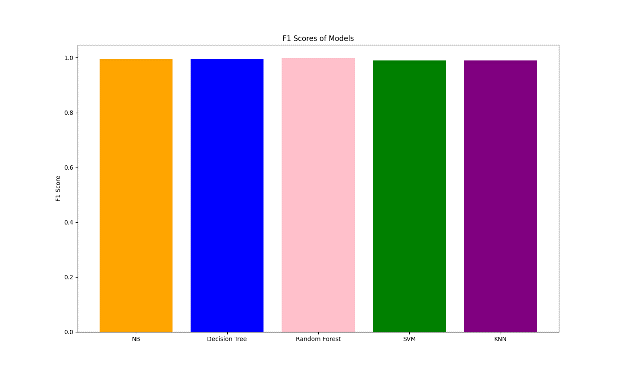
The fourth model executed was the Support Vector Machine model, with parameters kernel set to 'rbf' and gamma set to 'scale'. Upon completion, the model achieved an accuracy of 99.94%, an F1 score of 99.08%, a precision of 99.02%, and a recall score of 99.13%. According to the confusion matrix, there were 42 instances of Type 1 error (False Positive) and 37 instances of Type 2 error (False Negative).

### **K-Nearest Neighbors Classifier**



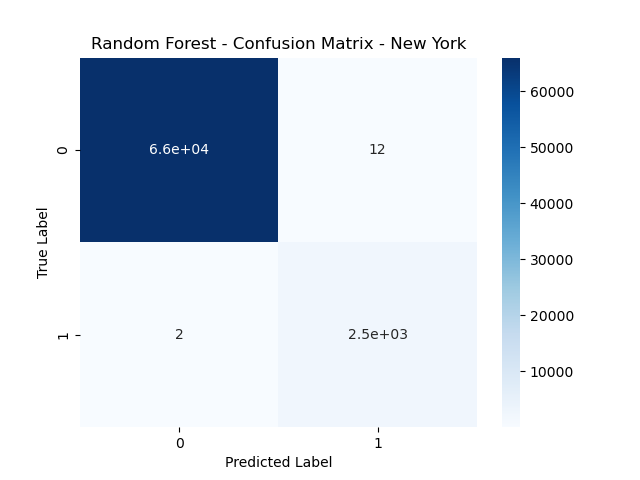
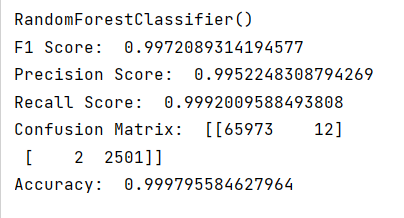
The fifth model executed was the KNN model. Upon completion, the model achieved an accuracy of 99.94%, an F1 score of 99.05%, a precision of 99%, and a recall score of 99.11%. According to the confusion matrix, there were 43 instances of Type 1 error (False Positive) and 38 instances of Type 2 error (False Negative).

### **Model Evaluation**



After thoroughly assessing the performance of the three models based on F1-Score, accuracy, and confusion matrix, we arrived at the conclusion that the Random Forest model outperformed the others in accurately predicting airline cancellations. Its results in terms of precision and recall, as well as the reduced number of errors in the confusion matrix, demonstrate its effectiveness in this specific task.

### **Evaluating Best Model – New York**



After evaluating three models and identifying the Random Forest model as the best one, we used it to predict cancellations in the Texas dataset, achieving 99.98%, accuracy, a 99.7% F-1 score, and 10 instances of Type 1 errors (False Positives) along with 12 instances of Type 2 errors (False Negatives). To further assess the effectiveness of our top-performing model, we applied it to the New York dataset, which had the same number of data points as the Texas dataset. The Random Forest model produced similar results, with an accuracy of 99.97%, an F1 score of 99.7%, 12 instances of Type 1 errors (False Positives), and 2 instances of Type 2 errors (False Negatives). The consistent performance across both datasets in terms of accuracy, F1 score, and confusion matrix indicates that the Random Forest classifier is an effective model for predicting flight cancellations.

# **Discussion**

## **Summary of Results**

The results presented in the analysis demonstrate the complexity airlines in Texas face in regard to flight logistics. Any disruption to a flight’s original schedule can result in delays or cancellations that “snowball” and can cause subsequent flights to face the same disruptions. Our analysis indicates various factors such as location, time, airline carrier, and other events impact flight delays and cancellations.

The practical importance of this research is its ability to highlight controllable and uncontrollable factors that leaders in aviation can leverage to improve air travel. It provides the necessary framework leaders need to create strategies that result in the least disruptive outcomes for flights.

Additionally, the CatBoost and Random Forrest prediction models are beneficial in creating proactive measures. Their high accuracy rates provide decision makers with time to adjust flight schedules, reallocate staff, or reschedule passengers to comparable flights before delays and cancellations occur. This can improve instances of delays and cancellations while saving money and keeping paying customers happy.

## **Answer Research Question**

Upon concluding the analysis of the obtained dataset, consisting of a total of 86 variables, and conducting a comprehensive descriptive analysis, it can be inferred that certain variables exhibit the strongest correlation with our dependent variables. Specifically, for the variable DEP DEL15, which indicates flight delays, the variables that demonstrate the highest correlation are DISTANCE, DEP DELAY, WHEELS ON, WHEELS OFF, and TAXI IN. Similarly, for the variable CANCELLED, which signifies flights that do not take place at all, the variables with the strongest correlation are DEP Delay, ARR DELAY, ARR DELAY GROUP, ARR DELAY NEW, DEP DELAY 15, DEP DELAY GROUP, DISTANCE, WHEELS OFF, and TAXI IN.

In summary, flight delays are impacted by factors, including flight distance and the time taken for flights to transport from the terminal to takeoff. These findings are supported by the analysis, which reveals that they are exacerbated by other variables such as flights scheduled during the summer or in the evenings, as well as issues related to the carrier where we have been able to pinpoint Hawaiian Airlines, Allegiant Air, JetBlue Airways, Frontier Airlines, and American Airlines as those with the most delayed flights. On the other hand, cancellations are influenced by delayed departures of previous flights, the duration of those delays, flight distance, the time of wheels off, and the type of taxi used. These factors are also significantly affected by external factors such as weather conditions and seasonality, with the months of December to March having the greatest impact on cancellations. We have been able to identify Mesa Airlines, Horizon Air, PSA Airlines, Allegiant Air, and American Airlines as airline carriers with high cancellation rates.

## **Exclusion of the Year 2020**

We excluded data from the year 2020 for a majority of the models due to the COVID-19 pandemic, which resulted in a significant number of flight cancellations and could potentially hinder our ability to analyze patterns in delays and cancellations over the years. By excluding data from this year, we aimed to eliminate the impact of this unique and unprecedented event on our analysis.

## **Limitations**

Due to hardware limitations, we used a sample of 10% of the total dataset to effectively run our delays predictive model. In the future, other variables could be added for further analysis. For example, weather details can be included to analyze historical weather data and determine which weather conditions have the greatest impact on flight operations. This information can help airlines and airports prepare better for potential weather-related disruptions. Another variable that could be added is aircraft type details, which can help identify which types of aircraft are more likely to experience delays or cancellations due to mechanical issues or maintenance requirements. By analyzing the relationship between aircraft type and delay/cancellation rates, airlines and airports can make more informed decisions about maintenance and repairs to minimize disruptions to flight operations.

For the cancellation predictive model, we used a subset of the data because the complete dataset would have been too time consuming to execute the algorithms. Another reason we decided to use a subset was due to the finding of some algorithms not being able to handle the large dataset. An example of this is when we tried to run Logistic Regression with solver as saga (as it’s fast and efficient on larger dataset) on our sample dataset, and we received a warning stating that the max iterations was reached. Additionally, we were unable to run algorithms like SVM and KNN with the complete dataset because of limitations with our existing hardware configurations. Future studies can solve this problem by utilizing higher configurated super computers.

The next limitation we faced with the cancellation models was dealing with the variables like Arrival time and Departure Time of the flight in reference to cancellations, because the dataset had null values for them. To better suit the algorithm, we had to impute these variables with value as 0, for the algorithm to execute and provide us the evaluation metrics.

The last limitation was that our project was dealing with cancellations from the year 2019 – 2022. With COVID-19 and lockdown restrictions impacting the year 2020, especially on the transportation industry, our data has class imbalance between cancelled and non-cancelled. Future research could focus more on how much the year 2020 impacted the results by running all models with and without the year 2020.

## **Future Directions**

Further research should analyze other factors that contribute to flight cancellations. Researchers could explore the effectiveness of different predictive models and identify the variables that are most predictive of flight cancellations. The proposed model could be extended further to predict flight cancellations based on a range of variables such as aircraft maintenance records, historical data, by using Machine learning and data analytics. Further research could explore strategies for minimizing the impact of flight cancellations on passengers and airlines; for example, developing more robust contingency plans, exploring alternative modes of transportation, improving communication with passengers during flight disruptions, etc.

Moreover, researchers can also use multivariate analysis techniques to identify underlying patterns or relationships between variables that may not be immediately apparent. For example, some common multivariate analysis techniques are multiple regression analysis, factor analysis, and structural equation modeling. Researchers can gain a more comprehensive understanding of the complex relationships that exist between different factors and can use this information to develop more effective interventions and strategies (Team, 2022).

## **Implications & Importance**

An implication of this study is that location, airline carrier, and time can impact flight delays and cancellations. These insights can be used by airports, airline carriers, or passengers to make better decisions. Airports and airlines can use the data provided to benchmark their ability of mitigating delays and cancellations against their competitors while adopting new strategies that improve the timeliness of air travel. An example of this would be spacing flight times out during peak delay time periods or to stop offering flights at a certain time if it often results in cancellations. Passengers benefit from this research too as they gain the ability to make more informed decisions on which cities, airlines, and time periods are most likely to face delays and cancellations.

Although many causes of delays and cancellations are intuitive, others were surprising. It was found that high cancellation numbers during the middle of the month were followed by a high number of flight delays. These delays were likely due to spillover from cancelled passengers needing to be rescheduled to the following day, however it is surprising that this is most apparent during the middle of the month. Location’s percentage of overall flight delays was also surprising because the top 3 highest locations had on average an almost 20% chance of delays. High probability of delays can indicate a location’s inability to handle its flight load or its difficulty in mitigating challenges that cause delays. Passengers at these locations may feel less satisfied with their experience and seek alternate transportation in the future. Overall, the ability to predict factors that impact delays and cancellations improves decision making and the customer experience.

# **Conclusion**

To assess which factors impacted flight delays and cancellations, linear regression was used to determine correlated factors. A drill down approach was used to evaluate the impacts of seasonality, specific day of week, generalized time of day, and hour on a flight's ability to adhere to its original schedule. Next, locations in Texas and airline carriers most associated with delays and cancellations were identified. Our analysis indicated that various factors such as location, time, airline carrier, and other events all played a part in flight delays and cancellations. Finally, a CatGBoost Classfier model and Random Forest model was built in order to predict delays and cancellations.

Our research adds to the already present extensive literature on the topic as it focused on the state of Texas, a highly air trafficked state with many large airports that serve as hubs for major airline carriers. Additionally, the findings in the analysis addressed seasonality and a drill approach to understand how the specific day of the month and time of flight impacted flight delays and cancellations. Our CatBoost Classifier predictive model was also somewhat unique and not mentioned in literature we reviewed.

The airline industry may benefit from this data by creating initiatives that help flights adhere to their originally planned schedule. Factors that reduce taxiing time or ways to mitigate the impacts of previous flight delays should be evaluated on an individual location and airline carrier basis to make the most beneficial improvements.

The next important research step to further answer what factors impact delays and cancellations should focus on different predictive models and extending the data to include plane maintenance records and historical weather data. It is likely that the complexity of air travel logistics will continue to be studied and improved upon.

# **References**

Atlioglu, M., & Bolat, M., Sahin, M., Tunali, V., & Kilinc, D. . (2020). Supervised Learning Approaches to Flight Delay Prediction. *Sakarya University Journal of Science*.

Bombelli, A. (2023). Analysis of the effect of extreme weather on the US domestic air network. A delay and cancellation propagation network approach. *Transport Geography*.

Cai, Kaiquan & Li, Yue & Fang, Yiping & Yanbo, Zhu. (2021). A Deep Learning Approach for Flight Delay Prediction Through Time-Evolving Graphs. *IEEE Transactions on Intelligent Transportation Systems*.

Dewi, O. (2018). Revenue management model based on capacity sharing and overbooking in the airline. *Journal of Engineering and Management in Industrial System*, 86-94.

Gruenwald, H. (2020). Covid-19 and Airlines.

Gu, Y., Yang, J., Wang, C., & Xie, G. . (2020). Early warning model for passenger disturbance due to flight delays. *PloS one*.

Matthews, A. W. (2023). *Why so many flight cancellations? here's what to know*. Retrieved from ViaTravelers: <https://viatravelers.com/many-flight-cancellations/>

Pagoni, I & Koumoutsidi, A. (2022). Efficiency of Regional Airports: Insights on the Effect of Airline Type and Seasonal Variations in Traffic. *Future Transportation*, 774-792.

Seyedmirsajad, M. (2022). Flight delay causality: Machine learning technique in conjunction with random parameter statistical analysis. *Transportation Science and Technology*.

Team, G. L. (2022, December 2). *Overview of multivariate analysis: What is multivariate analysis and model building process?* Great Learning Blog: Free Resources what Matters to shape your Career! Retrieved April 18, 2023, from <https://www.mygreatlearning.com/blog/introduction-to-multivariate-analysis/>

Venkatesh, V., Arya, A., Agarwal, P., Lakshmi, S., & Balana, S. (2017). Iterative machine and deep learning approach for aviation delay prediction. *4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics*, (pp. 562-567).

Wang, Z., Liao, C., Hang, X., Li, L., Delahaye, D., & Hansen, M. (2022). *Distribution Prediction of Strategic Flight Delays via Machine Learning Methods*. Retrieved from Sustainability: <https://doi.org/10.3390/su142215180>

Xuan, X., Khan, K., Su, C.-W., & Khurshid , A. (2021). *Will COVID-19 Threaten the Survival of the Airline Industry?* MDPI AG.

Yanying, Y., Mo, H., & Haifeng, L. (2019). A classification prediction analysis of flight cancellation based on Spark. *Procedia Computer Science*.

Yi, J., Zhang, H., Liu, H., Zhong, G., & Li, G. (2021). Flight Delay Classification Prediction Based on Stacking Algorithm. *Journal of Advanced Transportation*, 1-10.

Yimga, J. (2021). The Airline On-time Performance Impacts of the COVID-19 Pandemic. *Transportation Research Interdisciplinary Perspectives*.

Zámková, M., Rojík, S., Prokop, M., & Stolín, R. (2022). 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. *Sustainability*, 14763.

# **Author’s Contributions**

**Abhilash Bhuyan:** Conducted exploratory data analysis, including feature selection, extraction, and transformation, as well as data sampling. Used Python-based visualizations to identify factors affecting delays and cancellations. Developed three AI machine learning models to predict flight delays, enhancing the project's overall success.

**Alicia Ringo:** Project Manager, Comic Relief, Provider of Gum. Assisted in writing the Abstract, Introduction, Literature Review, Analysis, Discussion, and Conclusion. Created visualizations in Excel and Tableau. Organized collaborative workspace & scheduled meetings.

**Julio Martínez:** Downloading database information and concatenating it for proper application in models using Python. Implementation of data analysis through Tableau for its description and interpretation, creating visualizations.

**Sritha Darbha:** Gathered the initial dataset files from the original source for further analysis. Conducted exploratory data analysis, including feature selection, extraction, transformation and data sampling using Python. Developed five AI Machine Learning models to predict flight cancellations using Python. Assisted in writing Research Limitations.

**Van Nguyen:** Provision of profound research on the previous literature related to the project, involved in planning, and supervising the work. Helped shape the research, designed the figures, and contributed to the final version of the manuscript.

# **Appendix**

**Delay Prediction Texas Code:**

****

**Delay Prediction New York Code:**

****

**Cancellation Prediction Texas Code:**

****

**Cancellation Prediction New York Code:**

****