campusx-official / 100-days-of-machine-learning Public

Code Issues 3 Pull requests Actions Projects Security Insights

main

100-days-of-machine-learning / day27-one-hot-encoding / day27.ipynb

campusx-official Add files via upload History

1 contributor

1079 lines (1079 sloc) | 31.8 KB

In [92]:
```
import numpy as np
import pandas as pd
```

In [119…
```
df = pd.read_csv('cars.csv')
```

In [120…
```
df.head()
```

Out[120…

	brand	km_driven	fuel	owner	selling_price
0	Maruti	145500	Diesel	First Owner	450000
1	Skoda	120000	Diesel	Second Owner	370000
2	Honda	140000	Petrol	Third Owner	158000
3	Hyundai	127000	Diesel	First Owner	225000
4	Maruti	120000	Petrol	First Owner	130000

In [121…
```
df['owner'].value_counts()
```

Out[121…
```
First Owner             5289
Second Owner            2105
Third Owner              555
Fourth & Above Owner     174
Test Drive Car             5
Name: owner, dtype: int64
```

1. OneHotEncoding using Pandas

In [99]:
```
pd.get_dummies(df,columns=['fuel','owner'])
```

Out[99]:

	brand	km_driven	selling_price	fuel_CNG	fuel_Die
0	Maruti	145500	450000	0	
1	Skoda	120000	370000	0	
2	Honda	140000	158000	0	
3	Hyundai	127000	225000	0	
4	Maruti	120000	130000	0	

...	...	...	...	...
8123	Hyundai	110000	320000	0
8124	Hyundai	119000	135000	0
8125	Maruti	120000	382000	0
8126	Tata	25000	290000	0
8127	Tata	25000	290000	0

8128 rows × 12 columns

2. K-1 OneHotEncoding

In [100…

```
pd.get_dummies(df,columns=['fuel','owner'],drop_f
```

Out[100…

	brand	km_driven	selling_price	fuel_Diesel	fuel_LP
0	Maruti	145500	450000	1	
1	Skoda	120000	370000	1	
2	Honda	140000	158000	0	
3	Hyundai	127000	225000	1	
4	Maruti	120000	130000	0	
...	...	...	...	...	
8123	Hyundai	110000	320000	0	
8124	Hyundai	119000	135000	1	
8125	Maruti	120000	382000	1	
8126	Tata	25000	290000	1	
8127	Tata	25000	290000	1	

8128 rows × 10 columns

3. OneHotEncoding using Sklearn

In [122…

```
from sklearn.model_selection import train_test_sp
X_train,X_test,y_train,y_test = train_test_split(
```

In [111…

```
X_train.head()
```

Out[111…

	brand	km_driven	fuel	owner
5571	Hyundai	35000	Diesel	First Owner
2038	Jeep	60000	Diesel	First Owner
2957	Hyundai	25000	Petrol	First Owner
7618	Mahindra	130000	Diesel	Second Owner
6684	Hyundai	155000	Diesel	First Owner

In [123…

```
from sklearn.preprocessing import OneHotEncoder
```

In [137…

```
ohe = OneHotEncoder(drop='first',sparse=False,dty
```

In [138…

```
X_train_new = ohe.fit_transform(X_train[['fuel','
```

In [139…

```
X_test_new = ohe.transform(X_test[['fuel','owner'
```

In [140…

```
X_train_new.shape
```

Out[140…

```
(6502, 7)
```

In [141…

```
np.hstack((X_train[['brand','km_driven']].values,
```

Out[141…

```
array([['Hyundai', 35000, 1, ..., 0, 0, 0],
       ['Jeep', 60000, 1, ..., 0, 0, 0],
       ['Hyundai', 25000, 0, ..., 0, 0, 0],
       ...,
       ['Tata', 15000, 0, ..., 0, 0, 0],
       ['Maruti', 32500, 1, ..., 1, 0, 0],
       ['Isuzu', 121000, 1, ..., 0, 0, 0]], dtype=
object)
```

In [ ]:

In [ ]:

4. OneHotEncoding with Top Categories

In [143…

```
counts = df['brand'].value_counts()
```

In [144…

```
df['brand'].nunique()
```

```
threshold = 100
```

In [146…

```
repl = counts[counts <= threshold].index
```

In [150…

```
pd.get_dummies(df['brand'].replace(repl, 'uncommo
```