# 预测信用卡使用活跃者

报告人：熊欣

时间：2019 年 11 月 12 日

## 目录

# 1 问题描述

随着经济的发展，信用卡市场逐步壮大并日益繁荣，信用卡逐渐成为居民个人消费使用最为频繁的非现金支付工具。商业银行的信用卡产品要想在信用卡市场中脱颖而出，就必须及时发现信用卡使用活跃者，对客户数据进行分析并提供个性化产品和服务。利用分类算法，从其业务系统中累积的大量客户历史数据中挖掘，帮助银行提早发现预测可能存在的信用卡使用活跃者。

本实验通过构建决策树模型，将数据集划分为训练集和测试集，对训练集进行训练和学习，对测试集数据进行有效预测其信用卡使用活跃度。实验过程中，需要利用剪枝操作对训练得到的模型进行优化，防止模型过拟合。并且在该数据集上应用四种分类算法 KNN、DT、SVM、NB，计算常用的评测指标 Accuracy，precision，recall，F1 和 AUC 值来对模型效果进行验证，对比分析各算法性能。本次实验包括以下三个问题：

（1）利用数据建立决策树模型，并尝试通过剪枝操作对模型进行优化。

（2）可视化决策树模型和解释规则， 探讨预测信用卡使用活跃者的重要指标。

（3）利用数据构建 KNN、DT、SVM、NB 模型，计算比较其分类评价指标 Accuracy，precision, recall, F1 和 AUC 值。

# 2 数据描述

从整体上了解数据的存储方式，本次实验数据集包含信用卡使用者的个人基本情况信息以及其是否为活跃用户的类别信息，数据存储格式为 csv。其中 600 条数据记录中有 274 人属于信用卡使用活跃者，326 人属于信用卡使用非活跃者，数据集较为均衡。由于数据集并未划分训练集和测试集，因此需要对数据集进行划分，训练集用于分类模型的训练学习，测试集用于测试分类模型的有效性和泛化能力。

从属性上了解数据的存储结构，包括年龄、性别、地区、收入情况、是否有孩子、是否有汽车、是否储蓄、是否有现金交易、是否有房贷和信用卡使用是否活跃。前十项属于预测信用卡使用是否活跃的特征，可供选择；信用卡使用是否活跃属于有待预测指标。

从微观上了解数据的存储内容，数据集中除首行表示属性之外，每一行都是一个独立的样本。按照训练集：测试集=8：2 的比例划分数据集，则训练集共 480 条数据，测试集共 120 条数据。年龄和收入状况为连续性数值变量；性别（0 表示女性，1 表示男性）、是否有汽车（0 表示无汽车，1 表示有汽车）、是否储蓄（0 表示无储蓄，1 表示有储蓄）、是否有现金交易（0 表示无现金交易，1 表示有现金交易）、是否有房贷（0 表示无房贷，1 表示有房贷）、是否结婚（0 表示没结婚，1 表示已结婚）为二值变量；地区为分类变量（0 表示 INNER_CITY，1 表示 RURAL，2 表示 SUBURBAN，3 表示 TOWN），需要将其转化为数值变量；是否有孩子为数值类型的序次变量。最后一个属性“信用卡使用是否活跃”用 0 或 1 表示，其中 0 表示信用卡使用不活跃，1 表示信用卡使用活跃，属于二分类问题。

基于上述的数据描述情况，以前十项作为划分属性，构建决策树模型，并通过 KNN、SVM、NB 分类算法对数据进行分析验证，探究信用卡使用活跃预测最佳方法。

# 3 方法描述

根据实验要求，本次实验主要基于信用卡使用是否活跃的数据构建决策树分类模型，并利用 Graphviz 实现模型可视化展示，对决策树模型的规则进行解释，并进一步分析预测信用卡使用活跃者的重要指标。同时，本次实验对该数据集进行 KNN、DT、SVM、NB 四种分类模型的训练验证，并利用 Accuracy，precision，recall，F1 和 AUC 的评测指标比较分析分类算法的性能。具体实验流程如图 1 所示。

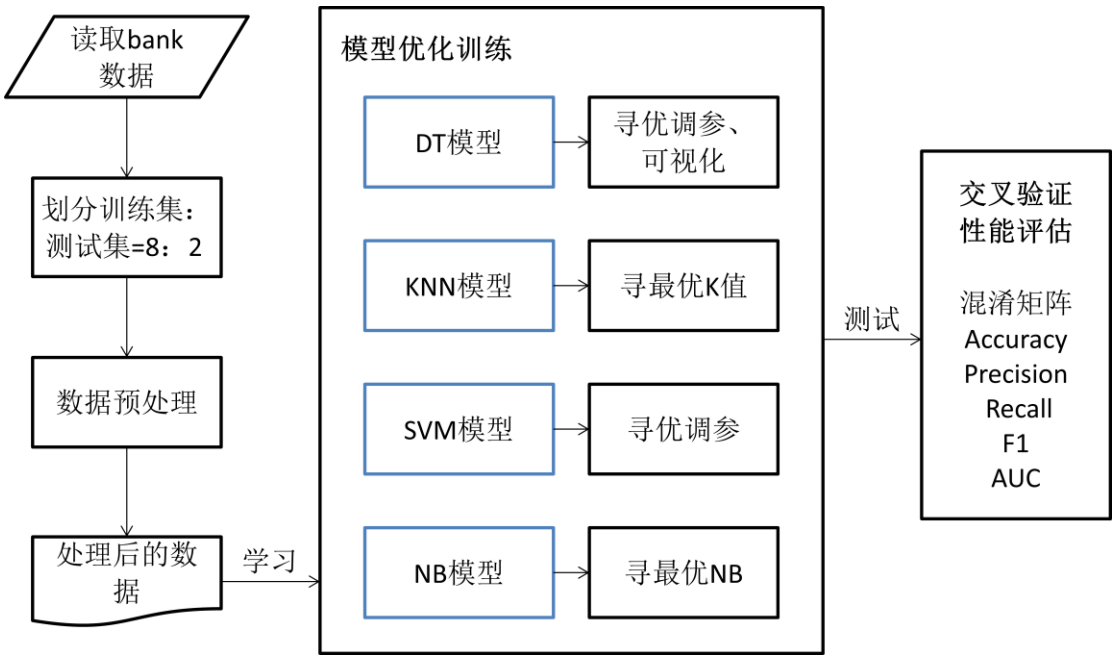


图 1 实验流程图

实验整体过程主要分为三个部分，第一个部分是对数据 bank.csv 的前期处理，包括数据的读取以及划分训练集和测试集，其中训练集用来进行模型的训练和学习，测试集用来对训练得到的模型的性能进行评估。数据预处理包括查看数据的类型和数据缺省值情况，本次实验数据无缺省值，主要是将非数值型的字符串或分类变量转化为机器可识别的数值分类变量，方便后续的数据分析。

第二个部分是模型的优化训练，对决策树模型进行寻优调参，并通过 Graphviz 实现决策树可视化展示；对 KNN 模型进行基于 F1 指标的最优 K 值训练；对 SVM 模型进行寻优调参并学习训练；对 NB 模型的三类算法——基于高斯分布、多项式分布及伯努利分布进行试验寻找最优 NB 训练模型。

第三个部分是通过五折交叉验证实现对四类分类模型的性能评估。对四类分类模型进行在测试集上的混淆矩阵计算；在训练集上进行五折交叉验证，比较计算四类分类模型的评测指标 Accuracy，precision，recall，F1 和 AUC 值，确定信用卡使用活跃者预测的最佳模型。

# 4 实验结果与分析

## 4.1 基于决策树的信用卡使用活跃者预测模型

### 4.1.1 模型构建及预剪枝优化

由于本案例数据集较小，故使用 GridSearchCV 网格搜索方法进行自动调参，只要把参数输进去，就能给出最优化的结果和参数。得到最优参数为{criterion= 'gini', max_depth= 7, min_samples_split= 14, splitter= 'best'}，即决策树模型的特征选择标准为 gini 指数，特征划分点选择标准为 best，决策树最大深度为 7，内部节点再划分所需最小样本数为 14，实现模型的预剪枝处理，在当前样本中训练效果较好，并以此进行后续的实验分析。

```
决策树模型的寻优调参：
hyperparameters={
        "criterion": ["entropy", "gini"],
        "splitter": ["best", "random"],
        "max_depth": range(1,20,2),
        "min_samples_split": range(3,16)
}
clf = DecisionTreeClassifier()
grid=GridSearchCV(clf, param_grid=hyperparameters, cv=10)
grid.fit(train[columns], train["pep"])
best_params=grid.best_params_
```

对上述构建的决策树进行可视化后，得到图 2 的效果图。其中，第一行表示当前节点的划分属性及其取值情况，左分支表示当前属性取值条件成立的情况，右分支表示当前属性取值条件不成立的情况；第二行表示当前节点数据的混杂度大小， 用 gini 指数表示在样本集合中一个随机选中的样本被分错的概率，值越小表示节点混杂度越低样本集合越纯；第三行表示当前节点的样本数，用 samples 标记；第四行表示当前节点样本标签取值的分布， 用 value 标记，第一个数值表示当前节点中标签为 0，类别 class 为 NO 的样本数， 第二个数值表示当前节点中标签为 1，类别 class 为 YES 的样本数；第四行表示当前节点样本所属类别，用 class 表示，当节点样本不纯显示样本量数目最多的标签。图中，节点的颜色显示了当前节点样本类别取值的情况， 不同颜色代表不同类别。蓝色表示 class=YES，即认为信用卡使用活跃的情况， 橙色表示 class=NO 的，即认为信用卡使用不活跃的情况。而节点颜色的深浅则显示了节点的混杂度大小，节点颜色越深说明当前节点样本的混杂度越低，样本越纯；节点颜色越浅则说

明当前节点样本的混杂度越高，样本纯度越低；特别地，当节点的颜色为白色的时候，说明节点混杂度最高，gini 取值为 0.5。

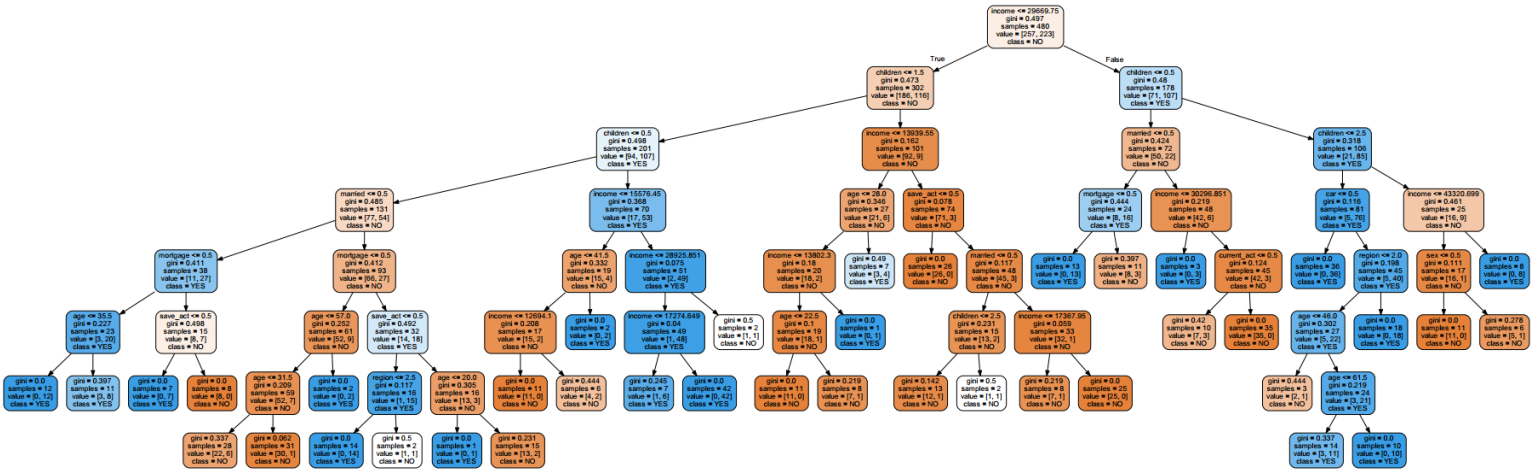

图 2 决策树可视化效果图

### 4.1.2 预测信用卡使用活跃的重要指标探讨

决策树模型可视化图中橙色和蓝色节点数量相当，与数据集中信用卡使用是否活跃的取值 NO:YES=326:274 的比例相符。详细查看每个节点的划分属性，可以发现收入（income）和孩子数量（children）是信用卡使用是否活跃预测的重要指标。使用这些指标进行预测可以有效降低样本的信息熵值，使得样本数据的不确定性降低。基于图 2 的可视化效果图，可以发现**判断信用卡使用是否活跃的首要指标是收入，以收入是否大于 29669.75 为划分节点，其次孩子数量成为进一步判断信用使用是否活跃的重要指标。**

### 4.1.3 决策树模型规则解读

决策树模型具有良好的可解释性。通过图 2，容易得到几条判断信用卡使用是否活跃的规则：

规则 1：if(收入 > 29669.75 and 孩子 > 0.5 and 孩子 > 2.5 and 收入 > 43320.699), then 信用卡使用活跃。即收入大于 43320.699 且孩子数 3 个的用户为信用卡使用活跃者。

规则 2：if(收入 > 29669.75 and 孩子 > 0.5 and 孩子 > 2.5 and 收入 <= 43320.699 and 性别 <= 0.5), then 信用卡使用不活跃。即收入大于 29669.75 而小于 43320.699 且孩子数 3 个且女性的用户为信用卡使用不活跃者。

规则 3：if(收入 > 29669.75 and 孩子 > 0.5 and 孩子 <= 2.5 and 汽车 <= 0.5), then 信用卡使用活跃。即收入大于 29669.75 且孩子数 1 或 2 个且无汽车的用户为信用卡使用活跃者。

规则 4：if(收入 > 29669.75 and 孩子 <= 0.5 and 结婚 <= 0.5 and 房贷 <= 0.5), then 信用卡使用活跃。即收入大于 29669.75 且无孩子且未结婚且无房贷的用户为信用卡使用活跃者。

规则 5：if(收入<= 29669.75 and 孩子 > 1.5 and 收入 > 13939.55 and 储蓄 <= 0.5), then 信用卡使用不活跃。即收入大于 13939.55 而小于 29669.75 且孩子数 2 或 3 个且无储蓄的用户为信用卡使用不活跃者。

……

### 4.1.4 决策树模型后剪枝优化

通过对图 2 决策树模型以及规则的进一步分析，发现决策树模型挖掘出来的规则存在冗余，决策树部分分支可能没用，如规则 1 中的孩子 > 0.5 and 孩子 > 2.5，收入 > 29669.75 and 收入 > 43320.699 等，因此在允许一定误差的范围内，可以采取后剪枝，通过规则裁剪来简化树结构。考虑在本案例实际应用中，**预测信用卡使用活跃者的价值较大**，即需要对判断信用卡使用活跃者的规则进行适当的精简，保证在某个可接受的误判概率下，有效预测出绝大部分存在的信用卡使用活跃者。如将规则 3 泛化为“if(收入 > 29669.75 and 孩子 > 0.5 and 孩子 <= 2.5), then 信用卡使用活跃”，此时能够识别出 85 条（38.12%）的信用卡使用活跃者，同时 21 条（8.17%）信用卡使用不活跃者会被误判。这样通过后剪枝操作可以有效减少过度拟合情况的发生，提高决策树模型的适用性和运行效率。

### 4.1.5 决策树模型效果评估

Accuracy（精准度），Precision（查准率），Recall（查全率），F1（F1 综合指标）和 AUC 值是分类算法效果评估的重要指标。 其中， Accuracy 能够从整体层面把控构建的分类模型对原始训练数据集的预测准确度； Precision 和 Recall 则能够从类别层面知悉构建的分类模型对每个类别下返回数据的预测准确度以及对每个类别应返回数据的实际掌握情况； F1 则综合 Precision 和 Recall，从类别层面给出构建的分类模型对每个类别数据特征的综合把握情况；AUC 值考虑消除样本倾斜的影响，能客观反映对正样本、负样本综合预测的能力。

表 1 决策树模型测试集混淆矩阵

| 混淆矩阵 | | 预测值 | |
|---|---|---|---|
| | | **Negative(0)** | **Positive(1)** |
| **真实值** | **Negative(0)** | 64 | 5 |
| | **Positive(1)** | 6 | 45 |

表 2 决策树模型评估结果

| 评估指标 / 类别名称 | **Accuracy** | **AUC** | **Precision** | **Recall** | **F1** | **Support** |
|---|---|---|---|---|---|---|
| **NO** | 0.908 | 0.905 | 0.91 | 0.93 | 0.92 | 69 |
| **YES** | | | 0.90 | 0.88 | 0.89 | 51 |
| **Avg/total** | | | 0.91 | 0.91 | 0.91 | 120 |

利用测试集数据对上述构建的决策树模型进行评估，得到表 1 和 2 所示的结果。从中可以发现，当前构造的决策树模型的 Accuracy 约为 0.91，说明当前模型能够准确预测测试集中 91%的数据。AUC 值约为 0.91，数值较高说明模型整体效果良好。其中，类别为 NO（标签为 0，即信用卡使用不活跃者）预测准确率达到 0.91，召回率 0.93，F1 综合指标达到 0.92；类别为 YES（标签为 1，即信用卡使用活跃者）预测准确率达到 0.90，召回率 0.88，F1 综合指标达到 0.89，说明决策数模型在犯错概率约 10%的情况下能够识别出 90%的实际情况中信用卡使用活跃者，效果较好。总体而言，**所构建的决策树模型效果较好，在测试集上的微平均值达到 0.91，能够有效预测信用卡使用是否活跃者。**

## 4.2 基于信用卡使用活跃者预测的分类模型性能对比

### 4.2.1 KNN 分类模型

KNN 最邻近分类算法的实现原理：为了判断未知样本的类别，以所有已知类别的样本作为参照，计算未知样本与所有已知样本 的距离，从中选取与未知样本距离最近的 K 个已知样本，根据少数服从多数的投票法则，将未知样本与 K 个最邻近样本中所属类别占比较多的归为一类。

本实验中选取 K 值为 1-50 进行实验，查看 KNN 模型在不同 K 取值情况下的 F1 指标变化情况，从而找出最优 K 值，如下图 3 所示。从图中可以看出，

KNN 模型随 K 值变化的 F1 指标波动相对较小，当 K 值取 27 时，F1 指标值最大，即后续实验 KNN 的参数 n_neighbors 设置为 27。

```
# 绘制折线图：基于 F1 值找出 KNN 模型的最优 K 值
def plot_dict(dictionary):
    pd.Series(dictionary).plot(figsize=(8,4), xlim=(1,50), ylim=(0.4,0.6), marker='o',
markersize=3)
    plt.xlabel("K")
    plt.ylabel("F1 score")
    plt.xticks(range(1,50,2))
    plt.show()

knn_f1 = dict()
for k in range(1,50,2):
    knn=KNeighborsClassifier(n_neighbors=k)
    knn.fit(train[columns], train['pep'])
    prediction = knn.predict(test[columns])
    f1=metrics.f1_score(test['pep'], prediction)
    knn_f1[k]=f1
plot_dict(knn_f1)
```

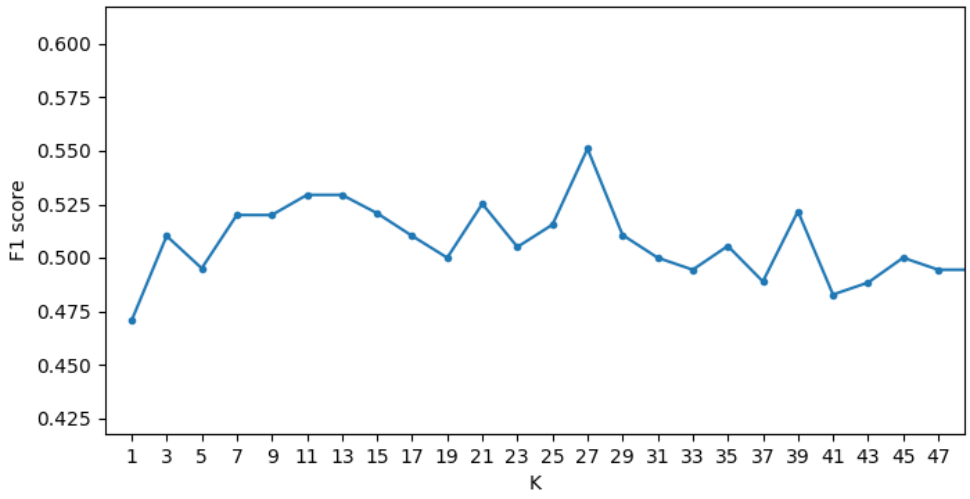


图 3 KNN 模型不同 K 值下的 F1 指标

对训练集进行 KNN 模型的学习后，在测试集上建立混淆矩阵，如下表 3 所示，同时在测试集上的各评测指标值如表 4 所示。从中可以发现，当前构造的 KNN 模型能够准确预测测试集中 63%的数据。AUC 值约为 0.62，说明模型整

体效果较差。其中，类别为 NO 预测准确率达到 0.67，召回率 0.71，F1 综合指标达到 0.69；类别为 YES 预测准确率达到 0.57，召回率 0.53，F1 综合指标达到 0.55，说明 KNN 模型仅能识别出约 55%的实际情况中信用卡使用活跃者，效果较差。

表 3 KNN 模型测试集混淆矩阵

| 混淆矩阵 | | 预测值 | |
|---|---|---|---|
| | | **Negative(0)** | **Positive(1)** |
| **真实值** | **Negative(0)** | 49 | 20 |
| | **Positive(1)** | 24 | 27 |

表 4 KNN 模型测试集评估结果

| 评估指标 / 类别名称 | **Accuracy** | **AUC** | **Precision** | **Recall** | **F1** | **Support** |
|---|---|---|---|---|---|---|
| **NO** | 0.63 | 0.62 | 0.67 | 0.71 | 0.69 | 69 |
| **YES** | | | 0.57 | 0.53 | 0.55 | 51 |
| **Avg/total** | | | 0.63 | 0.63 | 0.63 | 120 |

### 4.2.2 SVM 分类模型

SVM（support vector machine）又称为支持向量机，是一种二分类的模型。当然如果进行修改之后也是可以用于多类别问题的分类。支持向量机可以分为线性核非线性两大类。其主要思想为找到空间中的一个更够将所有数据样本划开的超平面，并且使得本本集中所有数据到这个超平面的距离最短。

本实验中，通过 GridSearchCV 网格搜索方法进行自动调参，得到最优参数为{kernel='rbf', C=2.0, gamma= 0.125, class_weight='balanced'}，在当前样本中训练效果较好，并以此进行后续的实验分析。对训练集进行 SVM 模型的学习后，在测试集上建立混淆矩阵，如下表 5 所示，同时在测试集上的各评测指标值如表 6 所示。从中可以发现，当前构造的 SVM 模型在测试集上预测所有数据为信用卡使用非活跃者，说明模型无法预测出信用卡使用活跃的用户。

表 5 SVM 模型测试集混淆矩阵

| 混淆矩阵 | 预测值 | |
|---|---|---|
| | **Negative(0)** | **Positive(1)** |

| | | | |
|---|---|---|---|
| **真实值** | **Negative(0)** | 69 | 0 |
| | **Positive(1)** | 51 | 0 |

表 6 SVM 模型测试集评估结果

| 评估指标 / 类别名称 | **Accuracy** | **AUC** | **Precision** | **Recall** | **F1** | **Support** |
|---|---|---|---|---|---|---|
| **NO** | 0.575 | 0.50 | 0.57 | 1.00 | 0.73 | 69 |
| **YES** | | | 0 | 0 | 0 | 51 |
| **Avg/total** | | | 0.57 | 0.57 | 0.57 | 120 |

### 4.2.3 NB 分类模型

朴素贝叶斯（Naive Bayesian）方法是一组基于贝叶斯定理的监督学习算法，其“朴素”假设是：给定类别变量的每一对特征之间条件独立。与其他更复杂的方法相比，朴素贝叶斯学习器和分类器执行速度非常快，它常见有三种不同的分布，分别是高斯 GaussianNB、多项式 MultinomialNB、伯努利 BernoulliNB。

本实验中，通过对三种不同的朴素贝叶斯算法进行训练集上的学习评估，最终选择分类效果最佳的高斯朴素贝叶斯算法作为该组实验的分类模型。对训练集进行 GaussianNB 模型的学习后，在测试集上建立混淆矩阵，如下表 7 所示，同时在测试集上的各评测指标值如表 8 所示。从中可以发现，当前构造的 GaussianNB 模型能够准确预测测试集中 64%的数据。AUC 值约为 0.63，说明模型整体效果较差。其中，类别为 NO 预测准确率达到 0.69，召回率 0.68，F1 综合指标达到 0.69；类别为 YES 预测准确率达到 0.58，召回率 0.59，F1 综合指标达到 0.58，说明 GaussianNB 模型在信用卡使用活跃预测性能上和 KNN 模型相当，效果较差。

表 7 GaussianNB 模型测试集混淆矩阵

| **混淆矩阵** | | **预测值** | |
|---|---|---|---|
| | | **Negative(0)** | **Positive(1)** |
| **真实值** | **Negative(0)** | 47 | 22 |
| | **Positive(1)** | 21 | 30 |

表 8 GaussianNB 模型测试集评估结果

| 评估指标 / 类别名称 | Accuracy | AUC | Precision | Recall | F1 | Support |
|---|---|---|---|---|---|---|
| **NO** | 0.64 | 0.63 | 0.69 | 0.68 | 0.69 | 69 |
| **YES** | | | 0.58 | 0.59 | 0.58 | 51 |
| **Avg/total** | | | 0.64 | 0.64 | 0.64 | 120 |

### 4.2.4 分类模型交叉验证比较

由于本次实验的数据集体量较小，通过交叉验证的方法来评估模型的预测性能，可以在一定程度上减小过拟合，更好的体现各模型在预测信用卡使用活跃问题上的效果。本次实验中将 KNN、SVM、NB、DT 四种分类算法在训练集上进行 5 折交叉验证，分别计算四个模型的评测指标 Accuracy, precision, recall, F1 和 AUC 进行比较分析，可以发现 SVM 模型无法预测信用卡使用活跃者，KNN 模型和 NB 模型的预测效果都较差，DT 模型能够较好的预测信用卡使用活跃者。其原因可以从算法的本质进行解释，本实验所构建的决策树模型是通过基尼系数来评估样本标签的混杂度，进而得到最优划分属性，因而不同标签的样本会不断区分，能够较好地预测信用卡使用活跃者。而 KNN 模型采用 K 近邻懒惰投票机制、SVM 模型寻找划分的最优超平面、NB 模型基于概率特征分布，它们在特征训练过程中可能对于特征的数量、质量等较为敏感，对于本实验未进行特征选择和工程的多特征二分类问题中表现不佳。

表 9 四类分类模型的评测指标比较

| 分类模型 / 评估指标 | KNN | NB | SVM | DT |
|---|---|---|---|---|
| **Accuracy** | 0.57 | 0.65 | 0.54 | 0.85 |
| **Precision** | 0.55 | 0.65 | 0.40 | 0.86 |
| **Recall** | 0.44 | 0.52 | 0.01 | 0.82 |
| **F1** | 0.48 | 0.58 | 0.02 | 0.84 |
| **AUC** | 0.60 | 0.71 | 0.50 | 0.88 |

## 5 实验总结

本实验首先通过对问题描述和数据描述对信用卡使用活跃者预测的数据集划分训练集和测试集，在此基础上对数据集的属性类型特征进行预处理。通过构建决策树模型，利用 Grid Search 网格搜索方法进行寻优调参，实现决策树

的预剪枝操作，并利用 Graphviz 实现决策树可视化。通过对可视化决策树进一步分析，可以得到信用卡使用活跃者预测的重要指标。判断信用卡使用是否活跃的首要指标是收入，以收入是否大于 29669.75 为划分节点，其次孩子数量成为进一步判断信用使用是否活跃的重要指标。在该数据集中，对训练集进行 KNN，DT，SVM 和 NB 四类分类模型的训练学习，对测试集建立混淆矩阵分析模型预测效果。最后，从 Accuracy, precision, recall, F1 和 AUC 这五个评测指标，对 KNN，DT，SVM 和 NB 四类分类模型在训练集上进行 5 折交叉验证，比较分析发现 DT 模型的预测效果最佳，其余分类模型效果均较差。

实验中遇到以下几个问题：对分类模型的调参使用不熟悉，通过 sklearn 学习模型及参数作用；对 SVM 模型底层算法理解不深，对于目标值的设置消耗不少时间。

本次实验中还存在一些不足：决策树模型的后剪枝分析仍不够深入，未给出全面剪枝策略；没有对数据集进行合理特征选择和特征工程，分类模型的训练过程可能未达到最优情况。

后续实验可以从以下几个方面进一步深入分析：尝试多种剪枝方案，包括预剪枝和后剪枝，寻找最优的剪枝方案；考虑对数据集进行有效的特征选择及特征工程，提升分类模型在该数据集预测上的准确性。

## 附：程序代码

见文件“信用卡活跃预测-熊欣.py”。