对银行信用卡客户流失情况预测的最佳模型探究

【摘要】本研究旨在通过使用 Kaggle 网站上公开的信用卡客户数据集，选取最佳机器学习模型预测潜在流失客户，帮助银行减少损失。该数据集包括 10127 名消费者的 23 个特征信息，实验采用逻辑回归、决策树、随机森林、支持向量机分类器、K-最近邻和神经网络六种模型进行二分类预测。通过混淆矩阵和 ROC 曲线等指标进行了评估。结果显示，决策树、随机森林和神经网络的预测性能较好，其中随机森林在综合性能上表现最佳，通过其输出的特征权重排行为银行提供进一步参考。然而，实验仍存在不足，数据集中可能存在异常值导致混淆矩阵高度相似，可进一步探究。此外，由于流失客户与未流失客户数量不平衡，本研究提供下采样方法提高流失客户预测召回率，以帮助银行识别潜在流失客户。

【关键词】二分类；机器学习；模型对比；银行客户流失

1 引言

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。一方面，银行客户的流失会导致银行业务量减少、带来资金损失，且可能致使银行市场份额下降、无法享受规模效应带来的盈利机会；另一方面，银行的资金常常被用来做再投资，因此未被预料到的客户流失可能会导致银行既有的投资组合产生风险和损失。因此，对于银行而言，提前预知可能流失的用户并采取必要的补救措施就显得尤为重要。

要想提前预知潜在的流失客户，一般采用机器学习或深度学习的模型来做预测。这些模型可以通过分析客户的个人特征数据、历史交易数据和其它行为数据，来判断该客户流失的概率。通过模型的预测结果，银行可以很快识别出可能流失的客户群体，并且采取有针对性的补救措施。面对可能流失的客户，银行能够提前与客户进行交流，给客户提供更好的服务；面对可能失去的资金来源，银行也可以及时调整自己的投资组合，从而减少因为突然的资金流失带来的财务损失。

本研究将采用包括 Lasso 回归、决策树、随机森林、支持向量机、K 最近邻和全连接神经网络等不同的机器学习和深度学习模型，对某一关于信用卡用户流

失情况的数据集进行分析。并且以混淆矩阵、ROC 曲线，以及精确率、召回率、准确率、F1-score 等指标作为性能评价指标，希望能够找到预测这一数据集的最佳模型。并以此为银行提供较精准的流失客户预测服务。

2 研究背景和意义

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（Peppard，2000）。

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（Neslin 等，2006）。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（Reichhold 和 Sasser，1990）。而开发新客户的成本是留住现有客户成本的 5 到 6 倍。

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（Bandam et al.，2022）以预测客户流失的行为（Günesen et al.，2021）。这其中包括逻辑回归（Kiguchi 等，2022）、决策树（Vezzoli 等，2020）、随机森林模型（Kuznietsova 等，2022）、SVM（Sánchez et al.，2022）、朴素贝叶斯（Jayadi et al.，2020；Rabiul-Alam et al.，2021）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，本研究将把这些模型全部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。

3 数据集介绍和预处理

3.1 数据集简介

本研究采用数据集来自于全球最大的数据集平台之一 Kaggle。这一数据集包含了从 10127 名消费者信用卡投资组合中收集的 23 个客户特征信息。这些特

征包括全面的人口统计信息，如年龄、性别、附属卡数量（可近似家属人数）、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、持有银行产品的总数、过去一年中的不活跃月份数和活跃月份数。此外，它还包含了关于客户流失前消费行为的关键数据，如信用额度、总循环余额、过去 12 个月开放购买的信用额度；以及其它一些可分析指标如第 4 到第 1 季度的总变化金额、过去 12 个月的总交易额、过去 12 个月的总交易数、平均利用率和朴素贝叶斯分类器的流失标志(信用卡类别与 12 个月期间的联系人数量、依赖数量、教育水平和不活跃月份相结合)。

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，另一类则是仍在稳定使用信用卡服务的客户。其中第一类有 1627 个样本，第二类有 8500 个样本。这一数据集可以使用监督学习和神经网络对其进行训练，来预测一个新的拥有全部特征的用户属于流失客户还是未流失客户。

3.2 数据预处理

3.2.1 缺失值查询

在导入数据集后，我们首先使用 python 中的 info()函数对数据的缺失值和每一个特征的数据类型做查询。

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10127 entries, 0 to 10126
Data columns (total 21 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   CLIENTNUM                 10127 non-null  int64
 1   Attrition_Flag            10127 non-null  object
 2   Customer_Age              10127 non-null  int64
 3   Gender                    10127 non-null  object
 4   Dependent_count           10127 non-null  int64
 5   Education_Level           10127 non-null  object
 6   Marital_Status            10127 non-null  object
 7   Income_Category           10127 non-null  object
 8   Card_Category             10127 non-null  object
 9   Months_on_book            10127 non-null  int64
 10  Total_Relationship_Count  10127 non-null  int64
 11  Months_Inactive_12_mon    10127 non-null  int64
 12  Contacts_Count_12_mon     10127 non-null  int64
 13  Credit_Limit              10127 non-null  float64
 14  Total_Revolving_Bal       10127 non-null  int64
 15  Avg_Open_To_Buy           10127 non-null  float64
 16  Total_Amt_Chng_Q4_Q1      10127 non-null  float64
 17  Total_Trans_Amt           10127 non-null  int64
 18  Total_Trans_Ct            10127 non-null  int64
 19  Total_Ct_Chng_Q4_Q1       10127 non-null  float64
 20  Avg_Utilization_Ratio     10127 non-null  float64
dtypes: float64(5), int64(10), object(6)
memory usage: 1.6+ MB
```

图 1-1 缺失值查询

可以在上图中看到，这一数据集的 23 个特征均无缺失数据。除了 Attrition_Flag（客户流失与否标签）、Gender（性别）、Education_Level（受教育程度）、Marital_Status（婚姻状况）、Income_Category（收入水平）和 Card_Category（信用卡类型）这几个特征的数据为 object 类型外，其它特征的数据均为整数或浮点数。

3.2.2 去除无效特征

经过分析，显然在所有特征中“CLIENTNUM”（客户编号）、以及最后两列朴素贝叶斯（即通过朴素贝叶斯分析客户是否会按照某些特定特征流失）是无效特征。这三个特征与客户是否会流失没有直接的关系，因此可以直接删除。

删除后，这一数据集变为 10127 行、20 列。每一行包含了一个银行信用卡用户的信息，每一个用户有 20 个相关特征，且所有数据都没有缺失值。

3.2.3 数据集分布分析

在开始研究前，我们首先对数据集的一些重要特征做简单的分析，来观察这一数据集是否分布合理并适合模型训练。

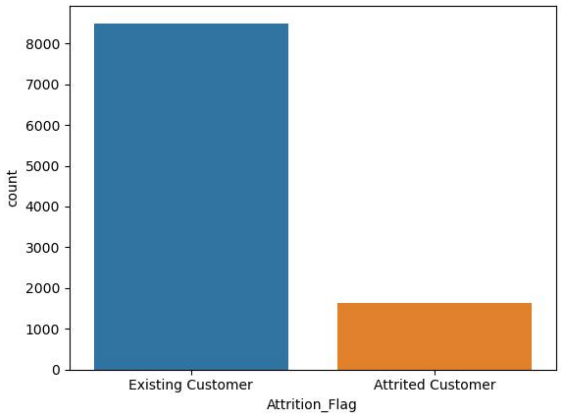

图 1-2-1 未流失客户与流失客户

由于该数据集将用来预测客户是否会流失，因此我们首先查看流失客户和未流失客户的数量是否平衡。可以看到，该数据集中未流失的客户数量远远大于已经流失的客户数量，这可能会在后续模型训练的时候导致模型对流失客户的分类效果不佳。在后期训练中应该特别注意这一点。

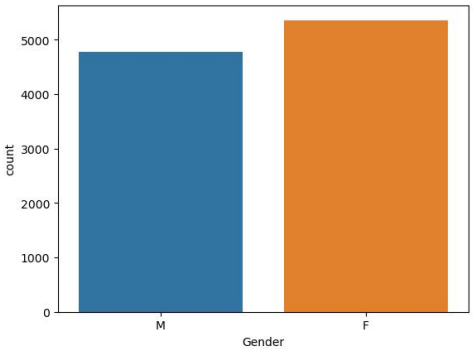

图 1-2-1 男女比例

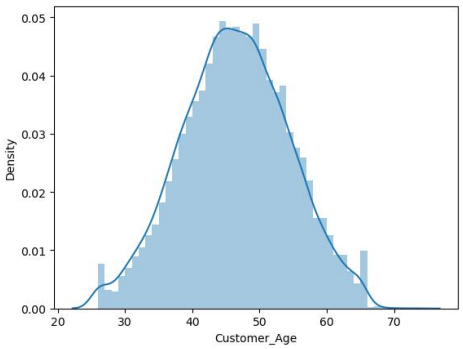

图 1-2-3 年龄分布

查看数据集中客户的性别和年龄分布。可以看到，虽然男性客户的个数要略少于女性客户的个数，但是总体而言分布比较均匀，接近 1：1；而数据集中客户的年龄分布接近于正态分布。

因此可以认为，该数据集所包含的用户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。因而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。

3.2.4 数据集格式转换

在使用 info()函数查看数据集中特征情况时，发现有 6 个特征的数据是非数值类型的。在去除无效特征时，这 6 个特征均未被去除。而在进行模型训练时，我们需要所有数据都是数值数据才能够进行训练。因此，需要将非数值类型的数据转换为数值类型。

由于这 6 个非数值类型的特征均对客户进行了分类，因此我们采用一个数字表示一个类别，来对这些特征的数据进行转换。

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

表 1 非数值类型特征转换

转换中采用的数值和类别对应如上图所示。由于 Attrition_Flag（客户流失与否标签）和 Gender（性别）均为二分类标签，两个标签直接没有直接联系，因此直接将其转换为数字 0 或数字 1。Education_Level（受教育程度）、Income_Category（收入水平）和 Card_Category（信用卡类型）的类别之间有逐级上升的关系，因此分别随着学历、收入水平和信用卡类型等级的上升，所分类的类别对应的数字也相应变大。Marital_Status（婚姻状况）的数据有部分的联系，离异和单身与已婚都有一定的联系，因此将单身设置为 1、离异设置为 2、已婚设置为 3，未知则被设置为 0。

3.2.5 特征归一化

在这一数据集的 20 个特征中，每个特征的数据的值跨越的范围很大，这会在模型训练中导致权重不一致的情况。因此，我们对数据做特征归一化，来让不同特征的数值取值范围统一到相同的尺度上。在本研究中，我们使用 python 中的 MinMaxScaler 将所有特征的取值范围缩放到[0,1]之间，以此避免某些特征对模型有过大的影响。

完成归一化后，所有特征的值的范围都将为[0,1]，但是每一个特征本身的分布情况是不会发生改变的。因此，特征归一化不会因为改变一些特征的数值而对数据集本身做出改变，它能够在保持数据集原有分布的情况下更加有利于模型的训练。

3.2.6 训练集和测试集的划分

我们将采用监督学习的方式完成这一二分类问题的预测，因此首先需要划分标签和数据。标签就是 Attrition_Flag（客户流失与否标签）这一列，数据则是其它的 19 个特征。如前文所述，在数据集中，未流失的用户数量有 8500，已经流失的用户数量只有 1627，两者的比例严重不均衡。而研究的目标是完成客户“流失”还是“未流失”的二分类预测。

因此，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。我们希望可以通过这一方式减少两类数据集巨大差值对模型性能的影响。

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。

3.2.7 探索性分析

在开始模型训练之前，我们希望对数据集各特征的重要程度有一个初步的认识。我们采用 python 的 seaborn 库中的 heatmap 函数绘制特征间的相关性系数热力图。这里采用的是 Person 相关性系数，是因为经考察发现，表格中的各项数据基本符合正太分布。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。（图见下一页）

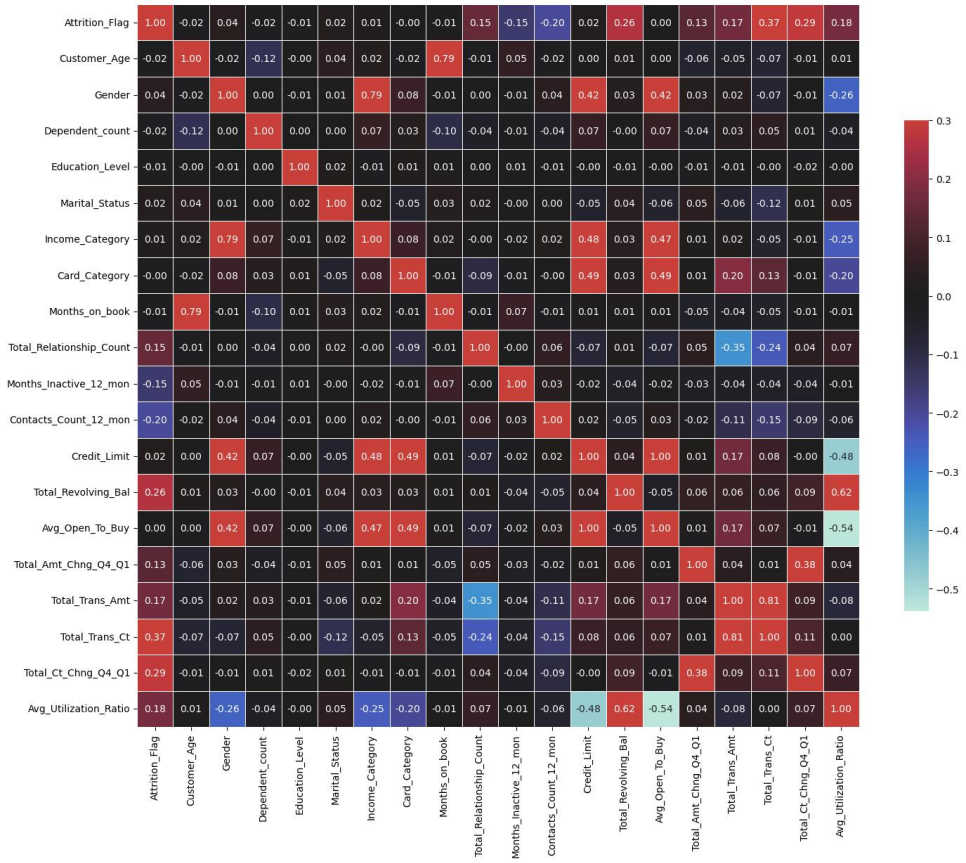

图 1-3 相关性系数热力图

通过观察第一行或者第一列，就可以看到不同特征与 Attrition_Flag（客户流失与否标签）之间的关系。可以看到，与其正相关度最高的特征是过去 12 个月的交易总数，相关性系数为 0.37。其次为相关性系数为 0.29 和 0.26 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。

通过对特征重要程度的初步认识，可以帮助我们在后续模型搭建的过程中，在合适的模型上适当增加相关性系数更高的特征的权重。

4 研究方法和性能评价指标

4.1 研究方法

在面对信用卡用户流失预测这一复杂的二分类问题时，我们选择了逻辑回归、决策树、随机森林、K-最近邻（KNN）、支持向量分类器（SVC）和多层感知机（MLP）这六种模型。这样的选择基于对这些模型特性的综合考虑，以期望覆盖各种数据分布和模型复杂性的情况。

首先，逻辑回归是一个简单而经典的模型，适用于线性可分的情况，其参数直观解释且计算效率高，为建模的初始选择提供了基准。决策树和随机森林对于处理非线性关系有较好的拟合能力。决策树通过学习规则和决策路径，而随机森林通过集成多个决策树来提高模型的鲁棒性和准确性。这两者的可解释性使其成为理解客户流失因素的强有力工具。KNN 是一种基于实例的学习方法，具有对局部模式的敏感性。在存在明显的局部特征时，KNN 能够捕捉这些模式，为模型提供更细致的决策依据。支持向量分类器（SVC）被引入，是因为它在可以高维空间中构建决策边界，适用于处理潜在的非线性关系，其核技巧能够有效地处理复杂的数据分布。最后，为了更好地适应复杂的非线性关系，引入了多层感知机（MLP）。MLP 是一种深度学习模型，通过多层结构自适应地学习数据的多层次特征表示，对于高阶关系的建模能力更强。

这六个模型之间存在一些相似性，例如决策树和随机森林都是基于树结构的模型，而逻辑回归和 SVC 都是线性分类器。这样的选择是为了确保我们在建模过程中能够考虑到数据的多样性，并在模型的选择上有更全面的侧重。

4.1.1 逻辑回归

逻辑回归是一种可以用于解决二分类问题的线性模型，基于以下假设：对于给定的输入特征 $X = (X_1, X_2, \ldots X_n)$，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（通常为 sigmoid 函数）将结果映射到[0, 1]的范围。逻辑回归模型的表达式如下：

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

sigmoid 函数的表达式为：

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

其中 $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$

模型的训练过程中，通过最大似然估计找到一组参数，找到一组最佳参数 $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$，使训练集的标签数据在给定模型下的概率最大。

结合预先进行的探索性分析，本数据集特征存在稀疏性，即数据集中存在一些关键的特征，对客户流失的预测起到关键作用（如过去 12 个月的交易总数），同时也存在一些特征对结果的影响相对较小（如受教育程度），故而选择使用

L1 正则化，也称为 Lasso 回归。 L1 正则化可以促使模型选择稀疏权重，将一些特征的权重推向零，使得模型更易于解释。

损失函数表达式：

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} log\,(\sigma() + (1 - y^{(i)})\,log\,(1 - \sigma()] + \lambda\sum_{j=1}^{n}|\beta_j|$$

4.1.2 决策树

决策树是一种常用的机器学习算法，用于解决分类和回归问题。它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信息增益或基尼指数等指标，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。此外，为了避免过拟合，可以对构建好的决策树进行剪枝操作，去除一些不必要的节点。

在该模型的设置中，我们选择熵（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。在选择分割特征时，决策树会尝试不同的分割方式，并计算每种分割方式下的熵。选择能够使得熵减小最多的分割方式，即选择能够最大程度降低混乱程度的特征进行分割。熵函数公式如下：

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

4.1.3 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，以确保每个树的训练数据略有不同，这样可以减少过拟合的风险，并提高模型的泛化能力。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。

4.1.4 SVM

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（margin）的超平面，将

不同类别的数据点分开。在 SVC 模型中，数据点被视为在 n 维空间中的向量，超平面则是一个 n-1 维的线性子空间。

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：

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

4.1.5 K-最近邻

K-最近邻（K-Nearest Neighbors，简称 KNN）是一种基于实例的监督学习算法，常用于多分类问题，也可用于二分类问题。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 KNN 的搭建中，首先我们要进行距离度量，在本模型中选取了欧式距离。

距离的计算公式为：

$$d(x,x^{'}) = \sqrt{\sum_{i=1}^{n} (x_i - x_i^{'})^2}$$

接着我们将邻居数设定为 5，也就是说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，也就是 5 个距离最近的点中出现类别次数最多的标签即为该数据的标签。

4.1.6 全连接神经网络

神经网络是一种模拟人脑神经网络结构的机器学习模型，本实验中使用的是‘MLPClassifier’，即多层感知机分类器，一种基于神经网络的分类模型。神经网络包含输入层、隐藏层和输出层，其中每一层由多个神经元组成。隐藏层中的神经元数量的选择是根据具体问题和经验进行调整的重要参数，这些神经元通过权重连接到下一层，形成网络结构。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。

对于神经网络中的每个神经元 j，其输入 a_j 和输出 z_j 之间的关系可以表示为：

$$z_j = f(\sum_{i=1}^{N} w_{ij} \cdot a_i + b_j)$$

其中，N 是上一层神经元的数量，w_{ij} 是连接神经元 i 和 j 的权重，b_j 是神经元 j 的偏差（bias），f 是激活函数。这个过程在神经网络中一层一层地重复，直到输出层。

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，形成了一个深层的神经网络结构。这样的结构有助于模型学习更复杂的特征和关系，但也需要更多的数据和调整。

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（Stochastic Gradient Descen，SGD）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。

4.2 性能评价指标

4.2.1 混淆矩阵

混淆矩阵是用来评价模型的一个重要指标，它分为 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。

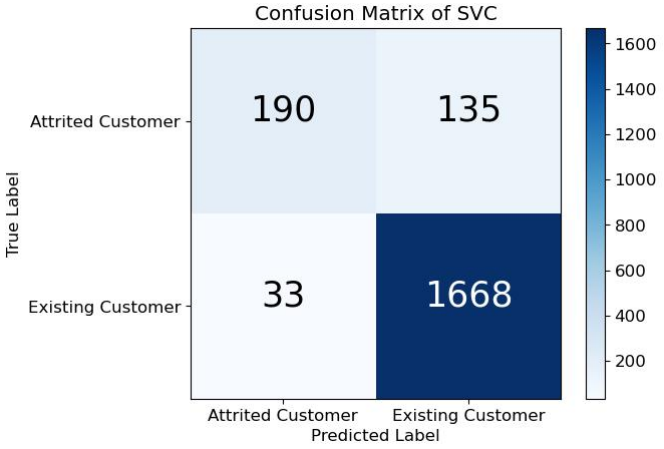

图 2 混淆矩阵示例（SVC）

这一矩阵可以帮助我们清晰地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。

4.2.2 精确率、召回率、正确率和 F1 Score

精确率、召回率、正确率和 F1 Score 都是判断模型性能的重要指标。需要说明的是，在这些指标的判断中，我们会设置流失的客户为正例、未流失的客户为反例，以及流失的客户为反例、未流失的客户为正例这两种情况。以便能够更为全面地分析模型的性能。

精确率(presicion)的计算公式为 TP/(TP+FP)。如：流失用户的精准率=（预测流失真实也流失 190）/（预测流失真实也流失 190+预测流失真实未流失 33）=85.20%。这一数值就代表了在模型预测为流失的所有用户中，真正流失的用户所占的比例。可以帮助我们判断模型面对某一类数据，判断正确的概率。

召回率(recall)的计算公式为 TP/(TP+FN)。如：未流失用户的召回率=（预测未流失真实也未流失 1668）/（预测未流失真实也未流失 1668+预测流失真实未流失 33）=98.06%。这一数值代表了在所有真正未流失的用户中，被预测为未流失的用户所占的比例。这一概率越高，代表错失的个数越少。

正确率(accuracy)的计算公式为(TP+TN)/ALL。如：正确率=（预测流失真实也流失 190+预测未流失真实也未流失 1668）/（总样本 2026）=91.7078%。它代表了在所有样本中，预测值和真实值一致的样本所占的概率。在这里可以看到，被准确预测的用户占总数的 91.7078%。

F1 Score 的计算公式为 2 (precision * recall) / (precision + recall)，它也就是调和平均数（即 P 和 R 的倒数之和的 1/2 的倒数）。只有在 P 和 R 都比较好的时候，F1 Score 才有可能比较高；如果 P 和 R 中仅有一者较好，而另一者与其相差较大，那么 F1 Score 也不会很高。它意味着，P 和 R 都较好的模型性能比 P 和 R 值相差过大的模型具有更好的性能。

4.2.3 ROC 曲线

ROC 曲线（Receiver Operating Characteristic curve）是一种用于评估分类模型性能的图形工具。它的横坐标为假阳性率（False Positive Rate），表示错误地判断为正例的概率（错误地预测为正的数量/原本为负的数量）；纵坐标为真阳性率（True Positive Rate，即 precision）（正确地预测为正的数量/原本为正的数量），表示正确地判断为正例的概率。

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。

5 实验过程

5.1 逻辑回归（Lasso 回归）实验结果

5.1.1 混淆矩阵和 ROC 曲线

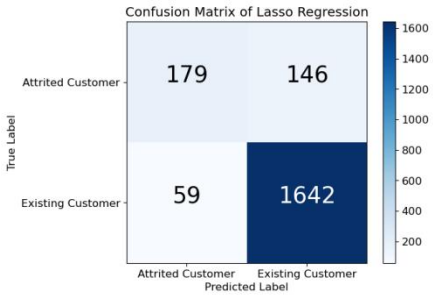

图 3-1-1 Lasso 回归混淆矩阵

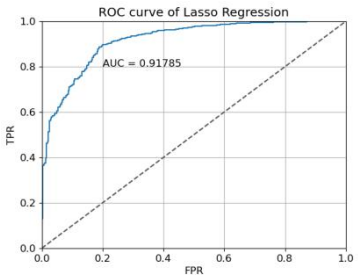

图 3-1-2 Lasso 回归 ROC 曲线

5.1.2 精确率、召回率、准确率、F1-score

Attrition_Flag	precision	recall	accuracy	F1-score
Attrited Customer	75.2101	55.0769	89.8815	0.635879
Existing Customer	91.8345	96.5315	89.8815	0.941244

表 2 Lasso 回归性能指标

5.1.3 结果分析

由上述图像和数据可以看出，Lasso 回归在非流失客户的预测方面表现相对较好，具有 91.83%的精确度和 96.53%的召回率，显示其较高的准确性和对非流失客户的较好预测能力。然而，Lasso 回归在流失客户的预测方面表现一般，其召回率仅为 55.07%，说明在捕捉流失客户方面存在一定的不足。

这种现象可能源于 Lasso 回归的线性假设和 L1 正则化的影响。Lasso 回归对特征的线性关系进行建模，可能无法充分捕捉数据中存在的复杂非线性关系，尤其是在涉及到客户流失的情况下，可能存在非线性的模式。此外，L1 正则化的引入可能导致模型更加偏向于将一些特征的系数缩小为零，从而影响了对流失客户的有效识别。

5.2 决策树实验结果

5.2.1 混淆矩阵和 ROC 曲线

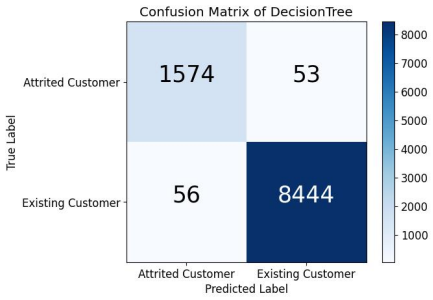

图 3-2-1 决策树混淆矩阵

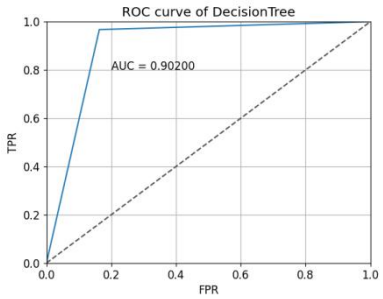

图 3-2-2 决策树 ROC 曲线

5.2.2 精确率、召回率、准确率、F1-score

Attrition_Flag	precision	recall	accuracy	F1-score
Attrited Customer	96.5644	96.7425	98.9237	0.966534
Existing Customer	99.3763	99.3412	98.9237	0.993587

表 3 决策树性能指标

5.2.3 结果分析

决策树的优势之一在于其非线性建模能力，能够捕捉复杂的数据关系。可以看出，它在银行客户流失预测中表现出色，主要体现在其在流失客户的预测方面取得了显著的成功，模型展现了高达 96.74%的召回率和 0.9665 的 F1 分数，这表明决策树在准确地识别真正流失客户方面非常强大。

除了性能指标之外，AUC 值为 0.902 也表明了决策树在 ROC 曲线下的性能表现相当不错。较高的 AUC 值暗示着模型在不同阈值下的性能稳健性，并在正例和负例之间提供了很好的区分度。

5.3 随机森林实验结果分析

5.3.1 混淆矩阵和 ROC 曲线

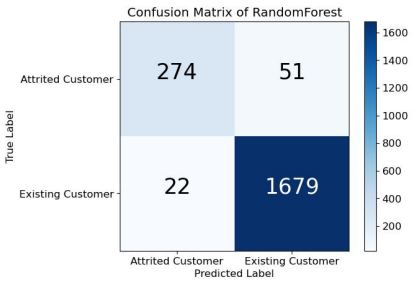

图 3-3-1 随机森林混淆矩阵

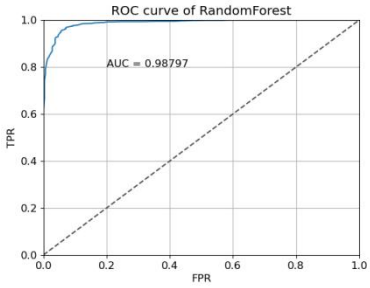

图 3-3-2 随机森林 ROC 曲线

5.3.2 精确率、召回率、准确率、F1-score

Attrition_Flag	precision	recall	accuracy	F1-score
Attrited Customer	92.5676	84.3077	96.3968	0.882448
Existing Customer	97.0520	98.7066	96.3968	0.978723

表 4 随机森林性能指标

5.3.3 结果分析

由以上图表和数据可以看出，随机森林模型在所有指标上表现都很好，随机森林在流失预测上表现更为出色，精确率高达 92.57%，且召回率较高，说明模型在检测流失客户时比较全面。对于未流失的预测同样非常准确，精确率达到 97.05%。随机森林通过集成多个决策树，有效克服了单个决策树的过拟合问题，具有更好的泛化能力。

5.4 支持向量机分类器实验结果分析

5.4.1 混淆矩阵和 ROC 曲线

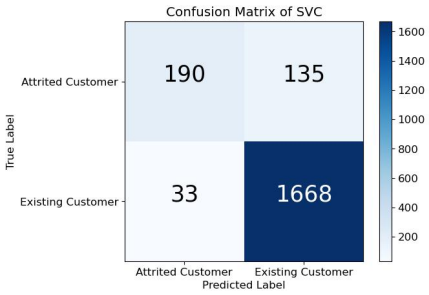

图 3-4-1 支持向量机混淆矩阵

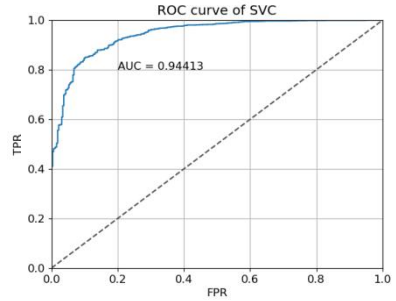

图 3-4-2 支持向量机 ROC 曲线

5.4.2 精确率、召回率、准确率、F1-score

Attrition_Flag	precision	recall	accracy	F1-score
Attrited Customer	85.2018	58.4615	91.7078	0.693430
Existing Customer	92.5125	98.0600	91.7078	0.952055

表 5 支持向量机分类器性能指标

5.4.3 结果分析

该支持向量分类器在非流失客户的预测上表现相对较好，然而，SVM 在流失客户的预测方面表现相对较差，其召回率近为 58.46%，表明在识别流失客户方面存在一定的挑战。

这一现象可能与 SVM 对噪声数据的敏感性，SVM 在高维度空间中表现良好，

但对于噪声较多或复杂关系较强的数据，其性能可能受到不良影响。

5.5 K-最邻近

5.5.1 混淆矩阵和 ROC 曲线

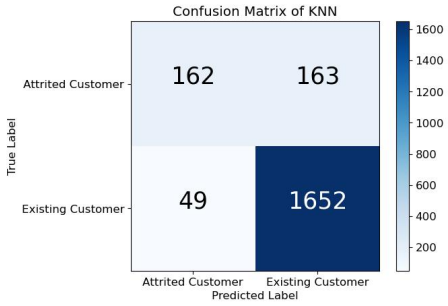

图 3-5-1 K-最近邻混淆矩阵

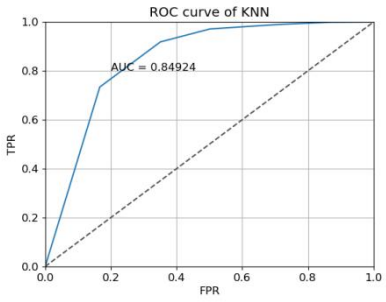

图 3-5-2 K-最近邻 ROC 曲线

5.5.2 精确率、召回率、准确率、F1-score

Attrition_Flag	precision	recall	accracy	F1-score
Attrited Customer	76.7773	49.8462	89.5360	0.604478
Existing Customer	91.0193	97.1193	89.5360	0.939704

表 6 K 最邻近性能指标

5.5.3 结果分析

可以看出，KNN 在非流失客户的预测方面表现相对不错，具有 91.02%的精确度和 97.12%的召回率，显示其对于非流失客户的较好预测能力。然而，在流失客户的预测方面表现较为一般，召回率仅为 49.85%。这可能是因为 KNN 对于比例敏感，且在处理大规模数据集时的计算成本相对较高。此外，KNN 对特征的比例敏感，因此在存在不同特征尺度的情况下，模型的性能可能受到影响。

5.6 神经网络（多层感知机分类器）

5.6.1 混淆矩阵和 ROC 曲线

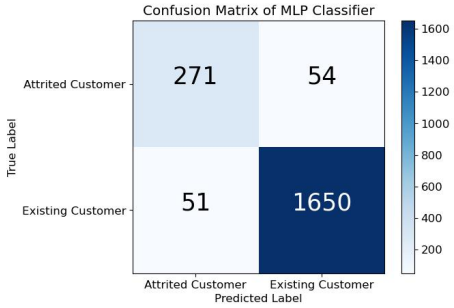

图 3-6-1 神经网络混淆矩阵

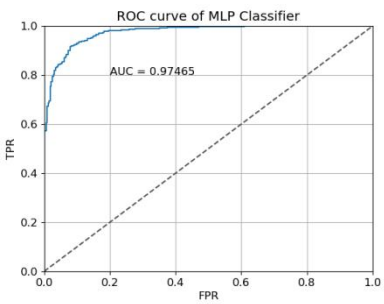

图 3-6-2 神经网络 ROC 曲线

5.6.2 精确率、召回率、准确率、F1-score

Attrition_Flag	precision	recall	accracy	F1-score
Attrited Customer	84.1615	83.3846	94.8174	0.837712
Existing Customer	96.8310	97.0018	94.8174	0.969163

表 6　神经网络性能指标

5.6.3 结果分析

由此可见，多层感知器在银行客户流失预测中表现卓越，具体体现在多个性能指标上。首先，MLP 呈现出高度的准确性，整体准确度高达 94.82%，说明模型在客户流失预测任务上取得了令人满意的整体性能。此外，MLP 在流失客户的召回率达到了 83.38%，精确度为 84.16%，显示出模型对于识别潜在流失客户方面的强大能力。

MLP 在非流失客户的预测方面同样表现卓越，在整个客户流失预测任务中的出色性能。此外，AUC（Area Under the Curve）值为 0.97465，进一步强化了 MLP 在分类问题中的优越性，即模型在正负样本之间能够有效区分。MLP 的这些卓越表现可能归因于其多层结构，允许模型学习并捕捉数据中的复杂非线性关系。

6 实验结果分析和结论

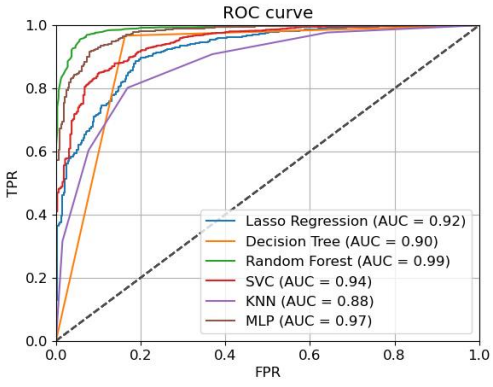

图 4-1 六个模型 ROC 曲线图

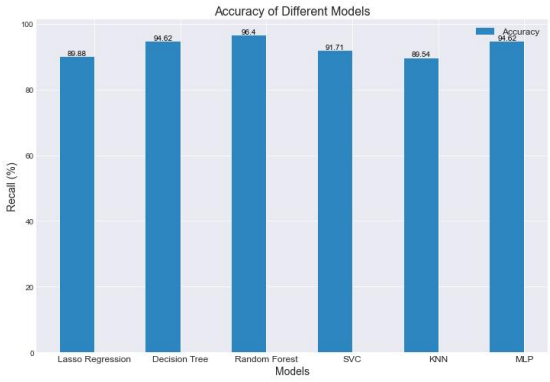

图 4-2　六个模型正准确率

models	Attrition precision	Existingprecision	Attritionrecall	Existingrecall	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

表 7　六个模型性能指标

6.1 流失客户预测

在本项目中，我们的目的是预测银行客户是否流失，因此我们以流失客户为阳性、未流失客户为阴性。一方面，我们需要关注实验结果的精确率，也即希望模型能够将实际流失的客户预测为流失；另一方面，对于错误预测的数据，我们认为假阴比假阳更重要，也即我们希望模型尽可能少地把实际流失的客户预测为未流失的，而相比之下将未流失的客户预测为已经流失的重要性优先级没有那么高。这主要是因为在实际情况中，流失流失客户对银行的业务影响通常更为显著。失去一个现有客户可能意味着失去了未来的利润，因为维护现有客户相对于获取新客户的成本更低。此外，流失客户可能带走的不仅仅是资金，还可能损害银行的声誉。如果某一客户已经流失，但却被认为未流失，银行就无法采取相应的措施留住他们，这样一来会造成很大损失。但如果只是将未流失的客户预测为已经流失的，相比于前者银行只是多一点客户维护的成本，而不会对整体的业务绩效

造成很大影响。

因此在模型的选择上，我们主要关注精确率和召回率，并希望这两个指标同处于高位，故对 F1 Score 的观察也可以直接帮助我们完成这一要求。基于这一目的，我们画出了 6 个模型的精确率、召回率和 F1 Score 的数值柱状图。

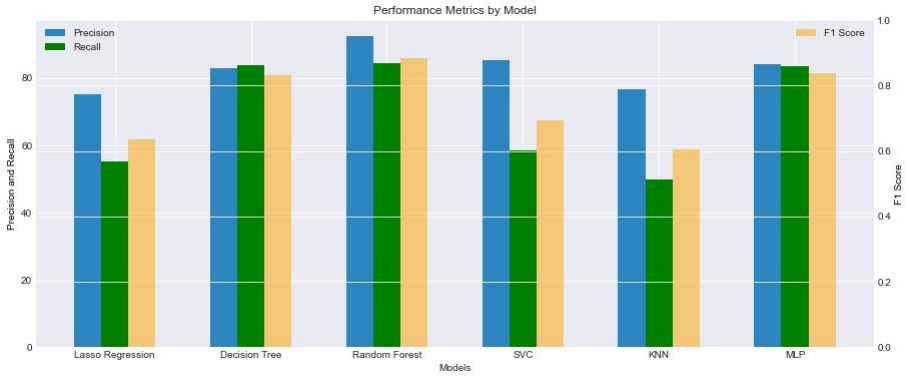

图 5-1　六个模型流失客户预测指标

通过对上图的分析，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。以下将对这两类算法进行更加详细的分析：

6.1.1 高性能类：

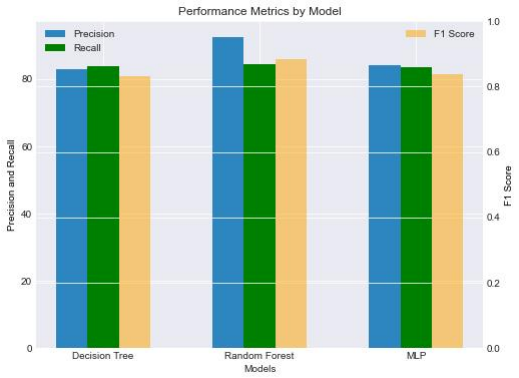

图 5-2 高性能模型流失客户预测指标

综合分析决策树、随机森林和多层感知机这三个高性能模型，可以发现这三个模型在所有指标中都表现良好，表现出较高的精确率，分别为 82.93%（Decision Tree）、92.57%（Random Forest）、84.16%（MLP），高准确率意味着模型在标识流失客户时有较低的误判率，对业务决策的可靠性有很大帮助。同时，这三个模型的召回率也都保持在较高水平，分别为 83.69%（Decision Tree）、

84.31%（Random Forest）、83.38%（MLP）。 高召回率表示模型对于实际流失客户的识别相对较全面，减少了漏诊的可能性。

在最终的选择模型时，可以针对不同模型的优势根据具体情况选择，例如决策树易于解释，随机森林适用于高维数据，而深度学习模型能够学习复杂的非线性关系。总的来说，这三个模型都适用于对于流失客户进行预测。

6.1.2 低性能类：

图 5-3 低性能模型流失客户预测指标

可以看出，逻辑回归、支持向量机和 K-最邻近算法，这三个模型在流失客户的预测中表现较差，召回率不超过 60%，F1-score 也只有不到 0.7.

低性能类别的模型可能对于流失客户的复杂关系建模能力不足，或者对数据特征的敏感性较低。逻辑回归和支持向量机在处理非线性关系方面可能相对受限，而 K 最邻近算法对于高维数据的处理可能较为困难。可以通过进一步调整参数进行优化。

6.2 模型的综合选择

上文中已然阐释，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，这一缺点使得在实际应用中，多层感知机的使用受到一定的限制。

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。

6.3 基于随机森林的特征分析

	0	1
0	Total_Trans_Amt	0.189420
1	Total_Trans_Ct	0.173149
2	Total_Ct_Chng_Q4_Q1	0.116235
3	Total_Revolving_Bal	0.093249
4	Total_Relationship_Count	0.070948
5	Avg_Utilization_Ratio	0.067318
6	Total_Amt_Chng_Q4_Q1	0.062129
7	Customer_Age	0.035123

表 8　随机森林模型输出权重前七的特征

上图是通过随机森林输出权重最大的特征排行截取。可以看到，权重在 0.05 以上的，从高到低分别为“过去 12 个月的交易总额”、“过去 12 个月的交易总数”“第 4 季度和第 1 季度相比交易数量的变化”、“信用卡上的循环余额总额”、“客户总持有银行产品的数量”、“平均卡片利用率”和“第 4 季度和第 1 季度相比交易金额的变化”。

通过以上特征可以知道，在过去的一年中交易金额多少、次数多少，相比一年前的交易次数和金额变化，这些对信用卡使用多少的数据，是判断客户是否即将流失的重要依据。此外，如果客户持有更多该银行的产品，也会更倾向于留下，反之则可能成为易流失的不稳定客户。

因此，我们认为，银行可以通过增加信用卡使用的便捷程度，比如建立界面简洁易操作的手机 APP 等，让用户可以即触即用。也可以定期发布相应的优惠政策，并且与客户积极沟通、研发对客户有利的产品，这些都有利于银行保留和扩大其信用卡用户。

7 不足与展望

7.1 挑战与不足

7.1.1 相似性混淆矩阵分析

通过决策树、随机森林和神经网络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。

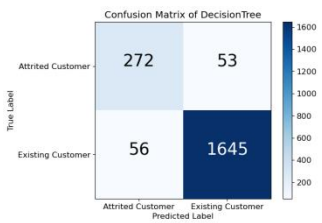

图 6-1 决策树混淆矩阵

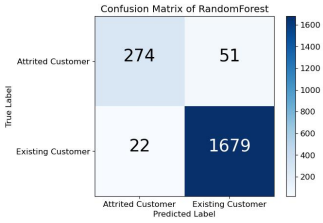

图 6-2 随机森林混淆矩阵

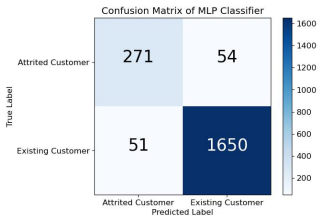

图 6-3 神经网络混淆矩阵

可以看到，这三个模型都把 50 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。但由于时间原因，这一过程我们还没有进行实验。希望在寒假中继续对这一部分工作的完善。

7.1.2 样本不平衡

前文数据预处理的阶段就提到，我们的数据集虽然基本温和事实，并且每一项数据也都吻合正态分布，但是作为标签的流失客户和未流失客户数量差距过大，可能由于这一不平衡导致最后训练所得的模型效果不佳。

针对这一点，我们最后采用了下采样的方式再次对所有模型进行训练。

7.2 针对数据集不平衡问题的优化——下采样

流失客户样本数为 1627，未流失客户的样本数为 8500，我们采取下采样的方法将未流失客户的数据减少到 1627，也即随机抽取 1627 份未流失客户数据集的样本，对这 3254 份数据进行训练。然后依然在原本划分的测试集上进行测试，结果如下图所示：

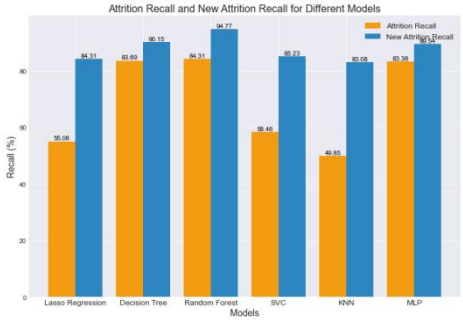

图 7-1 下采样前后流失客户召回率对比

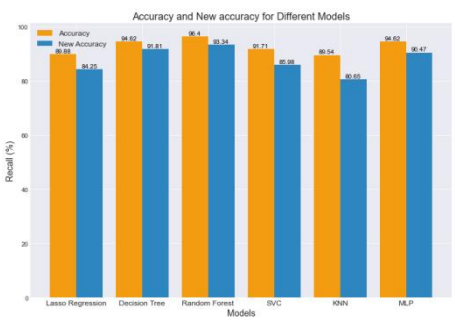

图 7-2 下采样前后模型准确率对比

可以看到，虽然样本数减少了，但是流失客户与未流失客户的数量平衡后，6 个模型的召回率几乎都上升到较高的水平，尤其是 SVC、KNN 以及 Lasso 回归的召回率有显著上升。而召回率对这一数据集来说非常重要，可见不同类别样本的平衡对于这一数据集的模型训练尤为关键。

此外，下采样后 6 个模型的准确率相比原来都有所下降，这一下降值相比于召回率的上升值较小。我们猜测，模型准确率下降主要是由于训练的样本总量变少所导致的。因此，我们建议银行能够收集尽可能多的数据或增加尽可能多的流失客户数据，这样一来模型就可以同时获得较高的召回率和准确率，从而达到比较好的预测效果。

这一实验也进一步让我们得出一个模型选择的结论，即对于 SVC、KNN 以及 Lasso 回归而言，输入样本类别的平衡大大可以减少发生预测中假阴的概率。故而，如果拿到的数据集恰好是一个分类问题，且需要分类的不同类别样本量不均衡，最好不要采用这三种模型。对于不平衡的样本，采用决策树、随机森林和神经网络会得到更好的预测效果。

8 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).

[4] Hughes, A. M. (1994). Strategic database marketing. Chicago: Probus Publishing Company.

[5] Kaymak, U. (2001). Fuzzy target selection using RFM variables. In IFSA World congress and 20th NAFIPS international conference, Vol. 2 (pp. 1038– 1043).

[6] Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., and Mason, C. H. (2006). Defection detection: measuring and understanding the predictive accuracy of customer churn models. Journal of Marketing Research 43(2), 204–211 (https://doi.org/10.1509/jmkr.43.2.204).

[7] Bhattacharya, C. B. (1998). When customers are members: customer retention in paid membership contexts. Journal of the Academy of Marketing Science 26(1), 31–44 (https://doi.org/10.1177/0092070398261004).

[8] Reichheld, F. F., and Sasser, W. E. (1990). Zero definitions: quality comes to services. Harvard Business Review 68(5), 105–111.

[9] Wu, Z., Li, Z. (2021). Customer churn prediction for commercial banks using customer-valueweighted machine learning models. Journal of Credit Risk, 17(4), 15-42.

[10] Bandam, A., Busari, E., Syranidou, C., Linssen, J., Stolten, D. (2022). Classification of building types in Germany: a data-driven modeling approach. Data, 7(4), 45.

[11] Günesen, S.N., Şen, N., Yıldırım, N., Kaya, T. (2021). Customer churn prediction in FMCG sector using machine learning applications, 82-103.

[12] Vezzoli, M., Zogmaister, C., Van den Poel, D. (2020). Will they stay or will they go? predicting customer churn in the energy sector. Applied Marketing Analytics, 6(2), 136-150.

[13] Kuznietsova, N., Bidyuk, P., Kuznietsova, M. (2022). Data mining methods, models and solutions for Big Data cases in telecommunication industry. In: [14]Babichev, S., Lytvynenko, V. (eds) Lecture Notes in Computational Intelligence and Decision Making. ISDMCI 2021. Lecture Notes on Data Engineering and Communications Technologies, 77. Springer, Cham.

[15] Sánchez, D.M., Moreno, A., López, M.D.J. (2022). Machine learning methods for automatic gender detection. International Journal on Artificial Intelligence Tools, 31(3).

[16] Xiahou, X., Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. Journal of Theoretical and Applied Electronic

Commerce Research, 17(2), 458-475.

[17] Huang, J. (2022). Real-time statistical method for marketing profit of Japanese cosmetics online cross-border e-commerce platform. In: Jiang, D., Song, H. (eds) Simulation Tools and Techniques. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 424. Springer, Cham.

[18] Jayadi, R., Kelvin, A., Jery, Rifyansyah, P., Mufarih, M., Firmantyo, H.M. (2020). Predicting customer churn of fire insurance policy: a case study in an Indonesian insurance company. Proceedings of the 6th International Conference on Science and Technology, ICST.

[19] Rabiul Alam, M.G., Hussain, S., Mim, M.M.I., Islam, M.T. (2021). Telecom customer behavior analysis using naïve bayes classifier. IEEE 4th International Conference on Computer and Communication Engineering Technology, CCET, 308-312.

[20] Kelley, K., Todd, M., Hopfer, H., Centinari, M. (2022). Identifying wine consumers interested in environmentally sustainable production practices. International Journal of Wine Business Research, 34(1), 86-111.

[21] Kiguchi, M., Saeed, W., Medi, I. (2022). Churn prediction in digital game-based learning using data mining techniques: logistic regression, decision tree, and random forest. Applied Soft Computing, 118.