

**（酒店预定需求分析&取消率预测）**

**姓 名： 杨茜雅**

**学 号： 10215501435**

**学 院： 数据科学与工程学院**

**课 程： 数据科学与工程导论**

**2023年 1 月**

**目录**

1. **实验引言**

**1.1 背景介绍**

**1.2 数据集简介**

1. **问题描述**
2. **实验过程**

**3.1 数据预处理**

**3.2 探索性分析EDA**

**3.2.1 总体概况**

**3.2.2 顾客背景**

**3.2.3 酒店经营状况**

**3.2.4 顾客行为**

**3.2.4.1 天数选择**

**3.2.4.2 特殊要求**

**3.2.5 取消率**

**3.2.5.1 渠道选择**

**3.2.5.2 历史取消订单量**

**3.2.5.3 提前预定时长**

**3.2.5.4 平均每日房价**

**3.2.5.5 房型匹配率**

**3.2.6 相关性分析**

**3.3 取消率预测**

**3.3.1 手动选取特征**

**3.3.2 特征分类**

**3.3.3 特征预处理**

**3.3.4 选取评估模型的统计方法**

**3.3.5 模型评分展示**

1. **模型优化**
2. **特征评分**
3. **实验引言**

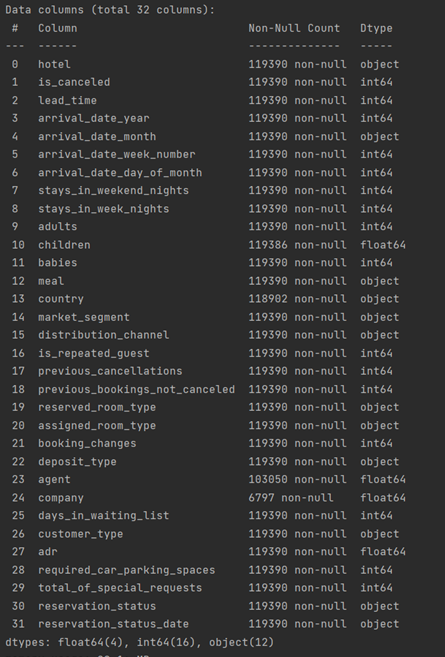
**1.1背景介绍**

**大数据时代，企业比以往任何时候都更加依赖数据分析用于优化和分析整体性能。数据分析提供了管理、保护、转换和交流信息的解决方案，以做出更有根据的决策。通过利用原始数据，业务数据分析生成可用和可理解的数据，企业利用这些数据来降低风险和成本，并提高盈利能力。本次项目运用到数据分析需要的关键技术之一——机器学习，作为人工智能的一个方面之一它允许软件发现并自动化业务模型，在本次酒店预定需求分析和取消率预测的项目中尽快归档并根据模型预测客人是否会取消订单。数据分析使酒店等服务场所可以使用数据来识别客人趋势并且更详细地了解客户需求，了解内部弱点，并创建各种解决方案来对标这些问题，提高整体盈利能力和客户口碑。**

**在疫情放开的当下，相信在不久的将来许多人会把因为疫情而推迟的旅行计划重新提上日程。提到旅行就不得不牵扯到酒店的预订，在旅游业因为疫情防控低迷了几年以后，酒店的订单量会因为人们增加的出游计划而有所回升。在订单量暴涨的情况下，分析顾客行为数据就显得更为重要。**

**1.2数据集简介**

**项目基于kaggle网站上名为Hotel booking demand的数据集进行研究，网址https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand。数据集包含了从2015年7月到2017年8月一家城市酒店和一家度假酒店的预定信息、包括诸如预定时间、逗留时间、成人、儿童、婴儿的数量，以及可用的停车位数量等32条信息。**

****

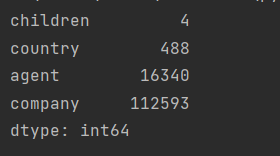
1. **问题描述**

**该项目中我首先对现有数据做了一些预处理，方便后续工作，再从总体概况、顾客背景、酒店经营状况、顾客行为、取消率和相关性分析的角度进行探索性分析。最后建立多个机器学习的模型来预测顾客是否会取消订单，并且针对表现最好的base model进行优化和特征评分。**

1. **实验过程**

**3.1 数据预处理**

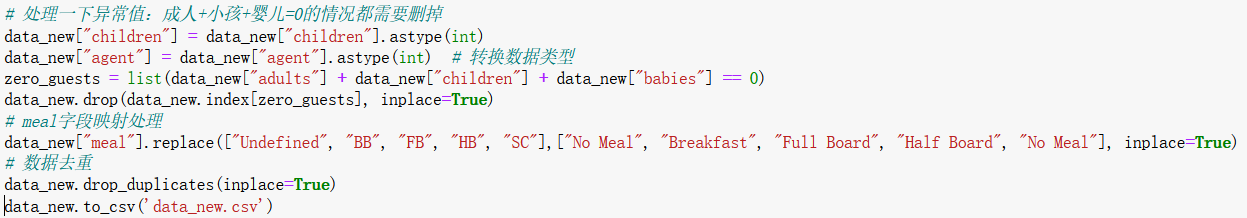
****

****

**通过查看缺失值可以发现company特征缺失过多，所以删去。对于缺失值较少的country和children，用字段内的众数填充。对于agent，则用0填充，表示没有指定任何机构。**

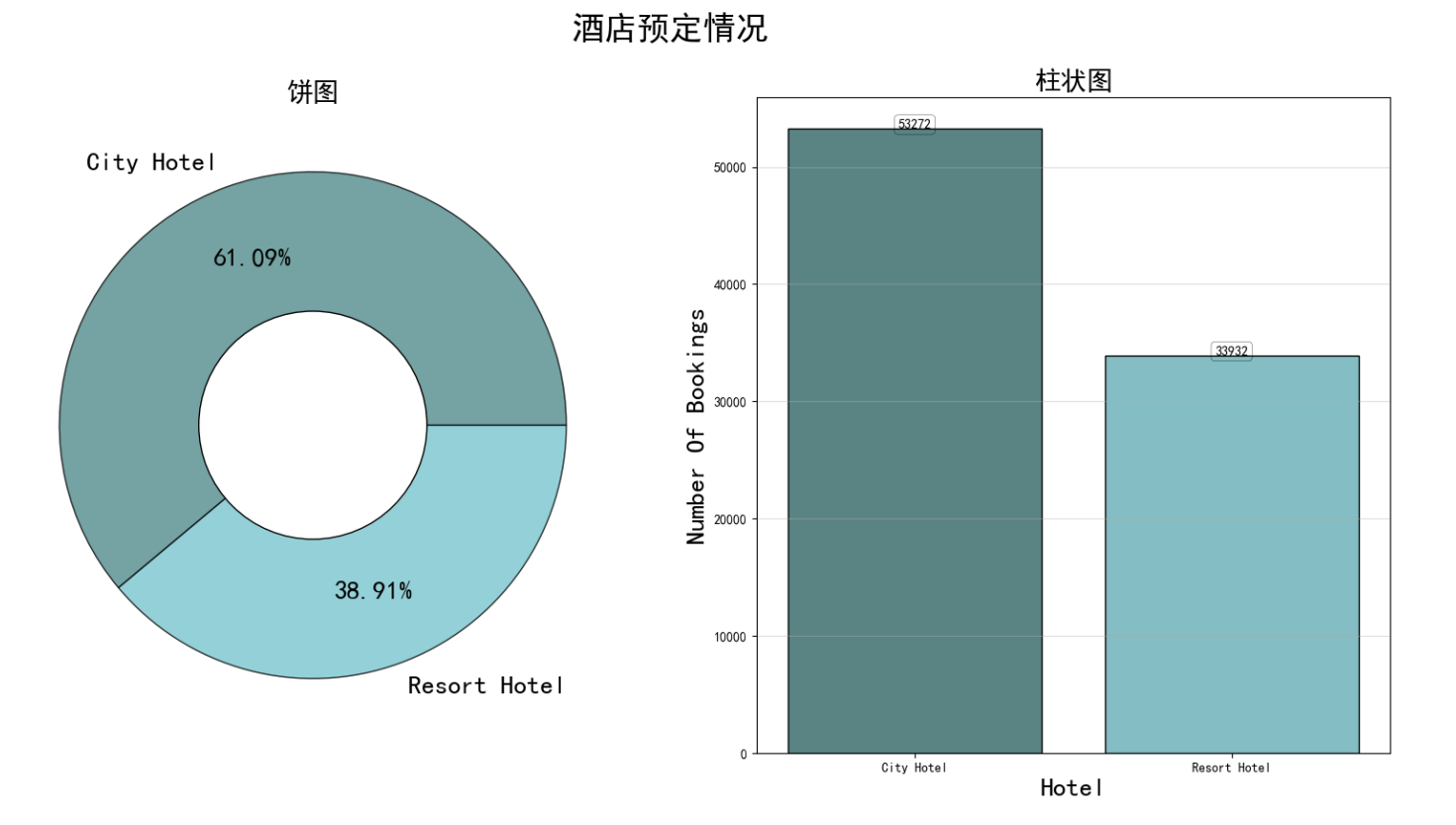
****

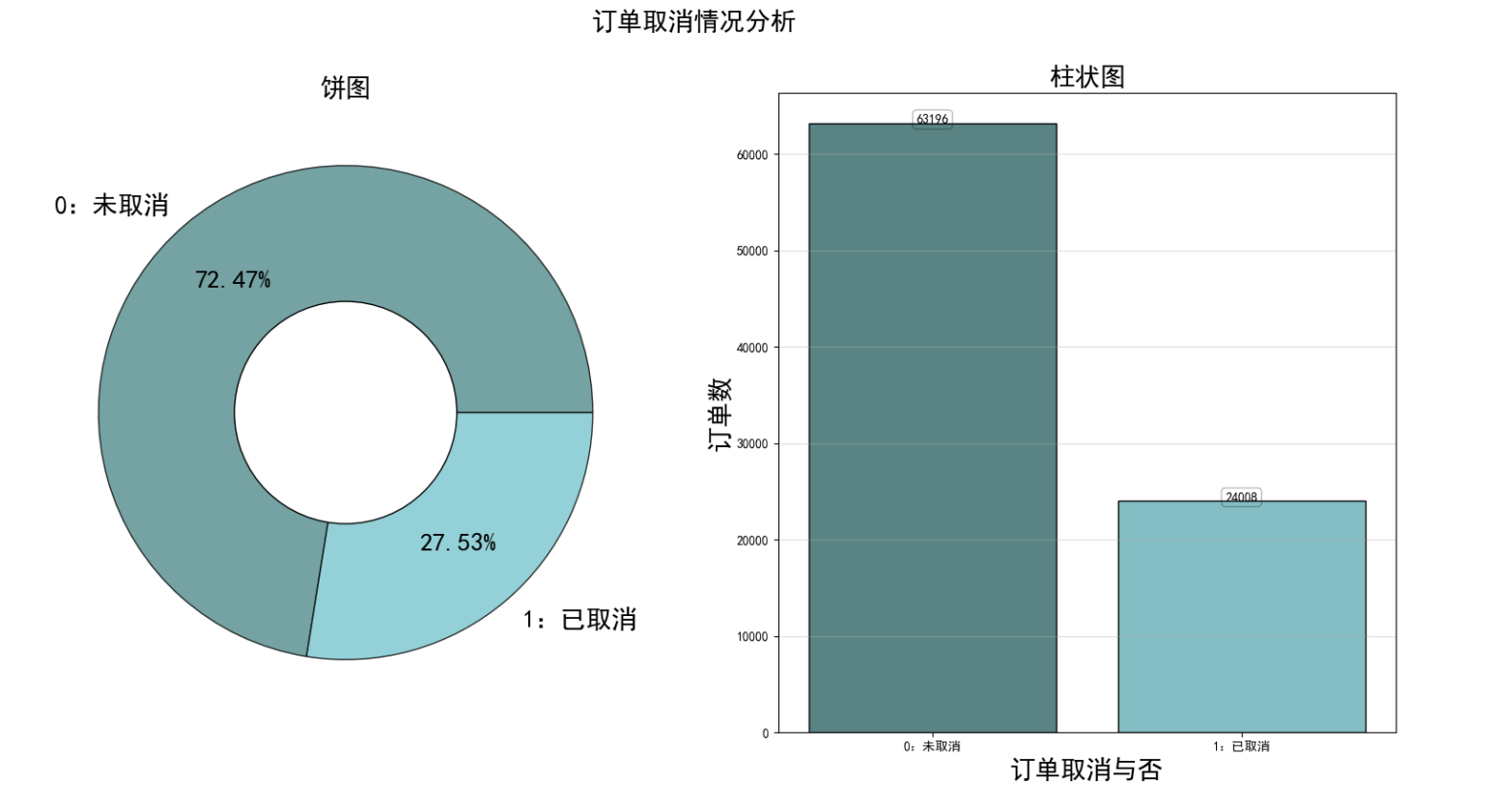
**同时对于异常数据进行删除，比如adults+children+babies=0的订单是不符合常理的，理应删除。对于用到简写的字段需要进行字段映射处理，增加可读性。最后形成新的数据集data\_new，规模从（119390，32）变为（87204，31）。**

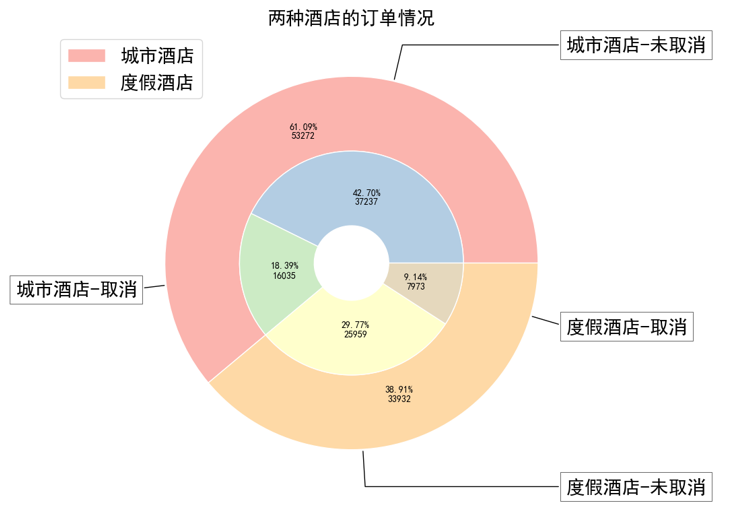


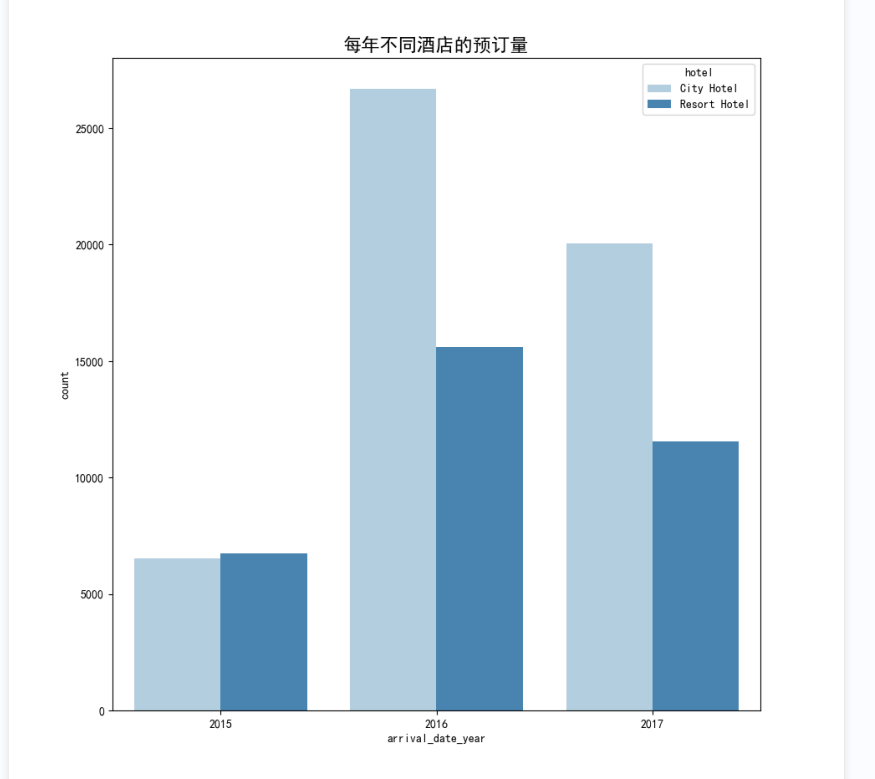
**3.2探索性分析**

**3.2.1总体概况（不同酒店的总预定量&&取消量&&年度订单走势）**

****

****

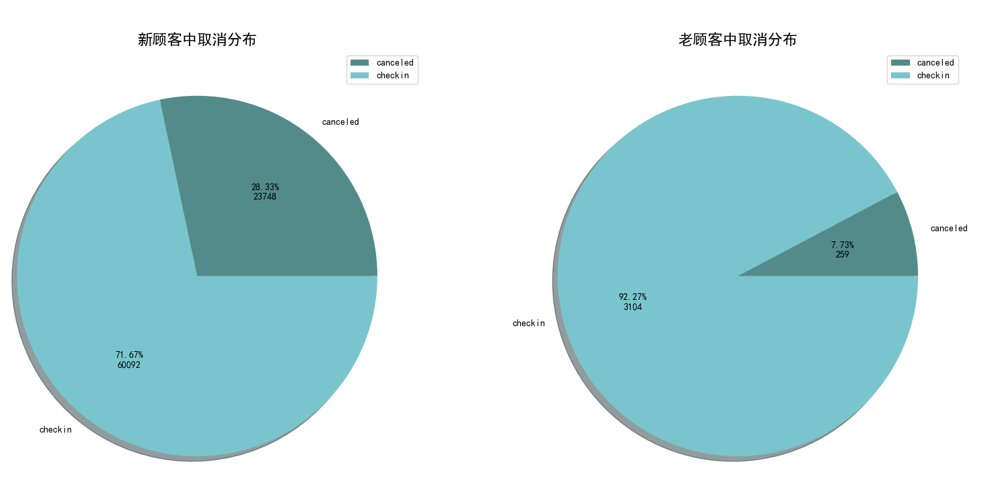


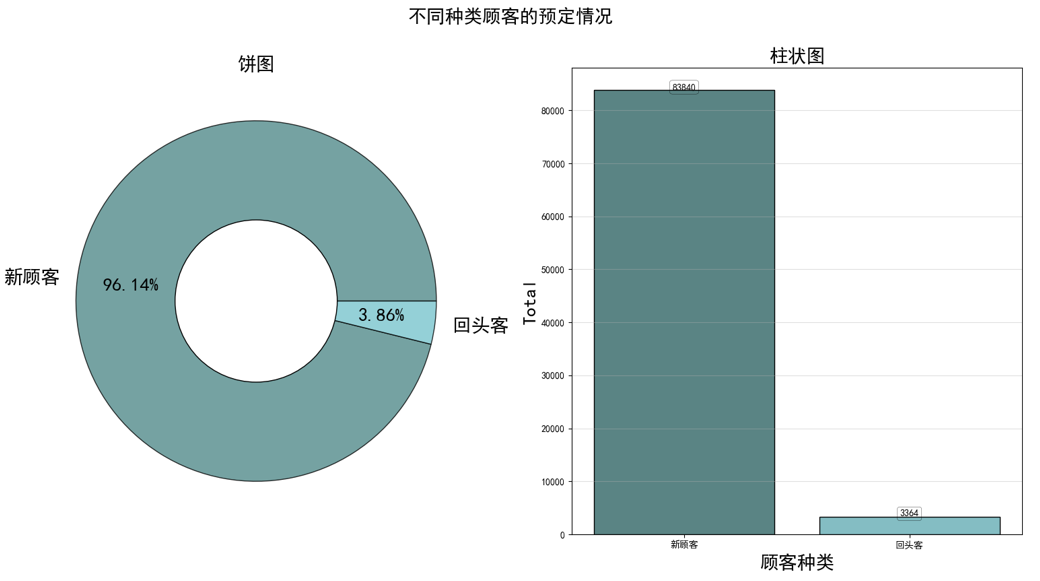
****

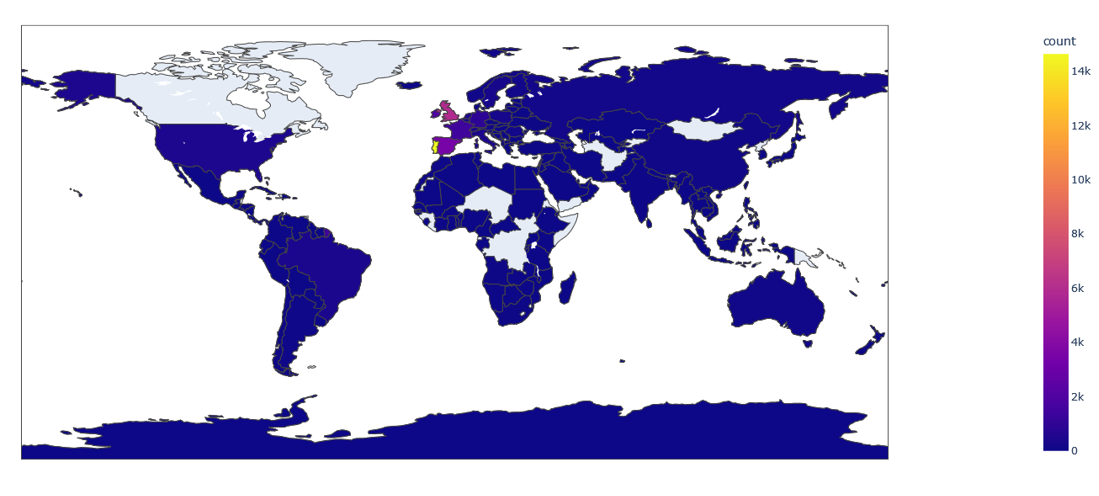
**结论：**

**城市酒店就取消量和预订量而言都高于度假酒店，年度订单量走势2016>2017>2015，订单取消和未取消比约为3：7，数据集存在样本不均衡的问题。**

**3.2.2 顾客背景（新老顾客分布&&取消率&&顾客地域分布）**



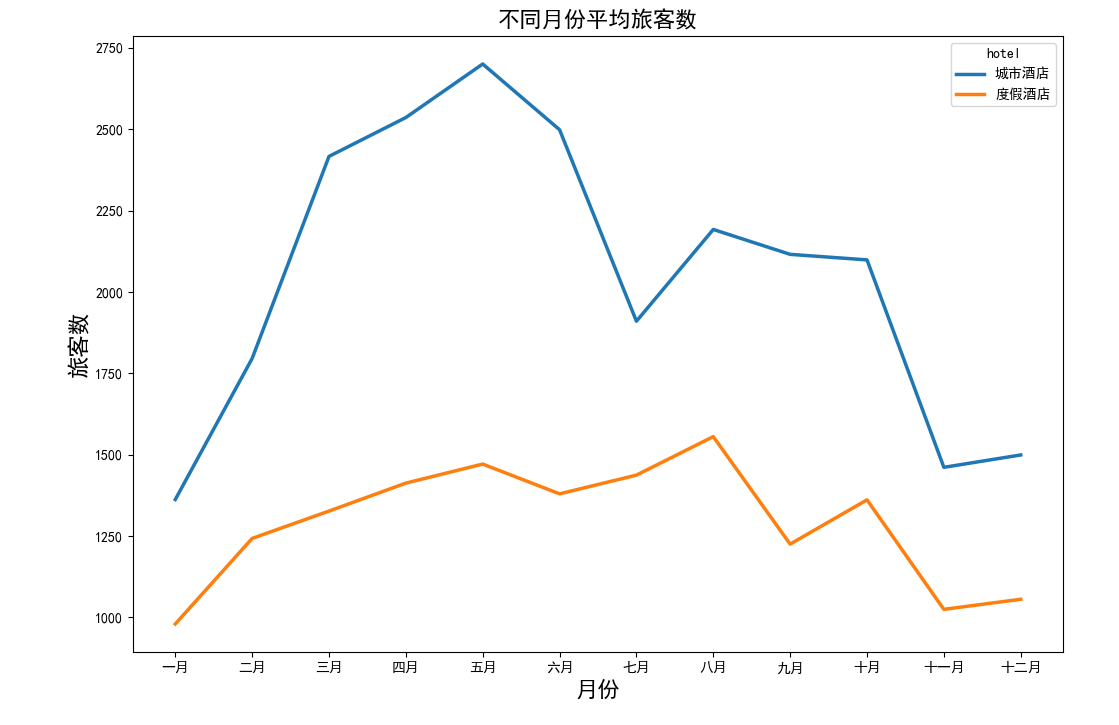


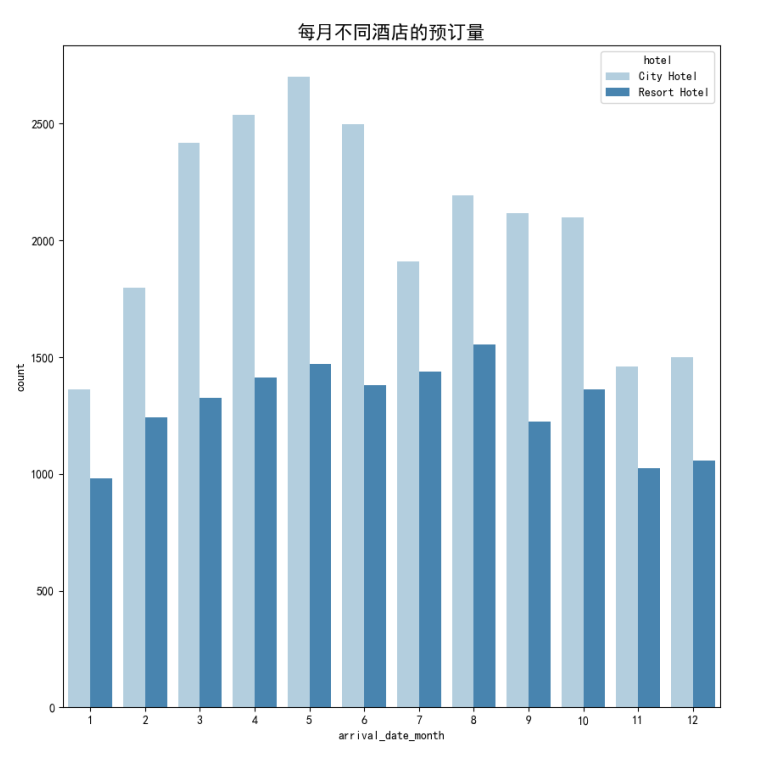


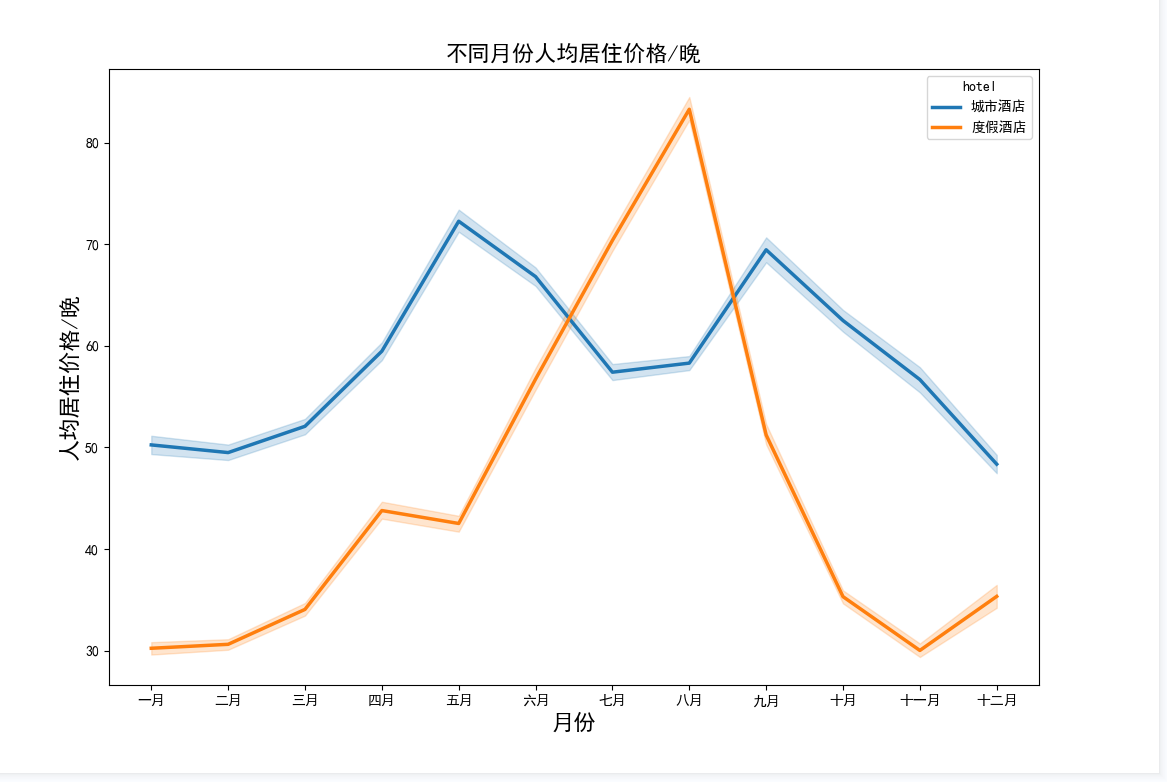
**结论：**

**老顾客几乎没有，所以就取消率而言也是新顾客>老顾客，地域分布情况而言顾客主要集中在葡萄牙、英国、法国、西班牙、德国等西欧地区。**

**3.2.3 酒店经营状况（每月不同酒店的平均旅客数&&预订量和取消率&&人均价格）**

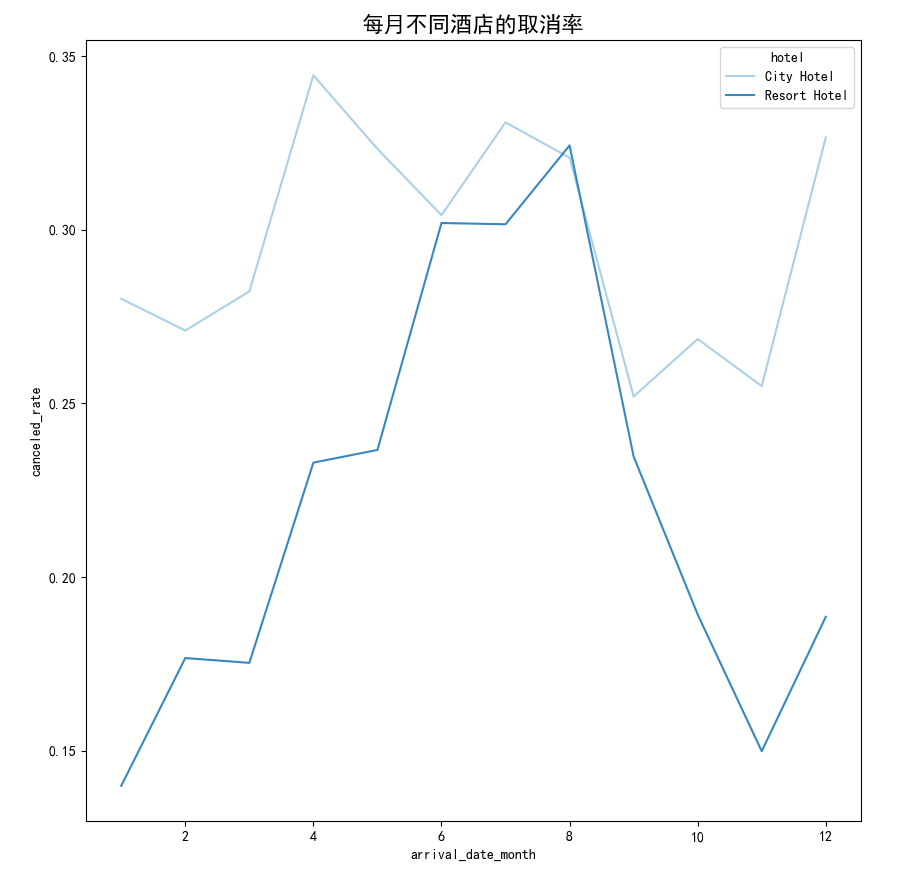
****

****

****

**旅客数和订单量的走势是大致相同的，此处选取订单量而不是人均价格作为衡量旅游淡旺季的指标。可以看出城市酒店在3-5月（春季）和8-10月（秋季）为预定旺季，房价也相应提高,订单数和总客流量也迎来峰值；度假酒店3-5月份（春季）和7-9月份（秋季）为预定旺季，房价稍有上浮。**

**度假酒店在7-8月房价很高，远高于其他月份，不排除有夏季人们扎堆出游的影响。两种酒店冬季生意都不太好。度假酒店全年订单量比较稳定。**

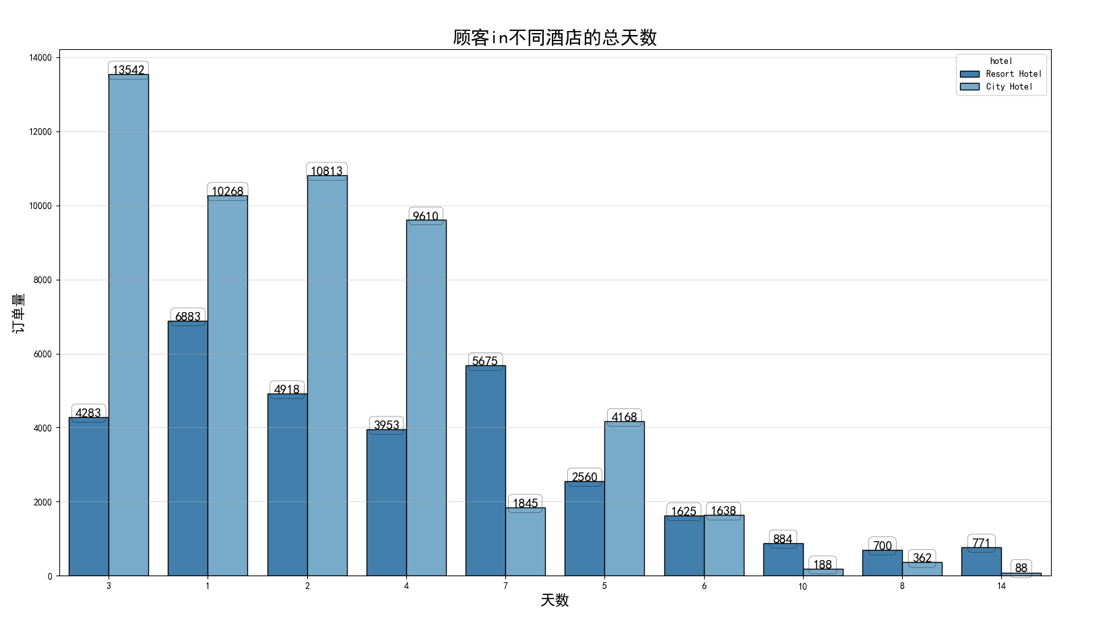
****

**但是就取消率而言，度假酒店就没有这么稳定了，全年起伏大，过年期间取消率极低，原因分析：1、冬季订单少。2、跨年旅游，不容易临时取消。**

**城市酒店取消率几乎全年都高于度假酒店，4月和7-8月和冬季取消率都偏高。**

**3.2.4 顾客行为**

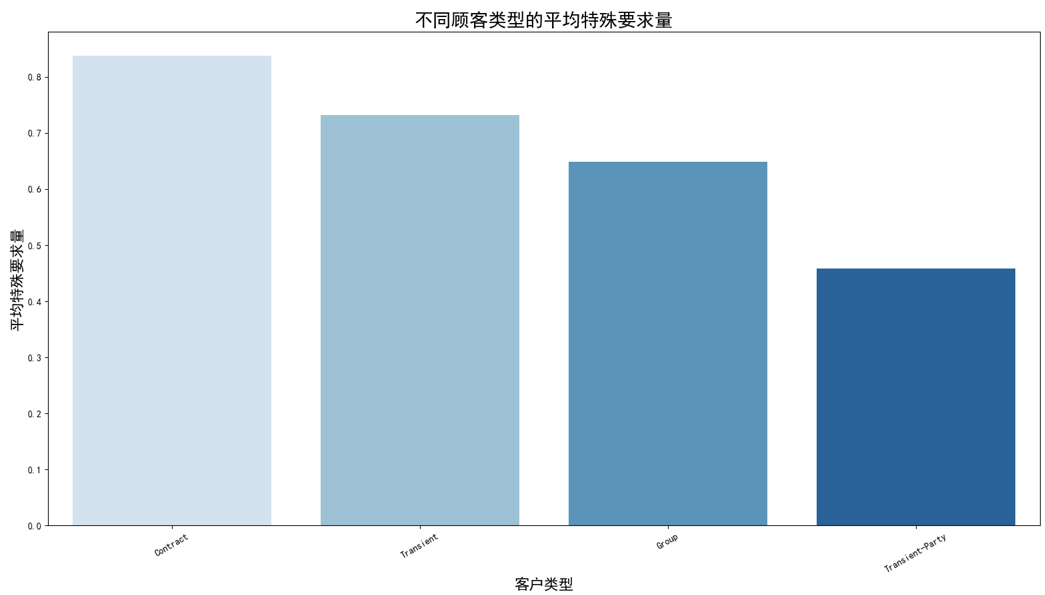
**3.2.4.1 天数选择**

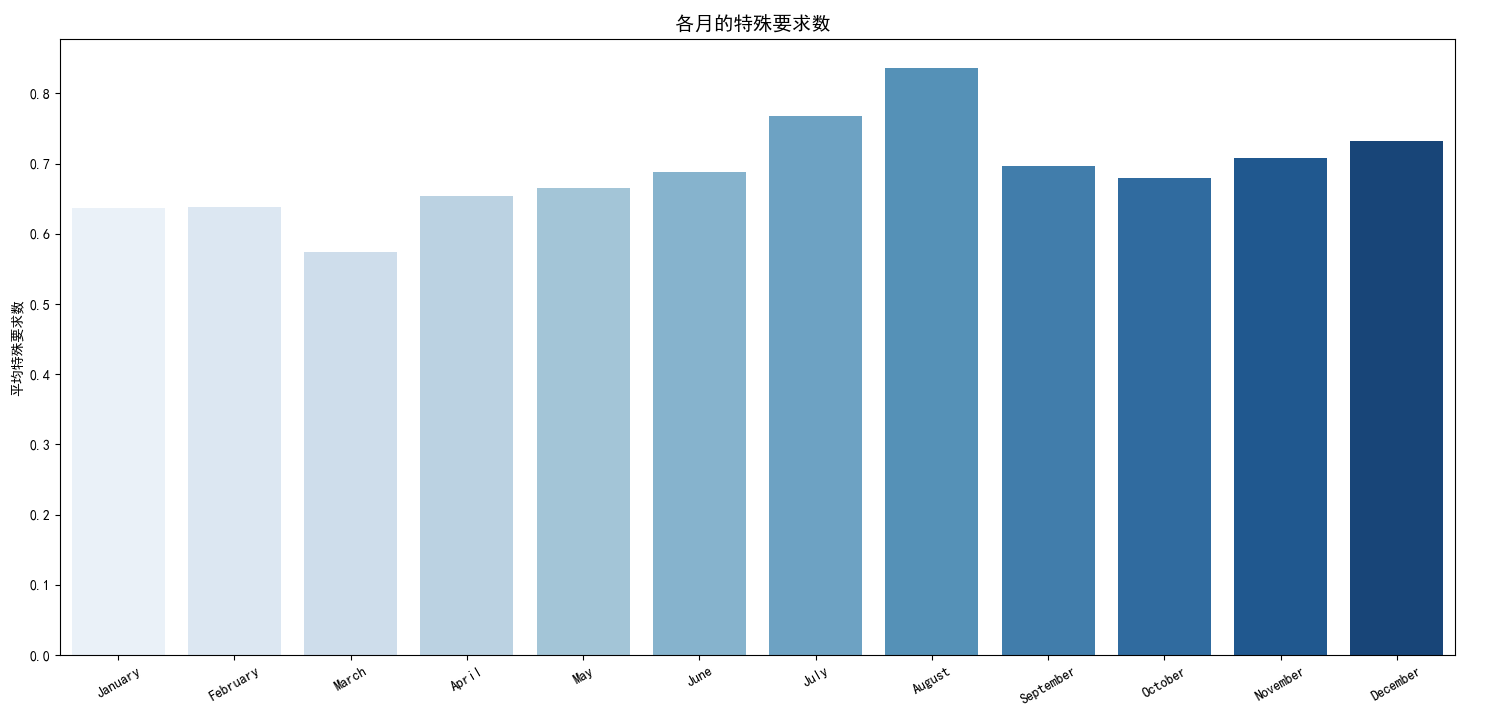
****

**度假酒店最受欢迎的住宿时间是一晚、七晚、两晚、三晚和四晚；城市酒店最受欢迎的住宿时间是三晚、两晚、一晚和四晚。**

**可以看出人们居住时间越长，越倾向于选择度假酒店。**

**3.2.4.2 特殊要求**

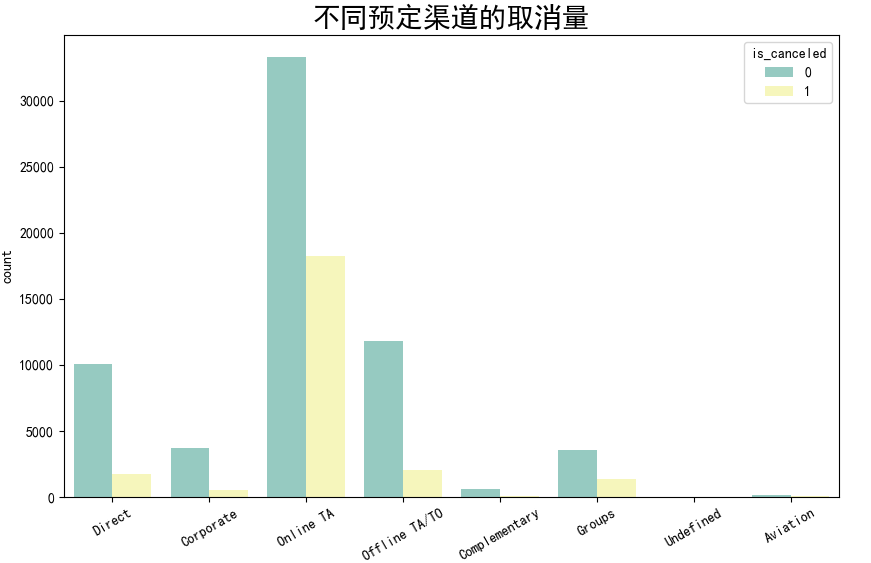
****

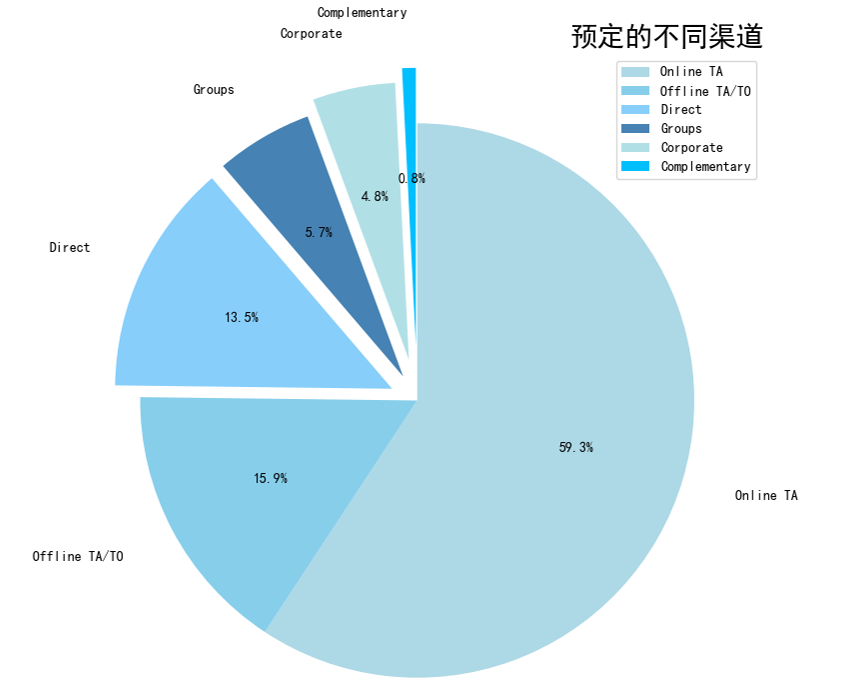
****

**跟团客人的特殊要求数量最多，临时派对客人的特殊要求数量最少。7、8和12月份这种旅游旺季或者有特殊节日意义的月份，特殊要求多一些。**

**3.2.5 取消率**

**3.2.5.1渠道选择**

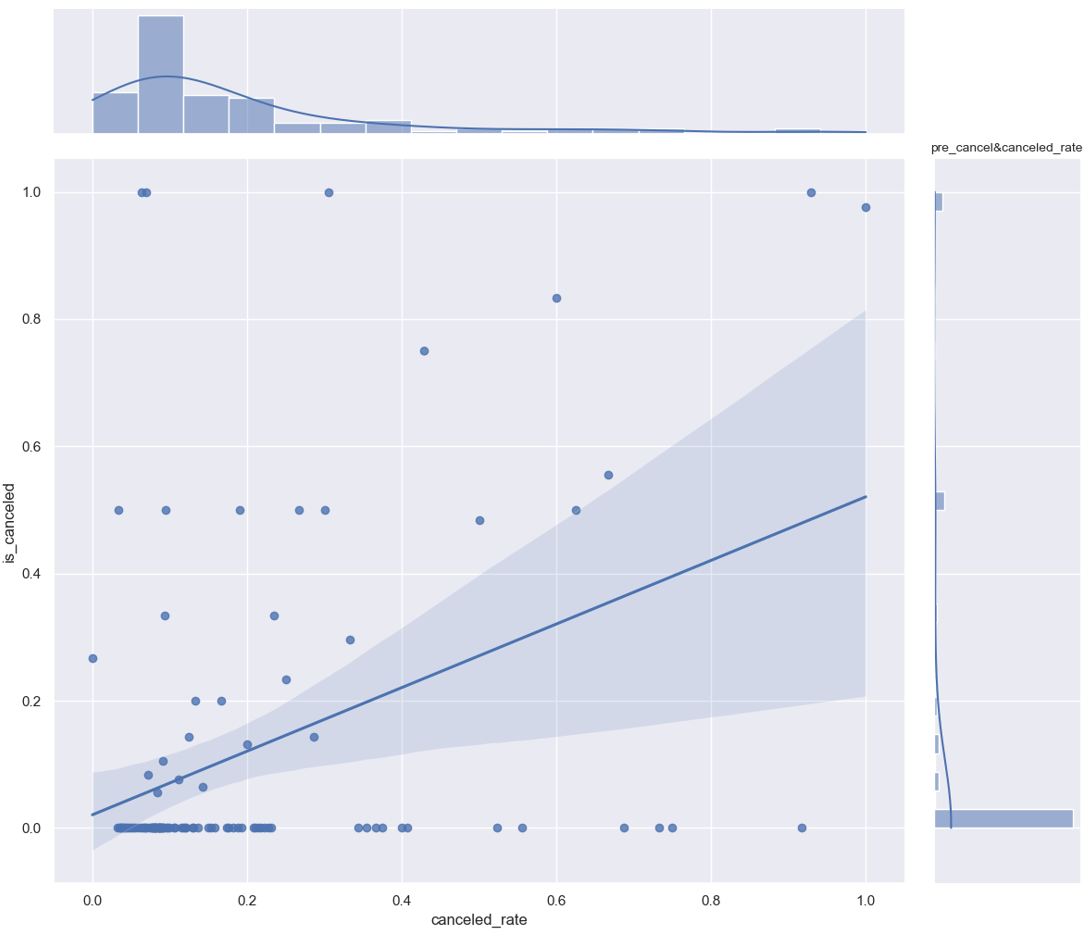
****

****

**就预定而言，不管是线上还是线下，旅行社都是人们的首选，总体站50%以上。**

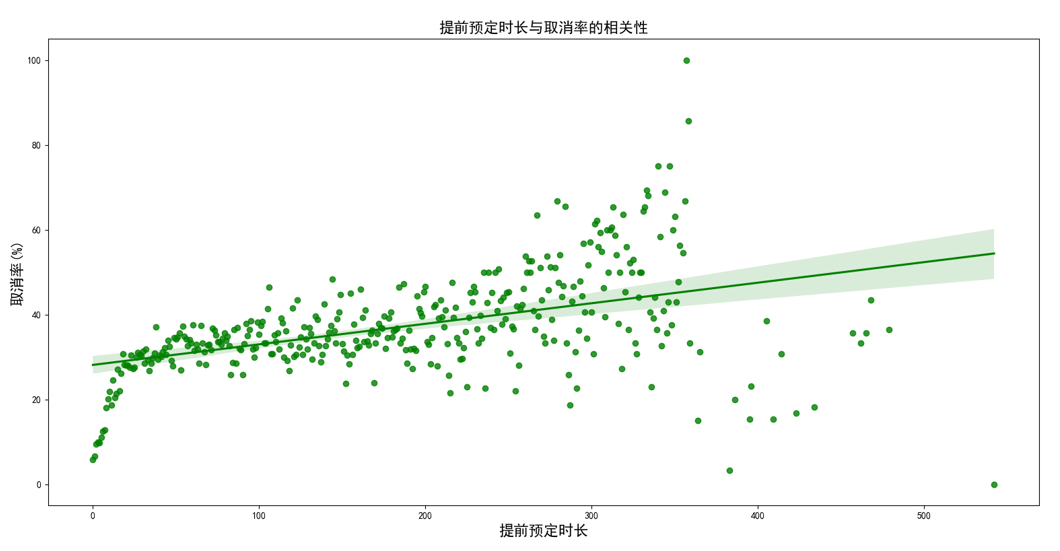
**就取消而言，线上TA比例最高，其次是线下TA/TO，最后是团体预定。**

**3.2.5.2 历史取消订单量**

****

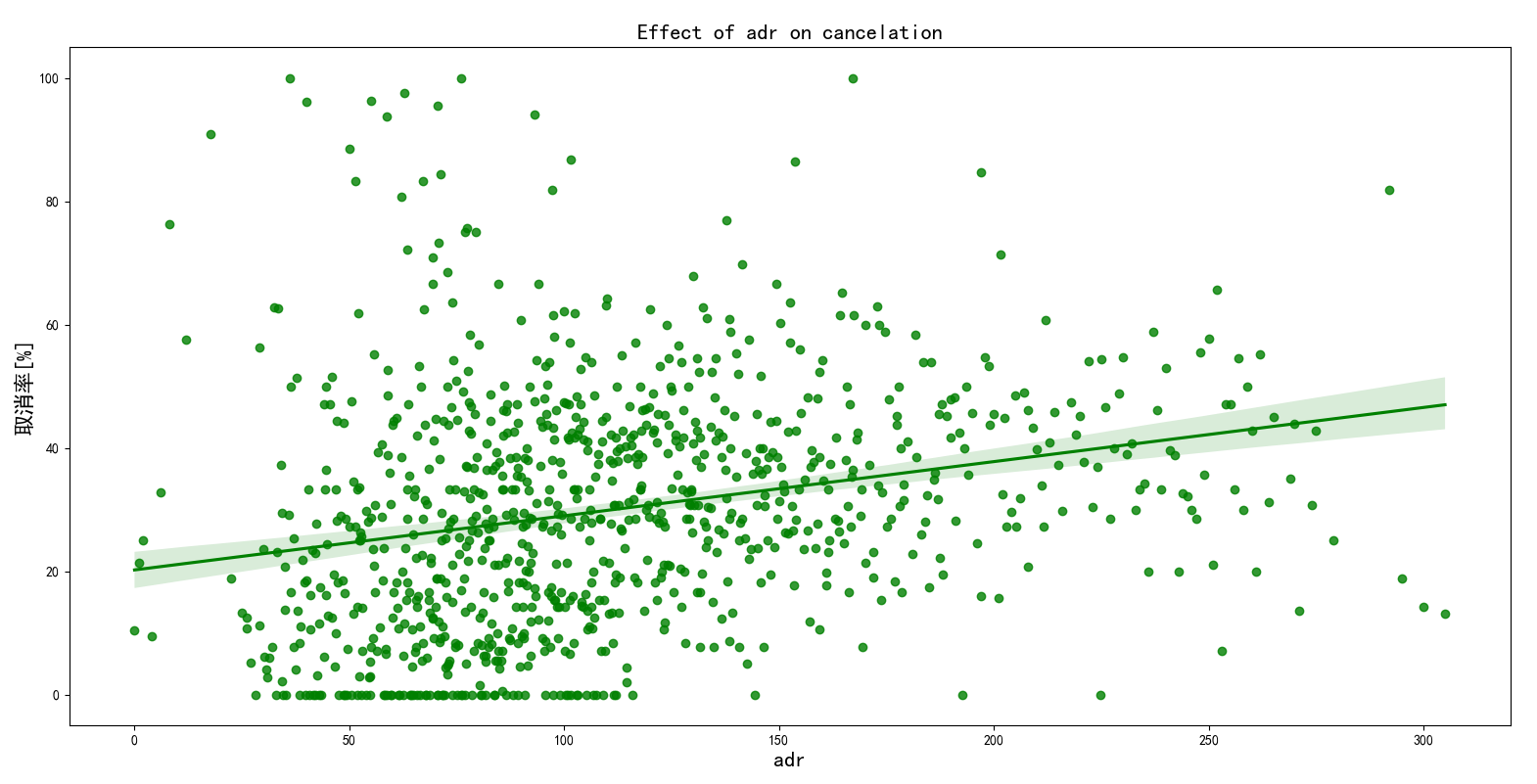
**由散点图和拟合的曲线可以看出顾客在此次订单之前，取消过的订单越多，越有可能取消本次订单。**

**3.2.5.3 提前预定时长**

****

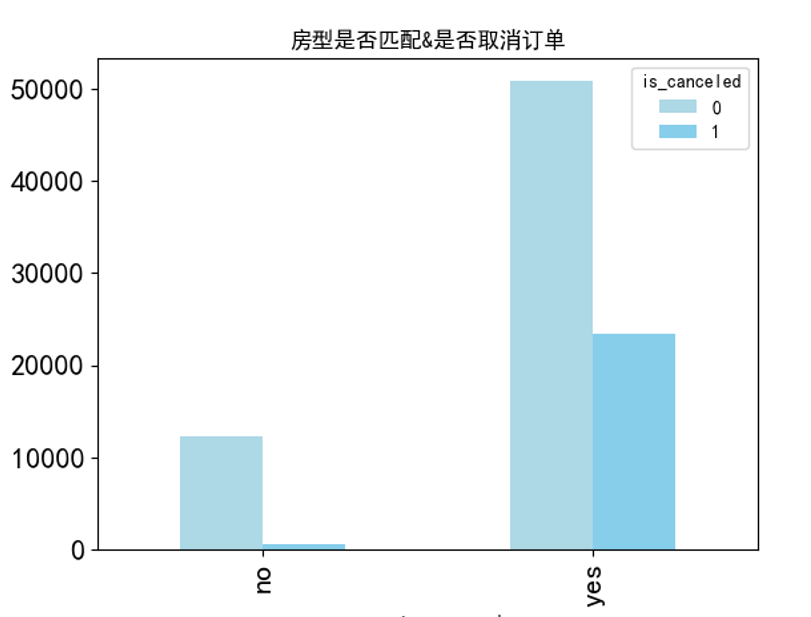
**由散点图和拟合的回归线可以看出，顾客提前预定的时间越长，计划改变的概率越高。**

**3.2.5.4 平均每日房价**

****

**每日房价越低，开支越小，取消订单的代价越小，取消率也越高。**

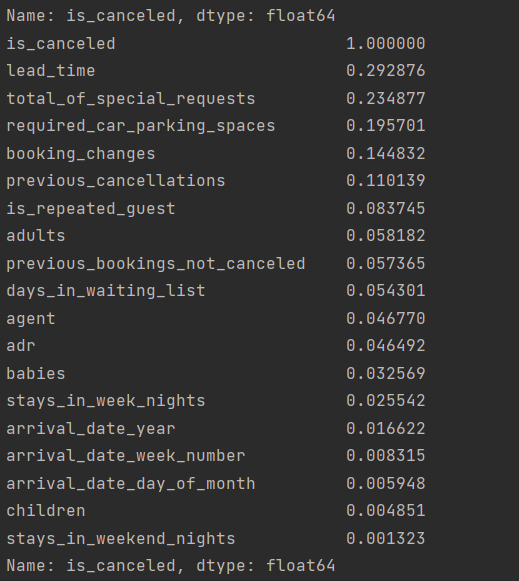
**3.2.5.5 房型匹配率**

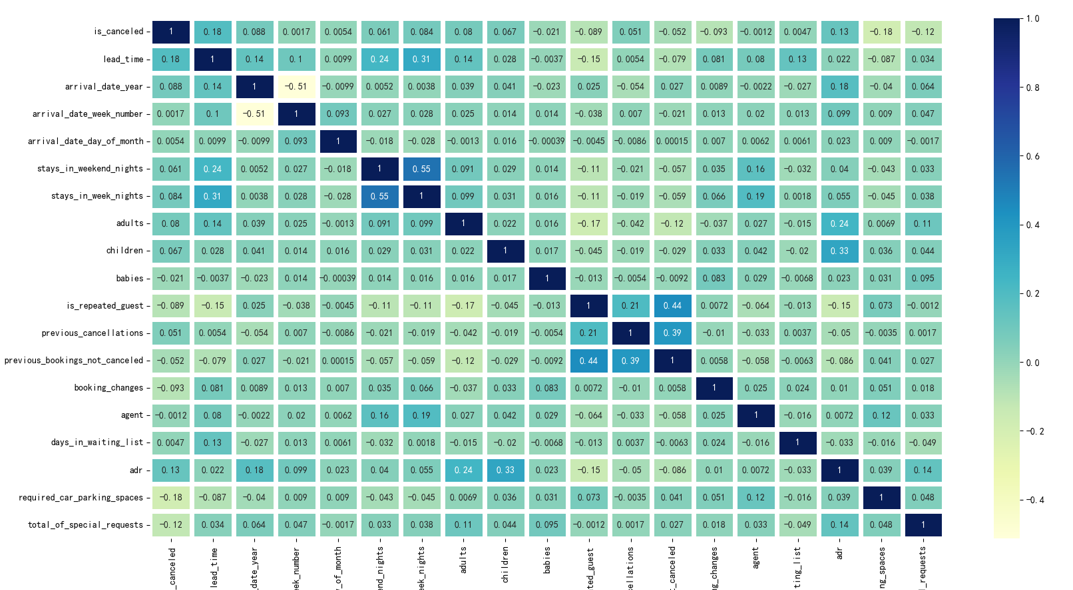
****

**数据集中有reserved\_room\_type和assigned\_room\_type两个特征，即用户下单时选择的房型和届时入住的房型有可能不一样，但是研究结果显示：房型不匹配的用户反而不倾向于取消订单。对此猜测的原因可能有：1、房型不一致属于酒店方房型升级的结果。2、因旺季酒店房源供不应求所以大众倾向于妥协。**

**3.2.6 相关性分析**

**查看每一个特征与is\_canceled的相关系数**

****

****

**与是否取消订单相关性top5高的特征分别是：提前预定的时长、特殊要求总量、是否需要停车位、对预定进行的更改、历史取消量。**

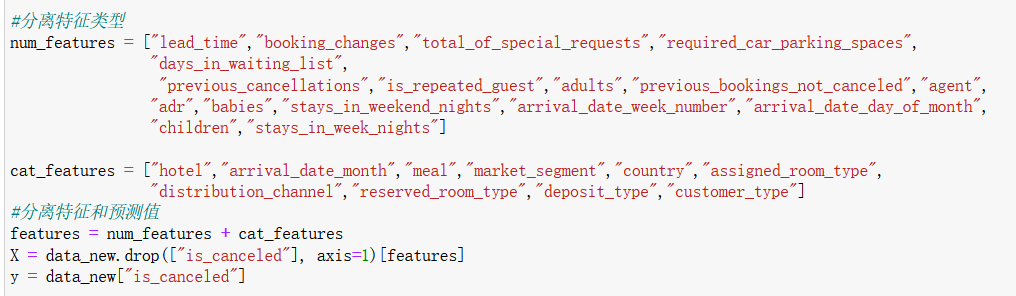
**3.3 取消率预测**

**预测顾客是否会取消订单时我试用了六个机器学习的base model，大体思路是将六个base model都尝试一遍，并选取表现最优的进行模型优化。我选取的base有：决策树、随机森林、逻辑回归、XGBoost、KNN和AdaBoost。**

**3.3.1 手动选取特征**

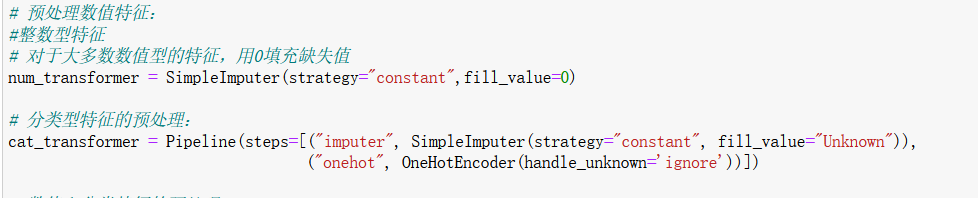
**此处删除所有时间类特征，arrival\_date\_year、arrival\_date\_month、arrival\_date\_week\_number。**

**3.3.2特征分类**

****

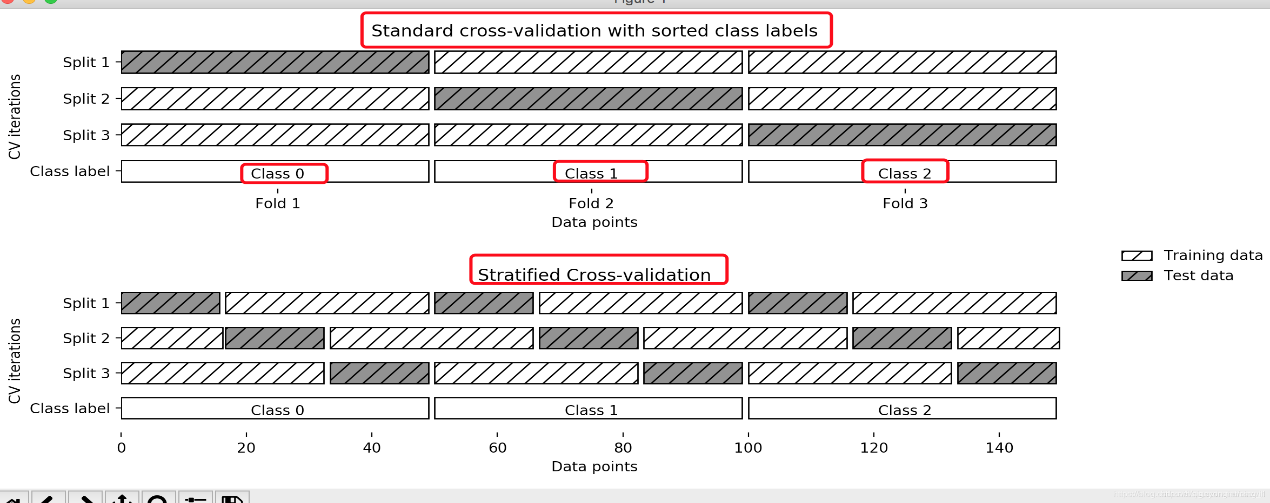
**数据集中有整数型特征和分类型特征，为了方便特征预处理，将两种特征分离到num\_features和cat\_features中。**

**3.3.3特征预处理**

****

**对于整数型特征，用对于最大多数特征都符合逻辑的0进行填充；对于分类型特征，用OneHotEncoder将其数字化，方便后续评分。**

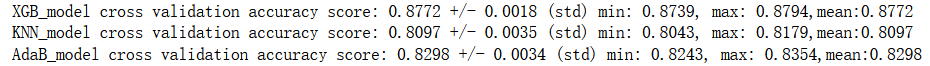
**3.3.4 选取评估模型的统计方法**

****

**选取分层K-Fold交叉验证方法，其优点是对于不平衡的数据集比较友好，该方法可以将数据集先分成k折（实验中选择了十折，在用train\_test\_split函数划分模型的训练集和测试集时，通常会用四折，也就是75%的数据作为训练集，25%的数据作为测试集；在五折交叉验证的时候，是80%的数据作为训练集，20%的数据作为测试集；十折交叉验证的时候，是90%的数据作为训练集，10%的数据作为测试集。显然，训练集的数据越多，模型的accuracy得分就越高。但是十折交叉运算也意味着六个模型都要训练十次，增加十倍的训练时间），然后再从每个k折中选择k折作为测试集合，这样就可以在口中程度上保证每次测试集合和训练集和都有各个类型的样本，特别是在当数据分类的类型数量相差悬殊的情况下，而本次实验所用的数据集有样本分布不均匀的问题，客户取消和未取消的比例约为3:7，分层k折交叉验证可以使随机划分测试集时不存在一个测试集中全部都是未取消顾客的情况。缺点是在该方法中样本是随机选取的，所以不适合时间序列型数据，因此在手动选取特征值的阶段删去了所有时间型特征，一是为了让模型更加通用，不受限于2015-2017的时间范围，二是为了与评估方法更加适配。**

**3.3.5 模型评分展示**





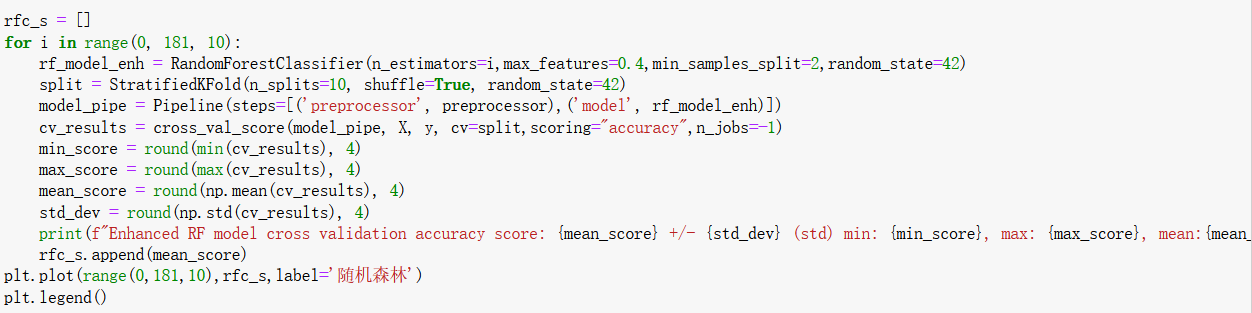
**可以看出随机森林是表现最好的模型，XGBoost次之，表现最差的是KNN。**

**对KNN表现不好的猜测：在样本不平衡的时候，对稀有类别的预测准确率低，比如一个类的样本容量很大，输入一个样本的时候，K个临近值中大多数都是大样本容量的那个类。假设实验中取消订单的顾客占比大约百分之30，即使KNN得分在80左右，也有可能把所有顾客都预测成了未取消订单的顾客；相比起决策树等模型，KNN模型的可解释性也不强，是慵懒学习法，基本上不学习，导致实验时的运行速度相对于其他算法更慢。**

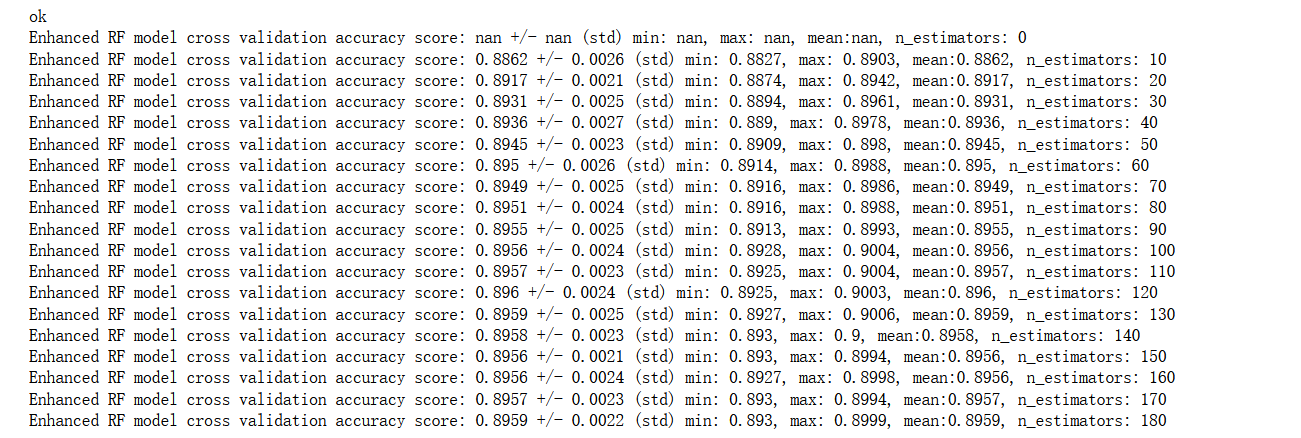
**对随机森林表现好的猜测：对多元共线性不敏感，对非平衡的数据比较稳健。通过bootstrap重采样技术，可以对占比高的样本降采样，也可以对占比少的样本过采样，以此弱化样本不平衡的问题。**

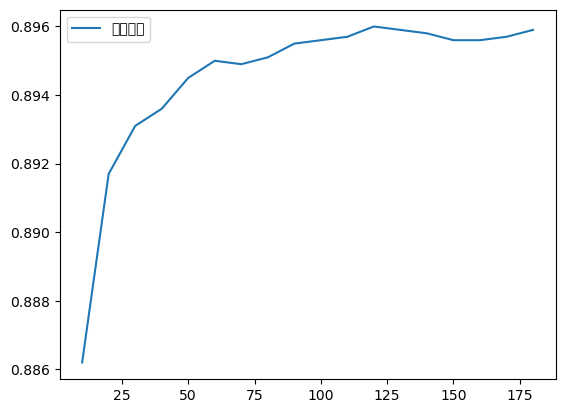
**4、 模型优化**

**选取表现最优的随机森林模型，并对其进行n\_estimators参数调优。**

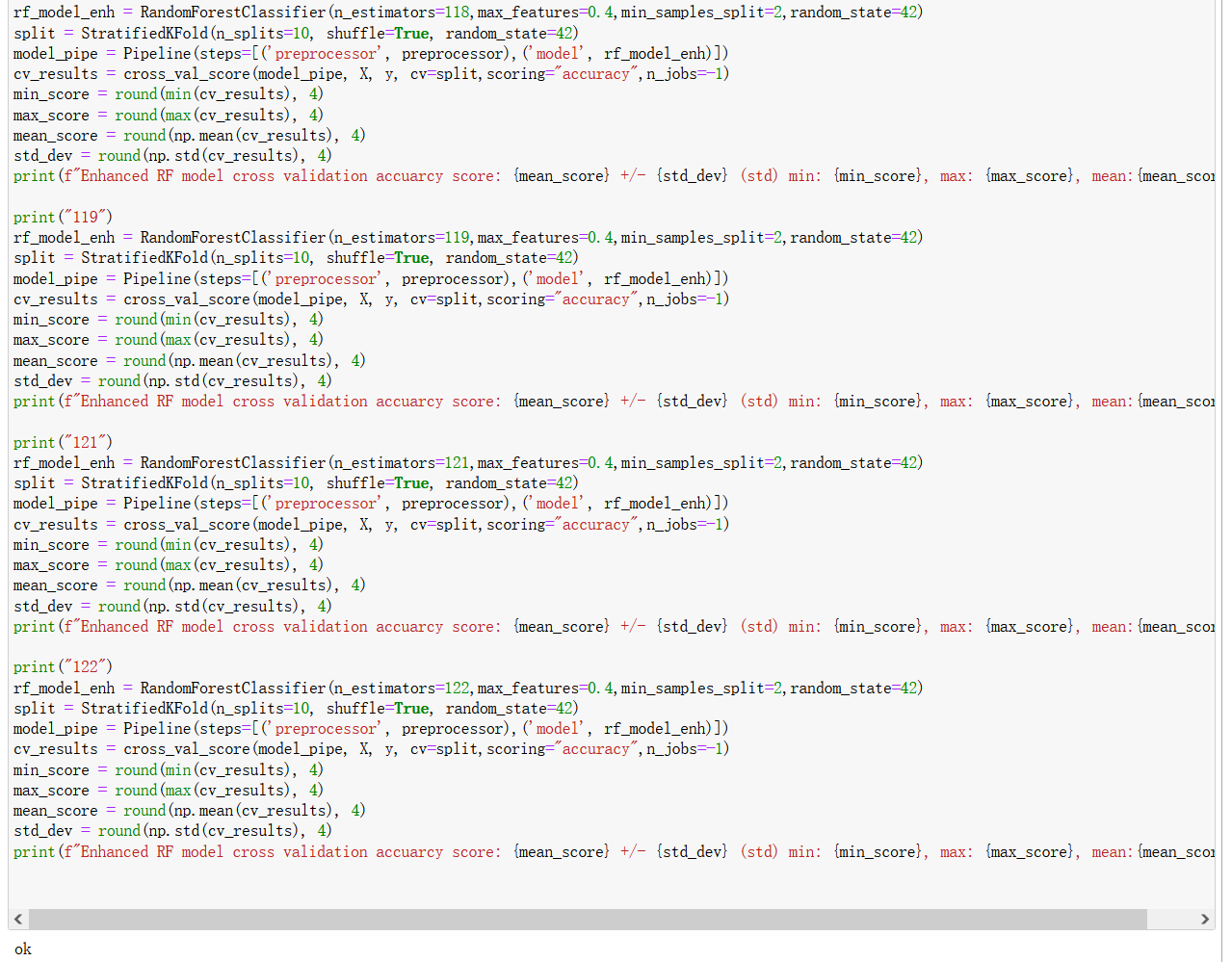


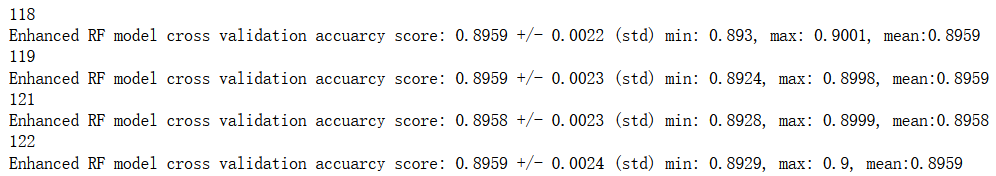
**将范围设置在0-180之间，步长为10，并绘制学习曲线。**

****

****

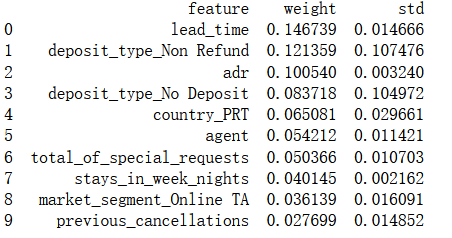
**通过学习曲线可以看出：模型在120左右的参数区间内表现最好，于是可以根据该图缩小参数范围，本次实验尝试118、119、121、122这些参数，结果显示模型在n\_estimators=120的时候表现最优。**

****



**5、特征评分**





**可以看出权重最大的三个特征是提前预定的时长、押金类型（不退回）和人均价格。其中提前预定的时长和人均价格都在前文的探索性分析中被证明与是否取消订单高度相关。**