+ Code + Text

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

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

1. OneHotEncoding

OneHotEncoding is a prpcess to convert string data into numeric data, but data should be categorical value.

The input to the transformer should be an array like of integers or string, denoting the value takenby categorical feature. This creates a binary columns and returns the spars matrix or dense array.

2. Multicollinearity

It is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with accuracy.

It neither reduce the predictive power nor reliablility of the model as whole, at least within the sample data set; it only affects calculations regarding individual predictors

3. Dummy Variable

It is a scenario where there are attributes which are highly correlated and one varable predicts the value of others. When we use one hot encoding for handling the categorical data, then one attribute can be predicted with the help of other dummy variables.

For example, Sex having two values male and female. Either they can be 1/0 or 0/1. Including both dummy variable can cause redundancy because if a person is not male in such case the person is female, hence, we don't need to use both the variables in model.

4. Nominal Variable

It describes a variable with categories that do not have order or sequence.

Example: blood type, sex

5. Ordinal Variable

It describes the variable with order or sequence.

Example: Sentiment("poor", "bad", "neutral", "good", "very good")

```
import pandas as pd
df = pd.read_csv("https://data.princeton.edu/wws509/datasets/salary.dat", delim_whitespace =
```

```
df.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

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 52 entries, 0 to 51
Data columns (total 6 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   sx      52 non-null     object
 1   rk      52 non-null     object
 2   yr      52 non-null     int64 
 3   dg      52 non-null     object
 4   yd      52 non-null     int64 
 5   sl      52 non-null     int64 
dtypes: int64(3), object(3)
memory usage: 2.6+ KB
```

```
df.columns
```

```
Index(['sx', 'rk', 'yr', 'dg', 'yd', 'sl'], dtype='object')
```

```
from sklearn.preprocessing import LabelEncoder
label = LabelEncoder()
```

```
df.sx = label.fit_transform(df.sx)
df.head()
```

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

```
df.dg = le.fit_transform(df.dg)
df.head()
```

```
df.head()
```

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