# PREDICTING DELAYS AT LAX

MASTER'S THESIS (CDSCO4001E) – COPENHAGEN BUSINESS SCHOOL

**NAME:** MAX LADEGAARD

**STUDENT ID**: 110166

**SUPERVISOR:** ROBERT KAUFFMAN

**DATE:** 15TH OF MAY 2023

**PAGES**: 49

**CHARACTERS**: 109,266

## Abstract

This paper uses different machine learning algorithms to investigate how flight delays can be predicted at Los Angeles International Airport (LAX). LAX is a large international airport located around 30 kilometers from Downtown Los Angeles. It serviced more than 80 million passengers in 2019, which made it the second busiest airport in the United States. Like most other large international airports, LAX also struggles with flight delays and how to limit their negative impact. One of the potential solutions is to use machine learning to predict which flights are going to depart on time and which flights are going to be delayed. We built three models using three different machine learning algorithms. The goal of the models was to predict whether a flight would belong to one of four delay categories: on-time, small delay, medium delay, or large delay.

It was found that we could train a Neural Network model with an accuracy score of 95%. This was our most accurate model as it beat out a Random Forest model and an XGBoost model, which had accuracy scores of 89% and 92% respectively. In the analysis, it was discovered that weather was less important a factor than anticipated. Except for the adjusted XGBoost model, the weather features have a limited impact on the output of the models. At the other end of the spectrum, the flight's airline was a surprisingly important feature. Similarly, the time of departure was consistently one of the most essential features of the models.

In terms of practical uses, we found that a flight delay prediction model can be used to enhance operational efficiency, improve passenger experience, and optimize pricing strategies. Furthermore, there is a vast potential for using machine learning and artificial intelligence in other airport domains in the future as the industry needs efficient and scalable solutions.

# Table of Contents

# 1. Introduction

Airports can often take on an important role in a city's economy. The labor required to run an airport leads to thousands of jobs, while the constant flow of passengers creates opportunities for tourism and local businesses. A great example is Los Angeles International Airport, which is one of the busiest airports in the world with over 80 million passengers annually. The airport has approximately 4,000 employees and is the workplace for over 50,000 people. All of these employees are tasked with transporting millions of passengers and ensuring operations run smoothly. One of the biggest obstacles to an airport's operations is flight delays. Flight delays can come from many different factors, but they usually result in high costs for the airport. A single delayed flight might start a ripple effect that can cause multiple flights on the same day to be delayed. The issue is also exacerbated by the limited ability to anticipate delays.

In 2007, it was estimated that flight delays cost the US economy $32.9 billion (Federal Aviation Administration, 2010). This means that there is an enormous economic opportunity to improve the area and limit the future costs of delays. One of the ways that airports can attempt to predict flight delays is through the use of machine learning. The usage of machine learning has grown exponentially over the last decade with many firms relying on its capabilities for critical business infrastructure. The algorithms can find patterns in vast amounts of data, which can result in insights for stakeholders that were previously not possible or sustainable to create.

This paper will seek to answer how flight delays can be predicted using machine learning. With a focus on Los Angeles International Airport (LAX), we will attempt to build a model that can accurately predict whether a flight departure will be delayed. To ensure the highest possible accuracy, we will test different machine learning algorithms and compare the results. The process behind the model building will be explained throughout the paper. After describing the modeling process, it will be discussed how airports can use the insights of these models to improve their operations and how it might be possible to leverage machine learning and artificial intelligence in other airport domains too.

# 2. Literature and Background

This section will provide the background information needed to fully understand the project. We will review some of the previous works on the subject, and how we can use the findings of these works in our project. In the latter part of this section, we will present information about airports and how the operations of these normally function. LAX's infrastructure, route network, competitors, and challenges will also be covered.

## 2.1. Literature Review

Flight delay prediction is not a unique topic as other work has already been done on the subject. Similar to this paper, previous works have mainly focused on developing a model that can predict whether a flight will be delayed. Various algorithms, from logistic regression to deep learning models, have been used to create models, but there appears to be a weak correlation between more advanced methods and higher accuracy. It is important to note that the datasets used for these models have varied significantly too, so it is difficult to conclude anything definitively.

Qiang Li and Ranzhe Jing used a Random Forest Classifier to predict flight delays based on spatial, temporal, and extrinsic factors. The model was tested on domestic flights in China from June to August 2016 and yielded an accuracy score of 92.39% (Li & Jing, 2021). Warittorn Cheevachaipimol, Bhudharhita Teinwan, and Parames Chutima took a different approach to the issue of flight delay prediction as they tried to use a hybrid deep learning model. They investigated the method's effectiveness on data from 10 major US airports and found that it was more accurate than a standard Neural Network model. However, it also turned out that it was less accurate than a pure machine learning model (Cheevachaipimol, Teinwan, & Chutima, 2021). A different study used a Neural Network algorithm and found great success. Hajar Alla, Lahcen Moumon, and Youssef Balouki used data from the US Bureau of Transportation and trained a multilayer perceptron model. It yielded an accuracy score of 95.6%, which was higher than the Gradient Boosting and Decision Tree models that were also trained (Alla, Moumoun, & Balouki, 2021).

In terms of research on flight delay causes, there are varying opinions on which factors influence delays the most. Some studies point to weather conditions as one of the most significant factors

impacting delays. For example, a study by Stefan Borsky and Christian Unterberger found an increase in departure delays of up to 23 minutes from weather shocks like rainfall, snow, and wind (Borsky & Unterberger, 2019). A different study by Christopher Mayer and Todd Sinai found that air traffic congestion was the primary cause of flight delays. Mayer and Sinai explain that congestion happens due to over-scheduling, and that it is especially prevalent in connecting flights. The airports are interested in servicing connecting flights as the network effects for the airport will be stronger with more travel destinations. However, airlines and airports are also interested in cutting down passenger travel time, so they try to schedule the connecting flights in clusters. This creates congestion at certain peak times, which ultimately can lead to flight delays (Mayer & Sinai, 2003). Another paper found that flight delays were usually the result of a propagation of delays from previous flights. Martina Zámková, Martin Prokop, and Radek Stolín analyzed flights from Spanish airports, and while the frequency and length of delays varied from airport to airport, there was a common picture of longer delays being caused by the delays of previous flights. This was especially common in the evening and afternoon (Zámková, Prokop, & Stolín, 2019).

LAX has also been the focus of various research papers over the years. The topics of previous works can broadly be put into six categories: operations, environment, economics, security, infrastructure, and passenger experience. Research on operations and efficiency at LAX has been done with the aim of optimizing the airport's overall performance. Factors such as runway capacity, gate utilization, and passenger flow have been analyzed as well as the potential benefits of implementing new technologies or systems (Clarke, et al., 2013). Other papers have looked at the environmental impact of LAX. These have explored topics such as the airport's impact on air and noise pollution, the energy consumption of the airport, and how the airport manages waste (Westerdahl, Fruin, Fine, & Sioutas, 2008).

The economic impact of LAX has naturally also been an area of study for research papers. This has been examined on both a local and regional level with specific topics ranging from how the airport serves as an employment generator in the local community to how it impacts tourism in southern California (Oster & Strong, 2021). Airport security is another common research topic. LAX operates at a massive scale, so it is especially interesting to look at how the airport can improve the

effectiveness of its security measures; both in terms of heightened security and speed (Hamilton, Schell, Stevens, & Mesic, 2004).

The area of infrastructure has been particularly interesting for researchers as LAX has undergone multiple expansions during the last two decades. This has prompted research on how these expansions have impacted various factors such as operations, the environment, and passenger experience (Vascik & Hansman, 2017). Passenger experience is also a major research topic of its own as multiple papers have been written on passenger satisfaction and how airports can take measures to improve it through terminal design, amenities, or route network (Yamamoto & Paternoster, 2017).

## 2.2. Airports

The importance of airports has greatly increased in the last couple of decades as the world has become more globalized and the demand for travel has grown. Airports have become catalysts for economic growth, international trade, and cultural exchange . By connecting different countries and cities, airports help enable business development, tourism, and the exchange of ideas. This makes them indispensable for sustaining the globalized world that we live in today. In the United States alone, there are over 19,000 airports (Statista, 2023).

The main task of an airport is to facilitate the movement of passengers, aircrafts, and cargo. Airports consist of multiple components that must work together to ensure smooth and efficient operations. It is possible to divide these components up into airside and landside operations. Airside operations include air traffic control, aircraft maintenance, ramp services, and runway operations. Landside operations include terminal management, shops, restaurants, passenger transportation, and parking facilities (Horn & Orman, 1975). To create the traditional airport experience, all of these components have to come together.

Revenue for airports can similarly be split into aeronautical revenue and non-aeronautical revenue. Aeronautical revenue pertains to income generated from activities related to aircraft operations and services. This could be landing fees, terminal fees, passenger service charges, and hangar fees.

Non-aeronautical revenue is everything else, which usually includes income from retail stores, restaurants, parking, advertising, and property rentals (Los Angeles International Airport, 2023).

Airports can often become a central cog in the economy of a city. It allows people to easily access and leave the city, which brings opportunities for tourism and business collaboration. Without a large international airport, a city is much less likely to become a destination for tourists. This is among the reasons why countries like the United Arab Emirates and Qatar have invested huge sums of money into creating large international airports and financing national airlines to service routes to and from the airports (Lohmann, Albers, Koch, & Pavlovich, 2009). Airports also create a high number of jobs, both direct and indirect. The direct jobs are the ones where the airport is the employer. This could be baggage handling, facility management, or air traffic controllers. The indirect jobs relate to the jobs created because of the airport's existence, but where the airport is not the employer. This could be pilots and stewardesses at an airline, receptionists at car rental companies, or waiters at airport restaurants.

The largest airports are capable of servicing tens of millions of passengers annually, while others are only capable of handling a few thousand passengers. An airport's capacity is determined by various factors such as runway capacity, terminal and gate capacity, air traffic control staffing, infrastructure constraints, weather conditions, and environmental restrictions (Sweet Jr., 1975). In this paper, we will be looking at an airport with one of the largest passenger capacities in the world.

### 2.3. Los Angeles International Airport

Los Angeles International Airport, also referred to as LAX, is an international airport that is located in the Westchester neighborhood of Los Angeles. It was opened in 1928 and has since grown to become one of the busiest airports in the world. In 2019, LAX set a new record with 88 million passengers (Los Angeles World of Airports, 2022). This made them the second busiest airport in the United States, and the third busiest airport in the entire world (Airports Council International, 2020). The passenger numbers naturally saw a huge drop during the COVID-19 pandemic, but it has since rebounded and is working its way back to pre-pandemic levels (Los Angeles International Airport, 2023).

### 2.3.1. Infrastructure

LAX has nine different terminals with 146 gates spread across them. The terminals are arranged in a characteristic horseshoe shape with terminals 1-3 on one side and terminals 4-8 on the other side. In the middle is Tom Bradley International Terminal, also called Terminal B, which handles the majority of international traffic at LAX. All of the terminals are connected via either walkways or busses. The terminals serve various purposes, and some are almost exclusively dedicated to one airline. For example, Terminal 7 and Terminal 8 are only used by United Airlines.

There are four runways at LAX, which all run parallel in an east-west configuration. 25L/07R and 25R/07L are placed on the airport's southern side, while 24L/06R and 24R/06L are placed on the northern side. The runways are of relatively similar width, while the length ranges from 2700 meters to 3747 meters. It is estimated that the runways have a capacity of 147-153 operations per hour (Federal Aviation Administration, 2014).

LAX has not declared an official passenger capacity loft, but it has previously been estimated that the airport can handle around 100 million passengers annually (Weikel, 2015). This number is significantly higher than its current annual passenger total, but this does not necessarily mean that the airport is not straining its current resources. The number of gates and terminals might be capable of handling 100 million passengers annually. However, the airport would still need more personnel and more efficient surrounding infrastructure to handle more demand. For example, one of the main critiques about LAX is its lack of connection to public transportation. Almost all of the traffic in and out of LAX comes from cars, and it can take more than an hour to drive around the horseshoe (Yellow Productions, 2022). There is no metro station directly at LAX, which means that you will have to take a 15-minute shuttle from the airport to the LAX Metro Rail Station before being able to get on the metro line.

### 2.3.2. Route Network

LAX is in a privileged position as it is the main airport for one of the largest cities in the world. Besides being a large city, Los Angeles is also a major tourist destination. People come from all over the

world to see Hollywood, Beverley Hills, Santa Monica, and Disneyland. This means there is a large natural demand for flight routes in and out of Los Angeles. LAX is the airport that services this demand. It offers almost every flight route imaginable as it has direct non-stop flights to 182 destinations in 41 different countries. LAX is also a major player in terms of domestic flights as it currently offers 105 different domestic flights (FlightConnections, 2023).

Several major airlines in the United States have their main hubs at LAX. For example, American Airlines, United Airlines, and Delta Airlines all operate their primary base out of LAX. This is a hugely positive thing for an airport as it will mean more flights out of the location from these airlines. Many legacy airlines like United Airlines use the hub-and-spoke network model, which means that it flies passengers from spoke airports to the hub airport, where they will change flights and then go to their final destination. This gives travelers more options for destinations and flight times (Wall Street Journal, 2023). For LAX, this is an advantageous relationship as the airlines have to rent and pay for various infrastructure like terminals and rooms at the hub airport. Furthermore, the airline's crew and pilots will usually also be living in the area surrounding the hub airport, so it deepens the airline's dependency on the airport. It can be argued that an airline's main base is also picked based on natural travel demand, but it is nonetheless a factor contributing to flight routes staying at an airport.

### 2.3.3. Delays

As with many other large international airports, delays are unfortunately also a common issue at LAX. In a study by Forbes in 2018, it was estimated that about 19.5% of flights from LAX were delayed (Bloom, 2018). Delays at LAX can arise due to a variety of different factors, such as weather, technical issues, air traffic congestion, and security concerns.

Weather is usually not a significant factor for LAX as southern California is not prone to adverse weather conditions. However, the challenge related to weather usually revolves around strong winds. The airport borders the Pacific Ocean, and it can therefore be windy at times. This fact is part of why the runways are configured in an east-west orientation. In what LAX calls 'Westerly Operations', the planes take off over the Pacific Ocean, while the planes arrive from the east (Los

Angeles World Airports, 2022). When the conditions do not allow for Westerly Operations, the operating efficiency might be less than usual as it is less common for the airport to operate in other modes.

Technical issues encompass incidents such as unscheduled aircraft maintenance or system glitches. These are tough to predict and can often throw a wrench in the wheel of a smooth operation, which makes them perfect causes for delays. However, as airports have to follow a rigorous protocol and must prioritize passenger safety, these issues have to be handled before allowing flights to takeoff. Air traffic congestion, which can happen in the skies and on the ground, also contributes to delays at LAX. The airport operates multiple runways, but the high volume of flights can lead to bottlenecks at peak times. Furthermore, the location of LAX also means that it shares airspace with other nearby airports, which further complicates the management of air traffic.

#### 2.3.4. Competitors

An airport's competitive environment is different from that of a traditional business. Most airports are in the privileged position of not having significant competitors nearby. This allows these airports to access a pool of customers with limited alternatives. Airports do not have to fear that a hot new startup will emerge and take away its customers as there are significant regulatory and economic barriers to starting an airport. While this is an advantageous situation, it does not mean that an airport like LAX can rest on its laurels and wait for customers to come to them. Airports need to be able to attract airlines and have them offer the routes and flight times that their potential passengers want to use. If these are not available, passengers might decide to use a different airport or seek alternative modes of transportation.

The competitors to LAX can be divided into direct competitors and indirect competitors. The direct competitors are other airports, while the indirect competitors are alternatives to air travel. Other airports offer the same service as LAX, but there might be factors such as location, routes, and comfort, which set them apart in the eyes of the customers. The direct competitors can also be further differentiated as some airports are more direct competitors to LAX than others. For example,

San Francisco International Airport is significantly more likely to take away passengers from LAX than Newark Liberty International Airport.

The most direct competitors to LAX are airports that are located close to the larger Los Angeles area. These would be John Wayne Airport, Hollywood Burbank Airport, and Ontario International Airport. Due to their geographical location, these airports can compete for the same customer segment as LAX. The other type of direct competitors are airports that might be competing with LAX for connecting passengers. These airports are more spread out geographically as the competition depends on where the passengers arrive from and where they intend to go. Examples include San Francisco International Airport, San Diego International Airport, Harry Reid International Airport, and Denver International Airport.

Indirect competitors to LAX are alternative modes of transportation like trains or buses. For some routes, these alternatives do not pose much of a threat. For example, it is doubtful that someone is going to choose to take a combination of buses, trains, and ferries for a trip to Paris. However, the rising awareness of carbon footprint has made an increasing number of people consider taking trains or busses for shorter trips like from Los Angeles to San Francisco (Conboye & Hook, 2019). This might be a minor issue for LAX, but it could grow into a major one with time.

#### 2.3.5. Challenges

One of the most significant challenges for LAX is keeping up with demand. From 2009 to 2019, the airport saw an increase in annual passengers from 56 million to 88 million. The airport has expanded its infrastructure during this time, but this type of increase will nonetheless put a strain on resources. The steady rise in demand makes it imperative for LAX to remain efficient in its operations. This goes along with the airport's infrastructure challenges. LAX is an old airport with an aging infrastructure that requires continuous maintenance. The reduced capacity in times of maintenance will be a problem if it is not coordinated well. Furthermore, the surrounding roads are often heavily congested, and this will only worsen with rising demand. The issue with increased demand is therefore heavily connected with the airport's infrastructure challenges.

Environmental concerns are also starting to become an issue for airports around the world (Conboye & Hook, 2019). While flying is unlikely to fall out of favor in the next couple of years, it could be advantageous for airports to begin to consider the environmental ramifications of their operations. LAX have started to invest in various environmentally conscious projects such as electric vehicles, onsite renewable energy production, and a transit shuttle for workers to limit carbon emissions from commutes (City News Service, 2022).

# 3. Research Context

Flight delays have been a significant issue in the aviation industry for many years. It affects millions of passengers annually and causes massive economic losses for airlines and airports. In recent years, machine learning has emerged as a promising and popular approach to tackle this problem by offering a sophisticated and advanced method of predicting flight delays. This master thesis aims to investigate and develop an effective machine learning model to predict flight delays for Los Angeles International Airport, thereby contributing to improved operational efficiency, enhanced passenger satisfaction, and reduced financial impacts for the airport.

## 3.1. Background and Motivation

Flight delays can be attributed to various factors, including weather conditions, air traffic control restrictions, mechanical issues, and crew availability. These delays not only inconvenience passengers but also result in increased operational costs for airlines, such as additional fuel consumption, crew expenses, and compensation payments. Furthermore, the knock-on effects of flight delays can propagate through the air traffic network, causing disruptions to flight schedules and resource allocation at airports.

Over the past few decades, researchers have explored several techniques to predict flight delays. Some of these approaches have shown promise in improving prediction accuracy and facilitating better decision-making for airlines and airports. However, there is still room for improvement. Machine learning has demonstrated exceptional performance in various applications, including image and speech recognition, natural language processing, and recommendation systems. The

potential of machine learning in flight delay prediction lies in its ability to handle complex, high-dimensional, and noisy data, which is typical of aviation operations.

Various machine learning techniques have been applied to the issue of flight delay prediction in recent years. This includes supervised techniques such as Decision Trees, Support Vector Machines, and Neural Networks, and unsupervised techniques such as clustering and dimensionality reduction. In addition, advanced techniques such as deep learning and reinforcement learning have also been tested to improve models' predictive accuracy and robustness.

### 3.2. Research Gap and Objectives

This paper aims to answer the research question: How can machine learning algorithms be utilized to predict flight delays at Los Angeles International Airport, and how can it be helpful in enhancing operational efficiency and passenger experience?

One of the unique aspects of this study is the focus on Los Angeles International Airport and the usage of data specific to the airport. We will attempt to build a flight delay prediction model tailored to LAX, which should yield higher accuracy than one that is more general. In previous studies, the scope of airports has mainly been on airports for an entire country like the United States. This makes sense if you wish to have a model that can be used to predict flight delays for multiple different airports. However, if you are only concerned with how flight delays impact one specific airport, it makes better sense to only consider data from that airport.

The paper will also contain a more comprehensive business review about the practical uses of predictive models than previous papers on flight delay prediction. This should make the paper increasingly relevant in a business context and for LAX stakeholders. In summary, this paper seeks to bridge the gap in flight delay prediction research by focusing on Los Angeles International Airport and employing the most relevant data available. By doing so, it aims to contribute to the existing knowledge base and provide practical insights that can help mitigate the detrimental effects of flight delays on passengers and the overall efficiency of airport operations.

# 4. Proposed Theory

In this section, the theoretical framework for the paper will be presented. This will include a presentation of the various machine learning algorithms that we will be using. These can often be seen as black boxes if they are not properly explained, so this section aims to give the reader an understanding of how these algorithms work along with an explanation of other vital concepts used in the paper.

## 4.1. Python

The majority of the model building will be done in Python. Python is a versatile and open-source coding language that is among the most popular worldwide. It is used in various domains such as web development, mobile application development, data analysis, and machine learning. One of the strengths of Python is the extensive list of advanced third-party libraries, which makes it easier to accomplish various tasks without having to code everything from the ground up. This project will make use of several different libraries in Python. Some will be used to a higher degree than others, but the most used ones will be pandas, NumPy, scikit-learn, and TensorFlow.

Pandas is a library commonly used for data cleaning and other data manipulation tasks. It allows you to use certain data structures that make the handling and processing of large datasets easier. Scikit-learn is a machine learning library that gives you access to various algorithms like logistic regression, k-nearest neighbors, and decision trees. TensorFlow is an open-source library created by Google. It is primarily used for deep learning tasks as it provides an efficient platform for large-scale computations. All of these libraries will be useful when building the models.

## 4.2. Machine Learning

Machine learning is the practice of training algorithms to learn and improve their own performance. It is about creating a system where the machine can learn from its previous experiences and use the knowledge to perform better in future tasks. This is often done by training algorithms on large datasets, where the algorithm identifies patterns and relationships within the data. Based on this, the algorithm can make predictions about new data. As the algorithm is utilized on more data, it will gain experience and should improve its performances over time.

There are generally two machine learning types: unsupervised and supervised learning. Unsupervised machine learning is where the algorithms are trained on unlabeled data. This means that the algorithm is not given the correct answer and will instead have to make sense of just the input data. This type of machine learning is commonly used for clustering or dimensionality reduction. Supervised machine learning is where the algorithms are trained on labeled data. The algorithm is given both the input data and the correct output, which means the task of the algorithm is to learn the relationship between the inputs and outputs. It is commonly used to make predictions or solve classification problems.

Machine learning is used for many different purposes across numerous industries. Some of the most well-known examples include speech recognition, fraud detection, forecasting, image classification, and medical diagnostics. For the airport industry, machine learning has the potential to improve various aspects of operations, which could lead to improved efficiency, safety, and passenger experience. Among the potential use cases are passenger flow optimization, queue management, maintenance forecasting, threat detection, and flight delay prediction.

### 4.3. Random Forest

Random Forest is a machine learning algorithm that can be used for classification and regression tasks. The algorithm builds a series of decision trees, which it can use to output either the mode or the mean of the values predicted by the individual trees. The mode of the values is considered for classification purposes, while the mean of the values is used for regression tasks. In our case, we will use the Random Forest algorithm for classification purposes.

The process behind the algorithm can be explained as follows:

1. Bootstrapping: Multiple subsets of the training dataset is created. This process is called bootstrapping. Each subset is a random sample of the dataset's observations.
2. Building Decision Trees: For each subset, a decision tree is built. Furthermore, only a random selection of features is considered during the construction of each tree. This randomness creates diversity among trees and should reduce the correlation between them.

3. Aggregating Predictions: When all of the decision trees have been built, the predictions are created by passing an observation through them. For classification tasks, the final prediction is determined by the option that the most decision trees have voted outputted. For regression tasks, the final prediction is the mean of all the decision tree predictions.

There are several advantages to using the Random Forest algorithm. It is well-equipped for datasets with lower data quality as it handles missing data and outliers well. Furthermore, it can directly show the different features' importance, heightening the interpretability and easing analysis. The disadvantages of the Random Forest algorithm would include limited effectiveness on high-dimensional data, propensity to bias on imbalanced datasets, and memory consumption.

### 4.4. XGBoost

XGBoost, or eXtreme Gradient Boosting, is an algorithm that has become popular among machine learning practitioners. It is a further development of the traditional gradient boosting methods and should therefore be able to provide greater accuracy, scalability, and efficiency. The algorithm uses gradient boosting to create a series of decision trees, which progressively gets more accurate as each decision tree attempts to eliminate the mistakes made by the previous tree. The process of a XGBoost algorithm can be explained as follows:

1. An initial prediction is made. This is the base for the decision trees, and it is usually the median or mean of the target variable.
2. The difference between the initial prediction and the target variable is calculated. This is also referred to as the residual.
3. A decision tree is fitted to the residuals to predict the errors.
4. The decision tree is added to the series of trees with a weight determined by the learning rate and the performance of the tree.
5. The weighted output of the decision tree is combined with previous predictions to make updated predictions for each data point.
6. Steps 2-5 are repeated until a stopping criterion is met or a pre-defined number of iterations have been run.

In addition to the described process, XGBoost has various built-in components that contribute to its effectiveness. For example, XGBoost uses a column block technique, which improves computational efficiency. Its support of distributed computing and parallel tree construction also allows it to train faster and on larger datasets. Lastly, regularization techniques are built into the algorithm to prevent overfitting.

### 4.5. Neural Networks

Neural Networks are a subset of machine learning. It can be used both supervised and unsupervised, but the most common applications are in supervised learning settings. The general concept of Neural Networks is to replicate the function of the human brain, which takes inputs from our senses, transmits it through the network of neurons, and then outputs an impression, interpretation, or decision.

The structure of a Neural Network usually consists of three different layers: input, output, and hidden layers. The input layer receives the data, and the output layer generates the final prediction or decision. Between the input and output layers are the hidden layers, which contain the artificial neurons that perform computations on the data to learn the underlying patterns. The neurons are interconnected, and the weights and biases determine the strength and influence of each connection.

Weights and biases are the parameters that Neural Networks learn during training. All connections between neurons have an associated weight, which influences how strongly one neuron's output will affect another neuron's activation. A positive weight will increase the strength of the input signal, while a negative weight will decrease the strength of the input signal. Biases are connected to individual neurons, and they control the ease with which a neuron is activated. The bias term is added to the weighted inputs before going through the activation function. This can create a shift for the neuron’s output, which allows the network to learn more complex patterns. The weights and biases are essentially used to regulate the output of a neuron.

During training, the weights and biases will constantly shift to find the optimal value to reduce the error between the network's predictions and the target outputs. This process will continue until a satisfactory accuracy level has been reached, a pre-defined stopping criterion has been met, or all of the training data has been processed. A learning algorithm and a loss function are used to guide the optimization process. A common learning algorithm could be gradient descent, while some standard loss functions are mean squared error and cross-entropy loss. Choosing the most appropriate learning algorithm and loss function for a task depends on several factors such as the type of problem, the nature of the data, and constraints.

Within the topic of neural networks exist different types such as recurrent neural networks (RNN), convolutional neural networks (CNN), and long short-term memory (LSTM). Again, the task at hand will determine the most appropriate type of neural network. For example, CNNs are often used for image processing tasks, while LSTMs are frequently used for sequential data tasks.

### 4.6. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is an oversampling technique that is used to balance class distribution. The balancing happens by generating synthetic examples of the minority classes. The technique is usually used for classification tasks in machine learning, where the dataset has a higher number of instances of one class. This imbalance might lead to a loss of accuracy, so it can be advantageous to even them out. This can happen by either using an even subset of the data or by generating more examples of the minority class; SMOTE helps with the latter option.

The generation of synthetic samples is done using a combination of nearest neighbor and interpolation. The oversampling ratio, which is the ratio of the minority class to the majority class, is first determined. It is usually set to create a completely balanced dataset, but it is also possible to give a desired ratio. The next thing is to choose the number of nearest neighbors. This parameter determines how many nearest neighbors are considered when generating the synthetic samples. A low value will create similar samples, while a higher number will lead to more diverse samples. Lastly, the kind of interpolation used to create the samples can vary. The most common option is linear interpolation, but certain datasets might fit other interpolation types better.

### 4.7. Evaluation Metrics

There are several different metrics that can be used to evaluate the performance of machine learning models. The most common include accuracy, precision, recall, and F1 score.

- Accuracy is the percentage of correctly classified predictions out of the total number of predictions. It is one of the most common metrics for classification problems as it is easy to calculate and understand. It can be misleading if the dataset is unbalanced.
- Precision is also known as the positive predictive value as it measures the percentage of true positive predictions out of the total number of positive predictions. It can be an especially useful metric if the costs of false positives are high.
- Recall measures the percentage of true positive predictions from the total number of positive observations. This can be useful when the costs of false negatives are high.
- F1 score is the harmonic mean of the precision and recall metrics. It can be used when you want a balance between the two or you are using an imbalanced dataset.

### 4.8. Model Sketch

The goal of the model is to correctly predict when a flight will depart. We have created four buckets, which indicate the level of delay a flight departure is going to experience: on-time, small delay, medium delay, and large delay. To build our model, we will follow a 7-step process:

1. Data Retrieval: The first step involves collecting the necessary data required for building the model. Data is the building stones of a machine learning model, so it is important to find data that is of high quality and with the right quantity. There are different sources for airport data, and we will be examining the different options and selecting the data that is deemed the most appropriate for our project.
2. Exploratory Data Analysis (EDA): This step is all about becoming familiar with the data. By performing an EDA, we can better understand the data's strengths and weaknesses, identify patterns, and uncover potential issues that may affect the model's performance. This section will also contain a significant amount of data visualization to represent the data graphically.

3. Data Cleaning: This step involves cleaning and transforming the raw data to make it suitable for use in machine learning algorithms. Data comes in all different shapes and forms, so it is essential to standardize it to ensure the algorithms process it correctly. Furthermore, it might be necessary to remove certain observations if the data is not rich enough. Typical tasks in the data cleaning step include handling missing or inconsistent values, removing duplicates, and correcting errors.
4. Feature Engineering: In this phase, we create new features or modify existing ones to enhance the performance of the machine learning algorithms. This step is similar to data cleaning, but it focuses on the data's features, also called columns. A dataset might contain hundreds of features for each row, which is great as it contains a lot of useful information. However, processing this data and using it in machine learning algorithms becomes very computationally heavy. Using dimensionality reduction is a way to combat this. It can speed up the training process of machine learning models and helps with making the data more interpretable. However, it does come with the risk of information loss, which can impact the performance of the model negatively. Other techniques in this stage are One-Hot Encoding and scaling.
5. Model Selection: This step entails exploring various machine learning algorithms and choosing the most appropriate for the given data, based on factors like complexity, interpretability, and training time. Understanding the problem at hand is an essential part of this step as the problem dictates which model will be the most useful for the situation.
6. Model Training: Once the suitable algorithms are selected, we train the models on our prepared data. This can involve adjusting the model's hyperparameters to optimize performance and minimize the error between predictions and actual outcomes. Some popular ways to optimize parameters include grid search and Bayesian optimization.
7. Model Evaluation: Finally, we evaluate the performance of the trained models by comparing them based on metrics such as accuracy, precision, recall, and F1 score. This assessment helps us determine the best model to deploy and provides insights into potential improvements for future iterations.

### 4.9. Hypotheses

We have formed three hypotheses about the results of our model. These can be used to test whether the model is giving us results as expected, or whether the results can be used to reject the hypotheses. Our hypotheses are:

1. **Weather will be the most crucial feature affecting the predictability of flight delays.**

Weather is considered an essential factor in a flight's delay. This makes sense as there can be certain restrictions on takeoffs and landings based on the local weather. It is not uncommon to see airports shutting down during snowstorms or blizzards. Similarly, strong winds can determine which runways to use, and which aircrafts are able to take off. For LAX, the wind is more likely to be an influencing factor than temperature as California is not particularly plagued by extreme temperature changes. However, it can be very windy in the area around LAX.

2. **No significant correlation will be found between the operating airline and the predictability of delays.**

We are not expecting a significant correlation between the airline of the flight and the predictability of delays. No airline should be given priority over others, nor should there be fewer maintenance requirements for specific airlines. This should result in an even playing field, where the specific airline does not influence the likelihood of delays. The only real difference between the airlines should stem from their selection of routes, where certain routes might be more prone to delays as the airspace could be more crowded or certain runways have to be used.

3. **The Neural Network will be the best performing model.**

The third hypothesis is that the Neural Network model will be the most accurate one of our models. The Neural Network algorithm is generally considered the most sophisticated of our three algorithms, and the study from (Alla, Moumoun, & Balouki, 2021) showed great results using Neural Networks to predict flight delays. While past performance is no guarantee of future results, it is noteworthy that the study used data from the same source as us. Therefore, we expect the model to perform better than the Random Forest and XGBoost models.

# 5. Methods and Data

This section will outline and describe the process of building our model. The structure is going to follow the process outlined in section 4.8.

## 5.1. Data Retrieval

We have chosen to work with a dataset from the US Bureau of Transportation. It contains all flights from US airports between January 2018 and May 2022 and has accompanying information about the airline, origin, destination, taxi time, distance, and delay time. The total number of columns is 61, and there are 29,193,782 rows. In Table 1, you can see descriptions for a selected group of the dataset's columns. A full list of column descriptions can be found in Appendix A.

| COLUMN | DESCRIPTION |
|---|---|
| FLIGHTDATE | The data of the flight. It is represented in the (yyyy/mm/dd) format. |
| MARKETING_AIRLINE_NETWORK | The airline of the flight. |
| TAIL_NUMBER | The tail number of the aircraft. |
| FLIGHT_NUMBER_OPERATING_AIRLINE | The flight number of the aircraft. |
| ORIGIN | The airport that the flight is departing from. |
| DEST | The airport that the flight is arriving to. |
| TAXIOUT | The number of minutes the aircraft moved on the ground at the origin airport before taking off. |
| TAXIIN | The number of minutes the aircraft moved on the ground at the destination airport before parking at the gate. |
| DEPDELAYMINUTES | The difference in minutes between departure time and scheduled departure time. |
| ARRDELAYMINUTES | The difference in minutes between arrival time and scheduled arrival time. |
| CANCELLED | Indicates whether the flight was cancelled. 1 represents cancelled. |

| | |
|---|---|
| DISTANCE | Distance in miles between the origin airport and the destination airport. |
| CRSDEPTIME | The scheduled departure time of the flight. |
| WHEELSOFF | The time when the flight has lifted off from the departure airport. |
| AIRTIME | The flight time in minutes. |

*Table 1: Description of selected columns from Bureau of Transportation dataset.*

In the search for a dataset, we mainly looked for high data quality, high data quantity, and a high number of features. The dataset from the Bureau of Transportation filled all of these criteria, so it was perfect for the project. Since we focus on LAX, we filtered the initial dataset to only contain flights that departed from LAX. This brought the total number of observations down to 833,491.

The dataset from the Bureau of Transportation did not contain information about the weather conditions at the departing airport. As one of our hypotheses is that weather plays a significant role in the likelihood of delays, we had to search for a weather dataset to merge with the flight dataset. We found this from Virtual Crossing, which is a service that provides weather data and forecasts. The dataset shows weather-related information in the area around LAX such as temperature and wind conditions for all days between January 2018 and May 2022. A description of the dataset's columns can be seen in Table 2.

| **COLUMN** | **DESCRIPTION** |
|---|---|
| NAME | The location of the weather observation. |
| DATETIME | The date of the observation. |
| TEMPMAX | The maximum temperature on that day. |
| TEMPMIN | The minimum temperature on that day. |
| TEMP | The mean temperature of the day. |
| FEELSLIKEMAX | How the maximum temperature of the day feels. |
| FEELSLIKEMIN | How the minimum temperature of the day feels. |
| FEELSLIKE | How the mean temperature of the day feels. |

| | |
|---|---|
| DEW | The dew points. |
| HUMIDITY | The humidity at the location. |
| PRECIP | The amount of rain (measured in mm). |
| PRECIPPROB | The probability of rain. |
| PRECIPCOVER | The percentage of rain cover. |
| PRECIPTYPE | Rain type. |
| SNOW | Whether it is snowing at the time of the observation. |
| SNOWDEPTH | The snow depth if it was snowing. |
| WINDGUST | The wind gust (measured in kph). |
| WINDSPEED | The wind speed (measured in kph). |
| SEALEVELPRESSURE | The sea level pressure (measured in mb). |
| CLOUDCOVER | The percentage of cloud cover on the day. |
| VISIBILITY | How far can you see during the day (measured in km)? |
| SOLARRADIATION | The radiation of the sun (measured in $W/m^2$). |
| SOLARENERGY | The amount of solar energy (measured in $MJ/m^2$). |
| UVINDEX | The UV indexes. |
| SEVEREVERISK | If there are severe weather risks. |
| SUNRISE | The sunrise time. |
| SUNSET | The sunset time. |
| MOONPHASE | Which phase the moon is in. |
| CONDITIONS | A short description about the weather. |
| DESCRIPTION | A description of the weather conditions of the day. |
| ICON | A weather icon. |
| STATIONS | The ID of the station that made the weather observation. |

*Table 2: Description of columns from Virtual Crossing weather dataset.*

## 5.2. Exploratory Data Analysis

The goal of the EDA is to become familiar with the data and examine how various features are related. It is also an important step in understanding which features to select for the model training. The first thing that we want to have a look at is the distribution in terms of minutes delayed. This should provide us with information regarding how frequently a flight is delayed a certain number of minutes. It was initially difficult to read the graph as the scaling was off due to there being significantly more flights with 0 minutes of delay than with delay. Consequently, we had to look at flights with at least a one-minute delay for the graph to be readable. We also have to apply an upper limit of 60 minutes as there are only a few flights with larger delays than an hour. After adjusting for this, we can see that the frequency drops the longer the delay.

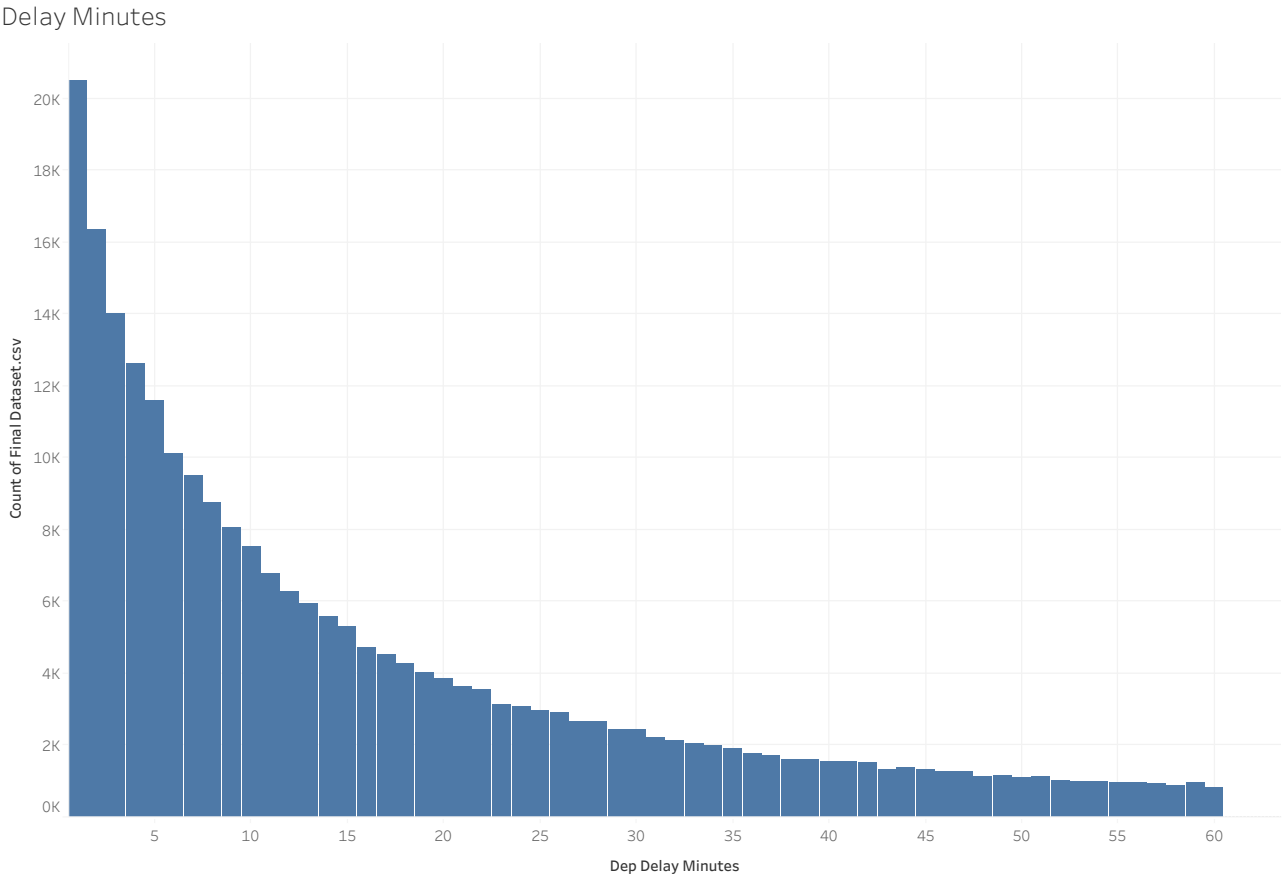


*Figure 1: Bar plot showing the delay frequency at minutes 1-60.*

To better analyze the distribution of delays, we create buckets for different delay lengths. There are four buckets: on-time, small delay, medium delay, and large delay. A flight can only qualify for on-time if the delay is zero minutes, while a small delay is a flight that is up to 15 minutes late. A medium delay is 15 to 45 minutes, and a large delay is everything above 45 minutes. To show the distribution, we create a bar chart with the total number of flights in each bucket.

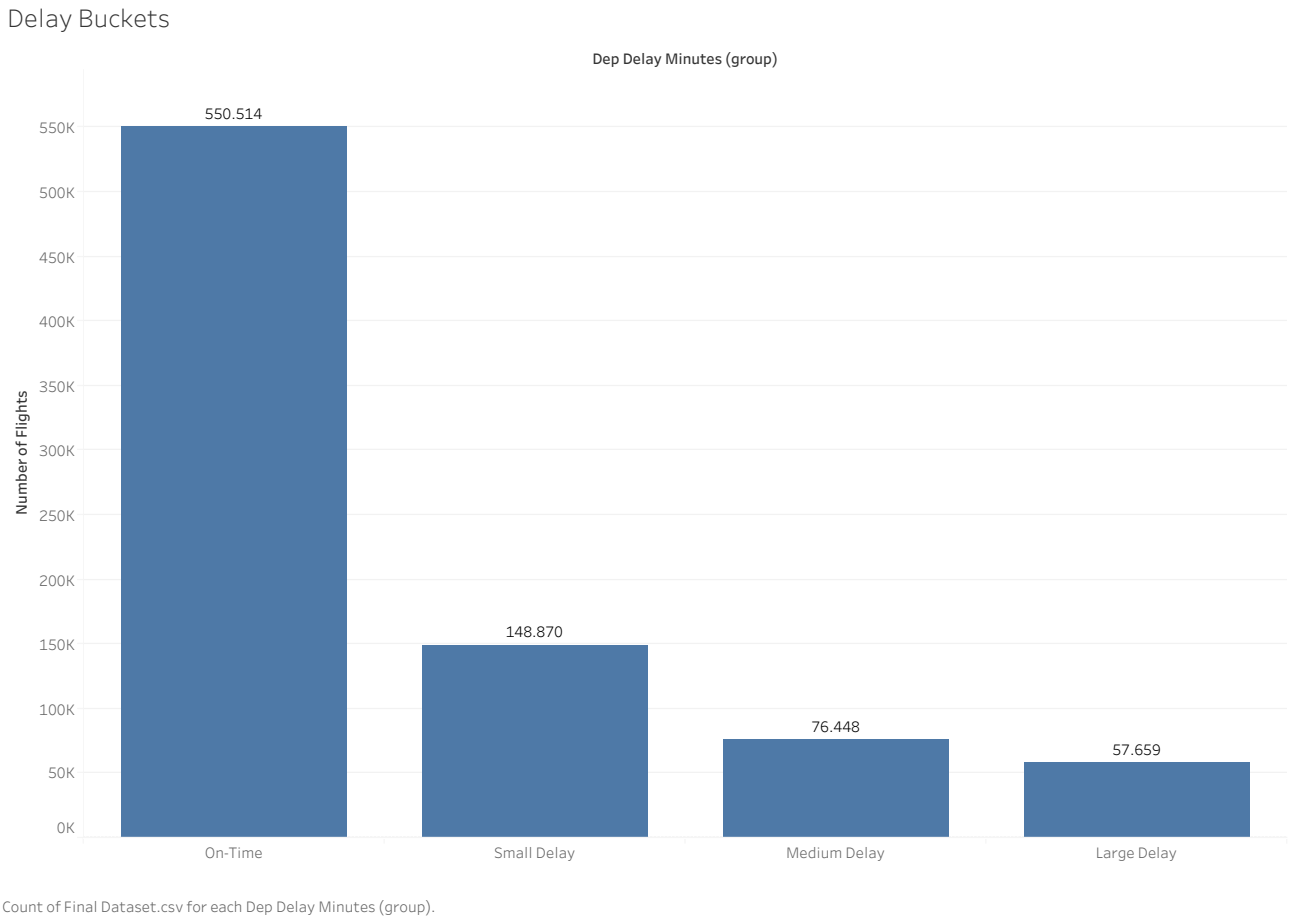


*Figure 2: Bar plot showing the total number of flights in the different delay buckets.*

The buckets confirm our initial assessment that the majority of flights depart on time. The on-time bucket is approximately three and a half times larger than the bucket for small delays, which is the second biggest bucket. Also, we can again see that shorter delays are more common than longer delays. Interestingly, there does not appear to be much difference in the frequency between medium and large delays though as the buckets are almost of equal size. Since these will also be our output classes, it is clear that the dataset is unbalanced. This can create issues for some algorithms

as it can introduce bias. There are multiple ways to deal with unbalanced datasets such as under-sampling, over-sampling, and SMOTE.

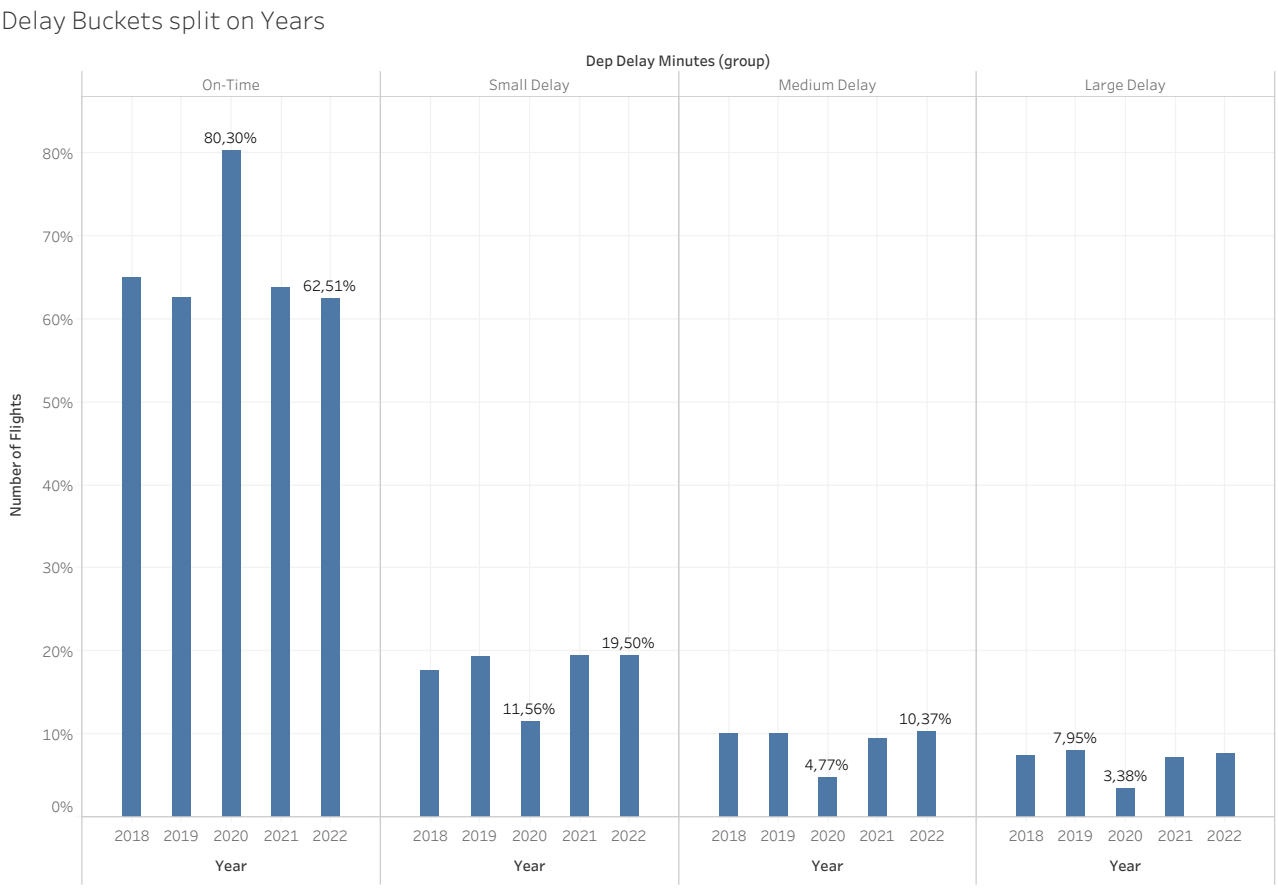


*Figure 3: Plot showing the percentage share of flights in each delay bucket for every year of the dataset.*

Figure 3 shows the bucket distribution of delayed flights across the dataset's years. In all of the dataset's years, the bucket with flights on-time is the largest. The highest percentage of flights on-time compared to total flights was in 2020 with 80.30%. This could be explained by the fact that many places had severe travel restrictions in place during 2020, which made traffic less of an issue for airports. Conversely, 2019 had the highest percentage of large delays with 7.95%. This was also the year that LAX set its annual passenger record, so it would make sense that resources and capabilities were stretched more than usual during this time.

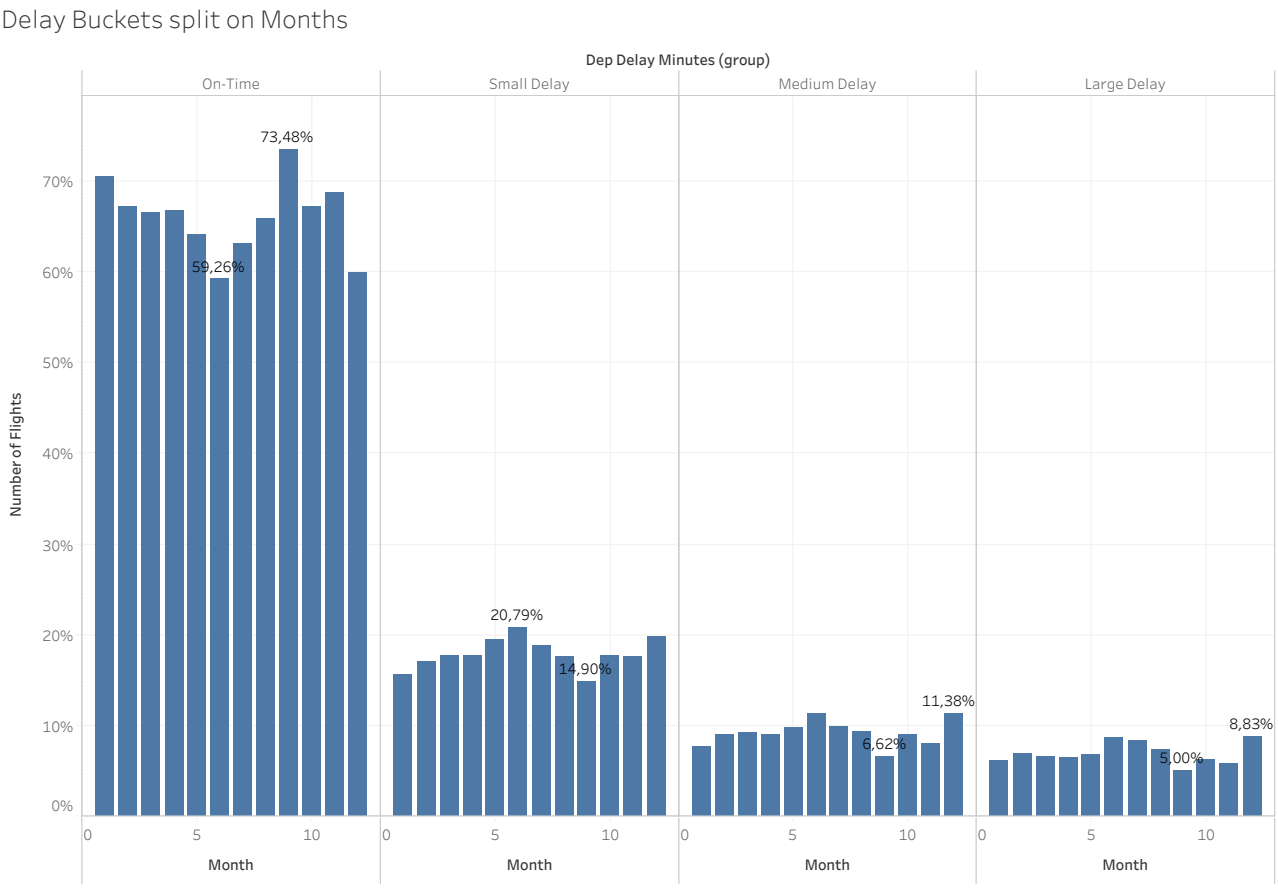


*Figure 4: Plot showing the percentage share of flights in each delay bucket for every month of the year.*

When we look at the delay distribution across the different months of the year, we can see that many of the summer months have a high percentage of delayed flights. Specifically, the month of June has the lowest percentage of on-time flights. Similarly, July and August have relatively low percentages of on-time flights with 63.08% and 65.87% respectively. It is unsurprising that the summer months have a higher percentage of delayed flights as travel demand is usually higher in these months, and thus the strain on the airport's resources will be greater. It is also noteworthy that December has the highest percentage of medium and large delays. This can possibly be explained by LAX having a lot of arrivals during this month as California is a typical holiday destination for people living in colder climates. At the other end of the spectrum, September appears to be the best month for flights that depart on time. 73.48% of flights depart on time in September, and it is simultaneously also the month with the lowest percentage of small, medium, and large departure delays.

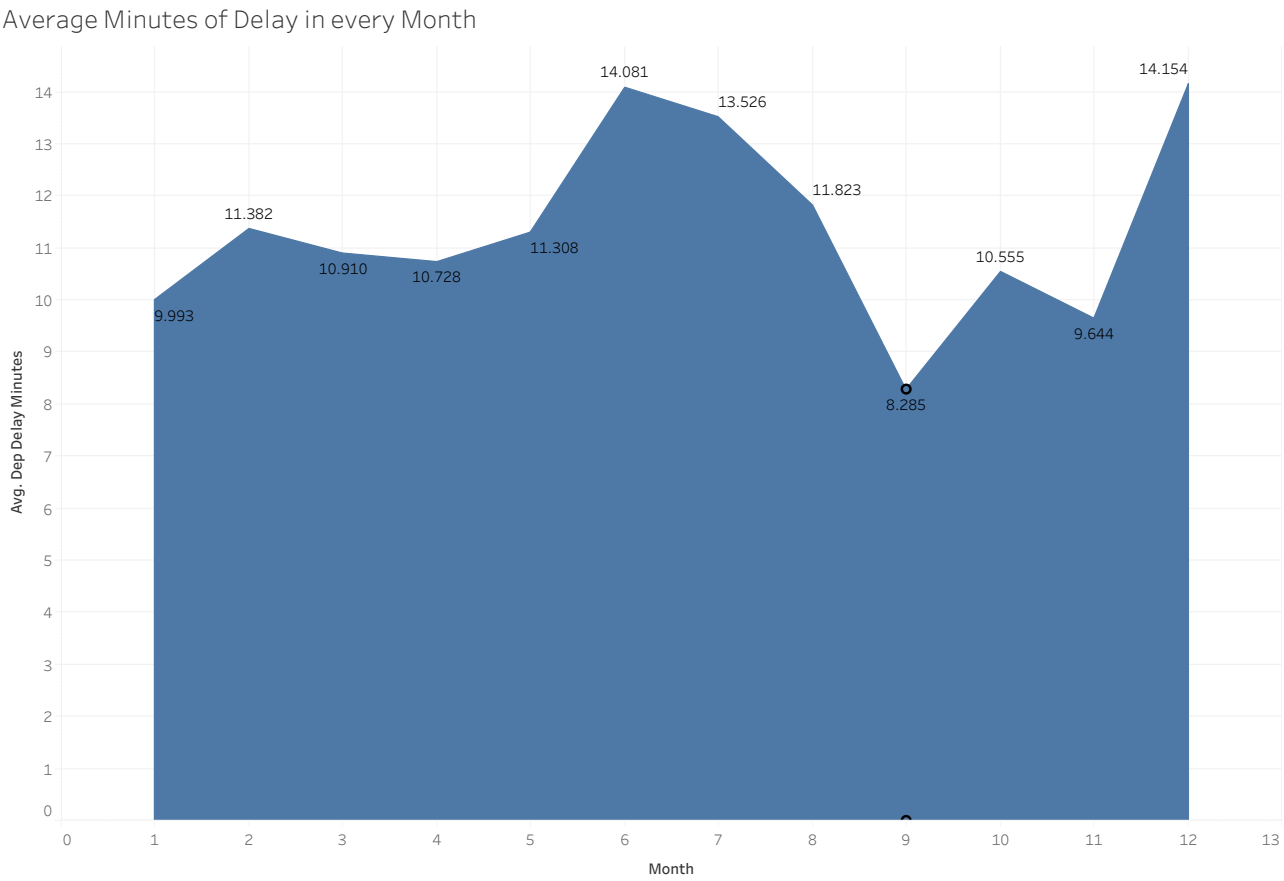


*Figure 5: Area chart showing the average minutes of delay in each month of the year.*

Looking at the average minutes of delay in every month, we can see a spike in the summer months and in December. This makes sense as travel demand is generally greater in these months than in others. It should be especially prevalent at LAX, where there are also many arrivals due to tourism in these months. On the other side, the autumn months seem to have shorter delays than other months, and September is the month least affected by delays with only an average delay time of around eight minutes.

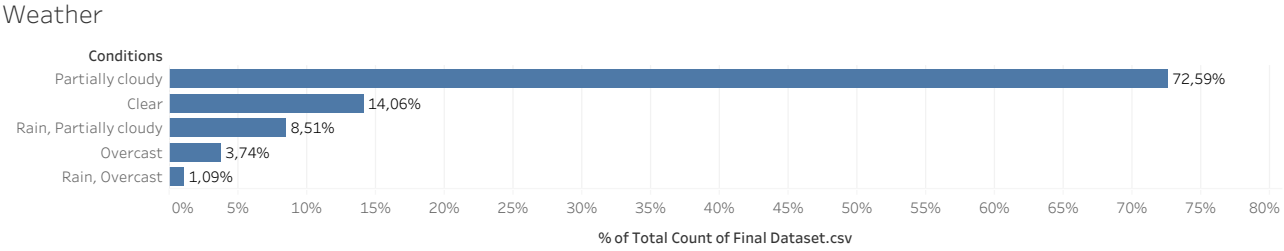


*Figure 6: Chart showing the different weather conditions and the share of flights in each category.*

Figure 6 shows the different weather conditions and how many flights departed under these conditions. We can see that 72.59% of departures had partially cloudy weather, while only 1.09%

took off with rain and overcast weather. This goes along with our expectations as southern California is not a place that is known for having a lot of rain fall.

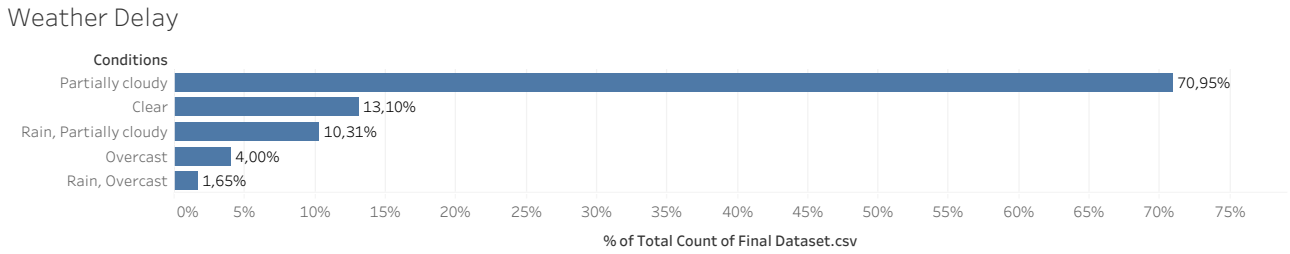


*Figure 7: Chart showing the different weather conditions and the share of flights in each company. This is filtered to only contain flights with delays.*

Surprisingly, the distribution does not change a lot when only filtering for flights with a delay. The percentage of departures with overcast and rain jumps 0.56 percentage points, while the percentage of partially cloudy departures fall 1.64 percentage points. The biggest difference is in the 'rain, partially cloudy' category with a percentage point increase of 1.8. This could indicate that weather is not as significant a factor in flight delays as previously assumed.

## 5.3. Data Cleaning

Our main dataset comes from an official government authority, so the data quality is higher than what would normally be expected of this large a dataset. Similarly, the weather dataset comes from a provider whose main business objective is to sell weather data. Therefore, the quality of that dataset is also very high. Due to these reasons, a significant amount of data cleaning is not needed. However, checking for duplicate values, missing values, and outliers is always good before deploying the data onto the models.

We first check for duplicate values. Duplicate observations can happen during the actual collection of data, or it might be the result of an error during scraping or concatenating. Nonetheless, it is preferable to remove one of the duplicate values from the dataset. We did not find duplicate values in our dataset, so the process can move along to check for missing values. If some rows have a lot of missing values, removing them might be advantageous as these can negatively impact the accuracy of the model. In our case, we found a few rows with missing values in different columns. As this number was relatively minuscule and our dataset was sufficiently big, we decided to drop these rows from the dataset.

During the EDA, we also checked for outliers. Outliers can have a negative effect on the accuracy of the model as their large values might skew the model in one direction. However, if they are not the result of data entry mistakes, they are a part of the real world, and the models might be missing out on valuable information if these are removed before training. In our case, we did not have any significant outliers and the data points that could have seemed out-of-the-ordinary were most likely not the result of mistakes. Therefore, we decided to keep them in the dataset for our models.

### 5.4. Feature Engineering and Selection

The first thing that we do in the feature engineering section is to bring our two datasets together. As previously mentioned, we did not have information about weather conditions for each observation in the flight dataset. Therefore, we need to create new features using the weather dataset. To do this, we do a simple mapping of matching the flight dataset's date with the weather dataset's date. This will give us the temperature and wind conditions at the day of departure for each row in our flight dataset. With these new features, we are able to include weather conditions in the training of our models and we will be able to see if it affects flight delays.

The next step is selecting the features that we wish to include in our model. This process can sometimes be more of an art than a science, but we will use domain knowledge and data exploration to pick our chosen features. In the end, we decided to look at airline, destination, departure time, taxi time, wheels-up time, airtime, distance, day of the week, weather, wind speed, and temperature. We also did test runs with other features, but the selected features appeared to be the best ones for accuracy.

One of the last steps is applying One-Hot Encoding to our categorical features. Most machine learning algorithms can only work with numerical values, so there is a need to transform all the categorical values to numerical values. This can be done using the One-Hot Encoding function. It takes every categorical value and creates a new binary feature. For example, if there is a categorical value called ‘American Airlines’ in a column, it will create a new column called ‘American Airlines’ and set it to 1 if the row had that categorical value or set it to 0 if it did not. Lastly, we split our

dataset into training and test sets. We use a 70/30 split, meaning 70% of the data goes to training, and the remaining 30% is used for testing.

## 5.5. Model Selection

Based on our literature review, there is no significant correlation between how advanced an algorithm is and its accuracy in predicting flight delays. Therefore, we have decided to build three models with three different algorithms and compare their results. The first uses a Random Forest Classifier, and it is arguably the least advanced of our three used algorithms. The accuracy of this model should be a good indicator of the need to use more advanced models. If it yields scores close to the other models, it might not be worth the time and effort to train more advanced models. The second model uses a XGBoost algorithm. It is more advanced than the Random Forest algorithm, so it should theoretically perform better on our data. However, there is always the risk that it might not fit our chosen dataset. The last model is a Neural Network model. It is arguably the most sophisticated algorithm of the three models, so it should give us the best results if there is a positive correlation between accuracy and how advanced the algorithm is. This rationale was also the basis for our hypothesis that the Neural Network model would be our best-performing model.

The Random Forest and XGBoost algorithms can be used directly from the Python packages, while we have to configure the Neural Network ourselves. We tried various Neural Network structures to optimize performance, and we saw the best performance with a model that had three fully connected layers. The first layer has 64 neurons and will take the input data and put it through a ReLu activation function. This function helps the model learn complex patterns by introducing non-linearity. The second layer has 32 neurons and also uses the ReLu activation function. In the last layer, there are four neurons, each representing one of our output classes, and it uses a softmax activation function. This function is commonly used for multi-class classification tasks like ours. It converts the outputs into probabilities, which makes it easier to interpret the results. The loss function is a categorical cross-entropy loss function, and we will use the Adam optimizer. We train the model on 20 epochs with a batch size of 32.

## 5.6. Model Training

We are going to train multiple versions of our different models. The first versions will be with no adjustments to the parameters and dataset. This should give us a general picture of how the models perform on the data. After we have done the training with no adjustments, we will attempt to improve the performance of the models by tuning parameters and adjusting the dataset further. It is our expectation that the models are going to perform better after the adjustments.

For the adjustments, the first thing that we are going to do is balance the dataset. As we saw in the EDA, significantly more flights arrived on time than flights with small, medium, or large delays. This can be a problem for some algorithms as there are simply more data to process for certain outputs. This can lead to bias and a loss of accuracy. Therefore, we balance the dataset using a SMOTE object. Next, we will use k-fold cross-validation to ensure the model is not overfitted to the training subset of the dataset. This method takes the dataset and divides it into a set number of equally sized folds. The models are then trained and tested on each fold. The average accuracy of all iterations is lastly given as the final performance estimate.

## 5.7. Model Evaluation

For the evaluation of our models, we will mostly be focusing on the accuracy score. There is a discussion to be had about whether it is preferred to have false positives or false negatives. In an ideal scenario, neither would be especially prevalent in a model's performance. However, it is important to determine which scenario is preferable as it can impact where to set the output thresholds and help mitigate the risks of potentially harmful predictions. For a flight delay prediction model, it can be argued that false positives are preferable as it might be better to err on the side of caution. With false positives, the model is wrongly predicting flights to be delayed. This might be preferred because passengers could have connecting flights or other time-sensitive plans upon their arrival, which could be jeopardized if the model falsely predicted the flight to be on-time. If it turns out that the flight is not going to be delayed, it is merely a pleasant surprise as opposed to something that can ruin travel plans. However, as neither of the two options are especially catastrophic, it makes the most sense to focus on accuracy.

In the first version of the models, both the Neural Network and XGBoost models performed well. These had accuracy scores of 95% and 89% respectively, which are impressive considering that limited tuning had been done. Both models excelled at predicting on-time flights and large delays, while they struggled with small delays. The Random Forest model performed significantly worse than the other two with an accuracy score of just 70%. It performed best on flights that departed on time, while both small and medium delays were a struggle. However, the limited accuracy and poor performance relative to the other models were not particularly surprising given that Neural Network and XGBoost are more sophisticated algorithms.

All models had accuracy scores above 89% after the dataset was adjusted. Curiously, the Neural Network model actually saw a drop in performance after balancing the dataset. This can be explained by the fact that Neural Networks usually require large datasets to be efficient. After we balanced the dataset, the volume of data was less than before and therefore the algorithm could not learn as much as previously. Both the Random Forest and XGBoost models improved upon the dataset adjustments, but the Random Forest model saw the biggest improvement. It jumped 20 percentage points in accuracy after the adjustments. The XGBoost model only improved by three percentage points, but it was actually the most accurate of the three models on the balanced dataset.

| MODEL | ACCURACY | ACCURACY (AFTER ADJUSTMENTS) |
|---|---|---|
| **RANDOM FOREST** | 0.70 | 0.90 |
| **XGBOOST** | 0.89 | 0.92 |
| **NEURAL NETWORK** | 0.95 | 0.90 |

*Table 3: Accuracy scores of the three models, both before and after adjustments.*

# 6. Analysis and Results

## 6.1. Analysis Preliminaries

We will look at each model's feature importance charts in this section. These are a built-in function to both the Random Forest and XGBoost algorithms, but it is unfortunately not available for the

Neural Network algorithm. The reason is because Neural Networks involve complex interactions between features, which makes it difficult to isolate the effect of a single feature on the output. Furthermore, the importance of a feature is not tied to a specific weight. It can influence the output through a myriad of different paths, which can enhance or counteract each other. There are different methods, which can be used to estimate the feature importance in Neural Networks. However, these methods have limitations, and the results are not always trustworthy.

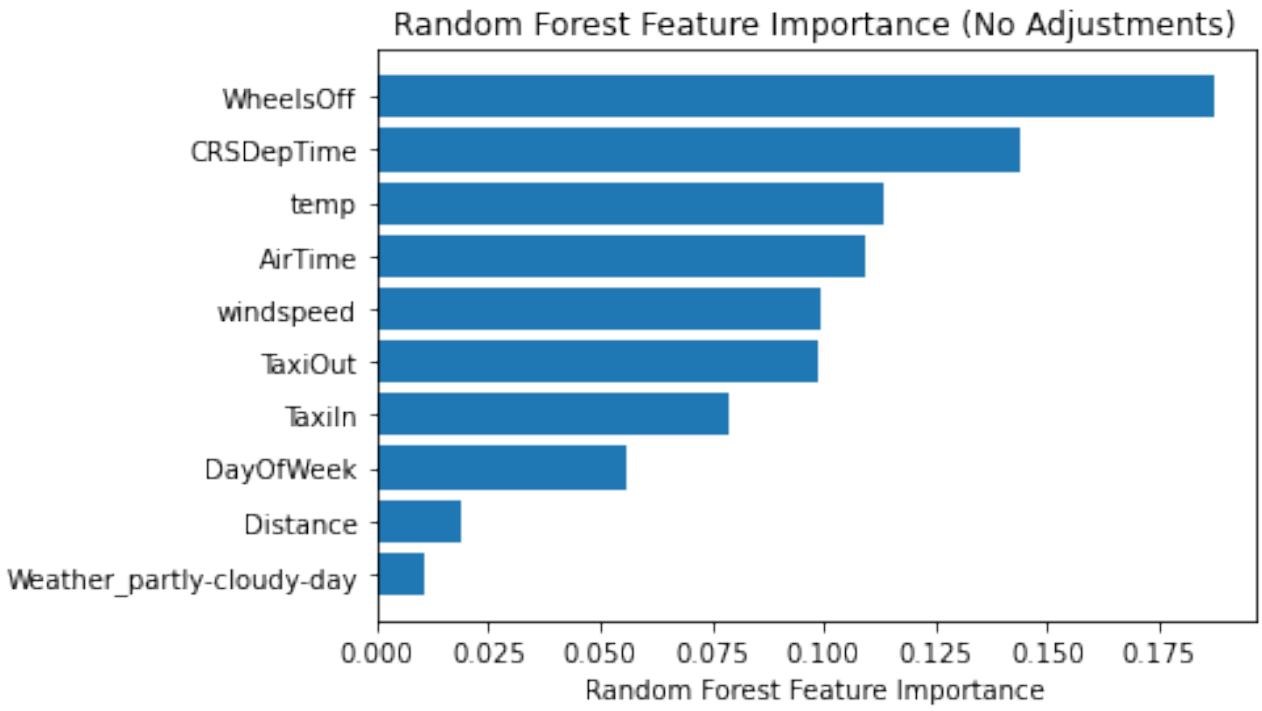


*Figure 8: Feature importance chart for Random Forest model with no adjustments.*

If we look at the feature importance for our Random Forest model with no adjustments, we can see that WheelsOff, CRSDepTime, and temp are the three most important features for the model. The WheelsOff feature signals the time at 'Wheels Up', which is an aviation term for when the aircraft lifts off from the origin airport. Similarly, the CRSDepTime is the scheduled departure time of the flight. This shows that the departure time is a highly important factor for the predictability of delays. This does make sense from a logical perspective as there are going to be times that are busier than others. If a flight is scheduled for take-off in a busy window, the unpredictability will most likely be higher because the airport's resources will be stretched.

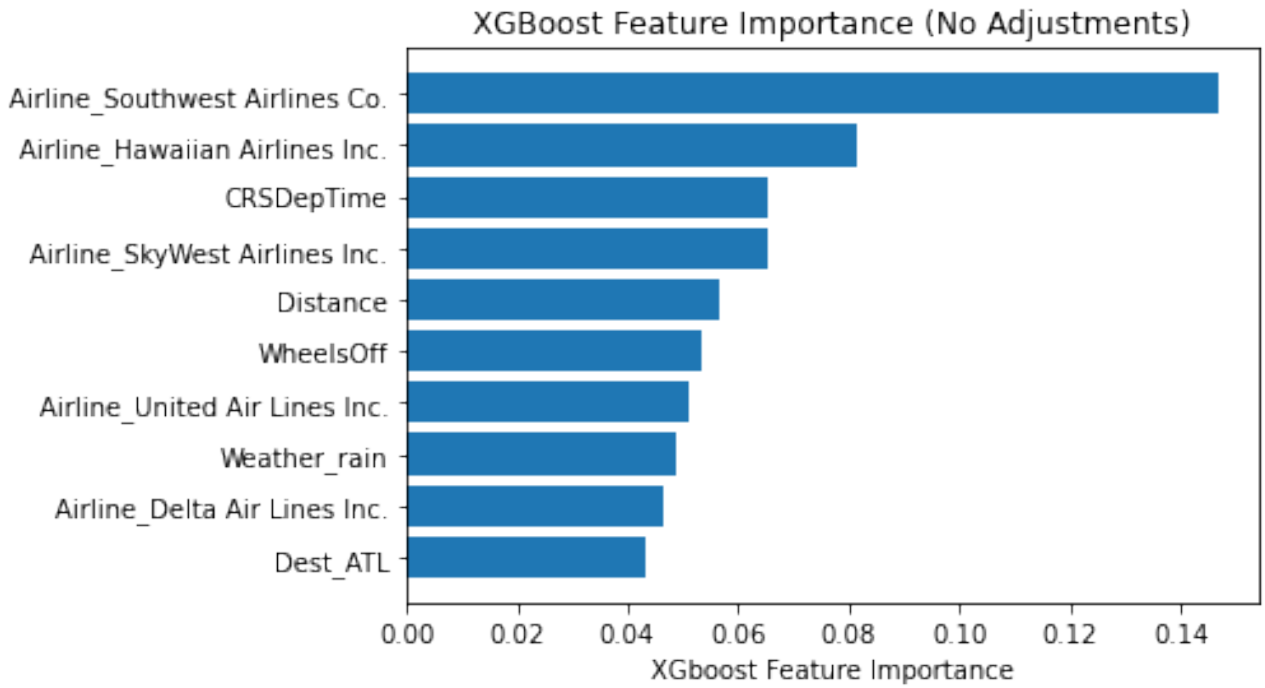


*Figure 9: Feature importance chart for XGBoost model with no adjustments.*

For the XGBoost model without adjustments, we can see that the two most important features are One-Hot Encoded features created from the 'Airline' feature. In addition, the fourth, seventh, and eighth most important features are also airline features. This suggests that certain airlines are more predictable than others in terms of delay time. CRSDepTime and WheelsOff are also important features for this model, but they are only the third and sixth most important features. Their presence does however confirm the suspicion that departure time is an important indicator.

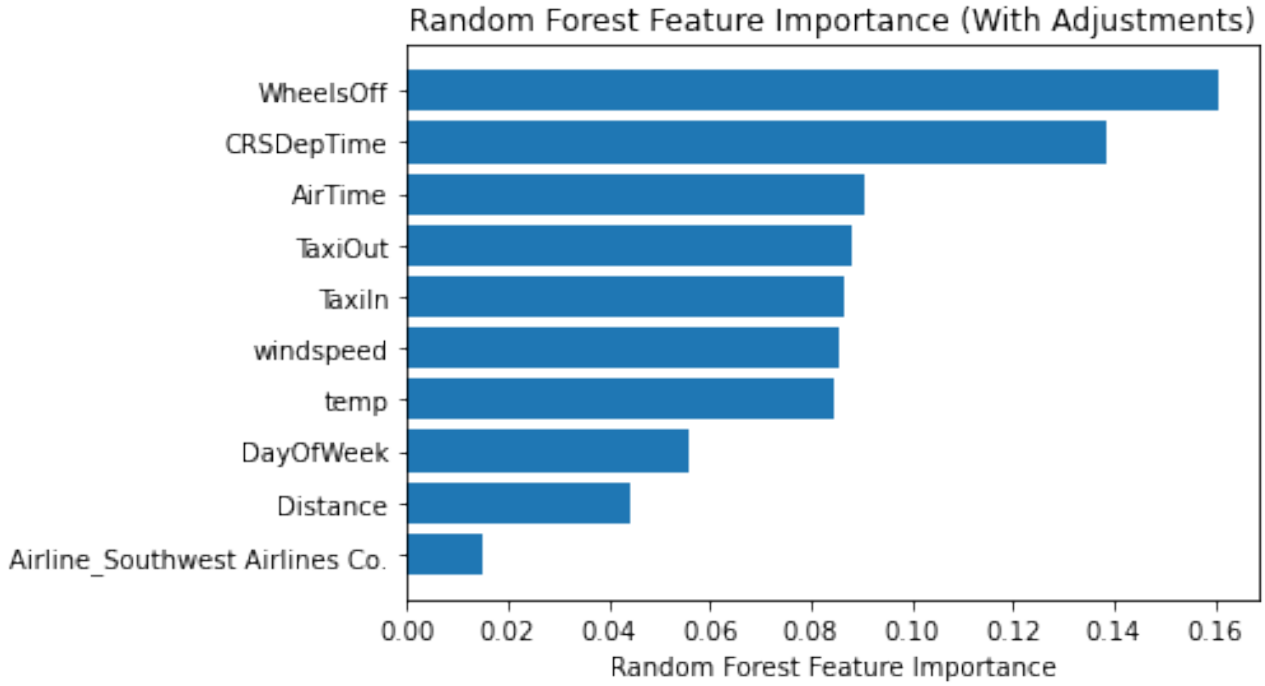


*Figure 10: Feature importance chart for Random Forest model with adjustments.*

The feature importance graph for the Random Forest model with adjustments looks similar to the one without adjustments. Most of the top features are the same as both WheelsOff and CRSDepTime are the two most important features. Temperature is no longer the third most important feature as Airtime has overtaken it. However, the feature weight difference is very small between the third and seventh most important feature. It is also noteworthy that the importance of each feature is higher than in the model without adjustments. This could indicate that the initial model had trouble filtering out the noise of the dataset.

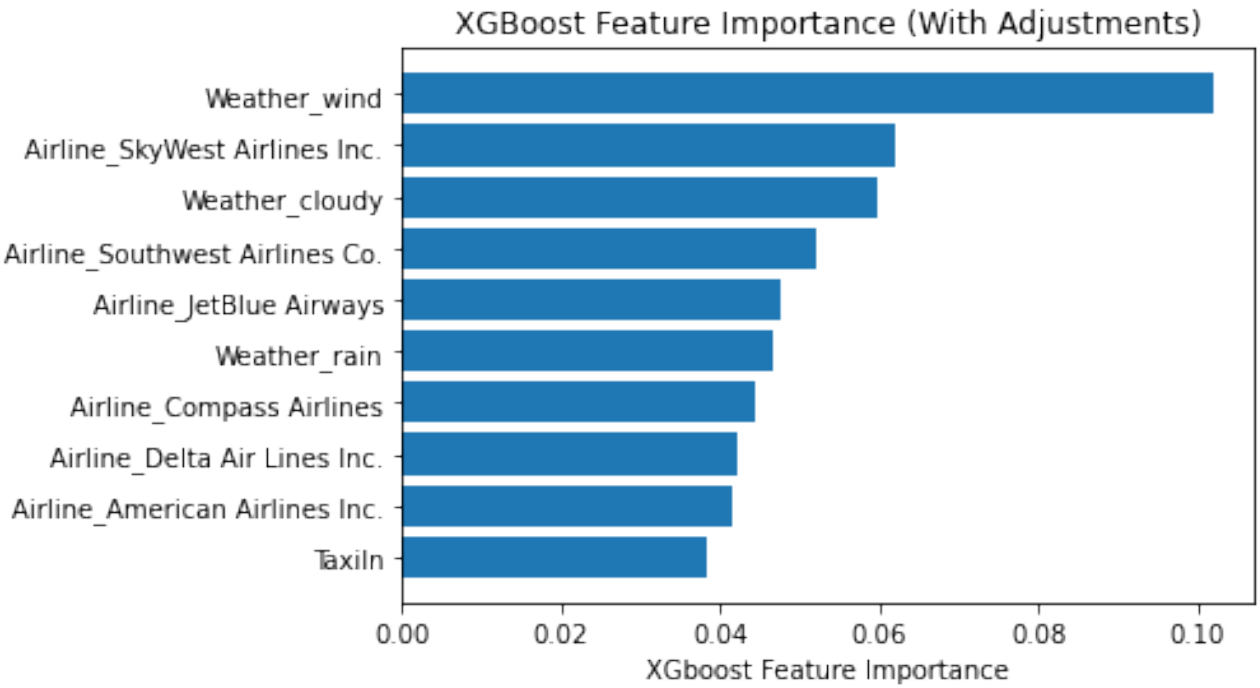


*Figure 11: Feature importance chart for XGBoost model with adjustments.*

The XGBoost model with adjustments still weighs certain airline companies highly, but the picture slightly differs from the model without adjustments. For example, it appears that the model with adjustments considers weather a more important factor for predictability than the unadjusted model. The One-Hot Encoded features of wind and cloudy weather are in the top three of most important features, whereas they were not even in the top 10 of the initial model's feature importance.

## 6.2. Analysis and Results for Hypothesis #1

The first hypothesis stated that weather would be the most important factor determining flight delays. The feature importance charts show that some weather features are among the most important for the adjusted XGBoost model. However, it is not a consistent theme across all of the models. For example, the weather features are not among the ten most important features for the Random Forest models. It can therefore be argued that this hypothesis should be rejected. However, it is noteworthy that one of the more accurate models considered certain weather features to be of high importance.

It is perhaps not surprising that LAX is not particularly affected by the weather conditions. As previously mentioned, the weather in the Los Angeles area is generally very stable. There are few extreme weather shocks like massive rainfall, snowstorms, or blizzards. Furthermore, airports have become increasingly robust at handling smaller weather obstacles, which means the effects of adverse weather is not going to be as severe as previously. This could contribute to why weather does not appear to be a highly important factor for our models.

### 6.3. Analysis and Results for Hypothesis #2

The second hypothesis stated that there would be limited correlation between the operating airline and the predictability of delays. This did not turn out to be true as most of the models had airlines as one of the most important features. For example, an airline feature was the most important feature for both of the XGBoost models. It was not as significant a picture for the Random Forest models, but both had one airline feature in the list of top 10 most important features.

The predictability of a flight's departure time could be affected by the airline for a variety of different reasons. Some airlines are simply more efficient than their competitors. This could involve things such as aircraft maintenance or ground handling, which in turn enables them to minimize delays. The fleet size or type of aircraft available to the airline might also play a role as these can help them navigate unexpected circumstances better than others. For example, if the aircraft for a flight has been severely delayed on the previous route, some airlines might be able to call up a reserve aircraft, which allows them to stay on-schedule and limit chain reactions.

It is noteworthy that some of the airlines that show up in the feature importance charts are budget airlines. These airlines usually operate with the point-to-point route network system. This system is beneficial because it offers more direct flight options for travelers, but it is also more susceptible to chain delays. Southwest Airlines is an example of how wrong it can go. The airline had a catastrophic period in late December 2022, where it had to cancel about 16,700 flights (Josephs, 2023). Multiple analysts have pointed to the system as one of the causes behind the meltdown (Wall Street Journal, 2023).

### 6.4. Analysis and Results for Hypothesis #3

The third hypothesis stated that the Neural Network model would be the most accurate of our three models. This did turn out to be true as the non-adjusted Neural Network model was the best performing model. It achieved an accuracy score of 95%, which was four percentage points better than the adjusted XGBoost model. Interestingly, the Neural Network model performed worse on the balanced dataset, while all of the other models saw improvements.

The Neural Network algorithm is generally considered more advanced than the two others, so it would make sense that it would achieve a higher accuracy score. While this is true, there is no guarantee that it will be true for your dataset. In our case, it turned out to be a correct prediction. This can probably be explained by our dataset being substantial enough for the algorithm to properly learn the patterns of the data. It had a decent number of features and observations, making it ideal for an algorithm that works best on large datasets. We could also see that the accuracy score for the Neural Network model dropped significantly after we shrunk the dataset to balance the output values, while the other two algorithms performed better. This is another indication of the large dataset being a key component in the model's success.

### 6.5. Summary of Hypotheses Results

In summary, not all of our hypotheses turned out to be correct. The only completely correct hypothesis was that the Neural Network model would perform best. The other two hypothesis could largely be rejected as the airline of the flight was a significant feature for multiple models, while the weather was not definitively the most significant factor for our models.

## 7. Discussion and Interpretation

In this section, we will discuss the practical use cases of flight delay prediction models, ethical concerns, and how airports could expand their use of machine learning and artificial intelligence in the future.

## 7.1. Practical Applications

A flight delay prediction model can have several practical use cases for airports. It can help improve efficiency in the airport's operations, enhance passenger experience, and optimize price settings. In this section, we will dive deeper into how an airport can utilize the knowledge obtained from our model, and how that might influence the overall airport experience for various stakeholders.

### 7.1.1. Operational Efficiency

There are several ways in which better anticipating flight delays can improve operational efficiency for airports. The first is being better to proactively adjust flight schedules. It becomes easier to plan around delays if you can accurately predict them. If an airport anticipates a flight to be delayed, it can make sense to adjust flight schedules to avoid causing a chain of delays. This will also help air traffic management as air traffic controllers can use this information to optimize flight routing and airspace management. This could help reduce congestion and improve the overall traffic efficiency.

Resource allocation is another area where flight delay predictions can make a significant impact. Airlines can use these predictions to allocate resources, such as ground staff, gate assignments, and maintenance crews, minimizing delay impacts and ensuring smooth operations. Crew management could benefit from accurate flight delay predictions as well. By utilizing these predictions, airlines can optimize crew scheduling to allocate flight crew members efficiently across flights.

Proactive maintenance is another aspect of the aviation industry that can be improved through flight delay predictions. Our model did not consider data related to aircrafts, but it should be possible to incorporate this information into similar models. This would allow airports to identify potential delays due to technical issues. This could be passed on to the airlines, which could schedule maintenance tasks more efficiently, reducing downtime and increasing aircraft availability. Finally, airport operations stand to gain from flight delay predictions. Airports can optimize the use of resources such as gates, security checkpoints, and baggage handling systems, resulting in improved overall efficiency and reduced congestion. In summary, the effective use of flight delay predictions can significantly enhance the operational efficiency of airports and airlines.

### 7.1.2. Customer Experience

Predicting flight delays with greater accuracy could help improve the passenger experience at airports. An example could be warnings about potential delays. While most people will be displeased with having their flight delayed, it makes it easier to stomach and plan around it if you are given a warning beforehand. Giving this type of information to passengers could lead to increased loyalty and customer satisfaction. It would also allow airports and airlines to give a more personalized service. Alternative travel suggestions or tailored compensation packages could serve as a consolation for a delayed flight.

There is an argument to be made that it is not in the airport's interest to reveal potential delays. If a passenger sees that a flight has a high risk of being delayed, he or she might decide to cancel or rebook the ticket. This could become expensive for airports and airlines as the predictability of passengers would decrease. The outrage over flights that were not predicted to be delayed could also increase, so the question is how much total customer satisfaction would actually improve?

### 7.1.3. Others

Machine learning can also play an integral role in enhancing the industry's efficiency and sustainability through its application in dynamic pricing and revenue management. By incorporating flight delay predictions into dynamic pricing models, airlines can optimize their pricing strategies based on the likelihood of delays or cancellations. This allows them to maximize revenue and improve the overall customer experience.

Moreover, these models can potentially contribute to reducing the environmental impact of aviation. By minimizing delays and enhancing operational efficiency, these predictions help decrease fuel consumption and the associated greenhouse gas emissions. As a result, the aviation industry can actively work towards mitigating its impact on the environment while maintaining high standards of service.

## 7.2. Limiting Delays at LAX

Delays are an inevitable part of airport operations as many moving parts has to come together. However, it should be possible to limit the number of delays by being proactive and improving the current systems. Our analysis showed that certain airlines are more predictable than others in terms of departing on time. LAX could use this information to review the processes for airlines that experience significant delays. The findings could also prove useful for when the airport needs to be expanded or renovated. There might be design proposals that could limit the number of delays.

The departure time also proved to be an important factor in predicting delays. (Mayer & Sinai, 2003) described how flights are often scheduled in clusters, so this is likely the result of congestion at certain peak times. It is difficult to avoid this type of delay as the airport is running at maximum capacity in these hours. However, it could be a sign that the airport should be looking for ways to expand its capacity. A more radical solution is to demand that airlines fly larger planes. This could reduce the number of aircrafts that would need to take off and land, while maintaining the same passenger numbers as before. This is probably only a suitable solution if the airport is prepared to change its route network. For example, it is unlikely that an airline would be willing to use a Boeing 747 or Airbus A380 on a short domestic flight.

Some of the factors in the analysis fall outside of LAX's control. For example, the airport has limited control over the general air traffic congestion, and it is also unable to influence the weather. The best thing an airport can do is take precautions to limit the impact of these outside factors.

## 7.3. Ethical Concerns

Using machine learning for flight delay predictions can create many advantages for airports. However, there are also several ethical concerns that one must consider when implementing such systems. One of the most discussed issues when dealing with systems built on machine learning is data privacy. A flight delay prediction model will require large amounts of data, and the accuracy of the model is likely to be greater the more data it is being fed. Our model only uses data that is publicly available, but it is not unthinkable that private data could be used to train similar models

and make them even more accurate. This can potentially create a dilemma for airports as they will have to weigh up whether to pursue more accuracy or protect the privacy of their customers.

Data bias and fairness is another aspect that must be considered when using machine learning models. The quality of predictions depends on the quality of data that is used to train the models. If the data is biased, the resulting predictions are going to be less accurate for some groups or situations. This is not an especially hurtful issue with our model as it does not contain passenger-specific information. However, there is still the risk of bias if certain flight routes, weather conditions, or airlines are over- or underrepresented.

Transparency is also an important topic when dealing with algorithms. Machine learning models can often feel like black boxes, where something goes in, and a result comes out. This is great for its simplicity in application, but it simultaneously makes it difficult for stakeholders to understand how or why a certain prediction was made. The complexity of algorithms is often so high that it will take meaningful time to explain thoroughly. However, it is important for trust that the inner workings of a model can be explained in broad strokes to a non-expert audience. This paper is an example of trying to open the black box.

A related topic is that of accountability. If a model fails to provide accurate predictions, who is going to be held accountable for its consequences? In the case of flight delay prediction, it is limited how much damage a failed prediction can cause. It could result in minor financial losses and negative customer reviews, but it is ultimately not a life-or-death situation like in other industries. As with other cases related to artificial intelligence, the role of governance is essential. The complex nature of machine learning systems allows for many different scenarios, where blame can be thrown around multiple different stakeholders. It is therefore important to have considered the most likely scenarios and determined who will shoulder responsibility for failures beforehand.

Lastly, increased automation is inevitably going to lead to displacement of employees in certain positions. This is not a new concept as technological advancements have historically made many jobs irrelevant while also creating many new ones. However, in an age with increased focus on

corporate social responsibility, it is important for airports to provide opportunities for retraining or upskilling to employees that are likely to see their jobs eliminated. It can be argued that it is even more important for an institution like LAX, which is a large employer of traditional blue-collar workers and prides itself on its diversity and social responsibility initiatives (Los Angeles World Airports, 2023).

### 7.4. Future Uses

Flight delay prediction is just one of many ways machine learning can help improve operations and passenger experience at airports. The potential of machine learning, and artificial intelligence as a whole, is almost limitless, and it makes sense for airports to be one of the first places to use these new technologies due to their need for scalable solutions. In this section, we will briefly cover some of the potential use cases for artificial intelligence at airports in the future.

The use of object detection and facial recognition is already prevalent at certain airports as tasks such as identifying potentially harmful objects and checking passports are being done using artificial intelligence. In the future, it might be possible to leverage this technology even further. For example, the Transportation Security Administration in the United States is testing out facial recognition technology at certain airports. The goal is to streamline the security process and identify individuals who should not be flying. This could potentially be a massive improvement for airports as security is a heavy task in terms of labor and equipment required. It is also a source of frustration for passengers as the waiting time is often unpredictable.

Other airport domains that could be helped by automation include facility management and air traffic control. Facility management is a huge task at airports since the buildings usually cover thousands of square meters, and there is a steady stream of traffic almost daily. This could be helped with future technological breakthroughs. The use of robotic vacuum cleaners is already common in domestic households, so it is not difficult to imagine a future where robots will take over some of the cleaning duties at airports. Air traffic control can potentially also use artificial intelligence to automate the repetitive processes of the job. For example, artificial intelligence can be used to interpret images taken from a camera installed on the top of the control tower. The system can

check for aircrafts and then notify controllers, so that the next plane can be signaled in on a cleared runway. This could be a useful way of freeing up capacity for existing air traffic controllers as there is also expected to be a shortage of trained air traffic controllers in the future (Ledsom, 2023).

Another interesting frontier is the use of generative artificial intelligence at airports. Services like ChatGPT have already given us a glimpse of how effective the technology can be, and it only feels like a matter of time before it becomes a major thing at airports. It is not difficult to envision how the technology might be able to help different stakeholders at airports. Travelers could use it as a personal travel assistant that can plan and schedule trips, while airport customer service could use it to answer chat questions.

While many of these ideas are great, there is a risk of overreliance on technology. As machine learning systems become more integrated into services and processes, there may be a risk of overreliance on the technology. This could lead to a lack of critical thinking or reduced human input in the process, which could be detrimental if the system produces inaccurate predictions or encounters unforeseen situations. It is therefore important that proper governance is instilled when implementing systems built on artificial intelligence.

# 8. Conclusion

In conclusion, LAX is a large international airport that is also affected by flight delays. Before the pandemic, the airport saw over 80 million passengers come through its gates annually. With this type of scale, the need to keep operations running smoothly is paramount. However, flight delays are difficult to predict, and they also tend to propagate through the system and cause a chain of delays. One of the most sophisticated ways to predict flight delays is using machine learning. We built three different models, which were tasked with predicting whether a flight would be on-time or have a small, medium, or large delay. It was found that we could achieve a 95% accuracy score using a Neural Network algorithm. The Neural Network model was able to beat out the Random Forest model and XGBoost model, which were able to achieve accuracy scores of 89% and 92% respectively. An interesting observation is that the highest accuracy score was achieved by not

balancing the dataset. This can be explained by the fact that the sheer data amount was larger with the unbalanced dataset, and that plays into the strengths of a Neural Network algorithm.

Some of the practical applications for a flight delay prediction model is to improve operational efficiency, customer experience, and pricing. However, while there are many advantages to using machine learning, one must also be aware of the ethical dilemmas regarding its use. For example, data privacy is an increasingly important topic in today's debate, and there is a scenario where airports have to choose between higher accuracy or greater data privacy for their passengers. Similarly, it is important to have open discussions about transparency, accountability, and job displacements. Lastly, there is a huge potential for future uses of machine learning and artificial intelligence at airports. Services like airport security, facility management, and readings of information are ripe for automation, and some airports are already testing out systems that can help carry part of that burden. Additionally, great advancements will likely be made in the generative artificial intelligence space in the next few years. This could open up a whole slew of opportunities for airports.

## 8.1. New Knowledge Contribution

The research gap identified was to bring flight delay prediction and a focus on LAX together. We have examined how various machine learning algorithms perform on recent flight data from LAX, and it was discovered that the Neural Network algorithm works particularly well on this data. The Random Forest Classifier and XGBoost algorithm also achieved good accuracy scores, but it was ultimately the more advanced Neural Network algorithm that performed the best.

Furthermore, we discovered that the operating airline of the flight played an important role in most of the models. This could be explained by factors such as differing operational efficiencies, varying fleet sizes, and different route networks. The departure time also turned out to be an important factor, which makes sense given the busyness of an airport fluctuates throughout the day. It was also discovered that weather was not a major factor for most of our models. The adjusted XGBoost model was the only model to have weather conditions among its most important features, which

was a surprise given previous studies on the topic. It could be explained by LAX's geographical location as the airport is located in an area with stable temperatures and few weather shocks.

## 8.2. Threats to Validity

This paper has certain limitations that will be addressed in this section. The first is that our models could have considered more factors such as air traffic congestion, number of employees at the airport, or the passenger composition for each flight. This was mostly not possible due to data availability. It is unknown whether adding these features would have led to increased accuracy, but it would nonetheless have given a more detailed picture of the conditions surrounding each flight. It would also have been interesting to look at the importance of these features for the models.

A different limitation is the weather information for the flights. As previously mentioned, the weather dataset was obtained in daily intervals. This lessens the data specificity as flights on the same day are going to have the same weather information. If we had obtained the weather dataset in hourly or minute intervals, this would not have been the case and there would most likely have been more data diversity. It is also possible that weather would have been a more important factor in each of the models, which could have led us to a different conclusion for our first hypothesis. However, this is just speculation, so it is also possible that it would not have made much of a difference.

# 9. References


Airports Council International. (2020). *ACI reveals top 20 airports for passenger traffic, cargo, and aircraft movements* . Retrieved from Airports Council International: https://aci.aero/2020/05/19/aci-reveals-top-20-airports-for-passenger-traffic-cargo-and-aircraft-movements/

Alla, H., Moumoun, L., & Balouki, Y. (2021). *A Multilayer Perceptron Neural Network with Selective-Data Training for Flight Arrival Delay Prediction .*

Bloom, L. B. (2018). *15 Best And Worst Airports In The US For Flight Delays .* Retrieved from Forbes: https://www.forbes.com/sites/laurabegleybloom/2018/10/22/15-best-and-worst-airports-in-the-us-for-flight-delays/?sh=39a8bf407bca

Borsky, S., & Unterberger, C. (2019). *Bad weather and flight delays: The impact of sudden and slow onset weather events .*

Bureau of Transportation. (2023). *On-Time: Marketing Carrier On-Time Performance (Beginning January 2018)* . Retrieved from Bureau of Transportation: https://www.transtats.bts.gov/Fields.asp?gnoyr_VQ=FGK

Cheevachaipimol, W., Teinwan, B., & Chutima, P. (2021). *Flight Delay Prediction Using a Hybrid Deep Learning Method.*

City News Service. (2022). *LAX makes clean energy commitments in agreement with workers' union.* Retrieved from https://spectrumnews1.com/ca/la-west/environment/2022/02/04/lax-makes-clean-energy-commitments-in-agreement-with-workers--union

Clarke, J. P., Brooks, J., Nagle, G., Scacchioli, A., White, W., & Liu, S. R. (2013). *Optimized Profile Descent Arrivals at Los Angeles International Airport.*

Conboye, J., & Hook, L. (2019). *Flight shame: Airlines are under rising pressure to cut their carbon emissions* . Retrieved from Los Angeles Times: https://www.latimes.com/business/story/2019-08-27/flight-shame-can-airlines-ever-reduce-their-emissions

Federal Aviation Administration. (2010). *Total Delay Impact Study.* Retrieved from https://news.berkeley.edu/2010/10/18/flight_delays/

Federal Aviation Administration. (2014). *LAX Capacity Profile 2014.*

FlightConnections. (2023). *Los Angeles International Airport*. Retrieved from Flight Connections: https://www.flightconnections.com/flights-from-los-angeles-lax

Hamilton, T., Schell, T. L., Stevens, D., & Mesic, R. (2004). *Near-Term Options for Improving Security at Los Angeles International Airport.*

Horn, R. E., & Orman, J. C. (1975). *AIRPORT AIRSIDE AND LANDSIDE INTERACTION.*

Josephs, L. (2023). *Southwest CEO maps out a recovery after holiday meltdown: 'We have work to do'* . Retrieved from CNBC: https://www.cnbc.com/2023/01/25/southwest-airlines-tries-to-improve-its-system-after-holiday-meltdown.html

Ledsom, A. (2023). *U.S. Pilot, Air Traffic Controller Shortage Leading To Fewer Flights* . Retrieved from Forbes: https://www.forbes.com/sites/alexledsom/2023/04/30/us-pilot-air-traffic-controller-shortage-leading-to-fewer-flights/

Li, Q., & Jing, R. (2021). *Flight delay prediction from spatial and temporal perspective.*

Lohmann, G., Albers, S., Koch, B., & Pavlovich, K. (2009). *From hub to tourist destination – An explorative study of Singapore and Dubai's aviation-based transformation.*

Los Angeles International Airport. (2023). 2022 Annual Financial Report.

Los Angeles World Airports. (2022). *LAX at a Glance: The Airfield.* Retrieved from FlyLAX: https://www.flylax.com/-/media/flylax/media-center/pdfs/fs---airfield-feb_2021

Los Angeles World Airports. (2023). *Strategies*. Retrieved from Los Angeles World Airports: https://www.lawa.org/lawa-sustainability/strategies

Los Angeles World of Airports. (2022). *10-Year Summary of Passengers*. Retrieved from Los Angeles World of Airports: https://www.lawa.org/lawa-investor-relations/statistics-for-lax/10-year-summary/passengers

Mayer, C., & Sinai, T. (2003). *Network Effects, Congestion Externalities, and Air Traffic Delays: Or Why All Delays Are Not Evil* .

Oster, C., & Strong, J. (2021). *Economic effects of shifting airport activity in the Los Angeles metro region.*

Statista. (2023). *Number of public and private airports in the United States from 1990 to 2021*. Retrieved from Statista: https://www.statista.com/statistics/183496/number-of-airports-in-the-united-states-since-1990/

Sweet Jr., C. P. (1975). *AIRSIDE AND OFF-AIRPORT FACTORS AND LANDSIDE CAP ACITY.*

Vascik, P. D., & Hansman, J. (2017). *Evaluation of Key Operational Constraints Affecting On-Demand Mobility for Aviation in the Los Angeles Basin: Ground Infrastructure, Air Traffic Control and Noise.* American Institute of Aeronautics and Astronautics, Inc.

Wall Street Journal. (2023). *United Airlines Explains How It Orchestrates 30,000 Weekly Flights | WSJ Travel Guides.* Retrieved from YouTube: https://www.youtube.com/watch?v=mDkSehPF0M8&list=LL&index=1

Weikel, D. (2015). *LAX could see more than 100 million travelers a year by 2040.* Retrieved from Los Angeles Times: https://www.latimes.com/local/california/la-me-adv-air-travel-forecast-20150726-story.html

Westerdahl, D., Fruin, S. A., Fine, P. L., & Sioutas, C. (2008). *The Los Angeles International Airport as a source of ultrafine particles and other pollutants to nearby communities.*

Yamamoto, B., & Paternoster, J. (2017). *Aligning the airport community to improve LAX guest satisfaction: Every journey begins with a very important first step.* Henry Stewart Publications.

Yellow Productions. (2022). *Why Everyone Hates LAX Airport*. Retrieved from YouTube: https://www.youtube.com/watch?v=KFdGejjV3mg

Zámková, M., Prokop, M., & Stolín, R. (2019). *A review of flight delay causes at Spanish airports based on statistical analysis of categorical data.*

# 10. Appendix

## Appendix A – Dataset Description (Bureau of Transportation, 2023).

| COLUMN | DESCRIPTION |
|---|---|
| Year | Year |
| QUARTER | Quarter (1-4) |
| MONTH | Month |
| DAYOFMONTH | Day of Month |
| DAYOFWEEK | Day of Week |
| FLIGHTDATE | Flight Date (yyyymmdd) |
| MARKETING_AIRLINE_NETWORK | Unique Marketing Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years. |
| OPERATED_OR_BRANDED_CODE_SHARE_PARTNERS | Reporting Carrier Operated or Branded Code Share Partners |
| DOT_ID_MARKETING_AIRLINE | An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation. |
| IATA_CODE_MARKETING_AIRLINE | Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code. |
| FLIGHT_NUMBER_MARKETING_AIRLINE | Flight Number |
| ORIGINALLY_SCHEDULED_CODE_SHARE_AIRLINE | Unique Scheduled Operating Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users,for example, PA, PA(1), PA(2). Use this field for analysis across a range of years. |

| | |
|---|---|
| DOT_ID_ORIGINALLY_SCHEDULED_CODE_SHARE_AIRLINE | An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation. |
| IATA_CODE_ORIGINALLY_SCHEDULED_CODE_SHARE_AIRLINE | Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code. |
| FLIGHT_NUM_ORIGINALLY_SCHEDULED_CODE_SHARE_AIRLINE | Flight Number |
| OPERATING_AIRLINE | Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years. |
| DOT_ID_OPERATING_AIRLINE | An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation. |
| IATA_CODE_OPERATING_AIRLINE | Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code. |
| TAIL_NUMBER | Tail Number |
| FLIGHT_NUMBER_OPERATING_AIRLINE | Flight Number |
| ORIGINAIRPORTID | Origin Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused. |
| ORIGINAIRPORTSEQID | Origin Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time. |
| ORIGINCITYMARKETID | Origin Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market. |
| ORIGIN | Origin Airport |

| | |
|---|---|
| ORIGINCITYNAME | Origin Airport, City Name |
| ORIGINSTATE | Origin Airport, State Code |
| ORIGINSTATEFIPS | Origin Airport, State Fips |
| ORIGINSTATENAME | Origin Airport, State Name |
| ORIGINWAC | Origin Airport, World Area Code |
| DESTAIRPORTID | Destination Airport, Airport ID. An identification number assigned by US DOT to identify a unique airport. Use this field for airport analysis across a range of years because an airport can change its airport code and airport codes can be reused. |
| DESTAIRPORTSEQID | Destination Airport, Airport Sequence ID. An identification number assigned by US DOT to identify a unique airport at a given point of time. Airport attributes, such as airport name or coordinates, may change over time. |
| DESTCITYMARKETID | Destination Airport, City Market ID. City Market ID is an identification number assigned by US DOT to identify a city market. Use this field to consolidate airports serving the same city market. |
| DEST | Destination Airport |
| DESTCITYNAME | Destination Airport, City Name |
| DESTSTATE | Destination Airport, State Code |
| DESTSTATEFIPS | Destination Airport, State Fips |
| DESTSTATENAME | Destination Airport, State Name |
| DESTWAC | Destination Airport, World Area Code |
| CRSDEPTIME | CRS Departure Time (local time: hhmm) |

| | |
|---|---|
| DEPTIME | Actual Departure Time (local time: hhmm) |
| DEPDELAY | Difference in minutes between scheduled and actual departure time. Early departures show negative numbers. |
| DEPDELAYMINUTES | Difference in minutes between scheduled and actual departure time. Early departures set to 0. |
| DEPDEL15 | Departure Delay Indicator, 15 Minutes or More (1=Yes) |
| DEPARTUREDELAYGROUPS | Departure Delay intervals, every (15 minutes from 180) |
| DEPTIMEBLK | CRS Departure Time Block, Hourly Intervals |
| TAXIOUT | Taxi Out Time, in Minutes |
| WHEELSOFF | Wheels Off Time (local time: hhmm) |
| WHEELSON | Wheels On Time (local time: hhmm) |
| TAXIIN | Taxi In Time, in Minutes |
| CRSARRTIME | CRS Arrival Time (local time: hhmm) |
| ARRTIME | Actual Arrival Time (local time: hhmm) |
| ARRDELAY | Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers. |
| ARRDELAYMINUTES | Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0. |
| ARRDEL15 | Arrival Delay Indicator, 15 Minutes or More (1=Yes) |
| ARRIVALDELAYGROUPS | Arrival Delay intervals, every (15-minutes from 180) |
| ARRTIMEBLK | CRS Arrival Time Block, Hourly Intervals |

| | |
|---|---|
| CANCELLED | Cancelled Flight Indicator (1=Yes) |
| CANCELLATIONCODE | Specifies The Reason For Cancellation |
| DIVERTED | Diverted Flight Indicator (1=Yes) |
| CRSELAPSEDTIME | CRS Elapsed Time of Flight, in Minutes |
| ACTUALELAPSEDTIME | Elapsed Time of Flight, in Minutes |
| AIRTIME | Flight Time, in Minutes |
| FLIGHTS | Number of Flights |
| DISTANCE | Distance between airports (miles) |
| DISTANCEGROUP | Distance Intervals, every 250 Miles, for Flight Segment |
| CARRIERDELAY | Carrier Delay, in Minutes |
| WEATHERDELAY | Weather Delay, in Minutes |
| NASDELAY | National Air System Delay, in Minutes |
| SECURITYDELAY | Security Delay, in Minutes |
| LATEAIRCRAFTDELAY | Late Aircraft Delay, in Minutes |
| FIRSTDEPTIME | First Gate Departure Time at Origin Airport |
| TOTALADDGTIME | Total Ground Time Away from Gate for Gate Return or Cancelled Flight |
| LONGESTADDGTIME | Longest Time Away from Gate for Gate Return or Cancelled Flight |

| | |
|---|---|
| DIVAIRPORTLANDINGS | Number of Diverted Airport Landings |
| DIVREACHEDDEST | Diverted Flight Reaching Scheduled Destination Indicator (1=Yes) |
| DIVACTUALELAPSEDTIME | Elapsed Time of Diverted Flight Reaching Scheduled Destination, in Minutes. The ActualElapsedTime column remains NULL for all diverted flights. |
| DIVARRDELAY | Difference in minutes between scheduled and actual arrival time for a diverted flight reaching scheduled destination. The ArrDelay column remains NULL for all diverted flights. |
| DIVDISTANCE | Distance between scheduled destination and final diverted airport (miles). Value will be 0 for diverted flight reaching scheduled destination. |
| DIV1AIRPORT | Diverted Airport Code1 |
| DIV1AIRPORTID | Airport ID of Diverted Airport 1. Airport ID is a Unique Key for an Airport |
| DIV1AIRPORTSEQID | Airport Sequence ID of Diverted Airport 1. Unique Key for Time Specific Information for an Airport |
| DIV1WHEELSON | Wheels On Time (local time: hhmm) at Diverted Airport Code1 |
| DIV1TOTALGTIME | Total Ground Time Away from Gate at Diverted Airport Code1 |
| DIV1LONGESTGTIME | Longest Ground Time Away from Gate at Diverted Airport Code1 |
| DIV1WHEELSOFF | Wheels Off Time (local time: hhmm) at Diverted Airport Code1 |
| DIV1TAILNUM | Aircraft Tail Number for Diverted Airport Code1 |
| DIV2AIRPORT | Diverted Airport Code2 |
| DIV2AIRPORTID | Airport ID of Diverted Airport 2. Airport ID is a Unique Key for an Airport |
| DIV2AIRPORTSEQID | Airport Sequence ID of Diverted Airport 2. Unique Key for Time Specific Information for an Airport |

| | |
|---|---|
| DIV2WHEELSON | Wheels On Time (local time: hhmm) at Diverted Airport Code2 |
| DIV2TOTALGTIME | Total Ground Time Away from Gate at Diverted Airport Code2 |
| DIV2LONGESTGTIME | Longest Ground Time Away from Gate at Diverted Airport Code2 |
| DIV2WHEELSOFF | Wheels Off Time (local time: hhmm) at Diverted Airport Code2 |
| DIV2TAILNUM | Aircraft Tail Number for Diverted Airport Code2 |
| DIV3AIRPORT | Diverted Airport Code3 |
| DIV3AIRPORTID | Airport ID of Diverted Airport 3. Airport ID is a Unique Key for an Airport |
| DIV3AIRPORTSEQID | Airport Sequence ID of Diverted Airport 3. Unique Key for Time Specific Information for an Airport |
| DIV3WHEELSON | Wheels On Time (local time: hhmm) at Diverted Airport Code3 |
| DIV3TOTALGTIME | Total Ground Time Away from Gate at Diverted Airport Code3 |
| DIV3LONGESTGTIME | Longest Ground Time Away from Gate at Diverted Airport Code3 |
| DIV3WHEELSOFF | Wheels Off Time (local time: hhmm) at Diverted Airport Code3 |
| DIV3TAILNUM | Aircraft Tail Number for Diverted Airport Code3 |
| DIV4AIRPORT | Diverted Airport Code4 |
| DIV4AIRPORTID | Airport ID of Diverted Airport 4. Airport ID is a Unique Key for an Airport |
| DIV4AIRPORTSEQID | Airport Sequence ID of Diverted Airport 4. Unique Key for Time Specific Information for an Airport |
| DIV4WHEELSON | Wheels On Time (local time: hhmm) at Diverted Airport Code4 |

| | |
|---|---|
| DIV4TOTALGTIME | Total Ground Time Away from Gate at Diverted Airport Code4 |
| DIV4LONGESTGTIME | Longest Ground Time Away from Gate at Diverted Airport Code4 |
| DIV4WHEELSOFF | Wheels Off Time (local time: hhmm) at Diverted Airport Code4 |
| DIV4TAILNUM | Aircraft Tail Number for Diverted Airport Code4 |
| DIV5AIRPORT | Diverted Airport Code5 |
| DIV5AIRPORTID | Airport ID of Diverted Airport 5. Airport ID is a Unique Key for an Airport |
| DIV5AIRPORTSEQID | Airport Sequence ID of Diverted Airport 5. Unique Key for Time Specific Information for an Airport |
| DIV5WHEELSON | Wheels On Time (local time: hhmm) at Diverted Airport Code5 |
| DIV5TOTALGTIME | Total Ground Time Away from Gate at Diverted Airport Code5 |
| DIV5LONGESTGTIME | Longest Ground Time Away from Gate at Diverted Airport Code5 |
| DIV5WHEELSOFF | Wheels Off Time (local time: hhmm) at Diverted Airport Code5 |
| DIV5TAILNUM | Aircraft Tail Number for Diverted Airport Code5 |
| DUPLICATE | Duplicate flag marked Y if the flight is swapped based on Form-3A data |