

神算子平台帮助文档

常见问题	2
如何在神算子上搭建模型 ?	2
如何进行数据处理 ?	4
平台支持哪些算法 ?	5
神算子平台简介	6
快速开始	8
登录/注册	8
平台概览	9
创建您的第一个工作流	10
组件参数设定	11
运行工作流	12
查看运行结果	12
从模板创建工作流	13
项目管理	14
新建项目	14
新建面板	15
查看详情、修改和删除项目	15
修改和删除面板	16
数据源管理	16
我的数据源	16
公共数据源	19
工作流搭建和控制	20
工作流搭建	21
组件操作	22
组件设置	22
面板控制	23
模型保存与使用	23

数据处理	25
常见数据处理方法	25
数据处理方案比对	29
模型管理	32
服务发布	33
发布新服务	33
调用服务	34
服务的启用、停用与更新	35
算法组件介绍	36

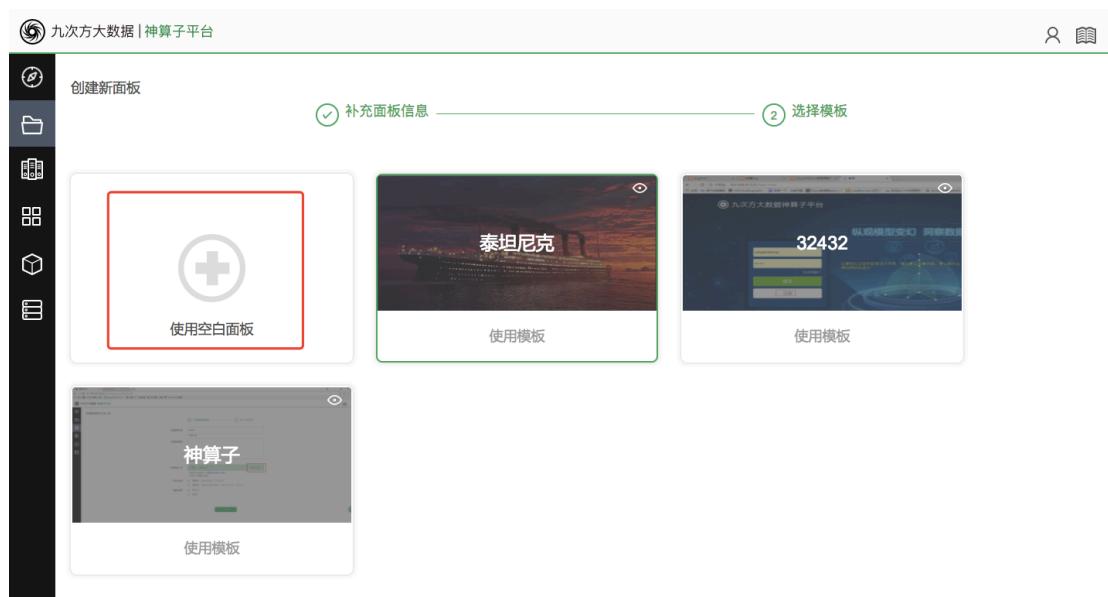
常见问题

如何在神算子上搭建模型？

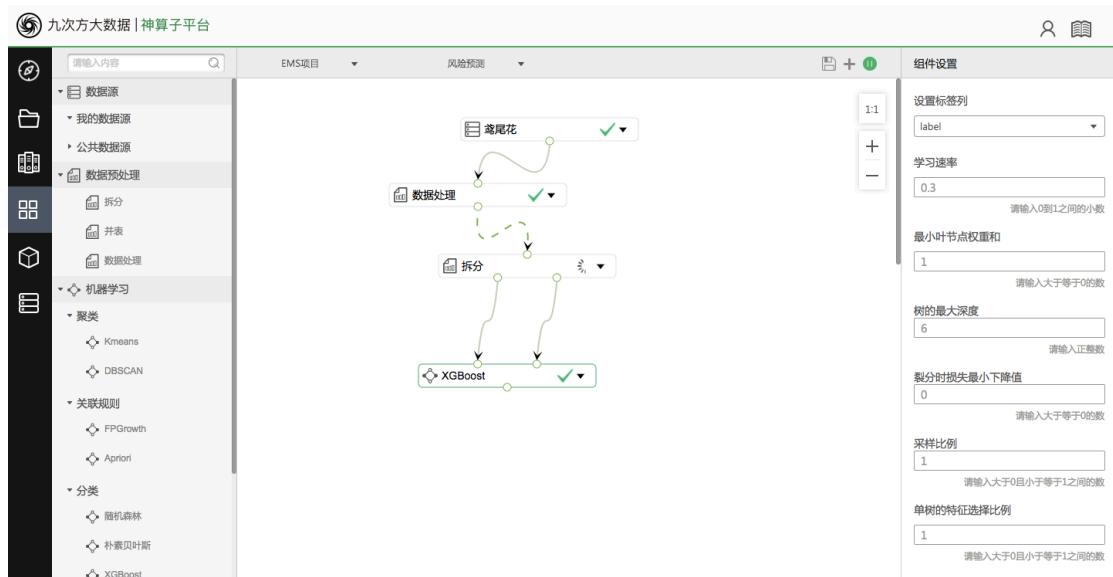
在神算子上搭建模型有两种方法：

1. 从空白面板开始创建模型

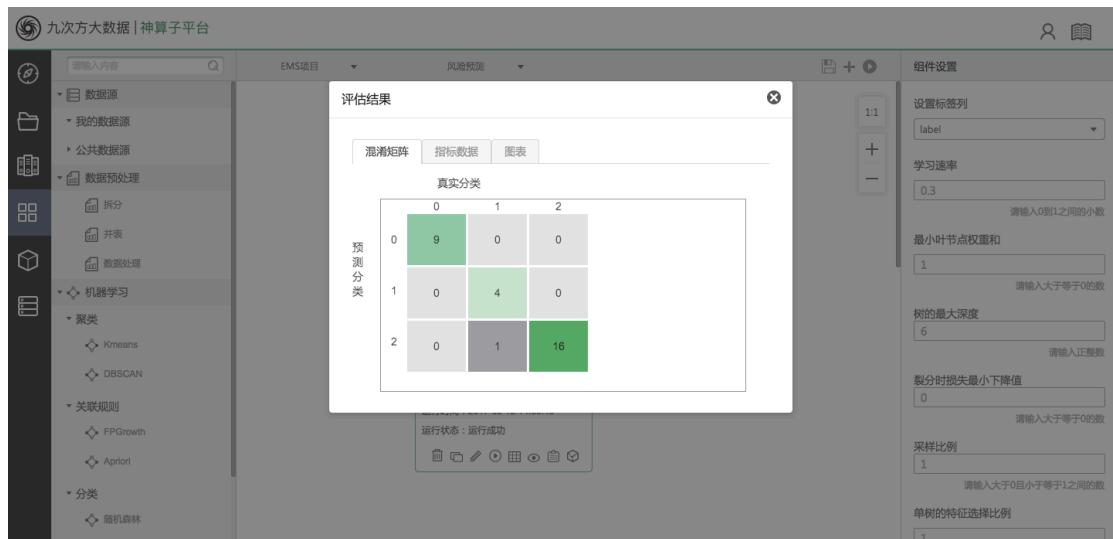
补充面板信息之后，选择“创建空白面板”，进入到一个新的空白面板中



将您需要的组件从左侧拖拽到面板中，根据需求连线，即可完成建模工作流的搭建。通常，一个典型的模型训练工作流由以下几个步骤组成：数据源——数据处理——拆分——算法

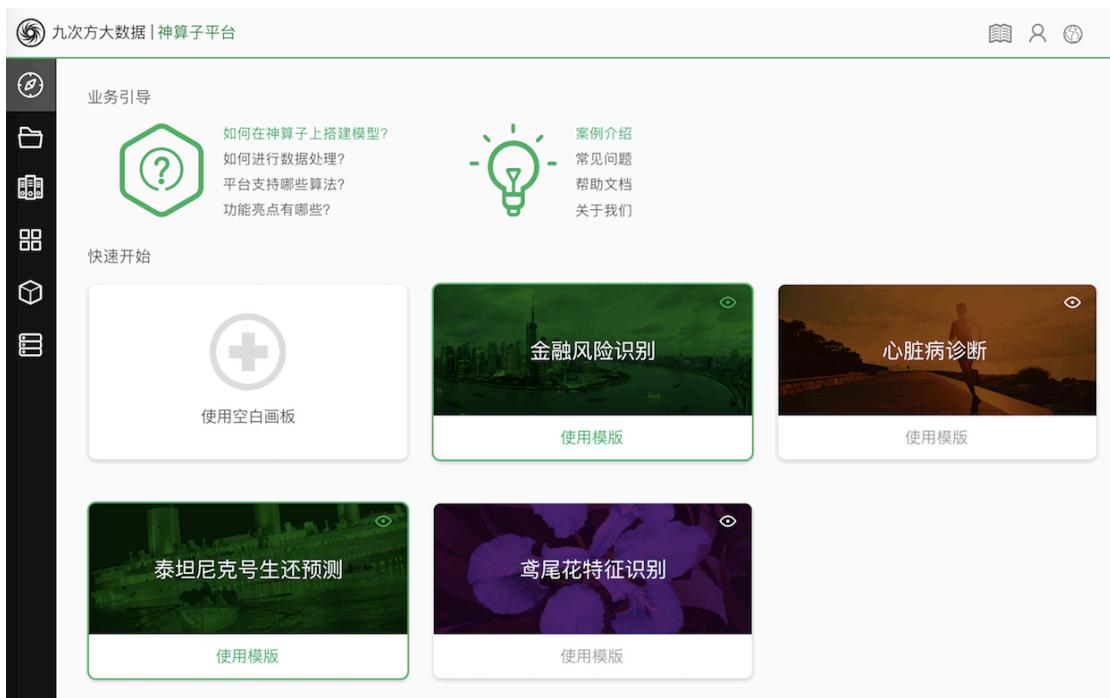


启动运行，完成以后可在算法组件详情中查看评估结果，根据评估结果调整算法或步骤，直至达到满意、稳定的结果



2. 从模板创建模型

如果您想要训练的模型与我们提供的模板吻合，也可以直接选择从模板创建，补充完面板信息后，选择对应的模板



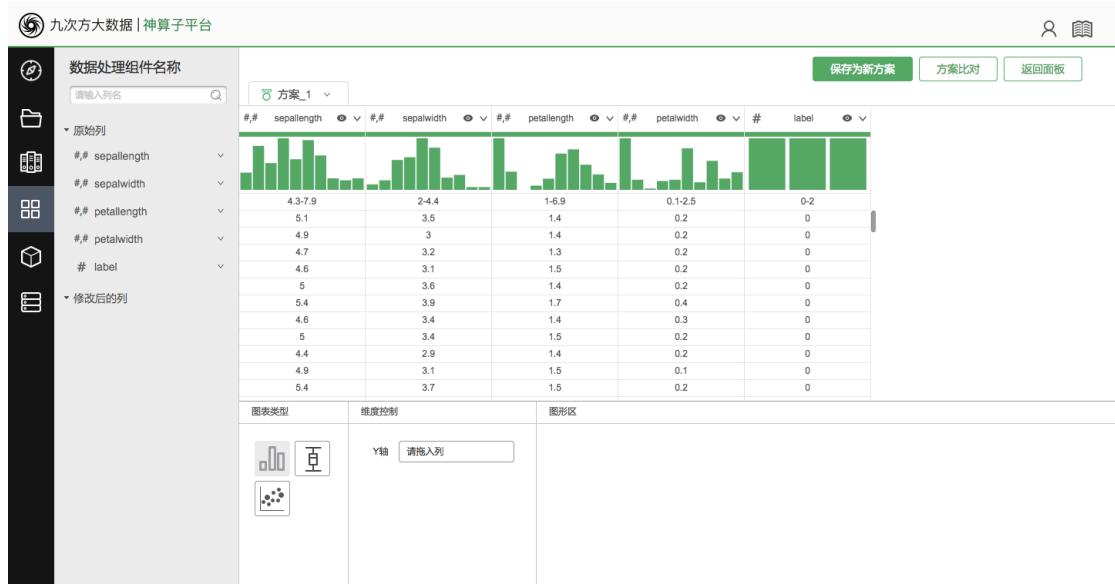
可以对模板提供的流程和组件进行修改，在模板基础上搭建需要的工作流

如何进行数据处理？

将数据源连接数据处理组件，可以进入数据处理模块

在数据处理模块中，平台提供类型转换、公式运算、缺失值处理、异常值处理、One-hot 编码、规则化、归一化等处理方式，按列处理

在处理过程中，可以随时查看特征列的详细汇总信息，例如缺失值信息、异常值信息，以及柱状图、箱线图、散点图等数据图表查看，辅助您做出有效决策



平台支持哪些算法？

平台的算法库正在不断扩充，目前已经支持的机器学习算法如下：

回归算法

1. 线性回归
2. 逻辑回归
3. XGboost 回归
4. 支持向量机回归

分类算法

1. 逻辑回归分类
2. 决策树分类
3. XGboost 分类
4. 支持向量机分类
5. 随机森林分类

- 6. 朴素贝叶斯
- 7. K 近邻
- 8. Adaboost 分类
- 9. 多层神经网络分类

聚类算法

- 1. K-means
- 2. DBSCAN

关联规则

- 1. FP-Growth
- 2. Apriori

时间序列算法

- 1. ARIMA
- 2. 一阶指数平滑
- 3. 二阶指数平滑
- 4. 三阶指数平滑

神算子平台简介

神算子平台是大数据科学院独立自主研发的一个大规模机器学习的全流程可视化建模平台。

平台提供了关于数据预处理，特征工程，模型构建，模型发布应用等一系列能力。

平台具有支持模型生命周期每个阶段的必要功能，专门用来管理和部署分析模型，平台使用项目对构建模型过程进行组织管理，不同的项目可对应于不同的业务用途或应用。在平台中用户可以通过有意义的业务过程数据，结合自己的业务目标进行人工智能模型调研、模型应用以及模型自学习的过程，自动化、智能化的帮助企业完成数据价值提升。

神算子平台是一个可针对非专业数据科学家的智能建模平台，相对于传统的人员专业要求高，数据处理慢以及建模周期长的建模方式，使用本平台不仅能够大大降低构建模型的门槛，还能够优化模型的效率。

作为机器学习平台，神算子具有以下特点：

低门槛——平台是国内首个成熟商用的人工智能全流程平台，用户只需通过简洁的可视化界面操作即可完成复杂的机器学习任务。同时神算子平台是一个针对非专业数据科学家的建模平台，相对于传统的人员专业要求高，数据处理慢以及建模周期长的建模方式，神算子平台能够大大降低构建模型的门槛，优化模型的效率。

高维度——平台内置了大数据科学院独有知识产权的高维度模型算法和特征工程算法，结合自主研发的高性能的分布式计算框架，在大数据和海量特征（支持万亿级以上）的场景下有很好的计算性能和计算效果。

全方位可视化——平台通过数据建模过程可视化、数据探索可视化、模型优化可视化实现了快速、高效的可视化建模，大大降低用户的建模成本和维护成本。

快速开始

登录/注册

注册入口 : senses.jusfoun.com/signup

如果您是一个新用户，用浏览器打开注册页面，填写邮箱、用户名、验证手机号码、设置密码，即可完成注册，其中公司名称和姓名为选填信息，注册成功后，系统会向您的注册邮箱发送一封激活邮件，从邮箱中点击邮件中的链接完成激活，即可登录使用神算子平台

 九次方大数据 | 神算子平台

欢迎注册神算子平台



The registration form consists of several input fields:

- 用户名 (Input field with placeholder: 请输入您的用户名)
- 邮箱 (Input field with placeholder: 请输入您的邮箱)
- 手机号 (Input field with placeholder: 请输入您的手机号码) - Includes a +86 prefix button and a "获取验证码" (Get Verification Code) button.
- 短信验证码 (Input field with placeholder: 请输入您的短信验证码) - Includes a "获取验证码" (Get Verification Code) button.
- 密码 (Input field with placeholder: 设置密码 (字母或数字至少6位))
- 真实姓名 (Input field with placeholder: 请输入您的真实姓名)
- 公司名称 (Input field with placeholder: 请输入您的公司名称)
- A checkbox labeled "我已阅读并同意《Jusfoun隐私条款》" (I have read and agree to the Jusfoun Privacy Policy).
- A "确定" (Confirm) button at the bottom.

登录入口 : senses.jusfoun.com/login

如果您已经有神算子平台账号，用浏览器打开登录页面，输入已经注册的账号密码，登录到平台。支持使用邮箱、用户名、手机号码登录

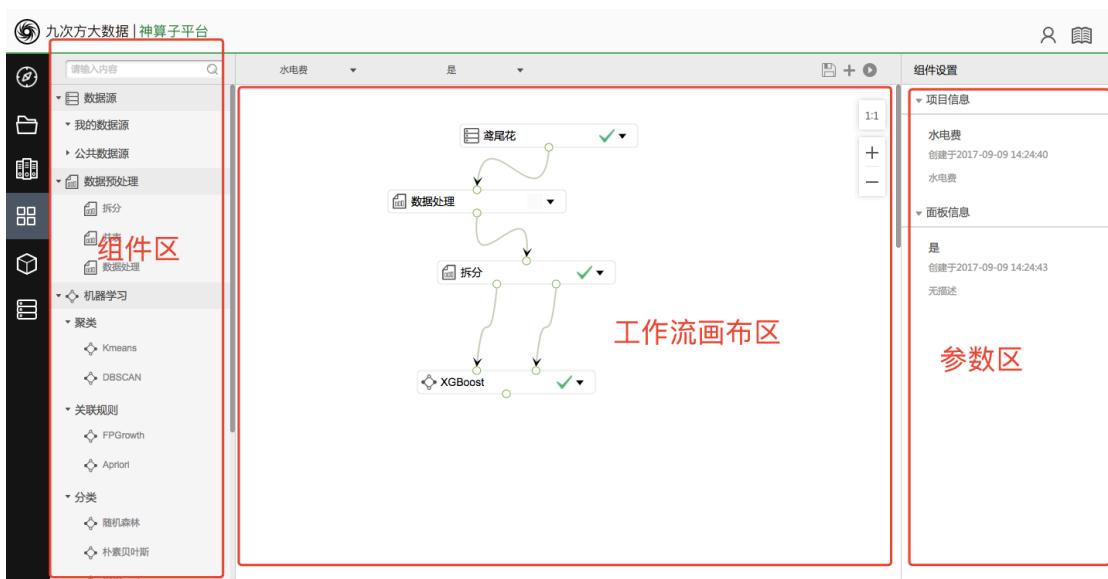


如果您忘记了密码，可点击密码输入框下方的“忘记密码”，验证核实您的信息后，可以进行密码重设

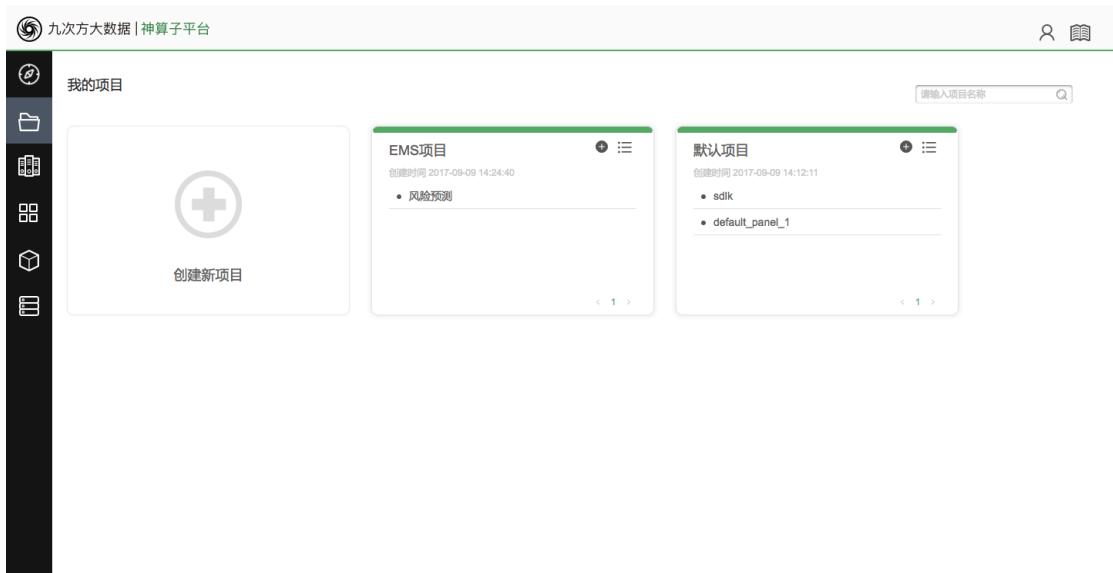
平台概览

神算子平台以项目为管理单位，每个项目下可以存放多个操作面板，每个面板对应一个建模流程

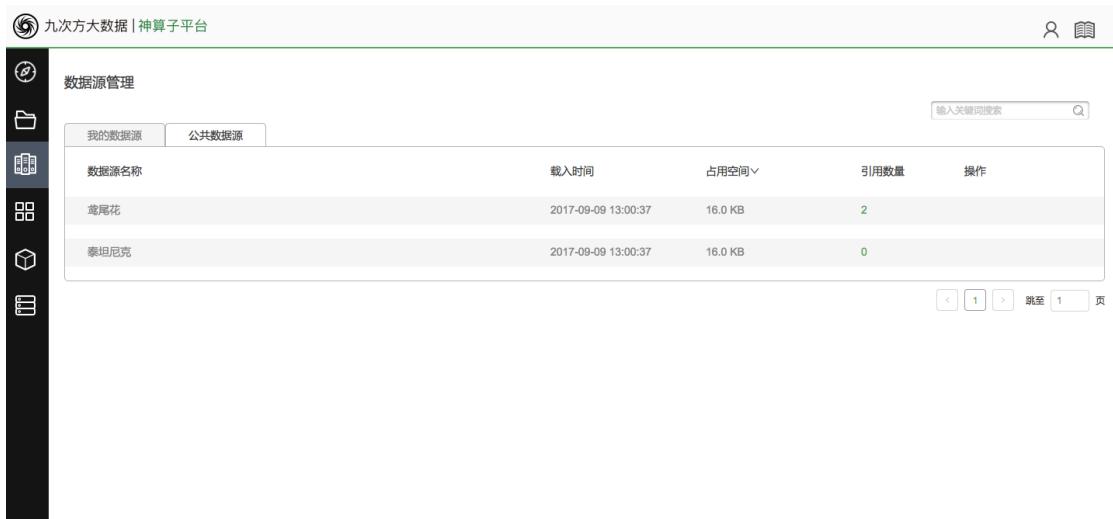
而工作流中需要用到的数据源、数据处理操作、算法等都作为组件出现，在面板中可以从左侧组件区域中拖拽到工作流中使用



项目管理模块可以进行项目和面板的管理，包括新建、修改、删除等操作

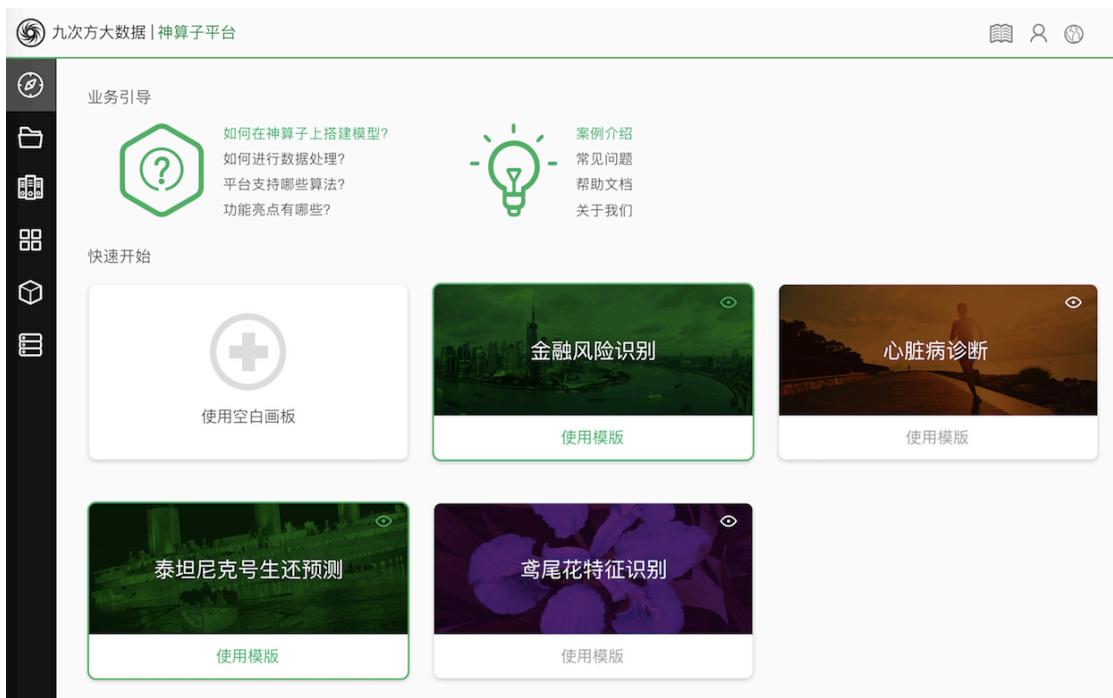


数据源模块可以进行自有数据源管理，包括数据源创建、修改、更新等操作



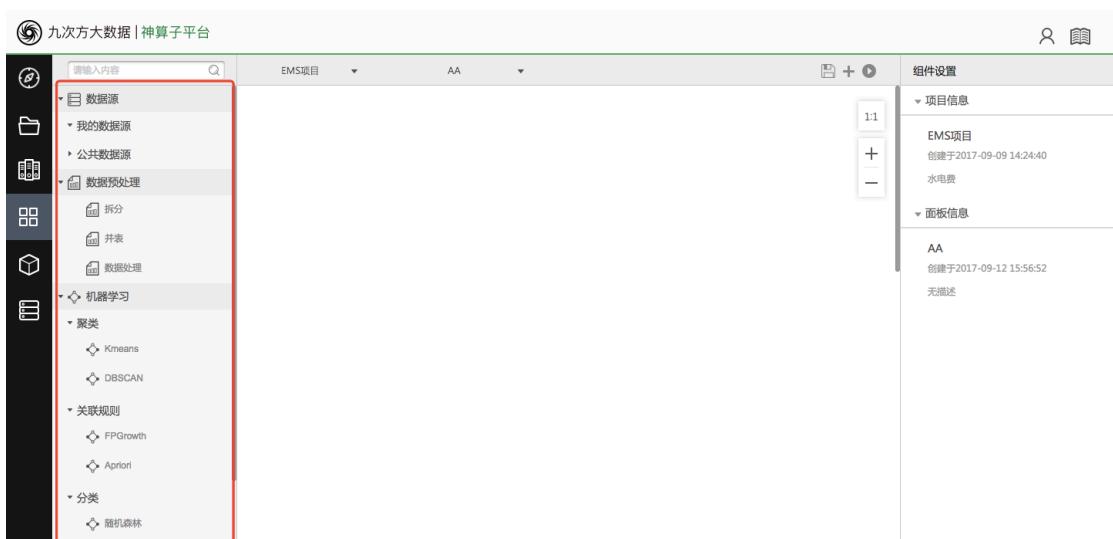
创建您的第一个工作流

登录到神算子平台，进入快速引导页面，选择“使用空白面板”



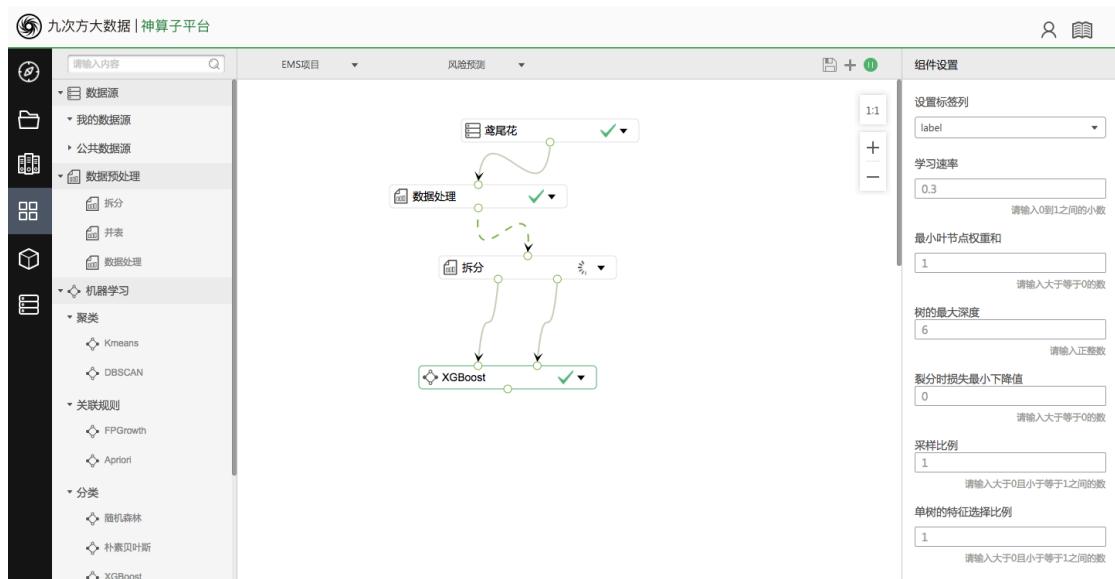
补充项目信息和面板信息以后，将进入到一个全新的空白面板中

从左侧组件区拖入必要的组件如：数据源，数据处理，拆分，算法，按照模型搭建顺序连线，即可完成工作流搭建。



组件参数设定

点击每个组件，可以从右侧的组件设置中调整组件对应参数，每个组件对应一组参数，参数的结果关系到模型效果，平台会根据经验给出参数的默认值



运行工作流

单击右上角的“运行”，工作流将按照您设定的顺序开始执行

查看运行结果

单击算法组件详情中的“数据集”可以预览算法输出的数据结果

sepalength	sepalwidth	petallength	petalwidth
5	3.4	1.5	0.2
4.9	3.1	1.5	0.1
5.4	3.7	1.5	0.2
5.7	4.4	1.5	0.4
4.6	3.6	1	0.2
5.1	3.3	1.7	0.5
4.9	3.1	1.5	0.1

单击算法组件详情中的“评估”可以查看算法结果评估情况，平台根据算法的不同提供了较为全面的评估信息，以分类算法为例，可以查看到分类结果的混淆矩阵，Acc、Precision、Recall等评估指标，ROC曲线，帮助您判断训练效果是否达到预期

		真实分类		
		0	1	2
预测分类	0	9	0	0
	1	0	4	0
	2	0	1	16

从模板创建工作流

您也可以从快速引导页中点击使用模板，系统将自动在面板中为您创建一个和模型相同的工作流，您可以直接使用模型结果，或者在此基础之上进行修改，调整为自己想要的模型

业务引导

如何在神算子上搭建模型?
如何进行数据处理?
平台支持哪些算法?
功能亮点有哪些?

案例介绍
常见问题
帮助文档
关于我们

快速开始

使用空白画板

金融风险识别

心脏病诊断

使用模版

泰坦尼克号生还预测

鸢尾花特征识别

使用模版

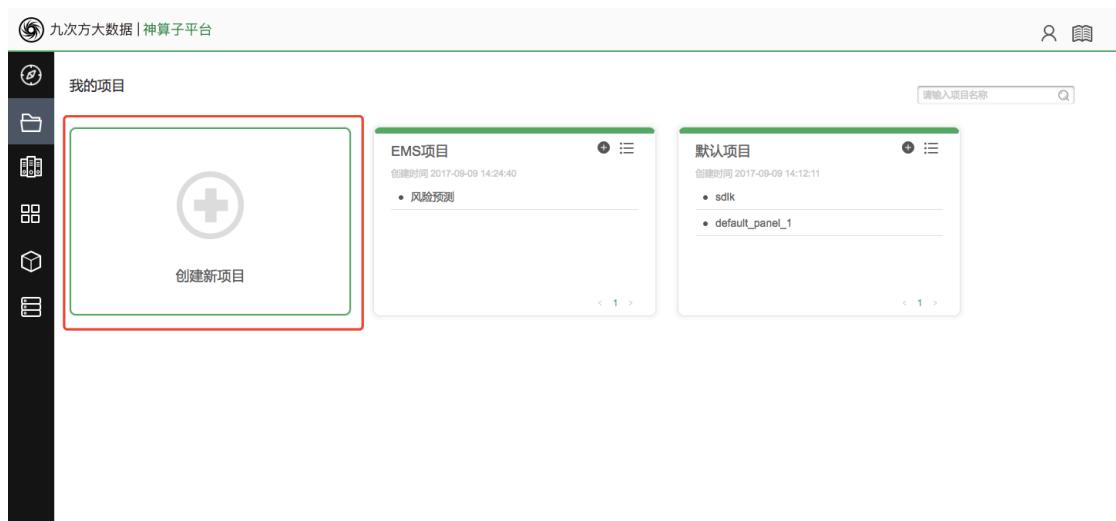
现在，尝试自己在平台上搭建属于自己的模型吧

项目管理

项目管理是用于管理面板的模块, 用户可以在这里完成项目和面板的创建、修改、删除

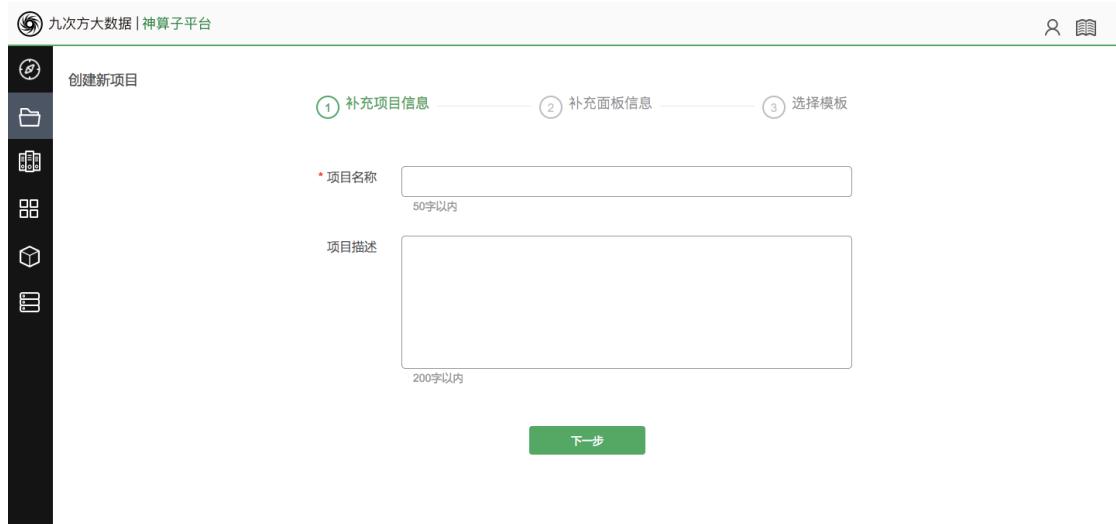
新建项目

1. 在项目管理中, 点击项目创建



The screenshot shows the 'My Projects' section of the platform. On the left is a sidebar with icons for project management. In the center, there's a search bar and a list of existing projects: 'EMS项目' (created 2017-09-09 14:24:40) and '默认项目' (created 2017-09-09 14:12:11). To the right of these is a large button labeled 'Create New Project' with a plus sign icon, which is highlighted with a red box.

2. 补充项目名称和项目描述, 点击下一步, 即可完成项目创建, 页面将自动进入面板创建



The screenshot shows the first step of the 'Create New Project' wizard. It has three tabs at the top: (1) 补充项目信息 (selected), (2) 补充面板信息, and (3) 选择模板. The '补充项目信息' tab contains fields for 'Project Name' (必填, 50字以内) and 'Project Description' (200字以内). Below these fields is a green 'Next Step' button.

新建面板

选择一个项目，在项目选项中点击创建，系统将把面板创建在选中的项目下

The screenshot shows the 'Create New Panel' interface. On the left, there is a sidebar with icons for creating new projects, managing files, and other operations. The main area displays two project options:

- EMS项目**: Created on 2017-09-09 14:24:40. Options: Edit (highlighted with a red box), Delete, and More.
- 默认项目**: Created on 2017-09-09 14:12:11. Options: Edit, Delete, and More.

A large red text overlay in the center says "从当前项目中创建面板" (Create panel from current project). At the bottom right of the interface is a green 'Next Step' button.

查看详情、修改和删除项目

在项目选项的更多中，可以进行项目详情查看（查看项目创建时间、项目名称和描述），编辑项目（修改项目名称和描述），删除项目（正在运行的项目不可删除）



修改和删除面板

鼠标停留在项目的面板信息中，可以进行删除面板，修改面板信息的操作



数据源管理

我的数据源

新建数据源-本地上传

The screenshot shows the 'Data Source Management' interface. On the left is a sidebar with icons for data source management. The main area has tabs for 'My Data Sources' and 'Public Data Sources'. A search bar at the top right contains a placeholder 'Search by keyword' and a magnifying glass icon. Below the search bar is a toolbar with a green '+' button labeled '+ New Data Source', a blue 'Local File' button (which is highlighted with a red arrow), and a grey 'Connect Database' button. The main table lists data sources with columns: Type, Data Source Name, Data Source Description, Creation Time, Size, Number of引用面版, and Operations. One entry is shown: 'Test' (Type: CSV), 'Test' (Name), 'Test' (Description), '2017-09-11 10:54:50' (Creation Time), '3.0 KB' (Size), '0' (Number of引用面版), and 'Operations' (button). At the bottom right are navigation buttons for pages 1-3 and a 'Jump to' input field.

1. 补充数据源信息

点击新建数据源按钮→本地文件，打开本地上传界面，*为必填项，键入数据源名称、数据源描述、选择上传文件，单击下一步；

The screenshot shows the 'Add Data Source - Local Upload' step 1. The sidebar on the left is identical to the previous screenshot. The main form has two tabs: ① 补充数据源信息 (Selected) and ② 确认字段信息 (Not Selected). The '补充数据源信息' tab contains fields: '数据源名称' (Data Source Name) with value 'iris' (必填项), '数据源描述' (Data Source Description) with value '鸢尾花数据' (200字以内), and '数据源上传' (Data Source Upload) with a green highlighted box containing '100% iris.csv' and a '重新选择' (Re-select) button. Below these are optional settings: '表头选项' (Header Options) with '有表头' (Has Header) selected, and '编码选项' (Encoding Options) with 'UTF-8' selected. At the bottom is a green '下一步' (Next) button.

2. 确认字段信息

预览数据查看上传数据源内容，点选字段名文本框可对字段名称进行修改，

点击字段类型下拉菜单可更换字段类型，完成以上操作后点击完成按钮；

The screenshot shows the 'Add Data Source' interface in the 'Local Upload' mode. It displays four fields with their current types: 'sepallength' is 'double', 'sepalwidth' is 'string', 'petallength' is 'string', and 'petalwidth' is 'double'. Below the table are two buttons: 'Preview Data' and 'Finish'.

连接数据库

1. 连接数据库

点击数据库标签页选择相应的数据块，*为必填项，键入数据块连接信息，

点击连接并选择数据库选择数据库，单击下一步。

The screenshot shows the 'Connect Database' step of the database connection wizard. The 'MySQL' tab is selected. Required fields are highlighted with red boxes: 'Host or IP Address' (192.168.15.143), 'Port' (3306), 'Username' (test), and 'Password' (empty). The 'Select Database' field contains 'dmp-data' and the 'Connect and Select Database' button is also highlighted with a red box. A green 'Next Step' button is at the bottom.

2. 选择数据表

点击预览按钮可查看数据库表数据，勾选需要上传的数据库表，单击下一步。

The screenshot shows the 'Select Data Table' step of a data import wizard. The top navigation bar includes '九次方大数据 | 神算子平台' and three tabs: '① 连接数据库', '② 选择数据表' (highlighted in green), and '③ 补充数据源信息'. On the left is a sidebar with icons for database connection, file management, and other operations. The main area displays a table titled '选择数据表(已过滤空表)'. The table lists several database tables with their names, column counts, and row counts. For each table, there is a checkbox and a 'Preview' button. The 'Preview' button for the first table is highlighted with a red box. A search bar at the top right allows inputting keywords to search for specific tables.

3. 补充数据源信息

补充数据源信息→导入为数据源

The screenshot shows the 'Supplement Data Source Information' step of the data import wizard. The top navigation bar includes '九次方大数据 | 神算子平台' and three tabs: '① 连接数据库', '② 选择数据表' (highlighted in green), and '③ 补充数据源信息' (highlighted in green). On the left is a sidebar with icons for database connection, file management, and other operations. The main area displays a table titled '数据源信息补充'. It has columns for '数据库名称' (Database Name), '表名称' (Table Name), '数据源名称' (Data Source Name), and '数据源描述' (Data Source Description). A single row is shown, with the '表名称' field containing 'dataset_012bf5aa37da4d1f88...' and the '数据源名称' field containing 'dmp-data_dataset_012bf5aa'. A '测试' (Test) button is also visible. At the bottom is a large green '导入为数据源' (Import as Data Source) button.

公共数据源

选择公共数据源标签页进入公共数据源管理界面，点击预览按钮可查看公共数据源的数据内容。

九次方大数据 | 神算子平台

数据源管理

我的数据源 公共数据源

输入关键词搜索

数据源名称	载入时间	占用空间	引用数量	操作
乳腺癌分类	2017-09-12 11:28:12	122.3 KB	0	
学生学习成绩挖掘数据	2017-09-11 16:35:14	37.1 KB	0	
人才流失预测	2017-09-11 16:33:37	222.6 KB	0	
房价预测	2017-09-11 16:32:32	2456.3 KB	0	
医院的综合信息和质量评价	2017-09-11 16:31:10	2597.4 KB	0	
鸢尾花	2017-09-09 13:00:37	16.0 KB	0	
泰坦尼克	2017-09-09 13:00:37	16.0 KB	0	

< 1 > 跳至 1 页

工作流搭建和控制

通过创建面板，可以进入工作流的搭建；选择一个已经存在的面板，可以进入对应工作流的编辑界面；

工作流是模型搭建的核心模块，也是您建模的主要操作界面，为了能最快开始建模，平台左侧提供了一个快捷入口

九次方大数据 | 神算子平台

请输入内容

EMS项目 AA

我的数据源

公共数据源

数据预处理

拆分

并表

数据处理

机器学习

聚类

工作流的主要功能结构如下：



工作流搭建

从一个空白面板开始，您可以搭建您的建模流程

在建模的常规方法中，一套完整的模型工作流为：对数据源进行数据清洗，筛选和创建特征，拆分数据，进行算法训练，然后根据结果不断调整特征、参数和算法，直至达到满意的训练结果

选定组件

组件是平台已经封装好的运算对象或者运算方法，包含

所有组件都分门别类的放在左侧组件区中，将需要的组件依次拖拽进入中间的画布区

连线

每个组件都有对应的输入点和输出点，输入点接收用于组件运算的数据，输出点用于向下一个组件输出运算数据

将一个组件的输出点连接到另一个组件的输出点，可以搭建起两个组件之间的运算关系；根据流程运行顺序连接好所有需要的组件，即可完成流程设定

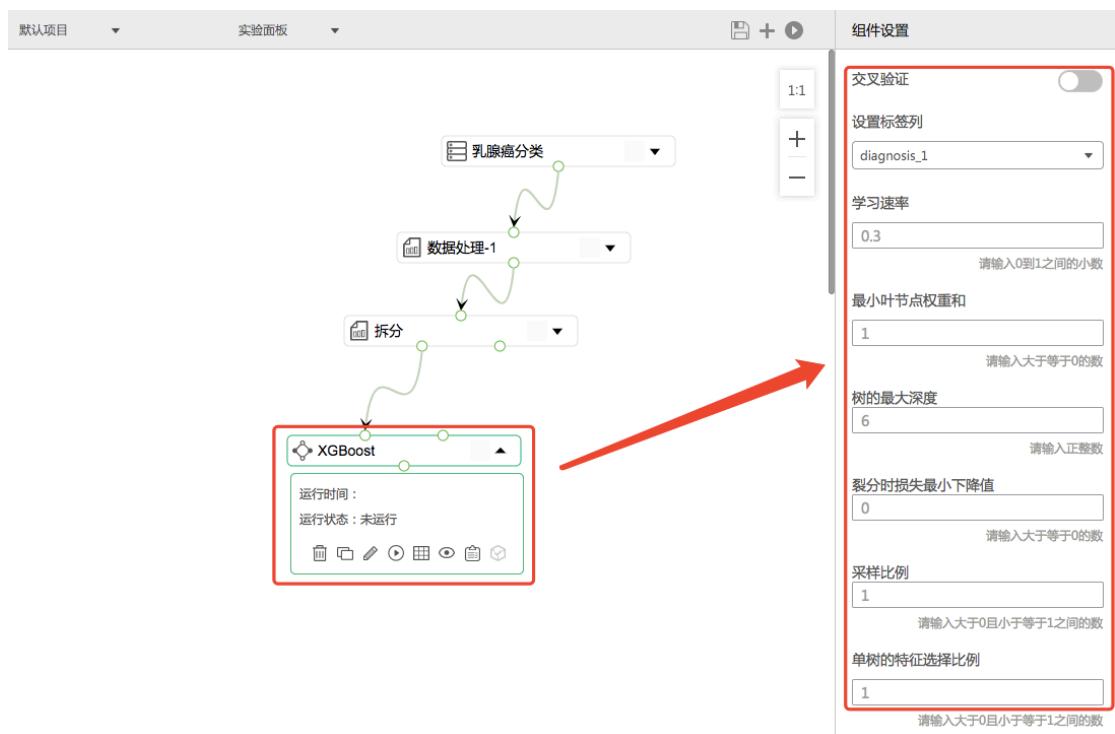
组件操作

在每个组件的下拉菜单中，有一组针对组件的操作

操作名称	图标	说明	适用组件
删除		删除目标组件实例	所有组件
复制		复制目标组件实例，包含组件对应的参数	所有组件
重命名		组件实例重命名	所有组件
运行选项		以目标组件为运行节点，有两个选项： 运行到此处：流程运行到目标组件实例即停止 从此处运行：从目标组件开始运行	所有组件
输出数据		预览组件的运行结果数据，默认显示前100行	所有组件
评估		算法类组件的效果评估，平台为不同的算法设定了不同的评估方法	算法类组件，关联规则算法除外
运行日志		目标组件在流程中的运行日志	所有组件
模型保存		将运行成功的算法组件保存为模型	算法类组件

组件设置

在画布区选中一个组件，右侧将显示对应的组件参数设置



面板控制

位于画布上方，可对工作流进行运行控制和切换面板的操作



操作名称	说明
项目切换	删除目标组件实例
面板切换	复制目标组件实例，包含组件对应的参数
自动保存	所有工作流内容将被实时保存，无需额外操作
另存为	将当前面板内容保存到一个新面板中
新建	在当前项目下创建一个新的面板
运行/停止	运行：开始完整运行整个工作流直至所有组件运行完成 停止：停止正在运行的工作流

模型保存与使用

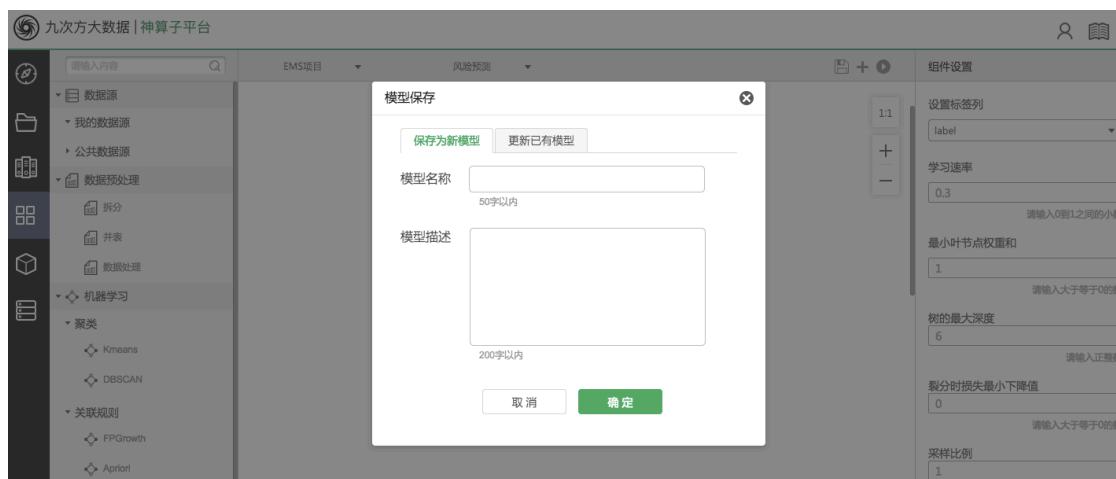
模型是指通过训练将算法的运算参数确定下来以后形成的一套计算方法，新的数据集输入到模型中可以直接预测出结果

模型保存

已经成功运行的算法组件，可以保存为模型进行复用，入口位于组件实例详情的操作中：



补充模型信息后创建为新模型，或者更新到一个已存在的模型中，当前模型将会替换已有模型



模型在工作流中的使用

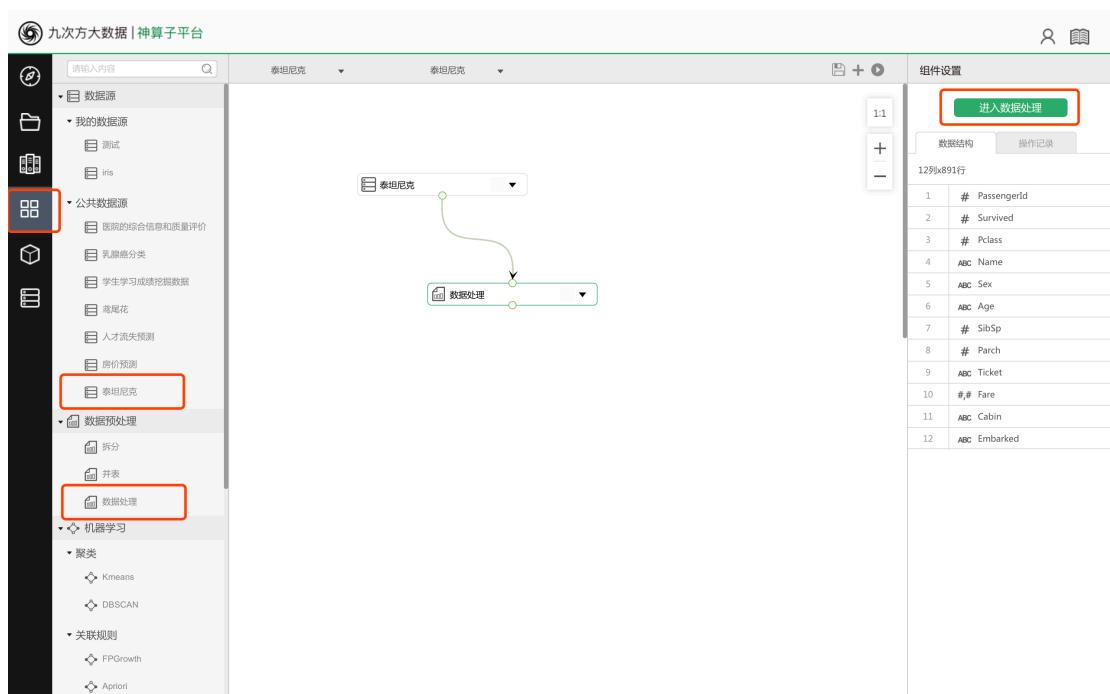
保存好的模型将出现在左侧组件区，可以当做组件被拖拽到工作流中使用



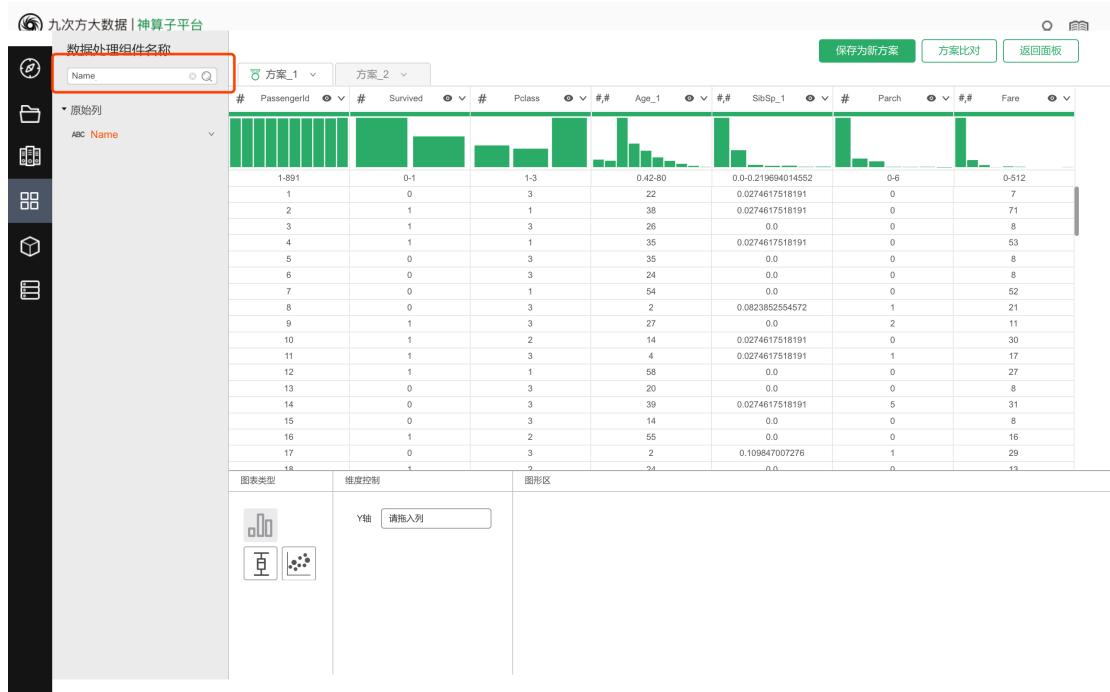
数据处理

常见数据处理方法

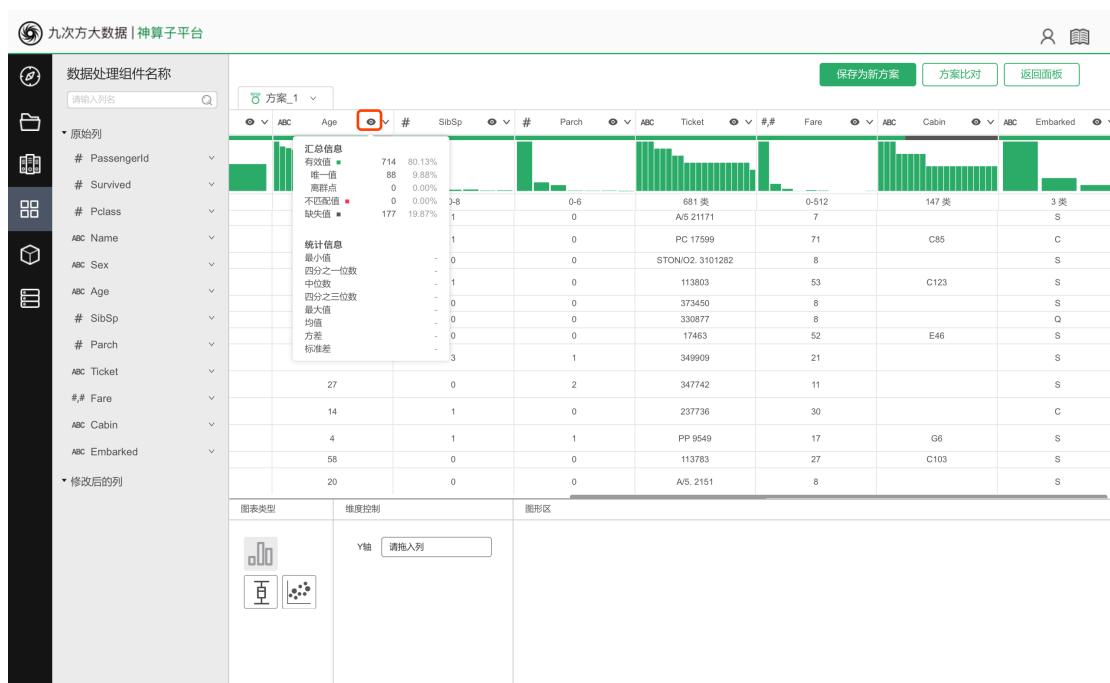
选择左侧数据源拖拽到面板区域创建组件示例，点击组件边缘圆点添加组件连线，进行数据关联，点击数据处理组件，页面右侧可查看数据结构和操作记录，点击进入数据处理按钮。



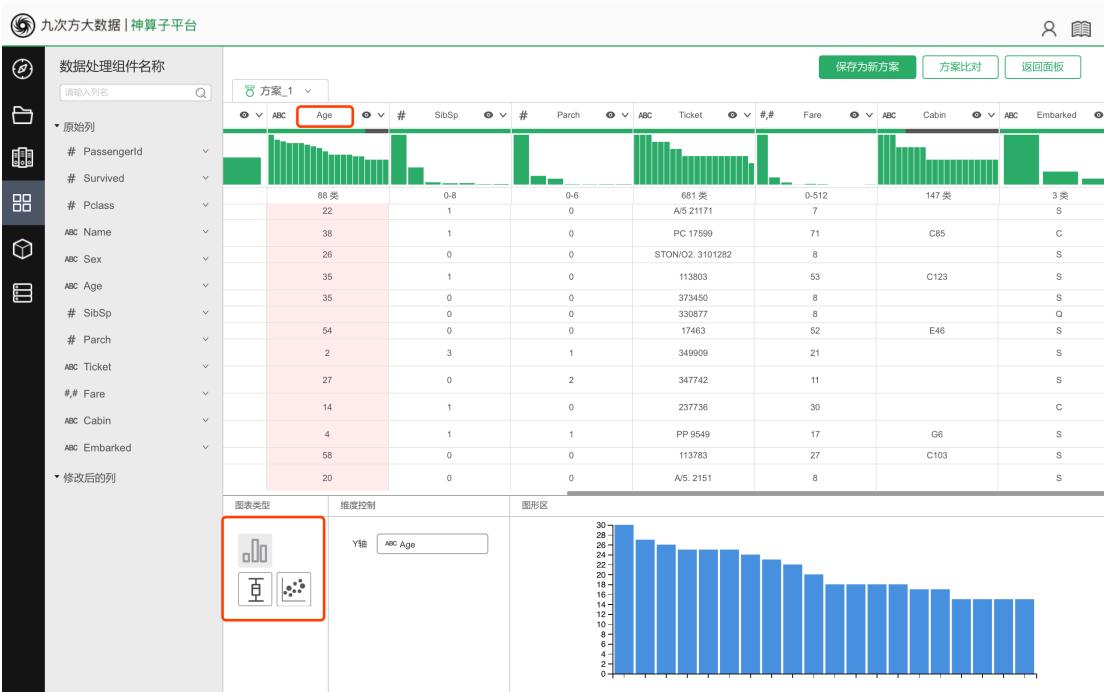
左侧文本框输入列名称→点击插叙按钮→查询相应列。



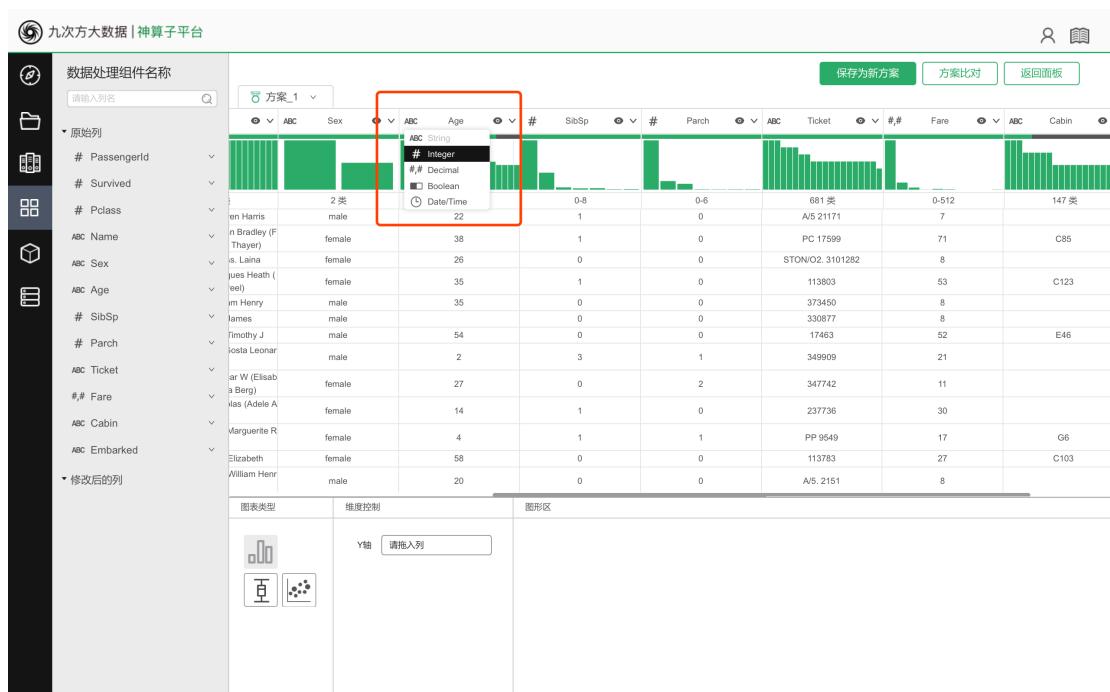
进入数据处理界面，每列数据下方以柱形图方式展示数据分布状态，中间横条以不同颜色展示数据状态，绿色区域表示有效值，红色区域表示异常值，黑色区域表示缺失值，点击数据透视按钮可查看每列数据的汇总信息。



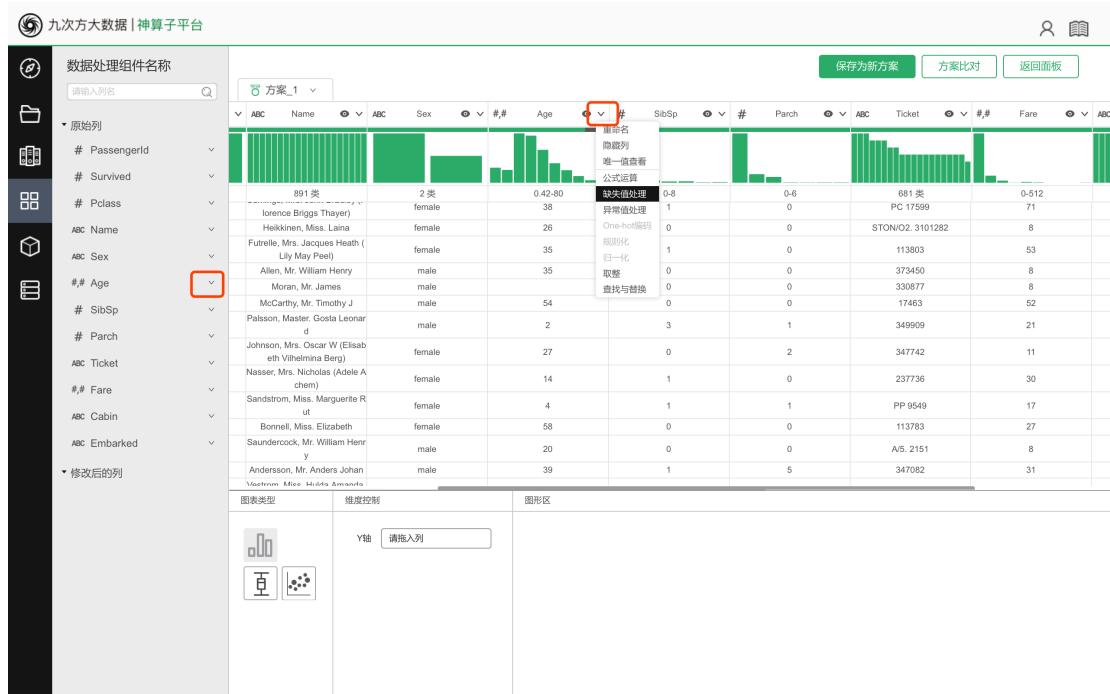
单机列名，以图表形式展示数据信息,选择不同的图标类型，可做相应的图标展示。将左侧展示列拖拽到维度控制，可做相应的列叠加的图标展示。



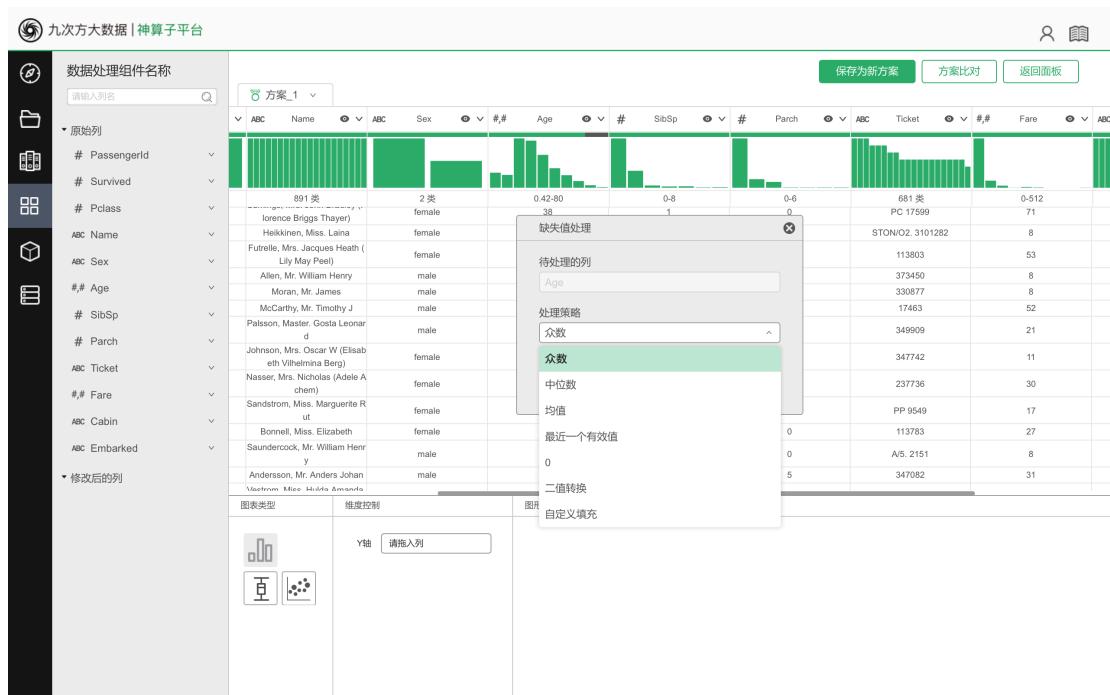
点击列名左侧列类型下拉菜单，选择数据类型进行相应数据类型转换



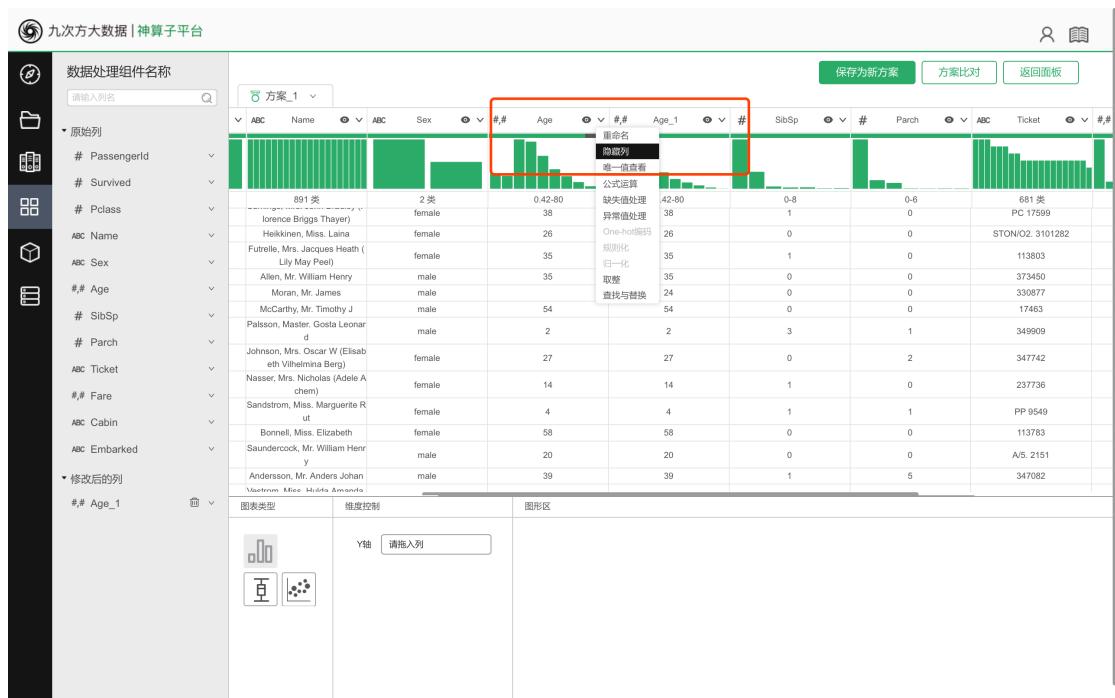
点击列名右侧按钮打开预处理下拉菜单。



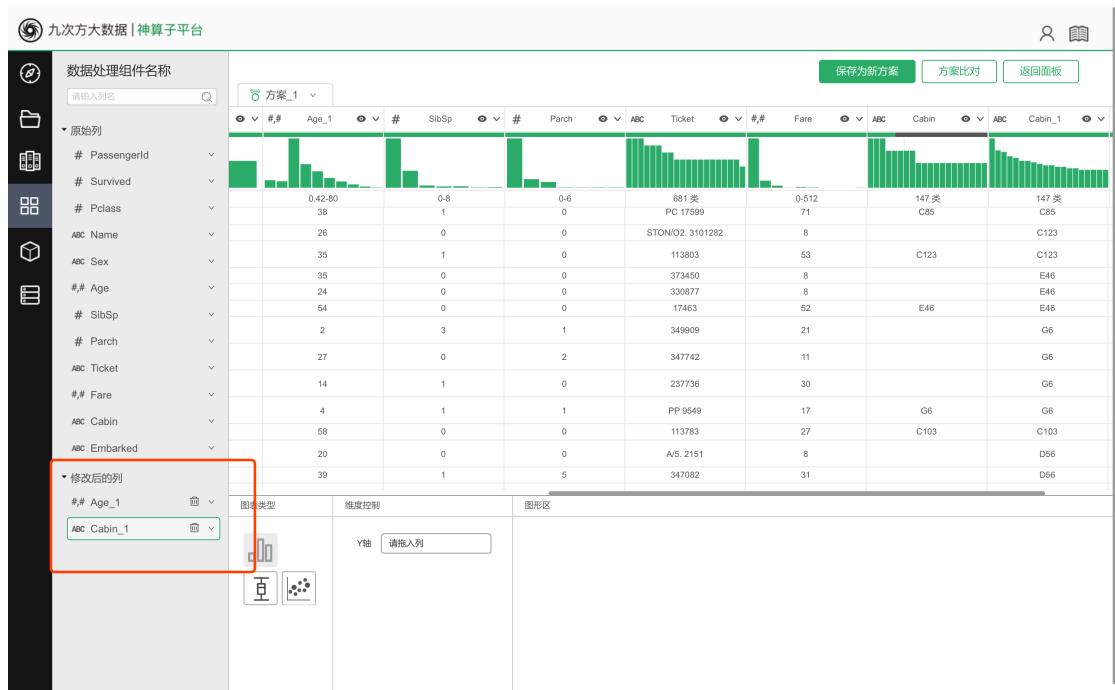
根据数据形态选择需要进行的数据预处理，操作完成产生新的数据列。



将原列隐藏不参与后续计算，也可根据需求隐藏掉无用的特征列。

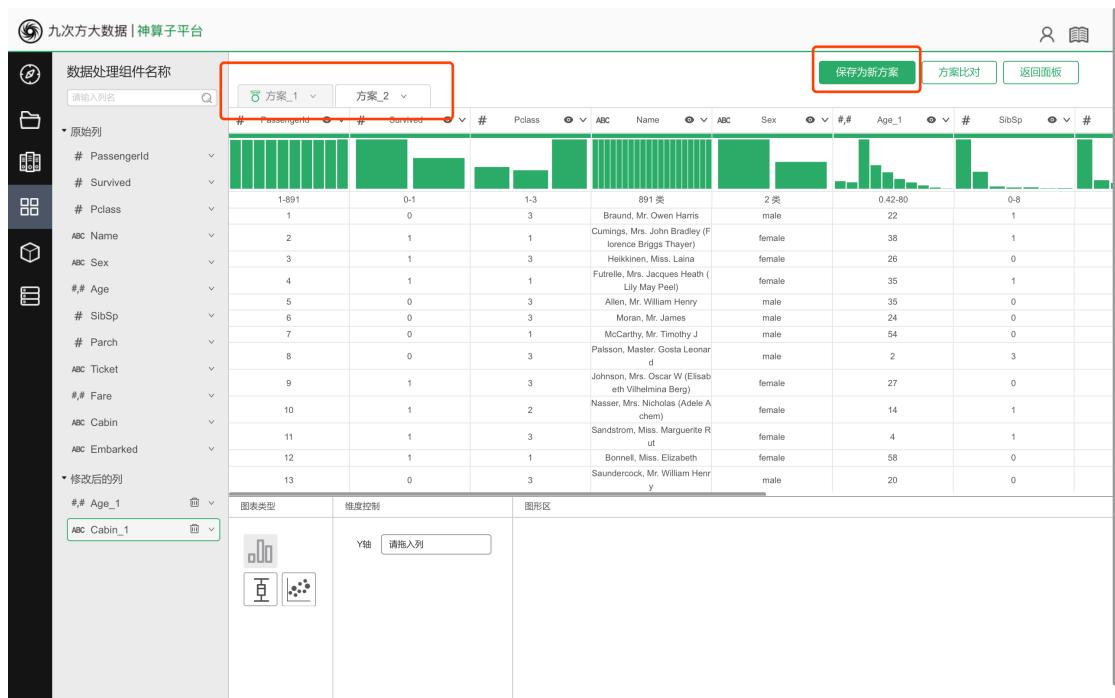


先隐藏修改后的列后，可在左侧菜单中对该列进行删除操作。

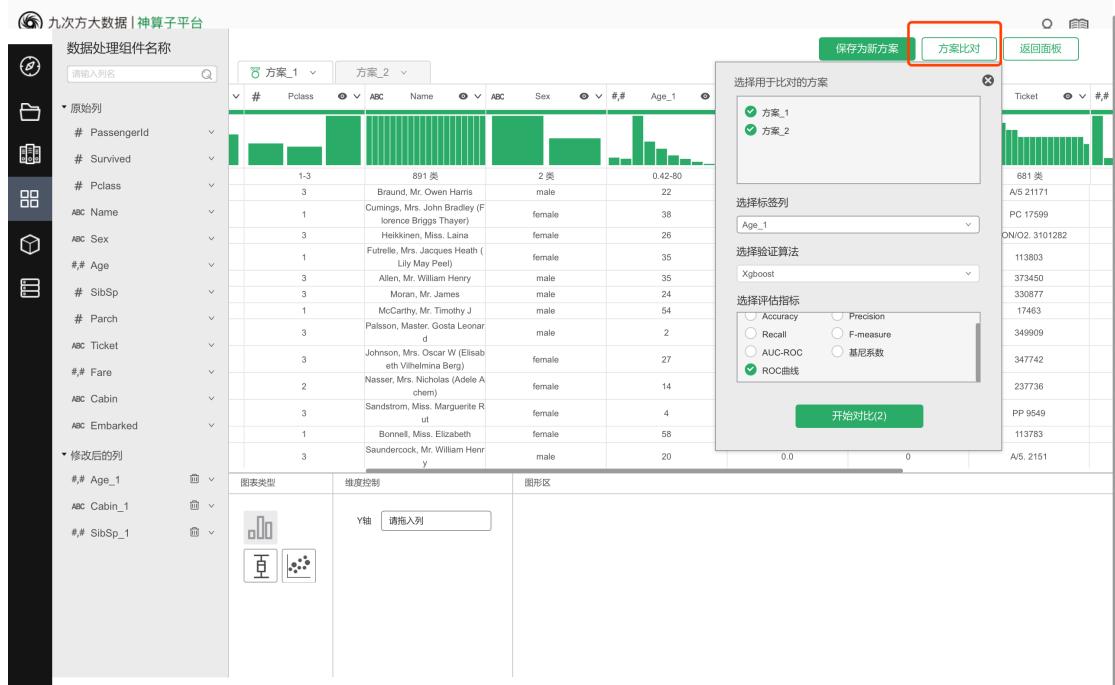


数据处理方案比对

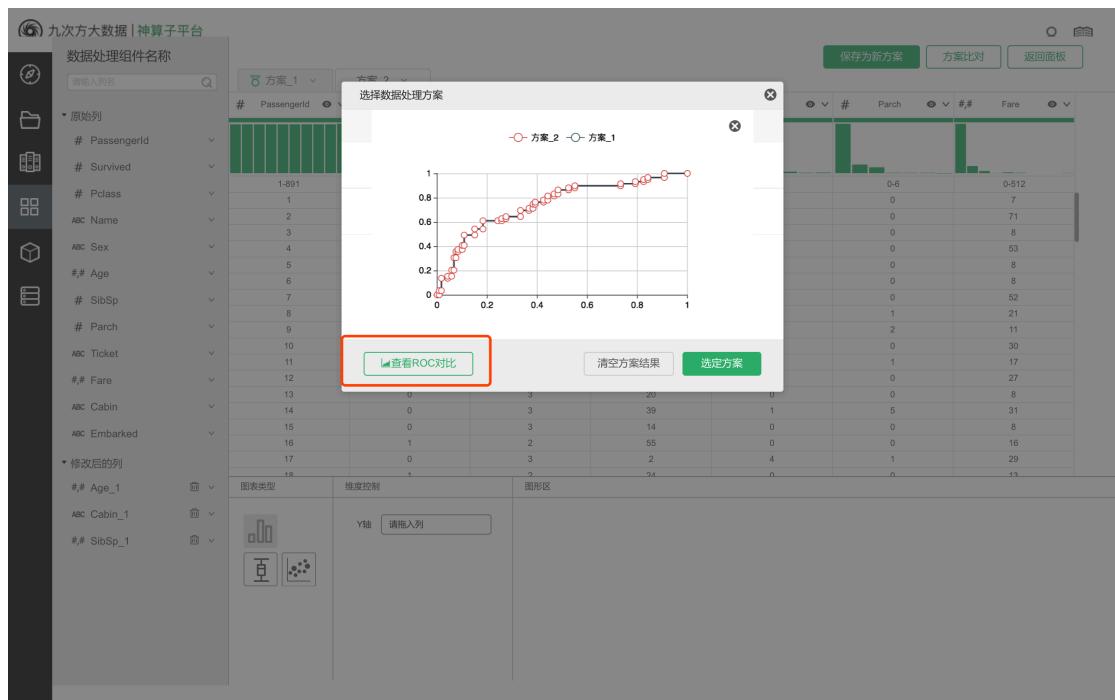
点击保存为新方案按钮可把当前进度保存为新方案。



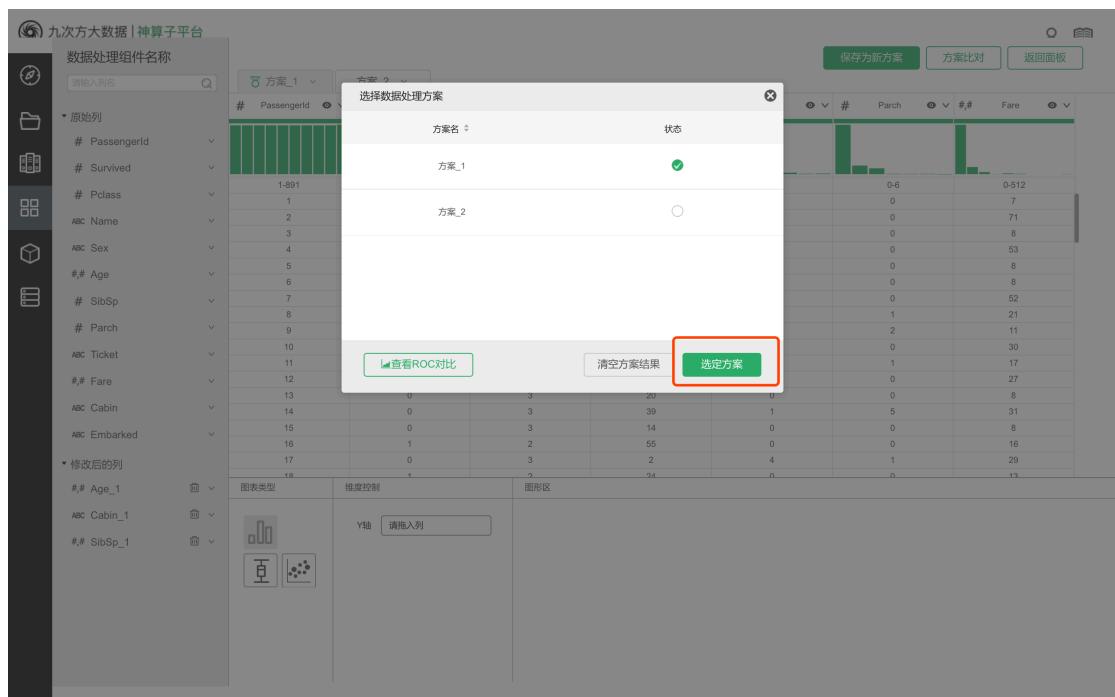
点击方案对比按钮，选择标签列、验证算法、评估指标→开始方案对比。



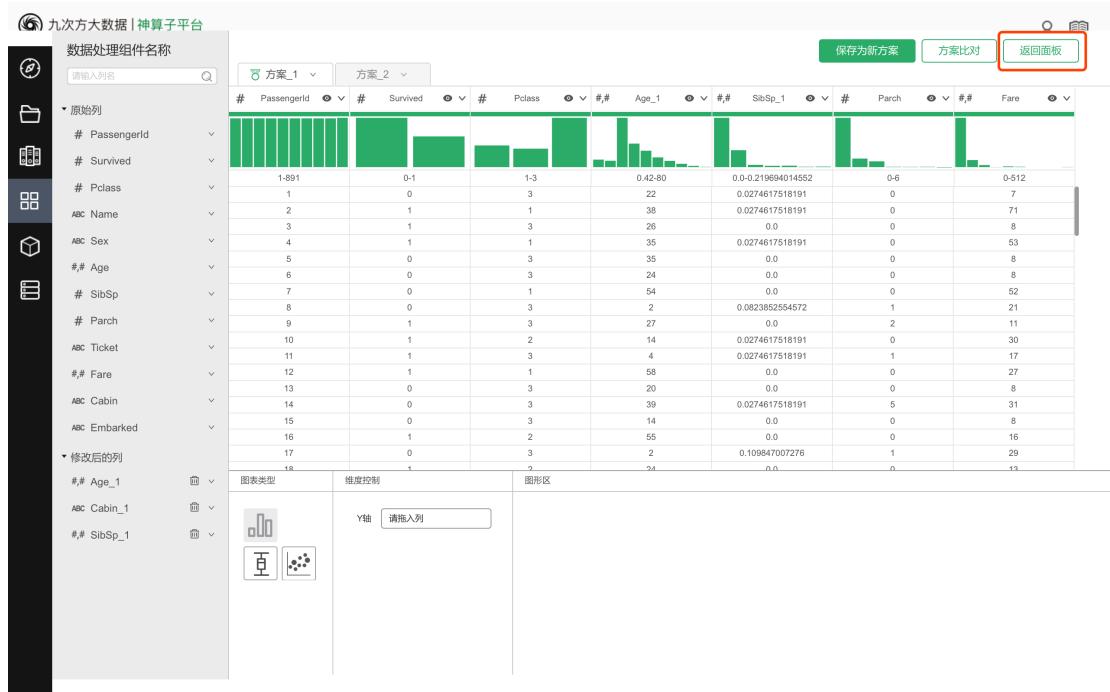
方案对比完成后，可查看 ROC 曲线。



选定方案



1. 返回面板进行后续处理。



模型管理

在模型管理模块可以对保存的模型进行查看和修改



操作中点击模型信息查看，将显示对应的模型信息，包括模型名称，模型描述，特征列，标签列，算法来源，超参数，来源面板，更新时间

The screenshot shows the 'Model Management' interface. On the left, there's a sidebar with icons for project management, data, components, and models. The 'Model Management' section is selected. A list of models is shown, with 'XGB3' selected. A modal window titled 'Model Information View' displays detailed information about the selected model:

模型信息查看	
模型名称	XGB3
模型描述	
特征列	4个特征列 #,# sepallength #,# sepalwidth #,# petallength #,# petalwidth
标签列	label
算法来源	XGBoost
超参数	colsample_bytree 1 min_child_weight 1 subsample 1 eta 0.3 max_depth 6 gamma 0
来源面板	风险预测
最后更新时间	2017-09-13 11:54:21
创建时间	2017-09-13 11:54:21

点击“编辑”图标能够修改模型的名称和描述

The screenshot shows the 'Model Management' interface. The 'Model Edit' dialog is open over the 'Model Management' list. It contains fields for 'Model Name' (XGB3) and 'Model Description'. The right side of the screen shows a list of panels and their usage counts.

点击删除可以删除选中模型，删除操作会将模型从面板左侧的组件区删除，但并不会删除面板中对应的组件实例

服务发布

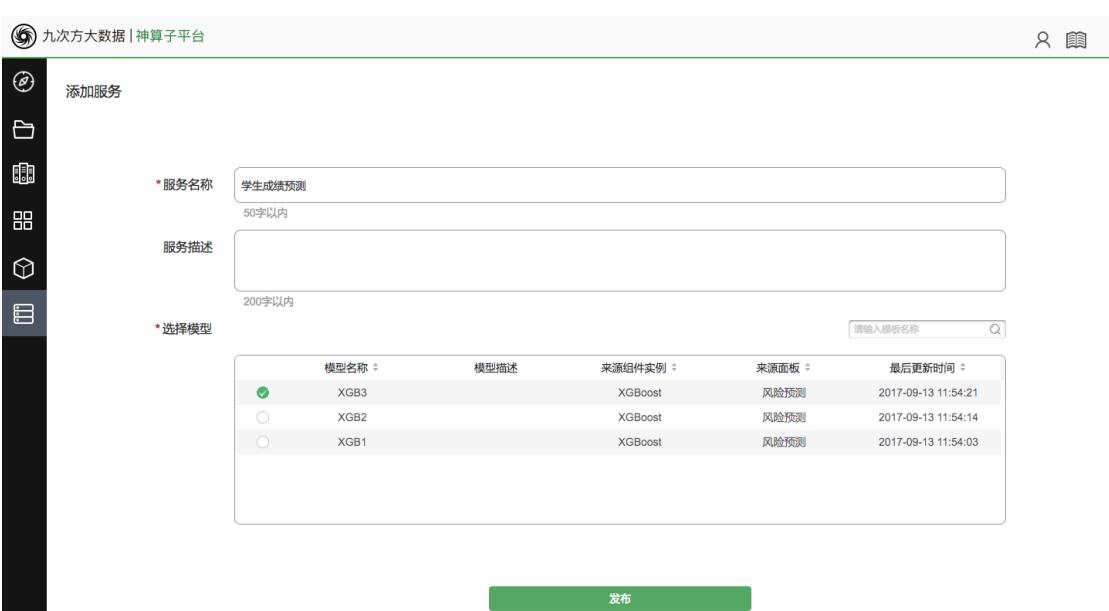
服务发布中，支持将已经训练好的模型形成接口，方便在实际业务中使用

发布新服务

点击右上角的“+添加服务”，补充服务信息，选择服务对应的模型，不超过 30 秒，就可以添加完成



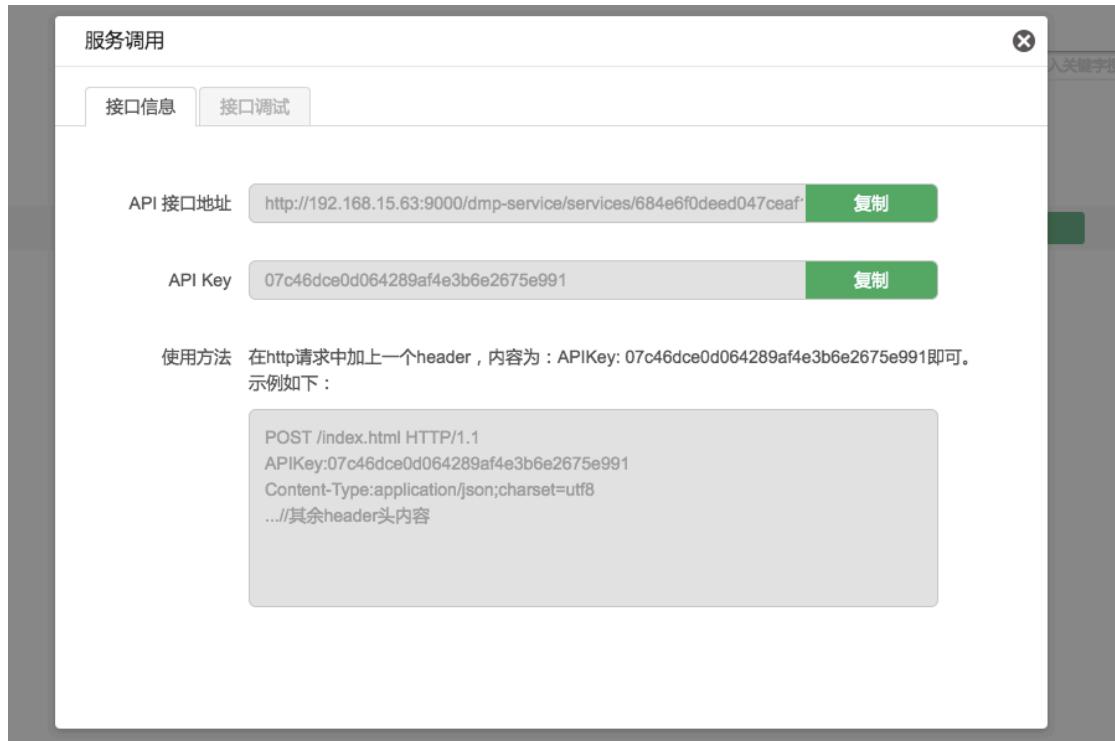
The screenshot shows the 'Service Release' section of the platform. It lists a single service entry: '学生成绩预测' (Student Score Prediction) with deployment time and last update both at '2017-09-13 13:51:49'. Below the table are buttons for 'Service Invocation', 'Service Update', and 'Deactivate'. A red box highlights the green 'Add Service' button in the top right corner.



The screenshot shows the 'Add Service' page. It has fields for 'Service Name' ('学生成绩预测') and 'Service Description'. Below is a table titled 'Select Model' listing three models: XGB3, XGB2, and XGB1. XGB3 is selected. A search bar for 'Model Name' is also present. At the bottom is a large green 'Release' button.

调用服务

在模型列表中点击对应模型的“服务调用”，可以查看服务接口信息，您可以直接在程序中通过调用该接口，得到模型预测数据



选择接口调试，可以对生成的接口进行测试，建议您确保接口可用、返回数据达到要求之后再正式环境中使用该接口



服务的启用、停用与更新

对于已经发布的服务，可以进行停用与更新操作：

操作名称	说明
服务停用	对于已生效的服务，该操作将使接口失效
服务启用	已停用的服务进行启用操作，恢复接口可用性，而不会变更接口信息
服务更新	更新服务对应的模型，更新以后再发送到接口的数据将根据新的模型运算出结果

算法组件介绍

回归算法

1. 线性回归

线性回归是利用数理统计中的回归分析，来确定两种或两种以上变数间相互依赖的定量关系的一种统计分析方法之一，应用十分广泛。变量的相关关系中最为简单的是线性相关关系，设随机变量与变量之间存在线性相关关系，则由试验数据得到的点(,)将散布在某一直线周围。

2. 逻辑回归

这里指的是用逻辑回归的方法处理回归问题
用给定的输入变量(X) 来预测二元的结果(Y) (1/0,是/不是, 真/假)。我们一般用虚拟变量来表示二元/类别结果。你可以把逻辑回归看成一种特殊

的线性回归，简单来说，逻辑回归是利用 logit 函数拟合数据来预测某一个事件发生的概率的。

3. XGboost 回归

XGBoost 是 “Extreme Gradient Boosting” 的简称，是在 GBDT 的基础上对 boosting 算法进行的改进，内部决策树使用的是回归。Xgboost 是 GB 算法的高效实现，xgboost 中的基学习器除了可以是 CART (gbtree) 也可以是线性分类器 (gblinear)

4. 支持向量机回归

分算法用支持向量机处理回归问题

支持向量机(Support Vector Machine, SVM)通过使用最大分类间隔采设计决策最优的划分超平面，以获得良好的泛化能力。SVM 通过核函数的方法将低维数据映射到高维甚至无限维空间，从而能够处理低维空间中线性不可分的数据。SVM 主要应用在模式识别领域中的文本识别、文本分类、人脸识别等问题中，同时也应用到许多的工程技术和信息过滤等方面。

分类算法

1. 随机森林

随机森林是由多棵决策树形成的“森林”，它是一个监督学习的提升模型。森林中的每棵树都是基于随机抽取的样本和特征，独立训练出来的模型，随机森林就像 是由很多个专家组成的团队，团队中的每个专家擅长不同的领域，进行分类或者 回归时，就由这些专家投票进行表决。

2. 逻辑回归分类

逻辑回归是一种分类的算法，它用给定的输入变量（X）来预测二元的结果

（Y）（1/0,是/不是， 真/假）。我们一般用虚拟变量来表示二元/类别结果。

你可以把逻辑回归看成一种特殊的线性回归，只是因为最后的结果是类别变量， 所以我们需要用胜算比取对数来作为因变量（Dependent Variable）。简单来说，逻辑回归是利用 logit 函数拟合数据来预测某一个事件发生的概率的。

3. 决策树分类

决策树（decision tree）是一个树结构（可以是二叉树或非二叉树）。其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果。

4. 朴素贝叶斯

朴素贝叶斯（naive Bayes）法是基于贝叶斯定理 和 特征条件独立假设的分类方法，对于给定的训练数据集，首先基于特征条件独立假设学习输入/输出的联合分布概率；然后基于此模型，对给定的输入 x ，再利用贝叶斯定理求出其后验概率最大的输出 y 。

5. 支持向量机分类

本算法采用支持向量机处理分类问题

支持向量机(Support Vector MacHine, SVM)通过使用最大分类间隔采设计决定最优的划分超平面，以获碍良好的泛化能力。SVM 通过核函数的方法将低维数据映射到高维甚至无限维空间，从而能够处理低维空间中线性不可分的数据。SVM 主要应用在模式识别领域中的文本识别、文本分类、人脸识别等问题中，同时也应用到许多的工程技术和信息过滤等方面。

6. XGboost 分类

XGBoost 是 “Extreme Gradient Boosting” 的简称，是在 GBDT 的基础上对 boosting 算法进行的改进，内部决策树使用的是回归。Xgboost 是 GB 算法的高效实现，这里用 XGBoost 来解决分类问题，也是 XGboost 的最经典应用

7. K 近邻

K 近邻(K-Nearest Neighbor, KNN)是一社最经典和简单的再监督学习方法之一 K 近邻(K-Nearest Neighbor, KNN)是一社最经典和简单的再监督学习方法之一。当对数据的分布只有很少或者没有任何先验知识时，K 近邻算法是一个不错的 选择。该方法有着非常简单的原理：当对测试样本进行分类时，首先扫描训练集，找到与该测试样本最相似的 K 个训练样本，根据这 K 个样本的类别进行投票确定 测试样本的类别。也可以通过 K 个样本与测试样本的相似程度进行加权投票。

8. Adaboost 分类

Adaboost 是一种迭代算法，其核心思想是针对同一个训练集训练不同的分类器(弱分类器)，然后把这些弱分类器集合起来，构成一个更强的最终分类

器(强分类器)。其算法本身是通过改变数据分布来实现的，它根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练，最后将每次训练得到的分类器最后融合起来，作为最后的决策分类器。使用adaboost 分类器可以排除一些不必要的训练数据特征，并放在关键的训练数据上面。

9. 多层神经网络分类

MLP (Multi-layer Perceptron)，即多层感知器，是一种前向结构的人工神经网络，映射一组输入向量到一组输出向量。MLP 可以被看做是一个有向图，由多个节点层组成，每一层全连接到下一层。除了输入节点，每个节点都是一个带有非线性激活函数的神经元（或称处理单元）。一种被称为反向传播算法的监督学习方法常被用来训练 MLP。MLP 是感知器的推广，克服了感知器不能对线性不可分数据进行识别的弱点。

聚类算法

1. K-means

K-means 算法是硬聚类算法，是典型的基于原型的目标函数聚类方法的代表，它是数据点到原型的某种距离作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则。K-means 算法以欧式距离作为相似度测

度，它是求对应某一 初始聚类中心向量 V 最优分类 J 使得评价指标 J 最小。算法采用误差平方和准则函数作为聚类准则函数。

2. DBSCAN

DBSCAN 是一种基于密度的空间聚类算法，该算法将具有足够密度的区域划分为 聚类，并在具有噪声的空间数据库中发现任意形状的聚类，它将聚类定义为密度相 连的点的最大集合。该算法利用基于密度的聚类的概念，即要求聚类空间中的一定 区域内所包含对象（点或其他空间对象）的数目不小于某一给定阈值。DBSCAN 算 法的显著优点是聚类速度快且能够有效处 理噪声点和发现任意形状的聚类。

关联规则算法

1. FP-Growth

FP-Growth 算法是韩家炜等人在 2000 年提出的关联分析算法，它采取如下分治策略：将提供频繁项集的数据库压缩到一棵频繁模式树（FP-tree），但仍保留项集关联信息。在算法中使用了一种称为频繁模式树（Frequent Pattern Tree）的数据结构。FP-tree 是一种特殊的前缀树，由频繁项头表和项前缀树构成。FP-Growth 算法基于以上的结构加快整个挖掘过程。

2. Apriori

Apriori 算法是一种挖掘关联规则的频繁项集算法，其核心思想是通过候选集生成和情节的向下封闭检测两个阶段来挖掘频繁项集。该算法的基本思想是：首先找出所有的频集，这些项集出现的频繁性至少和预定义的最小支持

度一样。然后由频集产生强关联规则，这些规则必须 满足最小支持度和最小可信度。然后使用第 1 步找到的频集产生期望的规则，产生只包含集合的项的所有规则，其中每一条规则的右部只有一项，这里采用的是中规则的定义。一旦这些规则被生成，那么只有那些大于用户给定的最小可信度的规则才被留下来。

时间序列

1. ARIMA

ARIMA (p, d, q) 称为差分自回归移动平均模型，是一个著名时间序列预测方法。AR 是自回归， p 为自回归项；MA 为移动平均， q 为移动平均项数， d 为时间序列成为平稳时所做的差分次数。所谓 ARIMA 模型，是指将非平稳时间序列转化为平稳时间序列，然后将因变量仅对它的滞后值以及随机误差项的现值和滞后值进行回归所建立的模型。ARIMA 模型根据原序列是否平稳以及回归中所含部分的不同，包括移动平均过程 (MA)、自回归过程 (AR)、自回归移动平均过程 (ARMA) 以及 ARIMA 过程。

2. 一阶指数平滑

一阶指数平滑实际就是对历史数据的加权平均，它可以用于任何一种没有明显函数规律但确实存在某种前后关联的时间序列的短期预测。它最突出的优点是方法非常简单，甚至只要样本末期的平滑值，就可以得到预测结果。能够跟踪数据变化。这一特点所有指数都具有。预测过程中添加最新的样本数据后，新数据应取代老数据的地位，老数据会逐渐居于次要的地位。

位，直至被淘汰。这样，预测值总是反映最新的数据结构。一次指数平滑有局限性。第一，预测值不能反映趋势变动、季节波动等有规律的变动；第二，这种方法多适用于短期预测，而不适合作中长期的预测；第三，由于预测值是历史数据的均值，因此与实际序列的变化相比有滞后现象。

3. 二阶指数平滑

二次指数平滑是对一次指数平滑的再平滑。它适用于具线性趋势的时间数列。虽然一次指数平均在产生新的数列的时候考虑了所有的历史数据，但是仅仅考虑其静态值，即没有考虑时间序列当前的变化趋势。如果当前的数据处于某种趋势中，那么当我们对下一期数据进行预测的时候，好的预测值不仅仅是对历史数据进行“平均”，而且要考虑到当前数据变化的上升趋势。同时考虑历史平均和变化趋势，这便是二阶指数平均。

4. 三阶指数平滑

三阶指数平滑是二次平滑基础上的再平滑，它比一阶和二阶多考虑了一个因素：季节性效应。（Seasonality）。这种平均模型考虑的季节性效应在股票或者期货价格中都会比较常见，比如在过年前 A 股市场通常会交易比较频繁，在小麦成熟的时候小麦期货价格也会有比较明显的波动。但是，模型本身的复杂度也增加了其使用难度，我们需要一定的经验才能比较合理地设置其中复杂的参数。