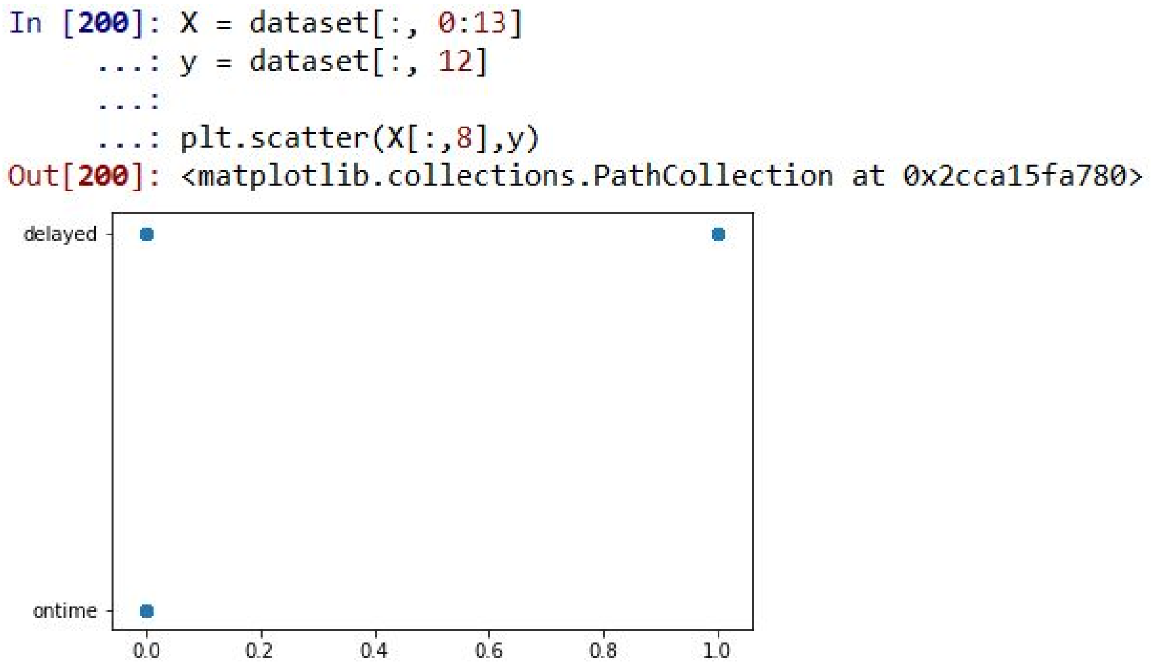
ASSIGNMENT 2-GNR 652

Questions/Answers

Ans1)Here are few plots using matplotlib library using scatter
Let’s analyze output with Weather



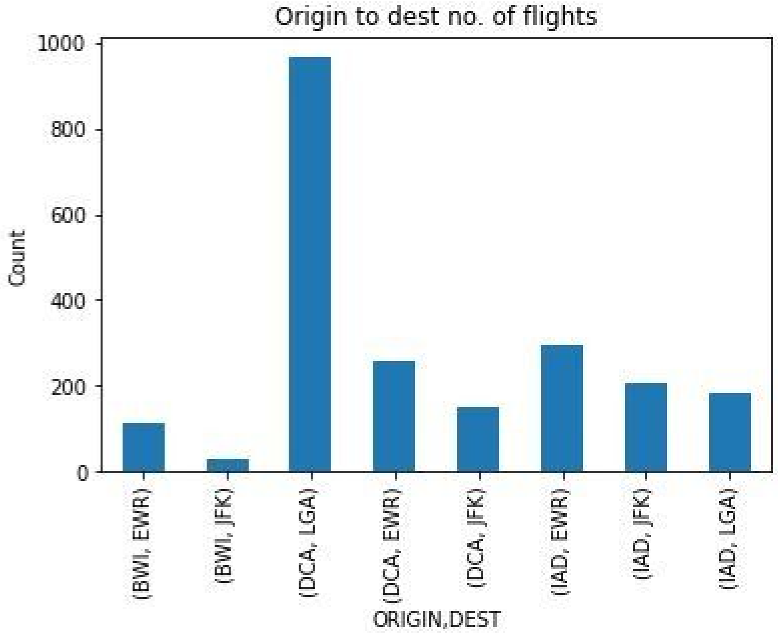
```
In [200]: X = dataset[:, 0:13]
     ...: y = dataset[:, 12]
     ...:
     ...: plt.scatter(X[:,8],y)
Out[200]: <matplotlib.collections.PathCollection at 0x2cca15fa780>
```

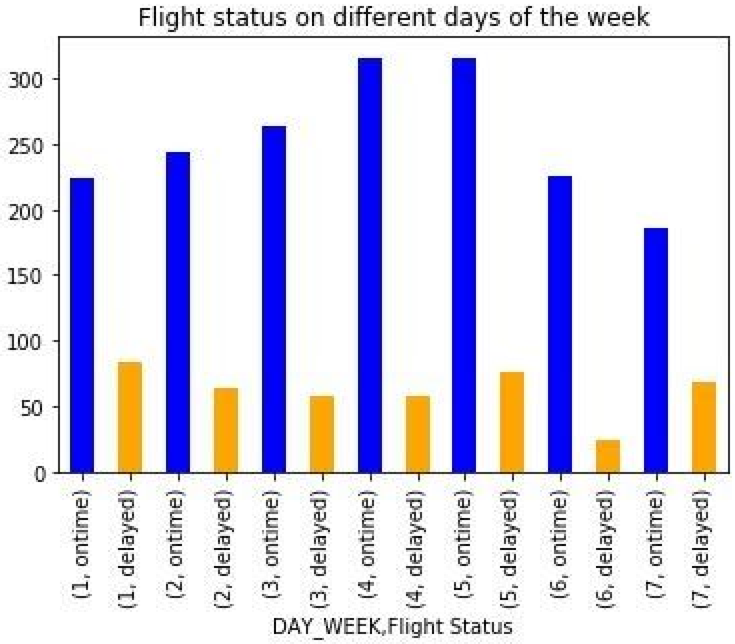
Here X axis is value of weather and Y axis is value of flight status,WE can easily observe when the weather has the value 1(bad) the flight is always delayed ,we can hardcode it by

```
indx=np.where(X_test[:,7]==1)
for i in indx:
    y_pred[i]=1
```

BUT WE WONT As We are interested in ML not classical programming.

Few more plots





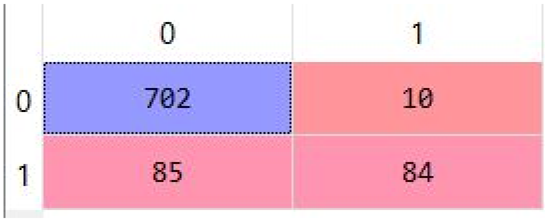
ANS2)The dataset is preprocessed by using the sklearn library and using label encoders to map strings to numerals and then i have used one hot encoding to convert these categorical data into numerical independent variables,Because of this there will be new dummy variables
The dataset is divided by following(60%,40%)...

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.4, random_state = 0)
```

ANS3)The model is basic logistic regression,the input X_train is a (1320,639) vector with all variables converted to numerical values by label encoding and extra variable 1 representing the intercept term.Weights initialized to 0 initially i.e W=np.zeros((639,1)). Model is trained by batch gradient descent and results are evaluated by ‘accuracy’ and ‘f1_score’
The accuracy obtained is 79% and f score is 0.32(bad)
.

ANS4)I have removed Flight date and day of the month as it is redundant as we already have the feature **day of the week** and it more reasonable,and also i have removed the FL_NUM and TAIL_NUM as it plays no role on deciding the status of the flight.Also i removed the actual departure time and kept CRS_DEPT_TIME because while predicting for new flights we do not have the info of actual dept_time.If we were to keep both then no need for training we can directly y_pred=(X[:,2]-X[:,0]>15).
Now the input matrix X_train is reduced to size
(1320,27)

ANS5)After fitting a new model,The accuracy has improved and also the F1_score.The new accuracy value is 89.55%
The new f score is 0.6567
Also the confusion matrix is



	0	1
0	702	10
1	85	84

Note> I have calculated the f score value because the given dataset has approx 80:20 distribution of class labels which is a bit skewed dataset.

ANS6)I’ve answered these on the basis of histogram plots y vs feature ,the ideal weather conditions for the highest chance of an ontime flight from DC to New York is (0(good),between 1234PM to 1311 PM , day 4, USairways)
It is a very good question.

Bonus Questions/answers

ANS1)AI’s made by Tony Stark are - Dummy,Veronica,Karen

ANS2)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'.
Intuitively, the data processing inequality says that no clever transformation of the received code (channel
output) Y can give more information about the sent code (channel input) X than Y itself.

ANS3)The Rule of Two

ANS4)**C-3PO** and **R2-D2**

ANS5) The creators of *Cards Against Humanity* are back for their annual Black Friday stunt, and this one is delightfully dystopian. Starting at 11AM ET today and lasting for the next 16 hours, the human writers on the *CAH* team are facing off against an artificial intelligence to see who can create the most popular new pack of cards, based on how many people pay for more $5 packs
They actually built a real ML algorithm that generates cards against humanity cards.