# 实验四报告

## 任务一

编写 MapReduce 程序，统计每个⼯作领域 industry 的⽹贷记录的数量，并按数量从⼤到⼩进⾏排序

第一个job先进行统计，map发送key（industry）、value（one）；reduce进行累加

第二个job先交换key和value，然后重写comparator，接受intwriable参数，然后比较数值大小，然后reduce交换key和value的位置输出



## 任务二

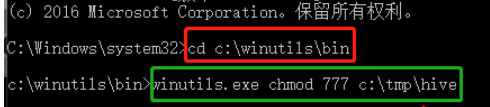
编写 Spark 程序，统计⽹络信⽤贷产品记录数据中所有⽤户的贷款⾦额 total\_loan 的分布情况。

以 1000 元为区间进⾏输出。

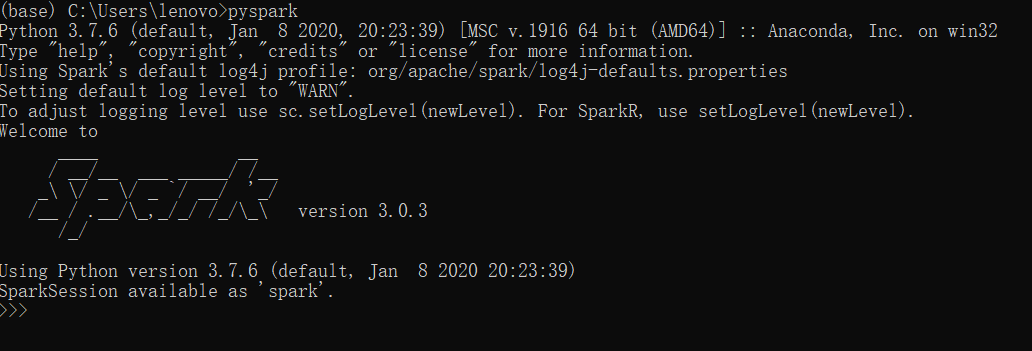
环境配置

安装sparl、配置环境变量

下载winutils.exe修改权限

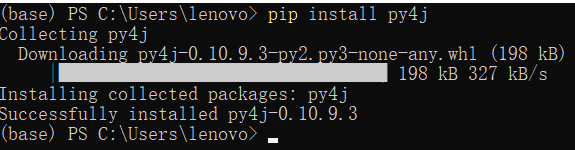


检验pyspark安装成功



将pyspark的包复制到anacoda里的lib文件夹下

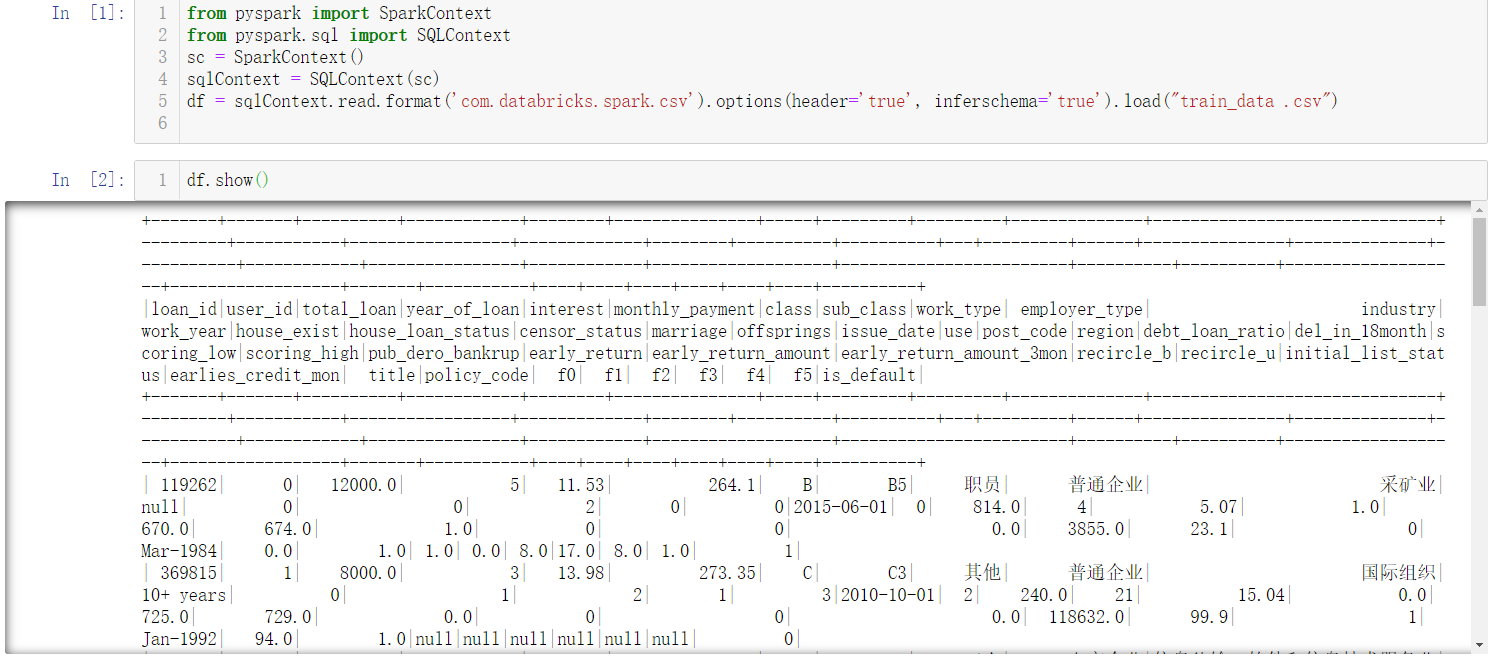
然后安装py4j包，环境配置成功

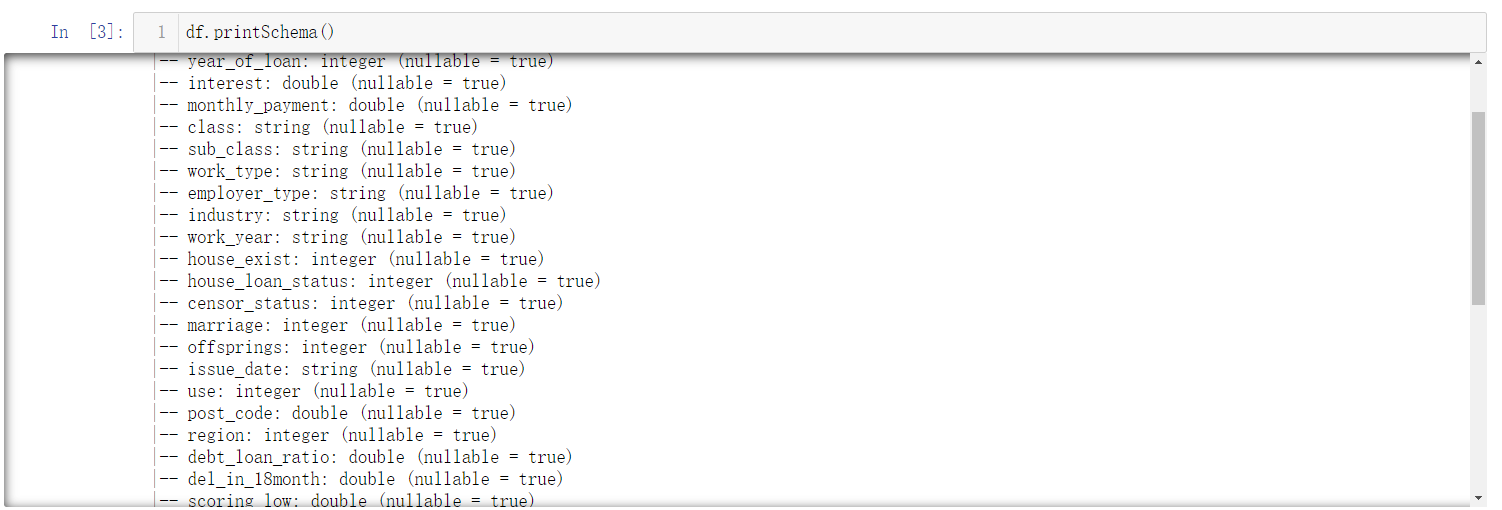


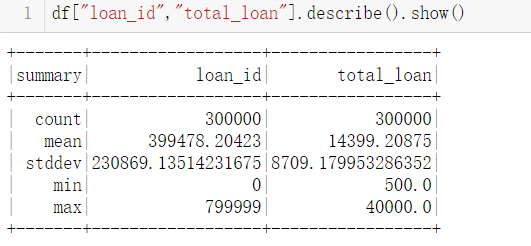
在jupter里完成任务

先进行pyspark操作的熟悉

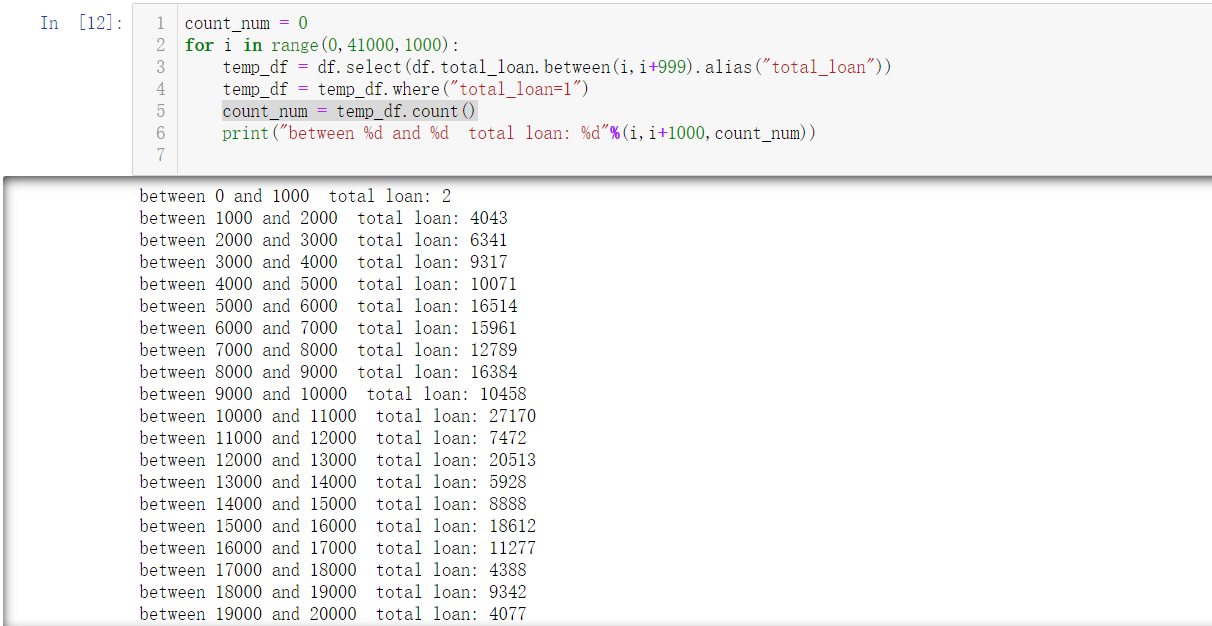
读取、展示数据







使用between函数统计数目，循环输出结果

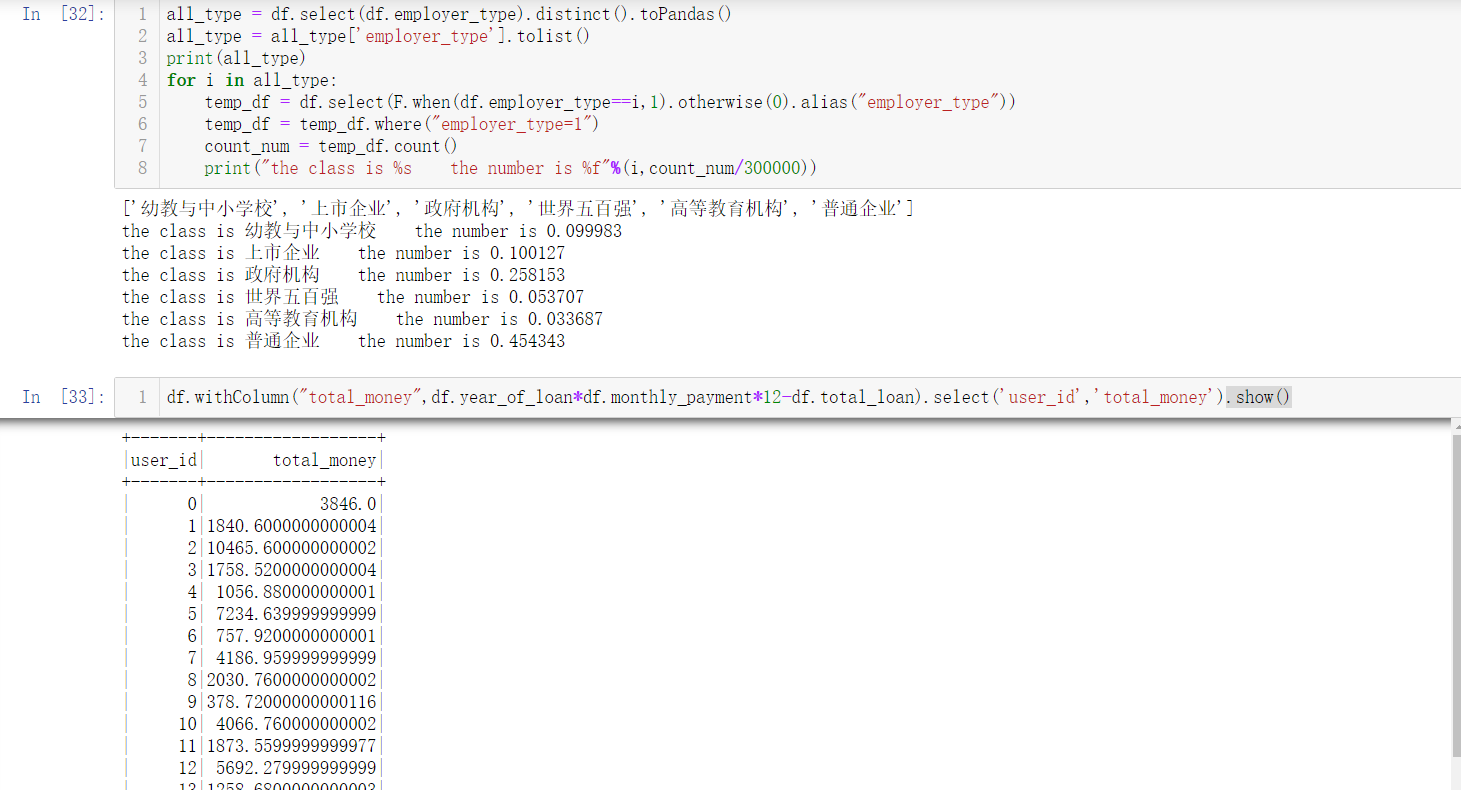


## 任务三

基于 Hive 或者 Spark SQL 对⽹络信⽤贷产品记录数据进⾏如下统计：

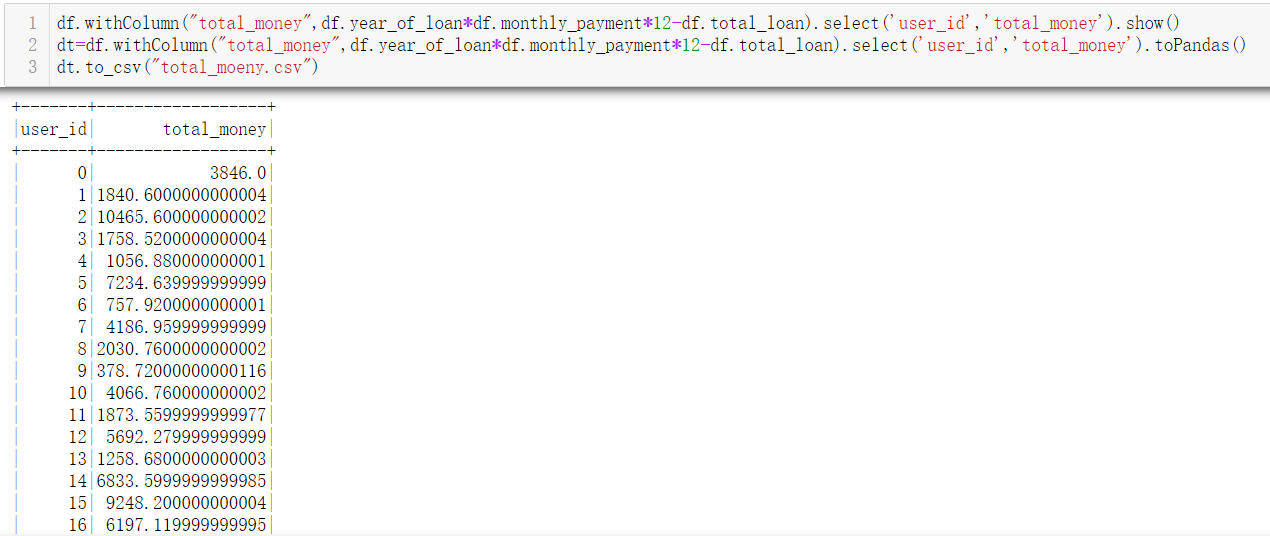
1. 统计所有⽤户所在公司类型 employer\_type 的数量分布占⽐情况。

使用sql库中的When完成任务，将符合条件的值置为1，然后统计数目



2.统计每个⽤户最终须缴纳的利息⾦额

直接通过context类型进行计算



3.统计⼯作年限 work\_year 超过 5 年的⽤户的房贷情况 censor\_status 的数量分布占⽐情况。

使用filter，过滤满足条件的数据



## 任务四

根据给定的数据集，基于 Spark MLlib 或者Spark ML编写程序预测有可能违约的借贷⼈，并评估实验结果的准确率。

数据集中 is\_default 反映借贷⼈是否违约。

可按⽐例（如 8:2）把数据集拆分成训练集和测试集，训练模型，并评估模型的性能，评估指标可以是 accuracy、f1 score 等等

先进行特征分析（分析需要进行的处理）和特征选取：

loan\_id ⽹络贷款记录唯⼀标识 不作为特征

user\_id ⽤户唯⼀标识 不作为特征

total\_loan ⽹络贷款⾦额 作为特征，标准化

year\_of\_loan ⽹络贷款期限（year） 作为特征，标准化

interest ⽹络贷款利率 作为特征，标准化

monthly\_payment 分期付款⾦额 作为特征，标准化

class ⽹络贷款等级 独热编码 作为特征

sub\_class ⽹络贷款等级之⼦级 独热编码 作为特征

work\_type ⼯作类型（公务员、企业⽩领、创业…） 独热编码 作为特征

employment\_type 所在公司类型（世界五百强、国有企业、普通企业…）独热编码 作为特征

industry ⼯作领域（传统⼯业、商业、互联⽹、⾦融…） 独热编码 作为特征

work\_year 就业年限（年） 独热编码 作为特征

house\_exist 是否有房 独热编码 作为特征

house\_loan\_status 房屋贷款状况（⽆房贷、正在还房贷、已经还完房贷） 作为特征

censor\_status 验证状态 作为特征

marriage 婚姻状态（未婚、已婚、离异、丧偶） 作为特征

offsprings ⼦⼥状态(⽆⼦⼥、学前、⼩学、中学、⼤学、⼯作) 作为特征

issue\_date ⽹络贷款发放的⽉份

数据说明

use 贷款⽤途 作为特征

post\_code 借款⼈邮政编码的前3位 不选取

region 地区编码 作为特征

debt\_loan\_ratio 债务收⼊⽐ 作为特征

del\_in\_18month 借款⼈过去18个⽉信⽤档案中逾期60天内的违约事件数 作为特征

scoring\_low 借款⼈在信⽤评分系统所属的下限范围 标准化

scoring\_high 借款⼈在信⽤评分系统所属的上限范围 标准化

pub\_dero\_bankrup 公开记录清除的数量 不作为特征

early\_return 提前还款次数 作为特征

early\_return\_amount 提前还款累积⾦额 不作为特征

early\_return\_amount\_3mon 近3个⽉内提前还款⾦额 不作为特征

recircle\_bal 信贷周转余额合计 不作为特征

recircle\_util 循环额度利⽤率，或借款⼈使⽤的相对于所有可⽤循环信贷的信贷⾦

额 不作为特征

initial\_list\_status ⽹络贷款的初始列表状态 不作为特征

earlies\_credit\_line ⽹络贷款信⽤额度开⽴的⽉份 不作为特征

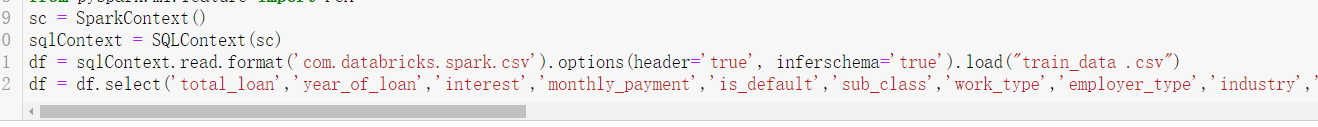
title 借款⼈提供的⽹络贷款名称 不作为特征

policy\_code 公开策略=1不公开策略=2 不作为特征

f系列匿名特征 匿名特征f0-f5，为⼀些⽹络贷款⼈⾏为计数特征的处理 不作为特征

PAC

筛选特征

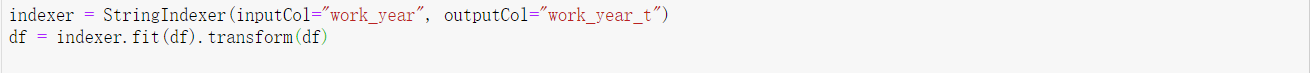


填充缺失值

先用均值填空，不能用均值填充的特征用0来填充



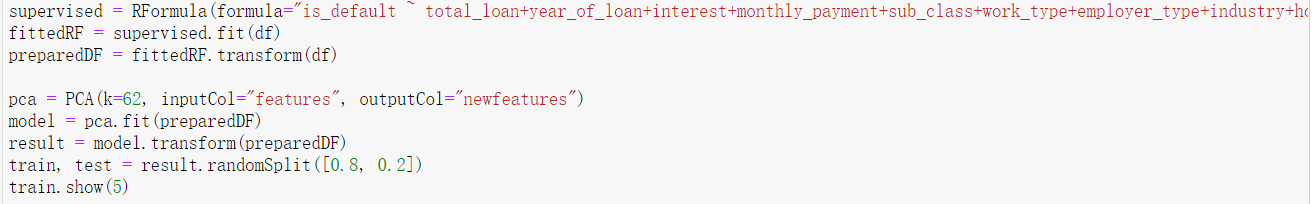
进行字符编码



通过Rfomula进行特征组合以及字符特征的字符编码

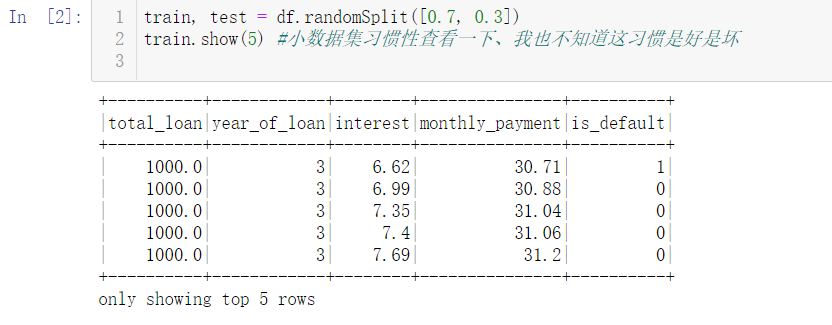
然后进行PAC主成分分析

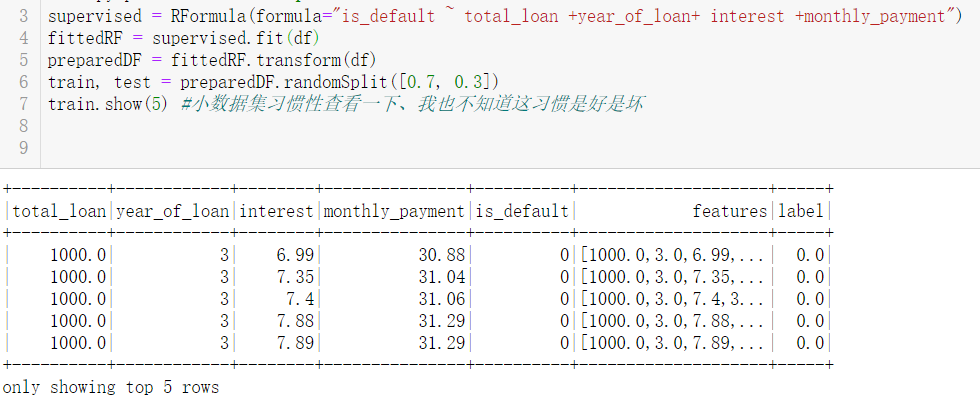
然后分割训练集和预测集



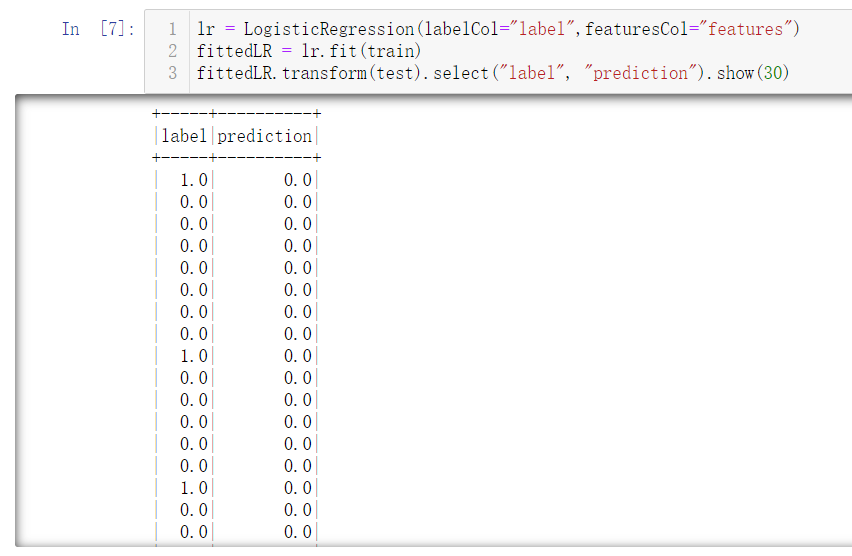
下面对比经过特征工程的结果和没经过特征工程的结果：

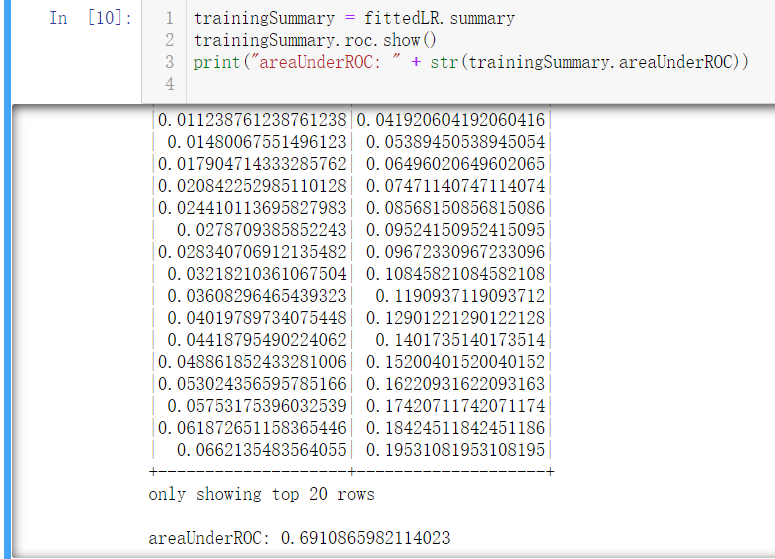
没有特征工程，使用逻辑回归：

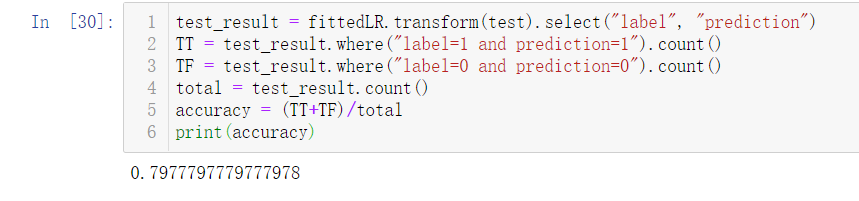




预测结果展示：

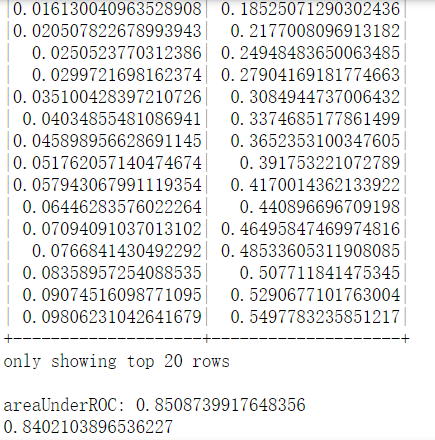






特征工程后的结果：

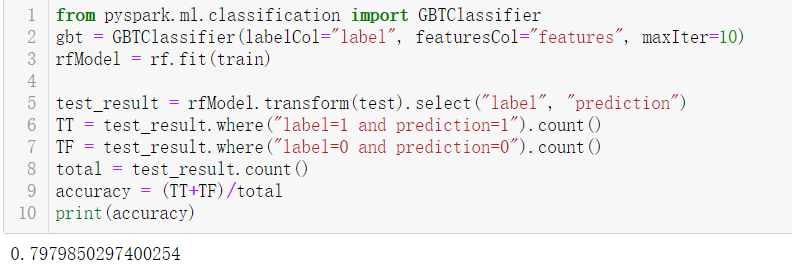
使用逻辑回归：



使用随机森林：



使用GBtree



准确率在降低，本问题中，越复杂的模型效果越差，说明遇到了过拟合的问题。