西安欧亚学院

《信用风险度量与管理》

题 目 个人征信报告

学生姓名	刘嘉玲
学生学号	18610608150094
指导教师	胡留所
导师职称	讲师
所在分院	金融学院
专　　业	数据科学与大数据技术
班　　级	统本大数据 1802 班
提交日期	二〇二〇年十二月

目 录

1 引言

1.1 选题的背景和意义

本文基于2005年台湾信用卡客户数据，建立逻辑回归、决策树、KNN及随机森林模型来探索影响客户信用的关键因素，包括个体特征及某些客观特征，通过比较模型的预测准确度对银行信用卡违约进行预测分析。

意义：信用卡对于银行来说是高收益和高风险并存的业务，伴随信用卡业务发展的是各大银行都在利用网络和移动端的数据来建立客户的信用评分系统。如何从客户所填的资料里对客户进行信用评估、如何鉴别所填资料的真假性及应该要求客户填什么类型的资料等对银行来说是至关重要的。信用卡违约预测模型的建立以及影响客户信用的关键因素的探索，对于银行选择客户和设计资料填写具有重要的指导价值，并且能够为信贷决策提供一定的理论支持，具有很强的理论和现实意义。

2 相关理论综述

2.1 数据平衡述评

不平衡数据，顾名思义，就是指在收集到的数据中各个分类之比并非为1:1，在对不平衡数据的研究中，普遍认为不平衡意味着少数类所占比例在10%到20%之间，但实际上，这种现象可能会更严重。机器学习中常常会遇到数据的类别不平衡（class imbalance），也叫数据偏斜（class skew）。在这种情况下，学习出好的分类器是很难的，而且在这种情况下得到结论往往也是很具迷惑性的。

处理不平衡数据的思路比较简单，那就是想办法让数据平衡，我们可以简单得分为以下几类：

- 更改数据集中各分类数据的量，使他们比例匹配——常用方法有采样、数据合成；
- 更改数据集中各分类数据的权重，使他们的量与权重之积匹配——常用方法为加权；
- 不修改数据集，而是在思路上将不平衡数据训练问题转化为一分类问题或者

异常检测问题（少数类就像是存在于多数类中的异常值）。

2.2 独热编码述评

独热编码即 One-Hot 编码，又称一位有效编码，其方法是使用 N 位状态寄存器来对 N 个状态进行编码，每个状态都由他独立的寄存器位，并且在任意时候，其中只有一位有效。

独热编码（哑变量 dummy variable）是因为大部分算法是基于向量空间中的度量来进行计算的，为了使非偏序关系的变量取值不具有偏序性，并且到圆点是等距的。使用 one-hot 编码，将离散特征的取值扩展到了欧式空间，离散特征的某个取值就对应欧式空间的某个点。将离散型特征使用 one-hot 编码，会让特征之间的距离计算更加合理。离散特征进行 one-hot 编码后，编码后的特征，其实每一维度的特征都可以看做是连续的特征。就可以跟对连续型特征的归一化方法一样，对每一维特征进行归一化。比如归一化到[-1,1]或归一化到均值为 0,方差为 1。解决了分类器不好处理属性数据的问题，在一定程度上也起到了扩充特征的作用。它的值只有 0 和 1，不同的类型存储在垂直的空间。

2.3 数据归一化述评

把数据变成(0，1)或者（1,1）之间的小数。主要是为了数据处理方便提出来的，把数据映射到 0～1 范围之内处理，更加便捷快速。把有量纲表达式变成无量纲表达式，便于不同单位或量级的指标能够进行比较和加权。归一化是一种简化计算的方式，即将有量纲的表达式，经过变换，化为无量纲的表达式，成为纯量。

2.5 逻辑回归述评

逻辑回归（Logistic Regression）是一种用于解决二分类（0 or 1）问题的机器学习方法，用于估计某种事物的可能性。这里用的是“可能性”，而非数学上的“概率”，logisitc 回归的结果并非数学定义中的概率值，不可以直接当做概率值来用。该结果往往用于和其他特征值加权求和，而非直接相乘。

2.6 随机森林述评

随机森林方法通俗的讲，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样

本应该属于哪一类(对于分类算法)，然后看看哪一类被选择最多，就预测这个样本为那一类。

在建立每一棵决策树的过程中有两点需要注意，采样与完全分裂。

2.7 决策树述评

决策树是一种机器学习的方法。决策树的生成算法有 ID3, C4.5 和 C5.0 等。决策树是一种树形结构，其中每个内部节点表示一个属性上的判断，每个分支代表一个判断结果的输出，最后每个叶节点代表一种分类结果。

决策树是一种十分常用的分类方法，需要监管学习（有教师的 Supervised Learning)，监管学习就是给出一堆样本，每个样本都有一组属性和一个分类结果，也就是分类结果已知，那么通过学习这些样本得到一个决策树，这个决策树能够对新的数据给出正确的分类。

2.8 KNN 述评

K 近邻算法，即是给定一个训练数据集，对新的输入实例，在训练数据集中找到与该实例最邻近的 K 个实例，这 K 个实例的多数属于某个类，就把该输入实例分类到这个类中。

KNN 是通过测量不同特征值之间的距离进行分类。它的思路是：如果一个样本在特征空间中的 k 个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别，其中 K 通常是不大于 20 的整数。KNN 算法中，所选择的邻居都是已经正确分类的对象。该方法在定类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。

3 数据说明

本文采用的数据是来自 UCI 网站中台湾地区信用卡客户数据，研究目的是对信用卡客户是否违约做一个预测，因此响应变量是一个二分类变量，违约记为 1，未违约记为 0。并且本文使用 23 个变量作为解释变量：

- ID：每个客户的 ID
- LIMIT_BAL：以新台币计的给定信用额度（包括个人和家庭/辅助信用额）
- 性别：性别（1 =男性，2 =女性）
- 教育程度：（1 =研究生院，2 =大学，3 =高中，4 =其他，5 =未知，6 =未知）
- 婚姻：婚姻状况（1 =已婚，2 =单身，3 =其他）
- 年龄：岁

- PAY_0：2005 年 9 月的还款状态（-1 =正常付款，1 =延迟一个月的付款，2 =延迟两个月的付款，…8 =延迟八个月的付款，9 =延迟 9 个月以上的付款）
- PAY_2：2005 年 8 月的还款状态（与上述相同）
- PAY_3：2005 年 7 月的还款状态（与上述相同）
- PAY_4：2005 年 6 月的还款状态（与上述金额相同）
- PAY_5：2005 年 5 月的还款状态（与上述相同）
- PAY_6：2005 年 4 月的还款状态（与上述相同）
- BILL_AMT1：2005 年 9 月的帐单金额（新台币）
- BILL_AMT2：2005 年 8 月的对帐单金额（新台币）
- BILL_AMT3：2005 年 7 月的帐单金额（新台币）
- BILL_AMT4：2005 年 6 月的帐单金额（新台币）
- BILL_AMT5：2005 年 5 月的对帐单金额（新台币）
- BILL_AMT6：2005 年 4 月的帐单金额（新台币）
- PAY_AMT1：2005 年 9 月的先前付款金额（新台币）
- PAY_AMT2：2005 年 8 月的先前付款金额（新台币）
- PAY_AMT3：2005 年 7 月的先前付款金额（新台币）
- PAY_AMT4：2005 年 6 月的先前付款金额（新台币）
- PAY_AMT5：2005 年 5 月的先前付款金额（新台币）
- PAY_AMT6：2005 年 4 月的先前付款金额（新台币）
- default.payment.next.month：默认付款（1 =是，0 =否）

4　探索与描述性分析

4.1　绘制整体自变量数据因子相关性系数矩阵

由图 4-1 自变量数据因子相关性系数矩阵图，我们得出账单之前有正的强相关，教育程度和婚姻状况有强的负相关。

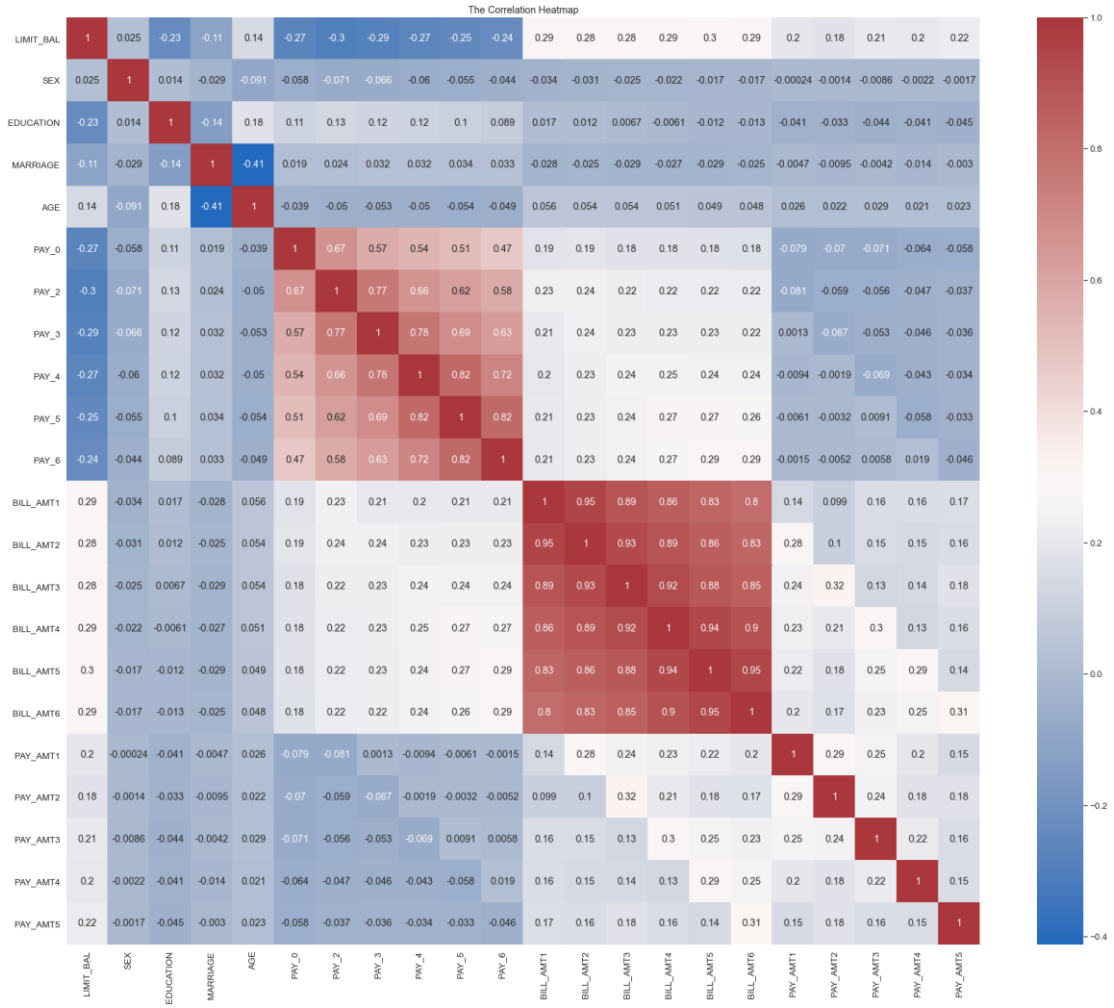


图 4-1 自变量数据因子相关性系数矩阵图

4.2　描述性分析

由下图 4-2 是否默认付款频率分布图得出，是默认付款的占 77.9%，主动付款的占 22.1%，由此可见，该二分类数据并不平衡，我们需要将数据进行一系列处理时期均衡，为后续建模做好准备。

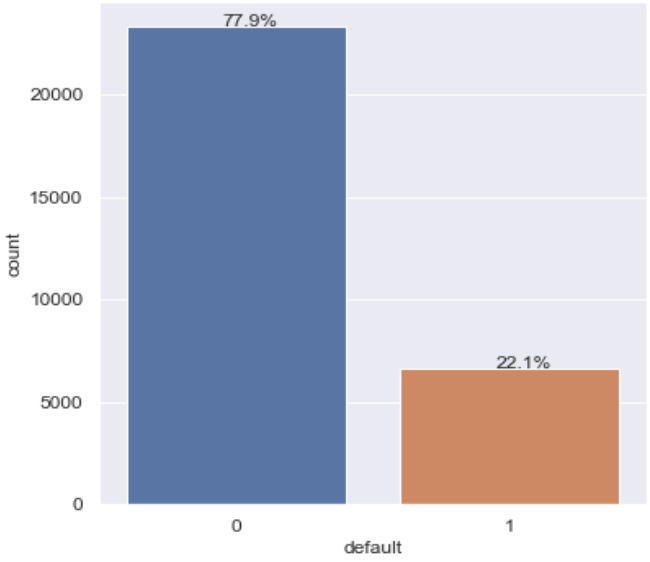


图 4-2 是否默认付款频率分布图

由下图 4-3 教育程度对是否默认付款频率分布图得出，大学中不论是否默认还款在其他教育程度中都是最多的，教育程度为其他中并没有是默认付款的。

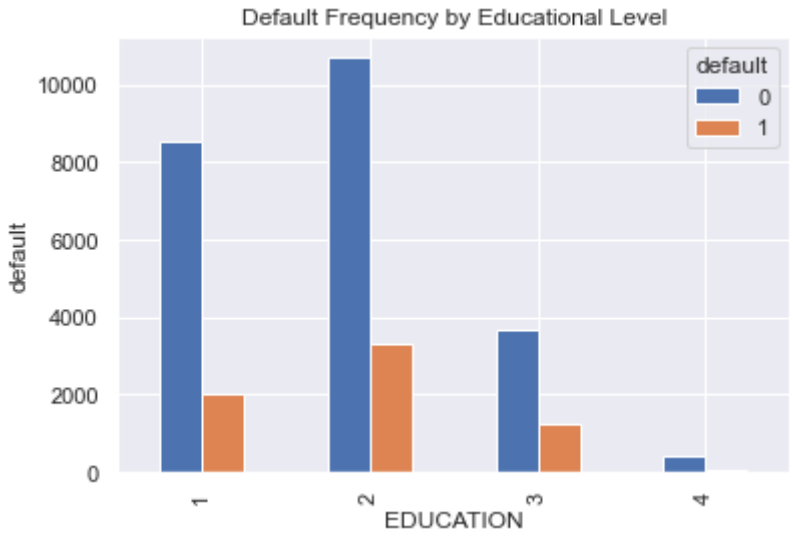


图 4-3 教育程度对是否默认付款频率分布图

由下图 4-4 婚姻状况对是否默认付款频率分布图得出，是否结婚对默认还款的影响不是很大，但是单身主动付款数还是最多的。

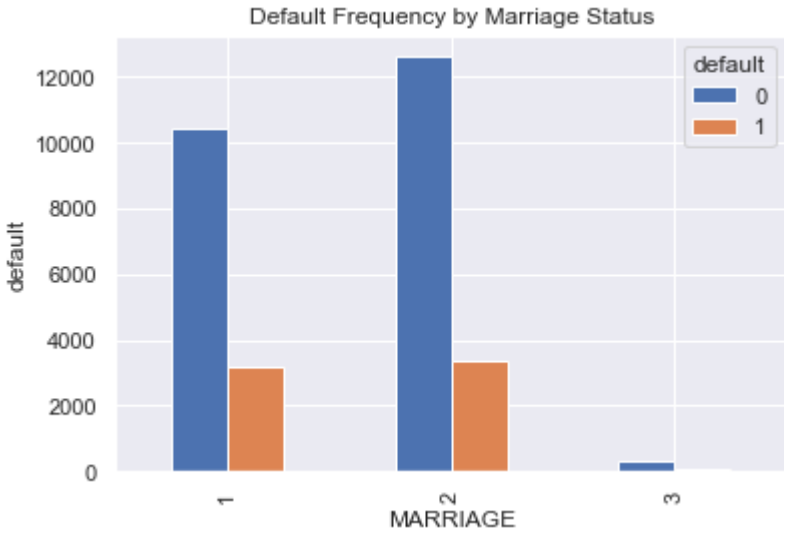


图 4-4 婚姻状况对是否默认付款频率分布图

由下图 4-5 性别对是否默认付款频率分布图得出，女性不论是默认付款还是主动付款都比男生多的，可见男生没有女生在平常更关注理财。

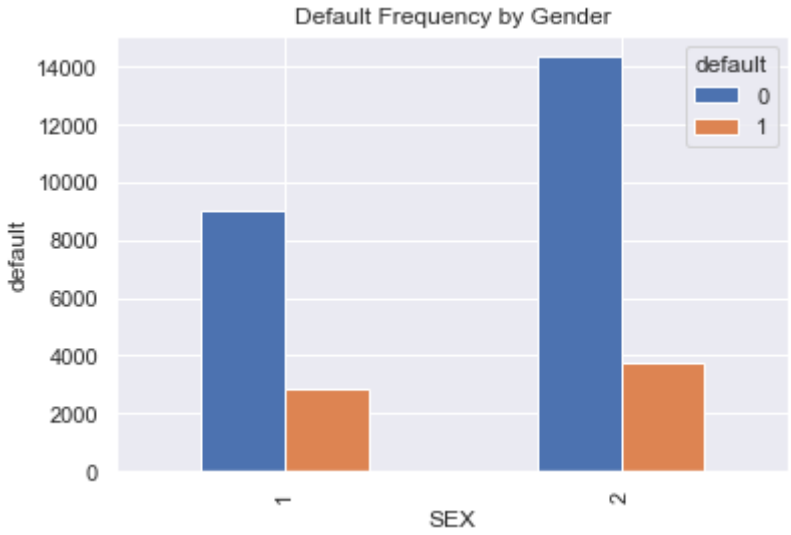


图 4-5 性别对是否默认付款频率分布图

由下图 4-6 LIMIT_BAL 的频率直方图得出，以新台币计的给定信用额度呈递减趋势，随着额度的减小，所占数量也越来越少。

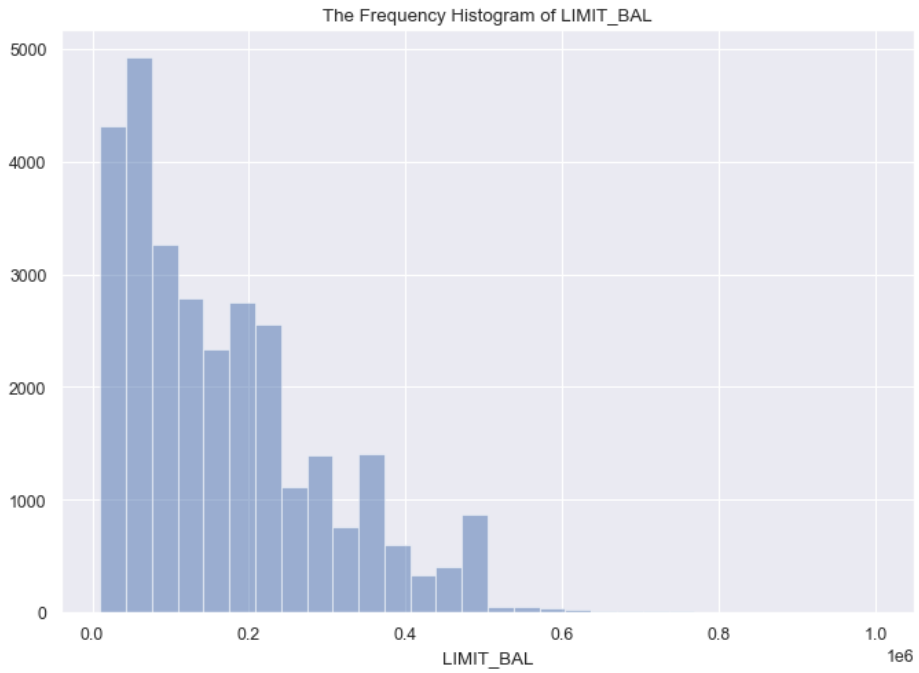


图 4-6 LIMIT_BAL 的频率直方图

由下图 4-6　年龄频率直方图得出，该自变量呈右偏分布。

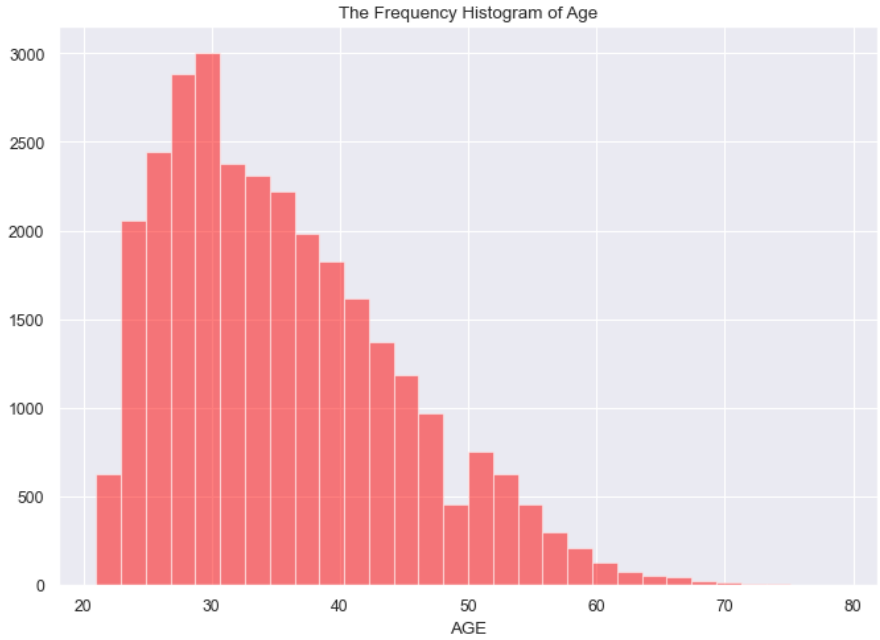


图 4-7 年龄频率直方图

图 4-8 各分类变量的频率直方图

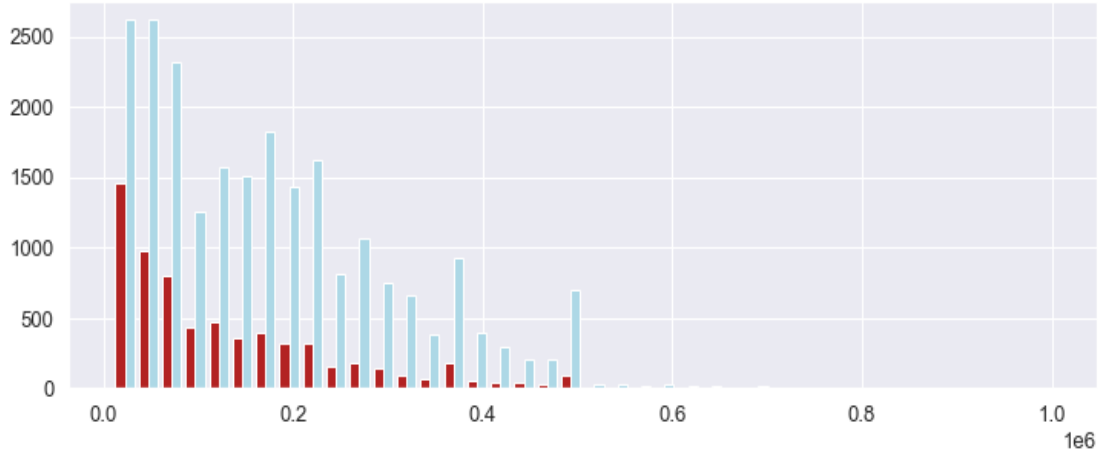


图 4-9 限制平衡的默认直方图

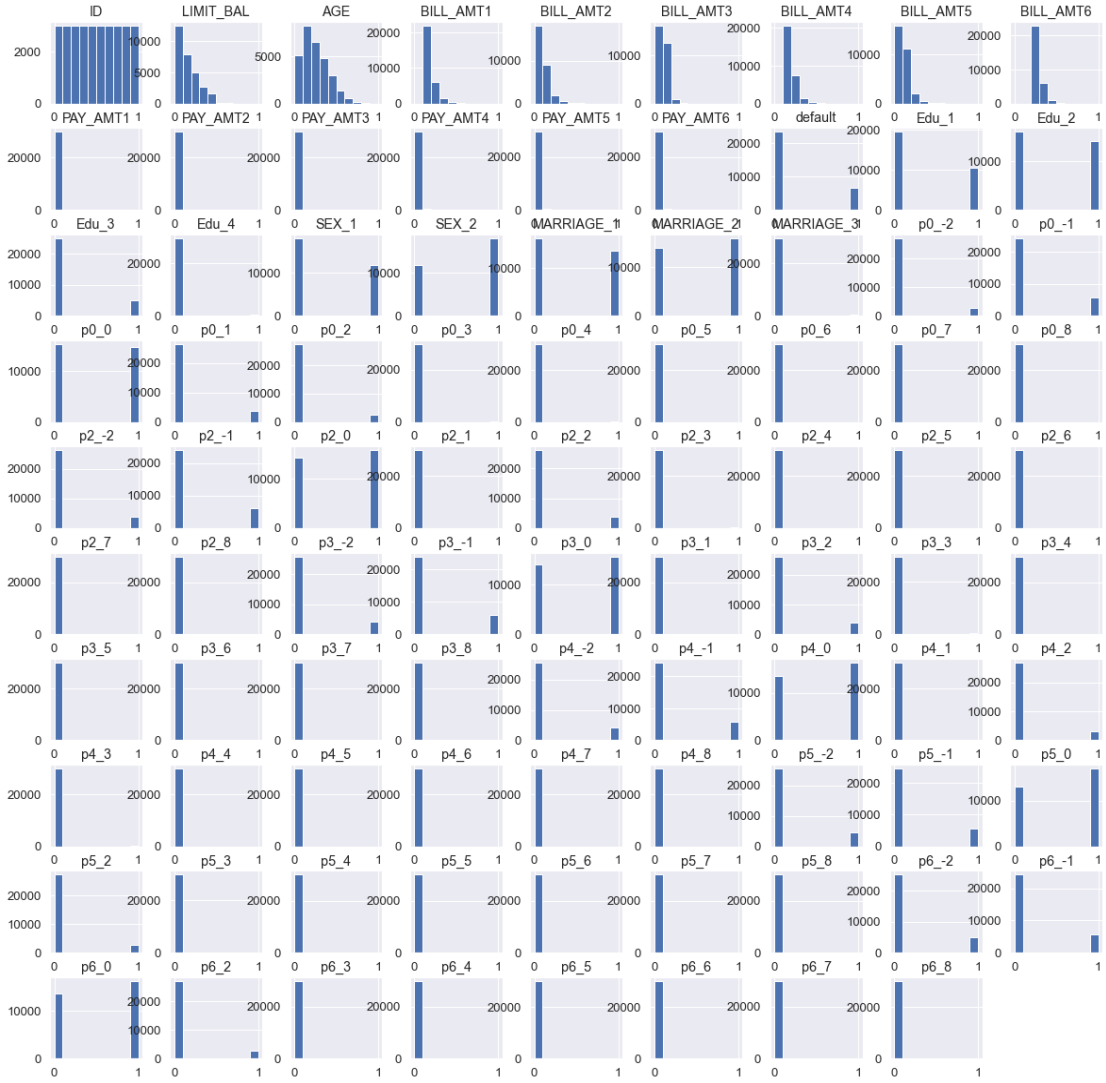


图 4-9 归一化:各因子应用最小最大标量展示

5 模型的建立

5.1 在不平衡数据上建立逻辑回归模型

混淆矩阵

4493	210
829	468

	precision	recall	f1-score	support
0.0	0.84	0.96	0.90	4703
1.0	0.69	0.36	0.47	1297
accuracy			0.83	6000
macro avg	0.77	0.66	0.69	6000
weighted avg	0.81	0.83	0.81	6000

5.2 在平稳数据上建立逻辑回归模型

混淆矩阵

3888	815
548	749

	precision	recall	f1-score	support
0.0	0.88	0.83	0.85	4703
1.0	0.48	0.58	0.52	1297
accuracy			0.77	6000
macro avg	0.68	0.70	0.69	6000
weighted avg	0.79	0.77	0.78	6000

5.3 建立随机森林模型

混淆矩阵

4267	436
737	560

	precision	recall	f1-score	support
0.0	0.85	0.91	0.88	4703
1.0	0.56	0.43	0.49	1297
accuracy			0.80	6000
macro avg	0.71	0.67	0.68	6000
weighted avg	0.79	0.80	0.79	6000

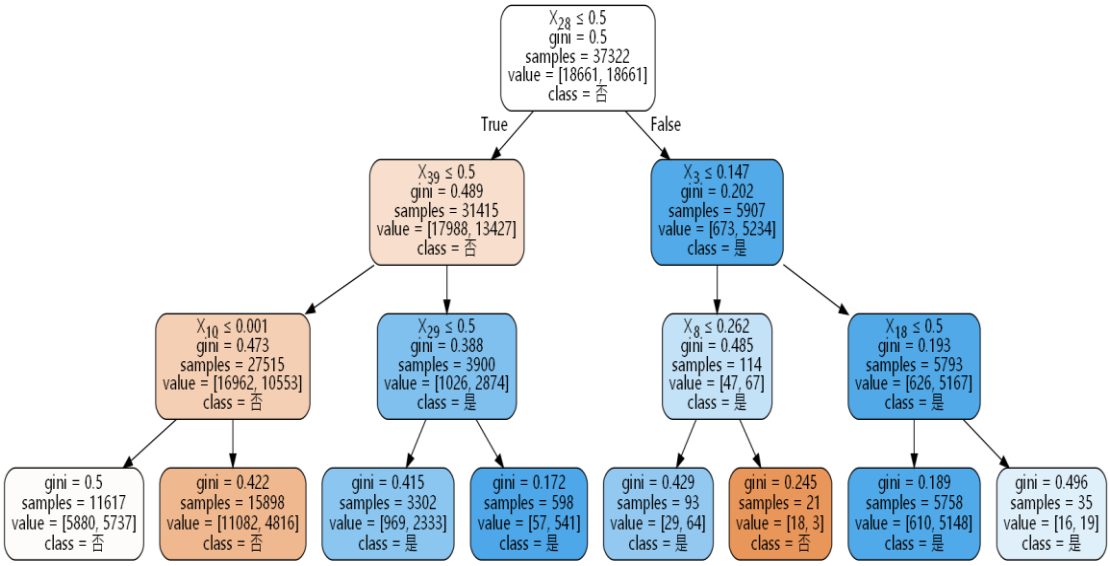
5.4 建立决策树模型

混淆矩阵

4288	415
742	555

The Classification Report of Decision Tree Model

	precision	recall	f1-score	support
0.0	0.85	0.91	0.88	4703
1.0	0.57	0.43	0.49	1297
accuracy			0.81	6000
macro avg	0.71	0.67	0.69	6000
weighted avg	0.79	0.81	0.80	6000



5.5 建立 KNN 模型

混淆矩阵

3231	1472
547	750

	precision	recall	f1-score	support
0.0	0.86	0.69	0.76	4703
1.0	0.34	0.58	0.43	1297
accuracy			0.66	6000
macro avg	0.60	0.63	0.59	6000
weighted avg	0.74	0.66	0.69	6000

6 模型的选择与精度评价

四个模型精度对比

模型	精度	标准差
逻辑回归	0.646637	0.111970
随机森林	0.933526	0.033366
决策树	0.882379	0.083252
KNN	0.720513	0.096524

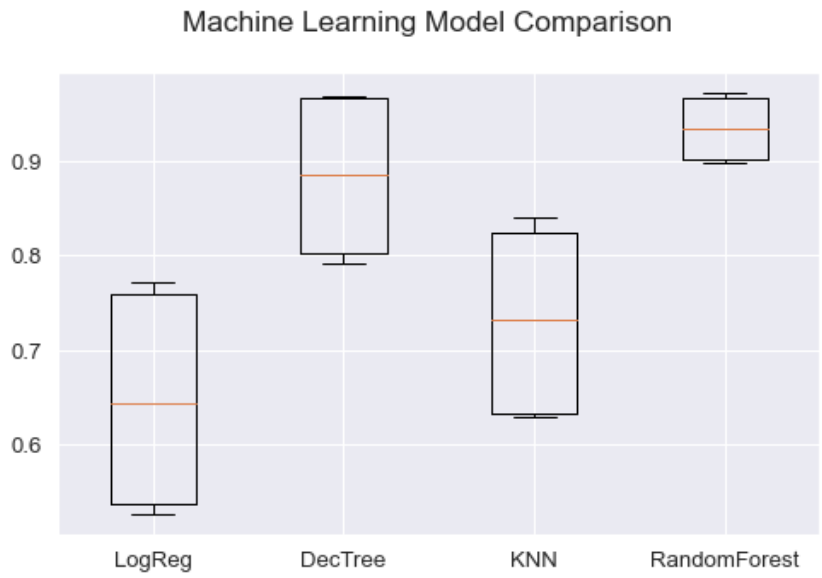


图 6-1 模型精度对比箱线图

本文建立了逻辑回归、决策树、KNN 和随机森林四种模型作为本文信用卡是否默认付款的预测模型，两种模型的预测效果即预测准确率(Accuracy)以及 F scores (Precise, recall)指标如上图所示：发现随机森林预测准确度以及 F 得分都比其他模型高。因此在本文信用卡是否默认付款预测模型分析中，随机森林预测效果要优于其他的模型。

随机森林在本文数据集的基础上表现良好，源于随机森林本身所具有的优势。随机森林能够处理高维度大数据，在处理过程中不需要做特征选择，但在训练完后，能够给出哪些特征变量比较重要，并且可以进行并行化的训练，训练速度快。因此数据挖掘技术在很多时候能解决经典统计方法在进行数据分析的时候所遇到的一些问题。

7 结论与建议

7.1 结论

本文主要针对信用卡客户的基本数据建立了逻辑回归、决策树、KNN 和随机森林四种模型客户违约预测模型，并进行了模型预测效果的对比，发现预测效果随机森林优于其他的模型。

综合以上分析，我们看到在预测信用卡客户违约中，客户的某些个体特征如教育水平、性别、年龄以及总信用额度、月份的消费及还贷情况等对信用违约与否有着明

显的影响。而这些客户的资料属于很容易收集到的基本信息，我们可以建立一种信用评分机制。

7.2 建议

目前国内申请信用卡填表资料与该数据所涉及到的变量相同，所以可以从该数据的研究结果适用于国内。基于以上研究，本文结合国内信用卡及公民信用相关知识从以下几个角度给出政策建议：

1) 严格审批、预防为主，银行应设立信用卡申请的初审制度，规范相关申请批准流程，严格把关，特别是对于被评估为潜在高违约的人群，应设置较高的审批标准，比如设定更高的手续费率、违约金、滞纳金等，在源头上控制风险，可以有效降低风险。

2) 实时监督，银行在信用卡发放后，可对持卡人的消费行为进行动态跟踪及评级，随时掌握客户的消费情况，建立动态机制，及时更新客户的信用特征信息。特别是高违约客户，一般消费额也高，对其账户实时监督，并且对于消费额高又长期逾期不还的客户可及时加入黑名单中，以此对风险进行很好的防范。

3) 征信系统完善，本文数据所涵盖的信息只局限于信用卡业务的信用信息(性别年龄、教育水平、婚姻状况及总额度、消费和还款情况)，还不能充分反映作为社会中的一个个体所有的信用。这些信用资料都散布在各个职能部门机关和相关单位，因此这些个人的信用数据征信困难，部门单位之间信用数据共享能力差，导致了银行只能依靠自身所具有的数据，建立的信用卡评分模型也就不能全面反映客户的信用情况。因此需要加强全国统一联网的个人征信体系的建设，完善征信系统平台，才能让我们的经济活动更安全，更有保障。

4) 信用卡业务创新，当信用卡违约率上升时，若只凭信用卡的某项业务的收益，是无法保证信用卡业务的持续发展。因此银行需要创新信用卡的业务，增加业务的多样性。

参考文献：

[1]隋孟琪. 基于混合特征提取和集成学习的个人贷款违约预测研究[D].电子科技大学,2020.

[2]陶艳丽. 随机森林改进模型对个人信贷违约预测的研究[D].河北经贸大学,2020.

[3]周瑞珍.基于 Python 的贷款违约预测[J].电子技术与软件工程,2019(22):157-158.

[4]刘开元.随机森林与逻辑回归模型在违约预测中的应用[J].信息与电脑(理论版),2016(21):111-112. [5]路颖颖. 基于非平衡数据集的 P2P 网络借贷违约预测研究[D].山西大学,2020.

[6]周翔,张文宇,江业峰.个人信贷违约预测模型的研究[J].辽宁科技大学学报,2020,43(03):223-230.

姓名：刘嘉玲

2020 年 12 月 27 日