数据挖掘

大作业报告

专业班级 软件工程

学生姓名

学　　号

指导老师 刘　彤

山东科技大学

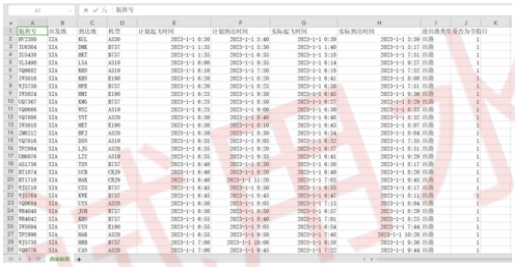
一、任务目的

构建一个多机器学习竞争模型，用以预测航班出港延误率并构建预测平台。

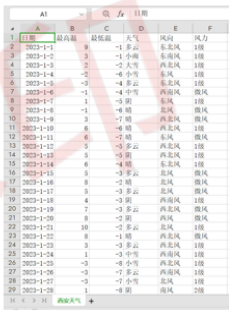
二、数据集

采用爬虫技术对网页数据进行爬取，获取西安市为出发地的航班信息，导出为“西安航班.csv”。从天气网下载西安市2023年1月到6月的天气数据，并导出为″西安天气.csv″文件。

“西安天气.csv”中共包括日期、最高温、最低温、天气、风向、风力等特特征信息，“西安航班.csv”中共含有航班号、出发地、到达地、机型、计划起飞时间、计划到达时间、实际起飞时间、实际到达时间、进出港类型、是否为节假日等特征信息。下图所示为原始数据集展示：



西安航班. csv



西安天气. csv

图 1　原 始 数 据 信 息

在航班数据集中，首先剔除缺失值信息。为了控制研究范围，这里忽略入港数据。原始数据中的出发地/进出港类型将被剔除。考虑时间特征的复杂性，从时间信息中分离出以下五个特征：几月、几号、星期几、第几周、出发时间段。航班是否延误的标签，按照行业标准的15分钟来进行计算后并入数据集中新列。

此时可以通过pandas库将天气数据集合并到航班数据集中并开始下文中的初步特征选择。在进行完初步特征选择后，将预计需要用到的航空公司、到达地、天气三个特征列使用preprocessing.LabelEncoder()转化为数值编码，得到航空公司_num、到达地_num、天气_num三个新特征列，用于后续模型的训练与预测。

保留75%的数据用于训练基模型，剩余数据用于检验模型。

三、任务内容

3.1 数据挖掘

（1）任务总览：

航班延误是民航业的一大难题，提前对航班的延误情况进行预测，以采取合理的应对措施，对缓解航班延误产生的负面影响有着重要意义。

根据以往的研究成果，LightGBM (李任坤,何元清.基于Catboost算法的航班延误预测研究)与CatBoost (丁建立,孙玥.基于LightGBM的航班延误多分类预测)两种基于GBDT的模型在航班延误领域的表现较好。但少有考虑使用集成学习的方式综合两种模型来进行航班离港延误预测的先例，我们考虑使用MLP集成两种模型。

在特征提取上，我们参照先例加入了多种特征选择的方法，从相关系数、机器学习模型本身、数学角度进行特征提取后融合得到特征最终重要性排序。结合数据可视化分析，最终敲定用于模型训练使用的特征。

本项目预测部分首先采取了 CatBoost 与 LightGBM 双模型进行预测，然后使用神经网络对两个模型的预测结果进行融合，得到更为准确的预测结果，如图 2 为模型的整体构建。

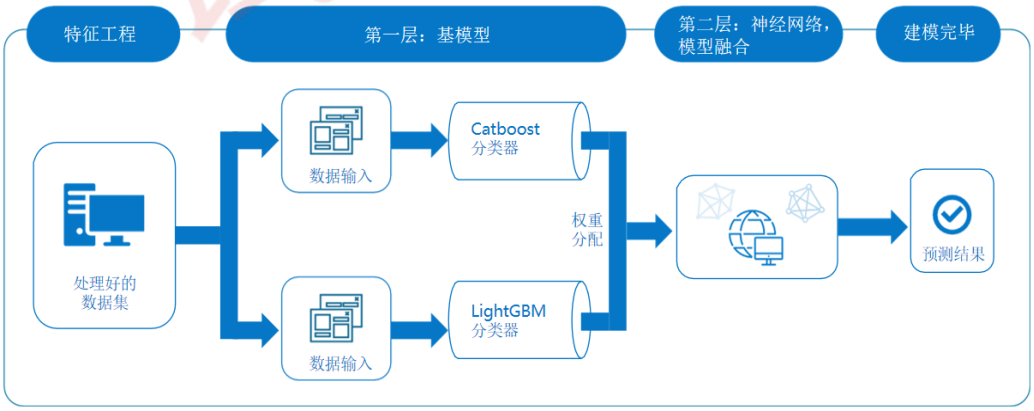


图 2 模型技术架构

在模型建立过程中，为了解决传统分类模型预测精准度低的痛点问题，这里的关键技术是将多个机器学习模型如 LightGBM、CatBoost 等模型进行融合操作，集成为一个性能更好且预测效果更优秀的复合模型。

（3）数据初步可视化与初步特征选择：

①原理：

在进行数据挖掘与数据分析时，选取不合理的特征可能会降低算法的准确度，对预测数据造成干扰，通过合理的数据特征选择使得预测的数据更加准确。本项目先通过数据可视化初步选择特征，再使用多种模型进一步选择与印证特征选取的合理。

②航空公司和延误情况的关系：

根据数据显示，存在某些航空公司经常发生延误，也存在某些公司航班延误较少，例如：根据航班管家 APP 的报告显示，从 2023 年 9 月 11 日到 2023 年 9 月 17 日中山东航空准点率为此期间内最高，准点率为 85.71%。

通过对表 1 的简单可视化分析，可以明确地观察到不同航空公司的延误之间存在明显的差异。这表明航空公司是影响航班延误的一个重要因素，因此需要将航空公司作为选择特征的一项考虑因素。

为了更全面地分析航班延误的影响因素，本项目将在可视化大屏中嵌入所有可视化展示部分，以便更直观地展示航空公司与各因素之间的关系。

表 1 航空公司与航班延误率对应表

平均延误率分档	航空公司代码
[0,0.2]	EG、HJ、IR
(0.2,0.4]	AS、BT、CU、IW、JE、JI、KW 、LJ TP、UH、UP、VP、WJ、WQ、XK、XS、YL、ZU
(0.4,0.6]	DV、FH、FX、GY、HZ、JH、JS、JX、MA NB、NL、SO、SY、TO、TZ、UQ、VI、VQ、WR、XR、ZM
(0.6,0.8]	DM、HF、HT、HV、IG、IU、LX、RN、YT
(0.8,1.0]	C3、EW、GI、ID、IG、IH、JV、KI

③到达地与航班延误的关系：

由于不同路线的拥挤程度和空中管制情况有所差异，因此图 3 中显示的信息完全符合假设，因此可以推测到达地也是影响延误的重要因素，因此到达地也是需要考虑选择的特征之一。

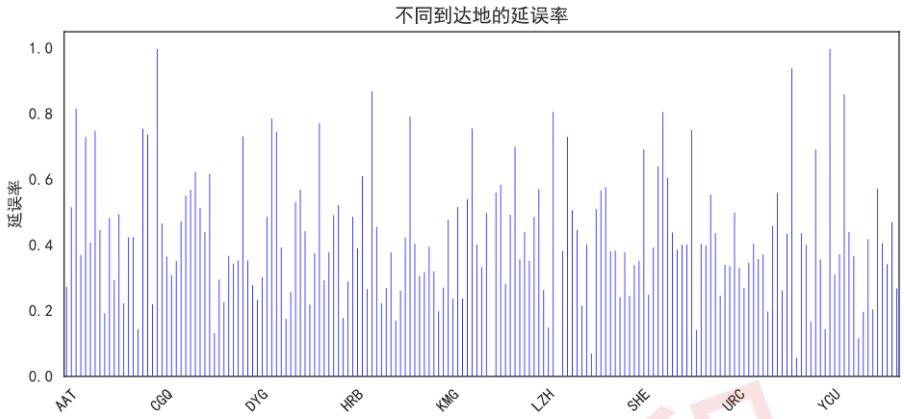


图 3 到达地与航班延误率可视化

④旅行时长与航班延误关系：

对于到达地与航班延误的关系，为了获得更准确的预测结果，本项目希望尽可能利用更多可用的特征，使模型预测的准确率越高的原则，另提出了一个构想，是否存在航班途径的路线距离越远，航班的延误率越高的关系。在图 4 中，通过用旅行时长刻画航班路线跨过的距离。此特征在下文中进行进一步选择。

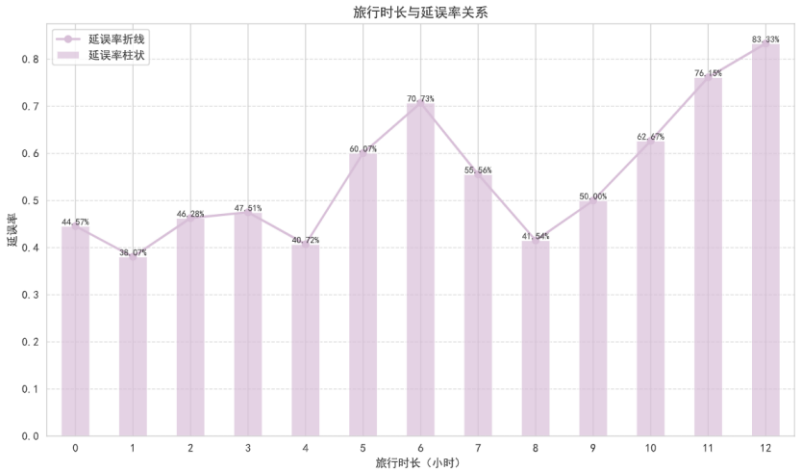


图 4 旅行时长与航班延误率的可视化

⑤飞机型号与航班延误率的关系：

前期猜测飞机型号可能也会影响航班延误率，因此对飞机型号与航班延误率也做一个可视化展示，发现飞机型号对于航班延误率的影响不够显著，因此大概率不用选用飞机型号作为特征之一但此特征在下文中也将进行进一步选择。

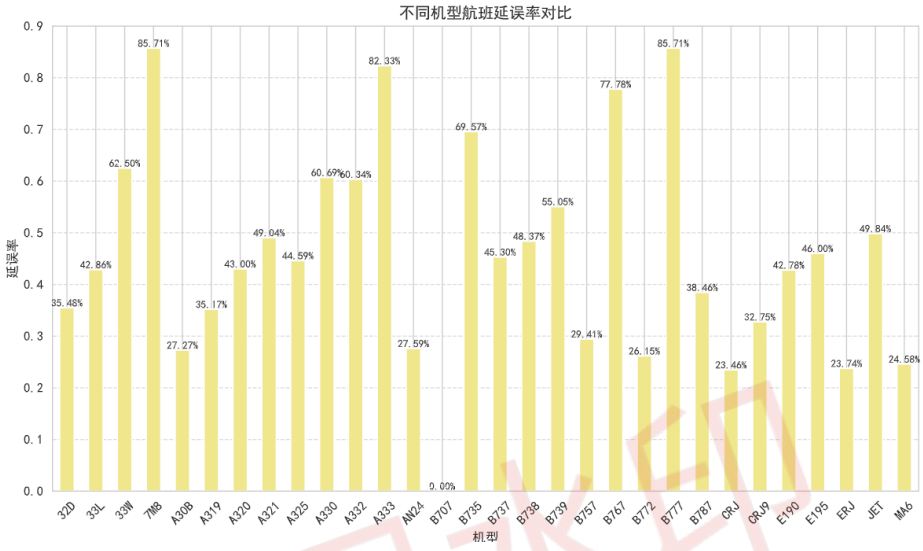


图 5 飞机型号与航班延误率的可视化

⑥出发时间段与航班延误率关系

推测白天、凌晨的航班延误率会有一定影响。从图 6 中可以明显看出，出发时间对于航班延误率具有比较明显影响，因此我们考虑选用出发时间作为航班延误率作为特征之一。

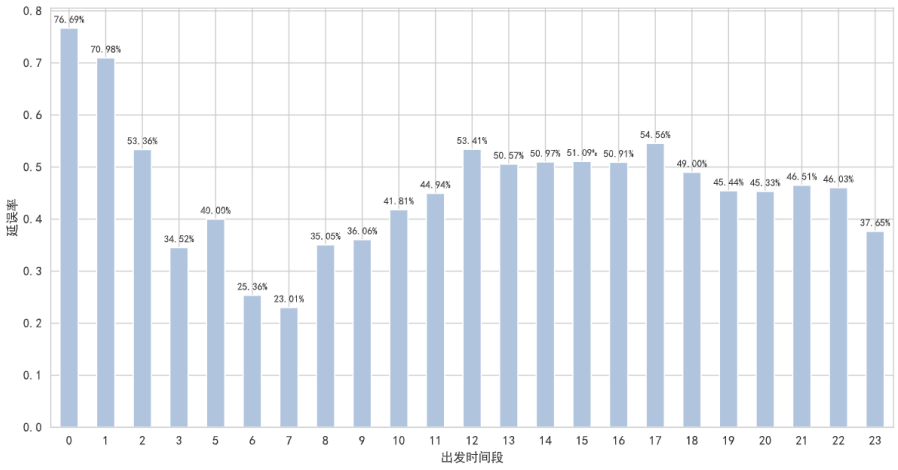


图 6 出发时间与航班延误率的可视化

⑦时间特征对航班延误的关系：

首先通过从计划出发的时间分离出四个特征：月份、日期、周数、星期进行可视化展示：

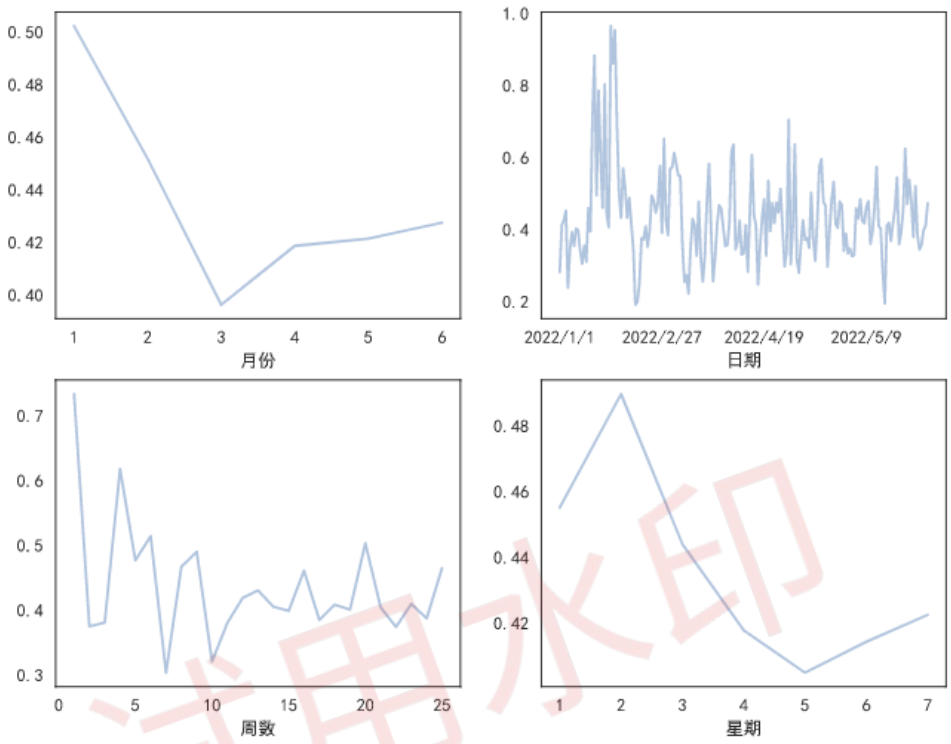


图 7 一些时间特征与航班延误率的可视化

从图 7 可知，1 月份至 2 月份航班延误率最高，由于冬天雨雪天气较多引导致概率较大；其次月初与月末的航班延误率也较高，每月中旬的延误率较低，由于机场的航空调度与社会活动有关导致概率较大，由此可知日期对于航班延误率有一定影响，也作为选用的特征之一；周数对于航班延误率也有较大影响，因此可能需要选用；从一周内来看，每周一到周天的延误率差别不大，故初步决定选择此特征。

⑧天气与航班延误率的关系：

由图 8 可知，在下雪天气中，观察数据飞机延误的概率均达到 80%以上，由此可知天气对于航班延误率有比较大的影响，并且在其它天气下的延误率均保持在 40%左右，因此考虑将此特征作为本项目选定的特征之一。

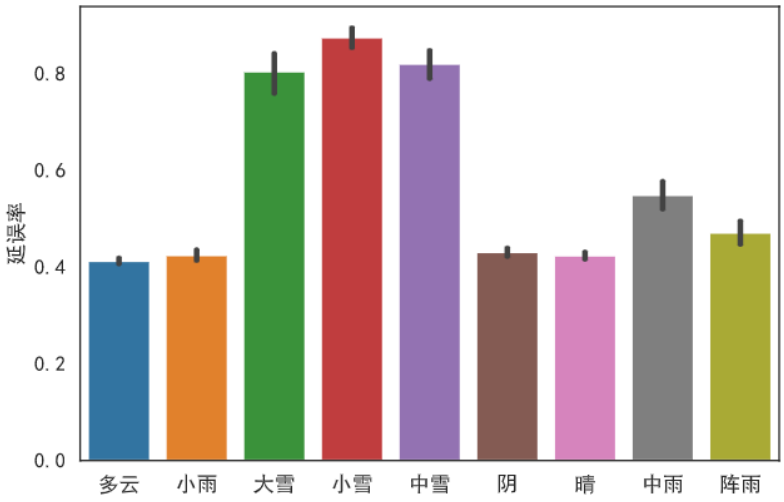


图 8 主要天气与航班延误率的可视化

但与天气有关的数据还存在风力、风向，由图 9 可观察到风力风向与航班延误率的关系基本可以忽略。

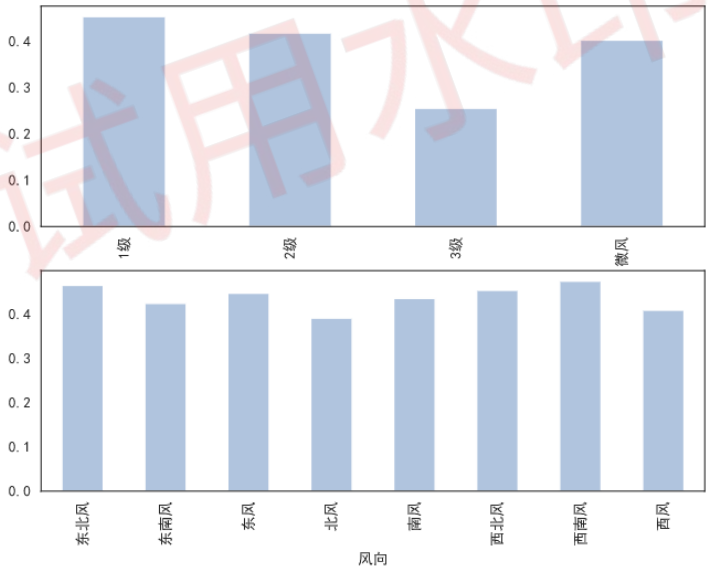


图 9 风力、风向与航班延误率的可视化

⑨节假日对航班延误的关系：

首先挑选出样本数据中节假日日期，接着通过统计并计算得到节假日航班数量占总航班数量的比例、节假日与非节假日的平均延误时长、延误航班处于节假日的比例等数据，最后得出下表。

表 2　节假日与航班延误率的数字可视化

节假日航班数量： 9205
非节假日航班数量： 69516
节假日航班数量占总航班数量： 13.24%
延迟时间超过 15 分钟的航班数量： 39099
节假日平均起飞延迟时间： 155.86 分钟
非节假日平均起飞延迟时间： 150.10 分钟
延迟时间超过 15 分钟的航班中，节假日航班所占比例： 12.14%

由上表数据可得出是否为节假日对航班延误影响不大。

⑩标签数值化：

最终根据以上可视化信息，本项目初步选择航空公司，到达地，出发时间段，日期，天气，第几周暂时作为新的特征作为预测模型的有效特征，并将所有特征再次通过量化的方法在下文中进一步进行特征提取。

（4）进一步提取特征

为了更加准确的提取特征，这里采用三种不同角度的方法进行模型进一步特征提取，分别是相关系数、嵌入法选择、秩和比综合评价法选择。

相关系数可以快速筛选出相关性强的特征，简单直观。嵌入法通过模型本身的学习过程来选择特征。这种方法不仅考虑了特征与目标的关系，还能处理特征间的多重共线性问题。嵌入法能够自动调整特征的权重，从而发现对预测最重要的特征。秩和比法提供了对非线性和不同类型数据的适应性。结合这三种方法，可以全面评估特征的重要性，避免单一方法可能带来的偏差或遗漏。

①基于相关系数的特征选择

Pearson 相关系数和Spearman 相关系数常用的两种相关系数。他们的计算公式分别如下：

$$r=\frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}}\quad,\quad \rho=1-\frac{6\sum d_i^2}{n(n^2-1)}$$

其中n代表样本数量，x、y为两个变量的值，d为观察值的秩差。

■综合相关系数：

首先，本项目分别基于 Pearson 相关系数和 Spearman 相关系数计算特征变量的相关性系数。然后，本项目将两种不同的相关系数都赋予 0.5 的权重，计算出综合相关系数（均保留 6 位小数）。

表 3 相关系数得分

影响因素	综合得分	斯皮尔曼系数	皮尔逊系数
是否为节假日	0.002665	0.002665	0.002665
航空公司	0.092119	0.184237	0.000000
目的地	0.589981	0.219090	0.960871
天气	0.582398	0.370750	0.794046
机型	0.000000	0.000000	0.000000
旅程时长	0.000000	0.000000	0.000000
几月	0.000000	0.000000	0.000000
几号	0.207322	0.220547	0.194097
第几周	0.000000	0.000000	0.000000
星期几	0.000000	0.000000	0.000000
出发时间	0.000000	0.000000	0.000000

最后，由于相关系数的绝对值表示相关性的强度，而与其符号无关，因此本项目对上表中的综合相关系数按照绝对值进行降序排列，又由上表可知只有“航空公司，目的地，天气，几号”四个影响因素所得综合相关系数绝对值均大于 0.01，其他影响因素所得综合相关系数绝对值均小于 0.01，其他影响因素相关性很弱，因此考虑对其他影响因素进行初步剔除。

② 基于嵌入法特征选择

■随机森林特征选择：

随机森林特征选择是一种通过随机森林算法计算每个特征在每个决策树中的重要性，从而确定数据集中最重要的特征的机器学习算法。它可以减少过拟合的风险，

提高模型的准确性，并减少训练时间。通过使用随机森林特征选择，可以确定哪些特征对于预测结果最重要，从而优化机器学习模型。

以随机森林模型为基础的嵌入式特征选择法，输出特征的重要性排序如下表所示：

表 4　随机森林特征选择

影响因素	重要性	排序	影响因素	重要性	排序
目的地	0.179110	1	机型	0.073807	7
出发时间段	0.164441	2	天气	0.057826	8
几号	0.136314	3	旅程时长	0.041510	9
航空公司	0.111117	4	几月	0.040980	10
第几周	0.097254	5	是否为节假日	0.009525	11
星期几	0.088115	6	/	/	/

■XGBoost 特征选择：

XGBoost 是一种基于梯度提升树的机器学习算法，其特征选择功能可以帮助用户自动识别和选择对模型预测最有影响的特征。通过使用 XGBoost 进行特征选择，可以有效地降低模型的复杂性，提高模型的泛化能力，加快模型训练和预测的速度。XGBoost 通过对特征的重要性进行评估，可以帮助用户更好地理解数据，找出最具预测能力的特征，从而提高模型的准确性和效率。通过 XGBoost 的特征选择功能，用户可以更加高效地进行特征工程，提高模型的性能，并且有效地处理高维度数据和噪声特征，使得模型更加鲁棒和可靠。

以 XGBoost 模型为基础的嵌入式特征选择法，输出特征的重要性排序如下表所示：

表 5　XGBoost 特征选择

影响因素	重要性	排序	影响因素	重要性	排序
出发时间段	0.144959	1	几号	0.076549	7
旅程时长	0.115834	2	几月	0.075751	8
第几周	0.114303	3	机型	0.074604	9
目的地	0.109010	4	星期几	0.056614	10
天气	0.091827	5	是否为节假日	0.056008	11
航空公司	0.084540	6	/	/	/

③ 秩和比综合评价法进行特征选择

秩和比综合评价法是一种基于特征排序的特征选择方法，它通过计算每个特征的秩和比值来评估特征的重要性。首先，对每个特征进行排序，然后计算每个特征在各个排序中的秩和，再计算各特征的秩和比值。通过秩和比值的比较，可以确定哪些特征对目标变量的影响最大，从而进行特征选择。秩和比综合评价法能够综合考虑特征在不同排序下的重要性，相对于单一的评价指标，更加全面客观地评估特征的重要性，适用于多特征多维度的数据集。该方法在特征选择过程中能够减少因特征排序不稳定而引起的误差，提高了特征选择的准确性和鲁棒性。

根据秩和比综合评价法得到的特征重要性排序如下表所示：

表 6　秩和比综合评价法进行特征选择

影响因素	重要性	排序	影响因素	重要性	排序
航空公司	3.416753	1	几号	1.672766	7
出发时间段	2.607955	2	第几周	1.618072	8
目的地	2.009769	3	几月	1.500761	9
天气	1.917373	4	星期几	1.498457	10
旅程时长	1.828081	5	是否为节假日	0.363887	11
机型	1.725677	6	/	/	/

④确定选择特征

首先，本项目对通过上述方法得到的数据进行标准化处理，以确保所有特征具有相似的尺度，然后计算其平均值作为该变量的最终得分，并进行降序排列，最后取的得分最高的六个特征作为最终选择特征。

表 7　特征最终得分

影响因素	平均值	排序	影响因素	平均值	排序
目的地	3.416753	1	旅程时长	1.672766	7
出发时间段	2.607955	2	机型	1.618072	8
天气	2.009769	3	星期几	1.500761	9
航空公司	1.917373	4	几月	1.498457	10
几号	1.828081	5	是否为节假日	-1.315735	11
第几周	1.725677	6	/	/	/

结合前面可视化的结果，取上表平均值最高的前六位作为本项目模型训练的最终选择特征，分别为“目的地、出发时间段、天气、航空公司、几号、第几周”。

（5）数据存储

经过以上结果，我们可以得到若干条按照以下形式存储的预测用的数据集：

表 8 预测模型训练数据存储示意表

航班号	航空公司	到达地	时间段	天气	日期	第几周	延误
HV2380	HV	KUL	0	晴	2023/1/1	1	0
IU6364	IU	DMK	1	晴	2023/1/1	1	0
IU5438	IU	HKT	1	晴	2023/1/1	1	1
YL3490	YL	LXA	6	晴	2023/1/1	1	0
VQ9682	VQ	KHN	6	晴	2023/1/1	1	0

（6）CatBoost：

CatBoost 是一种梯度提升决策树（Gradient Boosting Decision Tree）算法。相较于传统的梯度提升决策树算法，CatBoost 在对类别特征的处理上有一些独特的优势。CatBoost 能够自动处理类别特征，无需进行独热编码或手动转换,它使用一种特殊的编码方式，将类别特征转化为数值型特征，从而更好地利用这些特征。同时，CatBoost 通过克服梯度偏差和预测偏移的问题来减少过拟合现象，具有更强的泛化性。

CatBoost 模型通过基于先验值和统计量实现类别的划分，并对数据集进行随机排序，并加入权重系数，以此避免特征维度稀疏。本项目中CatBoost模型将进行600次迭代，树深度设置为7，学习率设定为0.1，随机种子设为18。

对样本随机排序，$\boldsymbol{\sigma} = (\boldsymbol{\sigma_1}, \boldsymbol{\sigma_2}, \ldots\ldots, \boldsymbol{\sigma}_6)$，假设排序后的样本$\boldsymbol{\sigma_p}$的第 k 个维度的特征$\mathbf{x}_{\boldsymbol{\sigma p,k}}$为类别特征，以下是将其转化成数值类特征的公式：

$$\hat{x}_k^i = \frac{\Sigma_{j=1}^{p-1}[X_{\sigma j,k}=X_{\sigma p,k}]Y_{\sigma j}+ap}{\Sigma_{j=1}^{p-1}[X_{\sigma j,k}=X_{\sigma P,k}]+a}$$

其中$\hat{\mathbf{x}}_{\mathbf{k}}^{\mathbf{i}}$是一个目标变量统计，$\mathbf{X}_{\boldsymbol{\sigma p,k}}$是一个类别特征，p 是一个先验值，表示统计样本中延误样本占总样本的比重，$\mathbf{Y}_{\boldsymbol{\sigma j}}$是对特征的标签值，a>0 是一个表示先验值的权重。

CatBoost 算法针对每个样本$\mathbf{x_k}$，对除了它们自己以外的所有数据都训练一个模型$\mathbf{M_k}$，以次来估计样本的梯度，克服梯度偏差。并根据结果对树进行分级，在每次迭代的过程中采用了无偏梯度估计，克服预测偏移，避免了过拟合现象的发生，提高了 catboost 算法的运行效率。

用于评价机器学习中分类模型的常见指标有准确率、召回率以及 F1 分数值、AUC。本项目在这里引入这四个指标对 CatBoost 以及以下用到的模型进行评价。

■准确率：预测结果与实际结果一致的数据条数占样本总条数的百分比。

■召回率：代表实际正样本中被预测为正样本的概率，可以用来反映对少数样本的捕获灵敏度，在我们的项目中，代表对少数延误航班的捕获灵敏度。

■F1 分数：即精确率与准确率的调和平均数，可以比较全面的客观的反应模型的综合性能，最小值为 0，越接近 1 可以说明该分类模型的性能越好。

■AUC：ROC 曲线下的面积，也可以用来用来衡量分类模型的性能。（下面详细说明）

首先定义 TP、TN、FP、FN 分别表示正样本被判断为正样本，负样本被判断为负样本，负样本被判断为正样本，正样本被判断为负样本。

准确率的公式可以表示为：

$$\mathbf{A} = \frac{\mathbf{TP} + \mathbf{TN}}{\mathbf{TP} + \mathbf{TN} + \mathbf{FP} + \mathbf{FN}}$$

下面的计算还可能用到一个概念——精确率，精确率是指被所有预测为正的样本中实际为正样本的概率，此指标可以反映出正样本结果中的预测准确程度；

精确率的公式可以表示为：

$$\mathbf{P} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}}$$

由于实际样本并不可能是完全均衡的，因此不能只参照准确率和精确率进行评估模型的性能，假设正样本的数目远小于负样本的数目，若某模型给出的预测结果全部为负样本数，那么准确率与精确率的参考价值就不大，此时本项目需要用到召回率进行评估模型的好坏。

召回率的公式可以表示为：

$$\mathbf{R} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}}$$

F1 分数的公式可以表示为：

$$\mathbf{F_1} = \frac{\mathbf{2} \cdot \mathbf{P} \cdot \mathbf{R}}{\mathbf{P} + \mathbf{R}}$$

AUC（Area Under the ROC Curve）是指 ROC 曲线下的面积，用来衡量分类模型的性能。在二分类问题中，通过改变分类的阈值，可以得到一组不同的 True Positive Rate（TPR）和 False Positive Rate（FPR）的值，这些值用来绘制 ROC 曲线。AUC 越接近 1，表示模型的性能越好；AUC 越接近 0.5，则表示模型的性能越差。计算 AUC 通常使用 TPR 和 FPR。

TPR 的计算公式可表示为：

$$\mathbf{TPR} = \mathbf{R}$$

FPR 的计算公式可表示为：

$$\mathbf{FPR} = \frac{\mathbf{FP}}{\mathbf{FP} + \mathbf{TN}}$$

通过计算 TPR 和 FPR，可以得到一组点，然后根据这些点绘制 ROC 曲线，并计算 ROC 曲线下的面积（AUC）。如下图为计算 AUC 的流程图：

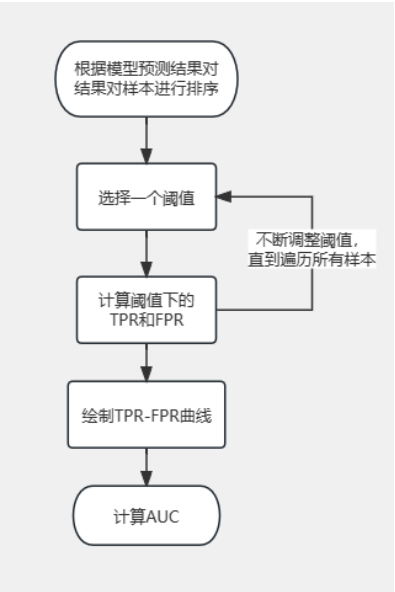


图 10 计算 AUC 流程图

对预测结果的以上指标进行输出得到下表：

表 9　CatBoost 模型的评价指标

准确率	召回率	F1 分数	AUC
71.11%	58.01%	63.88%	69.71%

（7）LightGBM：

LightGBM 是一种基于梯度提升决策树（Gradient Boosting Decision Trees）的机器学习算法。LightGBM 综合采用基于梯度单边采样（(Gradient-based One-Side Sampling, GOSS），互斥特征捆绑（(Exclusive Feature Bundling, EFB）和直方图算法等方法来提高算法整体的运行效率。

本项目使用的LightGBM模型的迭代次数为270，最大树深度设置为7，叶子节点数为20，最小子样本数为10，L1正则化系数为0.8，L2正则化系数为0.2，最小子样本权重设为7，帮助进一步降低过拟合风险。特征选择方面，每棵树使用70%的特征，并且在训练过程中采用60%的样本，学习率设置为0.71。

若采用浮点型特征（最终保留的特征中无浮点型数据），LightGBM 可以采用直方图算法确定最优分类点，直方图算法将连续的浮点特征值转化成分段函数，再将每一个分段函数分别映射成 m 个整数（箱），其中每个整数对应多个分段函数，从而将连续值转化成离散值，这方便我们后续改进模型添加更加复杂的选择特征。

在遍历数据时，根据特征值所对应的离散值进行梯度累积和样本个数的统计，最后遍历直方图的离散值寻找最优分类点，为了找到最优分类点，需要计算增益：

$$\mathbf{\Delta loss} = \frac{\mathbf{S_L^2}}{\mathbf{n_L}} + \frac{\mathbf{S_R^2}}{\mathbf{n_R}} - \frac{\mathbf{S_p^2}}{\mathbf{n_p}}$$

其中 $\mathbf{\Delta loss}$为增益值，$\mathbf{S_L}$为当前箱左侧的梯度之和，$\mathbf{S_R}$为当前箱右侧的所有箱的梯度之和，$\mathbf{S_P}$ 为父结点的梯度之和，$\mathbf{n_L}$，$\mathbf{n_R}$，$\mathbf{n_p}$分别为对应的样本数量，当最大增益最大时，其所对应的特征和箱为最优分类点。

采用基于梯度的单边采样算法减少训练样本数。因为较小梯度的样本意味着该样本已经得到较好的训练，所以可以丢弃部分梯度较小的样本来实现减少训练样本数的

目的。GOSS 算法首先对样本梯度的绝对值进行降序排序，然后选取前 a%的梯度较大的样本，记为样本 A。然后在剩余的 1-a%的梯度绝对值较小的样本中，随机选取 b%的样本（其中 b<100-a），记为样本 B。但这样会改变样本的分布，可以放大 B 中样本的梯度（1-a）/b 倍，保证该算法充分关注梯度较小的样本点。设 O 为原训练集，$\mathbf{g_i}$指梯度值，$\mathbf{n_0}$表示未分裂的叶子结点的样本总数，$\mathbf{n^j_{l|O}(u)}$和$\mathbf{n^j_{r|O}(u)}$分别表示分裂后的左叶子结点的样本总数和右叶子结点的样本总数，在 GOSS 算法中，第 j 个特征，值为 u 处进行分裂的增$\mathbf{V_{GOSS}(u)}$为：

$$\mathbf{V_{GOSS}(u)} = \frac{1}{n_0}\left(\frac{\left(G_{Al}+\frac{1-a}{b}G_{Bl}\right)^2}{n_l^j(u)} + \frac{\left(G_{Ar}+\frac{1-a}{b}G_{Br}\right)^2}{n_r^j(u)}\right)$$

其中，$\mathbf{A_l}$为左叶子结点中包含样本 A 的集合，$\mathbf{B_l}$，$\mathbf{A_r}$，$\mathbf{B_r}$同理，记$\mathbf{\Sigma_{xi\epsilon A_l} g_i}$为$\mathbf{G_{Al}}$，$\mathbf{G_{Bl}}$，$\mathbf{G_{Ar}}$，$\mathbf{G_{Br}}$同理，$\mathbf{n_l^j(u)}$为左叶子结点数，$\mathbf{n_r^j(u)}$为右叶子结点数。GOSS 算法提高了模型准确性和训练效率。

采用互斥特征捆绑算法用来减少训练的特征，其关键在于确定哪些特征可以合并，确定是否可以合并的依据是判断原有的不同的特征值在合并后的特征中是否能够被识别出来，如果可以被识别则原特征可以合并，反之则不可以。针对如何合并问题，可以采用贪心算法构造带权图，并对特征的度降序排列，并对排好的特征进行遍历。

此外，LightGBM 采用的是 Leaf-wise 生长策略，该策略每次从所有的叶结点找最优分类点进行分裂。相对于传统方法，该策略的缺点在于生成的树的深度很大，进而产生过拟合现象，所以 LightGBM 一般会设置最大深度进行限制，避免过拟合现象，但优点在于可以降低内存损失，提高算法的效率。

如下图，通过 LightGBM 进行预测的流程图：

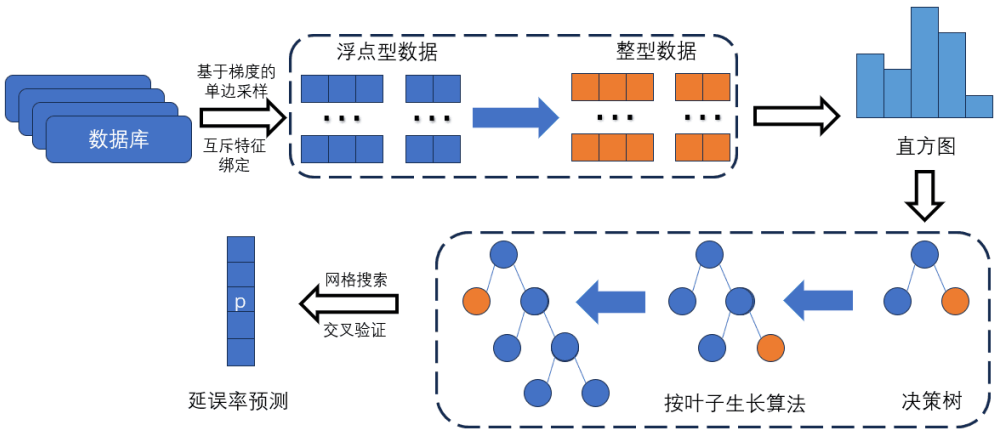


图 11 lightGBM 应用于航班延误率预测的模型结构

本项目通过访问数据预处理好的数据，向 LightGBM 模型输入航空公司、到达地、天气、航班延误率等数据进行训练。同以上的 CatBoost 模型，对预测结果的以上指标进行输出得到下表：

表 10　lightGBM 模型的评价指标

准确率	召回率	F1 分数	AUC
70.70%	60.79%	64.43%	69.64%

（8）神经网络融合模型

从以上两个模型的输出结果可以看出准确率均为 70%左右， 然后将两个分类器预测结果与真实的标签值整合，将分类器结果作为神经网络输入值，将真正的标签值作为训练神经网络过程中的标签值，采用神经网络分配权重。

神经网络模型是一种受到生物神经网络启发的人工智能算法。在神经网络模型中，输入层负责将输入向量传递给神经网络，隐藏层对输入进行多次变换以提高结果的准确性，输出层返回最终的输出结果，而训练过程主要通过反向传播算法来实现。该算法通过计算输出结果和实际结果之间的误差，并反向传播到网络中的每个神经元来更新权重和偏置。通过不断迭代和调整权重和偏置，神经网络可以逐渐提高输出结果的准确性。如下图，神经网络模型的结构图示意：

在本项目搭建的神经网络模型中，输入对应 CatBoost 与 LightGBM 模型预测的概率值。对于每一个样本，将真正的标签值作为训练神经网络过程中的标签值，假设 CatBoost 与 LightGBM 的预测概率值分别为$\mathbf{x_1}$、$\mathbf{x_2}$。则神经网络的结构如下图所示：

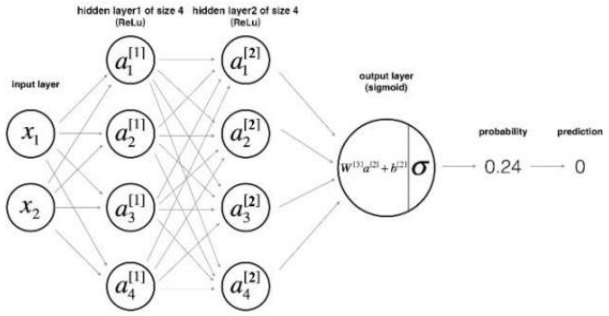


图 12 神经网络结构模型图

从最后一层输出层到第一层隐藏层的逐层反向传播公式如下所示：

$$
\begin{cases}
dA^{[3]} = \dfrac{1-1}{1-A^{[3]}} - \dfrac{1}{A^{[3]}} \\
dZ^{[3]} = dA^{[3]} * \dfrac{1}{1+e^{-Z^{[3]}}} * \left(1 - \dfrac{1}{1+e^{-Z^{[3]}}}\right) \\
dW^{[3]} = \dfrac{1}{N} dZ^{[3]} \cdot A^{[2]} \\
db^{[3]} = \dfrac{1}{N} \sum_{i=1}^{N} dZ^{[3](i)} \\
dA^{[2]} = W^{[3]^T} \cdot dZ^{[3]} \\
dZ_i^{[2]} = \begin{cases} dA_i^{[2]} & Z_i^{[2]} > 0 \\ 0 & Z_i^{[2]} \le 0 \end{cases} (i = 1, 2 \ldots\ldots, n_2) \\
dW^{[2]} = \dfrac{1}{N} dZ^{[2]} \cdot A^{[1]} \\
db^{[2]} = \dfrac{1}{N} \sum_{i=1}^{N} dZ^{[2](i)} \\
dA^{[1]} = W^{[2]^T} \cdot dZ^{[2]} \\
dZ_i^{[1]} = \begin{cases} dA_i^{[1]} & Z_i^{[1]} > 0 \\ 0 & Z_i^{[1]} \le 0 \end{cases} (i = 1, 2 \ldots\ldots, n_1) \\
dW^{[1]} = \dfrac{1}{N} dZ^{[1]} \cdot X \\
db^{[1]} = \dfrac{1}{N} \sum_{i=1}^{N} dZ^{[])(i)} \left(\dfrac{1}{1+e^{-Z^{[3]}}}\right)
\end{cases}
$$

隐藏层中在使用激活函数后，将上一层的输出加权求和，然后用到 ReLU 函数来进行非线性转换。ReLU 函数是一种常用激活函数，ReLU 函数公式如下：

$$\mathbf{ReLU(x)} = \begin{cases} \mathbf{0} & \mathbf{,x < 0} \\ \mathbf{x,} & \mathbf{x \geq 0} \end{cases}$$

其中，输出层使用 sigmoid 函数作为激活函数，sigmoid 函数也是一种常见的激活函数，sigmoid 函数的公式如下：

$$\mathbf{sig(x)} = \frac{\mathbf{1}}{\mathbf{1 + e^{-x}}}$$

在 Python 编译模型时，本项目选用的损失函数是二值交叉熵函数。二值交叉熵函数是一种常用的损失函数，主要应用于二分类问题中。它衡量了观测值与预测值之间的差异，并用于评估和优化二分类模型的性能。二值交叉熵函数公式如下：

$$\mathbf{LOSS} = -\frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{i=1}}^{\mathbf{N}} \mathbf{y_i \log(p(y_i)) + (1 - y_i)\log(1 - p(y_i))}$$

其中 N 表示样本数量，$\mathbf{y_i}$={0,1}是二元标签，p(y)=1 是输出属于 y 标签的概率。

二值交叉熵函数通过比较真实标签和模型预测概率的差异来衡量模型的性能。当真实标签为 1 时，函数关注预测为 1 的概率；当真实标签为 0 时，函数关注预测为 0 的概率。这种差异度量可以有效地指导模型进行分类预测的优化。二元交叉熵函数是一种负对数似然损失函数，其形式来源于最大似然估计的推导。通过最小化二元交叉熵函数，可以使得模型的预测概率尽可能接近真实的标签分布。二元交叉熵函数是连续且可导的，对于梯度下降等优化算法提供了良好的优化性质。

在该神经网络中，还采用了 Adam 优化器（Adaptive Moment Estimation）。这是一种基于梯度的优化算法，常用来调整神经网络的参数，以最小化训练过程中的损失函数。Adam 算法可以自适应地为不同参数计算并且应用学习率，并且考虑了梯度本身和参数的二阶矩估计（即梯度的平方的均值），相对于其他优化算法，Adam 优化器具有较好的性能和收敛速度，特别适用于大规模数据和复杂网络结构的训练任务。

以下为通过神经网络分配权重后的融合模型预测的评价指标结果：

表 11 组合模型的评价指标

准确率	召回率	F1 分数	AUC
71.37%	71.32%	63.99%	71.36%

通过观察融合模型的相关指标可以发现准确率、召回率、F1 分数、AUC 等相关指标综合来看有较大进步，特别是召回率，说明模型捕获延误航班的**能力大幅增强，**这体现了本项目预测分析系统所建模型的适用性；又因为 CatBoost 与 LightGBM 模型本身的性能比较优秀，本航班延误率预测分析系统还在一定程度上保证了软件响应的效率，为后续添加实时性的预测分析模块提供了基础。

3.2 工程构建

利用模型与本专业学习到的软件开发方面知识，构建航班延误预测平台，开发为java，框架选用SpringBoot与Vue，IDEA版本为2023.2.2，数据库采用MySQL数据库存储。以下为界面展示：

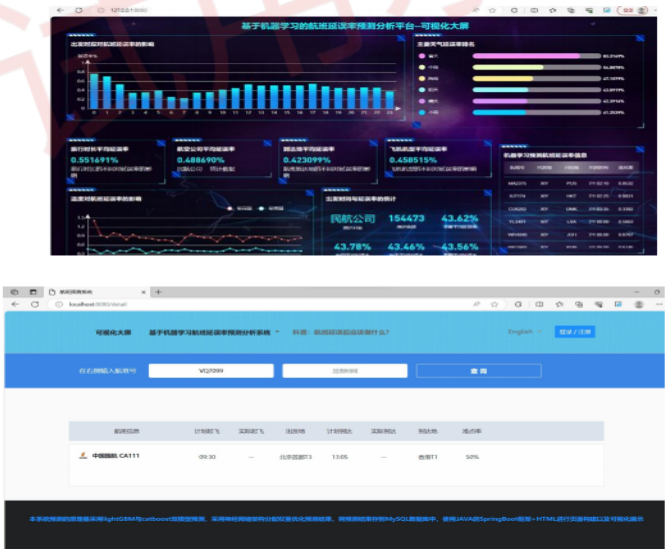


图12 界面展示

参考文献

[1] 陈云辉. 机器学习分类算法对航班延误数据的实证分析[D]. 兰州大学，2023.

[2] 张星园,孟怡辰,罗建龙,李言. 于 CatBoost 的玻璃成分与风化关系实证[J],玻璃,2023,No.8.

[3] 秦婉怡. 基于 CatBoost 算法的信用卡用户信用风险预测模型应用研究[D]. 重庆工商大学，2021.

[4] 李任坤, 何元清.基于 Catboost 算法的航班延误预测研究[J]. 现代计算机, 2022,V.28; No.3.

[5] 王鹏新,王 颖, 田惠仁,王 婕,刘峻明,权文婷. 基于 LightGBM 的冬小麦产量估测及可解释性研究 [J/OL] . 农业机械学报 , https://link.cnki.net/urlid/11.19 64.S.20231012.0846.002.

[6] 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,V.53; No.6.

[7] 晋百川,杨鸿波,侯 霞,罗 杰,胡大胆. 面向集成学习的航班离港延误状态预测 [J] . 传感器与微系统. 2023,V. 42; No.9.