**Data to graph – raw data**

图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成日程表

中度可信度描述已自动生成图形用户界面

中度可信度描述已自动生成图表, 散点图

描述已自动生成散点图

低可信度描述已自动生成

Column Min Max Mean Median

0 NonVIP\_average\_click 0.0 1.505106e+06 2.462947e+03 220.0

1 reviews\_num 0.0 8.420340e+05 2.613584e+02 0.0

2 collected\_num 0.0 1.208338e+06 7.059624e+02 1.0

3 nutrient\_num 0.0 2.226500e+06 5.237111e+02 0.0

4 credits 0.0 3.152578e+10 1.250009e+07 51959.5

5 progress -1.0 1.000000e+00 1.020132e-01 0.0

6 contract\_status 0.0 1.000000e+00 1.205962e-01 0.0

7 total\_word\_count 0.0 3.281276e+06 3.579993e+04 3447.0

8 days\_from\_min\_date 0.0 6.444000e+03 1.884934e+03 24.0

(progress

0 0.807007 -- ongoing

1 0.147503 -- finished

-1 0.045490 -- stopped

Name: proportion, dtype: float64,

contract\_status

0 0.879404 – not signed

1 0.120596 – signed

Name: proportion, dtype: float64)

**Clustering data analysis graph**

图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成

**Absolute Difference between prediction and real**

**Linear regression on credits v.s. NonVIP\_average\_click**

图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成

|  |  |  |
| --- | --- | --- |
|  | Prediction on credits | Prediction on NonVIP\_average\_click |
| MSE | 3147262948330286.0 | 123263078.13311392 |
| MSE^0.5 | 56100471 | 11102 |
| R^2 | 0.9907407802675333 | 0.8797413223108169 |
| Medium of real data | 51959 | 220.0 |
| MSE^0.5 : Medium of real data | 1079 | 50 |

**Linear Regression v.s. neuron network v.s. random forest**

**图表, 散点图

描述已自动生成** **图表, 散点图

描述已自动生成** 图表, 散点图

描述已自动生成

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Linear Regression** | **neuron network** | **random forest** |
| **MSE** | 120329312 | 2142283 | 1406065 |
| **R^2** | 0.807951012290418 | 0.7088330900507047 | 0.8088956557384459 |
| **MAE** | **10969** | **1463** | **1185** |

**Random Forest improving**

**K = 3/5/10**

**图表, 散点图

描述已自动生成**图表, 散点图

描述已自动生成**图表, 散点图

描述已自动生成**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **3** | **5** | **10** |
| **MSE** | 7433678 | 1406065 | 550615 |
| **R^2** | 0.8714027552351264 | 0.8088956557384459 | 0.7101768946867036 |

**Filter range -- 10+, 100+, 1000+, 10~100000, 10~1000000**

**图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成**

**图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **10+** | **100+** | **100+** | **10~100000** | **10~1000000** |
| **MSE** | 7433678 | 47095540 | 105302794 | 1283744 | 6185304 |
| **R^2** | 0.8714027552351264 | 0.7662261880085788 | 0.824400975112541 | 0.6692809499665875 | 0.7467951525211289 |

**K-fold k=5, Kmeans k=5, filter >10**

MSE scores for each fold: [1363196.76059947 1518224.70763707 1710628.93328473 1739091.11196168

1996251.1573239 ]

Average MSE: 1665478.534161368

R² scores for each fold: [0.81472208 0.76514415 0.76444384 0.8013634 0.74725925]

Average R²: 0.7785865460776179

K = 3

MSE scores for each fold: [ 7068044.54755345 10608554.18354652 6121813.11992536 7811594.39032592

10412572.01525727]

Average MSE: 8404515.651321704

R² scores for each fold: [0.87772796 0.79920202 0.88404796 0.83089972 0.78702552]

Average R²: 0.8357806345931185