**大数据条件下的共享单车短期需求预测**

**——基于机器学习的模型比较**

焦志伦[[1]](#footnote-1) 金红2 张子豪2（中文名待验证）

（1. 南开大学经济与社会发展研究院，天津，300071）

2. Analytic Partners, Inc. 纽约，NY 10017

**中文摘要：（待续）**。

**关键词：共享单车；大数据；需求预测；机器学习**

## 一、引言

数据革命对经济研究产生了深远的影响。Einav和Levin（2014）指出，信息技术使大规模的管理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机遇。借助互联网、传感器和云存储技术的发展，经济学家们有机会接触到海量的微观层面、实时产生和行为属性的数据，这使得当代经济学研究，尤其是实证研究开始具备之前无法想象的研究基础和研究优势。同时，随之而来的问题是对于大数据的评价、处理能力亟待提升，尤其是新的分析工具和分析方法的引入。

机器学习（Machine Learning，ML）是基于大数据的建模和分析方法，是“对能通过经验积累自动改进的计算机算法的研究”（Mitchell T.M.，1997）。机器学习分为监督学习与非监督学习，非监督学习没有训练集样本，主要关注“发现一般类型（Pattern）”的问题，即从海量非结构数据中搜索某种实现没被关注的结构或规律，如进行聚类、关联规则等；监督学习主要关注“预测”问题，包括分类和回归。首先通过训练集样本（Training Sample），由人为设定的“学什么”和“怎么学”来进行学习，然后将一定标准下的最优函数类型应用到测试集样本（Test Sample），以防止过渡关注对训练集样本的组内估计效率而造成组外估计效率的下降，即机器学习领域的“过度识别”问题（Varian，2014）。

与计量经济学一样，监督学习中的回归预测也是通过对一系列解释变量的拟合来预测被解释变量。但不同的是，在机器学习方法中，解释变量影响被解释变量的模型形式和具体参数都是无法得到的，或者说是基于高维空间拟合多种模型形式和参数的并行结构，因此机器学习的预测模型常被称为“黑箱”。这与经济学领域传统实证研究倡导的“理论驱动”的建模方式有所差异。可以说，机器学习首先抛开了变量间彼此影响的具体逻辑，以海量数据构建的多重联系模式综合实现模型结果，更多体现了数据驱动思维下的模型选择思维。例如，在监督学习中，模型主要通过将训练集划分子样本并以交叉验证（Cross-Validation）方式来选择模型“调整”参数，并主要关注预测效果而不在意估计量的渐进性质，并常常因为减少预测误差而牺牲模型估计量的无偏性（Athey和Imbens，2017）。此外，在计量经济学“可信性”革命背景下，计量经济模型的因果推断受到普遍重视，但监督学习则由于“黑箱”的存在，无法确定协变量的估计参数和模型形式，进而很难直接进行理论分析和因果推断。

尽管如此，在大数据条件下，机器学习在“发现一般类型”和“预测”方面仍然具有传统方法无法比拟的优势。以监督学习的“预测”模型为例，首先，机器学习模型的“样本内”（In-sample）拟合效果更好。Mullainathan和Spiess（2017）使用美国房产调查数据对自有房产的对数美元价格进行拟合，使用数据包含150个协变量和10000个随机抽样的训练集样本，拟合结果发现传统OLS的组内估计效果（R2）为47.3%，但使用随机森林（Random Forest）和集成学习（Ensemble Learning）两种机器学习方法的估计效果分别达到85.1%和80.4%。其次，样本外（Out-of-sample）预测效果更好。同样在Mullainathan和Spiess（2017）的例子中，作者在以训练样本拟合模型后，又以41808个样本作为测试集观测房产价格预测的样本外预测效果，发现两种机器学习模型的样本外预测效果比传统OLS预测高出3.8和4.2个百分点。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。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。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（Bajari等，2015），但应用机器学习中的LASSO模型等可以很好的应对模型协变量过多的问题（Belloni, Chernozhukov和Hanson，2014）。

通过整合机器学习和因果推断的各自优势，经济学家已经开始探索将机器学习应用在计量经济学研究领域（Athey，2015）。目前主要的应用集中在经济预测（Bajari等，2015）以及和预测相关的因果推断模型辅助估计（Athey和Imbens, 2017）方面。例如在经济预测方面，Chalfin等（2016）利用机器学习模型的预测优势讨论了警察雇佣和教师职位升级两个案例，并认为运用机器学习预测方法进行了员工雇佣和晋升设计可以提升生产率，带来社会福利效应。在辅助因果推断方面，Lee，Lessler和Stuart（2010）讨论了使用机器学习的分类与回归树（CART）方法提高倾向匹配评分的权重估计效果。Athey，Imbens和Wager（2016）基于机器学习方法开发了一类结合模型调整和平衡的估计量，调整后模型可用来估计处理组的反事实结果，进而获取平均处理效应。此外，在估计个体异质性因果效应中，机器学习方法也体现出一定优势，如Athey和Imbens（2016）提出的回归树分析，Wager和Athey（2015）提出的随机森林法，Athey，Tibshirani和Wager（2016）提出的Gradient Forest法，等等。总体来看，经济学研究领域应用机器学习方法的研究尚处于起步阶段，国内经济学领域应用机器学习方法的研究文献也极为少见。

预测是经济学的基本问题之一，计量模型理论研究对预测方法的探索也从未停止。在需求预测领域，随着行业、企业和个人行为层面大数据的不断积累和应用，需求预测研究正在从长期宏观总体预测转入更加具体的短期中微观预测领域（Deb，2017）。本文主要尝试将机器学习的方法应用于中短期的行业需求预测，并选取共享单车行业的需求作为分析对象。此外，本文将对比普通OLS计量模型与不同的机器学习模型在预测效果上的差异，论述不同模型在预测共享单车行业需求方面的各自优势和分析结论。本文第二部分回顾现有文献对共享单车需求影响因素的讨论，为模型变量选取提供理论支持。第三部分提出研究设计和数据说明，第四部分分析模型结果，最后提出本文研究结论，并简要提出机器学习在经济学领域的研究愿景和个人建议。

## 二、共享单车需求及其影响因素

互联网平台多边即时交易的特征使分时租赁和共享经济获得了规模化发展的技术基础。共享单车一般指以互联网平台为基础的自行车分时租赁和共享服务[[2]](#footnote-2)。在最近半个世纪中，共享单车经历过三代（DeMaio，2009），第一代出现在1965年的阿姆斯特丹，为无站模式（Free-floating bike sharing，FFBS），名为“白色单车”（White Bikes），但运营中出现诸多问题，如自行车被丢弃到运河中或沦为自用等，很快便结束项目；第二代相继出现在丹麦的几座城市，称为“城市单车”（City Bikes），为有站模式（Station-based Bike Sharing，SBBS），起初规模较小，1995年前后在哥本哈根发展壮大，已经采用实心轮胎和广告版设计，以投币方式租赁使用。第三代共享单车首先出现在1996年的英国的普次茅斯大学校园，采用磁卡租赁，并融入电子锁、电信系统、车载电脑等功能。目前，包括中国在内，以互联网、定位技术、移动设备、网络支付为基础的新一代（可称为第四代）共享单车获得蓬勃发展，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，同时超过300个项目正在计划和建设中（Meddin和De- Maio，2016）。

不同技术代际差异、运营模式和时空范围下，共享单车需求预测的影响因素存在差异。技术代际差异和运营模式会从根本上影响使用体验，进而决定或影响了需求的整体量级。长期和大范围的需求可能受到宏观经济条件变化的影响。例如基础的微观经济学会首先指出收入水平和价格等因素会影响需求总量。对于特定技术代际下的短期共享单车需求，起主要作用的因素会有所不同，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。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，同样指出天气条件影响的文献还包括Nosal和Miranda-Moreno（2014）、Gebhart和Noland（2014）等。Faghih-Imani等（2014）提出，时间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，持相同观点的文献包括Faghih-Imani等（2017）、Gebhart和Noland（2014）等。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（Rixey，2013；Wang等，2016）、自行车专用道设施情况（El-Assi等，2015；Fishman等，2014；Wang等，2016）、与城市CBD和大学的距离（Faghih-Imani等，2014；Wang等，2016）等。一些研究提出了特殊类型单车数量的预测，如Cock（2016）指出小型城市共享单车项目需求受到是否接近校园、商圈和城市地形的影响。Kaspi，Raviv和Tzur（2015）提出了预测某一站点存在无法使用单车数量的贝叶斯估计模型。

从方法上看，目前对共享单车需求预测的研究相对较少，对共享单车需求影响因素的相关研究也集中在传统线性OLS模型、二分类和多分类logit模型（如Tang，2016；Campbell等，2016）等。Bacciu等（2017）采用机器学习中的支持向量机和随机森林模型分析了SBBS模式下共享单车站点是否会在短时间内有单车归还的问题，结论发现预测效果为15分钟内的平均预测准确率达到0.73。

总体来看，在共享单车需求预测方面，目前经济学的研究集中在主要影响因素的搜索和验证方面，这些研究为本文变量和数据选取提供了支持。在共享单车需求预测的研究中，机器学习模型的应用也处在起步研究阶段。本文将尝试通过模型比较，讨论机器学习相关模型在共享单车需求预测领域的应用。

## 三、研究设计与数据分析

### （一）模型设计

本研究关注单一区域共享单车短期（以天为单位）需求预测，按照Faghih-Imani等（2017）、Gebhart和Noland（2014）等现有文献的研究基础，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和天气条件因素。对于固定地区，人口密度、单车基础设施、与城市CBD和大学的距、居民收入、使用者的人口统计特征等影响因素在短期内是相对固定的，可以作为固定常数加入模型，因此省略这些变量，这种处理也存在相关文献支持，如Campbell等（2016）。在设定时间因素时，需要确定相关日期的各类时间特征变量，包括当日的季节、月份、是否周末、是否法定假期、以小时区分的出行时段等，天气变量包括温度、降水、风力、气压等，最终加入OLS模型的解释变量为72个，加入其他机器学习模型的变量可能更多，因为可能包含了这些变量的交乘项、高阶项等。

在预测模型选取上，从需求量变量自身出发的纵向预测模型在短期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，如增长系数、周期系数、重力模型等，在大数据条件下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、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（刘涛雄和徐晓飞，2015）。

监督学习中有关预测的方法主要包括Lasso（套索回归）、Ridge（岭回归）、Regression Tree（回归树）、RF（Random Forest，随机森林）、GBDT（Gradient Boost Decision Tree，迭代决策树）等。其中，Ridge提供了应对多重共线问题（X为奇异矩阵）的解决方法（Hoerl和Kennard，1970），即提供一个二阶惩罚函数来获得精炼模型，

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（Tibshirani，1996），

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可以看做是改进的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为零，从而达到降低维度的目的。

回归树是将数据的特征空间划分为若干决策区间（叶子），使得每一片叶子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。从某种意义上讲，回归树是一种一般化的固定效应模型，其固定效应大小受到协变量X的影响。为了弥补单个回归树功能简单且容易出现过度拟合的问题，依托决策树进行的模型组合方法越来越受到关注，如RF和GBDT的方法。

随机森林是多个回归树组成的决策体系。RF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。RF方法也需要选择正则化的程度，即对算法进行调整以保证样本外预测的有效性。与RF类似，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。

本文将综合运用这些机器学习的方法进行共享单车短期需求预测，以测试集样本的均方根误差（标准误）进行估计模型的预测效果比较。在应用机器学习模型时，本文将样本划分为训练集样本（60%的观测值）和测试集样本（40%的观测值）。在对训练集样本并进行交叉验证时，Athey和Imbens（2017）建议通常划分k（例如k=10）组子样本，留下第m组，并将其余的子样本组进行拟合估计，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是交叉验证模型残差平方和最小的那个。本文对Lasso和Ridge设定k=5，对RF和GBDT设定k=10。在确立调整模型之后，最终模型效果的评估依靠测试集样本的R2和RMSE来衡量，R2越大和RMSE小的模型，在预测共享单车短期需求方面具有更大优势。

### （二）数据说明

本文的研究对象是同一区域的短期共享单车需求量，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（SF Bay 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SBBS有站模型管理，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天的无限制通行费为22美元和9美元[[3]](#footnote-3)。2016年8-9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，数据来源于项目公开数据[[4]](#footnote-4)。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2014年8月31日二年时间的运营数据，共669,959个观测值。具体的变量名称和描述统计见表1。

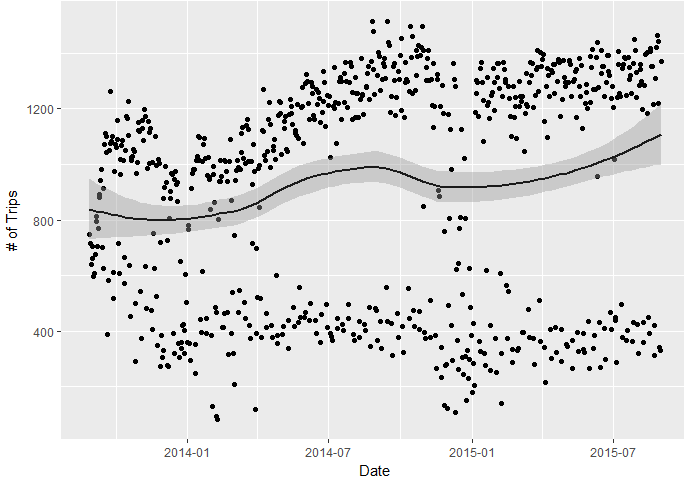
**表1 主要变量及其描述统计**

|  |  |  |  |
| --- | --- | --- | --- |
| **变量名称** | **观测数** | **平均数** | **含义** |
| *ID* | 669,959 | 460,382 | 记录编号，每条观测一个编号 |
| *Date* | 669,959 | - | 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、法定假日等其他日期特征 |
| *Station\_id* | 669,959 | 35 | 站点编号 |
| *Bike\_id* | 669,959 | 427.95 | 自行车编号 |
| *Station* | 669,959 | - | 具体站点，又分为两个变量，租车站点和还车站点 |
| *Subscription\_type* | 669,959 | - | 客户类型，分为注册用户和临时用户 |
| *Duration* | 669,959 | 1107.95 | 车辆使用的时间长度 |
| *Temper* | 669,959 | - | 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* | 669,959 | - | 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* | 669,959 | - | 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* | 669,959 | - | 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* | 669,959 | - | 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* | 669,959 | - | 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* | 669,959 | 2.78 | 当天云层覆盖率，取值范围为0-8 |
| *Precipitation* | 669,959 | 0.02 | 当天降水，取值范围为0-3.36 |
| *Wind\_dir* | 669,959 | 266.61 | 当天风向，取值范围为0-360 |
| *Event* | 669,959 | - | 天气事件，有五类取值，分别为雾、鱼、雾雨、雷雨和无事件 |

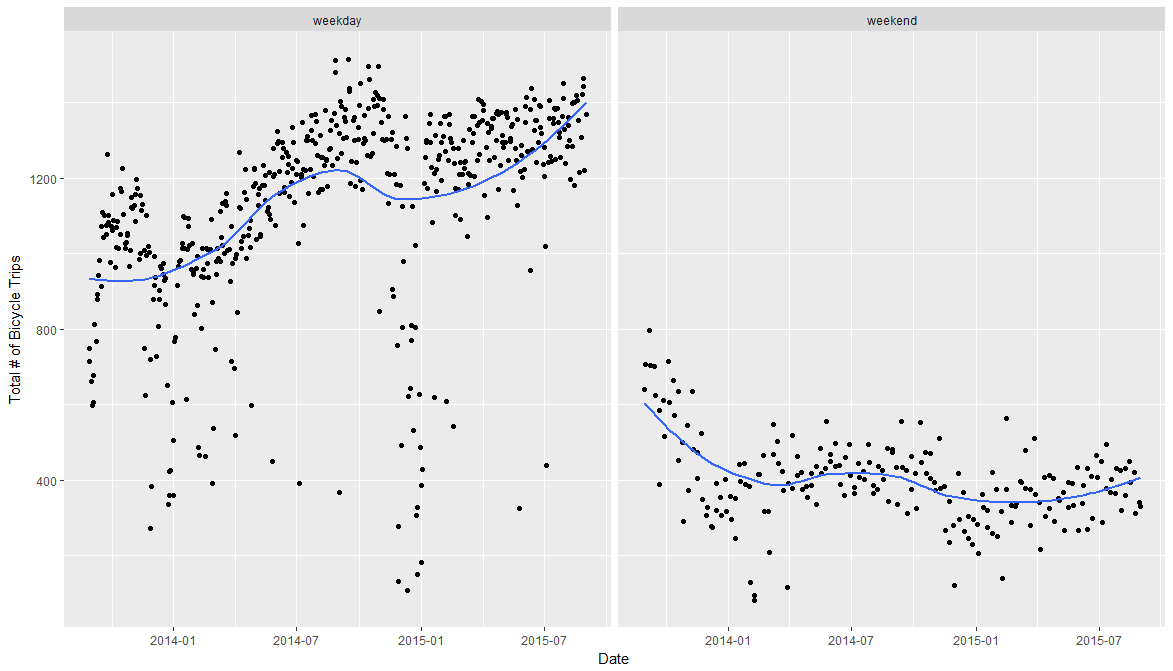
**注：其他与本文相关度不高的变量没有进行描述。**

### （三）数据分析

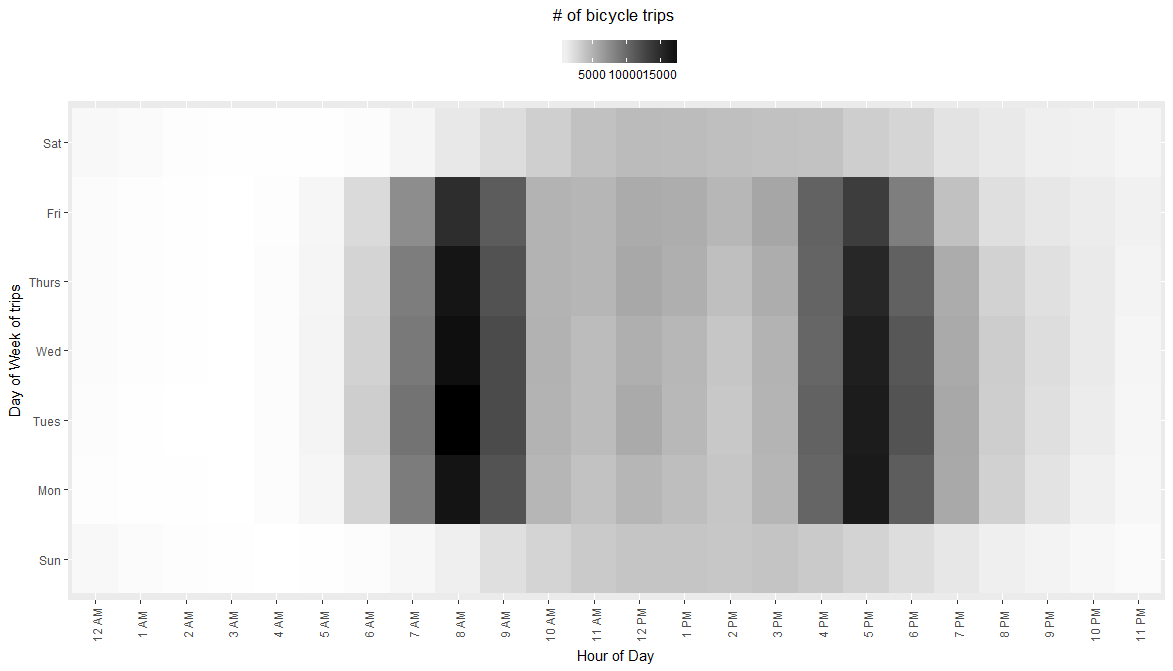
首先以实际数据初步考察共享单车需求的主要影响因素。图1是以日期和需求量为维度制作的散点图，并以局部加权回归画出了回归线，可以看到不同日期条件下，需求量总体上具有明显的二分化差异，引致这种差异的时间因素可能是工作日与周末的日期属性因素。图2进一步以工作日和周末划分样本，并绘制散点图即局部加权回归线，样本进一步体现了工作日与周末两个日期特征对需求的重要影响。



**图1 日期与共享单车需求量的散点图**



**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**



**图3 时间段与共享单车需求量的关系**

进一步考察不同时间段条件下共享单车需求的情况，如图3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，远远超过了其他时间段的需求，表明很多人使用单车是用于上下班通勤，即在上下班高峰期，单车需求最高。时间段也是影响单车需求的重要因素。

对于天气因素，我们预期很多天气变量与需求量存在一个舒适值区间，过高或过低的极端温度、湿度、风力等会对需求有负向作用。由于篇幅关系，对其他变量不再绘图验证。

## 四、预测结果与模型评估

表2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（具体模型调整参数选择过程可见附录），每个模型均采用R2和RMSE测量模型对训练集合测试集的预测效果。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R2为0.8035，RMSE为25.9729，将拟合模型应用到测试集样本，预测结果R2为0.8056，RMSE为25.9921。测试集与训练集估计效果的差异体现了模型样本外预测的能力，具体来看OLS模型样本内外预测的拟合优度基本相当，但样本外数据的预测误差明显加大。

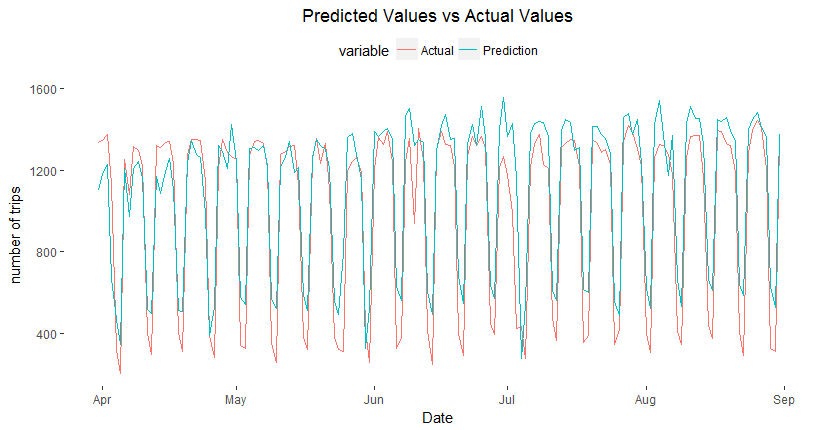
对比机器学习模型的估计结果，Lasso和Radge对训练集和测试集样本的预测效率都没有提升，具体来看，R2有所下升，拟合结果变差，RMSE有所提升，误差有所增加。因此总体上讲，这两种机器学习的方法对共享单车需求预测的效果并不好，原因在于Lasso与Radge两个模型在设计上主要是解决模型协变量共线和变量维度过多的问题，在对共享单车需求预测中，由于模型选取依据了经济学的基本理论，并非完全的数据驱动，因此其自变量之间的共线性和数量过多的问题并不突出，因此没有突出模型优势。

对比RF和GBDT两种决策树模型组合的机器学习模型，可以发现这两个模型在预测共享单车需求方面具有较好的预测效果，从训练集的样本内估计来看，两个模型的R2比普通的OLS回归模型提升均在10%以上，RMSE比OLS模型的误差降幅在20个单位以上。从样本外预测效果来看，虽然相对自身来讲，其样本外数据的拟合优度没有明显的差别，样本外预测误差上还有所增大，但由于其测试集样本预测误差在绝对值上大幅低于OLS回归（降幅同样在20个单位以上），因此可以说，两种机器学习模型在样本外预测能力方面比OLS回归具有较大的优势，体现出了机器学习算法在估计大数据需求预测模型上的明显优势。

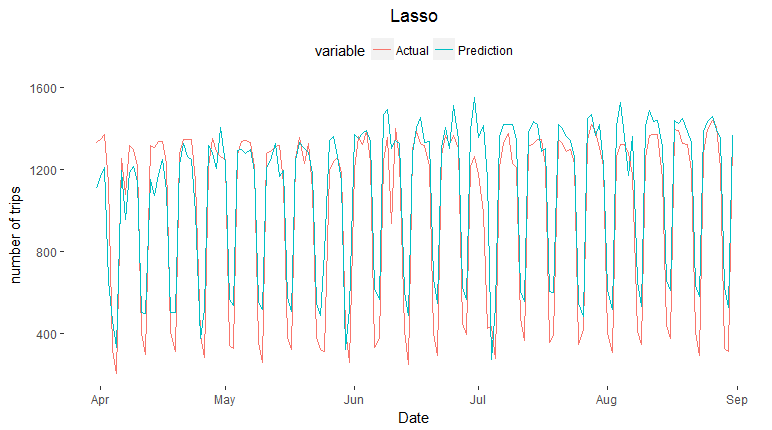
**表2 主要变量及其描述统计**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **模型** | **训练集样本** | | **测试集样本** | | **测试集-训练集** | |
| **R2** | **RMSE** | **R2** | **RMSE** | **R2** | **RMSE** |
| *OLS* | 0.8035 | 25.9729 | 0.8056 | 25.9921 | 0.0021 | 0.0192 |
| *Lasso* | 0.8015 | 25.9871 | 0.8036 | 26.0064 | 0.0021 | 0.0193 |
| *与OLS比较* | -0.0002 | 0.0142 | -0.0002 | 0.0143 | -- | -- |
| *Radge* | 0.7319 | 26.3650 | 0.7337 | 26.3943 | 0.0018 | 0.0293 |
| *与OLS比较* | -0.0716 | 0.3921 | -0.0719 | 0.4022 | -- | -- |
| *Random Forest* |  |  |  |  |  |  |
| *与OLS比较* |  |  |  |  | -- | -- |
| *Gradient Boost* | 0.9255 | 2.7160 | 0.9248 | 2.7822 | -0.0007 | 0.0662 |
| *与OLS比较* | 0.1220 | -23.2569 | 0.1192 | -23.2099 | -- | -- |

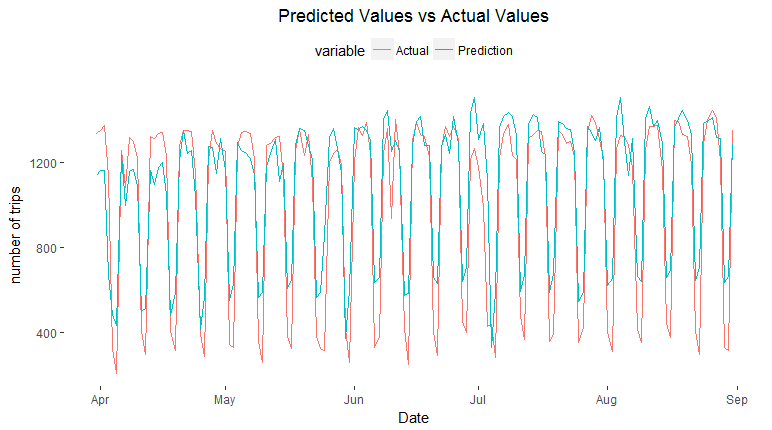
为了对不同模型样本外预测误差进行直观的比较，图4至图8显示了五个模型对测试集样本的预测结果，图中没有被蓝色样本覆盖的红色样本为预测值对真实值的预测误差（样本外预测误差），可以直观看出，RF和GBDT模型的样本外预测结果要明显优于其他模型的预测结果。



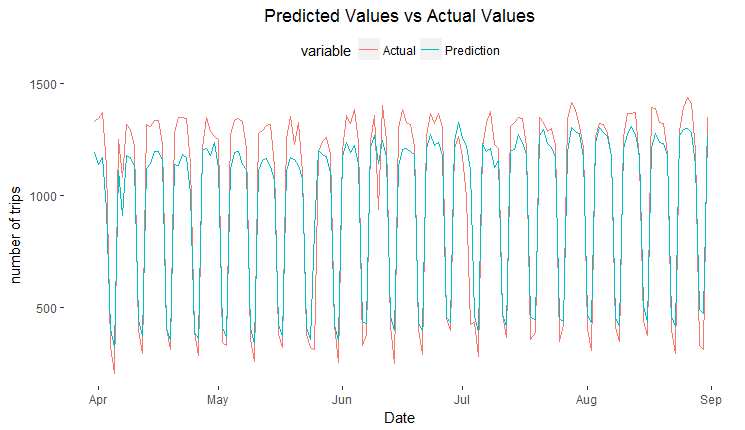
**图4 OLS模型测试集样本的预测值与真实值比较**



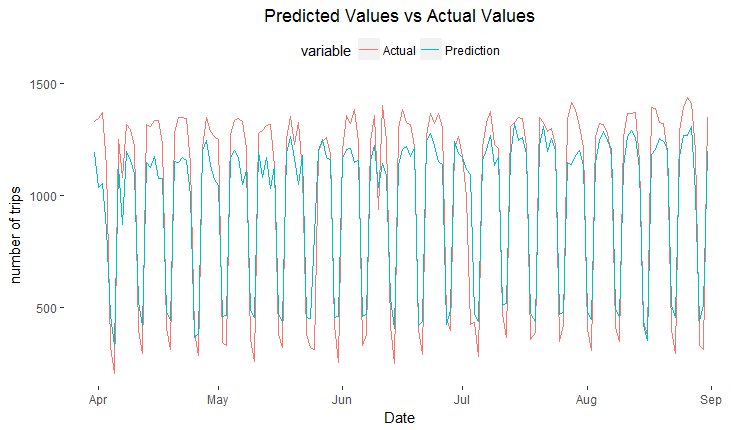
**图5 Lasso模型测试集样本的预测值与真实值比较**



**图6 Ridge模型测试集样本的预测值与真实值比较**



**图7 Random Forest模型测试集样本的预测值与真实值比较**



**图8 GBDT模型测试集样本的预测值与真实值比较**

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。模型预测结果来看，从对共享单车需求量影响最大的前五名变量来看（表3），OLS，Lasso和Ridge指向了相同的五个时间段变量，包括上午8点（hour8.AM）和下午5点（hour5.PM）两个上下班通勤高峰期。RF和GBDT则将工作日特征（wday）和“是否周末”（is\_weekend）作为仅次于上下班高峰时间段的重要的模型预测影响指标。

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（表4），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仍然都是时间虚拟变量，不过不同模型得出的最重要因素排序已经发生变化。RF和GBDT模型得出的重要性指标也存在差异，RF将最高气温变量（max\_temperature\_f）排名在第9位，认为高温特征对共享单车需求产生重要影响，GBDT模型结论则没有将高温因素纳入前十名的影响之中。

**表3 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 | 154.1214 | t统计量 | 295.169 |
|  | hour5.PM | 135.5266 | 259.504 |
|  | hour9.AM | 99.57691 | 189.173 |
|  | hour6.PM | 88.52317 | 167.684 |
|  | hour4.PM | 84.39095 | 160.164 |
| *Lasso* | hour8.AM | 96.9135 | 非零系数 | 96.9135 |
|  | hour5.PM | 78.3290 | 78.3290 |
|  | hour9.AM | 42.3482 | 42.3482 |
|  | hour6.PM | 31.3126 | 31.3126 |
|  | hour4.PM | 27.1566 | 27.1566 |
| *Radge* | hour8.AM | 73.44749 | 非零系数 | 73.44749 |
|  | hour5.PM | 55.75364 | 55.75364 |
|  | hour9.AM | 21.46749 | 21.46749 |
|  | hour6.PM | 10.99845 | 10.99845 |
|  | hour4.PM | 7.091371 | 7.091371 |
| *Random Forest* | hour8.AM | -- | IncNodePurity |  |
|  | hour5.PM | -- |  |
|  | wday | -- |  |
|  | is\_weekend | -- |  |
|  | hour9.AM | -- |  |
| *Gradient Boost* | hour8.AM | -- | 信息增益 | 0.2470 |
|  | hour5.PM | -- | 0.1816 |
|  | wday | -- | 0.0690 |
|  | is\_weekend | -- | 0.0672 |
|  | hour9.AM | -- | 0.0395 |

**表4 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  6-10 | hour8.AM, hour5.PM, hour9.AM, hour6.PM, hour4.PM,  hour7.AM, hour12.PM, hour3.PM, hour1.PM, hour11.AM |
| *Lasso* | 1-5  6-10 | hour8.AM, hour5.PM, hour9.AM, hour6.PM, hour4.PM  hour7.AM, month5, month4, month6, month8 |
| *Radge* | 1-5  6-10 | hour8.AM, hour5.PM, hour9.AM, hour6.PM, hour4.PM  month5, month6, month4, month8, month7 |
| *Random Forest* | 1-5  6-10 | hour8.AM, hour5.PM, wday, is\_weekend, hour9.AM  hour6.PM, hour4.PM, hour7.AM, max\_temperature\_f, hour8.PM |
| *Gradient Boost* | 1-5  6-10 | hour8.AM, hour5.PM, wday, is\_weekend, hour9.AM  hour4.PM, hour6.PM, hour10.AM, hour8.PM, hour1.PM |

## 五、研究结论

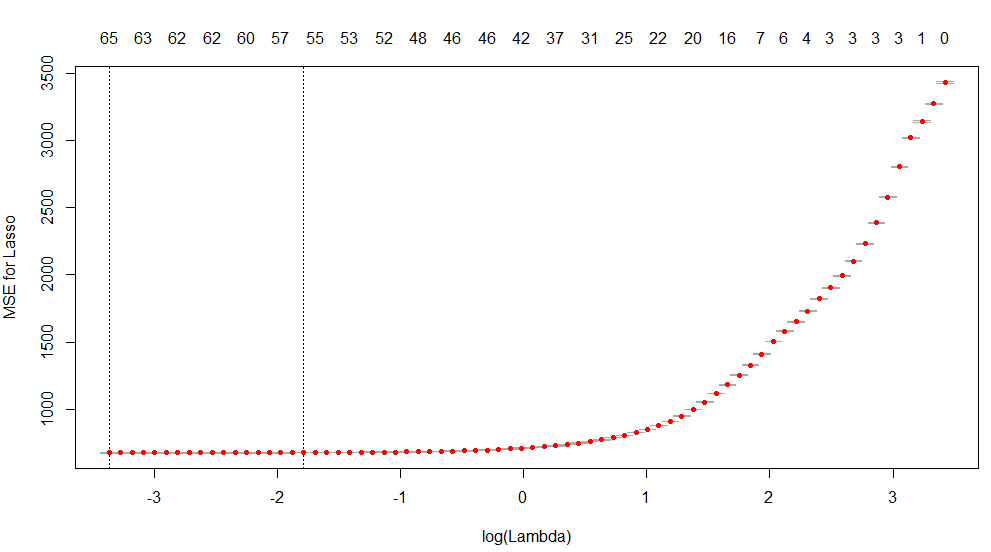
对于有站模式和无站模式的需求预测，影响因素可能相同，也可能存在其他重要的影响因素（如运营商对按照时间对供需不平衡区域进行调度的能力），需要

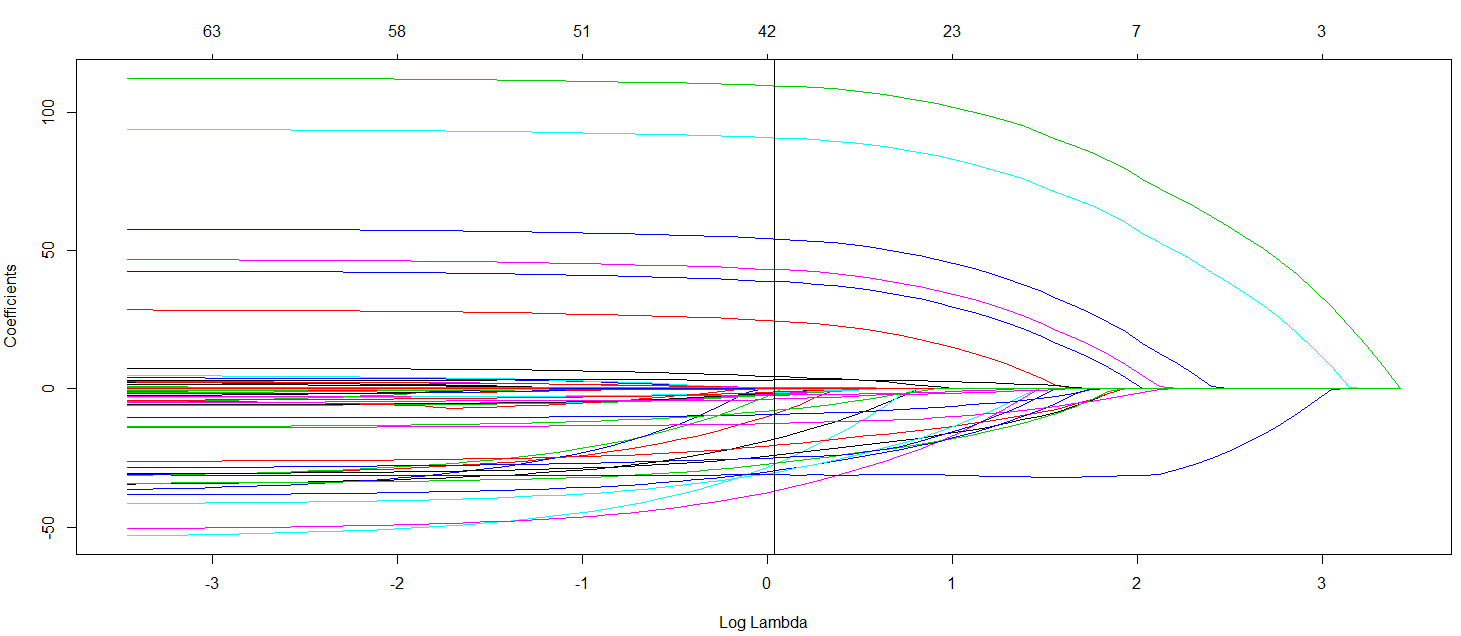
## 参考文献：

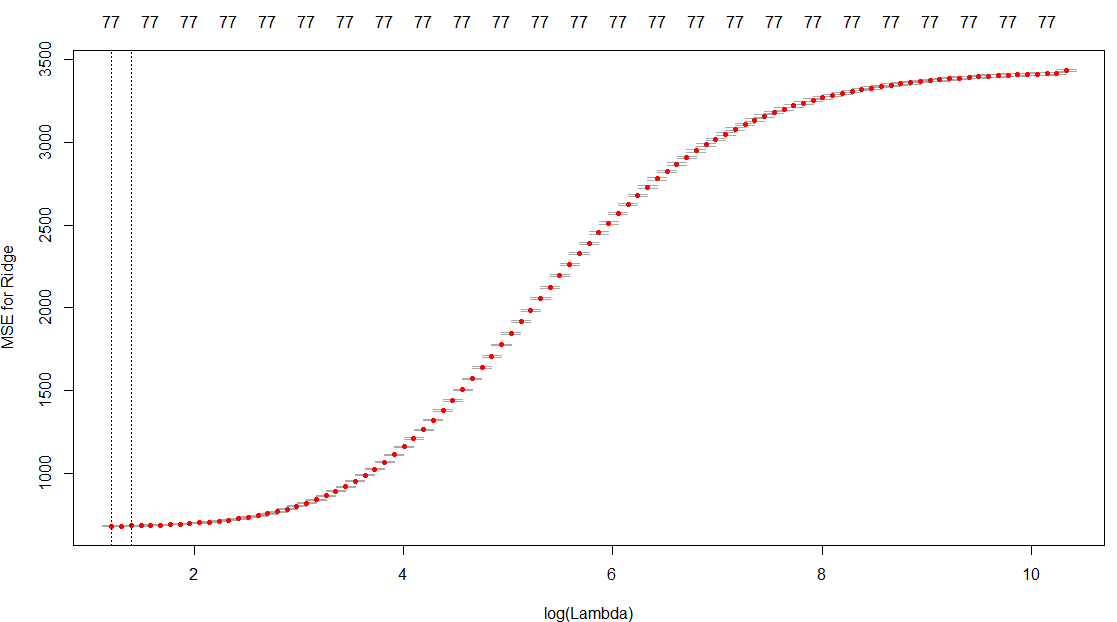
1. Liran Einav, Jonathan Levin. Economics in the age of big data [J]. Science, 2014, 346(6210): 1243089. DOI:10.1126/Science. 1243089
2. P. Bajari, D. Nekipelov, SP. Ryan, M. Yang. Machine Learning Methods for Demand Estimation [J]. American Economic Review, 2015, 105 (5): 481-485.
3. A. Belloni, V. Chernozhukov, C. Hansen. High-Dimensional Methods and Inference on Structural and Treatment Effects [J]. Journal of Economic Perspectives, 2014, 28(2):29-50.
4. S. Athey, Machine Learning and Causal Inference for Policy Evaluation [A]. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [C], 2015, ACM. pp.5-6. DOI: http://dx.doi.org/10.1145/2783258.2785466
5. S. Athey, G. Imbens. The State of Applied Econometrics: Causality and Policy Evaluation [J]. Journal of Economic Perspectives, 2017, 31(2):3-32.
6. Brian K. Lee, Justin Lessler, Elizabeth A. Stuart. Improving Propensity Score Weighting Using Machine Learning [J]. Statistics in Medicine, 2010, 29(3):337–346.
7. A. Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, Sendhil Mullainathan. Productivity and Selection of Human Capital with Machine Learning [J]. American Economic Review, 2016, 106(5):124-127.
8. S. Athey, G. Imbens, S. Wager. Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing [EB/OL]. 2016, arXiv: 1604.07125.
9. S. Athey, G. Imbens. Recursive Partitioning for Estimating Heterogeneous Causal Effects [J]. PNAS, 2016, 113(27): 7353–7360.
10. S. Wager, S. Athey. Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests [EB/OL]. 2015, arXiv:1510.04342
11. S. Athey, J. Tibshirani, S. Wager. Solving Heterogeneous Estimating Equations with Gradient Forests [EB/OL].2016, arXiv: 1610.01271.
12. S. Deb. Analytical Ideas to Improve Daily Demand Forecasts: A Case Study [A]. In NT. Nguyen, S. Tojo, LM Nguyen, B. Trawinski. Intelligent Information and Database Systems (Aciids 2017), Part II [C]. 2017, ACIIDS. pp. 23-32. DOI: 10.1007/978-3-319-54430-4\_3.
13. DeMaio, P. 2004. Will smart bikes succeed as public transportation in the United States? Journal of Public Transportation 7(2): 1-15.
14. DeMaio, P. 2009. Bike-sharing: History, impacts, models or provision and future.Journal of Public Transportation, 12 (4), 41–56.
15. Meddin, R., DeMaio, P. (2016). The Bike Sharing World Map [EB/OL]. <http://www.metrobike.net/the-bike-sharing-world-map/>
16. Mor Kaspi,Tal Raviv n, Michal Tzur. Detection of Unusable Bicycles in Bike-Sharing Systems [J]. The International Journal of Management Science, 2016,65:10-16
17. Hoerl A E, Kennard R W. Ridge regression: biased estimation for non- orthogonal problem [J]. Technometrics (S0040-1706), 1970, 12(1): 55-88.
18. Tibshirani RJ. Regression shrinkage and selection via the lasso [J]. Journal of the Royal Statistical Society, 1996,58:267-288．
19. A.Belloni, V. Chernozhukov, I. Fernández-Val, C. Hansen. Program Evaluation and Causal Inference with High-Dimensional Data [J]. 2013, arXiv: 1311.2645.

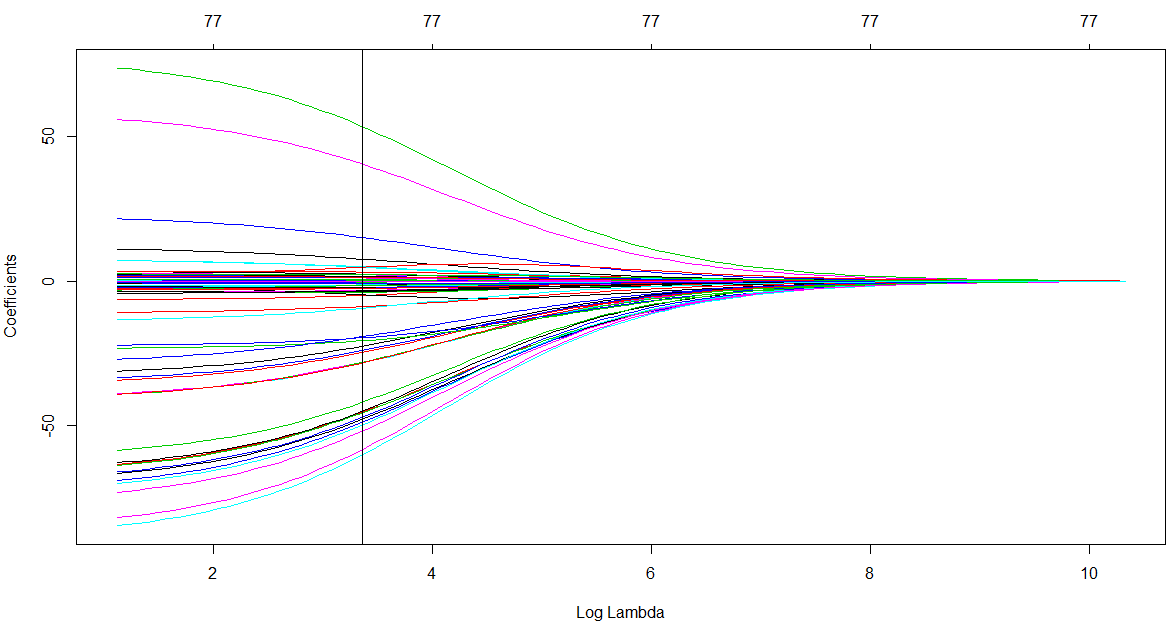
## 附录：机器学习模型的正则化参数调整

为了防止过度识别问题，机器学习模型需要进行模型的参数调整。待完成。









1. 基金项目：本研究受到“中国特色社会主义经济建设协同创新中心”资助，同时受到南开大学“人文社会科学青年教师研究启动项目”支持。

   作者简介：焦志伦（1979-）男，博士，讲师，研究方向为商业经济、国际物流、区域经济。  
   Email：zjiao@nankai.edu.cn [↑](#footnote-ref-1)
2. 共享经济概念由Felson 和 Spaeth首次提出，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”。 [↑](#footnote-ref-2)
3. 更详细的信息可见，http://kalw.org/post/sf-bay-area-bike-share-launches-thursday#stream/0 [↑](#footnote-ref-3)
4. 读者可以在<https://www.kaggle.com/benhamner/sf-bay-area-bike-share>上获取相关数据。 [↑](#footnote-ref-4)