Handled Categorical Missing Value Part 4

July 21, 2023

1 Handle Categorical Features

One Hot Encoding

[5]:
```
import pandas as pd
import numpy as np
```

[3]:
```
df= pd.read_csv("C:\\Users\\ssart\\Downloads\\train.csv",usecols = ['Sex'])
```

[5]:
```
df.head(5)
```

[5]:

	Sex
0	male
1	female
2	female
3	female
4	male

[8]:
```
pd.get_dummies(df,drop_first = True).head(5)
```

[8]:

	Sex_male
0	1
1	0
2	0
3	0
4	1

[11]:
```

```

[11]:
```
Index(['Sex'], dtype='object')
```

[14]:
```
df= pd.read_csv("C:\\Users\\ssart\\Downloads\\train.csv",usecols = ['Embarked'])
```

[16]:
```
df['Embarked'].unique()
```

[16]:
```
array(['S', 'C', 'Q', nan], dtype=object)
```

[18]:
```
df.dropna(inplace= True)
```

[20]:
```
pd.get_dummies(df,drop_first =True).head(5)
```

[20]:

	Embarked_Q	Embarked_S
0	0	1
1	0	0
2	0	1
3	0	1
4	0	1

Onehotencoding with many categories in a feature

[6]:
```
df=pd.read_csv("C:\\Users\\ssart\\Downloads\\Merceds.csv",usecols =␣
↪['X0',"X1","X2","X3","X4","X5","X6"])
```

[7]:
```
df.head(5)
```

[7]:

	X0	X1	X2	X3	X4	X5	X6
0	k	v	at	a	d	u	j
1	k	t	av	e	d	y	l
2	az	w	n	c	d	x	j
3	az	t	n	f	d	x	l
4	az	v	n	f	d	h	d

[10]:
```
for i in df.columns:
    print(len(df[i].unique()))
```

```
47
27
44
7
4
29
12
```

[16]:
```
df.X1.value_counts().sort_values(ascending = True).head(10)
```

[16]:
```
ab     3
d      3
q      3
g      6
p      9
k     17
n     19
j     22
f     23
y     23
Name: X1, dtype: int64
```

[27]:
```
lst_10=df.X1.value_counts().sort_values(ascending=False).head(10).index
lst_10=list(lst_10)
```

[28]:
```
lst_10
```

[28]:
```
['aa', 's', 'b', 'l', 'v', 'r', 'i', 'a', 'c', 'o']
```

[29]:
```
import numpy as np
for categories in lst_10:
    df[categories]=np.where(df['X1']==categories,1,0)
```

[30]:
```
lst_10.append('X1')
```

[31]:
```
df[lst_10]
```

[31]:

	aa	s	b	l	v	r	i	a	c	o	X1
0	0	0	0	0	1	0	0	0	0	0	v
1	0	0	0	0	0	0	0	0	0	0	t
2	0	0	0	0	0	0	0	0	0	0	w
3	0	0	0	0	0	0	0	0	0	0	t
4	0	0	0	0	1	0	0	0	0	0	v
…	..	..	..	..	..	..	..	..	..	..	..
4204	0	1	0	0	0	0	0	0	0	0	s
4205	0	0	0	0	0	0	0	0	0	1	o
4206	0	0	0	0	1	0	0	0	0	0	v
4207	0	0	0	0	0	1	0	0	0	0	r
4208	0	0	0	0	0	1	0	0	0	0	r

[4209 rows x 11 columns]

[ ]:
```

```