数据说明：simudata.csv

此数据为某公司用户在第三方交易平台的用户的日常刷卡行为数据，数据提供了一批用户行为记录，包括：年龄、银行卡数、借贷比率、交易笔数、所有行为均值、 所有行为最大值等。详细的变量含义及说明如下：

变量类型	变量名	变量含义	详细说明	取值范围	备注
因变量	black	是否违约	定性变量 （2 个水平）	1 表示违约 0 表示未违约	违约占比 33.37%
自变量	age	年龄	连续变量	14-47	只取整数
	cardnum	银行卡数	连续变量	0-16	只取整数
	creded	借贷比率	连续变量	0-0.82	用户所有行为中，采用贷记卡（或称为信用卡）交易的次数占所有采用贷记卡或者借记卡（或称为储蓄卡）交易次数的比率。
	billnum	交易笔数	连续变量	0-427	用户自从第一笔交易记录开始，被记录的交易总笔数。
	meanpay	所有行为均值	连续变量	0-1346594	用户被记录的所有交易行为的平均金额。
	maxpay	所有行为最大值	连续变量	9-6470316	用户自从第一笔交易开始，所有交易行为的金额的最大值，可以度量用户的极端行为情况。
	xiaofeiF	消费类 F	连续变量	0-19	详见下述说明
	jinkaF	金卡类 F	连续变量	0-44	
	youxiM	游戏类 M	连续变量	0-1939	
	debitM	借记类 M	连续变量	26-1436086	
	debitF	借记类 F	连续变量	0-202	
	gongjiaoR	公缴类 R	连续变量	206.3-365	
	gongjiaoF	公缴类 F	连续变量	0-7	
	gongjiaoM	公缴类 M	连续变量	0.02-55745.07	
	zhongxingR	中型银行 R	连续变量	0.17-365	
	zhongxingF	中型银行 F	连续变量	0-53	
	zhongxingM	中型银行 M	连续变量	3-1017770	
	sidaR	四大行 R	连续变量	0.16-365	
	sidaF	四大行 F	连续变量	0-144	
	sidaM	四大行 M	连续变量	10-1299826	
	zhuanzhangR	转账类 R	连续变量	0.28-704.05	
	zhuanzhangF	转账类 F	连续变量	0-124	
	zhuanzhangM	转账类 M	连续变量	38.2-1650050.4	
	xindaiR	信贷类 R	连续变量	0.02-365	
	xindaiF	信贷类 F	连续变量	0-20	
	xindaiS	信贷类 S	连续变量	8.1-485178.3	

在营销领域，RFM 模型是用来衡量客户的价值和客户的创利能力的重要工具和手段。这个模型通过一个客户的近期购买行为、购买的总体频率以及花了多少钱三项指标来描述该客户的综合价值，具体如下：

（1）**R（Recency）**，最近一次消费，指上一次购买的时间到现在的距离。理论上，上一次消费时间越近的用户应该是相对而言活跃的用户。

（2）**F（Frequency）**，消费频率，即用户在限定的期间内产生购买的总次数。产生购买越频繁的用户，忠诚度越高。

（3）**M（Monetary）**，某个用户所有消费金额的平均值。

（4）由于这三个指标并不能够衡量用户产生行为的波动性，所以我们增加一个指标 **S（Standard Deviation）**来衡量用户行为的波动性。

举例说明：对于购买游戏点卡类行为，我们可以定义 R 为用户最近一次购买游戏点卡距离数据提取时间的时间间隔，F 定义为一年内用户购买游戏点卡的次数，M 定义为一年内用户每次购买游戏点卡的平均金额，S 定义为用户每次购买游戏点卡金额的标准差。我们将所有变量记作类别名称加指标简称的形式，例如**游戏 R**，表示**游戏类的 Recency**。

用户行为的分类通过银行卡信息表，以及商户分类信息表，根据业务场景，我们提取了以下类别，每个类别都对应着以上我们已经定义好的 RFMS 四个指标。类别包括：

（1）借记类：刻画用户使用储蓄卡的交易行为。不同的用户习惯不同，采用储蓄卡和信用卡的倾向也可能不同。

（2）消费类：刻画用户的日常消费行为。日常消费行为的金额以及频次不同，用户的还款能力可能不同。

（3）信贷类：刻画用户之前的小额贷款类行为。用户之前如果有其他消费贷款类行为可能已经习惯进行消费贷款，从而可能具备更良好的信用状况。

（4）转账类：刻画用户的转账行为。经常通过熊小贷 APP 转账的用户可能与不转账的用户行为不同。

（5）话费类：刻画用户的话费充值交易行为。话费充值是否规律与充值金额多少都可能意味着用户群体不同。

（6）公缴类：刻画用户交水，电，煤气费等交易行为。公缴费用的多少与是否规律也可能说明用户群体的不同。

（7）游戏类：刻画用户购买游戏点卡的行为。经常玩游戏的用户群体可能与不玩游戏的人不同。

（8，9）四大行卡类以及中型银行卡类：四大行包括中国银行，中国农业银行，中国工商银行，中国建设银行，中型银行包括招商银行，浦发银行，兴业银行，平安银行等。这个指标的设定有以下两方面原因：a. 不同公司的工资卡不同，小型创业公司一般采用中型银行的银行卡；b. 四大行的信用卡发放较为保守，所以能够申请到四大行信用卡的人可能和采用其他银行信用卡的用户群体不同。

（10）白金及金卡类：通过卡标首可以对应到银行卡是属于哪家银行的哪种类型的卡，例如招商银行的金葵花卡。我们搜索整理了相应银行的金卡和白金卡卡种名称，并对应到每一个用户。我们初步认为，拥有白金卡和金卡的用户具备更高的还款能力，所以用户群体不同。

分析任务：

1. 读入数据并了解各个自变量的含义；

2. 对变量交易笔数和所有用户行为均值分别绘制违约组和非违约组的对比箱线图，并分析是否违约与这些变量之间的关系，给出解读；

3. 用全部样本数据，以是否违约为因变量建立逻辑回归模型，利用 BIC 准则进行变量筛选，观察最终得到的回归系数并尝试解释对系数进行解释；

4. 使用任务 3 的模型，对全部样本进行预测，计算 AUC 值，并绘制 ROC 曲线，对模型的效果进行评估；

5. 任务 4 中对每个训练样本预测出了非违约的概率，按照非违约率从高到低排序，将全部样本分为 5 组人群：非违约率最高的 20%用户、…、非违约率最低的 20%用户，计算五类人群的平均非违约概率，从高到低排序，绘制柱状图；对结果的商业应用进行解读。