

智能建模功能简介

Al Model



www.raqsoft.com.cn





目录 CONTENTS

本地数据文件
 数据库



数据源





智能建模支持txt、csv等格式的 数据文件。

Data Source foc	auon					
Local data	file	 Database type 		 Remote 	server	
Look <u>I</u> n: 📔	data		•			
🚞 tmp						
	et cev					
📄 titanic_te	SLUSY					
titanic_te	ain.csv					
titanic_te	ain.csv					
titanic_te	ain.csv					
titanic_te	ain.csv					
titanic_te	ain.csv					
titanic_te	ain.csv					
titanic_te	ain.csv					
titanic_te	ain.csv					
titanic_te	ain.csv					
file <u>N</u> ame:	ain.csv					



R

选择文件后,可以定义数据文件的参数配置。

reate data file name	e titanic_train.mtx		Preview data			Preview the top 100	lines 🕂 I	Reloa
Import the first I	line as variable name		Passengerld	Survived	Pclass	Name	Sex	Ag
Omit all quotation	on marks		1			Braund, Mr. Owen Harris		ļ
Check Column	Count		2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	
Delete a line wi	hen column count does not match value count at line 1		3	1	3	Heikkinen, Miss. Laina	female	
Use double que	otation marks as escape characters		4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	
Delimiter		T	5	0	3	Allen, Mr. William Henry	male	
	-		6	0	3	Moran, Mr. James	male	
Charset	GBK		7	0	1	McCarthy, Mr. Timothy J	male	
Date format	yyyy/MM/dd	•	8	0	3	Palsson, Master. Gosta Leonard	male	
ime format	HHimmiss		9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	
			10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	
Date time format	yyyy/MM/dd HH:mm:ss	•	11	1	3	Sandstrom, Miss. Marguerite Rut	female	
ocale	English		12	1	1	Bonnell, Miss. Elizabeth	female	
			13	0	3	Saundercock, Mr. William Henry	male	
issing values (bar		_	14	0	3	Andersson, Mr. Anders Johan	male	
issing values (bai	HOLLINA		15	0	3	Vestrom, Miss. Hulda Amanda Adolfina	female	
			16	1	2	Hewlett, Mrs. (Mary D Kingcome)	female	
			17	0	3	Rice, Master. Eugene	male	
			18	1	2	Williams, Mr. Charles Eugene	male	
			19	0	3	Vander Planke, Mrs. Julius (Emelia Maria Vandemoortele)	female	
			20	1	3	Masselmani, Mrs. Fatima	female	
			21	0	2	Fynney, Mr. Joseph J	male	
			-			1		10





下一步,可以定义变量类型、日期格式和选出状态。

变量类型既可以自动检测,也可以导入数据字典配置。 数据字典格式如下:

Name	Туре	DateFormat	Used	Importance
Passengerld	Identity		TRUE	0
Survived	Binary		TRUE	0
Pclass	Categorical		TRUE	0
Name	Text		FALSE	0
Sex	Binary		TRUE	0
Age	Numerical		TRUE	0
SibSp	Categorical		TRUE	0

<u>K</u> Load data Import data dictionary Note: Unselected variables won't be imported. NO. Variable name Date format Select Type Passengerld \checkmark 1 Automatic Survived 2 Automatic \checkmark Pclass 3 \checkmark Automatic Name \checkmark 4 Automatic 5 Sex \checkmark Automatic \checkmark 6 Age Automatic 7 SibSp \checkmark Automatic \checkmark 8 Parch Automatic 9 Ticket \checkmark Automatic 10 Fare \checkmark Automatic 11 Cabin \checkmark Automatic 12 Embarked \checkmark Automatic





在数据源窗口中,可以定义JDBC和ODBC两种数据源连接。

K Datasource	\times
	Connect
Database type	× nect
Database type JDBC ODBC 	<u>Q</u> K v <u>C</u> ancel te t
Encryption level Plaintext	<u>C</u> lose





JDBC数据源

Datasource		×
General properties Extended p	roperties	<u>о</u> к
Datasource name	Database vendor	Cancel
orcl	ORACLE	
Driver		
oracle.jdbc.driver.OracleDriver	•	
Datasource URL: Remember to repla	ace the contents in brackets	
jdbc:oracle:thin:@127.0.0.1:1521:or	cl 💌	
User	Password	
System	****	
Batch size	0	
Qualify object with schema	Enclose object name in quotes	

ODBC数据源

ODBC datasource		×
Datasource name	access	<u>о</u> к
ODBC name	•	Cancel
Username		
Password		
Qualify object with	nschema	
Case sensitive		
Enclose object na	ame in quotes	





接下来可以使用配置好的数据源,编辑SQL语句进行取数。

<u>K</u> Load data				\times	
Data source location					
O Local data file					
Create data file name scores mtx					
Table Field Where	e Group Having	Sort SQL]		
Available table			Selected table		
SCORES			SCORES		
		>			
		<			
Data source orcl	•	Schema	WN	T	
			<u>o</u> k	<u>C</u> ancel	

Data source location		
O Local data file	 Database type 	◯ Remote server
Create data file name sco	res.mtx	
Table Field Where G	oup Having Sort SQL	
ELECT * FROM SCORES		
•		
Data source	Schema WN	

目录 CONTENTS

基本特征 离散变量统计 连续变量统计 数据探索报告

5. 数据质量报告



数据探索

● 1. 基本特征

导入数据以后,显示了数据的基本 特征:

目标变量是Survived (需要用户设置),有12个变量,891条记录。

自动解析了各个变量的类型和推荐 的选出状态。

Model file 🛛 titanic_train.pcf 🛛 🚰 🕍 Model performance 📲 Model presentation 📝 Model options							
Data file titanic_train.mtx 😭 Reload data							
Target variable Survived Set 🍸 Variable filter 🛧 🦊							
NO.	Variable name	Туре	Date format 🗹 Select				
1	Passengerid	ID	\checkmark				
2	Survived	Binary variable	\checkmark				
3	Pclass	Categorical variable	\checkmark				
4	Name	ID					
5	Sex	Binary variable	\checkmark				
6	Age	Numerical variable	\checkmark				
7	SibSp	Categorical variable	\checkmark				
8	Parch	Categorical variable	\checkmark				
9	Ticket	Categorical variable	\checkmark				
10	Fare	Numerical variable	\checkmark				
11	Cabin	Categorical variable	\checkmark				
12	Embarked	Categorical variable	V				
Search variable	•		Import 891 rows, 12 variables				





智能建模的变量类型有以下几种:

变量类型	描述
数值变量	取值为实数的变量
单值变量	只包含一个类别的变量 (不含缺失值)
二值变量	只包含两个类别的变量 (不含缺失值)
计数变量	取值为自然数的变量
分类变量	分类数大于二的变量 (不含缺失值)
ID	唯一标识符
时间日期	日期、时间或日期时间变量
长文本	长度超过128字节且分类数特别多的变量

智能建模的目标变量支持二值变量、数值变量、计数变量和分类变量。

离散变量包括单值变量、二值变量和 分类变量。

缺失率:缺失值在全部数据中的占比。 势:离散变量可取值集合的成员数量。 饼图直观显示了各分类的占比。



● 2. 离散变量统计







目标变量是二值变量:分组目标频数表

分组目标频数表将样本按分类值分组, 观察每组样本的数量和正样本数,正 样本率以及赔率(发生比)。

二值目标的正样本是指样本数较少的 分类值。通过右图可以看到,在本例 中正样本是目标变量值为1的记录。

Pie chart Contingency table Odds wrt All								
Categorical Level	Frequency	Positive Frequency	Positive Ratio	Odds wrt All				
S	644	217	33.696%	0.878				
С	168	93	55.357%	1.442				
Q	77	30	38.961%	1.015				
NULL	2	2	100%	2.605				
All	891	342	38.384%	1				







R

目标变量是二值变量:分组目标频数表

Odds wrt All 图形显示了每组样本的 赔率和总的赔率。样本较少的分类 (样本数少于100个)不进行绘制。





R

目标变量是数值变量:分组目标统计量,分组目标统计图

分组目标统计量将样本按分类值分组, 观察每组样本的统计量。包括:频率, 平均值,标准差,中位数,最小值和 最大值以及Z-STAT。

分组目标统计图,使用箱线图的形式, 更直观的表现了每组样本的分布情况。 箱线图可以用来标记异常值。

Pie chart Relation with Target Relation graph with target								
Categorical variable	Frequency	Average	Standard deviation	Median	Minimum	Maximum	Z-STAT	
NULL	1369	183452.131	80667.145	165000	34900	755000	1.19	
Grvl	50	122219.08	34780.781	119000	52500	256000	-5.276	
Pave	41	168000.585	38370.375	171900	40000	265979	-1.051	





连续变量包括数值变量、计数变量 和时间日期变量。

描述性统计量显示了数据的基本统 计信息。

频数分布图,绘制了频数分布直方 图,正态分布曲线,以及箱线图。







目标变量是二值变量:分组描述性统计量

分组描述性统计量,将样本按目标变 量值分组,分别进行统计,并绘制相 应的箱线图。

statistics	Histogram	Relationship with target	Histogram with target		
		Target=0		Target=1	Total
lissing rat	e	22.769%	· · · · ·	15.205%	19.865%
Average		30.626		28.344	29.699
dard devia	ation	14.172		14.951	14.526
Skewness		0.584		0.18	0.388
Minimum		1.0		0.42	0.42
ower quarti	le	21.0		19.0	20.0
Median		28.0		28.0	28.0
pper quarti	le	39.0		36.0	38.0
Maximum		74.0		80.0	80.0
Z-STAT		1.675		-1.933	-
	Ilissing rat Average dard devia Skewness Minimum wer quarti Median oper quarti Maximum Z-STAT	statistics Histogram	Itatistics Histogram Relationship with target Issing rate 22.769% Average 30.626 dard deviation 14.172 Skewness 0.584 Minimum 1.0 wer quartile 21.0 Median 28.0 oper quartile 39.0 Maximum 74.0 Z-STAT 1.675	Itatistics Histogram Relationship with target Histogram with target Issing rate 22.769% Income and target Income and target	Itistogram Relationship with target Histogram with target Image: Target=0 Target=1 Itissing rate 22.769% 15.205% Average 30.626 28.344 dard deviation 14.172 14.951 Skewness 0.584 0.18 Minimum 1.0 0.42 wer quartile 21.0 19.0 Median 28.0 28.0 opper quartile 39.0 36.0 Aximum 74.0 80.0 Z-STAT 1.675 -1.933





目标变量是二值变量:分组频数分布图

分组频数分布图,将每个区间的样本 按目标变量值分组,频数用不同颜色 显示。







目标变量是数值变量:目标变量相关系数

Pearson相关系数:用于描述两个连续 变量之间的线性相关性。

Descriptive statistics	Histogra	m	Correlation	Scatte	r Plot	
Pearson			Spearman			
	0.7086				0.7313	3

Spearman**秩相关系数**:用于描述两个 连续变量之间的等级相关性。

上图是地下室面积和房价之间的相关系数。 可以看到两者有很强的相关性。

相关系数的绝对值越大,表示两个变量的相关性越大。



R

目标变量是数值变量: 单因素散点图

单因素散点图直观的展现了当 前变量(地下室面积)和目标 变量(房价)相关的分布情况。 其中黄线为回归线。







提供导出数据探索报告到excel文件的功能。第一页是各类变量的基本信息:

	А	В	С	D	E	F	G	Н	l I	J
1	1.Numerical Variable:									
2		Variable Name	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Missing Rate	Skewness
3		Age	0.42	20.0	28.0	29.699	38.0	80.0	19.865%	0.388
4		Fare	0.0	7.896	14.454	32.204	31.0	512.329	0.0%	4.779
5	2.Count Variable: None	e								
6	3.Categorical Variable	(Binary and Unary)								
7		Variable Name	Cardinality	Missing Rate						
8		Survived	2	0.0%						
9		Pclass	3	0.0%						
10		Sex	2	0.0%						
11		SibSp	7	0.0%						
12		Parch	7	0.0%						
13		Ticket	681	0.0%						
14		Cabin	148	77.104%						
15		Embarked	4	0.224%						
16	4.ID:									
17		Variable Name								
18		Passengerld								
19		Name								
20	5.Date Time: None									
21	6.Text String: None									
22										
23										
24										
25										
26										
27										
28										
29										
30										
	Variables	With Binary Tar	get Variable	+						•



R

第二页是各类变量与目标变量的关联性:

	А	В	С	D	E	F	G	Н	I	J	К
1	1.Numerical Variable:										
2						Ac	ge		-		
3		Target	Frequency	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Missing Rate	Z_STAT wrt Overall
4		1	342	0.42	19.0	28.0	28.344	36.0	80.0	15.205%	-1.933
5		0	549	1.0	21.0	28.0	30.626	39.0	74.0	22.769%	1.675
6											
7						Fa	re				
8		Target	Frequency	Minimum	Lower Quartile	Median	Mean	Upper Quartile	Maximum	Missing Rate	Z_STAT wrt Overall
9		1	342	0.0	12.475	26.0	48.395	57.0	512.329	0.0%	6.232
10		0	549	0.0	7.854	10.5	22.118	26.0	263.0	0.0%	-4.919
11	2.Count Variable: None										
12	3.Categorical Variable (B	<u> Binary, Una</u>	iry):								
13				Pclass		1					
14		Category	Frequency	Positive Frequency	Positive Ratio	Odds wrt All					
15		1	216	136	62.963%	1.64					
16		2	184	87	47.283%	1.232					
17		3	491	119	24.236%	0.631					
18		All	891	342	38.384%	1.0					
19											
20				Sex		1					
21		Category	Frequency	Positive Frequency	Positive Ratio	Odds wrt All					
22		female	314	233	74.204%	1.933					
23		male	577	109	18.891%	0.492					
24		All	891	342	38.384%	1.0					
25											
26				SibSp							
27		Category	Frequency	Positive Frequency	Positive Ratio	Odds wrt All					
28		0	608	210	34.539%	0.9					
29		1	209	112	53.589%	1.396					
30		2	28	13	46.429%	1.21					
	▶ Variables	With Bina	ary Target \	/ariable 🕂 🕂						: [•



提供导出数据质量报告到pdf文件的功能。部分内容如下:

These observations from 891 unique ID. Since the number of unique Id is equal to the number of observations, time sensitive information can not be studied using this data set. Variables that have all "empty" value are not exist.

Variables that have more than 99% of missing values are not exist.

Variables that have missing values between 95% and 99% are not exist.

Table 1 Missingness Analysis						
Missing Percentage	Number of Variables	% of All Numerical Variables				
100%	0	0%				
99% to 100%	0	0%				
95% to 99%	0	0%				
90% to 95%	0	0%				
80% to 90%	0	0%				
70% to 80%	0	0%				
60% to 70%	0	0%				
50% to 60%	0	0%				
30% to 50%	0	0%				
10% to 30%	1	20%				
Below10%	4	80%				

The highly positive skewness (with skewness > 10) numerical variables are not exist. The highly negative skewness (with skewness < -10) numerical variables are not exist.

Table 2 Skewness of Numerical Variables					
Skewness Range	Number of Variables	% of All Numerical			

Table 2 Skewness of Numerical Variables						
Skewness Range	Number of Variables	% of All Numerical Variables				
10 +	0	0%				
5 to 10	0	0%				
2 to 5	3	60%				
1 to 2	0	0%				
-1 to 1	2	40%				
-2 to -1	0	0%				
-5 to -2	0	0%				
-10 to -5	0	0%				
-10-	0	0%				
Total	5	100%				

All categorical variables with cardinality over 512 are Name, Ticket. The calculation of cardinality includes missing category.

The following categorical variables have cell frequency less than 100:

Name, TicketSurvived, Pclass, Sex, Embarked.

目录 CONTENTS

- 1. 自动预处理
 2. 预处理报告
 3. 预处理流程
- 4. 手动预处理

预处理





智能建模的预处理过程集成在建模的流程中,一键式自动预处理。

K Build model	×
INFO: Start checking data and categorical conversion.	
[2020-02-09 10:14:58]	D
INFO: Time for checking data and categorical conversion: 109 milliseconds	
[2020-02-09 10:14:58]	
INFO: Start preparing.	
[2020-02-09 10:14:58]	
INFO: Modeling data preparing10%	
[2020-02-09 10:14:59]	
INFO: Modeling data preparing20%	
[2020-02-09 10:14:59]	
INFO: Modeling data preparing30%	
[2020-02-09 10:14:59]	
INFO: Modeling data preparing40%	-
Tog View log 📑 Export report 😋 Model presentation 🛍 Model performance 对 Open model	directory





建模结束后可以导出模型报告,描述了预处理执行了哪些动作。部分内容如下:

Target variable: Survived, ID variable: PassengerId.

The number of fields before pretreatment: 12, the number of fields after pretreatment: 11. The number of fields with missing values before pretreatment: 3 and the number of fields with missing values after pretreatment: 0.

Total rows of data: 891, where deleted rows due to missing target: 0.

Variable selection table						
	Number of selections	Number not selected	Total number			
All variables	11	1	12			
Unary variables	0	0	0			
Binary variables	2	0	2			
Category variables	4	1	5			
Numerical variables	2	0	2			
Counting variables	2	0	2			
Datetime variables	0	0	0			

Variables Processing Information

Variable name: PassengerId. The type is ID Variable name: Pclass. The type is Category variables

Number of categories: 3

The variable fills the missing value by using the yimming intelligent filling algorithm. There are 3 categories are merged because of low frequency. Generation Category Derivative Variables: BI_Pclass_1, BI_Pclass_2 Variable name: Sex. The type is Binary variables Number of categories: 2 The variable fills the missing value by using the yimming intelligent filling algorithm. There are 2 categories are merged because of low frequency. Generation Category Derivative Variables: BI_Sex_1 Variable name: Age. The type is Numerical variables Skewness: 0 Average:29.699

Median:24Variance:13.002The variable fills the missing value by using the yimming intelligent filling algorithm.Variable name: SibSp. The type is Counting variablesSkewness: 0Average:0.523Median:0Variance:1.103The variable fills the missing value by using the yimming intelligent filling algorithm.Variable name: Parch. The type is Counting variables





(1) 检查变量值域

检查并记录所有变量的值域,若测试数据出现训练数据没有的分类或者超出数值范围, 进行针对性的处理。

(2) 时间日期变量处理

检查所有时间日期型变量,创建若干常用的衍生变量。并检测时间日期变量的关联性, 创建多日期联动的衍生变量。

(3) 缺失值信息提取

若数据存在缺失值,提取并记录缺失值模式,将缺失值所表现出的行为特征转换为衍 生变量加以利用。





(4) 缺失值填补

若数据存在缺失值,利用简单或个性化智能算法,填补缺失值。

(5) 分类变量降噪

针对分类变量可能存在的噪音,例如极少数分类,异常分类,疑似错误分类等情况,进行针对性处理。

(6) 分类变量数值化

将分类变量转换为可正常进行运算的数值型变量。主要方式是dummy variable和平滑化,由算法智能判断。





(7) 纠偏

针对部分存在正态性假设的模型,对高偏态变量进行数学变换,使偏度回到0附近,满 足模型假设。

(8) 异常值处理

探测并识别可能存在的异常值,并进行针对性处理。



以较宽松的门槛, 剔除掉对建模无用的变量, 降低时间成本和模型复杂度。





(10) 标准化/归一化

数据标准化/归一化, 消除口径差异。有利于神经网络等模型的寻优求解。

(11) 平衡样本

对于二分类数据,若正负样本比例严重不均衡,会按照指定的比例配平,并智能重采 样建模。





选择变量

根据变量类型去除一些无关的变量。 例如ID和长文本,没有缺失值的单值 变量等。



根据变量重要度筛选变量,只保留重 要度较高的变量。变量重要度可以由 数据字典导入,也可以通过建模得到。







衍生变量

用变量姐妹、配偶数量 "SibSp" 和 变量父母、子女数量 "Parch" 相加得到家庭成员数量 "Family"。可以看到家庭成员在1-3人时幸存率较高。

K Add computed variable						
Computed variable name Family						
Normal Advance						
'SibSp'+'Parch'						
Variable Function						
Variable name		Variable information				
Pclass	A .	Statistical method	Statistical value			
Name		Missing rate	0%			
Sex		Minimum	0			
Age		Maximum	6			
SibSp		Average	0			
Parch		Upper quartile	0			
Ticket		Median	0			
Fare		Lower quartile	0			
Embarked	Embarked Standard deviation 0.806					
Family	•	Skewness	2.744			
1		1	QK Cancel			

增加衍生变量Family



统计变量Family





衍生变量

可以将数值变量通过分箱离散化,转换为分类变量。以年龄为例,分为0,8,18,35,60几个年龄段, 生成衍生变量,并对其进行统计。





Pie chart Conting	gency table Oc	lds wrt All		
Categorical Level	Frequency	Positive Frequency	Positive Ratio	Odds wrt All
4.21	54	36	66.667%	1.737
47.5	195	78	40%	1.042
13.0	85	34	40%	1.042
26.5	358	137	38.268%	0.997
NULL	177	52	29.379%	0.765
70.0	22	5	22.727%	0.592
All	891	342	38.384%	1

统计变量AgeArea

可以看到0-8岁的少年幸存率最高,青少年、青年和中年的区分不大, 老年幸存率最低。



R

预处理选项

在模型选项中可以定义是否数据预处理 和是否智能填补。

如果数据已经进行过预处理,可以取消 数据预处理。

智能填补可以更好的对缺失值进行补缺, 但是会消耗更多的硬件资源和时间,当 数据量很大时不建议智能填补。不勾选 时会进行简单填补。

Model options		×
Normal Binary model Regression model Multiclassif	ication model	
☑ Data preparation ☑ Intelligent impute		
Resampling Number of samples 5	Best number of sample combinations	3 🔺
Balanced sampling ratio 1:1	Sample multiplier	150 🛓
Ensemble method Optimal model strategy	Best number of ensembles	0
Ensemble function np.mean	Model evaluation criterion	▼
Percentage of test data Automatic		
Adjust scoring results	✓ Set random seeds	0 🛓
		OK Cancel

目录 CONTENTS

建模流程
 2. 智能建模
 3. 专业建模







在使用传统工具时,通常需要有统计学基础的专业人员,不断选择算法,调整模型参数,最终 得到符合期望的模型。建模的流程如下:





智能建模工具无须统计学知识,一键式智能建模,优选模型组合和模型参数都在内部实现。

×

K Build model

2020-02-03 10.13.20,000 - ymmg model.cp37-wm amdo4.pygnine.sof - ini O. mioderinied successium 2020-02-09 10:15:20,085 - yiming model.cp37-win_amd64.pyd[line:90] - DEBUG: feature importance of YiModel: {'Rank_F are": 1.0, 'Pow0 69 Age": 0.6549645954182951, 'MI Age": 0.43832267557855187, 'Rank SibSp": 0.39012433562963306, Rank Parch': 0.0} 2020-02-09 10:15:20,085 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: performance of each base model in Yi Model: {'XGBClassification 1': 0.8122683142100618, 'RFClassification 1': 0.7518387761106208, 'FNNClassification 1': 0 .5. 'RidgeClassification 1': 0.757811120917917, 'TreeClassification 1': 0.7086201824065902, 'LogicClassification 1': 0.7 496322447778758, 'CNNClassification 1': 0.5, 'GBDTClassification 1': 0.7994998528979111} 2020-02-09 10:15:20,085 - interface_library.cp37-win_amd64.pyd[line:90] - INFO: Calculate predict value on test data 2020-02-09 10:15:20,132 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: predict value on test data: 2020-02-09 10:15:20,132 - interface_library.cp37-win_amd64.pyd[line:90] - INFO: Calculate ensemble performance 2020-02-09 10:15:20,132 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: ensemble performance: 0.820535 2020-02-09 10:15:20,132 - interface_library.cp37-win_amd64.pyd[line:90] - INFO: Writing out results 2020-02-09 10:15:20,132 - interface library.cp37-win_amd64.pyd[line:90] - DEBUG: writing out predict values 2020-02-09 10:15:20,132 - interface_library.cp37-win_amd64.pyd[line:90] - DEBUG: writing out model 2020-02-09 10:15:20,155 - interface library.cp37-win_amd64.pyd[line:90] - DEBUG: writing out feature importance 2020-02-09 10:15:20,155 - interface library.cp37-win amd64.pyd[line:90] - DEBUG: writing out modeling information 2020-02-09 10:15:20,155 - interface library.cp37-win amd64.pyd[line:90] - INFO: Build model finished

Tog View log 📑 Export report 🖙 Model presentation 🛍 Model performance 💕 Open model directory



智能建模开放了模型参数,提供给精通模型的专业用户使用。下面是模型的常规选项:

K Model options		×
Normal Binary model Regression model Multiclas	ssification model	
✓ Data preparation ✓ Intelligent impute		
☑ Resampling Number of samples 5 🛖	Best number of sample combinations	3 🛓
Balanced sampling ratio 1:1	Sample multiplier	150
Ensemble method Optimal model strategy	Best number of ensembles	0
Ensemble function np.mean	Model evaluation criterion	T
Percentage of test data Automatic	%	
☑ Adjust scoring results	✓ Set random seeds	0
		OK Cancel

● 3. 专业建模

R

智能建模支持图中几种二分类 算法模型,还可以设置每种模 型是否使用以及抽样次数。在 右侧可以设置各模型的参数值。 对于普通用户可以不用关心这 些设置。

K Me	K Model options								
Norr	mal Binary model	Regression model	Multiclassificati	on model					
NO.	Binary model	Number of samples	Select	NO.	Parameter name	Parameter value			
1	TreeClassification			1	criterion				
2	GBDTClassification	1	✓	2	splitter				
3	RFClassification	1	√	3	max_depth				
4	LogicClassification	1	\checkmark	4	min_samples_split				
5	RidgeClassification	1	\checkmark	5	min_samples_leaf				
6	* FNNClassification	1	\checkmark	6	min_weight_fraction_leaf				
7	XGBClassification	1	\checkmark	7	max_features				
8	* CNNClassification	1	✓	8	max_leaf_nodes				
9	PCAClassification	1	\checkmark	9	min_impurity_decrease				
				10	class_weight				
				11	presort				
The m	odel marked with * is a	a supplementary mod	el that can not b	e configur	ed	<u>O</u> K <u>C</u> ancel			





类似的,我们可以设置回归模型和多分类模型是否使用,以及各自的参数。

KM	odel options					×	K	Мо	del options						×
Nor	mal Binary model	Regression model	Multiclassificatio	on model			N	lorm	al Binary model Regress	sion model Multicla	ssification model				
NO.	Regression model	Number of samples	Select	NO.	Parameter name	Parameter value	N	0.	Multiclassification model	Number of samples	s 🗹 Select	NO.	Parameter name	Parameter v	/alue
1				1			1		XGBMultiClassification		V	1			4
2	GBDTRegression	1	\checkmark	2	splitter		2	2 '	* CNNMultiClassification	1	\checkmark	2	learning_rate		
3	RFRegression	1	✓	3	max_depth							3	n_estimators		
4	LRegression	1		4	min_samples_split							4	booster		
5	LassoRegression	1	✓	5	min_samples_leaf							5	gamma		
6	ENRegression	1	\checkmark	6	min_weight_fraction_leaf							6	min_child_weight		
7	RidgeRegression	1	V	7	max_features							7	max_delta_step		
8	* FNNRegression	1	\checkmark	8	max_leaf_nodes							8	subsample		
9	XGBRegression	1	✓	9	min_impurity_decrease							9	colsample_bytree		
10	* CNNRegression	1	\checkmark	10	presort							10	colsample_bylevel		
11	PCARegression	1	V									11	reg_alpha		
												12	ren lambda		¥
The m	odel marked with * is	a supplementary mod	el that can not be	e configur	ed	<u>O</u> K <u>C</u> ancel	The	e mo	del marked with * is a supple	ementary model that c	an not be configu	ired		<u>o</u> k	<u>C</u> ancel

各模型参数的详细文档: <u>http://doc.raqsoft.com.cn/AIModel/userrefer/jm20.html</u>

目录 CONTENTS

模型表现
 模型描述
 变量重要度







分类模型: 评价指标

智能建模提供了分类模型常用的3个 评价指标:

<u>K</u> Model pe	rformance	×
GINI	AUC	KS
0.785054	0.892527	0.670079

评价指标	描述
GINI	GINI指数在数值上等于2*AUC-1,用于表征模型对正负样本的区分能力。
AUC	AUC相当于ROC曲线下的面积。AUC值越大表示模型越好。
KS	KS值用于衡量模型区分正负样本的能力。KS值越大,模型区分正负样本的 能力越强。



分类模型: ROC曲线

ROC曲线是真正类率与"1-真负类 率"的关系图。ROC曲线可以被视 为评估给定模型所有可能决策性能 的可视化显示。





分类模型:提升度

提升度 (Lift) 表示使用关联规则可 以提升的倍数, 是置信度与期望置信 度的比值。

提升度特别适合有针对性的市场营销 等场景。





分类模型: 查全率

查全率图显示模型找到正样本的情况, 主要应用在数据不平衡的场景。累计 查全率是各组累计正样本数与总正样 本数的比值。





分类模型:准确率表

阈值:用来区分正负样本的值。 准确率:预测正确的样本占所有样本的 比率。

精确率:预测为正样本的结果中,预测正确的比率。

查全率:正确预测正样本的数量,在所 有正样本中的比率。

K Model perform	ance			×
GINI	4	AUC	KS	
0.785054	0.8	92527	0.670079	
ROC Curve Lift Re	call Accuracy			
Lower limit 0.05	Upper limit 0.95 🛉	Number of subsect	ions 19 📩 Set	
Threshold	Accuracy	Precision	Recall	
0.05	0.448	0.41	0.99	
0.1	0.552	0.46	0.942	
0.15	0.653	0.528	0.913	
0.2	0.728	0.599	0.883	
0.25	0.787	0.674	0.864	
0.3	0.813	0.715	0.854	
0.35	0.806	0.714	0.825	
0.4	0.836	0.776	0.806	
0.45	0.828	0.782	0.767	
0.5	0.843	0.843	0.728	
0.55	0.854	0.9	0.699	
0.6	0.854	0.944	0.66	
0.65	0.847	0.943	0.641	
0.7	0.84	0.955	0.612	
0.75	0.81	0.964	0.524	
0.8	0.799	0.962	0.495	
0.85	0.776	0.957	0.437	V
			Cic	ose



多分类模型

目标变量是分类 变量时,模型表 现通过切换预测 值查看每个分类 的模型表现。









回归模型:真实值和转换值

回归模型的表现,分为真实值表现和转换值表现(对数据预处理后的数值)。真实值看起来 比较直观,而转换值对于模型表现的评估更加准确。

K Model per	forma	ance				×
Model performance	evalua	tion type True	response values	•		
R2		MSE	RMSE	GINI	MAE	MAPE
0.921551	35017	70043.358849	18712.830982	0.196404	12929.571303	8.00067
Residual Resu	ilt comp	parison				
End value						
150000	•	150000				 Residual
Start value		100000			•	
-100000		R				
		s 50000				
		d o	والمعتد فسيعرف	and the loss		•
		u u a	1.1.1	100		
		-50000	•••			•
		-100000 50	000 138888.8	89 227777.778	316666.667	405555.556
				Scori	ng	
X-axis variable Sa	alePrice	e 🔻	Start value 5000	0 🔽	End value 450000	•
						Close





回归模型: 评价指标

智能建模提供了回归模型常用的 6个评价指标:

Model performance	evaluation type	True response values	•		
R2	MSE	RMSE	GINI	MAE	MAPE
0.921551	350170043.358	18712.830982	0.196404	12929.571303	8.00067

Model performance	evaluation type	ransformed response	e values 🔻		
R2	MSE	RMSE	GINI	MAE	MAPE
0.911286	0.011907	0.109121	0.016287	0.109121	0.016287

评价指标	描述
R ²	R²是预测值与观测值的误差平方和与观测值和观测均值之差的平方和的比值。
MSE	预测值与真实值偏差的平方和的平均数。
RMSE	MSE的平方根。数量级与真实值相同。
GINI	预测值与真实值偏差的绝对值的平均数。
MAE	预测值与真实值偏差的绝对值的平均数。
MAPE	预测值与真实值偏差比真实值的绝对值的平均数。



回归模型: 残差图

残差是观察值与预测值之差。残差 图是以残差为纵轴,以任一数值变 量为横轴的散点图。图中黄线为三 倍RMSE。

可以调整横轴变量和横纵轴的数值 范围进一步查看。







R

回归模型:结果对照图

结果对照图横轴为随机均分的样本, 纵轴为对应的观察值和预测值。其中 蓝色为观察值,红色为预测值。







模型描述列举了最终选出的模型组合以及每个模型的参数值。通过按钮可以将选中的 模型参数复制到模型选项中,可以进一步优化模型参数。

K Model presentat	tion					×
Ensemble performance	0.892527			Parameter name	Parameter val	ue
Madal nama	0110	C Oslart	max			
WODOLASSIESSIES 4	auc	v select	learn	iing_rate	0.1	
XGBCIassification_1	0.879464	 ✓ 	n_es	timators	150	
* FNNClassification_1	0.873433	\checkmark	objec	ctive	binary:logistic	
RidgeClassification_1	0.872050	\checkmark	boos	ter	abtree	
GBDTClassification_1	0.882200	\checkmark	gam	ma	0	
A ¥				shild weight	4	
Unused models	auc	Select	lill mu	child_weight	1	
RFClassification_1	0.846454		max_	_delta_step	0	
LogicClassification_1	0.865166		subs	ample	1	
* CNNClassification 1	0.786437		colsa	ample_bytree	1	
PCAClassification 1	0.851927		colsa	ample_bylevel	1	
			reg_a	alpha	0	
	Unselected model		reg_l	lambda	1	
TreeClassification			scale	e_pos_weight	1	
The model marked with * i	s a supplementary model t	hat can not be cor	figured	Copy selected mo	del to model options	Close

K Model presentation						×		
Ensemble performance	296270797.147141	6270797.147141 Parameter name			Parameter value			
Model name	mse	Select	•	loss	ls			
GBDTRegression 1	382347158 307895		d.	learning_rate	0.1			
	445564549.012427			n_estimators	100			
VORDagraasian 1	260020540.507562		ł.	subsample	1.0			
XGBRegression_1	309838540.587502	V		criterion	friedman_mse			
Unused models	mse	Select	1	min_samples_split	50			
ENRegression_1	445674473.7744		ı	min_samples_leaf	50			
PCARegression_1	596129767.9753		0	min_weight_fraction_leaf	0			
* CNNRegression_1	17585172054.03		1	max_depth	6			
RidgeRegression_1	551359413.2697		1	min_impurity_decrease	1e-08			
RFRegression_1	812798684.5407		1	max_features	null			
* FNNRegression_1	659157565.9669		1	alpha	0.9			
A V	_			max_leaf_nodes	null			
l	Unselected model			warm_start	false			
TreeRegression				presort				
LRegression								
The model marked with * is a supplementary model that can not be configured Copy selected model to model options Qlose								

Titanic模型最终使用的分类模型及参数

房价模型最终使用的回归模型及参数

● 3. 变量重要度



建模之后,可以得到本次建模时各变量的重要度信息。从titanic模型返回的重要度可以看到, 性别(女士优先)和年龄范围(儿童优先)对于幸存最为重要。

Target variable Survived Set 🍸 Variable f							
NO.	Variable name	Туре	Date format	Select	Importance		
1		Binary variable		V	1		
2	AgeArea	Categorical variable		\checkmark	0.726		
3	Pclass	Categorical variable		\checkmark	0.524		
4	SibSp	Categorical variable		\checkmark	0.443		
5	Age	Numerical variable		\checkmark	0.392		
6	Fare	Numerical variable		\checkmark	0.275		
7	Parch	Categorical variable		\checkmark	0.244		
8	Family	Numerical variable		\checkmark	0.197		
9	Cabin	Categorical variable		\checkmark	0.169		
10	Embarked	Categorical variable		\checkmark	0.146		
11	Passengerld	ID		\checkmark	0		
12	Survived	Binary variable		\checkmark	•		
13	Name	ID			0		
14	Ticket	Categorical variable		\checkmark	0		

	变量重要度的作用
1	参考变量重要度,有针对性的对数据重新处理。
2	使用重要度高的变量进行交互生成衍生变量,如路 程/时间=速度,速度*时间=路程等重新建模。
3	参考变量重要度,有针对性的对客户进行建议。

目录 CONTENTS

1. 批量预测
 2. 单条预测



● 1. 批量预测

创建模型以后,可以使用测试数据 进行预测。

对于二分类模型, 第一列是目标变 量为正样本的概率。

以titanic为例, 预测624号乘客幸 存的概率为21.584%。

	Batch scoring	Scoring					
_	Scoring data	C:\Users\w	vunan\OneDrive\c	lata\titanic_t	est.csv		
	Survived_1_p	ercentage	Passengerld	Survived	Pclass	Name	Sex
	21.584	1%	624	0	3	Hansen, Mr. Henry Damsgaard	male
	13.652	!%	625	0	3	"Bowen, Mr. David John ""Dai"""	male
	21.625	i%	626	0	1	Sutton, Mr. Frederick	male
	9.799	%	627	0	2	Kirkland, Rev. Charles Leonard	male
	95.103	%	628	1	1	Longley, Miss. Gretchen Fiske	female
	12.653	%	629	0	3	Bostandyeff, Mr. Guentcho	male
	6.248	%	630	0	3	O'Connell, Mr. Patrick D	male
	47.066	i%	631	1	1	Barkworth, Mr. Algernon Henry Wilson	male
	3.796	%	632	0	3	Lundahl, Mr. Johan Svensson	male
	63.63	%	633	1	1	Stahelin-Maeglin, Dr. Max	male
	7.895	%	634	0	1	Parr, Mr. William Henry Marsh	male
	17.128	1%	635	0	3	Skoog, Miss. Mabel	female
	87.152	!%	636	1	2	Davis, Miss. Mary	female
	30.58	%	637	0	3	Leinonen, Mr. Antti Gustaf	male
	26.677	'%	638	0	2	Collyer, Mr. Harvey	male
	33.02	%	639	0	3	Panula, Mrs. Juha (Maria Emilia Ojala)	female
	9.154	%	640	0	3	Thorneycroft, Mr. Percival	male
	23.667	'%	641	0	3	Jensen, Mr. Hans Peder	male
	97.429	1%	642	1	1	Sagesser, Mile. Emma	female
	50.589	1%	643	0	3	Skoog, Miss. Margit Elizabeth	female
	24.772	!%	644	1	3	Foo, Mr. Choong	male
	87.833	1%	645	1	3	Baclini, Miss. Eugenie	female

● 1. 批量预测

R

对于回归模型, 第一列是对目 标变量的预测值。

以房价预测为例,预测1461号 房屋的价格为129298.66。

Batch scoring	Scoring								
Scoring data	C:\Users\w	unan\On	eDrive\data\hou	se_prices_te	st.csv				
SalePrice pre	dictvalue	Id	MSSubClass	MSZonina	LotFrontage	LotArea	Street	Alley	LotShape
129298	66	1461	20	RH	80	11622	Pave		Reg
160807.	103	1462	20	RL	81	14267	Pave		IR1
191135.4	414	1463	60	RL	74	13830	Pave		IR1
198392.	522	1464	60	RL	78	9978	Pave		IR1
189149.3	272	1465	120	RL	43	5005	Pave		IR1
168192.3	263	1466	60	RL	75	10000	Pave		IR1
185509.	826	1467	20	RL		7980	Pave		IR1
158992.3	343	1468	60	RL	63	8402	Pave		IR1
200264.	592	1469	20	RL	85	10176	Pave		Reg
120879.	556	1470	20	RL	70	8400	Pave		Reg
200707.	594	1471	120	RH	26	5858	Pave		IR1
98007.4	84	1472	160	RM	21	1680	Pave		Reg
97726.5	94	1473	160	RM	21	1680	Pave		Reg
141444.4	443	1474	160	RL	24	2280	Pave		Reg
103150.	052	1475	120	RL	24	2280	Pave		Reg
354069.3	211	1476	60	RL	102	12858	Pave		IR1
255800.	709	1477	20	RL	94	12883	Pave		IR1
282393.	717	1478	20	RL	90	11520	Pave		Reg
313624.	796	1479	20	RL	79	14122	Pave		IR1
490093.3	299	1480	20	RL	110	14300	Pave		Reg
331277.3	321	1481	60	RL	105	13650	Pave		Reg





目标变量是分类变量时,预测后显示每个目标分类值的概率(总和为1)。例如第一条记录, 目标值为2的概率最高,为97.402%。

Batch scoring Scoring						
Scoring data C:\Program Fi	les\yimming\yimming\data\Fore	est_Covertype.mtx				
Cover_Type_1_percentage	Cover_Type_2_percentage	Cover_Type_3_percentage	Cover_Type_4_percentage	Cover_Type_5_percentage	Cover_Type_6_percentage	Cover_Type_7_percentage
0.448%	97.402%	0.169%	0.021%	1.745%	0.177%	0.038%
0.297%	98.152%	0.115%	0.015%	1.223%	0.172%	0.027%
1.875%	97.405%	0.594%	0.01%	0.088%	0.011%	0.017%
3.302%	94.912%	1.172%	0.014%	0.146%	0.429%	0.025%
0.319%	97.864%	0.091%	0.014%	1.546%	0.137%	0.027%
0.768%	96.389%	0.337%	0.034%	2.059%	0.359%	0.054%
0.699%	95.365%	0.171%	0.021%	3.529%	0.176%	0.039%
0.37%	96.957%	0.095%	0.015%	2.385%	0.148%	0.029%
0.511%	97.973%	0.107%	0.014%	1.211%	0.163%	0.021%
0.673%	98.115%	0.073%	0.013%	0.999%	0.103%	0.024%
0.421%	98.58%	0.137%	0.011%	0.708%	0.124%	0.019%
3.994%	95.644%	0.044%	0.022%	0.222%	0.026%	0.047%
2.927%	96.683%	0.178%	0.01%	0.155%	0.028%	0.019%
0.229%	98.33%	0.182%	0.011%	1.119%	0.11%	0.018%
0.318%	98.225%	0.133%	0.024%	0.802%	0.448%	0.05%
0.704%	94.935%	0.224%	0.041%	2.922%	1.099%	0.074%
1.336%	96.347%	0.178%	0.033%	1.74%	0.315%	0.052%
0.383%	96.798%	0.146%	0.027%	2.265%	0.329%	0.053%
0.234%	94.256%	0.108%	0.016%	5.058%	0.297%	0.03%
0.252%	96.695%	0.12%	0.019%	2.536%	0.335%	0.043%
0.681%	97.92%	0.139%	0.029%	0.718%	0.452%	0.06%
6.872%	92.693%	0.026%	0.018%	0.334%	0.021%	0.035%

● 1. 批量预测

通常预测数据中是不包含目标变量的。 当预测数据中包含目标变量时,可以 根据预测结果计算模型表现,用来评 估模型。



● 2. 单条预测

单条预测通过拖拽方式修改变量值, 即时查看预测结果。

变量是按重要度降序排列的,通常靠前的变量对于预测结果的影响更大。可以看到年龄较小的女性幸存率很高。



对于房价预测模型,可以看 到当房屋地下室面积从334 拖拽到5642时(其他变量 没有改变),房价有了大幅 提升。

Batch scoring Scoring									
Minimum value Sa			SalePrice_predictvalue			Maximum value			
59293.715			65927.772		226828.545				
NO.	Variable name	Importance			Edit			Value	
1	GrLivArea	1	о 334	ı 1190	15	54	ı 1954	1 334 5642	

Bat	Batch scoring Scoring									
Minimum value S			SalePrice_predictvalue			Maximum value				
59293.715			226828.545		226828.545					
NO.	Variable name	Importance			Ed	it			Value	
1	GrLivArea	1	л 334	ı 1190	ا 155	54	ı 1954	0 1 5642	5642	



目录 CONTENTS

1. 集算器外部库
 2. 集成框架

● 1.集算器外部库

集算器外部库提供了智能建模 的接口函数,可以通过SPL调用。 建模的SPL:

	Α	В
1	=file("titanic_train.csv").cursor@cqt()	/创建训练数据游标
2	=ym_env()	/初始化环境
3	=ym_model(A2,A1)	/加载数据
4	=ym_target(A3, "Survived")	/设置目标变量
5	=ym_build_model(A3)	/执行建模
6	=ym_save_pcf(A5,"titanic.pcf")	/保存模型文件
7	=ym_json(A5)	/导出模型信息为json串
8	=ym_importance(A5)	/获取变量重要度
9	=ym_present(A5)	/获取模型描述
10	=ym_performance(A5)	/获取模型表现
11	>ym_close(A2)	/关闭

A7

值 {"Importance":{"PassengerId":0,"Pcl ass":0,"Sex":0,""Age":0.433191...

A8	
Name	Importance
Passengerld	0.0
Pclass	0.0

A9		
name	value	properties
XGBClass	0.815	[[max_delt
XGBClass	0.777	[[max_delt

详细信息可以查看: <u>http://c.raqsoft.com.cn/article/1568163387677</u>

● 1. 集算器外部库

模型创建以后(也可以使用智能建模设计器创建的模型), 可以通过SPL调用智能建模外部 库进行预测。预测的SPL:

	Α	В
1	=ym_env()	/初始化环境
2	=ym_load_pcf("titanic.pcf")	/加载模型文件
3	=file("titanic_test.csv").import@cqt()	/加载预测数据
4	=ym_predict(A2,A3)	/执行预测,返回预测结果对象
5	=ym_result(A4)	/获取预测结果序表
6	=ym_json(A4)	/预测数据不少于20条批量预测时, 会根据预测数据评估导出模型表现 json信息。
7	>ym_close(A1)	/关闭

A5

Passengerld	Survived	Pclass	Name	Sex	
624	0	3	Hansen,	male	
625	0	3	Bowen,	male	
626	0	1	Sutton,	male	
627	0	2	Kirkland	male	

A6

3

值
"Model- Performance":"{\"GINI\":0.8369670542635659,\"AUC\": 0.9184835271317829,\"KS\":0.6867732558139534,\"R DC-Data\":[\"{\\\"1- specificity\\\":\\\\"0.0\\\",\\\"sensitivity\\\":\\\\"0.020833333 33333332\\\"}\",\"{\\\"1

创建模型有两种方式:

- 1. 使用智能建模设计器创建模型文件。
- 2. 通过SPL调用集算器外部库建模。

THANKS

创新技术 推动应用进步

www.raqsoft.com.cn