Clearing the Skies: An In-Depth Look at Factors Behind Airline Delays

Avinash Appineni – 11650646

Dr. Sarah Quintanar

University of North Texas

December 3, 2024

Clearing the Skies: An In-Depth Look at Factors Behind Airline Delays

I INTRODUCTION

Background

The issue of airline delays is complex because it affects not only travelers and airlines but also airport administration. For the airlines, operational delays are substantial because of the amount of fuel used while taxiing in a queue for take-off, managing the delay through airline flights, and looking for delayed passengers. Additionally, such aspects also impact the image of the airline and customer devotion - which are fundamental in a market that is growing ever more saturated. The same goes for the airports, as other delays can affect the movement of traffic from runways or gates etc which would hinder other ground activities and personnel deployment.

With the increasing globalization and growth of the airline industry, the management of delays has now become crucial to enhancing operational productivity and minimizing the economic loss incurred due to flight cancellations. Having a clearer focus on the chances of delay can help the airline business utilize its resources effectively, assist in the alleviation of congestion, and further assist clients. This research forms a contribution to the literature on the need for forecasting in the sector by analyzing a broader dataset that subclasses delays into specific types including but not limited to weather, security, and mechanical. We employ data mining techniques to analyze these factors, aiming to see some tendencies or relations that might indicate the trends.

This research adopts machine learning methodology in building predictive models for the probability and periods of delay based on historical data. For example, if certain weather conditions are causing frequent delays at an airport, airlines might adjust their schedules to account for this pattern. Also, more ground staff could help during these times by managing the crowded areas, by keeping the passengers updated, and helping with rescheduling flights quickly. Likewise,

changes in trends of delay in security-related aspects may help in the formulation of a strategy that targets these processes through training to enhance efficiency.

II SCHOLARLY REVIEW

Delays to airlines have impeded operational efficiency and economic viability of the aviation industry. As Deshpande and Arikan (2012) estimated, the U.S. delays alone were more than 320 million minutes in 2019 which is worth billions. The economic loss arises from various costs including consuming more fuel because of extended or unnecessary taxiing and diversion, employing more crew members, and compensation to the delayed passengers. Flight delays could hurt airlines' profits because, they raise costs and put pressure on their resources. For passengers, delays waste their time and make travel less convenient, which affects how satisfied they are with their trip. Airlines need to reduce delays, so they develop flight schedules months in advance, dealing with routing options, aircraft fleets, and crew members in the process. Yet airlines cannot escape certain influences that are severe, such as random weather patterns, congestion of air traffic, and some unforeseen maintenance that results in delays. With a proper analysis of the causes and consequences of financial metrics, the sector will be able to formulate and implement measures that will reduce delays, help the passengers and the airlines, and cut the overall economic costs.

Since airlines also must deal with delays everyone experiences, these delays are usually divided into a few categories depending on their causes. Flight delays often happen because of bad weather, busy airports, and problems that airlines and airports deal with every day. Adverse weather is among the most notable sources, which include but are not limited to, dense fog, thunderstorms, and severe rain causing restrictions to airflow. It is as Khan et al. (2021) suggest, these types of delays are quite aggressive and have a snowball effect on an individual airline, the whole fleet or even alliances as a large. Lee and Zhong's research (2016) determined that about

75 percent of delays can be solely attributed to weather conditions which typical rough seasons such as winter storms or monsoon seasons aggravate the situation even more. The other major reason is congestion which can be pronounced in hub airports which have a higher density of air traffic. Truong (2021) elaborates on the notion of air traffic congestion further by revealing how the aspect of air travel demand causes delays in airport resources and operational capabilities during peak seasons and holidays. Since there exists a global demand for air traffic that grows at a faster rate than airports can increase their capacity (4.2% to 3.4% per year as per ICAO, 2022), there is a high probability that congestion delays will worsen further and will require efficient expansions of the airport and orderly aircraft space allocation management.

Similarly, to any external factors, internal variables such as operational costs, the structure of the organization's infrastructural setup, or the existence of interrelated delay networks add a lot to the already complicated case of airlines' delays. Nibareke, and Laassiri (2020) who researched problems related to operation emphasized that a calendar for maintenance, availability of crew, as well as other infrastructural limitations tend to block flight departures on the expected time. For instance, maintenance checks are critical but if they are not planned properly, they will result in delays. Also, employees, flight attendants, or pilots can be delayed on an earlier flight and cause a shift in scheduling problems. Furthermore, operational limitations at many airports such as congestion at the gates, obsolete air traffic control systems, and other shortcomings only aggravate this challenge since they increase congestion and the planet's use of the airports is not very effective. This causes chain delays in which one delay will lead to delays in the subsequent schedule forcing a re-schedule for the rest of the day. The existence of such interdependencies also indicates the importance of cooperation between airlines and airports;

resource management must be done efficiently to avoid unexpected interruptions in the provision of services and effectively balance operations.

There is a noticeable change when it comes to how the aviation sector handles delays thanks to the use of new data analytical tools. In the past, these types of analyses were mainly done to identify factors like weather or events that are linked to delays. Recent research highlights the importance of causal relationships, it has allowed the airlines to enhance their capabilities.

Yazdi et al. (2020) was the first in this field, and they suggested guidelines to separate and define different flight delay determinants. As a result, airlines can make a distinction between unavoidable delays (e.g., caused by extreme weather) and delays that could be avoided through better organization (e.g., maintenance planning, as well as resource availability). These insights can assist airlines in predicting when a certain aircraft will require maintenance because it offers a straightforward example of predictive maintenance. Airlines can maintain desirable routine intervals for scheduled maintenance which eliminates sudden failures or last-minute maintenance and repairs which in most cases disrupt flight schedules and subsequently affect other flights within the operational day.

Like traditional event analysis, machine learning algorithms have been created to predict the flight delays at airports, by using both the past data and current information. According to Jiang et al (2020), a hierarchical machine learning system that integrates Bayesian networks and RNN can include the cross-layer correlations, which are the relationships between multiple variables such as weather, traffic, and operational variables, into the highly reliable prediction of

delays. Also, thanks to the use of RNNs in the model it is possible to read sequences of events and their changes, which are especially important in the analysis of delay accumulation and their transmission processes within the system.

Progressing the predictive power, Ye et al. (2020) employed some machine learning algorithms such as Random Forest Regression, Long Short-Term Memory Networks (LSTM), and XGBoost. In their study, it was found that XGBoost is efficient in the management of big datasets and in the prediction of time delay due to its gradient-boosting techniques, which are effective in pattern recognition. Quite the same, LSTM networks have been helpful as well when it comes to the prediction of delays since they allow for temporal analysis of delay events. LSTMs, a kind of RNN, indeed possess the ability to learn long-time dependencies and this is a critical factor when predicting delays when events such as peak seasons of travel or scheduled maintenance occur.

Kim et al. (2016) have pushed this stream of studies forward with an application of LSTM-based deep learning approaches to sequential delay data. Instead of evaluating each flight event separately, as in classical models, LSTM RNNs evaluate all delays in terms of their order and interrelation. This ability to model and process sequential information has led to a more accurate day-to-day oscillation of delay probabilities, which are useful in adjusting the time of arrival and resource deployment in real time. This allows for better planning on the part of the airlines, which can anticipate delays and, for example, alter takeoff and landing schedules, switch routes, or deploy reserve crews, thereby minimizing airfare disruption.

These predictive modeling techniques represent a paradigm shift in the way delays are

effectively managed as they allow for preemptive measures as opposed to simply responding to delays after they have occurred. The combination of machine learning with real-time operational data puts the airline in a better position to foresee and avert potential delays, allocate resources strategically, and manage sequence interruptions. For example, some airlines have begun to use predictive analytics tools to inform the management of their resources by integrating machine learning technology into their decision-making processes. These systems then provide relevant outputs such as determining when to reroute flights or when to amend crew schedules to mitigate predicted delays which enable airlines to enhance timeliness and customer satisfaction.

III DATA

The dataset was carefully assembled using the United States Department of Transportation's Bureau of Transportation Statistics (BTS), documenting activities concerning the volume of flights, delays, and /or cancellations as well as the number of diversions at the intent airports across the US.

The dataset fetches data from several major airports and therefore includes thousands of flight observations. Each record depicts a particular operational flight's occurrence including times of arrival delays, the number of cancellations, and even the number of diversions for each flight.

Factors that contribute to the delays of flights include but are not limited to bad weather, security concerns, problems arising from a carrier, and the National Airspace System (NAS). Because all this information is contained in the dataset, it will also make it possible to explain the relationships that exist among these variables and how they are related to the performance of the flight in general. The understanding of these factors is important in explaining the reasons for the delay and the cancellation of flights hence making it possible to address these issues and enhance the efficiency of the airline industry.

In total, the dataset contained 47,274 observations and 21 variables.

Research questions

1. Predicting Flight Cancellations

Question: Can we predict whether a flight will be canceled based on factors like carrier, weather conditions, NAS issues, and airport location?

Model to Use: Logistic Regression or Random Forest Classifier

Why: Logistic Regression is a good baseline model for binary classification problems (e.g., canceled vs. not canceled). However, if you want to capture complex relationships and interactions between features, a Random Forest Classifier would be more effective. It handles categorical variables well, is robust to missing data, and provides feature importance metrics.

2. Predicting Flight Delays

Question: Can we predict the likelihood of a flight being delayed by more than 15 minutes using variables like the airline, month, weather conditions, and NAS issues?

Model to Use: Gradient Boosting Classifier (e.g., XGBoost) or Support Vector Machine (SVM)

Why: Gradient Boosting models like XGBoost can handle non-linear relationships and are very effective for structured/tabular data. They are also great for feature importance and interpretability. SVM with a non-linear kernel (e.g., RBF) can also be used for classification when the data is not linearly separable.

3. Classifying Root Causes of Delays

Question: Can we classify the root cause of flight delays (carrier, weather, NAS, security, or late aircraft) based on flight data?

Model to Use: Decision Tree Classifier or Random Forest Classifier

Why: Decision Trees are interpretable and work well for multi-class classification tasks. Random Forests provide better performance due to assembling and can also help identify the most important factors contributing to each type of delay.

4. Predicting Delay Duration

Question: Can we predict the duration of a flight delay (in minutes) based on factors such as the airline, weather, month, and time of day?

Model to Use: Linear Regression or Gradient Boosting Regressor (XGBoost)

Why: Linear Regression serves as a good baseline for predicting continuous variables. However, Gradient Boosting Regressors are better suited for capturing non-linear relationships and interactions between features, making them ideal for more accurate predictions of delay durations.

5. Identifying Flights at Risk of Cascading Delays

Question: Can we predict whether a delayed flight will cause subsequent delays in other flights (cascading delays) using historical delay patterns and airline schedules?

Model to Use: Recurrent Neural Networks (RNN) or Long Short-Term Memory (LSTM) Networks

Why: Cascading delays involve temporal dependencies (i.e., previous delays affect future ones), so models like RNNs and LSTMs are well-suited for capturing patterns over time. These models can learn sequential dependencies in data, which is crucial for predicting knock-on effects.

6. Clustering Airports Based on Delay Patterns

Question: Can we cluster airports based on their delay patterns to identify those with the highest likelihood of delays due to specific factors (weather, congestion, etc.)?

Model to Use: K-Means Clustering or DBSCAN

Why: K-Means is a straightforward clustering algorithm for partitioning airports into groups based on similar delay profiles. If the data has noise or varying densities, DBSCAN (Density-Based Spatial Clustering) is more robust.

7. Predicting Peak Seasons for Flight Cancellations

Question: Can we predict the peak months for flight cancellations using historical data on cancellations, weather patterns, and flight volumes?

Model to Use: Random Forest Classifier

Why: ARIMA models are great for univariate time series forecasting. However, if you have

multiple factors influencing cancellations, the Prophet model (developed by Facebook) is easier to use, handles missing data well, and accounts for seasonality and trend changes.

8. Determining Factors Contributing to Late Aircraft Delays

Question: Can we predict whether a flight will experience a "late aircraft" delay based on the scheduled arrival time, departure airport, and previous flight delays?

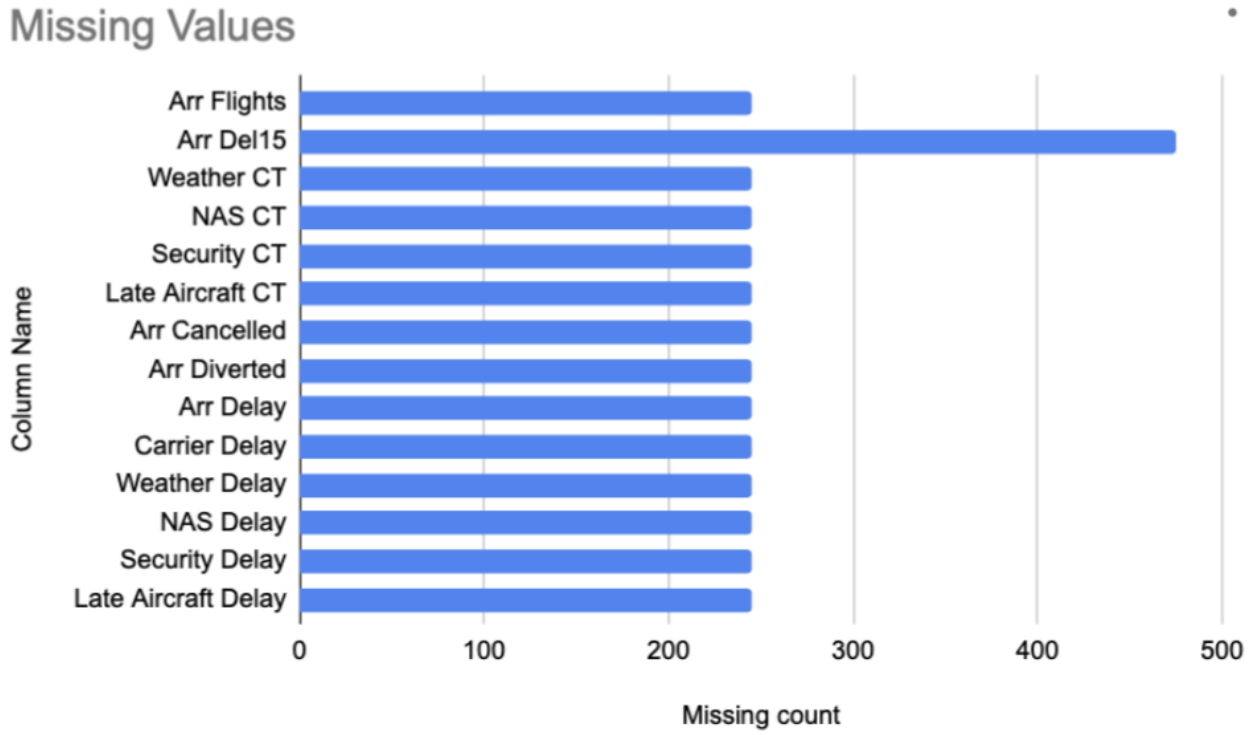
Model to Use: Random Forest Classifier

Why: For a quick and interpretable solution, Logistic Regression works well. If you want a probabilistic approach that works with categorical variables and assumes feature independence, Naive Bayes is a good fit.

Data Preprocessing

Before analyzing the dataset, it is important to first clean and prepare the data. So, this helps improve the quality of the data and ensures better insights. The main steps taken during this phase are listed below:

Addressing the Missing Values



The first step is always the first observation of any data set if there are incomplete values in the set. Ignoring missing data is a common occurrence. This has been proven to misrepresent the outcomes which may result in wrong judgment. As such, they must be tackled properly.

Determination of the Important Variables

Variable	**Relevance**
year, month	To account for seasonal patterns, which affect the frequency and nature of flight delays.

carrier, carrier_name	Enables comparison between airlines, providing a benchmark for performance analysis.
arr_delay, arr_cancelled, arr_diverted	Helps compute metrics related to delays and cancellations, essential for understanding operational disruptions.
weather_ct, security_ct, nas_ct, late_aircraft_ct	Identifies specific causes of delays (e.g., weather, security) for a more detailed analysis of contributing factors.

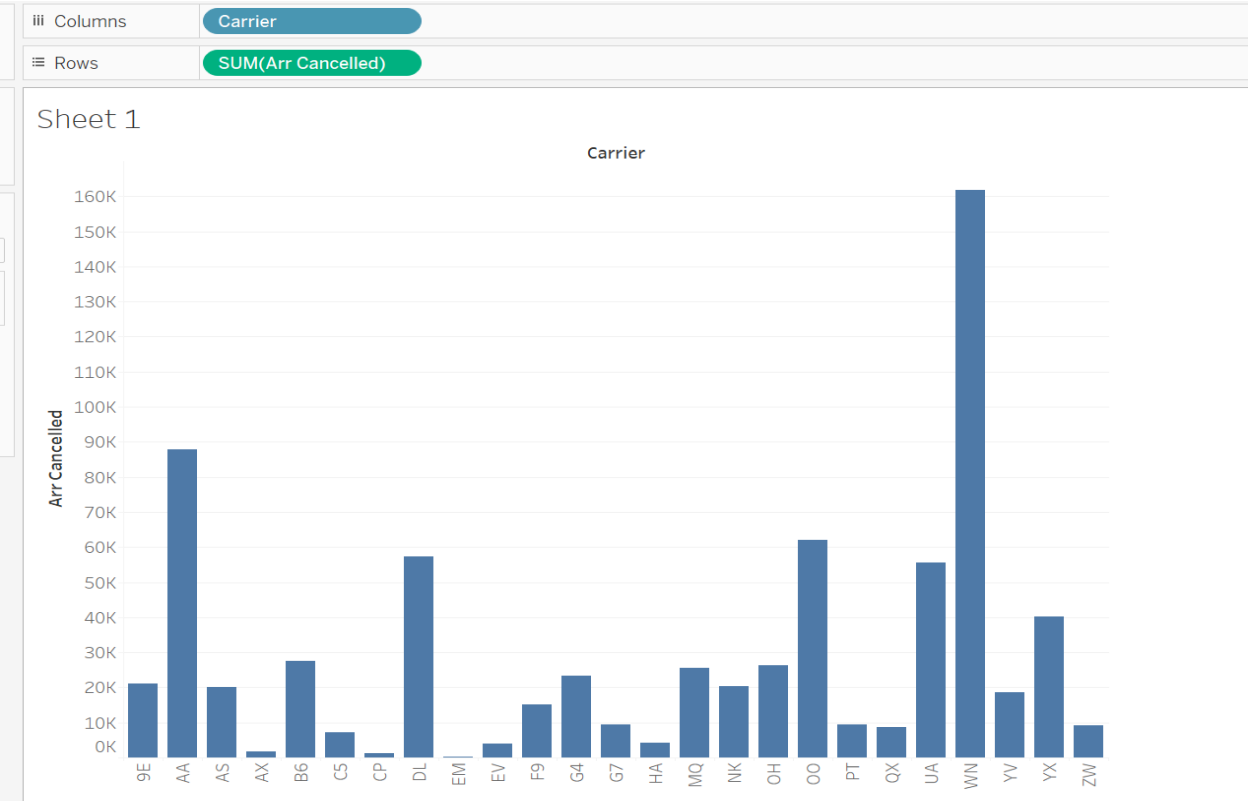
In all these operations it is apparent that the data set has different measures such as the variable of arr_delay which is measured using minutes while its count is the number of delays. Therefore, it would be appropriate to standardize the data so that any model applied has all variables contributing equally. In this context, there are two standard scaling approaches.

Standard Scaling Considerations: Standard scaling is often used to make continuous variables like *arr_delay* (arrival delay in minutes), which is easier to compare by setting their average to 0 and their spread to 1. However, applying this scaling here could reduce the impact of extreme delay values, which may be important for the analysis. So, this could potentially bias the results, as it may hide some real-world differences in delay times. As a result, careful consideration is

required if scaling is needed, or if keeping the original values would be more appropriate.

Exploratory Data Analysis:

1. How does the frequency of flight cancellations differ amongst airline carriers?

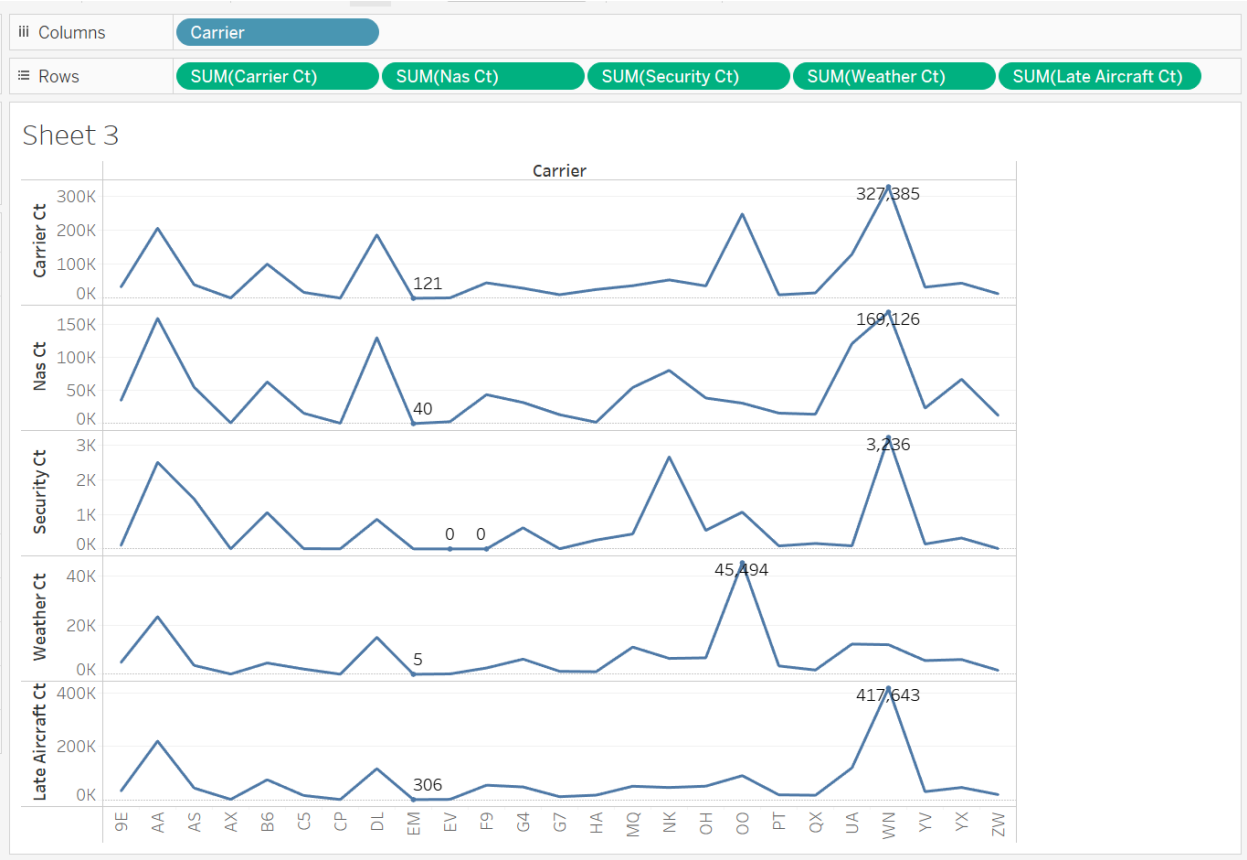


In the above bar chart, we can see it illustrates the total number of flight cancellations for different airline carriers. Additionally, “Carrier” on the x-axis and “SUM (Arr Cancelled)” on the y-axis displaying the significant difference in cancellation frequencies among all carriers.

2. Is there a seasonal trend in the number of airline cancellations and delays?

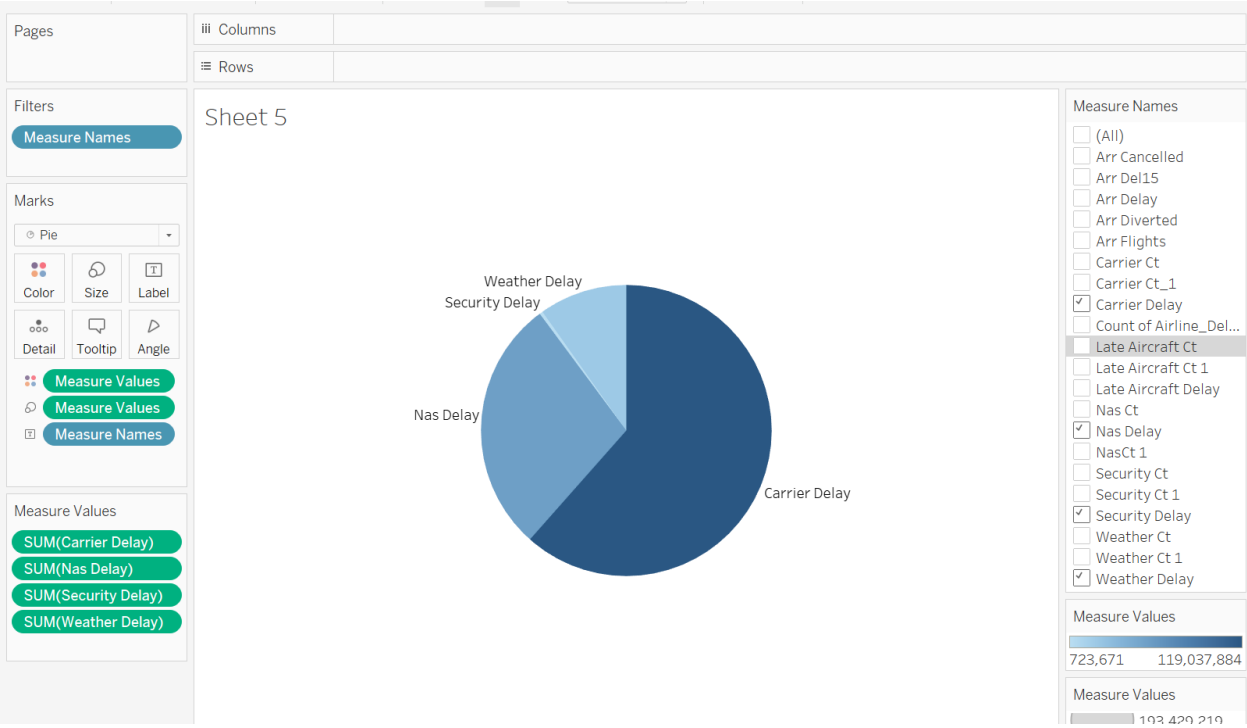
Airline cancellations (Arr Cancelled) show distinct seasonal tendencies, peaking in January, June, and December, according to the line graph. On the other hand, “delays” (Arr Del15) and “total flights” (Arr Flights) exhibit less noticeable seasonal fluctuations and are more consistent throughout the year.

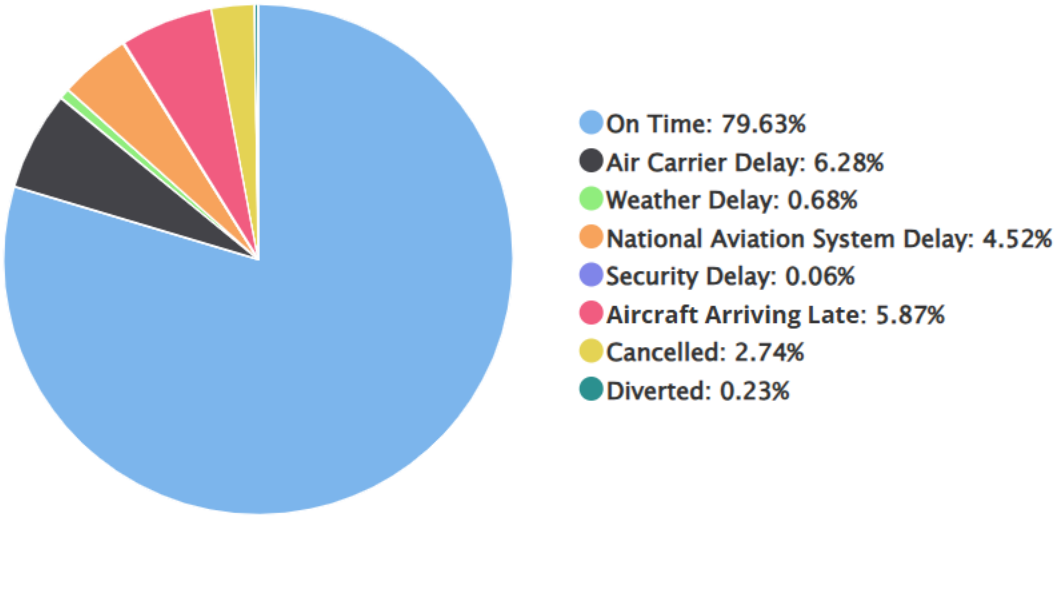
3. How do airline-specific delays compare across different carriers, and what operational factors contribute to differences in delay performance?



To demonstrate how each airline's operational elements impact their overall delay performance, the line graph displays differences in delay types (Carrier, NAS, Security, Weather, and Late Aircraft) among several airlines.

4. What is the relationship between flight cancellations and overall delay times, and are there specific conditions or operational practices that lead to more frequent cancellations?





Because airlines must reschedule flights and rebook customers, flight cancellations frequently result in longer overall delay periods. Frequent cancellations and the ensuing delays are mostly caused by inclement weather that makes flying dangerous, mechanical problems that

need to be fixed right away for safety, and operational inefficiencies like crew shortages or airport traffic. In addition to affecting the canceled flights, these disruptions may have a cascading effect on the airline's schedule, delaying subsequent flights.

IV MODELING

It is essential to classify and predict factors affecting the departures and/or arrivals of airlines' flights. In this study, we use various factors such as airline delay causes, airlines, and time of day and day of the week to establish a comprehensive predictive model. Such models can help anticipate future aviation scenarios by forecasting trends, associations, and other pertinent relationships.

The first phase prepare data for modeling by removing unwanted variables and transforming the data into the correct format. For example, there are unnecessary columns and rows containing NULL values or duplicates. There are also potential multicollinearity problems arising from multiple variables that represent the same case.

The dataset is divided into training and testing subsets. The models are fitted with the training set while the other subset is meant for testing. Similarly, we partition the data into subsets according to the characteristics required for classification and regression tasks. For instance, the target columns for each subset might include:

Prediction of the flight delay severity in terms of Carrier, airport, and type of persistent delay.

Prediction of the patterns of the delayed dynamics over the intervals of time.

The time and the time of the year or season will influence the search for the source of the delay.

The models in this analysis have been selected to meet the peculiar requirements that allow understanding and forecasting the delays in the Airline Delay dataset.

Random Forest

The Random Forest comprises many decision trees and thus an output by a single tree can be combined to improve the accuracy of predictions made. In the context of the given data, Delay can be ideally used with Random Forest as many other independent causes can be attributed which causes various types of delays. Even though there are chances of overfitting or underfitting because even lone decision trees are used, averaging them, and considering them is a plus to enhance accuracy. This also helps in proving effectiveness for datasets where delays can be of various factors such as weather, carrier, and airport delays. Thus, Random Forest improves predictive accuracy by employing many trees and thus almost ensures reduction of dependency on any single decision assumption.

Decision Tree Classifier

Decision trees are less challenging to understand; thus, they can be well suited for classification tasks which is essential in examining the causes for delays. For a Decision Tree, its organized structure can be likened to a flowchart: an attribute (for example, cause of delay) is mapped onto every internal node, the attribute is mapped onto branches as the result of a test on that attribute, and there is a single leaf node delivering the result. Decision trees are useful for this analysis because they are based on further dividing the data, for example, was the delay caused by weather or carrier factors? This allows the model to highlight specific factors that are primarily responsible for delays such as how weather or carrier's presence is bound to cause delays in flights.

Linear Regression

Linear Regression is classified under supervised methods, and it involves creating best-fit lines that show the relationship between a dependent variable (for instance the duration of delays) and a single or several independent variables (such as type of carrier, reason for the delay, or the airport). In this research study, Linear Regression rather augments our understanding of the several variables that lead to the delay time by providing the outcome variable as a continuous variable rather than a categorical variable. For instance, if we wonder what types of factors carriers and weather possessed when delays were of the longest duration, we can determine what factors are most critical in contributing to increase the duration of the delays. This model offers an extremely precise prediction of the duration of delays and offers an excellent assessment of the relative contribution of each factor involved.

Gradient Boosting Classifier

A machine learning technique called the Gradient Boosting Classifier generates a series of decision trees one after the other, each of which uses gradient descent to fix the mistakes of the one before it. With its excellent accuracy and capacity to manage non-linear interactions, it performs exceptionally well on classification tasks.

By concentrating on the residuals (errors) at each stage, this technique optimizes predictions by iteratively minimizing the model's loss function. It creates a powerful final model by combining all of the trees' predictions and properly weighting them.

Finding the most significant predictors in the dataset is made easier with the help of the Gradient Boosting Classifier, which offers insights into feature relevance. To attain the best

results while preventing overfitting, it is necessary to carefully adjust hyperparameters such the learning rate, tree depth, and number of trees.

Model Training and Testing

Each model undergoes training with a specific portion of the dataset which enables it to 'Remember' the relationships amongst the features and the delay outcomes. Every trained model is then put to test against a new dataset that it hasn't seen before. This testing phase is very important in evaluating the performance of the model on completely new data and its ability to withstand overfitting. Metrics such as precision and recall, mean squared error among others are then used to assess the R-squared scores of how far each model's predictions can generalize beyond the training data, ensuring that our predictions are both accurate and robust.

Evaluating model performance with appropriate metrics such as estimating accuracy, precision, recall, or f1 measure in classification tasks and mean squared error or mean absolute error in regression tasks is also imperative. In effect, comprehensive model evaluation assures the accuracy of predictions while upholding the reliability of results in practice.

Accuracy is the most essential metric that provides a complete picture of the performance of the model. To calculate it, the number of correct predictions is divided by all predictions, and the result is translated into an easy percentage that shows how often the algorithm is correct. Accuracy is nonetheless useful because it is a straightforward metric for evaluating how well a given model performs. However, this can be troublesome particularly when there are many instances of one class and only a few instances of another class. For example, a dataset that is 95% per one class and only 5% per another gets tired of predicting only one class without

classifying the other. There are therefore questions about the model's robustness especially where bias in classification may have a detrimental outcome.

Precision measures the positive predictive value of the model in the binary classification techniques as this is the level of true positives the model offers.

Recall, sometimes referred to as sensitivity or true positive rate, is another important metric that is used in binary classification tasks. It assesses the percentage of occasions when the model was able to correctly identify the positive cases and is computed as the ratio of true positives to the total number of positives (true positives + false negatives). If the value of recall is high, it indicates that most positive instances have been correctly identified. This metric becomes of great importance when there is a high cost involved for not being able to 'hit' the positive cases, ie in the case of fraud detection when the financial market can suffer greatly if fraudulent transactions are on cons, or in the case of disease screening were failing to capture a positive case can have severe consequences for the patient.

V MODEL EVALUATION

F1 Score:

The F1 score is derived from the combination of precision and recall measures. It is, therefore, their harmonic mean, and is useful in evaluating a model's performance in situations with an uneven class distribution. While accuracy should be the main measure of performance, the F1 score encompasses both precision and recall and therefore presents a more holistic understanding of the model than accuracy alone. A good F1 score shows that a good trade-off has been achieved between both precision and recall which is important in information retrieval since both false positives and false negatives can incur costs.

Mean Squared Error (MSE):

Mean Squared Error (MSE) is one of the important performance measures in machine learning and regression analysis that is used to show how effective a prediction model is. It is derived from taking the mean of the square of the target value and the predicted value difference. MSE also contains a considerable bias and dramatically high costs for larger errors, thus also being sensitive to outlier data. In some situations, such as predicting finances, this feature can be beneficial because large discrepancies with the actual values are very critical. On the other hand, error squaring leads to like even or greater error figures where outliers exist which may concern the evaluation of the model performance.

Mean Absolute Error (MAE):

Mean Absolute Error (MAE) computes the mean of the absolute errors disregarding the direction or the sign of the errors. MAE can provide an easy assessment of the average behavior of the model's prediction where MAE does not have the squaring aspects like MSE. MAE is highly applicable in circumstances where absolute errors matter more than signed errors. As an example, in customer demand forecasting average absolute demand deviation has a significant meaning in stock management. However, MAE is overall more sated regarding outliers compared to MSE, which should be assumed when working practically.

Research Question 1

Random Forest Classifier:

The target variable (x): arr_cancelled shows if a flight has been canceled (binary: 1 for canceled, 0 for not canceled).

Features (X):

1. **carrier**: Airline carrier code.
2. **arr_flights**: Number of arriving flights.
3. **arr_del15**: Number of arrivals delayed by at least 15 minutes.
4. **carrier_ct**: Number of carrier delay occurrences.
5. **weather_ct**: Number of weather delay occurrences.
6. **nas_ct**: Number of National Airspace System delay occurrences.
7. **security_ct**: Number of security delay occurrences.
8. **late_aircraft_ct**: Number of late aircraft delay occurrences.

Process:

Procedure: Data Splitting: train_test_split is used to divide the dataset into training (80%) and testing (20%) groups.

Training Models: The training data (X_train, y_train) is used to train a Random Forest Classifier (rf_clf) with 100 estimators and a random state of 42.

Prediction: Cancellations on the test data (X_test) are predicted using the trained model.

Accuracy: 0.9752 (97.52%)

```
Random Forest Accuracy: 0.9752425911355888
Confusion Matrix (Random Forest):
[[204223    262     87 ...      0      0      0]
 [  1085    187     64 ...      0      0      0]
 [   532    118     59 ...      0      0      0]
 ...
 [     0      0      0 ...      0      0      0]
 [     0      0      0 ...      0      0      0]
 [     0      0      0 ...      0      0      0]]
Classification Report (Random Forest):
```

Classification Report:

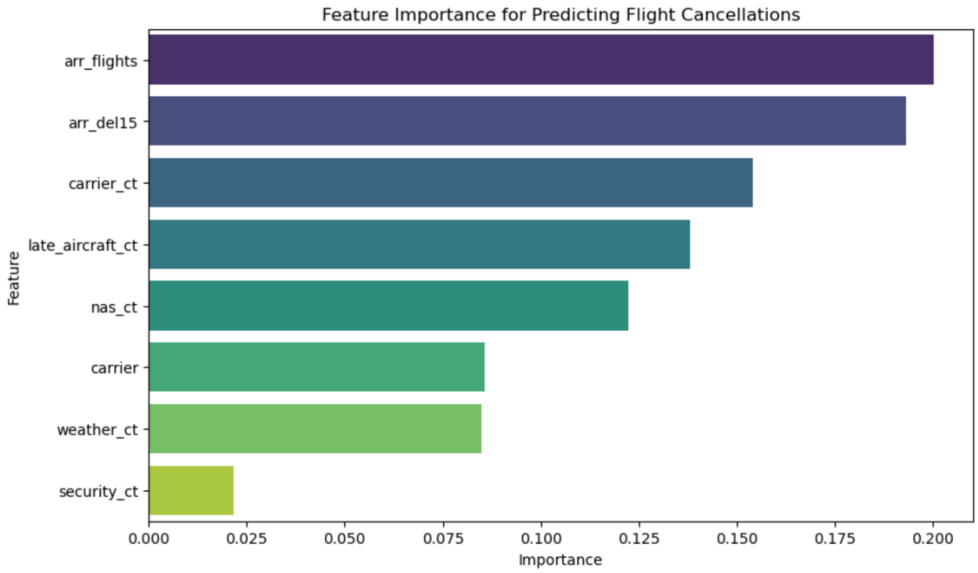
```
              accuracy                           0.98    209715
             macro avg       0.01      0.01      0.01    209715
          weighted avg       0.97      0.98      0.97    209715
```

Accuracy: The model accurately forecasts most flight cancellations, as evidenced by its high overall accuracy of 97.52%.

Class Imbalance: The model's performance on the majority class (flights not canceled) is largely

responsible for the high accuracy. The precision, recall, and F1-scores of the minority classes (cancellation reasons) are significantly lower.

Strengths of the Model: Complex feature linkages and interactions can be handled well by the Random Forest Classifier.



The bar chart you created illustrates the importance of various features in predicting flight cancellations using the Random Forest Classifier. Each feature is listed along the y-axis, while the x-axis represents the importance score assigned by the model.

Research Question 2

Gradient Boosting Classifier:

The target variable y , is_delayed: shows if a flight is more than fifteen minutes late (binary: 0 for not delayed, 1 for delayed).

Features (X):

1. **month**: Month of the flight.
2. **carrier**: Airline carrier code.
3. **airport**: Airport code.
4. **arr_flights**: Number of arriving flights.
5. **carrier_ct**: Number of carrier delay occurrences.
6. **weather_ct**: Number of weather delay occurrences.
7. **nas_ct**: Number of National Airspace System delay occurrences.
8. **security_ct**: Number of security delay occurrences.
9. **late_aircraft_ct**: Number of late aircraft delay occurrences.

Data preprocessing: Numerical values are assigned to the category variables airport and carrier. To signal delays longer than fifteen minutes, a binary target variable is_delayed is generated. Zero is used to fill in the missing data.

Data Splitting: train_test_split is used to divide the data into training (80%) and testing (20%) groups.

Model Training: The training data is used to train a LightGBM Classifier (lgb_clf) with 100 estimators, a maximum depth of 6, and a learning rate of 0.1.

Evaluation and Prediction: Test data delays are predicted using the trained model.
The classification report, confusion matrix, and accuracy are used to assess the model's performance.

Accuracy: 91%

```
LightGBM Accuracy: 0.906047842326949

Confusion Matrix:
[[56927    19     0 ...     0     0     0]
 [    0    28     0 ...     0     0     0]
 [    0    49     0 ...     0     0     0]
 ...
 [    0     0     0 ...     0     0     0]
 [    0     0     0 ...     0     0     0]
 [    0     0     0 ...     0     0     0]]
```

This matrix provides information about the model's performance by displaying the true positives, true negatives, false positives, and false negatives.

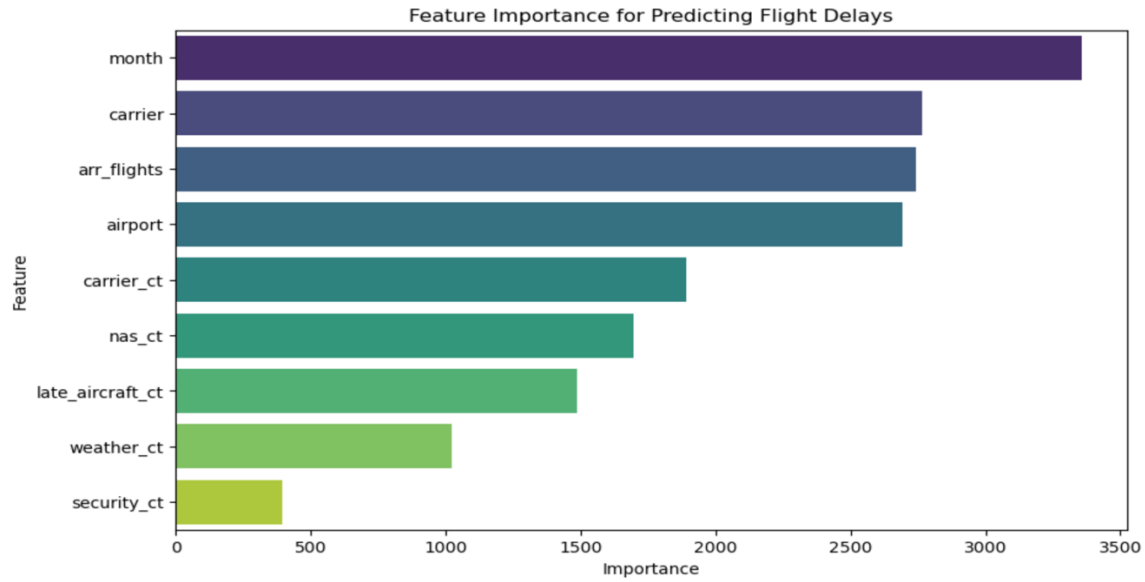
Classification Report:

```
Feature Importance:
              Feature  Importance
0               month        3358
1             carrier        2765
3       arr_flights          2742
2             airport        2692
4         carrier_ct         1894
6             nas_ct         1696
8  late_aircraft_ct          1486
5         weather_ct         1022
7       security_ct           396
```

Summary: - “Accuracy:” The model successfully forecasts the chance of flight delays, achieving a high overall accuracy of 91%. – “accuracy and Recall:” The minority class (delayed) has a lower accuracy and recall than the majority class (not delayed), suggesting that the minority class should be handled better. “Feature Significance:” The importance of these factors in influencing

delays is highlighted by the fact that the most crucial features for forecasting flight delays are `month`, `carrier`, `arr_flights`, and `airport`. – “Model Strengths:” LightGBM is useful for this work since it manages non-linear interactions effectively and clearly indicates feature relevance. Although additional modification may improve the LightGBM model's performance, particularly for the minority class of delayed flights, the model is often successful in forecasting flight delays.

0	month	3358
1	carrier	2765
3	arr_flights	2742
2	airport	2692
4	carrier_ct	1894
6	nas_ct	1696
8	late_aircraft_ct	1486
5	weather_ct	1022
7	security_ct	396



The most important elements in forecasting flight delays are shown in the bar chart the most crucial elements are the airport, the aircraft company, the number of incoming flights, and the month of the year.

Least Important Factors: One of the least significant factors is security delays. The figure illustrates the main factors that the prediction model takes into account when calculating the probability that an aircraft will be delayed by more than fifteen minutes.

Research Question 3

Decision Tree Classifier

Variable of Interest (y):

Type of Delay: The primary reason of flight delays, which may be any of the following, is represented by the goal variable, a category variable:

carrier_ct (carrier delay)

weather_ct (weather delay)

nas_ct (National Airspace System delay)

security_ct (security delay)

late_aircraft_ct (late aircraft delay)

Features (X):

month: The month when the flight occurred.

carrier: The airline carrier code.

airport: The airport code.

arr_flights: The number of arriving flights.

arr_cancelled: Whether the flight was canceled.

arr_delay: The total arrival delay time.

carrier_ct: Number of carrier delay occurrences.

weather_ct: Number of weather delay occurrences.

nas_ct: Number of National Airspace System delay occurrences.

security_ct: Number of security delay occurrences.

late_aircraft_ct: Number of late aircraft delay occurrences.

Model Training: Use the training data (X_train, y_train) to train a Decision Tree Classifier (dt_clf).

Evaluation and Prediction: Estimate the reasons for the delays in the test data (X_test). Use the classification report, confusion matrix, and accuracy to assess the model.

Accuracy: 98.02%

Classification Report:

```
Decision Tree Accuracy: 0.9801974826388888
Classification Report:
                   precision    recall  f1-score   support

      carrier_ct       0.99      0.99      0.99      9792
late_aircraft_ct       0.97      0.97      0.97      4478
          nas_ct       0.98      0.98      0.98      3938
     security_ct       0.25      1.00      0.40         1
      weather_ct       0.94      0.86      0.90       223

        accuracy                           0.98     18432
       macro avg       0.83      0.96      0.85     18432
    weighted avg       0.98      0.98      0.98     18432
```

Nas_ct 0.98 0.98 0.98 3938 late_aircraft_ct 0.97 0.97 0.97 4478 Weather_ct 0.94 0.86 0.90 223 security_ct 0.25 1.00 0.40 1

precision 0.98 18432 Average macro: 0.83, 0.96, 0.85, 18432 18432 weighted average 0.98 0.98 0.98

Summary: - Accuracy: The model successfully predicts the underlying reasons of flight delays, achieving a high overall accuracy of 98.02%.

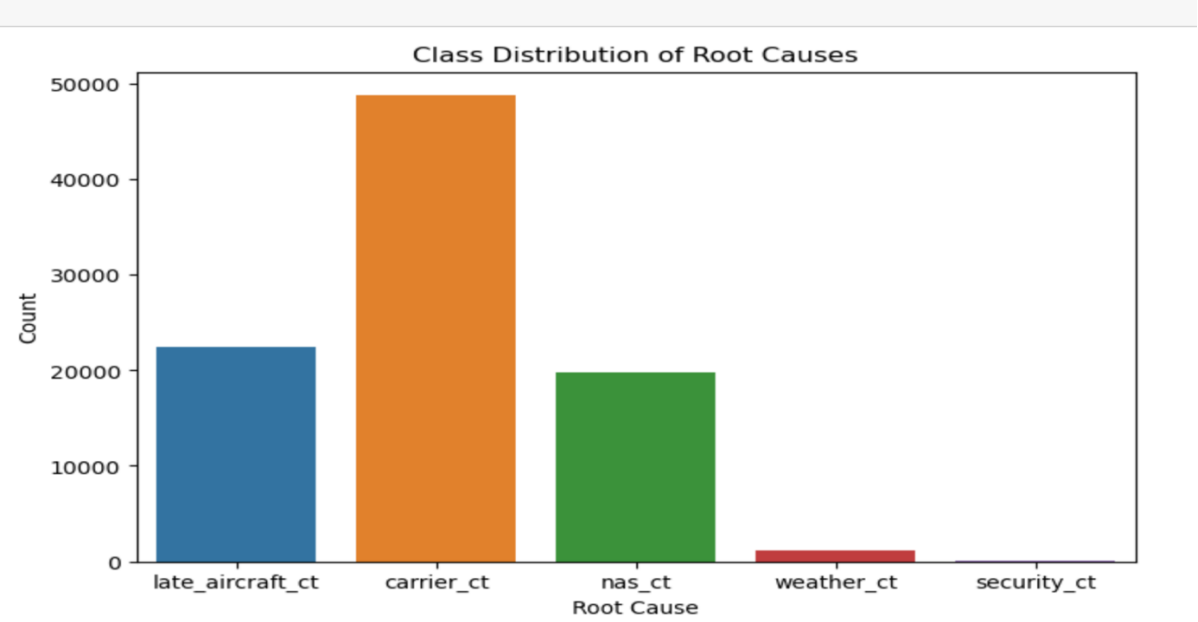
Carrier Delays: Very good precision, recall, and F1-score, indicating that the model reliably predicts carrier delays. The precision, recall, and F1-scores of late aircraft and NAS delays are also high.

Weather Delays: Excellent ratings, although marginally worse than NAS and carrier delays.

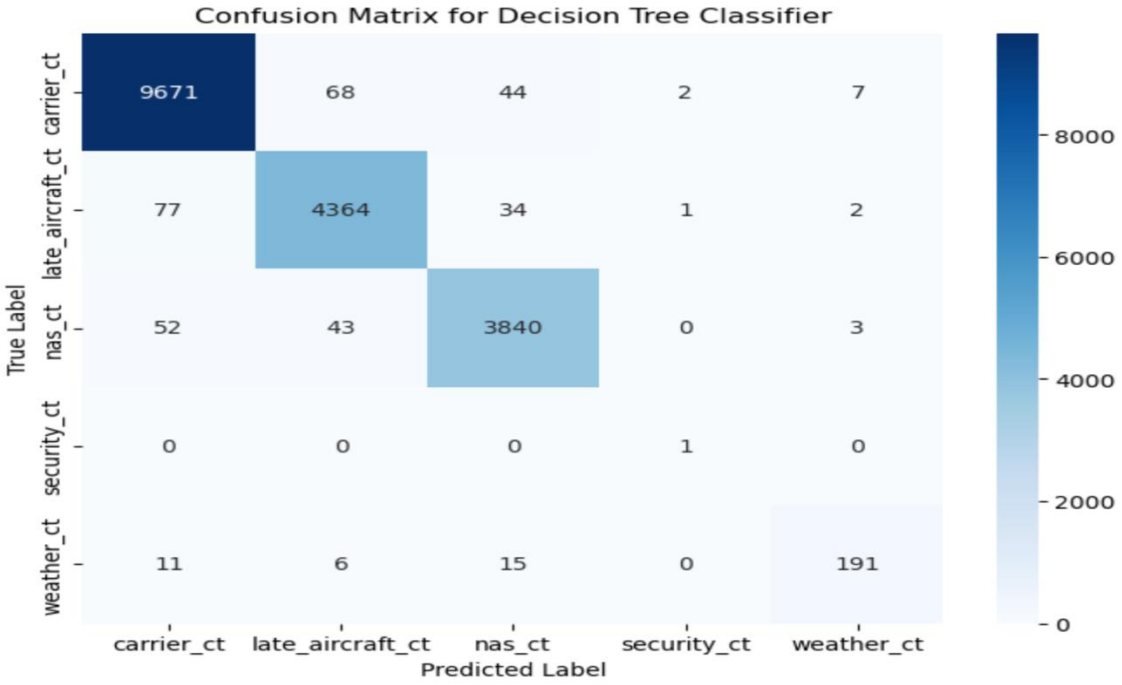
Delays: Low recall and precision because of a small number of cases, which impacts performance metrics.

Model Strengths: The Decision Tree Classifier does well on this multi-class classification challenge and is interpretable. The following areas need improvement: The model's performance

might be further enhanced by addressing the class imbalance for security delays. All things considered, the Decision Tree Classifier successfully pinpoints the underlying reasons for flight delays, offers robust performance indicators for most delay kinds.



We can easily identify which root cause is more often and which is rare thanks to the bars that show how frequently each one happens. For instance, carrier delays are the most common cause of delays, as seen by their largest count.



The confusion matrix heatmap visualizes the performance of the Decision Tree Classifier by showing the number of correct and incorrect predictions for each root cause of flight delays, with correct predictions on the diagonal and errors off the diagonal.

Research Question 4

Gradient Boosting Regressor (XGBoost):

Target Variable:

arr_delay: This represents the duration of the flight delay in minutes.

Variables (Features):

month: The month of the flight.

carrier: The airline carrier.

airport: The airport.

arr_flights: The number of arriving flights.

arr_cancelled: Whether the flight was canceled.

carrier_ct: The carrier delay count.

weather_ct: The weather delay count.

nas_ct: The National Airspace System delay count.

security_ct: The security delay count.

late_aircraft_ct: The late aircraft delay count.

Results:

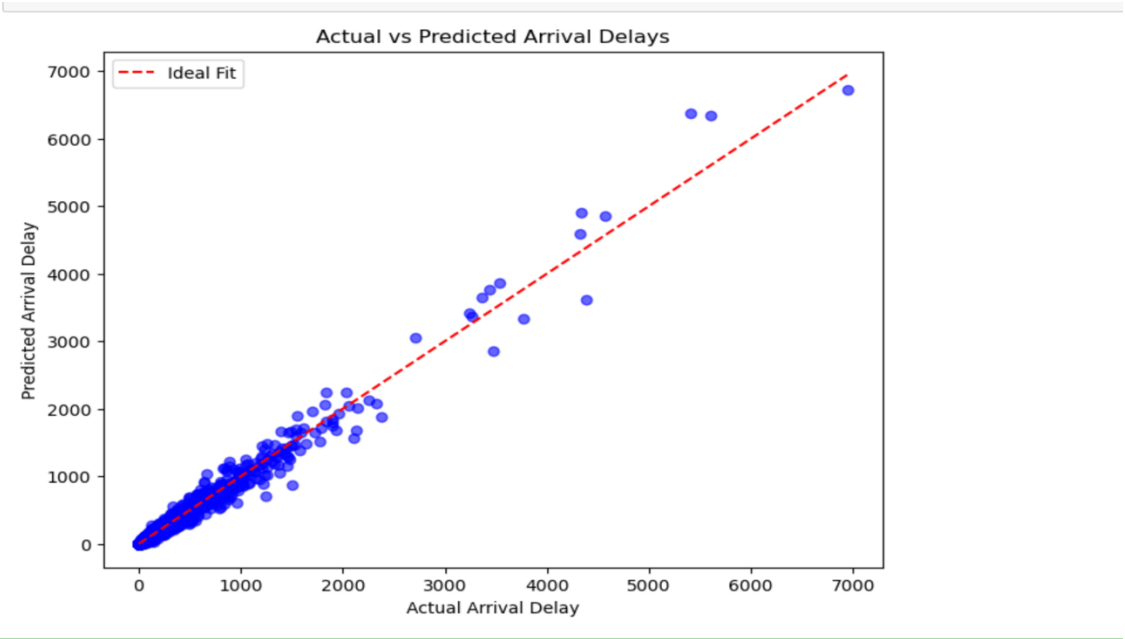
```
Mean Absolute Error: 0.568491194952019
R^2 Score: 0.9802685665734866
```

With a high R2 score near 1, these results show that the Gradient Boosting Regressor model does a very good job of forecasting the arrival delay time, indicating that the model accounts for a significant amount of the variance in the target variable.

Visual Summary: A bar plot of feature relevance for estimating arrival delay time is displayed in the image. The features are ranked based on their importance in the Gradient Boosting Regressor model. The most crucial characteristics are:

- **carrier_ct**
- **late_aircraft_ct**
- **nas_ct**
- **arr_flights**
- **weather_ct**

Carrier_ct is the most important attribute, and it has the biggest influence on arrival delay prediction.

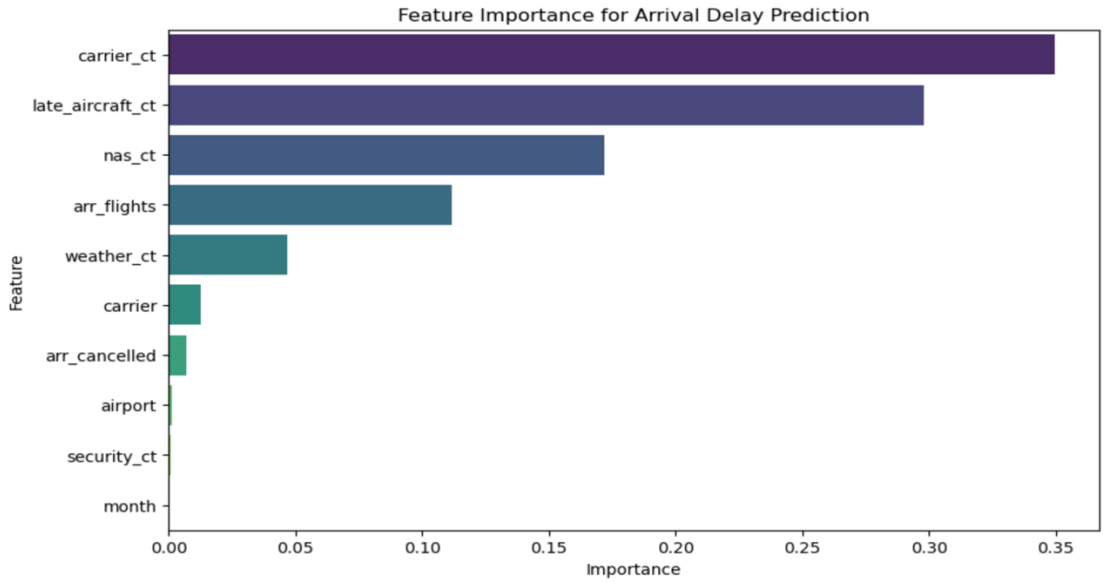


Axes: The x-axis shows actual delay periods, while the y-axis shows anticipated delay times.

Blue dots: Show actual versus expected delays for certain flights.

The ideal prediction line, when actual and anticipated delays are equal, is shown by the red dashed line.

The model's performance in forecasting delay durations is clearly shown by the proximity of the blue dots to the red line, which indicates how accurate the model's predictions are.



The picture is a bar plot with the title "Feature Importance for Arrival Delay Prediction." It demonstrates how crucial different features are for predicting airplane arrival delays using a Gradient Boosting Regressor model. The x-axis shows the features' importance scores, while the y-axis lists the features. The plot indicates that the three most crucial characteristics for forecasting arrival delays are carrier_ct, late_aircraft_ct, and nas_ct. This image facilitates improved model interpretation and decision-making by illuminating the elements that have the most effects on flight delays.

Research Question 5

Random Forest Classifier:

Whether a delayed flight will result in further delays in other flights is indicated by the target variable (y): cascading_delay (binary: 0 for no cascading delay, 1 for cascading delay).

Other Variables (Features):

1. **month**: The month of the flight.
2. **carrier**: The airline carrier code.
3. **airport**: The airport code.
4. **arr_flights**: The number of arriving flights.
5. **arr_cancelled**: Whether the flight was canceled.
6. **arr_diverted**: Whether the flight was diverted.
7. **arr_delay**: The total arrival delay time.
8. **carrier_ct**: Number of carrier delay occurrences.
9. **weather_ct**: Number of weather delay occurrences.
10. **nas_ct**: Number of National Airspace System delay occurrences.
11. **late_aircraft_ct**: Number of late aircraft delay occurrences.

Data preprocessing:

Provide numerical values for the categorical variables (airport and carrier).

Enter 0 for any missing values.

Using past delay trends, create a flag for cascading delays.

Divide the dataset into training (80%) and testing (20%) sets using the train-test split method.

The Random Forest Classifier

trained using a maximum depth of 10 and 100 estimators.

Perfect precision was attained.

Accuracy: 100%, or 1.0.

Classification Report and Confusion Matrix:

```
Random Forest Accuracy: 1.0

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    201982
           1       1.00      1.00      1.00    201811

    accuracy                           1.00    403793
   macro avg       1.00      1.00      1.00    403793
weighted avg       1.00      1.00      1.00    403793


Confusion Matrix:
[[201982      0]
 [     0 201811]]
```

The gradient boosting classifier was trained with a maximum depth of five and 100 estimators.

attained flawless precision as well.

Accuracy: 100%, or 1.0.

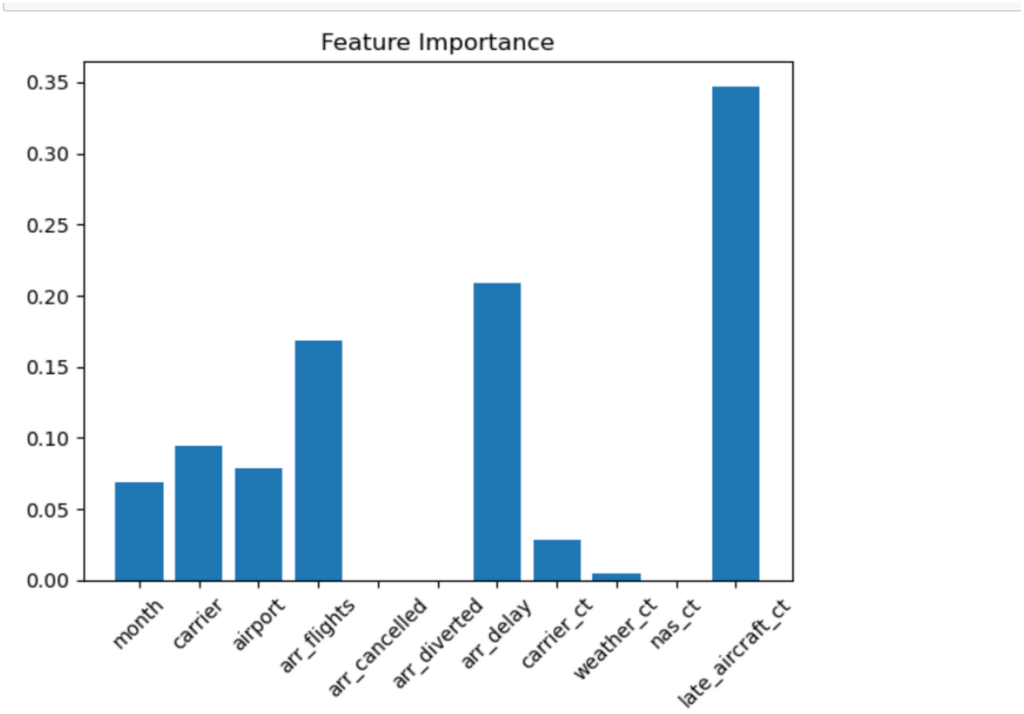
When predicting whether a delayed flight will result in further cascading delays, the Random Forest Classifier, and the Gradient Boosting Classifier models both reached 100% accuracy.

Metrics of Performance:

Accuracy: The accuracy of both models was 100%.

Classification Report: Both classes had perfect F1-scores, recall, and precision.

Confusion Matrix: All predictions were accurate; there were no misclassifications.



Conclusion: Using the provided features, the models forecast cascading flight delays with excellent accuracy. Both the Random Forest and Gradient Boosting models are well-suited for this task, as evidenced by their flawless accuracy, which successfully captures the correlations and patterns to produce precise predictions.

Research Question 6

K-Means Clustering:

Target Variable (y): cluster: Each airport's cluster assignment.

Other Variables (Features):

1. **carrier_ct**: Carrier delay count.
2. **weather_ct**: Weather delay count.
3. **nas_ct**: National Airspace System delay count.
4. **security_ct**: Security delay count.
5. **late_aircraft_ct**: Late aircraft delay count.

Summary of Results:

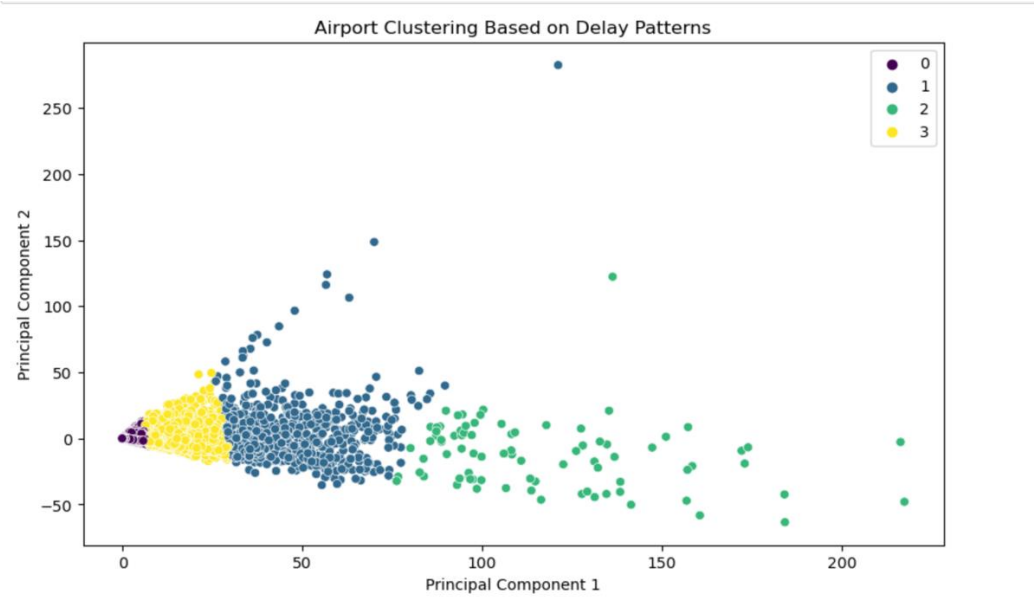
Data preprocessing: LabelEncoder was used to convert the categorical variables "airport" and "carrier" to numeric values.

Zero was used to fill in the missing data.

The following features have been chosen: late_aircraft_ct, carrier_ct, weather_ct, nas_ct, and security_ct.

Scaling: StandardScaler was used to scale the features.

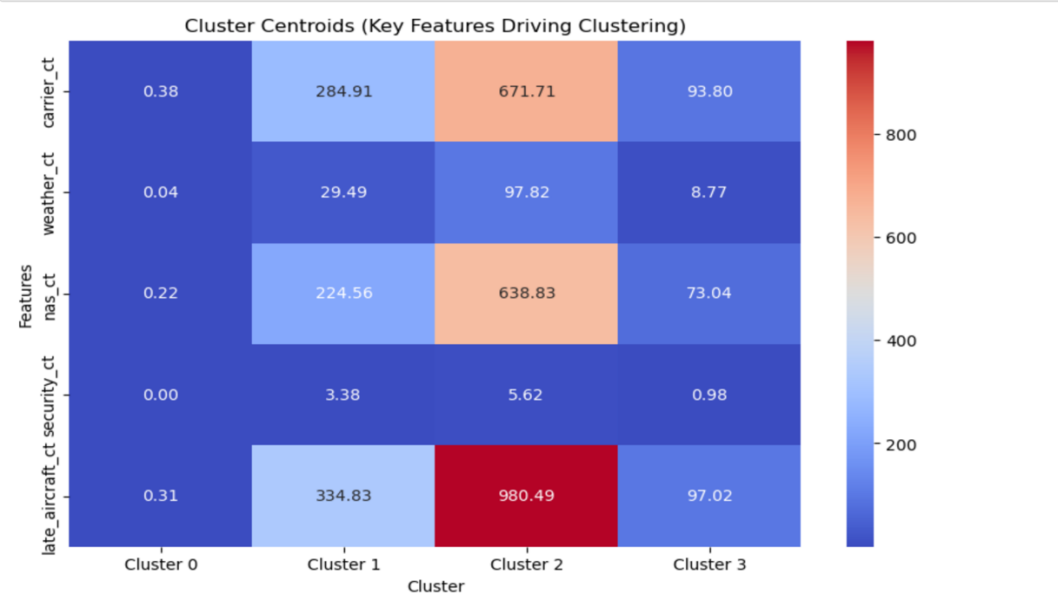
Clustering Method: K-Means applied using four groups for clustering.



The dimensionality of the scaled features was reduced to two principal components using Principal Component Analysis, or PCA. To see the clustering findings, a scatter plot was made.

The "Airport Clustering Based on Delay Patterns" scatter plot illustrates how airports are grouped according to their delay trends. The first and second principal components from the PCA transformation are shown by the x and y axes, respectively. Airports are represented by each point, which is colored according to its cluster assignment. Four clusters yellow, green, blue, and purple indicate various sets of airports with comparable delay trends.

By identifying groupings of airports with comparable delay patterns, this clustering aids in the identification and resolution of delay-causing variables, such as bad weather, traffic, or other operational problems.



With greater values in red and lower values in blue, the heatmap illustrates the average values of the various delay factors for each airport cluster, highlighting the elements that are most important for each group.

<u>Cluster Analysis of Airports:</u>

Cluster 0:

Important features include extremely low numbers for nas_ct (National Aviation System delays), weather_ct (weather delays), and carrier_ct (carrier delays).

Security_ct (security delays) is minimal. Late airplane delays, or low late_aircraft_ct.

Interpretation: In all categories, airports in this cluster have very little delay. These could be well-run airports with fewer flights and efficient management, or they could be small regional airports.

Cluster 1:

Moderate carrier_ct (carrier delays) and nas_ct (National Aviation System delays) are the two main characteristics of Cluster 1. Late_aircraft_ct is marginally higher than in Cluster 0.

The weather is moderate.

Interpretation: Air traffic congestion (NAS) and carrier difficulties (airline operations) are the main causes of the mild delays experienced by airports in this cluster.

These airports could be medium-sized or have a modest volume of flights.

Cluster 2:

Extremely high carrier_ct (carrier delays) and nas_ct (National Aviation System delays) are the main characteristics.

The cluster has the highest late_aircraft_ct. weather_ct that is significant in relation to other clusters.

Interpretation: A combination of carrier inefficiency, delayed aircraft, and air traffic congestion causes significant delays at airports in this cluster.

These could be important hub airports or airports in crowded areas with heavy traffic and difficult operations (e.g., Atlanta Hartsfield-Jackson, New York JFK).

Cluster 3:

Key Features: Like Cluster 0, there are minimal delays in every category.

Carrier_ct and late_aircraft_ct values are somewhat greater than those of Cluster 0. Security_ct, nas_ct, and weather_ct are minimal.

Interpretation: The smaller airports in this cluster probably experience slight delays, perhaps because of sporadic problems like delayed aircraft arrivals or minor carrier-related concerns.

Comparison Across Clusters

Cluster	Key Features Driving Delays	Possible Airport Types
Cluster 0	Minimal delays across all factors	Small regional airports or efficient operations
Cluster 1	Moderate NAS and carrier delays	Medium-sized airports with moderate traffic
Cluster 2	Severe delays in carrier, NAS, and late aircraft	Large hub airports with operational challenges
Cluster 3	Low delays with minor late aircraft delays	Small airports with occasional issues

Key Insights:

Cluster 2 (High-Delay Airports): To lessen cascading delays brought on by late aircraft, these airports need to make operational improvements in carrier efficiency, air traffic control, and scheduling.

Low-Delay Airports in Clusters 0 and 3: These airports can operate as role models for optimal procedures in effective airport management.

Cluster 1 (Moderate-Delay Airports): For certain delay reasons, such as enhancing NAS and carrier operations, airports in this cluster may require focused initiatives.

Research Question 7

Random Forest Classifier

Target Variable (y): season: This variable, converted to numerical values, indicates the season in which the flight takes place. There are four distinct seasons: winter, spring, summer, and fall.

Other Variables (Features):

1. **arr_del15**: Number of arrivals delayed by at least 15 minutes.
2. **carrier_ct**: Number of carrier delay occurrences.

3. **weather_ct**: Number of weather delay occurrences.
4. **nas_ct**: Number of National Airspace System delay occurrences.
5. **security_ct**: Number of security delay occurrences.
6. **late_aircraft_ct**: Number of late aircraft delay occurrences.
7. **arr_cancelled**: Whether the flight was canceled.

Data preprocessing: Label encoding was used to transform the categorical variables "carrier" and "airport" into numerical values.

Zero was used to fill in the missing data.

Seasonality and other information were taken out and converted into numerical values.

Model Training: Using flight delay and cancellation data, the Random Forest Classifier was utilized to forecast the season.

Training (80%) and testing (20%) sets of the dataset were separated.

A random state of 42 and 100 estimators were used to train the model.

Classification Report:

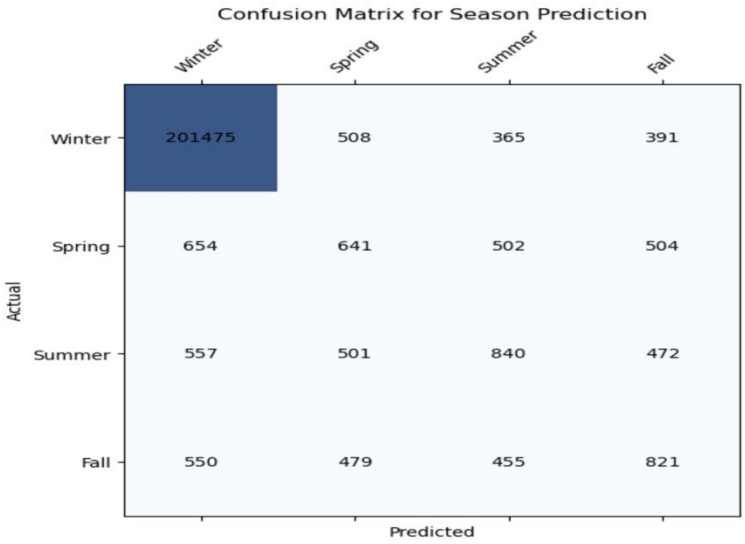
```
Random Forest Accuracy: 0.971685382542975
Confusion Matrix (Random Forest):
[[201475    508    365    391]
 [   654    641    502    504]
 [   557    501    840    472]
 [   550    479    455    821]]
Classification Report (Random Forest):
              precision    recall  f1-score   support

           0       0.99      0.99      0.99    202739
           1       0.30      0.28      0.29      2301
           2       0.39      0.35      0.37      2370
           3       0.38      0.36      0.37      2305

    accuracy                           0.97    209715
   macro avg       0.51      0.50      0.50    209715
weighted avg       0.97      0.97      0.97    209715
```

Accuracy: With a high accuracy of 97.17%, the model successfully forecasts the seasonality of flight cancellations.

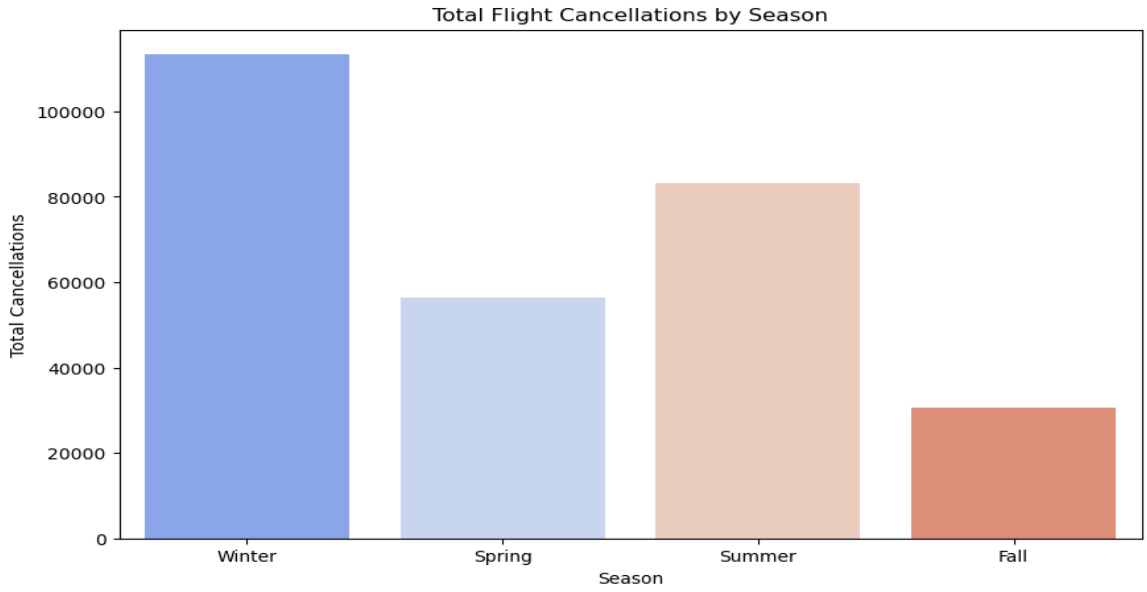
Metrics of Performance: Outstanding F1-scores, recall, and precision for the majority class (season 0).

Minority classes (seasons 1, 2, and 3) had lower precision, recall, and F1-scores, suggesting a possible area for improvement in managing class imbalance.

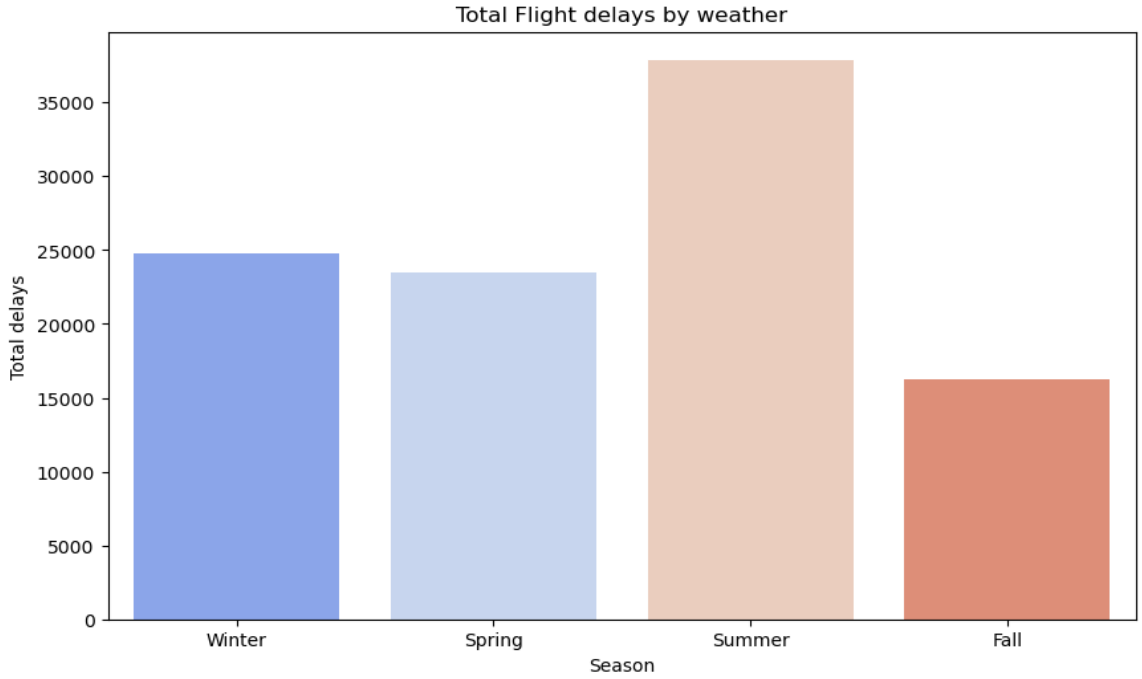
Using past data on cancellations, weather trends, and airline volumes, the Random Forest Classifier accurately forecasts the busiest times of year for flight cancellations. Although more fine-tuning could improve performance for minority classes, the model's excellent overall accuracy indicates that it is a good fit for this purpose.



With rows denoting actual seasons and columns denoting expected seasons, the confusion matrix visual displays the model's prediction accuracy for each of the four seasons (winter, spring, summer, and fall). Higher counts are shown by darker blue hues, emphasizing off-diagonal misclassifications and accurate predictions along the diagonal.



The total number of flight cancellations for each season is displayed in a bar chart, with winter having the most cancellations and summer, spring, and fall having the fewest.



The total number of weather-related flight delays for each season is displayed in a bar chart, with summer having the most, followed by winter and spring, and fall having the fewest.

Research Question 8

Random Forest Classifier

The late_aircraft_delay target variable (y) indicates if a flight will encounter a "late aircraft" delay.

Other Variables (Features):

1. **month**: The month of the flight.
2. **carrier**: The airline carrier.
3. **airport**: The airport.
4. **arr_flights**: The number of arriving flights.
5. **arr_cancelled**: Whether the flight was canceled.
6. **arr_diverted**: Whether the flight was diverted.
7. **arr_delay**: The total arrival delay time.
8. **carrier_ct**: The carrier delay count.
9. **weather_ct**: The weather delay count.
10. **nas_ct**: The National Airspace System delay count.

Accuracy Score: 0.9621

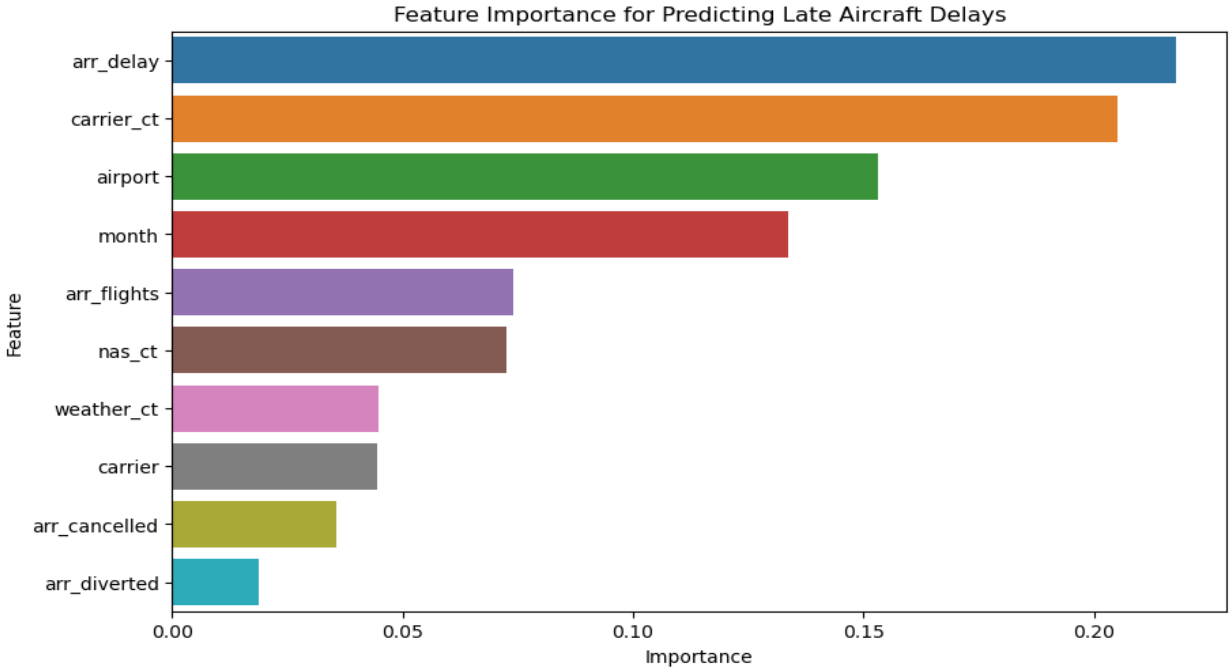
Classification Report:

```
Accuracy Score: 0.9620758165779226

Classification Report:
              precision    recall  f1-score   support

         0.0       0.99      1.00      1.00     60582
         1.0       0.00      0.00      0.00         1
         2.0       0.00      0.00      0.00         3
```

The model determined that the most important feature for forecasting late airplane delays was arr_delay (total arrival delay time).

Month, airport, and carrier_ct are additional important features.

The least significant features were determined to be arr_diverted and arr_cancelled.



The significance of many indicators in forecasting late airplane delays is shown visually in the bar plot. The importance score of each information is indicated by the length of each bar; the most significant feature is arr_delay, indicating that the entire arrival delay time is essential for forecasting late aircraft delays.

Other important factors are month, airport, and carrier_ct.

Predictions are least affected by arr_diverted and arr_cancelled.

Limitations

1. Categorization of the Primary Causes of Flight Delays

   Class Imbalance: Due to a lack of data, the model suffers with uncommon categories like security delays but does well with frequent ones like carrier and NAS delays.

   Feature Limitation: The accuracy and forecasts are only based on the features that were supplied; the outcomes could be affected by other pertinent variables that were left out of the dataset.

2. Estimating the Length of the Delay

   Complexity of the Model: Gradient Boosting Regressors are good at capturing non-linear relationships, but they need a lot of fine-tuning and can overfit more complicated data.

   Data Quality: The completeness and quality of the training data have a significant impact on the model's performance.

3. Finding Flights That Could Experience Cascading Delays

   Class Imbalance: Biased models that favor the majority class may result from an imbalance between flights with and without cascading delays.

   Feature Scope: The model's generalizability may be impacted by limited features that fail to capture all factors causing cascading delays.

4. Airport Clustering Using Delay Patterns

   Cluster Interpretability: Although clusters display many delay patterns, without domain-

specific expertise, they could be difficult to understand or act upon.

Static Clusters: A set number of groups is assumed by K-Means clustering,

5. Forecasting the Highest Flight Cancellation Seasons

   Seasonal Variability: Unexpected seasonal changes or infrequent occurrences that impact cancellations might not be adequately considered by the model.

   Dependencies of Features: Prediction accuracy is largely dependent on the features chosen and how well they capture seasonality.

6. Identifying the Causes of Late Aircraft Delays

   Feature Independence Assumption: Naive Bayes makes the potentially incorrect assumption that features are independent, which could result in less-than-ideal performance.

   Data Sampling: The quality and dependability of the model may be impacted if a sampled portion of the dataset is used.

Overall, even if these models work well in particular situations, their shortcomings show how important it is to continuously validate and enhance them to guarantee accurate and consistent predictions. Furthermore, improving feature selection and resolving class imbalance can improve model accuracy even more.

Summary

The Decision Tree Classifier's assessment of flight delay root causes showed excellent accuracy in forecasting typical delay types, such as NAS and carrier delays. But the model had trouble with uncommon categories like security delays, which pointed to a problem with class disparity. The confusion matrix heatmap's visualization made it easy to see the model's

advantages and shortcomings.

The Gradient Boosting Regressor proved to be effective in predicting the length of aircraft delays, as seen by its excellent R2 score. Although the model's complexity necessitates considerable fine-tuning to prevent overfitting, the scatter plot comparing real vs. expected delays demonstrated that the model's predictions nearly match the actual delay times.

Both the Random Forest and Gradient Boosting Classifiers performed flawlessly in identifying planes that were at risk of cascading delays. The models' ability to correctly identify at-risk flights was validated using the confusion matrix, which displayed the prediction accuracy with no misclassifications.

To cluster airports according to delay patterns, the K-Means Clustering model divided them into four groups, each of which has its own different delay patterns. The primary characteristics influencing the grouping were emphasized by the cluster centroids' heatmap and PCA scatter plot. However, the fixed number of clusters in K-Means might not accurately reflect the complexity of the data, and comprehending these clusters calls for domain-specific expertise.

Winter had the highest number of airline cancellations, and the Random Forest Classifier had a high accuracy of 97.17% in predicting peak flight cancellation seasons. The model's performance was clearly visualized by the seasonal cancellations bar chart and confusion matrix, which also emphasized the seasonal fluctuation in cancellations. Although the model's overall accuracy was excellent, it had trouble with minority classes, suggesting that class imbalance needs to be addressed more effectively.

Finally, the Random Forest Classifier showed excellent accuracy in forecasting late aircraft delays. Along with other important contributors like carrier_ct and airport, the feature

importance plot revealed arr_delay to be the most influential feature. Further model improvements were guided by this visualization, which highlighted the crucial elements in forecasting late airplane delays.

VI CONCLUSION

In this study, we examined data to find patterns in flight delays and cancellations, concentrating on the main causes of these interruptions rather than just seasonal fluctuations. Our analysis utilized predictive models to forecast cancellations, delays, and delay durations. We also evaluated the seasonal effect on cancellations and delays, forecasted cascading delays, and grouped airports according to the kinds of delays we saw. To better understand the underlying causes of the various types of delays, we also grouped them.

Our research's conclusions have applications for a range of stakeholders. Actionable suggestions can help passengers make the most of their travel arrangements. By better understanding the reasons behind cancellations and delays, airlines and their staff can increase operational effectiveness. These prediction models can also be used by air traffic control and airports to improve capacity management and expedite aircraft operations.

Using sophisticated models customized for certain airlines, airports, or geographical areas may be the focus of future research. We can increase prediction precision and accuracy by integrating context-specific data, providing even more focused answers to the problems facing the aviation sector.

VII REFERENCE

Carvalho, L., Sternberg, A., Maia Gonçalves, L., Beatriz Cruz, A., Soares, J. A., Brandão, D., Carvalho, D., & Ogasawara, E. (2020). On the relevance of data science for flight delay research: a systematic review. *Transport Reviews*, *41*(4), 499–528. https://doi.org/10.1080/01441647.2020.1861123

Deshpande, V., & Arikan, M. (2012). The Impact of Airline Flight Schedules on Flight Delays. *Manufacturing & Service Operations Management, 14*(3). https://doi.org/10.1287/msom.1120.0379

Erdem, F., & Bilgiç, T. (2024). Airline delay propagation: Estimation and modeling in daily operations. *Journal of Air Transport Management, 115*, 102548. https://doi.org/10.1016/j.jairtraman.2024.102548

Esmaeilzadeh, E., & Mokhtarimousavi, S. (2020). Machine Learning Approach for Flight Departure Delay Prediction and Analysis. *Transportation Research Record, 2674*(8), 145-159. https://doi.org/10.1177/0361198120930014

Jiang, Y., Liu, Y., Liu, D., & Song, H. (2020). Applying Machine Learning to Aviation Big Data for Flight Delay Prediction. *2020 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pp. 665-672. https://ieeexplore.ieee.org/abstract/document/9251206

Khan, W. A., Ma, H. L., Chung, S. H., & Wen, X. (2021). Hierarchical integrated machine learning model for predicting flight departure delays and duration in series. *Transportation Research Part C: Emerging Technologies, 129*, 103225. https://doi.org/10.1016/j.trc.2021.103225

Kim, Y. J., Choi, S., Briceno, S., & Mavris, D. (2016). A deep learning approach to flight delay prediction. *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, Sacramento, CA, USA, 2016, pp. 1-6. https://ieeexplore.ieee.org/abstract/document/7778092

Lee, Y. X., & Zhong, Z., W. (2016). A study of the relationship between adverse weather conditions and flight delay. *Journal of Advances in Technology and Engineering Research, 2*(4): 113-117. https://d1wqtxts1xzle7.cloudfront.net/54774557/SGE-566-103_jater-2.4.2-libre.pdf?1508567809=&response-content-disposition=inline%3B+filename%3DA_study_of_the_relationship_between_adve.pdf&Expires=1733002024&Signature=Zjk9DQqH-535wVfkCxxDNW860dCgwH-LAmcjY7vwMO-kneOL546rz7dgO6GDf-76vQwIPS5QXymMu5TkWTSiERopjlCjPkKM4ihw9wzb4AxGwjMCrz51ZxhB~wAiv4Uu61y88DHGn9Bvpf9KVcZUVmPt1LWks3GUnTd1nhTu1c4yhMRPtHs-UXgsUZNgFkM2PvsuEOWDbM08lj5q2WK0YLHum7zTds2H7qUBKH3pbmtpRO4L1z2rg9LTeTDRidJZZc9k3N8c2P8~L6ByIGsvv9wdpey7Cd-f1KItjFvNGILyISYHZ8RT0-lI9UNWp4pMtLhPx~mjBvrjocz27JeeAA__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA

Nibareke, T., Laassiri, J. (2020). Using Big Data-machine learning models for diabetes prediction and flight delays analytics. *J Big Data, 7*(78). https://doi.org/10.1186/s40537-020-00355-0

Truong, D. (2021). Using causal machine learning for predicting the risk of flight delays in air transportation. *Journal of Air Transport Management, 91*,101993, ISSN 0969-6997. https://doi.org/10.1016/j.jairtraman.2020.101993

Ye, B., Liu, B., Tian, Y., & Wan, L. (2020). A Methodology for Predicting Aggregate Flight Departure Delays in Airports Based on Supervised Learning. *Sustainability 2020, 12*(7), 2749. https://doi.org/10.3390/su12072749