Ban Qu

Final Report:
U.S. Flights Delay Analysis

Problem Statement

Flight delays and cancellations are always a major topic in air travel service. The Bureau of Transportation Statistics (BTS) has been recording airline customer data to help decision makers better understand statistics on transportation. To be specific, the goal is to improve operational efficiency, to examine emerging topics, and to create relevant and timely information that foster understanding of transportation and its transformational role in society.

The present study contains a 2015 dataset with 28 features that could be used to predict a potential flight delay in U.S. By using and analysing the data via Python and Tableau, various visualization and machine learning models are deployed to find patterns and better predict unseen delays.

The final dashboard of Tableau enables an overview summary of delayed information that involves an interaction among cities, airlines, and time. Machine learning techniques involves K-Nearest Neighbor, Logistic Regression, Random Forest, Support Vector Machine, Naïve Bayes, and Gradient Boosting. The best performances were among KNN, Random Forest, and Gradient Boosting, and the latter two agree on the most predicting feature of arrival delay of U.S. flights as departure delay. Air System delay is the second main reason for causing arrival delays.

Data Wrangling

The raw dataset involves 3 different files: airlines, airports, and flights. Airline data contains 14 unique airlines and their codes, and the table contains no missing data. Airports data contains 322 rows and 7 columns that are related to locations of airports. There are 3 missing values on latitude and longitude, respectively, and therefore removed from this dataset.

Flights data contains 5819079 rows and 31 columns. “ARRIVAL_TIME” and “SCHEDULED_ARRIVAL” are removed because “ARRIVAL_DELAY” simply summarizes these two features. Similarly, “SCHEDULED_DEPARTURE” and “DEPARTURE_TIME” are removed because “DEPARTURE_DELAY” summarizes the two. Columns not interested in the current study are also dropped: “TAXI_OUT”, “TAXI_IN”, “WHEELS_OFF”, “WHEELS_ON”, “AIR_TIME”, “SCHDULED_TIME”, and “ELAPSED_TIME”. Missing values under “TAIL_NUMBER” are dropped since missed aircraft identifier rows are useless.

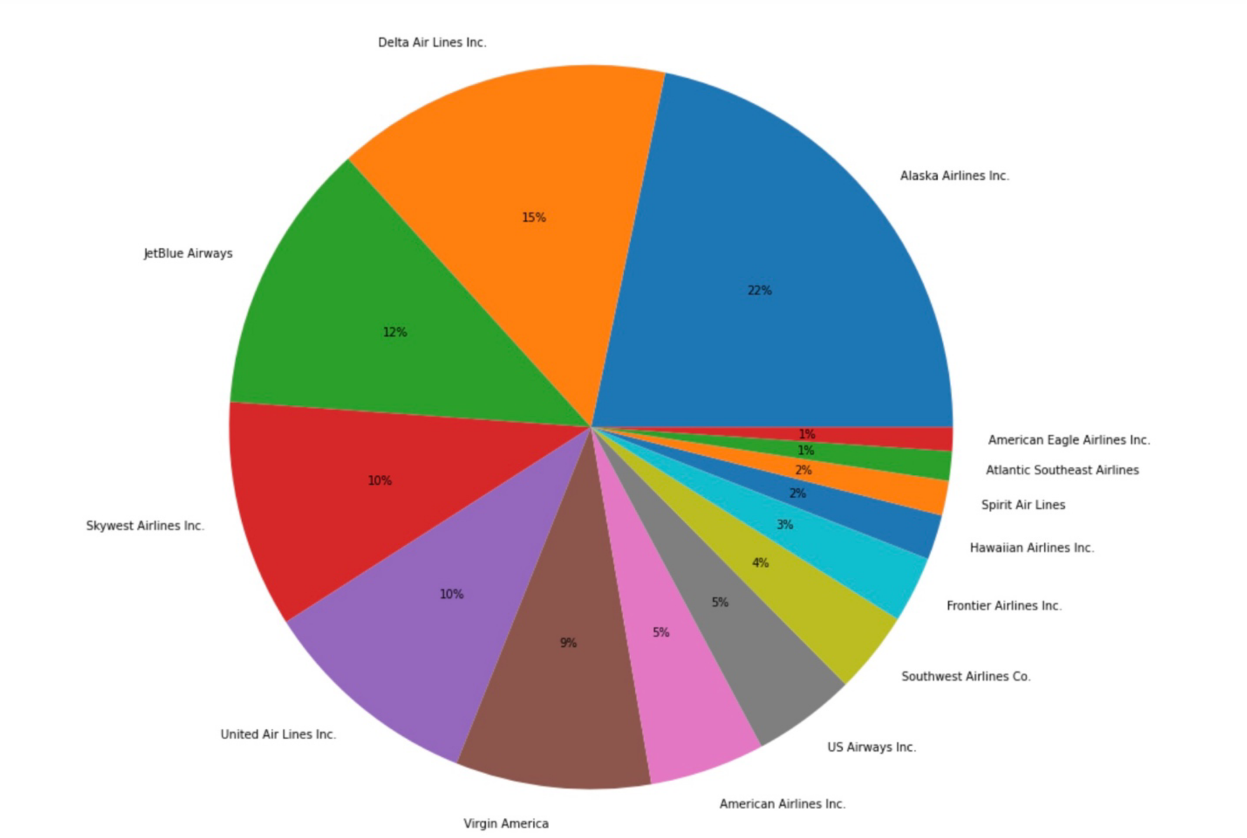
A large portion of missing values are found in “CANCELLATION_REASON”, “WEATHER_DELAY”, “LATE_AIRCRAFT_DELAY”, “AIRLINE_DELAY”, “SECURITY_DELAY”, and “AIR_SYSTEM_DELAY”. In this case, missing values under these columns are filled by zeros because a closer look at when some delays equal to zeros indicates that there’s no impact on the flight by these data that causes a delay. For “CANCELATION_REASON”, a relabelling is made that transforms “A” (Airline/Carrier), “B” (Weather), “C” (National Air System), “D” (Security) into “1”, “2”, “3”, and “4” respectively.

Three datasets are then merged based on an inner join of “IATA_CODE” that are unique identifiers of three tables. A check of duplicated columns indicates that “COUNTRY_x”, “COUNTRY_y”, “STATE_x”, “STATE_y”, “IATA_CODE_x”, and “IATA_CODE_y” are duplicated, and thus removed. Some column names are renamed and all column names are made in lower cases.

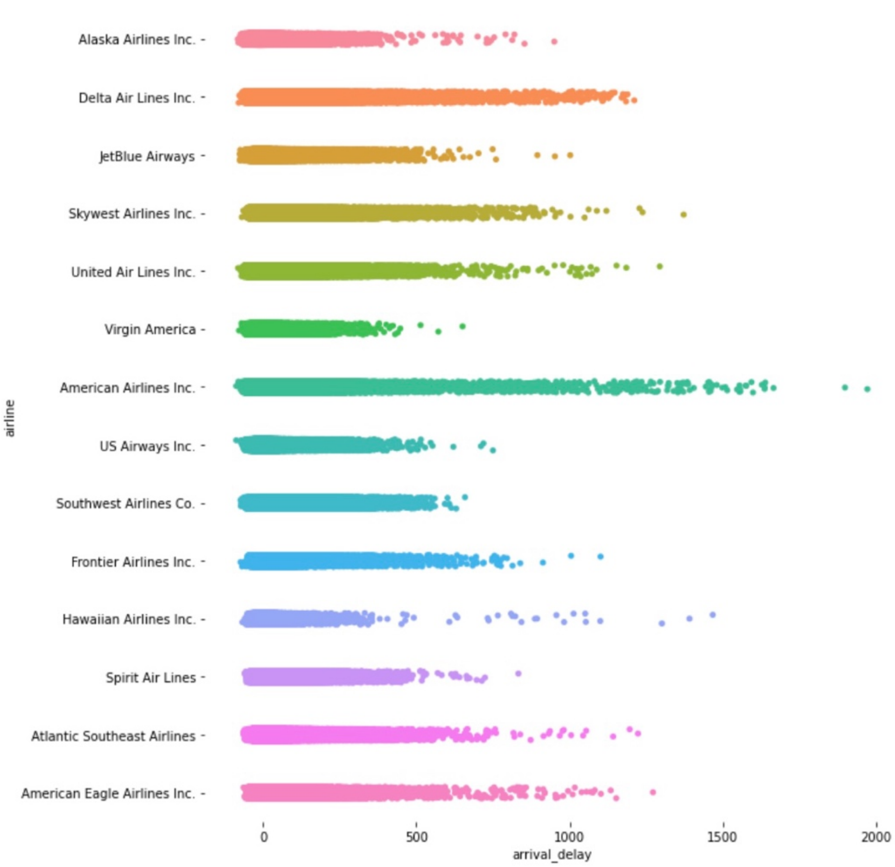
The final cleaned dataset contains 5309350 rows and 28 columns.

Exploratory Data Analysis

Using Seaborn, Matplotlib, and Tableau, some findings are worth noticing.



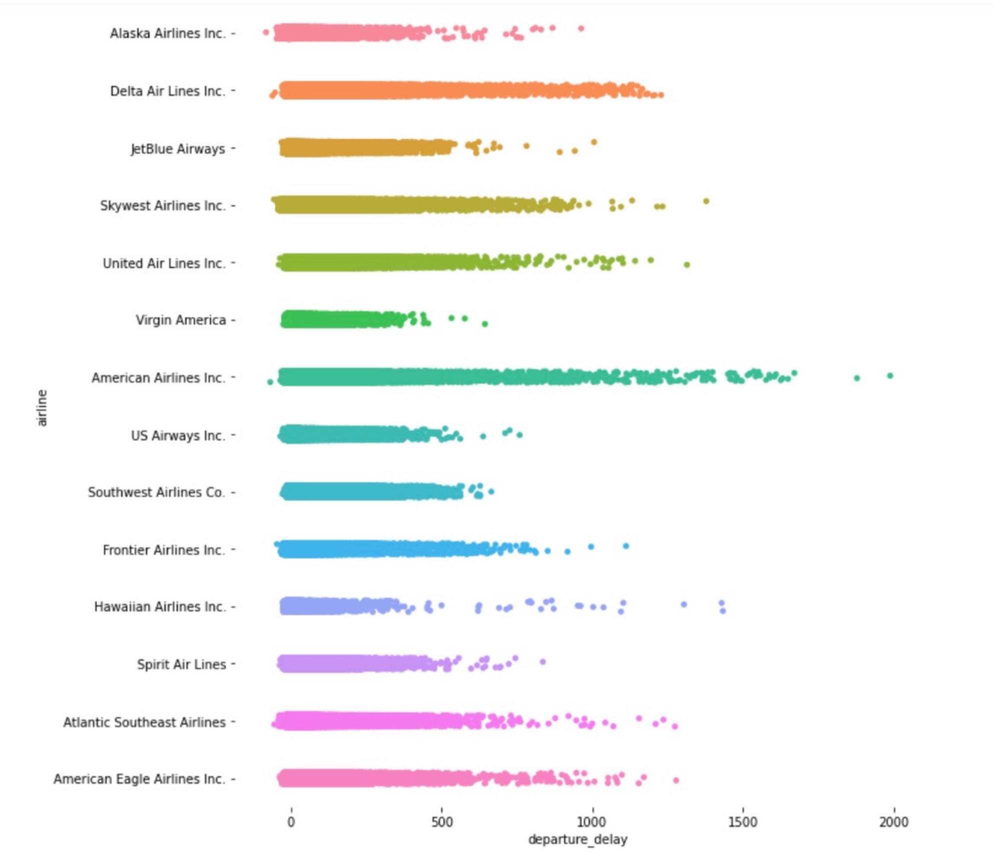
Airline Distributions



We can see that American Airlines ranks the top in terms of arrival delay. In contrast, Southwest Airlines and Virgin America have the lowest arrival delays.

	airline	total_time_arrival_delay	count_arrival_delay	avg_delay
9	Southwest Airlines Co.	5500557.0	1152731	4.771761
3	Atlantic Southeast Airlines	3548236.0	524825	6.760798
8	Skywest Airlines Inc.	3286960.0	539441	6.093271
12	United Air Lines Inc.	2870162.0	463778	6.188655
1	American Airlines Inc.	2504976.0	646583	3.874175
2	American Eagle Airlines Inc.	1881302.0	272602	6.901277
7	JetBlue Airways	1669954.0	245135	6.812385
10	Spirit Air Lines	1591094.0	106608	14.924715
5	Frontier Airlines Inc.	1124311.0	81908	13.726510
11	US Airways Inc.	719831.0	194825	3.694757
4	Delta Air Lines Inc.	534088.0	796532	0.670517
13	Virgin America	277923.0	56439	4.924308
6	Hawaiian Airlines Inc.	150930.0	69889	2.159567
0	Alaska Airlines Inc.	-124271.0	158054	-0.786257

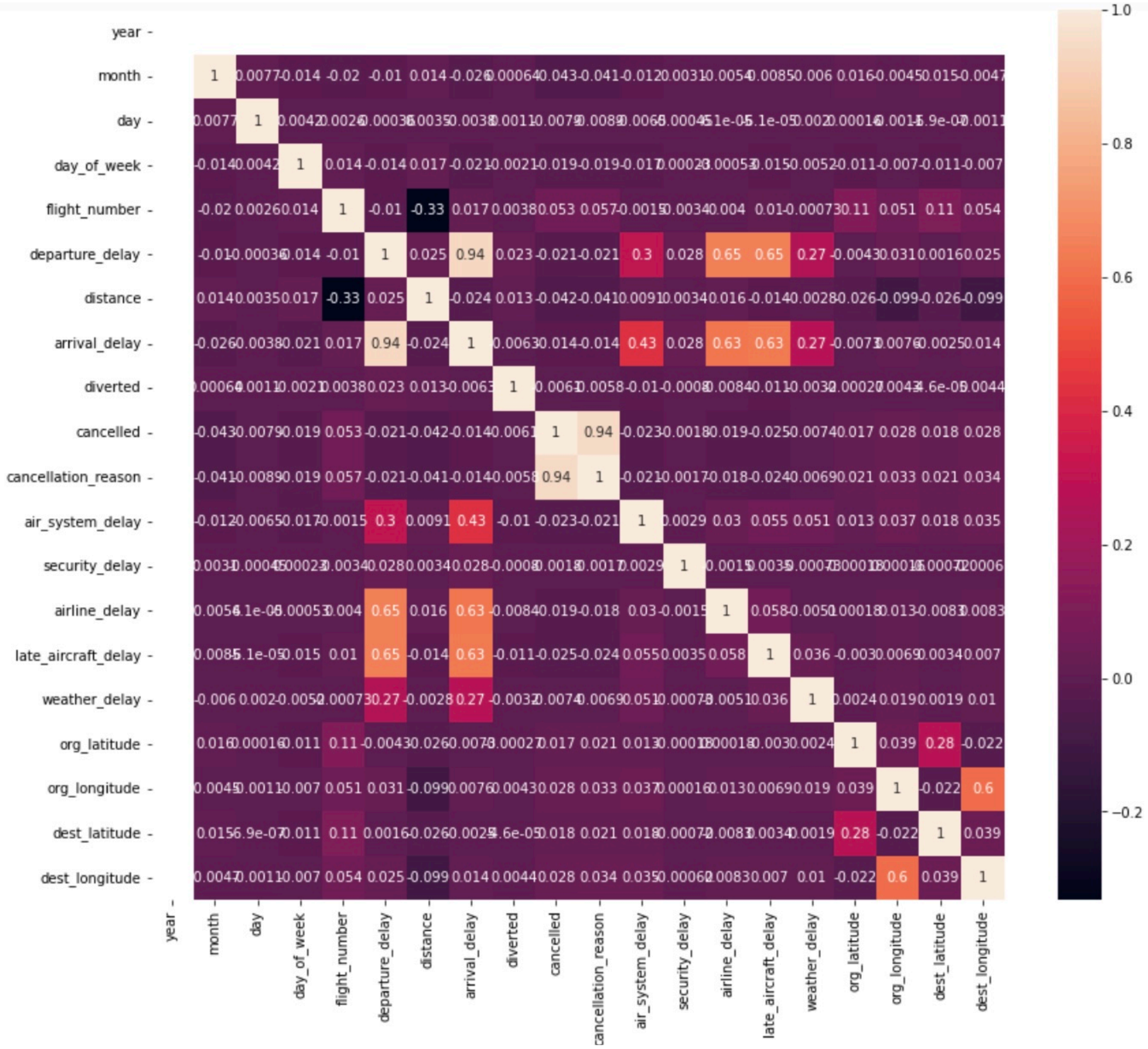
- We can see Southwest Airlines has the highest total delayed arrvial in terms of time and numbers.
- Spirit Air Lines ranks the top on average time per delay.
- American Airlines is the top third airline in terms of number of delays.
- Alaska Airlines has the least total delayed time and average time per delay (remember Alaska has the most records in our dataset yet it has the least arrival time delays).



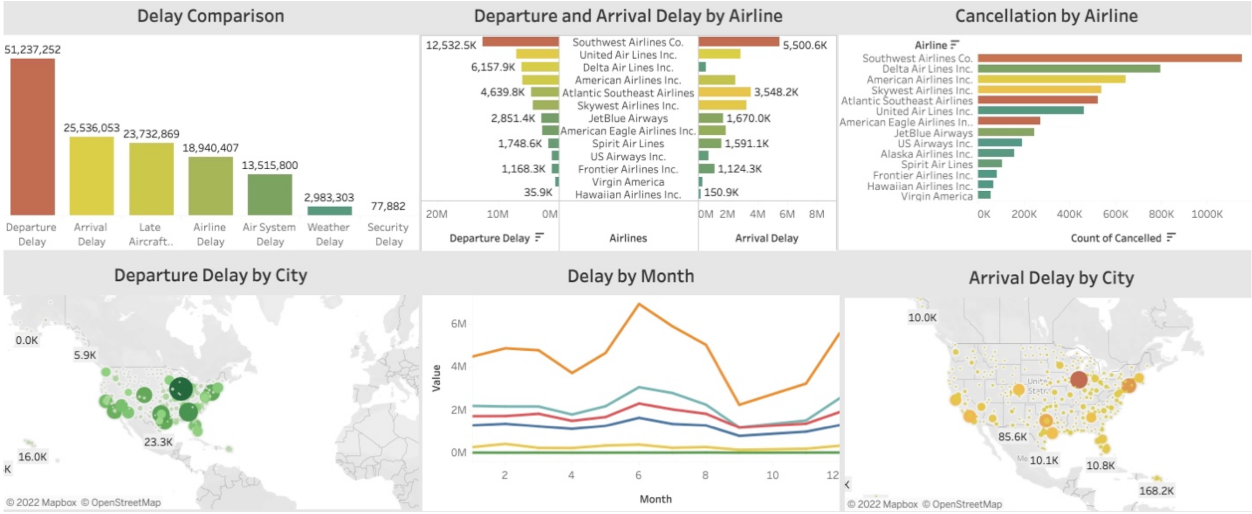
- American Airlines seem to have the most departure delayed time, followed by Delta Air Lines
- Virgin America seems to have the least departure delayed time

	airline	total_time_depart_delay	count_depart_delay	avg_delay
9	Southwest Airlines Co.	12532530.0	1152731	10.872033
12	United Air Lines Inc.	7017425.0	463778	15.131000
4	Delta Air Lines Inc.	6157918.0	796532	7.730911
1	American Airlines Inc.	5962452.0	646583	9.221480
3	Atlantic Southeast Airlines	4639795.0	524825	8.840652
8	Skywest Airlines Inc.	4340148.0	539441	8.045640
7	JetBlue Airways	2851368.0	245135	11.631827
2	American Eagle Airlines Inc.	2762785.0	272602	10.134867
10	Spirit Air Lines	1748583.0	106608	16.401987
11	US Airways Inc.	1196447.0	194825	6.141137
5	Frontier Airlines Inc.	1168293.0	81908	14.263479
13	Virgin America	515312.0	56439	9.130424
0	Alaska Airlines Inc.	308273.0	158054	1.950428
6	Hawaiian Airlines Inc.	35923.0	69889	0.514001

- Southwest Airlines again has the highest total time of departure delays and numbers of departure delays
- American Airlines ranks the 4th on total time of departure delays
- Spirit Air Lines again is the top on average time per delay
- Hawaiian Airlines has the least total delayed time and average time per delay



- Evident positive correlations between arrival_delay and:
  - departure_delay (0.94)
  - airline_delay (0.63)
  - late_aircraft_delay (0.63)
  - air_sytem_delay (0.43)
  - weather_delay (0.27)

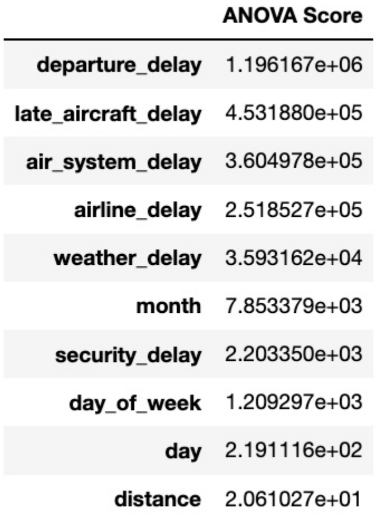


Please see Tableau workbook/dashboard for more details

Pre-Processing

A check of features that related to delays suggests that these features are not normally distributed. Therefore, a MinMaxScaler is applied to these features: “air_system_delay”, “airline_delay”, “weather_delay”, “late_aircraft_delay”, “security_delay”, “arrival_delay”, “departure_delay”, and “distance”. In comparison, “month”, “day”, and “day_of_week” are all normally distributed, and thus they are scaled via StandardScaler. The target feature “delayed” is also converted to binary values (0 and 1).

Since all predicting features are numeric, an ANOVA test is conducted for feature importance. It suggests that “distance” is not important and thus removed from our analysis.



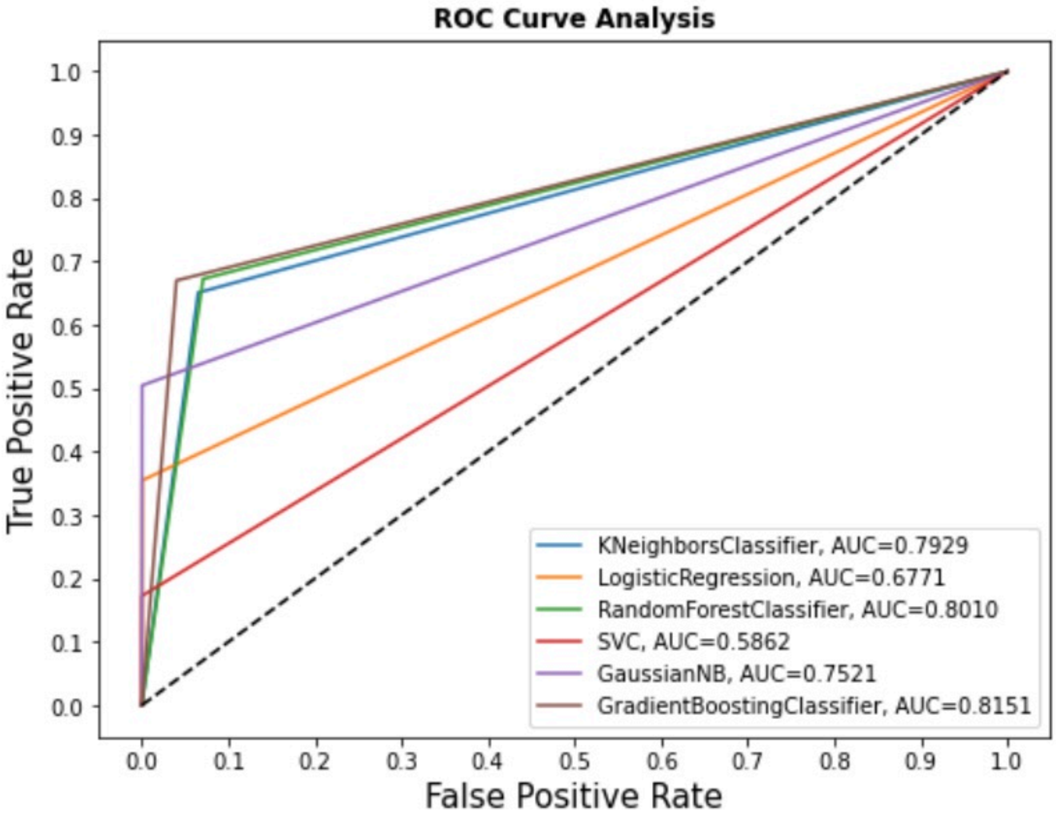
	ANOVA Score
departure_delay	1.196167e+06
late_aircraft_delay	4.531880e+05
air_system_delay	3.604978e+05
airline_delay	2.518527e+05
weather_delay	3.593162e+04
month	7.853379e+03
security_delay	2.203350e+03
day_of_week	1.209297e+03
day	2.191116e+02
distance	2.061027e+01

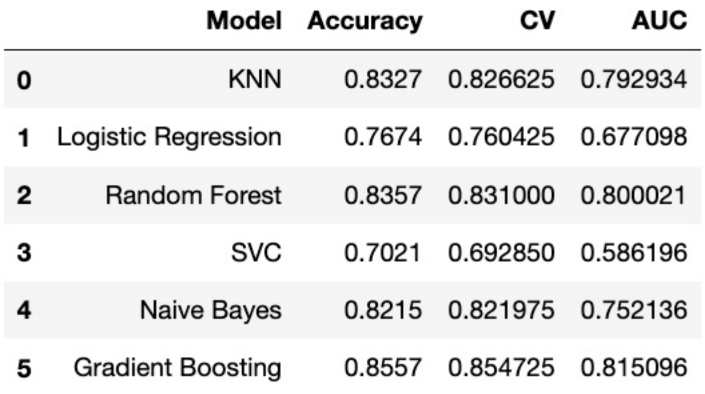
Since the dataset is huge, a sample of 50000 is selected for modelling. Final features for modelling are “month”, “day”, “day_of_week”, “air_system_delay”, “airline_delay”, “weather_delay”, “late_aircraft_delay”, “departure_delay”, and “security_delay”.

Finally, training data and testing data are split by 80% and 20%: X_train: (50000, 9); X_test: (40000, 9); y_train: (50000,); y_test: (40000,).

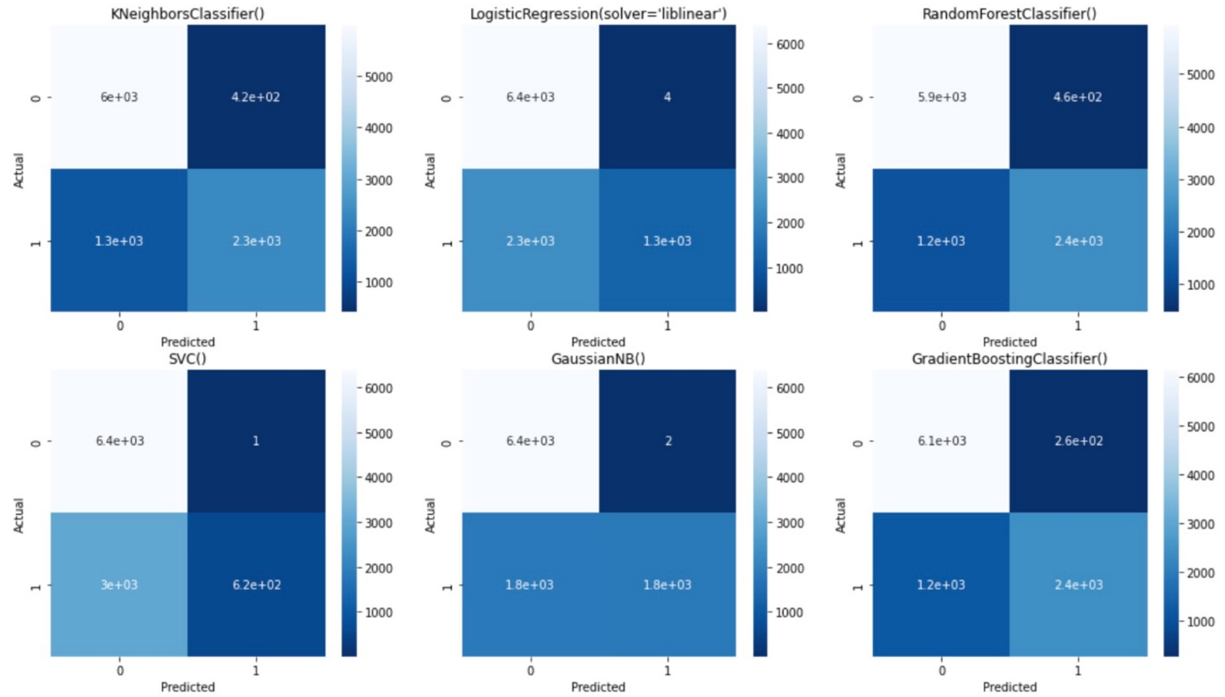
Modelling

There are 9 predictive features and 1 target feature (delayed: 0 and 1). Since this is a classification problem, the following machine learning models are selected: KNN, Logistic Regression, Random Forest, Support Vector Machine, Naïve Bayes, and Gradient Boosting. At first, a pipeline is built with default parameters for all models. The evaluation metrics used in this study are Accuracy Score, CV Score, ROC-AUC Score for model performance.





	Model	Accuracy	CV	AUC
0	KNN	0.8327	0.826625	0.792934
1	Logistic Regression	0.7674	0.760425	0.677098
2	Random Forest	0.8357	0.831000	0.800021
3	SVC	0.7021	0.692850	0.586196
4	Naive Bayes	0.8215	0.821975	0.752136
5	Gradient Boosting	0.8557	0.854725	0.815096

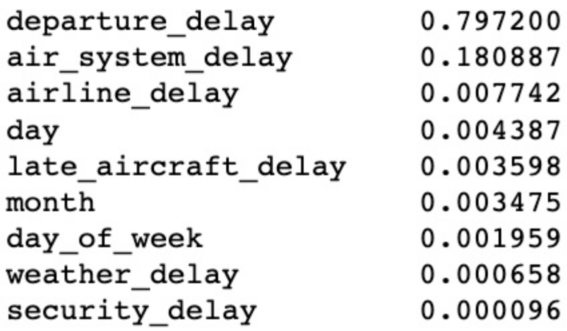


From the above analysis, we can notice that the Gradient Boosting has the highest accuracy score and AUC score, followed by Random Forest. Logistic Regression and SVC seem to be not doing well. KNN and Naïve Bayes are acceptable discrimination. But remember we only train each model with its default parameters; model performance might be different if applying different parameters. So, the next step is hyperparameter tuning.

To optimize model performance, three models are selected for hyperparameter tuning: Random Forest, Gradient Boosting, and KNN. For Random Forest and Gradient Boosting, “max_depth” and “n_estimators” are selected from GridSearchCV; for KNN, “n_neighbors” is selected. Models are trained again using optimized parameters.

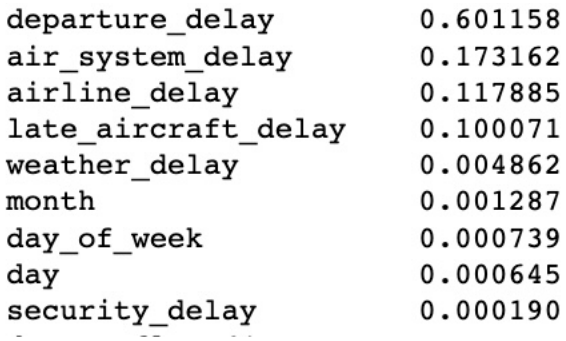
	Accuracy Score (Default)	Accuracy Score (Tuning)
Gradient Boosting	0.8557	0.8560
Random Forest	0.8357	0.8550
KNN	0.8327	0.8344

From the results above, we can see that hyperparameter tuning clearly improves our model performance. The highest accuracy score is Gradient Boosting (0.8557).



```
departure_delay        0.797200
air_system_delay       0.180887
airline_delay          0.007742
day                    0.004387
late_aircraft_delay    0.003598
month                  0.003475
day_of_week            0.001959
weather_delay          0.000658
security_delay         0.000096
```

Feature importance (Gradient Boosting)



```
departure_delay        0.601158
air_system_delay       0.173162
airline_delay          0.117885
late_aircraft_delay    0.100071
weather_delay          0.004862
month                  0.001287
day_of_week            0.000739
day                    0.000645
security_delay         0.000190
```

Feature Importance (Random Forest)

Gradient Boosting and Random Forest both agree on the most predicting feature of delay: “departure_delay”.

We can see that departure delay is the leading factor for U.S. airlines delay in 2015. It is also clear that weather, month, day, day of week are not that important when predicting delays. Departure delay would directly make airlines stay on air for more time, and thus consuming more fuel, leading to an extra cost of the company.

In order to cut the fuel cost and increase revenue, U.S. Department of Transportation (DOT) should focus on making measures to reduce the departure delay. To be specific, DOT should pay more attention to Southwest Airlines, which has the highest total departure delay time, arrival delay time, and the number of departure and arrival delays.

In the meantime, DOT could also consider how to handle delay caused by air system, as it is the second top reason causing arrival delay.