基于长短期记忆神经网络的量化投资

摘　要：

著名的投资大师Joel Greenblatt曾说过“如果你在选个股的过程中都不知道自己到底在找什么，那感觉就像你奔跑在一个快要被点燃的炸药厂。虽然你也有可能活下来，但那不管怎样你都是一个傻瓜。”股票市场是一个高风险高收益的市场，有的投资者在一夜之间暴富，有着极高的短期收益，但也有的投资者随着股市的起起伏伏，输的倾家荡产。在二十一世纪经济学的理论日渐成熟完善，量化投资，择时、套利、交易，配置，风控一套操作也形成体系。且随着社会经济的发展，计算机的技术水平也在爆炸性的增长，对于股市存在的风险与收益复杂的关系，在计算机大数据的分析之下正在逐渐揭开其神秘的面纱。

针对问题一，本文将各项指标中的缺失值进行分析预处理，并对各项指标数据采用皮尔逊相关系数处理，结合科普洛特热相关性图直观分析，从技术指标中提取出BBI、EXPMA、MA，其他板块信息中提取出快手概念、互联网电商、互联网,宏观数据指标提取出GDP现价等7项与数字经济板块呈现正相关系数较强的重要指标，最后运用探索性因子分析法从所选取指标中提取出4 个与数字经济板块相关的主要指标。

针对问题二，本文从问题一中选取相关性较强的主要指标，然后将这些指标对数字经济版块信息中的成交量指标进行预测。经过对股票的分析与研究，本文建立了LSTM神经网络预测模型。本文在SPSSPRO软件上对所选取的指标进行了**z-score**标准化处理，再将数据导入到MATLAB上实现LSTM神经网络预测，得到预测的成交量数据，并且实际值与预测值的曲线基本吻合，RMSE 的值为 **0.13389**，说明此模型得出的成交量预测结果精确可靠。

针对问题三，本文从问题一中选取相关性较强的主要指标 BBI(多空指数)，MA(移动平均线)和EXPMA(指数平均数)，然后将这些指标对数字经济版块信息中的收盘价指标进行预测。经过对股票的分析与研究，本文继续建立了LSTM神经网络预测模型，得到预测的收盘价数据，并且实际值与预测值曲线基本吻合，说明此模型得出的收盘价预测结果精确可靠。

针对问题四考虑到在股票的投资具有一定的特殊性，即将新进买入股票后，即转为股票价值，在卖出变现之前，是无法继续作为现金参与投资。是产生了一个现金流入与现金流出的差额，又因为本体的第三问已经对2022年1月4日至28日的收盘价进行预测，根据其导出结果来制定投资周期方案，符合净现值的概念，故运用 NPV 模型对初始资金100万进行合理的投资规划，计算出总收益率为 **4.9%**、信息比率为 **11.4**、最大回撤率为 **8.35%**。

最后，我们分析了我们的模型的优点和缺点，其结果表明我们的模型具有较高的鲁棒性、精度和准确性。之后，附上一份备忘录。

关 键 词：　LSTM神经网络；皮尔逊相关系数；探索性因子分析；NPV；RMSE；z-scor

目录

一、问题重述.....2
1.1 问题背景.....2
1.2 问题重述.....2
二、问题分析.....2
2.1 数据的分析.....2
2.2 对问题一的分析.....2
2.3 对问题二的分析.....3
2.4 对问题三的分析.....3
2.5 对问题四的分析.....3
三、符号和假设.....3
3.1 符号说明.....3
3.2 基本假设.....3
四、模型的建立与求解.....4
4.1 问题一的分析与求解.....4
4.2 问题二的分析求解.....8
4.2.1 LSTM 神经网络预测模型的建立与求解.....8
4.3 问题三的分析与求解.....10
4.3.1 LSTM 神经网络预测模型的求解.....10
4.4 问题四的分析与求解.....12
4.4.1 股票投资周期资金分配（EPV）模型的背景介绍.....12
4.4.2 股票投资周期资金分配（EPV）模型分析.....12
4.4.3 模型的结果与分析.....15
五、模型的优缺点.....17
5.1 模型的优点.....17
5.2 模型的缺点.....17
六、参考文献.....17
七、附录.....18

一、问题重述

1.1 问题背景

近年来,随着中国金融市场的逐渐完备，越来越多的金融投资者进入股票市场。除传统基本面投资外,量化投资交易也开始兴起。量化投资是指通过数量化方式及计算机程序化发出买卖指令，以获取稳定收益为目的的交易方式。随着大数据技术的迅速发展，量化投资已经成为当代投资者的热门手段。量化投资作为一种新型的投资方法，但在实际运用中，受到多种因素的限制，并存在多种风险。因此如何从海量的市场信息中提取出有效指标，并制订出合适的交易策略，是一个值得深入分析的工作。

本题附件中给出了宏观市场指标、国内股票市场指标、技术指标、国际股票市场指标、“数字经济”板块信息、其他板块信息的指标数据。

1.2 问题重述

请基于 2021 年 7 月 14 日至 2022 年 1 月 28 日每 5 分钟的“数字经济”板块给出的数据信息，解决以下问题：

问题一：对所提供的各项指标进行分析，从中提取出与“数字经济”板块有关的主要指标。

问题二：以 2021 年 7 月 14 日至 2021 年 12 月 31 日的每 5 分钟“数字经济”板块指数为 训练集，以 2022 年 1 月 4 日至 2022 年 1 月 28 日的每 5 分钟“数字经济”板块指数为测试集。根据问题一提取出来的各项指标对“数字经济”板块指数每 5 分钟成交量进行预测。

问题三：以 2021 年 7 月 14 日至 2021 年 12 月 31 日的每 5 分钟“数字经济”板块指数为 训练集，以 2022 年 1 月 4 日至 2022 年 1 月 28 日的每 5 分钟“数字经济”板块指数为测 试集。根据问题一和二建立模型对每 5 分钟的“数字经济”板块指数（收盘价）进行预 测。

问题四：假设以“数字经济”板块指数为交易对象（在实际交易中指数无法交易，只能 交易其中的个股），给定初始资金 100 万元，交易佣金为 0.3%，根据问题三得到的结果对 “数字经济”板块每 5 分钟频率价格进行买卖交易，计算在 2022 年 1 月 4 日至 2022 年 1 月 28 日期间交易的总收益率、信息比率、最大回撤率。

二、问题分析

2.1 数据的分析

1.对给出的指标数据进行了缺失值处理；
2.选取了数字经济板块信息中的 15.00 的数据
3.对成交量数据进行了 z-score 标准化处理

2.2 对问题一的分析

针对附表中的所有指标采用皮尔逊相关系数进行选取，并以热相关性图直观表示，筛选出 7 个指标，再结合探索性因子分析法最终筛选出 3 个与数字经济板块相关性较强

的指标，即 BBI(多空指数)，MA(移动平均线)，EXPMA(指数平均数)。

2.3 对问题二的分析

通过问题一所选取的 BBI(多空指数)，MA(移动平均线)，EXPMA(指数平均数)主要指标对数字经济模块中的成交量指标进行预测，本文通过分析研究股票与选取的指标，建立了 LSTM 神经网络预测模型。

2.4 对问题三的分析

通过问题一所选取的 BBI(多空指数)，MA(移动平均线)，EXPMA(指数平均数)主要指标对数字经济模块中的收盘价指标进行预测，本文通过分析研究股票与选取的指标，建立了 LSTM 神经网络预测模型。

2.5 对问题四的分析

依据第三问的 LSTM 神经网络预测模型对每五分钟的“数字经济”板块指数的预测数据（收盘价）所导出的数据进行分析处理，因为净现值法考虑了资金的时间成本，股票在买入期间股票的价值并非为现金进行投资，且根据第三问预测结果加入合理的风险评估，故运用净现值法（NPV）来预测股票买入（投资）的现金流量（一周期之内），在根据上一周期的资金积累，从新带入公式来对 NPV 值的大小进行评价是否继续进行买入，直至 2022 年 1 月 28 日的净终值（四周期累计的净收益折算的最终值）。再根据整体的数据对总收益率和信息比率及最大回撤率进行计算。

三、符号和假设

3.1 符号说明

符号	定义
α_a	a 的标准差
$\bar{a}$	样本均值
α_b	b 的标准差
r_{ab}	样本相关系数
$Cov(ab)$	样本协方差
S_a	a 的样本标准差
S_b	b 的样本标准差
i_t	控制信息是否流入
f_t	控制上一时刻 Memory 信息是否积累
o_t	控制当前时刻是否流入隐藏状态中
c_t	记忆单元，表示神经元状态的记忆
W_*	相应门限的递归链接权重

3.2 基本假设

1.假设题目所给的数据真实可靠

2.假设运用模型在成交量预测结果时未出现较大误差；

3.假设各项指标的提取不受外界因素影响，仅对本文数字经济板块产生影响；

4.假设买卖交易的进行真实可靠，交易对象的选取不干扰交易过程；

四、模型的建立与求解

4.1 问题一的分析与求解

4.1.1 皮尔逊相关系数模型与探索性因子分析的建立与求解

本文在问题一的分析中，首先用 SPSS 将国际市场指标和宏观市场指标中的缺失值运用均值插补法集中处理，再对各项指标中的数据运用皮尔逊相关系数分析各个变量同数字经济板块之间的正向相关性：

如果两组数据$a:\{a_1,a_2,\dots,a_n\}$和$b:\{b_1,b_2,\dots,b_n\}$是总体数据，那么

$$\beta_{ab}=\frac{Cov(a,b)}{\alpha_a\alpha_b}=\frac{\sum_{i=1}^{n}\frac{(a_i-E(a))(b_i-E(b))}{\alpha_a\quad\alpha_b}}{x} \tag{1}$$

$$\alpha_a=\sqrt{\frac{\sum_{i=1}^{x}(a_i-E(a))^2}{x}} \tag{2}$$

$$\alpha_b=\sqrt{\frac{\sum_{i=1}^{x}(b_i-E(b))^2}{x}} \tag{3}$$

即证：$|\beta_{ab}|\le 1$，且当$a=mb+k$时，

$$\beta_{ab}=\begin{cases}1 & ,m>0\\ -1 & ,m<0\end{cases} \tag{4}$$

假设有两组数据$a:\{a_1,a_2,\dots,a_n\}$和$b:\{b_1,b_2,\dots,b_n\}$

样本均值：

$$\bar{a}=\frac{\sum_{i=1}^{x}a_i}{x};\bar{b}=\frac{\sum_{i=1}^{x}b_i}{x} \tag{5}$$

样本协方差：

$$Cov(a,b)=\frac{\sum_{i=1}^{x}(a_i-\bar{a})(b_i-\bar{b})}{x-1} \tag{6}$$

样本 pearson 相关系数：

$$r_{ab}=\frac{Cov(a,b)}{S_aS_b} \tag{7}$$

其中$S_a(\text{sigma a})$是 a 的样本标准差：

$$S_a=\sqrt{\frac{\sum_{i=1}^{x}(a_i-\bar{a})^2}{x-1}} \tag{8}$$

同理：

$$S_b=\sqrt{\frac{\sum_{i=1}^{x}(b_i-\bar{b})^2}{x-1}} \tag{9}$$

由皮尔逊相关系数计算出的相关性用科普洛特热相关性图直观表示出来：

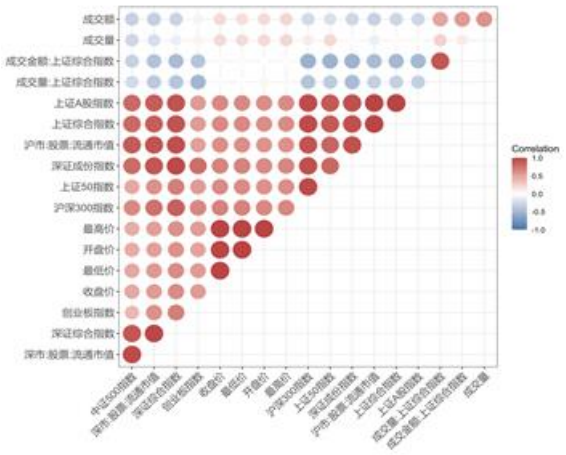


图 1 技术指标与经济板块图

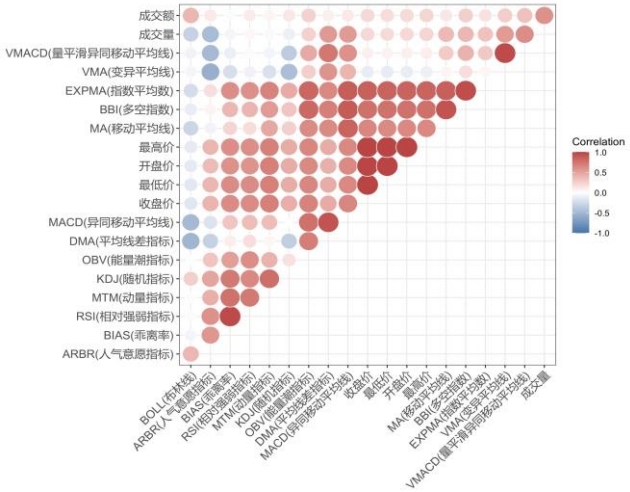


图 2 国内市场指标与经济板块图

由图 1 中的热相关性图可看出技术指标中的 EXPMA(指数平均数)、BBI(多空指数)以及 MA(移动平均线)对数字经济板块信息中对应的收盘价、开盘价、最低价和最高价呈现的正相关系数均为显著，且较为集中，可作为提取指标。图 2 中可以看出国内市场各项指标和数字经济板块中的指标显著性没有明显变化，较为平均。

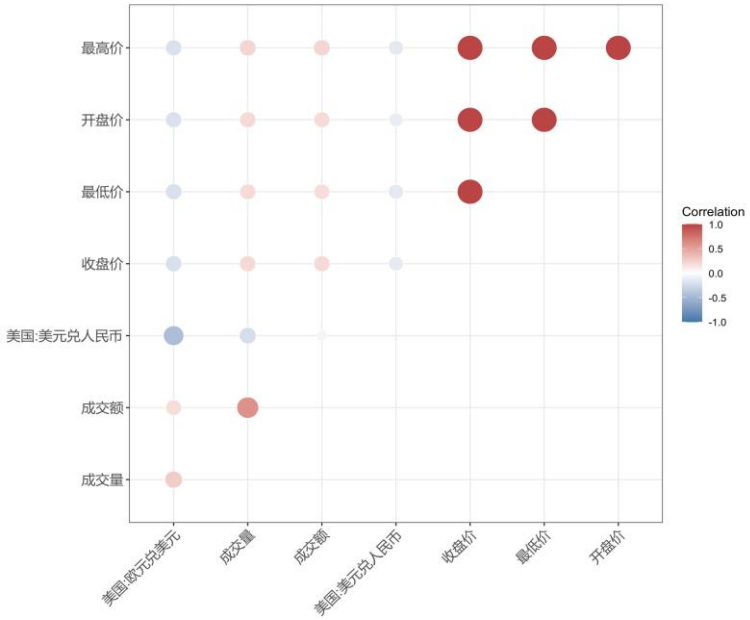


图 3 汇率指标与经济板块图

由图 3 中的热相关性图可看出汇率指标中的各项指标和数字经济板块中的指标相关性显著水平一般，影响较弱。

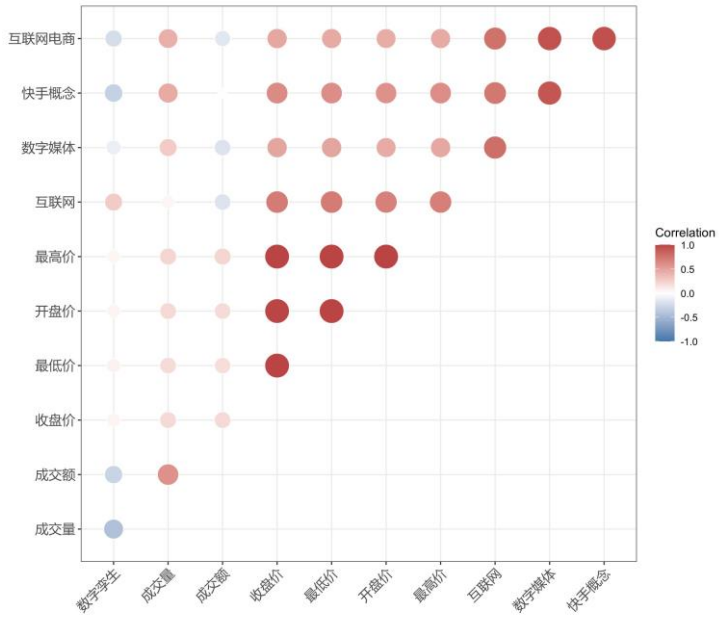


图 4 其他板块信息与经济板块图

由图 4 的热相关性图可看出其他板块信息中的快手概念、互联网电商和互联网等指标与数字经济板块中的指标正相关性较为显著，可作为提取指标。

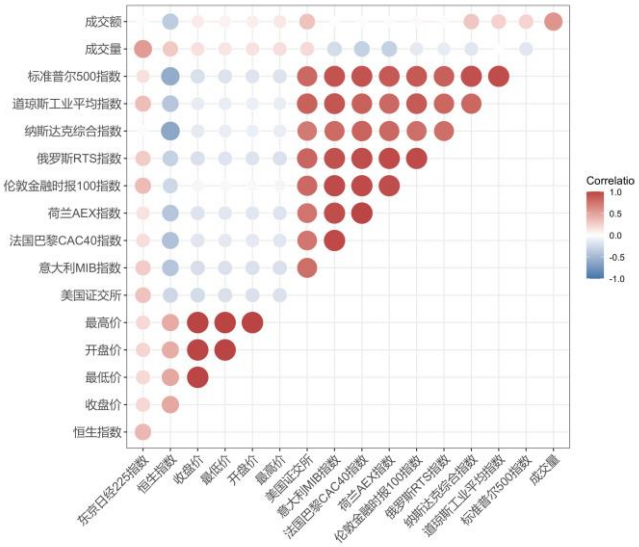


图 5 国际市场指标与经济板块图

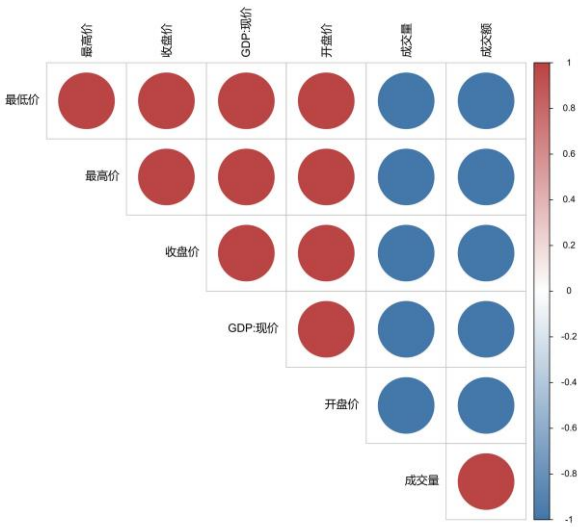


图 6 宏观市场指标 2 与经济板块图

由图 5 可以看出国际市场指标与数字经济板块中的指标显著性较弱，相关性不明显；图 6 中宏观市场指标中的 GDP-现价和数字经济指标中的最低价呈现正相关性，显著性较强，可作为提取指标。

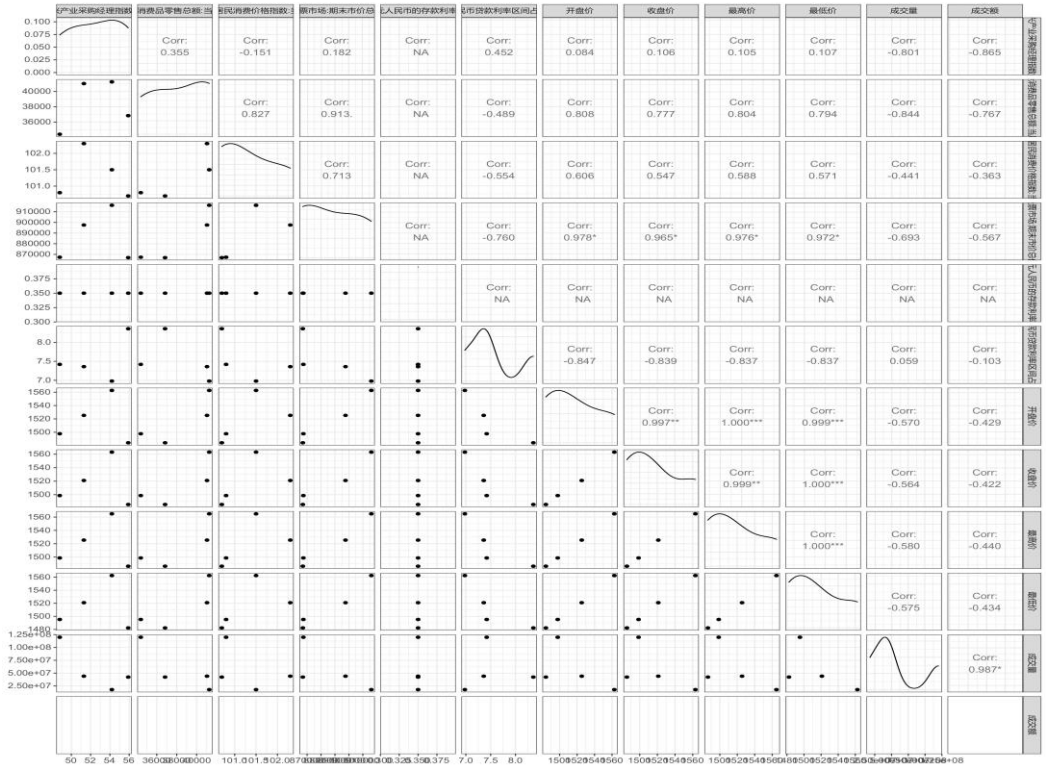


图 7 宏观市场指标 1 与经济板块配对图

由图 7 可看出股票市场(期末市价总值)和数字经济板块中的开盘价、收盘价和最高价正相关系数较高，可作为提取指标，其余指标相关系数较低。

提取出与数字经济板块有关的七个指标后，统一对七个指标:BBI(多空指数)、EXPMA(指数平均数)、MA(移动平均线)、快手概念、互联网电商、互联网和 GDP-现价进行因子分析法，在未知其中因子载荷情况下采用探索性因子分析，得出 KMO 系数为 0.9716，说明该指标信度较高，最后根据旋转因子载荷数得出因子载荷较大的 3 个指标分别是 BBI(多空指数)，MA(移动平均线)，EXPMA(指数平均数)，接着绘制股票 K 线图，并以数字经济板块中的开盘价、收盘价、最高价和最低价为主要指标进行绘制，如图 8 所示：

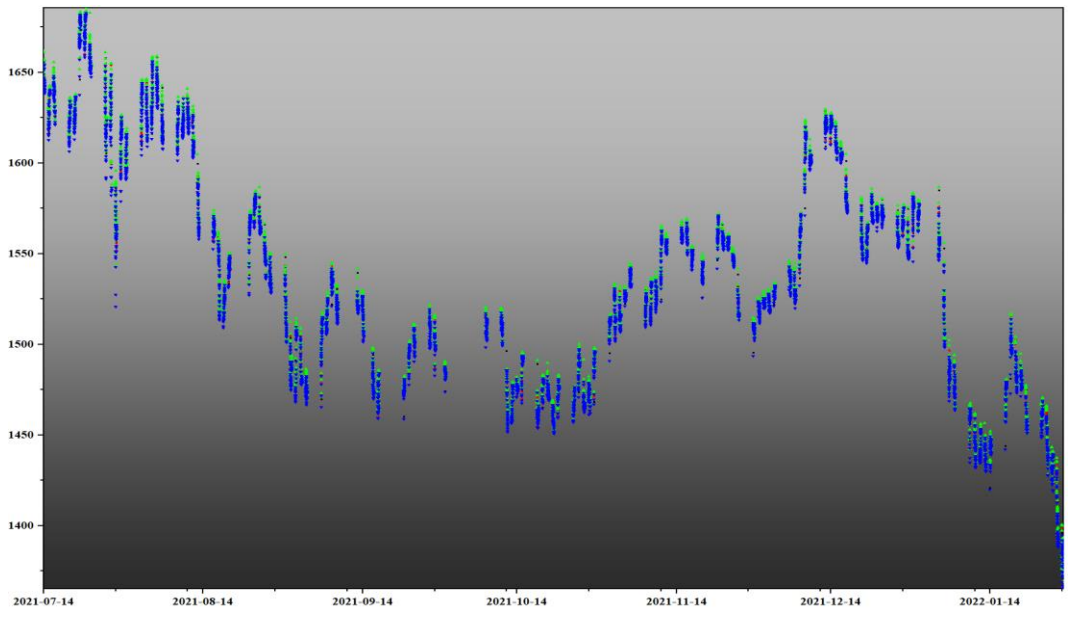


图 8 股票 K 线图

本文股票 K 线图为单根 K 线，从中可以看出在 21 年 7 月至八月这段时间阳线上涨幅度较大，同时在 12 月份阳线上涨幅度也比同邻月份迅速，可为后续提供参考。

4.2 问题二的分析求解

4.2.1 LSTM 神经网络预测模型的建立与求解

为了分析与研究股票的相关数据，更好地预测股票的成交量，本文选用长短期记忆神经网络预测模型（LSTM 神经网络）进行解决问题。LSTM 神经网络模型是由 Hochreiter 和 Schmidhuber 两位科学家 1997 年发表的一类传统循环神经网络基础上改进后的应用模型，通过在传统的循环模型中加入输入、输出和遗忘三个逻辑单元构成更为适合长期时序数据的模型。具体原理图如下：

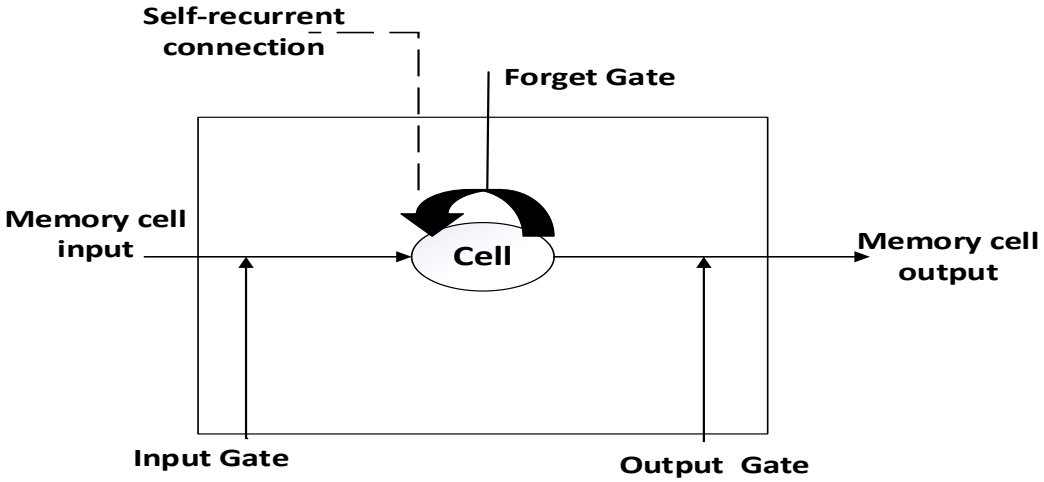


图 9 LSTM 神经网络模型原理图

其中 Input Gate、Output Gate 和 Forget Gate 分别为控制数据流入、控制数据流出和数据遗忘的意思。LSTM 神经网络预测模型又叫长短期记忆神经网络预测模型，为了更好地理解模型中的含义，可以通过 LSTM 神经网络的结构图（图 10）进行进一步的理解[3]。

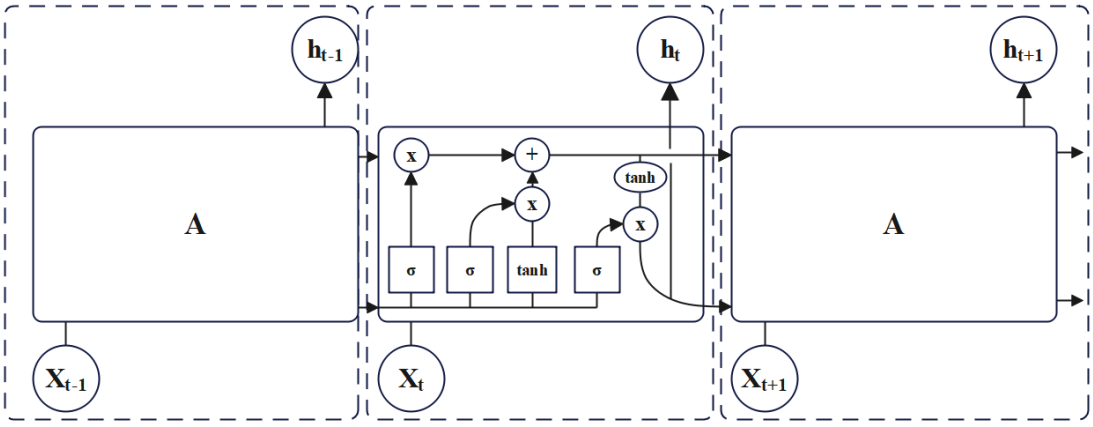


图 10 LSTM 神经网络结构

LSTM 神经网络预测模型可以大致分为以下四个步骤进行：

步骤一：从上一个细胞获得数据后，选择要遗忘的数据，遗忘数据计算：

$$f_t = \text{sigmoid}\left(W_f * \left[h_t -, x_t\right] + b_f\right) \tag{10}$$

步骤二：对学习的数据进行保存，数据的输入计算：

$$i_t = \text{sigmoid}\left(W_{i^*}\left[h_t - 1, x_i\right] + b_i\right) \tag{11}$$

步骤三：获取新的数据和上一细胞数据遗忘程度综合后，细胞数据更新：

$$C_{\text{t}} = \tanh\left(W_{\text{c}} * \left[h_t - 1, x_t\right] + b_c\right) \tag{12}$$

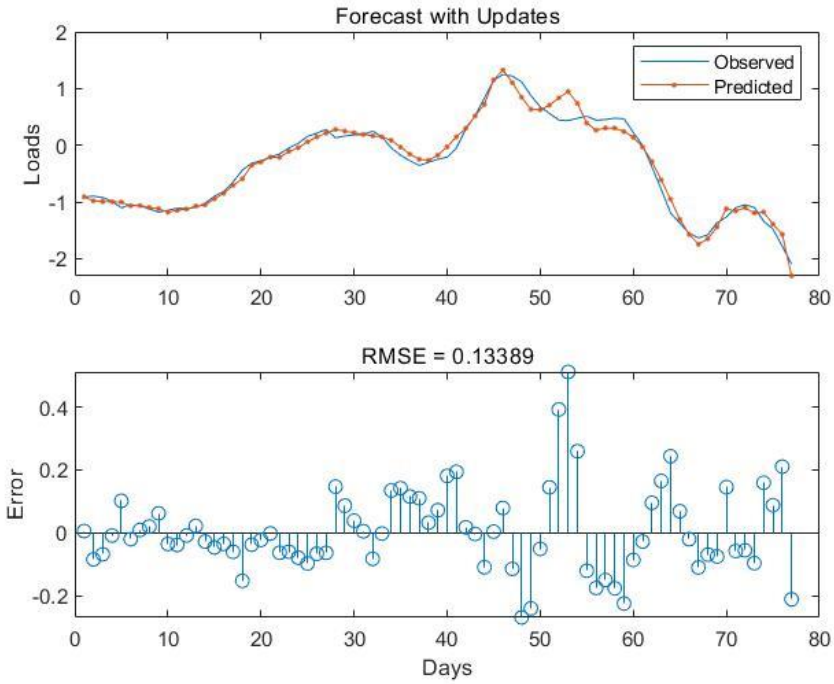
步骤四：将输入的数据和上一个细胞的影响进行未来的预测，输出数据计算：

$$O_{\text{t}} = \text{sigmoid}\left(W_o * \left[h_t - 1, x_t\right] + b_o\right) \tag{13}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{14}$$

由 LSTM 神经网络的原理可以看到，输出一个预测值可以通过对多个时序数据进行输入 LSTM 模型对数据进行导入后得出未来的预测值，在本文中，将股票成交量输入其中进行预测[1]。

对于数字经济指标模块，本文选取了 2021 年 7 月 14 日到 2022 年 1 月 28 日每天 10 点的成交量数据，再将成交量数据与问题一选取的 BBI(多空指数)，MA(移动平均线)，EXPMA(指数平均数)主要指标数据选取相同的时间区域，在软件 SPSSPRO 中将选取的指标数据进行 z-score 标准化处理。本文将经过 z-score 标准化处理后的指标数据导入到

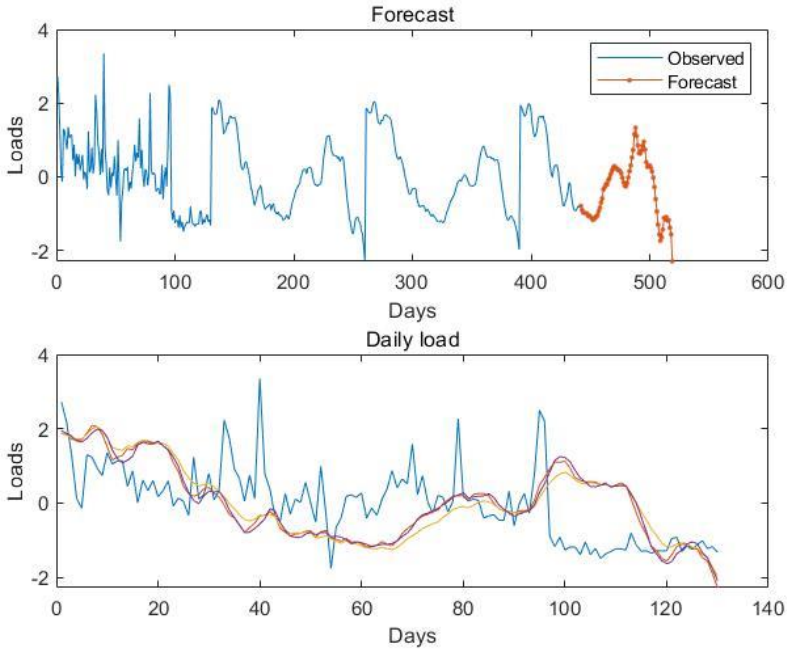


MATLAB 中，通过 LSTM 神经网络算法得出了通过导入大批量数据得出实际和预测得验证图如下：

图 11 成交量的实际值和预测值验证图

输入 2021 年 7 月 14 日到 2022 年 1 月 28 日每天 10 点的成交量与问题一选取的指标共有的时间域上一共有 130 个数据，本文以 2021 年 7 月 14 日至 2021 年 12 月 31 日的数据为训练集（111 个，占总数据的 85%），以 2022 年 1 月 4 日至 2022 年 1 月

28 日的数据为测试集（19 个，占总数据的 15%）。在 MATLAB 上进行对数字经济板块中的股票成交量进行预测，得到实际值和预测值的差异验证，图 11 中分析看到，实际值的成交量趋势和预测值得成交量趋势基本是吻合的，但是实际会出现一定的突发情况，导致预测的值和实际的值存在一定的差异。从趋势和误差综合分析可以得到，RMSE 的值为 0.13389（模型出错的概率很小），本文对成交量的预测是准确的。因此本文在



MATLAB 上进一步地绘制了成交量具体的成交量预测值以及预测趋势，图 12 所示。

图 12 股票成交量（标准化）预测

从图 12 中，我们可以清楚地看到成交量（标准化）的预测值和预测趋势走向，因此通过成交量作为输入在 LATM 神经网咯模型中可以较好的得到一个未来的预测走势图和未来成交量数据的预测值。

4.3 问题三的分析与求解

4.3.1 LSTM 神经网络预测模型的求解

通过问题一所选取的 BBI(多空指数)，MA(移动平均线)，EXPMA(指数平均数)主要指标对数字经济模块中的收盘价指标进行预测，本文通过分析研究股票与选取的指标，建立了 LSTM 神经网络预测模型。

对于问题三，为了分析与研究股票的相关数据，更好地预测股票的收盘价，本文继续选用长短期记忆神经网络预测模型（LSTM 神经网络）进行解决问题。

对于数字经济指标模块，本文选取了 2021 年 7 月 14 日到 2022 年 1 月 28 日每天 10 点的收盘价数据，再将收盘价数据与问题一选取的 BBI(多空指数)，MA(移动平均线)，EXPMA(指数平均数)主要指标数据选取相同的时间区域，再将指标数据导入到 MATLAB 中，通过 LSTM 神经网络算法得出了通过导入大批量数据得出实际和预测得验证图如下：

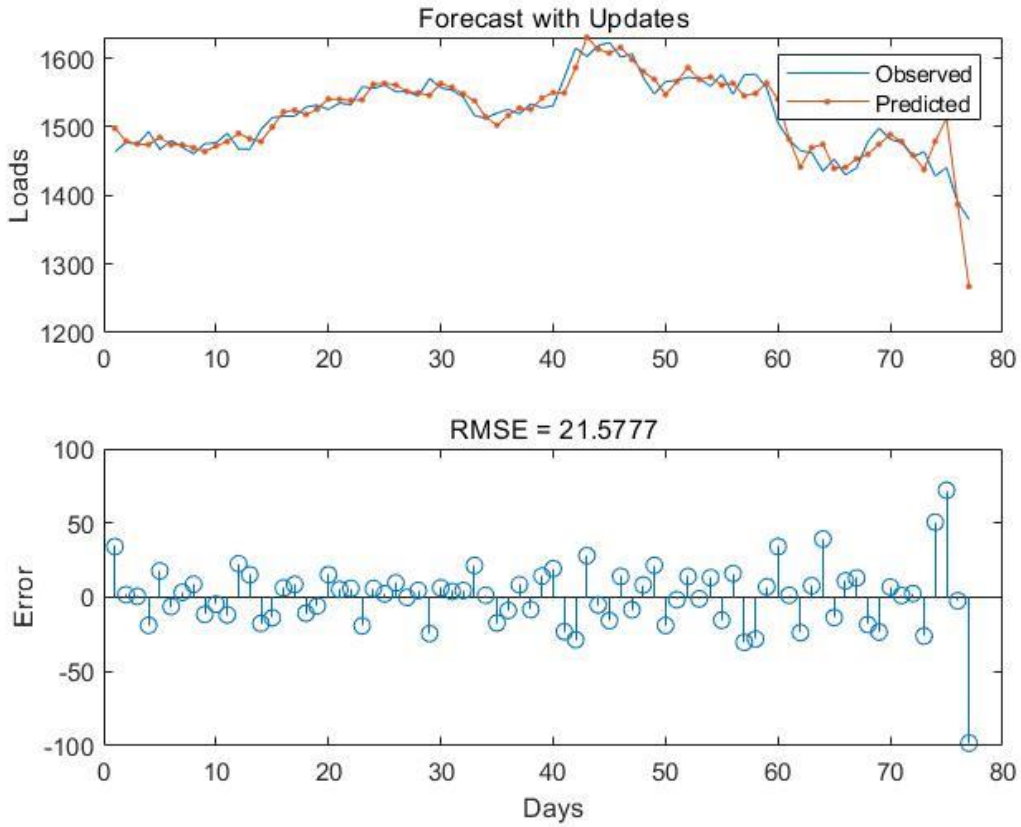


图 13 收盘价的实际值和预测值验证图

输入 2021 年 7 月 14 日到 2022 年 1 月 28 日每天 10 点的收盘价与问题一选取的指标共有的时间域上一共有 130 个数据，本文以 2021 年 7 月 14 日至 2021 年 12 月 31 日的数据为训练集（111 个，占总数据的 85%），以 2022 年 1 月 4 日至 2022 年 1 月 28 日的数据为测试集（19 个，占总数据的 15%）。在 MATLAB 上进行对数字经济板块中的股票收盘价进行预测，得到实际值和预测值的差异验证，图 13 中分析看到，实际值的收盘价趋势和预测值得成交量趋势基本是吻合的，但是实际会出现一定的突发情况，导致预测的值和实际的值存在一定的差异。从趋势和误差综合分析可以得到，本文对成交量的预测是准确的。因此本文在 MATLAB 上进一步地绘制了成交量具体的成交量预测值以及预测趋势，图 14 所示。

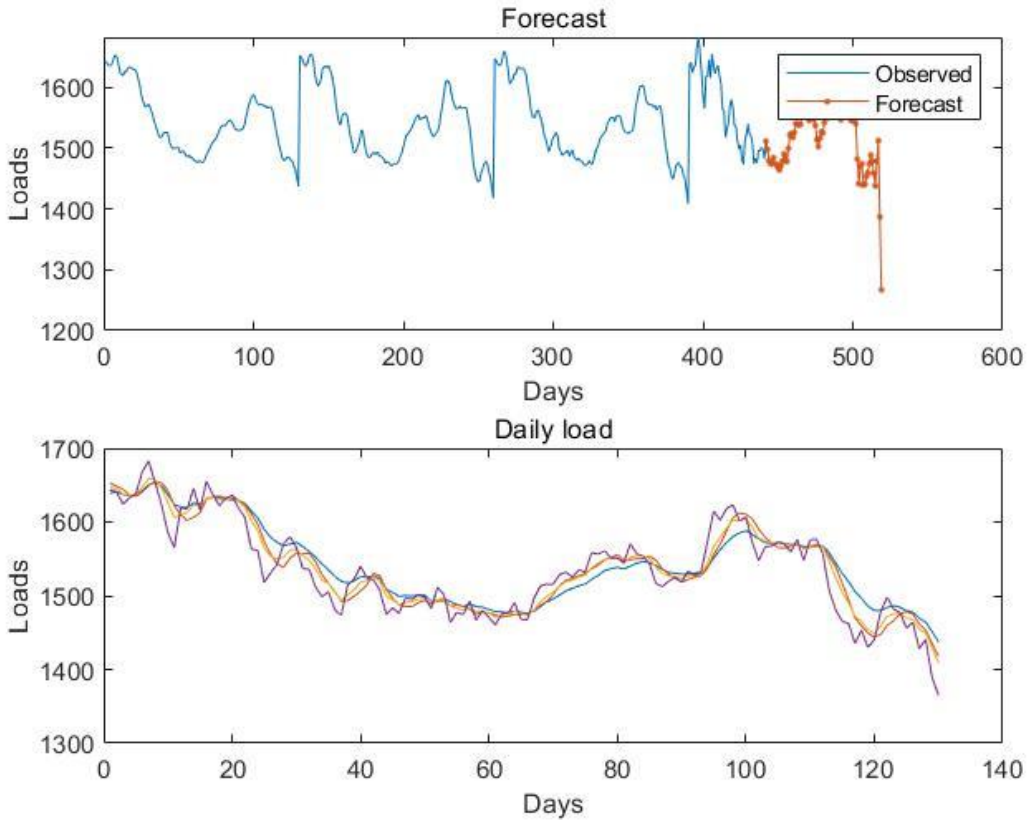


图 14 股票收盘价的预测

从图 14 中，我们可以清楚地看到收盘价的预测值和预测趋势走向，因此通过成交量作为输入在 LATM 神经网咯模型中可以较好的得到一个未来的预测走势图和未来成交量数据的预测值。

4.4 问题四的分析与求解

4.4.1 股票投资周期资金分配（EPV）模型的背景介绍

我们以“数字经济”板块指数作为交易对象，且给定初始资金在 100 万元，每次的交易佣金为 0.3%，我们团队选取的方法是净现值法，因为净现值法是现金效益量的总现值与净现金投资量算出净现值，然后根据净现值的大小来进行判断。考虑到股票价值并不为现金，其根本原因是会影响第二日股票开盘后对其买入的投资。且净现值法（NPV）为较为成熟的计算方法在经济学领域中，包括本次我们团队，选取 python 的 numpy_financial 库可对 npv 进行计算[6]。

4.4.2 股票投资周期资金分配（EPV）模型分析

NPV 模型

结合本题的情况，初始资金为 100 万元我们既定该资金不会以股票增值以外的方式来进行增加，即为资本限额。当面对 n（n≤4）个 NPV>0 的股票周期,但投资资本却有限时,一般的办法就是将投资周期各种组合情况进行一一比较,选择预期收益最大的投资方案。但是该方法显然不适合应用于本题，虽然我们可利用第三问预测结果进行计算参考，但是我们无法忽略资金有限和折现问题，且收盘价每个周期都在进行变化，存在时间成

本，不为同一时刻进行，故进行改进，建立投资分配模型。

股票投资周期资金分配模型

因收盘价的数据已通过预测得出，在金额有限的情况下，我们需对这四个周期进行选择。通常情况下，会选择 NPV 最大的周期进行买空，本模型以净终值为最大目标函数[2]。模型如下：

$$\begin{aligned} & Ma.r\sum_{r} AX - W_t \\ & s.t.\ (a)\sum_{j} a_t, X_j - W_t \le D_t \\ & \quad (b)\sum_{j} a_{t,} X_j + (1+r)W_{t-1} - W_t \leqslant D_t \\ & \quad (c)0 \leqslant X_j \leqslant 1 (d)W_t \ge 0 \end{aligned} \tag{15}$$

其中

A 为一个周期中的折算终值，即为净终值。

a_t 为在一个周期中第 t 天所发生的现金流量。若为正值则代表买入的股票，若为负值则代表卖出。

D_t 为预计第 t 天可支配资金，且不包括折算为股票价值的资金。

W_t 为第 t 天所将股票值变现金额。

r 则为进行股票交易所需支付的交易佣金。

X_j 为 0—1 变量，若 X=1 表示接收该周期，若 X=0 表示该周期拒绝。

此模型的解和 NPV 标准的关系

在最优化状态下,影子价格是指资源对目标函数的边际贡献。也即资源增加某-一微量时对最优化了的目标函数所起的影响。在以上模型中,被接受的方案,即 X=1 它的影子价格是一定大于零的(因为目标函数为极大)。而一个线性规划的影子价格.就是其对偶问题的最优解。所以通过对偶问题的分析.可以知道线性规划选择方案的标准[4]。

对偶模型为：

$$
\begin{aligned}
&\text{Min}\ \sum_{t-1}^{T}\rho_t D_t+\sum_{j-1}^{n}U_j\\
&s\cdot t\quad (a)\sum_{i-1}^{T}\rho_t a_t,+U_j\geq A\quad j=1.\cdots.n\\
&\qquad (b)\quad \rho_T\geq 1\\
&\qquad (c)-\rho_T\geq -1\\
&\qquad (d)\ \ \rho_{i-1}-(1+r)\rho_i\geq 0\quad t=2.\cdots.T\\
&\qquad (e)-\rho_{i-1}+(1+r)\rho_1\geq 0\quad t=2\cdots,T\\
&\qquad (f)\rho_t,U_j\geq 0
\end{aligned}
\tag{16}
$$

其中U_j原规划约束（c）的对偶变量，表示周期 j 的影子价格。

ρ_t原规划约束（a），（b）的对偶变量。

由（b）和（c）可推得$\rho_T=1$，记作$\rho_T^*=1$。

由（d）得$\rho_{t-r}/\rho_t\geqslant 1+r$。

由（e）得$\rho_t-1/\rho_t\leq 1+r$。

可推得$\rho_{t-1}/\rho_t=1+r\quad t=2.\cdots.T$

所以$\rho_t=(1+r)\rho_{t+1}=(1+r)^2\rho_{t+r}=\cdots=(1+r)^{t-t}\rho_T$

因为$\rho_T^*=1$说以$\rho_t^*=(1+r)^{r-t}$这是一个增量的复利用率，表示在第 t 天增加代为资金流量在第 T 天时的终值（周期末）。

由于a_t>0 表示支出，所以当周期被接受时，亦为X_j=1 时，约束（a）变成一个方程：$U_i=A_j-\sum_{i-1}^{T}\rho_i^* a_j$，将以上$\rho_t^*$值代入得到$U_i^*=A_j+\sum_{1-1}^{T}\left(-a_j\right)(1+r)^{T-1}$，$A_j$是净终值，也可假设此值是从第 n 天卖出股票所得

$$A_j=\sum_{t-T+1}^{n}\frac{\left(-a_j\right)}{(1+r)^{1-T}}=\sum_{t-T+1}^{n}\left(-a_j\right)(1+r)^{T-t}$$

这样 $U=\sum_{t-T+1}^{n}\left(-a_j\right)(1+r)^{T-1}+\sum_{t-1}^{T}\left(-a_{tj}\right)(1+r)^{T-t}$

$$=\sum_{t-1}^{\infty}\left(-a_{tj}\right)(1+r)^{T-t}=(1+r)^r\sum_{t-1}^{r}\left(-a_j\right)(1+r)^{-t}$$

而 $\text{NPV}=\sum_{t-1}^{n}\left(-a_t\right)(1+r)^{-1};\ \ (1+r)^T$ 为一个常数。

考虑最大回撤率下的风险投资决策优化

考虑回撤率风险下的投资决策优化。一般的 NPV 法为了衡量现金流的风险，采用各种办法在无风险利率基础上进行风险加成(风险溢价),然后直接以 NPV 的正负情况来判断投资项目可否执行。传统的风险溢价确定方法有 CAPM 法、确定当量法、风险偏好系数法等等。CAPM 法基于传统的资本资产定价模型,其前提假设过于严密,对于我们当前解决问题并无帮助。但是在本题所实践的情况下，我们为已经预测出收盘价。因此，我们只需要将在当前情况下不可避免走势所产生的最大回撤率作为所参考的依据，但实际上的最大回撤率并非为实际的影响,而是一钟对投资者参考数据，故计算各周期的 NPV;同时将各周期的风险(按马尔可夫链模型计算出的标准差分离出来，对该进行权重分配[5]。改进如下：

$$(GP)\begin{cases}\min T=\min\sum_{i=1}^{2}\left[\lambda_i\left(d_i^{+}+d_i^{-}\right)\right]\\ \sum_{i=1}^{n}X_i\times NPV_i+d_1^{-}-d_1+=f_1^{*}\\ \sum_{i=1}^{n}\frac{C_i}{\sum C_i}\times\sigma_i+d_2--d_2^{+}=f_2^{*}\quad NPV_i=\sum_{t=0}^{\infty}\frac{\overline{CF}(i,d}{(1+t)t}\\ \sum_{i=1}^{n}X_i\times C_i\leqslant C\\ d_1^{-}\times d_1^{+}=0,\text{且}d_2^{-}\times d_2^{+}=0\\ 0\leqslant X_i\leqslant 1,X_i\in Z\end{cases}\tag{17}$$

4. 4. 3 模型的结果与分析

根据第三问所预测出每日 15:00 的预测值，进行总收益率的计算

表 1 收盘价与当日收益率

时间	收盘价（15.00）	当日收益率
2021-12-31	1577	
2022-01-04	1568	-0.005707039
2022-01-05	1558	-0.006377551
2022-01-06	1544	-0.008985879
2022-01-07	1530	-0.009067358
2022-01-10	1519	-0.007189542
2022-01-11	1505	-0.00921659
2022-01-12	1496	-0.005980066
2022-01-13	1480	-0.010695187
2022-01-14	1480	0
2022-01-17	1484	0.002702703

2022-01-18	1486	0.001347709
2022-01-19	1484	-0.001345895
2022-01-20	1480	-0.002695418
2022-01-21	1478	-0.001351351
2022-01-24	1470	-0.00541272
2022-01-25	1464	-0.003389831
2022-01-26	1451	-0.008797814
2022-01-27	1437	-0.009648518
2022-01-28	1653	0.150313152

套入公式进行计算得出总收益率为 4.90%

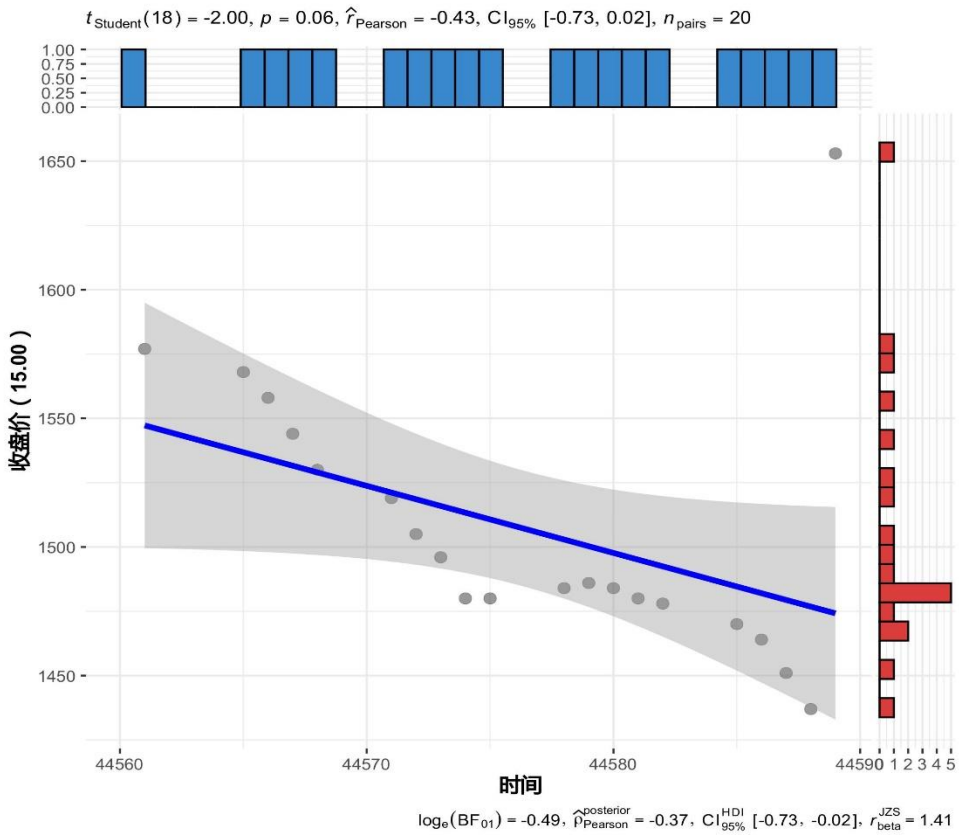


图 15 收盘价分析图

由 2022 年 1 月 4 日到 1 月 28 日的收盘价得出，直观的带入公式可得最大回撤率为 8.35%

将 NPV 的模型基于 python 实现，得出结果不难看出只有第四个周期 npv 值为正，

```
npv1=-3.841
npv2=-4.807
npv3=-3.846
npv4=10.576
```

图 16　python 运行结果

在结合预测值和模型中第 j 天代入公式得出信息比率为 11.4。

五、模型的优缺点

5.1 模型的优点

1、 LSTM 解决梯度反转由于技术不足导致的缩短过程

2、 LSTM 具有一定记忆效应，易处理时序相关敏感问题

3、 皮尔逊相关系数可准确反映两个变量之间的相关性

5.2 模型的缺点

1、 LSTM 在并行处理上落后于最新的网络

2、 当 LSTM 计算时间跨度很大时，计算量会非常大，且费时间

3、 LSTM 处理量级序列对大序列处理能力有所欠缺

4、 对本文后续数字经济指数相关预测值有较小误差

5、 当个数较大时，皮尔逊相关系数的绝对值容易偏小

六、参考文献

[1]何栩晗.基于 Matlab 的量化投资策略研究[J].商讯,2022(07):143-146.

[2]周子煜,杨超然.关于资本市场的量化投资策略及风控策略[J].商业文化,2021(21):114-115.

[3]郭家.应用 LSTM 神经网络实现机械钻速（ROP）预测及异常值检测的方法研究[J].石化技术,2022,29(03):107-108.

[4]杨晓燕,黄莹莹,喻聪,郭仲辉,赵莹.可溶性白细胞分化抗原 CD25、白细胞介素 2 与登革热相关性分析[J].标记免疫分析与临床,2020,27(09):1510-1515.

[5]袁国强,刘晓俊,朱建林.带有 VaR 方法的模糊 NPV 模型及其混合智能算法[J].计算机工程与应用,2015,51(14):35-39+50.

[6]张彩霞,李因果.不确定性投资评价中 NPV 模型的扩展[J].统计与决策,2006(19):26-27.

七、附录

神经网络预测成交量标准化代码（matlab）：

```
data = xlsread('D:\建模\华中杯\指标二问标准化.xlsx')
%%
%数据中前 85%用于训练，后 10%检验
numTimeStepsTrain = floor(0.85*numel(data));
dataTrain = data(1:numTimeStepsTrain+1);
dataTest = data(numTimeStepsTrain+1:end);
%数据预处理，得出方差和 lingo 均值。
mu = mean(dataTrain);
sig = std(dataTrain);
dataTrainStandardized = (dataTrain - mu) / sig;
%输入 LSTM 的时间序列交替
XTrain = dataTrainStandardized(1:end-1);
YTrain = dataTrainStandardized(2:end);
%%
%创建 LSTM 回归网络，指定 LSTM 层的隐含单元个数 179*3
%LSTM 预测，因此，输入一维，输出一维
numFeatures = 1;
numResponses = 1;
numHiddenUnits = 179*3;
layers = [ ...
sequenceInputLayer(numFeatures)
lstmLayer(numHiddenUnits)
fullyConnectedLayer(numResponses)
regressionLayer];
%进行 200 轮训练。
%梯度阈值设置为 1。初始学习率 0.005，在 100 轮训练后通过乘以因子 0.2 来降低学习率。
options = trainingOptions('adam', ...
'MaxEpochs',200, ...
'GradientThreshold',1, ...
'InitialLearnRate',0.005, ...
'LearnRateSchedule','piecewise', ...
'LearnRateDropPeriod',100, ...
'LearnRateDropFactor',0.2, ...
'Verbose',0, ...
'Plots','training-progress');
%训练 LSTM
net = trainNetwork(XTrain,YTrain,layers,options);
dataTestStandardized = (dataTest - mu) / sig;
XTest = dataTestStandardized(1:end-1);
```

```
YTest = dataTest(2:end);
net = resetState(net);
net = predictAndUpdateState(net,XTrain);
YPred = [];
numTimeStepsTest = numel(XTest);
for i = 1:numTimeStepsTest
[net,YPred(:,i)] =
predictAndUpdateState(net,XTest(:,i),'ExecutionEnvironment','cpu');
end
%通过以前的数据对未来进行预测
YPred = sig*YPred + mu;
%计算误差 (RMSE)。
rmse = sqrt(mean((YPred-YTest).^2));
%将预测值与实际值进行比较。
figure
subplot(2,1,1)
plot(YTest)
hold on
plot(YPred,'.-')
hold off
legend(["Observed" "Predicted"])
ylabel("Loads")
title("Forecast with Updates")
subplot(2,1,2)
stem(YPred - YTest)
xlabel("Days")
ylabel("Error")
title("RMSE = " + rmse)
figure
subplot(2,1,1)
plot(dataTrain(1:end-1))
hold on
idx = numTimeStepsTrain:(numTimeStepsTrain+numTimeStepsTest);
plot(idx,[data(numTimeStepsTrain) YPred],'.-')
hold off
xlabel("Days")
ylabel("Loads")
title("Forecast")
legend(["Observed" "Forecast"])
subplot(2,1,2)
plot(data)
xlabel("Days")
ylabel("Loads")
```

```
title("Daily load")
```

神经网络预测 15.00 收盘价代码（matlab）：

```
data = xlsread('D:\建模\华中杯\15.00 收盘价（无时间列）.xlsx')
%%
%数据中前 85%用于训练，后 10%检验
numTimeStepsTrain = floor(0.85*numel(data));
dataTrain = data(1:numTimeStepsTrain+1);
dataTest = data(numTimeStepsTrain+1:end);
%数据预处理，得出方差和 lingo 均值。
mu = mean(dataTrain);
sig = std(dataTrain);
dataTrainStandardized = (dataTrain - mu) / sig;
%输入 LSTM 的时间序列交替
XTrain = dataTrainStandardized(1:end-1);
YTrain = dataTrainStandardized(2:end);
%%
%创建 LSTM 回归网络，指定 LSTM 层的隐含单元个数 179*3
%LSTM 预测，因此，输入一维，输出一维
numFeatures = 1;
numResponses = 1;
numHiddenUnits = 179*3;
layers = [ ...
sequenceInputLayer(numFeatures)
lstmLayer(numHiddenUnits)
fullyConnectedLayer(numResponses)
regressionLayer];
%进行 200 轮训练。
%梯度阈值设置为 1。初始学习率 0.005，在 100 轮训练后通过乘以因子 0.2 来降低学习率。
options = trainingOptions('adam', ...
'MaxEpochs',200, ...
'GradientThreshold',1, ...
'InitialLearnRate',0.005, ...
'LearnRateSchedule','piecewise', ...
'LearnRateDropPeriod',100, ...
'LearnRateDropFactor',0.2, ...
'Verbose',0, ...
'Plots','training-progress');
%训练 LSTM
net = trainNetwork(XTrain,YTrain,layers,options);
dataTestStandardized = (dataTest - mu) / sig;
XTest = dataTestStandardized(1:end-1);
```

```
YTest = dataTest(2:end);
net = resetState(net);
net = predictAndUpdateState(net,XTrain);
YPred = [];
numTimeStepsTest = numel(XTest);
for i = 1:numTimeStepsTest
[net,YPred(:,i)] =
predictAndUpdateState(net,XTest(:,i),'ExecutionEnvironment','cpu');
end
%通过以前的数据对未来进行预测
YPred = sig*YPred + mu;
%计算误差 (RMSE)。
rmse = sqrt(mean((YPred-YTest).^2));
%将预测值与实际值进行比较。
figure
subplot(2,1,1)
plot(YTest)
hold on
plot(YPred,'.-')
hold off
legend(["Observed" "Predicted"])
ylabel("Loads")
title("Forecast with Updates")
subplot(2,1,2)
stem(YPred - YTest)
xlabel("Days")
ylabel("Error")
title("RMSE = " + rmse)
figure
subplot(2,1,1)
plot(dataTrain(1:end-1))
hold on
idx = numTimeStepsTrain:(numTimeStepsTrain+numTimeStepsTest);
plot(idx,[data(numTimeStepsTrain) YPred],'.-')
hold off
xlabel("Days")
ylabel("Loads")
title("Forecast")
legend(["Observed" "Forecast"])
subplot(2,1,2)
plot(data)
xlabel("Days")
ylabel("Loads")
```

```
title("Daily load")
```