Preprocessing-Number_of_kids_prediction

July 7, 2021

0.1 Importing necessary libraries

[1]:
```
import numpy as np
import pandas as pd
```

0.2 Reading the data

[2]:
```
data = pd.read_csv("data.csv")
```

0.3 Describing the data

[3]:
```
data.head()
```

[3]:

	Number of Kids	Working Experience(years)	Age	Salary	Blood Types
0	3	15.0	45	250000	A
1	1	5.0	30	200000	B
2	2	10.0	38	150000	AB
3	1	NaN	36	180000	O

[4]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4 entries, 0 to 3
Data columns (total 5 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Number of Kids             4 non-null      int64
 1   Working Experience(years)  3 non-null      float64
 2   Age                        4 non-null      int64
 3   Salary                     4 non-null      int64
 4   Blood Types                4 non-null      object
dtypes: float64(1), int64(3), object(1)
memory usage: 288.0+ bytes
```

0.4 Filling missing data with median

[5]:
```
data["Working Experience(years)"].fillna(data["Working Experience(years)"].
↪median(),inplace=True)
```

[6]:
```
data.head()
```

[6]:

	Number of Kids	Working Experience(years)	Age	Salary	Blood Types
0	3	15.0	45	250000	A
1	1	5.0	30	200000	B
2	2	10.0	38	150000	AB
3	1	10.0	36	180000	O

[7]:
```
data["Working Experience(years)"]=data["Working Experience(years)"].astype(int)
```

[8]:
```
data.head()
```

[8]:

	Number of Kids	Working Experience(years)	Age	Salary	Blood Types
0	3	15	45	250000	A
1	1	5	30	200000	B
2	2	10	38	150000	AB
3	1	10	36	180000	O

0.5 Finding Correlation between Number of Kids and Age, Number of Kids and Working Experience

[9]:
```
data["Number of Kids"].corr(data["Age"])
```

[9]:
```
0.9147673836616229
```

[10]:
```
data["Working Experience(years)"].corr(data["Number of Kids"])
```

[10]:
```
0.8528028654224419
```

Age is more related to Number of Kids - 0.91 correlation

0.6 One Hot Vectors

[11]:
```
blood_types_encoded,categories=data["Blood Types"].factorize()
blood_types_encoded
```

[11]:
```
array([0, 1, 2, 3], dtype=int64)
```

[12]:
```
from sklearn.preprocessing import OneHotEncoder
encoder = OneHotEncoder(sparse=False)
blood_type_cat_1hot = encoder.fit_transform(blood_types_encoded.reshape(-1,1))
blood_type_cat_1hot
```

[12]:
```
array([[1., 0., 0., 0.],
       [0., 1., 0., 0.],
       [0., 0., 1., 0.],
       [0., 0., 0., 1.]])
```

[13]:
```
data.head()
```

[13]:

	Number of Kids	Working Experience(years)	Age	Salary	Blood Types
0	3	15	45	250000	A
1	1	5	30	200000	B
2	2	10	38	150000	AB
3	1	10	36	180000	O

[14]:
```
one_hot=pd.get_dummies(data,columns=["Blood␣
↪Types"],drop_first=False,prefix='',prefix_sep='')
one_hot
```

[14]:

	Number of Kids	Working Experience(years)	Age	Salary	A	AB	B	O
0	3	15	45	250000	1	0	0	0
1	1	5	30	200000	0	0	1	0
2	2	10	38	150000	0	1	0	0
3	1	10	36	180000	0	0	0	1

[15]:
```
data=data.drop("Blood Types",axis=1)
data
```

[15]:

	Number of Kids	Working Experience(years)	Age	Salary
0	3	15	45	250000
1	1	5	30	200000
2	2	10	38	150000
3	1	10	36	180000

[16]:
```
data=data.join(one_hot["A"])
data=data.join(one_hot["B"])
data=data.join(one_hot["AB"])
data=data.join(one_hot["O"])
```

[17]:
```
data
```

[17]:

	Number of Kids	Working Experience(years)	Age	Salary	A	B	AB	O
0	3	15	45	250000	1	0	0	0
1	1	5	30	200000	0	1	0	0
2	2	10	38	150000	0	0	1	0
3	1	10	36	180000	0	0	0	1

0.7 Scaling the data

[18]:
```
from sklearn.preprocessing import StandardScaler
```

[24]:
```
scaler = StandardScaler()
data_scaled = pd.DataFrame(scaler.fit_transform(data),columns=["Number of␣
↪Kids","Working Experience(years)","Age",
                                                               "Salary","Blood␣
↪Type A","Blood Type B","Blood Type AB",
                                                               "Blood Type O"])
```

[25]:
```
data_scaled
```

[25]:
```
   Number of Kids  Working Experience(years)       Age    Salary  \
0        1.507557                   1.414214  1.446956  1.510966
1       -0.904534                  -1.414214 -1.353604  0.137361
2        0.301511                   0.000000  0.140028 -1.236245
3       -0.904534                   0.000000 -0.233380 -0.412082

   Blood Type A  Blood Type B  Blood Type AB  Blood Type O
0      1.732051     -0.577350      -0.577350     -0.577350
1     -0.577350      1.732051      -0.577350     -0.577350
2     -0.577350     -0.577350       1.732051     -0.577350
3     -0.577350     -0.577350      -0.577350      1.732051
```

[26]:
```
data_scaled.round(decimals=2)
```

[26]:
```
   Number of Kids  Working Experience(years)   Age  Salary  Blood Type A  \
0            1.51                       1.41  1.45    1.51          1.73
1           -0.90                      -1.41 -1.35    0.14         -0.58
2            0.30                       0.00  0.14   -1.24         -0.58
3           -0.90                       0.00 -0.23   -0.41         -0.58

   Blood Type B  Blood Type AB  Blood Type O
0         -0.58          -0.58         -0.58
1          1.73          -0.58         -0.58
2         -0.58           1.73         -0.58
3         -0.58          -0.58          1.73
```

[ ]:
```

```