银行信用卡客户流失情况预测

大数据引论课程项目汇报

徐朱玮　　刘琮璟

研究背景及研究目标

通过分析客户交易的数据来全面了解客户的价值、需求、期望和行为，以期改善与客户的关系的行为，被称为CRM（客户关系管理）。它是一种商业理念，旨在获取和留住客户，提高客户价值和忠诚度，并且实施以客户为中心的战略。

客户流失预测是CRM的一种，它通常被定义为客户在给定的时间段内停止与公司开展业务。对于商业银行而言，判断哪些客户可能流失显得尤为重要。研究表明，留住客户可以带来很大的经济效益，如果将客户流失率降低5%，可以给银行带来25%至85%的业绩提升。而开发新客户的成本是留住现有客户成本的5到6倍。负债业务是银行的重要业务之一，个体零售客户也是银行重要的客户组成部分。因此，银行信用卡客户的不定期流失往往是困扰银行经理的一大问题。

近年来，许多科学家提出了各种机器学习方法，其中很多方法能够用于分类，以预测客户流失的行为。这其中包括逻辑回归、决策树、随机森林、SVM、KNN、神经网络等。

这些模型在应用于不同的分类任务时表现各有千秋，并没有一个标准能够判断这些模型的优劣。但目前尚未有人将这些模型全部应用于银行用户流失的数据集，并且分析它们在这一数据集上的表现。因此，本研究将把上述六种模型应用于银行信用卡用户流失的数据集，并且找出最适用于这一特定数据集的模型，并为银行提供较精准的流失客户预测服务。

数据集介绍及数据预处理

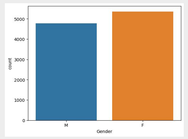
数据集介绍

该数据集源自Kaggle网站，包含了从10127名消费者信用卡的23个客户特征信息，包括年龄、性别、附属卡数量、受教育程度、婚姻状况和收入类别；以及每位客户与信用卡提供商关系的信息，如卡片类型、与银行交互的频率、信用额度、总循环余额、过去12个月开放购买的信用额度等。数据集中的10127名消费者有两类标签：一类为已经流失的客户（1627个样本），另一类未流失客户（8500个 样本）。

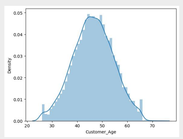
缺失值处理

去除无效特征

数据集分布分析



左图展示了该数据集的男女分布，接近1：1；右图展示出数据集中客户的年龄分布接近于正态分布。



因此，该数据集基本接近真实世界中银行信用卡用户的注册情况，数据集的分布较为合理。使用这一数据集训练的模型在面对新的用户时，鲁棒性预计较好。

数据集格式转换

将数据集中6个非数值类型转化为数值类型，具体转化方式如图所示。

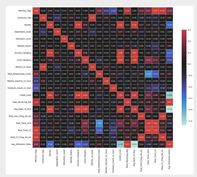
Attrition_Flag	Attrited Customer（流失客户）：0	Existing Customer（未流失客户）：1
Gender	F（女性）：0	M（男性）：1
Education_Level	Unknown（未知）：1	Uneducated（未受教育）：2
	High School（高中）：3	College（本科在读）：4
	Graduate（本科毕业）：5	Post-Graduate（硕士）：6
	Doctorate（博士）：7	
Marital_Status	Unknown（未知）：0	Single（单身）：1
	Divorced（离异）：2	Married（已婚）：3
Income_Category	Unknown（未知）：1	Less than $40K（少于4万美金）：2
	$40K - $60K（四万-六万美金）：3	$60K - $80K（六万-八万美金）：4
	$80K - $120K（8万-12万美金）：5	$120K + （大于12万美金）：6
Card_Category	Blue（蓝卡）：1	Silver（银卡）：2
	Gold（金卡）：3	Platinum（铂金卡）：4

特征归一化

训练集和测试集的划分

为了增加模型的鲁棒性，在划分训练集和测试集时，我们采用分层抽样的方式（也即设置stratify参数为y）。这样一来，在训练集和测试集中，流失客户和未流失客户的比例都和原数据集中这两类的比例相同。划分80%为训练集，20%为测试集，最终得到8101组训练集数据和2026组测试集数据。

探索性分析



利用python的seaborn库中的heatmap函数绘制特征间的相关性系数热力图。图中，方格的颜色越接近于红色，就表明特征的正相关程度越高；方格的颜色越接近于浅蓝色，就表明特征的负相关程度越高。

通过观察第一行或者第一列，可以看到不同特征与Attrition_Flag（客户流失与否标签）之间的关系，与其正相关度最高的特征是过去12个月的交易总数，相关性系数未0.37。其次为相关性系数为0.29和10.26的第4季度相比于第1季度的交易数量变化和信用卡上的循环余额总额。负相关程度最高的特征是相关性系数为-0.2的过去12个月与银行进行交互的数量。

实验模型搭建

逻辑回归

假设：对于给v定的输入特征 X = （X1,X2,...Xn),输出 Y为1（未流失客户）的概率可以用一个线性组合来表示，并通过逻辑函数（sigmoid函数）将结果映射到[0, 1]的范围。表达式如下：

$$P(Y=1|X)=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_nX_n)}}$$

由于数据稀疏性明显，本模型选择L1正则化，也称为Lasso回归。

决策树

决策树通过对数据集进行递归地划分，构建一个树形结构来进行决策。先通过计算不同特征的指标，选择最佳的划分特征。再根据选择的划分特征，将数据集划分成多个子集。后续对每个子集递归地重复前两个步骤，直到满足停止条件。

在该模型的设置中，选择熵（Entropy）作为不纯度度量。在每个节点上选择能够最大程度降低熵的特征进行分割。熵函数公式如下：

$$H(S)=-\sum_{i=1}^{C}p_i\log_2 p_i$$

随机森林

随机森林是一种集成学习方法，基本思想是通过构建多个决策树来进行预测，然后将这些决策树的结果进行综合，以获得更准确的预测结果。随机森林引入了两种随机性，即随机抽样和随机特征选择。对于每个决策树的训练数据，随机森林从训练集中随机抽样。此外，对于每个节点的特征选择，随机森林在节点分裂时从所有特征中选择一个随机子集，而不是使用全部特征。这样可以增加每个决策树的独特性，提高整体模型的多样性。

支持向量机

支持向量机分类器是一种用于分类问题的监督学习模型，主要目标是找到一个最优的超平面，尝试找到一个能够最大化类别间间隔的超平面，将不同类别的数据点分开。

由于该数据集的分类问题属于线性不可分问题，我们引入了径向基函数（Radial Basis Function，RBF）并设置grammar=0.2，从而将数据从原始特征空间映射到高维特征空间，以便在新空间中找到线性可分的超平面。

此外，该模型还设置正则化参数C=1.0，具体的损失函数如下：

$$f(w,b)=C\cdot L(w)+\frac{1}{2}\cdot\|w\|^2$$

K最近邻

在KNN中，给定一个新的数据点，算法会找到特征空间中最近的K个训练数据点，然后通过这些邻居的多数投票来确定新数据点的类别。

在该模型的搭建中，我们选择了欧式距离，计算公式为：$d(x,x')=\sqrt{\sum_{i=1}^{n}(x_i-x'_i)^2}$

接着我们将邻居数设定为5，记录与该数据点距离最近的5个距离最近的点的标签（流失客户或者未流失客户）。最后通过多数投票决定该数据的类别。

神经网络（多层感知机）

神经网络包含输入层、隐藏层和输出层，每一层由多个神经元组成。神经网络的输入通过每个连接进行传递，并在隐藏层中进行加权求和，最终通过激活函数进行转换，得到输出。

在该模型中共设置了三个隐层，分别含有256、256和128个神经元，选择ReLU函数作为激活函数，数学表达式为：f(x) = max(0, x)

选择随机梯度下降算法作为优化器，用于最小化损失函数。设置alpha=0.0001，对权重进行轻度的L2正则化。设置max_iter=100，最大迭代次数为100.

评价指标

混淆矩阵

混淆矩阵分为TP（真阳性）、FP（假阳性）、FN（假阴性）、TN（真阴性）。应用到这一数据集上时，TP指预测是流失的客户，结果也是流失的客户；FN指预测是未流失的客户，结果是流失的客户；FP指预测是流失的客户，结果是未流失的客户；TN指预测是未流失的客户，结果也是未流失的客户。这一矩阵可以帮助我们看到模型的预测结果中有多少预测正确、多少预测错误且是怎么样的错误形式，方便我们对模型性能进行判断。

ROC曲线

横坐标为假阳性率，纵坐标为真阳性率。通常，我们认为曲线的凸起程度越高，模型准确率越好。图中的虚线是对角线，表示随即猜测，因此ROC曲线越接近对角线，则模型的预测率越低。

ROC曲线下方的面积称为AUC，一般来说，AUC越大、分类器越好。AUC为0.5表示随机猜测。

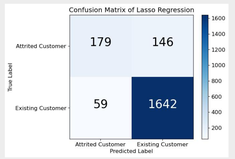
精确率、召回率、正确率、F1 Score

精确率(presicion) = TP/(TP+FP)
召回率(recall) = TP/(TP+FN)
正确率(accuracy) = (TP+TN)/ALL
F1 Score（调和平均数） = 2 (precision * recall) / (precision + recall)

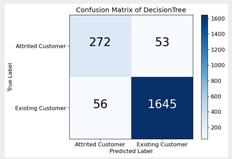
实验结果

混淆矩阵

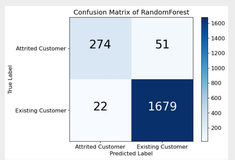
- **逻辑回归**



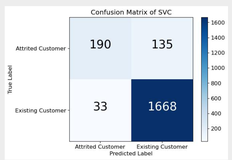
- **决策树**



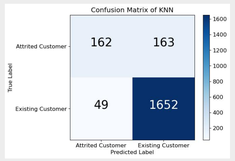
- **随机森林**



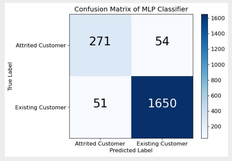
- **支持向量机**



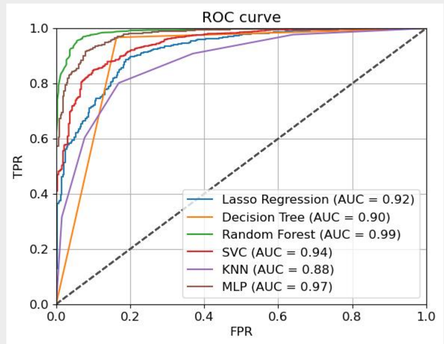
- **K最近邻**



- **神经网络**



ROC曲线



精确率、召回率、正确率、F1-score

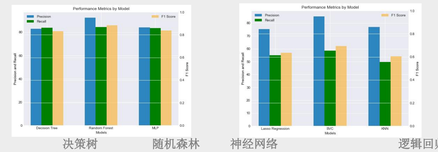
models	Attrition_precision	Existing_precision	Attrition_recall	Existing_recall	Attrition_F1-score	Existing_F1-score	accuracy
Lasso Regression	75.2101	91.8345	55.0769	96.5315	0.635879	0.941244	89.8815
Decision Tree	82.9268	96.8787	83.6923	96.7078	0.833078	0.967932	94.6199
Random Forest	92.5676	97.0520	84.3077	98.7066	0.882448	0.978723	96.3968
SVC	85.2018	92.5125	58.4615	98.0600	0.693430	0.952055	91.7078
KNN	76.7773	91.0193	49.8462	97.1193	0.604478	0.939704	89.5360
MLP	84.1615	96.8310	83.3846	97.0018	0.837712	0.969163	94.8174

结论

流失客户预测

在银行用户流失的预测中，相比未流失客户的预测，对于流失客户的预测更为重要，这主要是因为在实际情况中，流失客户对银行的业务影响通常更为显著。

对上述实验数据进行分析，我们可以发现在对于流失客户的预测中，六个模型大致可以分为两类，第一类包括决策树、随机森林和多层感知机，它们展现出较为良好的性能；第二类包括逻辑回归、支持向量机和K最邻近算法，它们的性能较差。



决策树　随机森林　神经网络　逻辑回归　SVC　KNN

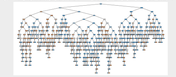
模型的综合选择

根据上述结论，决策树、随机森林以及多层感知机，这三种模型在流失客户的预测中表现出色。观察所有指标可以得出，它们在未流失用户的预测方面也取得了显著的成果，可以作为成靠的预测工具。

然而，在实验过程中发现多层感知机的计算复杂度较高，导致训练时间较长，不太适用于实时预测应用。

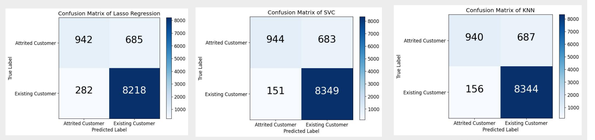
随机森林在流失客户预测中表现尤为出色，所有指标都位于六个模型之首，显示出其优越的性能。与此同时，通过利用多个决策树的集成，随机森林克服了决策树容易过拟合的问题，提高了模型的泛化能力。

综上所述，基于实验结果，推荐选择随机森林作为流失客户预测的最优模型。



该研究依然存在的不足

通过观察逻辑回归、支持向量机和K最邻近这三个模型的预测结果，可以发现在它们的混淆矩阵中出现了极高的相似性，但本研究并未探寻真正的原因。



逻辑回归　支持向量机　KNN

可以看到，当把数据集中的所有数据都进行测试的时候，这三个模型都把680余个流失客户预测为未流失客户，我们猜想数据集中可能存在一些具有迷惑性的特殊情况或异常值，使得模型受到干扰，共同遗漏或者误判了某些特定特征，导致难以正确地识别流失客户。

对此，我们提出设想的探究方案，可以找到三个模型错误判断的具体用户并取交集，如若交集覆盖面广泛，则可初步验证该猜想。随后可以引入特征工程，对这些数据进行更深入的分析，探究降低模型准确率的真实原因并寻找解决方案。

参考文献

[1] Peppers, D., & Rogers, M. (1996). The one to one future: Building relationships one customer at a time. NY: Doubleday.
[2] Kahreh, M. S., Tive, M., Babania, A., and Hesan, M. (2014). Analyzing the applications of customer lifetime value (CLV) based on benefit segmentation for the banking sector. Procedia: Social and Behavioral Sciences 109(8), 590 - 594 (https://doi.org/10.1016/j .sbspro.2013.12.511).
[3] Lopez, J., and Maldonado, S. (2019). Profit-based credit scoring based on robust optimization and feature selection. Information Sciences 500, 190 - 202 (https://doi.org/ 10.1016/j.ins.2019.05.093).
.....