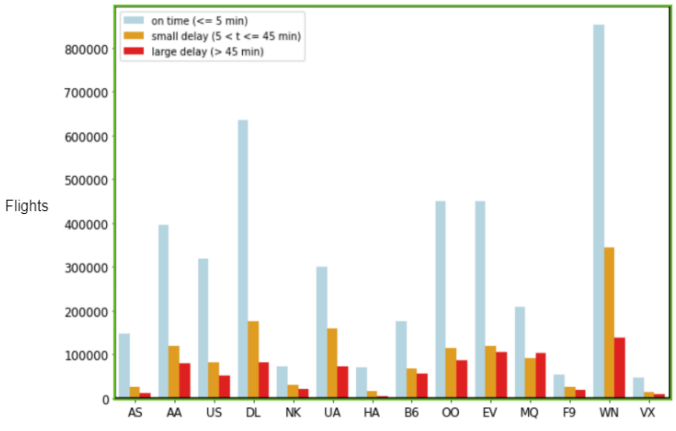
Flight Delay Prediction

Vyshnavi Adusumelli (vadusum)
Parth Katlana (pkatln)
Sunandini Medisetti (smedise)
Hsueh-Yang Yu (hyu25)

I. Problem Statement

The demand for flights has grown as a result of population expansion and the necessity to survive in a fast-paced world. Due to its effectiveness, swiftness, ease, accessibility, and efficient use of time, many choose flying as a mode of transportation. However, it is a well-known fact that weather, poor staffing, and dense air traffic are the leading causes of flight delays. Flight delays have an impact on many aspects of the travel ecosystem, such as consumer displeasure, financial losses, and crew costs. Customer interests and relationships suffer as a result.



Deep data analysis and precise flight delay forecasting are crucial components of our project. This forecasting tool can aid in enhancing airline operations, increasing efficiency, and enhancing customer satisfaction. More significantly, it could lessen the ”knock-on flight effect,” in which the delay of one aircraft causes a chain reaction of delays on subsequent flights. The company’s earnings are significantly impacted by this. Therefore, it is essential to be aware of any potential flight delays in advance and to estimate how long they could last. This can aid in the development of risk processing and mitigation strategies for both passengers and the aviation department.

II. Related Work

Many research endeavors have tackled the challenge of forecasting flight delays, utilizing a variety of machine learning techniques,and deep learning approaches. Nonetheless, the majority of these studies have concentrated solely on a particular airline or a limited number of airports in their investigations. My objective is to construct a model that can proficiently predict flight delays for all flights and destinations, regardless of the airline or airport involved. The subsequent papers are some of the most noteworthy contributions to this field.

The research conducted by Simon Briceno and Dimitri N. Mavris [1] introduced a Model that integrated Synthetic Minority Over-sampling Technique (SMOTE) to forecast airline delays caused by weather conditions, utilizing machine learning algorithms. The primary goal of using this paper is to acquire knowledge regarding the fundamental principles of implementing decision trees and random forests and explore how machine learning models can be improved for classification purposes.

Jerry Ye, Jyhherng Chow, Jiang Chen, and Zhaohui Zheng [2] created a model that utilized gradient boosted distributed decision trees to predict flight arrival and departure delays, following the analysis of the raw flight data. Their investigation aimed to delve into the research on the Machine Learning Algorithm Gradient Boosted Decision Tree and how it can be employed in predicting

flight delays. This paper is used to implement Gradient Boosted Algorithm for our process.

Ning Xu, George Donohue, Kathryn Blackmond Laskey, and Chun-Hung Chen [3] used Bayesian networks to estimate the delay in the propagation of the national aviation system. This study was used to comprehend how the Bayesian networkswere applied to complicated systems.

III. Approach

The implementation has been divided into multiple phases for the ease of training the model. In the primary stage the dataset is preprocessed. Later, a variety of models were employed for both regression and classification tasks.

A. Preprocessing

In order to prepare data for machine learning models, it is important to perform several preprocessing steps, which may include handling missing values, reformatting time data, selecting relevant features, encoding categorical data, normalizing data, and creating new features for classification purposes. The specific techniques utilized for each step are described below:

The original time data in the dataset is in the form of four-digit numbers and is not useful for modeling purposes. Therefore, it is transformed into HH:MM format and added as new columns for Departure time, Scheduled arrival, Scheduled departure, and arrival time. Some columns in the dataset, such as Departure delay and taxi out, have a small number of missing values. These rows are removed since they represent only a small portion of the overall dataset.

Certain features are represented as text, so they are converted into numerical values using a Label encoder. The values are assigned beginning from zero in order to make the dataset more suitable for machine learning, as models do not perform well with text features. Not all features are necessary for predicting delays. As a result, only certain features are retained for prediction purposes, including Airline operator, Origin Airport, Destination Airport, Distance, Actual Departure, Date, Day, Scheduled Departure, Departure Delay, Actual Arrival, Scheduled Arrival, Arrival Delay, Scheduled Time, Elapsed Time, Air Time, Taxi In, Taxi Out, and Diverted.

A new binary feature is created for the classification model. This feature is based on the delay time and has a value of 1 assigned if the delay time is greater than 0, and 0 assigned otherwise. The dataset is standardized using Python's standard scalar library, so that all features are on the same relative scale. This is important because features have different magnitudes, and distance metrics used by the algorithms are sensitive to large variations in magnitudes.

B. Models - Regression and Classification

1. Regression:

One of the simplest machine learning algorithms is simple linear regression, which attempts to establish a linear relationship between the value to be predicted and the attributes utilized to make the prediction. It operates by obtaining a formula for predicting one variable using others, provided there is a causal relationship between them.

Another approach for regression is random forest regression [5], which is an ensemble technique that can be used for both regression and classification tasks. It generates multiple decision trees using bagging, which involves training all the decision trees on different data samples. The final prediction is derived by combining the outcomes of all the decision trees instead of relying on a single one.

Boosted linear regression is an ensemble machine learning technique [6] merges several feeble models to create a single robust model. This process follows a step-by-step approach with the belief that each subsequent model, when combined with the previous models, will decrease the overall forecasting inaccuracies.

2. Classification:

The K-neighbors classifier algorithm [7] operates by determining the k nearest neighbors to the data point to be predicted, and then assigning a

class value based on the majority class of those neighbors. The K-nearest neighbor algorithm can be executed in the following steps: first, the k value, which is the number of nearest neighbors to consider, is determined. Then, the distances between the data to be predicted and the training data are calculated and sorted. Next, the k nearest neighbors are identified, and their class labels are used to predict the class of the data point.

Logistic regression employs a logistic model with an 'S' shaped curve to predict values, particularly for binary classification problems where there are only two possible output classes such as true or false. This algorithm calculates probabilities and transforms them into a function to estimate the likelihood of the data point belonging to one of the classes.

The Decision Trees algorithm functions by constructing a tree structure where each node represents a choice between two possible branches - left or right. The objective of this algorithm is to continue questioning and building the tree until the purest possible splits are achieved. While the induction process is typically slow, the deduction process is rapid as it merely involves traversing the constructed tree and reaching the leaf node.

XGBoost is known for its high accuracy and ability to handle large datasets. This is particularly important in flight delay prediction, where there are a large number of variables that can affect flight schedules.It can identify the most important features in a dataset, which can be useful for understanding the factors that contribute to flight delays. XGBoost is optimized for speed, which means it can process large amounts of data quickly.

IV. Rationale

The approach mentioned above is suitable for the given problem and dataset because it involves preprocessing the data, transforming text features into numerical values, removing unnecessary features, and standardizing the dataset to make it more suitable for machine learning models. Furthermore, it employs a variety of models, both for regression and classification tasks, to capture the complex relationships and patterns in the dataset.

Simple linear regression is a straightforward approach that works well when there is a clear linear relationship between the variables, and it is easy to interpret the coefficients. Random forest regression [5], on the other hand, is an ensemble technique that can handle non-linear relationships and is less prone to overfitting. Boosted linear regression is another ensemble technique [6] that combines multiple models to reduce the overall forecasting inaccuracies.

The K-neighbors classifier algorithm is suitable for this problem because it can handle multi-class classification problems and is easy to interpret. Logistic regression is also a suitable algorithm for binary classification problems, as it estimates the probability of the data point belonging to one of the classes. Decision Trees algorithm is a powerful approach that can handle both categorical and numerical data, and it can capture non-linear relationships and interactions between features.

Overall, the chosen techniques are appropriate for this problem and dataset because they can handle the complexity of the data and capture the underlying patterns and relationships. Additionally, the ensemble techniques and decision trees can handle non-linear relationships and interactions between features, which are common in airline delay prediction problems.

V. Dataset

The dataset [4] used in this study was sourced from Kaggle and has 4 Million data points. It was originally collected by the U.S. Department of Transportation. The dataset primarily focuses on the performance of domestic flights within the United States and contains information regarding flights from the year 2015. This dataset can be used to perform various analyses, including identifying trends and patterns in flight delays and cancellations, assessing the performance of different airlines and airports, and examining the impact of weather on flight operations.

Furthermore, this dataset can also be utilized to develop predictive models that can accurately forecast flight delays and cancellations, which can

be beneficial for both airlines and passengers. Overall, this dataset provides a wealth of information that can be harnessed for a variety of analytical purposes. The dataset is quite comprehensive, and it encompasses various attributes as listed below:

```
['YEAR',
 'MONTH',
 'DAY',
 'DAY_OF_WEEK',
 'AIRLINE',
 'FLIGHT_NUMBER',
 'TAIL_NUMBER',
 'ORIGIN_AIRPORT',
 'DESTINATION_AIRPORT',
 'SCHEDULED_DEPARTURE',
 'DEPARTURE_TIME',
 'DEPARTURE_DELAY',
 'TAXI_OUT',
 'WHEELS_OFF',
 'SCHEDULED_TIME',
 'ELAPSED_TIME',
 'AIR_TIME',
 'DISTANCE',
 'WHEELS_ON',
 'TAXI_IN',
 'SCHEDULED_ARRIVAL',
 'ARRIVAL_TIME',
 'ARRIVAL_DELAY',
 'DIVERTED',
 'CANCELLED',
 'CANCELLATION_REASON',
 'AIR_SYSTEM_DELAY',
 'SECURITY_DELAY',
 'AIRLINE_DELAY',
 'LATE_AIRCRAFT_DELAY',
 'WEATHER_DELAY']
```

Fig. 1. Dataset Features

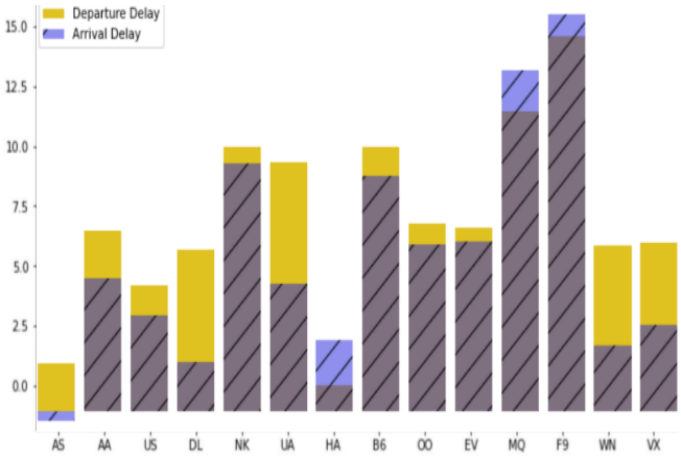


Fig. 2. Arrival and Departure Delay

VI. Hypotheses

Our model is capable of addressing one of the aviation industry's most interesting and challenging problems - the "Knock-on" effect. This phenomenon occurs when a delayed flight has a cascading effect on the entire network of flights, resulting in further delays and cancellations. Our model can be utilized to mitigate this effect and assist airlines in minimizing the impact of delays and disruptions on their operations, resulting in an improved travel experience for passengers.

Airlines can leverage our model by analyzing various data points such as weather conditions, air traffic congestion, and previous flight delays to identify flights that are at risk of being delayed. With this information, they can proactively take measures to minimize the impact on the rest of the network. By doing so, they can minimize the knock-on effect, thereby reducing the number of subsequent flight delays and cancellations.

VII. Experimental Design

A. Metrics Used for Regression

1. Mean Absolute Error - Indicates the differences between the anticipated and real values of the predictions.

2. Mean Squared Error - Evaluates the average of the sum of the squared errors.

3. Root Mean Squared Deviation - Measure of squared root of Mean Squared Error

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Where,

$\hat{y}$ – *predicted value of* y

$\bar{y}$ – *mean value of* y

B. Metrics Used for Regression

1. Precision - refers to the proportion of true positive predictions out of the total number of positive predictions made by the classifier.

2. Recall - measures the ability of the classifier to correctly identify the actual positive instances from the entire population of positive instances it has learned.

3. F1 score - represents the harmonic average of precision and recall, providing a single metric to evaluate the overall performance of a classifier.

$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$Precision = \frac{T_p}{T_p + F_p}$$

$$Recall = \frac{T_p}{T_p + T_n}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

VIII. Results

A correlation matrix is plotted to illustrate the relationships among the variables in the dataset. This matrix provides valuable insights into which variables are interconnected, with a particular focus on arrival delay.

Notably, we can observe a strong correlation between departure delay and arrival delay. However, it is worth mentioning that some flights still manage to arrive on time despite experiencing departure delays.

To start with, we retain the highly correlated variable as an attribute in the training set and analyze its performance. However, the most intriguing aspect of this analysis is how the removal of departure delay affects our model's accuracy.

A. Regression Analysis

1. Linear Regression - We begin with Linear Regression, one of the most basic regression models, which is trained on a preprocessed dataset that has been divided into separate training and testing subsets.

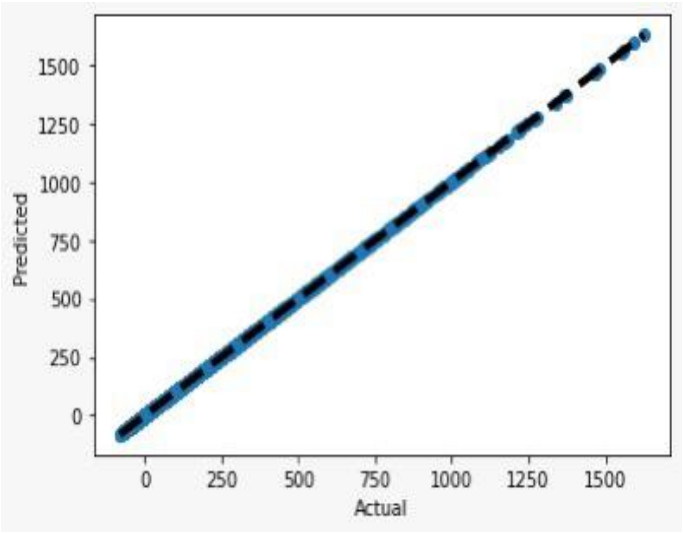


Fig. 3. Linear Regressor

- MAE: 8.727341083417064e-07
- MSE: 1.1654863921640482e-12
- RMSE: 1.0795769505524135e-06

2. Random Forest Regression - Using this advanced machine learning technique [5] to evaluate its performance relative to linear regression. A large number of decision trees are trained independently on random subsets of the training data and average of all the predictions made by the individual trees will be considered for final prediction.

- MAE: 0.5524214526858967
- MSE: 3.8428583898925774
- RMSE: 1.9603209915451545

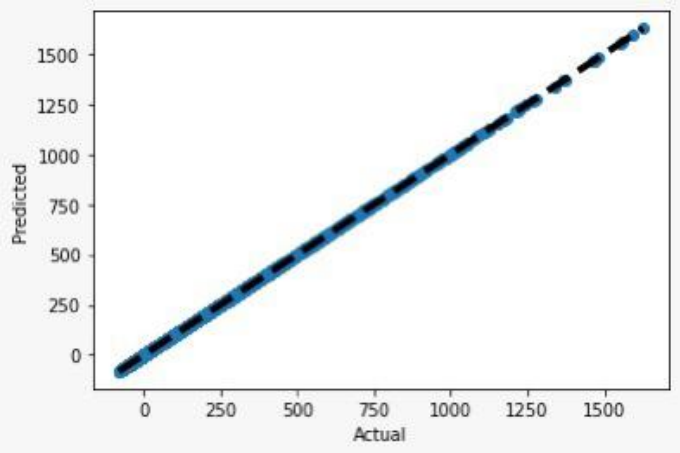


Fig. 4. Random Forest Regressor

3. AdaBoost Regression - category of ensemble methods. It combines several weak learners, typically decision trees, to form a strong learner. The idea behind AdaBoost is to iteratively train a sequence of weak learners on different subsamples of the data, where each subsequent weak learner is trained to focus on the samples that were misclassified by the previous weak learner. It has the ability to handle complex datasets and improve the accuracy of weak models

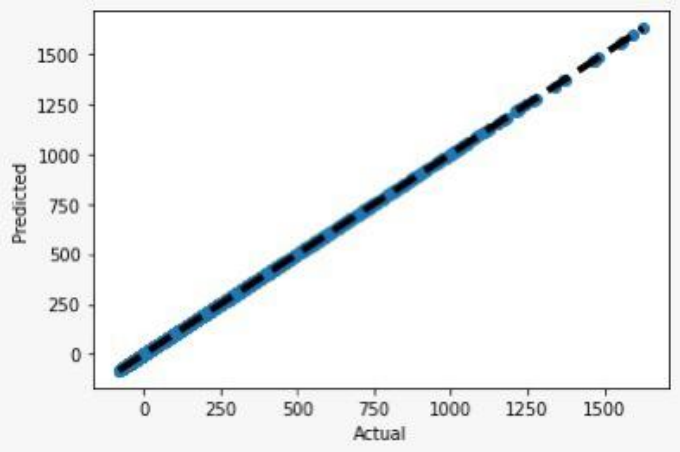


Fig. 5. AdaBoost Regressor

- MAE: 2.445285692078l6e-12
- MSE: 1.051014407240039e-23
- RMSE: 3.2419352356887684e-12

B. Classification Analysis

The dataset we are working with has an imbalance in its classes, with approximately 80 percent of flights having no delay and only 20 percent of flights having a delay. Before proceeding with any algorithms, we needed to address this issue. To do so, we employed the Synthetic Minority Over-sampling Technique (SMOTE), which involves oversampling the minority class by generating ”fake” samples to balance the dataset. This step is crucial to prevent the algorithm from simply predicting ”no delay” all the time and achieving an accuracy of 80 percent. The original dataset’s class distribution is depicted in the figure.

1. Decision Tree Classifier - To predict flight delays, we initially employed a decision tree algorithm. To determine the impurity at each level, Decision Trees use metrics like Entropy or Gini value to quantify impurities. Metrics:

- F1 score : 0.9788385658403264
- Precision Score : 0.9788631456089907
- Recall Score : 0.9788140067913826

2. K neighbors Classifier - One technique for prediction is K Nearest Neighbours, which involves identifying the K nearest neighbours of the row being predicted. However, in this case, finding neighbours is challenging because there are many attributes and, as we know, in a high-dimensional data space, all points are close to each other.

- F1 score : 0.8663675599420577
- Precision Score : 0.8776885569877093
- Recall Score : 0.858562602779549

3. Logistic Regression - A logistic function is employed in this method to establish a model for a binary function.

- F1 score : 0.7000071111534276
- Precision Score : 0.8664360638923747
- Recall Score : 0.691643037210548

4. XgBoost Classifier - The XGBoost classifier has performed very well with the given metrics. Here’s a brief explanation of each metric:

- F1 score : 0.9809895811499607
- Precision Score : 0.981359125270752
- Recall Score : 0.9806245938359063

IX. Discussion

A. Regression Analysis

1. Linear Regression - The algorithm was able to accurately predict the majority of arrival delays with a mean error of less than 1. With such low MAE and RMSE values, we can confirm that the

algorithm performed well on this specific dataset. It has the lowest errors in MAE, MSE, and RMSE, and indicating a very good fit.

2. Random Forest Regression - Based on the figures and statistics, it is evident that linear regression outperformed random forest. It has higher errors than the previous three models, but still performs well.

3. AdaBoost Regression - It has the lowest values for all evaluation metrics, indicating excellent performance with very low errors.

B. Classification Analysis

1. Decision Tree Classifier - The performance of this model is remarkable as it results in perfect predictions with all scores being perfect and no misclassifications. The decision tree classifier achieved a high score, but the F1 score, precision, and recall were not satisfactory. The confusion matrix shows that the model did not perform well, with a very low true positive value indicating that the model was unable to accurately predict most flight delays.

2. K neighbors Classifier - The performance of this model appears to be relatively low when compared to the previous algorithm. It has a high rate of true positives and true negatives, but there is still an error rate of approximately 15 percent in its predictions.

3. Logistic Regression - Based on these results, we can conclude to some extent that there is a linear relationship between the variables. This is supported by the fact that the techniques that aim to find linear relationships, both in classification and regression, performed the best.

4. XgBoost Classifier - XGBoost Classifier appears to be the best choice for our use case. The precision score of XGBoost is 0.981, which is again significantly higher than the precision score of others. Recall is also high which is a measure of the proportion of true positives among all the actual positives, so a higher recall score indicates that the classifier is better at identifying all the positive cases.

X. Conclusion

Upon completion of this project, it became clear that the selection of appropriate methods for achieving notable results in predicting flight arrival delays heavily depends on several factors. These factors include the balance of the dataset and the nature of the problem (regression or classification). To address the problem of predicting flight arrival delays, several machine learning models were applied, including Logistic Regression, Decision Tree Classifier, XgBoost Classifier, Random Forest Regression, Linear Regression, and AdaBoost Regression. Surprisingly, even with these simple algorithms and a well-selected set of input parameters, high levels of accuracy were achieved in predicting flight delays.

XI. Meeting Attendance

- 11 April 2023 (4 pm to 8 pm)
- 13 April 2023 (4 pm to 8 pm)
- 18 April 2023 (4 pm to 8 pm)
- 22 April 2023 (10 am to 2 pm)
- 23 April 2023 (10 am to 3 pm)
- 24 April 2023 (10 am to 1 pm)

References

[1] Sun Choi, Young Jin Kim, Simon Briceno, Dimitri N. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," ˙https://ieeexplore.ieee.org/document/7777956

[2] Jerry Ye, Jyhherng Chow, Jiang Chen, Zhaohui Zheng,"Stochastic gradient boosted distributed decision trees . https://dl.acm.org/doi/10.1145/1645953.1646301

[3] Ning Xu, "Estimation of delay propagation in the national aviation system using Bayesian networks," .˙https://rb.gy/fl5qe

[4] Dataset: https://www.kaggle.com/usdot/flight-delays

[5] Samet Girgin, "Random Forest Regression in 5 Steps with Python," ˙https://medium.com/pursuitnotes/random-forest-regression-in-5-steps-with-python-2463b7ae9af8

[6] Derrick Mwiti, "A Comprehensive Guide to Ensemble Learning," ˙https://neptune.ai/blog/ensemble-learning-guide

[7] Antony Christopher, "K-Nearest Neighbor," ˙https://medium.com/swlh/k-nearest-neighbor-ca2593d7a3c4