大数据驱动下的共享单车短期需求预测

——基于机器学习模型的比较分析

焦志伦[1],金 红[2],刘秉镰[1],张子豪[2]

(1. 南开大学 经济与社会发展研究院,天津 300071;2. 美国堪讯企业咨询服务有限公司,纽约 NY10017)

摘 要: 基于共享单车项目的多维度大样本数据,以套索回归、岭回归、随机森林和迭代决策树等机器学习模型,探讨了共享单车短期(基于小时)需求预测的主要影响因素,并对不同模型预测效果进行了比较分析。研究结果发现,影响共享单车小时需求的主要因素包括特定的位置因素、时间因素以及天气条件因素。同时,相比普通线性回归、套索回归和岭回归模型,随机森林和迭代决策树模型对共享单车短期即时需求预测的结果更精确,在样本内部拟合和样本外推预测中的拟合优度(R^2)更高,标准误差(RMSE)更低,是共享单车行业短期实时需求精准预测的更有效手段。

关键词: 共享单车;大数据;需求预测;机器学习

中图分类号: F251 **文献标志码:** A **文章编号:** 1000-2154(2018)08-0016-10

DOI:10.14134/j.cnki.cn33-1336/f.2018.08.002

Short Term Bike-sharing Ridership Prediction under the Big-data Condition: Comparison of Machine Learning Models

JIAO Zhi-lun[1], JIN Hong[2], LIU Bing-lian[1], ZHANG Zi-hao[2]

(*1. College of Economic and Social Development, Nankai University, Tianjin 300071, China;*
2. Analytic Partners, Inc. New York, NY 10017, USA)

Abstract: Using the large and multidimensional data released by the bike-sharing project, and employing the Machine Learning Models, this article discussed the factors influenced short-term demand prediction of a bike-haring business. The results showed that the major factors that affected the short-term demand of bike sharing include specific location, time, and weather conditions. Meanwhile, compared with the Ordinary Linear Regression, Lasso Regression and Ridge Regression model, the Random Forest and Gradient Boosting Decision Tree models had higher goodness of fit (R^2) and lower standard error (RMSE) in both in sample and the out sample predictions, which shed lights on the machine learning models and are more suitable for short-term precise demand predictions.

Key words: bike sharing; big data; demand prediction; machine learning

一、引 言

随着信息技术的高速发展和广泛应用,互联网平台实现了多边供需对接和即时交易,推动了分时租赁

收稿日期: 2018-03-25

基金项目: 国家自然科学基金项目"考虑消费者行为的O2O服务企业决策优化与供应链协同研究"(71772095);"中国特色社会主义经济建设协同创新中心"项目支持;南开大学人文社会科学青年教师研究启动项目"互联网革命与物流业态变革研究"

作者简介: 焦志伦,男,讲师,经济学博士,主要从事交通经济、区域经济、物流管理研究;金红,男,经济学博士,主要从事大数据模型应用研究;刘秉镰,男,教授,博士生导师,经济学博士,主要从事交通经济、区域经济、物流管理研究;张子豪,男,博士,主要从事大数据模型应用研究。

和共享经济的高速发展。其中,共享单车是以互联网平台为基础的自行车分时租赁和共享服务。①共享单车新兴业态的发展,填补了城市居民交通出行"最后一公里"的服务链条,为广大消费者带来了出行便利。到2016年,全球范围内主要城市中约有超过1000个正在运营的共享单车项目,超过300个项目正在计划和建设中[1]。同时,在快速发展过程中,共享单车行业也为城市经济和社会带来了一些"负外部性",例如,一些城市的共享单车废弃车辆堆成"百里坟场",造成对社会资源的浪费;又如,共享单车在街头无序投放,影响了正常交通通行和城市形象。造成这些问题背后的原因,除了风险投资推动下的市场竞争之外,也与共享单车企业对具体时间地点短期需求预测不精确、资源调度不合理有关。如何利用大数据精确预测共享单车短期需求,从而科学合理地确定单车投放和调度安排,不仅是企业运营效率问题,也与社会资源合理使用、城市空间合理布局以及共享单车企业政府监管等问题密切相关。

在共享单车需求预测方面,除了技术代际差异和运营模式差异,②共享单车的需求可能受到宏观经济条件、收入水平和价格等因素的影响。对于经济社会条件相对固定(价格固定)环境下的短期共享单车需求,更细节的因素会起到主要作用,如 Campbell 等(2016)通过对北京共享单车项目的调查指出,影响共享单车需求因素主要有距离、气温、降水、空气质量等,用户自身的人口统计特征(含收入、性别、职业等)对单车需求没有明显影响[3]。Matton 和 Godavarthy(2017)指出气温、风力、降水等气候条件是影响共享单车需求的主要因素[4];Faghih-Imani 等(2014)提出,时点因素也是影响共享单车需求的重要变量,包括每天的时间段、是否周末、高峰时间等[5];一些文献也同时讨论了天气因素和时点因素[6-8]。此外,现有文献也讨论了地点相关的影响因素,主要包括人口密度[9-10]、自行车专用道设施情况[10-12]、与城市 CBD 和大学的距离[5,10,13]等。

从预测方法上看,现有文献对共享单车需求的预测通常采用的方法集中在传统线性 OLS 模型、二分类和多分类 Logit 模型[3,14]等。Kaspi 等(2016)提出了一个贝叶斯估计模型,用来预测某一站点存在的无法使用的单车数量[15]。Einav 和 Levin(2014)指出,信息技术使大规模运营管理层数据和私人部门数据的获得性逐步提升,为经济学的研究提供了新的机遇。应用大数据对具体问题进行实证分析和处理,需要新的工具和方法[16]。机器学习(Machine Learning,ML)是基于大数据的建模和分析方法,是"能通过经验积累自动改进的计算机算法"[17]。其中,机器学习工具体系中的监督学习主要关注"预测"问题,在大数据条件下,监督学习在预测方面具有较为显著的优势。

首先,机器学习模型的"样本内"(In-sample)拟合效果更好。Mullainathan 和 Spiess(2017)使用美国房产调查的多维度大样本数据对自有房产的对数美元价格进行拟合,拟合结果发现传统 OLS 的组内估计效果(R^2)为47.3%,但使用随机森林(Random Forest)等机器学习方法的估计效果均超过80%[18]。其次,样本外(Out-of-sample)预测效果更好。Bajari 等(2015)使用 IRI 市场研究数据中连锁百货商店的837460条数据进行估计,结果显示线性回归、条件 Logit 模型对样本外数据预测的标准误差(RMSE)分别为1.193和1.234,而表现更好的机器学习模型,如随机森林和支持向量机(SVM)的 RMSE 分别达到0.965和1.068[19]。第三,机器学习模型更适合处理含有大量协变量的多维数据。在 Bajari 等(2015)的模型中,如果允许店铺和产品固定效应存在,那模型将包含上千个解释变量,使用传统计量模型将造成估计效率下降且存在大量共线性问题,严重降低组内组外样本的预测水平。对此,Belloni 等(2014)指出,应用机器学习中的套索(LASSO)模型等可以很好地应对模型协变量过多的问题[20]。

目前,在经济管理学框架下的共享单车实践领域引入机器学习模型进行短期需求预测的文献相对较

---

①共享经济概念由 Felson 和 Spaeth(1978)首次提出[35],主要特点是个体通过第三方平台实现点对点(Peer to Peer)的直接交易。目前对共享经济的概念仍存不同理解。本文将自行车的分时租赁和共享服务统称为"共享单车"。

②按照 DeMaio(2009)的总结,在现代互联网技术应用到单车租赁行业之前,共享单车在技术和模式上经历过三代发展,包括1965年首次在阿姆斯特丹出现的第一代"白色单车"(White Bikes),第二代1995年在丹麦迅速发展的"城市单车"(City Bikes)和第三代1996年在英国出现的磁卡单车。目前,以互联网、定位技术、移动设备、网络支付为基础的新一代共享单车可称为第四代。

少,Bacciu 等(2017)采用机器学习中的支持向量机和随机森林模型预测了共享单车站点是否会在短时间内有单车归还[21],但没有系统讨论单车使用的短期需求问题。总体来看,有关共享单车需求预测的相关文献成果十分有限,同时对不同机器学习模型进行应用比较的研究更加缺乏。

本文尝试将套索回归、岭回归、随机森林和迭代决策树等机器学习模型引入共享单车短期需求预测的分析中,并比较这些模型与普通 OLS 回归在预测精度等方面的差异。本文的贡献在于,一是将大数据机器学习方法引入共享单车行业的"小时级"短期需求预测,提升行业对即时性需求的预测效率,从而辅助企业的实时调度,提高单车资源的整体利用水平;二是通过模型比较,系统讨论机器学习模型对共享单车短期需求预测的适用性,识别和评价多种不同机器学习模型之间的预测精度和预测效果。

二、研究设计与数据分析

(一) 模型选择

本研究关注不同区域共享单车短期(每小时时间段内)需求预测,按照 Faghih-Imani 等(2017)、Gebhart 和 Noland(2014)等现有文献的研究基础[5,7],本文将影响共享单车需求的协变量选取为具体的时间特征因素、天气条件因素和站点位置因素等。在设定各类因素的具体变量时,需要详细考虑相关时点的细分特征,如月份、日期、是否周末、是否法定假期、以小时区分的出行时段、是否高峰期等,这些涉及到对原始时间数据进行大量的清洗和分析处理工作。同理,对于天气和位置因素,也需要进行细分特征的变量处理。最终加入 OLS 模型的解释变量为75个,加入其他机器学习模型的变量可能有变化,如 Lasso 模型会压缩协变量数量,而以决策树为基础的组合模型则会增加变量,即包含了原有变量的交乘项、高阶项等。

在预测模型选取上,从需求变量自身出发的时间序列类预测模型在短期实时预测方面存在明显缺陷,这些模型包括最后期、趋势外推、自回归、移动平均、指数平滑等。对于依托其他单一或少量相关变量的预测方法,如弹性系数、增长系数、周期系数、重力模型等,在实际操作中可能会损失很多维度信息。相对于这些方法,基于小数据样本的灰度预测、传统回归分析等可以进一步捕捉更多协变量信息,但仍无法满足大数据条件下的分析需要,而基于大数据的机器学习模型和算法则在提升预测精度、控制"数据维度灾难"上更有优势[22]。各类预测方法特点及其主要特点如表1所示。

表1 主要预测方法及其主要特点比较

预测方法	类别	优点	局限	适用条件	示例或综述
德尔菲法	定性	不依赖历史数据;多轮交叉验证;	依赖于预测者的经验、能力等因素	历史数据材料有限	曾照云和程晓康(2016)[23]
弹性系数法	定量	仅需小样本数据	仅预测平稳趋势	预测变量、类比变量均需相对稳定	刘卫东等(2016)[24]
时间序列法	定量	仅需自身小样本数据	基于自身趋势预测,忽略其他变量影响	预测变量相对稳定,较少依赖其他变量	张钠等(2014)[25]
回归分析	定量	通过多维变量预测	大样本条件下预测效率下降	低维度小样本数据	刘涛雄和徐晓飞(2015)[26]
灰色预测	定量	需要数据少,短期预测更精准	远期预测误差大	小样本短期预测	何国华(2008)[27]
机器学习	定量	自学习自适应,预测精准	无法解释预测机理,需要控制学习过程中的过度识别	多维度大样本数据	Bacciu 等(2017)[21]

机器学习中有关预测的模型、算法也存在多种选择，且新的模型算法还在不断发展。目前监督学习中的预测模型主要包括套索回归（Lasso）、岭回归（Ridge）、支持向量机（SVM）、回归树（DT）等。此外，一些集成方法同时训练多个模型，在预测效果上可能更有优势。这些集成模型包括随机森林（RF）、迭代决策树（Gradient Boost Decision Tree，GBDT）等。

大数据条件下的共享单车需求预测具有样本量大、细分影响因素多的特征，如果将单车短期实时需求预测的目标时段设定为每小时，则仅该小时的时间特征影响因素就包含24个时段变量、7个星期变量、是否周末、是否其他法定假日等多个维度。因此，本文在机器学习模型选择时重点考虑了这一条件，最终同时选择了4个机器学习模型进行同步预测和比较，包括 Ridge、Lasso 两个回归模型以及 RF、GBDT 两种集成模型。Ridge 和 Lasso 在减少模型估计维度方面具有优势。RF、GBDT 两种集成模型则具有融合优势，在工业生产和一般服务需求类预测中应用较多，也常常表现出比其他机器学习模型更高的预测精度、性能和稳定性。

具体来看，在减少估计维度方面，Ridge 提供了应对多重共线问题（X 为奇异矩阵）的解决方法[28]，即提供一个二阶惩罚函数来获得精炼模型。

$$\hat{\beta}^{ridge} = \underset{\beta}{argmin}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta^2\right\} \tag{1}$$

其中惩罚函数的系数 λ 越大，估计矩阵的奇异性影响越小，估计参数 β 的估计值也逐步稳定。类似的，Lasso 回归也提供了带有惩罚函数的回归结果[29]。

$$\hat{\beta}^{lesso} = \underset{\beta}{argmin}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\} \tag{2}$$

不同的是，Lasso 是针对系数绝对值而非系数平方项进行惩罚。Lasso 可以看作是改进的 Ridge 方法，在预测方程的协变量较多时，可以通过 Lasso 的惩罚函数迫使部分协变量的估计系数为零，从而达到降低维度的目的。

RF、GBDT 两类集成模型都依托回归树算法，回归树算法是将数据的特征空间划分为若干决策区间（叶子），使得每一个区间都是空间中不相交的区域，然后汇报每个划分模块的函数均值[30]。在回归树基础上，RF 和 GBDT 更好地弥补了单个回归树功能简单且容易出现过度拟合的问题[31]。RF 是多个回归树组成的决策体系[32]，其中每棵树的生成都依赖随机选出的少量变量，最终的决策则通过对潜在随机向量树进行“投票”表决生成。在回归预测条件下，“投票”机制就是对这些树的结果进行平均，得到因变量预测值。类似的，GBDT 也是通过对多棵树的结果进行综合，不同的是每一棵树是从之前所有树的残差中来学习的，并以新树每个叶子的信息增益来进行最后全局预测[33]。

（二）数据说明

本文选取的数据为旧金山湾区共享单车项目数据。湾区共享单车项目（SFBay Area Bike Share）自2013年8月开始运营，前期投资700万美元，由湾区空气质量管理机构和城市交通机构管理，在旧金山、圣何塞、帕洛阿尔托等五个湾区城市70个报刊亭附近推出700辆自行车，自行车一半数量投入到旧金山，另外一半投入到其他城市，采用会员注册和有站模式（Station-based Bike Sharing，SBBS）管理，会员年费88美元。会员在30分钟通勤时间内归还自行车享受免费待遇。为游客设计的非会员3天和1天的无限次通票为22美元和9美元。①2016年8至9月份期间，该项目转由福特公司运营，并重新命名为“Ford GoBike”。本文选取项目运营前两年的数据进行预测研究，数据均来源于公开发布数据。②由于项目运营第一个月时，不同城市站点的安装启用时间不同，因此，本文数据最终选取的时间范围为2013年8月29日至2015年8月31日二年时间的运营数据，共669959个观测值。具体的变量名称和描述统计见表2。

①更详细的信息可见 http://kalw.org/post/sf-bay-area-bike-share-launches-thursday#stream/0。

②读者可以在 https://www.kaggle.com/benhamner/sf-bay-area-bike-share 上获取相关数据。

表2 主要变量及其描述统计

变量名称	观测数	平均数	含义
ID	669959	460382	记录编号,每条观测一个编号
Date	669959	—	具体时间,又分为两个变量,租车时间和还车时间,包含日期和时间,精确到分钟,可推算是否周末、假日、时段等其他时间特征
Station_id	669959	35	站点编号
Bike_id	669959	427.95	自行车编号
Station	669959	—	具体站点,又分为两个变量,租车站点和还车站点,可以分离出空间位置变量
Subscription_type	669959	—	客户类型,分为注册用户和临时用户
Duration	669959	1107.95	车辆使用的时间长度
Temper	669959	—	当天温度,又分为三个变量,最高、最低和平均温度,平均温度取值范围为38~84
Dew	669959	—	当天露点,同上又分为三个变量,平均露点取值范围为13~65
Humidity	669959	—	当天湿度,同上又分为三个变量,平均湿度取值范围为24~96
Pressure	669959	—	当天气压,同上又分为三个变量,平均气压取值范围为29.43~30.41
Visibility	669959	—	当天能见度,同上又分为三个变量,平均能见度取值范围为4~20
Wind_speed	669959	—	当天风速,又分为三个变量,最高、平均和阵风风速,平均风速取值范围为0~23
Cloud_cover	669959	2.78	当天云层覆盖率,取值范围为0~8
Precipitation	669959	0.02	当天降水,取值范围为0~3.36
Wind_dir	669959	266.61	当天风向,取值范围为0~360
Event	669959	—	天气事件,有五类取值,分别为雾、鱼、雾雨、雷雨和无事件

注:其他与本文结论相关度不高的变量没有进行描述。

初步考察共享单车需求的主要影响因素。图1是以日期和需求量为横纵坐标制作的散点图,并以局部加权回归画出了一条回归线。可以看到,不同日期的需求量总体上具有明显的二分化差异,引致这种差异的时间因素可能是工作日与周末的日期属性因素。图2进一步以工作日和周末划分样本,并绘制散点图及局部加权回归线,可以看到工作日(图中用 weekday 表示)的共享单车需求频次明显高于周末(图中用 weekend 表示),进一步体现了工作日与周末两个日期特征对需求的影响。同时,我们也看到工作日图中存在很多需求量较小的点,我们猜测这些点的出现与两种情况相关:一是当天存在极端天气;二是当天虽然不是周末,但却是一些法定节日,可以视同周末处理。除了天气外,本文建立需求模型时也纳入了这些法定节日的影响。①

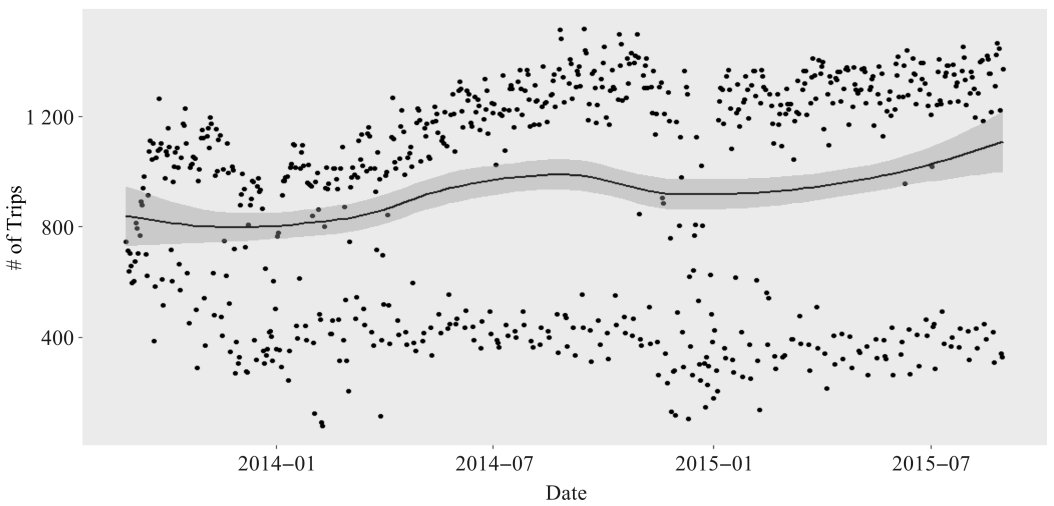

图1 日期与共享单车需求量的散点图和局部加权回归线

①最终纳入模型的有法定假期的节日每年有17天,分别按照样本对应的日期加入虚拟变量。

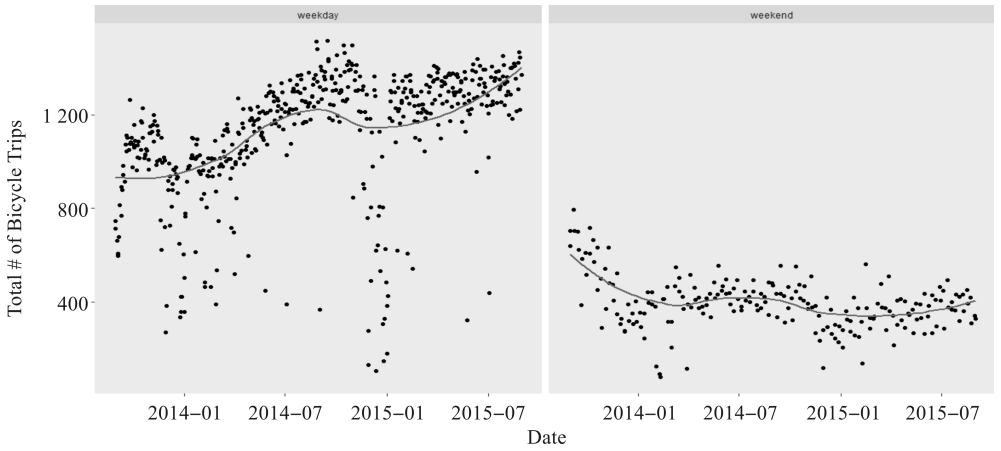

图2　工作日与周末两个日期属性与共享单车需求量的散点图和局部加权回归线

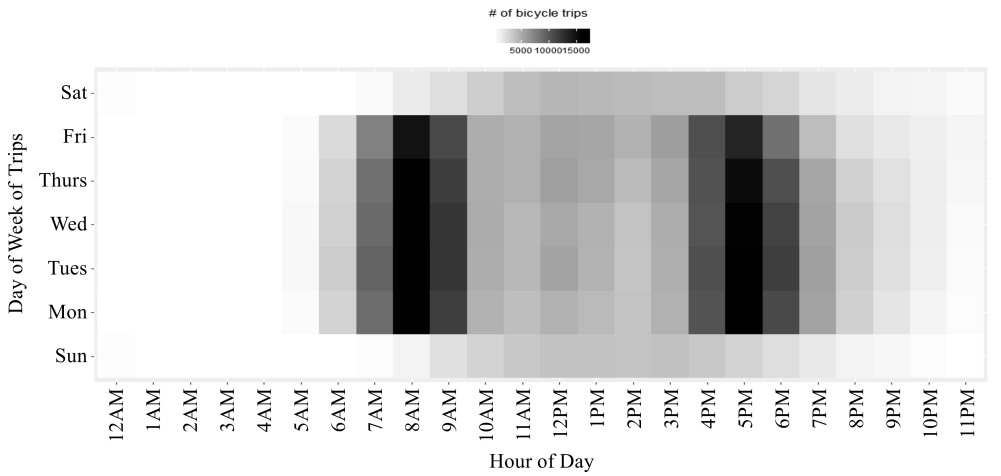

图3　时间段与共享单车需求量的关系

图3进一步考察了每天的不同时间段对共享单车需求的情况，如图所示，在每天的上午8点和下午5点的时候，需求量达到高峰，全样本每小时累积需求在15000车次左右，明显超过了其他时间段的需求，表明很多人使用单车是出于上下班通勤目的，时间段尤其是上下班高峰时间是影响单车需求的重要因素。此外，对于天气因素，我们预期很多天气变量存在一个舒适值区间，在区间内需求量较大，过高或过低的极端温度、湿度、风力等都会对需求有负向作用。对于地理位置因素，旧金山城区的需求量明显高于其他湾区城市。由于篇幅关系，对其他变量不再绘图分析讨论。

（三）算法实施与参数调整

本文关注基于机器学习模型的共享单车短期需求预测，对于 OLS 模型，本文暂没有采用引入动态因素的自回归、向量自回归、移动平均等时间序列因素，也没有采用面板估计模型。对于 OLS 的模型识别问题，引入模型的解释变量均为时间特征、天气特征和站点位置特征的变量，这些变量与模型残差相关性极为有限，且因果关系解释并不是本模型关注重点，因此没有采用相关识别策略。

在应用机器学习模型时，本文将样本划分为训练集样本（2013年8月29日至2015年3月30日样本，约占总观测值的78.19%）和测试集样本（剩余观测值）。在对训练集样本进行交叉验证时，Athey 和 Imbens（2017）建议划分 k（例如 k = 10）组子样本，留下其中第 m 组，并将其余的子样本组进行模型估计，并将拟合模型应用于留下的子样本组 m。重复迭代模型，最终选择的正则化调整参数为交叉验证模型残差平方和最小的模型参数[34]。本文对机器学习模型采用 k = 10的交叉验证设定。

为了防止过度拟合问题，机器学习模型需要进行模型的正则化调整，设置最佳模型参数。对于 Lasso

和 Ridge，首先通过10组子样本的交叉验证绘制 CV 曲线图寻找最佳的惩罚函数系数 λ，可见图4和图5，依次表示采用最小 MSE 判定获取的适度简洁模型。其中，Lasso 模型结果显示，采用最小 MSE 取值的 λmin = 0.005和采用1倍误差取值的 λ1se = 0.191，分别对应74个和47个解释变量，其余变量系数被设为0。相应的，Ridge 模型 CV 图计算的 λ 分别为1.822和3.184，Ridge 模型包含的变量个数均为75个。

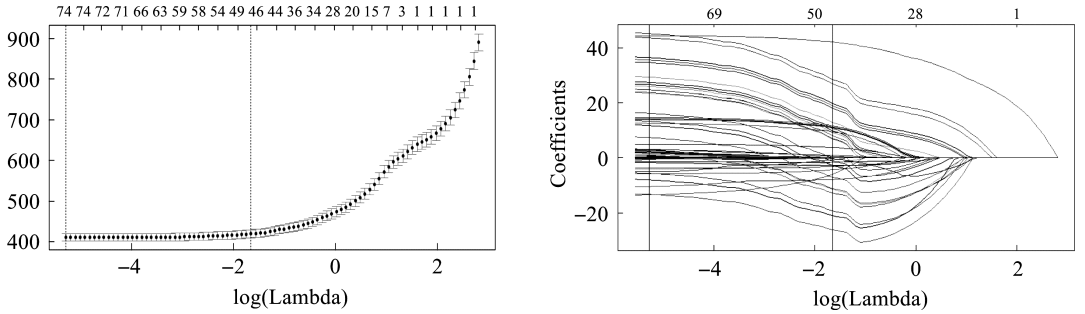

图4 Lasso 的 CV 交叉验证图(左)和系数变化图(右)

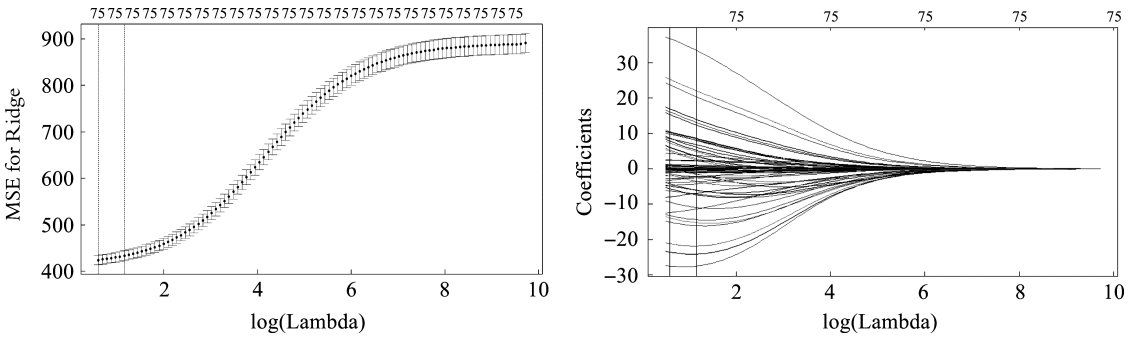

图5 Ridge 的 CV 交叉验证曲线图(左)和系数变化图(右)

对于 RF 和 GBDT，设定决策树的数量参数为 $n = 50$，每棵树参与分类选择的变量个数 m 分别设为 {10,20,30}，在逐个建立森林中的每棵树 i 时，将用自助法选择 $m_i \leqslant m$ 个预测变量，以信息增益最大原则为分类属性进行节点分割，建立一个无需修剪的深度最大的回归树。建立 n 棵回归树之后，由这些树最后结果的均值作为因变量预测值。对于 GBDT，设定学习速度的 eta 值为0.15。RF 和 GBDT 参数调整的指标选择为组内标准误差(RMSE)。最终模型效果的评估依靠测试集样本的 R^2 和 RMSE 来衡量，R^2 越大和 RMSE 越小的模型，表明拟合优度越高和标准误差越小，在预测共享单车短期需求方面具有更大优势。

三、模型评估与预测结果

表3汇报了不同模型下训练集样本与测试集样本的估计精确度。从比较结果可以看出，应用 OLS 模型，以训练样本拟合，拟合模型的可决系数 R^2 为0.5418，RMSE 为20.2086，将拟合模型应用到测试集样本，样本外预测结果 R^2 为0.3638，RMSE 为24.8286。具体比较，测试集的 R^2 下降17.8个百分点，RMSE 提升4.62个单位，表明与训练集结果相比，测

表3 主要变量及其描述统计

模型	训练集样本		测试集样本		测试集−训练集	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
OLS	0.5418	20.2086	0.3638	24.8286	−0.1780	4.6200
Lasso	0.5392	20.2106	0.3612	24.8261	−0.1780	4.6155
与 OLS 比较	−0.0026	0.0020	−0.0026	−0.0025	—	—
Radge	0.4470	20.5515	0.3021	25.4617	−0.1449	4.9102
与 OLS 比较	−0.0948	0.3429	−0.0617	0.6331	—	—
Random Forest	0.9383	3.4719	0.7495	8.4688	−0.1888	4.9969
与 OLS 比较	0.3965	−16.7367	0.3857	−16.3598	—	—
Gradient Boost	0.9812	0.8364	0.6576	10.8384	−0.3236	10.002
与 OLS 比较	0.4394	−19.3722	0.2938	−13.9902	—	—

试集预测拟合优度大幅下降、误差大幅提升,即 OLS 模型的样本外预测效果明显变差。

对比机器学习模型的估计结果,Lasso 和 Radge 对训练集样本的预测效率没有提升,对测试集样本的预测效果也没有明显变化。具体来看,Lasso 和 Ridge 组内预测的 R^2有所下降,RMSE 有所提升,预测总体效果下降。在组外预测方面,Lasso 预测的 RMSE 有所提升,但提升幅度不大,Ridge 预测效果在 RMSE 误差方面增大,预测效果欠佳。总体上讲,Lasso 与 Radge 两个模型在设计上主要是解决模型变量维度过多的问题,在对共享单车需求预测中,由于模型选取依据了经济学的基本理论,非直接影响的变量基本没有选取,其自变量之间的共线性问题也并不突出,因此没有体现出这类模型的优势。

从 RF 和 GBDT 两种基于决策树的机器学习模型来看,这两个模型在预测共享单车需求方面比之前的模型效果存在较为明显的改进。从训练集的样本内估计来看,两个模型的 R^2比普通 OLS 回归模型有大幅提升,RMSE 比 OLS 模型的误差有大幅降低。从样本外预测效果来看,RF 和 GBDT 模型在 R^2上提升分别达到约39和29个百分点,在 RMSE 误差降低方面分别达到16和14个单位。基于此可以认为,相比 OLS 等其他模型,这两个集成模型在样本内拟合和样本外预测方面都具有较大优势。

对 RF 和 GBDT 这两个模型进行比较,在当前样本条件和模型设置下,GBDT 比 RF 在样本内预测的效率更高,其中 GBDT 的样本内 RMSE 可以降低到0.8364。但从样本外预测效果来看,RF 比 GBDT 在样本外预测的效果更佳,其中 RF 模型的 R^2比 GBDT 模型高约10个百分点,RMSE 低约2.4个单位。此外,GBDT 模型的样本外误差与样本内误差的差距较大,达到10.002个单位,表明 GBDT 模型的设置存在一定的过渡拟合问题。

对于影响共享单车需求的主要因素,不同模型的结论也存在差异。OLS 回归、Lasso 和 Ridge 模型可以估计出模型协变量的回归系数,OLS 模型还能获得 t 统计量汇报的显著性指标,相比之下,RF 和 GBDT 不能确定各个参数的系数,但是可以通过算法实施过程中的某些指标获得变量的相对重要性。从模型结果来看,对共享单车需求量影响最大的前五名变量(表4)中,OLS、Lasso 和 Ridge 指向了相同的五个变量(在重要性次序上稍有不同),包括上午8点、9点(hour8. AM,hour9. AM)和下午4点、5点(hour4. PM,hour5. PM)两个上下班通勤高峰期的四个时间段变量和特定空间位置(旧金山城市)变量。RF 和 GBDT 则将工作日特征(weekday1)和周末(weekday7)作为仅次于特定城市、上下班高峰时间段的重要影响指标。

进一步拓展影响需求预测重要变量的范围,从对共享单车需求量影响最大的前十名变量来看(见表5),OLS、Lasso 和 Ridge 模型表明,对共享单车需求量影响最大的前十名变量主要是位置(旧金山城市)和时间特征变量,其中时间特征变量主要集中在高峰时间段方面。RF 和 GBDT 模型选

表4　影响共享单车需求的主要因素及其指标(前5名变量)

模型	变量名称	系数	重要性指标	重要性值
OLS	hour8. AM	46.6632	t 统计量	50.456
	hour5. PM	45.1450		49.199
	citySan Francisco	44.4856		104.425
	hour9. AM	37.9518		41.251
	hour4. PM	37.1078		40.289
Lasso	hour8. AM	45.3442	非零系数及其大小	45.3442
	citySan Francisco	44.3354		44.3354
	hour5. PM	43.8285		43.8285
	hour9. AM	36.6396		36.6396
	hour4. PM	35.7923		35.7923
Radge	citySan Francisco	36.8528	系数大小	36.8528
	hour8. AM	25.5069		25.5069
	hour5. PM	23.8375		23.8375
	hour9. AM	17.1463		17.1463
	hour4. PM	16.2616		16.2616
Random Forest	citySan Francisco	—	IncNodePurity	6728441.5
	hour8. AM	—		2663179.6
	hour5. PM	—		2289546.2
	Weekday7	—		1969220.4
	Weekday1	—		1820289.1
Gradient Boost	citySan Francisco	—	信息增益	0.2306
	hour8. AM	—		0.0825
	Weekday7	—		0.0731
	hour5. PM	—		0.0634
	Weekday1	—		0.0624

出的重要变量则综合包含了位置、时间和天气特征，在位置变量上，RF 和 GBDT 模型选择了旧金山和 San Jose，在时间变量上选择了高峰时段、工作日和周末变量，在天气特征上选择了风向（wind_dir_degrees）和最高气温（max_temperature_f）两个变量。

表5 影响共享单车需求的主要因素（前10名变量）

模型	排序	主要影响变量，按重要性排序
OLS	1~5	hour8. AM, hour5. PM, citySan Francisco, hour9. AM, hour4. PM,
	6~10	hour6. PM, hour7. AM, hour12. PM, hour3. PM, hour1. PM
Lasso	1~5	hour8. AM, citySan Francisco, hour5. PM, hour9. AM, hour4. PM
	6~10	hour6. PM, hour7. AM, hour12 PM, hour3 PM, hour1 PM
Radge	1~5	citySan Francisco, hour8. AM, hour5. PM, hour9. AM, hour4. PM
	6~10	hour6. PM, hour7. AM, weekday3, weekday4, weekday5
RandomForest	1~5	citySan Francisco, hour8. AM, hour5. PM, weekday7, weekday1
	6~10	hour9. AM, hour6. PM, hour4. PM, citySan. Jose, wind_dir_degrees
GradientBoost	1~5	citySan Francisco, hour8. AM, weekday7, hour5. PM, weekday1
	6~10	hour9. AM, hour4. PM, wind_dir_degrees, hour6. PM, max_temperature_f

整体来看，我们认为影响短期（每小时）共享单车需求的重要因素涵盖位置特征、高峰时段特征、工作日特征和天气特征。具体来看，各因素中都存在影响需求的主要变量，其中，位置特征主要变量为是否为旧金山站点；高峰时段特征的主要变量为上午8点、下午5点，其次是与这两个时间段相邻的时段；工作日特征主要变量是周日或周一（周六变量影响较弱），天气特征最主要变量是最高气温和风向。在模型选择上，通过以上因素预测基于小时的短期共享单车需求，采用 RF 和 GBDT 方法的模型预测效率更高，即在样本内拟合和样本外预测中得到的拟合优度较高，标准误差较小。进一步来说，在现有样本和模型设定下，RF 取得的预测效果最好，可以作为共享单车短期实时需求预测的重要工具。

四、结论及展望

通过旧金山湾区共享单车项目的669959个样本观测值，综合采用线性最小二乘回归（OLS）和机器学习模型中的套索回归（Lasso）、岭回归（Ridge）、随机森林（RF）和迭代决策树（GBDT）模型，本文探讨了共享单车短期（每小时）需求预测的影响因素和模型设计问题。研究发现，首先，从共享单车的需求影响因素来看，共享单车短期需求的主要影响因素包括位置因素（是否属于旧金山）、高峰时间段因素（上午8点和下午5点）、工作日因素（周日和周一）以及天气条件因素（最高温度和风向）。

其次，从预测模型比较来看，相比与 OLS、Lasso 和 Ridge 模型，RF 和 GBDT 两类集成模型在预测共享单车需求时具有较高的拟合优度和较低的标准误差。原因在于，RF 和 GBDT 模型在进行模型预测分析时能够综合考虑模型协变量之间的相互作用，因而观测到的影响因素也更加广泛。这是此类机器学习模型在算法上的优势，例如，OLS 模型能够观测到高峰时段（如 hour8. AM）的重要影响，但该变量在叠加周末、假日时的影响会有所减弱，在同时叠加周末和极端天气特征时，影响则进一步减弱。这是 OLS 模型在预测过程中无法考量的问题。即使按照一些理论的指导，在 OLS 模型中添加高峰时段与周末等因素的交乘项，但也可能忽略一些理论尚未发现的交互作用或高阶作用，因而限制了 OLS 模型的预测效果。此外，Lasso 和 Ridge 模型的优势在于处理协变量数量过多或变量之间存在多重共线的情况，对于变量之间的交互作用也缺乏处理，对于本研究的共享单车需求预测，这两类模型预测并不具有优势，因而预测效果与 OLS 相当。

共享单车已经成为多个国家的新兴经济业态，共享单车的需求预测问题，与企业车辆投放、调配及社会公共资源的合理利用密切相关。在该领域未来的研究中，可继续扩展到不同商业模式下共享单车的需求预测差异。例如，对于有站模式和无站模式的需求预测，主要影响因素可能有所不同，无站模式下，运营商对供需不平衡区域进行调度的能力可能成为单车需求的现实约束，需求的时点因素和空间因素可能存在更多类型的叠加机制。此外，加快将机器学习模型应用扩展到其他领域的预测研究中，发现更多细节因素变量的作用机制，也可能提升预测研究对其他经济解释型研究和因果效应研究的助力作用。

参考文献:

[1] MEDDIN R, DEMAIO P. The bike sharing world map[EB/OL]. (2007-01-01)[2017-09-10]. http://www.metrobike.net/the-bike-sharing-world-map/.

[2] DEMAIO P. Bike-sharing: history, impacts, models or provision and future[J]. Journal of Public Transportation, 2009, 12(4): 41-56.

[3] CAMPBELLA, CHERRY C, RYERSON M, et al. Factors influencing the choice of shared bicycles and shared electric bikes in Beijing[J]. Transportation Research Part C, 2016, 67(6): 399-414.

[4] MATTSON J, GODAVARTHY R. Bike share in Fargo, North Dakota: keys to success and factors affecting ridership[J]. Sustainable Cities and Society, 2017, 34(10): 174-182.

[5] FAGHIH-IMANI A, ELURU N, EL-GENEIDY A, et al. How landuse and urban form impact bicycle flows: evidence from the bicycle-sharing system(BIXI) in Montreal[J]. Journal of Transport Geography, 2014, 41(12): 306-314.

[6] NOSAL T, MIRANDA-MORENO. The effect of weather on the use of North American bicycle facilities: a multi-city analysis using automatic counts[J]. Transportation Research Part A, 2014, 66(8): 213-225.

[7] GEBHART K, NOLAND R. The impact of weather conditions on bikeshare tirps in Washington D C[J]. Transportant, 2014, 41(6): 1205-1225.

[8] FAGHIH-IMANI A, HAMPSHIRE R, MARLA L, et al. An empirical analysis of bike sharing usage and rebalancing: evidence from Barcelona and Seville[J]. Transportation Research Part A, 2017, 97(3): 177-191.

[9] RIXEYR. Station-level rorecasting of bikesharing ridership: station network effects in three U. S. systems[J]. Transportation Research Record: Journal of the Transportation Research Board, 2013, 2387(1): 46-55.

[10] WANG X, LINDSEY G, SCHONER J, et al. Modelling bike share station activity: effects of nearby businesses and jobs on trips to and from stations[J]. Journal of Urban Planing and Development, 2016, 142(1): 1-9.

[11] EL-ASSI W, MAHMOUD M, HABIB K. Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto[J]. Transportation, 2017, 44(3): 589-613.

[12] FISHMAN E, WASHINGTON S, HAWORTH N, et al. Barriers to bikesharing: an analysis from Melbourne and Brisbane[J]. Journal of Transport Geography, 2014, 41(12): 325-337.

[13] COCK J. Bike share in small and medium-sized cities[C]. Washington D C: Presentation at 2016 Transportation Research Board Tools of the Trade Conference, 2016.

[14] TANG Y, PAN H, FEI Y. Research on user's frequency of ride in Shanghai Minhang Bikesharing System[C]. Shanghai: World Conference on Transport Research, 2016.

[15] KASPI M, RAVIV T, TZUR M. Detection of unusable bicycles in bike-sharing systems[J]. The International Journal of Management Science, 2016, 65(12): 10-16.

[16] EINAV L, LEVIN J. Economics in the age of big data[J]. Science, 2014, 346(11): 715-721.

[17] MITCHELL T. Machine learning[M]. New York: McGraw Hill, 1997: 2.

[18] MULLAINATHAN S, SPIESS J. Machine learning: an applied econometric approach[J]. Journal of Economic Perspectives, 2017, 31(2): 87-106.

[19] BAJARI P, NEKIPELOV D, RYANS P, et al. Machine learning methods for demand estimation[J]. American Economic Review, 2015, 105(5): 481-485.

[20] BELLONI A, CHERNOZHUKOV V, HANSEN C. High-dimensional methods and inference on structural and treatment effects [J]. Journal of Economic Perspectives, 2014, 28(2): 29-50.

[21] BACCIU D, CARTA A, GNESI S, et al. An experience in using machine learning for short-term predictions in smart transportation systems[J]. Journal of Logical and Algeraic Methods in Programming, 2017, 87(2): 52-66.

[22] 刘涛雄,徐晓飞. 大数据与宏观经济分析研究综述[J]. 国外理论动态,2015(1):57-64.

[23] 曾照云,程晓康. 德尔菲法应用研究中存在的问题分析——基于 38 种 CSSCI(2014-2015)来源期刊[J]. 图书情报工作,2016(16):116-120.

[24] 刘卫东,仲伟周,石清. 2020 年中国能源消费总量预测——基于定基能源消费弹性系数法[J]. 资源科学,2016(4):658-664.

(下转第 35 页)