



Pune District Education Association's
College Of Engineering

Manjari (Bk.), Hadapsar, Pune-412307.

Accredited by NAAC



SBDAL

Assignment No:-05

Title:- Data Analytics II

Objective:- students should be able to data analysis using logistics regression using python for any open source dataset.

Aim:- Data Analytics II

1) Implement logistic regression using python to perform classification on social-network Ads. csv dataset.

2) compute confusion matrix to find TP, FP, TN, FN, Accuracy, Error rate, precision.

Requirements:- 1) Basic of python programming.
2) concept of Regression.

Theory:-

logistic Regression:- social network Ads.

This project will be a walk through of a simple logistic Regression model in an attempt to strategies a basic ad-targetting campigon for a social media network.

one of four sponsors advertisements seems to be particularly successful among

our older neither users but seemingly less- so with your younger.

- We like to implement an appropriate so that we know who our target audience is this specific advertisement, thus the maximizes click-through rate.

- our dataset contains some information about all of our users in social network.

- If we wanted to determine the effect more independent variables on the outcome we would have to implement a dimensionality reduction aspect to the model because only describe so many dimensions visually.

- worried about how the user's age & estimate salary effect their decision on click or not clients on adv.

extracting relevant vectors independent variables (x), dependent variables (y),

- Split data into two sets:- train set to learn m/c from & test set for m/c to execute on.

- This process is referred to as cross validation & we will be the implemented scikit learn's appropriately. Industry standard usually calls for a training set size of 70-80%. So well split the two.



Pune District Education Association's
College Of Engineering

Manjari (Bk.), Hadapsar, Pune-412307.

Accredited by NAAC



When we look at both the model together we can actually see that there is a shape to this data that's becoming increasingly apparent as the no. of observation increases.

- The best X-intercept, is probably, closer to the than is to be 2, and the y-intercept is likely between 2 & 3.

- This function will compare the calculated result in our y-pred vector to the actual observation results in y-test to determine how similar they are.

- The more values that match, the higher the accuracy of the classifier.

* Conclusion:-

The confusion matrix tells us that there were correct predictions and 11 incorrect ones, meaning the model overall accomplished an 89.1% accuracy rating.

* csv file / dataset - social_network_Ads.csv.

Required libraries-

import pandas as pd

import numpy as np

import matplotlib as plt

import seaborn as sns


```
from sklearn.preprocessing import StandardScaler.  
from sklearn.model_selection import train_test_split.  
from sklearn.metrics import confusion_matrix,  
classification_report, accuracy_score,  
precision_score, recall_score, f1_score.  
import warnings.
```

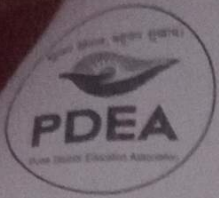
* function used -

```
df = pd.read_csv("social_network_ads.csv")  
df.head()  
df.shape  
df.info()  
df.describe()  
df.isnull().sum()  
histplot = sns.histplot()  
plt.show()  
Draw histogram for each column  
df["column name"].value_counts()  
countplot = sns.countplot()  
sns.heatmap()
```

- Data preparation
- model building
- Evaluation

Q.1) Explain confusion matrix with Accuracy error rate, precision & Recall.

→ It contains Actual value & predicted value.



Pune District Education Association's
College Of Engineering

Manjari (Bk.), Hadapsar, Pune-412307.

Accredited by NAAC



- Terms included in confusion matrix are -

i) True Negative (TN)

ii) True positive

iii) False positive

iv) False Negative

e.g. patient have disease with sample 165
confusion matrix predicted

165

No नरुजन हिताय, नरुह नरुखाय। Yes

No

50
[TN]

10
[FP]

60

Yes

5
[FN]

100
[TP]

105

55

110

PDEA
Pune District Education Association

Actual value - which are already true,
its reality.

predicted values - After some experiment

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}} = \frac{100 + 50}{165} = 0.91$$

$$\text{Error rate} = 1 - \text{Accuracy} \text{ or } \frac{FP + FN}{\text{Total}} = 0.09$$

$$\text{precision} = \frac{TP}{\text{predicted Yes}} = \frac{100}{110} = 0.64$$

$$\text{Recall} = \frac{TP}{\text{actual Yes}} = \frac{100}{105} = 0.95$$