**Scikit-Stack and Deep Learning - Predict diamond prices**

The goal is to use the given 9 independent variables (6 numerical, 3 categorical) to predict the diamond price (target variable). The dataset contains 53,940 entries.

Table

Description automatically generated

|  |  |  |
| --- | --- | --- |
| Chart, histogram  Description automatically generated | Chart, histogram  Description automatically generated | Price is Target Variable  Range: 326 ~ 18823  Mean: 3932.799722  Median: 2401  Mean > Median : right skew  High peak around 1,000 USD |

**Numerical variables** – carat, depth, table, x, y, z

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated | Chart, scatter chart  Description automatically generated | Chart, box and whisker chart  Description automatically generated |

**Categorical variables** – cut, color, clarity

|  |  |
| --- | --- |
| Diagram, schematic  Description automatically generated | Diagram  Description automatically generated with low confidence |
| Chart  Description automatically generated with medium confidence | |

Among the independent variables, we can inspect that there some independent variables have very strong correlation with target variable (price) : carat (0.92), x(0.88), y(0.87), z(0.86). Also high correlation among independent variables (x, y, z and carat).

|  |  |
| --- | --- |
| Calendar  Description automatically generated | A picture containing text, white  Description automatically generated |

Missing values are when the dimension of the diamonds x = length, y = width, z = depth equal to 0 as it is not normal when it is 0, we treated it as missing values and remove them.

**Results from 2 models:** Multiple linear regression and Random Forest Regressor are being build. Also, due to the wide range of variables. We scale the variables

Model Summary is listed as below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R2** | **Adjusted R2** | **MSE** | **MAE** |
| Linear Regression | 0.9195 | 0.9193 | 1,290,368 | 726.29 |
| Random Forest | 0.9828 | 0.9827 | 274,684 | 263.92 |
| Linear Regression *(standardized)* | 0.9195 | 0.9193 | 1,290,368 | 726.29 |
| Random Forest  *(standardized)* | 0.9830 | 0.9827 | 272,372 | 263.71 |

According to the result shown in the table, both models can explain more than 90% of the variability of the response data. As for the performance, the Random Forest has higher R2 and lower MSE and MSAE which indicated Random Forest models explain more of the response data and have smaller mean squared error.

**Findings so far:** Random forest regressor after standardization of numerical independent variables have the best performance among them.

**Task 4: Deep Learning model using multiple linear Regression**

We also build deep Learning sequential model to predict the price from diamonds dataset and the performance is also listed in below table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **R2** | **Adjusted R2** | **MSE** | **MAE** |
| Linear Regression | 0.9195 | 0.9193 | 1,290,368 | 726.29 |
| Random Forest | 0.9828 | 0.9827 | 274,684 | 263.92 |
| Linear Regression (standardized) | 0.9195 | 0.9193 | 1,290,368 | 726.29 |
| **Random Forest**  **(standardized)** | 0.9830 | 0.9827 | 272,372 | 263.71 |
| Deep Learning | 0.976 | 0.976 | 380,527 | 302.14 |

**To sum up**: the Random Forest with standardized independent variables are with the best result among these models.