東南大學

毕业设计(论文)报告

题　目　基于机器学习的交通违法时空演化趋势预测研究

交通　院（系）交通工程　专业

学　　号　21014101

学生姓名　魏　薇

指导教师　王　晨

起止日期　2018 年 2 月至 2018 年 6 月

设计地点　东南大学交通学院

东南大学毕业（设计）论文独创性声明

本人声明所呈交的毕业（设计）论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得东南大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

论文作者签名：　　日期：2018 年 6 月 1 日

东南大学毕业（设计）论文使用授权声明

东南大学有权保留本人所送交毕业（设计）论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权东南大学教务处办理。

论文作者签名：　　导师签名：

日期：2018 年 6 月 1 日　　日期：2018 年 6 月 1 日

摘要

我国近些年的综合实力快速增长，交通事业也正在快速发展，随之而来的是日益严重的交通安全问题。现有交通安全研究的关注点普通在交通事故上，却忽略了交通违法行为作为交通事故最重要的影响因素之一，若能对其从宏观上预测时空演化趋势，可以显著改善交通安全甚至整个道路交通环境。

本文首先利用多元统计分析理论中的对应分析方法，得到昆山市历史交通事故类型与交通违法类型的对应影响关系模型，从而针对特定事故类型确定所要分析研究和预测的违法类型对象，达到有效减少该类事故发生数量的目的。

其次，本文基于国内外对交通违法行为的研究现状，从驾驶员、车辆、道路交通环境、自然环境和社会环境等几个方面归纳提取出交通违法行为的主要影响因素，通过数据集成和特征工程的方法，构建出违法预测的特征集。

针对特征集与交通违法预测问题的内在特性，本文介绍并分析了线性回归、LASSO 回归、岭回归、梯度提升决策树、XGBoost 等常用机器学习模型的优劣势，从而选取出合适的机器学习模型进行预测，并且使用 Blending 的方法得到了融合模型。考虑到各区域之间在地理特征上具有高度相似性，本文进一步提出了基于聚类的分层融合模型。各模型在测试集上的验证结果，证明了本文的机器学习违法预测融合模型具有较高的准确度，且对比较为简单的回归模型有显著提高。

最后，基于模型得到的特征重要性和预测结果，分析得到违法数量在昆山市的时空演化趋势并且为交通监管部门提出有效预防管理该类违法行为的对策建议。这一系列交通违法行为的预测分析流程经过改进，可被应用于各地的交通违法监管中，使得执法部门可以预知各类型违法的时空分布，从而高效调配其监管资源，从根本上控制了交通事故的发生量。

关键词：机器学习模型，交通违法预测，融合模型，对应分析

Abstract

The transportation industry in China is developing rapidly. The focus of current traffic safety research is common in traffic accidents to improve traffic safety, but it neglects the traffic illegal behavior as one of the most important factors of traffic accidents. Thus, it is meaningful to predict the temporal and spatial evolution trend from the macroscopic.

This paper first uses the corresponding analysis method to obtain the corresponding relationship model between the historical traffic accident types and the traffic illegal types in Kunshan, so as to determine the illegal type objects which should be analyzed and studied and predicted in view of the specific type of accident, so as to achieve the purpose of reducing the number of such accidents.

Secondly, the main influencing factors of traffic illegal behavior are extracted from several aspects, such as drivers, vehicles, road traffic environment, natural environment and social environment. Through data integration and characteristic engineering methods, a feature set of illegal prediction is constructed.

In this paper, the advantages and disadvantages of linear regression, LASSO regression, ridge regression, GBDT, XGBoost and other common machine learning models are introduced in view of the intrinsic characteristics of the feature set and traffic illegal prediction. The suitable machine learning model is selected to predict the model, and the method of Blending is used to get the fusion model. Considering the high similarity of geographical features among different regions, a predicting model based on clustering is proposed. The validation results of each model on the test set show that the machine learning model in this paper has high accuracy.

Finally, based on the feature importance and prediction results obtained by the model, the temporal and spatial evolution trend of violations in Kunshan is analyzed, and the countermeasures and suggestions for effective prevention and management of this kind of illegal behavior are put forward for the traffic supervision department. This process of prediction analysis can be applied to traffic illegal supervision in various places. The enforcement departments can predict the temporal and spatial distribution of various types of illegal activities, so as to efficiently allocate their regulatory resources and fundamentally control the occurrence of traffic accidents.

KEY WORDS: Machine Learning Model; Traffic Violation Prediction; Ensembling ; Corresponding Analysis

目录

第一章　绪论

1.1　研究背景及意义

我国近些年的综合实力快速增长，经济发展日益繁荣，随着国家逐步制定并实施交通相关的战略，以及智能交通的发展普及，我国交通事业将进一步发展并同时促进国民经济的发展。然而中国在车辆数量和交通事故死亡人数上已经成为全球最高的国家。其中，道路交通事故造成的死亡已成为 5 至 44 岁人口死亡的三大死因之一，直接经济损失约占中国年度国内生产总值（GDP）的 1-3%（WHO，2009）。

从不同的方面系统地分析交通安全数据，应用先进的科学技术和方法，实施适当措施降低死亡率将是未来几年中国面临的紧迫和挑战。大量研究表明，与交通事故发生率和伤亡相关的最重要的因素是驾驶行为，例如超速驾驶，酒后驾车，疲劳驾驶以及不使用保护装置，因此，研究预测交通违法行为是十分必要的。国外对交通安全的研究时间比较久，已有五六十年的历史，研究内容也比较广泛。但是关于交通违法的研究却还未发展成熟，并且大多从单个违法事件的人为因素进行分析预测，比如文献只使用了驾驶员的个人属性和车辆属性对违法严重程度进行预测，很少有同时结合地理环境信息和社会经济变量从宏观角度来对其进行分析预测的相关研究。同时，其方法大多基于传统的统计学方法如贝叶斯网络模型[1]，还未有成熟的基于机器学习的方法来对交通违法进行时空演化分析。

交通违法研究在国内起步相对较晚，研究成果也相对较少，主要成果包括违规行为影响因素以及交通违法数量与社会发展的关系等[2]。对于交通违法行为的预测，同样的，大多研究只使用了个人的特征，未把时空因素考虑在内，比如文献[3]。而文献[4]虽然将时间特征考虑在内，根据 2011 年至 2015 年重庆市交通违法行为的年度历史数据，建立了交通违章的灰色动态系列预测模型，但忽略了地理特征的变化。

然而，从宏观上来讲，只有从整个城市的维度来进行违法行为的时空分布预测才能最大程度上为相关道路执法部门提供有效依据，以制定相应的道路交通违法治理措施以减少交通事故和经济损失。

在此背景下，将时间特征和地理特征都纳入数据集，研究并选取合适的机器学习模型来对违法行为的时空分布进行较为准确的预测，是非常具有现实意义的研究。相关部门可针对违法行为数据的分析结果作出相应管理，最终促进地区整体交通服务水平。

1.2　研究目标及内容

1.2.1　研究目标

在借鉴国内外文献综述的基础上，首先，采用对应分析的方法得出不同交通事故类型所对应的道路违法类型以及对应的影响程度，从而达到针对不同待整治的交通事故而预测分析相应类型违法行为的目的。同时，将昆山市不同交通中区的天气情况、规划信息、路网信息等数据作为自变量，结合当日交通量，采用合适的机器学习模型，如 LASSO 回归、岭回归、随机梯度提升树，以及模型融合和基于聚类的模型融合的方法，对昆山市交通违

法行为进行分析预测，并探究道路违法行为的内在影响因素。此外，采用随机森林的方法对短时交通流进行预测，从而实现提前数天预测不同区域的交通流并进一步分析预测该区域各类型的交通违法数量的时空演变。最后，达到分析违法行为时空演变并提出预防和整治违法行为措施的目的。

1.2.2 研究内容

本次基于机器学习的交通违法时空演化趋势预测的研究，主要内容有：

（1）文献与数据搜集整理。归纳交通违法研究和机器学习回归预测方面的研究现状和既有算法，总结出最合适违法预测的模型算法。并且根据模型需求，完成大量的原始数据融合处理。

（2）交通违法行为概述。描述并进一步定义交通违法的特征、分类。初步探索交通违法行为的影响因素，为模型搭建时的特征构造提供依据。

（3）交通事故类型与道路违法类型的关联分析。利用统计学中的多元对应分析的方法，构建模型探究各事故类型和违法类型之间的对应影响关系。

（4）回归预测机器学习模型概述。分析了共六个本文用到的机器学习模型的算法原理和各自优劣势，确认了最终预测所用的模型。

（5）搭建基于机器学习的融合回归预测模型。通过前期的特征工程和利用随机森林构造出的交通流特征，对违法数据的时空分布进行预测，并根据结果不断改进模型。

（6）交通违法行为预防对策研究。针对违法行为的分布以及预测模型中各特征的贡献度，提出预防和监管道路违法的管理措施。

1.3 研究方法及技术路线

本文的研究内容具有跨学科的特征属性，主要研究方法包括：交通违法行为的原理背景等概述部分主要运用到了交通工程学理论、道路安全理论等；事故违法类型对应分析运用到了统计学中的应用多元统计分析理论和 SPSS、R 语言等工具的运用；预测模型搭建部分运用到了各种机器学习算法理论，以及大量特征工程的技巧。

本文的技术路线如图 1-1 所示。

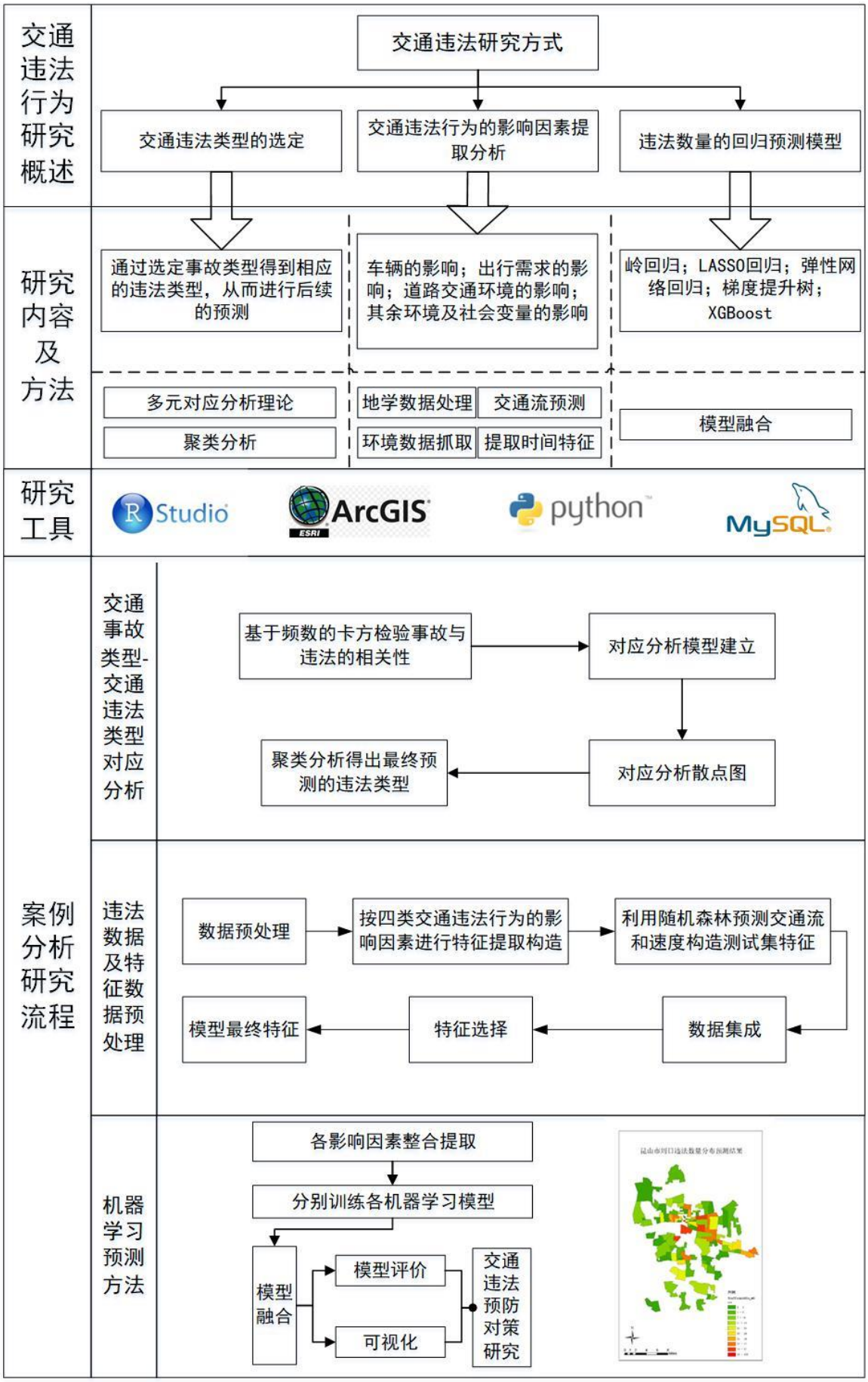

交通违法行为研究概述
交通违法研究方式
交通违法类型的选定
交通违法行为的影响因素提取分析
违法数量的回归预测模型
研究内容及方法
通过选定事故类型得到相应的违法类型，从而进行后续的预测
车辆的影响；出行需求的影响；道路交通环境的影响；其余环境及社会变量的影响
岭回归；LASSO回归；弹性网络回归；梯度提升树；XGBoost
多元对应分析理论
聚类分析
地学数据处理
交通流预测
环境数据抓取
提取时间特征
模型融合
研究工具
Studio
ArcGIS
python
MySQL
案例分析研究流程
交通事故类型-交通违法类型对应分析
基于频数的卡方检验事故与违法的相关性
对应分析模型建立
聚类分析得出最终预测的违法类型
对应分析散点图
违法数据及特征数据预处理
数据预处理
按四类交通违法行为的影响因素进行特征提取构造
利用随机森林预测交通流和速度构造测试集特征
模型最终特征
特征选择
数据集成
机器学习预测方法
各影响因素整合提取
分别训练各机器学习模型
模型融合
模型评价
可视化
交通违法预防对策研究

图 1-1 技术路线图

1.4 本文结构安排

本文通过对交通违法行为的现状和背景的阐述分析，提出研究课题，旨在以最终可以有效控制事故发生为目的，预测各类型违法发生量的时空分布。针对此课题，制定出了合理的研究方法和技术路线。全文共分为六章，组织安排如下：

第一章主要介绍了论文的研究背景和研究意义，针对研究目标，制定了相应的研究方法和技术路线。

第二章简要介绍了交通违法行为的基本概念，以及其影响因素，为构建特征集提供理论依据。然后介绍比较了几个机器学习模型，作为之后的预测基模型。

第三章介绍了事故类型和违法类型对应分析的模型建立方法，以及昆山市的分析结果。选定“伤人事故”作为目标改善事故类型，从而确定与其关联性最强的“违反交通标志标识”类型违法数据为本次研究的预测对象。

第四章详细介绍了本次实验准备内容，特征工程处理方法，以及利用随机森林模型预测交通流和速度构造测试集特征。最后阐述分析了实验训练过程和测试结果。

第五章基于第四章的预测结果，对违法行为的时空演化进行了可视化分析，并提出了相应的监管措施建议。

第六章对全文的工作进行总结，并提出了后续的研究改进方向。

第二章　国内外研究现状概述

本文主要针对交通违法行为的时空演化规律进行预测，为针对性的执法与监管提供科学依据，从而有效减少交通违法及事故。目前，针对交通违法行为及预测建模，国内外学者已经进行了大量的研究，并取得了一定的成果，本章将从交通违法行为研究和基于机器学习的预测模型研究两个方面对国内外研究现状进行归纳和叙述。

2.1　交通违法行为研究现状

道路交通违法行为是指行人、机动车驾驶员以及任何与道路交通活动有关的单位和个人，违反了《中华人民共和国道路交通安全法》及相关法律、法规以及国务院所属部委颁布的交通管理的规章，各省、直辖市、自治区制定的交通管理的地方性法规、地方性规章和各级政府及行政职能部门颁布的有关交通管理的规定、通知、通告和条例，同时对社会和道路交通环境产生了不良后果的交通行为[5]。

道路交通违法行为的特征主要体现在其对交通环境甚至社会各个方面的危害性上，而危害性又主要体现在三个方面[6]：

（1）导致交通秩序混乱。交通违法行为的发生一般都会直接导致侵占车道、处理违法事故现场等公共交通时空资源被浪费的现象，这会极大妨碍交通秩序和车辆畅通。

（2）影响市容，造成公害。

（3）影响交通安全，导致事故的发生。

并且，多项研究表明，在交通事故的人为因素中，除了少量无意识的危险驾驶行为，82%的属于驾驶者侥幸心理造成的有意识的危险驾驶行为[7]，这些危险驾驶行为极易导致驾驶者发生交通事故，而这类行为更多的记录在驾驶者历史交通违法中[8]。因此，交通违法是研究危险驾驶行为和各类交通事故间联系的重要切入点，其与交通事故的联系为交通事故分析、预防及控制提供新的视角。

在交通违法对事故影响方面，Mercedes Ayuso[9]等将交通事故轻重程度用经济损失来衡量，同时将各类交通违法对交通事故的影响大小折合成经济损失，使用多项式逻辑回归模型评估其影响大小，可以得到不同类型的交通违法对交通事故的影响程度，而不同种类的驾驶分心（打电话，交谈等）行为会造成不同类型的交通事故，对驾驶者惩罚及奖励测试[10]发现：经常违法的人没有意识到自己的交通违法行为会给自己带来危险。

目前，许多国家已采取在驾照上扣分的政策以规范人们的驾驶行为[11]。借助全新的行为控制理论[12]分析驾驶者交通违法行为，从控制交通违法行为的角度研究解决道路交通违法问题的方法，并在此基础上提出了解决交通违法问题的具体措施和政策建议。胡家兴等[13]人通过对交通违法行为的机理分析，从博弈论的角度对交通违法行为进行了探讨；根据事故因果连锁论，建立了交通违法导致交通事故的机理模型，分析了交通事故与交通违法行为之间的关系，指出了治理交通违法行为的工作重点。通过建立基于小波理论与最小二乘向量机的组合预测模型，为预测道路交通违法行为的发生提供了一种新的方法，对治理交通违法行为具有很强的实用性和借鉴意义，有助于交通安全决策预案的制定。其借鉴交通事故黑点分析方法，提出基于违法记分的当量违法总次数分析方法，能够快速准确地确定道路交通违法多发点段。通过对易引发交通事故的多发性交通违法行为的成因、危害及

治理对策的研究，从政府主导、公路执法、科技应用、宣传教育、集中治理等方面提出了强化交通违法整治的工作方法和途径。

而交通违法行为的影响因素也是很重要的一个研究方向。整个交通系统主要包括四大类：人、车、路和环境[14]；社会变量主要包括人口、用地属性等。而交通违法的影响因素研究也主要围绕这几方面展开。

交通违法的重要影响因素之一就是驾驶员：由于驾驶行为是根据生活水平和社会文化形成的习惯，它们自然与驾驶者的性别，年龄，受教育程度和收入等个人特征有关。具体而言，Sabey 和 Taylor[15]认为驾驶员的个人特征占导致事故发生的因素的 95%，在这些个人特征中，最重要的因素是年龄和性别。驾驶速度与驾驶员年龄呈负相关，与驾驶员的收入呈正相关；男性司机往往比女性司机开车更快；单婚姻状态的司机的平均驾驶速度比已婚司机的平均驾驶速度略快。Factor 等人[16]应用逻辑回归程序对以色列数据进行分析，得出结论认为，各种交通事故风险是由于不同的社会习惯和技术进步所致。无论驾驶员属于哪个社会部门，社会和文化特征都是影响交通安全的重要因素。他们发现，在以色列，穆斯林，分居，寡妇，男性，年轻，低技术工人和受过低等教育的司机的事故率往往较高。因此，驾驶员的年龄、性别、身体状况、心理素质都被认为是潜在的违法风险因素[17]。

交通违法的重要影响因素之二就是出行需求：驾驶过程就是一个交通系统和驾驶员之间“供需平衡”的过程，其中机动车驾驶员的驾驶需求主要包括安全需求等六方面的需求[18]，其中安全需求是驾驶员最基本的需求，但是在实际复杂的交通系统中，驾驶员的安全需求往往会妥协于其他需求如速度需求、时间需求等。这些需求的相对重要性影响了驾驶员对于交通违法行为的态度，进一步引发他们做出各种规范或不规范的行驶决策。出行需求可以具体定义为体现在出行时间的差异[19]，比如工作日与节假日，高峰时期与深夜，其二者的出行需求有明显区别。

交通违法的重要影响因素之三就是道路交通环境：研究表明，道路交通环境包括道路类型、路灯条件等是影响车辆交通违法和事故严重程度的重要因素[20]。影响机动车驾驶员违法行为的道路因素包括道路等级、道路结构、道路几何特征、路面所有附属设施。另外，道路交通流状况和路面车辆速度也是重要影响因素。复杂的交通流对驾驶员的驾驶水平有较高的要求，而拥挤的交通状况，较大的路面交通流量和缓慢的行驶速度会对驾驶员的心理状态有负面的影响。反之，特别畅通的道路状况和平稳不变的行驶速度可能会使驾驶员放松警惕，引发无意的交通违法行为。因此，交通流和车辆速度是影响违法行为的重要因素，并且影响模式复杂多变。

交通违法的重要影响因素之四就是其余环境因素及社会变量属性：其余环境因素包括八个方面：路灯状况，天气状况，能见度，事故发生在一周中的第几天，是否是公共假期，时间，季节和事故年份。比如：早晚高峰时期车流量较大，驾驶员对于时间、速度、停车的需求也会显著增强，在这种情形下，违反交通信号灯、乱变道、乱停车等行为会增多。而深夜时期，由于交通畅通，整体车辆超速的行为可能会显著增多。社会属性方面，土地使用类型等规划信息是研究关注的重点。土地使用类型包括商业用地、居住用地、建设用地等。地区的规划属性决定了该地区路面交通环境状况，比如建设用地就会存在大量道路被占用，车辆和行人无法正常通行的情况。因此，这些因素会对驾驶员和车辆都产生特定影响，从而间接引发交通违法行为的发生。

总结来说，目前针对交通违法的研究取得了一定的成果，但是缺乏系统性，尤其对于

一些实际的重点问题研究较少，包括：

（1）缺少交通事故类型与违法类型的量化对应关系研究。由于违法的种类众多，而执法资源与警力相对有限，因此需要确定与事故强相关的交通违法类型，优化交通执法对象与目标；

（2）目前尚缺乏对交通违法时空演化规律的宏观预测，因此无法为交通执法提供科学依据，使得目前主要执法策略依旧停留在经验层面，效率偏低。；

2.2 机器学习预测模型研究现状

交通违法行为的影响因素众多，因此传统的统计学预测模型很难精准捕捉多因素间的复杂耦合关系。机器学习作为一门专注于研究通过科学计算，利用得到的经验来提升系统自身性能的学科，尽管在某些应用方面的可解释性不如统计模型，但可以通过反复迭代学习发现隐藏在数据中的科学。并且，由于机器学习作用在真实的数据上，并不依赖于假设，相对于统计模型来说预测效果要明显更优[21]。

根据训练集是否拥有标签，机器学习一般可分为监督学习和非监督学习，其中监督学习以分类和回归作为代表[22]。LASSO 回归、随机森林、梯度提升决策树、XGBoost 都是常见的监督学习算法。

2.2.1 岭回归

（1）模型原理：

回归是用于建模和分析变量之间关系的一种技术，分析变量是如何影响结果的。线性回归是指完全由线性变量组成的回归模型。所有统计学习方法（机器学习方法）中最常用的回归预测模型便是多元线性回归模型，然而，当自变量之间存在高度相关性，即发生多重共线性时，通过该模型求出的自变量系数的绝对值容易过大，并且不一样的样本也会导致参数估计值变化非常大，即参数估计量的方差也增大，对参数的估计会不准确。因此，模型容易产生过拟合的现象[23]。

为了解决多元线性回归模型的特征高共线性所带来的影响，Hoerl 和 Knnard[24]提出了一个新的线性回归估计方法，岭回归。该方法通过在损失函数中引入正则化项λ（L1 范数）解决了两个问题：一是在通过正规方程方法求解 θ 的过程中出现的$X'X$不可逆的情况，二是线性回归出现的过拟合现象。岭回归模型求解对应的优化目标损失函数如下：

$$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)})-y^{(i)})^2+\lambda\sum_{j=1}^{m}\theta_j{}^2 \qquad (2\text{-}1)$$

其中，$h_\theta(x^{(i)})$为第 i 个样本的预测值，$y^{(i)}$为第 i 个样本的实际值，m 为样本数，θ为自变量对应系数，下同。λ称为正则化参数，用来控制系数θ的收缩量，λ越大，系数θ的收缩量也越大，如图所示，因此系数θ对变量共线现象的稳健性就会增强。

但是需要注意的是：如果λ选取过大，会把所有系数 θ 均最小化，造成欠拟合，如果λ选取过小，会导致对过拟合问题解决不当，因此λ的选取至关重要。

（2）模型特点

- 相较于最基本的多元线性回归，岭回归模型允许特征变量之间存在共线性问题
- 模型的稳定性较强，不容易过拟合

岭回归经常被用于对特征较多且存在共线性的数据集进行回归预测，比如涉及天气、环境等非实验性数据中[25]，于本文的违法预测来说，很多解释变量如“人口”和“该区域用地属性”存在共线性问题，岭回归可以解决这部分问题，并且提升模型的稳定性、精确性。

2.2.2 LASSO 回归

（1）模型原理

同样为了解决多元线性回归的多重共线性问题，Robert Tibshirani (1996)[26] 提出了 LASSO 模型，全称 Least Absolute Shrinkage and Selection Operator，该模型是在岭回归的基础上发展的。与岭回归一样，LASSO 也是通过在优化函数中增加一个正则项来进行有偏估计，从而减少模型方差和共线性的影响；但与岭回归不同的是，LASSO 构造的是一个一阶的惩罚函数（L1 范数），从而使得模型一些变量的系数为 0，这个特性可以帮助我们更好地理解数据，并且让程序自动实现变量剔除。这是 L1 范数的一个非常有用的属性，而 L2 范数不具有这种特性。这实际上因为 L1 范数倾向于产生稀疏系数。LASSO 回归的优化目标损失函数为：

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{m}|\theta_j| \qquad (2\text{-}2)$$

（2）模型特点

- LASSO 回归同样能解决特征共线性问题，并且减少模型过拟合的可能性。
- LASSO 模型可以自动剔除变量，非常适用于特征变量特别多的分析场景
- 由于 L1 范数的解具有稀疏性，这使得它可以和稀疏算法一起使用，计算效率很高。
- 模型的稀疏性可能会导致过多的特征被剔除，降低了模型的拟合效果

对于本文的违法预测来说，LASSO 回归模型可以同时解决共线性和过拟合的问题，并且剔除大量贡献较小的特征，使得模型在不牺牲准确性的前提下变得更简化和更清晰了。同时还能筛选出最重要的特征供后续分析。然而，可能存在模型被剔除的特征过多的情况，我们期望能中和一下 LASSO 回归与岭回归的模型效果。

2.2.3 弹性网络回归

（1）模型原理

弹性网络本质上是岭回归和 LASSO 回归的组合，也就是同时使用 L1 和 L2 正则项的线性回归模型，这种组合大多用于权重非零系数较少的稀疏模型。当多个特征和另一个特征相关的时候弹性网络模型效果较好，LASSO 倾向于随机选择其中一个，而弹性网络更倾向于选择两个。另外，在实践中，弹性网络模型在 LASSO 和 Ridge 之间权衡的一个优势是它允许在循环过程（Under rotate）中继承 Ridge 的稳定性[27]。弹性网络回归的优化目标损失函数为：

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda_1\sum_{j=1}^{m}{\theta_j}^2 + \lambda_2\sum_{j=1}^{m}|\theta_j| \qquad (2\text{-}3)$$

（2）模型特点

- 模型继承了 LASSO 回归和岭回归的所有优点，并且更稳定，变量选取也更合理。

对于本文的违法预测来说，弹性网络回归可以较好的拟合训练数据，还可以通过弹性

网络回归模型或者 LASSO 回归模型筛选重要特征。

2.2.4 梯度提升决策树 GBDT(Gradient Boosting Decision Tree)

为了进一步挖掘各特征变量与违法数据之间更为复杂的非线性关系，从而更准确的预测违法数据，本文进一步采用基于梯度提升原理的决策树回归模型。该模型的输入是训练集，输出是 n 棵 CART 回归决策树的组成。

（1）CART 回归模型原理

决策树（decision tree）是一个树结构（可以是二叉树或非二叉树）的预测模型，代表的是一种对象特征属性与对象目标值之间的映射关系。其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性在某个值域上的输出，而每个叶节点存放一个类别。使用决策树进行决策的过程就是从根节点开始，测试待分类项中相应的特征属性，并按照其值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果[28]。

CART 决策树（Classification and Regression Tree）是决策树的一种实现，常被用来进行回归预测。其主要思想和特点在于：将样本按基尼系数递归划分完成建树过程，生成结构简洁的二叉树，使得生成的任意一个非叶子节点都只存在两个分支。最后采用后剪枝的方法防止模型过拟合。

（2）GBDT 模型原理

为提高 CART 决策树的预测准确度，本文引入基于 CART 的梯度提升（Gradient Boosting）方法。它的主要思想是：通过不断更新训练样本的权重，迭代学习多个弱分类器也就是上述的 CART 决策树，并使用某种方式将弱分类器组合集成，达到最好的预测结果。而每次迭代建立新模型是在之前建立的 CART 模型损失函数的梯度下降方向，也就是利用了原先违法预测的损失函数在当前模型输出的梯度值作为回归提升树算法的残差值来拟合下一组回归树。从而不断降低损失函数值，提升模型性能。

（3）GBDT 模型特点

• 通过每一次迭代的残差计算，变相增大了训练集中预测不精确的样本的权重，使得模型准确性大大提升。

• 由于过于关注预测结果较差的样本，模型可能会过拟合

该模型具有分类、回归预测速度快，模型容易可视化、解释性强，拟合准确度高等特点，并且可以通过剪枝防止其过拟合，非常适用于本文的违法数据预测。

2.2.5 XGBoost 算法

为了解决 4.5 中梯度提升回归树存在的缺陷，以及提升树模型的迭代计算速度，本文使用了陈天奇提出的 XGBoost 算法框架[29]。

XGBoost 算法是由 GBDT 改进而来，相对于 GBDT 算法有两个不同点：1. XGBoost 在函数空间中利用了牛顿法而不是梯度下降法进行优化，这使得模型计算效率更高了。2. XGBoost 引入了 L1、L2 正则项，降低了每棵回归树的复杂度，从而有效防止模型过拟合。3. 可自动利用 CPU 的多线程并行计算，从而大大提高运算速度。

因此，XGBoost 在违法预测中，无论是模型准确度还是稳健性都有极大的优势。

第三章　事故类型和违法类型对应分析

从不同的方面系统地分析交通安全数据，将是未来几年中国面临的紧迫挑战。多项研究表明，道路违规行为是交通事故最重要的影响因素之一，但过往的研究只停留在探究“是否交通违法”这一变量对交通事故的影响，缺乏系统化且定量的不同事故与不同违法类型之间的影响模型分析。本文为了针对所需要关注的交通事故类型而预测相应需关注的交通违法行为，进行分析得到事故与违法类型之间的关联性。

由于事故与违法数据都是分类变量，根据分类型数据的特点，很多基于均值、方差和标准差的分析方法就不太适用了，通常使用的关联性分析方法是基于频数的卡方检验和逻辑回归等。然而面对变量数量较多的情况，卡方检验等方法的分析结果就不太精确了。

因此，本文先通过χ^2检验事故与违法类型两者之间的关联性，并提出进一步利用对应分析（Correspondence Analysis）的方法深入挖掘出两者的对应影响关系。

对应分析（CA），也称 R-Q 分析，是在因子分析的基础上发展起来的一种多元统计分析方法[30]，适用于探索定型变量（或分类数据）之间的相关关系。该方法原理为将一个联列表的行和列中各元素，被称为样品和变量，也就是本文中所指的交通违法数据和事故数据的比例结构以点的形式在较低维的空间中表示出来，以使行和列点的位置与其在表格中的关联一致，目标是对解释有用的数据进行全局观察。它最大特点是能把众多的样品和众多的变量同时作到同一张图解上，将样品的大类及其属性在图上直观而又明了地表示出来，非常直观。另外，它还省去了因子选择和因子轴旋转等复杂的数学运算及中间过程，可以从因子载荷图上对样品进行直观的分类，而且能够指示分类的主要参数（主因子）以及分类的依据，是一种直观、简单、方便的多元统计方法。

3.1　对应分析模型建立

步骤一、事故及违法数据的结构表示

将事故类型和违法类型作为两个属性变量，事故类型考虑$\mathrm{A}=(a_1,a_2,\dots,a_i,\dots a_n)$共n个类目，违法类型考虑$\mathrm{B}=(b_1,b_2,\dots,b_j,\dots b_p)$共p个类目。可得曾发生各类交通事故的车辆驾驶员的所有历史违法统计矩阵为$\mathrm{X}=\begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & x_{ij} & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$。其中，$\mathrm{i}=1,2,\dots,\mathrm{n};\ \mathrm{j}=1,2,\dots,\mathrm{p};$ x_{ij}为发生过a_i类事故的所有驾驶员犯b_j类违法事故的统计次数[31]。

步骤二、对应分析的数据阵变换得到Z

（1）由原始交通事故-违法分布数据阵X出发，分别按行和列求和，T为矩阵总和：

$x_{11}\ x_{12}\ \dots\ x_{1p}$	$\sum_{k=1}^{p} x_{1k} = X_{1\cdot}$
$x_{21}\ x_{22}\ \dots\ x_{2p}$	$\sum_{k=1}^{p} x_{2k} = X_{2\cdot}$
$\vdots\ \ \vdots\ \ \vdots$	$\vdots$
$x_{n1}\ x_{n2}\ \dots\ x_{np}$	$\sum_{k=1}^{p} x_{nk} = X_{n\cdot}$
$X_{\cdot 1}\ X_{\cdot 2}\ \dots\ X_{\cdot 3}$	$\sum_{l=1}^{n}\sum_{k=1}^{p} x_{lk} = X_{\cdot\cdot} \stackrel{\text{def}}{=} T$

（3-1）

（2）化矩阵 X 为规格化的“概率”矩阵 P，并称之为对应阵，令

$$\mathrm{P} = \frac{1}{T}X \stackrel{\text{def}}{=} (p_{ij})_{n\times p} \qquad (3\text{-}2)$$

其中，$p_{ij} = \frac{1}{T}x_{ij}(i = 1, \dots, n; j = 1, \dots, p)$，可理解为数据$x_{ij}$出现的概率。$P_{\cdot j} = \sum_{i=1}^{n} p_{ij}$可理解为第$j$个变量的边缘概率（$j = 1, \dots, p$）；$P_{\cdot i} = \sum_{j=1}^{p} p_{ij}$可理解为第$i$个样品的边缘概率（$i = 1, \dots, n$）。

（3）为消除原表中事故和违法各变量量纲和数量级不同等影响，利用以上得到的结果，对数据矩阵进行对应变换，也就是中心化和标准化处理，令新矩阵

$$\mathrm{Z} = (z_{ij})_{n*p} \qquad (3\text{-}3)$$

其中

$$z_{ij} = \frac{p_{ij} - P_{i\cdot}P_{\cdot j}}{\sqrt{P_{i\cdot}P_{\cdot j}}} = \frac{x_{ij} - \frac{X_{i\cdot}X_{\cdot j}}{T}}{\sqrt{X_{i\cdot}X_{\cdot j}}} \quad (\mathrm{i} = 1, \dots, \mathrm{n}; \mathrm{j} = 1, \dots, \mathrm{p}) \qquad (3\text{-}4)$$

步骤三、计算联列表的行轮廓与列轮廓分布

行轮廓矩阵 R 由原矩阵X的每一行除以各行总和得到，列轮廓矩阵C由原矩阵X的每一列除以各列总和得到。计算轮廓分布的目的在于分别消除行变量和列变量各事故、违法类型出现的“概率”不同的影响。

$$\mathrm{R} = (\frac{x_{ij}}{X_{i\cdot}})_{n*p} = (\frac{p_{ij}}{P_{i\cdot}})_{n*p} = {D_r}^{-1}P \stackrel{\text{def}}{=} [{R_1}' \dots {R_n}']^T \qquad (3\text{-}5)$$

$$\mathrm{C} = (\frac{x_{ij}}{X_{\cdot j}})_{n*p} = (\frac{p_{ij}}{P_{\cdot j}})_{n*p} = \mathrm{P}{D_c}^{-1} \stackrel{\text{def}}{=} (C_1, \dots, C_p) \qquad (3\text{-}6)$$

步骤四、计算总惯量

计算该联列表的总惯量 Q，首先要得到各行之间的平方距离（又称加权距离），比如第k与第l行之间的平方距离为：

$$D^2(k,l)=\sum_{j=1}^{p}\frac{\left(\frac{p_{kj}}{P_{k\cdot}}-\frac{p_{lj}}{P_{l\cdot}}\right)^2}{P_{\cdot j}}=(R_k-R_l)^{'}\ D_c{}^{-1}(R_k-R_l) \qquad (3\text{-}7)$$

然后，将所有行到平均行轮廓c的平方距离总和定义为该联列表的总惯量Q:

$$\mathrm{Q}=\sum_{i=1}^{n}P_{i\cdot}D^2\,(i,c)=\sum_{i=1}^{n}P_{i\cdot}\sum_{j=1}^{p}\frac{1}{P_{\cdot j}}\left(\frac{P_{ij}}{P_{i\cdot}}-P_{\cdot j}\right)^2$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{p}z_{ij}{}^2=\chi^2/T \qquad (3\text{-}8)$$

步骤五、对新矩阵Z作奇异值分解

$$\mathrm{Z}=U_1\Lambda_m{V_1}',\quad m=rank(Z)\le\min(n-1,p-1) \qquad (3\text{-}9)$$

其中

$$\Lambda_m=diag(d_1,\dots,d_m),\quad {V_1}'V_1=I_m,\ {U_1}'U_1=I_m \qquad (3\text{-}10)$$

求 Z 的奇异值分解（SVD）是通过求解交通事故-违法的协方差矩阵$S_R=Z'Z$的特征值和特征向量得到的。对于$\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_m>0$，相应的特征向量为$v_1,v_2,\dots,v_m$，其中m为所有公共因子即维度的个数。在本文以及其他的实际应用中[31]，都按照特征值累计贡献率$\frac{\lambda_1+\lambda_2+\cdots+\lambda_l}{\lambda_1+\cdots+\lambda_l+\cdots+\lambda_m}\ge 0.80$ (或 0.70，或 0.85)来确定所取公共因子的个数$l(l\le m)$，也就是主要维度的个数。

步骤六、计算行轮廓和列轮廓的坐标矩阵

计算行轮廓的坐标矩阵，即违法类型的坐标矩阵 G；以及列轮廓的坐标矩阵，即事故类型的坐标矩阵 F。具体计算方法如下：

令

$$a_i=D_c{}^{-1/2}v_i \qquad (3\text{-}11)$$

则该表的列轮廓坐标F（又称 R 形因子载荷矩阵）为：

$$\mathrm{F}=(d_1a_1,d_2a_2,\dots,d_ma_m)=D_c{}^{-1/2}V_1\Lambda_m$$

$$=\begin{bmatrix}\frac{d_1}{\sqrt{P_{\cdot 1}}}v_{11} & \frac{d_2}{\sqrt{P_{\cdot 1}}}v_{12} & \cdots & \frac{d_m}{\sqrt{P_{\cdot 1}}}v_{1m}\\ \frac{d_1}{\sqrt{P_{\cdot 2}}}v_{21} & \frac{d_2}{\sqrt{P_{\cdot 2}}}v_{22} & \cdots & \frac{d_m}{\sqrt{P_{\cdot 2}}}v_{2m}\\ \vdots & \vdots & & \vdots\\ \frac{d_1}{\sqrt{P_{\cdot p}}}v_{p1} & \frac{d_2}{\sqrt{P_{\cdot p}}}v_{p2} & \cdots & \frac{d_m}{\sqrt{P_{\cdot p}}}v_{pm}\end{bmatrix}, \qquad (3\text{-}12)$$

其中，$D_c{}^{-1/2}$为 p 阶矩阵，V_1为 p*m 矩阵。

该表的列轮廓坐标G（又称 Q 形因子载荷矩阵）为：

$$G=(d_1b_1,d_2b_2,\cdots,d_mb_m)=D_r^{-1/2}U_1\Lambda_m$$
$$=\begin{bmatrix}\frac{d_1}{\sqrt{P_{1\cdot}}}u_{11} & \frac{d_2}{\sqrt{P_{1\cdot}}}u_{12} & \cdots & \frac{d_m}{\sqrt{P_{1\cdot}}}u_{1m}\\ \frac{d_1}{\sqrt{P_{2\cdot}}}u_{21} & \frac{d_2}{\sqrt{P_{2\cdot}}}u_{22} & \cdots & \frac{d_m}{\sqrt{P_{2\cdot}}}u_{2m}\\ \vdots & \vdots & & \vdots\\ \frac{d_1}{\sqrt{P_{n\cdot}}}u_{n1} & \frac{d_2}{\sqrt{P_{n\cdot}}}u_{n2} & \cdots & \frac{d_m}{\sqrt{P_{n\cdot}}}u_{nm}\end{bmatrix}, \quad (3\text{-}13)$$

其中，$D_r^{-1/2}$为 n 阶矩阵，U_1为 $n*m$ 矩阵。

步骤七、事故-违法数据对应分析结果可视化以及类别间相似度计算

由于事故类型和违法类型都被映射到相同的 m 维因子轴上，因此可以求得行-行之间、行-列之间、列-列之间的距离。其中，为定义事故类型a_i和违法类型b_j间的相关度σ_{ij}，需同时考虑两者在某一主要解释维度η上的距离τ和该维度η对该类别的惯量$\mathrm{d}_{i\eta}$和$d_{j\eta}$（维度对点的惯量表示分类变量中每个类别在某个维度中所占信息的比例），本文提出类别间的相似度计算方法为：

$$\sigma_{ij}=\sum_{\eta=1}^{l}(1-\tau)(\mathrm{d}_{i\eta}+\mathrm{d}_{j\eta})/2 \quad (3\text{-}14)$$

事故类型和违法类型也可通过在主要维度构成的二维平面中进行可视化，从而直观的判断相近的类别。

3.2 数据准备

3.2.1 事故类型定义

本文选取昆山市 2017 全年的交通违法数据和交通事故数据，基于各自的车牌号进行数据连接，构建事故类型-违法类型的交叉频数表，从而进行对应分析。

其中，定义的事故类型与对应编号如表 3-1。

表 3-1 事故类型编号

事故类型	模型中使用的编号
伤人事故	1
死亡事故	2
财产损失事故	3

3.2.2 违法类型定义

交通违法行为基于过往理论可以用多种维度进行分类，如按照交通违法行为的主体、行为主体的主观状态、路权理论等进行分类。然而，由于本文的研究目的为预测交通违法行为的发生，考虑从违法行为的内在机理、产生原因来分类将会更有益于后续的研究过程。

原始违法数据集将交通违法行为细分为 60 小类，笔者按照违法属性和潜在的产生原因，将其大致归纳为 20 类，如表 3-2 所示。

表 3-2 违法类别编号

违法类型	原始违法数据库中所对应的类型编号	模型中使用的编号
货车超速	13523,17291,17292,17271,13511,16291,16311,1633,1723,17272,1635,1636,1725,13501；	1
客车超速	17262,17261,17281,17212,1722,17282,16281,16282,16301,1632,1634,17211,1724,13491；	2
违反标志指示	13441,1090,13451;	3
违反交通信号灯指示通行	16251,16252,1208;	4
车辆外型问题	10832,1718;	5
机动车驾驶人未系安全带	6011;	6
驾驶时拨打接听手持电话的	1223;	7
违规停放	10393,70053;	8
机动车逆向行驶的	1301;	9
遇前方机动车停车排队等候或缓慢行驶时不合规行为	12432,12433,12434,12435,12436,12431,10252;	10
通过路口遇停止信号时，停在停止线或路口以内	1211;	11
不按规定倒车	1074;	12
机动车不在机动车道内行驶的	1018;	13
遇行人正在通过人行横道时未停车让行的	1357,1358,1356;	14
通过路口向右转弯遇同车道内有车等候放行信号，不依次等候	1212;	15
变更车道影响正常行驶机动车	1043;	16
在禁止掉头标志标线的地点掉头	10441;	17
网状线区域内停车	10254;	18
掉头妨碍正常行驶车辆、行人	1046;	19
在禁止左转弯掉头标志标线的地点掉头	10442	20

3.3 实例分析

本文利用 R 语言，以及"gplots", "reshape2"等工具包以及 SPSS 软件对昆山市交通事故与历史违法数据进行对应分析，得到对各交通事故类型影响最大的违法类型。操作方法为：

首先，利用卡方检验确定事故类型与违法类型存在显著相关性，然后进一步使用对应分析（CA），详细分析出事故类型与违法类型之间的对应相关关系，针对所要关注的事故类型，如“伤人事故”，得到对其影响最显著的违法类型。

3.3.1 基于频数的卡方检验事故与违法的相关性

当联列表结构较大时，无法通过直观的可视化结果来判断行列之间的相关性，需要通过某些统计方法来定量判断行列变量之间的相关性。常用的联列表独立性检验方法为卡方检验，卡方检验是一种统计量的分布在零假设成立时近似服从卡方分布的假设检验，被用于检验表中的行变量与列变量之间是否显著相关[32]。

卡方检验的实质是将实际的频数分析与期望频数作对比，如果差距很大，超过界限值，那么就可以认为组成交叉表的两个

分类变量之间具有相关性。假设检验的过程为建立假设：

H_0：两个分类的变量之间是独立的

H_1：两个分类的变量之间是不独立的

检验的结果若是接受H_0就说明不能推翻两个分类的变量是独立的假设：反之，拒绝H_0接受H_1就说明它们之间是不独立的，检验的统计量就是χ^2[33]。

具体检验方法与结果如下：

（1）对于表中每一个单元，计算其期望值e；并计算出检验值χ^2

$$\chi^2 = \sum \frac{(o-e)^2}{e} \qquad (3\text{-}15)$$

其中，e是每单元的期望值，o是每单元的观测值。

（2）将得到的χ^2与统计表中在$\mathrm{df} = (\mathrm{r}-1)(\mathrm{c}-1)$的自由度和$\mathrm{p}-\mathrm{value} = 0.05$下查询得到的关键值进行比较，其中r 为联列表的行数，c 为联列表的列数。若χ^2 >关键值，则说明行向量与列向量有显著的相关性。

（3）对于本文所用的事故-违法数据表，卡方检验结果为：p-value=0.01958。由结果可得，事故与违法类型显著相关。但需要注意，尽管卡方检验可以建立事故与违法行为的相关性，但具体对应的相关关系却无法直接得出。因此，需通过对应分析进一步分析各事故和违法类型之间的对应关系。

3.3.2 对应分析结果分析

本节将利用对应分析，按照 3.2.1 的方法构建模型，得到各事故类型和违法类型在不同维度的坐标，并且在同一个低维空间中作可视化展示，然后计算得到各类型之间的距离/相似度，从而进行聚类分析，直观的判断事故与违法类别之间的关联。

通过 R 语言与 SPSS 进行对应分析处理，具体结果如下：

（1）对应分析摘要表

表 3-3 对应分析摘要表

维数	奇异值	惯量	卡方	Sig.	惯量比例		置信奇异值	
					解释	累计	标准差	相关 2
1	0.115	0.013			0.679	0.679	0.024	0.177
2	0.079	0.006			0.321	1.000	0.030	
总计		0.020	35.100	0.020[a]	1.000	1.000		

结果显示，通过标准化残差矩阵，将原数据降到了 2 维空间，总共提取了 2 个公因子，其累计惯量达到 100%，可以解释原事故-违法表中所有信息，信息的全面度达到要求。

（2）行/列点总览（变量分类降维表）（部分内容节选）

表 3-4 行点总览表

违法类型	质量	维中的得分		惯量	贡献				
		1	2		点对维惯量		维对点惯量		
					1	2	1	2	总计
1	0.080	0.786	-0.524	0.007	0.430	0.278	0.766	0.234	1.000
2	0.003	-1.595	-0.235	0.001	0.074	0.002	0.985	0.015	1.000
3	0.489	-0.135	-0.019	0.001	0.077	0.002	0.986	0.014	1.000
4	0.248	0.112	0.167	0.001	0.027	0.087	0.395	0.605	1.000
5	0.000								
6	0.108	-0.296	-0.066	0.001	0.082	0.006	0.967	0.033	1.000
7	0.018	0.692	0.397	0.001	0.077	0.037	0.816	0.184	1.000
8	0.001	-1.595	-0.235	0.000	0.012	0.000	0.985	0.015	1.000
9	0.000								
10	0.010	0.919	2.021	0.004	0.074	0.518	0.231	0.769	1.000
11	0.016	-0.658	0.121	0.001	0.061	0.003	0.977	0.023	1.000
12	0.000								
13	0.000								
14	0.020	0.433	-0.357	0.001	0.032	0.031	0.681	0.319	1.000
15	0.000								
16	0.007	-0.965	-0.647	0.001	0.054	0.035	0.764	0.236	1.000
有效总计	1.000			0.020	1.000	1.000			

a.对称标准化

表 3-5 列点总览表

事故类型	质量	维中的得分		惯量	贡献				
		1	2		点对维惯量		维对点惯量		
					1	2	1	2	总计
1	0.771	-0.184	-0.019	0.003	0.226	0.003	0.993	0.007	1.000
2	0.138	0.688	-0.410	0.009	0.567	0.294	0803	0.197	1.000
3	0.091	0.511	0.783	0007	0.206	0.703	0.383	0.617	1.000
有效总计	1.000			0.020	1.000	1.000			

a.对称标准化

这两张表给出了事故和违法的每个类别在这两个维度上的坐标值（维中的得分）以及相应的惯量。维数对点的惯量含义为各违法或事故类型在该维度所占信息的比例，比如事故类型 2“死亡事故”的信息量有 80.3%分布在第 1 维。两个维度总共解释“死亡事故”类型所有信息量。

（3）对应分析散点图

如对应图所示，总体观察，可以发现某些事故类型和违法类型的距离较近，比如四五事故（2）和货车超速（1），这代表这两者有很大的相关性。同样的，还有“伤人事故”（1）和违法标志指示（3），财产损失事故（3）和驾驶时拨打接听手持电话（7）。

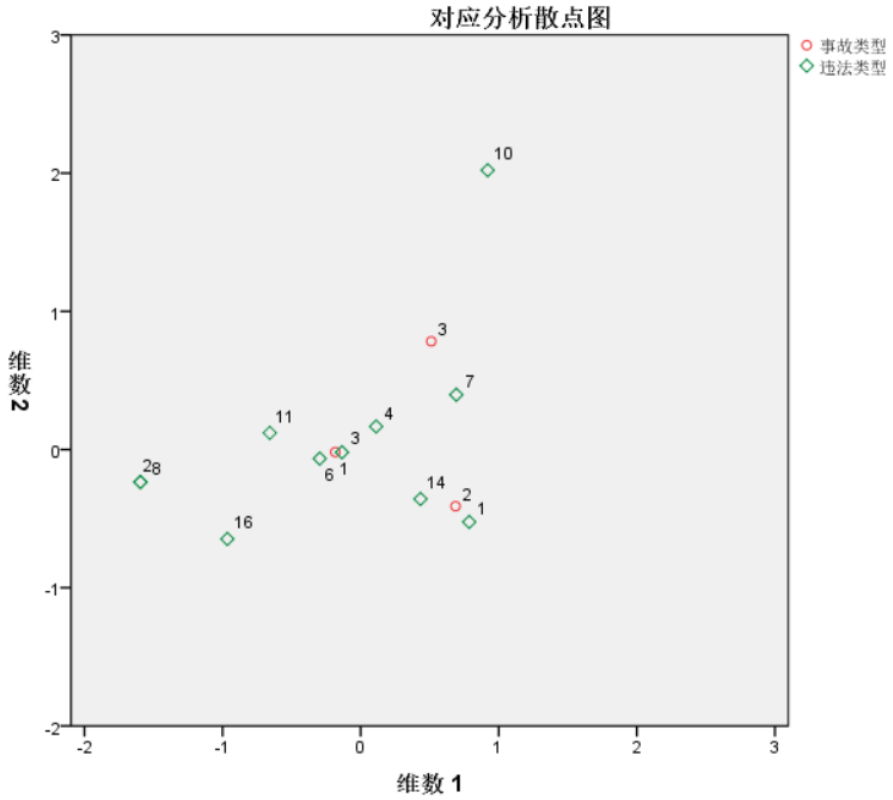


图 3-1 对应分析散点图

（4）聚类分析

由于对应图只能作一个直观上的展示，缺乏全面的量化总结。本文提出基于 K-means 的聚类方法，以 3.1.7 节提出的类别距离衡量方式σ_{ij}作为点之间的距离函数，对所有的事故、违法类别进行聚类，选出各种事故发生所对应的违法影响因素。

表 3-6 K-means 聚类算法

K-means 聚类算法
1：随机选取 k 个聚类质心点（cluster centroids）为$\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$
2：重复下面过程直到收敛{
3：　　对于每一个样例 i，计算其应该属于的类
4：　　$c^{(i)} \coloneqq \operatorname{argmin}_j \sigma_{ij}$
5：　　对于每一个类 j，重新计算该类的质心
6：　　$\mu \coloneqq \frac{\sum_{i=1}^{m} 1\{c^{(i)}=j\}x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)}=j\}}$
7：}

其中，k 是我们事先给定的聚类数，$c^{(i)}$代表样例 i 与 k 个类中距离最近的那个类，$c^{(i)}$的值是 1 至 k 中的一个。质心μ_j代表我们对属于同一个类的样本中心点的猜测。

最后选取聚类数为 4 时，得到的聚类结果如图 3-2 所示：

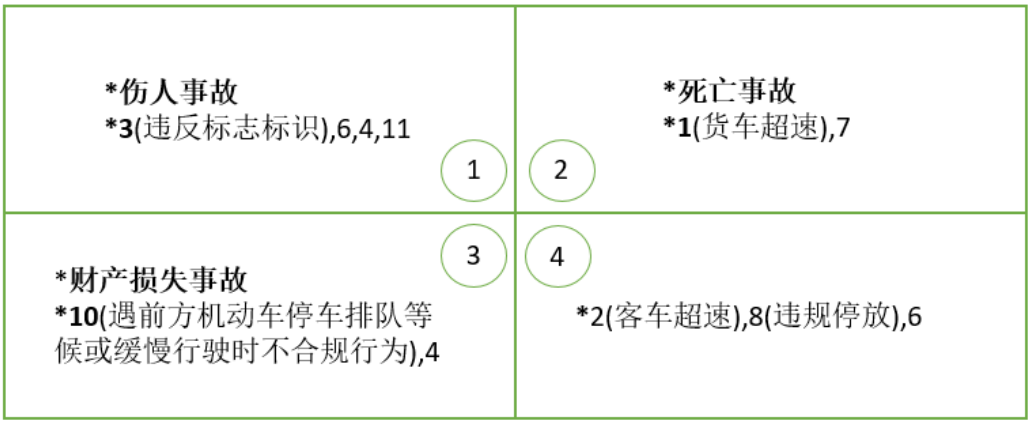


图 3-2 事故类型与违法类型聚类结果

以第一类聚类结果举例，可发现，事故类型“伤人事故”与编号为 3,6,4,11 的违法类型相关性很强，尤其是“违反标志标识”类违法。

因此，如果要控制伤人事故的发生，严加监测和预防“违反标志标识”的交通行为便可起到很大的效果。本文假设需要控制的就是“伤人事故”，因此，下文预测的是便是类型为“3-违反标志标识”的违法数据时空分布。

第四章　违法预测的实验与结果分析

本文采用 2017 年的昆山市各交通区域的城市规划信息，路网信息，人口等社会经济变量，微波数据得到的交通量、车速信息，以及监控设备得到的车辆违法信息作为训练集和测试集，利用随机森林模型构造测试集特征，通过融合岭回归、LASSO 回归、ENet 回归、梯度提升回归树、XGboost 模型和基于聚类的分层融合模型等，对昆山市不同区域不同时间段的交通违法数量进行预测。实验证明：该融合模型相比于传统的单一的线性回归模型，在精度上有明显提升，可以较为准确的预测某地区未来短时内的违法数量时空分布，并且给出影响违法分布的重要特征；从而验证了本文模型的价值。

4.1　实验准备

4.1.1　实验环境

本文实验环境为 Intel(R) Core(TM) i5-4200H CPU @ 2.80GHz 的 CPU，内存为 12.0GB，64 位操作系统。编程工具为 Jupyter Notebook，编程语言为 Python，使用的库和工具包主要包括 numpy, pandas, matplotlib, seaborn, scipy, sklearn, xgboost, lightgbm 等。具体环境如表 4-1 所示。

表 4-1 实验环境

实验环境	具体配置
操作系统	X64
CPU(处理器)	Intel(R) Core(TM) i5-4200H CPU @ 2.80GHz
编程工具	Jupyter Notebook
编程语言	Python
工具包	numpy, pandas, matplotlib, seaborn, scipy, sklearn, xgboost

4.1.2　实验原始数据

实验数据共分为 3 个来源：

来源一：交通流量、车辆速度数据

本实验使用的交通流量和车辆速度数据是中国江苏省昆山市部分区域的微波检测数据。本文选取 326 个检测器，包含 266 个路段，覆盖 83 个交通中区，由于计算资源有限，只使用五月数据作为训练集，六月第一周作为测试集。每条样本数据包括设备编号、采集日期、采集时间、流量、速度等 7 个数据字段以及每个设备的经纬度信息。通过 MySQL 处理，给微波数据加上 hourperiod（当天的第几个小时）和 weekday（一周第几天）两个时间属性，根据所需探究的不同时间粒度，将 5 月数据共 118,815,800 条数据作为训练集的一部分特征，将 6 月第一周共 20,833,015 条数据作为测试集的一部分特征，数据如附录 A 所示。

来源二：昆山规划数据集

本实验使用的规划、人口、经济变量来自昆山交通规划辅助决策平台。本文初步选取并构造了共 297 个中区(字段 MidTZoneId)的规划、人口、经济变量。由于昆山交通规划辅

助决策平台数据库中原始的规划数据是以交通小区为最小单位的，每条交通小区样本的有效数据包括小区总人口数、就业岗位数、用地面积、用地性质、机动车停车位等 13 个字段，如表 4-3 所示。通过数据处理（详见 4.1.2），将交通小区规划数据聚合为中区的规划信息，作为训练集和测试集的一部分重要特征。数据如附录 A 所示。

来源三：2017 年非现场执法数据

本实验采用江苏省昆山市非现场执法设备采集到的 2017 全年道路违法数据共 483569 条，选取其中 2017-05-01-00:00:00 至 2017-06-07-23:59:59 的违法数据作为训练集和测试集。每条样本数据包括违法车辆类型、设备编号、违法时间、违法类型等 8 条字段，数据如附录 A 所示。本文主要选取违法类型为“13441 机动车违反禁令标志指示”、“1090 机动车违反警告标志指示”和“13451 机动车违反禁止标线指示”的记录，并用时间和交通中区单位聚合得到本实验的预测目标——单位时间内某区域的“违法交通标志标识”违法总数。

4.2 特征工程

本实验的特征工程部分主要包括了数据清洗、特征提取（Feature Extraction）、特征构造（Feature Construction）、数据集成（Ensemble）、特征选择（Feature Selection）。

4.2.1 数据预处理

在实际研究应用中，我们的数据往往会有缺失值、异常值等，因此在对数据进行进一步特征处理前，需要先预处理清洗数据，主要内容包括处理缺失值、数据编码、数据标准化等。以下分别介绍本实验预处理阶段的几项主要工作：

（1）处理缺失值

处理缺失值一般采用的流程是先尽可能的利用均值法、建模法等方法或按照实际情况拟合缺失值。

例如：微波数据集中，交通流和速度属性为空值的样本自动将其填补为零值；规划数据集中，部分区域的总人口数为“空值/0/1”，这些空值或异常值可以用建模预测的方法来纠正，即：将缺失的人口数作为预测目标，利用该地区的其他正常规划属性比如地区面积、工作岗位数等，调用随机森林回归模型来拟合输出值。

但是当特征的缺失值达到近一半时，则无法填补该值，只能删除。

（2）数据编码

由于分类变量无法被模型直接处理，因此需要将其转化为哑变量。哑变量是非常有用的，因为它们使我们能够使用一个单一的回归方程来表示多组。这意味着我们不需要为每个亚组设置独立的方程模型。

例如：规划数据集中的特殊建筑属性（Remark）和天气情况就属于分类变量，在这里我们利用 sklearn 的 get_dummy 函数对其编码。若当时正在下雨，则“if_rain”字段为 1，反之为 0。若该区域有学校，则“school_yes”字段为 1；若该区域有大型市场，则“market_yes”字段为 1，以此类推。最后的部分数据结构见表 4-2。

表 4-2 部分哑变量数据结构

MidTZoneId	school_yes	hospital_yes	market_yes	biz_yes	res_yes	if_rain
JTZQ101	1	0	0	0	1	1
JTZQ103	0	0	1	1	1	0
JTZQ108	1	1	0	0	0	0
JTZQ109	0	0	0	1	1	0
JTZQ110	0	0	0	0	0	1

（3）数据标准化

数据标准化（normalization）是指将数据按比例缩放，使其落入一个特定区间内，在某些比较和评价的指标处理中经常会用到，去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。其中最常用的就是数据的归一化处理，统一将数据缩放到[0,1]区间内。

归一化有两个优势，第一个是提升模型的收敛速度；第二个是提升模型的准确度，由于各属性通常具有不同的量纲和数量级，当属性之间相差很大时，如果直接用原始数值进行分析建模，就会突出数值较高的指标在综合分析中的作用。因此，为了保证结果的可靠性以及防止模型求解过程中的梯度爆炸等问题，需要对原始数据进行标准化（归一化）处理。归一化处理的具体步骤如下：

对序列$x_1, x_2, \dots, x_n$进行变换：

$$y_i = \frac{x_i - \min\limits_{1\le j\le n}\{x_j\}}{\min\limits_{1\le j\le n}\{x_j\} - \min\limits_{1\le j\le n}\{x_j\}} \qquad (4\text{-}1)$$

则新序列$y_1, y_2, \dots, y_n \in [0,1]$且无量纲。

4.2.2 特征提取与构造

在数据挖掘中，使用合适的特征可以减少数据冗余度，提升模型的准确度。本文 2.2 节已经从交通参与者、车辆、出行需求、道路交通环境、其余环境因素及社会变量等五个方面详细分析了交通违法行为产生的可能影响因素。本节将从以上这些方面，人为的从原始数据集中提取并构造出有效特征来预测交通违法行为。

（1）车辆相关部分的特征集是以每个中区聚合，分别得到该地区注册的各种类型机动车的数量，包括”smlpas_sum”（小型及以下客车总量），”midpas_sum”（中型客车总量），”maxpas_sum”（大型客车总量）。具体数据结构如表 4-3 所示：

表 4-3 车辆因素相关特征数据结构

MidTZoneId	Smlpas_sum	Midpas_sum	Maxpas_sum
JTZQ101	0	26	29
JTZQ103	0	0	0
JTZQ108	766	5	5

然而，当特征被提取出来以后，我们发现车辆相关的特征缺失值达到了一半左右，根据 4.2.1 节中提到当缺失值达到近一半时，只能将这些特征从集合中删除。

（2）出行需求相关的特征主要以样本所处时间段（是否为高峰期/是否为工作日）来间接表示，因为在高峰期和工作日时，驾驶员的速度需求和时间需求会有明显增加，因此时间段的不同可能会显著影响交通违法行为的发生。本文定义高峰时间段为 07:00:00-09:00:00，17:00:00-19:00:00；周一至周五为工作日。时间特征是通过 MySQL 对每一条交通违法记录的时间做整体聚类而挖掘得到的。具体数据结构如表 4-4 所示。

表 4-4 出行需求因素相关特征数据结构

MidTZoneId	date	weekday	hourperiod	if_rush	其余特征
JTZQ101	2017/5/2	1	0-1	0	…
JTZQ103	2017/5/2	1	22-23	0	…
JTZQ108	2017/5/3	2	8-9	1	…

（3）道路交通环境相关的特征包括每个区域的停车设施情况（地上地下停车场数量）、各等级路段总长度、平均车道宽度、道路总通行能力，以及每个时间段内的交通流量和车辆平均速度。其中，每个区域某个时间段内的总交通流量在过往研究中没有统一的定义，经过试验，本文将其定义为平均每个路段的小时交通流*该区域路网总长度。具体数据结构如表 4-5 所示。

表 4-5 道路交通环境相关特征数据结构

MidTZoneId	date	hourperiod	tot_onpk	tot_unpk	tot_volume	avg_speed	…
JTZQ101	2017/5/2	0-1	67	0	486621.4912	34.56	…
JTZQ103	2017/5/2	22-23	234	86	467538.2955	27.60	…
JTZQ108	2017/5/2	8-9	566	887	508090.0864	33.54	…

（4）其余的环境和社会属性相关的特征则包括当天的气温、晴雨、该中区各种用地属性的面积划分、各种类型人口总数、是否有特殊建筑等。其中“该中区各种用地属性的面积划分”这一特征的划分原理是：中区由不同小区构成，每个小区有独立用地属性和用地面积，将中区内各小区的面积按照不同用地属性求和便可得到“该中区各种用地属性的面积划分”。具体数据结构如表 4-6 所示：

表 4-6 其余的环境和社会属性相关的特征数据结构

MidTZoneId	特殊建筑	tot_pop	…	A_acrg	U_acrg
JTZQ101	学校	15820	…	10554.13	4054.61
JTZQ103	商场	9347	…	156026.47	66463.28
JTZQ108	NULL	3372	…	69771.84	4074.68

综上，我们提取了共计 54 组特征，从四个方面入手较为全面的概括了预测违法行为所需要的重要特征。

4.2.3 数据集成

根据 4.1.2 中的描述，本实验的数据由昆山市 2017 年的微波数据、规划数据、非现场执法三个来源所集合而成。在进行数据集成时，要考虑到：该以什么维度和颗粒度来聚合出一条样本，如何定义输出值，又如何将多种数据库中不同存储结构的数据进行合并处理。因此，数据集成不仅仅是简单的数据相加，而是以某种规范化的流程解决原始数据的矛盾，并将其进行合并，形成可用的训练集和测试集。本实验融合三组数据来源和四种影响因素，给出了最终的训练和测试数据特征集。

训练集的每条样本输出代表昆山市某一交通中区在某天的某个小时段内的某类型违法行为发生总数，输入则包括该地区该时间段内的规划信息、路网信息、交通流和车辆速度信息、其余环境信息。数据集全部特征字段解释见附录 B。

4.2.4 特征选择

通过 4.2.3 的特征提取与构造，我们已经有了大量的备用训练特征，然而我们需要进一步的特征选择来确定哪些特征的解释性最强，预测结果的效果最好。因为特征选择可以减少模型特征数量，从而使模型泛化能力增加，减少过拟合的可能性；同时，能够增强我们对于特征和特征值之间的理解。

特征选择主要有两种方法：1. 过滤式（filter）特征选择策略。其主要思想是使用某种评价函数来检验各特征与预测值的相关性，并且降低相关特征之间的影响，主要方法有卡方检验、信息增益、Pearson 相关系数检验等。2. 集成式（Embedded）特征选择策略。其主要思想是在实际模型既定的情况下学习出对提高模型准确性最好的属性，也就是学习器自主选择特征，通常使用 Regularization 或决策树思想[33]。

特征选择过程一般包括产生过程、评价、停止准则、验证这四个部分，流程如图 4-1 所示：

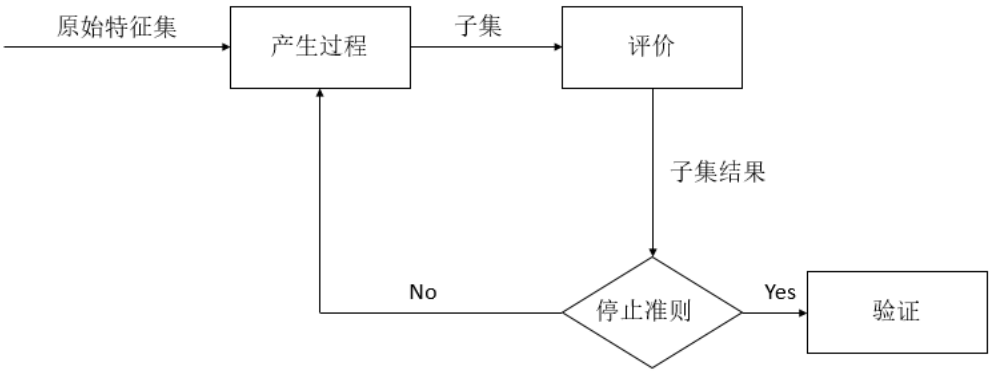


图 4-1 特征选择流程图

（1）产生过程：产生过程是搜索生成特征各子集的过程

（2）评价：评价某个特征子集的优劣准则数值

（3）停止准则：通常是一个阈值，当搜索到某一个特征子集的评价函数达到此阈值便可停止。

（4）验证：验证上述选出的特征子集的效果

由于本实验的特征和预测值之间不是简单的线性关系，因此无法使用简单且直观的过

滤式特征选择策略，只能使用集成式特征选择。为了节省计算资源，选用回归预测常用的的特征选择方法 XGBoost 模型训练，通过训练自动得到的 f-score 作为特征选择标准。每个特征的 f-score 代表该特征在 XGBoost 的树模型迭代构建中共被选中作为内部节点进行节点分裂多少次，可作为该特征在模型中的重要性指标[29]。最终得到的特征重要程度如图 4-2 所示。

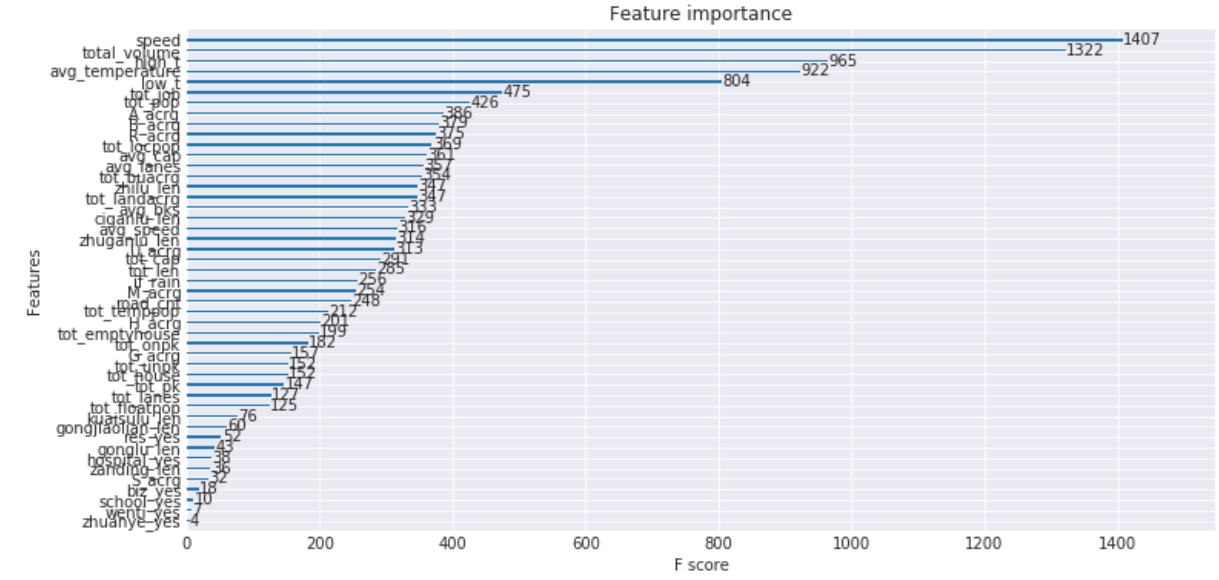


图 4-2 特征重要性排序

由此图可知，路网中的车辆速度、交通总量、天气情况等都是非常重要的特征，而浮动人口、路网中快速路的总长度等特征重要性相对较低，这与直觉相符。

由于没有相关研究给出使用 XGBoost 模型进行特征选择时所应选取重要特征的比例，而本实验的特征集并不算太大，因此我们选择只删除重要程度有明显下降的最后 11 个特征。

4.3 实验训练过程

本节将基于 2.3 节阐述的机器学习算法，介绍违法数量的训练过程。

4.3.1 准备数据

通过 5.2 节中的特征工程处理后，本文将 39 维来自昆山规划集、微波集的数据加上“小时时间段”和“周几”共 41 维数据作为每条样本的输入特征，将每个中区一周中每天每小时内的某违法类型发生次数作为预测值。需注意的是，六月第一周的样本由于被当作测试集，该时间段的“交通量”、“车辆速度”这两个特征值取的是 5.3 节中的时间序列预测值。

4.3.2 训练集与测试集划分

在建立机器学习模型时，一般需要将数据样本分成独立的三个部分：训练集（training set），验证集（validation set）以及测试集（test set）。其中训练集用来学习，估计模型参数；验证集用来对学习器的泛化误差进行评估并进而对模型的结构或超参数进行选择；而测试

集则检验最终训练得到的模型的性能如何。一个典型的划分是训练集占总样本的 60%，验证集和测试集各占 20%。

对于本实验来说，由于数据预处理的计算消耗很大所以只构造了五月份的训练样本，鉴于一个月的训练样本量不是非常大，本文将训练集和验证集合并后再采用交叉验证的方式来进行训练和调参。同时，鉴于违法数据有一定的时间规律，测试集不按比例选取，而是取六月第一周整周的数据。本文将特征工程处理过后的样本数据共 44,568 条划分为：

训练集+验证集 $Set1$：36,359 条样本，占 80%

测试集 *Set2*：8,209 条样本，占 20%

另外，在模型搭建过程中，训练集在学习和参数选择时采用 k 折交叉验证法来对其所包含的训练集和验证集作详细划分，具体方法为：

1. 将数据集$Set1$随机分为互斥的 k 个子集，一般取 k=10。
2. 将 k 个子集随机分为 k-1 个一组剩下一个为另一组，有 k 种分法。
3. 将每一种分组结果中，k-1 个子集的组当做训练集，另外一个当做测试集，这样就产生了 k 次预测，对其取平均。

4.3.3 基模型的训练

（1）岭回归、LASSO 回归、弹性网络回归预测训练

这三个回归模型的原理就是通过增加惩罚项来消除特征间的共线性问题，同时后两者还可以进行特征选择，从而提高模型的解释能力和预测精度。模型最关键的参数便是正则项系数，岭回归、LASSO 回归都是 alpha，弹性网络模型则为 alpha 和 l1_ratio。一般也只需调整这一/两个参数，可以运用交叉验证法得到测试分数 rmse 最低时的参数组合。训练结果如表 4-7、表 4-8、表 4-9 所示：

表 4-7 岭回归训练参数与结果

参数及范围	调参结果	模型 RMSE
alpha:{ 0.0005,0.001, 0.01,0.05, 0.1, 0.3, 1}	0.001	0.1604

表 4-8 LASSO 回归训练参数与结果

参数及范围	调参结果	模型 RMSE
alpha:{ 0.0005,0.001, 0.01,0.05, 0.1, 0.3, 1}	0.01	0.1613

表 4-9 弹性网络回归训练参数与结果

参数及范围	调参结果	模型 RMSE
alpha:{ 0.0005,0.001, 0.01,0.05, 0.1, 0.3, 1}	0.05	0.1613
L1_ratio:{ 0.0005,0.001, 0.01,0.05, 0.1, 0.3, 1}	0.5	

（2）梯度提升决策树预测训练及调参

在梯度提升决策树（GBDT）回归模型中,参数可以分为两类。一类是 Boosting 框架的超参数；一类是框架中单个弱学习器的参数，如 CART 决策树的超参数。

Boosting 框架相关的重要参数如下：

1）n_estimators：弱学习器的最大迭代次数。n_estimators 太小，容易欠拟合，n_estimators 太大，又容易过拟合，一般选择一个适中的数值，默认是 100。

2）learning_rate：每个弱学习器的权重缩减系数，也称作步长。一般可以从一个小一点的步长开始调参，默认是 1。对于同样的拟合效果，较小的步长意味着需要次弱学习器的迭代。因此，在实际调参的过程中，常常将 n_estimators 和 learning_rate 学习率一起调参来决定算法的拟合效果。

3）loss：损失函数类型。由于本模型存在部分噪音点，因此选择用抗噪音的损失函数"huber"。

弱学习器 CART 回归决策树的重要参数：

1）max_features：划分时考虑的最大特征数。由于本文的数据样本特征数不到 50 维，并且已经做过特征选择，用默认的"None"。

2）max_depth：决策树最大深度。由于本实验的特征数不多，默认可以不输入，决策树在建立子树的时候不会限制子树的深度。

3）min_samples_split、min_samples_leaf、max_leaf_nodes 等参数都是为了防止决策树过于复杂的，由于本文的特征数不太多，不会有太大的过拟合问题，因此可以不设置。

对于梯度提升决策树的调参过程，本文考虑到了 n_estimators、learning_rate 等参数的影响，利用交叉验证的流程和网格搜索的方法[34]，得到测试分数 rmse 最低时的最优参数组合。训练结果如表 4-10 所示。

表 4-10 GBDT 模型训练参数与结果

参数及范围	调参结果	模型 RMSE
n_estimators:{ 50,70,100,150,300}	70	
Learning_rate:{0.01,0.005,0.01,0.5,0.8,1.0}	0.01	
max_depth	8	0.1514
loss	Huber	
max_features	None	

（3）XGBoost 预测训练及调参

XGBoost 必须设置三种类型的参数：通用类型参数，用来对模型的总体结构作控制，包括使用的 booster 的类别以及在 boosting 时使用的最大共用线程数；booster 参数，具体用法与 GBDT 类似；学习任务参数，对于回归问题，一般采用均方根误差 eval_metric[29]来衡量学习目标的表现。

在 XGBoost 预测中，当数据和特征都不是很多的情况下，可采用网格搜索法来枚举所有参数进行调优。或者按照如下思路进行参数调节：

Step1：设置一个较高的学习步长，此处选定为 0.4，即 booster 参数中的 learning_rate。此时，保持其它参数不变，调节 n_estimators 的参数也就是最佳决策树的数量。

Step2：保持 n_estimators 和其他的 booster 参数不变，调节 learning_rate

Step3：保持 n_estimators 和 learning_rate 参数不变，调节 booster 的其他参数。从影响最大的树最大深度 max_depth 和最小叶子节点权重 min_child_weight 开始。

Step4：保持其余所有参数不变，缩小 learning_rate，得到最佳 learning_rate 值。

本文采用上述方式对本数据集进行调参，考虑到了 n_estimators、learning_rate、max_depth、min_child_weight 等参数的影响，利用交叉验证的流程和网格搜索的方法，得到测试分数 rmse 最低时的最优参数组合。训练结果如表 4-11 所示。

表 4-11 XGBoost 模型训练参数与结果

参数及范围	调参结果	模型 RMSE
n_estimators:{ 50,70,100,150,300}	100	
Learning_rate:{0.01,0.005,0.01,0.5,0.8,1.0}	0.01	
max_depth	8	0.1451
loss	Huber	
max_features	None	

4.3.4 改进的模型融合预测

构建并结合多个学习器来完成学习任务，我们把它称为模型融合或者集成学习。由于上述模型算法原理不同，在对本文的事故-违法数据集进行训练预测时结果具有一定差异性，而模型融合可以使所有模型发挥出各自的优势，可以通过某种策略将几个弱学习器结合，从何形成一个强学习器。

首先，模型融合有两个前提：1. 各个基学习器的性能不能太差。2. 各个学习器之间的准确度和算法原理都需要一定的区分度。在本文第五章的实例分析中，可以发现本文采用的上述几个模型可以满足模型融合的前提。

本文采用两种融合方式 Blending 和 Stacking，并对结果作比较，选取结果较优的一个。

（1）线性 Blending

这是一种减小估计方差的融合方式，操作较为简单，只需用之前已训练好的各基模型 $g_t(x)$（第四章中提到的所有模型）对训练集预测结果按照各自准确度分配大致权重，并对权重进行进一步调参：

$$G(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t \cdot g_t(X) \qquad (4\text{-}2)$$

其中，$\alpha_t \geq 0$。最后在预测测试集时，结果也按照相同比重来输出最终的预测值。本实验基模型的具体参数详见 4.3.3 节，各基模型的权重设置见表 4-12。模型最后的评价指

标 rmse 值为 0.1401。

表 4-12 各基模型 rmse 值以及权重值

模型$g_t(X)$	rmse	α_t值
岭回归模型	0.1604	0.15
LASSO 回归模型	0.1613	0.15
弹性网络回归模型	0.1613	0.15
梯度提升决策树（GBDT）	0.1504	0.25
XGBoost	0.1501	0.30

（2）Stacking

Stacking 实际上就是把 Blending 组合起来，Blending 只有一层，而 Stacking 有多层，它把各个基学习器的预测结果作为下一层新的训练集，来学习一个新的学习器[35]。图 4-3 为原理示意图，具体算法见表 4-13。

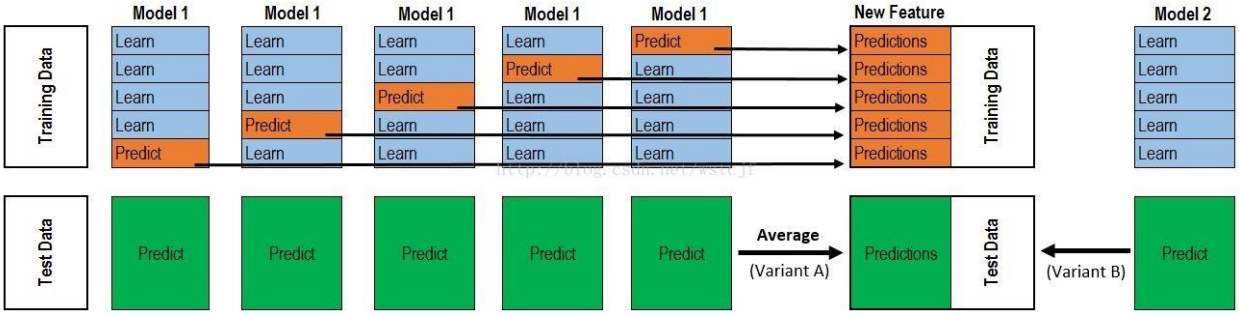


图 4-3 Stacking 算法流程图

表 4-13 Stacking 算法

Stacking 算法

1：输入：违法数据训练集D $= \{x_i, y_i\}_{i=1}^{m}$

2：输出：融合回归模型H

3：步骤一：训练基模型

4：for t=1 to T

5：　使用 D 训练h_t

6：end

7：步骤二：利用预测值构造新的训练集

8：for i=1 to m

9：　$D_h = \{x_i', y_i\},\ where\ x_i' = \{h_1(x_i), \dots, h_T(x_i)\}$

10：end

11：步骤三：学习新的融合模型

12：使用D_h训练H

13：返回H

本实验基模型的具体参数详见 4.3.3 节。Stacking 融合模型最后的 rmse 值为 0.1399。由于本数据集的样本量和特征数都不是非常大，使用 stacking 有过拟合的风险。而且根据

rmse 值来看，最终实验结果证明 stacking 方法并没有显著提升模型准确度，而消耗的计算资源却大得多。因此本实验最终采用加权 blending 的方法对模型进行融合。

4.3.5 基于聚类的分层融合模型

在前几节的实验基础上，本节提出使用基于地理信息聚类的分层融合模型以实现更为精确的违法预测。根据 4.2.4 节得到的特征重要性，可以看出，拥有不同规划数据和地理信息的地区发生交通违法数量有明显不同，并且具有相似地理属性的地区所发生的违法数量可能对交通流、天气等特征的变化有着更相似的变化规律。因此，本文提出了基于 K-means 聚类的分层融合模型，通过更深层次地挖掘时间与空间之间的联系，得到更好的预测结果[36]。

本节首先通过轮廓系数确定了最优聚类数，然后选取了根据 4.2.4 节得到的重要性在前八位的地理特征，如表 4-14 所示，对所有交通中区进行聚类，以对不同聚类中的交通中区分别重新训练模型，进行预测。

表 4-14 区域聚类所用特征

字段名	字段含义
tot_job	该中区总工作岗位数
avg_cap	该中区路段平均通行能力
U_acrg	公用设施用地
tot_pop	该中区总人口数
tot_buacrg	总建筑面积
B_acrg	商业服务业设施用地面积
A_acrg	公共管理与公共服务设施用地面积
tot_landacrg	总面积

（1）确定最优聚类数

K-means 具体算法已在 3.3.2 节中作说明，见表 3-6。但是由于昆山市交通分区众多，难以单靠目测确认最优聚类数，因此此处详细说明确定聚类算法最优聚类数的原理[37]。

本文采用平均轮廓系数法来确认最优聚类数。轮廓系数是类密集程度的评价指标。彼此相距很远，本身很密集的类，其轮廓系数较大；彼此集中，本身很大的类，其轮廓系数较小。轮廓系数是通过所有样本计算出来的，计算每个样本分数的均值，计算公式如下：

$$s = \frac{b-a}{\max(a,b)} \qquad (4\text{-}3)$$

其中，a 是每一个类中样本彼此距离的均值，b 是一个类中样本与其最近的那个类的所有样本的距离的均值。平均轮廓系数法通过计算 k 取不同值时的样本总平均轮廓系数，得到当轮廓系数最大时的 k 便为最优分类数。具体步骤如下：

Step1：对各区域使用不同 k 值进行 K-means 聚类操作，此处 k 取[2,3,4,5,6,7,8,9].

Step2：对每次 k 值聚类，计算样本总平均轮廓系数。

Step3：以 k 值为 x 轴，画出轮廓系数变化曲线图，如图 4-4。

Step4：曲线最大值处对应的 k 值则为最优聚类数。

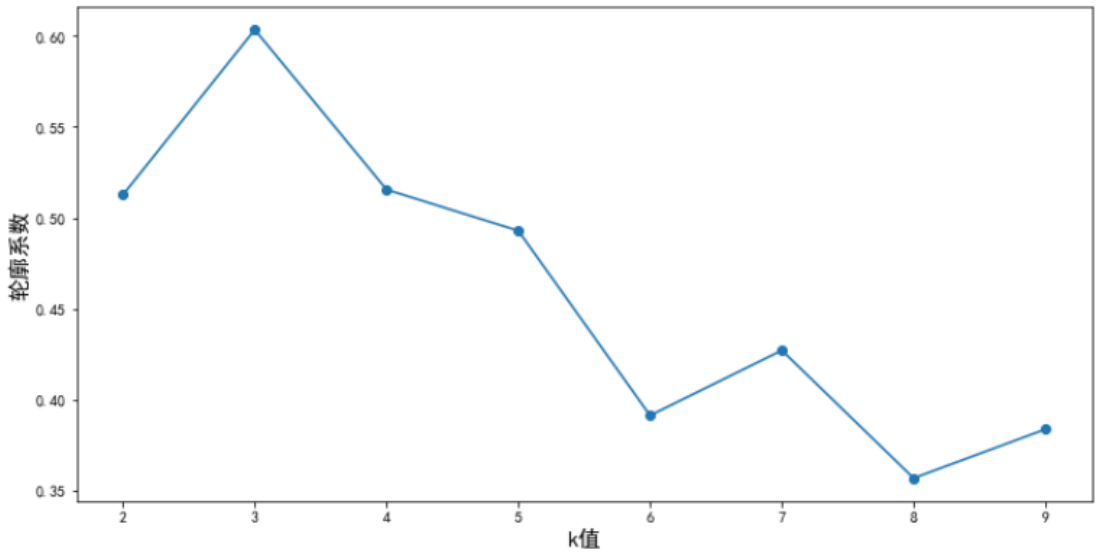


图 4-4 不同聚类数对应的平均轮廓系数

由此可知，k 值为 3 时，整体聚类效果最好，因此在后续的基于聚类的预测模型中，聚类数也取 3。

（2）基于聚类的分层融合模型建立

本文提出的该模型算法如表 4-15 所示，算法的输入是原始特征集和 4.3.4 节得到的融合模型，输出为各地区违法数据。通过交叉验证得到的 rmse 分数为 0.1387，优于融合模型的 rmse0.1399。因此，通过聚类的手段探索地理信息和交通流等特征的时序变化之间的关系，本文进一步提升了模型的预测准确度。

表 4-15 基于聚类的分层融合模型算法

基于聚类的分层融合模型算法
1：输入：违法数据训练集D = $\{x_i, y_i\}_{i=1}^{m}$
2：输出：基于聚类的分层融合模型 M1
3：步骤一：得到 D 的聚类结果 clusters
4：clusters=Kmeans(D)
5：步骤二：对于每一个 cluster 调用 4.3.4 中的融合模型 M2
6：for cluster in clusters:
7：　a）$M1_{(cluster)} = M2(cluster)$
8：　b）M2.add($M1_{(cluster)}$)
9：得到最终基于聚类的分层融合模型 M2

4.4 基于随机森林模型预测交通流和平均速度构造测试集特征

在机器学习中，随机森林是由 Leo Breiman（2001）提出的一种包含多棵分类回归树（Classification And Regression Tree, CART）的集成算法，它通过 bootstrap 重采样技术和

bagging 的思想[38]，从原始训练样本集 N 中有放回地重复随机抽取 n 个样本生成新的训练样本集合训练决策树，然后按以上步骤生成 m 棵决策树组成随机森林，新数据的结果按决策树预测结果的平均值而定。其实质是对决策树算法的一种改进，通过集合多个弱决策树来得到强学习器。随机森林包含多个决策树可以降低过拟合的风险，原理为：RF 分别训练一系列的决策树，所以训练过程是并行的，因算法中加入随机过程，所以每个决策树又有少量区别；通过合并每个树的预测结果可以减少预测的方差，从而显著提高模型在交通流预测的测试集上的性能表现。

4.4.1 RF 算法步骤

由于本文目的是预测交通流和区域平均速度，并且预测模型相似，因此以下只举例讨论随机森林回归模型预测交通流问题，模型最终的预测结果为各个树预测结果的平均值，具体算法如下：

(1) 给定一个训练集 S，测试集 T 和特征数量 K。

(2) 从训练集 S 中随机有放回的抽取特征数和 S 相同的训练子集 S(i)，作为最先开始训练的根节点。

(3) 若当前节点已经达到训练结束条件（该节点上的样本数≤ s或该节点上的信息增益 ≤ m），则设置当前节点为叶子节点，预测结果为当前节点中所有训练样本的平均值；然后以同样的方式训练其他节点。若当前节点还未满足终止条件，则从所有特征中随机选择 k 个特征，对选出的样本利用 k 个特征建立决策树（本实验采用 CART）。

(4) 重复(2) (3)步骤 m 次，即生成 m 棵决策树，形成最终随机森林模型。

4.4.2 特征构造

本文采用经预处理的江苏省昆山市各交通中区的五月份交通总量(每条路段的平均小时交通量*该地区路网总长度)和地区车辆平均行驶速度进行研究，数据来源于各路段的微波检测器，检测器全天候工作，每隔 30s 记录一次数据，通过处理得到各时段的交通流数据。

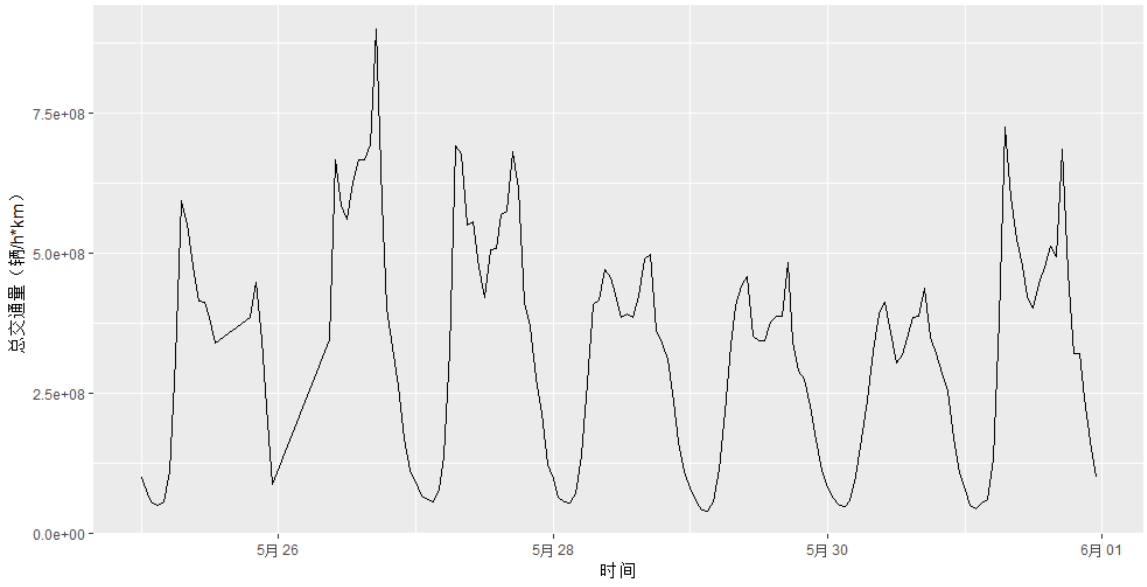


图 4-5 2017/05/25 至 2017/05/31 交通总量数据

如图 4-5 所示是昆山市所有微波设备检测地区的交通总量之和在五月最后一周的波动曲线。通过对数据集的观测可以得知，交通流量在一天之内和一周之内都有相似性波动。由此得到启发，如果要对路段未来的交通量进行预测，一天中的时段和一周第几天将会成为重要特征；同时，考虑到交通流有很强的时间序列特性，前一周以及之前紧邻时间段的交通量也应作为重要特征。通过随机森林模型的重要特征选择作用，挑选出以下所构造的有效特征：

表 4-16 交通流预测模型特征解释

变量	变量解释
$y = flow(t)$	预测值:某小时段t内的车流量
$x_0 =$ weekday	当前时刻t所在一周中的第几天
$x_1 = hourperiod$	当前时刻t所在一天中的时间段(hour)
$x_2 = flow(t-1)$	当前时刻 t 前 1 小时的车流量
$x_3 = flow(t-2)$	当前时刻 t 前 2 小时的车流量
$x_4 = flow(t-3)$	当前时刻 t 前 3 小时的车流量
$x_5 = flow(t-4)$	当前时刻 t 前 4 小时的车流量
$x_6 = flow(t-7\_0)$	当前时刻 t 前一周相同时刻 t 的车流量
$x_7 = flow(t-14\_0)$	当前时刻 t 前两周相同时刻 t 的车流量
$x_8 = flow(t-21\_0)$	当前时刻 t 前三周相同时刻 t 的车流量
$x_9 = weather$	若当前时刻 t 下雨则为 1，不下雨为 0

本文取 2017-04-06 至 2017-05-31 共 8 周的数据作为训练集，2017-06-01 至 2017-06-07 共一周的数据作为测试集。

4.4.3 预测结果

通过 Grid Search 调参方法，得到本交通流预测模型的最优超参数设置如表 4-17：

表 4-17 模型超参数设置

超参数取值	超参数含义
n_estimators=125	决策树迭代次数
criterion=mse	CART 回归树分裂的衡量标准，默认为平均方差
max_features=None	RF 划分时考虑的最大特征数，此处训练集特征较少不做限制
max_depth=50	决策树划分的最大深度
min_samples_split=2	内部节点再划分所需最小样本数

需要注意的是，通过模型在测试集上的误差变化来看，事实上随机森林模型对于参数调节的要求不高。

最终训练得到的 RF 模型对测试集的预测结果与真实值的比较分布如图 4-6 所示。其中折线图代表实际观测交通流，散点代表模型输出值。可看出输出值与真实值的波动基本一致，这反应了随机森林模型在拟合时间序列交通流方面表现良好。并且本预测模型的平均百分比误差(Mean Absolute Percent Error)只有 18%，这意味了模型预测准确率从某种方

面已经达到了 82%，已经可以满足后续预测需求。

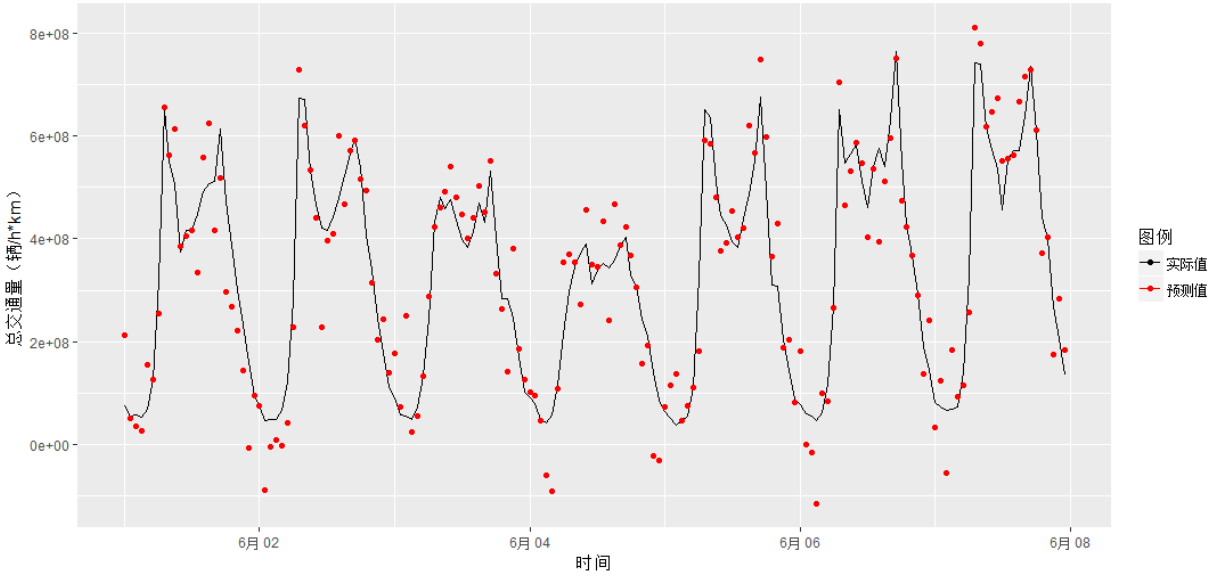


图 4-6 交通流预测结果

通过这个模型框架，可以得到昆山市各交通中区在 6 月第一周每天每小时的交通流和车辆平均速度的数据，从而有效为最终预测违法数量的测试集构建出交通流和车辆速度两个特征，这使得仅仅利用已有的数据来预测未来的违法行为得以实现。

4.5 测试结果与分析

4.5.1 评价指标

我们选取 RMSE，MAE，MAPE 三个指标对预测结果做测试评价。其中，均方根误差 RMSE 一直是机器学习回归预测问题中最常用的指标[39]，和平均绝对误差 MAE 相似，都表示预测值与真实值之间的差异。而现在研究者逐渐开始使用无量纲的评价指标，其中最常用的便是平均绝对百分误差 MAPE，用来衡量预测准确度，可以说明该预测结果的精确值为多少[40]。三个指标定义如下：

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{t=1}^{T}(\widehat{y_t}-y_t)^2}{T}} \tag{4-4}$$

$$\mathrm{MAE}=\frac{\sum_{t=1}^{T}|\widehat{y_t}-y_t|}{T} \tag{4-5}$$

$$\mathrm{MAPE}=\frac{100\%}{n}\sum_{t=1}^{T}\left|\frac{\widehat{y_t}-y_t}{y_t}\right| \tag{4-6}$$

其中，y_t是观测值，$\widehat{y_t}$是预测值。

4.5.2 模型性能

各基模型与融合模型在测试集上的性能评价如图 4-7 所示。通过该图，我们可以发现基于聚类的融合模型的性能较原本的基模型有了一定的提升，最终预测准确度达到了 85.95%，效果较好。

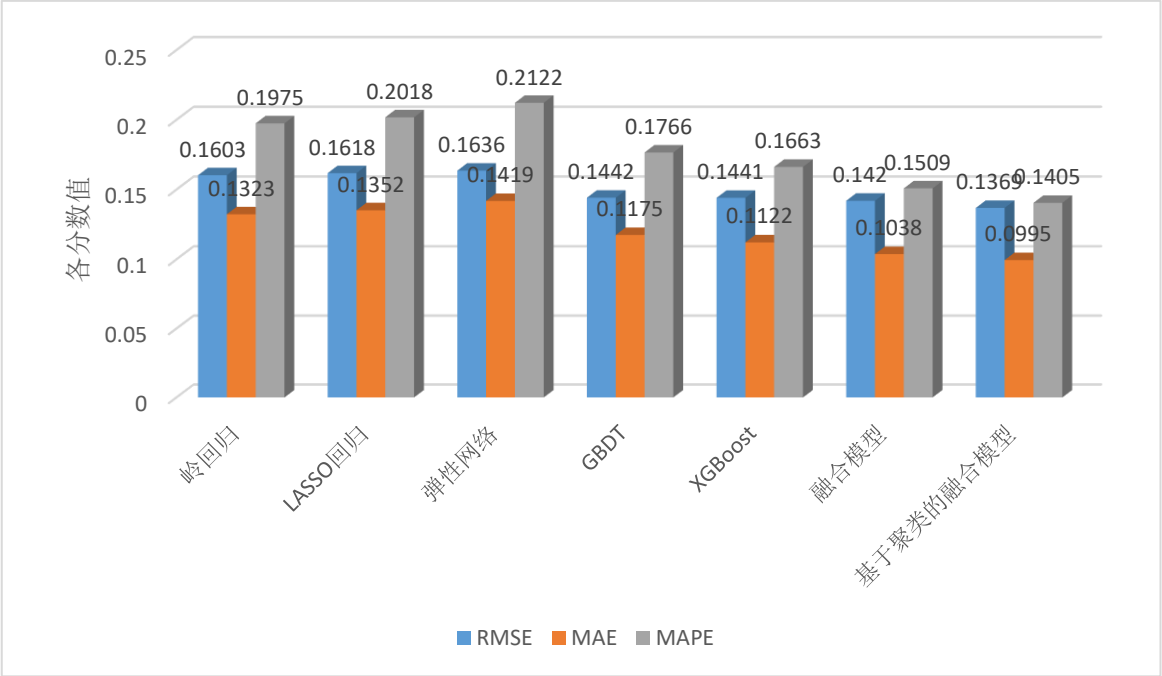


图 4-7 各模型在构造测试集上的性能评价

若测试集中的交通量/速度特征使用实际值，而非 4.4 节得到的预测值，各模型表现如图 所示，可以发现，模型结果指标普遍显著更优。因此，若能使用更好的方法提高交通流的预测精度，违法预测的模型结果还能进一步提升。

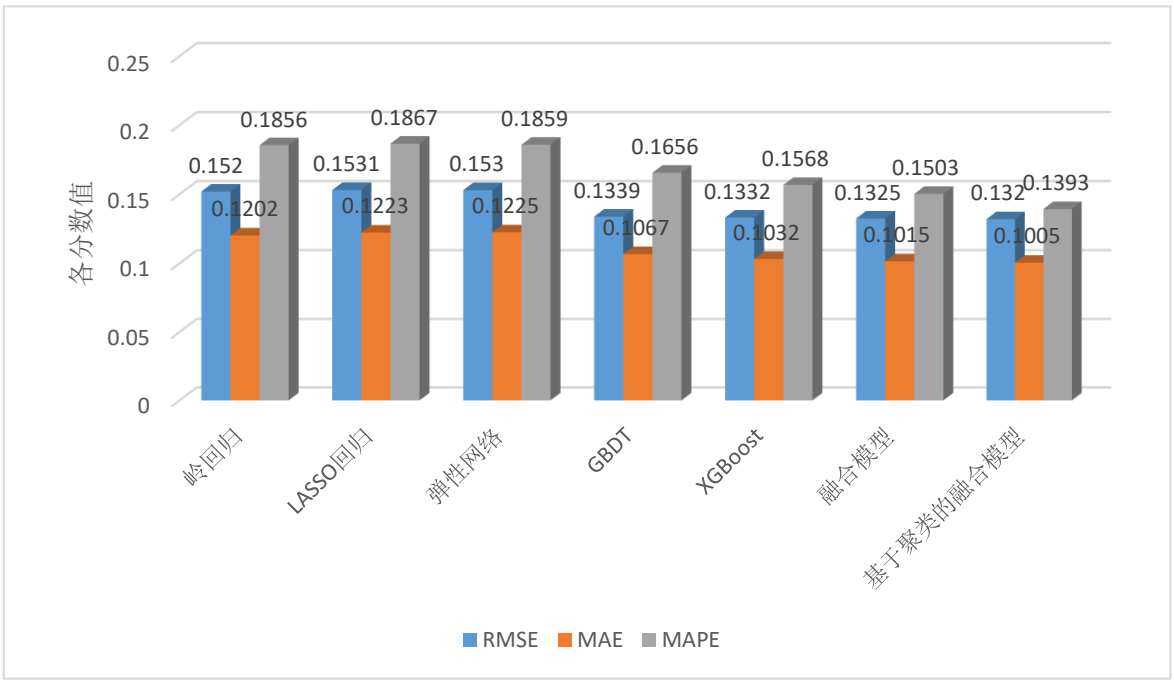


图 4-8 各模型在原始测试集上的性能评价

综上，本文所实现的基于机器学习的违法数量预测方法可以很好的考虑到时间特征以及空间特征，并且通过一系列的特征工程手段和随机森林构建测试集特征，最终训练给出较为准确的预测结果。并且根据模型训练过程中得到的特征重要度以及预测结果的可视化可以进一步对昆山市的交通违法行为做影响因素分析和时空演化分析，从而得到预防对策，最终降低某类交通事故如“伤人事故”的发生概率。

第五章　交通违法行为预防对策研究

5.1　违法行为主要影响因素

基于 2.1 节的影响因素分类和 4.2.4 节得到的特征重要性来看，“道路交通环境”和“其余环境因素及社会变量”（空间与环境因素）对于违法行为的影响比出行需求（时间因素）更大一些，如图 所示。

对于空间与环境因素来说，“道路交通环境”类中的“平均车速”、“总交通量”、“区域路段平均车道数”等特征以及“其余环境因素及社会变量”类中的“最高气温”、“平均气温”、“该地区工作岗位数”、“公共管理与公共服务设施用地面积”等特征对于交通违法行为的发生数量影响最大。与直觉不符的是，“是否下雨”以及“平均路段限速”等特征对于违法数量的多少并没有显著影响。

对于时间因素来说，“是否为高峰期”对违法数量的多少有较为明显的影响，而工作日与双休日之间并没有太大的区别。

然而，遗憾的是，由于本文基于机器学习模型来预测交通违法行为，无法得到特征与违法数量之间的线性关系。因此，若想得到它们之间的详细影响关系并提出具体的改善建议，需要在未来的研究中对每一个特征进行详细展开。

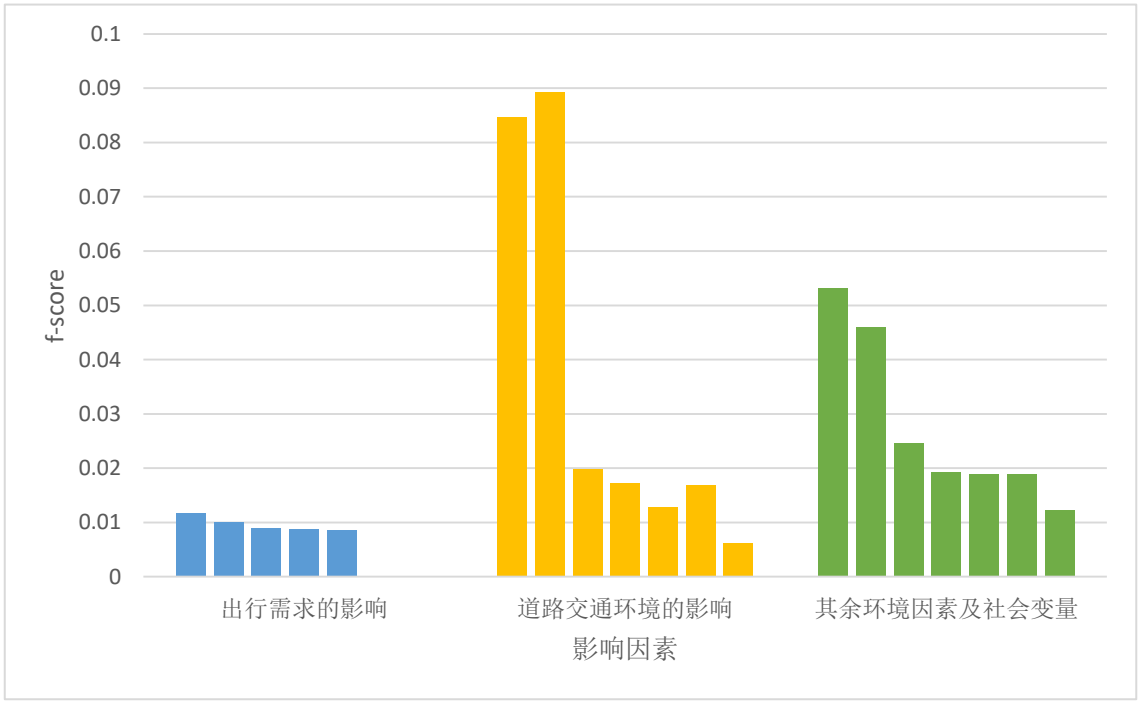


图 5-1 各类因素对违法数量的影响程度

所有特征对于违法数量的影响关系详见附录 C。

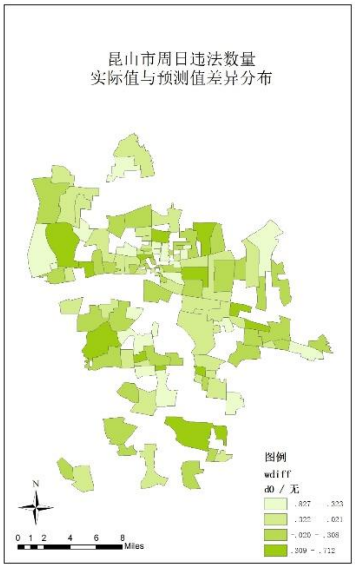
~~5.1~~ 5.2　违法行为时空演化分析

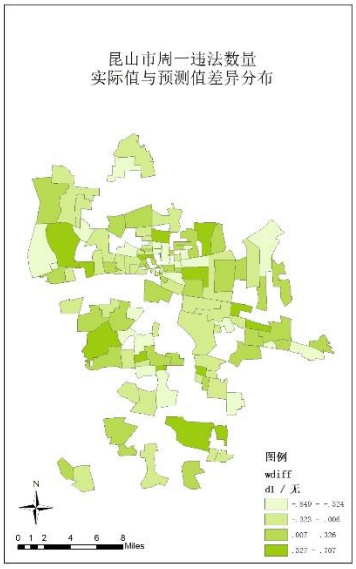
~~5.1.1~~ 5.2.1　违法实际数量与预测值的趋势一致性

为了证明违法数量的实际值与预测值在时空分布的趋势一致，更好的对违法时空演化进行分析，同时证明本文模型的有效性，本文从“周”和“小时”维度给出预测与实际差

值的分布图。

选取周日、周一、周四三天的预测与实际的差值作出差异图，如图，可以看到，差异分布图颜色较为一致，证明违法分布的预测值与实际观测值趋势在周的维度上一致。





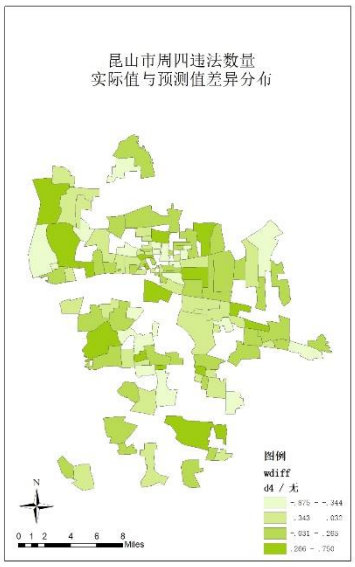
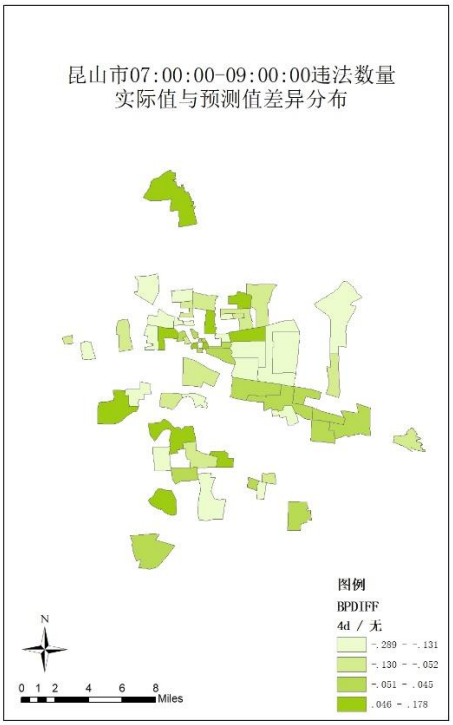


图 5-1 至 5-3 周日、周一、周四三天的预测值与实际值差异图

同理，选取 07:00:00-09:00:00、11:00:00-13:00:00、17:00:00-19:00:00、21:00:00-23:00:00 四个时间段的预测与实际差值作出差异图，如图，差异分布图颜色也较为一致，证明违法分布的预测值与实际观测值在小时的维度上趋势一致。



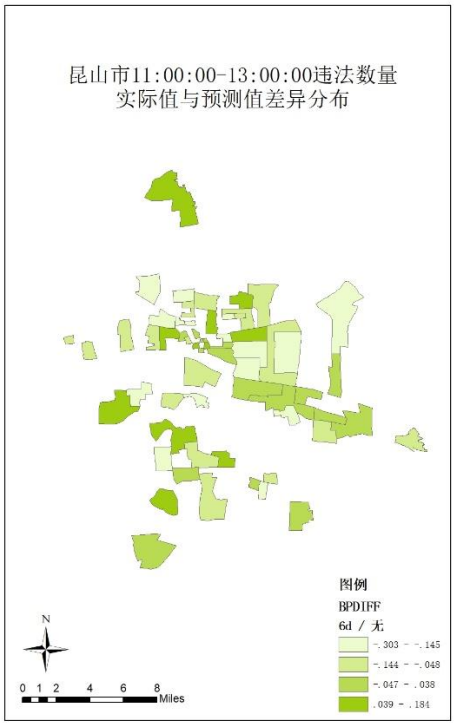


图 5-4 至 5-7 四个时间段的预测值与实际值差异图

综上，可以证明违法实际值与实验预测值的趋势一致，下用实验预测值进行可视化和进一步的分析。

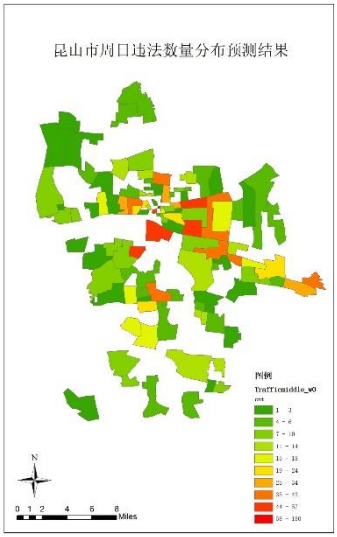
~~5.1.2~~ 5.2.2 违法预测值在各维度上的趋势

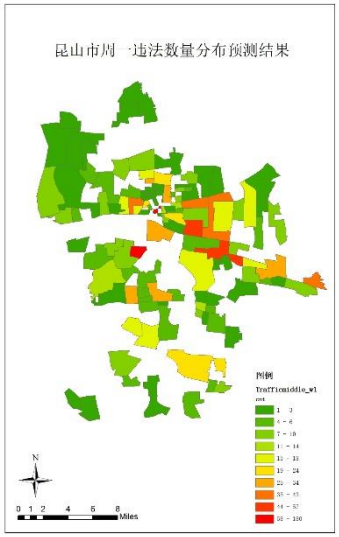
本节将从 “周”和“小时”两个维度来分别展示与分析昆山市各区域的违法数量时空演化。

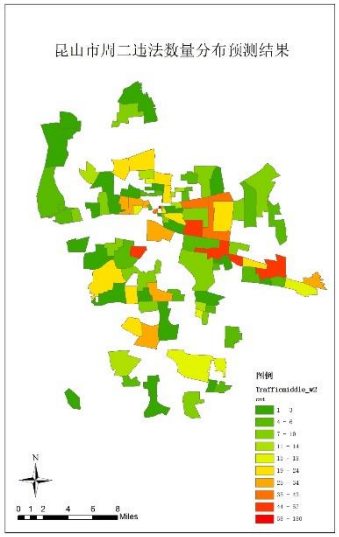
（1）“周”维度

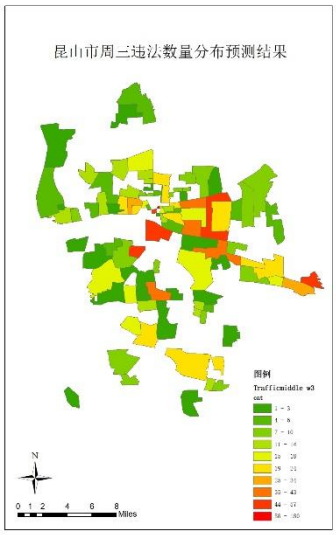
如图为昆山市六月第一周每天的预测结果。需要注意的是，由于部分区域没有违法监测设备，因此无法进行预测。可以得到以下结论：

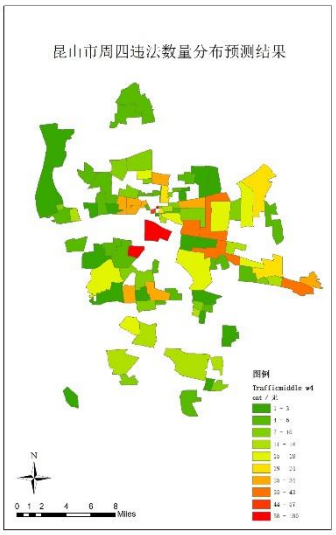
整体来看，周日和周一昆山市中心发生的违法行为较多，而周围地区的违法行为很少。从周二开始，市中心的违法行为逐渐减少，而周围地区的违法行为却逐渐增多，违法行为呈“由市中心向四周逐渐扩散”状。直到周六，违法行为又逐渐“向市中心聚集”。

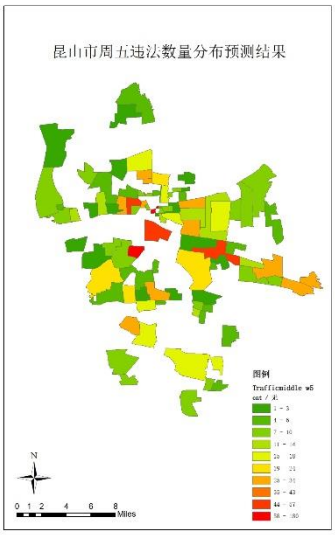












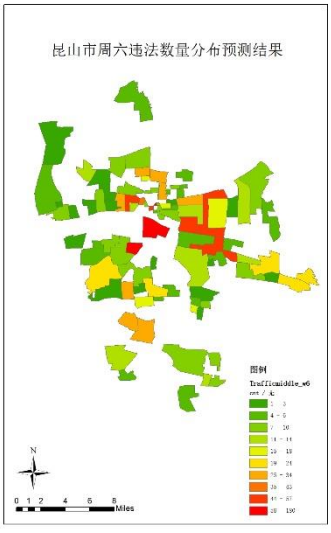


图 5-8 至 5-14 昆山市一周内违法预测结果

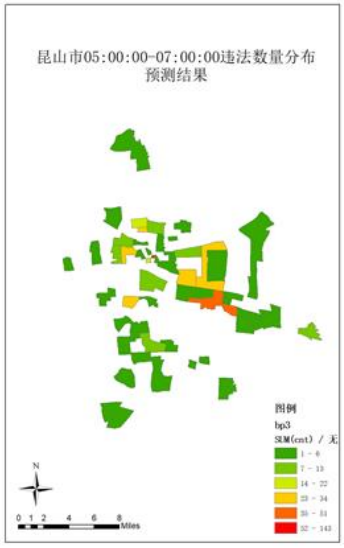
（2）“小时”维度

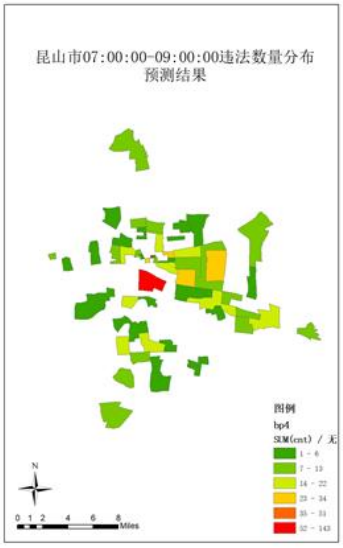
如图为昆山市六月第一周平均每两小时段的违法预测结果。需要注意的是，由于部分区域没有违法监测设备，部分区域违法行为记录时间的小时字段有缺失，因此无法进行预测。可以得到以下结论：

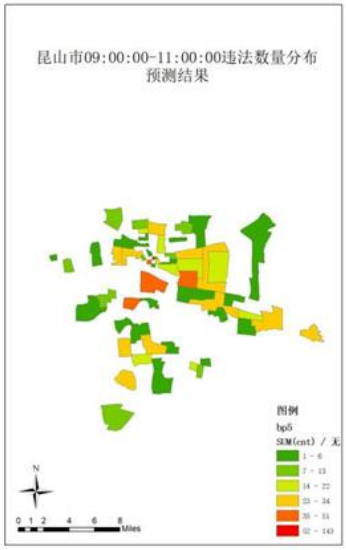
1. 每天 19:00:00 至次日 05:00:00 的违法行为都很少。

2. 早高峰时期违法行为较少，并且集中在市中心区域；晚高峰时期违法行为也较少，但分散在市区各地。

3. 与直觉相反，白天的“违反交通标志标识”行为比夜晚发生频率要高很多，可能是由于夜晚视线较差，司机反而会提高注意力。







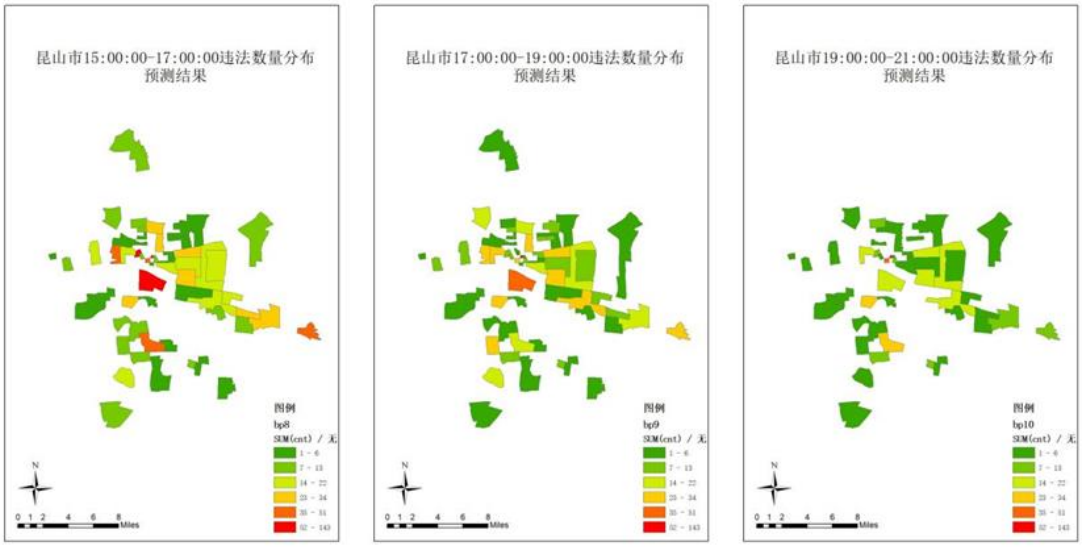


图 5-15 至图 5-23 昆山市一天内违法预测结果

~~5.2~~ 5.3 违法行为预防对策

根据 5.1 节和 5.2 节的分析，拟从以下几个方面提出对策以控制“违反交通标志标识”类违法行为从而进一步减少交通“伤人事故”。

（1）检查改进部分区域的交通标志设置

部分区域“违反交通标志标识”类违法数量在任何时间段都一直很高，如中山社区居委会、新南社区居委会、枫景苑社区居委会等区域，如图 5-24 蓝圈内的区域。

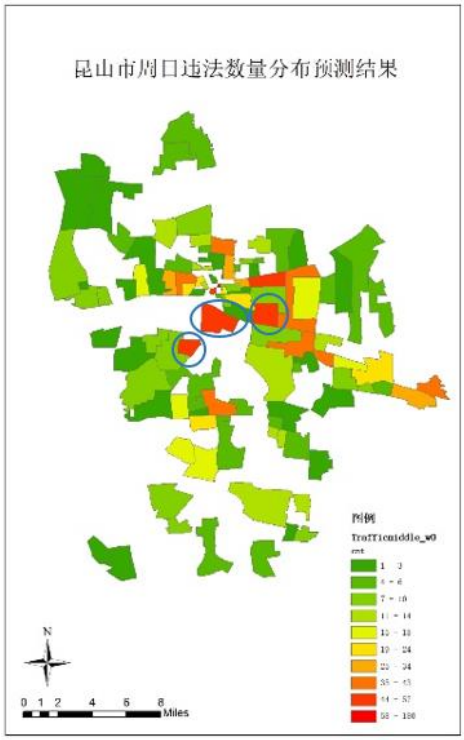


图 5-24 “违反交通标志标识”行为发生率很高的地区

通过和此类事故在所有地区发生次数占所有类型违法行为发生次数的比例进行比较，如图 5-25 所示，可以发现，这三个地区该类型违法发生频率高于平均值。执法部门需要对这三个地区的交通标志标识设置进行检查和改进。

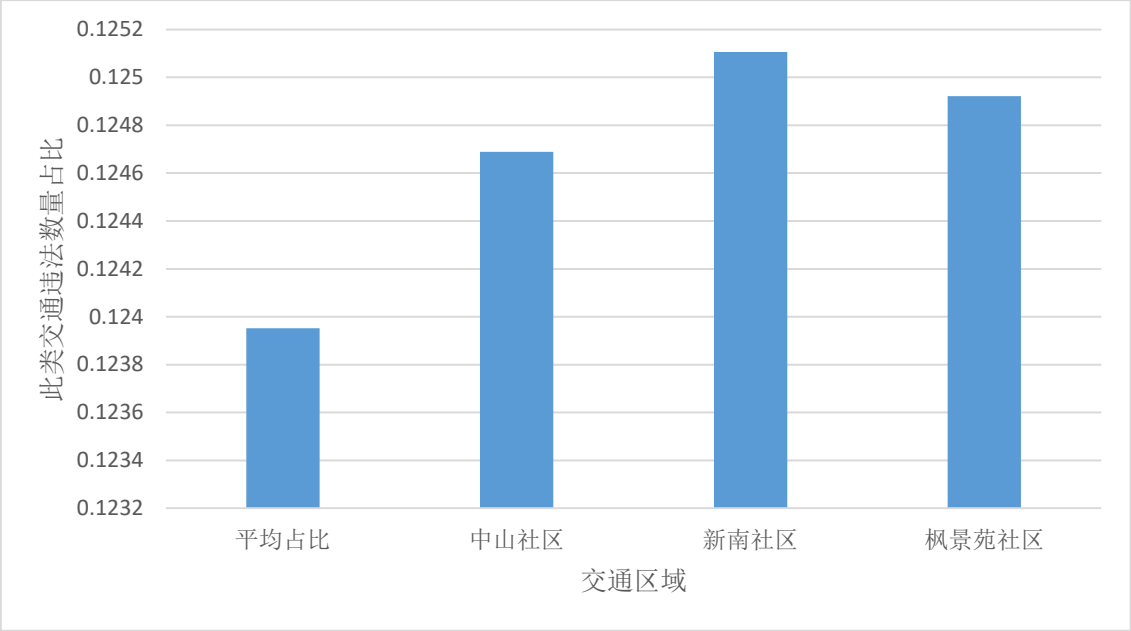


图 5-25 “违反交通标志”类数据占比

（2）应对天气变化的措施

根据 5.1 节的结论，“是否下雨”对该类型违法影响并不大，因此不必在雨天刻意增加对此类违法的监管资源。但是当气温变化尤其是气温骤升时，要注意该交通违法行为的监管查处。本文对所有区域的每天最高气温和该类型违法发生次数这两个变量作了相关性检验，得到显著相关（P-value<0.05）[41]并且相关系数较高的区域，如表 5-1 所示。这些区域的违法数量明显被高温影响，可能是因为高温会影响驾驶员的心理状态，而这些区域的标志标识设置的不够明显或不够人性化，所以驾驶员很容易违反交通标志。

表 5-1 违法数量与“最高气温”呈明显正相关关系的区域

区域	Pearson 系数	P-value
姜杭村	0.768	0.043
共青社区	0.401	0.038
蓬曦社区	0.393	0.013
…	…	…

（3）路网设置

区域的路网平均车道数和平均限速是路网的各个特征中对违法预测贡献最大的三个特征。当进行城区道路改造时，应注意控制各路段车道数，以减少该违法行为的发生。

由图 5-26 可知，当区域的平均车道数为 2 时，该区域发生“违反交通标志”类型的违法行为是最多的。同时对数据进行独立样本 t 检验，得到“2 组”与“1 组”、“3 组”的 t 检验 sig 值分别为 0.032 和 0.029（<0.05），这说明平均值在小于 5%的几率上是相等的，而在

大于 95%的几率上不相等[42]，从而得到“2 组”的平均值显著高于“1 组”、“3 组”的结论。因此，我们在双向车道数为 2 的路段要格外注意各种标志的设置规范，从而尽量减少该路段违法事故的发生。

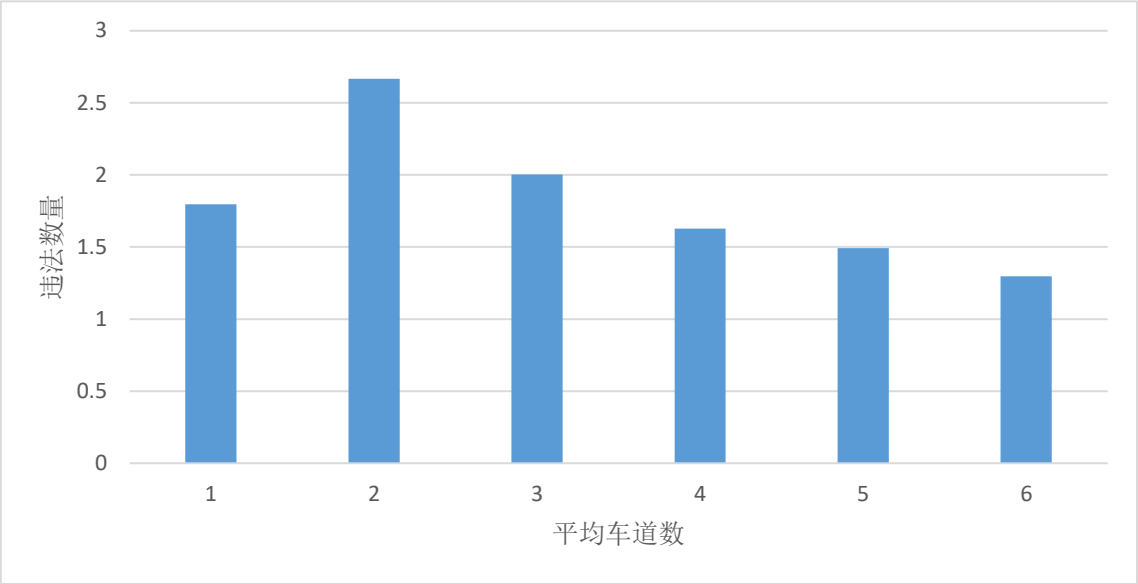


图 5-26 不同平均车道数的区域平均每小时违法数量

由图 5-27 可知，平均限速越低的区域发生“违反交通标志”类型的违法行为越多。同时对数据进行独立样本 t 检验，得到“30-35 组”与“35-40 组”的 t 检验 sig 值为 0.042，这说明平均值在小于 5%的几率上是相等的，而在大于 95%的几率上不相等[35]，从而得到“30-35 组”的平均值显著高于其他组的结论。因此，我们在限速较低的路段要格外注意各种标志的设置规范，从而尽量减少该路段违法事故的发生。

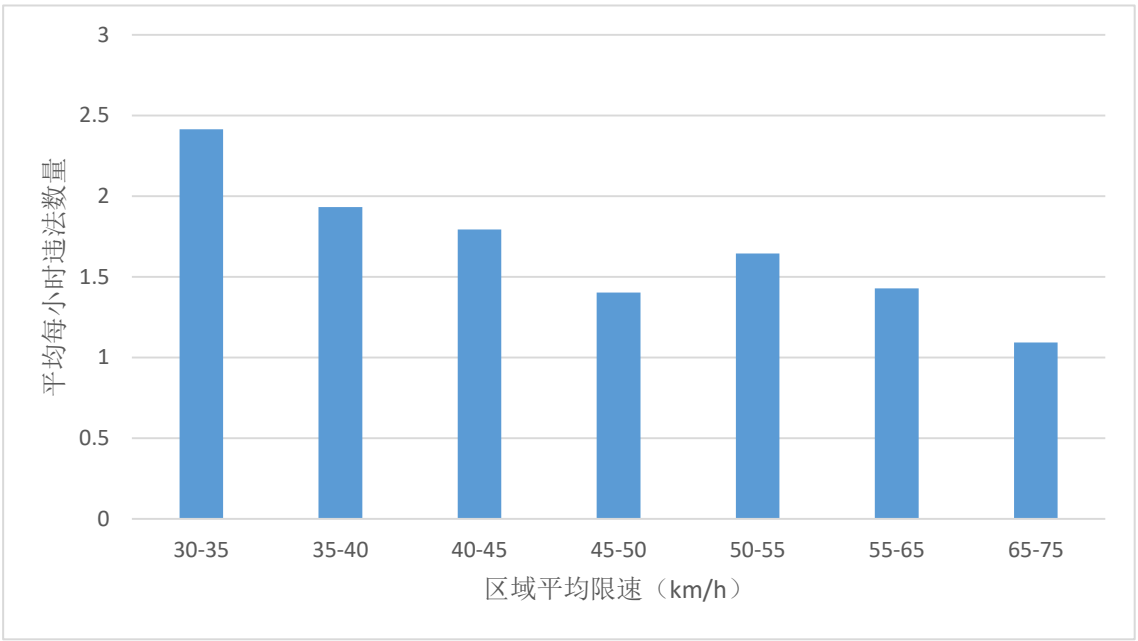


图 5-27 不同平均限速的区域平均每小时违法数量

（4）时间维度的执法管理资源调配

由 5.2 节的违法行为时空演化分析可得，整体来说，在周六、周日、周一这三天，交通

监管部门应尤其将重心放在对市中心区域的执法上。同时，整个城市白天的违法数量明显比夜晚多，因此应将夜晚期间的执法资源调配一部分到白天。

另外由 5.2.2 所有的时空演化图可知“JTZQ290 花安社区居委会”、“JTZQ289 新安社区”、“JTZQ245 虹桥社区”三个区域对于时间的敏感度很高，违法数据经常随时间的变化而变化。为了探究时间与空间的联系，本文将这三个区域的空间属性取平均值，与全昆山市的个属性平均值取相对误差（绝对误差/总体平均值），得到与整体差异最明显的特征如表 5-3 所示。

表 5-2 与全市平均值差异明显的特征

特征名	相对误差（%）
正在修建的道路长度	712.34
商业区数量	468.25
公用设施用地	390.79
支路长度	327.29

因此，这几个特征是影响地区违法数量时间敏感度的重要原因，后续研究和执法管理工作应单独针对“正在修建道路较长、商业区较多、公用设施用地较多、支路较多”的地区做进一步时间维度上的分析。

（5）监测交通流的变化

根据 5.1 节介绍，交通量为预测违法数量的重要特征，因此交通量的变化对违法行为的影响很大，根据图 可知，总交通量与违法数量大致呈现负相关关系。当总交通量大于 3,000,0000 辆/h*km 时，地区违法数量一直维持在较低水平。然而当总交通量降低时，违法数量开始上升，并且波动明显。因此，当交通量较少时路面较畅通时应更加注意“违反交通标志标识”类型违法数量的增多。相反的，高峰时期该类违法行为发生较少，可将交通管理资源调配到其他监管方向。

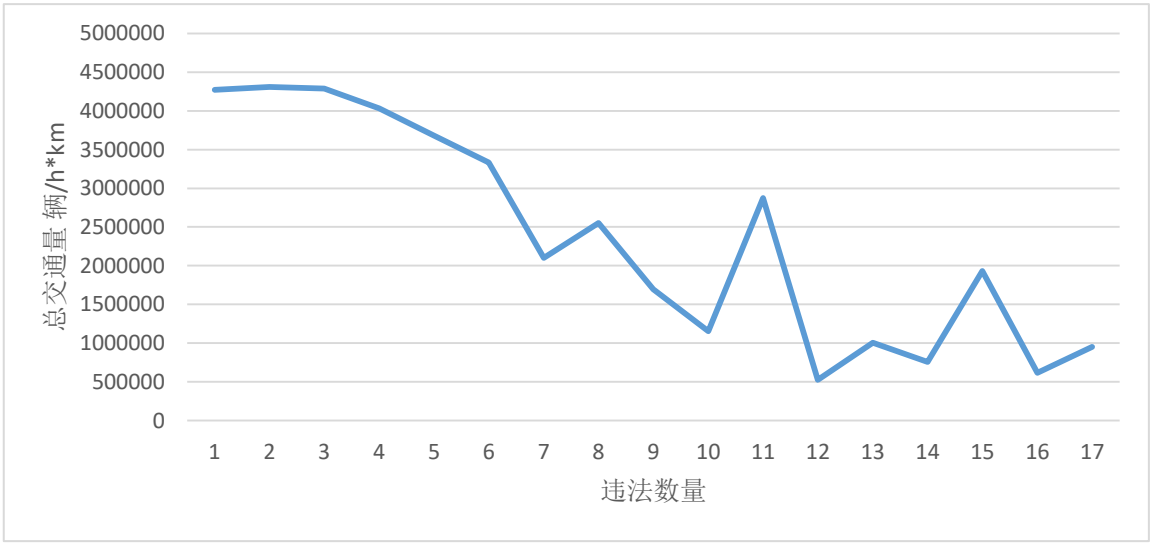


图 5-28 不同违法数量地区的平均总交通量

第六章　总结与展望

6.1　主要工作成果

本论文主要工作成果如下：

（1）参考交通安全相关文献，总结出交通违法行为的各类影响因素。然后通过集成各平台的规划、微波、违法等数据，以及一系列的特征工程处理，创新性地构建出交通违法数量预测的训练数据结构。

（2）通过多元统计分析中对应分析的方法，得到不同事故类型与违法类型的相关关系的检验方式和统计模型。

（3）分析并选取合适的机器学习回归预测模型，对某类型的交通违法行为的时空分布做了较为准确的预测。通过模型融合的方法，并提出基于聚类的分层融合模型，通过深层挖掘时间与空间之间的关联，进一步提高模型准确度。实验验证了所构建特征集的有效性和所用算法的预测效果。

（4）利用随机森林模型对交通流进行时序预测，从而构建了测试集的交通流和车速特征，为模型的测试提供较为可靠的数据。

（5）通过从模型中提取出较为重要的违法影响因素，以及对违法数量进行时空分布的可视化，总结规律，从而为交通执法部门的监督管理提供了客观依据和具体建议。

6.2　未来展望

本文研究了违法时空分布的预测方法，并得到了不错的预测结果。但本文仍存在不足之处，今后仍需改进：

（1）本实验的训练集只有五月份一个月的数据，之后有更多计算资源的情况下，应该将训练集数据扩展至一年，将月份季节等更多的时间特征纳入特征集。

（2）有许多违法行为的影响因素由于各种原因没有被纳入本特征集，例如驾驶员的相关数据；还有许多特征缺失值太多，填补效果很差，只能删去。因此，后续应进一步扩展有效特征集。

（3）可以选用更多的机器学习基模型进行训练预测，通过组合、调参，选出结果更优的组合模型。

（4）对于各违法行为影响较大的特征因素应被进一步详细分析，得到具体的非线性影响关系，使得违法行为的预防对策可以更直接有效。

致谢

在我的本科毕设论文即将完成之际，我却还没感受到畅快的情绪，心中思绪万千。在此，我想要感谢过去四年中在生活、学习中帮助过我的挚友、同学、亲人、老师们，是你们让我度过了这段特别又快乐的时光。

首先，我要感谢王晨教授和刘林等师兄指导我完成这次的毕设《基于机器学习的交通违法时空演化趋势预测研究》。由于我将不继续在东大读研也没有研究生导师，之前一直接触的课题组的毕设题目和我对自己预想的研究方向不太符合，当时选课题的时候再三纠结，直到王老师的这个课题的出现，我就觉得，“这个题目太适合我了”。事实证明，我没有选错。虽然我的进度一直都不快，但这次跟着王老师和学长尝试预测交通违法行为的过程还是挺愉快的。王老师给了我充分的自由去构思，也给了我很多有用的建议；而学长们也不厌其烦的给我讲数据，指导思路。对此，我非常感激。

其次，我还想感谢我在 tbs 实习的总工，王伟。虽然他没有给我提供直接的帮助，但我在公司弄毕设的东西，即使可能被发现了，他也没有指责我什么。知道我临近毕业事情比较多，近期的工作也一直不来催促我，对领导的这种宽容我抱一百分的感激之心。

还有，我要感谢我的同学，尤其是三位室友。感谢你们的快马加鞭，让经常不在学校，不和其他同学交流的我也能大致把握毕设的进度，不至于特别仓促。写论文写到头疼的时候，和你们唠唠嗑也是极好的。

另外，我还要感谢小柳。谢谢你对我生活上的百般照顾，你是我难过头疼时候的避风港，虽然我觉得我帮你毕设比你帮我多哈哈哈哈。

最后，感谢东南大学交通学院，祝母校越来越好，祝交院越来越好！

参考文献

[1] Jin J, Deng Y. A comparative study on traffic violation level prediction using different models[C]//Transportation Information and Safety (ICTIS), 2017 4th International Conference on. IEEE, 2017: 1134-1139.

[2] 汪贝. 电子执法环境下交通违法行为倾向模型研究[D].哈尔滨工业大学,2014.李晶. 基于数据挖掘与移动通信技术的高速公路违法分析研究[D].浙江工业大学,2009.

[3] 李晶. 基于数据挖掘与移动通信技术的高速公路违法分析研究[D].浙江工业大学,2009.

[4] 杜长海,杨民,魏丽丽.基于灰色模型的重庆市交通违法预测研究[J].警察技术,2016(05):89-91.

[5] 中华人民共和国道路交通安全法[N]. 人民日报,2011-08-06(006).

[6] 赵梨利. 道路交通违法行为的研究[D].西南交通大学,2014.

[7] 郭洪洋,韩雪松,刘澜,马亚峰.驾驶员交通安全行为可靠性风险度量研究[J].中国安全科学学报,2013,23(06):103-109.

[8] Ortet G, Ibáñez M I, Llerena A, et al. The underlying traits of the Karolinska Scales of Personality (KSP)[J]. European Journal of Psychological Assessment, 2002, 18(2): 139.

[9] Ayuso M, Guillén M, Alcañiz M. The impact of traffic violations on the estimated cost of traffic accidents with victims[J]. Accident Analysis & Prevention, 2010, 42(2): 709-717.

[10] Neyens D M, Boyle L N. The effect of distractions on the crash types of teenage drivers[J]. Accident Analysis & Prevention, 2007, 39(1): 206-212.

[11] Nallet N, Bernard M, Chiron M. Individuals taking a French driving licence points recovery course: Their attitudes towards violations[J]. Accident Analysis & Prevention, 2008, 40(6): 1836-1843.

[12] 贺超. 道路交通违法问题现状分析与对策研究[D].国防科学技术大学,2007.

[13] 胡家兴. 基于违法数据分析的道路交通安全管理决策研究与应用[D].大连海事大学,2011.

[14] Haddon Jr W. The changing approach to the epidemiology, prevention, and amelioration of trauma: the transition to approaches etiologically rather than descriptively based[J]. American journal of public health and the Nations health, 1968, 58(8): 1431-1438.

[15] Sabey B E, Taylor H. The known risks we run: The highway. Report SR567[J]. Transport Research Laboratory, Crowthorne, Berks, 1980.

[16] Fosgerau M. Speed and income[J]. Journal of Transport Economics and Policy (JTEP), 2005, 39(2): 225-240.

[17] 刘志强.道路交通安全研究方法[J].中国安全科学学报,2000(06):17-22.

[18] 郑世伟. 交通违章原因及预防[J]. 道路交通管理, 1998, 6.

[19] 孙轶轩. 基于数据挖掘的道路交通事故分析研究[D].北京交通大学,2014.

[20] Híjar M, Carrillo C, Flores M, et al. Risk factors in highway traffic accidents: a case control study[J]. Accident Analysis & Prevention, 2000, 32(5): 703-709.

[21] 周志华. 机器学习[M]. Qing hua da xue chu ban she, 2016.

[22] 邓磊. 基于机器学习的酒店价格预测分析[D].东南大学,2017.

[23] Hoerl A E, Kennard R W. Ridge regression: Biased estimation for nonorthogonal problems[J]. Technometrics, 1970, 12(1): 55-67.

[24] Price B. Ridge regression: Application to nonexperimental data[J]. Psychological Bulletin, 1977, 84(4):

759.

[25] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1996: 267-288.

[26] Zou H, Hastie T. Regularization and variable selection via the elastic net[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2005, 67(2): 301-320.

[27] Fonarow G C, Adams K F, Abraham W T, et al. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis[J]. Jama, 2005, 293(5): 572-580.

[28] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of statistics, 2001: 1189-1232.

[29] Chen T, Guestrin C. Xgboost: A scalable tree boosting system[C]//Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. ACM, 2016: 785-794.

[30] 高惠璇.应用多元统计分析[M].北京:北京大学出版社,2005,324-341

[31] 刘林,傅惠,吕伟韬等.交通事故类型和违法类型对应分析[J].中国安全科学学报,2017,27(11):79-84.

[32] 潘鸿,张小宇等.应用统计学[M].北京:人民邮电出版社,2011,256-275

[33] 王博,黄九鸣,贾焰等.适用于多种监督模型的特征选择方法研究[J].计算机研究与发展,2010,47(09):1548-1557.

[34] 柯国霖. 梯度提升决策树 (GBDT) 并行学习算法研究[D]. , 2016.

[35] Koren Y. The bellkor solution to the netflix grand prize[J]. Netflix prize documentation, 2009, 81: 1-10.

[36] Chiu, Stephen L. "Fuzzy model identification based on cluster estimation." Journal of Intelligent & fuzzy systems 2.3 (1994): 267-278.

[37] Sander, Jörg, et al. "Density-based clustering in spatial databases: The algorithm gdbscan and its applications." Data mining and knowledge discovery 2.2 (1998): 169-194.

[38] Liaw A, Wiener M. Classification and regression by randomForest[J]. R news, 2002, 2(3): 18-22.

[39] Hyndman R J, Koehler A B. Another look at measures of forecast accuracy[J]. International journal of forecasting, 2006, 22(4): 679-688.

[40] Armstrong J S, Collopy F. Error measures for generalizing about forecasting methods: Empirical comparisons[J]. International journal of forecasting, 1992, 8(1): 69-80.

[41] Pearson K. Note on regression and inheritance in the case of two parents[J]. Proceedings of the Royal Society of London, 1895, 58: 240-242.

[42] Fisher Box, Joan. Guinness, Gosset, Fisher, and Small Samples. Statistical Science. 1987, 2 (1): 45–52.

附录 A
原始数据字段解释

（1）微波数据字段

字段名	数据类型	备注
ID	NUMBER	编号 ID
FACILITY_ID	VARCHAR	设备编号
LANE_ID	NUMBER	路段编号
DATE_KEY	NUMBER	采集日期
TIME_KEY	NUMBER	采集时间
VOLUME	NUMBER	该段采集时间内通过的车辆数
SPEED	NUMBER	该段采集时间内车道平均速度（km/h）
WGS84_LNG	VARCHAR	设备位置经度（WGS84 坐标系）
WGS84_LAT	VARCHAR	设备位置纬度（WGS84 坐标系）

（2）交通中区的规划、人口、经济变量字段

字段名	数据类型	备注
TotalPop	Integer	总人口数
JobNum	Integer	就业岗位数
LANDACRG	Varchar2	用地面积
BUACREAGE	Varchar2	建筑面积
LocPopNum	Integer	户籍人口数
TempPopNum	Integer	暂住人口
floatPoNum	Integer	流动人口
moPkNum	Integer	机动车停车位
onGrdMoNum	Integer	地上机动车车位数
UnMotorNum	Integer	地下机动车停车数
HouseSum	Integer	总户数
EmptyHous	Integer	空置房数

字段名	数据类型	备注
LandChrtcs	Varchar2(128)	用地性质： U 公用设施用地 B 商业服务业设施用地 R 居住用地 A 公共管理与公共服务设施用地 H 建设用地 G 绿地与广场用地 M 工业用地 W 物流仓储用地 S 道路与交通设施用地

（3）非现场执法数据字段

字段	数据类型	备注
ID	VARCHAR	编号
PLATE_TYPE_ID	INT	违法车辆类型： 22 混合新能源小车 24 混合新能源大车 16 教练车 6 外籍车辆 1 大型汽车 2 小型汽车 51 大型新能源汽车 52 小型新能源汽车 21 纯电动新能源小车 23 纯电动新能源大车
PLATE_COLOR_ID	INT	违法车辆牌照类型： 0 蓝牌 1 黑牌 2 黄牌 3 白牌 99 其它
FACILITY_ID	VARCHAR	设备编号

字段	数据类型	备注
DEVICE_TYPE_ID	INT	设备类型： 1 电子警察 2 高清卡口 3 测速卡口 4 闭路电视 5 移动摄像 6 警务通 7 区间测速 8 卫星定位 9 其它设备 10 单行道
VIOLATION_TIME	DATETIME	违法时间
VIOLATION_TYPE_ID	INT	违法类型（共 60 种）： 详见 2.1.2
STATUS	INT	略 设备状态： 1 为正常 0 为故障

附录 B
模型特征集详细字段解释

特征来源	特征名	特征解释
时间特征	date	日期
	weekday	一周第几天
	hourperiod	当天第几个小时
	If_rush	是否为高峰期：“1”，是；“2”，否.下同
	If_workday	是否为工作日
2017 非现场执法数据	3vio_cnt	“违反标志指示”违法行为的发生次数
交通流和速度信息	tot_volume	总交通量=平均每条路的小时交通量*路网总长度
	speed	平均车速=该区域平均瞬时车速
规划信息	tot_pop	该中区总人口数
	tot_job	该中区总工作岗位数
	tot_landacrg	总面积
	tot_buacrg	总建筑面积
	tot_locpop	总当地人口
	tot_temppop	总暂住人口
	tot_floatpop	总浮动人口
	tot_pk	总停车位
	tot_onpk	地面总停车位
	tot_unpk	地下总停车位
	tot_house	房屋总数量
	tot_emptyhouse	空置房屋总数量
	A_acrg	公共管理与公共服务设施用地面积
	B_acrg	商业服务业设施用地面积
	G_acrg	绿地与广场用地面积
	H_acrg	建设用地面积
	M_acrg	工业用地面积
	R_acrg	居住用地面积
	U_acrg	公用设施用地
	Zhuanye_yes	该地区存在专业机构

路网信息	School_yes	该地区存在学校
	Wenti_yes	该地区存在文体场所
	Hospital_yes	该地区存在医院
	Market_yes	该地区存在大型市场
	Biz_yes	该地区存在商业区
	Res_yes	该地区存在居民场所
	Tot_len	该区域路网总长度
	Zhuganlu_len	主干道总长度
	Zhilu_len	支路总长度
	Ciganlu_len	次干路总长度
	Gongjiao_len	公交专用道总长度
	Gonglu_len	公路总长度
	Kuaisulu_len	快速路总长度
	Zanding_len	暂定路总长度
	Road_cnt	路段数量
	Tot_lanes	路网总车道数
	Avg_lanes	平均车道数
	Tot_cap	路网总通行能力
	Avg_speed	路段平均限速值
其余环境信息	If_rain	当天是否下雨
	High_t	当天最高气温
	Low_t	当天最低气温
	Avg_temperature	当天平均气温

附录 C
XGBoost 模型得到的特征重要性

字段名	字段含义	f-score 占比
tot_volume	总交通量	0.089208
speed	区域平均车速	0.084671
high_t	最高气温	0.053143
avg_temperature	平均气温	0.045989
low_t	最低气温	0.040216
tot_job	该中区总工作岗位数	0.024629
avg_cap	该中区路段平均通行能力	0.019815
U_acrg	公用设施用地	0.019705
tot_pop	该中区总人口数	0.019330
tot_buacrg	总建筑面积	0.019275
B_acrg	商业服务业设施用地面积	0.018943
A_acrg	公共管理与公共服务设施用地面积	0.018921
tot_landacrg	总面积	0.018800
zhuganlu_len	主干道长度	0.018546
avg_lanes	区域路段平均车道数	0.017265
avg_bks	平均公交车道数	0.016846
avg_speed	区域平均限速	0.016515
tot_len	路网总长度	0.016482
M_acrg	工业用地面积	0.016172
tot_locpop	当地人口数	0.016073
R_acrg	居住用地面积	0.015378
H_acrg	建设用地面积	0.015311
ciganlu_len	次干路长度	0.013446
zhilu_len	支路长度	0.012828
tot_cap	区域总通行能力	0.012750
if_rain	是否下雨	0.012375
road_cnt	路段数	0.012287
if_rush	是否在高峰期	0.011757
weekday_2	是否在周二	0.010035
tot_house	总房屋数	0.009582
tot_temppop	总暂住人口数	0.009405
tot_onpk	路面停车位数	0.009074
weekday_5	是否为周五	0.009008
tot_pk	总停车位数	0.008853
weekday_3	是否为周三	0.008688
weekday_0	是否为周日	0.008589
G_acrg	绿地与广场用地面积	0.008302
weekday_4	是否为周四	0.008103
weekday_6	是否为周六	0.008014
字段名	字段含义	f-score 占比

hourperiod_13	是否在第 13 个小时	0.007871
hourperiod_11	是否在第 11 个小时	0.007783
hourperiod_16	是否在第 16 个小时	0.007451
hourperiod_9	是否在第 9 个小时	0.007275
hourperiod_15	是否在第 15 个小时	0.007242
hourperiod_14	是否在第 14 个小时	0.007220
hourperiod_5	是否在第 5 个小时	0.007131
hourperiod_10	是否在第 10 个小时	0.006966
tot_lanes	总车道数	0.006933
hourperiod_12	是否在第 12 个小时	0.006911
weekday_1	是否为周一	0.006888
tot_emptyhouse	总空置房屋数	0.006822
hourperiod_7	是否在第 7 个小时	0.006624
hourperiod_23	是否在第 23 个小时	0.006546
hourperiod_8	是否在第 8 个小时	0.006381
tot_unpk	总地下停车位数	0.006259
hourperiod_17	是否在第 17 个小时	0.006116
hourperiod_2	是否在第 2 个小时	0.005740
hourperiod_1	是否在第 1 个小时	0.005575
hourperiod_3	是否在第 3 个小时	0.005520
gonglu_len	公路长度	0.005509
hourperiod_19	是否在第 19 个小时	0.005464
hourperiod_21	是否在第 21 个小时	0.005453
hourperiod_22	是否在第 22 个小时	0.005420
hourperiod_6	是否在第 6 个小时	0.005387
hourperiod_20	是否在第 20 个小时	0.005343
hourperiod_4	是否在第 4 个小时	0.005111
hourperiod_0	是否在第 0 个小时	0.005045
hourperiod_18	是否在第 18 个小时	0.005012
tot_floatpop	总浮动人口数	0.004890
hospital_yes	是否有医院	0.002594
res_yes	是否有居民区	0.002362
school_yes	是否有学校	0.001711
zhuanye_yes	是否有专业机构	0.001634
zanding_len	正修建道路长度	0.001181
kuaisulu_len	快速路长度	0.001060
biz_yes	是否有商业区	0.000883
wenti_yes	是否有文体中心	0.000188
market_yes	是否有大型市场	0.000033