****



**本 科 毕 业 设 计**

院 系 工程管理学院

专 业 计算机与金融工程实验班

题 目 基于数据挖掘的信用卡违约

风险及其影响因素的实证研究

年 级 2016级 学 号 161278015

学生姓名 李康

指导教师 方立兵 职 称 副教授

职 称

提交日期 2020年6月9日

**南京大学本科生毕业论文（设计、作品）中文摘要**

题目：基于数据挖掘的信用卡违约风险及其影响因素的实证研究

院系：工程管理学院

专业：计算机与金融工程实验班

本科生姓名：石垚

指导教师（姓名、职称）：李心丹教授 宋育全博士

摘要：随着我国经济的快速发展和居民消费观念的日益转变，近年来银行的信用卡业务处于快速增长阶段。但信用卡作为一项高风险与高收益并重的业务，对银行的风险监测与管理提出了更高的考验。本研究从信用卡违约预测出发，结合我国某城乡真实的信用卡客户数据，提取出综合了客户个人信息和交易记录的37个特征变量，利用Logistic回归、SVM、CART模型、随机森林和XGBoost五种机器学习算法，相应地构建了五个数据挖掘模型来对该行信用卡客户的违约风险进行预测，并分析违约风险背后的驱动因素。通过模型的分析与比较，我们发现随机森林和XGBoost是最适合该银行的信用卡违约风险预警模型，两者在违约预测的准确率和效率上都具有明显的优势。而Logistic回归作为其中最简单的线性模型，虽然预测能力和效率都不及上述两种集成学习模型，但其对于违约风险影响因子的解释性最好，对银行后续的风险管理实践有很强的指导意义。

关键词：信用卡违约风险；机器学习；数据挖掘

**南京大学本科生毕业论文（设计、作品）英文摘要**

THESIS：Empirical Study on Credit Card Default Risk and Its Driven Factors Based on Data Mining

DEPARTMENT：School of Management and Engineering

SPECIALIZATION: Fintech

UNDERGRADUATE: Shi Yao

MENTOR: Prof. Li Xindan & Dr. Song Yuquan

ABSTRACT：With the rapid development of China's economy and residents' changing consumption concept, the credit card business of banks has been in a rapid growth stage in recent years. However, as a high-risk and high-yield business, credit card business puts forward a higher test of risk monitoring and management for banks. Based on real credit card data of a city commercial bank in China, this study extracts 37 characteristic variables which synthesize customer's personal information and transaction records. Five machine learning algorithms, Logistic regression, SVM, CART model, random forest and XGBoost, are used to construct different data mining models for predicting the default risk of the credit card customers and for analyzing the driving factors behind the default risk. Through the analysis and comparison of the five models, we find that Random Forest and XGBoost are the most suitable early warning models for the prediction of credit card default risk of this bank. Both of them have obvious advantages in the accuracy and efficiency of default prediction. Logistic regression, as the simplest model among all five models, has the best explanatory effect on default risk factors, although its predictive ability and efficiency are not as good as those of the two integrated learning models mentioned above. It provides a strong guidance for this bank’s risk management practice in the future.

KEY WORDS: Credit Card Default Risk; Machine Learning; Data Mining

目录

[1 绪论 1](#_Toc40827107)

[1.1 研究背景 1](#_Toc40827108)

[1.2 研究问题 4](#_Toc40827109)

[1.3 研究意义 4](#_Toc40827110)

[1.4 论文结构安排 5](#_Toc40827111)

[2 理论与方法 7](#_Toc40827112)

[2.1 文献综述 7](#_Toc40827113)

[2.1.1 国内研究现状 7](#_Toc40827114)

[2.1.2 国外研究现状 10](#_Toc40827115)

[2.2 实验逻辑 14](#_Toc40827116)

[2.2.1 整体实验结构 14](#_Toc40827117)

[2.2.2 基金经理的私人信号 15](#_Toc40827118)

[2.2.3 使用私人信号预测股价崩盘 15](#_Toc40827119)

[2.2.4 预测效果比较 16](#_Toc40827120)

[2.2 模型简述 17](#_Toc40827121)

[2.2.1 股票减持指标 18](#_Toc40827122)

[2.2.2 DUVOL指标 20](#_Toc40827123)

[2.2.3 K-折交叉验证 22](#_Toc40827124)

[2.2.4 Logistic回归 23](#_Toc40827125)

[2.2.5 SVR模型 25](#_Toc40827126)

[2.2.6 CART模型 25](#_Toc40827127)

[2.2.7 随机森林 28](#_Toc40827128)

[3 数据处理 30](#_Toc40827129)

[3.1数据来源与数据示例 30](#_Toc40827130)

[3.2数据预处理 33](#_Toc40827131)

[3.2.1 特殊值处理 33](#_Toc40827132)

[3.2.2 缺失值处理 34](#_Toc40827133)

[3.2.3 重复值处理 35](#_Toc40827134)

[3.2.4 数据标准化 35](#_Toc40827135)

[3.3 数据处理逻辑 36](#_Toc40827136)

[3.4 描述性统计 38](#_Toc40827137)

[4 实证分析 41](#_Toc40827138)

[4.1 模型评价方法 41](#_Toc40827139)

[4.1.1 混淆矩阵 41](#_Toc40827140)

[4.1.2 ROC曲线 、AUC值 42](#_Toc40827141)

[4.2 模型结果 42](#_Toc40827142)

[4.2.1 基于Logistic回归的信用卡风险预测模型 42](#_Toc40827143)

[4.2.2 基于SVM的信用卡风险预测模型 45](#_Toc40827144)

[4.2.3 基于CART模型的信用卡风险预测模型 45](#_Toc40827145)

[4.2.4 基于随机森林的信用卡风险预测模型 46](#_Toc40827146)

[4.2.5 基于XGBoost的信用卡风险预测模型 47](#_Toc40827147)

[4.3 模型比较 48](#_Toc40827148)

[5 研究结论 50](#_Toc40827149)

[6 研究展望 52](#_Toc40827150)

[参考文献 I](#_Toc40827151)

[致谢 IV](#_Toc40827152)

# 1 绪论

## 1.1 研究背景

改革开放以来，我国的金融市场高速发展，随着金融开放的程度不断加大，越来越多的企业选择上市。但是，于此同时，越来越多的股票发生了突然的崩盘，如何对此崩盘现象进行预警，如何对此进行实时的监控成了时代发展的关键问题。

而学术界和实务界都把证券市场的众多波动，尤其是巨幅波动看作是十分值得人们研究的重要的金融问题，李心丹，陈莹（2010）。而这些问题当中便包括股价崩盘，股价崩盘是一种非理性的金融现象。其背后的原因十分复杂，引发了国内外众多学者的讨论。无论是亚洲金融危机，还是互联网泡沫，还是2008年的美国金融危机，股价崩盘对于整个经济社会生活是有着巨大的破坏性影响的。它不仅会打击投资者的进场信心，导致整个社会的股权融资的难度提升。同时也会造成资产的错误定价，从而导致资源的错误配置，阻碍整个实体经济的平稳运行。

同时，历年的股价崩盘，背后都可以看到以下行为的影子。比方说，相对一致的投资行为，机构投资者的大幅度集中减持。相对于股价上涨，股价崩盘发生的速度很快，大多数的投资者没有足够的时间去应对。因而，管理者和投资者都迫切地需要预测股价崩盘。以实现对市场的监控。同时对危机做出有效的应对。

Graham （2004)指出，公司的管理者可能会因为自己的薪资激励、职业声誉等原因，会有意去隐瞒一些对于公司不利的私人信号。但是对于公司有利的信号，他们会选择及时披露。而这些对于公司不利的私人信号，会随着时间的推移不断的累积。当外界环境变化触发信息向整个市场公开是，便会导致对股价的冲击，在严重的时候便会导致股价崩盘。

另一方面，近年来机构投资者的一些行为对预测股价崩盘也逐渐收到越来越多的关注。随着信息时代的到来，众多投资者之间的交流也逐渐频繁起来。个人投资者会根据自己的社交关系网络来进行投资。例如Liang 和Guo（2015）根据 2011 年的中国家庭金融调查得出的数据，指出社会互动推动家庭股市参与。而机构投资者之间也会根据自己的社交关系网络中的私人信号来投资股票。同时作为自己增持减持股票的一个重要的依据。

Baker （2012）和 Stefan （2015）建立了投资者相互之间的“相互信任网络”模型，利用仿真，演化了投资者之间的投资行为，研究了股市当中的羊群行为探究了社交关系网络对投资行为的一定影响。

郭白滢（2019）也指出，机构投资者之间除了相互竞争只外，也会选择信息共享。这样的行为提高了市场的定价效率。因而对股价崩盘的风险也有着重要的影响。在相当多的情况下，选择信息共享，其实是机构投资者基于自身的利益，审慎思考后做出的决定，陈新春（2017）。此外Han 和 Yang（2016）也指出投资者之间的社交关系网络是金融市场私人信息传播的有效途径，同时也是主要途径。网络对信息传播有着决定性的作用。同时，很多机构投资者会通过分享信息，一方面他人对于该信息的反应，可以作为博弈中，检测该信息是否真实的一个手段；另一方面，如果他人同意该观点，Stein（2008）； Crawford（2017）指出，其他人也会选择提供有价值的信息作为回报；此外，其他的投资者如果认同该私人信号，并选择跟随，Gary(2012)指出，会增加该投资的期望收益；最后，通过共享信息，机构投资者扩大了可以选择的资产的范围，Gray(2008)证实，使得他们的投资组合可以足够分散化，进而实现降低投资组合短期波动的目的，最终可以稳定该机构投资者的资金流。

Black，1986; Kumar，2009; 李广川等，2009; 胡昌生和池阳春，2013指出，长久以来，散户为主的投资者结构是造成资本市场的剧烈股价波动的主要诱导性因素。为提升金融市场的稳定性，我国也在逐步增加机构投资者的数量。证监会在2001年就发文并开始实施了“超常规发展机构投资者”的策略。在此背景下，我们的基金行业深入发展。随着我国金融的繁荣程度不断加深，机构投资者也逐步增加。其中不乏公募基金的影子。随着时间的推移，公募基金在整个市场上的影响已经开始举足轻重。因而，研究公募基金的基金经理之间的社交关系网络，逐渐有了实际意义。同时，作为市场中的重要的私人信号的发掘者，公募基金的私人信号，对于预测某个将要崩盘的公司有着重要的意义。

近年来，我国经济快速发展，随着居民消费水平不断升级，消费升级观念日益深入，消费金融行业前景广阔，我国信用卡业务进入到爆发式增长阶段。信用卡不仅为客户和商户提供了收付款便利，也逐渐成为银行提供全面金融、提升盈利水平的重要工具。2017年，中国信用卡累计发卡量同比增长26%，为近年来最高增速。但信用卡业务在带来高收益的同时，银行所面临的信用卡风险和损失也日益增长。数据显示，2018年第一季度中国信用卡逾期未还金额已达711亿元，是2010年（76.89亿）同一时期的8倍；第二季度，我国信用卡逾期半年未偿信贷总额756.67亿元，环比增长6.35%，占信用卡应偿信贷余额的1.21%。信用卡业务已然成为了银行零售业务的重要组成部分，然而，如何提高银行的信用卡风险管理也成为一个严峻的挑战。目前，国内银行的信用卡风险管理机制尚不完善，面临巨大的的操作风险、欺诈风险和信用风险等，其中信用违约风险带来的损失占到银行信用卡业务风险损失的九成以上。因此，如何更好地监测和管理信用卡违约风险对于商业银行有着重要的意义。

在这个信息爆炸的时代，银行本身已成为一个巨大数据库，涵盖了无数的客户信息和交易记录，再加之计算机技术的迅速发展，为数据挖掘提供了条件，也为银行信用卡风险管理提供了新思路。风险管理的核心是依据现有信息来预判客户是否存在违约风险，这是一个典型的类别预测问题，而机器学习和数据挖掘技术为我们提供了很多成熟的解决分类问题的算法，同时打破了大规模数据存储与计算能力的桎梏，拓展了可研究数据的宽度和深度。因此，利用企业级的数据，结合恰当的大数据挖掘算法构建风险预测模型，可以在很大程度上提高风险预测的效果以及效率，从而提高发卡行的信用卡风险管理能力。如今，大数据技术在信用卡风险管理中已有颇多应用，如信用卡申请时的诈骗识别、欺诈交易实时检测、信用评分等。利用机器学习与数据挖掘技术构建自动化的银行风险管理体系，是一个必然的发展方向。

## 1.2 研究问题

本研究从投资组合的分布在不同城市的权重中推测基金经理在不同城市的私人信号的精确度，其次利用该精确度和对某个特定股票的减持幅度，来预测该股票的崩盘风险。

## 1.3 研究意义

本研究建立在我们公募基金的真实数据之上，并从应对股价崩盘的实际需求出发，结合计算机与金融工程的理论研究，利用机器学习和数据挖掘，来利用基金经理的隐藏信息和市场公开数据，预测某个特定股票的崩盘风险。

该实验实现了股价崩盘的预警，具有较强的实际意义。同时我们的实验结果还表明，一些特定的机器学习方法在预测股价崩盘，监督市场风险方面，具有较强的实际意义。

本模型不仅可以为个人投资者和机构投资者，提供减持、平仓预警。同时还可以为市场监管部分，如证监会等提供对某些股票即将崩盘的预警。对于多只在同一个行业的股票假如同时存在崩盘风险，市场管理者还可以利用本实验来对整个市场进行宏观调控。同时也为政府出台金融政策提供了有利的帮助。

此外，本研究还可以为公司本身所利用。一些公司在利用本框架监测到股价即将崩盘的之前，可以积极调整战略，控制公司资金风险、运营风险等。为公司的长远发展做出及时积极的调整和打下坚实的基础。

本研究建立在我国某城乡行真实完整的信用卡客户数据之上，并由该行的实际风险管理需求出发，结合金融与计算机理论研究，将机器学习和数据挖掘技术应用于综合了客户个人信息和历史交易情况的变量组合上，来预测其信用卡违约情况，实现了该行信用卡客户违约风险的预警，具有很强的现实意义。研究结果还表明，机器学习方法在信用卡数据挖掘问题中有良好的表现，除了提供违约概率和信用风险的准确度量之外，模型还可用于分析和比较客户信用卡违约拖欠的驱动因素，为银行后续的信用卡风险管理提供可靠的理论与数据支撑。

本研究结合现实数据，刻画了金融数据的特殊性与复杂性，探讨了机器学习在银行信用卡业务这一特定金融领域的应用。从结果上看，预警模型的准确率和实证结果的精确度再次印证了机器学习在金融领域应用的有效性，并通过实证过程中对多种模型的不断调整与优化，为机器学习在本领域的发展提供了针对现实场景改进的可能性，具有一定的理论意义。

## 1.4 论文结构安排

不断调整与优化，为机器学习在本领域的发展提供了针对现实场景改进的可能性，具有一定的理论意义。

第一章为绪论，具体介绍了本研究的相关背景、具体问题以及实际意义。

第二章刻画了本文的理论和方法。首先，本文从文献出发，提供具体的前人理论支撑。接着，本文介绍了实验的逻辑，阐述了如何实现具体的理论。最后，本文提供了实验所需要用到的模型，阐述了开展实验所需要的具体工具。

第三章描述了实证研究的数据来源及数据处理逻辑，同时展示了本研究所需要用到的实验数据。首先，本文阐明了数据的来源，标明了数据的格式和具体变量。接着，本文介绍了对于一些特殊数据和极端数据的预处理方式，比如一些空缺值的处理。其次，本文描述了数据处理的具体逻辑，提供了具体的复现逻辑。最后，本文展示了实验数据的描述性统计。

第四章是本文的实证分析的部分。首先介绍了模型的一些评价的方法。接着，给出了实验的具体结果。最后基于不同的模型，给出了评价和对比分析。

第五章是本文的研究结论，其总结了本文实证研究所得到的具体的结论

第六章是本文的研究展望部分，其综合了以上的实验结果，和国内外已有文献和已有研究，对本实验的未来研究方向给出了一定的展望。

第一章为绪论，介绍了研究的背景、问题及意义；第二章综述了国内外理论研究现状，并具体介绍了研究所采用的模型与方法，包括EM算法、SVM、CART模型、随机森林和XGBoost这五类机器学习方法；第三章对本实证研究所采用的数据集进行了描述，包括数据来源、预处理、变量说明与描述性统计；第四章是实证分析部分，首先介绍了本研究所采用的模型评价指标，之后给出了各个模型的结果并进行比较与分析；第五章总结了实证研究所得到的结论；第六章综合上述研究结果，给出未来的研究展望。

# 2 理论与方法

## 2.1 文献综述

### 2.1.1 国内研究现状

基金经理可以通过自己的努力获得很多信息，按照是否向市场公开，一般分为公有信息和私有信息。其中公有信息指的是投资者可以自由获取的信息。陈新春(2017)指出，私有信息包含两大部分，一方面是投资者自己的主观信息，另一方面是投资者的社会关系网络中流存的信息。

申宇(2013)指出了私有信息是影响机构投资者持仓行为的主要原因。国内研究肖欣荣（2012）指出，“在熊市和震荡市中，基金重仓股票仓位的变化与其基金的网络结构有着显著的正相关关系。” 而投资者的行为具有传染性。投资者一方面会从市场资产价格中学习信息，同时也会观察网络（注：社交关系网络，下同）中其他人的行为和结果，彼此之间相互交流。因而两个投资者的重仓股票如果十分相似，或者在某一段时间内，增持减持步调保持一致。这很有可能意味着，他们获得了类似甚至相同的信号。也有可能意味着他们之间存在着直接或者间接的关联甚至交流。申宇等（2013）认为基金具有信息优势，私有信息促进了基金的隐性交易，并提高了业绩。

肖欣荣（2012）基于Pareek（2011）也通过实验证实了基金经理的行为传染与来自基金经理网络中的私人信息有关。“基金经理会根据来自网络中的信息而进行交易”，同时也证实了网络结构中投资者的羊群效应。 Scharfstein和Stein，1990；Prendergast(1996); 罗真和张宗成2004证实，基金业会存在排行榜现象。假如某只基金的表现是落后于市场或者同行业的。会收到相关投资者和同行的质疑和责备的。这样的压力，也会导致基金经理放弃原先的投资规则，在投资行为上呈现某种跟风。呈现处一种同行博弈行为，例如羊群效应。刘京军，苏楚林(2016)通过空间计量，检验了基金现金流的网络溢出对基金自身业绩的影响。

张雪峰（2017）指出：“大多数基金投资者只寻求短期的收益，缺乏长期投资的耐心，同时也缺乏长期投资的风险承受能力。”同时基金公司施行的末尾淘汰的竞争制度，也会使得基金经理由于职业竞争的压力，而寻求短期利益。追逐市场热点。另外“基金经理处于自身职业发展以及自身的薪资的考虑，会去追求自身利益的最大化。在追求业绩排名和薪资收入的双重刺激下，很难进行长期投资”，这些都会导致，基金经理在得到一个关于股票价格崩盘或者抛售的具体信号时，会在短期将其抛售，而不会等到整个市场全部反应过来，价格已经触底时才展开抛售行为。因而，公募基金的大规模减持行为，一般会发生在股价完全崩盘之前。所以，对于基金经理的减持行为的观测，有着于研究股价崩盘。同时其私人信号的准确度，也于预测股价崩盘的准确与否息息相关。

对于预测崩盘，国内的研究，大多基于基金经理的网络关系和股价崩盘之间的线性关系。但是，较为明显的时，有必要将在考虑到线性关系的基础上，同时考虑非线性关系，以获得更为准确的预测结果。

在我国，早期的信用卡违约风险研究主要围绕信用评分展开。赵刚(2007)概述了信用卡的起源和定义，从其金融特性出发，介绍了国外较成熟的信用卡信用风险管理的方法和技术，分析了信用评分法在信用卡风险管理中的作用，并应用Logistic回归模型，对信用卡征信审批评分卡的使用进行了实证研究。但信用评分法存在的问题是主观因素影响较大，且只考虑信用卡客户的个人信息，而忽略其真实交易行为。

之后，不少国内学者开始关注信用卡违约数据，探索信用卡违约风险的主要影响因素。陈雄（2011）研究了信用卡违约风险的成因，结合我国信用卡违约现况，利用因子分析法，得出了个人信用卡违约的主要影响因素，以及各个影响因子的权重情况，并据此给出加强信用卡违约风险管理的对策建议。邵亦明（2013）提出了基于主成分分析的信用风险预警模型，变量选取自信用卡用户的个人信息和交易记录，从中找到主要成分，并得到各个主成分对应的权重，从而反映各变量对信用风险的影响程度，以便更好地预测和预警。盛洁（2014）利用客户行为数据建立信用评分卡，用Logistic回归方法构建违约概率的预测模型，并得出影响信用卡客户违约概率的影响因素。陈莹等（2017）利用来自商业银行信用卡账户的消费数据，归纳出可以用来表示信用卡账户行为特征的变量，通过单因素事件法和多元线性回归，研究其对客户违约风险的影响程度，并给出影响信用卡违约风险的主要变量。该研究对于银行的风险管理实操性较强，但作为违约风险监测系统而言预测性较差。上述这些方法仅对变量使用了简单的线性回归，无法刻画变量间可能存在的复杂的非线性关系，且变量选取较单一，模型在实时风险预测的准确性有待考证。

随着计算机技术的迅速发展，海量数据存储与运算成为现实，为银行风险管理提供了新思路。鲁长东（2017）指出，银行业传统的信用卡风险管理由于数据的存储条件有限以及计算力的限制，形式单一、模式固定，效率低下。但随着计算机和互联网的迅速发展，银行本身已成为一个巨大的数据库，涵盖了无数的客户信息和交易记录，为数据挖掘和模式识别提供了条件。大数据技术在信用卡风险管理中有颇多应用，如信用卡申请时的诈骗识别、欺诈交易实时检测、信用评分及违约风险预测等等。

国内学者纷纷开始探索机器学习及数据挖掘等技术在信用卡违约风险预测中的应用。陈为民（2009）引入了基于支持向量机的信用卡客户的信用评分模型，用于信用卡审批和后续管理中。基于客户消费和还款行为数据，通过数据挖掘技术，建立信用评分及违约预警模型。范巍强等（2011）利用 BP神经网络进行信用卡违约风险预测，实证结果表明，此模型在该数据集上的预测准确性较高，具有很强的泛化能力。但是，由于神经网络的“黑箱”性质，模型的解释性较差，对于违约风险的影响因子分析不够，无法指导银行更好地进行信用卡风险管理。张双全（2018）采用智能算法和数据分析技术对信用卡客户数据进行了分析研究，找到对信用卡客户违约情况影响较大的因素，建立了有效的分类模型。李涵峰（2018）比较了多种机器学习模型对银行信用卡违约的检测能力，分别研究了KNN、决策树、SVM、随机森林和LightGBM五种机器学习算法，以银行信用卡用户的个人信息、历史交易交易和还款记录为数据进行挖掘分析，比较得出LightGBM模型的预测准确性最高。但该研究存在的问题是，在特征选取上较为单一， 无法细致全面地刻画用户特征，且忽略了金融市场上外部因素对信用卡客户还款行为的影响，对于现实世界实时违约风险的预测能力有限。

### 2.1.2 国外研究现状

由于社交关系网络存在的信息传递的作用。一个信息通过投资者的社交关系网络，例如一种主流观点，或者市场上的一种策略，会在不同的机构投资者之间相互传播。因而重仓持有的相同的股票的基金经理之间会彼此存在相互联系，Pareek(2012)。

国外的学者Bushee & Goodman（2007）；Jiang（2010）指出：基金经理持有的较大仓位的特定股票，主要源于基金经理的私人信息。投资者之间的社会互动与其之间的信息交流也会对其投资行为产生影响。Shiller 和Pound（1986）证实了“基金经理倾向于预期持有相同股票的其他基金经理之间进行交流，并且其实际的投资行为也会受到他们的影响。” Hong、Kubik和Stein（2004）证实了投资者的投资决策是会显著收到其社会互动的影响的。

另一方面，社会网络会显著影响投资者的行为。Cohen、Frazzini 和 Malloy（2008）证明了基金经理更加倾向于投资他们教育网络中存在的公司。例如基金经理会更加愿意去投资公司高层于基金经理有着相同教育背景的公司，例如公司高层和基金经理毕业于同一个大学同一个专业。同时，。Cohen、Frazzini 和 Malloy（2008）也证实了上述依赖社交关系网络的投资行为会有着很好的投资收益。在另一个角度上看，这样的更好的收益，也会促使投资者最终的投资行为是和自己的社交关系网络紧密相关的，尤其是重仓股票。投资者甚至会为了获得更好的投资收益去获得一些与重仓股票紧密相关的社交关系网络关系。其持仓的大幅变化也会参考该社交关系网络。

这就导致了，即便两个基金经理原先与其重仓股票背后并没有社交关系网络关系。更好的投资收益，也会促使他们去与该股票相关的社交关系网络节点，建立社交关系。从而两位基金经理，最后很可能间接，甚至直接地建立了社交关系。

Hong(2005)的研究指出，基金经理的持仓决策非常容易收到同一个城市内的基金经理的影响。他们在许多情况下都会通过口口相传来传递关于股价的信息。投资者个体在这样的一个行为下，最终形成了一个投资者的信息网络。成为帮助私人信息传播的一个有效的工具。Colla 和 Mele( 2010)中也证实了，社交关系网络中联系“紧密”投资者的投资行为之间会存在显著正相关性。

我们需要从一些可以预测投资行为的信息中去预测基金经理的社交关系网络。比方说从基金经理所持有的重复重仓股票上，或者同一个城市的重仓股票的投资上Harrison Hong和Jiangmin Xu(2019)。尽管有一些社交软件，但是其中关于投资的信息还是一直都很少。在社交账号上的伙伴也未必是投资上的伙伴，也未必对分析他的投资策略有帮助。

Harrison Hong和Jiangmin Xu(2019)指出私人信号的精确度会随着他在一个城市的non-iid（非独立同分布，下用non-iid替代）的关联数的增加而增加。关联数也就是他在该城市的社交关系网络的大小。

这种关联数意味着，一个基金经理在该城市的人际关系模型并不是简单的Erdös and Rényi (1959) null 模型（表征普通两个随机路人之间的关联），在这个null模型中，人与人之间的关联是独立同分布的。加入我们观测到一些隐藏网络的参数估计是显著地不平均分布，那就意味着在这群人里面，有些管理人对于某个特定的城市的关系是non-iid的。

这种非独立同分布的关系，是怎么来的呢。可能是他的亲戚，或者某个朋友在那个城市。可能在那个城市工作。他们也可能是大学的校友，因为他们有的人在那个城市工作，或者知道某个人在那边工作。

举例来讲，一个在哈佛上学的人，可能在波士顿有着不成比例的联系。由于他们是同样的水平同样的专业，所以关系更加稳定，也更可能会对他们的未来决策起着关键的作用。

而在Harrison Hong和Jiangmin Xu(2019)中，投资者是面临着short-sales constraints （卖空约束）和 quadratic trading costs（二次交易成本）的，这就意味着，投资者并不能很随意的得到资金或者随便处置自己管理的基金，在面临着投资哪一个股票时需要认真考虑，同时也不能频繁地更改自己的投资头寸。这就意味着，投资者会尽可能地利用自己的信息优势去投资自己有着丰富而又精确的私人信号的股票。而short-sales constraints （卖空约束）和 quadratic trading costs（二次交易成本）的假设是合理的，因为Chen et al.（2002)和Gârleanu and Pedersen（2013）中证实了基金经理是被机构禁止shorting（卖空）的。并且较大的风险头寸会导致价格的波动，进而触及机构所设置的风险管理约束。

Harrison Hong和Jiangmin Xu(2019)证明，依靠基经理在不同城市的关系，可以得到很精确的私人信号，同时依赖此指标投资，可以获得更多的收益。这也催使基金经理更多地利用自己在不同城市的关系，去发掘更加精确的私人信号。

在研究基金经理在不同城市的投资组合的权重，可以发掘其在该城市的隐藏社交关系网络，同时利用其网络关系的大小，来刻画其在不同城市的私人信号的相对精确度。

Pareek(2012)利用了基金的重仓股票的网络密度，发现在社交关系网络中，私人信息广泛传递，基金经理有着较为明显的羊群效应，同时这样的现象与基金经理各自的投资风格无关。这从另一方面也意味着，持有相同的重仓股票的基金经理之间具有较强的相互认识，并时常相互沟通的可能性。

Ozsoylev、Walden、Yavuz和Bildik（2011）研究证实了，股票市场上的投资者网络中，信息会通过网络扩散，使得一些投资者在市场其他人交易之前有所行动。进而获得较高的收益。这也意味着有着交为精确的大量私人信号的投资者，一般会在信息在整个市场中大规模扩散之前有所行动。进而该基金经理的调仓行为在某种程度上可以预示，将要到来的股价的变化。比方说大规模减持所预示的股价崩盘。因而对于一个有着较为精确信号的社交关系网络节点，其大幅减持行为一般是先与股价崩盘的。我们在选取时间区间时，可以选取一个时间区间，例如一年，来同时研究减持情况与崩盘，而不需要一定选取两个时间区间。

而根据私人信号的基金经理的减持行为是先于整体市场的行为的。因而我们要考虑的是，关于股票利空的私人信号对于基金经理的行为的影响，而不是基金经理抛售行为在人群中传染所导致的后续一系列关于股票影响。

许多统计方法，包括判别分析，逻辑回归，贝叶斯分类器和最近邻，已被用于开发信用预测模型，比如Hand和Henley（1997）所提出的基于数据统计的信用评分模型。随着人工智能和机器学习的发展，人工神经网络和决策树也被用来预测信用风险，如Koh和Chan（2002）研究中所应用的数据挖掘方法，以及Thomas（2000）提出的基于行为评分的信用预测法。Lee等（2002）通过将反向传播神经网络与传统的判别分析方法相结合，探讨了其模型在信用评分上的表现。他们所提出的混合方法比传统的神经网络模型收敛得快得多，且信用评分准确性也有所增加，并且混合方法优于传统的判别分析和逻辑回归。

从风险控制的角度来看，估计违约概率比将客户分类为二元结果（违约和不违约）更有意义，但需要仔细考量的是，利用数据挖掘方法预测出的概率是否能真正有效地代表现实违约概率。I-Cheng Yeh和 Che-hui Lien（2009）提出了全新的“排序平滑方法”来估计违约的实际概率，研究了台湾信用卡客户违约支付的情况，并比较了六种数据挖掘方法中违约概率的预测准确性。

近期的文献研究，在已有模型的基础上，考虑了海量数据集和更大的特征空间，对于信用卡违约风险的预测更为高效准确。Khandani等（2010）应用机器学习技术来构建消费者信用风险的非线性非参数预测模型，将传统的信用因素（如债务与收入比率）与消费者银行交易相结合，通过分析消费者支出、储蓄和债务支付的模式，能够识别大规模数据集中普通信用违约模型（如Logistic回归、判别分析）难以检测的微妙非线性关系，极大地提高了模型的预测能力，并发现机器学习预测具有更强的适应性，能够了解不断变化的信贷周期以及违约率的绝对水平。在此研究的基础上，Butaru等（2016）进一步扩大了特征变量空间，在考虑个人信用因素和消费情况之外，还加入了宏观经济变量，这种做法与Bellotti等（2013）提出的用卡借款人违约的离散时间生存模型相似，结果显示，包含这些行为和宏观经济变量的动态模型在数据拟合方面在统计上有显著改进，当应用于样本外数据集时，取得了更好的违约预测效果。此外，Butaru等（2016）还扩大了实证研究的数据集，综合了来自六家银行的截面数据，而不是单一银行的数据。他们的研究结果表明，不同银行的风险差异较大，模型预测结果存在优劣性的差异，在风险管理上也应有不同的侧重。

## 2.2 实验逻辑

### 2.2.1 整体实验结构

本研究将利用基金经理在不同城市的私人信号的精确度，以及在不同城市减持的情况来预测股价本盘。总体上本研究分为以下部分的实验。

首先是通过基金经理所持有的股票，利用股票所在相同城市和股票持有权重，来得到不同基金经理之间的关联程度。并通过标准化处理，用关联程度衍生的指标来刻画不同基金经理在不同城市的私人信号的精确度。

其次，由于本实验是检测基金经理的社交关系是否能很好地反馈股价崩盘。实验的本质目的是探究一个变量对另一个变量的相关性。同时给出一些刻画相关性的模型。因而实验并没有对股价崩盘的参考因素，做出时效性的考量。本文设置的时间界面的长度为一年。来检测指标之间的相关性。

此外，本文通过整体市场的收益率和单只股票的收益率之间关系，联系一只股票在一年内的收益率高于平均收益率的天数和低于平均收益率的天数，得到股价崩盘指标。

最后，本文对于研究基金经理的网络指标和股价崩盘指标之间的关系，将会通过机器学习的方法，把线性和非线性关系同时纳入其中。使得本研究结果有着较强的实际应用价值。

### 2.2.2 基金经理的私人信号

在本文中，我们利用基金经理的私人信号精确度和股票本身的性质来预测股价崩盘。

我们对基金经理在某个城市的关联数进行标准化，来表征其私人信号的准确度。

### 2.2.3 使用私人信号预测股价崩盘

在本文中，我们利用基金经理的私人信号精确度和股票本身的性质来预测股价崩盘。

首先是该股票的基本信息，诸如振幅，区间换手率，非系统风险，波动率等。用于预测该股价是否会崩盘。本实验将基金经理的数据进行分类，一部分用于训练模型，而另一部分用于测试模型。

我们在前人研究的基础上加了了基金经理对于一只股票是否会崩盘，以及崩盘程度的判断。

如果实验选取的是整个市场的基金经理，那么他们的私人信号的精确度，对于整个市场是有效的。同样的，选取的是特定市场的基金经理，那么其私人信号的相对大小对该特定市场有效。

本实验根据Wind基金分类，选取了中国公募基金市场作为研究。同时排除了被动跟踪的的ETF基金，因为该类基金调仓的依据是跟踪具体指数，而不是在卖空约束和二次交易成本下的依赖自己的私人信号和择时判断。由于选取的是公募基金市场，因而每个基金经理的私人信号的准确度大小，在该市场内是有效的。本文用于预测崩盘的股票也选取了公募基金池中的股票。

基金经理根据自己私人信号的相对准确度，减持一定量的股票，对于预测股价崩盘是有着内在联系的。在大幅减持一只特定股票的情况下，一个基金经理的私人信号的相对精确度越大，说明他越有把握该股票会有着崩盘风险。在减持一只特定股票的情况下，减持幅度越大，说明基金经理对于该股票将要崩盘的倾向越大。

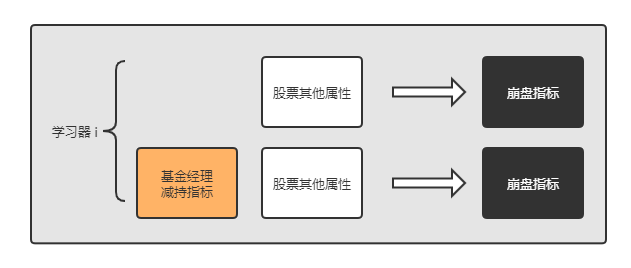
此外， 我们采用了对照组实验，比较在预测回归任务中，加入了基金经理崩盘指标和不加入，对回归的损失函数的影响。从而来反映，该崩盘指标对于实际预测股价崩盘的效果。

### 2.2.4 预测效果比较

本实验中，通过横向比较4个学习器，来确定那个学习器更加适合在实际应用中预测股价崩盘。

其中，股票的其他属性为本实验的控制变量，对于控制变量的选取，一定程度上参考借鉴了郭白滢（2019）。

控制变量的指标见数据处理，控制变量的示例见描述性统计。



不同学习器的对照实验示意图

## 2.2 模型简述

本研究采用的机器学习，机器学习为一门学科，机器学习利用研究如何通过计算的办法，使用已有的经验来改善整个系统自身的性能。而在计算机系统中，经验一般是以数据的形式存在的。机器学习的主要任务，就是在计算机的数据中产生“模型”，这样一个算法，也就是我们常说的“学习算法”。通过学习算法，我们可以把经验数据提供给计算机，让计算机执行学习算法，这样他就可以通过数据来产生新的模型。当这个算法在面对新的数据，也就是新的情况的时候。模型可以为我们提供一个基于学习算法的对应判断。如果说计算机学科是研究算法的，那么机器学习就是研究学习算法的。

机器学习的训练集为，输入数据，让模型通过计算得到模型具体参数的数据集。而机器学习的测试集为，输入数据，通过已经计算好的模型，得到预测结果，与现有结果进行比较，从而测试机器学习效果的数据集。

本研究的学习器均为有监督学习，通过输入不同的属性以及标签，让学习器学习，在输入未知的属性的情况下，如何输出一个正确的标签值。对于学习器的训练集而言，如果同时存在输入属性与输出标签，我们称该学习任务为有监督学习，典型的有监督学习有分类和回归。对于学习器的训练集，如果只存在输入属性而不存在输出标签，我们称之为无监督学习，典型的无监督学习有聚类等。

对于输出的标签值为离散数据时，我们称之为分类问题；对于输出的标签值为连续数据时，我们称之为回归问题。

监督学习的难点在于处理不属于训练集的数据。当输入的数据集，是学习器从未见过的时候，如何得到正确的预测结果。此问题，即如何提高模型泛性。归纳与演绎是科学研究与科学推理的两大基本手段。前者就是是从“特殊”到“一般”的“泛化”过程。即从具体事实到归纳处一般规律。

本研究中，我们的机器学习模型将采用逻辑回归（Logistic回归）、支持向量回归（SVR模型）、分类回归树（Classification And Regression Tree模型）和随机森林（Random Forest）。

本研究分实验对照组，一组为：将股票的特征属性作为输入属性，将股价崩盘指标作为输出标签。另一组为：将基金经理对于该股票的减持幅度和股票的特征属性作为输入属性，将股价的崩盘标签作为输出标签。对于不同的学习器，将

机器学习指的是一组专门用于解决大数据集中计算密集型模式识别问题的算法，具体方法包括支持向量机、基于树的分类器和各类神经网络等等。机器学习方法非常适合于本课题信用卡违约风险的预测与分析，因为其中涉及到较大的样本量和客户行为的高度复杂性。

本研究将采用监督学习的方法，向学习器呈现来自过去数据的输入/输出对。其中输入数据表示用于预测输出值的预先识别的特征属性，这样的输入数据通常表示为矢量，并且根据学习算法，可以包括具有或不具有缺失数据的连续和/或离散值。当输出是连续的时，监督学习问题被称为“回归”问题；当输出是离散的时，被称为“分类”问题。一旦提供了这样的输入和输出数据，学习器的任务就是找到一个能够正确地将这些输入向量映射到输出值的函数。完成此任务的一种方法是“记住”输入/输出对的所有先前值，虽然这可以正确映射训练数据集中的输入/输出对，但如果输入值与训练数据集中的输入值不同，或者训练数据集包含噪声，则不太可能成功在训练集外预测输出值。因此，监督学习的挑战是找到一个超出训练集的泛函，以便所得函数也能准确地将样本外输入映射到样本外结果。

本研究将使用Logistic回归、SVM、CART模型、随机森林、XGBoost这五类方法，以信用卡客户的个人信息和历史交易作为输入特征，次月违约情况作为输出，利用监督学习的方式来训练模型，进而得到违约风险预测模型，用来预测信用卡客户的违约情况。

### 2.2.1 股票减持指标

第一，在本文中，我们通过基金经理之间持有的相同的股票，来预测该基金经理在该城市的关联数。

在【Infer】中指出，如果基金经理受到卖空约束和二次交易成本，那么基金经理，会更加愿意持有自己有着相对更加精确的私人信号的城市的股票。而这种精确的私人信号，在【infer】中指出，是由于其有着相对于其他城市的更多的关联，【infer】中给出的例子是，大学同学在这个城市工作。所以与投资关联很大的私人关系的数量，可以在某种程度上很好地表征，该基金经理在该城市的私人信号的精确度。【infer】中通过实验证明，基金经理之间大量持有的相同的城市的相同的股票，在某种程度上意味着他们很可能直接相互认识，或者通在该城市之间的某个中间人相互认识。两个基金经理通过持有的重仓股票之间的关联，可以作为描述他们与投资相关联的私人关系的指标。本文中，将采用【infer】中给出的指标，用来刻画一个基金经理 在城市 的私人信号的精确度 。从而得到所有基金经理在所有城市之间的私人信号的精确度集合

对于，其具体计算逻辑如下：

本算法是一个三层循环算法

1. 首先对于一个基金经理i：
2. 在某个城市k:
3. 对于在城市k的其他基金经理l：
4. 通过基金经理i在该城市k的所有投资组合权重 与 基金经理i在该城市k的所有投资组合权重，比较两者之间的较小值（如果有一个权重为0，则返回0，意味着通过该股票，无法为两基金经理之间建立关联）求和后得到相对权重
5. 将（4）中得到的权重求和，即可得到基金经理i在城市k的关联数
6. 将关联数，对于整个集合标准化，即可得到基金经理的私人信号精确度

第二，我们选取了基金经理在过去一年中，减持幅度最大的股票作为研究对象。

对于该股票，我们记减持份数为 ，其对应的股价为 ，而该基金经理管理的基金总份额为 ，则其减持幅度为：



第三，本文结合基金经理的信息准确度 ，对于基金经理减持的最大幅度股票k，得到信息精确度 。一支股票的所有基金经理的减持幅度的加总，并进行标准化，则可以得到该股票的减持力度。

可以得到其对于股票k的减持力度：



线性回归是一般量化分析问题中常用的一种统计方法，对于分析各解释变量的影响权重有直观的表述，但由于线性回归的因变量通常是连续的数值，对于一些常见的分类问题并不直接适用。Logistic回归可以被认为是线性回归模型的一个特例，它的输出值不再是一个连续的数值，而是通过一个sigmoid函数，被转化为0-1之间的一个接近0或1的y值，因此可以很好地应用于二分类问题，其回归式如下：

但是，二元因变量违反了一般回归模型的正态性假设。Logistic回归模型表明，事件拟合概率是解释变量的观测值的线性函数。这种方法的主要优点是它可以产生一个简单的分类概率公式，缺点是无法正确处理解释变量的非线性关系和交互效应问题。

### 2.2.2 DUVOL指标

DUVOL指标为股价崩盘指标。国内国外研究者对于股价崩盘的指标都有着一定的研究，Hong and Stein(2003)指出了刻画股价崩盘的标准：

1. 股价反常的剧烈波动
2. 负向的价格变动
3. 崩盘的大规模传染

在数值方法刻画上，国内外也逐渐有了较为公认的指标。其中以收益上下波动比率的应用最为广泛。崩盘指标来自Chen(2001).作为本研究采用股价崩盘指标。

公司股票收益率的上下波动的比率的计算方法如下：

首先，通过公式（1）来消除市场整体收益率的影响。其中  为股票 i 在 t 日的收益率；  为整个市场在 t 日的平均收益率；残差项表示整个该股票的收益率当中，不能被市场的收益率所解释的部分。



接着，我们定义



为该股票该日的特有收益率。

最后，定义一个股票的收益率的上下波动比率。设为 DUVOL , 其具体实现如下：



对于上述模型，其中 为股票在第i日的回报率低于整年的总回报率的天数。

总体而言，DUVOL 的值越大，该股票收益左偏的程度越大，股价的崩盘风险越高。

如果实验选取的是整个市场的基金经理，那么他们的私人信号的精确度，对于整个市场是有效的。

### 2.2.3 K-折交叉验证

K-折交叉验证模型（k – fold cross validation），有被称作循环估计。是一种可以充分利用数据集对算法测试作用的模型。

K-折交叉验证的主要目的在于，减少过拟合和选择偏差的问题，使得整个模型可以在一个新的独立的数据集上实现通用化。

K-折交叉验证模型首先将原数据集划分为测试集和训练集，接着将原来的数据集划分为相等的K部分，也就是K折。然后，按顺序依次将数据的第i 部分作为测试集，其余的作为训练集。i从1至K，一共循环K次，每次训练模型后，测试模型训练结果。利用测试集获得训练的准确度。记录每次训练的准确度，将平均准确度作为模型的最终准确度。举例如下（五折交叉验证模型）

K-折交叉验证模型，主要有两个作用。首先，可以充分利用实验数据，实验的训练集相当于增加了K倍。在实验数据较为紧张的情况下，K-折交叉验证模型可以较好的实现数据的充分利用。其次，K-折交叉验证模型可以更好地评估算法性能的好坏。可以防止由于选择训练集和测试集有误，而导致地无法较为准确地反应算法的性能。



5-fold cross-validation

### 2.2.4 Logistic回归

Logistic回归，即logistic regression，逻辑回归，又被称作为对数几率回归。是回归分析里面常用的一种方法。同时它有着可以直观解释变量的优势。

逻辑回归的函数形式为：



XGBoost算法也是一种集成学习方法，它的思想是将许多弱分类器集成在一起形成一个强分类器。XGBoost是一种提升树模型，将许多树模型集成在一起形成一个很强的分类器，其中用到的基础树模型是CART模型。但与随机森林算法不同的是，XGBoost不是用Bagging策略来集成，而是采用Boosting策略，算法生成的多棵树之间是相关联的，而不是像随机森林那样生成的树之间彼此独立。该算法思想就是不断地添加树，不断地进行特征分裂来生长一棵树，每次添加一个树，其实是学习一个新函数，去拟合上次预测的残差。当训练完成得到k棵树的模型时，要预测一个新的样本，其实就是根据这个样本的特征，在每棵树上落入一个对应的叶节点，每个叶节点又对应一个分数，最后只需要将每棵树对应的分数加起来就是该样本的预测值。

### 2.2.5 SVR模型

SVR模型，即Support Vector Regression模型。又被称作为支持向量回归模型。其最初是SVM模型，为一种分类模型，随着时间的推移，逐步被改进，进而使用在回归任务上。为一种监督式的学习模型。

对于一个回归问题，给定一个模型的训练样本【】，目标是得到一个【】的回归模型。

那么问题可以转化为：

【svr】

其中

经过计算，SVR的解为【】

带入核函数，则最终的SVR可以表示为：

【】

### 2.2.6 CART模型

CART模型，即Classification And Regression Tree模型，又被称作为分类回归树模型。

CART模型可以检测到输入输出变量之间的非线性关系。使得在一些即便在OLS检测下，线性无关的变量，也可以在预测问题中发挥很大的作用。

此外，CART模型，由于采用了决策树算法，相比深度学习模型，比如ResNet还具有较好的可解释性，这也使得决策者可以打开模型“黑箱”，进一步去观测事件发生的原因。也可以利用该实验的最终结果，分析决策树的各个分支，来实现对模型的理解和调控，以实现控制的目的。

其次，CART模型，在多属性问题中，具有较好的表现。这就意味着，在满足模型的数据量的基础上，决策者可以在模型中加入其他次重要的属性，来实现更为精准的预测效果。这对于本实验而言，是有着重要的意义的。

CART模型的算法具体流程为：

输入：训练集，基尼系数阈值，样本数阈值。

输出：CART决策树

算法采用递归，从CART决策树的根节点开始

1. 对于当前节点对应的数据集，若样本数小于阈值，则返回决策树，当前节点递归停止。
2. 计算样本数据集的基尼系数，如果基尼系数小于阈值，那么就返回CART决策树，当前节点递归停止。
3. 计算当前节点各特征对应的基尼系数。
4. 对于计算得出的众多基尼系数，选取基尼系数最小的特征。并根据该特征将数据集划分为两个部分。
5. 对左右两个节点递归调用以上四个步骤，生成CART决策树。

对于上述算法中的基尼系数，其定义为：



本实验采用的是CART回归树模型。

模型的表达式为：



其中真实值与模型的输出值之间的误差可以表示为：



目标是上述的误差最小。

上述的模型的具体算法步骤可以表示为：

输入：训练集D

输出：回归树

1. 首先选择最优切分点，求解：



1. 用上述的求解结果对数据进行划分，并决定其相对应的输出：





1. 对上述的子集继续迭代步骤（1）以及步骤（2），直到满足（1）最小。
2. 生成决策树：



CART模型（Classification And Regression Tree，分类回归树）日益普及的原因在于它克服了标准模型的局限性，即因变量被迫在整个输入空间中拟合单个线性模型，例如上文提到的Logistic回归模型。相比之下，CART能够检测输入变量之间的非线性相互作用，这极大地增加了可捕获的关系类型和可以使用的独立变量的数量。此外，CART模型可以产生易于解释的决策规则，并将其逻辑清晰地布置在树中，这一优势使得它尤其适用于银行业的应用，因为在银行等领域的应用中，“黑盒”模型总是被怀疑。

CART模型可以轻松应用于高维特征空间的问题。假设有N个因变量及其对应的D维特征向量，通过递归地选择能最小化残差平方和误差的特征和参数 来估计训练数据集上的CART模型。此外，模型还必须设定一个“剪纸标准”来阻止树的扩展，以避免过度拟合训练数据。最常用的修剪措施之一是基尼系数，定义如下：

，

其中指的是CART模型的叶节点，指的是在叶节点处分配给第k类的训练样本的比例。CART模型T的剪枝标准定义如下：

，

其中指的是CART模型T中的叶节点数量，指的是通过交叉验证选择的正则化参数。一旦剪枝标准达到最小值，CART算法将停止扩展树。

### 2.2.7 随机森林

随机森林模型，即Random Forest模型，是一种典型的Bagging算法。同时也是Bagging算法的进一步延申与拓展。

首先，对于Bagging算法。在集成学习中，如果想要得到更强的泛化能力。那么集成学习中的个体应该尽量相互独立，但是独立这件事在现实当中很难实现。为了解决这个问题，我们设法让基学习器有着尽可能大的差异。但是为了集成的效果，我们同时也需要学习器的不能太差，如果采样的过程中，每个子集都是完全不同的，那么每个基学习器就只能使用到一小部分的数据，很多甚至都无法学习，也就无法保证学习的效果了。

为了解决这个问题，我们采用了Bagging算法。它基于自助采样法。给定数据集，我们随机采样一个样本放入采样集。并放回，进行下一次采样。多次采样之后，得到多个样本的采样集。重复上述的步骤，我们可得到多个采样集，然后通过每个采样机来训练对应的基学习器。最后再将这些基学习器相结合，就可以得到最终的学习器。

随机森林算法，建立在Bagging算法的基础上。再决策树的训练的过程中，我们加入了随机属性选择的过程。相较于传统的决策树，在选择划分属性的时候是在当前的集合中经过计算得到一个最优的属性。在随机森林中，对于整个学习的每个基学习器，是一个基决策树，对于这个树的每一个节点，我们都先从该节点的所有的属性中随机选取一个包含k个特征属性的一个属性子集。接着再用这个子集，从中选取一个最好的属性进行划分。这里的k是用于控制随机性的引入的。假如k等于原来属性集合的元素总个数，那么这就是传统的决策树算法。

具体的实现为：

随机森林计算开销较小，正好适用于本实验。随机森林中基学习器具有很强的多样性，一方面，这样的多样性来自于样本扰动，另一方面，这样的多样性来自于属性的扰动。这样就可以进一步提升模型的泛化能力。

对于随机森林的收敛性，它其实和Bagging算法很相似。虽然通过引入属性的扰动，个体学习器的能力在降低，但是随着学习器数量的增加，整体算法的泛化能力在增加，随即森林会收敛到更低的误差，具有较强的泛化能力。

随机森林是用于分类或回归任务的集成学习方法，模型在训练时会构建多棵决策树，最终输出占多数的类别（分类问题）或平均预测值（回归问题），其优势是减弱了决策树模型过度拟合训练集的问题。

随机森林模型大大提高了作为基础学习器的决策树模型的性能，其模型训练过程中体现了两个重要的思想。第一个思想是装袋（Bagging），又被称作自举聚合（bootstrap aggregation）。Bagging策略不再是传统地在训练集上学习得到单个决策树，而是采用T次重新采样的训练集，并在每个自举样本训练集上学习新的决策树模型。然后，分类模型允许所有这些T个决策树对分类进行投票，使用多数投票来决定预测的类别。Bagging策略的最大好处是它大大减少了决策树模型的方差，并且能够改善样本外数据分类的性能。随机森林的第二个关键思想是，通过人为地限制每个递归分割节点所考虑的特征集，来进一步减少每棵树之间的相关性。在学习每棵树时，当考虑每个递归分割时，模型随机选择一些特征子集（对于分类任务，通常是特征总数的平方根），并且在训练中仅考虑这些特征。在过去的研究和应用中，随机森林在许多样本外的分类基准测试中取得了巨大的成功，被认为是当今可用于一般任务的最佳的“开箱即用”的学习算法。

# 3 数据处理

## 3.1数据来源与数据示例

本研究的实验数据来源于Wind数据库和优矿数据库。

|  |  |  |  |
| --- | --- | --- | --- |
| 数据表 | 名称 | 释义 | 算法 |
| reduceShare | 重仓证券持仓变动 | 报告期末投资组合中重仓证券的持仓变动 | 报告期末重仓证券的持仓数量-上个季度末重仓证券的持仓数量 |
|  |  |  |  |

数据示例：

reduceShare表：

|  |  |
| --- | --- |
| 证券代码 | 673090.OF |
| 证券简称 | 西部利得个股精选 |
| 任职期限最长的现任基金经理 [名次] 第1名 | 刘荟 |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第1名 | 601318.SH |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第1名 [单位] 份 | 4,300.0000 |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第2名 | 600519.SH |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第2名 [单位] 份 | 200.0000 |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第3名 | 600036.SH |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第3名 [单位] 份 | 4,400.0000 |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第4名 | 600276.SH |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第4名 [单位] 份 | 1,200.0000 |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第5名 | 600030.SH |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第5名 [单位] 份 | 3,300.0000 |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第6名 | 600887.SH |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第6名 [单位] 份 | 2,600.0000 |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第7名 | 601601.SH |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第7名 [单位] 份 | 1,300.0000 |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第8名 | 601328.SH |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第8名 [单位] 份 | 12,700.0000 |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第9名 | 600016.SH |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第9名 [单位] 份 | 11,600.0000 |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第10名 | 600000.SH |
| 重仓证券Wind代码 [报告期]去年年报 [名次] 第10名 [单位] 份 | 5,100.0000 |

该实证研究所使用的信用卡客户数据来自国内某城商行，最初样本数据涵盖2018年8月至2019年1月六个月的时间内该银行各分行所有客户信用卡基本信息和交易记录，包含账户数量超过20万个。

首先，本研究只考虑个人信用卡客户的违约风险及其影响因素，故在实证数据集中，事先剔除了非个人持有的信用卡账户，不考虑那些由国家或城市、特定机构、企业等为主体的所有借款人。

其次，本研究剔除了信用卡使用频率低于特定标准的客户（连续三个月没有交易的账户）。因为本研究所采用的数据来源于单一银行，而同一客户有可能在不同的银行同时持有信用卡，我们无法得知用户是否在其他银行持有信用卡及其使用情况。在这种情况下，使用频率低的信用卡账户对于用户行为的描述和预测意义不大，甚至可能会造成负面影响，因此我们只保留了本银行信用卡使用频率较高的用户作为样本。

另外，本研究剔除了在本银行持有两张或两张以上信用卡的客户，除了同一客户在多家银行持有信用卡的情况，同一客户还有可能在同一家银行持有多张信用卡，这时常常会出现该客户的某张信用卡已有违约，而另外的信用卡记录显示没有违约。为了实证研究结果的准确性及说服力，在实证数据集中删除了这类同时持有两张及以上的该银行信用卡账户的客户。

最终，经过上述三个步骤，本实证研究的数据集最终包含18921个有效样本，每一个样本对应于该银行的一位信用卡客户及其持有的唯一的信用卡账户，每条样本数据包含了该信用卡客户的基本信息（用户特征）、交易记录及其月度违约（即信用卡逾期）情况。

## 3.2数据预处理

### 3.2.1 特殊值处理

在本实验中，有许多股票在从wind的数据库中导出时，并没有标明所在的城市。我们查询了其总部所在地，手动标明了其所在城市。

在本实证研究的数据集中，信用卡的交易记录及月度违约情况不存在缺失，数据全部来自于银行内部交易系统的准确记录，而客户的基本信息，由于需要信用卡持卡人在申请或办理特定业务时自行完善个人信息，所以在样本数据中存在不同程度的缺失。本研究将个人信息（用户特征）的缺失值，默认设置为该项的最低值，代表评价最差。因为当某客户的个人信息填写不完整时，我们有理由认为他/她是一个不太负责任的客户，其潜在的信用风险相应也会更大。

### 3.2.2 缺失值处理

对于reduceShare表中，有些股票在公示阶段并没有给出减持幅度。本文的解决方案是，只对持仓变化公示的股票进行排序，选取其中减持幅度最大的股票。

例如，对于0005.HK，实验开展过程中，并未在数据库中得到其总部城市地址，经过查询，手动设置其总部为“香港特别行政区”。

考虑到原样本中违约样本量过少（约占0.7%），在实际研究中我们将使用SMOTE（Synthetic Minority Oversampling Technique）算法，即通过人工合成少数类样本来进行过采样。传统的过采样方法思路很简单，仅仅是通过复制已有的少数类样本来增加其数量，从而实现样本平衡，但简单的复制很容易产生模型过度拟合的问题，因为本质上少数类样本所包含的信息依旧很少。因此，用这种传统的过采样方法训练出的模型，在样本外测试中往往表现很差，得到的模型过于特别而不够泛化。而SMOTE是基于这种简单过采样方法的一种改进，基本思想是通过对少数类样本所包含的变量信息进行分析，来人工合成新的少数类样本并添加到原数据集中，这样一来，不仅是少数类样本的数量增加了，而且其所包含的信息也成倍增加。SMOTE算法的具体实施步骤如下：

1. 对于少数类中每一个样本，计算它到所有其他少数类样本的欧式距离，比较得到其近邻；
2. 根据数据集中两类样本不平衡的比例设置一个采样倍率，对于每一个少数类样本，从其近邻中随机选择个邻近样本；
3. 将选出的邻近样本与原样本按如下的公式合成新的少数类样本：

.

通过SMOTE算法，以原有的违约样本为基础，我们人工合成了50倍的新违约样本，这样所形成的新数据集中，违约与未违约样本的比例约为1:3，对于机器学习模型的训练效果有很大的提升。

这里要特别说明的是，SMOTE算法只能运用于训练集的样本平衡，而不能对整个数据集做SMOTE，否则将产生模型过度拟合数据集的问题，即在样本外数据集上的表现很差。在实证过程中的具体做法是，先将数据集随机切分成训练集和测试集，之后仅在训练集上使用SMOTE算法，使得训练集中的两类样本相对平衡，然后再用新合成的训练集对模型进行训练，最后用初始切分得到的测试集对模型预测能力进行测试。

### 3.2.3 重复值处理

对于基金而言，很有可能会存在同一个基金前缀，但是后缀是A，B，C的情况、或者不同基金名称，在此情况下，基金经理是相同的。对于本实验而言，要研究基金经理在一个特定城市的信息的精确度。因而把他看作是不同的基金是不合理的。有两种解决办法：（1）将其合并成一个基金，作为一个基金处理。（2）只选取一个基金对其进行研究。本文采取的是第二种解决办法。本文对一个基金经理名下，不同的基金，只选取其中一个具有代表性的进行研究。

### 3.2.4 数据标准化

本研究，进行数据标准化处理，将实验数据均归于一个指定区间。这样可以去除数据绝对大小对实验的影响，同时去除不同单位的影响。在本研究中，对基金经理的减持指标进行了标准化处理，使之全部都介于0到1之间。

通过标准化处理，我们可以很好地消除原来的数据中单位和数量级的对模型预测能力的影响。使得模型的预测精确度提升。

数据的标准化是将数据按照一定的比例缩放至某一固定区间，去除不同类型数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标进行比较。在本实证研究的过程中，我们先把每一个变量都转化成均值为0，方差为1的标准化数值，再将其作为机器学习模型输入，从而消除原始数据的单位或量级对模型结果造成的影响。

## 3.3 数据处理逻辑

由于实验的样本量较小，因而本文选取了十折交叉验证。

对于所有的基金经理，我们选取了减持最大的股票的代码，以及其所对应的减持份额，通过其减持份额乘以其平均价格，在对应其基金净值，得到了减持幅度属性。最后生成了[基金经理，减持最大股票代码，减持幅度]数据表。

对于市场收益率指标，我们选取了两个市场的指数收益率作为代替，A股市场选取万得全A，香港市场选取恒生综指：

表一 指数信息表

|  |  |
| --- | --- |
| 指数名称 | 指数代码 |
| 恒生综指 | HSCI.HI |
| 万得全A | 881001.WI |

特征变量的选取是本实证研究中的一个重要内容，因为机器学习模型的表现与输入特征变量的选取有着极其密切的联系，因此特征变量的选取对于实证结果有着巨大的影响。一方面，我们想尽可能多地选取不同维度的变量，来更为全面地刻画银行信用卡客户的特征；另一方面，特征变量的数量也不宜过多，以避免特征空间维数过高带来的灾难（The Curse of Dimension）。

在选择特征变量时，本研究从两个维度出发，一是基于客户个人信息的用户特征，二是基于客户近期信用卡使用情况的历史交易。在用户特征方面，首先考虑信用卡客户的基础信息，包括年龄、性别、户籍、婚姻、受教育情况等；其次考虑客户的职业信息，包括所在行业、单位、职位职称、职业资格、从业稳定性等；接下来是客户的经济状况，包括汽车、住房状况、个人年收入、家庭年收入、抚养子女及赡养父母费用支出、负债比率和还款承受力，以及客户在该银行的办理其他金融业务情况，包括存款、贷款、购买国债、保险、投资金融产品、信用卡持有情况、借贷记卡消费情况、近一年金融资产状况、使用我行代理业务产品情况与本行业务关系历史等；最后是客户的信用信息，包括人行征信系统分值、品质状况、最近不良记录、公检法及仲裁机关个人不良记录和公用事业付费不良记录等。在历史交易方面，我们考察客户过去的信用卡违约情况，即过去每一期的月度违约金额。综合上述特征，整理归纳为如表一所示的变量描述：

表一 信用卡客户特征变量描述

|  |  |  |  |
| --- | --- | --- | --- |
| 输入变量 | | | |
| 用户特征 | | | |
| 变量名 | 解释 | 变量名 | 解释 |
|  | 年龄分值 |  | 个人年收入分值 |
|  | 性别分值 |  | 家庭年收入分值 |
|  | 户籍分值 |  | 抚养子女费用支出分值 |
|  | 健康状况分值 |  | 赡养父母费用支出分值 |
|  | 品质分值 |  | 负债比率分值 |
|  | 受教育程度分值 |  | 还款承受力分值 |
|  | 婚姻状况分值 |  | 人行征信系统情况分值 |
|  | 行业单位职位分值 |  | 最近不良记录距今分值 |
|  | 职位或职称职业资格分值 |  | 信用卡持有情况分值 |
|  | 从业稳定性分值 |  | 与本行业务关系历史分值 |
|  | 住房状况分值 |  | 我行借贷记卡消费情况分值 |
|  | 汽车状况分支 |  | 近一年在我行金融资产状况分值 |
|  | 存款状况分值 |  | 在我行贷款情况分值 |
|  | 购买国债情况分值 |  | 使用我行代理业务产品情况分值 |
|  | 保险状况分值 |  | 公检法及仲裁机关个人不良记录分值 |
|  | 投资金融产品分值 |  | 公用事业付费不良记录分值 |
| 历史交易 | | | |
|  | | 历史逾期金额（过去1-5个月） | |
| 输出变量 | | | |
| *y* | | 次月违约情况（0:未违约;1:违约） | |

## 3.4 描述性统计

本章节中，我们对实验所需要的实验数据进行了描述性统计，我们计算了岩本的均值、方差、最大值和最小值。

首先是市场收益率的描述性统计，这里我们使用两个指数的收益率作为代替。

表一 指数数据描述性统计

|  |  |  |
| --- | --- | --- |
|  | 万得全A | 恒生综指 |
| 计数 | 331 | 331 |
| 最大值 | 5.6213 | 5.5112 |
| 最小值 | -8.1583 | -4.9381 |
| 标准差 | 1.4426 | 1.2603 |

表二 基金经理、股票和城市描述性统计

|  |  |
| --- | --- |
| 变量名称 | 数量 |
| 基金经理 |  |
| 城市 |  |
| 股票 |  |

表二 样本数据描述性统计

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 输入变量 | | | | | | |
| 变量名 | 均值 | | 方差 | 最大值 | 最小值 | 极差 |
| 全部样本 | 违约样本 |
|  | 3.78 | 3.74 | 1.50 | 6.00 | 1.50 | 4.50 |
|  | 1.75 | 1.34 | 1.69 | 5.00 | 0.00 | 5.00 |
|  | 1.51 | 1.29 | 0.31 | 1.59 | 0.00 | 1.59 |
|  | 1.98 | 1.95 | 0.06 | 1.98 | 0.00 | 1.98 |
|  | 2.62 | 2.69 | 0.30 | 2.72 | 0.82 | 1.90 |
|  | 3.82 | 3.47 | 1.02 | 6.19 | 1.09 | 5.10 |
|  | 1.53 | 1.44 | 0.74 | 2.44 | 0.00 | 2.44 |
|  | 4.81 | 4.08 | 3.16 | 10.44 | 0.47 | 9.97 |
|  | 2.78 | 3.70 | 1.86 | 8.78 | 0.71 | 8.07 |
|  | 2.36 | 2.38 | 0.72 | 3.40 | 0.56 | 2.84 |
|  | 5.37 | 5.93 | 2.34 | 9.71 | 0.00 | 9.71 |
|  | 1.86 | 1.91 | 1.80 | 5.21 | 0.00 | 5.21 |
|  | 1.49 | 2.29 | 2.06 | 8.64 | 0.00 | 8.64 |
|  | 0.41 | 0.30 | 0.48 | 1.30 | 0.00 | 1.30 |
|  | 0.85 | 0.84 | 0.47 | 1.44 | 0.00 | 1.44 |
|  | 0.71 | 0.41 | 0.56 | 1.16 | 0.00 | 1.16 |
|  | 1.94 | 2.59 | 0.87 | 4.24 | 0.28 | 3.96 |
|  | 4.58 | 5.57 | 4.38 | 15.04 | 0.00 | 15.04 |
|  | 1.68 | 1.66 | 0.43 | 2.15 | 0.86 | 1.29 |
|  | 0.89 | 0.90 | 0.15 | 0.99 | 0.33 | 0.66 |
|  | 2.72 | 2.71 | 1.40 | 4.85 | 0.00 | 4.85 |
|  | 2.21 | 2.34 | 0.59 | 3.01 | 0.60 | 2.41 |
|  | 6.55 | 6.15 | 2.24 | 8.65 | 0.00 | 8.65 |
|  | 6.05 | 5.67 | 1.95 | 7.20 | 0.00 | 7.20 |
|  | 5.41 | 4.75 | 2.01 | 7.01 | 0.00 | 7.01 |
|  | 2.92 | 2.27 | 2.10 | 6.12 | 0.00 | 6.12 |
|  | 0.69 | 0.65 | 0.49 | 2.30 | 0.46 | 1.84 |
|  | 0.72 | 0.75 | 1.38 | 4.48 | 0.00 | 4.48 |

续表二 样本数据描述性统计

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0.79 | 0.73 | | 1.59 | 3.97 | | 0.00 | 3.97 |
|  | 1.53 | 1.29 | | 1.44 | 2.88 | | 0.00 | 2.88 |
|  | -0.01 | 0.00 | | 0.20 | 0.00 | | -5.00 | 5.00 |
|  | 0.00 | 0.00 | | 0.06 | 0.00 | | -2.00 | 2.00 |
|  | 127.06 | 18076.24 | | 3273.75 | 255753.42 | | 0.00 | 255753.42 |
|  | 119.14 | 16949.78 | | 3105.59 | 243772.89 | | 0.00 | 243772.89 |
|  | 140.72 | 20019.67 | | 4611.51 | 483299.18 | | 0.00 | 483299.18 |
|  | 114.56 | 16297.34 | | 3025.09 | 220432.44 | | 0.00 | 220432.44 |
|  | 2048.15 | 289725.33 | | 53373.03 | 4184760.11 | | 0.00 | 4184760.11 |
| 输出变量 | | | | | | | | |
|  | 133.46 | 18985.97 | | 3429.26 | 269108.63 | | 0 | 269108.63 |
| 总样本量 | | | 违约样本量 | | | 未违约样本量 | | |
| 18921 | | | 133 | | | 18788 | | |

由表二可看出，、、、几个变量在违约样本集和总样本集上的均值有较大差异，可初步猜测这几个特征可能对于信用卡客户违约风险有一定的影响，但具体的影响程度还需要根据后续的机器学习模型的结果来确定。

# 4 实证分析

## 4.1 模型评价方法

### 4.1.1 混淆矩阵

我们的信用卡违约预测模型的目标是将信用卡账户分为两类：下个月违约的账户和下个月不违约的账户。对于二元分类算法，机器学习和统计文献中常用的性能衡量指标是一个2×2列联表，通常称为“混淆矩阵”，如表三所示。

表三 二分类问题的混淆矩阵

|  |  |  |
| --- | --- | --- |
| 真实情况 | 预测结果 | |
| 正例 | 反例 |
| 正例 | TP（真正例） | FN（假反例） |
| 反例 | FP（假正例） | TN（真反例） |

我们根据混淆矩阵的条目，可以计算出多个特定的性能度量指标，包括查准率（Precision）、查全率（Recall）、真正例率（True Positive Rate）、假正例率（False Positive Rate）、F1度量（F1-Measure）和Kappa统计量（Kappa Statistic）。本研究中，查准率被定义为正确预测的违约帐户数除以全部预测的违约帐户数，而查全率被定义为正确预测的违约帐户数除以违约帐户的实际数量。查准率旨在衡量误报的数量，而查全率则衡量假负类的数量。真正例率和假正例率的含义不言自明。另外，我们还考虑了两个统计数据，F1度量和kappa统计量，它们是结合了查准率和查全率的衡量指标。F1度量被定义为查准率和查全率的调和平均值，并为在查准率和查全率之间实现合理平衡的方法分配更高的值。kappa统计量衡量模型相对于随机分类的性能，并且可以被认为是在给定正例和反例的分布的情况下对预期准确度的改进。根据Khandani（2010）等人的说法，kappa统计值高于0.6，则表示模型具有预测能力。以下总结了上述各分类性能统计度量的定义：

查准率： ，

查全率： ，

真正例率： ，

假正例率： ，

F1度量，

Kappa统计量： ，

where and

### 4.1.2 ROC曲线 、AUC值

通过将不同的分类阈值应用于我们的预测模型，可以观察到正确分类和误报之间的折衷。为了更直观地表现出这种折衷，我们以假正例率为横轴、真正例率为纵轴，绘制出一条曲线，即ROC（Receiver Operating Characteristic，接收者操作特征）曲线，它是各种分类阈值的真假正例率的成对图，通常情况下，随着阈值的增加，真正例率增加，假正例率也会随之增加。ROC曲线显示真假正例率之间的折衷不是线性的，这意味着真正例率的增加并不总是与假正例率的增加相称。最佳分类阈值是不同的，但可以很容易地从ROC曲线确定。

最后，ROC曲线下面积（AUC值，Area Under Curve）是一种广泛使用的用于比较机器学习模型表现的度量指标。AUC作为数值，可以直观地评价和比较模型，AUC值越大，模型分类准确率越高。AUC值是一个概率值，当我们随机挑选一个正样本以及负样本，当前的分类算法根据计算得到的将这个正样本排在负样本前面的概率就是AUC值，AUC值越大，当前分类算法越有可能将正样本排在负样本前面，从而能够更好地分类。

## 4.2 模型结果

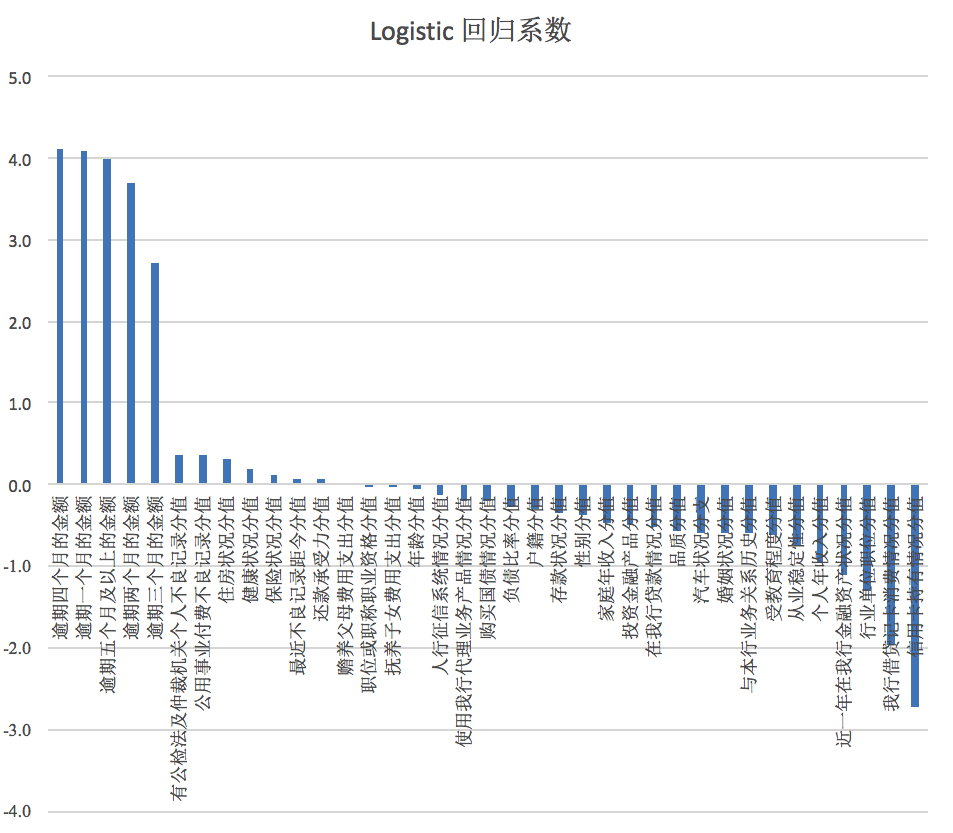
### 4.2.1 基于Logistic回归的信用卡风险预测模型

使用Logistic回归模型的预测结果如表四所示，此时，模型在训练集上的预测准确率为0.9897，在测试集上的预测准确率为0.9968。

表四 Logistic回归模型结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 混淆矩阵 | | | | | |
| 真实情况 | | 预测结果 | | | |
| 正例 | | 反例 | |
| 正例 | | 5623 | | 12 | |
| 反例 | | 6 | | 36 | |
| 分类性能统计度量指标 | | | | | |
| 查准率 | 查全率 | 真正例率 | 假正例率 | F1度量 | Kappa统计量 |
| 0.75 | 0.8571 | 0.9979 | 0.1429 | 0.8000 | 0.7984 |

Logistic回归模型在测试集的预测准确性较高，主要是由于数据集样本类别不平衡造成的，要考察其真实的预测能力，需要关注模型对于违约样本的预测准确性，即查准率和查全率。由混淆矩阵可看出，模型预测出假正例和假反例的现象时有发生，查准率只有75%，查重率也只有86%。这样，在真实的银行信用卡违约风险预警的场景中，模型的可靠性较差，并不适用。另一方面，综合模型查全率和查准率的F1度量为0.8，Kappa统计量0.7984也远高于0.6，表明Logistic回归模型的分类预测效果还是远优于随机分类的。



图一 Logistic回归模型系数

虽然Logistic回归模型的预测违约样本的能力较为有限，但作为一种广义的线性回归模型，它的优势是模型具有较强的解释性，可通过回归系数体现出各特征变量对于输出分类的影响权重。Logistic回归模型最终得到的回归系数，如图一所示。从图中，可以直观地看出，历史违约记录对于次月违约风险有较大的影响，若过去有违约情况发生，则其次月违约风险相应增大。除去历史交易的影响，在用户特征方面，对违约风险有显著正向影响的是公检法及仲裁机关个人不良记录和公用事业付费不良记录两项，表明用户在其他领域的不良表现与其使用信用卡的违约风险有很强的正相关性，因此，银行在发行信用卡时需要对客户的相关背景进行审查，关注其不良记录，以降低客户信用卡违约风险为银行带来的损失。住房状况对违约风险有较显著正面影响，主要来自于使用信用卡还房贷的这类客户，逾期的状况时有发生。其它一些对信用卡违约风险有较显著负向影响的因素，主要包括在本行贷款、存款、购买国债、投资金融产品、借贷记卡消费情况、信用卡持有情况、与本行业务关系历史等在该银行的其他金融业务情况，以及个人与家庭年收入、汽车状况、受教育程度、婚姻状况、就业稳定性等客户个人稳定性的指标。如果客户在该银行的其他金融业务方面有良好稳定的表现，则其相应的信用卡违约风险就会降低；如果客户个人生活和工作稳定性较高，则其信用卡违约风险也会较低。

### 4.2.2 基于SVM的信用卡风险预测模型

在使用SVM 模型时，我们选择了RBF核函数，并通过网格搜索，得到模型的最优参数，即惩罚系数C=8192.0，gamma=0.004。将训练得到的模型在数据集上做预测，结果如表五所示，此时，模型在训练集上的预测准确率为0.9999，在测试集上的预测准确率为0.9975。

表五 SVM模型结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 混淆矩阵 | | | | | |
| 真实情况 | | 预测结果 | | | |
| 正例 | | 反例 | |
| 正例 | | 5627 | | 0 | |
| 反例 | | 14 | | 36 | |
| 分类性能统计度量指标 | | | | | |
| 查准率 | 查全率 | 真正例率 | 假正例率 | F1度量 | Kappa统计量 |
| 1 | 0.72 | 1 | 0.28 | 0.8372 | 0.836 |

由表五可见，支持向量机模型在预测集上的表现并不是很好。虽然达到了100%的查准率，但查全率只有72%，这意味着SVM模型在预测违约时，很容易错判一些违约风险较高的样本，这是我们在真实的应用场景中很不愿看到的情况。虽然真正例率达到了100%，但假正例率也有28%，表明该模型过多地将样本预测为正例（未违约样本），对于违约样本的识别不够全面。但从模型的F1度量和Kappa统计量来看，SVM相较于随机分类还是有较强的预测能力。

### 4.2.3 基于CART模型的信用卡风险预测模型

CART分类回归树模型的预测结果如表六所示，此时，模型在训练集上的预测准确率为0.9990，在测试集上的预测准确率为0.9981。

表六 CART模型结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 混淆矩阵 | | | | | | | |
| 真实情况 | | | 预测结果 | | | | |
| 正例 | | | 反例 | |
| 正例 | | | 5621 | | | 8 | |
| 反例 | | | 3 | | | 45 | |
| 分类性能统计度量指标 | | | | | | | |
| 查准率 | 查全率 | 真正例率 | | 假正例率 | F1度量 | | Kappa统计量 |
| 0.8491 | 0.9375 | 0.9986 | | 0.0625 | 0.8911 | | 0.8901 |

由表六可见，CART模型虽然在训练集和预测集上的准确率较高，但究其原因，是原始样本集中的样本不平衡造成的，所以对于违约样本的预测误差无法仅通过模型的预测准确率来表现，而是应更多地关注混淆矩阵。根据CART模型预测的混淆矩阵，可以计算得到查准率和查全率指标，可以看出CART模型的查准率只有85%左右，表明模型在做样本外预测时，存在一定的可能性将正例（未违约样本）错误地预测为反例（违约样本）。而查全率相对较高，说明模型在做预测时能比较全地识别出违约样本。F1度量为0.8911，综合了查准率和查全率两个指标，表明模型的预测效果尚可；kappa统计量0.8901显著大于0.6，表明CART模型相对于随机分类而言，有很强的分类预测能力。

### 4.2.4 基于随机森林的信用卡风险预测模型

使用随机森林模型的预测结果如表七所示，此时，模型在训练集上的预测准确率为0.9989，在测试集上的预测准确率为0.9995。

表七 随机森林模型结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 混淆矩阵 | | | | | |
| 真实情况 | | 预测结果 | | | |
| 正例 | | 反例 | |
| 正例 | | 5636 | | 0 | |
| 反例 | | 3 | | 38 | |
| 分类性能统计度量指标 | | | | | |
| 查准率 | 查全率 | 真正例率 | 假正例率 | F1度量 | Kappa统计量 |
| 1 | 0.9268 | 1 | 0.0732 | 0.962 | 0.9618 |

由表七可见，随机森林模型相较于前一个CART模型在各方面的性能都有显著提升。由于预测准确性受初始样本本身的类别不平衡影响较大，所以即使数值较高也无法直接单独用来客观地评价模型。从混淆矩阵看，随机森林模型的查准率和查全率相较于单独的树模型（CART）都有明显提高，分别为100%和93%。但另一方面，不可否认的是，模型查准率和查全率都很高，也与测试集中违约样本数量较少有关。从F1度量0.9620和kappa统计量0.9618来看，随机森林模型的预测能力明显高于CART模型，由此得出，集成学习方法的应用可以显著提高单个树模型（CART）的表现。

### 4.2.5 基于XGBoost的信用卡风险预测模型

使用XGBoost模型的预测结果如表八所示，此时，模型在训练集上的预测准确率为0.9999，在测试集上的预测准确率为0.9994。

表八 XGBoost模型结果

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 混淆矩阵 | | | | | | | |
| 真实情况 | | | 预测结果 | | | | |
| 正例 | | | 反例 | |
| 正例 | | | 5633 | | | 1 | |
| 反例 | | | 2 | | | 41 | |
| 分类性能统计度量指标 | | | | | | | |
| 查准率 | 查全率 | 真正例率 | | 假正例率 | F1度量 | | Kappa统计量 |
| 0.9672 | 0.9534 | 0.9998 | | 0.0465 | 0.9647 | | 0.9644 |

作为另一种基于树模型的集成学习方法，XGBoost模型的总体表现也是显著优于CART模型的，但与随机森林模型相比较，没有绝对优势。由上表可见，XGBoost模型的查准率为97%，略低于随机森林模型，但查全率95%高于随机森林模型。综合查准率和查全率考虑，XGBoost模型和随机森林模型的F1度量很相近，Kappa统计量也相差很小，都为0.96左右。因此可以得到结论，作为集成学习的随机森林模型和XGBoost模型的预测能力都明显优于单个树模型（CART），但这两种模型总体的预测能力没有显著差别，只是随机森林模型的查准率更高，而XGBoost模型的查全率更高。

## 4.3 模型比较

由4.2节的模型结果可以看出，本实证研究所采用的五种分类模型在该银行信用卡客户违约数据集上都表现出不错的预测能力，其中决策树模型，尤其是使用集成学习方法改进后的树模型，都有着出色的预测能力。

首先，Logistic回归模型作为一种传统的线性回归模型，通过本实证研究，表现出一定的预测能力，但对于违约样本的查全和查准的能力都不高，若作为真实的银行信用卡违约风险预警模型来使用，容易产生误报的情况。但另一方面，虽然Logistic回归模型不适用于违约风险的预警，但其简单的线性回归逻辑使得模型具有较好的解释性，我们可以很直接地根据各个输入变量的回归系数来判断其对于信用卡客户违约风险的影响方向和影响程度。

SVM模型的预测效果，从F1度量和Kappa统计量来看，是优于Logistic回归模型的，查准率也达到了100%，但查全率却比Logistic回归更低。然而，在真实的银行信用卡违约风险预警中，模型的查全能力至关重要，所以SVM模型同样不适合用作该银行的风险预警系统。同时，需要指出的是，由于SVM模型需要在高维空间中寻找一个超平面来分隔两类样本，所以其时间开销很大，模型的预测效率很低，这一点也使得它不适合在实际中应用。

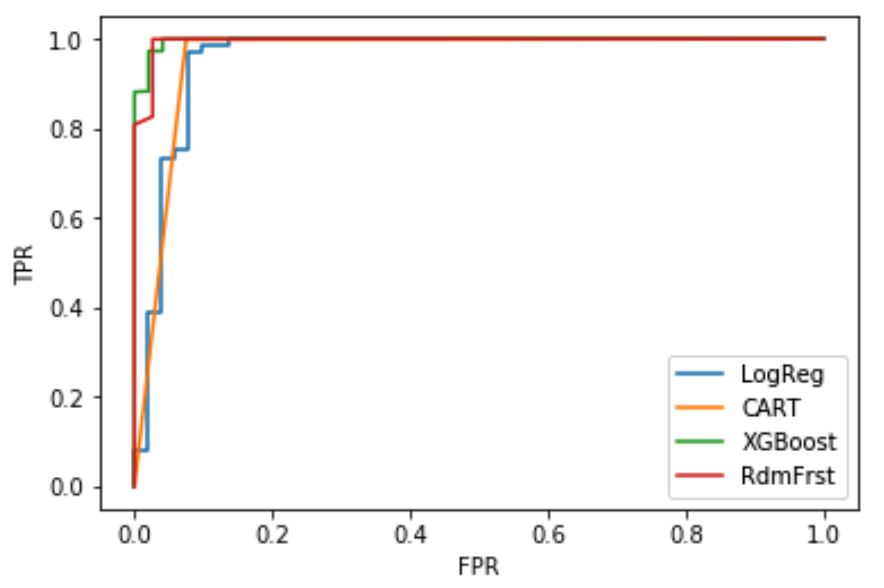
相较于Logistic回归模型和SVM模型，决策树模型CART的预测效果有了显著的提升。CART模型在测试集上同时达到了较高的查全率和查准率，F1度量的值近似达到了0.9。这一结果表明，CART模型在该银行数据集上有较好的预测效果，一定程度上可以有效预测该银行信用卡客户的违约风险，但在使用CART模型时，要注意剪枝，防止模型过拟合，从而达到较高的外本外预测准确性。

虽然上述CART模型已经表现出不俗的预测能力，但使用集成学习方法改进的树模型随机森林和XGBoost，在CART模型的基础上，达到了更优的表现。随机森林模型，通过Bagging策略，训练出多棵独立的分类回归树，综合来对样本进行分类预测。预测结果显示，随机森林模型同时达到了100%的查准率和92%的查全率，这一结果表明，随机森林模型可以准确而全面地检测出潜在的违约样本，可以考虑用作该银行的信用卡客户违约风险的预警模型。XGBoost模型，采用Boosting的方法，训练出一系列相关的分类回归树，来共同对样本进行分类预测。XGBoost的总体预测能力和随机森林不相上下，但表现出更加均衡的查准率和查全率，即查准率97%略低于随机森林模型，但查全率95%略高于随机森林模型。这样的预测结果表明，XGBoost也同样适合用作该银行信用卡客户违约的预警模型。

最后，为了更直观地比较各个模型对于客户信用卡违约风险的预测能力，我们给出了表九所示的上述五种模型的AUC值，并绘制了如图二所示的ROC曲线。结果显示，SVM模型表现相对较差，Logistic回归模型的预测准确性也一般，树模型CART以及以树模型为基学习器的随机森林和XGBoost都表现出较强的预测能力，再综合考虑查全率和查准率，随机森林和XGBoost是最适合该银行数据特征的两种信用卡客户违约预警模型。

表九 模型的AUC值

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 模型 | Logistic回归 | SVM | CART模型 | 随机森林 | XGBoost |
| AUC值 | 0.9275 | 0.86 | 0.9680 | 0.9634 | 0.9767 |



图二 模型的ROC曲线

# 5 研究结论

银行信用卡客户的风险管理对于银行来说至关重要，一方面，适当的违约风险预警模型可以有效地避免银行不必要的损失；另一方面，搞清楚信用卡客户违约风险背后的影响因素，才能更好地从根本上降低银行所面临的信用卡风险。本研究针对某城乡行的真实信用卡客户数据，运用了五种不同的机器学习方法进行数据挖掘，分别对样本数据进行了分类预测，并分析比较了各个模型的结果。最后，从文献研究和实证结果出发，我们总结得出了以下几条结论：

1. **特征变量的选取对风险预测模型的表现有很大的影响。**

机器学习和数据挖掘类的模型，其表现对数据的依赖性很高，尤其是模型输入的特征空间十分敏感。本研究在特征选取上颇费心思，尽可能选择多个差异化的特征变量来较全面地刻画信用卡客户行为，同时又努力控制特征变量的总数，避免维度过高带来的灾难。通过综合考量，最终确定了由37个不同变量组成的特征空间来作为预测模型的输入。从实证结果来看，这些变量能较准确而全面地刻画该银行信用卡客户的特征，训练出的模型都具有较好的预测能力。

1. **线性模型的解释性更强，对风险管理更具有启发性。**

Logistic回归模型，作为本实证研究中唯一的线性模型，在实现不错的预测能力的同时，由于其模型的简明性，可以很好地解释各输入变量对于违约风险的影响程度。据此，我们根据4.2.1中的结果，针对性地提出了一下几点关于该银行的信用卡客户违约风险管理的建议：

1. 关注信用卡客户的历史违约记录，若客户过去几个月频繁有违约情况发生，可考虑提醒或冻结客户的信用卡，以及时止损。
2. 关注信用卡客户过去的不良记录，例如公检法及仲裁机关个人不良记录和公用事业付费不良记录。银行在发行信用卡时，需要对客户的相关背景进行审查，尤其要关注其历史不良记录，面向优质客户群发行信用卡，以降低客户信用卡违约风险为银行带来的损失
3. 关注信用卡客户在本行的其他业务，包括贷款、存款、购买国债、投资金融产品、借贷记卡消费情况、信用卡持有情况、与本行业务关系历史等在该银行的其他金融业务情况。当客户在该银行的其他金融业务方面有良好稳定的表现时，则其相应的信用卡违约风险就会较低；反之，当客户本行的其他金融业务出现不良行为时，则应当高度关注其信用卡使用状况，预防违约情况发生。
4. 关注信用卡客户个人生活的稳定性，例如个人与家庭年收入、汽车状况、住房状况、受教育程度、婚姻状况、就业稳定性等。当客户个人生活状况良好稳定时，其信用卡使用行为就会相对正常；当生活出现变故，如突发失业或离异等，则需特别关注其信用卡违约风险。
5. **非线性模型的预测效率和准确率更高，且自动化程度高，适用于大数据集。**

本研究使用的非线性模型包括SVM支持向量机模型、CART分类回归树模型、随机森林模型和XGBoost模型四种，它们都是自动化程度很高的机器学习模型，给定模型输入，进行适当的调参，即可得到分类预测结果，适用于大规模的数据集，例如本研究使用的银行信用卡客户数据，包括大量的用户特征和交易记录。 实证结果表明，相较于简单的线性模型，这类非线性模型都在预测集上表现出更高的预测效率和准确性，更适合做为银行信用卡违约风险预警模型来使用，尤其是随机森林和XGBoost模型。

1. **相同模型在不同数据集上表现不同，没有哪种模型具有绝对优势。**

最后需要特别指出的是，由于本实证研究仅使用了单个银行的数据进行模型的训练和预测，模型的表现受该银行的数据特征的影响较大。本研究结果表明，随机森林和XGBoost模型对于该银行客户的信用卡违约风险有很好的预测效果，并不代表这两种模型在其他银行也有类似出色的表现。因为同样的模型，即使选取相似的特征空间，在不同的样本数据集上也会有不同的表现。因此，在信用卡违约风险预测方面，我们需要根据真实的数据特征来选取最适合的机器学习模型，没有哪种模型具有绝对的优势。

# 6 研究展望

1. 由于本实验是检测基金经理的社交关系是否能很好地反馈股价崩盘。实验的本质目的是探究一个变量对另一个变量的相关性。同时给出一些刻画相关性的模型。因而实验并没有对股价崩盘的参考因素，做出时效性的考量。本文设置的时间界面的长度为一年。来检测指标之间的相关性。同时，由于很多基金经理的大幅减持的行为都发生在一年中相对较短的一段时间内。因而在实际应用中，可以在本实验的基础上，适当地缩短时间截面，来达到预测的效果。但是如何选择一个相对恰当的时间截面是一个深入研究的问题。
2. 本实证研究中用到的数据集只涵盖了该银行六个月的信用卡交易数据，由于模型采用监督学习的方式，我们把最近一个月的违约金额作为输出变量，而之前五个月的违约情况视作历史交易记录，作为模型输入变量的一部分，所以我们在训练模型时只使用了截面数据，数据集包含的信息十分有限。若能获得时间段更长的信用卡交易数据，则可考虑把样本集处理为时序数据，这样在训练模型时，数据量更大，能发掘得到的信息也更多，从而提高模型准确性。
3. 本实证研究只采用了来自单一银行的信用卡客户数据，而无法同时得到来自多家银行的更全面的数据。而同样的模型在不同银行信用卡客户数据集上的表现可能存在较大差异，所以本研究得到的结果和结论仅适用于本行的信用卡风险管理。
4. 本实证研究分析并比较了五种机器学习模型，在未来的研究中，可以尝试更多数据挖掘方法，并考虑对已有的成熟模型进行进一步的升级和优化。

# 参考文献

[1] 周志华. 机器学习

[2] 罗真，张宗成，职业忧虑影响基金经理投资行为的经验分析[j]. 财经研究，2004，30（12）：111-120

[3] 张雪峰. 开放式基金投资行为及其影响因素研究[D].天津商业大学,2017.

[4] 申宇, 赵静梅, 何欣. 基金未公开的信息: 隐形交易与投资业绩 [J]. 管理世界,2013 (8): 53-66.

[5] 陈莹,袁建辉,李心丹,肖斌卿.基于计算实验的协同羊群行为与市场波动研究[J].管理科学学报,2010,13(09):119-128.

[6] 楼晓霞. 社交关系网络影响股市崩盘的机理[D].浙江大学,2019.

[7] 郭白滢,李瑾.机构投资者信息共享与股价崩盘风险——基于社会关系网络的分析[J].经济管理,2019,41(07):171-189.

[8] 陈新春,刘阳,罗荣华.机构投资者信息共享会引来黑天鹅吗?——基金信息网络与极端市场风险[J].金融研究,2017(07):140-155.

[10] Jiang H.，2010，“Institutional Investors，Intangible In⁃

formation and the Book-to-Market Effect”，Journal of Financial Economics，Vol. 96，Iss.1，pp. 98~126.

[11] Bushee，B. and T. Goodman，2007，“Which Institution⁃

al Investors Trade Based on Private Information about Earnings and Returns?”，Journal of Accounting Research，Vol. 45，pp. 289~ 321.

[12] Pareek，A.，2011，“Information Networks: Implications

for Mutual Fund Trading Behavior and Stock Returns”，Working Paper.

[13] Ozsoylev，H.，J. Walden，M. D. Yavuz，R. Bildik，2011，

“Investor Networks in the Stock Market”，Working Paper.

[14] Scharfstein, Stein, 1990, “Herd Behavior and Investment”, American Economic Review, Vol.80(3): PP465-479.

[15] Prendergast, 1996, “Stole Impetuous Youngsters and Jaded Old-Times: Acquiring a Reputation for Learning”, Journal of Political Economy, Vol.104(6): PP1105-1134.

[16] Chen, J., Hong, H., & Stein, J. C. Forecasting crashes: trading volume, past returns, and conditional skewness in stock prices[J]. Journal of Financial Economics, 2001, 61(3): 345-381.

[17]Hong, H., Stein, J. C. Differences of Opinion, Short-Sales Constraints, and Market Crashes[J]. Review of Financial Studies, 2003, 16(2): 487–525.

[1] 李涵峰. 基于数据挖掘的银行信用卡违约实证研究[D].兰州大学,2018.

[2] 陈莹,宋建华.中国城市居民信用卡违约行为的影响因素[J].金融论坛,2017,22(09):27-38.

[3] 鲁长东.大数据分析挖掘技术在信用卡风险管理中的应用[J].中国信用卡,2017(04):32-36.

[4] 范巍强,刘暾东.基于BP神经网络的信用卡违约风险预测[J].电脑知识与技术,2011,7(10):2348-2349.

[5] 赵刚. 商业银行信用卡业务信用风险管理研究[D].华东师范大学,2007.

[6] 张双全. 基于改进智能算法的信用卡客户违约预测研究[D].长春工业大学,2018.

[7] 陈栋栋. 信用卡客户风险预警与应用研究-基于主成分分析法[D].山东大学,2012.

[8] 邵亦明. 中国信用卡分期付款业务违约风险实证研究[D].上海交通大学,2013.

[9] 帅理. 个人信用风险评估理论与方法的拓展研究[D].电子科技大学,2015.

[10] 陈为民. 基于支持向量机的信用卡信用风险管理模型与技术研究[D].湖南大学,2009.

[11] 盛洁. 商业银行信用卡违约概率评估的实证研究[D].厦门大学,2014.

[12] 陈雄. 信用卡违约风险个人影响因素研究[D].浙江大学,2011.

[13] Florentin Butaru, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W. Lo, Akhtar Siddique. Risk and risk management in the credit card industry[J]. Journal of Banking & Finance,2016,72:218-239.

[14] Amir E. Khandani, Adlar J. Kim, Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms[J]. Journal of Banking & Finance,2010,34(11):2767-2787.

[15] I-Cheng Yeh, Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients[J]. Expert Systems with Applications,2009,36(2):2473-2480.

[16] Tony Bellotti, Jonathan Crook. Forecasting and stress testing credit card default using dynamic models[J]. International Journal of Forecasting,2013,29(4):563-574.

[17] Berry, M., & Linoff, G. Mastering data mining: The art and science of customer relationship management[J]. New York: John Wiley & Sons, Inc, 2000.

[18] Paolo, G. Bayesian data mining, with application to benchmarking and credit scoring. Applied Stochastic Models in Business and Society[J]. 2001(17): 69–81.

[19] Thomas, L. C. A survey of credit and behavioral scoring: Forecasting financial risk of lending to consumers[J]. International Journal of Forecasting, 2000(16): 149–172.

[20] Hand, D. J., & Henley, W. E. Statistical classification methods in consumer credit scoring: A review[J]. Journal of the Royal Statistical Society, Series A – Statistics in Society, 1997, 160(3), 523–541.

[21] Koh, H. C., & Chan, K. L. G. Data mining and customer relationship marketing in the banking industry[J]. Singapore Management Review, 2002, 24(2), 1–27.

[22] Lee, T. S., Chiu, C. C., Lu, C. J., & Chen, I. F. Credit scoring using the hybrid neural discriminant technique[J]. Expert Systems with Applications, 2002, 23(3), 245–254.

# 致谢

我首先要感谢我的导师李心丹老师和方立兵老师，在选题初期，两位老师就给了我很多建议和帮助，确定毕业设计题目后，他们及时跟进我的研究进度，多次与我沟通交流，并在遇到问题和困惑时第一时间帮我解决。其次，我还要感谢为我的实证研究提供数据和硬件支持的某基金此外，我还要感谢在毕业设计中，与我并肩作战的研究生同学谢贤，本研究的顺利完成离不开他几个月来无私的奉献与支持。最后，我要特别感谢南京大学计算机与金融工程实验班的全体老师和同学们，感谢大家一路以来的帮助与支持，让我的大学生活更加丰富而难忘！