所属类别	2023 年第二届全国大学生数学分析实践赛	参赛编号
专科组		spsspro23123041

基于 BO-LSTM 的黄金股价预测分析

摘要

本文针对黄金股价预测分析的问题，运用了贝叶斯优化算法（BO）、长短期记忆神经网络(LSTM)等方法，构建了基于贝叶斯优化的长短期记忆神经网络模型（BO-LATM）、长短期记忆神经网络（LSTM）及差分自回归移动平均模型（ARIMA），综合运用了 SPSSRPO、MATLAB 等软件编程求解，使用模型对未来四个月的黄金收盘价数据进行了预测，最后对模型的性能进行了评估，并对预测的结果进行了分析。

本文的特色是进行数据可视化，使研究过程、研究内容更加直观、清晰。体现在方法、编程、灵敏度分析等。

针对问题一，要求解决数据探索问题。首先，运用了频数分析、描述性统计和正态性分析，探索了数据集的变量特征。然后，运用了 SPSSPRO 软件进行求解，进行了可视化展示，得出了使用机器学习进行黄金收盘价的预测比较好。

针对问题二，要求解决数据预处理问题。首先，运用了数据标准化、缺失值处理和时序数据滑窗转换。然后，运用了 SPSSPRO 软件求解，将有缺失值的普通时间序列数据转为了无缺失值且具有零均值和单位方差的回归数据。

针对问题三，要求解决模型建立问题。首先，分析了 LSTM 网络的优势，然后，运用深度学习中的 LSTM 网络，构建了基于 LSTM 的黄金股价预测模型并运用 MATLAB 软件编程求解。

针对问题四，要求解决模型调优问题。首先，运用了 BO 算法对问题三建立的模型进行了超参数的调优，然后，得到了基于 BO-LSTM 的黄金股价预测模型，运用了 MATLAB 软件编程求解。

针对问题五，要求解决模型的预测与评估问题。首先，运用了闭环预测方法，使用之前的预测值作为输入来预测序列中的后续时间步，运用了 MATLAB 软件编程求解，得出了未来四个月的黄金收盘价数据，通过将预测出的数据和真实数据比较，计算出了决定系数（R^2）、均方误差（MSE）、均方根误差（RMSE）、归一化得均方根误差（NRME）、误差均值（Error Mean）、误差标准差（Error StD）等评价指标，根据评价指标和拟合图评估出模型的性能很好。

针对问题六，要求解决结果分析问题。首先，通过对模型的预测结果进行了分析，得出了黄金在金融市场的价值较为稳定，但有一定的波动性，受政策的影响较大。

本文最后还对模型进行了误差分析，对模型的优点和缺点进行了客观评价，基于经验模态分解（EMD）对存在的不足进行了改进，对模型数据的准确性进行了灵敏度分析，同时还对模型的使用范围进行了推广。最后，提出了模型的后续优化思路。

关键词:黄金股价预测，贝叶斯优化算法，长短期记忆神经网络，BO-LSTM 模型

一、 问题的重述

1.1 背景知识

1.1.1 引言部分

黄金作为一种特殊的有色金属，不仅可以起到制作珠宝饰品以装饰人的作用，又是一种常用的工业生产原料；更重要的是，它还兼具了商品属性和货币属性。稳定的化学性质和稀少的储藏量使得它受到了极大的追捧。故越来越多的人将其作为一种安全资产用以进行投资。近几年来，受全球局势不稳定和疫情肆虐得影响，黄金市场在金融市场中的地位逐步递增，因此，能否提前了解或预测黄金当日的收盘价成为了每一个投资者的必修课。本文便是在这样的前提下产生的——在知道以前黄金股价相关数据的前提下，建立适合本题的数学模型预测未来黄金的收盘价。

1.1.2 切入标题

黄金的价格也受到许多因素的影响，如全球经济发展形势、各国货币政策、地缘政治等因素。因此，认识并预测黄金价格的变化对于投资决策者来说至关重要。本题的主要目标便是通过题中所给以往的黄金各项数据，预测未来黄金收盘价的趋势，从而帮助投资者在复杂的金融市场环境中做出更加明智的决策。

1.1.3 研究意义

通过研究黄金股价的变化，我们可以较为准确的预测未来黄金的收盘价，从而帮助更多的人进行投资，防止不必要的损失。同时，我们还可以通过这种预测模型和投资策略帮助政府做出正确的经济预判，并助力企业制定更加合理的投融资计划，从而服务于投资者科学管理财富、实现增值获益，营造出更加良好、健康的投资氛围。

1.2 相关数据

1. data.xlsx

1.3 具体问题

问题一：该题要求我们将需要的数据从题中所给数据附件中提取出来，探索出数据集中各个变量的特征，并进行可视化展示。

问题二：该题要求我们将题中所给数据进行清洗、整理等处理，方便之后的使用。

问题三：分析问题二中已经处理过的数据，建立适合本题的数学模型，并阐述选择该模型的理由及过程。

问题四：在问题三已经建立数学模型的基础上，选择合适的方法对其进行调优处理，从而得到性能最佳的模型，并说明在调优方法选择上的理由及调优过程。

问题五：对问题四所建立好的数学模型，对未来四个月的黄金收盘价数据进行预测，并与这四个月的真实数据进行比较，分析并评估该数学模型的性能。

问题六：通过分析模型的预测结果，理解黄金在环境复杂的金融市场中的价值。

二、 问题的分析

2.1 研究现状综述

Persio[1]用两层隐藏层的长短期记忆网络（long short-term memory，LSTM）与循环神经网络（Recurrent Neural Networks, RNN）在预测股票价值预测效果进行对比，取得了良好的结果。Yang 和 Wang[2]使用了 BiLSTM 神经网络对金融时间序列进行预测，利用对沪深 300 指数收盘价进行实证分析对比了 ARIMA、SVR、LSTM 与 BiLSTM 模型在金融时序预测中的性能，结果验证了在预测有噪声的金融时间序列 BiLSTM 模型可以取得良好的预测结果。邓凤欣和王洪良[3]以实例展现了 LSTM 神经网络在通过学习与训练环节下能以较高精度和稳定性拟合股票数据并做出个股价格趋势预测。陈卫华[4]在 5 分钟高频交易的数据基础上，采用 LSTM 神经网络对波动率进行样本外预测，预测精度在不同损失函数下均有不错的提升。王嘉增[5]在 LSTM 神经网络模型从股价预判和构建股票投资组合两方面课题的研究中，其中通过 LSTM 的长期依赖性完成对股票价格的预测，结果表明从预测精度和时效性而言,模型具备可操作性。本文总结以往使用 BO-LSTM（Bayesian optimization，BO)模型进行黄金股价预测分析，并与其它模型对比，证明了 BO-LATM 模型具有较高精度。

2.2 对问题的总体分析和解题思路

黄金股价预测分析其实是一种时间序列预测问题，而这种股票的价格都是没有特殊规律的，一般通过深度学习、回归等方法来进行预测。我们建立了基于 LSTM 的预测模型，然后 LSTM 预测模型的基础上，选择了贝叶斯优化算法对模型进行调优，最终得到了性能最佳的模型。

在进行基本 LSTM 模型训练时，以黄金股价的历年数据作为输入，将当前时刻的黄金股价收盘价作为预测目标。LSTM 模型网络层中超参数的设置对模型的预测性能有很大影响，在预测中由于大部分的超参数需要人工调整，这就需要反复的试验，耗费大量时间才能获取较好的模型。因此，我们提出一种用贝叶斯优化来完成超参数优化的方法、实现自动选择超参数，以提高模型的泛化能力。主要是隐藏层层数、隐藏层神经元个数、LSTM 结构、LSTM 初始化学习率和正则化 L2 系数进行贝叶斯优化。超参数范围设置见表 1。

表 1 LSTM 网络超参数选择表

超参数	范围
双向网络	[True, False]
LSTM 结构	[LSTM, BILSTM](或[1,2])
神经元个数	[50,200]
初始学习率	[0.01,0.1]
正则化 L2 系数	[1e-10,1e-2]

2.3 对所用模型理论的介绍

2.3.1 长短期记忆神经网络（LSTM）

LSTM 是一种特殊的 RNN（循环神经网络）。相比于 RNN，它更适用处理和预测时间序列中间隔较长的数据。

传统的 RNN 结构是由许多重复的神经元构成的“回路”，每个神经元都可以接受输入信息并从而产生输出，然后将输出的结果作为下一个神经元的输入，从而依次传递下去。这种结构在序列数据上存在学习短时依赖关系，但是由于梯度消失和梯度爆炸等问题的影响，RNN 在处理长序列数据时难以发挥很好的性能。为了解决这些问题，我们引进了 LATM 网络。

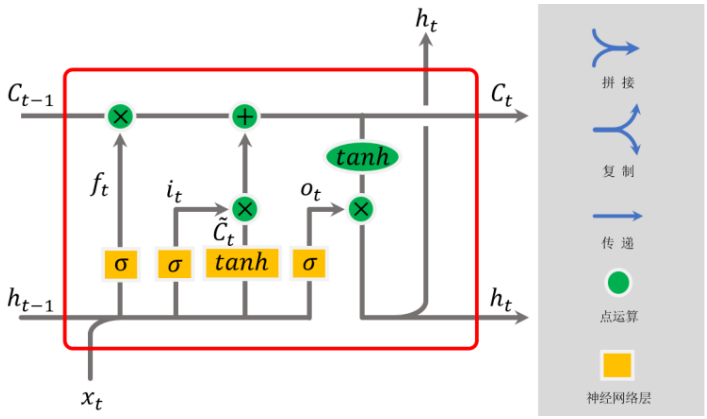


图 1 LSTM 结构图

如上图所示，那条贯穿整个结构的水平线就是神经元状态，也就是 LSTM 网络的核心，此外，LSTM 网络具有遗忘或保留神经元中信息的能力，这项功能被一种由门控单元组成的的结构管理。其中，门控单元是由一个 Sigmoid 网络层和逐点相乘器所组成的。Sigmoid 的取值范围是 0-1 之间的实数，它反映了一个神经元能够让多少信息通过——0 代表着全部信息都不能通过，1 则代表所有信息均能通过。一个 LSTM 网络是由三个这样的门控元所组成的，并借此对神经元状态的进行保护和控制。

LSTM 神经网络的第一个门控单元是根据输入数据来决定神经元应该遗忘哪些信息。故被称为“遗忘门”，它是由 Sigmoid 层所组成的。通过输入 t −1 时刻的隐藏态信息 h_{t-1} 和 x_t，可以输出 0-1 之间的数字，该数字就表示对于 C_{t-1} 的神经元所包含的信息量。计算公式如下所示：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

下一个门控单元是一个关系到决定后续要更新的数值，故被称为“输入门控”的 Sigmoid 层。数值经过 Tanh 层会生成一个新的候选数值 C_t，它会被增加到神经元状态中。接下来代入以下公式更新状态值。其计算公式如下所示：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

最后一个门控单元是需要更新旧的神经元状态C_{t-1}更新到新的神经元状态C_t。接下来需要给旧的状态乘以一个遗忘系数f_t，然后在此基础上，本研究增加$i_t \times \tilde{C}_t$，这就是新的神经元状态C_t。

最后需要控制输出结果及内容。此处的输出是建立在神经元状态C_t的基础上的，先用 Sigmoid 层计算输出因子O_t决定哪一部分的神经元状态被输出；接着让神经元状态C_t经过 tanh 层并且乘上 Sigmoid 门控的输出O_t，最终可以获得第t时刻的输出h_t。其计算公式如下所示：

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \times tanh(C_t)$$

2.3.2 贝叶斯优化算法（BO）[6]

贝叶斯优化是一种全局优化算法，利用较少的迭代次数和已知数据来获得最优解，现在多用于调节机器学习算法的超参数。贝叶斯优化算法的思路是利用目标函数的先验概率分布及已知观测点来更新后验概率分布，然后根据后验概率分布寻找下一个极小值点，使极小值不断减小，最终得到最优超参数。

观察数据=先验分布-后验数据

根据贝叶斯定理：

$$p(f \mid D) = \frac{p(D \mid f)p(f)}{p(D)}$$

式子中，f为未知的目标函数；D为已观测到的参数和观测值集合（$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$）；$x_i$为观测到的数据；$y_i = f(x_i) + \delta_i$为观测值；$p(f \mid D)$为$f$的后验概率；$p(f)$为$f$的先验概率，$p(D \mid f)$为y的似然分布；$p(D)$为$f$的边际似然分布。

概率代理模型是用来代理未知目标函数（即不同超参数下的 LSTM 模型预测误差）的概率模型，随着贝叶斯优化的迭代计算，不断修正先验概率，代理模型将会更接近未知目标函数，使代理模型更接近位置目标函数。我们选用高斯过程（Gaussian Processes，GPs）作为概率代理模型，这种模型已经广泛运用在非线性回归、分类以及许多需要推断黑箱函数的领域中[7]。高斯分布表示为

$$f(x) \sim GP(\mu(x), k(x, x))$$

式中，$\mu(x)=E\left(f(x)\right)$；$E\left(f(x)\right)$为$f(x)$的数学期望，通常均值函数设置为 0；$f(x)$为平均绝对误差；$k(x,x)$ x的协方差函数。

通过假设待优化 LSTM 模型超参数的集合符合高斯分布，接着通过观测到的超参数数据点去预测下一个更优的超参数点。

三、 模型的假设

1.假设待优化 LSTM 模型超参数的任意线性组合均符合高斯分布，即可通过已观测的超参数点去预测下一个最有“潜力”的超参数点[8]。

2.假设测试集上未来四个月为 2023 年 3 月 4 日到 2023 年 7 月 4 日，共 122 天。

3.假设预测测试集中未来四个月黄金收盘价时并不知道未来四个月的开盘价、最高价、最低价、变化（点数）和变化（百分比）。

4.假设每个星期的星期六股市都不开盘，即所有星期六都没有数据。

5.假设“data.xlsx”中的数据来源准确、可信、稳定、科学。

四、 名词解释与符号说明

4.1 名词解释

长短期记忆神经网络：是一种特殊的循环神经网络，相比于传统的循环神经网络，它更适用处理和预测时间序列中间隔较长的数据[9]。

贝叶斯优化：是一种用于求解函数最优值的算法，它普遍使用于在机器学习过程中对超参数进行调优。

时间序列：指将同一统计指标的数值按其发生的时间顺序排列而成的序列，其主要目的是根据已有的历史数据对未来数据进行预测。

循环神经网络：是一类以序列数据为输入，在序列的演进方向进行递归的递归神经网络。

超参数：运用模型前事先定义的，不随运行过程的改变而改变，定义模型属性或者定义训练过程的参数，称为超参数。

机器学习：它是一门多领域交叉的学科，涉及概率论、统计学、逼近论等多门学科，主要是通过让计算机学习模拟或实现人类的学习行为，从而获得新的知识或技能，并在已有的知识结构不断改善自身性能的一种方式。常见的方法主要有聚类、人工神经网络、贝叶斯学习等。

开盘价：即开市价，是指某种证券在证券交易所每个交易日当天开市后的第一笔每股买卖成交价格。

最高价：即在证券交易所交易日当天某种证券所成交的价格中的最高价位。

最低价：即在证券交易所交易日当天个股成交的最低价位。

收盘价：即证券交易日当天某种证券最后一笔交易前一分钟所有交易的成交量的加权平均价（包含最后一笔交易）；若当日没有成交的，以之前的收盘价为当日收盘价。

变化（点数）：即股票价格指数，它反映了在一定的时期内，股票市场总体价格变动

情况的指标。一般是通过对市场上一定数量的代表性股票价格进行加权平均计算得出，主要用来衡量整个股票市场的价格水平及其波动情况。

变化（百分比）：是股票市场对于某种股票的涨幅所做出的表述。它通常用来描述股票价格的变化情况。

模型参数：它是机器学习模型的基础，不仅可以人为设置，也可以由学习算法决定。同时，它也是机器学习模型的基本单位，通过许多参数一起工作，可以为模型控制输出结果。主要用于调整模型的准确性和精确度。

4.2 符号说明

序号	符号	符号说明
1	f	input 输入对应的函数
2	D	表示一个由若干对数据组成的数据集
3	x	一组超参数
4	y	表示一组超参数对应的结果，一组对应一个结果
5	$p(f \mid D)$	f 的后验概率
6	$p(f)$	f 的先验概率
7	$f(D)$	f 的边界似然分布
8	$p(D \mid f)$	y 的似然分布
9	h_{t-1}、x_t	“遗传门”t-1 时刻的隐藏态信息
10	C_{t-1}	旧神经元状态
11	C_t	“输入门控”层产生的候选数值
12	C_t	新神经元状态
13	O_t	Sigmoid 层产生的输出因子
14	h_t	LSTM 最终输出结果
15	f_t	遗忘系数

五、 模型的建立与求解

5.1 问题一的分析与求解

5.1.1 对问题的分析

数据源为 2021 年 7 月 4 日至 2023 年 7 月 4 日的黄金股价数据集，该数据集包括每天的日期、开盘价、最高价、最低价、收盘价、变化（点数）和变化（百分比）。描述数据集的变量特征，即对数据集进行描述性分析，包括频数分析、描述性统计和正态性分析。

5.1.2 对问题的求解

1. 频数分析

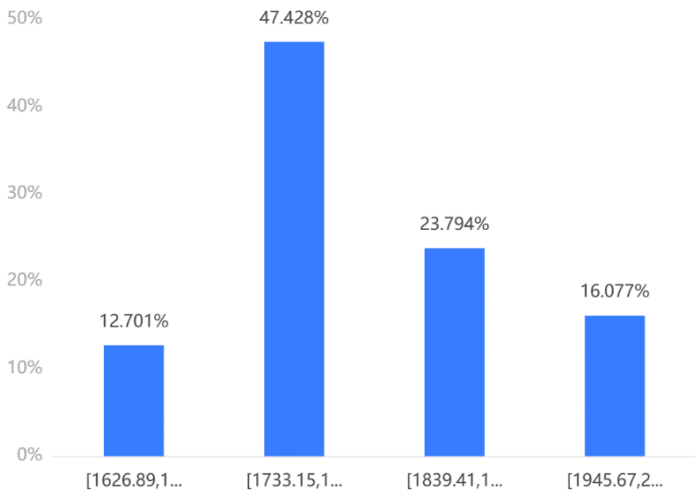


图 2 收盘价频数百分比柱形图

注：其余变量的频数百分比柱形图在附录中展示。

从图中可以看出，收盘价在[1733.15,1839.41)区间的日期最多，频数达到了 295，所占百分比 47.428%。

2. 描述性统计

表 2 变量总体描述结果

变量名	样本量	最大值	最小值	平均值	标准差
Close	622	2051.93	1626.89	1833.593	94.346

续表

中位数	方差	峰度	偏度	变异系数（CV）
1815.47	8901.146	-0.563	0.147	0.051

注：其余变量的总体描述结果在附录中展示。

从表中可以看到，Close 变异系数（CV）为 0.051< 0.15，当前数据中较小概率出现异常值。

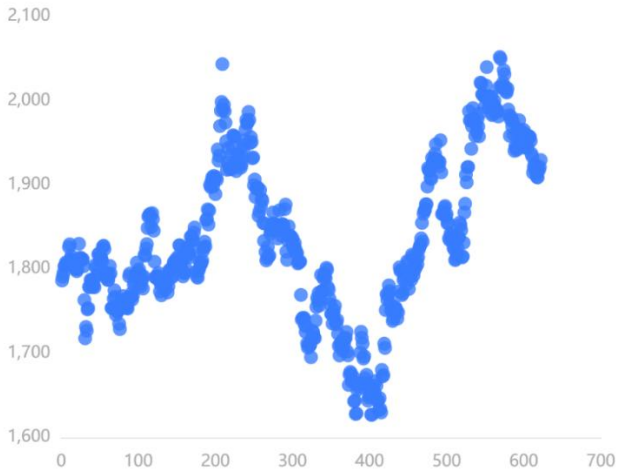


图 3 Close 散点图

注：其余变量的散点图在附录中展示。

集中趋势分析的结果从散点图中可以看出，Close 数据并不集中。

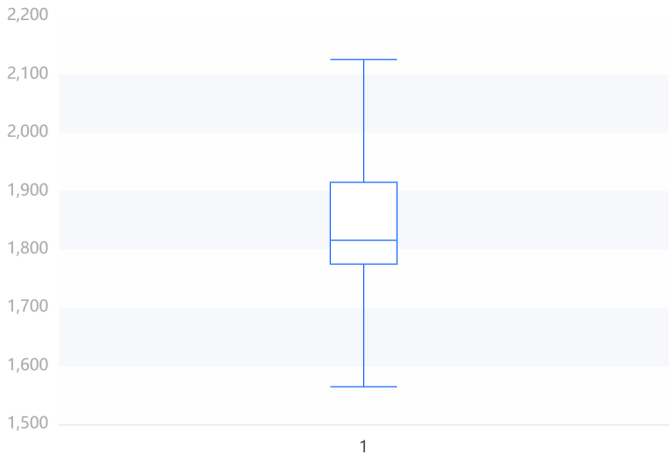


图 4 Close 箱型图

注：其余变量的箱型图在附录中展示。

箱线图的内限为极大值、极小值，超出箱线图的内限的点就是异常点，但是该极大值、极小值不是数据的最大值最小值。用极大值、极小值、中位数、25%分位数等统计指标对数据分布进行差异测量即是 Close 的离散趋势。

3. 正态性检验

表 3 正态性检验表

变量名	S-W 检验	K-S 检验
Close	0.977(0.000***)	0.082(0.000***)

Close 数据量 N<5000，使用 S-W 检验，显著性 P 值为 0.000***，水平呈现显著性，拒绝原假设，Close 数据不满足正态分布。

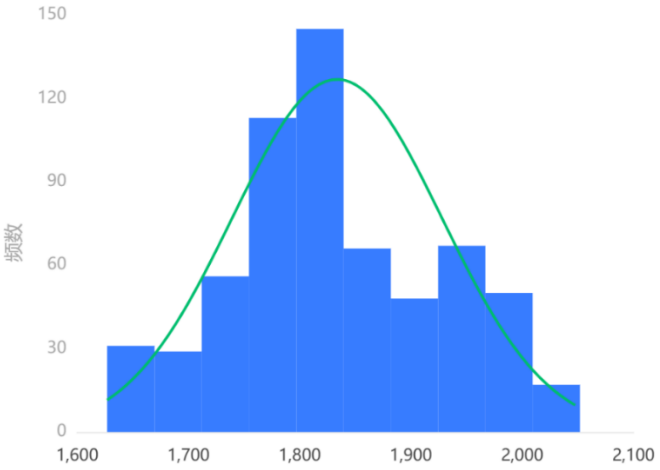


图 5 Close 正态性检验直方图

Close 正态性检验直方图基本上呈现出钟形（中间高，两端低），说明数据虽然不是绝对正态，但基本可接受为正态分布。

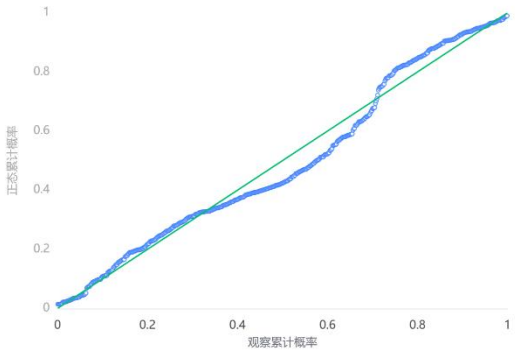


图 6 Close 正态性检验 P-P 图

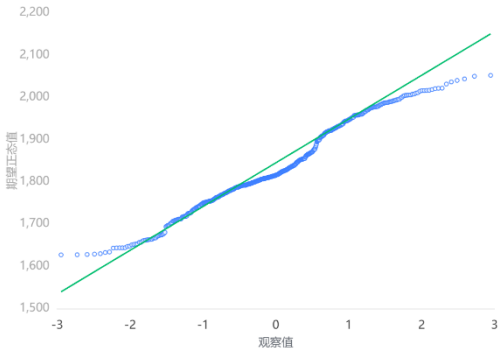


图 7 Close 正态性检验 Q-Q 图

Close 正态性检验 P-P 图表示累计概率（P）与正态累计概率（P）拟合程度较差，不太服从正态分布。

Q-Q 图用来比较观测值与预测值（假定正态下的分布）不同分位数的概率分布，从而检验是否吻合正态分布规律。同时将实际数据作为 X 轴，将假定正态时的数据分位数作为 Y 轴，作散点图，散点与直线重合度不高，说明 Close 不太符合正态分布。

5.2 问题二的分析与求解

5.2.1 对问题的分析

问题二是数据预处理。数据预处理一般采用的方法为数据标准化、异常值处理、异常值处理、缺失值处理等。通过分析附件数据，可以知道缺失了所有星期六的数据和2021-11-21，2021-12-05，2021-12-12，2022-01-16，2023-02-12 这 5 天的数据。后续模型建立时计划采用机器学习的方法，需要将时间序列数据转为回归数据。

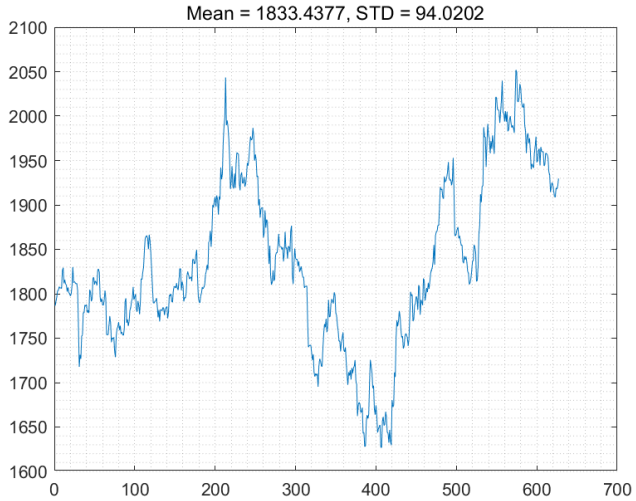


图 8 原始数据序列图

对于本题的数据，应该采用数据标准化、缺失值处理和时序数据滑窗转换。

5.2.2 对问题的求解

星期六股票不开盘，这是之前在假设中提到的，所以不考虑，直接删除。2021-11-21，2021-12-05，2021-12-12，2022-01-16，2023-02-12 这 5 天缺失的数据，股价易受其滞后性影响，因此缺失值用前一天的价格进行填充。

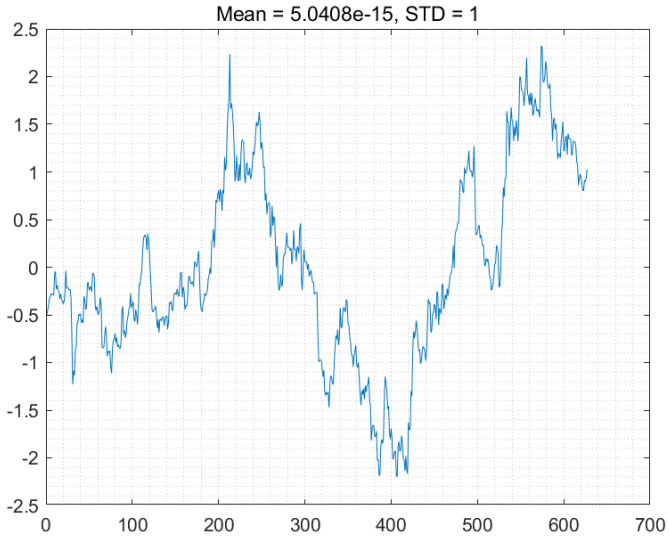


图 9 标准化数据序列图

数据标准化是将数据标准化为具有零均值和单位方差，目的是为了防止训练发散和

获得较好的拟合效果。数据标准化采用 z-score 标准化：

$$x' = \frac{x - \mu}{\sigma}$$

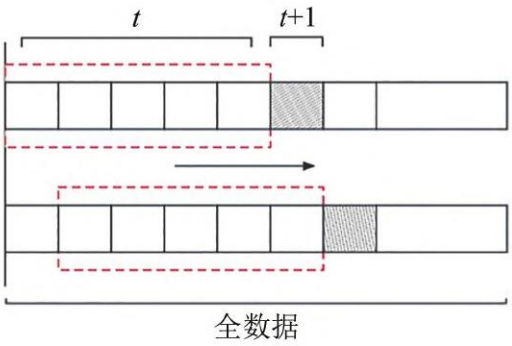


图 10 滑窗结构图

时序数据滑窗转换就是将收盘价这一序列数据转换为 $X \rightarrow Y$ 的回归数据，步阶为 30，代表 30 个 X，就是用第 1-30 天的数据预测第 31 天，用 2-31 天的数据预测第 32 天，以此类推[10]。

5.3 问题三的分析与求解

5.3.1 对问题的分析

因为预测测试集中未来四个月黄金收盘价时并不知道未来四个月的开盘价、最高价、最低价、变化（点数）和变化（百分比），这是在上面假设中所提到的，而且，收盘价历史数据没有明显的特征，所以，采用深度学习进行收盘价的时序预测。LSTM 网络是一种循环神经网络 (RNN)，可以学习序列数据的时间步之间的长期依存关系，它通过遍历时间步并更新网络状态来处理输入数据，网络状态包含在所有先前时间步中记住的信息。LSTM 网络可以避免模型发生梯度消失和梯度爆炸，而且可以同时考虑长期和短期信息。

黄金股票价格容易受到过去价格变化的影响，LSTM 能够记录过去的数据信息并用来对当前黄金股票价格进行预测。黄金股票价格有长期趋势和短期波动，LSTM 可以通过控制输入门、输出门、遗忘门和输入门来控制长期和短期信息的重要性，更好地处理黄金股票数据。所以，选择 LSTM 预测黄金股票收盘价。

5.3.2 对问题的求解

基于 LSTM 的黄金股价预测模型

首先，需要定义 LSTM 网络架构，包括创建 LSTM 回归网络和指定训练选项。将 LSTM 层的隐含单元个数设置为 200 个，求解器设置为“adam”，将训练轮数设置为 250 轮。将梯度阈值设置为 1，目的是防止梯度爆炸。将初始学习率设置为 0.005，设置因子 0.2，在 125 轮训练后通过乘以因子来降低学习率。

接着，训练 LSTM 网络，使用 MATLAB 中的 trainNetwork 函数根据前面指定的训练选项训练 LSTM 网络。

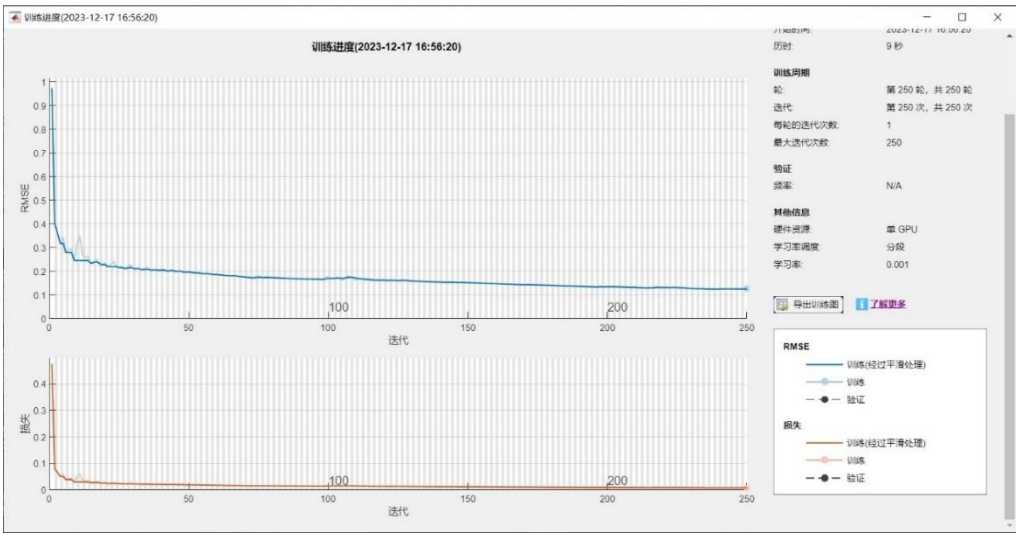


图 11 LSTM 网络训练图

然后，使用训练好的 LSTM 网络预测测试集中未来四个月的收盘价。在每次预测时先对训练数据进行预测，进行网络状态的初始化。接下来，使用训练响应的最后一个时间步进行第一次预测，循环其余预测并将前面一次预测出来的收盘价作为函数 predictAndUpdateState 的输入。对于大型数据集合、长序列或大型网络，在 GPU 上进行预测计算通常比在 CPU 上快，对于单时间步预测，使用 CPU 进行预测计算更快。将 MATLAB 中 predictAndUpdateState 函数的执行环境设置为“cpu”，使用 CPU 进行预测运算。

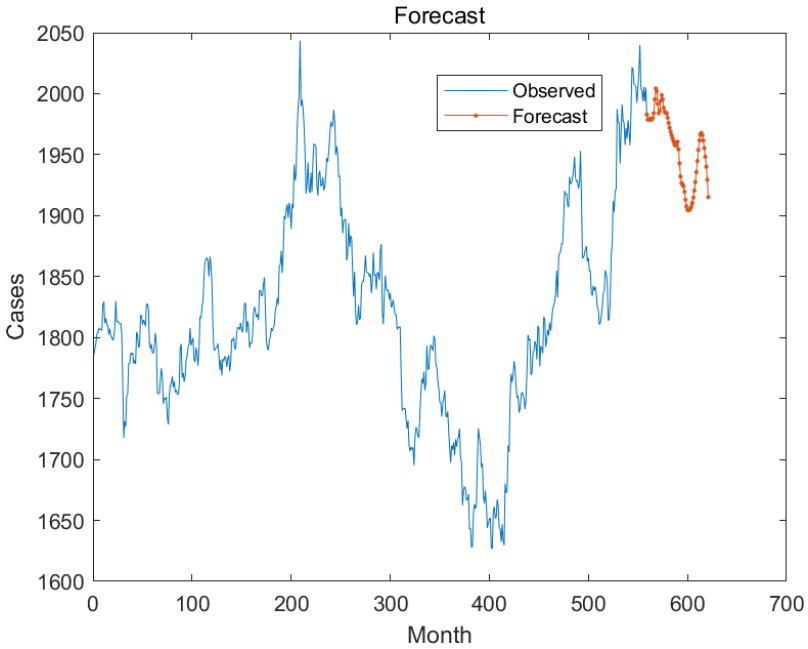


图 12 预测结果图

最后，将预测出来的数据与测试集进行比较。使用之前进行数据标准化的参数对预测出来的数据进行去标准化操作，根据去标准化后的预测出来的数据计算均方根误差（RMSE）。使用预测出来的数据绘制时序图与测试集中的数据进行比较。

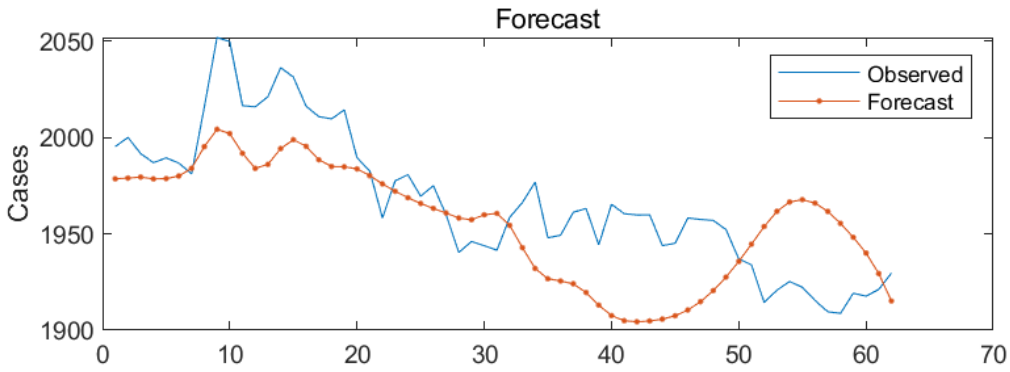


图 13 预测值和真实值比较

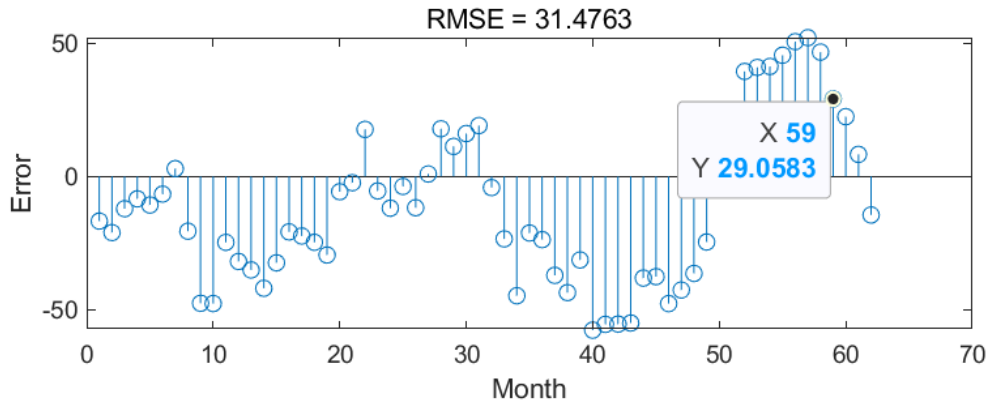


图 14 均方根误差

5.4 问题四的分析与求解

5.4.1 对问题的分析

使用上面建立的基于 LSTM 的黄金股票收盘价预测模型进行预测时，对模型的性能影响较大的因素是超参数，超参数一般是由认为调节的，但人为调节超参数一般无法调节到最优超参数点。

为了避免人为调参的缺点，利用贝叶斯（Bayesian）算法对 LSTM 模型的超参数进行优化选择，包括采取的隐含层层数、隐含层神经元数、正则化 L2 系数、初始学习率，这样可以使模型的性能得到提高。贝叶斯优化算法主要包含概率代理模型和采集函数两个部分，目标函数的后验概率分布是根据概率代理模型利用高斯过程寻找目标函数的近似，采集函数的作用是为目标函数取得极值，根据后验概率分布在最优的区域选择好的样本点。

5.4.2 对问题的求解

基于 BO-LSTM 的黄金股价预测模型

该模型分为三个阶段，第一阶段训练 LSTM 模型，第二阶段将需要进行优化的超参数传入贝叶斯优化模型，第三阶段利用优化后的 LSTM 模型输出预测结果。

首先，训练 LSTM 模型，在第三问建模的基础上，需要用到双向 LSTM（BILSTM），设置 BILSTM 的参数，为了避免过拟合，创建一个 dropout 层，设置概率为 0.5，进行

BILSTM 神经网络的训练，之后不断改变超参数，优化模型，配合第三问建立的模型就可开始后续的优化步骤。

接着，设置需要优化的超参数，并设置参数空间。LSTM 隐含层层数(1-4)，数据类型为整数；LSTM 隐含层神经元(50-200)，数据类型为整数；LSTM 结构，1 代表 LSTM，2 代表 BILSTM，数据类型为整数；LSTM 初始学习率(0.01-1)，数据类型为浮点型；LSTM 正则化 L2 系数(1e-10-1e-2)，数据类型为浮点型。

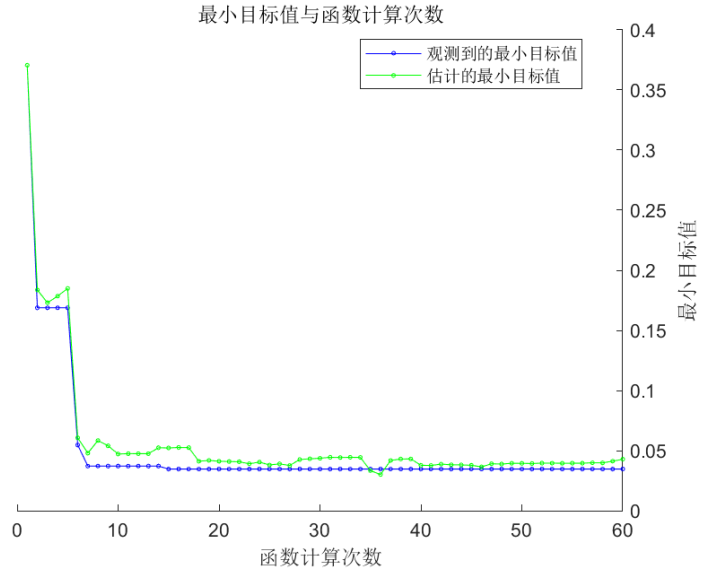


图 15 最小目标值与函数计算次数

然后，使用贝叶斯优化找到最佳 LSTM 参数，设置贝叶斯优化运行的最大时间为 14*60*60，最大迭代次数为 60，利用高斯过程求目标函数的后验概率分布，根据后验概率分布对超参数样本点进行采样，优先选择最优超参数，实现对超参数的更新。之后不断更新目标函数的后验概率分布和超参数，直至满足模型的要求。展示优化过程日志，将最佳优化输出保存为 mat 文件，这个 mat 文件是输入特征为 30，输出特征为 1 的贝叶斯算法优化后的 LSTM 网络。

图 16 优化结果

最后，使用建立好的网络结构，对测试集数据进行预测，误差为预测测试集输出数据与实际测试集数据的均方根误差。

5.5 问题五的分析与求解

5.5.1 对问题的分析

在测试集上，使用基于 BO-LSTM 的黄金股价预测模型对未来四个月即 2023 年 3 月 4 日到 2023 年 7 月 4 日，共 122 天的黄金收盘价数据进行预测，用测试集的真实值与模型预测的输出值进行比较，通过结果对模型的性能进行评估。

5.5.2 对问题的求解

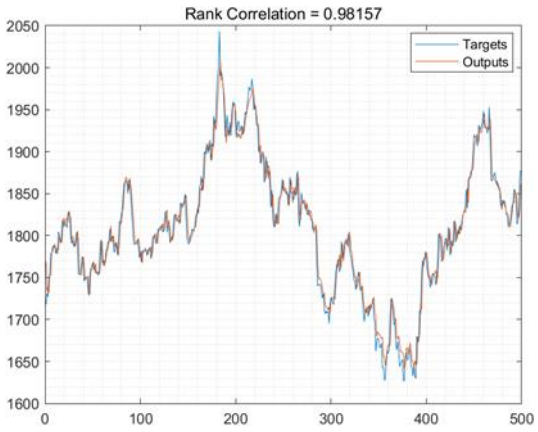


图 17 训练集预测结果

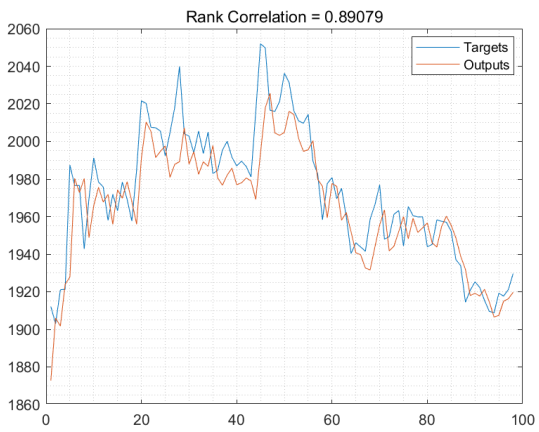


图 18 测试集预测结果

根据预测结果图可以看出，在训练集上，Rank Correlation=0.98157，预测值和真实值几乎完全重合，预测的准确度很高，在测试集上，Rank Correlation=0.89079，预测值和真实值有所差距，但趋势正确，总体来说预测结果较好。

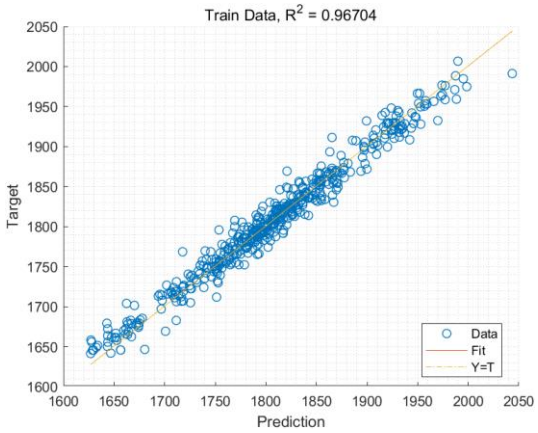


图 19 R^2 训练集线性拟合效果

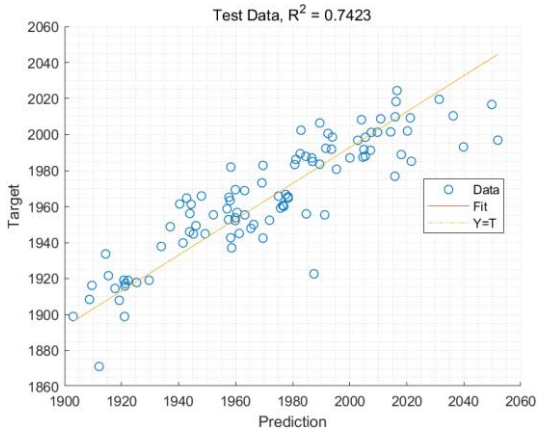


图 20 R^2 测试集线性拟合效果

R^2 是变量 x 引起变量 y 变异的回归平方和占 y 变异总平方和的比值，也称为“拟合

优度”，计算公式为：

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

其中，决定系数反应了变量 y 的波动有多少百分比可以被 x 的波动所描述。

在训练集上，根据 R^2 线性拟合效果图可知数据预测值和实际值的相似程度非常高，在测试集上，根据 R^2 线性拟合效果图可知数据预测值和实际值的相似程度较高，对于长期预测来说，效果已经非常好了。

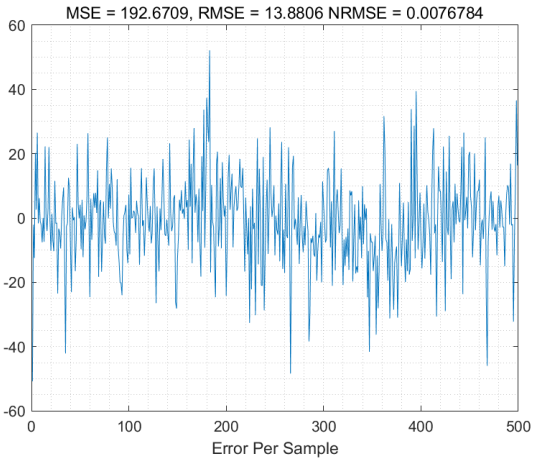


图 21 训练集拟合误差

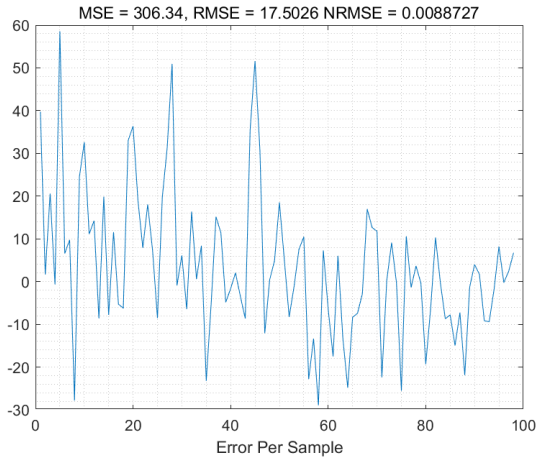


图 22 测试集拟合误差

MSE 用于表示参数估计值与参数真值之差平方的期望值，计算公式为：

$$\text{MSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\hat{y}_i)^2}$$

在训练集上，MSE=192.6709 说明预测值相对于真实值的误差较小，预测效果很好。在测试集上，MSE=306.34 说明预测值相对于真实值的误差不大，预测效果较好。

RMSE 预测值与真实值的偏差平方和同观测次数 m 比值的平方根，计算公式为：

$$RMSE(X,h)=\sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h(x^{(i)})-y^{(i)}\right)^2}$$

在训练集上，RMSE=13.8806 说明预测值与真实值之间的偏差较小，预测效果很好。在测试集上，RMSE=17.5026 说明预测值与真实值的偏差不大，预测效果较好。

NRMSE 通常使用 RMSE 值除以真值的均值，计算公式为：

$$\text{NRMSD}=\frac{\text{RMSD}}{y_{\max}-y_{\min}} \text{ 或 } \text{NRMSD} = \frac{\text{RMSD}}{\bar{y}}$$

在训练集上，NRMSE=0.0076784，在测试集上，NRMSE=0.0088727，说明训练集的 NRMSE 为较小的残差。

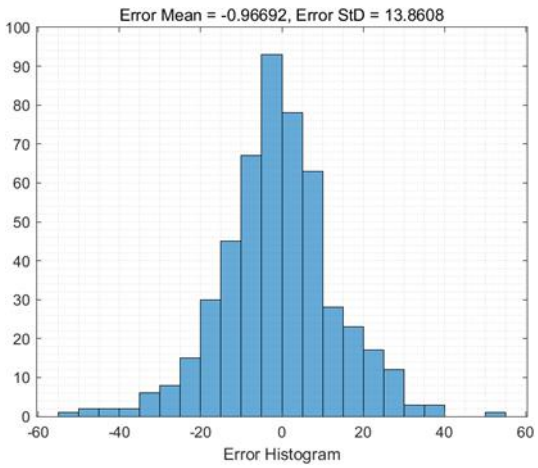


图 23 训练集误差直方图

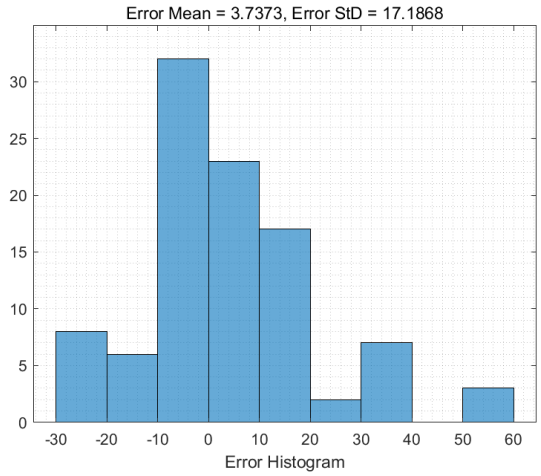


图 24 测试集误差直方图

Error 是误差，Mean 是均值，若给定一组包含 n 个样本的数据集合 $X = \{X_1, \cdots, X_n\}$ 时，将均值定义为这个集合中所有元素和的平均值，从而用于反映一个现象总体的一般水平，计算公式为：

$$\mu = \frac{1}{N}\sum_{i=1}^{N} A_i$$

在训练集上，Error Mean=-0.96692 说明训练集的平均误差较小，预测结果很好，在测试集上，Error Mean=3.7373 说明测试集的平均误差不大，预测结果较好。

StD 是“方差”的算术平方根，即样本总体各单位标志值与其算术平均数之间的平均离差，又称为均方差，计算公式为：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

在训练集上，Error StD=13.8608 说明训练集中个体的离散程度很小，即误差波动很小，预测结果很好，在测试集上，Error StD=17.1868 说明测试集个体的离散程度较小，误差波动较小，预测结果较好。

5.6 问题六的分析与求解

5.6.1 对问题的分析

对模型的预测结果进行分析，得出预测结果随时间的大体走势与真实值相近，可如果仅仅只看某日的收盘价后，可以知道这两个数值是有些许差异的。

5.6.2 对问题的求解

黄金一直以来都是金融市场中备受关注的一种资产，尽管如今各国的货币体系不再直接与黄金挂钩，但它独特的属性和价值依旧在投资、避险和资产保值方面扮演着重要的角色。特别是在如今全球经济不稳定和疫情蔓延期间，因为它的价值在经济动荡时相

对稳定，故被越来越多的人视为一种避险资产，投资者们通过购买黄金来对冲其他高风险资产的价值下跌。

黄金市场作为一个活跃的市场，投资者可以在这里进行买卖，不仅可以选择长期持有黄金并借此来稳健地积累财富，也可以用它进行短期交易从而赚取市场波动的利润。因此，对于投资者来说，能否提前准确的预测交易日当天黄金的收盘价变得极为重要。

这表明黄金的确可以作为一种安全资产被投资者所投资，不过，收盘价的走势也让我们知道黄金也受许多因素影响，诸如国家政策、国际局势、疫情蔓延等等。

六、 误差分析与灵敏度分析

6.1 误差分析

1.因为没有所有星期六的数据，可能因为数据缺失这一部分而造成预测的精度不高；

2.“data.xlsx”是相关部门统计的数据，统计过程中可能有误，造成预测模型的精度下降；

3.没有考虑到突发事件的影响，如疫情、政策等，模型对与突发事件影响下的预测精度不高；

4.黄金股价数据本身没有什么规律性，通过模型预测未来四个月的收盘价只是根据历史数据做出的统计预测，不能保证结果一定准确；

5.数据总共只有 622 组，数据量太少导致预测模型的精度不高。

6.2 灵敏度分析

问题三在设计具有动态监测功能的干预方案时，基于贝叶斯优化的 LSTM 神经网络模型中使用的数据是相关部门的统计数据，统计过程中可能有错误。因此，我们需要考虑数据不准确对模型解的影响，即逻辑权重的敏感性分析。X1、X2、X3、X4、X5、X6 对应权重的灵敏度按定义记为 S(X1,W1)、S(X2,W2)、S(X3,W3)、S(X4,W4)、S(X5,W5)、S (X6,W6)，从而得到：

$$\begin{Bmatrix} S(X_1,W_1) \\ S(X_2,W_2) \\ S(X_3,W_3) \\ S(X_4,W_4) \\ S(X_5,W_5) \\ S(X_6,W_6) \end{Bmatrix} = \begin{Bmatrix} \frac{dX_1}{dW_1} * \frac{W_1}{X_1} \\ \frac{dX_2}{dW_2} * \frac{W_2}{X_2} \\ \frac{dX_3}{dW_3} * \frac{W_3}{X_3} \\ \frac{dX_4}{dW_4} * \frac{W_4}{X_4} \\ \frac{dX_5}{dW_5} * \frac{W_5}{X_5} \\ \frac{dX_6}{dW_6} * \frac{W_6}{X_6} \end{Bmatrix} = \begin{Bmatrix} 0.052 \\ 0.047 \\ 0.043 \\ 0.030 \\ 0.032 \\ 0.039 \end{Bmatrix}$$

七、 模型的评价与推广

7.1 模型的评价

7.1.2 优点

（1）BO-LSTM 的优点是它可以自动地找到模型最佳的超参数配置，从而使该模型的训练过程和预测结果准确率和效率更高。

（2）BO-LSTM 对于不同的时间序列数据也有很好的实用性，在不同的问题上均可以取得良好的性能。

（3）运用 SPSSPRO 将数据可视化，更加便于理解；

（4）运用 SPSSPRO、MATLAB 等软件做出相关图表，更加直观、清晰、简捷；

（5）运用 MATLAB 拟合工具箱对数据进行拟合，提高了模型的准确性；

7.1.3 缺点

（1）贝叶斯优化算法的计算成本较高，特别是在超参数空间较大的情况下该缺陷会更加明显。

（2）BO-LSTM 的性能主要依赖于目标函数的选择和超参数空间的定义，故我们在使用该算法之前，必须花费大量的时间和精力去选择目标函数和超参数空间，以便取得更好的结果。

（3）研究不够细致。

（4）可用数据较少。

7.2 模型的推广

BO-LSTM 模型不仅可以用于黄金证券的等的预测，也可以用于水库移民安置意向智能预测、天然气处理厂负荷率预测、电池剩余使用寿命预测、短期风速的功率预测。

八、 模型的改进

我们提出一种 EMD-BO-LSTM 模型的想法：即将 EMD 与 LSTM 结合，建立黄金股价预测函数，并应用贝叶斯优化算法对 LSTM 超参数进行寻优。方法如下，

首先，获取原始序列，手动对数据预处理；对处理后的数据用 EMD 分解将国际黄金价格序列分解成若干个 IMF 项和残差项，将分解后的 IMF 分量与影响变量重组，并划分为训练集和测试集；将训练集作为输入建立 LSTM 模型，用贝叶斯优化对 LSTM 模型进行超参数寻优，并返回最优的超参数组；把测试集输入到上一步中训练好的模型，完成预测验证.

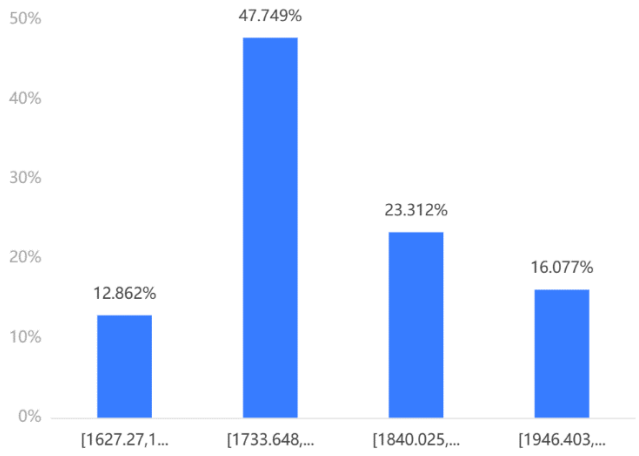
此方法是比较好的，但由于时间的限制，我们并未完成模型的创建，但由算法的良好性可以得知这种办法能更好的预测黄金价格。

参考文献

[1] Persio L D, Honchar O. Recurrent neural networks approach to the financialforecast
of Google assets[J]. International Journal of Mathematics andComputers in simulation, 2017, 11: 7-13.
[2] Mo Y A, Jing W. Adaptability of Financial Time Series Prediction Based on BiLSTM[J]. Procedia Computer Science, 2022, 199: 18-25
[3]王嘉增.基于改进 LSTM 神经网络的股价预测与股票投资组合研究[D].西安建筑科技大学,2023.DOI:10.27393/d.cnki.gxazu.2023.000894
[4] 邓凤欣, 王洪良. LSTM 神经网络在股票价格趋势预测中的应用———基于美 港股票市场个股数据的研究[J]. 金融经济, 2018(14): 96-98.
[5] 陈卫华. 基于深度学习的上证综指波动率预测效果比较研究[J]. 统计与信息 论坛, 2018, 33(05): 99-106
[6]毛昊然,叶发青,叶瑞辉等.基于 BO-LSTM 的水库移民安置意向智能预测模型[J/OL].水利发展研究:1-9[2023-12-22].
[7] WILLIAMS C K I, RASMUSSEN C E. Gaussian processes for machine learning[M]. Cambridge, MA: MIT press, 2006.
[8]毛昊然,叶发青,叶瑞辉等. 基于BO-LSTM的水库移民安置意向智能预测模型 [J/OL].水利发展研究, 1-9[2023-12-20].
[9]王嘉增.基于改进 LSTM 神经网络的股价预测与股票投资组合研究[D].西安建筑科技大学,2023.DOI:10.27393/d.cnki.gxazu.2023.000894
[10] Scientific Platform Serving for Statistics Professional 2021. SPSSPRO. (Version 1.0.11)[Online Application Software]. Retrieved from https://www.spsspro.com.

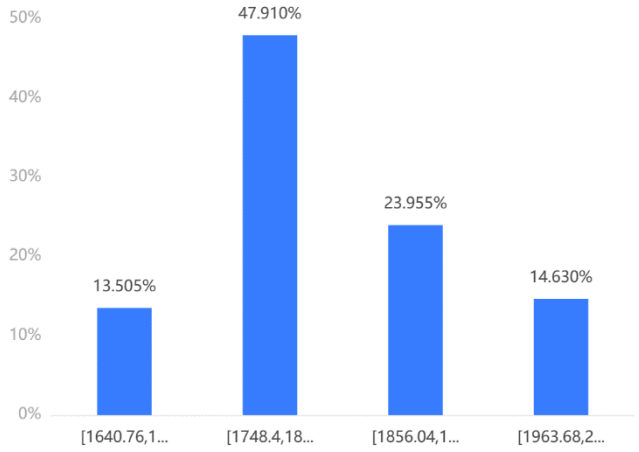
附录

附录 1：



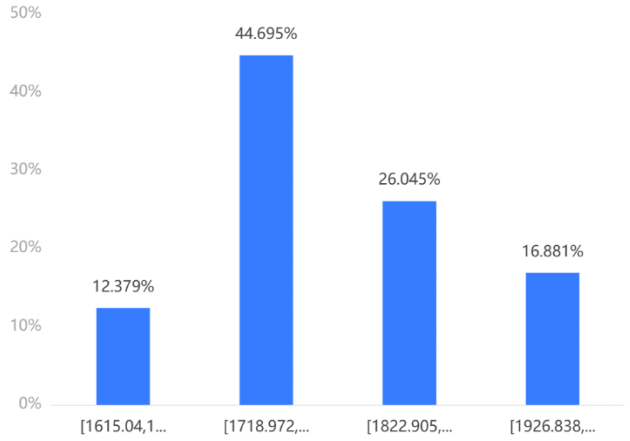
开盘价

附录 2：



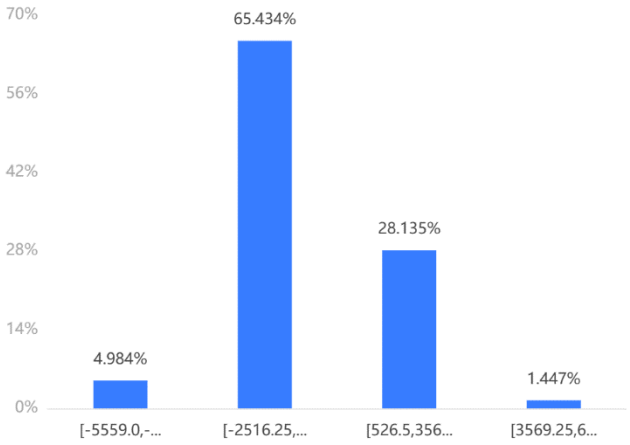
最高价

附录 3：



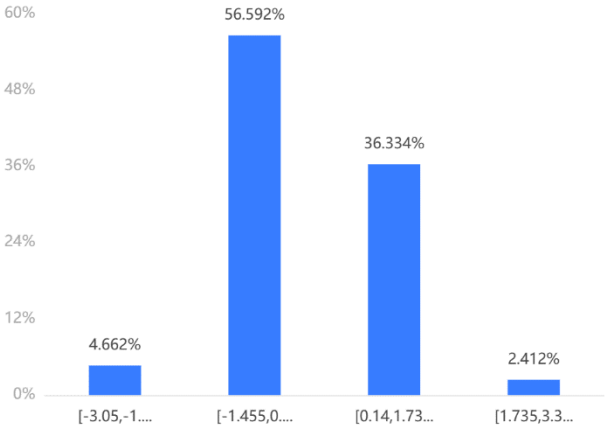
最低价

附录 4：



变化（点数）

附录 5：



变化（百分比）

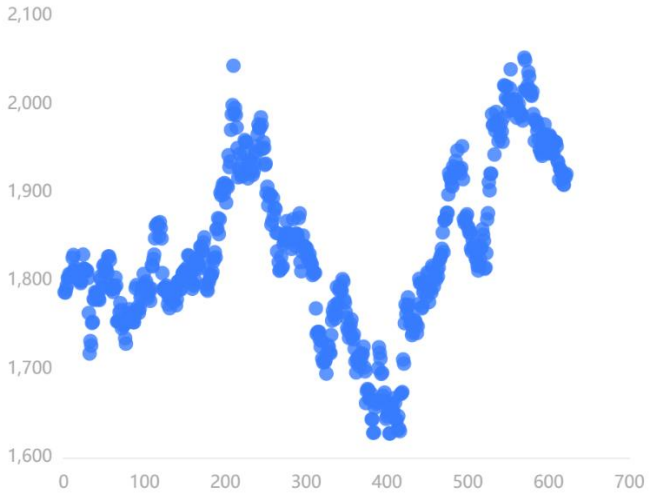
附录 6：

频数分析结果表

名称	选项	频数	百分比(%)	累计百分比(%)
Open	[1627.27,1733.648)	80	12.862	12.862
	[1733.648,1840.025)	297	47.749	60.611
	[1840.025,1946.403)	145	23.312	83.923
	[1946.403,2052.78]	100	16.077	100
High	[1640.76,1748.4)	84	13.505	13.505
	[1748.4,1856.04)	298	47.91	61.415
	[1856.04,1963.68)	149	23.955	85.37
	[1963.68,2071.32]	91	14.63	100
Low	[1615.04,1718.972)	77	12.379	12.379
	[1718.972,1822.905)	278	44.695	57.074
	[1822.905,1926.838)	162	26.045	83.119
	[1926.838,2030.77]	105	16.881	100
Close	[1626.89,1733.15)	79	12.701	12.701
	[1733.15,1839.41)	295	47.428	60.129
	[1839.41,1945.67)	148	23.794	83.923
	[1945.67,2051.93]	100	16.077	100
Change(Pips)	[-5559.0,-2516.25)	31	4.984	4.984
	[-2516.25,526.5)	407	65.434	70.418
	[526.5,3569.25)	175	28.135	98.553
	[3569.25,6612.0]	9	1.447	100
Change(%)	[-3.05,-1.455)	29	4.662	4.662
	[-1.455,0.14)	352	56.592	61.254
	[0.14,1.735)	226	36.334	97.588

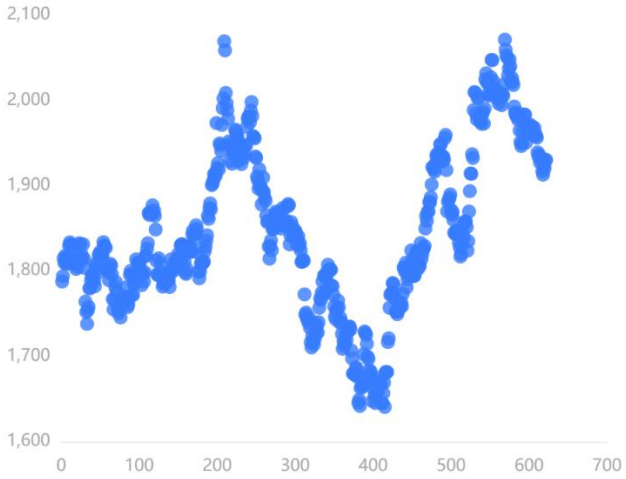
名称	选项	频数	百分比(%)	累计百分比(%)
	[1.735,3.33]	15	2.412	100
合计		622	100.000	100.000

附录 7：



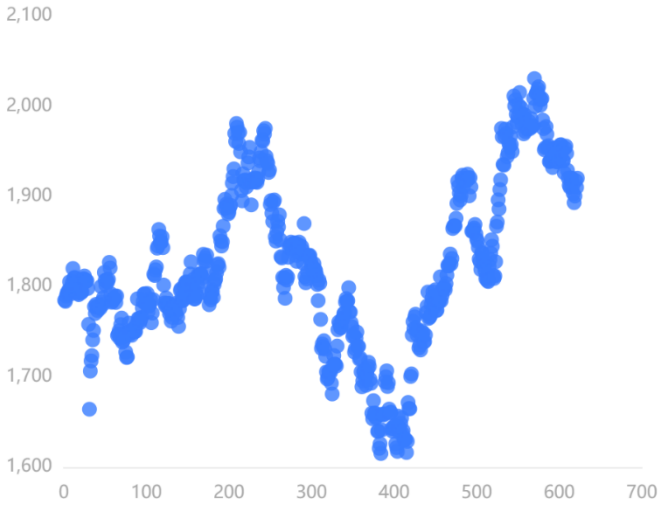
开盘价

附录 8：



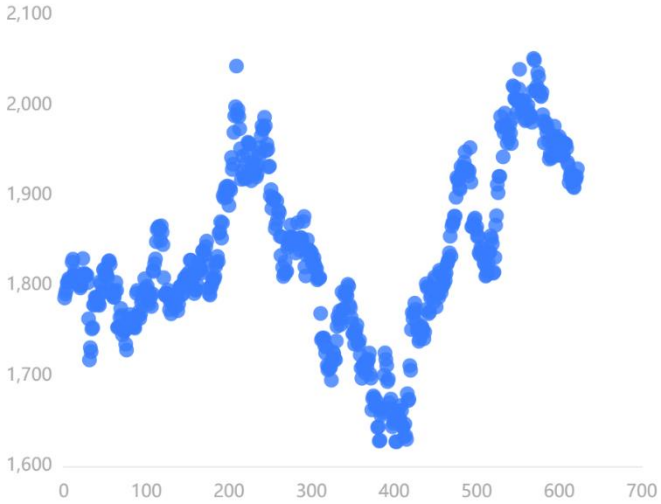
最高价

附录 9：



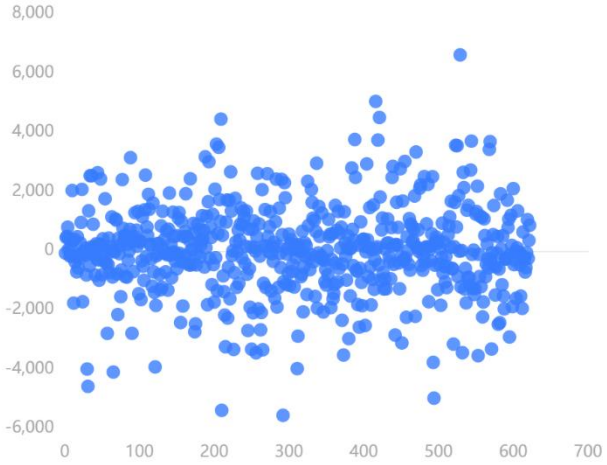
最低价

附录 10：



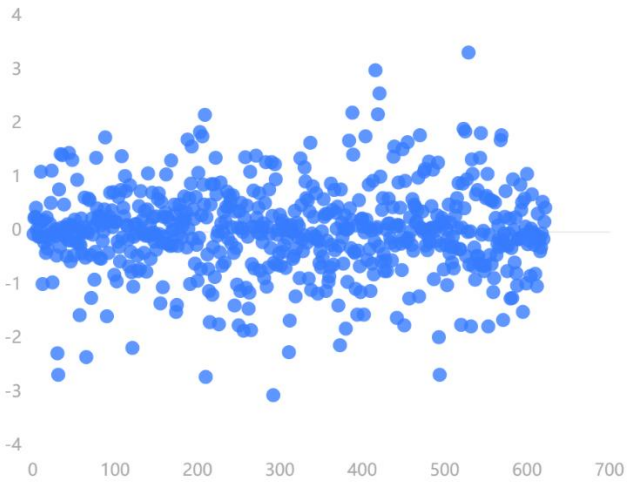
收盘价

附录 11：



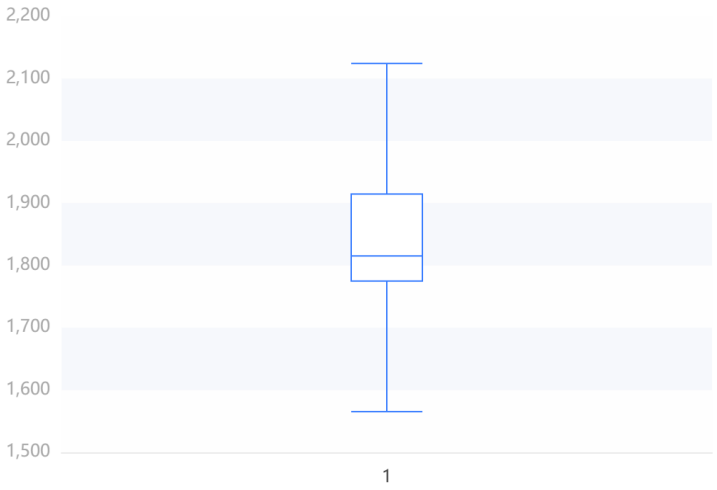
变化（点数）

附录 12：



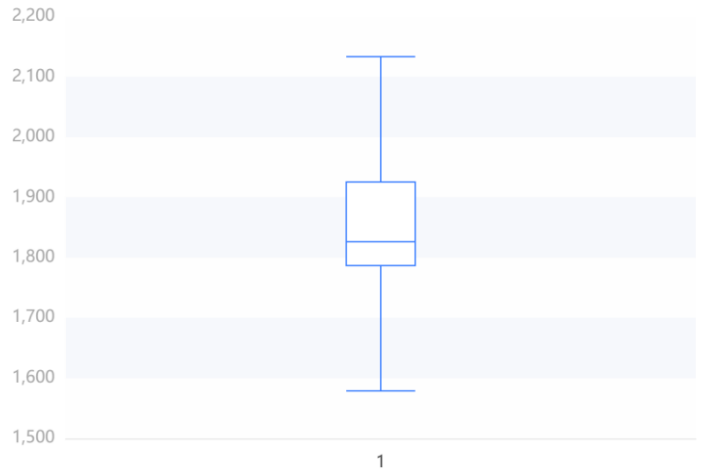
变化（百分比）

附录 13：



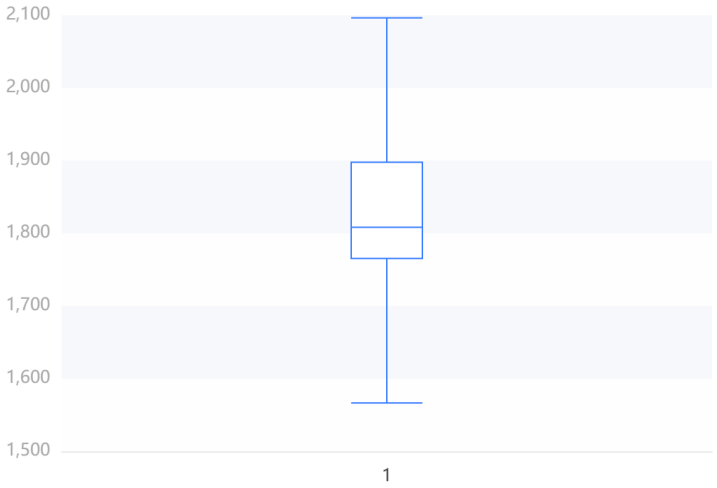
开盘价

附录 14：



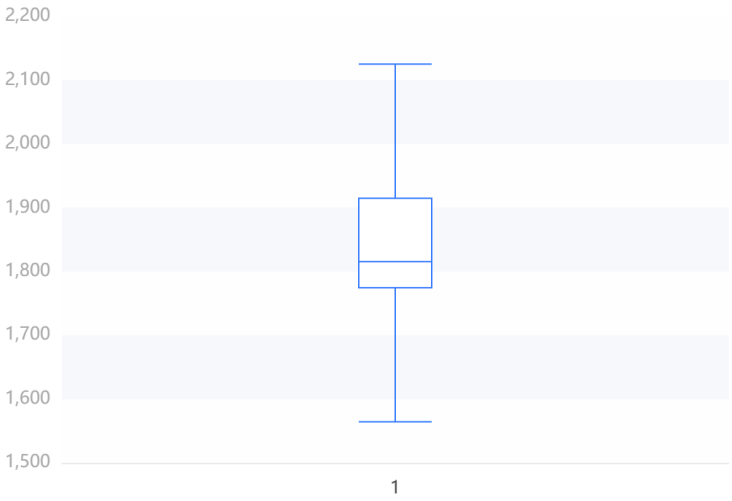
最高价

附录 15：



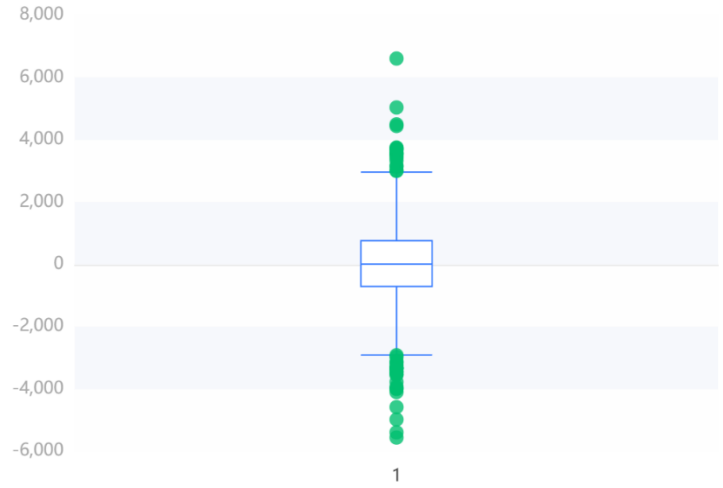
最低价

附录 16：



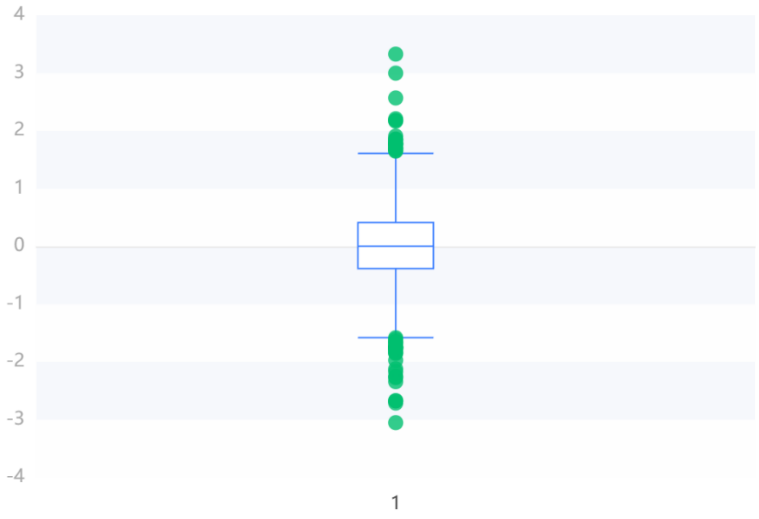
收盘价

附录 17：



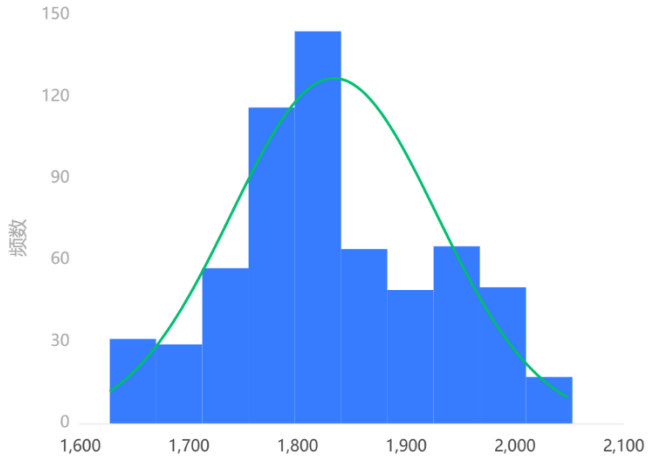
变化（点数）

附录 18：



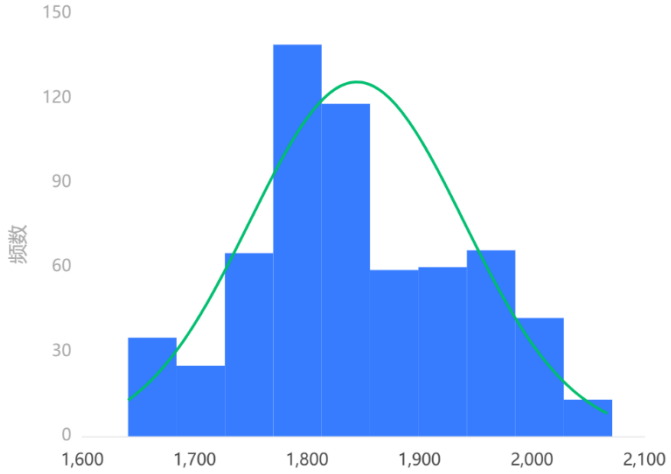
变化（百分比）

附录 19：



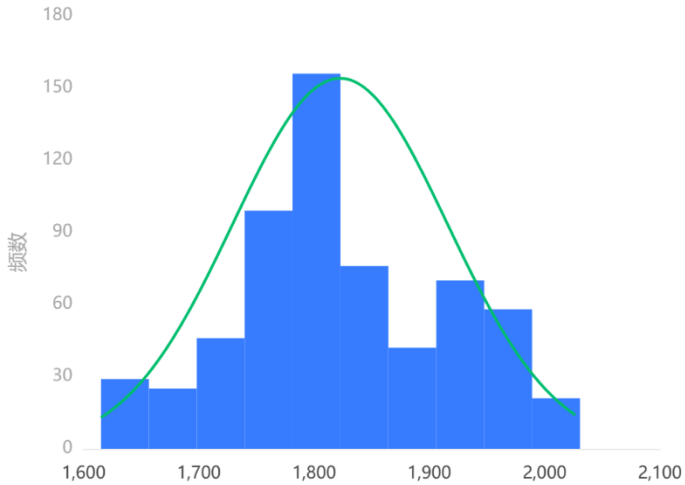
开盘价

附录 20：



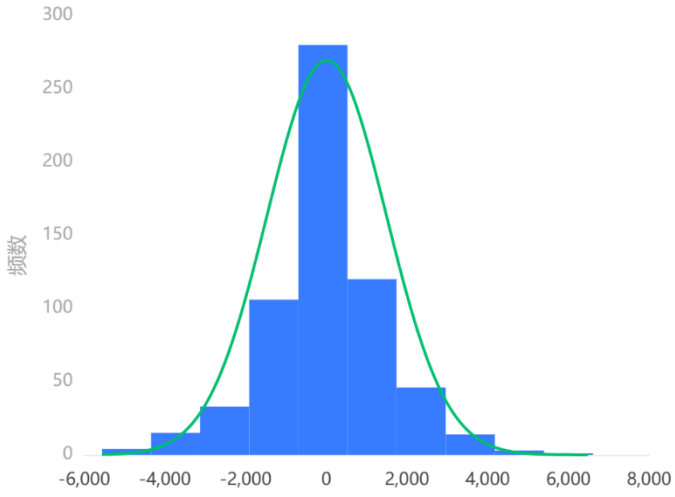
最高价

附录 21：



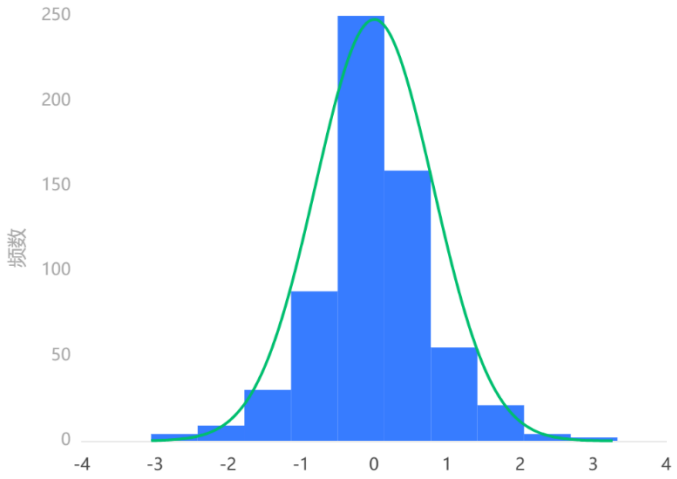
最低价

附录 22：



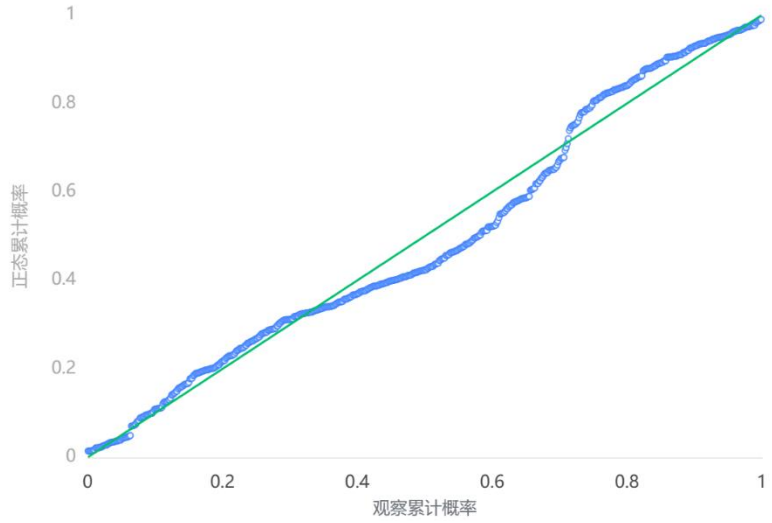
变化（点数）

附录 23：



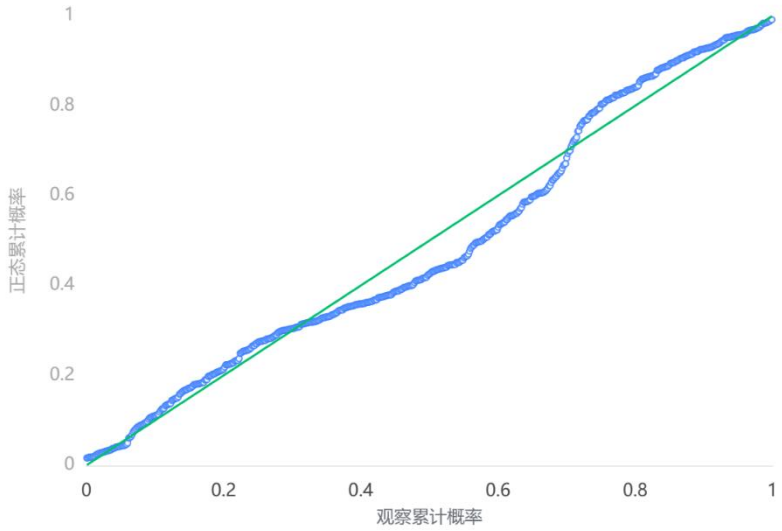
变化（百分比）

附录 24：



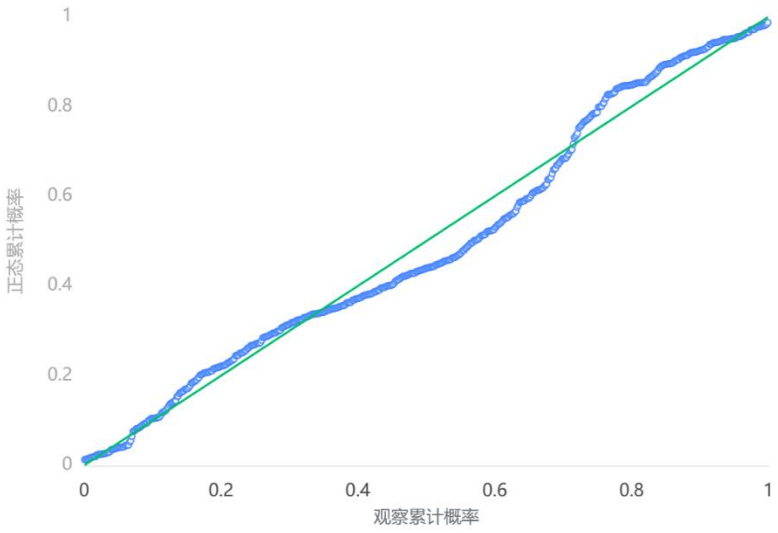
开盘价

附录 25：



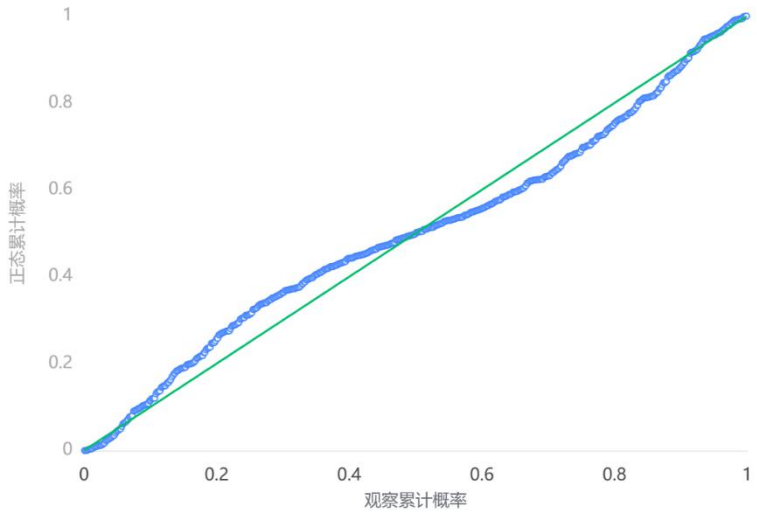
最高价

附录 26：



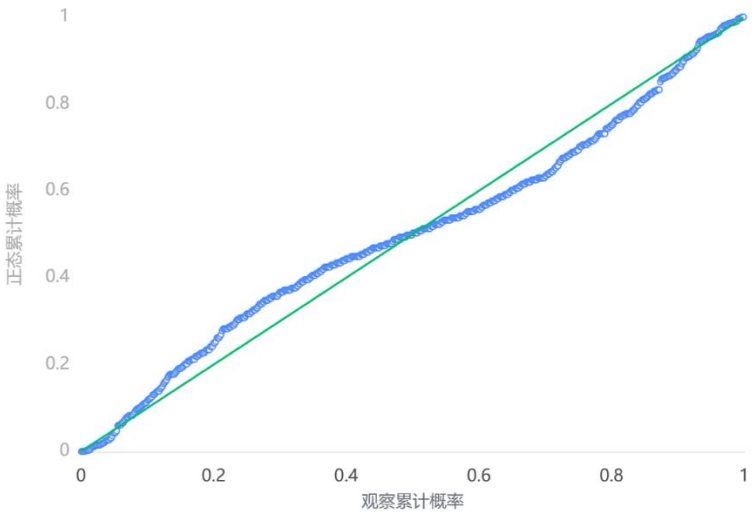
最低价

附录 27：



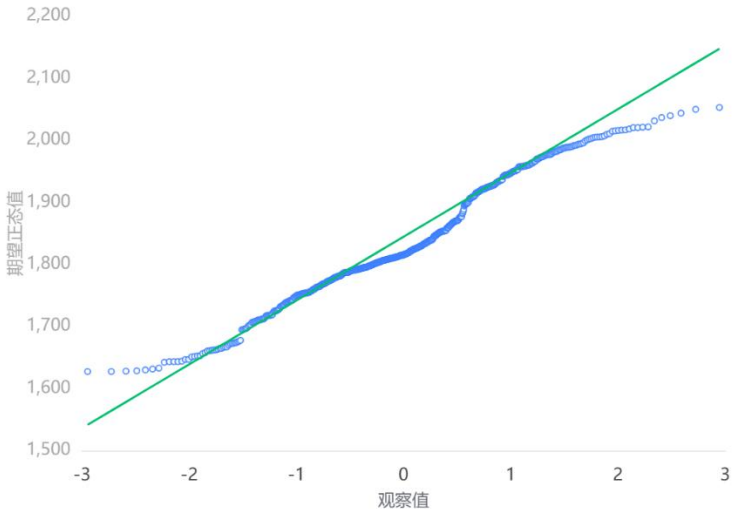
变化（点数）

附录 28：



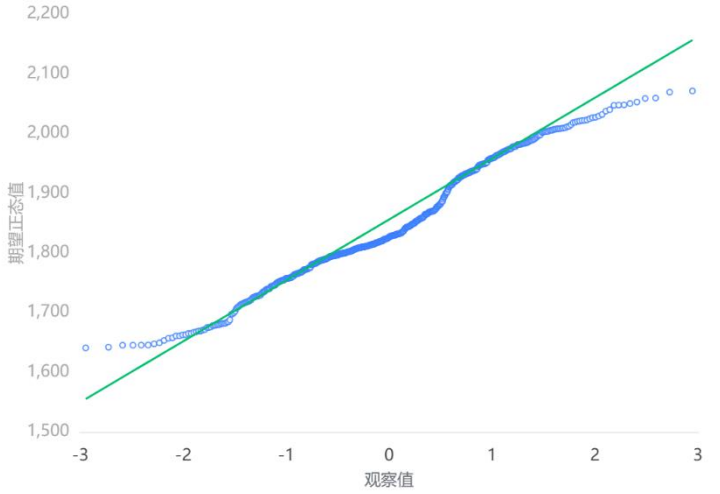
变化（百分比）

附录 29：



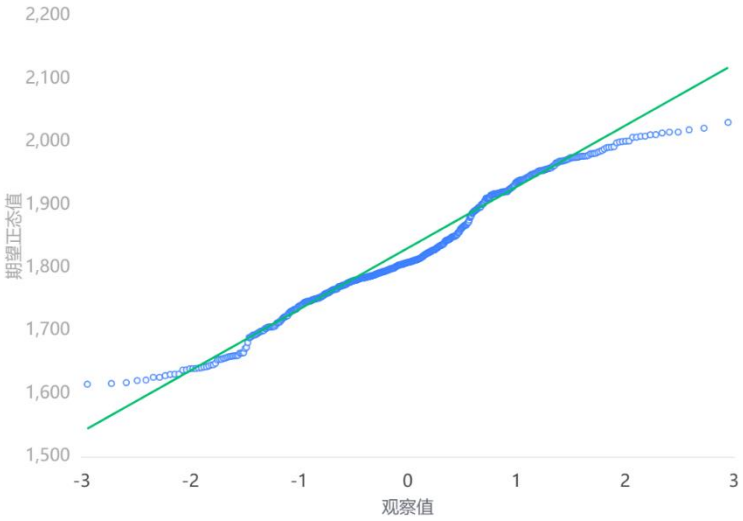
开盘价

附录 30：



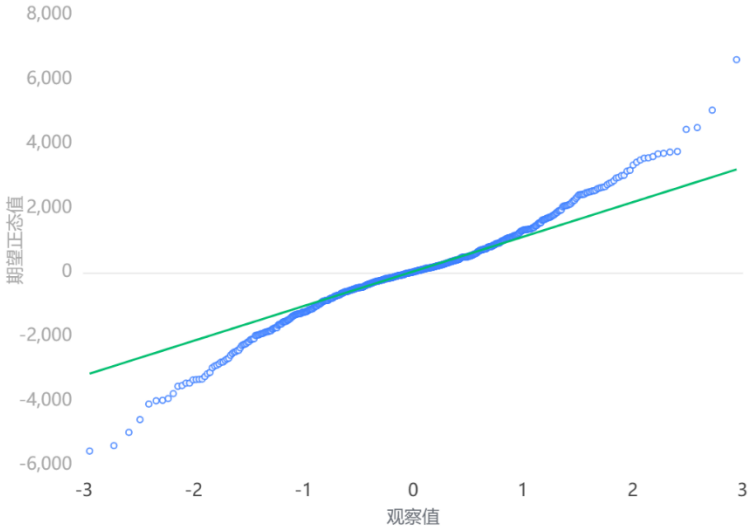
最高价

附录 31：



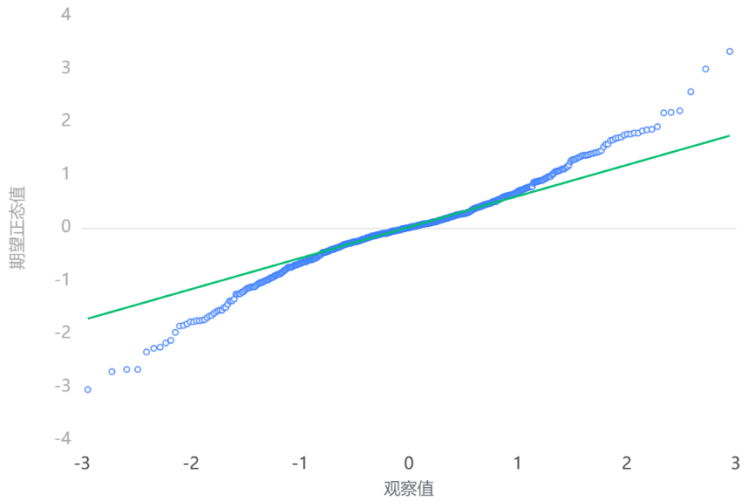
最低价

附录 32：



变化（点数）

附录 33：



变化（百分比）

附录 34：

正态性检验结果表

变量名	样本量	中位数	平均值	标准差	偏度	峰度	S-W 检验	K-S 检验
Open	622	1815.23	1833.47	94.418	0.153	-0.552	0.977(0.000***)	0.083(0.000***)
High	622	1826.375	1844.209	95.163	0.207	-0.562	0.976(0.000***)	0.093(0.000***)
Low	622	1808.335	1822.538	93.223	0.099	-0.55	0.978(0.000***)	0.079(0.001***)
Close	622	1815.47	1833.593	94.346	0.147	-0.563	0.977(0.000***)	0.082(0.000***)
Change(Pips)	622	11.5	12.24	1480.618	-0.026	1.745	0.974(0.000***)	0.072(0.003***)
Change(%)	622	0.005	0.004	0.805	-0.058	1.674	0.974(0.000***)	0.072(0.003***)

注：***、**、*分别代表 1%、5%、10%的显著性水平

附录 35：

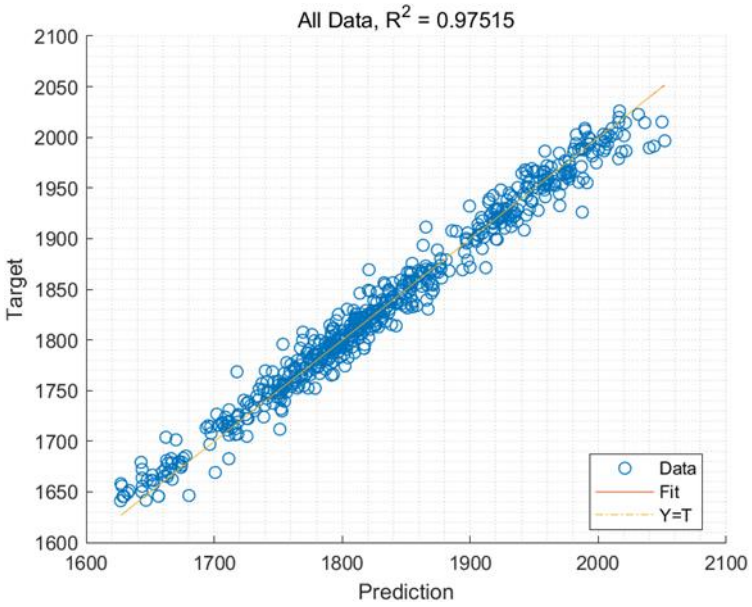


图 1 总数据拟合效果

附录 36：

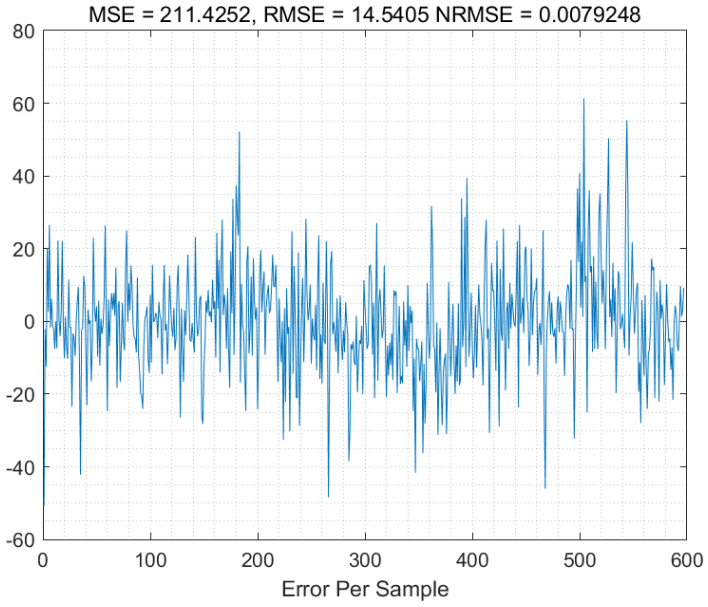


图 1 总数据拟合误差

附录 37：

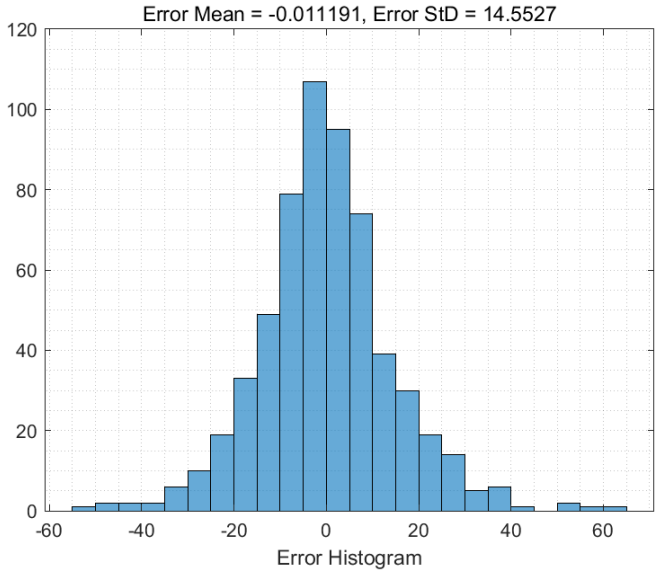


图 1 总数据误差直方图

附录 38：

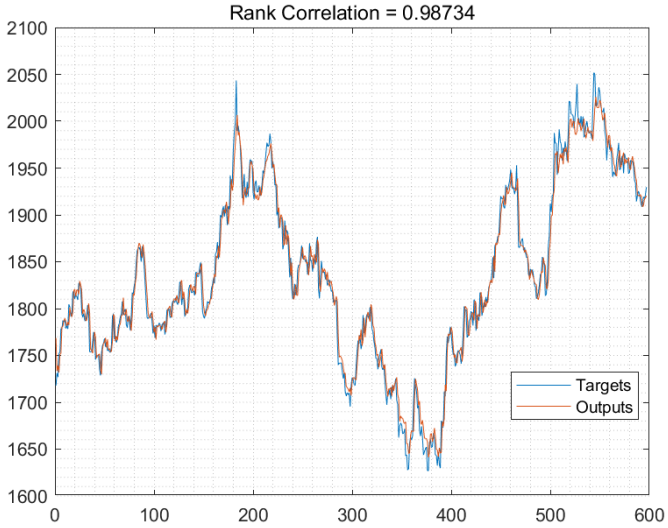


图 1 总数据线性拟合效果

附录 39：

变量名	样本量	最大值	最小值	平均值	标准差	中位数	方差	峰度	偏度	变异系数（CV）
Open	622	2052.78	1627.27	1833.47	94.418	1815.23	8914.708	-0.552	0.153	0.051
High	622	2071.32	1640.76	1844.209	95.163	1826.375	9056.054	-0.562	0.207	0.052
Low	622	2030.77	1615.04	1822.538	93.223	1808.335	8690.543	-0.55	0.099	0.051
Close	622	2051.93	1626.89	1833.593	94.346	1815.47	8901.146	-0.563	0.147	0.051

变量名	样本量	最大值	最小值	平均值	标准差	中位数	方差	峰度	偏度	变异系数（CV）
Change(Pips)	622	6612	-5559	12.24	1480.618	11.5	2192228.524	1.745	-0.026	120.97
Change(%)	622	3.33	-3.05	0.004	0.805	0.005	0.647	1.674	-0.058	222.407

附录 40：

变量名	样本量	中位数	平均值	标准差	偏度	峰度	S-W 检验	K-S 检验
Open	622	1815.23	1833.47	94.418	0.153	-0.552	0.977(0.000***)	0.083(0.000***)
High	622	1826.375	1844.209	95.163	0.207	-0.562	0.976(0.000***)	0.093(0.000***)
Low	622	1808.335	1822.538	93.223	0.099	-0.55	0.978(0.000***)	0.079(0.001***)
Close	622	1815.47	1833.593	94.346	0.147	-0.563	0.977(0.000***)	0.082(0.000***)
Change(Pips)	622	11.5	12.24	1480.618	-0.026	1.745	0.974(0.000***)	0.072(0.003***)
Change(%)	622	0.005	0.004	0.805	-0.058	1.674	0.974(0.000***)	0.072(0.003***)

注：***、**、*分别代表 1%、5%、10%的显著性水平