大数据与金融科技———数据分析项目

黄一鸣、潘演乐、周婧

广东金融学院金融硕士

2021/4/13

目录

注：本数据分析项目采用 R 语言

数据描述及数据展示

原始数据来自各大 P2P 贷款平台：

- 68 个变量
- 100000 个观测值

数据描述及数据展示

部分数据展示：

	userId	nickName	realName	gender	birthDay	
1	6192030	YanF_15191889841.yx	\u4e25*	\u7537	1981/6/14	32
2	9035861	ZhengQ_15413629497.yx	\u90d1*	\u5973	1993/6/4	42
3	1511029	all20130708.rrd	\u9f99**	\u7537	1987/10/24	42
4	6516550	LRF_20768902017051O.xd	\u674e**	\u5973	1990/9/5	41
5	6785812	LMY_22405962017071l.xd	\u5415**	\u7537	1975/4/14	33
6	672505	huangdeyi.rrd	\u9ec4**	\u7537	1986/6/3	43

	marriage	graduation
1	MARRIED	\u7814\u7a76\u751f\u6216\u4ee5\u4e0a
2	UNMARRIED	\u5927\u4e13
3	MARRIED	\u5927\u4e13
4	UNMARRIED	\u5927\u4e13
5	DIVORCED	\u7814\u7a76\u751f\u6216\u4ee5\u4e0a
6	MARRIED	\u5927\u4e13

	homeTown

本小组研究的主要问题

根据数据集中大量的变量，本小组选择研究三个问题：

1. 借款人特征对审批额的影响；
2. 借款人特征与审批状况；
3. 研究借款人的借款理由（尚未完成）。

本小组研究的主要问题

因为原始数据集的变量过多，本文选择了以下的重点变量：

- age
- marriage
- gender
- availableCredits
- sumCreditPoint
- region
- hasHouse, houseLoan, hasCar, carLoan
- officeDomain

数据清洗与数据预处理

数据清洗可以分为选择子集、列重命名、删除重复值、缺失值处理、一致化处理、数据排列、异常值处理；

- 选择子集：选择子集可以通过隐藏不需要的列数据来使整个数据集更加明了；
- 列重命名：将已存在的列明进行重命名，以方便理解；
- 删除重复值；
- 缺失值处理：人工补全、删除、平均值替代、统计模型计算替代值；
- 一致化处理：让数据整整齐齐、方便后续操作；
- 数据排序：数据排序可以让整组数据看起来更有序；
- 异常值处理：处理数值过大或过小的观测值；

本项目的数据清洗代码文件为：

dataclean.R

数据清洗与数据预处理

数据清洗部分代码展示：

```
loandata$gender[loandata$gender == " 男"] <- "male"
loandata$gender[loandata$gender == " 女"] <- "female"
loandata$gender <- factor(loandata$gender,
                          ordered = FALSE,
                          labels = c("male","female"))

loandata$officeDomain[loandata$officeDomain
                      == ", IT"] <- "IT"
loandata$officeDomain[loandata$officeDomain
                      == ", \u623f\u5730\u4ea7\u4e1a"]
                      <- " 房地产业"
```

数据清洗与数据预处理

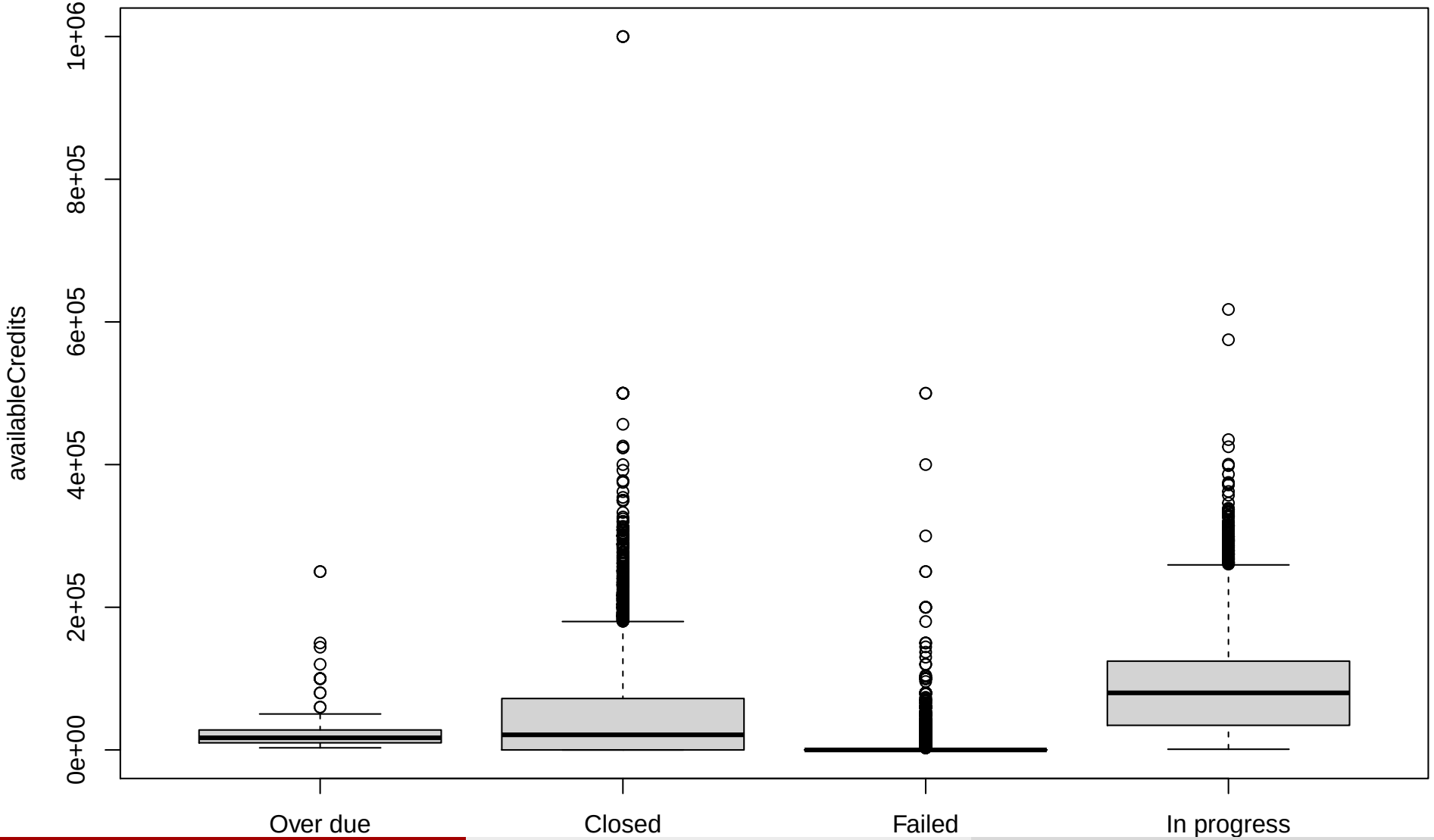
数据预处理：

本项目的数据预处理主要是删除异常值：

- 使用箱线图分析异常值情况；
- 删除异常值

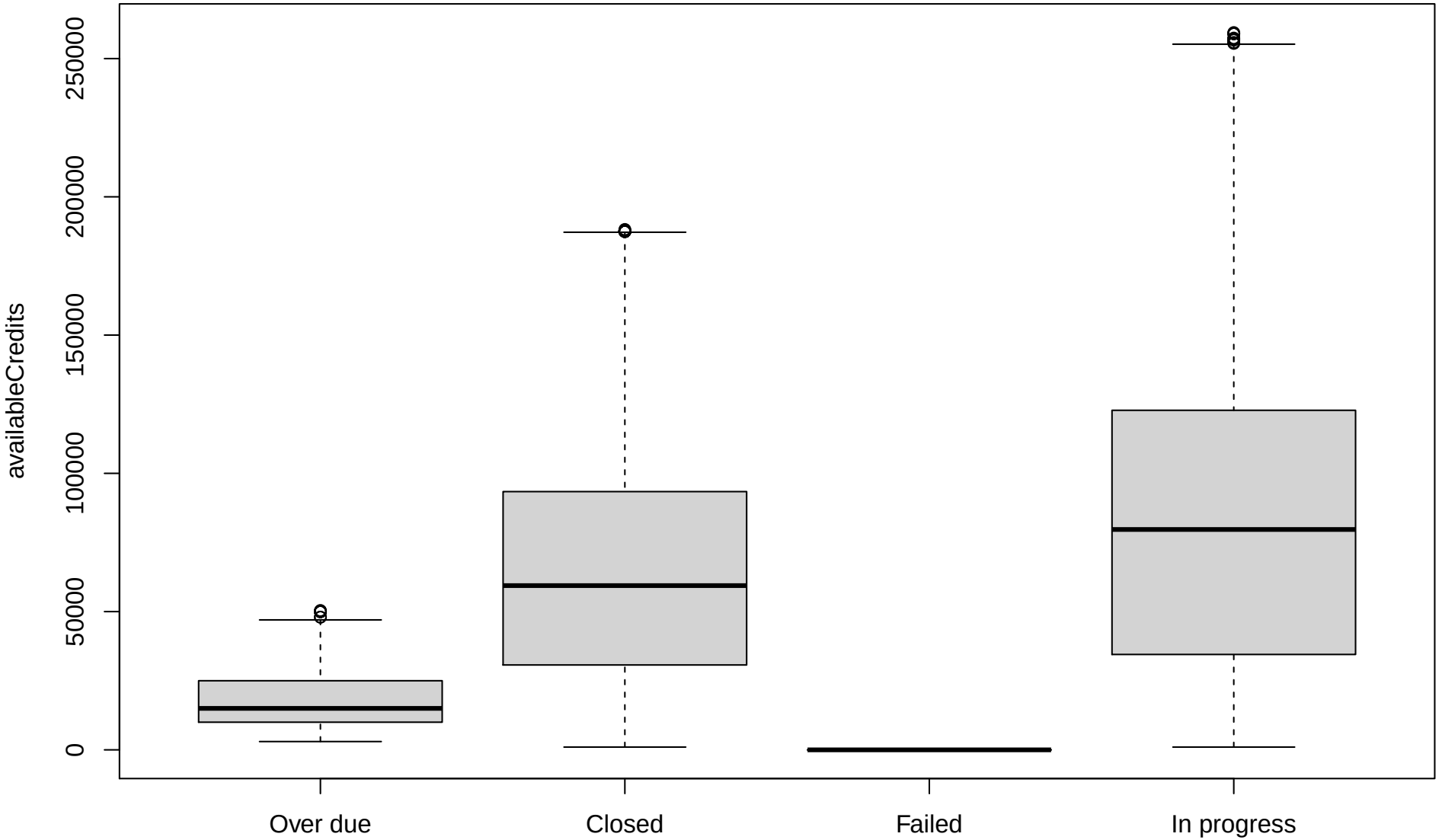
数据清洗与数据预处理

异常值的情况：



数据清洗与数据预处理

清理了异常值后的情况：



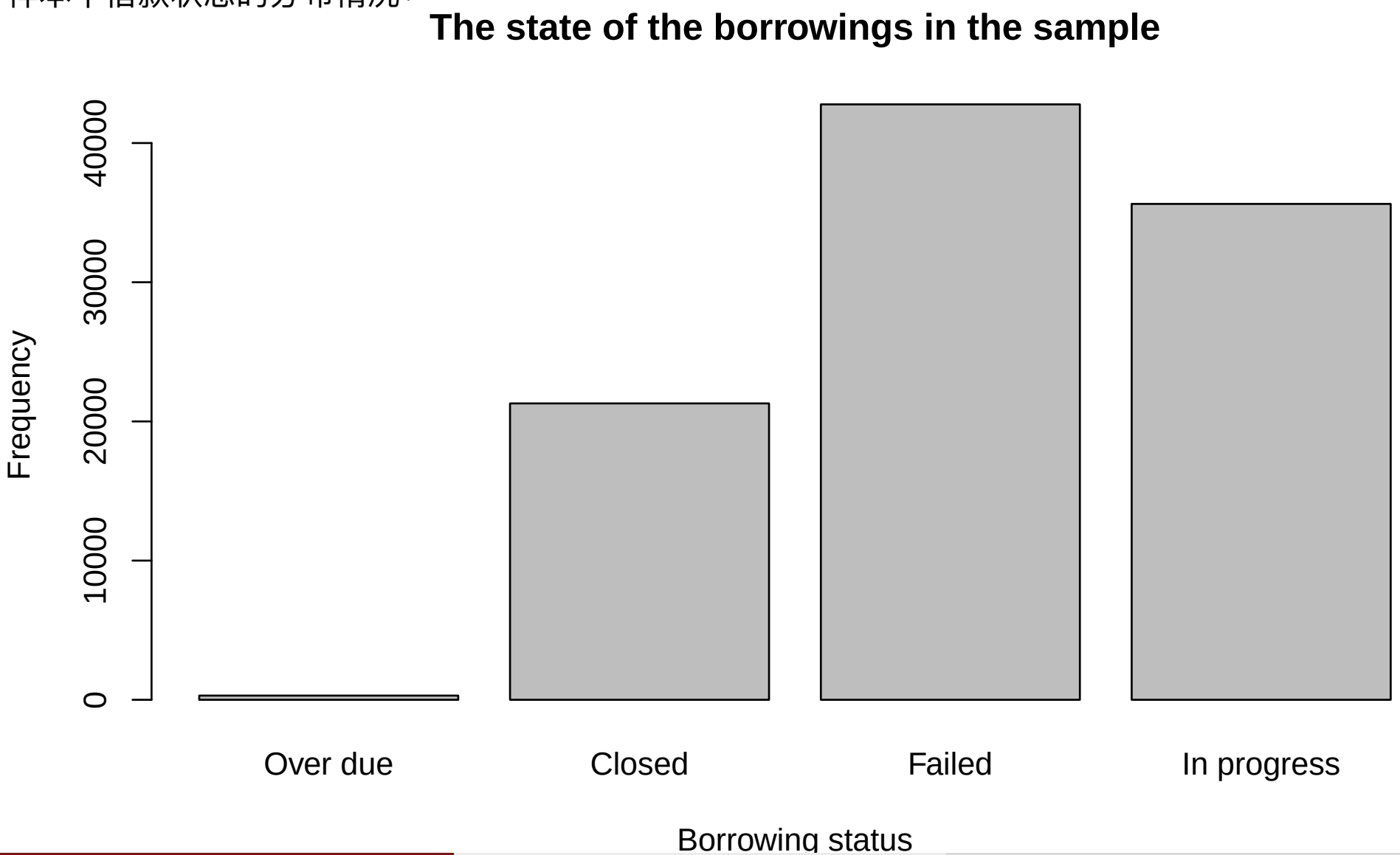
描述性统计

样本中借款状态的分布情况（百分比）：

Over due	Closed	Failed	In progress
0.301	21.293	42.779	35.627

描述性统计

样本中借款状态的分布情况：



描述性统计

审批额度的的描述性统计：

```
$`Over due`
   vars   n      mean        sd median  trimmed     mad  min   ma
X1    1 286 18419.23 10868.07  15000 17184.35 8895.6 3000 5040
   kurtosis     se
X1     0.67 642.64

$Closed
   vars      n    mean        sd median  trimmed       mad  min
X1    1 12381 64277.2 39790.04  59400 61127.77 44329.74 1000 1
   kurtosis    se
X1    -0.14 357.6
```

描述性统计

审批额度的的描述性统计：

```
$Failed
   vars     n mean sd median trimmed mad min max range skew ku
X1    1 42779    0  0      0       0   0   0   0     0  NaN

$`In progress`
   vars     n     mean       sd median  trimmed      mad  min
X1    1 35489 86451.01 54039.07  79700 82623.39 66272.22 1000
   skew kurtosis     se
X1 0.49    -0.61 286.85
```

回归分析与分类预测

回归分析：

本项目使用 age、region、hasHouse、houseLoan、hasCar、carLoan、gender、officeDomain、marriage 对 availableCredits 进行回归，因为受页面大小限制，不在 PPT 上展示。

回归分析与分类预测

分类预测：

本项目使用 officeDomain、region、归一化后的 sumCreditPoint 和 age 对 status 进行分类预测，使用的算法是随机森林。

由于样本中的 status 包含四种借款类型：借款完成、逾期、借款进行时、借款失败，所以用于随机森林中进行分类训练的只有借款完成和还款失败两种类型，使用这个两个类型的 70% 样本进行训练，剩余样本进行预测。接着使用训练好的模型，用借款进行时的数据进行预测，尝试预测这些借款人未来是否会逾期。

结果表明：

1. 训练模型的准确度达 95%；
2. 使用借款进行时的数据进行分类预测发现，这些样本未来都不会逾期

回归分析与分类预测

分类预测存在的问题：

逾期类型的样本量相对借款完成类型而言，样本量太少，模型可能无法捕捉到逾期的典型特征。

- 逾期类型的样本仅占 0.31%
- 借款完成类型的样本占 21.29%

谢谢

展示完毕

敬请各位老师同学批评指正

汇报人：黄一鸣