第十一届华中地区大学生数学建模邀请赛

承 诺 书

我们仔细阅读了第十一届华中地区大学生数学建模邀请赛的竞赛细则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与队外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的, 如果引用别人的成果或其他公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们将受到严肃处理。

我们的参赛报名号为：1502

参赛队员（签名）：

队员 1：_____

队员 2：_____

队员 3：_____

湖北省工业与应用数学学会

第十一届华中地区大学生数学建模邀请赛组委会

第十一届华中地区大学生数学建模邀请赛
编号专用页

选择的题号：B

参赛的编号：1502

（以下内容参赛队伍不需要填写）

竞赛评阅编号：

题目：就诊数据对糖尿病患者治疗效果影响的研究

【摘要】

越来越多的研究表明控制住院病人的血糖水平对降低发病率和死亡率具有重要作用，许多医疗机构已经把严格的血糖指标纳入到重症监护室 ICU 级别的正式协议。但是，许多医院对于大多数非 ICU 住院病人的血糖管理比较随意，导致或者完全没有治疗，或者患者的血糖波动很大。为了提高患者的安全性，有必要对现有的医院对收治病人的糖尿病治疗模式进行分析评估。

针对问题一，分析诊断数据的构成，利用数据中表征患者出入院情况的数据，构建出院时身体状况与入院时身体状况两个指标，以其差异来反应医院的治疗效果。评价系统既可以对医院的每一位患者的治疗效果进行定量评价，同时可以通过医院所有患者的治疗效果评分进而对医院进行整体的评价。

对于问题二和问题三，首先对数据进行预处理，剔除同一患者的多项诊断数据，保证患者唯一；剔除缺失率高、与分析无关的数据，筛选描述相同属性的多个特征。之后根据 ICD-9 编码规范对疾病进行分组，对各组就诊人数进行统计。最后将就诊病人根据不同的特征进行分类，并对每种分类进行特征变量概率分布、各类就诊病人的再次入院率及其置信区间进行统计。

对于问题四，首先使用简单的逻辑回归模型进行统计分析，剔除对再次入院率显著性差异几乎不构成影响的特征。之后利用相关性分析去除属性中存在的线性相关的属性，在此基础上进行训练提升决策树拟合，得出了再入院率的主要影响因素。考虑到多个因素共同影响，进行多因素方差分析，得出多对对再入院率联合影响的显著性水平较高的属性对，并再次使用决策树进行拟合，得到这些因素的作用情况。

对于问题五，综合问题一至问题四的分析，根据对医院治疗效果影响因素的研究，得出对医院糖尿病治疗模式的合理建议。

关键词：综合评价 糖尿病治疗 统计分析 决策树 逻辑回归

1　问题重述

控制住院病人的血糖水平对降低发病率和死亡率具有重要作用，许多医疗机构把严格的血糖指标纳入到重症监护室 ICU 级别的正式协议。但是，对于大多数非 ICU 住院病人的接收却没有这样做。

事实上，按传统的方式，住院病人的管理比较随意，导致或者完全没有处理，或者在血糖方面波动很大。为了提高患者的安全性，有必要对现有的医院收治的病人的糖尿病治疗模式进行分析评估。因此就医院对糖尿病的诊疗管理的问题，尝试建立模型解决以下问题：

1. 试分析确定合理的评价指标体系，用以评价医院对糖尿病患者的治疗效果。

2. 对数据进行预处理，剔除无效或无用数据。对各种疾病类型进行分组，并给出各组的诊断值（ICD-9 编码）范围以及各组就诊人数在所有就诊人数中所占的百分比。

3. 给出就诊病人的 HbA1c 检测、性别、年龄、种族、出院去处、入院来源、住院天数、诊疗医师的专业、初次诊断结果、葡萄糖血清检测等特征变量的概率分布，各类就诊病人的再次入院率及其置信区间。

4. 分析再次入院率与各特征变量之间的关系，分析 HbA1c 检测等因素对再次入院率的影响。

5. 对医院糖尿病治疗模式进行评价，就如何降低再次入院率，提高治病效率和效果，降低病人住院治疗成本给出合理建议。

2　问题分析

问题一要求建立一个用以评价医院对糖尿病患者的治疗效果的评价指标体系。通过查阅文献和分析诊断数据的构成，利用数据中表征患者出入院情况的数据，通过出院时身体状况与入院时身体状况的其差异来反应医院的治疗效果。评价系统应该既可以对医院的每一位患者的治疗效果进行定量评价，同时还可以通过医院所有患者的治疗效果评分进而对医院进行整体的评价。

问题二和问题三要求对数据进行预处理，对疾病类型进行分组并统计各组的就诊人数在所有就诊人数中所占的百分比。将就诊病人根据不同的特征进行分类，并对每种分类进行特征变量概率分布、各类就诊病人的再次入院率及其置信区间进行统计。

首先对数据进行预处理，剔除同一患者的多项诊断数据，保证患者唯一；剔除缺失率高、与分析无关的数据，筛选描述相同属性的多个特征。之后根据 ICD-9 编码规范对疾病进行分组，对各组就诊人数进行统计。最后将就诊病人根据不

同的特征进行分类，并对每种分类进行特征变量概率分布、各类就诊病人的再次入院率及其置信区间进行统计计算。

对于问题四，首先使用简单的逻辑回归模型进行统计分析，剔除对再次入院率显著性差异几乎不构成影响的特征。之后利用相关性分析去除属性中存在的线性相关的属性，在此基础上进行训练提升决策树模型，得出再入院率的主要影响因素。考虑到实际情况中可能存在的多个因素共同影响再入院率结果，对影响因素进行多因素方差分析，得出对再入院率联合影响的显著性水平较高的因素组，并再次使用提升决策树进行模型训练，从而进一步分析这些因素的作用情况。

对于问题五，综合问题一至问题四的分析，根据对医院治疗效果影响因素的研究，得出对医院糖尿病治疗模式的合理建议。

3 模型假设

1. 假设患者在一个诊疗期内进行至多一次 HbA1c 测量；
2. 假设患者在一个诊疗期内进行至多一次换药处理；
3. 假设患者的一个诊疗期内至少有一个诊断是糖尿病；

4 符号说明

符号	定义
N	总就诊人数
R_{mis}^{i}	第i个特征的缺失值占比
V_i^j	第 j 个诊疗记录第 i 个特征项的值
C_{Hb}	糖化血红蛋白含量（%）
C_{Glu}	血糖含量（单位：mg/dl）
$R_{ad}^{i,j}$	第i种分类中第j类的就诊人数占总就诊人数比值
$R_{re}^{i,j}$	第i种分类中第j类的再次入院率

5 模型建立与求解

5.1 问题一模型建立与求解

问题一要求根据病患治疗数据确定合理的评价指标体系，用以评价医院对糖尿病患者的治疗效果。

5.1.1 选取评价治疗效果的指标

选取评价治疗效果的指标，需要对患者进入医院治疗前后的健康情况变化作为参考依据，本文认为患者进入医院治疗前后的健康情况改善程度代表了医院的治疗效果，在原数据中，选取的评价指标如图所示：

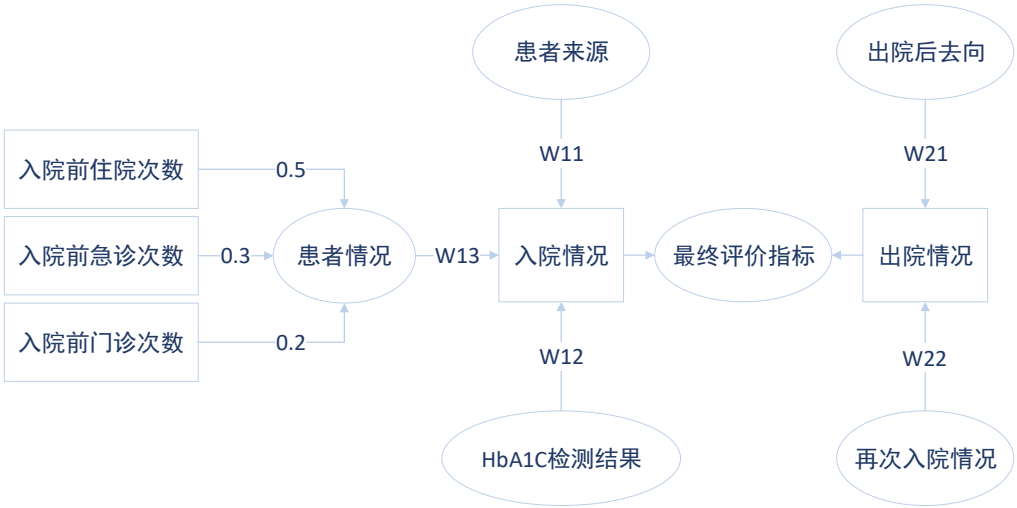

图 5-1 评价指标

1. 病患入院前身体状况

- **HbA1c 检测结果W_{Glu}**

HbA1c 是糖化血红蛋白（GHb）的主要组成成分，可反应糖尿病患者 8~12 周内血糖控制的情况，主要用作糖尿病控制的检测指标，将 HbA1c 检测结果分成三个等级：

若$C_{Glu} < 7$时，认为血糖控制状况正常，则$W_{Glu} = 0$

若$7 < C_{Glu} < 8$时，血糖偏高，认为血糖控制状况较差，则$W_{Glu} = 1$

若 $8< C_{Glu}$时，血糖高，认为血糖控制状况差，则$W_{Glu} = 2$

- **入院前门诊次数N_{out}**

N_{out}反映了病人在入院诊疗前的患病严重程度，二者正相关。

- **入院前急诊次数N_{eme}**

N_{eme}反映了病人在入院诊疗前的患病严重程度，二者正相关。

- **入院前住院次数N_{inp}**

N_{inp}反映了病人在入院诊疗前的患病严重程度，二者正相关。

- **病患来源W_{adm}**

将入院的病患来源依据暗示的疾病严重程度划分为 2 个等级：

若$W_{adm} = 1$，患病严重程度相对较低，来自其他医院、医疗机构、其他医生等转诊；

若$W_{adm}=2$，患病严重程度相对较高，来自急诊室转移。

2. 诊疗后身体状况

- **出院后去向W_{dis}**

将出院后的去向依据暗示的疾病严重程度划分为个 6 个等级：

若$W_{dis}=1$，患病严重程度最低，包括出院后回家；

若$W_{dis}=2$，患病严重程度较低，包括出院后转移到相关的复健机构；

若$W_{dis}=3$，患病严重程度中等，包括出院后需要定期复查；

若$W_{dis}=4$，患病严重程度较高，包括出院后转移到其他医院或医疗机构；

若$W_{dis}=5$，患病严重程度最高，包括转移到长期的护理机构；

若$W_{dis}=6$，治疗不成功，包括死亡和临终关怀。

- **出院后再次入院W_{re}**

同样依据暗示的诊疗出院后的疾病严重程度，将W_{re}分成 3 个等级：

若$W_{re}=0$，身体状况良好，出院后未再次入院；

若$W_{re}=1$，身体状况较好，出院后 30 天后入院；

若$W_{re}=2$，身体状况差，出院后 30 内后入院。

5.1.2 建立评价治疗效果的评价体系

1. 预处理

首先对N_{out}进行标准化处理，公式如下：

$$N_{out}^{i}{}' = \frac{N_{out}^{i} - \overline{N_{out}}}{S_{out}{}^{2}} \tag{5-1}$$

其中$\overline{N_{out}}$是入院前门诊次数的均值，计算如下：

$$\overline{N_{out}} = \frac{1}{N}\sum_{i=1}^{N} N_{out}^{i} \tag{5-2}$$

$$S_{out}{}^{2} = \frac{1}{N}\sum_{i=1}^{N}(N_{out}^{i} - \overline{N_{out}})^{2} \tag{5-3}$$

对N_{eme}，N_{inp}按相同的方法进行标准化处理。

考虑到真实数据的大量缺失值的不可避免性，对于上述指标中出现的缺失值，均将缺失值替换成：

$$X' = [\bar{X}] \tag{5-4}$$

即按照式(5-2)求均值后四舍五入取整。

患者在入院前的门诊、急诊、住院次数可以反映患者自身病情严重程度，同时三者之间也有一定的主次关系。认为相同的次数下，反应患者病情严重度：门诊<急诊<住院。

因此将三者赋权求和，得到P作为衡量患者入院情况指标之一：

$$P_i = 0.2{N_{out}^i}' + 0.3{N_{eme}^i}' + 0.5{N_{inp}^i}' \tag{5-5}$$

2. 确定指标权重

熵是对不确定性的一种度量，信息量越大，不确定性就越小，熵也就越小。根据熵的特性，通过计算熵来判断一个指标的离散程度，离散程度越大，该指标对于综合评价的影响越大，即权重越大。因此通过熵权法对指标进行赋权。

对N条诊断数据，有M_{in}个指标（$M_{in} = 3$），则$X_{i,j}$为第i条诊断数据的第j个指标的值（$j = 1,2,3$）。计算第j个指标下第i条诊断数据占该指标的比重：

$$p_{i,j} = \frac{X_{i,j}}{\sum_{i=1}^{N} X_{i,j}} \tag{5-6}$$

计算第j个指标的熵值：

$$e_j = -k \sum_{i=1}^{N} p_{i,j} \ln(p_{i,j}) \tag{5-7}$$

其中$k = 1/\ln(N) > 0$，满足$e_j \geq 0$

计算信息熵冗余度：

$$d_j = 1 - e_j \tag{5-8}$$

计算各项指标的权重：

$$w_j = \frac{d_j}{\sum_{j=1}^{M_{in}} d_j} \tag{5-9}$$

则入院情况指标为：

$${S_{in}}^i = \sum_{j=1}^{M_{in}} w_j p_{i,j} \tag{5-10}$$

以同样的方式计算出院情况指标，得到：

$${S_{out}}^i = \sum_{j=1}^{M_{out}} {w_j}' {p_{i,j}}' \tag{5-11}$$

则对医院的第i位糖尿病患者治疗效果的评价公式：

$$G_i = \frac{{S_{out}}^i}{{S_{in}}^i} \tag{5-12}$$

若对医院进行整体评价，则设定一个阈值T，对于每一条诊断数据：

$$G_i' = \begin{cases} 1 & G_i > T \\ 0 & G_i < T \end{cases} \tag{5-13}$$

则最终对医院的治疗效果整体评价公式：

$$G = \frac{\sum G_i'}{N} \tag{5-14}$$

5.1.3 评价结果与分析

由上述整体评价公式，将医院对糖尿病患者的治疗效果分为五个等级：

若$G \in [0,0.2)$，则认为医院对糖尿病患者的治疗效果很差；

若$G \in [0.2,0.4)$，则认为医院对糖尿病患者的治疗效果较差；

若$G \in [0.4,0.6)$，则认为医院对糖尿病患者的治疗效果一般；

若$G \in [0.6,0.8)$，则认为医院对糖尿病患者的治疗效果较好；

若$G \in [0.8,1)$，则认为医院对糖尿病患者的治疗效果很好；

在该评价系统中，充分利用了数据中的患者出入院情况数据，利用出院时身体状况与入院时身体状况的差异来反应医院的治疗效果。可以对医院的每一位患者的治疗效果进行定量评价，同时可以通过一家医院所有患者的治疗效果评分进而对医院进行整体的评价。

5.2 问题二模型建立与求解

问题二要求对数据进行预处理，并对疾病类型进行分组，给出各组的诊断值（ICD-9）编码范围及其就诊人数所占总就诊人数的比值。

5.2.1 病患就诊数据分析

附件一给出了糖尿病病患的诊疗数据，共有 101766 项诊疗数据，其中包含了 71518 个糖尿病患者，共由 50 个特征组成。我们将属性值为“？”以及表示为“未知”意义的数据统一归为缺失值后，计算数据缺失值占比。数据描述如表 5-1 所示。

数据缺失率意义如下：

$$R_{mis}^i = \frac{N_{mis}^i}{N} \tag{5-15}$$

其中N_{mis}^i表示第i个特征的缺失项个数；

表 5-1 诊疗数据描述

特征名	i	特征描述	R_{mis}^{i}
encounter_id	1	诊疗编号，诊疗的唯一标识	0
patient_nbr	2	病患编号，病患的唯一标识	0
payer_code	3	付款人代码，包括 BC、CH 等 18 个分类	39.6
race	4	种族，包括亚裔、非裔等五个分类	2.2
gender	5	性别	0
age	6	年龄，在区间[0,100)内以十年为间隔分组	0
weight	7	体重	97.0
admission_type_id	8	入院类型，共急诊、新生等 6 个分类	9.9
discharge_disposition_id	9	出院类型，包括转院、在家死亡等共 21 个分类	3.6
admission_source_id	10	病患来源，包括转诊等 14 个分类	6.8
time_in_hospital	11	住院时间，范围 1~14 天	0
medical_specialty	12	诊疗医生的医学专业，共 84 类	49.1
num_lab_procedures	13	就诊期间实验室测试次数	0
num_procedures	14	就诊期间诊疗次数	0
num_medications	15	就诊期间药物数量	0
diag_1	16	初步诊断（ICD-9 编码），共 848 个分类	0
diag_2	17	二级诊断（ICD-9 编码），共 923 个分类	0.4
diag_3	18	额外诊断（ICD-9 编码），共 954 个分类	1.4
number_diagnoses	19	单位时间内诊断总次数	0
max_glu_serum	20	最大血糖含量，包括正常（<200）、偏高（200~300）、高（>300）	94.7
A1Cresult	21	HbA1c 测量，包括正常（<7）、偏高（7~8）、高（>8）	83.3
medication	22-44	使用药物情况，包括糖尿病 23 种常用药的用药变化（未用、稳定、加量、减量）	0
change	45	是否更换药物或改变药物用量	0
diabetesMed	46	是否有药物处方	0
number_outpatient	47	入院前门诊总次数	0
number_emergency	48	入院前急诊总次数	0
number_inpatient	49	入院前住院总次数	0
readmitted	50	再入院时间，包括未入院、30 天内未入院和 30 天内再入院	0

缺失值的值域为：$V_{mis} = \{'Invalid','Unknow','?','Null','Not\ \ Available'\}$

特别需要指出的是，当某一特征的特征值$V = 'Not\ \ Mapped'$时，当作其他

值处理，而不是缺失值。

5.2.2 数据预处理

该数据为来自美国 130 家医院的真实糖尿病患者治疗数据，不可避免的出现了记录不完整或不一致的高维复杂数据，虽然可能有潜在的分析价值，但难以直接使用。

在本文中针对该原始数据进行一些预处理，以便于后期的处理和分析：

1. 对于数据中表示为“未知”的值，即$V \in V_{mis}$时，特征值均作为缺失值；
2. 数据集中的某些病患有多于一次的住院就诊记录，由于同一病患的数次就诊记录间很可能是相互关联的，使用这些数据可能会对分析结果造成影响。因此每个病患仅保留第一次出现的诊疗记录，即留下了 71518 位病患的一次诊疗记录；
3. 由于支付人代码（$i = 3$）对于糖尿病治疗相关分析无关，且数据的缺失较严重（$R_{mis}^{i} = 39.6\%$），因此将此数据删除；
4. 体重（$i = 7$）数据有 97%的缺失值（$R_{mis}^{i} = 97\%$），因此将体重数据删除；
5. 病患入院类型（$i = 8$）和病患来源（$i = 10$）属于同一特征的两种表示，由于病患入院类型数据缺失率比病患来源数据缺失率高，因此删除病患入院类型；
6. 在后续分析中，再次入院率是分析的重点，最终死亡的病患此项数据为无效值。为了避免最终死亡的病患就诊记录对分析结果的影响，在预处理时将死亡病患的数据删除，出院类型是临终关怀（$Hospice$）相关的数据项同样需要删除；
7. HbA1c 是糖化血红蛋白（GHb）的组成成分之一，C_{Hb}可反应糖尿病患者 8~12 周内血糖控制的情况，主要用作糖尿病控制的检测指标；C_{Glu}反应的是即刻血糖水平，此测量值在机体受到严重创伤等强烈刺激因素作用下出现应激性高血糖症（SHG）[1]时会高于正常值，同时在一般情况下，血糖也会呈现一定范围内的正常波动。相比之下C_{Hb}更适用于评价血糖控制，同时C_{Glu}有 94.7%缺失，因此将此特征（$i = 20$）删除；
8. 用药情况（$i \in [22, 44]$）体现了 23 种糖尿病患者常用药的用药变化（$V_i^j \in \{'no', 'steady', 'up', 'down'\}$），当一次诊疗中有至少一种药发生了用量变化，或换成了新药，则至少存在一个i，有：

$$V_i^j \notin \{'no', 'steady'\} \tag{5-16}$$

此时对应的用药变化（$i = 45$）有：

$$V_i^j = 1 \tag{5-17}$$

否则$V_i^j = 0$，该数值是逻辑值。

9. 为统计再次入院率，再次入院情况（$i = 50$）中，将大于三十天入院和小于三十天入院归为一类，最终再次入院情况的值包括“入院”和“未再入院”，即：

$$V_i^j = 1 \text{ 或 } 0,\ i = 50, \tag{5-18}$$

经过上述预处理步骤，最终保留了 69990 项就诊记录，共 23 个特征项。

5.2.3 疾病类型分组

根据 ICD-9（国际疾病伤害及死因分类标准）的编码规范，可将 ICD-9 编码根据疾病类型进行分类如表 5-2 所示。

在若干次诊断中，初步诊断（$i = 16$）作为病患入院后经过检查做出的第一步判断，也是一般的主要诊断结果，后续诊断通常是对后续发病的诊断以及对初步诊断的补充。因此取初步诊断结果作为分组依据，并：

1. 将不明症状编码（780-789）中详细描述的各编码与对应明确症状的编码范围合并；
2. 将占比小的疾病类型合并。对于就诊人数占比低于 3%的先天畸形、感觉器官疾病等七类疾病合并为一个分组；
3. 糖尿病属于“内分泌、营养、新陈代谢及免疫系统疾病”类，作为本文分析的重点疾病，同时其就诊人数占比高达 7.93%，因此糖尿病单独分成一类，其 ICD-9 编码小数点后数字表示对糖尿病的进一步细分，在这里不再细分。

最终得到疾病分组共九组，其中循环、呼吸、消化、糖尿病、受伤及中毒、肌肉骨骼系统及结缔组织疾病、泌尿、肿瘤八类疾病的就诊人数分别占总人数的 3%以上，因此各成一组。最终得到的各组疾病的 ICD-9 范围以及对应的就诊人数在总就诊人数中所占比值如表 5-3 所示。

从图 5-1 中可以看出患有糖尿病的病患主要诊断进行分组后，各组的就诊人数分布情况。

表 5-2 ICD-9 编码分类

ICD-9	描述
001-139	传染病和寄生虫疾病
140-239	肿瘤

240-279	内分泌、营养、新陈代谢及免疫系统疾病
280-289	血液及造血器官疾病
290-319	精神失常
320-359	神经系统疾病
360-389	感觉器官疾病
390-459	循环系统疾病
460-519	呼吸系统疾病
520-579	消化系统疾病
580-629	泌尿生殖系统疾病
630-676	妊娠、分娩和产后合并症
680-709	皮肤及皮下组织疾病
710-739	肌肉骨骼系统及结缔组织疾病
740-759	先天畸形
760-779	围产期引起的特定情况
780-799	症候、征候及不明情况
800-999	受伤及中毒
E 和 V	外伤及补充分类
780	一般症状
781	涉及神经和肌肉骨骼系统的症状
782	涉及皮肤和其他皮肤组织的症状
783	有关营养，新陈代谢和发育的症状
784	涉及头部和颈部的症状
785	涉及心血管系统的症状
786	涉及呼吸系统和其他胸部症状的症状
787	涉及消化系统的症状
788	涉及泌尿系统的症状
789	其他症状涉及腹部和骨盆

表 5-3 疾病分组

组别	ICD-9	描述	人数	人占比
Circulatory	390-459, 785	循环系统疾病	21390	30.57%
Respiratory	460-519, 786	呼吸系统疾病	9491	13.56%
Digestive	520-579, 787	消化系统疾病	6488	9.27%
Diabetes	250.xx	糖尿病	5548	7.93%
Injury	800-999	受伤及中毒	4696	6.71%
Musculoskeletal	710-739	肌肉骨骼系统及结缔组织疾病	4064	5.81%
Genitourinary	580-629, 788	泌尿生殖系统疾病	3441	4.92%
Neoplasms	140-239	肿瘤	2538	3.63%
Other	240-279(除 250), 783	内分泌、营养、新陈代谢及免疫系统疾病	2067	17.60%
	789-799, 780, 784	症候、征候及不明情况	2050	
	680-709, 782	皮肤及皮下组织疾病	1843	
	001-139	传染病和寄生虫疾病	1685	
	290-319	精神失常	1545	
	E 和 V	外伤及补充分类	919	
	320-359, 781	神经系统疾病	717	
	280-289	血液及造血器官疾病	652	
	630-676	妊娠、分娩和产后合并症	586	
	360-389	感觉器官疾病	219	
	740-759	先天畸形	41	

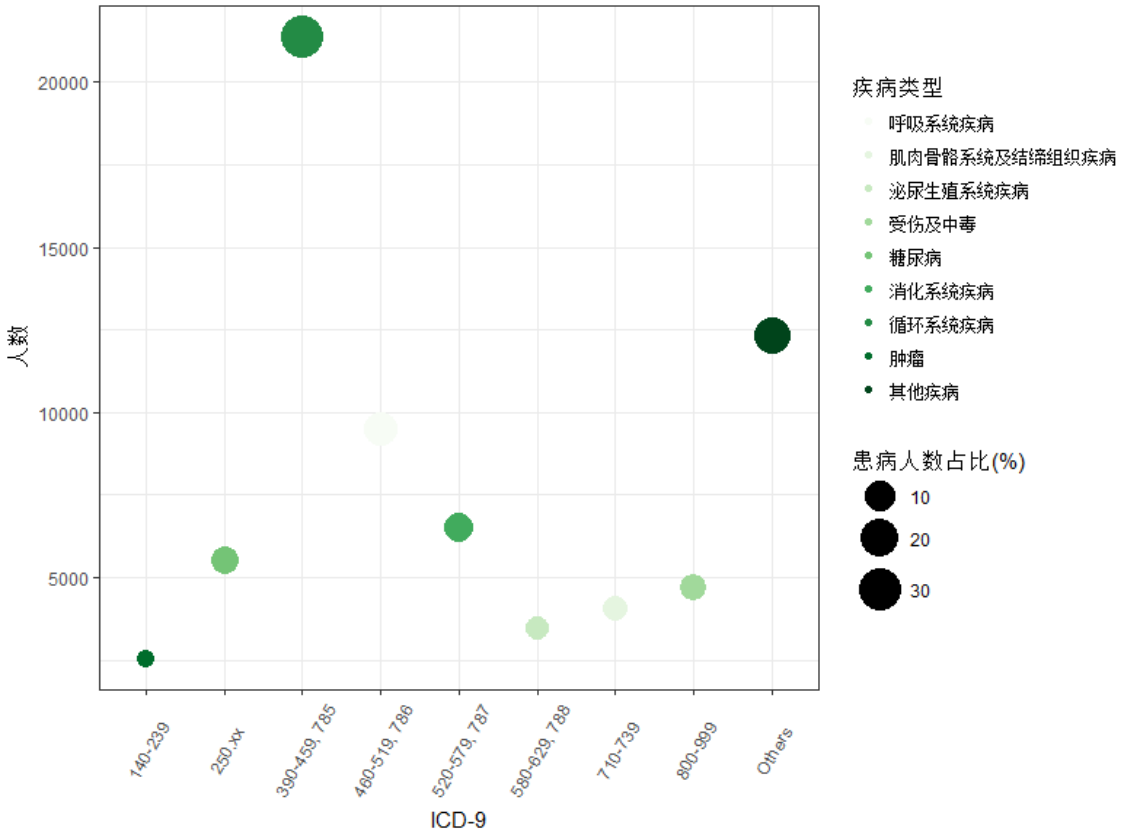

图 5-2 各组疾病人数分布情况

5.3 问题三模型建立与求解

问题三要求给出就诊病人的各项特征信息概率分布，并求出不同分类情况下，各类就诊病人的再次入院率及其置信区间。

5.3.1 特征变量概率分布

在数据中，病患年龄将[0,100)内的年龄以十年为间隔平均划分为十个分组，按照各年龄段的再次入院率水平将年龄重新分为三组（分别是[0,30)，[30,60)以及[60,100)）。

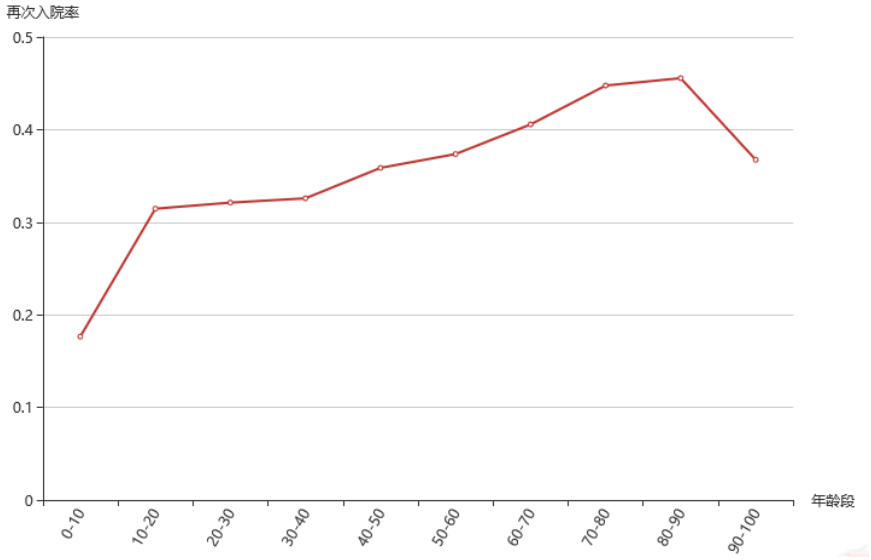

图 5-3 各年龄段患者再次入院率

在图 5-4 中，依据各人种的再次入院率水平，同样将人种分为三组（分别是高加索、非裔和其他）。

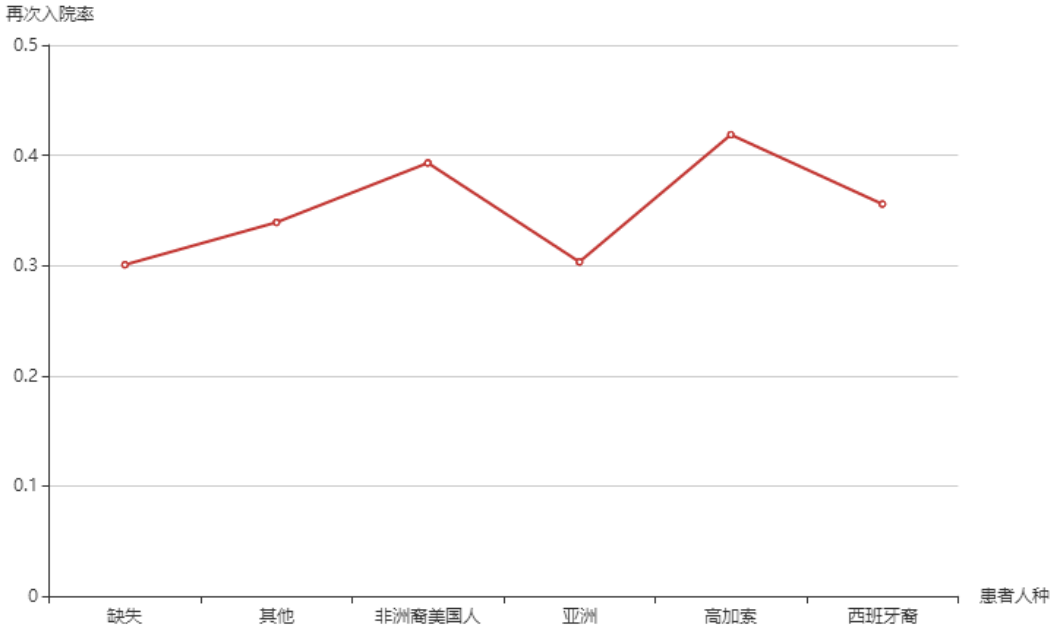

图 5-4 各种族患者再次入院率

分组后患者个人信息的各项特征变量的概率分布情况如表 5-4 所示。

表 5-4 个人信息各项特征变量的概率分布

特征名	特征值	就诊人数	人数占比(%)
人种	高加索人	52305	0.747321
	非裔美国人	12627	0.180411
	其他（包括未知）	5058	0.072267
性别	男	32748	0.467895
	女	37239	0.532062
年龄	[0,30)	1808	0.025832
	[30,60)	37560	0.536648
	[60,100)	30622	0.43752

依据本文 5.1 所述，根据暗示的疾病严重程度对于入院类型划分为三组，出院类型分为六组，得到的患者住院信息概率分布如表 5-5 所示。

表 5-5 住院信息各项特征变量的概率分布

特征名	特征值	就诊人数	人数占比(%)
出院类型	回家	44395	0.634305
	转康复机构	1962	0.028033
	待复查	11	0.000157
	转院	9883	0.141206
	转护理机构	10078	0.143992
	其他（包括未知）	3661	0.052307
入院来源	其他医疗机构	27631	0.394785
	急救室	37275	0.532576
	其他（包括未知）	5084	0.072639

住院时长的特征值范围$V_i^j \in [1,14]$, 其中$i = 11$, V_i^j为整数，其就诊人数占比概率分布如图 5-5 所示。

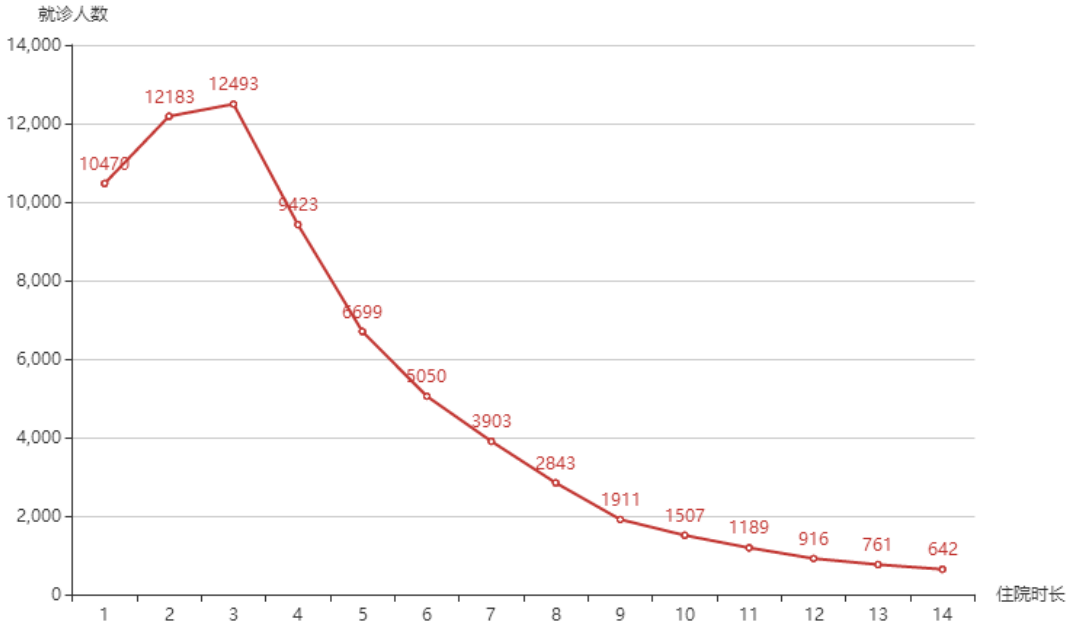

图 5-5 不同住院时长的就诊人数概率分布

对原有 84 个分类的诊疗医生专业（$i = 12$）按照医学上的外科、内科、全科、手术划分为四类，将无法明确分类且就诊人数占比少的专业与缺失值统一归为其他类。

表 5-6 诊疗信息各项特征变量的概率分布

特征名	特征值	就诊人数	人数占比(%)
诊疗医生专业	外科	8327	0.118974
	内科	16531	0.236191
	全科	4978	0.071124
	手术	1473	0.021046
	其他（包括未知）	38681	0.552665
HbA1c 含量	正常（＜7）	3741	0.05345
	偏高（7~8）	2866	0.040949
	高（＞8）	6239	0.089141
	未知	57144	0.816459
药物变换	否	38493	0.549979
	是	31497	0.450021

5.3.2 再次入院率及其置信区间

对于一个疾病分组，再次入院率为：

$$R_{re}^{i,j}=\frac{N_{re}^{i}}{N^{i}} \tag{5-19}$$

其中N^{i}表示该分组的所有就诊人数；N_{re}^{i}表示该分组中再次回院就诊的人数。

在按疾病分组时，每一个疾病分组中的病患再次入院值为：

$$V_{re}^{i,j}=1\text{ 或 }0 \tag{5-20}$$

其中i表示疾病分组，j表示在第i分组中第j项病患诊疗记录。0 表示未再次入院，1 表示出院后有再次入院。

由于总体方差未知，再次入院率的置信区间$(u(V_{re}^{i,j}),v(V_{re}^{i,j}))$为：

$$\left(\overline{V_{re}^{i}}-\frac{S}{\sqrt{N^{i}}}t_{\alpha/2}(N^{i}-1),\ \overline{V_{re}^{i}}+\frac{S}{\sqrt{N^{i}}}t_{\alpha/2}(N^{i}-1)\right) \tag{5-21}$$

其中$\overline{V_{re}^{i}}$是第i分组中再次入院值的均值，计算方法如下：

$$\overline{V_{re}^{i}}=\frac{1}{N^{i}}\sum_{j=1}^{N^{i}}V_{re}^{i,j} \tag{5-22}$$

S^{2}是样本方差，计算如下：

$$S^{2}=\frac{1}{N^{i}-1}\sum_{\mathrm{i=1}}^{\mathrm{n}}(V_{re}^{i,j}-\overline{V_{re}^{i}})^{2} \tag{5-23}$$

当置信水平$(1-\alpha)=95\%$时，即：

$$Pr\left(u(V_{re}^{i,j})<\omega<v(V_{re}^{i,j})\right)=0.95 \tag{5-24}$$

自由度为无限大（$N^{i}>120$）的 t-分布和正态分布等价，查表得到$t_{\frac{\alpha}{2}}(N^{i}-1)=1.96$，最终计算得到各个疾病分组的再次入院率的置信区间如表 5-7 所示。

表 5-7 疾病分组下各组就诊病人的再次入院率及其置信区间

疾病组名	再次入院率(%)	置信区间
Circulatory	0.425479196	(0.43%, 0.42%)
Respiratory	0.427879043	(0.44%, 0.42%)
Digestive	0.395961776	(0.41%, 0.38%)

Diabetes	0.436193223	(0.45%, 0.42%)
Injury	0.39246167	(0.41%, 0.38%)
Musculoskeletal	0.388259227	(0.40%, 0.37%)
Genitourinary	0.355807087	(0.37%, 0.34%)
Neoplasms	0.327029157	(0.35%, 0.31%)
Other	0.397679325	(0.41%, 0.39%)

利用相同的计算方法得到各种分类情况下的再次入院率置信区间结果如表5-8所示。

表 5-8 各种分类条件下各类就诊病人的再次入院率及其置信区间

特征名	特征值	再次入院率	置信区间
人种	高加索人	0.418679	(0.42%, 0.41%)
	非裔美国人	0.393047	(0.40%, 0.38%)
	其他（包括未知）	0.326018	(0.34%, 0.31%)
性别	男	0.414807057	(0.42%, 0.41%)
	女	0.398888584	(0.40%, 0.39%)
年龄	[0,30)	0.306969	(0.33%, 0.29%)
	[30,60)	0.380804	(0.39%, 0.38%)
	[60,100)	0.445856	(0.45%, 0.44%)
出院类型	回家	0.387251	(0.39%, 0.38%)
	转康复机构	0.476555	(0.50%, 0.45%)
	待复查	0.363636	(0.66%, 0.07%)
	转院	0.463118	(0.47%, 0.45%)
	转护理机构	0.448006	(0.46%, 0.44%)
	其他（包括未知）	0.351816	(0.37%, 0.34%)
入院来源	其他医疗机构	0.363867	(0.37%, 0.36%)
	急救室	0.434339	(0.44%, 0.43%)
	其他（包括未知）	0.445909	(0.46%, 0.43%)

诊疗医生专业	外科	0.404828	(0.42%, 0.39%)
	内科	0.40167	(0.41%, 0.39%)
	全科	0.411546	(0.42%, 0.41%)
	手术	0.432101	(0.45%, 0.42%)
	其他（包括未知）	0.291921	(0.32%, 0.27%)
HbA1c 含量	正常（＜7）	0.364074	(0.38%, 0.35%)
	偏高（7~8）	0.38695	(0.40%, 0.37%)
	高（＞8）	0.395736	(0.41%, 0.38%)
	未知	0.412484	(0.42%, 0.41%)
药物变换	否	0.391292	(0.40%, 0.39%)
	是	0.426993	(0.43%, 0.42%)

在按住院天数分类条件下，各类就诊病人的再次入院率及其置信区间如图 5-6 所示。

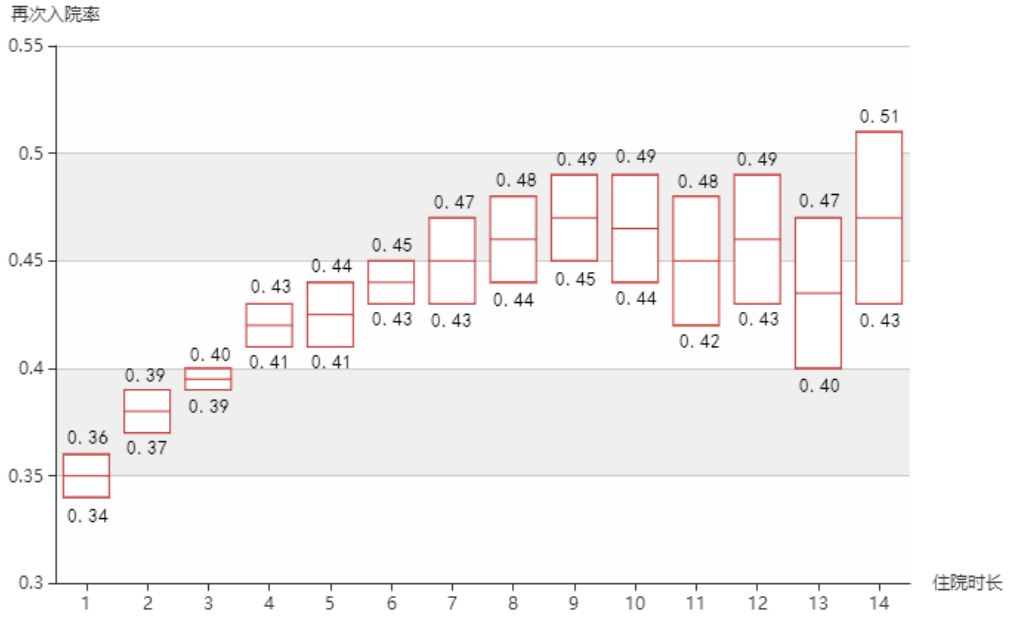

图 5-6 不同住院时长下就诊病人的再次入院率及其置信区间

5.4 问题四模型建立与求解

问题四要求分析再次入院率与各特征变量之间的关系，并分析 HbA1c 等因素对再次入院率的影响。

5.4.1 数据处理

1. 考虑到血糖浓度对换药与否可能存在关联，将 HbA1c 测量值（$i = 21$）和换药情况（$i = 45$）组合，分为未测量 HbA1c、测量显示正常、测量显示不正常但未换药、测量显示不正常并且换药；

2. 考虑到样本数据的平衡性对逻辑回归具有较大的影响，将出院去向（$i = 9$）分类进行合并，合并结果为：出院（63%）、转院（32%）、其他（5%）；

3. 对所有数值型变量进行归一化处理，以防止数值尺度通过印象信息增量的计算从而影响决策树模型的效果；

4. 针对名义和类别变量，创建虚拟变量来代替原有变量，特别地，对于具有m个类别的变量，在逻辑回归模型中，需要创建$\mathrm{m}-1$个虚拟变量以避免多重线性相关的发生。

5.4.2 逻辑回归

逻辑回归是一种广义的线性回归分类模型，适用于二分类问题的预测，并且能够通过给出自变量的权重和显著性检验结果来分析每个自变量在模型中所起的作用，因此在本文中使用逻辑回归作为初步分析手段，对给出的所有属性对于再次入院率的影响进行分析，以便于去除对结果显著性差异几乎不造成影响的属性并对各因素对于再次入院率的影响作出简单的概括。

逻辑回归使用的激活函数为：

$$\mathrm{g(z)} = \frac{1}{1+e^{-z}} \tag{5-25}$$

其中，训练数据为向量：

$$\mathrm{z} = \theta^T = \theta_0 x_0 + \theta_1 x_1 + \cdots + \theta_n x_n \tag{5-26}$$

损失函数为加入正则项的对数似然函数：

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{n}(h_\theta(x_i) - y_i)^2 + \lambda\sum_{j=1}^{n}\theta_j^2 \tag{5-27}$$

使用梯度下降法更新权重：

$$\theta_\mathrm{j} := \theta_\mathrm{j} - \frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)x_i^j - \frac{\lambda}{m}\theta_j \tag{5-28}$$

在训练过程中，每次训练随机抽取 30%的数据作为测试数据，其余数据作为训练数据，重复五次，取五次均值作为最终结果。使用网格搜索得到模型最佳参数$\mathrm{C} = 5$（$\mathrm{C} = \frac{1}{\lambda}$）。

由于未排除属性间的相关性影响以及样本不平衡带来的影响，此处采用显著性检验结果对属性重要性进行衡量，通过逻辑回归得到的显著性检验结果（部分）如表 5-9 所示。

表 5-9 显著性检验

变量	特征名称	p-value
gender	性别	1.021064e-01
time_in_hospital	住院时间	3.609355e-05
num_lab_procedures	实验室测试次数	9.764901e-02
num_procedures	诊疗次数	1.388421e-02
num_medications	药物数量	7.316712e-01

分析得到对于再入院率影响不明显的属性，并在后续的模型中去除：gender（$p = 0.102$）, num_medications（$p = 0.731$），race（$p > 0.1$），medical_specialty（$p > 0.5$）。

分析上述结果还能初步得到对于再入院率的影响较为显著的因素，其中包括：住院时间、门诊和急诊次数、住院次数、诊断次数、是否有糖尿病处方、老年群体以及急诊入院等。

得到的模型精度如表 5-10 所示。

表 5-10 精度评定

精度指标	精度
准确率	0.62058
精确率	0.5889
召回率	0.2309
f1	0.3317
AUC	0.6300

5.4.3 相关性分析

由于决策树模型要求变量之间相互独立，为避免线性相关和多重线性相关对训练结果造成影响，因此需要对数据中的变量进行相关性分析；同时，相关性分析也能够发现一些具有潜在的联合影响再次入院率的可能的变量组。

在衡量数值型变量的相关性时，使用皮尔逊相关系数（Pearson），计算公式如下：

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \tag{5-29}$$

在衡量有序分类变量的相关性时，使用斯皮尔曼相关系数（Spearman）或肯德尔相关系数（Kendall），计算公式如下：

$$\rho_{\text{spearman}} = 1 - \frac{6\Sigma d_i^2}{n(n^2-1)} \tag{5-30}$$

$$\text{Tau} = \frac{C-D}{N(N-1)/2} \tag{5-31}$$

表 5-11 数值型变量之间的相关系数（部分）

特征 1	特征 2	*pr*	p-value
time_in_hospital	num_lab_procedures	0.316842	######
num_lab_procedures	time_in_hospital	0.316842	######
time_in_hospital	number_diagnoses	0.234767	2.95E-88
number_diagnoses	time_in_hospital	0.234767	2.95E-88
time_in_hospital	num_procedures	0.210957	3.07E-71
num_procedures	time_in_hospital	0.210957	3.07E-71
number_emergency	number_inpatient	0.174387	6.50E-49
number_inpatient	number_emergency	0.174387	6.50E-49
num_lab_procedures	number_diagnoses	0.147352	2.85E-35
number_diagnoses	num_lab_procedures	0.147352	2.85E-35

首先计算数值型特征变量之间的相关系数得到表 5-1，通过观察表格可以发现，住院时间与实验室化验次数、手术次数、诊断次数之间存在较大相关性，急诊次数与住院次数之间也存在一定相关性。

这里为了避免对决策树结果造成影响，选择只保留住院时间和急诊次数这两个特征，将与其相关的特征删除，以消除线性相关和多重线性相关。

表 5-12 非数值变量之间相关性（部分）

f_1	f_2	*sr*	p-value
age	discharge_disposition_id	0.25975	######
discharge_disposition_id	age	0.25975	######
age	HdA1c	0.09586	9.38E-16
HdA1c	age	0.09586	9.38E-16

HdA1c	diabetesMed	0.08566	7.03E-13
diabetesMed	HdA1c	0.08566	7.03E-13

分析非数值变量之间的相关性得到表中所示的结果：相关性较高的特征并没有分类次序的递进关系，因此不能说明其中的相关性，而年龄与出院去向、年龄与 HdA1c 检测以及 HdA1c 检测与糖尿病处方之间的潜在联系会在后文中给出分析。

5.4.4 提升决策树训练

逻辑回归模型从根本上来说是一种线性加权拟合模型，对于属性间关系复杂的非线性问题解决能力有限，因此这里使用提升决策树（XGBoost）模型进行分析。

分类回归树（CART）使用基尼不纯度代替信息熵来进行数据集的划分，计算公式为：

$$\text{gini(T)} = 1 - \sum_{j=1}^{n} p_j^2 \tag{5-32}$$

$$\text{Gini}_{\text{split}}\text{(T)} = \sum \frac{N_i}{N} gini(T_i) \tag{5-33}$$

CART 是遍历每一个特征的特征值，每个特征值得到一个划分，然后计算每个特征的信息增益从而找到最优的特征；CART 每一个分支都是二分的，当特征值大于两个的时候，需要考虑特征值的组合来得到分支；之后便可以通过回溯法递归生成一颗决策树，在决策过程中，CART 会把输入根据输入的属性分配到各个叶子节点，而每个叶子节点上面都会对应一个实数分数。

提升决策树在 CART 的基础上通过限制最大树深度使 CART 成为弱分类器，利用 Boosting 的思想，通过叠加多个弱分类器来得到一个强分类器，并且引入了正则化项来限制模型的复杂度，因此，第 t 轮计算的模型目标函数可以表示为;

$$Obj^t = \sum_{i=1}^{n} l\left(y_i, y_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + Constant \tag{5-34}$$

$$\Omega(f_t) = \gamma \text{T} + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{5-35}$$

在 XGBoost 中损失函数可以根据需要更改，这里使用和逻辑回归模型中一样的 ROC 曲线下面积作为精度评价标准。将 5.4.3 中处理得到的数据输入训练模型，经过网格搜索和交叉验证进行参数调整后得到结果如图 5-7 所示：

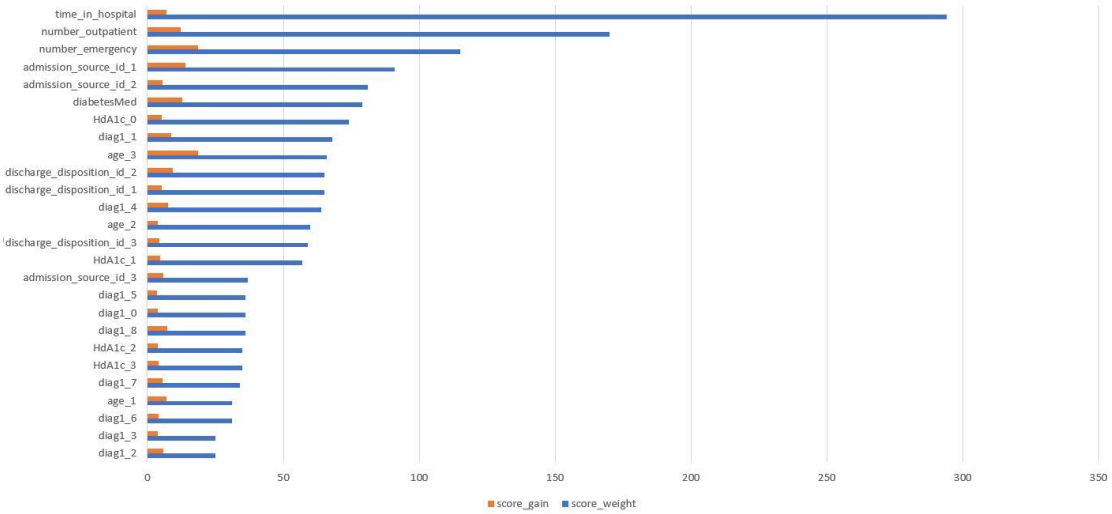

图 5-7 XGBoost 训练结果

从统计结果中可以发现，决策树中权重较高的因素包括：住院时间、门诊和急诊次数、入院方式、糖尿病处方药、HdA1c 的测量以及心脏病患者和老年群体等因素；分类信息增益较高的因素包括：老年群体、急诊次数、急诊入院以及糖尿病处方、心脏疾病诊断以及门诊次数等。

总体看来，住院时间的长短、来院次数（住院、急诊）、年龄和急诊入院等病患信息很大程度上的影响了再入院率，而 HdA1c 检测、初次诊断和糖尿病处方开具作为医院措施也对再入院率有一定的影响。

模型的精度评价如表 5-13 所示。

表 5-13 模型精度

精度指标	精度
准确率	0.655
AUC	0.694

5.4.5 多因素方差分析

方差分析用于多个样本均数的显著性检验基本原理是认为不同处理组的均数间的差别基本来源有两个：

(1) 实验条件，即不同的处理造成的差异，称为组间差异。用变量在各组的均值与总均值之偏差平方和的总和表示，记作SSb，组间自由度dfb。

(2) 随机误差，如测量误差造成的差异或个体间的差异，称为组内差异，用变量在各组的均值与该组内变量值之偏差平方和的总和表示，记作SSw，组内自由度dfw。

组内SSw、组间SSb除以各自的自由度（组内$dfw = n - m$，组间dfb=m-1，其中 n 为样本总数，m 为组数），得到其均方MSw和MSb，若$MSw/MSb \approx 1$，则变量的差异主要来自组内自身差异，与样本差异无关，另一种情况是$MSb >> MSw$，这种情况下变量的差异主要来自组间偏差，即样本间差异。

对所有因素针对再入院情况进行多因素方差分析得到 F 值和 P 值如下表：

表 5-14

特征 1	特征 2	F	PR(> F)
age	discharge_disposition_id	13.428624	2.496589e-04
age	admission_source_id	59.021488	1.773881e-14
age	time_in_hospital	29.603917	5.480422e-08
age	number_outpatient	29.957554	4.570169e-08
age	number_emergency	22.068727	2.681154e-06
age	diag1	3.792721	5.151612e-02
age	medical_specialty	1.371013	2.416781e-01
age	HdA1c	0.229185	6.321438e-01
age	diabetesMed	32.503185	1.238749e-08
discharge_disposition_id	admission_source_id	3.327705	6.816477e-02

从上表中分析可以得到对再次入院率影响较大的分类属性组合：

Age 和 admission_source_id，age 和 diabetesMed 以及 age 和 discharge_disposition_id，再综合 5.4.3 中得到可能对再入院率具有联合影响特征的属性组合，得到属性组检测表

age:admission_source_id

age:diabetesMed

age:discharge_disposition_id

diag1:HdA1c

age:HdA1c

diabetesMed:HdA1c

5.4.6 使用组合属性的决策树模型

使用 5.4.5 中的得到的 6 组属性重新生成虚拟变量，并将其用于训练提升决策树模型，得到如图 5-8 所示结果：

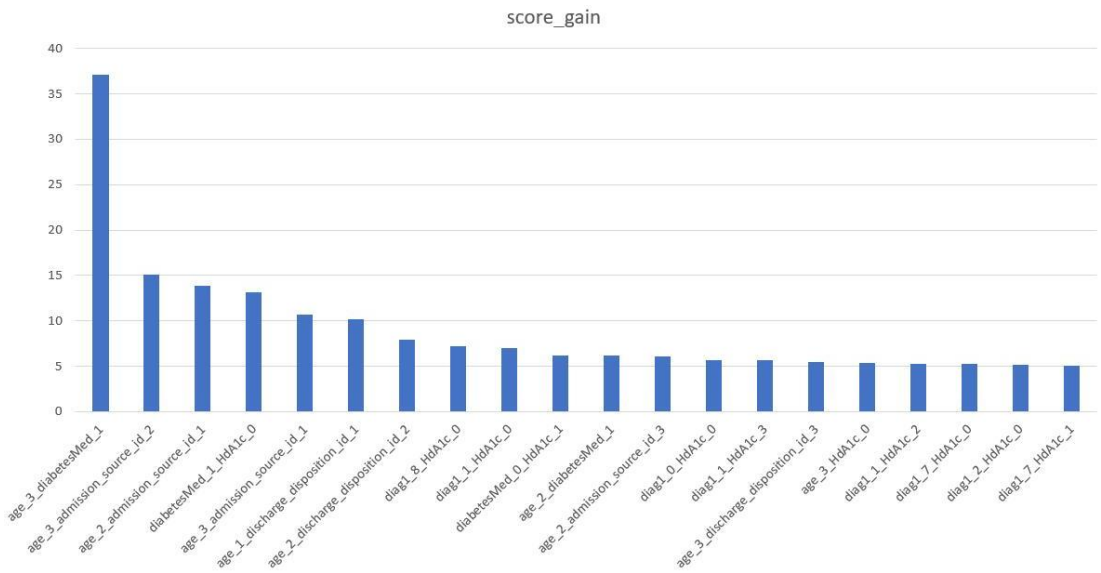

图 5-8 训练结果

通过观察各虚拟属性的分类信息增益可以发现：具有糖尿病药物处方的老年群体和急诊入院的中老年群体具有较高的再次入院率，同时有过糖尿病药物处方人群的 HdA1c 检测对再次入院也有一定的影响。

表 5-15 精度评定

精度指标	精度
准确率	0.621
AUC	0.653

5.4.7 结果分析

从单个因素的影响来看，患者的住院时间与再入院率具有显著的联系，随着住院时间增加，再入院率不断上升（与住院时间呈现显著相关的实验测试数、用药次数、诊断次数等因素均有同样的趋势）；来院次数（门诊、急诊、住院）和糖尿病处方与再入院率的相关性推测可能来自于病人的病史，即频繁入院的病人更倾向于患有长期性难以根治的疾病，从而导致了再次入院率的上升；另外，急诊病人和老年群体的再次入院率偏高，也符合实际情况。

考虑多个因素的影响，HdA1c 的检测与住院期间的换药情况构成的模式表现

出了对再次入院率的影响：在检测出 HdA1c 偏高之后及时换药有效地降低了在此入院的概率，而检测后指标平稳患者的再次入院率也显著低于未检测患者。同时，急诊入院的中老年人以及有过糖尿病药物处方的老年人，呈现出更容易在此住院的趋势。

5.5 问题五的解决

综合问题一至问题四的分析，可以得到如下结论：

（1） 根据数据中显示出的糖尿病医疗体系不够完善，对于数据表中的 7 万名糖尿病患者 HdA1c 的检测率只有 17%，血糖浓度检测率只有 5%，许多急诊患者和老年群体等高危人群并没有得到血糖方面的检测，总体再入院率超过 40%，7 万名患者中死亡率高达 3%。

（2） 医院应完善对于病人的病史、病例的管理和统计制度，从上文的分析中得知，患者的再入院率在患者的就诊记录（次数和类型）中有很大程度的体现，有效地管理病例数据可以针对不同情况的患者给予不同程度的建议或诊疗手段，例如，住院期间增加血糖指标的检测。

（3） 对于中老年病人等高危群体应该提高诊疗要求，增加如糖尿病这类在此类人群中高发疾病的检测措施，力图降低病人的再入院率，提高医院诊疗效率的同时也为病人节约医疗成本。

（4） 急诊病人、心脏疾病病人的血糖含量通常会有很大幅度的波动，从上文分析中也可以看出此类病人在进行 HdA1c 检测的情况下，再入院率有了明显的下降，因此应该着重对于此类病人的血糖监控。

6 模型评价与推广

6.1 模型优点

1. 在问题一构建的评价系统中，充分利用患者出入院数据，以患者出院时身体状况与入院前身体状况的对比表示医院的治疗效果，既可以为每一个患者的治疗效果进行评分，同时也可以对医院整体进行评价。

6.2 模型缺点

1. 对数据的分组是根据文献和相关经验人为划分的，可能存在一定的不合理性。

7　参考文献

[1] 汪珍珠, 杨海俊, 苏小琴,等. HbA1c 在鉴别应激性高血糖与糖尿病性高血糖中的作用的探讨[J]. 中国社区医师(医学专业), 2013, 15(1):255-256.

附录：相关源代码