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，以期能够找到预测这一数据集的最佳模型，并为银行提供较精准的流失客户预测服务。

1. 文献综述

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。

1. 数据集介绍和预处理

1、数据集简介

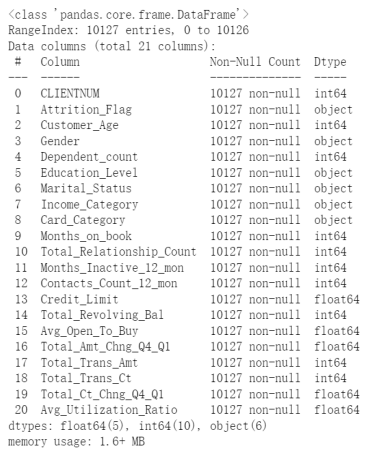
本研究采用数据集来自于全球最大的数据集平台之一Kaggle。这一数据集包含了从10127名消费者信用卡投资组合中收集的23个客户特征信息。这些特征包括全面的人口统计信息，如年龄、性别、附属卡数量（可近似家属人数）、受教育程度、婚姻状况和收入类别；以及每位客户与信用卡提供商关系的信息，如卡片类型、与银行交互的频率、持有银行产品的总数、过去一年中的不活跃月份数和活跃月份数。此外，它还包含了关于客户流失前消费行为的关键数据，如信用额度、总循环余额、过去12个月开放购买的信用额度；以及其它一些可分析指标如第4到第1季度的总变化金额、过去12个月的总交易额、过去12个月的总交易数、平均利用率和朴素贝叶斯分类器的流失标志(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。

2、数据预处理

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。



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。

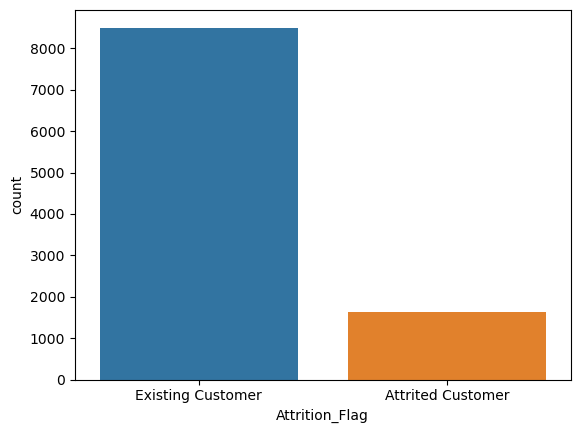
2.2 去除无效特征

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。

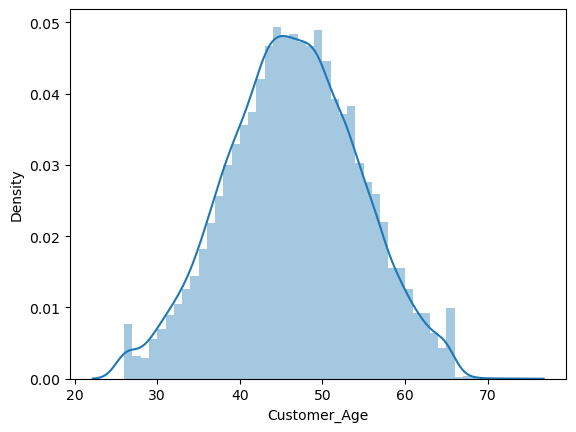
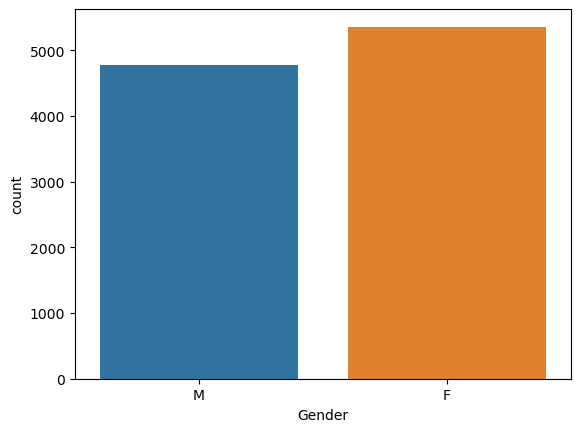
删除后，这一数据集变为10127行、20列。每一行包含了一个银行信用卡用户的信息，每一个用户有20个相关特征，且所有数据都没有缺失值。

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。



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。



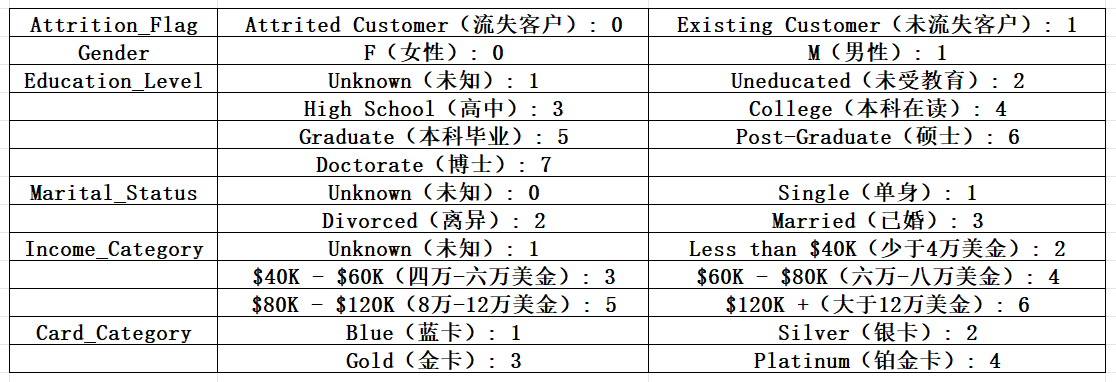
然后查看数据集中客户的性别和年龄分布。可以看到，虽然男性客户的个数要略少于女性客户的个数，但是总体而言分布比较均匀，接近1：1；而数据集中客户的年龄分布接近于正态分布。

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。

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。



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。

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。

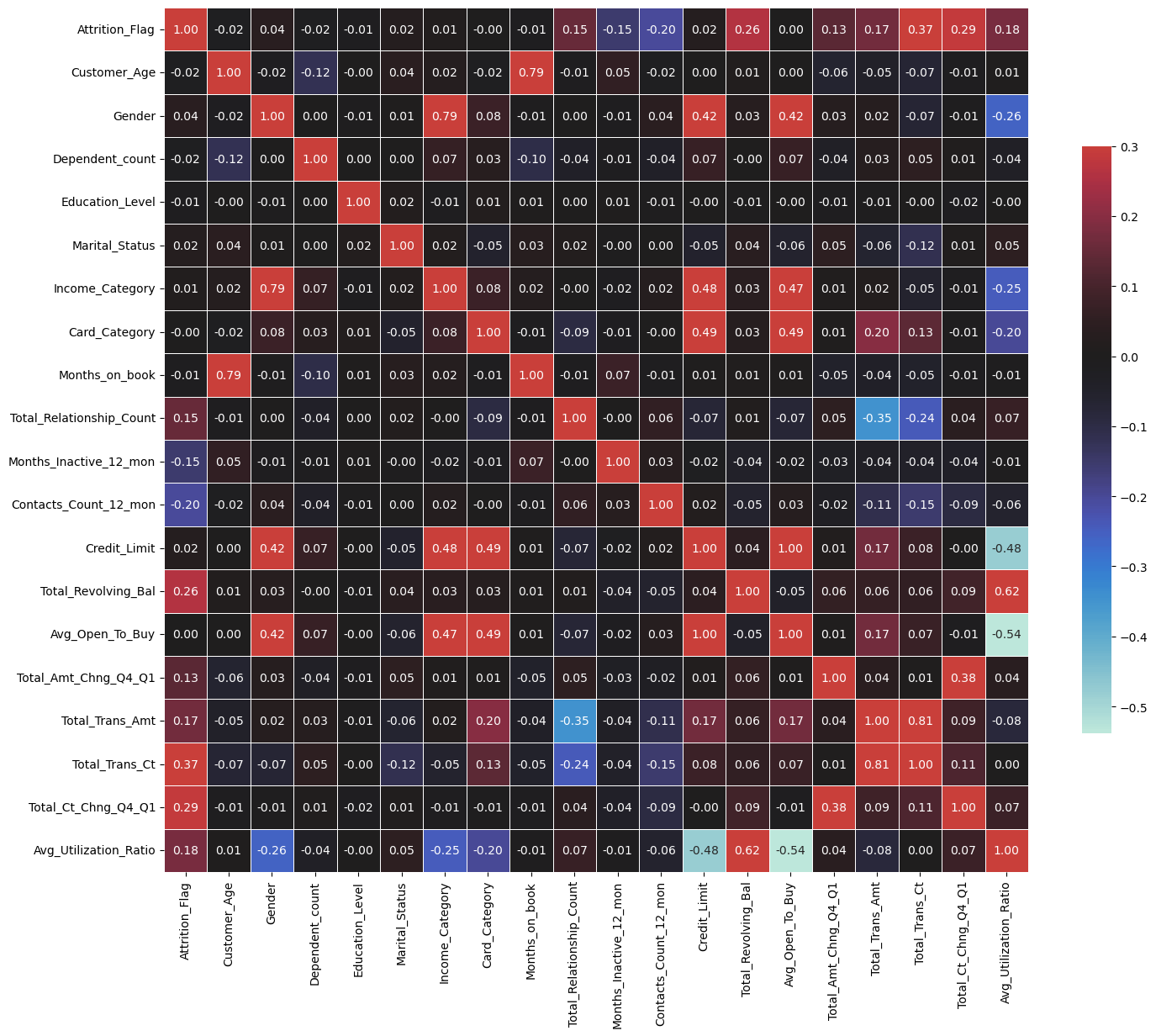
2.6 训练集和测试集的划分

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。

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。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，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，相关性系数未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。

1. 模型介绍和性能评价指标
2. 模型介绍

1.1 SVC

SVC（Support Vector Classifier）是一种用于分类问题的监督学习模型，它属于支持向量机（SVM）的一种。SVC模型的主要目标是找到一个最优的超平面，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。

SVC模型的工作原理是通过找到一个最大间隔超平面来进行分类。这意味着SVC试图找到一个能够最大化两个不同类别数据点之间的距离的超平面，以确保分类的准确性和泛化能力。在实际应用中，SVC模型可以使用不同的核函数来处理非线性可分的数据，如多项式核函数、高斯核函数等。

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。

决策树的构建过程通常包括以下步骤：

1）选择最佳的划分特征：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。

2）划分数据集：根据选择的划分特征，将数据集划分成多个子集。

3）递归构建子树：对每个子集递归地重复步骤1和步骤2，直到满足停止条件。

4）剪枝：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。

1.3 随机森林

随机森林(Random Forest)是一种集成学习(Ensemble Learning)方法，它通过构建多个决策树来进行分类或回归。随机森林是由Leo Breiman和Adele Cutler在2001年提出的，它是一种强大的机器学习算法，被广泛应用于数据挖掘、模式识别和预测等领域。

随机森林的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。在构建每棵决策树时，随机森林会随机选择一部分特征和样本进行训练，这样可以减少过拟合的风险，并提高模型的泛化能力。

1.4 朴素贝叶斯

朴素贝叶斯（Naive Bayes）是一种基于贝叶斯定理和特征条件独立假设的分类算法。它是一种简单而有效的分类方法，特别适用于文本分类和垃圾邮件过滤等领域。

朴素贝叶斯算法的基本思想是利用训练数据集中的特征和类别之间的关系，来对新的样本进行分类。它假设所有特征都是相互独立的，即给定类别的情况下，特征之间是条件独立的。这个假设虽然在现实中很难成立，但在实际应用中，朴素贝叶斯算法仍然表现出了很好的分类性能。

朴素贝叶斯算法的核心是贝叶斯定理，即根据先验概率和样本的特征来计算后验概率，然后选择具有最大后验概率的类别作为样本的分类结果。在实际应用中，通常使用拉普拉斯平滑来解决零概率问题，避免出现概率为零的情况。

1.5 KNN

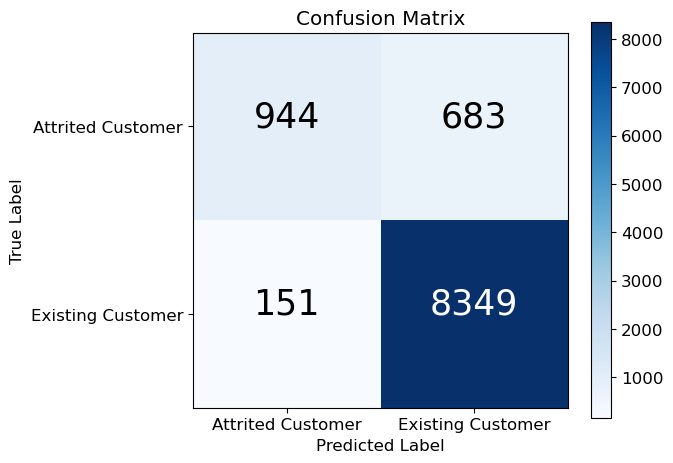
1.6 逻辑回归

1.7 全连接神经网络

1. 性能评价指标

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。



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。

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944）/（预测流失真实也流失944+预测流失真实未流失151）=86.210%。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8349）/（预测未流失真实也未流失8349+预测流失真实未流失151）=98.2235%。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944+预测未流失真实也未流失8349）/（总样本10127）=91.7646%。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91.7646%。

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。

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。

1. 实验结果及分析
2. 实验结果
   1. SVC
   2. 决策树
   3. 随机森林
   4. 朴素贝叶斯
   5. KNN
   6. 逻辑回归
   7. 全连接神经网络
3. 实验结果分析
4. 总结和不足
5. 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.

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.

[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.