厦门大数据安全开放创新应用大赛·交通专题

——厦门五一假期交通路网拥堵特征分析

参赛编号：JT-WYJT-0138

赛道类型：算法赛

赛题名称（三选一）：厦门五一假期交通路网拥堵特征分析

团队人数：4

# 摘要

五一假期厦门市交通供需紧张状况严重，交通拥堵治理工作面临严峻的考验和挑战，存在着规划不合理、城市车流激增、监管乏力、停车难、停车占道等方面的难题，我们团队将尝试利用大数据分析和机器学习等方法进行数学建模。主要工作包括：一、厦门市路网拥堵指数构建；二、重点旅游景点（厦大-演武路-大学路）的拥堵特征分析与对策。

# 问题需求

伴随着厦门城市的飞速发展，城市人口规模不断扩增，城市居民机动化出行也进入高速发展阶段。根据厦门交警发布的《2019年厦门市道路交通运行大数据》显示，2019年厦门市机动车保有量超过166万辆，车辆日均活跃度为95.6万辆，同比提升3.7%。机动车数量的增加极大地便利了居民以及游客的出行，但相应的城市交通拥堵现象不容忽视。尽管《2019年厦门市道路交通运行大数据》中提示2019年厦门的交通拥堵指数为1.29，同比2018年下降了1.5%，但是根据百度地图发布的《2019年度中国城市交通报告》，在汽车保有量100万~200万级城市中，厦门的通勤高峰拥堵指数达到了1.666，在拥堵排名中位列第六。尽管节假日期间的全市平均交通拥堵指数（1.20）比工作日（1.32）略低，但拥堵情况分布不均衡，城市主干道景区周边道路等主要通道拥堵尤为严重。2019年厦门十大拥堵路段中，除了全市交通主干道或连接进出岛“四桥一隧”的主要道路以及受占道施工围挡影响的路段外，还包括了几条岛内各热门景区的周边道路。如虎园路的钟鼓山隧道-同安路方向拥堵指数达到了1.68，镇海路的同安路-鹭江道方向拥堵指数为1.63，龙虎山路的文屏路-换到南路方向拥堵指数为1.42，均位列厦门前十大拥堵路段。岛内景区周边路网资源有限、游客集中、人车流量大，是造成拥堵的主要原因，为公众出行带来极大不便，影响了厦门在游客心中留下的形象，也为交通和旅游等相关部门带来不小的管理压力。

厦门作为海上丝绸之路核心区的一座现代化国际性港口风景旅游城市，2019年接待游客已超1亿人次，光“五一”小长假四天就接待国内外游客210.68万人次。基于游客在节假日期间集中的出行需求，节假日景区周边道路拥堵的现象尤为突出。而游客搭乘出租车、网约车既是加重道路拥堵的重要原因之一，同时也是拥堵的主要影响对象。根据《2019年厦门市交通运输发展报告》数据显示，去年全市出租车累计营运17458.97万趟次。其中，巡游车累计营运8052.19万趟次，日均营运趟次22.06万趟次，网约车累计营运9406.78万趟次，日均营运趟次25.77万趟次，较往年均有小幅增长。出租车热门起终点主要为高崎机场、火车站、厦门北站、莲坂新村、明发商业广场、轮渡中山路等交通枢纽、旅游景区以及商圈。

## 前沿研究简述

交通拥堵和管理并不是新话题。1868年英国伦敦出现第一盏交通灯；大约一个世纪之前机动车开始量产。大量出现的机动车革新了陆路交通，也带来了前所未见的交通大拥堵。实际上，早在古罗马时代的庞贝城，交通管理就已经出现了。

近年来，许多不同领域的研究者都投入到解决交通拥堵问题的研究之中，其中包括了数学、物理学和工程方面的专家。他们从不同的角度入手，建立了许多种数学模型，但至今尚未找到公认的建模过程中的首要法则。

很多研究者在研究车流时，会套用其他类型的流动，比如液体流、气体流和颗粒流。于是许多对车流的研究采用的是流体力学的方法。研究者也确实发现，在车流量很大但并不拥堵的道路上，一个很小的事件就会触发堵塞，比如某个司机突然做了一个转向动作。这个状况与流体力学中过冷液体的突然冻结非常相似。

德国汉堡大学数学学院的因格温·加瑟(Ingenuin Gasser)教授将目前的研究模型分为三类，分别是微观、运动和宏观模型。微观模型描述的是单个驾驶员的动力学；运动模型是用气体动力学的方法来考察概率分布；而宏观模型讨论的则是车流密度和速度等宏观物理量。

科学家建立这些模型的目的当然是为了理解复杂的车流现象并最终对其做到影响或控制。但在奥罗兹看来，“尽管这些类推可能会帮助科学家获得对车辆系统的理解，但也越来越明显地可以看到，车流与牛顿宇宙中的任何流动都不一样。”最明显的不同之处就在于，车流是会遇到各种指示牌、红绿灯和交叉路口的。

## 厦门市交通拥堵状况分析

目前，厦门交通供需紧张状况严重，交通拥堵治理工作面临严峻的考验和挑战，存在着规划不合理、城市车流激增、监管乏力、停车难、停车占道等方面的难题，因此必须科学管理和规划，构建大交通网络，夯实基础设施，创新理念，改善人文交通环境。

### 厦门城市交通面临的难点

1、经济社会快速发展拉动交通需求持续快速增长，出行总量大幅攀升

目前厦门市总人口已达4百余万。形成的庞大的出行总量，给厦门市交通带来极大压力。

同国内许多城市一样，随着城市人口的快速增长，厦门城区面积迅速扩张，增大了人们交通出行的距离和时间，增加了交通成本。在“摊大饼”的城市发展形态下，在便利性与可达性较高的环路周边进行高密度土地开发，不仅创造出巨大的交通需求，而且大量日常出行要通过市区主干道，从而造成主干道的交通拥堵及潮汐式通勤交通流，进而呈现拥堵范围由中心城区向周边扩散的态势,相应引发一系列城市交通问题。

与城市功能相联系，厦门的政治、文化、教育、医疗、商业、旅游等城市核心功能都集中于中心城区内，中心城区承载资源过多，形成单中心的城市空间格局，制约了城市综合承载能力的提升。特别值得注意的是，由于厦门市小学、中学甚至幼儿园的教育资源分布极不平衡，“教育移民”、“跨区择校”已成为相当普遍的现象，引发大量交通流。随着城市持续不断扩展，市中心城市功能的过度聚集和土地的超强度开发，造成人口与就业岗位高度集中，“职住分离”现象突出，并由此导致交通出行的高度集中，带来极大交通压力。

2、机动车增长过猛、使用过频，与停车泊位等交通资源产生尖锐矛盾

伴随着经济特区建设快速增长和人民生活水平的显著提升，自20世纪80年代末期开始厦门机动化进程加速，其发展速度在我国沿海城市中也属罕见。2019年全市机动车拥有量就突破166万辆，现数量仍然高速地增长着。不仅汽车拥有量急速增加，而且使用频率也很高，过多过频地占用了有限的道路资源，交通需求过度膨胀与城市道路承载能力产生尖锐矛盾，使得近年厦门的交通拥堵现象日趋严重。

随着机动车的迅猛增长，停车泊位缺口不断增大，尽管政府尽可能开建设立停车场（楼），几乎所有的市区次干道路边也都设置了停车位，但仍无法满足需求，非法停车、占道停车成为相当普遍的现象。

3、路网先天不足，完善结构难度较大，且道路建设速度远不及机动车增长速度

厦门市路网通行能力近年不断得到改善，但作为面积不大的岛城，路网条件具有特殊性，存在先天性不足和结构性缺陷。历史上旧城区可供汽车行驶的道路本就很少，之后形成的主干道、次干道、支路不匹配的格局，降低了行车路径的选择性，形成非稳态道路网结构。一些山、海及旧城风貌保护构成空间阻隔，损害了路网系统的整体性，加重了路网密度稀疏的先天性不足。次干路、支路严重短缺，路网稀，节点通行能力低，“微循环”系统薄弱，干路系统空间布局不均衡，道路通达性差等因素，严重制约了路网整体效能的发挥，将交通压力都挤向干道或次干道，致使市区道路面积率偏低，交通拥堵严重。前些年建了BRT快速公交专用道，也只是在短短几年一定程度上缓解拥堵而已。

### 交通出行结构呈现不合理性

在私人小汽车逐渐普及、机动化进程日益加速的形势下，近年来厦门市交通出行结构发生了重大变化。居民出行结构迅速由过去以步行、自行车为主转向以机动车出行为主，公共交通受到小汽车交通的挑战，自行车出行比例持续大幅下降，出租车出行比例保持平稳，导致交通出行结构呈现不合理性，道路交通压力过大，缓解交通拥堵的任务十分艰巨。

近年厦门市积极实施公交优先战略，公共交通出行比例持续增长，但总体看增幅仍不够大，与小汽车出行尚处于“平分秋色”格局。而国外发达国家城市核心区出行中，公共交通分担率（以轨道交通为主、地面公交为辅）一般占60％－80％，小汽车交通（包括私家车、公务车和出租车），仅占总出行量的12%－32%。

## 我们团队研究主要内容

随着GPS技术在城市出租车运营管理中应用推广，交通运输行业获取到了及时准确的浮动车数据，为道路交通运行状态的评价和缓解交通拥堵治理提供了数据支撑。

我们拟将采用五一假期期间多种交通运输车辆运行监测数据，完成厦门市交通运行指数模型，并重点分析厦大片区节假日的拥堵时空特征，通过大数据分析和机器学习建模为解决城市拥堵治理提供辅助决策。

# 数据应用

## 数据清单

1、五一假期厦门巡游车订单数据

车牌号、上车时间、上车经度、上车纬度、下车时间、下车经度、下车纬度、计程公里、空驶公里、等待计时时间

2、五一假期厦门巡游车辆GPS数据

车牌号、运营状态、GPS速度、行驶方向、卫星定位时间、经度、纬度、海拔高度

3、五一假期厦门网约车订单数据

订单编号、上车位置行政区划编号、车辆号牌、预计上车时间、等待时间、车辆出发经度、车辆出发纬度、上车地点、上车时间、车辆到达经度、车辆到达纬度、下车地点、下车时间、载客里程、载客时间、空驶里程、订单完成时间

4、五一假期厦门网约车辆GPS数据

车牌号、定位时间、经度、纬度、坐标系、方向角、海拔高度、GPS速度、运营状态（1：载客;2：接单;3：空驶; 4：停运）、订单号

5、其它数据（需要特别采集，字段信息待定）：

如：路网数据、天气、交通事故、旅客进岛、旅游景点、商家活动信息等。

## 数据分析细节

### 数据清洗

数据清洗是将重复、多余的数据筛选清除，将缺失的数据补充完整，将错误的数据纠正或者删除，最后整理成为我们可以进一步加工、使用的数据。

所谓的数据清洗，也就是ETL处理，包含抽取Extract、转换Transform、加载load这三大法宝。在大数据挖掘过程中，面对的至少是G级别的数据量，包括用户基本数据、行为数据、交易数据、资金流数据以及第三方的数据等等。选择正确的方式来清洗特征数据极为重要，除了让你能够事半功倍，还至少能够保证在方案上是可行的。

数据清洗的一般步骤：分析数据、缺失值处理、异常值处理、去重处理、噪音数据处理。在大数据生态圈，有很多来源的数据ETL工具，但是对于公司内部来说，稳定性、安全性和成本都是必须考虑的。

对于数据值缺失的处理，通常使用的方法有下面几种：

1、删除缺失值

当样本数很多的时候，并且出现缺失值的样本在整个的样本的比例相对较小，这种情况下，我们可以使用最简单有效的方法处理缺失值的情况。那就是将出现有缺失值的样本直接丢弃。这是一种很常用的策略。

2、均值填补法

根据缺失值的属性相关系数最大的那个属性把数据分成几个组，然后分别计算每个组的均值，把这些均值放入到缺失的数值里面就可以了。

3、热卡填补法

对于一个包含缺失值的变量，热卡填充法的做法是：在数据库中找到一个与它最相似的对象，然后用这个相似对象的值来进行填充。不同的问题可能会选用不同的标准来对相似进行判定。最常见的是使用相关系数矩阵来确定哪个变量（如变量Y）与缺失值所在变量（如变量X）最相关。然后把所有变量按Y的取值大小进行排序。那么变量X的缺失值就可以用排在缺失值前的那个个案的数据来代替了。

还有类似于最近距离决定填补法、回归填补法、多重填补方法、K-最近邻法、有序最近邻法、基于贝叶斯的方法等。

异常值通常被称为“离群点”，对于异常值的处理，通常使用的方法有下面几种：

1、简单的统计分析

拿到数据后可以对数据进行一个简单的描述性统计分析，譬如最大最小值可以用来判断这个变量的取值是否超过了合理的范围，如客户的年龄为-20岁或200岁，显然是不合常理的，为异常值。

2、3∂原则

如果数据服从正态分布，在3∂原则下，异常值为一组测定值中与平均值的偏差超过3倍标准差的值。如果数据服从正态分布，距离平均值3∂之外的值出现的概率为P(|x-u| > 3∂) <= 0.003，属于极个别的小概率事件。如果数据不服从正态分布，也可以用远离平均值的多少倍标准差来描述。

3、箱型图分析

箱型图提供了识别异常值的一个标准：如果一个值小于QL01.5IQR或大于OU-1.5IQR的值，则被称为异常值。QL为下四分位数，表示全部观察值中有四分之一的数据取值比它小；QU为上四分位数，表示全部观察值中有四分之一的数据取值比它大；IQR为四分位数间距，是上四分位数QU与下四分位数QL的差值，包含了全部观察值的一半。箱型图判断异常值的方法以四分位数和四分位距为基础，四分位数具有鲁棒性：25%的数据可以变得任意远并且不会干扰四分位数，所以异常值不能对这个标准施加影响。因此箱型图识别异常值比较客观，在识别异常值时有一定的优越性。

4、基于模型检测

首先建立一个数据模型，异常是那些同模型不能完美拟合的对象；如果模型是簇的集合，则异常是不显著属于任何簇的对象；在使用回归模型时，异常是相对远离预测值的对象

优缺点：1.有坚实的统计学理论基础，当存在充分的数据和所用的检验类型的知识时，这些检验可能非常有效；2.对于多元数据，可用的选择少一些，并且对于高维数据，这些检测可能性很差。

5、基于距离

通常可以在对象之间定义邻近性度量，异常对象是那些远离其他对象的对象

优缺点：1.简单；2.缺点：基于邻近度的方法需要O(m2)时间，大数据集不适用；3.该方法对参数的选择也是敏感的；4.不能处理具有不同密度区域的数据集，因为它使用全局阈值，不能考虑这种密度的变化。

6、基于密度

当一个点的局部密度显著低于它的大部分近邻时才将其分类为离群点。适合非均匀分布的数据。

优缺点：1.给出了对象是离群点的定量度量，并且即使数据具有不同的区域也能够很好的处理；2.与基于距离的方法一样，这些方法必然具有O(m2)的时间复杂度。对于低维数据使用特定的数据结构可以达到O(mlogm)；3.参数选择困难。虽然算法通过观察不同的k值，取得最大离群点得分来处理该问题，但是，仍然需要选择这些值的上下界。

7、基于聚类：

基于聚类的离群点：一个对象是基于聚类的离群点，如果该对象不强属于任何簇。离群点对初始聚类的影响：如果通过聚类检测离群点，则由于离群点影响聚类，存在一个问题：结构是否有效。为了处理该问题，可以使用如下方法：对象聚类，删除离群点，对象再次聚类（这个不能保证产生最优结果）。

优缺点：1.基于线性和接近线性复杂度（k均值）的聚类技术来发现离群点可能是高度有效的；2.簇的定义通常是离群点的补，因此可能同时发现簇和离群点；3.产生的离群点集和它们的得分可能非常依赖所用的簇的个数和数据中离群点的存在性；4.聚类算法产生的簇的质量对该算法产生的离群点的质量影响非常大。

噪音，是被测量变量的随机误差或方差。对于噪音的处理，通常有下面的两种方法：

1、分箱法

分箱方法通过考察数据的“近邻”（即，周围的值）来光滑有序数据值。这些有序的值被分布到一些“桶”或箱中。由于分箱方法考察近邻的值，因此它进行局部光滑。

用箱均值光滑：箱中每一个值被箱中的平均值替换。

用箱中位数平滑：箱中的每一个值被箱中的中位数替换。

用箱边界平滑：箱中的最大和最小值同样被视为边界。箱中的每一个值被最近的边界值替换。

一般而言，宽度越大，光滑效果越明显。箱也可以是等宽的，其中每个箱值的区间范围是个常量。分箱也可以作为一种离散化技术使用.

2、回归法

可以用一个函数拟合数据来光滑数据。线性回归涉及找出拟合两个属性（或变量）的“最佳”直线，使得一个属性能够预测另一个。多线性回归是线性回归的扩展，它涉及多于两个属性，并且数据拟合到一个多维面。使用回归，找出适合数据的数学方程式，能够帮助消除噪声。

### 数据挖掘

数据挖掘（Data Mining，简称DM），是指从大量的数据中，挖掘出未知的且有价值的信息和知识的过程。

与数据挖掘类似的有一个术语叫做”机器学习“，这两个术语在本质上的区别不大，如果在书店分别购买两本讲数据挖掘和机器学习的书籍，书中大部分内容都是互相重复的。具体来说，小的区别如下：

机器学习这个词应该更侧重于技术方面和各种算法，一般提到机器学习就会想到语音识别，图像视频识别，机器翻译，无人驾驶等等各种其他的模式识别，甚至于谷歌大脑等AI，这些东西的一个共同点就是极其复杂的算法，所以说机器学习的核心就是各种精妙的算法。

数据挖掘则更偏向于“数据”而非算法，而且包括了很多数据的前期处理，用爬虫爬取数据，然后做数据的清洗，数据的整合，数据有效性检测，数据可视化（画图）等等，最后才是用一些统计的或者机器学习的算法来抽取某些有用的“知识”。前期数据处理的工作比较多。

数据挖掘是一门交叉学科，覆盖了统计学、计算机程序设计、数学与算法、数据库、机器学习、市场营销、数据可视化等领域的理论和实践成果

从数据挖掘的角度看，都可以转换为五类问题：分类，聚类，回归，关联和推荐。

1、分类问题

简单来说，就是根据已经分好类的一推数据，分析每一类的潜在特征建立分类模型。对于新数据，可以输出新出具属于每一类的概率。

比如主流邮箱都具备的垃圾邮件识别功能：一开始，正常邮件和垃圾邮件都是混合在一起的，如果我们手工去点击哪些是垃圾邮件，逐渐的，垃圾邮件就会自动被识别放到垃圾文件夹。如果我们对于混在正常邮件中的垃圾持续进行判断，系统的识别率就会越来越高。我们人工点击判断，相当于预先分类（两类：垃圾邮件和非垃圾邮件），系统就会自己学习两类邮件的特征建立模式，对于新邮件，会根据模式判断属于每个类别的可能性。

2、聚类问题

和分类算法是不同概念，但是工作中业务人员经常误用。 聚类的的目的也是把数据分类，但类别并不是预先定义的，算法根据“物以类聚”的原则，判断各条数据之间的相似性，相似的就归为一类。

比如我有十万消费者的信息数据，比如包括性别，年龄，收入，消费等，通过聚类的方法事可以把这些数据分成不同的群，理论上每群用户内都是相似性较高的，就可以覆盖分群用户制定不同的策略

3、回归问题

回归问题和分类问题有点类似，但是回归问题中的因变量是一个数值，而分类问题，最终输出的因变量是一个类别

简单理解，就是定义一个因变量，在定义若干自变量，找到一个数学公式，描述自变量和因变量之间的关系

比如，我们要研究房价，然后收集房子距离市中心的距离，面积，收集足够多的房子的数据，就可以建立一个房价和距离，面积的公式，这样给出一个新的距离和面积数据，就可以预测这个房子的价格

4、关联分析

也就是很有名气的啤酒与尿布的故事，据说这是一个真实的案例：沃尔玛在分析销售记录时，发现啤酒和尿布经常一起被购买，于是他们调整了货架，把两者放在一起，结果真的提升了啤酒的销量。后来还分析背后的原因，说是因为爸爸在给宝宝买尿布的时候，会顺便给自己买点啤酒……这已经成为了数据挖掘领域的一个经典桥段。

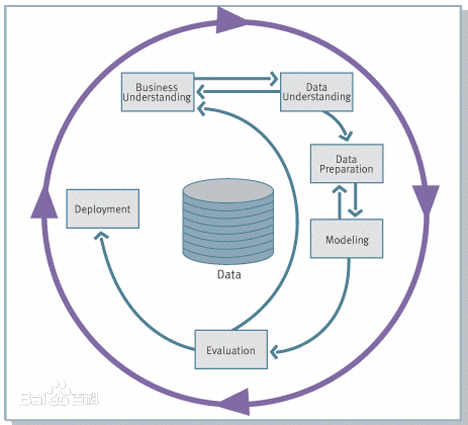
因此，关联分析就是基于数据识别产品之间潜在的关联，识别有可能频繁发生的模式

5、推荐系统

利用电子商务网站向客户提供商品信息和建议，帮助用户决定应该购买什么产品，模拟销售人员帮助客户完成购买过程也。也就是平时我们在浏览电商网站、视频网站、新闻App中"猜你喜欢"、“其他人也购买了XXX”等类似的功能。

6、数据挖掘的工作流程

数据挖掘的通用流程叫做CRISP-DM（Cross Industry Standard Process-Data Mining）数据挖掘方法论



CRISP-DM 模型为一个KDD工程提供了一个完整的过程描述。该模型将一个KDD工程分为6个不同的，但顺序并非完全不变的阶段。

商业理解（business understanding）

在这第一个阶段我们必须从商业的角度了解项目的要求和最终目的是什么，并将这些目的与数据挖掘的定义以及结果结合起来。

主要工作包括：确定商业目标，发现影响结果的重要因素，从商业角度描绘客户的首要目标，评估形势，查找所有的资源、局限、设想以及在确定数据分析目标和项目方案时考虑到的各种其他的因素，包括风险和意外、相关术语、成本和收益等等，接下来确定数据挖掘的目标，制定项目计划。

数据理解（data understanding）

数据理解阶段开始于数据的收集工作。接下来就是熟悉数据的工作，具体如：检测数据的量，对数据有初步的理解，探测数据中比较有趣的数据子集，进而形成对潜在信息的假设。收集原始数据，对数据进行装载，描绘数据，并且探索数据特征，进行简单的特征统计，检验数据的质量，包括数据的完整性和正确性，缺失值的填补等。

数据准备（data preparation）

数据准备阶段涵盖了从原始粗糙数据中构建最终数据集（将作为建模工具的分析对象）的全部工作。数据准备工作有可能被实施多次，而且其实施顺序并不是预先规定好的。这一阶段的任务主要包括：制表，记录，数据变量的选择和转换，以及为适应建模工具而进行的数据清理等等。

根据与挖掘目标的相关性，数据质量以及技术限制，选择作为分析使用的数据，并进一步对数据进行清理转换，构造衍生变量，整合数据，并根据工具的要求，格式化数据。

建模（modeling）

在这一阶段，各种各样的建模方法将被加以选择和使用，通过建造，评估模型将其参数将被校准为最为理想的值。比较典型的是，对于同一个数据挖掘的问题类型，可以有多种方法选择使用。如果有多重技术要使用，那么在这一任务中，对于每一个要使用的技术要分别对待。一些建模方法对数据的形式有具体的要求，因此，在这一阶段，重新回到数据准备阶段执行某些任务有时是非常必要的。

评估（evaluation）

从数据分析的角度考虑，在这一阶段中，已经建立了一个或多个高质量的模型。但在进行最终的模型部署之前，更加彻底的评估模型，回顾在构建模型过程中所执行的每一个步骤，是非常重要的，这样可以确保这些模型是否达到了企业的目标。一个关键的评价指标就是看，是否仍然有一些重要的企业问题还没有被充分地加以注意和考虑。在这一阶段结束之时，有关数据挖掘结果的使用应达成一致的决定。

部署（deployment）

部署，即将其发现的结果以及过程组织成为可读文本形式。模型的创建并不是项目的最终目的。尽管建模是为了增加更多有关于数据的信息，但这些信息仍然需要以一种客户能够使用的方式被组织和呈现。这经常涉及到一个组织在处理某些决策过程中，如在决定有关网页的实时人员或者营销数据库的重复得分时，拥有一个“活”的模型。

根据需求的不同，部署阶段可以是仅仅像写一份报告那样简单，也可以像在企业中进行可重复的数据挖掘程序那样复杂。在许多案例中，往往是客户而不是数据分析师来执行部署阶段。然而，尽管数据分析师不需要处理部署阶段的工作，对于客户而言，预先了解需要执行的活动从而正确的使用已构建的模型是非常重要的。

### 数据可视化

数据可视化，它是创造性设计美学和严谨的工程科学的卓越产物。它的美丽令人向往，而它的繁杂又使其蒙上层神秘的面纱。

对数据可视化的定义，似乎存在有“两派”说法。对于研究大规模数据人员而言，数据可视化指综合运用计算机图形学、图像、人机交互等技术，将采集或模拟的数据映射为可识别的图形、图像、视频或动画，并允许用户对数据进行交互分析的理论、方法和技术。而对于广大的编辑、设计师、数据分析师等需要呈现简单数据序列的人员而言，数据可视化是将数据用统计图表和信息图方式呈现，同样也符合“3+2”（文字、图表、图像+声音、动画）的基本构成元素。两种定义其实是从广义和狭义两个不同层面去理解，它们既不是对立的，也没有严格区分，仅是针对于不同的业务场景。

无论是哪种职业和应用场景，数据可视化都有一个共同的目的，即明确、有效的传递信息。图形能将不可见现象转化为可见的图形符号，并直截了当和清晰直观的表达出来。而人类右脑记忆图像的速度比左脑记忆抽象的文字快100万倍。因此，数据可视化能够加深人对于数据的理解和记忆。

广义的数据可视化包括数据的采集、分析、治理、管理、挖掘在内的一系列复杂数据处理，然后由设计师设计一种表现形式，最后由工程师创建对应的可视化算法及技术实现手段。

在信息可视化的视觉表达中，动态将相互分离的各种信息传播形式有机地融合在一起，进行有节奏的信息处理、传输和实现。通过造型和色彩的运动，满足受众的视觉感受，同时将信息内容更加深刻地传达给受众，使整个信息传达的过程更加轻松。

对于数据可视化有诸多工具，如：iCharts、Flot、Raphal等功能都十分强大。

## 数据安全保障

数据安全首先要成立专门的数据安全治理机构，以明确数据安全治理的政策、落实和监督由谁长期负责。该机构通常是虚拟机构，可称为数据安全治理委员会或数据安全治理小组，成员由数据的利益相关者和专家构成。其成立，标志着组织的数据安全治理工作正式启动，使组织内数据安全规范制定、数据安全技术导入、数据安全体系建设得以不断完善。该机构成立后，履行以下职责：

A．数据的分级分类原则的制定

B．数据安全使用（管理）规范的制定

C．数据安全治理技术的导入

D．数据安全使用规范的监督执行

E．数据安全治理的持续演进

### 数据安全政策

数据安全治理，最为重要的是实现数据安全策略和流程的制订，在企业或行业内经常被作为《某某数据安全管理规范》进行发布，所有的工作流程和技术支撑都是围绕此规范来制订、落实。

数据安全治理同样需要遵循国家级的安全政策和行业内的安全政策。举例如下：

1. 网络安全法；

2. 等级保护政策；

3. BMB17；

4. 行业相关的政策要求举例：

(a) PCI-DSS、Sarbanes-Oxley Act（SOX 法案）、HIPPA；

(b) 企业内部控制基本规范；（三会、财政、审计）

(c) 中央企业商业秘密保护暂行规定；

### 数据的分级分类

数据治理主要依据数据的来源、内容和用途进行分类；以数据的价值、内容敏感程度、影响和分发范围进行敏感级别划分。

1、数据使用部门和角色梳理

数据资产梳理中，明确数据如何被存储、数据被哪些对象使用、数据被如何使用。对于数据的存储和系统的使用，需要通过自动化的工具进行；对于部门、人员角色梳理，更多在管理规范文件中体现；对于数据资产使用角色的梳理，关键要明确不同受众的分工、权利和职责。

2、数据的存储与分布梳理

清楚敏感数据分布，才能知道需要对什么样的库实现何种管控策略；对该库运维人员实现怎样的管控措施；对该库的数据导出实现怎样的模糊化策略；对该库数据的存储实现何种加密要求。

3、数据的使用状况梳理

明确数据被什么业务系统访问，才能准确地制订业务系统工作人员对敏感数据访问的权限策略和管控措施。

### 数据的访问控制

针对数据使用不同方面，完成对数据使用的原则和控制策略，包括：数据访问的账号和权限管理、数据使用过程管理、数据共享（提取）管理、数据存储管理。

### 定期的稽核策略

定期稽核，保证数据安全治理规范落地，包括：

A、合规性检查；

B、操作监管与稽核；

C、风险分析与发现。

# 技术架构

没有哪个机器学习模型可以包打天下，如何找到当前问题的最优解是一个永恒的问题。幸运的是，结合/融合/整合 (integration/ combination/ fusion)多个机器学习模型往往可以提高整体的预测能力。这是一种非常有效的提升手段，在多分类器系统(multi-classifier system)和集成学习(ensemble learning)中，融合都是最重要的一个步骤。

一般来说，模型融合或多或少都能提高的最终的预测能力，且一般不会比最优子模型差。举个实用的例子，stacking方法就是模型融合，通过结合多个各有所长的子学习器，我们实现了更好的预测结果。基本的理论假设是：不同的子模型在不同的数据上有不同的表达能力，我们可以结合他们擅长的部分，得到一个在各个方面都很“准确”的模型。当然，最基本的假设是子模型的误差是互相独立的，这个一般是不现实的。但即使子模型间的误差有相关性，适当的结合方法依然可以各取其长，从而达到提升效果。

我们今天介绍几种简单、有效的模型结合方法。

1、案例分析

让我们给出一个简单的分析。假设我们有天气数据X和对应的标签y，现在希望实现一个可以预测明天天气的模型Ψ。但我们并不知道用什么算法效果最好，于是尝试了十种算法，包括：

* 算法1: 逻辑回归 – C1
* 算法2：支持向量机（SVM）– C2
* ...
* 算法10：随机森林– C10

结果发现他们表现都一般，在验证集上的误分率比较高。我们现在期待找到一种方法，可以有效提高最终预测结果。

2、平均法/投票法

一种比较直白的方法就是对让10个算法模型同时对需要预测的数据进行预测，并对结果取平均数/众数。假设10个分类器对于测试数据Xt的预测结果是[C1( Xt), C2( Xt),…, C10( Xt)]=[0,1,1,1,1,1,1,0,1,1,0]，那很显然少数服从多数，我们应该选择1作为Xt的预测结果。如果取平均值的话也可以那么会得到0.7，高于阈值0.5，因此是等价的。

但这个时候需要有几个注意的地方：

首先，不同分类器的输出结果取值范围不同，不一定是[0,1]，而可以是无限定范围的值。举例，逻辑回归的输出范围是0-1（概率），而k-近邻的输出结果可以是大于0的任意实数，其他算法的输出范围可能是负数。因此整合多个分类器时，需要注意不同分类器的输出范围，并统一这个取值范围。

* 比如可以先转化为如二分类结果，把输出的范围统一后再进行整合。但这种方法的问题在于我们丢失了很多信息，0.5和0.99都会被转化为1，但明显其可靠程度差别很大。
* 也可以转化为排序（ranking），再对不同的ranking进行求平均。
* 更加稳妥的方法是对每个分类器的输出结果做标准化，也就是调整到正态分布上去。之后就可以对多个调整后的结果进行整合。同理，用归一化也可以有类似的效果。

其次，就是整合稳定性的问题。采用平均法的另一个风险在于可能被极值所影响。正态分布的取值是[-∞，+∞]，在少数情况下平均值会受到少数极值的影响。一个常见的解决方法是，用中位数（median）来代替平均数进行整合。

同时，模型整合面临的另一个问题是子模型良莠不齐。如果10个模型中有1个表现非常差，那么会拖累最终的效果，适得其反。因此，简单、粗暴的把所有子模型通过平均法整合起来效果往往一般。

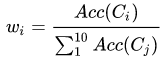
3、寻找优秀的子模型

不难看出，一个较差的子模型会拖累整体的集成表现，那么这就涉及到另一个问题？什么样的子模型是优秀的。

一般来说，我们希望子模型：准而不同（accurate but diversified）。好的子模型应该首先是准确的，这样才会有所帮助。其次不同子模型间应该有差别，比如独立的误差，这样作为一个整体才能起到互补作用。

因此，如果想实现良好的结合效果，就必须对子模型进行筛选，去粗取精。在这里我们需要做出一点解释，我们今天说的融合方法和bagging还有boosting中的思路不大相同。bagging和boosting中的子模型都是很简单的且基数庞大，而我们今天的模型融合是结合少量但较为复杂的模型。

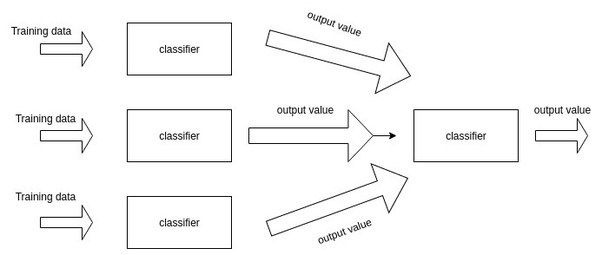
4、筛选方法：赋予不同子模型不同的权重

因此我们不能再简单的取平均了，而应该给优秀的子模型更大的权重。在这种前提下，一个比较直白的方法就是根据子模型的准确率给出一个参考权重ω，子模型越准确那么它的权重就更大，对于最终预测的影响就更强：。简单取平均是这个方法的一个特例，即假设子模型准确率一致。

5、更进一步：学习分类器权重

在4中提到的方法在一定程度上可以缓解问题，但效果有限。那么另一个思路是，我们是否可以学习每个分类器的权重呢？

答案是肯定，这也就是Stacking的核心思路。通过增加一层来学习子模型的权重。



简单来说，就是加一层逻辑回归或者SVM，把子模型的输出结果当做训练数据，来自动赋予不同子模型不同的权重。

一般来看，这种方法只要使用得当，效果应该比简单取平均值、或者根据准确度计算权重的效果会更好。

6、再进一步：挖掘局部关系

上面提到的方法，都有一个不可避免的问题，那就是对于问题处理是全局的(global)。一个分类器不一定在每个局部上表现都好，而我们赋予的全局权重会无差别的认为：”一个子模型在全局的表现上一致“，这是不现实的。

因此，另一个非常有效的融合方法就是：动态分类器选择（Dynamic Classifier Selection），简称DCS。DCS的思路是，当我们遇到一个新的数据需要去预测时，我们首先找到训练数据中和新数据临近的k个数据，一般这个搜寻可以通过k-近邻来实现。找到以后，我们只需要找到在k个相邻的训练数据构成的局部空间上，选择表现最好的分类器。它可能是逻辑回归，也可能是SVM，但思路是找到最好的那一个即可。

这种做法的最大优点是考虑到了不同分类器在不同局部的表现能力可能有差异，不该默认分类器的全局表现一样。在这个基础上，研究人员还提出 动态集成选择（Dynamic Ensemble Selection），DES的区别在于目标是找到局部上表现较好的几个子模型来共同预测，可以理解为集成上的集成。

# 预期成果

五一假期，厦门市交通供需紧张状况严重，交通拥堵治理工作面临严峻的考验和挑战，存在着规划不合理、城市车流激增、监管乏力、停车难、停车占道等方面的难题，我们团队将尝试利用大数据分析和机器学习等方法进行数学建模。预期成果包括：一、厦门市路网拥堵指数构建；二、重点旅游景点（厦大-演武路-大学路）的拥堵特征分析与对策。

**1、厦门市路网拥堵指数构建**

交通指数是[交通拥堵指数](https://baike.baidu.com/item/%E4%BA%A4%E9%80%9A%E6%8B%A5%E5%A0%B5%E6%8C%87%E6%95%B0)或交通运行指数（Traffic Performance Index，即“TPI”）的简称，是综合反映道路网畅通或拥堵的概念性指数值。相当于把拥堵情况数字化。

交通指数取值范围为0～10，分为五级。其中0～2、2～4、4～6、6～8、8～10分别对应“畅通”、“基本畅通”、“轻度拥堵”、“中度拥堵”、“严重拥堵”五个级别，数值越高表明交通拥堵状况越严重。   
　　具体数据规定如下：

交通指数 级别对应路况 出行时间

0 - 2：畅通，交通运行状况良好，基本没有道路拥堵，可以按道路限速标准[行驶](https://baike.baidu.com/item/%E8%A1%8C%E9%A9%B6)

2 - 4：基本畅通，有交通运行状况较好，少量道路[拥堵](https://baike.baidu.com/item/%E6%8B%A5%E5%A0%B5)，比畅通时多耗时0.2至0.5倍

4 - 6： 轻度拥堵，交通运行状况较差，部分环路、主干路拥堵，比畅通时多耗时0.5至0.8倍

6 - 8： 中度拥堵，交通运行状况差，大量环路、主干路拥堵，比畅通时多耗时0.8至1.1倍

8 - 10： 严重拥堵，交通运行状况很差，全市大部分道路拥堵，比畅通时多耗时1.1倍以上

**2、重点旅游景点（厦大-演武路-大学路）的拥堵特征分析与对策**

厦门作为海上丝绸之路核心区的一座现代化国际性港口风景旅游城市，2019年接待游客已超1亿人次，光“五一”小长假四天就接待国内外游客210.68万人次。基于游客在节假日期间集中的出行需求，节假日景区周边道路拥堵的现象尤为突出。而游客搭乘出租车、网约车既是加重道路拥堵的重要原因之一，同时也是拥堵的主要影响对象。

厦大-演武路-大学路路段在五一假期拥堵尤其明显，这里是仅次于鼓浪屿的最热门的景点。我们选取这个路段进行分析，主要考虑有以下几点：

1、此路段有厦大、演武观景平台、猫街、白城海滩、胡里山等旅游景点，是游客最集中的地方之一，旅游资源丰富，也是市区最拥堵的地方。

2、此路段连接了鹭江道-演武大桥、中山路-思明南路、植物园-钟鼓山隧道、成功大道、环岛干道、环岛南路，是厦门市路网比较复杂的区域。

3、此路段是团队成员主要生活工作区域，各种假期备受交通拥堵困扰，出行很不方便。

在主办方大数据的加持下，我们希望利用团队的专业知识和技术创新能力，对此路段的路网规划、车流、交通监管、交通标示、公共交通和游客等信息进行分析，期望能对交通拥堵的特征进行量化挖掘，提出我们的决策参考方案。

# 作品价值

我们认为作品价值主要表现在三方面：

1、提升交通管理效率，有助于交通管理提高效率，降低管理成本。

2、强化交通安全保障，有助于交通安全的保障，降低交通事故。

3、优化公众出行体验，有助于游客旅游体验，提升厦门市的旅游满意度。