E-Commerce Letters 电子商务评论, 2024, 13(4), 2570-2578
Published Online November 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.1341430

Hans汉斯

基于预测模型的金融产品选择影响因素研究

张　鑫

上海工程技术大学管理学院，上海

收稿日期：2024年7月26日；录用日期：2024年8月23日；发布日期：2024年11月12日

摘　要

随着金融市场的发展，金融产品的种类越来越多，消费者有了更多的选择。为了分析消费者对金融产品的选择并做出相关预测，一家金融数据咨询公司评估了客户对金融数据服务和数据库系统的需求和满意度。本文基于评估调查数据，分别采用随机森林分类算法(RF)、Adaboost分类算法、K近邻(KNN)分类算法建立预测模型；同时，针对每种模型开展特征重要性分析，探究不同因素对金融产品选择的影响程度。结果表明，K近邻(KNN)的分类模型的预测能力最佳；三种模型提供的可解释性基本符合实际规律，且特征重要性排序规律定性基本一致：存储指标2、使用时间指标和数据库大小对消费者选择金融产品影响显著，有无国家经济数据所占比例最低。

关键词

随机森林分类模型，Adaboost分类模型，K近邻(KNN)分类模型

Research on the Influencing Factors of Financial Product Selection Based on Predictive Models

Xin Zhang

School of Management, Shanghai University of Engineering Science, Shanghai

Received: Jul. 26th, 2024; accepted: Aug. 23rd, 2024; published: Nov. 12th, 2024

Abstract

With the development of the financial market, there are more and more types of financial products, and consumers have more choices. In order to analyze consumers' choice of financial products and make relevant predictions, a financial data consulting company assesses customers' demand and satisfaction with financial data services and database systems. Based on the survey data, Adaboost classification algorithm, K-nearest neighbor (KNN) classification algorithm and random forest classification

文章引用: 张鑫. 基于预测模型的金融产品选择影响因素研究[J]. 电子商务评论, 2024, 13(4): 2570-2578.
DOI: 10.12677/ecl.2024.1341430

algorithm (RF) were used to establish prediction models. At the same time, we carry out feature importance analysis for each model to explore the influence of different factors on the selection of financial products. The results show that K-nearest neighbor (KNN) classification model has the best prediction ability. The interpretability provided by the three models basically conforms to the actual law, and the characteristics of the importance ranking law are basically the same qualitatively: storage index 2, the use time index and the size of the database have a significant impact on consumers' choice of financial products, and the proportion of whether there is national economic data is the lowest.

Keywords

Random Forest Classification Model, Adaboost Classification Model, K-Nearest Neighbor (KNN) Classification Model

1. 引言

金融产品已经成为了现代人生活不可或缺的一部分，金融市场的发展也日趋成熟与完善。随着金融市场的发展，金融产品的种类越来越多，消费者有了更多的选择。同时，消费者的消费行为也受到了影响。消费者的需求不断变化，对金融产品的投资理念、风险偏好、消费心理等方面也提出了新的要求。因此，针对金融产品对消费者行为的分析预测就变得非常重要。

首先，金融机构需要通过对消费者行为的分析预测来制定相关的营销策略和产品策略。通过对消费者行为的研究，金融机构可以更好地了解消费者的需求和反馈，进而针对性地推出适合消费者的金融产品。同时，金融机构还需要根据消费者行为的变化来调整产品价格、投资期限、利率等方面的策略，以保持市场竞争力和用户黏性。

其次，消费者也需要通过对金融产品的分析预测来做出更明智的投资决策。对金融产品的分析预测可以帮助消费者更好地了解产品的优势和劣势，选择适合自己的产品。此外，通过对金融市场的分析预测，消费者可以更好地把握市场走势，严格控制风险，提升自身的投资收益。

因此，对金融产品对消费者行为的分析预测是金融机构和消费者都非常重要的研究方向，也是当前金融市场中的热门话题之一。

2. 文献综述

2.1. 国外相关文献综述

Bilias (2010)的研究表明，家庭结构对投资组合选择有显著影响，不同的家庭组成会导致投资决策存在较大差异[1]。Gollier (2011)认为，经验丰富的金融消费者会根据他们的投资经验调整理财策略和风险偏好，从而追求更高的收益[2]。El-Attar 和 Poschke (2011)发现，风险态度的差异直接影响理财方式，而金融消费者对信任的不同程度则会导致他们持有不同类型的理财产品。信任度较高的消费者往往更倾向于选择高风险的理财产品[3]。Georgarakos 和 Pasini (2011)指出，宗教文化对金融消费者的投资理财行为也有影响，不同的宗教文化背景会导致理财方式的差异[4]。Calvet 和 Sodini (2014)的研究进一步表明，教育程度对金融市场的参与度和理财行为具有影响，高学历的消费者参与金融市场的程度更高，并且会更合理地分配资金，而低学历的消费者则参与度较低[5]。Shih (2014)的研究发现，教育水平、金融知识和

金钱观念的不同都会对消费者的投资行为和资产配置产生影响[6]。

2.2. 国内相关文献综述

邹梦碧(2012)等人在研究消费者购买行为时指出，消费者的购买决策通常经历了从意向到行为的转化过程[7]。尹志超(2014)则指出，金融市场参与度受到知识水平的显著影响，同时也涉及经济水平、教育程度和风险偏好等因素的影响。他还强调，具有较高金融知识的消费者更倾向于参与金融市场，并且更愿意选择较高风险的金融资产[8]。汪瞻辰(2017)在探讨我国金融市场可持续发展的研究中，提出要推动我国绿色金融的长期发展，需要扩大金融产品的种类，优化市场供给的产品结构，从而满足不同风险偏好投资者的多元化需求[9]。丁嫚琪(2019)在分析金融投资者金融素养对投资行为的影响时，通过调研发现，投资者对金融产品知识的掌握至关重要，应将其视为投资者必备的关键素养之一。她还强调了关注投资者细分需求的必要性。金融投资产品的成功在于能否赢得投资者的认可[10]。

2.3. 文献评述

目前，关于金融产品选择意愿的研究，国外学者研究表明，家庭结构、经验、信任度、宗教文化、教育水平等因素显著影响金融消费者的投资决策和理财行为。国内学者则更多地研究投资者对互联网理财产品的选择意愿。通过整合技术接受与使用模型，国内研究强调了金融知识、风险偏好、市场结构等因素对投资决策和金融产品成功的关键作用，分析投资者对互联网金融产品的选择意愿的影响因素。大多数实证研究表明，这些因素对投资者的决策有显著的影响[11] [12]。

然而，针对用户选择互联网理财产品的专门预测模型尚未出现。通过对比分析已有的技术接受与使用模型，可以有效预测用户在选择金融产品时的决策因素。这种预测模型对传统机构和其他互联网理财平台具有重要的参考价值[13]。

3. 数据处理

3.1. 数据来源

数据为 2023 年第二届全国财经大数据处理综合技能大赛中赛题数据，数据记录了某公司是一家刚刚成立的金融数据咨询公司，主要业务包含：数据定制、数据咨询、数据分析等，现在公司收到客户记录的自己用户购买产品中包含的各项信息数据(数据已经过脱敏处理)，并预测下用户会购买的是哪个产品。全文所用数据来源于大赛数据。

3.2. 数据预处理

在进行数据集预处理时，首先要检查是否存在重复值，以免影响特征值的计算，导致模型预测出现偏差。客户信息表中的客户 ID 为唯一主键，经检查确认该表中无重复记录。然而，部分字段存在空值，为了保证用户画像分析的准确性，本文将这些空值记录直接删除。接下来需要检查数值型变量的异常值情况。数值型变量包括货币市场指标、是否支持多账号、以及网络状况等，检查后确认这些变量均无异常。本文将变量使用时间指标、有无货币市场指标、处理速度、是否支持多账号、网络情况、存储指标 1、数据范围指标、数据库大小指标、读取反应指标、数据质量指标、存储指标 2、搜索时长指标、有无国家经济数据、有无高频行情数据分别用 v1~v14 表示。最后，本文对分类变量进行了检查，确认也没有异常值存在。

3.3. 数据特征可视化

表 1 展示了各特征(自变量)的重要性比例。可以看出存储指标 2、使用时间指标、数据库大小等特征

对客户选择金融产品有较大的影响，有无国家经济数据对客户选择金融产品的影响较小。

Table 1. The proportion of importance of each feature
表 1. 各个特征的重要性比例

特征名称	特征重要性
使用时间指标	9.30%
有无货币市场指标	0.70%
处理速度	3.40%
是否支持多账号	0.50%
网络情况	0.90%
存储指标 1	4.60%
数据范围指标	2.70%
数据库大小指标	5.60%
读取反应指标	2.50%
数据质量指标	3.40%
存储指标 2	61.50%
搜索时长指标	3.70%
有无国家经济数据	0.60%
有无高频行情数据	0.60%

4. 模型建立

4.1. 随机森林分类模型

随机森林能够有效处理多维度、非线性数据，通过集成多棵决策树，捕捉复杂的特征间关系，并显著提升模型的预测准确性，其抗过拟合的能力使得模型在应对噪声数据时表现出色，保证了预测结果的稳健性。此外，随机森林天然支持特征重要性分析，能够识别并量化对投资决策影响最大的因素，为理解投资者行为提供了更深层次的洞见。综上，本文通过训练集数据来建立随机森林分类模型，将建立的随机森林分类模型应用到训练、测试数据，得到模型的分类评估结果。本文所用到的模型参数为：模型的训练时间为 0.157 秒，数据集以 8:2 的比例分为训练集和测试集，并且在训练前对数据进行了洗牌。模型未使用交叉验证，节点分裂的评价准则为 Gini 不纯度，随机森林中包含 100 棵决策树。构建每棵树时采用有放回抽样，未使用袋外数据进行测试。划分节点时考虑的最大特征比例为自动选择，每个内部节点分裂的最小样本数为 2，每个叶子节点的最小样本数为 1，叶子节点的最小样本权重为 0。决策树的最大深度为 10，叶子节点的最大数量为 50，节点划分的不纯度阈值为 0。

4.2. Adaboost 分类模型

Adaboost 通过迭代调整样本权重，逐步提升分类精度，并通过结合多个弱分类器形成强分类器，特别擅长处理复杂的分类问题，具有良好的偏差与方差平衡能力。本文通过训练集数据来建立 Adaboost 分类模型，将建立的随机森林分类模型应用到训练、测试数据，得到模型的分类评估结果。本文所用到的模型参数为：模型的训练时间为 0.25 秒，数据集以 8:2 的比例分为训练集和测试集，并且在训练前对数据进行了洗牌。模型未使用交叉验证，基分类器的数量为 100，学习率设置为 1。

4.3. K 近邻(KNN)分类模型

K 近邻分类模型(KNN)则以其基于实例的学习方式，通过计算测试样本与训练样本之间的距离进行分类，特别适合非线性数据分布。KNN 模型简单直观，适合处理小规模数据集，并能灵活应对多类别问题。本文通过训练集数据来建立 K 近邻分类模型(KNN)模型，将建立的随机森林分类模型应用到训练、测试数据，得到模型的分类评估结果。本文所用到的模型参数为：模型的训练时间为 0.003 秒，数据集以 8:2 的比例分为训练集和测试集，并且在训练前对数据进行了洗牌。模型未使用交叉验证，搜索算法为自动选择，叶节点的数量为 30，近邻数为 5，近邻样本的权重函数为均匀分布(uniform)，向量距离算法为切比雪夫距离(Chebyshev)。

5. 结果分析

5.1. 混淆矩阵热力图对比分析

5.1.1. 随机森林分类模型

图 1 提供了关于随机森林分类的直观展示，其中行代表实际类别，列代表预测类别，单元格中的值则表示观测数量。根据给出的测试数据混淆矩阵，分析可得：

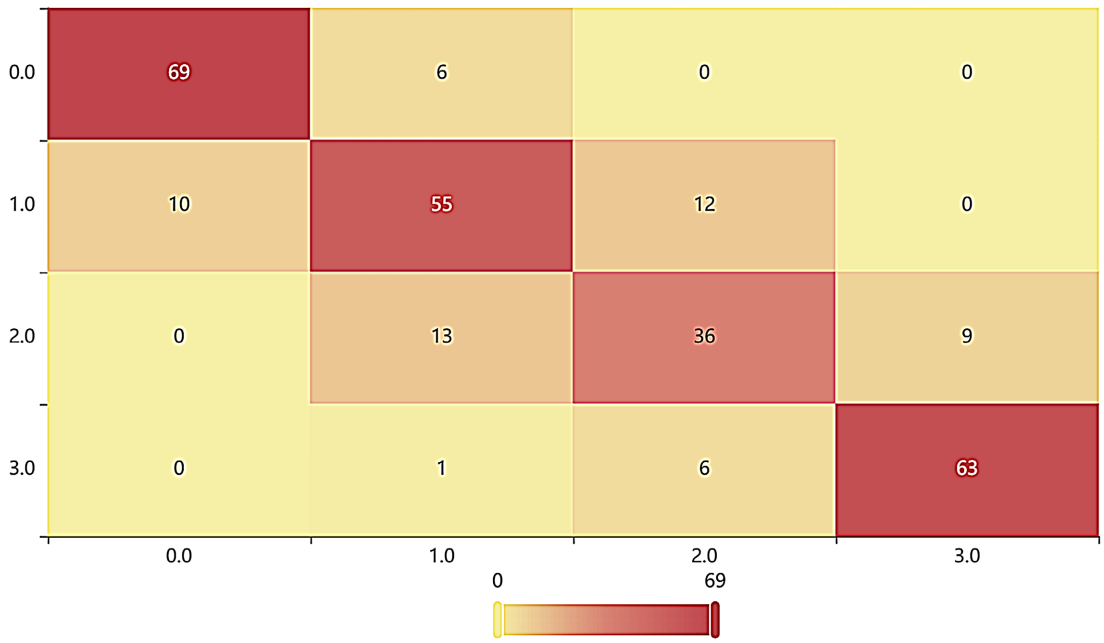

Figure 1. Confusion matrix heat map of random forest classification model
图 1. 随机森林分类模型的混淆矩阵热力图

首先，从主对角线上的值可以看出，分类器在大多数类别上的表现是准确的。例如，对于实际类别为 0 的样本，分类器正确预测了 75 个，占该类总样本的绝大多数。类似地，对于实际类别为 3 的样本，分类器也取得了 53 个正确预测，表现出较高的准确性。

然而，观察非主对角线上的值，我们发现分类器在某些情况下出现了误判。具体来说，对于实际类别为 0 的样本，有 6 个被错误地预测为类别 1；对于类别 1，有 10 个被错误地归类为类别 0，同时有 12 个和 0 个分别被误判为类别 2 和类别 3。这表明分类器在区分相邻类别时存在一定的困难，可能需要通过调整模型参数或特征选择来优化性能。

进一步分析混淆矩阵，我们发现类别 2 的样本在预测时出现了较为显著的误判。具体来说，有 13 个类别 2 的样本被错误地预测为类别 1，同时有 6 个类别 3 的样本被误判为类别 2。这提示我们，类别 2 与类别 1 和类别 3 之间的界限可能较为模糊，分类器难以准确区分。为了提高分类器在这些类别上的性能，可能需要进一步探索数据的特征空间，以寻找更有效的区分特征。

5.1.2. Adaboost 分类模型

图 2 提供了关于 Adaboost 分类的直观展示，其中行代表实际类别，列代表预测类别，单元格中的值则表示观测数量。根据给出的 Adaboost 分类模型测试数据混淆矩阵，分析可得：

首先，观察混淆矩阵热力图，可以看出矩阵中的数值以 0.0、1.0、2.0、3.0 为标签，分别代表不同的类别。矩阵的每一行表示实际类别，每一列表示模型预测的类别。矩阵中的数值则代表对应行列的样本数量。

从矩阵中可以看出，对角线上的数值(61、47、18、63)相对较大，这些数值表示模型正确分类的样本数量。以第一行第一列为例，数值 61 表示模型将 32 个标签为 0.0 的样本正确分类为 0.0。对角线上的高值反映了模型对各类别样本的较高分类准确率。

然而，矩阵中的非对角线元素也存在非零值，这些数值表示模型分类错误的样本数量。以第一行第二列为例，数值 9 表示模型将 9 个标签为 0.0 的样本错误地分类为 1.0。这些非零值反映了模型在分类过程中存在的误差。

进一步分析混淆矩阵，可以发现模型在类别 1.0 和 2.0 之间的分类误差较大。第二行第三列数值 10 表示模型错误地将 10 个标签为 1.0 的样本分类为 2.0。同时，第三行第二列的数值 19 也表明有 19 个标签为 2.0 的样本被错误地分类为 1.0。这些结果表明模型在区分 1.0 和 2.0 这两个类别时存在一定的困难。

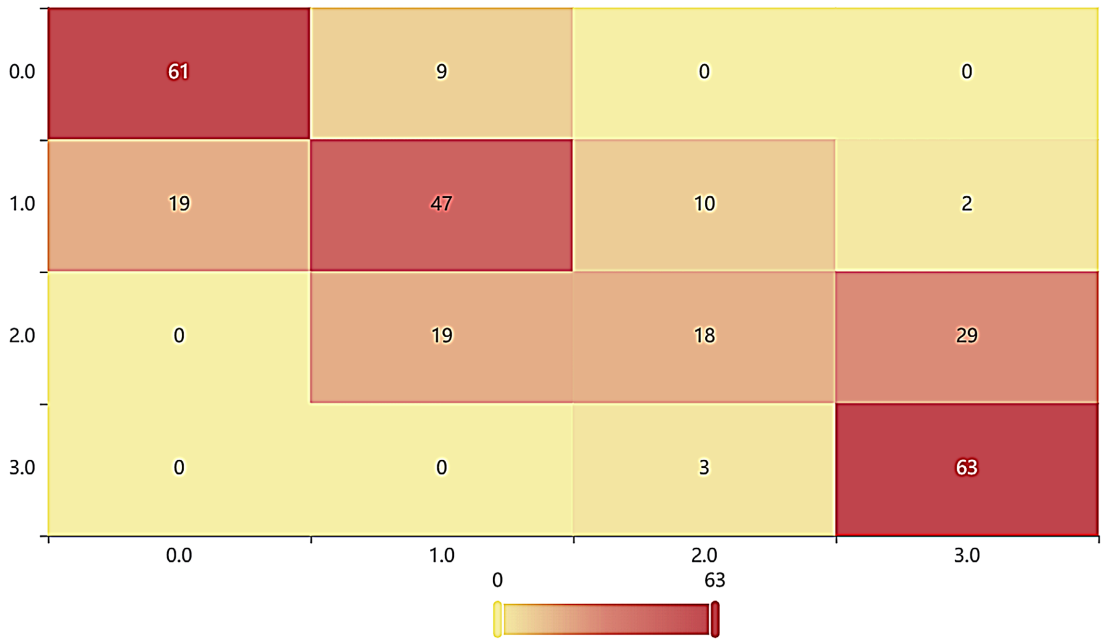

Figure 2. Confusion matrix heat map of Adaboost classification model
图 2. Adaboost 分类模型的混淆矩阵热力图

5.1.3. K 近邻(KNN)分类模型

图 3 提供了关于 K 近邻(KNN)分类的直观展示，其中行代表实际类别，列代表预测类别，单元格中

的值则表示观测数量。根据给出的 K 近邻(KNN)分类模型测试数据混淆矩阵，分析可得：

首先，观察混淆矩阵的对角线元素，它们表示正确分类的样本数量。具体来说，对于真实类别为 0 的样本，有 66 个被正确分类，而真实类别为 1、2、3 的样本中，分别有 54、55、52 个被正确分类。这表明 KNN 分类器在大多数情况下能够准确地识别样本的真实类别。

然而，也应注意到混淆矩阵中非对角线元素的存在，它们表示错误分类的样本数量。例如，在真实类别为 0 的样本中，有 8 个被错误地分类为类别 1；在真实类别为 1 的样本中，有 8 个被错误地分类为类别 0，同时有 11 个被错误地分类为类别 2；类似地，在真实类别为 2 和 3 的样本中，也存在错误分类的情况。这些错误分类揭示了 KNN 分类器在某些情况下的局限性，可能是由于数据分布的重叠、噪声数据的存在或 K 值的选择不当等原因导致的。

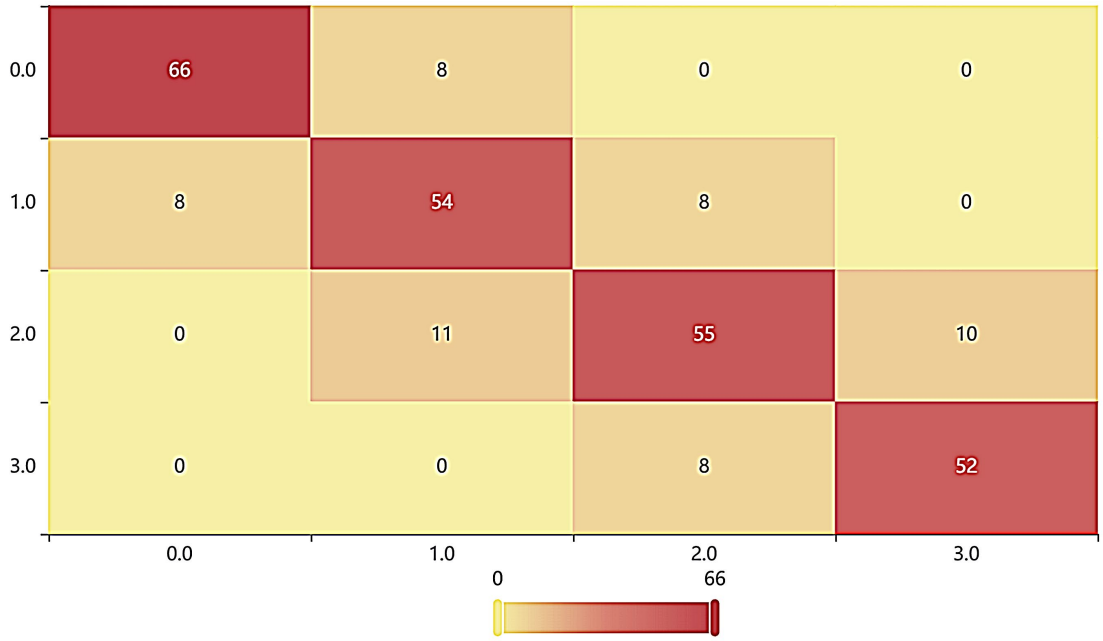

Figure 3. Confusion matrix heat map of K-nearest neighbor (KNN) classification model
图 3. K 近邻(KNN)分类模型的混淆矩阵热力图

5.2. 实验结果

随机森林分类模型的准确率为 0.796，召回率为 0.796，精确率为 0.792，F1 分数为 0.794；Adaboost 分类模型的准确率为 0.675，召回率为 0.675，精确率为 0.66，F1 分数为 0.648；K 近邻(KNN)分类模型的准确率为 0.811，召回率为 0.811，精确率为 0.811，F1 分数为 0.81 (见表 2)。综合来看，KNN 分类模型在各项指标上表现最好，准确率、召回率、精确率和 F1 分数均达到了 0.811，说明该模型在整体性能上

Table 2. Experiment results
表 2. 实验结果

模型名称	准确率	召回率	精确率	F1
随机森林分类模型	0.796	0.796	0.792	0.794
Adaboost 分类模型	0.675	0.675	0.66	0.648
K 近邻(KNN)分类模型	0.811	0.811	0.811	0.81

优于其他两个模型。随机森林分类模型紧随其后，各项指标也较为接近，但略逊于 KNN。而 Adaboost 分类模型在所有指标上的表现都相对较差，特别是 F1 分数，仅为 0.648，表明其在处理分类任务时的综合表现不如前两者。

6. 结论

本文为了分析消费者对金融产品的选择并做出相关预测，一家金融数据咨询公司评估了客户对金融数据服务和数据库系统的需求和满意度。基于评估调查数据，本文分别采用 Adaboost 分类算法、K 近邻(KNN)分类算法和随机森林分类算法(RF)建立预测模型，同时对每种模型进行特征重要性分析，以探讨不同因素对金融产品选择的影响。结果显示，K 近邻(KNN)分类模型的预测能力最强；三种模型提供的解释性与实际规律基本一致，特征重要性排序定性基本相同：存储指标 2、使用时间指标和数据库大小对消费者选择金融产品的影响显著，而有无国家经济数据的影响最小。基于上述研究结果，本文提出如下建议：

第一，加强金融产品的透明度与公平度。首先，它可以提高消费者对金融产品的信任度，促进金融市场的健康发展。其次，透明度和公平度可以帮助消费者更好地理解金融产品，从而更好地做出决策。此外，加强透明度和公平度还可以减少金融市场中的不当行为和欺诈行为，保护消费者的权益。可以通过以下措施来加强金融产品的透明度与公平度：首先，加强监管：政府和监管机构应该加强对金融产品的监管，确保金融产品的信息公开和销售行为的公平性。其次，提高消费者教育：消费者应该了解金融产品的基本知识和风险，从而更好地做出决策。最后，加强信息披露：金融机构应该提供充分的信息披露，包括产品的费用、风险、收益等方面的信息，让消费者更好地了解产品。总之，加强金融产品的透明度与公平度是非常重要的，可以促进金融市场的健康发展，保护消费者的权益。政府、监管机构和金融机构应该共同努力，采取有效措施，加强金融产品的透明度与公平度。

第二，强化消费者的金融素养教育。消费者的金融素养是评估他们在金融产品市场中的决策能力和风险认知能力的重要指标。消费者的金融素养水平越高，就越能够理性地选择适合自己的金融产品，防范金融风险。金融机构应该加强对消费者的金融素养教育。具体来说，金融机构应该制定并实施针对不同群体、不同阶段的消费者的金融素养计划，包括金融知识普及、理财技巧、风险管理和投资策略等方面的内容。政府部门也应该加强对消费者的金融素养教育，通过举办各种金融知识普及的讲座和培训课程，提高消费者的金融素养水平。

第三，完善金融监管制度与法律法规。金融监管制度与法律法规是保障金融市场健康发展的基础。完善金融监管制度与法律法规，对于促进金融市场的稳定和规范发展具有非常重要的意义。首先，要加强金融监管的力度，建立完善的监管制度和监管机制，从源头上规范金融产品的设计、销售和运作。同时，加强对金融机构的监督和执法，对于那些违反法律法规、损害消费者利益的金融机构，要严格依法处罚。其次，加强法律法规的制定和完善。要及时跟进国内外金融市场的发展，总结和借鉴其他国家的经验和做法，不断完善现有的法律法规，特别是针对新兴的金融产品，加强监管和规范。最后，金融监管部门应该加强与其他部门的协作，形成金融监管的合力。例如，与消费者权益保护部门合作，共同保障消费者的利益；与工商、税务等部门合作，共同打击金融诈骗等违法行为。

参考文献

[1] Bilias, Y., Georgarakos, D. and Haliassos, M. (2010) Portfolio Inertia and Stock Market Fluctuations. *Journal of Money, Credit and Banking*, **42**, 715-742. https://doi.org/10.1111/j.1538-4616.2010.00304.x

[2] Gollier, C. (2011) Portfolio Choices and Asset Prices: The Comparative Statics of Ambiguity Aversion. *The Review of Economic Studies*, **78**, 1329-1344. https://doi.org/10.1093/restud/rdr013

[3] El-Attar, M. and Poschke, M. (2011) Trust and the Choice between Housing and Financial Assets: Evidence from

Spanish Households. *Review of Finance*, **15**, 727-756. https://doi.org/10.1093/rof/rfq030

[4] Georgarakos, D. and Pasini, G. (2011) Trust, Sociability, and Stock Market Participation. *Review of Finance*, **15**, 693-725. https://doi.org/10.1093/rof/rfr028

[5] Calvet, L.E. and Sodini, P. (2014) Twin Picks: Disentangling the Determinants of Risk-Taking in Household Portfolios. *The Journal of Finance*, **69**, 867-906. https://doi.org/10.1111/jofi.12125

[6] Shih, K. (2016) The Impact of International Students on US Graduate Education. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2832996

[7] 邹梦碧, 陈文晶. 网络购买“意向-行为”过程及影响因素研究[J]. 北京邮电大学学报: 社会科学版, 2012, 14(3): 59-64.

[8] 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, 49(4): 62-75.

[9] 汪瞻辰. 我国绿色金融可持续发展的长效机制探索[J]. 农家参谋, 2017(22): 236.

[10] 丁嫚琪, 张立. 金融素养对我国家庭金融资产配置的影响研究[J]. 上海金融, 2019(3): 81-87.

[11] 魏银平. 基于认知偏差的金融消费者理财产品购买行为研究[D]: [硕士学位论文]. 蚌埠: 安徽财经大学, 2017.

[12] 邱紫莹. 个人投资者选择蚂蚁基金平台的影响因素研究[D]: [硕士学位论文]. 上海: 上海外国语大学, 2022.

[13] 庄湘霞, 朱长春, 杨盛. 互联网时代商业银行金融产品营销模式探讨[J]. 现代经济信息, 2016(19): 270.