基于机器学习模型的汇率预测实证分析

1. 研究问题与数据介绍

本作业以美元指数的日度对数收益率为研究对象，考察多种机器学习方法在汇率预测中的样本外表现。数据来源为英为财情（Investing.com），最终样本区间为 2006 年 3 月至 2025 年 11 月，频率为日度。通过选取美元指数锚定的汇率币种、相关国家利率、国际大宗商品收益率、以及在原始汇率收益率基础上构造多阶滞后收益率共 36 个解释变量。具体变量定义、数据清洗与构造步骤，各变量的时间趋势、频数分布和相关性热力图见附录 A。

2. 研究方法

为避免信息泄露，本文按照时间顺序将样本划分为训练集、验证集和测试集，在验证集上调参并在测试集上评估样本外预测能力。在线性模型方面，选取 Ridge 和 LASSO 回归，在非线性模型方面，采用随机森林、XGBoost 以及 LSTM 模型进行比较，部分扩展模型如 Attention-LSTM、Transformer 的结果列于附录 C。模型性能主要通过测试集的 RMSE、MAE 与 R^2进行衡量，各模型超参数设定与搜索范围见附录 B。

3. 描述性统计

表 1 报告了目标变量及主要解释变量的描述性统计结果。其中，USD_index 所代表的前五个变量为汇率变量，GOLD 与 WTI 代表大宗商品变量、US_10Y 等代表各国十年期国债利率、US_index 等代表各国市场收益率。更详细的统计结果见附录 A 表 A1。

表 1 目标变量及主要解释变量的描述性统计

	count	mean	std	min	25%	50%	75%	max
USD_index_Close	3456	-2.90	485.44	-3064.57	-280.57	0.00	273.44	2407.47
USD_EUR_Open	3456	-6.43	583.52	-3631.46	-334.00	0.00	306.00	3137.21
USD_JPY_Open	3456	2.19	552.31	-3889.67	-276.97	9.14	301.93	3433.27
USD_GBP_Open	3456	-7.42	607.79	-2983.84	-337.93	-15.70	316.89	9200.58
USD_CNY_Open	3456	-1.23	197.64	-1589.15	-74.12	-1.61	61.98	2444.78
GOLD_Open	3456	29.56	1151.93	-9039.81	-521.50	57.58	645.78	10537.33
WTI_Open	3456	-40.67	2436.14	-26188.93	-1258.53	19.07	1248.40	16694.57
US_10Y_Open	3456	2.95	1.13	0.51	2.03	2.82	3.87	5.29
CN_10Y_Open	3456	3.31	0.64	1.62	2.93	3.32	3.68	4.73
UK_10Y_Open	3456	2.75	1.49	0.08	1.42	2.68	4.18	5.58
JP_10Y_Open	3456	0.76	0.63	-0.29	0.09	0.73	1.33	2.01
GER_10Y_Open	3456	1.81	1.51	-0.79	0.39	1.85	3.07	4.69

	count	mean	std	min	25%	50%	75%	max
VIX_Open	3456	19.73	9.01	9.10	13.84	17.30	22.81	82.69
US_index_Open	3456	27.78	1138.51	-9114.04	-394.86	80.22	560.42	10140.64
CN_index_Open	3456	-13.22	1772.22	-11878.16	-736.78	-15.93	819.62	13492.15
UK_index_Open	3456	-4.46	1143.05	-11512.43	-490.55	42.69	561.80	8666.81
JP_index_Open	3456	5.01	1314.71	-10318.81	-686.80	8.55	734.83	11052.20
GER_index_Open	3456	33.35	1305.93	-10380.90	-571.91	83.90	708.36	9283.15

注：汇率、大宗商品和市场指数变量均计算收益率，并放大为原数量级的 10^5，利率数据单位是百分比，Open 代表开盘价、Close 代表收盘价.

4. 实证结果

表 2 比较了各模型在测试集上的预测表现。整体来看，多数模型的样本外 R^2 为负值，表明其预测误差并未优于均值基准。值得注意的是，Ridge 回归在测试集上取得了约 0.2%的正向 R^2，说明在信号极弱且分散的情形下，L2 正则化通过温和收缩系数仍能保留少量可预测成分。图 1 给出了 Ridge 回归模型的预测曲线，更多模型的结果见附录 C。

表 2 各模型的样本外表现

模型名称	**评价指标**		
	MAE	**RMSE**	**R^2**
LASSO	34.63	46.01	-0.0235
RIDGE	34.46	45.95	0.2442
Random Forest	35.82	47.47	-6.4831
XGBoost	34.54	46.34	-1.4708
LSTM	46.32	34.90	-0.4451

注：MAE 和 MSE 放大为原数量级的 10^5，R^2 单位是百分比.

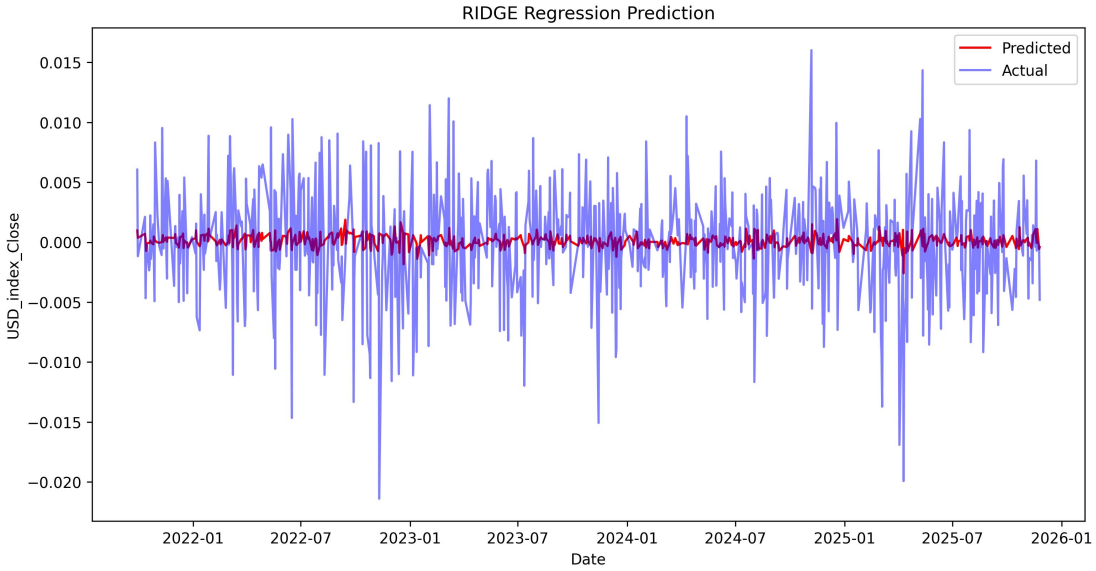

图 1 Ridge 回归模型预测与实际汇率收益率

5. 研究结论

综上，在构造 36 个潜在预测因子的基础上，本文比较了多种机器学习方法在汇率对数收益率预测中的样本外表现。结果表明，在典型高噪声的外汇市场环境中，多数复杂模型难以显著战胜简单基准，复杂模型并不必然优于适度收缩的线性模型。

附录 A：数据来源、变量定义与清洗

A.1 数据来源

本文使用的数据全部来自于英为财经新闻网（Investing.com）。

A.2 变量定义

在本研究中，我们使用了一系列金融数据变量来构建汇率预测模型。以下是各变量定义：

1.汇率

表示某一货币相对于其他货币的比率。本文主要用到了美元和欧元、美元和英镑、美元和日元、美元和人民币的汇率数据。

2.大宗商品价格

大宗商品价格是市场中大批量交易的、广泛应用于工业和农业生产的商品。本文主要用到了黄金(GC1)和石油期货(CLC1)的商品价格。

3.无风险利率

无风险利率是指在没有任何风险的情况下，投资者能够获得的预期回报率，通常以国债收益率作为参考。本文主要选取美国、中国、英国、日本和德国的十年期国债收益率作为主要的无风险利率。

4.市场收益率

市场收益率是指在特定市场中，所有投资资产在一定时间内的平均回报率，是衡量市场整体表现的重要指标。本文主要选取标准普尔 500 指数、沪深 300 指数、日经 225 指数、DAX30 指数和富时 100 指数作为各个国家的股票市场收益率。

5.市场波动率

市场波动率是指特定市场中，金融资产价格的波动程度，反映了资产的风险程度和市场情绪高低。本文选取了芝加哥场商品交易所发布的 VIX 指数作为市场波动率的衡量。

A.3 数据清洗

数据清洗是数据分析的基础，我们首先对原始数据进行了以下处理：

1.缺失值处理

根据各个变量的可得性，我们没有采取填充的方式补全空值，而是简单剔除了所有包含空值的数据记录。因此最终得到的数据集是各个变量可得数据的交集。

2.时间对齐

所有的数据集都进行了时间对齐，确保每一时期的各个变量值能够一致。并且在之后的模型训练和预测时，都关注了时间先后的问题，避免训练偏差（look-ahead bias）。

A.4 变量构造

在完成数据清洗之后，我们对原始数据进行了如下变量构造：

1.对数收益率计算

对美元指数、汇率、大宗商品价格变量，进行了对数收益率的转换。对数收益率通过以下公式计算：

$$\log\_\text{Return}_t = \log\frac{P_{t+1}}{P_t}$$

其中，P_{t+1} 是第 $t+1$ 期的汇率数据，P_t 是前一期的汇率数据。

2.变量做差计算

对 VIX 指数自身进行了差分计算，得到变化量序列；并对各国利率之间进行了做差，得到各国间共 10 个利率差序列。

3.滞后变量构造

考虑到汇率变化的滞后效应，我们构建了若干滞后变量。例如，构建了美元指数过去 1 天、3 天、7 天、14 天滞后序列，用来捕捉短期内汇率变动的影响。

4.滚动窗口特征

为了进一步捕捉时间序列数据的动态变化，我们使用了滚动窗口方法构造了美元指数过去 3 天、7 天 14 天期的标准差序列。

最后，表 A1 中汇报了所有变量的描述性统计。图 A1、图 A2、图 A3 分别汇报了各变量的时间序列图、频数分布图和相关热力图。

表 A1 变量描述性统计

	count	mean	std	min	25%	50%	75%	max
USD_index_Close	3456	-2.90	485.44	-3064.57	-280.57	0.00	273.44	2407.47
USD_EUR_Open	3456	-6.43	583.52	-3631.46	-334.00	0.00	306.00	3137.21
USD_JPY_Open	3456	2.19	552.31	-3889.67	-276.97	9.14	301.93	3433.27
USD_GBP_Open	3456	-7.42	607.79	-2983.84	-337.93	-15.70	316.89	9200.58
USD_CNY_Open	3456	-1.23	197.64	-1589.15	-74.12	-1.61	61.98	2444.78
GOLD_Open	3456	29.56	1151.93	-9039.81	-521.50	57.58	645.78	10537.33
WTI_Open	3456	-40.67	2436.14	-26188.93	-1258.53	19.07	1248.40	16694.57
US_10Y_Open	3456	2.95	1.13	0.51	2.03	2.82	3.87	5.29
CN_10Y_Open	3456	3.31	0.64	1.62	2.93	3.32	3.68	4.73
UK_10Y_Open	3456	2.75	1.49	0.08	1.42	2.68	4.18	5.58
JP_10Y_Open	3456	0.76	0.63	-0.29	0.09	0.73	1.33	2.01
GER_10Y_Open	3456	1.81	1.51	-0.79	0.39	1.85	3.07	4.69
VIX_Open	3456	19.73	9.01	9.10	13.84	17.30	22.81	82.69
US_index_Open	3456	27.78	1138.51	-9114.04	-394.86	80.22	560.42	10140.64
CN_index_Open	3456	-13.22	1772.22	-11878.16	-736.78	-15.93	819.62	13492.15
UK_index_Open	3456	-4.46	1143.05	-11512.43	-490.55	42.69	561.80	8666.81
JP_index_Open	3456	5.01	1314.71	-10318.81	-686.80	8.55	734.83	11052.20
GER_index_Open	3456	33.35	1305.93	-10380.90	-571.91	83.90	708.36	9283.15
US_10Y-CN_10Y_Open	3456	-0.37	1.36	-2.56	-1.45	-0.71	0.51	3.13
US_10Y-GER_10Y_Open	3456	1.14	0.84	-0.90	0.34	1.34	1.79	2.79
US_10Y-UK_10Y_Open	3456	0.19	0.62	-1.17	-0.25	0.07	0.57	1.72
US_10Y-JP_10Y_Open	3456	2.19	0.77	0.51	1.70	2.17	2.85	4.14
CN_10Y-GER_10Y_Open	3456	1.50	1.55	-1.09	0.07	1.72	3.03	3.98
CN_10Y-UK_10Y_Open	3456	0.56	1.64	-3.16	-0.70	1.18	1.94	3.15
CN_10Y-JP_10Y_Open	3456	2.56	0.86	0.00	2.12	2.73	3.14	4.12
GER_10Y-UK_10Y_Open	3456	-0.94	0.48	-2.34	-1.19	-0.89	-0.61	0.14
GER_10Y-JP_10Y_Open	3456	1.05	0.94	-0.75	0.28	1.02	1.90	3.03
UK_10Y-JP_10Y_Open	3456	2.00	0.98	0.06	1.20	1.99	2.90	4.29
VIX_Delta	3456	-0.04	1.92	-17.45	-0.80	-0.10	0.59	24.86
USD_index_Close_Lag1	3456	-2.93	485.47	-3064.57	-280.57	0.00	273.44	2407.47
USD_index_Close_Lag4	3456	-2.92	485.51	-3064.57	-281.49	0.00	273.44	2407.47
USD_index_Close_Lag7	3456	-3.13	485.38	-3064.57	-281.49	0.00	273.08	2407.47
USD_index_Close_Lag14	3456	-2.79	485.29	-3064.57	-280.20	0.00	273.08	2407.47
USD_index_Close_Std3	3456	4.08	2.55	0.06	2.27	3.64	5.38	20.88
USD_index_Close_Std7	3456	4.42	1.97	0.71	3.09	4.11	5.30	17.76
USD_index_Close_Std14	3456	4.53	1.74	1.22	3.38	4.21	5.32	13.77

注：汇率、大宗商品和市场指数以及美元指数的滞后项均计算收益率，并放大为原数量级的 10^5；美元指数的波动率序列放大为原数量级的 10^4。利率和利率差数据不变，单位是百分比。

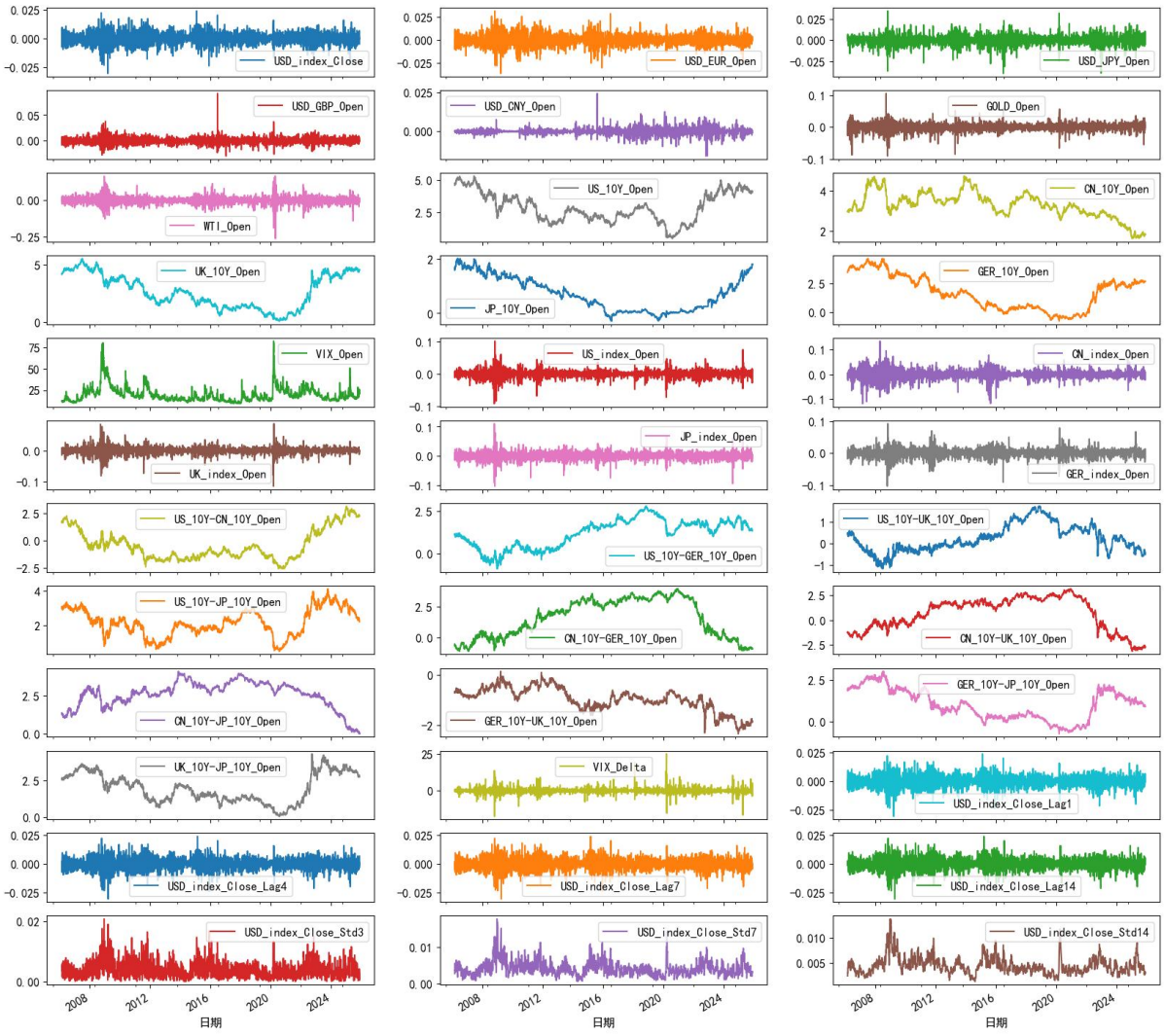
USD_index_Close
USD_EUR_Open
USD_JPY_Open
USD_GBP_Open
USD_CNY_Open
GOLD_Open
WTI_Open
US_10Y_Open
CN_10Y_Open
UK_10Y_Open
JP_10Y_Open
GER_10Y_Open
VIX_Open
US_index_Open
CN_index_Open
UK_index_Open
JP_index_Open
GER_index_Open
US_10Y-CN_10Y_Open
US_10Y-GER_10Y_Open
US_10Y-UK_10Y_Open
US_10Y-JP_10Y_Open
CN_10Y-GER_10Y_Open
CN_10Y-UK_10Y_Open
CN_10Y-JP_10Y_Open
GER_10Y-UK_10Y_Open
GER_10Y-JP_10Y_Open
UK_10Y-JP_10Y_Open
VIX_Delta
USD_index_Close_Lag1
USD_index_Close_Lag4
USD_index_Close_Lag7
USD_index_Close_Lag14
USD_index_Close_Std3
USD_index_Close_Std7
USD_index_Close_Std14
2008
2012
2016
2020
2024
日期

图 A1 变量时间序列图

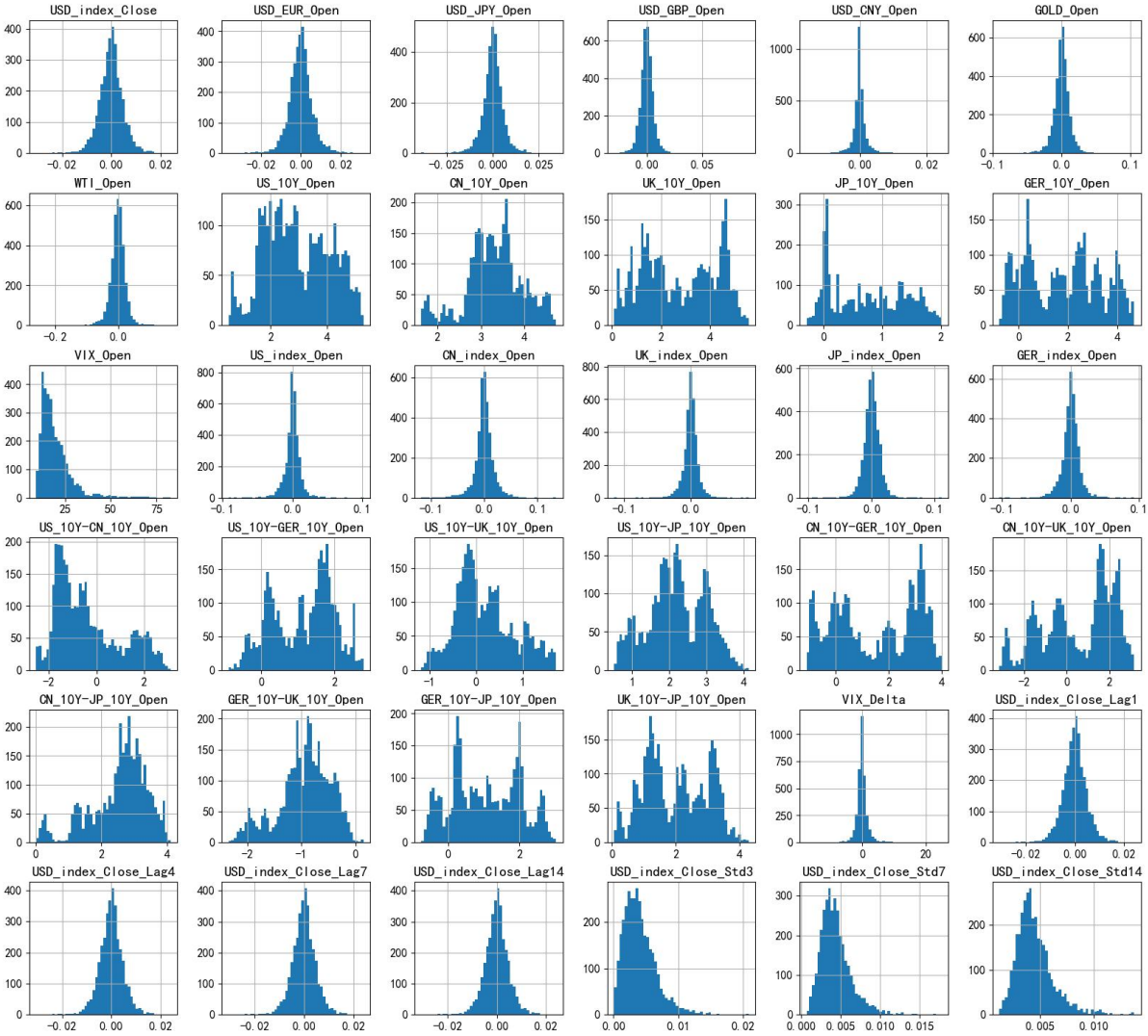
USD_index_Close
USD_EUR_Open
USD_JPY_Open
USD_GBP_Open
USD_CNY_Open
GOLD_Open
WTI_Open
US_10Y_Open
CN_10Y_Open
UK_10Y_Open
JP_10Y_Open
GER_10Y_Open
VIX_Open
US_index_Open
CN_index_Open
UK_index_Open
JP_index_Open
GER_index_Open
US_10Y-CN_10Y_Open
US_10Y-GER_10Y_Open
US_10Y-UK_10Y_Open
US_10Y-JP_10Y_Open
CN_10Y-GER_10Y_Open
CN_10Y-UK_10Y_Open
CN_10Y-JP_10Y_Open
GER_10Y-UK_10Y_Open
GER_10Y-JP_10Y_Open
UK_10Y-JP_10Y_Open
VIX_Delta
USD_index_Close_Lag1
USD_index_Close_Lag4
USD_index_Close_Lag7
USD_index_Close_Lag14
USD_index_Close_Std3
USD_index_Close_Std7
USD_index_Close_Std14

图 A2 变量频数分布图

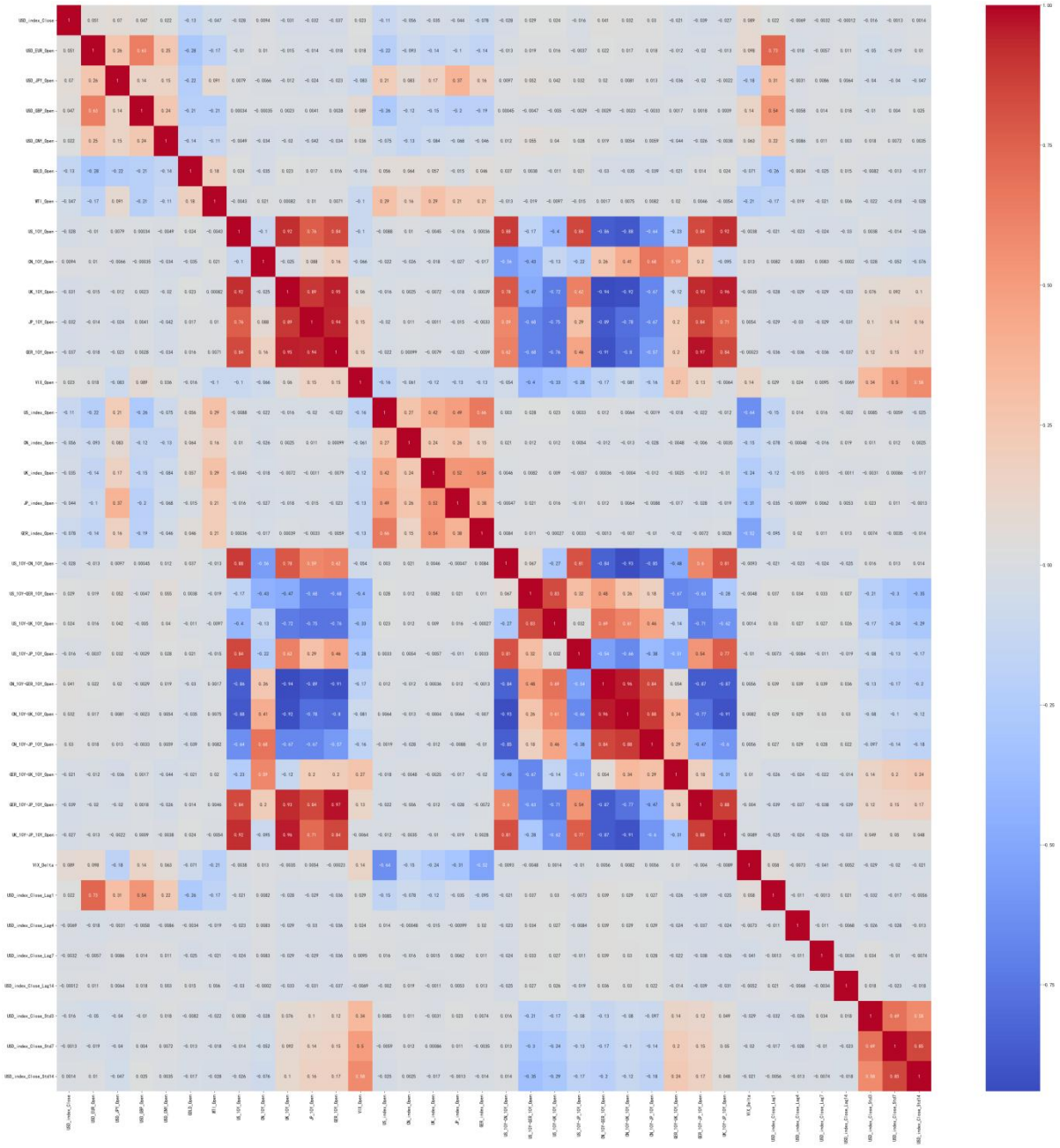

图 A3 变量相关热力图

附录 B：模型选择与训练

B.1 线性正则化模型（Ridge 与 LASSO）

本文首先考虑基于 L1 与 L2 正则项的回归，分别得到 LASSO 与 Ridge 模型。模型的超参数为惩罚项系数α。为了选定最终模型，首先根据网格选取候选的模型超参数，其次按照 0.8 和 0.2 的比例划分训练集和测试集，最后在训练集内进行 5 折拓展窗口的交叉验证，确定最有模型。

具体实现上，首先对 36 个解释变量统一进行标准化处理（减去训练集均值并除以标准差），目标变量不作缩放。其次通过网格搜索选取，搜索范围为 10^{-5} 与 10^{-3} 之间（LASSO）、10^{0} 到 10^{5} 之间（Ridge），并将按照时间顺序，将训练集平均成六份。第 i 轮利用 $1-i$ 共 i 份数据进行训练，利用第 $i+1$ 份数据进行验证，根据验证的均方根误差（RMSE）选择最优模型。确定最终模型后，会在整个训练集上重新进行一次训练，随后通过测试集进行测试。线性模型的主要超参数的搜索范围和最优值见表 B1。

表 B1 模型超参数范围以及最优值

模型名称	超参数	超参数范围	最优值
LASSO	alpha	10^{-5}–10^{-3}	7×10^{-4}
Ridge	alpha	10^{0}–10^{5}	4329

注：其中 alpha 是惩罚项系数。

B.2 树模型与集成学习模型（随机森林与 XGBoost）

为刻画潜在的非线性关系与高维交互项，本文引入随机森林与梯度提升树（XGBoost）两类树模型。两者均以 36 个标准化特征为输入，目标变量为原始对数收益率。

随机森林模型的关键超参数包括：决策树数量 n_estimators、单棵树最大深度 max_depth、内部节点最小样本数 min_samples_split、叶节点最小样本数 min_samples_leaf 以及特征采样比例 max_features 等。

梯度提升回归模型的核心超参数包括：树的数量 n_estimators、单棵树最大深度 max_depth、学习率 learning_rate、子样本比例 subsample、特征列采样比例 colsample_bytree。树类模型的主要超参数的搜索范围和最优值见表 B2。

具体的训练方法和线性模型类似。

表 B2 模型超参数范围以及最优值

模型名称	超参数	超参数范围	最优值
Random Forest	max_depth	[3,5,10, None]	None
	max_features	[“sqrt”, 0.30, 0.35]	0.30
	min_samples_leaf	[1,2,3,5]	2
	min_samples_split	[2,3,4]	3
	n_estimators	[100, 200, 250, 300]	200
XGBoost	max_depth	[6, 7, 9]	7
	subsample	[0.65, 0.70, 0.75]	0.7
	colsample_bytree	[0.65, 0.70, 0.75]	0.70
	n_estimators	[150, 200, 250]	200
	learning rate	[0.005, 0.01, 0.05]	0.01

B.3 神经网络模型（LSTM 及扩展模型）

在神经网络部分，本文采用 LSTM，Attention-LSTM 和一个简单的 Transformer 回归模型对数收益率进行序列预测。基准 LSTM 模型采用 LSTM 结构接全连接输出层。主要超参数包括：隐藏单元数（hidden_size）、窗口长度 lookback、丢弃率 dropout、学习率 learning_rate、批次大小 batch_size 以及训练轮数 epochs 等。

对于扩展模型，Attention-LSTM 的输出会通过一个时间注意力层进行加权处理，使得模型能够自动选择最重要的时刻特征。而 Transformer 模型部分采用仅包含编码器的结构。具体地，首先通过线性层将每期 36 维特征映射到高维空间，并叠加基于正余弦函数的绝对位置编码；随后输入若干层基于多头自注意力和前馈网络构成的 TransformerEncoder，在时间维度上进行全局平均池化，得到整段窗口的表示，最后通过一层全连接层映射为标量预测值，实现对下一期汇率对数收益率的预测。

在模型的训练上，与之前的机器学习模型不同，出于计算资源的限制，不进行交叉验证寻找最优模型。而是将完整样本按照 0.6，0.2，0.2 的比例，按照时间顺序划分为训练集、验证集和测试集。在给定模型和超参数候选集合的前提下，先在训练集上训练模型，在验证集上测试模型。同样根据均方根误差确认最优模型，随后重新在训练集和验证集上进行训练，最后在测试集上检验最优模型的性能。神经网络模型超参数的范围和最优值汇总于表 B3。

表 B3 神经网络模型超参数设定

模型名称	超参数	超参数范围	最优值
LSTM	lookback	[10, 20, 40]	20
	Huidden_size	[16, 32, 64]	32
	Num_layers	[1, 2]	1
	Drop_out	[0.2, 0.4, 0.5]	0.4
	lr	[5×10^{-4}, 1×10^{-4}, 1×10^{-3}]	1×10^{-4}
	Batch_size	[64, 128, 256]	256
Attention-LSTM	lookback	[10, 20, 40]	40
	Huidden_size	[16, 32, 64]	32
	Num_layers	[1, 2]	1
	Drop_out	[0.2, 0.4]	0.2
	lr	[1×10^{-4}, 1×10^{-3}, 5×10^{-3}]	1×10^{-3}
	Batch_size	[64, 128]	64
Transformer	Look_back	[10, 20, 30]	20
	Dim_model	[16, 32]	16
	Nhead	[2, 4]	2
	Num_layers	[1, 2]	1
	Dim_feedforward	[128, 256]	256
	Drop_out	[0.2 ,0.3]	0.2
	lr	[1×10^{-5}, 5×10^{-5}, 1×10^{-4}]	1×10^{-5}

附录 C：模型预测结果

D.1 预测结果展示

在本问中，我们使用了多个机器学习模型对汇率进行预测。各个模型的预测结果经过训练和验证后，样本外的表现效果汇报在表 C1 中，预测图象汇报在图 C1、图 C2 中。从样本外测试来看，各个模型对汇率的预测表现相对较差，其 R^2值普遍为负数，表明模型在训练集和测试集上都未能有效拟合数据。仅有 Ridge 回归在预测汇率时表现出较好的性能，R^2值可达 0.2%。

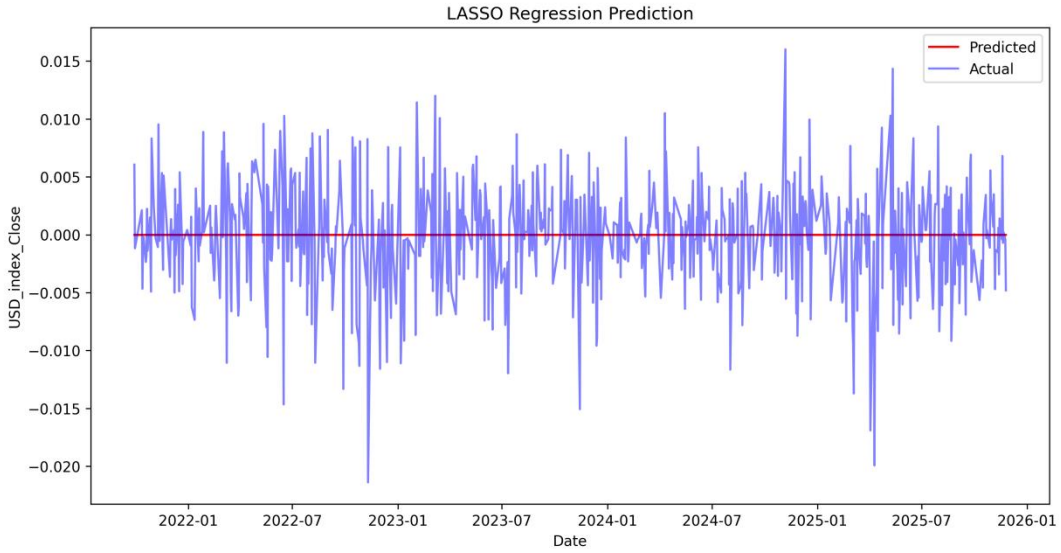

图 C1 LASSO 回归

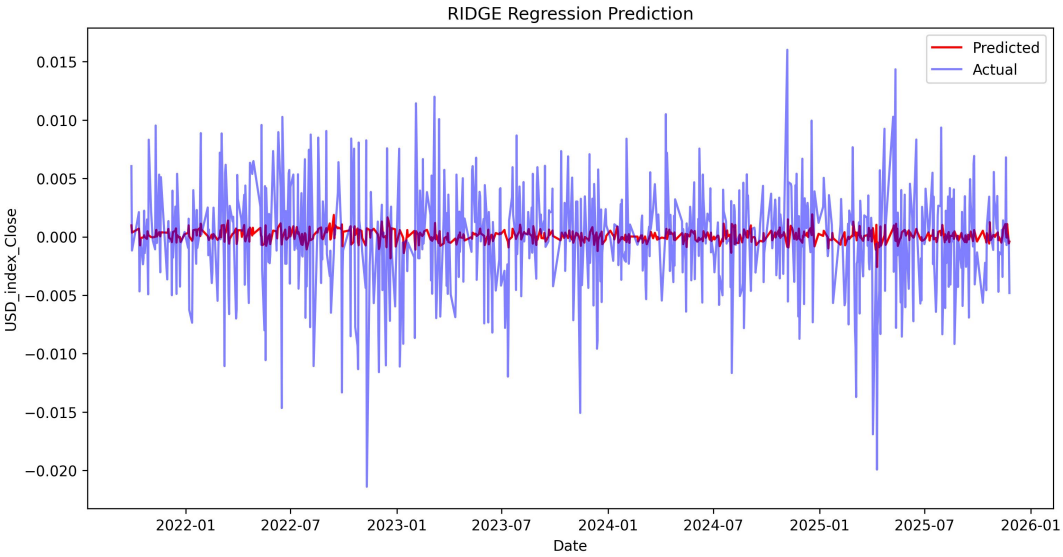

图 C2 Ridge 回归

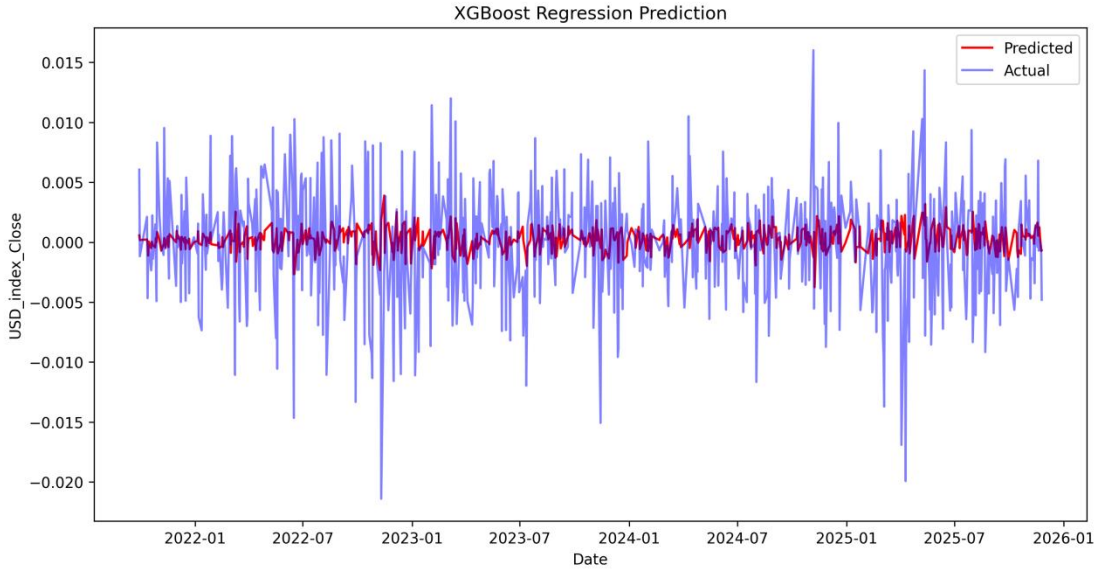

图 C3 XGBoost 回归

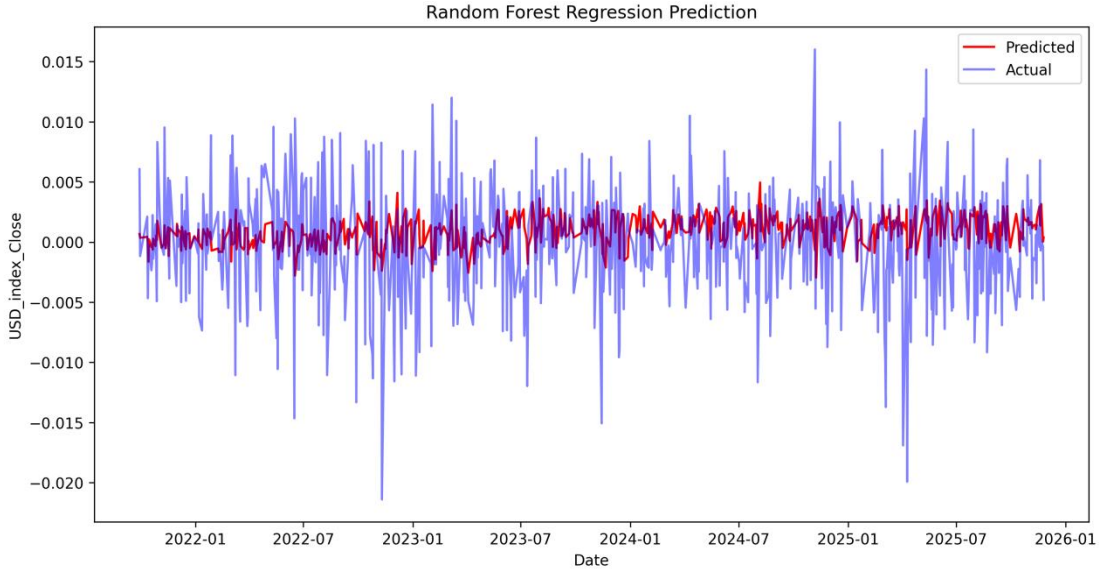

图 C4 RandomForest 回归模型

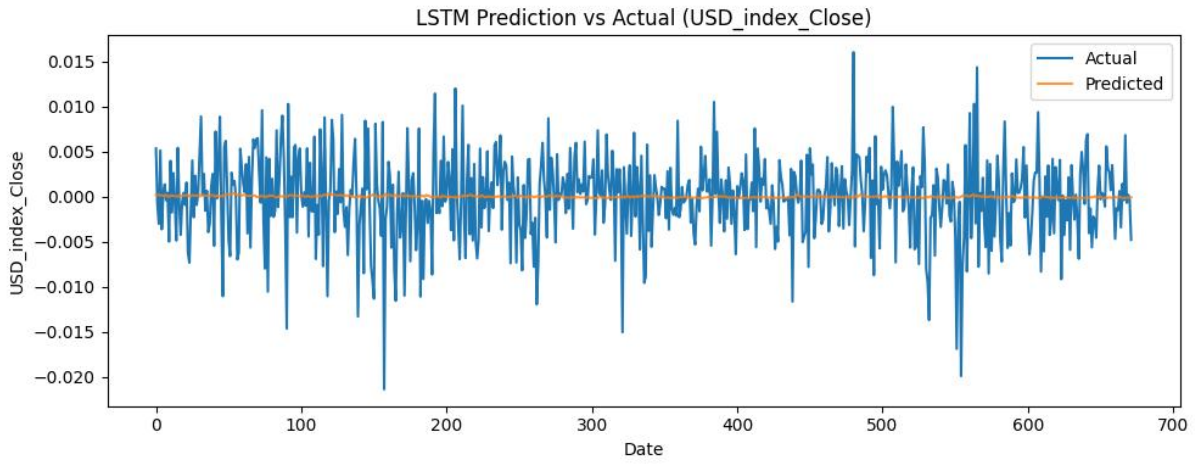

图 C5 LSTM 回归模型

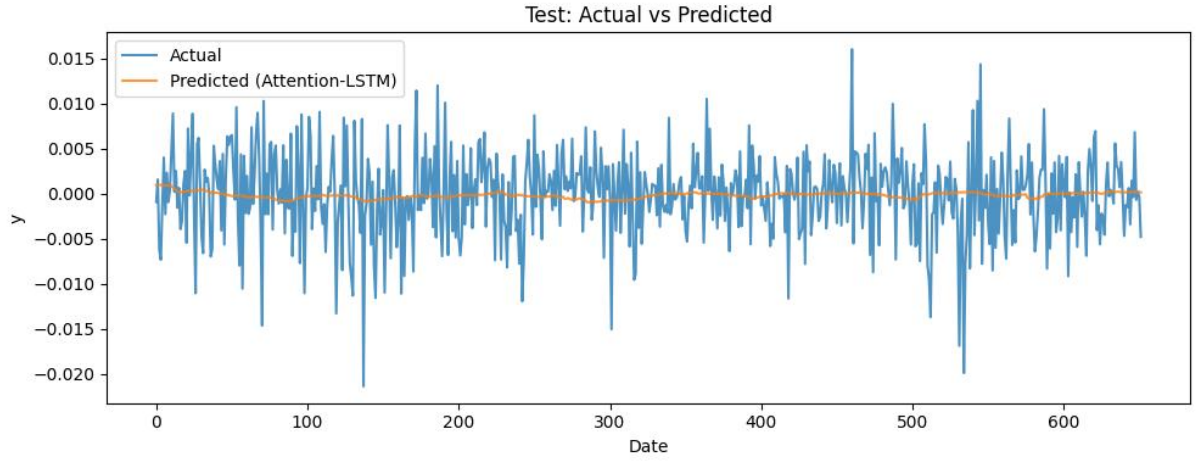

图 C6 Attention-LSTM 模型

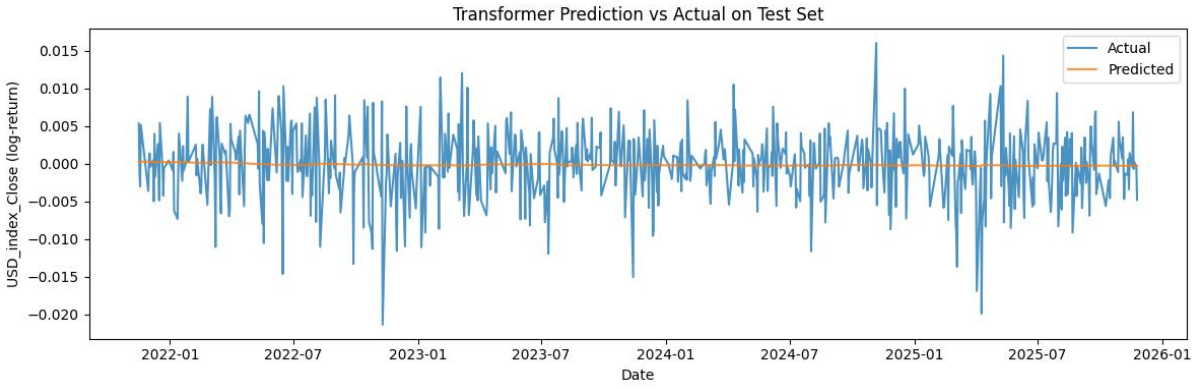

图 C7 Transformer 回归模型