短期金融市场多步使用时间融合进行预测Transformer：与传统综合损失的模型与探索功能

作者：郝泽阳，硕士

已提交至诺丁汉大学

2024年9月，

部分满足授予学位的条件

计算机科学硕士（人工智能）

我声明本论文全部为我自己的作品，除非文中另有说明

抽象的

本文旨在研究时间融合变换器 (TFT) 模型在预测金融股票指数方面的表现,主要侧重于预测趋势。在评估中,我们选取了大约五年的每日股市数据来评估 TFT 在短期内进行多步预测的表现。与 LSTM 和 GRU 模型相比,我们发现 TFT 模型在从数据中捕捉趋势变化方面具有更好的稳定性和准确性。但是,当应用于同一数据集时,LSTM 和 GRU 的性能并不完全匹配。此外,该研究还探讨了如何结合各种损失函数,并分析了在损失函数中包含相关性度量的效果。研究发现,在损失函数中添加 Spearman 的 rho (ρ) 有助于获得一致的趋势预测结果。本文还提出了使用自动技术调整损失函数中权重分布方式的想法,这表明这可能是一个有前途的领域。不过,在测试中,GradNorm 方法未能如预期般提升性能,这可能与损失函数组合的选择有关。该研究为时间序列数据的短期多步预测提供了新方法,强调了使用 TFT 模型预测金融市场趋势的可能好处,并为未来的研究提供了宝贵的见解。

关键字:
时间融合变换器;深度学习;LSTM;GRU;损失函数

致谢

我非常感谢我的导师 Jamie Twycross 教授,他的宝贵指导和深刻建议对我的研究方向起到了重要作用。

目录

1. 简介

世界政局动荡,新冠疫情后全球经济复苏艰难,新旧产业碰撞,影响到广大的金融市场（Goldstein等，2021）。随着金融市场波动性的增加,对股票等金融数据的预测变得越来越困难。由于市场不再呈现明显的趋势,人们的短期投资不再有稳定的收益,这让人们对新闻、政策和数据更加敏感,再次加剧了微观金融市场的波动,投资者不再有信心（Ullah,2022）。

尤其是2022年至2023年,全球金融市场经历了多次不稳定事件,暴露了银行业和衍生品市场的脆弱性。尽管这些危机事件彼此独立,但它们有共同的触发因素,例如通胀上升、利率上升和个别紧急情况。这些事件凸显了金融市场在面对不确定性时的脆弱性,并表明当金融稳定性减弱时,恐慌会引发自我实现的危机（《金融稳定报告》,2023年5月）。

同样,欧洲央行在2023年5月的《金融稳定报告》中提到,高通胀和货币紧缩给市场发展带来重大压力,导致收益率上升和资产价格波动（欧洲央行,2023年）。

在金融领域,深度学习的应用范围越来越广。深度学习方法不仅在预测金融资产价格方面发挥着重要作用,而且在分析市场动向、发现潜在投资机会、优化投资组合等领域也得到越来越多的应用（Sezer 等人，2019 年）。这些工具为机构和投资者提供了强大的能力,帮助他们应对经常面临的复杂金融状况（Heaton 等人,2016 年）。

股票市场预测、基于算法的交易和风险控制都大量采用了深度学习。研究和应用深度学习系统的人可以处理大量财务数据和非结构化信息,例如新闻媒体的文章或社交媒体帖子。这反过来又可以导致对市场结果的预测更加准确（Olorunnimbe & Viktor,2022 年）。

随着数据收集越来越频繁,深度学习中新的模型不断涌现,它们可以通过各种方式在数据中发现复杂的特征（Sezer et al., 2019）。时间融合 Transformers 是一种主要用于分多步预测时间序列的深度学习模型（Lim et al., 2019）。TFT 汇集了传统 Transformer 结构中的内容,并将其与长短期记忆 (LSTM) 网络相结合,增加了注意力机制,以更好地关注数据中的长程依赖关系。TFT 不仅可以处理单变量和多变量时间序列,还可以应用于不同类型的时间序列,并在保持其可解释性的同时做出更准确的预测。

此外,TFT 在处理更复杂的基于时间的财务数据方面具有几大优势,模型分析复杂数据的能力是其主要优势（Laborda 等人,2023 年）,可解释性被视为关键优势。经过训练后,模型可以显示哪些已知变量在训练期间具有更高的重要性,以及哪些未知变量在预测期间很重要。这有助于涵盖主要

深度学习模型的弱点是它们通常具有高性能但解释结果的能力较低。

TFT还可以实现训练过程中注意力学习的可视化,这对于识别和研究重大事件和经济模型具有重要的参考价值（Laborda等,2023）。

损失函数的应用一直是深度学习的一个重要方面。
选择正确的损失函数可以显著提高模型性能（Jaiswal & Singh,2023）。不同的损失函数决定了模型在学习过程中受到的惩罚和奖励机制。例如,TFT的默认损失函数是分位数损失,相当于优化模型预测的高估和低估指标,范围可以手动调整。它在金融风险管理中非常有帮助,因为它有助于理解极端损失（Terven等,2023）。

MSE 也是一种常见的损失函数,它可以惩罚异常值过高的预测,对模型的预测有一定的稳定作用（Dosovitskiy, A. 2020）。
平均方向准确度 (MDA) 是衡量预测模型准确度的指标。MDA 计算模型正确预测的上升或下降趋势的比例 (Vishwesh, 2023)。Spearman 相关系数是一种非参数统计数据,用于衡量两个变量之间的单调关系 (MacFarland & Yates, 2016)。

组合损失的研究同样具有重要意义,对于一些属性不同或者互斥的目标,可能需要多个损失函数来表达期望。

使用具有适当权重的损失函数可以使模型的预测结果表现出均衡的性能（Dosovitskiy,A.2020）。

本研究旨在以日为单位预测短期周期框架内的股票收盘价,并在不同数据集上应用和比较不同的模型。然后,单独使用 TFT 模型探索组合损失函数,并在相同训练模式下评估不同的组合损失函数。详细信息将在后续章节中展示。

1.1 问题描述

本研究主要关注Temporal Fusion Transformer（TFT)模型与其他经典模型如LSTM、GRU等的性能差异,以及组合损失函数是否能提供更优的性能。

数据来源为三只热门股票（NVDA,TSLA,AAPL）,使用三个模型进行预测。通过随机化编码器长度等各种不稳定参数（较长的编码器将获得较长的历史数据,这可能带来更好的性能,但可能会过拟合）,固定随机种子和超参数,然后训练多次并取平均值作为评估值。

其次,除了TFT默认的损失函数外,对TFT模型的不同损失函数进行了比较。为了探索其他损失函数,便于组合损失
后续研究分别实现了MSE与MDA融合的损失函数、以及采用MSE、Spearman相关系数、分位数损失加权融合的损失函数。

上述组合损失函数均依赖于手动初始化权重来设置不同损失目标的重要性。GradNorm 是一种自动权重控制方法,可在训练过程中实现权重的动态变化（Chen et al.,2017）,可实现

动态权重的目的。研究结束时的目标是利用此方法实现自动化动态权重。

2. 文献综述

使用模型预测数据一直是研究人员的热门选择。通常,使用深度学习技术预测数据需要不同领域的专业知识（Jordan & Mitchell,2015）。许多文献支持特定数据在金融股票市场中的重要性、深度学习技术的可行性、TFT模型的优势以及该技术在各个领域的卓越性能。此外,一些研究探讨了损失函数在深度学习模型中的关键作用。在本节中,我们将回顾这些领域的主要参考文献,分析它们提供的理论基础,并探索它们在应用中的相互联系。

在财务方面的理论支持方面,Lev（1983)在其研究中探讨了影响公司盈利时间序列特征的经济因素。文章详细分析了不同财务指标（如收入、净利润等)的动态行为及其对盈利预测的影响。Lev的研究表明,公司的盈利能力不仅受到内部管理效率的影响,还受到外部经济环境、市场竞争程度以及行业的周期性等因素的影响。

Lev 通过建立数学模型展示了这些因素如何影响企业盈利的稳定性和可预测性,并提出了时间序列分析在财务预测中的重要应用。该研究为理解企业盈利行为的经济驱动因素提供了理论支持,为后续财务分析和企业估值领域的研究奠定了基础（Lev,1983）。其中一项研究引起了人们对不同股票和指数的统计特性的重要性的关注

标准差、偏度和峰度　影响风险评估和投资决策（Campbell 等,1997）。此外,由于深度学习模型在训练过程中可以自动学习市场动态中如此复杂的非线性趋势,这增强了它们分析和预测股市数据的能力（Fischer & Krauss,2018）。这些研究成果在时间序列历史数据之外的特征数据收集中发挥了至关重要的作用。在后续研究中,我们将探讨这些数据的重要性以及预测性能的改进。在各种预测方法中,深度学习模型可以取得良好的预测效果。

Olorunnimbe 和 Viktor (2022) 在其系统综述中详细讨论了深度学习在股票市场中的应用,涵盖实践、回测和具体应用领域。文章分析了各种深度学习模型在股价预测、投资组合优化、市场趋势分析等方面的应用,并深入讨论了这些模型在实际操作中的表现和挑战。研究指出,尽管深度学习在处理大规模金融数据方面表现出巨大潜力,但仍面临数据质量、模型可解释性和过度拟合等问题。通过对现有文献的系统回顾,作者总结了深度学习在股票市场应用中的成功案例和不足,并指出了未来研究的方向,特别是如何提高模型在股票市场中的可解释性和鲁棒性。

实际应用（Olorunnimbe & Viktor,2022）。可以看出,深度学习技术在股市预测中是可行的,而深度学习预测的主要方向是可解释性和更好的性能。因此,TFT因其独特的可解释性和优异的性能成为本研究的重点。

前沿算法模型是本研究的重点,Lim et al. (2021) 在他们的研究中提出了时间融合变换器 (TFT) 模型,这是一种用于多水平时间序列预测的深度学习模型,旨在解决传统时间序列模型在处理复杂依赖关系和非线性关系时的局限性。与早期基于 LSTM 或 GRU 的模型相比,TFT 不仅像旧模型一样捕获长期依赖结构,而且还通过结合多头注意力和静态特征嵌入来努力提高模型的可解释性。通过这种方式,TFT 模型可以在每个预测步骤解释这些关键驱动因素的同时动态选择关键的时间序列特征,从而为预测过程提供透明度,保持高精度（这在金融市场和需求预测应用中至关重要）,并让领域专家认为它是正确的。TFT 模型针对 10000 个时间步的公共金融数据的损失进一步优化,取得了最先进的结果。损失值随 epoch 数的增加而减小,最差表现出现在一个 epoch 之后。广播公司可以使用这项自动化服务在多个平台上更快、更轻松地分发体育内容。TFT 综合利用了深度学习和增强的可解释性,代表了时间序列预测领域的重要进步（Lim et al.,2021）。在本研究中,以 TFT 模型为核心的预测技术表现出很强的可解释性,填补了深度学习领域的空白,这也是本研究的重点。然而,在调试模型时,损失函数的设置可能会产生重大影响。

不同的损失函数会对模型产生显著的影响。在 Jaiswal 和 Singh (2023) 的研究中,他们针对深度神经网络在时间序列分析中的各种损失函数进行了比较研究。文章系统地分析了不同损失函数对模型性能的影响,特别是在处理复杂时间序列数据时。研究涵盖了常见的损失函数,如均方误差 (MSE)、平均绝对误差 (MAE)、对数损失 (Log Loss) 等,并深入讨论了这些损失函数在不同应用场景中的优缺点。通过实验结果,可以发现某种损失函数在特定情况下能够更好地捕捉数据的特征,从而提高模型预测精度。此外,论文还讨论了损失函数选择对建模过程的影响,特别是对收敛速度和模型鲁棒性的影响。Jaiswal 和 Singh 的研究为时间序列研究的损失函数正确选择提供了有用的信息,为该领域的未来研究铺平了道路 (Jaiswal & Singh, 2023)。在金融领域,还有一些特殊的评估方法值得注意。

财务预测的两种评估方法通常使用 MDA 和 Spearman 等级相关系数。谈到 Spearman 等级相关系数，McFarland 和 Yates (2016) 深入讨论了这种非参数统计方法,该方法通常用于评估大多数非线性关系中的变量之间的排名关系。Spearman 相关系数比较两个变量的排名之间的差异,这对于分析资产的相关性非常有用

金融市场预测模型或验证模型预测结果的有效性（McFarland & Yates,2016）。另一方面,Vishwesh（2023）将MDA作为时间序列预测的重要评价指标。MDA主要用于评估模型预测方向的准确性,即预测值与实际值在方向上的一致性。它在金融预测中特别有用,因为方向性预测（如价格上涨或下跌的趋势)往往比具体的数值预测更重要。通过结合MDA和Spearman相关系数,这两种方法可以提供更全面的评价视角,不仅可以衡量预测的方向准确性,还可以评估变量之间的关系,从而为金融市场的预测模型提供更有力的理论和实践支持（Vishwesh，2023）。将这两种能够反映趋势的评价方法添加到损失函数中进行探索,旨在使模型对趋势的预测更加准确。因此,组合损失函数可能是一个可行的解决方案。

组合损失函数的探索也是本研究的重要部分。

Dosovitskiy (2020) 在他的研究博客中讨论了多个损失函数优化问题,并提出了一种损失条件训练的方法。该方法旨在解决机器学习模型训练过程中多个损失函数之间存在冲突或竞争时的优化挑战。传统方法通常通过加权求和来组合多个损失函数,但这种方法很难找到不同任务之间的最佳权衡点。Dosovitskiy 的损失条件训练方法引入了一个控制变量,使模型能够根据不同的训练阶段或数据特征动态调整损失函数的权重,从而更有效地平衡多个目标的优化。这项研究证明了该方法在实际应用中的优势,特别是在需要同时优化多个目标的复杂任务中。该方法可以显著提高模型的性能,并提供更灵活的优化策略 (Dosovitskiy, 2020)。本文探讨了在复杂目标任务上组合损失函数的可行性。在金融预测领域,模型经常由于过度拟合或不遵循趋势而导致性能不佳。这通常是由于损失函数不合适造成的。灵活的损失函数组合和组合损失函数可能会由于应用或选择错误而表现出更好的性能。

在金融时间序列预测研究中,曾等 (2023) 深入分析了卷积神经网络 (CNN) 和 Transformer 模型在捕捉金融数据时间依赖性方面的表现。研究表明,CNN 通过其强大的局部特征提取能力可以有效识别金融数据中的短期模式,而 Transformer 模型则凭借其自注意力机制在处理长期依赖性方面表现出显著优势 (Zeng et al., 2023)。作者还比较了两种模型在不同金融市场数据集上的预测准确率,发现结合 CNN 和 Transformer 的混合模型在多次实验中取得了更高的准确率。这项研究为金融时间序列预测提供了新的方法论视角,展示了深度学习模型在复杂金融环境中的巨大潜力,并为未来的研究指明了方向。

Lim 等人提出的 Temporal Fusion Transformer (TFT) 结合了 LSTM 的门控机制和 Transformer 的自注意力机制,证明了其

强大的处理复杂数据的能力。将 LSTM 与 Transformer 相结合可以在捕获长期依赖关系和实现高性能融合方面表现出显著的优势。对 TFT 模型的解释性理解尤为关键，尤其是对 LSTM 的门控机制的深刻理解。Karpathy 等人（2015）使用可视化技术揭示了 LSTM 在处理自然语言和序列数据时的工作原理，展示了它如何捕获和存储长期依赖关系。他们的研究为 LSTM 和 RNN 的可解释性提供了新的见解，这对于金融时间序列预测中的可解释性需求至关重要。

在《宏观经济预测》中，Laborda、Ruano 和 Zamanillo（2023）探讨了 TFT 模型在多国 GDP 预测中的应用。研究表明，TFT 模型可以通过多头注意力机制和静态特征嵌入有效捕捉复杂的时间序列依赖关系，在不同时间范围内表现出出色的预测性能。该研究不仅证明了 TFT 在单个国家 GDP 预测中的高准确率，还证明了其在同时预测多个国家经济数据时的稳定性和可靠性。这为 TFT 在跨国经济分析和宏观经济预测中的应用提供了新的参考，凸显了其在处理复杂经济数据方面的有效性（Laborda 等，2023）

Santos 等（2022）研究了 TFT 在光伏发电量提前预测中的应用。研究表明，TFT 通过多头注意力机制和静态特征嵌入可以有效捕捉光伏发电数据中的时间依赖性和非线性关系，特别是在处理天气变化带来的不确定性时，TFT 可以提供更准确的预测结果。实验验证了 TFT 在提高光伏发电量预测精度和模型稳定性方面的优势，并指出了其在可再生能源管理中的潜力，为未来的研究提供了重要参考（Santos 等，2022）。

Giacomazzi、Haag 和 Hopf (2023) 探讨了 TFT 在短期电力负荷预测中的应用，重点研究了电网层级和数据源选择对模型性能的影响。研究发现，TFT 在处理复杂电力负荷时间序列数据时可以有效捕捉数据中的时间依赖性，并在不同电网层级中表现出良好的适应性。作者还指出，多种数据源的融合有助于提高预测精度，并展示了 TFT 在复杂能源系统中的广阔应用前景 (Giacomazzi et al.，2023)。该研究为短期电力负荷预测提供了新的方法参考，也为 TFT 在更大规模电网中的应用奠定了基础。

这些研究展示了TFT在处理复杂时间序列数据方面的强大能力，在金融、能源、经济等领域展现出明显优势。但研究也指出了TFT在某些应用场景下可能遇到的挑战，为进一步的模型优化和实际应用提供了宝贵经验。

3. 方法论与技术

本研究将利用一种新的深度学习模型Temporal Fusion Transformer，预测三家不同公司股票数据未来一周的收盘价。接下来，我们将介绍TFT的结构、算法的原理以及设计细节和步骤

本研究的最后部分。此外，我们将探讨组合损失函数的可行性，并展示整体的评估结果。

3.1 经典模型回顾

3.1.1 LSTM

LSTM（长短期记忆网络）是为了解决传统 RNN 模型中的“长程依赖”问题而设计的。在训练传统 RNN 时，随着时间步长的增加，模型的梯度可能会消失或爆炸，从而难以有效捕捉长时间序列中的依赖关系。LSTM 通过引入记忆单元和三个核心门控机制（输入门、遗忘门、输出门）有效地解决了这个问题（Karpathy et al., 2015）。这些门控机制帮助 LSTM 选择性地记住或遗忘信息，从而保证梯度能够沿着时间轴平滑地传播，避免出现梯度消失或爆炸的现象。

此外，LSTM 的一些单元可以自动学习和识别特定的可解释特征。例如，行长度计数单元通过检测输入字符逐渐降低其激活值，直到检测到换行符，表明它在跟踪文本行的长度。同样，引语检测单元在遇到引语时激活，离开引语后返回非激活状态。这种现象在图 1 中很明显，它显示了 LSTM 单元的激活状态如何根据输入内容而变化，验证了其强大的处理结构化信息的能力（Karpathy 等，2015）。

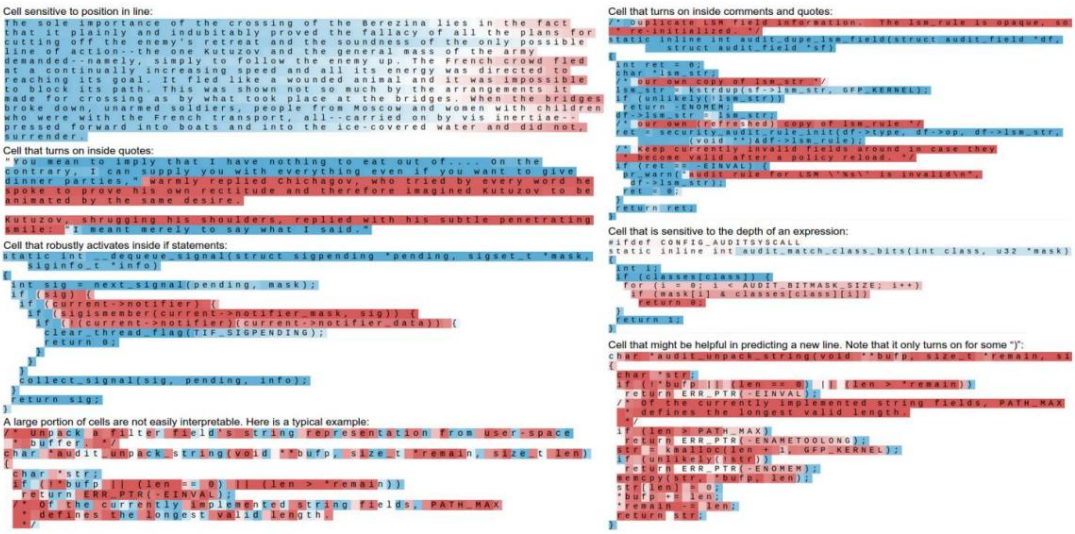

图 1：具有可解释激活的细胞的几个示例（Karpathy 等人，2015 年）

3.1.2 格鲁吉亚单元

门控循环单元（GRU）是循环神经网络（RNN）的一种变体，旨在通过简化模型结构来提高计算效率，同时保持捕获长期序列数据的能力。GRU 的核心在于两个门控机制：重置门和更新门（Jozefowicz 等，2015）。重置门控制从前一个状态传递到当前状态的信息量，更新门控制从前一个状态传递到当前状态的信息量。

gate 决定了当前状态与前一个状态的混合程度,通过这两个门控机制,GRU 可以有效地处理长期依赖问题,而不需要依赖复杂的内部结构。

相比于LSTM,GRU引入了LSTM的输入门和遗忘门,使得GRU计算更加简洁,但同时保留了捕获长期依赖信息的能力。如图2所示,在处理时间序列数据时,GRU通过重置门和更新门的协同作用,选择性地保留或丢弃信息,以更新隐藏状态H_t。Jozefowicz等（2015)指出,尽管结构简化,GRU在很多任务上的表现与LSTM相当,甚至在某些情况下优于LSTM（Jozefowicz等,2015）

该图清晰地展示了GRU的内部机制,帮助我们直观地理解GRU是如何通过这两个门控机制来控制信息流动的。
图中的重置门r_t和更新门z_t分别控制如何将之前的隐藏状态H_{t-1}和当前输入X_t结合起来生成新的隐藏状态H_t。
这样的设计使得GRU在训练时更加高效,同时保持了其处理复杂时间序列数据的能力。

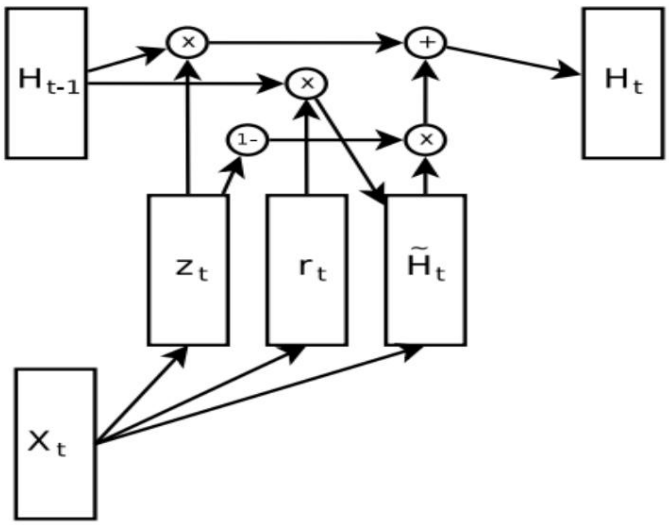

图 2 门控循环单元（Jozefowicz 等人,2015 年）

3.2 时间融合变压器（TFT）

Temporal Fusion Transformer（TFT)是一个为多步时间序列预测而设计的模型。它结合了 LSTM 的序列处理能力和 Transformer 的注意力机制,可以在各种复杂条件下做出高效的预测。基于这两个图,我们可以深刻理解 TFT 的工作原理以及它为什么适合时间序列预测,以及它的可解释性来源（Lim et al.,2019）。

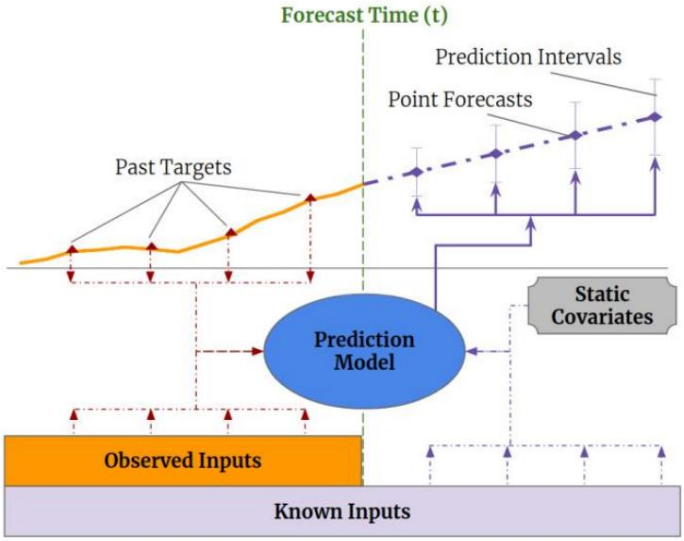

图 3 具有静态协变量、过去观察到的和先验已知的未来时间相关输入的多视野预测图解（Lim et al.,2019）

图 3 概述了 TFT 的整体框架,展示了模型如何处理过去的目标值、已知的未来输入和观察到的输入,并结合静态协变量来生成预测。TFT 模型将这些不同类型的数据输入到预测模型中,以生成点预测和预测区间。这种设计使 TFT 能够处理复杂的时间序列数据,尤其是在处理具有多个变量和不同时间范围的预测任务时。

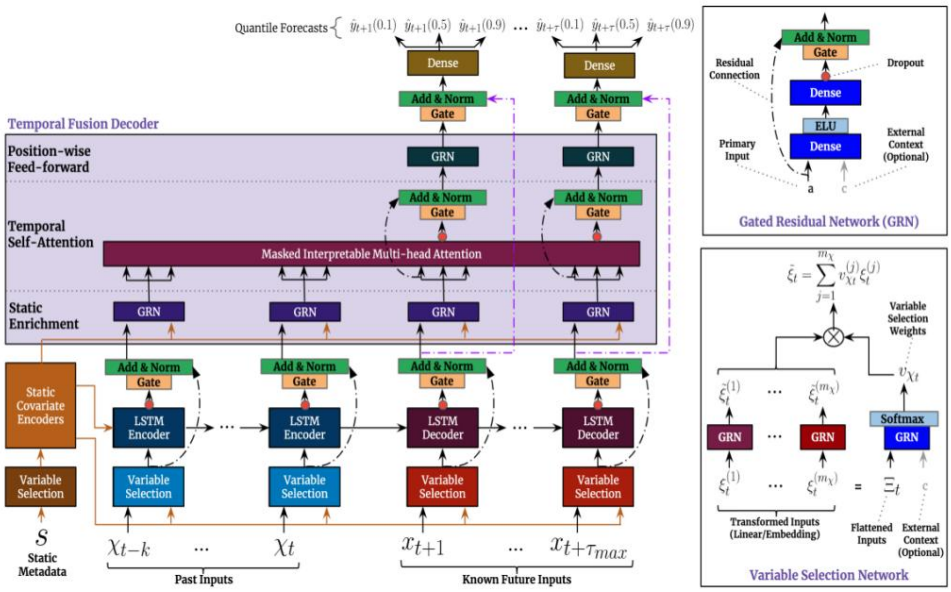

图 4 TFT 架构（Lim et al.,2019）

图 4 展示了 TFT（Temporal Fusion Transformer)的核心结构,包括静态协变量编码器、LSTM 编码器和解码器、多头注意力机制和变量选择网络等关键模块。这些组件共同处理

不同类型的输入数据。LSTM 用于捕获时间序列中的长期依赖关系,而多头注意力机制可帮助模型关注最相关的特征,以提高预测准确性。

TFT 在时间序列预测方面具有优势,可以灵活处理多个输入特征并捕捉时间步骤之间复杂的依赖关系。LSTM 模块从历史数据中提取与时间相关的特征,注意力机制动态选择最重要的特征。变量选择网络在每个时间步骤动态选择相关输入,这对于处理复杂数据的时间序列任务特别有效。

此外,TFT 的门控残差网络（GRN)通过决定是否跳过某些层来保持模型的高效。

TFT的可解释性源自其多模块设计:静态特征模块有助于理解静态变量的作用,注意力机制通过权重分布解释模型的关注点,变量选择网络提供每个时间步骤输入特征的重要性分析。这样的设计不仅提升了模型的预测能力,也增强了模型决策过程的透明度和可解释性。

3.3 损失函数

1. 分位数损失

TFT 的默认损失函数是联合最小化分位数损失（Wen et al.,2017）
并输出所有分位数输出的总和作为总体损失。

2. 平均方向精度

平均方向准确度（MDA)是评估时间序列预测模型趋势方向的关键指标,衡量模型预测的趋势方向（如上升或下降）与实际情况的一致程度,如果预测方向与实际趋势一致,则认为预测准确。MDA 的取值范围为 0 至 1,值越接近 1,模型在方向性判断上的准确度越高。

Vishwesh（2023)进一步强调了MDA在处理小样本和噪声数据集方面的作用,指出MDA的独特之处在于它不依赖于特定的预测值与实际值之间的差异,而是关注趋势方向的正确性。因此,MDA已成为评估趋势预测模型有效性的重要工具,尤其是对于数据波动较大或不稳定的场景。

3. Spearman 等级差异相关系数

Spearman秩差相关系数（Spearman s ρ)是一个非参数统计指标,用于衡量两个变量之间的单调关系。
Spearman s ρ的计算步骤如下:

1. 对数据进行排序:首先,对每个变量中的数据进行排序,并将每个值替换为其
数据集中对应的排名。

2. 计算秩差:对于每对观测值,计算秩
两个变量之间的差异。

3. 秩差平方和:将所有秩差的平方和得出

结果。

Spearman ρ 的取值在 -1 到 1 之间,+1 表示完全正相关，-1 表示完全负相关,0 表示不相关。MacFarland 和 Yates（2016)强调 Spearman ρ 的优点在于它不要求数据满足正态分布或线性关系,这使得它在处理非线性关系或排序数据时特别有用。在本研究中,尝试将 Spearman ρ 添加到组合损失函数中。Spearman ρ 指标对于股票的涨跌非常有用,可以显示趋势相关性。

3.4 方法论

本节将介绍模型训练、验证和测试的内容、超参数的使用以及其他细节。

3.4.1 研究总体设计

本研究中模型性能对比部分采用固定超参数、可控编码器长度、固定解码器长度、随机化数据子集（见图5）、固定全局训练随机数的方式使用模型进行预测。对输入数据进行分类测试,加入开盘价、最高价、最低价、收盘价、调整收盘价、成交量等单只股票时间序列数据,加入Revenue、Net Income、ESP、Operating Income、EBITDA、Current Ratio等财务特征的时间序列数据,以及“AI”、“Nvidia”、“Chatgpt”等与当前公司相关的市场趋势特征数据。

对按编码器长度分类的数据分别进行测试,得到在不同时间长度、不同数据复杂度下TFT模型是否更具有优势的结果。

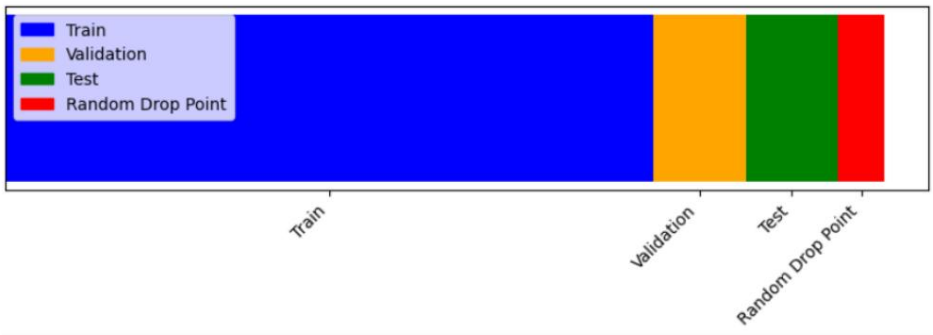

图5 数据集设计

3.4.2 数据来源

本研究中三家公司的股票数据来自雅虎财经网站，
具体来说是 NVIDIA Corporation (NVDA)、Apple Corporation (AAPL) 和 Tesla Corporation (TSLA) 的历史股价数据页面。数据于 2024 年 8 月 26 日从以下链接访问和下载（雅虎财经,2024 年）：

NVIDIA:https://uk.finance.yahoo.com/quote/NVDA/history

其他季度公司财务数据,比如Revenue、Net Income、ESP、Operating Income、EBITDA、Current Ratio等,均来自matrotrends网站,具体页面来自股票详情界面,大家可以点击不同的公司进行选择查看:

英伟达:https://www.macrotrends.net/stocks/charts/NVDA/nvidia/financial-statements

此外,我们还从Google获取了一些社交热门关键词的数据信息
趋势。

“人工智能”:https://trends.google.com/trends/explore?date=today%205-y&geo=GB&q=AI&hl=zh-CN

本研究中,数据集被分为[data]-5y、[data]_data。[data]代表不同的公司如NVIDIA:NVDA;APPLE:AAPL;TESLA:TSLA。-5y后缀为纯股票数据,_data后缀为添加了金融属性特征的数据,NVDA-AI代表添加了社交热门“AI”关键词特征的数据。

3.4.3 工具和环境

本研究采用Python编程语言进行开发,整个开发过程在Anaconda环境下进行,以Jupyter Lab作为集成开发环境(IDE)。为实现高效的模型训练,采用了不同的深度学习框架:Temporal Fusion Transformer(TFT)模型使用PyTorch框架构建并训练,用于对比的LSTM和GRU模型则基于TensorFlow实现。所有模型的训练过程均在支持GPU加速的环境下进行,以提高训练效率,加速模型优化。

3.4.4 模型设计细节

为了获得更好的模型性能,TFT模型的超参数为
调试完毕。使用Pytorch_forecasting包的optimize_hyperparameters()方法调试参数后,似乎无法获得最佳性能,因此进行了手动调试。参数如表1所示。另外为了获得更好的性能,在TFT,LSTM和GRU中设置了早期停止机制(Early Stopping),以防止模型在训练过程中过拟合(Ferro et al.,2023)。

具体来说,如果在连续几轮训练中验证集的性能不再提高,训练将自动停止。这种机制有助于防止模型对训练数据过拟合,从而提高模型的泛化能力

测试数据。在模型设计中,early stopping 的patient参数设置为12。当12个epoch后loss不再减小时,学习停止。当模型停止训练时,Trainer的回调会回溯并记录loss值最好的模型作为最终

模型。

值得注意的是,Gradient_clip_val也是一个非常重要的参数,它将
将在下一节中讨论。

表 1 TFT 的超参数。

主要超参数	价值
辍学	0.1
喷头数量（TFT用)1	2（更多功能）
批次大小	128
隐藏大小	128
规格化	最小最大缩放器
分位数（用于 TFT）	[0.1,0.2, 0.5,0.8,0.9]
优化器	亚当
Trainer_Gradient_clip_val	2.0
最大周期数	50

此外,为了进一步优化训练过程,学习率调度器
在模型设计中引入学习率降低策略,当监测到的性能指标（比如验证集的损失)在指定的训练轮次中不再提升时,学习率会自动按一定比例降低。通过逐步降低学习率,模型可以更好地探索损失函数的局部最优解,从而提高训练的稳定性和最终模型的性能（Sutskever et al., 2013）。在模型设计中,学习率降低设置的参数如下表2：

表 2 学习率调度器超参数

参数	价值
初始学习率	0.01
减小因子	0.8
耐心	3
最小学习率	1e-6

3.4.5 评估细节

模型比较 ：

在模型间性能比较中,相同参数（若可设置)与
使用相同的数据格式进行训练。在LSTM和GRU中,使用相同的提前停止机制和自动学习率下降功能。为了保证一致性,使用相同的损失函数和一致的参数进行训练,并使用测试集的结果进行模型评估。为了在更多维度上获取更多信息,使用平均绝对百分比误差（MAPE)和均方根误差（RMSE)来评估拟合值的准确性。使用Spearman相关系数来评估拟合趋势的准确性,因为对于金融预测来说,预测的趋势结果非常重要,甚至比预测值的准确性更重要。

在TFT中,为了便于模型识别,可以将数据划分为不同的属性,因此提取月、日数据,并创建time_idx作为历史数据下标。TFT可以对不同类别的数据分别进行训练,但本研究中每次训练都是分开的,因此group_id是一致的。TFT的详细数据类型如表3所示:

表 3 TFT 数据结构

参数	特征
time_idx	‘时间 IDX’
目标	关闭
time_varying_known_categoricals	[月 , 日]
time_varying_known_reals	[time_idx , 年份 , group_ids]
time_varying_unknown_reals	[调整收盘价 , 成交量 , 最低价 , 开盘价 , 最高价]

在LSTM和GRU中,还会提取日期数据,从日期数据中提取年、月、日与其他时间序列数据一起作为特征,这两个模型的数据结构由表4组成:

表 4 LSTM 和 GRU 数据结构

参数	特征
特征	[开盘价 , 最高价 , 最低价 , 调整收盘价 , 成交量 , 年 , 月 , 日]
目标	关闭

探索组合损失函数 :

利用TFT模型研究组合损失函数的可行性,并比较
不同损失函数的结果对比,以分位数损失与MSE作为损失函数单独训练,以参与融合MDA、Spearman相关系数等其他趋势评估指标的损失函数作为组合损失函数。

丙二醛 :

MDA 预测值的计算将分为几个单独的步骤。结果
MDA 的值为 0 或 1。1 代表正确的趋势,0 代表错误的预测趋势。
通常MDA计算的最终输出值是整体正确预测的比例。但本研究在每一步都进行了计算。通过转换预测评估结果,符合趋势时值为1,不符合趋势时值为3,然后将每一步的值乘以MSE。这样,当趋势一致时,loss值就是MSE本身。当趋势不一致时,loss值就是3倍的MSE,以达到融合MDA的评估效果。值得

提到,如果“符合趋势”的 MDA 值变成负数,再乘以 MSE,模型就会误以为趋势是正确的,但

MSE 值显著的预测点将是正确的,如果“符合趋势”的 MDA 值乘以 MSE 后变为 0,则预测值的准确率将被忽略。因此需要将评估结果转化为:“正确”趋势的 MDA 为 1,“错误”趋势的 MDA 为 3,即在保留预测值准确率的同时,根据不同的趋势结果分配惩罚权重。

Spearman 相关系数 :

由于斯皮尔曼相关系数(Spearman s ρ)是基于整体数据计算的,无法单独将其剔除,本研究将Spearman s ρ作为单独的损失函数,在整体损失函数的计算过程中,通过权重控制该结果在整体损失函数中的重要性。值得一提的是,Spearman s ρ在多步预测的评估结果上更有优势,因为在数据较多的情况下,其对于趋势的评估结果更加稳定,而次数较少的短期预测可能会导致评估结果波动过大。本研究将Spearman s ρ与MSE结合起来形成损失函数,目的在于前者控制趋势拟合,后者控制精度拟合。

自动损失函数权重GradNorm方法:

GradNorm旨在训练过程中平衡各个损失函数的权重,对于学习速度过快(梯度较小)的任务,会降低其梯度;对于学习速度较慢(梯度较大)的任务,会提高其梯度,从而平衡各个任务的训练过程,避免某个损失函数过度主导整个训练任务(Chen et al.,2017)。本研究采用MSE、分位数损失和Spearman s ρ作为损失函数组,在对损失函数组进行实时监控实验后,将三者缩放为[mse_loss/200,quantile_losses/3,50*spearman_corr_loss]。使得初始的三个损失函数值处于同一数量级。

表5给出了损失函数的组成和分类。

表 5 损失函数组成

损失函数	内容
单一损失函数	分位数损失
单一损失函数	微分方程
组合损失函数	MDA_LOSS * 微型和中型企业
组合损失函数	SPEARMAN_CORR_LOSS + MSE
组合损失函数的自动加权	[MSE_LOSS/200,QUANTILE_LOSSES/3, 50*SPEARMAN_CORR_LOSS]

4.结果

为了避免极端情况影响结果,任何与大多数测试结果有显著偏差的个别结果都会被排除。除非模型本身不稳定,产生太多极端值。同样,在选择随机数据个数时,如果丢弃后产生的数据使得所有模型都无法有效预测,则不会将其纳入测试,因为这会导致对模型整体性能的评估不正确。

模型预测结果比较:

使用MSE作为损失函数。使用50、300和600编码器长度对模型进行10轮训练,固定随机种子为42。使用历史股票序列数据（没有特定的金融特征）。由于Spearman的p_value结果区间很低,方差很难具有参考性质,因此将最佳值作为附带信息。

值得注意的是,gradient_clip_val 的值会显著影响拟合
模型的效果。将gradient_clip_val设置为2,有更好的学习能力。

使用 NVDA-5y 数据集删除 0 个尾部

表6 不同数据复杂度下的模型结果

		检验的平均值（方差）				Spearman 的 p_value（最佳）
模型编码器	长度	甲基丙烯酸甲酯	均方根误差	MAE Spearman 的 rho (ρ)		
薄膜晶体管	50	6.50 (2.16) 8.82	(4.09) 8.13 (3.95) 5.20 (1.08) 4.31		0.53(0.02) 0.45(0.12)	
长短期记忆（LSTM）		(0.62) 6.16 (2.02) 5.30 (0.70) 4.45 (0.50) 5.89 (0.90)			0.20(0.17)0.64(0.37)	
格鲁乌		7.21 (0.41) 9.96 (0.62) 9 .09 (0.77) 4.87 (0.69) 7.06			-0.04(0.15)0.59(0.39)	
薄膜晶体管	300	(2.02) 5.89 (1.27) 4.77 (0.80) 6.39 (1.02) 5.71 (1.17)			0.70(0.03) 0.38(0.037)	
长短期记忆（LSTM）		7.63 (1.16) 10.38 (1.95) 9.67 (2.18) 5.28 (1.81) 7.48			0.04(0.21)0.48(0.037)	
格鲁乌		(3.77) 6.26 (2.16) 4.63 (1.39) 6.57 (3.61) 5.51 (1.62)			-0.09(0.29)0.52(0.10)	
薄膜晶体管	600				0.78（0.04)0.09（1.4e-24）	
长短期记忆（LSTM）					0.10（0.13)0.24（1.4e-24）	
格鲁乌					0.17(0.21)0.624(0.03)	

在NVDA-5y数据集的测试结果中,三个模型的MAPE、RMSE和MAE整体上表现出很强的相关性。TFT在任何Encoder长度下对模型的预测精度影响都很差（例如,与其他模型相比,MAPE很高,如图6所示）。但它在Spearman s rho（ρ)值上表现出断崖式的领先,分别为0.53、0.7和0.78。随着Encoder长度的增加,TFT和LSTM的预测精度变低（MAPE增加）,但所有模型的Spearman s rho（ρ)值均变好（图6）。另外,TFT和LSTM的Spearman s p_value都有不同程度的下降,但GRU却没有。值得注意的是,在数据复杂度较低的情况下,GRU的Spearman s rho（ρ)在0以下。此外,TFT的Spearman rho（ρ)值的平均方差明显低于其他两个模型。

总体来看,当encoder长度较长时,整体评价指标变好,例如TFT和LSTM的Spearman p_value在Encoder长度为600时均较低（趋势表现较好）,但预测值指标MAPE、RMSE、MAE有所提升（预测准确率表现下降）。

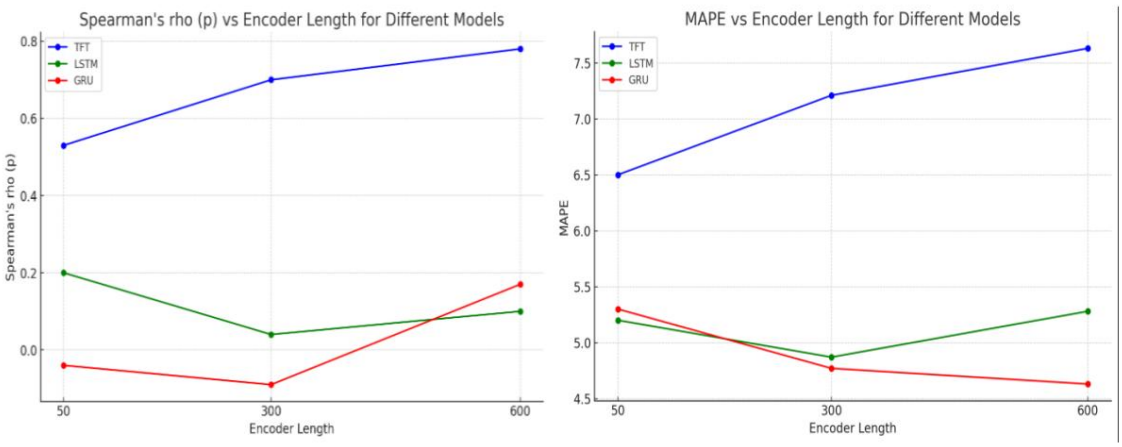

图 6 预测结果对比

图7是对RMSE、MAPE、MAE进行相互处理后的结果,图中所有评价指标随着数值的增大都代表模型性能的提高。整体评价结果下,TFT在Spearman’ s rho(ρ)上优势明显,其他指标略低于另外两个模型,而LSTM的Spearman’ s rho(ρ)在0.1左右,而GRU趋近于0。

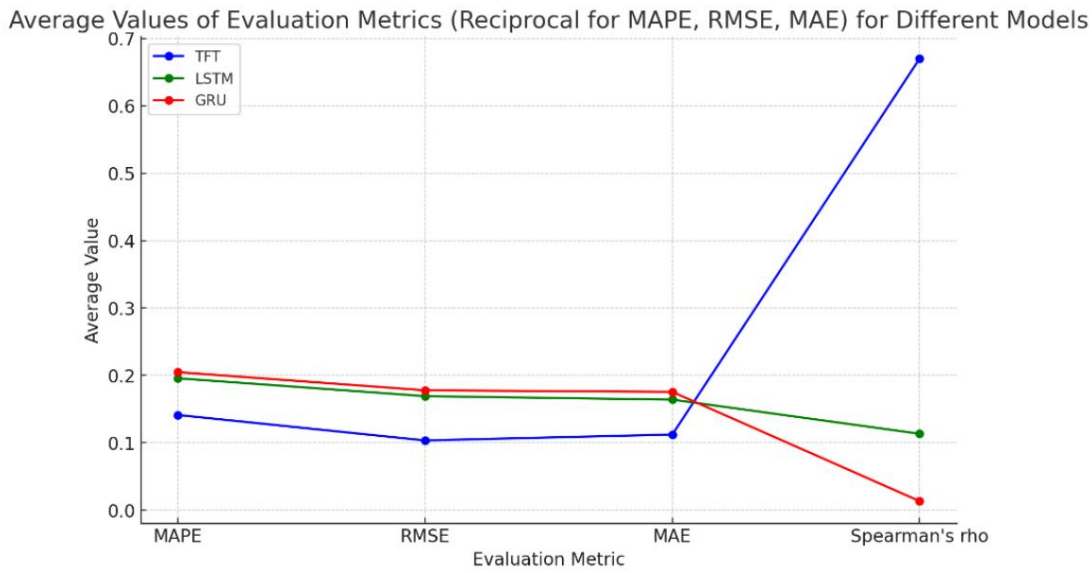

图 7 平均值

获得最佳编码器长度后,随机切换数据集以测试模型的稳定性:切换到另外两家公司的数据,并随机删除随机数量的尾部数据。为了使实验可重复,将记录随机值

使用 AAPL-5y 数据集删除 57 个反面 ：

表7 比较结果

		检验的平均值（方差）				Spearman 的 p_value（最佳）
模型编码器	长度	甲基丙烯酸甲酯	均方根误差	MAE Spearman 的 rho (ρ)		
薄膜晶体管	600	1.41(0.17) 2.92(0.67)		2.71(0.65)	0.81(0.01) 0.11(0.03)	
长短期记忆（LSTM）		1.42(2.04) 3.13(8.14) 0.95(0.13)		2.61 (6.23)	-0.22(0.14) 0.57(0.10)	
格鲁乌		2.24(0.62)		1.80 (0.45)	0.15（0.18)0.59 (0.10)	

TABLE 7给出了不同数据集下的评估结果，当Encoder长度为600时，使用AAPL-5y数据集，去掉57个尾部数据后，TFT表现依然显著，在MAPE、RMSE、MAE上均没有表现不佳的迹象，另外Spearman的rho(ρ)达到了0.81，Spearman的p_value只有0.11。而LSTM在这个数据集上表现不佳，MAPE、RMSE、MAE值均弱于GRU，Spearman的rho(ρ)为-0.22，而GRU为0.15。

使用 TSLA -5y 数据集删除 176 个反面 ：

表8 比较结果

		检验的平均值（方差）				Spearman 的 p_value（最佳）
模型编码器	长度	甲基丙烯酸甲酯	均方根误差	平均动脉介入	斯皮尔曼的 rho (ρ)	
薄膜晶体管	300	6.60 (2.68) 13.88 (8.43) 13.38 (10.14) 0.73 (0.05) 0.19 (0.03)				
长短期记忆（LSTM）		3.62 (3.00) 8.52 (13.27) 7.75 (12.85) 0.07 (0.20) 0.53 (0.10)				
格鲁乌		2.61 (1.49) 6.60 (9.32) 5.61 (6.67) 0.13 (0.36) 0.33 (0.03)				

TABLE 8 是使用TSLA-5y数据集，在encoder长度为300的情况下，去掉176个尾部数据的结果，这里出现的结果和图7类似，TFT模型在MAPE、RMSE、MAE结果上表现较差，比其他两个模型高出近50%，不过Spearman的rho（ρ)表现不错，达到了0.73，方差极低，为0.05，此时Spearman的p_value为0.19，最好结果是0.03。

使用附加特征进行预测的比较：

由于 TFT 允许非常灵活的特征分类，因此使用 TFT 可以探索更多相关功能可能会提高性能。

使用 NVDA -5y/NVDA_data/NVDA-AI 数据集删除 0 个尾部：

表9 测试结果

数据	检验的平均值（方差）				Spearman 的 p_value（最佳）
	甲基丙烯酸甲酯	均方根误差	平均动脉介入	斯皮尔曼的 rho (ρ)	
纯股票	7.21（0.41）	9.96（0.62）	9.09（0.77)0.70	(0.03)	0.38(0.037)

凭借金融特征	6.85（1.02）	9.76（1.26）	8.61（1.84）	0.64（0.006）	0.25（0.1）
和 社会的趋势	7.78（0.47）	11.25 (0.57)	9.91 (0.87)	-0.37 (0.04)	0.55（0.05）

为了探究TFT是否更适合附加复杂特征的数据集并具有更好的性能，我们使用了上面提到的不同数据集进行训练和测试。添加金融特征之后，模型拟合效果似乎并没有发生明显变化。MAPE、RMSE和MAE几乎没有波动，Spearman s rho（ρ）只下降了0.06（从0.7下降到0.64），但实际上整个结果变得非常平滑，如图8所示（与图10相当）。值得注意的是，当使用带有社交趋势的数据时，所有评估标准都变得更糟。

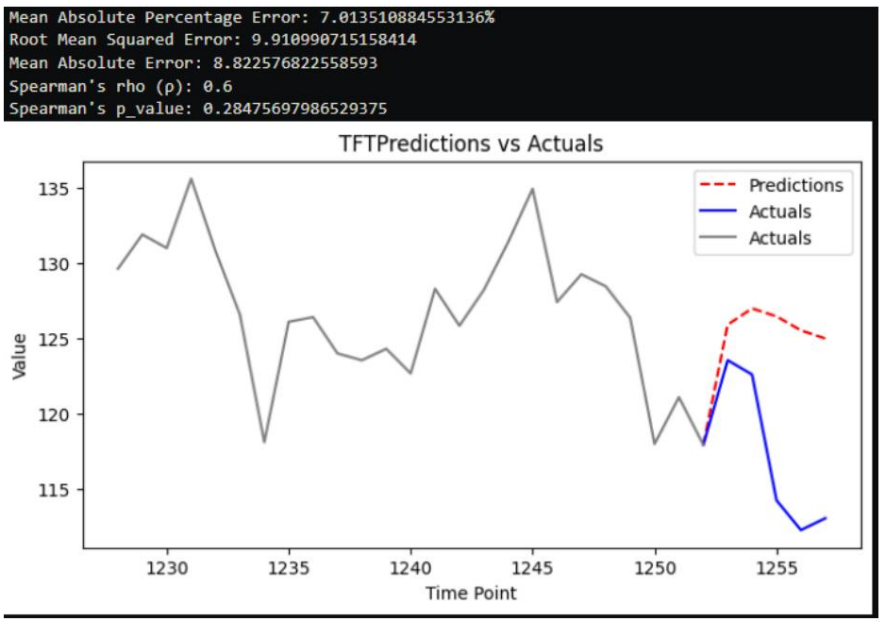

图 8含财务特征的数据结果

不同损失函数之间的结果差异。

本研究中，在探索如何提升模型性能时，损失函数成为非常重要的原因之一。这部分研究在统一的标准下测试了不同的损失函数，并探索了组合损失函数的可能性。在Quantiles = [0.1, 0.2, 0.5, 0.8, 0.9]的Quantile损失下，模型并没有表现出更好的性能，各项指标下的数值都比较差。MSE在这个数据集上表现最好，RMSE、MAE和Spearman s rho(ρ)是所有损失函数中最好的值，所有的评价值的方差都很低，但Spearman的p_value并不是最好的，为0.38。MDA-MSE损失函数的表现比较普通，没有什么亮点。GradNorm损失函数在RMSE、MAE和Spearman s rho(ρ)的结果中方差最大，没有体现出什么性能优势。

值得一提的是，SPEARMAN 损失函数在 MAPE 中结果最差，

RMSE、MAE 与其他损失函数相比,方差非常大,但 Spearman 的 rho(ρ) 表现良好,为 0.66,方差仅为 0.01,并且其 Spearman p_value 是所有损失函数中最低的。

使用 NVDA-5y 数据集删除 0 个尾部:

表10 测试结果

损失函数	检验的平均值（方差）				Spearman 的 p_value（最佳）
	甲基丙烯酸甲酯	均方根误差	平均动脉介入	斯皮尔曼的 rho (ρ)	
分位数	8.99 (1.29) 12.44 (2.48) 11.59 (2.63) 0.25 (0.13) 7.21 (0.41) 9.96 (0.62)				0.58（0.09）
微分方程	9.09 (0.77) 0.70 (0.03) 5.39 (4.39) 12.25 (19.13) 10.97 (16) .42) 0.40				0.38(0.28)
MDA含量	(0.13) 9.46 (6.14) 19.41 (20.85) 18.51 (18.97) 0.66 (0.01) 5.21 (6.06)				0.43（0.1）
斯皮尔曼	11.37 (21.99) 10.67 (23.34) 0.42 (0.13)				0.24(0.03)
梯度标准差					0.35(0.28)

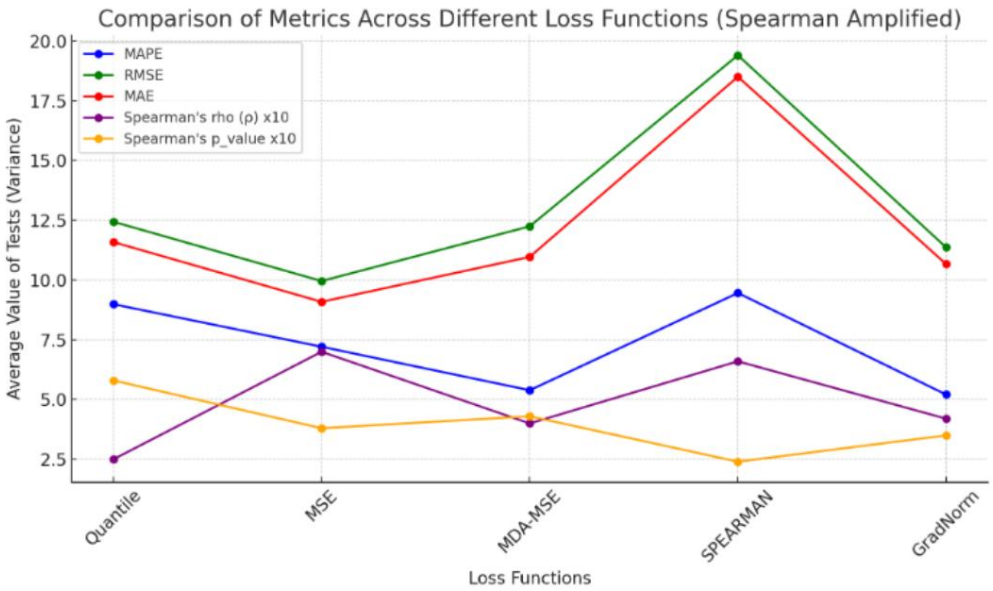

图 9 表 10 的可视化

5.讨论

模型预测结果比较:

在研究的第一部分主要研究了不同模型在同一数据集上的表现差异,可以发现TFT在使用NVDA-5y数据集drop 0 tails时表现出了明显的预测优势,虽然在MAPE、RMSE和MAE值上表现不佳,但极高的Spearman’s rho(ρ)和低的Spearman’s p_value代表TFT对趋势的拟合极其准确,数据可信度良好,这在短期多步预测中具有良好的应用前景,在金融领域尤其重要

预测。如图9所示，尽管LSTM在MAPE、RMSE和MAE的结果上具有优势，但在中期拟合中存在较大的异常，导致Spearman s rho（ρ）值非常低，仅为0.19。虽然TFT在这三个评估值上表现不佳，但拟合曲线在预测趋势方面相当准确，这证实了相关性指标在某些时候更具有参考价值（MacFarland&Yates，2020）。

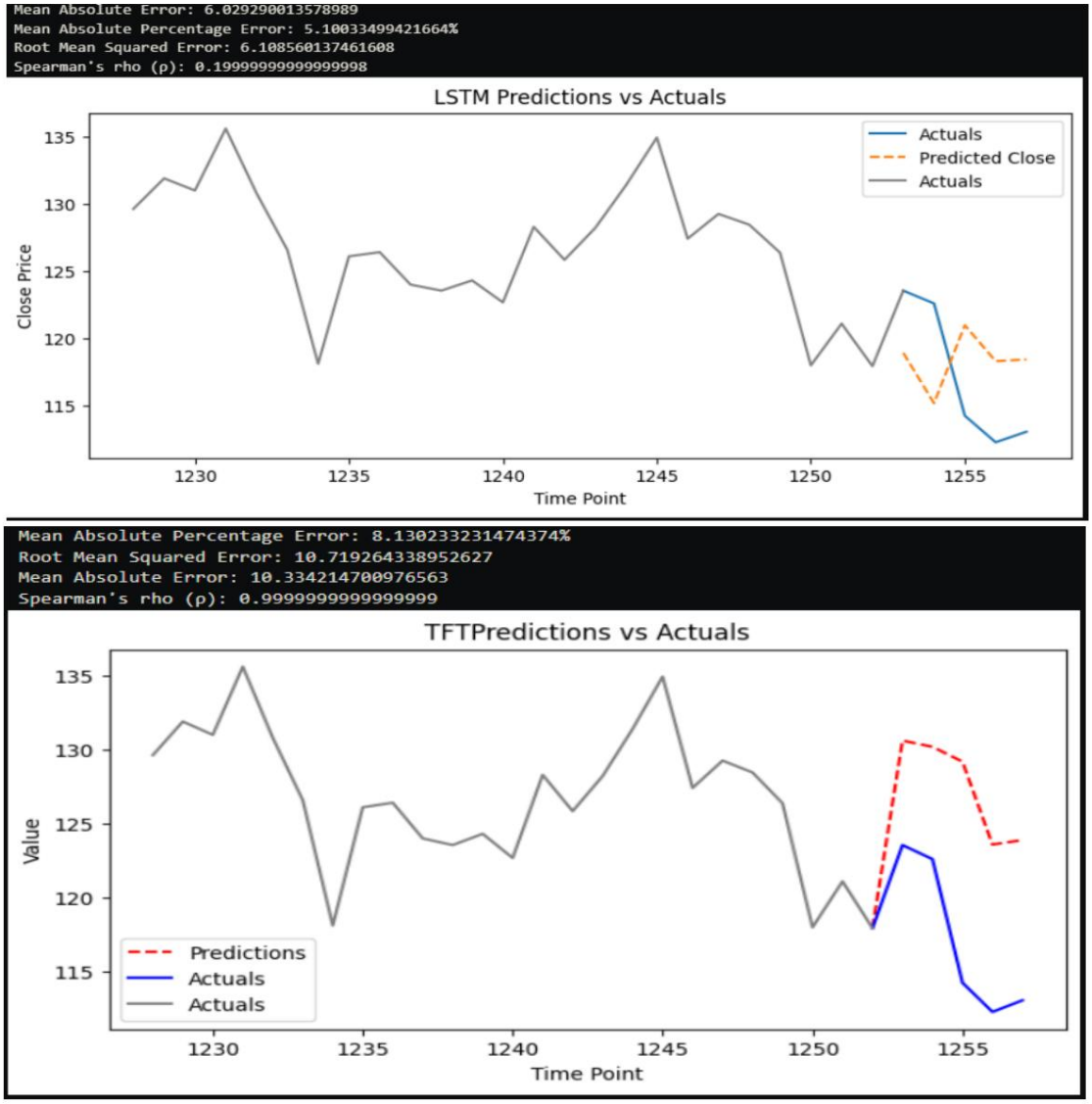

图10 同一数据下两种模型的预测结果示例

表6的结果中，随着编码器长度的增加，TFT和LSTM都出现了预测值准确率下降但预测趋势准确率上升的情况，这意味着更多的历史数据可以让模型更好地学习到未来可能的趋势。而随着GRU的预测趋势准确率的上升，预测值准确率也下降，但变化很小，这可能是由于模型本身的局限性。

当编码器长度达到 600 时，TFT 的性能提升变小但仍有提升，这可能意味着更大的数据会产生更好的结果，但提升幅度有限。而且在编码器长度较长的情况下，所有模型都不会产生错误的趋势拟合（GRU 在编码器长度为 50 时趋势拟合结果小于 0

和300)。

值得注意的是,TFT在不同情况下的Spearman s rho（ρ)方差非常低,代表预测的稳定性。如前所述,当模型拟合结果出现明显误差时,可以舍弃异常结果来提高性能,这和随机森林的投票原理类似,所以稳定的模型是提高性能的关键。如果模型输出的结果大部分都是高度随机的,没有明显的趋势,研究人员就无法分辨哪些结果是异常的。因此,TFT预测的稳定性进一步提高了TFT的准确性。

在使用其他随机数据重复测试之后,三个模型的结果保持稳定。使用 TSLA -5y 数据集 drop 176 tails 的结果显示（TABLE 8）,TFT 仍然具有很好的趋势拟合优势,其 Spearman’s rho（ρ)达到 0.73,方差为 0.05,代表了 TFT 的稳定性。这里 Spearman’s p_value 也达到了 0.19,优于在编码器长度为 300 的 NVDA 数据集中测试的值 0.38。这说明该模型在不同数据集中都有实现良好趋势拟合的能力,但不同的数据可能造成评估可信度的波动,这可能是数据本身的异常所致。

同样的,TABLE 7是使用AAPL-5y数据集drop 57 tails,encoder长度为600的结果。在这次测试中,TFT的趋势拟合效果很惊人,达到了0.81,方差仅为0.01,在预测准确率方面也毫不逊色,MAPE、RMSE、MAE的值与另外两个模型接近。而LSTM在这个数据集上的表现很差,趋势拟合甚至出现了反方向的情况,Spearman’s rho(ρ)为-0.22,虽然MAPE等预测值的准确率比较正常,但是趋势拟合误差说明了LSTM的不稳定性。GRU的Spearman’s rho(ρ)为0.15,证明了经典模型依然能够获得一定的拟合效果,数据本身也是

没有反常。

使用附加特征进行预测的比较:

TFT拥有多头注意力机制,理论上引入季度、年度信息等相关静态变量会提升模型的表现。TABLE 9展示了添加不同特征数据时的训练结果。在加入金融特征数据的测试结果中,预测精度指标如MAPE并没有发生明显变化,但是Spearman’s rho（ρ)似乎变得更加稳定了,因为与纯股票数据相比,方差从0.03降低到了0.006,p_value从0.38提升到了0.25,这意味着在金融特征的支持下,模型对趋势的预测能力似乎变得更加稳定了。图8展示了加入金融特征数据测试的结果相对有代表性的截图,数据变得比纯股票更加平滑,这意味着使用季度信息作为静态变量在天数预测上的效果可能没那么好。结果变得稳定了,但是一些可以提升表现的细节会丢失,完整的趋势拟合曲线也会丢失。

，和斯皮尔曼的

使用带有社会趋势的数据进行预测的结果更差。Spearman 的 rho (ρ) 为 -0.37,拟合了相反方向的趋势效应。这可能是因为所选数据与整体数据没有微相关性。

它们具有较强的宏观相关性,但不适合本预测模型。

不同损失函数的结果差异:

损失函数的检验代表了不同目标下的预测趋势。
在这一部分,Quantile loss并没有表现出有效的性能,猜测可能是数据集选择的问题,有时候不同的损失函数对不同类型的数据有不同的效果,Quantile loss更适合那些容易训练但整体结果有偏差的数据模型。

MSE效果最好,由于MSE会严厉惩罚异常值,所以它可能更适合规则的数据集,惩罚那些异常的波动。

MDA-MSE 表现平平,没有展现出预期的效果。这可能是因为惩罚的效果不够明显。例如,模型可以识别出趋势错误时损失值很大,但模型无法区分惩罚是由 MSE 还是 MDA 引起的。这可能会导致模型训练过程中的不稳定。或许需要更多的训练层和耐心才能获得更好的性能。

值得一提的是,SPEARMAN的结果与TFT的结果类似,见表6。在损失函数中加入Spearman’s rho(ρ)后,模型的MAPE、RMSE、MAE都产生了较大的误差,但Spearman’s rho(ρ)表现非常出色,为0.66。不仅方差只有0.01,而且Spearman’s p_value也只有0.24,是所有损失函数中最低的值。这证明了模型在SPEARMAN损失函数下产生的结果在趋势表达上是最稳定和可信的。在不同的权重比例下,或许模型可以在趋势和拟合细节之间取得更好的平衡,以应对不同的预测目标。这也得出一个结论,模型对趋势的拟合程度越高,预测值的准确性就越难控制。

使用自动化的方法来控制组合损失函数的权重分布是一种理想的技术,但 GradNorm 似乎表现并不好。为了探究原因,在训练 GradNorm 损失函数的过程中,时刻监控权重分布情况,如图 11 所示。在分配权重的过程中,GradNorm 不断向那些变化率很低的权重倾斜权重。这就导致如果某个损失函数本身的变化率不高或很难改变,GradNorm 仍然会疯狂地增加其权重。例如,在损失函数中加入 Spearman s rho(ρ) 的意义在于控制一些辅助的趋势拟合效果,GradNorm 会不断向其倾斜权重,直到无限接近 1。但模型单靠这个辅助损失函数并不能真正训练出模型,训练效果会越来越差。因此,GradNorm 的问题在于它不能完全自动区分权重比例的重要性,而损失函数越来越极端的权重分布会导致模型进一步欠拟合。也许控制权重的下限并降低变化速度,以及不要将Spearman的rho(ρ)等辅助评价值单独放在GradNorm损失函数列表中,可能会取得更好的效果。

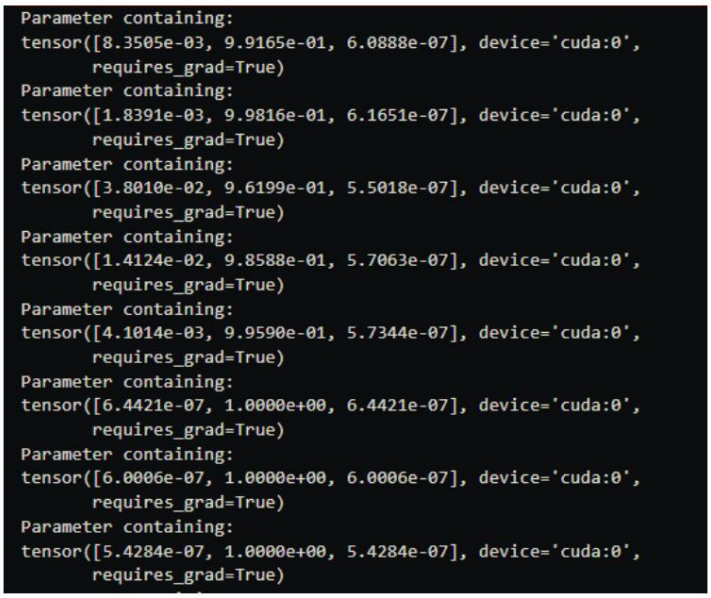

图11 训练过程中GradNorm权重分布

TFT的可解释性优势：

TFT 的一大特点就是能够将一些具有解释原理的数据可视化。图 12 中的灰线展示了在预测 TEST 集时，数据对预测的贡献。通过这种注意力机制，TFT 关注那些在预测中的人。

对特定时刻的预测结果比较关键的数据点。可以看出最近50个单位的数据对模型的影响非常大，但是根据以往的实验结果，长期的数据同样重要。

类似的图13显示了哪些Encoder数据在TEST集的预测中是最重要的。可以看到时间序列下标，Open数据和Month数据都很重要，具有很好的参考价值。

图14表明了哪些Decoder数据在TEST集的预测中是最重要的，代表了未来数据的参与。Day的重要性显著，说明数据具有一定的周期性。

图15是训练过程中数据拟合的细节，大家可以通过这个方法来看一下哪些特征影响了你的拟合效果。

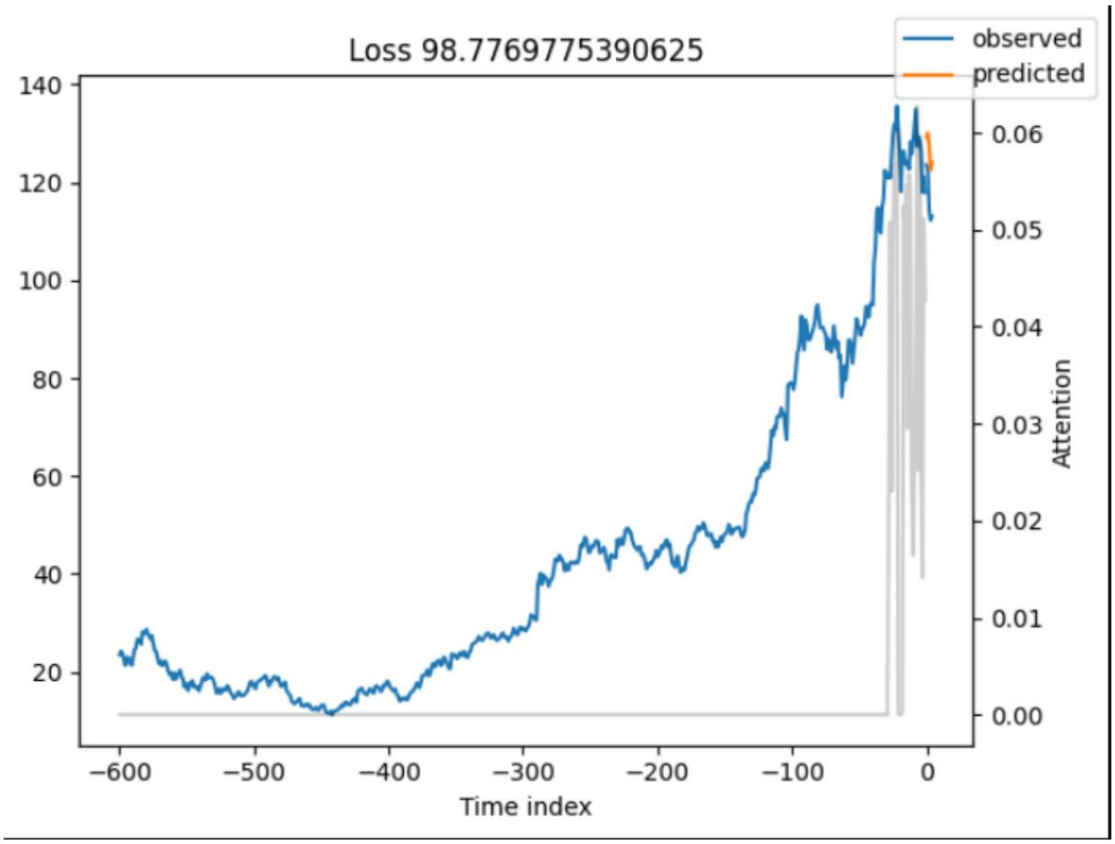

图 12 注意力分布示例

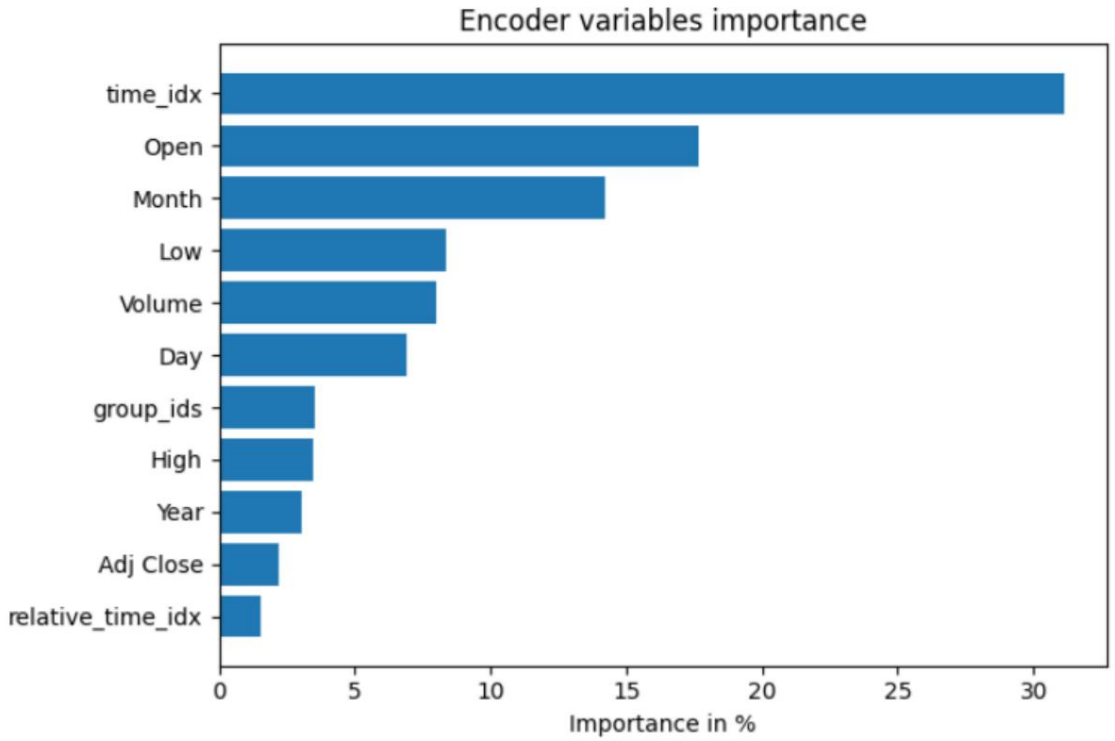

图 13 编码器变量重要性

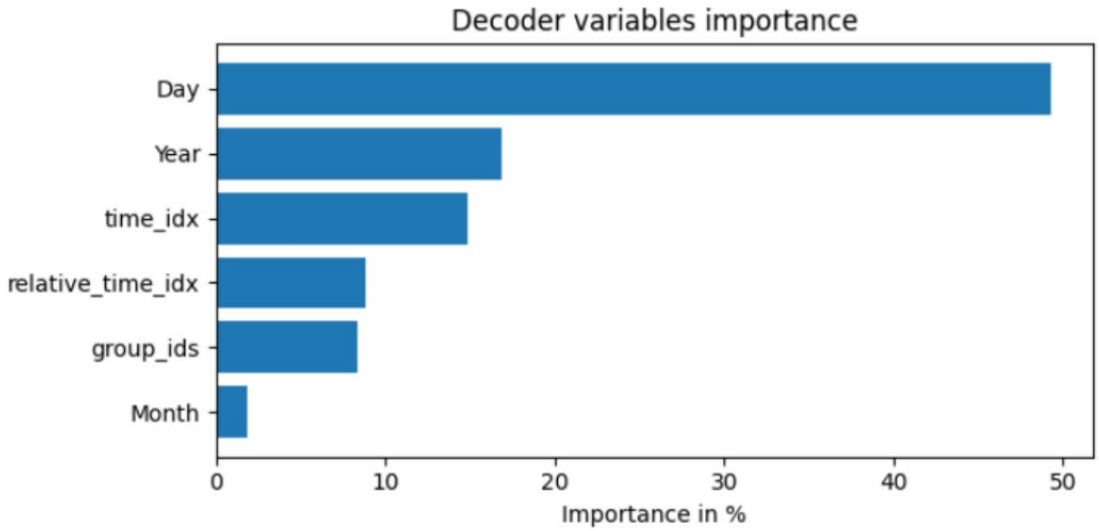

图 14 解码器变量重要性

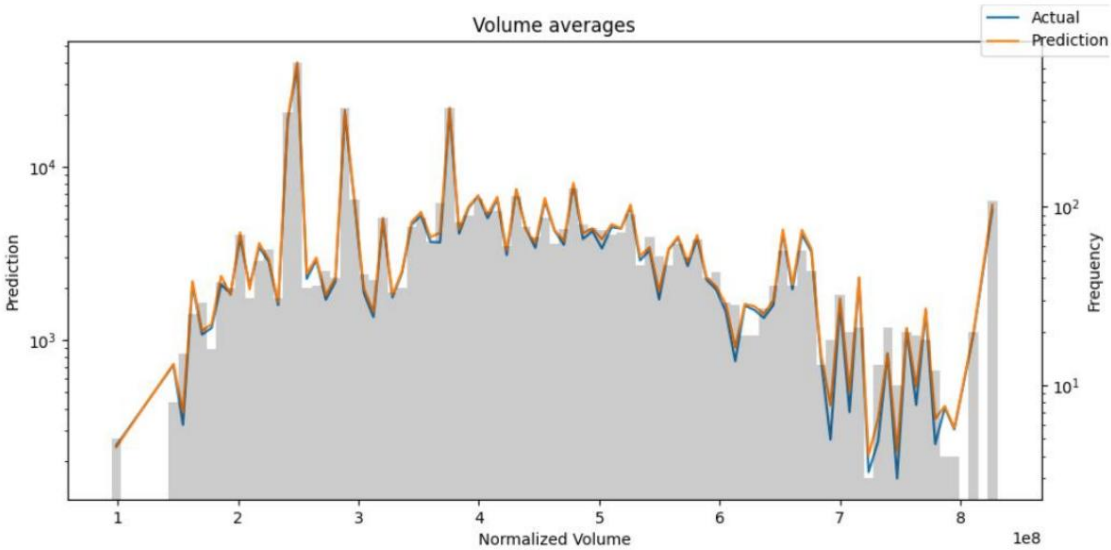

图15 训练过程中的特征拟合

6. 结论

本研究指出了时间融合变换器（TFT)在预测金融股票指数,尤其是在预测趋势方面的有效性。本研究利用近五年的每日股票数据进行模型训练和对未来的短期多步预测。利用TFT设计了金融股票数据预测模型,并使用不同的数据集和参数评估TFT的性能,并与LSTM和GRU预测模型进行了比较。结果表明,TFT预测模型可以有效地预测金融股票数据,尤其是在趋势表现上。此外,还探索了一些组合损失函数的有效性,并得出结论:季度特征不能作为有效的特征来提升短期多步预测模型的性能。而且,在本研究环境下使用GradNorm无法取得更好的性能。

在研究过程中,探索的数据集数量不够,如果数据集更多一些,比如小公司的数据集,或者数据本身没有长期趋势,可能更能体现实验的可靠性。实验次数也不足。TFT 比 LSTM 和 GRU 更稳定,所以几十次训练结果就可以算作实验结果。对于 LSTM 和 GRU,需要进行更多次实验才能达到稳定的结果。但模型的稳定性同样重要。

应使用更多的评价标准来验证模型的稳定性,这对实验结果的可靠性有参考意义。添加合理的特征并没有达到预期的效果,应该在更多不同的数据集上进行测试,以增加实验结论的可靠性。使用GradNorm方法并没有达到预期的性能提升,这可能是由于损失函数组的选择,如果将MSE、RMSE、MAPE等性质相近的损失值组合起来,可能会达到合理动态分配的效果。

在未来的研究中,利用TFT模型预测趋势数据目标是很有意义的。

短期多步预测加剧了预测的不稳定性,如果将预测跨度改为几周或者几个月,性能可能会进一步提升,此时使用一些相关的季度数据作为特征也可能变得有效。同样,对 Combined 损失函数的研究也可以更加深入,如何将 Spearman 的 rho(ρ) 更好地融入到经典损失函数中或许是一个非常有前景的方向。最后,Combined 损失函数的自动权重分配也值得深入探索,或许这里可以加入一些简化的机器学习方法,让自动权重分配更加有效。

参考

1. MacFarland, TW 和 Yates, JM (2020)。相关性、关联性、回归性、似然性和预测。在 Springer 电子书中（第 427-584 页）。https: //doi.org/10.1007/978-3-030-62404-0_7

2. Chen, Z.、Badrinarayanan, V.、Lee, C. 和 Rabinovich, A. (2017 年 11 月 7 日)。GraDNorM：深度多任务网络中自适应损失平衡的梯度归一化。arXiv.org。https: //arxiv.org/abs/1711.02257

3. Sutskever, I.、Martens, J.、Dahl, G. 和 Hinton, G. (2013)。论初始化和动量在深度学习中的重要性。第 30 届国际机器学习会议论文集，机器学习研究论文集 28(3):1139-1147 可从https://proceedings.mlr.press/v28/sutskever13.html 获取。

4. Ferro, MV、Mosquera, YD、Pena, FJR 和 Bilbao, VMD (2023)。通过关联神经网络中的在线指标实现早期停止。神经网络，159，109–124。https: //doi.org/10.1016/j.neunet.2022.11.035

5. 雅虎财经。（2024 年）。NVIDIA Corporation (NVDA)、Apple Inc. 的历史数据。(AAPL) 和特斯拉公司 (TSLA)。2024 年 8 月 26 日检索自https://uk.finance.yahoo.com/quote/NVDA/history/

6. Wen, R.、Torkkola, K.、Narayanaswamy, B. 和 Madeka, D. (2017 年 11 月 29 日)。多视野分位数递归预测器。arXiv.org。https ://arxiv.org/abs/1711.11053\

7. Jozefowicz, R.、Zaremba, W. 和 Sutskever, I. (2015 年 6 月 1 日)。PMLR 的实证探索。复发性 网络 架构。https://proceedings.mlr.press/v37/jozefowicz15.html

8. Karpathy, A.、Johnson, J. 和 Fei-Fei, L. (2015b 年 6 月 5 日)。可视化和理解循环网络。arXiv.org。https: //arxiv.org/abs/1506.02078

9. Fischer, T. 和 Krauss, C. (2018)。利用长期短期记忆网络进行深度学习以预测金融市场。《欧洲运筹学杂志》，270(2)，654–669.https ://doi.org/10.1016/j.ejor.2017.11.054

10. Campbell, JY、Lo, AW 和 MacKinlay, AC (1997)。金融计量经济学（第 34 页）。市场 21）。 普林斯顿 大学出版社。https: //press.princeton.edu/books/hardcover/9780691043012/the-econometrics-of-

金融市场

11. Jordan, MI 和 Mitchell, TM (2015)。机器学习:趋势、观点和
前景。 《科学》， 349(6245),255–260。https ://doi.org/10.1126/science.aaa8415

12. Heaton, JB、Polson, NG 和 Witte, JH (2016)。深度学习在金融领域的应用:深度投资组合。《商业和工业
中的应用随机模型》， 33(1),3–12。https ://doi.org/10.1002/asmb.2209

13. Ullah, S. (2022)。COVID-19 疫情对金融市场的影响:全球视角。《知识经济杂志》,14(2),982–1003。https: //
doi.org/10.1007/s13132-022-00970-7

14. Goldstein, I.、Koijen, RSJ 和 Mueller, HM (2021)。COVID-19 及其对金融市场和实体经济的影响。《金融研究评
论》,34(11),5135–5148。https: //doi.org/10.1093/rfs/hhab085

15.金融稳定报告 - 2023 年 5 月。 （nd）。美联储理事会https://www.federalreserve.gov/publications/2023-
系统。报告 may-financial-stability-
目的和框架.htm

16. Olorunnimbe, K. 和 Viktor, H. (2022)。股票市场的深度学习 实践、回测和应用的系统调查。人工智能评论， 56(3)，
2057–2109。https ://doi.org/10.1007/s10462-022-10226-0

17. Lim, B.、Arik, SO、Loeff, N. 和 Pfister, T.（2019 年 12 月 19 日）。用于可解释多视野时间序列预测的时间融合变
换器。arXiv.org。https : //arxiv.org/abs/1912.09363

18. Laborda, J.、Ruano, S. 和 Zamanillo, I. (2023)。使用时间融合变换器进行多国和多视野 GDP 预测。数学， 11(12)，
2625。https ://doi.org/10.3390/math11122625

19. Terven, J.、Cordova-Esparza, DM、Ramirez-Pedraza, A.、Chavez-Urbiola, EA 和 Romero-Gonzalez,
JA (2023 年 7 月 5 日)。深度学习中的损失函数和指标。arXiv.org。https : //arxiv.org/abs/2307.02694

20. Jaiswal, R. 和 Singh, B. (2023)。时间序列分析中深度神经网络损失函数的比较研究。电气工程讲义（第 147-163
页）。https: //doi.org/10.1007/978-981-99-3481-2_12

21. Dosovitskiy, A.（2020 年 4 月 27 日）。使用损失训练优化多个损失函数。
条件 谷歌 研究 博客。来自https://research.google/blog/检索
optimizing-multiple-loss-functions-with-loss-conditional- training/

22.科恩克 R.和巴塞特， G. （2007 回归 分位数。
年）。https : //www.semanticscholar.org/paper/Regression-Quantiles-Koenker-Bassett/
c09db7439505f49a0958f68e782df94b3807341a

23. MacFarland, TW 和 Yates, JM (2016)。Spearman 等级差异相关系数。在Springer 电子书中（第 249-297
页）。https ://doi.org/10.1007/978-3-319-30634-
6_8

24. Vishwesh,K.（2023 年 4 月 29 日）。时间序列预测的平均方向准确度 - 数据科学特技。数据科学特技。
https: //datasciencestunt.com/mean-directional-accuracy-of-time-series-forecast/

25. Lev, B. (1983). 收入时间序列特性的一些经济决定因素。
《会计与经济学杂志》， 5,31-48。https: //doi.org/10.1016/0165-
4101(83)90004-6

26. Giacomazzi, E.、Haag, F. 和 Hopf, K. (2023)。使用时间 Fusion Transformer 进行短期电力负荷预测:电网
层次结构和数据源的影响。arXiv.org。https : //doi.org/10.1145/10.1145/3575813.3597345

27. Santos, ML、García-Santiago, X.、Camarero, FE、Gil, GB 和 Ortega, PC (2022)。
时间融合Transformer在日前光伏电力预测中的应用。
能源， 15(14),5232。https ://doi.org/10.3390/en15145232

28. Zeng, Z.、Kaur, R.、Siddagangappa, S.、Rahimi, S.、Balch, T. 和 Veloso, M. (2023 年 4 月 11 日)。使用
CNN 和 Transformer 进行金融时间序列预测。arXiv.org。https : //arxiv.org/abs/2304.04912

29. Karpathy, A.、Johnson, J. 和 Fei-Fei, L. (2015 年 6 月 5 日)。可视化和理解
循环网络。arXiv.org。https : //arxiv.org/abs/1506.02078

30. Sezer, OB、Gudelek, MU 和 Ozbayoglu, AM (2019 年 11 月 29 日)。基于深度学习的金融时间序列预测:系
统文献综述:2005-2019 年。arXiv.org。https : //arxiv.org/abs/1911.13288