Flight Delay Prediction

Optimizing Flight Times from O'hare by Predicting Flight Delays

CS 418- Introduction to Data Science, Professor Brian Ziebart

Sohum Bhole, Mohammad Nusairat, Nahom Yohanes, Richa Rameshkrishna, Vageesh Indukuri

01

Introduction

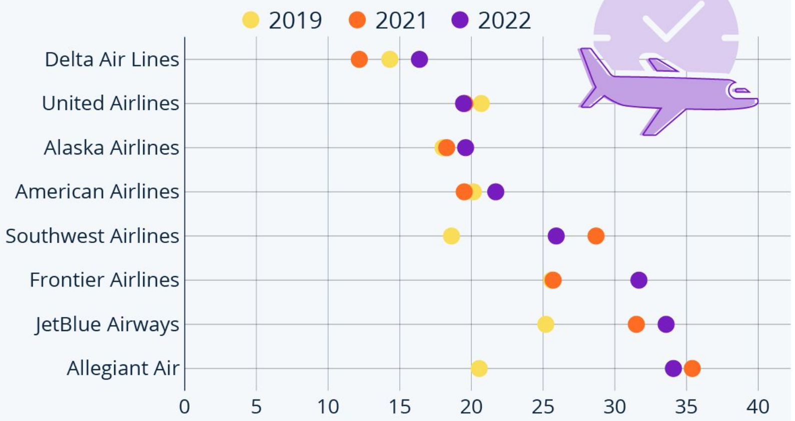
Abstract

A Brief Introduction of Our Project

Flight delays pose significant challenges to the aviation industry globally. This project leverages machine learning to analyze demographic and operational data to predict flight delays and provide actionable insights for improving scheduling efficiency, resource allocation, and overall punctuality in air travel.

Delayed Flights – the New Post-Pandemic Normal?

Share of late arrivals by North American airlines (in percent of all flights)



No report for 2020. Excludes flights by partner airlines. Where data was available
Source: Cirium On-Time Performance Review

statista

Project Introduction

The problem we aim to solve and its relevance to machine learning

This project develops a machine learning-based model to predict flight delays by analyzing operational and scheduling factors such as departure times, carrier performance, and weather conditions. By leveraging techniques like Gradient Boosting and Random Forest, the model identifies critical predictors and offers actionable insights to optimize flight schedules, enhance resource allocation, and minimize delays. These insights aim to improve operational efficiency and passenger satisfaction within the aviation industry.

Index Terms: Flight Delays, Prediction, Operational Efficiency, Machine Learning, Gradient Boosting, Random Forest, Aviation Analytics, Scheduling Optimization, Resource Allocation, Data-Driven Insights, Predictive Modeling.

02

Dataset

Dataset Description and Relevant Pre-processing Conducted

Describing the dataset used, including its source and key characteristics, as well as data preprocessing steps

Dataset Source & Key Characteristics:

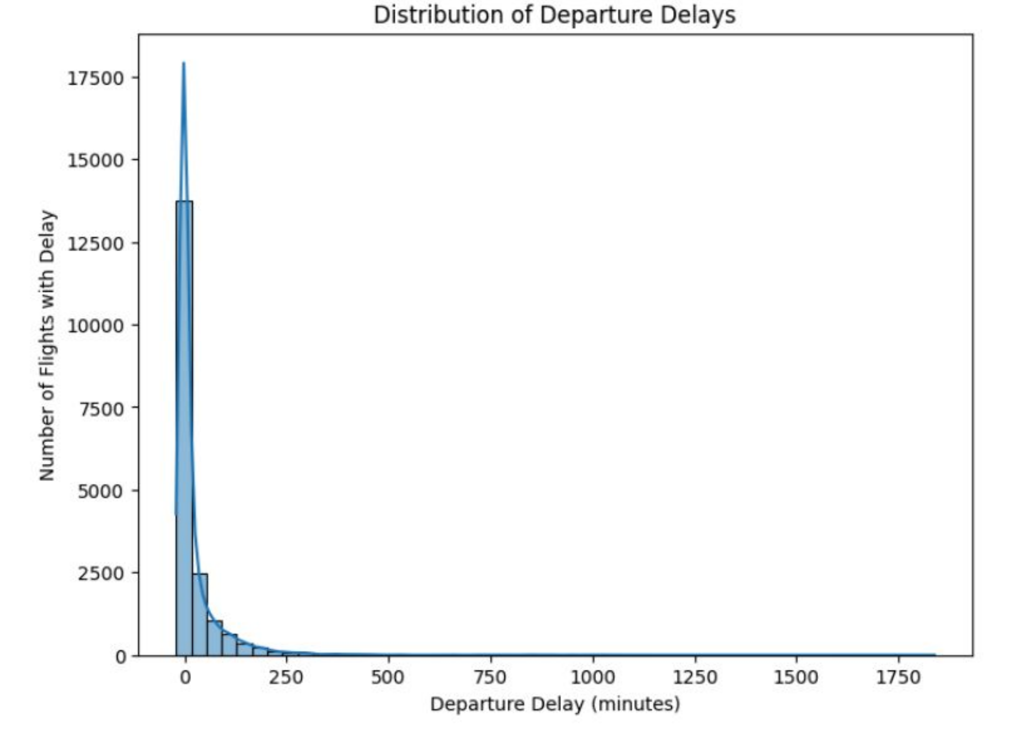
- Source: U.S. Department of Transportation (Bureau of Transportation Statistics)
- The dataset includes features such as ORIGIN, DEST, CRS_DEP_TIME, DISTANCE, CARRIER_DELAY, WEATHER_DELAY, NAS_DELAY, SECURITY_DELAY, and LATE_AIRCRAFT_DELAY.
- Dimensionality: 18,901 rows & 9 primary features

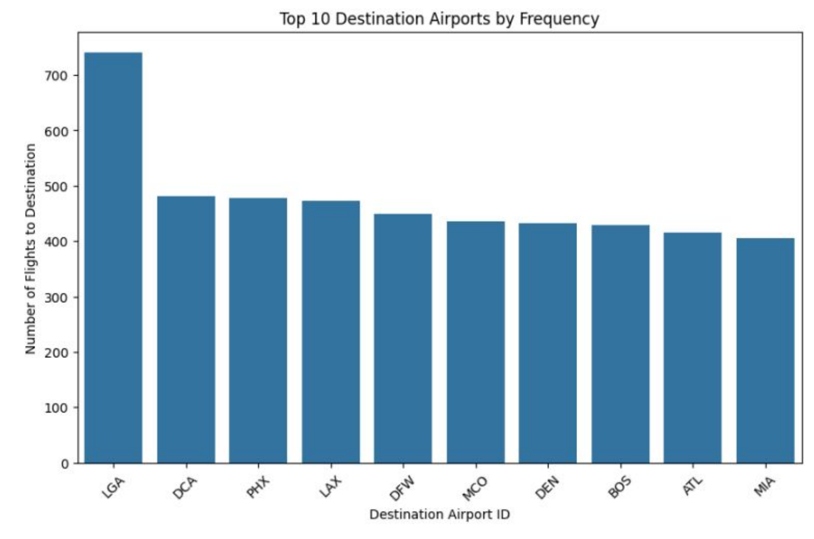
Data Preprocessing Steps:

- Encoded categorical variables (e.g., origin and destination airports).
- Removed duplicates and irrelevant records (e.g., canceled flights).
- Normalized numerical features using StandardScaler.
- Imputed missing values in delay-related columns.
- Created a binary target column Delay to classify flights as on-time or delayed.

Dataset Description and Relevant Pre-processing Conducted

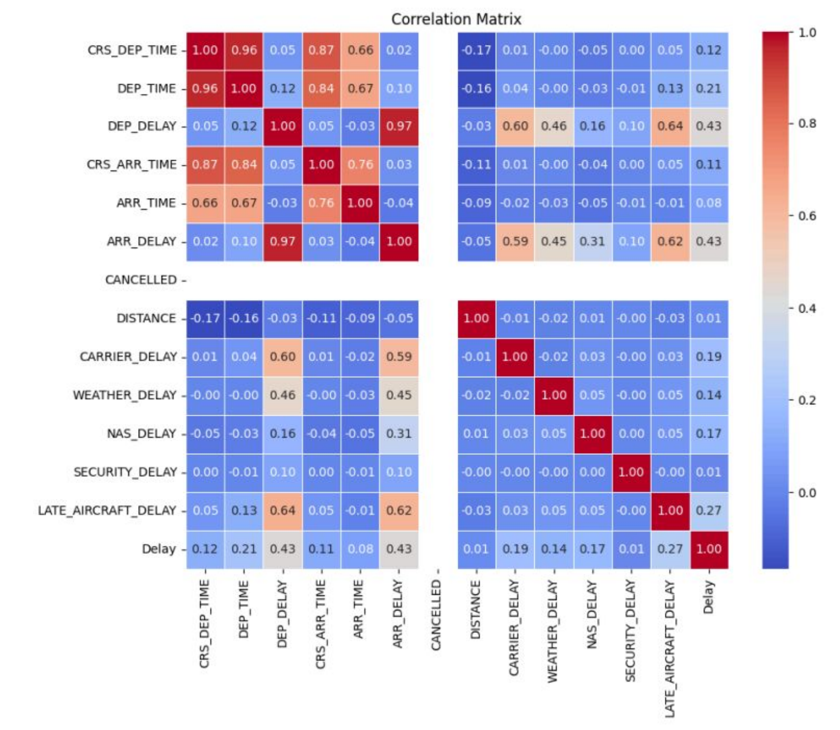
Describing the dataset used, including its source and key characteristics, as well as data preprocessing steps





Dataset Description and Relevant Pre-processing Conducted

Describing the dataset used, including its source and key characteristics, as well as data preprocessing steps



Correlation Matrix

	CRS_DEP_TIME	DEP_TIME	DEP_DELAY	CRS_ARR_TIME	ARR_TIME	ARR_DELAY	CANCELLED	DISTANCE	CARRIER_DELAY	WEATHER_DELAY	NAS_DELAY	SECURITY_DELAY	LATE_AIRCRAFT_DELAY	Delay
CRS_DEP_TIME	1.00	0.96	0.05	0.87	0.66	0.02		-0.17	0.01	-0.00	-0.05	0.00	0.05	0.12
DEP_TIME	0.96	1.00	0.12	0.84	0.67	0.10		-0.16	0.04	-0.00	-0.03	-0.01	0.13	0.21
DEP_DELAY	0.05	0.12	1.00	0.05	-0.03	0.97		-0.03	0.60	0.46	0.16	0.10	0.64	0.43
CRS_ARR_TIME	0.87	0.84	0.05	1.00	0.76	0.03		-0.11	0.01	-0.00	-0.04	0.00	0.05	0.11
ARR_TIME	0.66	0.67	-0.03	0.76	1.00	-0.04		-0.09	-0.02	-0.03	-0.05	-0.01	-0.01	0.08
ARR_DELAY	0.02	0.10	0.97	0.03	-0.04	1.00		-0.05	0.59	0.45	0.31	0.10	0.62	0.43
CANCELLED														
DISTANCE	-0.17	-0.16	-0.03	-0.11	-0.09	-0.05		1.00	-0.01	-0.02	0.01	-0.00	-0.03	0.01
CARRIER_DELAY	0.01	0.04	0.60	0.01	-0.02	0.59		-0.01	1.00	-0.02	0.03	-0.00	0.03	0.19
WEATHER_DELAY	-0.00	-0.00	0.46	-0.00	-0.03	0.45		-0.02	-0.02	1.00	0.05	-0.00	0.05	0.14
NAS_DELAY	-0.05	-0.03	0.16	-0.04	-0.05	0.31		0.01	0.03	0.05	1.00	0.00	0.05	0.17
SECURITY_DELAY	0.00	-0.01	0.10	0.00	-0.01	0.10		-0.00	-0.00	-0.00	0.00	1.00	-0.00	0.01
LATE_AIRCRAFT_DELAY	0.05	0.13	0.64	0.05	-0.01	0.62		-0.03	0.03	0.05	0.05	-0.00	1.00	0.27
Delay	0.12	0.21	0.43	0.11	0.08	0.43		0.01	0.19	0.14	0.17	0.01	0.27	1.00

03

Methodology

Methodology

Our chosen approaches to predictive modeling of the dataset

Random Forest

Effective for handling large, complex datasets and capturing nonlinear relationships between features like carrier delays and late aircraft delays.

Encoded categorical variables and scaled numerical features to optimize model performance.

Logistic Regression

Handles high-dimensional data efficiently and provides interpretable results for binary classification of on-time vs. delayed flights.

Applied normalization to numerical features and one-hot encoding to categorical variables.

Gradient Boosting

Combines iterative learning techniques to improve prediction accuracy for flight delays.

K-Nearest Neighbors (KNN)

Simpler, instance-based learning approach, effective for smaller subsets of flight data.

Found the optimal number of neighbors (k) through parameter tuning and used scaled features for accurate distance-based classification.

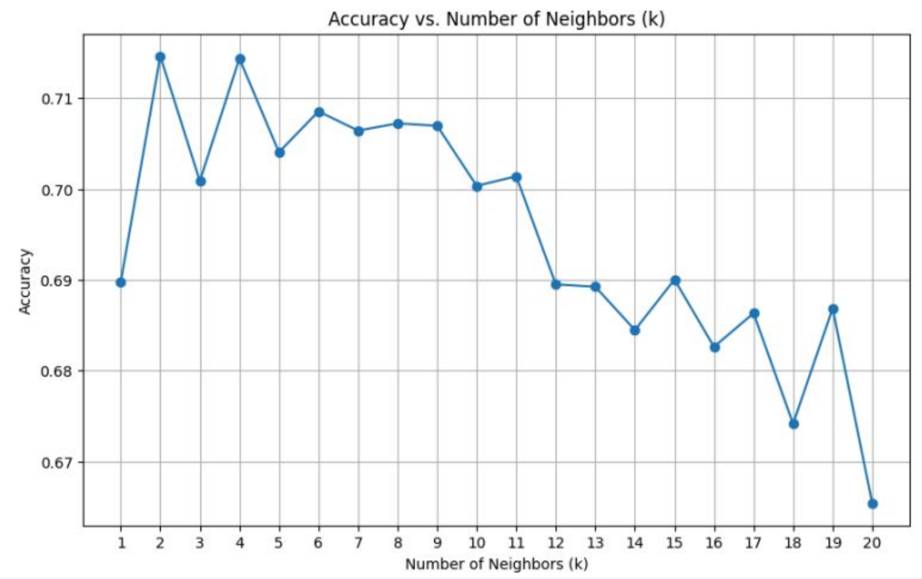
04

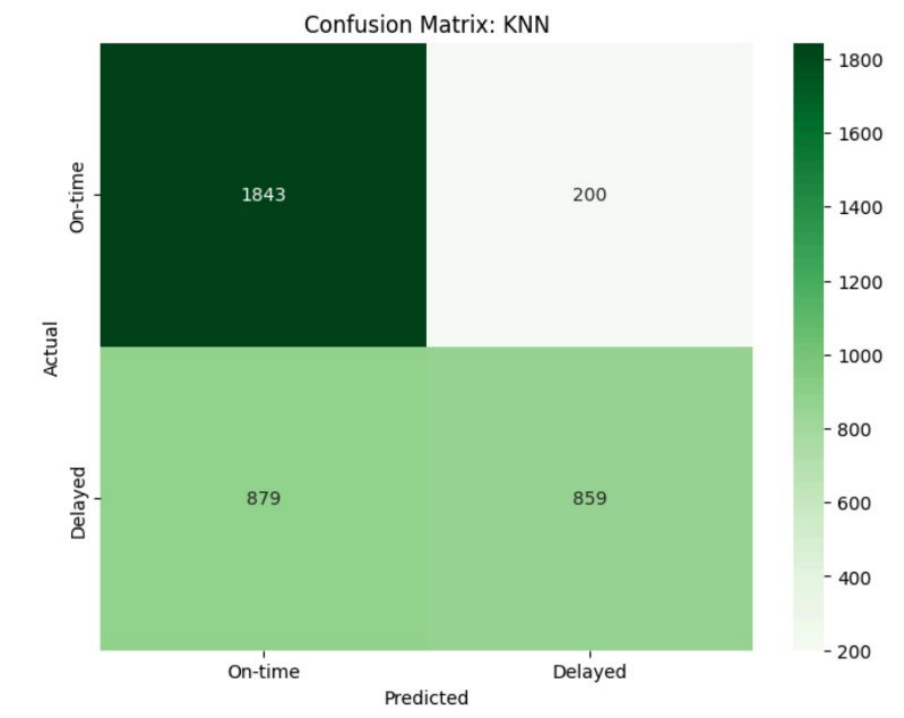
Results

Results

Key metrics obtained from model implementations

K-Nearest Neighbors (KNN)

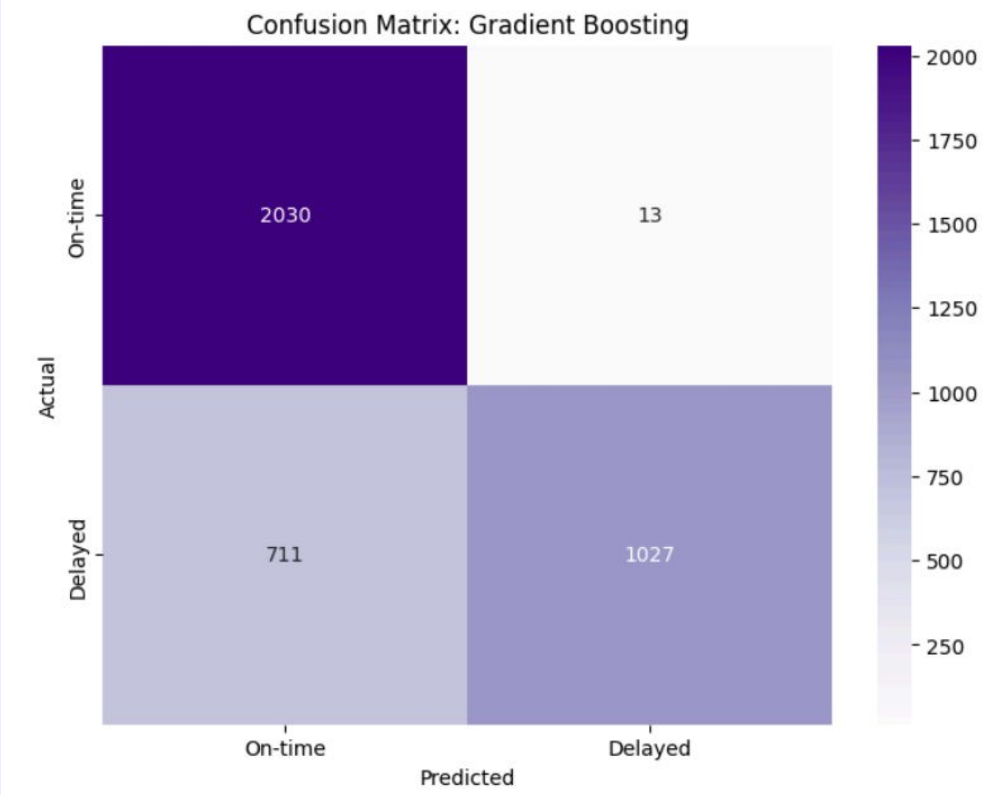




Results

Key metrics obtained from model implementations

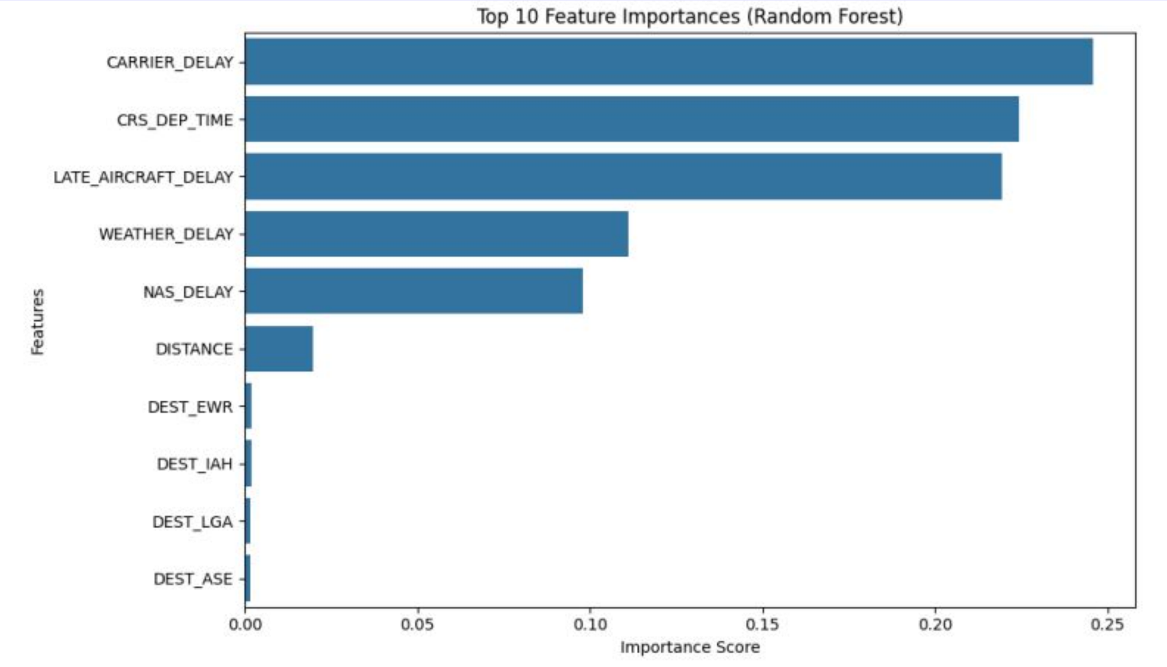
Gradient Boosting

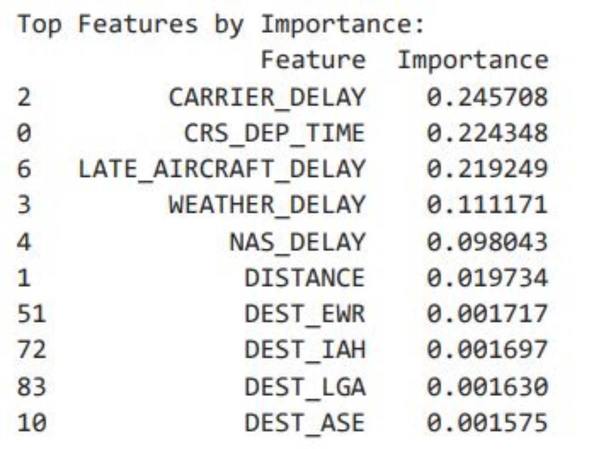


Results

Key metrics obtained from model implementations

Random Forest



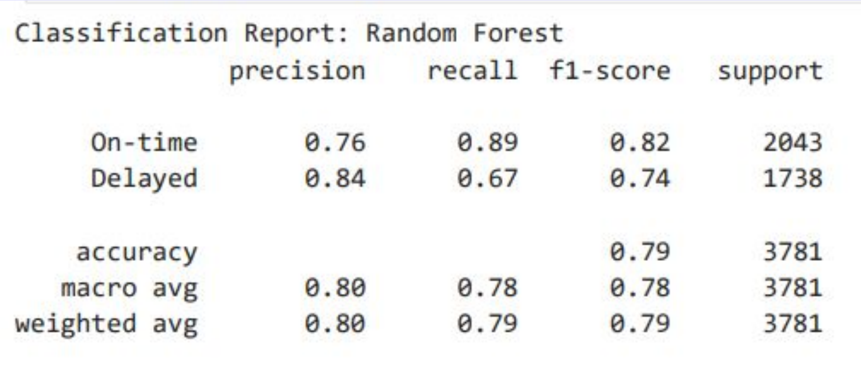


```
Top Features by Importance:
               Feature  Importance
2        CARRIER_DELAY    0.245708
0         CRS_DEP_TIME    0.224348
6  LATE_AIRCRAFT_DELAY    0.219249
3        WEATHER_DELAY    0.111171
4            NAS_DELAY    0.098043
1             DISTANCE    0.019734
51            DEST_EWR    0.001717
72            DEST_IAH    0.001697
83            DEST_LGA    0.001630
10            DEST_ASE    0.001575
```

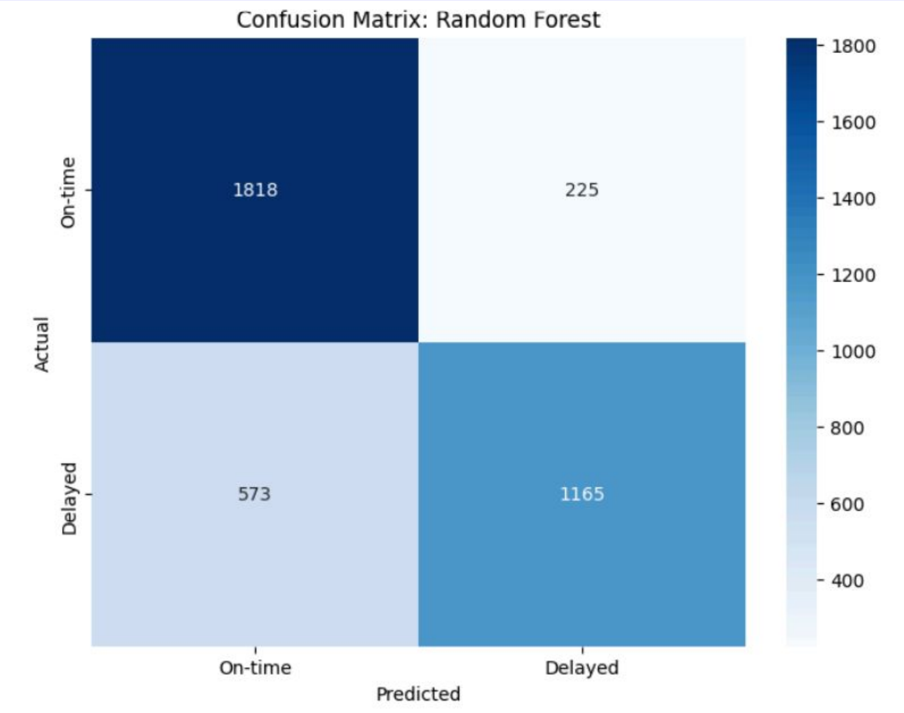
Results

Key metrics obtained from model implementations

Random Forest



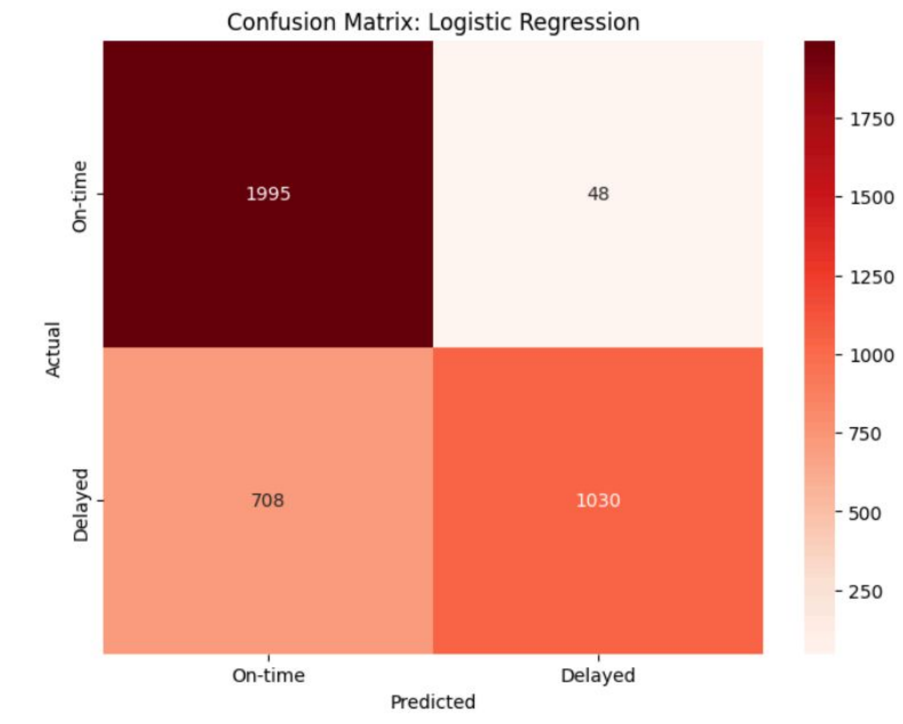
```
Classification Report: Random Forest
              precision    recall  f1-score   support

     On-time       0.76      0.89      0.82      2043
     Delayed       0.84      0.67      0.74      1738

    accuracy                           0.79      3781
   macro avg       0.80      0.78      0.78      3781
weighted avg       0.80      0.79      0.79      3781
```



Results

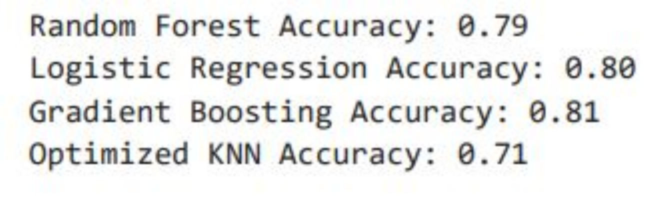
Key metrics obtained from model implementations

Logistic Regression



Results

Key metrics obtained from model implementations



```
Random Forest Accuracy: 0.79
Logistic Regression Accuracy: 0.80
Gradient Boosting Accuracy: 0.81
Optimized KNN Accuracy: 0.71
```

05

Challenges and Adaptations

Challenges and Adaptations

Problems faced and methods used to tackle them

Challenges and Adaptations:

- Challenge: Data was imbalanced, ie. imbalanced classes in the target variable, with significantly more on-time flights compared to delayed ones, led to potential model bias and reduced generalization.
- Adaptation: Applied evaluation metrics such as F1-scores to balance precision and recall, ensuring the model performs effectively on both classes. Additionally, optimized thresholds for certain models to handle imbalance better.

Preprocessing Complexity:

- Challenge: Categorical variables (e.g., origin and destination airports) and numerical features (e.g., delay times, distances) required different preprocessing techniques.
- Adaptation: Used a categorical variable transformation to streamline feature preparation, ensuring consistent encoding for categorical variables and scaling for numerical ones.

06

Conclusion

Conclusion

Evaluation and review of the results obtained

Findings and Implications:

- The Gradient Boosting model achieved the highest accuracy (81%), outperforming other models such as Random Forest, Logistic Regression, and KNN.
- This project highlighted key predictors of flight delays, such as carrier delays, scheduled departure times, and late aircraft arrivals, providing actionable insights to optimize flight operations.
- These results can support aviation industry efforts by enabling predictive tools that improve resource allocation, minimize delays, and enhance passenger satisfaction.

Future Work and Improvements:

- Enhancements:
  - Incorporate additional datasets to include external factors such as air traffic and weather patterns.
  - Experiment with deep learning models to better capture complex interactions between delay factors.
- Develop a user-friendly dashboard for real-time flight delay predictions and operational insights.
- Integrate the predictive model into airline scheduling systems to proactively address potential delays.

THANK YOU

Sohum Bhole, Mohammad Nusairat, Nahom Yohanes, Richa Rameshkrishna, Vageesh Indukuri