Flight Delays Prediction

Datasci-261-team-7-1 (2023 Fall)

Zukang Yang, Nishika Abeytunge, Sam Meng, Joey He, Weijie Yang

Berkeley
UNIVERSITY OF CALIFORNIA

Agenda & Project Outline

1	Project Introduction
2	Exploratory Data Analysis (EDA)
3	Featuring Engineering & Top Features
4	Feature Transformation
5	Overview of Machine Learning Pipeline
6	Machine Learning Pipeline 1: Model & Performance
7	Machine Learning Pipeline 2: Model & Performance
8	Extra Credits Phase 2: Full Join of Dataset, EDA, Feature Processing, Modeling Phase 3: Clean Dataset Preparation for Flight Delays & Weather for Years 2020 - 2022 Phase 3: A different Neural Network
9	Project Conclusion
10	References

Berkeley
UNIVERSITY OF CALIFORNIA

Project Introduction

Abstract

Problem: How we can mitigate the economic losses and customer inconvenience due to unexpected flight delays?

Objective: Provide a scalable machine learning solution based on historical data to forecast flight delays

Berkeley
UNIVERSITY OF CALIFORNIA

Exploratory Data Analysis (EDA)

	path	name
1	dbfs:/mnt/mids-w261/HW5/	HW5/
2	dbfs:/mnt/mids-w261/OTPW_12M/	OTPW_12M/
3	dbfs:/mnt/mids-w261/OTPW_1D_CSV/	OTPW_1D_CSV/
4	dbfs:/mnt/mids-w261/OTPW_36M/	OTPW_36M/
5	dbfs:/mnt/mids-w261/OTPW_3M/	OTPW_3M/
6	dbfs:/mnt/mids-w261/OTPW_3M_2015.csv	OTPW_3M_2015.csv
7	dbfs:/mnt/mids-w261/OTPW_60M/	OTPW_60M/
8	dbfs:/mnt/mids-w261/airport-codes_csv.csv	airport-codes_csv.csv
9	dbfs:/mnt/mids-w261/datasets_final_project/	datasets_final_project/
10	dbfs:/mnt/mids-w261/datasets_final_project_2022/	datasets_final_project_2022/

Target Variable - `DEP_DEL15`

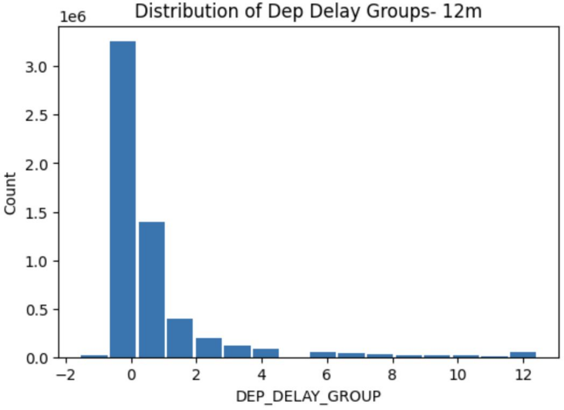
`'CARRIER_DELAY','WEATHER_DELAY','NAS_DELAY','SECURITY_DELAY','LATE_AIRCRAFT_DELAY'`

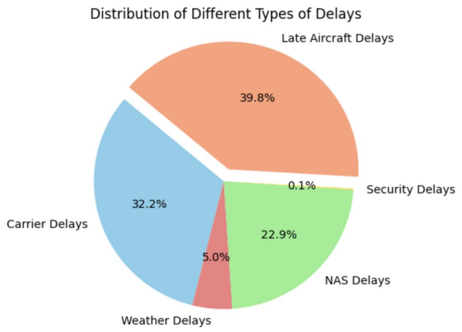
Property	Fact 12M	Fact 3M
Total number of rows	11623708	1401363
Total number of rows after removing duplicates	5811854	1401363
Number of dep delays >15 min	1055735	277302
% of dep delays >15 min	18.2%	19.8%
% of cancel	1.5%	3.0%
% of dep ontime	80.3%	77.2%
Date range start date	01/01/2015	01/01/2015
Date range end date	12/31/2015	03/31/2015
Number of unique carriers	14	14
Number of unique airports	320	313
Total number of columns	216	216
Number of columns with <60% records of missing data	107	107
Number of columns added back to list*	5	5
Number of columns with similar or not relavent data	76	76
Number of columns selected for analyis	39	39

Berkeley
UNIVERSITY OF CALIFORNIA

Exploratory Data Analysis (EDA)

Imbalanced Data





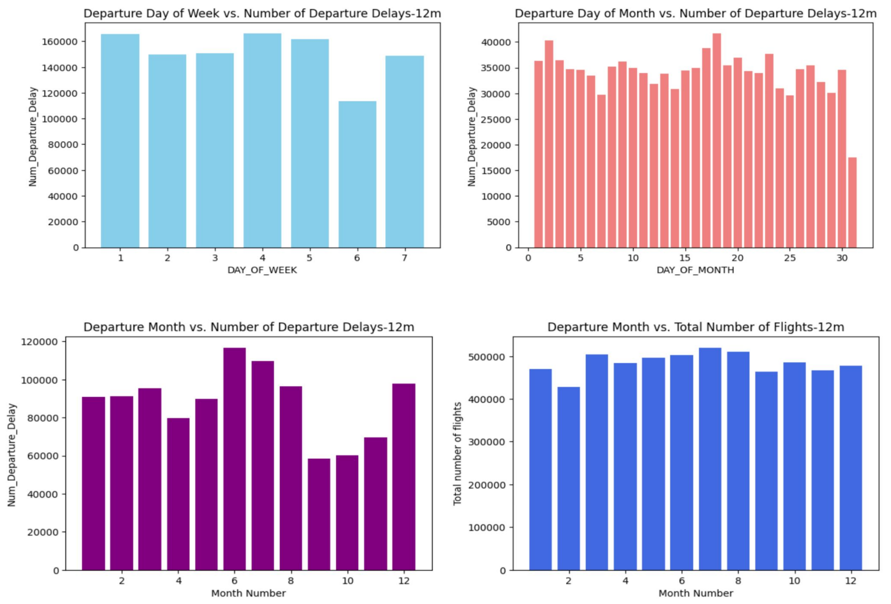
Total delays are only 19.8% out of total trips

Weather related delays are only 5% out of all the delays

Berkeley
UNIVERSITY OF CALIFORNIA

Exploratory Data Analysis (EDA)

Data Seasonality



Relatively lower number of delays on saturdays

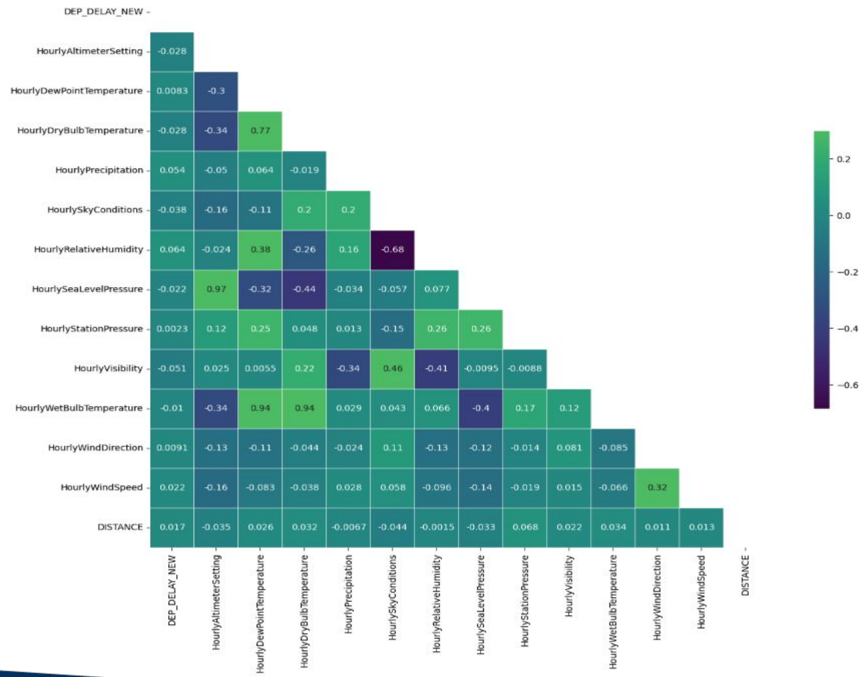
Beginning and mid of the month slightly higher number of delays. Last day of the month is 31 and occur only in 7 months

Summer and december has more number of delays, however there is no significant variability in the total number of trips by month

Berkeley
UNIVERSITY OF CALIFORNIA

Exploratory Data Analysis (EDA)

Correlation of Weather Features



There is no significant correlation of the weather features with the departure delay

'*HourlySkyConditions*' and '*HourlyAltimeterSetting*' has the highest negative correlation which is negative.68

'*HourlySeaLevelPressure*' and '*HourlyAltimeterSetting*' has the highest positive correlation which is .97

Berkeley
UNIVERSITY OF CALIFORNIA

Feature Engineering and Top Features

US Holiday Dataset

Natural Disaster Dataset (NCEI)

Joined by Date & Origin State

OTPW Dataset

New Features

Holiday
Disaster

Our Expectation:

- ➔ Flight volume is higher in Holiday -> Delay prob. increase
- ➔ Weather Disaster related to flights -> Delay prob. increase

Notes:

A. Only weather related disasters are selected

B. Both features are binary features

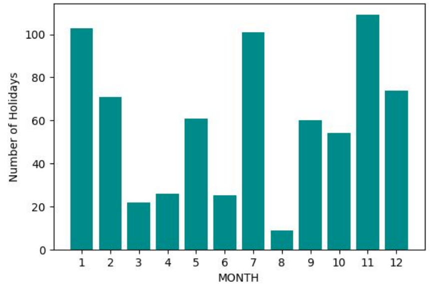
Time-Based Feature Generation

Time-based Features
Holiday & Disaster

Given origination state, how recent was the last holiday/disaster?

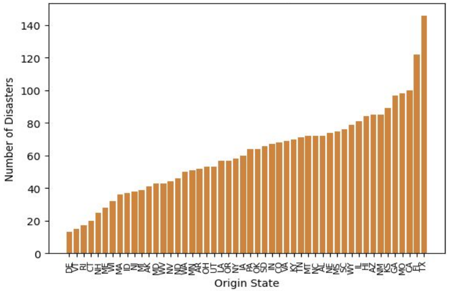
Given origination state, how many holidays/disasters were per month?

US Holiday Count by Month



- ➔ Holiday has seasonal effects. Jan, Feb, Nov & Dec have more holidays
- ➔ More holidays refer to high volume of flights

Weather Disaster Count by State



- ➔ Texas, Florida, California seems to have more disasters
- ➔ More disasters normally leads to flight delays

Berkeley
UNIVERSITY OF CALIFORNIA

Feature Engineering and Top Features

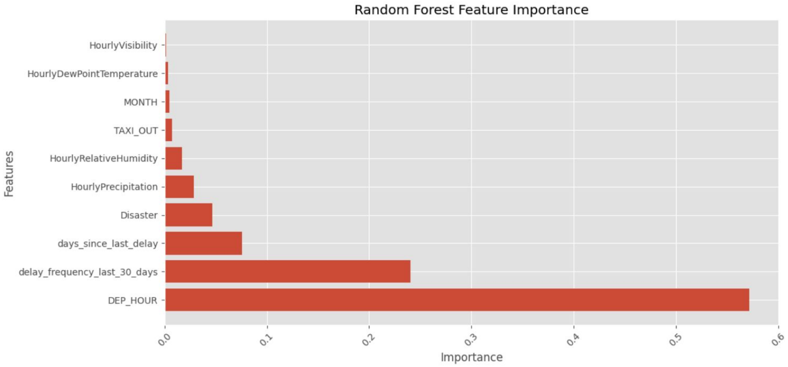
19 total features used

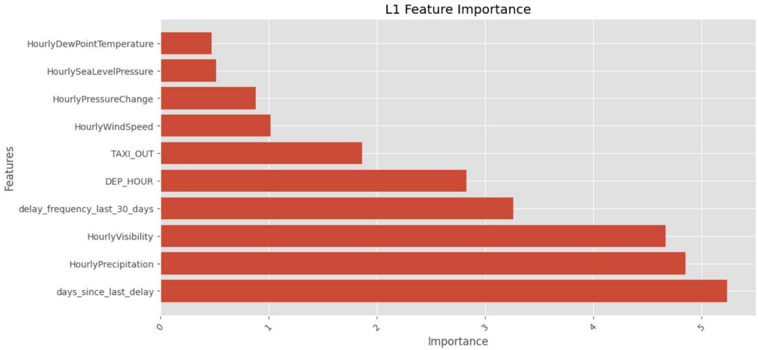
- ➔ 3 + 2 categorical features (Two for Holiday & Disaster) - Phase III
- ➔ 12 + 2 numerical features (2 time-based features created)

DEP_DEL15 as Label (Y)

- ➔ “1” representing delays exceeding 15 minutes
- ➔ “0” representing delays not exceeding 15 minutes or no delays happen.

Top Factors Selected





Berkeley
UNIVERSITY OF CALIFORNIA

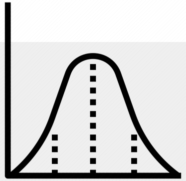
Data Cleaning

- Data type conversion
- Joining flight data with additional data
  - Weather
  - Disaster
  - Holiday
- Imputation with median value of the feature column

Berkeley
UNIVERSITY OF CALIFORNIA

Feature Transformation

➔ Normalization
- ◆ Re-scale each feature into [0, 1].



➔ Train-test split
- ◆ First 3 quarters as training set
- ◆ Last quarter as test set

Overview of Modeling Pipelines

Model A Logistic Regression (Baseline)

Model C Multilayer Perceptron

Model B Gradient Boosting Classifier

Pipeline 1: Model & Performance

- Logistic regression (Baseline)

AUC	0.676
Accuracy	0.839
Precision	0.839
Recall	0.999
F1	0.766

Pipeline 2: Model & Performance

- Gradient Boosting Classifier

AUC	0.673
Accuracy	0.84
Precision	0.784
Recall	0.84
F1	0.793

Pipeline 3: Model & Performance

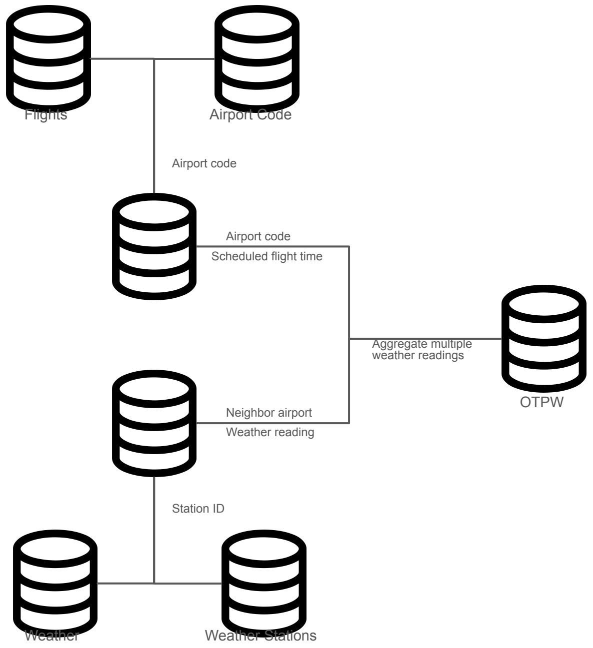
- Multi-layer Perceptron

AUC	0.52
Accuracy	0.85
Precision	0.85
Recall	0.98
F1	0.79

Phase II Extra Credits

Full Join of Dataset, EDA, Feature Processing, Modeling

- Flights dataset
  - Contain flight information, including delays
- Airport code dataset
  - Has the mapping of the airport code of the Flights dataset to the airport code in the Weather Stations dataset
- Weather dataset
  - Contains weather readings from weather stations
- Weather stations dataset
  - Contains the weather stations in close proximity to airports. Has the station ID of the weather data in the Weather dataset



Berkeley
UNIVERSITY OF CALIFORNIA

Phase III Extra Credits

Clean Dataset Preparation for Flight Delays & Weather for Years 2020 - 2022

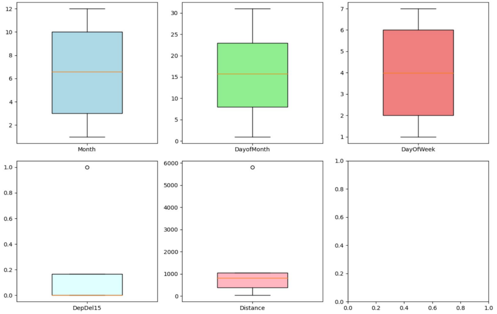
Flight Delays Cleaned Dataset Summary

Data Cleaning

- ➔ **Step 1:** Select useful columns
- ➔ **Step 2:** Drop missing values based on a subset of columns
- ➔ **Step 3:** Drop duplicates
- ➔ **Step 4:** Set data types for each columns

- **Column Numbers:** 65
- **Time Range:** 2020Y - 2022Y
- **Records:** 16809806

Cleaned Data Summary



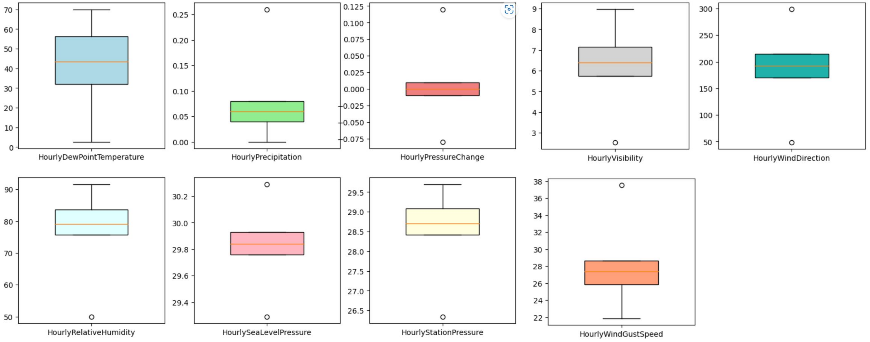
Weather Cleaned Dataset Summary

Data Cleaning

- ➔ **Step 1:** Select useful columns
- ➔ **Step 2:** Drop missing values based on a subset of columns
- ➔ **Step 3:** Drop duplicates
- ➔ **Step 4:** Set data types for each columns
- ➔ **Step 5:** Aggregate data by date

- **Column Numbers:** 26
- **Time Range:** 2020Y - 2022Y
- **Records:** 1096

Cleaned Data Summary



Berkeley
UNIVERSITY OF CALIFORNIA

Phase III Extra Credits

A different Neural Network

Spark MLlib MLP has limited deep learning functionality

Instead, we used PyTorch to implement a neural network

- 2 hidden layers
- Relu activation functions at the hidden layers
- Sigmoid function at the output layer

However, the performance did not improve over the baseline

Berkeley
UNIVERSITY OF CALIFORNIA

Project Conclusions

- All models converged in accuracy score with varying AUCs.
    - Suggest that there may be some variance unexplained by our features
- Our baseline model (logistic regression) is the most well-rounded model
    - Best performance
    - Simple structure
    - Fast implementation
- We've only tested on the 2015 OTPW dataset, and experiments on dataset of longer period is recommended.

References

- R. Nigam and K. Govinda, "Cloud based flight delay prediction using logistic regression," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2017, pp. 662-667, doi: 10.1109/ISS1.2017.8389254.
- Yuemin Tang. 2021. Airline Flight Delay Prediction Using Machine Learning Models. In 2021 5th International Conference on E-Business and Internet (ICEBI 2021), October 15-17, 2021, Singapore, Singapore. ACM, New York, NY, USA, 7 Pages. https://doi.org/10.1145/3497701.3497725
- "RFM Analysis - Understanding Customer Behavior" by SuperOffice: https://www.superoffice.com/blog/rfm-analysis/
- "RFM (recency, frequency, monetary) analysis" by IBM: https://www.ibm.com/cloud/learn/rfm-analysis
- US Holiday Dataset: https://www.timeanddate.com/holidays/us/2015
- US Natural Disaster Dataset: https://www.ncdc.noaa.gov/stormevents/ftp.jsp

Thank you