Assignment-2-Priyangka-Roy

June 19, 2021

[1]:
```
import pandas as pd
import numpy as np
df=pd.read_csv('agora.csv')
df.head()
```

[1]:

	Marketing Spend	Administration	Transport	Area	Profit
0	114523.61	136897.80	471784.10	Dhaka	192261.83
1	162597.70	151377.59	443898.53	Ctg	191792.06
2	153441.51	101145.55	407934.54	Rangpur	191050.39
3	144372.41	118671.85	383199.62	Dhaka	182901.99
4	142107.34	91391.77	366168.42	Rangpur	166187.94

[2]:
```
df.isnull().sum()
```

[2]:
```
Marketing Spend    0
Administration     0
Transport          1
Area               0
Profit             0
dtype: int64
```

[3]:
```
#HANDLE NULL VALUES
median=df.Transport.median()
median
```

[3]:
```
214634.81
```

[4]:
```
df.Transport=df.Transport.fillna(median)
df.isnull().sum()
```

[4]:
```
Marketing Spend    0
Administration     0
Transport          0
Area               0
Profit             0
dtype: int64
```

1 ENCODING

[5]:
```
df=pd.read_csv('agora.csv')
df.head()
```

[5]:

	Marketing Spend	Administration	Transport	Area	Profit
0	114523.61	136897.80	471784.10	Dhaka	192261.83
1	162597.70	151377.59	443898.53	Ctg	191792.06
2	153441.51	101145.55	407934.54	Rangpur	191050.39
3	144372.41	118671.85	383199.62	Dhaka	182901.99
4	142107.34	91391.77	366168.42	Rangpur	166187.94

[6]:
```
df.Area.unique()
```

[6]:
```
array(['Dhaka', 'Ctg', 'Rangpur'], dtype=object)
```

[7]:
```
df.Area = df.Area.replace(['Dhaka', 'Ctg', 'Rangpur'],[3,2,1])
```

[8]:
```
df.Area.head()
```

[8]:
```
0    3
1    2
2    1
3    3
4    1
Name: Area, dtype: int64
```

2 LABEL ENCODING

[9]:
```
from sklearn.preprocessing import LabelEncoder
df.head()
```

[9]:

	Marketing Spend	Administration	Transport	Area	Profit
0	114523.61	136897.80	471784.10	3	192261.83
1	162597.70	151377.59	443898.53	2	191792.06
2	153441.51	101145.55	407934.54	1	191050.39
3	144372.41	118671.85	383199.62	3	182901.99
4	142107.34	91391.77	366168.42	1	166187.94

[10]:
```
Label=LabelEncoder()
df.Area=Label.fit_transform(df['Area'])
df.head()
```

[10]:

	Marketing Spend	Administration	Transport	Area	Profit
0	114523.61	136897.80	471784.10	2	192261.83

```
1        162597.70       151377.59  443898.53     1  191792.06
2        153441.51       101145.55  407934.54     0  191050.39
3        144372.41       118671.85  383199.62     2  182901.99
4        142107.34        91391.77  366168.42     0  166187.94
```

[12]:
```
#loop
for column in df.columns:
    if df[column].dtype==np.number:
        continue
    df[column]=LabelEncoder().fit_transform(df[column])
```

[13]:
```
df.Area.head()
```

[13]:
```
0    2
1    1
2    0
3    2
4    0
Name: Area, dtype: int64
```

3 One Hot Encoding

[15]:
```
df=pd.read_csv('agora.csv')
df.head()
```

[15]:
```
   Marketing Spend  Administration  Transport     Area     Profit
0        114523.61       136897.80  471784.10    Dhaka  192261.83
1        162597.70       151377.59  443898.53      Ctg  191792.06
2        153441.51       101145.55  407934.54  Rangpur  191050.39
3        144372.41       118671.85  383199.62    Dhaka  182901.99
4        142107.34        91391.77  366168.42  Rangpur  166187.94
```

[16]:
```
pd.get_dummies(df['Area'])
```

[16]:
```
    Ctg  Dhaka  Rangpur
0     0      1        0
1     1      0        0
2     0      0        1
3     0      1        0
4     0      0        1
5     0      1        0
6     1      0        0
7     0      0        1
8     0      1        0
9     1      0        0
```

```
10    0       0         1
11    1       0         0
12    0       0         1
13    1       0         0
14    0       0         1
15    0       1         0
16    1       0         0
17    0       1         0
18    0       0         1
19    0       1         0
20    1       0         0
21    0       1         0
22    0       0         1
23    0       0         1
24    0       1         0
25    1       0         0
26    0       0         1
27    0       1         0
28    0       0         1
29    0       1         0
30    0       0         1
31    0       1         0
32    1       0         0
33    0       0         1
34    1       0         0
35    0       1         0
36    0       0         1
37    1       0         0
38    0       1         0
39    1       0         0
40    1       0         0
41    0       0         1
42    1       0         0
43    0       1         0
44    1       0         0
45    0       1         0
46    0       0         1
47    1       0         0
48    0       1         0
49    1       0         0
```

[17]:
```
dummy_variables = pd.get_dummies(df['Area'],drop_first=True)
dummy_variables.head()
```

[17]:
```
   Dhaka  Rangpur
0      1        0
1      0        0
```

```
2      0        1
3      1        0
4      0        1
```

[18]:
```
df.head()
```

[18]:
```
   Marketing Spend  Administration  Transport     Area     Profit
0        114523.61       136897.80  471784.10    Dhaka  192261.83
1        162597.70       151377.59  443898.53      Ctg  191792.06
2        153441.51       101145.55  407934.54  Rangpur  191050.39
3        144372.41       118671.85  383199.62    Dhaka  182901.99
4        142107.34        91391.77  366168.42  Rangpur  166187.94
```

[19]:
```
new_df= df.drop('Area', axis=1)
```

[20]:
```
new_df.head()
```

[20]:
```
   Marketing Spend  Administration  Transport     Profit
0        114523.61       136897.80  471784.10  192261.83
1        162597.70       151377.59  443898.53  191792.06
2        153441.51       101145.55  407934.54  191050.39
3        144372.41       118671.85  383199.62  182901.99
4        142107.34        91391.77  366168.42  166187.94
```

[21]:
```
df = pd.concat([new_df,dummy_variables],axis=1)
```

[22]:
```
df.head()
```

[22]:
```
   Marketing Spend  Administration  Transport     Profit  Dhaka  Rangpur
0        114523.61       136897.80  471784.10  192261.83      1        0
1        162597.70       151377.59  443898.53  191792.06      0        0
2        153441.51       101145.55  407934.54  191050.39      0        1
3        144372.41       118671.85  383199.62  182901.99      1        0
4        142107.34        91391.77  366168.42  166187.94      0        1
```

4 Ordinal Encoder

[23]:
```
df=pd.read_csv('agora.csv')
df.head()
```

[23]:
```
   Marketing Spend  Administration  Transport     Area     Profit
0        114523.61       136897.80  471784.10    Dhaka  192261.83
1        162597.70       151377.59  443898.53      Ctg  191792.06
2        153441.51       101145.55  407934.54  Rangpur  191050.39
3        144372.41       118671.85  383199.62    Dhaka  182901.99
```

```
4        142107.34        91391.77  366168.42  Rangpur  166187.94
```

[24]:
```
from sklearn.preprocessing import OrdinalEncoder
df.Area.unique()
```

[24]:
```
array(['Dhaka', 'Ctg', 'Rangpur'], dtype=object)
```

[25]:
```
city_list = ['Dhaka', 'Ctg', 'Rangpur']
```

[26]:
```
ordinal = OrdinalEncoder(categories=[city_list])
```

[27]:
```
encoded_values = ordinal.fit_transform(df[['Area']]) # number of sample &␣
↪number of feature
```

[28]:
```
new_area = pd.DataFrame(encoded_values, columns= ['Area'])
```

[29]:
```
df.head()
```

[29]:

	Marketing Spend	Administration	Transport	Area	Profit
0	114523.61	136897.80	471784.10	Dhaka	192261.83
1	162597.70	151377.59	443898.53	Ctg	191792.06
2	153441.51	101145.55	407934.54	Rangpur	191050.39
3	144372.41	118671.85	383199.62	Dhaka	182901.99
4	142107.34	91391.77	366168.42	Rangpur	166187.94

[30]:
```
new_area
```

[30]:

	Area
0	0.0
1	1.0
2	2.0
3	0.0
4	2.0
5	0.0
6	1.0
7	2.0
8	0.0
9	1.0
10	2.0
11	1.0
12	2.0
13	1.0
14	2.0
15	0.0
16	1.0
17	0.0
18	2.0

```
19   0.0
20   1.0
21   0.0
22   2.0
23   2.0
24   0.0
25   1.0
26   2.0
27   0.0
28   2.0
29   0.0
30   2.0
31   0.0
32   1.0
33   2.0
34   1.0
35   0.0
36   2.0
37   1.0
38   0.0
39   1.0
40   1.0
41   2.0
42   1.0
43   0.0
44   1.0
45   0.0
46   2.0
47   1.0
48   0.0
49   1.0
```

[31]:
```
new_df=df.drop('Area',axis=1)
new_df.head()
```

[31]:

	Marketing Spend	Administration	Transport	Profit
0	114523.61	136897.80	471784.10	192261.83
1	162597.70	151377.59	443898.53	191792.06
2	153441.51	101145.55	407934.54	191050.39
3	144372.41	118671.85	383199.62	182901.99
4	142107.34	91391.77	366168.42	166187.94

[32]:
```
df=pd.concat([new_df,new_area],axis=1)
df.head()
```

[32]:

	Marketing Spend	Administration	Transport	Profit	Area
0	114523.61	136897.80	471784.10	192261.83	0.0

```
1        162597.70       151377.59  443898.53  191792.06   1.0
2        153441.51       101145.55  407934.54  191050.39   2.0
3        144372.41       118671.85  383199.62  182901.99   0.0
4        142107.34        91391.77  366168.42  166187.94   2.0
```

5 Hashing Encoder

[33]:
```
df  = pd.read_csv('agora.csv')
df.head()
```

[33]:
```
   Marketing Spend  Administration  Transport     Area     Profit
0        114523.61       136897.80  471784.10    Dhaka  192261.83
1        162597.70       151377.59  443898.53      Ctg  191792.06
2        153441.51       101145.55  407934.54  Rangpur  191050.39
3        144372.41       118671.85  383199.62    Dhaka  182901.99
4        142107.34        91391.77  366168.42  Rangpur  166187.94
```

[34]:
```
df.Area.unique()
```

[34]:
```
array(['Dhaka', 'Ctg', 'Rangpur'], dtype=object)
```

[35]:
```
# ! pip install category-encoders
```

[36]:
```
import category_encoders as ce
```

[37]:
```
encoders = ce.HashingEncoder(cols='Area',n_components=3)
```

[38]:
```
encoders.fit_transform(df)
```

[38]:
```
    col_0  col_1  col_2  Marketing Spend  Administration  Transport     Profit
0       0      1      0        114523.61       136897.80  471784.10  192261.83
1       0      0      1        162597.70       151377.59  443898.53  191792.06
2       1      0      0        153441.51       101145.55  407934.54  191050.39
3       0      1      0        144372.41       118671.85  383199.62  182901.99
4       1      0      0        142107.34        91391.77  366168.42  166187.94
5       0      1      0        131876.90        99814.71  362861.36  156991.12
6       0      0      1        134615.46       147198.87  127716.82  156122.51
7       1      0      0        130298.13       145530.06  323876.68  155752.60
8       0      1      0        120542.52       148718.95  311613.29  152211.77
9       0      0      1        123334.88       108679.17  304981.62  149759.96
10      1      0      0        101913.08       110594.11  229160.95  146121.95
11      0      0      1        100671.96        91790.61  249744.55  144259.40
12      1      0      0         93863.75       127320.38  249839.44  141585.52
13      0      0      1         91992.39       135495.07  252664.93  134307.35
14      1      0      0        119943.24       156547.42  256512.92  132602.65
```

15	0	1	0	165349.20	122616.84	261776.23	129917.04
16	0	0	1	78013.11	121597.55	264346.06	126992.93
17	0	1	0	94657.16	145077.58	282574.31	125370.37
18	1	0	0	91749.16	114175.79	294919.57	124266.90
19	0	1	0	86419.70	153514.11	NaN	122776.86
20	0	0	1	76253.86	113867.30	298664.47	118474.03
21	0	1	0	78389.47	153773.43	299737.29	111313.02
22	1	0	0	73994.56	122782.75	303319.26	110352.25
23	1	0	0	67532.53	105751.03	304768.73	108733.99
24	0	1	0	77044.01	99281.34	140574.81	108552.04
25	0	0	1	64664.71	139553.16	137962.62	107404.34
26	1	0	0	75328.87	144135.98	134050.07	105733.54
27	0	1	0	72107.60	127864.55	353183.81	105008.31
28	1	0	0	66051.52	182645.56	118148.20	103282.38
29	0	1	0	65605.48	153032.06	107138.38	101004.64
30	1	0	0	61994.48	115641.28	91131.24	99937.59
31	0	1	0	61136.38	152701.92	88218.23	97483.56
32	0	0	1	63408.86	129219.61	46085.25	97427.84
33	1	0	0	55493.95	103057.49	214634.81	96778.92
34	0	0	1	46426.07	157693.92	210797.67	96712.80
35	0	1	0	46014.02	85047.44	205517.64	96479.51
36	1	0	0	28663.76	127056.21	201126.82	90708.19
37	0	0	1	44069.95	51283.14	197029.42	89949.14
38	0	1	0	20229.59	65947.93	185265.10	81229.06
39	0	0	1	38558.51	82982.09	174999.30	81005.76
40	0	0	1	28754.33	118546.05	172795.67	78239.91
41	1	0	0	27892.92	84710.77	164470.71	77798.83
42	0	0	1	23640.93	96189.63	148001.11	71498.49
43	0	1	0	15505.73	127382.30	35534.17	69758.98
44	0	0	1	22177.74	154806.14	28334.72	65200.33
45	0	1	0	1000.23	124153.04	1903.93	64926.08
46	1	0	0	1315.46	115816.21	297114.46	49490.75
47	0	0	1	0.00	135426.92	0.00	42559.73
48	0	1	0	542.05	51743.15	0.00	35673.41
49	0	0	1	0.00	116983.80	45173.06	14681.40

[ ]: