【航空科学与工程研究】

基于随机森林算法的航班延误时间预测模型研究

许振腾[1]，王　琪[2]

（1. 南京工业职业技术大学 航空工程学院，江苏 南京 210046；
2. 上海机场（集团）有限公司，上海 201300）

摘　要：航班延误一直是影响航空公司运行效率和经济效益的关键问题。航班延误时间预测的方法较多，但是存在准确率不高、影响因素考虑不全面等问题。为了解决上述问题，提出了一种基于数据驱动的航班延误时间间接预测模型。该模型以机场协同决策系统的数据为依据，采用随机森林算法，直接预测航班在场停留时间和最终起飞时间，然后计算得出航班延误时间。通过实验数据进行验证，证明该预测模型按照 15 mim 航班延误标准进行评估的准确率达 100%。该模型可以为航空公司的航班延误预测提供支持，从而有针对性地优化机队运行流程，提高运行效率。

关键词：航班延误；预测模型；随机森林；数据编码

中图分类号：V 35　　**文献标识码**：A　　DOI：10.13486/j.cnki.1673-2618.2024.02.004

0　引言

航空运输业一直以来都面临着诸多挑战，其中之一便是航班延误问题。航班延误不仅对航空公司和机场运营商造成巨大的经济损失，还给旅客带来了诸多不便。尽管航空公司已经采取了各种措施来减少延误，但仍然存在许多因素（如气象条件、机械故障、航空管制等），可能导致航班延误的发生。根据全球航班信息和航空数据的在线平台 FLIGHTSTATS 提供的信息，2023 年 8 月份全球航班准点率最高的 10 个机场的平均准点率为 83.15%，亚洲地区航线的平均准点率只有 79.62%，北美地区航线的平均准点率最低，只有 71.82%[1]。

为了降低航班延误的影响，学者们对航班延误时间进行预测，建立了不同的预测模型。文献[2]结合 DenseNet 和 SENet，提出了深度 SE-DenseNet 算法模型进行延误预测；文献[3]使用支持向量机（Support Vector Machine，SVM）的模型来探索飞行延误结果之间的非线性关系；文献[4-5]使用了多种常用的机器学习算法（如多元线性回归算法、决策树算法、随机森林算法、梯度决策树算法等）进行对比分析，旨在提高延误预测的准确性。虽然在航班延误预测的方法选择上有较高的自由度，但是影响航班运行的因素偏多，大部分研究没办法做到影响因素全面覆盖。如文献[6]未考虑特殊航空公司、飞机注册号和起始/目的地等；文献[7]未使用空中交通控制数据等。

基于上述原因，本文提出了以航班运行数据为基础的数据驱动航班延误预测模型。该模型的数据来

收稿日期：2024-03-04

基金项目：南京职业技术大学教育研究课题（ZBYB22-01）

第一作者简介：许振腾（1989—），男，山东聊城人，讲师，硕士，主要从事载运工具运用工程研究。

E-mail：xuzhenteng@163.com

源为机场协同决策系统（ACDM），采用的算法为随机森林算法，预测过程采用间接预测，即通过预测相关数据，计算出航班延误时间，而不是直接预测航班延误时间；通过对真实航班运行数据进行预测，证实该方法切实可行。

1　数据处理及分析

本节主要对从 ACDM 中下载的数据进行初步处理和分析，并对一些基本参数进行定义。通过数据的处理，一方面对数据格式进行修改，增强数据的易读性，另一方面保证用于预测数据的准确性。数据分析主要是对航班运行数据进行初步计算分析，确定航班运行的基本趋势和规律，为进一步的研究提供依据。

1.1　数据选取

ACDM 旨在建立一个高度协作的机场运营环境，确保机场资源的高效利用、增加容量、减少延误、增强乘客体验、降低运营成本，并促进整个航空业的可持续性。该系统在全球范围内得到广泛应用，为各种规模和类型的机场提供支持。ACDM 中的数据包含很多时间节点，如计划到达时间（STA）、最晚周转时间（LTOT）、计划出发时间（STD）、登机口开启时间（GOT）等。

为了进一步确定实验数据选取的时间，对 2019 年至 2023 年某机场的航班运行数据进行了分析，结果如图 1 所示。

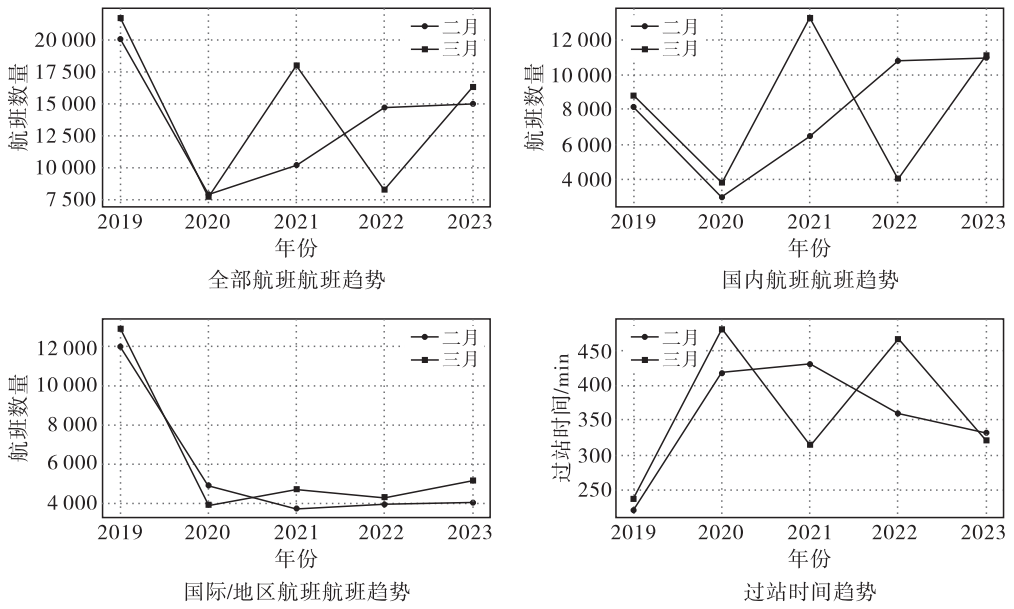

图 1　2019-2023 年的航班量及过站时间

由图 1 可以看出，2023 年总航班量与 2019 年还有较大差距，但是国内航班量已超过疫情期的水平，差别较大的是国际/地区航班量。国际/地区的航班数量虽然还没有达到疫情期的水平，但是也在逐年递增，伴随时间的推移，必将会超过疫情前的航班量。2023 年的过站时间相较 2020 年已有明显降低，但是依然比 2019 年用时要长。导致这一现象的原因主要是疫情期间为航班周转增加了一些新的操作流程，而这些流程在疫情结束后依然保留，因此过站时间变长。结合上述分析，本文选用了某机场 2023 年 6 月份的运行数据。

1.2　数据处理

从 ACDM 导出的数据包含很多信息，其中有些信息如果不处理掉，可能会对分析造成负面影响，因此需要先进行数据预处理。本文的数据处理过程共包含 6 个步骤：第 1 步，将节点时间转换成时长；第 2 步，根据到港状态指示数据（STUA 和 STUD），筛选出实际到达机场的航班；第 3 步，根据到港时间（ALDT）选择时间为 07:00－19:00 的航班；第 4 步，根据目标停机位（TAR）选择 1 号停机坪的航班，TAR＝Nr. 1-

12.14-32;第 5 步,提取机型(ITY)、停机位(TAR)、实际到达时间(ALDT)、计划起飞时间(STD)、上轮挡、开客舱门、登机口开、撤轮挡、实际起飞时间(ATOT)、最终起飞时间(LTOT)、延误时长;第 6 步,删除表格中的空数据、删除或修正错误数据。在上述步骤中,第 1 步和第 6 步需要进行计算,其余步骤需要根据条件进行数据筛选。

ACDM 中的大部分时间节点数据都是日期和时间的形式,为了方便后续的处理,需要设定一个起始时间点,然后将其他时间转化成时间差,以分钟为单位。除了已知的时间节点之外,预测分析用到的时间也是以时间差的形式给出。大多数学者进行数据清洗时优先选择删除错误数据,或者根据已知数据补充空数据。本文提出了对错误数据进行修正的新方法,即通过相关数据的数值特点,来判断错误数据产生的原因,根据原因对错误数据进行纠正。

因为航班运行数据中采集了所有的航班,也包括取消航班和备降航班,但是这部分数据会大大降低延误预测的精确度,因此需要通过 STUA 和 STUD 来进行筛选。筛选过程中,保留 STUA 和 STUD 为“到达”的数据,删除其他数据。本文重点解决正常过站航班的延误问题,不考虑夜间停场航班的延误,因此通过 ALDT 时间进行了筛选。这样做的一个好处是,时间具有线性关系,可以直接进行编码。如果考虑 24 h 内的航班延误情况,则需要对时间进行三角变换[8]。

1.3 数据分析

本节内容主要对平均服务时间进行统计分析。平均服务时间用来衡量一个航班在机场接受地面勤务的总时间,该时间并非净服务时间,而是包含等待勤务的时间,其计算公式为

$$T_{serv}=\frac{\sum_{1}^{n}(t_{ob_i}-t_{ib_i})}{n}(i=1,2,\cdots,n)。$$

式中,T_{serv} 为平均服务时间,t_{ob_i} 为撤轮档时间,t_{ib_i} 为上轮档时间,n 为同一机型航班数量。经计算后的平均服务时间如表 1 所示。

表 1 平均服务时间统计数据

机型	航班数量	中位时间/min	最长时间/min	最短时间/min	平均时间/min
B737	26	68.0	185	43	86.00
B738	276	72.0	406	40	93.42
A20N	225	97.0	407	45	121.40
A319	35	113.5	330	46	128.83
A320	43	104.0	444	45	118.35
A321	109	85.5	469	46	139.38
A332	56	130.0	518	67	174.20
A333	63	156.0	419	73	183.41
A359	26	78.5	484	62	122.27
B772	13	100.0	170	80	106.54
B773	8	81.5	93	72	81.50
B77W	79	208.0	511	62	236.23
B788	21	120.0	140	78	112.29
B789	8	184.0	437	137	236.38

为了进一步阐述各数据之间的关系,将计算结果绘制成 K 线图,如图 2 所示。根据上述统计可知,不

同机型的最少用时非常接近，因此这一数据的参考意义不大；最大用时因机型不同差别较大，可作为极限条件下的时间参考。除了 B773 和 B788 的结果为阴线外，其他机型均为阳线，即平均用时大于中位用时，说明多数航班的服务时间少于平均用时，因此如果在后续计算中选用平均用时，会给预测带来较大的裕度。

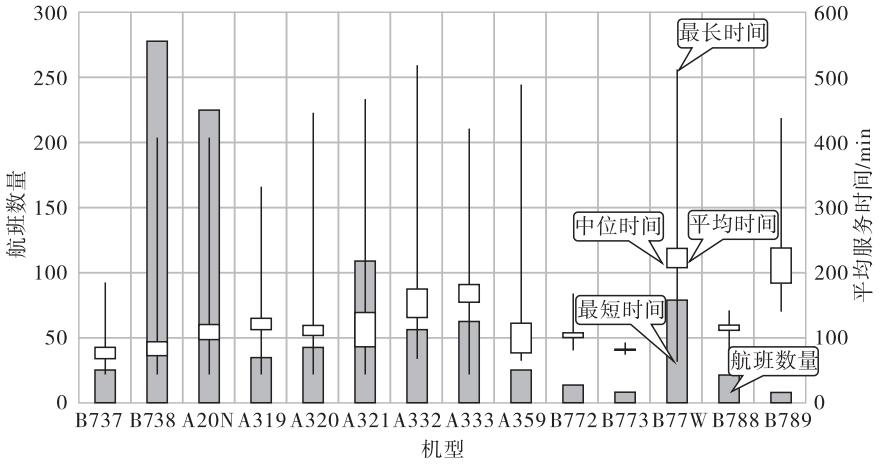

图 2　平均服务时间 K 线图

2　预测模型构建

本节将对不同的算法进行分析，并给出选择随机森林算法进行模型建立的原因，随后结合航班延误预测问题，使用随机森林算法进行建模。

2.1　算法选择

在预测模型领域，目前比较流行的算法主要有两类，一类是机器学习算法，另一类是深度学习算法。两种算法的应用领域略有不同，机器学习算法主要针对的是大数据模型的预测，深度学习算法则更倾向于处理一些文本、图形类的数据。因此，在进行过站模型的预测时，大部分选用的是机器学习算法。常用于机器学习的算法有：随机森林模型、广义线性模型(GLM)、梯度提升模型(GBM)、K-means、Prophet。

上述算法都有各自适用的领域，也都存在一些不足。广义线性模型需要相对较大的数据集，并且容易受到异常值的影响；梯度提升模型按顺序构建每棵树时，往往需要更长的时间；K-means 算法广泛应用于医疗保健领域的预测分析中；Prophet 算法则在容量规划中非常有用。

本文选用的预测模型为随机森林模型，主要考虑该模型具有以下优点：(1)模型采用多棵树的运算方式，可以有效减小单棵树的误差；(2)随机森林模型可以有效抵制过度拟合，并且可以同时处理较多的数据；(3)对于多变量的预测，可以估计变量的重要性，并在出现数据丢失时保持预测的准确性。

2.2　建立模型

本模型的预测目标是航班延误时间，因此有两种预测方案：直接预测和间接预测。直接预测是指，根据已知的特征变量预测出航班延误时长。间接预测需要先寻找航班延误时间的计算依据，确定与其计算相关的参数，通过模型预测出上述参数，并计算得到航班延误时长。直接预测和间接预测各有优劣：直接预测只预测一个结果，所以对特征变量要求较高，需要找到能够准确预测出结果的特征变量，难度较大；间接预测因为有多个预测结果，所以会造成预测误差的叠加。通过大量实验发现，直接预测难以获得满意的预测效果，而间接预测的正确率更高，因此最终采用了间接预测的方法。随机森林预测模型如图 3 所示。

由图 3 可以看出，随机森林预测过程主要包括三部分内容：数据拆分、训练和测试。在数据拆分环节，数据集被分成两个子集，分别是训练集和测试集，各占总数据集的 50%。随机森林回归(RFR)是一类基于决策树的机器学习算法，在使用训练集进行模型训练的过程中，大部分数据将被直接打包，并通过不同

的决策树进行拟合，另外有 10％的数据用于检测模型是否存在过拟合。

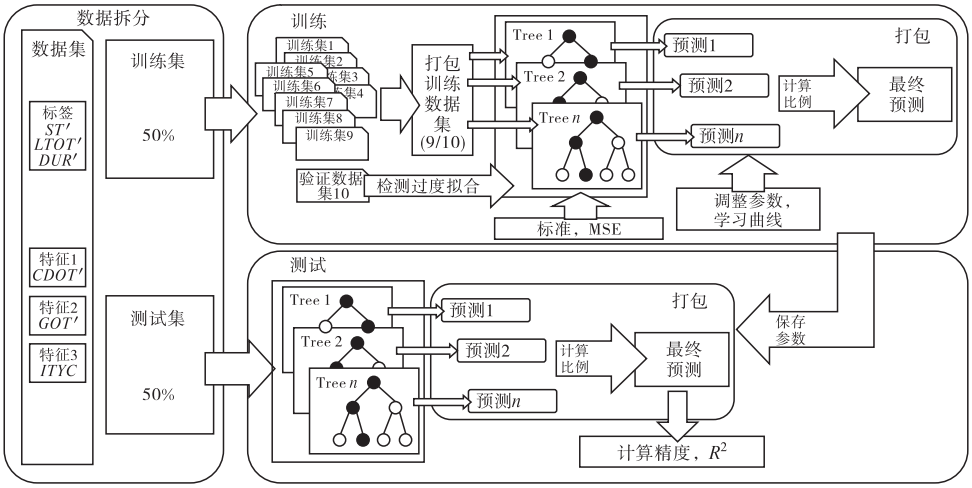

图 3 随机森林模型预测过程

决策树评价回归质量的标准选用的是均方误差(Mean Squared Error，*MSE*)，计算公式为

$$MSE = \frac{1}{m}\sum_{1}^{m}(y_{f_i} - y_{c_i})^2 (i = 1,2,\cdots,n)。$$

式中：m 是一棵决策树上的节点数量，y_{f_i} 是父节点的数值，y_{c_i} 是子节点的数值。在决策树每一个节点的选择中，*MSE* 更小的节点将被视为回归质量高的节点。

每棵树都是随机从 3 个特征中选择固定数量的特征子集，且都尽最大可能地生长，并且没有剪枝的过程。最终每棵树都会获得对应的预测结果，根据预测结果占比，可以获取最合理的结果，并作为最终预测结果。在这一过程中，根据学习曲线，不断整理参数，以获取最佳结果。

测试集的主要作用是验证模型的效果。在测试环节所用的决策树不需要再进行训练，直接使用在训练环节已经确定的树。同样，打包所用的参数也是训练过程中保存下来的，当获取到测试数据集的最终预测结果后，可以计算本模型的精度。本文选用确定系数 R^2 来对模型进行评估，计算公式为

$$R^2 = 1 - \frac{\sum_{1}^{n}(\hat{y}_i - y_i)^2}{\sum_{1}^{n}(\overline{y}_i - y_i)^2}(i = 1,2,\cdots,n)。$$

式中：n 数据集的数据量，y_i 是原始数据，$\hat{y}_i$ 是预测数据，$\overline{y}_i$ 是原始数据均值。R^2 的取值范围为[0，1]，如果结果是 0，说明模型拟合效果很差，如果结果是 1，说明模型无错误。

2.3 数据编码及计算过程

ACDM 中的数据包含分类、时间相关和数值三大类。针对上述类型的数据，主要的编码方法有：目标编码、三角编码和数字编码[9]。本文在编码过程中用到了目标编码和数字编码，同时针对时间相关特征数据提出了时间转换编码方法。时间转换编码方法的核心原理是建立时间基线，将日历时间转变成时长。本文使用的时间基线是上轮挡时间 t_{ib}。

根据 ACDM 提供的数据可知，航班延误时间(FDT)的计算公式为

$$FDT' = t_{\mathrm{ATOT}} - t_{\mathrm{LTOT}}。 \tag{1}$$

将最终起飞时间(LTOT)转换成时长的计算公式为

$$LTOT' = t_{\mathrm{LTOT}} - t_{\mathrm{ib}}。 \tag{2}$$

航班在场停留时间(DUR)的计算公式为

$$DUR' = t_{\mathrm{ATOT}} - t_{\mathrm{ib}}。 \tag{3}$$

航班服务时间(ST)的计算公式为

$$ST' = t_{ob} - t_{ib}。\tag{4}$$

将客舱门开启时间(CDOT)转换成时长的计算公式为

$$CDOT' = t_{\mathrm{CDOT}} - t_{\mathrm{ib}}。\tag{5}$$

将登机口开启时间(GOT)转换成时长的计算公式为

$$GOT' = t_{\mathrm{GOT}} - t_{\mathrm{ib}}。\tag{6}$$

公式(1)～(6)中，$(\cdot)'$所表示的时间均为时长，单位为 min，$t_{(\cdot)}$所表示的时间为日历时间，可直接在 ACDM 中获得。根据公式(1)～(3)可以得出

$$FDT' = DUR' - LTOT'。\tag{7}$$

根据公式(4)和公式(7)，确定了预测模型的标签，分别是 ST'、DUR'和 $LTOT'$。特征变量的选择则是通过大量的实验来完成的，通过实验确定了 DUR'、$CDOT'$和 $ITYC$(机型编码)三个特征变量。其中 DUR'、$CDOT'$可以通过公式(3)和公式(5)计算得出，$ITYC$ 则需要进行目标编码。

$ITYC$ 是对 ACDM 中的 ITY 进行编码。ITY 对应的数据为机型编号。在本文所涉及的数据中，与机型关系最密切的是勤务时间，因此将表 1 中机型的平均服务时间作为目标编码的原始数据，最终确定的 $ITYC$ 如表 2 所示。

表 2　$ITYC$ 对照表

ITY	B737	B738	A20N	A319	A320	A321	A332
$ITYC$	0.86	0.9342	1.214	1.2883	1.1835	1.3938	1.742
ITY	A333	A359	B772	B773	B77W	B788	B789
$ITYC$	1.8341	1.2227	1.0654	0.815	2.3623	1.1229	2.3638

3　预测过程及结果

本节主要对前文所述模型和方法进行验证。使用 python 中的 sklearn 工具包，结合章节 2 中的计算过程，实现随机森林预测模型的搭建，随后进行参数设置。经过大量实验和调试后，最终确定参数数值：test_size＝0.5，random_state＝1，max_depth＝20，n_estimators＝1000，n_jobs＝－1。

本文使用的总数据集为某机场 2023 年 6 月份(不含 6 月 20 日)的运行数据，原始数据 1092 条，经过处理后的可用数据共计 974 条。在模型训练过程中，测试集占比 50%，使用经实验确定的模型参数，训练出的模型效果如表 3 所示。

表 3　随机森林模型效果

特征	标签	R^2_train	R^2_test
DUR'、$CDOT'$、$ITYC$	ST'	0.997	0.985
	DUR'	0.979	0.911
	$LTOT'$	0.996	0.974

根据表 3 的结果可以看出，训练集的 R^2 基本接近 1，说明模型误差很小，测试集除了 DUR'之外，其他两个标签的 R^2 也都接近 1。对于一般训练模型，R^2 值大于 0.8，说明模型的训练效果较好，因此本模型对 DUR'的预测远远高于一般标准。下面使用本模型对 6 月 20 日共 20 个有效航班的延误情况进行预测。

本文选用的 20 个航班的到场时间最早为 7:35，最晚为 18:40，包含了当日除停场航班之外的所有过站航班。另外，从 15:28 至 15:55 共有 5 个航班到达机场，航班密度较大。因此从时间跨度和航班密度两个维度来分析，本次选择的数据非常有代表性。使用训练后的模型对上述航班进行延误预测，预测结果如

图 4 所示。

通过图 4 中的折线图可以看出，预测值和真实值基本一致，通过柱状图可以看出，两者误差均在±9 min 以内。本次预测最大正向误差约为 6.7 min，最小正向误差约为 1 min，最大反向误差约为－8.2 min，最小反向误差约为－0.3 min，误差跨度（正向误差减反向误差）保持在 1.3～14.9 min。按照中国民用航空局 15 min 定义航班延误的标准[10]，本次预测结果的正确率为 100%，即使将时间限制在±8 min 以内，本预测模型的正确率也高达 95%。

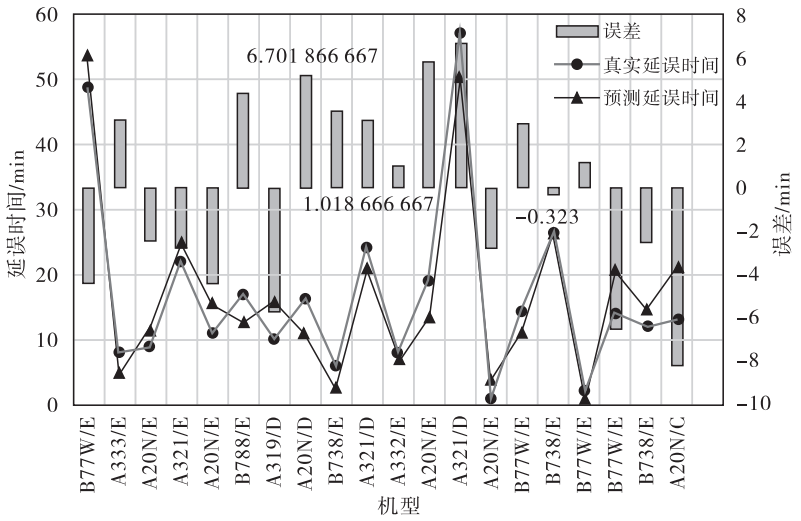

图 4 模型预测结果与真实航班运行数据对比

4 结论

经过上述理论研究与实践测试，本文实现了预设目标，主要取得如下成果：(1)分析了 2019 年以来航班流量及过站时间情况，从数据层面证实了疫情后航空产业的不断恢复；(2)通过计算得出了不同机型的平均服务时间，创造性提出了以服务时间为参考的机型编码方法；(3)根据航班延误时间难预测的特点，首次提出了以 ACDM 数据为驱动，以随机森林算法为手段的间接预测方法；(4)利用在场停留时间 DUR' 和最终起飞时间 $LTOT'$ 预测准确率高的特点，通过计算得出航班延误时间，提升了整体模型的预测准确率。

针对此类研究，本文需进一步完善，即减少预测模型的输入参数，提升模型的预测准确率，探索以大模型为基础的航班延误预测方法，降低对 ACDM 数据的依赖。

参考文献：

[1] CIRIUM. The on-time performance monthly report-airlines[EB/OL]. [2024-03-02]. https://www.cirium.com/thoughtcloud/cirium-monthly-on-time-performance-report-2023/.

[2] 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.

[3] ESMAEILZADEH E, MOKHTARIMOUSAVI S. Machine learning approach for flight departure delay prediction and analysis[J]. Transportation research record, 2020, 2674(8): 145-159.

[4] YW B, LIU B, TIAN Y, et al. A methodology for predicting aggregate flight departure delays in airports based on supervised learning[J]. Sustainability, 2020, 12(7): 2749.

[5] WANG Z, LIAO C, HANG X, et al. Distribution prediction of strategic flight delays via machine learning methods[J]. Sustainability, 2022, 14(22): 15180.

[6] BANSAL J C，SHARMA H，JADON S S，et al. Spider monkey optimization algorithm for numerical optimization[J]. Memetic computing，2014，6：31-47.

[7] YU B，GUO Z，ASIAN S，et al. Flight delay prediction for commercial air transport：a deep learning approach[J]. Transportation research part E：logistics and transportation review，2019，125：203-221.

[8] HORIGUCHI Y，BABA Y，KASHIMA H，et al. Predicting fuel consumption and flight delays for low-cost airlines[C]//Proceedings of the AAAI conference on artificial intelligence. The Twenty-ninth Innovative Applications Artificial Intelligence conference，February 4-9，2017，San Francisco，California USA. AAAI Press，2017，31(2)：4686-4693.

[9] ZOUTENDIJK M，MITICI M. Probabilistic flight delay predictions using machine learning and applications to the flight-to-gate assignment problem[J]. Aerospace，2021，8(6)：152.

[10] 中国民用航空局，CCAR-300，航班正常管理规定[Z/OL]. [2023-02-10]. https：//www. caac. gov. cn/XXGK/XXGK/MHGZ/2 01706/t20170621_44917. html.

Research on Flight Delay Time Prediction Model Based on Random Forest Algorithm

XU Zhenteng[1]，WANG Qi[2]

(1. *School of Aeronautical Engineering*，*Nanjing Vocational University of Industry Technology*，*Nanjing* 210046，*China*；
2. *Shanghai Airport* (*Group*) *Co.*，*Ltd.*，*Shanghai* 201300，*China*)

Abstract：Flight delay has long been a critical issue affecting the operational efficiency and economic performance of airlines. While there are various methods for predicting flight delay times，they often suffer from challenges such as low accuracy and incomplete consideration of influencing factors. To address these issues，a data-driven indirect prediction model for flight delay time is proposed. This model，based on data from the Airport Collaborative Decision Making(ACDM) system，employs the random forest algorithm to predict directly the aircraft's dwell time on the apron and the final departure time，from which the flight delay time is calculated. Validation using experimental data demonstrates a 100% accuracy rate when evaluated against the 15-minute flight delay standard. This model can support airlines in predicting flight delays，enabling targeted optimization of fleet operation processes to enhance operational efficiency.

Keywords：flight delay；prediction model；random forest；data encoding

(责任编辑：唐立平)