program 21

October 8, 2024

[2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[3]:
```
salary_df=pd.read_csv("C:/Users/DSU-CSE513-16/Downloads/Salary_Data.csv")
salary_df
```

[3]:

	YearsExperience	Salary
0	1.1	39343
1	1.3	46205
2	1.5	37731
3	2.0	43525
4	2.2	39891
5	2.9	56642
6	3.0	60150
7	3.2	54445
8	3.2	64445
9	3.7	57189
10	3.9	63218
11	4.0	55794
12	4.0	56957
13	4.1	57081
14	4.5	61111
15	4.9	67938
16	5.1	66029
17	5.3	83088
18	5.9	81363
19	6.0	93940
20	6.8	91738
21	7.1	98273
22	7.9	101302
23	8.2	113812
24	8.7	109431
25	9.0	105582
26	9.5	116969
27	9.6	112635
28	10.3	122391

```
29              10.5  121872
```

[4]:
```
salary_df.shape
```

[4]:
```
(30, 2)
```

[5]:
```
salary_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30 entries, 0 to 29
Data columns (total 2 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   YearsExperience  30 non-null     float64
 1   Salary           30 non-null     int64  
dtypes: float64(1), int64(1)
memory usage: 612.0 bytes
```

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

[6]:

	YearsExperience	Salary
0	1.1	39343
1	1.3	46205
2	1.5	37731
3	2.0	43525
4	2.2	39891

[7]:
```
salary_df.describe()
```

[7]:

	YearsExperience	Salary
count	30.000000	30.000000
mean	5.313333	76003.000000
std	2.837888	27414.429785
min	1.100000	37731.000000
25%	3.200000	56720.750000
50%	4.700000	65237.000000
75%	7.700000	100544.750000
max	10.500000	122391.000000

[8]:
```
x=salary_df.loc[:,'YearsExperience'].values
y=salary_df.loc[:,'Salary'].values
```

[9]:
```
from sklearn.model_selection import train_test_split
```

[10]:
```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3,␣
↪random_state=0)
```

[11]:
```
x_train.shape,x_test.shape,y_train.shape,y_test.shape
```

[11]:
```
((21,), (9,), (21,), (9,))
```

[12]:
```
plt.scatter(data=salary_df,x='YearsExperience',y="Salary")
plt.title("Salary based on thr years of experience")
plt.xlabel("Years of Experience")
plt.ylabel("Salary")
plt.show()
```

[13]:
```
type(x_train)
```

[13]:
```
numpy.ndarray
```

[14]:
```
from sklearn.linear_model import LinearRegression
reg_model=LinearRegression()
reg_model.fit(x_train.reshape(-1,1),y_train.reshape(-1,1))
```

[14]:
```
LinearRegression()
```

[15]:
```
reg_model.coef_
reg_model.intercept_
```

[15]:
```
array([26777.3913412])
```

[16]:
```
reg_model.coef_
```

[16]:
```
array([[9360.26128619]])
```

[17]:
```
y_predicted=reg_model.predict(x_test.reshape(-1,1))
y_predicted
```

[17]:
```
array([[ 40817.78327049],
       [123188.08258899],
       [ 65154.46261459],
       [ 63282.41035735],
       [115699.87356004],
       [108211.66453108],
       [116635.89968866],
       [ 64218.43648597],
       [ 76386.77615802]])
```

[18]:
```
from sklearn.linear_model import LinearRegression
```

[19]:
```
reg_model=LinearRegression()
```

[20]:
```
reg_model.fit(x_train.reshape(-1,1),y_train.reshape(-1,1))
```

[20]:
```
LinearRegression()
```

[21]:
```
y_test
```

[21]:
```
array([ 37731, 122391,  57081,  63218, 116969, 109431, 112635,  55794,
        83088], dtype=int64)
```

[22]:
```
from sklearn.metrics import mean_squared_error, r2_score
r_square = r2_score(y_test, y_predicted)
r_square
```

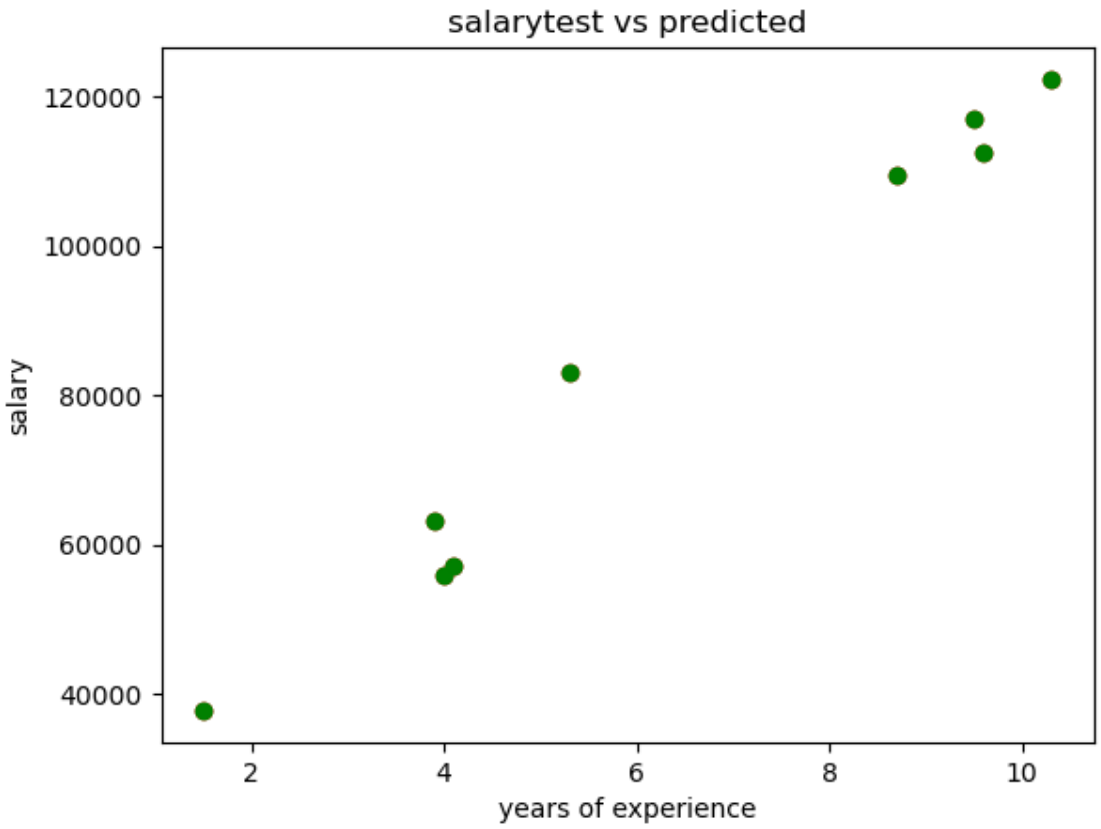
[22]:
```
0.9740993407213511
```

[23]:
```
rmse = mean_squared_error(y_test, y_predicted)
rmse
```

[23]:
```
23370078.800832972
```

[27]:
```
plt.scatter(x=x_test,y=y_test,color="red")
plt.scatter(x=x_test,y=y_test,color="green")
plt.title("salarytest vs predicted")
plt.xlabel("years of experience")
plt.ylabel("salary")
```

```
plt.show()
```



[25]:
```
from sklearn.metrics import mean_squared_error
mse=mean_squared_error(y_test,y_predicted)
rmse=np.sqrt(mse)
```

[26]:
```
rmse
```

[26]: 4834.260936361728

[ ]: