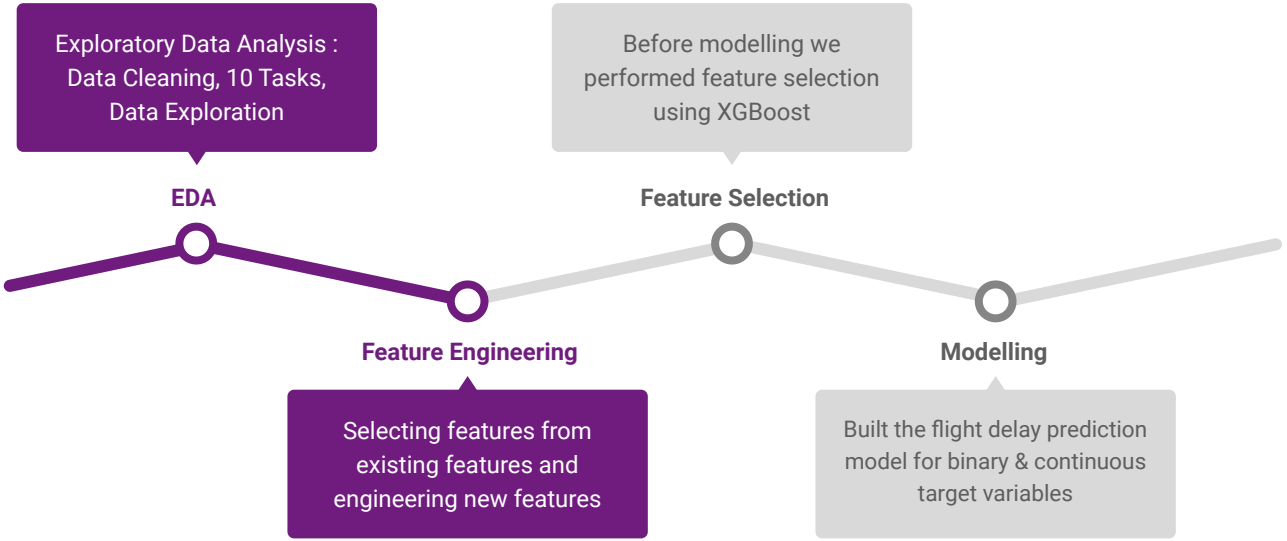
Predicting Flight Delays

Creating value through data-enabled travel technology

Mid-Term Project

Isaac, Pavel, Nakul

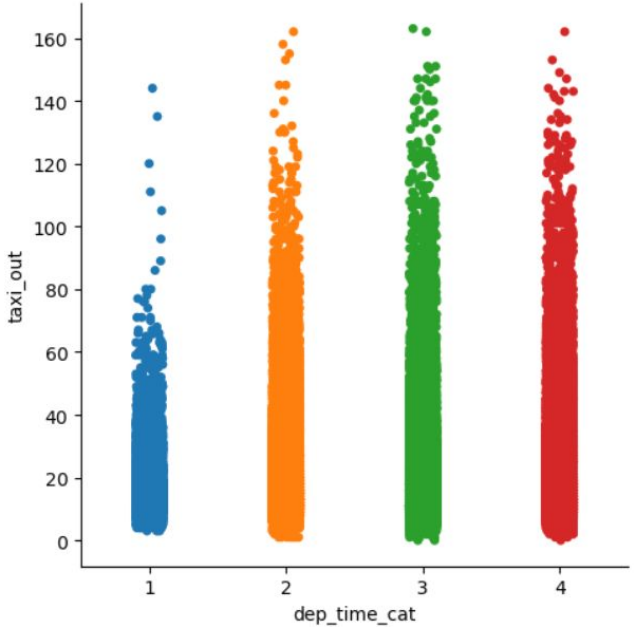
Project Flow



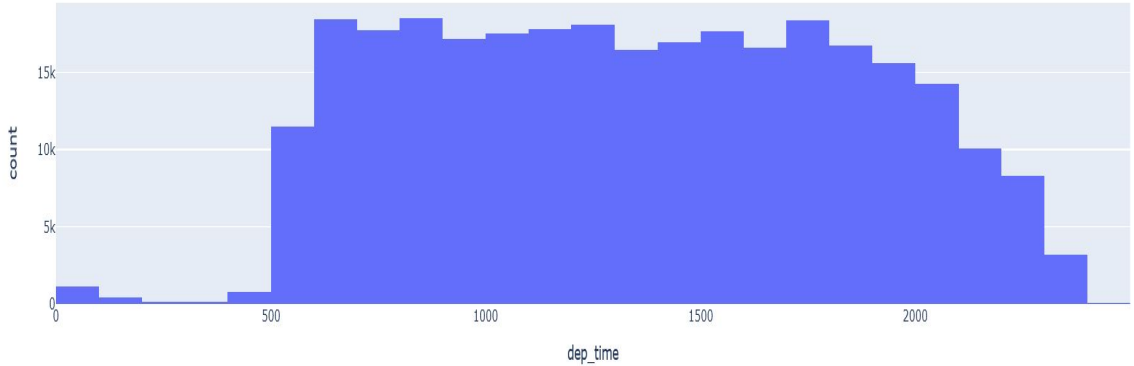
Insights and Relationships

- List of top 10 airports in terms of passenger traffic is not the same as that of top 10 airports in terms of flights traffic [2018-19 data window]
- 'LONG', 'SHORT', 'MEDIUM' haul flights take off at 17:00, 7:00 and 12:00 hrs respectively
- Difference between speeds of flights when delay was less than 10 (including negative {which means in advance} and when delay was over 30 was not significant
- Mean taxi outbound time are less during late night (12:00AM - 6AM)
    - Coincides with levels of traffic

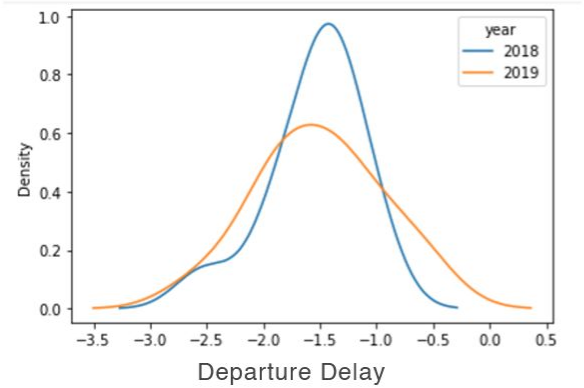
Taxi outbound times for different times of day

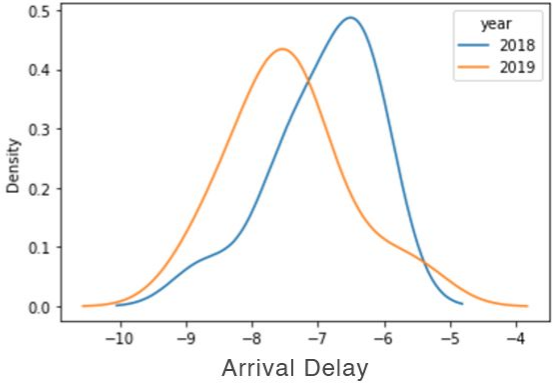


Departure Time Histogram



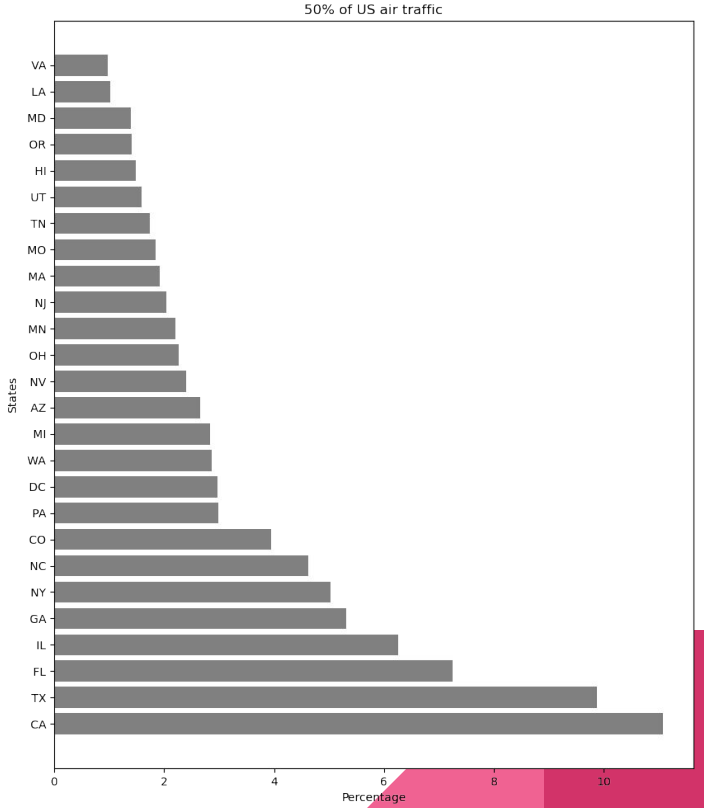
Departure Vs Arrival Delay





50 % of all US traffic by destination

CA	11.07%
TX	9.87%
FL	7.25%
IL	6.25%
GA	5.31%



Results (Regression)

Decision Tree Modeling

- Accuracy: 0.0200
- Accuracy: 0.0292 (criterion="entropy", max_depth=3)

Random Forest Modeling

- Model accuracy score with 10 decision-trees : 0.0202
- Precision: 0.4300
- Recall: 0.2708

XGBoost Modeling

- RMSE: 49.77

Results (Classification)

Decision Tree Modeling

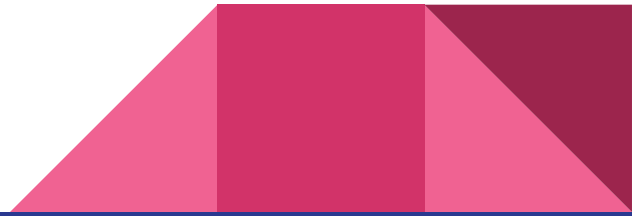
- Accuracy: 0.5750
- Precision: 0.3905
- Recall: 0.3819
- Accuracy: 0.6500 (criterion="entropy", max_depth=3)

Random Forest Modeling

- Model accuracy score with 10 decision-trees : 0.6191
- Precision: 0.4300
- Recall: 0.2708
- Model accuracy score with 100 decision-trees : 0.6270
- Precision: 0.4477
- Recall: 0.2809

XGBoost Modeling

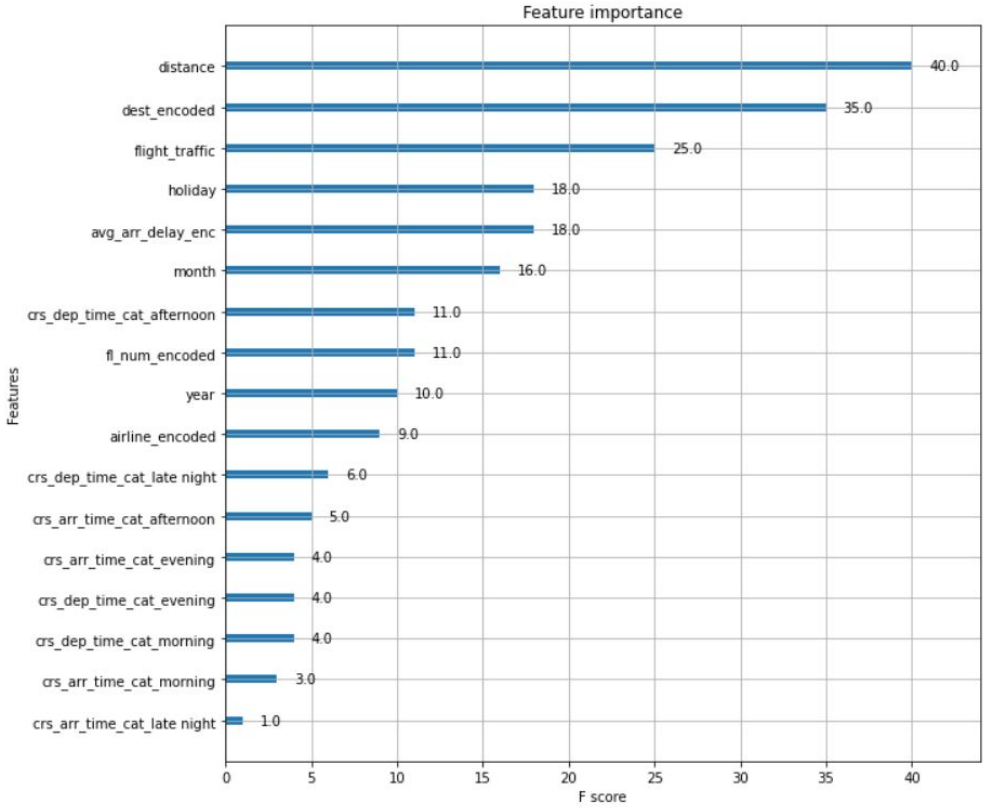
- RMSE: 49.77



Feature Importance (Classification)

Interesting Features: Holidays, use of passenger and flight traffic as new features

Most Important Features: Distance, flight traffic, holiday (whether a given day was a holiday or not), flight number, year, airline, departure time



Challenges

Explain the biggest challenges:

1. API limits (eg: Weather API - 1000 queries allowed from free account)
2. Granularity of Passenger data being monthly and that for Flights data being daily

What would you do if you have a bit more time:

1. Spend more time on Feature Engineering. For example we wanted to calculate the number of flights in an hour just before and just after the flight into consideration. While computing this we were running out of RAM hence for finding a suitable alternative some more time would have had been helpful.
2. Try more combinations of modelling techniques
3. Attempt to optimize regression results