运营商流失用户的分析和预测

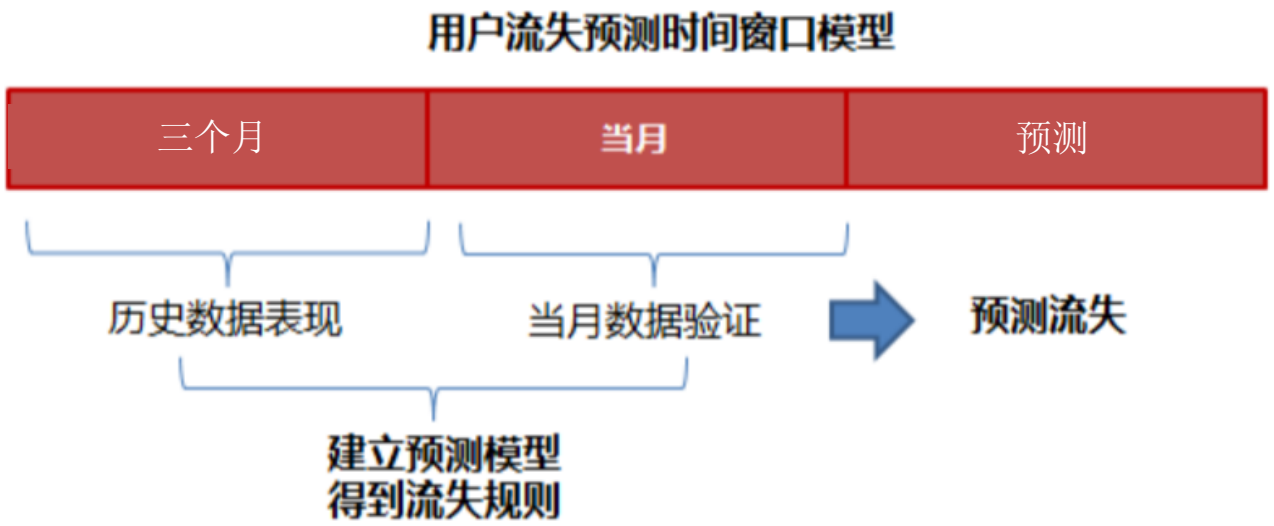
目录

背景

随着业务的快速发展、移动业务市场的竞争愈演愈烈。如何最大程度地挽留在网用户、吸取新客户，是电信企业最关注的问题之一。竞争对手的促销、公司资费软着陆措施的出台和政策法规的不断变化，影响了客户消费心理和消费行为，导致客户的流失特征不断变化。对于电信运营商而言，流失会给电信企业带来市场占有率下降、营销成本增加、利润下降等一系列问题。在发展用户每月增加的同时，如何挽留和争取更多的用户，是一项非常重要的工作。

背景

- 随着大数据挖掘技术的不断发展和应用，移动运营商希望能借助数据挖掘技术识别哪些用户可能流失，什么时候会发生流失。而通过建立流失预测模型，分析用户的历史数据和当前数据，提取辅助决策的关键性数据，并从中发现隐藏关系和模式，进而预测未来可能发生的行为，就可以帮助移动运营商实现这些要求。



数据说明

名称	字段描述
MONTH_ID	月份
USER_ID	用户ID
INNET_MONTH	在网时长
IS_AGREE	是否合约有效用户
AGREE_EXP_DATE	合约计划到期时间
CREDIT_LEVEL	信用等级
VIP_LVL	VIP等级
ACCT_FEE	本月费用（元）
CALL_DURA	通话时长(秒)
NO_ROAM_LOCAL_CALL_DURA	本地通话时长(秒)
NO_ROAM_GN_LONG_CALL_DURA	国内长途通话时长(秒)
GN_ROAM_CALL_DURA	国内漫游通话时长(秒)
CDR_NUM	通话次数（次）
NO_ROAM_CDR_NUM	非漫游通话次数（次）
NO_ROAM_LOCAL_CDR_NUM	本地通话次数（次）
NO_ROAM_GN_LONG_CDR_NUM	国内长途通话次数（次）
GN_ROAM_CDR_NUM	国内漫游通话次数（次）
P2P_SMS_CNT_UP	短信发送数（条）
TOTAL_FLUX	上网流量(MB)
LOCAL_FLUX	本地非漫游上网流量(MB)
GN_ROAM_FLUX	国内漫游上网流量(MB)

运营商用户的基础信息和使用行为信息

user_info_m.csv。

90W条记录，30W用户3个月的数据。

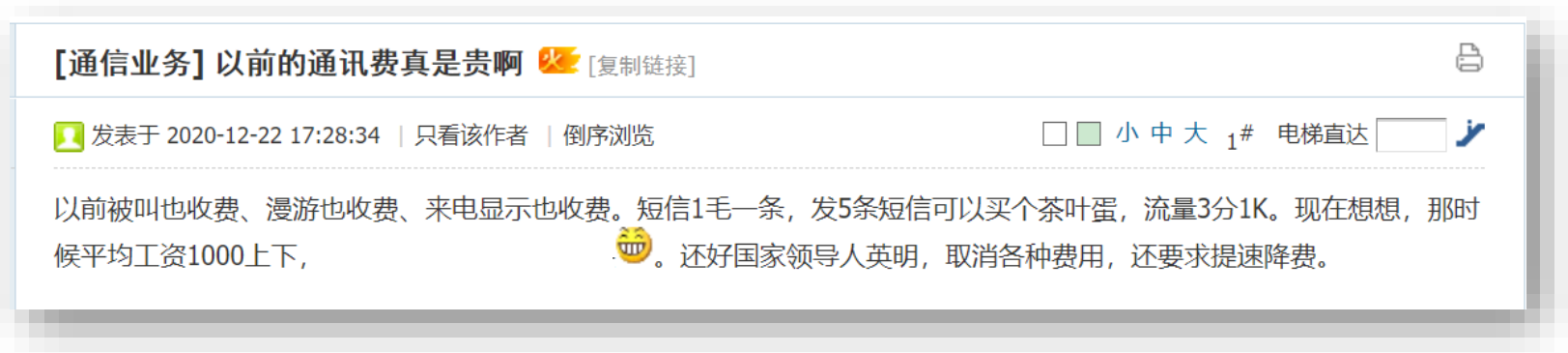
35个特征。

数据说明

名称	字段描述
CALL_DAYS	有通话天数
CALLING_DAYS	有主叫天数
CALLED_DAYS	有被叫天数
CALL_RING	语音呼叫圈
CALLING_RING	主叫呼叫圈
CALLED_RING	被叫呼叫圈
CUST_SEX	性别
CERT_AGE	年龄
CONSTELLATION_DESC	星座
MANU_NAME	手机品牌名称
MODEL_NAME	手机型号名称
OS_DESC	操作系统描述
TERM_TYPE	终端硬件类型(4=4g、3=3g、2=2g)
IS_LOST	用户在3月是否流失标记（1=是，0=否），1月和2月值为空

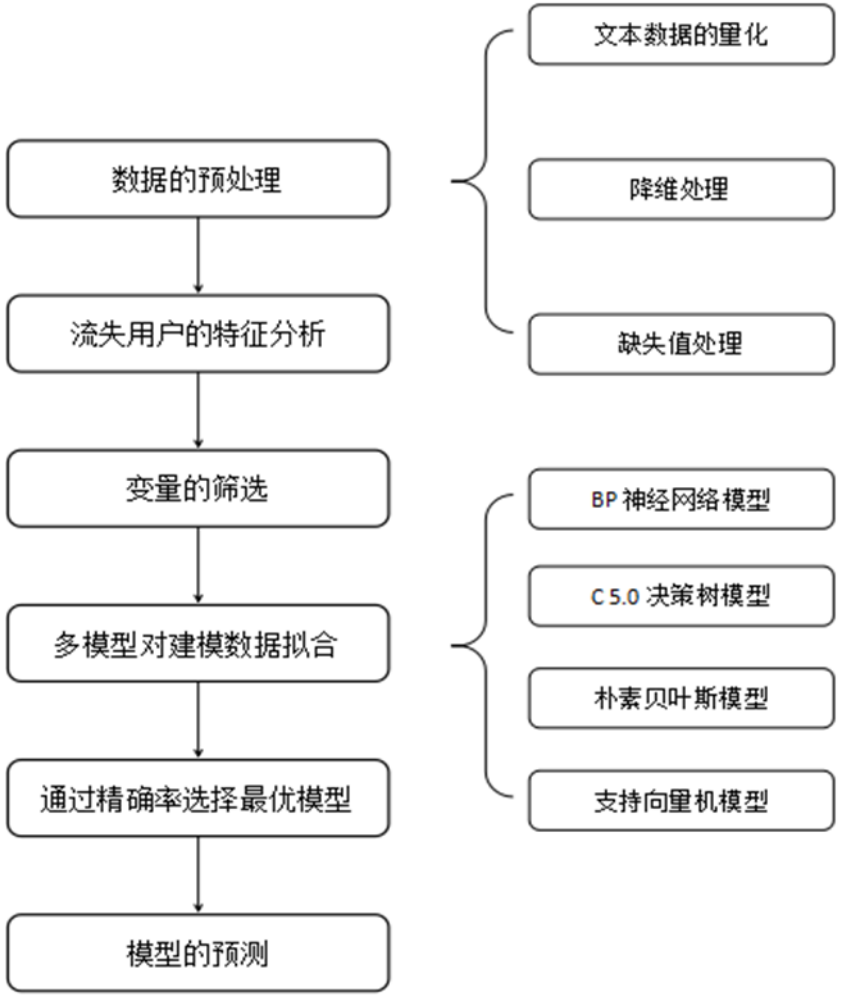
数据说明

- 漫游：指移动电话用户在离开开户省区或本国时，仍可以在其他一些省区或国家继续使用他们的移动电话的一种业务。2017年9月1日起，三大运营商中国电信、中国联通、中国移动全面取消国内手机长途和漫游通话费（不含港澳台），所以用户拨打国内电话是按套餐标准资费进行收取的。
- 高昂的网络运行维护成本、营改增、国资委考核、村村通以及扶持TD-LTE的大背景下，通信企业再次降价的空间十分有限。

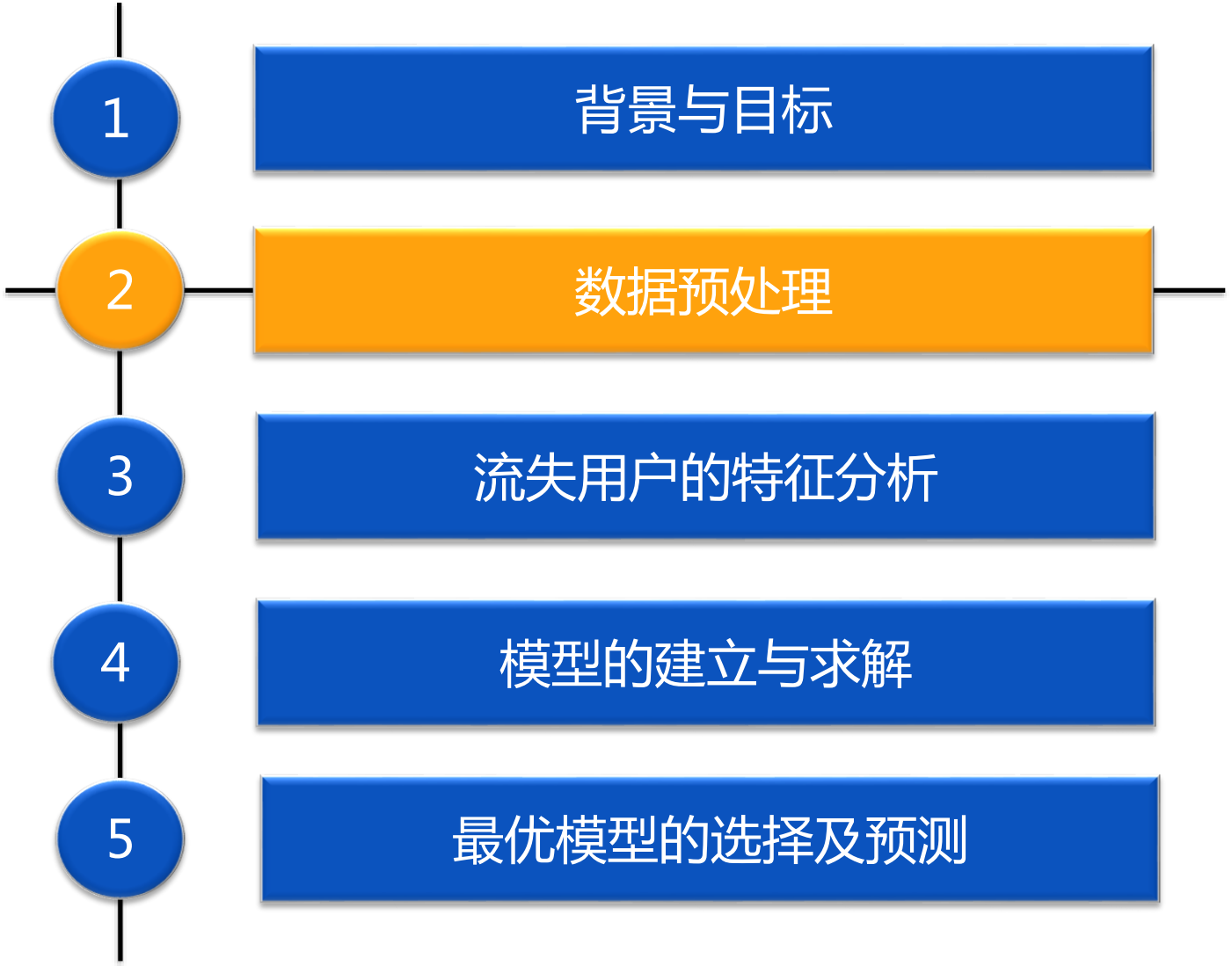


目标

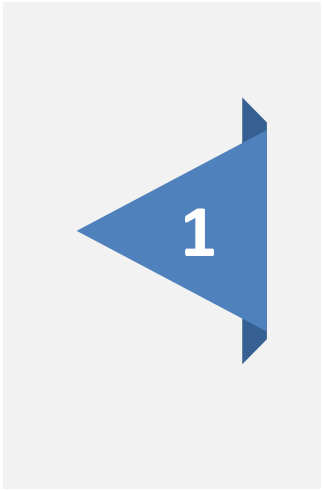
- 利用用户的短信、流量、通话、消费的使用情况以及客户基本信息的数据，采用数据挖掘技术对用户进行分群。
- 分析不同群体用户的使用规律，识别各群体客户流失的重要特征。
- 建立不同群体用户流失模型，结合结果为运营商提供差异化的意见和建议。



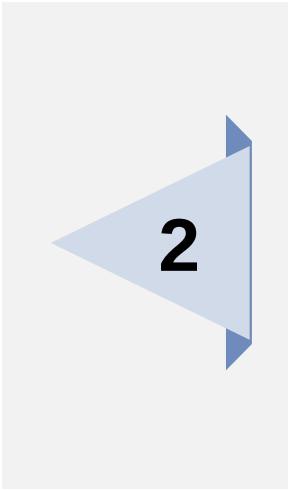
目录



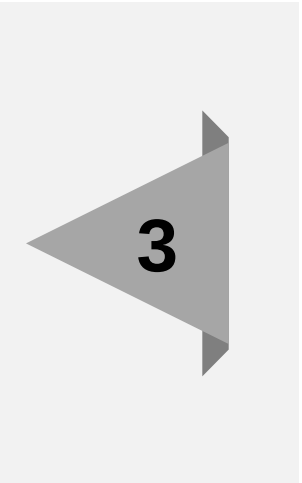
数据预处理



重复数据处理

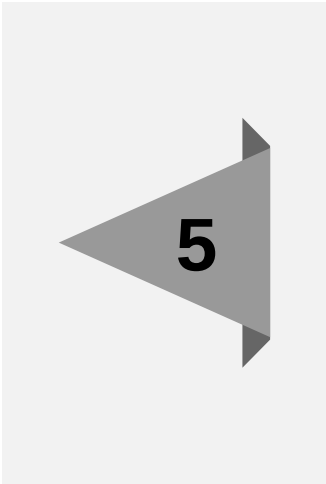


降维



数据的提取和整合

缺失值的处理



异常值的处理

查找并删除重复数据

- 本文采用python在900000条原始建模数据中查找出重复的用户信息，如表所示。

月份	用户ID	在网时长	有否合约	合约到期时间	信用等级	VIP等级	本月费用	通话时长	…
201601	U3114071724581571	61	0		67		165.1	6343	…
201601	U3114071724581571	61	0		67		165.1	6343	…
201601	U3114071324562424	19	0		66	4	220	27921	…
201601	U3114071324562354	19	0		65	4	124.5	1205	…
201601	U3114071224555356	19	1	201607	65	4	116	6611	…
201601	U3114071224554921	43	0		65	99	30	6022	…

查找并删除重复数据

- 由于使用重复数据会对下面的分析和建模产生不利影响，因此，将所有重复数据视为异常值剔除。删除重复个案的数据如表所示。

月份	用户ID	在网时长	有否合约	合约到期时间	信用等级	VIP等级	本月费用	通话时长	…
201601	U3114071324562424	19	0		66	4	220	27921	…
201601	U3114071324562354	19	0		65	4	124.5	1205	…
201601	U3114071224555356	19	1	201607	65	4	116	6611	…
201601	U3114071224554921	43	0		65	99	30	6022	…

降维及数字化处理

降维处理原始数据的变量有33条（不包含用户ID和“用户在三月是否流失”），变量较多。其中，与用户手机情况相关的数据均为文本数据。

原始数据（部分数据）如表所示。

月份	用户ID	…	星座	手机品牌	手机型号	操作系统描述
201601	U3115020568568480	…	金牛座	小米	2012061	ANDROID 4.1.1
201601	U3115071634539940	…	巨蟹座	赫比	YD201	ANDROID
201601	U3114071224554830	…	处女座	苹果	A1700	IOS 9.0
201601	U3115061833296551	…	白羊座	华为	PE-TL10	ANDROID 4.4
201601	U3114112267516766	…	天蝎座	华为	PE-TL10	ANDROID 4.4
201602	U3115050669440719	…	水瓶座	锤子	SM705	ANDROID 4.4.2

数据的提取与整合

- 在原始的建模数据中，每个用户有三行的数据，对应着同一个ID三个月的信息。这种数据格式不便于进行建模和分析。因此，本文对数据进行提取和整合，将每个用户信息处理为一行数据。具体的处理方式如下:

（1） 移动客户基本信息：由于同一个用户的基本信息在三个月内是相同的。因此，用户ID、性别、星座、年龄以及终端硬件类型只取1月的数据。

（2） 在网时长：由于每个月的在网时长等于上个月的在网时长加一（为零的除外），故第三个月的在网时长的数据就包含了这三个月在网时长的全部信息。因此，每个ID只提取第三个月的数据。

数据的提取与整合

（3）是否合约有效:

将这三个月的合约有效的情况规整为一个数据，处理如下：当三个月不全为1时，用第三个月的值减去前两个的均值；当三个月的值都是为1时，取值为1.5。所有取值情况为-1、-0.5、0、0.5、1、1.5。

数值	含义
-1	前两个月合约有效，第三个月合约无效
-0.5	前两个月其中一个月合约有效，另一个月和第三个月合约无效。
0	三个月均合约无效
0.5	前两个月其中一个月合约无效，另一个月和第三个月合约有效
1	前两个月合约无效，第三个月合约有效
1.5	三个月均合约有效

数据的提取与整合

（4）合约计划到期时间：取第三个月的数据作为合约计划到期时长，将空值赋值为-1，不是空值的到期时间以201603为基准0，每增加一个月，数据为1，例如日期为201607，那么该月份的到期时间为4，以此类推。

（5）信用等级：取3个月的平均值。

（6）VIP等级：将空值赋值为0 ，同一个ID值，三个月的数值相等，则取第三个月的数值；三个月的数值都不相等，则第三个月的数据减去前两个月的均值。这样的处理可以突出用户在第三个月的变化情况。

数据的提取与整合

（7）本月费用：取三个月的平均值。

（8）平均每次通话时长：将各种通话时长除以通话次数，得到各类的平均通话时长。

（9）其余的变量处理方式相同：同一个ID下，取三个月的平均值。

数据的提取与整合

- 经上述规则提取和整合后，得到的新数据集。

用户ID	在网时长	性别	年龄	星座	终端硬件	合约到期时间	信用等级	是否合约有效	VIP等级	费用	…
U3114031824148707	25	1	22	12	3	-1	67	0	99	-4.575	…
U3114031824148874	25	2	30	4	4	0	65	1.5	99	79.22	…
U3114031824148975	25	2	42	2	4	0	65	1.5	99	-0.05	…
U3114031824149138	25	2	22	5	3	-1	65	0	99	66.15	…
U3114031824149150	25	1	30	10	4	-1	65	0	99	-7	…

缺失值与异常值处理

- 缺失值处理

用python对新数据集进行缺失值查找，共发现存在缺失值的数据14645条，缺失值集中在用户的基本信息——星座、年龄和性别上。对缺失的数据进行赋值，性别缺失的赋值为3，年龄缺失的赋值为0，星座缺失的赋值为0。

- 异常值处理

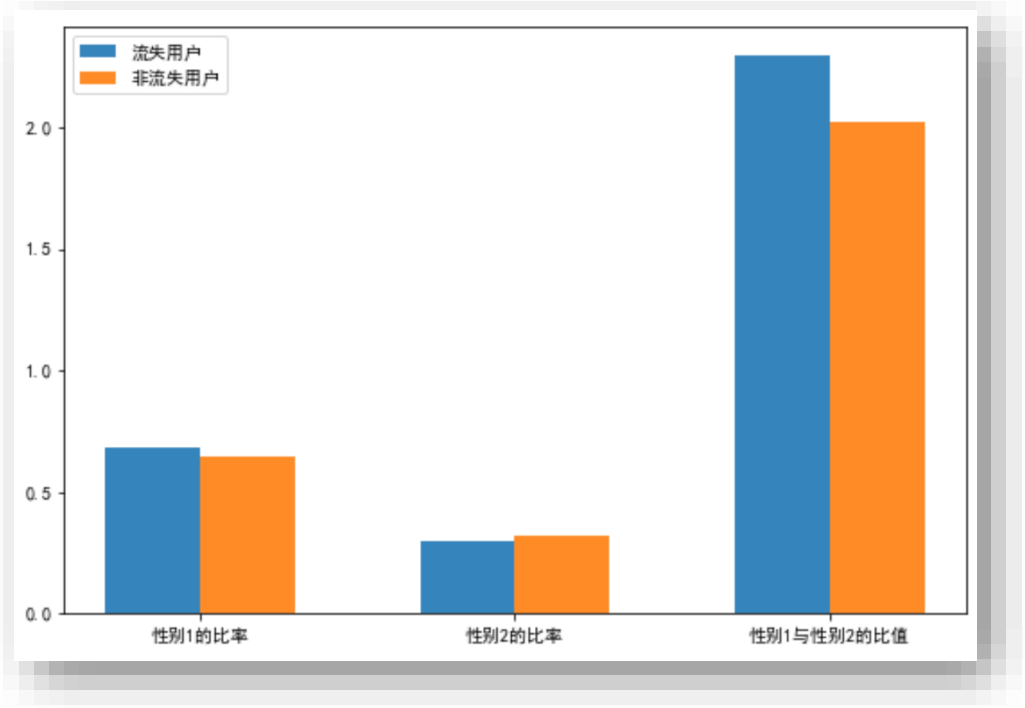
对新数据集进行异常值的查找，查找出在网时长小于0的数据，费用大于4万元的数据。这些异常数据有很大的可能是因为人为录入时出现失误，直接利用异常数据建模会导致模型结果与实际情况出现严重的偏差。

目录

基本信息分析

性别分析

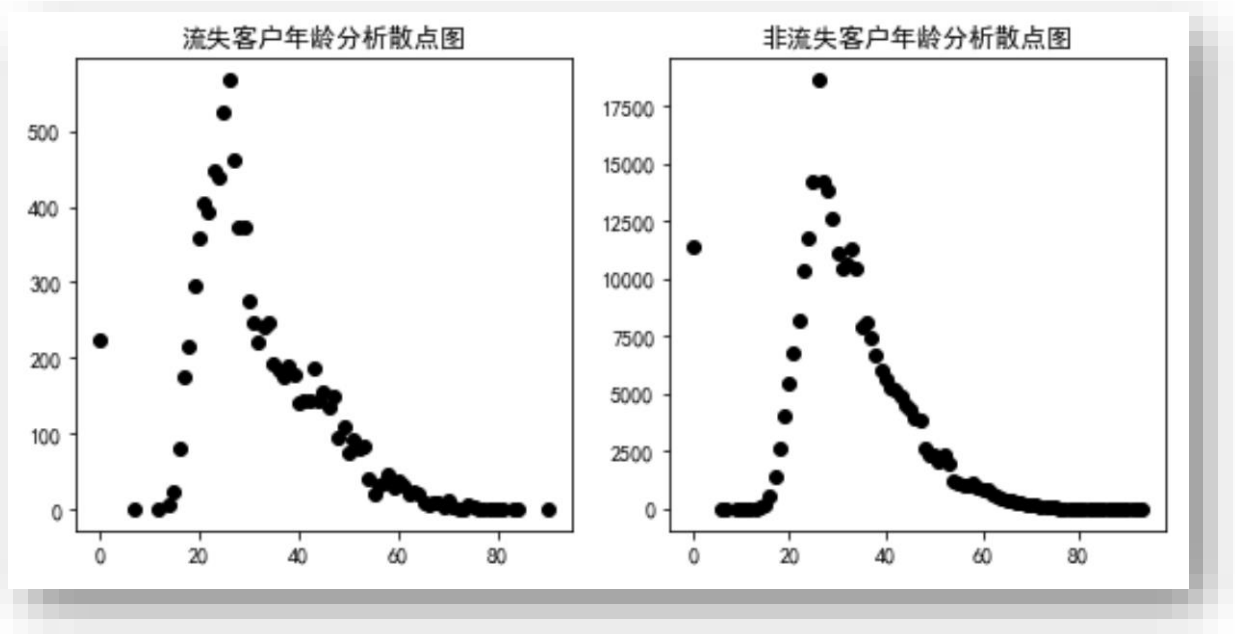
性别为1和性别为2占所在群体中比例的差别不大，但从性别1与性别的比例中可以看到，流失用户群体中性别的比值要比非流失用户群体中性别的比值相对更大，即虽然流失用户与非流失用户中的性别比例相当，但是在流失用户中，性别1比性别2的用户更易流失。



基本信息分析

年龄分析

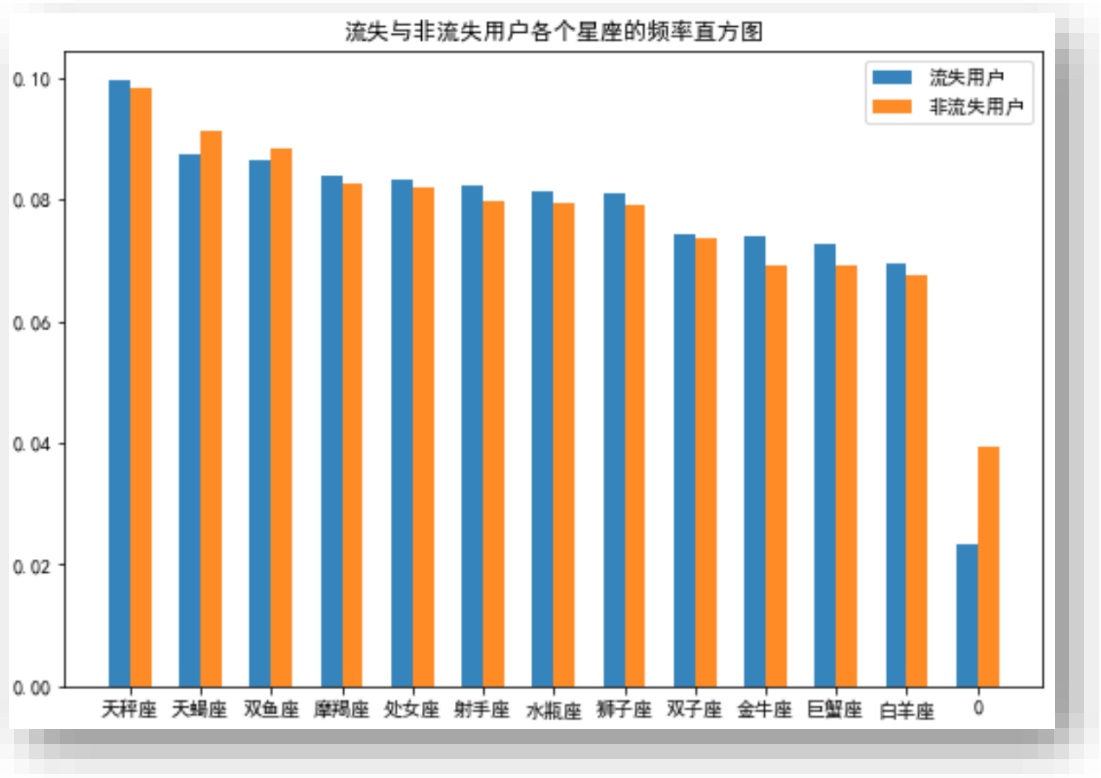
流失用户群体与非流失用户群体各个年龄使用的人数呈偏态分布，其中年龄为26岁的用户是最多的。流失用户群体与非流失用户群体在各个年龄上均有用户流失，由于两个群体关于年龄的特征相似，因此，认为年龄与用户是否在3月流失无重要关系。



基本信息分析

星座分析

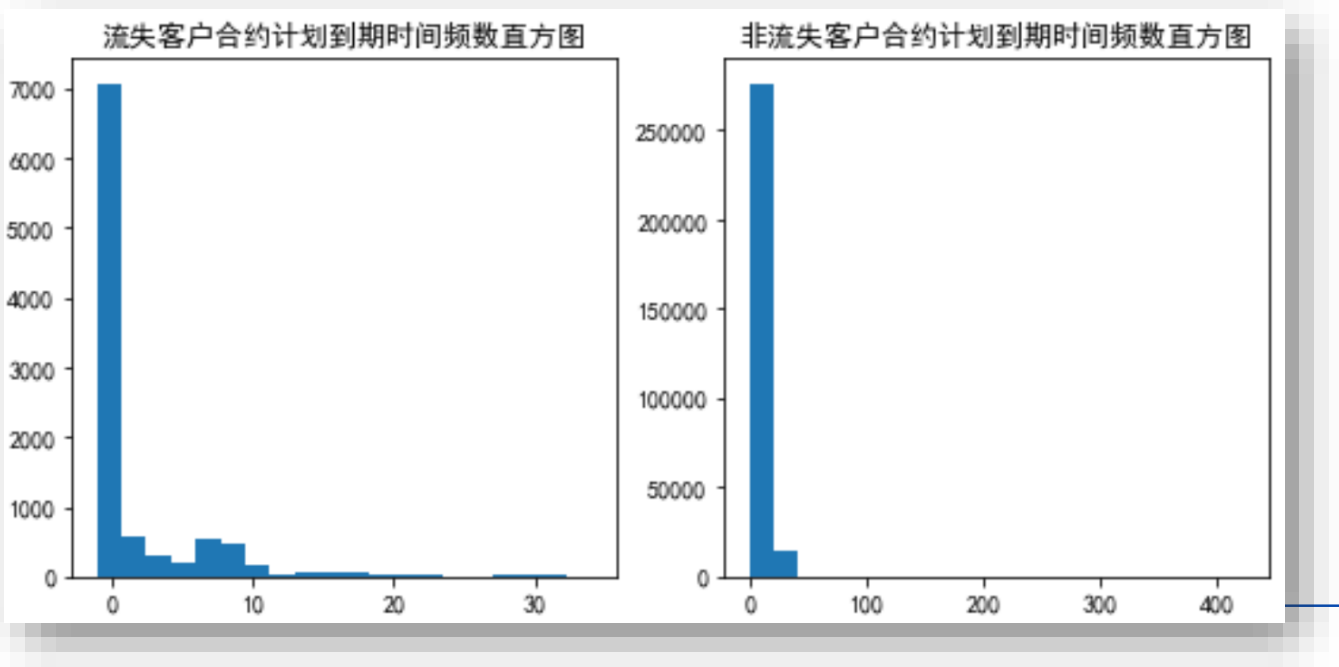
- 流失用户群体与非流失用户群体各个星座占所在群体的比例相当，因此星座的特征在流失用户群体以及非流失用户群体不明显，认为星座与用户是否流失无重要关系。



基本信息分析

在网时长分析

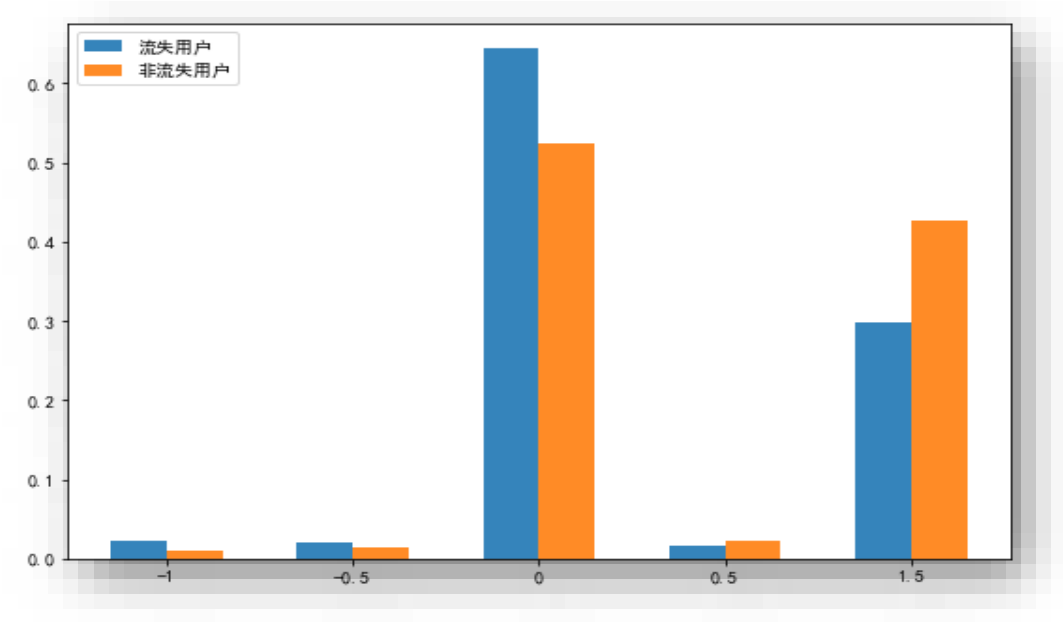
在流失客户中，在网时长高于20个月的用户数都不超过100，而大多数流失用户的在网时长都低于20个月，数量占了流失客户的85.19%；在非流失客户中，有44.95%的用户在网时长低于20个月；流失用户在网时长低于20个月的比例约为非流失用户的2倍，因此可以认为在网时长对用户是否在3月流失有重要的影响，同时，在网时长低于20个月的用户易于流失。



基本信息分析

合约是否有效

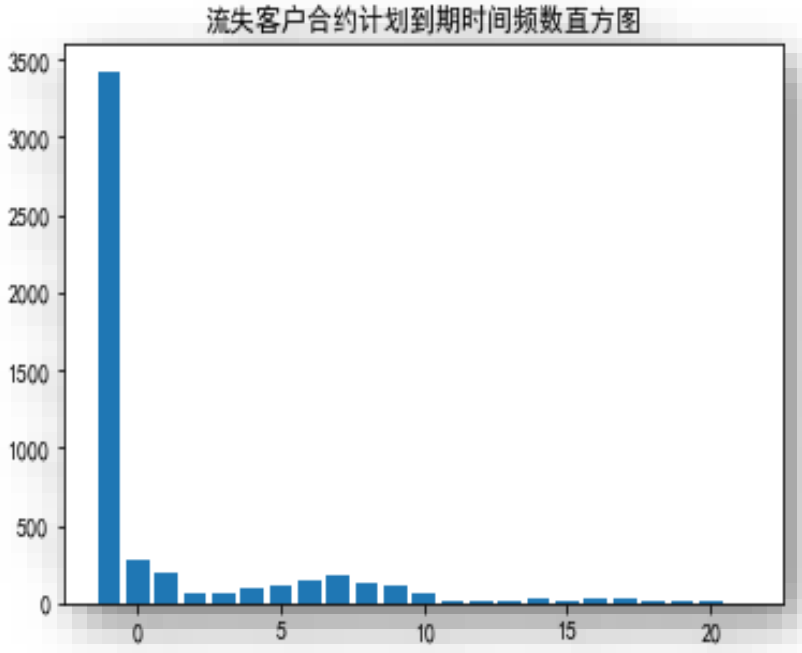
- 可以知道无效用户占流失与非流失用户群体的大多数，有效值为0.5占流失用户和非流失用户的比率很少，即只有少数人在前2个月有一个月是签订合同的，第3个月仍签订合同，成为有效用户；观察到流失和非流失用户中均没有有效值为1（只有第3个月签订合同，成为有效用户）。因此，无效用户更容易流失。

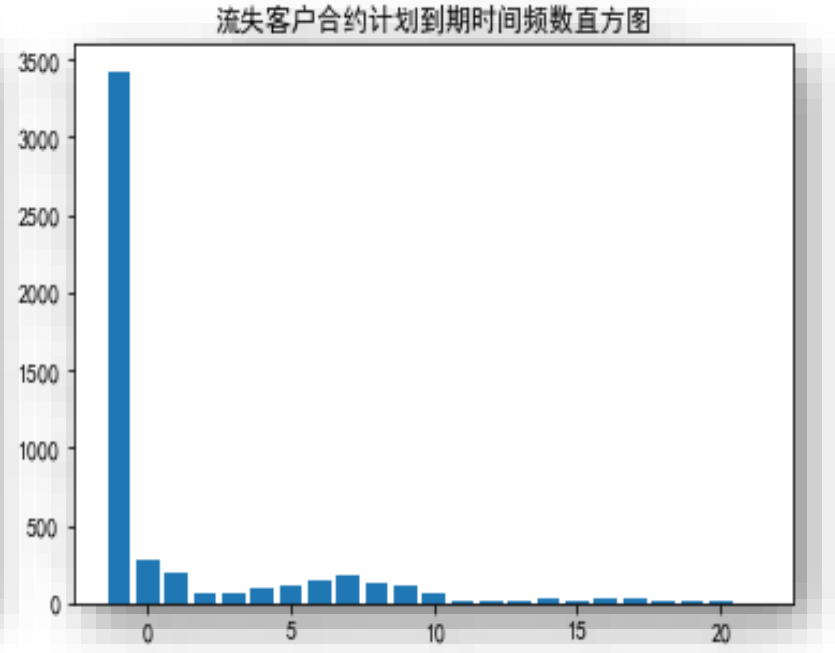


基本信息分析

合约到期时间分析

- 无论是流失用户群体与非流失用户群体，用户数量最大的都为-1到9月，9月到20个月时数量减少，30个月以上在所在群体占的比重不大。而其中合约计划到期时间为-1（没有签订合同）占的比例最大都是，即流失用户群体与非流失用户群体中大部分都是无效用户。

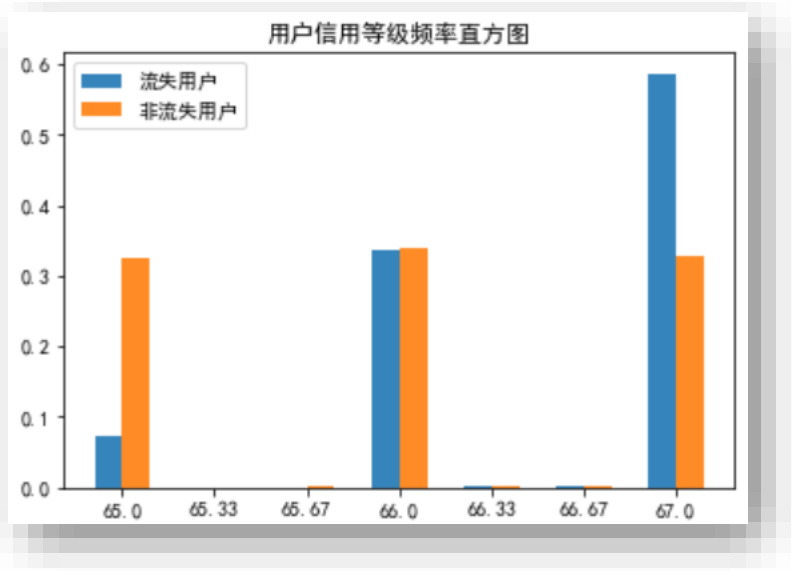




基本信息分析

信用等级分析

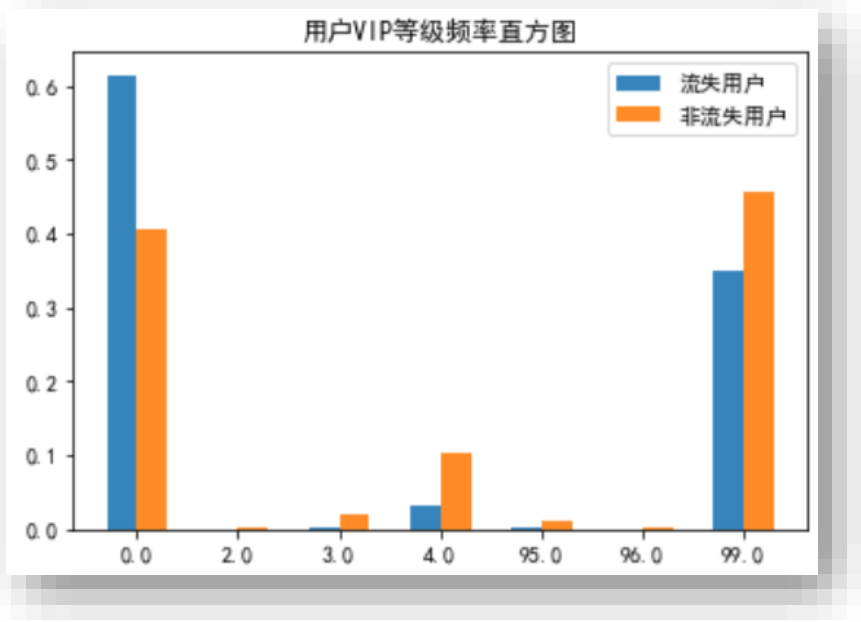
- 不管是流失用户还是非流失用户，他们在三个月中的信用等级普遍是不变的，在非流失用户中，三个信用等级的比例相当，而在流失用户中，信用等级为67的用户占过半的比例，信用等级为65的用户占比例较少。因此，可认为信用等级是用户流失的一个特征变量，信用越高的用户反而越容易流失。



基本信息分析

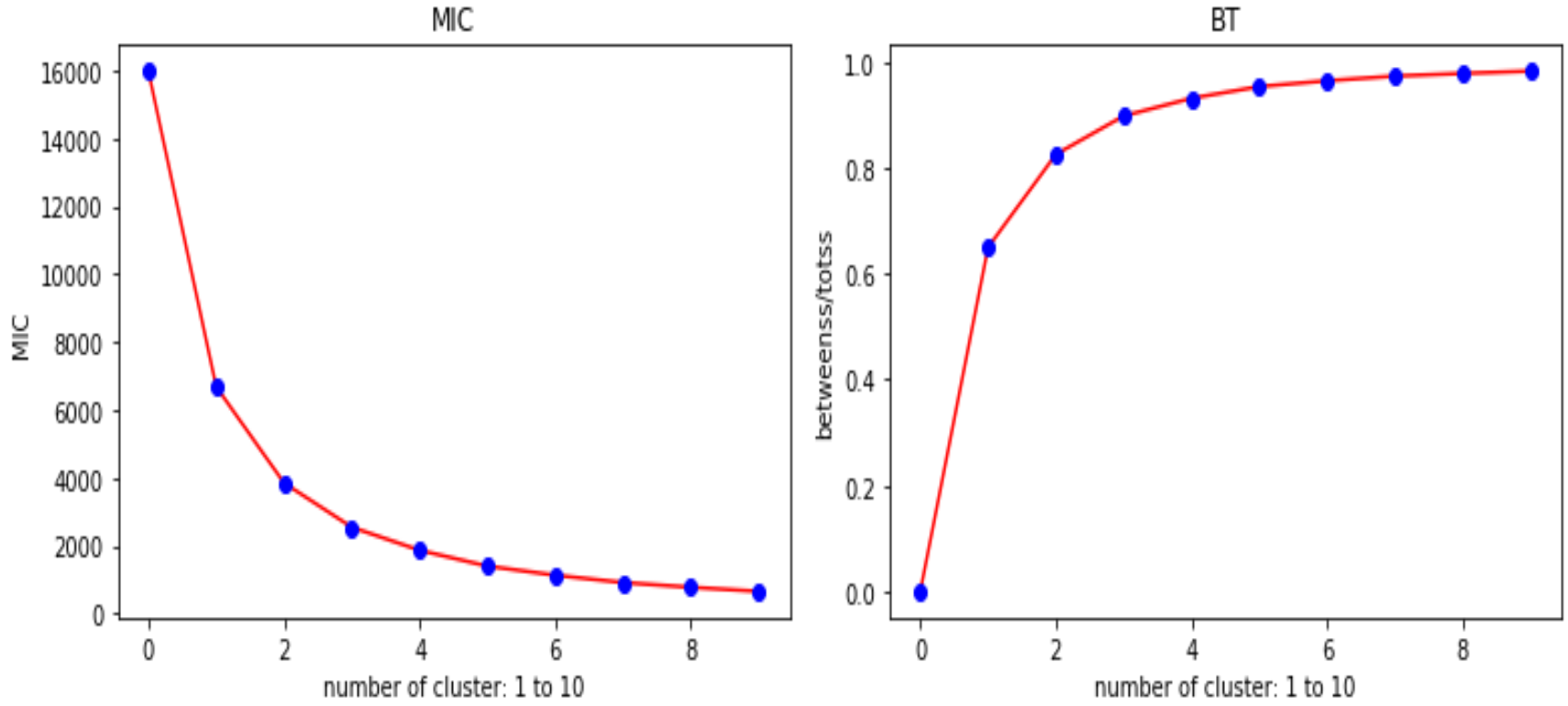
VIP等级分析

- 流失用户中VIP等级缺失的情况比非流失的更多，且VIP等级为4的用户更不容易流失。因此，可认为VIP等级是用户流失的一个特征变量。



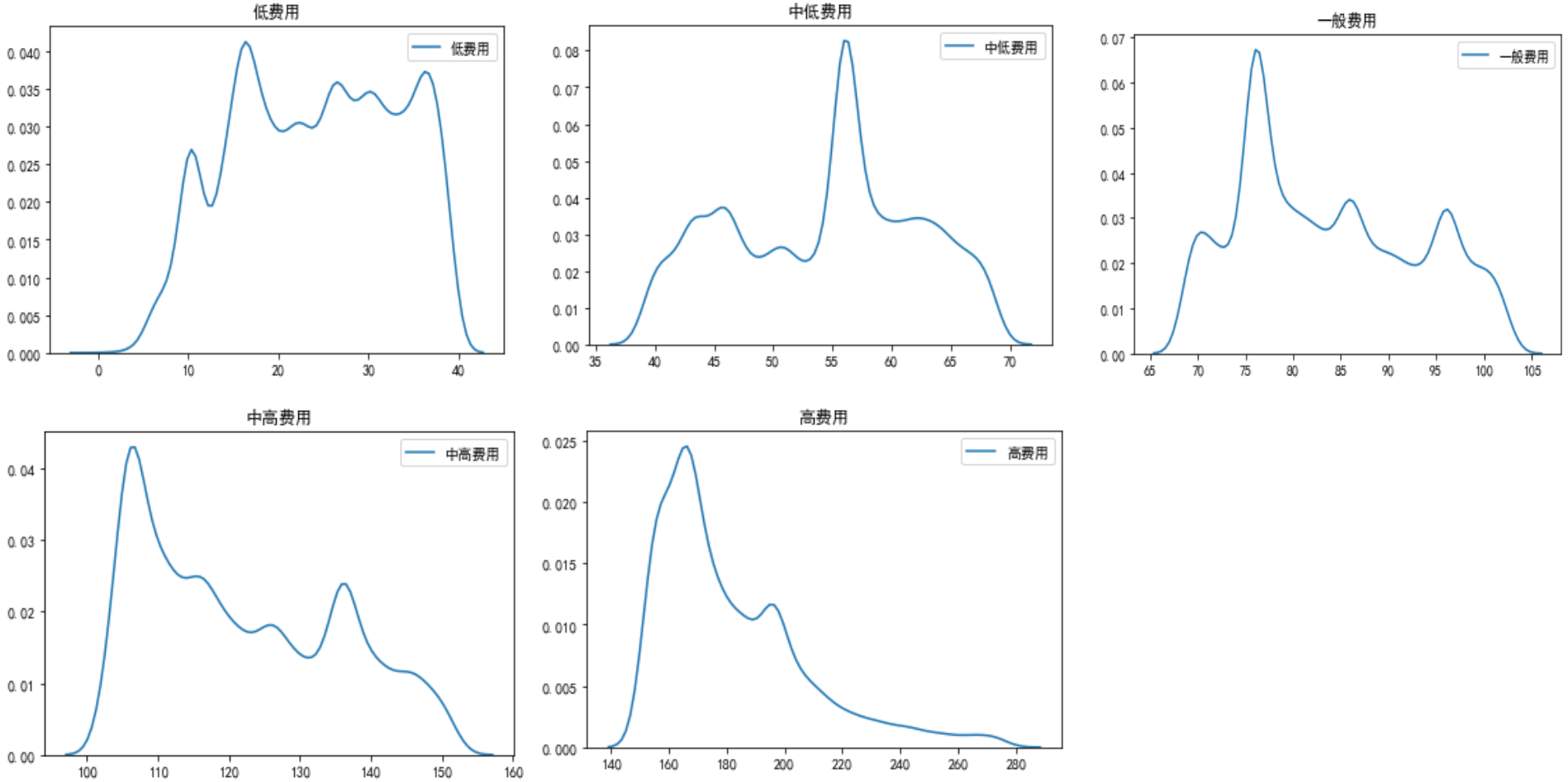
K-Means聚类分析

- 首先绘制聚类优度的选择图，如图所示。越多的聚类数目，能够体现每一类越多的特征。在聚类数目取5时，特征的体现程度已经达到90.0%，而且随着聚类数目的增加，体现程度的增加幅度愈加缓慢。



K-Means聚类分析

- 把本月话费的平均值分成五类，以观察各类用户的具体情况。



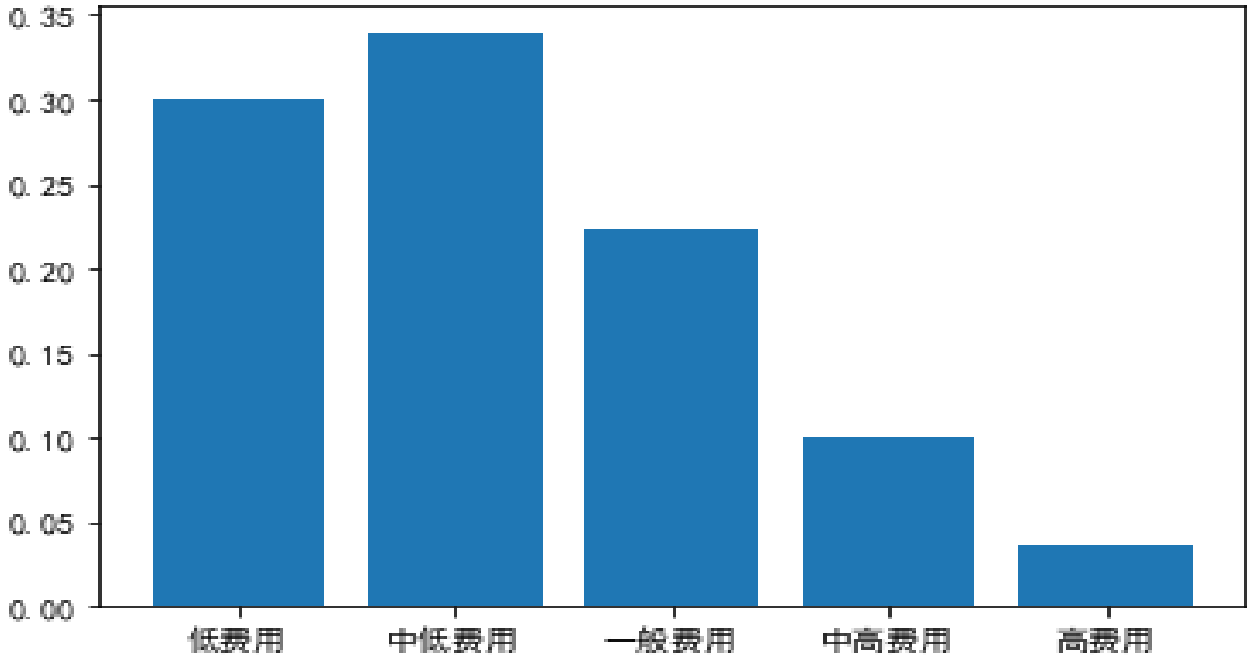
K-Means聚类分析

- 根据实际情况和其数据特征为这五类分别命名，如表所示。

输出类型	First	Second	Third	Fourth	Fifth
命名	中高费用	高费用	低费用	中低费用	一般费用

K-Means聚类分析

- 从图中可以看到，中低费用与低费用类的流失用户群体都分别约占了所有流失用户的40%，即中低费用和低费用类别的流失用户占了所有流失用户的80%左右，即低费用人群更容易流失。因此，以下针对低费用以及中低费用人群进行通话记录信息等信息的特征分析。



特征值的提取

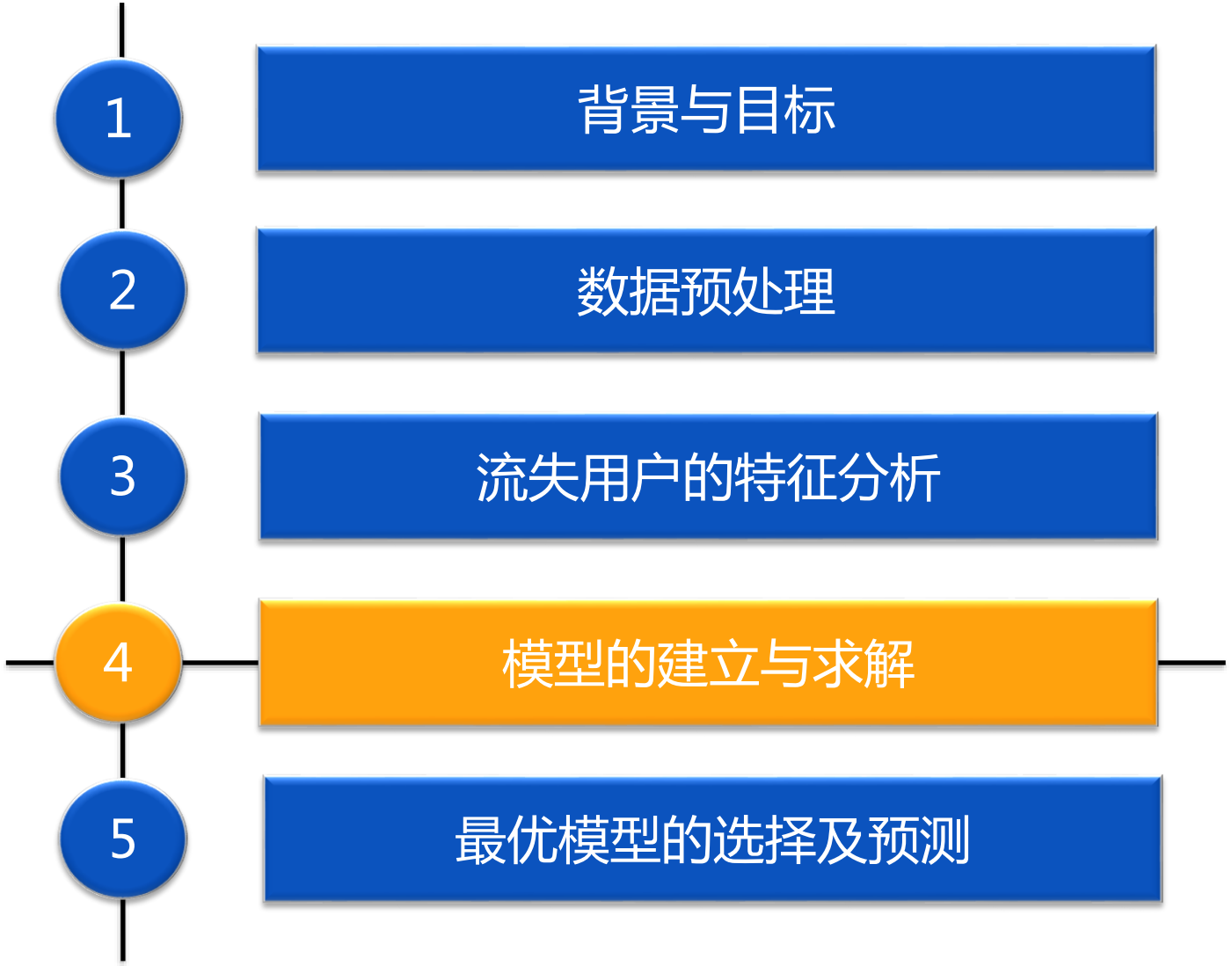
- 特征抽取是数据挖掘任务最为重要的一个环节，一般而言，它对最终结果的影响要高过数据挖掘算法本身。只有先把现实用特征表示出来，才能借助数据挖掘的力量找到问题的答案。特征选择的另一个优点在于：降低真实世界的复杂度，模型比现实更容易操纵。
- 特征选择的原因有：(1)降低复杂度；(2)降低噪音；(3)增加模型可读性。
- 单个特征和某一类别之间相关性的计算方法有很多，比较有效的有卡方检验（chi2）以及互信息和信息熵，本文我们选择基于信息熵的方法来选取特征变量。
- 信息熵是在决策树中广泛使用的一个变量，用以获取最优划分的节点。基于树的预测模型能够用来计算特征的重要程度，因此能用来去除不相关的特征。
- 因此选择基于树的特征选择(Tree-based feature selection)来获取特征变量。

特征值的提取

- 选取重要程度前10的变量为特征变量，得到的结果如表所示。

	选取的变量
低费用	'CERT_AGE', 'INNET_MONTH', 'ACCT_FEE', 'NO_ROAM_LOCAL_CDR_NUM', 'NO_ROAM_GN_LONG_CDR_NUM', 'GN_ROAM_CDR_NUM', 'CALL_DAYS', 'CALLING_DAYS', 'CALLED_DAYS', 'CALLING_RING'
中低费用	'CERT_AGE', 'INNET_MONTH', 'ACCT_FEE', 'NO_ROAM_LOCAL_CDR_NUM', 'NO_ROAM_GN_LONG_CDR_NUM', 'GN_ROAM_CDR_NUM', 'TOTAL_FLUX', 'CALL_DAYS', 'CALLING_DAYS', 'CALLED_DAYS'
一般费用	'CERT_AGE', 'INNET_MONTH', 'NO_ROAM_LOCAL_CDR_NUM', 'NO_ROAM_GN_LONG_CDR_NUM', 'GN_ROAM_CDR_NUM', 'TOTAL_FLUX', 'LOCAL_FLUX', 'CALL_DAYS', 'CALLING_DAYS', 'CALLED_DAYS'
中高费用	'CERT_AGE', 'NO_ROAM_LOCAL_CDR_NUM', 'NO_ROAM_GN_LONG_CDR_NUM', 'GN_ROAM_CDR_NUM', 'TOTAL_FLUX', 'LOCAL_FLUX', 'CALL_DAYS', 'CALLING_DAYS', 'CALLED_DAYS', 'CALLED_RING'
高费用	'CERT_AGE', 'INNET_MONTH', 'NO_ROAM_GN_LONG_CALL_DURA_MEAN', 'NO_ROAM_GN_LONG_CDR_NUM', 'GN_ROAM_CDR_NUM', 'LOCAL_FLUX', 'CALL_DAYS', 'CALLING_DAYS', 'CALLED_DAYS', 'CALL_RING'

目录



建模思路及数据准备

➢ 训练数据集

- 分别对分类后的每个数据集使用神经网络模型、决策树模型、朴素贝叶斯模型和支持向量机模型进行建模，使用测试数据集计算出模型的错误率，选择错误率最低的模型作为最优模型，并使用最优的模型对该类别的预测数据集进行预测。
- 分类后的五个数据集作为建模数据集，将各个建模数据集按照8:2的比例随机抽取80%的数据作为待处理训练数据集，20%的建模数据作为测试数据集，待处理训练测试集和测试数据集中均包含流失和非流失用户的信息。

建模思路及数据准备

- 抽取数据集

- 由于原始数据流失与非流失用户的比例有严重差距，导致待处理数据集中流失用户与非流失用户的数量差别很大，使用不放回抽样会导致模型的误判率较高。因此，我们对于流失用户采用放回式简单随机抽样的方法从待处理训练测试集的流失用户中抽取用户数据，直到所抽到的样本与原数据中非流失用户的数量相等。
- 将抽样得到的流失用户数据与不作处理的非流失用户数据合并，并根据筛选出所属类别对应的特征变量作为自变量，将用户是否流失作为模型的因变量，得到训练数据集。

建模思路及数据准备

- 测试数据集

建模数据集中抽样得到的测试数据集直接筛选出所属类别对应的特征变量作为自变量，将用户是否流失作为模型的因变量，作为模型的测试数据集。

建模思路及数据准备

- 预测数据集
- 将预测数据按照前面的数据预处理方法进行处理，并将数据集按照前面聚类的标准分成五类，取所在类别对应的特征变量作为预测模型的输入。

CART决策树模型

- 决策树是一种树状结构，它的每一个叶节点对应着一个分类，非叶节点对应着在某个属性上的划分，根据样本在该属性上的不同取值将其划分成若干个子集。对于非纯的叶节点，多数类的标号给出到达这个节点的样本所属的类。
- 分类与回归树CART模型采用与传统统计学完全不同的方式构建的预测树比常用的统计方法构建的代数学预测准则更加准确，且数据越复杂，变量越多，算法的优越性就越显著。CART算法是一种二分递归分割技术，即把当前样本划分为两个子样本，是生成的每个非叶子结点都有两个分支，它在每一步决策时只能是“是”和“否”。

CART决策树模型

- 采用Python分别对每一类建立CART决策树模型，以低费用群体为例：

（1）使用低费用群体的训练数据集建立决策树模型，构建出决策树，用测试数据集对模型的结果进行测试，分别输出测试和训练结果的混淆矩阵、分类结果，并绘制出ROC曲线，判断模型建立的效果。

（2）将用户基本信息和呼叫信息的指标变量设置为自变量，将3月用户是否流失的指标设置为目标变量，利用训练数据建立C5.0决策树模型。

（3）分别输出训练数据以及测试数据的分类预测结果。

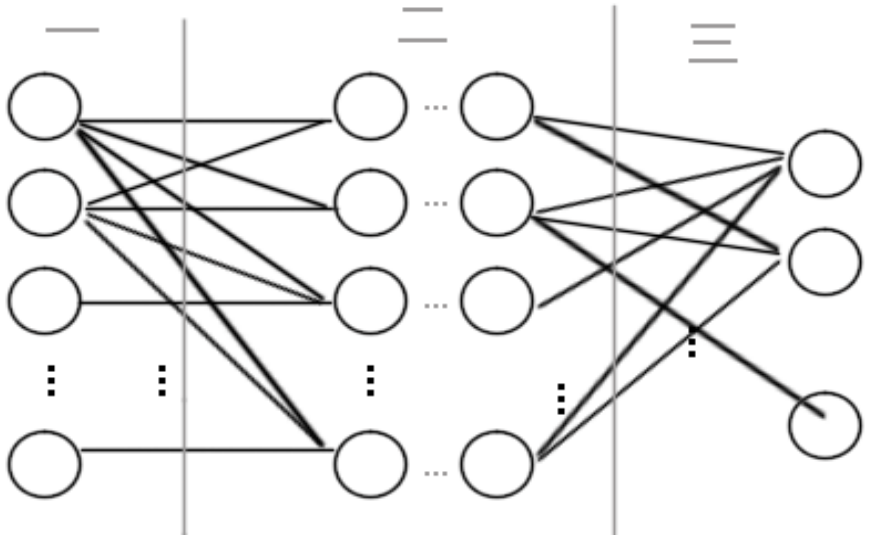
（4）输出测试数据的混淆矩阵，判断模型建立的效果。

神经网络模型

- 人工神经网络是在现代神经科学的基础上提出和发展起来的，这种网络能够依靠系统的复杂程度，通过调整内部大量节点之间相互连接的关系，从而达到处理信息的目的。人工神经网络具有很强的自学习和自适应的能力，可以通过一系列的输入和输出数据，分析两者之间潜在的关系和规律，根据这些关系和规律对新的输入数据进行推测，从而推测出输出结果，这种学习分析的过程被称为“训练”或“学习”。
- BP神经网络的输入层（第一层）将刺激传递给隐藏层（第二层），隐藏层通过神经元之间联系的强度（权重）和传递规则（激活函数）将刺激传到输出层（第三层），输出层整理隐藏层处理的后的刺激产生最终结果。若有正确的结果，那么将正确的结果和产生的结果进行比较，得到误差，再反过来对神经网中的链接权重进行反馈修正，从而来完成学习的过程。

神经网络模型

- 将建模数据集按照7:3的比例随机抽取70%的数据作为训练数据集，30%的建模数据作为测试数据集。输入层为用户基本信息和呼叫信息的指标变量，输出层为3月是否流失的指标，设置隐含节点数以及权值衰减参数建立BP神经网络模型。分别输出训练数据以及测试数据的分类预测结果。输出测试数据的混淆矩阵，判断模型建立的效果。



朴素贝叶斯模型

- 朴素贝叶斯分类是一种十分简单的分类算法，它根据给出的待分类项，求解出在此项出现的条件下各个类别出现的概率，根据概率最大原则判断待分类项的类别。
- 建立朴素贝叶斯的具体步骤是：进行首先确定特征属性，并形成训练样本集合，接着计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计，生成分类器，最后使用分类器对待分类项进行分类，输出是待分类项与类别的映射关系。
- 将建模数据集按照7:3的比例随机抽取70%的数据作为训练数据集，30%的建模数据作为测试数据集。
- 将用户基本信息和呼叫信息的指标变量设置为自变量，将3月用户是否流失的指标设置为目标变量，利用训练数据建立朴素贝叶斯模型。
- 分别输出训练数据以及测试数据的分类预测结果。
- 输出测试数据的混淆矩阵，判断模型建立的效果。

支持向量机模型

- 支持向量机(support vector machine)是一种分类算法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。通俗来讲，它是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，即支持向量机的学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解。原理如下：

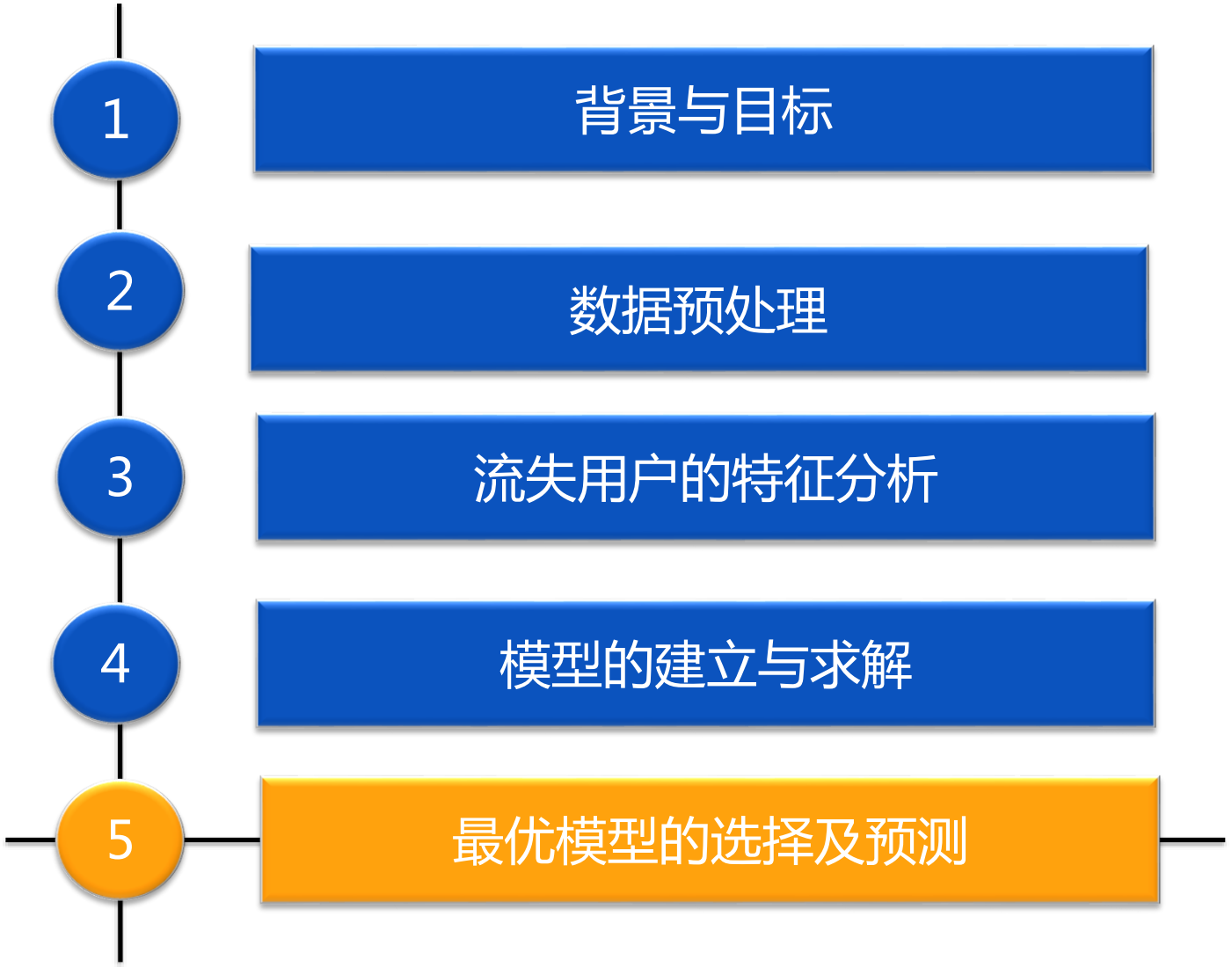
（1）在n维空间中找到一个分类超平面，将空间上的点分类。

（2）一般而言，一个点距离超平面的远近可以表示为分类预测的确信或准确程度。

（3）线性不可分时，把样例特征映射到高维空间中去。

（4）使用松弛变量处理数据噪音。

目录



预测数据集处理

- 将预测数据集USER_INFO_P进行重复值、异常值处理，删除不正常月数的数据，并将通话时长大于本地通话时长、国内漫游通话时长、国内长途通话时长的总和的100的数据删去，得到新的数据后按照建模数据的处理方式进行降维和整合的处理。
- 对整合数据的本月消费进行K-Means聚类，将数据分为5类，分别为高费用、中高费用、一般费用、中低费用、低费用，结合表7选出的关于每类的最优模型，对预测数据的每一类进行流失预测。

各类预测的基本情况

- 通过对各个类别进行预测，得到各个类别的预测基本情况如表所示。

	建模数据集			预测数据集		
	流失数	非流失数	流失率	流失率	流失预测数	非流失预测数
高费用	405	29260	0.013652	0.0352	472	12937
中高费用	107	13188	0.008048	0.013271	373	27733
一般费用	1330	69670	0.018732	0.01932	1372	69641
中低费用	3738	99617	0.036167	0.028216	2917	100464
低费用	4032	65753	0.057777	0.072905	5093	64765

- 从建模数据和预测数据的各类流失率中可以看到，预测数据各类的流失率与建模数据集的流失率的误差在2%左右，因此，认为预测效果良好。

小结

- 通过对用户流失特征的掌握和流失客户的预测，电信公司可以根据不同类型用户的特点来推出不同的优惠套餐，改变收费策略，满足客户的需要，以防止现有客户的流失，以实现公司利润的最大化。

Thank you!