集算器

高性能计算专家

SPL实现自动建模和预测

润乾软件出品

主要内容

1、建模序言

随着互联网经济的蓬勃发展，商业决策对基于大数据依赖越来越强烈。正确而连贯的数据流对商业用户做出快速、灵活的决策起到决定性的作用，数据建模迫切需求正顺应时代潮流而生，紧随AI时代步伐而起。

秉承建模应用“智能、高效、易用”的全新设计理念，易明建模打通了“从数据到模型，从预测到场景化应用”的全新操作流程。凭借大数据处理能力与独特算法引擎，构建智能、易用的人工智能分析与应用平台，有助于提升公司建模效率、降低建模成本。

千里之行始如足下，为开拓在大数据领域的新天地，让我们从SPL自动建模开始。

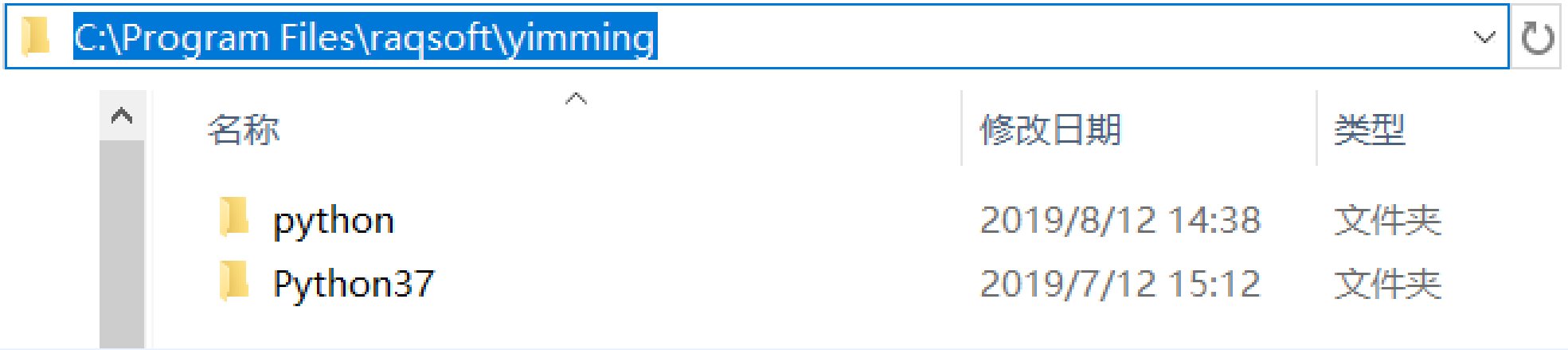
2、环境设置

SPL建模由易明智能建模软件、集算器SPL外部库Yimming两部分组成，通过配置文件userconfig.xml关联起来。

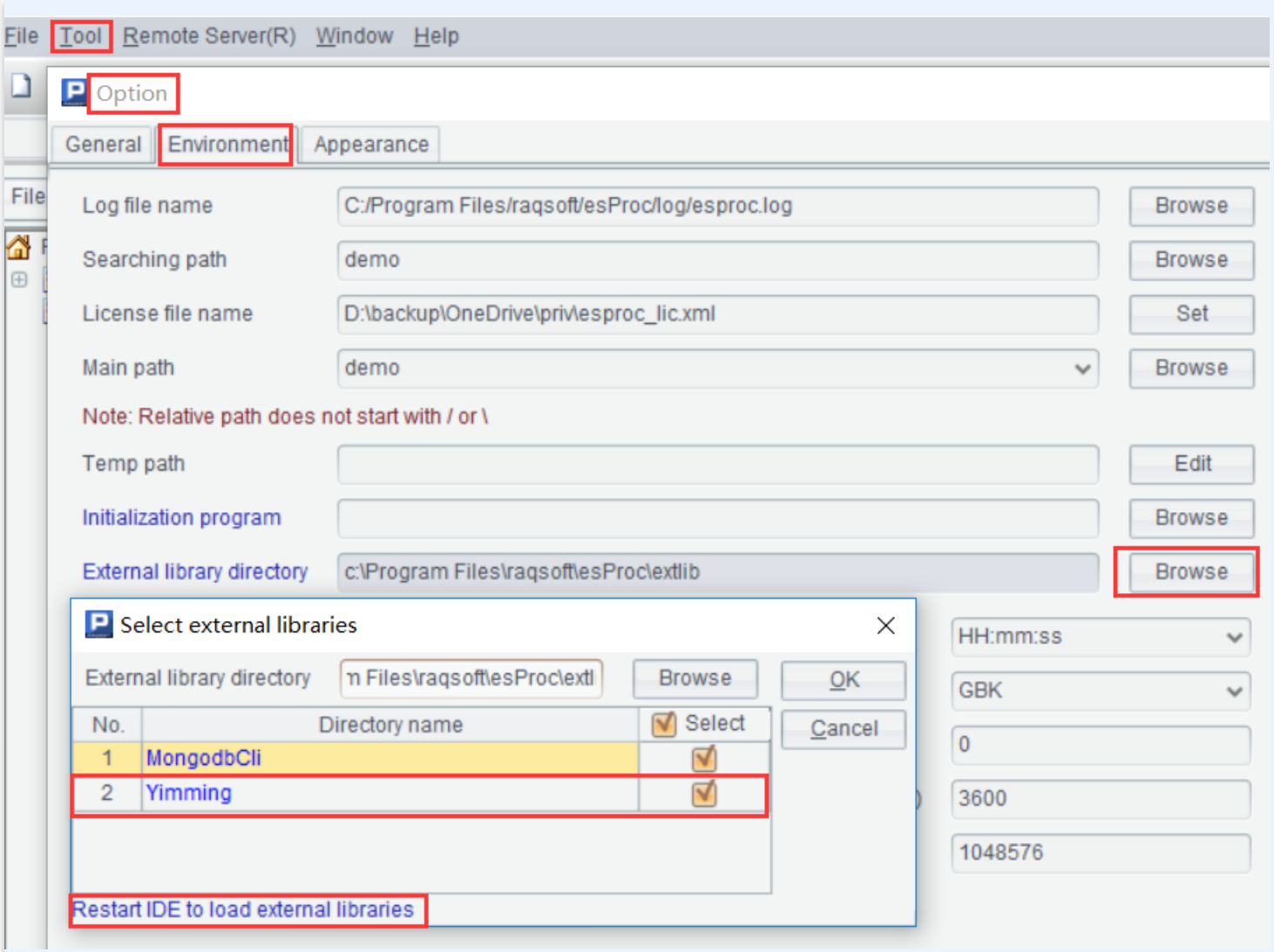
A、建模软件安装：

下载地址： http://download.raqsoft.com.cn/yimming/yimming-V2018-install.zip

安装易明智能建模软件，记下安装目录，如C:\Program Files\raqsoft\yimming。



2、环境设置



B、外部库安装： 缺省安装在集算器SPL软件的esProc\extlib\Yimming路径下，在集算器的外部库设置中勾选Yimming项让其生效。

2、环境设置

C、**配置文件**：SPL建模软件要有效地运作起来，需要在外部库目录esProc\extlib\Yimming下的userconfig.xml文件中设置参数，主要参数如下：

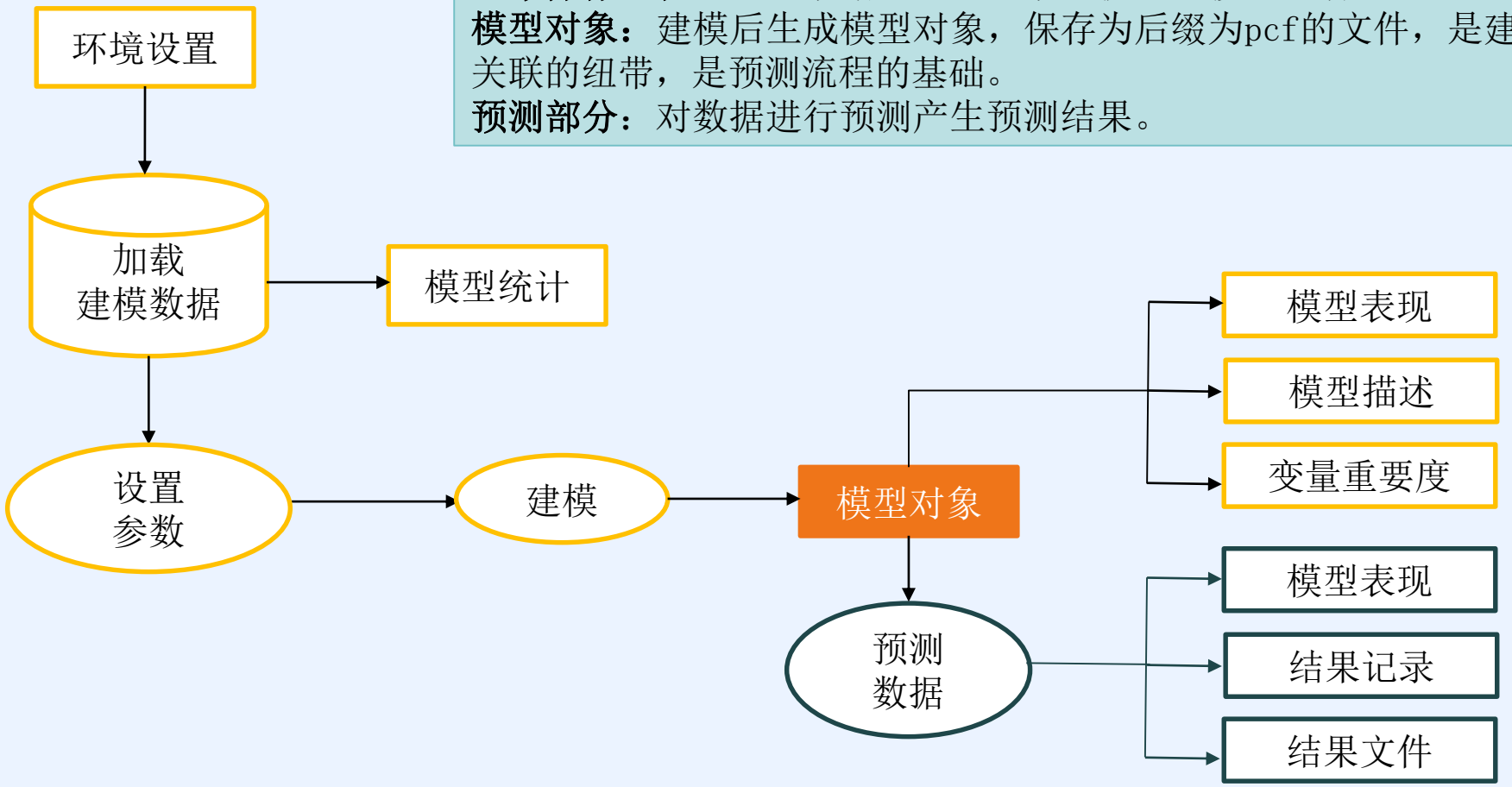
选项	名称	说明
sAppHome	C:\Program Files\raqsoft\yimming	应用程序目录
sLicenseFile	D:\backup\OneDrive\priv\yimming_lic.xml	智能建模授权
sEsprocLicenseFile	D:\backup\OneDrive\priv\esproc_lic.xml	集算器授权
sPythonHome	c:\Program Files\raqsoft\yimming\Python37\python.exe(for windows)	Python文件
	/raqsoft/yimming/Python37/bin/python3.7(for linux)	
bAutoDecideImpute	true	智能补缺
bLogConsole	true	日志在终端显示
iResampleMultiple	150	重采样次数

其中sAppHome为建模软件安装目录。

3、建模部分

3.1、建模流程图

基本操作过程：加载数据->设置参数->执行建模->查看结果
模型对象：建模后生成模型对象，保存为后缀为pcf的文件，是建模与预测关联的纽带，是预测流程的基础。
预测部分：对数据进行预测产生预测结果。



3.2、建模SPL样例

	A	注释
1	=file("train.csv").cursor@tq(;,",")	建模数据
2	="passenger.pcf"	要生成的模型文件
3	=ym_env()	初始化环境
4	=ym_model(A3, A1)	加载数据
5	=ym_target(A4, "Survived")	设置目标变量
6	=ym_setparam(A4, "intelligence":true, "Balance":2)	设置建模参数
7	=ym_statistics(A4,"Age")	获取变量统计信息
8	=ym_build_model(A4,A2)	执行建模(过程)
9		
10	=ym_importance(A8)	获取变量重要度信息
11	=ym_present(A8)	获取模型描述
12	=ym_performance(A8)	获取模型表现
13	=ym_result(A8, "train_t.csv")	生成预测结果
14	>ym_close(A3)	关闭

建模主要过程为黄色标记部分，A13预测也可在建模后执行，其它的主要是查看建模相关信息。

3.3、建模数据

建模所用的数据可来自关系型数据库、文本文件、序表、游标、csv、mcf文件等，要求是结构化的数据。如CSV文件数据：

```
PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,113803,53.1,C123,S
5,0,3,"Allen, Mr. William Henry",male,35,0,0,373450,8.05,,S
6,0,3,"Moran, Mr. James",male,,0,0,330877,8.4583,,Q
7,0,1,"McCarthy, Mr. Timothy J",male,54,0,0,17463,51.8625,E46,S
```

文件首行是字段信息数据，其它行是数据记录。
mcf数据文件为建模预处理数据后的数据文件，是二进制格式，加载数据时比较快。
数据接入支持多类型数据源接入，实现各类数据统一接入与管理，为建模应用奠定坚实的数据基础。
建模数据预处理时，可进行缺失值处理、高基数变量处理、数据平滑处理、数值变量筛选、添加衍生变量、DOC变量等内部数据清洗。

3.4、变量统计信息

Name	Value
VarName	Age
Miss	0.2102728731942215
Imp	0.0
Card	0
GraphData	
GroupDescStatisticsTable	
GroupFrequencyTable	
Upquar	38.0
Median	28.0
Lwquar	21.0
Sd	14.378831499148678
Max	71.0
Min	0.75
Avg	29.78048780487805
Sk	0.3387264693285246
OuterValues	[64.0,64.0,65.0,65.0,65.0,66.0,70.5,71.0,71.0]
Pearson	NaN
Spearman	NaN
Target0	0
Target1	1
bGraphStatistics	true
bStatistics	true
bTargetStatistics	true

获取指定变量的具体信息，主要返回最大值，最小值，重要度，缺失率，偏度等信息，有助于进行数据探索与分析。如查看Age变量，返回信息如左所示：

3.5、建模参数设置

设置建模变量参数，相关参数说明：（更详细规则可参考易明智能建模手册）

Key	Value Type	说明
balance	int	配平参数
Target	String	目标参数
id	String	ID变量名
intelligence	boolean	是否智能填补
misformat	String	缺失值格式
optimal	boolean	是否使用最优参数配置
parallel	int	预处理并行数
resample	boolean	是否简单模型
resamplemul	int	抽样倍数
resamplenum	int	重抽样次数
testpercent	int	测试数据百分比0-99
vartypes	ArrayList< Byte>	变量类型
ModelFields	ArrayList<String>	建模的字段名顺序

3.6、模型信息

A。模型描述

智能建模功能中包含多种算法，将返回当前模型使用了哪些算法及其相关模型参数等。

Index	name	value	properties
1	RidgeClassification_1	0.80441281989954 56	[[random_state,0],[alpha,0.5],[m...
2	LogicClassification_1	0.80381487682372 63	[[C,1.0],[random_state,0],[verbo...
3	RFClassification_1	0.78850753408275 53	[[min_samples_leaf,50],[n_esti...
4	FNNClassification_1	0.75442477876106 21	[[warm_start,false],[random_sta...
5	XGBClassification_1	0.83126046400382 68	[[max_delta_step,0],[base_scor...
6	GBDTClassification_1	0.81661085864625 68	[[min_samples_leaf,50],[learnin...
7	TreeClassification_1	0.79239416407558	[[min_samples_leaf,50],[splitter...

3.6、模型信息

B、模建表现

此功能展示该模型相关信息，可查看GINI、AUC、KS指数等

Index	Name	Value
1	GINI	0.6627601052379812
2	AUC	0.8313800526189906
3	KS	0.5908873475245157
4	AccuTable	[[0.05000000074505806,0.4919786096256685,0.4...
5	RocTable	[[0.0,0.01351351351351514],[0.0,0.02702702702...
6	LiftAndRecallTable	[[1,2.5270270270270268,0.1216216216216163, ...

3.6、模型信息

C、变量重要度

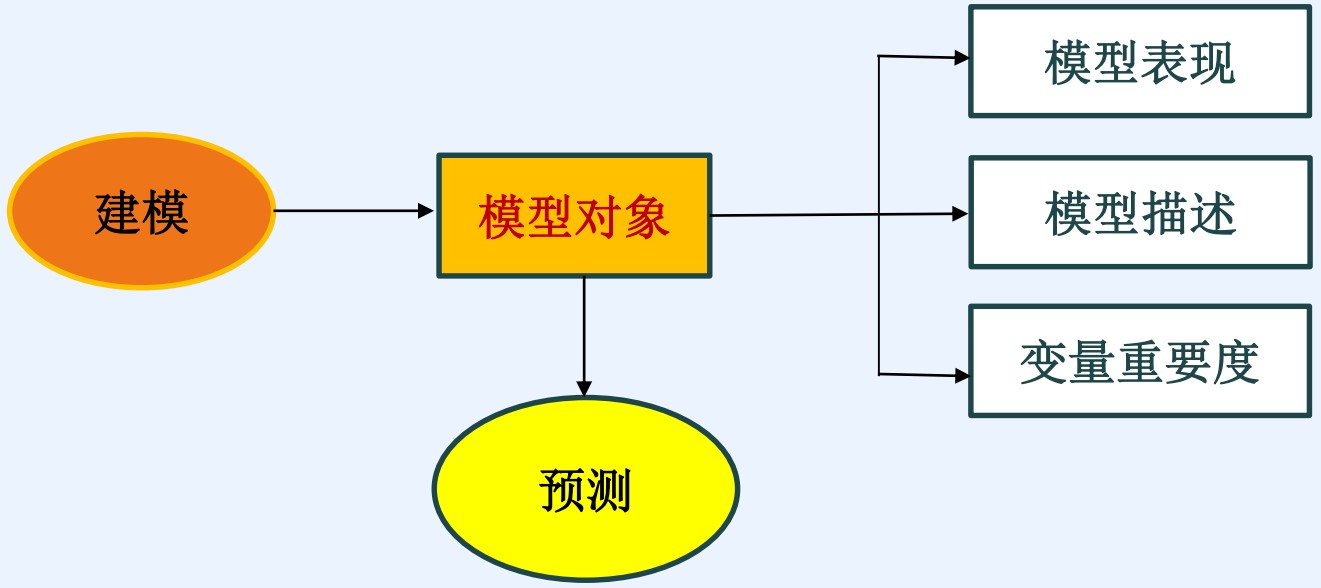
查看各个变量的重要度信息。

Index	Name	Importance
1	PassengerId	0.0
2	Pclass	0.3348135805855989
3	Sex	1.0
4	Age	0.19204237684722372
5	SibSp	0.14110517904914055
6	Parch	0.08141316846013069
7	Ticket	0.0
8	Fare	0.18767660989418544
9	Cabin	0.0
10	Embarked	0.08088429746924328
11	Survived	0.0

3.6、模型信息

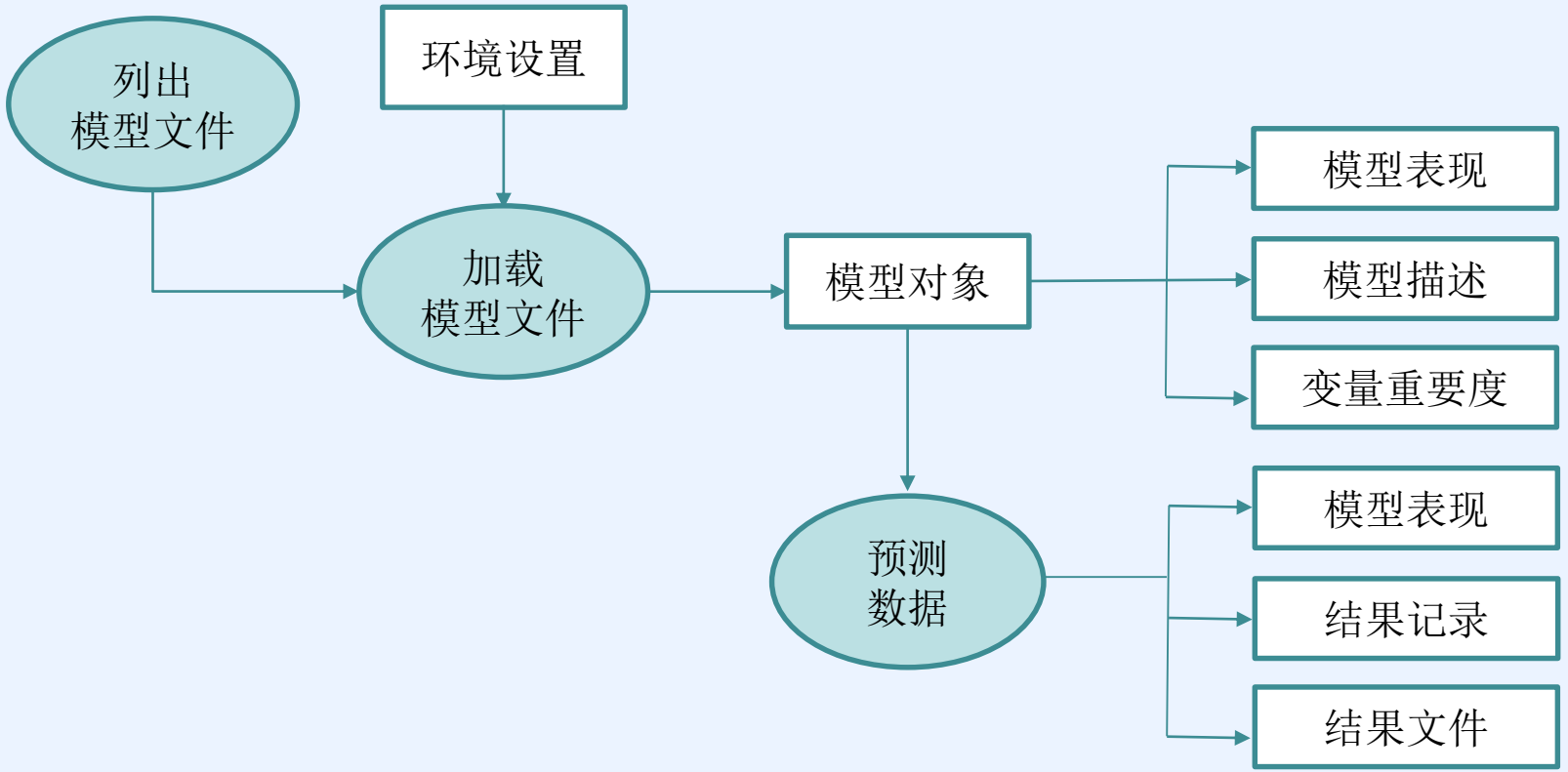
D、建模结果

建模后产生模型对象，以pcf建模文件形式保存，是建模与预测关联的纽带，为预测处理提供模型文件，预测流程也可以直接从建模文件开始。



4、预测部分

4.1、预测流程图



基本操作过程：加载模型文件->预测数据->查看结果。

4、预测部分

4.2、预测SPL样例

	A	注释
1	=ym_env()	初始化环境
2	=ym_list()	列出模型文件
3	=ym_predict(A2(1))	根据模型文件生成模型对象
5	=ym_result@s(A3, "D:/dev/train.csv", "D:/dev/train_res.txt")	生成预测结果，train.csv为预测的数据，train_res.txt为生成的结果文件
6	>ym_close(A1)	关闭

4.3、模型对象

列出模型文件

通过ym_list()列出当前有哪些pcf模型文件，或列出指定目录下的模型文件。此项主要是为了方便用户预测时，知道有哪些可用的模型文件，是可选项。

Index	FileName
1	C:\Program Files\raqsoft\yimming\store\predict\emp_101.pcf
2	C:\Program Files\raqsoft\yimming\store\predict\train.pcf
3	C:\Program Files\raqsoft\yimming\store\predict\train101.pcf

ym_predict () 根据模型文件生成模型对象.
有了模型对象后，除了可预测外，用户也可以通过它获取相关的建模信息，如模型表现等。

4.4、预测结果

根据模型对有效数据进行预测后，生成预测结果。
要预测的数据与建模所用的数据类似，可来自数据库、序表、csv等。

A、预测结果

返回预测结果信息，同时也可保存为文本文件或csv文件。

Index	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket
1	624	0	3	Hansen, M...	male	21	0	0	350029
2	625	0	3	"Bowen, Mr...	male	21	0	0	54636
3	626	0	1	Sutton, Mr. ...	male	61	0	0	36963
4	627	0	2	Kirkland, R...	male	57	0	0	219533
5	628	1	1	Longley, Mi...	female	21	0	0	13502
6	629	0	3	Bostandyef...	male	26	0	0	349224
7	630	0	3	O'Connell, ...	male	(null)	0	0	334912
8	631	1	1	Barkworth, ...	male	80	0	0	27042
9	632	0	3	Lundahl, M...	male	51	0	0	347743
10	633	1	1	Stahelin-M...	male	32	0	0	13214
11	634	0	1	Parr, Mr. Wi...	male	(null)	0	0	112052
12	635	0	3	Skoog, Mis...	female	9	3	2	347088

4.4、预测结果

B、预测模型表现

当预测数据中包含目标变量时，可以根据预测结果查看模型表现。该功能是通过预测结果反推出模型表现，可以对比此处的模型表现与模型文件中的模型表现，评估模型的质量。

返回指数GINI、AUC、KS等信息

Index	Name	Value
1	AUC	0.830062984496124
2	GINI	0.6601259689922481
3	KS	0.5666182170542636
4	AccuTable	[[0.05000000074505806,0.5074626865671642,0.41818181...
5	RocTable	[[0.0,0.020833333333333332],[0.0,0.041666666666666664]...
6	LiftAndRecallTable	[[1,2.791666666666667,0.13541666666666666, ...],[2,2.362...

5、最后总结

我们通过SPL自动建模操作流程，可见建模在已有的专业理论知识与算法技术基础上，结合公司自身的业务需求，通过简单的操作也能方便使用。建模预测整体操作流程归结如下：