杭州电子科技大学

《数据挖掘课程设计》

课 程 设 计 报 告

专　　业	信息与计算科学
班　　级	18070111
学生姓名	王耀亮
学　　号	18072127
指导教师	邵新平
实验地点	6 教 402
完成日期	2022 年 7月 4日

目录

1. 背景 **3**

1.1 数据挖掘背景 3

1.2 数据集介绍 3

2. 数据预处理 **3**

2.1 缺失值处理 4

3. 多元线性回归 **4**

3.1 模型原理 4

3.2 相关系数矩阵 5

3.3 逻辑回归 6

3.3.1模型原理 6

3.3.2模型结果 6

3.4 随机森林 7

3.4.1模型原理 7

3.4.2模型结果 8

1. 背景

1.1 数据挖掘背景

银行的主要收入来自贷款。但它往往与风险有关。借款人可能会拖欠贷款。为了缓解这个问题，银行决定使用机器学习来克服这个问题。他们收集了关于贷款借款人的过去数据，希望您开发一个强大的分类算法模型来分类是否有任何新的借款人可能违约。该数据集庞大，由多个决定性因素组成，如借款者的收入、性别、贷款用途等。

1.2 数据集介绍

该数据集中有 34 个特征向量，DataFrame 并没有显示所有特征向量。为了解决这个问题，我们可以使用 pandas 设置设置要显示的列数，如下单元格所示：

```
pd.set_option("display.max_columns", dataset.shape[-1])
dataset.head(5)
```

	ID	year	loan_limit	Gender	approv_in_adv	loan_type	loan_purpose	Credit_Worthiness	open_credit	business_or_commercial	...	credit_type	Credit_Score	co-applicant_credit_type	age	submission_of_application	LTV	Region	Security_Type	Status	dtir1
0	24890	2019	cf	Sex Not Available	nopre	type1	p1	l1	nopc	nob/c	...	EXP	758	CIB	25-34	to_inst	98.728814	south	direct	1	45.0
1	24891	2019	cf	Male	nopre	type2	p1	l1	nopc	b/c	...	EQUI	552	EXP	55-64	to_inst	NaN	North	direct	1	NaN
2	24892	2019	cf	Male	pre	type1	p1	l1	nopc	nob/c	...	EXP	834	CIB	35-44	to_inst	80.019685	south	direct	0	46.0
3	24893	2019	cf	Male	nopre	type1	p4	l1	nopc	nob/c	...	EXP	587	CIB	45-54	not_inst	69.376900	North	direct	0	42.0
4	24894	2019	cf	Joint	pre	type1	p1	l1	nopc	nob/c	...	CRIF	602	EXP	25-34	not_inst	91.886544	North	direct	0	39.0

5 rows × 34 columns

图 1　数据集介绍

2. 数据预处理

数据预处理是准备原始数据并使其适合机器学习模型的过程。这是创建机器学习模型的第一步，也是至关重要的一步。在创建机器学习项目时，我们并不总是会遇到干净且格式化的数据。在对数据进行任何操作时，必须对其进行清理并以格式化的方式放置。为此，我们使用数据预处理任务.

真实世界的数据通常包含噪声、缺失值，并且可能是无法直接用于机器学习模型的不可用格式。数据预处理是清理数据并使其适合机器学习模型所需的任务，这也提高了机器学习模型的准确性和效率。数据预处理包括以下步骤：

- 获取数据集
- 导入库
- 导入数据集
- 查找缺失数据
- 编码分类数据

2.1 缺失值处理

缺失数据在真实数据集中并不罕见。事实上，随着数据集大小的增加，至少一个数据点丢失的可能性增加。缺失数据可能以多种方式出现，在这里去，使用如下三种方法对缺失数据进行处理：

- 删除数据
- 编码缺失
- 插补方法

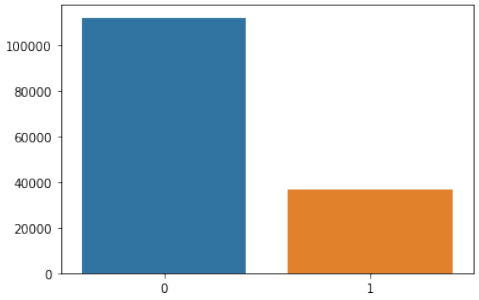

图 2　类 1 和类 0 分布

3. 多元线性回归

线性回归 (linear regression) 是试图学得一个线性模型以尽可能准确地预测实值的输出标记。线性回归一般被用在回归问题上，但是对线性回归做出简单的变换，即可用于进行二分类问题。此时的线性回归模型也就变成了我们熟悉的逻辑斯蒂回归 (logistic regression)。而一个多分类问题又可以分解成多个二分类问题，于是我们同样可以将逻辑斯蒂回归模型用于解决多分类问题上。以下将顺序介绍线性回归，逻辑斯蒂回归，以及多分类问题如何分解成多个二分类问题。

3.1 模型原理

首先我们定义一下字符。一个样本我们用 x 来表示，数据集中第 k 个样本则为 x_k 。一个样本中存在 d 个特征值，我们用一个列向量来表示一个样本，即 $x_k = \left(x_k^1; x_k^2; x_k^3; \ldots; x_k^d\right)$ 。w 是待学习的权重，因为每个样本中有 d 个特征，因此 w 是一个 d 维的列向量，记为 $w = \left(w^1; w^2; w^3, \ldots, w^d\right)$ 。在 W 和 b 确定的的情况下，模型就确定了下来 $\hat{y} = w^T x + b$ 。我们用均方误差来衡量模型的性能，则 $L(w,b) = \sum_{i=0}^{m} \left(y_i - \hat{y}_i^2\right) = \sum_{i=0}^{m} \left(y_i - w^T x_i - b\right)^2$ 。其中 m 为样本总数。简单梳理一下，在训练过程 (在这里我们把确定 W 和 b 的过程称为训练过程) 中我们的目标是找出令 $L(w,b) = \sum_{i=0}^{m} \left(y_i - w^T x_i - b\right)^2$ 最小的 w 和 b 。

而在测试过程（在这里我们把根据求得的 w 和 b 计算 y 的过程称为测试过程) 中，我们输入 x 的输出为 $\hat{y} = w^T x + b$ 。目标明确后，我们如何计算得到 w 和 b 使得 L(w, b) 最小. 因为这是一个线性模型，而且求解的目标是均方误差最小化，因此我们可以用最小二乘法来求解 W 和 b。p.s. 为了方便讨论，我们在下面的推导中将 x 的维度设为 1 ，即 d =1 首先将损失函数 L 分别对 w 和 b 求偏导，得:

$$\frac{\partial L(w,b)}{\partial b} = -2\sum_{i=1}^{m}(y_i - wx_i - b) \tag{1}$$

$$\frac{\partial L(w,b)}{\partial w} = -2\sum_{i=1}^{m}\left(x_i y_i - wx_i^2 - bx_i\right) \tag{2}$$

令 $\frac{\partial L(w,b)}{\partial b} = 0 b = \frac{1}{m}\sum_{i=1}^{m}(y_i - wx_i)$ 又因为 $m\bar{y} = \sum_{i=1}^{m} y_i, m\bar{x} = \sum_{i=1}^{m} x_i$ 所以

$$w = \frac{\sum_{i=1}^{m} x_i y_i - \bar{y}\sum_{i=1}^{m} x_i}{\sum_{i=1}^{m} x_i^2 - \bar{x}\sum_{i=1}^{m} x_i} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \tag{3}$$

3.2 相关系数矩阵

相关性热图测量数据集列之间的零相关性。它显示了一个特征的存在或不存在对另一个特征的影响有多大。

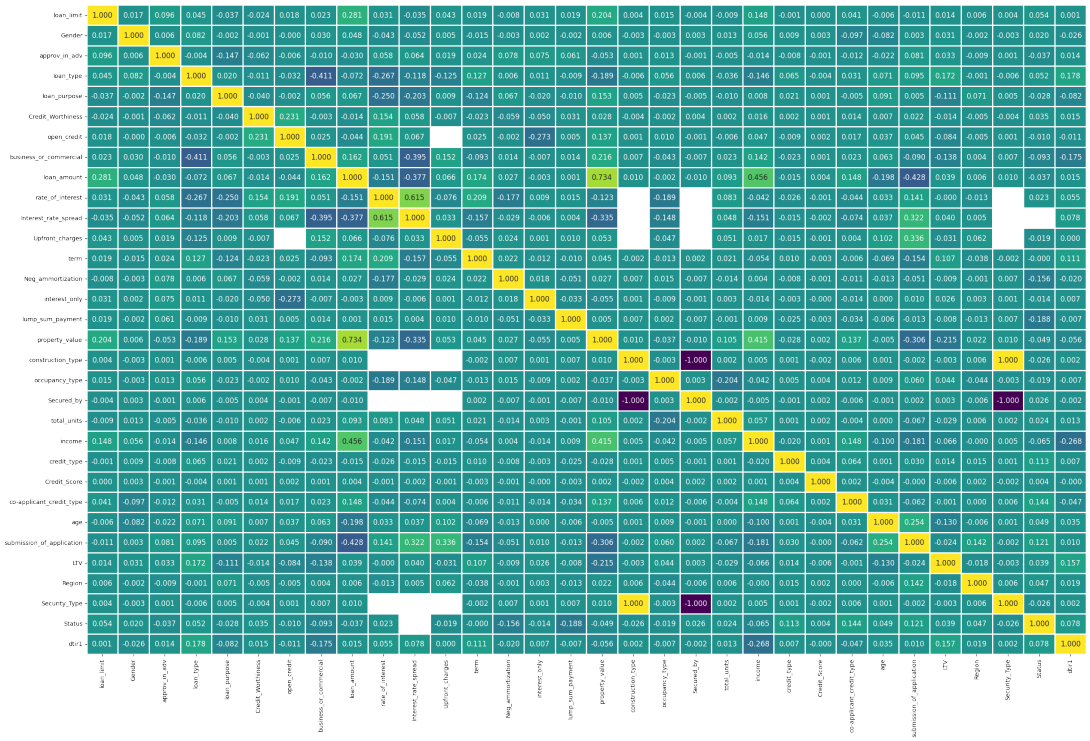

图 3　相关系数矩阵

loan_amount 与 property_value 为变量的散点图如下：

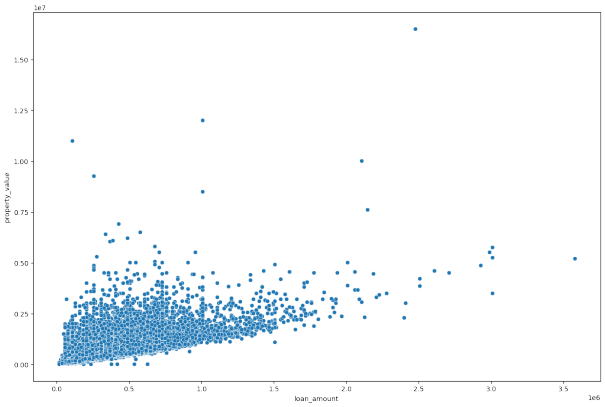

图 4　loan_amount 与 property_value 散点图

从图中我们可以看到两个特征之间存在线性关系，但线性回归不能很好地拟合数据。虽然线性回归不能很好地拟合数据，但可以尝试使用线性回归，

3.3 逻辑回归

3.3.1 模型原理

二分类问题: 对样本点 x，其标签 $y \in \{0,1\}$，想要知道的是概率

$$\Pr(y=1 \mid x) \tag{4}$$

对于二分类问题，显然可以建模假设 $y \mid x$ 服从参数为 p 的 Bernoulli 分布，因此只需要估计 $p=E(y \mid x)$

Bernoulli 分布是二项分布 n=1 的特殊情形，显然也属于指数分布族。使用 GLM 建模：选择 Bernoulli 分布 + 正规连接，得到

$$p=\frac{1}{1+e^{-\beta^T x}}=s\left(\beta^T x\right) \tag{5}$$

其中 $s(\cdot)$ 称为 Sigmoid 函数

$$s(z)=\frac{1}{1+e^{-z}} \tag{6}$$

3.3.2 模型结果

使用逻辑回归得到的结果如下：

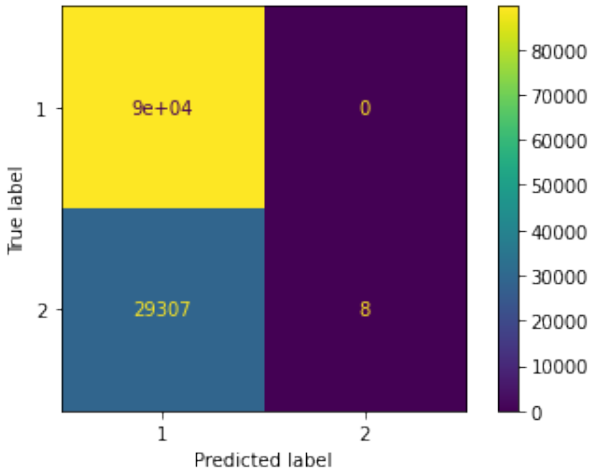

图 5　混淆矩阵

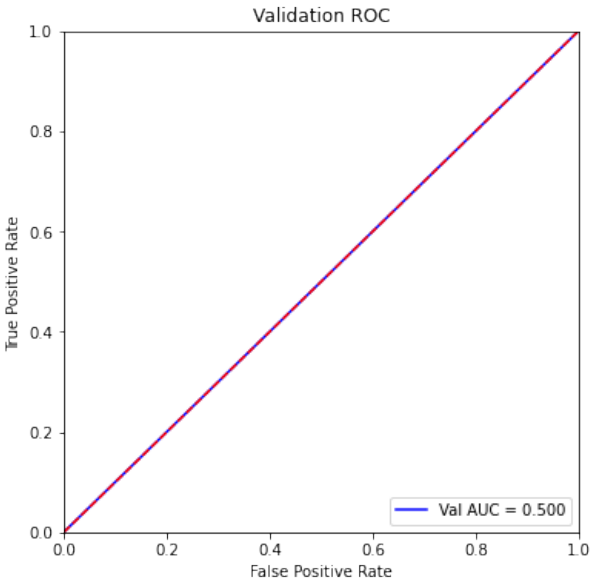

图 6　AUC 曲线

可以看到，逻辑回归的 AUC 为 0.5，效果很差。

3.4 随机森林

3.4.1 模型原理

生成单棵决策树：（1）训练总样本的个数为 N，则单棵决策树从 N 个训练集中有放回的随机抽取 n 个作为此单颗树的训练样本。

（2）令训练样例的输入特征的个数为 M，m 远远小于 M，则我们在每颗决策树的每个节点上进行分裂时，从 M 个输入特征里随机选择 m 个输入特征，然后从这 m 个输入特征里选择一个最好的进行分裂。m 在构建决策树的过程中不会改变。这里注意，要为每个节点随机选出 m 个特征，然后选择最好的那个特征来分裂。

（3）每棵树都一直这样分裂下去，直到该节点的所有训练样例都属于同一类。不需要剪枝。由于之前的两个随机采样的过程保证了随机性，所以就算不剪枝，也不会出现 over-fitting。

3.4.2 模型结果

使用随机森林得到的结果如下：

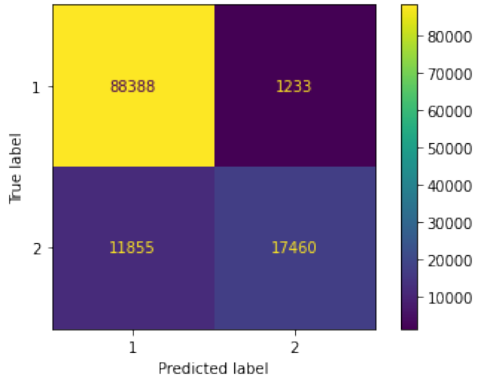

图 7　混淆矩阵

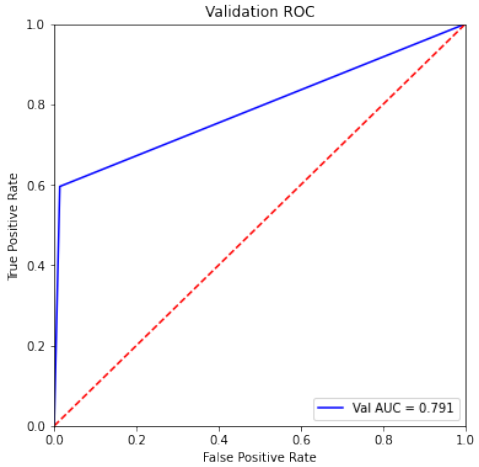

图 8　AUC 曲线

可以看到，随机森林的 AUC 为 0.791，效果比逻辑回归更好。