ASSIGNMENT/ TASK 10

GO_STP_12574
SHWETA JHA

Q)Discuss the concept of One-Hot-Encoding, Multicollinearity and the Dummy Variable Trap. What is Nominal and Ordinal Variables ?

Salary Dataset of 52 professors having categorical columns. Apply dummy variables concept and one-hot-encoding on categorical columns.

Dataset Link- Click here
ANS.

One Hot Encoding –

It refers to splitting the column which contains numerical categorical data to many columns depending on the number of categories present in that column. Each column contains “0” or “1” corresponding to which column it has been placed.

Multicorllinearity

Multicollinearity occurs when two or more independent variables (a.k.a. features) in the dataset are correlated with each other.

There are several methods using which we can measure the degree and direction of correlation for bivariate cases (more information on measures of correlation), while multicollinearity is generally measured using Variance Inflation Factor (more information on measures of multicollinearity).

In a nutshell, multicollinearity is said to exist in a dataset when the independent variables are (nearly) linearly related to each other.

Dummy Variable Trap

The dummy variable trap manifests itself directly from one-hot-encoding applied on categorical variables. As discussed earlier, size of one-hot vectors is equal to the number of unique values that a categorical column takes up and each such vector contains exactly one ‘1’ in it. This ingests multicollinearity into our dataset.

Nominal Variables

A nominal variable is a type of variable that is used to name, label or categorize particular attributes that are being measured. It takes qualitative values representing different categories, and there is no intrinsic ordering of these categories.

You can code nominal variables with numbers, but the order is arbitrary and arithmetic operations cannot be performed on the numbers. This is the case when a person's phone number, National Identification Number postal code, etc. are being collected.

A nominal variable is one of the 2 types of categorical variables and is the simplest among all the measurement variables. Some examples of nominal variables include gender, Name, phone, etc.

Ordinal Variables

Ordinal variable is a type of measurement variable that takes values with an order or rank. It is the 2nd level of measurement and is an extension of the nominal variable.

They are built upon nominal scales by assigning numbers to objects to reflect a rank or ordering on an attribute. Also, there is no standard ordering in the ordinal variable scale.

In another sense, we could say the difference in the rank of an ordinal variable is not equal. It is mostly classified as one of the 2 types of categorical variables, while in some cases it is said to be a midpoint between categorical and numerical variables.

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import sklearn
from sklearn.linear_model import LinearRegression
```

```
dataset = pd.read_table('/content/salary.txt',delim_whitespace=True)
dataset.head()
```

	sx	rk	yr	dg	yd	sl
0	male	full	25	doctorate	35	36350
1	male	full	13	doctorate	22	35350
2	male	full	10	doctorate	23	28200
3	female	full	7	doctorate	27	26775
4	male	full	19	masters	30	33696

```
dataset.add_prefix
```

```
dataset.dtypes
```

```
sx        rk yr        dg yd    sl    object
dtype: object
```

```
dataset.describe()
```

	sx rk yr dg yd sl
count	52
unique	52
top	male associate 3 masters 17 23725
freq	1

```
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer

# creating one hot encoder object with categorical feature 0
# indicating the first column
columnTransformer = ColumnTransformer([('encoder',OneHotEncoder(),[0,1,3])],remainder='pas
data=np.array(columnTransformer.fit_transform(dataset),dtype=str)
print(data[:6])
```

```
[['0.0' '1.0' '0.0' '0.0' '1.0' '1.0' '0.0' '25.0' '35.0' '36350.0']
 ['0.0' '1.0' '0.0' '0.0' '1.0' '1.0' '0.0' '13.0' '22.0' '35350.0']
 ['0.0' '1.0' '0.0' '0.0' '1.0' '1.0' '0.0' '10.0' '23.0' '28200.0']
 ['1.0' '0.0' '0.0' '0.0' '1.0' '1.0' '0.0' '7.0' '27.0' '26775.0']
 ['0.0' '1.0' '0.0' '0.0' '1.0' '0.0' '1.0' '19.0' '30.0' '33696.0']
 ['0.0' '1.0' '0.0' '0.0' '1.0' '1.0' '0.0' '16.0' '21.0' '28516.0']]
```

✓ 0s completed at 6:22 AM