

面向恶劣天气的航班延误 Stacking集成预测模型

孙玥¹, 丁建立²

1. 中国民航大学宁河校区管理中心, 天津 300300;
2. 中国民航大学计算机科学与技术学院, 天津 300300

摘要

天气因素作为影响航班延误的首要因素, 对航班延误预测有重要影响。面向恶劣天气, 对航班延误时长进行多分类预测, 并针对传统单一模型预测精度低、稳定性差等问题, 提出一种基于Stacking的航班延误集成预测模型, 融合航班数据与天气数据特征, 采用LightGBM、XGBoost等多个异质分类器作为基学习器, SVM作为元学习器, 构建堆叠式的双层集成学习框架。为验证模型有效性, 构建多个单一模型与集成模型进行比较。实验结果证明, Stacking集成预测模型性能最优, 总体准确率达到95.25%, F1分数达到0.9527。

关键词

航班延误预测; Stacking集成学习; 多模型融合; 恶劣天气

中图分类号: TP181

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2025012

Flight delay stacking ensemble prediction model for severe weather

SUN Yue¹, DING Jianli²

1. The Management Center of Ninghe Campus, Civil Aviation University of China, Tianjin 300300, China
2. College of Computer Science and Technology, Civil Aviation University of China, Tianjin 300300, China

Abstract

Weather factors, as the primary factors affecting flight delays, have an important impact on flight delay prediction. Confronting the severe weather, multi-classification prediction of flight delay duration was made, and a Stacking-based integrated flight delay prediction model was proposed for the problems of low prediction accuracy and poor stability of traditional single model. Combining flight data and weather data features, multiple heterogeneous classifiers such as LightGBM and XGBoost were used as base learners, and SVM was used as the primary learner. A stacked, two-layer integrated learning framework was constructed. To verify the model validity, multiple single models were constructed for comparison with the integrated model. The experimental results demonstrate that the Stacking integrated prediction model has the best performance with an overall accuracy of 95.25% and an F1 score of 0.9527.

Key words

flight delay predicting, Stacking ensemble learning, multi model fusion, severe weather

0 引言

航空出行以其快速便捷、轻松舒适等优势,成为人们长途旅行的重要出行方式。而航班延误作为航空出行的一大难题,一直难以被彻底解决,不仅影响机场与航空公司的正常运行,还会造成旅客的麻烦与损失。随着科技的快速发展,飞机起飞前的保障工作、安检工作的效率都有了大幅度的提升,机械故障、设备失效的问题也愈发减少,而天气因素由于其客观性,难以被人为干预。因此,结合天气对航班延误进行预测,对进一步提升预测准确率有着重要意义。

关联天气的航班延误预测研究已有一定的研究成果。Zheng 等^[1]分析了对流天气下出发机场与目的机场的延误情况与延误传播的相关影响因素;Rodríguez 等^[2]通过识别所需的天气信息,预测机场的延误情况与吞吐量,为机场的相关决策提供帮助;Zanin 等^[3]基于数据驱动系统研究了中国航空运输延误的演变和原因,证实了雷暴的存在与季节变化是影响航班延误的重要因素;Škultety 等^[4]针对欧洲雷暴次数与强度逐年增加的现象,对夏季恶劣天气下的欧洲航空运输情况进行分析,并使用 Holt-Winters 模型对延误进行中期预测;Zhu 等^[5]结合 METAR 天气报文与 ADS-B 等多个数据源,采用长短期记忆(long short-term memory, LSTM)网络对航班的飞行时间进行预测,并完成了燃料装载方法的优化;Li 等^[6]运用聚类算法评估天气特征与非天气特征对航班延误的影响,并对不同特征的优先级进行了具体分析;Borsky 等^[7]评估了突发性天气事件与慢性天气事件对起飞延误的影响,并对延误时

间进行了小时级别的预测;Chen^[8]等以中国高铁和航空为例,分析了恶劣天气对不同交通系统可靠性的影响;陈可嘉等^[9]设计了两阶段求解算法,减少总航行时长,验证危险天气下航班等待时长与改航路径优化的有效性;王洪等^[10]通过 K-means 聚类算法识别终端区对流天气相似性,将历史运行场景分类,分析不同对流天气对空域通行能力的影响;王春政等^[11]结合天气数据提出了一种基于 Agent 的机场网络延误模型,有效表现了复杂机场网络中组分子系统间的交互特性及空中交通中的延误传播效应;安洋等^[12]融合交通流量、城市区域间的天气、位置关系等多个模态的时空数据,提出一种基于多源异构时空数据融合的交通流量预测模型 MHF-STNet,使用深度学习方法对异构数据统一建模,预测未来时刻的交通流量。

针对航班延误的分类预测大多选取单一的分类器,往往会存在泛化能力不佳的情况,而集成学习通过构建并结合多个学习器来完成学习任务,可以进一步提升模型的预测精度,同时保证模型的稳健性,确保大多数情况下模型能拥有良好的预测性能。相比常见的集成学习方法,Stacking 算法组合策略没有选用简单的投票法或求平均法,而是采用堆叠式的双层学习法,将基分类器的输出作为元分类器新的输入,从而实现模型的融合。近年来,多项研究证明了 Stacking 模型在金融、化学、电力工业、网络安全等众多领域的有效性。Li 等^[13]充分利用了卷积神经网络(convolutional neural network, CNN)算法的特征提取能力和 Stacking 算法的泛化能力,建立更适用于网络贷款平台的风险预测系统;Jiang 等^[14]将历史交易数据和宏观经济信息结合到 Stacking 框架中,预测股票价格指数的变动方向;李泉伦等^[15]

结合近红外光谱分析与Stacking框架的异构集成特性，实现对油页岩含油率的检测；马良玉等^[16]提出一种基于Stacking多模型融合的风电机组齿轮箱异常状态监测方法，以实际风场的齿轮箱历史故障为例，验证模型根据阈值进行故障预警的能力；Khan^[17]以人工神经网络（artificial neural network，ANN）和长短期记忆网络为基础模型，结合深度学习与Stacking集成的优势，对太阳能的光伏发电量进行预测与分析评估；Khraisat等^[18]利用Stacking框架对C5决策树分类器和一类支持向量机（one-class support vector machine，OC-SVM）进行结合，提出了一种混合入侵检测系统（host-based intrusion detection system，HIDS），用于实现对计算机资源的有效监控，并生成有关异常或可疑活动的报告；Kardani等^[19]提出一种基于Stacking的混合叠加集成方法，用于提升边坡稳定性的预测性能，并利用线性矢量量化（learning vector quantization，LVQ）方法研究了变量对边坡稳定性的重要性。

本文对基于Stacking集成航班延误预测模型进行研究，可以实现对航班延误时长的多等级预测。首先，融合航班数据与天气数据，并筛选出恶劣天气条件下的样本，进行特征选择与不平衡处理，得到可用于训练的数据集。基于Stacking集成框架，以LightGBM（light gradient

boosting machine）、XGBoost（extreme gradient boosting）、GBDT（gradient boosting decision tree）、RF（random forest）、KNN（k-nearest neighbor）为基础学习器，以SVM为元学习器，进行分类结果预测，采取多种评价指标，与单一模型和其他集成模型进行比较，充分验证模型在航班延误问题上具有良好的预测性能。

1 航班延误数据集的构建

1.1 航班数据与天气数据关联融合

纽瓦克自由国际机场在美国主要机场中延误情况较严重，且容易受到暴风雪、雷暴等恶劣天气影响，相关航班的历史数据来源于美国交通运输统计局（Bureau of Transportation Statistics，BTS），天气的历史数据来源于美国国家海洋和大气管理局（National Oceanic and Atmospheric Administration，NOAA）。

选取2019年全年的样本，将航班数据与天气数据以半小时为时间节点进行融合，对数据进行缺失值填充、异常值筛选等预处理操作。将航班延误的严重程度按照延误时长划分为5个等级，分别为未延误、轻度延误、中度延误、高度延误以及重度延误^[20]，对应0至4的分类编码，用作模型预测的标签，划分标准见表1^[20]。

1.2 恶劣天气的样本筛选

初始数据集的样本多达15万余条，经过不平衡处理，数据量会进一步增多，给后续集成模型的训练造成很大的时间开销。恶劣天气与航班延误的严重程度关系密切，对恶劣天气下的数据样本进行筛选，作为用于训练的典型数据，有利于在不影响预

表1 航班延误等级划分

延误等级	延误时间	分类编码
未延误	$T \leq 15 \text{ min}$	0
轻度延误	$15 \text{ min} < T \leq 60 \text{ min}$	1
中度延误	$60 \text{ min} < T \leq 120 \text{ min}$	2
高度延误	$120 \text{ min} < T \leq 240 \text{ min}$	3
重度延误	$T > 240 \text{ min}$	4

测性能的前提下，节省训练成本，降低模型过拟合的可能性。

机场使用最低标准只写明不允许起飞或着陆的限制条件，规定了最低安全保障，但不意味着能依此判定航班的延误情况，且不同机场、不同机型执行标准不同，常会发生在同一时间、同一机场，有的航班可以准时起飞，有的航班却会延误的情况。因此，恶劣天气下典型样本的筛选主要以能见度、低空云等5个容易影响航班延误的天气因素为指标，剔除对航班的正常运行没有影响或影响很小的样本，具体判定标准见表2。

能见度对飞行路径的选择与飞行操作的安全性都有重要影响，气象台的天气预报通常将其划分为1 km、2 km、4 km、6 km、8 km、10 km和10 km以上几个等级，达到10 km及以上的值意味着能见度非常良好，基本不会对航班的正常运行造成不利影响，因此被剔除，而余下能见度等级的样本对航班延误的影响，还需要结合其他航班信息与天气信息进行进一步分析，予以保留。

在飞机的起飞、降落以及巡航过程中，云底高度为2 500 m以上的中云与云底高度为5000 m以上的高云通常不会直接影响飞机的视野或飞行条件，而低空云接近地面，更容易导致能见度下降，加大起降过程中的风险，因此航空公司通常会对低空云采取更谨慎的态度，延误或取消航班。

降雨和强风同样会对飞机的飞行环境产生影响。小时降水量小于3 mm时可被归类为轻微降雨，通常不会对飞机造成直接阻碍，在降雨量逐渐增大后，跑道上的积水会影响飞机起降的安全性，并在飞行过程中导致周围能见度降低、飞行视野变差。降雨本身也会改变空气流动的模式，导致风速和风向在短距离内发生急剧变化，从而产生低空风切变，对飞机的飞行性能产生重大影响。风速大于24 m/h时可以被归为强风，在强顺风情况下，飞机起飞所获得的升力会降低，降落时也需要更长的跑道距离用于减速到停止，在环境设施有限的机场很容易出现飞机起降困难的情况；在强侧风的情况下，飞行条件更复杂苛刻，飞行员需要更精确的操作以保持飞机朝向和稳定性，避免偏离跑道或预定航线。

异常天气可分为雾、雪、冰雹、沙暴、雷电等多种情况，雾会降低能见度，使飞行员难以看清周围环境；雪会降低跑道的摩擦力，延长起飞和着陆时的滑行距离，甚至影响飞机的气动性能；冰雹可能会导致机身结冰，并造成严重的机体损伤；沙暴会产生严重的能见度问题，并且可能引入沙尘到飞机发动机中，导致发动机故障；雷电可能会导致飞机的电子系统受到影响而遭受损坏，直接被雷击中也可能导致机体结构受损，从而影响飞行器的控制和飞行安全。

表2 恶劣天气下的样本判定标准

指标	可能造成的影响	判定标准
能见度	难以判断飞机与跑道之间的相对关系	能见度小于8 km
低空云	云中含水量大,能见度差,阻碍飞行	云底高度低于2 500 m
降雨	跑道积水,能见度差,低空风切变	小时降雨量大于3 mm
强风	顺风减低升力,侧风增加操作难度	风速大于24 m/h
异常天气	道路湿滑,机身结冰,航空器受损	雾、雪、冰雹、沙暴等

1.3 特征选择与不平衡处理

原始数据集包含的95个特征中，航班相关特征为75个，天气相关特征为20个，其中存在部分无关特征与冗余特征，这些特征会增加模型的复杂程度，造成过拟合的后果。采用方差过滤与递归特征消除的方式对特征进行筛选^[21]，设定方差过滤的阈值为1，剔除掉样本差异较小的特征，减少后续递归特征消除的运算成本，再将LightGBM作为基模型进行多次训练，每次训练都对特征的重要程度进行排序，并剔除排名最靠后的特征，直至特征个数满足要求。最终筛选出30个对航班延误预测重要性最高的特征，其中包括20个航班相关特征与10个天气相关特征，部分特征选择结果见表3。

在实际的航班运行情况中，未延误的航班通常占据总航班的大多数，这也导致了在划分的5个延误等级中，未延误等级的样本数量远多于其余4个等级，使用这样带有不平衡特性的数据集进行预测，会使模型的预测结果更偏向于多数类，即更容易将未知的航班样本预测为“未延误”，影响模型对其他等级延误航班的预测能力。

采用SMOTE+Tomek Links组合采样

的方式，根据少数类样本的最近邻随机生成新样本，再删除互为最近邻的分属不同类别的样本对，可以在保持边界清晰的情况下，使不同标签的样本占比均衡^[21]。不平衡处理前后各等级标签的航班占比分布如图1所示。

2 基于Stacking的航班延误集成预测模型

2.1 Stacking集成学习总体思路

集成学习的思路是构建并结合多个学习器来完成学习任务，相比单一模型直接对结果进行输出，集成模型会对不同学习器的结果先按照某种策略进行组合，再经过综合判断来决定最终的分类，往往能取得比单一模型更好的性能。

Stacking集成学习的组合策略没有采用简单的投票法或求平均法，而是将模型分为两层结构进行训练，对航班延误进行预测的Stacking集成学习总体思路如图2所示。对于每个单一的基分类器，所用的训练集都是由K折交叉验证的方法构造的，将用于训练的航班延误数据划分为数量接近的K个子集，并保证每个子集的样本不重叠，依次选用其中的一个子集作为测试

表3 特征选择结果(部分)

特征名称	特征描述	重要程度排名	选择结果
Quarter	季度	15	剔除
Day_of_month	每月第几天	1	选用
Scheduled Departure Time	计划起飞时间	1	选用
Distance	飞行距离	1	选用
Destination City	目的城市	9	剔除
Sky Conditions	云层状况	1	选用
Sea Level Pressure	海平面气压	2	选用
Station Pressure	天气站气压	10	剔除
Wind Speed	风速	1	选用

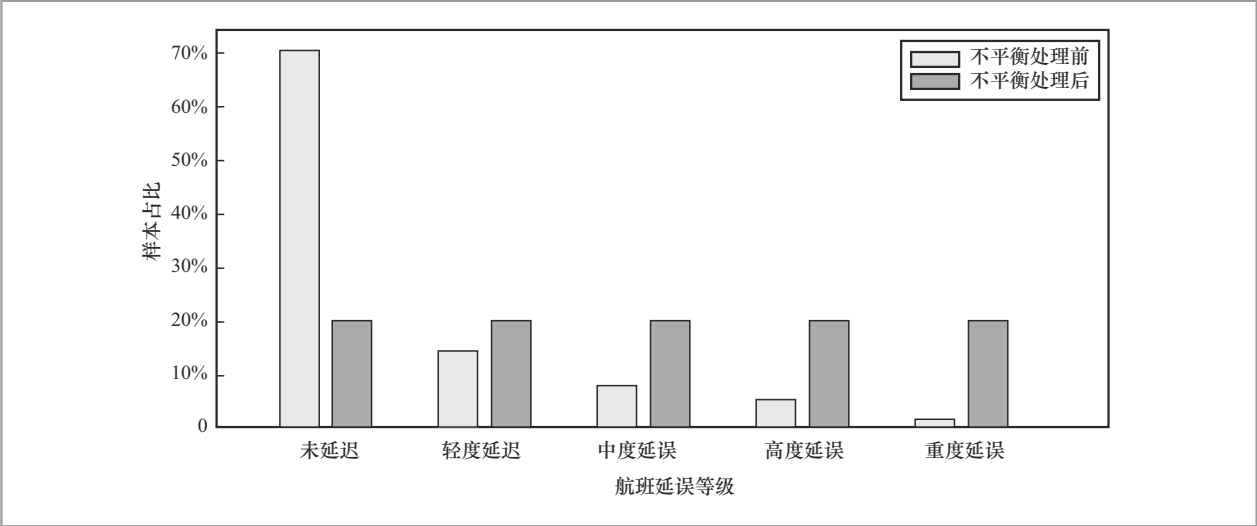


图1 不平衡处理前后样本占比分布

集，余下的 $K-1$ 个子集作为当次的训练集，可以得到 K 次的训练结果。将不同分类器的训练结果进行堆叠合并，可以形成一个与初始训练集样本数量相同的特征矩阵，作为第二层的新的输入。在第二层的训练中，由于前面的过程中已经使用了复杂的非线性变换，通常会选择较简单的分类器，以线性分类器为主，将结合了多种基分类器预测结果的输入特征矩阵进行

训练，输出最终的预测结果。

2.2 个体分类器的选择

Stacking 算法作为异质集成算法，基分类器的算法各不相同。为使整体模型的预测性能更优，对基分类器的选取要求应为“好而不同”，即基分类器本身要有一定的准确性，使整体模型可以更好地学习到

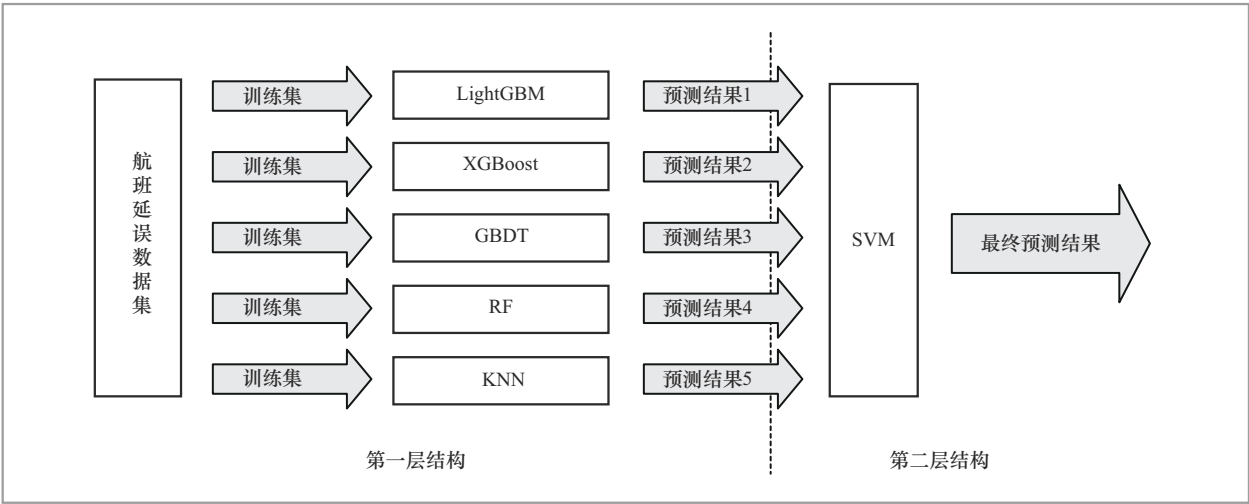


图2 Stacking集成学习总体思路

基分类器的优势,为后续的融合提供稳健的基础,且各个基分类器之间要有一定的差异,从而提升模型的整体泛化水平与预测能力。

在航班延误的多等级预测中,选择表现良好的5个分类器作为基分类器,分别为LightGBM、XGBoost、GBDT、RF、KNN。其中,GBDT是一种经典的Boosting算法,其采用一阶泰勒展开的误差计算方式,适合处理非线性关系和高维特征;XGBoost在GBDT的基础上引入了二阶泰勒展开的误差计算方式,以及在决策树构建阶段加入正则项的方法,有效提高了模型的泛化能力和鲁棒性,同时也降低了过拟合的风险;LightGBM采用基于直方图的加速算法,能够快速高效地处理大规模数据集,其决策树生长策略采用带有深度限制的按叶生长,结合单边梯度采样与互斥特征,能够有效减少训练所需的数据与特征,提升训练速度和准确性;RF采用Bagging的集成学习方式,通过构建多个决策树并对其进行平均来降低模型的方差,有效抑制了过拟合,并且由于个体学习器间不存在强依赖关系,可以并行化处理,提高了训练速度;KNN通过测量不同特征值之间的距离对样本进行分类,是一种更简单直观的分类方法,且基于局部邻域进行决策,具有对异常值不敏感的优势。5种基分类器对航班延误的整体预测准确率都在80%以上,且构成方式有一定差别,同时满足准确率与多样性的选择条件,选择以上分类器作为基分类器,能够充分利用它们各自的优势,从而在航班延误预测任务中获得更好的性能和泛化能力。

在元分类器的选取中,拟定AdaBoost、LR(logistic regression)、SVM、朴素贝叶斯4种较简单的分类器,

分别对LightGBM、XGBoost、GBDT、RF、KNN 5个基分类器进行Stacking集成,最终根据不同分类器实验结果,选定性能最优的SVM作为原分类器。

2.3 Stacking集成学习预测流程

航班延误数据的Stacking集成学习预测流程如图3所示,采用5折交叉算法对原始数据集进行划分,这样做的目的是实现数据的特征转换,保证所有数据在训练过程中只会被使用一次,有效避免过拟合,整体预测流程如图3所示。

步骤1:融合航班数据与天气数据,并按照恶劣天气的筛选指标,对不符合要求的数据进行剔除,并进行缺失值填充、数据编码等预处理。

步骤2:结合方差过滤与递归特征消除方法进行特征选择,保留30个对航班延误预测影响较大的相关特征,将延误时长划分为5个等级,运用SMOTE+Tomek Links进行不平衡处理,得到用于训练的数据集。

步骤3:将原始数据集进行划分,其中将75%作为训练集 D ,余下的25%作为测试集 T 。采用5折交叉验证的方法,将训练集划分为5个数量接近的子集 $D_1, D_2, \dots, D_k, D_k$ 与 D_{-k} 分别为交叉验证中第 k 折的测试子集与训练子集,其中 $D_{-k}=D-D_k, k=1, 2, \dots, 5$ 。

步骤4:运用LightGBM选用 D_{-k} 为训练集, D_k 为测试集进行训练,可以得到预测结果 α_k 。将5次的预测结果按 $\alpha_1, \alpha_2, \dots, \alpha_5$ 列堆叠到一起,可以构成与训练集 D 样本数量相同的新向量 A_1 。同时对测试集 T 进行预测,将5次的预测结果 $\beta_1, \beta_2, \dots, \beta_5$ 取平均,可以构成与测试集 T 样本数量相同的新向量 B_1 。

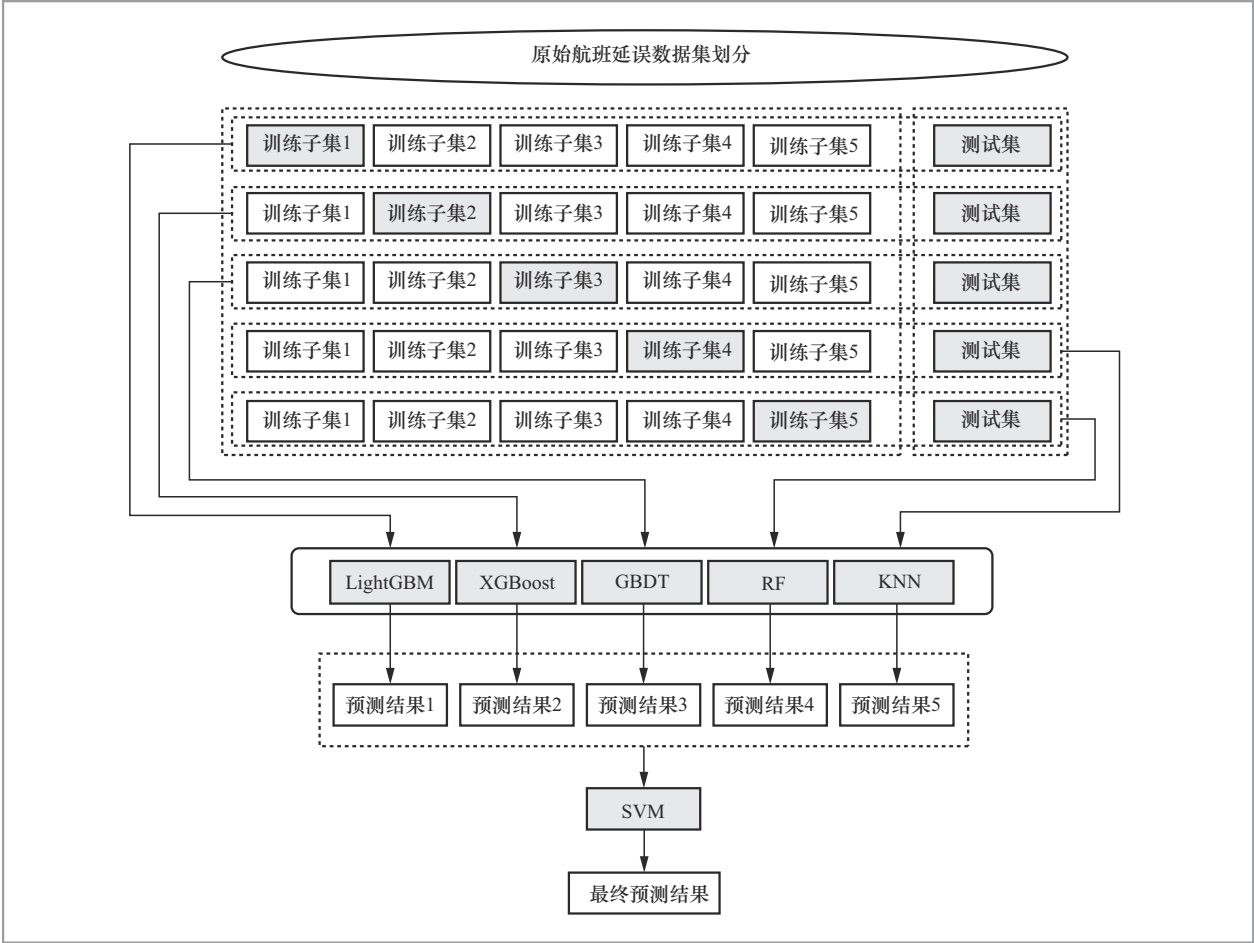


图3 Stacking集成学习预测流程

步骤5：依次使用XGBoost、GBDT、RF、KNN取代LightGBM重复上述步骤，可以得到训练集 D 所产生的 A_2, A_3, A_4, A_5 与测试集 T 所产生的 B_2, B_3, B_4, B_5 。

步骤6：将5个基分类器的预测结果 A_2, A_3, A_4, A_5 与训练集 D 的原标签 Y_D 合并为新的训练集， B_2, B_3, B_4, B_5 与测试集的原标签 Y_T 合并为新的测试集，使用元分类器SVM对航班延误数据进行预测，生成最终的预测结果。

2.4 评价指标

以航班延误预测问题为例，将延误时

长划分为 n 个等级进行预测，可得到多分类混淆矩阵，见表4。用于评估预测性能的指标主要有4种，分别为准确率（accuracy），精确率（precision），召回率（recall）以及F1分数（F1-score），其中准确率代表模型对延误航班样本的整体预测准确程度；精确率代表检出的延误航班样本中有多少是正确的；召回率代表在所有延误航班样本中有多少可以被检出；F1分数则综合了精确率和召回率的结果，更加客观全面地反映模型性能，其数值最小为0，越接近1代表模型的性能越好。

为更好地计算不同指标，可对应上述混淆矩阵，定义TP为实际延误等级为 i 且

表4 多分类混淆矩阵

延误等级	预测延误等级 1	预测延误等级 2	...	预测延误等级 n
实际延误等级 1	N_{11}	N_{12}	...	N_{1n}
实际延误等级 2	N_{21}	N_{22}	...	N_{2n}
...	...	...	...	...
实际延误等级 n	N_{n1}	N_{n2}	...	N_{nn}

预测延误等级为 i 的样本，即 N_{ii} ；TN 为实际延误等级非 i 且预测延误等级非 i 的样本，即 $\sum_{j \neq i} N_{jj}$ ；FP 为实际延误等级非 i 而预测延误等级为 i 的样本，即 $\sum_{j \neq i} N_{ji}$ ；FN 为实际延误等级为 i 而预测延误等级非 i 的样本，即 $\sum_{j \neq i} N_{ij}$ 。类精确率、类召回率、类 F1 分数的计算方法如式（1）至式（3）所示。

$$P_i = \frac{TP}{TP + FP} \quad (1)$$

$$R_i = \frac{TP}{TP + FN} \quad (2)$$

$$F_i = \frac{2 \times P_i \times R_i}{P_i + R_i} \quad (3)$$

对航班延误的每个等级分别进行评估有助于评价模型针对不同等级的预测性能，衡量模型的稳定性。如果一个模型只针对某个等级的性能较好，而其他等级的性能很差，同样无法对航班延误问题提供有效参考。除此之外，还需要对模型的整体性能进行评估，按照 Macro-Average 规则，先计算每一个等级的性能指标，再对所有等级的结果求算术平均值，可以得到整体的性能指标。整体的准确率、精确率、召回率、F1 分数的计算方法如式（4）至式（8）所示。

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

$$P_{Macro} = \frac{1}{n} \sum_{i=1}^n P_i \quad (5)$$

$$R_{Macro} = \frac{1}{n} \sum_{i=1}^n R_i \quad (6)$$

$$F_{Macro} = \frac{1}{n} \sum_{i=1}^n F_i \quad (7)$$

$$F_{Macro} = \frac{2 \times P_{Macro} \times R_{Macro}}{P_{Macro} + R_{Macro}} \quad (8)$$

3 实验结果与分析

3.1 实验环境与参数设置

实验所用平台为 Jupyter Notebook，个人计算机配置为 Windows 10（64 位）操作系统，Intel(R) Core(TM) i5-7200U 处理器，主频 2.50 GHz，运行内存 8 GB。调参方式选择贝叶斯调参，在寻找可以使目标函数达到全局最大的参数时，会考虑已有的先验信息，从而更好地调整当前的参数。相比于传统的网格调参，贝叶斯调参可以一次调整多个参数，只需设定大概的范围，而不对每个参数都设定不同的间隔来进一步细分，具有更少的运行次数与

表5 不同算法的超参数集

基分类器	超参数集
LightGBM	boosting_type='gbdt', objective='multiclass', metric='multi_logloss', num_leaves=254, max_depth=12, max_bin=147, learning_rate=0.4678, min_split_gain=0.0077, bagging_fraction=0.7524, feature_fraction=0.8775
XGBoost	objective='multi: softprob', eval_metric='mlogloss', learning_rate=0.6386, max_depth=15, subsample=0.9
GBDT	loss='deviance', learning_rate=0.4915, max_depth=15, subsample=0.9,
RF	criterion='gini', max_depth=15, max_features=0.5644
KNN	weights='distance', n_neighbors=5

更快的运行速度，同时避免了维度爆炸，最终的调参结果见**表5**。

3.2 基于 Stacking 集成的航班延误预测性能分析

用于实验的航班延误数据集结合了航班信息与天气信息，共包含 202 061 条样本、30 个特征变量，类别标签以航班延误的时长为标准，划分为 5 个延误等级，将数据的 75% 作为训练集、25% 作为测试集，对航班延误进行多等级预测。

选用 LightGBM、XGBoost、GBDT、RF、KNN 作为基分类器，其中 RF 与 KNN 无法单独进行多分类预测，需要与 OneVsRest Classifier 相结合，将多分类问题转化为多个二分类子问题，并为每个类别训练一个分类器，从而完成相应的训练任务。

元分类器的选择需要选用较简单的分类器防止过拟合，对 SVM、LR、NB、AdaBoost 分类器分别进行测试，其中 SVM 达到了最好效果，被选为最终的元分类器，具体测试结果见**表6**。

基于 Stacking 集成的航班延误预测模型的准确率为 95.25%，精确率为 95.31%，召回率为 95.27%，F1 分数为 0.9527，达到较高的预测水平。运用混淆矩阵图对模

型的总体性能进行更直观的评估，结果如**图4**所示。图中的每一列代表数据的预测类别标签，每一行代表数据的真实类别标签，对角线的数值则代表被预测正确的样本数量，对角线的颜色越深，意味着模型的分类性能越好。

ROC（receiver operating characteristic）即受试者工作特征曲线，是评价连续变量敏感度和特异度的综合指标，敏感度越高，漏查的延误航班样本就越少，特异度越低，误判的延误航班样本就越少。曲线图最左上角的一点代表敏感度和特异度能在两者平衡的状况下取得最佳效果，曲线越靠近这一点，模型的性能就越好。对预测的 5 个航班延误等级分别进行 ROC 曲线图的绘制如**图5**所示，模型不仅总体的分类预测性能较强，各等级的分类预测性能也都保持在较高水平，每条曲线的 AUC 都达到了 0.98 以上，预测情况稳定。

实验面向恶劣天气对航班延误进行预测，天气相关属性对于预测性能至关重要，

表6 不同分类器实验结果对比

分类器	准确率	精确率	召回率	F1 分数
SVM	95.25%	95.31%	95.27%	0.9527
LR	94.74%	94.76%	94.76%	0.9475
NB	93.71%	93.79%	93.73%	0.9375
AdaBoost	91.48%	92.13%	91.52%	0.9153

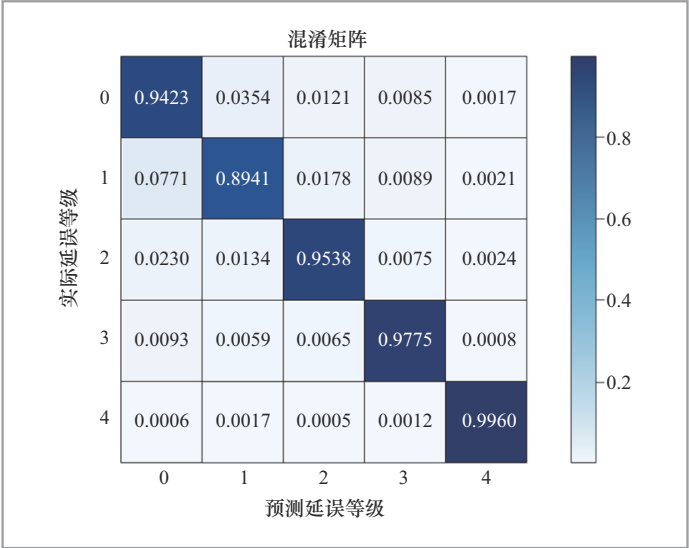


图4 混淆矩阵图

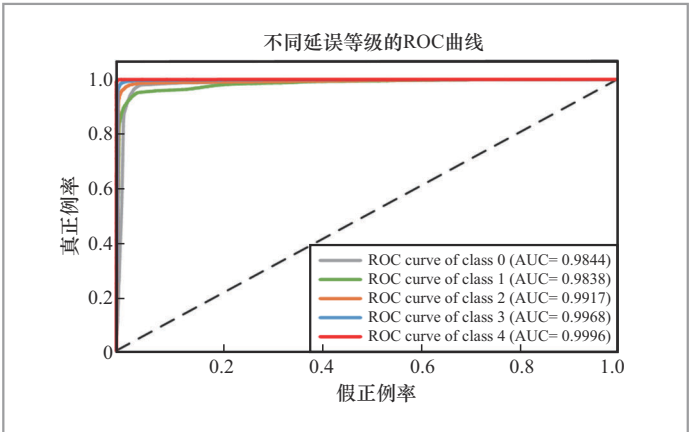


图5 ROC曲线图

表7 天气特征对预测性能的影响

数据集	准确率	精确率	召回率	F1分数
包含天气特征	95.25%	95.31%	95.27%	0.9527
不包含天气特征	92.88%	92.91%	92.90%	0.9287

设立不包含天气相关特征的数据集进行对比实验，见表7。实验结果证明，天气特征可以进一步提升模型的预测准确性，模型增加天气相关特征后，准确率、精确率、

召回率以及F1分数分别提升了2.47、2.40、2.37与2.40个百分点。

3.3 不同模型的预测性能对比

为充分验证模型性能，另外构建基于硬投票法、软投票法、加权算法3种集成预测模型与5种基学习器构成的单一预测模型用于比较。硬投票法中，每个基分类器的预测结果代表一票，根据少数服从多数的原则，得票最高的类别为最终的输出结果；软投票法中，每个基分类器预测结果为各个类别的概率值，将所有基分类器的结果求平均，最终概率最高的类别为最终的输出结果；加权算法中，参考每个基分类器准确率与错误率的比值，为不同基分类器赋予相应权重，使性能较好的分类器可以得到更多重视。不同模型各个类别的精确率、召回率于F1分数的具体对比结果见表8~表10。

在 LightGBM、XGBoost、GBDT、RF、KNN 5个基分类器构成的单一预测模型中，基于 LightGBM 的分类预测模型效果更优，5个等级的各项指标都稳定在80%以上，其中等级2、3、4的F1分数都达到了90%以上。不同单一模型对不同等级的预测精准程度不同，例如KNN在等级2、3、4、5的精确率上都优于总体性能最好的 LightGBM，但在等级1的精确率仅有53.16%，这说明单一模型对不同等级的预测性能并不稳定，预测性能有待进一步提高。

基于 Stacking 的集成预测模型针对这一问题进行了提升，综合各个基分类器的优势，采用堆叠式的双层集成学习框架进行多分类预测，在精确率、准确率以及F1分数这3个评价指标上，各个类别的评估数值均高于性能最优的单一预测模型与另

表8 不同模型的精确率对比

延误等级	分类模型								
	Stacking	软投票	硬投票	加权计算	LightGBM	XGBoost	GBDT	RF	KNN
0	94.23%	90.43%	91.12%	91.60%	91.91%	90.60%	82.14%	85.05%	53.16%
1	89.41%	87.15%	85.37%	86.99%	82.00%	81.03%	69.74%	69.63%	91.51%
2	95.38%	92.45%	89.41%	92.66%	88.92%	87.08%	76.08%	69.65%	95.07%
3	97.75%	96.63%	94.89%	96.71%	94.40%	93.01%	85.50%	87.63%	97.07%
4	99.60%	99.52%	99.02%	99.56%	99.27%	98.34%	95.70%	96.44%	99.09%

表9 不同模型的召回率对比

延误等级	分类模型								
	Stacking	软投票	硬投票	加权计算	LightGBM	XGBoost	GBDT	RF	KNN
0	89.44%	88.54%	86.80%	88.32%	85.58%	85.67%	81.59%	79.39%	93.37%
1	94.15%	90.40%	87.69%	91.02%	87.46%	85.71%	73.29%	74.51%	79.97%
2	96.24%	94.01%	93.23%	94.33%	91.45%	89.47%	78.53%	83.25%	85.74%
3	97.40%	95.17%	94.49%	95.49%	93.92%	92.16%	83.01%	81.03%	88.02%
4	99.30%	97.93%	97.67%	98.33%	98.09%	96.84%	91.70%	89.58%	92.41%

表10 不同模型的F1分数对比

延误等级	分类模型								
	Stacking	软投票	硬投票	加权计算	LightGBM	XGBoost	GBDT	RF	KNN
0	0.9177	0.8947	0.8891	0.8993	0.8863	0.8806	0.8186	0.8213	0.6775
1	0.9172	0.8875	0.8652	0.8896	0.8464	0.8330	0.7147	0.7199	0.8535
2	0.9581	0.9322	0.9128	0.9349	0.9017	0.8826	0.7729	0.7584	0.9016
3	0.9758	0.9590	0.9469	0.9610	0.9416	0.9258	0.8424	0.8420	0.9232
4	0.9945	0.9872	0.9834	0.9894	0.9868	0.9758	0.9366	0.9288	0.9563

外3种集成模型，且是唯一一个5个类别的F1分数都达到0.90以上的分类模型，充分证明该模型在航班延误分类预测问题上准确性与稳定性。

不同模型整体预测性能见表11，包含整体的准确率、精确率、召回率与F1分数，采用Macro-average的方式进行计算。从表11可以看出，基于Stacking的航班延误预测模型有良好的综合性能，其中整体准确率为95.25%，分别比软投票、硬投票、加权计算法分别高出2.03、3.31、

表11 不同模型的整体预测性能对比

分类模型	准确率	精确率	召回	F1分数
Stacking	95.25%	95.31%	95.27%	0.9527
软投票	93.22%	93.21%	93.24%	0.9321
硬投票	91.94%	91.98%	91.96%	0.9194
加权计算	93.49%	93.50%	93.51%	0.9348
LightGBM	91.27%	91.30%	91.30%	0.9125
XGBoost	89.98%	89.97%	90.01%	0.8995
GBDT	81.79%	81.63%	81.83%	0.8171
RF	81.64%	81.55%	81.68%	0.8141
KNN	87.19%	87.90%	87.18%	0.8624

1.76个百分点,比LightGBM、XGBoost、GBDT、RF、KNN分别高出3.98、5.27、13.46、13.61、9.06个百分点。实验证明,无论是对各个类别的预测性能,还是模型的整体预测性能,Stacking预测模型相比单一模型都有了较大幅度的提升,且在多个集成学习分类器的对比中仍处于最优,可以准确、稳定地对航班延误的严重程度进行预测。

4 结束语

对航班延误进行多等级预测,不仅有助于相关部门根据不同的严重程度做出决策,进行相应的资源调度与分配,还可以帮助旅客做好心理准备,提前对自己的行程规划进行调整,尽量减少航班延误带来的负面影响。基于Stacking的航班延误集成预测模型所进行的工作主要分为以下4点。

- 在航班历史数据中融合天气历史数据,设立恶劣天气的判定标准,更全面地考虑天气因素对航班延误的影响。
- 利用方差过滤与递归特征消除进行特征选择,降低模型的复杂度与训练成本。
- 将航班延误的严重程度划分多个等级,运用SMOTE+Tomek links重采样,消除数据的不平衡特性。
- 采用LightGBM、XGBoost、GBDT、RF、KNN作为基学习器,SVM作为元学习器,构建堆叠式的双层集成学习框架,并建立多项对比试验,充分证明天气对于航班延误预测的重要性,以及模型在航班延误预测问题上的准确性与稳定性。

未来的工作主要基于两点展开:一是实验所用的真实数据来自国外,与国内情

况会有一定差异,后续争取能与国内机场展开合作,更加结合实际地对航班延误问题展开研究;二是考虑空域限制、流量管理以及到达机场天气等相关因素,进一步提升模型的预测性能。

参考文献:

- [1] ZHENG Z, WEI W B, HU M H. A comparative analysis of delay propagation on departure and arrival flights for a Chinese case study[J]. Aerospace, 2021, 8(8): 212.
- [2] RODRÍGUEZ-SANZ Á, CANO J, RUBIO FERNÁNDEZ B. Impact of weather conditions on airport arrival delay and throughput[J]. Aircraft Engineering and Aerospace Technology, 2022, 94(1): 60-78.
- [3] ZANIN M, ZHU Y B, YAN R, et al. Characterization and prediction of air transport delays in China[J]. Applied Sciences, 2020, 10(18): 6165.
- [4] ŠKULTÉTY F, JAROŠOVÁ M, ROSTÁŠ J. Dangerous weather phenomena and their effect on en-route flight delays in Europe[J]. Transportation Research Procedia, 2021, 59: 174-182.
- [5] ZHU X T, LI L S. Flight time prediction for fuel loading decisions with a deep learning approach[J]. Transportation Research Part C: Emerging Technologies, 2021, 128: 103179.
- [6] LI Q, JING R Z, DONG Z S. Flight delay prediction with priority information of weather and non-weather features[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(7): 7149-7165.
- [7] BORSKY S, UNTERBERGER C. Bad weather and flight delays: the impact of

- sudden and slow onset weather events [J]. *Economics of Transportation*, 2019, 18: 10–26.
- [8] CHEN Z H, WANG Y X. Impacts of severe weather events on high-speed rail and aviation delays[J]. *Transportation Research Part D: Transport and Environment*, 2019, 69: 168–183.
- [9] 陈可嘉, 陈琳琳. 危险天气下航班等待与改航的实时集成优化[J]. *南京航空航天大学学报*, 2019, 51(6): 763–771.
- CHEN K J, CHEN L L. Aircraft holding and rerouting real-time integrated optimization under adverse weather[J]. *Journal of Nanjing University of Aeronautics & Astronautics*, 2019, 51(6): 763–771.
- [10] 王洪, 彭瑛, 郭聪聪, 等. 基于k-means的机场终端区对流天气场景识别[J]. *哈尔滨商业大学学报(自然科学版)*, 2021, 37(6): 695–702.
- WANG H, PENG Y, GUO C C, et al. Recognition of convective weather scenes in airport terminal area based on k-means[J]. *Journal of Harbin University of Commerce (Natural Sciences Edition)*, 2021, 37(6): 695–702.
- [11] 王春政, 胡明华, 杨磊, 等. 基于Agent模型的机场网络延误预测[J]. *航空学报*, 2021, 42(7): 324604.
- WANG C Z, HU M H, YANG L, et al. Airport network delay prediction based on Agent model[J]. *Acta Aeronautica et Astronautica Sinica*, 2021, 42(7): 324604.
- [12] 安洋, 孙健玮, 李倩, 等. 基于多源异构时空数据融合的交通流量预测模型[J]. *大数据*, 2023, 9(4): 69–82.
- AN Y, SUN J W, LI Q, et al. Urban traffic flow prediction based on the multi-source heterogeneous spatio-temporal data fusion[J]. *Big Data Research*, 2023, 9(4): 69–82.
- [13] LI M X, YAN C, LIU W. The network loan risk prediction model based on Convolutional neural network and Stacking fusion model[J]. *Applied Soft Computing*, 2021, 113: 107961.
- [14] JIANG M Q, LIU J P, ZHANG L, et al. An improved Stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms[J]. *Physica A: Statistical Mechanics and Its Applications*, 2020, 541: 122272.
- [15] 李泉伦, 陈争光, 焦峰. 基于Stacking集成学习的近红外光谱油页岩含油率预测[J]. *光谱学与光谱分析*, 2023, 43(4): 1030–1036.
- LI Q L, CHEN Z G, JIAO F. Prediction of oil content in oil shale by near-infrared spectroscopy based on stacking ensemble learning[J]. *Spectroscopy and Spectral Analysis*, 2023, 43(4): 1030–1036.
- [16] 马良玉, 耿妍竹, 梁书源, 等. 基于Stacking多模型融合的风电机齿轮箱油池温度异常预警[J]. *中国电机工程学报*, 2023, 43(S1): 242–251.
- MA L Y, GENG Y Z, LIANG S Y, et al. Early warning of abnormal temperature in gearbox oil pool of wind turbine based on Stacking multi-model fusion[J]. *Proceedings of the CSEE*, 2023, 43(S1): 242–251.
- [17] KHAN W, WALKER S, ZEILER W. Improved solar photovoltaic energy generation forecast using deep learning-based ensemble stacking approach[J]. *Energy*, 2022, 240: 122812.
- [18] KHRAISAT A, GONDAL I, VAMPLEW P, et al. Hybrid intrusion detection system based on the stacking ensemble of C5 decision tree classifier and one class support vector machine[J]. *Electronics*, 2020, 9(1): 173.
- [19] KARDANI N, ZHOU A N, NAZEM M, et

al. Improved prediction of slope stability using a hybrid stacking ensemble method based on finite element analysis and field data[J]. Journal of Rock Mechanics and Geotechnical Engineering, 2021, 13(1): 188-201.

[20] 吴仁彪, 赵婷, 屈景怡. 基于深度 SE-DenseNet 的航班延误预测模型[J]. 电子与信息学报, 2019, 41(6): 1510-1517.

WU R B, ZHAO T, QU J Y. Flight delay prediction model based on deep SE-DenseNet[J]. Journal of Electronics & Information Technology, 2019, 41(6): 1510-1517.

[21] 丁建立, 孙玥. 基于 LightGBM 的航班延误多分类预测[J]. 南京航空航天大学学报, 2021, 53(6): 847-854.

DING J L, SUN Y. Multi-classification prediction of flight delay based on LightGBM[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2021, 53(6): 847-854.

作者简介



孙玥（1997-），女，中国民航大学宁河校区管理中心助理工程师，主要研究方向为人工智能、数据分析。



丁建立（1963-），男，博士，中国民航大学计算机科学与技术学院教授，主要研究方向为大数据与人工智能在民航领域的应用。

收稿日期: 2024-04-23

基金项目: 国家自然科学基金民航联合基金重点项目(No.U2033205, No.U2233214)

Foundation Items: The National Natural Science Foundation and Civil Aviation Joint Key Project (No.U2033205, No.U2233214)