DATA SCIENC ESSENTIALS FINAL PROJECT ON

Survey of Labour and Income Dynamics

By

MANIDEEP MADDIPATLA
UIN :- 650193002

AIM :-

To create machine learning model to predict the wages of the labour.

Technique used:

We are given with data file, contains information about 7425 labour data examined and their wages,age,sex and their education. The main idea here is building multiple models with different sample and different initial variables from given data set. The attribute that which provides more information to analyze the wages should be determined. This model will take the categorial variables as sex,age,education it will predict the wages of the labour and this process will continue until all the set of variables are compared and the wages is predicted.

In the case i considered i got 62% accuracy when i applied the logistic regression model When i used the logistic regression method .

Procedure :-

First i started with the histograms and plotted all the given data in the bar graphs. Then i plotted the graph by giving certain value like df[(df.age <17) & (df.sex==‘Male')].wages.value_counts() .plot(kind=‘barh’). then i checked the mean , max and average value fot the attributes having the integer value.

Missing Values :-

The SLID data set contains many NAN values. When we build the model the data in the one attribute is been compared with the data in the other attributes so we need all the data to be of same data type and there should not be any null value in the data set.
I replaced the Null values by taking the average values in that current attribute and apply the mean value to all the data that which is NAN.

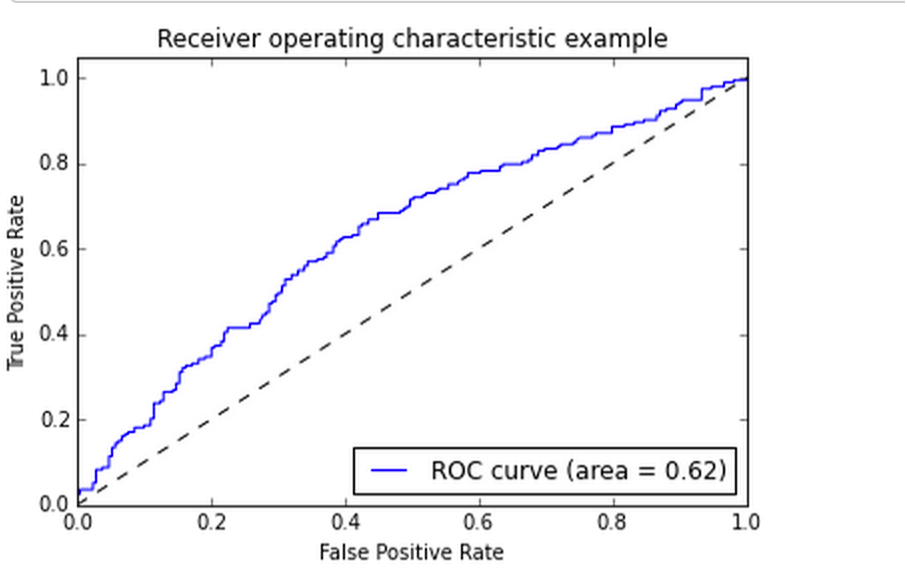
	Unnamed: 0	wages	education	age
count	7425.000000	4147.000000	7176.000000	7425.000000
mean	3713.000000	15.553082	12.496084	43.982761
std	2143.557207	7.883066	3.362506	17.694554
min	1.000000	2.300000	0.000000	16.000000
25%	1857.000000	9.235000	10.300000	30.000000
50%	3713.000000	14.090000	12.100000	41.000000
75%	5569.000000	19.800000	14.525000	57.000000
max	7425.000000	49.920000	20.000000	95.000000

In wages and education the data is of type decimal and the age is integer so i decided to make all the data into the type integer and i converted it using the command.
pd.options.display.float_format = '{:,.0f}'.format

And then i took the attribute what needed to be compared as the categorical variables and dropped the attributes what ever is not necessary . Here in this project i felt that the language has no any link with the wages so i dropped the attribute language. Then i had cleaned the data by deleting the null values that are not needed for the model prediction .

Performance of the model :-

```
---Base Model---
Base Rate AUC = 0.50
             precision    recall  f1-score   support

        0.0       0.52      1.00      0.68       209
        1.0       0.00      0.00      0.00       193

avg / total       0.27      0.52      0.36       402


---Logistic Model---
Logistic AUC = 0.62
             precision    recall  f1-score   support

        0.0       0.65      0.55      0.60       209
        1.0       0.58      0.68      0.63       193

avg / total       0.62      0.61      0.61       402
```



The problem i faced while doing the project is i was always getting the error
ValueError: continuous format is not supported and
ValueError: Can't handle mix of continuous and binary
I was not able to find the solution for long time and after that i taught that if i change all the data to the same type and then start comparing the data.
For this i changed all the decimal values to integer and in the sex attribute i assigned the binary values for Female and male values . After converting into the integer value i then changed substituted all the integer value with the binary value then the error was cleared and i was able to get the result.