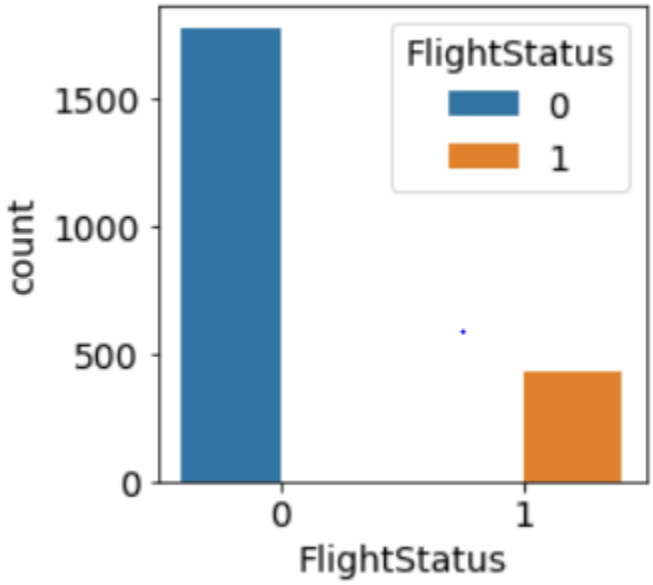
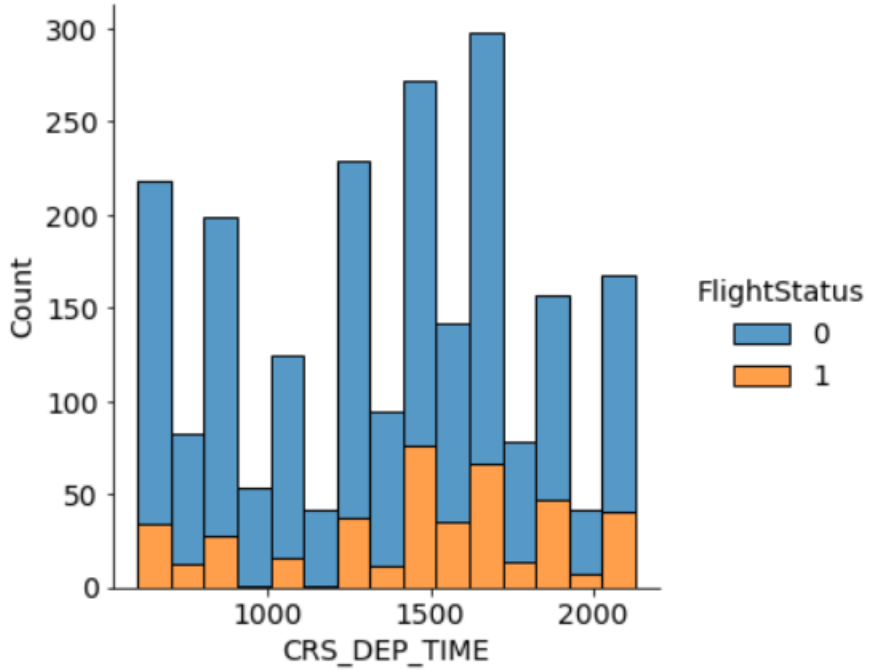
Q1).

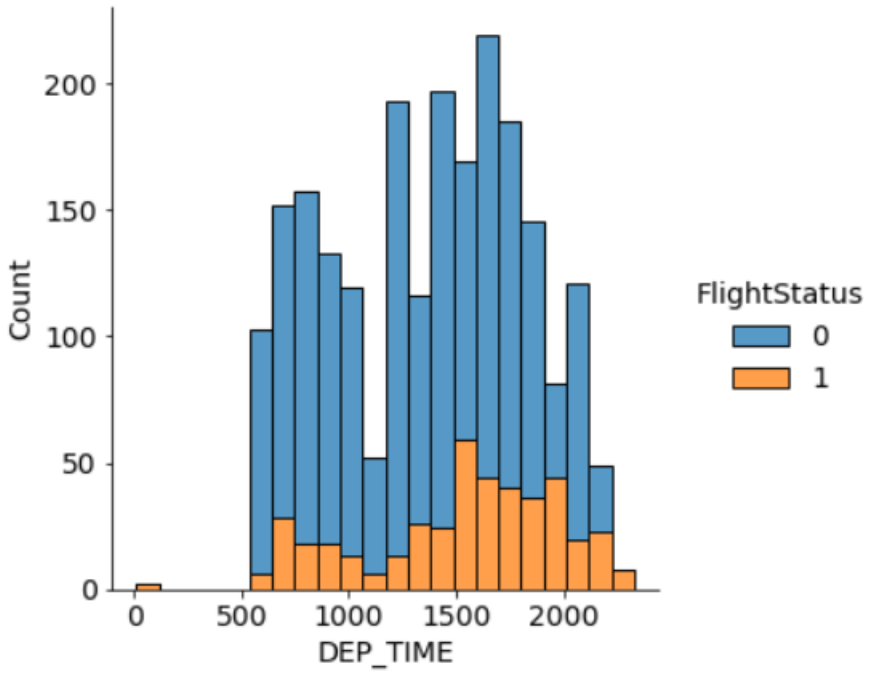
Number of on-time and delayed flights in the dataset

[0=on-time, 1=delayed]

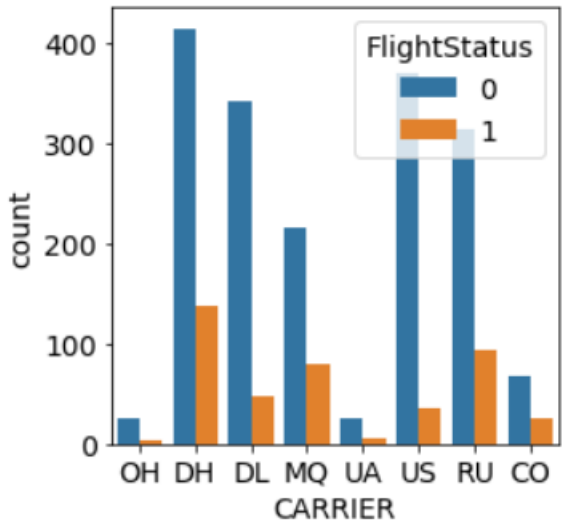




CRS_DEP_TIME is a good feature and thus can be used to train the logistic regression model for the dataset.

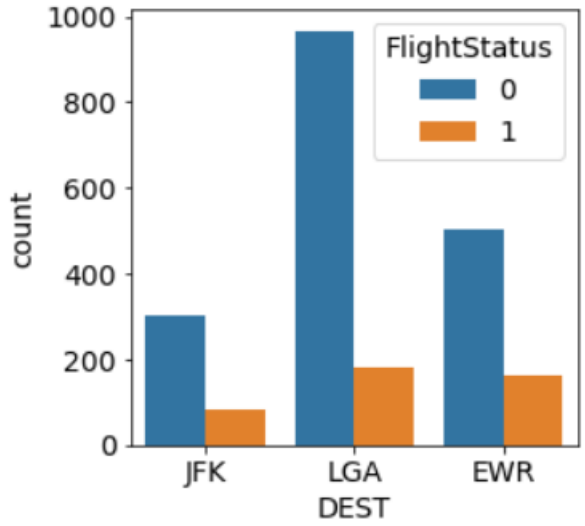


DEP_TIME is very variable and changes for flight every other time depending on other features like weather, carrier, etc. These is not a good feature as its quite dependent on the scenario thus will not improve the accuracy of the model, and we should drop these feature's column from the data set.



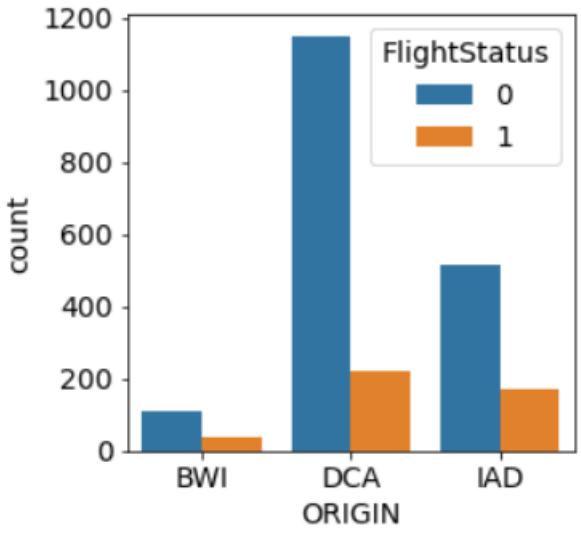
There are 8 unique carrier in the dataset.

CARRIER is a good feature and thus can be used to train the logistic regression model for the dataset.



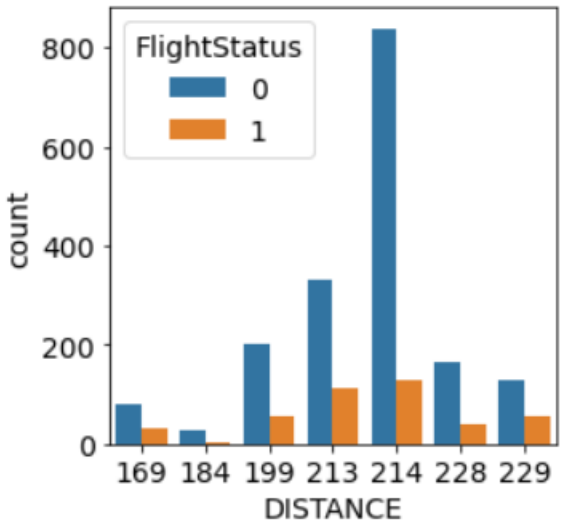
There are 3 unique destination in the dataset.

DEST is a good feature and thus can be used to train the logistic regression model for the dataset.



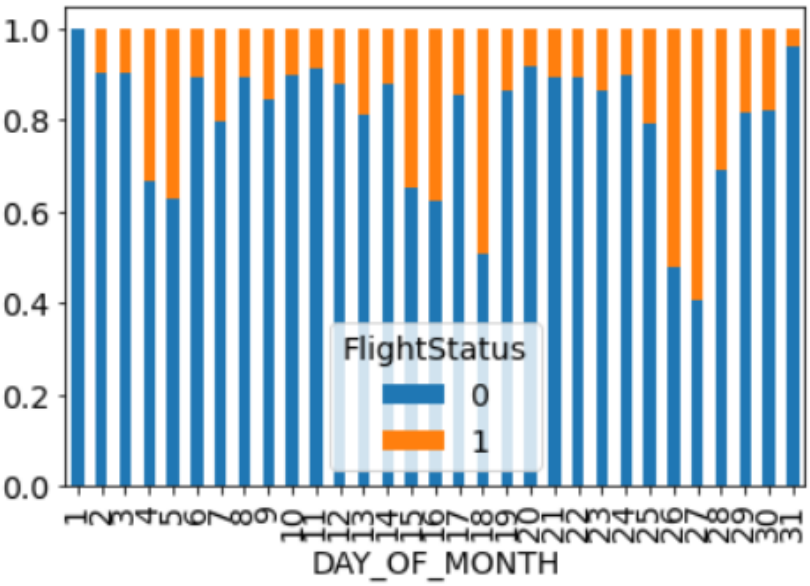
There are 3 unique origin in the dataset.

ORIGIN is a good feature and thus can be used to train the logistic regression model for the dataset.

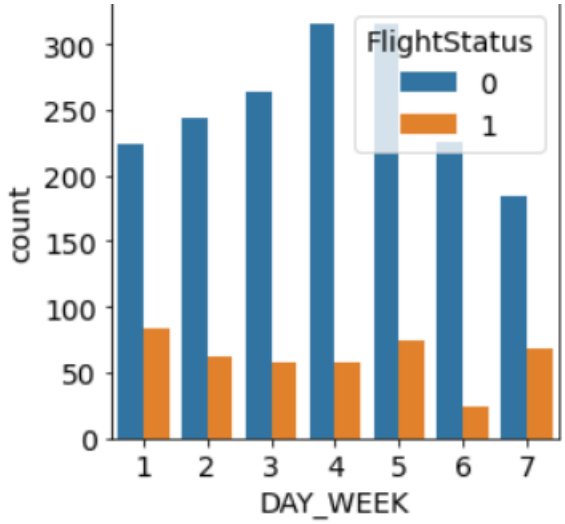


There are 9 unique distance as there are 3 origin and three destination thus 3x3.

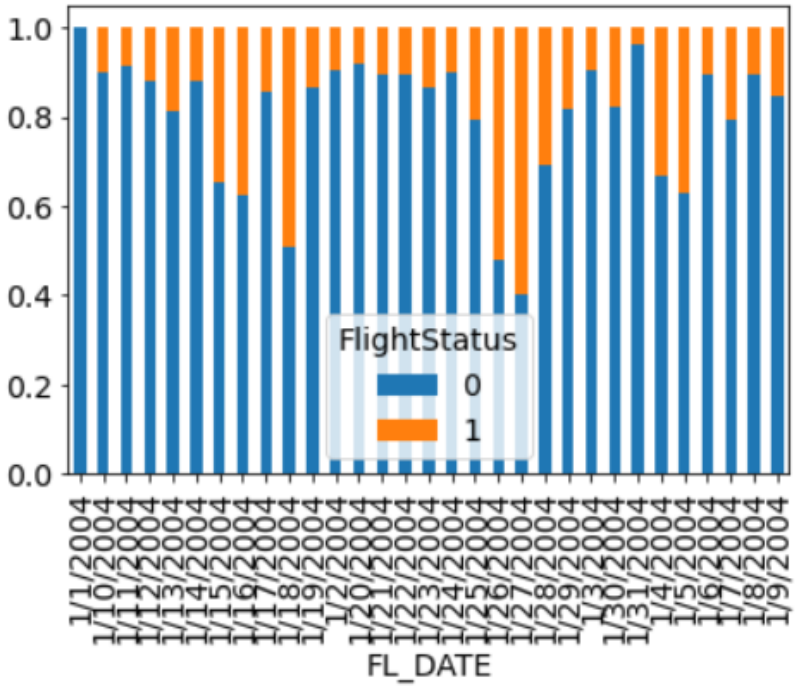
DISTANCE feature can be dropped from the dataset as we are using both destination and origin to train the model so basically, we are considering the distances through that features in our training model.



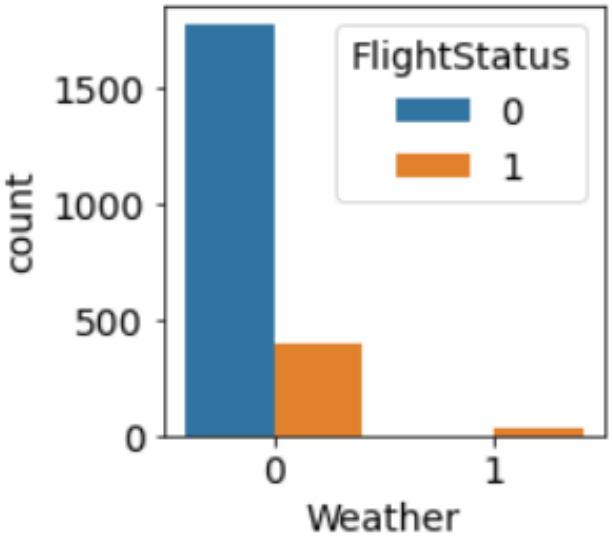
DAY_OF_MONTH is a good feature and thus can be used to train the logistic regression model for the dataset. It can be used to train a particular model as from these we can see that is their delay on certain particular day every week due to the schedule of the flight for particular destination or from particular origin or for a particular carrier.



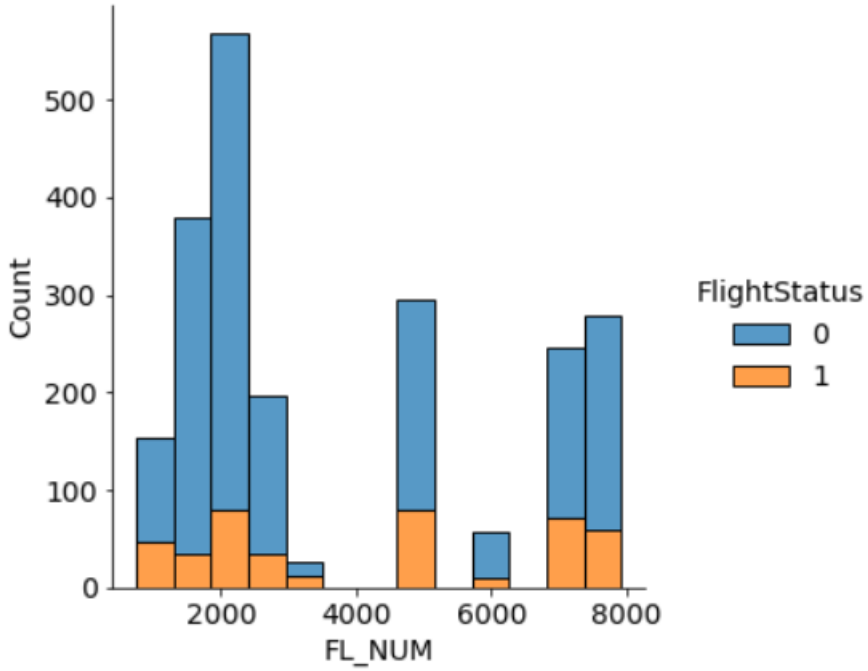
DAY_WEEK is also a good feature to train the logistic regression model for the dataset as for the same reason DAY_OF_MONTH is.



As we are considering both DAY_OF_MONTH and DAY_WEEK feature, we can drop FL_DATE feature as it's not providing any new information which will help the training model to increase its accuracy.

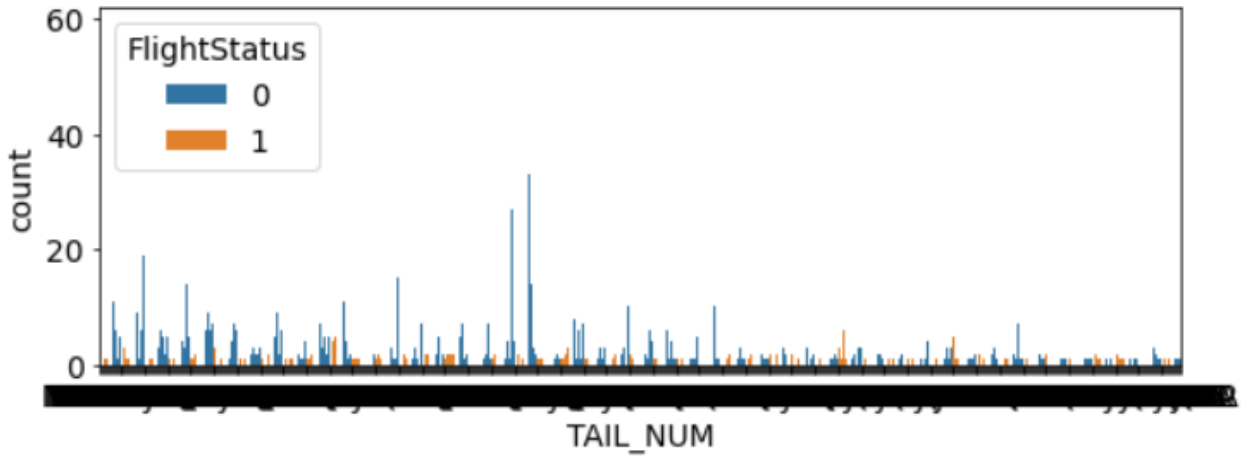


Weather is a very goof feature as we can see from the graph that whenever the weather is bad (1) there is a delay in flight. Thus, we should use it to train our model.



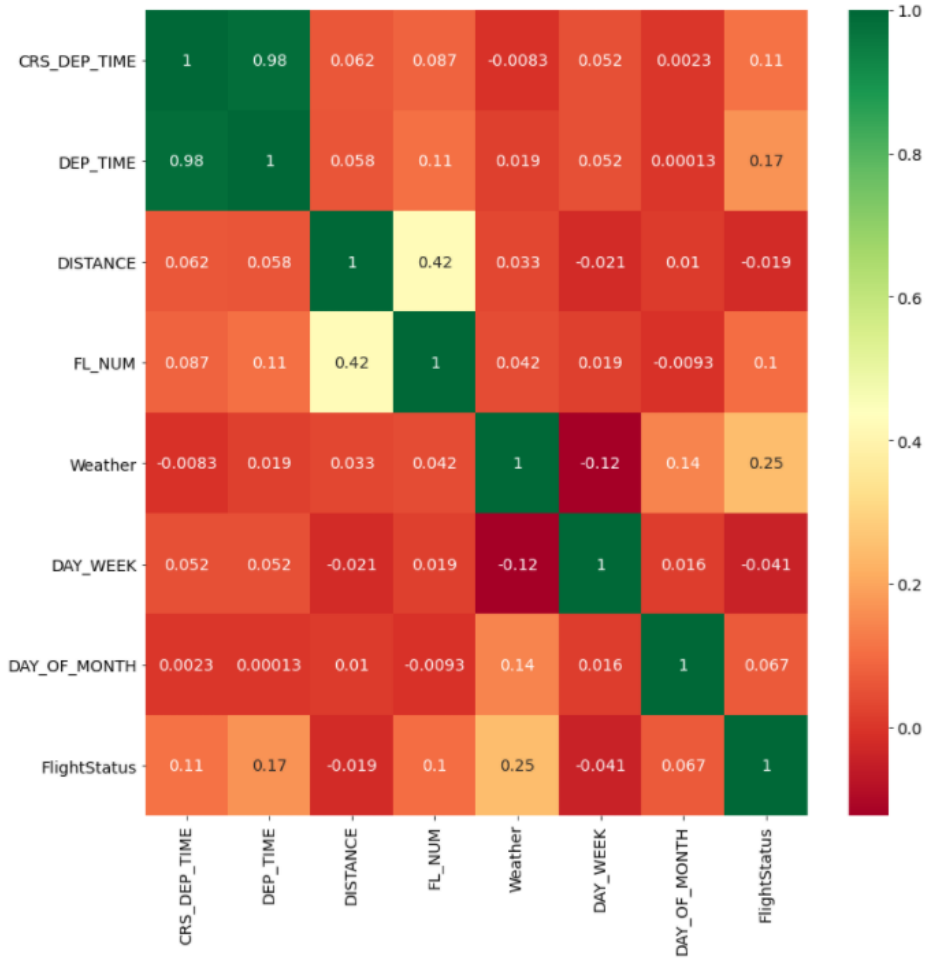
FL_NUM as we can see from the graph can be a good feature to train the model.

FL_NUM can give information about particular flight, about its daily and weekly schedule.



TAIL_NUM is not at all a good feature to train the model as there are a lot of values for this feature and thus it will not provide any helpful insight for the model to predict the flight status for test cases and even for any new cases.

HEAT MAP FOR THE DATASET



Q3).

So, after looking at all the graph and the heat map and after doing the Exploratory Data Analysis I dropped distance, TAIL_NUM, FL_DATE, DEP_TIME.

Accuracy of logistic regression classifier with all the features and dummy variables on the test data is 0.81

	coef
Weather	0.763874
DAY_OF_MONTH_27	0.352189
DAY_OF_MONTH_16	0.342942
DAY_OF_MONTH_18	0.336099
CRS_DEP_TIME_1515	0.287328
...	...
FL_NUM_2176	-0.286167
FL_NUM_2181	-0.301685
CRS_DEP_TIME_1645	-0.305673
FL_NUM_7810	-0.305673
DAY_OF_MONTH_1	-0.513250

These are the logarithmic odd and are quite not that useful and does not provide us with much insight/information.

Therefore, we will take exponent of these logarithmic odds to find normal probabilistic odds of the features, variables and dummy variables.

These below are the proper odds for the algorithm and thus it shows how one feature dominantly affects the target column flight status with respect to the other features.

Weather plays the most important role in determining the flight status for our data set with respect to other variables.

	coef
Weather	2.146577
DAY_OF_MONTH_27	1.422177
DAY_OF_MONTH_16	1.409087
DAY_OF_MONTH_18	1.399478
CRS_DEP_TIME_1515	1.332861
...	...
FL_NUM_2176	0.751137
FL_NUM_2181	0.739571
CRS_DEP_TIME_1645	0.736627
FL_NUM_7810	0.736627
DAY_OF_MONTH_1	0.598547

	precision	recall	f1-score	support
0	0.85	0.93	0.89	712
1	0.51	0.30	0.38	169
accuracy			0.81	881
macro avg	0.68	0.62	0.63	881
weighted avg	0.78	0.81	0.79	881

Interpretation of classification report: The report gives us precision and recall score of the algorithm we have trained for the test data.

Precision and **recall** are two extremely important model evaluation metrics.
While **precision** refers to the percentage of your results which are

relevant, **recall** refers to the percentage of total relevant results correctly classified by your algorithm.

Q4)

Total Variables after creating dummy are

215 and target variable ‘FlightStatus’


```
['Weather', 'CRS_DEP_TIME_600', 'CRS_DEP_TIME_630', 'CRS_DEP_TIME_640', 'CRS_
DEP_TIME_645', 'CRS_DEP_TIME_700', 'CRS_DEP_TIME_730', 'CRS_DEP_TIME_735', 'C
RS_DEP_TIME_759', 'CRS_DEP_TIME_800', 'CRS_DEP_TIME_830', 'CRS_DEP_TIME_840',
'CRS_DEP_TIME_845', 'CRS_DEP_TIME_850', 'CRS_DEP_TIME_900', 'CRS_DEP_TIME_925
', 'CRS_DEP_TIME_930', 'CRS_DEP_TIME_1000', 'CRS_DEP_TIME_1030', 'CRS_DEP_TIM
E_1039', 'CRS_DEP_TIME_1040', 'CRS_DEP_TIME_1100', 'CRS_DEP_TIME_1130', 'CRS_
DEP_TIME_1200', 'CRS_DEP_TIME_1230', 'CRS_DEP_TIME_1240', 'CRS_DEP_TIME_1245'
, 'CRS_DEP_TIME_1300', 'CRS_DEP_TIME_1315', 'CRS_DEP_TIME_1330', 'CRS_DEP_TIM
E_1359', 'CRS_DEP_TIME_1400', 'CRS_DEP_TIME_1430', 'CRS_DEP_TIME_1455', 'CRS_
DEP_TIME_1500', 'CRS_DEP_TIME_1515', 'CRS_DEP_TIME_1520', 'CRS_DEP_TIME_1525'
, 'CRS_DEP_TIME_1530', 'CRS_DEP_TIME_1600', 'CRS_DEP_TIME_1605', 'CRS_DEP_TIM
E_1610', 'CRS_DEP_TIME_1630', 'CRS_DEP_TIME_1640', 'CRS_DEP_TIME_1645', 'CRS_
DEP_TIME_1700', 'CRS_DEP_TIME_1710', 'CRS_DEP_TIME_1715', 'CRS_DEP_TIME_1720'
, 'CRS_DEP_TIME_1725', 'CRS_DEP_TIME_1730', 'CRS_DEP_TIME_1800', 'CRS_DEP_TIM
E_1830', 'CRS_DEP_TIME_1900', 'CRS_DEP_TIME_1930', 'CRS_DEP_TIME_2000', 'CRS_
DEP_TIME_2030', 'CRS_DEP_TIME_2100', 'CRS_DEP_TIME_2120', 'CRS_DEP_TIME_2130'
, 'CARRIER_CO', 'CARRIER_DH', 'CARRIER_DL', 'CARRIER_MQ', 'CARRIER_OH', 'CARR
IER_RU', 'CARRIER_UA', 'CARRIER_US', 'DEST_EWR', 'DEST_JFK', 'DEST_LGA', 'FL_
NUM_746', 'FL_NUM_806', 'FL_NUM_808', 'FL_NUM_810', 'FL_NUM_814', 'FL_NUM_816
', 'FL_NUM_846', 'FL_NUM_1479', 'FL_NUM_1740', 'FL_NUM_1742', 'FL_NUM_1744',
'FL_NUM_1746', 'FL_NUM_1748', 'FL_NUM_1750', 'FL_NUM_1752', 'FL_NUM_1754', 'F
L_NUM_1756', 'FL_NUM_1758', 'FL_NUM_1760', 'FL_NUM_1762', 'FL_NUM_1764', 'FL_
NUM_1766', 'FL_NUM_1767', 'FL_NUM_1768', 'FL_NUM_2097', 'FL_NUM_2156', 'FL_NU
M_2160', 'FL_NUM_2162', 'FL_NUM_2164', 'FL_NUM_2166', 'FL_NUM_2168', 'FL_NUM_
2170', 'FL_NUM_2172', 'FL_NUM_2174', 'FL_NUM_2176', 'FL_NUM_2178', 'FL_NUM_21
80', 'FL_NUM_2181', 'FL_NUM_2182', 'FL_NUM_2184', 'FL_NUM_2186', 'FL_NUM_2188
', 'FL_NUM_2216', 'FL_NUM_2229', 'FL_NUM_2254', 'FL_NUM_2261', 'FL_NUM_2267',
'FL_NUM_2303', 'FL_NUM_2336', 'FL_NUM_2361', 'FL_NUM_2367', 'FL_NUM_2385', 'F
L_NUM_2403', 'FL_NUM_2497', 'FL_NUM_2582', 'FL_NUM_2603', 'FL_NUM_2664', 'FL_
NUM_2675', 'FL_NUM_2692', 'FL_NUM_2703', 'FL_NUM_2761', 'FL_NUM_2855', 'FL_NU
M_2879', 'FL_NUM_3276', 'FL_NUM_3372', 'FL_NUM_4752', 'FL_NUM_4760', 'FL_NUM_
4771', 'FL_NUM_4784', 'FL_NUM_4952', 'FL_NUM_4954', 'FL_NUM_4956', 'FL_NUM_49
60', 'FL_NUM_4964', 'FL_NUM_4966', 'FL_NUM_4968', 'FL_NUM_4970', 'FL_NUM_4972
', 'FL_NUM_4976', 'FL_NUM_5935', 'FL_NUM_6155', 'FL_NUM_7208', 'FL_NUM_7211',
'FL_NUM_7215', 'FL_NUM_7299', 'FL_NUM_7302', 'FL_NUM_7303', 'FL_NUM_7304', 'F
L_NUM_7305', 'FL_NUM_7307', 'FL_NUM_7371', 'FL_NUM_7684', 'FL_NUM_7790', 'FL_
NUM_7792', 'FL_NUM_7800', 'FL_NUM_7806', 'FL_NUM_7808', 'FL_NUM_7810', 'FL_NU
M_7812', 'FL_NUM_7814', 'FL_NUM_7816', 'FL_NUM_7818', 'FL_NUM_7924', 'ORIGIN_
BWI', 'ORIGIN_DCA', 'ORIGIN_IAD', 'DAY_WEEK_1', 'DAY_WEEK_2', 'DAY_WEEK_3', '
DAY_WEEK_4', 'DAY_WEEK_5', 'DAY_WEEK_6', 'DAY_WEEK_7', 'DAY_OF_MONTH_1', 'DAY
_OF_MONTH_2', 'DAY_OF_MONTH_3', 'DAY_OF_MONTH_4', 'DAY_OF_MONTH_5', 'DAY_OF_M
ONTH_6', 'DAY_OF_MONTH_7', 'DAY_OF_MONTH_8', 'DAY_OF_MONTH_9', 'DAY_OF_MONTH_
10', 'DAY_OF_MONTH_11', 'DAY_OF_MONTH_12', 'DAY_OF_MONTH_13', 'DAY_OF_MONTH_1
4', 'DAY_OF_MONTH_15', 'DAY_OF_MONTH_16', 'DAY_OF_MONTH_17', 'DAY_OF_MONTH_18
', 'DAY_OF_MONTH_19', 'DAY_OF_MONTH_20', 'DAY_OF_MONTH_21', 'DAY_OF_MONTH_22'
, 'DAY_OF_MONTH_23', 'DAY_OF_MONTH_24', 'DAY_OF_MONTH_25', 'DAY_OF_MONTH_26',
```

'DAY_OF_MONTH_27', 'DAY_OF_MONTH_28', 'DAY_OF_MONTH_29', 'DAY_OF_MONTH_30', '
DAY_OF_MONTH_31']

Variables Selected [SIGNIFICANT]

['Weather', 'CRS_DEP_TIME_630', 'CRS_DEP_TIME_930', 'CRS_DEP_TIME_1040', 'CRS
_DEP_TIME_1330', 'CRS_DEP_TIME_1515', 'CRS_DEP_TIME_1525', 'CRS_DEP_TIME_1645
', 'CRS_DEP_TIME_1900', 'CARRIER_DL', 'CARRIER_US', 'FL_NUM_746', 'FL_NUM_806
', 'FL_NUM_814', 'FL_NUM_1479', 'FL_NUM_2166', 'FL_NUM_2170', 'FL_NUM_2174',
'FL_NUM_2182', 'FL_NUM_2761', 'FL_NUM_3372', 'FL_NUM_4760', 'FL_NUM_4970', 'F
L_NUM_7211', 'FL_NUM_7299', 'FL_NUM_7810', 'DAY_OF_MONTH_1', 'DAY_OF_MONTH_4'
, 'DAY_OF_MONTH_5', 'DAY_OF_MONTH_15', 'DAY_OF_MONTH_16', 'DAY_OF_MONTH_18',
'DAY_OF_MONTH_26', 'DAY_OF_MONTH_27', 'DAY_OF_MONTH_28', 'DAY_OF_MONTH_31']

Variables Not Selected [NOT SIGNIFICANT]

['CRS_DEP_TIME_600', 'CRS_DEP_TIME_640', 'CRS_DEP_TIME_645', 'CRS_DEP_TIME_70
0', 'CRS_DEP_TIME_730', 'CRS_DEP_TIME_735', 'CRS_DEP_TIME_759', 'CRS_DEP_TIME
_800', 'CRS_DEP_TIME_830', 'CRS_DEP_TIME_840', 'CRS_DEP_TIME_845', 'CRS_DEP_T
IME_850', 'CRS_DEP_TIME_900', 'CRS_DEP_TIME_925', 'CRS_DEP_TIME_1000', 'CRS_D
EP_TIME_1030', 'CRS_DEP_TIME_1039', 'CRS_DEP_TIME_1100', 'CRS_DEP_TIME_1130',
'CRS_DEP_TIME_1200', 'CRS_DEP_TIME_1230', 'CRS_DEP_TIME_1240', 'CRS_DEP_TIME_
1245', 'CRS_DEP_TIME_1300', 'CRS_DEP_TIME_1315', 'CRS_DEP_TIME_1359', 'CRS_DE
P_TIME_1400', 'CRS_DEP_TIME_1430', 'CRS_DEP_TIME_1455', 'CRS_DEP_TIME_1500',
'CRS_DEP_TIME_1520', 'CRS_DEP_TIME_1530', 'CRS_DEP_TIME_1600', 'CRS_DEP_TIME_
1605', 'CRS_DEP_TIME_1610', 'CRS_DEP_TIME_1630', 'CRS_DEP_TIME_1640', 'CRS_DE
P_TIME_1700', 'CRS_DEP_TIME_1710', 'CRS_DEP_TIME_1715', 'CRS_DEP_TIME_1720',
'CRS_DEP_TIME_1725', 'CRS_DEP_TIME_1730', 'CRS_DEP_TIME_1800', 'CRS_DEP_TIME_
1830', 'CRS_DEP_TIME_1930', 'CRS_DEP_TIME_2000', 'CRS_DEP_TIME_2030', 'CRS_DE
P_TIME_2100', 'CRS_DEP_TIME_2120', 'CRS_DEP_TIME_2130', 'CARRIER_CO', 'CARRIE
R_DH', 'CARRIER_MQ', 'CARRIER_OH', 'CARRIER_RU', 'CARRIER_UA', 'DEST_EWR', 'D
EST_JFK', 'DEST_LGA', 'FL_NUM_808', 'FL_NUM_810', 'FL_NUM_816', 'FL_NUM_846',
'FL_NUM_1740', 'FL_NUM_1742', 'FL_NUM_1744', 'FL_NUM_1746', 'FL_NUM_1748', 'F
L_NUM_1750', 'FL_NUM_1752', 'FL_NUM_1754', 'FL_NUM_1756', 'FL_NUM_1758', 'FL_
NUM_1760', 'FL_NUM_1762', 'FL_NUM_1764', 'FL_NUM_1766', 'FL_NUM_1767', 'FL_NU
M_1768', 'FL_NUM_2097', 'FL_NUM_2156', 'FL_NUM_2160', 'FL_NUM_2162', 'FL_NUM_
2164', 'FL_NUM_2168', 'FL_NUM_2172', 'FL_NUM_2176', 'FL_NUM_2178', 'FL_NUM_21
80', 'FL_NUM_2181', 'FL_NUM_2184', 'FL_NUM_2186', 'FL_NUM_2188', 'FL_NUM_2216
', 'FL_NUM_2229', 'FL_NUM_2254', 'FL_NUM_2261', 'FL_NUM_2267', 'FL_NUM_2303',
'FL_NUM_2336', 'FL_NUM_2361', 'FL_NUM_2367', 'FL_NUM_2385', 'FL_NUM_2403', 'F
L_NUM_2497', 'FL_NUM_2582', 'FL_NUM_2603', 'FL_NUM_2664', 'FL_NUM_2675', 'FL_
NUM_2692', 'FL_NUM_2703', 'FL_NUM_2855', 'FL_NUM_2879', 'FL_NUM_3276', 'FL_NU
M_4752', 'FL_NUM_4771', 'FL_NUM_4784', 'FL_NUM_4952', 'FL_NUM_4954', 'FL_NUM_
4956', 'FL_NUM_4960', 'FL_NUM_4964', 'FL_NUM_4966', 'FL_NUM_4968', 'FL_NUM_49
72', 'FL_NUM_4976', 'FL_NUM_5935', 'FL_NUM_6155', 'FL_NUM_7208', 'FL_NUM_7215
', 'FL_NUM_7302', 'FL_NUM_7303', 'FL_NUM_7304', 'FL_NUM_7305', 'FL_NUM_7307',
'FL_NUM_7371', 'FL_NUM_7684', 'FL_NUM_7790', 'FL_NUM_7792', 'FL_NUM_7800', 'F
L_NUM_7806', 'FL_NUM_7808', 'FL_NUM_7812', 'FL_NUM_7814', 'FL_NUM_7816', 'FL_
NUM_7818', 'FL_NUM_7924', 'ORIGIN_BWI', 'ORIGIN_DCA', 'ORIGIN_IAD', 'DAY_WEEK
_1', 'DAY_WEEK_2', 'DAY_WEEK_3', 'DAY_WEEK_4', 'DAY_WEEK_5', 'DAY_WEEK_6', 'D
AY_WEEK_7', 'DAY_OF_MONTH_2', 'DAY_OF_MONTH_3', 'DAY_OF_MONTH_6', 'DAY_OF_MON
TH_7', 'DAY_OF_MONTH_8', 'DAY_OF_MONTH_9', 'DAY_OF_MONTH_10', 'DAY_OF_MONTH_1
1', 'DAY_OF_MONTH_12', 'DAY_OF_MONTH_13', 'DAY_OF_MONTH_14', 'DAY_OF_MONTH_17
', 'DAY_OF_MONTH_19', 'DAY_OF_MONTH_20', 'DAY_OF_MONTH_21', 'DAY_OF_MONTH_22'

, 'DAY_OF_MONTH_23', 'DAY_OF_MONTH_24', 'DAY_OF_MONTH_25', 'DAY_OF_MONTH_29', 'DAY_OF_MONTH_30']

Q5.)

So, after doing the recursive feature elimination and selecting the significant variable we trained a new logistic regression model for our test data.

Accuracy of logistic regression classifier with the selected significant features on the test data is 0.84.
So, the accuracy of the classifier after doing selection of variable increased to 0.84 from 0.81.

Classification report for this model is:

	precision	recall	f1-score	support
0	0.87	0.94	0.91	714
1	0.63	0.40	0.49	167
accuracy			0.84	881
macro avg	0.75	0.67	0.70	881
weighted avg	0.82	0.84	0.83	881

Q6).
CARRIER – US AIRWAYS
DAY OF MONTH – 1
DAY OF WEEK – 6
CRS_DEP_TIME – 1645

BONUS QUESTIONS

1. KAREN, JOCASTA, VERONICA, HELEN, TADASHI.
2. The **Data processing inequality** is an information theoretic concept which states that the information content of a signal cannot be increased via a local physical operation. This can be expressed concisely as 'post-processing cannot increase information'.
3. Sheev Palpatine
4. C-3PO And R2-D2
5. This year for Black Friday, we taught a computer how to write Cards Against Humanity c ards. Now we put it to the test. Over the next 16 hours, our writers will battle this powerfu l card-writing algorithm to see who can write the most popular new pack of cards. If the writers win, they'll get a $5,000 holiday bonus. If the A.I. wins, we'll fire the writers. 😊