# 中央财经大学

Central University of Finance and Economics


课　　程　大数据计算机基础

论文题目　信用卡用户群信用评级分析

学　　校　中央财经大学

姓　　名　王思雨

学　　号　2017210761

论文日期　2017/12/14

# 第一章数据的说明与变量选取

## 1.1 研究目的与意义

近年来 psp 网贷业务以及信用卡业务慢慢的普及浸入到每个普通人的生活中，对于银行来讲，合理的评价每一个信用卡使用用户的信用等级，更加的划分其用户群体，并合理的评价每一个用户群体相应的信用风险程度，预测该类用户群体为银行所会带来的可能损失，将会更好的指导银行的管理者和销售人员对客户进行理财产品的推荐，信用贷款发放以及信用额度的分配等业务。会帮助银行进行更好的管理服务优化，提升银行规避风险，创造效益的能力。

因此本文基于银行信用卡数据，利用大量用户信息数据分别进行描述式特征数据挖掘分析和推断式数据挖掘建模来划分每一个用户的所属用户群体并对每个用户群体的信用状况，风险状况以及所能带来的损失情况进行分析和描述，得到结论希望能够帮助银行解决现实业务处理中的一些问题。

## 1.2 数据来源及说明

数据来源于大数据计算机基础课上老师所提供的银行信用卡违约数据表，数据一共包含了 65535 条 12 类违约用户的违约记录。并删除了每一条记录的用户敏感信息，仅用于实现统计学建模分析。变量说明如下表 1-1 所示：

表 1-1 信用卡违约记录变量说明表

| 变量类型 | 变量名 | 取值范围 | 详细说明 | 备注 |
|---|---|---|---|---|
| 人口信息学特征 | 性别<br>Sex | 女，男 | 分类变量<br>反映顾客性别状况 | 通过观察性别变量反映性别影响顾客分类和信誉 |
| | 年龄<br>Age | 0~890 岁 | 数值型变量<br>反映顾客年龄状况 | 数据中出现并不合理的数据，这会在预处理中进行数据的处理 |
| | 受教育程度<br>Education | 未上学儿童<br>小学<br>初中<br>高中或中专<br>大学及以上 | 分类变量<br>反映顾客的教育程度 | 受教育程度可能会影响顾客收入和信誉，进而影响顾客分类， |

| | | | | |
|---|---|---|---|---|
| | **职业**<br>Occupation | **在校学生**<br>**农牧渔水利业人员**<br>**商业、服务业人员**<br>**专业技术人员**<br>**办事人员和有关人员**<br>**家务、待业、其他等** | 分类变量<br>反映顾客职业状况 | 不同的客户分类中的从事各种工作的比例也可能不同进而影响各类顾客评分。 |
| **违约特征** | **违约发生时间**<br>pen_time | **20160101~20171213** | 日期型变量 | 可进一步转化为违约时长，用于判断违约程度 |
| | **违约严重程度**<br>degree | **轻度、中度、重度** | 分类变量<br>共有三个违约分类 | 代表着顾客的单词的违约程度 |
| **用户分组** | **顾客代码**<br>code | **顾客 1,2，…,12** | 分类变量 | 标记顾客所属的顾客群体 |

## 1.3 影响个人信用的因素分析

CreditCard 的数据表中提供了关于每一条顾客违约信息的多种分类变量和度量变量。本文想通过先构造对个人信用的评价得分来进一步的确定群体的信用得分。但每个指标具体对个人信用的影响有多大程度，通过查询资料可以得知，如下信息[1]。

现有的个人信用评分主要是由基于反映个人信用信息的微观因素构成，以下是对几种最为常见的影响个人信用的微观因素进行定性分析。也是此次数据挖掘提供的数据集中所包含的：

（1）年龄。达到法定年龄是对借款人的第一要求。国内外研究表明，年纪轻的人收入较低，稳定性差，而且思想更容易冲动，使违约概率增高。随着年龄增大，个人违约风险逐渐减小；而在不具备健全的福利制度的国家则是具有稳定收入的、中年人群违约率最低，年轻人群和退休后的老年人群的违约率较高。

（2）性别。一般认为女性的违约风险比男性的违约风险低，但在大部分国家，女性的收入不及男性，所以是否把性别作为违约因素考虑是有争议的。目前越来越多的国家认识到，在信用评分中考虑性别是带有歧视色彩。国内的许多商业银行也正在取消该项指标。

（3）教育程度。高学历意味着容易就业、高收入以及相对较高的道德水平，因此违约概率也较低。

[1] 参考资料《个人信用评分组合模型研究与应用》 作者：向辉

（4）职业情况。这里包括单位类型、职位以及工作时间长短等因素，它们体现了贷款人的收入稳定性、收入水平以及社会地位。另外，工作的地域流动性也会影响到违约概率。

## 1.4 数据预处理

通过 python 的描述性分析中，我们可以从信用卡违约记录数据中看到，存在这许多数据异常值和缺失值，有的记录是缺失部分值而有的记录缺失的记录较多。有的记录虽然并没有缺失统计量，但是通过尝试可以判断，一些记录是不合常理的，是异常值。例如有些记录的年龄部分超过 100 岁，如果职业是离退休的人员可以理解，但是有很多人是目前还在从事工作，所以在记录表中应该是错误的统计值。所以在原始的信用卡数据基础之上，首先进行了数据预处理和数据清理，用以获得整洁数据来进行下一步的数据挖掘工作。

### 1.4.1 缺失值的处理

原始信用卡违约记录的统计表中含有大量的缺失记录，由于原始数据量较大，并且缺失的部分经过 python 统计只有 331 条，缺失记录在数据框中所占比例并不大，删除后并不影响原始数据的分布，所以选择删除这些记录。

### 1.4.2 年龄变量处理

通过 python 的统计了命令查询，我们可以清楚的看到如下表所示，由于共有 9 条记录是大于 100 岁的，其中有几条记录是大于 150 岁的明显不符合常理所以删除这些错误异常记录，其余有几条记录实在 100 岁左右的，职业是离退休人员，是比较可靠的，为了在后面的描述性统计分析以及数据挖掘中处理起来比较方便将这些年龄大于 100 岁的可靠记录的年龄全部修改为 100 岁，也就是说，本次分析的年龄上界为 100。还包含了许多并不合理的数据比如年龄有部分未满 18 周岁，这按相关的法律法规的规定是不准办理银行卡的，而有些记录办理行用卡的年龄过大，甚至超过了 120 岁的，这些记录也是不可信的，删除了这些违约记录，进行处理。

由于需要进行与违约程度的信息增益率的计算，所以将连续变量年龄，根据决策树 C4.5 的跟据最大信息增益率的方法将连续变量离散化，将年龄变量进行各年龄段的划分。划分代码请见附录，划分结果使得根据年龄段划分的条件信息增益率最大。最终将数据集划分为 5 份。分割点分别为 25.5,30.5,33.5,40.5,55.5 岁五个分割点。

### 1.4.3 职业变量处理

同样的又进行了对从事职业不合理数据的删除，比如依据相关法规，“学龄前儿童”是不允许办理信用卡的，所以在职业一类中排除记录为“学龄前儿童”的错误记录。

### 1.4.4 日期变量处理

最后对日期变量进行处理。因为考虑到顾客违约对现在的信用评价的影响，所以，首先，利用 datetime 库的函数求出每条违约记录发生时间距离 2017 年 11 月 1 日的时间间隔，“时间间隔”变量的单位为天。距离现在时间越长的违约记录所带来的影响较小，并作为一个新的变量存储到数据框中。又由于作业要求只采取 2016 年 1 月 1 日到 2017 年 11 月 1 日的数据，所以，将不在考察范围的数据记录（也就是时间间隔大于 670 天的和小于 0 天的）进行删除。

## 1.5 研究方法及思路

1、对 creditcard 的数据首先进行数据预处理后，对关心的人口特征变量，进行单变量和多变量的描述性统计分析，并对描述性分析结果进行相关问题的分析。并根据人口学特征分析各类用户群和违约程度的直观关系。

2、客户群信用程度评价体系的建模。在此部分的思路：第一步,计算性别，年龄，职业，教育程度四个变量与违约程度的信息增益率，根据信息增益率来划分其他变量与违约风险的相关程度，从而定义一个人口学特征变量之间的风险相对比例，然后进行归一化，获得各变量的加权系数。

3、各变量内部的分箱风险比率。根据的是每一个变量中的各个分箱中不同违约程度所占比例。轻度为 2 分，中度为 5 分，重度为 10 分，由各违约程度所占比例与违约程度得分相乘分类相加所得到变量各分箱风险得分。

4、定义时间衰减函数，将计算获得的风险得分进行按时间的衰减转化，越靠近近期也就是 11 月 1 日的违约衰减比率越小，也就是风险越大，对现在的信誉影响较大。

5、进行每个用户风险得分公式的各部分组合。计算每一个记录的风险得分，然后根据用户群分组求和求平均作为每一类用户群体的风险评价系数。

# 第二章 信用卡特征描述式数据挖掘

## 2.1 人口学特征化描述

### 2.1.1 违约情况与年龄情况分析

由于 creditcard 数据量较为大，所以无法直观的判别和比较用户的年龄和违约程度的关系，通过 python 的绘图功能来进行尝试，观察有违约情况的人群中在哪一个年龄段的分布比较密集：如下图 2-1 所示，是本次违约数据的年龄分布直方图，分布密度曲线以及分布象限图：

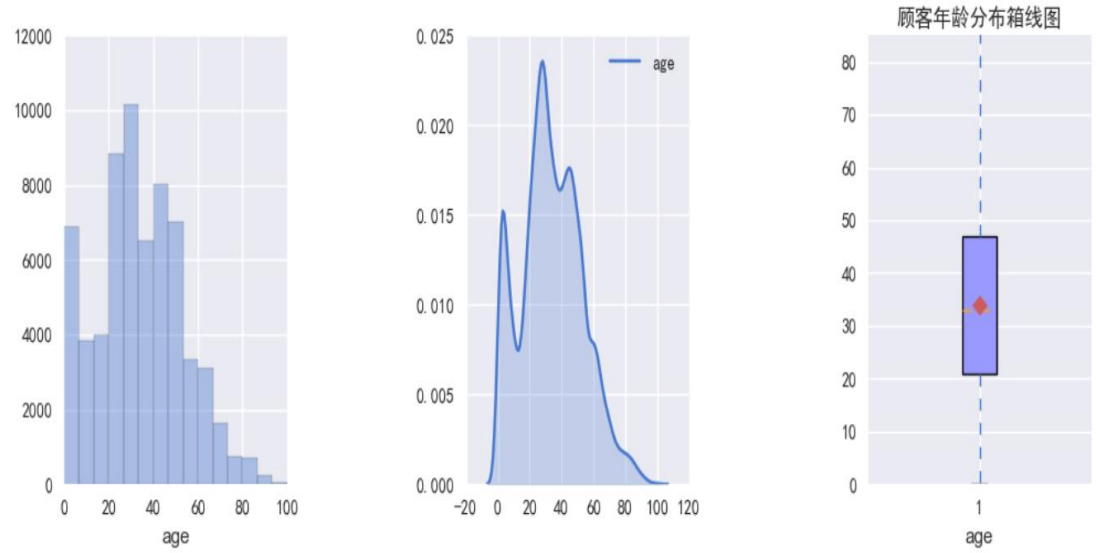


图 2-1 违约状况与年龄情况示意图

从图中直方图和分布密度曲线，可以看出违约状况整体呈右偏分布，右拖尾。说明年龄较大的违约者还是存在的，但所占比例比较小。违约年龄的中位数在和平均数都在 33 岁左右，去掉异常值点后，最大的违约年龄是 100 岁。出现违约情况最多的年龄段为 20 岁-50 岁之间，其中 30-40 岁更是高发期。可能是因为在该年龄段中，正处于事业和家庭的上升期，所以贷款较多，同样出现违约的次数就会很多。

由于本次数据挖掘任务是针对每一类顾客的违约评价体系的构建，所以自然地想到观测各个顾客群体的违约年龄分布，如下图 2-2 所示：

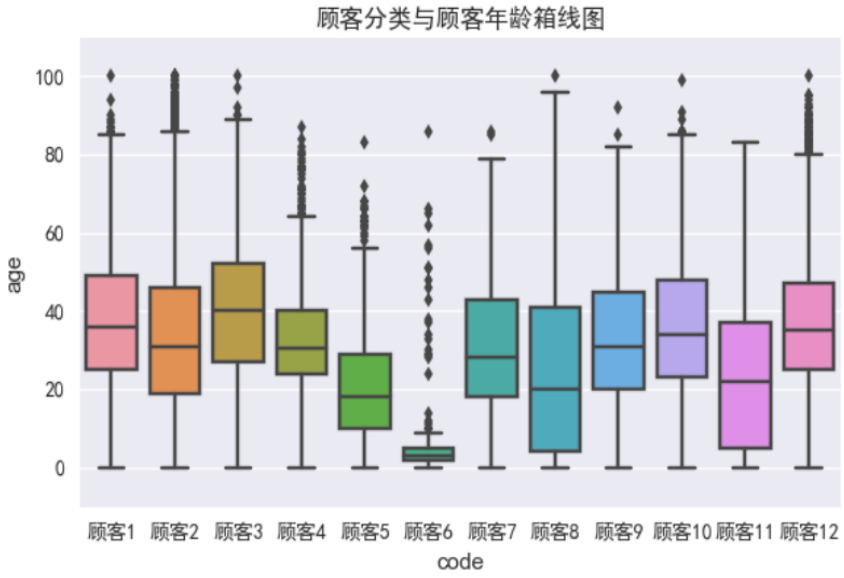


图 2-2 顾客群体违约年龄分布

从图 2-2 中可以看出每一类顾客群体的违约年龄分布状况，可以看到不同的顾客群体之间的违约年龄状况还是不尽相同的。从集中程度上来看，顾客 8 的违约年龄分布比较分散，而顾客 4 和 6 的分布比较集中，但顾客 6 存在着很多利群点。从分布年龄上看，其中顾客 6 的分布年龄较小，而顾客 1,3,10,12 的年龄分布较高。总而言之，顾客年龄上的分布还是有比较明显的差异的。

### 2.1.2 违约记录的职业分布状况

由于一个用户的职业情况在一定程度上就决定了其收入，也就间接影响着用户的信用状况。因此统计该数据集的违约状况与职业的相关关系比较有必要，数据量较大，故采用 python 可视化的方法来整理数据，并作出违约记录职业分布饼图，如下图 2-2 所示：

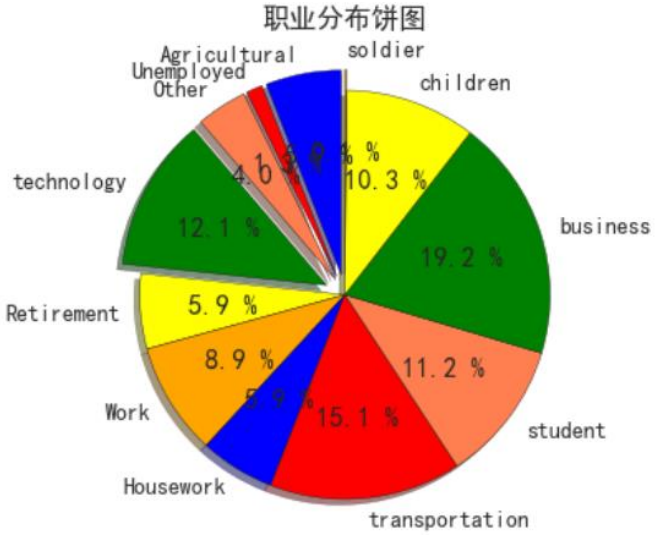


图 2-3 违约记录的职业分布情况

可以从图中看出，违约记录的各年龄段的职业分布差距还是有的，其中出现违约状况最多的职业是从事商务活动的人，占到总体的 19,2%。其次是从事交通运输活动的人，占到总体的 15.1%。随后是从事技术活动的人，占总体的 12.1%。出现最少的违约记录的职业是军人。说明军人的信用状况比较好，只有 1%左右的违约率。

### 2.1.3 违约程度与性别

在 creditcard 数据中不难发现性别这一变量，一般认为女性的违约风险比男性的违约风险低，但在大部分国家，女性的收入不及男性，所以是否把性别作为违约因素考虑是有争议的。虽然现在世界各国，针对性别这一变量对信用状况的影响作用都持有谨慎态度，目前越来越多的国家认识到，在信用评分中考虑性别是带有歧视色彩。在此只根据数据的状况展示一下违约程度与性别之间的统计状况，如下图 2-4 所示：

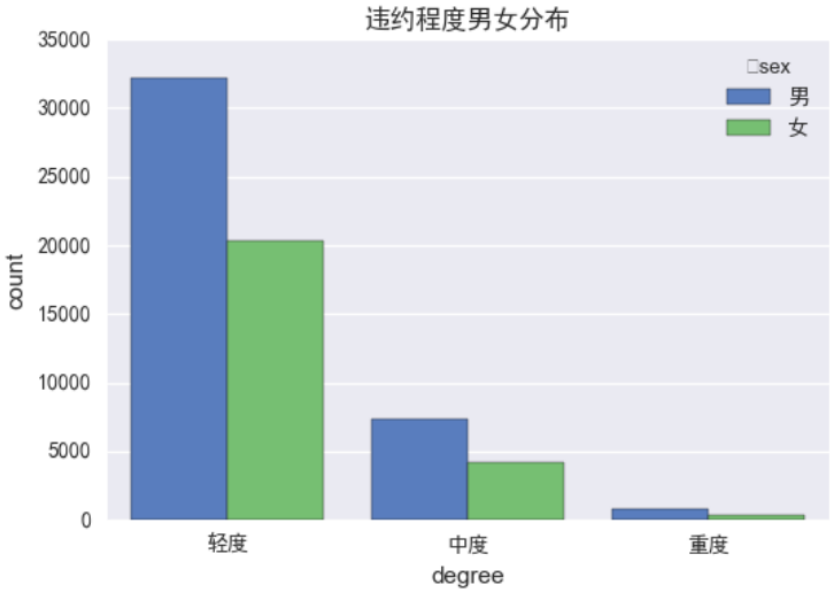


图 2-4 违约程度性别分布状况

从图中可以清楚地看到，不论哪个违约状况下，男性的违约数量都要大于女性的违约记录，其中轻度违约记录上男女之间的比例接近与 1.5:1 左右。而到了中度违约状况下，男女发生的违约比例有所增加，达到了 1.6:1 左右。在重度违约情况下，男女违约发生比为 2:1 左右。可见在信用卡上，男性更容易违约，因此在后面的用户群体评分上需要考虑这一因素。

### 2.1.4 违约记录的严重程度与性别分面统计

为了更加清晰到观察违约事件所发生的年龄分布，通过违约程度和性别等分类变量作为分面变量，做出了分面直方统计图，如下图 2-5 所示：

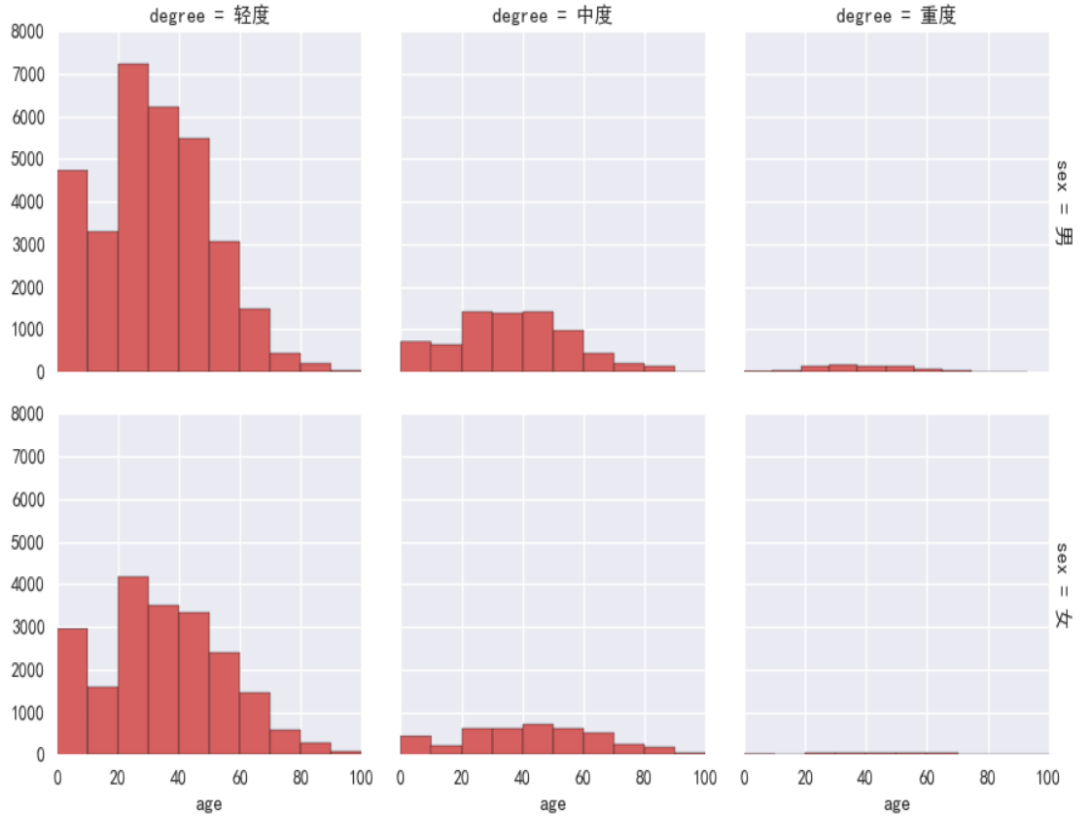


图 2-5　违约年龄分面统计图

从图中可以清晰的看到违约事件发生的年龄状况分布情况，可以结合图 2-1 分析得出，在轻度违约状况下，不论男女性别发生的人数都是最多的，且年龄分布状况与整体的年龄分布状况非常接近，都是出于右偏分布，而到了中度的违约状况下，人数急剧减少，年龄分布又处于以 40 岁为中心的正态分布状况。到了重度违约的情况下男女性别的数量分布几乎接近与为零。只有少量分布在 30-50 岁之间。可以看到违约年龄的分布与性别和违约程度还是有很大的差别的。

### 2.1.5 违约记录与受教育状况统计

依据所收集的资料来看，如上一章 1.3 节所讲述的内容上来看，我们不难得出结论每一名顾客的受教育情况，一定程度上也就影响了一个人的信誉状况和收入状况，也就进而的影响了该名顾客的违约状况，如图 2-6 所示;

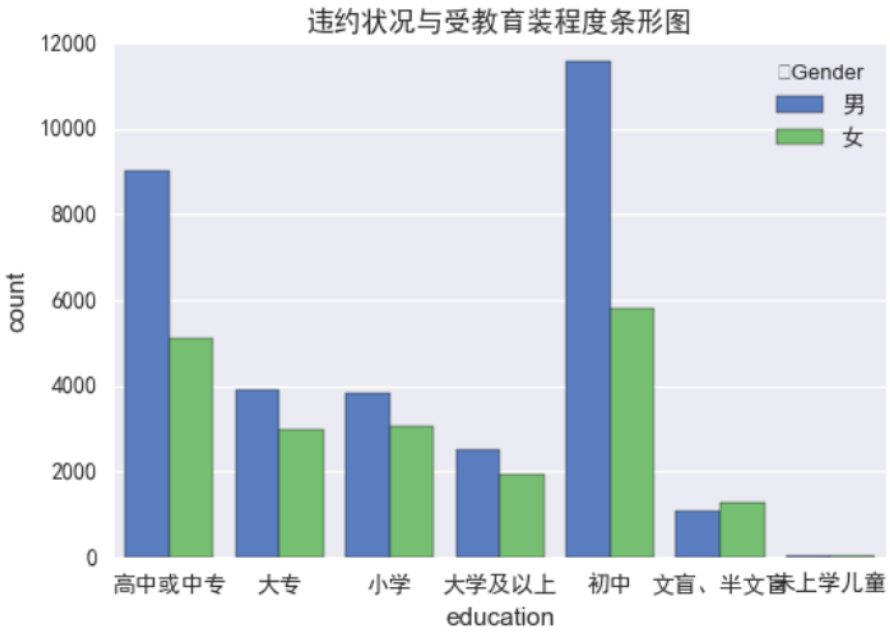


图 2-6 学历与违约情况柱状图

从柱状图中可以看出教育程度为初中的顾客，违约的数量较多，其次是高中和大专，这与受教育程度越高信誉状况越好的印象不尽相同，而是一种出乎意料的根据受教育程度违约风险先上升后下降的情况。在第三章的评价体系的构建上，也要考虑到这一与通常认知不相符的一点。

# 第三章 用户群体信用评价体系

依照数据挖掘知识来建立用户群体信用评价体系。评分体系主要由五部分构成，也可以说每类用户群体的评价公式主要由四个部分构成。分别是：变量间的权重，变量内部水平得分，时间指数衰减因子确定，个人风险函数的合成，群体平均风险指数的计算

## 3.1 信息增益率确定变量间权重

性别，年龄，学历，职业，违约程度这五个变量主要是衡量每一个用户的违约风险的重要变量，而违约程度又是和其他的四个变量有着一定的区别，因为违约程度是直接影响违约风险的变量，而其他的四个变量是间接影响变量。因此将五个变量分为两类，一类为违约程度，一类为人口学特征四个变量。两类指标富裕相同的权重 0.5。

但人口学信息所带来的违约风险又如何确定呢？在这里我们用违约程度作为监督变量来衡量人口学信息的 4 个变量。也就是说，衡量 4 个人口学信息与违约程度的相关关系，来确定每一个变量对违约风险的影响程度。

相关关系的衡量变量利用的数据挖掘知识是信息论中的信息增益率。信息增益率也叫信息增益比例，是信息增益的改进衡量指标，采用增益比例作为选择属性的标准，克服了信息增益度量的缺点，并不受单个指标的内部变量水平数量的影响。可以统一衡量各变量与违约程度的关系。

计算信息熵: $\text{entropy(S)} = -\sum_{i=1}^{n} p_i log_2 p_i$

计算条件信息熵： $\text{entropy}(S, A) = \sum_{i=1}^{m} \frac{|S_i|}{|S|} entropy(S_i)$

计算信息增益： $\text{gain}(S, A) = \text{entropy(S)} - \text{entropy}(S, A)$

计算信息分裂比例： $\text{split\_info}(S, A) = -\sum_{i=1}^{m} \frac{|S_i|}{|S|} log_2 \frac{|S_i|}{|S|}$

计算信息增益比例： $\text{gain}_{\text{radio}(S,A)} = \frac{gain(S,A)}{split\_info(S,A)}$

可以这样理解如果按一个人口学变量进行分类，使得违约程度的信息熵（也就是信息混乱程度）减小，那么就说明该变量在一定程度上与违约程度有着较强关系，那么也就和违约风险有着较强关系。

通过 python 计算的四个人口学特征变量的信息增益率如下表 3-1 所示：

表 3-1 人口学特征信息增益率

| 排序 | 人口学特征 | 信息增益率 | 归一化权重 |
|---|---|---|---|
| 1 | 职业 | 0.08603 | 0.4629 |
| 2 | 年龄 | 0.06440 | 0.4000 |
| 3 | 学历 | 0.01462 | 0.0786 |
| 4 | 性别 | 0.01080 | 0.0581 |

## 3.2 变量内部水平得分确定

变量内部水平的确定也是划分为两类变量，仍然是一类为违约程度，一类为人口学特征的四个变量。违约程度分为“轻度”“中度”“重度”三个水平，根据载资料中查找到的巴塞尔协议的相关内容，将三个违约程度的违约风险得分分别赋予 2:5:10。

人口学特征变量的内部水平风险得分的确定方式。以职业为例，首先计算每一个职业中轻度，中度，重度违约情况所占比例（如表 3-2 所示）。

| | occupation | degree | count | radio | weight | score |
|---|---|---|---|---|---|---|
| 0 | 专业技术人员 | 中度 | 1551 | 0.199306 | 5 | 0.996530 |
| 1 | 专业技术人员 | 轻度 | 6032 | 0.775122 | 2 | 1.550244 |
| 2 | 专业技术人员 | 重度 | 199 | 0.025572 | 10 | 0.255718 |
| 3 | 其他/不清楚 | 中度 | 508 | 0.198205 | 5 | 0.991026 |
| 4 | 其他/不清楚 | 轻度 | 2014 | 0.785798 | 2 | 1.571596 |
| 5 | 其他/不清楚 | 重度 | 41 | 0.015997 | 10 | 0.159969 |
| 6 | 军人 | 中度 | 16 | 0.210526 | 5 | 1.052632 |
| 7 | 军人 | 轻度 | 58 | 0.763158 | 2 | 1.526316 |
| 8 | 军人 | 重度 | 2 | 0.026316 | 10 | 0.263158 |
| 9 | 农牧渔水利业生产人员 | 中度 | 888 | 0.232096 | 5 | 1.160481 |

表 3-2 职业内部水平风险得分表（部分）

然后依据 2：5：10 的权重与各类所占比例进行相乘相加，最终得到每一类职业的风险得分。各职业得分表如下表所示:

表 3-3 各职业风险得分系数表

| occupation | count | score |
|---|---|---|
| 专业技术人员 | 7782 | 2.802493 |
| 其他/不清楚 | 2563 | 2.722591 |
| 军人 | 76 | 2.842105 |
| 农牧渔水利业生产人员 | 3826 | 2.997386 |
| 办事人员和有关人员 | 5712 | 2.607843 |
| 商业、服务业人员 | 12340 | 2.539141 |
| 在校学生 | 1820 | 2.534615 |
| 家务 | 3659 | 2.733534 |
| 待业 | 787 | 2.961881 |
| 生产运输设备操作人员 | 9625 | 2.699948 |
| 离退休人员 | 3867 | 3.029997 |

其余的人口学信息变量与职业变量计算方式相似。

## 3.3 时间指数衰减因子确定

很直观的想到距离时间节点 2017 年 11 月 1 日越近的违约记录越是对现在的影响较大，所以通过分析定义了一个违约风险关于距离时间节点的时间间隔的指数衰减因子 w，定义为：

$$\mathrm{w} = e^{-\frac{inv}{670}}$$

inv 为某一条记录的时间间隔，670 为 2016 年 1 月 1 日到 2017 年 11 月 1 日的时间间隔量。w 即为所定义的违约风险的时间衰减因子。

## 3.4 银行损失分析评价模型

根据以上 3.1-3.3 的损失分析评价模型的分段阐述，由此合成单个用户的信用风险评价模型，如下面所示：

$$\left[\frac{1}{2}\times(\textit{违约程度})+\frac{1}{2}\times(0.46\ \textit{职业分}+0.4\ \textit{年龄分}+0.08\ \textit{学历分}+0.06\ \textit{性别分})\right]e^{-\frac{inv}{670}}$$

依据如上公式所计算出的风险得分，只是每个用户（每个记录）的得分，依据顾客群体分组求和取平均，最后计算出最终得分才是用户群体的得分。总后计算的得分状况如下表 3-4 所示：

表 3-4 用户群体信用风险得分

| 排序 | 顾客群 | 得分 |
|---|---|---|
| 1 | 顾客 6 | 3.264 |
| 2 | 顾客 8 | 3.025 |
| 3 | 顾客 2 | 2.856 |
| 4 | 顾客 1 | 2.658 |
| 5 | 顾客 10 | 2.553 |
| 6 | 顾客 12 | 2.409 |
| 7 | 顾客 5 | 2.326 |
| 8 | 顾客 9 | 2.113 |
| 9 | 顾客 11 | 1.957 |
| 10 | 顾客 7 | 1.862 |
| 11 | 顾客 3 | 1.773 |
| 12 | 顾客 4 | 1.505 |

通过最终计算汇总的表 3-5 的用户群风险评价来看，用户群 6 获得了较高的风险评价分数，也就是说被标记为用户群 6 的用户信用状况比较令人堪忧，在信用管理中管理决策者在制定营销策略和给用户发放透支额度时需要对被分类为客户 6 的用户更加谨慎，而对于那些信用风险得分较低的用户群体则可以制定更为宽松的营销策略，发放的可透支额度更加的高。利用用户群风险评价体系，可以更加高效和有目标性的管理银行信用状况。

# 第四章 结论及展望

结合从网上搜集的影响顾客信用的因素的各方面资料以及结合第二章的描述性统计分析上来看，不论是性别，年龄，职业和受教育程度来看，都对违约程度有一定影响，同时也就对违约风险产生间接影响，因此将人口信息变量也作为衡量违约风险的变量。并通过违约程度进行监督，利用信息论中的信息增益率来衡量 4 个人口信息学变量与违约程度的关系，进而判断人口信息变量与目标量违约风险的系数关系，经过归一化就确定了权重系数。变量间的分箱得分是通过与违约程度比例来确定。

然后通过变量的初步组合获得违约风险，并通过时间衰减因子来获得最终的每个记录的信用得分。并经过依据顾客群进行分组求和求平均进行用户群得分的计算。

有最终结果可以看出，用户群的得分差异很大。且被分类为顾客 6 的用户信用风险较高，在银行的经营过程中需要注意，而用户 4，3,7 则拥有较好的信用风险评分，所以可以在银行经营过程中，提供较高的信用额度。合理的应用用户群信用评分可以保证银行的高效率运转。

本文也有些不足，比如违约程度与人口信息变量间的 1:1 的赋权比例师认为指定的，没有详细的科学依据。