



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



本文主要介绍了建模软件运行所需要的环境设置、建模、预测三部分，以详实的样例来说明具体如何建模、建模数据要求、模型信息、建模结果等，同时用户也可以利用建模后的模型文件或模型对象进行预测处理、生成预测结果。了解建模流程、预测流程后，用户在集算器中可熟练使用建模与预测操作。

# 主要内容



## 1、建模序言

## 2、环境设置

## 3、建模部分

### 3.1、建模流程

### 3.2、建模SPL样例

### 3.3、建模数据

### 3.4、建模统计信息

### 3.5、参数设置

### 3.6、建模信息

### 3.7、建模结果

## 4、预测部分

### 4.1、预测流程

### 4.2、预测SPL样例

### 4.3、预测模型对象

### 4.4、预测结果

## 5、最后总结



# 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。



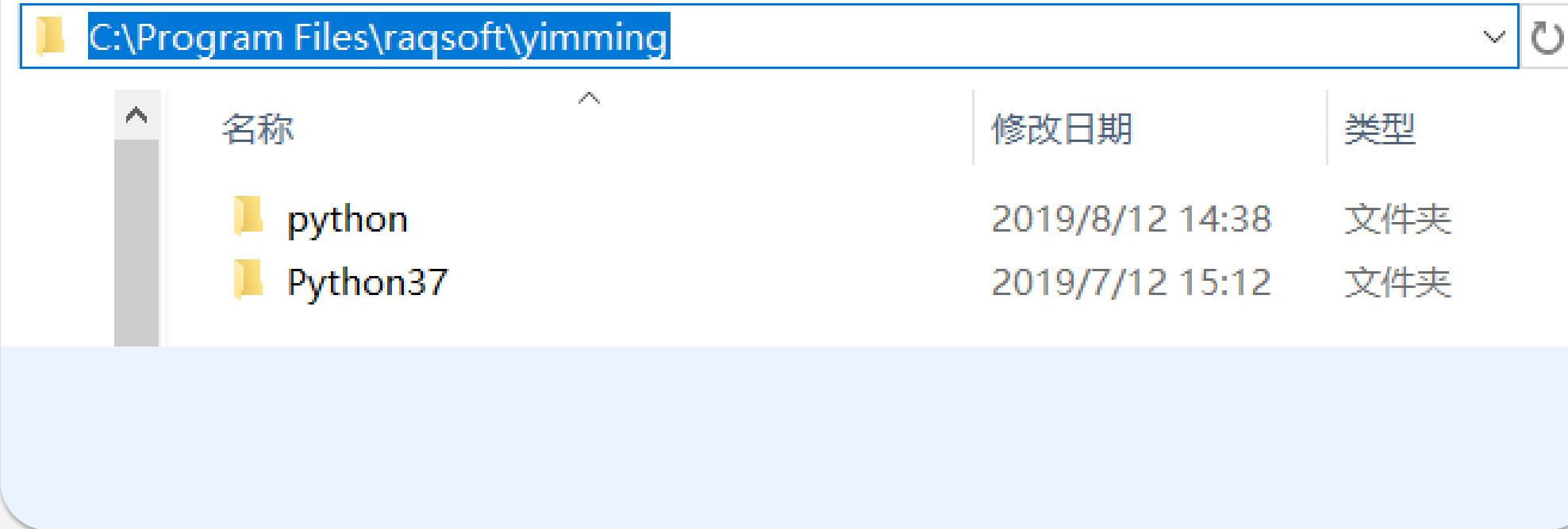


## 2、环境设置

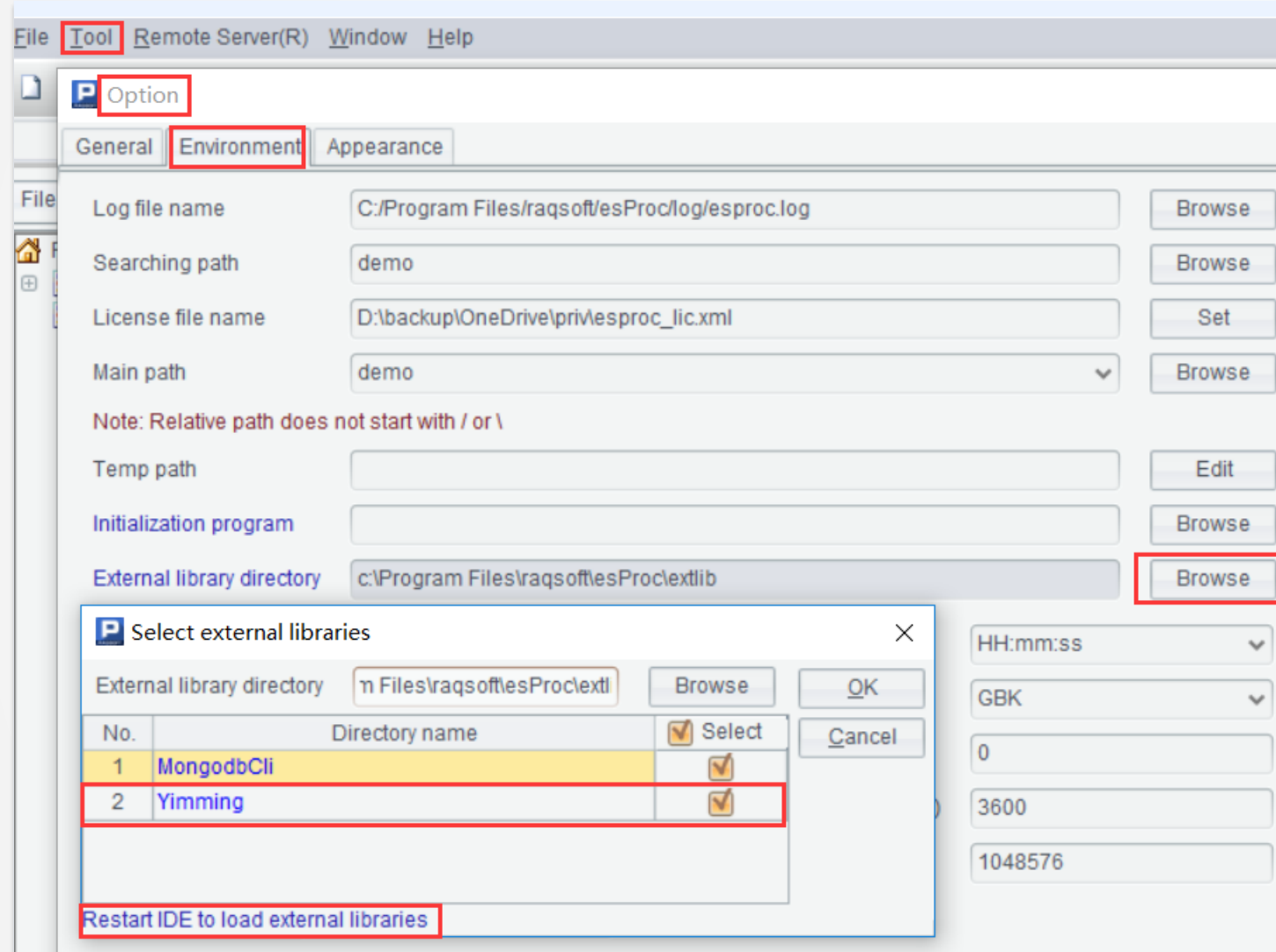
SPL建模由[易明智能建模软件](#)、集算器SPL外部库Yimming两部分组成，通过配置文件userconfig.xml关联起来。

### 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	Python文件 (for windows)
	/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

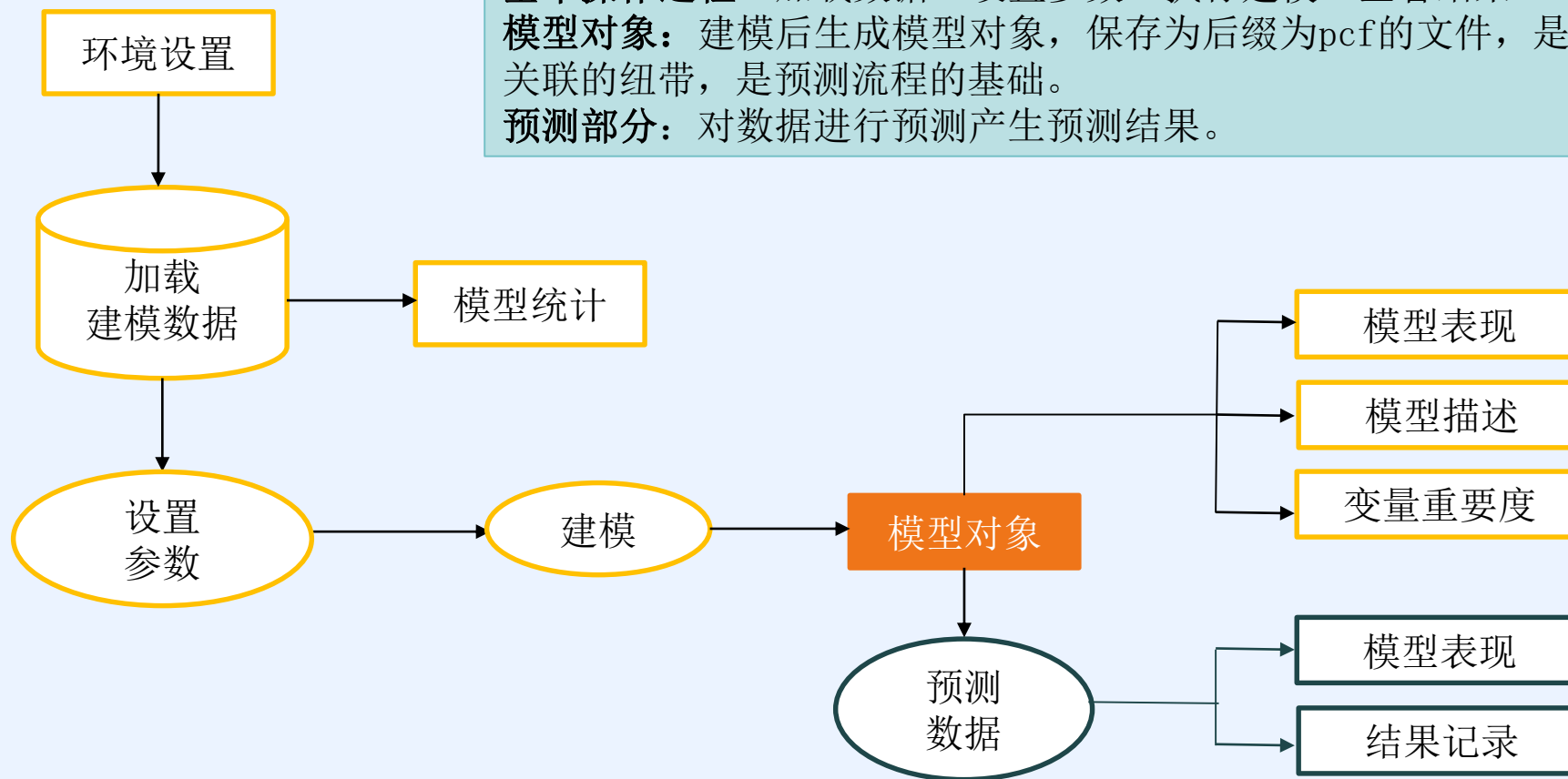
其中sAppHome为建模软件安装目录。



# 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	<code>=file("train.csv").cursor@tq(;;, ",")</code>	建模数据
2	<code>=ym_env()</code>	初始化环境
3	<code>=ym_model(A2, A1)</code>	加载数据
4	<code>=ym_target(A3, "Survived")</code>	设置目标变量
5	<code>=ym_setparam(A3, "intelligence":true, "Balance":2)</code>	设置建模参数
6	<code>=ym_statistics(A3, "Age")</code>	获取变量统计信息
7	<code>=ym_build_model(A3)</code>	执行建模(过程)
8	<code>=ym_save_pcf(A7, "demo.pcf")</code>	将模型对象保存为文件
9	<code>=ym_json(A7)</code>	导出模型信息为json串
10	<code>=ym_importance(A7)</code>	获取变量重要度信息
11	<code>=ym_present(A7)</code>	获取模型描述
12	<code>=ym_performance(A7)</code>	获取模型表现
13	<code>&gt;ym_close(A2)</code>	关闭

建模主要过程为黄色标记部分，A8将建模结果保存为pcf模型文件，在预测时可利用模型文件进行预测处理，A9-A12输出模型相关信息。



## 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等，要求是结构化的数据。如CSV文件数据：

```
1 PassengerId,Survived,Pclass,Name,Sex,Age,SibSp,Parch,Ticket,Fare,Cabin,Embarked
2 1,0,3,"Braund, Mr. Owen Harris",male,22,1,0,A/5 21171,7.25,,S
3 2,1,1,"Cumings, Mrs. John Bradley (Florence Briggs Thayer)",female,38,1,0,PC 17599,71.2833,C85,C
4 3,1,3,"Heikkinen, Miss. Laina",female,26,0,0,STON/O2. 3101282,7.925,,S
5 4,1,1,"Futrelle, Mrs. Jacques Heath (Lily May Peel)",female,35,1,0,113803,53.1,C123,S
6 5,0,3,"Allen, Mr. William Henry",male,35,0,0,373450,8.05,,S
7 6,0,3,"Moran, Mr. James",male,,0,0,330877,8.4583,,Q
8 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。

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

模型信息主要包括模型描述，模型表现及变量重要度信息，可通过接口`ym_json()` 导出成json字符串。

### 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	<a href="#">RidgeClassification_1</a>	<a href="#">0.8044128198995456</a>	[[random_state,0],[alpha,0.5],[m...
2	<a href="#">LogicClassification_1</a>	<a href="#">0.8038148768237263</a>	[[C,1.0],[random_state,0],[verbo...
3	<a href="#">RFCClassification_1</a>	<a href="#">0.7885075340827553</a>	[[min_samples_leaf,50],[n_esti...
4	<a href="#">FNNClassification_1</a>	<a href="#">0.7544247787610621</a>	[[warm_start,false],[random_sta...
5	<a href="#">XGBClassification_1</a>	<a href="#">0.8312604640038268</a>	[[max_delta_step,0],[base_scor...
6	<a href="#">GBDTClassification_1</a>	<a href="#">0.8166108586462568</a>	[[min_samples_leaf,50],[learnin...
7	<a href="#">TreeClassification_1</a>	<a href="#">0.79239416407558</a>	[[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	<a href="#">GINI</a>	0.6627601052379812
2	<a href="#">AUC</a>	0.8313800526189906
3	<a href="#">KS</a>	0.5908873475245157
4	<a href="#">AccuTable</a>	[[0.05000000074505806,0.4919786096256685,0.4...
5	<a href="#">RocTable</a>	[[0.0,0.013513513513513514],[0.0,0.02702702702...
6	<a href="#">LiftAndRecallTable</a>	[[1,2.5270270270270268,0.12162162162162163, .....



## 3.6、模型信息

### C、变量重要度

查看各个变量的重要度信息。

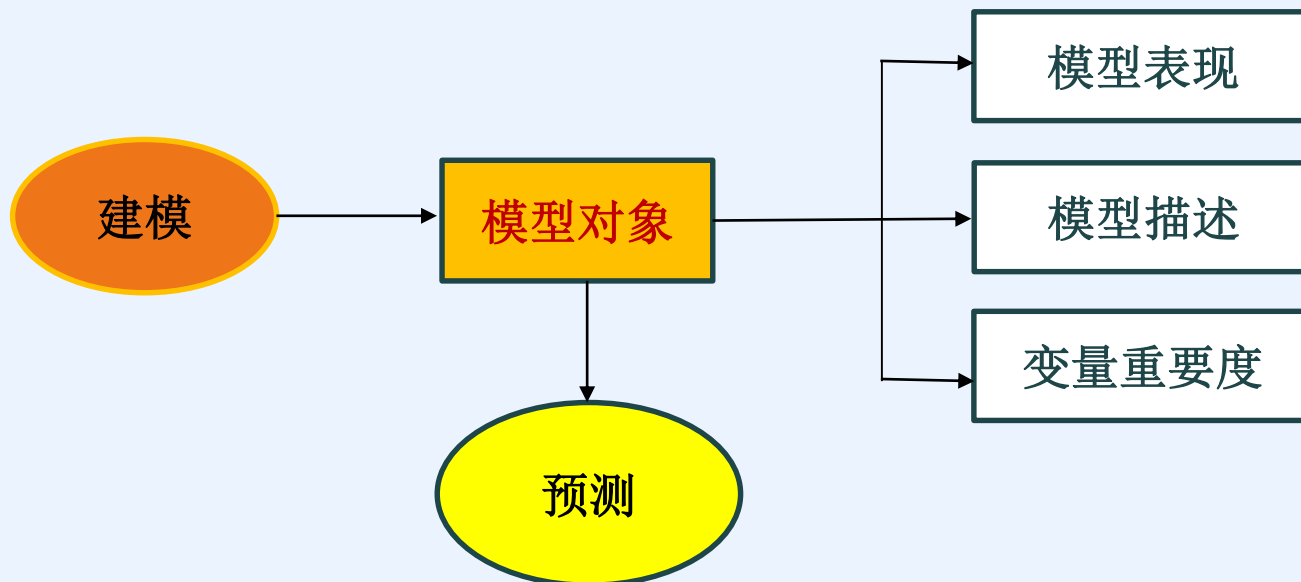
Index	Name	Importance
1	<a href="#">PassengerId</a>	0.0
2	<a href="#">Pclass</a>	0.3348135805855989
3	<a href="#">Sex</a>	1.0
4	<a href="#">Age</a>	0.19204237684722372
5	<a href="#">SibSp</a>	0.14110517904914055
6	<a href="#">Parch</a>	0.08141316846013069
7	<a href="#">Ticket</a>	0.0
8	<a href="#">Fare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



## 3.6、模型信息

### D、建模结果

建模后产生模型对象，以pcf模型文件形式保存，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模型文件开始。

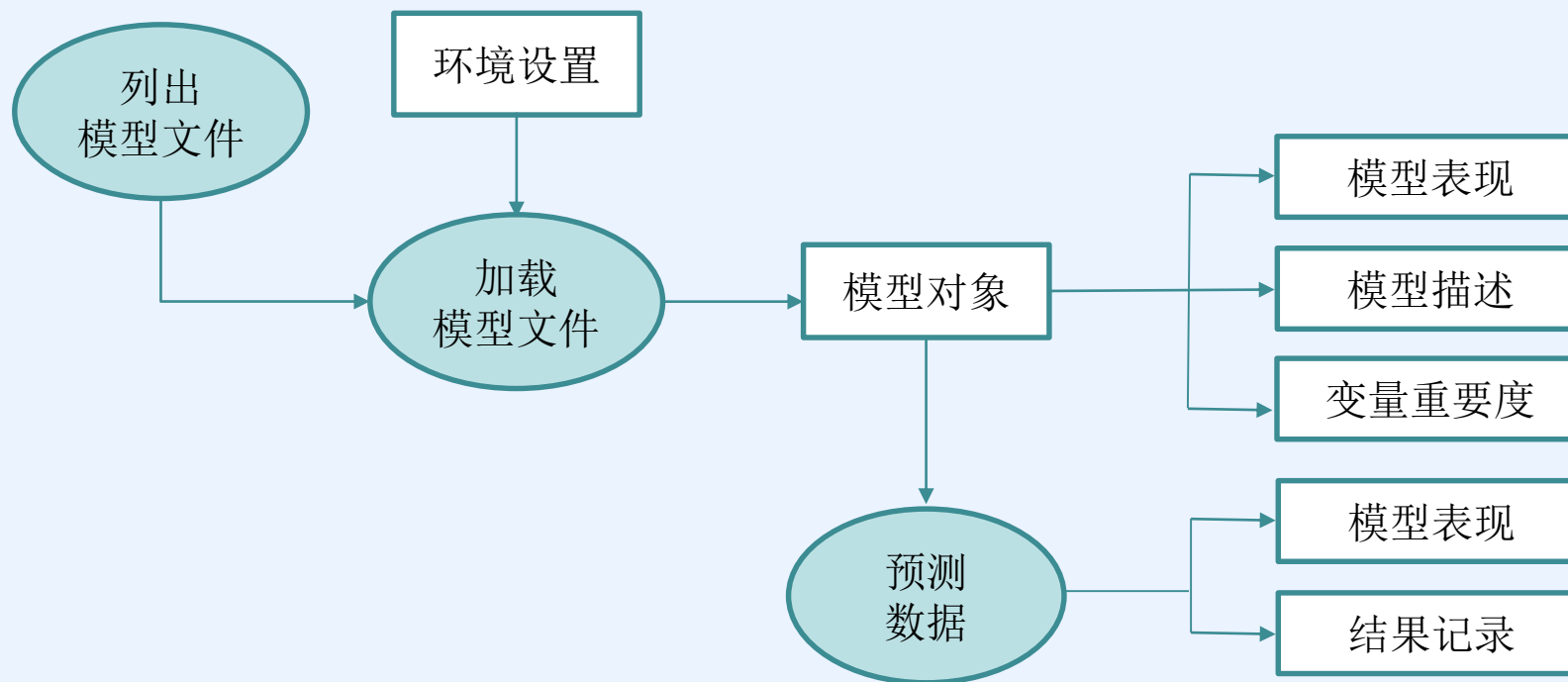






# 4、预测部分

## 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	<code>=ym_env()</code>	初始化环境
2	<code>=ym_load_pcf("demo.pcf")</code>	根据pcf模型文件生成模型对象
3	<code>=file("D:/dev/test_t.csv").import@tq(;",",")</code>	通过文件加载要预测的数据，返回序表。
4	<code>=ym_predict(A2, A3)</code>	执行预测处理，返回预测结果对象。
5	<code>=ym_result(A4)</code>	返回预测结果序表，
6	<code>=ym_json(A5)</code>	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	<code>&gt;ym_close(A1)</code>	关闭

A2加载由建模时生成的pcf模型文件, A4执行预测并返回预测结果对象。



## 4.3、预测模型对象

### 加载模型文件

接口 `ym_load_pcf()` 根据加载的模型文件生成预测模型对象。对于同一个模型来说，预测模型对象与建模后生成的对象功能上完全一致。

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，如模型描述，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。

Index	PassengerId	Survived	P...	Name	Sex	Age	...	...	Ticket	Fare	Ca...	Emb...	Survived_1_ratio
1	624	0	3	Hansen, M...	male	21	0	0	350029	7.8542	(null)	S	0.25707278881670026
2	625	0	3	"Bowen, Mr...	male	21	0	0	54636	16.1	(null)	S	0.5680230522833624
3	626	0	1	Sutton, Mr. ....	male	61	0	0	36963	32.3208	D50	S	0.23271421717823756
4	627	0	2	Kirkland, R...	male	57	0	0	219533	12.35	(null)	Q	0.5919246159939965
5	628	1	1	Longley, Mi...	female	21	0	0	13502	77.9583	D9	S	0.4284559885569835
6	629	0	3	Bostandyef...	male	26	0	0	349224	7.8958	(null)	S	0.298675652025568
7	630	0	3	O'Connell, ...	male	(null)	0	0	334912	7.7333	(null)	Q	0.06259809523413039
8	631	1	1	Barkworth, ...	male	80	0	0	27042	30.0	A23	S	0.20693993819873205
9	632	0	3	Lundahl, M...	male	51	0	0	347743	7.0542	(null)	S	0.04720448510306918
10	633	1	1	Stahelin-M...	male	32	0	0	13214	30.5	B50	C	0.5263867128075925
11	634	0	1	Parr, Mr. Wi...	male	(null)	0	0	112052	0.0	(null)	S	0.04965320978124586
12	635	0	3	Skoog, Mis...	female	9	3	2	347088	27.9	(null)	S	0.5855400248475947
13	636	1	2	Davis, Mis...	female	28	0	0	237668	13.0	(null)	S	0.38185936013549643
14	637	0	3	Leinonen, ...	male	32	0	0	STON/O 2. ...	7.925	(null)	S	0.1385752420653835
15	638	0	2	Collyer, Mr....	male	31	1	1	C.A. 31921	26.25	(null)	S	0.7216118101299388
16	639	0	3	Panula, Mr...	female	41	0	5	3101295	39.6875	(null)	S	0.338041641614491
17	640	0	3	Thorneycro...	male	(null)	1	0	376564	16.1	(null)	S	0.4347392114485213
18	641	0	3	Jensen, Mr...	male	20	0	0	350050	7.8542	(null)	S	0.25707278881670026



## 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，输出json串。

Index	Name	Value
1	<a href="#">AUC</a>	0.830062984496124
2	<a href="#">GINI</a>	0.6601259689922481
3	<a href="#">KS</a>	0.5666182170542636
4	<a href="#">AccuTable</a>	[[0.05000000074505806,0.5074626865671642,0.41818181...
5	<a href="#">RocTable</a>	[[0.0,0.020833333333333332],[0.0,0.04166666666666664]...
6	<a href="#">LiftAndRecallTable</a>	[[1,2.791666666666667,0.1354166666666666, ...],[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：

