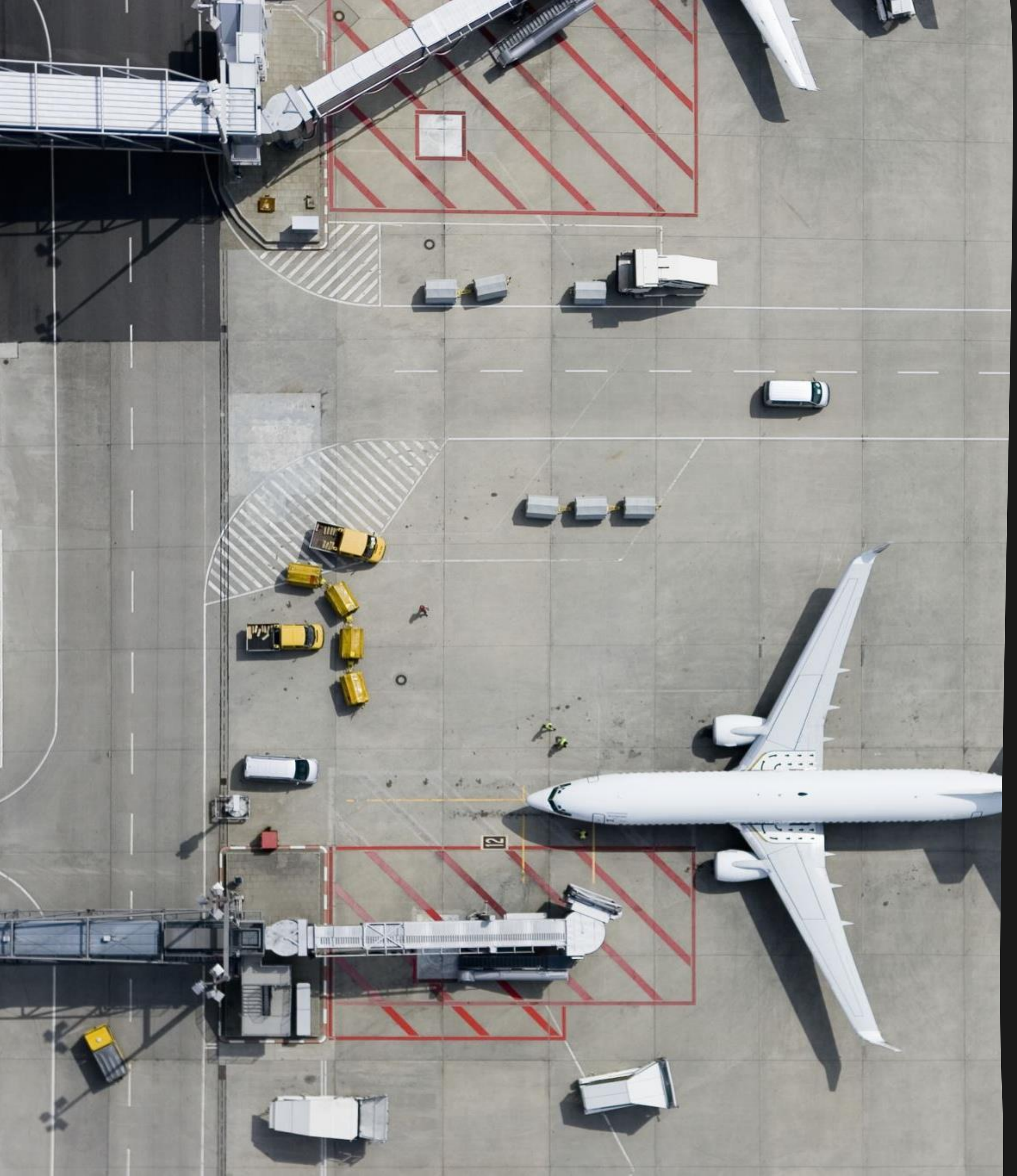


Predicting Airport Delays

By Jonathan Hooper,
Evan Burch, and Anurag Dwivedi

Problem Statement

- As global population grows and time becomes more precious, air travel is essential for long-distance mobility. The IATA predicts 4.7 billion air passengers in 2024, up from 4.5 billion in 2019. Yet, flight delays remain a significant challenge. According to the Bureau of Transportation and the FAA, flight delays increased from ~19% to ~21% in 2023, costing over $30 billion annually.

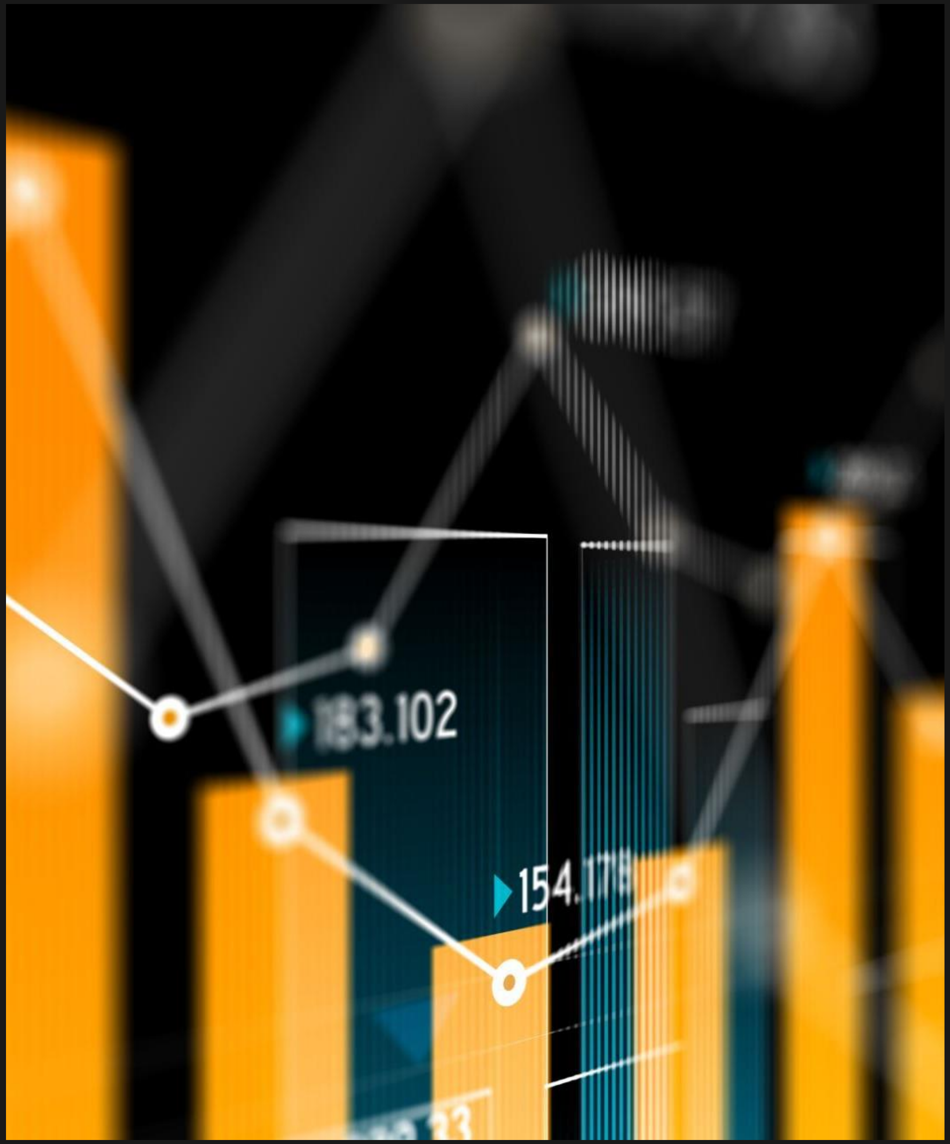
Related Works

- Meel P, et. al [3] designed 5 models to predict flight delay based on machine learning models such as Logistic Regression, Decision Tree Regression, Bayesian Ridge, Regression and Gradient Boosting Regression.
  - Domestic flights in 2015
  - Strength: utilization of data visualizations
  - Weakness: meteorological statistics

- Chakrabarty, Navoneel, et al [4] proposed a machine learning model using Gradient Boosting Classifier for predicting flight arrival delay in 2019.
  - Strength: handling imbalance of dataset

Data Set

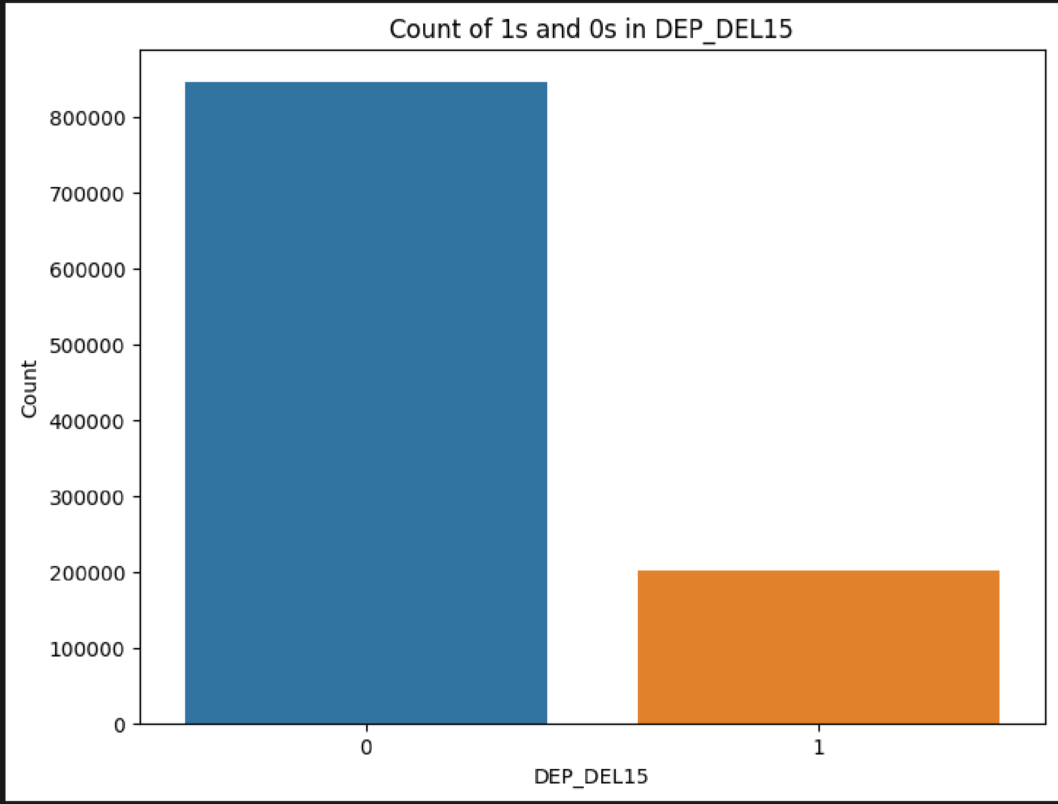
- This dataset is a comprehensive collection of flight-related information for the year 2019, encompassing a wide array of attributes that describe various aspects of flight operations.
- The dataset consists of 26 attributes, offering a balanced blend of different data types.



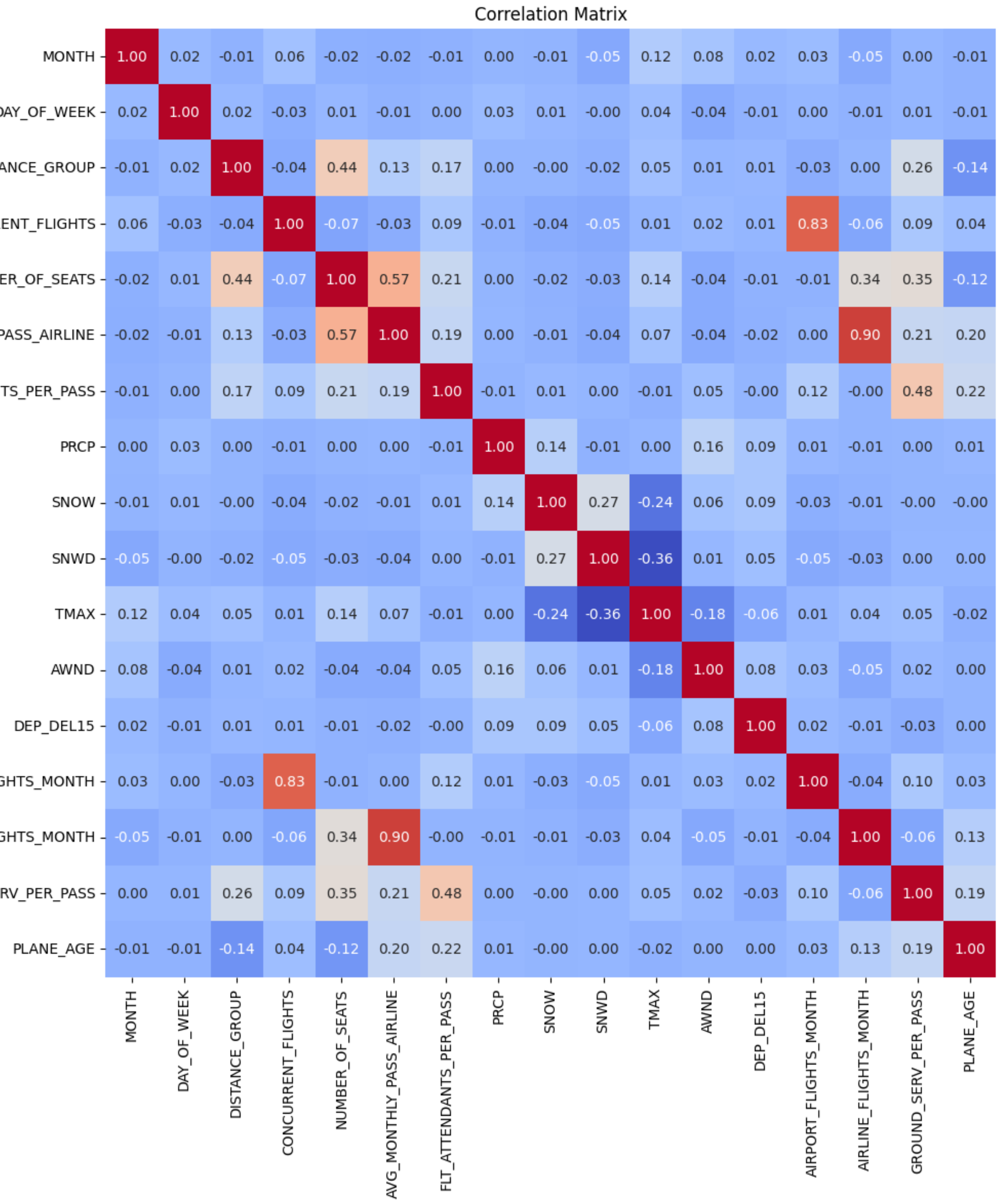
Data Set Features

Month	Month
DAY_OF_WEEK	Day of Week
DEP_DEL15	TARGET Binary of a departure delay over 15 minutes (1 means there was a delay)
DEP_TIME_BLK	Departure time block
DISTANCE_GROUP	Distance group to be flown by departing aircraft
SEGMENT_NUMBER	The segment that this tail number is on for the day
CONCURRENT_FLIGHTS	Concurrent flights leaving from the airport in the same departure block
NUMBER_OF_SEATS	Number of seats on the aircraft
CARRIER_NAME	Carrier
AIRPORT_FLIGHTS_MON	Avg Airport Flights per Month
AIRLINE_FLIGHTS_MON	Avg Airline Flights per Month
AIRLINE_AIRPORT_FLIGHTS_MONTH	Avg Flights per month for Airline AND Airport
AVG_MONTHLY_PASS_AIRPORT	Avg Passengers for the departing airport for the month
AVG_MONTHLY_PASS_AIRLINE	Avg Passengers for airline for month
FLT_ATTENDANTS_PER_PASS	Flight attendants per passenger for airline
GROUND_SERV_PER_PASS	Ground service employees (service desk) per passenger for airline
PLANE_AGE	Age of departing aircraft
DEPARTING_AIRPORT	Departing airport
LATITUDE	Latitude of departing airport
LONGITUDE	Longitude of departing airport
PREVIOUS_AIRPORT	Previous airport that aircraft departed from
PRCP	Inches of precipitation for day
SNOW	Inches of snowfall for day
SNWD	Inches of snow on ground for day
TMAX	Max temperature for day
AWND	Max wind speed for day

How many Flights Delayed



0 = On-time **1 = Delayed**



Data Collection/Analysis

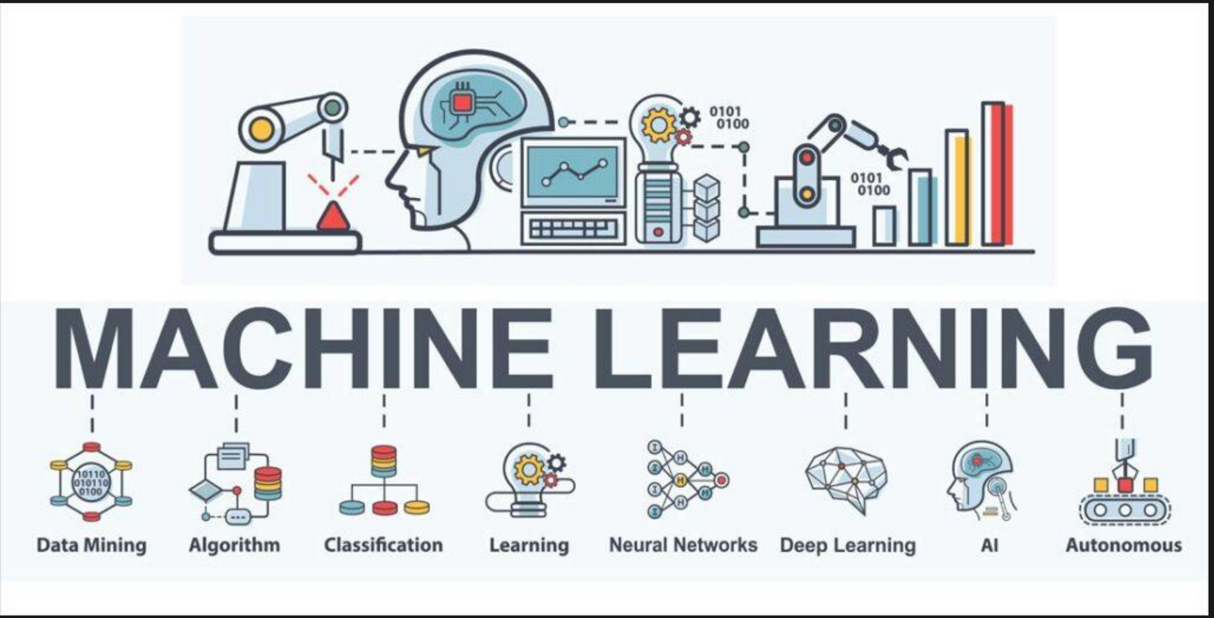
- The dataset used in our analysis was acquired from Kaggle [5]; the data was acquired and provided by Bureau of Transportation and NOAA
- In our data analysis process, we focused primarily on data visualization, cleaning, and wrangling methods to prepare the dataset for exploration.
- Utilization of box plots and histograms for data visualization, gaining insights into distribution and tendencies.

Data Preprocessing

- Preprocessing steps:
  - Fill in missing values
  - Outlier Analysis
  - Feature Engineering
    - Used LabelEncoder(): Converting Categorical labels into numerical variables

Machine Learning Models

- Logistic Regression: This is a fundamental algorithm for binary classification problems.
- K-Nearest Neighbors (KNN): KNN is a supervised machine learning algorithm used to solve classification and regression problems.
- Decision Tree Classifier: This algorithm uses a tree-like model (set of rules) to make decisions; similar to how humans make decisions.



Logistic Regression

- Target feature imbalance likely causing poor F1 score
- Attempted to use RSMOTE to solve this

Accuracy Measures

accuracy: 80.83%

f1_score: 0.03

roc_auc_score: 50.60%

K-Nearest Neighbors (KNN)

- Hyper-Parameters via GridSearchCV
- n_neighbors: [3, 5, 7, 9],
- weights: ['uniform', 'distance'],
- metric: ['euclidean', 'manhattan', 'minkowski']

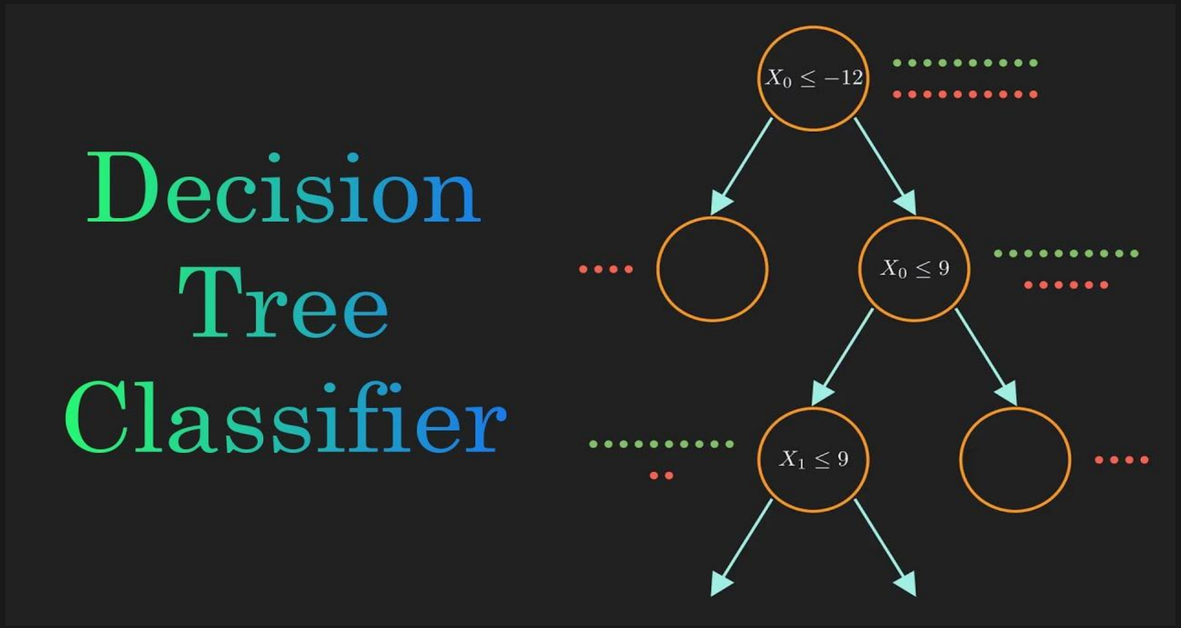
Accuracy Measures

accuracy: 80.06%

f1_score: 0.31

roc_auc_score: 58.60%

Decision Tree Classifier



Hyper-Parameters via GridSearchCV

- random_state: range(100)

Accuracy Measures

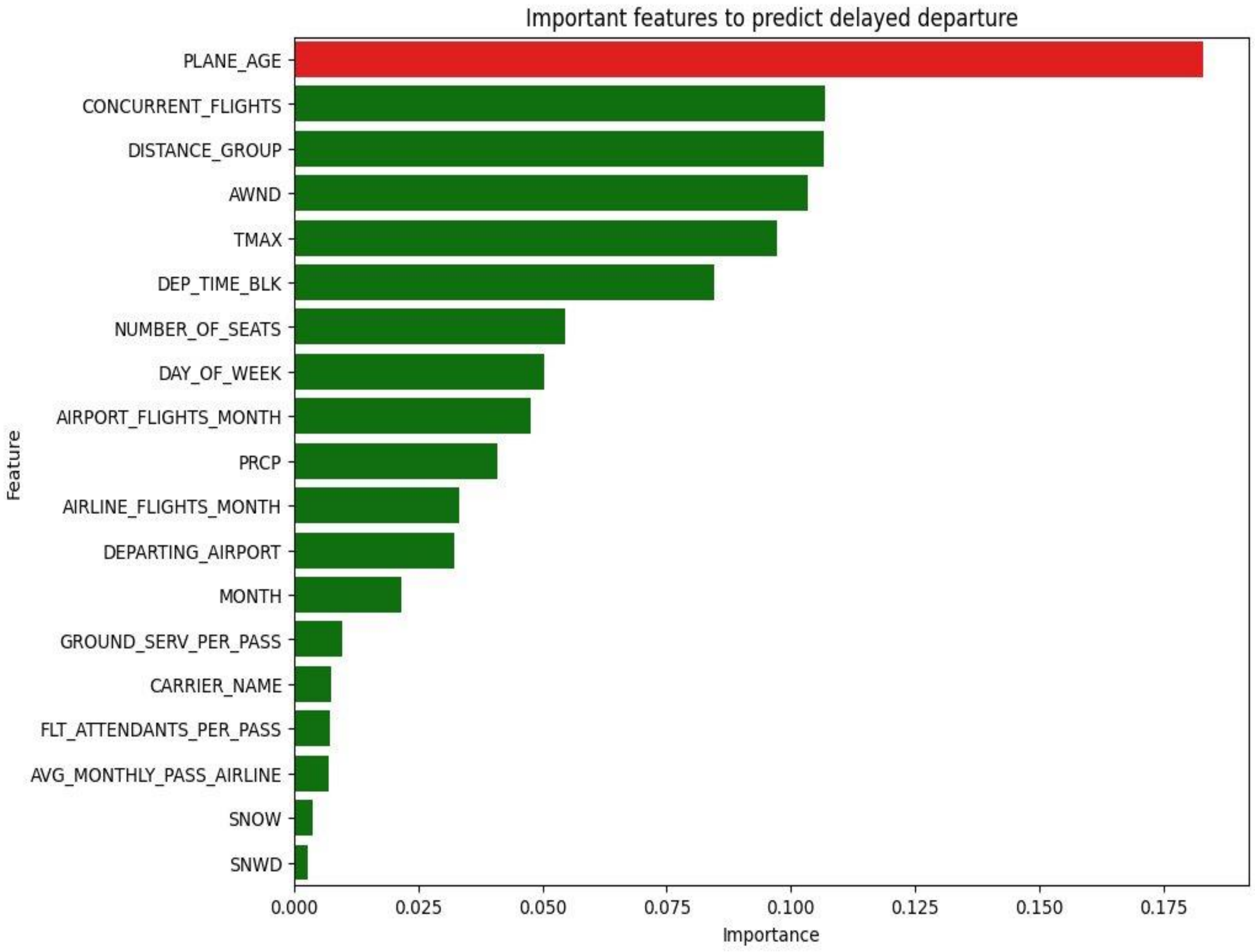
accuracy: 76.80%

f1_score: 0.32

roc_auc_score: 58.60%

Results

Machine Learning Algorithms	Test Accuracy	F1 Score	ROC Accuracy
Logistic Regression	80.83%	0.03	50.60%
KNN	80.06%	0.31	58.60%
Decision Tree Classifier	76.80%	0.32	58.60%



Feature Importance

Conclusion

- We are able to build a few machine learning models to accurately predict whether a flight will be delayed or not. By leveraging our extensive dataset of flight schedules, weather, airport operations, and passenger trends, these models discern key delay factors and enhance forecasting accuracy, aiding stakeholders in preempting and addressing disruptions. Such advancements improve reliability for passengers, optimize airline operations, and enhance resource utilization for airports, fostering a more sustainable and economically beneficial airline industry.

Future Improvements

- More in-depth data visualizations

- *Different Models:* Try more complex models such as Random Forests, Gradient Boosting Machines, or Neural Networks which might be better at handling complex patterns.

- *Algorithmic Approaches:* Use algorithms specifically designed to handle imbalanced datasets, such as SMOTE for over-sampling or ensemble methods like Balanced Random Forest.

- *Evaluation Metrics:* Since the dataset is imbalanced, metrics like f1-score, precision-recall curve, and AUC-PR might be more appropriate for evaluating model performance than accuracy.

References

- [1] Dooley, R. (2024, February 20). Survey predicts air travel boom for 2024: What it means for passengers. Forbes. https://www.forbes.com/sites/rogerdooley/2023/12/06/air-travel-boom-predicted-for-2024/?sh=4c20537fabf7
- [2] OST_R: BTS: Transtats. BTS. (n.d.). https://www.transtats.bts.gov/homedrillchart.asp
- [3] P. Meel, M. Singhal, M. Tanwar and N. Saini, "Predicting Flight Delays with Error Calculation using Machine Learned Classifiers," 2020 7th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2020, pp. 71-76, doi: 10.1109/SPIN48934.2020.9071159.
- [4] Esmaeilzadeh, E., & Mokhtarimousavi, S. (2020). Machine Learning Approach for Flight Departure Delay Prediction and Analysis. Transportation Research Record, 2674(8), 145-159. https://doi.org/10.1177/0361198120930014
- [5] Wadkins, J. (2022, January 17). 2019 airline delays w/weather and airport detail. Kaggle. https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations/data