文章编号：2095-2163(2024)03-0116-07 中图分类号：TP391 文献标志码：A

基于 Stacking 集成学习的高校录取分数线预测

干 霞，魏嘉银，卢友军，秦信芳，来小孟
（贵州民族大学 数据科学与信息工程学院，贵阳 550025）

摘 要：针对如何准确预测高校录取分数线，帮助高考生做出更加准确的志愿填报决策问题，提出一种基于 Stacking 集成思想的双层模型。该模型采用机器学习算法暴露特征重要性，融合 3 个单一算法并使用交叉检验法和网格搜索法进行参数优化。通过在贵州省 2018-2022 五年高考高校录取数据上进行实验结果表明，该双层融合模型的预测效果优于支持向量回归、决策树、随机森林等单一模型和其他集成模型；预测误差在 5 分以内的精度超过 95%，平均绝对值误差低于 2.43；较单一模型中表现最好的梯度提升指标分别提升 44%和 19%，提升了预测效果，为未来分数线预测提供了新的方向。

关键词：集成学习；Stacking；交叉检验法；网格搜索法；高考分数线

Prediction of college admission scores based on Stacking ensemble learning

GAN Xia, WEI Jiayin, LU Youjun, QIN Xinfang, LAI Xiaomeng
(School of Data Science and Information Engineering, Guizhou Minzu University, Guiyang 550025, China)

Abstract: Aiming at how to accurately predict the college admission score line and help college entrance examination students make more accurate voluntary filling decisions, a two-layer model based on the idea of Stacking ensemble is proposed. The model uses machine learning algorithms to expose the importance of features, fuses three single algorithms, and uses cross-checking and grid search methods for parameter optimization. Through experiments on the admission data of colleges and universities in Guizhou Province in 2018—2022, the experimental results show that the prediction effect of the two-layer fusion model is better than that of single models such as Support Vector Regression, Decision Tree, Random Forest and other ensemble models. The accuracy of the prediction error within 5 points exceeds 95%, and the average absolute error is less than 2.43, which is 44% and 19% higher than the best performing gradient improvement index in a single model, respectively. The research improves the prediction effect, and provides the new direction for future score line prediction.

Key words: ensemble learning; Stacking; cross-validation method; grid search method; college admission scores

0 引 言

高考是中国学生升学的重要门槛，根据考生的高考分数和高校往年的录取分数线来选择和填报志愿是非常重要的环节。而高考的不断改革和高校招生政策的不时调整，高校录取分数线变得越来越难以预测，给考生填报志愿带来了很大的困扰。简单的统计估计存在低估或高估的风险，若低估分数线可能导致录取不上，若高估分数线则会导致与理想高校失之交臂。因此，如何准确预测高校录取分数线对于考生填报志愿至关重要。

已有学者运用不同的方法对高考数据进行挖掘分析，并为考生填报志愿提供有参考价值的信息或建议。早期的研究利用数理统计方法对院校最低分数线和最低位次进行分数线预测[1-2]，此类方法简单，仅利用对分数的分析来预测分数线，而不考虑其他因素是远远不够的。鉴于单一模型的预测误差较大，周帆[3]利用变权组合预测法，对不同的线性模型添加不同系数以提高模型的预测精度。实验结果表明，加权后的模型预测精度高于单一模型，这是早

基金项目：贵州省省级科技计划项目资助（黔科合基础[2018]1082 号、[2019]1159 号）；贵州省教育厅自然科学研究项目（黔教技[2022]015 号、黔教技[2022]047 号）。

作者简介：干 霞（1997-），女，硕士研究生，主要研究方向：海量数据统计与分析，Email：2280225211@qq.com；卢友军（1985-），男，博士，副教授，硕士生导师，主要研究方向：大数据分析与处理、网络传播动力学研究。

通讯作者：魏嘉银（1986-），男，博士，副教授，硕士生导师，主要研究方向：大数据分析与处理、推荐算法设计与分析。Email：weijiayin05@sina.com

收稿日期：2023-11-04

期利用简单加权组合不同的预测模型，进而提高预测精度的研究。

随着科学技术的快速发展，机器学习和深度学习被广泛运用于高考数据的分析研究。王振如[4]使用自适应增强集成算法（Adaptive boosting，Adaboost）预测高考批次线，并运用深度学习来预测专业分数线，通过对神经网络的设计，使得模型对高考数据有良好的表现，预测差值最小在 5 分左右。支持向量机回归（Support Vector Machine Regression，SVR）因具有较强的泛化能力，被广泛运用于交通流预测[5]。一些学者将其引入分数线预测领域，组合其他机器学习算法构建分数线预测模型。如：王英英[6]训练 4 个线性模型作为第一层学习器，并将预测结果传入第二层 SVR 和多层感知机（Multilayer Perceptron，MLP）。实验证明，经过组合后的模型更稳定、鲁棒性更好，SVR 组合模型的平均绝对值误差为 6.41，对比 K 近邻（K-Nearest Neighbor，KNN）组合模型的平均绝对值误差降低了 0.34。何晶晶[7]获取学校、年份、招生计划、批次线、最低投档线等 7 个特征，以灰色 $GM(1,1)$ 算法和多元线性回归的预测值为输入，训练 SVR 作为最终预测器构建组合模型。通过反复实验，其预测值与真实值的平均相对误差为 4.27%。任建涛[8]仅利用往年的录取平均分数线一个特征，使用 SVR 回归对院校专业线进行预测，然而不管是专业录取分数线、还是高校录取分数线都受多种因素影响，对单一因素分析并不能达到最好的预测效果。在已有的研究中，吴凯[9]、Zhang 等学者[10]、胡如明[11]选取了年份、科类、批次、省控线、平均分、最高分、最低位次、省控线排名等 16 种因素，使用神经网络算法对分数线进行研究。神经网络中有大量的网络节点可以快速高效地处理多个维度数据中的非线性关系，上述分数线的预测较以往的传统预测模型均取得了一定的提升。郭孝文等学者[12-13]改进反向传播神经网络（Back Propagation，BP）使其符合分数线变化规律，并将其运用到西安工业大学录取分数线的预测上，预测误差在 5～10 分内，预测精度较传统线性模型提高 20%。

通过对已有研究分析可知，单一机器算法和简单的加权模型在分数线预测上的性能表现不够理想，已不能满足分数线预测的精度要求。

本文基于机器学习算法和集成学习思想，旨在通过优化预测模型的设计和算法选择，提高对高校录取分数线的准确预测。本文研究包括以下几点：对数据进行清洗整理和特征工程处理，提取出最具预测能力的特征；优化模型的参数，设计基础预测器和元学习器；将单一算法的预测结果进行融合，构建双层模型，进一步提高预测精度和准确性，从而为高考生填报志愿提供更为科学、准确的参考依据。

1　模型设计

1.1　研究思路

构建高校录取分数线预测模型时，需要考虑到数据特征的多样性和模型回归原理的要求，以开发适合高校录取分数线预测的组合方法。本文首先从教育服务网站上收集高校往年录取数据并进行相应的数据预处理；然后，基于文献研究和算法原理确定预测器；运用多个机器学习方法在数据集上进行实验以观察不同算法的预测效果；在同样的基础预测器上比较本文元学习器和其他元学习器的预测效果，以证明本文模型的回归性能优于其他研究算法。总体的预测路线如图 1 所示。

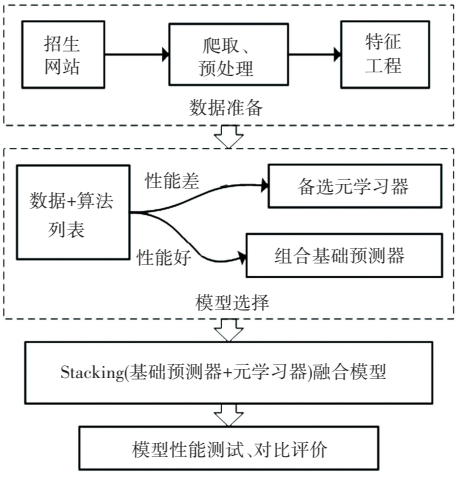

图 1　高校录取分数线预测路线图

Fig. 1　College admission score prediction line map

由图 1 可知，总体分为 3 个阶段：

（1）数据准备阶段。主要完成数据采集、数据预处理和提取特征；

（2）模型选择阶段。使用多种预测器在数据集上进行训练，得到预测器的性能差异；

（3）模型集成与性能测试阶段。使用 Stacking 集成技术融合预测器，进行融合模型的性能测试。

1.2　模型框架构建

本文是在机器学习的基础上，利用 Stacking 学习策略构建高校录取分数线预测模型。Stacking 是一种把初级预测器的预测结果作为第二层学习器的输入的方法，称为学习法[14]。在 Stacking 中，基础

学习器的质量和多样性非常重要,直接影响到最终集成模型的性能表现。不同的学习器可使用多折交叉检验拆分训练集,在训练数据上进行训练并使用多个预测器来做预测,得到多个预测结果[15]。把初级预测器的预测结果当作次级学习器(又称元学习器)的输入,相当于特征转换了一次,经过次级学习器的训练学习,输出最终预测结果。如果某个初级学习器错误地学习了特征空间的某个区域,那么次级学习器通过结合其他初级学习器的学习行为,可以适当地纠正这种错误。通过多层预测器的组合学习来解读数据中包含的信息,最终的组合模型不仅预测精度高,而且泛化性能很好。

为了提高 Stacking 融合模型的预测能力,需要确保基础预测器性能出色且具有差异,同时保证元学习器简单易懂,能够充分学习基础预测器的优点,又避免模型出现过度学习。相比于传统的树回归器,XGBoost 在很多细节上进行了优化,例如采用加权迭代方法、交叉验证选择最优步长等,Bentéjac 等学者[16]对 XGBoost、随机森林和梯度增强等集成算法做了全面的比较,证明 XGBoost 在效率和准确率两个维度都有较大的提升。而 HistGradientBoosting 使用更加高效的直方图算法加速训练,对于数据样本大于 10 000 的数据集,此估计器要比其他回归器快得多,而且对缺失值友好,可以提高回归模型的抗噪声以及局部扰动的能力,已被证明在广泛的机器学习问题上有不错的表现。经过对上述文献研究和算法回归性能的考虑,以及本文数据量和特征情况,构建了 XGBoost 和 HistGradientBoosting 作为基础预测器,并调用网格搜索,优化模型参数。通过实验,验证了 XGBoost 和 HistGradientBoosting 作为基础预测器时集成模型的良好性能。

为避免过拟合,本文使用五折交叉检验拆分训练集,并使用具有变量筛选和复杂度调整的 Lasso 回归作为元学习器,确保元学习器充分学习基础预测器的同时不会发生过拟合。最后,通过 Stacking 学习将 3 个预测器融合,用于各个高校的录取分数线回归模型,模型框架如图 2 所示。

预测高校录取分数线的集成模型实现步骤如下:

(1)将训练集进行五折划分,其中 4 份用作训练集,1 份作为测试集;

(2)用训练集平行训练第一层学习器 XGBoost 和 HistGradientBoosting,分别用 5 份测试集进行测试,每个模型得到 5 份预测值;

(3)将每个模型的 5 份预测值进行平均,得到每个模型的预测值;

(4)把第一层学习器的预测值作为元学习器 Lasso 回归的输入训练元学习器,得到融合模型。

测试时,将测试集的特征输入到每个基础模型中,并将其预测结果作为元模型的输入,经过元模型的学习,得到最终的预测结果。

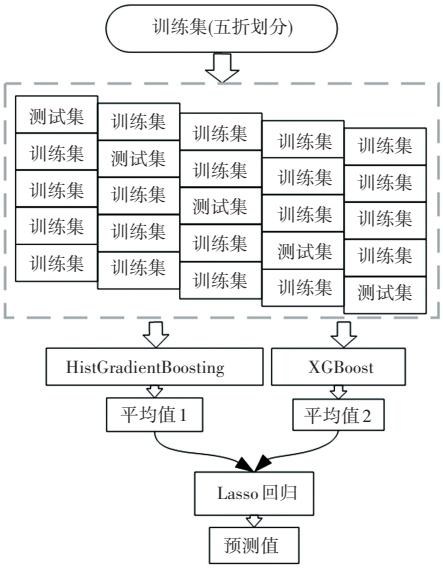

图 2 高校录取分数线预测模型框架

Fig. 2 College admission score prediction model framework

2 实验结果分析

2.1 实验数据及预处理

本文的数据为 2018-2022 五年间全国各高校在贵州省理科一批、理科二批、文科一批、文科二批的招生数据。使用数据挖掘技术从高考网、掌上高考和贵州省招生考试院官网等高考服务网站爬取收集而来。数据清洗内容主要包括去除因在多个网站中收集导致的重复值、异常值、某些年份的数据为图片格式或 pdf 格式,使用识别软件将其识别出来,存在识别不够准确,如误将数字 3 识别为数字 8、数字 0 识别为字母 c 等问题,只能通过对比原数据来排查错误数据。数据清洗内容还包括缺失值处理,某些高校在某地区的招生不是固定的,如川北医学院在贵州省 2021 年没有招生,其他年份正常招生,则该缺失值采用前 3 年的平均值填补。经过数据清洗后,共得到 11 078 条数据,其中 8 825 条(2018-2021 年)用作训练集,2 253 条(2022 年)作为测试集,用以验证模型性能。

原始数据包括学校名称、学校代码、专业类别、

计划数、投档比例、最高分、最低分、最低位次共 8 个特征。这里给出阐释分述如下。

(1)学校代码和学校名称具有相同的意义，学校代码是招生地区给各高校的数字表示，方便高考生查询招生信息，且学校代码是数字型数据。

(2)专业类别主要分为普通类、民族班、定向西藏、中外合作办学等类别，是对该招生专业的类别说明。普通类占据 90%以上的比例，其他各种类别占极少数，所以此特征不作为本次分析对象。

(3)计划数是高校在该地区的招生计划，每年的招生计划可能不同。

(4)投档比例是招生计划数和实际投档人数的比值，大多数高校会按照招生计划数招满即停，投档比例 100%，只有极少数高校招不满或者多招。因此此特征不具有分析必要性。

(5)最高分表示高校在该地区录取分数最高的考生分数。当考生分数超过最低分数时表示考生可以报考该校，但是被录取的概率通常会参考录取平均分。由于考生报考的不合理性，导致最高分高对平均分的影响较大，进而影响考生报考。所以很有必要将最高分作为分析特征。

(6)最低位次是指录取结束后，将录取最低分数所对应的高考排名作为录取最低位次。考生填志愿时会对应自己的高考排名，把高校录取最低位次作为录取概率的重要参考因素，所以直接将最低位次作为分析特征。

(7)本文的研究对象为高校录取分数线，故把最低分作为目标值。

通过对原数据的处理，得到可用于分析的特征有学校代码、计划数、最高分、最低位次、最低分。根据特征对目标值的贡献率不同，采用随机森林算法暴露特征重要性，其结果见表 1。由表 1 可知，学校代码和计划数对目标值的重要程度极低，所以本文不考虑将此特征作为分析对象。接着进行特征补充，补充省控线、省控线排名两个特征，通过特征工程得到省控线与录取分数线的差值、简称线差，省控线排名与录取分数线排名的差值、简称排名差。加入线差和排名差继续分析其特征贡献，得到结果见表 2。

表 1　原始特征对最低分的重要程度

Table 1　Importantce degree of the original feature to the lowest score

特征	学校代码	计划数	最高分	最低位次
贡献值	0.001 8	0.003 1	0.020 6	0.974 3

表 2　补充特征对最低分的重要程度

Table 2　Importance degree of the other features to the lowest score

特征	省控线	最高分	最低位次	线差	排名差
贡献值	0.860	0.960	−0.940	0.520	0.078

高校录取分数线受多种因素影响，但不能将所有数据特征都用于建模，过多的特征可能会带来信息冗余，影响模型精度。为了选择合适的特征，本文经过特征工程等特征处理技术，对特征进行加减转换，从原始数据中分离出对目标变量有影响的特征，根据相关性系数大小进行筛选，并确保这些特征在不同的基学习器中表现都不一样，以便让 Stacking 能够从各特征差异中受益。由表 2 可知，排名差与最低分相关性的绝对值未超过 0.5，其他特征与最低分都呈强相关性，说明其对目标值很重要。结合本文的预测目标是最低分，因此选择相关性强的 4 个特征作为输入变量，分别是：省控线、最高分、最低位次、线差。

2.2　评价指标

为了评估模型预测性能，本文评价指标采用平均绝对值误差（*MAE*），数学定义公式如下：

$$MAE = \frac{1}{m}\sum_{i=1}^{m} | (y_i - \hat{y}_i) | \tag{1}$$

其中，$y_i - \hat{y}_i$ 为测试集上真实值-预测值。当平均绝对值误差的值越小时，说明模型学习的数据信息越多，预测误差越小。

模型预测精度采用 $Accuracy(Error < 5)$ 为预测分数值和真实值的残差绝对值小于 5 分时的预测精度，$Accuracy(Error < 1)$ 为预测分数值和真实值的残差绝对值小于 1 分时的预测精度。由此推得：

$$Accuracy(Error < 5) = \frac{num(| y_i - \hat{y}_i | < 5)}{num(y_i)} \tag{2}$$

$$Accuracy(Error < 1) = \frac{num(| y_i - \hat{y}_i | < 1)}{num(y_i)} \tag{3}$$

其中，$num(| y_i - \hat{y}_i | < 5)$ 表示预测残差绝对值小于 5 的数量；$num(| y_i - \hat{y}_i | < 1)$ 表示预测残差绝对值小于 1 的数量；$num(y_i)$ 表示测试集的样本数量。当两者数值越大时，说明模型预测精度越高。

2.3　模型参数选择

在回归问题中，参数的设置对于模型的表现至关重要。当数据确定时，模型和参数的选择就决定了预测结果的优劣。模型可以根据数据量、特征数量和以往经验等大致确定。为寻找最佳参数组合以

使模型性能达到最佳,对其重要参数通过网格搜索法进行调整优化以获得最佳值。部分参数设置见表3,未展示的模型和其余参数使用默认参数。

表 3 部分模型的参数

Table 3 Partial model parameters

算法名称	参数设置
Random Forest	*oob_score* = True
	estimators = 50
	Max_depth = 13
	Min_samples_split = 20
HistGradientBoosting	*L2_regularization* = 0.1
	Min_samples_leaf = 7
	Learning_rate = 0.1
GradientBoosting	*n_estimators* = 10 000
	Learning_rate = 1
	Min_samples_split = 5
	Loss = 'ls'
XGBoost	*Learning_rate* = 0.02
	n_estimators = 500
Lasso	*alphas* = [0.5,1.0, 2.0]
	max_iter = 10 000
	tol = 0.001

2.4 元学习器

为了体现模型预测性能的差异,首先建立多种单一模型用于预测分数线,分别是 Lasso 套索、SVR、K 近邻、决策树、随机森林、XGBoost、GradientBoosting、HistGradientBoosting,经过在测试集上验证各个预测模型的性能,得出的预测误差 *MAE* 结果见表4。

表 4 单一模型预测的 *MAE*

Table 4 *MAE* for single model predictions

模型名称	*MAE*
决策树	4.446 515
SVR	11.569 039
K 近邻	8.610 617
Lasso	19.955 478
XGBoost	3.351 931
Random Forest	3.969 035
GradientBoosting	3.015 891
HistGradientBoosting	3.473 456

由表 4 可知,XGBoost、RandomForest、HistGradientBoosting、GradientBoosting 等集成模型的预测结果均优于其他的机器模型。这些集成模型的预测误差均低于 4 分,且较其他机器学习算法表现显著,说明使用集成思想预测高校录取分数线是可行的。

在集成模型中,元学习器的选择很重要。若元学习器是很复杂的算法则容易学习过度,导致过拟合的问题,若太过简单则不能充分学习基础学习器的信息,这样一来预测误差就会太大。基于此,本文把 Lasso 作为元学习器。Lasso 是一种线性回归技术,但其不同于简单线性回归,是通过在损失函数中添加惩罚项来执行正则化,有助于防止过度拟合,增强模型的泛化能力。为了验证 Lasso 作为元学习器的可行性,随机挑选 2 个算法,使用 Stacking 方法将其融合。通过随机选择的基础预测器 XGBoost 和 GradientBoosting,验证 Lasso 元学习器对集成模型性能的影响,实验结果见表5。

表 5 不同元学习器的预测结果

Table 5 Prediction results of different meta-learners

模型名称	*MAE*
Stacking(XG_K 近邻)	4.298 063
Stacking(XG_决策树)	5.220 595
Stacking(XG_Lasso)	**2.436 846**
Stacking(XG_SVR)	3.600 916

注:Stacking(XG_K 近邻)表示基础学习器为 XGBoost 和 GradientBoosting,元学习器为 K 近邻,其余模型同理。

以上实验中,模型单独预测时的 *MAE* 指标大小为:Lasso > SVR > K 近邻 > 决策树,Lasso 的预测误差最大,为 19.95,决策树的预测误差较小,为 4.44;当将其作为元学习器融合到集成模型中时的 *MAE* 指标大小为:Stacking(XG_决策树) > Stacking(XG_K 近邻) > Stacking(XG_SVR) > Stacking(XG_Lasso)。结果表明,单一模型预测误差最大的 Lasso 作为元学习器时的预测效果最好,而决策树作为元学习器时误差反而增大,精度不如其他 3 个基础算法。此实验验证了集成模型中对元学习器的要求,即元学习器不宜复杂,随着元学习器复杂度的上升,集成模型过拟合的风险在增大。本文选择具有复杂度调整因子的 Lasso 作为元学习器,集成模型预测的 *MAE* 为 2.436 846,显著降低了单一模型的预测误差。据此,可以选择 Lasso 套索回归作为元学习器构建预测模型。

2.5 基础预测器

将性能优异的预测器进行差异融合,得到不同的集成模型见表6,用以探究本文模型与其他融合模型各项指标的差异。首先基础预测器在数据集上并行训练、且进行预测,把预测的结果输入元学习器,元学习器接收的特征维度就是基础预测器的个数,通过元学习器的再次学习得到最终预测结果。

实验结果如图 3 所示。

表 6　不同融合模型

Table 6　Different ensemble models

融合模型	基础学习器	元学习器
Stacking_RF_GB	RandomForest、GradientBoosting	Lasso
Stacking_RF_HGB	RandomForest、HistGradientBoosting	Lasso
Stacking_GB_HGB	GradientBoosting、HistGradientBoosting	Lasso
Stacking_XGB_GB	XGBoost、GradientBoosting	Lasso
本文模型	XGBoost、HistGradientBoosting	Lasso

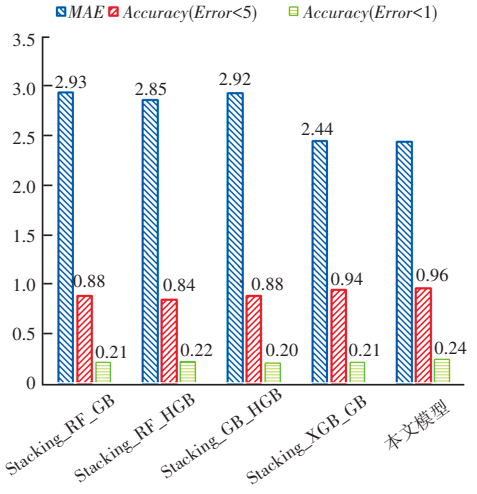

图 3　融合模型的回归性能比较

Fig. 3　Comparison of regression performance of ensemble models

由表 4 和表 6 可知，本文的 5 个集成模型均是由不同的优秀算法融合，保证了基础预测器之间的差异性。单一模型与本文组合模型预测误差比较如图 4 所示。由图 4 可知，本文模型的预测误差比单一模型均有所下降，预测性能优于单一模型。这是因为 Stacking 策略能够有效融合每个基础学习器的优势，元学习器学习预测误差大的样本，可减少模型陷入局部最优点的风险，从而提高预测精度。另外，本文使用的高考数据较为稳定，缺失特征通过均值填充，而且缺失值很少，根据数据特征的具体情况，选择预测性能较为平稳的学习器进行融合，可构建性能更好的模型。

近几年的高考分数线研究已取得了不错的成果，如王振如[4]用单个集成学习模型和神经网络分别预测，预测误差达到了 20 分。吴凯[9]用人工神经网络对第三批次预测的 *MAE* 为 5.82；胡如明[11]用神经网络预测的结果 *MAE* 为 6.75，5 分的百分比为 93.56%。任祥旭[17]用神经网络预测的 *MAE* 为 8.92。相比之下，本文构建的双层模型其误差 *MAE* 为2.433 354，5 分以内的预测精度达到 95%以上，精度最高，模型更稳定。较单一模型中表现最好的 GradientBoosting，其 *MAE* 提升 19%，5 分精度指标提升 44%。本文模型预测值和真实值的拟合情况如图 5 所示，本文模型预测情况随机见表 7。

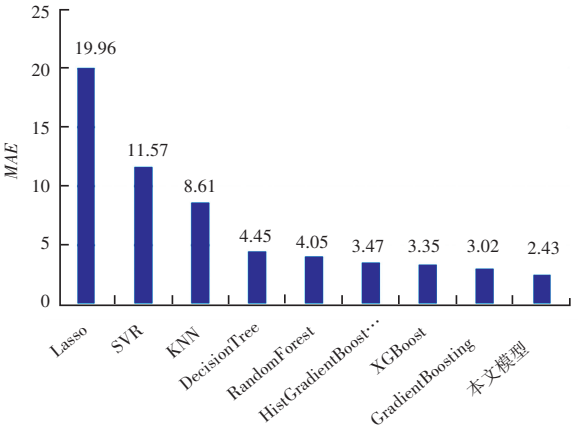

图 4　单一模型与本文组合模型预测误差比较

Fig. 4　Comparison of prediction error between single models and ensemble model in this paper

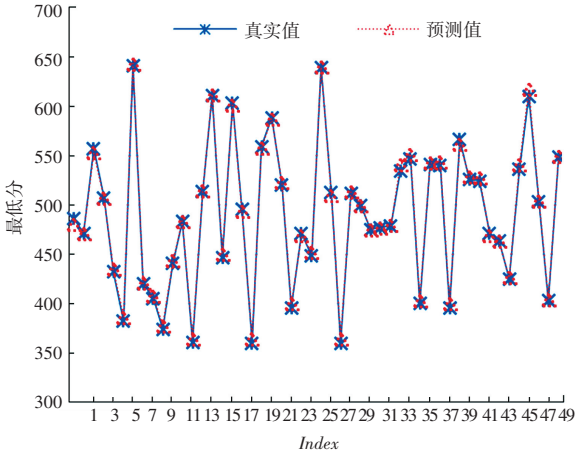

图 5　预测值和真实值的拟合情况(随机)

Fig. 5　Fitting of predicted values and true values(random)

表 7　本文模型预测情况(随机)

Table 7　Forecast of the model in this paper(random)

Index	真实值	预测值	残差
1 348	559	557.442 335	−1.557 665
2 014	475	475.075 214	0.075 214
390	541	542.489 703	1.489 703
1 964	471	469.271 697	−1.728 303
1 834	503	503.954 953	0.954 953
…	…	…	…
740	403	404.531 195	1.531 195
337	513	510.002 346	−2.997 654
2 169	499	499.442 334	0.442 334
1 786	512	513.095 401	1.095 401
1 023	432	435.038 508	3.038 508

3 结束语

本文针对高考数据特征构造 XGBoost、HistGradientBoosting 和 Lasso 三个回归器,网格搜索等技术优化超参数组合,并基于 Stacking 集成思想将其融合成双层模型。第一层基于 Boosting 类的 2 个预测器通过充分学习数据隐藏的信息,经过特征转化生成 2 个维度的预测值作为第二层的输入。为了降低过拟合的风险,构建第二层正则化学习器,学习第一层预测器的预测优点,纠正其因过度学习而导致的错误,进行训练再预测。最后,通过 Stacking 集成 3 个预测器,实现了高校录取分数线预测,为高考生提供更加准确的志愿填报策略。相对传统的单个预测模型以及简单模型组合的预测方法而言,本文构建的 Stacking 融合模型通过集成回归性能优异的算法对其预测结果进行二次学习,一定程度上弥补了单一模型预测误差大的缺陷。

在后续的工作中,考虑加入更多高校因素(如:软科排名、地理位置、建校年数等特征),以提高预测精度。还可以在此基础上对动态排名志愿填报进行研究,构建面向动态排名的实时预测模型,为已经实行动态排名志愿填报的地区提供实时的填报参考。

参考文献

[1] 王亚珊. 基于历史数据的研招信息分析与预测系统的设计与实现[D]. 沈阳:沈阳工业大学,2018.

[2] 刘金伟. 概率方法建模预测分数线[J]. 平原大学学报,2004,21(3):50-51.

[3] 周帆. 变权重组合预测法预测重庆市高考分数线[J]. 科教文汇(上旬刊),2009(9):287-288.

[4] 王振如. 基于机器学习的高考分数线预测系统的研究与实现[D]. 北京:北京邮电大学,2017.

[5] HU Jianming, GAO Pan, YAO Yunfei, et al. Traffic flow forecasting with particle swarm optimization andsupport vector regression [C]//Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC). Qingdao, China:IEEE,2014: 2267-2268.

[6] 王英英. 高考志愿智能填报系统的关键技术研究与实现[D]. 新乡:河南师范大学,2021.

[7] 何晶晶. 基于灰色神经网络的高考批次线预测[D]. 湘潭:湘潭大学,2018.

[8] 任建涛. 推荐算法在高考志愿填报中的应用研究[D]. 昆明:云南财经大学,2018.

[9] 吴凯. 基于人工神经网络的中外合作办学专业高考分数线预测研究[D]. 南昌:江西财经大学,2019.

[10] ZHANG Yue, XIWEI F, XIANGLI Q, et al. Research on the prediction method of the college professional admission scores [C]//International Seminaron Computer Science and Engineering Technology(SCSET). New York,USA:IEEE,2022:406-409.

[11] 胡如明. 基于深度学习的高考分数预测模型与算法研究[D]. 武汉:武汉工程大学,2022.

[12] 郭孝文,梁向阳. 改进的 BP 神经网络在分数线预测中的应用[J]. 西安工业大学学报, 2018, 38(3):286-292.

[13] 王泽卿,季圣鹏,李鑫,等. 基于分数线预测的多特征融合高考志愿推荐算法[J]. 计算机科学,2022,49(S2):254-260.

[14] ZHOU Zhihua. Ensemble methods: Foundations and algorithms [M]. New York: Chapman and Hall/CRC,2012.

[15] KROGH A, VEDELSBY J. Neural network ensembles cross validation, and active learning [C]// Proceedings of the 7th International Conference on Neural Information Processing Systems. Denver Colorado: ACM,1995: 231-238.

[16] BENTÉJAC C, CSÖRGÖ A, MARTINEZ-MUÑOZ G. A comparative analysis of gradient Boosting algorithms[J]. Artificial Intelligence Review, 2021, 54(3): 1937-1967.

[17] 任祥旭. 基于人工神经网络的高考分数线预测研究[D]. 南昌:江西财经大学,2018.