城市路口交通拥堵预测

2019大数据分析大作业

董颖 曹立 翟文祺 宋怡然

2019

1. **选题背景**

我们都有过被困在红绿灯前，却只被有几秒钟的短暂时间通过十字路口的经历。如果城市规划者可以提前预测交通热点，就可以有效减少通勤者的等待压力和经济损失。本次大作业的灵感正是来自于对于交通堵塞情况数据进行分析，希望能够进行预测，减少城市道路交通压力。

通过十五周的大数据课程学习，我们对Python编程，数据处理，机器学习中回归、分类、聚类、集成学习方法以及深度学习有了初步的了解，结合小组的自身能力情况，我们讨论后决定选取《城市路口交通拥堵预测》作为期末大作业，希望通过这个作业能够应用老师讲的理论知识，同时通过作业来检验学习成果以及发现不足。

1. **作业结构**

《城市路口交通拥堵预测》作业主要分成四个部分。 第一部分介绍选题背景和我们数据分析的目标与假设。第二部分为我们的数据分析，包括三大块：第一块是针对原数据，我们利用可视化进行探索性数据分析，为之后的模型搭建提供思路，可视化包括地图、热力图等。第二块是进行主成分分析和聚类分析。第二块是通过构造变量和模型，并且针对参数进行优化，最终输出了我们的预测结果，得到了一个相对较好的数据。第三部分为我们对这个大作业的总结以及思考。

1. **课堂理论知识应用**
2. python编程
3. 可视化分析
4. PCA主成分分析
5. kmeans聚类分析
6. LightGBM模型
7. **探索性数据分析——可视化和降维与聚类**

我们进行了探索性数据分析，包括可视化和降维与聚类，得到许多有趣的结论，并应用于之后的数据建模中。

【方法】

因为训练集的变量数较多有28个，我们希望从可视化中得到一些有趣的结论，来消除无关变量在拟合和预测时数据的影响。我们小组使用matplotlib和seaborn和mplleaflet绘制了折线图，计数折线图，按城市分类的柱状图，动态地图，相关因子热力图探究了如下的10种变量关系和一些结论。之后我们使用主成分分析和kmeans聚类进行降维和分类，进一步观察数据之间的特征。

【结论】

1. 我们定义了df函数，对数据进行总结，了解其特征。

整体上看，train.csv 中共有28个特征，857409列数据，test.csv中共有13个特征，1920335列数据。我们发现：在训练和测试集中，数据缺失的列相同：EntryStreetName和ExitStreetName。在计算丢失的数据仅占1%比例后，我们发现大多数列似乎没有任何丢失的数据。而且这两列只丢失了很小一部分的数据，因此在模型拟合之前可直接删除这些观测值。

2. 通过观察数据集中城市的柱状图数据分布，我们发现：

Philadelphia费城数据最多，为45.29%， 剩下三个城市Boston(21.23%), Atlanta(17.89%), Chicago(15.59%)，数据量差别不大，对我们的预测不会有较大影响。

3. 通过对于进入路口街道，驶出路口街道和通行量的三张分布图进行分析，我们发现：

1） 最频繁出现的街道是North Broad Street

2） 25名之后的街道数据量基本持平

3） 绝大多数车辆都保持直行

4. 通过对数据整体的时间分布分析（以小时和天为计），我们发现：

在小时图中：我们发现每个城市都有不同的时间模式。具体来看，

1. 费城在所有时段中都拥有最高的数据量。仅在凌晨5点，通行数据小于波士顿。
2. 所有城市的高峰时间为08:00至17:00，但费城为08:00至24:00。
3. 亚特兰大白天的通行量没有明显的波峰波谷，但从下午17点至凌晨4点，这是第二个汽车通行数据较多的城市，分布较为平均。

在月份图中：

1) 数据大约只有6个月（一月和五月的值很少），可能对上半年预测不够精准的风险。

2) 波士顿市的通行数据随月份增大而增大，其他城市的通行数据基本保持不变。

5. 通过对所有城市的月份和在路口等待时间的关系分析观察总体趋势，我们发现：

1) 7月份的数据中看到最小堵车时间，之后平均等待时间逐渐变长，在10月左右达到最大值，然后11月和12月逐渐减小。

我们的猜测是：随着夏季假期的临近，大部分人会在家休息或者度假，这可以解释7月份的最短时间。同意年底的假期的到来会使等待时间第二次出现下降。

分别观察每个城市，我们发现：

1) 总体趋势相同，但特定情况下，最大和最小的位置有所不同。

2) 我们可以观察到季节/月次数对平均停车时间的间接影响。而气候对交通的直接影响似乎无法单独从给定的数据中分离出来。

根据以上发现，我们产生了新的问题：驶入驶出数据的相似性的原因和方向变化和等待时间的关系。

7. 观察统计转角频率与转弯角度和停止时间关系，我们发现：

1) 每个转角的频率：0度转数计数为所有转数的0.70，所有其他转角的频率（45、90、135和180）的总和约为0.30。

2) 转过的角度的大小与等待时间之间存在线性关系，转角越大，等待时间越长。其中180°掉头时间最长。可以创建成新的变量添加进数据集中。

8. 通过mplleaflet包实现动态地图可视化，我们发现：

1) 白天/夜晚和天气会影响流量。我们可以添加城市的温度和降雨量来进一步研究。

2) 我们的数据主要集中在市区：因为乡村的编号较少且稀疏。

3) 交叉路口周围是商圈可能会在周中和周末影响交通。

4) 每个城市的集中度和地形也有关，如果再给出一些数据值可以进一步探究。

9. 等待时间和时间分布的关系,我们发现一些普遍的堵车规律：

1) 早晚高峰分别出现在：早8:00--9:00和晚上15:00-17:00

2) 晚高峰都比早高峰更堵一点20%

10. 对于我们想预测的堵车相关6个指标，从heatmap中，我们发现：

1) 第一次停车的距离和20%总时长有强相关关系。

11. 从kmeans聚类分城市图中，我们发现：

1) 我们发现城市之间最大的区别在第三簇（编号为2）。

2) 而亚特兰大和费城在这点上拥有比较相同的模式，并且其他簇形状都相似。

1. **数据建模**

下面我们对于数据建模过程中的难点和创新点进行总结。

【变量构建】

根据数据处理部分得到的灵感，我们大胆尝试，不拘泥于数据集中已给出的特征，而是从公开渠道找了一些关于气候的补充数据，加入我们的模型中，提高预测精度。

首先我们增添了一个0-1变量，专门用来表示车辆是否直行，直行为1，反之为0。并且由于不同时段的车流量的巨大区别，我们对堵车时段也进行了划分。并且我们加入了气候特征。季节/月份、城市温度、降雨量可能会对堵车时间产生间接影响，因此我们收集了这四个城市不同月份的温度、降雨量、降雪量数据，也作为模型的变量。对于不用的特征，我们大胆舍弃。以上均来自于我们在可视化分析中观察到的结论。

【算法选择】

我们首先使用了线性回归模型，通过与其他进阶模型均方差的比较，作为评价模型好坏的参照。进而我们选用LightGBM模型。它是个快速的，分布式的，高性能的基于决策树算法的梯度提升框架。可用于排序，分类，回归以及很多其他的机器学习任务中。与常见的机器学习算法对比，速度非常快。

我们总结LightGBM有以下优势：

　　1）更快的训练效率

　　2）低内存使用

　　3）更高的准确率

　　4）支持并行化学习

　　5）可以处理大规模数据

【调参优化】

在创建完模型后，我们必须进行参数调优，才能找到最好的匹配。即使模型还可以，如果它的参数设置不匹配，同样无法输出好的结果。常用的调参方式有Grid search 和 Random search。 Grid search 是全空间扫描，所以比较慢，Random search 虽然快，但可能错失空间上的一些重要的点，精度不够。而Hyperopt是一种通过贝叶斯优化来调整参数的工具，该方法有较快的速度，并有较好的效果。通过使用Hyperopt，我们得到了最优超参数（也就是找到了最好的那棵树），并把这些超参数运用到真实的目标函数上。

贝叶斯优化的原理是这样的：

1. 建立一个目标函数的代理概率模型（当实际问题——高精度模型的计算量很大、不容易求解时，使用计算量较小、求解迅速的简化模型来替代原模型，加速优化过程。）
2. 找到使这个模型最优的超参数
3. 把这些超参数运用到真实的目标函数上
4. 根据运行结果更新代理模型
5. 重复2-4步，直到到达设定的最大遍历数

【解决周期数量需要手动设置的问题】

我们的超参数之一是迭代周期的数量：也就是应当完整遍历数据集多少次（树的深度），一次为一个周期。如果周期数量太少，网络有可能发生欠拟合（即对于训练数据的学习不够充分）；如果周期数量太多，则有可能发生过拟合（即网络对迭代数据中的“噪声”而非信号拟合）。我们分析继续训练会导致测试集上的准确率下降的原因，可能是1）过拟合；2）学习率过大导致不收敛。

所以我们选择使用早停法（early stopping），解决周期数量需要手动设置的问题。它也可以被视为一种能够避免网络发生过拟合的正则化方法。在模型中我们设置早停周期数为250，为迭代周期的一半。如果超过超过这个数字效果还没有提升，则早停。

通过使用LightGBM模型，与线性回归模型相比，预测误差有了大幅度减小，说明了LightGBM模型对这个问题的适用性。

【思考与展望】

建模部分的亮点在于：

1) 很好地利用了EDA探索性数据分析中的结论，发现了一些给定数据集以外的信息。如在第一部分中得出堵车情况和气候有关的结论，在建模时加入了气候相关特征，使得模型更全面。又比如在KMeans聚类时得到原数据可分为4类的结论，运用到了K折交叉验证中K的选择。给我们的启发是在解决实际问题时要放开思维，搜集不同维度的数据，可能会对建模效果产生意想不到的作用。

2) 在建模时用了一些归一化方法，将非连续性数值/文本变量转化成更易操作的连续型变量，强迫症患者的福音。

3) 用贝叶斯优化进行调参，实现超参数的自动调优，达到了效率和准确率的平衡。

建模可优化部分：

1) 通过mplleaflet可视化地图发现每个城市的集中度和地形也有关，但因为地形数据不太方便收集，故本次建模并没有加入地形数据。如果再搜集一些地形方面的数据值可进一步探究。

2) 由于数据量大(几百万条)，跑模型时间长，所以只运用了LightGBM一种模型与baseline线性回归，如果时间更充裕，可以补充其他模型进行比较。

6. 驶入驶出的方向,观察总体趋势，我们发现：

1) 车辆驶入驶出的方向具有相似的数据量

2) 东西两个方向是最受欢迎的两个方向

1. **数据来源**

<https://www.kaggle.com/c/bigquery-geotab-intersection-congestion>