MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

Skill Based Mini Project Report

on

Salary Prediction Model with Machine Learning

Submitted By:

Yashvardhan Malve

0901CS201137

Faculty Mentor:

Dr. Ranjeet Kumar Singh

ASSISTANT PROFESSOR

Submitted to:

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

MADHAV INSTITUTE OF TECHOLOGY & SCIENCE
GWALIOR - 474005 (MP) est. 1957

JAN-JUN (2022)

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

CERTIFICATE

This is certified that **Yashvardhan Malve**(0901CS201137) has submitted the project report titled **Salary Prediction Model with Machine Learning** under the mentorship of **Dr. Ranjeet Kumar Singh, Assistant Professor** in partial fulfilment of the requirement for the award of degree of Bachelor of Technology in Computer Science and Engineering from Madhav Institute of Technology and Science, Gwalior.

DR RANJEET KUMAR
Faculty Mentor
Assistant Professor
Computer Science and Engineering

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

DECLARATION

I hereby declare that the work being presented in this project report, for the partial fulfilment of requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering at Madhav Institute of Technology & Science, Gwalior is an authenticated and original record of my work under the mentorship of **Dr. Ranjeet Kumar Singh, Assistant Professor, Computer Science and engineering**. I declare that I have not submitted the matter embodied in this report for the award of any degree or diploma anywhere else.

Yashvardhan Malve
0901CS201137
2nd Year, 4th semester
Computer Science and Engineering

MADHAV INSTITUTE OF TECHNOLOGY & SCIENCE, GWALIOR

(A Govt. Aided UGC Autonomous & NAAC Accredited Institute Affiliated to RGPV, Bhopal)

ACKNOWLEDGEMENT

The full semester project has proved to be pivotal to my career. I am thankful to my institute, **Madhav Institute of Technology and Science** to allow me to continue my disciplinary/interdisciplinary project as a curriculum requirement, under the provisions of the Flexible Curriculum Scheme (based on the AICTE Model Curriculum 2018), approved by the Academic Council of the institute. I extend my gratitude to the Director of the institute, **Dr. R. K. Pandit** and Dean Academics, **Dr. Manjaree Pandit** for this.

I would sincerely like to thank my department, **Department of Computer Science and Engineering, for allowing** me to explore this project. I humbly thank **Dr. Manish Dixit**, Professor and Head, Department of Computer Science and Engineering, for his continued support during the course of this engagement, which eased the process and formalities involved.

I am sincerely thankful to my faculty mentors. I am grateful to the guidance of **Dr Ranjeet Kumar, Assistant Professor** for his continued support and guidance throughout the project. I am also very thankful to the faculty and staff of the department.

Yashvardhan Malve
0901CS201137
2nd Year, 4th semester
Computer Science and Engineering

TABLE OF CONTENTS

ABSTRACT

The goal of this paper is to predict salary of a person after a certain year. The graphical representation of predicting salary is a process that aims for developing computerized system to maintain all the daily work of salary growth graph in any field and can predict salary after a certain time period. This application can take the database for the salary system from the organisation and makes a graph through this information from the database. It will check the salary fields then import a graph which helps to observe the graphical representation. And then it can predict a certain time period salary through the prediction algorithm. It can also be applied in some other effective prediction also.

INTRODUCTION

Machine learning (ML) offers an alternative approach for salary prediction by automating the process and potentially reducing bias. ML algorithms can learn from a dataset of employee information and make predictions based on pattern in the data. In this study, we aim to their job title , industry , education level

Salary Prediction is an important task for both employees and employers. For employees, knowing their expected salary can help with salary negotiations and career advancement planning. For employers, salary prediction can aid in budgeting and talent management. Traditional methods for salary prediction often rely on manual calculations and expert judgement, which can be time-consuming and prone to bias.

Hardware Requirements: System with Internet
Software Requirements: Goolge Colab

Problem Statements

To build a machine learning model and predict the salary of the employees based on year of experience

Dataset description

- This dataset is randomly created to show you how we can use machine learning technique and build a Linear Regression model to predict the salary of an employee based on years of experience.
- This dataset consists of two columns
- Salary- Represent the salary of a person.
- Years- Years of experience
- Note: - It's a small dataset which is being used for only demo purpose.

Salary Prediction

In [1]:
```
#import warnings
import warnings
warnings.filterwarnings('ignore')
```

In [27]:
```
#import libraries

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
import statsmodels.api as sm
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
```

In [3]:
```
#import data
data = pd.read_csv('Salary_Data.csv')
data.head()
```

Out[3]:

	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

Let us inspect the data:

Let us inspect the data:

In [4]:
```
data.shape
```

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

We hardly have 30 data points.

In [5]:
```
data.describe()
```

Out[5]:

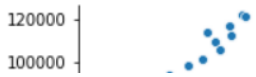
	YearsExperience	Salary
count	30.000000	30.000000
mean	5.313333	76003.000000
std	2.837888	27414.429785
min	1.100000	37731.000000
25%	3.200000	56720.750000
50%	4.700000	65237.000000
75%	7.700000	100544.750000
max	10.500000	122391.000000

Min years of experience is 1.1 with a salary of 27414.4 and max years of experience is 10.5 with a salary of 122391.

Let us visualise the data:

In [6]:
```
sns.pairplot(y_vars = 'Salary', x_vars = 'YearsExperience' ,data = data)
```

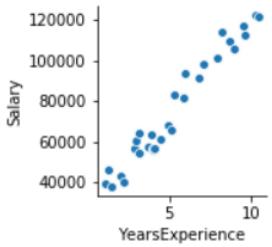
Out[6]:
```
<seaborn.axisgrid.PairGrid at 0xc86c5c0>
```



Let us visualise the data:

In [6]:
```
sns.pairplot(y_vars = 'Salary', x_vars = 'YearsExperience' ,data = data)
```

Out[6]:
```
<seaborn.axisgrid.PairGrid at 0xc86c5c0>
```



Salary looks linear related to years of experience.

In [7]:
```
# checking the correlation of the data
data.corr()
```

Out[7]:

	YearsExperience	Salary
YearsExperience	1.000000	0.978242
Salary	0.978242	1.000000

0.97 is highly correlated.

Data preparation:

In [8]:
```
X = data['YearsExperience']
y = data['Salary']
```

Model Building:

In [13]:
```
X_train_sm = sm.add_constant(X_train)
model = sm.OLS(y_train, X_train_sm).fit()
```

In [14]:
```
print(model.summary())
```

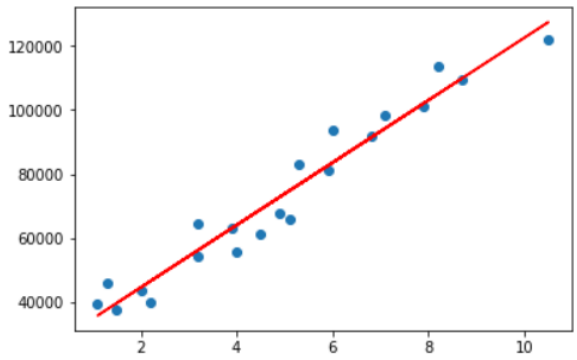
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Salary   R-squared:                       0.949
Model:                            OLS   Adj. R-squared:                  0.946
Method:                 Least Squares   F-statistic:                     352.9
Date:                Tue, 22 Oct 2019   Prob (F-statistic):           9.91e-14
Time:                        19:38:18   Log-Likelihood:                -211.80
No. Observations:                  21   AIC:                             427.6
Df Residuals:                      19   BIC:                             429.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.52e+04   2875.387      8.765      0.000    1.92e+04    3.12e+04
YearsExperience  9731.2038    517.993     18.786      0.000    8647.033    1.08e+04
==============================================================================
Omnibus:                        1.843   Durbin-Watson:                   1.749
Prob(Omnibus):                  0.398   Jarque-Bera (JB):                1.106
Skew:                           0.219   Prob(JB):                        0.575
Kurtosis:                       1.964   Cond. No.                         12.3
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

R2, p value looks fine.

So our linear regression equation is :

Salary = 25200 + YearsExperience x9731.2038

In [32]:
```
# Let us show the line fitting:
plt.scatter(X_train,y_train)
plt.plot(X_train, 25200 + X_train * 9731.2038,'r')
plt.show()
```



Residual analysis:

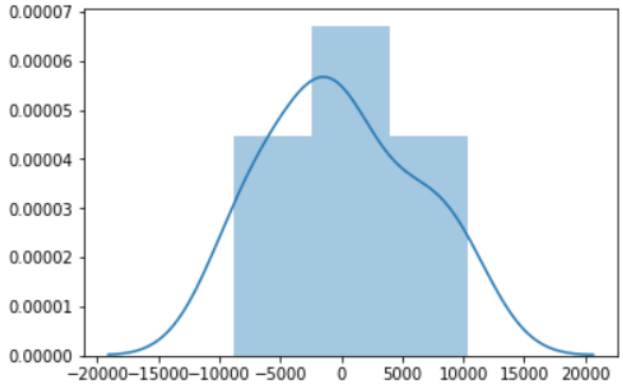
In [15]:
```
y_train_pred = model.predict(X_train_sm)
```

In [16]:
```
y_train_pred.head()
```

Out[16]:
```
17     76778.268129
22    102079.398108
11     64127.703139
4      46611.536230
29    127380.528088
dtype: float64
```

In [17]:
```
residual = (y_train - y_train_pred)
```

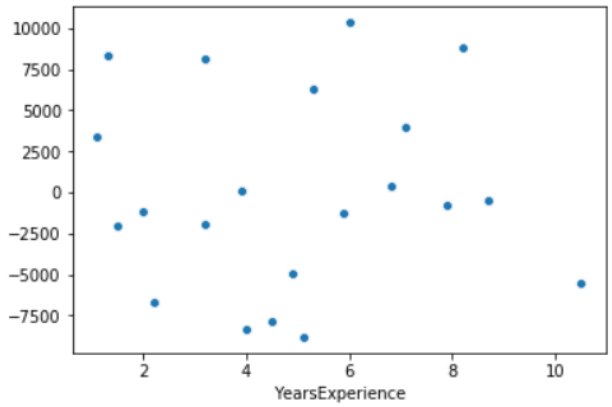
Out[22]: <matplotlib.axes._subplots.AxesSubplot at 0xcd57cc0>



Residual looks normally distributed.

In [24]:
```
sns.scatterplot(X_train,residual)
```

Out[24]: <matplotlib.axes._subplots.AxesSubplot at 0xcdd1278>



Predictions on the test data:

In [25]:
```
X_test_sm = sm.add_constant(X_test)
```

In [26]:
```
y_pred = model.predict(X_test_sm)
```

Root mean squared error:

In [29]:
```
RMSE = np.sqrt(mean_squared_error(y_test,y_pred))
RMSE
```

Out[29]:
```
5505.479002176114
```

RMSE won't suggest anything as we do not have any model to compare.

R2 score:

In [30]:
```
r2_score(y_test,y_pred)
```

Out[30]:
```
0.9627668685473272
```

The model explains 96.27% of variance.

In [31]:
```
# Let us show the line fitting:
plt.scatter(X_test,y_test)
plt.plot(X_test, 25200 + X_test * 9731.2038,'r')
plt.show()
```

RMSE won't suggest anything as we do not have any model to compare.

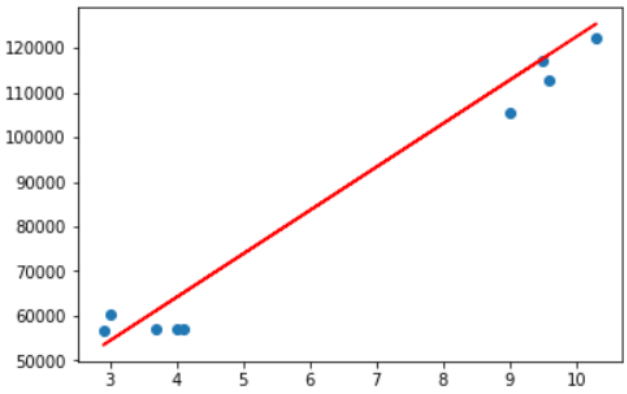
R2 score:

In [30]:
```
r2_score(y_test,y_pred)
```

Out[30]:
```
0.9627668685473272
```

The model explains 96.27% of variance.

In [31]:
```
# Let us show the line fitting:
plt.scatter(X_test,y_test)
plt.plot(X_test, 25200 + X_test * 9731.2038,'r')
plt.show()
```



In [ ]:
```

```

RESULT

There are many different ways to build a salary prediction model, but most of them involve using machine learning techniques to analyze data about salaries and the factors that might affect them. The goal of the model is to make predictions about the salary a person is likely to earn based on the input data provided. It's important to note that the accuracy of a salary prediction model will depend on the quality of the training data and the complexity of the model. More complex models might be able to make more accurate predictions, but they may also be more difficult to understand and interpret.

CONCLUSION

salary prediction model is a tool that uses machine learning techniques to analyze data about salaries and the factors that might affect them in order to make predictions about the salary a person is likely to earn. The accuracy of the model will depend on the quality of the training data and the complexity of the model. Such a model can be useful for employers in determining salary ranges for job openings and for job seekers in understanding what salary they might expect for a given position. However, it is important to note that salary prediction models are not perfect and should not be the only factor considered when making decisions about hiring or negotiating salary.

REFRENCES

[1] Manjula K. A., Karthikeyan P, "Gold Price Prediction using Ensemble based Machine Learning Techniques", Third International Conference on Trends in Electronics and Informatics, 2019

[2] R. Hafezi*, A. N. Akhavan, "Forecasting Gold Price Changes: Application of an Equipped Artificial Neural Network", AUT Journal of Modeling and Simulation, 2018.

[3] Shian-Chang Huang and Cheng-Feng Wu, Energy Commodity Price Forecasting with Deep Multiple Kernel Learning, MDPI Journal, 2018.