离散制造过程中典型工件的质量符合率预测

不愿意透露队名队

PB17081542 徐扬，PB17081544 张焰舒，PB17071417 毕超

University of Science and Technology of China,
Hefei, Anhui, China

Abstract: 由于在实际生产中，同一组工艺参数设定下生产的工件会出现多种质检结果，如何根据工艺参数预测产品质量，对离散制造企业意义重大。对于类似的纯数值分类问题，一般的解题思路是根据给出的工艺参数使用一系列基于 Gradient Boosting 框架的集成学习算法直接对类别进行预测，是单阶段的模型。本文对单阶段模型的基础上给出双阶段模型的两种不同思路，即先利用工艺参数回归预测工件属性，再利用工件属性预测样本类别，并通过结果证明双阶段模型在不同的测试集上具有更好的适应能力。作为数据科学导论的入门实验，我们也熟悉了数据科学的基本流程，对数据科学的问题分析、解答有了一定的理解。

Keywords: 离散制造；分类问题；双阶段模型。

1. Introduction

离散制造型企业生产过程是由不同零部件加工子过程或并联或串连组成的复杂的过程，其过程中包含着更多的变化和不确定因素。从这个意义上来说，离散制造型企业的过程控制更为复杂和多变。

离散制造型的企业的产能不像连续型企业主要由硬件决定，而主要以软件（加工要素的配置合理性）决定。同样规模和硬件设施的不同离散型企业因其管理水平的差异导致的结果可能有天壤之别，从这个意义上来说，离散制造型企业通过软件方面的改进来提升竞争力更具潜力。

而在高端制造领域，随着数字化转型的深入推进，越来越多的数据可以被用来分析和学习，进而实现制造过程中重要决策和控制环节的智能化，例如生产质量管理。从数据驱动的方法来看，生产质量管理通常需要完成质量影响因素挖掘及质量预测、质量控制优化等环节，本赛题将关注于第一个环节，基于对潜在的相关参数及历史生产数据的分析，完成质量相关因素的确认和最终质量符合率的预测。在实际生产中，该环节的结果将是后续控制优化的重要依据。

由于在实际生产中，同一组工艺参数设定下生产的工件会出现多种质检结果，所以

我们针对各组工艺参数定义其质检标准符合率，即为该组工艺参数生产的工件的质检结果分别符合优、良、合格与不合格四类指标的比率。相比预测各个工件的质检结果，预测该质检标准符合率会更具有实际意义。

本赛题要求参赛者对给定的工艺参数组合所生产工件的质检标准符合率进行预测。训练集集共 6000 个工件样本，每个样本有 10 种工艺参数，10 种工件属性共 20 种特征以及工件所符合的质检指标，共分四种类别 Excellent，Good，Pass，Fail。测试集共 6000 个工件样本，每个样本有 10 种工艺参数以及对应组别（每 50 个样本分为一组，共 120 组）。目标是对测试集中每组的质检标准符合率进行预测，即产生一个 120×4 的概率矩阵。

我们采用三种思路对每个样本的质检指标进行预测，再通过组内质检指标计算每组质检标准符合率，以下工艺参数称为 Parameter，工件属性称为 Attribute，质检指标称为 Label。

2. Experiment

2.1. 解题思路

对于此次赛题，由于测试集中的 Attribute 未给出，因此存在三种解题思路：

1. 忽略训练集、测试集中的 Attribute，将其当作预测 Label 中的黑箱而处理，直接建立 Parameter 与 Label 的关系，用 Parameter 预测 Label；

2. 先寻找训练集中 Attribute 与 Parameter 之间的关系，将两者回归拟合，用训练集得到的回归拟合模型预测测试集中的 Attribute。这之后，建立 Attribute 与 Label 的关系，用 Attribute 预测 Label。

3. 思路近似 2，但在预测 Label 时，使用 Attribute+Parameter 预测 Label。

在官方解读题目的直播视频中，也提及了上述三种思路。因此，我们在本次实验中，对上述三种思路均进行了尝试。

2.2. 实验步骤

对本次实验，采取数据预处理、特征工程、模型选择、训练、线上验证等基本流程。

在数据预处理阶段，我们对参数的数值特征进行观察，决定采用何种预处理方式。

在特征工程阶段，我们分别进行一维、二维、三维的数据 EDA，判断参数的离散/连续性，寻找参数分布特征。

在建模、训练阶段，我们根据三种不同的思路，分别选用不同的模型及参数进行拟合及预测，通过线上验证判断方案的可行性、进行超参数的优化等。

3. Results and Discussion

3.1. 数据预处理

计算参数的数值特征，数据的绝对大小差异极大（Tab. A1-A2）。故此考虑对参数进行 log10 变换，得到较好的数据分布范围（Fig. 1-2）。
初赛数据无空白值、无效值，无需特异处理。

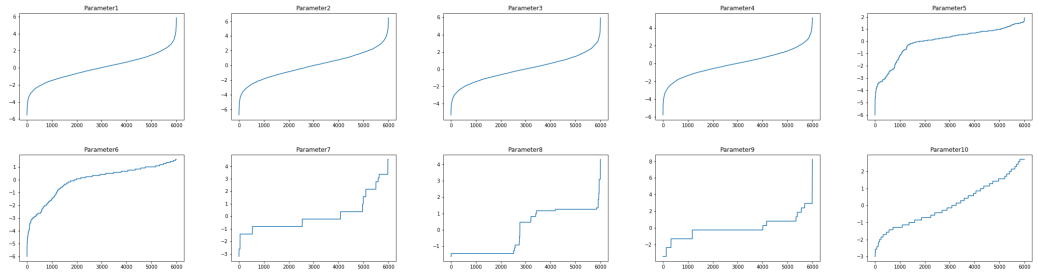

Fig. 1. Parameter 经 log10 变换后分布图象

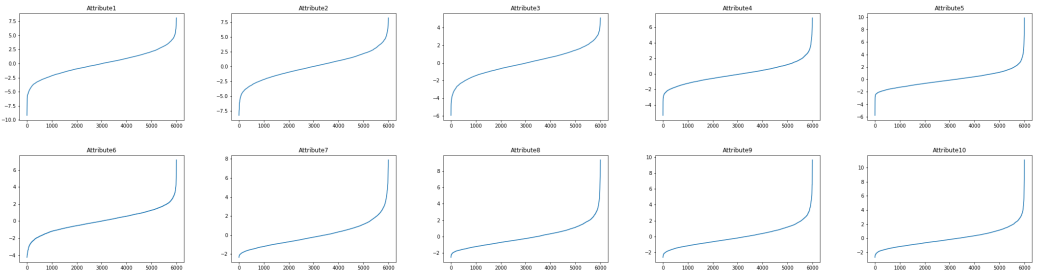

Fig. 2. Attribute 经 log10 变换后分布图象

3.2. 特征工程

对所给参数分布可视化（Fig. 1-2），Para5-10 为离散分布，Para1-4 及 Attr1-10 为连续分布。
受官方比赛群中讨论的启发，对所给参数进行 EDA，尝试寻找参数之间的关联。二维参数 EDA 的关联性分析见 Fig. 3。由数据分布可见，Attr1-3 与 Para1-10 对

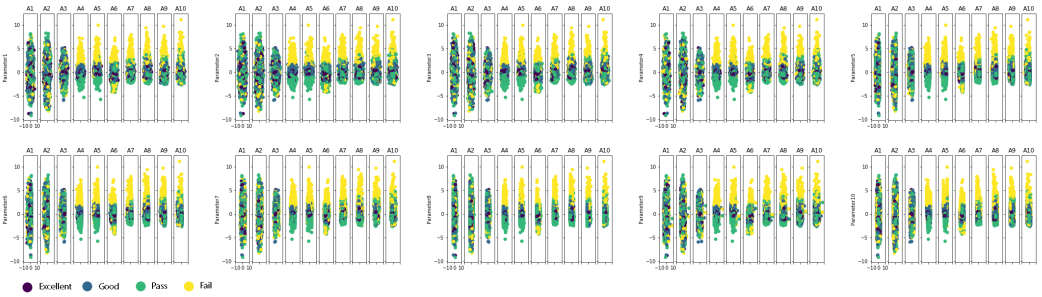

Fig. 3. 二维 EDA 可视化-Parameter~Attribute

Label 的分布关系极为混杂，Attr4-6 与 Para1-10 对 Label 呈现出较强的可划分性，Attr7-10 与 Para1-10 仅在 Fail 的区域上具有可划分性，其余 Label 的分布仍呈现出混杂状态。

三维参数 EDA 选取二维参数 EDA 中，Paremeter 与 Attribute 呈现较强关联的参数进行。由 EDA 结果可发现，Attribute4-6 几乎分布在一个平面上（Fig. 4），且具有较显著的可划分性（Fig. 5），PCA 之后可进行划分（Fig. 6），说明落在某一区间的数据可划分为该 Label。而其他 Attr 的三维分布不具有相应显著的分布特征。鉴于小组成员能力有限，暂时无法实现用划分直接对数据结果进行分类的方案，因此我们根据 Attr4-6 具有对 Label 的分布显著性，对 Attr4-6 直接分类，或对 Attr4-6 进行二维 PCA 后分类。

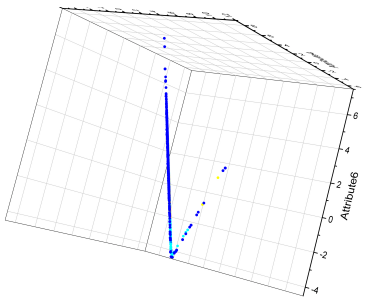

Fig. 4. Attri4-6 三维侧视图

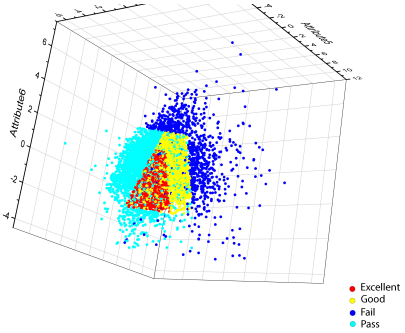

Fig. 5. Attri4-6 三维投影图

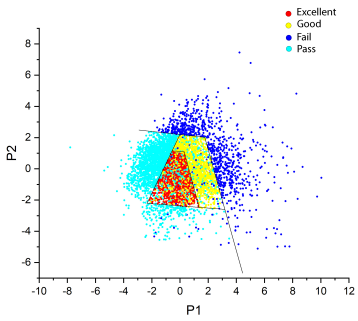

Fig. 6. Attri4-6 PCA 图

3.3. 模型选择

使用 Catboost、Xgboost、Lightboost（Kfold-4）进行测试集 Attribute 的预测。
使用 Catboost（取三个不同的训练轮数，对训练结果取平均）进行测试集 Label 的预测。

3.4. 训练及验证

在思路一（Tab. 1）中，我们选择离散参数 Para5-10，用 Ctb 预测 Label。
在思路二（Tab. 2）中，我们选择多种 Para 组合（见 Tab. 4）最终确定使用 Para5-10 效果最好），用 Lgb/Xgb 预测 Attr4-6，使用 Attr4-6/P1-2 组合预测 Label。
在思路三（Tab. 3）中，我们选择 Para5-10，用 Lgb/Xgb 预测 Attr4-6，再使用 Attr4-6/P1-2+Para5-10 预测 Label。

Table 1. 思路一训练及得分

训练	得分
Ctb Para5-10 预测 Label（初赛 A 榜）	0.68960580
Ctb Para5-10 预测 Label（初赛 B 榜）	0.67109910
Ctb Para5-10 预测 Label（训练赛）	0.68977440
Ctb Para5-10 预测 Label（训练赛）（Kfold-5）	0.68694454

Table 2. 思路二训练及得分

训练	得分
Xgb Para5-10 预测 Attr4-6	
Ctb Attr4-6 预测 Label（训练赛）	0.68318310
Xgb Para5-10 预测 Attr4-6	
Ctb Attr4-6 预测 Label（训练赛）（Kfold-10）	0.68725780
Xgb Para5-10 预测 Attr4-6+PCA-2D	
Ctb P1-2 预测 Label（训练赛）	0.67440420
Xgb Para5-10 预测 Attr4-6+PCA-2D	
Ctb P1-2 预测 Label（训练赛）（Kfold-10）	0.68734413

Table 3. 思路三训练及得分

训练	得分
Lgb Para5-10 预测 Attr4-6	
Ctb Para5-10+Attr4-6 预测 Label（训练赛）	0.68547034
Xgb Para5-10 预测 Attr4-6	
Ctb Para5-10+Attr4-6 预测 Label（训练赛）（Kfold-10）	0.68907860
Ctb Para5-10 预测 Attr4-6	
Ctb Para5-10+Attr4-6 预测 Label（训练赛）（Kfold-10）	0.69120800
Xgb Para5-10 预测 Attr4-6+PCA-2D	
Ctb Para5-10+P1-2 预测 Label（训练赛）	0.68553627
Xgb Para5-10 预测 Attr4-6+PCA-2D	
Ctb Para5-10+P1-2 预测 Label（训练赛）（Kfold-10）	0.69069266
Ctb Para5-10 预测 Attr4-6+PCA-2D	
Ctb Para5-10+P1-2 预测 Label（训练赛）（Kfold-10）	0.69236770

Table 4. Para 组合选择比较

训练	得分
Lgb Para1-4 预测 Attr4-6	
Ctb Attr4-6 预测 Label（训练赛）	0.47944313
Lgb Para5-10 预测 Attr4-6	
Ctb Attr4-6 预测 Label（训练赛）	0.68453664
Xgb Para1-10 预测 Attr4-6	
Ctb Attr4-6 预测 Label（训练赛）	0.67990386
Xgb Para5-10 预测 Attr4-6	
Ctb Attr4-6 预测 Label（训练赛）	0.68318310

Table 5. 预测 Attr 模型比较

训练	得分
Xgb Para5-10 预测 Attr4-6	
Ctb Attr4-6 预测 Label（训练赛）	0.68318310
Lgb Para5-10 预测 Attr4-6	
Ctb Attr4-6 预测 Label（训练赛）	0.68453664
Xgb Para5-10 预测 Attr4-6+PCA-2D	
Ctb Para5-10+P1-2 预测 Label（训练赛）(Kfold-10)	0.69069266
Ctb Para5-10 预测 Attr4-6+PCA-2D	
Ctb Para5-10+P1-2 预测 Label（训练赛）(Kfold-10)	0.69236770

Table 6. 预测 Label 模型比较

训练	得分
Xgb Para5-10 预测 Attr4-6+PCA-2D	
Lgb Para5-10+P1-2 预测 Label（训练赛）(Kfold-10)	0.68238056
Xgb Para5-10 预测 Attr4-6+PCA-2D	
Ctb Para5-10+P1-2 预测 Label（训练赛）(Kfold-10)	0.69069266

对于调参，我们测试之后发现，调参的提升效果有限，故将重点放在模型比较、思路优化上，调参结果此处不再赘述。

对模型的比较分析表明，在预测 Label（Tab. 6）中，Ctb 性能显著优于 Lgb，因此在预测 Label 中主要使用 Ctb；在预测 Attr 中 Ctb 的性能同样显著优于 Lgb/Xgb

(Tab. 5)。

最后，我们获得最高分的思路是 Ctb 预测 Attr4-6，将 Attr4-6 进行二维 PCA 后，取 PCA1-2+Para5-10，Ctb 预测 Label。

线上结果表明，思路一可以取得较高的线上得分，但经 Kfold-5 折训练后得分反而下降，说明思路一的高得分是由于过拟合引起的，在初赛的 B 榜表现较差也说明了过拟合的存在。总之，思路一的鲁棒性较差，我们认为此思路的提升空间有限，因此后续不再对此思路进行深入探究。

思路二、三在未取 Kfold 训练时，得分低于思路一，但在取 Kfold-10 折训练之后得分上升，进一步逼近或超过了思路一的最高得分，说明思路二、三具有较好的可迁移性及鲁棒性，应当作为主要提升的方案。特别地，思路二中的 Attr4-6 及对 Attr4-6 进行二维 PCA 的结果，虽然维度较小，但仍可提供接近或超过仅使用 Para5-10 的预测结果，说明 Attr4-6 可以稳定地显示 Label 的特征，与 EDA 的结论相一致。

由于 Attr4-6 变换后的数值特征呈现连续性，因此对连续变量直接进行分类的效果可能仍欠缺显著性，是制约线上得分进一步上升的因素，需要进一步将 Attr4-6 进行离散化，以取得更好的分类效果。

在实际工业生产中，工件的属性 Attribute 对工件的制造参数 Parameter 遵循一定的分布，而不是一一对应的关系，亦即，即使确定了一定的 Parameter，工件的 Attribute 也是在一个分布内取值的。我们基于这样的背景来讨论思路一-三的优缺点。结合实际工业背景来看，思路一将工件的 Attribute 当作黑箱子，直接跳过工件的 Attribute，通过工件的 Parameter 直接预测工件的 Label，属于单阶段模型。虽然这一思路实现简易、工作量较小，但结果较不稳定，也缺乏现实意义——思路一相当于忽略了工件 Attribute 对 Parameter 的分布，直接认为某一些工件 Parameter 可以确定工件的 Label，是难以让人信服的。

思路二、三先回归预测工件 Attribute 对应工件 Parameter 的分布，再将工件的 Attribute/Parameter+Attribute 同时用于预测工件的质量，属于双阶段模型。双阶段模型虽然实现复杂，工程量较大，但与单阶段模型相比较，泛化性能更好，显著地具有现实意义，符合工业背景。

只是，在此题中为什么会出现大量冗余的、不利于预测的参数（例如 Para1-4、Attr1-3/7-10），为何 Attr4-6 会完美地符合一个 V 字形的分布，又具有良好的可划分性，这一点是我们目前所不能理解的。对冗余的参数，一个初步的解释是，这些参数在工业中是难以控制的、随机变化的，因此这些参数是可以忽略的、在工业生产中弱化重要性的。但对于 Attr4-6 几乎完美的分布情况，目前暂时缺乏较好的

解释。

4. Conclusion

在本次数据科学导论的实验中，我们依照数据科学的基本流程，对工件质量预测题目进行了分析、作答。在这一过程中，我们从零开始，熟悉了数据科学的基本流程，积累了一定的数据分析的经验，从最开始面对题目的无从下手，到可以遵循一定的流程进行分析，这是本次实验的最大收获。

在这次实验中，我们也积累了一些失败的教训，例如，在实验的初期，没有较好地分析参数的特征，在测试的时候也缺乏清晰的思路，不能判断不同参数的重要性，导致初期训练中对数据缺乏预处理的措施，训练参数的选择也几乎全靠猜测；我们也没有注重参数之间的关联分析，导致我们在训练中引入了大量冗余参数，使结果始终原地踏步，不能进一步上升；在实验的后期，我们发现遗漏了在训练预测 Label 时进行 K-fold 检验，导致本该有用的思路没有继续优化。当然，这也是作为初学者，在缺乏一定的知识储备及代码能力，甚至在阅读别人的 Baseline 都具有困难，连 K-fold 的意义都不理解时，所必须经历的过程。在未来的数据科学学习中，我们应当更加谨慎，尽量遵循数据的基本特征进行。

对于本次赛题，我们通过分析、验证，确认了双阶段模型的性能好于单阶段模型。其次，我们认为这道题目本身是具有一定缺陷的。题目对工件参数的意义缺乏直接阐述，导致我们只能通过结果倒推工件参数的重要性，而无法更进一步地结合工业背景来理解赛题，这也是制约我们提高结果的一个因素。当然，在选题调研时，基于我们对数据科学的基础知识非常匮乏的事实，这道题参数数值简单，也没有异常值、无效值，工作量较小，用于入门，是可以接受的。在后续选题时，我们应当选择参数意义明确，可结合现实进行分析的题目，以进一步提高问题分析的能力。

5. Acknowledgements

在本次实验中，徐扬负责报告撰写，张焰舒、毕超负责代码工作，三人均参与了思路的讨论、完善及对思路的测试。在此对小组内成员的通力合作致谢。

在本次实验中，我们也获得了刘淇老师的指导，和李佳桐团队进行了一定的讨论，发现了我们实验中的思维盲区，积累了更多的经验，在此对刘淇老师、李佳桐团队做出致谢。

Appendix A. Appendices

Table A1. Parameter 数值特征

	Parameter1	Parameter2	Parameter3	Parameter4	Parameter5	Parameter6	Parameter7	Parameter8	Parameter9	Parameter10
count	6000	6000	6000	6000	6000	6000	6000	6000	6000	6000
mean	4.85E+02	1.95E+03	4.06E+02	1.89E+02	5.87E+00	5.60E+00	2.72E+02	2.21E+01	5.82E+04	3.59E+01
std	1.14E+04	5.42E+04	1.20E+04	2.62E+03	8.63E+00	7.69E+00	2.03E+03	3.02E+02	3.17E+06	1.02E+02
min	2.98E-06	1.68E-07	5.04E-06	1.62E-06	9.99E-07	9.81E-07	6.24E-04	2.37E-02	3.96E-04	1.02E-03
25%	0.09	0.05	0.10	0.12	0.74	0.38	0.15	0.04	0.59	0.10
50%	1.05	0.94	1.04	1.07	2.23	2.68	0.60	2.93	0.59	0.73
75%	10.94	15.69	10.67	9.90	6.68	7.12	2.37	17.85	6.78	13.97
max	7.14E+05	3.03E+06	8.65E+05	1.40E+05	8.37E+01	4.12E+01	3.57E+04	2.01E+04	1.74E+08	5.17E+02

Table A2. Attribute 数值特征

	Attribute1	Attribute2	Attribute3	Attribute4	Attribute5	Attribute6	Attribute7	Attribute8	Attribute9	Attribute10
count	6000	6000	6000	6000	6000	6000	6000	6000	6000	6000
mean	5.35E+04	8.84E+04	2.03E+02	6.27E+03	1.29E+06	3.40E+03	3.17E+04	4.55E+05	7.86E+05	1.99E+07
std	1.89E+06	2.85E+06	2.75E+03	2.59E+05	9.85E+07	2.00E+05	1.12E+06	2.85E+07	5.89E+07	1.53E+09
min	6.46E-10	5.49E-09	1.12E-06	4.34E-06	1.64E-06	5.46E-05	4.41E-03	3.01E-03	2.37E-03	1.97E-03
25%	0.03	0.03	0.10	0.13	0.12	0.14	0.11	0.11	0.13	0.13
50%	0.94	1.03	0.95	0.83	0.75	0.99	0.59	0.59	0.66	0.67
75%	27.85	30.01	9.80	6.05	5.74	7.42	4.64	4.77	5.57	4.91
max	1.20E+08	1.62E+08	1.28E+05	1.57E+07	7.63E+09	1.49E+07	7.79E+07	2.19E+09	4.56E+09	1.18E+11