

3.4.4 回归

3.4.4.1 线性回归

图标: 

描述: 线性回归是利用数理统计中回归分析, 来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。

字段属性

特征列: 请选择数值型数据, 如果勾选了非数值类型数据, 则会自动过滤, 下个组件可能无法获取所有列。

标签列: 请选择数值型数据。

输出

表结果: 线性回归预测结果。

报告: 模型拟合效果。

示例

下列对某数据进行线性回归算法:

- 选择自变量, 因变量, 均选择数值型数据。如图 205 所示。
- 运行成功后, 选择查看数据, 如图 206 所示。
- 运行成功后, 选择查看报告, 如图 207 所示。

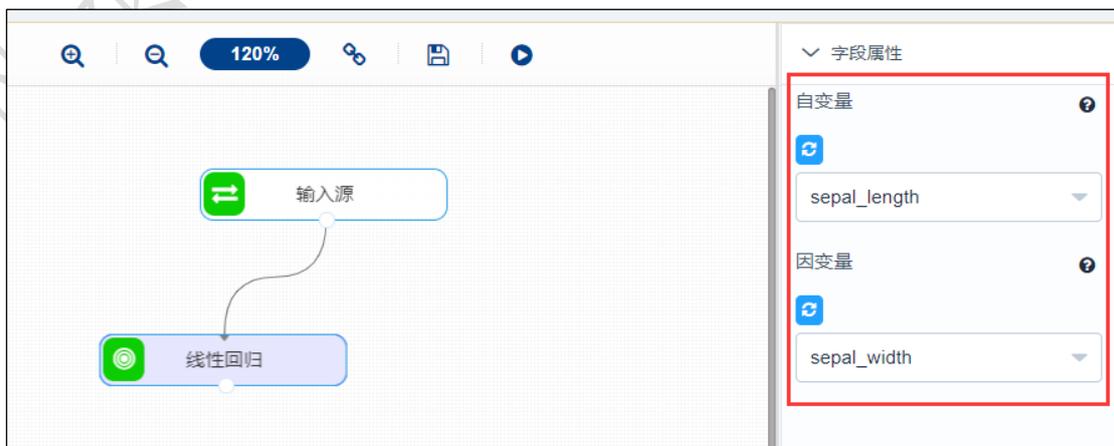


图 205

sepal_length	sepal_width	predict_label
5.1	3.5	3.103
4.9	3	3.116
4.7	3.2	3.128
4.6	3.1	3.134
5	3.6	3.11
5.4	3.9	3.085
4.6	3.4	3.134
5	3.4	3.11

共 150 条 25 条/页 < 1 2 3 4 5 6 > 前往 1 页

图 206

算法运行报告

线性回归模型评价

模型拟合效果

Coefficients(系数): [-0.0618848]
intercept(截距项): 3.418946836103816
Score(R²): 0.013822654141080748

图 207

3.4.4.2 广义最小二乘法

图标: 广义最小二乘法

描述: 广义最小二乘法是一种常见的消除异方差的方法。它的主要思想是为解释变量加上一个权重, 从而使得加上权重后的回归方程方差是相同的。因此在 GLS 方法下我们可以得到估计量的无偏和一致估计。

字段属性:

自变量: 选择自变量所在列, 请选择数值型数据。

因变量：选择响应变量所在的列，请选择数值型数据。

输出

表结果：无。

报告：GLS Regression Results。

示例

下列对某数据使用广义最小二乘法：

- 选择自变量，因变量，均选择数值型数据。如图 208 所示。
- 运行成功后，选择查看报告，如图 209、图 210 所示。



图 208

算法运行报告			
广义最小二乘法结果			
结果			
GLS Regression Results			
=====			
Dep. Variable:	sepal_width	R-squared:	0.957
Model:	GLS	Adj. R-squared:	0.956
Method:	Least Squares	F-statistic:	3277.
Date:	Wed, 07 Mar 2018	Prob (F-statistic):	2.42e-103
Time:	15:43:20	Log-Likelihood:	-146.83
No. Observations:	150	AIC:	295.7
Df Residuals:	149	BIC:	298.7
Df Model:	1		
Covariance Type:	nonrobust		

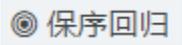
图 209

算法运行报告						
Df Model: 1						
Covariance Type: nonrobust						
=====						
	coef	std err	t	P> t	[0.025	0.975]

sepal_length	0.5118	0.009	57.246	0.000	0.494	0.529
=====						
Omnibus:	17.098	Durbin-Watson:	0.433			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7.933			
Skew:	0.352	Prob(JB):	0.0189			
Kurtosis:	2.121	Cond. No.	1.00			
=====						
Warnings:						
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.						

图 210

3.4.4.3 保序回归

图标: 

描述: 保序回归是定了一个无序的数字序列, 通过修改其中元素的值, 得到一个非递减的数字序列, 要求是使得误差 (预测值和实际值差的平方) 最小。

字段属性

自变量: 请选择数值型数据。

因变量: 请选择数值型数据。

输出

模型

表结果: 保序回归算法结果。

报告: 模型拟合效果、Isotonic regression。

示例

下列对某数据进行保序回归算法:

- 选择自变量, 因变量, 均选择数值型数据。图 211 所示。
- 运行成功后, 选择查看数据, 图 212 所示。
- 运行成功后, 选择查看报告, 如图 213、图 214 所示。
- 模型预测配置如图 215 所示。
- 模型预测结果如图 216 所示。

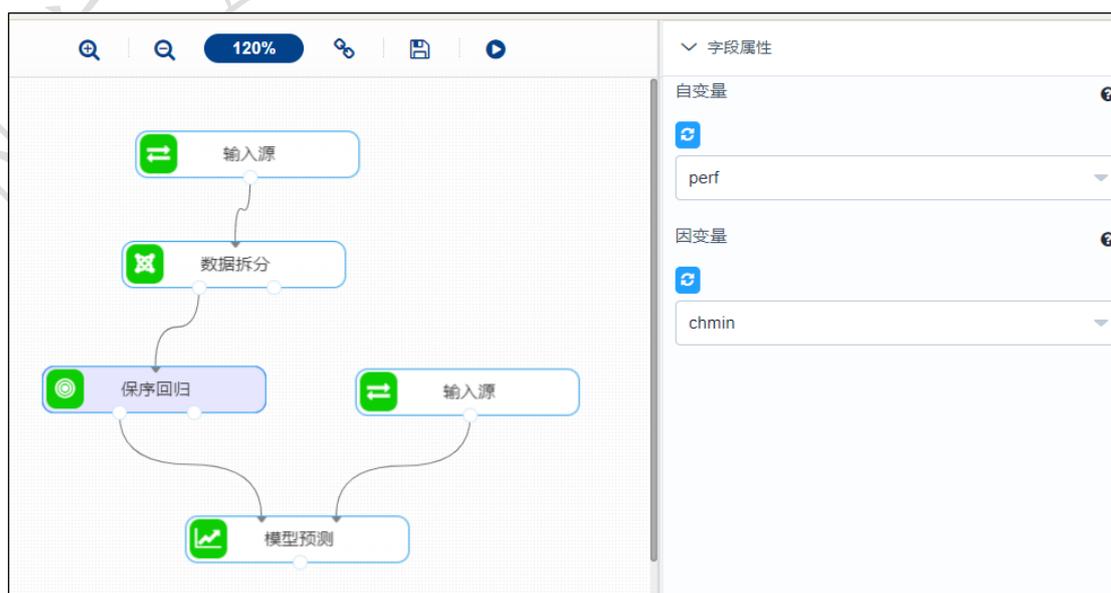


图 211

预览数据

perf	chmin	predict_label
397	52	22.5
636	16	22.5
36	3	2.611
18	1	1.05
45	3	2.611
80	1	2.611
22	1	1.4
248	12	10.909

共 156 条 25 条/页 < 1 2 3 4 5 6 7 > 前往 1 页

图 212

算法运行报告

保序回归模型评价

模型拟合效果

Score(R²): 0.56

图 213

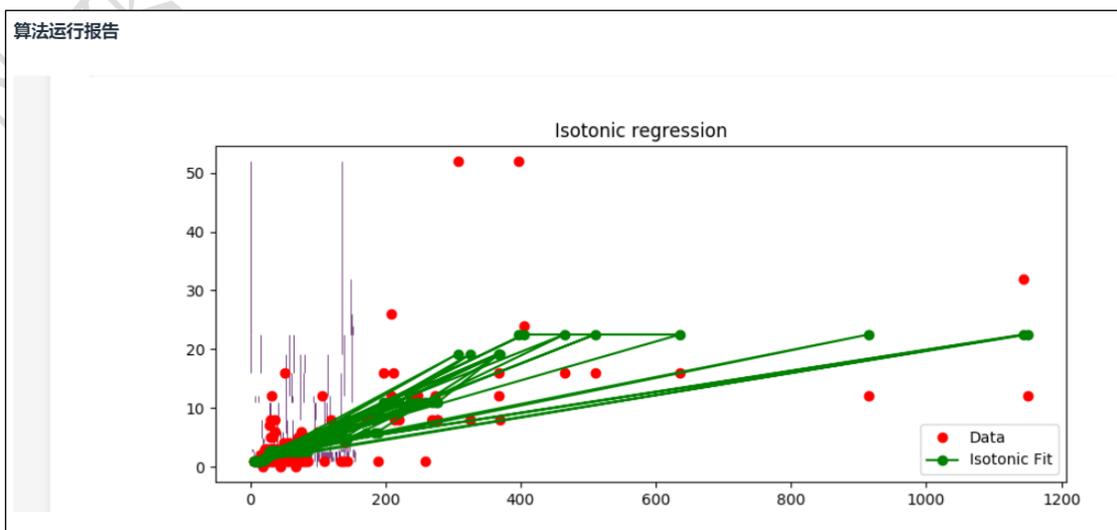


图 214

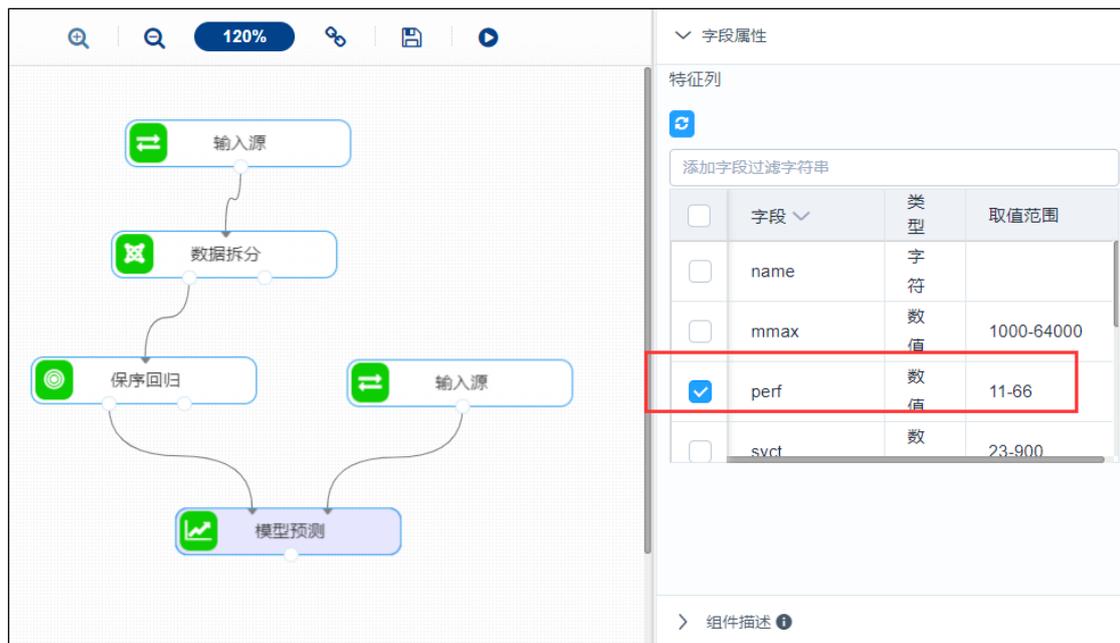


图 215

预览数据

perf	predict_label
198	10.909090909090908
269	10.909090909090908
220	10.909090909090908
172	5.666666666666667
132	4.75
318	19.2
367	19.2
489	22.5

共 209 条 25 条/页 < 1 2 3 4 5 6 ... 9 > 前往 1 页

图 216

3.4.4.4 岭回归

图标: 

描述: 岭回归是一种专用于共线性数据分析的有偏估计回归方法，实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法，对病态数据的拟合要强于最小二乘法。

字段属性

特征列：通过勾选的方式选择特征所在列。

标签列：选择响应变量所在的列。

输出

模型

表结果：岭回归算法结果。

报告：模型拟合效果。

示例

下列对某数据进行岭回归算法：

- 选择自变量，因变量，均选择数值型数据，如图 217 所示。
- 运行成功后，选择查看数据，如图 218 所示。
- 运行成功后，选择查看报告，如图 219 所示。
- 模型预测配置如图 220 所示。
- 模型预测结果如图 221 所示。



图 217

预览数据

e_pay	package_type	leave	educational_level	predict_label
1	2	1	3	3.146390283031271
1	4	1	3	3.4103142307932988
0	3	1	3	2.6102011102154608
0	3	0	4	2.2361321684074706
0	3	0	1	1.327650956576315
1	1	0	3	2.76201348845444
1	2	1	3	3.326114748971215
0	2	0	3	2.1961912496563345

共 597 条 25 条/页 < 1 2 3 4 5 6 ... 24 > 前往 1 页

图 218

算法运行报告

岭回归训练结果

模型拟合效果

Coefficients(系数): 2.7217171635838056
Residual sum of squares(均方误差): 2.7217171635838056
Score(R²): 2.7217171635838056

图 219



图 220

预览数据

wifi	e_pay	package_type	leave	predict_label
0	0	1	1	2.517733656944667
0	0	3	0	2.410079137442234
0	0	3	0	2.185801669553528
0	0	3	0	1.9794352523492829
0	0	2	0	2.3084376806694125
0	1	1	1	3.3701491677950592
0	0	3	0	2.1972744050121356
0	0	1	0	2.260510016826419

共 796 条 25 条/页 < 1 2 3 4 5 6 ... 32 > 前往 1 页

图 221

3.4.4.5 CART 回归树

图标:  CART回归树

描述: 使用 Cart 决策树算法的回归树。

字段属性

特征列: 通过勾选的方式选择特征所在列, 仅支持数值型数据。

标签列: 选择响应变量所在的列, 仅支持数值型数据。

参数设置

切分时的评价准则：包括均方误差、平均绝对误差，默认均方误差。

切分原则：包括选择最优的切分、随机切分，默认选择最优的切分。

输出

模型

表结果：CART 回归树算法结果。

报告：Regression model evaluation。

示例

下列对某数据进行 CART 回归树算法：

- 选择自变量，因变量，均选择数值型数据。如图 222 所示。
- 保留默认参数，切分时的评价准则为均方误差，切分原则为选择最优的切分，如图 223 所示。
- 运行成功后，选择查看数据，如图 224 所示。
- 运行成功后，选择查看报告，如图 225 所示。
- 模型评估配置如图 226 所示。
- 模型评估结果如图 227 所示。
- 模型评估报告如图 228 所示。
- 模型预测配置如图 229 所示。
- 模型预测结果如图 230 所示。

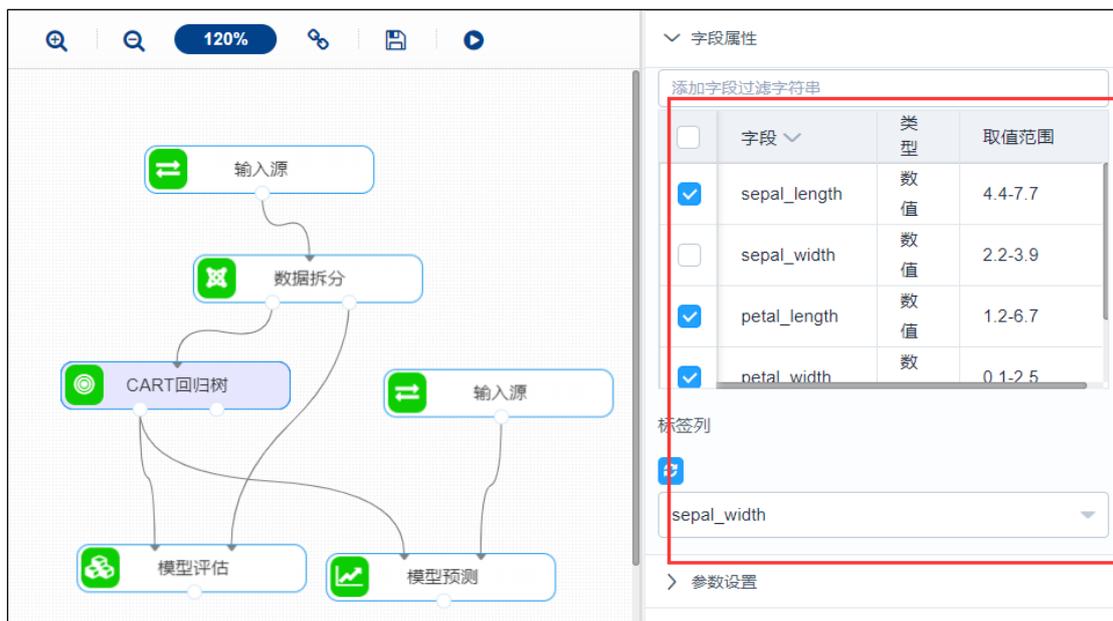


图 222

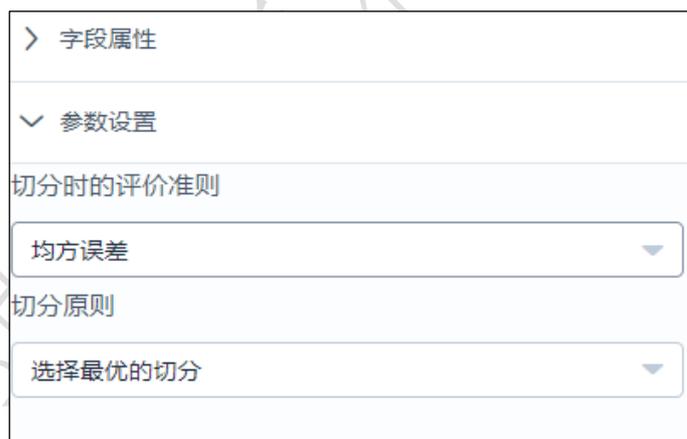


图 223

预览数据

sepal_length	petal_length	petal_width	sepal_width	predict_label
5.9	4.2	1.5	3	3
5.8	4	1.2	2.6	2.6
6.8	5.5	2.1	3	3
4.7	1.3	0.2	3.2	3.2
6.9	5.1	2.3	3.1	3.1
5	1.6	0.6	3.5	3.5
5.4	1.5	0.2	3.7	3.7
5	3.5	1	2	2

共 112 条 25 条/页 < 1 2 3 4 5 > 前往 1 页

图 224

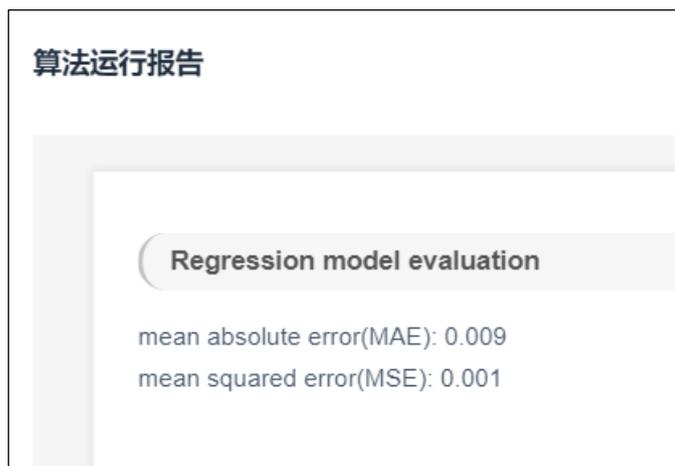


图 225



图 226

预览数据

sepal_length	petal_length	petal_width	sepal_width	predict_label
5.8	5.1	2.4	2.8	2.5
6	4	1	2.2	3
5.5	1.4	0.2	4.2	3.5
7.3	6.3	1.8	2.9	3.2
5	1.5	0.2	3.4	3.45
6.3	6	2.5	3.3	3.3
5	1.3	0.3	3.5	3.5
6.7	4.7	1.5	3.1	3.1

共 38 条 25 条/页 < 1 2 > 前往 1 页

图 227

算法运行报告

Regression model evaluation
mean absolute error(MAE): 0.3
mean squared error(MSE): 0.144

图 228



图 229

预览数据

petalwidth	sepalwidth	petalwidth	predict_label
1	5	0	2.4
1	5	0	2.4
1	5	0	2.4
2	5	0	2.4
1	5	0	2.4
2	5	0	2.4
1	5	0	2.4
2	5	0	2.4

共 150 条 25 条/页 < 1 2 3 4 5 6 > 前往 1 页

图 230

3.4.4.6 偏最小二乘回归

图标:

描述: 偏最小二乘回归 (PLSR) 是一种多因变量 Y 对多自变量 X 的回归建模方法, 该算法在建立回归的过程中, 既考虑了尽量提取 Y 和 X 中的主成分 (PCA—Principal Component Analysis, 主成分分析的思想), 又考虑了使分别从 X 和 Y 提取出的主成分之间的相关性最大化 (CCA 的思想)。简单的说, PLSR 是 PCA、CCA 和多元线性回归这三种基本算法组合的产物。

字段属性

自变量：请选择数值型数据，如果勾选了非数值类型数据，则会自动过滤，下个组件可能无法获取所有列。

因变量：请选择数值型数据。

参数设置

保留的主成分数量：整数型，默认为 2。

是否归一化数据：是/否：，默认是。

最大迭代数：整数型，默认 500。

输出

表结果：偏最小二乘回归算法结果。

报告：Regression model evaluation。

示例

下列对某数据进行偏最小二乘回归算法：

- 选择自变量，因变量。如图 231 所示。
- 保留默认参数，保留的主成分数量为 2，设置“是否归一化数据”为是，最大迭代数为 500，如图 232 所示。
- 运行成功后，选择查看数据，如图 233 所示。
- 运行成功后，选择查看报告，如图 234 所示。
- 模型评估配置如图 235 所示。
- 模型评估结果数据如图 236 所示。
- 模型评估运行报告如图 237 所示。
- 模型预测配置如图 238 所示。
- 模型预测结果数据如图 239 所示。



图 231



图 232

预览数据

e_pay	package_type	leave	educational_level	predict_value
1	2	1	3	3.2155253840803404
1	4	1	3	3.574540332932164
0	3	1	3	2.7067975549242758
0	3	0	4	2.098062415399615
0	3	0	1	1.396030847616374
1	1	0	3	2.8573978116648924
1	2	1	3	3.3362201134673173
0	2	0	3	2.2084169463764303

共 597 条 25 条/页 < 1 2 3 4 5 6 ... 24 > 前往 1 页

图 233



图 234



图 235

预览数据

e_pay	package_type	leave	educational_level	predict_label
0	3	0	1	1.9096767014107574
0	1	0	3	2.243351029345906
1	2	1	3	3.398886017962484
0	4	0	5	2.9229486259286332
0	3	0	1	2.28503575001661
1	1	0	1	3.5553221429465474
1	4	0	3	3.1854888615296884
1	2	0	2	2.8497799325270754

共 199 条 25 条/页 < 1 2 3 4 5 6 ... 8 > 前往 1 页

图 236

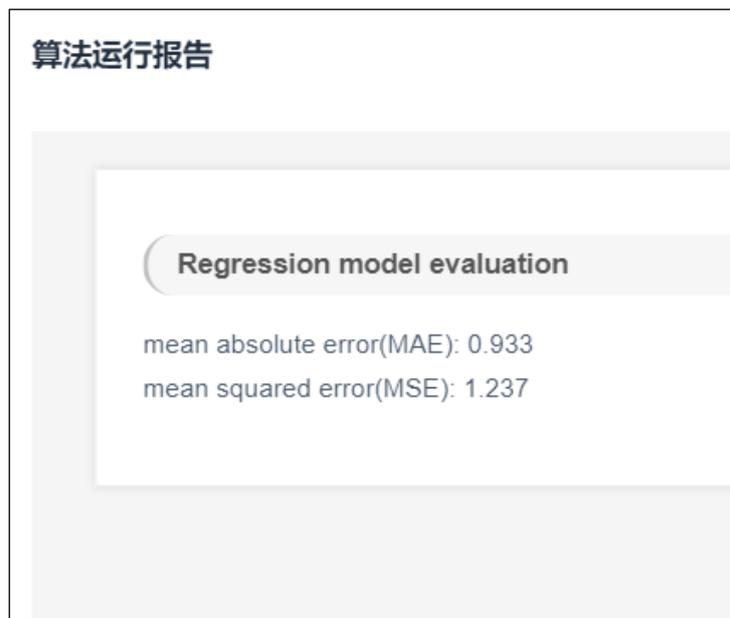


图 237

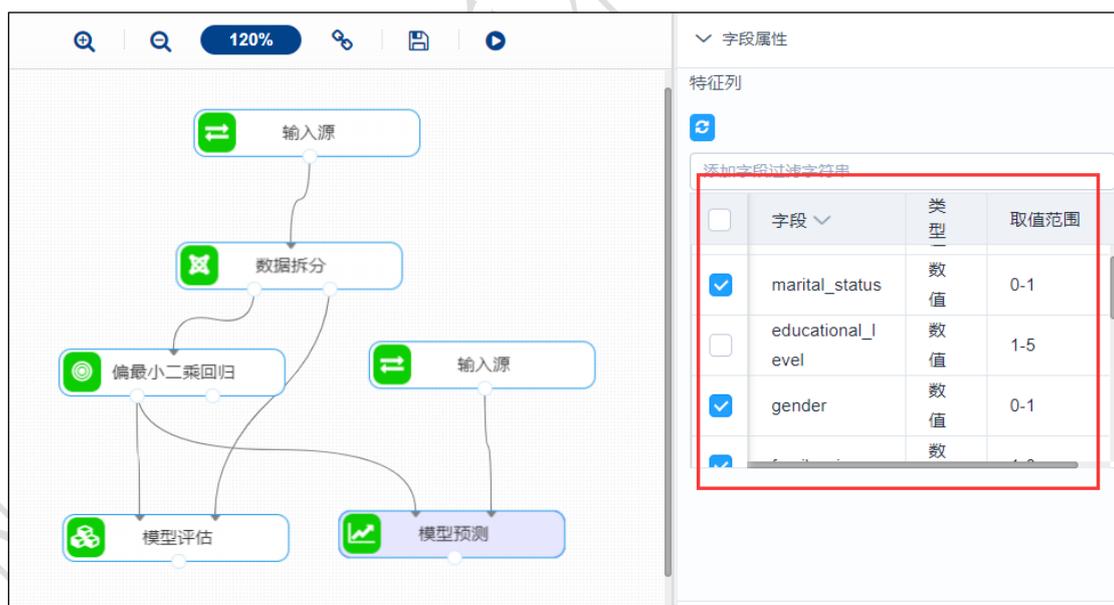


图 238

预览数据

free	wifi	package_type	leave	predict_label
0	0	1	1	19.381052127919805
0	0	3	0	9.658980956888641
19	0	3	0	23.66184763521481
29	0	3	0	49.175774674165666
0	0	2	0	21.261980649129196
0	0	1	1	38.78315003453123
22	0	3	0	24.824218973492858
0	0	1	0	11.045878028083042

共 796 条 25 条/页 < 1 2 3 4 5 6 ... 32 > 前往 1 页

图 239

3.4.4.7 Lasso 回归



图标:

描述: Lasso 回归是一个用于估计稀疏参数的线性模型，特别适用于参数数目缩减。

字段属性

特征列：通过勾选的方式选择特征所在列。

标签列：选择响应变量所在的列。

参数设置

alpha: 浮点型，默认 1.0。

最大迭代次数: 整数型，默认 500

输出

表结果: Lasso 回归预测结果。

报告: 模型拟合效果。

示例

下列对某数据进行 Lasso 回归算法:

- 选择自变量，因变量。如图 240 所示。
- 保留默认参数，alpha 为 1.0，最大迭代次数为 500，如图 241 所示。
- 运行成功后，选择查看数据，如图 242 所示。
- 运行成功后，选择查看报告，如图 243 所示。

- 模型预测配置如图 244 所示。
- 模型预测结果数据如图 245 所示。

The screenshot displays a workflow on the left and a configuration panel on the right. The workflow consists of the following steps: 输入源 (Input Source) -> 数据拆分 (Data Splitting) -> LASSO回归 (LASSO Regression) -> 模型预测 (Model Prediction). A second 输入源 (Input Source) node is connected to the 模型预测 (Model Prediction) node. The configuration panel on the right is titled '字段属性' (Field Properties) and contains a table with the following data:

字段	类型	取值范围
income	数值	9-115
educational_level	数值	1-5
gender	数值	0-1

Below the table, the '因变量' (Dependent Variable) is set to 'educational_level'. The '参数设置' (Parameter Settings) section is partially visible at the bottom.

图 240

The screenshot shows the '参数设置' (Parameter Settings) section of the configuration panel. It includes the following parameters:

- alpha: 1.0
- 最大迭代次数 (Maximum Iterations): 500

图 241

预览数据

e_pay	package_type	leave	educational_level	predict
1	2	1	3	2.511665096641166
1	4	1	3	2.9099671407662324
0	3	1	3	2.5497598942332087
0	3	0	4	2.3668517221160155
0	3	0	1	1.7312006085807385
1	1	0	3	2.486560174313897
1	2	1	3	2.5562199145982922
0	2	0	3	2.517115873107118

共 597 条 25 条/页 < 1 2 3 4 5 6 ... 24 > 前往 1 页

图 242

算法运行报告

Lasso回归模型评价

模型拟合效果

intercept(截距项): 2.8317241284955323
 Coefficients(系数): [-0. -0.00658748 -0. 0.00176153 -0. -0. -0.00472689 0. -0. -0.00698292 0.02021693 0. 0. 0.]
 Residual sum of squares(均方误差): [-0. -0.00658748 -0. 0.00176153 -0. -0. -0.00472689 0. -0. -0.00698292 0.02021693 0. 0. 0.]
 Score(R^2): [-0. -0.00658748 -0. 0.00176153 -0. -0. -0.00472689 0. -0. -0.00698292 0.02021693 0. 0. 0.]

图 243



图 244

预览数据

wifi	e_pay	package_type	leave	predict_label
0	0	1	1	2.5931635936971142
0	0	3	0	2.5782271848909155
0	0	3	0	2.3857340321559763
0	0	3	0	2.3202627521774177
0	0	2	0	2.367039087265935
0	1	1	1	2.810894647820224
0	0	3	0	2.555372705175261
0	0	1	0	2.452773978216377

共 796 条 25 条/页 < 1 2 3 4 5 6 ... 32 > 前往 1 页

图 245

3.4.4.8 多项式回归

图标:

 多项式回归

描述: 多项式回归是研究一个因变量与一个或多个自变量间多项式的回归分析方法。

字段属性

特征列: 通过勾选的方式选择特征所在列。如图 246 所示。

字段属性

字段

添加字段过滤字符串

<input checked="" type="checkbox"/>	字段	类型	取值范围
<input checked="" type="checkbox"/>	sepal_length	数值	4.4-7.7
<input checked="" type="checkbox"/>	sepal_width	数值	2.2-3.9
<input checked="" type="checkbox"/>	petal_length	数值	1.2-6.7
<input checked="" type="checkbox"/>	petal_width	数值	0.1-2.5

图 246

参数设置

degree: 多项式的阶数，默认为 2。

`interaction_only`: 是否产生相互影响的特征集, 默认为 `False`。

`include_bias`: 是否包含偏差列, 默认为 `True`。

如图 247 所示。



参数设置

多项式的阶数 ?

2

是否产生相互影响的特征集

否

是否包含偏差列

是

图 247

输出

表结果: 特征矩阵。

报告: 无。

示例

下面对某两列数据拟合多项式回归。

- 选择两列待拟合序列, 数据必须为数值型。如图 248 所示。
- 点击参数设置, 设置如图 249 所示。
- 运行该组件, 对组件右击, 选择查看报告, 结果如图 250 所示。

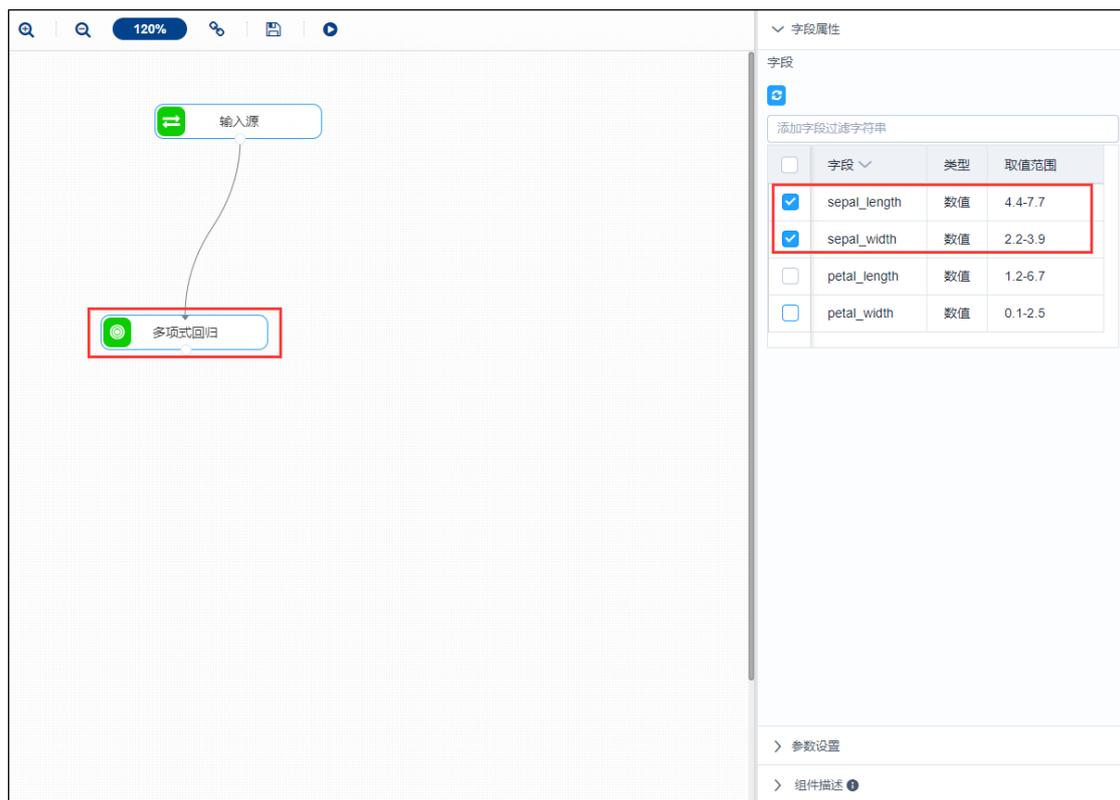


图 248

The '参数设置' (Parameter Settings) panel for the polynomial regression model includes the following configuration options:

- 多项式的阶数 (Polynomial Degree): 2
- 是否产生相互影响的特征集 (Generate interacting feature sets): 否 (No)
- 是否包含偏差列 (Include bias column): 是 (Yes)

图 249

0	1	2	3
1	5.1	3.5	17.85
1	4.9	3	14.7
1	4.7	3.2	15.04
1	4.6	3.1	14.26
1	5	3.6	18
1	5.4	3.9	21.06
1	4.6	3.4	15.64
1	5	3.4	17

共 150 条 25 条/页 < 1 2 3 4 5 6 > 前往 1 页

图 250

3.4.4.9 SVR

图标: 

描述: SVR (支持向量回归) 是使用支持向量机解决回归问题。支持向量回归假设我们能容忍的 $f(x)$ 与 y 之间最多有 ϵ 的偏差, 当且仅当 $f(x)$ 与 y 的差别绝对值大于 ϵ 时, 才计算损失, 此时相当于以 $f(x)$ 为中心, 构建一个宽度为 2ϵ 的间隔带, 若训练样本落入此间隔带, 则认为被预测正确的。

字段属性

特征列: 通过勾选的方式选择特征所在列。

标签列: 选择分类标签所在的列。

参数设置

罚项系数: 浮点型, 默认 1.0。

核函数: 支持线性核、多项式核、高斯核、sigmoid, 默认高斯核。

输出

表结果: SVR 回归算法结果。

报告: Regression model evaluation。

示例

下列对某数据进行 SVR 回归算法:

- 选择自变量，因变量。如图 251 所示。
- 保留默认参数，罚项系数为 1.0，核函数为高斯核，如图 252 所示。
- SVR 运行成功后，选择查看数据，如图 253 所示。
- SVR 成功后，选择查看报告，如图 254 所示。
- 模型评估配置如图 255 所示。
- 模型评估运行成功后，选择查看数据，如图 256 所示。
- 模型评估运行成功后，选择查看报告，如图 257 所示。
- 模型预测配置如图 258 所示。
- 模型预测运行成功后，选择查看数据，如图 259 所示。



图 251

> 字段属性

∨ 参数设置

罚项系数

1.0

核函数

高斯核

图 252

预览数据

wifi	package_type	leave	educational_level	predict_value
0	2	1	3	2.900089350037783
28	4	1	3	2.900165396438463
0	3	1	3	2.899454378940668
0	3	0	4	3.640831314689791
0	3	0	1	1.6390969989871498
0	1	0	3	2.9000916983609377
0	2	1	3	2.899892927341849
0	2	0	3	2.900096154261487

共 597 条 25 条/页 < 1 2 3 4 5 6 ... 24 > 前往 1 页

图 253

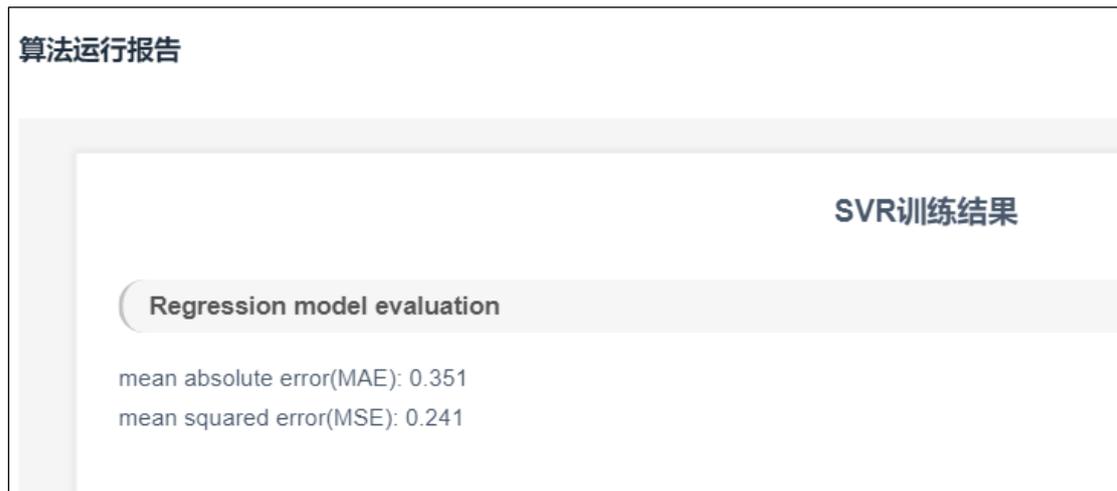


图 254



图 255

预览数据

wifi	package_type	leave	educational_level	predict_label
0	3	0	1	2.638469386649839
0	1	0	3	2.639096998987146
0	2	1	3	2.675724358634812
33	4	0	5	2.618765413080748
0	3	0	1	2.6398329951047184
28	1	0	1	2.6390687750467077
0	4	0	3	2.519184995521833
0	2	0	2	2.389312389681224

共 199 条 25 条/页 < 1 2 3 4 5 6 ... 8 > 前往 1 页

图 256

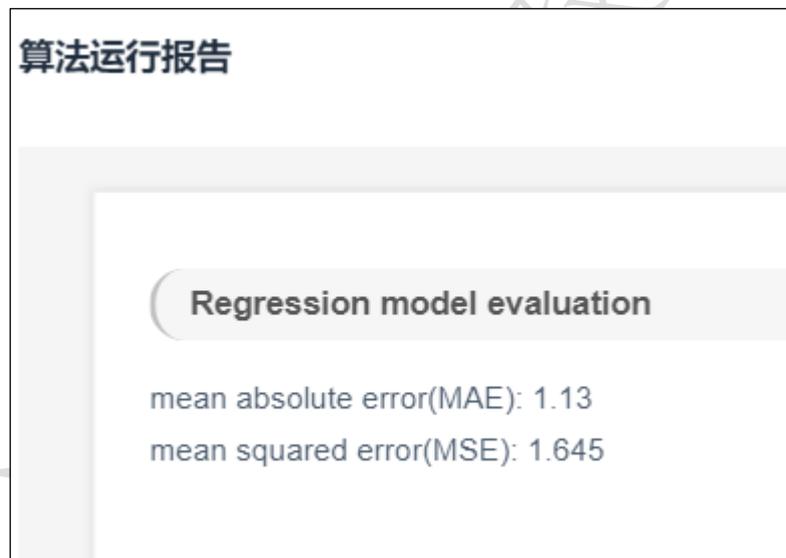


图 257

字段属性

特征列

添加字段过滤字符串	字段	类型	取值范围
<input checked="" type="checkbox"/>	age	数值	20-64
<input checked="" type="checkbox"/>	marital_status	数值	0-1
<input type="checkbox"/>	educational_level	数值	1-5
<input checked="" type="checkbox"/>	gender	数	0-1

图 258

预览数据					
	free	wifi	package_type	leave	predict_label
	0	0	1	1	3.6385220234615288
	0	0	3	0	2.3884333396494406
	19	0	3	0	2.639087524707017
	29	0	3	0	3.639098388126521
	0	0	2	0	1.6357231740355462
	0	0	1	1	3.6394269312282623
	22	0	3	0	2.099688958452499
	0	0	1	0	2.1000826991644876

共 796 条 25 条/页 < 1 2 3 4 5 6 ... 32 > 前往 1 页

图 259

3.4.4.10 KNN 回归

图标:



描述: KNN 进行回归

字段属性:

特征列: 通过勾选的方式选择特征所在列。

标签列: 仅支持数值型数据。

参数设置:

最近邻个数 K: 整数型, 通常不大于 20, 默认 5。

投票权重类型: 权重相等或权重与距离成反比, 默认权重相等。

计算最近邻的算法: 包括 自动、BallTree、KDTree、暴力搜索法, 默认自动。

输出

表结果: KNN 回归算法结果。

报告: Regression model evaluation。

示例

下列对某数据进行 KNN 回归算法:

- 选择自变量, 因变量。如图 260 所示。
- 保留默认参数, 最近邻个数为 5, 投票权重类型为权重相等, 计算最近邻的计算为

自动，如图 261 所示。

- 运行成功后，选择查看数据，如图 262 所示。
- 运行成功后，选择查看报告，如图 263 所示。
- 模型评估配置如图 264 所示。
- 模型评估运行成功后，选择查看数据，如图 265 所示。
- 模型评估运行成功后，选择查看报告，如图 266 所示。
- 模型预测配置如图 267 所示。
- 模型预测运行成功后，选择查看数据，如图 268 所示。

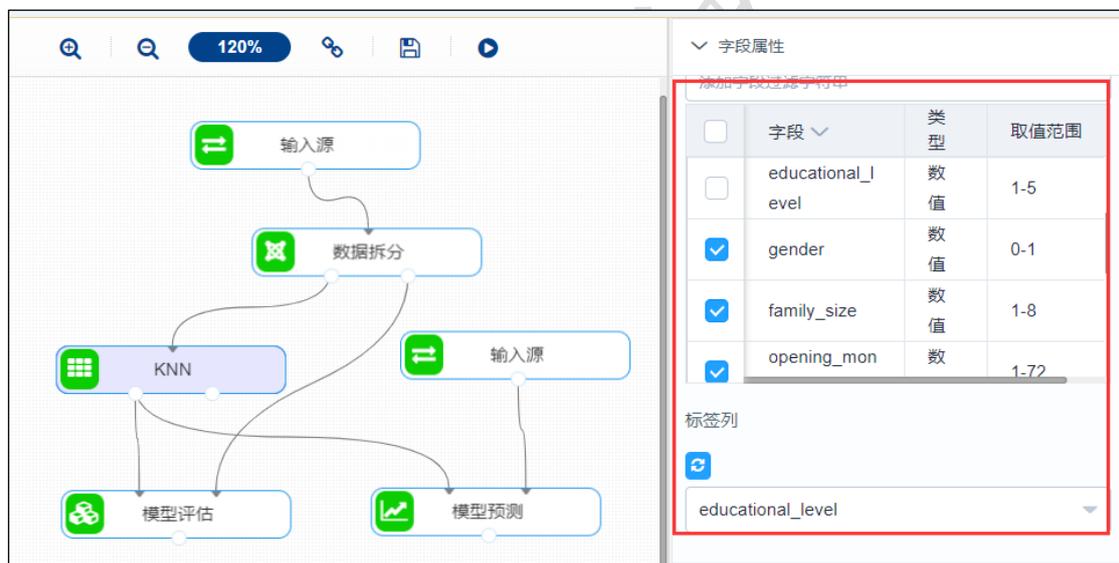


图 260

> 字段属性

∨ 参数设置

最近邻个数K

5

投票权重类型

权重相等

计算最近邻的算法

自动

图 261

预览数据

e_pay	package_type	leave	educational_level	predict_label
1	2	1	3	2.6
1	4	1	3	2.8
0	3	1	3	2.4
0	3	0	4	2.8
0	3	0	1	1.6
1	1	0	3	2.2
1	2	1	3	2.8
0	2	0	3	1.8

共 597 条 ...

图 262

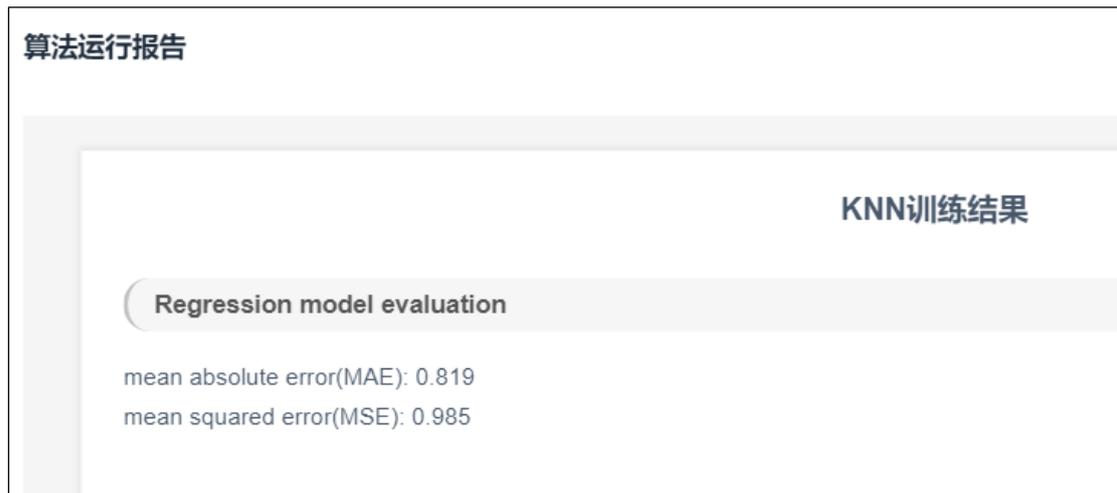


图 263

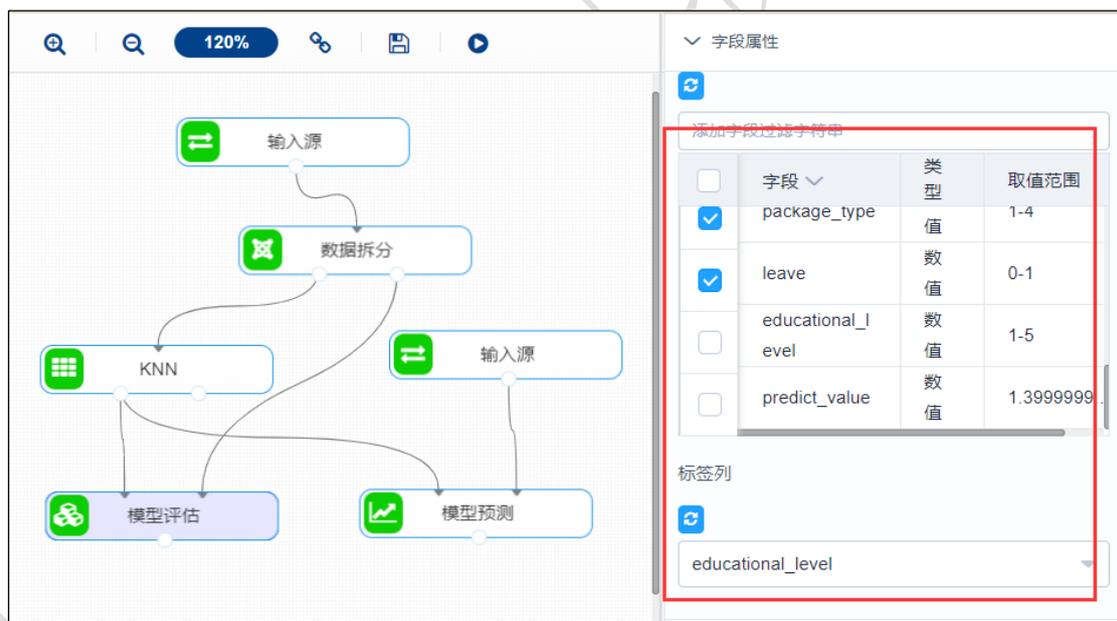


图 264

预览数据

	e_pay	package_type	leave	educational_level	predict_label
	0	3	0	1	2.6
	0	1	0	3	3
	1	2	1	3	2.8
	0	4	0	5	2.8
	0	3	0	1	2.2
	1	1	0	1	3
	1	4	0	3	2.2
	1	2	0	2	2.2

共 199 条 25 条/页 < 1 2 3 4 5 6 ... 8 > 前往 1 页

图 265

算法运行报告

KNN测试结果

Regression model evaluation

mean absolute error(MAE): 1.034
mean squared error(MSE): 1.568

图 266



图 267

预览数据

wifi	e_pay	package_type	leave	predict_label
0	0	1	1	3
0	0	3	0	2
0	0	3	0	2.4
0	0	3	0	2.4
0	0	2	0	2.4
0	1	1	1	4
0	0	3	0	3.2
0	0	1	0	2.8

共 796 条 25 条/页 < 1 2 3 4 5 6 ... 32 > 前往 1 页

图 268