# 机器学习与商业数据挖掘——期末报告

**1901211508 林芯羽**

## 1 摘要

本报告旨在通过酒店预定的相关信息比如每晚平均房价、住宿日期、房型及顾客类型等变量构建该预定是否取消的预测模型，为酒店的收益管理提供可参考之依据。为此，本报告首先提取 27 个描述酒店预定的指标，通过数据训练逻辑回归模型、决策树模型、随机森林模型、XGBoost 模型以及多层感知器神经网络模型。根据各个模型的外样本预测结果显示，XGBoost 模型整体展现的预测精度最佳，以此为基础，本报告亦对于影响模型结果的重要特征进行了延伸讨论。

## 2 背景介绍

随着全球化的发展以及经济水平的提升，休闲度假、商务出差等酒店需求日趋增长，为酒店行业带来了很多的挑战和机会。为了最大化酒店总收益，收益管理是所有酒店业者都需面临的问题，除了房价的动态管理之外，有效控制客房的空房率亦是酒店运营管理工作的重中之重。

对于酒店来说，空置的房间就像是服饰店积压的库存，空耗成本却没有收益。然而，酒店不像服饰店能够将今天未卖完的库存放置在仓库中，在接下来的几天继续贩售；酒店的库存有时间限制，今天卖不掉就过期作废，没办法把今日的空房存到明日。因此，酒店“清理剩余库存”的需求更加紧急和迫切。

为了控制空房率，酒店经常会与旅行团、企业及线上订房平台合作。一般来说，旅行团的预定取消率相对较低，即使取消酒店也可以根据协议收取相关的费用；商务散客则是行程较不固定，取消情况往往取决于当次出差是否成行；而以线上订房来说，消费者容易因为个人行程变动或是找到价格更实惠的其他酒店便取消订房，取消率相对较高，由于酒店线上平台可能会采用免取消费用的方式吸引顾客，且虽然在预定时客人会告知信用卡卡号，但是在诸多不利因素比如虚报卡号、客人投诉、客人拒付的情况下，当预定取消时酒店在很多时候是收不到当晚的房费的。

在这个情况下，准确预测酒店预定取消情况便变得至关重要，若是能够提前了解预定的取消情况，酒店便能事先做出应对措施，像是采取超额预定的策略，使得饭店能将顾客临时

取消住宿造成的损失降到最低。因此，本报告将通过 Antonio 等 [1] 构建的酒店需求数据集，主要使用 XGBoost 模型 [2] 以及多层感知器神经网络模型研究每晚平均房价、住宿日期、房型、顾客类型等多个相关变量对预定是否取消的影响，并以逻辑回归模型为基准，决策树模型及随机森林模型作为对比模型，比较不同模型预测能力的差异，同时尝试探讨哪些指标是影响预测结果的重要因素，为酒店预定取消情况的预测提供理论模型。

# 3 方法

## 3.1 数据集准备

本报告使用的数据集为 Antonio 等 [1] 发表的酒店需求数据集，该数据集源自于一家度假酒店以及一家城市酒店从 2015 年 7 月 1 号到 2017 年 8 月 31 号的住房预定信息。原始数据集的变量包含了订单最后的状态以及订单最后状态日期，由于从订单最后的状态便能知道该笔订单是否取消，因此将这两个变量予以去除，同时针对其他变量进行缺失和异常值检测、合并部分变量及构建新变量等数据预处理。

表 1为本报告最终采用之变量，总共 118,901 笔样本。模型的预测变量为该笔预定是否取消，输入变量为酒店类型、每日平均房价、总人数、代理商、抵达日期（日/月/周/年）、是否为家庭入住、预定改变次数、国家、顾客类型、于等待名单天数、是否有押金、预定渠道、是否为重复顾客、前置时间、市场划分、餐点、过往预定未取消次数、过往预定取消次数、要求停车格数量、预定房型、分配房型、周末住宿天数、平日住宿天数及特殊要求总数。分类变量的处理方面，因为国家以及代理商数量众多，采用了 Label Encoding 的方式将其编码成数值；其他分类变量则是通过虚拟变量的设置进行处理。

**表 1: 数据变量说明表**

| 变量名 | 数据类型 | 详细说明 |
|---|---|---|
| 是否取消（**IsCanceled**） | 分类变量 | 1 为取消，0 为未取消 |
| 酒店类型（**Hotel**） | 分类变量 | 酒店类型：度假酒店（Resort Hotel）、城市酒店（City Hotel） |
| 每日平均房价（**ADR**） | 数值型变量 | 房价总额/住宿天数 |
| 总人数(**TotalCustomers**) | 数值型变量 | 婴儿数量 + 儿童数量 + 成人数量 |
| 代理商（**Agent**） | 分类变量 | 旅行社 ID |
| 抵达日期-日（**ArrivalDateDayOfMonth**） | 数值型变量 | 抵达日期中的日 |
| 抵达日期-月（**ArrivalDateMonth**） | 分类变量 | 抵达日期中的月，以 January～December 表示 |

表 1 （接续前页）

| 变量名 | 数据类型 | 详细说明 |
| --- | --- | --- |
| 抵达日期-周（**ArrivalDateWeekNumber**） | 数值型变量 | 抵达日期中的周 |
| 抵达日期-年（**ArrivalDateYear**） | 分类变量 | 抵达日期中的年 |
| 是否为家庭入住（**IsFamily**） | 分类变量 | 1 为是家庭入住（成人与儿童/婴儿同行），0 为非家庭入住 |
| 预定改变次数（**BookingChanges**） | 数值型变量 | 到实际入住或是取消之前的预定改变次数 |
| 国家（**Country**） | 分类变量 | 来自国家，以 ISO 3155—3:2013 格式表示 |
| 顾客类型（**CustomerType**） | 分类变量 | 顾客类型：Contract（合同旅客）、Group（团体旅客）、Transient（与其他预定无关的短期住宿旅客）、Transient-party（与其他预定相关的短期住宿旅客） |
| 于等待名单天数（**DaysInWaitingList**） | 数值型变量 | 确认预定前于等待名单上的天数 |
| 是否有押金（**HasDeposit**） | 分类变量 | 1 表示预定需要押金，0 表示不需押金 |
| 预定渠道（**DistributionChannel**） | 分类变量 | 预定分销渠道：旅游代理商（TA, Travel Agents）、旅游经营商（TO, Tour Operators） |
| 是否为重复顾客（**IsRepeatedGuest**） | 分类变量 | 1 表示是重复旅客，0 表示非重复旅客 |
| 前置时间（**LeadTime**） | 数值型变量 | 预定日期与抵达日期相隔天数 |
| 市场划分（**MarketSegment**） | 分类变量 | 市场细分类型：直接（Direct）、企业（Corporate）、线上旅游代理商（Online TA）、线下旅游代理商/经营商（Offline TA/TO）、辅助（Complementary）、团体（Groups）、航空（Aviation）、未定义（Undefined） |
| 餐点（**Meal**） | 分类变量 | 预定餐点类型：Undefined/SC（没有附餐点）、BB（附早餐）、HB（附早餐以及其他一餐）、FB（附早、午、晚餐） |

表 1 （接续前页）

| 变量名 | 数据类型 | 详细说明 |
| --- | --- | --- |
| 过往预定未取消次数（**PreviousBookingsNotCanceled**） | 数值型变量 | 在本次预定之前预定未取消次数 |
| 过往预定取消次数（**PreviousCancellations**） | 数值型变量 | 在本次预定之前预定取消次数 |
| 要求停车格数量（**RequiredCarParkingSpaces**） | 数值型变量 | 旅客要求的停车格数量 |
| 预定房型（**ReservedRoomType**） | 分类变量 | 预定时指定的房间类型代码 |
| 分配房型（**AssignedRoomType**） | 分类变量 | 最后被分配的房间类型代码 |
| 周末住宿天数（**StaysInWeekendNights**） | 数值型变量 | 于星期六日入住酒店的天数 |
| 平日住宿天数（**StaysInWeekNights**） | 数值型变量 | 于星期一～五入住酒店的天数 |
| 特殊要求总数（**TotalOfSpecialRequests**） | 数值型变量 | 特殊要求总次数（比如双人床或是高楼层） |

## 3.2 数据基本描述

本报告接着针对数据做描述性统计以再次确认数据质量，同时初步判断预定是否取消与其他解释变量的关联。由于使用的变量个数较多，碍于报告篇幅仅针对部分变量详细说明。

### 3.2.1 数值型变量

从表 2呈现的各数值型变量样本量、最大值及最小值来看，当前数据并没有缺失值以及异常情况存在。

进一步查看过往预定取消次数、过往预定未取消次数、预定改变次数、于等待名单天数、要求停车格数量、特殊要求总数等变量的中位数以及平均值，可以发现大部分的数据值皆为 0，数据分布处于十分不均衡的状态，大部分的数据取值偏低，但样本的最大值可以很高，数据呈现右偏的情况。

另一方面，从前置时间、于等待名单天数以及每日平均房价的标准差来看，数据的波动性较大，比如每日平均房价最低可为 0 元（可能为促销活动），平均及中位数在 100 元上下，

表 2: 数值型变量描述性分析

| 变量名 | 样本数 | 平均值 | 标准差 | 最小值 | 中位数 | 最大值 |
|---|---|---|---|---|---|---|
| 前置时间（**LeadTime**） | 118901 | 104.31 | 106.90 | 0.00 | 69.00 | 737.00 |
| 抵达日期-日（**ArrivalDateDayOfMonth**） | 118901 | 27.17 | 13.59 | 1.00 | 16.00 | 53.00 |
| 抵达日期-周（**ArrivalDateWeekNumber**） | 118901 | 15.80 | 8.78 | 1.00 | 28.00 | 31.00 |
| 周末住宿天数（**StaysInWeekendNights**） | 118901 | 0.93 | 1.00 | 0.00 | 1.00 | 16.00 |
| 平日住宿天数（**StaysInWeekNights** | 118901 | 2.5 | 1.9 | 0.0 | 2.0 | 41.0 |
| 过往预定取消次数（**PreviousCancellations**） | 118901 | 0.09 | 0.85 | 0.00 | 0.00 | 26.00 |
| 过往预定未取消次数（**PreviousBookingsNotCanceled**） | 118901 | 0.13 | 1.48 | 0.00 | 0.00 | 72.00 |
| 预定改变次数（**BookingChanges**） | 118901 | 0.22 | 0.65 | 0.00 | 0.00 | 21.00 |
| 于等待名单天数（**DaysInWaitingList**） | 118901 | 2.33 | 17.63 | 0.00 | 0.00 | 391.00 |
| 每日平均房价（**ADR**） | 118901 | 102.00 | 50.49 | 0.00 | 95.00 | 5400.00 |
| 要求停车格数量（**RequiredCarParkingSpaces**） | 118901 | 0.06 | 0.24 | 0.00 | 0.00 | 8.00 |
| 特殊要求总数（**TotalOfSpecialRequests**） | 118901 | 0.57 | 0.79 | 0.00 | 0.00 | 5.00 |
| 总人数(**TotalCustomers**) | 118901 | 1.97 | 0.72 | 0.00 | 2.00 | 55.00 |

但是最高房价可以到 5400，或许是不同等级的酒店类型所导致的价格差异。

接下来，本报告进一步针对前置时间、总人数、过往预定取消次数、过往预定未取消次数、要求停车格数量、特殊要求总数、每日平均房价以及预定改变次数与预定是否取消的关系做简单分析。如同先前所述，许多变量的数据分布较为不均衡，若是使用箱型图会造成图中出现过多异常值，影响数据的可读性，因此在这里使用条形图（bar plot）呈现各个变量与是否取消的关系，如图 1所示。

从前置时间、过往预定取消次数及过往预定未取消次数来看，取消的预定中平均前置

天数以及过往预定取消次数都比未取消的预定高出许多，而未取消的预定中平均过往预定未取消次数较高，显示在预定日期相隔抵达日期较远的情况，消费者可能更容易因为种种因素改变原有计划，因此更改订单；另一方面，过往预定取消/未取消次数能一定程度上反映出消费者的行为，对于过往经常取消预定的消费者，在当次预定取消的可能性也会比较高。

以要求停车格数量、特殊要求总数以及预定改变次数的角度来看，可以看到未取消的预定其平均要求停车格数量、特殊要求总数以及预定改变次数相较取消的预定来的高，表示当消费者对于旅馆有较多要求的时候，可能较不倾向取消原有预定，因为这意味着消费者需要再耗费时间精力去寻找符合需求的其他旅馆。

从总人数以及每日平均房价的条形图来看，取消以及未取消的样本之间并没有显著的差别，可能需要通过模型进一步考察。

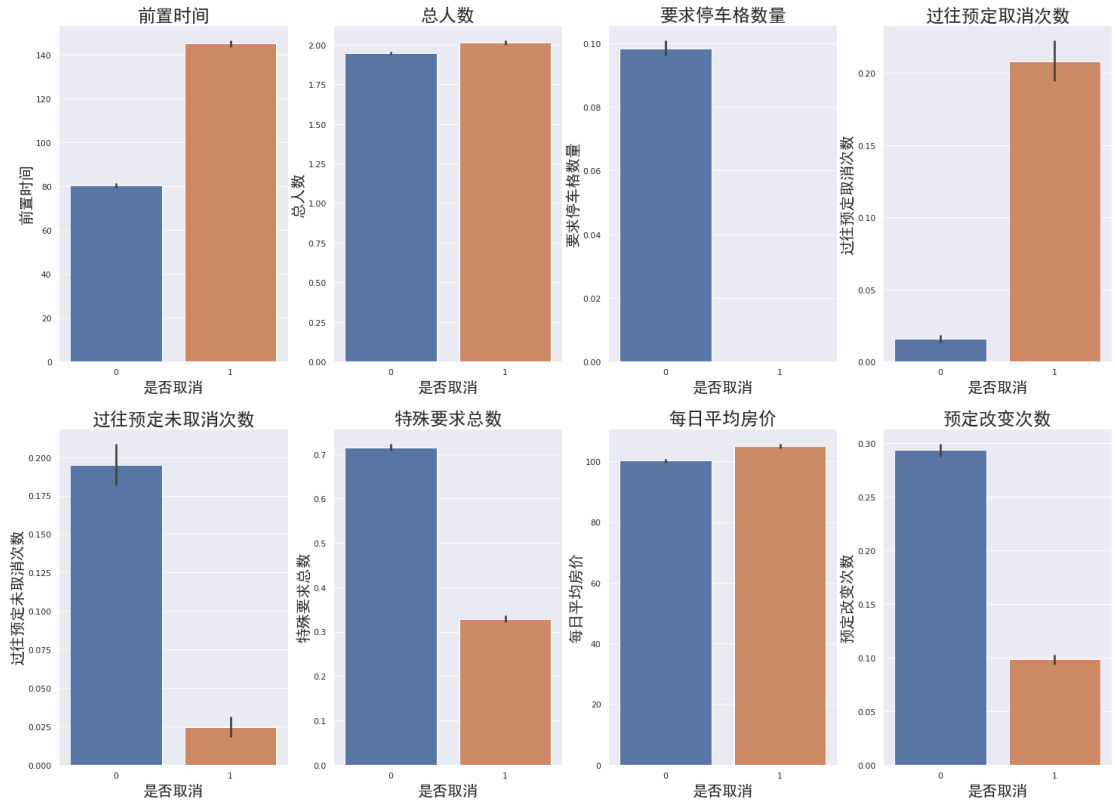


图 **1:** 前置时间、总人数、过往预定取消次数、过往预定未取消次数、要求停车格数量、特殊要求总数、每日平均房价以及预定改变次数与预定是否取消的条形图（条形的高度显示组内数据的平均值；是否取消为 **0** 代表未取消，为 **1** 代表取消。）

### 3.2.2 分类变量

在分类变量部分，首先从本报告的预测目标是否取消开始，从图 2可以看到数据中预定取消的样本数为 44,157 笔，占总样本量的 37.14%，数据并没有严重分布不均的问题。接着本报告将针对是否为重复顾客、国家、抵达日期-年、是否有押金、市场划分、顾客类型等变量进行描述性统计，简单查看数据的分布情况及与预定是否取消的影响关系。

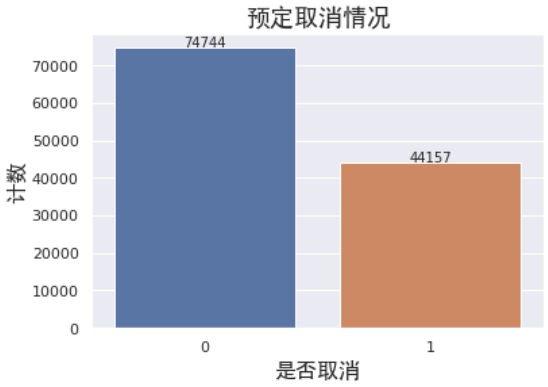


**图 2: 预定取消情况**

表 3为是否为重复顾客、国家、抵达日期-年及是否有押金的预定取消情况。从是否为重复顾客的数据分布来看，可以看到是重复顾客的预定取消占比相较非重复顾客来的低，这个结果是合乎预期的，若是顾客选择先前住过的酒店可能代表顾客对于该酒店的满意度较高，取消预定的可能性或许较低；各个国家消费者的预定取消情况这里仅展示预定次数最高的前五名国家，分别为葡萄牙（PRT）、英国（GBR）、西班牙（ESP）、法国（FRA）和德国（DEU）。其中，来自葡萄牙的旅客取消预定的情况相较其他国家高出许多；从抵达日期-年的数据来看，可以看到 2015 年以及 2017 年预定取消的占比相较 2016 年稍高，但并没有显著的区别；而是否有押金的预定取消情况则与一般预期的不同，需要押金的预定取消次数反而有 14,516 笔，比未取消的次数高出许多，因此本报告接着会尝试针对此结果查找可能原因。

表 4为不同市场划分类型中是否有押金的数据分布情况，可以看到有押金的预定主要集中在团体、线下旅游代理商/经营商以及企业市场划分类型。对于这类需要押金但仍然取消的预定，可能原因是因为团体旅客一旦无法成行便会集体取消，或是企业会为员工负担被没收的押金，因此没有退款上的压力，然而详细的原因可能还需要与酒店工作人员进一步的调研才能知晓。

从市场划分的预定取消情况来看，如表 5所示，可以看到预定主要集中在线上旅游代理商，同时团体、线上旅游代理商以及线下旅游代理商/经营商等市场划分类型的预定取消占比较高。对于顾客来说，线上预定的可变动性较大，顾客能够很轻易的在网上找到丰富的酒店选择，因此线上旅游代理商的预定取消率较高是十分合理的。

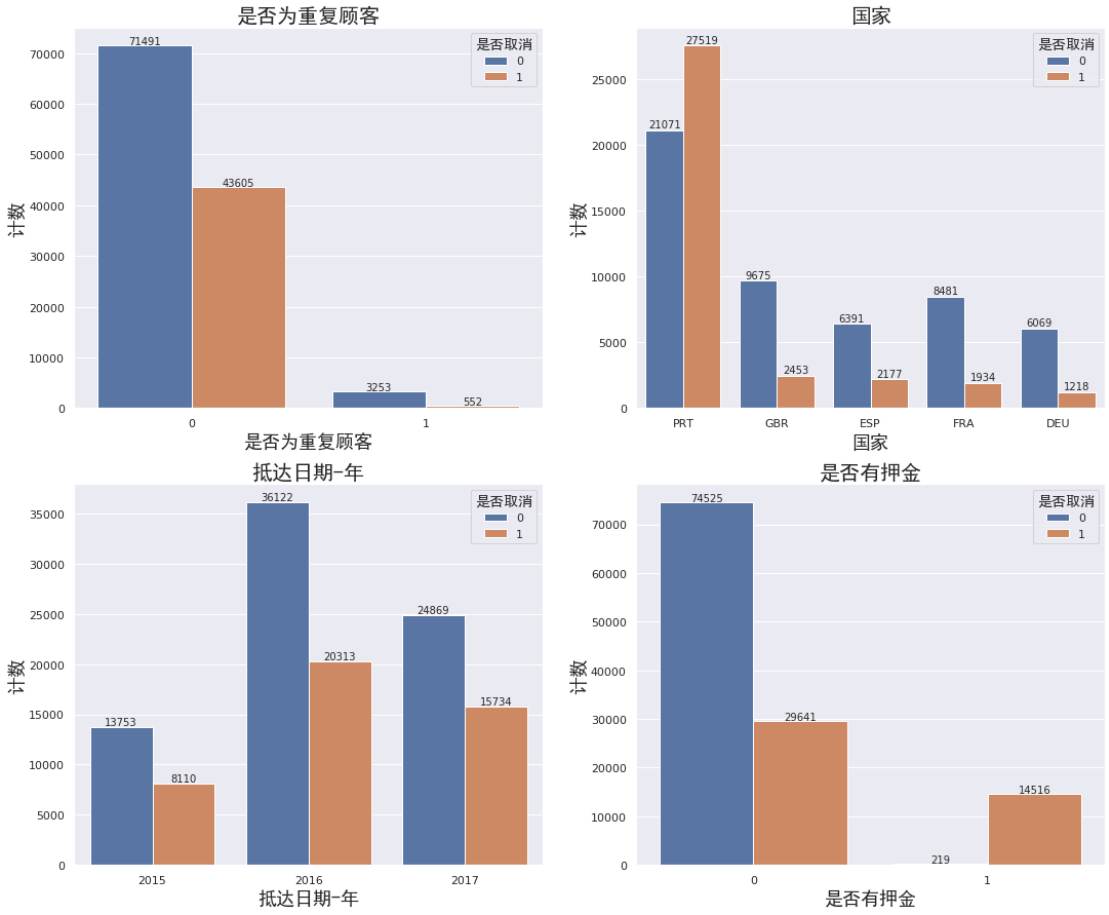


图 3: 是否为重复顾客、国家、抵达日期-年及是否有押金的预定取消情况

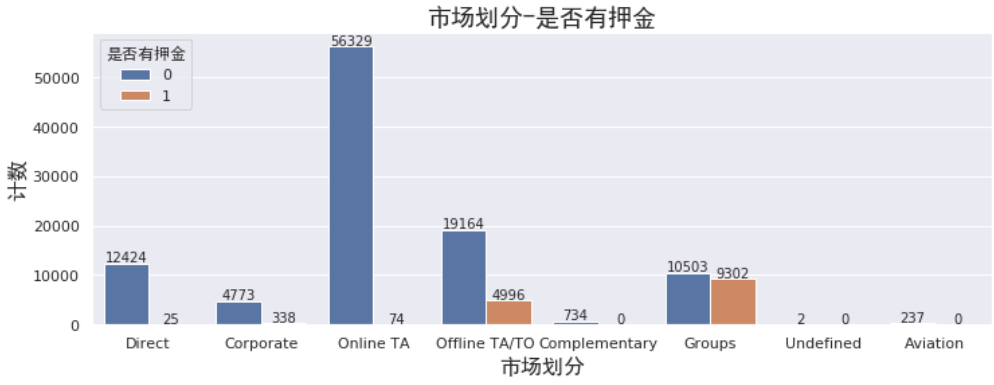


图 4: 市场划分-是否有押金数据分布情况

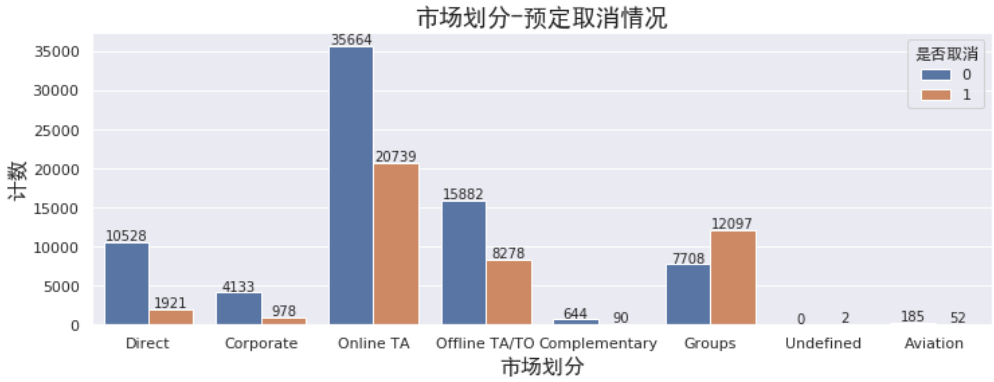


图 5: 市场划分-预定取消情况

最后，表 6为顾客类型的预定取消情况，可以看到预定的顾客主要为短期住宿旅客，其中与其他预定无关的短期住宿旅客预定取消的占比相对其他类型较高，可能是因为该类旅客的行程自由度较高，会因为各种个人因素取消预定。同时，可以发现顾客类型中也有“团体”的类别，然而其与“与其他预定相关的短期住宿旅客”以及市场划分变量中“团体”的区别并没有在原始数据集论文中详细说明，因此无法进行更深入的探讨。

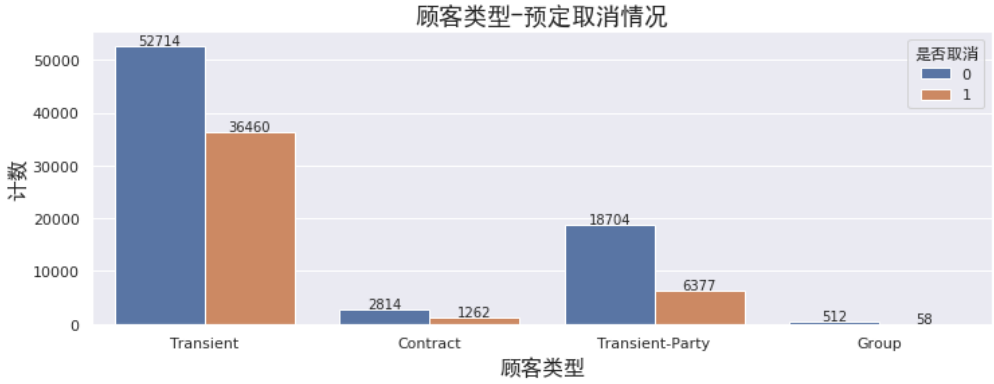


图 6: 顾客类型-预定取消情况

## 3.3 预测模型

自从人工智能围棋程序 AlphaGo 在与韩国围棋九段棋手李世石及中国围棋九段棋手柯洁的围棋人机大赛中取得出色表现之后，深度学习的话题突然火爆了起来；相比之下，横扫 Kaggle 大赛各个奖项的 XGBoost 模型名气却是小了许多。然而，究竟两种方法孰优孰劣? 对此，本报告在模型选择上，采用了 XGBoost 模型以及多层感知器神经网络模型作为主要的预测模型。

XGBoost 模型由陈天奇等 [2] 提出，XGBoost 模型是基于梯度提升决策树（Gradient Boosting Decision Tree, GBDT）[3] 的改良与延伸，然而其与 GBDT 模型不一样的地方在于：(1) 将损失函数从平方损失推广到二阶可导的损失。(2) 在目标函数中添加了惩罚函数，通过正则化限制模型的复杂度。(3) 使用随机森林模型中的列抽样方法提升模型效果。

而多层感知器神经网络模型主要源于第一个神经网络模型——感知器 [7]，表 7为感知机网络结构示意图，单个神经元感知机会将前层输入的所有神经元做线性运算，再把结果输入激活函数，即 $y = f(\sum_{i=1}^{n} w_i x_i + b)$，其中 f 为激活函数。但是单层感知器学习能力非常有限，无法解决线性不可分问题，因此就需要多层感知器，也就是本报告采用的多层感知器神经网络模型，感知器隐层越多，理论上就能拟合越复杂的函数。

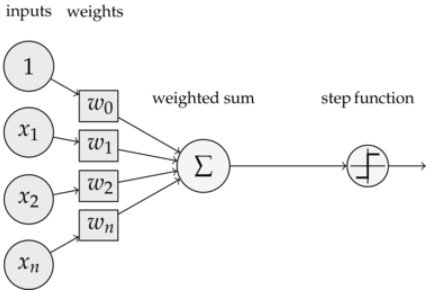


**图 7: 单层感知器示例**

多层感知器神经网络模型的网络架构如图 8所示，每一个神经元会把前一层所有神经元的输出作为输入，其输出又会给下一层的每一个神经元作为输入，相邻层的每个神经元都有“连接权”，神经网络所学习到的东西隐含在连接权和偏置项中。由于引入了非线性的激活函数，深度神经网络中的损失函数往往是非凸的，因此通常会使用梯度下降法求得数值解。

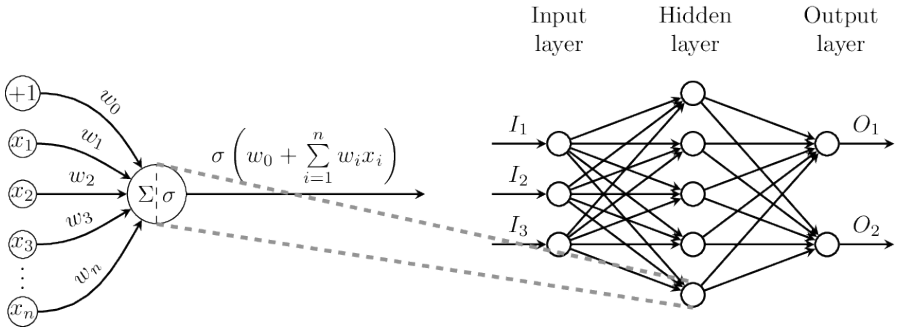


**图 8: 多层感知器神经网络模型架构示例**

本报告将通过外样本预测精度的评比，尝试比较 XGBoost 模型以及多层感知器神经网络模型的预测能力，同时以逻辑回归模型作为基准，决策树及随机森林模型作为对比模型，一同进行预测精度的评估。

# 4 实验

## 4.1 模型训练

本报告的目标便是在给定输入变量的情况下，建立输出为预定是否取消的预测模型。本报告随机抽取了数据集中 80% 的样本作为训练集，剩下 20% 作为测试集。接着，使用训练集针对先前提到的逻辑回归模型、决策树模型、随机森林模型、XGBoost 模型以及多层感知器神经网络模型进行 5 折交叉验证，从而找到各个模型的最佳参数组合。在模型的实现方面，XGBoost 模型使用了 xgboost 函数库 [2]；多层感知器神经网络模型使用 pytorch 函数库 [5]；其他模型使用 scikit-learn 函数库 [6]，各个模型的最终细节如表 3所示。

**表 3: 各模型细节**

| 模型 | 详细说明 |
|---|---|
| **XGBoost 模型** | 通过 xgboost 中的 XGBClassifier 实现，将参数 learning_rate 设为 0.1，max_depth 设为 10，n_estimators 设为 500，其余参数维持预设值。 |
| 多层感知器神经网络模型 | 模型网络设计部分使用两层隐藏层，维数分别为 128、64，在每层隐藏层后都会使用 ReLU 激活函数、批标准化以及 Dropout（比例为 0.5）处理数据。损失函数为负对数似然损失函数（Negative Log Likelihood），优化器使用 Adam[4]，学习率为 0.001，训练循环次数为 300，批大小为 128。 |
| 逻辑回归模型 | 通过 scikit-learn 中的 LogisticRegression 实现，将参数 C 设为 10，其余参数维持预设值。 |
| 决策树模型 | 通过 scikit-learn 中的 DecisionTreeClassifier 实现，将参数 ccp_alpha 设为 0，max_depth 设为 13，其余参数维持预设值。 |
| 随机森林模型 | 通过 scikit-learn 中的 RandomForestClassifier 实现，将参数 max_depth 设为 13，min_samples_split 设为 2，n_estimators 设为 500，其余参数维持预设值。 |

## 4.2 模型结果对比

在通过交叉验证方法选出各方法的最佳模型之后，本报告使用测试集针对各模型的预测能力进行评估，表 4列出了各个模型预测结果的 Accuracy、Precision、Recall 以及 F1-score，从结果可以看到 XGBoost 模型在 Accuracy、Recall 以及 F1-score 的表现最好；随

机森林模型在 Precision 的表现最佳；而多层感知器神经网络模型的表现略差于 XGBoost 模型，在 Accuracy、Recall 以及 F1-score 的表现都位居第二名。整体而言，XGBoost 模型在五个模型中展现的预测能力是最好的。根据 XGBoost 模型的发明人陈天奇所述，不同的机器学习模型适用于不同类型的任务：深度神经网络通过对时空位置建模，能够很好地捕获图像、语音、文本等高维数据；而基于树模型的 XGBoost 模型则能很好地处理表格数据，同时还拥有一些深度神经网络所没有的特性比如模型的可解释性、输入数据的不变性、更易于调参等等。由于本报告使用的数据为表格数据，XGBoost 模型在此数据中取得的良好表现更验证了前述说法。

表 4: 模型预测精度评估结果

| 模型 | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| **XGBoost 模型** | **89.46%** | 87.87% | **83.41%** | **85.58%** |
| 多层感知器神经网络模型 | 86.92% | 85.84% | 77.98% | 81.72% |
| 逻辑回归模型 | 80.51% | 81.06% | 62.67% | 70.69% |
| 决策树模型 | 85.07% | 82.08% | 77.01% | 79.46% |
| 随机森林模型 | 85.89% | **90.37%** | 69.84% | 78.79% |

## 4.3 模型重要特征

基于在测试集预测表现最好的 XGBoost 模型，本报告进一步查看各输入变量的特征重要性。图 9显示了特征重要性最高的前十名变量，从高到低分别为：是否有押金、要求停车格数量、市场划分（线上旅游代理商）、过往预定取消次数、抵达日期-年（2015）、消费者类型（与其他预定无关的短期住宿旅客）、特殊要求总数、过往预定未取消次数、国家、消费者类型（与其他预定相关的短期住宿旅客）。总的来说，这些变量与第二部分描述性统计呈现的结果大致相同。

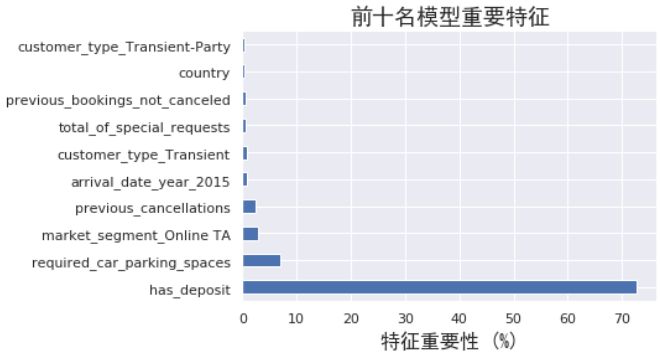


图 9: 前十名模型重要特征

# 5 结论

本报告基于酒店预定相关数据，以预定是否取消作为预测变量，并通过 27 种输入变量刻画描述酒店预定的指标，接着以各变量的数据为基础，以 XGBoost 模型以及多层感知器神经网络模型为主要模型，逻辑回归模型作为基准，决策树模型及随机森林模型作为对比模型，尝试训练出预测能力最佳的模型。根据各个模型的外样本预测结果，XGBoost 模型整体展现的预测精度最佳。基于 XGBoost 模型特征重要性的结果，本报告主要归纳出了以下结论: 是否有押金、要求停车格数量、市场划分（线上旅游代理商）、过往预定取消次数、抵达日期-年（2015）、消费者类型（与其他预定无关的短期住宿旅客）、特殊要求总数、过往预定未取消次数、国家、消费者类型（与其他预定相关的短期住宿旅客）为特征重要性最高的前十名变量。

由于影响预定取消的原因非常多元，因此在未来的研究中可以考虑在模型中添加更多的输入变量，比如酒店的交通便捷度、当地竞争酒店个数，并且可以考虑增加样本量，以更好的训练模型，对模型进行优化。

# 参考文献


[1] Nuno Antonio, Ana [de Almeida], and Luis Nunes. Hotel booking demand datasets. *Data in Brief*, 22:41 – 49, 2019.

[2] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.

[3] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[4] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

[5] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[7] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, pages 65–386, 1958.