各位老师，同学们，早上好

今天我汇报的项目主题是酒店预订需求分析和取消率预测

在疫情放开的当下，我相信很多人都会把旅游的计划提上日程。提到出游就不得不牵扯到酒店的预定。在旅游业因为疫情防控低迷了几年后，在不久的未来许多酒店的订单量肯定会因为人们增加的出游计划而有所回升，对于酒店服务方而言，分析用户行为可以更好的针对需求做出服务调整，增加收益；对于游客而言知道酒店订单量的时间分布和服务优缺点可以更好的调整出游时间和规划目标酒店的选取。

我的汇报将从一下四个方面展开：分别是。。。。。

首先是数据集介绍和准备工作：数据集是kaggle上一些包含酒店预定信息的数据集，时间跨度是2015 7月到 2017年8月，看一下info我们可以知道这个数据集特征值比较多，有32条

接下来是预处理

一些不同类型不同程度的缺失值我用删除、用众数、或者用符合逻辑的0进行填充

删除异常字段：比如说大人+小孩+婴儿=0的订单是肯定不能存在的

餐食选择国家等用了简写的可以进行字段填充方便分析

再经过数据去重以后，这是数据的变化

接下来是探索性分析，从。。。。。。

这是一些可视化的图，由于时间关系所有图的详细分析会在报告里呈现，这里大概讲一下结论

有样本不均衡的问题

顾客背景

经营状况

这里衡量旅游淡旺季的指标选取订单量而不是人均价格

城市酒店和度假酒店都是春季和秋季这些比较对于体感而言比较舒适的季节是旺季，房价也会相应变高。冬季生意都不好，度假酒店全年订单量波动小于城市酒店

就取消量，城市酒店几乎都比度假酒店高，但是度假酒店波动大，就过年期间低的原因猜测是

顾客行为上而言

可以发现住的时间越长越倾向于选择度假酒店，这点符合常识

特殊要求：跟团客人比临时排队的客人特殊要求多，7、8、12月多

不同渠道和取消率而言：线上线下旅行社预订量和取消量都是很高的，线上是人们最青睐的方式，有可能是操作便捷的原因

顾客行为：

历史取消行为多，取消率越高

越早预定取消率越高

有趣的是数据集里有顾客预定的房型和实际住到的房型，房型不匹配反而取消率低，猜测是……..有待进一步研究

相关性分析：热力图，通过查看与取消率的相关系数我们可以知道最相关的五个变量

接下来是取消率预测，我选择了六个base model 基本思路是先在六个base上跑一边然后找出表现最好的模型再对它进行优化

准备工作是导入一堆机器学习的包然后进行特征分类。这里我是手动去掉了一些特征的，去掉哪些的标准跟我选取的评估方法有关

分离整数型和分类型特征

接下来是选取评估模型的统计方法，我选的是分层k-fold交叉验证，为什么选它就是因为前面探索性研究的时候发现样本分布有不平衡的情况，缺点是不适合有时间序列分布的数据，所以在前面手动选择特征列的时候我就把一些跟时间相关的去除了