基于 XGBoost 与弹性网络回归的集成模型对上证指数日极差的预测分析*

周 亚

摘 要：股票指数的日极差可以反映市场风险，对其进行预测有助于监管部门提前发现风险，采取相应举措。本文提出了基于 XGBoost 和弹性网络回归模型的集成模型，以多阶滞后的日极差和日收益率及其非线性项作为模型的输入，将其应用于上证指数日极差的预测分析。实验结果表明，该集成模型的预测能力要优于 XGBoost、弹性网络回归和 ARIMA 模型，能够有效地识别影响日极差预测的重要因素。该模型一方面可以为监管部门提供一个风险预估的新工具，另一方面所揭示的重要影响因素有助于科学研究和政策制定。

关键词：上证指数 价格极差 风险预测 XGBoost 弹性网络回归

一、引言

防范金融风险、维护国家金融安全是金融监管部门的重要职责之一。中国人民银行（2020）强调，防范化解重大风险是党的十九大确定的三大攻坚战之一，是决胜全面建成小康社会的重要举措。在党中央和国务院的指引下，相关部门不断强化风险意识，建立风险决策预警机制，取得了显著成效。

随着国家深化金融改革开放，中国证券市场在企业融资、国民资产配置等方面发挥着举足轻重的作用，在全球金融系统中占据着越来越重要的地位。作为我国市场经济的重要组成部分，证券市场风险极易扩散到其他国民经济部门，造成巨大的经济损失。为了有效识别风险，人们提出了多种风险度量指标，其中股票指数日波动方面的指标应用最为广泛。

中外学者提出了许多度量股票指数波动风险的方法。Parkinson（1980）提出用极差（最高价和最低价之差）来估计证券收益率波动。随后，许多学者都意识到极差的重要性，对 Parkinson（1980）的工作进行了拓展。一个方向是利用极差改进波动率的估计和预测，另一个方向是提出了更多的度量风险的指标，比如谢海滨等（2009）提出用价格波动幅度变化率来度量市场风险。

由于极差可以反映市场风险，许多学者致力于日极差序列预测的研究。Lee 等（2016）观察到日极差序列的异方差性（Heteroskedasticity）及波动聚集性（Volatility clustering），提出用高维时变模型来预测纳斯达克综合指数每日高低价的极差序列。与传统的自回归模型相比，Lee 等（2016）在其模型中考虑了更多的特征变量，包括多阶滞后的非线性项，这有利于模型获得更多的数据信息。

对于预测问题来说，不管是经典的统计模型方法，还是单个的机器学习模型，可能都不如集成模型的效果好。经典统计模型往往比较保守，强调模型的结构和理论性质，具有一定的可解释性，但对复杂数据结构的拟合不够。相对而言，机器学习放松了对结构和理论的严格要求，着重于模型的实际效果。而集成模型比单个的机器学习模型更灵活，能更好地适应金融领域内复杂的数据结构。Chen 和 Guestrin（2016）提出的 XGBoost 方法是一种基于分类回归树的集成模型，已被成功应用于多个领域。黄卿等（2018）将 XGBoost 应用到股指期货的变动方向的预测上，验证了该方法在我国金融领域的应用价值。而基于 XGBoost 和其他模型的进一步集成有可能得到一个预测能力更强的模型。

本文基于 XGBoost 和弹性网络回归模型构建了一个集成模型，以多阶滞后的日极差和日收益率及其非线性项作为模型的输入，将其应用于上证指数日极差序列的预测分析，并与 XGBoost、弹性网络回归及 ARIMA 模型进行比较。

* 基金项目：部分内容得到中国国家留学基金管理委员会资助。

二、模型及背景

（一）XGboost 和弹性网络回归

对于现实生活中的大多数机器学习和数据挖掘问题而言，梯度提升决策树（GBDT）算法是标准的解法之一。结合前人提出的二阶导方法，Chen 和 Guestrin（2016）对传统的 GBDT 进行了一些改进，提出了可以高效执行的 XGBoost。经过一段时间的发展，现已有多个实现 XGBoost 的开源包（Open source package），它们可以在多个语言环境下（如 R、python 等）稳定运行。

线性模型是最常用的统计模型之一，被广泛应用于许多领域。结合 LASSO 筛选重要变量的特性和岭回归防止多重共线性的优点，Zou 和 Hastie（2005）提出了弹性网络回归。该方法在实践中有较好的可解释性，是统计学里面的一个基准方法。著名的软件包 Glmnet 可以在多个语言环境下（如 R、python 等）高效执行弹性网络回归的估计和预测。

（二）模型的集成

大量研究表明，集成模型可以提高预测精度。常见的模型集成方法有三种：比例相加法、投票法、学习法。比例相加法是指将不同的模型线性组合在一起，它包括许多子方法，如简单平均法、加权平均法和直接相加法等等。前面提到的 XGBoost 就是采用了直接相加方法。从逼近误差（Approximation error）和估计误差（Estimation error）的角度来分析，XGBoost 有较低的逼近误差，而弹性网络回归有较低的估计误差。因此，考虑将这两个模型通过比例相加法结合起来。由于更关注预测精度，将利用验证集误差最小化准则对模型进行组合集成。

三、实证分析

记 C_t, H_t, L_t 分别为上证指数在第 t 个交易日的收盘价、最高价和最低价。基于这些记号，定义第 t 个交易日的收益率和极差分别为 $R_t = log\, C_t - log\, C_{t-1}$ 和 $Y_t = log\, H_t - log\, L_t$。先将时间序列的问题转化为监督学习的问题，然后用集成模型对上证指数的日极差序列进行预测分析，并与 XGBoost、弹性网络回归和 ARIMA 模型的预测结果进行比较。

（一）数据及监督学习框架

选取上证指数 1995 年 12 月 27 日至 2020 年 9 月 21 日期间的数据（共 6000 个交易日）进行分析，数据来源于网易财经。与 Lee 等（2016）的方法相似，考虑的数据集由时间的变动记号 t、日收益率 R_t 和日极差 Y_t 组成，即 $\{(t, R_t, Y_t), t = 1, \cdots, 6000\}$，其中 $t = 1$ 对应 1995 年 12 月 27 日，$t = 2$ 对应下一个交易日，以此类推，$t = 6000$ 对应 2020 年 9 月 21 日。

上述的数据类型不能直接用于 XGBoost 和弹性网络回归模型的训练。第一步需要将时间序列预测的问题转化成一个监督学习的问题。换句话说，需要构建模型的输入向量（或变量），这个过程有时也被称为“特征工程”。这个步骤非常重要，因为，只有当选择了与 Y_t 有相关性的变量组成输入向量时，才有可能建立一个有预测能力的模型，并且，相关性的强弱将直接决定模型预测能力所能达到的上限。Lee 等（2016）提出可以使用提前 20 个交易日（包含当前交易日）的信息来预测未来一个交易日的极差；此外，在构建输入向量时，考虑了日收益率的非线性项和时间变化，构建了 121 维的输入向量 x_t，x_t 由以下变量组成：$t/6000$，$Y_{t-1}, \cdots, Y_{t-20}$，$R_{t-1}, \cdots, R_{t-20}$，$R_{t-1}^2, \cdots, R_{t-20}^2$，$I_{R_{t-1}<0}, \cdots, I_{R_{t-20}<0}$，$R_{t-1}\, I_{R_{t-1}<0}, \cdots R_{t-20}\, I_{R_{t-20}<0}$，$R_{t-1}^2\, I_{R_{t-1}<0}, \cdots, R_{t-20}^2\, I_{R_{t-20}<0}$，其中 $I_{(\cdot)}$ 为示性函数（当括号中的式子成立时，它是 1，否则是 0）。本文采用 Lee 等（2016）的方法，将上证指数时间序列数据转化为监督学习类型的数据 $\{(x_t, Y_t), t = 21, \cdots, 6000\}$。基于此设定，极差序列的预测问题被转化成为监督学习的预测问题。把从 2019 年 9 月 10 日结束起，到 2020 年 9 月 21 日结束为止的 250 个交易日的数据当成测试集，即 $\{(x_t, Y_t), t = 5751, \cdots, 6000\}$。

（二）模型的训练与集成

XGBoost 和弹性网络回归模型都包含有超参数。超参数不能够通过算法学习得到，需人为预先设定，并且不同的应用场景需要不同的超参数。

对于时间序列的问题，不可以用常见的随机交叉验证方法来选择超参数。因为，当样本带有时间属性时，如果不考虑时间先后顺序而随机分配训练集和验证集，则在建模过程可能会存在数据泄漏（Data leakage）问题。在这里，数据泄漏意味着错误地将用于验证模型的部分信息作为已知信息对模型进行训练，这将直接导致最终模型的无效。为避免这个问题，截取一个时间点，该时间点之前出现的数据用于训练模型，该时间点之后出现的数据将作为一个用于选择超参数的验证集。选择 $t = 5500$ 作为

分界点，也就是说，将集合 $\{(x_t, Y_t), t = 5501, \cdots, 5750\}$ 的 250 个交易日作为验证集，集合 $\{(x_t, Y_t), t = 21, \cdots, 5500\}$ 作为训练集。由于更关注预测效果，在验证集选择参数时采取最小均方误差（MSE）准则。

如前所述，通过模型集成，可以得到更好的预测结果。记 $\widehat{\varphi}$ 和 $\widehat{\psi}$ 分别为 XGBoost 和弹性网络回归训练好的模型。把最终的集成模型表示为 $\widehat{m} = a\widehat{\varphi} + b\widehat{\psi}$，其中 a，b 是如下优化问题的解:

$$argmin_{a,b} \sum_{t=5501}^{5750} \{ Y_t - a\widehat{\varphi}(x_t) - b\widehat{\psi}(x_t) \}^2$$

除了 XGBoost 和弹性网络回归外，还考虑了支持向量机和随机森林等方法的集成，但是没有发现更好的预测效果。考虑到篇幅大小，本文仅呈现 XGBoost 和弹性网络回归的集成模型，以下将其简称为集成模型。

（三）评价与分析

在预测效果方面，本文以上述 MSE 作为评价指标。为评估集成模型的预测能力，除了比较它与 XGBoost 和弹性网络回归模型外，还考虑了 ARIMA 模型的日差预测结果。使用 R 语言的 auto. arima 函数将第 t 天以前的所有日极差信息用于 ARIMA 模型的参数估计，并使用 Forecast 函数来预测第 t 天的日极差。从表 1 中可以看出，XGBoost、弹性网络回归和集成模型的预测结果均优于 ARIMA 模型，说明它们能够有效地利用历史信息，更适合对日极差序列进行预测。对比 XGBoost 和弹性网络回归，发现 XGBoost 对训练集的拟合效果明显优于弹性网络回归，但它们在验证集和测试集的表现很相似。这间接地说明 XGBoost 的估计误差较大，而弹性网络回归的逼近误差较大。当把它们两者结合以后，预测效果提高约 4%，这证明了集成模型的有效性。

表 1　模型预测结果

	训练集	验证集	测试集
XGBoost	5.685×10^{-5}	4.670×10^{-5}	4.002×10^{-5}
弹性网络	9.577×10^{-5}	4.516×10^{-5}	4.026×10^{-5}
集成模型	7.481×10^{-5}	4.285×10^{-5}	3.856×10^{-5}
ARIMA	无	无	4.143×10^{-5}

用集合模型中的 XGBoost 和弹性网络回归，对输入特征在预测中的重要性进行了评估，结果如图 1 所示，其中，特征 1，2，…，120 分别对应于: Y_{t-1}，…，Y_{t-20}，R_{t-1}，…，R_{t-20}，R_{t-1}^2，…，R_{t-20}^2，$I_{R_{t-1}<0}$，…，$I_{R_{t-20}<0}$，$R_{t-1}\, I_{R_{t-1}<0}$，…$R_{t-20}\, I_{R_{t-20}<0}$，$R_{t-1}^2\, I_{R_{t-1}<0}$，…，$R_{t-20}^2\, I_{R_{t-20}<0}$。图 1（左）为 XGBoost 的增益图，纵坐标与特征相对应，横坐标与重要性相对应；图 1（右）为弹性网络回归的系数图，纵坐标与系数大小相对应，横坐标与特征相对应。从 XGBoost 的增益图中可以看出，对于预测未来 1 天的极差来说，历史日极差比日收益率更重要，5 个交易日内的信息最重要，而且时间越近，信息的重要性程度越高。根据弹性网络回归的系数图，可以得到三个结论: 第一，第 t 天的极差与前 20 天的日极差呈正相关性，且时间间隔越近，正相关性越强；第二，第 t 天的极差与前几天的负收益率呈负相关性，且时间间隔越近，负相关性越明显；第三，提前 5 个交易日内的信息对于预测日极差最为重要。总体来看，弹性网络回归的系数图与 XGBoost 增益图的结果相似，且相辅相成，说明该集成模型可以有效地识别影响预测的重要因素。

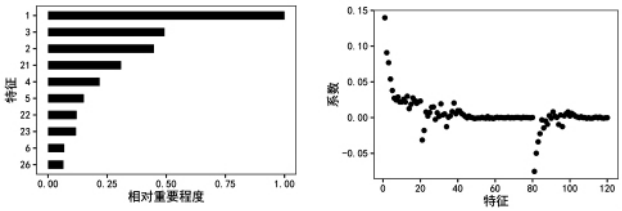

图 1　XGBoost 的特征增益（左）及弹性网络回归的系数（右）

四、结论与建议

本文基于 XGBoost 和弹性网络回归模型，利用机器学习的集成技术，对日极差时间序列建立了一个集成预测模型，并将其应用于上证指数日极差的预测分析。

实验结果表明，本文提出的集成模型的预测精度优于 XGBoost、弹性网络回归和 ARIMA 模型，并能有效识别影响日极差的重要因素。本文一方面可以提供一个日极差预测工具用于证券市场的风险预警，另一方面所揭示的重要影响因素有助于科学研究和政策制定。

基于本文的预测分析发现，集成模型表现出优异的预测效果，日收益率和日极差都与未来的日极差有相关关系，且 5 个交易日内的相关性最为明显。如果日收益率显著为负，或者日极差较大，那么未

来的日极差可能会显著增大。因此，给出以下建议:

监管部门可以采用集成模型预测未来的日极差，并将其作为一个风险预警指标，用于监控证券市场的风险。当该风险指标显著增大时，监管部门应及时介入，采取相应措施，防范风险扩散。

如果日极差或日收益率出现异常波动，监管部门需要提前做好准备应对未来一段时间内市场可能出现的风险。当异常持续出现，监管部门应该制定风险管理方案，并基于方案采取相应举措。从模型的角度来看，本文建议监管部门观测 20 个交易日以上的市场表现，至少连续 5 个交易日都没有明显的异常才能确保方案的有效性。

参考文献:

黄卿，谢合亮．机器学习方法在股指期货预测中的应用研究——基于 BP 神经网络、SVM 和 XGBoost 的比较分析［J］．数学的实践与认识，2018，48（8）：297－307.

谢海滨，邹国华，汪寿阳．价格波动幅度变动率——一个新的市场风险度量指标［J］．系统科学与数学，2009，29（11）：1460－1466.

中国人民银行．中国金融稳定报告（2020）［R/OL］，（2020－11－07）［2020－12－07］．http://www.pbc.gov.cn/goutongjiaoliu/113456/113469/4122054/index.html.

Chen T，Guestrin C. XGBoost: A Scalable Tree Boosting System［C］//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 785－794.

Lee E R，Mammen E. Local Linear Smoothing for Sparse High Dimensional Varying Coefficient Models［J］. Electronic Journal of Statistics，2016，10（1）：855－894.

Parkinson M. The Extreme Value Method for Estimating the Variance of the Rate of Return［J］. Journal of Business，1980，53：61－65.

Zou H，Hastie T. Regularization and Variable Selection Via the Elastic Net［J］. Journal of the Royal Statistical Society: Series B（Statistical Methodology），2005，67（2）：301－320.

作者单位：中国人民大学统计与大数据研究院

（上接第 64 页）

保护问题突出。金融发展的智能化和数字化使金融数据安全和个人数据隐私保护的问题日益突出，个人隐私数据被许多犯罪分子通过不法渠道获取，严重损害公众的利益。此外，肖钢等（2019）指出，由于行业应用中金融数据分散在不同机构、不同部门，成为割裂的“小数据”，从而形成了“数据孤岛”。

（二）政策建议

1. 增强对“程序员”的培养

本文把“程序员”的含义广义化，它不仅仅是传统意义上从事程序开发、程序维护的专业人员，同时也包括其他行业、职业以及专业兼修计算机知识的相关工作人员、学生和老师。加速推进人工智能技术的发展最终仍要落脚于研发人员，不仅要培养专业的程序员，而且要培养跨学科的复合型人才，从而在深入计算机领域研究的基础上，加速人工智能与其他领域的融合。作为科技领域的有力武器——人工智能，在将来的各个领域将发挥举足轻重的作用。

2. 适时、适度、有效监管

技术的问题仍需技术来解决，加速推进人工智能在监管领域的运用至关重要。比如，利用大数据追踪可疑人员信息，对其行为进行预测分析，提前把控风险，但同时要避免陷入“个人不是为他们的行为而是倾向负责”的陷阱，保护公正的概念。此外，可以参考国外实施监管沙盒机制，为新事物的发展提供空间，不能一管就死，同时监管要紧跟智能金融发展的步伐，修订完善监管原则。

3. 实现数据的均等化和安全性

大数据既是天使，也是魔鬼，关键在于谁使用、如何使用以及造成的结果如何。目前，数据的隐私保护问题极其突出，本文认为监管部门一方面应该设置数据使用责任制，对数据的非法使用者追责，而不再依赖于传统的个人许可。另一方面，对于数据孤岛现象，相关部门可以制定数据信息的反垄断条例，把传统的反垄断法更新转移到网络产业及其技术信息领域。此外，也可以建立一个大的金融云，通过“联邦学习”的方式，即各机构在不交换数据的情况下，完成人工智能算法模型训练，实现数据隐私保护下的共享。

参考文献:

李建军，李俊成．普惠金融与创业“授人以鱼”还是“授人以渔”？［J］．金融研究，2020（1）：69－87.

李建军，彭俞超，马思超．普惠金融与中国经济发展：多维度内涵与实证分析［J］．经济研究，2020（4）：37－48.

肖钢．中国智能金融发展报告（2019）［M］．北京：中国金融出版社，2020.

周利，封大威，易行健．数字普惠金融与城乡收入差距“数字红利”还是“数字鸿沟”［J］．经济学家，2020（5）：100－101.

张勋，万广华，张佳佳，等．数字经济、普惠金融与包容性增长［J］．经济研究，2019（8）：71－73.

朱太辉．智能金融发展的潜在风险及监管应对［J］．国际金融，2020（2）：30－34.