**银行客户行为分析与流失预测**

**赵甫贤**

**摘要：**对于银行来说，根据客户的理财行为预测未来的理财需求并对客户的流失进行预测是十分重要的，银行可以根据数据挖掘可用信息，为客户未来理财需求及客户流失提供参考与建议。在这种背景下，本文从银行的角度出发，基于银行真实客户理财行为与流失数据，建立银行客户客户流失预测模型，旨在有效识别银行潜在的流失客户，为银行防止客户流失提供数据参考与建议。首先对原始数据进行数据清洗，主要为缺失值和重复值的处理，分别从数据意义与数据质量两个方面对缺失值与重复值进行处理。在模型预测方面，首先使用不同的方法对客户进行聚类，然后使用XGBOOST、随机森林等方法进行预测，对不同的模型结果进行评估与可视化展现，并选择预测性能最好的模型。本文对银行客户数据进行深度挖掘，构建合适的模型对银行客户流失情况进行预测，为银行及时识别流失客户并进行挽留提供了合适的工具。

**关键词：银行客户流失分析；XGBOOST；随机森林；预测**

**一．文献综述**

通过对于银行客户行为的量化分析，银行可以制定更具针对性个性化的精准营销方式，挖掘潜力用户的同时规避高风险用户。近年来针对于银行客户行为分析的研究逐渐完善，以下将列举国内外的研究现状。

张明星（2022）认为商业银行利用大数据挖掘技术实施精准营销，金融服务业务模式也有所改变，从一定数量的传统客户，高效地扩展到初具规模的长尾用户。可以说，大数据挖掘技术已逐渐成为商业银行实施营销与规范化管理的着力点和主要生产力。[1]实现精准营销依托于对于用户的准确画像，朱静(2022)认为运用大数据技术，经过模型构建和海量的计算，分析客户信息及购买历史，有助于识别客户特征，了解客户潜在需求，降低了解客户信息的成本，实现客户的精准识别和分组，从而进行差异化营销策略，促进精准营销。[2]麦德奇和保罗·布朗（2014）从市场营销的角度出发，介绍了企业如何用大数据精准定位客户，提高企业在市场营销中的策略。[3]Zhongkai Tian（2018）认为大数据代表了一种获取信息并应用于行业的新方法[[1]](#endnote-1)，是下一代经济发展的动力来源。[4]

目前，主流的数据挖掘算法主要是于2006年在ICDM会议上由145名与会者评选出的十大经典算法，包括C4.5, k-Means, SVM, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes, and CART。[5]将以上算法应用于银行客户行为分析在国内外都有一定的研究基础，侯洪涛，黄有方（2012）利用Apriori算法探究银行客户各属性间的关联度，并对未知属性进行预测。[6]郭松（2018）利用K-均值聚类的方法进一步完善客户在银行的画像系统，便于针对性的进行精准营销和行为预测。[7]汤继杭（2020）在某银行建成的来宾上网系统基础上利用CRISP-DM标准数据挖掘流程和WEB文本挖掘，对客户“线上认证数据”和“线上浏览数据”两类数据进行挖掘，通过统计和分类分析等方法形成客户画像[8] 陈建成，马文扬（2007）利用COX模型对银行客户流失概率进行预测。[9] 王屹聪(2013)提出了在银行理财产品研究中应用的一种负关联规则的GA\_PNAR算法，首先利用Apriori算法生成频繁项集，之后利用基于相关系数的NRGA算法生成含有所有负项的关联规则，在所有规则生成后，利用遗传算法优选生成的规则。[10] Abbasimehr, Hossein和Mostafa Shabani（2019）提出了一种基于时间序列聚类分析银行客户行为的新方法。[11] Ogwueleka和Francisca Nonyelum（2012）提出通过聚类和神经网络识别国际银行用户行为的方法.[12]

近年来，以决策树为基础的随机森林算法在数据挖掘中得到广泛应用。随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定，[13] 它可以用于分类和回归任务，再加上其非线性特性，使其能够高度适应各种数据和情况，这一词最初是由何天琴于 1995 年提出的。何天琴开发了一个使用随机数据创建预测的公式。然后在 2006 年，利奥·布雷曼和阿黛尔·卡特勒扩展了算法，创造了我们今天所知的随机森林。

随后，Biau, Gérard和 Erwan Scornet（2016）提出了随机森林使用的基本流程。[14]应维云（2012）针对实际客户流失数据中正负样本数量不平衡而且数据量大的特点，提出一种改进的平衡随机森林算法，并将其应用于某商业银行的客户流失预测。[15]刘玲和郑建国（2020），提出了一种基于随机森林的组合分类算法用于分析银行用户信用风险，通过K-means算法对标签进行类分解。而对于哪个类分成的簇数(ki)以及随机森林算法自身的参数：树数、特征数使用改进的遗传算法对其进行优化选取。[16]金铮使用商业银行数据分别构建基于逻辑回归算法和随机森林算法的理财产品响应预测模型，并进行结果对比分析，得出经过梯度下降优化后的逻辑回归模型效果更好的结论，并将此模型预测结果应用于实际理财产品营销中[17]。

**二．数据描述**

本文使用的数据取自2021年6月至8月百万余客户的真实理财数据，原始数据的主要属性包括：

（1）客户基础信息：包括客户年龄、性别、客户等级、客户类型、风险等级、开户行是否在省内和收入区间。

（2）理财信息：包括近3个月购买理财次数（手机银行）、近3个月购买理财金额（手机银行）、近3个月赎回理财次数（手机银行）、近3个月赎回理财金额（手机银行）、近3个月理财账户/类资产增加额、最近一次购买理财距离8.31时长、近3月理财/总资产占比、9月购买理财次数、8月月日均理财持有、7月月日均理财持有、6月月日均理财持有、8月末理财持有总额和7月末理财持有总额。

（3）基金信息：近3个月购买基金次数（手机银行）、近3个月购买基金金额（手机银行）、近3个月赎回基金次数（手机银行）、近3个月赎回基金金额（手机银行）、近3个月基金账户/类资产增加额、最近一次购买理财距离8.31时长、近3月基金/总资产占比、9月购买基金次数、8月月日均基金持有、7月月日均基金持有、6月月日均基金持有、8月末基金持有总额和7月月末基金持有总额。

（4）存款信息：近3月存入金额、近3月汇出次数（非本人账户）、近3月汇出金额（非本人账户）、近3个月跨行转账次数、近3个月存款净增加额（和5月31日比、8月月日均存款余额、7月月日均存款余额、6月月月日均存款余额、8月末存款余额和7月月末存款余额。

（5）借记卡/储蓄卡信息：卡片持有数量、近3月转账金额和近3月转账次数。

（6）贷款信息：截止8.31贷款余额、客户是否申请过随e贷（以放款为准）和最近一次办理贷款距离8.31时长。

（7）信用卡信息：名下信用卡持续数量（开卡且有余额）、是否绑定微信/支付宝、近3月累计消费金额和近3月累计消费次数。

（8）资产信息：过去一年平均资产总额（年均值）、本月月末总资产、上月月末总资产、上上月月末总资产、8月末资产总额和7月月末资产总额。

1. **预测模型**

1、数据预处理

将数据输入模型进行训练之前首先要进行数据的预处理，基本的有值类型的转化（例如将性别转变为0-1变量），以及引入被解释属性列——流失情况（0代表未流失，1表示流失）。另外，原始数据存在属性列杂乱、缺失值过多的问题，难以直接用于模型的构建，特别是过多的缺失值传入模型后将会对模型的训练造成极大的影响，因此缺失值的处理是数据预处理工作的重中之重。以下是我们小组缺失值处理的过程：

1. 首先尝试能否减少行数或列数。因为所有列都存在空值，直接去除所有含空值的列不可取；若去除所有含空行，那么百万行数据将只剩下三千行左右，信息损失太多，显然也不可取。另外，发现有大批量多属性为空值的行，因此尝试将空值占比50%以上的行去除，共去除了1041行。但意外发现，这1041条记录中流失客户达到33%左右，而总体流失客户只占比6%，显然有偏差。

表格

描述已自动生成

上述结果暗示了一种信息：因为某种原因无法获得或给出相关信息的客户，可能是办理后未操作或操作过少，这类长期未操作或不活跃用户流失概率显然偏高。

1. 然后从特征的意义上判断，账号与客户流失与否逻辑上无关，因此去除掉账户信息。
2. 其次，去除掉空值大于97%的属性列，例如理财赎回金额。再根据具体列尝试能否去除，我们采取了两种判断方法，一种去除空值方法如下：假设将某属性为空值的客户划分为A类，非空划分为B类，若空值对结果影响大，A类和B类未流失客户占比应该有一定差距，也就是可以通过进行卡方检验，当空值影响大时删去该属性。例如，使用这种方法对贷款余额检验属性如下：

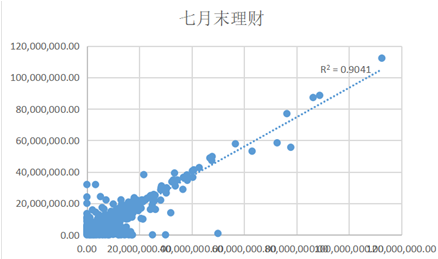
表格

描述已自动生成

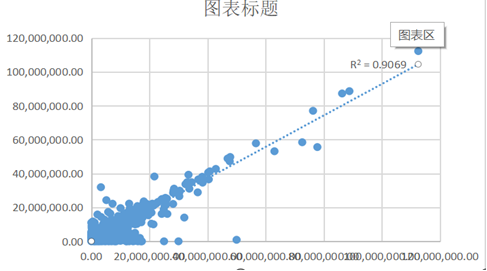
显然卡方值较大，判断空值对结果的影响大，应剔除该属性。其余属性经检测空值影响小，或空值少于设定阈值不需检验。

1. 第二种方法如下：计算两列间相关系数，若相关系数大，说明信息关联度高，存在重复，则去除其中一列。以八月末理财与七月末理财信息为例：

带有缺失数据时（所有数据对）R方值=0.9041。



不带缺失数据时（包括所有两属性其一非空的数据对）R方值=0.9069



显然，两个属性关联度较高，因此选择剔除七月末理财保留八月末理财。

（5）另外，由于原始数据正例数太少、学习偏倚负例太大，因此我们选择复制正例，将流失客户数据复制了16倍。

（6）最终经过上述步骤的数据预处理后，原始客户理财数据保留了17列属性：性别、年龄、客户等级、客户风险等级、理财购买金额、理财占比、月日均理财持有、近三月存入、近三月汇出金额、近三月跨行转账次数、近三月存款净增加额、卡近三月转账金额、八月末理财、八月末存款、八月末资产、七月末存款以及七月末资产

2、预测模型

我们小组的预测基于机器学习，采用XGBOOST和随机森林模型，得到不同属性对客户流失情况的解释程度并进行了可视化展示。

（1）XGBOOST模型

XGBOOST是一种boosting迭代型集成学习算法，适用于大规模数据的预测分析，因此我们小组首选该模型来进行预测，将处理过后的数据集输入模型进行训练，结果如图：

文本

描述已自动生成

可视化结果展示如下：

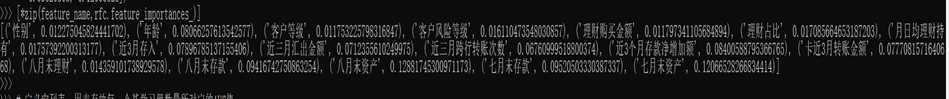
图表, 直方图

描述已自动生成

横轴从左往右与上一张图从上到下一一对应。可以看出，八月末资产的解释程度最高且远高于排名第二的八月末存款，性别的解释程度最低，比之稍微高出一点的是年龄、客户风险等级等客户基本信息。

（2）随机森林模型

另外，我们小组还试验了随机森林（RF）模型，RF是一种高度灵活的bagging型集成学习算法，适用于高纬度数据的预测分析。将数据输入模型进行训练，调参过后得到如下结果：



可视化效果如图：

图表, 条形图

描述已自动生成

可以看到，无论是XGBOOST还是RF，得到的结果大致上都是一样的——即八月末资产是重要性最高的一项特征。

**四．结论**

在银行客户流失预测中，采用XGBOOST和随机森林模型对预处理后的客户理财数据集进行分析，并对特征按重要性排序，发现客户八月末资产（统计截止日期为8.31）的解释程度最强，且近两个月的存款、理财数据的重要性普遍高于性别、年龄、客户等级这类基本信息，由此可得出结论：银行客户流失预测的关键在于近期资产、存款等特征。因此，银行应重点关注客户的近期理财行为，得出客户流失趋向，及时做好风险防范措施和挽留服务，保持与老客户的稳定联系并发掘出有价值的新客户。

**参考文献**

[1]张明星.浅析大数据时代商业银行实施精准营销的策略[J].黑龙江金融,2022(11):19-22.

[2]朱静.基于大数据的银行客户画像与精准营销研究[D].东南大学,2020.DOI:10.27014/d.cnki.gdnau.2020.000826.

[3]麦德奇，保罗·布朗.大数据营销：客户定位[M].机械工业出版社，2014

[4] Yi, Guo, and Tian Zhongkai. "Examining factors influencing customer intention to use Islamic home financing: a study in UTM." The Journal of Social Sciences Research (2018): 648-653.

[5] Data mining algorithm wikipedia

[6]侯洪涛, 黄有方.“基于 Apriori 算法的银行客户行为分析.”硅谷3 (2012).

[7]郭松.商业银行基于客户行为数据的客户画像研究[J].智库时代,2018(47):4-6.

[8]汤继杭.基于客户线上行为分析的Z银行精准营销策略研究[D].宁波大学,2019.DOI:10.27256/d.cnki.gnbou.2019.000813.

[9]陈建成,马文扬.“商业银行零售管理及目标客户流失概率预测研究.”金融论坛 12.12 (2007): 29-36.

[10]王屹聪.数据挖掘算法在银行理财产品营销中的应用研究.MS thesis.郑州大学,2013.

[11] Abbasimehr, Hossein, and Mostafa Shabani. "A new methodology for customer behavior analysis using time series clustering: A case study on a bank’s customers." Kybernetes (2019).

[12] Ogwueleka, Francisca Nonyelum, et al. "Neural network and classification approach in identifying customer behavior in the banking sector: A case study of an international bank." Human factors and ergonomics in manufacturing & service industries 25.1 (2015): 28-42.

[13] random forest wikipedia

[14] Biau, Gérard, and Erwan Scornet. "A random forest guided tour." Test 25.2 (2016): 197-227.

[15]应维云.随机森林方法及其在客户流失预测中的应用研究[J].管理评论,2012,24(2): 140-145.

[16]刘玲,郑建国.一种基于随机森林的组合分类算法设计与应用[J].电子设计工程,2020, 28(16): 54-57.

[17]金铮.基于机器学习算法的商业银行精准营销模型研究[J].

1. [↑](#endnote-ref-1)