Flight Delays Prediction

Datasci-261-team-7-1

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

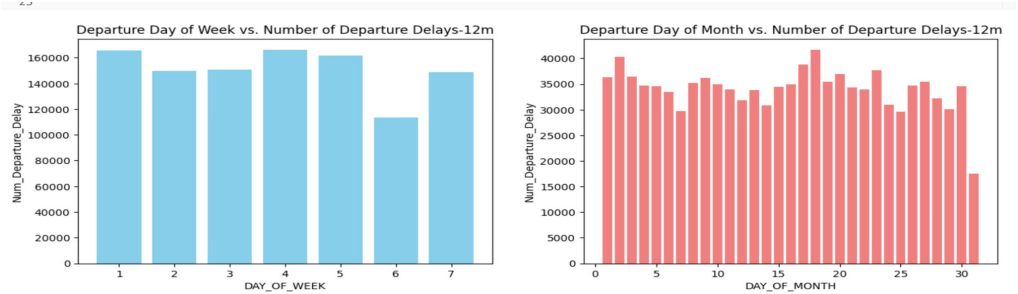
Exploratory Data Analysis (EDA)

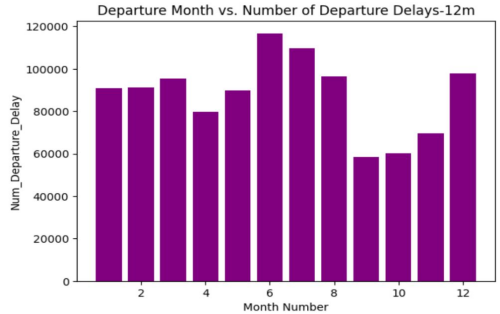
	path	name
1	dbfs:/mnt/mids-w261/HW5/	HW5/
2	dbfs:/mnt/mids-w261/OTPW_12M/	OTPW_12M/
3	dbfs:/mnt/mids-w261/OTPW_1D_CSV/	OTPW_1D_CSV/
4	dbfs:/mnt/mids-w261/OTPW_36M/	OTPW_36M/
5	dbfs:/mnt/mids-w261/OTPW_3M/	OTPW_3M/
6	dbfs:/mnt/mids-w261/OTPW_3M_2015.csv	OTPW_3M_2015.csv
7	dbfs:/mnt/mids-w261/OTPW_60M/	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/

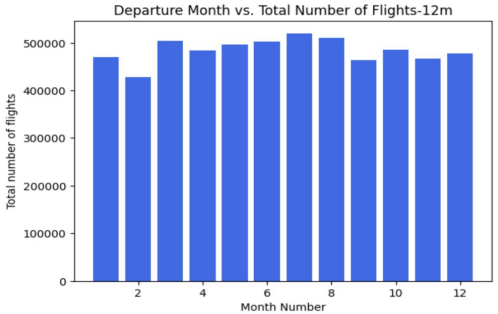

```
'CARRIER_DELAY','WEATHER_DELAY','NAS_DELAY','SECURITY_
DELAY','LATE_AIRCRAFT_DELAY'
```

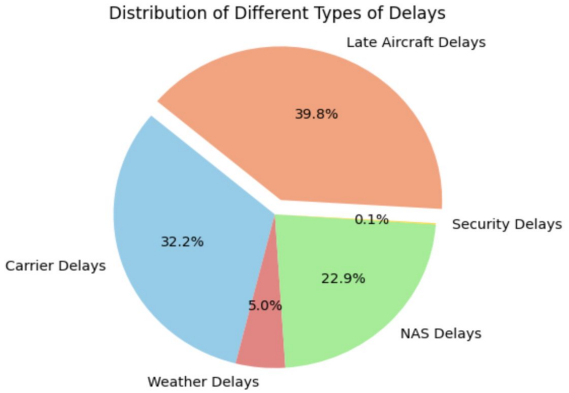
Property	Fact 12M	Fact 3M
Total number of rows	11623708	1401363
Total number of rows after removing duplicates	5811854	1401363
Number of dep delays >15 min	1055735	277302
% of dep delays >15 min	18.2%	19.8%
% of cancel	1.5%	3.0%
% of dep ontime	80.3%	77.2%
Date range start date	01/01/2015	01/01/2015
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

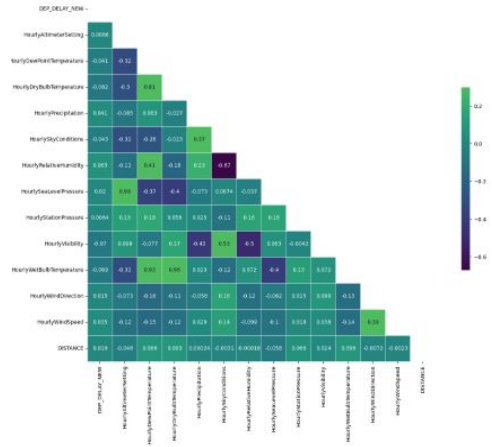
Exploratory Data Analysis (EDA)











Feature Preparation & Selection - Basic Features

20 total features used

- ➔ 7 categorical features
- ➔ 13 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.

Basic Factors Selected

Flight Factor

- Distance Numerical
- Taxi out time Numerical
- Origination Categorical
- Destination Categorical
- Carrier No. Categorical
- Flight Date Categorical

Longer Taxi out time will take more time and result in higher prob. of delays

Carrier is one of the Flight Delay reason (EDA result)

Temporal Factor

- Month Categorical
- Day of Month Categorical
- Day of Week Categorical

June ~ August might have more delays than other months (EDA result)

Climate Factor - Hourly

- Dew Point Temperature Numerical
- Precipitation Numerical
- Pressure Change Numerical
- Relative Humidity Numerical
- Sea Level Pressure Numerical
- Visibility Numerical
- Wind Direction Numerical
- Wind Gust Speed Numerical
- Wind Speed Numerical

Huge Pressure Change also contributes to flight delays

Bad visibility might contribute to flight delays

Feature Preparation & Selection- Time-based Feature

RFM

RFM stands for Recency, Frequency, and Monetary Value, which is a marketing analysis technique used to segment and target customers based on their transaction history. In our project, we introduced this concepts to create time-based features to enrich the feature data.

Take 'carrier' as the object to generate time-based features. The time interval we set is one month.

Recency	Frequency	Monetary
Example: How recent was the last flight delay for the same carrier with the same origin and destination	Example: How frequent the flight delay was per month for the same carrier with the same origin and destination	Example: What is the total flight distance per month for the same carrier with the same origin and destination
Current Delay Time - Previous Delay Time = Delay Recency (Delay recency is summarized in days)	**Summarize the number of delayed flights per month given group by carrier, origin and destination**	**Summarize the distance for each flight per month group by carrier, origin and destination**

Feature Engineering and select time-based features above

Feature Engineering

- Hour at which a flight takes off
  - Certain hours of a day tend to have more delays than others
- Two time-based features (RFM)
  - Recency
    - How recent was the last flight delay for the same carrier with the same origin and destination
  - Frequency
    - Relative frequency of delays for the same carrier with the same origin and destination in the past 30 days

Data Transformation

- Imputation
  - Filled missing numerical values with the median of the feature
  - Filled missing categorical values with “NA”
- Min-max scalar
  - Ensured all features on the same scale of [0, 1]
- One-hot encoding
  - For nominal categorical features
- PCA
  - One-hot encoding resulted in ~600 features, which makes dimensionality reduction necessary
- Train-test Split
  - Train: first 3 quarters
  - Test: last quarter

Machine Learning Pipeline

- **Logistic Regression** for binary classification.
- Time-series K-fold cross validation for hyper-parameter tuning
  1. Split the training set into K folds.
  2. At each K_n fold, combine the previous K_1, ..., K_(n-1) fold as training data to predict on the K_n fold data.
- Evaluated with Accuracy, Recall, Precision and F1-score.
- Evaluated on **last quarter** of the 12-month OTPW dataset.
- **Precision** is most important because
  - False positive is more costly
  - We don't want to inform passengers that the flight will be late, causing them to arrive late at the airport, yet the flight arrives on time.

Performance Report

AUC	0.676
Accuracy	0.839
Precision	0.839
Recall	0.999
F1	0.766

Next Steps

- More thorough EDA
  - More feature engineering
  - Rigorous feature selection with regularization, etc
- Create more models against the baseline
  - Random forest
  - Xgboost
  - Multi-layer perceptron
  - Possibly ensemble models
- Final presentation / Report

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

Thank You!