基于 LightGBM 模型的人民币短期汇率决定因素研究

莫云杰

内容摘要：本文从经典汇率决定理论以及国内外对人民币汇率决定因素研究的结论出发，使用 LightGBM 模型研究诸多日频经济变量与人民币汇率之间的关系，并且对模型进行了 K 折交叉验证，训练集和验证集准确率均较高。通过分析因变量在模型中的权重，我们发现利率平价、购买力平价以及其他货币走势构成了人民币短期汇率最主要的决定因素。

关键词：LightGBM　机器学习　人民币汇率　利率平价　购买力平价

一、 引言

2022 年初至今，随着美联储实行美元加息政策，全球货币市场出现巨大波动，人民币对美元汇率呈现大幅度上升趋势。例如 2022 年 8 月 31 日，离岸人民币对美元即期汇率收盘价为 6.9069，离岸人民币相对美元较上月末贬值 2.28%；在岸人民币对美元即期汇率收于 6.8905，较上月末贬值 2.25%。[i]人民币汇率的大幅度波动无疑引发了市场的担忧，其中暗含了人民币汇市风险传导至其他金融领域的悲观预期。由于汇率影响因素众多，且涉及不同金融领域之间的联动，分析汇率波动的因素进而预测汇率波动的研究方法尚未出现，因此值得进一步探究。

图表 1 2019 年至 2022 年 10 月，人民币对美元中间价

为了进一步理解人民币短期汇率波动的主要影响因素，本文使用基于 LightGBM 的决策树模型，通过分析包括美元指数、国债收益差、沪深 300 指数等高频数据，得出不同变量队人民币汇率的影响程度，并且对模型进行了 K 折验证。

二、文献综述

（一） 国内关于人民币汇率的研究

张兵等（2008）[ii]运用格兰杰因果检验和多变量协整检验，发现汇率和股价存在着长期均衡的协整关系，而且股市与汇市存在短期的交互影响。吴丽华和傅广敏（2014）[iii]使用 TV-P-SV-VAR 模型分析 2002 年 1 月至 2014 年 3 月的汇率、短期资本流动、股价的相关月度数据，发现了人民币汇率、短期资本与股价之间的互动关系随时间变化而变化，即在不同时期不同背景下有不同的影响。陈梦根和牛华（2016）[iv]考察了 77 个经济体的购买力平价相关数据，检验了传统的购买力平价理论，发现高收入经济体货币购买力平价年均变化相对较小，低收入特别是下中等收入经济体货币购买力平价年均变化相对较

大。朱孟楠等（2017）[v]同样使用 TV-P-SV-VAR 模型，着重分析 2010 年 7 月至 2016 年 12 月的人民币汇率与房价相关数据，发现人民币汇率预期升值会促进短期国际资本流入，而流入的短期国际资本对房价的影响却与汇率预期的波动强度有关。

（二） 国外关于人民币汇率的研究

Ji et.al(2018)[vi] 分析了 WTI 原油与美国和中国汇率之间的动态依存关系,发现原油对中国和美国汇率市场存在明显的风险溢出，而且溢出效应在中国汇率市场对石油收益率上升和下降的反应是显著不对称的。Wang & Xie(2012)[vii] 研究了人民币与人民币货币篮子中四种主要货币（美元、欧元、日元和韩元）之间的交叉关系，使用与 Ljung-Box 检验相类似的统计检验，人民币与上述货币存在明显的交叉关系，并且用 DCCA 交叉相关系数 ρDCCA 来量化交叉相关的水平，发现人民币货币篮子中的货币权重是按照美元>欧元>日元>韩元的顺序排列。Iqbal et al.（2020）[viii]研究了武汉的天气、COVID-19 疫情和中国经济之间的关系，采用小波变换相干性（WTC）、部分小波相干性（PWC）和多小波相干性（MWC）的方法来分析每日数据，研究发现人民币汇率和 COVID-19 在特定的时间频率点上显示出相位的一致性，表明武汉的 COVID-19 爆发对中国出口经济有负面但有限的影响。

三、模型说明

本文使用基于梯度提升决策树（GBDT，Gradient Boosting Decision Tree）的改进版决策树模型，即 LightGBM（Light Gradient Boosting Machine）。以下是对模型的简单说明。

（一） GBDT 模型简介

梯度提升决策树（GBDT）模型结合了决策树模型与 Boosting 方法，其基本思路是使用 CART 或 C4.5 等决策树作为基函数，通过在信息增益（用梯度表示）最大的节点进行分裂，即增加额外的基函数进行判断，从而达到预测效果提升的目的，体现为损失函数的下降。

对于一个叶节点数为 J 的决策树而言，可以将树模型表示成：

$$tree_m(x) = argmin_{J-node} \sum_{i=1}^{N} |y_i - F_{m-1}(x_i) - tree_{x_i}|$$

因此，可以写出其加总形式：

$$h\left(x; b_j{}_1^J, R_j{}_1^J\right) = \sum_{j=1}^{J} b_j 1(x \in R_j)$$

此处的$R_j{}_1^J$表示共同覆盖了预测变量 x 的所有节点值空间的不相交区域，这些区域由相应树的终端节点表示。其中的参数则是系数$b_j{}_1^J$以及规定 x 界限的R_j，这些变量被成为分裂变量，变量的值表示了在非叶节点上的分裂状况。由于$R_j{}_1^J$表示的空间是不相交的，所以上式也可以表示为，若$x \in R_j$，则有$h(x) = b_j$。

而更新参数的过程，可以表示为

$$F_m(x) = F_{m-1}(x) + \rho_m \sum_{j=1}^{J} b_{jm} 1(x \in R_{jm})$$

，此处的R_{jm}表示第 m 次迭代中的第 j 个叶节点的实数空间，$F_m(x)$表示回归树中基函数的权重，而ρ_m表示换算系数。这个过程可以看作在每一步$1(x \in R_{jm})_1^J$中分别加总 J 个

基函数。因此，通过解出$\left(\rho_m b_{jm}\right)_1^J = argmin_{\left(\rho_m b_{jm}\right)_1^J} \sum_{i=1}^N L\left(y_i, F_{m-1}(x) + \rho_m \sum_{j=1}^J b_{jm} 1\right)$，我们可以得到令损失函数最小化的基函数参数来优化拟合效果。由于$R_j{}_1^J$表示不相交空间，因此对于每一个叶节点而言，最优的参数则是

$$\rho_m b_{jm} = \text{argmin}_{\rho_m b_{jm}} \sum_{x \in R_{jm}} \text{L}\left(y_i, F_{m-1}(x) + \rho_m b_{jm}\right)$$

，基于损失函数 L 和当前的参数估计$F_{m-1}(x)$。对于 LAD（Least Absolute Deviation）情况而言，此时$\text{L}\left(y_i, F_{m-1}(x) + \rho_m b_{jm}\right) = \sum_{i=1}^N |y_i - F_{m-1}(x_i) - \rho_m b_{jm}|$，所以最优参数可以简化为$\rho_m b_{jm} = \text{meadian}_{x \in R_{jm}}\left(y_i - F_{m-1}(x)\right)$，也就是第 m 次迭代时第 J 个叶节点上的残差的中值。[ix]

（二） LightGBM 模型简介

在 GBDT 的基础之上，XGBoost 算法针对稀疏数据和大规模数据进行了优化。[x]但是 XGBoost 算法在特征维度高、数据量大的情况下，其效率和可扩展性仍然存在问题，其中很重要的原因在于，XGBoost 遵循了 GBDT 按叶计算增益(Leaf-wise Learning)的方法，对于每一个叶节点都计算增益，来判断是否需要在该节点进行分裂，从而造成了过大的计算代价。

针对 GBDT 以及 XGBoost 所面临的问题，LightGBM 模型对此进行了改进，主要体现为基于梯度的单边采样（GOSS）以及排他性特征捆绑（EFB）方面的创新。[xi]

（1） 基于梯度的单边采样（GOSS）

GOSS 方法针对的是 GBDT 模型中随机采样的改动，GOSS 方法有限对梯度较大的数据实例进行抽样，这样就使得未受训练的参数更容易收敛至最优值。GOSS 方法将训练实例的梯度降序排列，选取前$\alpha(\alpha \in [0,1))$的实例作为子集 A，在剩余梯度较小的实例随机抽取大小为$b * |A^c|$的样本，最后，根据方差增益的估计值判断是否对一层的节点进行分裂。方差增益的估计值$\tilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{n_r^j(d)}\right)$，其中$A_l = x_i \in A: x_{ij} \leq d, A_r = x_i \in A: x_{ij} > d, B_l = x_i \in B: x_{ij} \leq d, B_r = x_i \in B: x_{ij} > d$。

（2） 排他性特征捆绑（EFB）

高维数据通常是非常稀疏的。特征空间的稀疏性为我们提供了一种设计几乎无损精度的方法来减少特征数量的可能取值。具体来说，在一个稀疏的特征空间中，许多特征是相互排斥的，所以可以安全地将排他性特征捆绑成一个单一的特征（我们称之为排他性特征捆绑）。通过精心设计的特征扫描算法，我们可以从特征束中建立与单个特征相同的特征直方图。这样一来，构建直方图的复杂度就从O（*数据量* × *特征数*）变为O（*数据量* × *束数*），而*束数* ≪ *特征数*。这样 EFB 就可以在不损害精度的情况下大大加快 GBDT 的训练速度。

四、实证研究

（一） 变量选取

本文使用 LightGBM 模型对人民币汇率进行回归，人民币汇率用 2019 年 1 月 1 日至 2022 年 9 月 30 日的每日人民币对美元中间价表示。解释变量主要选择频率较高的每日

数据，用于考察其在短期人民币汇率的决定过程中的作用，解释变量内容广泛，按照性质大致可以分为其他货币关联指标、证券市场指标、商品价格指标、全球市场指标、货币政策指标。

（1） 货币关联指标

货币关联指标主要衡量对于人民币汇率有潜在影响的相关货币汇率，主要选取了美元指数（usd_index）与人民币对 SDR 货币的汇率（CNY/SDR）。

（2） 证券市场指标

证券市场指标此处分为了股票市场指标与债券市场指标，股票市场指标选用了沪深 300 指数（300_index），债券市场指标选用了中美两国一年期国债的收益率之差（interest_diff）。根据凯恩斯（J. M. Keynes）的利率平价理论，证券市场的套利过程对于短期汇率的产生起了决定性作用，因此需要格外关注。

（3） 商品价格指标

根据卡塞尔的购买力平价理论，长期汇率应该由货币购买力的比值决定。因此本文使用了高频的商品价格指数，包括中国大宗商品指数（cci）、欧佩克一揽子油价（opec）。

（4） 全球市场指标

人民币汇率与全球市场的贸易和投资活动密切相关，本文使用 BEISL 全球集装箱运价指数（FBX00）来表示全球经贸的活跃程度。另外，考虑到 2019 年末蔓延的 Covid-19 疫情，本文加入了 Covid-19 病毒的新增确诊人数（covid）作为解释变量。

（5） 货币政策指标

由于人民币实行有管理的浮动汇率制度（dirty-floating），因此中国人民银行所采用的货币政策也会对短期人民币汇率产生重要影响。本文用央行的公开市场操作（operation）来代表央行实行的货币政策。

（二） 特征工程

考虑到汇率市场中存在传导时滞，因此对以上变量构造了 1 到 3 阶的滞后项与差分项。变量大致分布如下：

变量名	有效观测数	均值	标准差.	最小值	最大值
CNYSDR	912	95.47741	3.547224	89.56	102.97
covid	912	189.3103	752.9867	0	15152
cci	912	242.3984	58.44018	168.9585	357.4083
_index	912	4377.228	583.3331	2964.842	5807.719
operation	912	19.73684	140774	-1120000	1160000
usd_index	907	96.75069	4.637345	89.436	114.106
FBX00	903	4358.986	3351.15	1223	11137
opec	912	67.64934	23.36872	12.22	128.27
inter~t_diff	877	1.180849	1.21075	-2.322	2.859
CNYSDR_lag_1	912	95.47055	3.545818	89.56	102.97
CNYSD~1_diff	911	0.001756	0.058398	-0.72	1.07
CNYSDR_lag_2	911	95.47143	3.545403	89.56	102.97
CNYSD~2_diff	909	0.005237	0.219391	-2.05	1.37
CNYSDR_lag_3	910	95.47395	3.54469	89.56	102.97
covid_lag_1	912	187.6875	745.4443	0	15152
covid~1_diff	911	8.054885	584.9902	-10062	13137
covid_lag_2	911	179.8386	706.8681	0	15152

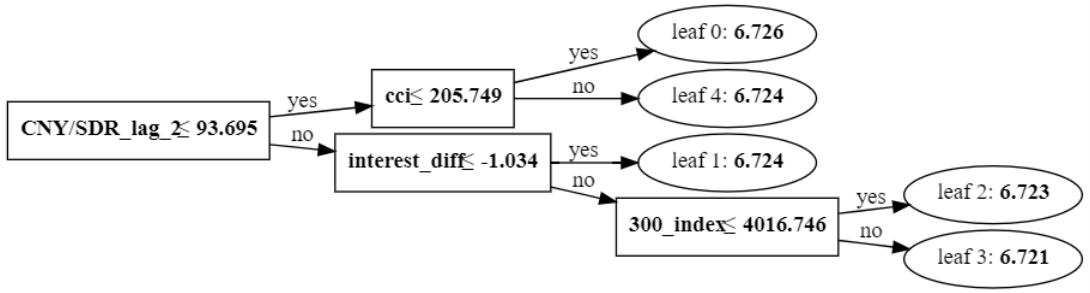
covid~2_diff	909	9.141914	464.9751	-3081	12674
covid_lag_3	910	167.3824	516.2307	0	3896
covid~3_diff	909	-17.6975	495.0318	-13143	3014
cci_lag_1	718	242.4021	58.44025	168.9585	357.4083
cci_lag_1_~f	527	0.072737	2.70999	-12.6249	10.89602
cci_lag_2	528	242.3315	58.41493	168.9585	357.4083
cci_lag_2_~f	164	0.223854	3.52374	-12.481	14.07232
cci_lag_3	512	241.9652	58.08287	169.3566	356.9029
cci_lag_3_~f	322	0.015186	4.836613	-35.1991	17.60193
_index_lag_1	718	4376.841	583.8465	2964.842	5768.381
_inde~1_diff	527	0.378885	49.33783	-228.216	172.27
_index_lag_2	528	4376.158	581.8135	2969.535	5625.923
_inde~2_diff	164	3.388447	63.55257	-164.816	192.2123
_index_lag_3	512	4378.768	579.191	3035.874	5778.842
_inde~3_diff	322	-0.85638	77.80459	-261.99	282.09
operation_~1	912	0	0	0	0
opera~1_diff	911	0	0	0	0
operation_~2	911	0	0	0	0
opera~2_diff	909	2420.242	139346.1	-1160000	1120000
operation_~3	910	-2967.03	137721.7	-1120000	1160000
opera~3_diff	909	-906.491	195567.5	-1160000	1160000
usd_index_~1	731	96.74129	4.641221	89.436	114.106
usd_i~1_diff	545	0.022506	0.405739	-1.698	1.595
usd_index_~2	547	96.72895	4.644369	89.436	114.106
usd_i~2_diff	177	0.032723	0.57925	-1.499	3.091
usd_index_~3	537	96.73056	4.604325	89.436	114.106
usd_i~3_diff	352	0.034151	0.570489	-3.674	3.242
FBX00_lag_1	728	4373.984	3352.004	1223	11137
FBX00~1_diff	544	3.329044	128.825	-1365	1313
FBX00_lag_2	545	4390.077	3357.915	1223	11134
FBX00~2_diff	174	9.770115	138.1081	-1400	679
FBX00_lag_3	533	4340.488	3351.235	1223	11134
FBX00~3_diff	350	10.76286	185.8169	-952	1681
opec_lag_1	912	67.62503	23.37486	12.22	128.27
opec_lag_1~f	911	0.030505	1.561392	-13.63	13.36
opec_lag_2	911	67.61153	23.37379	12.22	128.27
opec_lag_2~f	909	0.104819	1.966123	-13.63	14.78
opec_lag_3	910	67.58103	23.31254	13.3	127.74
opec_lag_3~f	909	0.061309	2.311836	-15.17	14.78
interest_d~1	689	1.179846	1.20961	-2.3092	2.859
inter~1_diff	496	0.000121	0.049169	-0.2464	0.2384
interest_d~2	504	1.183808	1.209781	-2.3092	2.859
inter~2_diff	142	0.000906	0.072768	-0.2441	0.3179

interest_d~3	484	1.194537	1.206094	-2.322	2.859
inter~3_diff	288	0.002758	0.075025	-0.3329	0.3683

图表 2 输入特征概览

（三） 模型训练

本文构造了基于 GOSS 方法的 LightGBM 模型，由于研究任务是回归，故采用平均绝对



图表 3 LightGBM 的决策树

误差（mean absolute error）作为损失函数。关于学习率和迭代次数的参数设定，这里选择了较小的学习率 0.01 配合较大的迭代次数 16000，为了有效防止过拟合问题，额外加入了每个叶节点的最小数据量限制以及最大模型深度限制，除此之外，还在 K 折验证的过程中根据验证集损失函数设置了早停（early stopping），早停的 patience 设置为 100。经过训练之后，我们可以得到如下模型：[1]

训练完模型的同时，我们还进行了 4 折的交叉验证，验证效果如下：

训练集准确率	验证集准确率
0.997413255	0.99576563
0.99741572	0.984065507
0.998417677	0.982836671
0.998535158	0.979850602

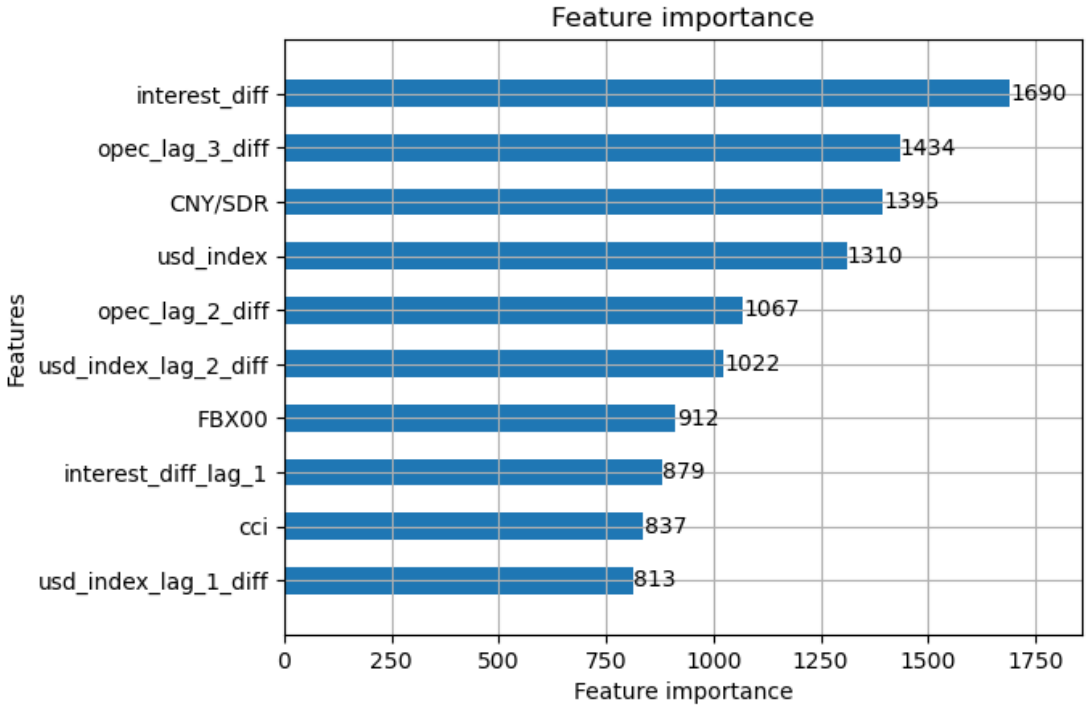
图表 4 K 折验证的准确率结果

其中准确率的计算方法为$Accuracy = \Sigma\left(1 - \frac{|y_{true} - y_{pred}|}{y_{true}}\right) * \left(\frac{y_{true}}{\Sigma y_{true}}\right)$，由此可见，模型的拟合效果极佳，过拟合程度尚可接受。

（四） 结论与建议

有了拟合效果较好的模型，我们可以着手分析通过不同变量的重要性来发掘决定短期人民币汇率的重要因素。

[1] 模型训练代码以及特征构造数据已发布至开源平台 Github



图表 5 重要性最高的 10 个变量

由此可见，在模型中被赋予了最大权重的变量依次是国债利差、OPEC 油价的 3 期差分、人民币对 SDR 汇率、美元指数、OPEC 油价的 2 期差分、美元指数的 2 期差分、BEISL 全球集装箱运价指数、国债利差的 1 阶滞后项、中国大宗商品指数、美元指数的 1 期差分。

从国债收益率差额的重要影响来看，利率平价理论对于短期人民币汇率的决定仍然具有重要指导意义。OPEC 一揽子油价和大宗商品指数对人民币汇率也有显著影响，不过 OPEC 一揽子油价的影响存在 2~3 天的迟滞，因此不能忽略购买力平价对于人民币短期汇率的影响。人民币对 SDR 汇率以及美元指数的影响也各外显著，可见人民币相对其他货币的走势是外汇市场套利的重要参考指标。

---

[i] 温梦瑶，2022：《2022 年 8 月境外人民币市场综述》，《中国货币市场》，第 09 期：第 92－93 页.

[ii] 张兵、封思贤、李心丹、汪慧建，2008：《汇率与股价变动关系_基于汇改后数据的实证研究》，《经济研究》 43，第 09 期 第 70-81+135 页.

[iii] 吴丽华 、 傅广敏，2014：《人民币汇率、短期资本与股价互动》 《经济研究》 第 49 卷，第 11 期 第 72－86 页.

[iv] 陈梦根、 牛华 ，2016： 《购买力平价变动影响因素研究：国际视角》［J］.《金融研究》 ，第 09 期:第 82-98 页.

[v] 朱孟楠、丁冰茜、闫帅，2017：《人民币汇率预期、短期国际资本流动与房价》 ， 《国际金融》，第 7 期 第 17－29 页，https://doi.org/10.13516/j.cnki.wes.2017.07.002.

[vi] Qiang Ji, Bing-Yue Liu, and Ying Fan, "Risk Dependence of CoVaR and Structural Change between Oil Prices and Exchange Rates: A Time-Varying Copula Model," *Energy Economics* 77 (January 2019): 80–92, https://doi.org/10.1016/j.eneco.2018.07.012.

[vii] Gang-Jin Wang and Chi Xie, "Cross-Correlations between Renminbi and Four Major Currencies in the Renminbi Currency Basket," *Physica A: Statistical Mechanics and Its Applications* 392, no. 6 (March 2013): 1418–28, https://doi.org/10.1016/j.physa.2012.11.035.

[viii] Najaf Iqbal et al., "The Nexus between COVID-19, Temperature and Exchange Rate in Wuhan City: New Findings from Partial and Multiple Wavelet Coherence," *Science of The Total Environment* 729 (August 2020): 138916, https://doi.org/10.1016/j.scitotenv.2020.138916.

[ix] Jerome H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine.," *The Annals of Statistics* 29, no. 5 (October 1, 2001), https://doi.org/10.1214/aos/1013203451.

[x] Tianqi Chen and Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco California USA: ACM, 2016), 785–94, https://doi.org/10.1145/2939672.2939785.

[xi] Guolin Ke et al., "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," n.d., 9.