We have a dataset consisting of experience in years for the candidates, the score of their admission test out of 10, their interview scores out of 10, based on which we have different records for the salary column in $. So our problem statement here is to predict the salary based on the other features in the dataset.

We have cleaned the data by:

1. Checking outliers
2. Filling the missing values of the dataset
3. Encoding of the categorical string values to numeric values (with the help of ordinal encoding)

Next, we have partitioned the dataset into x and y variables, where x is our set of independent variables and y is our dependent variable or target variable.

We have splitted x and y into train and test datasets, train dataset to build the model and test dataset to evaluate the model performance.

Because the target variable, y, is continuous so this problem is a regression problem and the model we have applied is linear regression.

The model summary is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| OLS Regression Results | | | |
| **Dep. Variable:** | salary($) | **R-squared (uncentered):** | 0.958 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.915 |
| **Method:** | Least Squares | **F-statistic:** | 22.55 |
| **Date:** | Sat, 30 Sep 2023 | **Prob (F-statistic):** | 0.0147 |
| **Time:** | 06:53:31 | **Log-Likelihood:** | -65.351 |
| **No. Observations:** | 6 | **AIC:** | 136.7 |
| **Df Residuals:** | 3 | **BIC:** | 136.1 |
| **Df Model:** | 3 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **test\_score(outof10)** | 3772.7688 | 4282.537 | 0.881 | 0.443 | -9856.174 | 1.74e+04 |
| **interview\_score(outof10)** | 3047.3775 | 3787.646 | 0.805 | 0.480 | -9006.601 | 1.51e+04 |
| **encoded\_experience** | 783.0030 | 1750.050 | 0.447 | 0.685 | -4786.436 | 6352.442 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | nan | **Durbin-Watson:** | 2.261 |
| **Prob(Omnibus):** | nan | **Jarque-Bera (JB):** | 0.494 |
| **Skew:** | 0.240 | **Prob(JB):** | 0.781 |
| **Kurtosis:** | 1.678 | **Cond. No.** | 10.4 |

Notes:  
[1] R² is computed without centering (uncentered) since the model does not contain a constant.  
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

After applying the Linear Regression Model, we have checked the regression metrics for evaluation of the model performance which looked satisfactory and then we predicted for the future y values (by passing the input data)