CEDB 1260
Big Data Analytics

A regression model to predict on-time and delayed flights

by Arwa Sheraky & Tiffany Eversley

Mission:

To predict average delay times based on select attributing factors

The problem

With flight information being readily available online, certain factors such as airline carrier, airport location, and/or historical delay and cancellation details, may be expected to increasingly influence passenger travel decisions.

Data set

This 2015 dataset summarizes US airline flight delay and cancellation information as collected and published by the DOT's Bureau of Transportation Statistic.

Attributes: Drawing airport and airline information from two additional datasets helped expand the original source file by pulling from, and merging, relevant attributes. The dataset is now characterized by 28 representative features and includes over a million instances. Features include airport origin, time of the flight,, actual and scheduled departure times, arrival times, flight number, as well as cancellation and delay reason.

Approach

Data Cleaning

Outliers were removed, missing values filled in, columns renamed, duplicates and unused columns removed and csv files were merged.

Visualization

Attributes were plotted against delay reason categories and average delay time to identify trends and draw conclusions

Modelling

A regression model was chosen to predict average delay time based on size of the dataset and desired output values.

Prediction App/API

A simple web application was created to deploy the machine learning model.

Data Cleaning

Cleaning involved:

- Merging columns
- Removing irrelevant and duplicated columns
- Renaming columns
- Change date and time format *convert from 'HHMM' string to datetime.time*
- Replace Cancellation Reason with a description
- Remove missing values
- Remove outliers

```
------- Main Dataset, Flights -------
(5819079, 31)
Index(['YEAR', 'MONTH', 'DAY', 'DAY_OF_WEEK', 'AIRLINE', 'FLIGHT_NUMBER',
       'TAIL_NUMBER', 'ORIGIN_AIRPORT', 'DESTINATION_AIRPORT',
       'SCHEDULED_DEPARTURE', 'DEPARTURE_TIME', 'DEPARTURE_DELAY', 'TAXI_OUT',
       'WHEELS_OFF', 'SCHEDULED_TIME', 'ELAPSED_TIME', 'AIR_TIME', 'DISTANCE',
       'WHEELS_ON', 'TAXI_IN', 'SCHEDULED_ARRIVAL', 'ARRIVAL_TIME',
       'ARRIVAL_DELAY', 'DIVERTED', 'CANCELLED', 'CANCELLATION_REASON',
       'AIR_SYSTEM_DELAY', 'SECURITY_DELAY', 'AIRLINE_DELAY',
       'LATE_AIRCRAFT_DELAY', 'WEATHER_DELAY'],
      dtype='object')
   YEAR  MONTH  DAY  DAY_OF_WEEK AIRLINE  FLIGHT_NUMBER TAIL_NUMBER  \
0  2015      1    1            4      AS             98      N407AS
1  2015      1    1            4      AA           2336      N3KUAA
2  2015      1    1            4      US            840      N171US
3  2015      1    1            4      AA            258      N3HYAA
4  2015      1    1            4      AS            135      N527AS

  ORIGIN_AIRPORT DESTINATION_AIRPORT  SCHEDULED_DEPARTURE      ...        \
0            ANC                 SEA                    5      ...
1            LAX                 PBI                   10      ...
2            SFO                 CLT                   20      ...
3            LAX                 MIA                   20      ...
4            SEA                 ANC                   25      ...

   ARRIVAL_TIME  ARRIVAL_DELAY  DIVERTED  CANCELLED  CANCELLATION_REASON  \
0        408.00         -22.00         0          0                  NaN
1        741.00          -9.00         0          0                  NaN
2        811.00           5.00         0          0                  NaN
3        756.00          -9.00         0          0                  NaN
4        259.00         -21.00         0          0                  NaN

   AIR_SYSTEM_DELAY  SECURITY_DELAY  AIRLINE_DELAY  LATE_AIRCRAFT_DELAY  \
0               nan             nan            nan                  nan
1               nan             nan            nan                  nan
2               nan             nan            nan                  nan
3               nan             nan            nan                  nan
4               nan             nan            nan                  nan

   WEATHER_DELAY
0            nan
1            nan
2            nan
3            nan
4            nan

[5 rows x 31 columns]
```

****Before cleaning****

Data Cleaning

Cleaning involved:

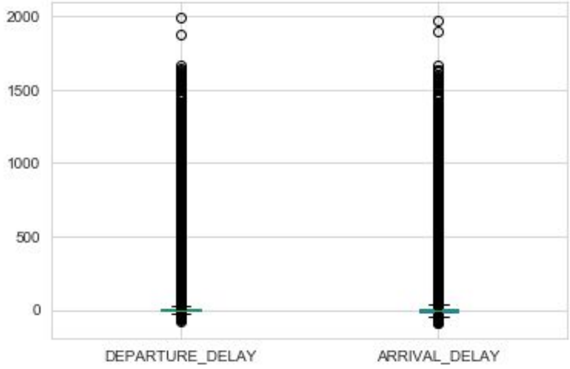
- Merging columns
- Removing irrelevant and duplicated columns
- Renaming columns
- Change date and time format *convert from 'HHMM' string to datetime.time*
- Replace Cancellation Reason with a description
- Remove missing values
- Remove outliers

```
#Replace cancellation reason with meaningful values
df_delayed_flights["CANCELLATION_REASON"].replace({'A':'Airline',
                                                   'B':'Weather',
                                                   'C':'National Air System',
                                                   'D':'Security'}, inplace=True)

df_delayed_flights["CANCELLATION_REASON"].value_counts()
```

```
Weather                48851
Airline                25262
National Air System    15749
Security                  22
Name: CANCELLATION_REASON, dtype: int64
```

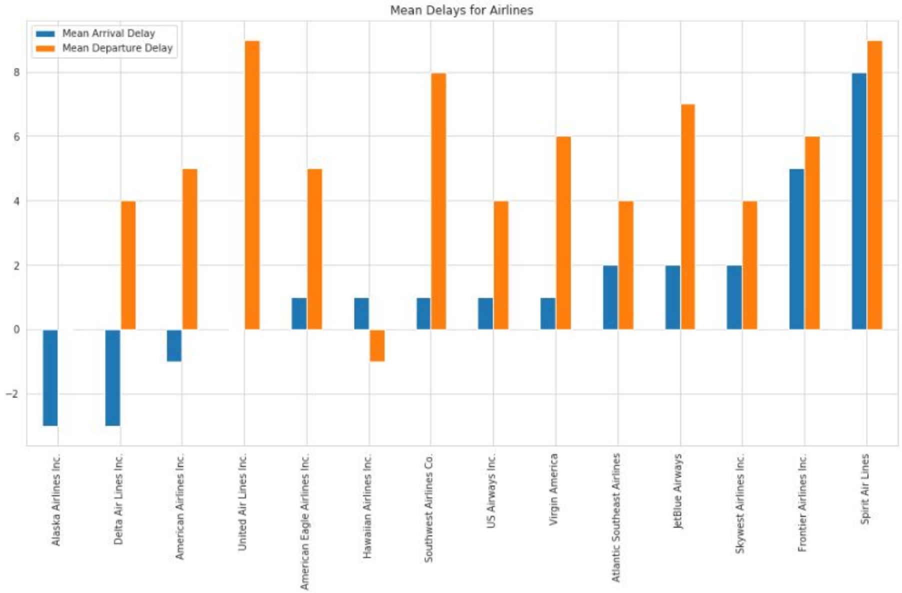
```
df_delayed_flights[["DEPARTURE_DELAY","ARRIVAL_DELAY"]].plot.box()
plt.show()
```



Visualization

Exploring the data involved:

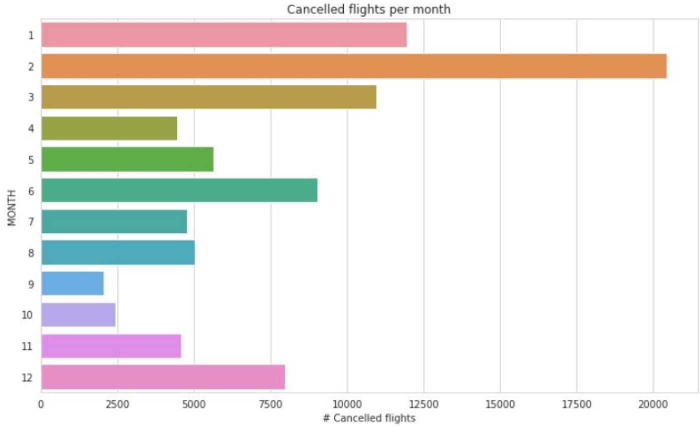
- Plotting numeric and categorical variables
- Answer specific business questions relevant to the data set such as:
    - What is the average delay for each airline ?
    - What is the average arrival and departure delay times based on airport ?
    - What is the impact of the weather on flights?

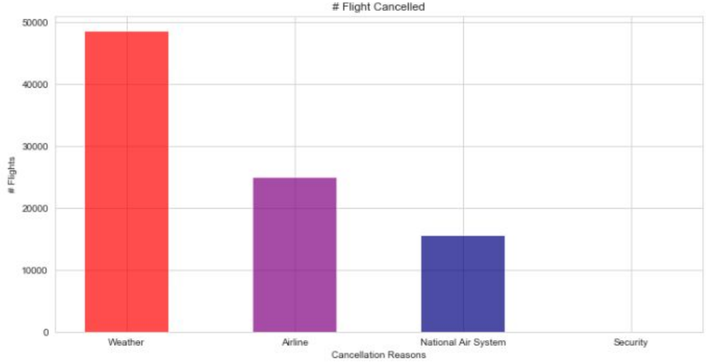


Visualization

Exploring the data involved:

- Plotting numeric and categorical variables
- Answer specific business questions relevant to the data set such as:
  - What is the average delay for each airline ?
  - What is the average arrival and departure delay times based on airport ?
  - What is the impact of the weather on flights?



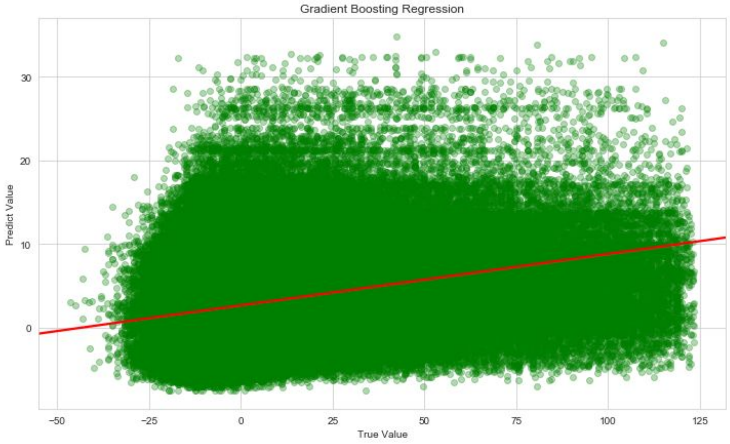


Modelling

After pre-processing, a subset of the data was then split in two, a training and testing set.

We examined and compared the performances of the KNN , Random Forest and Gradient Boosting classifiers. Among the 4 classifiers, Gradient Boosting produced the most reliable prediction model with the lowest root mean square error: RMSE 20.38

Identifying the most important features allowed us to work on improving the model by focusing on the important variables and removing x-variables that were deemed insignificant.




```
features_imp_0001 = features_imp[features_imp[1] > 0.0001]
features_imp_0001
```

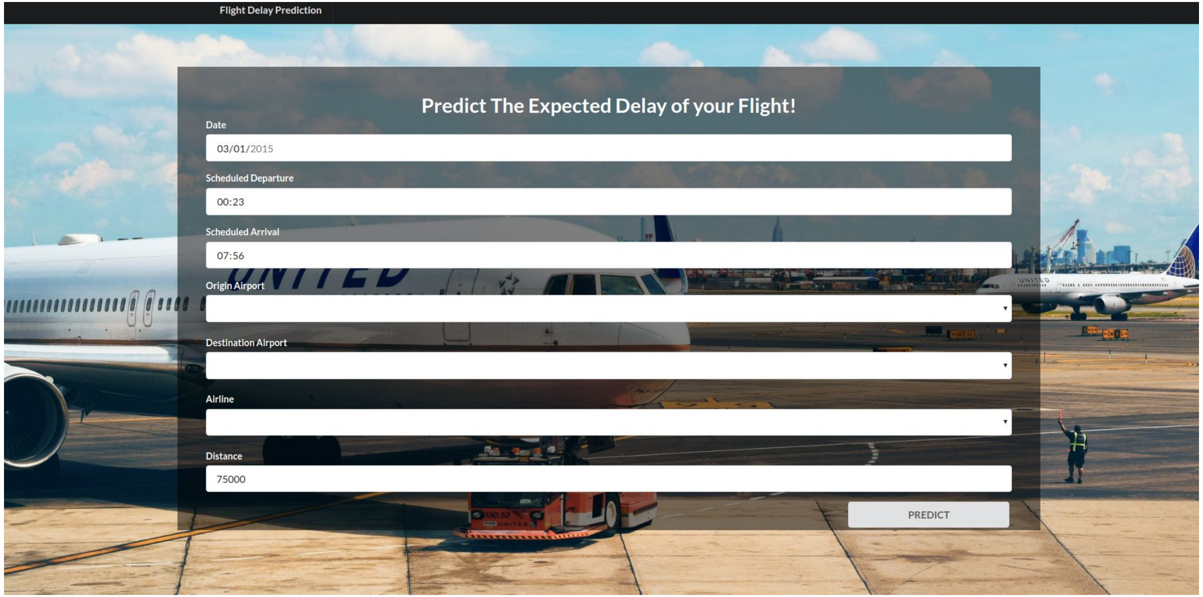
	0	1
1	DATE	0.41
0	SCHEDULED_DEPARTURE	0.32
2	SCHEDULED_ARRIVAL	0.08
3	AIRLINE_NAME_Southwest Airlines Co.	0.06
4	AIRLINE_NAME_Delta Air Lines Inc.	0.04
5	AIRLINE_NAME_Spirit Air Lines	0.02
6	MONTH_6	0.02
7	AIRLINE_NAME_Alaska Airlines Inc.	0.02
10	ORIGIN_AC_ORD	0.01
9	AIRLINE_NAME_JetBlue Airways	0.01
13	ORIGIN_STATE_IL	0.01
12	DESTINATION_AC_LGA	0.01
8	MONTH_2	0.01
14	ORIGIN_AC_DFW	0.01
11	DEST_STATE_NY	0.00
15	ORIGIN_AC_SEA	0.00

Results

Flask, a python based microframework, was used to deploy our chosen model.

To collect the data an index.html form was created containing the different attributes of the model.

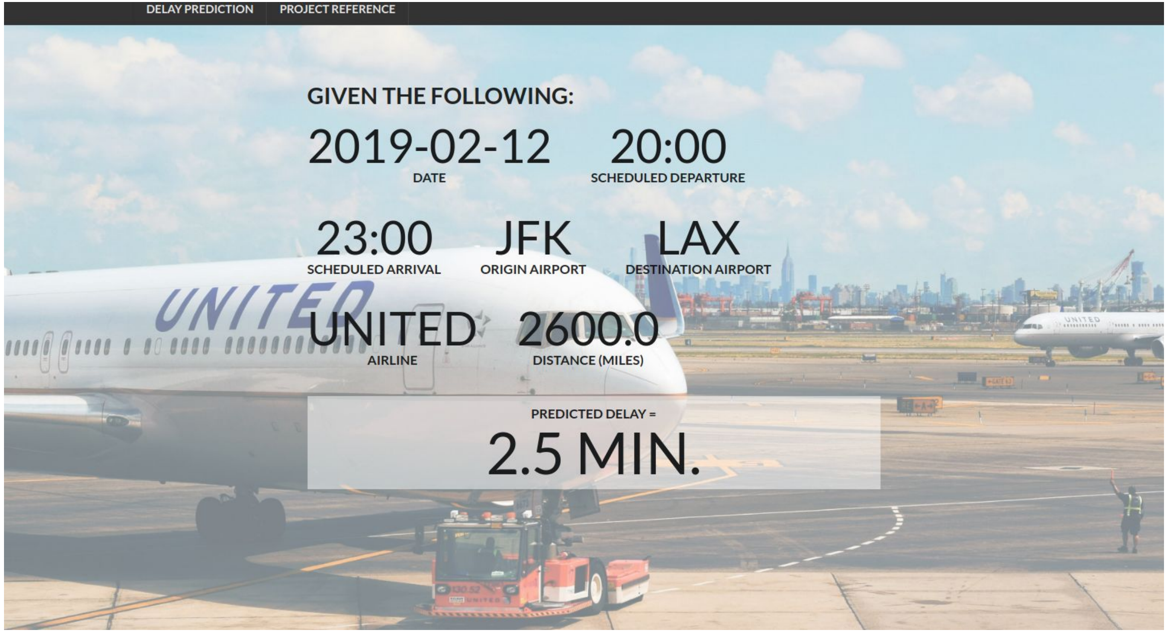
Upon completing the index.html form the predicted value for flight delay time will be calculated based on the model file we created.



Results

The model is then able to be used to predict new data.

https://predict-flight-delay.herokuapp.com/



Next Steps

Improvements

The following steps were identified as areas which the model could be improved:

- Include complementary data from sin datasets to increase significance of important features including weather details, aircraft characteristics, IATA delay codes etc..
- Try different subsample values with lower learning rates and higher number of trees (include cross validation to prevent overfitting).

```
GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse', init=None,
             learning_rate=0.1, loss='ls', max_depth=3, max_features=None,
             max_leaf_nodes=None, min_impurity_decrease=0.0,
             min_impurity_split=None, min_samples_leaf=1,
             min_samples_split=2, min_weight_fraction_leaf=0.0,
             n_estimators=100, n_iter_no_change=None, presort='auto',
             random_state=None, subsample=1.0, tol=0.0001,
             validation_fraction=0.1, verbose=0, warm_start=False)
```

Business Opportunities

- Understand whether certain airports are better equipped to deal with extreme weather conditions.
- Determine which time frames are the most at risk for delays and cancellations for the months that experience the most delays (February).
- Optimize flight departure times based on ideal time frames.
- Price ticket sales according to cancellation and delay likelihood.
- Understand whether the seasonal increase in flight delays is due to higher flight traffic.
- Determine whether crew availability is adjusted based on higher flight traffic.
- Determine whether the airlines with the highest ratio of delays is due to the higher volume of flights and similarly if the opposite shows to be true for airlines with smaller flight network.