第十六届中国大学生服务外包创新创业大赛

【A？】名称

项

目

详

细

方

案

队伍：奇迹编码团队

按报名顺序

目录

[1. 前言 1](#_Toc30052)

[1.1研究意义 1](#_Toc30140)

[1.2当前亟待解决的问题 1](#_Toc6777)

[1.3创意突破 2](#_Toc11174)

[2.问题分析与解决思路 3](#_Toc26599)

[2.1问题分析 3](#_Toc23036)

[2.2解决思路 3](#_Toc27811)

[2.2.1详细解决思路 4](#_Toc8618)

[2.2.1使用的方法和技术 4](#_Toc1167)

[2.3总体设计 5](#_Toc15145)

[3.数据预处理与特征降维 6](#_Toc2543)

[3.1数据预处理 6](#_Toc21648)

[3.1.1.缺失值处理 6](#_Toc25536)

[3.1.2异常值处理 7](#_Toc27943)

[3.1.3基于Smote算法的样本平衡与上采样 8](#_Toc24447)

[3.2特征降维 8](#_Toc637)

[3.2.1基于曼慧尼特检验的特征分析 9](#_Toc31327)

[3.2.2基于Spearman相关性分析的特征筛选 11](#_Toc17239)

[3.2.3基于包装法的特征筛选 12](#_Toc7428)

[4基于XGBoost算法的多层次医疗保险欺诈识别融合模型 13](#_Toc24306)

[4.1反常识检测层 14](#_Toc9784)

[4.2总体差异检测层 15](#_Toc31593)

[4.3深度分类层 18](#_Toc17468)

[5.医疗保险欺诈实时监测模型 20](#_Toc29762)

[5.1设计API获取数据 20](#_Toc22941)

[5.2数据获取 21](#_Toc19057)

[5.3实时数据监测 21](#_Toc20921)

[6评估模型与消融实验 22](#_Toc14666)

[6.1评估指标 22](#_Toc10832)

[6.2消融实验 23](#_Toc16445)

[6.3实验结果和分析 23](#_Toc25315)

[7.Web设计与GUI展示 23](#_Toc12124)

[7.1Web系统需求分析 24](#_Toc18340)

[7.2数据库设计 25](#_Toc29926)

[7.2.1实体的构建与设计（CDM） 25](#_Toc32064)

[7.2.2物理模型的构建（PDM） 25](#_Toc7292)

[7.2.3数据字典 26](#_Toc7134)

[7.3系统功能WEB展示 26](#_Toc13478)

[7.3.1医疗保险欺诈识别模块 27](#_Toc23283)

[7.3.2医疗保险欺诈监测模块 28](#_Toc13775)

[7.3.3用户管理模块 30](#_Toc7393)

[7.3.4系统评估模块 32](#_Toc7657)

[7.3.5帮助模块 33](#_Toc24707)

[8.系统测试 34](#_Toc20048)

[9总结与展望 34](#_Toc4107)

# 

# 前言

## 1.1研究意义

近年来，国家对民生越来越重视，医疗保险也成为重中之重，2021年全国医疗保险总支出24043.10亿元，比上年增长14.3%，2022年全国医疗保险总支出24597.24亿元，比上年增长2.3%，但随着信息技术的日新月异，出现了一些针对医疗保险报销系统的技术，医疗保险欺诈问题日益突出。同时，一些个人或机构为了获取不当利益，通过虚假诊疗、编造病历、套取医保基金等方式，进行医保欺诈，获取利益。仅2019年，国家医保局共组织69个飞行检查组赴30个省份，对177家定点医药机构进行检查，共查出涉嫌违法违规金额22.32亿元。这些行为违法行为严重影响了国家的医疗保险政策，影响了医疗保障的公平性和可持续性，甚至危害了正常的医疗秩序。有必要设计实时，安全，智能功能的医疗保险欺诈识别监测系统。

## 1.2当前亟待解决的问题（痛点）

当前使用的的医疗保险欺诈识别与监测系统大多采用个人医疗消费记录关联的模式进行欺诈识别，将不符合个人日常消费行为的记录定义为异常医疗消费记录并予以报警，这种方法存在极大的误判率，并且这种方法所输出的结果中，无法直接显示用户被定义为欺诈的原因。我们设计的系统拟实现对医疗保险欺诈行为的识别和实时监测，同时给出详细的欺诈原因。当前系统主要有以下问题需要解决：

（1）医疗保险消费数据维度高，严重影响欺诈识别效率。医疗保险消费数据当中往往具有大量的消费条目，这些消费条目在监测过程中可能无记录，有空值。当这些条目的数据无法获取时就会对识别和监测工作带来极大的不便，因此如何动态灵活的使用现有特征是医疗保险欺诈识别和监测系统的一大难题；

（2）数据质量不高，异常缺损严重。医疗保险消费数据基本都是由人工录入存储系统，而人为精力有限，时常会出现录入错误和缺失的现象，因此如何克服医疗保险消费数据质量较低的问题也是医疗保险欺诈识别和监测系统急需解决的问题；

（3）系统功能有限，无法实时监测：由于医疗保险消费数据在产生时往往具有先后顺序，因此现存的系统只能针对过往的医疗保险消费数据进行欺诈识别，无法做到监测正在进行的医疗保险欺诈行为。故此，使用怎样一种监测机制能够利用少量数据就达到识别欺诈行为的目的也是医疗保险欺诈识别与监测系统的一大痛点。

## 1.3创意突破

本系统融合多种技术，设计了医疗保险欺诈识别监测系统，系统具有如下创意：

（1）多层次数据预处理与分析。我们采用基于Smote数据预处理与曼慧尼特-Spearman-包装法进行多层次特征降维。原始数据仅有794条样本数据具有欺诈行为，仅占总样本数量的4.96%，正负样本数量极其不平衡，因此我们采用smote算法进行上采样，使得正负样本比例达到1:1，以减少模型的偏向性；82维度的原始数据对于实时监测系统来说维度较高，不能满足时效性，因此我们分别采用曼慧尼特检验、Spearman相关性分析、包装法对数据进行降维，最终获得12维的特征数据以此来建立融合模型。多层次降维弥补单一方法特征提取方法的局限性，对于提高模型实时监测性能和解释数据都具有积极的作用。

（2）利用熵权法确定不同特征对欺诈行为的权重。考虑到不同的特征会对欺诈行为的监测和识别具有不同的贡献率，我们利用熵权法对降维后的特征进行权重确定，然后通过包装法获得的特征重要性进行加权求和获得最终评分权重，利用最终评分权重与数据正态分布原则进行异常检测，全面考虑了不同特征的差异性。而确定原始数据列的权重对解决实时监测具有重要意义。

（3）我们根据相关文献和医疗保险消费过程中消费项之间的关联关系，设计了反常识监测层。通过反常识检测与熵权法差异检测，建立基于XGBoost算法的多层次融合模型。为了提高模型的准确性和可解释性，我们在梯度提升决策树算法XGBoost的基础上，增加了反常识检测层与总体差异检测层，以此建立了基于XGBoost算法的医疗保险欺诈识别融合模型。该融合模型欺诈识别准确率高达96.37%，且可以输出欺诈行为的异常信息。

（4）融合数据列的不同权重，构建多层次识别模型，通过API接口获取实时数据实现监测功能。我们的系统能对医疗保险欺诈行为进行实时监测与预警。为了保证监测的准确性，我们结合数理分析中的综合评价、异常监测与权威文献当中依据描述性统计分析的监测方法进行监测模型和功能的设计，而且达到了在仅有极少量数据的情况下仍能实现准确监测的性能要求。

# 2.问题分析与解决思路

## 2.1问题分析

在数据集有限且需要可解释性的前提下，传统的机器学习已经无法满足当今医疗保险欺诈识别的要求。一个实用的医疗保险欺诈识别与监测系统，不仅仅要从数据本身出发，还要考虑到医疗消费与报销关系之间的的协同关联特征。此外，为了减少欺诈行为的发生，全新的医疗保险欺诈识别算法还应具有欺诈预警功能，即在犯罪分子作案过程中，该算法应当需要具备识别并监测到其行为异常的功能。

本系统需要解决以下五个问题：

1. 利用均值填充解决解决数据集存在缺失值的问题。
2. 将脱敏数据与取值单一数据剔除。
3. 利用smote算法进行上采样解决样本不平衡问题。
4. 使用曼慧尼特-Spearman-包装法对82维特征数据进行多层次特征降维。
5. 为了识别医保欺诈行为建立基于XGBoost算法的医疗保险欺诈识别融合模型。

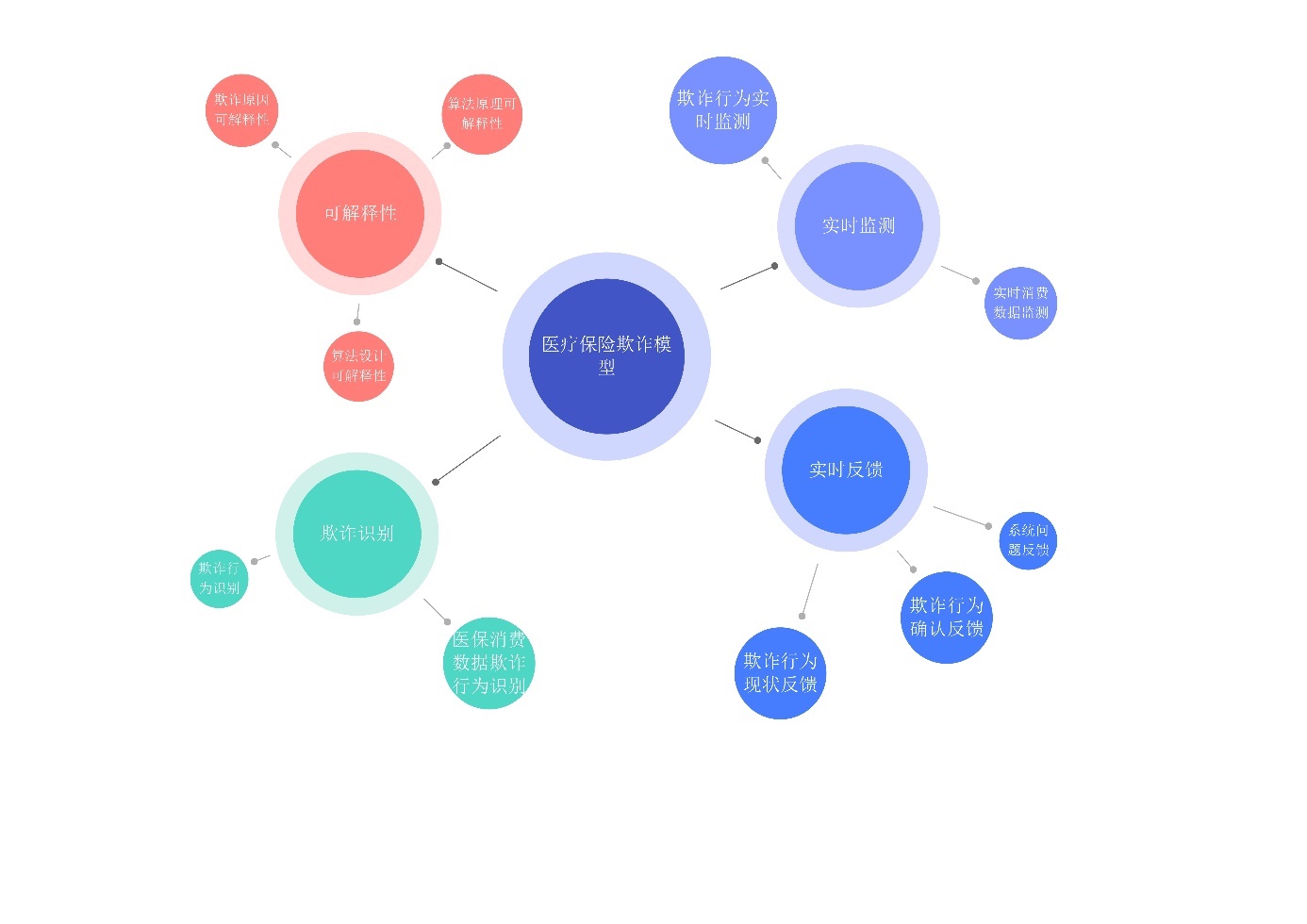


图1系统需具备的功能与特征

## 2.2解决思路

基于以上分析，本项目旨在开发一种全新的能依据少量数据判别和监测医疗保险欺诈行为的系统，系统具有医疗保险报销欺诈和监测功能，并能实时反馈系统识别结果。

### 2.2.1详细解决思路

为了解决这些问题，本项目通过依次进行数据预处理、特征降维、融合模型建模、数据库设计、Web编程，详细的解决思路如图2所示。

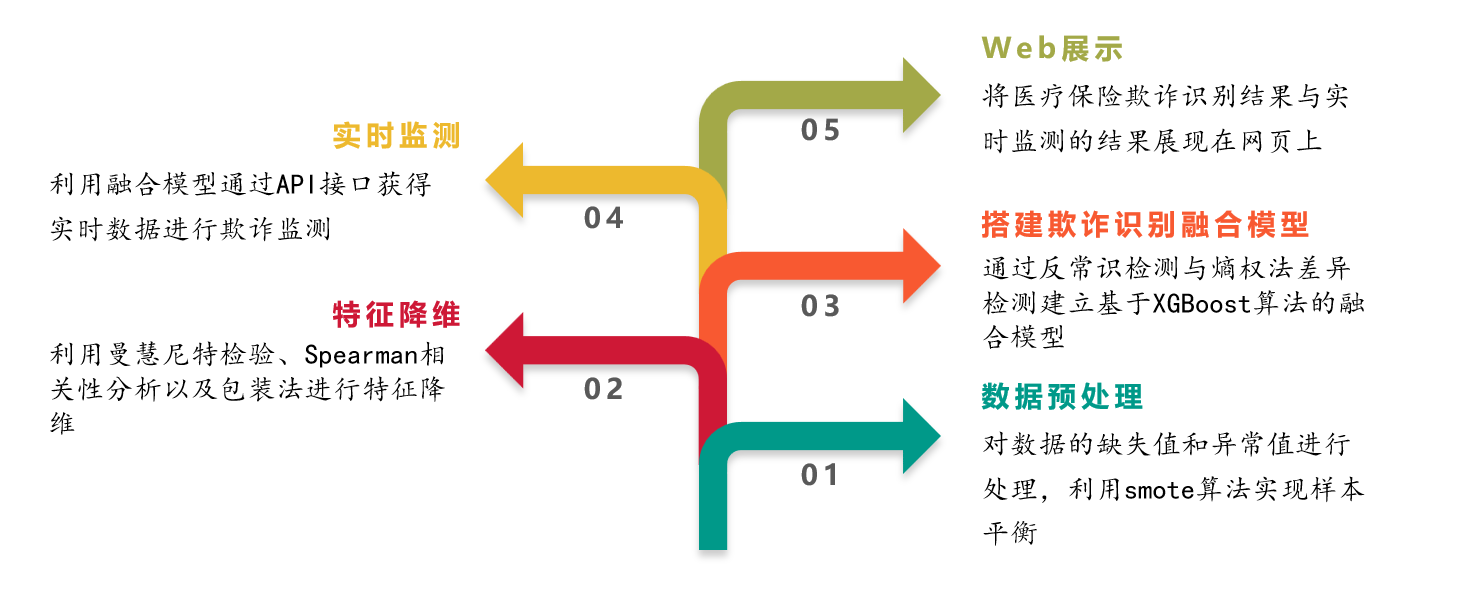


图 解决思路

### 2.2.1使用的方法和技术

基于以上结局思路，我们搭建了医疗保险欺诈识别与监测系统，系统的技术路线如图3所示。

（1）数据分析方法。对脱敏数据与缺失值数据进行处理，利用smote算法进行上采样将样本平衡，使用曼惠尼特检验U型差异性检验、Spearman相关性分析和包装法进行特征降维。

（2）人工智能算法。通过反常识检测与熵权法差异检测建立基于XGBoost（极限梯度提升）算法的融合模型。

（3）数据库技术：使用mysql进行数据存储，创建了四个数据库分别是医保部门，管理部门，原始医保数据，特征压缩后的医保数据。

（4）Web技术。然后AJAX进行前后端数据传输以及页面展示。

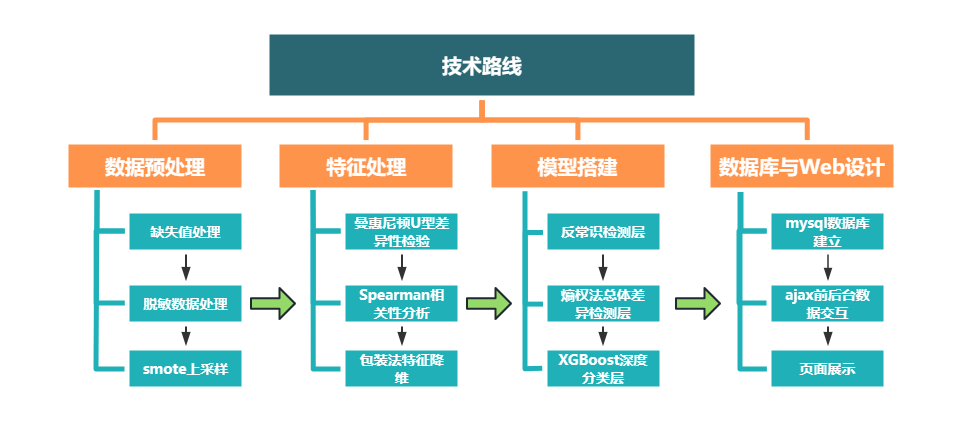
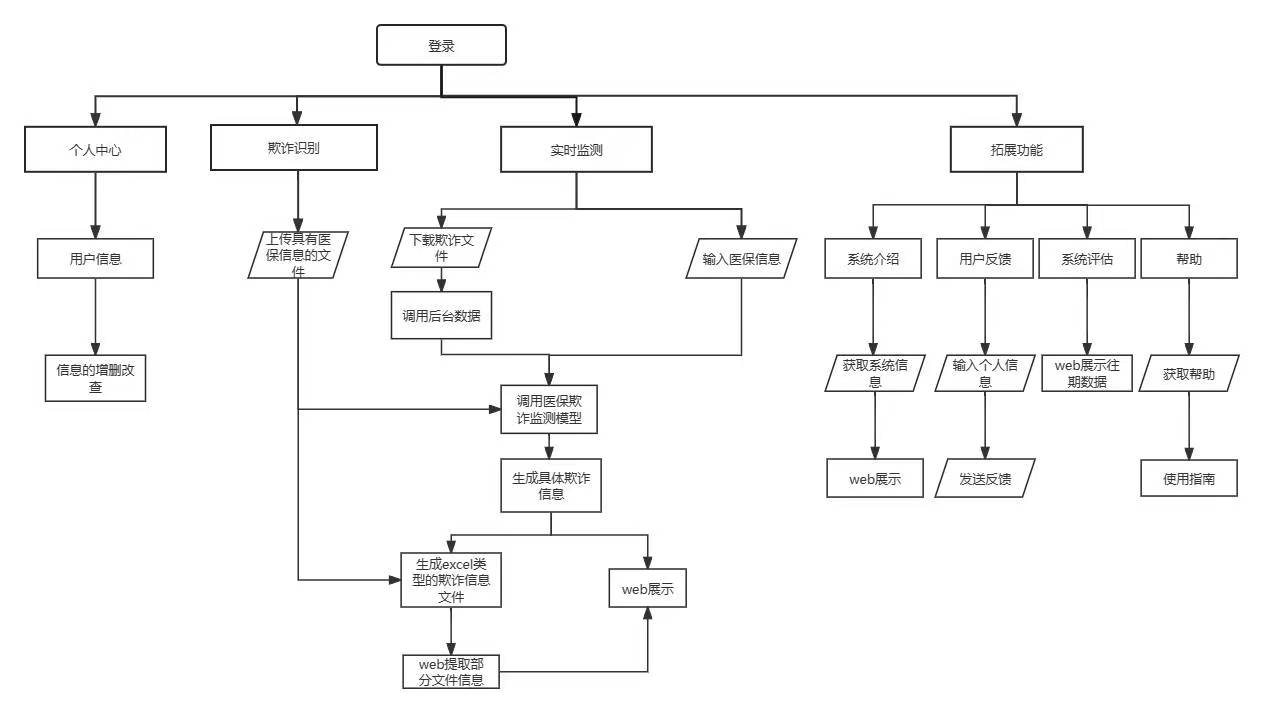


图3技术路线图

## 2.3~~总体设计~~

项目的具体设计流程为：首先对官方给16000条医疗保险消费数据进行数据预处理，然后对82维特征进行特征探索，最终基于确定12维特征建立层级模型。然后通过flask+vue+springboot技术实现算法模型与前后端的交互，用户可以上传需要识别的医疗保险消费数据文件进行欺诈识别，也可以在实时监测模块查看医疗保险欺诈的实时信息。

具体实现流程通过：异常特征的筛选与处理、进行样本平衡与上采样、对特征进行差异性检验、基于Spearman相关性分析的特征筛选、基于包装法的特征筛选、建立反常识检测层、建立总体差异检测层、建立深度分类层模型、将系统功能模块进行扩充，实现系统功能多样化，最终进行数据库与Web系统设计。本项目总体设计如图所示：



图总系统设计图

通过Mapper、Service 和 Controller 三层架构，依靠 MyBatis 实现 CRUD 操作。本系统基于 Vue + Spring Boot，通过前后端分离技术实现前后端数据交互。以 Spring Boot 为框架，后端使用 Java 更好地契合了 Spring Boot 项目的构建。此外，本系统使用 MySQL 进行数据存储，并利用了Python 算法编写的医保欺诈信息识别分析。

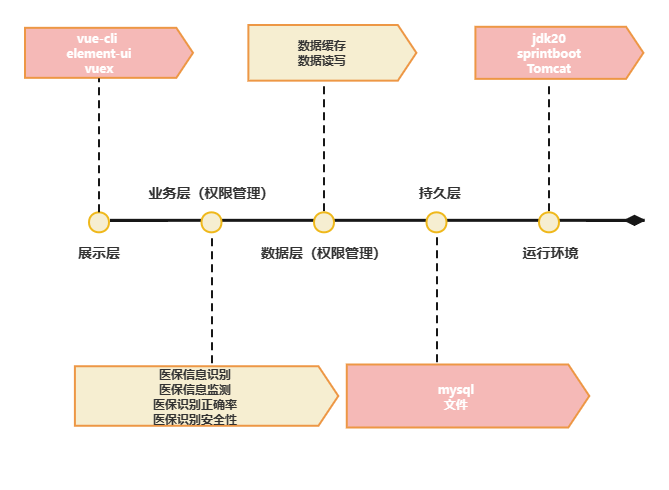


图5系统总体系结构框架图

# 3.数据预处理与特征降维

在机器学习等方法中，数据的质量尤为重要，好的数据质量将会决定模型设计的上限，因此在进行算法设计之前，首先需要对原始16000条医疗保险欺诈数据进行数据预处理。以保证原始数据当中的缺失值、异常值都能通过一定方法转化为合理的数值。同时，过多的特征维度也将导致模型在实际使用的时候受到数据量的限制，因此尽可能减少特征数量也是数据预处理需要解决的一大难题。

## 3.1数据预处理

经过初步统计，原始数据当中存在794条缺失数据，这些缺失数据会导致模型在训练和测试过程中因出现缺失值而删除缺失值所在的样本，进而影响实验结果，因此在进行算法设计前需要进行缺失值处理。

### 3.1.1.缺失值处理

我们首先对原始数据进行了数据探索，经过统计，原始数据中的不同维度特征空缺值如表1所示。

表原始数据缺失值

|  |  |  |  |
| --- | --- | --- | --- |
| 缺失值所在特征维度 | 缺失值数量 | 欺诈样本缺失值数量占比 | 为欺诈样本缺失值数量占比 |
| 出院诊断LENTH\_MAX | 355 | 7.89% | 92.11% |

从表1中可以看出“出院诊断LENTH\_MAX”这一特征列具有共计355条空缺值，其中欺诈样本数量占不到10%。

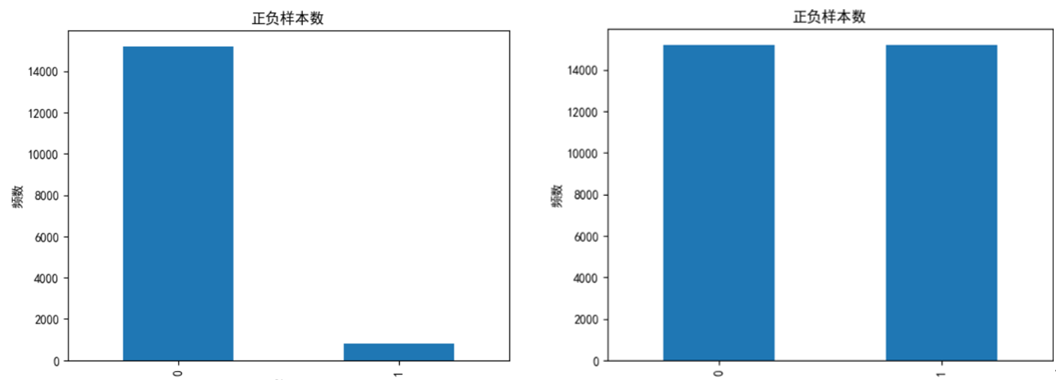
为了保证后续数据处理的准确性并尽量降低干扰，我们对欺诈样本和非欺诈样本的孔雀之分别取欺诈样本群体的平均值、非欺诈样本群体的平均值对其进行填充。

### 3.1.2异常值处理

在原始数据当中，有部分数据由于涉及到个人隐私，因此在企业给出这些数据之前，首先对其进行了不可预知的处理，让其无法反映原本的真实信息，因此在正式进行缺失值处理之前，我们经过相关文献求证，原始82维特征当中，有如表2所示的特征维度为脱敏特征。

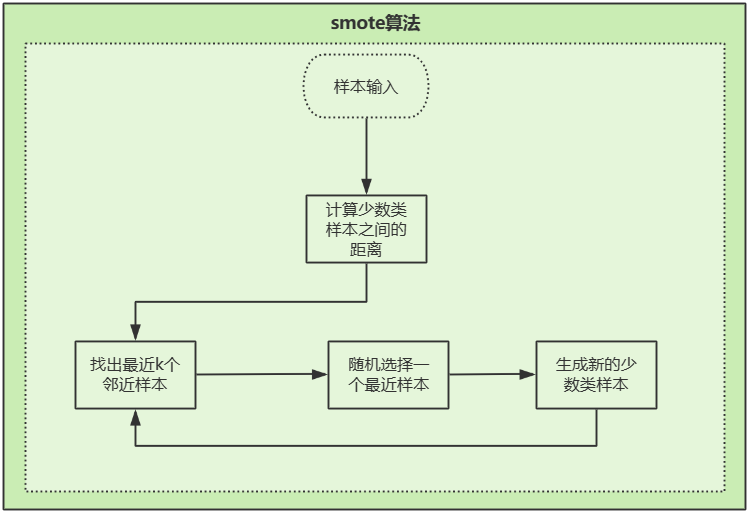
### 3.1.3基于Smote算法的样本平衡与上采样

如图6所示，原始数据汇总仅有794条样本数据具有欺诈行为，仅占总样本数量的4.96%，正负样本数量极其不平衡，这可能导致识别模型偏向于正样本，因此为了减少模型的偏向性需要对样本进行上采样，以保证正负样本数量基本满足1：1。



**图6****基于smote算法采样前后正负样本数量对比**

由于医疗保险数据的特殊性，一般途径很难确保获取到正确的医疗保险数据集，因此我们采用smote算法，对原有标识为欺诈的样本进行样本扩充，具体的样本扩充流程如**图7**所示。



**图7基于smote算法的样本上采样**

经过以上流程，我们最终将欺诈样本数量由794条扩充至15207条，样本总量因此也由16000条扩充至30415条。

## 3.2特征降维

原始数据中，共有特征维度82维，对脱敏和取值异常维度进行处理之后，仍剩余75维特征，针对这个75维度的特征，并不是每一次医疗保险消费的过程都能够获取到相关的数据，很有可能会存在部分数据缺失的情况，因此在进行分类算法的构建之前，首先需要针对预处理之后的数据进行数据特征探索，以尽可能合理的减少数据的特征维度。

为了实现特征降维，我们依次采取了曼惠尼特U型检验、Spearman相关性分析、包装法特征筛选。经过以上步骤，特征列数量变化如图8所示。

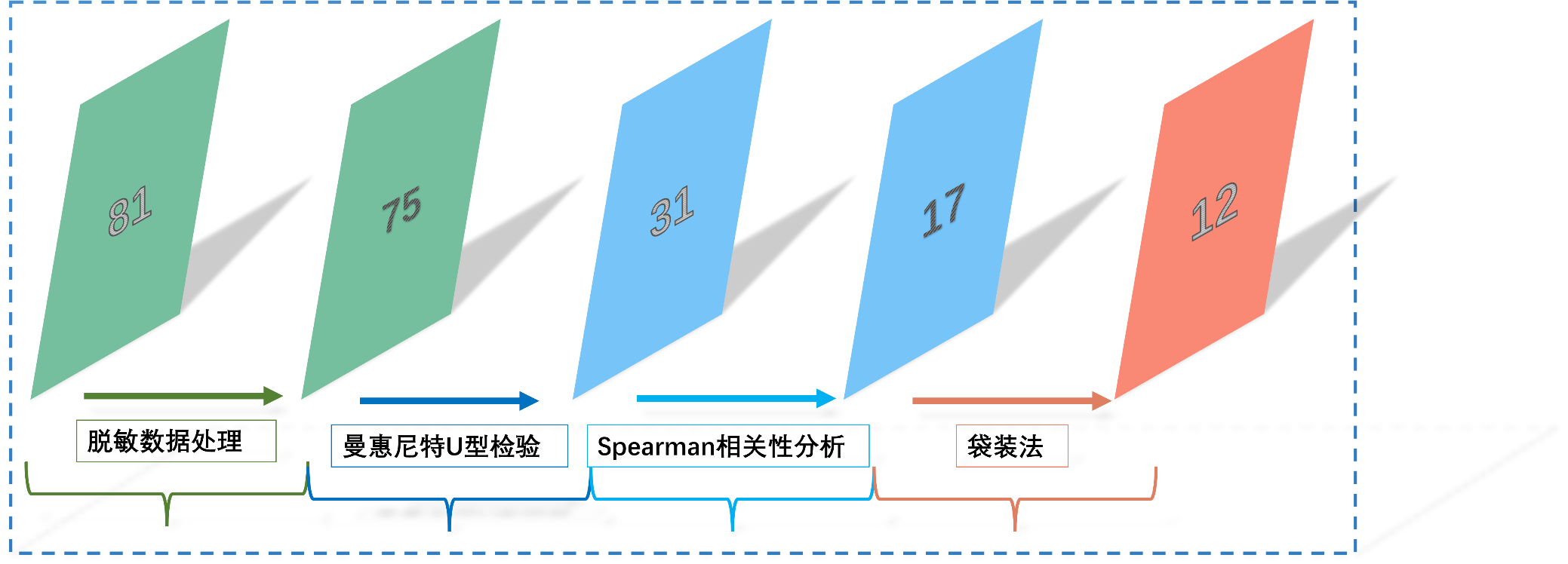


图8 特征列的变化

### 3.2.1基于曼慧尼特检验的特征分析

在原始数据文件中，列res表示是否有欺诈行为，其余列则为患者的相关消费和报销信息。为了保证所筛选的特征能够尽可能多的提供识别欺诈行为相关的信息，我们首先对所有共计82维度的特征进行描述性统计的分析，查看其在标识为欺诈和未欺诈的样本中，是否存在显著的差异性。

### 3.2.2基于Spearman相关性分析的特征筛选

为了进一步对这31维度的特征进行融合与分析，我们采用Spearman相关性分析对这31维度的特征和是否欺诈列（res）进行关联分析，筛选关联性强的特征作为进一步研究和分析的对象。

Spearman相关性分析能衡量两个变量之间关联程度的非参数统计方法，通过对原始数据进行排序，然后根据这些排序值来评估两个变量之间的相关性，而不直接基于原始数据的值 。Spearman相关性分析的性能针对数据类型多样的数据尤为适用。其具体计算公式如下所示。

|  |  |
| --- | --- |
|  | （1） |

其中为Spearman相关系数，是两个特征的等级之差，是样本点的数量。

通过Spearman相关性性分析后，我们选取相关性大于0.1的特征维度进行进一步的探索，通过相关性分析，最终筛选出目标特征如 所示（完整分析结果详见附件一：Spearman相关性分析结果）

表4关联性较强的特征

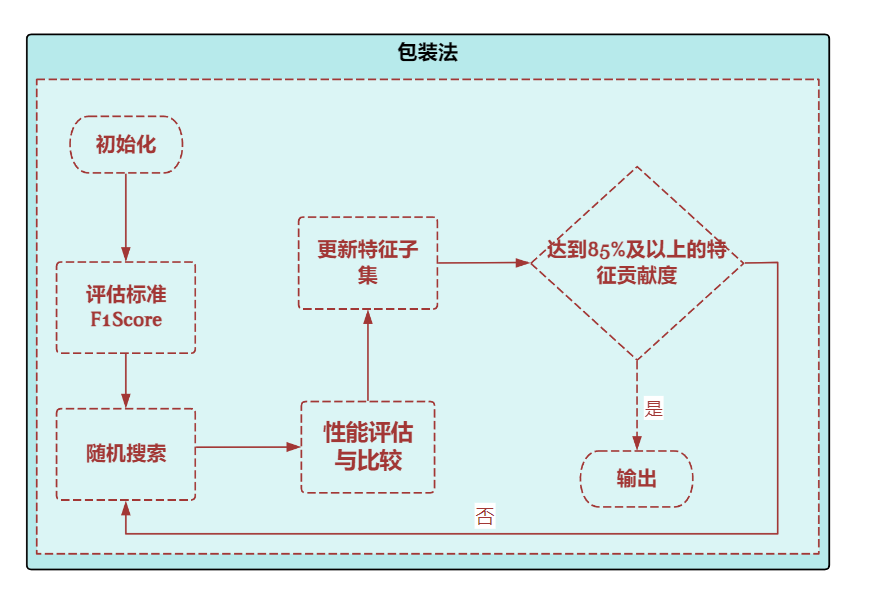
|  |  |
| --- | --- |
| 特征名 | 与是否欺诈的相关性 |
| 一天去两家医院的天数 | 0.129（\*\*\*） |
| 月就诊天数\_MAX | 0.156\*\*\* |
| 月就诊天数\_AVG | 0.128\*\*\* |
| 就诊次数\_SUM | 0.1\*\*\* |
| 月就诊次数\_MAX | 0.159\*\*\* |
| 月就诊次数\_AVG | 0.132\*\*\* |
| 月统筹金额\_MAX | 0.173\*\* |
| 月统筹金额\_AVG | 0.129\*\*\* |
| 月药品金额\_MAX | 0.134\*\* |
| 医疗救助个人按比例负担金额\_SUM | 0.297\*\*\* |
| 起付标准以上自负比例金额\_SUM | 0.108\*\*\* |
| 个人账户金额\_SUM | 0.113\*\*\* |
| 可用账户报销金额\_SUM | 0.111\*\* |
| 非账户支付金额\_SUM | 0.114\*\*\* |
| 医疗救助医院申请\_SUM | 0.298\*\*\* |
| 治疗费申报金额\_SUM | 0.115\* |
| 治疗费发生金额\_SUM | 0.114\*\* |
| 注：\*，\*\*，\*\*\*分别表示90%，95%，99%的显著水平 | |

经过Spearman相关性分析，我们进一步将原始特征维度减少至17，大大减少了模型的运算量和数据的依赖程度。

### 3.2.3基于包装法的特征筛选

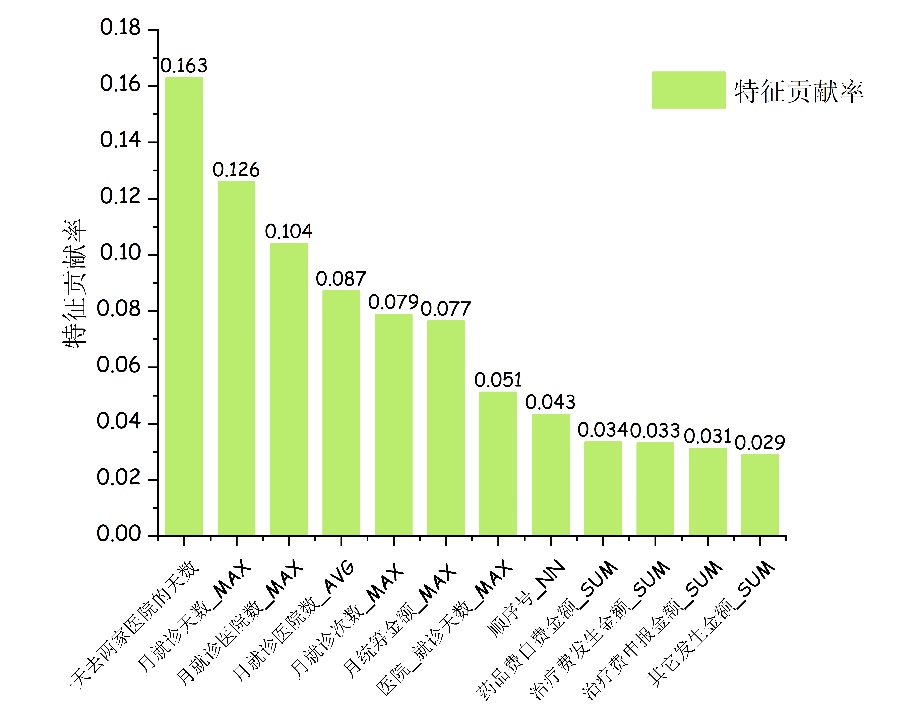
经过Spearman相关性分析差异性检验与相关性性分析后，特征维度已经降低到了17维，但特征维度仍然较高，并且由于医疗保险数据实时采集的特殊性，有时候很难一次性采集到所有有关的数据，因此我们仍然需要进一步地减少特征维度，同时，还保证原始数据的信息不会被遗漏太多，确保算法的灵活性和实用性。经过前面的筛选，我们已经获得17维在欺诈样本和非欺诈样本间能够提供较多分类信息的的特征，我们进一步使用包装法对这17维特征进行进一步筛选。

我们应用包装法对这17维特征对医疗保险欺诈识别与检测任务的特征重要性进行分析，并通过搜索不同的特征组合，使用模型评估每种组合的效果，以选择最佳的特征子集。具体的算法流程图如图9所示



图包装法特征筛选算法流程

结果分析经过包装法后，我们最终获得特征重要性累计大于85%的特征重要性占比如图10所示。

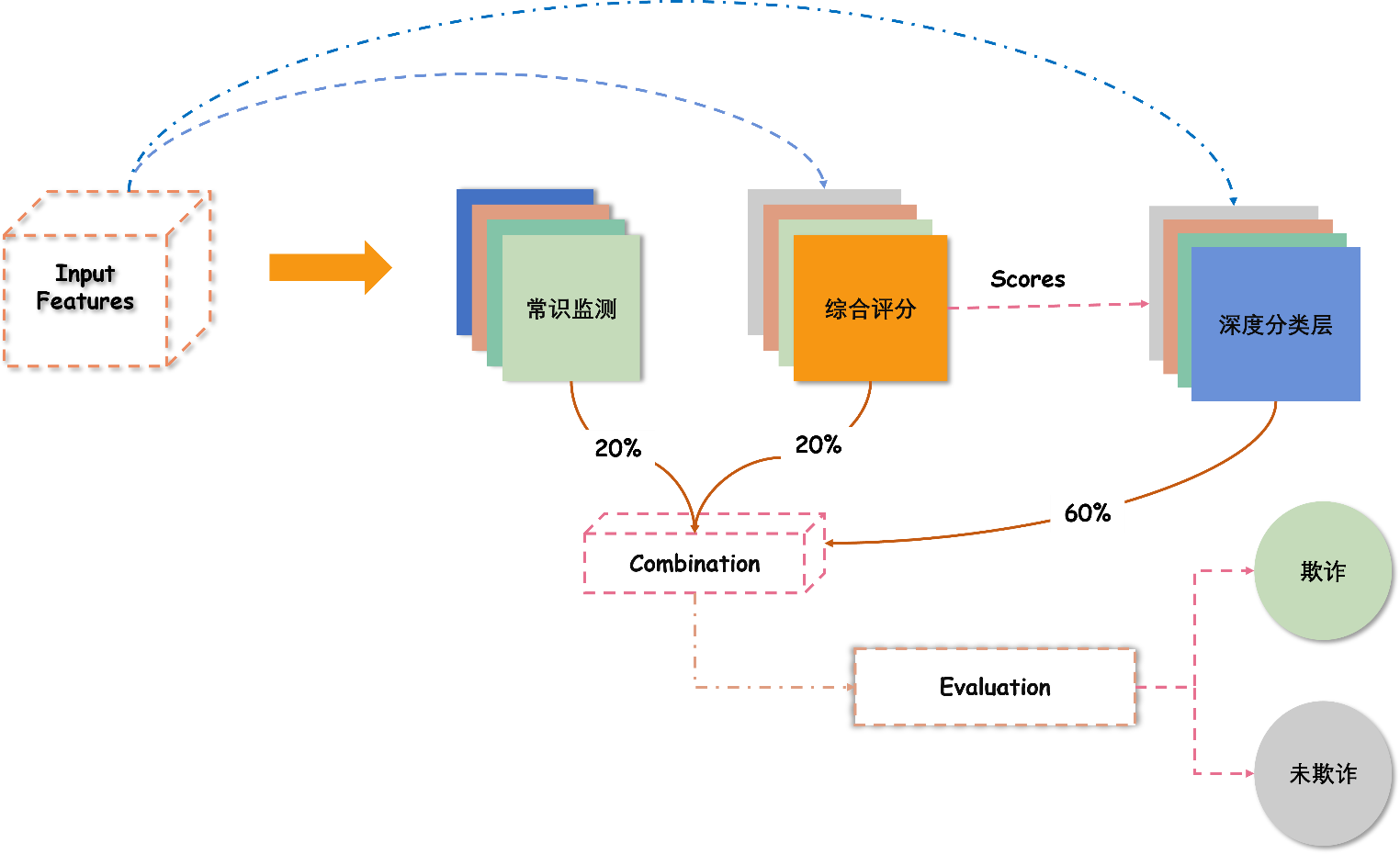


图特征重要性累计超过85%的特征

经过包装法筛选，我们从原始的数据特征中最终筛选出12维度的特征作为构建欺诈识别算法的依据。

# 4基于XGBoost算法的多层次医疗保险欺诈识别融合模型

如上所述，经过数据分析和预处理后，我们获得了共计12维度的原始数据作为源数据的特征因子，由于医疗保险欺诈的特殊性，单纯使用传统深度学习和机器学习基于大量数据集学习数据集分布特征的方法可能很难识别新型的欺诈手段。我们针对原始数据的分布特征，提出了一种基于XGBoost算法的多层次的医疗保险欺诈识别融合模型。我们的融合模型依赖多个层级或阶段的数据处理和决策过程，在每一个层级，模型专注于分析数据的某一个特定方面或者维度，然后将结果传递给下一层。经过查阅大量文献后，我们最终确定算法层级为3。三个层级分别为：反常识检测层、总体差异检测层、深度分类层。具体层级构建如图11所示。



图基于XGBoost算法的医疗保险欺诈融合模型网络结构图

## 4.1反常识检测层

通过对原始数据中的维度特征进行分析，我们发现了许多相关联的特征，比如病人若进行了手术，即存在手术费用，那么将必然会存在床位费和住院费；相反，若一个人进行了手术，但是却没有床位费和住院费，这就说明这位报销者的医疗报销将会具有一定的反常识特征，而这种特征往往无法被现有的深度学习或机器学习模型捕捉，因此我们将其全部统计出来并设计成反常识检测层。报销者的每一条反常识行为都将会为其增添20%的第一层欺诈概率。根据反常识行为，确认反常识检测层。反常识层的具体检测规则与依据如下：

①手术费发生金额\_SUM与一次性医用材料申报金额\_SUM、床位费发生金额、住院天数。众所周知，如果病人进行了手术，那么势必会产生相对应的材料费和床位费，一般的小型手术在结束之后都需要住院消炎，并且在手术期间将会使用大量的一次性用材，这些一次性用材也将会纳入到医疗保险的报销记录单里面。因此，只要用户进行了手术，即存在手术费发生金额，那么势必会产生对应的床位费和住院天数以及一次性材料的使用费用，如果存在部分小型手术，那么这些相关联的项中也必然会出现至少一项；

②贵重药品发生金额\_SUM、贵重检查费金额\_SUM与是否挂号。病人在进行贵重检查之前，必须首先经过医生开具相关的检查项目单，而医生开具相关的检查项目单则大部分情况下会需要患者首先进行挂号，如果患者没有进行挂号，则这条消费记录很有可能存在欺诈行为；

③个人账户金额\_SUM和基本个人账户支付\_SUM。如果病患的个人账户可支付金额低于账户的支付金额，这说明患者存在很大的嫌疑开具了非法的单票和结算记录，因此很有可能存在非法骗保的行为；

④可用账户报销金额\_SUM和检查费申报金额\_SUM、治疗费申报金额\_SUM等申报金额。如果患者的可用账户报销金额\_SUM小于了各项申报的金额，很有可能是欺诈者不了解账户本身可用于报销的金额数量，从而导致的违背常识性的错误。如果出现此类现象，则很有可能会是欺诈者在进行医疗保险欺诈。

⑤成分输血申报金额\_SUM和床位费发生金额。只要进行了成分输血或者输血，那患者必然将会需要在医院进行正规的受血流程。按照正规的受血流程，患者必须在医院的住院区或者观察房中进行输血，而这也将势必产生床位费，如果患者产生了输血费用但是没有产生床位费，这说明患者很有可能并没有进行输血，此条报销记录很有可能存在欺诈的行为；

⑥就诊的月数和医疗救助医院申请\_SUM。经过对欺诈样本的单独分析，我们发现，欺诈样本中都存在一个很明显的现象，那就是就诊月数多为6月并且和医疗救助医院申请存在一定的关系。因此，如果患者的就诊月数大于4月的同时医疗救助医院申请过高或者过低，都有可能预示着患者报销数据的不正常，该条消费记录可能存在欺诈的行为。

⑦月就诊天数\_MAX和就诊次数\_SUM经过关联分析发现，存在欺诈行为的消费记录中，月就诊天数的最大值和就诊次数之间具有一定的关系，如果就诊次数超过了就诊天数的3倍，则很有可能存在欺诈行为。

构建好以上反常识检测原则后，当患者的消费记录中，存在一条满足，则为其增加20%的欺诈概率，患者所有的消费记录经过这一层之后将会得到一个累计欺诈概率，该概率将会以20%的比例纳入最终的欺诈概率当中。

## 4.2总体差异检测层

经过第一层的常识检测之后，还需要对患者的报销记录进行综合的评价，如果用户的报销记录综合评分低于或着超过一定的数值，则很有可能是存在刻意控制报销评分的行为，有可能存在欺诈行为。我们首先利用熵权法对特征的重要性进行计算，然后对其记性加权求和，最后进行正态分布异常值判断。

1.基于熵权法与特征重要性的权重计算。为了给患者的报销记录进行更加合理的评分，我们首先需要对医疗保险报销期间的各项特征进行权重分析，为不同的特征设计其对应的权重。从之前的数据预处理和特征分析中我们已经将原始的82维度特征降低至12维，并计算了其位识别欺诈所能够提供的特征贡献度。然后，我们对提取的12维度进行进一步分析，获取其在数据分布层面所能够提供的信息。

熵权法能通过衡量数据指标变异性大小来客观地确定权重，通常一个指标的信息熵越小，表明其指标变异程度越大，提供的信息量越多，在模型中就越重要，因此权重也就越大。在熵权法中，首先需要将数据全部转化为正向指标或者负向指标，由于我们的模型主要用于识别保险欺诈的行为，因此严格意义上并不存在绝对的正向或者负向特征，因此我们将所有的特征都定义为正向特征。首先计算每个特征的信息熵为：

|  |  |
| --- | --- |
|  | （2） |

其中为样本的数量，为归一化后的属性，具体的计算公式如下：

|  |  |
| --- | --- |
|  | （3） |

确定信息熵后，通过信息熵计算各指标（特征）的权重如下：

|  |  |
| --- | --- |
|  | （4） |

我们通过熵权法，最终获得每一个特征的评分权重如表5所示。

表熵权法特征权重

|  |  |  |  |
| --- | --- | --- | --- |
| 特征项 | 信息熵值e | 信息效用值d | 权重(%) |
| 一天去两家医院的天数 | 0.809 | 0.191 | 17.815 |
| 月就诊天数\_MAX | 0.982 | 0.018 | 1.704 |
| 月就诊医院数\_MAX | 0.955 | 0.045 | 4.239 |
| 月就诊医院数\_AVG | 0.945 | 0.055 | 5.088 |
| 月就诊次数\_MAX | 0.971 | 0.029 | 2.724 |
| 月统筹金额\_MAX | 0.98 | 0.02 | 1.836 |
| 医院\_就诊天数\_MAX | 0.979 | 0.021 | 1.931 |
| 顺序号\_NN | 0.98 | 0.02 | 1.893 |
| 药品费自费金额\_SUM | 0.857 | 0.143 | 13.354 |
| 治疗费发生金额\_SUM | 0.831 | 0.169 | 15.738 |
| 治疗费申报金额\_SUM | 0.832 | 0.168 | 15.702 |
| 其它发生金额\_SUM | 0.807 | 0.193 | 17.975 |

为了保证获取到的权重的综合性，我们进一步将其与原始特征的通过包装法获得的特征重要性进行加权求和，将熵权法获得的权重设置为70%的比重，包装法获得的特征贡献率占比30%，最终获得用于综合评价的权重如表6所示

表最终评分权重

|  |  |
| --- | --- |
| 特征项 | 权重(%) |
| 一天去两家医院的天数 | 12.53 |
| 月就诊天数\_MAX | 1.24 |
| 月就诊医院数\_MAX | 3.00 |
| 月就诊医院数\_AVG | 3.59 |
| 月就诊次数\_MAX | 1.93 |
| 月统筹金额\_MAX | 1.31 |
| 医院\_就诊天数\_MAX | 1.37 |
| 顺序号\_NN | 1.34 |
| 药品费自费金额\_SUM | 9.36 |
| 治疗费发生金额\_SUM | 11.03 |
| 治疗费申报金额\_SUM | 11.00 |
| 其它发生金额\_SUM | 12.59 |

获得评分权重后，我们使用加权求和的方式，将每一个样本的综合评分求出，具体求解公式如下：

|  |  |
| --- | --- |
|  | （5） |

其中为权重矩阵，为所筛选出来的12项特征。

（3）获得样本的最终评分后，我们采用数据正态分布原则进行异常判断，当样本的评分处于评分均值之外的区间，则认为用户的行为具有较大的欺诈可能，将为其在第二层增加（3-4）40%的欺诈概率，如果用户的行为在之间，则为其增添60%的欺诈概率，如果在之外，则为其增添第二层100%的欺诈概率。利用以上规则，我们对每一列特征进行异常值判断，使用其异常监测概率乘以权重并求和，作为第二层的欺诈概率，再将样本在第二层的欺诈概率按20%的权重比例累加至最终的欺诈概率。

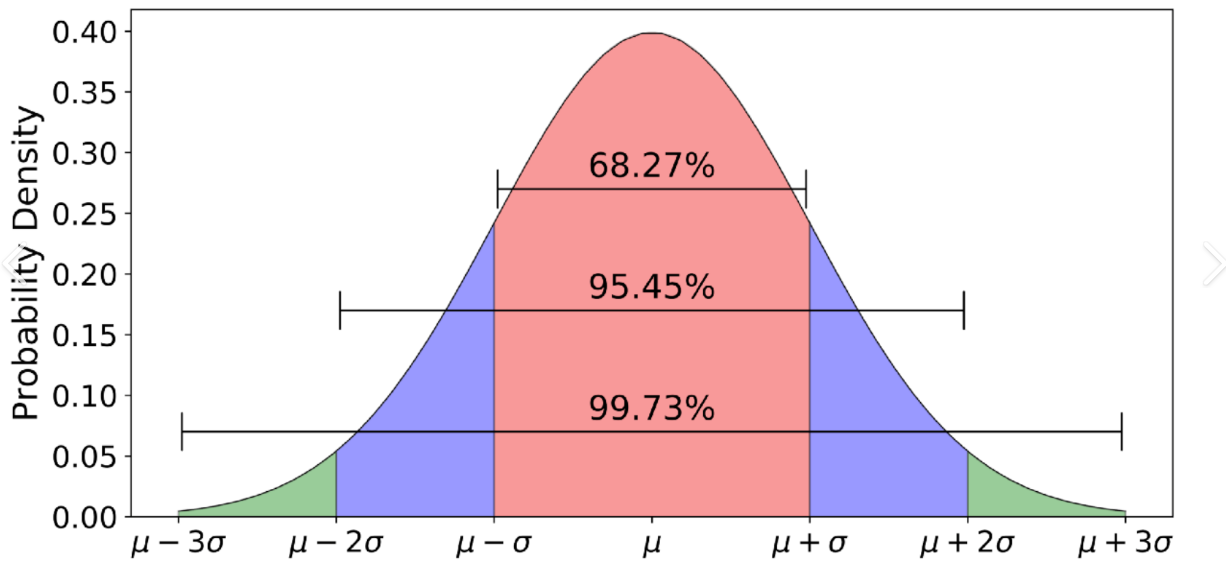
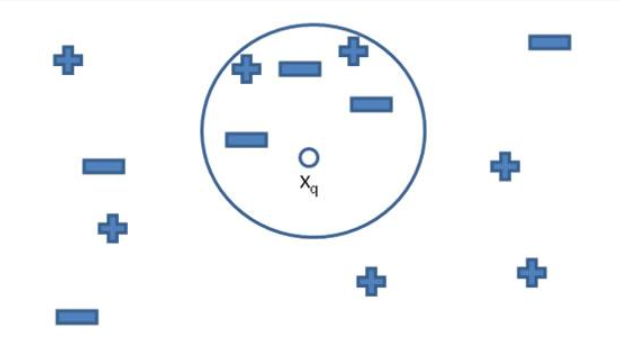


图12正态分布原则

## 4.3深度分类层

为了进一步提取到更加深度层次的特征，在深度分类层 ，我们分别采用了K-近邻算法（KNN）、随机森林（Random Forest）和XGBoost算法进行训练与分析，最终确定XGBOOST作为我们第三层的分类器。

（1）K-近邻算法（KNN）。KNN是一种基本分类与回归方法，其核心原理是给定一个训练数据集，其中的样本已经被标记为属于两个或多个类别或者已经被赋予一个具体数值。对于新的输入实例，在训练集中找到与这个新实例最邻近的K个样本，这K个样本的多数属于某个类别，则新实例也属于这个类别；或者这K个样本的数值的平均值被用作新实例的输出（如图13所示）。



图KNN算法示意图

KNN算法最核心的部分就是距离的度量，在一般的KNN算法中，距离的度量有三种方式，一种是欧式距离，另一种是曼哈顿距离，还有一种是余弦相似度（Cosine Similarity）。其中余弦相似度通常用于度量两个向量的相似度，转换成距离时可以用1减去余弦相似度的值。两外两个度量方式的其具体计算公式如下：

欧氏距离：

|  |  |
| --- | --- |
|  | （6） |

曼哈顿距离：

|  |  |
| --- | --- |
|  | （7） |

综合考虑医疗保险数据的特殊性和欺诈行为的不确定性，我们采用余弦相似度进行距离的度量。

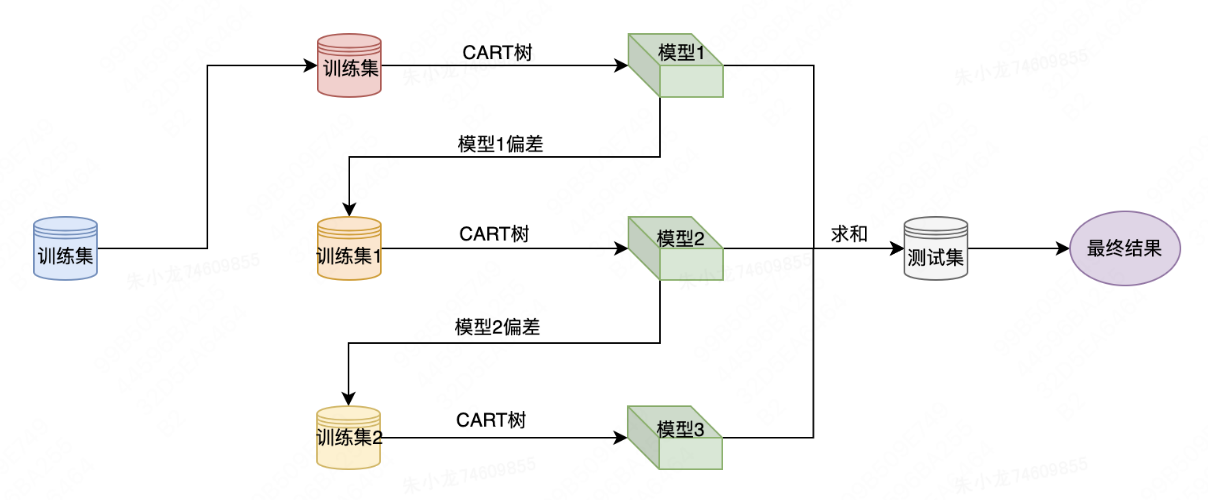
（2）随机森林（Random Forest）。RF是一种集成学习方法，特别用于分类、回归和其他任务，通过构建多个决策树（森林）并输出作为分类的模式或平均预测（回归）的类别。随机森林由多个决策树组成。每个决策树在训练时，会采取随机的样本（bootstrap抽样）和随机的特征来构建树，从而确保树之间的多样性。在进行分类或回归时，随机森林会综合所有决策树的结果，通过投票或平均的方式产生最终的预测结果。

在使用随机森林进行拟合之前我们首先确定部分超参数如表7所示

表部分随机森林算法中部分超参数

|  |  |
| --- | --- |
| 超参数 | 初始值 |
| 学习率 | 0.3 |
| 最大深度 | 10 |
| 最大叶子节点数量 | 50 |
| 节点分裂评价指标 | Gini |
| 内部节点分裂最小样本数量 | 2 |
| 叶子节点中样本的最小权重 | 0 |

（3）XGBOOST是梯度提升树（Gradient Boosting Decision Tree, GBDT）的优化实现，广泛应用于分类、回归、和排名问题中。其核心原理是核心原理基于GBDT，通过逐步添加树模型来最小化损失函数。其算法流程如图14所示。



图XGBOOST算法流程图

利用以上三种方法，我们以筛选出的12维特征为分类识别的特征维度，是否欺诈为识别目标，最终获得每种模型的识别准确率如表8所示：

表分类器准确率

|  |  |
| --- | --- |
| 分类器 | 准确率(%) |
| K-近邻算法 | 81.5% |
| 随机森林 | 92.1% |
| XGBoost | 94.4% |

从表中可以看出，XGBOOST算法的识别准确率最高，因此我们选择该种算法作为我们第三层的分类器。

在最终的欺诈概率计算中，第三层的欺诈概率按照60%的权重比例累加至最终的欺诈概率。最终输出样本是否为欺诈的判断基于反常识检测层20%权重、总体差异检测层20%权重以及深度分类层得到的概率60%权重进行加权比对。融合模型的总体设计流程图如图14所示。

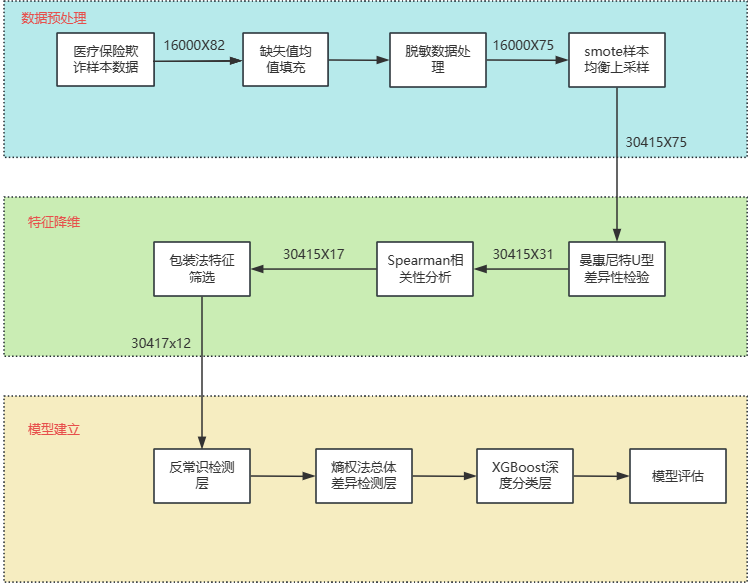


图 数据预处理与融合模型流程图

# 5.医疗保险欺诈实时监测模型

根据4.2中熵权法确定的欺诈特征权重，然后利用第四章中基于XGBoost算法的医疗保险欺诈识别融合模型，通过API接口实时获取医疗保险消费数据实现了欺诈识别实时监测功能。利用基于XGBoost算法的医疗保险欺诈识别融合模型，我们的系统实现了欺诈识别实时监测功能。

## 5.1设计API获取数据

我们向医疗部门提供API接口，通过该接口系统可以实时获取病人医疗保险报销的信息。经过模型分析后，可以得知该用户是否存在欺诈行为，并获得欺诈数据的异常信息，如图 16 所示。

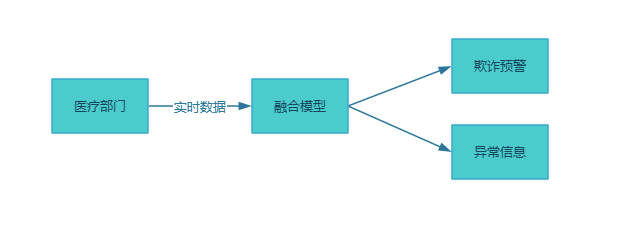


图17实时数据监测流程图

## 5.2数据获取

考虑到系统使用需求，不可能让用户输入完整的82维特征，会让系统使用效率低下，因此我们将筛选出来的12特征提供给用户输入。对用户的每一次输入都进行欺诈监测，一旦发现异常就可以立即预警并且给出异常信息。

由于融合模型至少需要12维特征作为输入，所以当用户输入特征时，我们会利用0值对未输入空值进行填充，确保未填充的值不会对结果造成影响，从而实现对输入数据进行实时预警，并得到预警的原因输出。如图18所示。

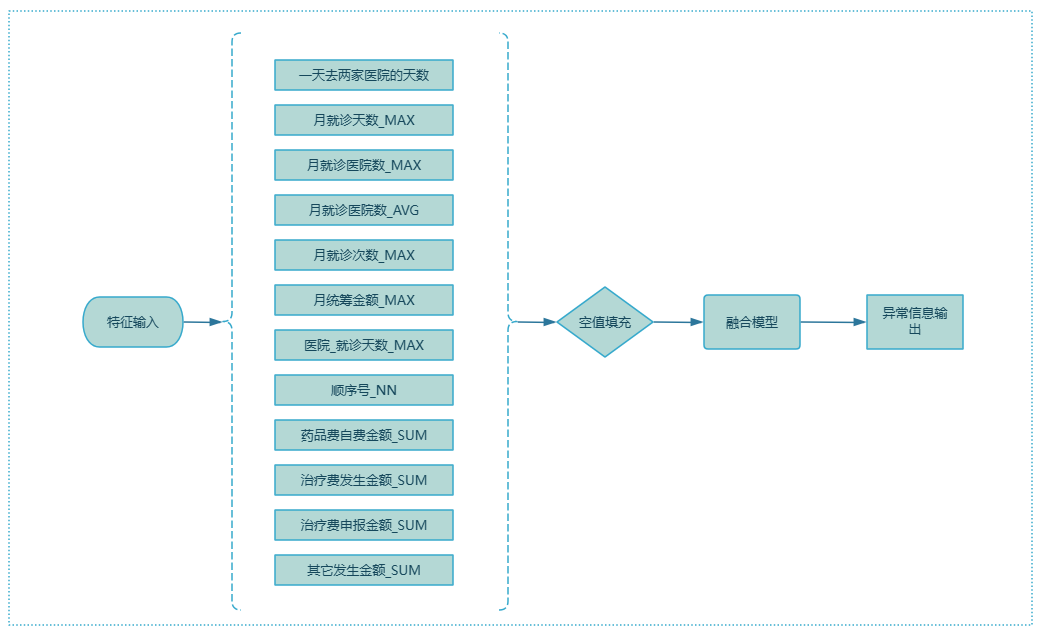


图19手动输入数据监测流程图

## 5.3实时数据监测

通过API接口获得数据之后即可利用构建好的监测算法进行欺诈行为的实时监测。

**6评估模型与消融实验**

为了验证先前的数据预处理、数据降维、模型设计与优化等工作对解决医疗保险欺诈识别与监测所带来的效益，我们进一步对模型训练结果进行评估。

**6.1评估指标**

为了衡量模型的准确率和识别效率，我们采用准确率、精确率、召回率和F1系数来衡量模型的分类效果，并且采用消融实验来评估模型的稳健性。

在表示整个模型的衡量指标之前，我们定义如所示的具有实际含义的符号：

**表9评估指标符号**

|  |  |
| --- | --- |
| 符号 | 实际含义 |
|  | 实际为欺诈，模型预测也为欺诈的样本数量 |
|  | 实际为欺诈，但是模型预测为未欺诈的样本数量 |
|  | 实际为未欺诈，模型预测也为未欺诈的样本数量 |
|  | 实际为未欺诈，但是模型预测为欺诈的样本数量 |

精确率是指所有被预测为正的样本中实际为正的样本的概率，即被预测为欺诈的样本中，真正是欺诈的样本占比，具体计算公式如下：

|  |  |
| --- | --- |
|  | （5） |

准确率是指算法模型的预测样本中，预测正确的样本在所有样本中的占比，即预测未欺诈或者未欺诈的样本中，预测正确的样本数量占比，具体的计算公式如下：

|  |  |
| --- | --- |
|  | （6） |

召回率是指实际为正的样本中被预测为正样本的概率，即实际为欺诈的样本中，被预测欺诈的样本的数量，具体的计算公式如下：

|  |  |
| --- | --- |
|  | （7） |

F1系数是综合衡量精确率、召回率的综合指标，即整个算法在数据集上面的综合表现，具体计算公式如下：

|  |  |
| --- | --- |
|  | （8） |

**6.2消融实验**

为了验证前文所进行的缺失值、异常值、特征降维和模型设计等工作队医疗保险欺诈识别和监测的促进效益，我们针对各项处理和设计之后的数据进行消融实验，并以XGBoost为基准模型，获得消融实验结果如**表10**所示。

**表10 消融实验结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 模型 | 精确率 | 召回率 | 准确率 | F1Score |
| XGBoost+未处理缺失值 | 96.21% | 96.11% | 95.89% | 96.16% |
| XGBoost+未处理异常值 | 96.67% | 96.34% | 96.07% | 96.5% |
| XGBoost+未进行特征降维（85维） | 97.07% | 96.34% | 96.88% | 97.2% |
| XGBoost+进行部分特征降维（75维） | 96.81% | 96.56% | 96.02% | 96.18% |
| XGBoost+进行部分特征降维（31维） | 95.6% | 95.6% | 95% | 94.9% |
| XGBoost+进行完整的数据处理（12维） | 97.11% | 97.82% | 96.37% | 97.09% |

从表中可以看出，在没有进行异常值处理、缺失值处理、降维等操作时，模型的效果相较基准模型XGBoost均有不同程度的下降，这说明我们所进行的一系列数据预处理操作是有效的。

**6.3实验结果和分析**

我们将构建好的XGBoost算法的医疗保险欺诈识别融合模型应用于最终所给定的医疗保险数据，并将通过smote算法的扩充的样本作为实时录入数据传输到模型，最终获得模型的各项检验指标对比如表11所示。

**表11模型评估**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 评估指标 | 精确率 | 召回率 | 准确率 | F1Score | Time |
| 数值 | 97.11% | 97.82% | 96.37% | 97.09% | 0.285ms/sample |

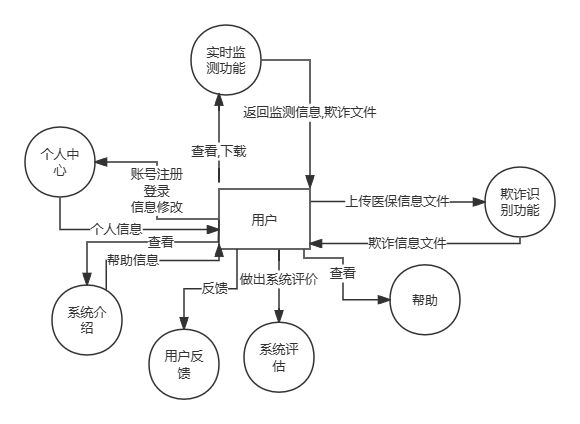
从表中可以看出，基于XGBoost算法的医疗保险欺诈识别融合模型准确率等各项检验指标均呈现较高水平，模型识别效果良好。进一步观察返现，模型识别单个样本仅需0.285ms,模型识别效率高。

**7.Web设计与GUI展示**

我们融合为了实现医疗保险欺诈识别与监测系统所需要的监测和识别功能，并能实时反馈用户系统识别结果。系统由用户管理模块、识别模块、监测模块、系统评估模块、系统帮助模块五个模块构成。

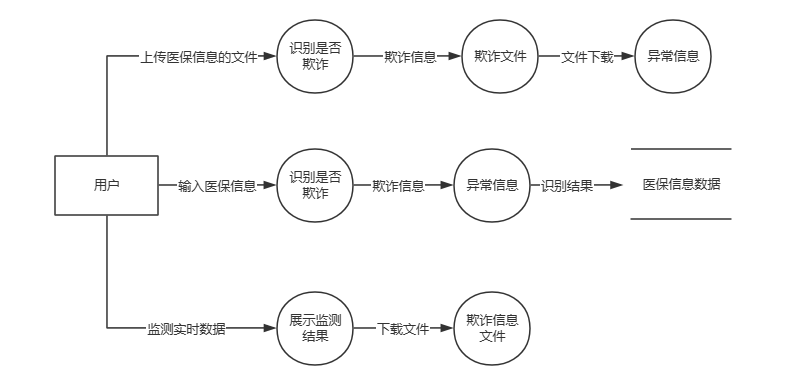
**7.1Web系统需求分析**

通过对医疗保险欺诈识别与监测系统进行需求分析，获得如图20所示的顶层数据流图。用户在个人中心进行注册与登录后，可以上传医保信息文件，通过医保欺诈识别功能获取包含欺诈信息文件，通过医保欺诈监测功能可以实时查看医保欺诈信息。在系统帮助页面获取系统帮助同时也能在系统评估页面对系统进行评价反馈。



**图20顶层数据流图**

进一步对顶层数据流图中的欺诈识别功能与欺诈监测功能进行分解，让其功能更加细化，便可得到0层数据流图如图21所示，识别与监测模块中的数据流向。



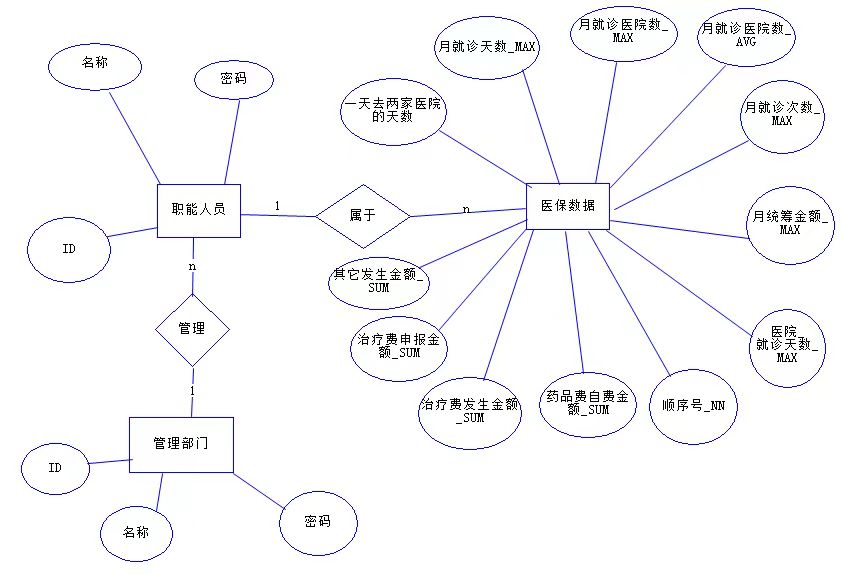
**图21 0层数据流图**

通过对系统数据流的分析可以发现，医疗保险欺诈识别系统除最关键的欺诈识别和监测功能以外，还需要用户的登录管理以及后续的维护模块，并且系统还需要设计多级用户进行管理。

**7.2数据库设计**

**7.2.1实体的构建与设计（CDM）**

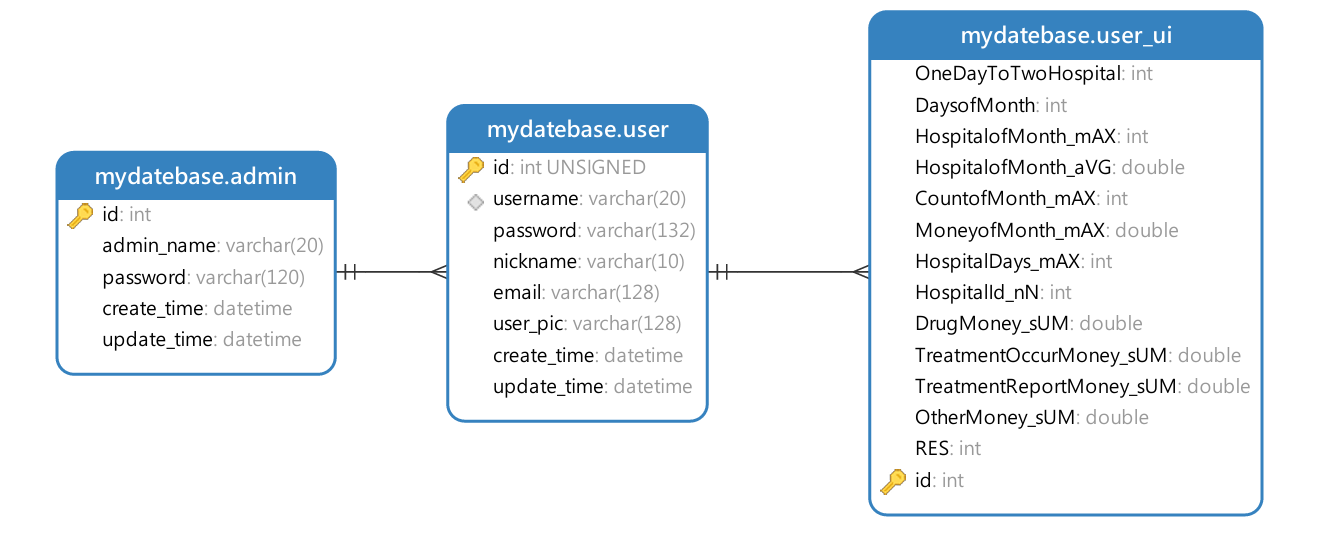
根据医疗保险欺诈识别监测系统的数据字段间的逻辑结构特点，构建用户、欺诈识别信息表，创建数据库表格，建立数据库ER图如图22所示。



**图22 系统ER图**

**7.2.2物理模型的构建（PDM）**

根据概念模型的中实体和他们之前的联系关系，我们最终确定系统的物理模型如**图23**所示。



**图23 物理模型**

**7.2.3数据字典**

进一步分析系统的各项字段，可将其设计为如**表12**所示的数据字典。

**表12数据字典**

|  |  |  |  |
| --- | --- | --- | --- |
| TABLE\_NAME | COLUMN\_NAME | COLUMN\_TYPE | Description |
| user\_ui | OneDayToTwoHospital | int | 一天去两家医院的天数 |
| user\_ui | DaysofMonth | int | 月就诊天数\_MAX |
| user\_ui | HospitalofMonthMAX | int | 月就诊医院数\_MAX |
| user\_ui | HospitalofMonthAVG | double | 月就诊医院数\_AVG |
| user\_ui | CountofMonthMAX | int | 月就诊次数\_MAX |
| user\_ui | MoneyofMonthMAX | double | 月统筹金额\_MAX |
| user\_ui | HospitalDaysMAX | int | 医院\_就诊天数\_MAX |
| user\_ui | HospitalIdNN | int | 顺序号\_NN |
| user\_ui | DrugMoneySUM | double | 药品费自费金额\_SUM |
| user\_ui | TreatmentOccurMoneySUM | double | 治疗费发生金额\_SUM |
| user\_ui | TreatmentReportMoneySUM | double | 治疗费申报金额\_SUM |
| user\_ui | OtherMoney\_SUM | double | 其它发生金额\_SUM |
| user\_ui | RES | int | 是否欺诈 |
| user\_ui | id | int | 医保id |
| user | username | varchar(20) | 用户名 |
| user | id | int | 用户id |
| user | password | varchar(132) | 用户密码 |
| user | email | varchar(128) | 用户邮箱 |
| admin | admin\_name | Varchar(120) | 管理员姓名名 |
| admin | password | varchar(120) | 管理员密码 |
| admin | id | int | 管理员id |

**7.3系统功能WEB展示**

根据模块化思想系统整体由用户管理模块、识别模块、监测模块、系统评估模块、系统帮助模块五个模块构成。

用户正确登录系统后可以上传医保信息文件，通过医保欺诈识别功能获取包含欺诈信息文件，通过医保欺诈监测功能可以实时查看医保欺诈信息。在系统帮助页面获取系统帮助同时也能在系统评估页面对系统进行评价反馈。Web功能总览如图24所示。

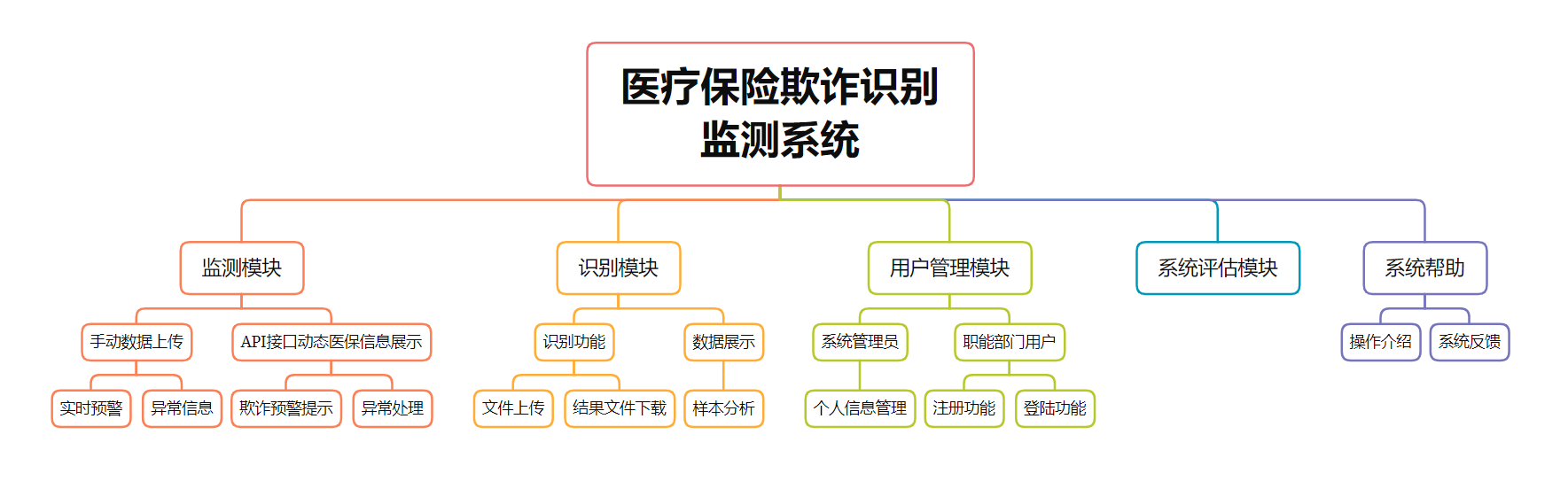
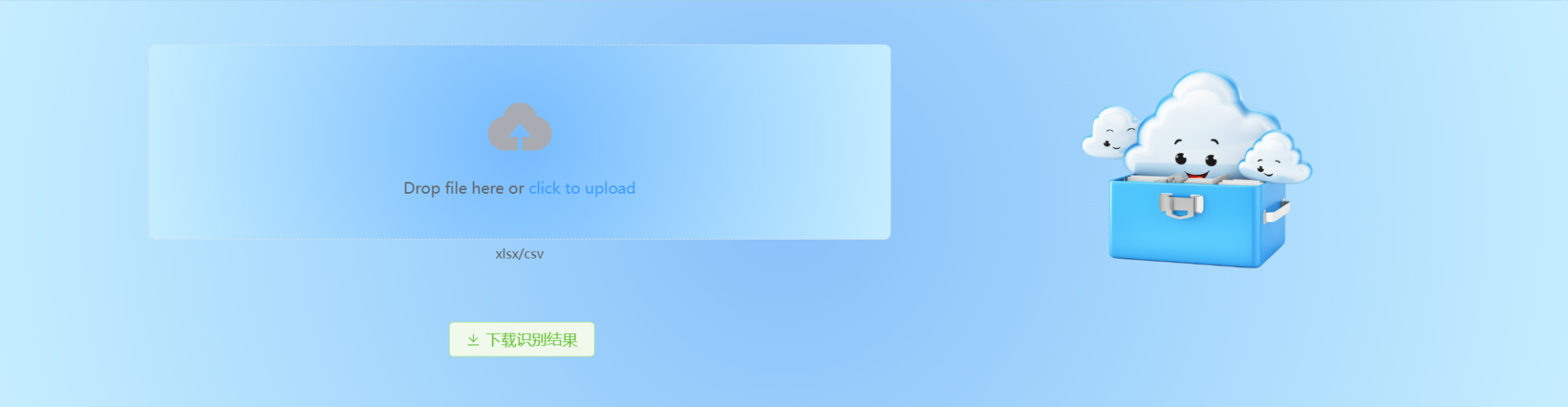


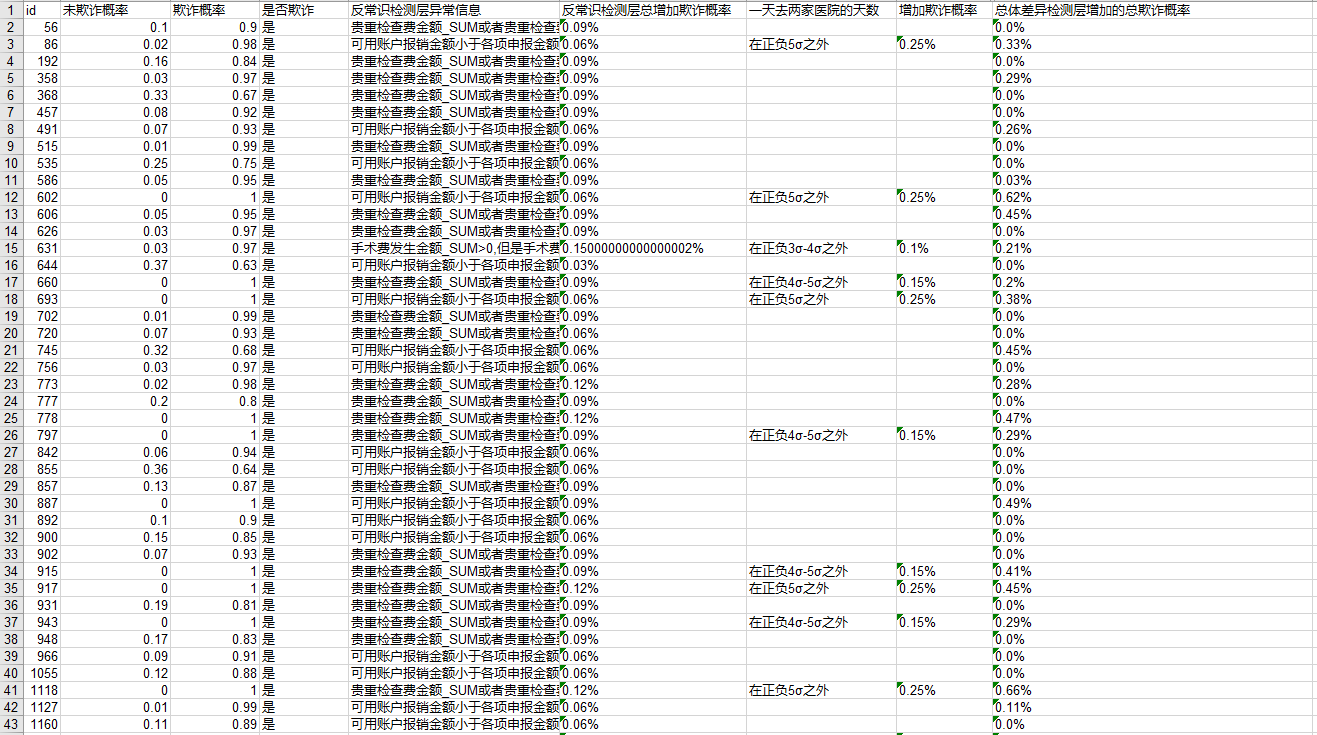
图24web功能总览图

**7.3.1医疗保险欺诈识别模块**

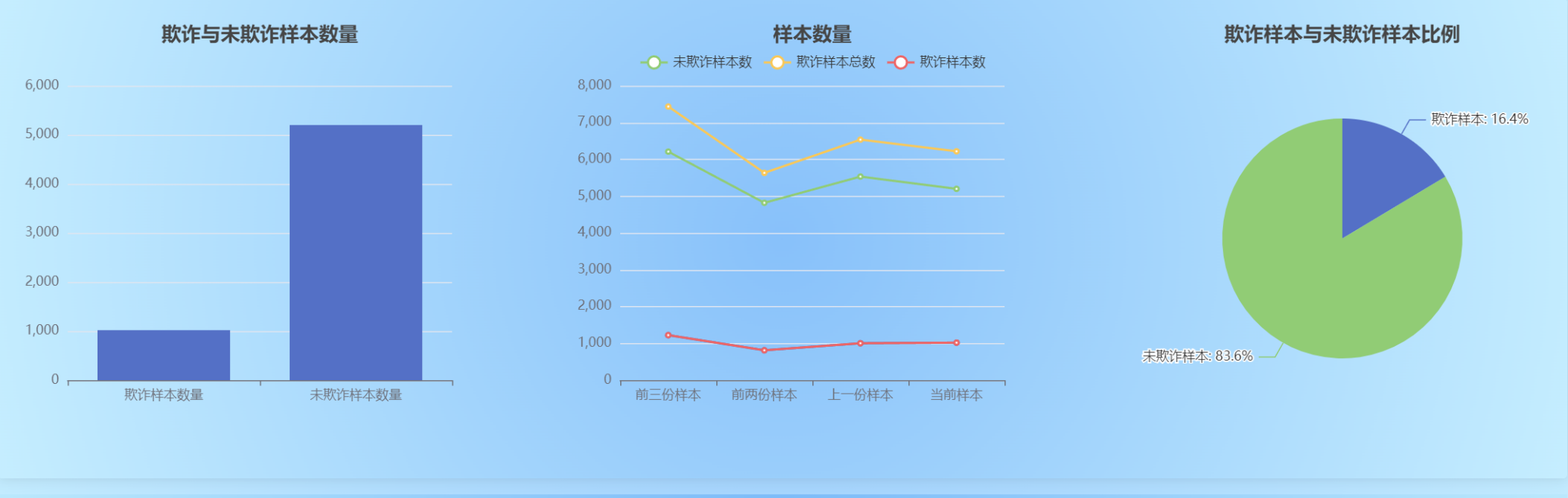
用户将文件拖拽到此处完成医疗保险数据信息文件上传，然后模型返回识别结果，点击下载识别结果即可获取识别信息文档，其中包含详细的异常信息与欺诈概率，还能展示当前医保信息数据中所存在的欺诈数据与未欺诈数据数量及占比，并和前面上传的数据作比较。



**图25医保信息文件上传**



**图 26识别结果文件**



**图27欺诈与为未欺诈统计**

**7.3.2医疗保险欺诈监测模块**

相较于目前的医疗保险欺诈识别和监测系统，我们系统的监测模块最大的特色就是可以连接医院的医疗保险消费系统，通过对医疗保险消费系统传输过来的实时数据进行分层分析，在消费期间就判断其是否存在可能为欺诈的消费行为并将其展示到监测模块当中（如**图28**所示），用户可以随时下载被分类为有欺诈行为的样本并对其欺诈行为进行实时监测与及时阻止。



**图28动态展示医保信息**

除了进行实时监测以外，考虑到部分医院可能不具备有很完备的医疗保险消费系统，我们还设计了用户可以手动输入指定数量的特征来进行医疗保险欺诈识别与监测的功能模块（如**图29**所示）。通过输入实时医疗保险消费数据，也可以做到对医疗保险报销行为的欺诈性进行监测目的。



**图29输入医保信息**

通过上述输入的医保信息，我们可以实时得到某个特征的异常信息与其对欺诈的贡献率，以此来提示用户（如**图30**所示）。



**图30特征贡献率**

**7.3.3用户管理模块**

普通用户和管理员根据不同的角色进行登录、注册，管理员可以进行对普通用户的增加删除修改

|  |  |
| --- | --- |
|  |  |
| **图（a）登录界面** | **图（b）注册界面** |

**图31登录注册**

|  |  |
| --- | --- |
|  | |
| 图（a）查看个人信息 | |
|  |  |
| 图（b）修改个人信息 | 图（c）修改密码 |

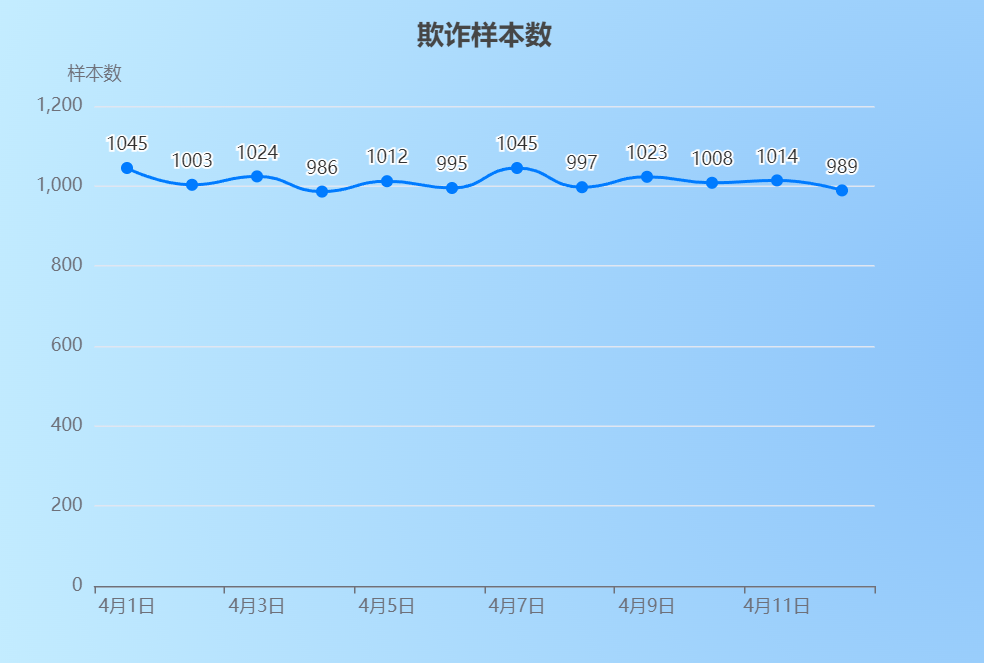
**图 32 个人信息的查看更新**



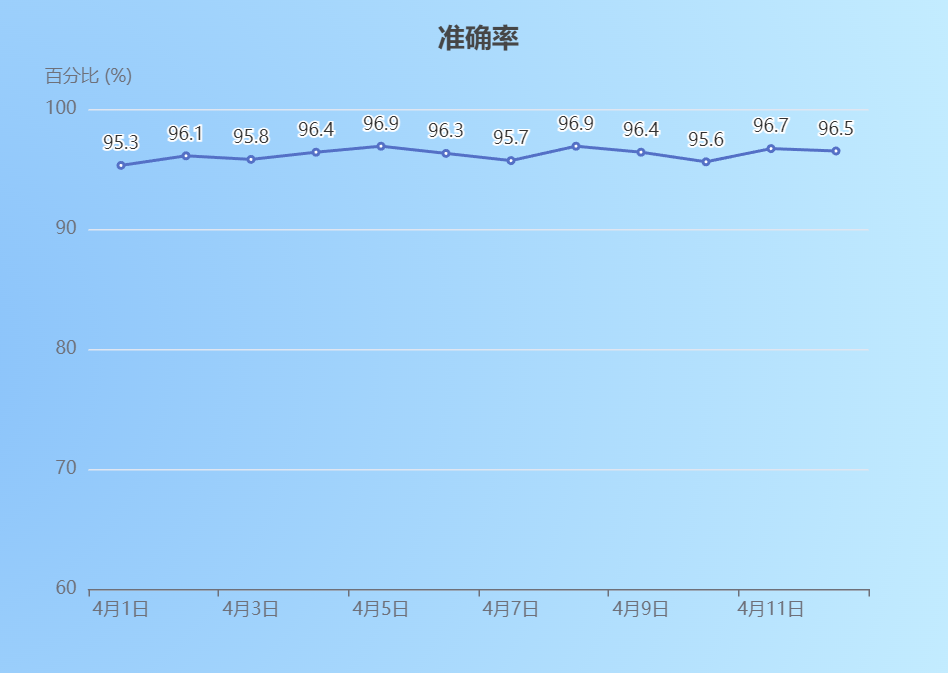
**图 33 用户管理**

**7.3.4系统评估模块**

由于科技水平的发展和进步，医疗保险欺诈行为也变得越来越隐蔽，因此没有任何一个系统或者算法是可以永远适用当时的社会现状的，因此我们在系统设计当中添加一个系统评估模块，从该模块中，用户不仅可以看到每一个天当中的各大医院实时传输和手动传输的样本数量中，欺诈样本的数量（如**图34**所示），还可以通过人工核实的情况，对模型标注的准确率进行实时评估，以判断算法是否需要更新迭代（如**图35**所示）。



**图34每天欺诈样本数**



**图35 月准确率**

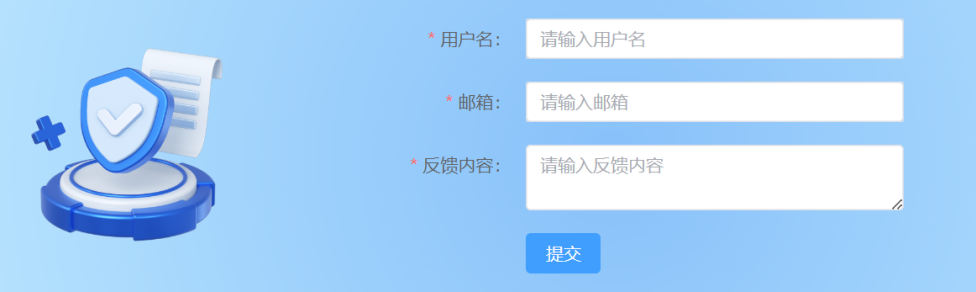
**7.3.5帮助模块**

用户可以查看系统的相关介绍帮助其更好地使用该系统，同时可以通过系统向开发人员反馈自己的使用感受以及建议等。用户点击系统帮助按键即可跳转至如**图36**所示的系统介绍界面。

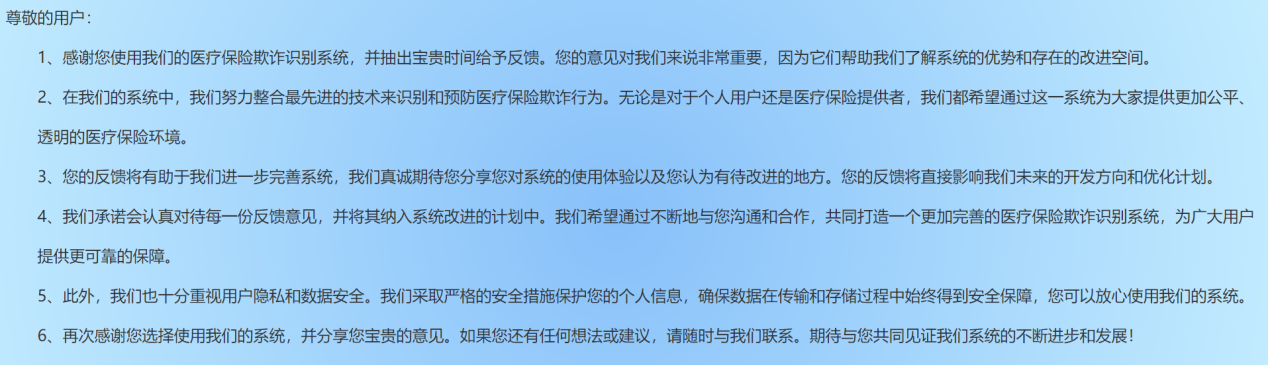


**图36系统介绍**

此外用户在使用该系统时，若发现了部分系统的设计和功能缺陷，可以在用户反馈部分进行登录之后反馈所发现的系统缺陷（如**图37**所示），便于系统后期的维护与升级。



**图37用户填写反馈信息**



**图38用户寄语**

# 8.系统测试

详情请见测试报告。

# 9总结与展望

为了解决医疗保险欺诈识别与监测问题，我们用python搭建基于XGBoost算法的医疗保险欺诈识别融合模型，使用mysql数据库存储数据，通过Flask、Springboot、Vue框架连接前后端与模型，从而搭建了医疗保险欺诈识别与监测系统，系统具有监测和识别功能，并能实时反馈给用户系统识别结果。系统有用户管理模块、识别模块、监测模块、系统评估模块、系统帮助模块五个模块构成。系统功能全面，算法模型效率高，可解释性。

本系统还存在很多不足。在我们的模型中，最终确定12维特征，模型没有考虑的其他特征可能对识别率也有影响，未来我们将进一步探索其他数据处理方法，探求其他分类模型，寻找到更准确的医疗保险欺诈因子并提高系统的性能。