Introduction To Machine Learning and Algorithms

Course Project : Flight Arrival Predictions

Abishek Umesh Gavali
Syracuse University
augavali@syr.edu
SUID : 302921506

Akshay Kumar Huliyar Prabhakara
Syracuse University
ahuliyar@syr.edu
SUID : 238301680

Shripad Sunil Mathekar
Syracuse University
ssmathek@syr.edu
SUID : 782755419

Tejas Manjunatha Devang
Syracuse University
tdevang@syr.edu
SUID : 593038108

ABSTRACT:

In this project, a machine learning model was developed to predict the arrival times of flights flying into Syracuse. A variety of data, such as the scheduled and actual flight times from the past, was used to aid these predictions. The Random Forest Classifier model was chosen because it handles complex data well and avoids overfitting, which means it doesn't just memorize the data but learns from it to make better predictions. The model was tested using cross-validation, which helps ensure that it works well not just on the data it was trained on but also on new, unseen data. The aim was to create a reliable tool that airlines and travelers can use to predict flight times more accurately, helping users to plan their trips better.

INTRODUCTION:

The purpose of this project was to develop predictive models to determine the arrival status of flights coming to Syracuse Hancock International Airport (SYR) in New York. The focus was on predicting that each flight would arrive early, on time, or late. To achieve this objective, pre-flight information such as scheduled and actual arrival times, along with weather conditions upon arrival, was utilized. The resulting model is intended to assist airlines and travelers in managing policies and making informed travel decisions.

DATA COLLECTION AND PREPROCESSING:

Data collected from the Bureau of Transportation Statistics website for the years 2019 – 2023 was utilized. The dataset includes historical aviation data, featuring actual projected times. Two datasets were collected, one for earlier flights and the other for latter flights. In preparing the flight data for predictive modelling, several essential steps were taken to ensure accuracy and effectiveness. Initially, necessary libraries were imported, and settings were adjusted to aid data handling and visualization. The raw data was loaded from CSV files and filtered to include specific airlines and airports, with carrier codes standardized for consistency. To get the data ready for modelling, columns with missing values were removed, and new features were created from date and time fields, such as extracting month, day, year, and converting 'Day of the Week' into a numerical format.

Additional cleaning involved standardizing various time formats into a consistent datetime format and dropping columns that wouldn’t be available at prediction time to avoid data leakage. Weather data was also merged based on day, month, and year to include factors like temperature and precipitation, which could impact flight delays. Categorical variables such as 'Carrier Code' and 'Origin Airport' were converted using one-hot encoding to make them suitable for use in machine learning models, and time data were transformed into minutes past midnight for uniformity.

The dataset was further refined by eliminating unnecessary columns and examining correlations to pinpoint the most impactful features. This thorough preprocessing ensures that the data fed into the machine learning models is clean, relevant, and properly formatted, which is crucial for enhancing model performance and reliability.

FEATURE SELECTION:

Before conducting feature selection, a thorough examination of relationships and distributions within the data was performed using various graphs and a correlation matrix. The correlation matrix is a crucial tool that quantifies the degree to which different variables are related. It provides insights into potential multicollinearity issues and helps identify which features might have significant predictive power regarding the target variable, 'Arrival Delay (Minutes)'. This matrix was visualized using a heatmap, enhancing the interpretability of the relationships by providing a color-coded representation of the correlation coefficients. Figure 1 shows the correlation matrix of different features with respect to Arrival Delay.

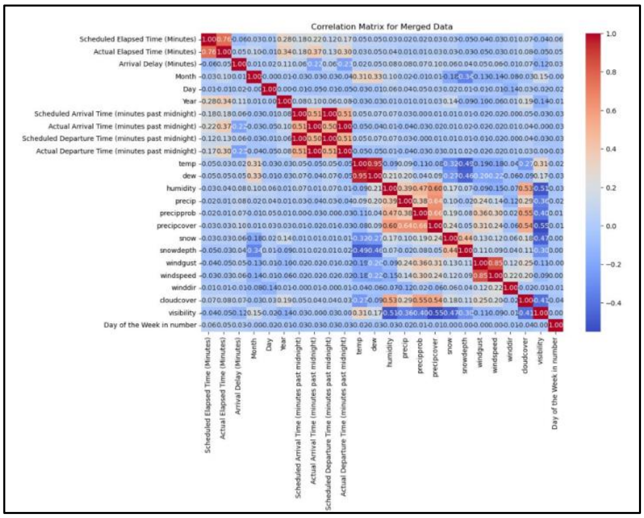

Fig. 1: Correlation matrix

The correlation matrix in figure 1 illustrates the relationships between flight and weather-related variables using a color-coded display. Strong positive correlations, shown in red, are evident between 'Scheduled Elapsed Time' and 'Actual Elapsed Time', as well as between 'Scheduled Arrival Time' and 'Actual Arrival Time', indicating that flights generally adhere to their schedules. Weather-related features like temperature, dew point, and humidity also show interrelated moderate to strong correlations. This matrix helps identify key relationships that could be pivotal for further detailed analysis and predictive modelling.

Additionally, to understand how 'Arrival Delay' interacts with other features individually, scatter plots were generated. These plots are valuable for visualizing the relationship between 'Arrival Delay' and continuous or ordinal variables, helping to identify trends, outliers, or peculiar patterns in the data. For each feature like 'Month', 'Day', 'precipitation levels', 'wind speed', and others, a scatter plot illustrated how changes in these variables might influence flight delays. The following scatter plots show the relationship between Arrival delay and various features.

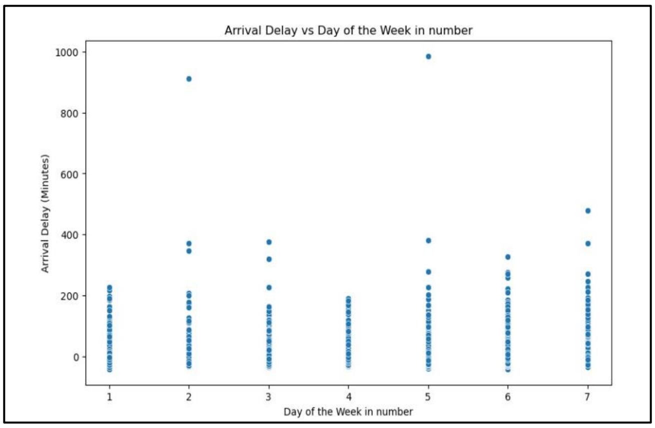

Fig. 2: Arrival delay vs Day of the week

Figure 2 evaluates the connection between aircraft delays and the day of the week. It aids in determining whether there are more delays on some days (such as weekends or weekdays) than on others.

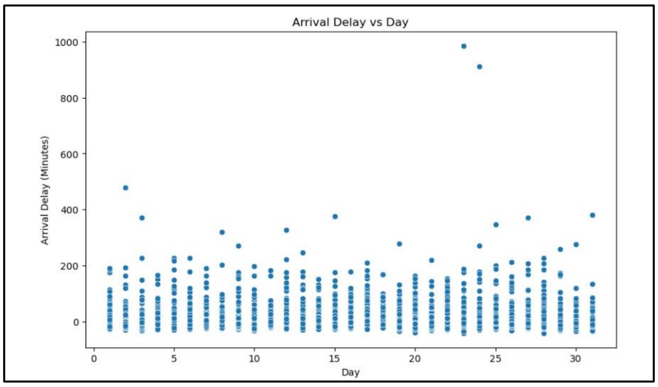

Fig. 3: Arrival delay vs date

Figure 3 shows a scatter plot of flight delays against dates is displayed in this graph. It could show trends of lateness at periods of the month or year, such as the holidays.

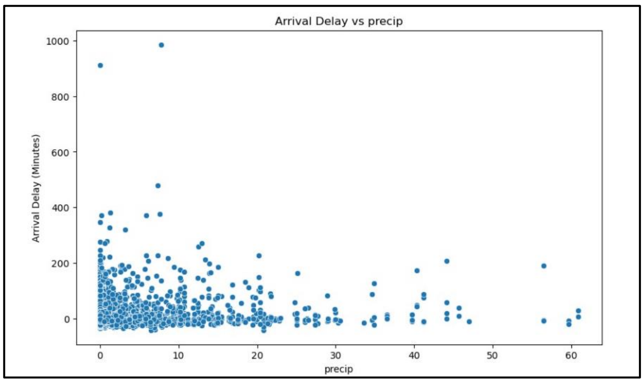

Fig. 4: Arrival delay vs precipitation

Figure 4 shows how variations in the amount of precipitation could impact flight delays. It investigates the possibility that lengthier delays are related to more precipitation.

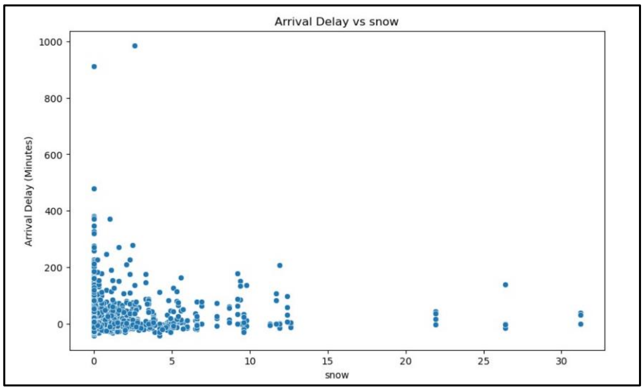

Fig. 5: Arrival delay vs snow

Figure 5: Arrival Delay vs Snow This scatter figure explains how snow affects flight delays, much as the precipitation graph. It focuses on whether snowfall causes major delays.

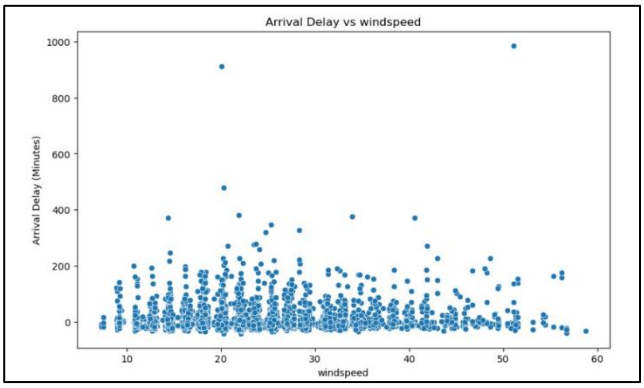

Fig. 6: Arrival delay vs windspeed

The figure 6 below shows how wind speed affects flight arrival timings and looks at the relationship between wind speed and delays.

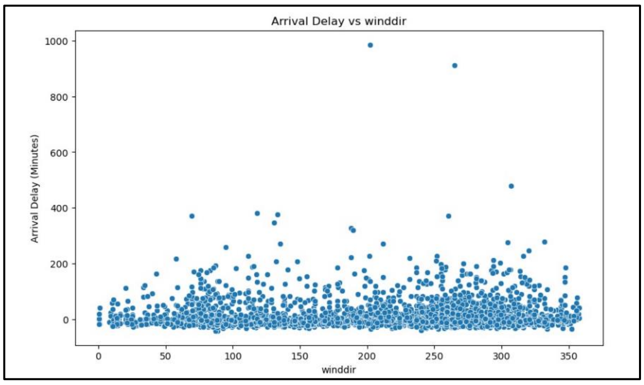

Fig. 7: Arrival delay vs winddir

Figure 7 shows Arrival Delay vs. Wind Direction: This graph examines if wind direction affects flight delays, revealing whether some wind directions are more difficult for aircraft.

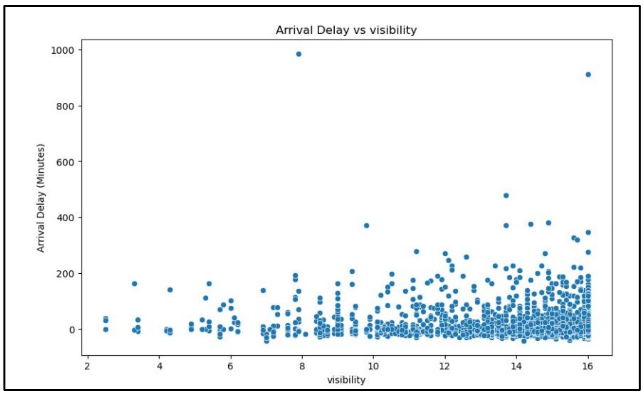

Fig. 8: Arrival delay vs visibility

Figure 8 evaluates the relationship between visibility and aircraft delays. It helps in determining whether poor sight typically causes delays that are more severe.

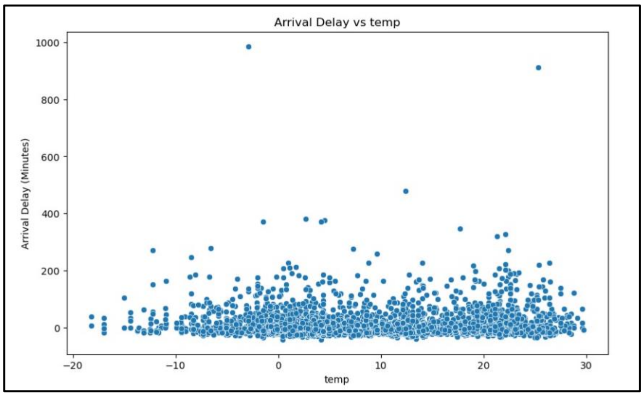

Fig. 9: Arrival delay vs temp

This graph in figure 9 examines the relationship between temperature changes and flight delays, to see if higher temperatures correlate with longer delays.

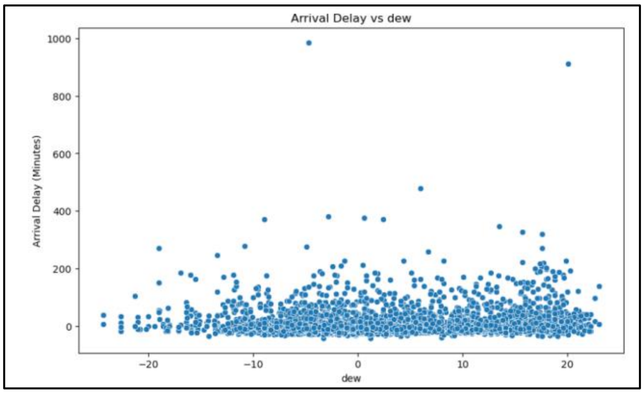

Fig. 10: Arrival delay vs dew

The scatter plot in figure 10 of airplane arrival delays versus dew point temperatures, which range from -20 to 20 degrees, is shown.

Figure 11 shown below represents the scatter plot of Arrival delay vs humidity. All humidity levels show that the data largely clusters at lower delay times.

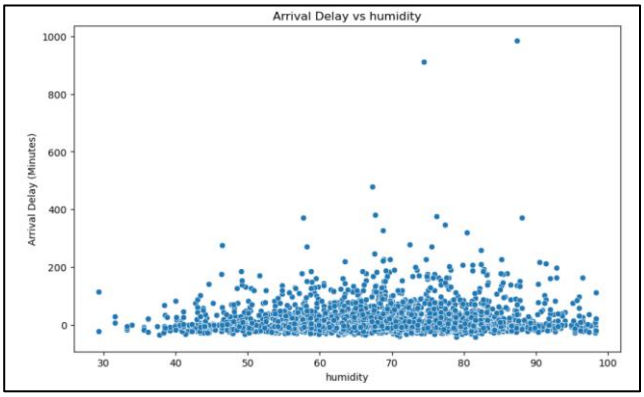

Fig. 11: Arrival delay vs humidity

Following the exploratory data analysis, feature selection was conducted to refine the model's input data. This process involved choosing the most relevant predictors for the model based on their correlation with 'Arrival Delay' and their contribution to model accuracy. Features with little to no correlation or redundant features exhibiting high multicollinearity were considered for removal. This step is critical in enhancing model performance by reducing complexity and focusing on the most informative attributes, ensuring that the final model is both efficient and robust.

METHODOLOGY:

In the approach to this work, the focus was on developing reliable machine learning models to predict flight arrival times using a systematic and comprehensive method. The process began with data preparation, which involved selecting relevant parts of the data, addressing missing values, and creating additional features that aid in accurate forecasting. This preparation is essential for extracting information such as month, day, and year from date lines. The Random Forest Classifier was chosen due to its proficiency in handling complex, nonlinear data and its resistance to overfitting. This framework is particularly suitable as it performs well with various datasets by combining different types of decision trees and their results to obtain more accurate predictions and a deeper understanding of the data.

To enhance the performance of the model on both the training data and unseen data, cross-validation was employed. This method divides the data into several segments, training the model on some segments and testing it on others, which aids in validating the model's performance across different data blocks. Additionally, emphasis was placed on feature engineering, where further adjustments were made to the existing data to derive deeper insights about flight time and status, factors that can influence delay times. During the model training phase, a detailed examination of the factors that most significantly affect the model's predictions was conducted. This scrutiny allowed for further refinement of the model by concentrating on the variables with the most substantial impact, thus enhancing the model's simplicity and effectiveness. The approach continued to be refined based on insights obtained from the model's performance and material analysis.

MODEL BUILDING AND EVALUATION:

In the model building phase, the Random Forest Classifier from the scikit-learn library was chosen primarily for its robustness against overfitting and its capability to effectively manage non-linear data. This decision was strategic, given the project's focus on complex, multifaceted data involving various predictors of flight delays.

Measurement efforts were concentrated on accuracy, which measures the proportion of correctly predicted statuses relative to the entire dataset. The dataset was divided into two parts: Earlier and Latter flights, and two models were built. The models achieved reasonable average cross-validation scores of 51.61% and 57.29%, respectively, demonstrating their proficient predictive abilities. This metric is particularly significant as it confirms the model's effective generalization, indicating that it can perform consistently across a variety of unknown datasets and real-world scenarios. This thorough review approach enables the identification of areas where the model performs well and where improvements are needed.

Accuracy Table:

Average Cross-Validation Score	57.29
Training Accuracy	78.38
Testing Accuracy	58.66

Average Cross-Validation Score	51.61
Training Accuracy	79.49
Testing Accuracy	51.74

FUTURE SCOPE:

Introducing new data sources, such as social media posts that can affect flight times, can enhance forecasting by incorporating multiple data sources. Utilizing real-time data processing allows the model to adjust forecasts based on the latest available information, enhancing accuracy and responsiveness. Additionally, the adoption of advanced machine learning or other artificial intelligence techniques can further improve the accuracy and efficiency of predictions. While currently focused on specific locations, the model can be adapted to predict flight times for other airports around the world, expanding its applicability and utility.

CONCLUSION:

In conclusion, the project developed a machine learning algorithm using Random Forest Classifier that predicts flight arrival times with reasonable accuracy. This model handles complex data effectively and prevents overfitting, making it applicable in real-world scenarios. Cross-validation confirmed the model's effectiveness on both known and new data. By homing in on critical factors affecting flight delays, the model now delivers reliable predictions. This tool aids airlines and passengers by offering detailed flight status information for informed decision-making. Looking ahead, there is potential to enhance accuracy further and broaden its application scope to boost overall flight performance.

REFERENCES:

1] https://www.transtats.bts.gov/ontime/

2] https://www.visualcrossing.com/weather-history/