



杭州地铁数据预测分析

数据挖掘课程项目报告

答辩人：王梓茗 孙赳志 陈润澎



项目背景

目前我国有很多城市在大力发展**轨道交通**这一出行方式，但是随着轨道交通的发展，很多重要的交通枢纽处都出现了不同程度的**拥堵**。

因此为了能够解决这一问题，需要对其客流量能够进行**有效的预测**。

实时观测和把握客流量的**趋势**，向组织部门和运力部门的工作提供科学的数据，做到能够**预防和缓解**拥堵现象的发生。





实验数据来源

我们选取了来自**阿里云天池数据源**的共**7000万**条地铁刷卡数据，希望对其**人流量**数据进行分析与预测。

覆盖了1月1日~1月25日的**所有**地铁出入站信息。

原数据规模量很大、特征繁杂

该如何处理？

特征种类：刷卡时间、刷卡站点、刷卡线路、刷卡设备ID、进出站状态、用户ID、支付方式等

	A	B	C	D	E	F	G
1	time	lineID	stationID	deviceID	status	userID	payType
2	2019/1/13 0:00	C		35	1672	0 Ca6924a5e	2
3	2019/1/13 0:00	C	64	2981	0 B61dfe8df	1	
4	2019/1/13 0:00	C		35	1672	0 Bb3ec6817	1
5	2019/1/13 0:00	B		16	894	0 D9856b18	3
6	2019/1/13 0:00	C		35	1671	0 B8d8345f6	1
7	2019/1/13 0:00	C		35	1673	0 B3af4051f3	1
8	2019/1/13 0:00	C		64	2979	0 D4d7d126	3
9	2019/1/13 0:00	C		66	30		1
10	2019/1/13 0:00	C		35	1		1
11	2019/1/13 0:01	B		5			3
12	2019/1/13 0:01	B		6			3
13	2019/1/13 0:02	C		65	30		2
14	2019/1/13 0:02	C		34	1		1
15	2019/1/13 0:02	C		34	1		1
16	2019/1/13 0:02	C		65	30		1
17	2019/1/13 0:02	C		37	1		0
18	2019/1/13 0:02	C		65	30		2
19	2019/1/13 0:02	A		77	3529	0 Daab52a3d1	3
20	2019/1/13 0:02	A		77	35	大小:	1.52 GB (3)
21	2019/1/13 0:02	A		77	35		3
22	2019/1/13 0:02	B		28	14	占用空间:	1.52 GB (3)
23	2019/1/13 0:02	B		28	14		3
24	2019/1/13 0:02	C		65	30		1



大规模原数据带来挑战



数据条目数量过大
需要进行**数据聚合**处理

模型复杂，影响变量过多
需要进行**多任务学习**

冗余特征信息过多
需要进行**特征工程**处理



数据聚合处理

我们的原数据达到了**1.5GB**之大，因此我们对于原数据进行了**数据聚合处理**，从海量数据中提取**有效特征**进行处理。

思路：加粗统计粒度，将**单一时刻**的刷卡记录，转化为单位**时间段**内进出的**吞吐量**数据。

```
time, lineID, stationID, deviceID, status, userID, payType  
2019-01-09 00:00:18 C, 35, 1674, 0, B6e2f8a4498af5df1a75064  
2019-01-09 00:00:27 C, 35, 1673, 0, C409614c3ad090164ca7fa2  
2019-01-09 00:00:29 C, 59, 2784, 1, D837fc448740932e32ea82a  
2019-01-09 00:00:37 C, 35, 1674, 0, B49fe907c9a112449e68247
```

数据聚合处理
吞吐量统计

```
stationID, startTime, endTime, inNums, outNums  
1, 2019-01-09 07:10:00, 2019-01-09 07:20:00, 19, 15  
1, 2019-01-09 07:20:00, 2019-01-09 07:30:00, 83, 15  
1, 2019-01-09 07:30:00, 2019-01-09 07:40:00, 22, 25
```

结果：压缩300倍以上！
极大地加速了训练过程





多任务学习：周末/工作日

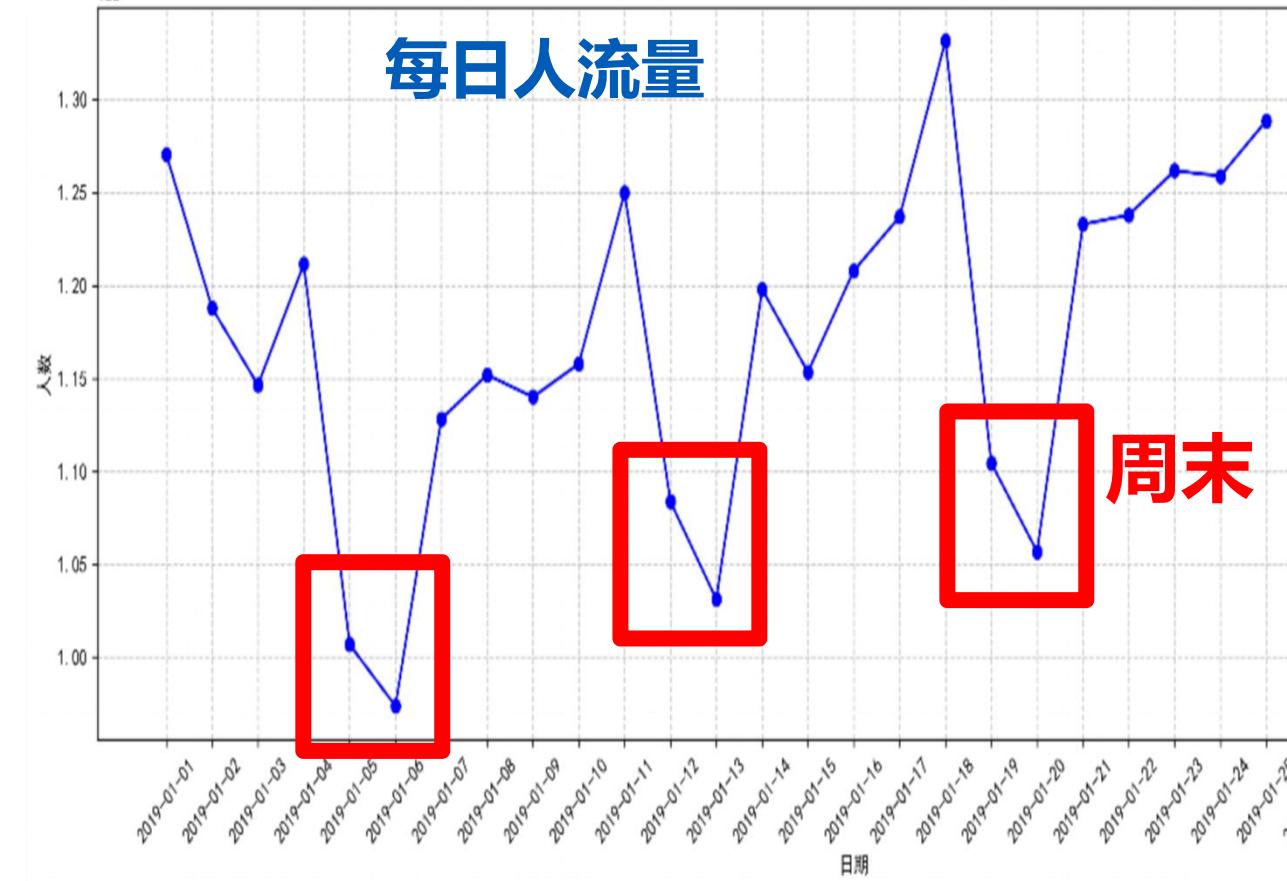
我们对于初筛后的数据进行了统计。

发掘规律：

进出站人数在**周末**和**工作日**呈现明显的差异，每周的进出站人数呈现**周期性**。

确立目标：

分别对于**周末**和**工作日**的人流量预测。





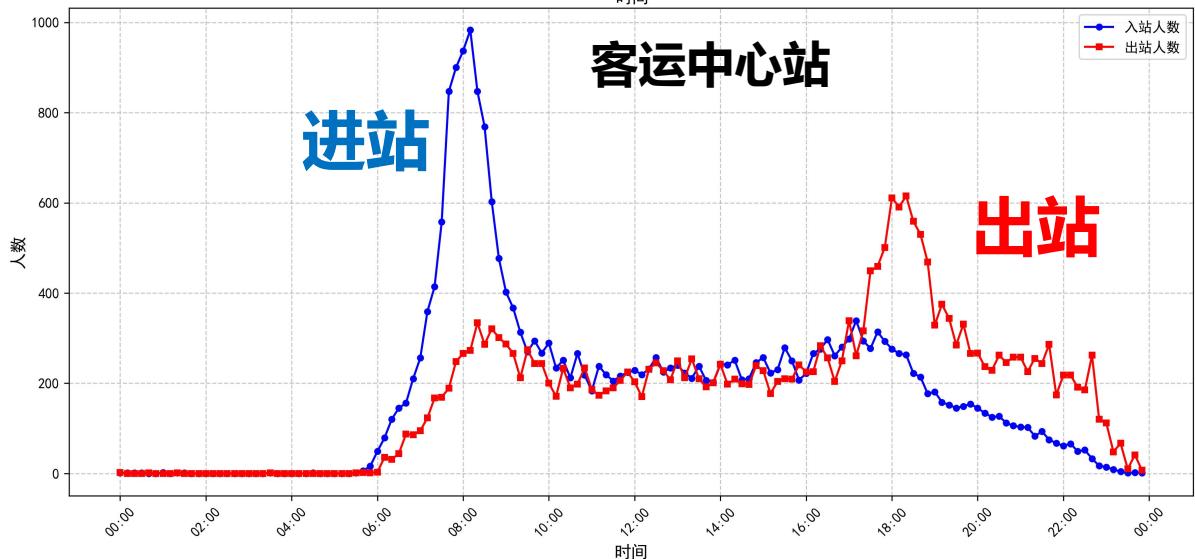
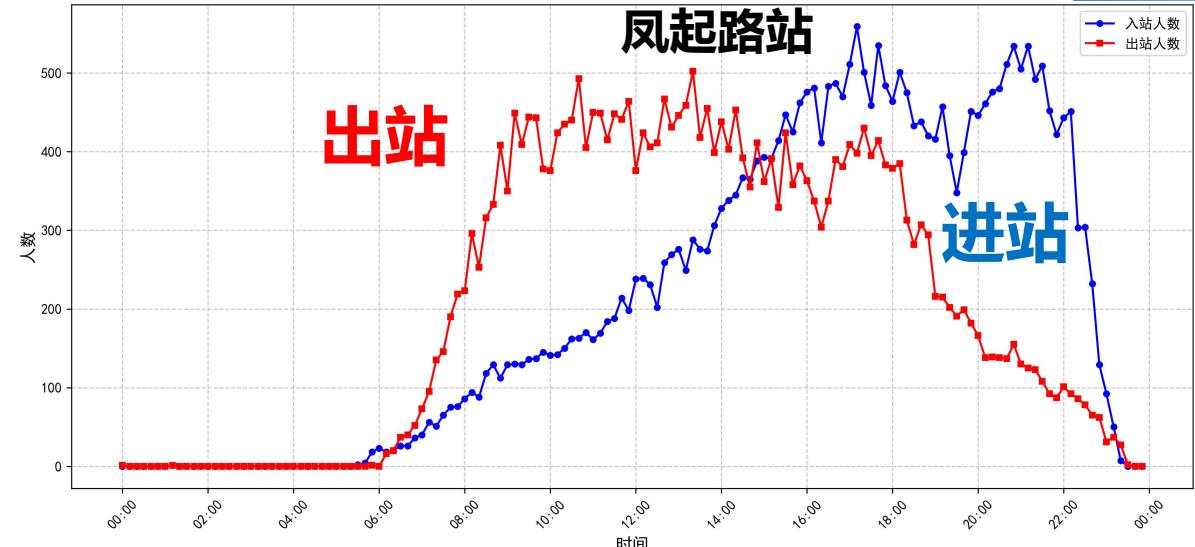
多任务学习：出入站人数

发掘规律：

单日内的**进站人数**与**出站人数**之间存在差异性

一天内的**不同时间、不同站点**，出入站的数据也存在差异。

确立目标：需要将**出/入站信息**与**站点信息**纳入多任务预测范围内。





训练策略

至此，我们得出了**训练策略**：

1. 对于**周末和工作日**，分别进行模型训练
2. 对于**入站**数据与**出站**数据，分别进行模型训练
3. 对于不同的**典型站点**，分别进行模型训练

此外，考虑到每个时间点的特征存在时序连续性，我们使用**滑动窗口**方法，基于过去24小时内的数据来更新特征。



预测模型选择

选用了**RandomForestRegressor**（随机森林回归器）作为主要预测模型

选择原因：

能处理**非线性关系**、对异常值**不敏感**、可以评估**特征重要性**、
不需要对数据做严格的分布假设、**较少**的参数调优需求

另外，还选用了**GRB**（梯度提升回归树）和**DecisionTree**
（决策树）作为对比。



评估指标

评估指标：

使用**R²分数**衡量模型拟合优度，代表模型能解释的**变异比例**。

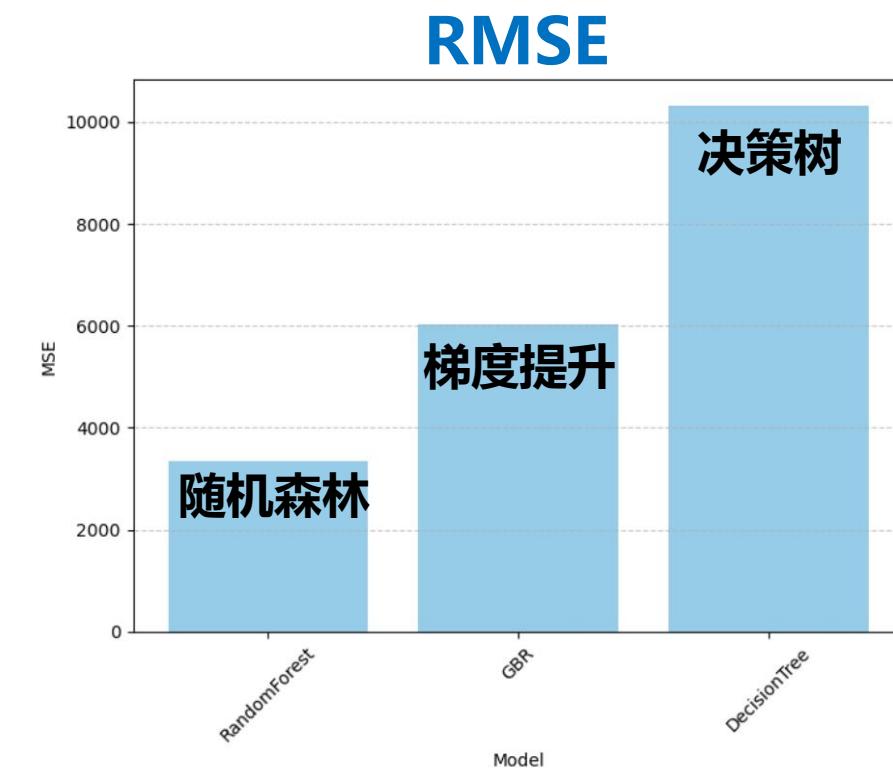
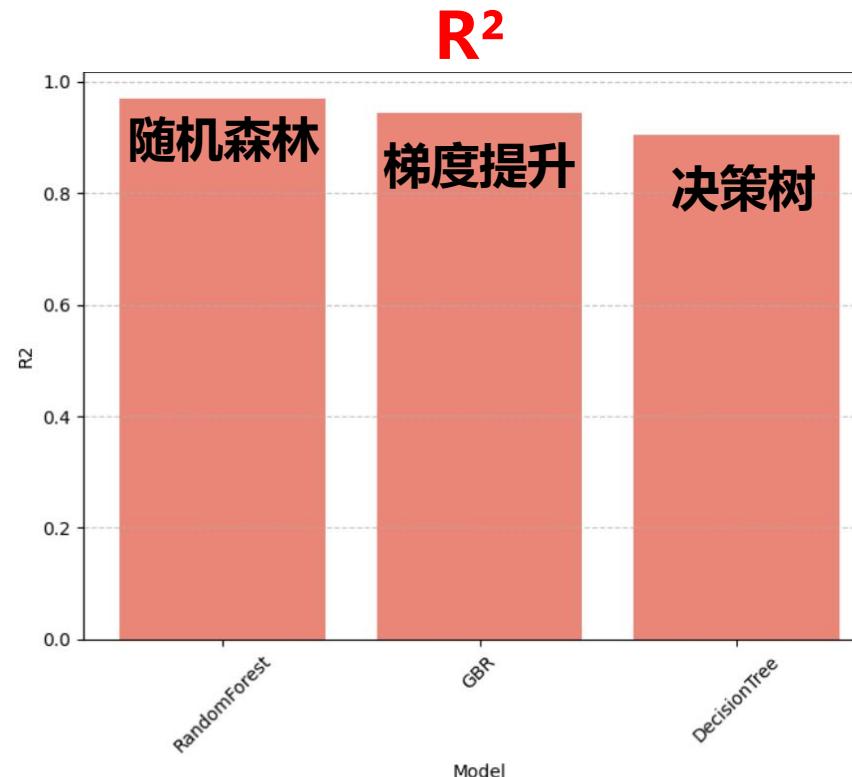
使用**RMSE**衡量预测值和实际值之间的偏差，单位与原始数据相同，使得结果更容易理解。

特征重要性评估，显示每个输入特征对预测结果的**贡献程度**，帮助我们理解哪些因素对预测更重要，对后续的特征工程和模型优化很有帮助。

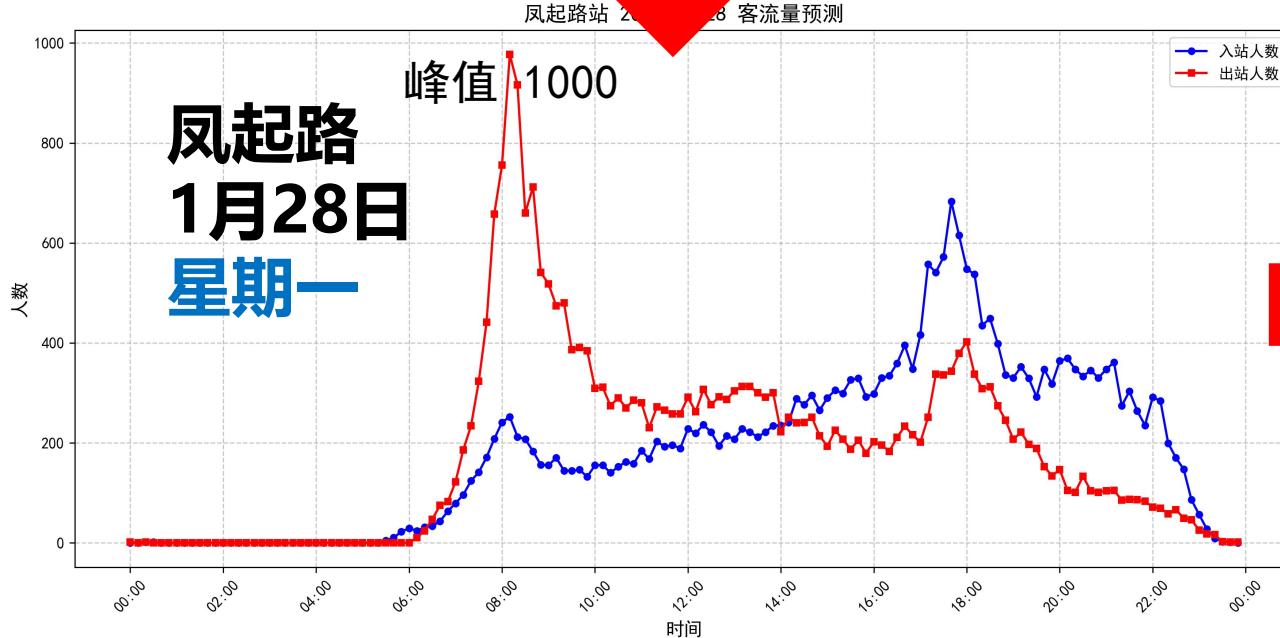
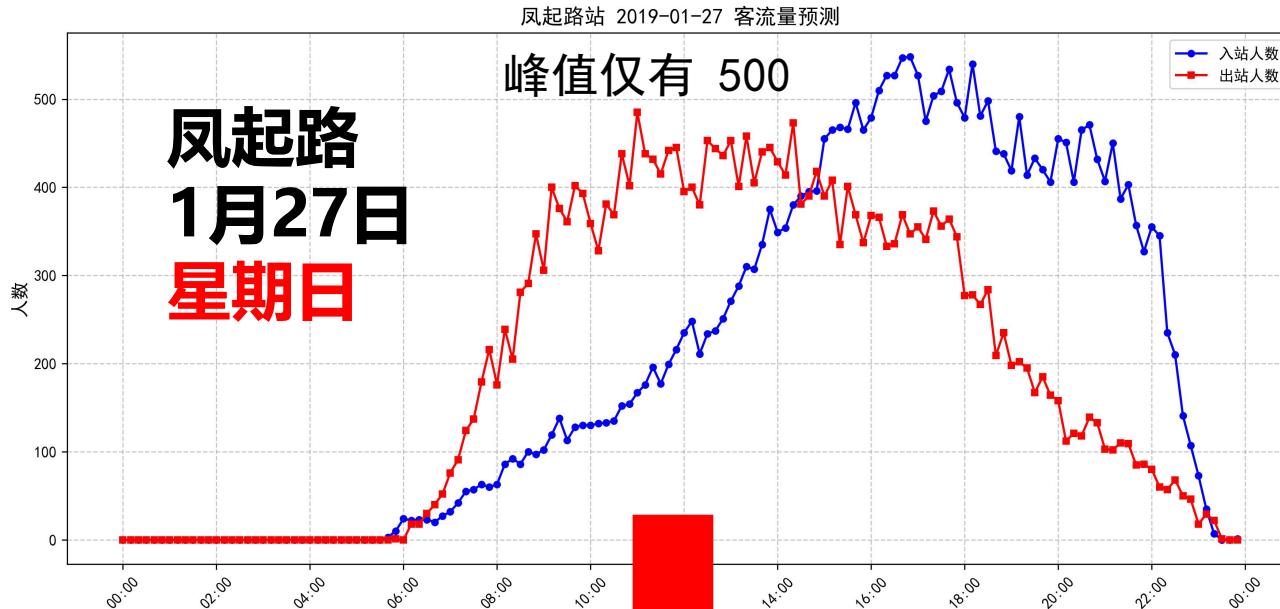


训练结果展示

训练结果如下，可以看到，**随机森林**在RMSE和R²上的表现均优于其他两个模型。这证明了我们模型选择的**合理性**。



不同日期预测结果展示



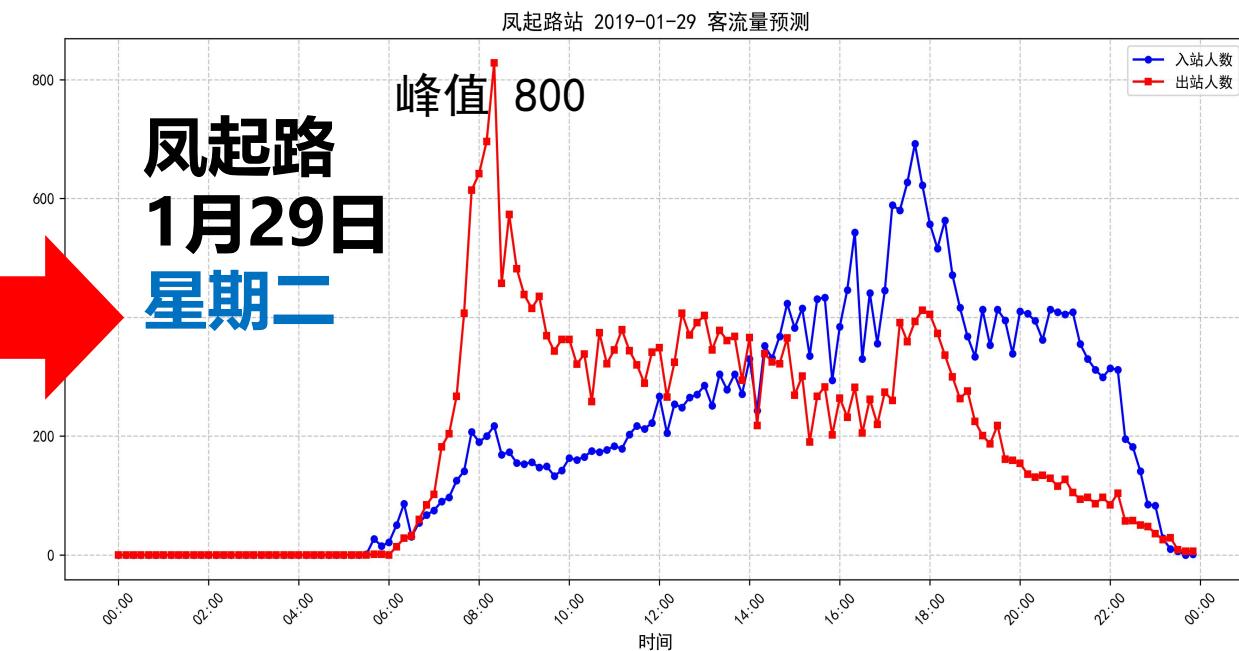
凤起路1.27~1.29期间预测如图。

周末与工作日流量趋向完全不同。

同类型的工作日之间流量类似，仅

存在峰值区别。

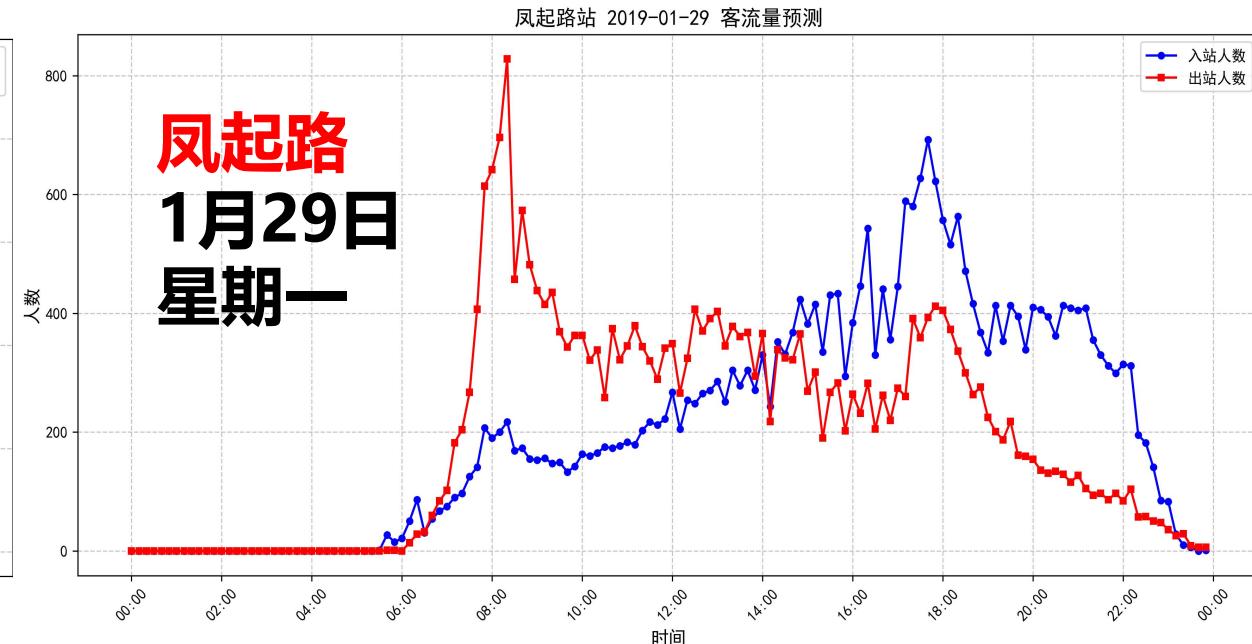
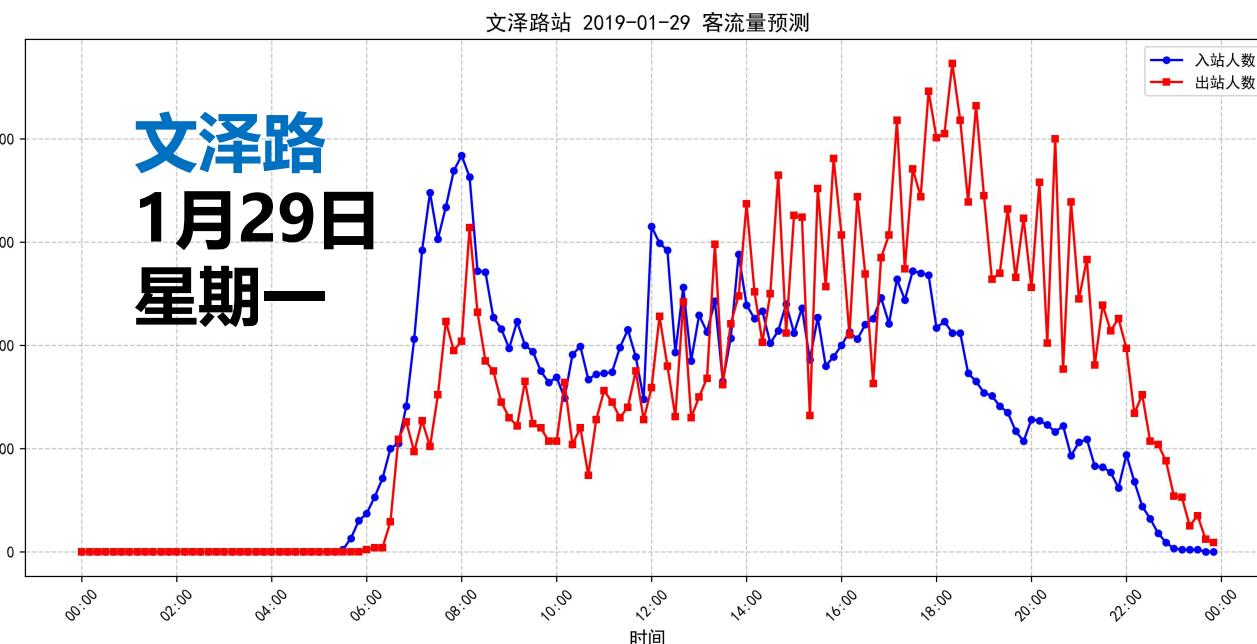
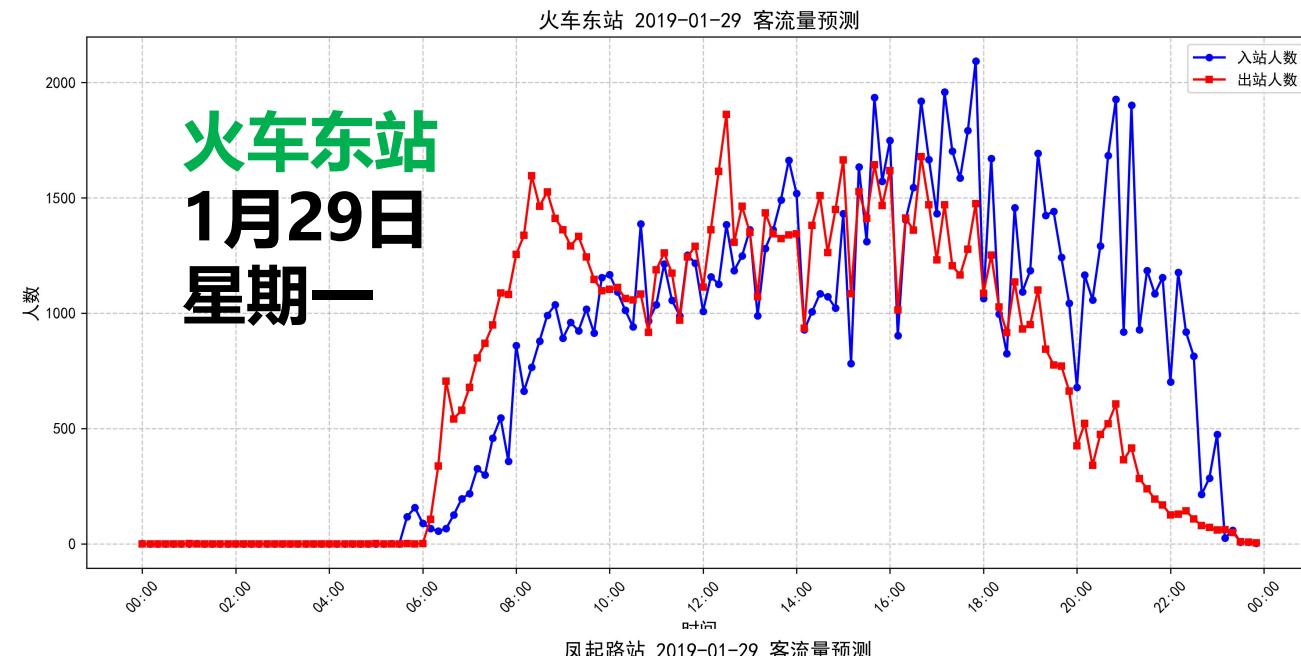
这完全符合我们的预期。



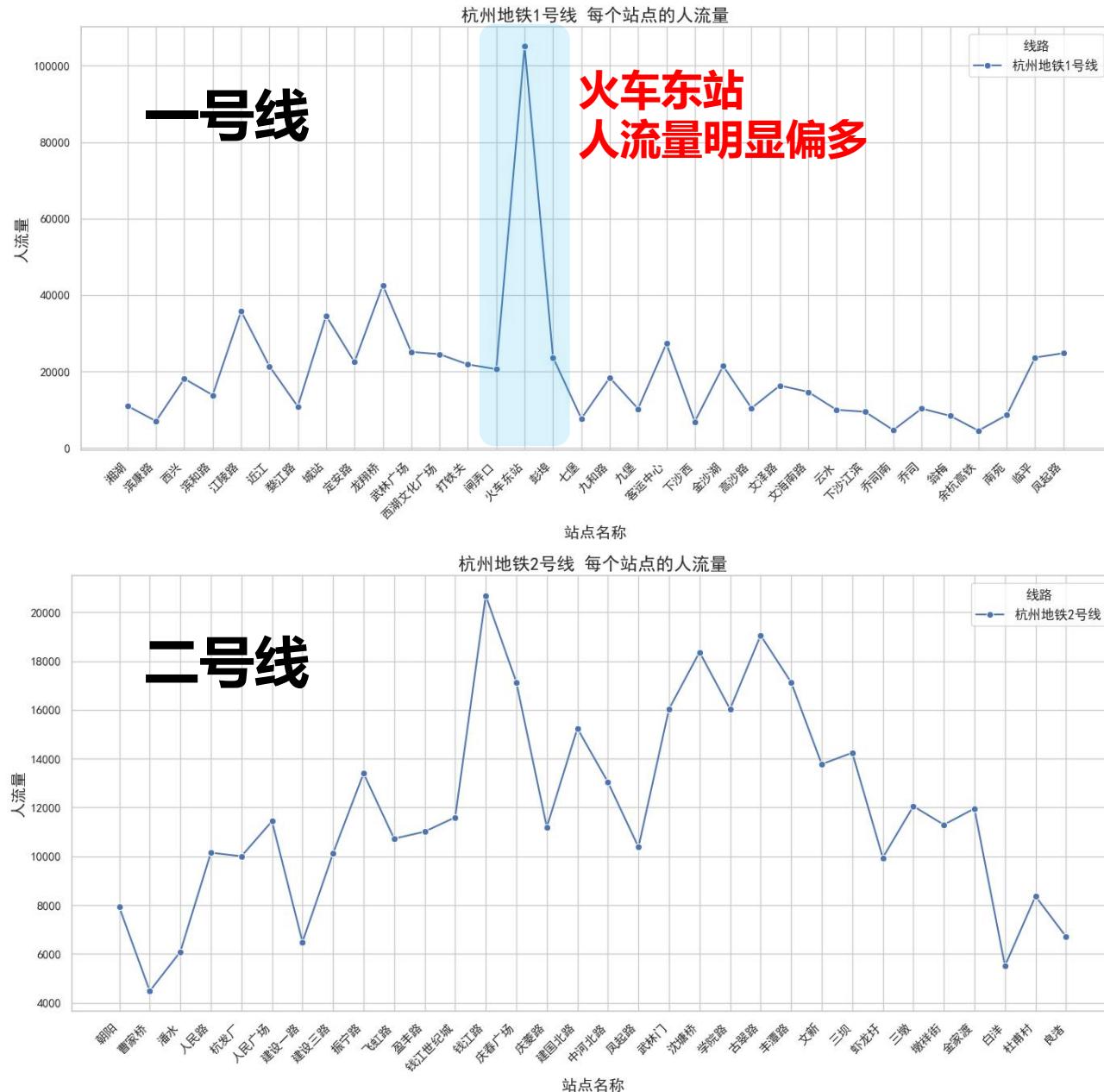
不同站点预测结果展示

火车东站、文泽路、凤起路三站对比

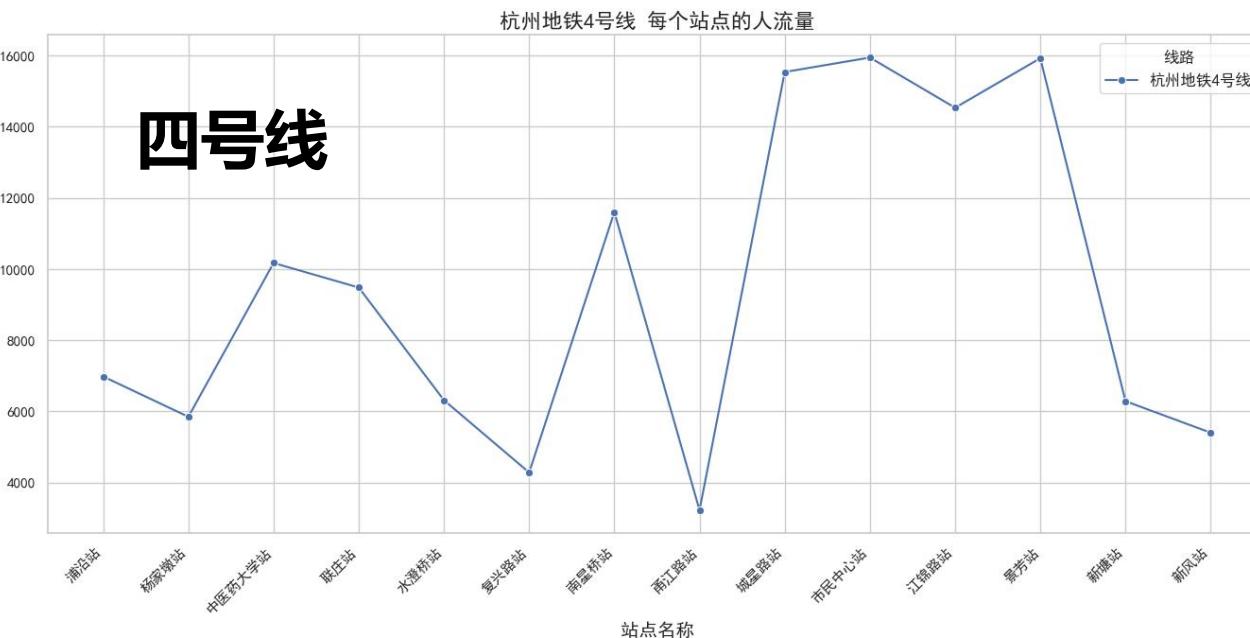
可以发现，尽管日期相同，但不同站点间的预测结果也会产生差异。
这也符合我们的预期。



不同路线、站点流量统计



此外，我们对于**不同路线、不同站点**的流量进行了统计。
部分站点，如**一号线的火车东站**人流量**明显偏多**，可以进行适当的**分流**。





总结

在本次实验中，我们对于**数据量庞大的**地铁人流量信息进行了数据挖掘，并对于人流量趋势进行了预测。

在预测建模过程中，我们从**出入站，周末或工作日，站点，时间段**等多个方面进行细分，以**多任务学习**的方式对其分别进行建模与预测。

训练后的模型对于人流量做出了良好预测，在输出的图像中我们进一步挖掘了不同**日期、站点**的流量信息，发现它们很好地符合了预期。

总体而言，我们的模型的确能够**胜任**一部分杭州地铁人流量预测和分流的工作，假如加以应用，该模型或许能够拥有一定的**实践意义**。



谢谢大家

THANK YOU FOR YOUR CRITICISM AND CORRECTION

答辩人：王梓茗 孙赳志 陈润澎