How can I use Pandas' `get_dummies` function to create dummy variables for categorical features in a dataset?

Authored by

stats writer

May 1, 2024

Pandas' `get_dummies` function is a powerful tool for creating dummy variables in a dataset. This function allows users to easily convert categorical features into numerical values, which can then be used for further analysis. By using this function, users can efficiently handle categorical data in their dataset and avoid any potential errors that may arise from using non-numerical values. Additionally, the `get_dummies` function offers various parameters and options for customizing the creation of dummy variables, providing further flexibility and control for users. Overall, the `get_dummies` function is an essential tool for accurately and effectively processing categorical features in a dataset.

Use Pandas Get Dummies - pd.get_dummies

Often in statistics, the datasets we're working with include .

These are variables that take on names or labels. Examples include:

Marital status ("married", "single", "divorced")Smoking status ("smoker", "non-smoker")Eye color ("blue", "green", "hazel")Level of education (e.g. "high school", "Bachelor's degree", "Master's degree")

When fitting machine learning algorithms (like , , , etc.), we often convert categorical variables to dummy variables, which are numeric variables that are used to represent categorical data.

For example, suppose we have a dataset that contains

the categorical variable Gender. To use this variable as a predictor in a regression model, we would first need to convert it to a dummy variable.

To create this dummy variable, we can choose one of the values ("Male") to represent 0 and the other value ("Female") to represent 1:

Income	Age	Gender
$45,000	23	Male
$48,000	25	Female
$54,000	24	Male
$57,000	29	Female
$65,000	38	Female
$69,000	36	Female
$78,000	40	Male
$83,000	59	Female
$98,000	56	Male
$104,000	64	Male
$107,000	53	Male

→

Income	Age	Gender_Dummy
$45,000	23	0
$48,000	25	1
$54,000	24	0
$57,000	29	1
$65,000	38	1
$69,000	36	1
$78,000	40	0
$83,000	59	1
$98,000	56	0
$104,000	64	0
$107,000	53	0

How to Create Dummy Variables in Pandas

To create dummy variables for a variable in a pandas DataFrame, we can use the function, which uses the following basic syntax:

```
pandas.get_dummies(data, prefix=None, columns=None, drop_first=False)
```

where:

data: The name of the pandas DataFrameprefix: A string to append to the front of the new dummy variable columncolumns: The name of the column(s) to convert to a dummy variabledrop_first: Whether or not to drop the first dummy variable column

The following examples show how to use this function in practice.

Example 1: Create a Single Dummy Variable

Suppose we have the following pandas DataFrame:

```
import pandas as pd

#create DataFrame
df = pd.DataFrame({'income': ,
'age': ,
'gender': })

#view DataFrame
df

income age gender
0 45 23 M
```

1 48 25 F

2 54 24 M

3 57 29 F

4 65 38 F

5 69 36 F

6 78 40 M

We can use the pd.get_dummies() function to turn gender into a dummy variable:

#convert gender to dummy variable
pd.get_dummies(df, columns=, drop_first=True)

income age gender_M
0 45 23 1
1 48 25 0
2 54 24 1
3 57 29 0
4 65 38 0
5 69 36 0
6 78 40 1

The gender column is now a dummy variable where:

A value of 0 represents "Female"A value of 1 represents

"Male"

Example 2: Create Multiple Dummy Variables

Suppose we have the following pandas DataFrame:

import pandas as pd

#create DataFrame
df = pd.DataFrame({'income': ,
'age': ,
'gender': ,
'college': })

#view DataFrame
df

income age gender college
0 45 23 M Y
1 48 25 F N
2 54 24 M N
3 57 29 F N
4 65 38 F Y
5 69 36 F Y
6 78 40 M Y

We can use the pd.get_dummies() function to convert gender and college both into dummy variables:

#convert gender to dummy variable
pd.get_dummies(df, columns=, drop_first=True)

income age gender_M college_Y
0 45 23 1 1
1 48 25 0 0
2 54 24 1 0
3 57 29 0 0
4 65 38 0 1
5 69 36 0 1
6 78 40 1 1

The gender column is now a dummy variable where:

A value of 0 represents "Female"A value of 1 represents "Male"

And the college column is now a dummy variable where:

A value of 0 represents "No" collegeA value of 1 represents "Yes" college