Computer Science and Application 计算机科学与应用, 2024, 14(2), 249-259
Published Online February 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.142025

Hans汉斯

基于LSTM、Transformer和LightGBM的机构备付金预测方法

冀乃庚，周昕博

中国银联股份有限公司，上海

收稿日期：2024年1月15日；录用日期：2024年2月16日；发布日期：2024年2月23日

摘　要

机构备付金是金融机构的重要指标之一，对于评估其稳定性和偿付能力具有重要意义。在第三方支付机构备付金集中存管的背景下，准确预测支付机构备付金的变动对于监管机构风险管理等方面具有重要价值。笔者提出了一种基于LSTM、Transformer和LightGBM的机构备付金预测模型。利用树模型针对表格数据的快速性和准确性，选取交易日志的关键特征；利用Transformer的全局上下文建模能力捕捉财务文件的局部特征；最后采用LSTM算法获取结合后的数据的长期依赖关系。实验结果表明：该模型在机构备付金方面的预测准确性优于ARMA算法、LSTM算法和时序预测Transformer模型。

关键词

深度学习，时间序列预测，长短期记忆网络，Transformer，LightGBM

A Prediction Method for Institutional Reserve Based on LSTM, Transformer, and LightGBM

Naigeng Ji, Xinbo Zhou

China UnionPay Co., Ltd., Shanghai

Received: Jan. 15th, 2024; accepted: Feb. 16th, 2024; published: Feb. 23rd, 2024

Abstract

Institutional reserve is one of the important indicators for evaluating the stability and solvency of financial institutions. Accurate prediction of changes in payment institution reserves is of significant value for risk management and regulation by regulatory authorities in the context of centralized

文章引用: 冀乃庚, 周昕博. 基于 LSTM、Transformer 和 LightGBM 的机构备付金预测方法[J]. 计算机科学与应用, 2024, 14(2): 249-259. DOI: 10.12677/csa.2024.142025

custody of reserves for third-party payment institutions. This paper proposes a prediction model for institutional reserve based on LSTM, Transformer, and LightGBM. The LightGBM model is utilized to extract key features of transaction logs, because tree-based model is fast and accurate in tabular data. The Transformer model is utilized to capture the local features of financial documents with its ability to model global context. Lastly, the LSTM algorithm is employed to capture the long-term dependencies of the combined data. Experimental results demonstrate that the proposed algorithm outperforms ARMA, LSTM, and Transformer models in predicting institutional reserves.

Keywords

Deep Learning, Time Series Forecasting, LSTM, Transformer, LightGBM

1. 引言

机构备付金是支付机构用于应对风险和支付义务的储备资金。准确预测机构备付金的变动对于支付机构和监管机构具有重要意义。机构备付金金额受到工作日周期、季节周期、交易活动、政策影响、市场趋势、特殊事件等多方面复杂因素影响，很难人工对各个因子的影响下进行综合量化分析[1]。

机构备付金预测问题属于标准的时间序列预测问题。目前，类似的研究主要包含以下几类[2]：1) 基于统计的方法；2) 传统时间序列分析方法[3] [4]，如 ARMA；3) 机器学习方法[5]，如 SVM；4) 深度学习方法，如 LSTM [6]、Prophet [7]等方法。

金融领域的时间序列预测往往存在以下特点：长期的预测需要考虑趋势信息，短期的预测需要考虑细粒度的波动性，一些周期规律不太明显的场景，依赖关系会随着时间动态变化。由于备付金数据的复杂性和非线性特征，传统方法在捕捉备付金变动中的复杂关系方面存在局限性。随着这些年来深度学习技术的不断发展，深度学习方法能够自动学习和提取复杂的时间序列模式，具有强大的上下文理解能力，用于时间序列预测具有显著的优势。

但是，现有的机构备付金预测方法往往仅能引入节假日、季节等简单特征，依赖在历史时序趋势中添加少量“拐点”实现对金额曲线的逼近。对于真正重要的多方面复杂因素影响，未能提出可靠方案引入，难以获得更准确的预测结果。如何将政策影响、市场趋势、特殊事件等高维度的信息引入预测模型，如何将复杂交易日志引入模型获取细粒度波动信息，是本文研究的关键点。为此，我们提出了一种由 Transformer、LightGBM 和 LSTM 三个模型堆叠实现的机构备付金预测模型。Transformer 模型负责分析财务报表、文件等的局部特征，LightGBM 模型负责提取交易日志的核心特征，LSTM 模型获取前两者的输出作为输入，结合历史数据、时间区间特征、季节周期等特征进行全局建模，挖掘备付金变动中的长期依赖关系和全局因素，完成机构备付金预测。

2. 模型关键技术设计

2.1. 基于 LightGBM 的交易特征提取模型

在考虑机构备付金时，不仅仅聚焦于最后的汇总值。各个类型、不同主体、不同内容的交易日志与之息息相关。其中隐含着机构特点、市场因素、特殊事件等多方面复杂因素，有些交易与最终机构备付

金额度直接相关，有些交易有隐性的关联。在纷乱复杂的交易中提取出关键信息，结合交易参与方等相关基本参数，是备付金预测模型的关键特征。

对于此类特征不均匀、样本量小、极值较大的表格类数据，综合考虑准确性和效率性，采用树模型往往优于 NN 模型。GBDT (Gradient Boosting Decision Tree)是机器学习中一个长盛不衰的模型，在各大机器学习竞赛中占据优势地位。GBDT 的主要思想是利用弱分类器(决策树)迭代训练以得到最优模型。

LightGBM (Light Gradient Boosting Machine) [8]是一个由微软开源的 GBDT 算法工程实现框架，并引入了基于梯度的单侧采样(GOSS)和独占特征绑定(EFB)等创新性技术，支持高效率的并行训练，具有更快的训练速度、更低的内存消耗、更好的准确率。在大规模数据和高维特征的处理中表现出色。本文使用 LightGBM 构建交易特征提取模型。

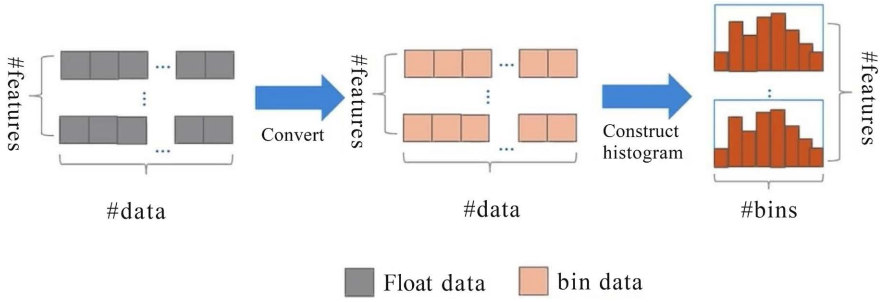

Figure 1. Principle of histogram statistics in LightGBM
图 1. LightGBM 直方图统计原理

如图 1 所示，LightGBM 采用直方图算法和直方图作差的计算方法，将连续特征以分桶的形式转为离散特征。在梯度提升时采用最小化损失函数加上正则化项作为决策树的优化目标：

$$\mathcal{L}(\theta)=\sum_{i=1}^{n}\ell(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{1}$$

其中，$\mathcal{L}(\theta)$ 是模型的损失函数，θ 是模型的参数，n 是样本数量，y_i 是真实标签，$\hat{y}_i$ 是预测标签，$\ell(y_i,\hat{y}_i)$ 是损失函数，K 是树的数量，f_k 是第 k 个树，$\Omega(f_k)$ 是正则化项。

损失函数的具体形式取决于任务类型，在这里，我们采用回归问题常用的平方损失函数，将预测的备付金汇总值 score 与真实值的均方根误差 RMSE (Root Mean Squared Error)作为评价指标：

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2} \tag{2}$$

接下来，按如下步骤构建和训练模型。

2.1.1. 数据预处理

对金额类字段进行归一化处理、对部分类型参数字段执行 one-hot 转化、对部分字段执行 Word2Vec 处理嵌入特征。

2.1.2. 特征工程

按照类型分类，包含以下特征：

1) 基本信息特征：参数信息、交易基本信息；

2) 聚合特征：sum/mean/max/min/median/std/skew 包括金额的各维度分类聚合、交易笔数的各维度分类聚合、时间衰减加权金额聚合；

3) SVC 特征：交易序列 Tfidf 矩阵降维；

4) 时间区间特征：计算时间差、lag 等信息；

5) 交互特征：历史交易和当日新交易的简单交互计算；

6) meta 特征：融合历史交易数据，将预测值 score 作为新的特征加入表中，计算 min/sum。

2.1.3. 特征选取

1) 使用数据可视化方式构建图表，分析交易字段业务含义，进行人工选取；

2) 使用 Boruta 特征选择方法[9]进行进一步筛选；

3) 通过 lightGBM 的特征重要性选取特征。

2.1.4. 模型训练

采用五折交叉验证(KFold)进行训练，经过不同特征选择后，训练得到最佳模型。

最后，根据 LightGBM 的特征重要性方法，选取出重要性 Top20 的特征，与预测值 score 一起作为每日交易局部特征，作为最终 LSTM 模型的输入值(每日)，用于后续训练和预测。

2.2. 基于 Transformer 的财务特征提取模型

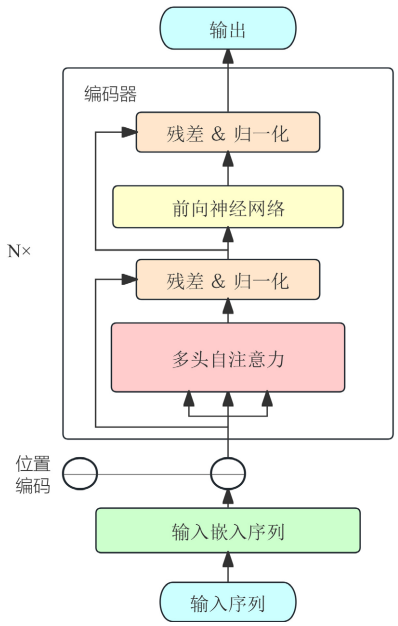

Figure 2. Structure of Transformer Encoder
图 2. Transformer Encoder 结构图

Transformer [10]是一种基于自注意力机制(Self-Attention)的神经网络架构，用于解决传统的序列到序列(sequence-to-sequence)模型在处理可变长序列时遇到的问题。Transformer 的设计旨在解决传统循环神经网络(RNN)在处理长序列时遇到的一些问题，如难以捕捉长距离的依赖关系和并行计算的限制。Transformer 通过引入自注意力机制，使得模型能够在不同位置之间直接建立关联，从而更好地捕捉序列中的依赖关系。除此之外，Transformer 还引入了残差连接和层归一化等技术，以加快训练和提高模型的表达能力，使用了位置编码来为序列中的每个位置提供信息，以保留序列的顺序信息。Transformer 模型在自然语言处理(NLP)领域获得了巨大的成功，并逐渐衍生到其他各类领域，颇有一种大一统的趋势。

本文只采用 Transformer 模型的编码器部分，依靠 self-attention 机制提取财务数据特征，编码器结构

如图 2 所示。参考[11]中嵌入向量概率分类方式，同时借助 BERT 预训练模型实现目标。

财务报表、文件、公告中，包含了政策因素、市场因素等复杂信息。但是不同类别的财务数据即包含表格类数据，又包含文本数据，难以处理。为获取更多的关键特征用于后续训练，综合考虑性能和资源，本文的 transformer 模型采用两个模型，分步实现：

1) 财务数据文本特征向量提取(单语句)

如图 3 所示，采用 BERT 预训练模型。将财务数据的文本经过清洗和预处理，去除空格、标点符号等无关信息，并将文本转换为统一的格式。准备一个包含文本和标签(财务指标、财务风险)的数据集，使用 BERT 模型进行训练，得到文本分类模型。将 BERT 模型输出的特征向量作为特征，用于第二步训练。特征向量包含了文本中所有单词的特征信息，可以反映文本的整体语义。

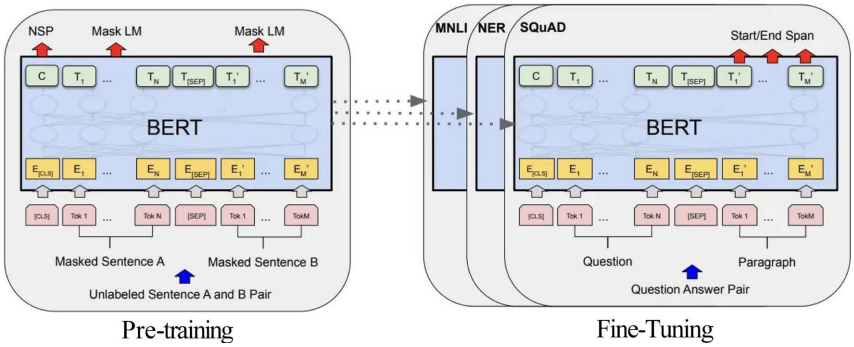

Figure 3. BERT single-sentence classification model
图 3. BERT 单语句分类模型

2) 财务表格类数据概率分类

财务表格数据包含财务报表以及第一步模型获得的特征向量。如图 4 所示，首先，将表格数据转换为适合 Transformer 输入的形式(表格数据的每一行表示一个样本，每一列表示一个特征)，使用嵌入层将每个特征的离散值映射为连续的向量表示。采用 Transformer 编码器提取其中潜在的嵌入向量，然后将其传递到最终具有 softmax 激活的多层感知器(MLP)，以生成符号分类输出。输出采用每个类别的概率(增长、持平、降低)，三个类别的概率总和为 1。

嵌入向量概率分类方式，同时借助 BERT 预训练模型实现目标。

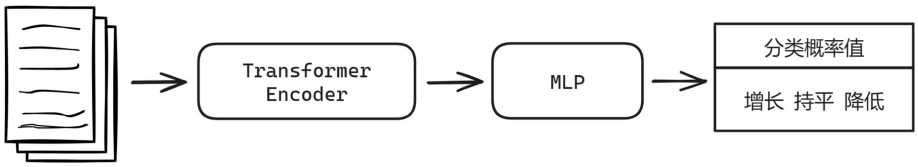

Figure 4. My Transformer model architecture
图 4. 模型整体架构

最后，将分类概率值作为特征，作为最终 LSTM 模型的输入值，用于后续的训练和预测。

2.3. 基于 LSTM 的备付金时序预测模型

长短时记忆网络(Long Short Term Memory，简称 LSTM)模型[12]，本质上是一种特定形式的循环神经网络(Recurrent Neural Network，简称 RNN)。LSTM 模型在 RNN 模型的基础上通过增加门限(Gates)来解决 RNN 短期记忆的问题，使得循环神经网络能够真正有效地利用长距离的时序信息。LSTM 在 RNN

的基础结构上增加了输入门限(Input Gate)、输出门限(Output Gate)、遗忘门限(Forget Gate) 3 个逻辑控制单元，且各自连接到了一个乘法元件上，通过设定神经网络的记忆单元与其他部分连接的边缘处的权值控制信息流的输入、输出以及细胞单元(Memory cell)的状态。其具体结构如下图(图 5)所示：

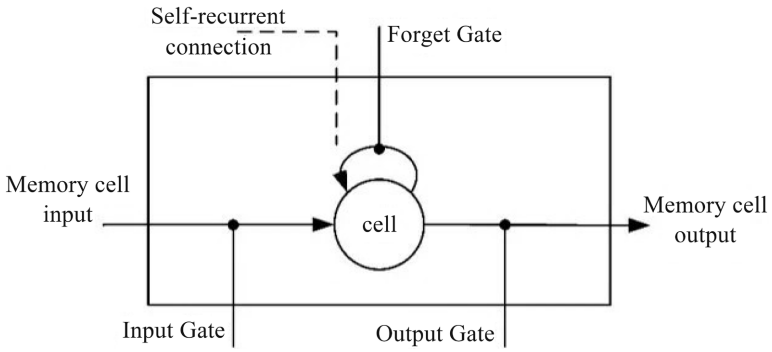

Figure 5. LSTM specific structure
图 5. LSTM 具体结构

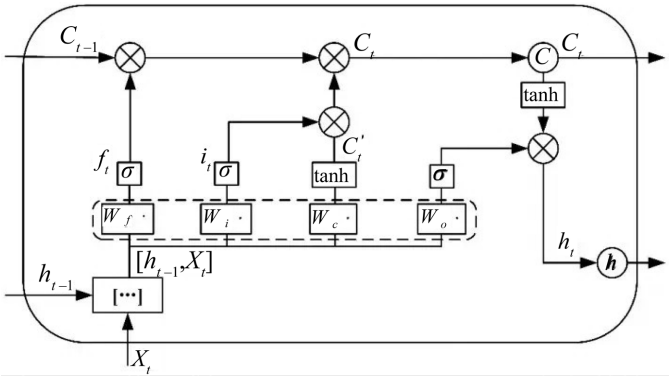

Figure 6. LSTM hidden layers structure
图 6. LSTM 隐藏层结构图

t 时刻，LSTM 神经网络公式如下：

$$\begin{aligned} f_t &= \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \\ i_t &= \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_i\right) \\ \tilde{C}_t &= \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ o_t &= \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \\ h_t &= o_t \odot \tanh\left(C_t\right) \end{aligned} \tag{3}$$

其中，x_t 是输入 h_t，是输出，C_t 是状态，f 是遗忘门，i_t 是输入门，$\tilde{C}_t$ 是新的状态 o_t，是输出门，W 和 b 是权重和偏置，σ 是 sigmoid 函数，tanh 是双曲正切函数。

隐藏层 cell 结构图如图 6 所示。在 LSTM 神经网络的训练过程中，首先将 t 时刻的数据特征输入至输入层，经过激励函数输出结果。将输出结果、t − 1 时刻的隐藏层输出和 t − 1 时刻 cell 单元存储的信息输入 LSTM 结构的节点中，通过 Input Gate，Output Gate，Forget Gate 和 cell 单元的处理，输出数据到下一隐藏层或输出层，输出 LSTM 结构节点的结果到输出层神经元，计算反向传播误差，更新各

个权值。

LSTM通过使用门控单元解决了传统的RNN在处理长期依赖性时可能会出现梯度消失或梯度爆炸的问题，使其特别适合用于时间序列预测问题。另一方面，实践下来，相较于当下火热的Transformer，在备付金预测问题上LSTM表现也更优。Transformer Encoder依靠基于位置编码的特性，在时序性的建模方面没有LSTM等RNN算法框架直接，无法很好地捕获这些时序关系，难以调参获得比RNN更优的结果[13]。在后续的模型实验结果中，本文列出了LSTM模型和Transformer时序预测类模型的测试对比。因此，本文选择使用LSTM算法完成备付金时序预测。

我们将时序预测的时间步长设置为14，即：使用过去14天的数据预测未来值。

按如下步骤构建和训练模型。

2.3.1. 模型构建

建立一个三层的线性堆叠神经网络(Sequential模型)，包含两个LSTM层和一个Dense层，如图7所示：

第一个LSTM层返回每个时间步长的输出序列；

第二个LSTM层返回最后一个时间步长的输出序列；

最后的Dense层输出模型的预测值。

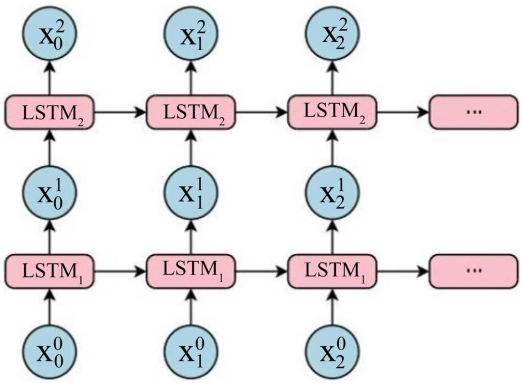

Figure 7. Dual-layer LSTM structure
图 7. 双层 LSTM 结构

2.3.2. 数据预处理

对金额类字段进行归一化处理、对部分类型参数字段执行one-hot转化。

2.3.3. 输入维度

包含以下特征：

1) 机构备付金历史数据；
2) 节假日特征、季节周期性特征等；
3) 基于LightGBM提取的交易特征；
4) 基于Transformer提取的财务特征。

2.3.4. 模型训练

使用网格搜索选取和优化网络的超参数，采用五折交叉验证(KFold)进行训练。

将传统时间序列预测数据、特征与LightGBM、Transformer模型提供的关键特征相结合后，使用LSTM模型进行训练，得到了最终的机构备付金预测模型。最终模型架构图如下(图8)。

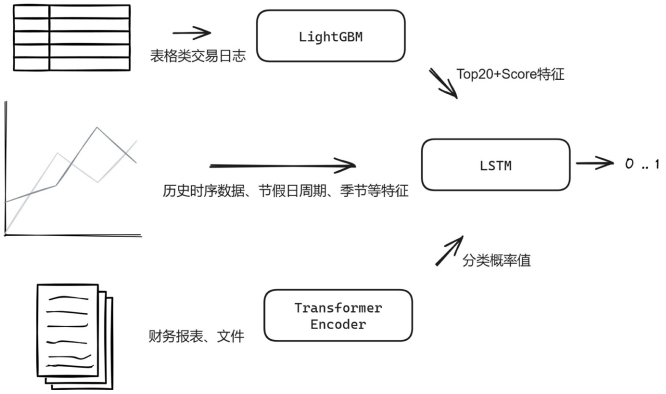

Figure 8. Overall architecture of my institutional reserve prediction model
图 8. 机构备付金预测模型整体架构

3. 模型预测结果分析

3.1. 数据和结果

本实验采用的数据为 2020.01~2023.05 时间内，某金融机构的全量相关交易数据、关联的账务核算报表、每日财务公告，数据来源于中国银联大数据数据仓库中，所有交易数据的业务字段经过脱敏处理。

节假日特征包含两部分，数据来源分别为：1) 全国法定节假日列表；2) 中国人民银行大额系统节假日表、中国人民银行大小额系统例行维护计划。以引入和区分节假日交易影响、大小额系统对业务办理的影响。季节周期性特征等采用 one-hot 进行转化而来。

将全量数据以 7:2:1 的比例划分训练数据集、测试数据集和验证数据集，从而更全面准确地评估模型准确率。

借助 Prophet 模型对机构备付金趋势、节假日、周期性进行可视化分析，如下图(图 9~12)所示：

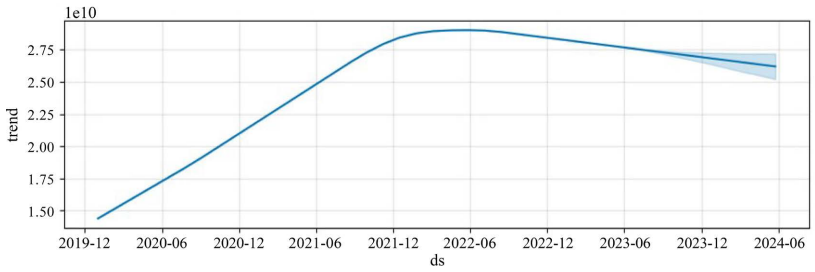

Figure 9. Growth trend on Institutional reserve
图 9. 增长趋势

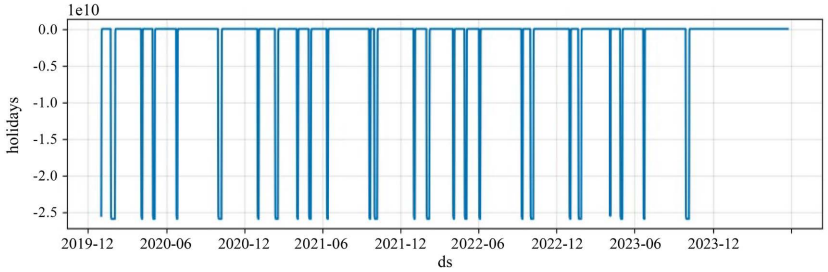

Figure 10. Impact of holidays on Institutional reserve
图 10. 节假日影响

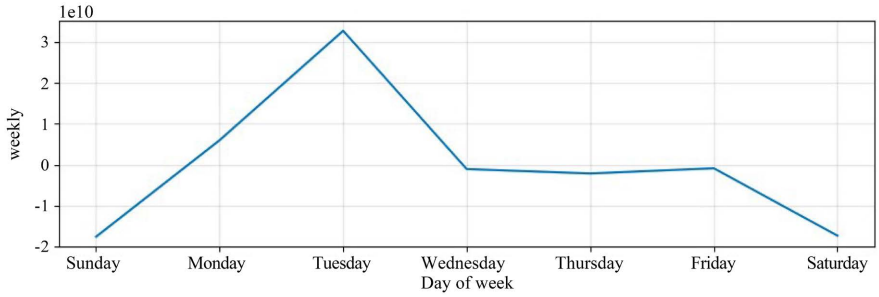

Figure 11. Impact of weekly cycles on Institutional reserve
图 11. 每周周期影响

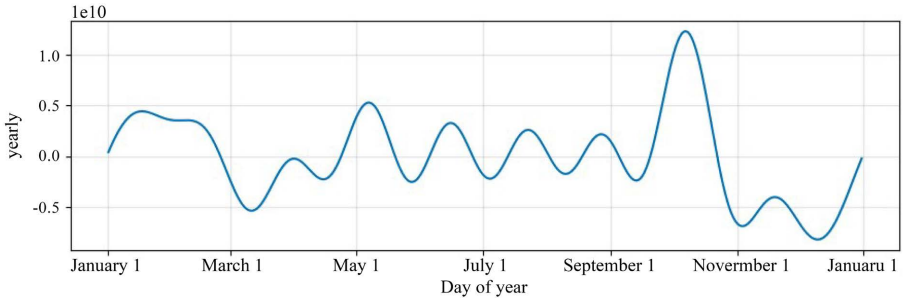

Figure 12. Impact of annual cycles on Institutional reserve
图 12. 每年周期影响

可以看出，机构备付金整体趋势逐渐增加直至平缓，符合备付金集中收缴的业务趋势。节假日期间备付金金额很小，而节假日后的第一、二个工作日，机构备付金额度有较大波动，符合业务预期。

结合以上所有特征，最终预测结果如下图(图 13)所示，较好地贴合了机构备付金额度变化曲线。

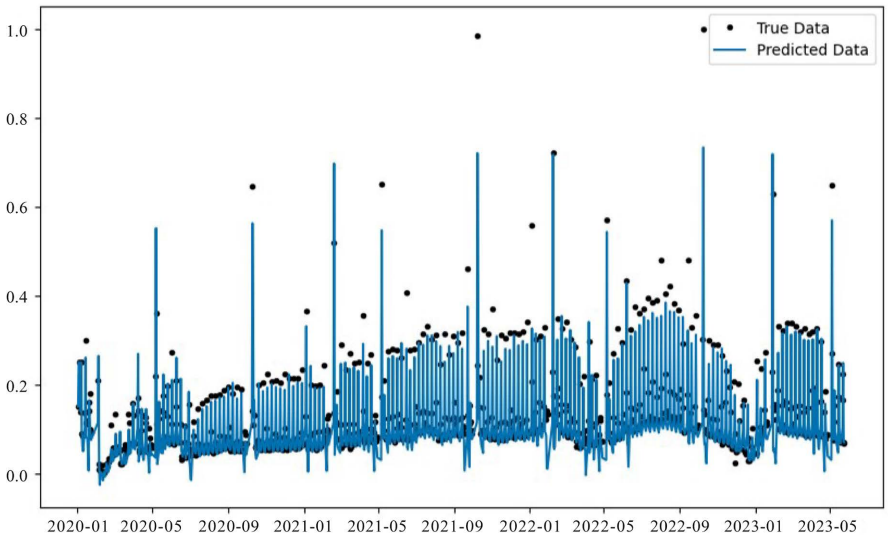

Figure 13. Institutional reserve prediction results
图 13. 机构备付金预测结果

3.2. 结果分析

为评估最终模型在备付金预测场景中的效果，将其与以下模型进行预测效果对比。

1) ARMA 模型；

2) LSTM 模型：纯时间序列预测的 LSTM、包含节假日等多维度特征的 LSTM (不包含本文 LightGBM 和 Transformer 模型特征)；

3) Prophet 模型；

4) Transformer 模型：时序预测模型，包括 LogSparse Transformer [14]和 TCCT [15]。

结合业务需求和数据特征，采取均方根误差(RMSE)作为指标，对结果进行评估。如表 1 和图 14 所示，本模型在机构备付金方面的预测准确性优于 ARMA 算法、LSTM 算法和常见的时序预测 Transformer 模型。其中深度学习模型均优于传统 ARMA 模型，体现了神经网络对于复杂时间序列模式的学习能力和上下文理解能力。而 Transformer 模型在该业务场景下的表现未能优于 LSTM 算法。

Table 1. Comparison table of accuracies
表 1. 各模型的准确度对比表

算法	训练数据	测试数据	验证数据
本模型	0.0377	0.0452	0.0444
ARMA		0.1105	
LSTM1	0.0522	0.0631	0.0626
LSTM2	0.0416	0.0524	0.0536
Prophet	0.0564	0.0701	0.0723
LogSparse Transformer	0.0665	0.0789	0.0811
TCCT	0.0635	0.0773	0.0792

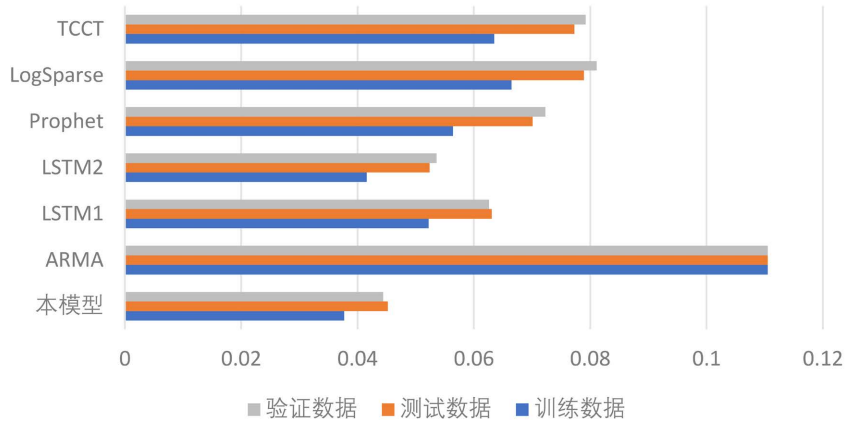

Figure 14. Comparison graph of accuracies
图 14. 各模型的准确度对比图

4. 结语

本文提出了一种基于 Transformer、LightGBM 和 LSTM 堆叠实现的机构备付金预测模型，成功将部分市场趋势、交易因素等高维度特征以财务报表和交易日志等数据的形式，抽象降维后，引入了备付金预测问题中。并验证了，增加此类高维度特征后，可以更准确地预测机构备付金数据。

本研究提供了详细的模型构建思路、细节和实验结果，为机构备付金预测问题提供了有效的思路和方向。下一步计划加入更多维度、更多数据，建立更准确的、更泛化的机构备付金预测模型。

参考文献

[1] 黄平, 周晋. 银行日常风险管理中备付金问题研究[J]. 系统管理学报, 2013(2): 212-216.

[2] 李丽萍, 段桂华, 王建新. 基于 Prophet 框架的银行网点备付金预测方法[J]. 中南大学学报: 自然科学版, 2019, 50(1): 75-82.

[3] Liu, C., Hoi, S.C.H., Zhao, P., *et al*. (2016) Online Arima Algorithms for Time Series Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, **30**. https://doi.org/10.1609/aaai.v30i1.10257

[4] 洪玮. ARMA 时间序列模型在自助设备现金需求预测中的应用[J]. 中国金融电脑, 2012(9): 87.

[5] Lu, C.J., Lee, T.S. and Chiu, C.C. (2009) Financial Time Series Forecasting Using Independent Component Analysis and Support Vector Regression. *Decision Support Systems*, **47**, 115-125. https://doi.org/10.1016/j.dss.2009.02.001

[6] Liu, Y., Dong, S., Lu, M. and Wang, J. (2019) LSTM Based Reserve Prediction for Bank Outlets. *Tsinghua Science and Technology*, **24**, 77-85. https://doi.org/10.26599/TST.2018.9010007

[7] Taylor, S.J. and Letham, B. (2018) Forecasting at Scale. *The American Statistician*, **72**, 37-45. https://doi.org/10.1080/00031305.2017.1380080

[8] Ke, G., Meng, Q., Finley, T., *et al*. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, **30**, 3149-3157.

[9] Kursa, M.B. and Rudnicki, W.R. (2010) Feature Selection with the Boruta Package. *Journal of Statistical Software*, **36**, 1-13. https://doi.org/10.18637/jss.v036.i11

[10] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**, 5998-6008.

[11] Zeng, Z., Kaur, R., Siddagangappa, S., *et al*. (2023) Financial Time Series Forecasting Using CNN and Transformer. arXiv: 2304.04912.

[12] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[13] Zeng, A., Chen, M., Zhang, L., *et al*. (2023) Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*, **37**, 11121-11128. https://doi.org/10.1609/aaai.v37i9.26317

[14] Li, S., Jin, X., Xuan, Y., *et al*. (2019) Enhancing the Locality and Breaking the Memory Bottleneck of Transformer on Time Series Forecasting. *Advances in Neural Information Processing Systems*, **32**, 5243–5253.

[15] Shen, L. and Wang, Y. (2022) TCCT: Tightly-Coupled Convolutional Transformer on Time Series Forecasting. *Neurocomputing*, **480**, 131-145. https://doi.org/10.1016/j.neucom.2022.01.039