北京大学

硕士研究生学位论文

题目：个体异质性特征与证券投资收益——基于 2015-2016 年“中国经济生活大调查”

姓　　名：	孔晓婷
学　　号：	1601214579
院　　系：	国家发展研究院
专　　业：	金融学
研究方向：	金融学
导师姓名：	胡大源　教授

二〇一八 年 四　月

版权声明

摘要

2015 年我国沪深股市出现了历史罕见的股灾，基于 2015~2016 年 CCTV 财经频道“中国经济生活大调查”数据发现，在 2015 年，参与证券市场投资的受访者中仅有 16%的个体获得了赢利，绝大多数均亏损且收益分化较大。那么个体投资者受教育程度、性别、年龄等异质性特征会不会以及会如何影响个体证券市场的投资收益状况呢？

为探究这一学界尚未有较多研究的问题，本文首先结合行为金融学的过度自信、过度反应和心理账户理论提出了个体不同异质性特征影响个体投资收益状况的假说，并进一步利用 2015~2016 年 CCTV“大调查”的微观截面数据进行有序多分类逻辑回归，发现高受教育程度、女性、企业管理人员和行政事业单位人员、高家庭收入的投资者相比于受教育程度低、男性、自由职业者和进城务工人员、低家庭收入的投资者，获得各种赢利情况尤其是赢利 20%以内的概率显著提高，且发生极端亏损（亏损 50%以上）的可能性显著降低；相比于租房的投资者，拥有农村住房的投资者发生极端亏损可能性的提高而获得赢利的可能性降低；年龄因素对投资收益的影响则不显著。

在回归分析的基础上，本文还通过计算有序多分类逻辑回归模型预测结果与真实数据的一致性来评估模型预测准确度，从而评判其实际应用价值。考虑到变量间的交互性和为提高模型预测准确度，本文进一步使用随机森林模型训练建模和预测分析，与有序多分类逻辑回归模型形成优势互补，发现随机森林模型的预测准确度为 42.60%，相比有序多分类逻辑回归模型的 37.87%有显著提高且随机森林模型的泛化能力较好；此外通过随机森林模型的变量相对重要性曲线，发现职业因素可能对个体投资者的投资收益状况有相对更重要影响，而性别因素的相对影响较小。

总之，本文的研究内容对学界相关领域内的研究形成较好补充，使用的预测准确度评估方法有助于提高学界对计量经济模型现实应用价值的重视程度，基于样本量大、节点独特、代表性强的“大调查”数据和创新性使用的随机森林模型得到的实证结论也为中低收入个体投资者更好地参与证券市场投资、业界金融机构更有针对性地服务证券投资者提供了有价值的启示和参考。

关键词：异质性特征；投资收益；行为金融学；有序多分类逻辑回归；随机森林

Individual Heterogeneous Characteristics and Returns from Securities Investment: Based on 2015-2016 *China Economic Life Survey*

Xiaoting Kong (Finance)

Directed by Professor Dayuan Hu

ABSTRACT

In 2015, a rare historical disaster shook China's Shanghai and Shenzhen stock markets, and based on data of the *China Economic Life Survey* conducted by the CCTV Finance Channel from 2015 to 2016, only 16% of the respondents who participated in the stock market in 2015 gained profits, whereas most of the participates lost money and their returns differentiated a lot from person to person. Thus, whether investor's individual heterogeneous characteristics, such as education level, gender, age, etc., will affect individual's returns from securities investment? And if any, how will them affect?

In order to explore this issue which has not got much attention from scholars, this thesis firstly puts forward hypotheses on the effects of individual heterogeneous characteristics to individual's returns from securities investment, which are based on theories of Overconfidence, Overreaction and Mental Account from Behavioral Finance. Then, using the cross-sectional data of the *China Economic Life Survey* from 2015 to 2016 and employing Ordered Multi-classified Logistic regression, this thesis finds out, compared to their counterparts, it is the individual who gets higher education or is female or serves as a company's manager/an administrative institution employee or has higher household income that gets significantly higher chance to gain profits at all levels, especially within 20%, and gets significantly lower chance to lose extremely (more than 50%); and compared to the investor who rents house, the investor who owns rural house gets significantly higher chance to lose extremely with in the meantime significantly lower chance to profit; while age does not affect individual's returns significantly.

On the basis of regression analysis, this thesis further evaluates the accuracy of model's prediction via calculating the consistency between the predicted results of Ordered Multi-classified Logistic regression model and the real data, which contributes to make judgements about the model's practical application value. Considering the interaction between variables and in order to improve the accuracy of model's prediction, this thesis takes further steps to use Random Forest to train, to model and to predict, which complements the advantages of Ordered Multi-classified Logistic regression model. The findings are that Random Forest model has an accuracy of 42.60%, which is a significantly improved outcome compared to 37.87% of Multi-classified Logistic regression model, and the generalization ability of Random Forest model is also much better; besides, through variables' relative

importance curve of Random Forest model, it might be the case that career factor may have a relatively more important impact on individual's returns, while the relative impact of gender factor may be smaller.

In sum, this thesis serves as a good complement to relevant academic fields, and the assessment method of prediction accuracy it used may help the academic community to put more emphasis on econometric model's practical application value. Based on the survey data which features large sample, unique nodes, and strong representations and using the novel method of Random Forest model, the empirical findings of this thesis may also help individual investor to better participate in the securities investment and supply financial companies with valuable enlightenments and references when it comes to providing individualized service.

KEY WORDS： Heterogeneous Characteristics, Investment Returns, Behavioral

Finance, Ordered Multi-classified Logistic Regression, Random Forest

目录

第一章　引言

1.1　选题背景与现实意义

1.1.1　选题背景

自 2014 年 7 月 10 日起至 2015 年 6 月 12 日，我国沪深证券市场开启了一轮历史罕见的牛市行情，以收盘价统计，上证综指从 2038.34 点上涨至 5166.35 点，涨幅达 153.46%，深证成指从 7159.63 点上涨至 18098.27 点，涨幅达 152.78%，两市均达到 2007 年金融危机以来的最高点。然而，自 2015 年 6 月 15 日开始的几个月内，沪深股市出现了历史罕见的急速大幅暴跌，即记入中国证券市场历史的“2015 年股灾”：上证综指从 6 月 15 日的 5062.99 点跌至 7 月 3 日的 3686.92 点，跌幅达 27.18%；深证成指从 6 月 15 日的 17702.55 点跌至 7 月 3 日的 12246.06 点，跌幅达 30.82%；三周内 A 股市值缩水约 15 万亿，相当于希腊 2014 年全国 GDP 总值的 10 倍之多；三个月内下跌股票的比例高达 95%，其中跌幅超过 70% 的股票占比达 60%。

为稳定市场情绪、避免证券市场进一步失控，2015 年 7 月 1 日，国务院联合证监会、央行等部委机构启动强力的联合救市行动，并以证金公司为主要救市实施机构，大幅增持沪深蓝筹股及部分中小盘股票，为股市注入流动性。国家层面的强力救市行动虽然在一定程度上避免了股市发生更剧烈的下跌，但并没有完全扭转股市的下跌趋势，从 6 月 19 日至 8 月 25 日出现了中国 26 年股市历史上从未有过的累计 11 天“千股跌停”现象，8 月 26 日上证综指触及 2850.37 点的年度最低点，深证成指也达到 9776.21 点的年度低点。之后，随着市场情绪的逐渐平稳以及相关救市措施的逐步实施，至 2015 年 12 月 31 日年终收盘时，上证综指缓幅上升至 3539.18 点，深证成指上升至 12664.89 点。

图 1 2014 年 6 月初~2016 年 6 月初我国上证综指走势

数据来源：Wind 资讯

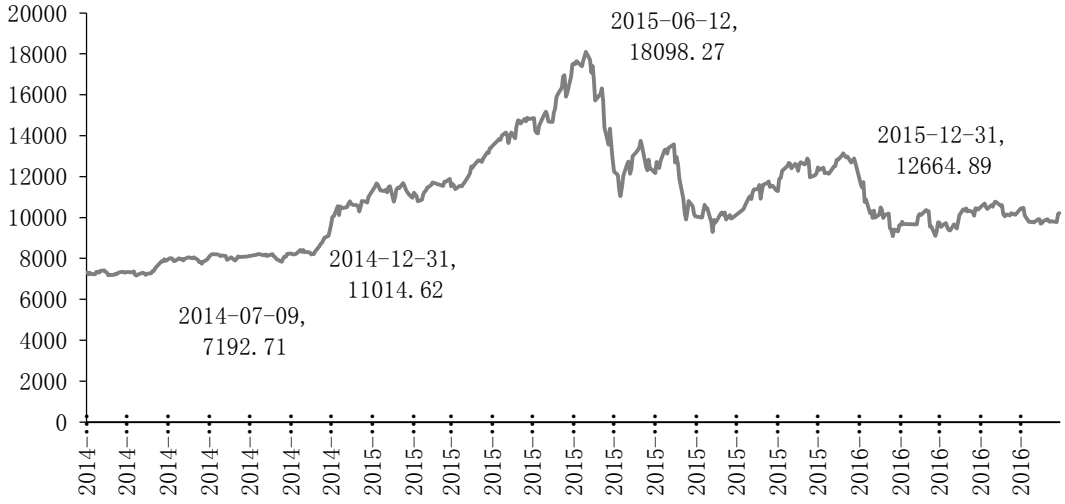

图 2 2014 年 6 月初~2016 年 6 月初我国深证成指走势

数据来源：Wind 资讯

1.1.2　选题现实意义

与外国证券市场主要由机构投资者构成不同，我国沪深股市的参与主体长期以来是广大个体投资者（“普通股民”，“散户”），而基金及国家队（“证金、汇金、社保基金等”）等组成的机构投资者占比较低。根据《2015 年中国证券登记结算

统计年鉴》的数据，至 2015 年末，沪深股市的个体投资者共 9882.15 万人，占全部投资者数量的 99.71%，其中持有股票流通市值在 50 万元以下的共 4679.59 万人，占全部投资者数量的 47.22%。在 2015 年我国股市的罕见巨幅波动中，广大个体投资者的投资收益受到较大影响且呈现出较大的分化。基于 2015~2016 年 CCTV“中国经济生活大调查”数据，发现参与证券市场投资的个体中仅有约 16%获得了赢利，而亏损 50%以上的个体则占到约 21%。

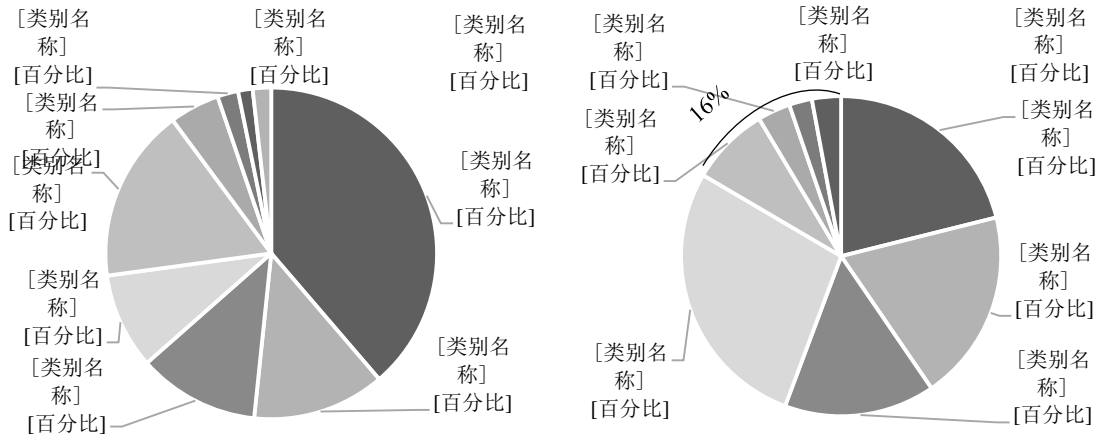

图 3 问卷受访者证券市场投资收益状况的分布情况（左：整体；右：炒股群体）

数据来源：2015-2016 年 CCTV 财经频道“中国经济生活大调查”

导致不同个体投资者在 2015 年证券投资收益产生较大差异的原因是多样的，那么诸如投资者受教育程度、性别、年龄等的异质性特征是否影响了个体投资者的证券投资收益呢？又分别有怎样的影响呢？

通过解答这些问题，将不仅有利于帮助个体证券投资者对自身证券投资收益产生更合理的预期，从而更审慎适当地参与证券市场投资，更好地保障中低收入群体的利益；同时，也有利于为相关金融机构使用个体证券投资者的异质性特征预测其投资收益情况提供有价值的参考，为丰富诸如保险、投资咨询等有针对性的金融产品提供实证支撑，帮助业界金融机构更好地服务于个体投资者；此外，在 2015 年我国沪深股市的历史罕见大幅波动行情中，个体异质性特征对证券投资收益的或有影响将会凸显，因而对这一时期进行研究也有助于提高获得有价值结论的可能性。

1.2　研究必要性

在进行上述问题的探究之前，有必要通过梳理和对比已有的相关研究，找到本文的研究必要性。总体来看，基于微观调查数据探究我国证券投资者个体异质性特征对其投资收益影响的研究在近一两年来呈现增多的趋势但整体仍较少，且

没有得到一致的结论，同时已有研究中使用的数据不仅数据量较小而且代表性不够强。因而，使用 2015~2016 年 CCTV“中国经济生活大调查”这一覆盖面广、代表性强的大样本横截面数据，有助于为证券投资者个体异质性特征影响投资收益的研究提供有益的补充，具有较大的研究必要性。

具体来看，彭星辉、汪小虹（1995）针对上海市的投资者进行了调查分析，发现高反应性个体（感受性高、活动性低）倾向于选择较为保守和低风险的投资策略，而低反应性个体（感受性高、活动性低）则倾向于选择较为冒险和高风险的投资策略，但是两种个体的投资收益情况并没有显著差异。李心丹等（2002）基于某证券营业部 7894 名个体投资者的交易数据，使用时间分析法实证发现投资者的过度自信心理会降低其投资收益。王垒等（2003）通过问卷调查的方式对全国 7 个城市的 1063 位投资者证券投资者的行为与心理特征进行了研究，发现投资者的赢利可能性与对投资对象的了解程度、投资知识的多少、独立性和自我效能呈正相关关系。赵振华等（2010）基于 2008 年底来自国内 43 家主要基金公司的 35866 份问卷调查数据，通过分组和多分类评定模型进行实证回归，发现投资者家庭收入、投资年限、资金时间约束对基金投资者的投资收益有显著影响，但投资者年龄和投资规模对投资收益没有显著作用。张腾文等（2016）基于 2015 年四川省《中小投资者权益保护调查问卷》数据，使用分位数回归方法进行实证探究，发现投资者的收入水平对投资收益呈现显著的正向影响，而投资经验与投资收益呈现倒 U 形关系。袁典（2016）使用 2013 年 CHFS 问卷受访者中参与股市投资的 1829 名投资者数据，采用逐步回归和 Probit 回归方法，实证研究发现投资者性别对投资收益具有显著影响，女性比男性获得赢利的可能性更大。窦松博（2016）使用某证券营业部 10103 名个人投资者数据，结合行为金融学理论和通过相关系数分析，发现投资者性别对投资收益没有显著影响，而投资者年龄对投资收益有显著的负向影响。

1.3　研究框架与方法

在本章引言部分后的第二章，本文将首先介绍行为金融学的相关理论，进而分别提出本文重点关注的个体异质性特征影响证券投资收益的理论假说。

随后在第三章，将对本文使用的 2015~2016 年 CCTV“中国经济生活大调查”数据进行介绍和统计性描述，同时设定后续实证部分所使用的被解释变量、解释变量及控制变量。

在对理论层面、数据和变量层面进行介绍和设定后，本文的第四章将使用传统计量经济学方法中的有序多分类逻辑回归模型进行实证回归，并在分析回归结果时与第二章提出的理论假说进行呼应，然后再进一步分析回归模型的预测效果。

考虑到个体异质性特征之间存在一定的相关性从而会影响逻辑回归结果的准确性，在接下来的第五章将使用机器学习算法中的随机森林模型对本文的问题进行重新分析。由于随机森林模型具有对解释变量交互性不敏感、预测准确度高、泛化能力强等优点，但解释变量的影响程度难以量化分析，因而可与有序多分类逻辑回归模型互为补充。本章将首先对随机森林模型的原理进行介绍，再使用与逻辑回归部分相同的数据和变量训练建模，最终基于构建的模型进行预测分析并与逻辑回归部分进行对比。

最后的第六章是对全文的总结。

1.4　创新与不足

1.4.1　创新与贡献

本文的创新和贡献之处可能在于以下几个方面：

首先，在选题层面，本文着眼于探究个体投资者的异质性特征对证券投资收益的影响，这方面的已有研究还较少且没有得到一致的结论；

其次，在理论层面，本文提出的理论假说结合了行为金融学的过度自信、过度反应及心理账户理论，并通过实证回归方法进行逐个检验；

此外，在数据层面，本文使用的 2015~2016 年 CCTV“中国经济生活大调查”问卷数据具有样本覆盖面广、数据量大、可靠性强等特点，对构成我国证券市场参与主体的中低收入个体投资者群体具有良好的代表性，同时所处的时间区间涵盖我国股市历史上罕见的股灾时期，有助于发现个体异质性特征对证券投资收益的影响；

最后，在实证研究方法层面，本文在使用有序多元逻辑回归模型充分挖掘数据信息的基础上，创新性地进行了预测准确度评估，有助于提高学界对计量经济学模型应用价值的重视程度，并进一步地考虑到解释变量间可能的交互作用和为提高模型的预测效果，探索性地使用了机器学习中性能良好的随机森林算法进行建模和预测评估，与逻辑回归模型优势互补，为学界和业界的实际应用提供启发和有益参考。

1.4.2　不足之处

当然，本文也存在一定的不足之处。首先，由于行为金融学尚未形成一个完整有机的理论体系，同时作者在心理学、行为金融学等方面的知识还存在欠缺，本文在第二章选取的行为金融学分支理论可能并不全面，假说设定也可能存在一定缺陷；其次，本文使用的数据为横截面数据，而更为理想的数据为面板数据，

但受限于大样本问卷追踪调查的现实可操作性，目前尚缺乏后续年度的数据，可在今后数据可得性的基础上进一步使用面板数据进行研究；最后，受限于篇幅，本文没有进一步探讨收入预期、房价预期等因素对个体证券市场投资收益的影响。这些方面也是未来进一步研究和完善的方向。

第二章　理论分析与假说设定

2.1　行为金融学理论

在实证探究个体异质性特征对证券投资收益的影响之前，本文先对相关的行为金融学理论进行简要介绍，然后在此基础上，结合有关文献，提出个体异质性特征影响证券投资收益的理论假说。

传统的有效市场理论认为，证券投资者应当是完全理性的，其投资决策和行为应当是在最大化自身效用的基础上作出的，但是现实中的情形与此并不相符，在金融市场尚不发达的发展中国家尤其如此。为此，许多学说在修正传统有效市场理论的基础上发展起来，其中行为金融学（Behavioral Finance）是最重要的一个理论，它结合了金融学、心理学、社会学和实验经济学等多学科方法，至今还没有完整有机的理论体系，从整体来看，其研究对象主要集中在 3 个层面——有限理性个体、群体行为和非有效市场。1979 年，Kahenman 和 Tversky 创立展望理论（Prospect Theory），其成为行为金融学的重要理论基础。1985 年，行为金融学奠基人塞勒（Thaler H R，2017 年诺贝尔经济学奖得主）与 De Bondt 发表论文“Does the Stock Market Overreact?”，揭开了行为金融学研究的新开端，随后 2002 年，Barberis 与 Thaler 对行为金融学的有关理论进行了系统的阐释和细分，并尝试提出了行为金融学的理论框架，在此框架下，有关有限理性个体投资行为的重要理论包括过度自信、过度反应、心理账户等。

2.1.1　过度自信（Overconfidence）

过度自信是一种普遍的心理学现象，对于过度自信的表现和定义，学界有多重表述：Weinstein（1980），Frank（1935）和 Taylor、Brown（1988）等研究发现，人们常常高估自己的能力；Wolosin 等（1973）研究发现，人们总是倾向高估自身的知识及能力水平和其对成功几率的影响程度；Fischhoff 等（1977）发现人们通常高估自身所掌握知识和信息的准确性，实际只有 80%可能性发生的事件会认为必然发生，而实际不可能发生的事件会认为有大约 20%的可能性发生；Mahajan（1992）认为过度自信是人们决策时高估一些事件的行为，并在该事件实际发生后会认定自己估计的正确性从而进一步巩固增强这种心理行为。

过度自信在证券市场投资者中普遍存在，直观表现为大多数投资者认为凭借自身能力可以实现超越市场平均水平的投资收益，但却往往事与愿违。Griffin 和 Tversky（1992）研究发现了过度自信的“难度效应”，即当面临难度较大问题时，投资者较容易产生过度自信的现象。Zakay 和 Tuvia（1998）发现投资者

越快确定投资决策和行为，则对该决策正确性的自信心就越大。

2.1.2　过度反应（Overreaction）

过度反应是常体现在证券市场投资者身上的一种非理性行为，主要由投资者在不确定情况下的系统性心理认知偏差所导致。De Bondt 和 Thaler（1985,1987）研究发现，股票市场上的大多数投资者对其未预期到的重大事件或信息呈现出过激或过度的反应，导致股市的超涨和超跌，并认为这一行为缘起于 Kahenman 和 Tversky（1974）所提出的“代表性直觉”理论，即投资者在进行未来预测时，倾向于对最近信息赋予高过较远信息的权重，而不是在考虑整体表现的基础上遵循贝叶斯规则来做出合理反应，从而使其对最近的信息过度反应，预测的结果也会与现实产生较大偏差。比如，当某支证券超常下跌时，投资者会高估这一状态持续的时间而大量卖出，进而推动价格进一步下跌，等到投资者基于新的信息意识到之前的预计过于悲观而重新买入时，价格又会超常上涨，最终价格被修正到合理区间。

后续的国内外诸多研究表明，过度反应在国内外股票市场上普遍存在（陈国进、范长平，2006）。

2.1.3　心理账户（Mental Account）

心理账户理论是指人们习惯于在头脑中把资金按照来源、用途等划分为不同的类别而形成的“思维账户”，分账户进行预算和支出决策；且对不同账户中资金的风险偏好不同，对于归入保值账户的资金往往风险厌恶，而对于归入升值账户的资金则风险厌恶程度较弱甚至风险偏好。由于心理账户的存在，在现实生活中，人们会同时购买了保险和彩票（Friedman M & Savage J L，1948）。心理账户的突出特征是非替代性（Thaler H R，1985），即不同账户之间的资金不能完全替代和转移，这又可以细分为三个方面（李爱梅，凌文辁，2007）：为不同来源的财富而设立的账户之间，为不同消费项目而设立的账户之间，不同存储方式的账户之间。

在投资决策领域，心理账户突出表现于投资组合配置方面，Shefrin 和 Statman（2000）认为包括个人及基金等投资组合管理公司在内的投资者倾向于把资金分为安全账户（保障财富水平）与风险账户（用于风险投资提高收入），且不同账户之间的资金较难流动。

2.2　假说设定

结合前述行为金融学理论，将个体不同异质性特征对证券投资收益影响的假

说设定如下。

2.2.1　受教育程度

一方面，受访者的受教育程度与其所具有的证券市场投资知识紧密相关。Vissing-Jørgensen（2002）研究发现，更高的受教育水平使居民更易学习和理解股票投资知识，从而推动其参与股票投资。另一方面，有研究发现拥有证券投资知识的投资者和证券分析师常常在证券投资中过度自信，但自信程度与成功投资的几率并不相关（黄莲琴，2009）；Ben-David 等（2007）的研究也发现，过度自信程度会随受教育程度的提高而提高。因而，当个体受教育程度越高进而掌握的证券投资知识越多时，我们认为这会对个体的证券投资收益产生两个相反的影响：一方面，丰富的投资知识有助于个体作出更合理的投资决策，从而使其在股市普遍下跌的情形下更可能获得较高的投资收益；但另一方面，拥有较多的投资知识也可能会放大个体的过度自信，从而在股市急速下跌时容易错过减持的良好时机而由赢利转为亏损。由此提出下面的假说 1。

H1：受教育程度越高，越有助于提高获得赢利的可能性，但不一定有助于减小发生亏损的概率。

2.2.2　性别

虽然有少数研究认为性别与风险偏好并没有必然联系（De Mel 等，2009；Booth A L & Nolen P，2012），但已有研究中的大部分均认为女性比男性更加厌恶风险。Hartog 等（2002）通过计算 Arrow-Pratt 风险测度数据发现，女性规避风险的倾向更加明显；Watson 和 McNaughton（2007）基于澳大利亚一大学养老基金数据，发现在控制收入和年龄因素后，女性仍有更强的风险规避倾向，偏好低风险养老金；周业安等（2013）总结已有研究后发现，抽象博彩实验研究得出女性相比男性更加风险厌恶的结论。本文认为，当股市大幅下跌时，厌恶风险的女性由于更倾向利用近期的信息来预测未来而产生更强烈的过度反应，提前止损而降低了发生极端亏损的可能性，同时，在股市普遍下跌的情况下风险厌恶也有利于提高获得赢利的可能性。由此提出本文的假说 2。

H2：女性相比于男性，在股市大幅下跌时，发生极端亏损的可能性较低而获得赢利的可能性较高。

2.2.3　年龄

有关投资者年龄与投资收益关系的已有研究还较少且没有得到一致的结论。赵振华等（2010）基于 2008 年底基金投资调查问卷数据，使用分组分析和多元分类评定模型，发现投资者年龄对投资收益没有显著影响。窦松博（2016）使用

某证券公司营业部提供的个人投资者数据，发现 30-40 岁之间个体的赢利均值最高，而随着年龄增长，由于综合分析能力和思维活跃度下降，证券投资收益率也呈现下降的趋势。尹志超等（2014）使用 CHFS 数据进行实证研究发现，投资经验的增加有助于个体投资者在证券投资中获得赢利。谭松涛和陈玉宇（2012）使用某证券营业厅数据，实证发现随着投资经验的增加，个体投资者的证券投资收益呈现显著上升的状态，且不受年龄、性别的影响。

结合本文所调查的受访者年龄段划分，我们认为，相比于 18~35 岁青壮年群体，年龄越大，投资经验可能更加丰富但同时分析能力和思维活跃度会出现下降，这两方面的因素呈现相互抵消的关系。由此提出下面的假说 3。

H3：投资者年龄对投资收益的影响并不显著。

2.2.4　职业

职业选择与个体的风险偏好有显著关联，冯•诺依曼和摩根斯坦的预期效用最大化理论认为，风险偏好的人倾向于选择高收入高风险的工作，而风险厌恶的人偏好选择低收入低风险的工作，这在诸多实证研究中也得到了印证。Bellante 和 Link（1981）使用 Michigan PSID 数据结合保险、安全带使用及喝酒吸烟情况等构造的风险回避指数，研究发现越厌恶风险的个人越倾向求职于公共部门；Hartog 等（2002）通过实证研究发现公务员较为风险厌恶，而自我雇佣者的风险厌恶程度较低；Cramer 等（2002）和 Ekelund 等（2005）的研究也均发现选择自我雇佣职业的个体更为风险偏好；丁小浩等（2009）使用可控制的仿真实验方法，发现风险偏好的个体通常选择外资企业、民营企业和国际组织的工作，而风险厌恶的个体通常选择政府部门、事业单位和高校工作；廖娟（2011）使用北京市的调查数据进行实证研究发现，风险偏好的个体选择国有部门工作的概率较低。

对于本文的职业分类来说，相比于进城务工人员，我们认为行政事业单位和企业管理人员更为风险厌恶，而自由职业者更为风险偏好。当股市大幅下跌时，风险厌恶的个体由于更倾向利用近期的信息来预测未来而产生更强烈的过度反应，提前止损而降低了发生极端亏损的可能性，同时在股市普遍下跌的情况下，风险厌恶也有利于提高赢利的概率。由此提出本文的假说 4。

H4：相比于进城务工者，行政事业单位和企业管理人员获得较高投资收益的可能性较高，而自由职业者获得较高投资收益的可能性较低。

2.2.5　家庭收入

家庭收入在很大程度上决定着个体的风险偏好和投资行为。由于证券市场相比于货币市场具有更高的风险，从我国现实来看，只有当家庭收入达到能承受一定风险时，居民才开始投资高风险资产，且对于中低等收入水平的家庭，其可选

择的高风险资产主要就是证券类资产（李潇潇，2015）。随着家庭收入水平的提高，一方面，投资者的风险承受能力会提高，从而显著提高其参与股市的可能性（李涛、郭杰，2009），同时也有更多的资金可用于投资证券资产，有利于投资组合的多样化，便于分散投资风险（王渊等，2016），从而在股市大规模下跌时有利于提高赢利的可能性同时避免极端亏损；另一方面，由于家庭收入较高的家庭常常是较为成功的家庭，从而导致其过度自信程度较高，在发生股灾时可能由于出清止损不及时而转赢为亏，提高了发生小幅亏损的可能性。由此提出假说 5。

H5：随着家庭收入水平的提高，投资者获得更高投资收益的可能性越高，但对降低发生投资亏损可能性的作用不确定。

2.2.6　住房状况

由于我国重房产的特殊国情，家庭住房状况也会影响证券市场个体投资者的风险偏好。李心丹等（2011）认为，与金融资产相比，自住房具有双重属性，一是作为消费品为居民提供居住服务，二是作为投资品具有非流动性特征，且转换为其他资产形式的自有交易成本很高。结合史代敏和宋艳（2006）的研究和心理账户理论，本文认为一方面，在储蓄既定且住房与金融资产都是由家庭储蓄形成的前提下，由于住房投资和证券投资均属于风险性投资而被归入风险心理账户，同时由于在安全账户的资金无法与风险账户的资金形成完全替代，因而在住房上的投资会对证券资产投资形成挤压效应（吴卫星等，2007，2010）。另一方面，对于（城市）自住房而言，由于其具有增值属性，在财富效应的作用下，拥有自住房的投资者又会增加对证券资产的需求（王聪和田存志，2012），因而（城市）自住房与证券资产之间又存在互补关系；对于租房的投资者，由于未来生活的不稳定性较大，因而风险规避倾向较强，不仅较少地投资于证券资产且在股市大规模下跌的情况下也倾向过度反应快速出清止损，这有助于降低发生极端亏损的可能性，同时在普遍亏损的情况下也有助于提高获得赢利的可能性；对于农村住房拥有者而言，由于鉴于我国的现实国情，农村住房还不具备城市房产的增值属性，农村住房与金融风险资产之间仅存在挤压关系而不存在互补关系。由于挤压和互补关系的相对大小关系不确定，由此提出假说 6。

H6：相比于租房，拥有（城市）自有房的投资者在投资收益的概率方面没有显著差异，而拥有农村住房的投资者获得较高投资收益的可能性较低。

第三章　数据来源与变量设定

3.1　数据来源与特点

为探究个体投资者异质性特征对证券投资收益的影响，本文采用 2015~2016 年 CCTV“中国经济生活大调查”（“大调查”）的社会问卷调查数据，该数据为微观截面数据，尚处于未公开状态，但经准许可被用于本次学位论文研究。

问卷设计方面，“大调查”问卷由 CCTV 财经频道联合北京大学国家发展研究院、国家统计局中国经济景气监测中心等研究机构共同设计，并与国家统计局、中国邮政共同推出。自 2006 年开始，“大调查”已连续 10 年每年年底组织 1 次，是国内覆盖最广的民间调查（李潇潇，2015）。问卷中的问题在保证延续性和可比性的同时，每年不断充实，从 2006 年的 8 道扩展至 2015 年的 19 道。2015~2016 年“大调查”于 2015 年底展开，除主要包括 2015 年已有确定性结果的问题外，还包括了受访者对 2016 年的预期性问题，并对每位受访者个体层面的社会属性特征进行了统计，如受教育程度、性别、年龄、职业、收入等。相比于 2014~2015 年“大调查”，2015~2016 年“大调查”的调查时间区间覆盖了我国证券历史上难得一见的 2015 年股灾，并相应地新增了受访者证券市场投资收益的问题，为本文提供了研究基础。

问卷调查对象和问卷投放方面，国家统计局严格按照抽样原则筛选出分布在全国 31 个省、自治区和直辖市、104 个城市、300 个县中的 10 万个受访家庭，通过在免邮资明信片上印制调查问卷、借助覆盖广泛的邮政网络进行投递的方式投放，还通过开设绿色投递通道和由邮递员直接担任调查员等方式保证问卷调查投放的高时效性和高回收率。历年来这种调查方式的问卷回收率始终都超过 80%，其中 2015~2016 年“大调查”共回收问卷 88415 份。

总体来看，相比已有研究的数据，2015~2016 年“大调查”的数据具有时间节点独特、样本覆盖面广、数据量大、时效性高、可靠性强等特点，是本文研究的优势之一。

3.2　数据统计性描述

为了更直观地展示问卷中受访者的整体状况，表 1 列示了受访者的社会属性特征情况。

表格 1 全样本受访者的基本资料构成情况

基本资料	选项	占比	基本资料	选项	占比
受教育程度	小学及以下	7.73%	**年龄**	18～25 岁	11.74%

	中学及中专	41.47%		26～35 岁	30.65%
	大专	35.00%		36～45 岁	37.45%
	本科	13.66%		46～59 岁	16.86%
	硕士	1.40%		60 岁及以上	3.13%
	博士	0.68%			
性别	男	57.40%	**住房状况**	自有房（大产权）	30.04%
	女	40.62%		自有房（小产权）	33.25%
				农村住房	20.26%
				自租房	5.32%
				公租房	10.38%
家庭收入	1 万以下	7.01%	**职业**	城市户籍企业职工	25.23%
	1～2 万	13.74%		企业管理人员	16.10%
	3～4 万	21.16%		行政事业单位人员	16.06%
	5～6 万	21.84%		进城务工人员	14.83%
	7～8 万	15.04%		自由职业者	10.01%
	9～10 万	10.60%		务农农民	6.38%
	11～15 万	5.59 %		待业/失业	3.25%
	16～20 万	2.75%		在校学生	3.10%
	21～30 万	1.13%		离退休人员	2.28%
	30 万及以上	1.14%			
常住地	城市	65.69%	**婚姻状况**	未婚无恋人	11.49%
	农村	30.73%		未婚有恋人	13.82%
				已婚	68.53%
户口所在地	城市	59.04%		离异	4.41%
	农村	39.62%		丧偶	1.52%

经过观察，可以发现问卷的受访者有如下的特征：

（1）受教育程度以中等及偏上学历为主，尤其集中在中学至大专，合计占比 76.47%，而硕士和博士高学历受访者仅占 2.08%；

（2）受访者向男性倾斜，男性受访者占 57.40%，而在国家统计局 2015 年人口普查数据中，我国男性人口比例为 51.22%；

（3）年龄结构偏中青年人群，26～45 岁的受访者合计占比为 68.10%，60 岁及以上的只有 3.13%；

（4）受访者职业主要分布于城市户籍企业职工、企业管理人员、行政事业单位人员、进城务工人员和自由职业者等，合计约占 82%，而剩下约 18%的受访者的职业包括务农农民、待业/失业、在校学生和离退休人员等；

（5）家庭收入的分布主要集中在中低收入，其中 2 万元及以下的受访者占比超过 20%，6 万元及以下的占比超过 60%，20 万元以上仅占 2.27%，这表明问

卷调查中参与证券投资的受访者可以在统计意义上被认定为我国证券市场中的小散户，与本文所要研究的个体投资者群体较契合；

（6）家庭住房状况分布较分散，以自有房（大、小产权房）和农村住房为主，合计占比达 84.3%；

（7）常住地以城市为主，超过 65%的受访者来自城市；

（8）接近 60%受访者的户口所在地为城市，且超过 68%的受访者为已婚。

由以上特征可知，问卷受访者的构成与我国人口整体的现实情况存在一定偏差，如文化程度偏低、男性和中青年较多、收入偏中低水平、城市人口偏多等，但由于一方面本文所关注的对象是我国参与证券市场投资的个体投资者，而这些群体在现实中就与我国人口的总体情况存在一定偏离；另一方面，“大调查”问卷受访者具有构成我国证券市场参与主体的中低收入个体投资者群体的突出特征，与本文的研究对象较为契合；此外，考虑到财力、物力及时间等因素的限制，尽管“大调查”受访者不能完全代表我国人口的整体全貌，但对于本文的研究内容来说已经是难能可贵的优良数据。

3.3　变量设定

3.3.1　被解释变量

本文关注的被解释变量是我国个体投资者在 2015 年证券市场（股市和基金）的投资收益情况，这一问题在 2015 年首次被纳入“大调查”问卷内容，对应于问卷中的“问题 9”，是依据亏损/赢利状况进行排序的有序选择数据，其中“0”代表不炒股，“1”至“8”分别代表“亏损 50%以上”、“亏损 20%-50%”、“亏损 20%以内”、“不赔不赚”、“赢利 20%以内”、“赢利 20%-50%”、“赢利 50%-100%”和“赢利一倍以上”。我们将这一数据设定为排序虚拟变量securities，并对应分别赋值 0-8。

图 2 显示了受访者 2015 年证券市场投资收益的分布情况，可以发现有约 39%的受访者没有参与证券市场投资（图 2 左）。而在剩余参与证券市场投资的群体中，绝大部分受访者都没有实现赢利（图 2 右），其中不赔不赚的受访者占比最高，约 28%；其次为亏损 50%以上的受访者，约 21%；实现赢利的仅占 16%，且其中的半数仅赢利 20%以内。整体来看，参与证券市场投资的群体中，亏损：不赔不赚：赢利的比例约为 6：3：1，这一分布情况也与我们的现实观察较一致。

由于问卷中“问题 9”为缺失值的个体仅占总样本的 6.12%，且后文有序多分类逻辑回归模型对缺失值较为敏感，故在进行实证分析之前，我们删除了“问题 9”为缺失值的个体，剩余的样本数为 83,000。

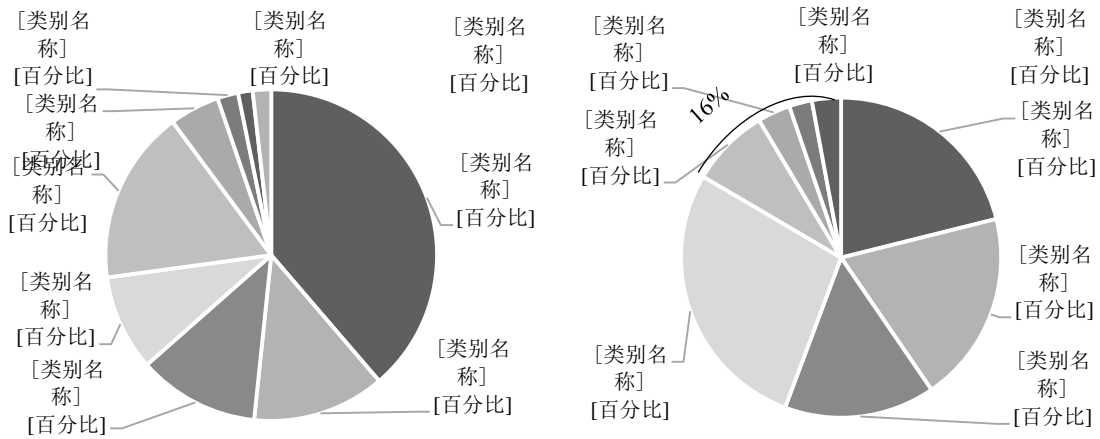

图 4 问卷受访者证券市场投资收益状况的分布情况（左：整体；右：炒股群体）

3.3.2　解释变量和控制变量

基于前文设定的理论假说和可得的问卷数据，本文关注的解释变量如下：

（1）受教育程度（education）：“大调查”问卷中的选项为“小学及以下、中学及中专、大专、本科、硕士和博士”，是一个类别变量（Categorical Variable）由于“大调查”中属性数据较多且若直接使用类别变量进行回归，结果解释存在困难，因而参照尹志超等（2014）的处理方法，将受教育程度折为所接受的教育年限均值，并依次赋值 6、12、15、16、19 和 21。

（2）性别（sex）：将男性受访者作为对照组，赋值 0，女性赋值 1。

（3）年龄（age）：参照 Tin（1998）的分类，将 18~35 岁受访者作为对照组，其他组别分别单独设为 0-1 虚拟变量，并赋值 1。

（4）职业（career）：将进城务工人员作为对照组，而其他的主要职位分别单独设置 0-1 虚拟变量，并赋值为 1。

（5）家庭收入（hincome）：问卷将家庭收入按收入升序划分为 10 个选项并分别用数字 0～9 来代表。与受教育程度变量类似，本文将家庭收入由类别变量折为以每一类收入选项均值为代表的数值变量，即“1 万以下”赋为 0.5，“1～2 万”赋为 1.5，“3～4 万”赋为 3.5，以次类推，考虑到个体投资者的家庭收入集中在中低收入水平，所以将最后一类选项“30 万以上”赋为 35。在以属性数据为主的“大调查”数据结构中，受教育程度和家庭收入变量的转换有助于增加更多的数值信息，对本文的实证研究有很大必要性。

（6）住房状况（house）：参考李潇潇（2015），将问卷中“自租房”和“公租房”合为“租房”变量，并以此为参照组，其他 3 个组别分别单独设为 0-1 虚拟变量，并赋值 1。

此外，为了控制地区差异，本文还纳入了常住地（residence）作为控制变量：将常住在城市的受访者作为对照组，赋值 0，而常住在农村的赋值 1。

表格 2 主要变量的基本统计结果

变量名	变量类型	观测数	均值	标准差	最小值	最大值
证券市场投资收益 securities	虚拟变量	83000	1.9368	2.0530	0	8
受教育程度 education	数值变量	82605	13.3107	2.7877	6	21
性别 sex（男性为对照组）	虚拟变量	83000	0.4333	0.4955	0	1
年龄 age（18~35 岁为对照组）						
36~45 岁	虚拟变量	83000	0.3262	0.4688	0	1
46~59 岁	虚拟变量	83000	0.1437	0.3508	0	1
60 岁及以上	虚拟变量	83000	0.0269	0.1618	0	1
职业 career（进城务工人员为对照组）						
行政事业单位人员	虚拟变量	83000	0.1658	0.3719	0	1
企业管理人员	虚拟变量	83000	0.1669	0.3729	0	1
城市户籍企业职工	虚拟变量	83000	0.2563	0.4366	0	1
自由职业者	虚拟变量	83000	0.1007	0.3009	0	1
家庭收入 hincome	数值变量	82458	6.2788	5.3784	0.5	35
住房状况 house（租房为对照组）						
自有房（大产权）	虚拟变量	83000	0.2956	0.4563	0	1
自有房（小产权）	虚拟变量	83000	0.3281	0.4695	0	1
农村住房	虚拟变量	83000	0.1960	0.3970	0	1
常住地 residence（城市为对照组）	虚拟变量	79318	0.3174	0.4655	0	1

第四章　有序多分类逻辑回归模型

4.1　模型设定

本节将基于 2015~2016 年“大调查”数据，采用经典的计量经济学方法对本文所探究的问题进行实证分析。鉴于本文被解释变量（证券市场投资收益 securities）为离散数据，因而不能使用适用于连续变量的传统 OLS 方法，而需要使用离散选择模型（Discrete Choice Model）。而按照被解释变量取值数量的不同，离散选择模型又可以分为二值选择模型和多元选择模型，其中前者是指被解释变量只有两种取值的模型，而后者是指被解释变量至少有三种取值的模型。本文的被解释变量有“0-8”共 9 种取值，虽然也可以将其组合成涨、跌两类变为二值数据进而使用二值选择模型对其进行实证分析，但鉴于使用多元选择模型不仅往往可以得到与二值选择模型基本一致的回归显著性及符号方向（王小勇，2011），更由于捕捉到不同涨跌情况之间的信息差异，而可以在后续的边际效应分析和预测分析中提供更为精细的结果，从而为现实应用提供更有价值的参考，因而本文将使用更为合适的多元选择模型进行实证分析。

常用的多元选择模型包括无序选择模型（Unordered Choice Model）和有序选择模型（Ordered Choice Model）两种。二者在直观上常通过观察被解释变量是否存在排序关系来区别，而在实证研究中，常通过平行性检验来判别。平行性检验是指检验在不同被解释变量水平的回归模型中，各个解释变量的系数是否相等，若统计意义上不能拒绝检验（即满足平行性）则应使用有序选择模型。依据对随机扰动项分布假设的不同，有序选择模型又可以细分为有序多分类概率比回归模型和有序多分类逻辑回归模型，由于在通常情况下，两模型的结果较为接近且后者的估计参数更容易解释，因而实际中常用有序多分类逻辑回归模型。

鉴于本文使用的证券市场投资收益变量本身具有排序关系，同时使用 STATA 软件中的 brant 命令发现平行性检验不能被拒绝，因而选择使用有序多分类逻辑回归模型进行实证研究。本文有序多分类逻辑回归模型的具体设定如下：

$$y_i^* = \beta X_i + \varepsilon_i, \varepsilon_i \sim Logistic\ Distribution$$

$$y_i = \begin{cases} 0, & if \quad y_i^* \le r_0 \\ 1, & if \quad r_0 < y_i^* \le r_1 \\ 2, & if \quad r_1 < y_i^* \le r_2 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ 6, & if \quad r_5 < y_i^* \le r_6 \\ 7, & if \quad r_6 < y_i^* \le r_7 \\ 8, & if \quad y_i^* > r_7 \end{cases} \tag{1}$$

其中X_i是前文设定的解释变量和控制变量，各变量之间应无多重共线性；β为待估参数，ε_i为随机扰动项，在有序多分类逻辑回归模型中假定满足 Logistic 分布；y_i^*是不可观测的潜在变量或隐变量（Latent Variable）；$r_0 < r_1 < \cdots r_7$为切点或门限值（Cutoff Point），均为待估参数且应独立于β，由于其取值对本文没有显著意义，故在后文回归结果中不予汇报；$(r_0, r_1, \cdots, r_7)$与y_i^*的关系将决定可观测的被解释变量y_i（即$securities_i$）被归入的排序值。综上，y_i各种取值的概率可由下列式子决定：

$$\begin{cases} \mathrm{P}(y_i = 0|X_i, \beta, \mathrm{r}) = F(r_0 - \beta X_i) \\ \mathrm{P}(y_i = 1|X_i, \beta, \mathrm{r}) = F(r_1 - \beta X_i) - F(r_0 - \beta X_i) \\ \cdots\cdots \\ \mathrm{P}(y_i = 8|X_i, \beta, \mathrm{r}) = 1 - F(r_7 - \beta X_i) \end{cases} \quad (2)$$

其中$F(\cdot)$为ε_i的累积分布函数，在此即逻辑分布函数。然后写出似然函数并使用 MLE 求解即可得到各待估参数的估计值。

总之，有序多分类逻辑回归模型实际上得到的是可观测的被解释变量被归入不同排序值的概率，因而回归模型的参数估计值只能反映各解释变量对被解释变量的相对影响程度而不是数量上的绝对影响，仅能提供显著性和符号方向方面的有用信息。而若要阐释各解释变量对被解释变量在数量上的绝对影响程度，需要另外求解边际效应，即当其他解释变量不变时，某解释变量变动 1 单位对被解释变量的边际概率影响。按照变量所处数值的不同，常用的边际效应又有平均边际效应、样本均值处的边际效应和在某代表值处的边际效应这三种，其中平均边际效应（Average Marginal Effect）为 STATA 默认使用的方法，最有代表意义（陈强，2014）且实际应用最广（Green，2017），是通过计算在每个观测值上的边际效应再进行简单算术平均得到的。设总观测数为 n，则变量 k 的平均边际效应的具体设定如下：

$$\begin{cases} \dfrac{\partial \mathrm{P}(\hat{y} = 0|\mathrm{X})}{\partial X_k} = \dfrac{1}{n} * \sum_{i=1}^{n} \left(-F\left(r_0 - \hat{\beta} X_i\right) * \widehat{\beta_k}\right) \\ \dfrac{\partial \mathrm{P}(\hat{y} = 1|\mathrm{X})}{\partial X_k} = \dfrac{1}{n} * \sum_{i=1}^{n} \left(F\left(r_0 - \hat{\beta} X_i\right) * \widehat{\beta_k} - F\left(r_1 - \hat{\beta} X_i\right) * \widehat{\beta_k}\right) \\ \cdots\cdots \\ \dfrac{\partial \mathrm{P}(\hat{y} = 8|\mathrm{X})}{\partial X_k} = \dfrac{1}{n} * \sum_{i=1}^{n} \left(F\left(r_7 - \hat{\beta} X_i\right) * \widehat{\beta_k}\right) \end{cases} \quad (3)$$

4.2　模型回归结果

4.2.1　整体回归结果

使用 STATA14.0 软件，有序多分类逻辑回归模型的回归结果如下表 3 所示。

我们首先考虑了不加入家庭收入和住房状况的模型，即模型一；之后在模型二中又加入了家庭收入及其平方项，以探究是否存在收入的倒 U 形关系；而在模型三中又进一步加入了住房状况变量。

表格 3 个体异质性特征与证券市场投资收益状况的有序多分类逻辑回归模型结果

	模型一	模型二	模型三
受教育程度	0.0290***	0.0164***	0.0154***
	(12.59)	(6.84)	(6.41)
性别（男性为对照组）	0.0320**	0.0379***	0.0354***
	(3.00)	(3.55)	(3.31)
年龄（18~35 岁为对照组）：			
36~45 岁	0.0204	0.0087	0.0084
	(1.42)	(0.60)	(0.58)
46~59 岁	0.0556**	0.0316	0.0308
	(2.93)	(1.65)	(1.62)
60 岁及以上	0.1070**	0.0674	0.0585
	(2.66)	(1.68)	(1.45)
职业（进城务工人员为对照组）：			
行政事业单位人员	0.1680***	0.1410***	0.1160***
	(8.66)	(7.23)	(5.84)
企业管理人员	0.2880***	0.2630***	0.2330***
	(15.12)	(13.75)	(11.88)
城市户籍企业职工	0.0259	0.0176	-0.0096
	(1.48)	(1.01)	(-0.53)
自由职业者	-0.1590***	-0.1810***	-0.1850***
	(-6.66)	(-7.55)	(-7.71)
家庭收入		0.0251***	0.0248***
		(8.49)	(8.36)
家庭收入平方项		0.0001	0.0001
		(-4.27)	(-4.33)
住房状况（租房为对照组）：			
自有房（大产权）			-0.0080
			(-0.40)
自有房（小产权）			0.0017
			(0.09)
农村住房			-0.1650***
			(-7.77)
常住地（城市为对照组）	0.0947***	0.1090***	0.1280***
	(10.44)	(11.89)	(13.61)
Observations	81521	81190	81190
Pseudo R-squared	0.003	0.004	0.004

注：括号内是系数估计值的标准差；***表示 p<0.001, ** 表示 p<0.05, *表示 p<0.1。

由前述，表 3 的回归结果仅能在显著性和符号方向方面提供有用的信息，因而为了探究各解释变量对被解释变量在数量上的绝对影响程度，还进一步求解了各解释变量的平均边际效应。下表 4 仅列示了在表 3 中显著的解释变量在被解释变量分别取“1~8”时的平均边际效应结果，从整体来看，下表 4 中各解释变量的平均边际效应均在 1%的显著性水平上显著，与表 3 中的显著性相一致；且对每一个解释变量来说，从平均边际效应符号来看，在证券市场投资收益亏损 50%以上时，其与在其他收益状况下的均相反，而从平均边际效应绝对值来看，在不赔不赚情况下的绝对值均最大。

表格 4 模型三显著解释变量的平均边际效应

证券市场投资收益	受教育程度	性别	职业			家庭收入	住房状况
	年	1=女	1=行政事业单位	1=企业管理	1=自由职业者	万	1=农村住房
亏 50%以上	-0.0002***	-0.0004***	-0.0013***	-0.0026***	0.0021***	-0.0015***	0.0019***
亏 20%-50%	0.0003***	0.0006***	0.0020***	0.0042***	-0.0033***	0.0023***	-0.0029***
亏 20%以内	0.0005***	0.0012***	0.0038***	0.0077***	-0.0061***	0.0042***	-0.0054***
不赔不赚	0.0016***	0.0037***	0.0112***	0.0245***	-0.0195***	0.0135***	-0.0174***
赢利 20%以内	0.0006***	0.0015***	0.0048***	0.0096***	-0.0077***	0.0053***	-0.0068***
赢利 20%-50%	0.0003***	0.0006***	0.0021***	0.0043***	-0.0034***	0.0023***	-0.0030***
赢利 50%-100%	0.0002***	0.0005***	0.0015***	0.0031***	-0.0024***	0.0017***	-0.0022***
赢利一倍以上	0.0003***	0.0006***	0.0020***	0.0041***	-0.0032***	0.0022***	-0.0029***

注：***表示 p<0.01, ** 表示 p<0.05, *表示 p<0.1。

4.2.2 回归结果分析

在参考表 3 显著性和参数符号方向的基础上，基于表 4 的平均边际效应结果，下面将对不同异质性特征影响个体证券投资收益的情况进行逐个分析。在没有特殊说明的情况下，对单一解释变量平均边际效应结果进行分析时，均假定其他解释变量及控制变量保持不变。

1、受教育程度

首先，受教育程度在表 3 的三个模型中均显著为正，说明受教育程度会对个体证券投资收益产生显著的正向影响。而从表 4 的平均边际效应来看：亏损 50%以上的平均边际效应为-0.0002，说明接受教育的年数每增加 1 年，证券市场投资亏损超过 50%的可能性降低 0.02%；亏损 20%-50%的平均边际效应为 0.0003，说明接受教育的年数每增加 1 年，证券市场投资亏损 20%-50%的可能性增加 0.03%；不赔不赚的平均边际效应为 0.0016，说明接受教育的年数每增加 1 年，投资不赔不赚的可能性增加 0.16%；赢利 20%以内的平均边际效应为 0.0006，说明接受教育的年数每增加 1 年，赢利 20%以内的可能性增加 0.06%；以此类推，赢利一倍以上的平均边际效应为 0.0003，说明接受教育的年数每增加 1 年，投资赢利一倍以上的可能性增加 0.03%。由此可见，受教育程度提高有助于证券投资知识的学习，有利于显著增加投资于证券市场获得各种赢利情况的可能性，尤其是提高赢利 20%以内的概率，并有效降低发生极端亏损（亏损 50%以上）的可能性；但另一方面，也会在一定程度上由于过度自信的作用，导致在股灾中出售股票不及时，进而提高亏损 50%以内的可能性。

2、性别

投资者性别在表 3 的三个模型中均显著为正，说明男女性别投资者的证券市场投资收益在统计意义上存在显著差异，且女性获得较高投资收益的概率更高。而从表 4 的平均边际效应来看，亏损 50%以上的平均边际效应为-0.0004，说明女性相比于男性，证券市场投资亏损超过 50%的可能性降低 0.04%；亏损 20%-50%、亏损 20%以内的平均边际效应分别为 0.0006 和 0.0012，说明女性相比于男性，证券市场投资亏损 20%-50%、亏损 20%以内的可能性分别提高 0.06%和 0.12%；以此类推，不赔不赚、赢利 20%以内、赢利 20%-50%、赢利 50%-100%和赢利一倍以上的平均边际效应分别为 0.0037、0.0015、0.0006、0.0005 和 0.0006，说明女性相比于男性，证券市场投资不赔不赚和各种赢利情况的可能性均较高，分别提高 0.37%、0.15%、0.06%、0.05%和 0.06%。总之，女性除了发生极端亏损（亏损 50%以上）的可能性要比男性低 0.04%外，获得各种赢利情况和亏损 50%以内情况的可能性均高于男性。由于女性投资者对风险较为厌恶，在股市大幅下跌时往往呈现过度反应，急于止跌而出清股票，这虽然降低了发生极端亏损的可能性，但也导致发生 50%以内亏损可能性的提高。而女性投资者获得各种赢利情况的可能性均较男性投资者高则很可能与证券市场投资者普遍亏损有关，由前文数据统计性描述可知，2015 年仅有 1 成的证券市场参与者获得了赢利，在这种大背景下，风险厌恶的女性比风险偏好的男性更可能获得赢利。

3、年龄

投资者年龄方面，由于表 3 中除了模型一外，模型二和三中的 3 个年龄虚拟

变量系数均不显著，因而可以认为投资者年龄并不会对证券投资收益产生显著的直接影响。这可能由于随着投资者年龄的增大，一方面投资经验更丰富从而有利于获得较好投资收益，但同时思维活跃度和分析能力也下降从而导致在股市发生异常波动时投资决策易出现较大偏差而影响投资收益，在两方面作用相互抵消的情况下，投资者年龄对证券投资收益的影响不显著。

4、职业

职业方面，从表 3 的回归结果中可以发现，相比于进城务工人员，行政事业单位和企业管理人员虚拟变量的系数均显著为正，自由职业者虚拟变量的系数显著为负；而城市户籍企业职工虚拟变量的系数在 3 个模型中均不显著，原因可能在于城市户籍职工为一类较宽泛的职业，分类内的工作性质差异较大（李潇潇，2015）。从表 4 的边际效应来看，依次从纵向来看，相比于进城务工人员，行政事业单位人员的证券市场投资亏损 50%以上的可能性降低 0.13%，而亏损 20%-50%、亏损 20%、不赔不赚、赢利 20%以内、赢利 20%-50%、赢利 50%-100%、赢利一倍以上的可能性提高 0.20%、0.38%、1.12%、0.48%、0.21%、0.15%、0.20%；企业管理人员相应的情况分别为降低 0.26%，提高 0.42%、0.77%、2.45%、0.96%、0.43%、0.31%、0.41%；自由职业者的情况分别为提高 0.21%，降低 0.33%、0.61%、1.95%、0.77%、0.34%、0.24%、0.32%。从横向来看，在同样的赢亏情况下，企业管理人员的平均边际效应绝对值均约为行政事业单位人员的 2 倍，而自由职业者的平均边际效应绝对值则介于前两者之间。由于相比于自由职业者和进城务工人员，选择作为企业管理人员和行政事业单位人员的投资者往往更为风险厌恶，在股灾发生时易于过度反应提前止损，在降低发生极端亏损可能性的同时，也导致发生 50%以内亏损可能性的提高。同时，在证券投资者普遍亏损的情况下，风险厌恶的企业管理人员和行政事业单位人员比风险偏好的自由职业者和进城务工人员更可能获得赢利。

5、家庭收入

从表 3 来看，模型二和三中，家庭收入变量在 1%的显著水平上显著为正，但家庭收入的平方项均不显著，说明家庭收入水平对证券投资收益有显著的正向影响，且未呈现出倒 U 型的关系，但未呈现倒 U 型也可能是由于本文的收入区间上限未达到倒 U 型的拐点。从表 4 的平均边际效应来看，随着家庭收入水平的提高，证券市场投资亏损 50%以上的可能性降低 0.15%，而亏损 20%-50%、亏损 20%、不赔不赚、赢利 20%以内、赢利 20%-50%、赢利 50%-100%、赢利一倍以上的可能性分别提高 0.23%、0.42%、1.35%、0.53%、0.23%、0.17%、0.22%。由于收入水平提高，投资者的股票投资组合多样性提高，从而有利于尽可能地分散和降低投资风险，在证券市场参与者普遍亏损时，有助于降低发生极端亏损可

能性并提高获得赢利的可能性，但由于过度自信程度较高，发生 50%以内亏损的可能性提高。

6、住房状况

住房状况方面，由表 3 模型三的回归结果可知，相比于租房，大、小产权自有房虚拟变量的系数均不显著，这可能是由于自住房投资对高风险金融资产投资的挤压和互补效应可能相互抵消；而农村住房虚拟变量的系数在 1%的水平上显著为负。由表 4 平均边际效应结果可知，相比于租房，拥有农村住房的证券市场投资者亏损 50%以上的可能性提高 0.19%；而亏损 20%-50%、亏损 20%、不赔不赚、赢利 20%以内、赢利 20%-50%、赢利 50%-100%、赢利一倍以上的可能性分别降低 0.29%、0.54%、1.74%、0.68%、0.30%、0.22%、0.29%。相比于租房者，农村住房拥有者由于未来生活更为确定而风险偏好程度较高，使得拥有农村住房的投资者在股市前期大幅下跌时倾向于继续持有，在提高发生极端亏损可能性的同时也降低了发生 50%以内亏损的可能性，且在证券投资者普遍亏损的情况下，风险更为偏好也使其获得赢利的可能性更低。

4.3　预测分析

4.3.1　从回归到预测

对于截面数据来说，经典横截面计量模型的主要任务是以经济理论为基础，通过回归分析，估计解释变量与被解释变量之间的显著性关联关系，也即探究历史截面数据中的经济规律，并对模型假说进行检验（李子奈，2007；洪永淼，2007）。在经典横截面计量模型的基础上，针对探究微观个体行为和使用微观调查数据的研究，以研究问题为导向，发展出微观计量经济学这一现代计量经济学分支，但实质仍是经典计量经济学的模型理论，只是更多地考虑了模型对所研究问题的适用性（李子奈、刘亚清，2010）。

学界对计量经济学模型的批判一直存在。李子奈（2007）认为，在设定总体计量模型时，由于研究者往往以先验的经济理论为导向而不是以经济现实为导向，使得基于不同的理论会设定出不同的模型，且模型的随机扰动项常常违背高斯-马尔科夫假设等。此外，冯燮刚和李子奈（2006）认为，计量经济学模型的总体设定是基于独立的理性经济人，而忽视了现实经济主体与理性人的差异以及经济主体之间、经济主体与所处环境之间的动力学相互作用，但这一作用过程才是模型中各种经济变量作用和变动的根本原因，因而建立起的计量模型往往与现实产生较大的背离。洪永淼（2007）认为，现代计量经济学模型建立在两大公理之上——“任何经济系统都可以看作是服从一定概率分布的随机过程”和“任何经济

现象都可以看作是这个随机数据生成过程的实现”，而这两大公理本身就决定了任何现代计量模型都无法囊括现实经济中的全部随机因素，因而不可能精确阐释各经济变量间的数量关系。

随着全球进入大数据和人工智能时代，机器学习（Machine Learning）模型又对计量经济学模型产生新的冲击，这突出体现在机器学习模型的建立和选择是以现实为导向，目的是解决现实中的应用问题，并普遍使用预测的方法来评估模型效果。在这样的时代变革背景下，弥补现代计量模型的缺陷对于计量经济学的应用和发展至关重要。洪永淼指出（2007），虽然几乎不能使用经济数据来检验作为计量模型基础的经济理论假设是否与现实相符，但可以退而求其次，通过考察在此假设上建立的计量模型的预测结果与真实数据之间的一致性来判断理论的合理性。但在已有的应用有序多分类逻辑回归模型进行实证分析的研究中，还鲜有见到在进行回归分析和边际效应分析后又进行预测分析的研究。

与已有研究不同，本文将在进行前文回归分析的基础上，更进一步研究，对构建的回归模型进行预测准确度评估，以判断模型预测结果与现实结果的一致性程度，从而评判理论和模型的现实应用价值。这一创新性研究环节不仅有助于提高学界对计量经济学模型应用价值的重视程度，同时也可以为业界的实际应用提供有价值的参考。

4.3.2　预测准确度评估

基于回归得到的模型三，参考张晓峒（2007）、易丹辉（2014）的方法，将通过计算有序多分类逻辑回归模型的预测精度来评估模型。具体过程如下：

（1）基于模型三，使用每位投资者所对应的一组解释变量数据进行回归，得到被解释变量的预测概率；

（2）将预测概率最大值所属类别作为该投资者投资收益的预测类别；

（3）将预测类别与该投资者实际的投资收益类别进行对比，若类别相同则视为分类准确；

（4）遍历所有观测，重复上述（1）-（3）过程，通过累加分类准确的观测数得到最终的分类准确观测数，再与全部观测数求比值即可得到模型的预测准确度。

依照上述方法，可以得到本文构造的有序多分类逻辑回归模型的预测准确度为 37.87%。

表格 5 有序多分类逻辑回归模型预测结果

	观测数	占比
预测类别=实际类别（分类准确）	31,428	37.87%

预测类别≠实际类别（分类不准确）	51,572	62.13%
合计	83,000	100.00%

鉴于本文所使用的被解释变量（证券投资收益变量）有“0-8”共 9 种类别，即进行随机预测的准确度均值为 1/9≈11.11%，所以尽管 37.87%的预测准确度看似较低，但与随机预测的 11.11%相比，仍有超过 26 个百分点的显著提高；而在业界的现实应用中，预测准确度提高 1%的水平就能带来巨大的经济利益和现实意义。总之，由于有序多分类逻辑回归模型比二值选择模型利用了数据的更多信息，在对比模型的预测类别与实际类别后，可以认为，本文构建的有序多分类逻辑回归模型三的预测效果较好。

第五章　随机森林模型

上文的有序多分类逻辑回归模型表明，多个个体异质性特征可显著影响个体的投资收益状况。但考虑到上述异质性特征之间必然存在内在关联，使得有序多分类逻辑回归模型的解释变量需满足的无多重共线性、独立性条件难以完全成立，从而可能导致回归结果存在偏误并降低了预测的精确度。为此，下文将基于相同的数据和变量设定，采用目前在机器学习等领域中被广泛应用并具有对解释变量间交互性不敏感、预测准确度较高、泛化能力强等优点的随机森林模型（Random Forest，“RF”）对本文所关注的问题进行进一步分析，以期在考虑影响个体投资收益状况的异质性特征间的相关性基础上，提高对个体投资收益状况的预测准确度，进而在与前文的有序多分类逻辑回归模型形成互补的同时，也为学界和业界的现实应用提供启发。

概括性地来看，随机森林是一种将决策树这种基分类器（Base Classifier）进行集成学习（Ensemble Learning）或分类器组合（Classifier Combination）所得到的一种组合分类器（Combined Classifier），从而同时使用多个决策树对样本进行训练并预测，能够被应用于分类、回归等多种问题，属于非参数的统计方法。目前，学界已有诸多研究将随机森林与决策树（(Classification and Regression) Decision Tree，“CART”或“DT”）、人工神经网络（Artificial Neural Network，“ANN”）、支持向量机（Support Vector Machine，“SVM”）、逻辑回归（Logistic Regression，“LR”）、 K 近邻（K Nearest Neighbor，“KNN”）等多种机器学习算法进行比较，发现随机森林在预测准确度和外推泛化能力等方面有显著的优势（Ballings M 等，2015）。

由于对于研究分类问题的随机森林模型，其基础（基分类器）是因变量为分类变量的决策树——分类树，因而本章将首先对分类树模型的相关原理进行介绍，随后再从原理和应用两大方面对随机森林模型进行阐释。

5.1　分类树模型的原理

5.1.1　分类树概述

分类树是依据自变量的值进行递归分割（Recursive Partitioning），并使用得到的分类规则来预测因变量分类的一种有监督学习方法。直观来看，分类树是一种类似流程图的树结构。

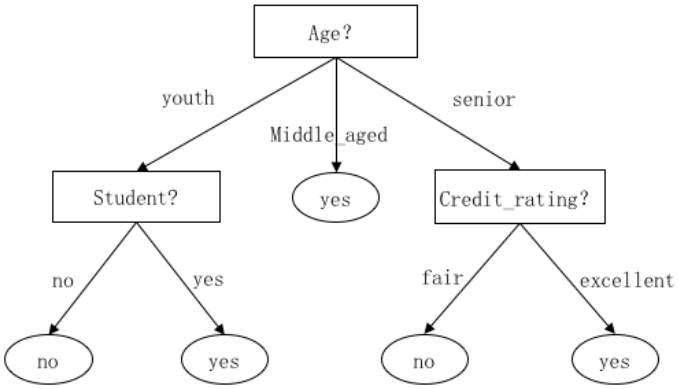

图 5 分类树示意图（是否放贷问题）

如上图有关是否放贷问题的一个简化分类树所示，参考张俊妮（2009）的命名规则：分类树最上层的节点叫做“根节点”，包含所有的观测，通常是所研究问题的一个最重要特征变量的测试；除了最上层节点和最下层节点之外的节点被称为“内部节点”，表示变量的一个测试；连接节点之间的线称作“分支”，表示变量的测试输出；最下层的节点被称作“叶节点”，表示最终的分类结果；分类树从根节点到叶节点的每一条路径代表一个分类规则，利用此规则可以基于新数据进行预测。

5.1.2　分类树的构建与修剪

1、基本思路

在得到用以预测因变量类别的分类规则之前，需要首先得到分类树，而要得到分类树则需要完成分类树的构建和修剪这两大任务。基本思路如下：

（1）首先，将原始数据放入分类树的根节点，并将原始数据随机分成两组，一部分为训练数据，另一部分为测试数据；

（2）然后，使用训练数据来构建分类树，在每个内部节点选择最优特征作为分割的依据，并对最优特征选择最优的划分规则，这个过程又称节点分割（Splitting Node），划分完成后判断下层新节点是否为叶节点，如果不是，则以新节点作为根节点继续建立新的分枝，如果是，则停止分割，最终生成层次和叶节点均足够多的分类树，以保证其对训练数据集的预测精度；

（3）此后，使用测试数据来修剪分类树，选取对测试数据预测性能最好的子树；

（4）最后，将（2）-（3）步不断递归，一直到所有内部节点都是叶节点为止，并对所有叶节点预测分类；

（5）基于上述过程形成的分类树，可从“根节点-每个叶节点”路径萃取出所需的分类规则。

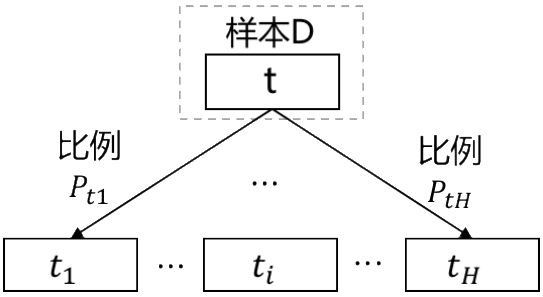

图 6 分类树核心问题图示

2、核心问题

由上述分类树构建和修剪的基本思路，又可以通过细分并提取出如下几个核心问题（张俊妮，2009）。

（1）如何选择作为每个内部节点分割依据的最优特征？

常用的方法包括三种，即信息论（Information Theory）中的信息增益最大（或熵最小）、信息增益率最大和基尼指数最小（张晓玉，2014），这三种方法分别被分类树算法中的 ID3 算法、C4.5 算法和 CART 算法所使用。在每一种方法中，分类树均会把特征变量遍历一遍，然后优先选取能使分类效果达到最优的那个特征变量作为节点分割依据。三种方法的本质均是找到使分类误差率最小的特征变量，而不同之处在于评估分类效果最优（分类误差率最小）的方法。

ID3 算法使用的信息增益（Information Gain）最大或熵（Entropy）最小，从直观来看，是指以该特征变量作为分割依据进行分割后，能使样本不确定性减少最多；从公式来看，如图 6 所示，在样本 D 的内部节点处，目标就是找到使信息增益$Gain()$最大的特征 t：

$$Gain(\mathrm{t}) = Info(\mathrm{D}) - Info_t(D)$$

其中：$Info(\mathrm{D}) = -\sum_{i=1}^{H} p_{ti}\log_2(p_{ti})$，$Info_t(D) = \sum_{j=1}^{m} |D_j|/|\mathrm{D}| * Info(D_j)$，样本 D 根据特征 t 的取值被划分成$D_1$，$D_2$，$D_j$ …，D_m个样本子集，$|D_j|$为样本子集D_j中的样本个数，p_{ti}为依据特征 t 进行分割后样本属于下层类别t_i的比例。由计算公式可知，若某个$p_{ti} = 1$而其它类别比例为 0，则信息增益最大或熵最小。

C4.5 算法使用的信息增益率最大与 ID3 算法使用的信息增益最大类似，只是采取了比值的形式，避免了 ID3 算法中偏向选择类别较多属性的问题。如图 6 所示，在样本 D 上，以特征 t 进行节点分割的信息增益率计算公式如下：

$$Gain_R(t) = Gain(\mathrm{t})/Info(\mathrm{D})$$

CART 算法使用的基尼指数（Gini）最小，从直观来看，是指以该特征变量作为依据进行节点分割后，能使从样本 D 中随机抽取两个子样本时类别标记不一致的概率最小，本质与信息增益最大、信息增益率最大相同，只是通过求解基尼指数来判别分类误差率；从公式来看，如图 6 所示，在样本 D 上，以特征 t

进行分割的基尼指数计算公式如下：

$$Gini(\mathrm{t}) = \sum_{i \neq j} p_{ti} p_{tj} = 1 - \sum_{i=1}^{H} p_{ti}^2 =$$

由计算公式可知，若某个$p_{ti} = 1$而其它类别的比例为 0，则基尼系数最小。

（2）如何对最优特征选择最优划分规则？

由上述最优特征的选取方式可知，最优特征的划分规则同样影响着分类误差率，在选取最优特征的同时也需要选择其划分规则：

首先，寻找所有可能的划分规则，构成候选划分集 S，对于定序变量 t，可将训练数据集中该变量的取值按大小顺序排列，构造不重叠的取值序列$t_1 < t_2 < t_i < \cdots < t_H$，j=1, ⋯;$H$，并将满足$t_{j-1} < \mathrm{t} \leq t_j$的观测划分入第 j 个下层节点；

再根据节点的分类误差率，从候选划分集 S 中选择能使分类误差率下降最多的划分规则，也即最优划分规则。

常用的衡量分类误差率的方法与前面用于选择最优特征的方法相同，即信息增益最大和基尼指数最小。具体而言，若划分前节点的分类误差率为Q(t)，划分后节点的分类误差率为$\sum_{h=1}^{H} p_{th} Q(t_h)$，则最优划分规则为使$\mathrm{Q(t)} - \sum_{h=1}^{H} p_{th} Q(t_h)$最大的候选划分。

以上为选择最优划分规则的基础方法，近年来还衍生出诸多在此基础上进行优化的方法，但限于篇幅不再详述。

（3）如何判断内部节点是否为叶节点？

一般情况下，以下情况之一发生，则相应内部节点已经是叶节点，无法继续进行划分：节点内训练数据的样本数达到某个最小值；树的深度达到一定限制；达到停止分割的最极限状态（该样本的每个数据都已被归到同一类别，或没有办法再找到新的特征变量进行节点分割，或已经没有任何尚未处理的数据）。

（4）为何以及如何修剪分类树？

首先，由于分类树的构建是基于训练数据的，且构建的目标是生成层次和叶节点均足够多的分类树，以达到对训练数据的预测性能最好，但这容易带来较多的训练数据噪音，产生过度拟合。为了提高分类树对新数据的预测性能，因而需要对分类树进行修剪，以达到分类树对训练数据和测试数据预测性能的平衡。

其次，在使用测试数据对分类树进行修剪时，判断预测性能的标准为分类误差率，分类误差率越低，则分类树对测试数据的预测性能越好。具体而言，令$\mathcal{D}$表示测试数据集，$N_{\mathcal{D}}$表示其中的样本数，Y_i和$\widehat{Y_i}$分别表示$\mathcal{D}$中样本 i 的因变量的实际值和预测值，则$\mathcal{D}$的分类误差率为：$1/N_{\mathcal{D}} * \sum_{i=1}^{N_{\mathcal{D}}} \mathcal{I}(Y_i \neq \widehat{Y_i})$。

最后，一般使用“最小成本-复杂度”方法对分类树进行修剪，依次考察树 T 的每个子树，直至修剪至使分类树对测试数据的预测性能最好为止。对树 T 的

以 t∈T 为根节点的任意子树T_t（包含节点 t 及 T 中节点 t 的所有后代节点），定义它的复杂度$|T_t|$为T_t中叶节点的个数，α为复杂度参数，则成本-复杂度度量为：

$$C_\alpha(T_t) = C(T_t) + \alpha|T_t|$$

其中，成本$C(T_t) = \sum_{t_i \in T_t} \tilde{p}_{ti}\, Q(t_i)$为树$T_t$对测试数据的分类误差率。

若$C_\alpha(T_t) > C_\alpha(\mathrm{t})$，则应从 T 中修剪掉分支$T_t$，即只留下分支$T_t$的根节点 t 而把节点 t 的所有后代节点都修剪掉，修剪后的子树为 T-T_t。

（5）如何确定叶节点的预测分类？

在构建和修剪分类树之后、提取分类规则之前，还需要预测每个叶节点的分类，这一过程主要基于概率，即将叶节点 t 归入使训练数据计算出的p_{ti}达到最大的类别 i。

5.1.3　与逻辑回归的区别

通过上述对分类树原理和构建、修剪过程的阐述，可以发现分类树显著区别于逻辑回归，这主要体现在以下几方面（Efron B & Hastie T，2016；张俊妮，2009；李航，2012）。

首先，分类树是逐个特征进行迭代分割，最后再对最终的每个叶节点预测分类并提取分类规则，具有模型图形化的突出优势，更接近人的思维习惯；而逻辑回归是将所有特征变量放入方程右侧，通过线性映射变换为方程左侧的概率后，将满足某一概率阈值的划分为一类，较为抽象化。

其次，分类树内在考虑了解释变量间的相关性和交互性，对变量的量纲、缺失值、异常值均不敏感；而逻辑回归内在假定解释变量间相互独立、不存在多重共线性，且在进行回归之前需要尽可能做变量标准化、处理缺失值或异常值等预处理以保障回归结果的稳健性。

最后，分类树可找到非线性分割，拟合出来的是分区间的阶梯函数；而逻辑回归除非对自变量进行多维映射，否则只能找到线性分割，拟合出的是线性函数。

5.2　随机森林模型的原理

5.2.1　随机森林概述

尽管分类树算法包括了旨在提高对新数据集预测性能的修剪过程，但由于一个分类树模型仅有一棵树，分类结果仍不稳定、泛化能力有限，且分类准确性有赖于样本量的多少，基于庞大样本建构出的分类树的预测分类结果往往符合期望，而基于较小样本量的往往与预期有较大偏离；此外，较复杂的分类树虽然也可提取出分类规则，但由于较繁冗而不易解释（Efron B & Hastie T，2016）。

为了克服分类树的部分缺陷，在借鉴贝尔实验室 Ho（1995,1998）提出的随机决策森林（Random Decision Forests）模型基础上，加州大学伯克利分校 Breiman 教授于 2001 年提出了随机森林（Random Forest）模型，在没有显著增加计算量的情况下大幅提高了预测结果的精度和泛化能力。

随机森林模型有两大核心特征：一是“随机性”（Efron B & Hastie T, 2016）——除了通过装袋算法（Bagging 算法）在训练数据层面引入随机性之外，随机森林还在选择分割变量层面引入了随机性，使单个分类树之间的相关性大大减弱，从而使其预测效果及泛化能力远好于单个分类树模型（Breiman L，2001）；二是“组合性”——这主要通过装袋算法实现，即在多个分类树单独生成各自的预测结果后，通过取单个分类树结果的众数或者通过其他投票机制进行结果的组合平均，从而最终得到随机森林模型的预测分类。

5.2.2　装袋算法

如前所述，随机森林模型在训练数据层面引入随机性和模型的组合性均有赖于装袋算法的应用。装袋算法又称“引导聚集算法”（Bootstrap Aggregating），其同样由加州大学伯克利分校 Breiman 教授提出（1996），其基础是斯坦福大学教授 Efron（1979）提出的自助抽样法（Bootstrap Sampling）。

自助抽样法是一种均匀有放回抽样，常被用于估计统计量方差和进行区间估计，基本形式是在原始样本中进行 n（自己给定）次均匀有放回的简单随机抽样，从而得到样本量为 n 的样本子集。其与蒙特卡洛（Monte Carlo）抽样的显著区别在于，自助抽样法与样本的分布形式无关，而蒙特卡洛抽样需要首先基于样本得到其分布形式。

装袋算法基于上述自助抽样法，首先使用自助抽样法对原始数据集进行抽样，形成多个样本量为 n 的训练数据集，再基于每个训练样本集，使用分类、回归等基分类器算法，得到多个子模型，最后再对全部子模型的结果进行组合平均，从而得到模型的最终输出结果。

Breiman（1996）研究指出，当基分类器是决策树或神经网络这类不稳定模型时，装袋算法通常可降低模型的方差，避免发生过度拟合，从而提高模型性能；而当基分类器是 k 近邻法这类稳定模型时，装袋算法反而可能降低模型性能。

5.2.3　随机森林的构建

概括来看，随机森林是在组合使用装袋算法和随机选择分割变量的基础上构建的，其基本思路如下：

（1）首先，采用自助抽样法（均匀有放回抽样）对原始样本随机采样，从样本数为 M 的原始样本中每次随机抽取 M 个样本作为单个分类树的训练数据集，

从而引入训练样本的随机性和弱相关性；

（2）其次，对于单个分类树，从所有 P 个属性变量中随机选取 K 个属性，基于抽取的特定训练数据集和使用的算法，依次从 K 个属性中找出能够提供最佳分割效果的最优属性，在不修剪的情况下经过递归分割生成单个分类树，再根据使用的训练数据集预测叶节点分类；

（3）最后，重复上述（1）-（2）过程 H 次，且在每次生成单个分类树时可使用不同的算法（如决策树、支持向量机、逻辑回归等），最终采用投票机制（一票否决、少数服从多数、加权多数等）从全部分类树预测类别中选出最终的预测分类作为随机森林模型的输出类别。

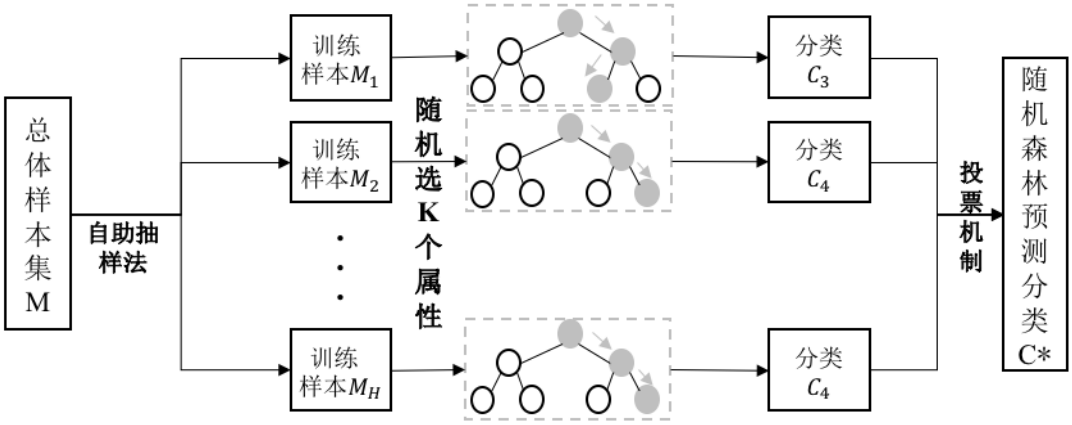

图 7 随机森林基本思想

随机森林模型的可调参数通常包括每次选取的属性个数 K（在分类问题中通常取全部属性的平方根），单个分类树的最大深度、最小分裂样本数以及类别比例，分类树的个数（或重复次数）H。

5.2.4　袋外误差

袋外误差（out-of-bag error，OOB error）是随机森林模型特有的用以估计模型误差的统计量，在构建出随机森林模型后，可以同时得到其袋外误差。

由前述装袋算法部分对自助抽样法的阐述可知，对于随机森林中的每棵分类树，由于其使用的训练样本集是对原始样本进行有放回抽样后生成的，原始样本中的每个样本未被抽到的概率是$(1-1/\mathrm{M})^{\mathrm{M}}$，且当 M 足够大时该值将收敛于$1/\mathrm{e}\approx 0.368$，即原始样本中有近 36.8%的数据不在用于生成分类树的训练样本集整体中出现，这些数据也被称作“袋外数据（out-of-bag，OOB）”。袋外误差就是使用袋外数据得到的随机森林模型误差，对于分类问题，这一误差也即模型对袋外数据的预测分类误差率的均值，具体计算过程如下：

（1）对每个样本，计算它作为袋外数据时，分类树对它的分类情况；

（2）以简单多数投票的分类树结果作为随机森林模型对该样本的最终分类

预测结果，若与该样本的实际分类不一致则计作 1 个误分类；

（3）遍历全部原始样本并累计误分类个数，其占原始样本总数的比率即为随机森林的袋外误差。

由袋外数据的定义可知，其与用来构建随机森林模型的训练数据形成原始数据中的一对补集，袋外数据也因而自然地成为训练数据的对照数据集（验证数据集），这使得随机森林模型本身不需要另外预留部分数据做交叉验证。同时，袋外误差也成为模型预测误差的无偏估计（Breiman L，2001），可视作对模型泛化能力的度量。

5.2.5　变量相对重要性

基于构建的随机森林模型，尽管无法得到解释变量影响被解释变量的量化数值程度及方向信息，但可以衡量解释变量的相对重要性，基本思想是对解释变量加入噪声值，然后考察加入前后的模型预测精度变化情况。根据具体考察方法的不同，随机森林模型对变量相对重要性的衡量方法主要有两种（李欣海，2013）：

一种是 MDA（Mean Decrease Accuracy）方法，该方法基于袋外误差，取值为当某个变量变为随机数后所有分类树袋外误差降低程度的平均值，该值越大则表示该变量的重要性越大；

另一种是 MDG（Mean Decrease Gini）方法，该方法基于基尼指数，取值为当某个变量作为分割特征时所有分类树 Gini 值降低程度的平均值，该值越大则表示该变量的重要性越大。

在随机森林模型的实际应用中，对于解释变量特别多的情况，可以先使用全部变量训练构建随机森林模型，再通过求解变量的相对重要性，筛选出相对更重要的解释变量子集，然后再次构建随机森林模型，从而提高模型效率。由于本文所使用的解释变量并不多，因而在后文的应用中不再进行二次训练建模。

5.2.6　模型优缺点

1、优点

随机森林模型既继承了分类树模型的优点，同时又兼有组合算法的内在优势，具体而言，主要有以下突出优点：

（1）首先，不需要对数据进行标准化等预处理（李建更、高志坤，2009），对缺失、异常离群值和非平衡数据较稳健，在随机干扰较多的情况下仍表现稳健（李欣海，2013），相比于适合小样本的支持向量机和相比于对缺失、离群值敏感的逻辑回归模型等均有显著优势；

（2）其次，不需要预先检查变量间是否存在显著的交互性和共线性（Cutler R D 等，2007），且能同时处理连续型和离散型变量，并可以对解释变量进行相

对重要性度量（Efron B & Hastie T，2016），可以很好地预测多达几千个解释变量的作用（Breiman L，2001），相比于依赖变量独立性条件的逻辑回归模型有显著优势；

（3）此外，不需要预留用于交叉验证的数据集，随机森林模型的袋外误差是模型预测误差的无偏估计（Breiman L，2001），相比于分类树模型，较少出现过度拟合情况，泛化能力和预测准确度显著提高（李欣海，2013）；

（4）最后，相比于人工神经网络模型等其他机器学习算法，随机森林模型不需要进行过多的调参和手动修改，原理和过程均较简洁，训练学习过程较快，在运算量没有显著提高的情况下提高了预测精度（李欣海，2013）。

2、缺点

然而，随机森林算法也存在一定缺点，主要体现在以下几个方面（张俊妮，2009）：

（1)无法解读模型生成时的黑箱过程，如构建单个分类树时使用的数据集、具体算法和分割属性等，但相比于人工神经网络算法仍更透明一些；

（2）预测结果倾向于样本较多的类别，样本较多分类的特征变量比样本较少分类的特征变量对模型的影响程度更大（Deng H 等，2011）；

（3）只能得到解释变量对被解释变量影响的相对重要性，而不能得到影响的量化数值和方向；

（4）较复杂的随机森林模型会因为评估速度慢而结束，但这一般仅出现在非常庞大的模型中。

5.3　随机森林模型的应用

5.3.1　金融应用的文献综述

在机器学习的诸多算法中，随机森林模型由于具备诸多优点，而已被生物信息学、医学、人工智能、金融等各行业较多地应用（Cutler R D 等，2007；Genuer R 等，2010；方匡南等，2010）。下面就对金融经济学界中应用随机森林模型的主要研究进行文献梳理和综述，主要分为预测银行信贷风险、P2P 网贷风险、企业财务风险、股票价格、基金评级与收益等几个方面，并几乎一致地认为随机森林模型相比于其他预测分析模型均有显著优势。但截至目前，学界尚未有研究将随机森林模型应用于分析个体投资者异质性特征对证券投资收益的影响。

1、银行信贷风险预测

方匡南等（2010）使用中国某大型商业银行的 70532 笔个人信用卡数据和用户特征，利用随机森林模型对信用卡用户的信用评分进行预测分析，发现相比于

逻辑回归模型，随机森林的预测准确性和外推稳定性均较高。萧超武等（2014）基于加州大学尔湾分校机器学习库公开的德国某商业银行的 1000 条个人信贷数据，分别使用随机森林模型以及单分类器模型 K 近邻、径向基网络（RBF-NET）、支持向量机等和组合模型梯度提升决策树（GBDT）对个人银行信用状况进行预测，发现随机森林模型具有更高的预测精度和稳定性，处理噪声的能力强且泛化能力好，并在随机森林模型的基础上对特征变量的重要性进行了相对排序。

2、P2P 网贷风险预测

周玉琴等（2016）基于“人人贷”平台一季度的订单数据，通过构建随机森林模型考察了借款人的 33 个特征变量对 P2P 网络借贷成功率的影响和相对重要性，并与决策树、神经网络、逻辑回归、贝叶斯和支持向量机等其他算法的预测结果进行对比，发现随机森林模型在预测准确度等方面要显著优于其他所有模型，同时对不同状态的非平衡数据的预测表现均较稳健，且对比例越高的数据集的预测准确度越高。Malekipirbazari 和 Aksakalli（2015）使用国际知名的美国 P2P 借贷平台 Lending Club（简称“LC”）在 2012 年 1 月-2014 年 9 月之间的共约 6.8 万个已到期借贷记录，基于 23 个借贷者特征变量对借贷违约情况进行分类预测，分别使用了 K 近邻、逻辑回归、支持向量机及随机森林模型，发现随机森林模型具有最高的预测准确率和最小模型误差，同时也比美国官方使用的 FICO 评分及 LC 平台评分的预测效果更佳。

3、企业财务风险预测

孟杰（2014）使用锐思金融数据库提供的 2011 年为 ST 或*ST 的公司在 2007-2011 年的数据，基于 17 个指标和随机森林分类模型，对公司在 2011 年是 ST 公司还是正常公司进行预测，同时基于同样的数据和指标，另外构建逻辑回归、决策树、支持向量机和人工神经网络模型进行预测，发现随机森林模型的预测精度最优且外推性能最好，而逻辑回归模型的预测能力较差。连晓丽（2014）基于 A 股 234 家正常公司和 78 家 ST 公司，通过构建随机森林模型和 Lasso-logistic 模型，在不同市场情况下以公司个体特征指标对其发生财务危机并被 ST 的情况进行预测，发现随机森林模型具有较高的预测精度和稳定性，具有较高实用价值。

4、股票价格预测

股票价格预测方面，曹正凤等（2014）使用 2012-2013 年两个行业中 360 余只非 ST 股票的数据，基于价值投资策略选取 9 个因子作为选股指标，在此基础上构建随机森林模型选择优质股票，发现预测精度超过 78%。王淑燕等（2016）使用计算机、通信和其他电子设备制造业中 200 只股票在 2013 年 3 月的数据，首先基于相关系数提出八因子选股指标，并在此基础上构建随机森林量化选股模

型对 4 月的涨跌情况进行预测，通过与实际涨跌情况进行对比发现预测精度高于 75%。Ballings 等（2015）使用在欧洲上市的涵盖 19 个行业的 5767 只股票的年度数据，基于量价、偿债能力、赢利能力等多个指标对价格波动方向进行了分类预测，分别使用了单一分类器方法——逻辑回归、神经网络、K 近邻、支持向量机，以及集成分类方法——随机森林、自适应增强、核函数工厂，通过对比各方法的预测结果，发现随机森林的效果最佳，其次分别为支持向量机、核函数工厂、自适应增强、神经网络、K 近邻和逻辑回归，并对价格波动方向分类的置信区间进行了稳健性检验。Patel 等（2015）使用印度股市两只股票和两个股票指数的 10 年数据作为总体样本，并使用其中 20%的数据筛选变量，再利用总体样本分别构建将变量视为连续变量和离散变量的人工神经网络、支持向量机、随机森林和朴素贝叶斯模型，对两只股票和两个股票指数的价格变动方向进行预测，发现随机森林模型在对连续变量的预测中效果最佳，同时所有模型在对离散变量的预测中，预测精度都得到提升。

5、基金评级与收益预测

基金评级方面，王志红和王华珍（2009）采用被国内 4 家主流评级机构一致评级的基金数据作为训练样本，以 18 个指标建立了基金绩效综合评级体系，并构建随机森林模型对基金级别进行预测分类，发现模型的预测精准性和稳定性均较为优良。基金收益预测方面，方匡南等（2010）使用成长型封闭式股票基金“裕隆基金”自 1999 年 6 月至 2008 年 11 月的数据，选取量价及收益指标的滞后变量作为解释变量，分别构建随机游走、自回归移动平均、支持向量机和随机森林模型对基金相对上证 180 指数的超额收益方向进行预测，并基于训练出的模型和得到的预测结果，以 2006-2008 年国内股市的数据构建交易策略，发现随机森林模型的预测效果最好且据此构建的交易策略显著优于其他策略。

总之，随机森林模型在评估个体信用、企业金融风险方面的应用已十分广泛，在证券及基金投资方面也有较多研究，但国内外学界还未有研究将随机森林模型用于分析个体异质性特征对证券投资收益的影响。鉴于本文所使用“大调查”数据样本量较大，且样本的代表性较好，同时已有诸多现实应用证实随机森林模型拥有对解释变量间交互性不敏感、预测准确度高、泛化能力强等优良特性，因而本文下面将创新性地使用随机森林模型对个体异质性特征影响证券投资收益的问题进行进一步探究，以期在与前文有序多分类逻辑回归模型形成互补的同时，对学界和业界今后的研究及应用提供一定启发。

5.3.2　数据、变量与模型构建

为了使随机森林模型的结果能与前文有序逻辑回归模型的结果进行有效对比，随机森林模型使用的原始数据和变量将与前文回归部分保持一致，未进行额

外的数据处理或变量删减。同时，我们首先随机抽取原始数据集 70%的数据作为训练集用以后续构建随机森林模型，而另外 30%的数据作为测试集用以更直观地展示随机森林模型的泛化能力和预测精度。本部分将使用开源的 R 软件及 randomForest 包构建随机森林模型，代码详见附录。

由于已有研究发现，在大多数情况下，随机森林模型的缺省参数设置可以给出最优或接近最优的结果，同时，在计算负荷可以接受的情况下，分类树的数量越大越好（李欣海，2013）。因此，本文在构建随机森林模型时均采用软件包提供的默认选项，并在未遇到提前结束程序的情况下不限制分类树的数量，由此得到的模型分类树总数为 500 个，每个分类树的解释变量个数为解释变量总数的平方根。

5.3.3　模型评估

评估随机森林模型效果的方法主要有两种，一种是混淆矩阵方法，另一种是袋外误差。

混淆矩阵是一个展示样本实际分类（行标签）与模型预测分类（列标签）的 K*K 表格，其中 K 代表类别数（本文为 8），表格(i，j)位置（除行列标签外）的数字代表实际分类为 i、但模型预测分类为 j 的样本个数，由此可知，表格对角线上的数字即为各类正确预测的样本数。同时，对于实际分类 i=0,1，⋯8，可以分别计算出其分类误差率=$1-(i,i)/\sum_{j=0}^{8}(i,j)$。本文构建的随机森林模型的混淆矩阵如下表 6 所示，可以直观发现实际分类的样本量越大则分类误差率越低。

袋外误差的基本原理已在前文进行过阐释，在此不再赘述。由下表 6 可知，构建的随机森林模型的袋外误差为 57.40%，对应的综合预测分类准确度为 42.60%，比有序多分类逻辑回归模型的 37.87%要高 4.73%。

表格 6 随机森林训练集混淆矩阵与袋外误差

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**分类误差率**
0	18404	459	109	37	183	29	2	4	5	0.0431
1	4447	1119	105	38	155	5	3	8	3	0.8098
2	4614	363	412	32	131	8	2	3	4	0.9260
3	3860	228	67	181	109	7	1	5	2	0.9594
4	7389	391	87	40	426	14	1	6	6	0.9490
5	2198	82	25	15	53	67	2	1	0	0.9726
6	886	47	16	3	23	3	8	2	1	0.9919
7	519	72	16	9	39	1	1	5	2	0.9925
8	701	72	23	14	22	2	4	3	21	0.9756

	袋外误差（OOB error）：57.40%

随后基于测试数据集，基于上述随机森林模型进行预测，通过与实际分类进行比较得到如下的混淆矩阵，可以发现综合预测分类准确度为 41.65%，比有序多分类逻辑回归模型高 3.78%，并与基于训练集得到的随机森林模型预测准确度（42.60%）相差不到 1%，体现随机森林确实具有较强的泛化能力和预测精度。

表格 7 随机森林测试集混淆矩阵与袋外误差

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**分类误差率**
0	7743	1945	2061	1730	3220	929	370	236	336	0.5830
1	211	502	179	105	171	33	27	30	33	0.6112
2	56	53	194	25	35	8	13	8	7	0.5138
3	19	11	17	79	16	8	3	2	4	0.5031
4	78	71	56	57	186	12	13	18	16	0.6331
5	5	3	4	2	5	32	0	1	0	0.3846
6	3	1	1	1	3	1	1	0	0	0.9091
7	0	3	0	1	1	2	0	5	1	0.6154
8	0	0	0	1	0	0	0	1	12	0.1429
	综合分类准确度：41.65%									

5.3.4　变量相对重要性评估

基于构建的随机森林模型，还可以进一步评估各个解释变量对个体证券投资收益影响的相对重要程度，结果由下图 8 的重要性曲线所示，变量对应的圆圈越靠右则代表 MDA 或 MDG 值越大，也即该变量相对越重要。

图 8 随机森林变量相对重要性曲线

由上图可以发现，在两种变量相对重要性评估方法中，职业变量career均位列前三位之一，因而可以初步推测其对个体投资者证券投资收益的影响相对较大；而性别变量sex均位于最后一位，因而可以初步推测其对个体证券投资收益的影响相对不重要。

5.3.5 与有序多分类逻辑回归模型的对比

由上述随机森林模型的预测结果和评估分析可知，与有序多分类逻辑回归模型相比，随机森林模型的突出优势在于预测准确度显著提高且泛化能力较强。这可能与本文使用的解释变量之间存在一定的内在关联有关，由于有序多分类逻辑回归模型对解释变量的交互性非常敏感，要求解释变量之间相互独立；而作为随机森林模型基分类器的分类树算法中已经自然地包括了变量间的交互作用（Cutler R D 等，2007），即 X1 的变化会导致 X2 对 Y 的作用发生改变，不需要自变量之间相互独立这个前提条件，与现实中解释变量间的情形更为契合，预测准确度也因而提高。

但是，相比于有序多分类逻辑回归模型，上述随机森林模型也有一定的不足。首先，由于模型的生成机制为黑箱，使得无法确定随机森林中每棵分类树使用的数据集、变量和分类规则；更为重要的是，随机森林模型无法对解释变量影响被解释变量的程度进行量化数值和方向分析。

总之，当需要对影响因素进行绝对量化分析时，有序多分类逻辑回归模型是更好的选择；而当数据较为不标准、存在大量缺失及异常值等，或当影响因素间存在较为明显的交互关系，或当需要从庞大的影响因素中筛选出相对重要的变量作为解释变量，或当需要得到更为精准的预测结果时，随机森林模型都不失为一种较好的选择。本文认为，有序多分类逻辑回归模型与随机森林模型各有所长也各有所短，在实际应用中应当灵活选取二者中更为合适的方法或者将二者结合使用，从而更大程度地挖掘数据中的信息，为学界或业界创造更大的现实应用价值。

第六章　结论

我国沪深股市在 2015 年上半年延续了自 2014 年 7 月开始的牛市，但在 6 月中旬突然出现历史罕见的急速大幅暴跌，受这一轮股灾的影响，作为我国股市主要参与者的个体投资者在 2015 年普遍亏损且收益状况差异较大。为了探讨个体异质性特征对个体证券投资收益状况的影响，本文基于行为金融学过度自信、过度反应和心理账户理论，首先分别提出个体异质性特征影响个体投资收益的假说，然后利用 2015~2016 年 CCTV 财经频道“中国经济生活大调查”这一微观截面数据，使用有序多分类逻辑回归模型实证探究上述影响，并在回归分析之后创新性地进行预测分析以评估模型预测结果与真实数据的一致性、评判现实应用价值；此外，考虑到解释变量的交互性和为提高模型的预测准确度，本文进一步使用随机森林模型训练建模并预测，与有序多分类逻辑回归模型优势互补，完善了对本文所研究问题的探究并为学界和业界的实际应用提供启发。

在有序多分类逻辑回归部分，本文实证发现投资者年龄对证券投资收益没有显著影响，而受教育程度、性别、职业、家庭收入和住房状况均会对个体投资者的证券投资收益产生显著影响，且影响的方式存在一定差异。具体来看，受教育程度提高有利于证券市场投资者提高获得各种赢利情况的可能性，尤其是提高赢利 20%以内的概率，并能有效降低发生极端亏损（亏损 50%以上）的可能性，但也会在一定程度上由于过度自信导致在股灾中出售股票不及时，进而转赢为亏，提高了亏损 50%以内的可能性；女性除了发生极端亏损的可能性要比男性低 0.04%外，亏损 50%以内和各种赢利情况的可能性均高于男性，由于股灾时期风险厌恶的女性投资者往往过度反应，从而导致发生小幅亏损的概率较高而发生极端亏损的概率较低，同时在投资者普遍亏损的情况下，风险厌恶的女性比风险偏好的男性更可能获得赢利；相比于自由职业者和进城务工人员，选择作为企业管理人员和行政事业单位人员的投资者较为风险厌恶，从而在股灾发生时易于过度反应提前止损，导致在降低发生极端亏损可能性的同时，发生 50%以内亏损的可能性也提高，同时风险厌恶也使二者更可能获得赢利；随着收入水平的提高，投资者的股票投资组合多样性提高，可以尽可能地分散投资风险，在降低发生极端亏损可能性的同时也提高获得赢利的可能性，但由于过度自信，导致发生 50%以内亏损的可能性提高；由于相较租房者而言，拥有农村住房的投资者的未来生活更稳定而风险偏好程度更高，使其在股市前期大幅下跌时倾向于继续持有，在提高发生极端亏损可能性的同时也降低了发生 50%以内亏损的可能性，同时风险更为偏好也使其获得赢利的可能性更低。此外，对回归模型进行预测分析发现，模型预测准确度达到 37.87%，鉴于被解释变量有 9 种分类，可以认为是较好的模型。

随机森林模型具有对变量交互性不敏感、预测能力和泛化能力强等有序多分类逻辑回归模型所不具备的诸多优点，但也有只能衡量解释变量对被解释变量的相对影响程度而不能衡量绝对量化影响程度的缺陷。考虑到解释变量的交互性和为提高模型的预测准确度，从而为学界和业界现实应用提供更有价值的参考，在梳理随机森林模型基本原理的基础上，本文进一步使用与有序多分类逻辑回归模型相同的数据和变量，训练构建随机森林模型并进行评估分析。通过求解混淆矩阵和袋外误差，本文发现随机森林模型相比于有序多分类逻辑回归模型的预测准确度提高了 4%-5%，同时泛化能力较强；同时，基于变量相对重要性曲线，本文发现职业因素对个体投资者证券投资收益的相对影响程度较大，而性别因素的相对影响程度较小，但受限于随机森林模型本身的缺点，相关影响均难以量化分析。

总之，上述实证研究结果表明，个体异质性特征确实会显著影响个体证券市场的投资收益，但在影响方向和强度方面有较大的差异，且同时使用有序多分类逻辑回归模型和随机森林模型可以达到优势互补、尽可能挖掘数据信息、提高现实应用价值的效果。

本文的研究内容和方法不仅为学界相关领域内的研究提供了良好的补充，也有助于提高学界对计量经济模型现实应用价值的重视程度；另外，本文的研究结果为个体投资者、尤其是中低收入投资者今后投资证券市场提供了一定的启示，有助于其对未来的投资收益产生更审慎合理的预期，从而更好地保障相关群体的利益；此外，通过使用机器学习算法中的随机森林算法搭建预测性能、泛化能力均较好的模型，在与有序多分类逻辑回归模型优势互补的同时，也具有更高现实应用价值，为业界金融机构向证券市场个体投资者提供更有针对性和更丰富的金融服务产品提供了有价值的实证参考。

然而，不容否认的是，本文的研究也存在诸多不足。首先，由于行为金融学还没有形成完整有机的理论体系，且作者在心理学、行为金融学等方面的知识还尚为浅显，选取的行为金融学分支理论可能不够完整，使得提出的理论假说可能不够严密；其次，本文使用的数据样本为横截面数据，在数据可得性的基础上，今后可以使用更理想的面板数据进行探究；最后，由于篇幅的原因，本文没有进一步展开探讨个体收入预期、房价预期等因素对个体证券市场投资收益的影响。上述方面也是今后进一步研究和完善的方向。

参考文献

Ballings M, et al．Evaluating Multiple Classifiers for Stock Price Direction Prediction[J]．Expert Systems with Applications, 2015, 42: 7046-7056.

Barberis N, Thaler H R．A Survey of Behavioral Finance[R]．NBER Working Paper No. 9222, 2002.

Bellante D, Link A．Are Public Sector Workers More Risk Averse Than Private Sector Workers? [J]．ILR Review, 1981, 34(3)：408-412.

Ben-David I, Graham R J, Harvey R C．Managerial Overconfidence and Corporate Policies[R]．NBER Working Paper No. 13711, 2007.

Booth A L, Nolen P. Gender Differences in Risk Behavior: Does Nurture Matter? [J]. The Economic Journal, 2012, 122(558)：56-78.

Breiman L．Bagging Predictors[J]．Machine Learning, 1996, 24(2)：123-140.

Breiman L．Random Forests[J]．Machine Learning, 2001, 45(1)：5-32.

Breiman L. Statistical Modeling: The Two Cultures[J]. Statistical Science, 2001, 16(3)：199-231.

Cramer J S, et al．Low Risk Aversion Encourages the Choice For Entrepreneurship: An Empirical Test of A Truism[J]．Journal of Economic Behavior & Organization, 2002, 48：29-36

Cutler R D, et al. Random Forests for Classification In Ecology[J]. Ecology , 2007, 88(11)：2783–2792.

De Bondt M F W, Thaler H R．Does the Stock Market Overreact? [J]．The Journal of Finance, 1985, 40(3)：793-805.

De Bondt M F W, Thaler H R．Further Evidence on Investor Overreaction And Stock Market Seasonality[J]．The Journal of Finance, 1987, 42(3)：557-581.

De Mel S, McKenzie D, Woodruff C. Are Women More Credit Constrained? Experimental Evidence on Gender and Microenterprise Returns[J]. American Economic Journal, 2009, 1(3)：1-32.

Deng H, Runger G, Tuv E．Bias of Importance Measures for Multi-valued Attributes and Solutions[J]. Proceedings of the 21st International Conference on Artificial Neural Networks, 2011：293-300.

Efron B．Bootstrap Methods: Another Look at the Jackknife[J]．The Annals of Statistics, 1979：1-26.

Efron B, Hastie T．Computer Age Statistical Inference：Algorithms, Evidence and Data Science[M]．Cambridge：Cambridge University Press，2016.

Ekelund J, et al．Self-employment And Risk Aversion--Evidence From Psychological Test Data[J]．Labor Economics, 2005, 12(5)：649-659.

Fischhoff B, Slovic P, Lichtenstein S．Knowing with Certainty: The Appropriateness of Extreme Confidence[J]．Journal of Experimental Psychology: Human Perception and Performance, 1977, 3(4)：552-564.

Frank D J．Some Psychological Determinants of the Level of Aspiration[J]．The American Journal

of Psychology, 1935, 47(2)：285-293.

Friedman M, Savage J L. The Utility Analysis of Choices Involving Risk[J]. The Journal of Political Economy, 1948, 56, 4：279-304.

Genuer R, Poggi J M， Tuleau-Malot C. Variable selection using random forests[J]. Pattern Recognition Letters，2010，31(14)：2225-2236.

Green W H. Econometric Analysis (Eighth Edition) [M]. London: Pearson plc，2017.

Griffin D, Tversky A . The Weighing of Evidence And The Determinants of Confidence[J]. Cognitive Psychology, 1992, 24(3)：411-435.

Hartog J, Ferrer-i-Carbonell A, Jonker N. Linking Measured Risk Aversion to Individual Characteristics[J]. Kyklos, 2002, 55(1)：3-26.

Ho K T. Random Decision Forests[J]. Proceedings of the 3rd International Conference on Document Analysis and Recognition, 1995, 1：278-282.

Ho K T. The Random Subspace Method for Constructing Decision Forests[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8)：832-844.

Kahenman D, Tversky A. Judgment under uncertainty: Heuristics and biases[J]. Science, 1974, 185(4157)：1124-1131.

Kahenman D, Tversky A. Prospect Theory: An Analysis of Decision Under Risk[J]. Econometrica, 1979, 47(2)：263-291.

Mahajan J. The Overconfidence Effect in Marketing Management Predictions[J]. Journal of Marketing Research, 1992, 29(3)：329-342.

Malekipirbazari M, Aksakalli V . Risk Assessment in Social Lending Via Random Forests[J]. Expert Systems with Applications, 2015, 42: 4621-4631.

Patel J, et al. Predicting Stock and Stock Price Index Movement Using Trend Deterministic Data Preparation and Machine Learning Techniques[J]. Expert Systems with Applications, 2015, 42: 259-268.

Shefrin H, Statman M. Behavior Portfolio Theory[J]. Journal of Financial and Quantitative Analysis, 2000, 35(2): 127-151.

Taylor S, Brown J D. Illusion And Well-being: A Social Psychological Perspective on Mental Health[J]. Psychological Bulletin, 1988, 103(2)：193-210.

Thaler H R. Mental Accounting And Consumer Choice[J]. Marketing Science, 1985, 4(3): 199-214.

Tin J. Household demand for financial assets: A life-cycle analysis[J]. The Quarterly Review of Economics and Finance, 1998, 38(4): 875-897.

Vissing-Jørgensen A. Limited Asset Market Participation and the Elasticity of Intertemporal Substitution[J]. Journal of political Economy, 2002, 110(4): 825-853.

Watson J, McNaughton M. Gender Differences In Risk Aversion And Expected Retirement Benefits[J]. Financial Analysts Journal, 2007, 63(4)：52-62.

Weinstein N D. Unrealistic Optimism About Future Life Events[J]. Journal of Personality and Social Psychology, 1980, 39(5)：806-820.

Wolosin R J, Sherman S J, Till A．Effects of Cooperation And Competition on Responsibility Attribution After Success And Failure[J]．Journal of Experimental Social Psychology, 1973, 9(3)：220-235.

Zakay D, Tuvia R. Choice Latency Times as Determinants of Post-Decisional Confidence[J]. Acta Psychologica, 1998, 98(1)：103-115.

曹正凤，纪宏，谢邦昌．使用随机森林算法实现优质股票的选择[J]．首都经济贸易大学学报，2014，2：21-27.

陈国进，范长平．我国股票市场的过度反应现象及其成因分析[J]．南开经济研究，2006，3：42-53.

陈强．高级计量经济学及 Stata 应用（第二版）[M]．北京：高等教育出版社，2014.

丁小浩，孙毓泽，Hartog J．风险态度与教育和职业选择行为——一个实验方法的研究案例[J]．北京大学学报（哲学社会科学版），2009，46(1)：140-144.

窦松博．现阶段个人投资者特征与投资收益率相关性研究[D]．济南：山东大学硕士学位论文，2016.

方匡南等．信贷信息不对称下的信用卡信用风险研究[J]．经济研究，2010，S1：97-107.

方匡南，朱建平，谢邦昌．基于随机森林方法的基金收益率方向预测与交易策略研究[J]．经济经纬，2010，2：61-65.

冯燮刚，李子奈．经济学的关系论转向[J]．经济学动态，2006，7：17-25.

洪永淼．计量经济学的地位、作用和局限[J]．经济研究，2007，5：139-153.

黄莲琴．管理者过度自信与公司融资行为研究[D]．厦门：厦门大学博士学位论文，2009.

李爱梅，凌文辁．心理账户：理论与应用启示[J]．心理科学进展，2007，15（5）：727-734.

李航．统计学习方法[M]．北京：清华大学出版社，2012.

李建更，高志坤. 随机森林针对小样本数据类权重设置[J]. 计算机工程与应用，2009，45(26)：131-134.

李涛，郭杰．风险态度与股票投资[J]．经济研究，2009，2：56-66.

李潇潇．中国居民股市和基金参与意愿的影响因素研究——基于 2014 年底中国经济生活大调查数据[D]．北京：北京大学硕士学位论文，2015.

李心丹，王冀宁，傅浩．中国个体证券投资者交易行为的实证研究[J]．经济研究，2002，11：54-63.

李心丹，肖斌卿，俞红海．家庭金融研究综述[J]．管理科学学报，2011，14(4)：74-85.

李欣海. 随机森林模型在分类与回归分析中的应用[J]. 应用昆虫学报，2013，50(4)：1190-1197.

李子奈．计量经济学模型方法论的若干问题[J]．经济学动态，2007，10：22-30.

李子奈，刘亚清．现代计量经济学模型体系解析[J]．经济学动态，2010，5：22-31.

连晓丽. 我国 A 股上市公司财务危机预警模型实证研究[D]. 厦门：厦门大学硕士学位论文，2014.

廖娟．收入风险、教育水平与职业选择——基于北京市部分用人单位职工的调研[J]．教育发展研究，2011，17：6-9.

孟杰．随机森林模型在财务失败预警中的应用[J]．统计与决策，2014，4：179-181.

彭星辉，汪小虹．上海股民的投资行为与个性特征研究[J]．心理科学，1995，18(2)：94-98．

史代敏, 宋艳．居民家庭金融资产选择的实证研究[J]．统计研究, 2006，10：43-49．

谭松涛，陈玉宇．投资经验能够改善股民的收益状况吗？[J]．金融研究，2012，5：164-178．

王聪, 田存志．股市参与, 参与程度及其影响因素[J]．经济研究，2012，10：97-107.

王垒等．中国证券投资者的投资行为与个性特征[J]．心理科学，2003，26(1)：24-27．

王淑燕，曹正凤，陈铭芷．随机森林在量化选股中的应用研究[J]．运筹与管理，2016，25(3)：163-168．

王小勇，我国幸福感的变化趋势与地区差异——经济生活大调查（2007-2010）数据分析[D]．北京：北京大学硕士学位论文，2011．

王渊，杨朝军，蔡明超．居民风险偏好水平对家庭资产结构的影响——基于中国家庭问卷调查数据的实证研究[J]．经济与管理研究，2016，37(5)：50-57．

王志红，王华珍．基于随机森林的基金评级模型选择[J]．财务与金融，2009，1：65-70．

吴卫星, 齐天翔．流动性, 生命周期与投资组合相异性——中国投资者行为调查实证分析[J]．经济研究，2007，2：97-110．

吴卫星, 易尽然, 郑建明．中国居民家庭投资结构: 基于生命周期, 财富和住房的实证分析[J]．经济研究，2010，S1：72-82．

萧超武等．基于随机森林的个人信用评估模型研究及实证分析[J]．管理科学，2014，6：111-113．

易丹辉．数据分析与 EViews 应用（第二版）[M]．北京：中国人民大学出版社出版，2014．

尹志超, 宋全云, 吴雨．金融知识, 投资经验与家庭资产选择[J]．经济研究，2014，4：62-75．

袁典．个人投资者风险态度与股票投资盈亏[D]．成都：西南财经大学硕士学位论文，2016．

张俊妮．数据挖掘与应用[M]．北京：北京大学出版社，2009．

张腾文，王威，玉翠婷．金融知识、风险认知与投资收益[J]．会计研究，2016，7：66-73．

张晓玉．大数据时代的数据挖掘分析实例——模型选择、比较、评估和提升[D]．北京：北京大学硕士学位论文，2014．

张晓峒．EViews 使用指南与案例[M]．北京：机械工业出版社，2007．

赵振华，刘淳，廖理．是谁获得了更高的基金投资收益？——对个人投资者问卷调查的实证分析[J]．金融研究，2010，5：166-178．

周业安，左聪颖，袁晓燕．偏好的性别差异研究：基于实验经济学的视角[J]．世界经济，2013，7：3-27．

周玉琴，张晓玫，罗璇．基于随机森林的 P2P 网络借贷成功率预测研究[J]．东北农业大学学报（社会科学版），2016，6：11-17．

附录 随机森林 R 代码

```
# Install and load packages
install.packages("mboost")
install.packages("rpart")
install.packages("maptree")
install.packages("randomForest")
install.packages("DMwR")
install.packages("ipred")
install.packages("caret")
install.packages("e1071")
library(mboost)
library(rpart)
library(maptree)
library(cluster)
library(lattice)
library(grid)
library(randomForest)
library(DMwR)
library(ipred)
library(caret)
library(ggplot2)
library(e1071)
# Read dataset
data <-
read.csv("/Users/apple/Desktop/dofile_no_marriage_hincome2_RF.csv",header=T)
ndata <- na.omit(data)
ndata$securities <- as.factor(ndata$securities)
# Set trainging and testing set
set.seed(2)
index <- sample(2, nrow(ndata), replace=T, prob=c(0.7,0.3))
traindata <-ndata[index==1,]
testdata <-ndata[index==2,]
# Random forest
model.forest <-randomForest(securities~., data=traindata,importance=T)
model.forest
fit.forest <- predict(model.forest, data.frame(testdata))
tabc <- confusionMatrix(fit.forest,testdata$securities)
tabc
# Importance of Variables
importance(model.forest)
varImpPlot(model.forest)
```

致谢

在本次学位论文完成之时，我想向为我提供过诸多帮助和支持的老师、家人和同学好友表达深深的感谢之情！

首先，我要真挚地感谢胡大源老师，胡老师为我在课程学习以及论文时间安排、选题、实证方法选择等各方面都提供了宝贵的建议和悉心的指点，使我获益匪浅。在我研一下学期修读高级计量经济学期间，胡老师建议我借写作课程论文的机会为学位论文探路，受益于胡老师的建议，我提前完成了毕业论文的部分内容，为后续攻克论文难点留出充足的时间；进入研二上学期，胡老师建议我修读光华管理学院的《数据挖掘及应用》课程，通过该门课程的学习，我接触到诸如人工神经网络、K 近邻、随机森林等前沿机器学习算法，并在课程作业的应用中亲身体会到各种方法的优缺点，进而为本次学位论文的写作找到了随机森林模型这一合适的研究方法；同时，在论文写作期间，经过与胡老师的多次面对面交流，我不仅在论文的选题、问卷调查数据及变量的设定等方面得到诸多启发，也在今后的职业选择和未来发展方面获得诸多体会，并被胡老师渊博的学识、丰富的阅历、严谨求精的态度和平易谦逊的人格所折服，使我开阔了自身视野、升华了人生目标。总之，胡老师的言传身教使我受益匪浅，在此谨向我的导师胡大源老师表示最诚挚的感谢和最由衷的敬意！

其次，我还想感谢在我求学南大经济系期间为我提供宝贵指导和谆谆教诲的本科导师沈坤荣老师和在我写作保研论文过程中为我不辞辛劳地提供宝贵修改意见的皮建才老师，你们是我求学深造道路上不可或缺的引路人；还要感谢在我求学国家发展研究院以来，为我提供帮助和指点的国家发展研究院老师们，使我夯实提高了金融和经济学专业知识，强化了科研学术能力；同时，还要感谢光华管理学院的张俊妮老师，帮助我直观了解和系统认识了机器学习和数据挖掘的前沿基本算法。

最后，我还要感谢我的父母、男友和同门、好友，你们在我写作论文的过程中为我提供了源源不断的鼓励和支持，使我在遇到困难时有了克服的勇气和信心。

我的研究生在校生涯即将结束，接下来又要奔向人生的另一段旅程，我将会努力不负各位的帮助与鼓励，再接再厉。再次由衷地感谢你们！

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

学位论文使用授权说明

（必须装订在提交学校图书馆的印刷本）

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；
- 因某种特殊原因需要延迟发布学位论文电子版，授权学校□一年/□两年/□三年以后，在校园网上全文发布。

（保密论文在解密后遵守此规定）

论文作者签名：　　导师签名：

日期：　　年　　月　　日