PREDICTING Flight ARRIVAL Delay

Team 9, Project 4

Angelina Wright, Eylem Yildirim, Pushpa Chhetri

OBJECTIVE

To develop a model to predict flight arrival delays for flights departing from Arizona airports using 3 years of flight data.

QUESTIONS

- Can we predict if a flight will be delayed upon arrival?
- What are the biggest factors contributing to flight delays?
- Can machine learning models help airlines optimize scheduling?

Dataset

Bureau of Transportation Statistics

On-Time: Reporting Carrier On-Time Performance

December 2021 - November 2024

613,556 rows × 35 columns

Website:

https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr

Tableau

Airline carriers chosen originating from Arizona state cities to create visuals using filters for Unique carrier, Season and Time of the day.

-Op Unique Carrier Vs Count of Op Carrier Fl Num

-Op Unique Carrier Vs Avg Total Delay causes(weather, security, carrier, Nas delay, late aircraft delay)

-Avg Dep Delay Vs Origin city

-Avg Arr Delay Vs Day of the week

-Avg Arr Delay Vs 24 hour window

Links to Tableau public

https://public.tableau.com/app/profile/pushpa.chhetri/viz/Project4Story1/Story1?publish=yes

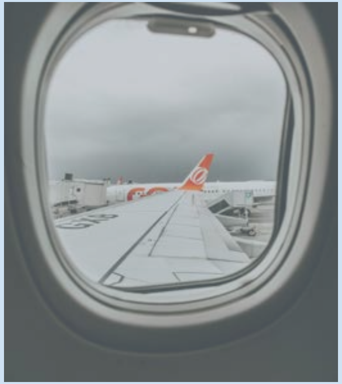
https://public.tableau.com/app/profile/pushpa.chhetri/viz/Project4Story2/Story2?publish=yes

https://public.tableau.com/app/profile/pushpa.chhetri/viz/Project4story3_17424949752860/Story3?publish=yes

PREPROCESSING & DATA PREPARATION

- DATA INTEGRATION
  - Merged Monthly Files: Combined monthly csv files into a single dataset.
  - Filtered for Arizona Departures: Retained columns where ORIGIN_STATE_ABR = "AZ".
- FEATURE ENGINEERING ENHANCEMENTS
  - Introduced new columns to enhance analysis:
    - ARR_DELAY: 0 = on-time, 1 = delayed
    - DAY_PART: Departure Time to “Early Morning”, “Morning”, “Midday”, “Afternoon”, “Evening”, “Night”, “Late Night”
    - FLIGHT_TRAFFIC: Count of flights every hour leaving the origin airport
    - SEASON: Based on month - Fall, Winter, Spring, Summer
    - SC_DEP_TIME: Scheduled departure time
    - SC_HOUR: Scheduled hour flight departing
- DATA CLEANING & ENCODING
  - DEPT_TIME - date time format & removed missing values.
  - Mapped Carrier Codes: Added airline names for improved readability.
  - Encoding Data: Applied label encoding to categorical variables.

Target & features



- **TARGET**
    - ARR_DELAY: 0 = on-time, 1 = delayed
- **FEATURES (18)**
    - **ORIGINAL**
        - YEAR
        - MONTH
        - DAY_OF_MONTH
        - DAY_OF_WEEK
        - AIR_TIME: Flight time (minutes)
        - DEP_DELAY_NEW: Delay time (minutes)
        - DEST: Destination airport, city name
        - DEST_STATE_ABR: Destination state
        - DISTANCE: Distance between airports (miles)
        - OP_CARRIER_FL_NUM: Flight number
        - ORIGIN: Origin Airport
        - TAXI_OUT: Taxi out time (minutes)
        - WHEELS_OFF: Time aircraft took off
    - **NEW COLUMNS**
        - DAY_PART: Departure Time to “Early Morning”, “Morning”, “Midday”, “Afternoon”, “Evening”, “Night”, “Late Night”
        - FLIGHT TRAFFIC: Count of flights every hour leaving the origin airport
        - OP_UNIQUE CARRIER: Carrier mapped from carrier code to carrier name
        - SC_HOUR: Scheduled hour flight departing
        - SEASON: Based on month - Fall, Winter, Spring, Summer

Model choice

- Models:
  - Machine Learning (Random Forest, Logistic Regression, Decision Tree,Gradient Boosting, K-Neighbors)
  - Neural Network
- Methods/Tools Used to Optimize the Results:
  - Hyperparameter Tuning
  - SMOTE to handle class imbalance
  - Feature Importance
  - Spearman Correlation
  - Out-Of-Bag Score (RandomForest)
  - Class-weights (RandomForest) to handle imbalanced data
    - Prevents the model from favoring the majority class (on-time flights)

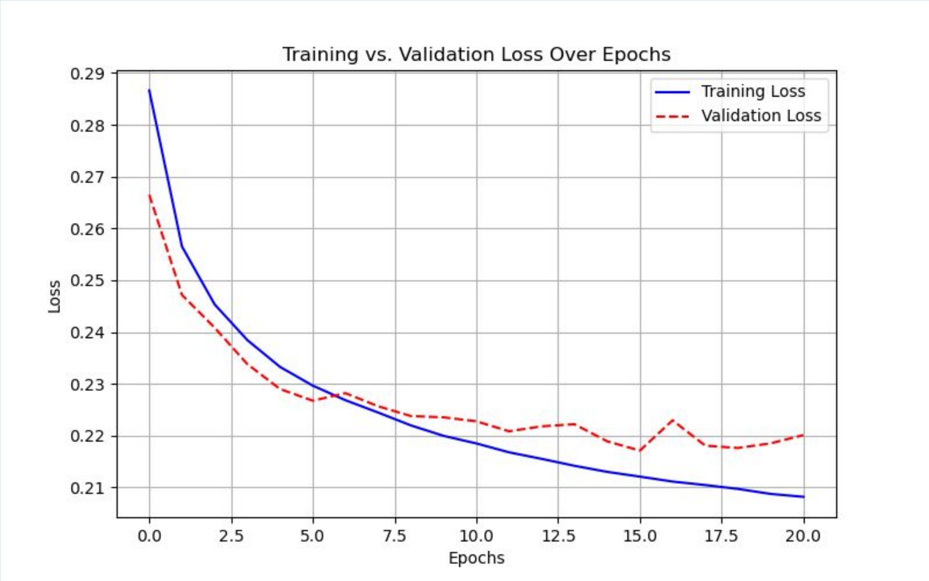
Machine Learning models

Model	Accuracy	Precision (0)	Precision (1)	Recall (0)	Recall (1)
Random Forest	**0.89**	**0.88**	**0.91**	**0.96**	**0.77**
Logistic Regression	0.88	0.87	0.89	0.95	0.75
Gradient Boosting	0.87	0.86	0.90	0.95	0.74
Decision Tree	0.85	0.88	0.79	0.88	0.80
K-Neighbors	0.76	0.77	0.74	0.90	0.52

Winner:

Random Forest!

Neural Network



- Applied Keras Hyperband tuner(activation,regularization, num of layers, dropout, etc)
- Training data does not generalize well to new data
- Possible overfitting
- Optimal epoch about 8-9
- Inconsistency on each model run

Accuracy: 0.91

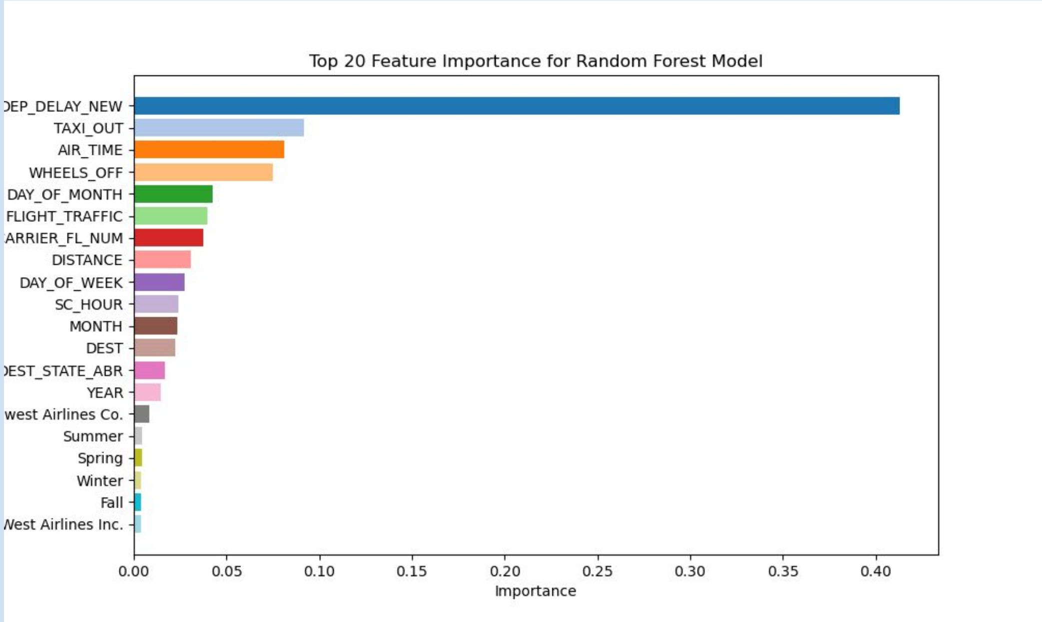
Random Forest Model Optimization

1. **Number of Trees:** Number of trees (50) increased but key statistics did not improve (100,200)

2. **Feature Importance:** Calculated how much each ‘Feature’ contributed the results and built a model

- Feature Importance > 0.01
- Feature Importance > 0.04

Result: Classification report results did not improve.



Random Forest MODEL OPTIMIZATION

3. SMOTE: Handling class imbalance in the dataset.

4. Hyperparameter Tuning: Randomized search to find best Hyperparameters
Max_depth: None, min_samples_split: 2, min_samples_leaf: 1, bootstrap:True

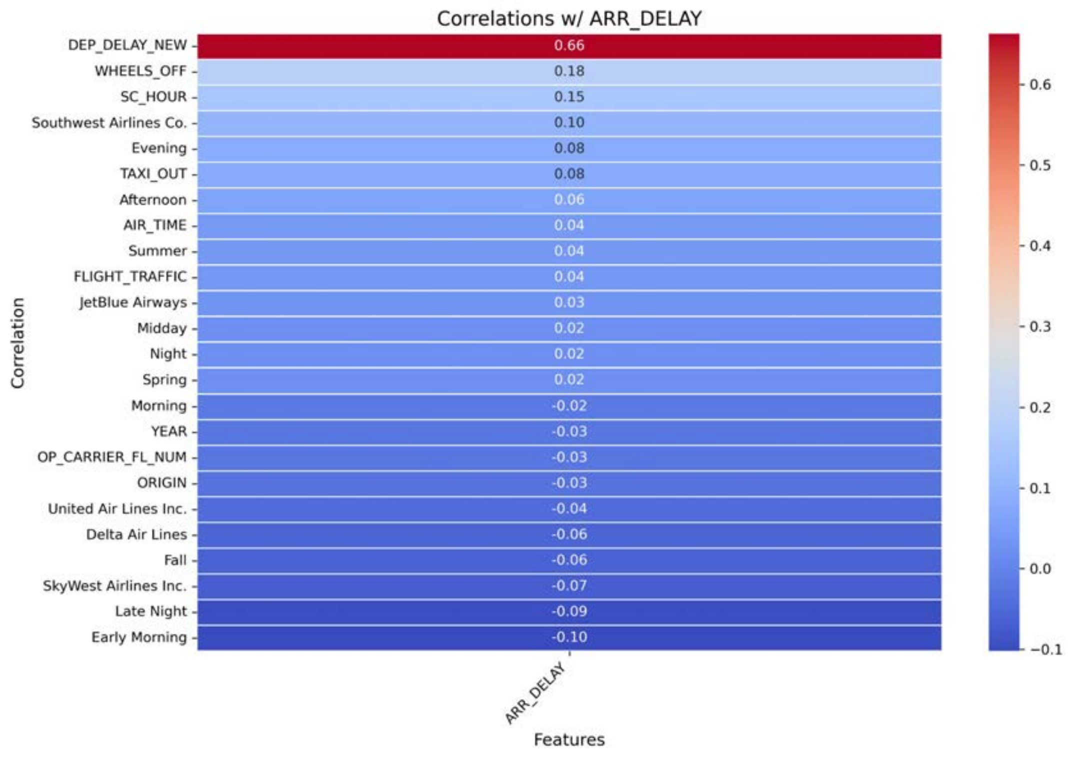
	Initial Model		**After SMOTE**		**SMOTE+Hyperparameter Tuning**	
	On-Time (0)	Delayed (1)	On-Time (0)	Delayed (1)	On-Time (0)	Delayed (1)
Accuracy	0.89		0.89		0.89	
OOB Score	0.89		0.91		0.91	
Precision	0.88	0.91	0.89	0.87	0.89	0.87
Recall	0.96	0.77	0.93	0.81	0.93	0.81
Class Dist	347,426	199,100	347,426	347,426	347,426	347,426

Random Forest MODEL OPTIMIZATION

5. Spearman Correlation:

Feature selection based on |corr| > 0.03

- Results did not improve



Random Forest BEST Model

Overall Accuracy: Model accurately classifies 89% of flights.

OOB Score: A high 91% suggests strong performance on unseen data. Model generalizes well.

Precision: When predicting flight status:

- **On-time flights:** Correct 89% of the time
- **Delayed flights:** Correct 87% of the time

Recall: The model successfully identifies:

- 93% of actual **on-time flights**
- 81% of actual **delayed flights**

After SMOTE		
	On-Time (0)	Delayed (1)
Accuracy	0.89	
OOB Score	0.91	
Precision	0.89	0.87
Recall	0.93	0.81
Class Dist	347,426	347,426

CONSIDERATIONS & STRATEGIES TO ENHANCE MODEL ACCURACY

- Weather conditions
  - To observe correlations with delays
- Holidays
  - Flight may have more delays around the holidays

QUESTIONS?

TIME!
TIME!
DELAYED
DELAYED!
TIME!
AIRPORT
DELAYT?
DELAYED
AIRPORT