参赛队号：

2021年（第七届）全国大学生统计建模大赛

参赛学校：	天津商业大学
论文题目：	基于集成学习的共享单车异常检测的研究
参赛队员：	徐晨恒 李姗姗 崔琳芳
指导老师：	刘高生 马云鹏

目 录

基于集成学习的共享单车异常检测的研究

摘 要

近年来，共享单车的快速发展在方便了人们出行的同时，也对城市交通产生了一定的负面影响，其主要原因为单车资源配置的不合理。本文通过建立单车租赁数量的预测模型和异常检测模型，以期能够帮助城市合理配置资源。

首先，进行探索性数据分析。主要步骤为数据预处理、描述性统计和回归分析。其中，分位数回归能够表现出输入变量与输出变量各分位点间的线性关系。

其次，建立单车预测模型。分别运用集成学习中的 Bagging、Boosting 和模型融合算法 Stacking 进行建模。实验结果显示，Boosting 算法中的 CatBoost 模型对单车租赁数量的预测效果最好。

最后，建立异常检测模型。运用孤立森林算法检测单车租赁数量的异常值，并利用支持向量机分析各输入变量对租赁异常的影响程度。研究表明，租赁异常可能与城市意外事件的发生、节假日的到来、温度与风速以及湿度的突变和极端恶劣天气的产生有关。

预测模型能够帮助城市合理规划共享单车的投放数量，而异常检测模型则有助于城市及时处理突发事件，希望本文的研究能够为城市资源合理配置提供参考。

关键词：共享单车 分位数回归 集成学习 模型融合 异常检测

Abstract

In recent years, the rapid development of shared bicycles has not only facilitated people's travel, but also has a certain negative impact on urban traffic. The main reason is the unreasonable allocation of bicycle resources. This paper establishes a prediction model and anomaly detection model for the number of bicycle rentals in order to help the city allocate resources rationally.

First, conduct exploratory data analysis. The main steps are data preprocessing, descriptive statistics and regression analysis. Among them, quantile regression can show the linear relationship between the input variables and the quantile points of the output variables.

Second, establish a bicycle prediction model. Use Bagging, Boosting and model fusion algorithm Stacking in ensemble learning to model. The experimental results show that the CatBoost model in the Boosting algorithm has the best predictive effect on the number of bicycle rentals.

Finally, establish an anomaly detection model. The isolated forest algorithm is used to detect the abnormal value of the number of bicycle rentals, and the support vector machine is used to analyze the impact of each input variable on the rental abnormality. Studies have shown that rental abnormalities may be related to the occurrence of urban accidents, the arrival of holidays, sudden changes in temperature and wind speed and humidity, and the occurrence of extreme weather.

Predictive models can help cities rationally plan the number of shared bicycles, while anomaly detection models can help cities deal with emergencies in a timely manner. It is hoped that the research in this article can provide a reference for the rational allocation of urban resources.

Keywords: bike sharing, quantile regression, ensemble learning, model fusion, anomaly detection

1 引言

1.1 研究背景

1.1.1 历史沿革

1995 年，世界第一辆共享单车于哥本哈根诞生，那时已引入押金概念，且设置固定停车点；2007 年，世界第一家共享单车公司于巴黎成立。从国外共享单车产业的发展来看，发源地哥本哈根有着完善的共享单车系统与人性化的配套设施。对于我国来说，2012 年，北京和上海等一线城市才开始提供公用单车；我国共享单车产业真正诞生于 2016 年，相比于国外起步较晚，有一定国外经验可供借鉴。

我国共享单车的发展可以分为三个阶段。2007 年至 2010 年为第一阶段，共享单车主要由政府引入，实行因地制宜的管理模式；2010 年至 2014 年为第二阶段，共享单车主要由经营性公司提供；2014 年至今为第三阶段，共享单车受共享经济和互联网经济的共同作用，无桩单车取代了有桩单车[1]。目前，我国共享单车的现状为有桩单车与无桩单车共存[1]。

由于巨大市场空间提供的有力支撑、雄厚投资力量引发的催生效应和我国一系列鼓励共享经济政策都为共享单车的发展奠定了基础；而随着“绿水青山就是金山银山”口号的提出，人们的环保意识不断增强，提倡低碳的出行方式也进一步促进了共享单车产业的发展[2]。

1.1.2 问题提出

然而，随着共享单车的发展，各种问题也纷至沓来，共享单车资源配置不合理以及城市单车过度投放问题已成为社会的焦点话题。人民日报显示，2017 年，上海无桩共享单车总量为 45 万辆，有桩共享单车总量为 8 万辆，城市对单车的承载力已近饱和，随处可见大量闲置单车，资源浪费严重，“低碳出行”已变得

不在低碳[3]；此外，已经投放过一段时间的旧车，其损坏率较高，且并未及时维修，也在一定程度上造成了资源浪费。共享单车盲目投放使得不少城市的单车数量过饱和，严重影响了人们的日常出行和城市的交通管理。

1.2 研究意义

共享单车已成为人们日常出行的主流方式之一，本文通过对共享单车租赁数量影响因素及其异常情况的研究，以期能够为解决城市资源配置问题提供帮助。

从经济层面，共享单车企业作为单车经营的主体。企业可以利用共享单车数据进行精细化管理与智慧化运营，带动企业发展，也可以进行智能监控，提升用户体验。此外，共享单车数据可以应用于智慧产业，有利于智慧产业的发展[3]。

从社会层面，利用共享单车数据与全市交通信息建立智慧交通[3]。疫情期间，传统的城市交通无法解决出行困难，通过大数据可以保证交通的正常运行。同时，深度挖掘共享单车数据，可以有效治理单车乱摆乱放等问题。

从民生方面，依靠数据挖掘技术，政府可以准确把握共享单车的动态，及时缓解交通压力，为人们出行提供保障[4]。

1.3 研究方法

首先，本文对该共享单车数据集进行了探索性数据分析。探索性数据分析用于解释原始数据，并挖掘数据的潜在规律[5]。第一步，进行数据预处理，即对数值特征的归一化处理和对类别特征的哑变量处理[5]。该预处理方法有利于数据建模分析[5]。第二步，进行描述性统计，本文绘制了小提琴图和饼状图，借以表现各特征变量的概率分布情况[9]。第三步，进行回归分析，其中，相比于线性回归，分位数回归能够表现出解释变量与被解释变量各分位点之间的线性关系，其解释数据效果更好[8]。

其次，本文对该数据集进行了集成学习研究。集成学习是指将若干个弱学习器通过一定的策略组合得到一个强学习器，其基本分类为 Bagging、Boosting 和 Stacking。第一步，运用随机森林、XGBoost、LightBoost 和 CatBoost 四种模型分别对该数据集进行训练，并探究其特征变量重要性。随机森林是集成决策树的 Bagging 算法，其学习结果由其决策树的投票产生[6]。XGBoost、LightGBM 和 CatBoost 均属于 Boosting 算法，其中，XGBoost 对损失函数进行改进，并利用正则化减少过拟合，提高了模型的泛化能力[13]。LightGBM 支持并行化学习，在处理多维问题时其计算效率更高[14]；CatBoost 在处理类别特征问题上进行了优化，其模型精度往往比 XGBoost 和 LightGBM 更高[15]。第二步，运用 Stacking 方法将上述四种算法进行模型融合，以期得到一个泛化能力更好的模型。Stacking 是模型融合的学习框架，其核心思想是将不同模型的优点进行有机结合，从而提高模型的泛化能力[11]。

最后，本文以上述模型的训练误差为样本对单车租赁数量进行异常检测研究。第一步，运用孤立森林模型检测样本中的异常点，以期通过误差异常来反映共享单车租赁数量异常。孤立森林是一种异常检测方法，可以精准识别分布稀疏的独立离群点。第二步，运用支持向量机对上述异常情况进行训练，支持向量机非常擅长分类及回归问题，以期通过其向量空间特征系数来反映各变量对单车租赁数量异常的影响程度[16]。

集成学习所建立的预测模型能够帮助城市合理规划共享单车的投放数量，而孤立森林和支持向量机所建立的异常检测模型则有助于城市及时处理突发事件。本文研究方法的流程图如下图 1 所示：

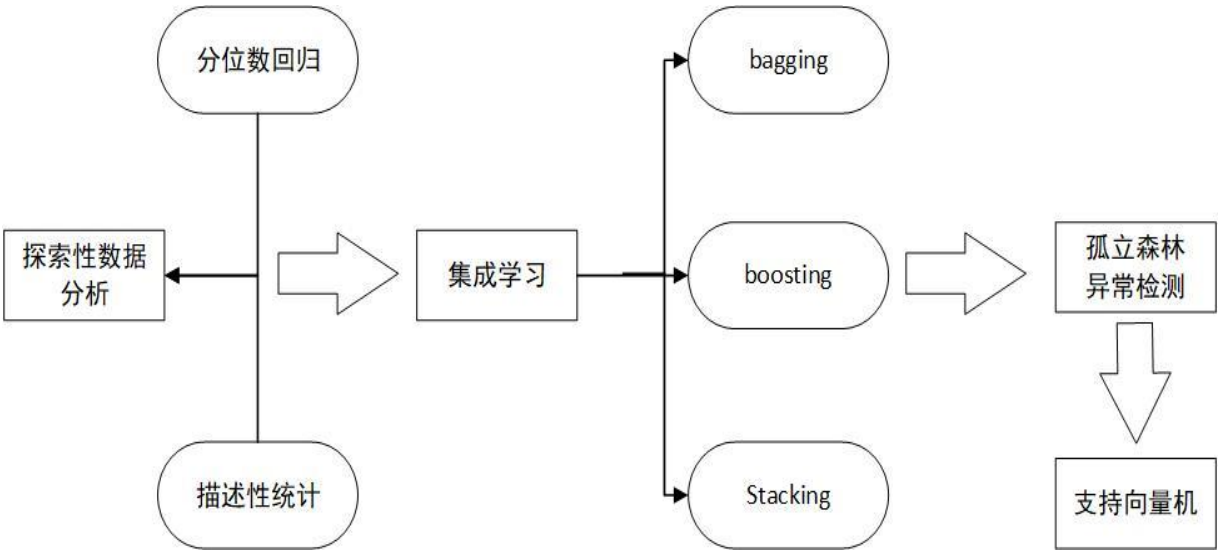

图 1 研究方法流程

其中，探索性数据分用于解释数据，集成学习用于构造预测模型，孤立森林和支持向量机用于构造检测模型。

2 探索性数据分析

2.1 数据来源

本文研究数据由波尔图大学人工智能与决策支持实验室（LIAAD）提供。

2.2 数据预处理

对于该数据集中的数值型变量，即数值特征，本文对其进行数据归一化，即统一映射到$[0,1]$区间上[10]；数据无量纲化有利于提升机器学习模型的训练精度和收敛速度，其公式如下：

$$x' = \frac{x-min}{max-min}, \tag{1}$$

其中，x代表原始数据，x'代表归一化后的数据，max代表原始数据中最大值，min代表原始数据中最小值[10]。

数值型变量的代表符号如下表 1 所示：

表 1 数值型变量

代表符号	变量名称
Y	单车租赁数量
X_1	时间
X_2	温度
X_3	湿度
X_4	风速

对于该数据集中的分类型变量，即类别特征，本文采用哑变量的处理方式，将其统一变换为 0 或 1 变量。

分类型变量的代表符号如下表 2 所示：

表 2 分类型变量

代表符号	变量名称
X_5	节假日
X_6	季节
X_{61}	春季
X_{62}	夏季
X_{63}	秋季
X_{64}	冬季
X_{71}	工作日
X_{72}	周末
X_8	天气
X_{81}	晴天
X_{82}	多云
X_{83}	雨雪

本文探索性数据分析和集成学习所用数据均为上述预处理数据。

2.3 描述性统计

探索性数据分析（Exploratory Data Analysis，EDA）用于解释原始数据，探索数据之间的潜在规律[9]。EDA 在数据科学工作过程中，能够对多个环节产生影响，是不可或缺的重要步骤[10]。本文采用描述性统计和分位数回归两种方法进行 EDA。

小提琴图是箱线图与密度图的结合，可以同时反映出变量的概率密度及分布情况，其中，箱线图的信息在中间部分，密度图的信息在两侧部分。本文主要运

用小提琴图对数值型变量进行了 EDA，其结果如下图 2 所示：

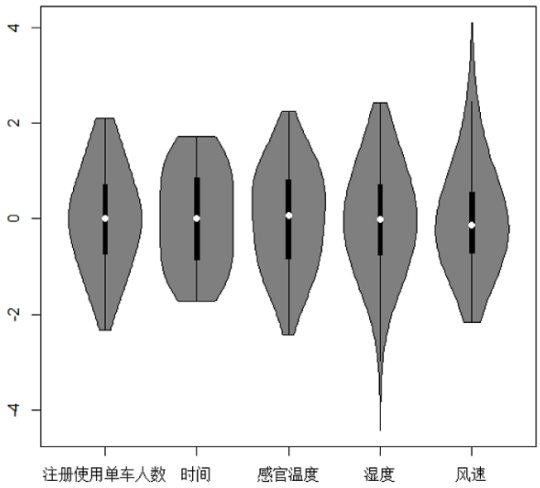

图 2 数值特征小提琴图

从图 2 中可以看出，单车租赁数量与时间和温度这两个变量主要集中分布在中间区域，各数据间的取值无明显差异；而湿度与风速这两个变量存在极值点，可初步判断这两个变量与单车租赁数量有较大相关性。

对于该数据集中的分类型变量，本文主要运用饼状图进行 EDA，其结果如下图 3 所示：

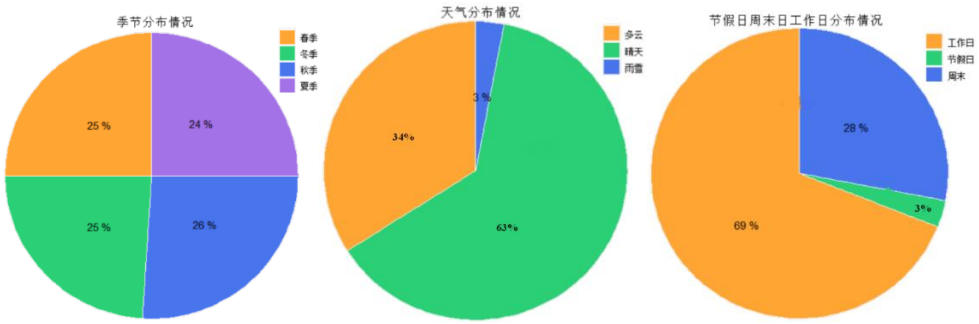

图 3 类别特征饼状图

从季节变量饼状图中可以看出，春夏秋冬四季约各为总体的四分之一，说明该数据集的季节变量分布较为均匀；从天气变量饼状图中可以看出，晴天出现的次数最多，约为总体的 63%，而雨雪等极端天气出现的次数最少，仅为 3%；除此之外，工作日所占比重最大，约为总体的 69%，而节假日所占比重最少，仅 3%。

2.4 分位数回归

2.4.1 基本原理

分位数回归（Quantile Regression，QR）用于解释自变量X与因变量 Y不同分位点之间的线性关系[8]，因此，其解释数据的能力更强。

假设存在样本序列$\{(X_i, Y_i), (i = 1, \dots, n)\}$满足下列回归模型：

$$Y = g(X) + \varepsilon \quad X \in R. \tag{2}$$

假设误差项$\varepsilon_i\{i = 1, \dots, n\}$为独立同分布，且其分布情况未知，设自变量$X$的分布函数为[8]：

$$F(x) = \mathrm{p(X \le x)}. \tag{3}$$

因变量Y的τ阶条件分位数$g_\tau(x)$满足$\tau = P[Y \le g_\tau(X) \mid X = x]$，其公式如下：

$$g_\tau(x) = \mathrm{argmin}_{\theta \in R}\, E\{\rho_\tau(Y - \theta) \mid X = x\}, \tag{4}$$

其中，$\rho_\tau(u) = u[\tau I(u \ge 0) - (1 - \tau) I(u < 0)]$为检验函数，$I()$为示性函数，而不包含示性函数的损失函数为：

$$\rho_\tau(\mu) = \begin{cases} \tau\mu, \mu \ge 0, \\ (\tau - 1)\mu, \mu < 0. \end{cases} \tag{5}$$

$E\rho_\tau(x - \hat{X})$越小分位数回归效果越好，其公式如下：

$$E\rho_\tau(x - \hat{X}) = (\tau - 1)\int_{-\infty}^{\hat{X}} (x - \hat{X}) dF(x) + \tau \int_{\hat{X}}^{\infty} (x - \hat{X}) dF(x) = 0, \tag{6}$$

其中，当$E\rho_\tau(x - \hat{X})$为 0 时可以得到最优解，即：

$$0 = (1 - \tau)\int_{-\infty}^{\hat{X}} dF(x) - \tau \int_{\hat{X}}^{\infty} dF(x) = F(\hat{X}) - \tau, \tag{7}$$

用样本经验分布函数代替$F(\mathrm{x})$，则：

$$F(x) = \frac{\sum_{i=1}^{n} I(X_i \le x)}{n}, \tag{8}$$

$$\int \rho_\tau(x - \hat{X}) dF_n(x) = n^{-1} \sum_{i=1}^{n} \rho_\tau\left(x_i - \hat{X}\right). \tag{9}$$

2.4.2 仿真实验

为了方便进行回归分析，本文对季节和天气两个类别特征进行编码化处理，其他数据仍为预处理数据，并进行了线性回归分析，其结果如下图 4 所示：

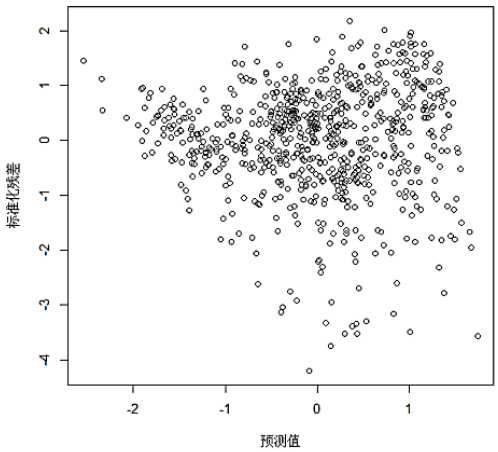

图 4 线性回归残差图

由图 4 可观察出，利用线性回归进行分析效果并不理想，因此，本文接下来运用分位数回归进行分析，其回归模型如下：

$$q(\mathrm{Y}) = \&\beta_{\alpha 0} + \beta_{\alpha 1}X_1 + \beta_{\alpha 2}X_2 + \beta_{\alpha 3}X_3 + \beta_{\alpha 4}X_4 + \beta_{\alpha 5}X_5 + \beta_{\alpha 6}X_6 + \beta_{\alpha 7}X_7 + \beta_{\alpha 8}X_8 + \varepsilon_{\alpha} \quad (10)$$

本文选取五个分位点，即 0.05、0.25、0.5、0.75 和 0.95，分别进行分位数回归，其结果如下表 3 所示：

表 3 分位点回归系数

特征变量	0.05	0.25	0.5	0.75	0.95
X_1	0.167(*)	0.509(*)	0.654(*)	0.697(*)	0.754(*)
X_2	0.349(*)	0.353(*)	0.402(*)	0.461(*)	0.484(*)
X_3	-0.024	-0.08	0.099(*)	0.099(*)	0.096(*)
X_4	-0.07	-0.093(*)	0.077(*)	0.056(*)	0.056(*)
X_5	-0.125(*)	-0.066(*)	-0.07(*)	-0.041	-0.06(*)
X_6	0.091	0.05	0.003	-0.018	-0.009
X_7	0.027	0.101(*)	0.114(*)	0.147(*)	0.139(*)
X_8	-0.158(*)	-0.137(*)	0.117(*)	0.106(*)	0.065(*)

其中，*表示 p 值小于 0.05，即该回归系数通过了 p 值检验。

从表 3 中可以看出，季节变量 X_6 各个分位点均未通过 p 值检验，而其他通过 p 值检验的特征变量，在不同分位点下对单车租赁数量的影响机制不同。除了 X_3 在 0.05 与 0.25 分位点处、X_4 在 0.05 分位点处、X_5 在 0.75 分位点处和 X_7 在 0.05 分位点处未通过 p 值检验，其他各分位点均通过 p 值检验，都可以合理解释因变量Y。其中X_1、X_2 和 X_7 这三个变量的回归系数为正数，X_3、X_4、X_5 和 X_8 这

四个变量回归系数为负数。为了利于观测各变量各分位点的回归系数的变化趋势，本文运用 R 语言对其进行了可视化，其结果如下图 5 所示：

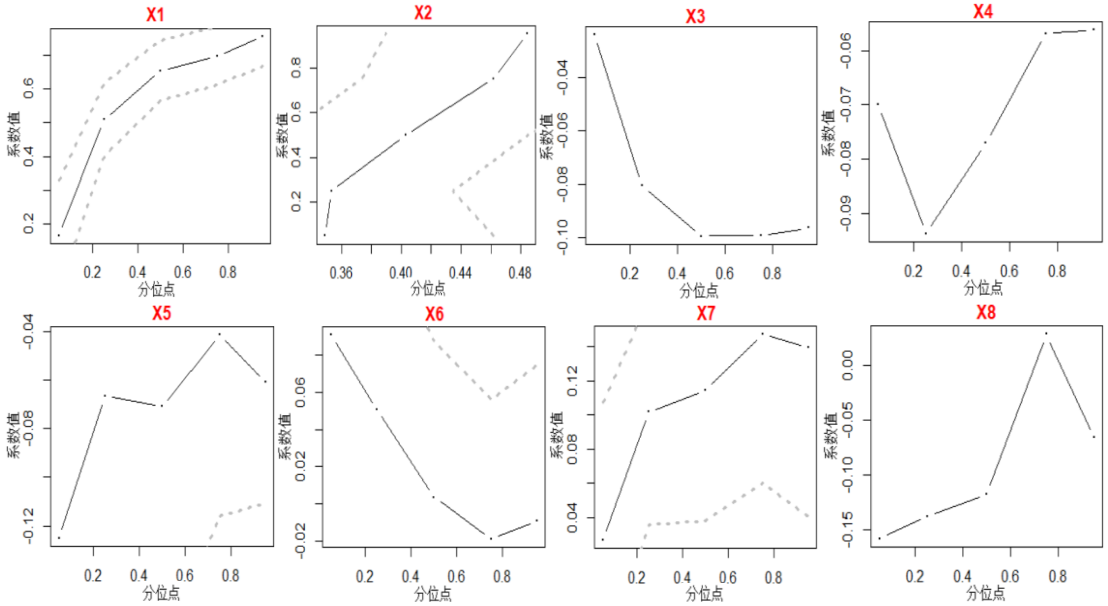

图 5 回归系数变化趋势

本文对各变量所出现的变化趋势给出了解释说明，具体如下表 4 所示：

表 4 特征变量解释说明

特征变量	变化趋势	解释说明
X_1	正增长	随着共享单车的发展，使用人数越来越多，其中，其初级发展阶段，单车租赁数量增长幅度最大
X_2	正增长	随着温度的升高，使用共享单车出现的人数越来越多
X_3	负增长	随着风速的增大，使用共享单车出现的人数越来越少
X_4	负增长	随着湿度的增大，使用共享单车出现的人数越来越少
X_5	负增长	节假日时使用共享单车出行的人数较少
X_6	\	\
X_7	负增长	工作日时使用共享单车出行的人数较多，周末则较少
X_8	先增后降	恶劣天气对共享单车租赁数量的影响较大

因季节变量 X_6 各分位点均未通过 p 值检验，故不作考虑。

3 集成学习

为了能够精准的预测不同条件下租赁单车的数量，本文使用集成学习算法构建单车租赁数量预测模型。

所谓集成学习（Ensemble Learning，EL），是指将若干个弱学习器通过一定

策略组合得到一个强学习器[11]。

EL 的核心思想如下：结合多个弱学习器的学习结果，通过汇总得到一个综合学习结果，从而获得比单个弱学习器更好的学习表现。

EL 的基本分类如下：装袋法（Bagging）、提升法（Boosting）和堆叠法（Stacking）；其核心思想和代表模型如表 5 所示。

表 5EL 的基本分类

基本分类	核心思想	代表模型
Bagging	构建多个相互独立的弱学习器，综合学习结果为每个弱学习器的算术平均值	随机森林（RF）
Boosting	不断训练上个弱学习器的学习结果，综合学习结果为每个弱学习器的加权平均值[12]	XGBoost、LightGBM 和 CatBoost
Stacking	将基学习器的学习结果汇总到元学习器中，综合学习结果为元学习器的学习结果[12]	\

Bagging 算法流程如下图 6 所示：

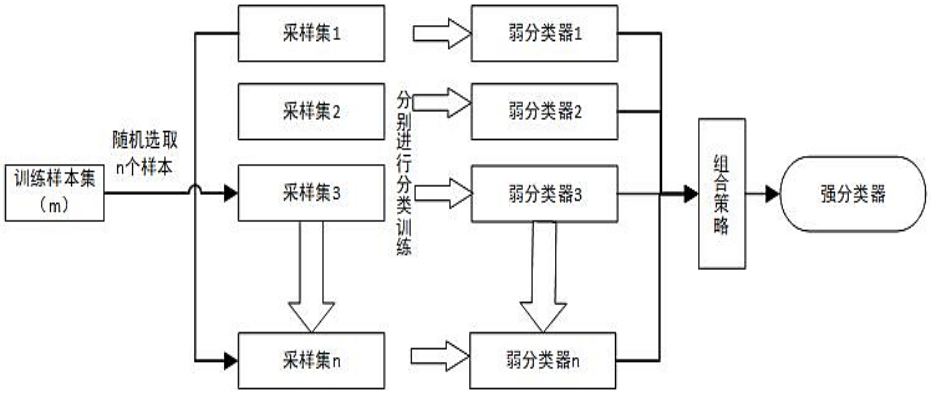

图 6Bagging 算法图示

Boosting 算法流程如下图 7 所示：

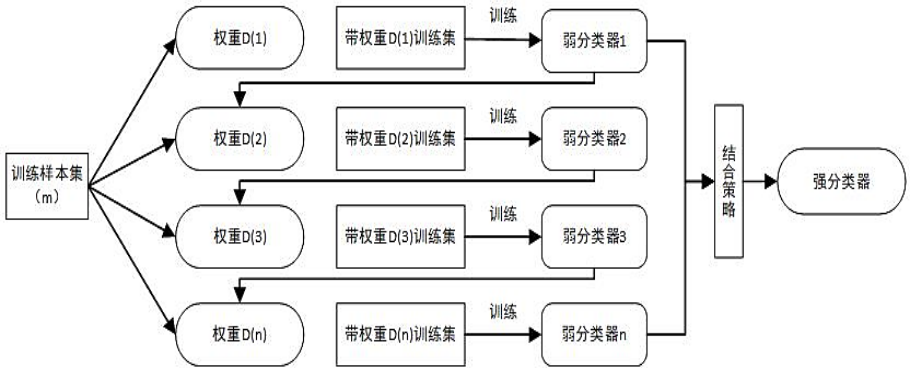

图 7Boosting 算法图示

Stacking 算法流程如下图 8 所示：

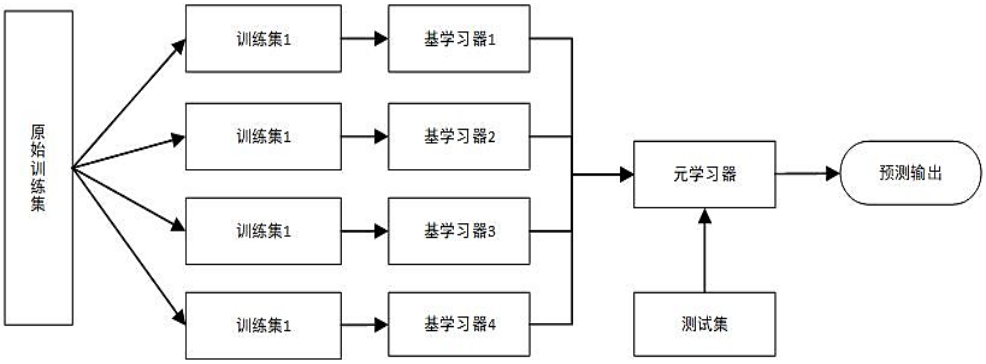

图 8Stacking 算法图示

本文运用 RF、XGBoost、LightGBM 和 CatBoost 四种模型分别对单车数据集进行了训练[19]。然后，并运用 Stacking 算法对上述四种模型进行了融合。最后的训练结果证明了集成学习算法对该数据集有很强的泛化能力，其中 CatBoost 模型对该数据集的训练效果最好。

3.1 随机森林

Bagging 算法主要运用有放回随机抽样技术（Bootstrap）[20]。抽样数据与原数据集大小相等，因此样本数据可能包括重复数据并且可能无法遍历原数据集中的所有数据[20]。Bagging 算法通过 Bootstrap 技术引入了随机样本扰动，提高了模型的泛化能力，并产生了袋外数据（out of bag data，oob）。

随机森林(RF)由是决策树集成而来[21]。相比于传统的 Bagging 算法，RF 引入了随机属性扰动，使其基学习器之前的差异不在仅来自于随机样本扰动，从而提高了模型的泛化能力[12]。

贪心算法是一种通过控制局部最优来达到全局最优的算法，即假设每片叶子最优，则整个树结构最优，因此无需枚举所有可能的树结构。本文运用的四种模型均运用贪心算法进行求解[12]。

RF 的优点如下：（1）适用于高维数据；（2）适用于大样本数据；（3）高度并行化，训练速度快。RF 的缺点如下：不适用于噪音较多的数据。

本文运用 RF 对单车数据集进行了训练，其中，RF 的学习曲线如下图 9 所

示：

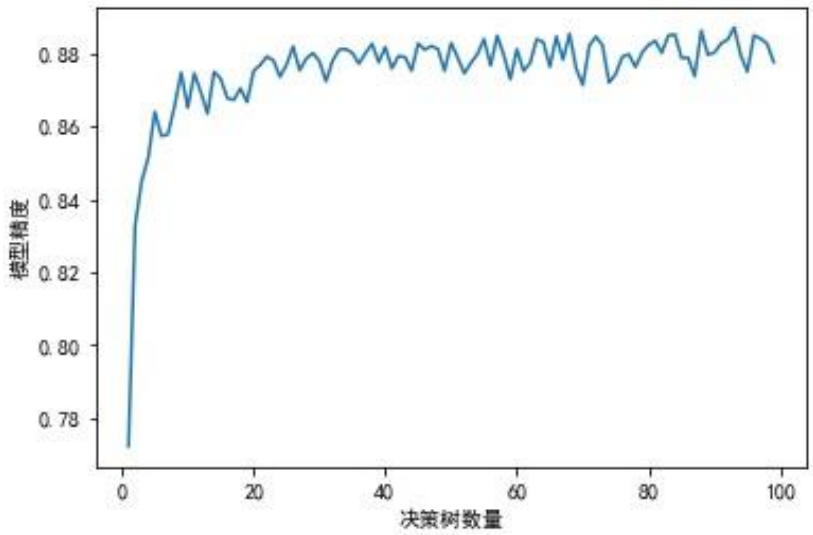

图 9RF 学习曲线

从图 9 中可以看出，随着决策树数量的增加，RF 的模型精度也越来越高，直到决策树增加到 20 棵时，RF 模型开始收敛。

3.2 XGBoost

XGBoost 是 Boosting 算法的一种，以下简称 XGB，其基本构成单位是 CART，即表现为二叉树的决策树。XGB 在梯度提升树（GBDT）的基础上进行了改进，即在每个叶子节点上设置一个叶子权重，其值为所有在这个叶子节点上的样本在这棵树上的回归值，从而提高了模型的泛化能力[13]。

XGB 的目标函数由损失函数和正则化项两部分组成，其公式如下：

$$\begin{aligned} obj &= \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \\ where\, \Omega(f) &= \gamma\sqrt{T} + \frac{1}{2}\lambda \parallel w \parallel^2, \end{aligned} \tag{11}$$

引入损失函数可以减小模型的偏差和方差；引入正则化项可以限制叶子节点的个数和分数，从而降低发生过拟合的概率[14]。

假设生成t棵树，则：

$$\begin{aligned}\hat{y}_i^{(0)} &= 0,\\ \hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i),\\ \hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i),\\ &\cdots\\ \hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i),\end{aligned} \tag{12}$$

则目标函数为：

$$Obj^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \Omega(f_t). \tag{13}$$

将目标函数进行二阶泰勒展开，则：

$$Obj^{(t)} \approx \sum_{i=1}^{n} \left[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t)\Big], \tag{14}$$

这种做法一方面可以减小预测值与真实值之间的误差，另一方面可以简化树模型的结构。

令梯度统计量g_i和h_i等于 0，可得叶子节点的最优解为：

$$w_j = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}, \tag{15}$$

则目标函数的最优解为：

$$Obj = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T. \tag{16}$$

上述公式的符号说明如下表 6 所示：

表 6XGB 符号说明

符号	符号说明
T	叶子节点数
w	叶子节点的分数
γ	用于控制叶子结点的个数
λ	用于控制叶子节点的分数
g_i	一阶梯度统计量
h_i	二阶梯度统计量

本文运用 XGB 模型对单车数据集进行了训练，其中，XGB 特征变量重要性如下图 10 所示：

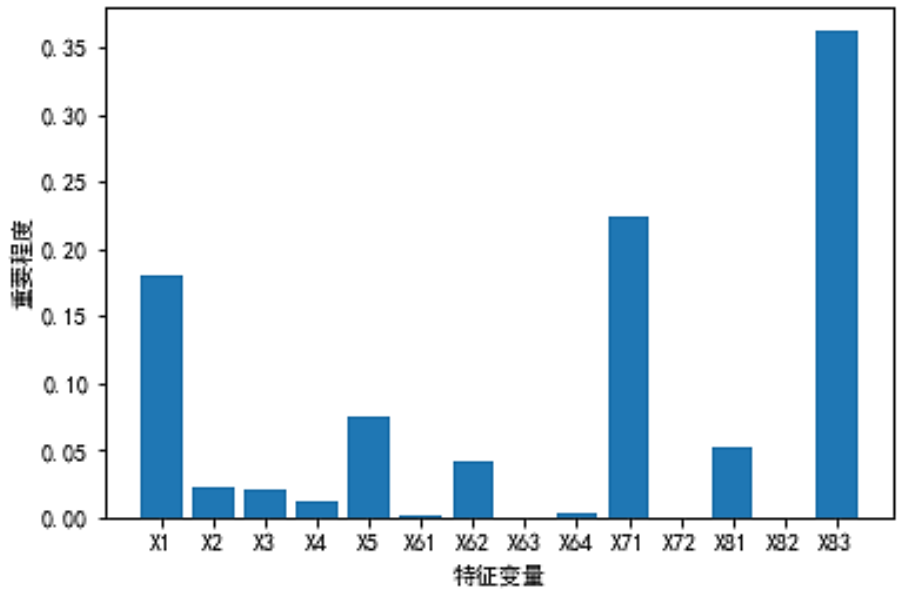

图 10XGB 特征变量重要性

从图 10 中可以看出，在影响共享单车租赁数量的特征变量中，最重要的为天气变量，其次为时间变量和工作日与周末变量。

XGB 的优点如下：（1）模型复杂度较低；（2）模型精度更高；（3）不易陷入局部最优；XGB 的缺点如下：对计算机运行内存消耗大。

3.3 LightGBM

LightGBM 也是 Boosting 算法的一种，可简称（LGB）[15]。相比于同为 GBDT 改进算法的 XGB，LGB 的训练速度更快，内存消耗更少[14]。

LGB 采用直方图算法（Histogram），通过对连续型数据进行分箱处理，有效减少了运算时间。此外，这种粗糙分割方式也在一定程度上降低了发生过拟合的概率，而数据离散化也可以减小预测值与目标值之间的误差。

LGB 支持类别特征识别。对于分类型数据，通常需要进行哑变量处理，然而这种方式可能造成数据稀疏性，无法保证数据的平衡。LGB 通过 m-v-m 的切分方式，实现类别特征的最优切分，即先对类别特征均值进行排序，然后根据排序依次枚举[15]，以此寻优。

LGB 支持高度并行化。在特征并行上，分别保存全部数据并基于此前得到的最佳划分条件执行接下来的训练。在数据并行上，使用分散规约的方法，把直方图合并后的数据分摊到不同的学习器进行做差，从而减少了计算量。在投票并行上，只合并部分特征的直方图，通过找出重要程度高的特征，基于投票法来确定出最优的分割点，提升了运算效率[14]。

LGB 的优点如下：（1）用遍历直方图代替遍历样本，减少了运算时间；（2）用单边梯度算法过滤小梯度样本，减少了计算量；（3）用互斥特征捆绑算法，降低了内存消耗。LGB 的缺点如下：（1）不适用于噪音较多的数据；（2）可能没有考虑全部特征变量。

3.4 CatBoost

CatBoost 同为 Boosting 算法的一种，以下简称 CatB，其基本构成单位是对称决策树。相比于同为 GBDT 改进算法的 XGB，CatB 采用了梯度步长的无偏估计方法，有效减少了梯度偏差[15]。

CatB 以对称树作为基模型，将特征变量进行了离散化保存。此外，CatB 对类别特征使用完美哈希进行按位压缩，主要采用分布式学习方式，实现了对多个数据集的并排计算[10]。

CatB 嵌入了一种自动将类别特征处理为数值特征的算法，即改进的 Greedy TBS 算法。在将标签平均值作为节点分裂标准的基础上，通过添加先验分布项，降低了数据噪声，其公式如下：

$$\hat{x}_k^i = \frac{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k}=x_{\sigma_p,k}\right]Y_{\sigma_j}+a*P}{\sum_{j=1}^{p-1}\left[x_{\sigma_j,k}=x_{\sigma_p,k}\right]+a}, \tag{17}$$

其中，P代表添加的先验项，a为大于 0 的权重系数。

CatB 采用了一种新型的叶子节点的计算方式，可以避免数据集排列过程中直接计算可能出现的过拟合问题。在选择第一个节点时，先只选择一个特征，当

生成至第二个节点时，则考虑前一个特征和其它任意一个类别特征的组合，直到其所有分割点都被视为具有两个值的类别特征。

CatB 的优点如下：（1）参数较少，具有噪声鲁棒性；（2）对数值特征和类别特征的学习效果均较好；（3）模型灵活度较高。CatB 的缺点如下：学习结果受参数设定的影响较大。

本文运用 CatB 模型对单车数据集进行了训练，其中，CatB 测试集的仿真结果如下图 11 所示：

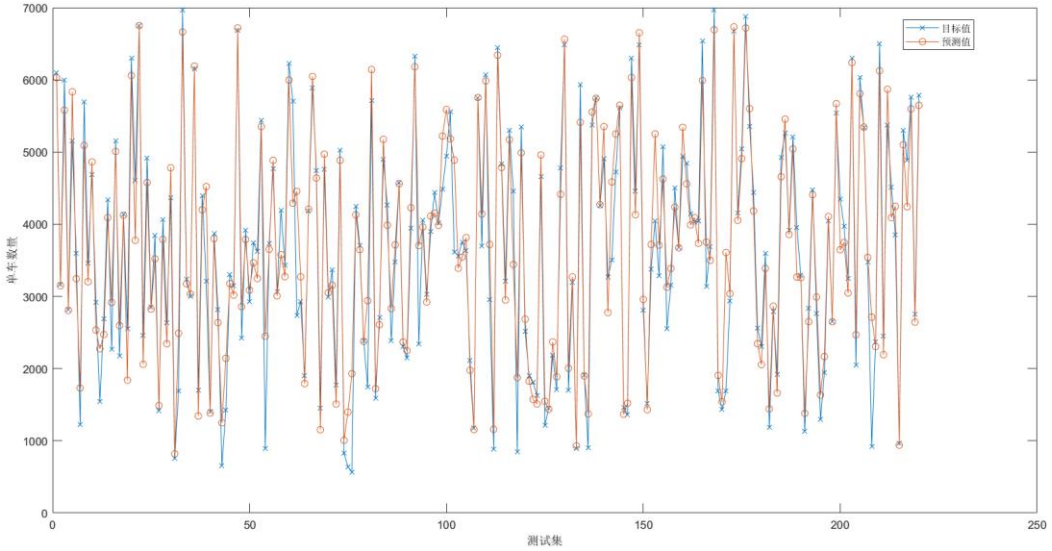

图 11CatBoost 仿真结果

从图 11 可以看出，CatB 模型的训练效果非常为好，预测值与目标值之间的误差较小。

3.5 模型融合

单个模型可能具有局限性，而融合模型往往泛化能力更好[11]。本文运用的模型融合算法为 Stacking。

假设$S=\{(x_i,y_i),i=1,2,\cdots,N\}$为训练集，$(x_i,y_i)$为训练集中的第$i$个样本，$x_i$为样本的特征属性，$y_i$为样本对应的预测值，$K$为$x_i$的长度，则：

$$x_i=\{x_1,x_2,\cdots,x_K\}. \tag{18}$$

把S随机等分为M份，得到$\{S_1\},\{S_2\},\cdots,\{S_M\}$，在第$m$次交叉训练过程中，定

义$\{S_m\}$为测试集，$\{S_{-m}\}=S-\{S_m\}$为训练集。第一层M个基学习器对训练集$\{S_{-m}\}$学习后得到的基模型分别为$L_1,L_2,\cdots L_m$。

M个模型根据每一个样本(x_i,y_i)的特征x_i得到训练值为：

$$\bar{y}_{1,i},\bar{y}_{2,i},\cdots,\bar{y}_{M,i},$$

与 y_i 组合后得到：

$$z_i=\{y_i,\bar{y}_{1,i},\bar{y}_{2,i},\cdots,\bar{y}_{M,i}\}, \tag{19}$$

其中 $i=1,2,\cdots,\frac{N}{M}$，将 $Z=\left\{z_1,z_2,\cdots,z_{\frac{N}{M}}\right\}$作为第二层元学习器的训练数据得到原模型Y。

本文将已训练好的模型 RF、XGB、LGB 和 CatB 作为基学习器，运用 Stacking 算法对四种模型进行了融合，得到一个元学习器，其融合框架如下图 12 所示：

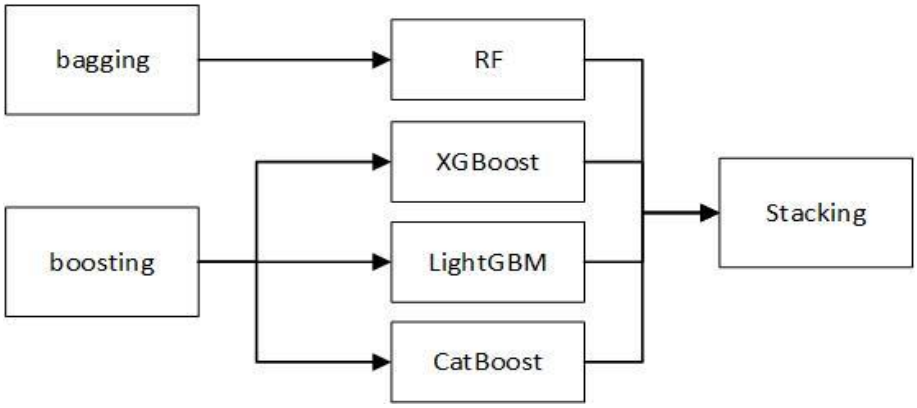

图 12Stacking 融合框架

Stacking 测试集的训练效果如图 13 所示：

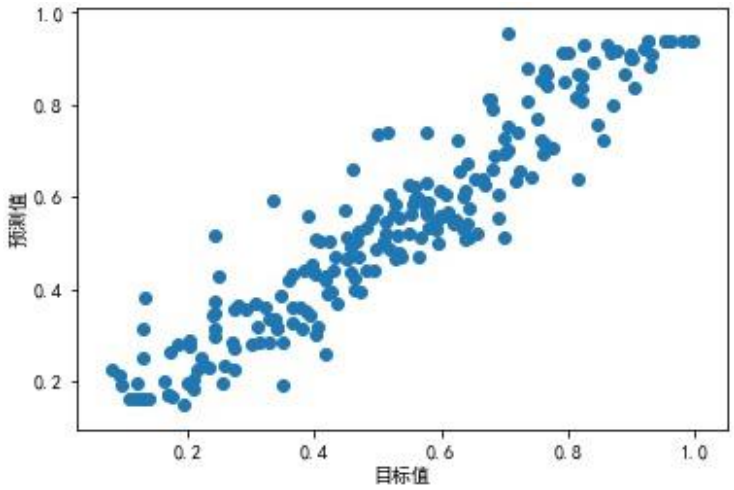

图 13Stacking 测试集的拟合优度

从图 13 可以看出，stacking 训练效果较好，图中所示的点基本都集中在对角线上，拟合优度较高。

3.6 模型评价

本文以拟合优度 R^2、均方误差 MSE 和绝对平均误差 MAE 作为上述五种集成模型的评价标准，其具体结果如下表 7 所示：

表 7 模型评价

集成学习		R^2	MSE	MAE
Bagging	RF	0.8846	0.0063	0.0543
Boosting	XGBoost	0.8986	0.0055	0.0499
	LightGBM	0.8860	0.0062	0.0538
	CatBoost	0.9204	0.0044	0.0448
Stacking		0.8824	0.0066	0.0606

从表 7 可知，CatB 模型的训练效果最好，而 Stacking 模型融合的效果一般，其原因可能是发生了过拟合现象，故本文接下来进行异常检测的数据集将以 CatB 测试集各点误差构造。

4 异常检测

异常点是数据集中显著不同的点。异常检测是指找出异常点的过程[17]。本文以 CatB 测试集的各点误差构造数据集，利用孤立森林算法找出该数据集中的异常点，通过单车数据是否异常判断该城市状态是否异常，以此达到监控城市健康状况的目的。

4.1 孤立森林

孤立森林（Isolation Forest, IF)通过隔离数据集中的异常点来识别异常，因其识别异常点的精准度较高且速度较快而被广泛应用[16]。

IF 的基本思想如下：选取一个随机超平面对某个数据空间进行不断切割，直到每个切割子空间里只有一个数据点为止[16]。

IF 与大多数集成算法原理相同，其步骤均可分为创建树、训练树和集成树三

步，其伪代码如下表 8、9 和 10 所示：

表 8 创建孤立树

算法 1：*iForest(X, t, ψ)*

输入：*X - input data, t - number of trees, ψ - subsampling size*

输出：*a set of t iTrees*

1：**Initialize** *Forest*
2：*set height limit l = ceiling($log_2 ψ$)*
3：**for** *i = 1 to t* **do**
4：　　*X'* ← *sample(X, ψ)*
5：　　*Forest* ← *Forest* ∪*iTree(X', 0, l)*
6：**end for**
7：**return** *Forest*

表 9 训练孤立树

算法 2：*iTree(X')*

输入：*X' – input data*

输出：*an iTree*

1: **if** *X' cannot be divided* **then**
2:　**return** *exNode{Size* ← *| X' |}*
3: **else**
4:　let Q be a list of attribute in *X'*
5:　randomly select an attribute q ∈ Q
6:　randomly select a split point p between the max and min values of attribute q in *X'*
7:　X_l ← *filter(X', q < p)*
8:　*Xr* ← *filter(X', q ≥ p)*
9:　**return** *inNode{Left* ← *iTree(X_l),*
10:　　*Right* ← *iTree(X_r),*
11:　　*SplitAtt* ← *q,*
12:　　*SplitValue* ← *p}*
13: **end if**

表 10 集成孤立树

算法 3：集成孤立树 *PathLength(x, T, hlim, e)*

输入：*x – an instance, T – an iTree, hlim – height limit, e – current path length;*
To be initialized to zero when first called

输出：*path length of x*

1: **if** *T is an external node or e ≥ hlim* **then**
2:　**return** *e + c(T, size) {c(.) is defined in Equation 1}*
3: **end if**
4:　a ← *T.splitAtt*
5: **if** x_a < *T.splitValue* **then**
6:　**return** *PathLength(x, T.Left, hlim, e +1)*
7: **else** {x_a ≥ *T.splitValue*}
8:　**return** *PathLength(x, T.Right, hlim, e +1)*
9: **end if**

IF 的优点如下：（1）相比于其他异常检测方法，IF 不需要计算有关距离和密度的指标，模型运行速度较快；（2）具有时间复杂度，模型较为稳定；（3）支持大规模分布式运算。

本文以 CatB 测试集各点误差构造数据集，利用孤立森林算法找出该数据集中的异常点，通过单车数据是否异常判断该城市状态是否异常，以此达到监控城市健康状况的目的。模型运算结果如下：数据集样本数量共计 220 个，异常点数量共计 43 个，可视化结果如图 14 所示：

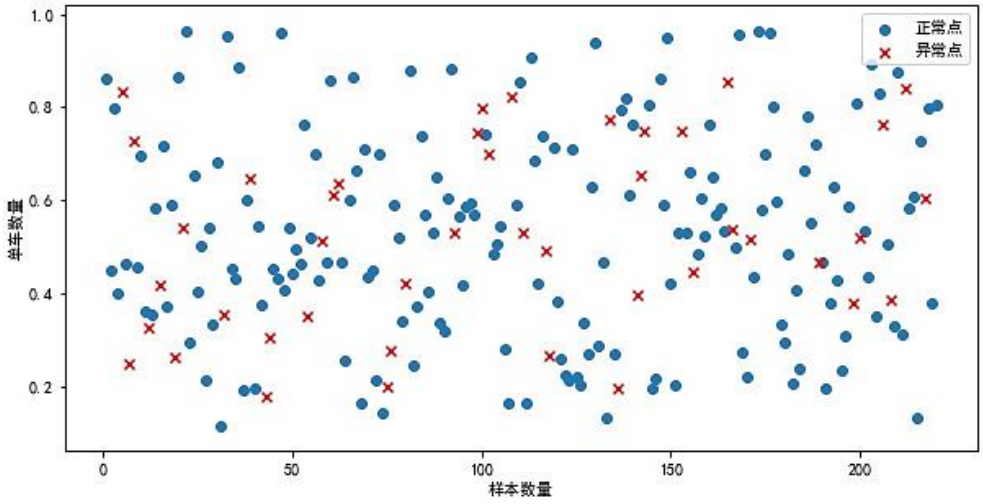

图 14 异常点检测结果

其中，城市健康状况异常的出现可能与城市突发意外事件的发生、节假日的到来、温度与风速以及湿度的突变、极端恶劣天气的产生有关。

4.2 支持向量机

支持向量机（SVM）常被应用于分类和回归问题[17]。

将给定的训练集记为：

$$T = \{(a_1, c_1), (a_2, c_2), \cdots, (a_N, c_N)\}, \tag{20}$$

其中，$a_i \in \Omega \subset \mathrm{R}^n$，$\Omega$称为输入空间，输入空间中的每一个点$a_i = [a_{i1}, a_{i2}, \ldots, a_{in}]$由n个属性特征组成；$C_i \in \mathrm{R}, \mathrm{i} = 1,2, \ldots, \mathrm{N}$[23]。

支持向量回归根据损失函数最小原则，采用ε −不敏感损失函数。在分类中，

损失函数中计入每一个预测的误差，在进行支待向量回归的过程中，忽略误差函数值小于指定值$\varepsilon(\varepsilon > 0)$的观测值。

支待向量回归的具体公式如下：

$$\min_{\omega,b} \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{N} L_\varepsilon(f(a_i) - c_i), \tag{21}$$

其中，$C \geq 0$为惩罚系数，$f(x) = \omega^{\mathrm{T}} x + b$，$L_\varepsilon$为损失函数，其定义为：

$$L_\varepsilon(z) = \begin{cases} 0, & |z| \leqslant \varepsilon, \\ |z| - \zeta, & |z| > \varepsilon, \end{cases} \tag{22}$$

更进一步， 引入松弛变量ζ_i, η_i，则新的最优化问题为：

$$\begin{aligned} &\min_{\omega,b,\zeta_i,\eta_i} \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{N} (\zeta_i + \eta_i), \\ &\text{s.t.} \begin{cases} f(a_i) - c_i \leqslant \varepsilon + \zeta_i, \\ c_i - f(a_i) \leqslant \varepsilon + \eta_i, \end{cases} \\ &\zeta_i \geqslant 0, \eta_i \geqslant 0, i = 1,2,\cdots,N. \end{aligned} \tag{23}$$

定义 Lagrange 函数：

$$\begin{aligned} L(\omega, b, \alpha, \beta, \zeta, \eta, \mu, \nu) = {} & \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{N} (\zeta_i + \eta_i) - \sum_{i=1}^{N} \mu_i \zeta_i - \sum_{i=1}^{N} \nu_i \eta_i \\ & + \sum_{i=1}^{N} \alpha_i (f(a_i) - c_i - \varepsilon - \zeta_i) + \sum_{i=1}^{N} \beta_i (c_i - f(a_i) - \varepsilon - \eta_i). \end{aligned} \tag{24}$$

同样地可以得到其 Lagrange 对偶问题如下：

$$\begin{aligned} &\max_{\alpha,\beta} \sum_{i=1}^{N} [\varepsilon(\beta_i + \alpha_i) - c_i(\beta_i - \alpha_i)] + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\beta_i - \alpha_i)(\beta_j - \alpha_j)(a_i \cdot a_j), \\ &\text{s.t.} \begin{cases} \sum_{i=1}^{N} (\beta_i - \alpha_i) = 0, \\ 0 \leqslant \alpha_i, \beta_i \leqslant C, i = 1,2,\cdots,N. \end{cases} \end{aligned} \tag{25}$$

假设最终解为$\alpha^* = [\alpha_1^*, \alpha_2^*, \cdots, \alpha_N^*]^{\mathrm{T}}, \beta^* = [\beta_1^*, \beta_2^*, \cdots, \beta_N^*]^{\mathrm{T}}$，在$\alpha^* = [\alpha_1^*, \alpha_2^*, \cdots, \alpha_N^*]^{\mathrm{T}}$中，找出$\alpha^*$的某个分量$C > \alpha_j^* > 0$，则有：

$$\begin{aligned} &\begin{cases} \omega^* = \sum_{i=1}^{N} (\beta_i^* - \alpha_i^*) a_i, \\ b^* = c_j + \varepsilon - \sum_{i=1}^{N} (\beta_i^* - \alpha_i^*) \boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{a}_j, \end{cases} \\ &f(\boldsymbol{x}) = \sum_{i=1}^{N} (\beta_i^* - \alpha_i^*) \boldsymbol{a}_i^{\mathrm{T}} \boldsymbol{x} + b^*. \end{aligned} \tag{26}$$

本文运用支持向量机对上述检测到的异常点进行训练，以期通过其向量空间特征分布来反映各变量对单车租赁数量异常的影响程度，其中，异常的出现可能与城市突发意外事件的发生、节假日的到来、温度与风速以及湿度的突变、极端恶劣天气的产生有关，向量空间分布如图 15 所示：

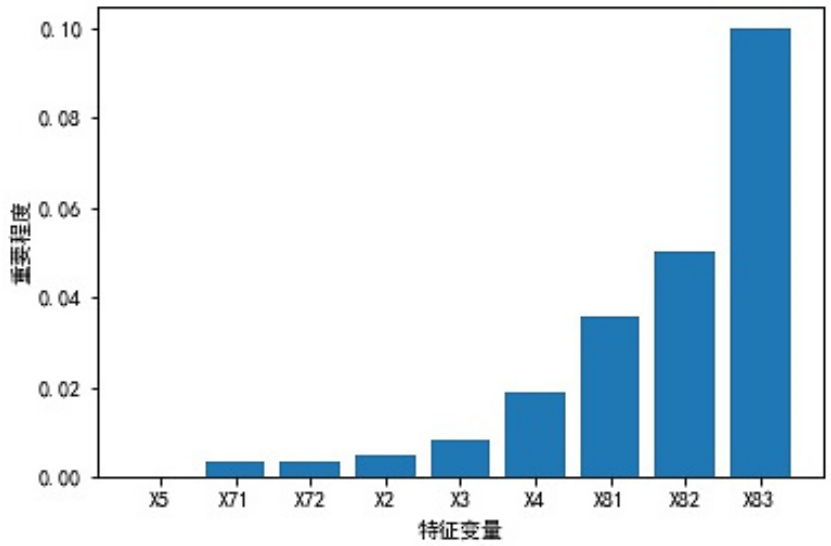

图 15SVM 向量空间分布

从图 15 可以看出，共享单车租赁数量的异常受天气因素影响最大，而对于其他影响因素，由上文分位数回归可知，对其影响具有一定的规律性和周期性。

5 结论

首先，本文通过对该单车数据集进行探索性数据分析，运用分位数回归得到了各特征变量与共享单车租赁数量各分位点之间的线性关系，有效解释了各特征变量的概率分布情况以及对单车数量的影响情况。

其次，本文通过集成学习算法训练出了精度较高的共享单车租赁数量预测模型。该模型可以推广到其他各地区进行单车需求分析，以提高资源利用率，避免因城市资源配置不合理而造成的浪费。此外，本文运用了多种集成算法对单车数据集进行了训练，其训练效果均较好，一方面证明了集成算法本身的优势，另一方面也为集成算法应用于城市交通其他领域提供了理论依据。

最后，本文利用孤立森林进行异常检测，并运用支持向量机进行回归，得到了各变量对单车租赁数量异常的影响程度，共享单车租赁数量异常在一定程度上可以反映城市健康状况异常，其中，异常的出现可能与城市突发意外事件的发生、节假日的到来、温度与风速以及湿度的突变、极端恶劣天气的产生有关。

预测模型能够帮助城市合理规划共享单车的投放数量，而异常检测模型则有助于城市及时处理突发事件，希望本文的研究能够为城市资源合理配置提供参考。

参考文献

[1]. 李盼道,宋晔琴.共享单车的供给模式及政府规制[J].重庆交通大学学报(社会科学版),2019,19(06):47-53+59.

[2]. 许新塬.“互联网+”背景下对共享单车可持续发展的思考与建议[J].现代商业,2021(01):49-51.

[3]. 柯湾. 共享单车大学生用户持续使用意愿影响因素研究[D].西南交通大学,2019..

[4]. 耿淑文. 共享经济平台下的用户持续使用意愿影响因素分析[D].江西师范大学,2020.

[5]. 陈艺坤. 基于粗糙集和增量 SVM 的入侵检测方法研究[D].西安科技大学,2012..

[6]. 刘如辉. 半监督约束快速密度峰值聚类算法研究及其在空调控制上的应用[D].浙江大学,2018.吴胜武.建设智慧城市必须消除“信息孤岛”[J].居业,2014(01):46.

[7]. 张秋月. 消费金融公司个人信用评价方法研究[D].云南财经大学,2018.

[8]. 张晨子. 中国老年消费的地区差异及影响因素分析[D].云南财经大学,2020.

[9]. 朱炜玉. 基于数据驱动模型的突发水污染预警技术与应急管理研究[D].哈尔滨工业大学,2018.

[10]. 范诗语,耿子悦,田芮绮,杜永强.基于集成学习的上市企业违约风险评价[J].统计与管理,2021,36(02):62-68.

[11]. 李阳,黄伟,席建忠.基于 Stacking 算法集成模型的电厂 NO_x 排放预测[J/OL].热能动力工程,2021(05):73-81[2021-05-31].https://doi.org/10.16146/j.cnki.rndlgc.2021.05.012.

[12]. 易茂祥,宋晨钰,于金星,宋钦,鲁迎春,黄正峰.基于随机森林的集成电路适应性测试方法研究[J/OL].郑州大学学报(工学版):1-6[2021-05-31].https://doi.org/10.13705/j.issn.1671-6833.2021.02.016.

[13]. 刘岩,王玉君,杨晓坤,李文文,郭磊.基于 KPCA 和 XGBoost 算法的非侵入式负荷辨识方法[J/OL].电测与仪表:1-9[2021-05-31].http://kns.cnki.net/kcms/detail/23.1202.TH.20210520.1648.002.html.

[14]. 李梅芳.基于 LightGBM 的上市公司财务困境预测[J].轻工科技,2021,37(05):129-130+164.

[15]. 张涛,范博.基于 CLPSO-CatBoost 的贷款风险预测方法[J].计算机系统应用,2021,30(04):222-226.

[16]. 王一大.基于孤立森林算法的分布式服务故障分析模型研究与应用[J].信息通信技术,2021,15(02):72-78.

[17]. 郑奇,李凤岐,李原,邹彦纯,朱洪明.基于支持向量机和实物期权法的工业大数

据资产价值研究[J].新型工业化,2020,10(05):170-172.
[18]. 郑奇,李凤岐,李原,邹彦纯,朱洪明.基于支持向量机和实物期权法的工业大数据资产价值研究[J].新型工业化,2020,10(05):170-172.
[19]. 王浩. 基于特征价格理论和CatBoost的旧机动车价值评估模型研究[D].天津商业大学,2019.
[20]. 王飞. 集成分类器及其在个人信用评估的应用[D].中南大学,2012.
[21]. 杜臻. 基于特征提取和异常分类的网络流量异常检测方法[D].南京邮电大学,2019.
[22]. 陈彬,朱臻涛,张翔.基于物联网的供水管网智慧运维系统设计[J].现代信息科技,2020,4(10):171-175.
[23]. 吴兴惠,周玉萍,邢海花,龙海侠.机器学习分类算法在糖尿病诊断中的应用研究[J].电脑知识与技术,2018,14(35):177-178+195.

致谢

在论文撰写完成之际，回首建模历程，感慨万千。这一路上，压力与进步同在，磨练与挑战并存。这一路上，我们是无比幸运的，因为有许多的人给与我们帮助与关心，借此机会，对每一个帮助过我们的人表达我最诚挚的谢意。

首先，我们要由衷的感谢两位指导老师，感谢他们一直以来对我们的耐心指导与帮助。在论文选题与数据收集等方面，老师都为我们付出了宝贵的时间与精力，为我们提供修改意见。当我们在研究中遇到棘手的问题时，老师也悉心指导，使我们克服重重的困难。

其次，由衷感谢我们的家人与朋友。在研究过程中遇到困难时给予我们安慰，并在精神与物质上给予支持。让我们顺利的完成这篇论文。在这里，向我们的家人表达最诚挚的谢意。

最后，由衷的感谢系主任以及学校的比赛负责人，感谢他们帮助我们顺利的提交论文。

建模期间，受益良多。一段旅程结束代表另一段旅程的开始，我们将怀着一颗感恩之心，继续学习，继续探索。