ST2195 PART 2 COURSEWORK

Jupyter Python Airport Data sets

Abstract

This Jupyter Python report covers data analysis of departure delays of the various airlines in America and to detect the probability of any such delays in the near future.

Clement Tan

Ctnk2

Contents

Introduction 2
Airport Data sets and the types of departure delays 2
Procedure to do the Questions for Part 2 2
Part A best time of day to fly. 3
Part A Best day of week to fly 3
Part B Do older planes suffer on a year-to-year basis? 4
Part C For each year, create a logistic regression model for the probability of diverted US flights 6
Bibliography 8

Introduction

Since the dawn of the 1900s, air travel has been advancing in various technological aspects. Examples are the improvements made in turbo-jet engines to improve passenger capacity in aircrafts as aircrafts get larger in size. Furthermore, improvements in telecommunications and data analysis have enabled tracking of aircraft movements and flight operations to be smoother. Henceforth, the writing of this airport data analysis report is to examine the best times of the 24-hour day and days of the week to fly the aircrafts and minimize the delays in departure.

In addition, the data analysis will be to determine if older planes suffer in yearly operations. Lastly, a logistic regression model for the probability of diverted US flights is determine the airlines that have problems with the aircrafts. Then aircraft manufacturing and maintenance teams can better inspect and repair the respect aircrafts.

The airport data sets will be from the year 1997 until 2000. These years were important as there have been several notable departure delays as the 20th century drew to the end due to errors in the reading gauges. That data analysts can better advice the pilots and air traffic control officers to minimise departure delays in the later years. The coding process for all of the questions are done using Jupyter Python codes.

Airport Data sets and the types of departure delays

Establish the 4 csv files from the years 1997 to 2000 by pairing the years 1997 to 1998 and 1999 to 2000 in to 2 pairs. After that, count the different delays and group them in to 3 different delay types. For observation counts with 15 minutes and below, they were grouped in minimum delays 'min-delay.' For observation counts with delays between 15 and 180 minutes, they were grouped with normal delays 'normal_delay,' and otherwise big delays for those with 180 minutes and above.

From 1997 to 1998, there were a total of 105549292 CRS departure delay counts and from 1999 to 2000, there were a total of 10869130 CRS departure delay observations. Hence, there were over 10 million flights with different departure delays for each of these 2 years.

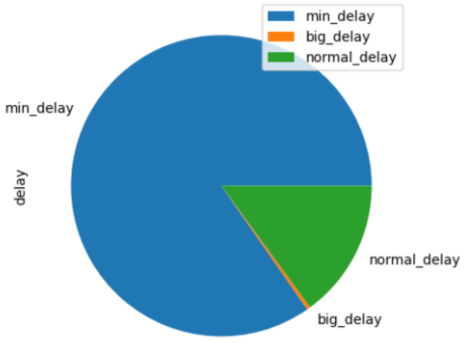


Fig.1.0.1: Pie chart of the delay from 1997 till 1998

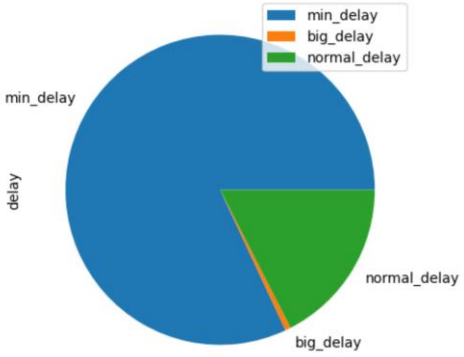


Fig.1.0.2: Pie chart of the delay from 1999 till 2000

From 1997 till 1998, there were 8936976 minimum delay count variables, 40751 big delay count variables and 1576565 normal delay count variables as shown in Fig.1.0.1. In Fig.1.02, there were 8908009 minimum delay count variables, 62789 big delay count variables and 1898332 normal delay count variables from 1999 till 2000.

Procedure to do the Questions for Part 2

In Part A, to find the best time of day to fly with the least delays, a stack bar plots per hour of 3 different delays was plotted over the 24-hour cycle for the years 1997-1998 and 1999-2000. From there, determine which hour has the most proportion of minimum delay flights and normal delay flights. In Part A to find the best day of week to fly with the least delays, plot a box-plot diagram of delay variance count for each day. A stacked bar plot is to show which of the days has the most minimum delay counts.

In part B, a histogram and the stacked bar for each of the 4 years from 1997 to 2000 is plotted to find the flight age categories that has the most delay counts. The flight chart head count of the 5 observations for each year is created. The head count is to show which airlines have the most departure delays counts that were in the top 5 and of that particular age of flight.
Lastly, for Part C, a logistic regression model is created to predict the probability of US diverted flights using the given flight data parameters as x-axis. The number of observations count of diverted flights is used as y-axis and using the x and y-axes to create a logistic regression model graph.

Part A best time of day to fly.

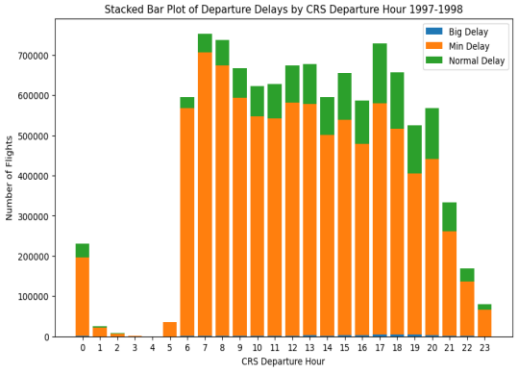


Fig.1.1.1: Stacked Bar Plot for 1997-1998

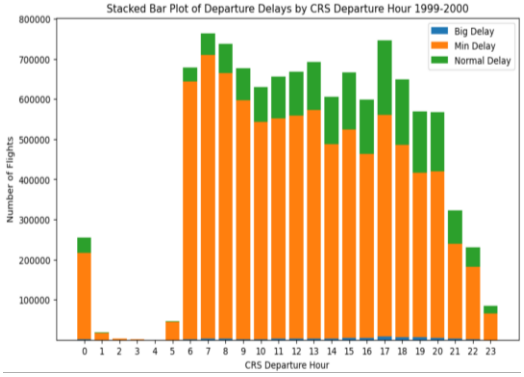


Fig.1.1.2: Stacked Bar Plot for 1999-2000

In both Fig.1.1.1 and Fig.1.1.2, the best time of the day to fly with minimum delays is around 6.00am. This is indicated that both bar plots at 6.00am have the greatest proportion of minimum delays of less than or equal to 15 minutes and normal delays of between 15 and 180 minutes.

Part A Best day of week to fly

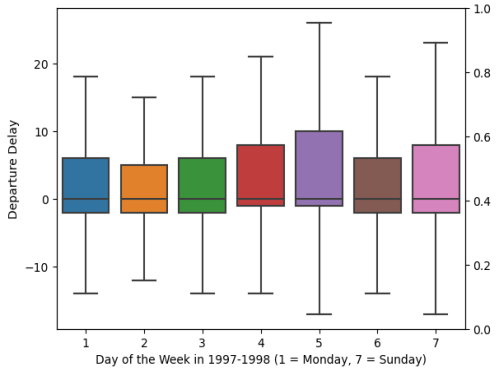


Fig.1.2.1: Box Plot Diagram for 1997-1998

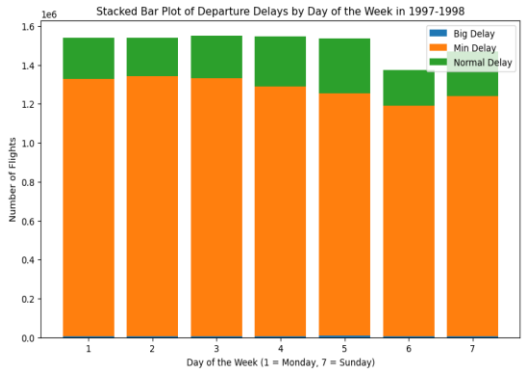
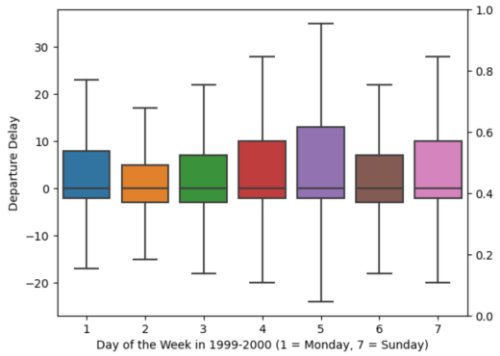


Fig.1.2.2: Stacked Bar Plot from 1997-1998



FigU.1.2.3: Box Plot Diagram for 1999-2000

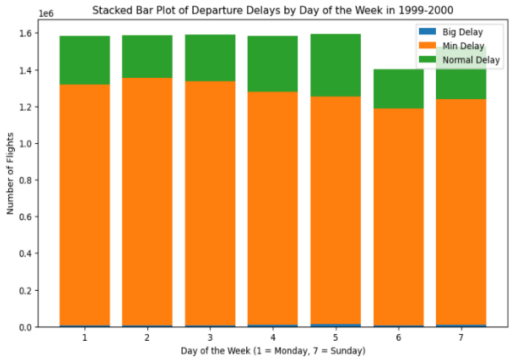


Fig.1.2.4: Stacked Bar Plot from 1999-2000

In both Box Plot Diagrams of Fig.1.2.1 and Fig.1.2.3, the range of departure delay that has the least variation in Tuesday (2nd day of the week). In Fig.1.2.2 and Fig.1.2.4, from Monday to Friday, there were approximately equal number of departure flights made for each working day. However, Tuesday has the greatest proportion of minimum delays of 15 minutes and below. Tuesday is the best day to fly to reduce any departure delays.

Part B Do older planes suffer on a year-to-year basis?

A histogram for each year from 1997 to 2000 has been created. The histogram for each year shows the number of flight delays for the respective age of flight category. The stacked bar plots of age of flight category with the number of flight departure delays in millions.

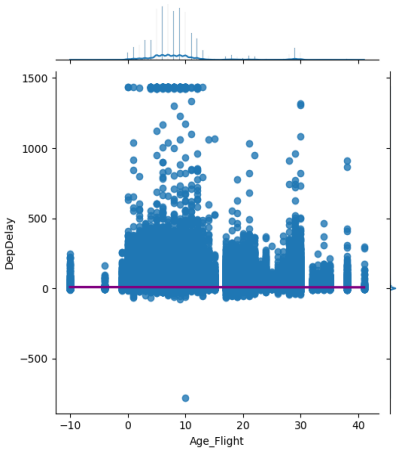


Fig.2.1.1: Histogram for year 1997

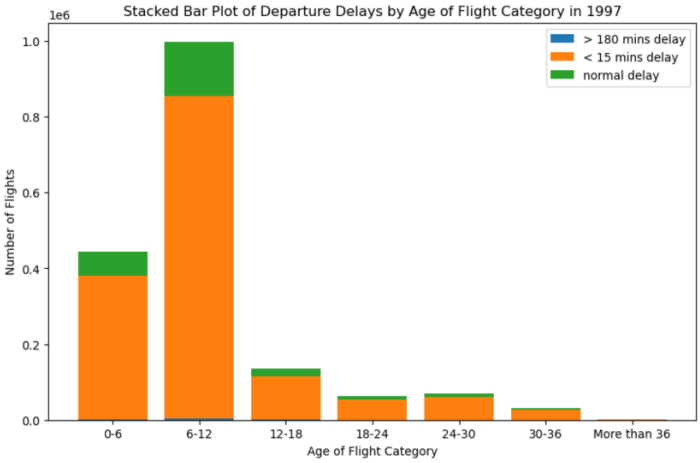


Fig.2.1.2: Stacked Bar Plot for flight delays in 1997

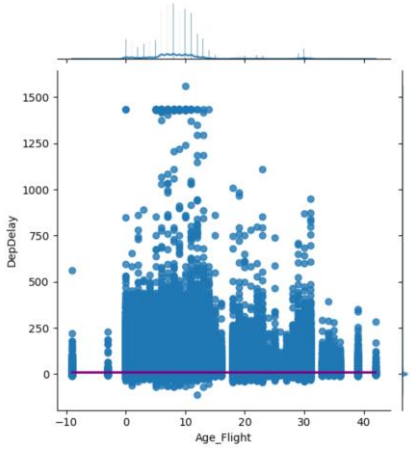


Fig.2.1.3: Histogram for year 1998

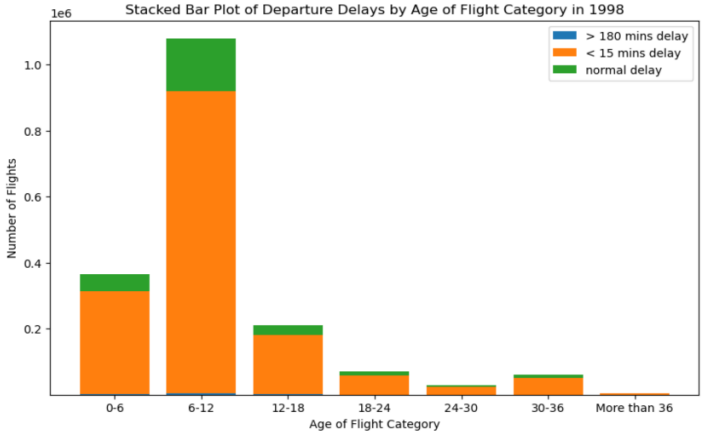
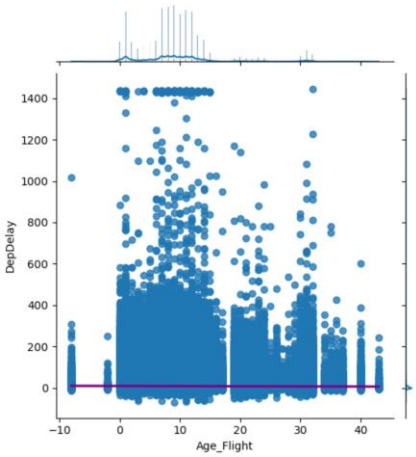


Figure.2.1.4: Stacked Bar Plot for flight delays in 1998

In Fig.2.1.1 and Fig.2.1.2, the histogram and the stacked bar plot respectively have shown that the majority of the flight delays in 1997 were in the 0-6 years and 6-12 years categories. This was seen in the histogram Fig.2.1.1 that the observations clustered from 0 to 12 years with flight delays that were closer to 1500 minutes. There was an outlier of approximately -750 minutes at the age of flight of 10 years seen in Fig.2.1.1. In Fig.2.1.2, the category 6-12 years have the greatest number of flight delays followed by category 0-6 years. This indicated that after 6 years of age in flight, aircrafts tend to have more CRS flight delays thus leading to the 1 million flights recorded with delays in 6-12 years category.

In Fig.2.1.3 and Fig.2.1.4, the histogram and the stacked bar plot respectively have shown that the majority of the flight delays in 1998 were in the 0-6 years and 6-12 years categories. Like in the year 1997, the flight observations with delays clustered in the 0 to 12 years range. In Fig.2.1.4, the category 6-12 years have the greatest number of flight delay observations in 1998 as in with the Fig.2.1.2 of the year 1997.



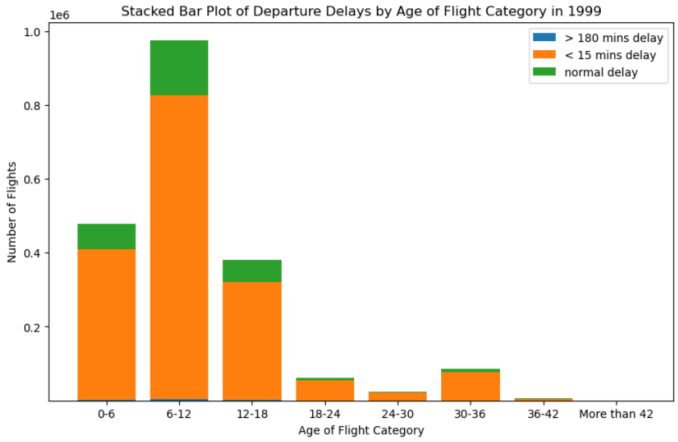
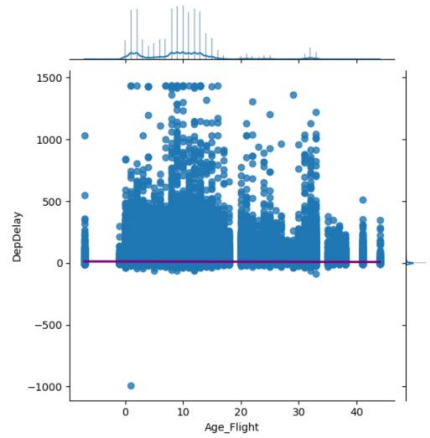


Fig.2.1.5: Histogram for year 1999 *Fig.2.1.6: Stacked Bar Plot for flight delays in 1999*

In Fig.2.1.5 histogram for the year 1999, the flight delay observations recorded clustered from 0 to 12 years of age of flight. This was the same as in Fig.2.1.1 and Fig.2.1.3. In Fig.2.1.8, the not only the categories 0-6 years and 6-12 years have recorded more flight delays, but the 12-18 years category has recorded a somewhat higher delay count. In 1999, aircrafts with 12-18 years age of flight have been recorded with more flight delays than in 1997 and 1998. This was possibly due to bad weather and an antiquated air traffic control system, which was administered by the FAA. (Neil Irwin, 1999)



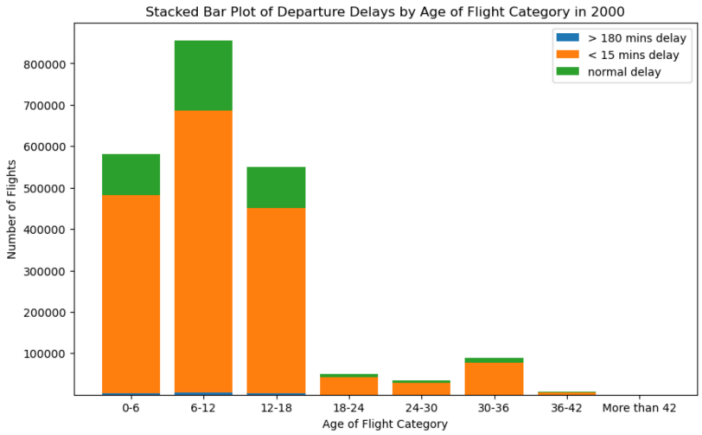


Fig.2.1.7: Histogram for year 2000 *Fig.2.1.8: Stacked Bar Plot for flight delays in 2000*

In the year 2000, there were significantly more flight delays for aircrafts with age of flight to be from 0 to 18 years indicated by histogram Fig.2.1.7 and stacked bar plot diagram Fig.2.1.8. In Fig.2.1.7, most of the flight delay data observations clustered in the 0 to 18 years range. In Fig.2.1.8, aircrafts in the 6-12 years age of flight category were recorded to have the most flight delays. This was followed close behind by aircrafts in the 0-6 years category and the 12-18 years category. That there were more than 500000 aircraft flight delays each were recorded in the 0-6 years and 12-18 years categories. The reason behind there was in 2000, runway and taxiway construction contributed to delays at airports like Boston, Dallas-Ft. Worth, Houston Intercontinental, LaGuardia and Phoenix as stated by FAA. (FAA: Record flight delays in 2000, 2001)

Using the aircraft flight data for the years 1997 to 2000 in Fig.2.1.9 to 2.1.12, the flight delays were aircrafts of Delta Airlines (DL), John Glenn Columbus International Airport (CMH), Hawaiian Airlines (HP) and United Airlines (US). The airports of origins of the aircrafts with such delays were Atlanta Airport (ATL), Boston Airport (BOS) and Southwest Florida International Airport (RSW). The destination airports were Phoenix Airport (PHX) and Philadelphia International Airport (PHL). The result of such departure delays was due to constructions of runways and taxiways for Boston Airport in 2000 that caused such delays. (FAA: Record flight delays in 2000, 2001)

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime
0	1997	1	1	3	2212.0	1930	2334.0
1	1997	1	23	4	1738.0	1735	1944.0
2	1997	1	13	1	1156.0	1155	1330.0
3	1997	1	26	7	2151.0	2140	2300.0
4	1997	1	14	2	1933.0	1900	2132.0

	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime
0	2136	DL	345	N128DL	202.0
1	1945	DL	689	N128DL	246.0
2	1325	DL	999	N128DL	94.0
3	2228	DL	437	N128DL	129.0
4	2122	DL	197	N128DL	299.0

	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn
0	246.0	190.0	118.0	162.0	ATL	PHX	1587	4
1	250.0	224.0	-1.0	3.0	ATL	PHX	1587	4
2	90.0	75.0	5.0	1.0	ATL	RSW	515	4
3	108.0	94.0	32.0	11.0	SLC	SAN	626	12
4	322.0	275.0	10.0	33.0	ATL	SEA	2182	4

	TaxiOut	Cancelled	CancellationCode	Diverted	CarrierDelay	WeatherDelay
0	8	0	NaN	0	NaN	NaN
1	18	0	NaN	0	NaN	NaN
2	15	0	NaN	0	NaN	NaN
3	23	0	NaN	0	NaN	NaN
4	20	0	NaN	0	NaN	NaN

	NASDelay	SecurityDelay	LateAircraftDelay	type	manufacturer
0	NaN	NaN	NaN	Corporation	BOEING
1	NaN	NaN	NaN	Corporation	BOEING
2	NaN	NaN	NaN	Corporation	BOEING
3	NaN	NaN	NaN	Corporation	BOEING
4	NaN	NaN	NaN	Corporation	BOEING

Fig.2.1.9: Flight data list of aircrafts in 1997

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime
0	1998	1	1	4	935.0	935	1304.0
1	1998	1	23	5	931.0	933	1312.0
2	1998	1	20	2	1111.0	1115	1330.0
3	1998	1	10	6	1715.0	1638	1922.0
4	1998	1	14	3	1638.0	1638	1844.0

	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime
0	1319	HP	2082	N901AW	329.0
1	1317	HP	2082	N901AW	341.0
2	1330	HP	93	N901AW	259.0
3	1846	HP	2031	N901AW	247.0
4	1846	HP	2031	N901AW	246.0

	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn
0	344.0	311.0	-15.0	0.0	BOS	PHX	2300	2
1	344.0	310.0	-5.0	-2.0	BOS	PHX	2300	18
2	255.0	247.0	0.0	-4.0	CMH	PHX	1671	6
3	248.0	231.0	36.0	37.0	CMH	PHX	1671	7
4	248.0	233.0	-2.0	0.0	CMH	PHX	1671	3

	TaxiOut	Cancelled	CancellationCode	Diverted	CarrierDelay	WeatherDelay
0	16	0	NaN	0	NaN	NaN
1	13	0	NaN	0	NaN	NaN
2	6	0	NaN	0	NaN	NaN
3	9	0	NaN	0	NaN	NaN
4	10	0	NaN	0	NaN	NaN

	NASDelay	SecurityDelay	LateAircraftDelay	type	manufacturer
0	NaN	NaN	NaN	Corporation	BOEING
1	NaN	NaN	NaN	Corporation	BOEING
2	NaN	NaN	NaN	Corporation	BOEING
3	NaN	NaN	NaN	Corporation	BOEING
4	NaN	NaN	NaN	Corporation	BOEING

Fig.2.1.10: Flight data list of aircrafts in 1998

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime
0	1999	1	2	6	604.0	605	719.0
1	1999	1	16	6	635.0	635	915.0
2	1999	1	19	2	634.0	635	905.0
3	1999	1	26	2	626.0	635	912.0
4	1999	1	11	1	1733.0	1735	2024.0

	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime
0	715	US	549	N427US	75.0
1	911	US	1028	N427US	160.0
2	911	US	1028	N427US	151.0
3	911	US	1028	N427US	166.0
4	2011	US	1092	N427US	171.0

	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn
0	70.0	51.0	4.0	-1.0	ROC	PHL	257	8
1	156.0	139.0	4.0	0.0	RSW	PHL	992	11
2	156.0	131.0	-6.0	-1.0	RSW	PHL	992	10
3	156.0	145.0	1.0	-9.0	RSW	PHL	992	10
4	156.0	152.0	13.0	-2.0	RSW	PHL	992	10

	TaxiOut	Cancelled	CancellationCode	Diverted	CarrierDelay	WeatherDelay
0	16	0	NaN	0	NaN	NaN
1	10	0	NaN	0	NaN	NaN
2	10	0	NaN	0	NaN	NaN
3	11	0	NaN	0	NaN	NaN
4	9	0	NaN	0	NaN	NaN

	NASDelay	SecurityDelay	LateAircraftDelay	type	manufacturer
0	NaN	NaN	NaN	Corporation	BOEING
1	NaN	NaN	NaN	Corporation	BOEING
2	NaN	NaN	NaN	Corporation	BOEING
3	NaN	NaN	NaN	Corporation	BOEING
4	NaN	NaN	NaN	Corporation	BOEING

Fig.2.1.11: Flight data list of aircrafts in 1999

	Year	Month	DayofMonth	DayOfWeek	DepTime	CRSDepTime	ArrTime
0	2000	1	28	5	1647.0	1647	1906.0
1	2000	1	25	2	846.0	826	1053.0
2	2000	1	6	4	1755.0	1747	2155.0
3	2000	1	19	3	1220.0	1223	1630.0
4	2000	1	1	6	1644.0	1644	1920.0

	CRSArrTime	UniqueCarrier	FlightNum	TailNum	ActualElapsedTime
0	1859	HP	154	N808AW	259.0
1	1039	HP	2027	N808AW	247.0
2	2135	HP	2442	N808AW	360.0
3	1610	HP	2792	N808AW	370.0
4	1904	HP	535	N808AW	276.0

	CRSElapsedTime	AirTime	ArrDelay	DepDelay	Origin	Dest	Distance	TaxiIn
0	252.0	233.0	7.0	0.0	ATL	PHX	1587	15
1	253.0	230.0	14.0	20.0	ATL	PHX	1587	3
2	348.0	325.0	20.0	8.0	BOS	PHX	2300	5
3	347.0	337.0	20.0	-3.0	BOS	PHX	2300	19
4	260.0	258.0	16.0	0.0	CLE	PHX	1737	6

	TaxiOut	Cancelled	CancellationCode	Diverted	CarrierDelay	WeatherDelay
0	11	0	NaN	0	NaN	NaN
1	14	0	NaN	0	NaN	NaN
2	30	0	NaN	0	NaN	NaN
3	14	0	NaN	0	NaN	NaN
4	12	0	NaN	0	NaN	NaN

	NASDelay	SecurityDelay	LateAircraftDelay	type	manufacturer
0	NaN	NaN	NaN	Corporation	AIRBUS INDUSTRIE
1	NaN	NaN	NaN	Corporation	AIRBUS INDUSTRIE
2	NaN	NaN	NaN	Corporation	AIRBUS INDUSTRIE
3	NaN	NaN	NaN	Corporation	AIRBUS INDUSTRIE
4	NaN	NaN	NaN	Corporation	AIRBUS INDUSTRIE

Fig.2.1.12: Flight data list of aircrafts in 2000

Part C For each year, create a logistic regression model for the probability of diverted US flights

In Fig.3.1.1, list the variables for the logistic regression model. Establish the flight data variables for the logistic regression. The variables are 'numerical_values', 'categorical_info', 'numerical_transformer', 'categorical_transformer' and 'data_transformer'. In Fig.3.1.2, declare the parameters for inputting to the x-axis and to convert the 'Day of Week' and 'Flightnum' variables to string. Declare in put variables from the flight data for the x-axis. After that, convert 'Day of week' to string and 'flight num' to string. Define 'Diverted flights' present as the y-axis.

In Fig.3.1.3, split the x and y axis components into the training and testing sets before using y_train as the target variable. Use 'data_transformer' and logistic regression to create a pipeline for GridSearchCV. Lastly,

plot the graph for the GridSearchCV with the False Positive Rate as the x-axis using stated flight data parameters and True Positive Rate as the y-axis with 'Diverted' flights. Referencing from Fig.3.1.4, the GridSearch CV is created using the coding information set in Fig.3.1.3. In Fig.3.1.5, the AUC curve is drawn from the GridSearchCV with the AUC value being 0.68. The probability of the US diverted flights from the logistic regression model is 0.68.

```
# Declare variables for the Logistic regression

numerical_values = ['CRSDepTime','CRSArrTime','CRSElapsedTime','Distance']
categorical_info =['DayOfWeek','UniqueCarrier',"FlightNum",'TailNum',
            'Origin','Dest']
numerical_transformer= Pipeline(steps=[
    ('imputer',SimpleImputer()),
    ('scaler',StandardScaler())])
categorical_transformer= Pipeline(steps=[
    ('imputer',SimpleImputer()),
    ('onehot',OneHotEncoder(handle_unknown='ignore'))])
data_transformer= ColumnTransformer(
    transformers=[
        ('numerical',numerical_transformer,numerical_values),
        ('categorical',categorical_transformer,categorical_info)])
```

Fig.3.1.1: The variables for Logistic regression model

```
# Declare parameters to input for x axis
parameters = ['DayOfWeek','CRSDepTime','CRSArrTime','UniqueCarrier',"FlightNum",'TailNum','CRSElapsedTime',
            'Origin','Dest','Distance']

flight_data[categorical_info].dtypes
# Convert 'Day of week' to string
flight_data['DayOfWeek'] = flight_data['DayOfWeek'].astype(str)

# Convert 'flight num' to string
flight_data['FlightNum'] = flight_data['FlightNum'].astype(str)
```

Fig.3.1.2: Jupyter codes for running the parameters in axis and to conversion of data parameters to strings

```
from sklearn.utils import column_or_1d
from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import Pipeline

# Assuming you have defined data_transformer, X, and y earlier
x = flight_data[parameters].copy()
x.head()
y = flight_data[['Diverted']].values


# Split the data into training and testing sets
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

# Assuming y_train is your target variable
y_train_reshaped = column_or_1d(y_train.ravel(), warn=False)

# Create a pipeline with Logistic regression and data transformer
pipe_lr = Pipeline(steps=[
    ('data_transformer', data_transformer),
    ('pipe_lr', LogisticRegression(max_iter=10000, solver='lbfgs', penalty='l2'))
])

# Define the parameter grid
param_grid = {
    'data_transformer__numerical__imputer__strategy': ['mean', 'median'],
    'data_transformer__categorical__imputer__strategy': ['constant', 'most_frequent'],
    'pipe_lr__penalty': ['l2']  # Only 'l2' penalty is supported for lbfgs solver
}

# Create the grid search
grid_lr = GridSearchCV(pipe_lr, param_grid=param_grid, error_score='raise')
grid_lr.fit(x_train, y_train_reshaped)
```

Fig.3.1.3: Using data transformer and training and testing data sets to create grid search

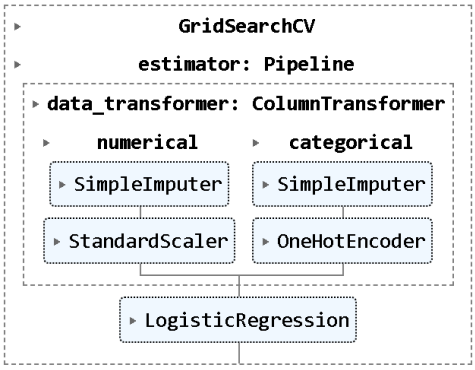


Fig.3.1.4: GridSearchCV

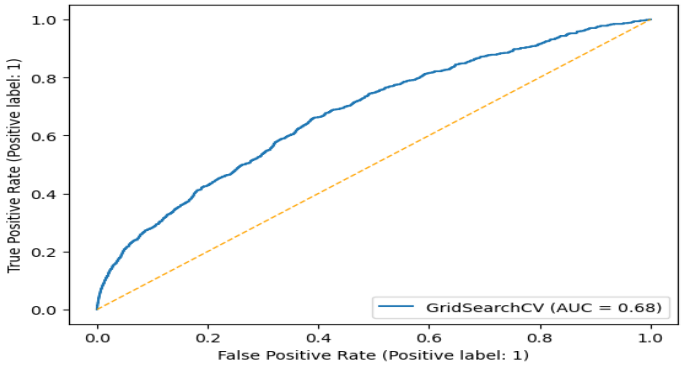


Fig.3.1.5: The AUC for the probability of diverted flights

Bibliography

FAA: Record flight delays in 2000. (2001, February 1). Retrieved from Edition CNN: https://edition.cnn.com/2001/TRAVEL/NEWS/02/01/flight.delays/

Neil Irwin, D. P. (1999, July 25). Airline Delays Grow, but Agencies Differ on Data. Retrieved from The Washington Post : https://www.washingtonpost.com/archive/business/2000/07/26/airline-delays-grow-but-agencies-differ-on-data/8214a7b7-9ef8-4113-8fae-600fbbfe45f7/