中山大學
SUN YAT-SEN UNIVERSITY

数据挖掘导论大作业

题目 Title：客户流失及客户行为偏好分析

院 系 School (Department)：计算机学院

专 业 Major：软件工程

学生姓名 Student Name：杨玲

学 号 Student No.：18342115

时间：2021 年 7 月 5 日

Date: Month July Day 5th Year 2021

【摘　要】

了解客户流失及客户行为偏好对各大服务类应用而言都十分重要，应用软件可以通过数据深入了解用户画像及行为偏好，挖掘出影响用户流失的关键因素，并通过算法预测客户访问的转化结果，从而更好地完善产品设计、提升用户体验。本文中主要针对携程用户的数据做出分析，在经过数据预处理之后，我们就这些经过处理的数据，从客户流失、客户价值和客户转化三个方面对客户流失及客户行为偏好进行了分析。

【关键词】：随机森林；逻辑回归；特征提取；数据挖掘

[ABSTRACT]

Understanding customer churn and behavior preferences is very important for major service applications. Application software can use data to understand user portraits and behavior preferences, unearth key factors that affect user churn, and use algorithms to predict the conversion results of customer visits so as to better improve product design and enhance user experience. This article mainly analyzes the data of Ctrip users. After data preprocessing, we analyze customer churn and customer behavior preferences from three aspects: customer churn, customer value, and customer conversion.

[Keywords]: random forest; logistic regression; feature extraction; data mining

目　录

第一章 引言

1.1 问题的描述

了解客户流失及客户行为偏好对各大服务类应用而言都十分重要，应用软件可以通过数据深入了解用户画像及行为偏好，挖掘出影响用户流失的关键因素，并通过算法预测客户访问的转化结果，从而更好地完善产品设计、提升用户体验。

1.2 本文的工作

我们通过数据挖掘分析影响用户流失的关键因素、深入了解用户行为偏好以此做出调整，提升客户留存率，增强客户黏性，并通过随机森林算法预测客户流失。

首先对数据进行预处理，为此需要探索数据分布、数据缺失情况，针对性的进行缺失值填补，对于缺失较少的重要特征选择随机森林缺失填补法，使用 3sigma、箱型图分析等对异常值进行处理，对分类型变量进行编码。

随后，使用方差过滤、F 检验过滤掉一部分特征，进行 WOE 分箱，对每个特征分箱结果进行可视化，分析每个特征分箱情况并以此分析用户行为偏好，使用各个特征的 IV 值进一步筛选特征。

最后，训练逻辑回归模型，通过其算法可解释性强的特点来对用户流失关键因素进行阐述；并训练随机森林模型，进行模型调参、评估、输出模型，以此模型对用户流失进行预测，以便针对性地挽留用户。

1.3 论文结构简介

本文第一章主要是阐述研究客户流失及客户行为偏好分析的背景和意义以及本文的大致工作内容；第二章主要综述本文在完成客户流失及客户行为偏好分析的过程中所使用到的一些模型；第三章介绍的是如何对数据进行预处理；第四章将会介绍我们在解决问题过程中所作的工作和所提出的方法；第五章是模型构建结果的评估；第六章是对本文的总结；最后是参考文献。

第二章 相关工作综述

2.1 随机森林提取特征重要性

随机森林的算法可以用如下几个步骤概括：

a. 用有抽样放回的方法（bootstrap）从样本集中选取 n 个样本作为一个训练；

b. 用抽样得到的样本集生成一棵决策树；

c. 在生成的每一个结点：

   a) 随机不重复地选择 d 个特征；

   b) 利用这 d 个特征分别对样本集进行划分，找到最佳的划分特征（可用基尼系数、增益率或者信息增益判别）；

   c) 重复步骤 1 到步骤 2 共 k 次，k 即为随机森林中决策树的个数；

d. 用训练得到的随机森林对测试样本进行预测，并用票选法决定预测的结果。

2.2 嵌入式模型 SelectFromModel

SelectFromModel 是一个基础分类器，其根据重要性权重选择特征。可与拟合后具有 coef_或 feature_importances_属性的任何估计器一起使用。如果相应的 coef_或 feature_importances_值低于提供的 threshold 参数，则这些特征可以认为不重要或者删除。除了指定数值阈值参数，还可以使用字符串参数查找阈值，参数包括：“mean”，“median” 以及这两个参数的浮点数乘积，例如“0.1*mean”。与 threshold 标准结合使用时，可以通过 max_features 参数限制选择的特征数量。

第三章　数据预处理

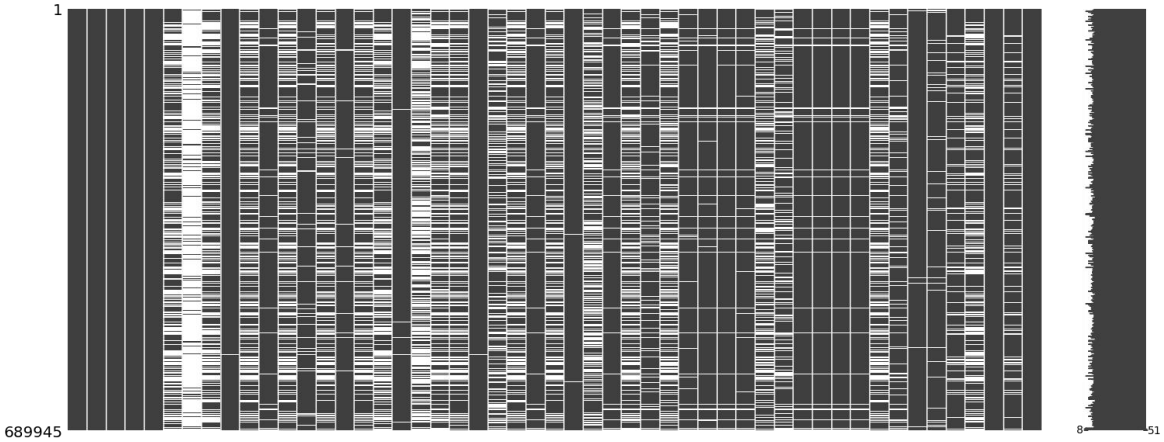
3.1　清洗数据

数据清洗主要包括处理缺失值和处理异常值这两个操作。

3.1.1 处理缺失值

首先计算出该数据集中的空值占所处字段的比例，然后将其按由小到大的顺序排序，图片 1 对数据集的缺失情况做出图示。针对缺失情况，我们采用以下方法对数据集缺失值进行处理：

a. 分类型变量用“众数”填补：分类型变量有 decisionhabit_use（决策习惯）；
b. 含有负数的特征用“中值”填补：这一类特征多为价格；
c. 方差大于 100 的连续型变量用“中值”填补；
d. 缺失大于等于 35%且小于 80%用“常数-1”填充单独做一类；
e. 超过 80%直接删除变量：从得到的结果中我们发现，historyvisit_7ordernum（近 7 天用户历史订单数）缺失比例最大，为 0.879824，所以我们删去 historyvisit_7ordernum 这一变量；
f. 其余变量用“均值”填补。



图片 1：可视化缺失比

3.1.2 异常值处理

首先处理异常值。在我们使用的数据集中，最低酒店定价有小于 0 的值，有等于 1 的值，明显属于异常值，我们使用盖帽法处理这些异常值。算法 1 显示了盖帽法的具体实现，同时处理前后的结果在表格 1 和表格 2 中显示。

算法 1：盖帽法

```
def block_lower(x):
# x 是输入的 Series 对象,替换 1%分位数
    ql = x.quantile(.01)
    out = x.mask(x<ql,ql)
    return(out)

def block_upper(x):
# x 是输入的 Series 对象,替换 99%分位数
    qu = x.quantile(.99)
    out = x.mask(x>qu,qu)
    return(out)
```

表格 1：异常处理前

（注：lowestprice 为当前酒店最低价格，lowestprice_pre 为 24 小时内已访问次数最多酒店可订最低价）

	count	mean	std	min	1%	25%	50%	75%	99%	max
lowestprice	687931	318.806242	575.782415	-3	37	116	200	380	1823	100000
lowestprice_pre	659689	315.954583	463.723643	1	38	118	208	385	1750	100000

表格 2：异常处理后

	count	mean	std	min	1%	25%	50%	75%	99%	max
lowestprice	687931	305.025771	297.382838	37.0	37	116	200	380	1823	1823
lowestprice_pre	659689	304.439507	287.1925123	38.0	38	118	208	385	1750	1750

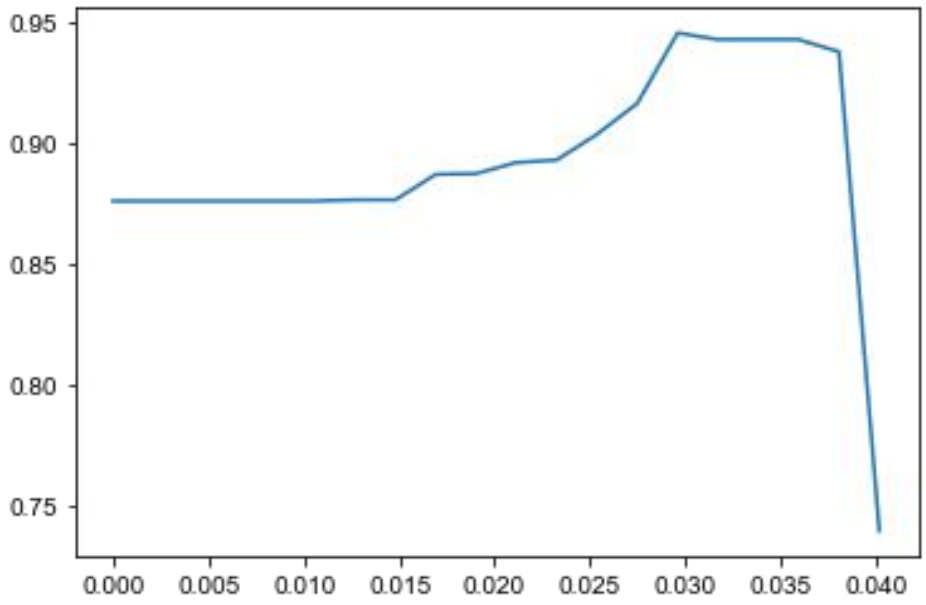
3.2 特征提取

3.2.1 剔除无关变量

计算所定义的计算两个总体（训练集 X 和训练集 Y）之间的 F 检验。根据计算发现，与标签没有显著关系的变量有 6 个。此时，剔除与标签没有关系的变量，并重置索引、用新生成的列表取代原来的训练集和测试集。

3.2.2 提取重要特征

利用随机森林特征重要性属性 feature_importances_定义阈值范围，以嵌入选择模型 SelectFromModel 为基础，通过交叉验证 cross_val_score 得到每个阈值下模型得分情况，图片 2 显示了不同阈值下模型的得分情况。我们找出最高分对应的阈值，并保留大于该阈值的部分。依照这种方法，总共选择出 8 个特征，分别是：lasthtlordergap, cityuvs, cityorders, lastpvgap, cr, sid, visitnum_oneyear, h。表格 3 显示了只保留重要特征的新训练集的表头。



图片 2：学习曲线（横轴为阈值，纵轴为得分）

表格 3：

index	lasthtlordergap	cityuvs	cityorders	lastpvgap	cr	sid	visitnum_oneyear	h
0	-1	2.267	0.133	0	1.139221	27	1050	0
1	23823	25.08	3.353	1753	1.139221	257	5780	14
2	30793	10.747	1.613	30792	1.2	205	3616	11
3	355113	0.913	0.14	806	1	184	1064	13
4	161858	0.107	0.007	19664	1	286	1468	14

第四章　模型构建

4.1　模型选择

我们主要从三个方面来对客户流失及客户行为偏好进行分析，即客户流失、客户价值和客户转化。

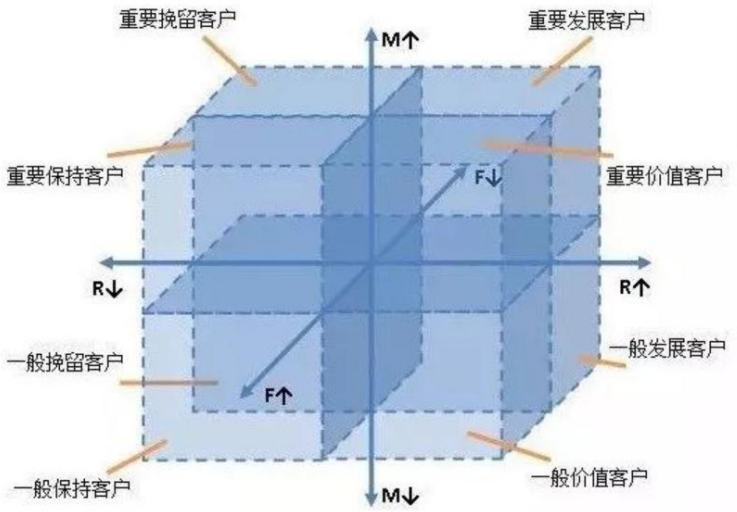
4.1.1 客户流失

客户流失预测模型的实现方法属于分类算法，常用算法包括逻辑回归、支持向量机、随机森林等。大部分情况下，流失客户的样本分类是少数类，需要注意处理样本不均衡问题。

我们用目标变量 label 来表示是否流失，这是一个 0-1 二分类问题，目的是需要挖掘出关键因素，拟选用逻辑回归做模型训练及预测。

4.1.2 客户价值

为了更加细致地挖掘客户价值，我们选择 RFM 客户价值模型进行分析。



图片 3：RFM 客户价值模型

RFM 客户价值模型是根据客户最近一次的购买时间 R(Recency)、购买频率 F(Frequency)、购买金额 M(Monetary)计算得出 RFM 得分，通过这三个维度来评估客户的订单活跃价值，常用来做客户分群或价值区分。

4.1.3 客户转化

预测客户转化率，是一个连续型变量预测问题，所以我们拟选择集成数模型--随机森林回归来进行预测。

4.2 客户流失预测模型

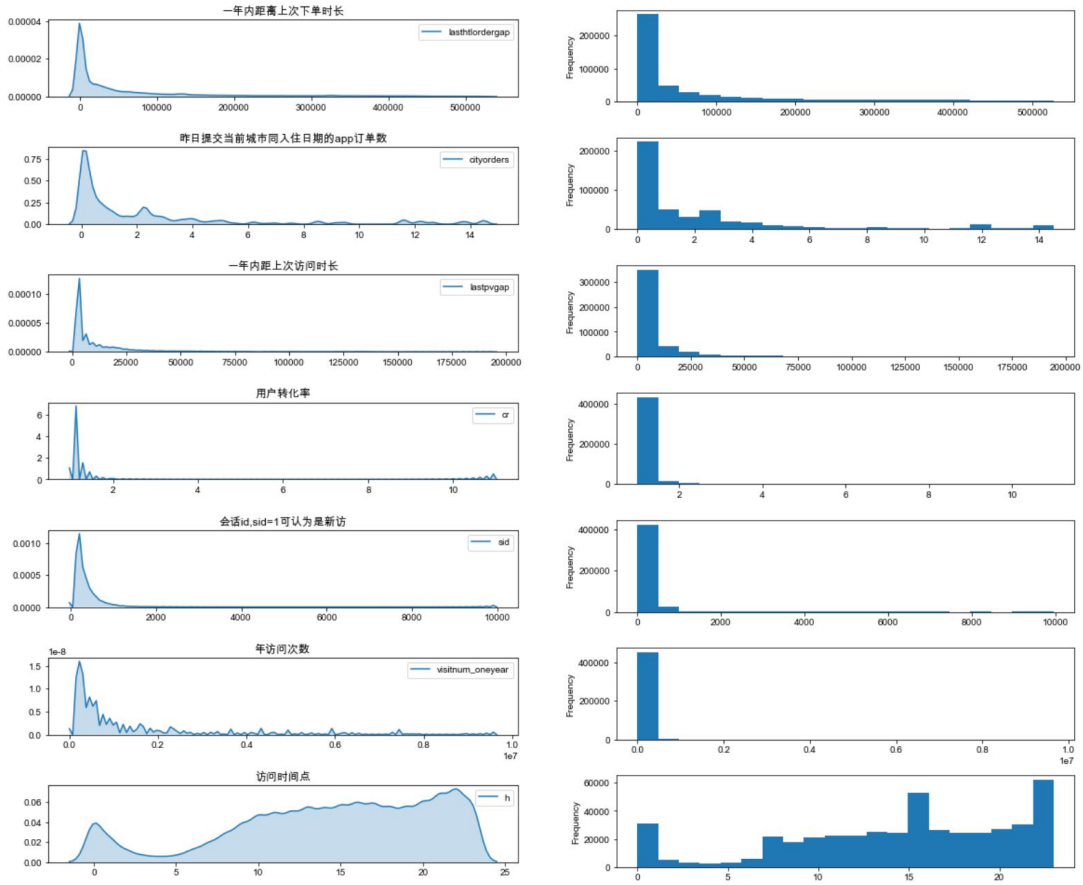
因为客户流失预测模型需要知道关键因素，要求模型需要有很好的可解释性，所以经过数据预处理后，我们决定利用逻辑回归了解用户画像及行为偏好，挖掘出影响用户流失的关键因素。

但从模型评价结果（ROC 曲线面积）来看，逻辑回归并不是很理想，考虑到要同时追求模型预测精确度，需要选取集成模型或其他强学习模型，我们选择辅以随机森林分类器来进行预测客户流失。

4.2.1 逻辑回归模型

首先，在将数据用于模型训练之前，需要先对变量进行深入分析。分析变量间是否存在高度相关性，连续性变量是否需要离散化，离散变量是否需要编码等等。

不难发现，除访问时间点外，特征均成右偏分布。因此将连续型、分类型变量做分箱处理，此处我们选用 WEO 分箱处理。



图片 4：特征变量分析

我们按一以下步骤总结进行分箱:

- 先把连续型变量分成一组数量较多的分类型变量，比如，将几万个样本分成 100 组，或 50 组(尽量有监督的分箱)；并确保每一组中都要包含两种类别的样本，否则 IV 值会无法计算。

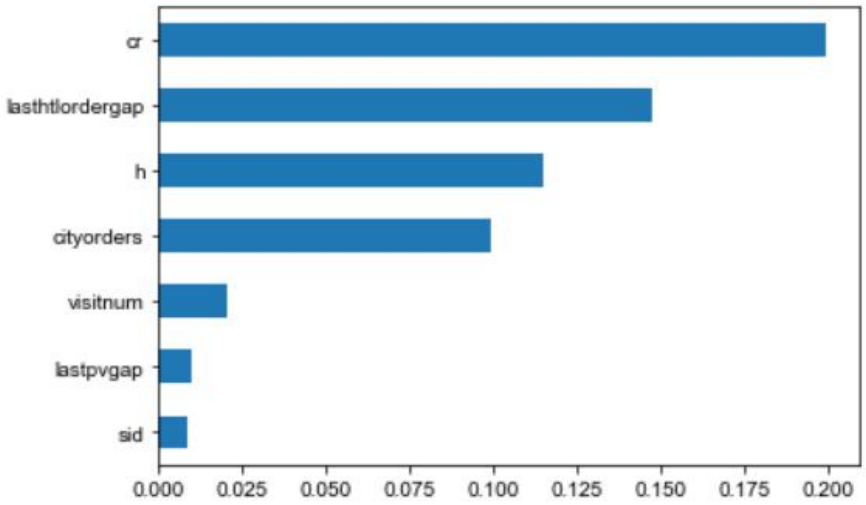
- 再对相邻的组进行卡方检验，并对卡方检验的 P 值很大的组进行合并，直到数据中的组数小于设定的 N 箱为止。

- 让一个特征分别分成[2,3,4.....20]箱，观察每个分箱个数下的 IV 值如何变化，最终找出最适合的分箱个数。

- 分箱完毕后，计算每个箱的 WOE 值，观察分箱效果。

这些步骤都完成后，我们可以对各个特征都进行分箱，然后观察每个特征的 IV 值，以此来挑选特征。在计算每个变量的 IV 值并排序后，我们绘制了如下的

条形图。通过对比分析并去掉 IV 值最小，即对模型基本没有贡献的两个特征——sid, lastpvgap。



图片 5：特征变量 IV 值统计

特征工程完毕后我们建立了逻辑回归模型，但经测试，发现效果并不理想，ROC 曲线面积只达到 0.69。

4.2.2 随机森林模型

逻辑回归的效果并不是很理想，考虑到要追求模型预测的精确度，我们选择辅以随机森林分类器来进行预测客户流失。

随机森林分类器目的是辅助预测客户流失，因此利用清洗好的数据直接利用网格搜索进行调参数，得到最佳参数组合如下：

最佳参数组合

```
model = RFC(n_estimators=180
            ,max_depth=20
            ,min_samples_leaf=1
            ,min_samples_split=2
            ,random_state=0)
```

经测试，随机森林分类器在保证未出现过拟合的情况下，ROC 曲线面积可达到 0.97，效果远好于逻辑回归模型。

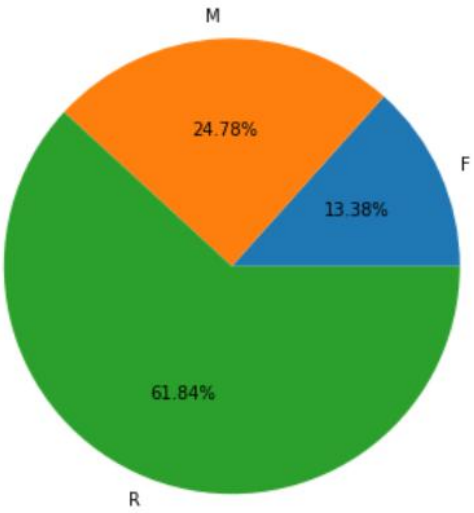
4.3 客户价值评估模型

为了进一步挖掘客户价值，提升用户体验，我们运用了 RFM 客户价值模型，我们定义一年内距离上次下单时长为 R，年订单量为 F，平均价格为 M。

RFM 模型是基于一个固定时间点来做模型分析，因此今天做的 RFM 得分跟 7 天前做的结果可能不一样，原因是每个客户在不同的时间节点所得到的数据不同。

得到 RFM 得分后，可以基于 3 个维度做用户群体划分和解读，对用户价值做分析；也可以基于汇总得分评估所有会员的价值价值，做活跃度排名；还可以作为维度输入和其他维度一起做输入变量，为数据挖掘和分析建模提高基础。

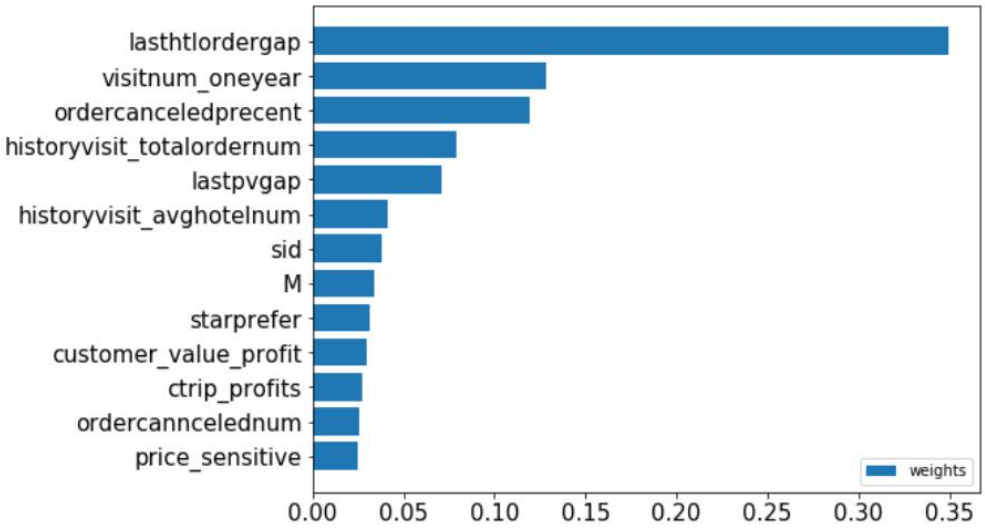
在进行数据预处理过后，我们以客户转化率做目标变量，利用随机森林回归模型计算出各价值指标权重，然后分布计算出每个用户的 RFM 得分，分别以权重加和，及标签组合来表示价值得分。



图片 6：RFM 结果

4.4 客户转化预测模型

由客户流失预测模型的分析结果可知，影响客户流失的两大因素是用户转化率和一年内距离上次下单时长，因此我们以客户转化率为目标标签，进行进一步预测分析。



图片 7：客户转化模型数据预处理结果

经过数据预处理，我们选择出了 13 个对模型贡献度较大特征，可以发现一年内距离上次下单时间对用户转化率的影响最高。而相比对客户流失影响的特征，多了历史订单数，历史取消订单数及星级偏好，客户价值等。

运用网格搜索方法对随机森林分类器进行调参，得到最佳参数如下：

最佳参数组合

```
model = RFR(n_estimators=180
            ,max_depth=25
            ,min_samples_leaf=1
            ,min_samples_split=2
            ,random_state=0)
```

第五章结果与模型评价

5.1 数据集

数据集来源于网络下载。此次数据是携程用户一周的访问数据，数据已经经过脱敏操作。数据规模为 689945*51，数据字段和每个字段代表的含义在表格 1 中列出，考虑到数据大小，这里只给出前三个数据作为示例。

表格 4：样本数据集

字段	解释	data1	data2	data3
label	目标变量	0	1	0
sampleid	样本 id	24636	24637	24641
d	访问日期	2016/5/18	2016/5/18	2016/5/18
arrival	入住日期	2016/5/18	2016/5/18	2016/5/19
iforderpv_24h	24 小时内是否访问订单填写页	0	0	0
decisionhabit_user	决策习惯：以用户为单位观察决策习惯	NULL	NULL	NULL
historyvisit_7ordernum	近 7 天用户历史订单数	NULL	NULL	NULL
historyvisit_totalordernum	近 1 年用户历史订单数	NULL	NULL	NULL
hotelcr	当前酒店历史 cr	1.04	1.06	1.05
ordercanceledprecent	用户一年内取消订单率	NULL	NULL	NULL
landhalfhours	24 小时内登陆时长	22	0	3
ordercanncelednum	用户一年内取消订单数	NULL	NULL	NULL
commentnums	当前酒店点评数	1089	5612	256
starprefer	星级偏好	NULL	NULL	NULL
novoters	当前酒店评分人数	1933	6852	367
consuming_capacity	消费能力指数	NULL	NULL	NULL
historyvisit_avghotelnum	近 3 个月用户历史日均访问酒店数	NULL	NULL	NULL
cancelrate	当前酒店历史取消率	1261	3205	194

字段	解释	data1	data2	data3
historyvisit_visit_detailpagenum	7 天内访问酒店详情页数	NULL	NULL	NULL
delta_price1	用户偏好价格-24 小时浏览最多酒店价格	NULL	NULL	NULL
price_sensitive	价格敏感指数	NULL	NULL	NULL
hoteluv	当前酒店历史 uv	102.607	278.373	16.133
businessrate_pre	24 小时历史浏览次数最多酒店商务属性指数	0.25	0.51	0.61
ordernum_oneyear	用户年订单数	NULL	NULL	NULL
cr_pre	24 小时历史浏览次数最多酒店历史 cr	1.03	1.07	1.12
avgprice	平均价格	NULL	NULL	NULL
lowestprice	当前酒店可定最低价	49	619	312
firstorder_bu	首单 bu	NULL	NULL	NULL
customereval_pre2	24 小时历史浏览酒店客户评分均值	3.2	4.9	3.9
delta_price2	用户偏好价格-24 小时浏览酒店平均价格	NULL	NULL	NULL
commentnums_pre	24 小时历史浏览次数最多酒店点评数	724	5610	4721
customer_value_profit	客户价值_近 1 年	NULL	NULL	NULL
commentnums_pre2	24 小时历史浏览酒店点评数均值	844	3789	4341
cancelrate_pre	24 小时内已访问次数最多酒店历史取消率	0.03	0.21	0.52
novoters_pre2	24 小时历史浏览酒店评分人数均值	1335	5430	5353
novoters_pre	24 小时历史浏览次数最多酒店评分人数	1249	7829	7324
ctrip_profits	客户价值	NULL	NULL	NULL

字段	解释	data1	data2	data3
deltaprice_pre2_t1	24 小时内已访问酒店价格与对手价差均值，t+1	29	-56	8
lowestprice_pre	24 小时内已访问次数最多酒店可订最低价	46	111	413
uv_pre	24 小时历史浏览次数最多酒店历史 uv	58.027	249.347	133.093
uv_pre2	24 小时历史浏览酒店历史 uv 均值	74.956	224.92	112.063
lowestprice_pre2	24 小时内已访问酒店可订最低价均值	615	513	382
lasthtlordergap	一年内距离上次下单时长	NULL	NULL	NULL
businessrate_pre2	24 小时内已访问酒店商务属性指数均值	0.29	0.53	0.6
cityuvs	昨日访问当前城市同入住日期的 app uv 数	12.88	17.933	3.993
cityorders	昨日提交当前城市同入住日期的 app 订单数	3.147	4.913	0.76
lastpvgap	一年内距上次访问时长	NULL	NULL	NULL
cr	用户转化率	NULL	NULL	NULL
sid	会话 id，sid=1 可认为是新访	7	33	10
visitnum_oneyear	年访问次数	NULL	NULL	NULL
h	访问时间点	12	14	19

5.2 用户画像

我们总结出易流失人群和留存客户人群特征，对客户进行一个简单的画像。针对易流失人群，推荐具体业务可以从时间维度、数量维度和价格维度三个维度将本次分析结果落地。

表格 5：用户画像

易留存人群特征	易流失人群特征
➢ 一年内距上次下单时长在（1，1.075）区间 ➢ 用户转化率在（ 1，1.075 ）区间 ➢ 访问时间在晚上 ➢ 订单数在 2.294 以下 ➢ 年访问次数超过 15003 ➢ 年消费越小越容易留存 ➢ 入住日期与访问日期间隔越长	➢ 一年内距上次下单时长在（2.5，1327）区间 ➢ 用户转化率在（1.505，1.925）区间 ➢ 访问时间在上午 ➢ 订单数在 2.61 以上 ➢ 年访问次数在小于 15000 ➢ 年消费越大越容易流失 ➢ 入住日期与访问日期间隔越短

5.3 客户流失预测模型评估

在进行数据预处理后，我们建立了逻辑回归模型，并计算了召回率，假正率，得到了训练集和测试集的分数。

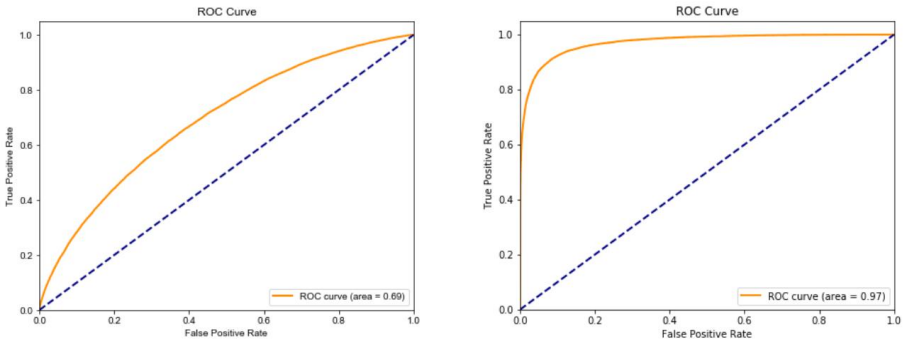
表格 5：逻辑回归模型评估指标

acu	accuracy	precision	recall	fl
0.676	0727	0.741	0.951	0.833

表格 6：逻辑回归模型训练和测试分数

	训练集	测试集
分数	0.728283	0.726898

由于逻辑回归模型所测得的结果不是非常理想，考虑到要追求模型预测的精确度，我们辅以随机森林分类器来进行预测客户流失。发现经改进，ROC 曲线面积从 0.69 到了 0.97

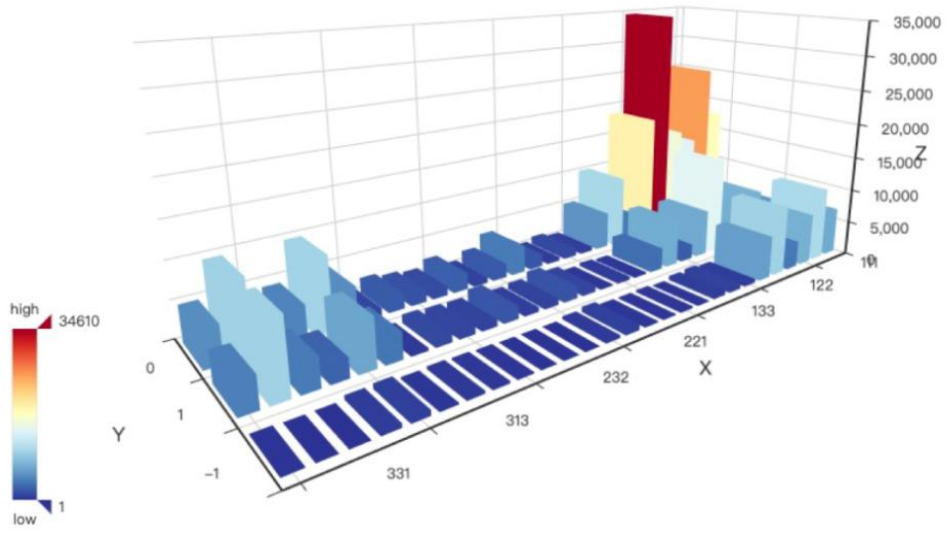


逻辑回归模型预测分数　　随机森林模型预测分数

图片 8：客户流失预测模型评价结果

5.4　客户价值模型评估

为了使结果更直观，我们对其进行了可视化，我们可以看到在以 R 就是一年内距离上次下单时长为轴，其两端留存和流失客户均很多，且 R 等于 1 就是距离时间越久，客户流失就越少，与我们用 WOE 分箱分析结果一致。



图片 9：客户价值模型可视化结果

图中标签 0 是留存用户，标签 1 是流失用户，标签-1 是留存减流失乘以优势比。

5.5　客户转化预测模型评估

该模型在测试集预测得分 93.16%，可以保存模型以供模型部署使用。

第六章 总结

本文中主要针对携程用户的数据做出分析，就客户行为做出流失分析和偏好分析。我们首先对数据进行预处理，将较大的数据集“小化”，使其在减小数据规模的同时，保留重要特征，做到使用具有代表性数据集来反映整体内容的效果。在做完数据预处理之后，我们就这些经过处理的数据，从客户流失、客户价值和客户转化三个方面对客户流失及客户行为偏好进行分析。我们选择了逻辑回归模型和随机森林模型作为客户流失预测模型，并通过绘制 ROC 曲线证明随机森林模型是更好的客户流失预测模型。我们使用 RFM 模型作为客户价值评估模型，同时应用以客户转化率为目标标签的客户转化预测模型评估，并且证明这两种模型在使用时会有比较好的效果。

第七章 任务分工

姓名	完成部分
张欢	数据集查找、数据预处理
杨玲	模型构建及模型评价

参考文献：

[1]. Pal, M. (2005). Random forest classifier for remote sensing classification. International journal of remote sensing, 26(1), 217-222.

[2]. Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS journal of photogrammetry and remote sensing*, *114*, 24-31.

[3]. Wei, J. T., Lin, S. Y., & Wu, H. H. (2010). A review of the application of RFM model. African Journal of Business Management, 4(19), 4199-4206.

[4]. Cheng, C. H., & Chen, Y. S. (2009). Classifying the segmentation of customer value via RFM model and RS theory. Expert systems with applications, 36(3), 4176-4184.