******武汉纺织大学**

**毕 业 设 计 [ 论 文 ]**

**题 目 ： 基于python的数据挖掘代码生成器设计与实现**

**学 院 ： 计算机与人工智能学院**

**专 业 ： 软 件 11705**

**姓 名 ： 李 政 辉**

**指 导 教 师 ： 胡 兵**

2021 年 5 月 29 日

武汉纺织大学

毕业设计（论文）任务书

课题名称： **基于Python的数据挖掘代码生成器**

完成期限：  **2020年12月10日至 2021年5月29日**

学院名称 **计算机与人工智能学院** 专业班级  **软件11705**

学生姓名  **李政辉**  学 号 **1704240108**

指导教师 **胡兵** 指导教师职称 **讲师**

学院领导小组组长签字

1. 课题训练内容

1. 分析数据挖掘代码生成器的需求，得出初步的设计思想；

2. 收集课题相关资料、查询相关文献、总结相关文件资料；

3. 通过不断的阅读与总结，整理得到课题的信息，规划处课题的总设

计；

4. 了解有关 Python数据挖代码生成器的发展背景和现状；

5. 学会利用 Python 语言，使用scikit-learn库搭建模型；学会利用django+vue+mongodb完成系统的实现;

6. 了解数据预处理、数据探索相关的思路和方法;

7. 用 Python 实现整体代码，实现Python数据挖掘的代码生成器，写出论文报告。

二、设计（论文）任务和要求（包括说明书、论文、译文、计算程序、图纸、作品等数量和质量的具体要求）

1. 按进度完成相关模块，并主动向导师汇报，征求导师意见。

2. 大量查阅相关资料、收集有关数据，整合成表格。

3. 熟练掌握数据挖掘相关技术以及分析流程，将流程进行整理以及模板化，基于Python的web框实现代码生成系统。

4. 按照要求完成不少于 2000 字的开题报告。

5. 详述实验过步骤并按照流程实验，完成报告。

6. 找一篇与论文技术相关的外文文献进行阅读，并翻译外文文献。

7. 完成毕业论文，论文内容要求层次清楚，语言准确，文字顺畅，逻辑性强。论文格式符合毕业论文格式要求。字数达到 15000 字以上，且文章查重率低于 30%。

三、毕业设计（论文）主要参数及主要参考资料

[1].王博;舒新峰;王小银;陈锐. 自动代码生成技术的发展现状与趋势. 西安邮电大学学报[J]

[2].张艳伟,陈正鸣,吕嘉,张霞. 主流代码生成方法的分析与比较[J]. 计算机与现代化,2015(11):103-108.

[3].李陈军. 基于MongoDB的SNS平台设计与实现[D].南京理工大学,2013.

[4].牛岩.作物产量指标综合评价的数据标准化处理[J].农村经济与科技,2017,28(19):16-19.

[5]. 叶之放. 基于多源公交数据和LSTM的公交到站时间预测研究[D].华南理工大学,2019

[6]. 刘海航. 基于XGBoost和BP神经网络的会员流失预测及内容推荐方法的研究[D].内蒙古大学,2019.

[7]. 陈颖. 电商商品评论情感分析方法及优化研究[D].南昌大学,2017.

[8]. 陈家伟. 支持向量机在人体健康状态预测中的研究与应用[D].中国科学技术大学,2014.

[9]. 许小芬. 基于时间序列的磨削颤振预测方法研究[D].浙江理工大学,2019.

[10]. 吕鹏滨. 社交网络匹配算法研究与改进[D].北京邮电大学,2017.

[11]. 任禹丞,徐超,赵磊,贾静,彭路,周子馨.基于自适应特征权重聚类算法的用电问题分析[J].计算机系统应用,2020,29(01):29-39.

[12]. 鲁志军,邱雪涛,赵金涛.数据挖掘在交易欺诈侦测中的应用[J].软件产业与工程,2014(05):24-28+38.

[13]. 刘超. 业界专家的媒体发言对公司股价影响的分析[D].上海师范大学,2016.

[14]. 潘抱一. 基于动态程序失效执行块的软件错误定位方法研究[D].浙江理工大学,2020.

[15]. 徐京京. 基于聚类和神经网络的异常数据识别算法研究[D].华北电力大学(北京),2019.

[16]. 王东岳,苗乃树,王忠举,卢绪振,屈慧星,张波,马天行.适用于夜晚环境下多个车牌的定位算法[J].农业装备与车辆工程,2018,56(05):69-71.

四、毕业设计（论文）进度表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 序  号 | 起止  日期 | 计划完成内容 | 实际  完成情况 | 检查人签名  检查日期 |
| 1 | 2020/12/10-2020/12/15 | 选题 | 完成 |  |
| 2 | 2020/12/16-2021/01/10 | 完成开题报告 | 完成 |  |
| 3 | 2021/01/11-2021/01/30 | 模型可行性分析 | 完成 |  |
| 4 | 2021/01/31-2021/02/10 | 数据收集 | 完成 |  |
| 5 | 2021/02/11-2021/02/25 | 模型建立 | 完成 |  |
| 6 | 2021/03/01-2021/03/27 | 完善系统 | 完成 |  |
| 7 | 2020/03/28- 2020/04/15 | 初步编写论文 | 完成 |  |
| 8 | 2020/04/16- 2020/04/20 | 毕业设计中期检查 | 完成 |  |
| 9 | 2020/04/22- 2020/05/14 | 完成论文并交指导教师批阅 | 完成 |  |
| 10 | 2020/05/14- 2020/05/29 | 毕业设计答辩 | 完成 |  |

**武汉纺织大学毕业设计（论文）开题报告**

|  |  |  |  |
| --- | --- | --- | --- |
| 课题名称 | 基于python的数据挖掘代码生成器设计与实现 | | |
| 学院名称 | 数学与计算机学院 | 专 业 | 软件工程 |
| 班 级 | 软件11705班 | 学生姓名 | 李政辉 |
| |  | | --- | | 1. **研究的背景和意义**   近年来，数据挖掘引起了信息产业界的极大关注，其主要原因是存在大量数据，可以广泛使用，并且迫切需要将这些数据转换成有用的信息和知识。获取的信息和知识可以广泛用于各种应用，包括商务管理、生产控制、市场分析、工程设计和科学探索等。数据挖掘利用了来自如下一些领域的思想：来自统计学的抽样、估计和假设检验；人工智能、模式识别和机器学习的[建模技术](https://baike.baidu.com/item/%E5%BB%BA%E6%A8%A1%E6%8A%80%E6%9C%AF/19041163)和学习理论。数据挖掘也迅速地接纳了来自其他领域的思想，这些领域包括最优化、进化计算、信息论、信号处理、可视化和信息检索。一些其他领域也起到重要的支撑作用。特别地，需要数据库系统提供有效的存储、索引和查询处理支持。源于高性能（并行）计算的技术在处理海量数据集方面常常是重要的。分布式技术也能帮助处理海量数据，并且当数据不能集中到一起处理时更是至关重要。  随着互联网各种的技术的成熟以及应用的落地，数据量的与日俱增，不同行业对数据分析的需求也因此产生了巨大的增长。此外，在目前的各种互联网产品中，后台存储的大量数据拥有着巨大的挖掘价值，随着机器学习技术在各领域的广泛应用，使得对大批量数据的建模分析变得容易。  然而，在企业实际运作中，由于不同数据源产生的数据往往有着不同字段，以及不同的预处理逻辑，使得面对不同的数据集时需要针对特定场景进行代码的重新编写。鉴于数据挖掘流程有着明确的处理流程，高可复用的逻辑，即数据探索、数据清洗、数据建模，模型评估等步骤。本文着眼于将数据挖掘流程模板化，通过用户在界面选择对应方法，个性化定制，生成以用户上传数据集为分析目标的python数据挖掘代码，提高数据分析流程的效率。   1. **发展状况**   自动代码生成技术从20世纪末的萌芽阶段发展之今，已经趋于平稳和成熟，“自动代码生成”从1991年开始出现相关研究，2015年达到最热。如图1所示（统计图中的纵轴代表研究成果，文中的数据统计于百度学术）    图1自动代码生成技术趋势图  自动代码生成是一个多领域多学科交叉的学科,涉及的面比较广。随着研究的不断深入，出现了越来越多与“代码生成”相关的研究点,形成了庞大的研究网络，该网络囊括：软件开发、编译器、模型驱动、嵌入式系统、设计模式和编译程序等领域的知识、国内的软件企业在自动代码生成和模型驱动构架的研究和开发中也有不少成果、比如金蝶软件公司是首家实现支持模型驱动构架软件开发的商业工程工具金蝶EAS4.0、是一个基于模型驱动架构、采用“业务模型驱动的全自动化软件工厂”开发的企业应用平台。楚凡科技是专业复杂软件及系统开发解决方案的供应商。为软件开发的整个生命周期提供集成的产品和专业化服务，楚凡科技工具KantStudio集合了UML建模和数据库建模、实现数据库模型和UML模型的双向转换；实现代码和模型之间的迭代转换。  一些中小企业和开源产品在上自动代码生成方面也有较好表现,开发了小型的代码生成器支持从模型或者模板到不同程序设计语言的转化。如动软.Net代码生成器是基于关系数据库元数据库的C#自动代码生成器,Codematic生成的代码,基于面向对象思想和三层架构设计。BBOSS自动代码生成工具是一款为专有框架和平台生成前端、后端代码、WEB服务代码､SQL配置､IOC配置的可视化配置管理工具｡CodeGenerator生成工具是基于Freemarker和Velocity模板生成代码的引擎,该生成器是根据关系实体的元数据来生成源代码。  国内的众多高校和研究机构也在代码生成、尤其是基于模型驱动骨架的代码生成软件工程也进行了长期的研究,取得了众多的成果。下面给出国内的在自动代码生成方面研究的典型问题，文献。在基于模型转换和代码生成、对象关系映射方面的自动代码生成,基于模式的用户界面自动代码生成技术，在逆向工程和程序流程图到自动代码生成提出一种基于图的生成算法，在平台无关模型到平台相关模型，web用户界面建模和自动生成,用户界面代码自动生成等等[1]   1. **研究内容**   本文主要研究通过python实现基于Web端的数据挖掘代码生成器。代码的生成主要采用基于基于模板的代码生成方法。基于模板的代码生成的基本原理是把软件的需求分成两部分,一部分是相对来说固定不变的部分,称为静态部分，即后台系统中的模板文件,另外一部分是根据外部输入变化的部分，即来自前端用户的输入（如选择哪些列做数据预处理，选择什么模型构建数据等等），如图2所示。    图2基于模板的自动代码生成  基于数据挖掘的流程，预先设置好数据预处理、数据可视化、数据建模对应的模板文件。为了能够取得更直观的界面以及更佳的用户体验，本文采用Django+mongoDB+Vue+Redis的架构实现代码生成器的系统功能。  用户通过上传数据集，系统读取数据存入mongodb，前端可展示该用户所有数据集，通过预览按钮解析上传文件并做对应前端展示。数据展示端提供一键生成数据分析报告功能，以交互式网页展示所选数据集每一列的类型、缺失值、异常值、强相关变量，并以高亮的形式展现，用以提示用户，作为数据预处理的参考。   1. **研究步骤**   1.自行查找资料，做好设计前的理论与技术准备。阅读数据分析、数据挖掘、机器学习相关的书籍文献等资料，汇总常见的数据预处理的方法，不同的机器学习回归、分类、聚类模型，以及常用的Python可视化库。  2.在步骤1的基础上，写出数据预处理、数据建模的模板文件。优化代码结构，提高代码的可读性，寻找合适的方法串联数据预处理与数据建模的代码逻辑。  3.实验方案的可行性。实现代码生成脚本，通过传入不同模型参数，确保能以运行脚本的形式生成数据挖掘的代码且能正常运行。  4.前后端交互。完成前端界面的设计，完成后端接口的交互。在web端不同的输入框中输入数据预处理/数据建模的参数，前端能够以弹出框的形式展示生成的代码，并支持生成代码的导出（下载）。同时构建代码运行引擎能够实时运行出生成代码的结果，并给出各个模型的性能比较。  5.用户权限划分。在功能完成的基础上，实现用户的登录/注册，对用户上传的数据集大小，数据集数量做出限制，确保代码生成器后台的正常运作，防止数据爆仓。   1. **参考文献**   [1].王博;舒新峰;王小银;陈锐. 自动代码生成技术的发展现状与趋势. 西安邮电大学学报[J]  [2].张艳伟,陈正鸣,吕嘉,张霞. 主流代码生成方法的分析与比较[J]. 计算机与现代化,2015(11):103-108.  [3].李陈军. 基于MongoDB的SNS平台设计与实现[D].南京理工大学,2013.  [4].牛岩.作物产量指标综合评价的数据标准化处理[J].农村经济与科技,2017,28(19):16-19.  [5]. 叶之放. 基于多源公交数据和LSTM的公交到站时间预测研究[D].华南理工大学,2019  [6]. 刘海航. 基于XGBoost和BP神经网络的会员流失预测及内容推荐方法的研究[D].内蒙古大学,2019.  [7]. 陈颖. 电商商品评论情感分析方法及优化研究[D].南昌大学,2017.  [8]. 陈家伟. 支持向量机在人体健康状态预测中的研究与应用[D].中国科学技术大学,2014.  [9]. 许小芬. 基于时间序列的磨削颤振预测方法研究[D].浙江理工大学,2019.  [10]. 吕鹏滨. 社交网络匹配算法研究与改进[D].北京邮电大学,2017.  [11]. 任禹丞,徐超,赵磊,贾静,彭路,周子馨.基于自适应特征权重聚类算法的用电问题分析[J].计算机系统应用,2020,29(01):29-39.  [12]. 鲁志军,邱雪涛,赵金涛.数据挖掘在交易欺诈侦测中的应用[J].软件产业与工程,2014(05):24-28+38.  [13]. 刘超. 业界专家的媒体发言对公司股价影响的分析[D].上海师范大学,2016.  [14]. 潘抱一. 基于动态程序失效执行块的软件错误定位方法研究[D].浙江理工大学,2020.  [15]. 徐京京. 基于聚类和神经网络的异常数据识别算法研究[D].华北电力大学(北京),2019.  [16]. 王东岳,苗乃树,王忠举,卢绪振,屈慧星,张波,马天行.适用于夜晚环境下多个车牌的定位算法[J].农业装备与车辆工程,2018,56(05):69-71.  同意开题□ 不同意开题□  指导教师签名：  年 月 日 | | | | |

# 目 录

## 概述

本项目主要研究通过Python实现基于Web端的数据挖掘代码生成器。代码的生成主要采用基于基于模板的代码生成方法[2]。关于基于模板的代码生成，该项技术的原理主要是为了将功能的需求进行拆解，从而拆分成两个部分。第一部分是相对来说固定不变的部分,即静态部分，来自后台系统中的模板文件；另一部分是根据外部输入变化的部分，即来自前端用户的输入（如选择哪些列做数据预处理，选择什么模型构建数据等等），如图1-2所示。



图1-2基于模板的自动代码生成

本系统基于数据挖掘的流程，通过预先设置好数据预处理、数据可视化、数据建模对应的模板文件。经过代码生成引擎调用，结合前端数据的输入，动态生成可运行的Python脚本，同时运行脚本生成可视化报告，用于前端展示。

为了能够取得更直观的界面以及更佳的用户体验，本文采用Django + MongoDB + Vue的架构实现代码生成器的系统功能。

用户通过上传数据集，系统读取数据存入MongoDB，前端可展示该用户所有数据集，通过预览按钮解析上传文件并做对应前端展示。数据展示端提供一键生成数据分析报告功能，以交互式网页展示所选数据集每一列的类型、缺失值、异常值、强相关变量，并以高亮的形式展现，用以提示用户，作为数据预处理的参考。

## 系统功能简述

本系统以数据驱动。通过保存用户上传的数据集到系统中用户的数据集管理界面，以供用户选择需要分析的数据集，并进行数据挖掘相关操作。

①、数据预览。数据预览将根据一个数据集以表格的形式展示,并且支持根据一个数据集具体内容自适应,从而制作出探索性数据分析报告。数据分析研究报告是针对大量数据整体进行的探究和分析，将数据集中存在缺失值、异常值、高相关度的关键信息予以高亮展示。同时针对数据集各列的维度进行详细分析以及分布特征的可视化图表。数据分析报告给出的高亮信息用以为数据清洗提供参考方向

②、数据清洗。用户在已上传数据集列表中，选择需要分析的数据集，自定义选择不同的列，在提供的数据清洗方法中自由组合不同的数据清洗规则以达到清洗“脏”数据的目的。系统支持的数据清洗规则主要包括缺失值填充、排序、模糊匹配、数据标准化等，允许用户自定义调换规则顺序，后台引擎根据用户输入的规则进行校验，并展示清洗后的数据表格。

③、数据建模。用户选择指定数据集，选择数据集中的若干列作为数据特征，选择一列作为目标列。提供分类、回归、聚类问题的若干种训练模型，以及对应的模型评估方法，模型评价结果以可视化图表的形式展示。参数选择完毕通过前端界面弹窗显示生成的代码。

# 数据挖掘代码生成系统分析

## 整体架构架构

本系统基于Python的Web框架django运行，前端使用vue框架以达到前后端分离的效果，系统架构如图2-1所示。

图形用户界面

描述已自动生成

图2-1 数据挖掘代码生成系统架构图

## 系统需求分析

### 需求分析概述

1.普通用户：普通用户可上传数据集（最大支持上传5份数据集），用户可在上传的数据集中选择需要分析的数据集进行预览，根据预览界面的数据分析报告生成详细数据探索分析报告，该报告为用户进行数据清洗提供指导方向。若数据集无异常数据，则可通过数据建模界面选择指定列需要的模型。根据用户配置，前端弹出窗口展示生成的数据文件。由代码生成器生成的代码支持导出（不支持运行）。

2.高级用户：高级用户可上传数据集（最大支持上传10份数据集），用户可在上传的数据集中选择需要分析的数据集进行预览，根据预览界面的数据分析报告生成详细数据探索分析报告，该报告为用户进行数据清洗提供指导方向。若数据集无异常数据，则可通过数据建模界面选择指定列需要的模型。根据用户配置，前端弹出窗口展示生成的数据文件。由代码生成器生成的代码支持导出，同时支持生成代码的运行，并以可视化html的形式展示模型评估报告。

### 代码生成系统流程图

用户通过代码生成系统的操作流程如图2-2所示

图示

描述已自动生成

图2-2 代码生成系统流程图

## 本章小结

本章就代码生成系统的整体架构和功能需求进行了详细的介绍和图解。使用Internet应用程序的B/S结构来实现网站，采用前后端分离的架构方式，详细的介绍了代码生成系统的存储，引擎，展示结构。并通过流程图的形式来介绍了下整个网站系统的流程，以及各个模块所具有的功能。此外，本章就Web安全的角度出发，考虑了本系统在实际场景下可能出现的情况做了预防。

# 数据库分析与设计-MongoDB

## 数据集合的具体设计

以下就是本代码生成系统中所用到的所有数据集合

1. user\_model：用户集合：存储使用用户信息以及用户上传数据集的列表

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 备注 |
| \_id | ObjectId | 默认自增id |
| username | String | 用户名 |
| password | String | 密码(base64加密) |
| email | String | 注册邮箱 |
| isVip | Boolean | 高级用户标识 |
| vipEndtime | Double | Vip到期时间 |
| dataset | Array | 用户上传的数据集名称 |

表3-1用户集合字段详情

1. dataset\_model；数据集的集合，用来保存整个平台上传的全部数据集

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 备注 |
| \_id | ObjectId | 默认自增id |
| username | String | 用户名 |
| dataset\_name | String | 数据集名称 |
| columns | Array | 数据集的各列名称 |
| data | Object | 数据集内容 |

表3-2 数据集集合字段详情

1. temp：临时集合，用于存储验证码内容

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 备注 |
| \_id | ObjectId | 默认自增id |
| address | String | 发送目标地址邮箱 |
| check\_code | Int32 | 验证码 |
| send\_time | Double | 验证码发送时间 |

表3-3临时集合字段详情

1. history\_model构建历史集合

|  |  |  |
| --- | --- | --- |
| 字段名称 | 字段类型 | 备注 |
| \_id | ObjectId | 默认自增id |
| dataset\_name | String | 数据集名称 |
| target | String | 模型训练的目标列 |
| feature | Array | 模型训练的特征列 |
| model\_type | String | 使用的模型类型 |
| model\_name | Array | 使用模型的名称 |
| evaluate\_methods | Array | 模型评价方法 |
| username | String | 用户名 |

表3-4构建历史集合字段详情

# 数据挖掘代码生成器模板代码设计

代码的生成实现方式主要采用基于模板的代码生成方法[2]。原理是把功能的需求拆分为二,第一部分是相对固定不变的部分,即静态部分，大多是来自后台系统中的模板文件；另一部分是根据外部输入变化的部分，即来自前端用户的输入（如选择哪些列做数据预处理，选择什么模型构建数据等等)，整体流程如图4-1所示



图4-1基于模板的自动代码生成

一般而言，生成代码需要经由三个维度进行考虑。第一个维度是寻找复用性高的代码，也就是可以直接使用的代码；第二个维度是有规律的代码（即结构相似的代码），此类型可通过自定义配置进行生成，例如不同的数据集的不同名称，不同字段，字段有着不同的类型，但是读取数据，切割数据，训练模型的代码却是相同的。第三个维度是没有规律的代码，即我们在代码生成中需要考虑的特殊场景。

## 数据清洗代码生成器

### 数据清洗代码模板设计

数据清洗的模板相对简单且固定。通过固定的读取方法，对不同数据列调用不同的数据清洗方法，清洗方法通过预定义的参数文件获取并填充。数据清洗的代码均通过pandas库完成。

数据清洗方法我们通过将上一小节提及的清洗方法以json文件形式存储，构造数据清洗方法—方法代码的映射关系，达到快速构建的效果。每当用户添加一条规则时，通过映射关系文件找到对应代码，进行逐行添加。映射关系文件如图4-2所示。（以下json文件中的df表示通过pandas读取的数据集文件,cols为需要操作的列，其他部分为对应数据清洗方法的调用）

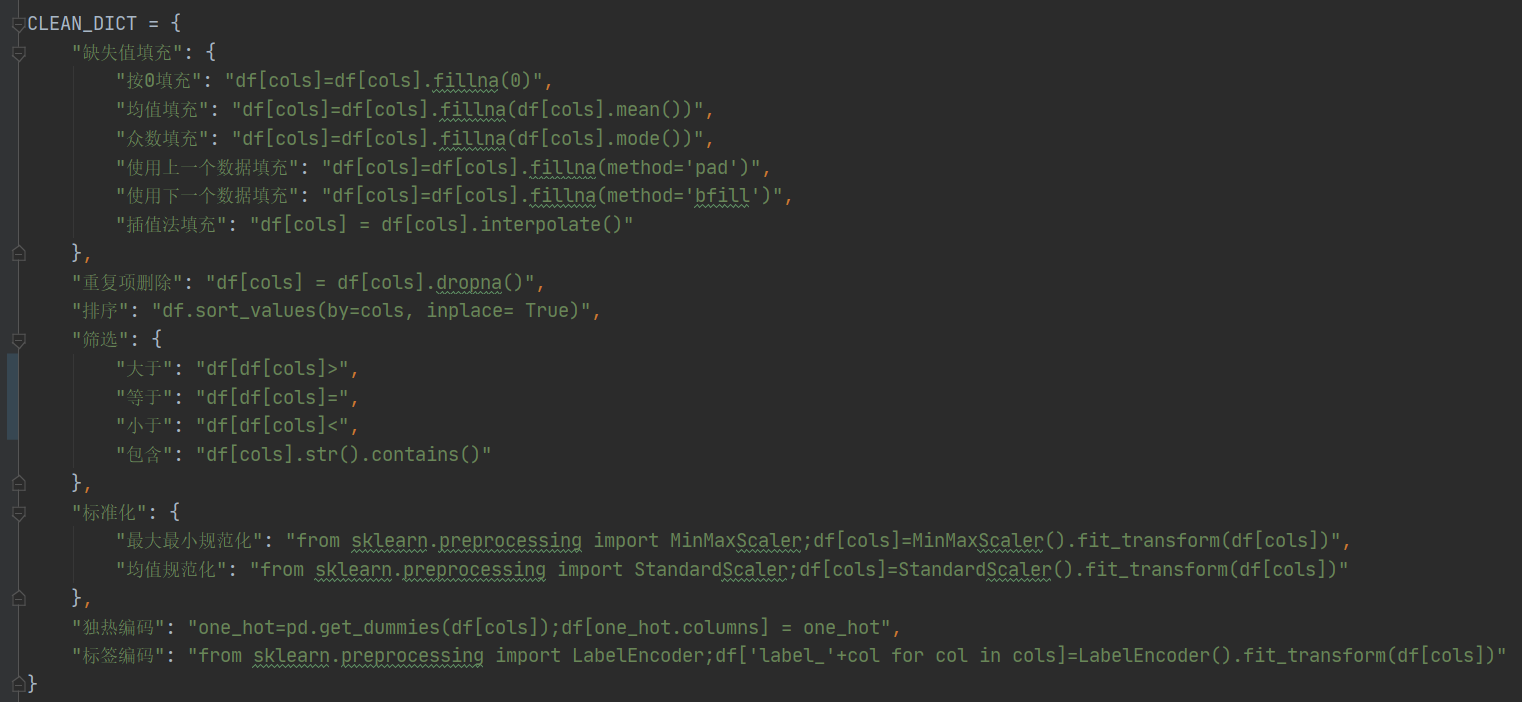


图4-2 数据清洗方法选择关系映射

### 数据清洗代码生成交互

前端页面通过提供下拉框以供用户选择，通过自定义多条清洗规则（规则形式:对xxx列进行xxx操作，具体包含的操作方法在上节中已做介绍）进行代码生成，具体流程图如图4-3所示，界面原型如图4-4所示

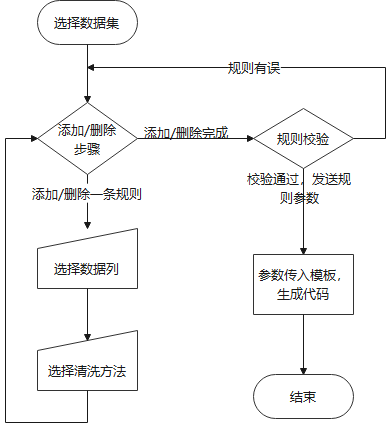


图4-3 数据清洗代码生成交互流程图



图4-4数据清洗代码生成原型图

## 数据建模（机器学习）代码生成器

### 数据建模（机器学习）代码生成模板设计

在自动生成数据模型建模代码之前，我们需要给用户提供代码生成的指导方向，即根据不同的问题需要传入对应的参数。数据建模时我们往往需要知道数据集的特征列、目标列以构建合适的模型解决问题。因此特征列，目标列以及模型类型是代码生成器的必要参数，基于此，我们的代码生成器构建的流程图如图4-6所示：

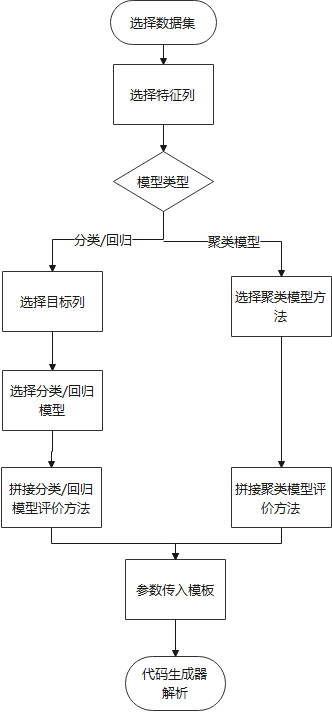


图4-6 数据建模代码生成器用户使用流程图

用户传入参数后至代码生成器后台，后台将参数解析，并在对应模板中扩充代码。传递的参数主要包括数据集名称，模型类型（分类、回归或者聚类），特征列名称，目标列名称（仅分类，回归任务包含），应用模型（一个或者多个）等参数传至后台数据建模的代码生成的逻辑主要如图4-7所示。

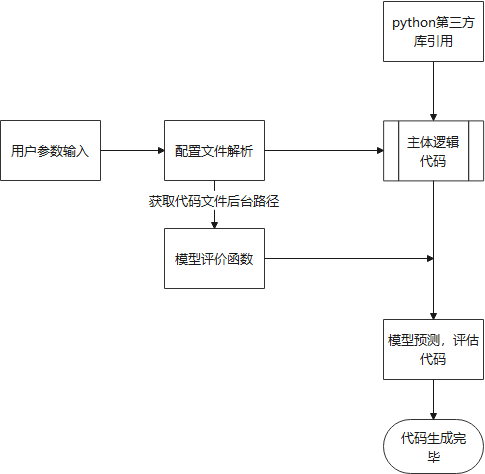


图4-7 数据建模代码生成逻辑流程图

首先需要引入建模需要的python第三方库，包括用于数据分析的pandas，numpy，用于机器学习建模的sklearn。

通过读取用户的参数输入获取需要构建的模型类型，进行主体逻辑的代码拼接。主体代码逻辑包括读取数据集文件、选择特征列，目标列，并拆分数据集为训练集与测试集。拆分完成后，将训练集作为模型的输入进行训练，并将训练的好的模型（模型支持多选）以列表的形式保存，作为后续预测模型方法的输入。

根据用户选择的分类/回归/聚类任务下的不同模型，进行主体逻辑拼接。生成器获取模型的映射关系以json形式储存，如下key-value结构所示。其中key为具体模型的名称，value为该python调用sklearn中封装的api的具体代码，在生成代码的过程中会优先导入这部分方法，导入方法完成后会对这部分字符串进行分隔，转换成python中实际运行的语法,配置如下图4-8所示：

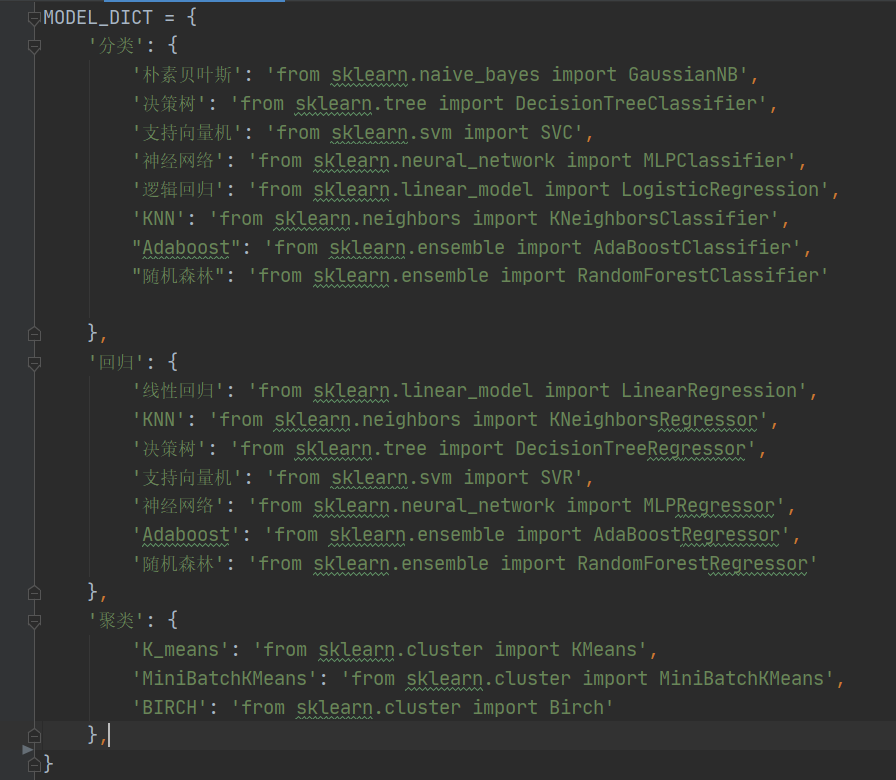


图4-8 数据建模模型选择关系映射

此外，针对分类模型与回归模型分别有着不同的模型评价参数，计算过程代码以单独模块的形式保存在后台。通过任务类型读取模块文件，将文件内容拼接于主体逻辑代码之后。至此，数据建模的代码生成已经完成。

基于上述流程，数据挖掘模型的交互界面(用户参数选择)原型图如图4-9所示

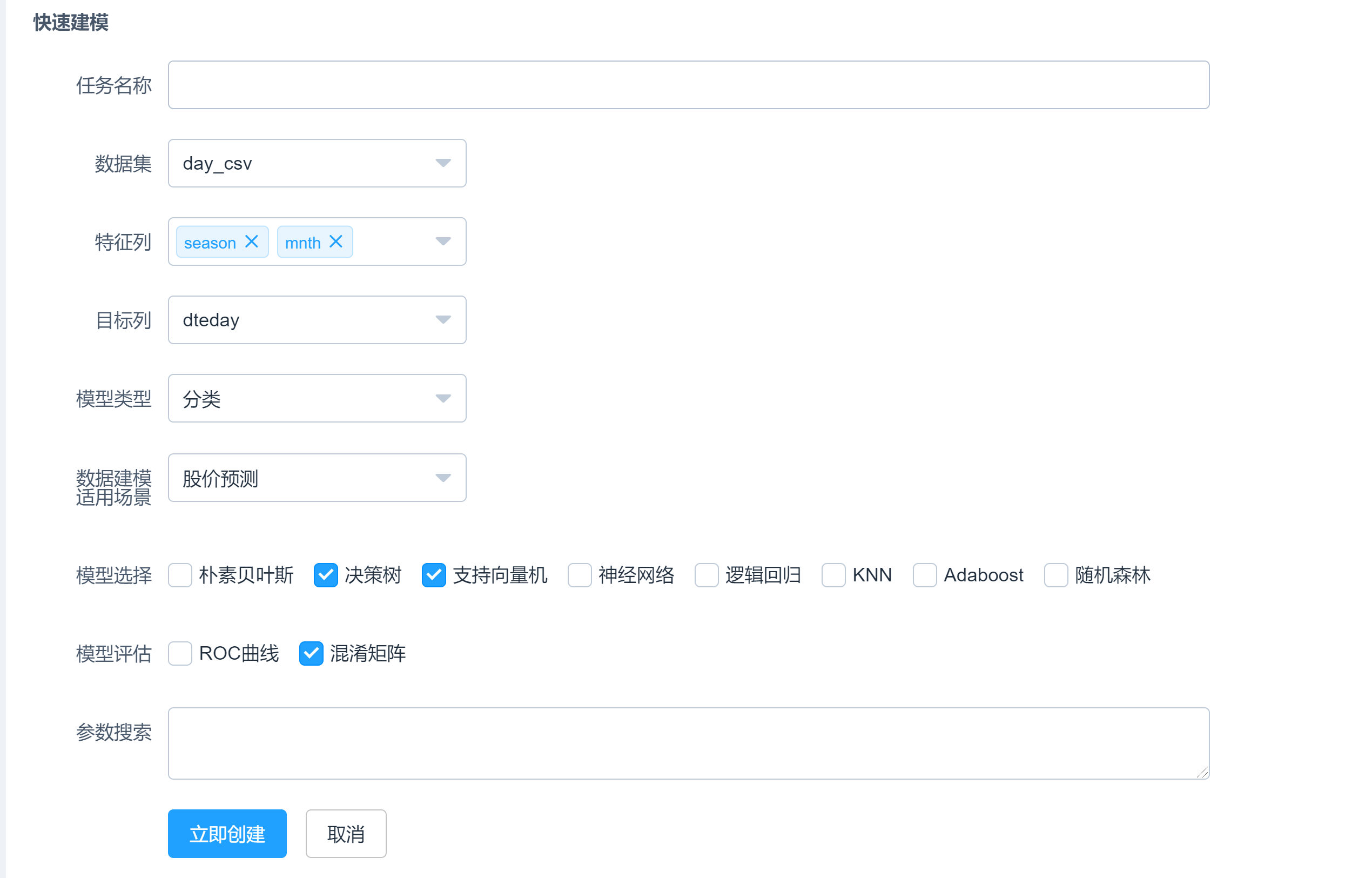


图4-9 数据挖掘代码生成器模型选择交互界面原型图

代码生成界面的效果如图4-10所示。

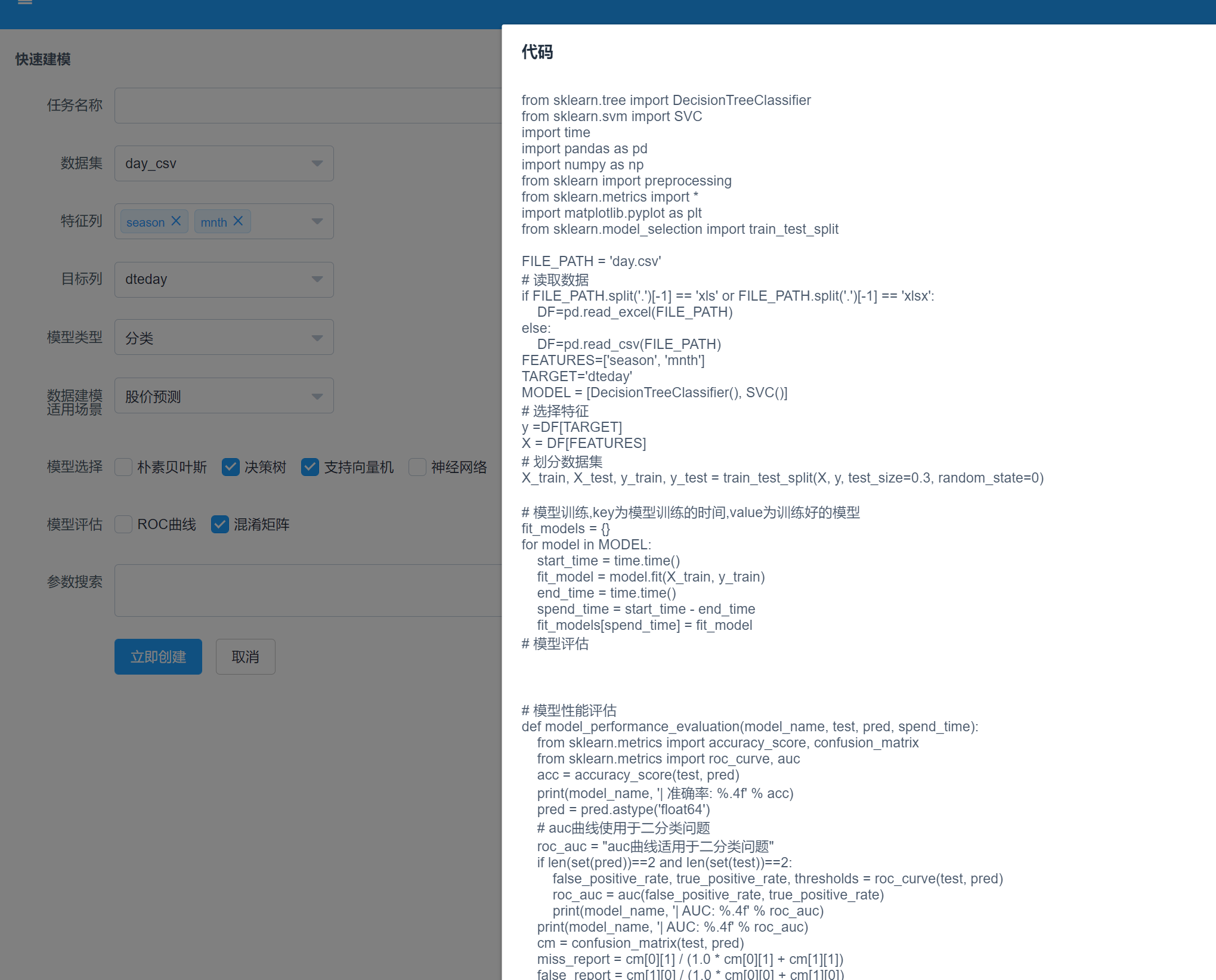


图4-10 代码生成效果图

在快速建模界面，提供了一个用户可选下拉框，以供用户填写当前分析的数据集适用的场景。下拉框的选项为本系统的内置场景，选项可选股价预测、天气预测、价格预测、属性(动物/植物/疾病)分类等等。该场景分类属性内置为后续功能——根据用户上传数据集能够实现推荐模型做出准备。

# 数据分析系统引擎设计

为了实现一键式探索性分析，后台引擎使用了pandas\_ Profiling库。 pandas\_Profiling库使用了基于pandas的DataFrame数据结构，利用DataFrame可以简单快速地进行探索性数据分析,并生成可交互的数据分析报告。分析报告从数据集分析结果概览（Overview）、各变量单列描述性统计(Variables)、变量与变量之间的可视化关系(Interactions)、所有变量之间的相关矩阵热力图(Correlations)、所有数据列缺失特征(Missing values)以及数据集部分数据展示(Sample)等多个维度进行展示。数据分析报告概览如图5-1所示

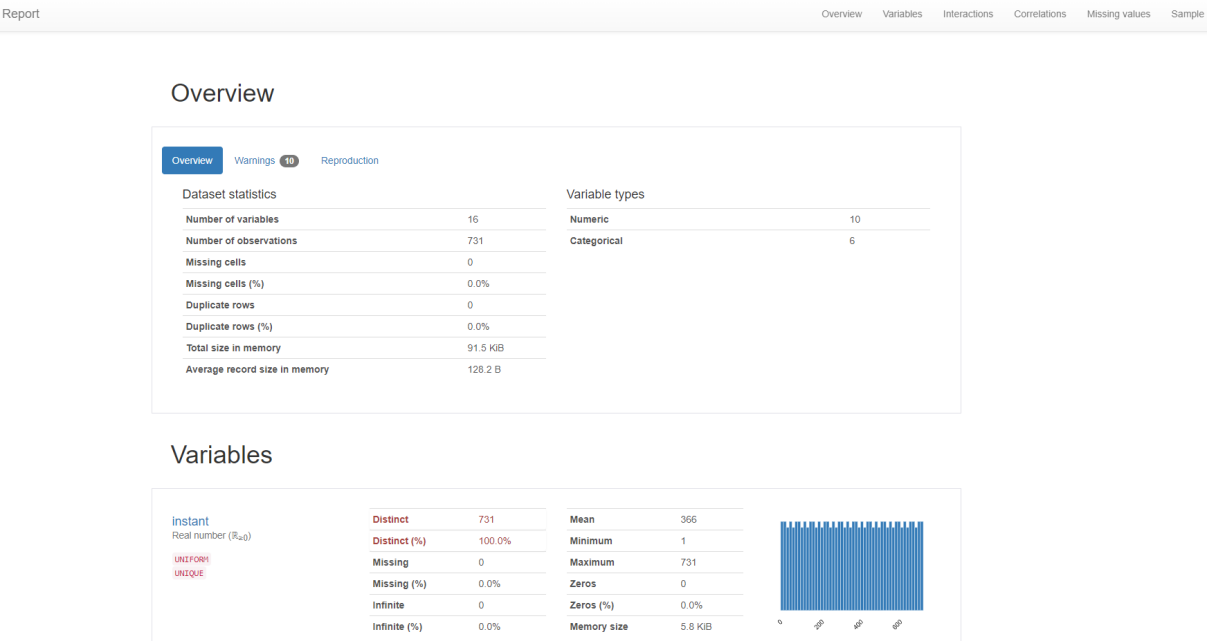


图5-1 探索性数据分析报告概览

其中在报告的概览Overview界面会提供数据集的基本信息如图5-1所示。其中Variables统计了数据集中每一列的详细分析情况，图5-1中即截取了样例数据集instant列的详情统计。当点击图5-1中的Warnings Tab页可以看到当前数据集中可能有重要意义的数据如图5-2所示，此类特征包括相关性高的两个特征，服从正态分布的特征，特征是否是离散值或者连续值，是否某列特征存在缺失值，或者无意义值（0值，空字符串等）过多的情况。具体的详情可通过右上角的五个页签跳转，查看各列特征的统计情况，分布情况等。

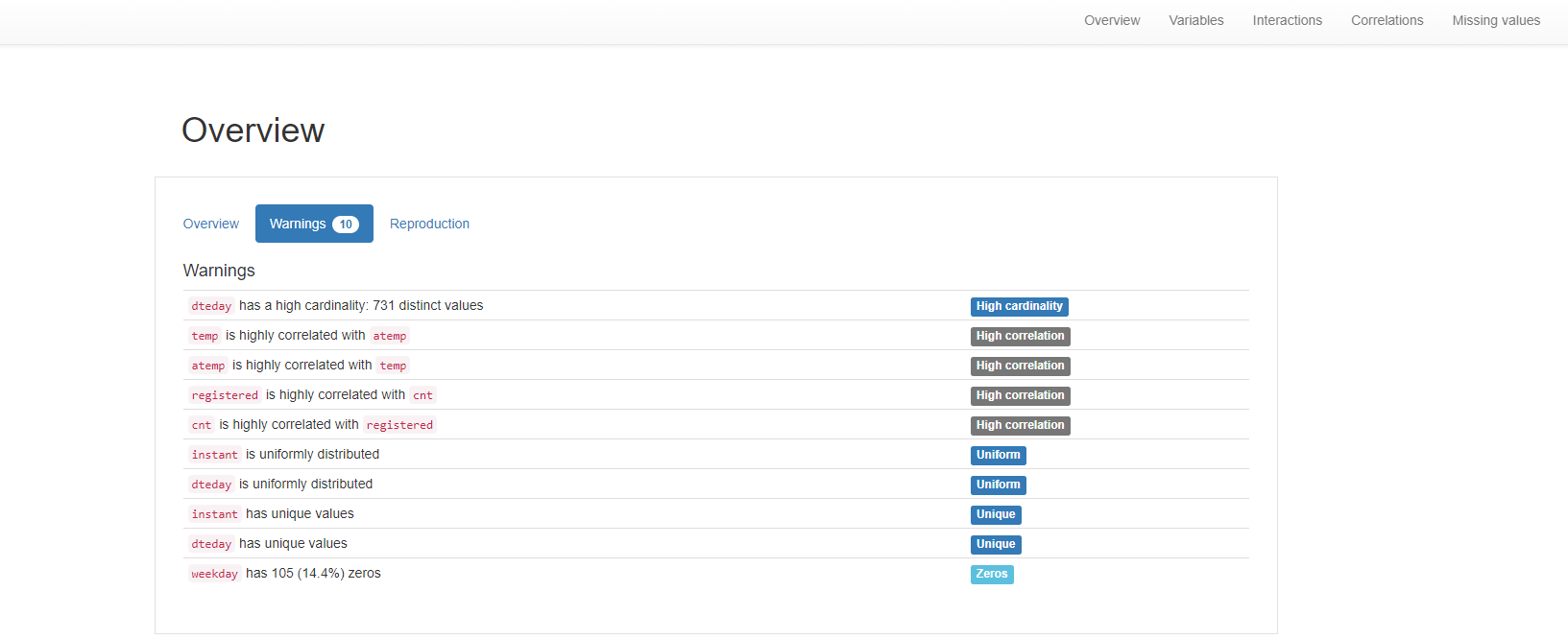


图5-2 探索性数据分析关键特征

## 数据清洗引擎设计

由于不同的数据集有着不同的清洗步骤和方法，这意味着针对不同来源的数据所遇到的困难也是不同的。因此，固定顺序的数据清洗模板往往很难应用于所有的数据集，数据清洗引擎的功能主要是使得用户能够自定义进行数据清洗规则的排序，同时根据数据清洗规则进行处理，对规则的合理性进行校验，对于能生效的清洗规则，输出对应的新数据，而不仅仅是生成一份看似合理的清洗代码。

所以，数据清洗引擎主要在于规则合理性的校验，当所有清洗规则校验通过时，生成一份新的数据集，即清洗完毕的数据集。对于校验不通过的规则，则解析后台错误信息，在前端给出具体错误原因。基于此，设计如下引擎的工作流程如图5-3所示。

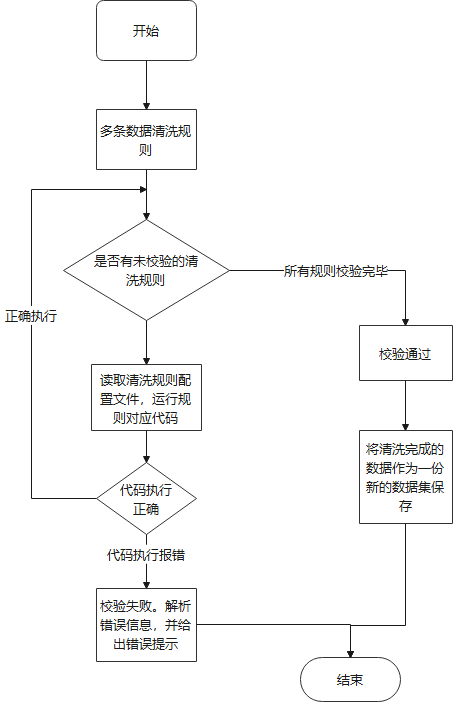


图5-3 数据清洗引擎流程图

## 数据建模（机器学习）引擎设计

在用户上传不同的数据集文件时，引擎通过读取数据集内容判断需要使用的模型。考虑到很多用户在实际应用场景中会对分类模型构建与回归模型的具体使用产生混淆，此外也需要考虑到用户可能产生的误操作，例如，模型类型将回归选择成了分类，由于回归模型的目标列往往是连续的，如若误操作选择成了分类模型，此时模型的输入会将所有的输入数据当作一个类别，从而产生几千上万个“分类类型”，如此可能导致系统性能的高负荷运转，甚至系统崩溃，在影响性能的同时，也使得用户的使用体验变得尤其糟糕。为了解决这个问题，数据建模引擎在代码运行前会对模型构建参数进行检查，这里我们预定义了如下模型检测规则:当检测到用户发起代码生成请求时，首先校验数据的样本总数否大于 50。如果样本小于 50 则提示需要收集更多的样本数据，以确保训练的模型有着足够的输入，从而获得有着良好效果的模型。如果样本数量大于 50， 则需要先判断预测的目标列的数据分布，即数据分布是离散型或者连续型。如果数据是离散型或者是标签类型数据，则判定该目标列是用于分类模型；若数据是连续型，则判定该目标列是一个用于回归模型的特征。当用户上传的数据集没有标签列，则提示用户使用聚类方法找到合适的标签，从而进行后续模型的构建。

当代码生成器运行生成了代码，引擎主要负责运行代码，并根据运行结果生成一份报告，用以展示模型结果。当有多个模型存在时，展示各个模型的结果对比，通过可视化数据表格的形式展示。

由于生成的代码从使用者的角度出发，在用户拿到生成的代码时，读取数据的步骤是通过加载本地数据文件到内存，进行后续建模操作。考虑到数据建模引擎运行在后端，可通过数据库直接读取数据集文件，达到节省服务器资源开销的目的，因此，引擎运行文件前会修改原始生成代码的数据源，即从本地文件系统加载修改为数据库读取。数据建模引擎的工作原理如图5-4所示

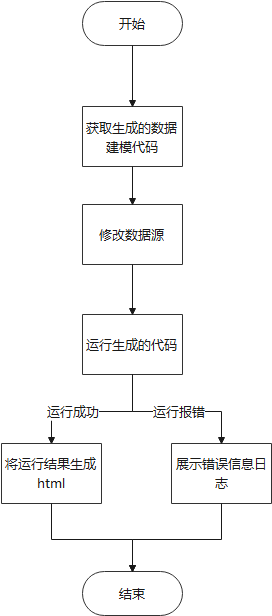
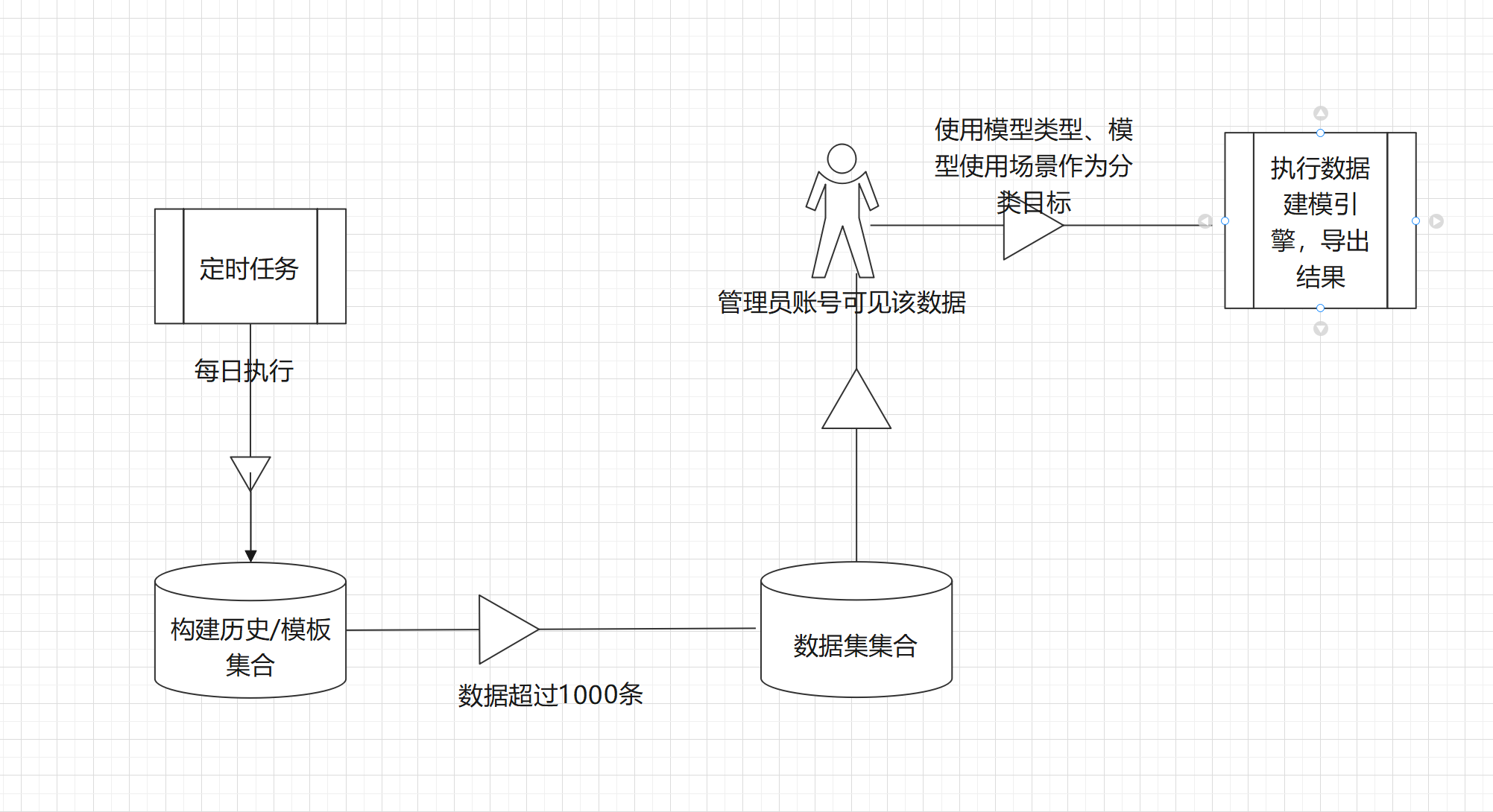


图5-4数据建模引擎流程图

## 常用模型参数模板化/推荐

由于当前系统并没有真实用户使用，因此对于每种模型的构建的频次 暂时无法预知。因此在当前系统实现过程中，在模型构建的地方已提前做好埋点统计，每当用户生成一次代码并执行导出代码操作时，系统会将此次用户的模型构建参数存入数据库中的模板集合。当数据量增长到一定程度时，通过将数据库中的模板集合导出为csv文件，将该csv文件作为一份新的数据集推送至管理员用户的数据集管理界面。通过使用数据建模，将构建学习模型，达到能够根据用户数据集识别出适配的模型。上述过程通过系统预定义定时任务来执行。定时任务每日执行一次，当数据集模板集合数量达到1000条时触发。工作流程如图5-5所示



**图5-5常用模型参数学习流程图**

根据模型的结果在后期加入数据建模界面定义快速填充参数下按钮，从而使得系统形成闭环，达到推荐参数，智能选择参数的效果。