Data Science Capstone

United States Airlines Analysis

Business Scenario

Problem statement:

According to air travel consumer reports, a large proportion of consumer complaints are about frequent flight delays. Of all the complaints received from consumers about airline services, 32% were related to cancellations, delays, or other deviations from the airlines' schedules.
Unavoidable delays can be caused by air traffic, no passengers at the airport, weather conditions, mechanical issues, passengers coming from delayed connecting flights, security clearance, and aircraft preparation.

Objective:

The objective of this project is to identify the factors that contribute to avoidable flight delays. You are also required to build a model to predict if the flight will be delayed.

Dataset Snapshot

Airlines.xlsx

id	Airline	Flight	AirportFrom	AirportTo	DayOfWeek	Time	Length	Delay
1	CO	269	SFO	IAH	3	15	205	1
2	US	1558	PHX	CLT	3	15	222	1
3	AA	2400	LAX	DFW	3	20	165	1
4	AA	2466	SFO	DFW	3	20	195	1
5	AS	108	ANC	SEA	3	30	202	0
6	CO	1094	LAX	IAH	3	30	181	1
7	DL	1768	LAX	MSP	3	30	220	0
8	DL	2722	PHX	DTW	3	30	228	0
9	DL	2606	SFO	MSP	3	35	216	1
10	AA	2538	LAS	ORD	3	40	200	1
11	CO	223	ANC	SEA	3	49	201	1
12	DL	1646	PHX	ATL	3	50	212	1
13	DL	2055	SLC	ATL	3	50	210	0
14	AA	2408	LAX	DFW	3	55	170	0
15	AS	132	ANC	PDX	3	55	215	0
16	US	498	DEN	CLT	3	55	179	0
17	B6	98	DEN	JFK	3	59	213	0
18	CO	1496	LAS	IAH	3	60	162	0
19	DL	1450	LAS	MSP	3	60	181	0
20	CO	507	ONT	IAH	3	75	167	0

Dataset Description

Airlines.xlsx

Variables	Description
id	Flight number
Airline	Type of commercial airlines
Flight	Type of aircraft
AirportFrom	Source airport
AirportTo	Destination airport
DayOfWeek	Day of the week
Time	Departure time measured in minutes from midnight (range is from 10 to 1439)
Length	Duration of the flight in minutes
Delay	If the flight is delayed

Dataset Snapshot

airports.xlsx

id	ident	type	name	latitude_d	longitude_	elevation_	continent	iso_countr	iso_region	municipali	scheduled_	gps_code
6523	00A	heliport	Total Rf Helig	40.0708008	-74.933601	11	NA	US	US-PA	Bensalem	no	00A
323361	00AA	small_airport	Aero B Ranch	38.704022	-101.47391	3435	NA	US	US-KS	Leoti	no	00AA
6524	00AK	small_airport	Lowell Field	59.947733	-151.69252	450	NA	US	US-AK	Anchor Point	no	00AK
6525	00AL	small_airport	Epps Airpark	34.8647995	-86.770302	820	NA	US	US-AL	Harvest	no	00AL
6526	00AR	closed	Newport Hos	35.6087	-91.254898	237	NA	US	US-AR	Newport	no	
322127	00AS	small_airport	Fulton Airport	34.9428028	-97.818019	1100	NA	US	US-OK	Alex	no	00AS
6527	00AZ	small_airport	Cordes Airpo	34.3055992	-112.165	3810	NA	US	US-AZ	Cordes	no	00AZ
6528	00CA	small_airport	Goldstone (G	35.35474	-116.88533	3038	NA	US	US-CA	Barstow	no	00CA
324424	00CL	small_airport	Williams Ag A	39.427188	-121.76343	87	NA	US	US-CA	Biggs	no	00CL
322658	00CN	heliport	Kitchen Creel	32.7273736	-116.45974	3350	NA	US	US-CA	Pine Valley	no	00CN
6529	00CO	closed	Cass Field	40.622202	-104.344	4830	NA	US	US-CO	Briggsdale	no	
6531	00FA	small_airport	Grass Patch A	28.6455002	-82.219002	53	NA	US	US-FL	Bushnell	no	00FA
6532	00FD	closed	Ringhaver He	28.8466	-82.345398	25	NA	US	US-FL	Riverview	no	
6533	00FL	small_airport	River Oak Air	27.2308998	-80.9692	35	NA	US	US-FL	Okeechobee	no	00FL
6534	00GA	small_airport	Lt World Airp	33.7675018	-84.068298	700	NA	US	US-GA	Lithonia	no	00GA
6535	00GE	heliport	Caffrey Helip	33.887982	-84.736983	957	NA	US	US-GA	Hiram	no	00GE
6536	00HI	heliport	Kaupulehu He	19.832881	-155.97835	43	OC	US	US-HI	Kailua-Kona	no	00HI
6537	00ID	small_airport	Delta Shores	48.1453018	-116.214	2064	NA	US	US-ID	Clark Fork	no	00ID
322581	00IG	small_airport	Goltl Airport	39.724028	-101.39599	3359	NA	US	US-KS	McDonald	no	00IG
6538	00II	closed	Bailey Gener	41.644501	-87.122803	600	NA	US	US-IN	Chesterton	no	

Dataset Description

airports.xlsx

Variables	Description
id	This is an identifier for the airport. It will stay persistent even if the airport code changes.
ident	This is the text identifier used in the *OurAirports* URL. This will be the International Civil Aviation Organization (ICAO) code, if available. Otherwise, it will be a local airport code (if there is no conflict) or an internally generated code starting with the ISO2 country code followed by a dash and a four-digit number.
type	This shows the type of the airport. The values allowed here are *closed_airport*, *heliport*, *large_airport*, *medium_airport*, *seaplane_base*, and *small_airport*.
name	This shows the official name of the airport, including *Airport* and *Airstrip*
latitude_deg	This shows the latitude of the airport in decimal degrees (north is positive).
longitude_deg	This shows the longitude of the airport in decimal degrees (east is positive).

Dataset Description

airports.xlsx

Variables	Description
elevation_ft	This shows the elevation MSL of the airport in feet (not meters).
continent	This shows the code for the continent where the airport is (primarily) located. The allowed values include *AF* (Africa), *AN* (Antarctica), *AS* (Asia), *EU* (Europe), *NA* (North America), *OC* (Oceania), and *SA* (South America).
iso_country	This shows the two-character ISO 3166:1-alpha2 code for the country where the airport is (primarily) located. A handful of unofficial, non-ISO codes are also in use, such as *XK* for Kosovo.
iso_region	This is an alphanumeric code for the high-level administrative subdivision of a country where the airport is primarily located (e.g., province and governorate), prefixed by the ISO2 country code and a hyphen. *OurAirports* uses ISO 3166:2 codes whenever possible, preferring higher administrative levels, but also includes some custom codes.
municipality	This shows the primary municipality that the airport serves (when available). Note that this is not necessarily the municipality where the airport is physically located.

Dataset Description

airports.xlsx

Variables	Description
scheduled_service	This shows *yes* if the airport currently has scheduled airline service and *no* if otherwise.
gps_code	This shows the code that an aviation GPS database (such as Jeppesen's or Garmin's) would normally use for the airport. This will always be the ICAO code if one exists. Note that, unlike the *ident* column, this is not guaranteed to be globally unique.
iata_code	This shows the three-letter IATA code for the airport (if it has one).
local_code	This shows the local country code for the airport if it’s different from the *gps_code* and *iata_code* fields (used mainly for US airports).
home_link	This shows the URL of the airport's official home page on the web if one exists.
wikipedia_link	This shows the URL of the airport's page on Wikipedia if one exists.
Keywords	This field contains other keywords or phrases to assist with the search. These are separated by a comma. It may also include former names for the airport, alternate codes, names in other languages, and nearby tourist destinations.

Dataset Snapshot

runways.xlsx

id	airport_ref	airport_ident	length_ft	width_ft	surface	lighted	closed	le_ident	le_latitude_d	le_longitude_	le_elevation_	le_heading_c
269408	6523	00A	80	80	ASPH-G	1	0	H1				
255155	6524	00AK	2500	70	GRVL	0	0	N				
254165	6525	00AL	2300	200	TURF	0	0	1				
270932	6526	00AR	40	40	GRASS	0	0	H1				
322128	322127	00AS	1450	60	Turf	0	0	1				
257681	6527	00AZ	1700	60	GRAVEL	0	0	15				
245528	6528	00CA	6000	80	ASPH	0	0	4	35.3493004	-116.893		50
250597	6529	00CO	3900	20	TURF-G	0	0	16				
247972	6531	00FA	3200	100	TURF	0	0	8				
265037	6532	00FD	74	74	TURF	0	0	H1				
250414	6533	00FL	4090	100	TURF	0	0	12				
253429	6534	00GA	2600	80	TURF	0	0	9				
265038	6535	00GE	125	95	ASPH	1	0	H1				
265039	6536	00HI	1155	45	ASPH-G	0	0	H1				
246648	6537	00ID	3300	40	TURF	0	0	8				
246649	6537	00ID	2700	40	TURF	0	0	11				
252182	6539	00IL	2500	75	TURF-F	0	0	18				
265040	6540	00IN	40	40	MATS	1	0	H1				
254597	6541	00IS	1600	70	TURF	0	0	9				
256603	6542	00KS	2600	85	TURF	0	0	17				

Dataset Description

runways.xlsx

Variables	Description
id	This shows the internal *OurAirports* integer identifier for the runway. This will stay persistent even if the runway numbering changes.
airport_ref	This shows the internal integer foreign key matching the *id* column for the associated airport in **airports.csv**. Here, *airport_ident* is a better alternative.
airport_ident	This shows the externally visible string foreign key matching the *ident* column for the associated airport in **airports.csv**.
length_ft	This shows the length of the full runway surface (including displaced thresholds and overrun areas) in feet.
width_ft	This shows the width of the runway surface in feet.
surface	This shows the code for the runway surface type. This is not a controlled vocabulary yet, but it will be soon (probably). Some common values include *ASP* (asphalt), *TURF* (turf), *CON* (concrete), *GRS* (grass), *GRE* (gravel), *WATER* (water), and *UNK* (unknown).

Dataset Description

runways.xlsx

Variables	Description
lighted	This shows *1* if the surface is lit at night and *0* if not. Note that this is inconsistent with **airports.csv**, which uses *yes* and *no* instead.)
closed	This shows *1* if the runway surface is currently closed and *0* if not.
le_ident	This shows the identifier for the low-numbered end of the runway.
le_latitude_deg	This shows the latitude of the center of the low-numbered end of the runway in decimal degrees (north is positive) if available.
le_longitude_deg	This shows the longitude of the center of the low-numbered end of the runway in decimal degrees (east is positive) if available.
le_elevation_ft	This shows the elevation above MSL of the low-numbered end of the runway in feet.
le_heading_degT	This shows the heading of the low-numbered end of the runway in degrees true (non-magnetic).

Dataset Description

runways.xlsx

Variables	Description
le_displaced_threshold_ft	This shows the length of the displaced threshold (if any) for the low-numbered end of the runway in feet.
he_ident	This shows the identifier for the high-numbered end of the runway.
he_latitude_deg	This shows the latitude of the center of the high-numbered end of the runway in decimal degrees (north is positive) if available.
he_longitude_deg	This shows the longitude of the center of the high-numbered end of the runway in decimal degrees (east is positive) if available.
he_elevation_ft	This shows the elevation above MSL of the high-numbered end of the runway in feet.
he_heading_degT	This shows the heading of the high-numbered end of the runway in degrees true (non-magnetic).
he_displaced_threshold_ft	This shows the length of the displaced threshold (if any) for the high-numbered end of the runway in feet.

Project Task: Week 1

Data science

1. Import and aggregate data:
    a. Collect information related to flights, airports (e.g., type of airport and elevation), and runways (e.g., *length_ft*, *width_ft*, *surface*, and number of runways). Gather all fields you believe might cause avoidable delays in one dataset.

Hint: In this case, you would have to determine the keys to join the tables. A data description will be useful.

    b. When it comes to on-time arrivals, different airlines perform differently based on the amount of experience they have. The major airlines in this field include US Airways Express (founded in 1967), Continental Airlines (founded in 1934), and Express Jet (founded in 1986). Pull such information specific to various airlines from the given Wikipedia page link.

    https://en.wikipedia.org/wiki/List_of_airlines_of_the_United_States.

Hint: Here, you should use web scraping to learn how long an airline has been operating for.

Project Task: Week 1

Data science

c. You should then get all the information gathered so far in one place.

d. The total passenger traffic may also contribute to flight delays. The term *hub* refers to busy commercial airports. *Large hubs* are airports that account for at least 1 percent of the total passenger enplanements in the United States. Airports that account for 0.25 percent to 1 percent of total passenger enplanements are considered medium hubs. Pull passenger traffic data from the Wikipedia page given below using web scraping and collate it in a table.

https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States

2. You should then examine the missing values in each field, perform missing value treatment, and justify your actions.

Project Task: Week 1

Data science

3. Perform data visualization and share your insights on the following points:
    a. According to the data provided, approximately 70% of Southwest Airlines flights are delayed. Compare it with other airlines' data by visualizing it.
    b. Flights were delayed on various weekdays. Which day of the week is the safest for travel?
    c. Which airlines should be recommended for short-, medium-, and long-distance travel?
    d. Do you notice any patterns in the departure times of long-duration flights?
4. How many flights were delayed at large hubs compared to medium hubs? Use appropriate visualization to represent your findings.

Project Task: Week 1

Data science

5. Use hypothesis testing strategies to discover:
    a. If the airport's altitude has anything to do with flight delays for incoming and departing flights
    b. If the number of runways at an airport affects flight delays
    c. If the duration of a flight (length) affects flight delays

Hint: Test this from the perspective of both the source and destination airports

6. Find the correlation matrix between the flight delay predictors, create a heatmap to visualize this, and share your findings

Project Task: Week 1

Machine learning

1. Use OneHotEncoder and OrdinalEncoder to deal with categorical variables
2. Perform the following model-building steps:
   a. Split data into train and test
   b. Standardize data
   c. Apply logistic regression (use stochastic gradient descent optimizer) and decision tree models

Note: Make sure you use standardization effectively, ensuring no data leakage and leverage pipelines to have a cleaner code

d. Check the accuracy report of the model on the train and test data

e. Take care of overfitting of decision tree model

Note: The final prediction will rely on the majority class voted by five models generated through the stratified five-fold method.

   g. Compare the results of logistic regression and decision tree classifier

3. Build and validate the models using the Gradient Boosting classifier, compare all methods, and share your findings

Project Task: Week 2

SQL

1. Determine the number of delayed flights on various days of the week
2. Determine the number of delayed flights for various airlines
3. Determine how many delayed flights land at airports with at least 10 runways
4. Compare the number of delayed flights at airports higher than average elevation and those that are lower than average elevation for both source and destination airports

Project Task: Week 2

Tableau

1. Create a dashboard in Tableau by selecting appropriate chart types and metrics for the business

Note: Emphasize on data storytelling

Thank you