【A04】航空旅客付费选座意愿识别【东软】

1.问题：关于数据的几个问题

1. 'emd_lable'特征是什么意思？和‘emd_lable2’有什么关系？

业务上有两种规则用来确定是否为付费选座行为。用emd_lable1确认的付费选座范围较大，emd_lable2被包含在其中。
本次竞赛中采用emd_lable2为付费选座业务准则。

2. 'gender'特征中”U“和”0“是什么意思？

“U”、“0”为无效数据，请忽略。

3. 'marital_stat'特征的各个值代表什么意思？

“M”，已婚；“S”，单身；“U”、“0”为无效数据，请忽略。

4. 'ffp_nbr'有会员编号，但是会员等级为0，是什么意思？有会员等级，但是没有会员编号又是什么意思？

这是两个不同会员体系管理系统的数据整合时的问题。请酌情使用。

5. 'member_level'会员级别的排序是什么样的？

会员级别说明不能提供。请酌情使用。

6. 'pref_orig_m3_1'特征的解释中，没有最后’_1‘、’_2‘...的解释

统计期内，出行次数排名前5的偏好出发地。

2.问题：时间描述

关于特征列中 最近Y年和过去Y年 的意思存在歧义，具体意思请出题方明确一下。

两个说法意思一样。

3.问题：a04

提供数据中有 过去Y年总里程和最近Y年总里程，请问这两个数据是如何区分的。

两个说法意思一样。

4.问题：航空累计次数与非航累积次数

请问航空累积次数和非航累积次数具体含义是什么，与flt_cnt_m3最近Y年飞行次数有什么内在关联？

比如航空累积总次数的统计量pit_accu_cnt_m6，pit_accu_cnt_y1，pit_accu_cnt_y2,pit_accu_cnt_y3等全部是空值，与最近Y年飞行次数并不相符

非航累积总次数 和 航累积总次数，指航空里程积累次数。非航空里程积累，常见有信用卡积分换里程。flt_cnt,指乘坐飞机的次数，有的舱位等级不能积累里程。

5.问题：问题汇总（14个）

2、数据中出发均为浦东，达到为纽约、洛杉矶和悉尼，是否会提供多一点出发达到数据?

以已提供的数据集的准。

3、能否告知舱位中C、F、J、W、Y含义，或确定C舱为公务舱公布价，F舱为头等舱公布价，J舱为公务舱公布价，W舱为普通舱35折，Y舱为普通舱（经济舱）公布价是否准确?

请参考：https://baike.baidu.com/item/%E9%A3%9E%E6%9C%BA%E8%88%B1%E4%BD%8D/4764328?fr=aladdin

4、性别中0是什么含义（训练样本中0为18000左右）、U是未知；年龄层中0是什么含义？（训练样本为0为19000左右）；生日为0为19000左右；定居国家总0为19000左右；婚姻状态中0为22000左右，数据缺失较多，是否为重新给数据?

“U”、“0”为无效数据，请忽略。

5、机票费是否包含机票税费?

不含。

6、会员级别含义?

比如“普通”、“白银”、“黄金"|等。会员级别说明不能提供。请酌情使用。

7、偏好机型0、1、2含义

统计期内，搭乘次数排名前5的偏好机型。

8、prebuy_d_cnt提前购买次数-国内数据中有0、1、2、3、4、5、9含义，另外出发均为上海浦东，达到为纽约、洛杉矶和悉尼，应该均为国际数据，为何有国内数据中有值?

旅客可以搭乘国内和国际航班。本次竞赛数据集仅提供了国际航班的乘机记录，但统计期内，旅客提前购买国内和国际的统计数据都提供了。

9、购买付费SSR的次数是否为付费选座次数?

不是。

10、常飞月分值为何有14?

数据处理错误，请忽略。

11、select_seat_cnt最近Y年优选座位次数和购买付费SSR的次数区别?

这两个字段不包含在竞赛数据中。请忽略。

12、flt_leg_d_cnt最近Y年乘坐国内航段次数含义，另外出发均为上海浦东，达到为纽约、洛杉矶和悉尼，应该均为国际数据，为何有国内数据中有值?

同理，tkt_d_amt最近Y年机票金额-国内是否也不应该有值?

旅客可以搭乘国内和国际航班。本次竞赛数据集仅提供了国际航班的乘机记录，但统计期内，旅客提前购买国内和国际的次数都提供了。

13、noshow_rate最近1年NOSHOW率含义?

买了机票但没有实际乘机。

14、pit_accu_air_amt航空累积余额和pit_accu_non_amt非航累积余额含义?

非航累积总次数 和 航累积总次数，指航空里程积累次数。非航空里程积累，常见有信用卡积分换里程。flt_cnt,指乘坐飞机的次数，有的舱位等级不能积累里程。

15、mdl_influence影响力指标含义?

数据集中没有提供此内容。请忽略。

6.问题：关于评价指标、数据集

题目中【用户期望】强调模型的准确性和简单性。识别出的有付费选座意愿的旅客，准确率越高越好，用于模型训练的特征因子越少越好。

通过分析标签分布， 我们小组成员认为不应该从准确率指标上来评判模型的好坏。而是应该从精确率 回归率 F1值、PR曲线、ROC曲线、AUC值等这些指标来观察模型。

其次题目中要求提交的数据也是不合理的--“要求提供不超过500人的预测结果旅客名单，按照有付费选座意愿的概率倒排序。“ 既然要求了不多于500人，并且要求准确率越高越好，是否可以只提交一人 这样准确率不是0就是100%，这样也满足了赛题的条件。

最后，我们团队分析了所给数据，数据大部分信息是0值，可用的特征列很少，非常容易造成过拟合，即使训练出一个好的模型，在日后真正投入使用中，效果会相差甚远

请以提供数据为准。

7.问题：0的含义

在所给的CSV文件中，有很多0值，这里的0所代表的含义 是否是不一致的？例如在gender的列中， 0代表的应该是未知，在次数列 0代表的是未知还是就是0次？

在Gender中表示未知；在次数列中表示0.

8.问题：数据重复？

过去Y年总飞行次数与最近Y年飞行次数是不是相同的含义？

是的。相同。

fit_cnt_y2和fit_nature_cnt_y2，其中

dist_cnt_y2和fit_nature_cnt_y2数据为0，明显存在数据错误，是否可以通过平均非常航程计算得到？

这两列数据为0是数据预处理时的错误。

9.问题：休息日与非工作日、节假日的区别？

数据中存在休息日飞行次数这一概念，是否为非工作日与节假日的并集？这些数据在训练集中都是0值，是否需要重新提供？

休息日为周六、周日。节假日为法定节假日。

10.问题：税费非常高是否正常

有的税费达到几十万、或者税费比票价高很多倍，是否是正常的

以提供的数据为准。

11.问题：前面机箱是否为某旅客某一次出行的信息

数据前面几列，包括出发、到达、航班号、舱位、航班日期、机票费、机票税费是不是旅客的某次飞行记录

是的。

12.问题：关于会员编号的问题

会员编号和会员级别为什么无法对应？例如有的记录里面有会员编号，但是没有会员级别，另外有的记录里面有会员级别但是没有会员编号

这是不同会员体系管理系统的数据整合时的问题。请酌情使用。

13.问题：出现星号或者井号的意思是?

city_name和province_name中都出现“****”，是否和0（未知）的含义一样

是的。

14.问题：提前购买次数

提前购买次数里面的d3、d7、d14、d30、d99分别表示什么意思

d3指提前3天；d7指提前7天。以此类推。

15.问题：关于样本

能否多给些选座的样本?

以提供的数据为准。

16.问题：关于特征名

样本的特征名，可以解释的更详细一些吗?

以提供的数据为准。

17.问题：数据中机票价格问题

机票价格中有部分数据过大，数十万乃至百万，数据是否正确?

以提供的数据为准。

18.问题：会员等级问题

1：验证集的会员等级类型比训练集数据类型更多。

以提供的数据为准。

2：会员等级类型表示什么含义

比如“普通”、“白银”、“黄金"|等。会员级别说明不能提供。请酌情使用。

19.问题：出发点与到达点问题

出发点与到达点什么意思？是否指代某地方？

出发点指乘机地点，到达点指落地地点。

20.问题：累计月份

例如：mouth_m数字a_数字b

其中数字b表示什么意思？

你指的是_m3,_m6吗？指统计期间为最近3个月，最近6个月。

21.问题：属性列相关问题

有些属性列意思不明，导致数据预处理工作难以开展

以提供的数据为准。

22.问题：关于数据的处理

进行数据分析时时要将650个特征因子都进行分析还是自己选取几个因子分析

尽量提供使用较少的因子的模型。

23.问题：关于开发工具

一般用什么软件操作

开发工具软件没有限制。

24.问题：U和0的含义

Gender 列中，F和M分别代表女性和男性，那U和0的各自的含义是什么，都表示未知吗？婚姻状态列也有同样的问题。

“U”、“0”为无效数据，请忽略。

25.问题：Age列

Age列表示的是年龄层，中间有大量的2020年11月20日是什么意思?

Excel错误。应为年龄层11-20。

26.问题：city name和province name

city name中大量的****是什么意思？还有少量数字。province name中也有少量的****

****和少量数字为无效数据。

27.问题：关于数据方面

数据中有很多0出现不知道代表什么意思，是数据缺失的意思吗，比如性别那里，有的是M，有的是F，有的是0，这个0究竟是代表什么意思

“U”、“0”为无效数据，请忽略。

28.问题：最终提交文本方面问题

最后提交的文本上面需不需要诠释我们的代码逻辑，比如为啥模型上面使用lgb不用xgb，或者通过准确率依次验证多种模型最终选出来一个模型，还是只需要说明我们最终选择的模型就可以了

尽量说明您的建模思路。

29.问题：特征工程方面问题

1.假设使用了100个原始特征使用PCA降维到20维，然后训练模型，那最终认为的使用到的特征数量是100个还是20个

算20个。但请尽量说明这20维的含义。

2.假设使用20个原始特征通过交互衍生得到100个特征，函数训练模型，那最终认为的使用到的特征数量是20个还是100个

算100个。

1.问题：我们需要做出完整的车载中控吗?

我们需要做完整的车载中控那一整套系统吗，还是说只要实现赛题要求的两个功能即可?

只要实现赛题要求的两 个功能

2.问题：系统需要做登录模块吗?

需要针对不同的用户建立登录系统吗? 还是说只是需要一个无登录状态即可，也就是无论放到哪一台车都可以使用的样子。

系统可以做登陆也可以不做，这个不做强求。