Predicting Departure Delays

Hawa Abdo, David Burke, Brianna Mitri, &
Casey Wright

Webportal

Homepage & Dashboards

Prediction Problem

Our goal was to predict flight delays using a combination of numeric and categorical features.

We investigated 3 options for the target variable.

- 6 classes:
  - Early
  - On Time
  - 1-10 min late
  - 11-30 min late
  - 31-60 min late
  - 60+ min late
- 4 classes:
  - Early
  - On Time (0-15 min late)
  - 16-60 min late
  - 61+ min late
- Binary
  - Not delayed: up to 15 min late
  - Delayed: 16+ min late

Data Sources

Source	Data
U.S. Bureau of Transportation Statistics	Flights originating from LAX, 2020-2024
Federal Aviation Administration (FAA)	Aircraft specifications by tail number
National Weather Service	Historical weather observations for origin (LAX) and destination airports

Data Cleaning

- Removed columns that contained a single value overwhelmingly (>99.9%)
- Condensed categorical columns
  - Low value counts merged into 'Other'
  - Synonymous categories combined
    - For example: Airbus and Airbus SAS
- Converted date and time columns into cyclical format (sin, cos) to facilitate machine learning
- Verified no nulls remained

Splitting for Machine Learning

- Loaded from SQLite database with SQLAlchemy
- 80% training & 20% testing
- Applied stratification
- Processed data
  - Categorical→ One Hot Encoder (will handle new values better)
  - Numerical→ standard scaler
- Tried rebalancing with SMOTE
  - Generated very large files
  - Took longer to run with decreased accuracy

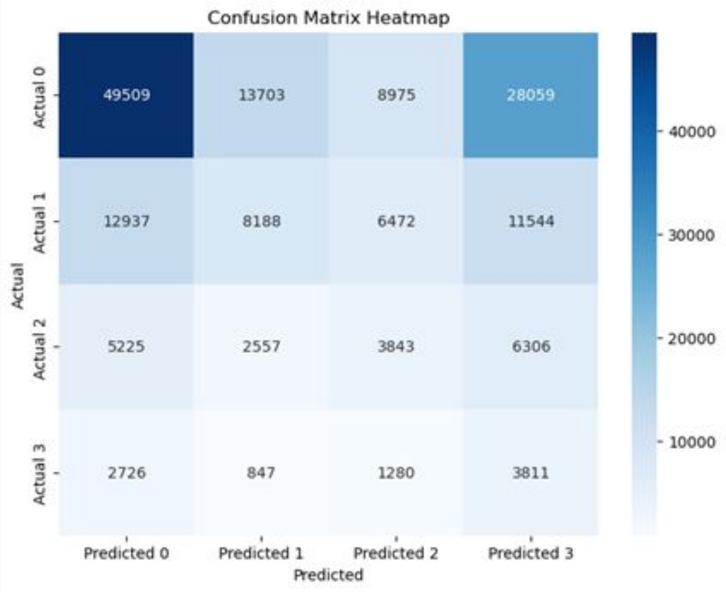
Modeling

We explored several modeling types with different assortments of hyperparameters.

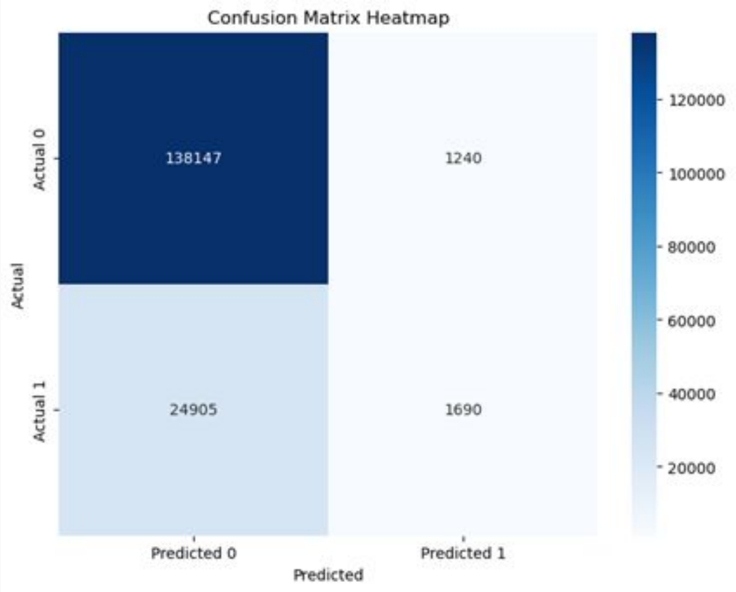
- Random Forest
- Support Vector Machine
- Logistic Regression
- K-Nearest Neighbors
- Neural Network

Random Forest

Multiclass (4 classes)



Binary Classification



	Multiclass	Binary Classification	Binary (Tuned)
Training Accuracy	0.3938	0.9999	0.5970
Testing Accuracy	0.3937	0.8424	0.5968

Random Forest

Binary Classification, full tree

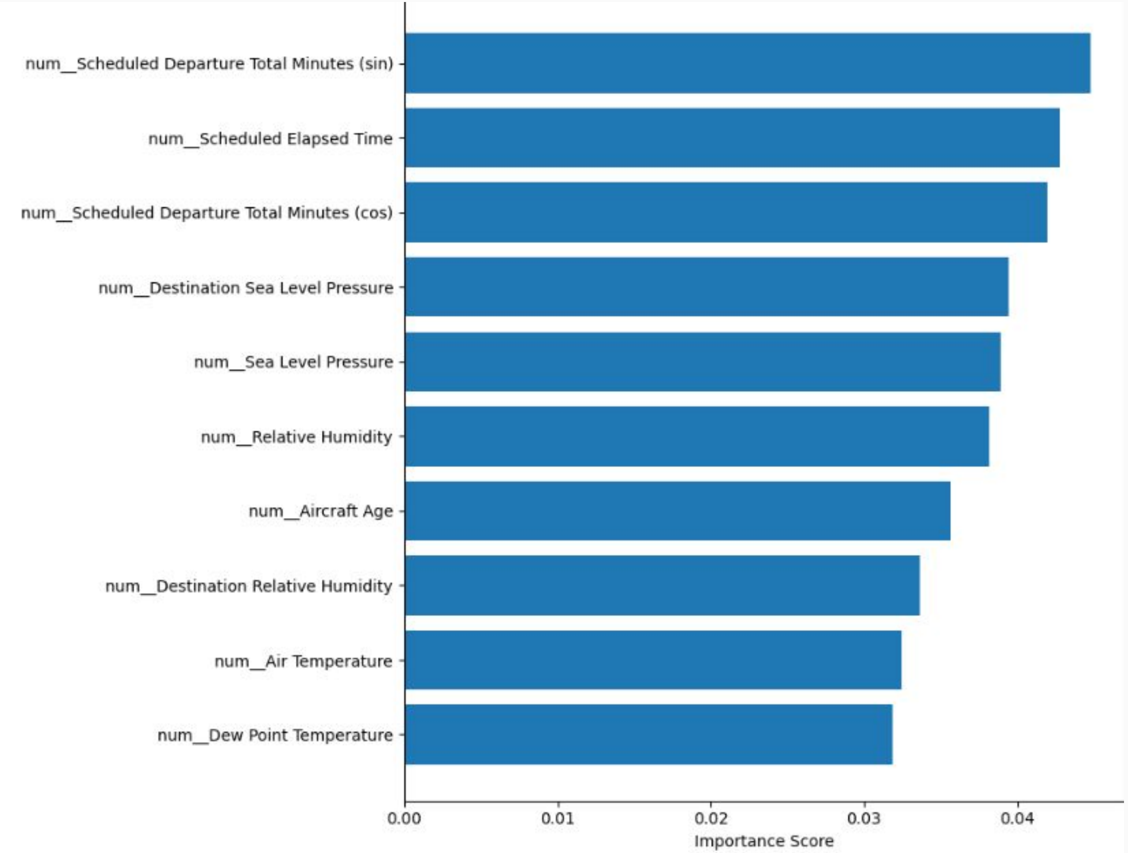
Label	Precision	Recall	F1
0 <= 15 min	0.85	0.99	0.91
1 > 15min	0.36	0.06	0.11
Accuracy			0.84

Binary Classification, hyperparameters determined by RandomSearchCV

Label	Precision	Recall	F1
0	0.88	0.59	0.71
1	0.22	0.60	0.32
Accuracy			0.59

Random Forest

Top 10 Feature Importances

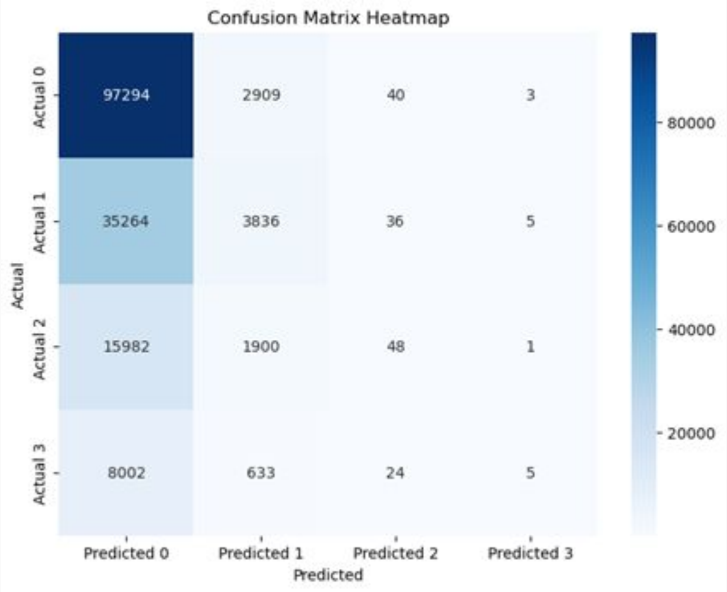


Support Vector Machine

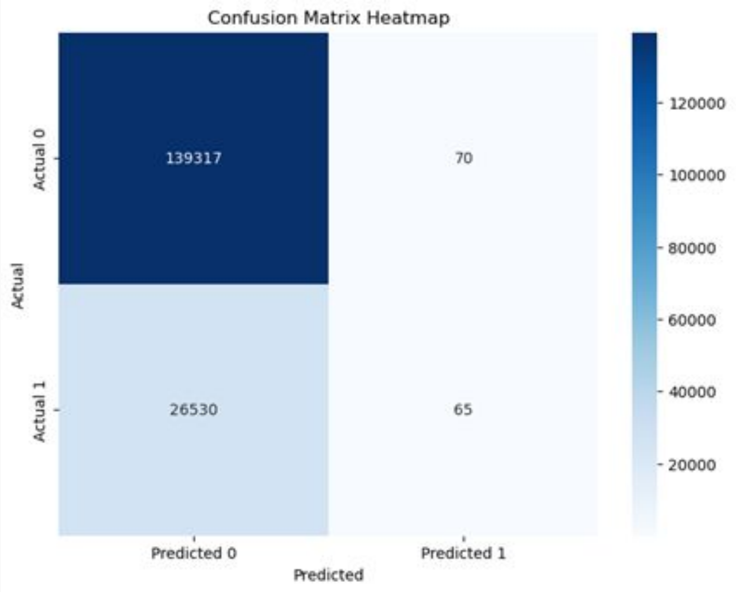
- Model did not complete
- Run time over 20 hours!
- Local machines and colab tried

Logistic Regression

Multiclass (4 classes)



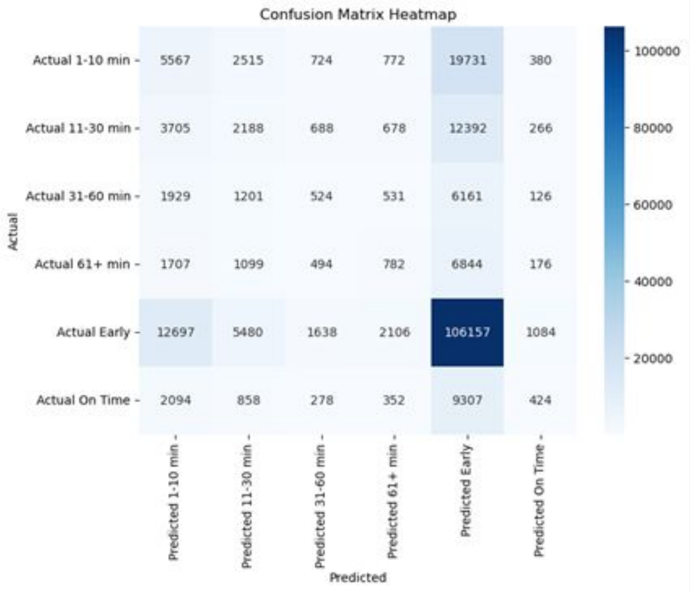
Binary Classification



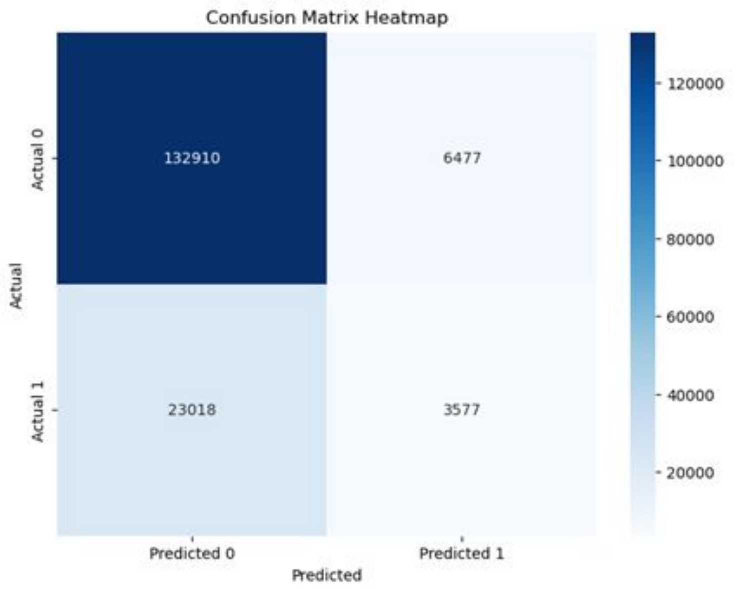
	Multiclass	Binary Classification
Training Accuracy	0.6105	0.8397
Testing Accuracy	0.6096	0.8397

K-Nearest Neighbors

Multiclass (6 classes)



Binary Classification



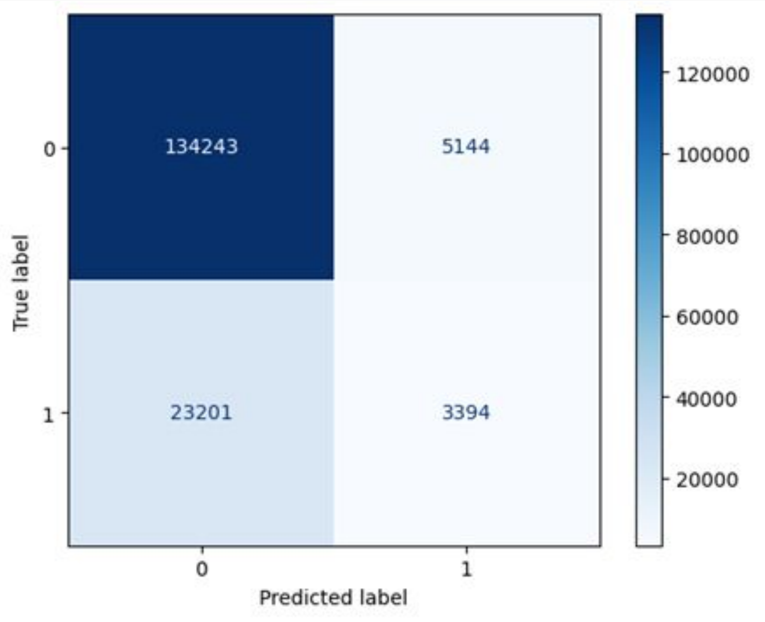
	Multiclass	Binary Classification
Training Accuracy	0.6613	0.8619
Testing Accuracy	0.5412	0.8223

Neural Network

Label	Precision	Recall	F1
Not Delayed (0)	0.85	0.96	0.90
Delayed (1)	0.40	0.13	0.19
Accuracy			0.83

	Binary Classification
Training Accuracy	0.8516
Testing Accuracy	0.8292

Binary Classification



83%

The Neural Network model was selected as the best option as it reached an overall accuracy of 83%, with the best precision and recall for the Delayed category.

Deployment

Backend: Flask App
holds the neural network model and preprocessor

Frontend: Web App
run predictions

Thank you!

Questions/ Comments?

Thank you for listening! Let us know if you have any questions.

“This is a super-important quote”

- From an expert

Final point

A one-line description of it

Prediction Problem

What's this presentation about? Use this slide to introduce yourself and give a high level overview of the topic you're about to explain.

This is the most important takeaway that everyone has to remember.