Advances in Applied Mathematics 应用数学进展, 2022, 11(5), 2858-2867
Published Online May 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.115303

Hans汉斯

基于两层Stacking模型的累积索赔额预测及定价研究

司晶硕

河北工业大学理学院，天津

收稿日期：2022年4月25日；录用日期：2022年5月19日；发布日期：2022年5月26日

摘　要

在传统的索赔额预测中，广义线性模型(GLM)是一种常用的方法。近年来，机器学习算法在该领域也取得了良好的效果，为索赔额预测提供了新的选择。在大数据时代，如何更准确地进行预测，是亟待解决的问题。为了解决该问题，本文利用两层Stacking模型，两种其他集成学习算法和广义线性模型对累积索赔额进行预测。通过比较各算法的均方根误差及平方绝对误差，可发现包括Stacking的集成算法精度全部优于传统广义线性模型。最后，本文利用累积索赔额建立了奖惩系统的转移规则，将之与集成学习结合可以更合理地开发新的保险产品。

关键词

Bagging，Boosting，Stacking，索赔额预测，奖惩系统

Prediction and Pricing of Cumulative Claims Based on Two-Layer Stacking Model

Jingshuo Si

School of Science, Hebei University of Technology, Tianjin

Received: Apr. 25th, 2022; accepted: May 19th, 2022; published: May 26th, 2022

Abstract

In traditional claim amount prediction, generalized linear model (GLM) is a commonly used method. Recently, machine learning algorithms have also achieved good results in the field of it, providing a new choice for prediction. In the era of big data, how to make predictions more accurately is an urgent problem to be solved. To solve this problem, a two-layer Stacking model, two other

文章引用: 司晶硕. 基于两层 Stacking 模型的累积索赔额预测及定价研究[J]. 应用数学进展, 2022, 11(5): 2858-2867.
DOI: 10.12677/aam.2022.115303

integrated learning algorithms and a generalized linear model were used to predict the cumulative claim amount. By comparing the root mean square error and squared absolute error of each algorithm, it can be found that the accuracy of ensemble algorithms including Stacking are better than that of traditional generalized linear model. Finally, the paper established the transfer rules of the reward and punishment system based on the cumulative claim amount, which can be combined with the two-layer Stacking model to develop new insurance products more reasonably.

Keywords

Bagging, Boosting, Stacking, Claim Amount Forecast, Bonus-Malus System

1. 引言

随着我国经济的发展，私家车的数量越来越多，汽车保险在产险公司的比重也变得越来越大。关于车险定价方法的研究始终处于重要位置[1]。在传统的车险费率厘定中，广义线性模型(GLM)为主流方法。该方法最早由 Nelder 和 Wedderburn 提出[2]，1989 年被 McCullagh 和 Nelder 引入到精算学领域，促进了非寿险精算法发展。随着我国一系列车险改革政策的出台，保险公司在车险产品定价上的主动权变得越来越大，车险行业的发展越来越快，行业之间的竞争也越来越激烈。与此同时，在信息时代的浪潮下，越来越多的车险数据集变得更加复杂和高维。这就给统计研究者带来了新的挑战，传统的广义线性模型无法准确地对庞大而复杂的数据进行精准的刻画。所以要寻找更加有效的方法来解决这些问题。

随着机器学习的发展，国内外学者开始将相关算法引入车险损失预测。在国外，Liu 等(2014)利用 AdaBoost 预测索赔强度，经过与广义线性模型、BP 神经网络和 SVM 的对比，发现 AdaBoost 的预测精度最优，方差相对较小[3]。Noll 等人(2018)利用 GLM、回归树、提升法和神经网络对法国某第三方责任保险数据集的索赔频率进行预测，结果表明机器学习算法可以更好地处理特征变量间的交互影响[4]。在国内，张连增等(2018)利用提升算法对回归树及广义线性模型进行改进，得到更为精准的车险索赔频率预测模型[5]。曾宇哲等人(2019)为了更全面地比较广义线性模型与机器学习方法在车险索赔频率预测问题上的效果，利用深度学习、随机森林、支持向量机、XGBoost 等机器学习方法在 7 个车险数据集进行了测试。研究结果显示在所有的数据集上 XGBoost 的预测效果一致地优于广义线性模型[6]。

奖惩系统为根据历史索赔数据对保费进行调整的方法。传统的奖惩系统存在一些不足，如只考虑了索赔次数，转移规则较为简单。Tan (2015)根据投保人目前所处的奖惩等级及历史索赔次数制定了一种动态转移规则，使得奖惩系统更为合理[7]。近年来将索赔额加入到奖惩系统中的相关研究越来越多，如 Gómez-Déniz (2016)给索赔额划定一个固定额，并以该固定额为界限，基于二元分布假设对给定转移规则下的奖惩系数进行计算[8]。国内关于该方法的研究较少，孟生旺(2013)比较了不同分布假设下的最优奖惩系统，发现基于负二项—贝塔分布的奖惩系统具有最优的应用价值[9]。2018 年，孙志强指出我国现有奖惩系统存在仅考虑索赔次数、转移规则简单及惩罚比较温和的问题，利用累积索赔额构建更为合理的奖惩系统[10]。本文利用包含两层 Stacking 模型在内的三种集成算法和广义线性模型对累积索赔额进行了预测，并对各算法的 RMSE 和 MAE 进行比较，本文还利用累积索赔额建立了一种新的奖惩系统转移规

则，将之与 Stacking 模型结合可以更合理地开发新的保险产品。

2. 理论基础

本文的主要工作是预测车险保单的累积索赔额。假设有 N 份汽车保单，每份保单的观测值为 (x_i, a_i)，其中 x_i 为第 i 份保单的解释变量，假设其为 p 维向量，a_i 为保单的累积索赔额，大于 0 且连续。

2.1. 建模使用方法

2.1.1. 广义线性模型

广义线性模型是对线性回归模型的进一步推广，广义线性模型因变量的分布为指数分布族，在经过一个函数变换后，拟合值可被表示为参数的线性组合。

本文选取伽马回归为广义线性模型的代表，对累积索赔额进行预测。伽马回归的因变量 a 服从伽马分布[11]，即：

$$\text{Gamma}(a \mid \alpha, \pi) = \frac{\pi^{\alpha} a^{\alpha-1} \mathrm{e}^{-\pi a}}{\Gamma(\alpha)},$$

其中参数 α 为形态参数，决定分布曲线的形状，参数 π 为尺度参数，决定分布曲线的陡度。

而伽马回归的定义如下：

$$\begin{cases} p(a_i; u_i, \phi) = \exp\left(\dfrac{-a_i/u_i - \ln u_i}{\phi} + \dfrac{1-\phi}{\phi} \ln a_i - \dfrac{\ln \phi}{\phi} - \ln \Gamma\left(\dfrac{1}{\phi}\right) \right) \\ g(u_i) = x_i^{\mathrm{T}} \beta \end{cases}$$

其中 u_i 为 a_i 的均值，ϕ 为离散参数，与分布的方差有关。

2.1.2. 随机森林

随机森林(RF)是在 Bagging 的基础上实现的[12]。Bagging (bootstrap aggregating)是由 Breiman 提出的一种并行集成算法[13]。该算法通过对数据进行多次有放回的抽样得到新样本，将每个样本经过弱学习器训练后的结果整合，生成一个强学习器。

随机森林先利用 Bootstrap 抽样从原始数据集中有放回地抽取多个不同的数据集。不同的是，RF 构建的树是“不相关的”。即 RF 在建立决策树时所用的特征是从所有的特征中随机选取的。本文以该算法作为 Bagging 的代表，其具体步骤见下：

1) 对包含 n 个样本的训练集 T 进行 Bootstrap 抽样，得到 B 个样本容量为 n 的训练样本集，用于构建决策树；

2) 在树的每个节点，从所有 p 个随机变量中选择 m ($m < p$)个随机变量，然后从中选择最优分裂变量。重复以上操作，直到节点的样本大小达到指定的最小限制；

对于所研究的回归问题，最终的预测结果为所有决策树的预测结果取均值。

2.1.3. GBDT

Boosting 是串行算法的一种，它首先利用基学习器训练初始训练集，然后根据其性能对样本分布进行调整，使有误差的样本在后续得到更多的关注[14]。然后再利用基学习器对调整后的样本进行训练，重复以上操作，直到基学习器的数量达到预定值。常见的 Boosting 包括 AdaBoost 和 GB，本文选取 GBDT 作为 Boosting 的代表。

GBDT 是梯度提升方法与决策树的结合[15]。在 GBDT 的迭代中，有以下假设：

1) 在 $j-1$ 轮迭代得到的强学习器是 $f_{j-1}(x)$，损失函数是 $L\left(y, f_{j-1}(x)\right)$；

2) 第 j 轮迭代的目标是找到一个弱学习器 $h_j(x)$，使得本来的损失 $L\left(y, f_j(x)\right)=L\left(y, f_{j-1}(x)+h_j(x)\right)$ 最小。其中，利用损失函数的负梯度来拟合本轮损失的近似值。

对于本文研究问题，我们选取的损失函数为平方损失：

$$L\left(y, f(x)\right)=\left(y-f(x)\right)^2,$$

在这种情况下，第 b 棵树损失函数的负梯度为：

$$r_{b,i}=-\left|\frac{\partial L\left(y_i, f(x_i)\right)}{\partial f(x_i)}\right|_{f(x)=f_{b-1}(x)}=y-f(x_i)$$

各叶子节点的最佳负梯度拟合值为：

$$c_{b,j}=\frac{1}{K}\sum_{x_i\in R_{b,j}}\left(y_i-f(x_i)\right)$$

K 表示第 b 棵树的第 j 个节点中的样本数量。

2.2. 两层 Stacking 模型

Stacking (Stacked Generalization)是一种对异质学习器进行集成的分层模型[16]。它通常由两层组成，一般把在第 0 层的学习器称为初级学习器，在第 1 层的学习器称为元学习器。对于该模型，我们首先按照以下原则训练初级学习器：

1) 将训练集 T 划分为 K 部分 $T_1, T_2, \cdots, T_K$；

2) 在第 k 次训练时，取出 $T_k\ (1<k<K)$，用各基学习器训练 $T^{-k}=T-T_k$；

3) 迭代 K 次，利用 K 次的结果得到与训练集行数相同的预测结果。

在完成上述步骤后，将初级学习器的预测结果作为新的特征输入元分类器，可以得到最终结果。

在初级学习器的选取上，本文遵循减小学习器间相关性和增强可比性的原则，选择了随机森林和 GBDT。因为在上一层已经存在复杂的非线性转换，所以在选择元学习器时可以选择较为简单的广义线性模型，还可以避免过拟合的发生。

综上可以得到 Stacking 优于其他模型的原因，一是在训练初级学习器时采用了交叉验证的思想，从而可以充分利用数据，增强算法的鲁棒性；二是集成不同的学习器，使模型泛化能力得到提高。

2.3. 评价指标

本文选取均方根误差(RMSE)及平方绝对误差(MAE)为模型的评价指标。均方根误差实际上描述了数据的离散程度，它可以解决数据中量纲不一致的问题，从而更好地对数据进行感知。平方绝对误差可以更好地反映预测值误差的实际情况。这两个指标常用来作为衡量机器学习模型预测结果的标准。一般情况下，RMSE 及 MAE 的值越小，代表模型的预测精度越高。

在样本容量为 n 的数据集 T 上，假设 $a_i\ (i=1,2,\cdots,n)$ 为真实的累积索赔额，$\hat{a}_i$ 为学习器 h 预测出的累积索赔额。RMSE 表示 $a_i\ (i=1,2,\cdots,n)$ 与 $\hat{a}_i$ 之间的关系，可以用来衡量 $\hat{a}_i$ 和 $a_i\ (i=1,2,\cdots,n)$ 的偏差，能够将预测的精密度很好地反映出来。

RMSE 的计算公式为：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(a_i-\hat{a}_i\right)^2}.$$

MAE 的计算公式为：

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{a}_i-a_i\right|.$$

3. 数据介绍

3.1. 数据集描述

本文数据来源于法国某保险公司。数据集中涵盖了许多与所研究问题相关的特征，与人的相关的因子包括驾驶执照年龄、性别、婚姻状况等，与被保险车辆相关的因素包括车的品牌、车的用途及引擎类型等。对变量的具体描述见表 1。

Table 1. Feature variable information table for the dataset
表 1. 数据集的特征变量信息表

变量名称	变量描述	变量类型
LicAge	驾驶执照年龄，以月为单位	数值型变量
VehAge	被保险车的车龄，分为 9 级	分类型变量
Gender	驾驶人性别	分类型变量
MariStat	婚姻状况，有 2 类	分类型变量
VehUsage	被保险车的用途，分为 4 类	分类型变量
DrivAge	驾驶人年龄，介于 15 岁~100 岁	数值型变量
BonusMalus	奖励或惩罚，介于 50~350 之间	数值型变量
VehBody	被保险车的类别，有 9 类	分类型变量
VehEngine	车载引擎类型，有 6 类	分类型变量
VehEnergy	被保险车的燃料类型，共 4 类	分类型变量
VehMaxSpeed	被保险车的最大速度，共 10 类	分类型变量
VehClass	被保险车的品牌，共 6 类	分类型变量

本文对数据集的特征进行进一步的分析，得到数值型变量及分类型变量的描述性统计分析，分别如表 2 和表 3 所示。

Table 2. Descriptive statistics for numerical variables
表 2. 数值型变量的描述性统计分析

变量名称	最小值	中位数	最大值	均值
LicAge	0.00	282.00	940.00	301.00
DrivAge	18.00	45.00	97.00	46.25
BonusMalus	50.00	54.00	272.00	64.27

Table 3. Descriptive statistics for categorical variables
表 3. 分类型变量的描述性统计分析

变量名称	每一类别的名称/样本量					
VehAge	0	1	2	3	4	5
	4722	4723	4902	3856	3348	2766
	6~7	8~9	10+			
	2926	1794	1558			
Gender	女性	男性				
	11,570	19,025				
MariStat	单身	其他				
	7424	23,171				
VehUsage	私人用车	私人办公车	专用车	运输用车		
	9956	13,522	6523	594		
VehBody	公共汽车	敞篷车	跑车	微型厢式货车	厢式轿车	SUV
	159	1343	1328	1374	20,140	1858
	旅行车	面包车	其他			
	1629	1085	1679			
VehEngine	油器式	直喷式	导电式	GPL	喷射式	其他
	516	7037	6	2	20,821	2213
VehEnergy	柴油	电	GPL	其他		
	9438	6	2	21,149		
VehMaxSpeed	1~130	130~140	140~150	150~160	160~170	170~180
	215	1081	1291	3863	5297	4830
	180~190	190~200	200~220	220+		
	4677	3672	3331	2338		
VehClass	0	A	B	H	M1	M2
	759	2991	9567	4894	7745	4639

因变量的分布频率直方图如图 1。

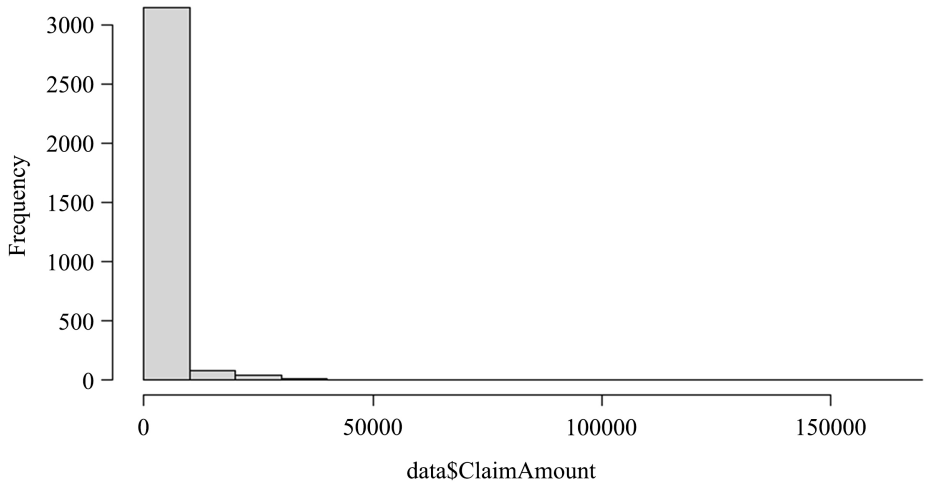

Figure 1. The distribution frequency of the dependent variable
图 1. 因变量的分布频率

3.2. 数据预处理

为了消除量纲不一致带来的不利影响，本文对数值型变量进行了离差标准化，将结果映射到 0-1。离差标准化的公式为：

$$\hat{x}_{i,m} = \frac{x_{i,m} - \min(x)}{\max(x) - \min(x)}$$

$x_{i,m}$ 为原始数据集中第 i 个样本第 m 维的数值，$\hat{x}_{i,m}$ 为标准化后的数值，$\min(x)$ 为原始数据集中的最小值，$\max(x)$ 对应原始数据集中的最大值。

对于因变量，本文对其进行了单位化处理：

$$\hat{a}_i = \frac{a_i}{e_i}$$

其中 $i = 1,2,\cdots,n$，e_i 为第 i 份保单所对应的暴露数，即保单存续期，$\hat{a}_i$ 为剔除 e_i 后的模型因变量。

4. 实证结果分析

本文将数据集按照五种不同的比例进行训练集及测试集的划分，在每种比例下，利用四种算法建模得到的 RMSE (以万为单位)如表 4，利用四种算法建模得到的 MAE (以万为单位)如表 5。其中 GR 表示伽马回归，RF 表示随机森林，Sta 表示两层 Stacking 模型。

Table 4. RMSE of four algorithms under different partition ratios
表 4. 不同划分比例下四种算法的 RMSE

	5:5	6:4	7:3	8:2	9:1
GR	0.656	0.690	0.610	0.430	0.349
RF	0.550	0.529	0.527	0.295	0.257
GBDT	0.656	0.692	0.607	0.426	0.336
Sta	0.605	0.566	0.523	0.288	0.249

Table 5. MAE of four algorithms under different partition ratios
表 5. 不同划分比例下四种算法的 MAE

	5:5	6:4	7:3	8:2	9:1
GR	0.240	0.239	0.233	0.221	0.211
RF	0.211	0.200	0.190	0.168	0.162
GBDT	0.235	0.236	0.230	0.219	0.207
Sta	0.218	0.212	0.185	0.164	0.146

对比各算法在不同划分比例下的表现，可以看出：

1) 在任一划分比例下，无论是以 RMSE 还是 MAE 为指标，三种集成学习算法的效果均优于传统广义线性模型。

2) 在训练集与测试集的划分比例为 9:1 时，各模型的算法精度最高。随着训练集的比重增加，模型的预测效果变得越来越好。

3) 在划分比例为 8:2 和 9:1 时，在以 RMSE 及 MAE 为评价指标时，所构建的两层 Stacking 模型效果最优。

5. 定价研究

在进行车险的产品定价时，需要考虑奖惩系统。奖惩系统由费率等级、转移规则和奖惩系数三个要素组成。传统的转移规则仅考虑了索赔次数，存在一些不足，如可能会使消费者感到不公平，难以形成良性的激励，还可能使保险公司的盈利降低。针对此问题，本文利用累积索赔额制定了一种新的转移规则，并与传统只考虑索赔次数的转移规则做了比较。

假设一个奖惩系统中的费率等级从 1 开始，最高可到 S 等，当费率等级越高时，惩罚力度越大。新投保的车龄进入一个处于中间水平的等级，即初始费率等级。现有文献通过索赔次数确定的奖惩系统的转移规则如下：如果被保险车辆未发生索赔，则在续保时，费率等级下降一等，直到最低等级。在被保险车辆发生索赔时，每发生一次索赔，费率等级就上升两级，直到最高等级[17]。

Table 6. A comparison of the two rules
表 6. 两种规则的比较

初始费率等级	索赔次数	累积索赔额	原规则下的等级	新规则下的等级
1	1	3295	3	2
1	1	57,037	3	7
1	2	975	5	2
1	2	53,477	5	7
1	3	1245	7	3
1	3	35,012	7	6

本文制定的转移规则如下：假设被保险车辆发生 z 次索赔，当前所处的费率等级为 s，$s=1,2,\cdots,S$，累积索赔额为 A，当 $A=0$ 时，表示被保险车辆未在保单年发生索赔，续保时费率等级会下降一级，即下一保单年的等级 $s'=s-1$。当 $A>0$ 时，假设该保单组合内的平均累积索赔额为 $\overline{A}$，则续保时，费率等级按照如下公式进行计算：

$$s'=s+\left[\frac{A}{\overline{A}}+\frac{z}{2}\right]$$

直到 s' 达到最高等级，$[\,]$ 表示四舍五入。

我国现行车险奖惩系统包括 8 个等级，假设某车险保单的平均累积索赔额为 10,476.65，并可以得到各保单的索赔次数及累积索赔额，则可以利用保单数据对两个转移规则进行比较如表 6。

从表 6 可看出，本文制定的转移规则对累积索赔额较高的车主惩罚力度较大，可增加保险公司的盈利，同时也可以使消费者感到更加公平。

在不断进行车险改革的大数据时代背景下，保险公司可先利用两层 Stacking 模型将累积索赔额预测出来，再利用该转移规则制定更为合理的奖惩系统。

6. 小结

在车险领域，随着行业大数据的积累，车险数据集的样本容量变得更加庞大，特征变得更加高维，传统的广义线性模型无法对车险数据进行充分精准地刻画，本文利用集成学习对车险的累积索赔额进行预测，发现包含所提出模型在内的三种集成方式均优于传统的广义线性模型。本文还利用累积索赔额制定了一种新的转移规则，使奖惩系统变得更为合理。在利用 Stacking 对累积索赔额进行预测后，再利用本文所制定的转移规则，可以帮助保险公司研发更适应时代发展的车险产品。但本文还存在一些不足，如数据集中的特征变量依旧相对较少，未来可以寻找特征更为丰富的车险数据集；现今还衍生了一些新的集成学习算法，可以将其加入到研究中。

参考文献

[1] 张连增, 王缔. 保险大数据条件下车险费率厘定的研究——基于 SOM 神经网络方法的车险索赔强度建模[J]. 保险研究, 2018(9): 56-65.

[2] McCullagh, P. (1989) Generalized Liner Models. Routledge, London. https://doi.org/10.1007/978-1-4899-3242-6

[3] Liu, Y., Wang, B.J. and Lv, S.G. (2014) Using Multi-Class AdaBoost Tree for Prediction Frequency of Auto Insurance. *Journal of Applied Finance and Banking*, **4**, 45-53.

[4] Noll, A., Salzmann, R. and Wuthrich, M.V. (2018) Case Study: French Motor Third-Party Liability Claims. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3164764

[5] 张连增, 申晴. 提升算法对传统车险索赔频率建模模型的改进——基于我国五省交强险保单数据[J]. 保险研究, 2019(7): 67-78.

[6] 曾宇哲, 吴媛博, 郑宏远, 等. 基于机器学习的车险索赔频率预测[J]. 统计与信息论坛, 2019, 34(5): 69-78.

[7] Tan, C.I., Li, J., Li, J.S.H., *et al*. (2015) Optimal Relativities and Transition Rules of a Bonus-Malus System. *Insurance*: *Mathematics and Economics*, **61**, 255-263. https://doi.org/10.1016/j.insmatheco.2015.02.001

[8] Gomez-Deniz, E., Hernandez-Bastida, A. and Fernandez-Sanchez, M.P. (2014) Computing Credibility Bonus-Malus Premiums Using the Total Claim Amount Distribution. *Hacettepe Journal of Mathematics and Statistics*, **43**, 1047-1061.

[9] 孟生旺. 考虑个体保单风险特征的最优奖惩系统[J]. 数理统计与管理, 2013, 32(3): 505-510.

[10] 孙志强. 我国现行汽车保险奖惩系统研究[D]: [硕士学位论文]. 郑州: 郑州大学, 2018: 24-31.

[11] 孟生旺. 回归模型[M]. 北京: 中国人民大学出版社, 2015: 30-32.

[12] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[13] Breiman, L. (1996) Bagging Predictors. *Machine Learning*, **24**, 123-140. https://doi.org/10.1007/BF00058655

[14] Schapire, R.E. (1990) The Strength of Weak Learn Ability. *Machine Learning*, **5**, 197-227. https://doi.org/10.1007/BF00116037

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

[16] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 97-119.

[17] Arthur, C. (2014) Computational Actuarial Science with R. CRC Press, Boca Raton.