**Computer Science and Application** 计算机科学与应用, 2019, 9(11), 2115-2128
Published Online November 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.911237

Hans汉斯

# Medical Expenses Prediction Based on Boosting Algorithms
## —Using Data of Nasopharyngeal Carcinoma (NPC)

**Lei Cao[1,2], Yihui He[1,2], Yuelin Liu[1,2], Yushan Jiang[1,2*]**

[1]Mathematics and Statistics School of Northeastern University at Qinhuangdao, Qinhuangdao Hebei
[2]Institute of Data Analysis and Intelligence Computing, Northeastern University at Qinhuangdao, Qinhuangdao Hebei
Email: *lhospital@foxmail.com

Received: Nov. 1st, 2019; accepted: Nov. 14th, 2019; published: Nov. 21st, 2019

## Abstract

**The data of this paper come from 2064 cases of NPC in a Cancer Hospital of Guangdong Province. We mine the data and predict the medical cost per patient. This paper studies the data through the following four steps. First, we select the characteristics of patients' age, gender, TNM diagnosis stage and length of stay as the prediction variables. Then, we build the cost prediction model based on the regression decision tree algorithm (CART). Then, two boosting algorithms, AdaBoost and gradient boosting, are used to improve the existing model. Then, through the visual comparison and regression evaluation index, the effect of the prediction model established by the three algorithms is analyzed and compared, and the best DBRT (gradient boosting decision tree) prediction model is obtained, with the prediction accuracy of about 85%. Finally, the significance of the model based on boosting algorithm is explained through the feature importance and partial dependency graph, which provides a reference for the allocation of medical insurance resources and the expected cost of a single case.**

## Keywords

**CART, NPC, AdaBoost, Gradient Boosting, DBRT, Regression Valuation Index, Feature Importance, Partial Dependency**

# 基于Boosting算法的医疗费用预测
## ——以鼻咽癌为例

**曹　蕾[1,2]，何轶辉[1,2]，柳岳霖[1,2]，姜玉山[1,2*]**

*通讯作者。

文章引用: 曹蕾, 何轶辉, 柳岳霖, 姜玉山. 基于 Boosting 算法的医疗费用预测[J]. 计算机科学与应用, 2019, 9(11): 2115-2128. DOI: 10.12677/csa.2019.911237

[1]东北大学秦皇岛分校数学与统计学院，河北 秦皇岛
[2]东北大学秦皇岛分校数据分析与智能计算研究所，河北 秦皇岛
Email: *lhospital@foxmail.com

收稿日期：2019年11月1日；录用日期：2019年11月14日；发布日期：2019年11月21日

## 摘 要

**本文的数据来源于广东省某肿瘤医院，共计2064个鼻咽癌病案，我们对其进行数据挖掘，并预测病人的医疗费用。本文通过以下四步对数据进行研究。首先，我们选取了病人的年龄、性别、TNM诊断分期以及住院天数等特征为预测变量。然后，基于回归决策树算法(CART)建立费用预测模型。其后，分别使用两种Boosting算法，AdaBoost和Gradient Boosting对已有模型进行改进。接着，通过直观比照和回归评价指标，分析三种算法建立的预测模型的效果并进行比较，得到效果最好的DBRT (Gradient Boosting Decision Tree)预测模型，其预测准确率约为85%。最后，通过特征重要度和部分依赖关系图，解释基于Boosting算法的模型的现实意义，为医疗保险资源的分配和单个病例预期费用提供了参考。**

## 关键词

**CART，鼻咽癌，AdaBoost，Gradient Boosting，DBRT，回归评价指标，特征重要度，部分依赖关系**



## 1. 引言

尽管医疗保险资金再分配是医疗保险机制中重要的一环，但在日常的运行过程中常常遇到资源分配不合理的问题。而对患者来说，难以预先得知较为准确的治疗费用，也将对其治疗过程产生不利的影响。所以，研究医疗费用预测对建立单病种的支付额度和患者的治疗十分重要。但是，由于很多因素的影响，单病种支付额度不可能是始终不变的，尤其是癌症这类医疗费用极差较大的病种来说，其治疗过程中产生的费用更是难以准确预测。究竟该如何准确给出治疗费用是摆在研究人员面前的棘手问题。

国内外对于该问题已有一定的研究，近些年，Robert B. Fetter 等人提出的 DRGs (Prospective Payment System Based On Diagnosis Related Groups) [1]，即诊断相关分类，有着迅速发展，它是当今世界公认的比较先进的支付方式之一。它以病例组合为基本依据，考虑了患者的个体特征以及并发症和合并症情况等因素，将诊疗过程相似、费用支出相近的病例分到同一个组，进而接受统一标准的诊疗预付费。这一方法通过统一的诊断分组定额支付，激励医院加强质量管理、优化资源利用。而国内对于医保赔付和医疗费用预测问题也有一些成果，林倩、杜剑亮、Ai-Jing Luo 等人[2] [3] [4]利用决策树实现 DRGs，并对医保赔付做出指导。张凯、王若佳[5] [6] [7]引入数据挖掘技术解决该问题，提供了新的思路。

本文对大量鼻咽癌患者病历进行数据挖掘，根据影响疾病治疗费用的主要因素，基于回归决策树构建了针对鼻咽癌(NPC)患者个体的费用预测模型，并运用 Boosting 算法进行模型改进，最后解释模型的现实意义，为医疗保险资源的分配和单个病例预期费用提供参考。

## 2. 描述性统计

### 2.1. 数据预处理

本文处理的鼻咽癌患者数据采样于广东省某肿瘤医院，但由于原始数据存在缺失、不规范等原因，需要对数据进行预处理。针对此次课题，本小组利用莱文斯坦编辑距离算法、正则表达式、余弦相似性等方法对原始数据进行了如下处理：剔除罕见的接受手术的样本、限制 ICD 编码范围、去除费用大于三倍标准差的样本(如表 1 所示)、TNM 分期标准化、费用明细整合，最终获得 2064 例标准化数据[8]。

**Table 1.** Removing data out of three times the standard deviation
**表 1.** 去掉大于三倍标准差的数据

| | 标准差 | 平均值 |
|---|---|---|
| 删除前 | 168,842.4 | 244,894.1 |
| 删除后 | 121,342 | 226,751 |

### 2.2. 描述性统计

#### 2.2.1. 基本情况

对鼻咽癌病例数据进行描述性统计，具体统计情况如表 2 [8]。

**Table 2.** Basic statistics
**表 2.** 基本情况统计

| 项目 | 分类 | 人次(人) | 平均费用(元) | 占比(%) |
|---|---|---|---|---|
| 年龄 | ≤30 | 212 | 153,781 | 10.27 |
| | 31~40 | 480 | 160,551 | 23.25 |
| | 41~50 | 740 | 170,860 | 35.85 |
| | 51~60 | 492 | 156,222 | 23.83 |
| | ≥61 | 140 | 158,057 | 6.78 |
| 性别 | 男 | 1605 | 162,233 | 77.76 |
| | 女 | 460 | 162,761 | 22.28 |
| TNM 分期 | 一期 | 660 | 169,298 | 31.97 |
| | 二期 | 120 | 175,484 | 5.81 |
| | 三期 | 528 | 169,177 | 25.58 |
| | 四期 | 752 | 149,461 | 36.43 |
| | 未知 | 4 | 144,164 | 0.19 |
| 治疗时长 | 0~1 年 | 1149 | 106,585 | 55.66 |
| | 1~2 年 | 410 | 195,115 | 19.86 |
| | 2~3 年 | 227 | 243,938 | 10.99 |
| | 3~4 年 | 138 | 277,088 | 6.69 |
| | 4 年以上 | 140 | 278,693 | 6.78 |

#### 2.2.2. 社会学数据分布情况

在 2064 例鼻咽癌患者病例中，男性共 1605 例，占病例总数的 77.76%，女性 460 例，占病例总数的

22.24%。由年龄和支付的密度散点图，即图 1，可见年龄多分布于区间[41, 50]岁，支付总费用在 10 万元到 20 万元的鼻咽癌患者是密度最大的[8]。

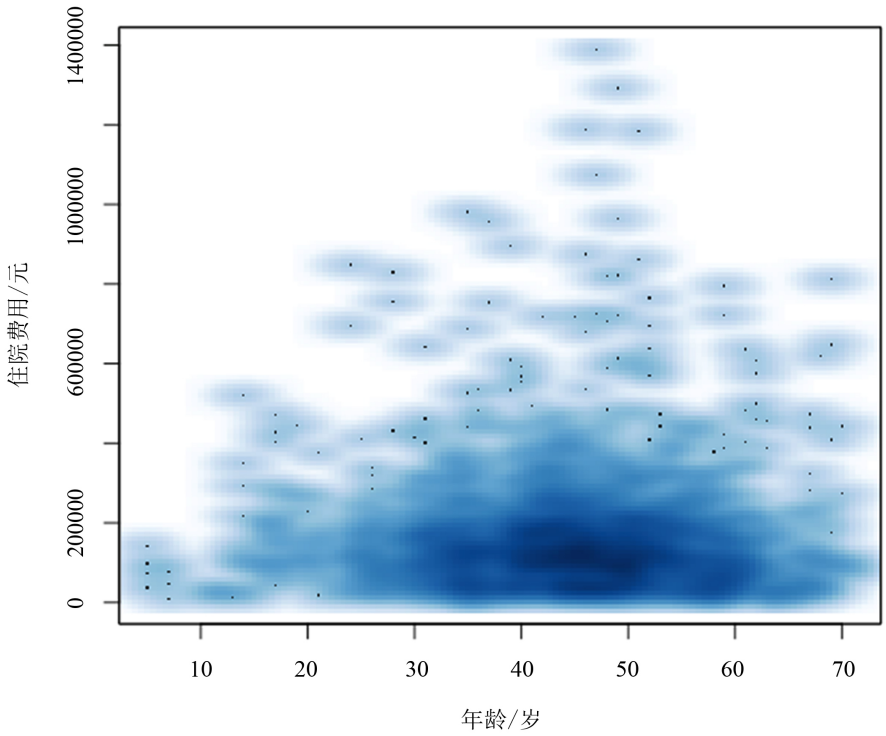



**Figure 1.** Density scatter plot of hospital expenses varying with age
**图 1.** 住院费用随年龄变化的密度散点图

#### 2.2.3. TNM 分期数据分布情况

观察病例数据在不同分期中的分布，发现男性患者较多的现象普遍存在，并且鼻咽癌分期主要集中在三期、四期中，占比 62%。

观察每个 TNM 分期与平均费用的关系，可知一期费用普遍较低，而二期费用最高，我们推测导致该现象的原因为二期鼻咽癌症状相对一期更为严重，但相较于三期、四期更容易被治愈，存活率相对较高，导致二期患者疗程更长，花费更多[8]。

#### 2.2.4. 对数据集的概况性描述

1) 男性更容易患鼻咽癌，占总患者数 78%；
2) 患者多为 40 至 60 岁的中老年人；
3) 患者的治疗费用集中在 20 万左右，其中二期的鼻咽癌患者花费最多；
4) 就诊患者主要为鼻咽癌中晚期(三期和四期)患者。

## 3. 基于 CART 算法的费用预测

### 3.1. CART 回归树生成算法原理

CART 回归树的生成算法基于最小二乘法。在训练数据集所在的输入空间中，递归地将每个区域划分成两个子区域并决定每个子区域上的输出值，构建二叉决策树。例如给定数据集 $D$，输出回归树 $f(x)$：

1) 第一步、选择数据集 $D$ 中最优切分变量 $x_j$ 与切分点 $s$：求解：

$$\min_{j,s} = \left[ \min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2 \right]$$

遍历变量$x_j$，对固定的切分变量$x_j$扫描切分点$s$，选择使上式达到最小值的二元有序数对$(j, s)$。

2) 第二步、用选定的有序数对$(j, s)$划分区域并决定相应的输出值：

$$R_1(j,s) = \{x | x^{(j)} \leq s\}, \quad R_2(j,s) = \{x | x^{(j)} > s\}$$

其中$R_m$是一个用二元有序数对(切分变量及其取值)表示的定义域。由此进一步可以解得枝结点和叶节点的因变量的预测值(回归拟合值)为其结点内样本均值：

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m(j,s)} y_i, \quad x \in R_m, \quad m = 1, 2$$

3) 第三步、对两个子区域(枝结点)重复前两步，直到满足停止条件(最大树深度、结点内最小样本量、复杂度参数阈值等)。

4) 第四步、输出回归决策树$f(x)$：

$$f(x) = \sum_{m=1}^{M} \hat{c}_m I(x \in R_m)$$

$f(x)$将输入空间划分为$m$个区域$R_i (i = 1, 2, \ldots, m)$，即$m$个叶节点($m$类) [8]。

### 3.2. 基于 CART 算法的费用预测

我们使用 Python 的 sklearn.tree 库实现 CART 回归预测，建立并训练最大深度为 4 的 CART 回归树，之后进行费用预测并与真实数据进行对比，得到如下预测费用与年龄分布散点图，由图 2 可知，CART 算法预测的费用集中于 1 万至 3 万，对于费用较低或费用较高的病例预测效果较差。

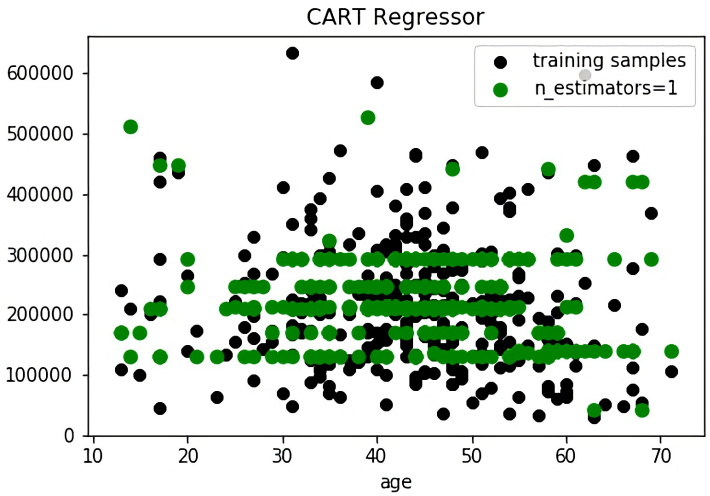



**Figure 2.** Comparison between predicted and actual expenses based on CART
**图 2.** 基于 CART 算法的预测费用与真实费用对比

## 4. 基于 Boosting 算法的费用预测

Boosting (提升方法)是 Ensemble Learning 算法的一类。这类算法的思想是将弱学习器提升为强学习

器。即先用初始训练集训练出一个基学习器，再根据基学习器表现对训练样本分布进行调整，那些基学习器判断错误的样本，在之后会被赋予更大的权值，然后基于调整后的样本来训练下一个基学习器，一直反复进行，直到达到理想效果。在 2000 年左右，Friedman 的几篇论文提出 Boosting 算法的成功实现[9] [10]。

Boosting 常用的两种方法为 AdaBoost (Adaptive Boost)和梯度提升(Gradient Boosting)。本章及下一章选取已出院鼻咽癌病例样本进行如下的费用预测，即所得到的费用可以认为是该病例治疗的最终花费。其中，取性别、年龄、住院次数、住院天数和首诊分期为自变量，建立并训练模型预测最终费用，并比较 CART 算法、AdaBoost 算法、梯度提升算法的预测效果。经过反复的实验，最大树深度为 4 且弱学习器为 300 个的梯度提升算法表现最好。

## 4.1. AdaBoost 算法原理

由于 CART 预测效果较差，我们选用集成算法中的提升方法进行改进。首先研究基于 AdaBoost 算法的费用预测。

### 4.1.1. AdaBoost 算法思路

AdaBoost 的思路为，先从初始训练集中训练出一个基学习器，再根据基学习器的表现对训练样本分布进行调整，对所有基学习器采用加权结合，增大分类或回归误差小的基学习器的权值，减少误差率大的基学习器的权值。

### 4.1.2. AdaBoost 算法步骤

输入：训练集 $D=(x_i,y_i)_{i=1}^m$，其中，$x_i\in\chi\subseteq\mathbb{R}^n$；基学习算法 $\mathfrak{L}$；基学习器个数 $T$。

过程：

1) 初始化训练样本的权值分布

$$D_1=(w_{11},\ldots,w_{1i},\ldots,w_{1m}),w_{1i}=1/m,i=1,2,\ldots,m\text{。}$$

2) 对迭代轮次 $t=1,2,\ldots,T$。

a) 使用具有当前分布 $D_t$ 的训练数据集训练基学习器 $h_t=\mathfrak{L}(D,D_t)$；

b) 计算训练集上的样本最大误差：

$$E_t=max|y_i-h_t(x_i)|,i=1,2,\ldots,m$$

计算每个样本的相对误差：

如果是线性误差，则 $e_{ti}=|y_i-h_t(x_i)|/E_t$；

如果是平方误差，则 $e_{ti}=(y_i-h_t(x_i))^2/E_t^2$；

如果是指数误差，则 $e_{ti}=1-\exp\{-|y_i-h_t(x_i)|/E_t\}$；

c) 基学习器在训练数据集上的回归误差率：

$$\varepsilon_t=\sum_{i=1}^m w_{ti}e_{ti}$$

d) 基学习器的权重系数：

$$\alpha_t=\frac{\varepsilon_t}{1-\varepsilon_t}$$

e) 更新训练集的样本分布：

$$D_{t+1}=(w_{t+1,1},\ldots,w_{t+1,i},\ldots,w_{t+1,m})$$

$$w_{t+1,i} = \frac{w_{ti}}{Z_t}\alpha_t^{1-e_{ti}}$$

其中， $Z_t$ 是规范化因子：

$$Z_t = \sum_{t=1}^{m} w_{ti}\alpha_t^{1-e_{ti}}$$

f) 构建基学习器的线性组合，得到最终的强学习器：

$$\sum_{t}^{T} \alpha_t h_t(x)$$

$$H(x) = \sum_{t=1}^{T} \left( ln\frac{1}{\alpha_t} \right) g(x)$$

输出：强学习器 $H(x)$ [11] [12]。

## 4.2. Gradient Boosting 算法原理

由于对 AdaBoost 算法对异常样本敏感，异常样本在迭代中可能获得较大权重，影响强学习器的预测准确性。接下来将使用 Gradient Boosting 算法进一步改进费用预测模型。

### 4.2.1. Gradient Boosting 算法

由 Freidman 提出的梯度提升(gradient boosting, GB)算法[10]，若将梯度提升算法里的弱算法选择为决策树，则的到 GBDT 算法(Gradient Boosting Decision Tree)。该算法主要利用最速下降法的近似方法，利用损失函数的负梯度在当前模型的值

$$-\left[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)}$$

作为回归问题提升树算法中的残差的近似值，拟合一个回归树。

### 4.2.2. Gradient Boosting 算法步骤

接下来是该算法的训练步骤[11] [12]：

给定训练数据集

$$T = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}, \quad x_i \in X \subseteq \mathbb{R}^n, \quad y_i \in Y \subseteq \mathbb{R}。$$

其中 $n$ 是变量自变量的个数。损失函数 $L(y, f(x))$。

1) 初始化

$$f_0(x) = arg\min_c \sum_{i=1}^{N} L(y_i, c)。$$

2) 对 $m = 1, 2, \ldots, M$。

a) 对 $i = 1, 2, \ldots, N$，计算

$$r_{mi} = -\left[\frac{\partial L(y, f(x_i))}{\partial f(x_i)}\right]_{f(x)=f_{m-1}(x)};$$

b) 对$r_{mi}$拟合一个回归树，得到第$m$棵树的叶节点区域$R_{mj}$，　$j=1,2,\ldots,J$；

c) 对$j=1,2,\ldots,J$，计算

$$c_{mj}=arg\min_{c}\sum_{x_i\in R_{mj}}L(y_i,f_{m-1}(x_i)+c)$$

；

d) 更新$f_m(x)=f_{m-1}(x)+\sum_{j=1}^{J}c_{mj}I(x\in R_{mj})$；

3) 得到回归树

$$\hat{f}(x)=f_M(x)=\sum_{m=1}^{M}\sum_{j=1}^{J}c_{mj}I(x\in R_{mj})$$

。

### 4.3. 基于 Boosting 算法的费用预测

我们使用 Python 的 sklearn.ensemble 库分别实现 AdaBoost 算法、GBDT 算法，最终建立如下两种费用预测模型模型：

基于 AdaBoost 算法：以最大深度为 4 的 CART 回归树作为基学习器，基学习器 300 个，学习速度为 1，损失函数为 linear 函。

基于 GBDT 算法：以最大深度为 4 的回归树作为基学习器，基学习器 100 个，学习速度为 0.1，损失函数为 l s 函数。

下面进行费用预测并与真实数据进行对比，得到预测费用与年龄分布的散点图，如图 3、图 4。

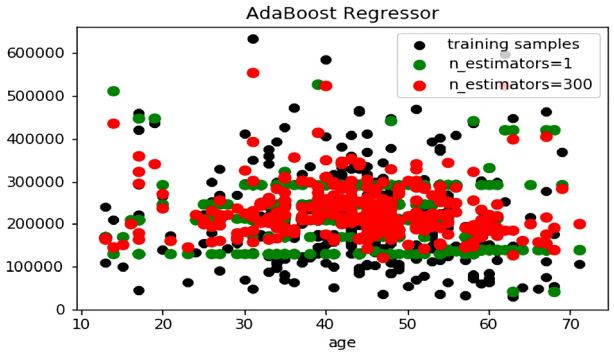



**Figure 3.** Comparison between predicted and actual expenses based on AdaBoost
**图 3.** 基于 AdaBoost 算法的预测费用与真实费用对比

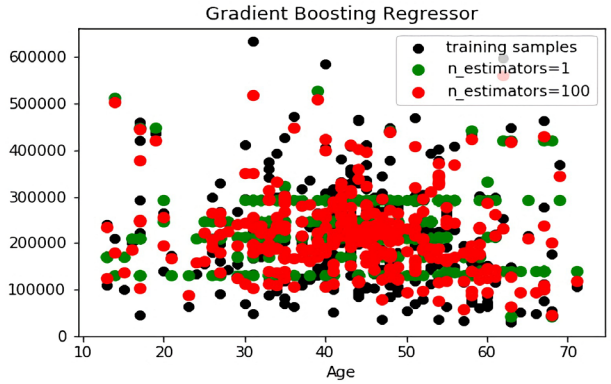



**Figure 4.** Comparison between predicted and actual expenses based on Gradient Boosting
**图 4.** 基于 Gradient Boosting 算法的预测费用与真实费用对比

## 5. 预测结果分析及评价

### 5.1. 三种算法对总费用的预测结果与真实结果的比较

现对比研究基于 CART 算法、AdaBoost 算法、GBDT 算法对总费用预测的结果。

直观地，由图 5 可知，三种算法预测效果有明显不同。尤其值得指出 Boosting 算法的预测结果较 CART 算法的预测结果更加分散，其中 GBDT 算法的预测结果最为分散，且预测结果的分布更接近于真实值，Boosting 算法对于预测效果有着明显的改进。

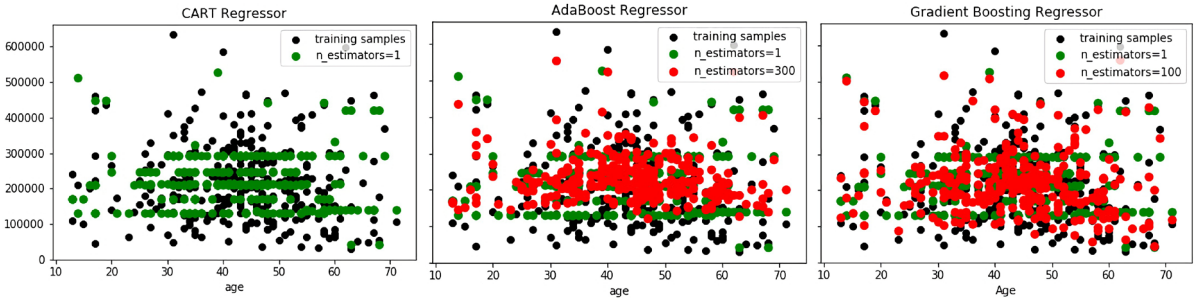



**Figure 5.** Comparison among predicted results by using CART, AdaBoost and GBDT
**图 5.** 三种算法费用预测效果对比

图 6 是基于 AdaBoost 算法和 GBDT 算法的费用预测结果与真实值比较的折线图，黑色的圆形散点为对应患者的总费用的真实值(经过排序)，蓝色曲线为基于 AdaBoost 算法的预测结果，红色曲线为基于 GBDT 算法的预测结果。显然红色曲线更趋近于真实值的分布，即基于 GBDT 算法的预测效果的效果最好。

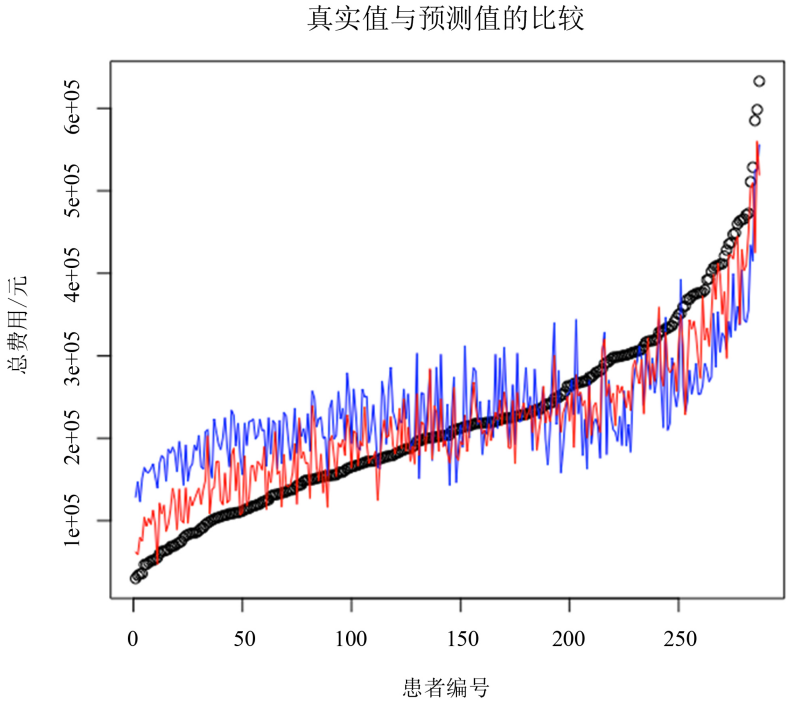



**Figure 6.** Comparison between predicted results by using AdaBoost and GBDT
**图 6.** 基于 AdaBoost 算法、GBDT 算法的预测效果对比

### 5.2. 对集成学习算法的 5 个回归评价指标

我们将通过 5 个回归评价指标，量化对比三种模型的预测效果。下面先引入这 5 个指标，分别是均方对数误差平均值(Mean squared logarithmic error, MSLE)、平均绝对误差(Mean Absolute Error, MAE)、中值绝对误差(Median Absolute Error, MedAE)、确定系数(The Coefficient of Determination, $R^2$)、解释方差分(Explained Variance Score, EVS)。

1) 均方对数误差平均值(Mean squared logarithmic error, MSLE)。由于本文中预测值较大，均方误差受单位影响较大，于是采用 MSLE 来代替均方误差。

$$\mathrm{MSLE}(y,\hat{y}) = \frac{1}{n_{\mathrm{samples}}}\sum_{i=0}^{n_{\mathrm{samples}}-1}(\ln(1+y_i)-\ln(1+\hat{y}_i))^2$$

2) 平均绝对误差(Mean Absolute Error, MAE)是指预测值与真实值之间的平均差值。

$$\mathrm{MAE}(y,\hat{y}) = \frac{1}{n_{\mathrm{samples}}}\sum_{i=0}^{n_{\mathrm{samples}}-1}|y_i-\hat{y}_i|$$

3) 中值绝对误差(Median Absolute Error, MedAE)对异常值有很好的稳定性，它通过获取预测值和真实值之间差值的绝对值的中值来计算损失。

$$\mathrm{MedAE}(y,\hat{y}) = \mathrm{median}(|y_1-\hat{y}_1|,\ldots,|y_n-\hat{y}_n|)$$

4) 决定系数(The Coefficient of Determination)即 $R^2$，用来反映模型拟合效果的好坏。最大取值为 1，但它也可能是负数。一般来说，$R^2$ 越大，表示模型拟合效果越好，由于随着样本数量的增加，$R^2$ 必然增加，无法真正定量说明模型拟合效果，只能大概定量。

$$R^2(y,\hat{y}) = 1-\frac{\sum_{i=0}^{n_{\mathrm{samples}}-1}(y_i-\hat{y}_i)^2}{\sum_{i=0}^{n_{\mathrm{samples}}-1}(y_i-\bar{y})^2}$$

其中 $\bar{y} = \frac{1}{n_{\mathrm{samples}}}\sum_{i=0}^{n_{\mathrm{samples}}-1}y_i$

5) 解释方差分(Explained Variance Score, EVS)是用来解释回归模型方差大小的分数，取值范围为 0 到 1，EVS 越大，表示回归效果越好。

$$\mathrm{EVS}(y,\hat{y}) = 1-\frac{var(y-\hat{y})}{var(y)}$$ 。

### 5.3. 基于 5 个回归评价指标的三种算法预测效果及比较

我们使用 Python 语言的 sklearn 库 metrics 模块计算上述五个回归评价指标，三种算法预测效果的量化数值如表 3、表 4 所示。

**Table 3.** Evaluation indicators of prediction models
**表 3.** 三种预测模型的回归评价指标

| 算法 | MSLE | EVS | MAE | MedAE | $R^2$ |
|---|---|---|---|---|---|
| CART | 0.20 | 0.41 | 67,412.85 | 61,480.30 | 0.41 |
| AdaBoost | 0.21 | 0.55 | 66,946.16 | 71,903.27 | 0.53 |
| Gradient Boosting | 0.07 | 0.85 | 34,588.13 | 29,508.20 | 0.85 |

**Table 4.** Evaluation indicators of adjusted prediction models
**表 4.** 三种预测模型拟合效果的改进

| 算法改进 | MSLE | EVS | MAE | MedAE | $R^2$ |
|---|---|---|---|---|---|
| CART→AdaBoost | +0.01 | +0.14 | −466.69 | +10,422.97 | +0.12 |
| AdaBoost→Gradient Boosting | −0.14 | +0.30 | −32,358.03 | −42,395.07 | +0.32 |

由上表中数据可知，虽然 AdaBoost 算法在 MSLE、EVS、MAE、$R^2$ 评价指标上的表现优于 CART 算法，但 AdaBoost 算法和 CART 算法的预测效果都较差，其拟合效果均低于 60%，即这两种算法的预测结果与真实值并不接近。而 GBDT 算法在五个指标上的表现远远优于 AdaBoost 算法和 CART 算法，其拟合效果达到了 85%，表明 GBDT 算法的预测结果与真实值较为相近，在三种算法中效果最好。该结论与先前对比的直观结论相符合。

为了检验费用预测模型是否过拟合，我们从总数据集中抽取 20%的样本作为测试集，代入进行预测效果评价并与之前的评价结果进行比较。综合表 5 和表 6 可知，GBDT 的拟合程度达到了 85%，在测试集中拟合效果 83%。基于 GBDT 算法建立的费用预测模型没有过拟合。

**Table 5.** Evaluation indicators of prediction models on test set
**表 5.** 将测试集分别带入三种预测模型的拟合效果

| 算法 | MSLE | EVS | MAE | MedAE | $R^2$ |
|---|---|---|---|---|---|
| CART | 0.25 | 0.29 | 79,480.59 | 70,992.09 | 0.29 |
| AdaBoost | 0.23 | 0.57 | 71,578.61 | 74,328.32 | 0.56 |
| Gradient Boosting | 0.09 | 0.83 | 40,957.83 | 33,752.57 | 0.83 |

**Table 6.** Evaluation indicators of adjusted prediction models on test set
**表 6.** 三种预测模型的对于测试集拟合效果的改进

| 算法改进 | MSLE | EVS | MAE | MedAE | $R^2$ |
|---|---|---|---|---|---|
| CART→AdaBoost | −0.02 | +0.28 | −7901.98 | +3336.23 | +0.27 |
| AdaBoost→Gradient Boosting | −0.14 | +0.26 | −30,620.78 | −40,575.75 | +0.27 |

## 5.4. 费用预测模型的可解释性

通过简单地可视化树结构，可以轻松解释单个决策树。然而，集成算法模型包括数百个回归树，因此通过对各个树的可视化无法轻易解释它们。然而可以通过计算特征重要程度以及绘制部分依赖图，来直观的概括和解释 Boosting 模型。

### 5.4.1. 特征重要程度

Friedman 在论文中提出在使用决策树集成算法(Tree Ensemble)时评价特征在模型中重要程度的方法[8]，便于更好的理解特征和模型。接下来介绍一下特征重要程度的计算原理，并给出已建立的三种回归模型的特征重要性。

单个决策树通过选择适当的分裂点本质上执行特征选择。此信息可用于衡量每个功能的重要性；基本思想是：在树的分裂点中使用特征的次数越多，特征就越重要。通过简单地平均每棵树的特征重要性，可以将这种重要性概念扩展到决策树集成算法中。

特征$x_j$在整个模型中的重要程度为：

$$\hat{J}_j^2 = \frac{1}{M}\sum_{m=1}^{M}\hat{J}_j^2(T_m)$$

其中，$M$ 是模型中树的数量。特征 $x_j$ 在单独一个树上的特征重要度为：

$$\hat{J}_j^2(T) = \sum_{m=1}^{L-1}\hat{I}_t^2(v_t = j)$$

其中，$L-1$ 是树中非叶子节点数量，$v_t$ 表示在内部节点 $t$ 进行分裂时选择的特征，$\hat{I}_t^2$ 是内部节点分裂后平方损失的减少量。

接下来给出五个特征分别在三种算法下的特征重要程度，由表 7 可知三种算法均将年龄、住院次数、住院天数作为训练模型时较为重要的特征，然而性别和 TNM 分期对预测模型的贡献较少。

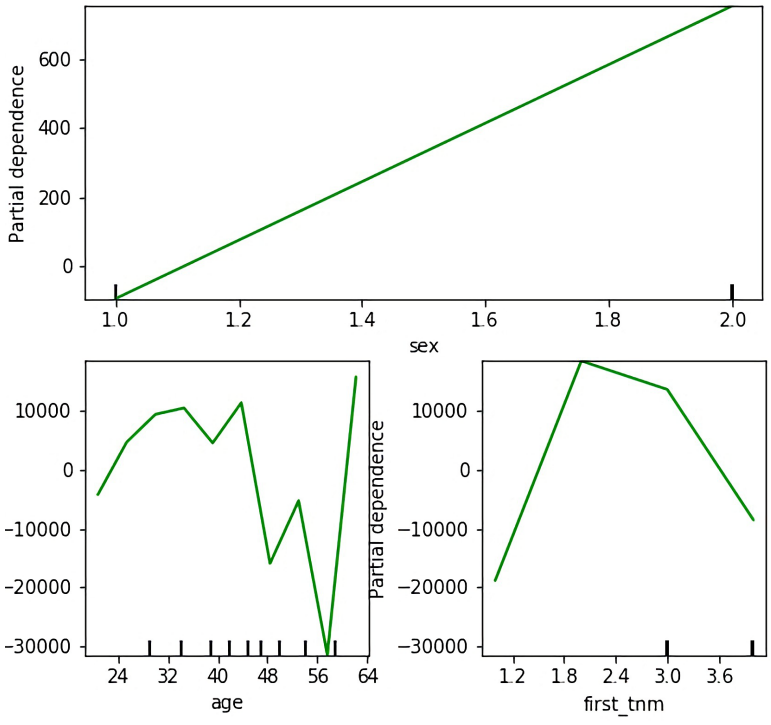
**Table 7.** Magnitude of each attribute by using CART, AdaBoost and GBDT
**表 7.** 五个特征分别在三种算法下的特征重要程度

| 算法 | 性别 | 年龄 | TNM 分期 | 住院次数 | 治疗天数 |
|---|---|---|---|---|---|
| CART | 0 | 0.205 | 0.022 | 0.393 | 0.380 |
| AdaBoost | 0.017 | 0.293 | 0.049 | 0.177 | 0.463 |
| Gradient Boosting | 0.034 | 0.268 | 0.037 | 0.220 | 0.442 |

### 5.4.2. 部分依赖关系

部分依赖图显示了训练模型得到的目标函数与一组特征之间的依赖关系。单项部分依赖图只研究一个特征，边缘化了所有其他特征的值，直观的展示一个特征对模型的影响。

由图 7，即 Gradient Boosting 预测模型中五个特征的单项部分依赖图可以得到以下关系：



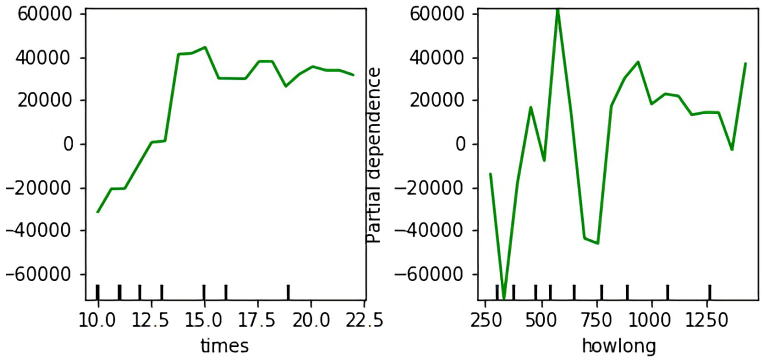



**Figure 7.** Single partial dependency graph based on two target characteristics
**图 7.** 基于两个目标特征的单项部分依赖图

1) 男性患者(sex = 2)倾向于产生更多费用；

2) 小于 45 岁或大于 60 岁的患者，费用呈现年龄越大花费越多的趋势，而大于 45 岁小于 60 岁的患者，费用呈现年龄越大花费越少的趋势；

3) 住院次数与模型有较为规律的依赖关系，住院次数小于 15 次，次数越多费用越高，而住院次数大于 15 次，则费用趋于稳定；

4) 治疗天数对于模型的依赖关系较为复杂。

多项部分依赖图研究多个特征，边缘化了其他特征的值，直观的展示了多个特征对模型的综合影响。由图 8，即 Gradient Boosting 预测模型中四个特征的多项部分依赖图可以得到以下关系：

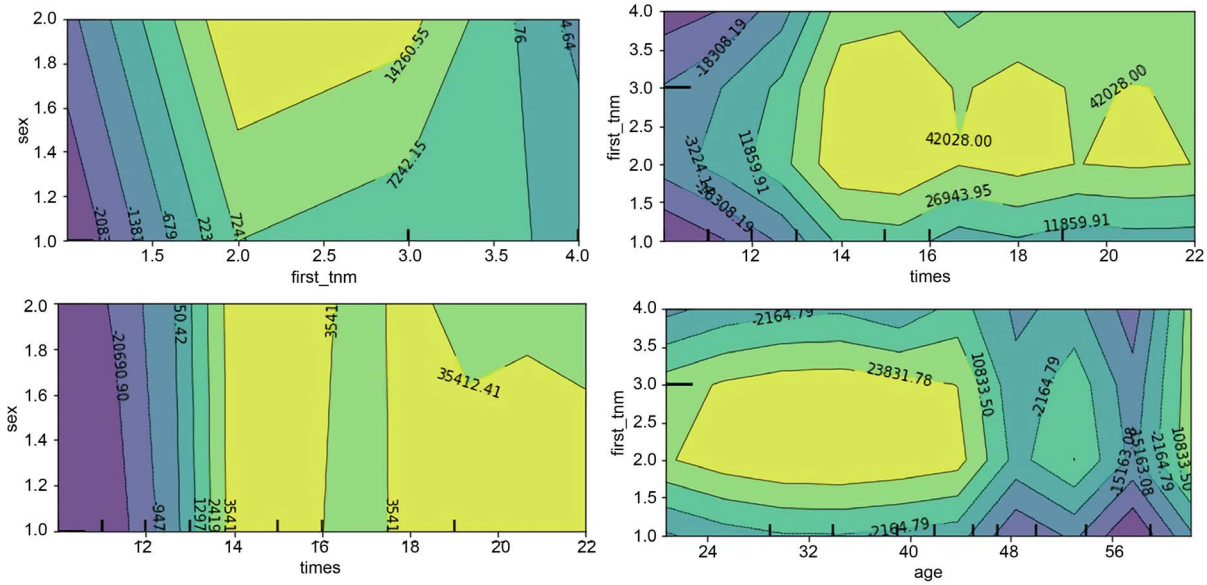



**Figure 8.** Multi-partial dependency graph based on two target characteristics
**图 8.** 基于两个目标特征的多项部分依赖图

1) TNM 分期在二期和三期的男性患者群体将产生大量费用；

2) 住院次数大于 13 次的患者，无论男女，都易产生大额医疗费用，其中二、三期的患者中这种依赖关系更显著；

3) 小于 45 岁的患者，费用呈现年龄越大花费越多的趋势，其中二、三期的患者中该关系更显著。

通过分析部分依赖图，能够直观的理解和解释基于GBDT算法预测模型中复杂的目标函数，这是分析基于Boosting算法以及其他集成算法的模型的有效方法之一。

## 致　谢

在该项目研究阶段，东北大学秦皇岛分校以及东软集团医疗事业部给予团队许多支持，特在此致谢。同时该项目受到东北大学秦皇岛分校创新训练项目，教育部科技发展中心科研创新基金的支持，使得项目有充足的资金得以顺利进行。最后感谢姜玉山博士为项目指明了研究方向，同时对论文的撰写提出了许多建议，使得项目及论文撰写都圆满成功。

## 基金项目

东北大学秦皇岛分校创新训练项目，教育部科技发展中心科研创新基金(2018A03031)。

## 参考文献

[1] Fetter, R.B., Shin, Y., Freeman, J.L., Averill, R.F. and Thompson, J.D. (1984) Case Mix Definition by Diagnosis-Related Groups. *Medical Care*, **18**, 1-53.

[2] 林倩, 王冬, 郭煜, 詹志颖, 吴志明. 基于CHAID算法的阑尾炎患者DRGs分组研究[J]. 卫生经济研究, 2017(8): 29-32.

[3] 杜剑亮, 刘骏峰, 陈倩. 不同决策树算法建立DRGs模型的差异[J]. 中国病案, 2014, 15(7): 38-41.

[4] Luo, A.-J., Chang, W.-F., Xin, Z.-R., Ling, H., Li, J.-J., Dai, P.-P., Deng, X.-T., Zhang, L. and Li, S.-G. (2018) Diagnosis Related Group Grouping Study of Senile Cataract Patients Based on E-CHAID Algorithm. *International Journal of Ophthalmology*, **11**, 308-313.

[5] 张凯. 数据挖掘技术在医疗费用数据中的应用研究[D]: [硕士学位论文]. 北京: 北京邮电大学, 2015.

[6] 王若佳, 魏思仪, 赵怡然, 王继民. 数据挖掘在健康医疗领域中的应用研究综述[J]. 图书情报知识, 2018(5): 114-123+9.

[7] 徐昆. 业健康保险与医疗大数据对接交互系统研究[J]. 金融理论与实践, 2018(7): 103-108.

[8] 曹蕾, 柳岳霖, 何轶辉, 姜玉山. 基于决策树的DRGs制度研究——以鼻咽癌为例[J]. 应用数学进展, 2019, 8(6): 1121-1132.

[9] Friedman, J.H., Hastie, T. and Tibshirani, R. (2000) Additive Logistic Regression: A Statistical View of Boosting. *Annals of Statistics*, **28**, 337-407. https://doi.org/10.1214/aos/1016218223

[10] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics*, **29**, 1189-1232. https://doi.org/10.1214/aos/1013203451

[11] 李航. 统计学习方法[M]. 北京: 清华大学出版社, 2012.

[12] 吕晓玲, 宋捷. 大数据挖掘与统计机器学习[M]. 北京: 中国人民大学出版社, 2016.