# 数学学院案例分析报告

案例名称：德国信贷数据集分类

报告负责人：魏丹怡（202032164）

完成时间：2021.04.05

# 目　录

# 一、数据来源和基本情况

## 1、数据来源

该数据集来源德国一家银行收集的客户信息，描述了客户在贷款前的状态。共收录了1000 个用户的数据，其中 700 个客户正常还款，300 个客户产生了违约。每一个用户采集了20 个特征，其中包含离散型的特征（例如贷款目的）13 个，连续型特征（例如贷款金额）7个。

## 2、基本情况

此数据集共有 1000 个样本，每个样本有 20 个特征，可以分为两类：无信用用户和有信用用户，分别用“0”和“1”来表示无信用和有信用。其中有信用有 700 例，无信用有 300例。

数据集的 20 个特征具体展式如下：

表 1 案例特征

| 特征 | 解释 |
| --- | --- |
| Creditability | 是否有信用：1:有信用（即无违约用户），0:无信用（即违约用户） |
| Account Balance | 现存支票账户状态，1：月流水为 0，2：月流水小于 200 马克，3：月流水大于等于 200 马克，4：未创建账户 |
| Duration of Credit (month) | 贷款时长 |
| Payment Status of Previous Credit | 历史信用情况。0：没有历史贷款或全部如期归还，1：在该银行的贷款已被全部归还，2：存在仍需还款的贷款，3：出现过延期归还情况，4：在其他银行仍存在贷款 |
| Purpose | 贷款目的。0：汽车（新），1：汽车（二手），2：家具装备，3：广播设备电视，4：家用电器，5：维修需要，6：教育，7：度假，8：再培训深造，9：生意，10：其他 |
| Credit Amount | 贷款金额 |
| Value Savings/Stocks | 储蓄账户状态。1：储蓄金额小于 100 马克，2：储蓄金额大于等于 100 马克小于 500 马克，3：储蓄金额大于等于 500 马克小于 1000 马克，4：储蓄金额大于等于 1000 马克，5：未知或不存在储蓄账户 |
| Length of current employment | 就职持续时间。1：暂无工作，2：就职小于 1 年，3：就职大于等于 1 年小于 4 年，4：就职大于等于 4 年小于 7 年，5：就职大于等于 7 年 |
| Instalment per cent | 分期付款百分比 |
| Sex & Marital Status | 性别与婚姻状况。1：男性离婚或分居，2：女性离婚、分居或已婚，3：男性单身，4：男性已婚或丧偶，5：女性单身 |
| Guarantors | 担保人。1：无担保人，2：联保，3：有担保人 |
| Duration in Current address | 当前住所的定居时间 |
| Most valuable available asset | 最有价值的资产。1：房地产，2：社保或人寿保险，3：汽车或其他，4：未知或无财产 |
| Age (years) | 年龄 |

| | |
|---|---|
| Concurrent Credits | 其他分期付款情况。1：银行，2：商铺，3：无 |
| Type of apartment | 住房类型。1：租赁，2：自己拥有，3：可免费居住 |
| No of Credits at this Bank | 在这家银行的信用额度 |
| Occupation | 职业。1：失业/非技术人员-非居民，2：非技术人员-居民，3：技术人员/官员，4：管理层/个体经营/高素质员工/管理人员 |
| No of dependents | 供养人口数量 |
| Telephone | 是否有以客户名称注册的电话.1：没有，2：有 |
| Foreign Worker | 外籍工作者。1：是，2：不是 |

# 二、关联性分析

## 1、关联性规则形式化描述

设$I=\{I_1,I_2,\ldots,I_m\}$是一个项集（item set），m 为项的个数，其中$I_i$表示第 i 个项，对应于一个个个人信息特征。事务（Transaction）$t_i$表示 I 的一个子集，对应于一个个订单。事务组成的集合记做 TID，每个事物中都包含若干个项。

关联性规则是形如$X \longrightarrow Y$的蕴含式，其中，X 和 Y 分别称为关联性规则的先导(antecedent 或 left-hand-side, LHS)和后继(consequent 或 right-hand-side, RHS) 。其中，关联规则 X Y，存在支持度和置信度，定义如下：

支持度$(X \longrightarrow Y)=\frac{\text{同时包含}X\text{和}Y\text{的事务数量}}{\text{所有事务数量}}$，理解为某一个项出现的概率。通常设置一个阈值 minsupport，当支持度不小于该值时认为是频繁项；

置信度$(X \longrightarrow Y)=\frac{\text{同时包含}X\text{和}Y\text{的事务数量}}{\text{包含}X\text{的事务数量}}$，理解为在 X 事务的基础上，X 和 Y 均出现的条件概率。

因此关联规则实际上包含两个子任务：

频繁模式发现：也称频繁模式挖掘、频繁项挖掘等，是指从一系列候选的项中选择频繁的部分，通常衡量频繁的程度可以是对每一项出现的频率，当超过某一阈值是则任务这个项是频繁的。

生成关联规则：在已经发现的最大频繁项目集中，寻找置信度不小于用户给定的 minconfidence 的关联规则。

## 2、Apriori 算法

Apriori 算法是经典的关联规则算法，其主要思路如下图所示：


算法流程：

（1）首先对数据库中进行一次扫描，统计每一个项出现的次数，形成候选 1-项集；

（2）根据 minsupport 阈值筛选出频繁 1-项集；

（3）将频繁 1-项集进行组合，形成候选 2-项集；

（4）对数据库进行第二次扫描，为每个候选 2-项集进行计数，并筛选出频繁 2-项集；

（5）重复上述流程，直到候选项集为空；

（6）根据生成的频繁项集，通过计算相应的置信度来生成管理规则。

Apriori 算法的特点：

简单且易于实现，是最具代表性的关联规则挖掘算法。随着数据集规模的不断增长，逐渐显现出一定的局限性：

（1）需多次扫描数据库，很大的 I/O 负载，算法的执行效率较低；

（2）产生大量的候选项目集，尤其是候选 2-项集占用内存非常大，会消耗大量的内存；

（3）对于每一趟扫描，只有当内存大小足够容纳需要进行计数的候选集时才能正确执行。如果内存不够大，要么使用一种空间复杂度更小的算法，要么只能对一个候选集进行多次扫描，否则将会出现“内存抖动”的情况，即在一趟扫描中页面频繁地移进移出内存（页面置换算法也无法避免内存抖动问题），造成运行时间的剧增。

## 3、实证研究

### （1）问题提出与指标选取

利用 R 语言 arules 包库中的 Apriori 算法进行该操作。由于电脑运行内存有限，故我们需要将自己感兴趣的研究项目先根据自己的认识，预先筛选变量，再利用 Apriori 算法，完成关联性分析。

本次主要探索的是婚姻状况的影响因素，由于这里对于女性的分类只有单身和非单身，所以只将男性作为考虑对象。针对男性婚姻状况的影响因素作者认为主要可分为存款、年龄、不动产、职业、国籍。

表 2 男性婚姻状况影响因素

| 目标层 | 标准层 | 指标层 | 指标解释 |
|---|---|---|---|
| 男性婚姻状况 | 财务状况 | Value Savings/Stocks | 储蓄账户金额: 1：小于 100，2：金额大于等于 100 小于 500，3：大于等于 500 小于 1000 4：大于等于 1000， 5：未知或不存在储蓄账户 |
| | | Account Balance | 现存支票账户月流水: 1： 0, 2: 小于 200, 3：大于等于 200，4：未创建账户 |
| | | Creditability | 是否有信用：1:有信用，0:无信用 |
| | 年龄 | Age (years) | |
| | 不动产 | Type of apartment | 住房类型。1：租赁，2：自己拥有，3：可免费居住 |
| | 职业 | Occupation | 职业。1：失业/非技术人员-非居民，2：非技术人员-居民，3：技术人员/官员，4：管理层/个体经营/高素质员工/管理人员 |
| | | Length of current | 就职持续时间。1：暂无工作，2：小于 1 |

| | | | |
|---|---|---|---|
| | | employment | 年，3：大于等于 1 年小于 4 年，4：大于等于 4 年小于 7 年，5：大于等于 7 年 |
| | 国籍 | Foreign Worker | 外籍工作者。1：是，2：不是 |

**（2）数据预处理**

由于数据集中年龄为连续型变量，不符合 Apriori 算法中对于数据需为离散型数据，故应将年龄数据进行分类，由于年龄数据范围是 20-75 岁，则我们将 20-25 定义为第一类，25-35 第二类，35-45 第三类，45 以上为第四类。

## 4、运行结果

根据 Apriori 算法筛选出与男性婚姻相关的且提升度大于 1.25 的关联规则自己，并根据支持度从大到小排列，输出前五条结果：

```
> inspect(sort(x1,by="support")[1:5])     # 排序后，查看前5条关联规则
    lhs                                rhs                    support confidence   coverage     lif
t count
[1] {Lengthofcurrentemployment=5,

     Age=3,

     Occupation=3}                => {MaritalStatus=3} 0.05942029  1.0000000 0.05942029 1.25912
4    41
```

从输出结果可以看出：其关联性规则主要针对于类别标签为 3 的男性单身以及类别为 4 的男性已婚或丧偶。其中外籍、年龄在 35-45 岁之间职业为技术人员或官员且在职持续时间大于七年的男性倾向于单身。年龄在 25-35 岁之间拥有自己住房的男性、年龄在 25-35 岁之间无违约记录的男性已婚或离异者居多。

# 三、特征分析

## 1、数据的统计特征分析

对样本的 20 个特征进行描述性统计，如下表所示：

表 3 特征描述性统计

| 特征 | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|
| Account Balance | 2.58 | 1.26 | 1.00 | 1.00 | 2.00 | 4.00 | 4.00 |
| Duration of Credit (month) | 20.90 | 12.06 | 4.00 | 12.00 | 18.00 | 24.00 | 72.00 |
| Payment Status of Previous Credit | 2.55 | 1.08 | 0.00 | 2.00 | 2.00 | 4.00 | 4.00 |
| Purpose | 2.83 | 2.74 | 0.00 | 1.00 | 2.00 | 3.00 | 10.00 |
| Credit Amount | 3271.25 | 2822.75 | 250.00 | 1365.50 | 2319.50 | 3972.25 | 18424.00 |
| Value Savings/Stocks | 2.11 | 1.58 | 1.00 | 1.00 | 1.00 | 3.00 | 5.00 |
| Length of current employment | 3.38 | 1.21 | 1.00 | 3.00 | 3.00 | 5.00 | 5.00 |
| Instalment per cent | 2.97 | 1.12 | 1.00 | 2.00 | 3.00 | 4.00 | 4.00 |
| Sex & Marital Status | 2.68 | 0.71 | 1.00 | 2.00 | 3.00 | 3.00 | 4.00 |
| Guarantors | 1.15 | 0.48 | 1.00 | 1.00 | 1.00 | 1.00 | 3.00 |
| Duration in Current address | 2.85 | 1.10 | 1.00 | 2.00 | 3.00 | 4.00 | 4.00 |
| Most valuable available asset | 2.36 | 1.05 | 1.00 | 1.00 | 2.00 | 3.00 | 4.00 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Age (years) | 35.54 | 11.35 | 19.00 | 27.00 | 33.00 | 42.00 | 75.00 |
| Concurrent Credits | 2.68 | 0.71 | 1.00 | 3.00 | 3.00 | 3.00 | 3.00 |
| Type of apartment | 1.93 | 0.53 | 1.00 | 2.00 | 2.00 | 2.00 | 3.00 |
| No of Credits at this Bank | 1.41 | 0.58 | 1.00 | 1.00 | 1.00 | 2.00 | 4.00 |
| Occupation | 2.90 | 0.65 | 1.00 | 3.00 | 3.00 | 3.00 | 4.00 |
| No of dependents | 1.16 | 0.36 | 1.00 | 1.00 | 1.00 | 1.00 | 2.00 |
| Telephone | 1.40 | 0.49 | 1.00 | 1.00 | 1.00 | 2.00 | 2.00 |
| Foreign Worker | 1.04 | 0.19 | 1.00 | 1.00 | 1.00 | 1.00 | 2.00 |

上表展式了 20 个特征的平均值，标准差，最大最小值以及四分位数。对于标准差来说，Credit Amount、Duration of Credit (month)和 Age (years)的值较大，分别为 2822.75、12.06 和 11.35，也就是说不同样本之间这几个特征的差异最大，在之后选择分析特征时可以给予着重考虑。

## 2、特征相关性

在本案例中，每个样本有 20 个特征，若将所有特征纳入考虑，无疑大大增加了计算复杂度，因此，首先对特征相关性进行分析。绘制特征相关性热力图如下：

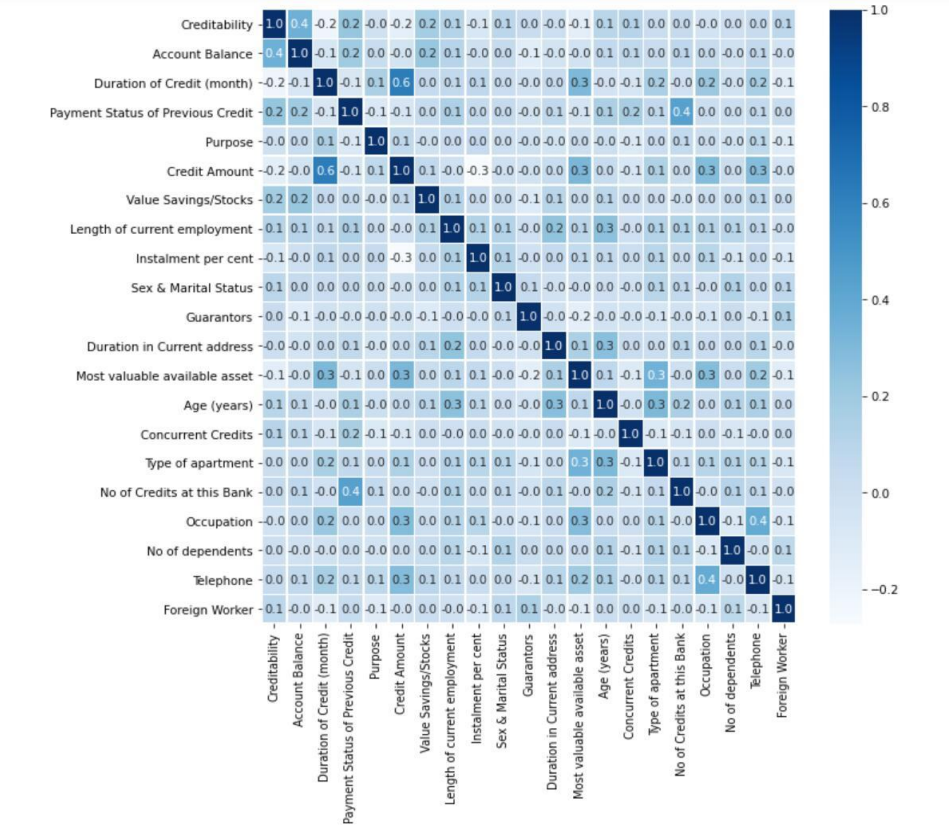



图 1 特征相关性

上图表现出了样本 20 个特征之间的相关性。首先，每个特征与自身的完全相关性是毋庸置疑的；其次，各个特征之间也有一定的相关性，但其相关性普遍不高，因此我们无法直接从中挑选出对信用情况影响较大的特征。可以通过 PCA 方法进行降维处理。

### 3、特征对信用影响分析

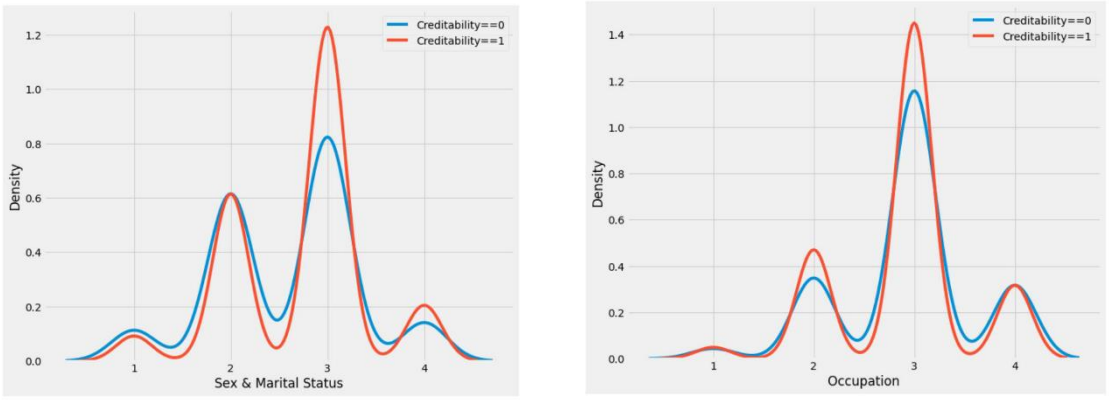



图 2 核密度估计图

绘制了不同性别及婚姻状况以及不同职业对应的信用状况的核密度估计图。从图像可以看出，单身男性信用差的人数明显多于信用好的人数，而在已婚（离婚）男性及女性中，有信用和无信用的比例基本为 1：1。因此，单身男性信用相对较差。失业和管理层/个体经营/高素质员工/管理人员中有无信用差别不大，但非技术人员以及技术人员和官员中信用好的人更多。

## 四、特征工程

### 1、PCA

采取 PCA 提取特征，设置主成分个数为 3，得到其贡献率如下图所示：

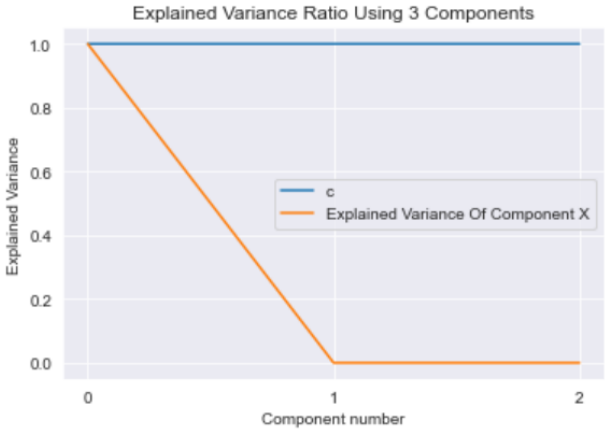



图 3 主成分贡献率

可以看出，所有主成分的累计贡献率为 99.99%，其中第一主成分的贡献率最大，可以以这三个主成分进行后续分类。绘制主成分热图如下：

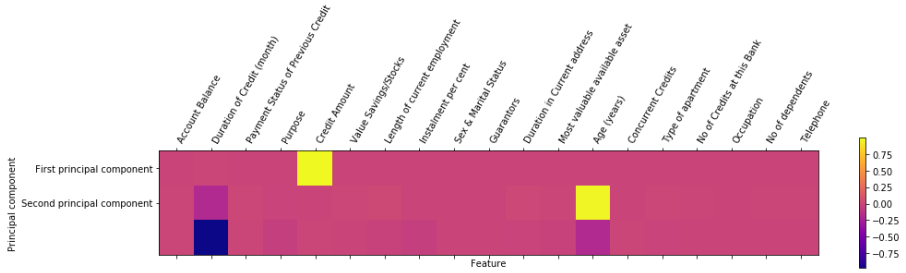



图 4 主成分热图

上图显示了各个特征在主成分中的占比，可以看出 Credit Amount、Duration of Credit (month)和 Age (years)三个特征对主成分的影响最大。并且这三个特征均为连续性变量，因此在 SVM 及 LogisticRegression 算法中，也可以采用这三个特征来训练模型。

## 2、结论

(1)绘制散点图矩阵及三维图分析三个主成分之间的关系：

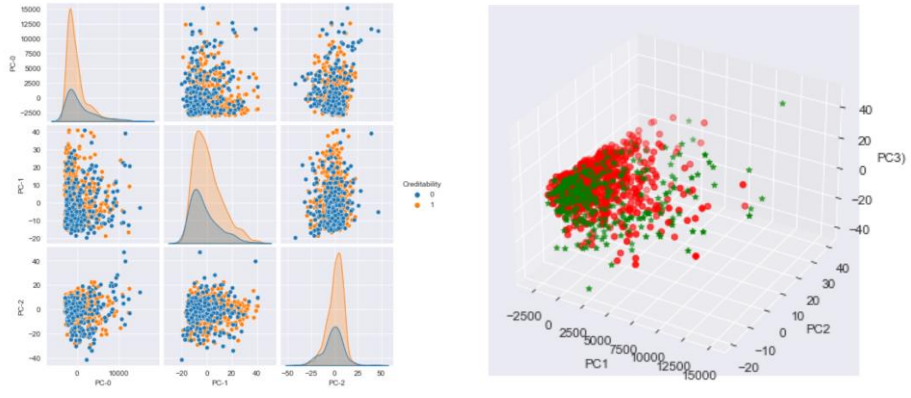



图 5 主成分之间的关系

由上面两幅图可以看出，在使用三个主成分进行分类时，分类边界不够明显。

(2)绘制散点图矩阵及三维图分析 Credit Amount、Duration of Credit (month)和 Age (years)这 3 个特征之间的关系：

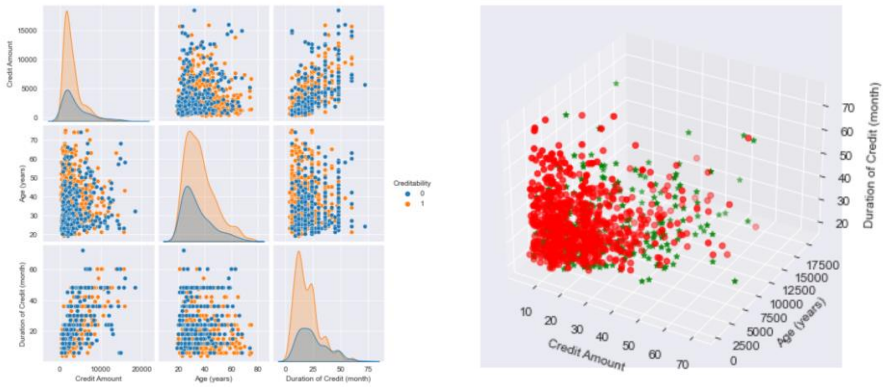



图 6 特征之间的关系

由上面两幅图可以看出，在选择这 3 个特征时，不同类别之间样本存在分类边界，但不够清晰，因此在后续分类中，可以采用这 3 个特征，但使用线性分类器分类时，结果精确度

可能不够高。

综上，选择 Credit Amount、Duration of Credit (month)和 Age (years)这 3 个特征进行后续实验。

# 五、德国信贷数据的分类

## 1、形式化和分类器数学推导

### （1）SVM 算法

SVM 的核心在于找到一个超平面将两类样本准确的分开，同时保证间隔尽可能的大，这样会有更好的泛化能力。

设超平面方程为:$w^T x + b = 0$,我们需要做的找到这样一个超平面划分两类信用状况并使得德国信贷训练集数据上的点到这个超平面的间隔距离尽可能的远。接着将原问题通过拉格朗日对偶算法转换为其对偶问题，其主要原因是自然引入核函数从而降低求解复杂度。其对偶优化问题如下，我们解决该优化问题即可。

$$max\ L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$s.t.$$

$$\begin{cases} \sum_{i=1}^{n} \alpha_i y_i = 0 \\ \alpha_i > 0, i = 1,2...n \end{cases}$$

### （2）Logistic Regression 算法

逻辑回归是监督学习中的一种，它根据大量带有分类标签的特征变量来训练优化模型，在根据模型来预测只有特征变量的分类标签。在德国信贷案例中，我们通过许多带有分类标签（信用状况有两种类别）的特征变量数据来训练预测信用状况类别的模型。为实现预测分类问题，我们使用了 Logistic 模型及 Sigmoid 函数:

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

Sigmoid 函数有一些特点，比如当$z = 0$是$log_i(\mathrm{z}) = 0.5$当$z < 0$时，$0 < log_i(\mathrm{z}) < 0.5$;当$z > 0$时，$0.5 < log_i(\mathrm{z}) < 1$。所以$log_i(\mathrm{z})$函数的取值范围为$(0,1)$。

其中回归的基本方程为:

$$z = w_0 + \sum_i^N w_i x_i$$

我们可以把$log_i(\mathrm{z})$的函数值看成类别为 1 的概率预测值，当$log_i(\mathrm{z}) < 0.5$时，我们预测的分类为 0；当$log_i(\mathrm{z}) >= 0.5$时，我们预测的分类为 1。这样我们就可以很容易的对德国信贷数据集分类，即有信用或者无信用。

### （3）随机森林算法

随机森林则是由多棵决策树组合而成的一个分类器。因为如果只有一棵决策树，预测的

结果可能会有比较大的偏差，而利用多棵决策树进行决策，再对所有决策树的输出结果进行统计，取票数最多的结果作为随机森林的最终输出结果。

随机森林由Leo Breiman（2001）提出，它通过自助法重采样技术。对于本次关于德国银行信贷的数据集来说，就是从原始训练样本集N=800中有放回地重复随机抽取K=10个样本生成新的训练样本集合，然后根据自助样本集生成10个决策树组成随机森林，测试集的分类结果按决策树投票多少形成的分数而定。随机森林实质是对决策树算法的一种改进，将多个决策树合并在一起，每棵树的建立依赖于一个独立抽取的样品，森林中的每棵树具有相同的分布，分类误差取决于每一棵树的分类能力和它们之间的相关性。

## 2、计算机仿真过程

### （1）SVM

在jupyter notebook中调用svc构建基于支持向量机的德国信贷分类模型。

### （2）Logistic Regression

在jupyter notebook中调用sklearn.linear_model包构建基于逻辑回归的德国信贷分类模型。

### （3）随机森林

在jupyter notebook中调用sklearn.ensemble包构建基于随机森林的德国信贷分类模型。

## 3、评估方法和数据划分

### （1）混淆矩阵

混淆矩阵向我们展式了查准率(准确率)与查全率(召回率)：

查准率(P) = $\frac{TP}{TP+FP}$ ，即在被判别为正类别的样本中，确实为正类别的比例是多少；

查全率(R) = $\frac{TP}{TP+FN}$ ，即在所有正类别样本中，被正确判别为正类别的比例是多少。

### （2）ROC曲线

模型训练完成之后,每个样本都会获得对应的两个概率值,一个是样本为正样本的概率,一个是样本为负样本的概率。把每个样本为正样本的概率取出来，进行排序，然后选定一个阈值，将大于这个阈值的样本判定为正样本，小于阈值的样本判定为负样本，可以得到两个值，一个是真正率，一个是假正率：

真正率（TPR）= $\frac{TP}{TP+FN}$ ，即模型判定为正样本且实际为正样本的样本数与所有的正样本数之;

假正率（FPR）= $\frac{FP}{TN+FP}$ ，即模型判定为正样本实际为负样本的样本数与所有的负样本数之比。

我每选定一个阈值，就能得到一对真正率和假正率，由于判定为正样本的概率值区间为[0,1]，那么阈值必然在这个区间内选择，因此在此区间内不停地选择不同的阈值，重复这个过程，就能得到一系列的真正率和假正率，以这两个序列作为横纵坐标，即可得到ROC曲线了。而ROC曲线下方的面积，即为AUC值。

### (3)数据划分

在本案例中，共有 1000 个样本，其中正例 700 个，负例 300 个。将样本的 80%划分为训练集，20%划分为测试集，并训练集、测试集中正、负例的比例与原数据集尽量一致。划分结果如下：

```
1    562
0    238
Name: Creditability, dtype: int64
1    138
0     62
Name: Creditability, dtype: int64
```

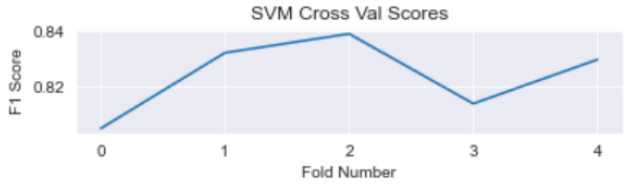
## 4、分类结果和比较统计

### （1）SVM

通过 Sklearn 的 SVC 模型对德国信贷数据集进行分类，分类结果如下：

```
预测 - 测试结果精确率: 0.7195767195767195
预测 - 测试结果召回率: 0.9855072463768116
预测 - 测试结果f1_score: 0.8318042813455658
预测 - 测试结果AUC: 0.5653342683496961
```

对分类结果进行 5 折交叉验证，其 F1 值展示如下：




```
[0.8046875  0.83206107 0.83895131 0.81368821 0.82962963]
均分:  0.82380354442422
```

图 7　SVM 交叉验证

可以看出，模型分类的 F1 值平均为 82.4%，通过混淆矩阵来观察预测分类和实际分类情况，绘制混淆矩阵的热点图如下：

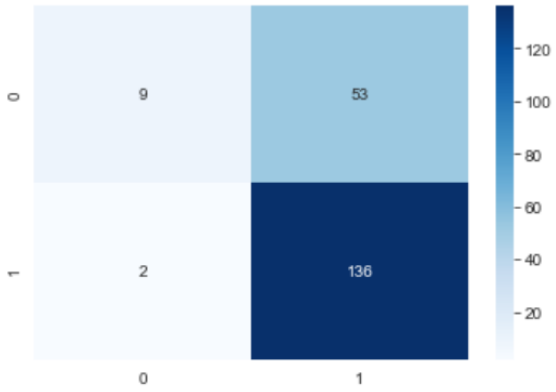



图 8　SVM 混淆矩阵

由上图可知，此模型将本应分类为无信用（标签为 0）的 53 个样本误分为有信用（标签为 1），同时将本应分类为有信用（标签为 1）的 2 个样本误分为无信用（标签为 0）。对比来看，此模型对于有信用（标签为 1）样本预测表现更好。

之后，绘制 ROC 曲线对此模型进行评估，如下所示：

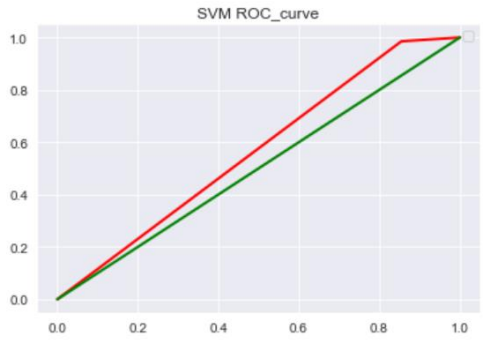



图 9　SVM ROC 曲线

由于将大量无信用（标签为 0）客户分类为有信用（标签为 1）客户，用 ROC 曲线来评价模型时，模型表现不能让人十分满意。

### （2）Logistic Regression

通过 Sklearn 的 LogisticRegression 模型，取 C=1000，solver='lbfgs'，对德国信贷数据集进行分类，结果如下：

```
混淆矩阵列 [[ 11  46]
 [  4 139]]
准确率 0.75
              precision    recall  f1-score   support

           0       0.73      0.19      0.31        57
           1       0.75      0.97      0.85       143

    accuracy                           0.75       200
   macro avg       0.74      0.58      0.58       200
weighted avg       0.75      0.75      0.69       200
```

可以看出，模型分类的精度达到了 75%，通过混淆矩阵来观察预测分类和实际分类情况，绘制混淆矩阵的热点图如下：

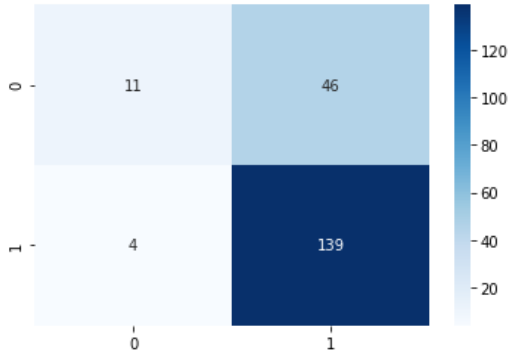



图 10　LR 算法混淆矩阵

由上图可知，此模型将本应分类为无信用（标签为 0）11 个样本误分为有信用（标签为 1），同时将本应分类为有信用（标签为 1）的 4 个样本误分为无信用（标签为 0）。也就是说，此模型在无信用的分类中，查准率(P)达到了 73%，查全率（R）只达到 19%；在有信用的分

类中，查准率(P)达到了 75%，查全率（R）达到了 97%。

之后，绘制 ROC 曲线对此模型进行评估，如下所示：

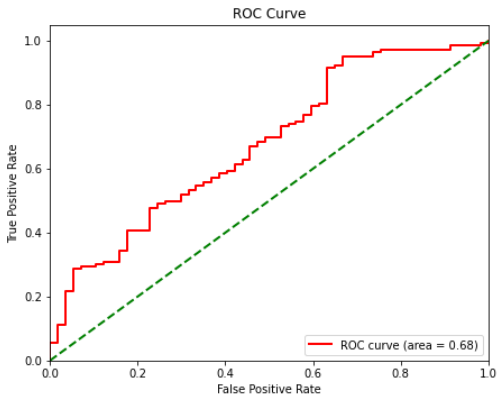



图 11　LR 算法 ROC 曲线

由上图可知，曲线下方的面积（AUC）为 0.68。

**（3）随机森林**

在随机森林算法中，无需进行特征选择，通过 Sklearn 的 RandomForest Classifier 模型对德国信贷数据集进行分类，各个特征的重要性如下图所示：

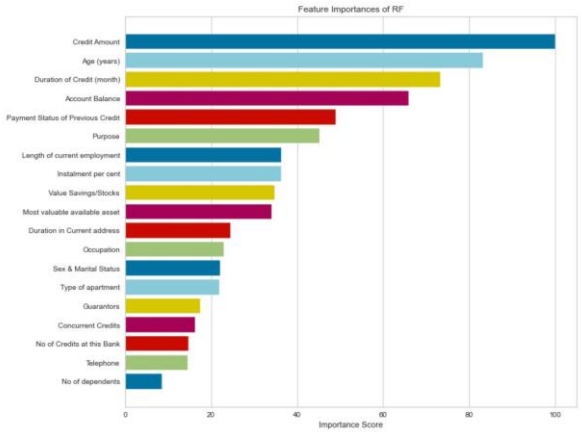



图 12 特征重要性

设置树的数目为 10，对数据进行分类，分类结果如下：

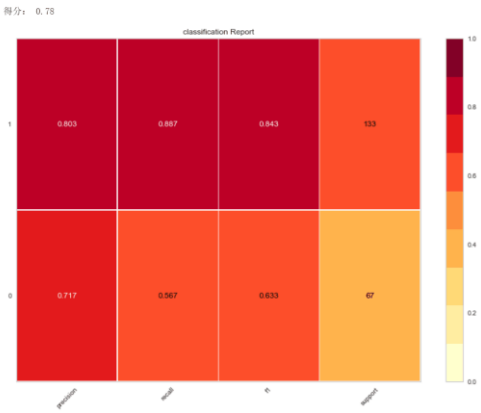



图 13 RF 分类结果

由上图可知，此模型的准确率达到了 78%。在有信用（标签为 1）客户的分类中，查准率(P)达到了 80%，查全率（R）达到了 89%；在无信用（标签为 0）客户的分类中，查准率(P)达到了 72%，查全率（R）达到了 57%。对比来看，此模型对于有信用（标签为 1）样本预测表现更好。

# 六、结果分析和得出结论

本文以德国信贷数据集为实验数据，采用 SVM、逻辑回归和随机森林模型对数据集进行分类。使用 pca 选择出三个对信用度影响最大的特征，然后构建 SVM、逻辑回归模型进行分类。其中，SVM 的分类准确率为 72%，逻辑回归的分类准确率是 75%，准确率较低，分类结果不是很好。由于随机森林可以自动选取特征，所以我使用所有的特征构建随机森林模型，随机森林的分类准确率为 78%，准确率较高于 SVM 和逻辑回归模型的准确率。随机森林是 bagging 下的集成算法，因此有较好的分类效果。

# 七、实践总结

本次实践通过逻辑回归、SVM 以及随机森林算法实现了对于德国银行信用贷款人员的有无信用的分类。由于信贷数据很多为离散型数据，所以考虑运用随机森林算法进行分类。又由于 PCA 主成分分析中提取的均为信贷数据中的连续型变量，所以还可以用这些变量进行逻辑回归及 SVM 进行分类。在实现分类算法的过程中，我加深了对逻辑回归、SVM 以及随机森林以及 PCA 主成分分析理论知识的理解，同时在不断的尝试与纠错中，增加了许多编程经验。同时，我还利用了关联性分析中的 Apriori 算法，对于男性婚姻状况影响因素进行数据挖掘。

但是这次针对于德国银行信用贷款人员的有无信用的分类中，所有算法的准确率都在 80%一下。鉴于此，我应在以后的工作中找到问题原因，修改算法或调整参数从而对于准确率进行提升。