💡 ChangTianML简介

逸思长天致力于人工智能基础软件的开发，在自动化机器学习领域，长期关注持续学习背景下的高阶自动化特征工程、自动化机器学习方法，以在规律不断变化的生产环境上对机器学习建模空间高效探索、高效更新模型、构建持续学习平台，从而极大降低机器学习使用门槛、降低机器学习场景探索与落地成本、极大提升建模效率。让用户在0成本情况下，搭建机器学习建模团队，积极探索机器学习场景。加速智能化创新业务的探索与落地。当前已在制造、金融、通信、农业、水利、互联网等多个场景应用落地。

当前，数字化转型与数字经济是现今各行业发展的共同诉求，亦成为国家发展重要规划，也是全球产业关注的焦点之一，中央发布的十四五规划中明确提出要推动产业数字化发展，据IDC预测，到2027年全球数字化转型支出将达到近3.9万亿美元。数据作为新型的生产要素，会有更多挖掘形成数据资产的需求。而利用机器学习人工智能技术，就可以挖掘数据规律，帮助各个行业形成数据资产。

并且近年来，各行各业的科研工作当中，涌现出大量的机器学习研究成果，比如：分子动力学、气象运动、流体结构、智能制造，甚至是金融市场投资，这些场景都可以利用机器学习进行建模。这类被称为**AI for Science**的技术应用，不同于感知数据识别和AIGC的生成，它能够帮助我们发掘更多自然科学和社会科学的规律，从而助力和推动产业数字化的发展。

但是AI for Science的落地之路也不是一帆风顺的，机器学习的传统建模方式非常依赖专家经验去进行手工建模的，这会有技术门槛高、人才成本高、建模周期长等等的问题。长天ML自动化建模技术平台的诞生就是为了解决这些痛点，通过自研的自动化特征工程、自动化机器学习、自动持续学习等核心技术，可以面向一个具体的行业问题的数据集，全自动化的去构建对应的机器学习模型，实现了零门槛操作、无学习成本，并且大大提升了建模效率。

本平台支持在线按月订阅使用，网址为：

持续学习平台，持续探索创新，逸思长天旗下全自动持续学习工具ChangTianML ›

https://www.changtianml.com/

本平台也支持本地化私有化部署，提供更丰富的持续学习能力的同时，保护您和企业的数字资产安全。逸思长天的官方内容发布渠道（B站账号：逸思长天）同步持续发布机器学习入门知识、AI4science等行业观察、机器学习研究进展等硬核内容，目前已有5000+粉丝和10w+播放，欢迎关注并联系我们（微信：freedoooo）加入机器学习社群。

下一页
ChangTianML核心特性

最后更新于3天前

✨ ChangTianML核心特性

ChangTianML致力于面向AI for Science的场景的技术应用，帮助用户以低门槛的形式发掘更多自然科学和社会科学的规律，助力和推动产业数字化的发展。

产品核心功能

本平台在产品侧主要有以下核心功能：

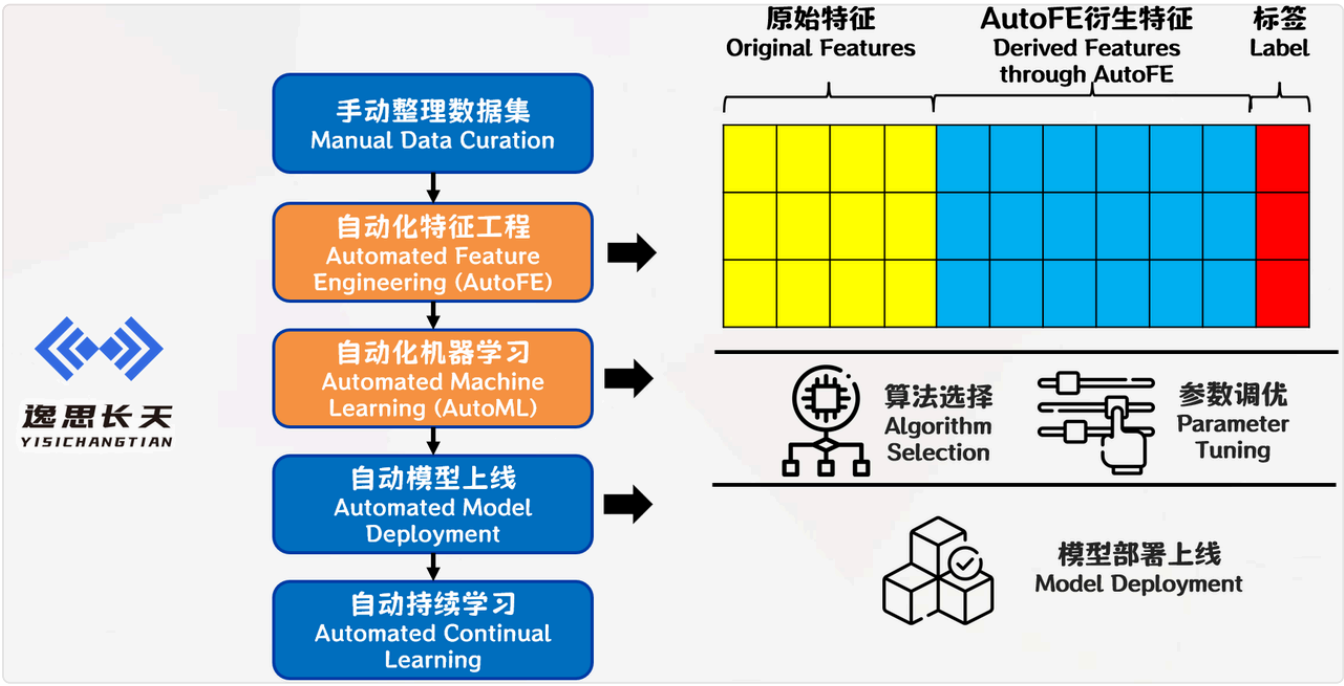
① 自动化机器学习建模；

② 自动探索构建高级特征组；

③ 自动化时间序列机器学习任务构建；

④ 建模报告自动生成导出。

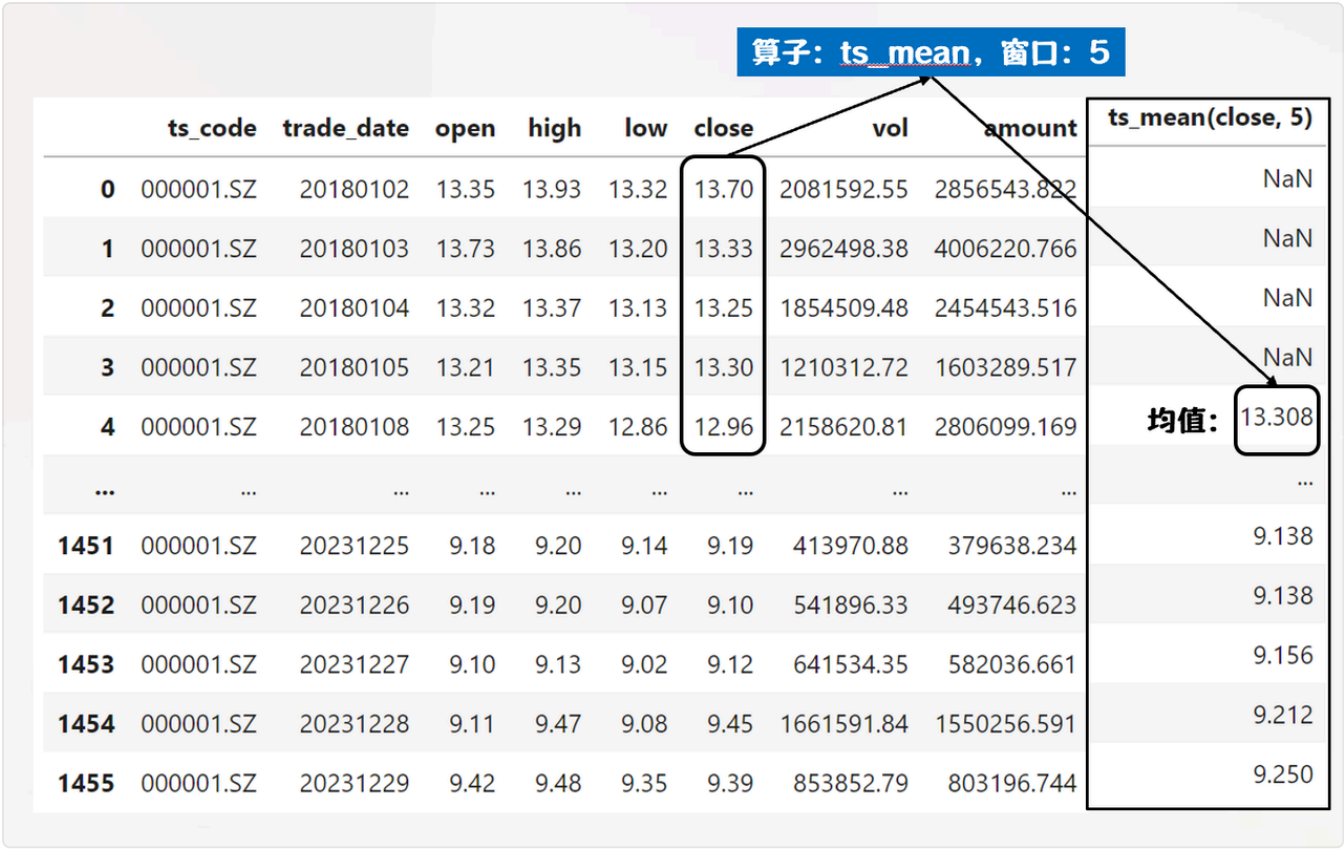
本平台支持以无代码自动化的形式进行机器学习建模，用户只需要上传定义好建模问题的数据集，然后设置待建模的特征标签以及时间预算，ChangtianML将自动在后台进行AutoFE（自动化特征工程）、AutoML（自动化模型选择和超参调优），自动构建出可解释性较强的高级特征组和性能优异的机器学习模型。整个建模过程中不需要写一行代码，大大降低了建模门槛，提高了建模效率。



本平台的AutoFE定义了一套标准通用的建模算子，通过ChangtianML自研的自动化特征工程算法，能够更高效地探索特征空间，同时最优化每个特征表征的方向，从而确定共同增益最大的特征组。这些高级特征有显式的公式展示，有更好的可解释性，给用户提供了解释数据的新角度新方法。

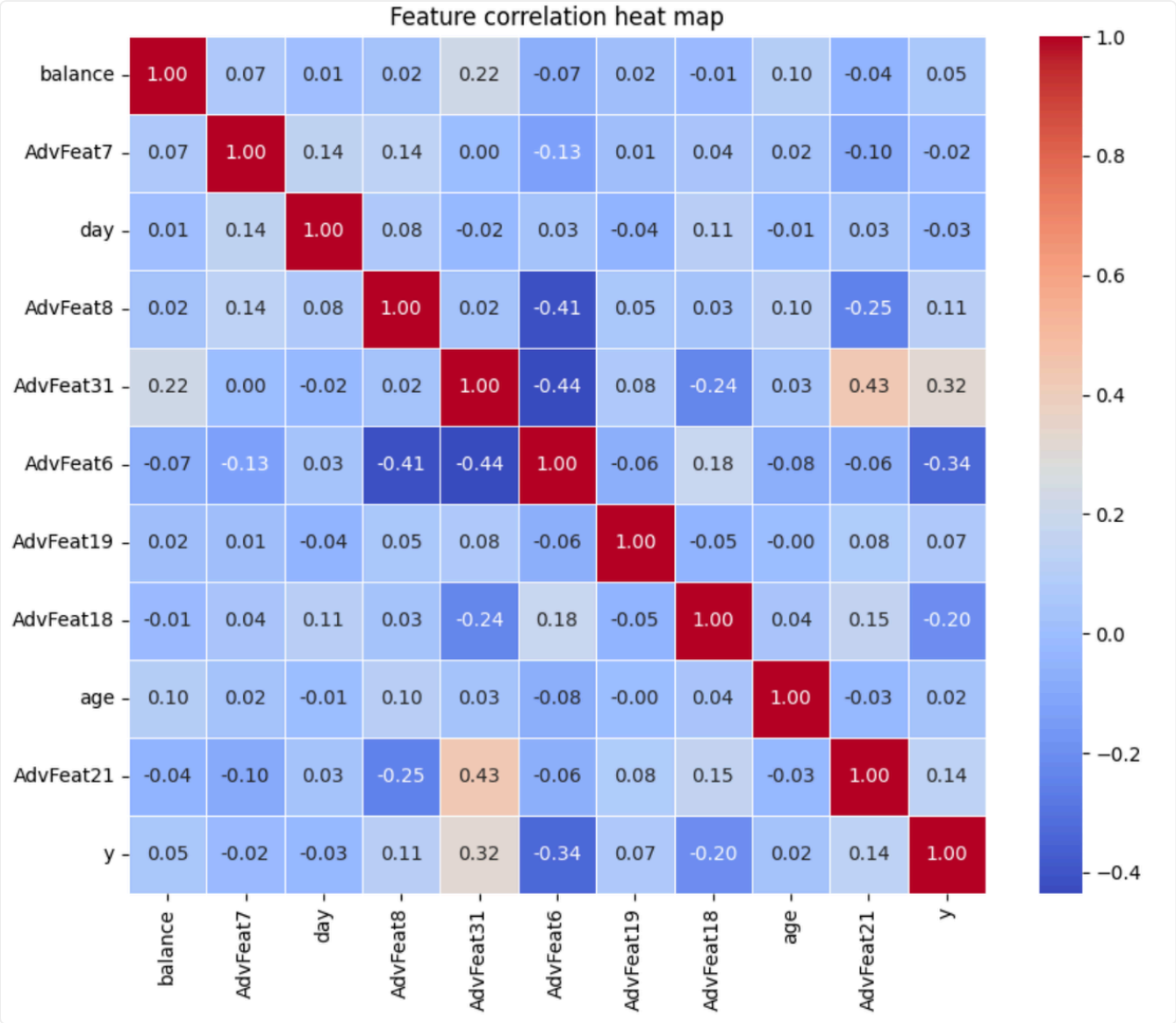
此外，本平台还支持自动化时间序列建模。时间序列数据的规律也存在在时间模式趋势和周期性中，那么在建模时就不仅仅要用当前时刻的特征信息来预测Y，更要回望当前时刻过去的特征信息，但是如何回望？回望多少？怎样挖掘出隐藏在时序中的高级特征，这些问题都显著增加了建模探索空间，更有必要使用自动化手段高效探索。

本平台将自动化机器学习方法扩展到时序机器学习领域，此外还创新性的内置了时序算子，在用户根据实际业务设置窗口长度后，会自动化的搜索构建时间序列高级特征，在有效提取过去信息的同时，还保证了特征的可解释性。



	ts_code	trade_date	open	high	low	close	vol	amount	ts_mean(close, 5)
0	000001.SZ	20180102	13.35	13.93	13.32	13.70	2081592.55	2856543.822	NaN
1	000001.SZ	20180103	13.73	13.86	13.20	13.33	2962498.38	4006220.766	NaN
2	000001.SZ	20180104	13.32	13.37	13.13	13.25	1854509.48	2454543.516	NaN
3	000001.SZ	20180105	13.21	13.35	13.15	13.30	1210312.72	1603289.517	NaN
4	000001.SZ	20180108	13.25	13.29	12.86	12.96	2158620.81	2806099.169	13.308
...	...	...	...	...	...	...	...	...	...
1451	000001.SZ	20231225	9.18	9.20	9.14	9.19	413970.88	379638.234	9.138
1452	000001.SZ	20231226	9.19	9.20	9.07	9.10	541896.33	493746.623	9.138
1453	000001.SZ	20231227	9.10	9.13	9.02	9.12	641534.35	582036.661	9.156
1454	000001.SZ	20231228	9.11	9.47	9.08	9.45	1661591.84	1550256.591	9.212
1455	000001.SZ	20231229	9.42	9.48	9.35	9.39	853852.79	803196.744	9.250

建模报告会记录建模人员的工作步骤、模型选择、特征工程、模型评估等关键步骤和结果，无论在科研还是工作中，都是建模人员的重要产物。本平台会在用户建模结束后自动生成建模示例报告，包含但不限于相关系数热力图分析、单特征分布、模型效果可视化分析等等。还提供了生成报告图表的API代码，供用户下载模型后快速完善报告内容。



技术核心能力

技术核心能力方面，逸思长天团队在自动化特征工程、持续学习样本增强领域不断探索，基于学界的最新成果，进行了进一步的研究，构建了自研的强化学习和演化计算框架。 这些技术使得本平台在经典案例中的表现，相对于开源的或者经典的工具，能够提高30%以上的精度。这使得普通的数据处理人员，无需具备机器学习建模专家的能力，就可以去使用我们的工具构建出不输于甚至远超专家建模表现的机器学习模型。建模效率，也得以从几十人天甚至几百人天，变到分钟级甚至秒级，大大节省人力成本。

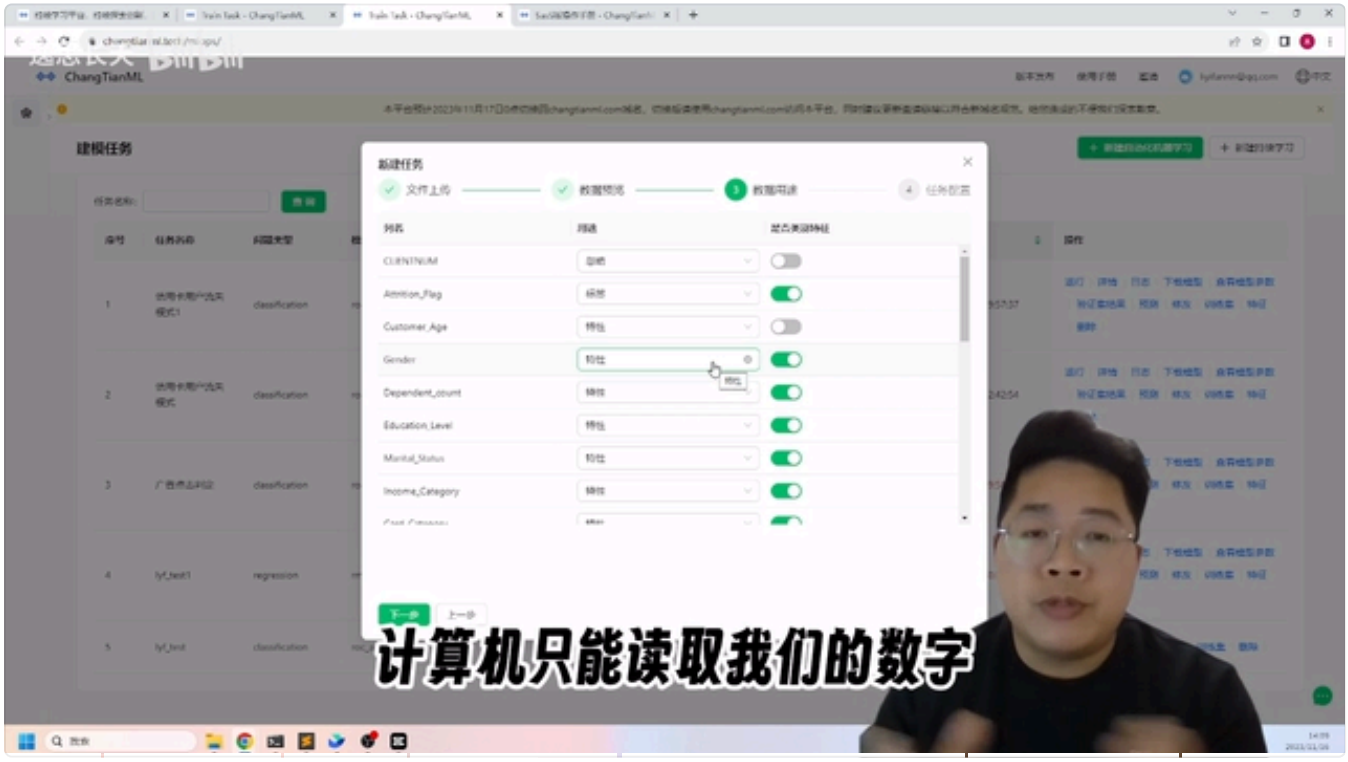
上一页
ChangTianML简介

下一页
SasS版操作手册

基本介绍

本工具实现了基于**持续学习**（Continental Machine Learning）的机器学习模型演化与持续服务的过程。在学界，持续学习指根据数据的变化自动进行机器学习的过程，并且解决数据变化带来的概念漂移、灾难性遗忘等问题。该过程要求机器学习过程中使用的特征工程方法、建模方法与参数调优策略可以被自动化执行，同时可以根据数据特征自动探索到表现更好的特征、模型与参数。这种自动化特性使得该工具可以作为快速验证机器学习场景可行性的工具。

目前已经基于本工具提供了SaaS版平台服务，您可以观看以下视频学习如何创建你的第一个机器学习模型！



ChangTianML基础版教学

如果您已经具备了一些机器学习相关的知识，本平台也提供如下进阶版视频让您体会更多新功能。

ChangTianML进阶版教学

使用方法

一.准备数据

为了用户方便理解，当前版本支持输入csv格式的数据文件，并且要求数据列分隔符为逗号，同时每个csv数据文件需要包含更新数据的表头。

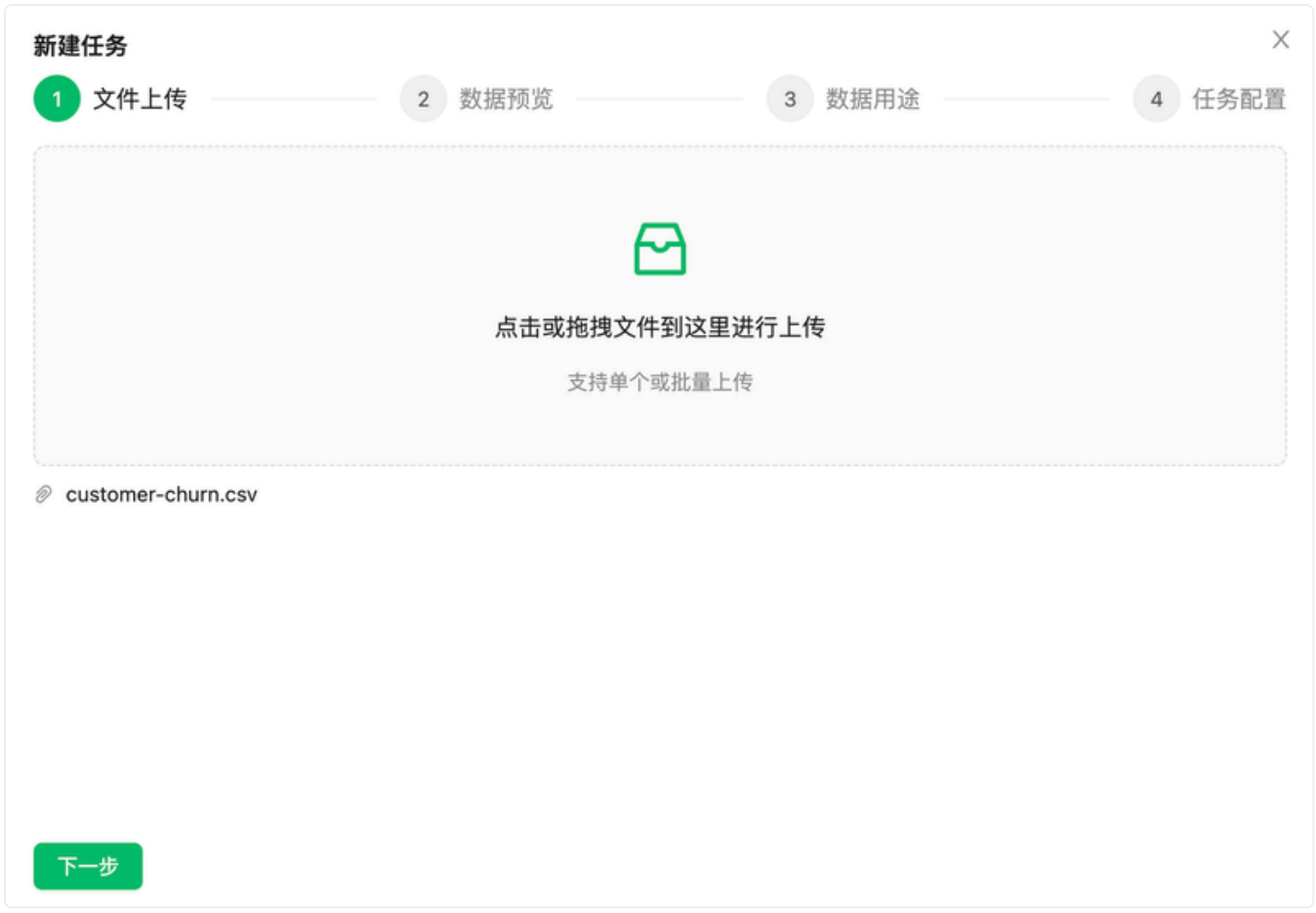
> 在正式版中有多种数据输入形式，可以搭配特有数据存储增强数据吞吐性能。 当前平台为免费用户开放了数据表形式的结构化数据输入方式。数据表形式的序列、以及更多数据形式会在后续版本中陆续开放。

二.任务配置

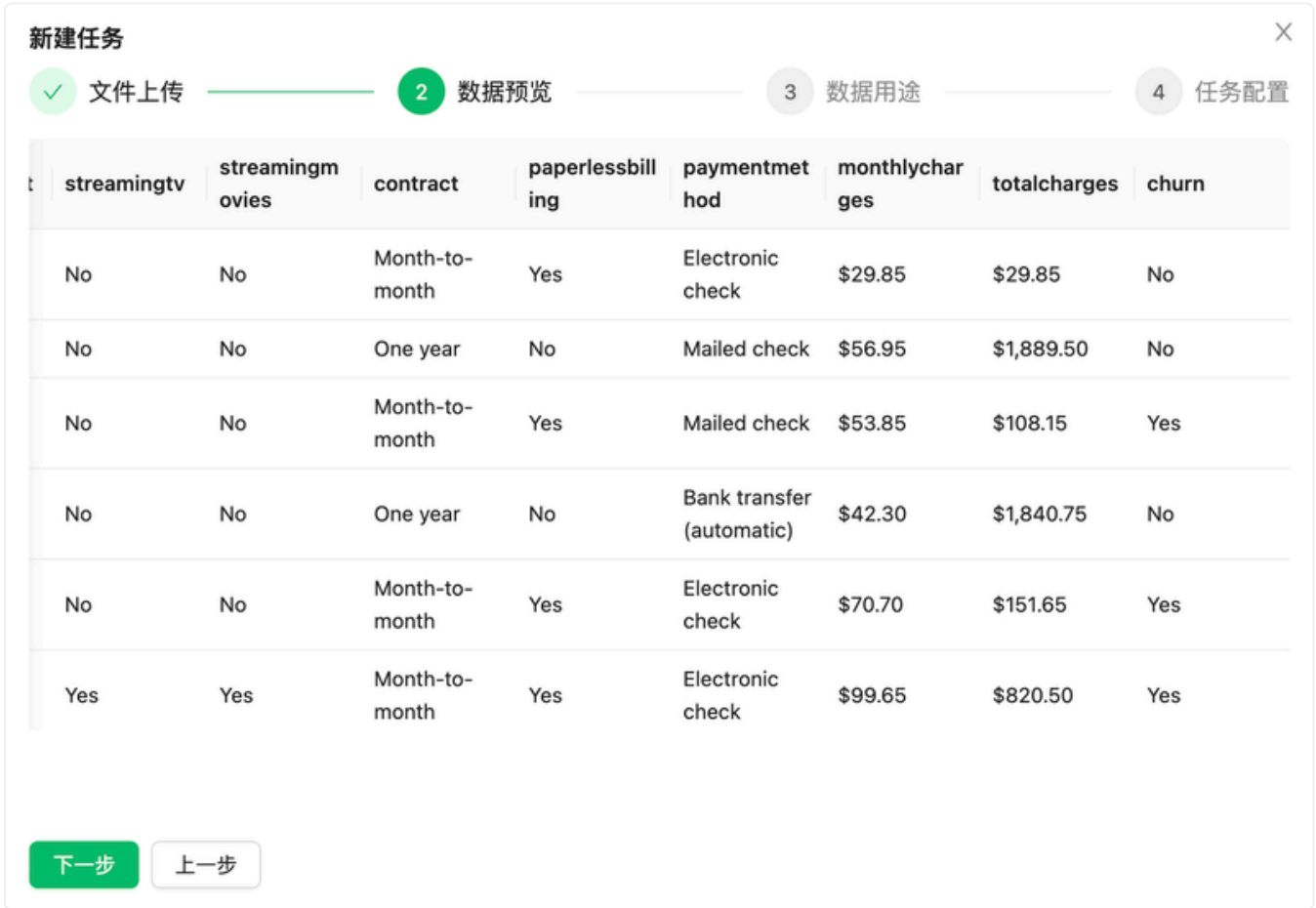
登录平台后，点击页面右上方【新建自动化机器学习】按钮，弹出【新建任务】对话框。

首先将准备好的数据文件上传到平台，这里可以上传多个数据文件，并且它们的数据格式与数据字段必须相同。上传完毕之后点击【下一步】。

当前上传数据文件类型仅支持CSV格式，如果上传的文件类型是Excel表格，需另存为CSV格式，并且推荐采用不含BOM的UTF-8字符编码保存。操作方法可到常见问题-使用相关查看。



点击【下一步】，可以预览所上传的数据。



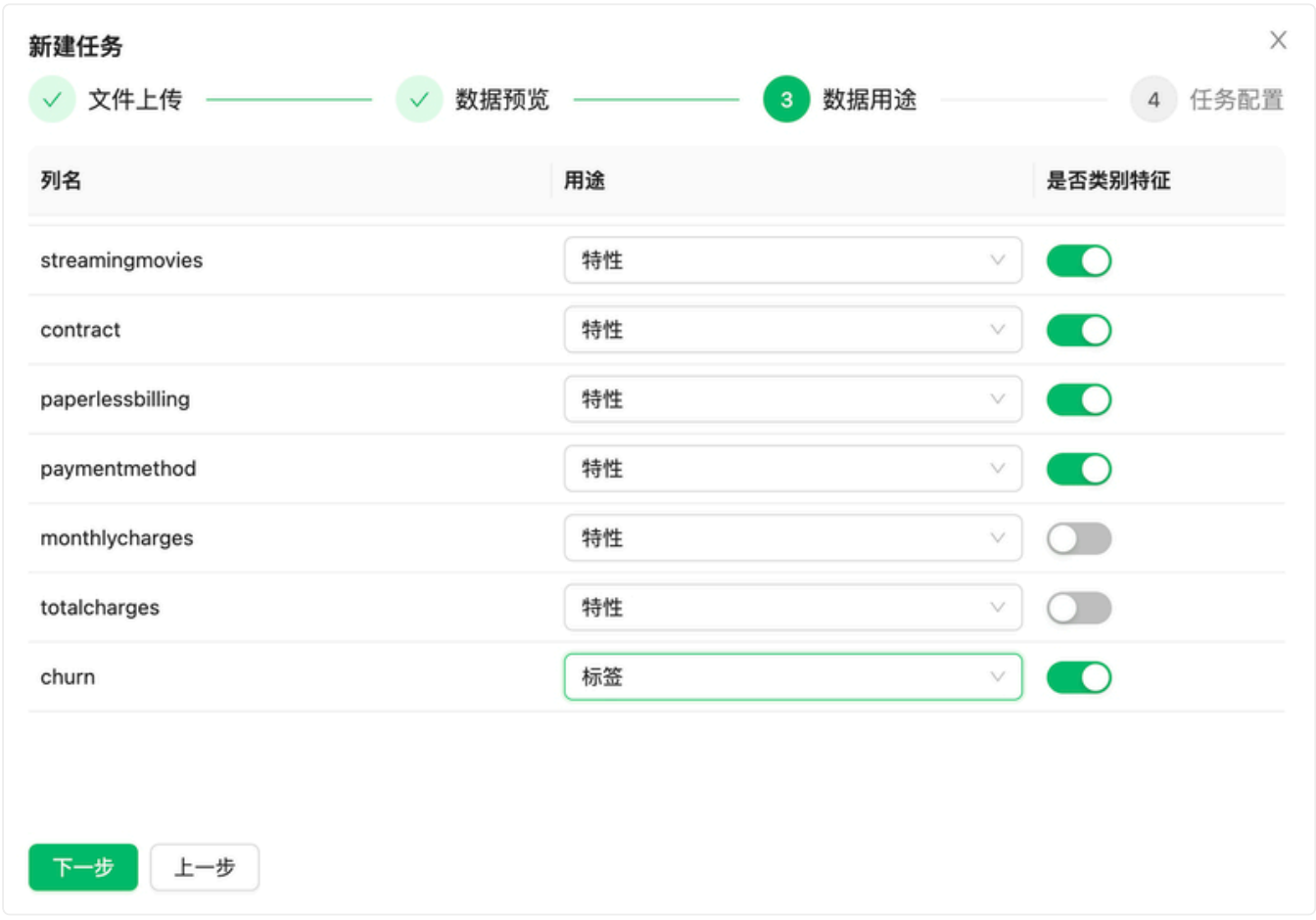
streamingtv	streamingmovies	contract	paperlessbilling	paymentmethod	monthlycharges	totalcharges	churn
No	No	Month-to-month	Yes	Electronic check	$29.85	$29.85	No
No	No	One year	No	Mailed check	$56.95	$1,889.50	No
No	No	Month-to-month	Yes	Mailed check	$53.85	$108.15	Yes
No	No	One year	No	Bank transfer (automatic)	$42.30	$1,840.75	No
No	No	Month-to-month	Yes	Electronic check	$70.70	$151.65	Yes
Yes	Yes	Month-to-month	Yes	Electronic check	$99.65	$820.50	Yes

继续点击【下一步】，平台会自动解析文件中每个字段的类型。此时需要用户确认如下内容：

⊖ 确认该数据集中需要拟合的字段，在对应字段下拉列表中选择“标签”，默认情况为“特征”。

⊖ 确认该数据集中不需要用到的字段，在对应字段下拉列表选择“忽略”。

⊖ 确认标签和特征所对应的字段中属于非数值的类别特征字段，并将类别特征的字段设置为开启状态。置为开启状态。



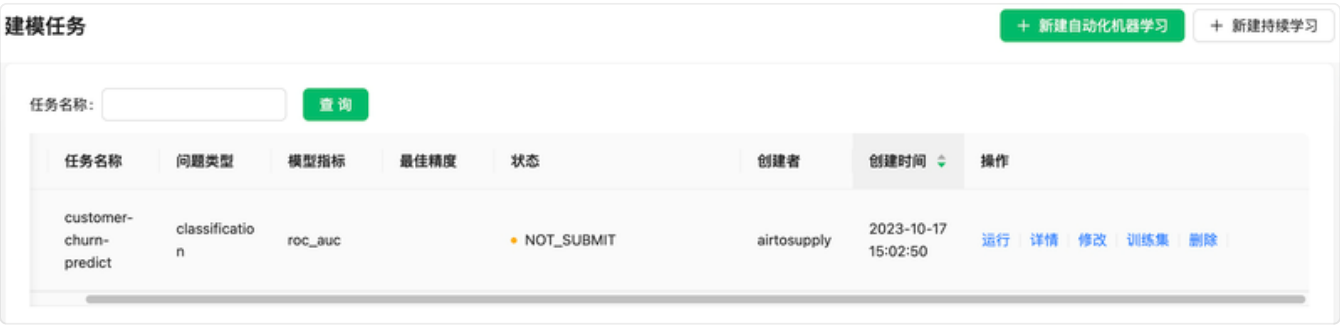
该案例中选择预测字段“churn”。点击【下一步】设置任务名称、自动化特征工程尝试次数、自动化机器学习尝试时间，备注等信息，然后点击【确定】则任务创建完毕。

【特征-问题类型】下拉选项中提供classification（分类）和regression（回归）两种机器学习任务类型的选择。如果标签为类别特征则是分类问题；如果标签为数值则是回归问题。本案例中选择“classification”。

【特征-指标】下拉选项中提供了多种机器学习评判指标。如果当前解决的是分类问题则选择“roc_auc”和“log_loss”作为评判指标；如果当前解决的是回归问题则选择“rmse”作为判断指标。

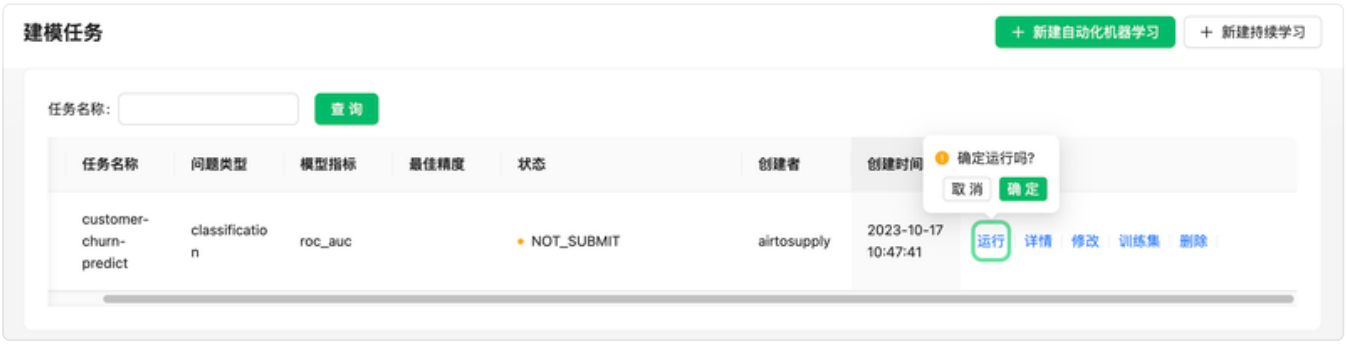
> 一般而言，当log_loss取值小于0.693即表示模型有拟合能力，且越小越好；当roc_auc取值范围为0.5～1，越大越好，当大于0.7即表示模型有良好分辨能力。

完成任务配置之后在【建模任务】列表中可以查看到所创建的任务。

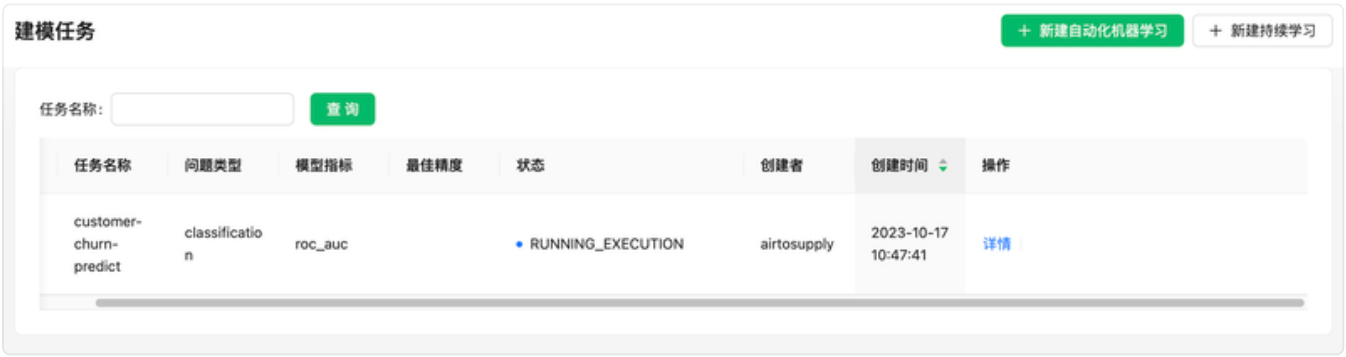


三.模型训练

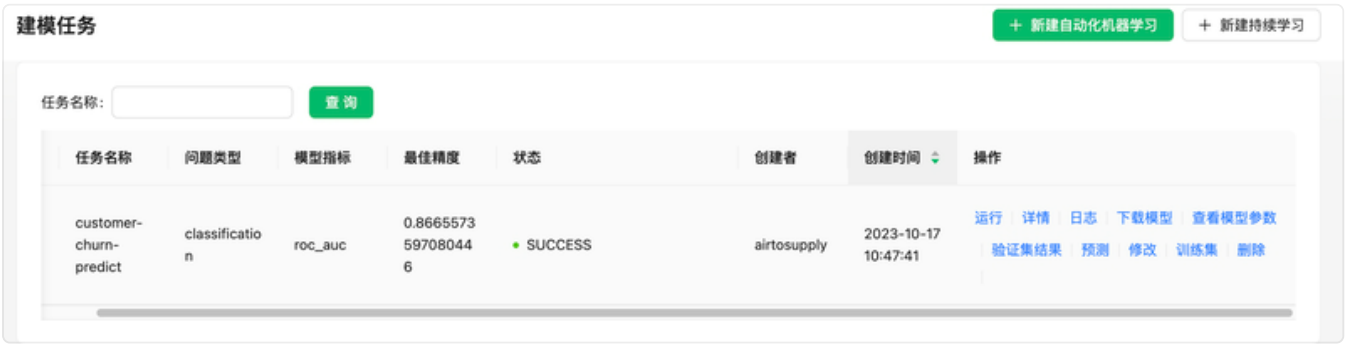
在【建模任务】列表选择创建好的任务点击【运行】确定即可，整个建模过程是全自动化的。



任务状态为“RUNNING_EXECUTION”表示训练任务正在运行。

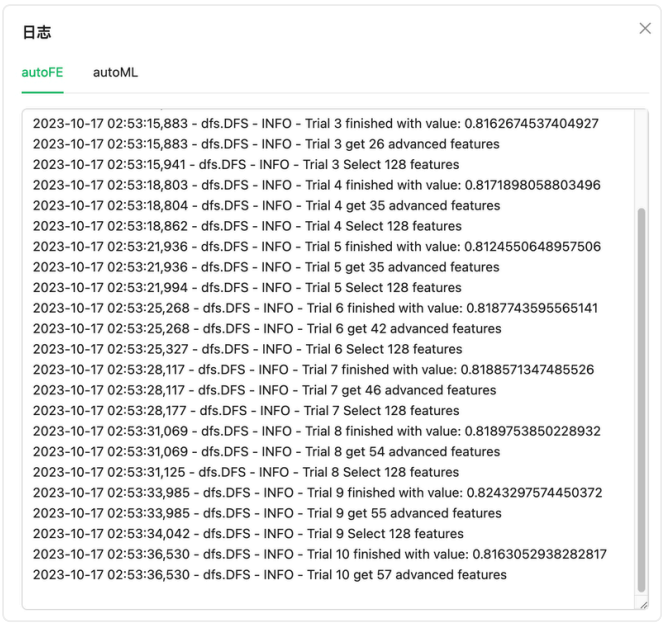


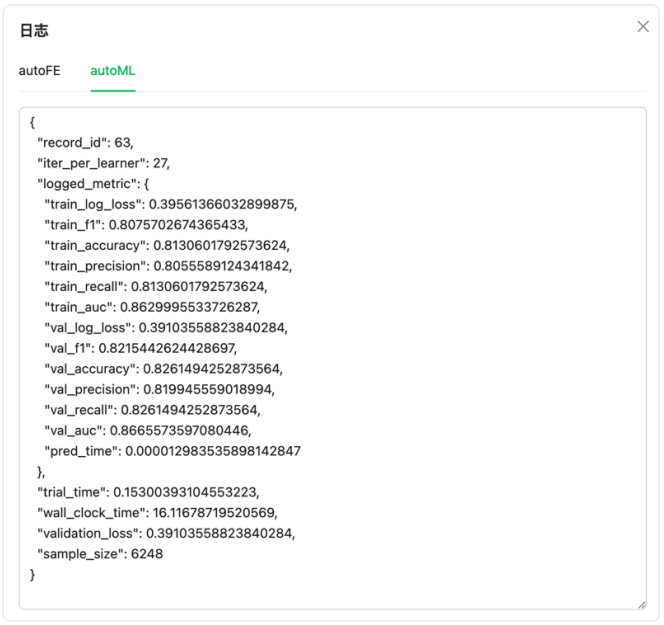
任务运行完毕之后会显示当前模型训练的最佳精度，本次训练的最佳精度大约为0.86，并且任务状态为“SUCCESS”，表示训练任务成功。



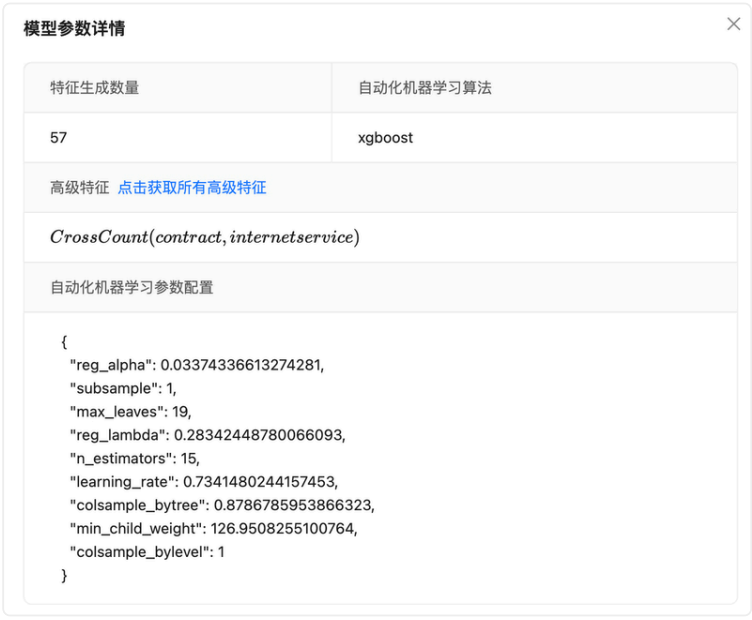
四.模型查看

在【建模任务】列表选择对应任务点击【日志】按钮可以查看相关整个持续学习过程中的自动化特征工程和自动化机器学习的相关细节。





用户可以通过这些日志更好的了解机器学习模型演化与持续服务的过程。也可以点击【查看模型参数】查看模型相关参数。



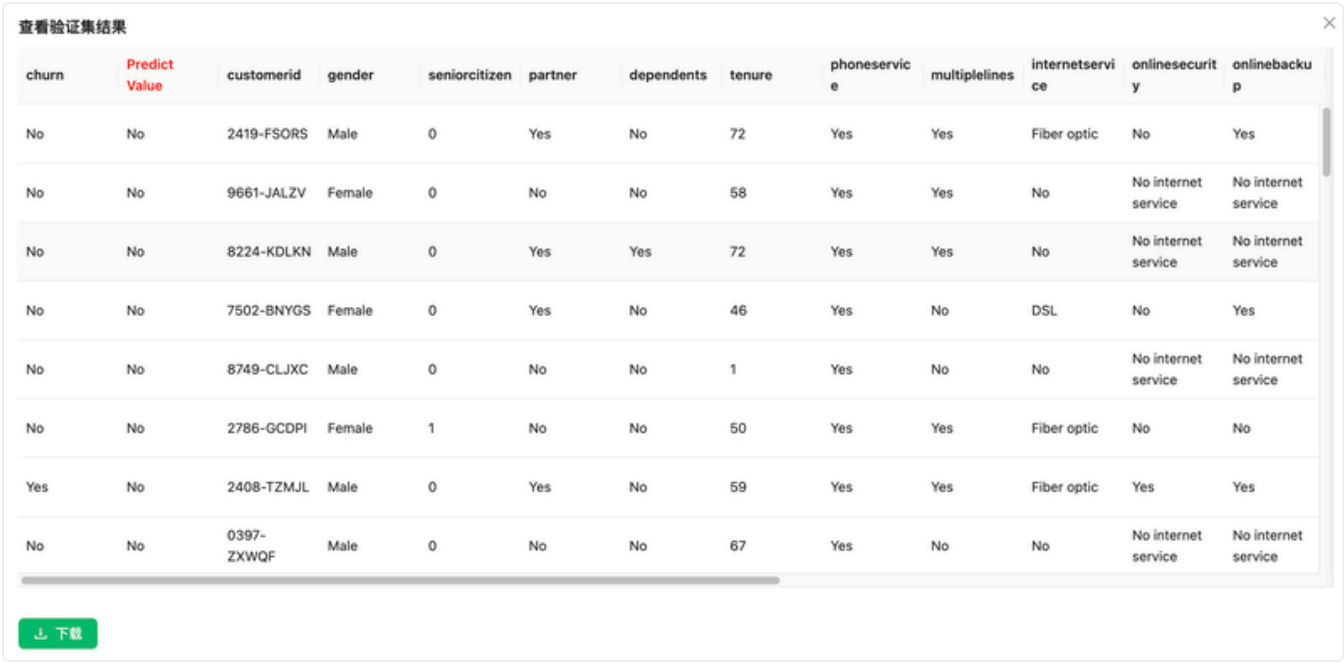
特征生成数量表示在自动化特征工程过程中探索到的特征数量，本次实验探索到57个特征。

高级特征表示在自动化特征工程过程中探索到的组合特征（高级特征），并且采用Latex风格的语法表达。

自动化机器学习算法表示在自动化机器学习过程中所采用的算法，本次实验采用的算法是：xgboost。

自动化机器学习参数配置表示在机器学习自动化调参过程中，自动化机器学习算法的最优参数，例如：xgboost学习率等超参数。

用户可以点击【验证集结果】查看该模型对上传数据文件中的数据所预测的结果。

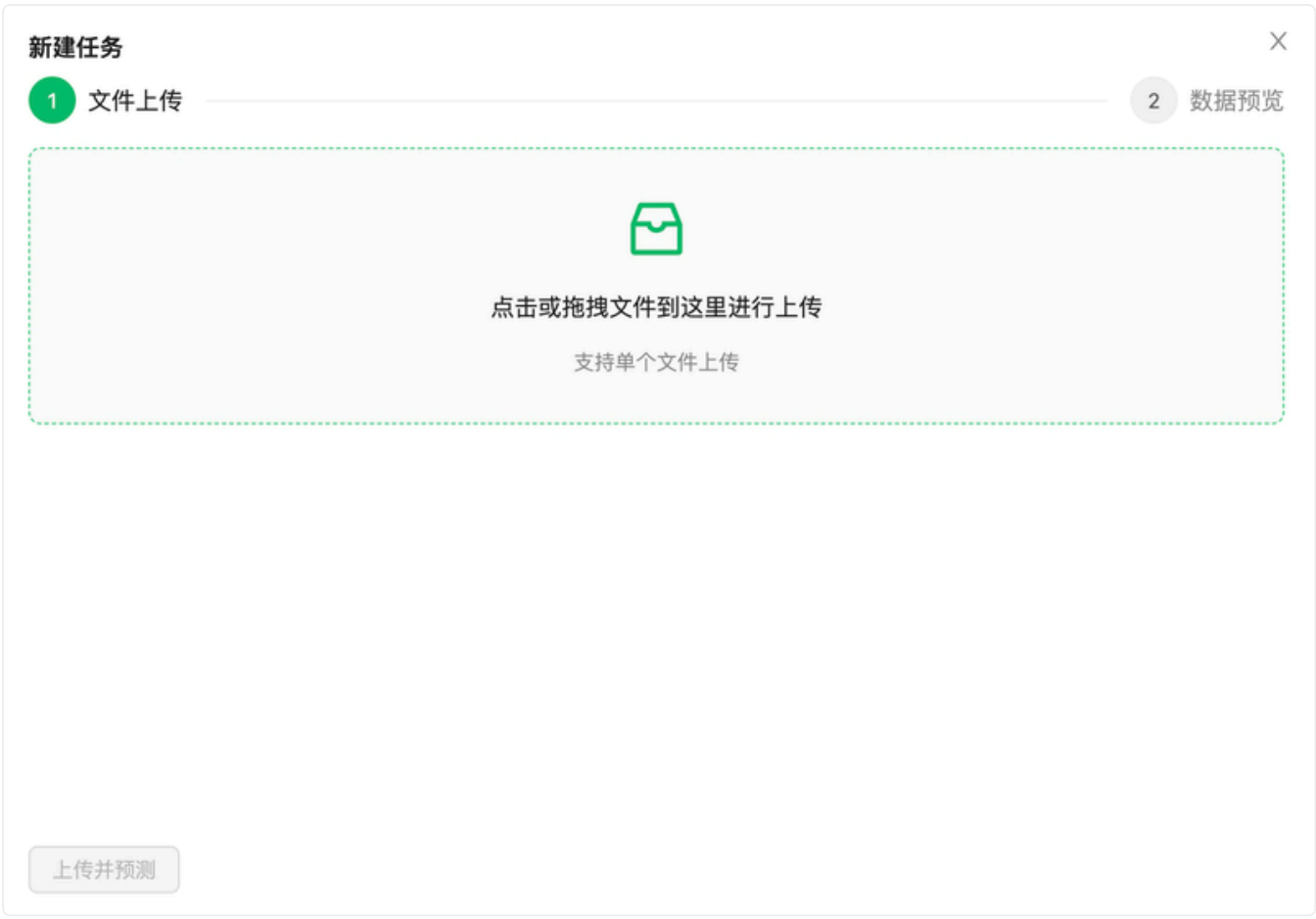


churn	Predict Value	customerid	gender	seniorcitizen	partner	dependents	tenure	phoneservice	multiplelines	internetservice	onlinesecurity	onlinebackup
No	No	2419-FSORS	Male	0	Yes	No	72	Yes	Yes	Fiber optic	No	Yes
No	No	9661-JALZV	Female	0	No	No	58	Yes	Yes	No	No internet service	No internet service
No	No	8224-KDLKN	Male	0	Yes	Yes	72	Yes	Yes	No	No internet service	No internet service
No	No	7502-BNYGS	Female	0	Yes	No	46	Yes	No	DSL	No	Yes
No	No	8749-CLJXC	Male	0	No	No	1	Yes	No	No	No internet service	No internet service
No	No	2786-GCDPI	Female	1	No	No	50	Yes	Yes	Fiber optic	No	No
Yes	No	2408-TZMJL	Male	0	Yes	No	59	Yes	Yes	Fiber optic	Yes	Yes
No	No	0397-ZXWQF	Male	0	No	No	67	Yes	No	No	No internet service	No internet service

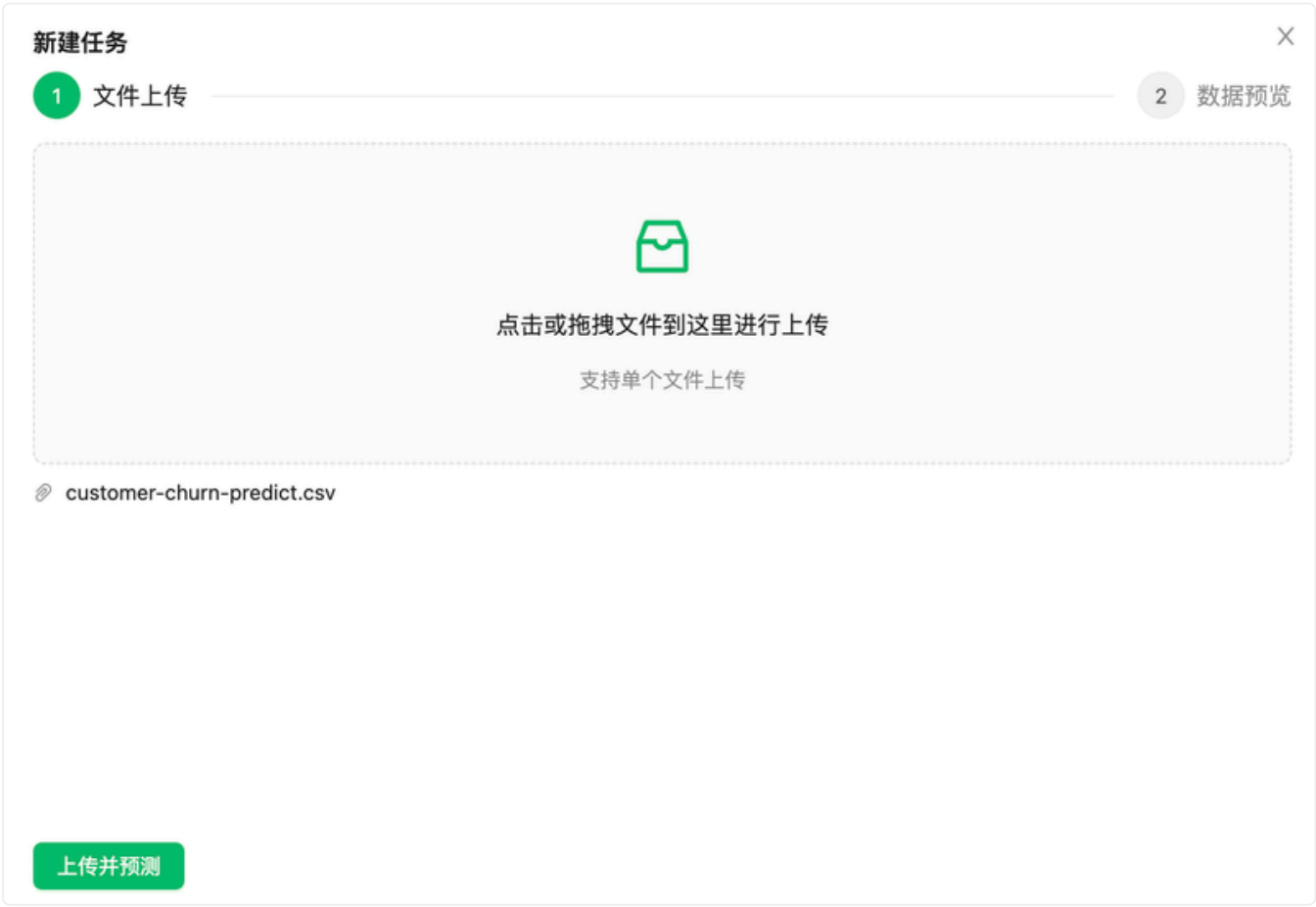
在本案例中模型所预测的字段是“churn”这一列，红色高亮标识的“Predict Value”这一列为模型预测的结果。

五.模型预测

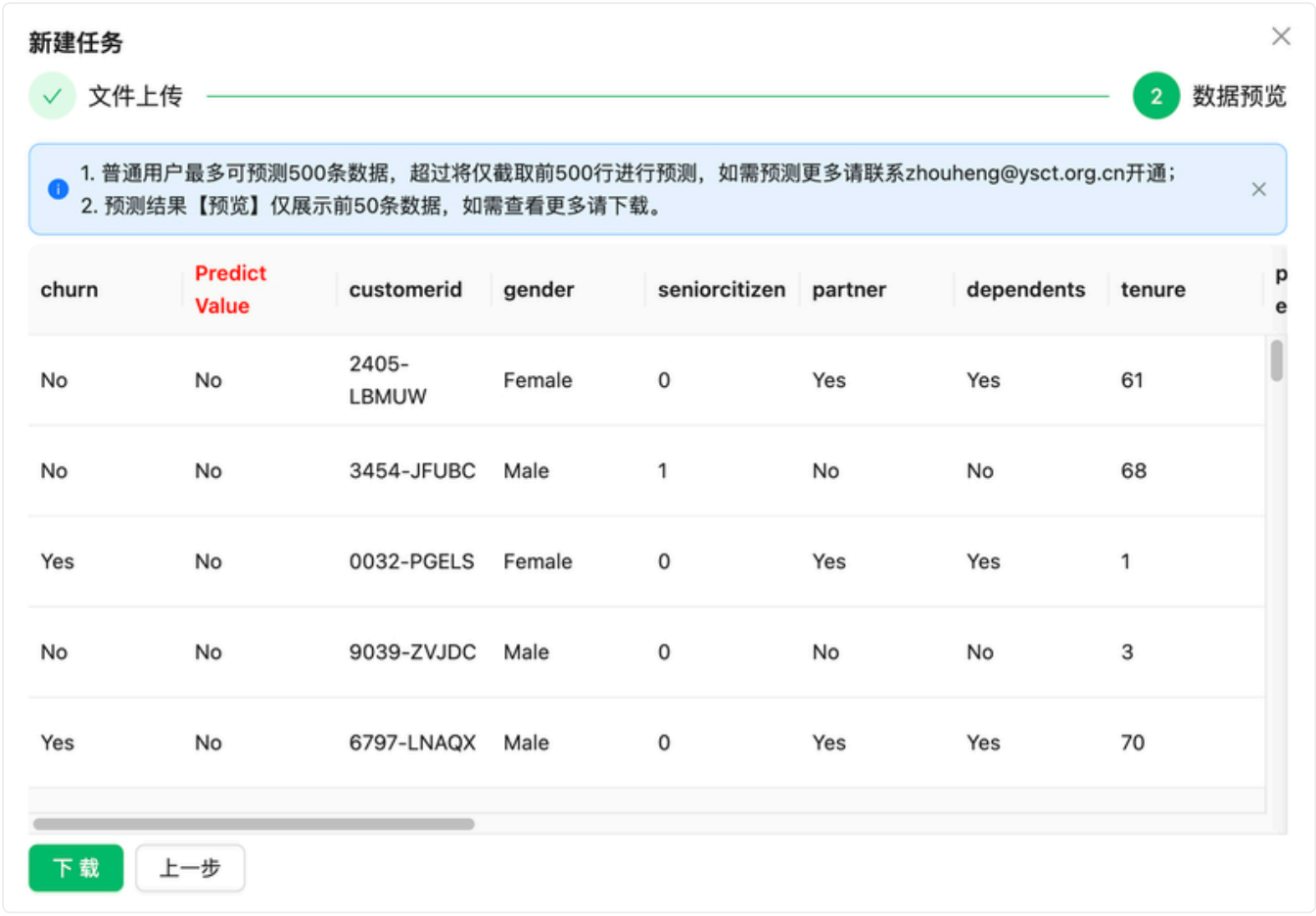
在【建模任务】列表选择对应任务点击【预测】可以进行模型预测。



此时可以将需要预测的数据文件上传到平台，这里对上传数据文件的要求和“数据准备”环节是数据文件要求是一致的。



点击【上传并预测】，稍等片刻在【数据预览】中所展示的数据则是通过训练好的模型所预测的结果。



红色高亮表示的“Predict Value”这一列为模型预测的结果，同时提供下载功能将预测结果下载到本地。

至此，通过该平台完成了一个分类任务自动化建模的整个过程的演示。同时该平台通过自动建模工具可以让用户都能够更加便捷，高效的建立机器学习模型并完成机器学习相关的工作。

上一页
ChangTianML核心特性

下一页
AutoFE算子手册

最后更新于4天前

📎 AutoFE算子手册

算子列表

基础算子

> f代表的是数字特征，c代表类别特征。

AggMin(f, c)：特征c各类别中f的最小值

AggMax(f, c)：特征c各类别中f的最大值

AggMean(f, c)：特征c各类别中f的平均值

AggMedian(f, c)：特征c各类别中f的中位数

AggVar(f, c)：特征c各类别中f的方差

CrossCount([c1, c2, ..])：根据特征list聚合的计数，list长度大于等于2

Nunique(c1, c2)：特征c2各类别中c1的唯一值计数

Entropy(c)：特征c各类别的熵

Percentile(f)：特征f各个数据的百分位

Combine(c1, c2)：特征c1和特征c2的字符结合

Count(c)：特征c各类别的计数

Equal(f1, f2)：判断特征f1和特征f2是否相等

Min(f1, f2)：取特征f1和特征f2相比的较小值

Max(f1, f2)：取特征f1和特征f2相比的较大值

Sigmoid(f)：对特征f进行sigmoid非线性变换

Round(f)：对特征f进行四舍五入

Residual(f)：保留特征f求小数点后的数

Softmax(f)：有限项离散概率分布的梯度对数归一化

时序算子

> **f**代表的是数字特征，**w**代表窗口数。

stddev(f, w)：计算窗口内特征f的标准差

ts_max(f, w)：计算窗口内特征f的最大值

ts_min(f, w)：计算窗口内特征f的最小值

ts_mean(f, w)：计算窗口内特征f的平均值

ts_sum(f, w)：计算窗口内特征f的加和值

ts_rank(f, w)：计算特征f当前值在在窗口内的排名（降序）

ts_argmax(f, w)：计算窗口内特征f最大值位置索引（从0计数）

ts_argmin(f, w)：计算窗口内特征f最小值位置索引（从0计数）

delay(f, w)：获取窗口内特征f最早时间所对应的值

decay(f, w)：计算窗口内特征f线性衰减和

delta(f, w)：计算窗口内特征f最晚和最早时间所对应值的差值

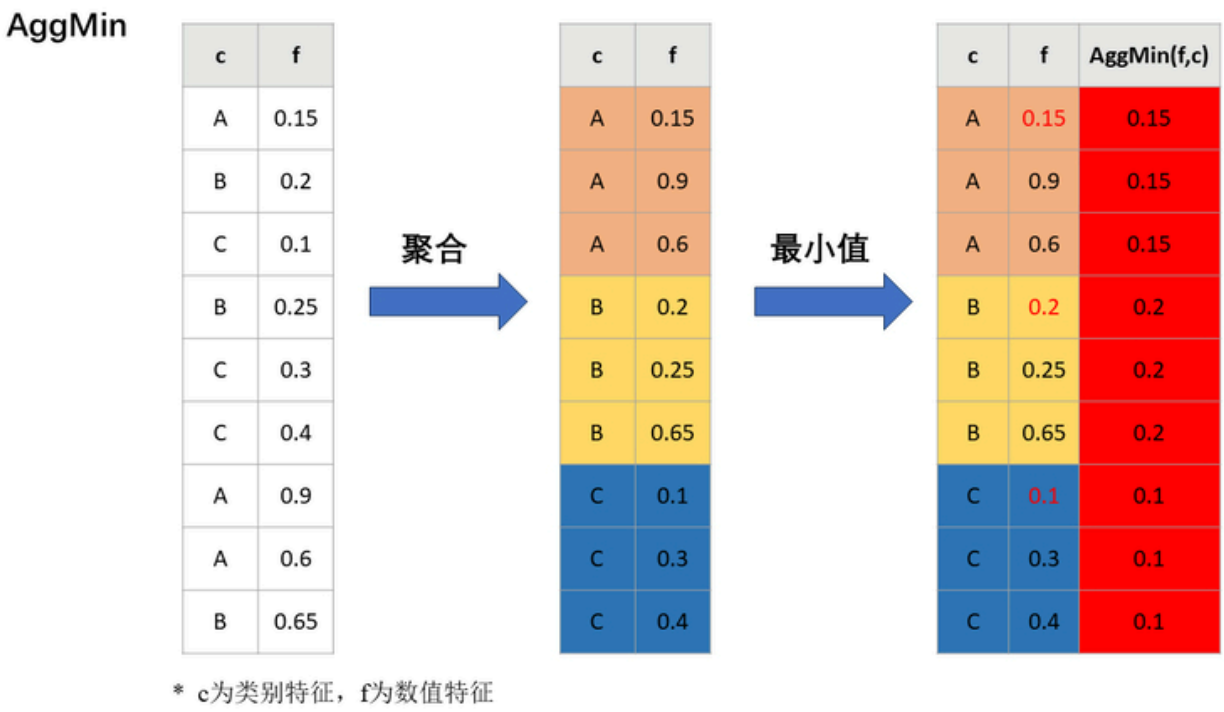
correlation(f1, f2, w)：计算窗口内特征f1和特征f2的相关系数

covariance(f1, f2, w)：计算窗口内特征f1和特征f2的协方差

算子图示

AggMin(f, c)

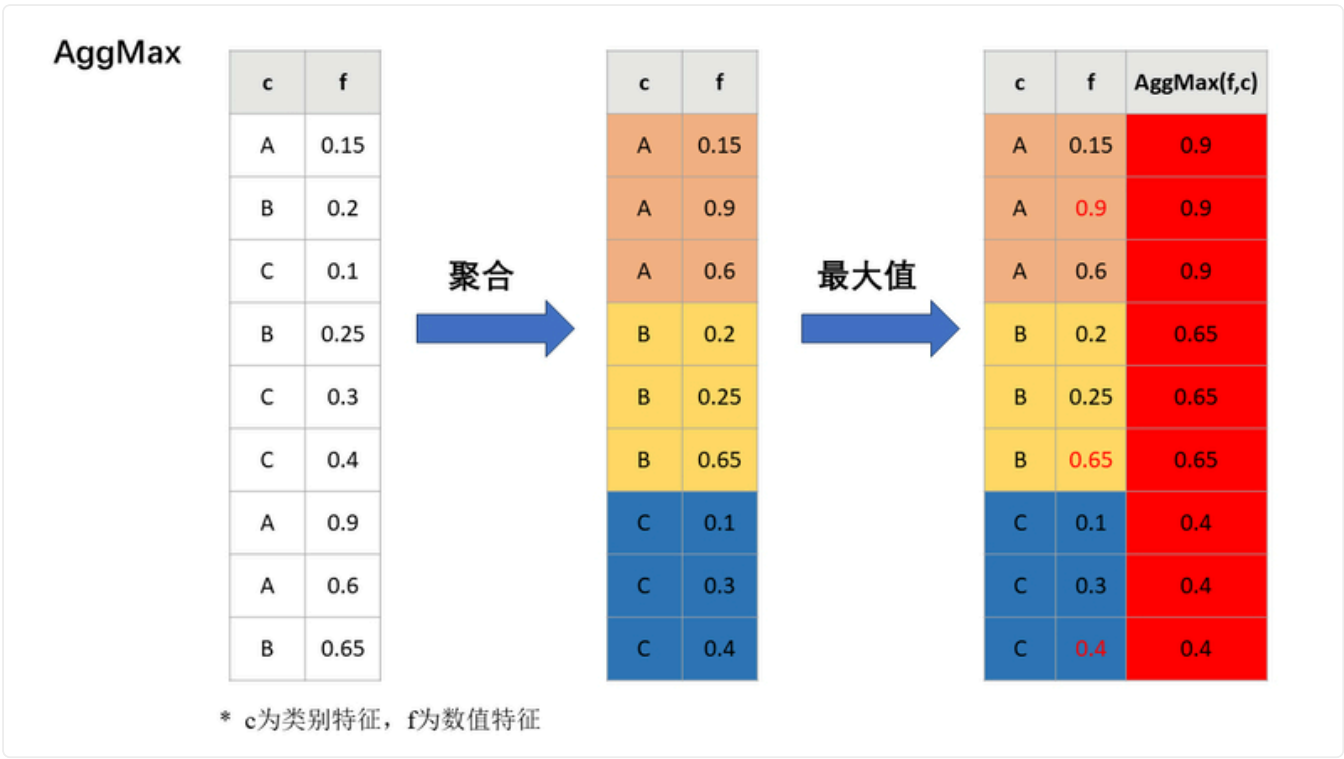
表示特征c各类别中f的最小值



AggMin

AggMax(f, c)

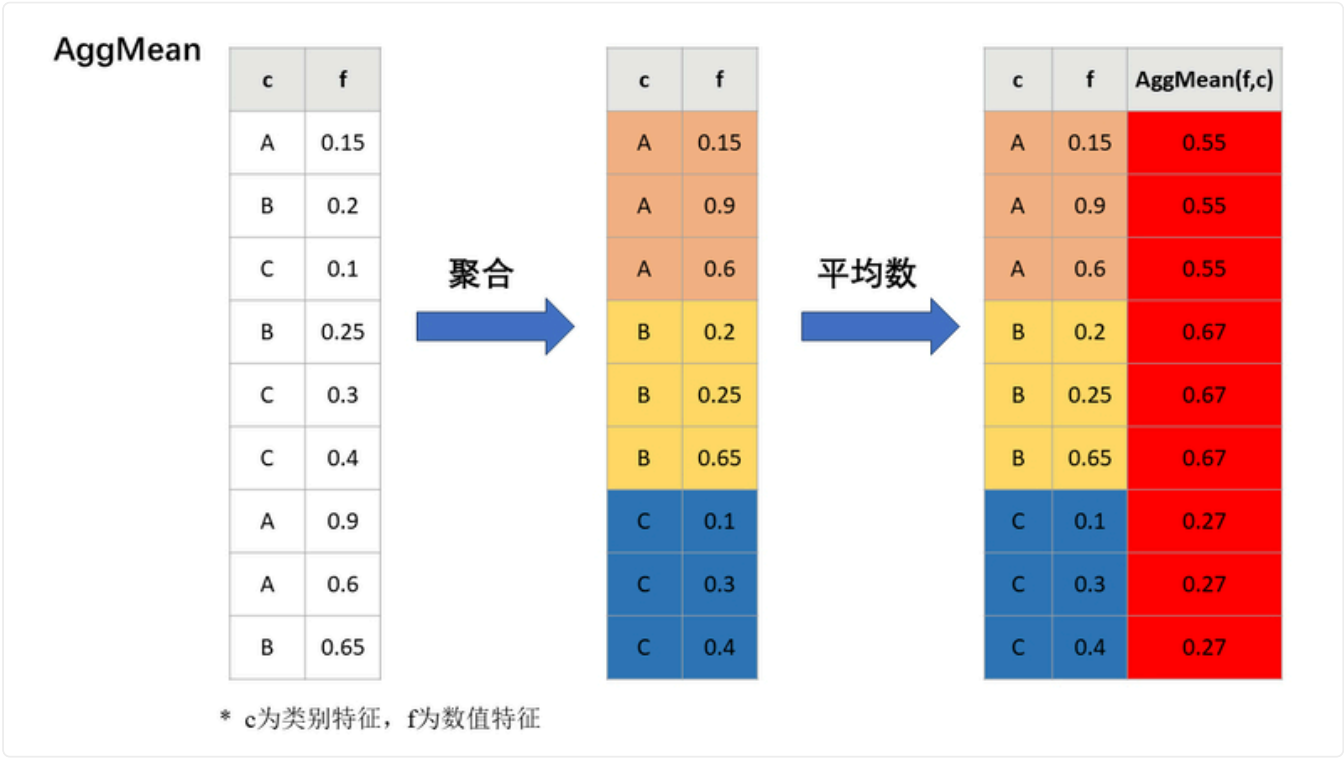
表示特征c各类别中f的最大值



AggMax

AggMean(f, c)

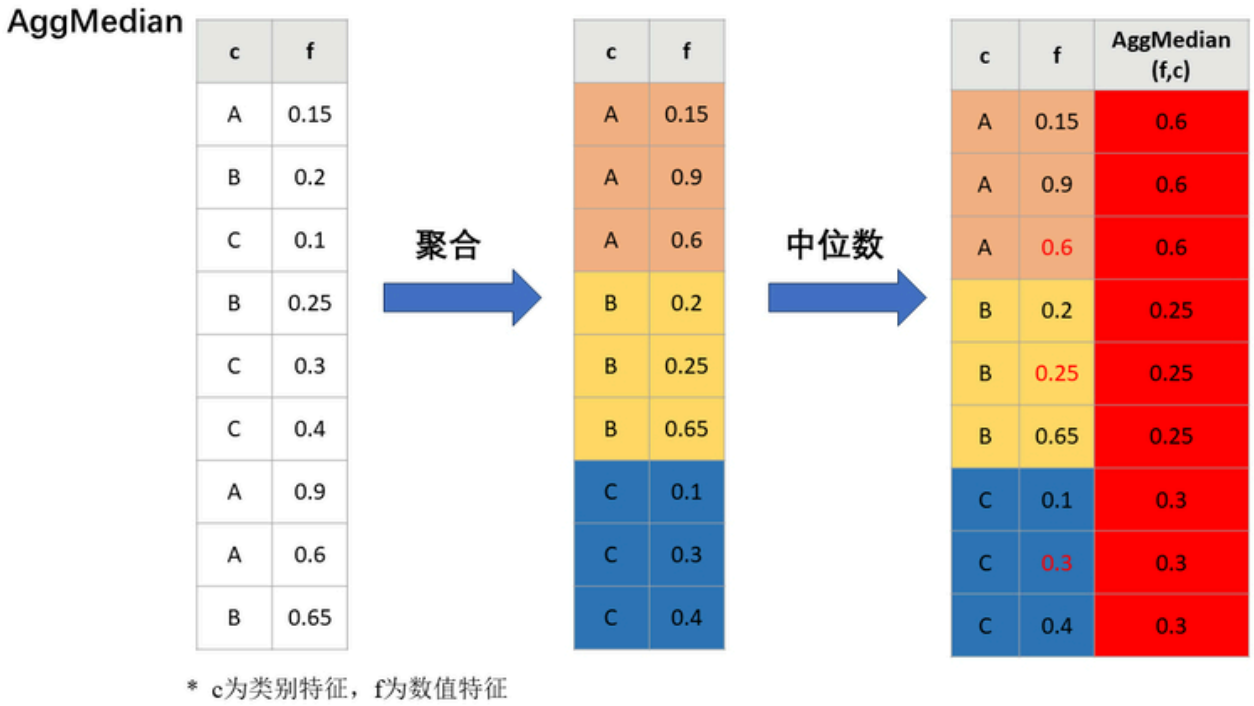
表示特征c各类别中f的平均值



AggMean

AggMedian(f, c)

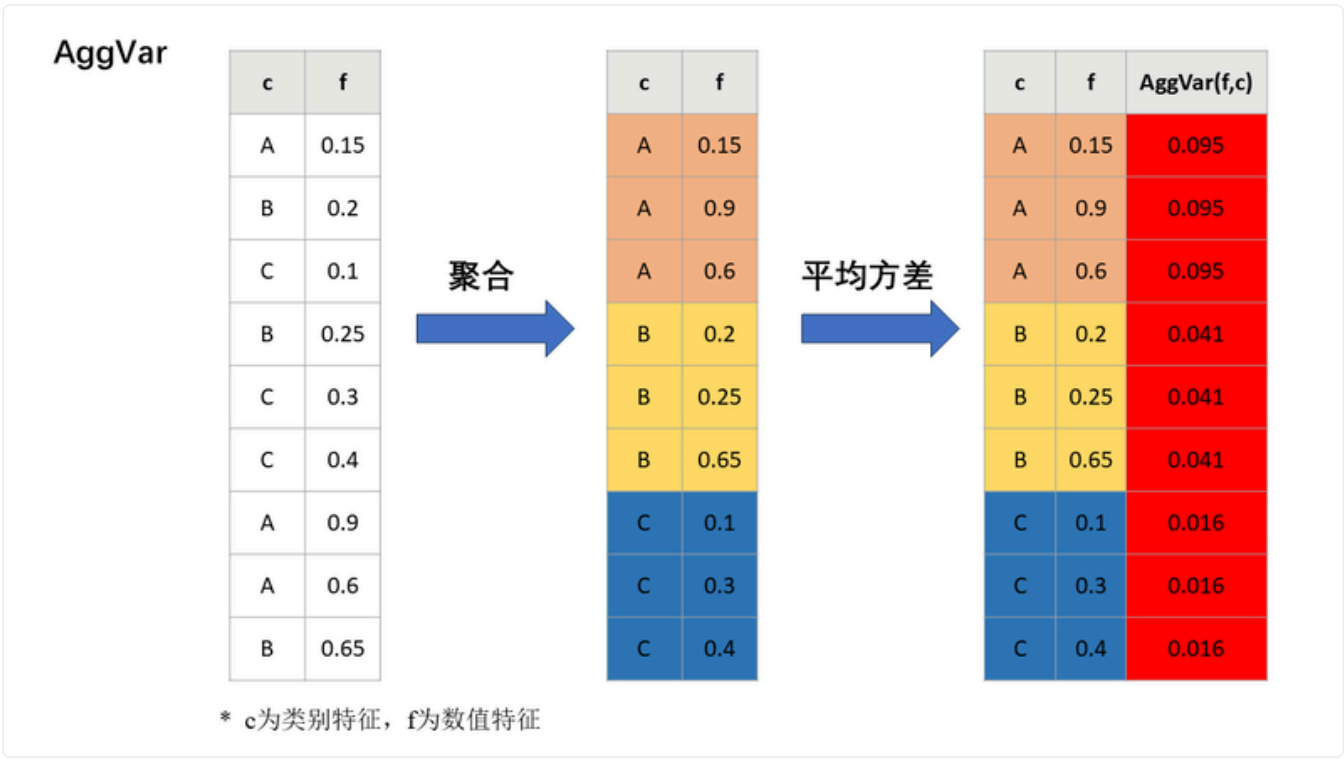
表示特征c各类别中f的中位数



AggMedian

AggVar(f, c)

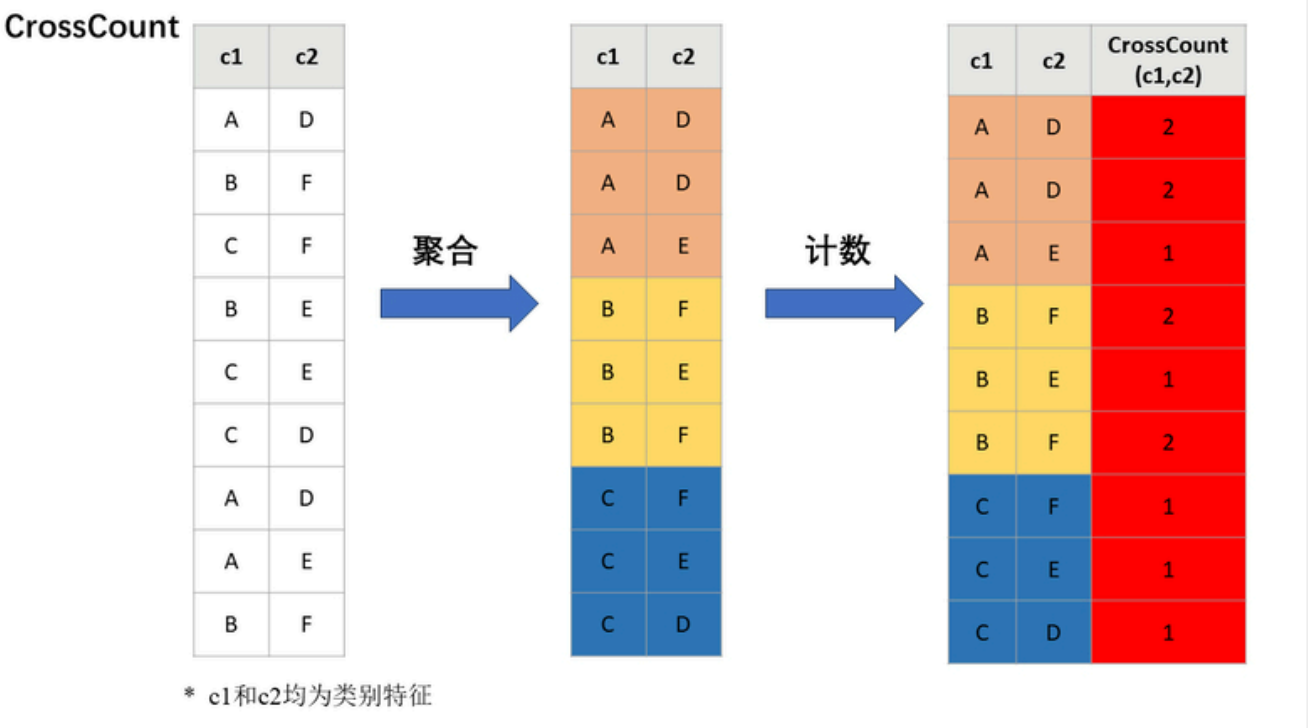
表示特征c各类别中f的方差



AggVar

CrossCount([c1, c2, ..])

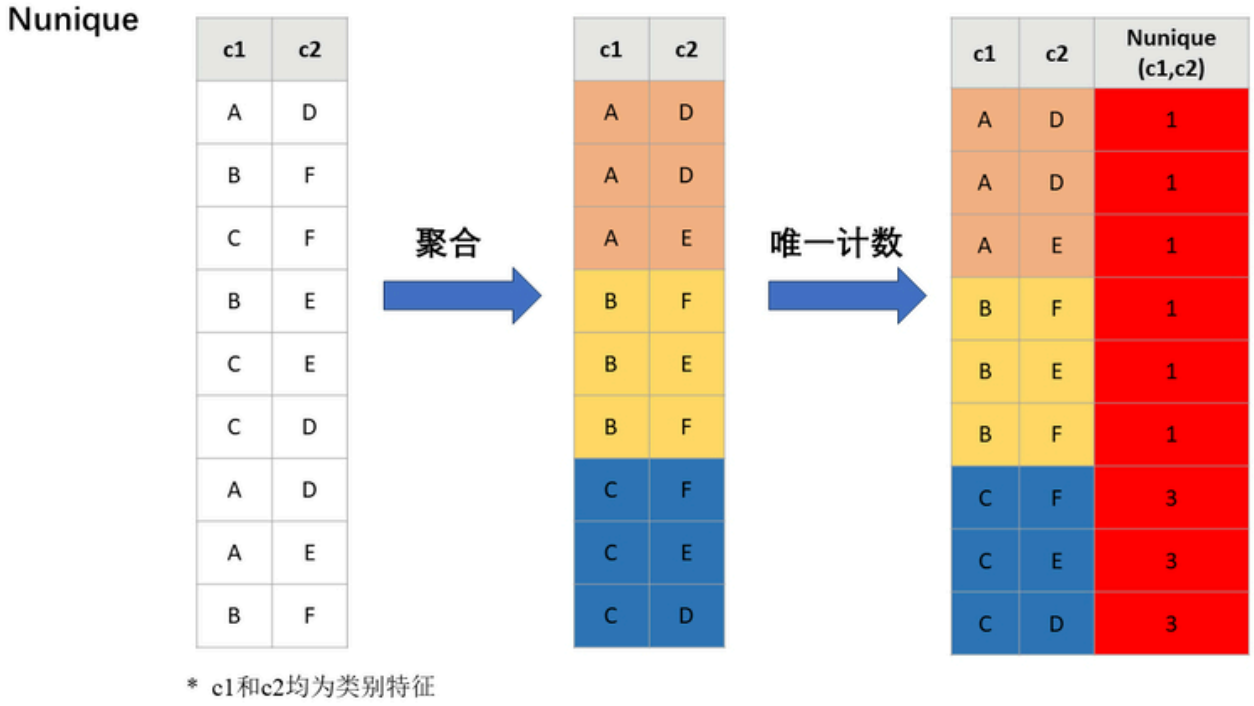
根据特征list聚合的计数，list长度大于等于2



CrossCount

Nunique(c1, c2)

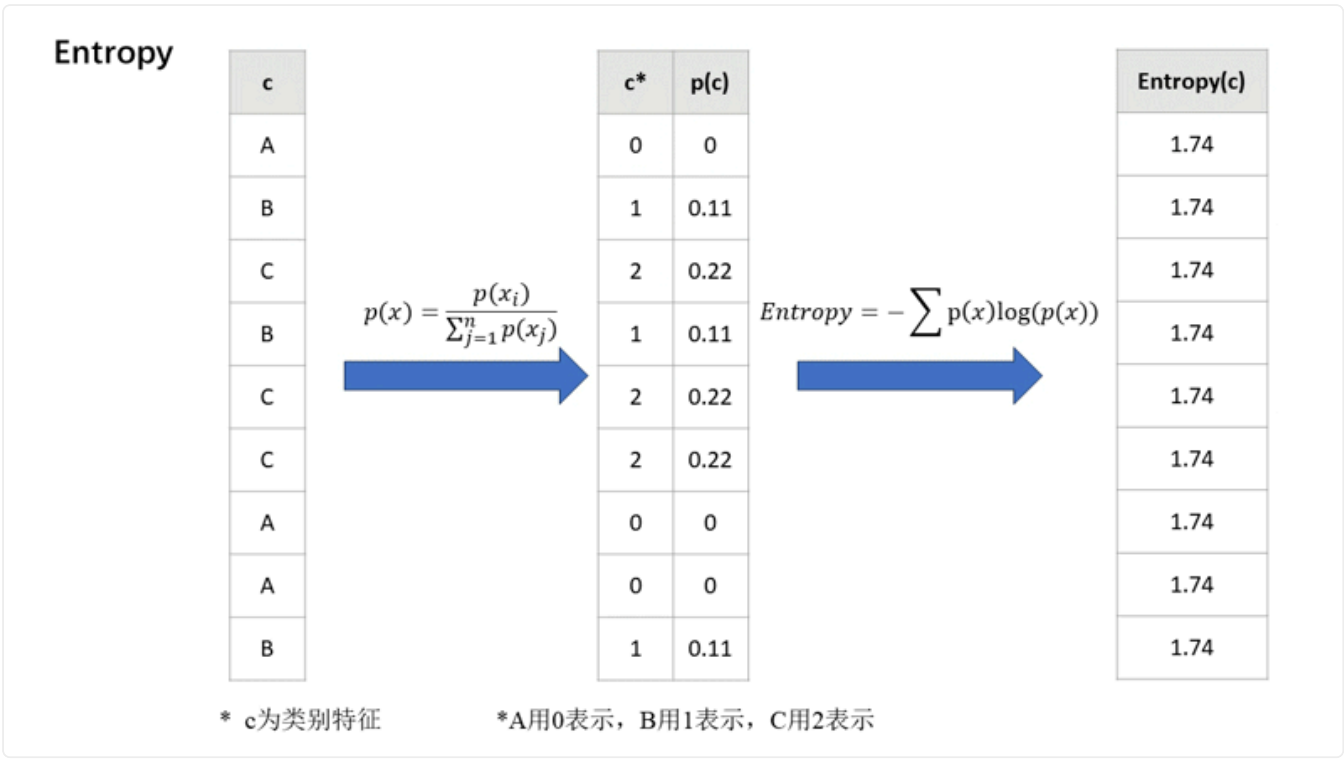
表示特征c2各类别中c1的唯一值计数



Nunique

Entropy(c)

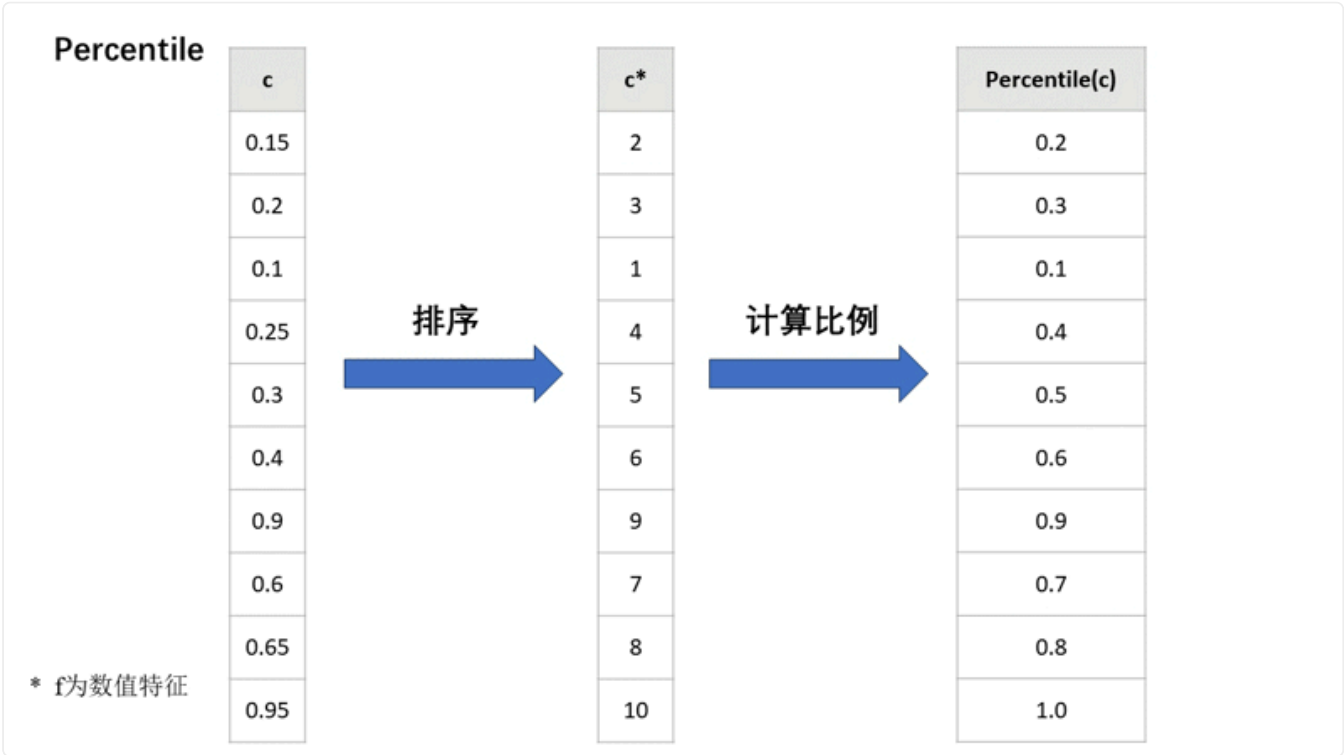
表示特征c各类别的熵



Entropy

Percentile(f)

表示特征f各个数据的百分位



Percentile

Combine(c1, c2)

表示特征c1和特征c2的字符结合

Combine

c1	c2
A	D
B	F
C	F
B	E
C	E
C	D
A	D
A	E
B	F

结合

c1	c2	Combine (c1,c2)
A	D	A_D
B	F	B_F
C	F	C_F
B	E	B_E
C	E	C_E
C	D	C_D
A	D	A_D
A	E	A_E
B	F	B_F

* c1和c2均为类别特征

Combine

Count(c)

特征c各类别的计数

Count

c
A
B
C
B
C
C
A
A
D

计数

c	Count(c)
A	3
B	2
C	3
B	2
C	3
C	3
A	3
A	3
D	1

* c为类别特征

Count

Equal(f1, f2)

判断特征f1和特征f2是否相等

Equal

f1	f2
0.72	0.15
0.3	0.3
0.88	0.1
0.12	0.25
0.6	0.3
0.2	0.2
0.5	0.9
0.9	0.9
0.4	0.65

是否相等

f1	f2	Equal(f1,f2)
0.72	0.15	flase
0.3	0.3	true
0.88	0.1	flase
0.12	0.25	flase
0.6	0.3	flase
0.2	0.2	true
0.5	0.9	flase
0.9	0. 9	true
0.4	0.65	flase

* f1和f2均为数值特征

Equal

Min(f1, f2)

取特征f1和特征f2相比的较小值

Min

f1	f2
0.72	0.15
0.3	0.2
0.88	0.1
0.12	0.25
0.6	0.3
0.2	0.4
0.5	0.9
0.9	0.6
0.4	0.65

最小值

f1	f2	Min(f1,f2)
0.72	0.15	0.15
0.3	0.2	0.2
0.88	0.1	0.1
0.12	0.25	0.12
0.6	0.3	0.3
0.2	0.4	0.2
0.5	0.9	0.5
0.9	0.6	0.6
0.4	0.65	0.4

* f1和f2均为数值特征

Min

Max(f1, f2)

取特征f1和特征f2相比的较大值

Max

f1	f2
0.72	0.15
0.3	0.2
0.88	0.1
0.12	0.25
0.6	0.3
0.2	0.4
0.5	0.9
0.9	0.6
0.4	0.65

最大值

f1	f2	Max(f1,f2)
0.72	0.15	0.72
0.3	0.2	0.3
0.88	0.1	0.88
0.12	0.25	0.25
0.6	0.3	0.6
0.2	0.4	0.4
0.5	0.9	0.9
0.9	0.6	0.9
0.4	0.65	0.65

* f1和f2均为数值特征

Max

Sigmoid(f)

对特征f进行sigmoid非线性变换

Sigmoid

f
0.15
0.2
0.1
0.25
0.3
0.4
0.9
0.6
0.65

$$\frac{1}{1+e^{-x}}$$

f	Sigmoid(f)
0.15	0.54
0.2	0.55
0.1	0.52
0.25	0.56
0.3	0.57
0.4	0.60
0.9	0.71
0.6	0.65
0.65	0.66

* f为数值特征

Sigmoid

Round(f)

对特征f进行四舍五入

Round

f
3.15
1.2
4.1
1.25
5.3
9.4
2.9
6.6
3.65

四舍五入

f	Round(f)
3.15	3
1.2	1
4.1	4
1.25	1
5.3	5
9.4	9
2.9	3
6.6	7
3.65	4

* f为数值特征

Round

Residual(f)

保留特征f求小数点后的数值

Residual

f
3.15
1.2
4.1
1.25
5.3
9.4
2.9
6.6
3.65

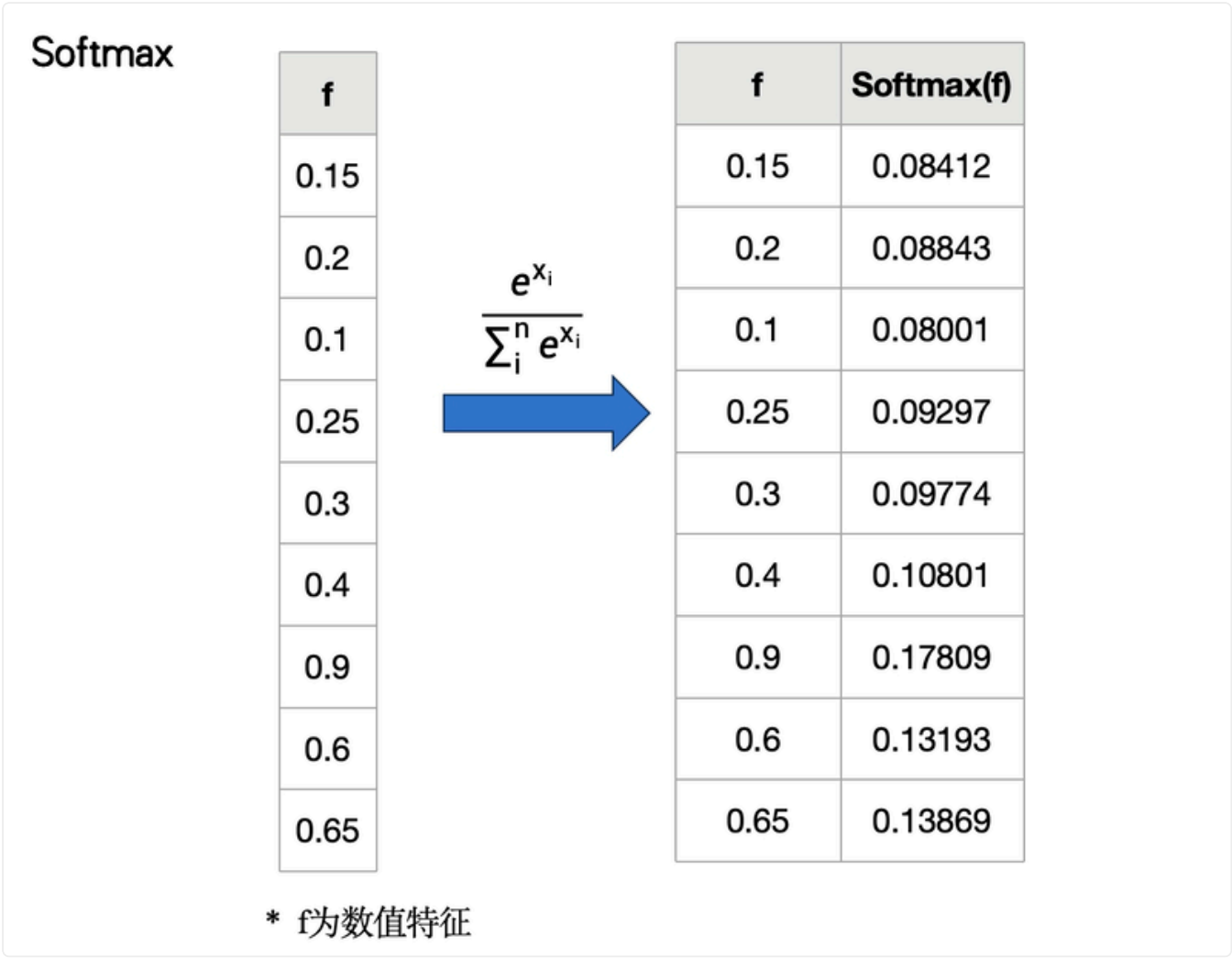
f	Residual(f)
3.15	0.15
1.2	0.2
4.1	0.1
1.25	0.25
5.3	0.3
9.4	0.4
2.9	0.9
6.6	0.6
3.65	0.65

* f为数值特征

Residual

Softmax(f)

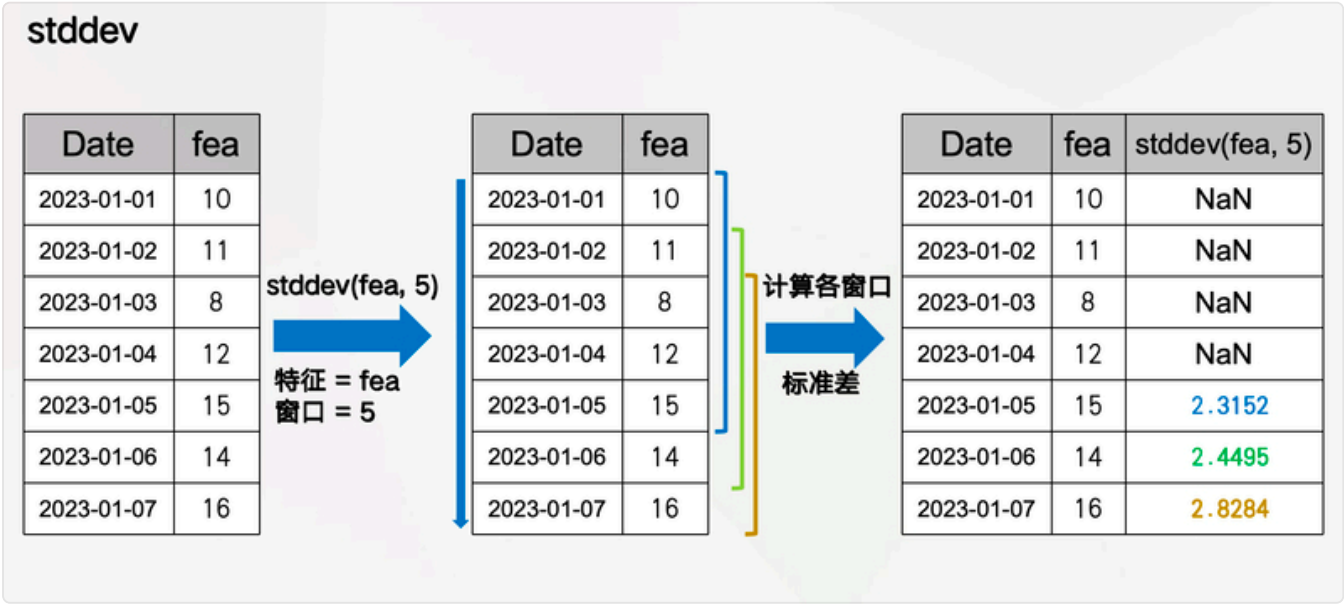
有限项离散概率分布的梯度对数归一化



Softmax

stddev(f, w)

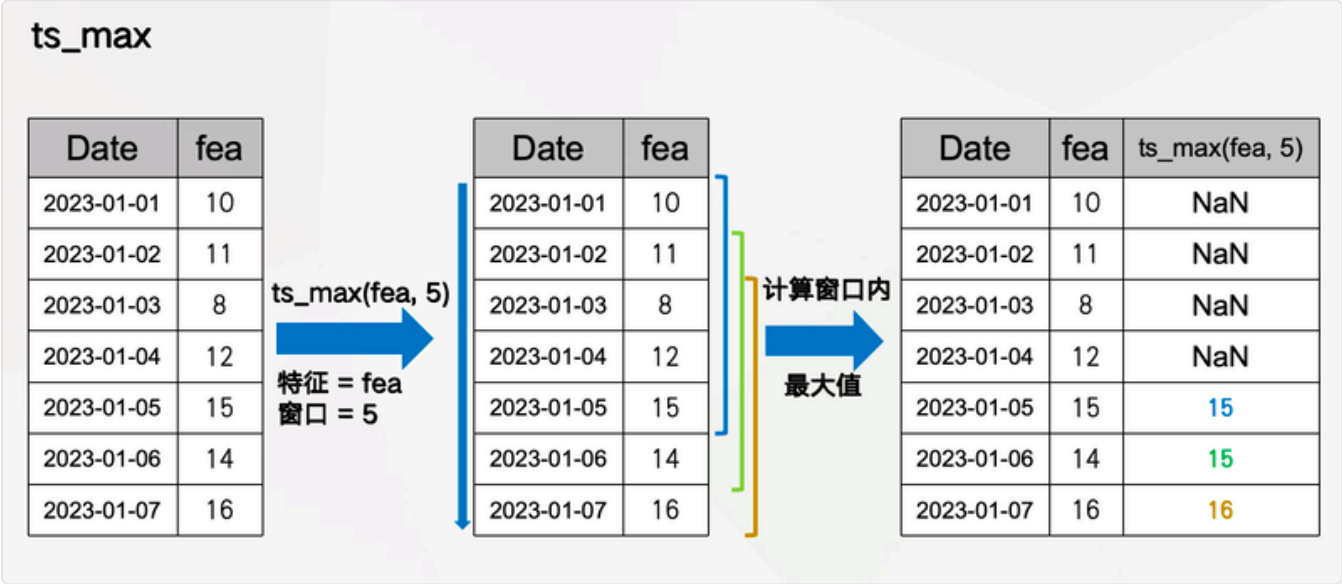
计算窗口内特征f的标准差



stddev

ts_max(f, w)

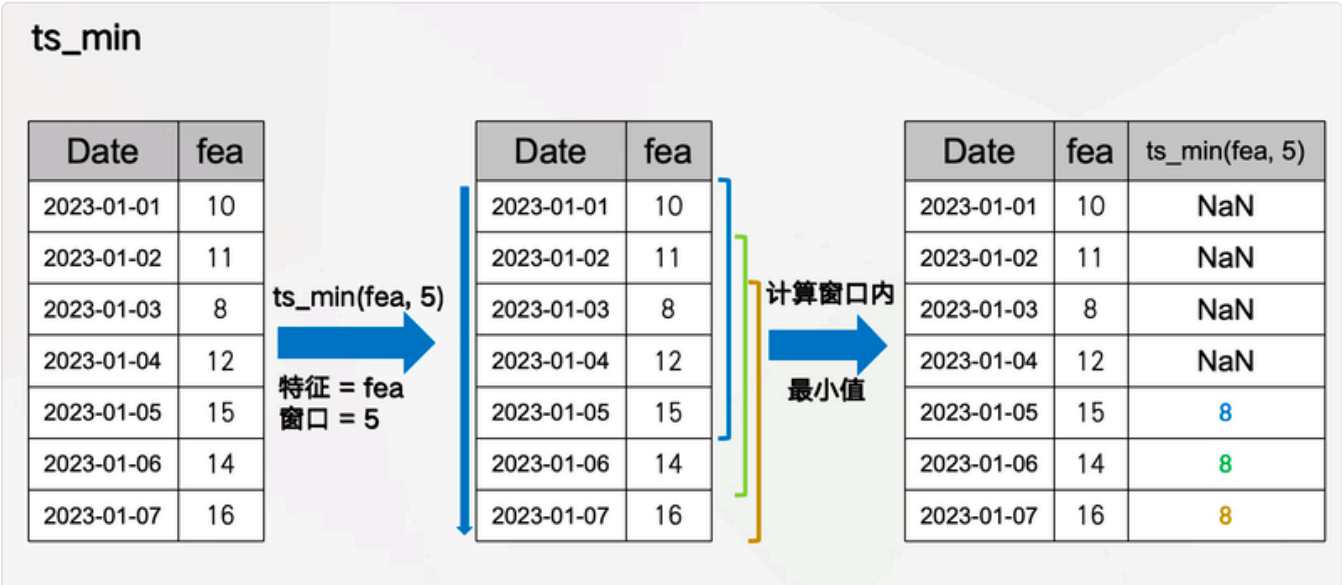
计算窗口内特征f的最大值



ts_max

ts_min(f, w)

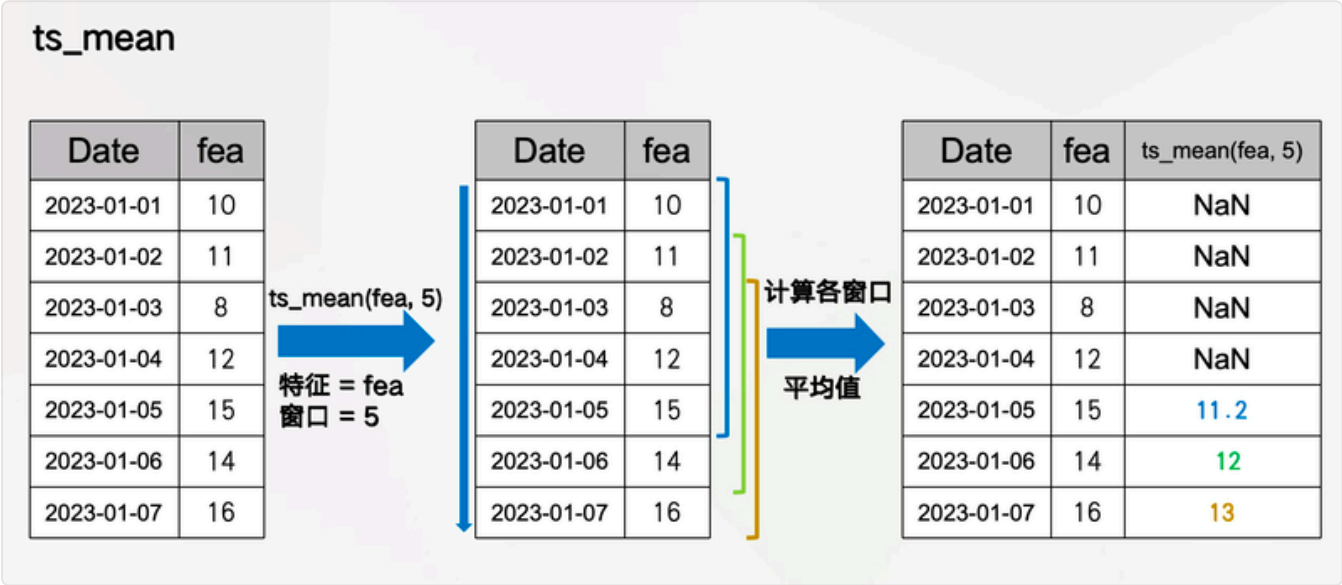
计算窗口内特征f的最小值



ts_min

ts_mean(f, w)

计算窗口内特征f的平均值



ts_mean

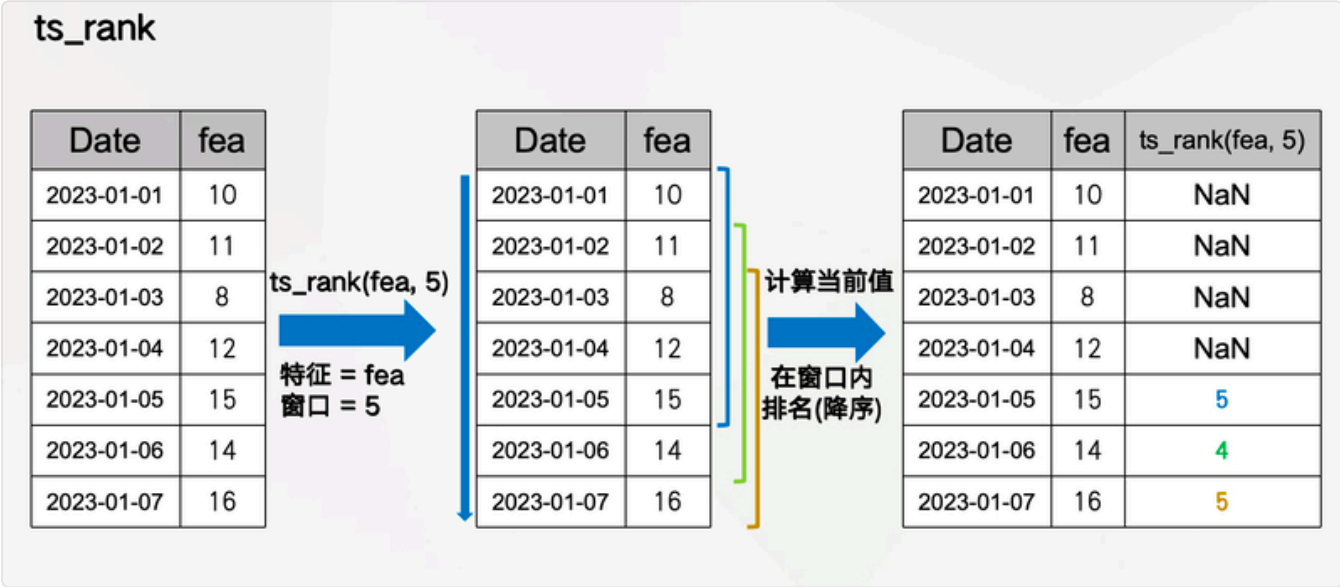
ts_sum(f, w)

计算窗口内特征f的加和值

ts_sum

ts_rank(f, w)

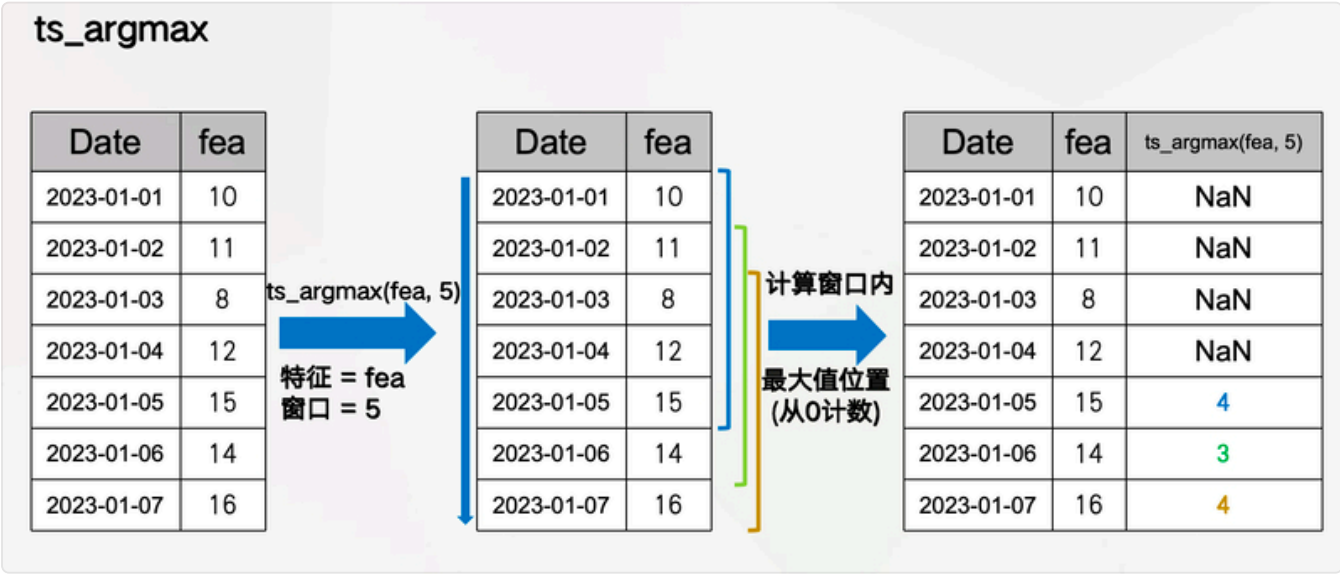
计算特征f当前值在在窗口内的排名（降序）



ts_rank

ts_argmax(f, w)

计算窗口内特征f最大值位置索引（从0计数）



ts_argmax

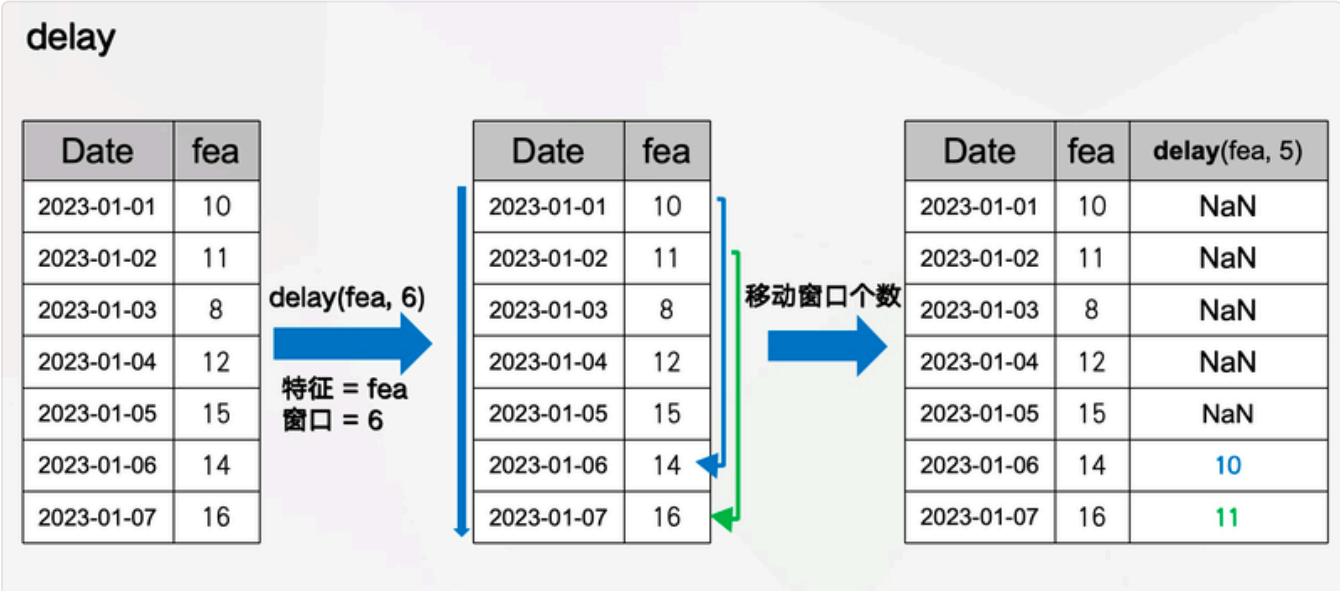
ts_argmin(f, w)

计算窗口内特征f最小值位置索引（从0计数）

ts_argmin

delay(f, w)

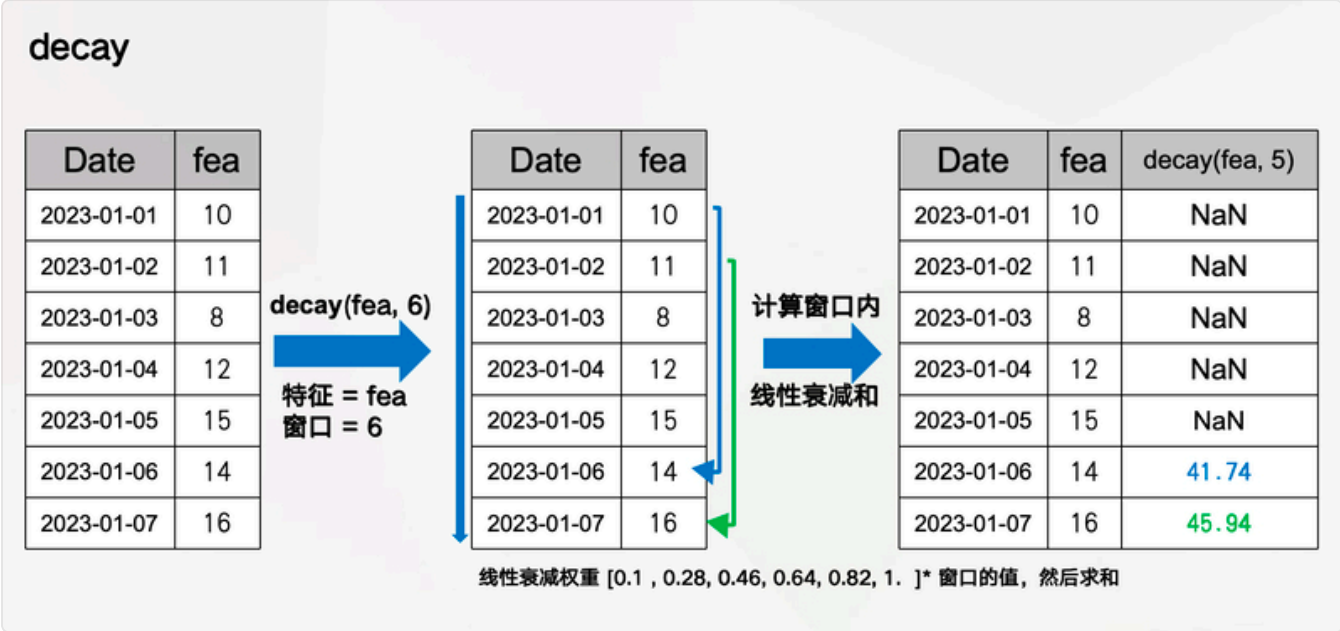
获取窗口内特征f最早时间所对应的值



delay

decay(f, w)

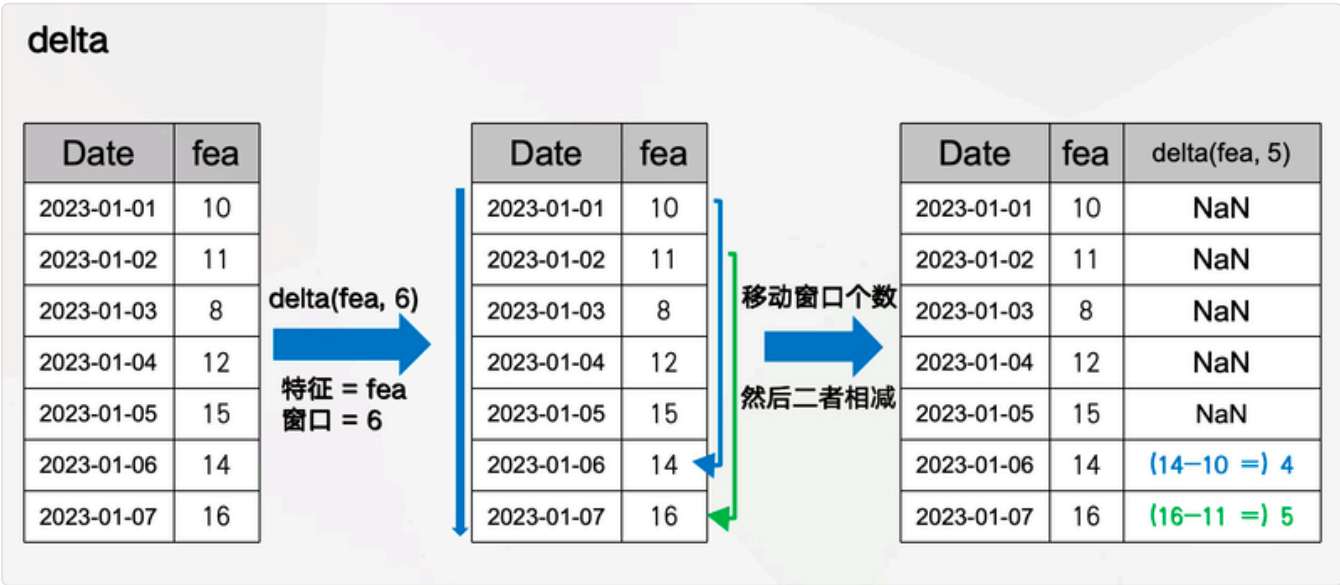
计算窗口内特征f线性衰减和



decay

delta(f, w)

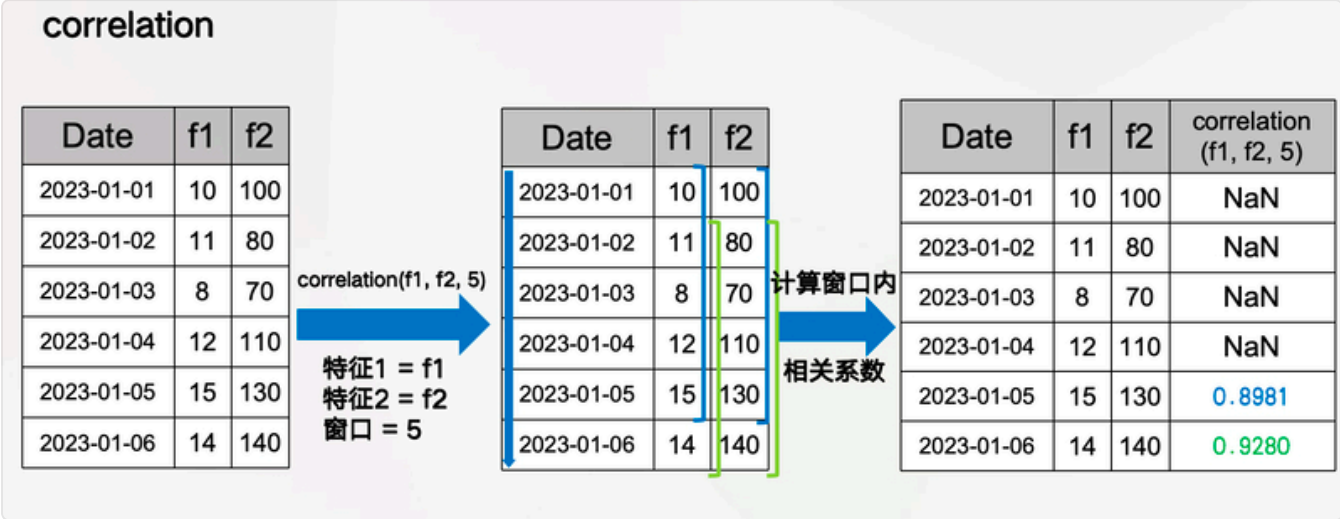
计算窗口内特征f最晚和最早时间所对应值的差值



delta

correlation(f1, f2, w)

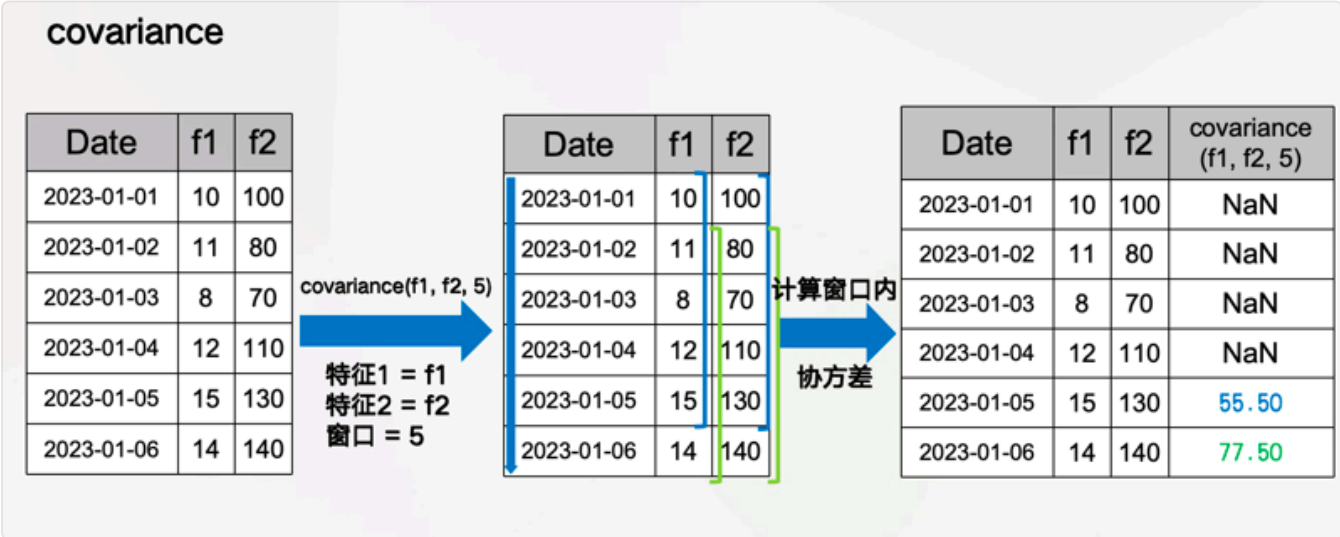
计算窗口内特征f1和特征f2的相关系数



correlation

covariance(f1, f2, w)

计算窗口内特征f1和特征f2的协方差



covariance

上一页
SasS版操作手册

下一页
词汇和术语

词汇和术语

任务类型

在工程实践中，常见的机器学习建模任务包含回归、分类和聚类这三大任务。

- **回归任务**：旨在预测连续数值，比如房价预测或销售量预测。
- **分类任务**：旨在将数据分为不同类别，如垃圾邮件检测或疾病诊断。
- **聚类任务**：旨在将数据分成类似的组，帮助发现数据中的潜在结构，比如市场细分或者客户群体分析。需要强调的是，在建模前我们可能不知道存在多少个细分市场和客户群体。

其中回归和分类任务又被称为**有监督学习**，在建模时需要用户对每条数据样本相应的标签，这样算法在学习过程中就受到了监督和指导。这种指导有助于算法理解输入和输出之间的关系，并在面对新的、未见过的数据时做出准确的预测或分类。

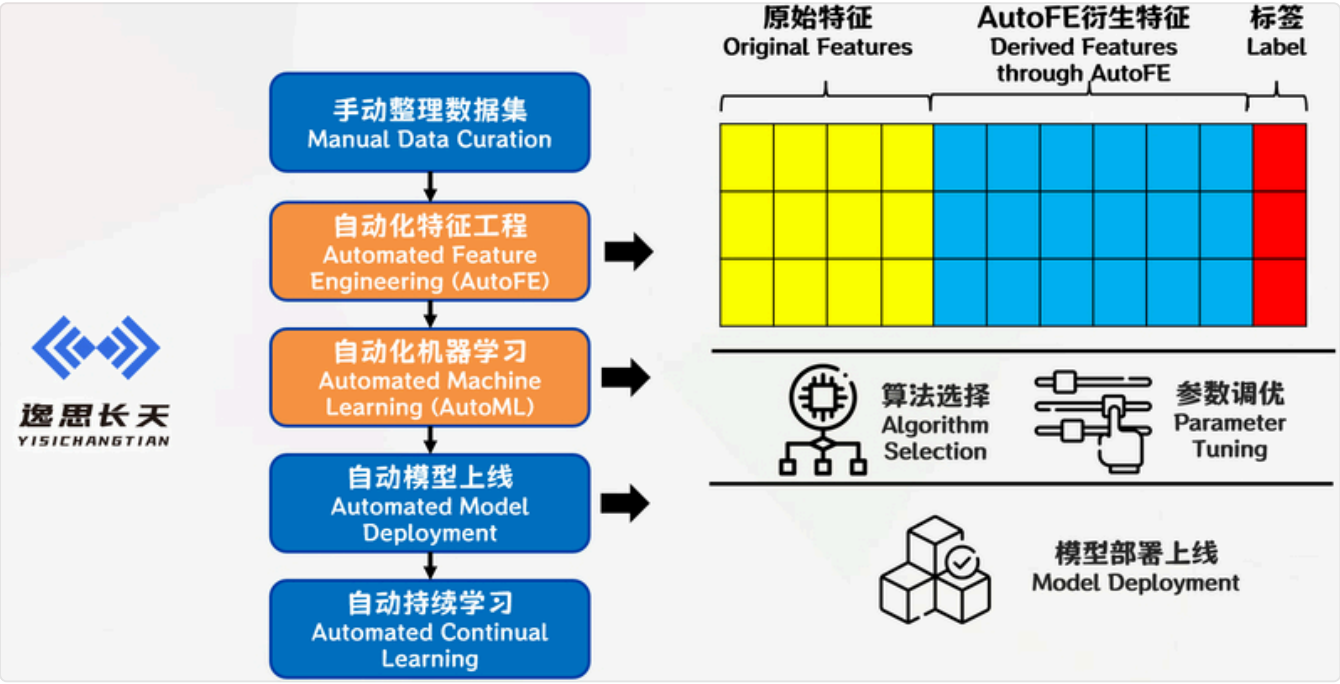
而聚类任务被称为**无监督学习**，因为在这种任务中，模型并不依赖预先标记或已知的输出来进行学习。相反，算法需要发现数据中的潜在结构或模式，将数据样本分成类似的组，而不需要预先知道这些类别的标签。

在工程实践中，有监督学习的任务通常更为常见，因为它适用于许多现实世界的问题，并且有标签的数据更容易获取和使用，因此本平台当前仅支持回归和分类这两种建模任务。

对于建模任务是回归还是分类需要用户自己根据实际情况进行定义，比如在量化投资场景中，我们可以预测明天股价的数值，这样标签是一个连续数值，就是回归任务。我们也可以选

择预测明天股价的涨跌方向，这样标签就是上涨下跌的2个类别，就是二分类任务。当然我们还可以根据上涨下跌的不同程度，细分成更多的类别，这就是多分类任务。

建模模式



用户在上传完毕手动整理数据集，并配置和启动建模任务后，本平台默认会在后台先进行自动化特征工程（AutoFE），再进行自动化机器学习（AutoML）。用户也可以在高级配置中取消勾选。

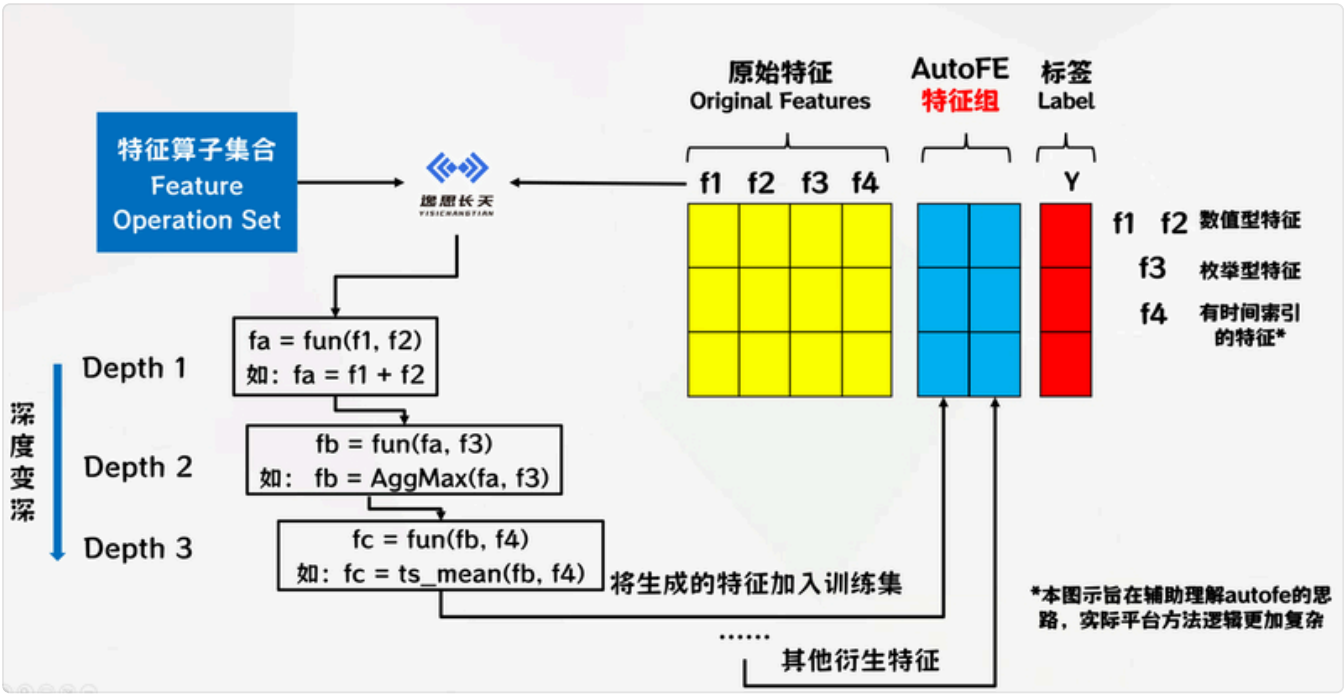
在AutoFE环节，本平台将根据原始特征和平台算子构建并高效探索特征空间，从而自动进行特征生成衍生，构建共同增益最大的特征组，为用户提供更多解释数据的新角度和新方法。

在AutoML环节，本平台将根据上游的特征和标签自动进行算法的选择和超参数的调优。如果上游进行了AutoFE过程，则此处的特征包含了原始特征和AutoFE衍生特征，否则仅包含原始特征。我们推荐用户同时使用AutoFE和AutoML，这将可能获得更好的建模效果。

AutoFE

AutoFE的中文名称是**自动化特征工程**（Automated Feature Engineering）。本平台内置了逸思长天自研的自动化特征工程搜索算法，会根据用户上传的数据集原始特征和算子集合（详见：AutoFE算子手册）构建并高效探索特征空间，逐步进行特征生成衍生。

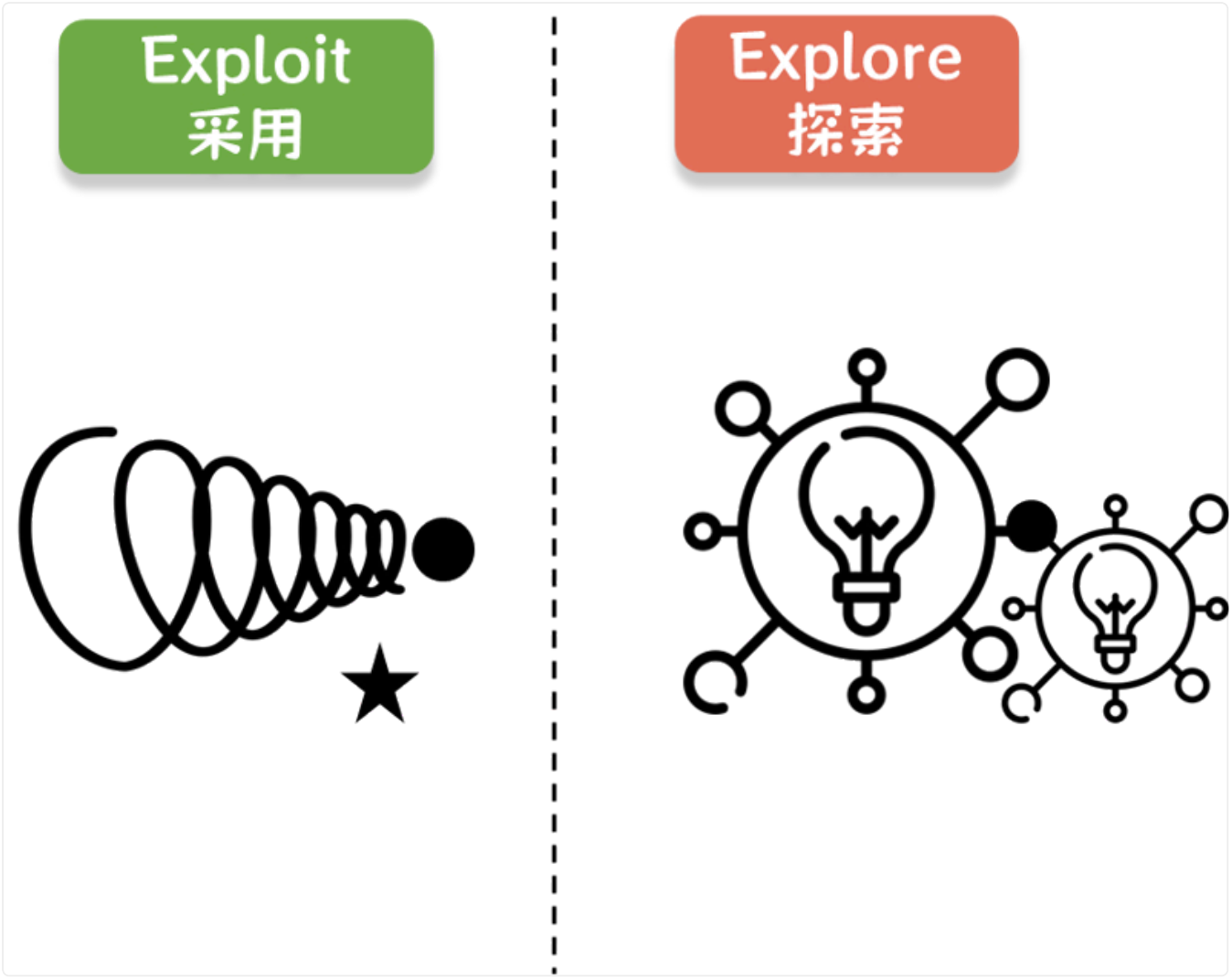
的特征组，这些高级特征有显式的公式展示，有更好的可解释性，给用户提供了解释数据的新角度新方法。



在自动化特征工程技术领域，目前仍然存在着特征探索空间巨大且不好把控探索过程的挑战，所以我们逸思长天团队设计并平衡了采用和探索这两种模式。

在采用模式下，AutoFE会根据之前探寻过的最优路径进行搜索，而不去考虑那些次优路径，这样的好处是可以加快搜索过程的收敛，但也可能因为注重“短期利益”而错失远期效果更好的特征构造结果。

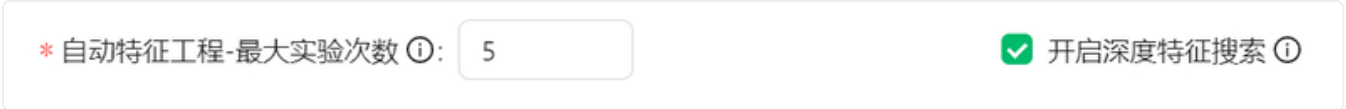
而探索模式下则恰恰相反，AutoFE会发散性的产生更多的路径，然后在部分表现较好的路径中进一步发散性的搜索。这样的好处是会探寻更多可能，但是也会消耗大量的时间。



由于这两种方式是你中有我、我中有你。因此本平台提供给用户三个选项：更有创造力、更平衡和更精确。其中的更精确和更有创造力就分别对应着“更多采用、少量探索”和“更多探索，少量采用”，而更平衡则是两者的折中，我们的默认配置也是更推荐更平衡的模式。

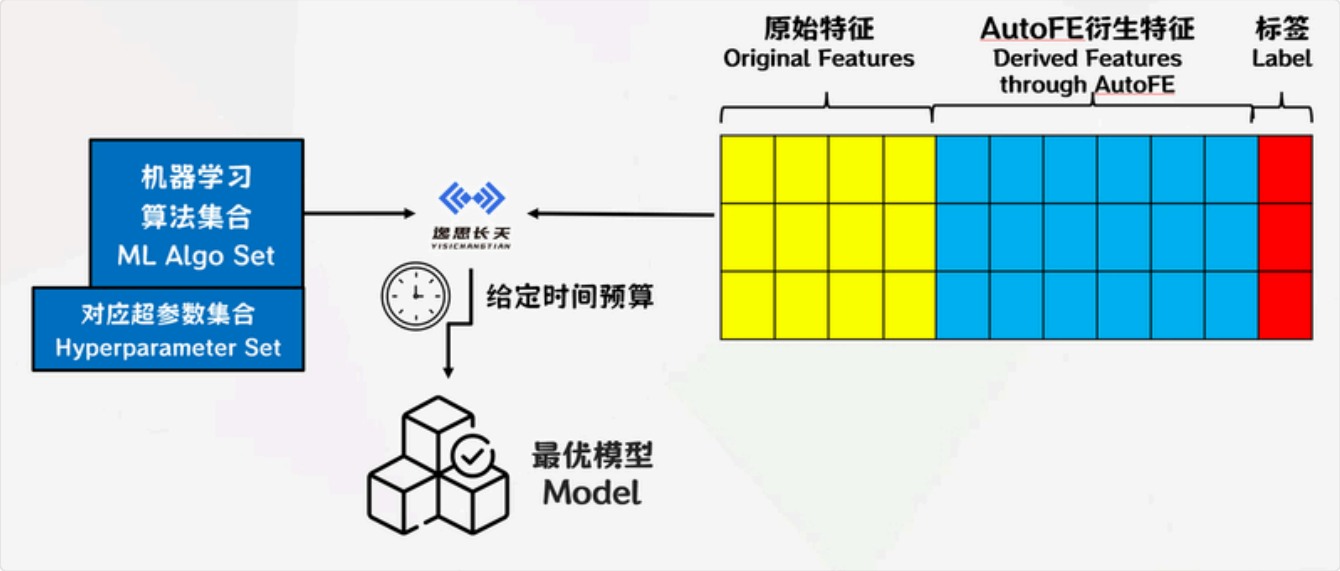
用户可以在新建任务时选择 【高级设置】 通过“特征生成模式”选项进行设置，如下图：

在建模任务的常规配置部分增加了“开启深度特征搜索”选项，默认是勾选的。这意味着我们会对特征进行更深层的嵌套搜索，这也会花费更多的时间。否则就会浅层嵌套，比如：嵌套一层或者两层。如果用户觉得过于复杂的特征没有可解释性，这里可以取消勾选。



AutoML

AutoML的中文名称是**自动化机器学习**（Automated Machine Learning）。本平台内置了逸思长天自研的自动化机器学习算法，会根据上游的数据集，和机器学习算法及其对应的超参数集合，在用户给定的时间预算下搜索出最优的机器学习模型。



随机种子

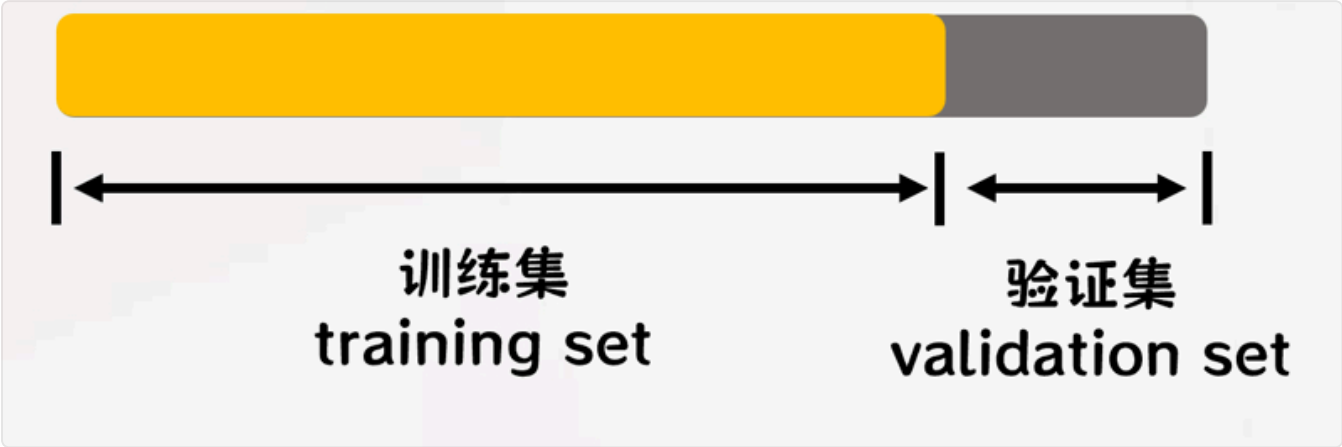
随机种子是用于初始化随机数生成器的起始点或种子值，一般为正整数。在机器学习和统计建模中，许多算法和模型会使用随机性，例如在数据拆分、权重初始化或样本抽样等过程中会用到随机数生成器。这些随机数的生成是基于初始种子值进行的。

随机种子能够控制模型在不同运行中的随机性，使得实验结果可以重现（在其他配置不变的情况下）。如果不设置随机种子，每次运行时生成的随机数可能不同，这会导致每次运行模型时得到不同的结果。而设置了随机种子之后，即使重新运行模型或算法，只要种子值相同，生成的随机数序列也将保持一致，从而确保了结果的可重复性。

保持结果的可重复性对于实验复现、模型评估和调试都非常重要。因此，在建模过程中，设置随机种子是一种常用的实践，以确保结果的一致性和可重现性。

验证集分割

验证集分割是为了评估模型的性能和泛化能力。本平台在建模初始环节会根据用户设定的分割方式和分割比例将数据集分成训练集和验证集两个部分。一般来说，训练集的比例要更多，以此让模型在训练阶段获取更多信息。



平台将使用训练集来训练模型，但最终模型的效果评估是对新的、未见过的验证集，平台页面中用户看到的建模结果的分数就是基于验证集的结果。

验证集分割将有助于提升模型的泛化能力，即模型对新数据的适应能力。如果模型在训练集上表现良好但在验证集上表现较差，可能存在过拟合的问题，即模型过度适应了训练数据，而无法很好地推广到新的数据上。

本平台支持常见的验证集分割方式如下：

- **随机抽样**： 随机从整个数据集中抽取样本作为验证集和训练集，没有特定的分布要求。
- **分层抽样**：根据标签或类别的分布，在切分验证集和训练集时保持相似的标签分布。这对于标签类别不平衡的情况特别有用，以确保训练集和验证集中都包含各个类别的代表性样本。
- **时间分割**：在时间序列数据中，按照时间顺序切分验证集和训练集（防止扰乱时间顺序，影响时序特征构造）。这对于时间相关性强的数据集，例如：股票数据、天气数据等，是一种常用的切分方式。
- **手动分割**：将训练集和验证集分割由用户手动上传数据集文件。

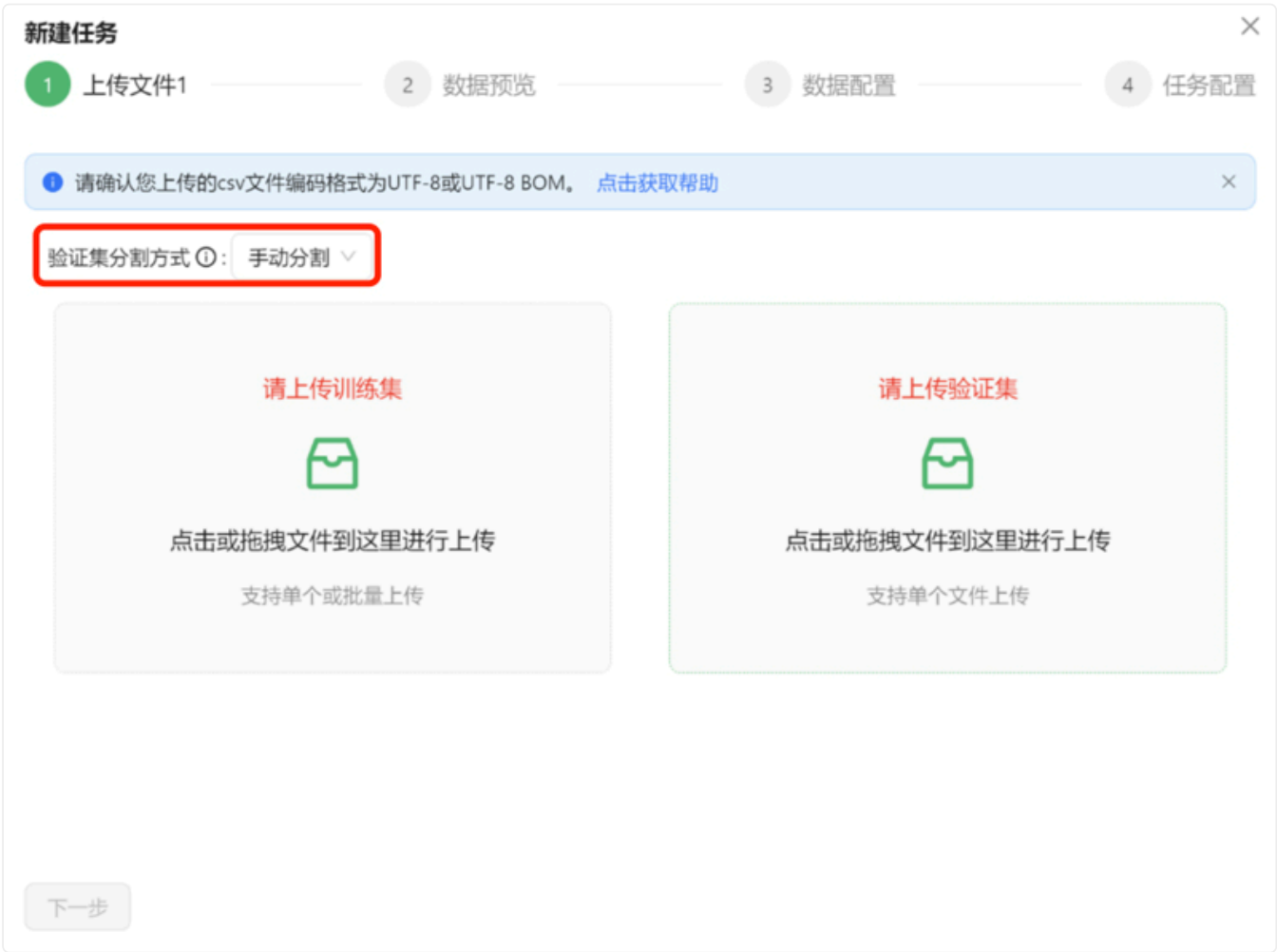
在工程实践中，分类任务的样本标签分布往往不够均衡时，比如在金融领域中，欺诈交易的发生率往往相对较低，此时如果还是用随机抽样方式进行数据集的分割，有不小的概率会导致验证集没有欺诈交易的样本，从而验证集评估的分数的有效性大大降低。

此时使用分层抽样的方式将避免上述情况发生。通过将总体数据根据样本标签先行分层，然后再分别对每类标签按比例单独抽样，从而能确保训练和验证集都有足够的欺诈交易样本。

当用户选择**传统经典机器学习**时，验证集分割方式支持：分层抽样、随机抽样和手动分割。

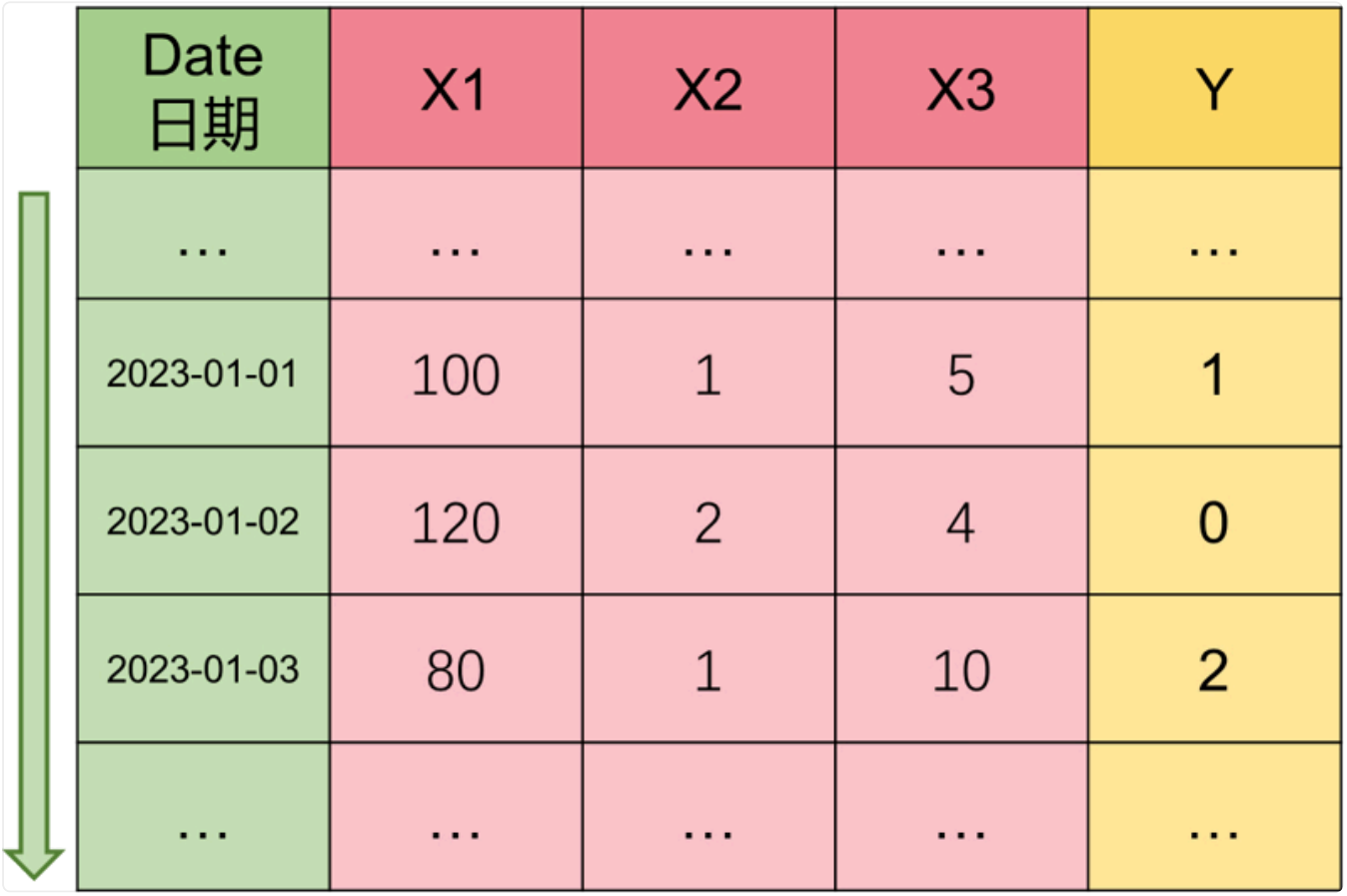
当用户选择**时间序列机器学习**时，验证集分割方式支持：时间分割和手动分割。

除了手动分割之外的验证集分割方式都具有一定的随机性，有时很难按指定的要求去分割数据集。用户可以选择手动分割的方式，将训练集和验证集分割由用户自行决定，此时需要分别单独上传训练集和验证集这两份数据文件，如下图：



时间序列

时间序列顾名思义是按时间顺序排列的一系列数据点的集合。这些数据点通常以等间隔的时间间隔进行采集或观察。时间序列可以是各种类型的数据，如股票价格、气温、销售数据、传感器读数等。这种数据一般都会带有一个时间列标签，据此可以**按时间进行升序排序**。



Date 日期	X1	X2	X3	Y
…	…	…	…	…
2023-01-01	100	1	5	1
2023-01-02	120	2	4	0
2023-01-03	80	1	10	2
…	…	…	…	…

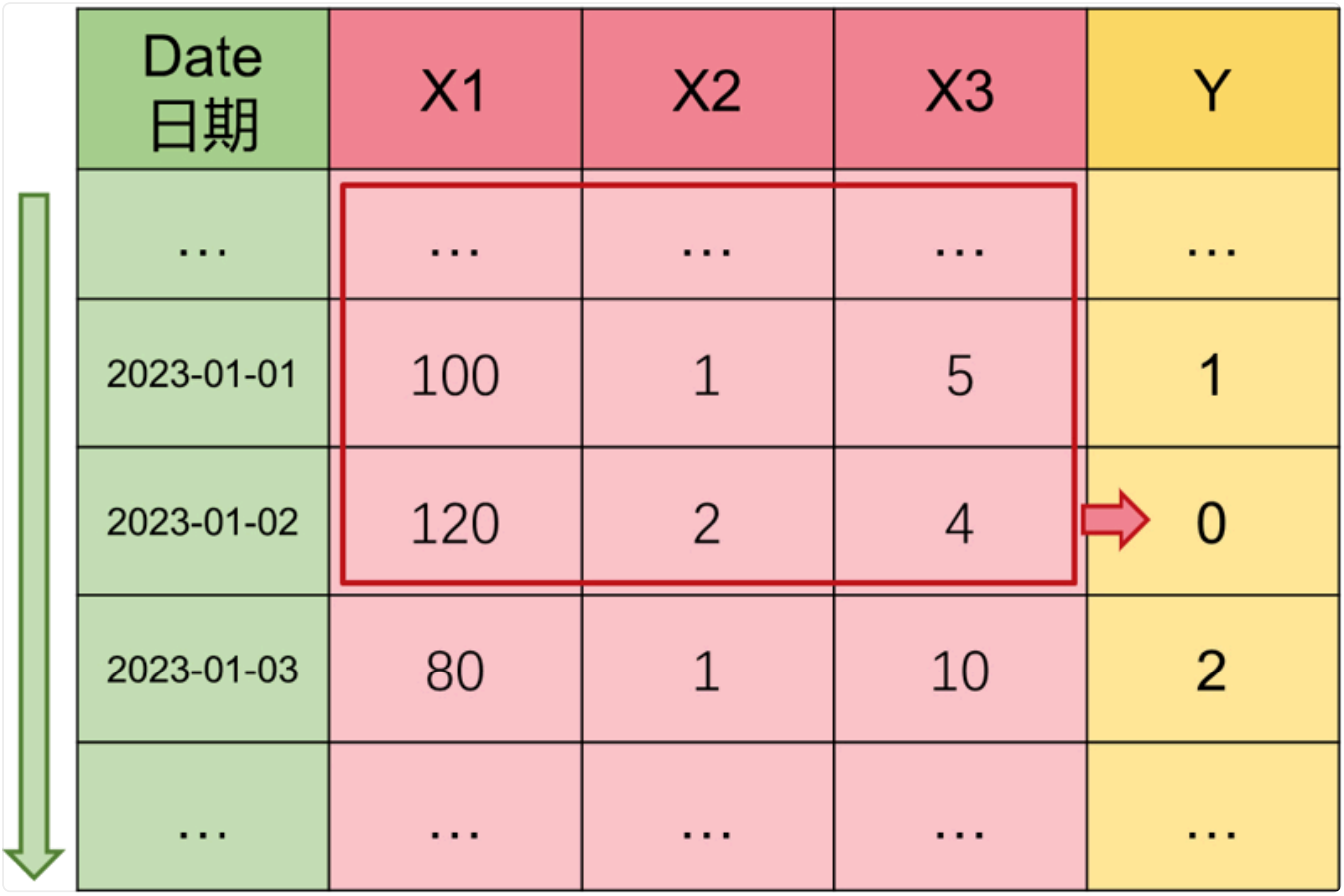
对于某些特殊的时间序列数据可能还会有**时间组标识**，用来区分同一时间的这条数据是由谁产生的，如同一时间有多个股票的特征数据，或者传感器数据等等。

	ts_code	trade_date	open	high	low	close	vol	amount
0	000001.SZ	20180102	13.35	13.93	13.32	13.70	2081592.55	2856543.822
1456	000002.SZ	20180102	31.45	32.99	31.45	32.56	683433.50	2218502.766
2912	000004.SZ	20180102	22.29	22.49	22.00	22.34	6261.81	13951.004
4340	000005.SZ	20180102	4.15	4.50	4.15	4.32	71539.34	30529.757
8609	000008.SZ	20180102	8.75	8.82	8.70	8.77	82899.99	72523.239
...	...	...	...	...	...	...	...	...
6134669	873679.BJ	20231229	22.97	23.59	22.82	23.48	16154.23	37624.974
6134715	873693.BJ	20231229	42.06	42.98	41.01	41.54	14218.37	59524.049
6134734	873703.BJ	20231229	38.49	39.16	38.11	38.48	20310.30	78364.940
6134786	873726.BJ	20231229	39.43	47.96	37.90	43.64	46619.68	200408.958
6134824	873833.BJ	20231229	14.86	15.11	14.53	15.06	26039.82	38753.033

在传统经典的机器学习建模任务中，我们往往用一行数据的一个或多个特征(X1、X2、X3)来预测这行的Y标签，我们默认预测Y所需要的信息都已经包含在了这行的特征当中了。

X1	X2	X3	Y
…	…	…	…
100	1	5	1
120	2	4	0
80	1	10	2
…	…	…	…

但是对于时间序列的建模场景，如果我们认为当前的**时间序列存在一定的模式、趋势和周期性**，那么建模时就不仅仅要用当前时刻的特征信息来预测Y，更要**回望当前时刻过去的特征信息去预测Y。**



Date 日期	X1	X2	X3	Y
…	…	…	…	…
2023-01-01	100	1	5	1
2023-01-02	120	2	4	0
2023-01-03	80	1	10	2
…	…	…	…	…

这里需要强调的是，尽管有的数据的表现是时间序列的形态，但是由于其不存在趋势周期等历史规律，所以这类建模任务也不属于时间序列建模任务，比如彩票号码中奖预测，显然是和过去信息是无关的。所以，**对于一个建模任务是属于经典机器学习还是时间序列机器学习，需要各位自行去判断了，但是显然没有时间列索引一定不能是时间序列。**

接着我们细化到时间序列机器学习的操作，主关注两个问题：

① 计算机如何识别各种格式不同的时间列呢？

② 平台是如何吸收历史特征信息的呢？

对于时间列格式，**由于其的确存在较高的灵活度，所以我们将其交给用户进行定义**，用户在上传数据集后，会选择哪一列为时间列，然后此时会有配置按钮供用户配置。

以金融领域的期货数据日频数据为例，由于其数据为日级别，所以用户在配置时选择日期格式即可，并且这里的"2018-11-20"也满足"yyyy-mm-dd"格式，如果你的格式不是这种日期表现，而是"11/20/2018"，那么你可以自定义修改为"mm/dd/yyyy"。

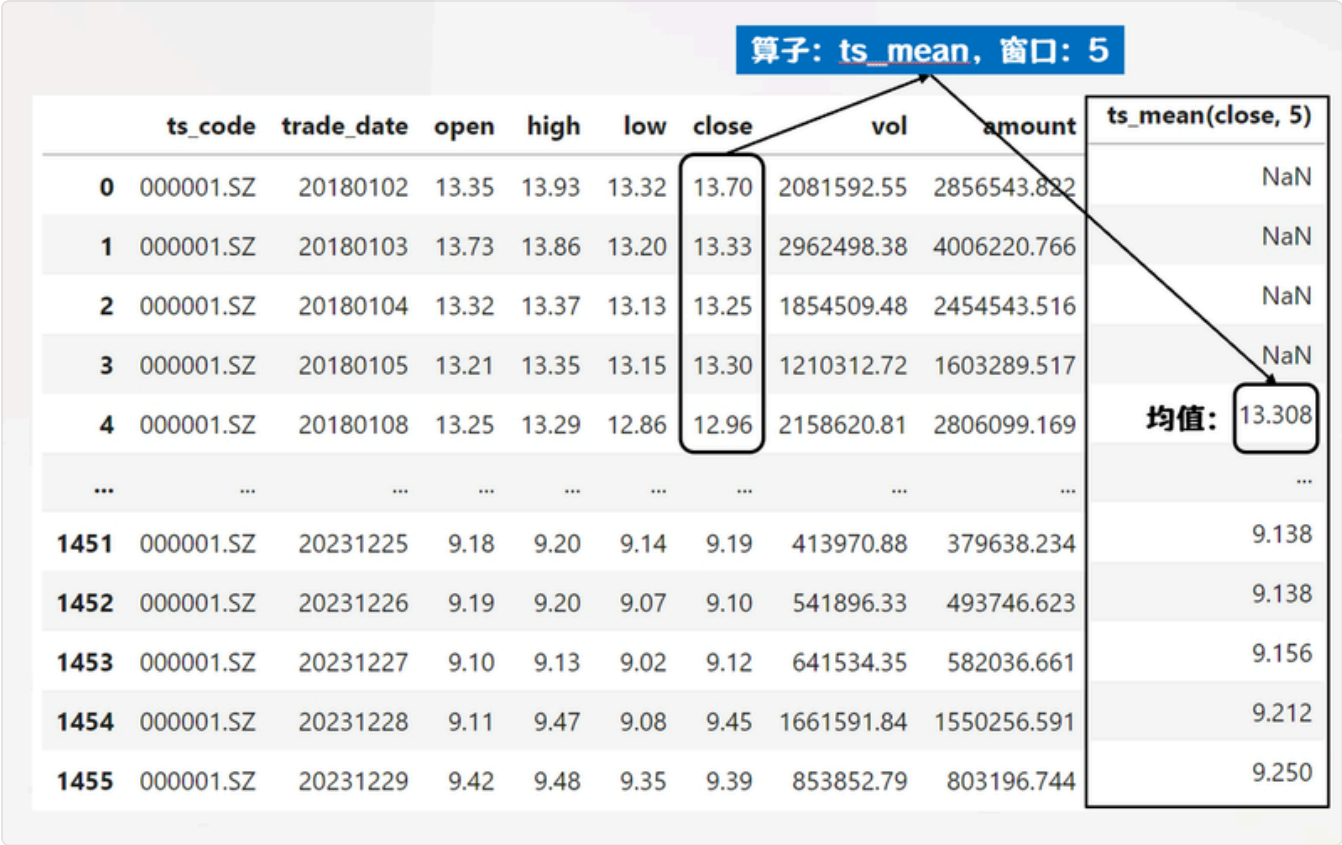
这里提供为大家了一个关于时间的对照表供大家参考：

接着我们看如何吸收历史特征信息，在刚才介绍时间序列建模的数据提到，要回望当前时刻过去的特征信息去预测Y，**那么需要回望多少呢，其实这需要根据业务的具体情况而定。**

比如在金融投资中的股票量价数据，过去5天、10天的数据包含了短期市场信息，20天、60天则是中期，120天和240天则是长期，240天在股票市场中也接近一年的交易天数。但是对于水利场景中的水位预测，可能就没必要考虑这么多天，过去1、2、3、4、5这5个回望窗口的信息就足够了。

因此我们也将这个配置开放给用户自己去定义，用户需要输入一串连续递增的数字，来表示不同的时间窗口，并使用英文逗号隔开。自动化特征工程将据此回望吸收过去的特征信息，如5,10分别对应着过去5条、10条数据的特征。需要注意的是**所填数字表示按时间顺序排列的先前数据条目数**，如果设置了时间组标识，则会考虑相同标识下之前的数据条目数。

比如：我们可能搜索出过去5天的收盘价平均值：**ts_mean(close, 5)**，还会结合窗口生成可能：**ts_mean(close, 10) 以及 ts_mean(close, 20)**等等，并会随着搜索深度的提升，对特征进行更深层的嵌套构造（更多算子可以参考：**AutoFE算子手册**中的**时序算子**部分）。



	ts_code	trade_date	open	high	low	close	vol	amount	ts_mean(close, 5)
0	000001.SZ	20180102	13.35	13.93	13.32	13.70	2081592.55	2856543.822	NaN
1	000001.SZ	20180103	13.73	13.86	13.20	13.33	2962498.38	4006220.766	NaN
2	000001.SZ	20180104	13.32	13.37	13.13	13.25	1854509.48	2454543.516	NaN
3	000001.SZ	20180105	13.21	13.35	13.15	13.30	1210312.72	1603289.517	NaN
4	000001.SZ	20180108	13.25	13.29	12.86	12.96	2158620.81	2806099.169	13.308
...	...	...	...	...	...	...	...	...	...
1451	000001.SZ	20231225	9.18	9.20	9.14	9.19	413970.88	379638.234	9.138
1452	000001.SZ	20231226	9.19	9.20	9.07	9.10	541896.33	493746.623	9.138
1453	000001.SZ	20231227	9.10	9.13	9.02	9.12	641534.35	582036.661	9.156
1454	000001.SZ	20231228	9.11	9.47	9.08	9.45	1661591.84	1550256.591	9.212
1455	000001.SZ	20231229	9.42	9.48	9.35	9.39	853852.79	803196.744	9.250

此外，在时序数据建模中，面对外部环境不断变化的场景，在验证集分割方面仅支持**根据时间进行比例分割的方法，防止随机抽样和分层抽样导致数据的时间间隔不连续。**这种方法能够有效应对时间变化导致的数据动态性，允许模型在不同时间段上进行训练和验证，更好地反映真实环境下的变化情况。

比如：在信贷风控的预测用户未来是否逾期的场景中，工程师一般会根据时间将数据升序排列，然后选择某一个时点，将数据切分成远期和近期两个部分，接着用远期的数据训练、近期的数据验证，以此来评估过去的某种行为规律在当前是否仍然还适用。

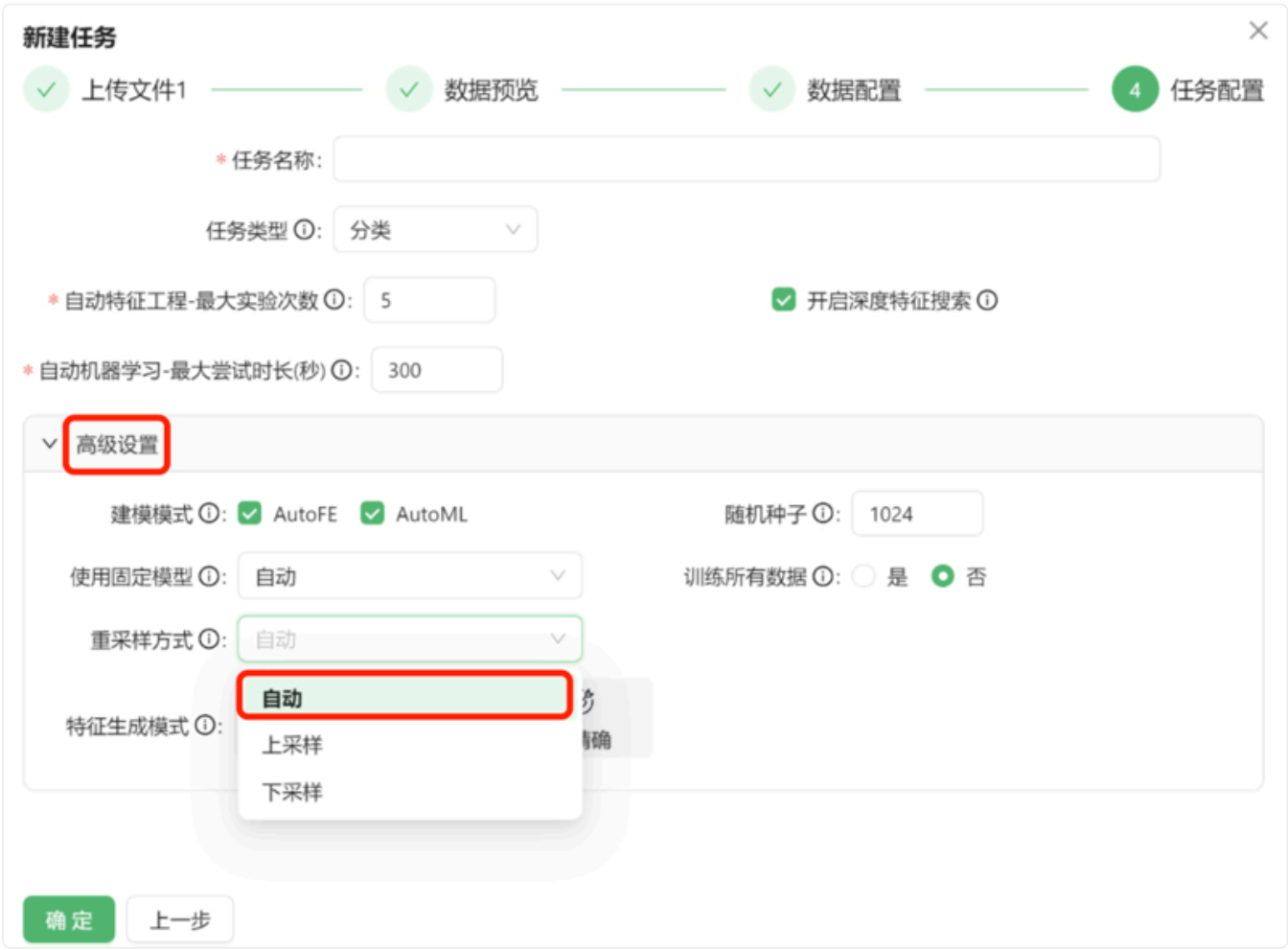
为了兼容其他模式的分割比例参数，本平台对于时间分割模式下的分割比例也是采用0到1的小数输入形式，默认是0.25，即代表75%的远期数据用于训练，25%的近期数据用于预测。用户可以根据实际情况手动调整。

重采样

在机器学习的分类任务中，经常会出现数据集标签样本不均衡的问题，尽管我们可以用**roc_auc**等方法（详见：**模型评价指标**）去合理评估样本不均衡下的建模结果，但是在模型训练时还是会更多吸收标签样本较多类的数据信息，因此我们仍可以有很多选择去调整数据集分布，一般包括上采样、下采样这两种方法，它们又统称为**重采样（Oversampling）**。

- **上采样（Upsampling）**：通过增加样本数量来平衡数据集中不同类别的样本，比如复制、SMOTE合成等。
- **下采样（Downsampling）**：通过减少某个类别的样本数量来平衡数据集，比如随机或选择性删除等。

由于下采样可能会损失样本的有效信息，上采样的合成方法造出的结果也是非真实的数据，加上不同建模任务的标准也不同。目前用户可以在高级设置中选择自动、上采样、下采样这3种采样模式，默认是自动，如下图：



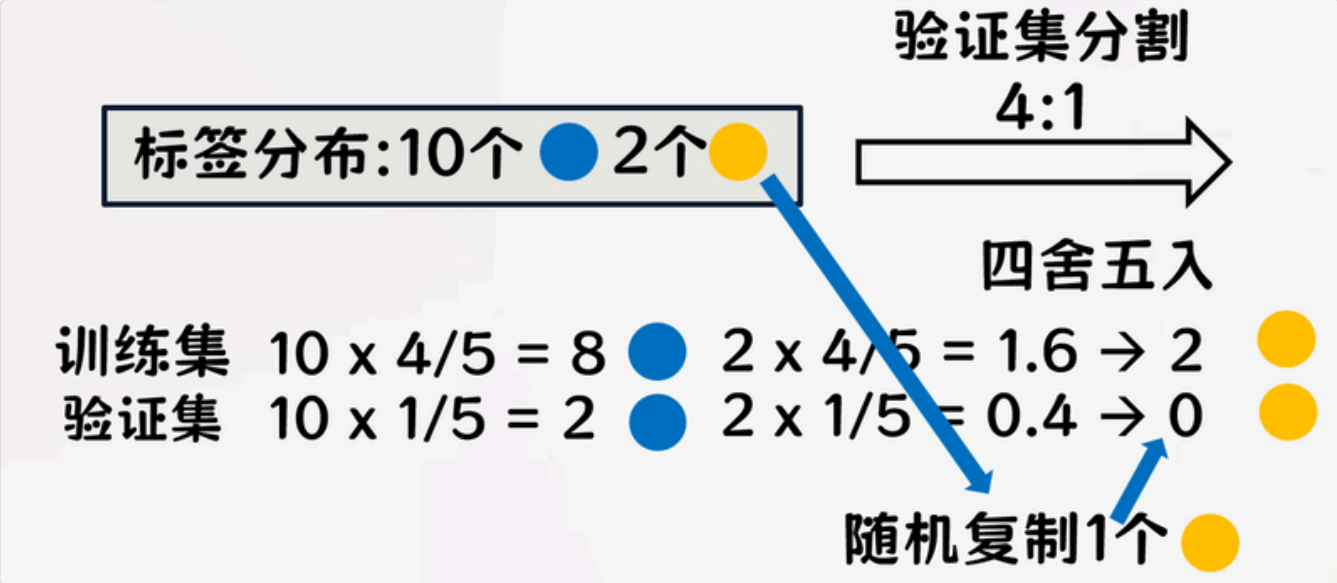
如果当前任务类型是回归任务，则不会进行采样。如果当前任务类型是分类任务，而大多数分类标签是不均衡的，例如：有的标签中True和False的比例是100:10。

① 如果选择上采样，则会将False标签对应的数据通过复制扩充到100条，变成100:100。

② 如果选择下采样，则会把True标签对应的数据采样10条，变成10:10。

③ 如果选择自动，则不会触发采样操作，除非最小类别对应的数据非常少，在乘上验证集分割比例后的数据条数小于1条，此时会对最小类别数据进行复制，以满足分割后有整数条数据的条件。

除非用户的数据集在分割数据集之后，由于数据点样本过少，出现了验证集中不存在少标签样本的情况，此时会触发平台的重采样机制，该过程会将用户该类别样本进行复制，使得验证集至少可以有一个该类别的样本。模型最终结果可能会受到些许影响。

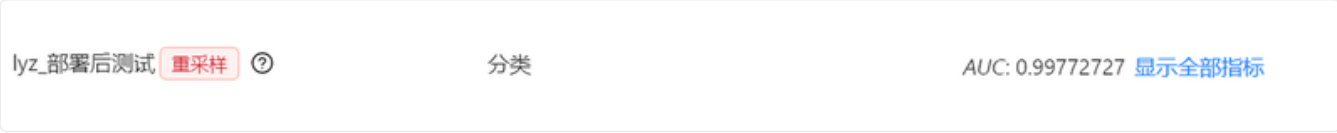


在这种情况下，平台会提示触发重采样的建模结果，此时建议有两种方式进行解决：

方法一：最优的解决方法就是您可以收集到足量的数据进行训练。

方法二：您可以调整高级设置中的验证集分隔比例，使得至少有一个样本可以被分到验证集。

同时触发重采样会在建模任务列中对应的任务出现“重采样”提示，如下图：



上一页
AutoFE算子手册

下一页
模型评价指标

⚙️ 模型评价指标

> **评价指标是**用于评估和比较模型性能的关键工具。该评价指标是根据模型对验证集的预测以及验证集的真实标签的差异计算出来。

一.分类

Log Loss

> 该指标也称为**对数损失**或**逻辑损失**（Logarithmic Loss），是评估分类模型性能的一个重要指标，特别是在二元分类和多类别分类任务中，它衡量的是模型预测概率的准确性。

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)]$$

- N：样本数量。
- y_i：样本i的实际标签。
- p_i：样本i被预测为类别1的概率。
- log：自然对数。

Accuracy

> 该指标称为**准确率**，它衡量的是模型预测正确的样本数占总样本数的比例。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- TP（True Positives）：真正例的数量，即模型正确预测为正类的样本数。

- TN（True Negatives）：真负例的数量，即模型正确预测为负类的样本数。
- FP（False Positives）：假正例的数量，即模型错误地将负类预测为正类的样本数。
- FN（False Negatives）：假负例的数量，即模型错误地将正类预测为负类的样本数。

Precision

> 该指标称为**精确率**，它衡量的是在所有被模型预测为正类的样本中，实际为正类的样本所占的比例。
>
> 简单来说，它回答了这个问题："在所有被预测为正的样本中，有多少是正确的？"。

$$Precision = \frac{TP}{TP + FP}$$

- TP（True Positives）：真正例的数量，即模型正确预测为正类的样本数。
- FP（False Positives）：假正例的数量，即模型错误地将负类预测为正类的样本数。

Recall

> 该指标称为**召回率**，它衡量的是在所有实际为正类的样本中，被模型正确预测为正类的样本所占的比例。
>
> 简单来说，它回答了这个问题："在所有真正的正样本中，有多少被模型正确地识别出来了？"

$$Recall = \frac{TP}{TP + FN}$$

- TP（True Positives）：真正例的数量，即模型正确预测为正类的样本数。
- FN（False Negatives）：假负例的数量，即模型错误地将正类预测为负类的样本数。

F1 Score

> 该指标为**精确率**（Precision）和**召回率**（Recall）的调和平均数，因此同时考虑了模型预测的精确性和完整性。

$$f_1 = \frac{2TP}{2TP + FP + FN}$$

- TP（True Positives）：真正例的数量，即模型正确预测为正类的样本数。
- FP（False Positives）：假正例的数量，即模型错误地将负类预测为正类的样本数。
- FN（False Negatives）：假负例的数量，即模型错误地将正类预测为负类的样本数。

AUC-ROC

> **ROC**代表接收者操作特征（Receiver Operating Characteristic），而**AUC**代表曲线下面积（Area Under the Curve）。
>
> 这个指标衡量模型区分类别（通常为“正类”和“负类”）的能力。

（1）**计算TPR和FPR**：通过改变分类阈值，计算在每个阈值下的真阳性率（TPR）和假阳性率（FPR）。

- **横轴**：假阳性率（False Positive Rate, FPR）：$FPR = \frac{FP}{FP+TN}$。
- **纵轴**：真阳性率（True Positive Rate, TPR）：$TPR = \frac{TP}{TP+FN}$。
- 通过改变分类阈值来计算不同的TPR和FPR，绘制成曲线。

（2）**绘制ROC曲线**：以FPR为横轴，TPR为纵轴。

（3）**计算AUC**：计算ROC曲线下的面积。这通常通过数值方法（如梯形规则）实现。

二.回归

R-square

> 该指标称为**决定系数**，是回归任务中常用的评价指标之一。它衡量的是模型预测值的变异性与实际值（目标变量）的变异性之间的比例。

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- n：样本数量。
- y_i：实际值。
- $\hat{y}_i$：预测值。
- $\bar{y}$：实际值的平均值。

MSE

> 该指标称为**均方误差**（Mean Squared Error），是回归任务中常用的评价指标之一。它衡量的是模型预测值与实际值之间差异的平均大小。
>
> MSE通过计算预测值与实际值之差的平方的平均值来实现这一点。这个指标对异常值（即那些偏离实际值很远的预测值）非常敏感，因为差异的平方会放大这些值的影响。

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- n：样本数量。
- y_i：实际值。
- $\hat{y}_i$：预测值。

RMSE

> 该指标称为**均方根误差**（Root Mean Square Error），是回归任务中常用的评价指标之一。
>
> RMSE 测量的是模型预测值与实际观测值之间差异的标准偏差。它是均方误差（Mean Square Error, MSE）的平方根，用于衡量预测误差的大小。

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

-

- n：样本数量。
- y_i：实际值。
- $\hat{y}_i$：预测值。

MAE

> 该指标称为**平均绝对误差**（Mean Absolute Error），是回归任务中常用的评价指标之一。
>
> 它测量的是模型预测值与实际观测值之间差异的平均绝对值。MAE 提供了一个直观的误差度量，表示预测值与实际值之间的平均绝对偏差。

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- n：样本数量。
- y_i：实际值。
- $\hat{y}_i$：预测值。

上一页
词汇和术语

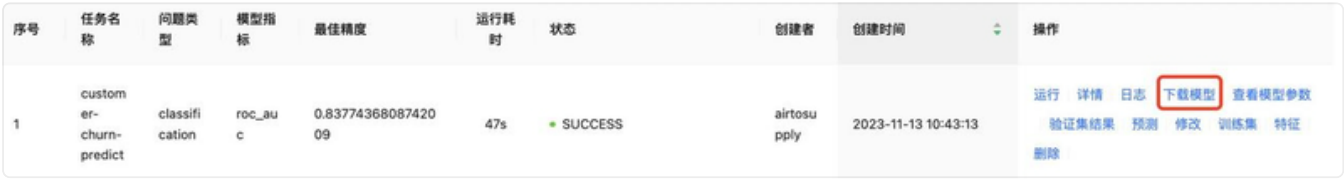
下一页
高级特征

最后更新于3天前

📸 模型下载

如何下载模型?

用户可以点击建模任务列中对应任务的 【下载模型】 按钮下载该任务训练完成的模型。如下图:



模型文件结构

当用户下载模型文件之后会获取到一个名称为model_*.zip的压缩文件，该压缩文件中记录了所有关于本次建模的基本信息和相关元数据。

通过解压工具解压可获取一个名称为result的文件夹，如下图:

文件夹中大致可分为如下几个部分:

- **example.py文件**：该文件为本平台提供的机器学习框架的代码样例。

- **result文件夹**：该文件夹存放本次模型训练过程中的元数据等相关文件，包括自动化特征工程（AutoFE），自动化机器学习（AutoML）以及建模任务配置文件等相关信息。

> 下载完毕的模型可以通过本平台提供的预测框架库完成预测等相关工作。
>
> 本平台已经在Python公有仓库 PyPI 中发布了预测框架库 changtianml，通过该依赖库可以更加便携的方便用户完成预测等相关工作。目前该工具库仍然在不断迭代和完善中，敬请期待！

模型元数据

建模任务配置

关于本次建模过程中所涉及到的配置信息以及高级参数等均存放在config.yaml文件中。

自动化特征工程

文件名称	文件描述
features.csv	高级特征列表（Latex公式风格）
record.csv	高级特征列表详情（CSV格式）
result	高级特征列表详情（JSON格式）
labelencoder.pkl	自动化特征工程模型二进制文件
dfs_log	自动化特征工程阶段探索日志

自动化机器学习

文件名称	文件描述
result	自动化机器学习模型二进制文件
val_res	验证集预测结果文件
val_logs	验证集评估指标文件
logs	自动化机器学习阶段探索日志

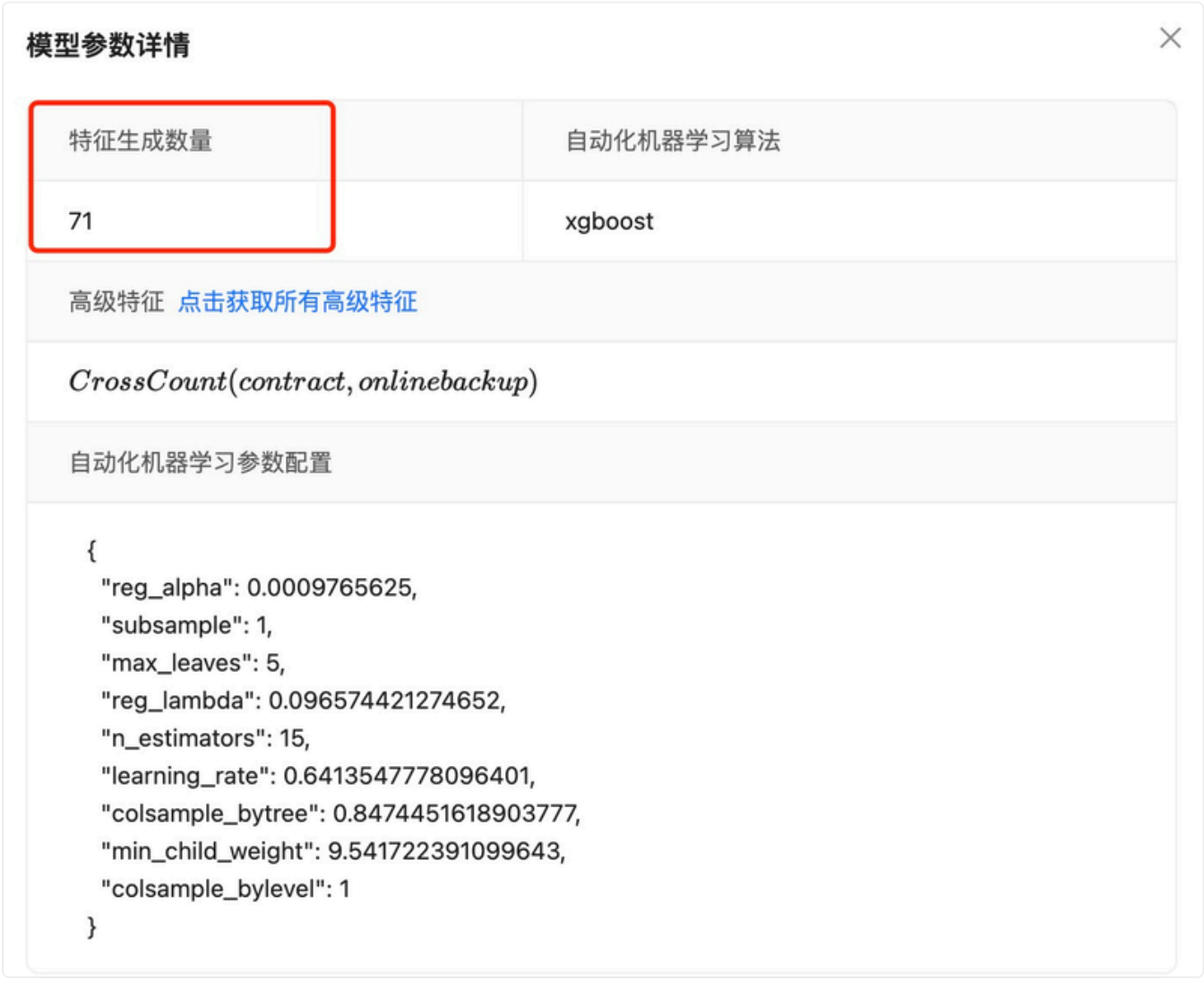
高级特征

什么是高级特征?

构造**高级特征**是更好发挥机器学习拟合能力的重要手段。构造的高级特征能从不同方向提炼出对原始数据集的高级解释，彼此同样重要。一起构成对模型精度增益最大的特征组。

如何查看高级特征?

用户可以点击建模任务列中对应任务的 【查看模型参数】 按钮查看训练模型的基本信息，如下图:



这里显示生成了71个高级特征，采用的自动化机器学习算法是xgboost。

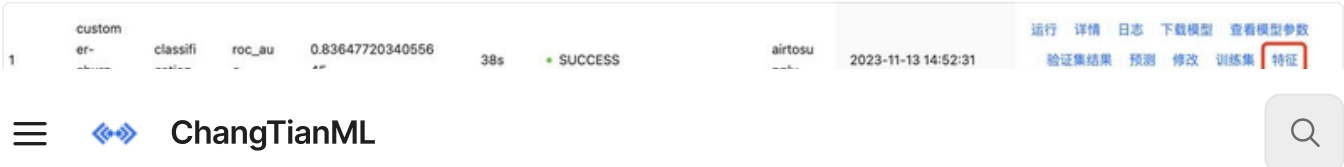
默认情况下系统智只会展示第一个高级特征，当特征生成数量大于1时，可以点击 【点击获取所有高级特征】 按钮获取所有的高级特征，系统会列出当前模型所有高级特征的具体表达形式，同时系统还提供高级特征下载用于后续的机器学习工作。

1	$CrossCount(contract, onlinebackup)$
2	$CrossCount(paymentmethod, techsupport)$
3	$CrossCount(contract, onlinesecurity)$
4	$AggMedian(tenure, totalcharges)$
5	$CrossCount(onlinebackup, streamingtv)$
6	$CrossCount(contract, streamingmovies)$
7	$CrossCount(contract, paymentmethod)$
8	$CrossCount(deviceprotection, monthlycharges)$
9	$CrossCount(monthlycharges, onlinebackup)$
10	$CrossCount(dependents, paperlessbilling)$

< 1 2 3 4 5 ••• 8 > 10 条/页

下载

与此同时，用户也可以点击建模任务列中对应任务的 【特征】 按钮直接跳转到高级特征详情页面，即可查看。



上一页
模型评价指标

下一页
模型下载

模型报告

如何获取模型报告?

本平台不仅可以将训练完成的模型下载下来进行后续的机器学习工作，还可以通过可视化的方式更加直观的展示当前模型的相关重要信息，这对机器学习模型的研究提供了重要依据。

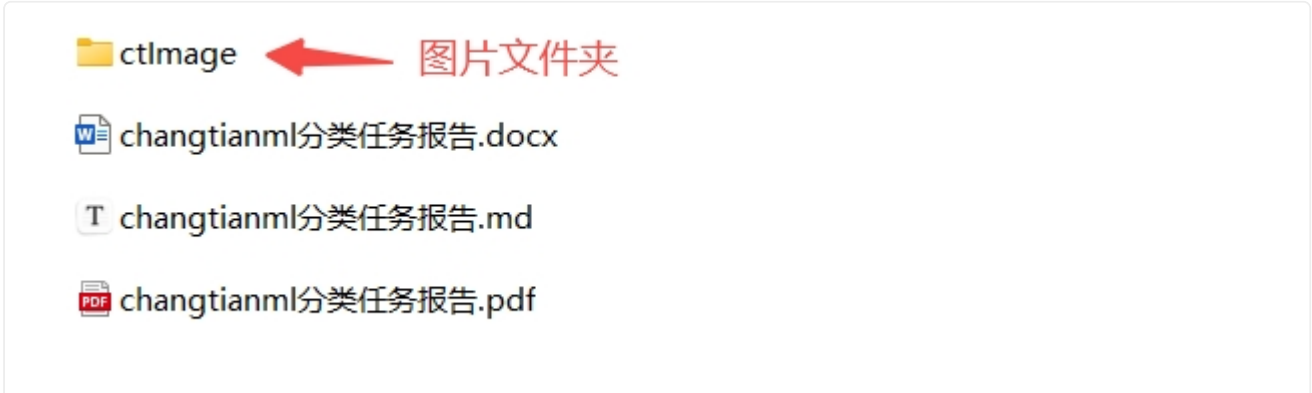
用户可以在建模列表中对应的建模任务中点击 【下载报告】 即可获取模型报告，如下图：

下载之后可以获取到一个名称为report_*.zip压缩文件，通过解压工具解压之后即可获取模型报告内容。

模型报告下载示例可参考：《附录-2：模型报告示例》

模型报告解读

本平台会自动为用户生成压缩之后的模型报告，解压缩之后的内容大致如下：



其中包括了原始图片文件夹、以及不同格式的报告文件（支持Word，PDF和Markdown）。

下面我们会通过一个分类任务的模型报告为大家如何查阅报告中的相关内容。

本平台提供的报告包括了训练数据的**特征空间分布**、**模型可视化**、**验证数据**的各类指标三个方面。

一、特征空间分布

高级特征是从不同的方向更深入地解释数据。下面使用的几种描述特征指标的手段不能衡量出特征表达能力的全部维度，特征彼此之间并不是替代关系，拥有同等的重要性。

特征空间分布可以简单分为两个方面，可以简单地分为训练前和后。训练前，通过特征空间的分布，从而去选择合适的特征和模型 进行训练。例如可以通过大部分特征是否有线性关系去选择模型。训练后，如果使用的是树模型的话，我们将得到每个特征的贡献值，可以通过这个贡献程度来反复调整我们训练的模型，也可以通过特征重要性来体现生成特征的有效性。

出于报告篇幅和简洁性的考虑，我们这里只对**处理过后的数据**进行分析，并只输出部分数据的。

如需要对**原始数据**或者**编码**分析，那么可以使用C**hangtianML工具**进行分析。

1. 训练前的数据分析

1.1 热力图 (Heat Map)

热力图，表示的是**特征之间的相关性**

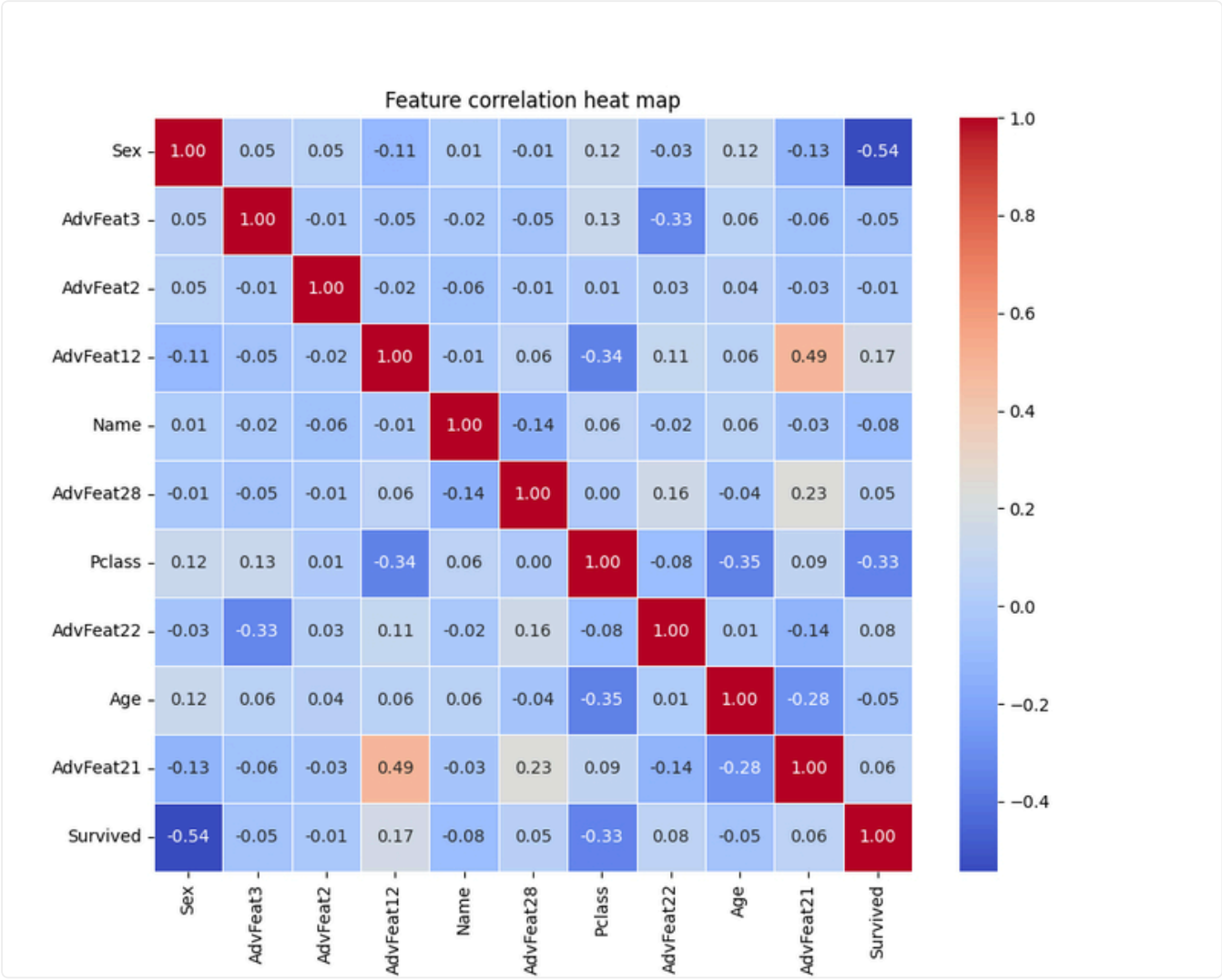
这里展示的是模型**特征重要性前10**的特征与标签的相关性

这里计算的相关系数是皮尔逊相关系数（Pearson correlation coefficient），其公式如下

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

皮尔逊相关系数衡量了两个变量之间的线性相关性，其取值范围为-1到1。具体而言：

- 1 表示完全正相关：一个变量增加，另一个变量也以相等的比例增加。
- 0 表示无相关性：两个变量之间没有线性关系。
- -1 表示完全负相关：一个变量增加，另一个变量以相等的比例减少。



1.2 特征核密度图 (Kernel Density Estimation, KDE)

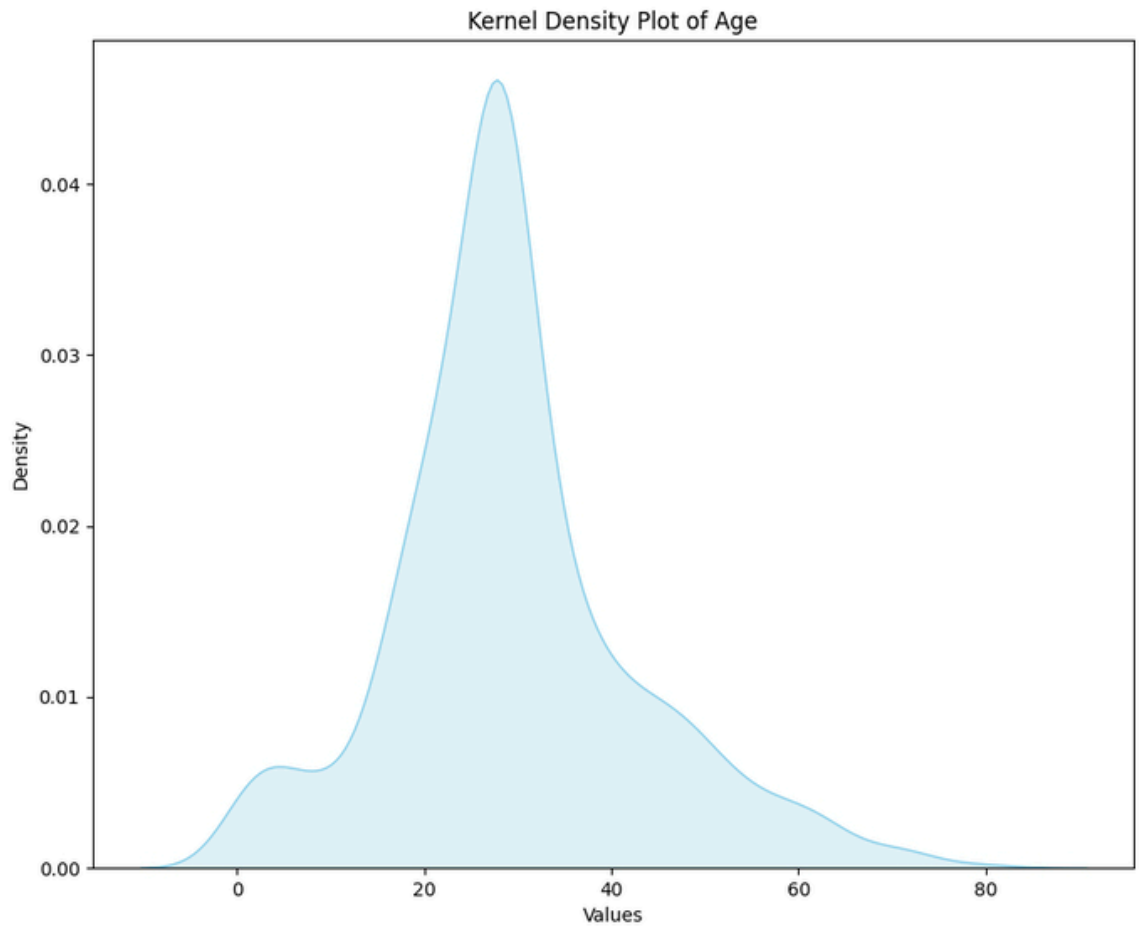
> 核密度估计（Kernel Density Estimation，简称KDE）是一种用于估计概率密度函数的非参数方法。它通过在每个数据点周围放置一个核（通常是正态分布的核），然后将这些核叠加起来，形成平滑的估计概率密度函数。
>
> 目前我们选择的是特征贡献度最大的高级特征或者原始特征进行核密度估计。

核密度估计图可以帮助你了解**单个变量**的**分布情况**，包括峰值和分布的形状。可以从以下二个方面进行对数据的分析。

1. **峰值（Peak）：** 峰值表示概率密度估计图中最高点的高度。
   - 在KDE中，峰值**越高**，表示在该位置的**数据点密度越大**，即该位置附近的数据点更为集中。

- 峰值的高度并不直接给出概率值，但可以用于比较不同位置的数据密度。

2. **整体形状：** 概率密度估计图的整体形状反映了数据分布的趋势。
   - 图形的平滑性和波动性可用于判断**数据的变异性**。
   - 例如，平坦的KDE图表明数据相对均匀地分布，而具有峰值和波动的图形可能表示数据在某些区域更为密集。



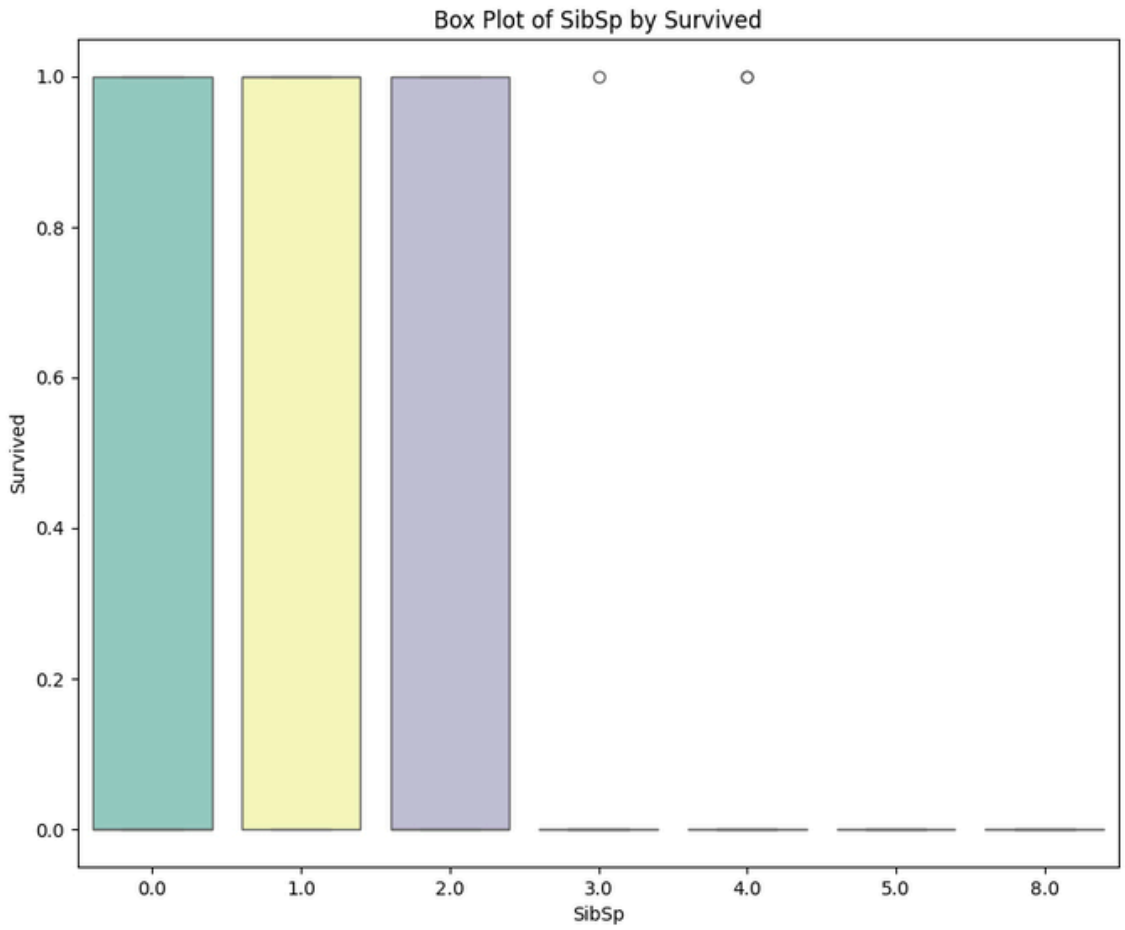
1.3 箱线图 (Box Plot)

> 箱线图是一种用于可视化数据分布和离群值的有效工具
>
> 目前我们选择的是特征贡献度最大的高级特征或者原始特征进行箱线图的绘制。

在分析箱线图时，你可以关注以下几个关键的要素：

1. **箱体（Box）：** 箱体显示了数据的四分位数范围，即数据的中间50%。
   - 箱体的底部和顶部分别表示第1四分位数（Q1，下四分位数）和第3四分位数（Q3，上四分位数），而箱体内部的线表示中位数（Q2）。

2. **须（Whiskers）：** 须延伸从箱体的两端，表示数据的最大值和最小值，但不考虑异常值。
   - 须的长度通常基于数据的分布，具体的计算方式可能有所不同。
3. **异常值（Outliers）：** 在箱线图中，通常将超出须的1.5倍四分位距的数据点定义为异常值，并用点表示。
   - 异常值可能是数据集中的离群值。
4. **整体形状：** 观察箱线图的整体形状可以提供关于数据分布的信息。
   - 例如，箱体的长度和位置，须的延伸情况，以及异常值的分布等。

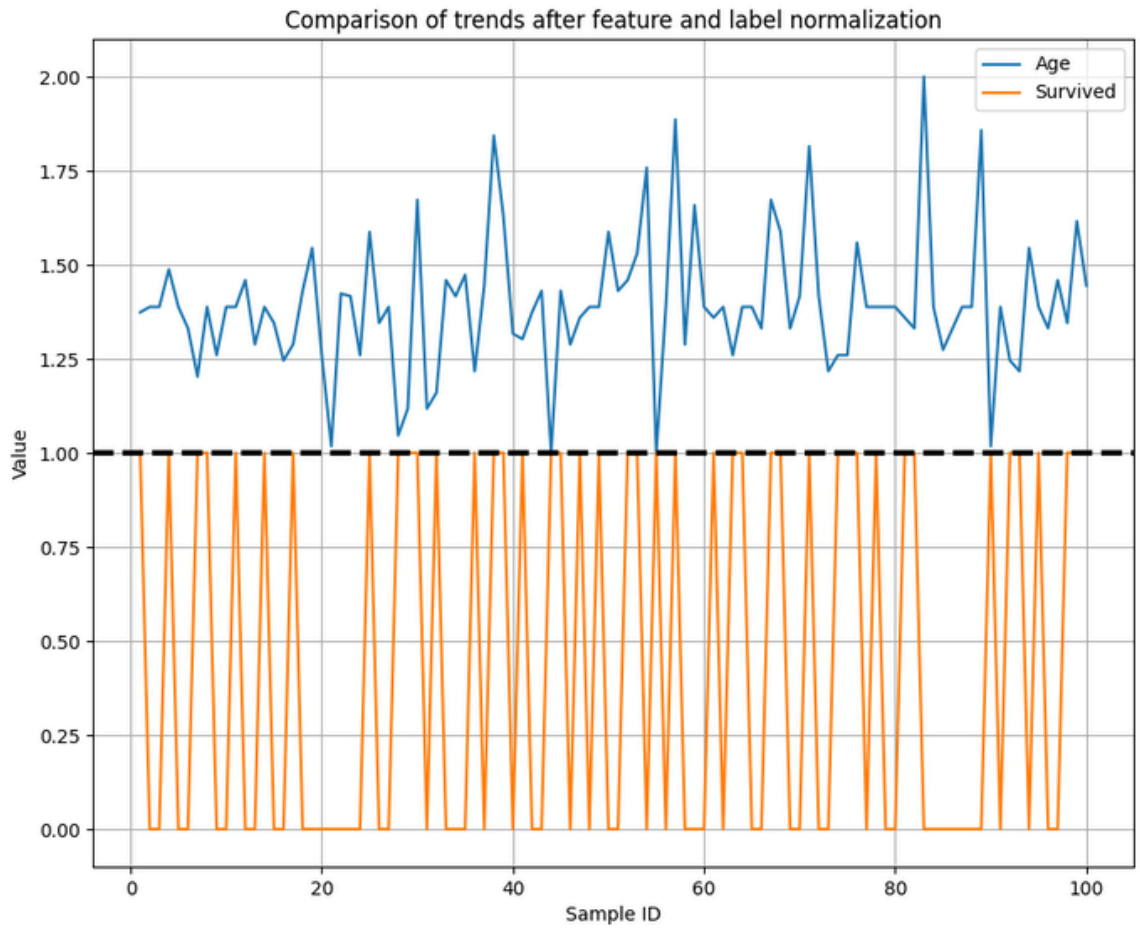


1.4 特征和标签的特征曲线图

查看特征和标签的趋势变化是否一致有助于构建更有效、解释性更强、泛化性能更好的机器学习模型。

目前我们选择的是特征贡献度最大的高级特征或者原始特征进行曲线绘制的，并选取前100个样本显示。

为了方便比较，我们将特征归一化到[1, 2]之间，把标签归一化到[0, 1]之间。



2.训练后的特征分析

2.1 特征重要性

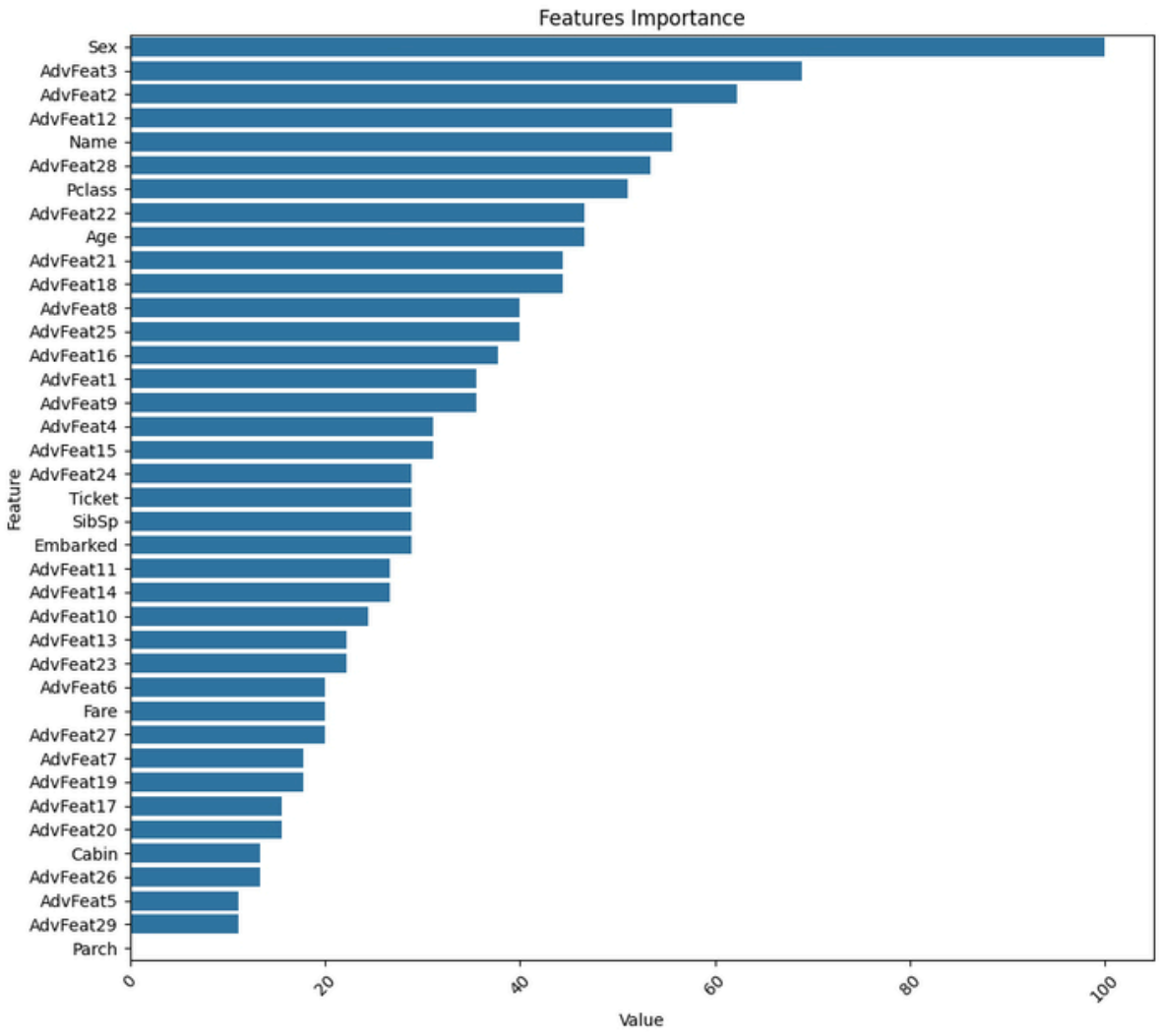
每个特征对于树模型提供的贡献程度，以下是常见树模型的特征重要性分数的简介。

1. **LightGBM:**
   - LightGBM使用基于树的学习算法，特征重要性主要基于**分裂增益（Split Gain）**。
2. **CatBoost:**
   - CatBoost也是基于树的学习算法，其特征重要性计算方式与LightGBM相似，主要基于**分裂增益**。
3. **XGBoost:**
   - XGBoost通过计算特征的**增益（Gain）**来评估特征的重要性。
   - XGBoost还提供了一种基于覆盖度（Coverage）的特征重要性计算方式，覆盖度表示每个特征在树中的使用频率。
4. **随机森林:**

- 随机森林使用**基尼不纯度（Gini impurity）**或信息增益（Information Gain）等指标来选择最佳的分裂特征。
- 这些次数的总和或者平均值可以用于衡量特征的重要性。

5. **极限树（Extra Trees）：**
   - 极限树是随机森林的一种变体，它在节点分裂时使用了更多的随机性。
   - **特征重要性的计算方式与随机森林类似。**

> 树模型在分裂的时候，可能会多次用到同一个特征，其特征重要性是这几个节点加权平均，不同模型的策略不同。



3. 如何数据分析

机器学习的一般**人工构建流程**如下：

（1）**数据分析**，了解数据的分布情况，例如：

- **查看是否有冗余：**一般情况下，如果**热力图**中特征之间的相关系数绝对值大于0.8，就应该慎重考虑特征选择。
- **查看特征和标签的相关性：**一般情况下，如果**热力图**中特征和标签之间的相关系数越高特征重要性可能越高。
- **查看特征的概率密度分布：**判断数据是否密集，密集可能导致区分度很小。
- 其他方法

（2）**特征工程**

通过第一步的数据分析进行特征的生成，例如：**特征A**和**特征B**相关度很低，那么将**特征A**和**特征B**简单的组合（比如加减乘除）得到**特征C**，那么**特征C**可能对于模型的帮助会很大。

举一个简单的例子，例如我们评测身体的肥胖程度。我们有身高h，体重w这两个特征，通过特征组合发现了$w \div h^2$这样一个特征。

这个特征对于分类有重大的贡献，那这就是**BMI**，给他赋予身体质量指数的物理含义，用于衡量人体肥胖程度。

（3）**模型调优**

模型调优主要分为两部分：**选择最合适的模型**和**选择最优的模型参数。**

特征工程和模型调优一般是一起的，可以简单理解，不同的数据集，那么就应该对应不同的最优模型。那这两个步骤将需要花费大量的时间和精力去实现。

> **使用ChangtianML的数据分析：**
>
> （1）您已经得到了有效的特征生成组合，那么可以反向推理**新特征的物理含义**，例如：以BMI为例（$w \div h^2$），那我们可以解释：为了考虑不同身高的肥胖程度我们采用了除法，平方项的引入使得该指标对于身高的变化更为敏感，从而更好地反映体重相对于身高的比例。
>
> （2）如果需要更加详尽的数据详情，那么可以使用C**hangtianML工具**进行分析。

二、模型可视化

> 模型可视化目前只支持LightGBM，如果最终模型不是LightGBM，那么可能没有生成这部分的图片。

其他模型的可视化正开发中，敬请期待！

1. 树模型的结构可视化

树模型可视化的重要性：

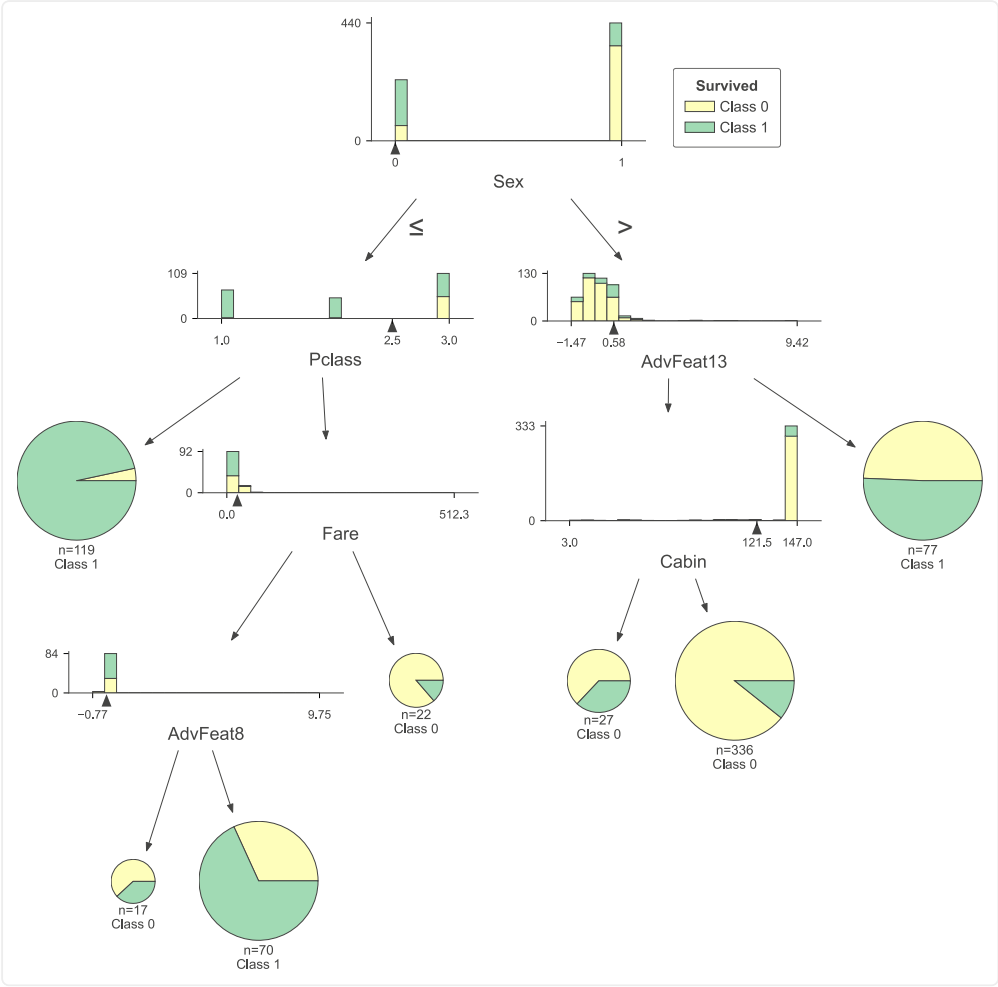
1. 树模型作为一种可解释的模型，你可以清晰地看到每个决策节点的分裂条件以及叶子节点的预测结果。这有助于理解模型是如何基于输入特征进行预测的，使得模型的工作过程更加透明和可解释。
2. 可以了解特征对于预测结果的影响，例如特征A在某个阈值下是否把数据分到正确的类别中了。
3. 可视化树模型有助于向非专业人员、利益相关者或团队成员解释模型的工作原理。

下面介绍多种树模型的显示方式。

树模型图的解释

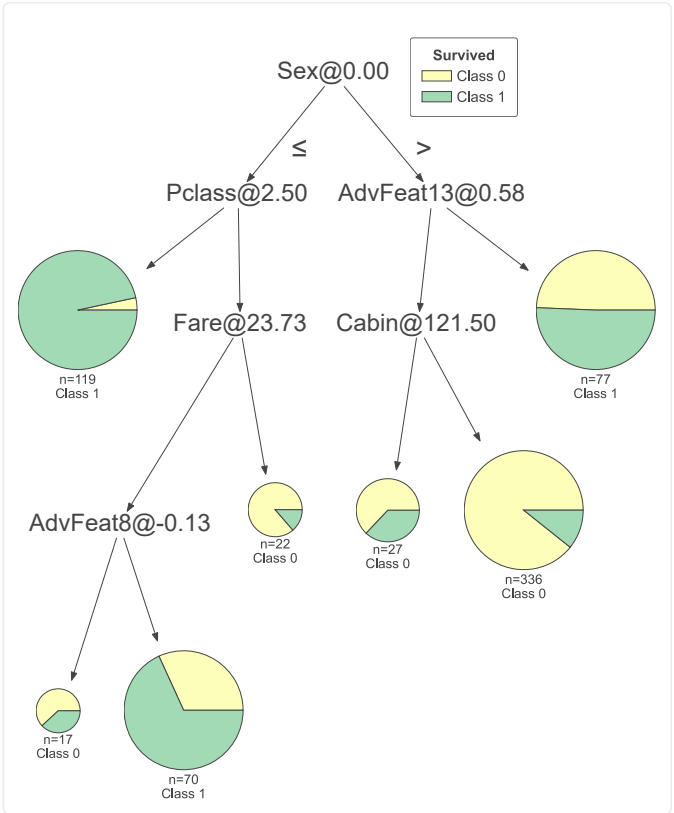
- 选择特征X_i，特征X_i会根据阈值S_i进行分裂，左边是小于等于阈值S_i，右边则反之。
- 如果分裂的样本中得到较好的结果，则停止分裂，反之继续选择特征X_i分裂。
- 重复步骤上面两个步骤直到所有样本都被分配完成。

（1）树模型的竖向展开图

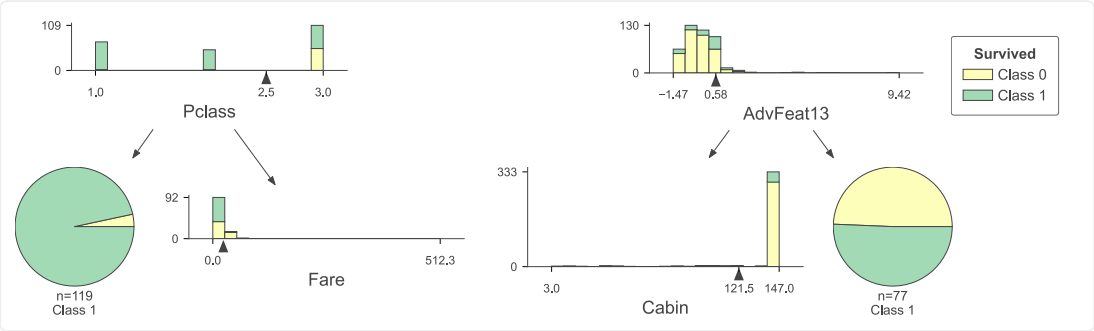


(2) 树模型的横向展开图

(3) 树模型的简化展开图



（4）树模型的浅层展开图

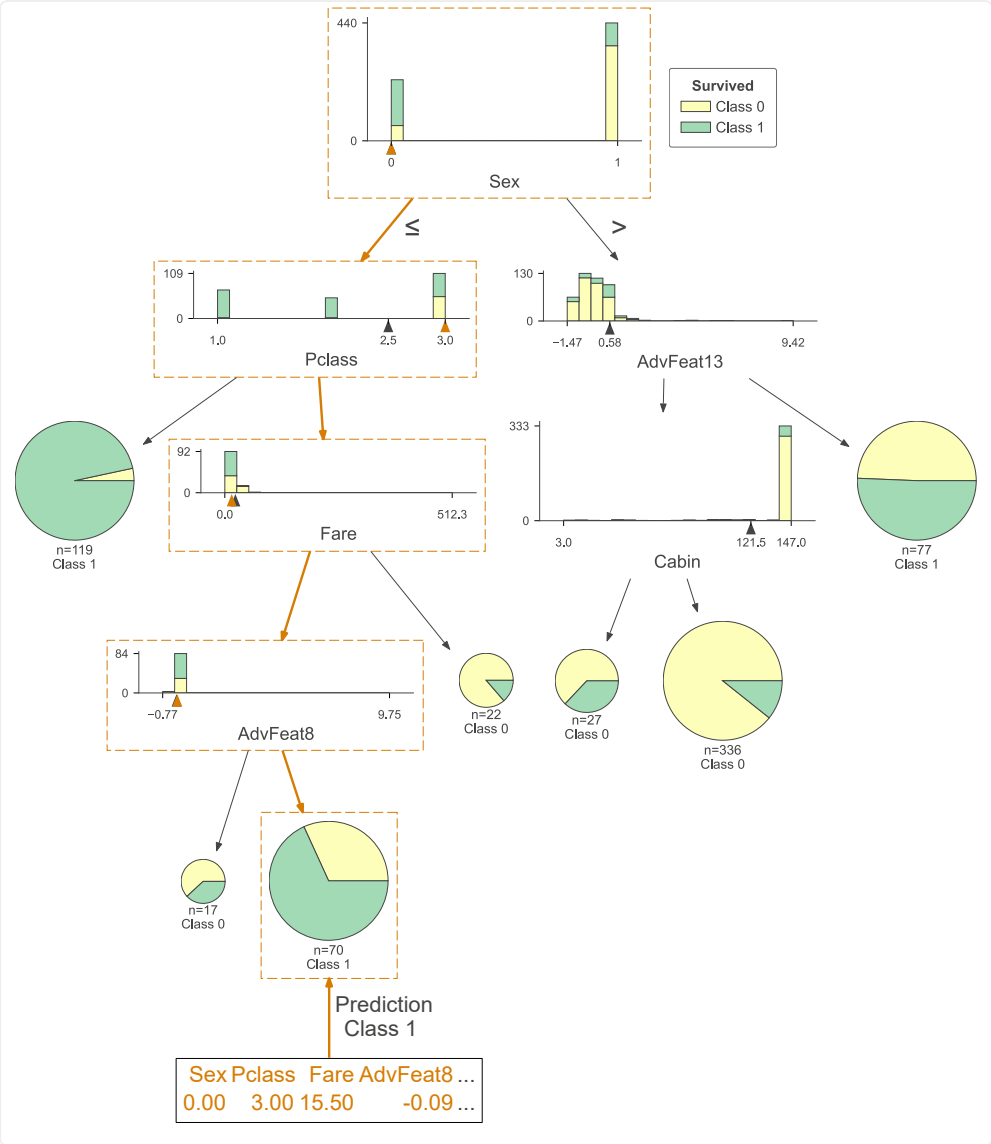


2. 决策树的预测路径

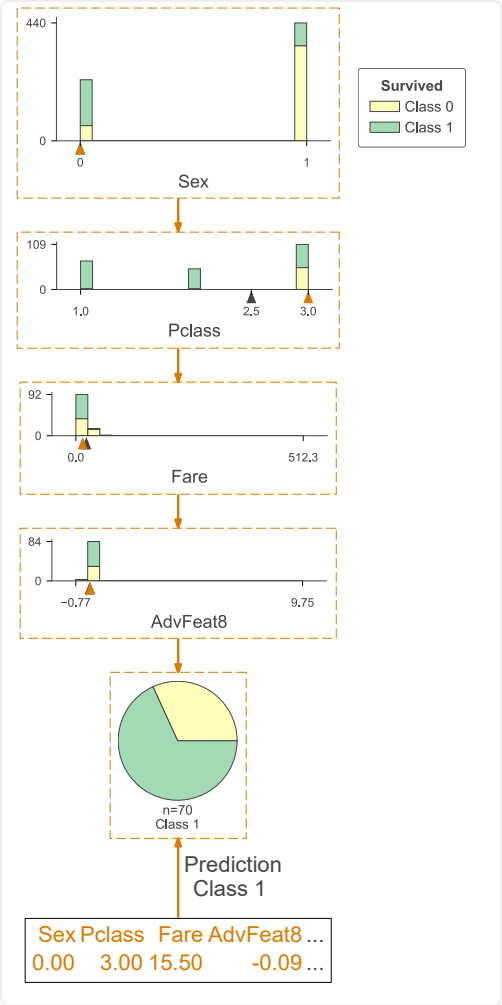
> 决策树的预测路径可以具象化看到模型是如何判断某个样本属于哪个类别的过程。

决策树的全部预测路径

预测路径已用橙色的框标示出来，随机抽一个测试集的样本并在下方列出样本的具体信息。

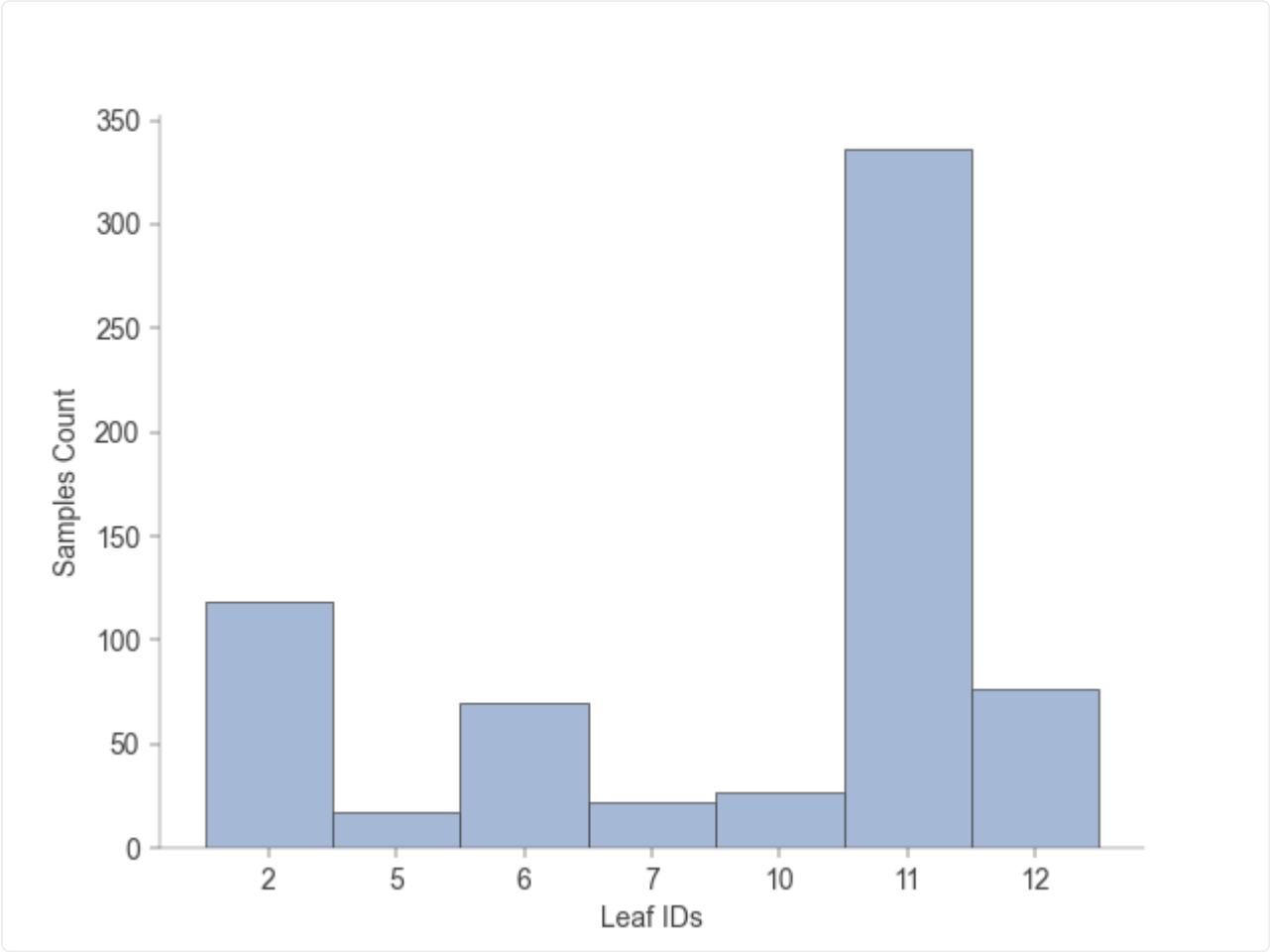

440
0
0
1
Sex
Survived
Class 0
Class 1
≤
>
109
0
1.0
2.5
3.0
Pclass
130
0
−1.47
0.58
9.42
AdvFeat13
92
0
0.0
512.3
Fare
n=119
Class 1
333
0
3.0
121.5
147.0
Cabin
n=77
Class 1
84
0
−0.77
9.75
AdvFeat8
n=22
Class 0
n=27
Class 0
n=336
Class 0
n=17
Class 0
n=70
Class 1
Prediction
Class 1
Sex Pclass Fare AdvFeat8 ...
0.00 3.00 15.50 -0.09 ...

决策树的预测路径简化

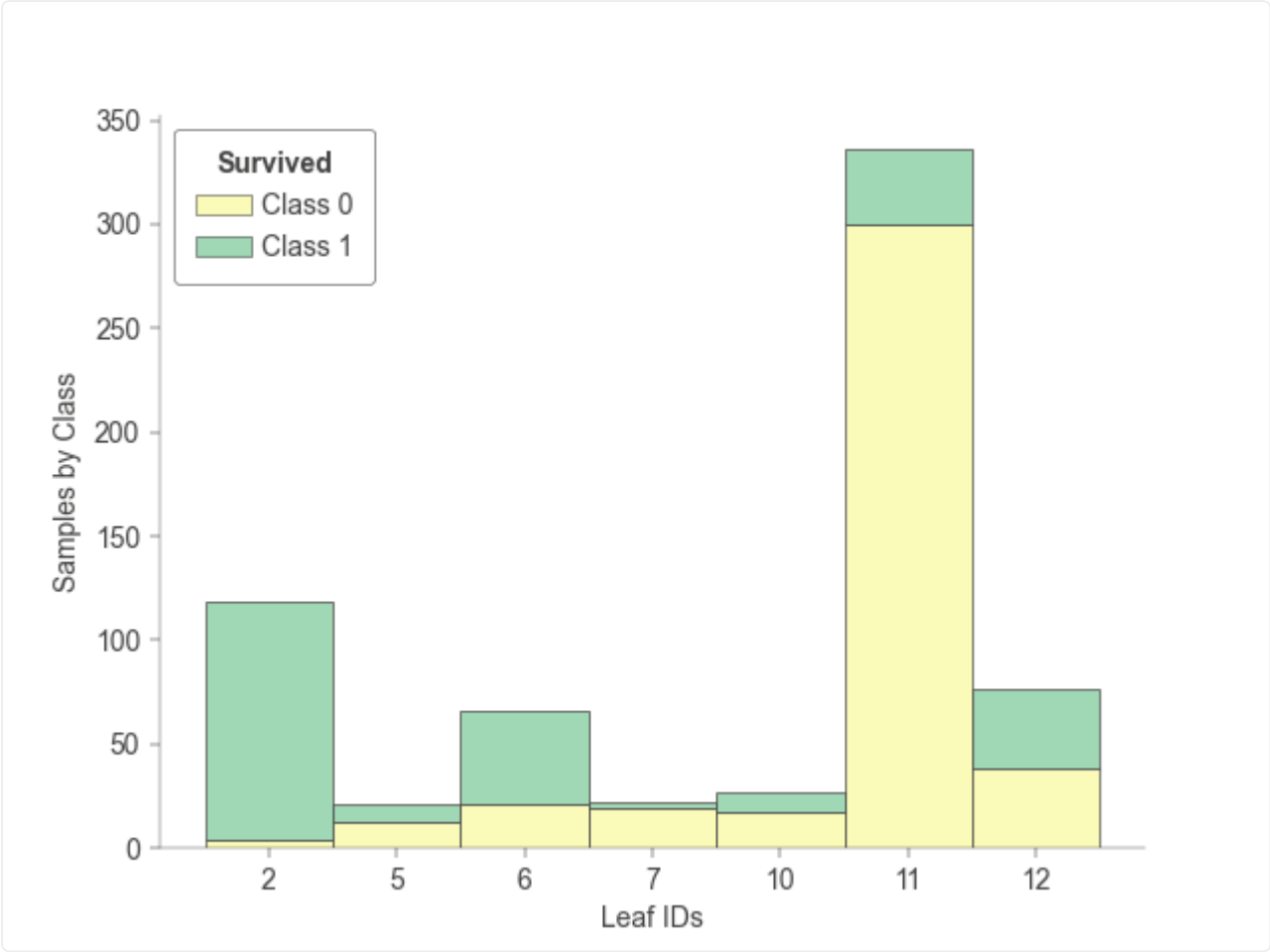


3. 树叶子节点样本树的统计

树模型的各个叶子节点训练的时候包含了多少样本的统计。



树模型的各个叶子节点训练的时候包含了多少样本的统计，以及每个叶子节点有多少属于各个类别。



三、验证集指标

ROC曲线和混淆矩阵目前只支持输出类别小于等于5种的分类任务。

1. ROC曲线

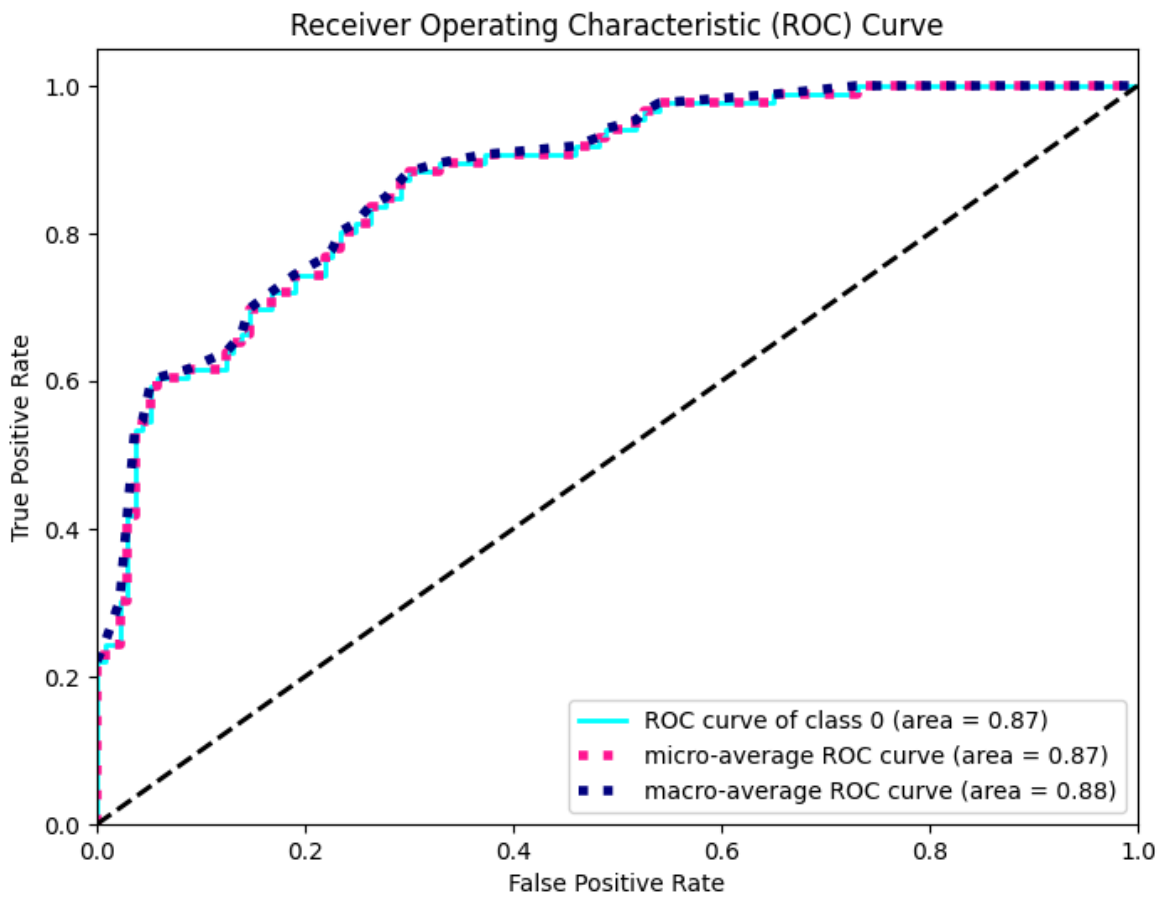
ROC曲线（Receiver Operating Characteristic Curve）是一种用于评估分类器性能的图形工具。主要关注两个重要的性能指标：True Positive Rate (**TPR**) 和 False Positive Rate (**FPR**)。

True Positive Rate (敏感度或召回率): TPR表示在所有实际正例中，模型成功识别为正例的比例。在ROC曲线上，TPR对应纵轴，范围从0到1。

False Positive Rate (假正例率): FPR表示在所有实际负例中，模型错误识别为正例的比例。在ROC曲线上，FPR对应横轴，范围同样从0到1。

意义和用途：

1. **模型比较：** ROC曲线提供了一种比较不同模型性能的可视化工具。曲线下的面积（AUC）用于量化不同模型在整个ROC空间下的性能，AUC越大，表示模型性能越好。
2. **阈值选择：** ROC曲线可以帮助选择分类阈值。不同应用场景可能对False Positive Rate和True Positive Rate有不同的关注程度，通过调整分类阈值，可以在这两者之间找到平衡。
3. **诊断性能：** ROC曲线能够展示模型在不同工作点下的性能，从而帮助你理解模型的诊断性能，特别是在二分类问题中。
4. **不受类别不平衡影响：** ROC曲线对于类别不平衡的问题更为鲁棒，因为它是基于真正例率和假正例率的比例关系。



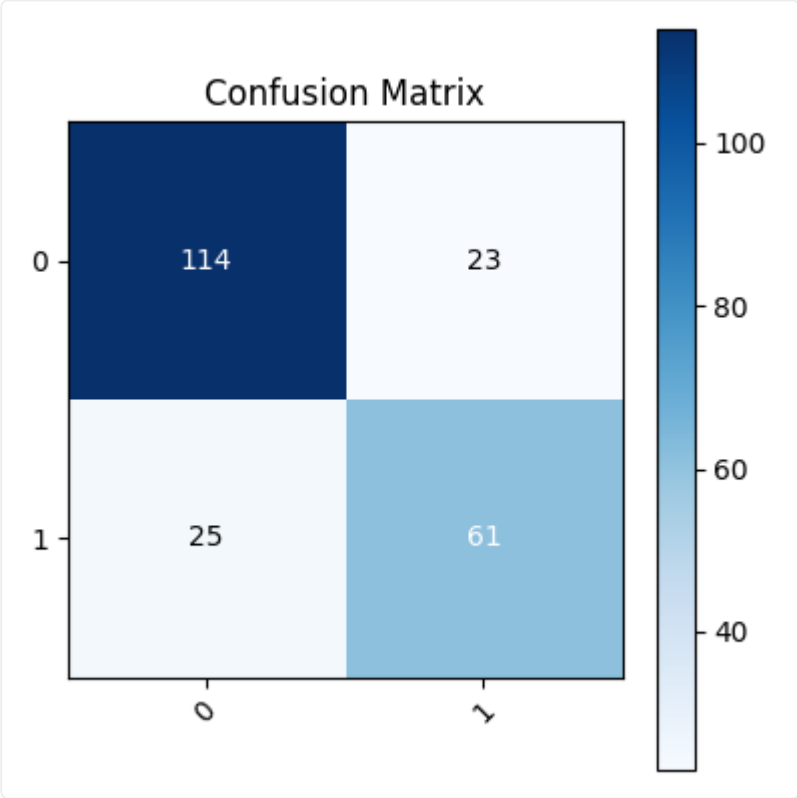
2. 混淆矩阵

验证集指标(Accuracy, Recall，Precision，F1)可以在ChangTianML任务的日志的AutoML得到。

1. **真正例（True Positives，TP）：** 模型正确地将正例样本预测为正例。
2. **真负例（True Negatives，TN）：** 模型正确地将负例样本预测为负例。
3. **假正例（False Positives，FP）：** 模型将负例样本错误地预测为正例。
4. **假负例（False Negatives，FN）：** 模型将正例样本错误地预测为负例。

混淆矩阵的意义和用途：

1. **性能评估：** 混淆矩阵提供了一个全面的性能评估，可以直观地了解模型在不同类别上的预测准确性。
2. **精确度（Accuracy）计算：** 精确度是分类器正确预测的总样本数占总样本数的比例，可以通过混淆矩阵计算。$Accuracy = \frac{TruePositives+TrueNegatives}{TotalSamples}$，该任务的准确率为: 0.7847533632286996。
3. **召回率（Recall）计算：** 召回率（也称为敏感度或真正例率）表示模型正确识别的正例样本在所有实际正例样本中的比例。$Recall = \frac{TruePositives}{TruePositives+FalseNegatives}$，该任务的召回率为: 0.7847533632286996。
4. **精确度（Precision）计算：** 精确度表示模型预测为正例的样本中，实际为正例的比例。$Precision = \frac{TruePositives}{TruePositives+FalsePositives}$，该任务的精确率为: 0.7839107317605237。
5. **F1分数计算：** F1分数是精确度和召回率的调和平均，用于综合考虑模型的准确性和全面性。$F1 = \frac{2\times Precision\times Recall}{Precision+Recall}$，该任务的f1为: 0.7842670856605463。



附录-1：高级特征与特征贡献度对应表

模型在可视化的时候捆绑特征可能看不清楚，可以通过下表查询相应的高级特征以及特征贡献度。

高级特征

- $AdvFeat1$：$(Ticket \div Fare) \div (cos(Ticket))$
- $AdvFeat2$：$(cos(Fare)) \div (Ticket \div Age)$
- $AdvFeat3$：$((cos(Fare)) \div (sin(Ticket))) \div (Min((Fare \div Age)(AggMax(Age, Cabin))$
- $AdvFeat4$：$(cos(Fare)) \times (Ticket \div Age)$
- $AdvFeat5$：$Min(((sin(Ticket)) - (Age + Ticket)))((cos(Fare)) \div (sin(Ticket))))$
- $AdvFeat6$：$((sin(Ticket)) \times (cos(Fare))) \times (Max((Age \times Ticket)(Ticket \div Fare)))$
- $AdvFeat7$：$(cos(Fare)) \div (Fare \times Ticket)$
- $AdvFeat8$：$(cos(Ticket)) \div (Fare + Ticket)$
- $AdvFeat9$：$(cos(Ticket)) \div (Fare \times Ticket)$
- $AdvFeat10$：$(Ticket \div Age) \div (cos(Fare))$
- $AdvFeat11$：$(Fare \times Ticket) \div (cos(Ticket))$
- $AdvFeat12$：$(Fare + Ticket) - (sin(Ticket)))$
- $AdvFeat13$：$(Ticket \div Age) + (Fare + Ticket))$
- $AdvFeat14$：$(cos(Fare)) - (Fare + Ticket))$
- $AdvFeat15$：$(Ticket \div Fare) \div (cos(Fare))$
- $AdvFeat16$：$(sin(Ticket)) \div (Ticket \div Age)$
- $AdvFeat17$：$(cos(Fare)) \times (Fare \times Ticket)$
- $AdvFeat18$：$(Fare + Ticket) \div (cos(Ticket))$
- $AdvFeat19$：$(Fare + Ticket) \div (cos(Fare))$
- $AdvFeat20$：$Max(((cos(Fare)) + (cos(Ticket))))(Min((Fare \times Ticket)(Fare \div Age))))$
- $AdvFeat21$：$((cos(Ticket)) - (sin(Ticket)))) - (Max((Age \times Ticket)(Ticket \div Fare))))$
- $AdvFeat22$：$Max(((cos(Fare)) - (cos(Ticket))))((cos(Ticket)) \div (sin(Ticket))))$
- $AdvFeat23$：$(cos(Fare)) \div (Fare + Ticket)$
- $AdvFeat24$：$((cos(Fare)) \times (cos(Ticket))) \times (sin(sin(Ticket)))$
-

- $AdvFeat25$：$(sin(Ticket)) \div (Fare \times Ticket)$
- $AdvFeat26$：$(Fare \times Ticket) \div (cos(Fare))$
- $AdvFeat27$：$(cos(Fare)) + (Fare + Ticket))$
- $AdvFeat28$：$Max((cos(Fare))(cos(Ticket)))$
- $AdvFeat29$：$Count(Sex)$

特征贡献度

- $AdvFeat1$：35.55555555555556
- $AdvFeat2$：62.22222222222222
- $AdvFeat3$：68.88888888888889
- $AdvFeat4$：31.11111111111111
- $AdvFeat5$：11.11111111111111
- $AdvFeat6$：20.0
- $AdvFeat7$：17.77777777777778
- $AdvFeat8$：40.0
- $AdvFeat9$：35.55555555555556
- $AdvFeat10$：24.444444444444443
- $AdvFeat11$：26.666666666666668
- $AdvFeat12$：55.55555555555556
- $AdvFeat13$：22.22222222222222
- $AdvFeat14$：26.666666666666668
- $AdvFeat15$：31.11111111111111
- $AdvFeat16$：37.77777777777778
- $AdvFeat17$：15.555555555555555
- $AdvFeat18$：44.44444444444444
- $AdvFeat19$：17.77777777777778
- $AdvFeat20$：15.55555555555555
- $AdvFeat21$：44.44444444444444
- $AdvFeat22$：46.666666666666664
- $AdvFeat23$：22.22222222222222
- $AdvFeat24$：28.888888888888886

- $AdvFeat25$：40.0
- $AdvFeat26$：13.333333333333334
- $AdvFeat27$：20.0
- $AdvFeat28$：53.333333333333336
- $AdvFeat29$：11.11111111111111

附录-2：模型报告示例

2MB

report_example.zip
archive

上一页
模型预测

下一页
模型复现

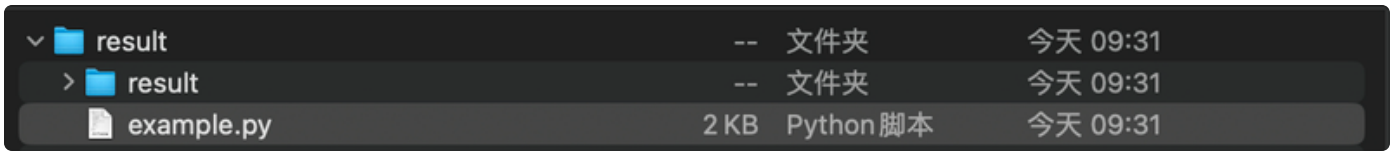
最后更新于1个月前

模型预测

模型预测的主要作用是：用户可以将建模平台中训练完成的模型下载下来，用于预测和当前模型相匹配的业务数据。具体使用步骤如下：

步骤一：下载模型

用户点击建模任务列中对应任务的 【下载模型】 按钮下载该任务训练完成的模型，此时会获取到一个压缩文件，解压之后大致内容如下：



步骤二：搭建预测环境

这里假定用户已经按照好了Anaconda环境，该环境是用于快速安装Python环境的工具，然后执行如下命令搭建预测环境：

```
# 创建Python虚拟环境
conda create -n changtian python==3.10 -y
# 激活虚拟环境
conda activate changtian
# 安装预测框架依赖
pip install changtianml -i https://pypi.tuna.tsinghua.edu.cn/simple/
```

至此预测环境搭建完毕！

步骤三：模型预测

> 本平台目前已经在Python公有仓库 PyPI 中发布了预测框架库 changtianml，通过该依赖库可以更加便携的方便用户完成预测等相关工作。

模型运行可以参考平台提供的example.py样例文件，该文件中的注释部分详细说明了预测框架库的功能和基本使用，请用户详细阅读。下面提供一段代码仅供参考：

```
# 导入平台相关依赖
from changtianml import tabular_incrml

# 指定模型目录加载模型
res_path = r'result'
ti = tabular_incrml(res_path)

# 指定需要预测数据集开始预测
test_path = r'test_data.csv'
pred = ti.predict(test_path)

# 输出预测结果
print(pred)
print(ti.predict_proba(test_path))
```

该平台为付费用户提供了一套模型预测框架，该框架后续会提供更多功能，敬请期待！

上一页
模型下载

下一页
模型报告

最后更新于4天前

调优指南

机器学习的训练过程相对比较复杂。

生成不同的特征，不同的模型，不同的模型参数，不同的验证集等等都会导致模型的效果不同。 这里主要介绍如何使用ChangTianML调优，以及如何得到基于您数据集最优的模型。

您可以按照顺序依次调整ChangTianML中的配置。

一. 设置合适的验证集比例

由于验证集是由整体数据集切分得来的，剩下的数据就是训练集，所以验证集和训练集就是**零和关系。**

验证集变大那么训练集就会变小，如果训练集太小那么可能会导致**模型拟合不足**甚至**性能不佳。**反之，验证集如果太小，它可能无法准确反映现实世界数据的真实分布。 这可能会导致**过度拟合**，即模型在训练集上表现出色，但在验证集上表现不佳，更重要的是，在未见过的数据上表现不佳。

调整方式

默认的是**0.25**的验证集比例，这是根据常见的数据所集设置的。

如果模型拟合的效果不好，可以首先调整这个比例，这里建议**调整的范围在0.1-0.3**之间。

二. 自动机器学习-最大尝试时长(秒)

ChangTianML自动机器学习是按照**时间**进行最佳模型的探索，增加**最大尝试时长**，模型探索的路径会更加深入，可能会找到更优的模型。

由于资源限制，目前**自动机器学习-最大尝试时长(秒)** 最高支持600秒。

三. 自动特征工程-最大实验次数

ChangTianML自动化特征工程会根据设置的最大实验次数去深度捆绑特征，当设置得越大，那么搜索的范围越大，越有可能搜索到有效特征。

由于资源限制，目前**自动特征工程-最大实验次数**最高支持10次。

四. 特征生成模式

更有创造力模式可能会探索并生成更多特征组合。

更精确模式在探索特征时以准确性为优先。

更平衡模式介于二者之间。

如果高级特征生成得较少，那么可以选择**更有创造力模型**，这会让算法更主动地寻找高级有效特征，为模型提供更多信息，从而得到更好的模型效果。

如果高级特征生成得较多，那么可以选择**更有精确模型**，这会让算法更严谨地寻找高级有效特征，降低模型过拟合的风险，提高模型的鲁棒性。

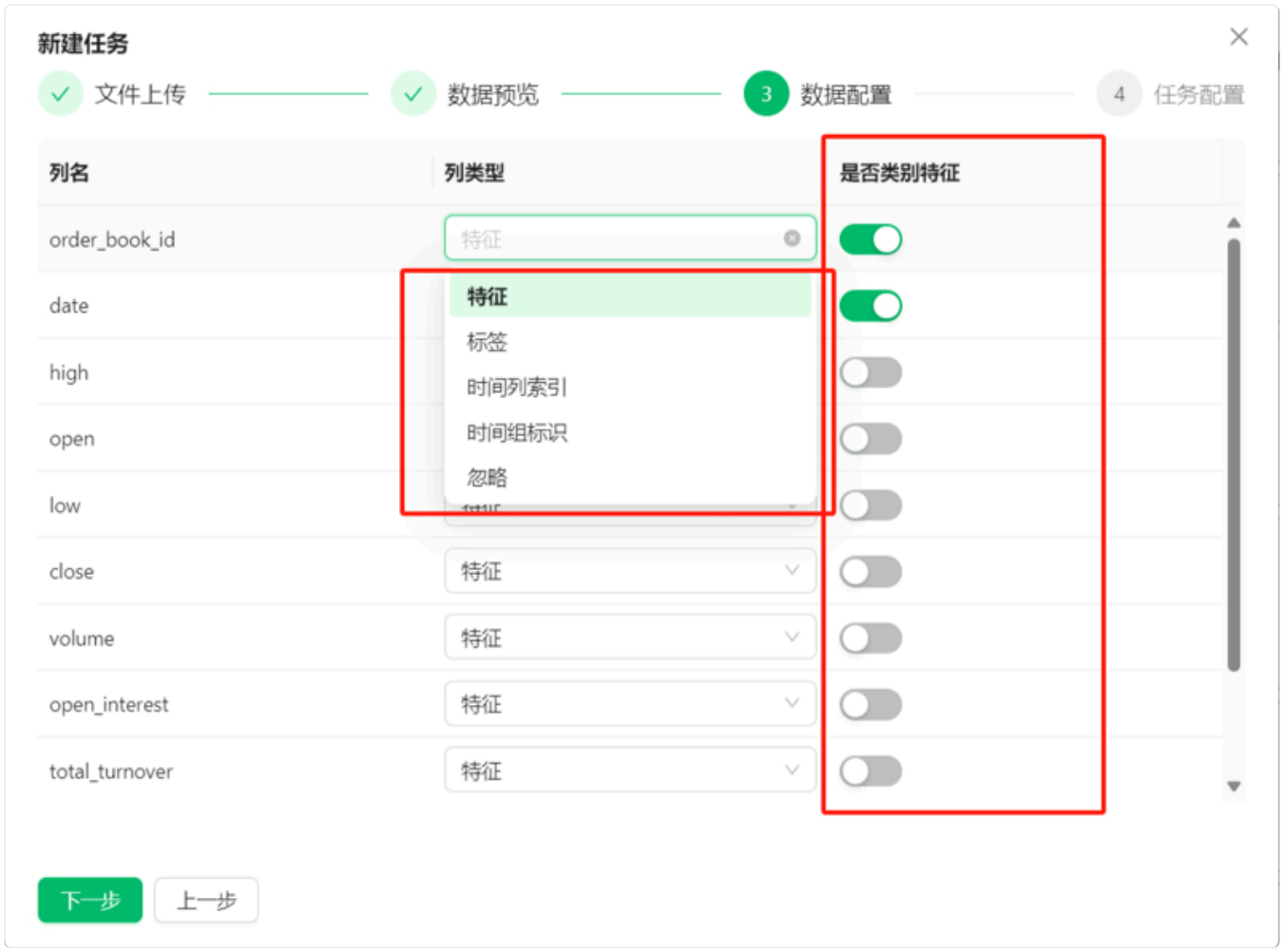
五. 深度特征搜索

特征搜索是一个无限空间内进行的，例如：特征A和B组合成一个C的特征，特征C和A还可以组合成D特征等等，那么开启深度搜索可能得到更深层次的特征。

六. 数据配置

检查数据是一个非常重要的步骤，平台上会根据整体的数据从统计学上给出一个初步判断，但从实际物理意义上不一定是完全正确的。检查可以从两个方面。

- **列特征**：特别需要注意的是，id类有多种，**唯一标识的id列需要忽略**否则可能造成过拟合。对于时间序列的任务，时间列索引和时间组标识需要配置好。
- **是否类别特征**：由于特征捆绑中类别和数值特征的方式不同，以及对机器学习模型也有较大的影响，所以请确认特征是否为类别特征。



上一页
模型复现

下一页
changtianml库

最后更新于4天前

📕 changtianml库

> `tabular_incrml`类提供了一系列用于处理表格数据、模型训练和预测的方法。以下是该类的公开API及其使用说明。

构造函数

`init`

```
init(path: str)
```

【说明】

初始化 `tabular_incrml` 类的实例。

【参数】

- `path (str)`: 指向 changtianml 训练结果所在目录的路径。

【返回】

无

预测函数

`predict`

```
predict(test_path: Union[str, pd.DataFrame]) -> np.ndarray
```

【说明】

使用模型对测试数据进行预测。

【参数】

- `test_path (str or pd.DataFrame)`: 提供三种加载方式：

  （1）输入csv数据集的地址；

  （2）输入文件夹，文件夹中仅包含csv文件；

  （3）直接输入pd.DataFrame。

【返回】

- `np.ndarray`: 预测结果的一维 numpy 数组。

`predict_proba`

```
predict_proba(test_path: Union[str, pd.DataFrame]) -> np.ndarray
```

【说明】

获取分类任务中，模型对每个类别的预测概率。

【参数】

- `test_path (str or pd.DataFrame)`: 提供三种加载方式：

  （1）输入csv数据集的地址；

  （2）输入文件夹，文件夹中仅包含csv文件；

  （3）直接输入pd.DataFrame。

【返回】

- `np.ndarray`: 预测概率的二维 numpy 数组。

数据处理函数

`preprocess`

```
preprocess(input: Union[str, pd.DataFrame], return_y: bool = False) ->
pd.DataFrame
```

【说明】

对输入的数据进行预处理，包括特征组合等操作。

【参数】

- `input (Union[str, pd.DataFrame])`: 输入数据，可以是CSV文件的路径或已加载的DataFrame。
- `return_y (bool, optional)`: 指示是否返回标签列，默认为 False。如果为True，则返回的DataFrame包含标签。如果数据没有标签，那么始终都不会返回标签信息。

【返回】

- `pd.DataFrame`: 预处理后的数据。

`labelencode`

```
labelencode(X: Union[pd.Series, pd.DataFrame], feature_name: Optional[str]
= None) -> pd.Series
```

【说明】

对分类特征进行编码，转换为模型可以处理的格式。

【参数】

- `X (Union[pd.Series, pd.DataFrame])`: 需要进行编码的数据。
- `feature_name (Optional[str], optional)`: 特征名，当 X 为 DataFrame 时必填。

【返回】

- `pd.Series`: 编码后的数据。

labeldecode

```
labeldecode(X: Union[pd.Series, pd.DataFrame], feature_name: Optional[str]
= None) -> pd.Series
```

【说明】

对编码后的特征进行解码，还原为原始格式。

【参数】

- `X (Union[pd.Series, pd.DataFrame])`: 需要进行解码的数据。
- `feature_name (Optional[str], optional)`: 特征名，当 X 为 DataFrame 时必填。

【返回】

- `pd.Series`: 解码后的数据。

获取函数

get_train_validation_data

```
get_train_validation_data(path: Optional[str] = None) ->
tuple(pd.DataFrame, pd.DataFrame, pd.Series, pd.Series)
```

【说明】

获取训练集和验证集。

【参数】

- `path (str, optional)`: 提供四种加载方式：

  （1）输入csv数据集的地址；

  （2）输入文件夹，文件夹中仅包含csv文件；

（3）输入None，并在此之前调用了load_training_data加载数据；4.直接输入DataFrame。

【返回】

- `tuple(pd.DataFrame, pd.DataFrame, pd.Series, pd.Series)`: 返回包含X_train, X_val, y_train, y_val的DataFrame。

`obtain_model_configuration`

```
obtain_model_configuration()
```

【说明】

获取并返回 changtianml 模型的配置信息。此函数提取模型的类型以及其最佳配置参数。以确保参数信息的准确性和可用性。

【返回】

- `dict`: 包含以下键值对的字典：
  - `model_type`: AutoML 选择的最佳模型类型。
  - `model_params`: 该模型的最佳配置参数。对某些参数进行了适当的名称转换和格式调整。
  - `task`: 模型的任务类型，主要是分类和回归。

画图函数

`plot_box`

```
plot_box(X: pd.DataFrame, y: pd.DataFrame, feature_name: str, target_name: str, path: Optional[str] = None)
```

【说明】

绘制箱线图，用于分析特征和目标之间的关系。

【参数】

- `X (pd.DataFrame)`: 包含特征的DataFrame。
- `y (pd.Series)`: 目标变量。
- `feature_name (str)`: 要分析的特征名称。
- `target_name (str)`: 目标变量的名称。
- `path (Optional[str], optional)`: 保存图表的路径，如果为 None，则不保存图表。

【返回】

无

plot_kde

```
plot_kde(X: pd.DataFrame, feature_name: str, path: Optional[str] = None)
```

【说明】

绘制核密度估计图，以估计特征的分布。

【参数】

- `X (pd.DataFrame)`: 包含特征的DataFrame。
- `feature_name (str)`: 要进行核密度估计的特征名称。
- `path (Optional[str], optional)`: 保存图表的路径，如果为 None，则不保存图表。

【返回】

无

plot_trend_comparison

```
plot_trend_comparison(X: pd.DataFrame, feature_name: str, target_name: str,
path: Optional[str] = None)
```

【说明】

绘制特征与目标之间的趋势图。

【参数】

- `X (pd.DataFrame)`: 包含数据的DataFrame。
- `feature_name (str)`: 要分析趋势的特征名称。
- `target_name (str)`: 目标变量的名称。
- `path (Optional[str], optional)`: 保存图表的路径，如果为 None，则不保存图表。

【返回】

无

plot_roc_curve

```
plot_roc_curve(y_true: Union[pd.Series, np.ndarray], y_score:
Union[pd.Series, np.ndarray], n_classes: Optional[int] = None, roc_type:
str = 'both', path: Optional[str] = None)
```

【说明】

绘制接收者操作特征(ROC)曲线，用于评估分类模型性能。

【参数】

- `y_true (pd.Series , np.ndarray)`: 真实的标签值。
- `y_score (pd.Series , np.ndarray)`: 模型预测的概率值或得分。
- `n_classes(int, optional)`: 分类任务中的类别数。如果不指定，将根据数据自动计算。
- `roc_type (str, optional)`: ROC曲线类型。可选项包括 'both'（同时绘制宏观和微观ROC曲线），'macro'，'micro'。默认为 'both'。
- `path (Optional[str], optional)`: 保存图表的路径，如果为 None，则不保存图表。

【返回】

无

plot_confusion_matrix

```
plot_confusion_matrix(y_true: Union[pd.Series, np.ndarray], y_pred:
Union[pd.Series, np.ndarray], normalize: bool = False, title: Optional[str]
= None, cmap = plt.cm.Blues, path: Optional[str] = None)
```

【说明】

绘制混淆矩阵，用于展示分类模型的预测准确性和错误类型。

【参数】

- `y_true (pd.Series , np.ndarray)`: 真实的标签值。
- `y_pred (pd.Series , np.ndarray)`: 模型的预测结果。
- `n_classes (int, optional)`: 分类任务中的类别数。如果不指定，将根据数据自动计算。
- `normalize (bool, optional)`: 是否对混淆矩阵进行归一化。默认为False。
- `title (str, optional)`: 图表的标题。默认为None。
- `cmap (Colormap, optional)`: 使用的颜色映射。默认为plt.cm.Blues。
- `path (Optional[str], optional)`: 保存图表的路径，如果为 None，则不保存图表。

【返回】

无

plot_regression_scatter

```
plot_regression_scatter(y_true: Union[pd.Series, np.ndarray], y_pred:
Union[pd.Series, np.ndarray], model_name: str = 'Model', path:
Optional[str] = None)
```

【说明】

绘制回归任务的散点图，比较模型的预测值和实际值。

【参数】

- `y_true (pd.Series , np.ndarray)`: 真实的目标值。
- `y_pred (pd.Series , np.ndarray)`: 模型的预测值。
- `model_name (str, optional)`: 图例中使用的模型名称。默认为'Model'。
-

- `path (Optional[str], optional)`: 保存图表的路径，如果为 None，则不保存图表。

【返回】

无

上一页
调优指南

下一页
数据集和经典案例

最后更新于4天前

🏈 数据集和经典案例

为了帮助用户更好地了解和使用该建模平台，我们提供了一些网上公开的数据集链接，以及与我们平台相关的建模视频。

本文中提供的数据集均公开可用（无版权限制），我们鼓励您利用这些数据集进行实践和学习，以更深入地了解我们的建模平台。此外，我们提供的操作视频旨在帮助用户掌握软件操作技巧，并促进学习与交流。我们强烈建议用户在使用这些资源时遵守所有适用的法律法规，并尊重数据集和视频的原始所有者的权利。

Meet-up建模直播系列

Meet-up建模直播系列会更详细介绍建模案例的背景和应用价值，并通过手动编写代码建模，对长天ML建模平台有更深入的应用介绍。

用户流失预测

> 数据：https://www.kaggle.com/datasets/whenamancodes/credit-card-customers-prediction
> 视频：https://www.bilibili.com/video/BV1ZC4y1776Q/

心脏病患者预测

> 数据：https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data
> 视频：https://www.bilibili.com/video/BV1XN411j7GQ

航班价格预测

> 数据：https://www.kaggle.com/datasets/shubhambathwal/flight-price-prediction

> 视频：https://www.bilibili.com/video/BV1Y94y1A7U4

Kaggle建模实战系列

Kaggle建模实战系列会通过3分钟的时间快速介绍Kaggle数据集情况，并使用长天ML建模平台快速进行建模演示。

酒店预定预测

> 数据：https://www.kaggle.com/datasets/youssefaboelwafa/hotel-booking-cancellation-prediction
> 视频：https://www.bilibili.com/video/BV1DN4y1a7Q7

公司破产预测

> 数据：https://archive.ics.uci.edu/dataset/572/taiwanese+bankruptcy+prediction
> 视频：https://www.bilibili.com/video/BV1Bb4y157cG

中风患者预测

> 数据：https://www.kaggle.com/datasets/zzettrkalpakbal/full-filled-brain-stroke-dataset视频：https://www.bilibili.com/video/BV1sw411t78c

水质健康预测

> 数据：https://www.kaggle.com/datasets/adityakadiwal/water-potability
> 视频：https://www.bilibili.com/video/BV1vb4y157qL

英雄联盟预测

> 数据：https://www.kaggle.com/datasets/bobbyscience/league-of-legends-diamond-ranked-games-10-min
> 视频：https://www.bilibili.com/video/BV1Se411k728

设备故障预测

> 数据：https://www.kaggle.com/datasets/shivamb/machine-predictive-maintenance-classification
> 视频：https://www.bilibili.com/video/BV1xc411y7yX

电梯维护预测

> 数据：https://www.kaggle.com/datasets/shivamb/elevator-predictive-maintenance-dataset
> 视频：https://www.bilibili.com/video/BV1oG411r7Mn

引擎健康预测

> 数据：https://www.kaggle.com/datasets/parvmodi/automotive-vehicles-engine-health-dataset
> 视频：https://www.bilibili.com/video/BV1E64y1W7RU

军舰维护预测

> 数据：
> https://archive.ics.uci.edu/dataset/316/condition+based+maintenance+of+naval+propulsion+plants
> 视频：https://www.bilibili.com/video/BV17C4y1u7gu

心力衰竭预测

> 数据：https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data
> 视频：https://www.bilibili.com/video/BV1bb4y1G7nk

糖尿病预测

> 数据：https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset视频：https://www.bilibili.com/video/BV1hw41157pa

肺癌风险预测

> 数据：https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer
> 视频：https://www.bilibili.com/video/BV1Hw41157

肝硬化存活预测

> 数据：https://www.kaggle.com/datasets/joebeachcapital/cirrhosis-patient-survival-prediction
> 视频：https://www.bilibili.com/video/BV12g4y1k7XT

甲状腺癌症预测

> 数据：https://www.kaggle.com/datasets/zsinghrahulk/differentiated-thyroid-cancer-recurrence
> 视频：https://www.bilibili.com/video/BV1Fg4y1C7wN/

分析师工资预测

> 数据：https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries/data
> 视频：https://www.bilibili.com/video/BV1xg4y1r7fs

银行存款预测

> 数据：https://www.kaggle.com/datasets/thedevastator/bank-term-deposit-predictions
> 视频：https://www.bilibili.com/video/BV1W5411z7dF

睡眠质量预测

> 数据：https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset
> 视频：https://www.bilibili.com/video/BV1v5411q7Y6

房屋价格预测

> 数据：https://www.kaggle.com/datasets/shubhammeshram579/house
> 视频：https://www.bilibili.com/video/BV1XT4y1H7D8

研究生录取预测

> 数据：https://www.kaggle.com/datasets/mohansacharya/graduate-admissions
> 视频：https://www.bilibili.com/video/BV1oe411e7gE

上一页
changtianml库

下一页
长天ML公开数据集

⚾ 长天ML公开数据集

长天ML公开数据集包含了许多我们逸思长天团队调研的各行业的开源数据集，每个数据集都包含了数据本以及数据的源信息。

如何获取长天ML数据集?

方式一：通过链接直接下载

在浏览器中数据输入相应的网址，即可进行下载。

方式二：使用changtianml python包进行数据下载

这里假定用户已经按照好了Anaconda环境，该环境是用于快速安装Python环境的工具，然后执行如下命令搭建预测环境。

```
# 创建Python虚拟环境
conda create -n changtian python==3.10 -y
# 激活虚拟环境
conda activate changtian
# 安装预测框架依赖
pip install changtianml -i https://pypi.tuna.tsinghua.edu.cn/simple/
```

然后创建一个python文件（以.py为后缀的文件）。

```
from changtianml import load_dataset

# 指定数据集对应的字符串，具体可查询附录中表格进行下载
obj = load_dataset('stock_info') # load_dataset方法会返回的是一个类
# 获取数据
data = obj.data
# ...
```

附录

序号	数据集名称	changtianml本地包加载字符串
1	股票信息	stock_info
2	机票价格	flight_price

上一页
数据集和经典案例

下一页
ChangTianML v1.2

最后更新于4天前

💰 购买相关

为了提供更优质的产品服务，长天ML在基础功能之上，面向个人用户推出付费套餐服务，用户完成支付购买后可享受更多功能权益。具体权益如下：

	基础版	付费版
功能		
创建模型	✅	✅
运行模型	✅	✅
下载模型	❌	✅
下载模型报告	❌	✅
查看模型高级特征	❌	✅
下载模型高级特征	❌	✅
用新数据进行预测	500条	根据服务器负载最大化数据量

ℹ 付费用户的更多权益将持续上线，敬请关注！

付费功能

1.查看高级特征

PAGE
高级特征 ›

2.查看模型报告

PAGE
模型报告

3.下载模型

PAGE
模型下载

4.模型预测（线上）

付费后，用户可以点击建模任务列中对应任务的 【预测】 按钮完成对业务数据的预测功能。对于付费用户可预测超过500条数据，同时也可以上传更多数据，如下图：

5.模型预测（线下）

PAGE
模型预测

如何购买?

用户可点击页面右上放的【购买套餐】即可进入支付入口，选择合适的购买模式点击【下一步】进入扫码支付。目前仅支持支付宝扫码购买，后续会支持PayPal支付，支付成功之后即可享用相关权益。

如果对支付订单有异议，用户点击页面右上放的用户头像选择【订单】前往订单中心查看相关订单。

常见问题

Q：月卡到期后会怎么样?

A：月卡模式到期后，用户可前往订单中心查看有无其他订单状态为【服务中】的订单。如果有相关订单，则会继续享有付费权益；如果没有相关的订单，则需要继续付费，反之相关付费功能将不再享有相关权益。

Q：开通月卡并支付完成后，模型仍然无法享有权益?

A：只有当建模任务的创建时间在月卡模式下订单所对应的服务起止时间内，才可享受相关权益；否则无法享受对应付费权益。用户在这种情况下可以重新构建当前任务即可。

Q：多次开通月卡并支付完成后，权益如何分配?

A：如果用户完成多次月卡支付，会根据卡种对应的服务时间区间然后按照订单支付的先后顺序在服务时间长完成累加。

举个例子：用户A在当前时间完成一笔单月卡订单A的支付，在大约1小时之后再次完成一笔季度卡订单B的支付，则月卡模式下享有权益的时间如下：

支付订单号	订单类型	订单状态	支付金额	服务开始时间
B	季度卡	付款成功	￥825	2023-12-12 10:42:12
A	单月卡	服务中	￥300	2023-11-12 10:42:12

上一页
ChangTianML v1.4

下一页
使用相关

最后更新于4天前

活动相关

为了帮助用户更好的使用平台的高级功能，本平台将陆续开展一系列优惠活动。用户可点击页面右上角 【福利】 按钮关注 【优惠活动】 了解相关活动细则。

声明：

（1）本平台所涉及的所有优惠活动相关内容的最终解释权归**逸思长天**所有。

（2）本平台所涉及的所有优惠活动规则变更另行通知，您可关注**逸思长天官方微信公众号**以及**逸思长天官方bilibili账号**，敬请关注我们的ChangTianML官方平台中的【优惠活动】，更多优惠活动即将推出，敬请期待...

一.邀请新用户注册免费获取次卡

即日起推荐新用户注册，您将获赠免费次卡使用模型高级功能的特权！邀请好友加入，轻松解锁自动化机器学习和高级特征工程体验，可获取模型的深度解析报告。

什么是次卡?

次卡是一种基于单模型计费方式下的全免消费券。即：平台用户只想使用某个建模任务下所有的高级功能。

规则和玩法

- **当用户携带邀请码注册时，当前用户和所携带邀请码的宿主用户同时获得一张次卡。**例如：用户B携带了邀请码abcd17，同时邀请码abcd17是属于用户A的，当用户B注册成功后，用户A和用户B两者同时获得一张次卡。
- **每张次卡均有有效期限制，默认30天有效期。**有效期从获得当前次卡之后开始计算。
- **使用次卡购买模型训练付费功能之后，该模型任务的所有功能将会获得永久使用权。**
- **用户在近30天内最多只能通过邀请其他用户获得5张次卡，无论这些次卡是否过期或者使用与否。**

用户在获取次卡之后可以前往 【福利】 中的 【我的卡券】 查看用户可用的次卡。

如果您已经拥有次卡，可在建模任务列表中点击相关高级功能按钮选择 【次卡抵扣】 按钮完成面额抵扣即可享用。

上一页
使用相关

下一页
联系我们

最后更新于4天前

联系我们

欢迎使用我们的服务！如有疑问或需要技术支持，您可通过以下方式联系我们，我们将尽快回复您，感谢您的支持！

1. 商务合作：zhouheng@ysct.org.cn
2. 技术支持：futureguo@ysct.org.cn
3. 扫码关注微信公众号：

wechat

上一页
活动相关

最后更新于4天前