Analysis and prediction of US flights delay

by Korchagina Evgeniya

(CEBD 1260)

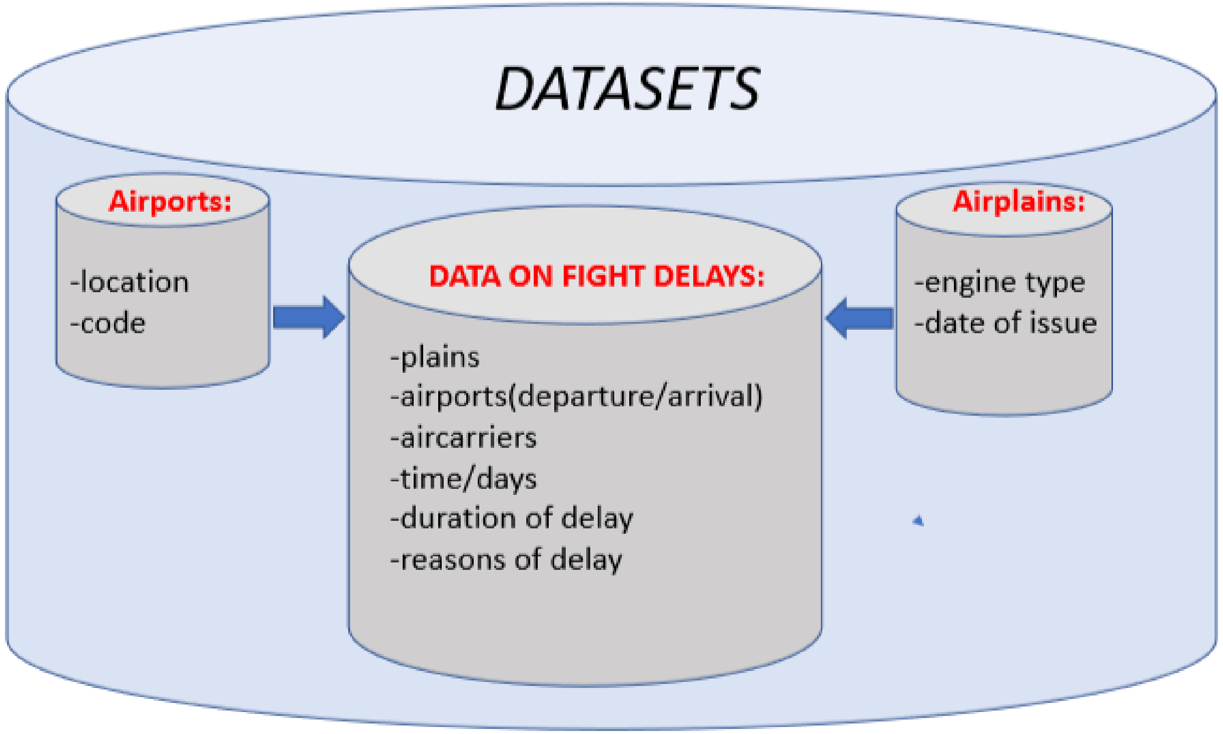
Datasets

Passengers

-search (day, time, delay, airport, etc.)
-history
-choice of suitable flight

Aircompanies

-opportunities to improve service
-decision on contracts with aircarriers
-decision on better using different airports



DATASETS

Airports:

-location
-code

DATA ON FIGHT DELAYS:

-plains
-airports(departure/arrival)
-aircarriers
-time/days
-duration of delay
-reasons of delay

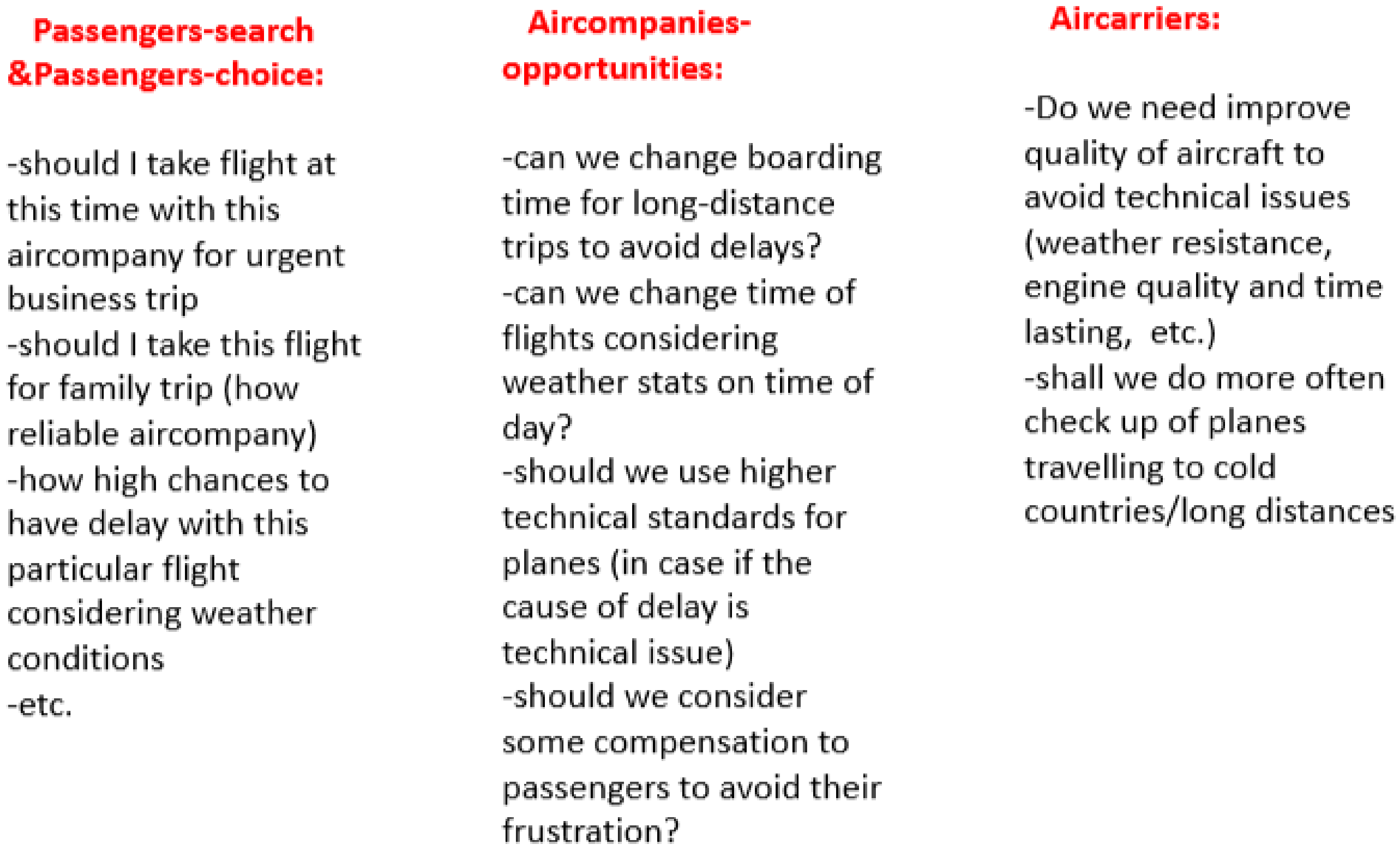
Airplains:

-engine type
-date of issue

Aircarriers

-history
-info
-trend and relationship with plane

Some use cases



Passengers-search &Passengers-choice:

-should I take flight at this time with this aircompany for urgent business trip

-should I take this flight for family trip (how reliable aircompany)

-how high chances to have delay with this particular flight considering weather conditions

-etc.

Aircompanies-opportunities:

-can we change boarding time for long-distance trips to avoid delays?

-can we change time of flights considering weather stats on time of day?

-should we use higher technical standards for planes (in case if the cause of delay is technical issue)

-should we consider some compensation to passengers to avoid their frustration?

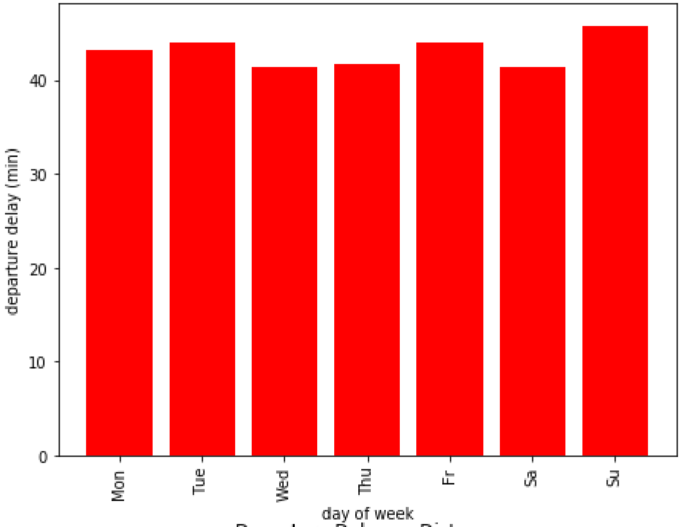
Aircarriers:

-Do we need improve quality of aircraft to avoid technical issues (weather resistance, engine quality and time lasting, etc.)

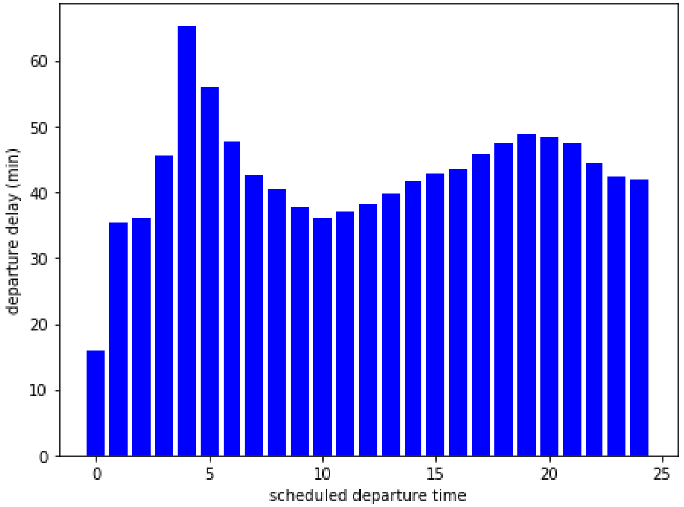
-shall we do more often check up of planes travelling to cold countries/long distances

Preprocessing and EDA

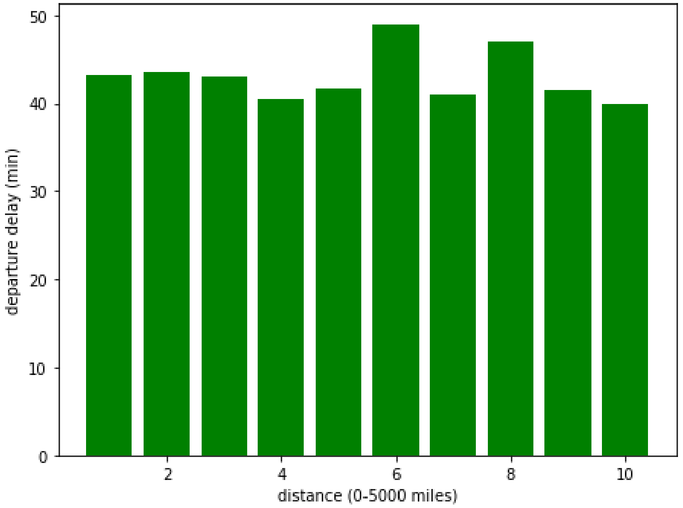
Departure Delay vs Dday of Week



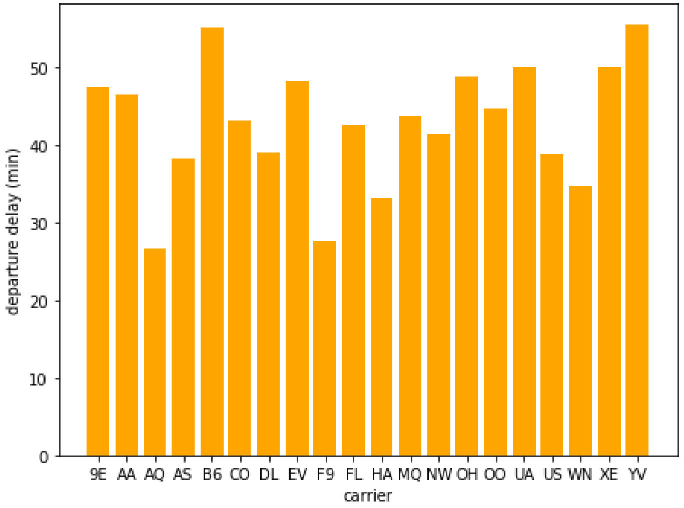
Departure Delay vs Sheduled Departure Time



Departure Delay vs Distance



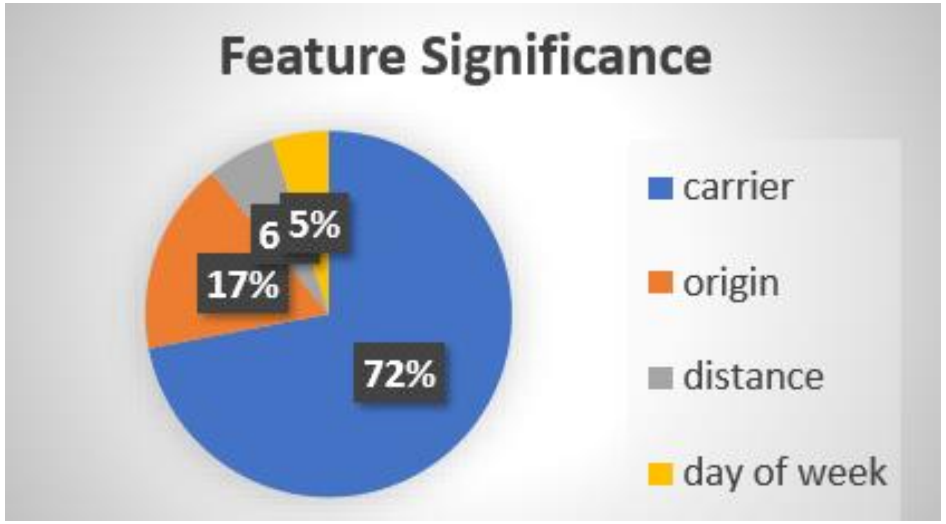
Departure Delay vs Carrier



Based on EDA:

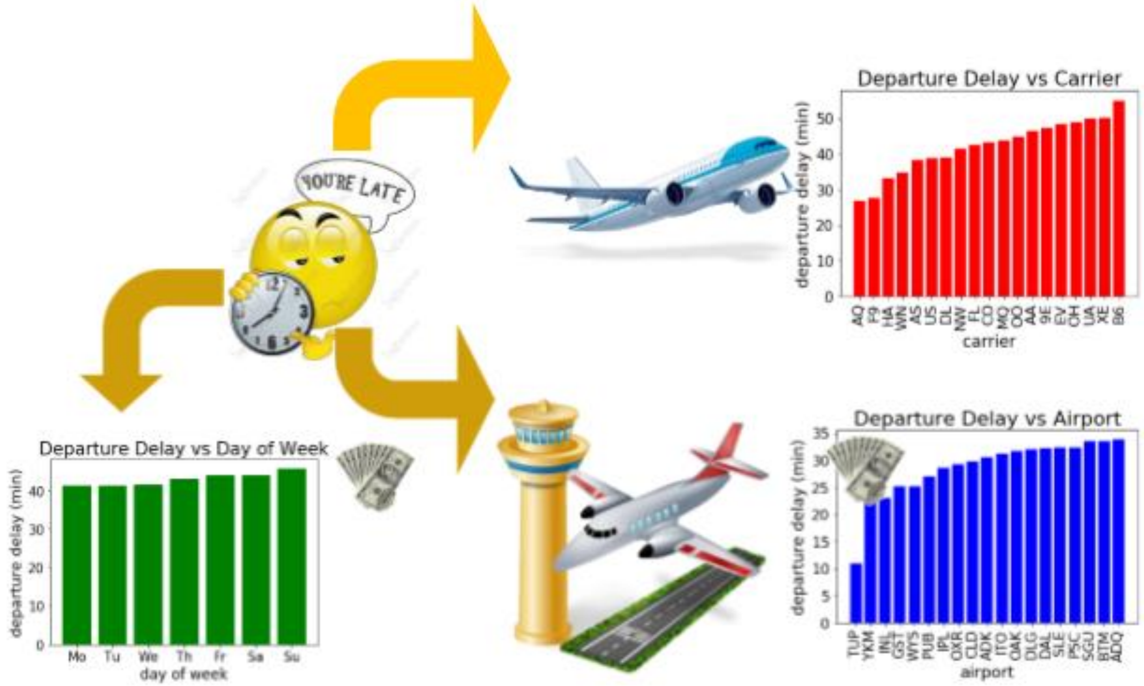
Current project will be focused on business use case concerning passengers search of flights covering some specific criteria. Result of data analysis suppousably will lead to an application that will predict delay time based on carrier, distance of flight, time of departure, airport of origin, tail number and day of week. Modeling will be based on regression analysis of these variables.

Technical and business perspective



DecisionTree Regressor and RandomForestRegressor (used for categorical data) showed the same absolute errors and standard deviations. Whiskers plot were visually undistinguishable.

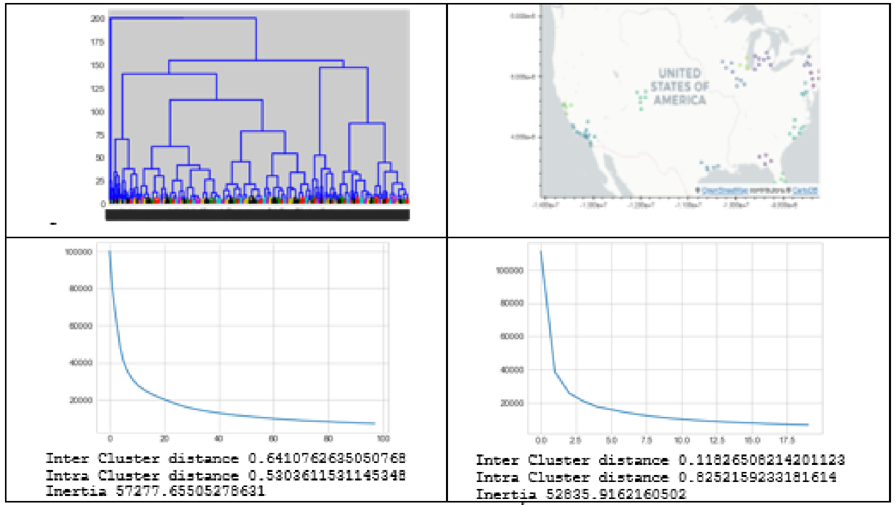
>>any of the methods can be applied as a regression model



Scheme1. *Scheme representing the main contributing factors to flight delay. The lower plots indicate money, as a factor affecting bar plot tendency.*

Here we are focusing on prediction of possible flight delay based on the flight features (from passenger' perspective).

Technical challenges

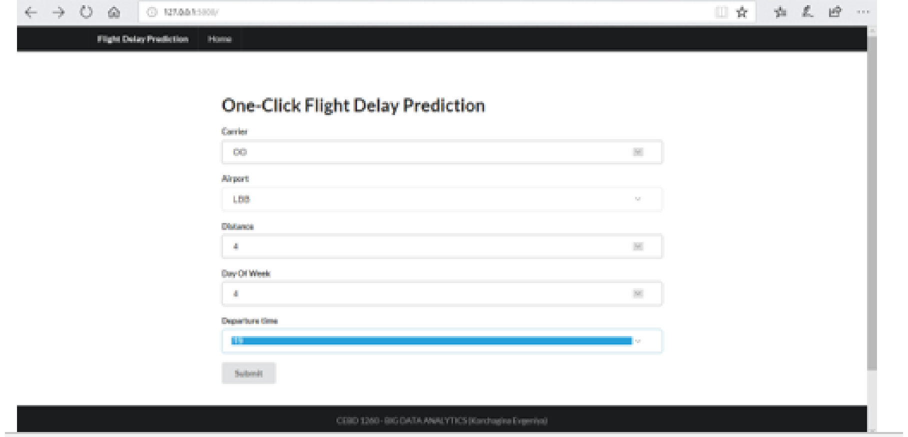


```
Inter Cluster distance 0.6410762635050768
Intra Cluster distance 0.5303611531145348
Inertia 57277.65505278631
```

```
Inter Cluster distance 0.11826508214201123
Intra Cluster distance 0.8252159233181614
Inertia 52835.9162160502
```

Fig.1. Top left – dendrograms of hierarchical clustering; top right – map of US airport created via DBSCAN; bottom left – inertia of hierarchical clustering; bottom right – inertia of K-means clustering mechanism.

APP:



One-Click Flight Delay Prediction

Carrier: OO

Airport: LBB

Distance: 4

Day Of Week: 4

Departure time: 15

CEBD 1260 - BIG DATA ANALYTICS (Korchagina Evgeniya)

- *Big size of dataset*

- *Parsing categorical value to dummy numerical ones*

- *time consuming regression processing on big set of data (impossible to apply models with categorical data of high variety)*

- *Hierarchical clustering and K-means do not for chosen data*

- *Due to the lack of time DBSCAN was used only for simple practice*

- *APP to be finished*