

# R2-Learn 用户手册

## R2-Learn 用户手册

12 October 2018

### 目录

- 1. 概览
  - 1.1. 机器学习
  - 1.2. 使用 R2-Learn 进行机器学习
- 2. R2-Learn 入门
  - 2.1. 软件要求
  - 2.2. 导入数据
    - 2.2.1. 从数据库导入数据
    - 2.2.2. 导入本地文件
  - 2.3. 项目主页
  - 2.4. 模型部署主页
- 3. 开始一个新项目
  - 3.1. 创建一个项目
  - 3.2. 描述您的业务问题
  - 3.3. 处理您的数据
    - 3.3.1. 上传您的数据
    - 3.3.2. 清洗您的数据
      - 数据预览
      - 数据类型
    - 3.3.3. 目标变量
    - 3.3.4. 数据质量
      - 数据的常见问题
      - 自动修复目标变量
      - 手动修复目标变量
      - 所有数据质量
  - 3.4. 建模
    - 3.4.1. 自动建模
    - 3.4.2. 高级建模
    - 3.4.3. 建立您的模型
      - 模型选择
- 4. 部署模型

- 4.1. 部署
  - 4.1.1. 用数据源预测
  - 4.1.2. 用 API 预测
- 4.2. 监控已部署的模型
  - 4.2.1. 运行监控
  - 4.2.2. 性能设定
  - 4.2.3. 性能状态
- 附录 A: 数据质量修复
  - A.1. 修复异常值
  - A.2. 修复缺失值
- 附录 B: 高级建模
  - B.1. 高级变量设置
  - B.2. 高级模型设置
    - B.2.1. 默认设置下创建/编辑模型设置
    - B.2.2. 选择算法
    - B.2.3. 设置最大模型集成大小
    - B.2.4. 训练验证留出和交叉验证
    - B.2.5. 重采样设置
    - B.2.6. 设置度量指标
    - B.2.7. 设置最大优化时间
    - B.2.8. 随机种子
- 附录 C: 二分类问题的模型选择
  - C.1. 简化视图
  - C.2. 高级视图
    - C.2.1. 顶部
    - C.2.2. 模型表
    - C.2.3. 其他模型细节
- 附录 D: 回归问题的模型选择
  - D.1. 简化视图
  - D.2. 高级视图
    - D.2.1. 顶部
    - D.2.2. 模型表
    - D.2.3. 其他模型细节

## 1. 概览

R2-Learn 可帮助公司快速将数据转换为机器学习模型，而无需 AI（人工智能）专业知识。R2-Learn 采用最先进的技术，用 AI 全程助您使用您的数据创建模型，部署模型，实时迭代模型。

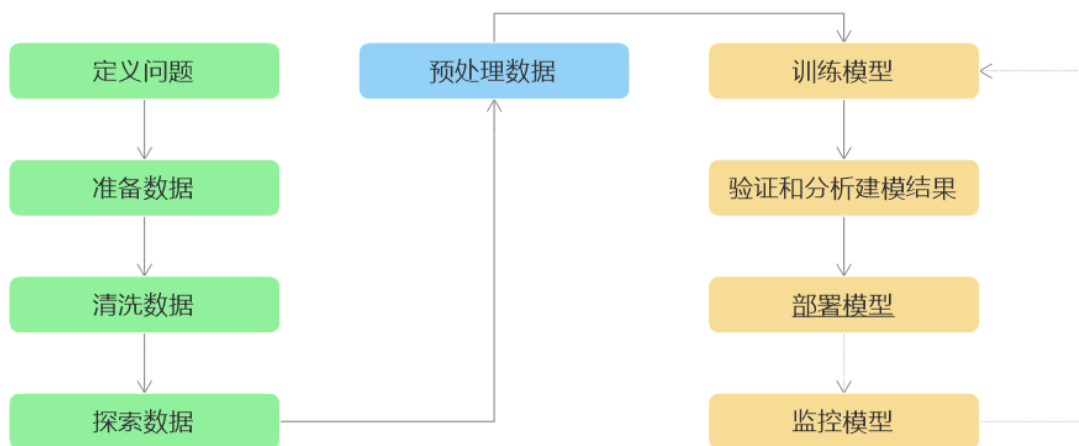
本节介绍机器学习的关键概念及其与 R2-Learn 的关系。

### 1.1. 机器学习

机器学习是 AI 的一个分支，其本质是训练计算机从数据中学习。随着算力的提升，机器学习可发现海量数据中深藏的模式和关系。目前正在开发机器学习模型，以对物体进行分类，检测异常，预测结果，并增加人的总体能力。现今机器学习模型已广泛地应用到物体分类，异常检测，预测结果，辅助增强人体能力等领域。

如何用机器学习辅助业务决策？常见工作流程如下：

1. **定义问题:** 决定要用数据解决的问题。
2. **准备数据:** 获取您要用于训练机器学习模型的数据集。
3. **清洗数据:** 检查数据集中是否存在数据缺失和数据错误，用合适的方法修复数据集，提升数据质量。
4. **探索数据:** 对数据有基本的了解，用于指导后续的分析 and 建模。
5. **预处理数据:** 修改数据集的格式，使其适合后续生成机器学习模型。
6. **训练模型:** 将数据用机器学习工具进行学习，生成模型。
7. **验证和分析建模结果:** 检查生成的模型是否能达到预期的预测效果。
8. **部署模型:** 模型达到可用标准后，将模型进行部署。模型部署通常会使用 REST API 端口进行部署。当您有新的数据希望预测时，您可调用此端口，它会为您返回预测结果。
9. **监控模型:** 随着时间的推移，机器学习模型的表现会受到影响，从而影响其预测准确性。因此，您需要监控其性能，来确保预测结果的表现。

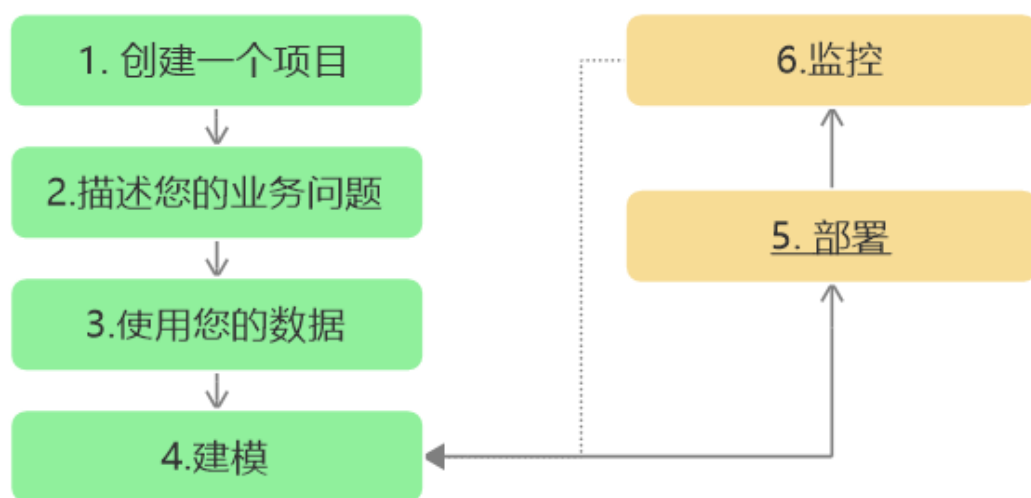


一个经验丰富的数据科学家会经历以上所有流程。这些流程中每一步和每一个关键点做出的决策对于机器学习模型的质量都至关重要，模型的质量进而决定了预测的表现。目前的数据科学家们要靠经验和直觉找到最佳的模型，这个过程可能会非常的繁琐和缓慢，耗费极大，且可能会产生缺陷。

## 1.2. 使用 R2-Learn 进行机器学习

R2-Learn 是一个用 AI 进行大规模，自动化，标准化机器学习建模的平台，使机器学习的业务落地更方便，快捷，准确。

R2-Learn 机器学习工作流程如下：



1. **创建一个项目**: 在 R2-Learn 中新建一个项目。
2. **描述您的业务问题**: 描述您想解决的业务问题, 帮助您理清项目目标和预期结果。
3. **使用您的数据**: 将您的训练数据集上传到 R2-Learn。R2-Learn 使用机器学习来帮助您检查和清洗数据集。
4. **建模**: 数据集加载至 R2-Learn 平台后, 您可选择以下方式进行建模:
  - **自动建模**: R2-Learn 使用训练集自动创建机器学习模型。
  - **高级建模**: 您可自行选择变量, 创建新变量, 选择算法等创建新模型。
5. **部署**: 机器学习模型创建后, R2-Learn 将自动在您的服务器上部署模型, 并允许您:
  - **使用数据源预测**: 将 R2-Learn 连接到您的数据库或上传 CSV 文件进行预测。
  - **使用 API 预测**: 使用 REST API 进行预测。

模型部署后, 您可监控模型的预测性能, 并在模型低于指定阈值时更新模型。

## 2. R2-Learn 产品入门

本章将带您创建您的第一个 R2-Learn 项目。要开始您的第一个项目, 请点击[新建项目](#)。

### 2.1. 软件要求

- 谷歌 Chrome 浏览器 65 或者更新版.

### 2.2. 导入数据

您可以将数据导入 R2-Learn 中, 可用于:

- **构建机器学习模型**, 以及
- 对导入的数据进行**预测**。

R2-Learn 支持以下导入数据的方法:

- **从数据库导入数据**
- **导入本地文件**




### 2.2.1. 从数据库导入数据

R2-Learn 支持从以下数据库导入数据:

- Oracle Database 11G
- Microsoft SQL Server
- MySQL

您的数据库必须支持 ODBC 11.2 链接。

Connect to Your SQL Database



Hostname:

e.g, db.abc.com

Port:

e.g, 12345

db-type

oracle

Database name:

Type in your database name

Table name:

Type in your table name

Database encoding (optional)

utf8

Username:

Type in your database username

Password:

Type in your database password

☐ Remember My Password

☒ Remember My Connection Profile

Connect

Cancel

### 2.2.2. 导入本地文件

您可导入 UTF-8 编码的 CSV 文件。该文件为训练集或输入数据。

您的 CSV 文件必须满足以下条件：

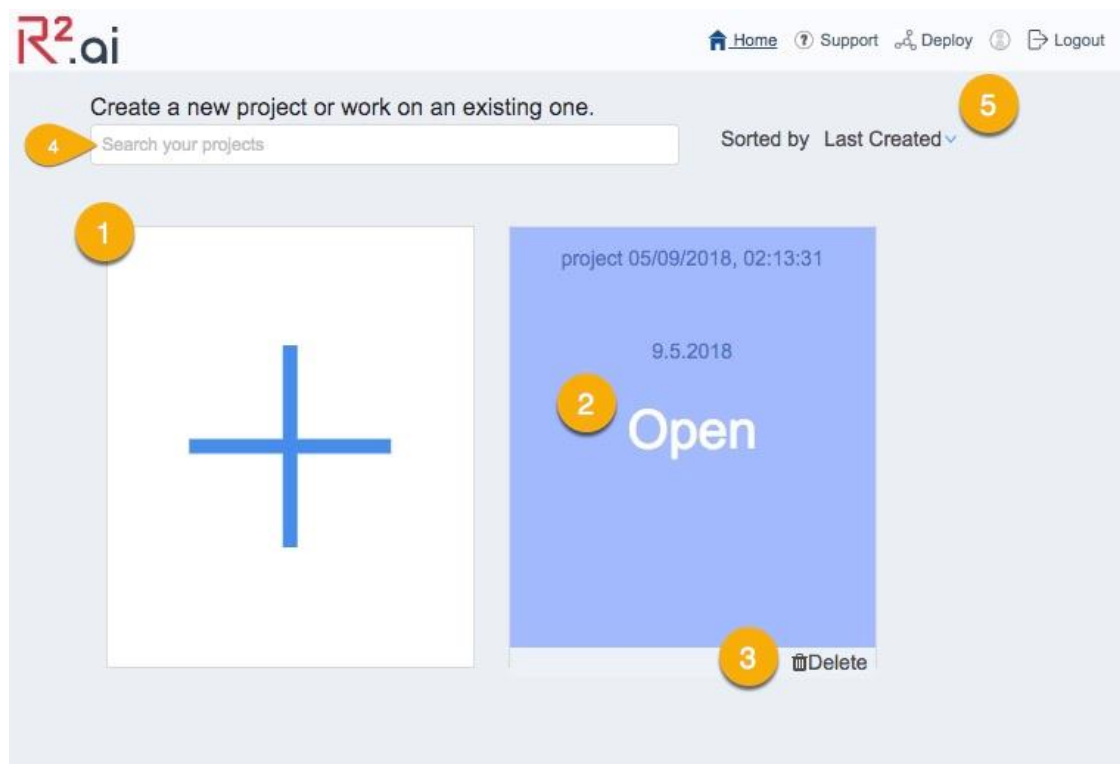
- 有一个标题行。
- 使用 UTF-8 编码。

**备注** 如果您导入一个用于模型部署的 CSV 文件，则该 CSV 文件需满足以下条件：

- 需包含已部署的模型所需的所有变量，以及
- 文件中的变量名需要与训练集中的变量名相同。

## 2.3. 项目主页

当您登录到 R2-Learn，呈现在您眼前的是项目主页：



在项目主页里，您可以：

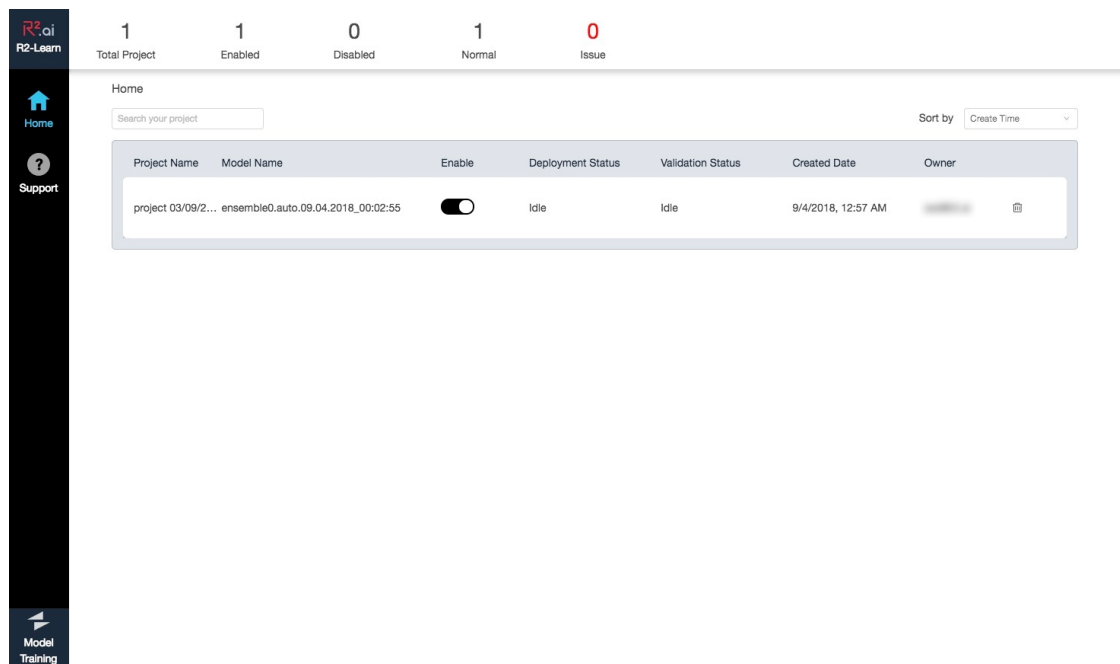
1. **创建** 新项目
2. **打开**和**编辑**已存在的项目

3. 删除一个项目
4. 查找一个项目
5. 对项目排序排序.

点击页面顶端的**部署**，您会进入模型部署主页。

## 2.4. 模型部署主页。

模型部署主页显示您创建的所有项目以及这些项目的部署/验证状态：



Summary:

Total Project	1	Enabled	1	Disabled	0	Normal	1	Issue	0
---------------	---	---------	---	----------	---	--------	---	-------	---

Home

Search your project

Sort by Create Time

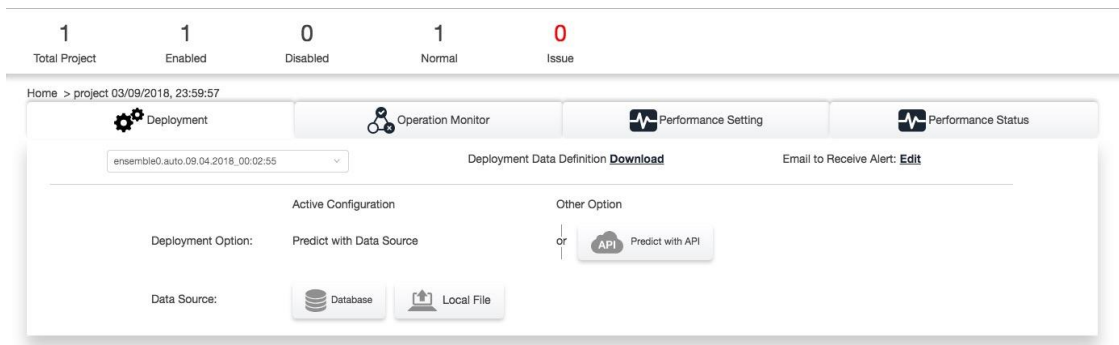
Project Name	Model Name	Enable	Deployment Status	Validation Status	Created Date	Owner
project 03/09/2...	ensemble0.auto.09.04.2018_00:02:55	<input checked="" type="checkbox"/>	Idle	Idle	9/4/2018, 12:57 AM	

**部署/验证状态：**显示部署或验证任务的当前状态。会有：

- **正在运行：**任务当前正在运行。.
- **空闲：**项目没有正在运行的任务。
- **问题：**项目在运行任务时遇到问题。请创建一个新案例。
- **已取消：**项目下的案例已取消。

单击项目可打开：





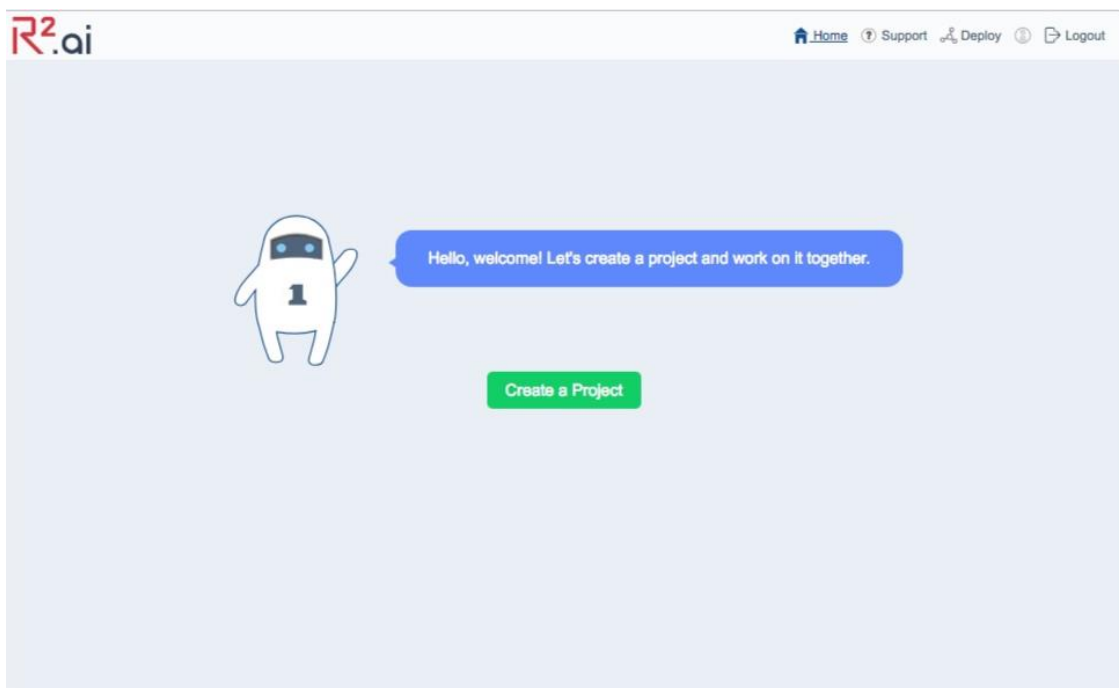
有关如何部署模型的更多信息，请参阅[部署模型](#)。

### 3. 开始一个新项目

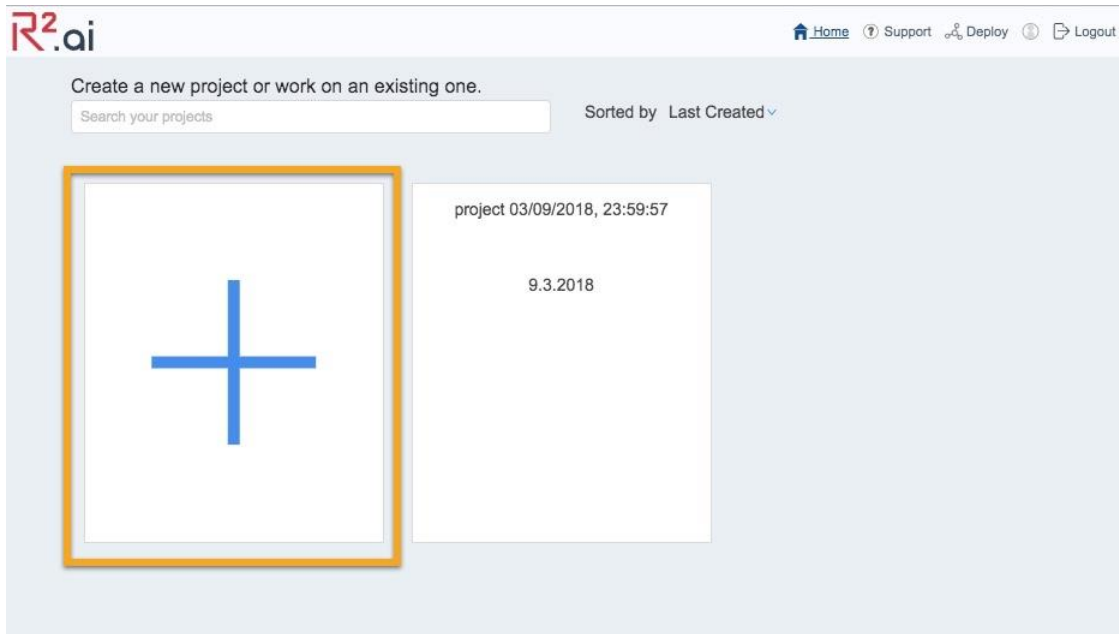
本章将向您介绍如何使用 R2-Learn 的自动建模功能建立一个新的机器学习项目。

#### 3.1. 创建一个项目

当您首次登陆 R2-Learn 时，您将看到一个空的项目主页。单击“**创建项目**”以启动新项目。



如果您在 R2-Learn 上有现有项目，请单击加号以启动新项目。



您会进入项目部分，R2-Learn 会指引您：

- **命名您的项目。**每个项目都有一个默认的名称，其命名格式为《日/月/年，时：分，秒》，您可按需修改项目名称。
- **描述您的项目（可选）。**请输入您的简要项目说明。这会帮助您明确该项目的目的。项目细节可记录在接下来的**业务问题**中的**问题描述**部分。

完成后，请点击**继续**按钮。

### 3.2. 描述您的业务问题

在**业务问题**部分，您将记录该项目的问题陈述和问题类型。

您可在项目的问题描述（可选）纷纷记录问题陈述和业务价值。这些记录有助于相关部门的持续跟踪和评估。

在**问题类型**中，您可选择希望项目模型预测的类型：

- **对错（二分类）：**项目建立的模型将用于预测时间是否会发生。例如客户是否会购买该产品。
- **连续值（回归）：**预测连续值/数值。例如根据给定的变量，预测转化一个客户的成本是多少。

一旦您设置好您的**业务问题**和**问题类型**后，点击**继续**进入下一步，将您的数据上传到 R2-Learn 上。

### 3.3. 处理您的数据

现在我们开始处理您的数据，**数据**部分将引导您一起完成：

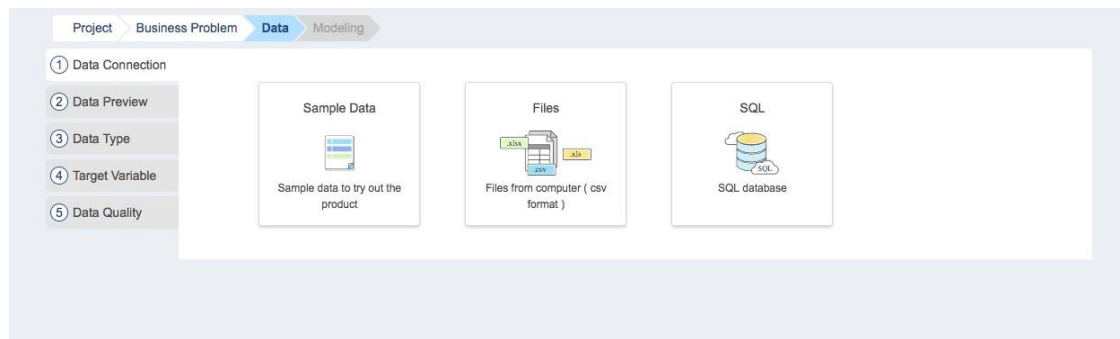
1. 首先在**数据连接**中**上传您的数据**。
2. 根据以下步骤**清洗您的数据**：
  - a. **数据预览**，检查数据的标题行。
  - b. **数据类型**, R2-Learn 会为您自动判断每列数据的数据类型。
3. 接下来，您可给 R2-Learn 指定您希望机器学习模型预测的**目标变量**。
4. **数据质量**：

高质量的训练集对机器学习是至关重要的。您的训练集需包含尽量多的与目标相关的信息。它应只包含与**业务问题**相关的信息，且样本量尽可能大。您可请专家确认训练集的质量。R2-Learn 可用高质量的数据集完成机器学习其他所有工作。

当您上传数据后，R2-Learn 会检测数据类型，检查数据质量，并自动修复问题或提醒您手动修复。

#### 3.3.1. 上传您的数据

您可在**数据连接**选项中上传数据集。“数据连接”选项允许您选择要用于机器学习模型的数据集。



如上图所示，您可将训练集加载至 R2-Learn 中。您可以用以下方式完成：

- 连接**数据库**，
- 上传**本地文件**。

若您刚开始了解 R2-Learn，您可使用 R2-Learn 提供的训练数据集。

**重要**            您上传的数据集必须包含标题行。

### 3.3.2. 清洗您的数据

#### 数据预览

训练集加载完成后，R2-Learn 会在[数据预览](#)中随机展示训练集中的数据供您检查数据质量。您可按需进行以下修改：

- 编辑标题行
- 更改 R2-Learn 用于机器学习建模的最大样本数。

**重要** 您必须确认每一列的列名都是**唯一的**。

Project Business Problem **Data** Modeling

① Data Connection Please edit the default header row (1st row) if necessary. If your data doesn't have a header, please prepare a dataset that has one.

② Data Preview ① Maximum Data Sampling Size: 100,000,000 rows [Edit](#)

1	2	3	4	5	6	7	8	9	10
age	job	marital	education	default	balance	housing	loan	contact	day
300	unemployed	married	primary	no	1787	no	no	cellular	19
	services	married	secondary	no	4789	yes	yes	cellular	11
35		single	tertiary	no	1476	yes	no	cellular	16
30	management	married	tertiary	no	1476	yes	yes	unknown	11
59	blue-collar	married	secondary	no	3	yes	no	unknown	5
35	management	single	tertiary	no	747	12	no	cellular	23
36	self-employed	married	tertiary	no	307	yes	no	cellular	14
39	technician	married	secondary	no	147	yes	no	cellular	6
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14

< 1 2 > Goto

[Continue](#)

完成以上操作后，请单击**继续**，会进入[数据类型](#)。

#### 数据类型

您需要确认 R2-Learn 为训练集中每列数据自动检测的**数据类型**。若某列数据类型有误，请从该列的下拉菜单中选择正确的数据类型。

Project Business Problem **Data** Modeling

☒ Data Connection  
☒ Data Preview  
☒ **Data Type**  
☐ Target Variable  
☐ Data Quality

Please verify the automatically inferred data types and make necessary changes. ⓘ

1	2	3	4	5	6	7	8	9	10
age	job	marital	education	default	balance	housing	loan	contact	day
Numerical ▾	Categorical ▾	Categorical ▾	Categorical ▾	Categorical ▾	Numerical ▾	Categorical ▾	Categorical ▾	Categorical ▾	Numerical ▾
300	unemployed	married	primary	no	1787	no	no	cellular	19
	services	married	secondary	no	4789	yes	yes	cellular	11
35		single	tertiary	no	1476	yes	no	cellular	16
30	management	married	tertiary	no	1476	yes	yes	unknown	11
56	blue-collar	married	secondary	no	3	yes	no	unknown	5
35	management	single	tertiary	no	747	12	no	cellular	23
36	self-employed	married	tertiary	no	307	yes	no	cellular	14

< 1 2 > Goto

Continue

每列有 2 种可能的数据类型：

- **数值型**: 当该列的数据为一系列数值时。
- **分类型**: 当该列中的数据为不同的类别时。

重要 数据类型的精确识别会影响机器学习模型。当您不确定某列数据的类别时，请咨询数据集的提供者。

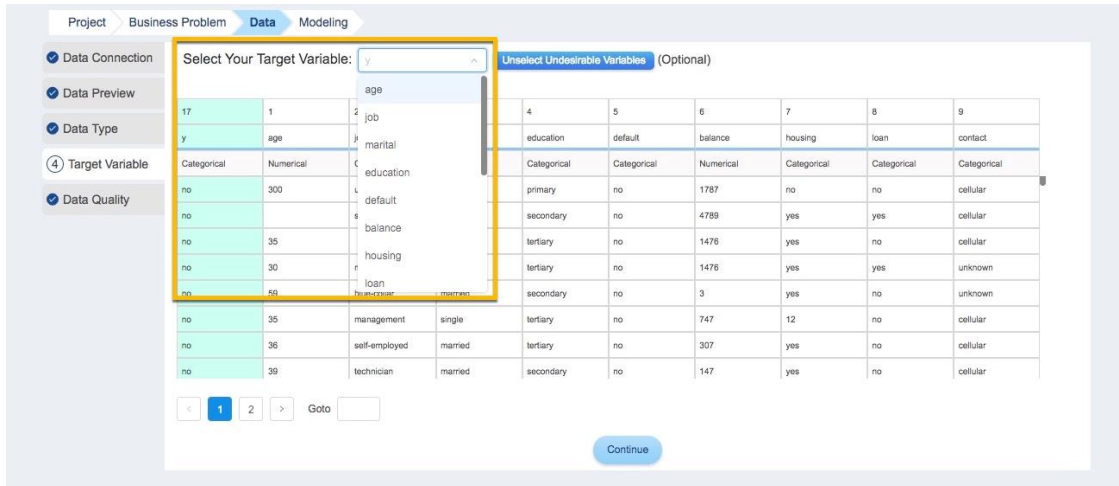
R2-Learn 会根据您上传数据方式的不同，用不同的方式识别数据类型：

- 当您从[数据库](#)中上传数据时，R2-Learn 将展示数据库中提供的数据类型。
- 当您从[本地](#)上传数据时，R2-Learn 会自动推断每列的数据。

完成以上操作后，单击**继续**，进入[目标变量](#)。

### 3.3.3. 目标变量

在**目标变量**中，我们需要选择我们希望机器学习模型预测的变量。目标变量的数据类型应与您在[业务问题](#)模块中设置的**问题类型**相匹配。



单击**选择目标变量**的下拉菜单，选择变量。

您可单击**反选不需要的变量**，该操作用于移除对目标变量没有贡献的变量，和与目标变量拥有一对一映射关系的变量，例如：

- ID
- 人名
- 产品名
- 目标变量的衍生变量
- 直接生成目标变量的变量

例如，在预测一个人的年收入的数据集中，我们可以直接移除人名，月收入。被移除的变量将变暗，并在后续模型训练中被移除。

完成以上操作后，单击**继续**，检查**数据质量**。

### 3.3.4. 数据质量

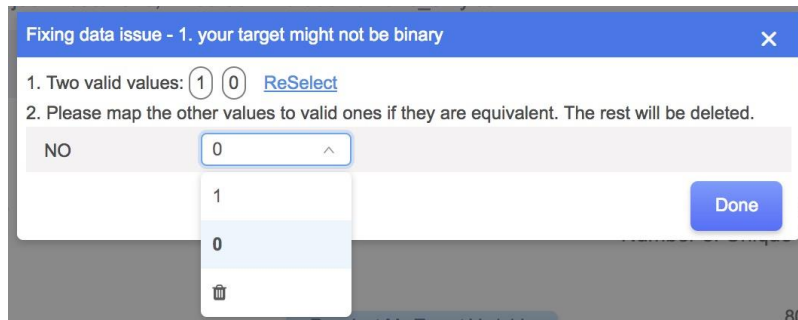
在**数据质量**中，R2-Learn 会检测数据中影响建模的问题，它会从以下两个方面检查数据质量：

- **目标变量质量**: R2-Learn 会分析处理并显示目标变量的问题。
- **所有数据质量**: 目标变量质量问题修复后，R2-Learn 将会检查数据集的其他部分是否存在问题。

### 数据的常见问题

R2-Learn 会识别和修复以下两种问题：

- **目标变量有 2 个以上唯一值:** 在二分类业务问题中，目标变量必须有且只有 2 个不同的唯一值（通常为“是”和“否”）。



- **数据类型错配:** 当变量的数据类型与之前设置的数据类型不匹配时，会有数据类型错配警告。例如，某列标记为数值型的数据里有一些文本值。
- **缺失值:** 产生缺失值的原因既可能是数据未搜集，也可能是该数据点包含空值。

注意          空值表示缺少数据，没有数据存在，与 0 值不同。

- **异常值:** 当某列为连续值时，若某数据点超出预期范围，会被认为是异常值。检查异常值是不良的数据点还是真正的偏差非常重要。

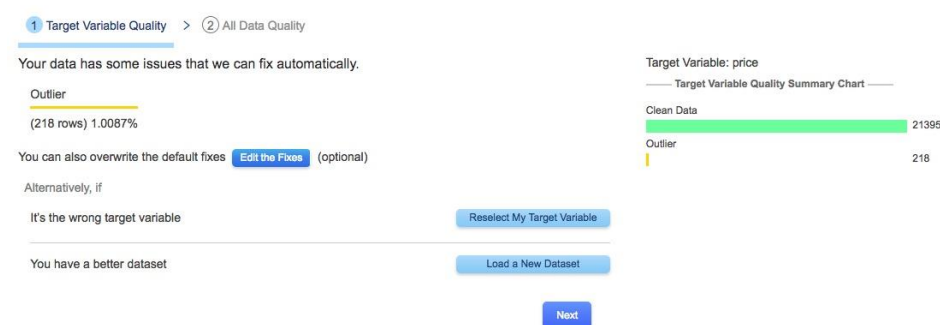
以上这些问题的可修复方法，请参阅[附录：数据质量修复](#)。

## 目标变量质量

R2-Learn 会识别并修正目标变量中存在的问题。

## 自动修复目标变量

当 R2-Learn 发现目标变量中存在的问题是，它会这些问题展示在**数据质量**中：



单击下一步，R2-Learn 会自动修复问题，或：

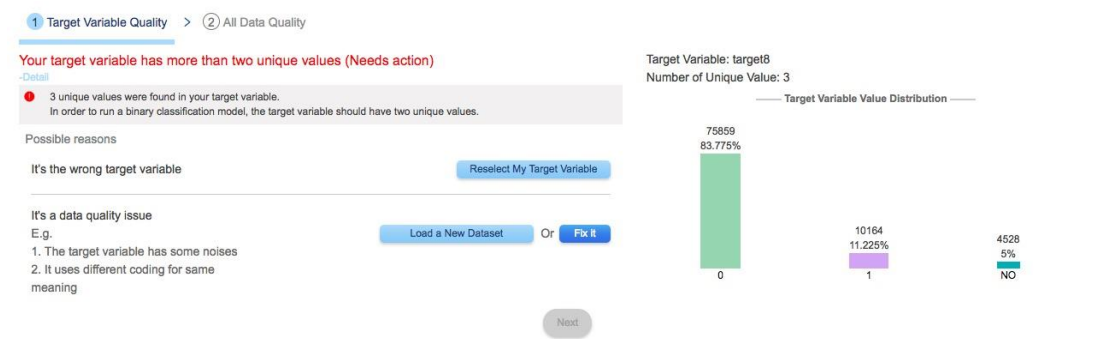
- 单击**编辑修复**，可修改 R2-Learn 修改目标变量中存在问题的方法。具体方法详见[附录：数据质量修复](#)。

- 单击**重新选择目标变量**，将会返回目标变量选项中，您可选择新的目标变量。
- 单击**上传新数据集**，将返回至**数据连接**，您可选择要使用的新数据集（要取消此操作，请单击**数据质量**即可返回至当前页面）。

完成目标变量修复问题后，单击**下一步**继续所有**数据质量**部分。

### 手动修复目标变量

当 R2-Learn 在目标变量中遇到无法自动修复的问题时，它会在数据质量中将这些问题标记为需要操作，并在您进行下一步之前提示您进行修复。



如上图所示，您会被提示用以下方法修复目标变量。

- 单击**重新选择我的目标变量**，可返回目标变量并选择新的目标变量。
- 单击**上传新数据集**，可返回数据连接，然后选择要使用的新数据集。（取消加载新数据集，请单击**数据质量**）
- 单击**修复**，可打开一个对话框，修复目标变量。

以上完成后，单击**下一步**，进入所有**数据质量**。

### 所有数据质量

目标变量质量问题解决后，R2-Learn 会继续检查其余所有变量的质量问题，并在此处显示结果。

R2-Learn 以提供默认修复方法。当您要自定义修复方法时，可进行以下操作：

- 单击**加载新数据集**可返回至**数据连接**，您可选择要是使用的新数据集。（要取消加载新数据集，请单击**数据质量**）。



- 选择**编辑修复**，会打开一个对话框，您可以自行选择修复数据的方式。

How R2-Learn Will Fix the Issues

Data Table

Missing Value

Variable Name	Why Missing?	Data Type	Quantity of Missing Value	Mean	Median	Most Frequent	Fix
age	I don't know	Numerical	1 (0.02%)	42.11	39.00	N/A	replace with mean value (Default)
job	Left blank on purpose		1 (0.02%)	N/A	N/A	management	replace with most frequent value (De...
	Failed to Collect or Data Error						
	I don't know						

Done

Cancel

单击**数据表**，可查看数据中遇到的问题样例。

How R2-Learn Will Fix the Issues

Data Table

Data Type Mismatch

Missing Value

age	job	marital	education	default	balance	housing	loan	contact	day
Numerical	Categorical	Categorical	Categorical	Categorical	Numerical	Categorical	Categorical	Categorical	Numerical
0.02%	0.02%								
	services	married	secondary	no	4,789	yes	yes	cellular	11
35		single	tertiary	no	1,476	yes	no	cellular	16
300	unemployed	married	primary	no	1,787	no	no	cellular	19
30	management	married	tertiary	no	1,476	yes	yes	unknown	11
59	blue-collar	married	secondary	no	3	yes	no	unknown	5
35	management	single	tertiary	no	747	12	no	cellular	23
36	self-employed	married	tertiary	no	307	yes	no	cellular	14
39	technician	married	secondary	no	147	yes	no	cellular	6
41	entrepreneur	married	tertiary	no	221	yes	no	unknown	14
43	services	married	primary	no	-88	yes	yes	cellular	17

Done

Cancel

关于常见修复方法，请参阅[附录：数据质量修复](#)。

注意 保存对数据的更改可能会花 1-2 分钟，请耐心等待。

以上完成后，请单击**继续**，开始**建模**。

### 3.4. 建模

- 自动建模，或
- 高级建模

#### 3.4.1. 自动建模

**自动建模**会根据您处理过得训练集，自动为您构建机器学习模型。单击**自动建模**，R2-Learn 将为您构建模型。

#### 3.4.2. 高级建模

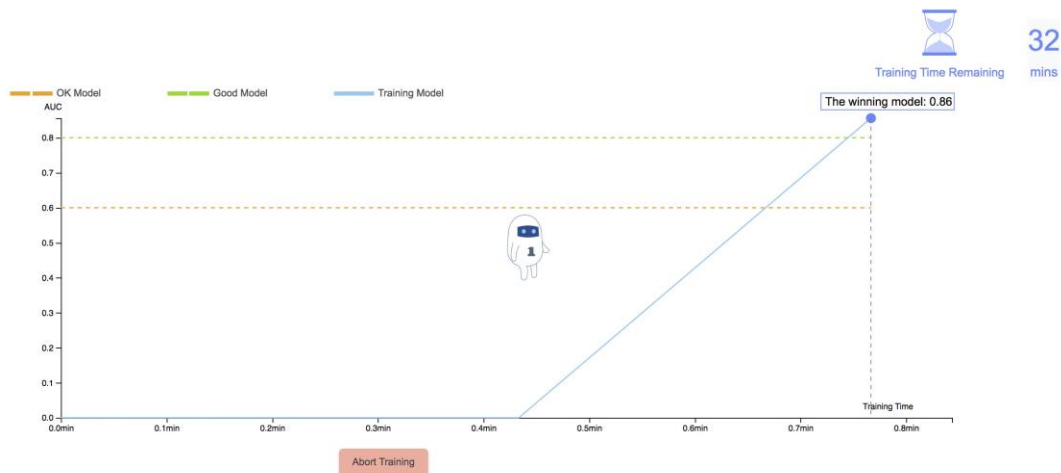
当您选择**高级建模**时，您可以更详细的控制建模过程。具体信息请参阅[附录：高级建模](#)。

### 3.4.3. 建立您的模型

完成上述所有建模前的配置，R2-Learn 将开始建造您的机器学习模型。

R2-Learn 将在建模过程中向您显示性能时间图：

- **二分类模型**：随着训练时间的递增，AUC(曲线下面积)的变化：
  - **"OK Model" 行**: 超过此行，表明机器学习模型已经超过了通常可接受的性能水平
  - **"Good Model" 行**: 超过此行，表明机器学习模型的已经拥有了良好的性能水平。
  - **"Training model" 行**: 这条线显示了目前性能最佳的机器学习模型。
- **回归模型**: 随着训练时间地增，MSE（均方误差）的变化。
  - **"Training Model" 行**: 此行显示了目前性能最佳的机器学习模型的表现。



单击中止训练，可取消魔性训练。

模型训练所需的剩余时间在页面最右侧。训练完成时，请给 R2-Learn 一些时间构建最终模型，并生成模型训练报告。

如果 R2-Learn 无法从给定的训练数据集中训练出足够高效的机器学习模型，则模型训练过程可能会失败。如果模型训练失败，则必须重新配置项目以修复数据集质量问题，或选择新的或更大的数据集。

### 模型选择

模型训练完成后，在**模型选择**部分，R2-Learn 会展示模型训练结果。根据您的不同的**问题陈述**，您可看到不同的图表：

- [附录：二分类问题的模型选择](#)
- [附录：回归问题的模型选择](#)

模型选择后，单击部署，进行[模型部署](#)。

## 4. 部署模型

通过部署模型，您可以对新的数据进行预测。这经常涉及到 REST API 端口。您可以将数据发送到该端口，并从中接收预测。

R2-Learn 为您自动处理模型部署。一旦您成功地创建了机器学习模型，您可以点击其中包含的项目来访问已部署的模型：

Home

Search your project Sort by Create Time

Project Name	Model Name	Enable	Deployment Status	Validation Status	Created Date	Owner
project 11/09/...	Ridge1.auto.09.11.2018_18:05:24	<input checked="" type="checkbox"/>	Idle	Idle	9/11/2018, 8:40 PM	
project 10/09/...	GradientBoostingClassifier5.auto.09.11...	<input checked="" type="checkbox"/>	Idle	Idle	9/11/2018, 6:04 PM	
project 05/09/...	ensemble0.auto.09.05.2018_04:43:28	<input checked="" type="checkbox"/>	Idle	Idle	9/5/2018, 10:43 AM	

在开放的项目中，您可看到以下部分：

- [部署](#)
- [操作监控](#)
- [性能设置](#)
- [表现状态](#)

此外，您可访问以下：

Home > project 11/09/2018, 18:04:13

Deployment	Operation Monitor	Performance Setting	Performance Status
<div>Ridge1.auto.09.11.2018_18:05:24</div> <div>Deployment Data Definition <a href="#">Download</a> Email to Receive Alert: <a href="#">Edit</a></div>			

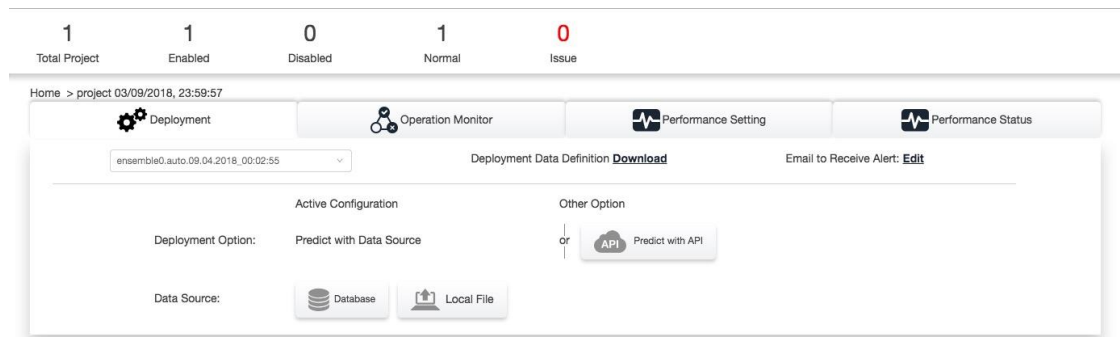
- **正在使用的模型：**选择要运行部署或其他操作的模型。
- **部署数据定义：**单击“下载”，可以下载包含模型使用的变量列表的 CSV 文件。
- **电子邮件接收警报：**单击可以输入电子邮件地址，可发送与部署相关的警报。

## 4.1. 部署

当您在[模型部署主页](#)打开一个项目时，您会进入模型部署。您可以选择使用已部署模型的方式：

- **使用数据源预测**: 使用本地文件或数据库中的数据进行预测。
- **使用 API 预测**: 使用 R2-Learn 的 REST API 将数据发送到 R2-Learn 进行预测。

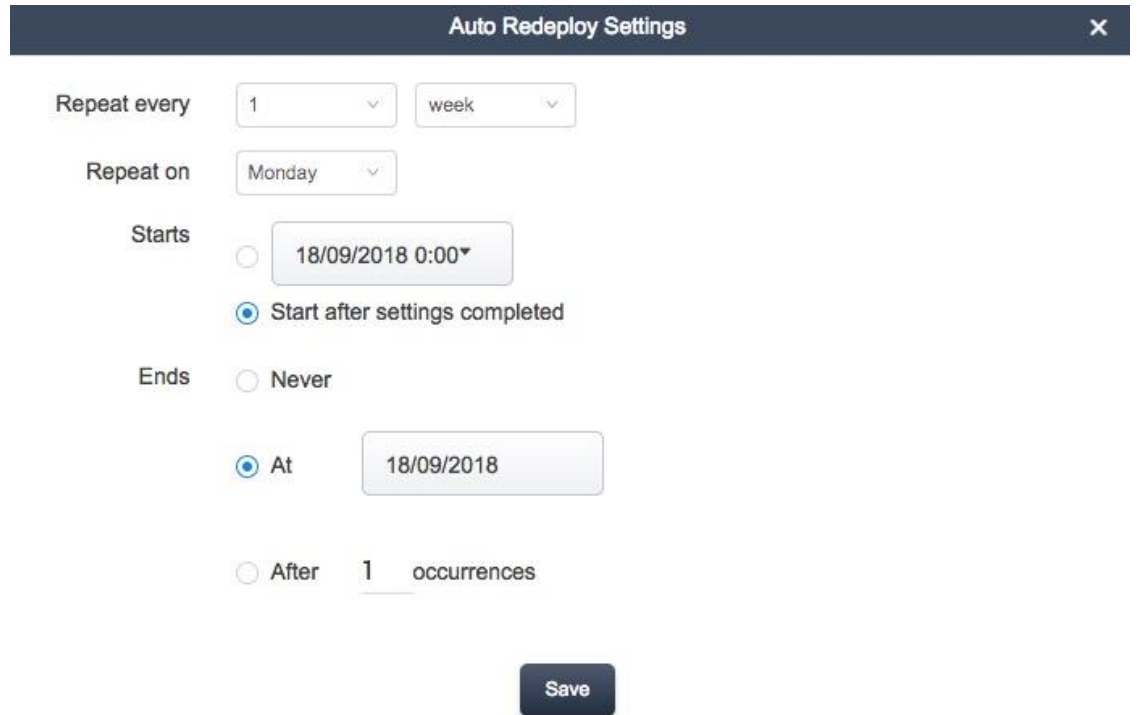
重要 导入的数据必须与所使用的机器学习模型具有相同的变量。要下载包含导入数据集中所需变量列表的文件，请单击[部署数据定义](#)旁边的[下载](#)。



### 4.1.1. 用数据源预测

要对从数据源导入的数据进行预测：

1. 单击**使用数据源预测**。
2. 选择要从以下位置导入数据的数据源：
  - **数据库**: 请提供链接的详细信息。请参阅[从数据库导入数据](#)。
  - **本地文件**: 将文件上载到 R2-Learn，其中包含要在其上运行预测的数据。有关更多信息，请参阅[导入本地文件](#)。
3. 选择**结果位置**，模型预测的结果会保存在结果位置上：
  - **在 App 中**: 在 R2-Learn 中保存并显示结果。
  - **上传到数据库**: 将结果写入给定数据库。
4. 选择**部署频率**。R2 会根据此频率进行部署：这告诉 R2-Learn 如何安排部署：
  - **一次性**
  - **自动重复**: 请为您的部署设置一个定期重复计划。

The image shows a 'Auto Redeploy Settings' dialog box with a dark blue header and a close button (X) in the top right corner. The settings are as follows: 'Repeat every' is set to '1' with a dropdown arrow, and the unit is 'week'; 'Repeat on' is set to 'Monday' with a dropdown arrow; 'Starts' has two options: a date/time picker set to '18/09/2018 0:00' (which is unselected) and a radio button labeled 'Start after settings completed' (which is selected); 'Ends' has two options: a radio button labeled 'Never' (which is unselected) and a radio button labeled 'At' (which is selected) followed by a date picker set to '18/09/2018'; there is also an unselected radio button labeled 'After' followed by a text input set to '1' and the word 'occurrences'. A 'Save' button is located at the bottom right of the dialog.

- **故障时自动禁用:** 若 R2-Learn 出现故障时, 请选择此项, 会停止部署。

#### 4.1.2. 用 API 预测

在 R2-Learn 中部署机器学习模型时, 会自动生成 REST API 端点, 您可据此发出请求。

使用 R2-Learn 的 REST API 进行预测前, 您需要有:

- R2-Learn API KEY;
- 您的 R2-Learn 用户名;
- 您的 R2-Learn 项目名; 例如, 项目 `project 11/09/2018, 18:04:13`。
- 您的 R2-Learn 模型名称。例如, `Ridge1.auto.09.11.2018_18:05:24`。

使用带有 cURL 的 R2-Learn 的 REST API 进行预测, 并将结果保存到 `output.csv` 时, 请在终端运行以下命令:

```
curl -F 'data=@path/to/local/file' http://service2.newa-tech.com/api/user_name/project_name/model_name/api_key -o output.csv
```

您需要项目名称和模型名称的 URL 编码版本, 并将其写入请求中。例如, 项目 `11/09/2018, 18:04:13` 的 URL 编码字符串为 `project%2011%2F09%2F2018%2C%2018%3A04%3A13`, 模型 `Ridge1.auto.09.11.2018_18:05:24` 的 URL 编码字符串为 `Ridge1.auto.09.11.2018_18%3A05%3A24`。

您生成的 cURL 请求格式应如下所示：

```
curl -F 'data=@/home/r2user/input_data.csv' http://service2.newa-tech.com/api/r2user/project%2011%2F09%2F2018%2C%2018%3A04%3A13/Ridge1.auto.09.11.2018_18%3A05%3A24/apikey912ec803b2ce49e4a -o /home/r2user/output.csv
```

您也可用 Python 写 API 请求：

```
# API code
import requests
# Make predictions on your data
data = {"file": open('/path/to/your/data.csv', 'rb')}
Response = requests.post('http://service2.newa-tech.com/api/<user_name>/<project_name>/<model_name>/<api_key>', data)
```

## 4.2. 监控已部署的模型

您可使用以下功能监控和调整已部署的模型：

- [运行监控](#)
- [性能设定](#)
- [性能状态](#)

### 4.2.1. 运行监控

此页面展示了所有与模型有关的运行信息。对于每次运行，都会显示以下信息：

- **模型名称:** 运营模型名称；
- **开始时间:** 运行开始时间；
- **结束时间:** 运行结束时间；
- **部署方式:** R2-Learn 如何连接到输入数据：通过数据库连接，或上传本地文件，或 API 请求；
- **状态:** 运行状态；
- **结果:** 运行结果。

此外，每个操作都有以下选项：

- **下载结果:** 操作完成后，您可以下载结果数据。届过数据中附加了额外的列，其名字为"<target\_variable\_name>\_pred"。它是对每个数据点的预测值。
- **取消正在进行的操作:** 中止运行。

#### 4.2.2. 性能设定

您可以上载验证数据集来验证该模型的性能。

您可用以下方式上传验证数据集：

- [从数据库导入数据](#)
- [导入本地文件](#)

验证指标有以下几种：

- **度量指标:** 用验证数据集运行模型时，您可使用度量指标来衡量模型的表现
  - 二分类模型：选择精度或 AUC 作为度量指标。
  - 回归模型：选择 RMSE 或  $R^2$  作为度量指标。
- **标准阈值:** 您可为已部署数据的性能设置一个标准阈值。若验证模型的性能低于此阈值，则 R2-Learn 会发出警报。
- **部署频率：** 您可以在此处安排重复操作。
- **故障时自动禁用:** 启用此选项可将操作设置为在遇到任何问题时终止。

#### 4.2.3. 性能状态

您可在此处监控任何正在运行或已完成的操作的状态。此处显示以下信息：

- **模型名称:** 运营模型名称；
- **开始时间:** 运行开始时间；
- **结束时间:** 运行结束时间；
- **性能:**
  - 回归模型：显示模型的  $R^2$  和 RMSE.
  - 二分类模型：显示 AUC 和精度
- **阈值:** 您在[性能设定](#)里设置的阈值
- **结果 Results:** 操作完成后，您可以下载结果数据。届过数据中附加了额外的列，其名字为"<target\_variable\_name>\_pred"。它是对每个数据点的预测值。

## 附录 A: 数据质量修复

R2-Learn 可能标记的数据质量问题包括：

- **数据类型错配:** 当变量的数据类型与之前标记的数据类型不匹配时，它会发出数据类型不匹配警告。例如，如果之前标记该变量为分类变量，但其实际为数字变量。
- **缺失值:** 数据中的缺失值可能是由于数据收集时的错误或空值。

注意            空值表示缺少数据（“无值”），与“零值”不同。

- **异常值:** 数据点超出给定数据集的预期范围。检查异常值的数据点并确定它们是否与数据或坏数据点中显示的趋势完全不同是很重要的。

在数据质量中，R2-Learn 可帮助您使用以下方法解决这些问题：

- [修复异常值](#)
- [修复缺失值](#)

### A.1. 修复异常值

- **编辑有效范围**
  - 使用我们的规则引擎，系统会自动的从数据集中推断出变量的合理范围，并将范围外的值视为异常值。手动编辑有效范围，可以扩展或收缩 R2-Learn 接受的值范围。
- **边界值替换**
  - 用边界值替换异常值。
  - 仅适用于数值型变量。
  - 例如，如果有一列的值为[1, 50, 60, 70, 1000]，其合理边界为（40, 80），那么**边界值替换**会将异常值[1] 和[1000]替换为[40]和[80],则这列的值变成了[40, 50, 60, 70, 80]。
- **保留**
  - 保留异常值。
  - 仅适用于数值变量。
- **均值替换**
  - 将异常值替换为该列中有效范围内所有值的平均值。
  - 仅适用于数值变量。



- 例如，如果有一列的值为[1,50,60,70,1000]，其有效范围为（40,80），则异常值[1]和[1000]将替换为值[60]，则这列的值变成了[60,50,60,70,60]。
- **中值替换**
  - 将异常值替换为该列中有效范围内所有值的中位数值。
  - 仅适用于数值变量。
  - 例如，如果有一列的值为[1,6,5,8,50]，其有效范围为（3,10），则异常值[1]和[50]将替换为值[6]，则这列的值变成了[6,6,5,8,6]。
- **删除行**
  - 删除包含异常值的行。
- **0 值替换**
  - 将异常值替换为 0。
  - 仅适用于数值变量。

## A.2. 修复缺失值

- **替换最常见的值**
  - 仅适用于分类变量。
  - 例如，如果有一列的值为[1,2,5,3,2]，则所有缺失值将设置为类别 2。
- **替换为新的唯一值**
  - 用一个全新的唯一值作为一个新类别，来替换缺失值。
  - 仅适用于分类变量。
- **平均值替换**
  - 将缺失值替换为列中剩余值的平均值。
  - 仅适用于数值变量。
  - 例如，如果有一列具有以下非缺失值[4,6,8]，则所有缺失值将替换为值 6。
- **中值替换**
  - 将缺失值替换为列中剩余值的中位数。

- 仅适用于数值变量。
  - 例如，如果有一列具有以下非缺失值[6,5,8]，则所有缺失值将替换为值 6。
- **删除行**
  - 删除包含缺失值的行。
- **0 值替换**
  - 用 0 替换缺失值。
  - 仅适用于数值变量。
- **为什么缺失？**
  - 您可以要求 **R2-Learn** 为您修复缺失的值。选择以下三个选项之一来告知 **R2-Learn** 缺少值的原因，让 **R2-Learn** 自动修复缺失值：
    - “我不知道”
    - “故意留空”
    - “无法收集数据”

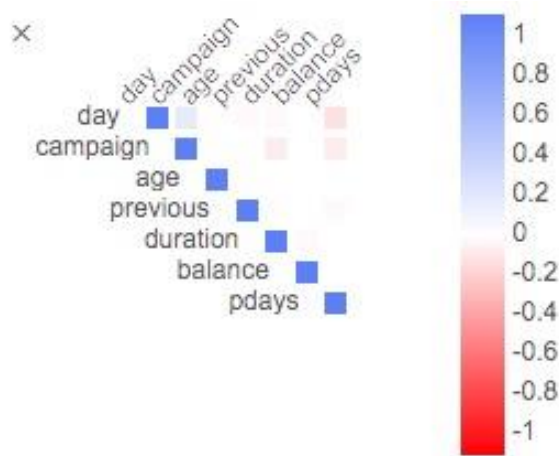
## 附录 B: 高级建模

当您选择**高级建模**时，可以配置以下设置：

### B.1.高级变量设置

这里您可以修改数据中的变量以更改模型。您可以在**高级变量设置**中更改以下内容：

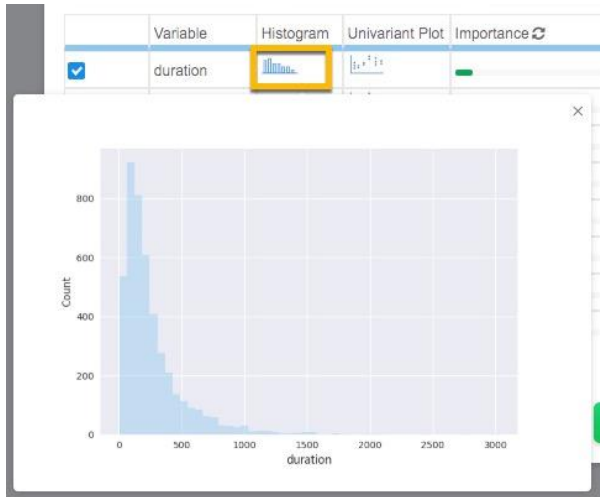
- **选择/取消选择变量：**您可以选择其他变量或从数据集中删除变量。构建机器学习模型后，您可以在[模型选择](#)部分中查看变量选择的更改如何影响模型。
- **检查关联矩阵：**单击**检查关联矩阵**以显示一个小图，该图显示数据集中具有最强相关关系的 15 个变量之间的相关强度。相关性越强，颜色越冷（更接近蓝色）。冷色（蓝色阴影）表示正相关，而暖色（红色阴影）表示负相关。



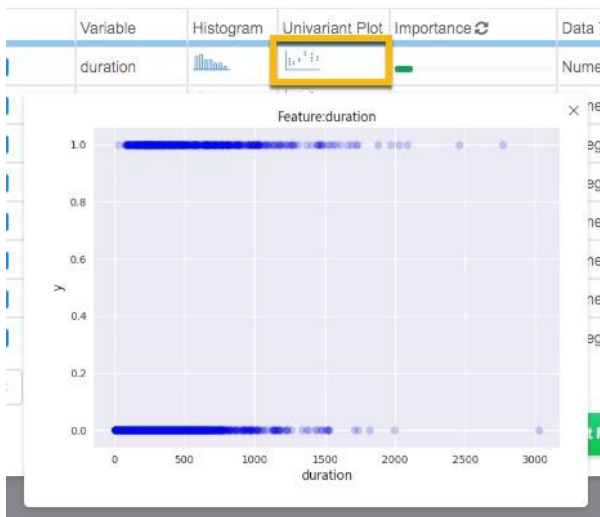
- **创建新变量：**单击**创建新变量**向模型添加新变量。此新变量可以是现有变量的组合，也可以是对现有变量执行的操作。要创建新变量：
  - a. 单击**创建新变量**。
  - b. 输入新变量的唯一名称。
  - c. 单击等号（“=”）后面的字段，打开列出所有可能操作的下拉菜单。您可以对现有变量执行以下操作以创建新变量：
    - 分组和计算
    - 组合变量
    - 比较值
    - 累积值
    - `log()`
    - 其他运算: +, -, \*, /

**注意** 有关这些操作的详细信息，请在选择操作时单击“**查看提示**”。

- **直方图/条形图：**单击变量的“直方图”会显示一个图形，显示变量的所有值的分布。数值变量显示为直方图，分类变量显示为条形图。



- **单变量图：**单击变量的“单变量图”会显示一个图形，显示该变量值的分布。



- **变量重要性：**变量重要性部分表示目标变量对预测变量的统计学显著性;变量的“重要性”越高，该变量的变化就越会影响预测的目标变量。默认情况下，高级变量设置中显示的变量从最高到最低“重要性”排序。

## B.2. 高级建模设置

在这里，您可以为模型构建配置其他设置：

### B.2.1. 默认状态下创建/编辑模型

单击**高级建模**时，R2-Learn 会创建一个新的“模型设置”。在**高级建模设置**中，新模型设置为：

1. 默认名称"custom.<MM.DD.YYY\_HH:MM:SS>"
2. 将{select}区域设定为此新模型的设置。

您可在{select}字段中选择它来构建先前定义的“模型设置”。在“命名模型设置”字段中更改名称，将修改后的模型设置另存为新模型设置。

### B.2.2. 选择算法

R2-Learn 默认使用以下所有算法来构建模型。您可以删除一些算法，加快建模速度。以下是针对不同问题可选择的算法列表：

算法	二分类	回归
Adaboost	✓	✓
ARD (Automatic Relevance Determination) Regression		✓
Bernoulli Naive Bayes	✓	
Decision Tree	✓	✓
Extra Trees	✓	✓
Gaussian Naive Bayes	✓	
Gaussian Process		✓
Gradient Boosting	✓	✓
K Nearest Neighbor	✓	✓
Linear Discriminant Analysis	✓	
Multinomial Naive Bayes	✓	
Passive Aggressive	✓	
Quadratic Discriminant Analysis	✓	
Random Forest	✓	✓
SGD (Stochastic Gradient Descent)	✓	✓
Logistic Regression	✓	
Ridge Regression		✓

### B.2.3. 设置模型集成大小

您可在**此**设置模型集成的算法数量。例如，当**最大模型集成数量**设置为 3 时，最多会有 3 种算法集成在一个模型里。您最终生成的模型中所包含的算法数量还取决于您设置的**训练时长**。

#### B.2.4. 训练验证留出和交叉验证

您可以选择使用以下方法构建机器学习模型：

- **训练验证留出：** 通过将训练数据集划分为三个子集来构建机器学习模型：
  - **训练集：** 用于构建机器学习模型；
  - **验证集：** 用于调整分类器的超参数，以达到更高的准确性；
  - **留出集：** 用于评估模型的准确性。它并不参与模型构建，只用于验证已构建的模型。

您可以拖动的方式设置每个子集占数据集的百分比。

- **K-fold 交叉验证留出：** 通过以下方法构建机器学习模型：
  - a. 将训练数据集划分为“k”个“折叠”或“k”个相同大小的子样本集；
  - b. 在建模过程中，每个‘k’子样本集将作为验证集，其余‘k-1’个子样本集作为训练集；
  - c. 建模过程将重复‘k’次，给出‘k’个模型；
  - d. 我们会平均‘k’个模型，以生成单个模型。

此方法会比**训练验证留出**花费更长的时间。您可设置折叠数，和留出百分比。

#### B.2.5. 重采样设置

注意                      仅适用于二分类模型

若目标变量的结果分布不均衡，则数据集不平衡。

R2-Learn 可通过对训练集进行**上采样**或**下采样**来平衡数据集。对于上采样和下采样，您可调整比例。

#### B.2.6. 设置度量指标

您可选择判断机器学习模型性能的度量指标。

二分类模型，您可选择：

- **AUC（曲线下面积）（默认）**
- **Accuracy**
- **F1**

回归模型，您可选择：

- **MSE（均方差）（默认）**
- **R<sup>2</sup>**

B.2.7. 设置最大优化时间

您可以设置优化模型构建参数的最大优化时间。设置的优化时间越长，生成的机器学习模型的性能越好。对于较大的训练数据集，我们建议您设置更高的最大优化时间。

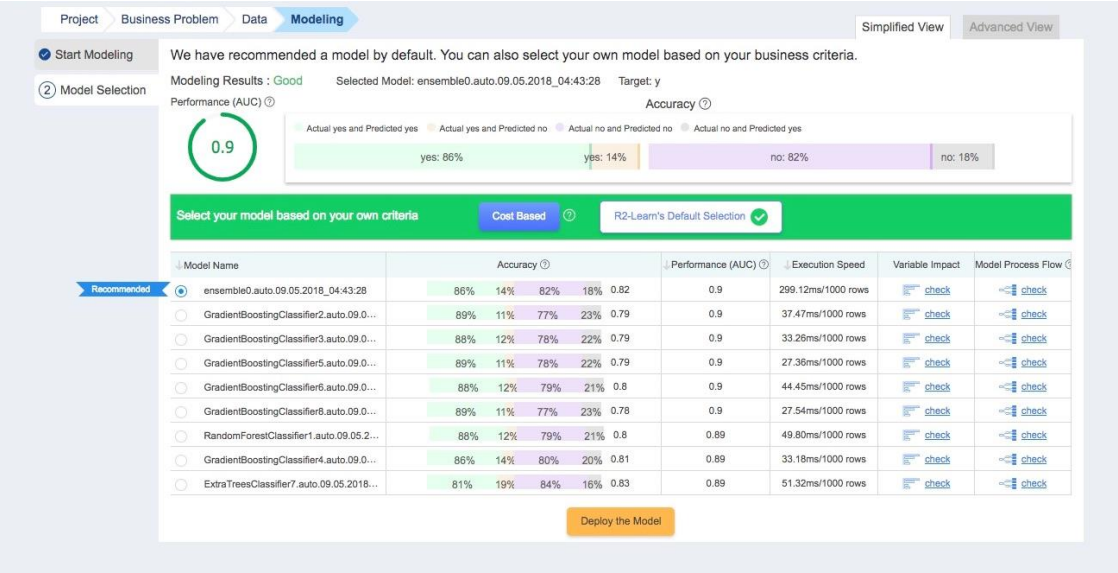
系统会为您自动计算一个最低的优化时间。

B.2.8. 随机种子

训练机器学习模型时，将会在生成随机数时将随机数设置为您设定好的随机种子。此举可使模型可复现。

附录 C: 二分类问题的模型选择

C.1. 简化视图



在二分类问题中的模型选择中，会显示以下内容：

- **建模结果:** 显示了模型的性能水平。
- **所选模型:** 您当前选择的机器学习模型。这是从下面显示的经过训练的机器学习模型列表中选择的。
- **目标:** 机器学习模型预测的目标变量。

- **性能：**显示所选模型的性能。二分类模型中通常是 AUC（曲线下面积）。
- **准确性：**显示所选模型预测的准确性。
- **根据您的标准选择模型：**R2-Learn 从模型列表中自动选择最适合您业务问题的模型。您可以选择其他方式：通过选择以下选项之一来修改其选择推荐的方式：
  - **R2-Learn 的默认推荐：**R2-Learn 会根据执行时间和性能之间的平衡，给出推荐。
  - **基于成本：**此处，您可量化假阳性（第一类错误）和假阴性（第二类错误）会产生的业务损失，以及从真阳性和真阴性获得的收益。R2-Learn 将据此给您推荐模型。

Select your model based on your own criteria

Cost Based ☒ R2-Learn's Default Selection

Please estimate the business benefit or cost of each prediction result (0 ~ 100) ×  
 Note: If a prediction result causes you to lose resource or money, it should be a cost; If a prediction result brings you revenue or income, it should be a benefit. All inputs should be measured at the same unit.

Predict "yes" correctly  
 Benefit: 0 unit

Predict "yes" to be "no"  
 Cost: 0 unit

Predict "no" to be "yes"  
 Cost: 0 unit

Predict "no" correctly  
 Benefit: 0 unit

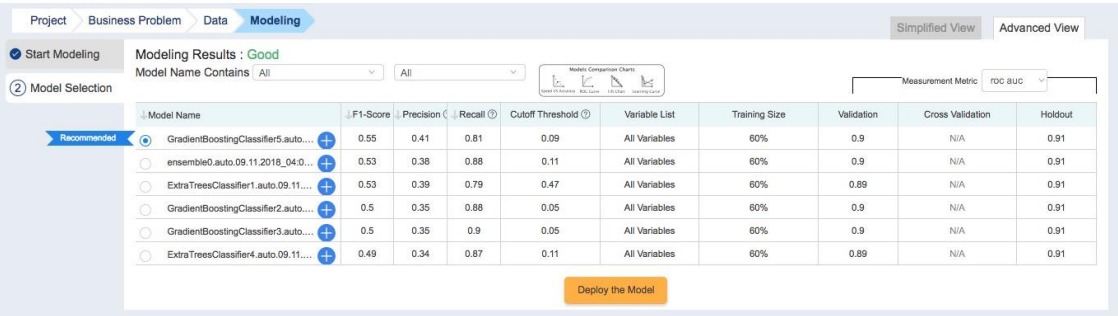
Submit

Model Name	Execution Speed	Variable Impact	Model Process FI...
ensemble0.auto.09.11.20	9.53ms/1000 rows	<a href="#">check</a>	<a href="#">check</a>
GradientBoostingClassifi	1.93ms/1000 rows	<a href="#">check</a>	<a href="#">check</a>
GradientBoostingClassifi	1.56ms/1000 rows	<a href="#">check</a>	<a href="#">check</a>
GradientBoostingClassifi	1.52ms/1000 rows	<a href="#">check</a>	<a href="#">check</a>
ExtraTreesClassifier1.au	1.99ms/1000 rows	<a href="#">check</a>	<a href="#">check</a>
ExtraTreesClassifier4.au	1.63ms/1000 rows	<a href="#">check</a>	<a href="#">check</a>

- **训练模型列表：**此处还显示了使用训练数据构建的模型列表。列出的模型各有一个：
  - **模型名称：**此模型的名字。通常为 "`<algorithm_name>.<model_setting_name>`". 若您选择 **自动建模** 时，`<model_setting_name>` 会自动为您创建。若您选择了 **高级建模**，您可自行设定名称。
  - **准确度：**显示模型预测结果与实际结果的匹配程度。准确度越高，模型预测的效果越好。
  - **性能：**显示此模型的性能指标。二分类模型通常为 AUC（曲线下面积），回归模型通常为 MSE（均方差）。
  - **执行速度：**每处理 1000 行数据，模型需要花费的时间。
  - **变量影响：**单击 **查看**，可查看每个模型变量对目标变量的影响。该值越高，说明该变量的变化对模型预测的结果的影响越大。
  - **模型流程：**单击 **查看**，可看到此模型构建的每一步流程和详细处理方法。



## C.2. 高级视图



用户可以使用**高级视图**查看有关可用于部署的模型的更多详细信息。

### C.2.1. 顶部

顶部显示以下内容：

- **建模结果：**模型的泛化表现。
- **模型内容包含：**您可根据算法和模型设置筛选模型
- **模型比较图表：**每个模型的表现，可用于进行模型性能比较：
  - **速度 v.s.准确性：**每个模型的速度和准确性的表
  - **提升图：**通过比较使用模型和不是用模型预测结果的比率，可用于确定模型的有效性。
  - **ROC 曲线：**根据不同阈值，确定二分类模型的分类能力。

### C.2.2. 模型表

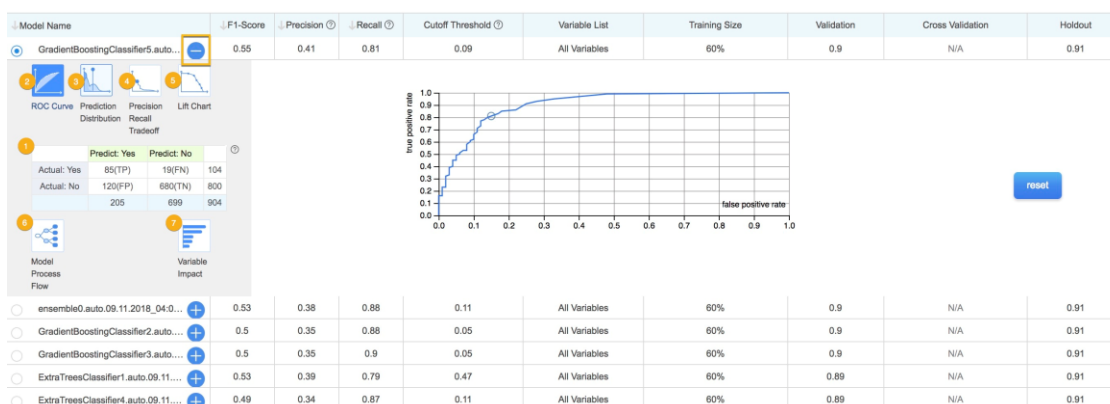
下表显示了使用数据集训练的所有模型。每个列出的模型都有：

- **模型名字：**此模型的名字。通常为 "`<algorithm_name>.<model_setting_name>`". 若您选择**自动建模**时，`<model_setting_name>`会自动为您创建。若您选择了**高级建模**，您可自行设定名称。
- **F1 值：**F1 值是一个综合指标，结合了准确率和召回率。当准确率和召回率都高时，F1 值也会高。其计算方法为： $F1 = 2 * \text{准确率} * \text{召回率} / (\text{准确率} + \text{召回率})$ 。
- **准确率：**描述的是在被识别为正类别的样本中，确实为正类别的比例。其计算方法为： $\text{Precision} = TP / (TP + FP)$ 。
- **召回率：**描述的是在所有正类别样本中，被正确识别为正类别的比例。其计算方法为： $\text{Recall} = TP / (TP + FN)$ 。

- **截止阈值:** 一些模型会返回概率值，将这些值映射到二元类别时，需指定分类阈值。若值高于该阈值，则认为是正类别。若值低于该阈值，则认为是负类别。
- **变量列表:** 显示训练数据集中的哪些变量包含在模型中。
- **验证/交叉验证:** 验证/交叉验证数据集是用于微调初始模型，可查看变量之间的关系是否正确操作，以创建可能的最佳模型的数据。
- **留出:** 留出集是为模型的最终测试留出的一部分原始数据，用于评估模型的执行情况。

### C.2.3. 其它模型细节

要查看其他模型详细信息，请单击**模型名称**旁边的**+**。

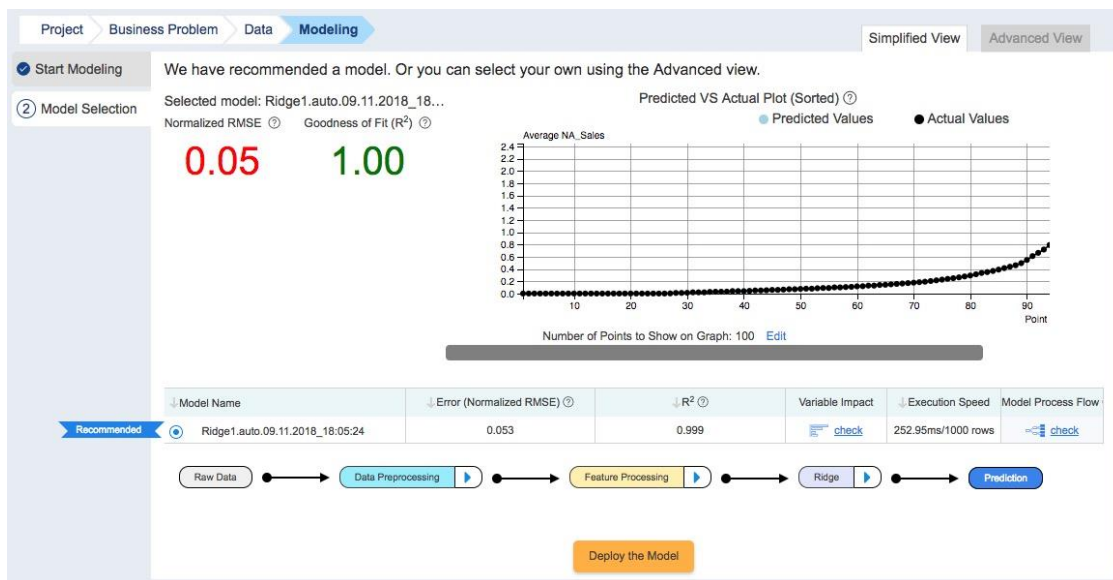


上图显示以下图表:

1. **误差/混淆矩阵:** 此矩阵将预测值与真实值进行比较。是一个显示了真阳性（TP），假阳性（FP），真阴性（TN）和假阴性（FN）的数量的矩阵。
2. **ROC 曲线 (可调整):** 随着阈值的调整，二分类模型分类能力。单击滑块可调整阈值。
3. **预测分布 (可调整):** 这张图里是概率密度分布，概率阈值可调。
4. **精确找回曲线 (可调整):** 这张图是召回率和精确度的关系图。单击滑块可调整。
5. **提升曲线:** 提升曲线图反映了使用模型和不使用模型获得的结果的提升。反映了模型的有效性。
6. **模型流程图:** 这张图显示了模型时如何逐步构建的。
7. **变量影响:** 模型中的每个变量对预测目标的实际影响。值越高，表示对模型预测的影响越大。

## 附录 D: 回归问题的模型选择

### D.1. 简化视图



在回归问题中的**模型选择**中，会显示以下内容：

- **所选模型**：您当前选择的机器学习模型。这是从下面显示的经过训练的机器学习模型列表中选择。
- **归一化 RMSE（均方根误差）**：可帮助您比较模型的性能。其值越小，模型预测的数据越好。
- **拟合质量（R²）**：R² 是一个表示回归线与给定数据拟合程度的度量指标。R² 值越高，模型越接近数据。
  - **变量影响**：是模型中每个变量对预测目标变量的影响。值越高，变量的变化对模型预测结果的影响越大。
- **预测 v.s. 实际（排序）**：显示模型预测值与实际值的接近程度。您可单击图表下的“编辑”更改显示的数据点。
- **训练模型列表**：此处还显示了使用训练数据构建的模型列表。列出的模型各有一个：
  - **模型名称**：此模型的名字。通常为 "`<algorithm_name>.<model_setting_name>`". 若您选择**自动建模**时，`<model_setting_name>`会自动为您创建。若您选择了**高级建模**，您可自行设定名称。

- **误差（归一化 RMSE）**: RMSE（均方根误差）是模型标准误差的平方根，显示了模型预测值与实际目标变量值的接近程度。
- **执行速度**: 每处理 1000 行数据，模型需要花费的时间。
- **变量影响**: 单击**查看**，可查看每个模型变量对目标变量的影响。该值越高，说明该变量的变化对模型预测的结果的影响越大。
- **模型流程**: 单击**查看**，可看到此模型构建的每一步流程和详细处理方法。

## D.2. 高级视图

Model Name	RMSE	RMSLE	MSE	MAE	R <sup>2</sup>	Adjust-R <sup>2</sup>	Variable List	Sample Size	Validation	Cross Validation	Holdout
Ridge1.auto.09.11.2018_18:05:24	0.005	0.004	0	0.002	0.999	0.999	All Variables	60%	0.005	N/A	0.005

用户可以使用**高级视图**查看有关可用于部署的模型的更多详细信息。

### D.2.1. 顶部

顶部显示以下内容：

- **模型名字**: 您可根据算法和模型设置筛选，快速找到模型。

### D.2.2. 模型表

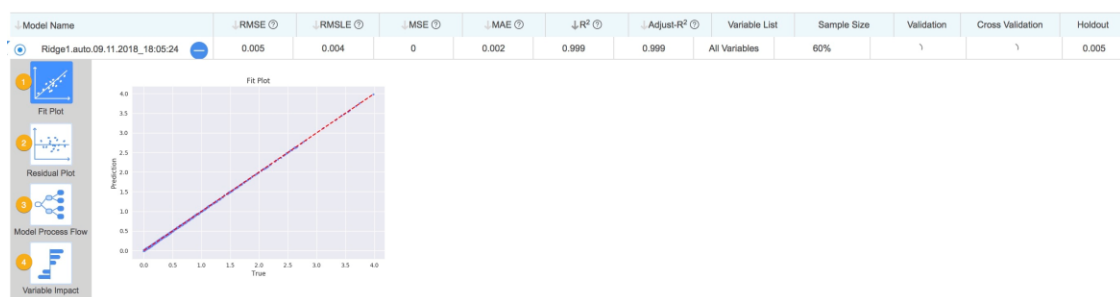
下表显示了使用数据集训练的所有模型。每个列出的模型都有：

- **模型名字**: 此模型的名字。通常为 "`<algorithm_name>.<model_setting_name>`". 若您选择**自动建模**时，`<model_setting_name>`会自动为您创建。若您选择了**高级建模**，您可自行设定名称。
- **RMSE（均方根误差）**: RMSE 是 MSE 的平方根，通常用于比较不同模型之间的预测误差。
- **RMLSE（均方根对数平方误差）**: RMLSE 是 MSE 对数值的平方根。
- **MSE（均方误差）**: MSE 是模型预测值与实际值之间误差平方的平均值。
- **MAE（平均绝对误差）**: 模型预测值与实际值之差的绝对值的平均数。
- **R<sup>2</sup>**: R<sup>2</sup> 是一个表示回归线与给定数据拟合程度的度量指标。R<sup>2</sup> 值越高，模型越接近数据。
- **R<sup>2</sup> 调整**: 调整后 R<sup>2</sup> 是 R<sup>2</sup> 的调整版，它对模型中的预测变量数进行了调整。它总是低于 R<sup>2</sup>。

- **变量列表：**显示训练数据集中的哪些变量包含在模型中。
- **验证/交叉验证：**验证/交叉验证数据集是用于微调初始模型，可查看变量之间的关系是否正确操作，以创建可能的最佳模型的数据。
- **留出：**留出集是为模型的最终测试留出的一部分原始数据，用于评估模型的执行情况。

### D.2.3. 其它模型细节

要查看其他模型详细信息，请单击**模型名称**旁边的**+**。



上图显示以下图表：

1. **拟合图：**拟合图显示了模型的预测精度。横轴为目标变量的实际值，纵轴为目标变量的预测值。**45 度线 ( $y=x$ )** 表明所有的目标变量都被准确的预测了。图中接近或在 **45 度线** 上的点越多，表示模型的预测性能越好。
2. **残差图：**残差是给定数据点目标变量的实际值和预测值之间的差值。横轴是残差，纵轴是实际目标变量值。观察残差图可发现当模型应用于给定数据集时可能会出现的问题。单击**诊断**可将当前模型的残差图与数种常见的残差图进行比较，给您提供如何改进模型的思路。当您在打开的对话框种选中**最接近模型残差图**的残差图形状后，**R2-Learn** 会给出修复方法：
  - 随机分布的残差图
  - Y 轴不平衡的残差图
  - X 轴不平衡的残差图
  - 异常值残差图
  - 非线性残差图
  - 异方差残差图
  - 大 Y 轴数据点残差图

3. **模型流程图:** 这张图显示了模型时如何逐步构建的。
4. **变量影响:** 模型中的每个变量对预测目标的实际影响。值越高，表示对模型预测的影响越大。