科技战疫·大数据公益挑战赛

2020 北京数据开放创新应用大赛

主办单位：
北京市经济和信息化局
中国计算机学会大数据专家委员会

重点区域人群预测

自由之翼

麦多健　洪永团　王梓

团队简介

队长: 麦多健（大三）厦门大学

队员: 洪永团（大三）厦门大学

王　梓（大三）厦门大学

参赛经历：

2019 全国数学建模大赛　　福建省一等奖

赛题描述

1. 问题描述：本次比赛，选取北京市100个不同类别的重点区域，提供各区域历史多天分小时人群密度数据。同时，提供北京六环内历史多天分小时的网格（200*200）人群密度、北京市迁入和迁出指数、网格间联系强度指数。参赛者需要根据这些已知数据，结合自己从互联网上获取的其他任何数据，来预测接下来每天分小时北京市重点区域的人群密度。

2. 数据：共997个重点区域，1月17日至2月15日

3. 评估指标：

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$score = \frac{1}{1 + RMSE}$$

解题思路

特征构造

	特征	说明
时间特征	月份	
	一年的第几天	
	一个月的第几天	
	一周的第几天	
	是否是周末	
	小时	
	节日划分	除夕前、除夕到元宵之间和元宵后
	一天的划分	采取简单的早中晚划分
天气特征	最高温	在疫情和春节两个因素的影响下，天气对人们的出行影响变得微乎其微
	最低温	
	气候类别	

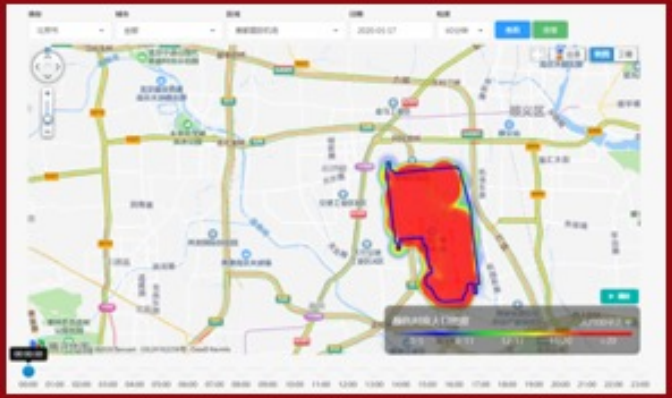
特征构造

	特征	说明
窗口特征（历史特征）	平均值	由上一周工作日、上一周周末、上一周今天计算
	最小值	
	标准差	
区域特征	区域名称	后面采取分区域进行模型预测，不考虑该部分特征
	区域类别	
	...	
其它特征	北京市网格联系强度表格得到的各个小时的区域人流入度和出度	

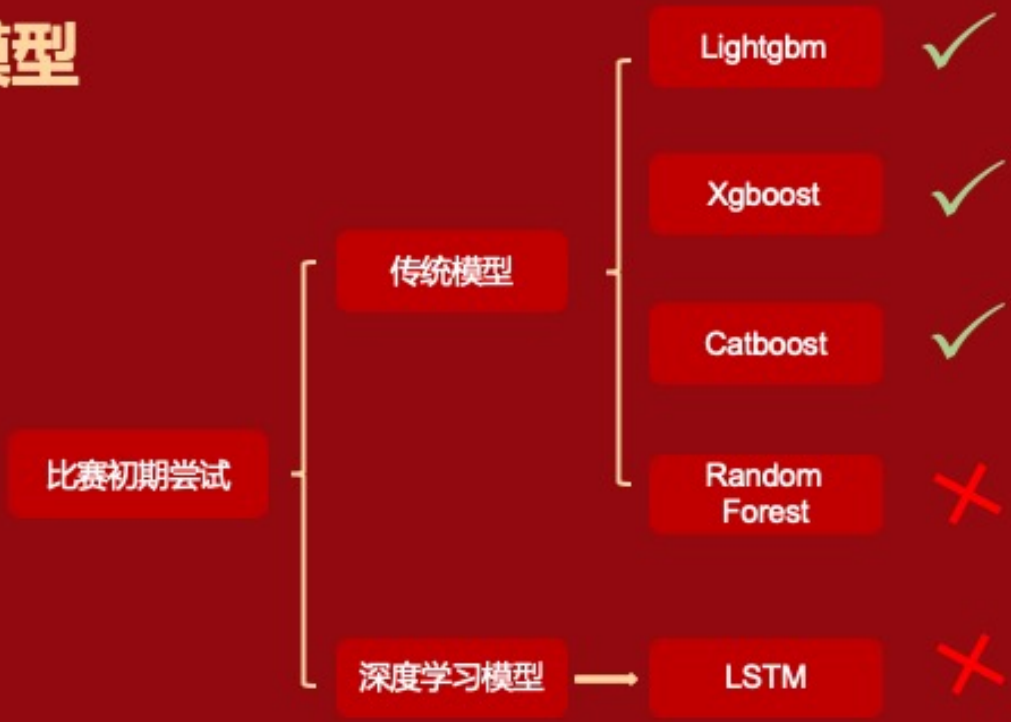
特征构造

	特征	说明
百度迁徙数据特征	北京迁入指数	该数据可以让lgb单模A榜提升0.01
	北京迁出指数	
	北京城内出行强度	
腾讯位置大数据特征（部分区域）	位置流量趋势	该数据可以让lgb单模A榜提升0.01
	最大热力值	
	最小热力值	

特征构造

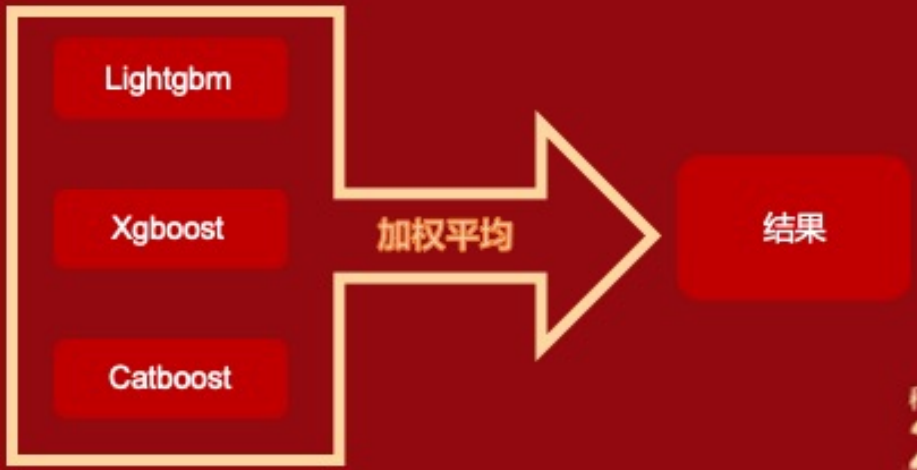


算法模型


比赛初期尝试
传统模型
Lightgbm
Xgboost
Catboost
Random Forest
深度学习模型
LSTM

训练思路

具体训练中我们的思路主要是先进行总体训练，在总体训练较好的参数基础上，分区域ID单独训练，根据各区域的特点对参数进行调整。



模型结果


block_predict_result.csv
所在赛程:
状态 / 得分
提交时间
第一阶段 - B榜
0.14417052000
查看日志
2020/04/30 11:55
备注: 无备注信息

比赛总结

该赛题的官方数据十分干净，省去了很多数据清洗的操作，可以花更多的时间进行数据的思考。模型融合很重要，最后加入Catboost模型进行融合，效果意外的好。有些和人流密度相关性很大的特征在后期进行了舍弃（如区域类别、区域名等），很遗憾没有发现一个好的处理方式。

THANKS

科技战疫·大数据公益挑战赛
2020 北京数据开放创新应用大赛