Machine Learning

Aircraft Delays and Cancellations

Nicole Thapa (20935982)
Harpreet Ghotra (20951321)
Dharsaa Bhagdueva (20954664)
Taylor Liew (20939880)
Kathryn Percy-Robb (20964151)

Table of Contents

Introduction

Why are flight delays and cancellation important?

- Mitigate Passenger Disruption
- Improve Passenger Satisfaction
- Optimize Resource Implications
- Crew Scheduling
- Maintenance Planning
- Reduce general costs and
- Reduce carbon emission, noise pollution and resource consumption.

Data Pre-Processing

Data Overview

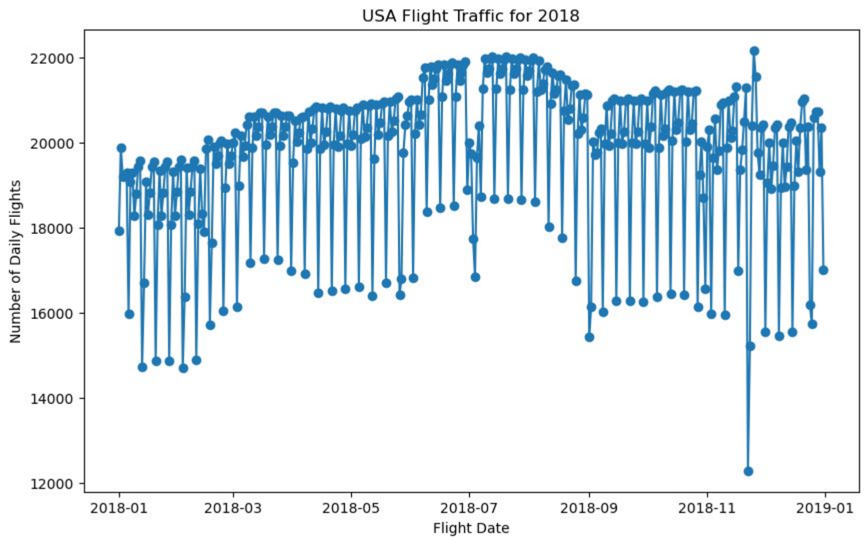
- Approx. 7.2 million rows of data, 28 columns

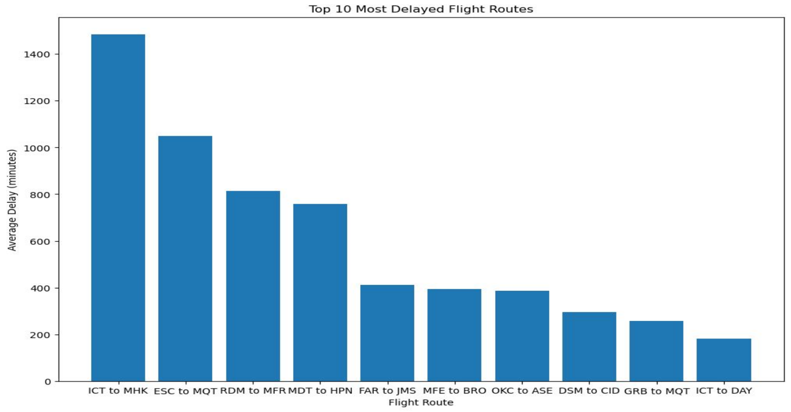
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7213446 entries, 0 to 7213445
Data columns (total 28 columns):
 #   Column               Dtype
---  ------               -----
 0   FL_DATE              object
 1   OP_CARRIER           object
 2   OP_CARRIER_FL_NUM    int64
 3   ORIGIN               object
 4   DEST                 object
 5   CRS_DEP_TIME         int64
 6   DEP_TIME             float64
 7   DEP_DELAY            float64
 8   TAXI_OUT             float64
 9   WHEELS_OFF           float64
 10  WHEELS_ON            float64
 11  TAXI_IN              float64
 12  CRS_ARR_TIME         int64
 13  ARR_TIME             float64
 14  ARR_DELAY            float64
 15  CANCELLED            float64
 16  CANCELLATION_CODE    object
 17  DIVERTED             float64
 18  CRS_ELAPSED_TIME     float64
 19  ACTUAL_ELAPSED_TIME  float64
 20  AIR_TIME             float64
 21  DISTANCE             float64
 22  CARRIER_DELAY        float64
 23  WEATHER_DELAY        float64
 24  NAS_DELAY            float64
 25  SECURITY_DELAY       float64
 26  LATE_AIRCRAFT_DELAY  float64
 27  Unnamed: 27          float64
dtypes: float64(20), int64(3), object(5)
```

Data Cleaning

1. Drop Dummy column
2. Replace missing values in time duration columns with 0s
3. Replace missing values in timestamp columns with 0s
4. Convert Flight Date column to datetime for seasonal analysis by day, month, year

Data Exploration





High-Risk of Cancellation by Airline/Carrier:

1. WN - Southwest Airlines
2. OH - PSA Airlines (American Eagle)
3. AA - American Airlines

Top 3 Causes for Delay:

1. Weather
2. Airline
3. Airport Security

ANALYSIS 1

Flight Delay Cause Prediction

Supervised Learning → Category Prediction → Logistic Regression

CATEGORIES

Weather

```
Accuracy: 0.7125
Confusion Matrix:
[[1356  663]
 [ 487 1494]]
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.67      0.70      2019
           1       0.69      0.75      0.72      1981

    accuracy                           0.71      4000
   macro avg       0.71      0.71      0.71      4000
weighted avg       0.71      0.71      0.71      4000
```

Security

```
Accuracy: 0.934
Confusion Matrix:
[[1755  264]
 [   0 1981]]
Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.87      0.93      2019
           1       0.88      1.00      0.94      1981

    accuracy                           0.93      4000
   macro avg       0.94      0.93      0.93      4000
weighted avg       0.94      0.93      0.93      4000
```

Carrier

```
Accuracy: 0.61525
Confusion Matrix:
[[1230  789]
 [ 750 1231]]
Classification Report:
              precision    recall  f1-score   support

           0       0.62      0.61      0.62      2019
           1       0.61      0.62      0.62      1981

    accuracy                           0.62      4000
   macro avg       0.62      0.62      0.62      4000
weighted avg       0.62      0.62      0.62      4000
```

Late Aircraft

```
Accuracy: 0.655
Confusion Matrix:
[[1252  767]
 [ 613 1368]]
Classification Report:
              precision    recall  f1-score   support

           0       0.67      0.62      0.64      2019
           1       0.64      0.69      0.66      1981

    accuracy                           0.66      4000
   macro avg       0.66      0.66      0.65      4000
weighted avg       0.66      0.66      0.65      4000
```

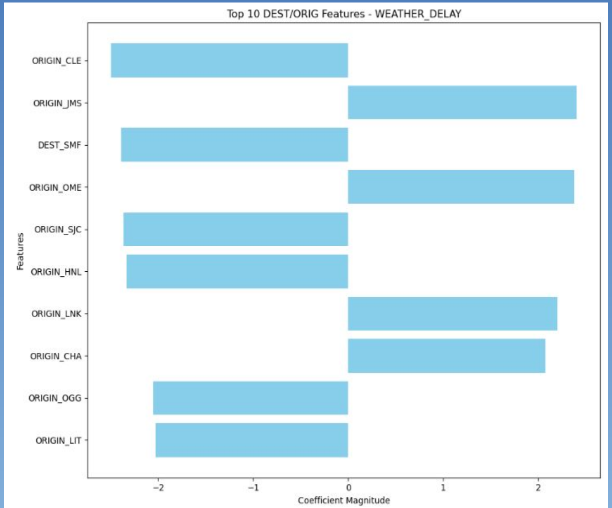
NAS

```
Accuracy: 0.60775
Confusion Matrix:
[[1198  821]
 [ 748 1233]]
Classification Report:
              precision    recall  f1-score   support

           0       0.62      0.59      0.60      2019
           1       0.60      0.62      0.61      1981

    accuracy                           0.61      4000
   macro avg       0.61      0.61      0.61      4000
weighted avg       0.61      0.61      0.61      4000
```

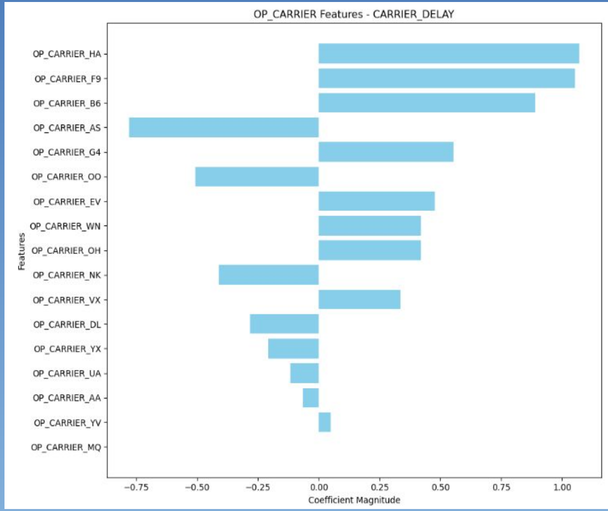
ANALYSIS 1

Flight Delay Cause Prediction

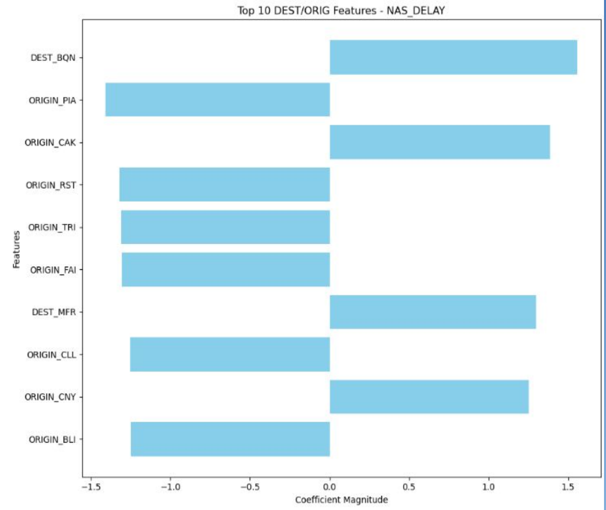
Weather



Carrier



NAS



- Positive Coefficient: greater likelihood of delays
- Negative Coefficient: smaller likelihood of delays
- Magnitude: feature's strength of influence on prediction outcome of model

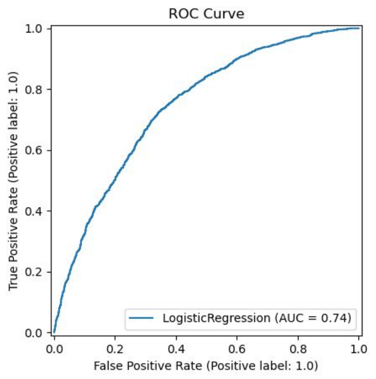
ANALYSIS 2

Flight Cancellation Classification

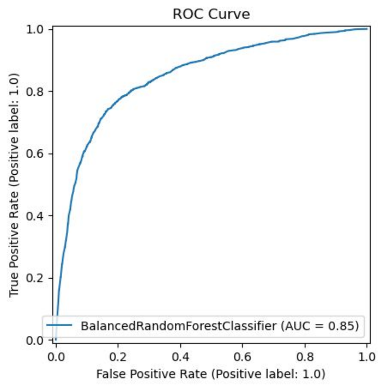
- Target: 'CANCELLED' (1=Cancelled Flight, 0 = Non-Cancelled Flight)
- Features: 'FL_DATE', 'OP_CARRIER', 'ORIGIN', 'DEST', 'CRS_DEP_TIME', 'CRS_ARR_TIME', 'DISTANCE'

Logistic Regression Vs **Balanced Random Forest**

```
Accuracy: 0.68725
Confusion Matrix:
[[1359  660]
 [ 591 1390]]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.70      0.67      0.68      2019
         1.0       0.68      0.70      0.69      1981

    accuracy                           0.69      4000
   macro avg       0.69      0.69      0.69      4000
weighted avg       0.69      0.69      0.69      4000
```




```
Accuracy: 0.772

Confusion Matrix:
[[1479  540]
 [ 372 1609]]

Classification Report:
              precision    recall  f1-score   support

         0.0       0.80      0.73      0.76      2019
         1.0       0.75      0.81      0.78      1981

    accuracy                           0.77      4000
   macro avg       0.77      0.77      0.77      4000
weighted avg       0.77      0.77      0.77      4000
```



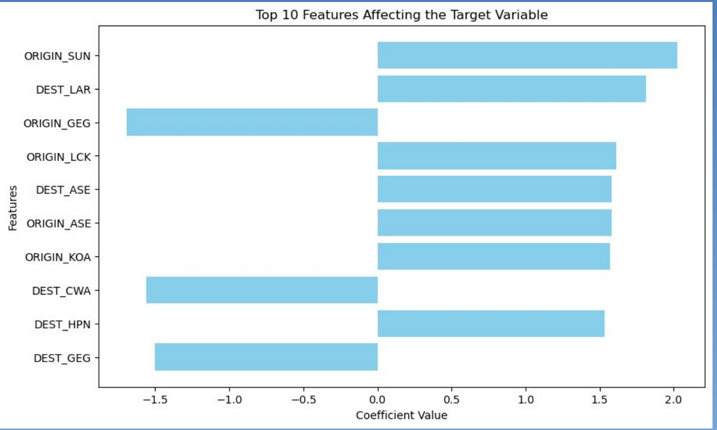
ANALYSIS 2

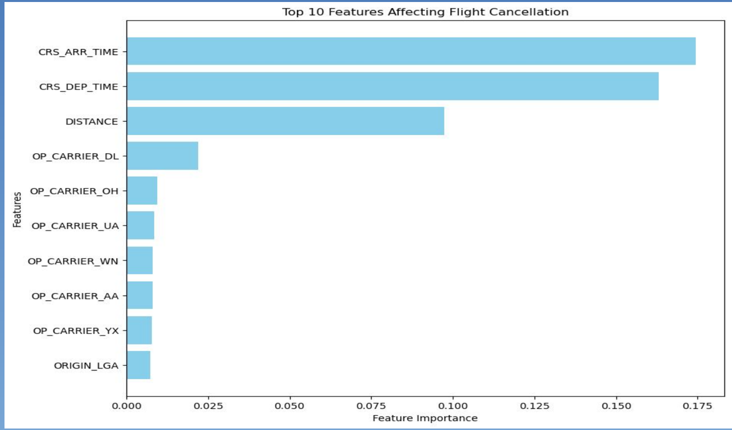
Flight Cancellation Classification

Logistic Regression

Vs

Balanced Random Forest





- Positive Coefficient: greater likelihood of cancellations
- Negative Coefficient: smaller likelihood of cancellations

- Higher Feature Importance Percentage: Greater usefulness in reducing uncertainty in the trees

Flight Anomaly Detection

Goal: Classify flights based on similar characteristics to help identify anomalous flights

Model Features: 'DEP_DELAY', 'TAXI_OUT', 'TAXI_IN', 'ARR_DELAY', 'CARRIER_DELAY', 'WEATHER_DELAY', 'NAS_DELAY', 'SECURITY_DELAY', 'LATE_AIRCRAFT_DELAY'

K-Means Cluster **Model Benefits:**

- ➢ Able to handle large datasets with many features
- ➢ Easy to interpret findings/clusters
- ➢ Unsupervised method (helpful for anomaly flight detection)

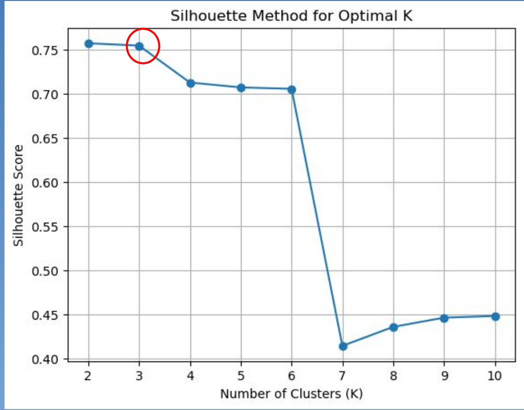
Key Anomalous Cluster (Cluster 1) Findings:

- Cluster 1 tends to have higher taxi out times and lower taxi in times than flights in Cluster 0 and 2
- Cluster 1 flights tend to experience far more delays due to carriers or weather than flights in Cluster 0 and 2
- Cluster 1 flights face more National Airport Security delays than Cluster 0
- Cluster 1 flights are most impacted by arrival delays and aircraft delays than other cluster flights

Model Results:

- Cluster 1 group had highest variance
- Cluster 2 group had an average of 100+ minutes for security delays
- **1,160** anomalous flights detected (Cluster 1) out of 25,000 flights

K-Means Cluster



Silhouette Analysis for parameter selection (score=0.76)

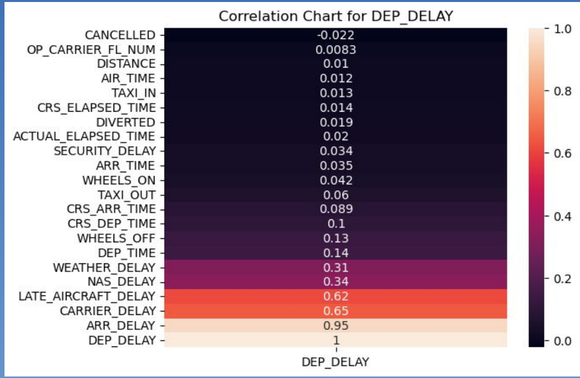
Anomalous Flight Detection Methods:
1) Cluster Variance Analysis
2) Feature PairPlot Cluster Analysis

ANALYSIS 4

Flight Delay Amount Prediction

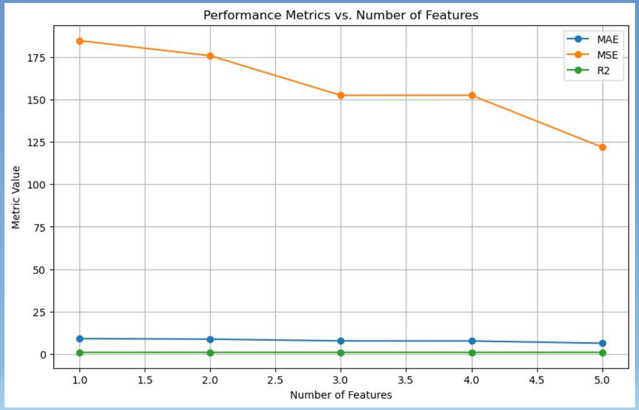
Part One: Multiple Linear Regression

- Step 1: Identify features to include in regression
  - 30% correlation threshold
- Step 2: Split data and fit models
  - Starting with most correlated feature
  - Continue to add features until threshold
- Step 3: Performance Evaluation
  - Calculate MAE, MSE, R-Squared for each model
  - Select best model based on R-squared
- Step 4: Conclusion
  - Model with all threshold features is optimal



Columns for Analysis:
['DEP_DELAY', 'ARR_DELAY', 'CARRIER_DELAY', 'LATE_AIRCRAFT_DELAY', 'NAS_DELAY', 'WEATHER_DELAY']

Features	MAE	MSE	R-Squared
1	9.007054	184.70862	0.907141
2	8.709634	175.816587	0.911611
3	7.670247	152.472994	0.923347
4	7.640934	152.503092	0.923332
5	6.291322	121.926685	0.938703

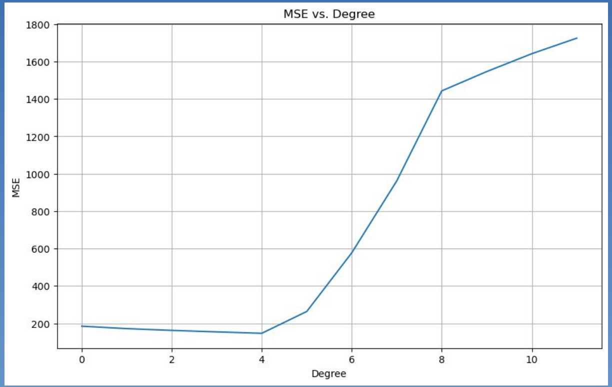


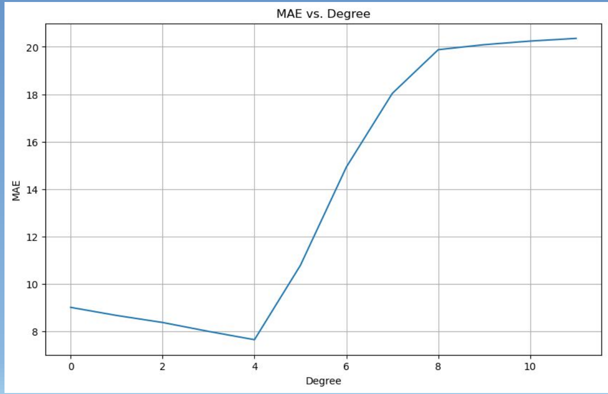
ANALYSIS 4

Flight Delay Amount Prediction

Part Two: Polynomial Regression

- Step 1: Identify Best Feature
  - For ARR_DELAY vs DEP_DELAY
  - With k=1, select k best features
- Step 2: Fit Model with Incrementing Degrees
  - For loop with degrees 1 to 12
  - Print evaluation metrics MAE and MSE
- Step 3: Identify Optimal Degree
  - Plot MAE and MSE
  - Select model where test errors are minimized
- Step 4: Conclusion
  - Degree 4 is optimal; contains smallest errors compared to all other degreed models

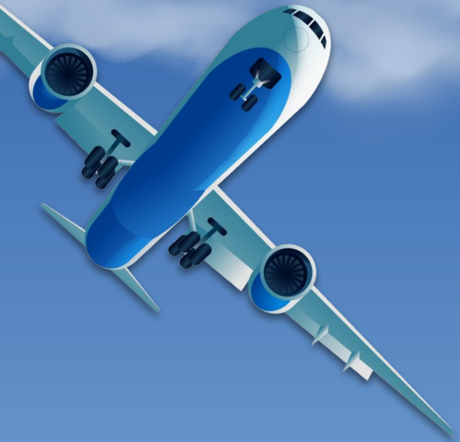




Part Three: Comparison

- Comparing linear and polynomial regression ARR_DELAY vs DEP_DELAY
  - Polynomial has better performance

	Linear	Polynomial
MAE (test)	9.007054366	7.993325383
MSE (test)	184.7086196	154.2529259
R-squared	0.90714098	0.922452046



Thanks!

Kaggle Dataset:
https://www.kaggle.com/datasets/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018?select=2018.csv