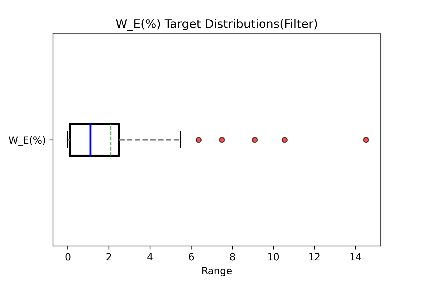
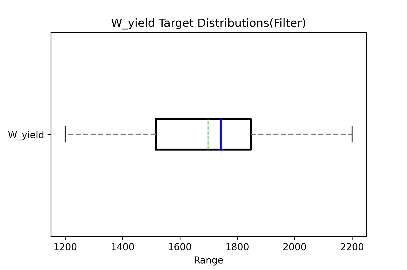
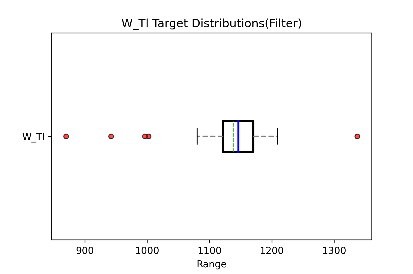
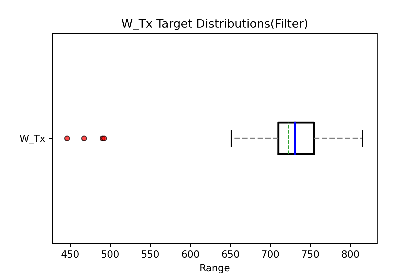
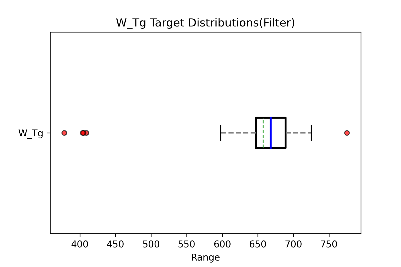
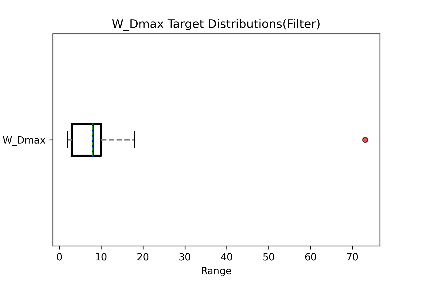
## ZrCu系列非晶合金机器学习实验报告

### 原始整理数据分析

#### 1.1 数据分析

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Dmax** | **Tg** | **Tx** | **Tl** | **yield** | **Modulus** | **Ε(%)** |
| **count** | **75** | **75** | **72** | **64** | **56** | **19** | **49** |
| **mean** | 7.994667 | 657.448 | 722.9708 | 1137.872 | 1697.911 | 121.8474 | 2.096286 |
| **std** | 8.756587 | 68.1413 | 69.63291 | 64.7071 | 219.624 | 69.43528 | 3.030064 |
| **min** | 2 | 378 | 446 | 870 | 1200 | 73 | 0 |
| **25%** | 3 | 647 | 710 | 1121.75 | 1516.25 | 81 | 0.11 |
| **50%** | 8 | 668 | 730.85 | 1146.5 | 1742 | 88 | 1.1 |
| **75%** | 10 | 689 | 754.7 | 1170.175 | 1847 | 107.5 | 2.5 |
| **max** | 73 | 775 | 815 | 1337 | 2200 | 279.2 | 14.5 |

该表格展示了7个目标性能（Dmax，Tg，Tx，Tl，yield，Modulus，Ε）的统计概况，包括count（观测数量）、mean（平均值）、std（标准差）、min（最小值）、25%（第一四分位数）、50%（中位数）以及75%（第三四分位数）和 max（最大值）。

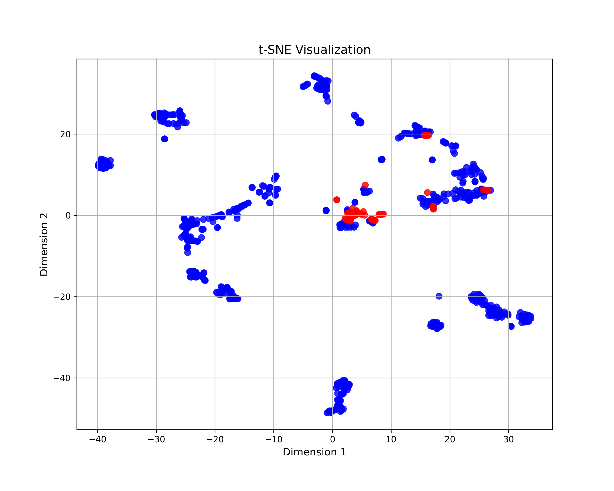


1. **Dmax**：分布范围为2到**73**，中位数为8，平均值约为7.99。观测值集中在较低值区间，右侧尾部较长，分布可能呈现正偏态。**从箱式分布中可以看到有一个异常点，仅有一个值来到了70，这也是后续预测的难点。**
2. **Tg**：分布范围为378到775，中位数为668，平均值约为657.45。分布相对较为集中，**标准差较小（约为68.14），呈现近似对称的分布形态。**
3. **Tx**：分布范围为446到815，中位数为730.85，平均值约为722.97。分布相对较为集中，**标准差较小（约为69.63），呈现近似对称的分布形态。**
4. **Tl**：分布范围为870到1337，中位数为1146.5，平均值约为1137.87。分布集中，**标准差较小（约为64.71），呈现近似对称的分布形态。**
5. **yield**：分布范围为1200到2200，中位数为1742，平均值约为1697.91。分布相对集中，**标准差略大（约为219.62）**，呈现近似对称的分布形态，可能存在一些异常值。
6. **Modulus**：分布范围为73到279.2，中位数为88，平均值约为121.85。观测值集中在较低值区间，右侧尾部较长**，分布可能呈现正偏态，标准差较大（约为69.44）**。
7. **Ε（%）**：分布范围为0到14.5，中位数为1.1，平均值约为2.1。观测值集中在较低值区间，**右侧尾部较长，分布可能呈现正偏态，标准差较大（约为3.03）**。

* 对于那些近似对称，标准差较小，且数据集中的特征（Tg、Tx、Tl），机器学习预测的难度相对较小。这是因为这些特性的值在一个较小的范围内变化，并且分布相对均匀。在这种情况下，机器学习模型可以更容易地识别数据之间的关系并进行预测。
* 而对于正偏态分布的特征（Dmax、Modulus、Ε、yield）且标准差交大的目标，机器学习预测的难度可能会稍大。**这是因为在这些分布中，观测值集中在较低值区间，而较高值的观测较少。**在这种情况下，机器学习模型可能在捕捉较高值区间的数据规律方面表现不佳。**针对这类特征，可以尝试对数据进行预处理，如对数变换或Box-Cox变换，以减少偏度并使数据更接近正态分布。这有助于提高模型的预测性能。**

#### 1.2 迁移学习ML预测新数据

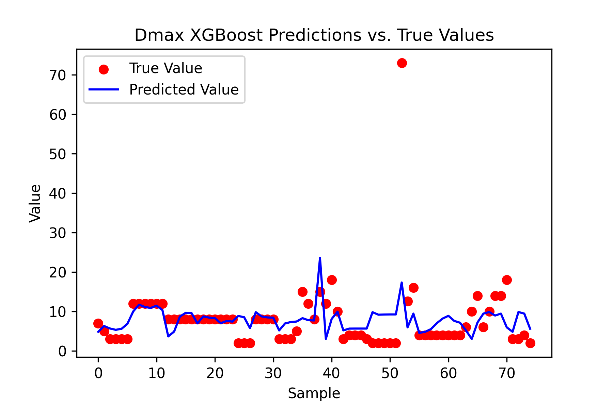
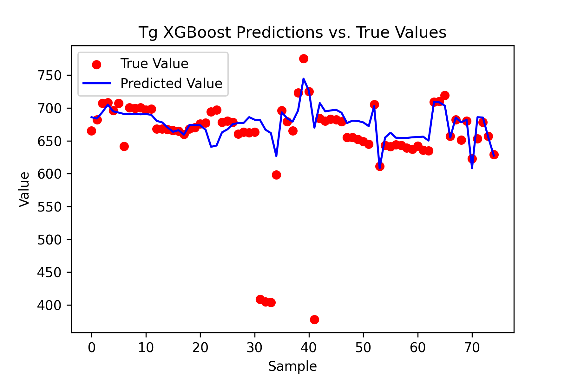
首先使用t-SNE方法将两个数据集的成分特征降维到2维后绘制特征分布图，**红色的点是新数据，蓝色的是原始数据**，可以发现红色的点多聚集在原始数据的一个团簇中。

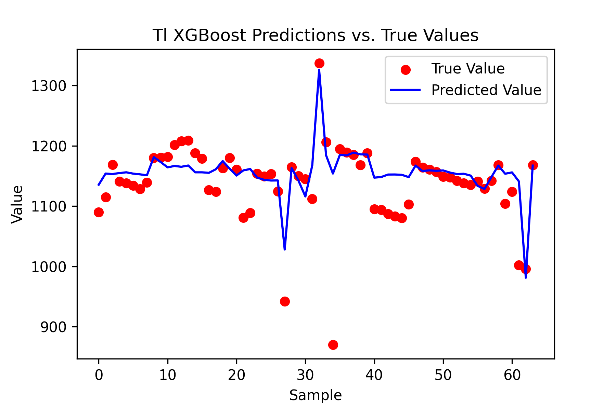
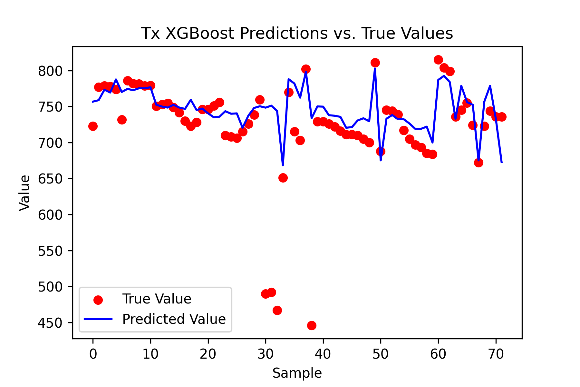


**实验结果：**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **r2** | **MAPE** | **RMSE** |
| **Dmax** | **0.207309** | 0.803753 | 7.744121 |
| **Tg** | **0.07367** | 0.05594 | 65.14461 |
| **Tx** | **0.043335** | 0.055065 | 67.63281 |
| **Tl** | **0.371379** | 0.028114 | 50.90104 |

Dmax决定系数（R2）为0.2073，表明模型只解释了约20.73%的数据变异,**特别在73的异常点预测差距特别大**。总体来看，Dmax的预测效果不佳。Tg,Tx,Tl因为数据分布和原始数据分布有较大差距，从下图看，**整体的预测效果还可以，但是在一些特殊点，模型的偏差特别大，导致最后的平均指标也被拉低了很多**，所以也这个迁移学习模型在预测新数据集时的效果不佳。

#### 新数据5-Fold交叉验证结果

* **Dmax**

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean R2 Score** | **Mean RMSE Score** |
| **Ridge** | 0.06930657 | 4.84766121 |
| **Lasso** | 0.18078265 | 4.58210114 |
| **ElasticNet** | 0.18327746 | 4.59795483 |
| **SVR** | 0.42126007 | 3.90119126 |
| **RandomForestRegressor** | **0.55704548** | 3.37575318 |
| **GradientBoostingRegressor** | 0.53278036 | 3.47344088 |
| **AdaBoostRegressor** | 0.32182854 | 4.15983809 |
| **KNeighborsRegressor** | 0.44340933 | 3.66646819 |
| **XGBRegressor** | 0.5222843 | 3.48984557 |

* **Tg**

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean R2 Score** | **Mean RMSE Score** |
| **Ridge** | 0.94557985 | 40.2159964 |
| **Lasso** | 0.94568677 | 40.1718798 |
| **ElasticNet** | 0.94557351 | 40.2258831 |
| **SVR** | 0.50173577 | 123.243753 |
| **RandomForestRegressor** | 0.96751559 | 30.1150116 |
| **GradientBoostingRegressor** | 0.97414161 | 27.281239 |
| **AdaBoostRegressor** | 0.92973592 | 46.2169687 |
| **KNeighborsRegressor** | 0.9802268 | 23.9100756 |
| **XGBRegressor** | **0.98378611** | 22.0632923 |

* **Tl**

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean R2 Score** | **Mean RMSE Score** |
| **Ridge** | 0.95902939 | 53.3392889 |
| **Lasso** | 0.96208462 | 51.3444538 |
| **ElasticNet** | 0.96445794 | 49.8040892 |
| **SVR** | 0.4099417 | 205.955115 |
| **RandomForestRegressor** | 0.97794774 | 39.3613163 |
| **GradientBoostingRegressor** | 0.97923496 | 38.1878788 |
| **AdaBoostRegressor** | 0.94232656 | 64.4289903 |
| **KNeighborsRegressor** | 0.97893927 | 38.2336608 |
| **XGBRegressor** | **0.98093027** | 36.4335728 |

* **Tx**

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean R2 Score** | **Mean RMSE Score** |
| **Ridge** | 0.85086822 | 60.6446084 |
| **Lasso** | 0.92973464 | 47.8997237 |
| **ElasticNet** | 0.932827 | 46.6294711 |
| **SVR** | 0.4842034 | 130.033105 |
| **RandomForestRegressor** | 0.97879164 | 25.91182 |
| **GradientBoostingRegressor** | 0.98115523 | 24.5134145 |
| **AdaBoostRegressor** | 0.92949016 | 48.0565943 |
| **KNeighborsRegressor** | **0.98326914** | 23.2354407 |
| **XGBRegressor** | **0.98041186** | 24.7671783 |

* **Yield**

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean R2 Score** | **Mean RMSE Score** |
| **Ridge** | 0.11238077 | 194.428377 |
| **Lasso** | 0.15621694 | 189.381448 |
| **ElasticNet** | 0.26020696 | 178.573599 |
| **SVR** | 0.23345372 | 182.321859 |
| **RandomForestRegressor** | 0.18475735 | 184.631225 |
| **GradientBoostingRegressor** | 0.20094268 | 187.236491 |
| **AdaBoostRegressor** | 0.2414352 | 181.407161 |
| **KNeighborsRegressor** | **0.34522651** | 168.004501 |
| **XGBRegressor** | 0.11335424 | 192.551276 |

* **Modulus**

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean R2 Score** | **Mean RMSE Score** |
| **Ridge** | -191.20931 | 107.352871 |
| **Lasso** | -247.40438 | 120.020582 |
| **ElasticNet** | -21.778607 | 69.5927429 |
| **SVR** | -0.3985699 | 19.7591294 |
| **RandomForestRegressor** | -7.1619123 | 49.212028 |
| **GradientBoostingRegressor** | -11.386971 | 48.1256294 |
| **AdaBoostRegressor** | -24.926495 | 38.5312618 |
| **KNeighborsRegressor** | -13.674777 | 55.3394845 |
| **XGBRegressor** | **-0.4977218** | 28.2635779 |

* **E(%)**

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean R2 Score** | **Mean RMSE Score** |
| **Ridge** | -1.6714499 | 3.12185956 |
| **Lasso** | -1.193705 | 2.79844991 |
| **ElasticNet** | -1.193705 | 2.79844991 |
| **SVR** | -0.1736526 | 2.51181851 |
| **RandomForestRegressor** | -1.6426775 | 2.73644429 |
| **GradientBoostingRegressor** | -0.6862932 | 2.71406269 |
| **AdaBoostRegressor** | -2.2742936 | 3.15700393 |
| **KNeighborsRegressor** | -0.6693847 | 2.53531106 |
| **XGBRegressor** | **-0.0840844** | 2.48329371 |

### 整理文献新增数据分析

#### 2.1 数据分析

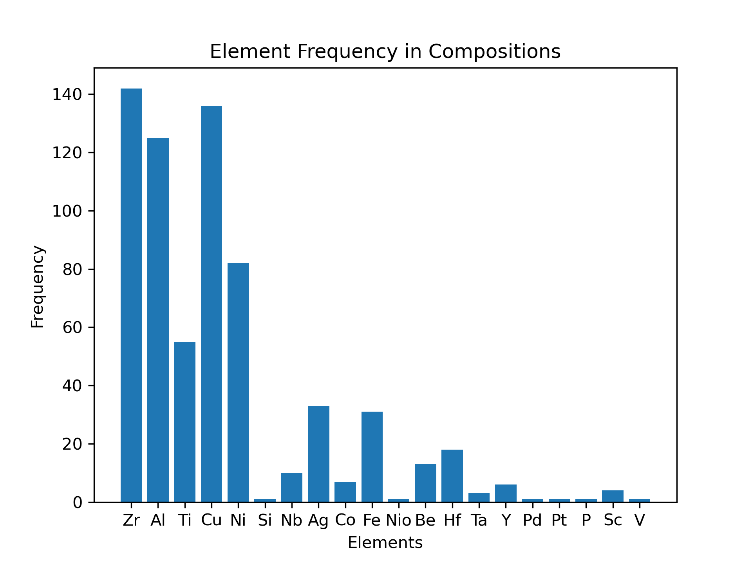
相同成分不同性能数据处理：

* 相同成分的如果都找到了论文支撑，Dmax取较大值，其余值取平均，如果有未找到文献支撑的，则仅保留找到的
* 对未找到支持文献的特殊值进行删除（如σy(MPa) 2200的值）
* 相同成分，如果存在一篇文献的体系数据更加完整，则仅保留该文献中的结果（怕实验方法不同，测量误差，如一个体系随着Ti的成分增加，Tx呈现下降趋势，另一篇文献中也有该成分，但是Tx偏高，破坏了这个趋势则不予保留）

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Critical Diameter/thickness(mm)** | **Tg(K)** | **Tx(K)** | **Tl(K)** | **σy(MPa)** | **Modulus (GPa)** | **Ε(%)** |
| **count** | **127** | **116** | **111** | **93** | **84** | **32** | **64** |
| **mean** | 9.892126 | 652.2888 | 720.0171 | 1132.503 | 1668.774 | **122.925** | **3.541875** |
| **std** | 10.36924 | 84.78649 | 87.81938 | 83.53289 | 246.6157 | **67.94109** | **5.241194** |
| **min** | 1.5 | 378 | 446 | 820 | 919 | 73 | 0 |
| **25%** | 3 | 648 | 710 | 1104 | 1516.25 | 86.525 | 0.375 |
| **50%** | 8 | 670 | 738.4 | 1144 | 1705 | 95.55 | 1.565 |
| **75%** | 12 | 696.85 | 777 | 1179.9 | 1839.25 | 105.25 | 5.425 |
| **max** | 73 | 802 | 818 | 1337 | 2204 | 284.3 | 32.9 |

总共143条数据，每一条数据都经过2次数据验证，并且验证过每一条数据成分之和都为100%，自己寻找文献增加了60多条数据，在每个目标性能上普遍增加了50%的数据，数据整理仓库[**https://github.com/yuyouyu32/BMG\_Data\_SHU**](https://github.com/yuyouyu32/BMG_Data_SHU)**。**

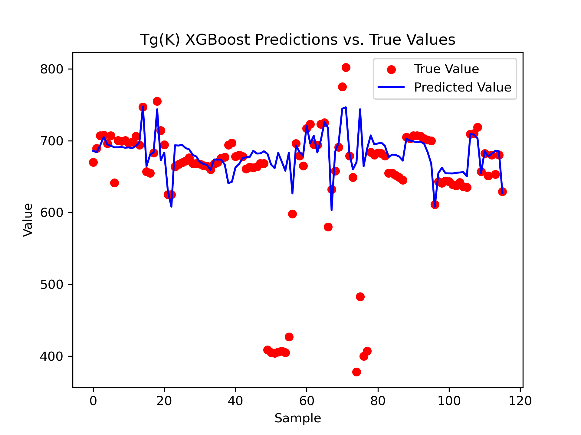
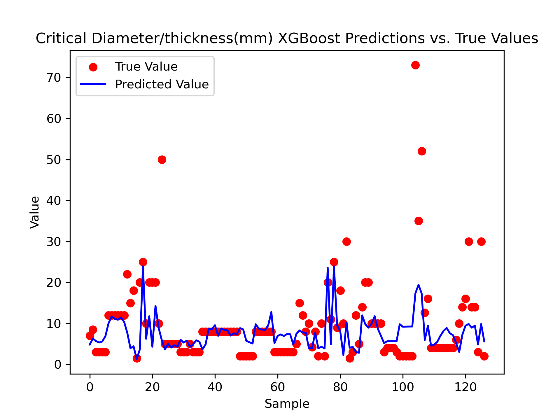
**成分元素分布柱形图：**

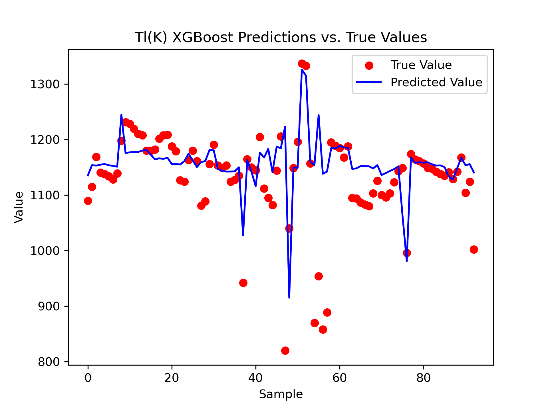
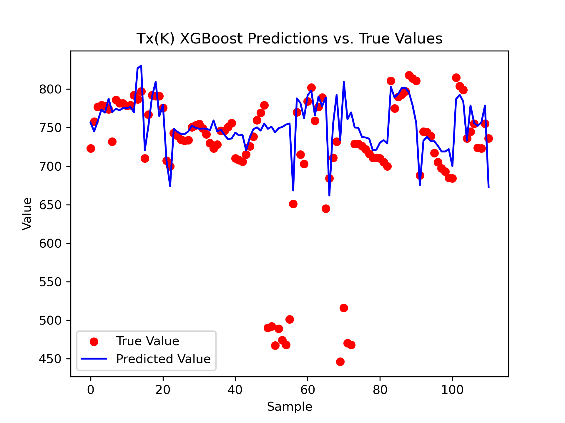


整理的体系都为ZrCuNiAl体系的非晶合金。

#### 2.2 迁移学习ML预测新数据

|  |  |  |  |
| --- | --- | --- | --- |
|  | **r2** | **MAPE** | **RMSE** |
| **Dmax** | **0.220966** | **0.694589** | 9.116089 |
| **Tg** | **-0.02094** | **0.083073** | 85.29936 |
| **Tx** | **-0.08513** | **0.08123** | 91.0679 |
| **Tl** | **0.043162** | **0.043943** | 81.26979 |

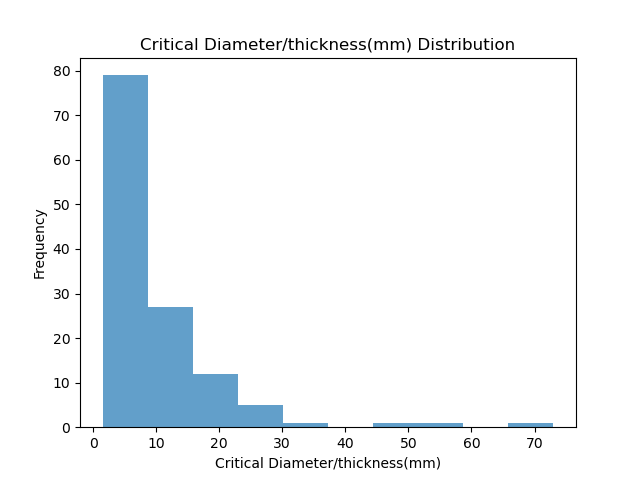
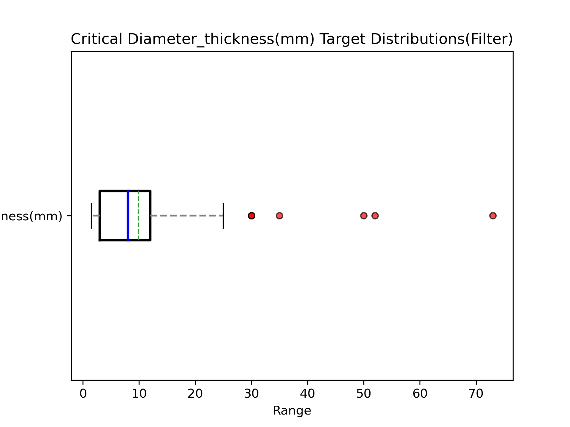


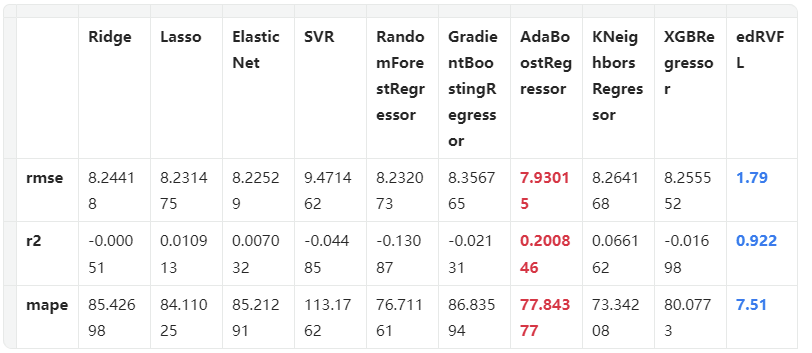


**移学习的效果不尽人意，这可能是由于目标域和源域之间的偏差较大，导致模型在新数据集上的表现受到影响。源域和目标域之间的分布差异可能会导致模型不能很好地泛化到目标域。此外，目标域数据的成分比较集中，这意味着它们都属于同一个体系，具有相似的特征和相互关联。在这种情况下，预测偏差可能会较大。**

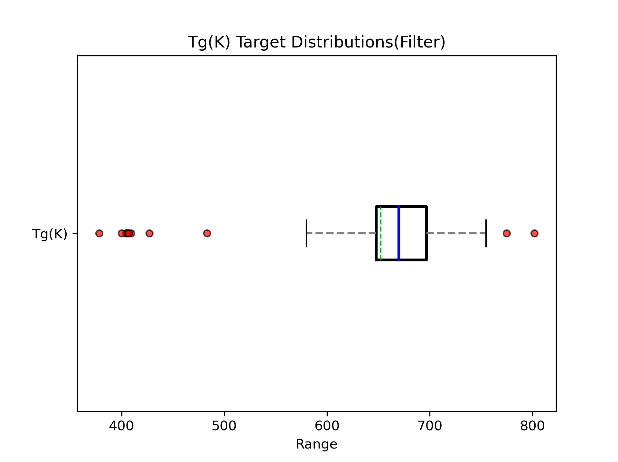
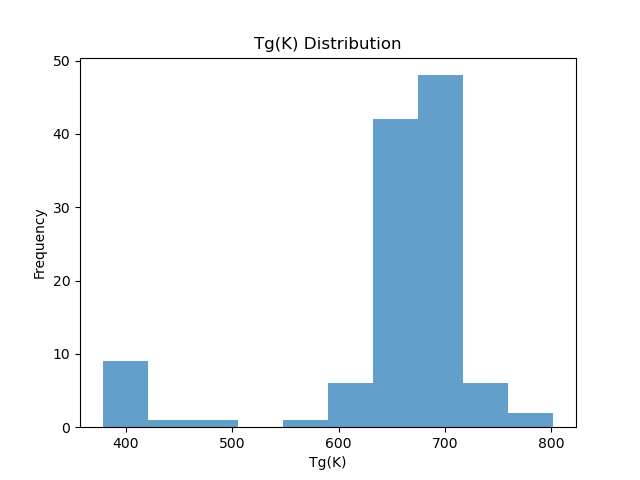
#### 新增数据5-Fold交叉验证结果（3-fold和10-fold也有结果，这里只放5-fold）

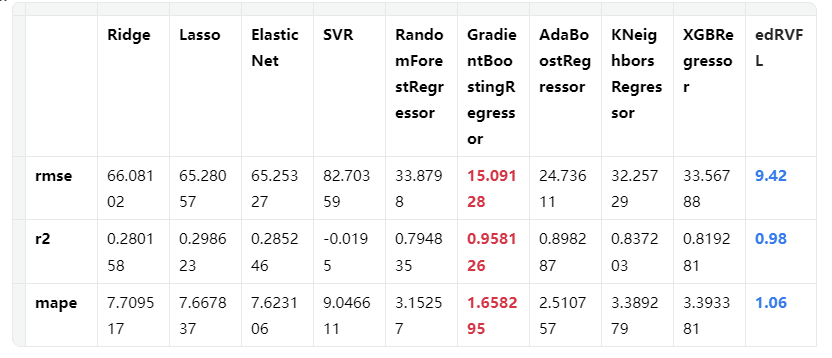
* **Dmax**

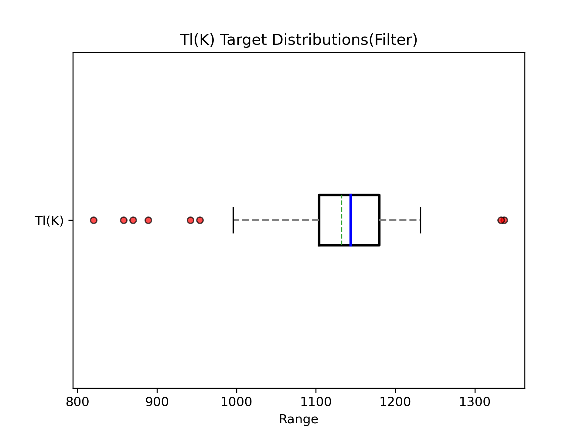
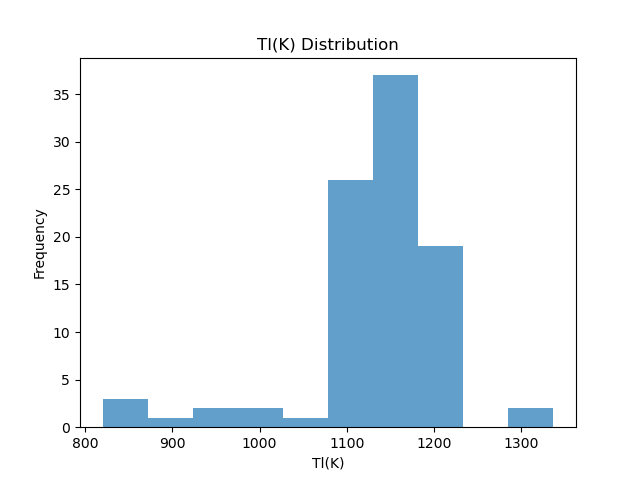


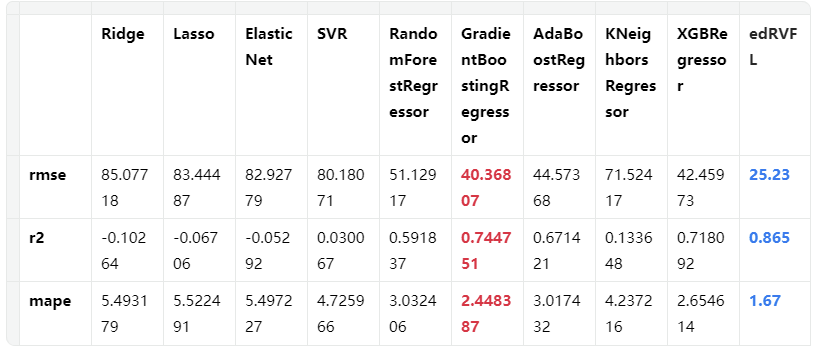
* **Tg**



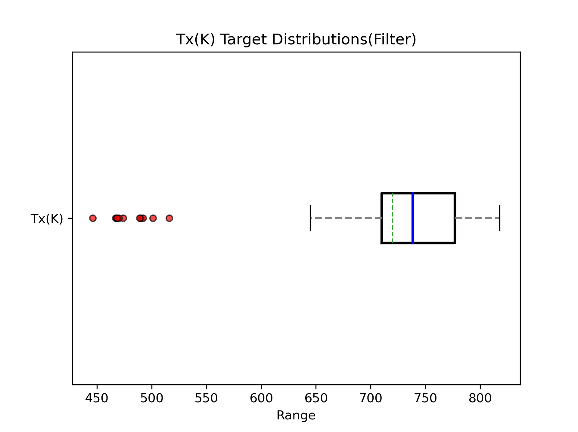
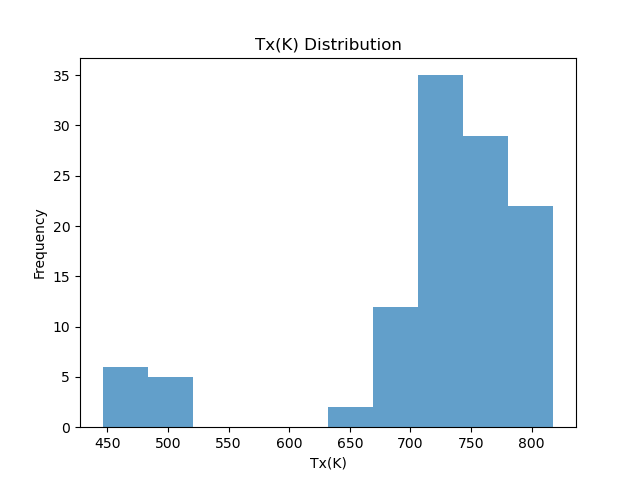


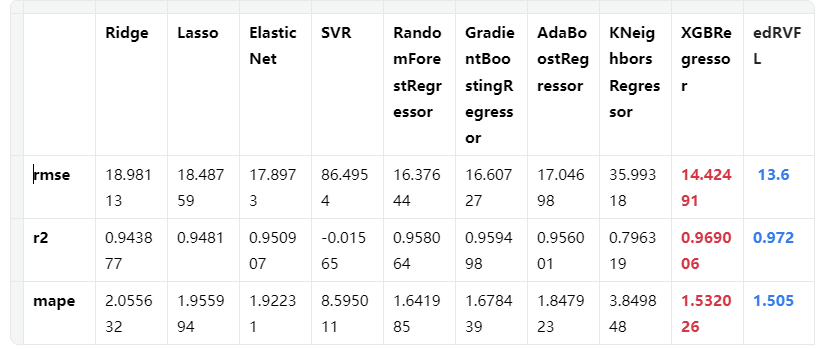
* **Tl**



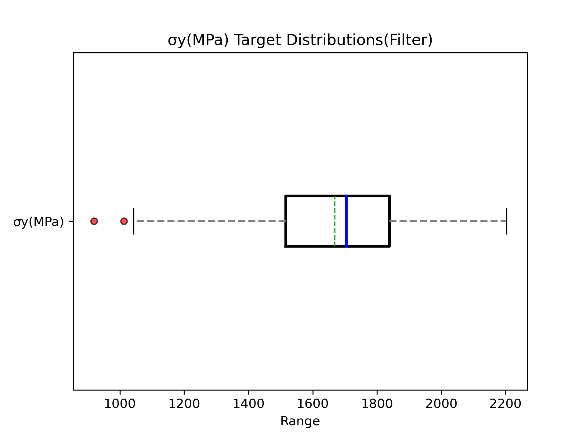
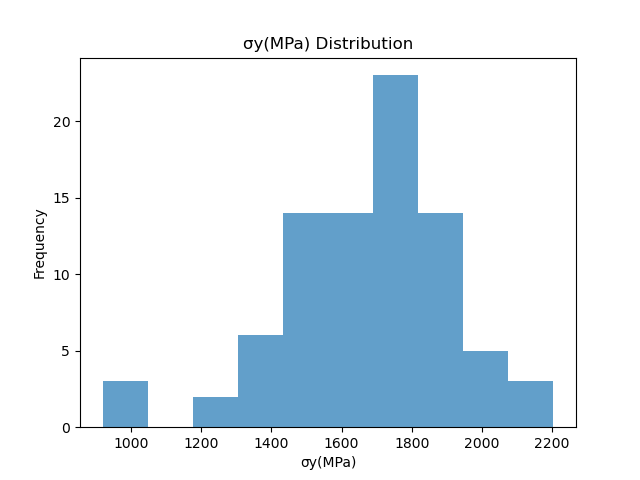


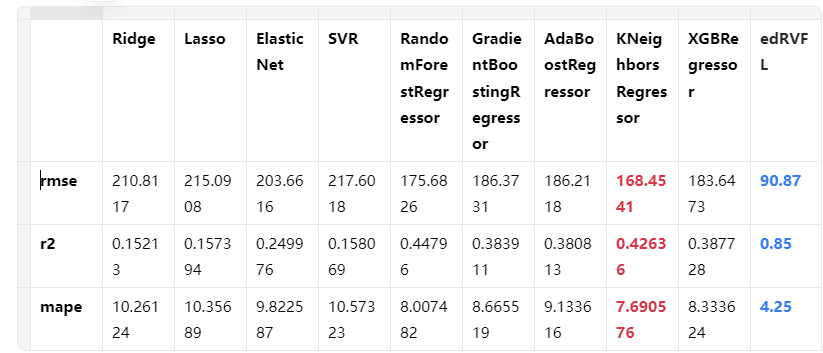
* **Tx**



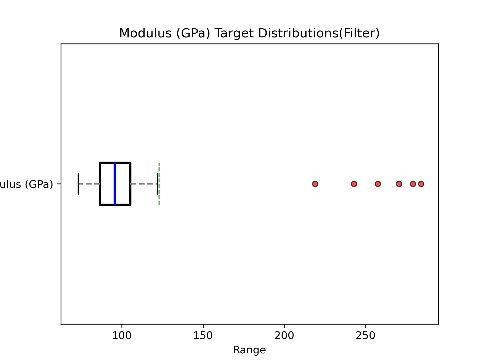
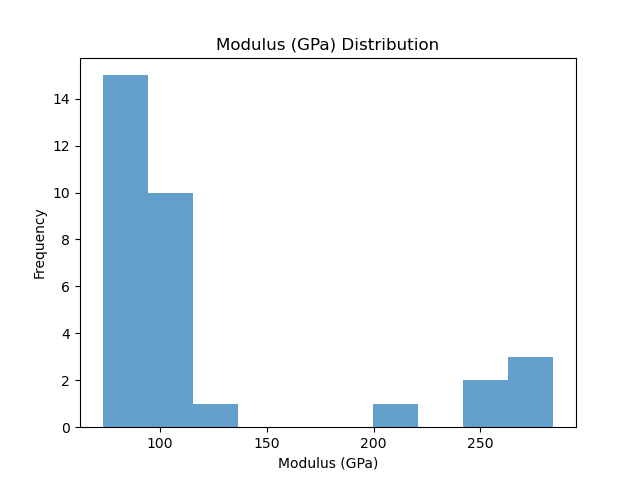


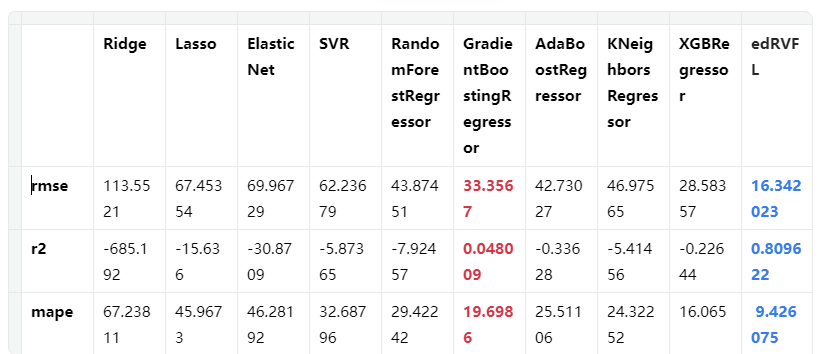
* **Yield**



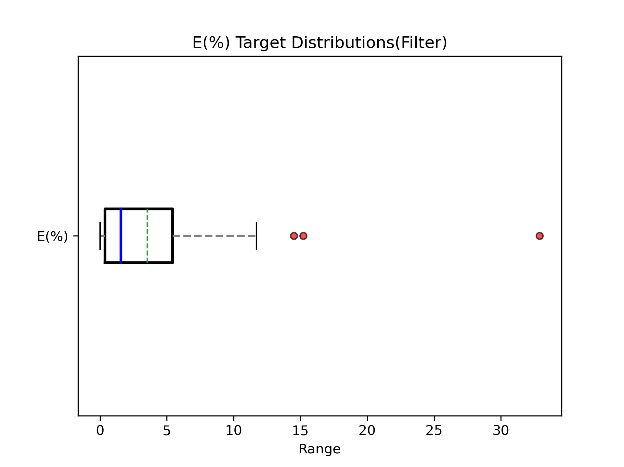
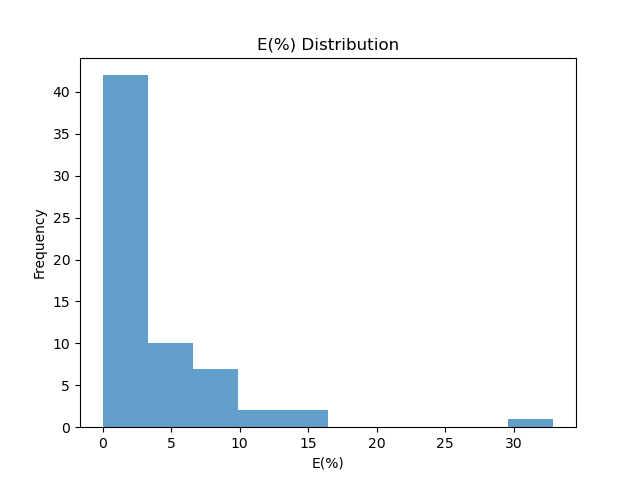


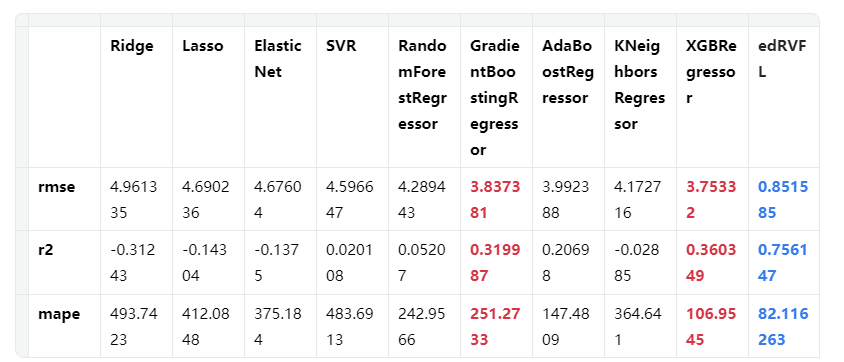
* **Modulus**





* **E(%)**



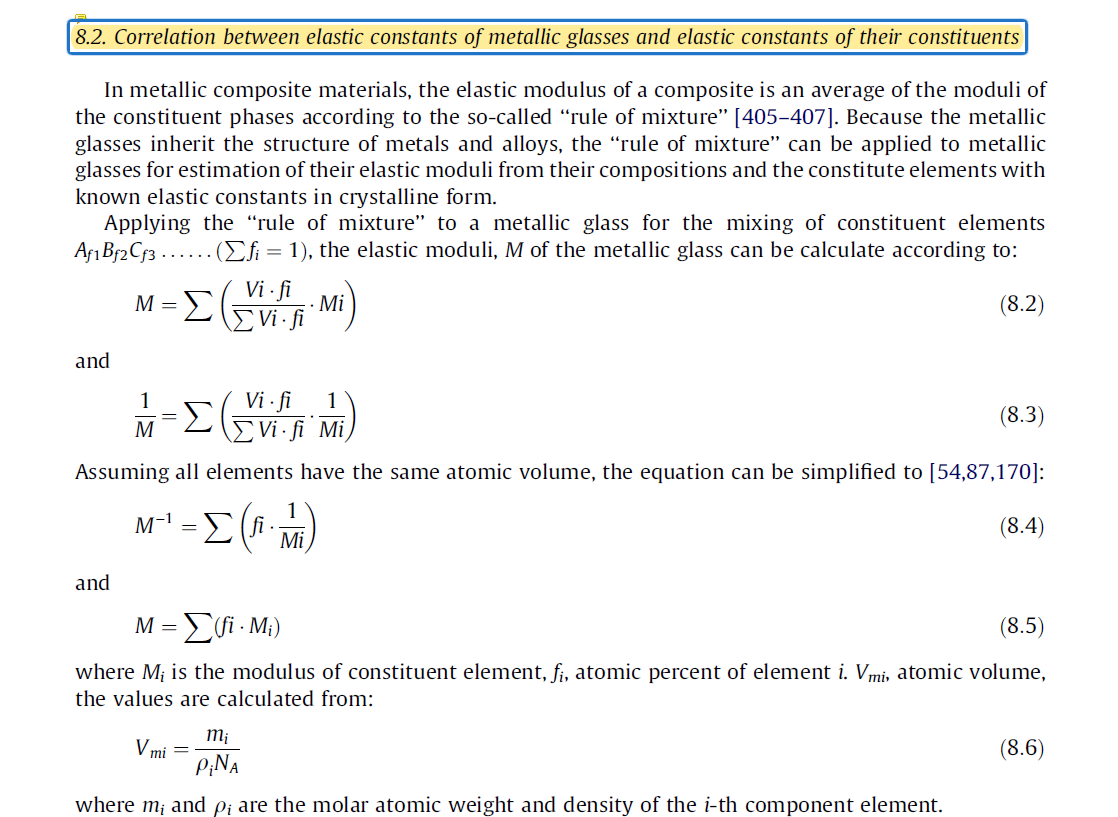


* **总结**

**使用edRVFL模型搜索超参数后可以得到5-fold结果，最差的E(%)的R2来到了0.75,但是MAPE比较差，仍有82%，可能是因为E(%)值都比较小的原因，这个也能从箱图中看出来，大部分都集中在0附近，但也有30%左右的值。其余的R2都达到了0.8以上。**

**现在希望可以得到卞老师这里的指导，主要指导7个目标参数优化方向，强化学习智能体学习的奖励函数设计等，推进后续的强化学习实验选点。**

* **熊老师推荐书籍中的公式**



* **在金属复合材料中，复合材料的弹性模量是各组成相模量的平均值。由于金属玻璃继承了金属和合金的结构，因此 "混合物法则 "可应用于金属玻璃，以估算其弹性模量。根据其成分和晶体中具有已知弹性常数的构成元素来估算其弹性模量。**

