西南财经大学

Southwestern University of Finance and Economics

商务大数据分析与决策

课程报告

学年学期：2021-2022 学年第二学期

课程名称：商务大数据分析与决策

学生学号：41951020

学生姓名：曾诚

学　　院：工商管理学院

年级专业：2019 级市场营销（金融服务与营销）

基于机器学习技术的酒店客户预订取消数据分析与预测

摘要：本项目使用了来自 Kaggle 网站上的 Hotel booking demand 项目数据集，该数据集包含了一家城市酒店和一家度假酒店的预订信息。本项目使用 Python 语言，通过机器学习等数据分析技术，首先对数据进行了描述性的统计，完成了对数据的预处理；其次利用数据集对酒店运营状况、市场情况、客户画像进行了可视化分析；最后根据数据集建立客户是否会取消预订的预测模型。

关键词：酒店管理　预订取消预测　机器学习

一、引言

大数据时代下，酒店旅游市场的快速发展使得行业内的竞争愈发激烈，客户对产品与服务信息获取渠道越来越多，酒店企业面临着产品同质化严重、同行竞争加剧，从而导致新增客户获取困难、成本增加等行业问题。同时，酒店市场客源较不稳定，客户流失量较高。因此维护好留存客户，保持客户忠诚度，降低客户流失率是当前面临的主要问题。

数据挖掘在这个实际课题中发挥了很大的作用，数据挖掘就是从大量杂乱无章的原始数据中得出一些有用的信息。在通信行业、银行等一些机构存在大量的数据。数据处理人员从大量的数据中进行分类、整理、归纳、总结,对客户进行分析、整理,和现在计算机、数据库、数据仓库普遍应用相比,过去那种分析、整理、归纳的方法都是和那个时代相匹配的，都是很简单的归纳，只能得到很少的信息，并且都是以数学理论为起点,然后向实际推广。随着计算机的大规模应用，人们可以在大量的数据中，运用数据挖掘技术寻找数据中蕴藏的规律，并对未来。

酒店行业的从业者为了更好地规划酒店的经营，他们有极强的意愿来了解消费者的行为特点，例如在消费者的视角里，什么时候是一年中预定酒店房间的最佳时间，为了获得最好的房价折扣而选择的最佳入住时间，酒店是否可能会收到不成比例的高数量的特殊请求。客户流失预测可以帮助酒店预测流失趋势，构建适合酒店行业数据特性的影响因素指标体系，有针对性地提出挽留措施，提高商家利润，因此，酒店业的客户流失预测是酒店管理领域的一个重要研究方向。

本研究使用的数据来自 Nuno Antonio 等人的数据文章 Hotel booking demand datasets.，该文章描述了两个包含酒店需求数据的数据集。其中一家酒店(H1)是度假酒店，另一家是城市酒店(H2)。两个数据集共享相同的结构，通过 31 个特征，形成了描述 H1 的 40060 次观测数据和 H2 的 79330 次观测数据，每一条观测数据都代表一次酒店预订。这两个数据集包含了 2015 年 7 月 1 日至 2017 年 8 月 31 日之间的预订，包括有效到达的预订和被取消的预订。

二、数据预处理与描述型分析

2.1 数据预处理

缺失值处理。发现一共有 4 个特征出现了缺失值，由于 children 和 country 缺失值的数量占总体比例非常小，且这两列是类别型变量，我们直接用对应列的众数进行填充；agent

列的缺失值数量较多，故将缺失值单独作为新类别，标记为 0；company 列几乎全为缺失值，包含了极少量的有效值，因此删除此列。

异常值处理。通过在 Kaggle 网站上的数据集说明，仔细观察数据可以发现，存在入住总人数为 0 和入住总天数为 0 的数据，即异常数据，我们需要对这些数据做筛选和清理。此外，数据集说明里，我们还可以发现 meal 列的 Undefined/SC 均表示未预定餐食，我们需要把其合并为同一类。

2.2 可视化分析

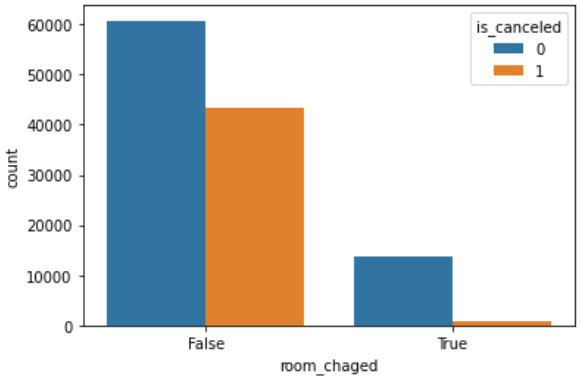

图 1 房间类型变更对取消预订的影响

如图 1 所示，房型变更过的客户取消预定的概率远远小于未变更过的客户。对此，可能有以下原因：客户到达酒店后临时更改房型，多数客户会选择不取消预定，直接入住；客户自行更改房型，相对取消预定而言，这类客户更愿意更改房间类型而保证正常入住。

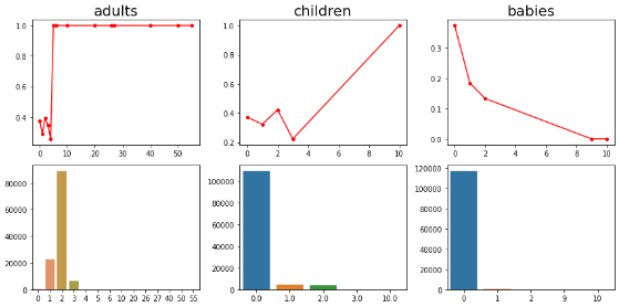

图 2 入住人数对取消预订的影响

如图 2 所示，多数预定订单显示没有儿童和婴儿入住，其中单人入住和双人入住是主要的预定人数模式；有婴儿入住时预定取消率大幅下降；超过 5 人以上入住的订单基本全部取消，这部分可能是刷单等异常订单，酒店需要注意。针对不同酒店，重点分析下列几种入住人数情况的取消率：单人入住：adults=1, children, babies=0；双人入住：adults=2, children, babies=0；家庭入住：adults>2, children, babies>0。

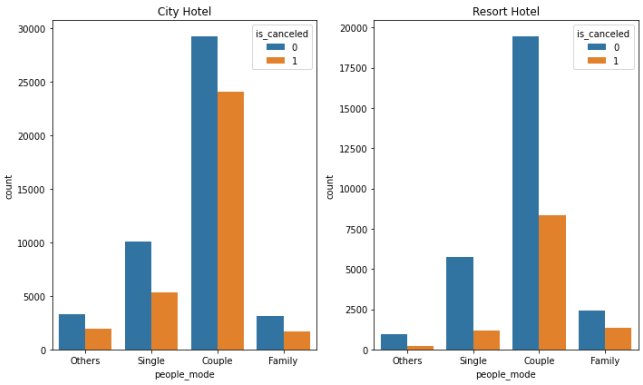

图 3 房间类型、入住人数对取消预订的影响

对于城市酒店，取消预定概率：双人>>单人≈家庭，应注意双人入住客户的高取消率现象，改善酒店对于双人入住客户的配套服务以降低取消率。对于度假酒店，取消预订概率：家庭>双人>单人，酒店可适当针对家庭客户提供相应的优惠折扣，提高家庭客户入住率。

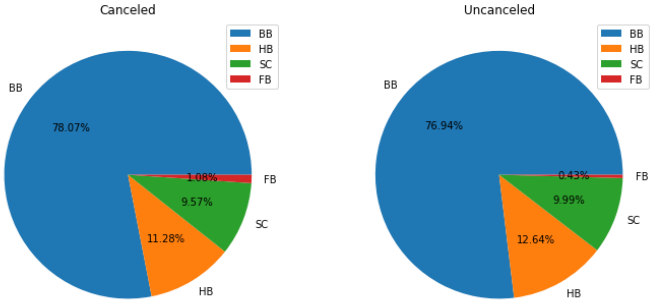

图 4 是否取消预订和饮食的影响

可以看到无论是否取消预订，餐食类型之间差异不大。

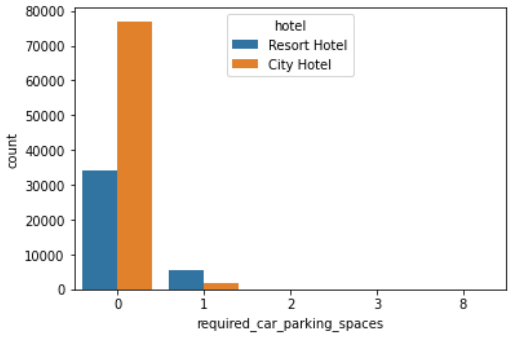

图 5 不同酒店类型和车位需求

多数客户不需要停车位，相比之下，度假酒店客户需要停车位的比例远大于城市酒店。

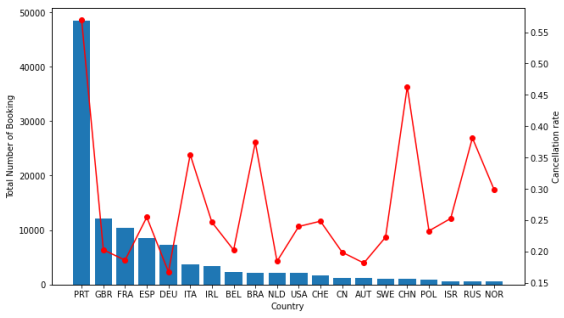

图 6 预定数前 20 的国家/地区

统计可知，前 20 名国家/地区数据量占据全部数据的 94%，客户主要来自葡萄牙，英国，法国，西班牙等欧洲国家，不同国家之间预定取消率的差距非常显著，取消率较高的国家有葡萄牙、意大利、巴西、中国、俄罗斯，以发展中国家为主。

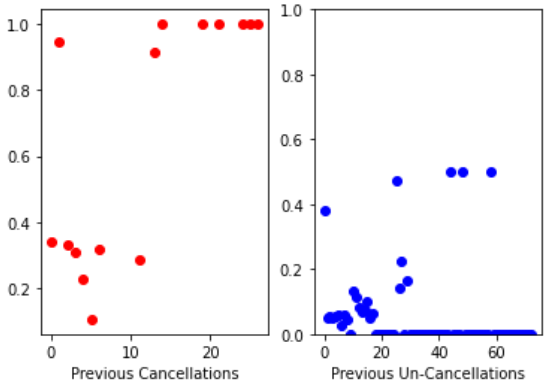

图 7 客户预定历史

客户预定历史指客户之前预定过的订单的取消情况，可以一定程度上反映客户当前订单的取消意愿。大多数预定来自于新客，而熟客取消预定的概率远远小于新客；先前取消过预定的客户本次预定取消的概率较大，尤其是取消过预定 15 次以上的客户，基本上不会选择入住，可以计入酒店的“黑名单”；先前预定并入住过的客户相对来说信用较好，高入住次数（>20 次）客户基本不会取消预订。

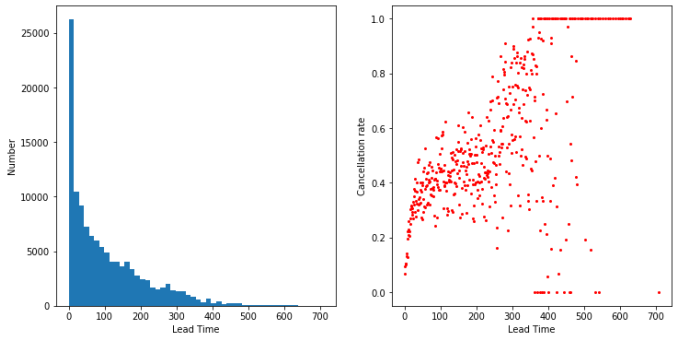

图 x 提前预订时长分布

从预定提前时长分布明显可以看出，客户倾向于选择与入住时间相近的时间预定，并且随着预定提前时长的增大，取消率呈现上升趋势。

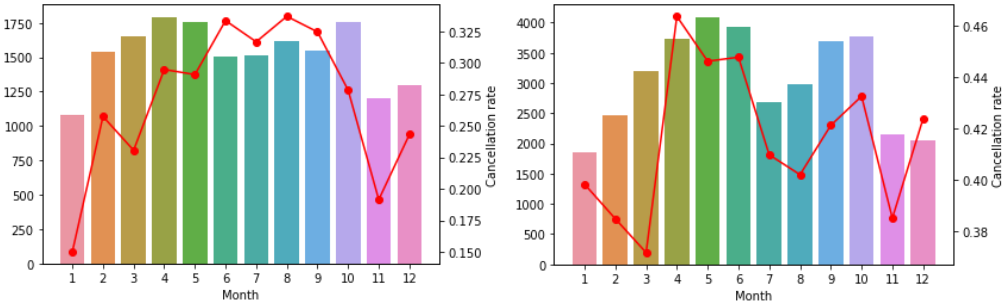

图 8 入住时间

预定量上，城市酒店 7/8 月出现大幅下滑，同期度假酒店变化较小，整体而言，度假酒店月度客流量变化较小；取消率上，两家酒店冬季取消率相对较低，城市酒店夏季取消率降

低，度假酒店却处于高峰。

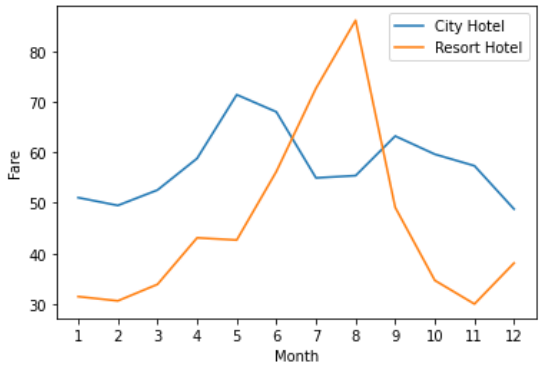

图 9 客房价格趋势变化

两种酒店客房价格趋势主要差异在于 7-8 月，此时期城市酒店价格小幅下跌，度假酒店价格却急速上涨，一度超过城市酒店。结合预定量和取消量分析，7-8 月度假酒店客流减少，取消率大幅上升，经营者应考虑调整价格策略以增加营收。对于用户而言，应考虑避免 8 月预定度假酒店，此时酒店价格处于高位，而 9 月价格便会下跌近一半，气候/环境差异不大，是入住的好时期。

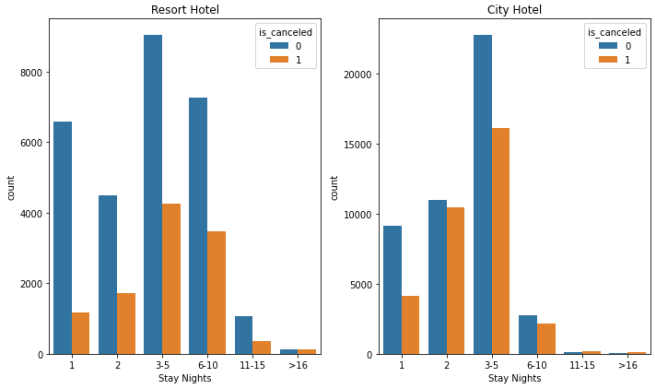

图 10 客房价格趋势变化

度假酒店客户入住时长集中在 1-10 晚，其中入住 1 晚的客户取消概率最低；城市酒店客户入住时长多在 5 晚以内，其中入住 2 晚的客户取消概率最高；整体而言，度假酒店客户平均入住天数明显高于城市酒店，可以考虑推出长租优惠方案吸引顾客。

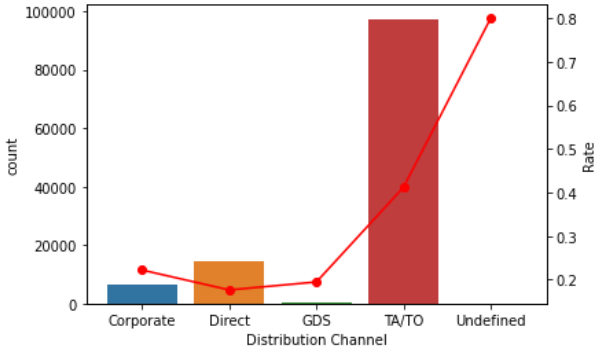

图 11 预订渠道

预定主要来自于旅行社(TA/TO)，个人直接预定(Direct)和团体预定(Group)；旅行社取消预定的概率远大于其他渠道，可能是由于旅行社出于盈利考虑会取消利润较低的订单。

三、模型的构建与评价

3.1 简单的特征工程

首先进行特征的筛选，模型目标是对订单是否取消进行预测，挑选特征变量应满足： 特征必须为客户预定时就能获得的数据， 因此排除 booking_changes （预定更改），reservation_status（结账状态），assigned_room_type （最终分配房型）等特征。考虑一定的信息脱敏和通用性，排除 country（国籍），arrival_date_year（入住年份）等特征。完成数据列处理后，先查看一下各数据参数与取消预订之间的相关性。

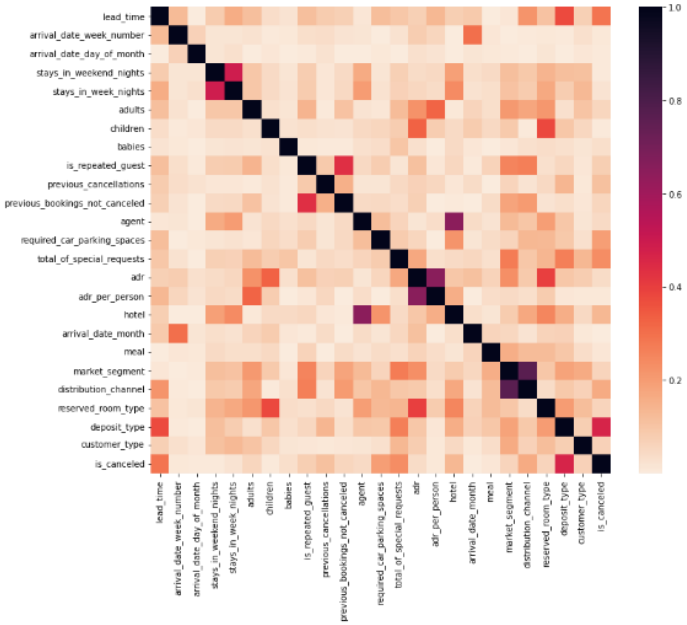

图 12 相关性矩阵

与取消预订相关性较高的特征有：提前预定时长，车位需求数，特殊需求数，押金类型，酒店类型等。

3.2 简单机器学习模型

接下来，进行模型训练，引入多个简单的分类模型，均采用默认参数，简单比较模型之间的准确率，表 1 展示了所选用的部分简单模型的效果。

表 1 简单模型的泛化效果比较

模型	验证集上的准确率
决策树	**0.824**
K 近邻	0.812
线性支持向量机	0.785
随机梯度下降	0.766
岭回归	0.765
感知机	0.755
高斯朴素贝叶斯	0.491

不难发现，决策树、K 近邻方法的泛化能力和预测准确率明显高于其他简单模型。虽然如此，简单模型的预测正确率仍然不算非常高，本文考虑使用一些更复杂的模型进行处理。

3.3 较复杂的模型

本文使用了 Logistic Regression，Random Forest，XGBoost，LightGBM，CatBoost 五个

模型进行训练，鉴于篇幅原因，本文不再对模型的原理进行解释，仅展示训练结果，模型的构建请参见附录代码。

表 2 较复杂的模型泛化效果评估

模型	AUC	Accuracy	Recall	F1-Score	Precision
Logistic Regression	0.721333	0.779618	0.484670	0.623722	0.874663
Random Forest	**0.843414**	**0.864382**	0.761137	**0.807028**	**0.858807**
XGBoost	0.808278	0.787040	**0.894517**	0.759958	0.660588
LightGBM	0.813518	0.842733	0.694890	0.769072	0.860985
CatBoost	0.838313	0.855777	0.767400	0.800420	0.836409

基于以上指标，我们可以对不同的模型进行有效的评估，我们发现 5 个指标中，随机森林模型有四个都是最高的，因此随机森林模型可能是最适合进行本项目预测分析的模型。

3.4 应用随机森林模型进行预订取消预测的示例

```
from sklearn import metrics
#选择待预测样本
pre_x=X[1:2,]
#相应的真实值
true_y=y[2]
#预测结果
pre_y=clf.predict(pre_x)
print(true_y)
print(pre_y)
```

```
0
[0]
```

发现随机森林模型的预测是正确的，观察随机森林模型的特征重要度，如表 3。

表 3 随机森林模型的特征重要度（前 7）

特征	重要度
lead_time	0.15
deposit_type	0.13
adr_per_person	0.08
adr	0.08
arrival_date_day_of_month	0.07
total_of_special_requests	0.07
arrival_date_week_number	0.06

四、结论与应用

本项目使用了来自 Kaggle 网站上的 Hotel booking demand 项目数据集，该数据集包含了一家城市酒店和一家度假酒店的预订信息。本项目使用 Python 语言，通过机器学习等数据分析技术，首先对数据进行了描述性的统计，完成了对数据的预处理；其次利用数据集对酒店运营状况、市场情况、客户画像进行了可视化分析；最后根据数据集建立客户是否会取消预订的预测模型。

本文基于机器学习的评价指标，发现随机森林模型能有效预测客户是否取消预订，其中 lead_time,deposit_type,adr_per_person,adr,arrival_date_day_of_month,total_of_special_requests,arrival_date_week_number7 个特征的重要性较大，酒店可

以考虑从这 7 个特征的角度进行服务改善和预订取消预测。

参考文献：

[1] Antonio, N., de Almeida, A., & Nunes, L. (2019). Hotel booking demand datasets. *Data in brief*, *22*, 41-49.

[2] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[3] Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert systems with applications*, *29*(2), 472-484.

[4] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[5] Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Arfi, W. B. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change*, *166*, 120658.

[6] Noble, W. S. (2006). What is a support vector machine?. *Nature biotechnology*, *24*(12), 1565-1567.

[7] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

[8] 赖院根, & 刘砺利. (2011). 基于客户价值的信息用户流失预测研究. *情报理论与实践*, *34*(7), 67-70.

[9] 王若佳, 严承希, 郭凤英, & 王继民. (2022). 基于用户画像的在线健康社区用户流失预测研究. *数据分析与知识发现*, *6*(2/3), 80-92.

[10] 钱苏丽, 何建敏, & 王纯麟. (2007). 基于改进支持向量机的电信客户流失预测模型. *管理科学*, *20*(1), 54-58.

[11] 周艳聪, & 郝园媛. (2022). 基于机器学习的运营商客户行为分析. *科学技术与工程*.

附录（含代码）