

# 復旦大學

## 本科毕业论文（设计）



论文题目： 机器学习模型的高频数据趋势预测研究

院 系： 大数据学院

专 业： 物理学（数据科学与大数据技术方向）

姓 名： 张川 学 号： 16307110288

指导教师： 李祥林 职 称： 教授

单 位： 复旦大学大数据学院

日 期： 2020 年 6 月 9 日

## 毕业论文（设计）撰写人诚信承诺

### 《学位论文作假行为处理办法》（节选）

（中华人民共和国教育部令第34号发布，自2013年1月1日起施行）

**第二条** 向学位授予单位申请博士、硕士、学士学位所提交的博士学位论文、硕士学位论文和本科学生毕业论文（毕业设计或其他毕业实践环节）（统称为学位论文），出现本办法所列作假情形的，依照本办法的规定处理。

**第三条** 本办法所称学位论文作假行为包括下列情形：

- （一）购买、出售学位论文或者组织学位论文买卖的；
- （二）由他人代写、为他人代写学位论文或者组织学位论文代写的；
- （三）剽窃他人作品和学术成果的；
- （四）伪造数据的；
- （五）有其他严重学位论文作假行为的

**第四条** 学位申请人员应当恪守学术道德和学术规范，在指导教师指导下独立完成学位论文。

**第五条** 指导教师应当对学位申请人员进行学术道德、学术规范教育，对其学位论文研究和撰写过程予以指导，对学位论文是否由其独立完成进行审查。

**第七条** 学位申请人员的学位论文出现购买、由他人代写、剽窃或者伪造数据等作假情形的，学位授予单位可以取消其学位申请资格；已经获得学位的，学位授予单位可以依法撤销其学位，并注销学位证书。取消学位申请资格或者撤销学位的处理决定应当向社会公布。从做出处理决定之日起至少3年内，各学位授予单位不得再接受其学位申请。

前款规定的学位申请人员为在读学生的，其所在学校或者学位授予单位可以给予开除学籍处分；为在职人员的，学位授予单位除给予纪律处分外，还应当通报其所在单位。

**第八条** 为他人代写学位论文、出售学位论文或者组织学位论文买卖、代写的人员，属于在读学生的，其所在学校或者学位授予单位可以给予开除学籍处分；属于学校或者学位授予单位的教师和其他工作人员的，其所在学校或者学位授予单位可以给予开除处分或者解除聘任合同。

**第九条** 指导教师未履行学术道德和学术规范教育、论文指导和审查把关等职责，其指导的学位论文存在作假情形的，学位授予单位可以给予警告、记过处分；情节严重的，可以降低岗位等级直至给予开除处分或者解除聘任合同。

### 论文撰写人承诺：

本毕业论文是本人在导师指导下独立完成的，内容真实、可靠。本人在撰写毕业论文过程中不存在请人代写、抄袭或者剽窃他人作品、伪造或者篡改数据以及其他学位论文作假行为。

本人清楚知道学位论文作假行为将会导致行为人受到不授予/撤销学位、开除学籍等处理（处分）决定。本人如果被查证在撰写本毕业论文过程中存在学位论文作假行为，愿意接受学校依法作出的处理（处分）决定。

承诺人签名：張川

日期：2020年6月9日

# 目录

目录.....	i
摘要.....	iii
Abstract.....	iv
第一章 引言.....	1
1.1 研究背景与相关工作.....	1
1.2 本文贡献.....	3
第二章 理论.....	4
2.1 量化多因子的理论基础.....	4
2.1.1 资本资产定价模型.....	4
2.1.2 套利定价理论与多因子模型.....	4
2.2 逻辑回归与归一化指数回归.....	5
2.3 随机森林.....	6
2.4 长短期记忆神经网络.....	7
2.4.1 长短期记忆神经网络基本原理.....	7
2.4.2 堆叠长短期记忆神经网络.....	8
2.5 日内效应.....	9
2.5.1 日内效应的影响.....	9
2.5.2 日内平均法.....	9
第三章 数据.....	11
3.1 数据概况.....	11
3.2 数据处理与特征构建.....	12
3.2.1 特征处理.....	12
3.2.2 特征添加.....	13
3.2.3 标准化.....	15
3.2.4 数据集划分和输入输出.....	15
3.3 高频数据特征.....	16

3.3.1 高频数据基本特征.....	16
3.3.2 高频数据的日内效应.....	17
第四章 实验与结果分析.....	20
4.1 回测框架.....	20
4.2 日内效应调整.....	21
4.3 神经网络方法.....	23
4.4 模型对比总结.....	25
第五章 总结与展望.....	27
第六章 参考文献.....	28
第七章 致谢.....	31

## 摘要

随着计算机发展，电子高频交易正在逐步取代传统人工交易成为主流，而机器学习算法的预测能力为高频交易策略提供了新的思路。目前，将基于机器学习算法的高频交易策略应用在期货市场上的研究比较成熟，但应用在国内股票市场上的研究较少。因此，本文从三个主要机器学习流派中各挑选一个代表模型，包括“统计学习”的逻辑回归模型、“符号主义学习”的随机森林模型和“连接主义学习”的LSTM神经网络模型，通过实验对它们在国内股票高频快照数据上的趋势预测能力进行了比较分析。同时，由于部分高频数据的日内效应对前两种模型训练的影响较大，为了使模型具有更好的可比性，本文还运用了日内平均法来消除日内效应。回测结果表明，消除日内效应能够有效提升逻辑回归和随机森林模型的趋势预测能力；随机森林的预测准确率最差；逻辑回归的预测准确率最好但捕获率最低；LSTM最灵活，通过调节超参数可以调整预测准确率与捕获率大小，但对训练成本和训练样本量的要求也最高。

关键词：机器学习，LSTM神经网络，日内效应，高频交易，量化投资

## Abstract

With the development of computing technology, electronic and high-frequency trading are gradually replacing the traditional manual trading, and the predicting power of algorithms based on machine learning helps to build new high-frequency trading strategies. The research on the application of the high-frequency trading strategies based on machine learning algorithms for the futures market has become relatively mature, but there is little research coverage for the application on the domestic stock. Therefore, this paper selects one representative model from three mainstream machine learning schools, including logistic regression from statistical learning, random forest from symbolism and LSTM from connectionism, and then compares their trend prediction ability. Besides, because the Intraday Effect of high-frequency data will reduce the effectiveness of the first two models, to make them more comparable, this paper also applies the Intraday Average Method method to eliminate the Intraday Effect. The back-test results show that the elimination of Intraday Effect can effectively improve trend prediction ability of logistic regression and random forest models. It is also found that random forest model has worst precision, logistic regression has highest precision but worst recall and LSTM shows greatest flexibility, which means it can balance the precision and recall by changing the structure but meanwhile requires more training samples and higher training cost.

Key words: Machine Learning, LSTM, Intraday Effect, High-frequency Trading, Quantitative Investment

# 第一章 引言

## 1.1 研究背景与相关工作

随着计算机技术的发展和大数据概念的出现，基于机器学习、人工智能等前沿技术的量化投资新方法开始兴起。量化投资旨在通过数学模型和计算机技术制定一套严格的交易策略，克服人工交易中的情绪因素，同时能够捕捉到市场中稍纵即逝的微小交易机会。而利用机器学习技术的预测能力，可以辅助量化策略获得更优的交易决策能力，进而打败交易对手捕捉到更多的获利机会。

市场中存在大量交易者，他们实行着许多不同的交易策略。在一些市场中，利用少量理想化的假设和严谨的数学推导，交易者们能够为资产进行实时定价，例如由 Black, Scholes 和 Merton 提出的期权定价公式[1]以及李祥林提出的利用 copula 函数对信用衍生品进行定价的方法[2]，因而交易者们可以参考模型定价结果对于偏离定价的资产进行套利交易；但在另一些市场，例如期货市场或股票市场，资产价格波动主要受到各对手方供需关系的影响，因此难以通过数学模型进行实时定价。在这些市场中，根据 Ross 提出的套利定价理论[3]，只要市场未达到均衡状态，市场上就存在无风险的套利机会。他还假设风险资产均衡收益可以由多个因素来近似线性地解释，这为量化多因子模型获取市场超额回报（alpha）提供了理论基础。因子挖掘是目前量化领域的热门，但是易于发现的有效因子往往会受到大量资金的追捧，进而波动率加大，出现收益的减弱甚至大幅回撤，这被称为因子拥挤现象（factor crowding）。在低频数据中找寻新的有效因子越来越困难，投资者需要从更高频、更详尽的数据中获取到更多信息。

分笔数据（tick data）是一类频率高至秒甚至毫秒级的市场数据，它通常被高频交易者们利用以进行持仓时间很短、换仓频率很高的交易，这样的交易每笔收益通常很低，但依赖高性能计算机强大的算力和高速网络的通信能力，通常能以远超人力的速度完成大量订单交易，从而积少成多获得不错的收益。高频交易策略最早诞生于二十世纪末的美国，伴随着电子化自动交易替代手工交易的时代进程而出现，发展到目前在美国股票市场已经相当成熟，市场份额占据日均交易量的半数以上。在国内由于市场特征不同，高频交易方兴未艾。

机器学习（machine learning）是一门致力于研究如何通过计算的手段，利用经验来改善系统自身性能的学科。自上世纪五十年代诞生以来，机器学习领域涌现出众多流派，目前依然盛行的主要包括符号主义学习（symbolism）、

统计学习 (statistical learning) 和连接主义学习 (connectionism) 等 [4]。符号主义学习的代表是决策树 (decision tree) 和基于决策树的随机森林 (random forest) 等, 优点是逻辑清晰; 统计学习的代表性技术包括各种回归模型 (regression)、支持向量机 (Support Vector Machine, SVM) 以及更一般的核方法 (kernel methods), 优点是可解释性强、有坚实的数学理论基础; 连接主义就是基于神经网络的学习方法, 现在也被称为“深度学习”

(deep learning) [5], 常见的模型包括卷积神经网络 (Convolutional Neural Network, CNN), 循环神经网络 (Recurrent Neural Network, RNN), 长短期记忆神经网络 (Long Short-Term Memory, LSTM) 等 [6], 优点是能够自动提取深层特征, 模型可塑性强。机器学习技术被广泛应用于图像识别 [7] [8] 和自然语言处理 [9] [10] 等领域, 又因为其良好的预测效果, 近年来被越来越多地应用到量化交易中。

早在 1988 年, White 就将神经网络技术应用在日度股票数据的预测上 [11]。之后量化研究者们广泛尝试了各种基于机器学习的交易策略, 由于利益原因只有一部分非常经典的模型被发表了出来。Dunis 和 Nathani 在 2007 年发现多层感知机 (Multilayer Perceptron, MPP)、高阶神经网络 (Higher Order Neural Networks, HONN) 和 K 近邻方法 (k-nearest) 在黄金和白银期货上具有比线性模型更好的超额收益 [12], 这显示该市场的时间序列中存在着非线性关系。Persio 和 Honchar 在此基础上将多层感知机、卷积神经网络和循环神经网络应用到标普 500 指数上进行价格预测 [13]。Choundhry 等人则将支持向量机模型成功应用于美国股票市场的趋势预测 [14]。Maknickiene 和 Maknickas 证明了 LSTM 在欧元兑美元的汇率上具有良好的预测效果 [15]。在国内, 相关研究近年来也在增多, 比如袁祥枫研究了 LSTM 模型在国内商品期货高频数据上的预测表现 [16], 张贵勇验证了 CNN 和 SVM 的混合模型在国内股票指数和汇率上的预测有效性 [17], 但机器学习在股票高频领域的研究尚且较少。

魏瑾瑞、朱建平和谢邦昌分析了国内股票市场高频数据的统计特征, 指出受市场微结构噪声、跳跃等因素影响, 高频数据也并非优质的时间序列, 而是具有离散价格、价格惰性和同时交易等特点 [18]。王维国和余宏俊则研究了国内股票市场超高频数据的日内效应, 并比较了日内加权平均法 (Intraday Average Method, IAM)、核估计法和自组织映射 (Self Organizing Mapping, SOM) 神经网络等日内效应调整的方法 [19]。这些研究都为进一步进行股票高频交易策略的研究提供了参考。

目前国内对机器学习高频交易策略的研究主要针对期货市场, 例如孙达昌和毕秀春将基于 CNN 和 LSTM 的高频交易策略应用于沥青期货主力合约, 获得了



较好的泛化能力[20]。随着国内金融市场开放，未来 A 股市场的做空机制和交易接口会更加完善，而股票市场的资金容量远高于期货市场，从国外成熟市场的经验看来国内股票市场高频策略还有很大的发展空间，因此做一些前瞻性的研究很有价值。

## 1.2 本文贡献

本文使用了信息比较完整、采集质量较好的国内股票市场 L2 高频数据，重点研究了以下两点：第一，简要分析国内股票市场高频数据的特点；第二，探究不同流派的机器学习方法在这一时间序列中的预测效果。

本文以上交所交易的贵州茅台（600519.SH）这支股票为例，简要分析了高频数据尖峰厚尾分布、日内效应等特点，并添加了新的技术指标特征；针对在训练机器学习模型中影响很大的日内效应，采取了日内平均法进行消除处理，去除周期性因素，改善了部分特征的数据分布。

本文采用了简单的固定持仓时间交易策略进行回测，旨在探究不同流派机器学习模型在该金融时间序列上的预测效果。本文选取的机器学习模型包括“统计学习”的逻辑回归模型、“符号主义学习”的随机森林模型和“连接主义学习”的长短期记忆神经网络模型，涵盖了目前主流的三个流派。在逻辑回归和随机森林模型上尝试结合日内平均法消除日内效应，使预测效果大大提升；给长短期记忆神经网络设计了更复杂的堆叠结构，获得了更高的多头信号捕获率。本文最后对比了各模型在该高频数据集上的预测性能，分析了各自的优点和不足。

本文使用 Python 语言分析沪市高频数据并构建交易策略和回测框架，机器学习模型调用 Sklearn 和 Keras 工具包进行搭建。

## 第二章 理论

### 2.1 量化多因子的理论基础

#### 2.1.1 资本资产定价模型

资本资产定价模型 (Capital Asset Pricing Model, CAPM) 最早由 Sharpe[21]、Lintner[22]和 Mossin[23]等人在现代投资组合理论的基础上发展而来。它假设在理想的市场中, 所有投资者都是理性的且他们的信息相同: 他们都用期望收益率和标准差来构建投资组合, 同时他们对于各证券的期望收益率、标准差及相关性都有相同的预期, 资本和信息的流通自由。投资者都会偏向更小的风险和更高的回报 (risk averse), 但是不同的人对风险有不同的承受能力。由于货币的时间价值, 资产收益一部分是无风险的, 而另一部分则是风险收益。对于那些愿意冒险的人, 承担更高风险的预期回报是更高的期望收益率。CAPM 告诉我们, 决定单个资产或证券组合的期望收益率与风险之间均衡关系的定价模型为:

$$r_i = r_f + \beta_i(r_m - r_f) \quad (2.1)$$

$$\beta_i = \frac{\text{cor}(i, m)}{\sigma_m^2} \quad (2.2)$$

其中 $r_i$ 是该资产或组合的预期收益率 (expected return),  $r_f$ 是无风险回报率 (risk-free rate),  $\beta_i$ 反映该资产或组合与整体市场组合的联动性,  $r_m$ 是整体市场组合的期望回报率 (expected market return),  $(r_m - r_f)$ 称为该资产或组合的市场风险溢价 (market risk premium)。CAPM 通过简单的模型描述了收益与风险之间的定量关系, 从而成为现代金融学的基石之一。

#### 2.1.2 套利定价理论与多因子模型

CAPM 虽然形式简单, 但前提假设太多: 理性人假设与事实偏离, 整体市场组合难以度量等等。因此, Ross 于 1976 年发展出了假设更简单的套利定价理论[3]。该理论只有三个基本假设:

- 1) 因素模型能够描述证券收益;
- 2) 市场上有足够多的证券来分散风险;
- 3) 完善的市场不允许任何套利机会的存在。

因素模型在数学上类似于统计学中的线性回归模型，它假设证券收益受一种或多种因素的影响，数学上写作：

$$r_i = a_i + b_{i,1}F_1 + b_{i,2}F_2 + \cdots + b_{i,n}F_n + \epsilon_i \quad (2.3)$$

其中 $F_j$ 表示各种因素， $b_{i,j}$ 表示证券 $i$ 对因素 $F_j$ 的敏感度， $\epsilon_i$ 表示证券 $i$ 特有的扰动项。假设因素模型成立，那么就可以通过挖掘市场中潜在的有效因子，从而对证券进行合理定价。

套利定价理论的第（2）条假设中，如果市场上证券数量趋近于无穷多，那么非系统性风险可以被完全分散，系统的非系统扰动 $\epsilon_i$ 也会趋近于 0，此时通过整理单因素模型可以得到与 CAPM 模型相同的公式和结论。

套利定价理论的第（3）条假设条件说明，非均衡状态的市场上总是会存在无风险套利机会，套利行为会最终帮助形成均衡价格。由于现实中的市场几乎不可能是完全均衡状态的，因此就始终存在大量套利机会。前面提到因素模型是套利定价理论的基本假设之一，如果能够寻找到合适的市场因素，从而正确定价证券资产，那么就可以找到市场价格偏离的证券，从而在其价格回归后实现套利。

多因子模型因此成为量化投资领域的主流方法之一。量化多因子的主要工作是挖掘新的因子，但实际上大量因子之间往往都存在多重共线性，使用简单的线性多因素模型并不能很好地解决这一问题，而筛选掉相关性高的因子又不可避免地会损失一部分信息，因此如何有效解决这一问题也是重要的研究方向。而机器学习的一些方法，例如具有自动挖掘深层特征特点的神经网络方法等，为建立更有效的新模型提供了思路。

## 2.2 逻辑回归与归一化指数回归

逻辑回归（logistic regression）是一种常用的分类模型。逻辑回归是广义线性模型（generalized linear model）的一种，广义线性回归的数学公式是：

$$y = g^{-1}(w^T x + b) \quad (2.4)$$

其中 $g^{-1}(\cdot)$ 称为联系函数（link function），当联系函数是逻辑斯蒂函数（logistic function）时，模型就是逻辑回归模型。其数学形式是：

$$y = \frac{1}{1 + e^{-(w^T x + b)}} \quad (2.5)$$

它的两端极限值分别趋近于 0 和 1，因此非常适合处理二分类问题。但是当面对多分类问题时，需要在逻辑回归的基础上进行推广，新的函数称为归一化指数函数（softmax function）：

$$h_w(\mathbf{x}) \equiv \begin{bmatrix} p(y=1|\mathbf{x}; \mathbf{w}) \\ p(y=2|\mathbf{x}; \mathbf{w}) \\ \vdots \\ p(y=k|\mathbf{x}; \mathbf{w}) \end{bmatrix} = \frac{1}{\sum_{i=1}^k e^{\mathbf{w}_i \mathbf{x} + b_i}} \begin{bmatrix} e^{\mathbf{w}_1 \mathbf{x} + b_1} \\ e^{\mathbf{w}_2 \mathbf{x} + b_2} \\ \vdots \\ e^{\mathbf{w}_k \mathbf{x} + b_k} \end{bmatrix} \quad (2.6)$$

归一化指数函数的输出是一个  $k$  维的向量，对应了第  $k$  个类别的概率大小，预测结果通常就取其中最高概率对应的类别。利用最大似然估计，建立模型的对数似然代价函数（log-likelihood loss function）为：

$$J(\mathbf{w}, \mathbf{b}) = -\frac{1}{m} \sum_{j=1}^m \sum_{l=1}^k \mathbb{I}\{y^{(j)} = l\} \log \left( \frac{e^{\mathbf{w}_l \mathbf{x} + b_l}}{\sum_{i=1}^k e^{\mathbf{w}_i \mathbf{x} + b_i}} \right) \quad (2.7)$$

其中  $m$  代表训练样本个数， $\mathbb{I}\{\cdot\}$  是示性函数。容易发现在二分类时，该代价函数就是交叉熵代价函数（cross-entropy loss function）。该代价函数没有显式解，因此通过梯度下降方法进行优化，公式为：

$$\frac{\partial J(\mathbf{w}, \mathbf{b})}{\partial \mathbf{w}_i} = -\frac{1}{m} \sum_{j=1}^m (\mathbf{x}^{(j)} - p(y^{(j)} = i | \mathbf{x}^{(j)}) * \mathbf{x}^{(j)}) \quad (2.8)$$

$$\mathbf{w}_i = \mathbf{w}_i - \alpha \frac{\partial J(\mathbf{w}, \mathbf{b})}{\partial \mathbf{w}_i} \quad (2.9)$$

## 2.3 随机森林

随机森林（random forest）是 Breiman 在 2001 年提出的一种基于决策树（decision tree）的集成学习（ensemble learning）模型[24]，因其简单的结构和良好的泛化能力而受到欢迎。

随机森林的个体单元是决策树。决策树基于树模型，模拟人类做判断时的思维流程，在每个结点上通过某个特征进行判断，再通过层层判断最终实现分类。在判断的过程中，优化的目标是结点的纯度（purity），即希望经过该结点分类之后的叶子结点上的样本都尽量属于同一类。纯度有多种评价指标，常用的包括信息熵（information entropy）和基尼系数（Gini index），使用后者的决策树通常称为 CART 决策树（Classification and Regression Tree）[25]。

但是对于复杂的问题，单个决策树的效果通常并不好，被称为弱学习器（weak learner）。虽然个体分类器是弱学习器，但是如果有许多不同的弱分类器，将它们的分类结果再通过投票法（voting）来得到最终结果，就可能会得到更好的分类器。同时产生不存在强依赖关系的若干弱分类器再投票进行集成学习的代表方法就是随机森林。

具体来说，随机森林在生成个体决策树时，对训练样本集和参与训练的属性集都进行随机采样，在个体决策树的生成上引入了很大的随机性。由于这种方法是通过随机的方式构建一系列互不关联的决策树，因此被形象地称作“随机”的“决策树森林”。Breiman 证明了随机森林在理论上的有效性，同时指出划分随机属性集时推荐的取样数量是  $\log_2 d$ ，其中  $d$  是总属性的个数。

## 2.4 长短期记忆神经网络

### 2.4.1 长短期记忆神经网络基本原理

长短期记忆神经网络（Long Short-Term Memory, LSTM）是一种特殊的循环神经网络（Recurrent Neural Network, RNN），最早由 Gers 等人在 1999 年提出[26]。由于它对于之前的信息具有记忆和选择性遗忘效果，因此常用于处理序列数据。LSTM 相较于普通 RNN 有长期记忆和短期记忆两种传输状态，其中  $c^t$ （cell state）改变很慢，而  $h^t$ （hidden state）改变很快；同时由于增加了忘记门控和选择记忆门控，因此能够选择性地“记住”重要信息，从而有效解决长序列训练过程中的梯度消失和梯度爆炸问题。

$$\begin{aligned}
 z^f &= \sigma \left( W^f \begin{bmatrix} x^t \\ h^{t-1} \end{bmatrix} \right) & z^i &= \sigma \left( W^i \begin{bmatrix} x^t \\ h^{t-1} \end{bmatrix} \right) \\
 z^o &= \sigma \left( W^o \begin{bmatrix} x^t \\ h^{t-1} \end{bmatrix} \right) & z &= \tanh \left( W \begin{bmatrix} x^t \\ h^{t-1} \end{bmatrix} \right)
 \end{aligned}$$

图 2.1 LSTM 的四个状态计算单元示意图

LSTM 的内部结构首先是四个状态  $z^f$ 、 $z^i$ 、 $z^o$  和  $z$  的计算（如图 2.1）。前三个状态通过 sigmoid 激活函数映射到 0 到 1 之间，作为门控；最后一个  $z$  通过 tanh 激活函数映射到 -1 到 1 之间，作为输入数据。

运用到这四个状态，LSTM 每个单元的具体结构如图 2.2。内部计算流程可以分为三个阶段：阶段一选择性忘记前一个节点的输入，应用的门控是  $z^f$ ；阶

段二对新的输入进行选择记忆，应用的门控是 $z^i$ ，同时输入数据 $z$ ；阶段三决定当前状态的输出，应用的门控是 $z^o$ 。

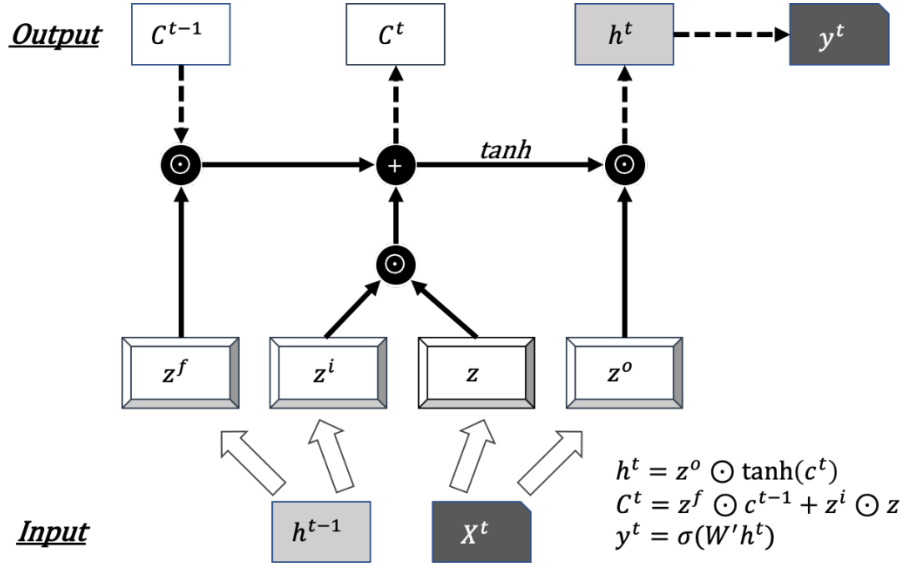


图 2.2 LSTM 内部结构示意图

#### 2.4.2 堆叠长短期记忆神经网络

单层的 LSTM 神经网络用于处理复杂的问题往往泛化能力不足，因此可以考虑使用多层 LSTM 神经网络提高模型的复杂程度，从而提升模型泛化能力。

对于多层 LSTM 神经网络有多种构造方式，常见的包括堆叠长短期记忆神经网络（Stack LSTM）[27]、网格长短期记忆神经网络（Grid LSTM）[28]等。本文采用了堆叠长短期记忆神经网络。

堆叠 LSTM 的思想就是运用多个单层 LSTM 作为隐藏层，上一层 LSTM 输出一个三维数组作为下一层 LSTM 的输入，最后再使用一层全连接层（Dense layer）将结果转化为一维输出，其基本结构如图 2.3。

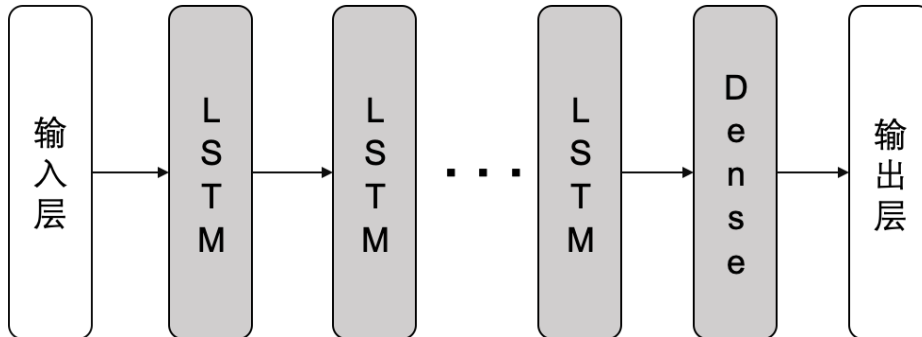


图 2.3 堆叠 LSTM 结构示意图

## 2.5 日内效应

对于高频数据来说，交易量、交易额等数据普遍存在一种以交易日为单位的周期性日内运动模式。Wood 等人最早发现纽交所在每日开盘和收盘时段的交易频率往往明显高于其他时段[29]，因此每日内这类数据几乎都呈现出“U”形的分布，且以交易日为单位存在周期性，这种特点被称为日内效应（Intraday Effect）。在 A 股市场，由于存在午间休市，因此更合适的描述是“W”形特征[19]。

### 2.5.1 日内效应的影响

考虑一个单因素模型

$$y_t = b_x x_t + \epsilon_t \quad (2.10)$$

$$y_t = b_x x_t + b_s d_t + \epsilon_t \quad (2.11)$$

其中  $d_t$  表示周期性虚拟变量。方程（2.10）是不考虑日内效应的单因素模型，方程（2.11）是考虑了日内效应周期性因素的模型。对比发现，假如日内效应没有被考虑，那么单因素模型中的扰动项实际上等于

$$\epsilon'_t = b_s d_t + \epsilon_t \quad (2.12)$$

这时待估参数  $b_x$  的一阶矩和二阶矩分别为：

$$E(b_x) = b_x + E((x'_t x_t)^{-1} x'_t d_t) b_s \quad (2.13)$$

$$Var(b_x) = E((x'_t x_t)^{-1}) \sigma^2 + [E((x'_t x_t)^{-1}) \sigma_s^2 + Var((x'_t x_t)^{-1} x'_t d_t)] b_s^2 \quad (2.14)$$

由此可见，当存在日内效应，即  $b_s \neq 0$  时，待估参数的一阶矩和二阶矩都会由于日内效应而有偏。多因素模型也会得到相似的有偏结果。因此对于线性因素模型来说，消除日内效应对于提升模型精度很有必要。

### 2.5.2 日内平均法

日内效应的调整方法中，比较简单的是日内平均法。因为日内效应是以交易日为单位周期性出现的，因此最简单的消除办法就是找到日内效应的周期性规律并在数据上减去规律变化的部分。由于高频数据的时间间隔是非连续变化的，不同日的数据难以在真实时间刻度上对齐，因此一种常见的处理办法是将每日交易时间分成  $m$  个时段，然后计算数据序列  $X_t$  在每个时段内的平均值

$$\bar{X}_k = \frac{1}{K} \sum_{i=k}^{k+K} X_i, \quad k = 1, 2, \dots, m \quad (2.15)$$

对变量进行日内效应调整时只需要把对应时段该变量的平均值减去即可。



## 第三章 数据

### 3.1 数据概况

本文使用的数据来自通联数据提供的沪深 L2 高频行情数据库。该数据库记录了沪深两市股票交易的行情数据、交易数据、订单数据等，3 秒左右记录一个数据快照。本文从中抽取出了单支股票贵州茅台（600519.SH）的市场快照数据进行研究，该支股票的流动性较好。

本文选取了 2017 年 6 月 7 日到 2017 年 7 月 7 日共记 23 个交易日的 112560 条数据作为训练样本（训练集和验证集），将 2017 年 7 月 8 日到 2017 年 8 月 8 日共 22 个交易日的 106242 条数据作为测试样本。数据示例如表 3.1。

表 3.1 数据示例

字段	数据 1	数据 2	数据 3
上传时间	09:30:02.0	09:30:06.0	09:30:09.0
市场价格	449.54	449.43	449.43
总交易笔数	60	75	75
总交易量	17450	20350	20350
总交易额	7838327	9141389	9141389
总委买量	167550	182550	187550
总委卖量	116984	139284	140984
加权平均委买价	441.971	442.333	442.514
加权平均委卖价	458.207	457.696	457.696
买一价	449.50	449.43	449.43
买一量	200	200	200
卖一价	449.60	449.60	449.50
卖一量	100	100	100

训练集上的价格趋势如图 3.1。可以看到价格的短期波动非常明显，因此为高频交易赚取短期价差提供了可能。

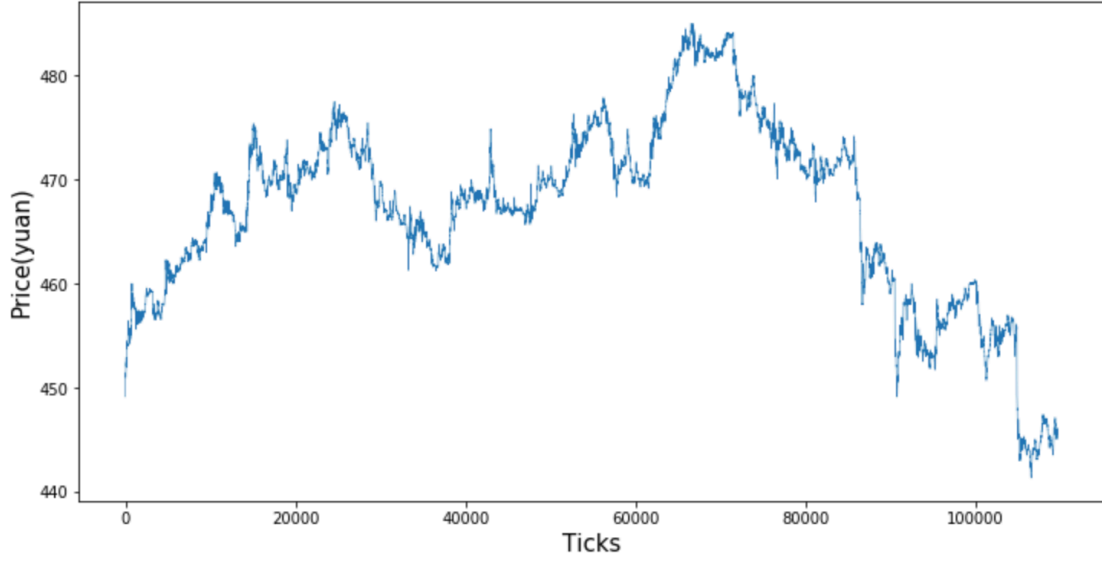


图 3.1 训练集上股票价格走势

## 3.2 数据处理与特征构建

### 3.2.1 特征处理

原始数据中一些特征需要进行预处理，方便后续模型训练。特征处理主要包括两方面，第一是对一些累积型的特征做差分处理，第二是给训练数据贴上标签 $y$ 。

首先需要将累积值转换为每个 tick 对应的变化值，从而满足模型假设、方便模型训练。因此对总交易量、总交易额、总交易笔数、总买单量和总卖单量进行差分处理。

其次，为了训练有监督机器学习模型，需要给训练数据提前加上标签。固定时间 $\tau$ 之后的股票价格涨幅记为

$$\Delta p_{percent}^{t_0} = \frac{p_{t_0+\tau} - p_{t_0}}{p_{t_0}} \times 100\% \quad (3.1)$$

本文采取了固定持仓时间 $\tau$ 的交易策略，因此只需要根据固定时间 $\tau$ 之后的股票价格涨幅 $\Delta p_{percent}^t$ 是否超过设定的阈值 $\Delta p_{threshold}$ 来对 $t$ 时刻贴标签：

$$\begin{cases} \Delta p_{percent}^t \geq \Delta p_{threshold}, & y_t = 1 \\ \Delta p_{percent}^t < \Delta p_{threshold}, & y_t = 0 \end{cases} \quad (3.2)$$

其中 $y_t$ 表示 $t$ 时刻对应的标签，标签值为 1 代表买多信号，标签值为 0 代表保持持仓不变。如果考虑到买空操作可行的情况，则变成三类信号：买多、买空、保持。此时标签相应地变为：

$$\begin{cases} \Delta p_{percent}^t \geq \Delta p_{threshold}, & y_t = 1 \\ \Delta p_{percent}^t \leq -\Delta p_{threshold}, & y_t = -1 \\ -\Delta p_{threshold} < \Delta p_{percent}^t < \Delta p_{threshold}, & y_t = 0 \end{cases} \quad (3.3)$$

其中 $y_t = -1$ 代表买空信号。只买多的情况下问题简化为一个二分类问题，同时买多和买空的情况下则是一个多分类问题。

关于固定持仓时间 $\tau$ 的选择，我们希望持仓时间尽量短的同时保证足够多的交易机会。根据随机游走的模型，预测的时间间隔越大，预测值分布的方差也越大；增加持仓时间也会增大持仓的风险敞口，因此持仓时间越短越好。但是由于交易成本的存在，只有资产涨幅超过交易成本后才能获利，就必须适当增加持仓时间来捕捉更多涨幅超过交易成本的交易机会。若只考虑交易的固定成本，目前上交所的经手费为千分之 0.05，印花税为千分之 1，券商的佣金普遍在千分之 0.2 左右，因此每笔买卖的总成本约为千分之 1.25，再考虑到冲击成本等非固定交易成本，本文假设只有当资产价格涨幅超过千分之 2 时，才能够稳定获利，即 $\Delta p_{threshold}$ 取值为 0.2%。

数据中 tick 之间的时间间隔大多在 3 至 5 秒，选择 50 个、100 个和 200 个 tick 作为固定持仓时间进行测试，在训练集上的结果如表 3.2。在保证多空头信号占比较多的情况下希望持仓时间尽量短，因此本文选取 100 个 tick 作为固定持仓时间，持仓时间在自然时间上约为 5 至 7 分钟，在训练集上买多信号、持仓不变信号和买空信号数量的比例大约为 1:8:1。

表 3.2 不同固定持仓时间下信号数量

固定持仓时间 (ticks)		50	100	200
买多信号	数量	5989	12471	19816
	占比	5.5%	11.4%	18.1%
买空信号	数量	5522	11860	19388
	占比	5.0%	10.8%	17.7%

### 3.2.2 特征添加

除了原数据集提供的 12 个特征，为了提高模型精度，同时反映出时间序列顺序演变的特点，需要再构建一些新的指标。新添加的指标主要是技术分析指标，如表 3.3 所示。

表 3.3 新添加的技术指标

变量	指标	释义	变量	指标	释义
交易量	MA5	5 期移动平均	价格	MA5	同左
	MA10	10 期移动平均		MA10	
	MA20	20 期移动平均		MA20	
交易笔数	MA5	同上		K	KDJ 指标
	MA10			D	
	MA20			J	
委买数量	MA5	同上		DIF	MACD 指标
	MA10			DEA	
	MA20			MACD	
委卖数量	MA5	同上			
	MA10				
	MA20				

移动平均 (Moving Average, MA) 指标是处理时间序列最常用的方法之一, 对于  $t$  时刻的变量  $X_t$ , 他的  $k$  期移动平均通过下式计算

$$MA(X_t, k) = \frac{1}{k} \sum_{i=0}^k x_{t-i} \quad (3.4)$$

如果用指数递减的加权方式代替等权重, 则变成指数移动平均 (Exponential Moving Average, EMA), 计算公式变为

$$EMA(X_t, k) = \frac{2}{k+1} \sum_{i=0}^{\infty} \left(\frac{k-1}{k+1}\right)^i x_{t-i} \quad (3.5)$$

KDJ 指标又叫做随机指标, 它通过一个特定周期  $n$  内变量出现的最高值、最低值和最后一个计算周期的最后值以及三者间的比例关系来计算最后一个计算周期内的未成熟随机值 RSV, 再通过移动平均计算 K 值、D 值和 J 值。数学上计算公式写作

$$RSV_t = \frac{X_{n,close} - X_{n,low}}{X_{n,high} - X_{n,low}} \quad (3.6)$$

$$K_t = MA(RSV_t, m_1) \quad (3.7)$$

$$D_t = MA(K_t, m_2) \quad (3.8)$$

$$J_t = 3K_t - 2D_t \quad (3.9)$$

本文中参数取值为  $n = 9$ ,  $m_1 = 3$ ,  $m_2 = 3$ 。

指数平滑移动平均线 (Moving Average Convergence/Divergence, MACD) 指标由差离值 (Difference, DIF)、指数平均差离值 (Difference Exponential Average, DEA) 和 MACD 三个值构成。DIF 具体指的是  $N^{short}$  期快速指数移动平均值与  $N^{long}$  期慢速指数移动平均值的差值, 用以表示短期平均值与较长期平均值之间的关系, 对于涨跌趋势有一定的指示作用; DEA 是 DIF 的

$N^{mid}$ 期指数移动平均，它是对差离值的进一步平滑，代表 DIF 值的中长期变化；最后 MACD 就是 DIF 与 DEA 差值的倍数，表示了 DIF 值短期相较于中长期的变化趋势。具体计算公式如下

$$DIF_t = EMA(X_t, N^{short}) - EMA(X_t, N^{long}) \quad (3.10)$$

$$DEA_t = EMA(DIF_t, N^{mid}) \quad (3.11)$$

$$MACD_t = (DIF_t - DEA_t) \times 2 \quad (3.12)$$

本文中参数取值为  $N^{short} = 12$ ,  $N^{mid} = 9$ ,  $N^{long} = 26$ 。

添加后共有 33 个特征，包括 12 个原始特征和差分后的特征，以及 21 个新增的衍生特征。

### 3.2.3 标准化

标准化与归一化都旨在将不同特征映射到一定范围内增加可比性，同时使得梯度下降的收敛速度更快、更容易收敛到全局最优解，有助于提高分类器的精度。

标准化方法中最常用的是 z-score 标准化，方法类似于构造 z 统计量，计算方法如下

$$X^* = \frac{X - \mu}{\sigma} \quad (3.13)$$

其中  $X$  是原始变量， $\mu$  是原始变量的算术平均值， $\sigma^2$  是原始变量的样本方差。通过这样的映射，数据的分布会接近于标准正态分布，从而使不同取值范围的变量具有更好的可比性。

比较标准化方法，另一种常用的处理方法是归一化。归一化最常用的是最大最小值归一化，方法如下

$$X^* = \frac{X - x_{min}}{x_{max} - x_{min}} \quad (3.14)$$

通过这样的映射，可以将变量的数值缩放到  $[0,1]$  范围之内。

通过章节 3.3 中的讨论，发现高频数据的特征分布具有厚尾（heavy-tailed）的特点，容易出现异常值，如果使用归一化方法会使得中值附近的数据过于拥挤，因此本文采用 z-score 标准化方法进行处理。

### 3.2.4 数据集划分和输入输出

对于逻辑回归和随机森林模型，将训练集整体放入模型训练即可；但对于 LSTM 模型，它需要输入一个三维的数组，同时神经网络模型的训练需要将训练

数据划分为训练集 (training set) 和验证集 (validate set)，因此要求对输入数据进行预处理。

LSTM 的输入数据结构是这样一个三维数组：

(样本数, 时间步长, 特征个数)

本文中时间步长设置等于预测步长，即 100 个 tick；样本数等于时间窗口的个数；特征个数就是放入模型的变量个数，本文中为 33 个。

所有模型的输出都是一个一维向量  $(y_{t+1}, y_{t+2}, \dots, y_{t'})$ ，表示从预测的起始点  $t$  到预测的最后一个点  $t'$  对应的预测信号， $y_t$  的取值如章节 3.2.2 中提到的。需要注意的是，在高频数据中  $t$  表示的是 tick 值，因为每个 tick 之间的时间间隔往往不同，因此不等于自然时间。

训练 LSTM 模型时，选取训练数据的前 80% 共 87548 个样本作为训练集，后 20% 共 21811 个样本作为验证集。

### 3.3 高频数据特征

#### 3.3.1 高频数据基本特征

高频数据的基本特征通常包括以下几点：

- 1) 数据点的时间间隔不同；
- 2) 包含市场微观结构，包括交易量、交易额、买一价、卖一价、买一量、卖一量等等；
- 3) 价格变化更接近于随机游走；
- 4) 交易量等数据存在日内分布不均的周期性特点。

其中第 4 点被称作高频数据的日内效应，在章节 3.3.2 中专门讨论。

由于数据点之间的时间间隔不同，因此高频数据对于自然时间来说是离散的数据。在研究高频数据时一个自然的做法就是用 tick 来代替自然时间，通常 tick 指的是订单变化的时刻，因此 tick 行情又被称为逐笔行情。tick 行情虽然在自然时间上是离散的，但它很好地反映了市场的变化，因为市场只在新的订单发生时才变化。本文采用的是对 tick 行情数据在某些时刻切片的数据，严格意义上应称为快照数据 (Snapshot)，但在国内通常不区分二者的叫法，而把它们都统称为 tick 数据，因此后文在描述时也把每个采样点称为一个 tick。本文的数据集上每个 tick 之间的时间间隔大多在 3~5 秒之间。

高频数据记录了市场的微观结构，包括订单的详情，订单的数据越详细则数据的质量越好。通常认为价格最高的委买订单以及价格最低的委卖订单对于市场走势的影响最直接，因此本文主要使用了每个 tick 的买一价、买一量、卖

一价和卖一量，同时使用所有订单根据委托交易的股数进行加权平均得到的加权平均委买价和加权平均委卖价来反映委买和委卖双边的整体行情。

高频数据的市场价格接近于随机游走。比如每 100 个 tick 的收益率分布，如图 3.2，是一个典型的尖峰厚尾分布，大致关于 0 对称。这说明虽然市场价格近似随机游走，但并非布朗运动（Brownian motion），而是更接近于具有尖峰厚尾步长分布的莱维飞行（Levy flight）。

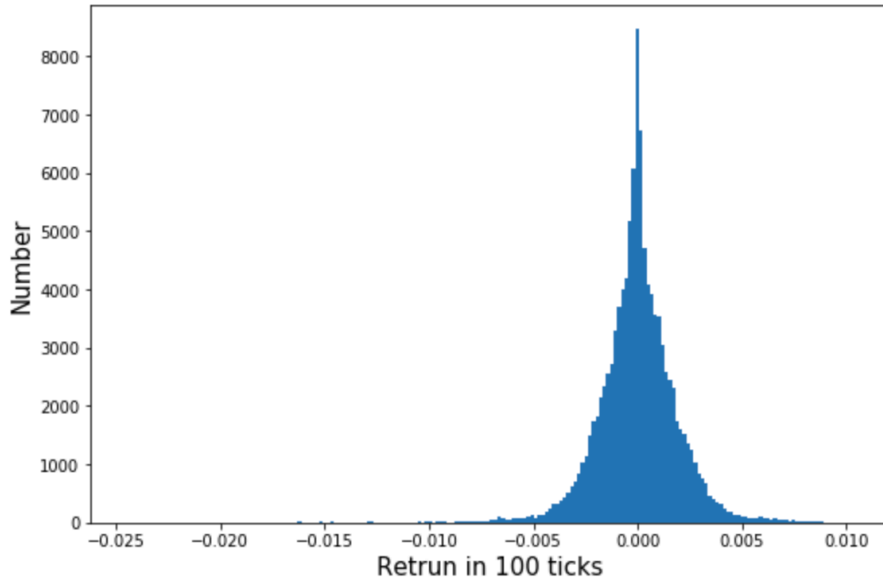


图 3.2 100 个 tick 后收益率分布图

### 3.3.2 高频数据的日内效应

如章节 2.5 中讨论的，如果不消除周期性因素的影响，将会对统计预测的结果产生很大影响。

在本文的数据集上同样可以观察到明显的日内效应现象。以 2017 年 6 月 7 日的交易额数据为例，每个 tick 交易额的分布如图 3.3，可以看到在上下午交易时段的开盘和收盘阶段交易额明显高于其他时段。

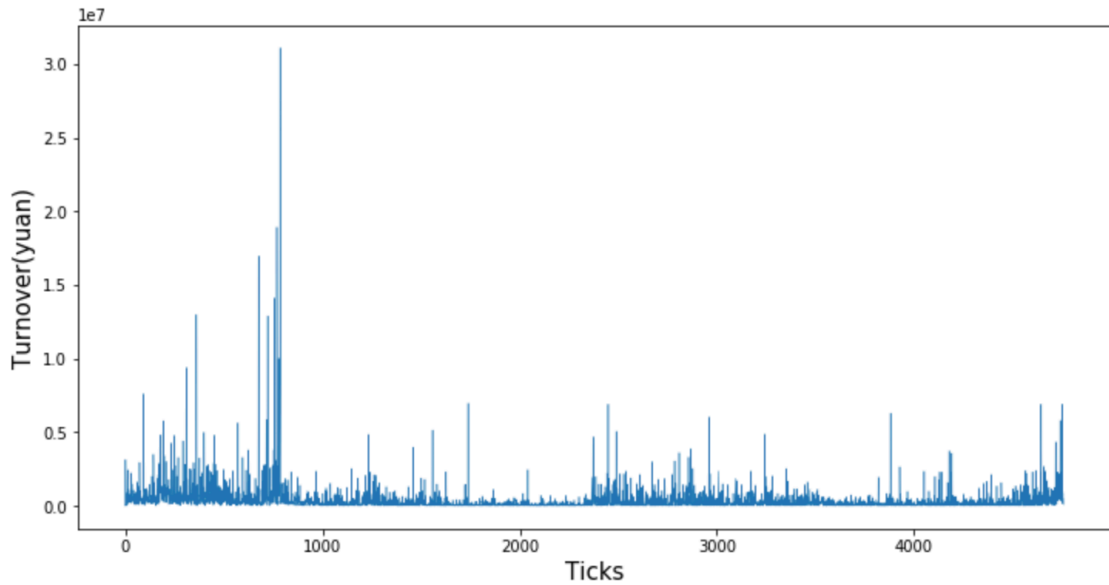


图 3.3 2017 年 6 月 7 日贵州茅台交易额变化图

如果将每日的交易时间按照 30 分钟的间隔分为 8 个时段，计算出整个训练数据在每个时段内交易额的平均值，结果如图 3.4，可以更明显地观察到“W”形状的分佈规律。

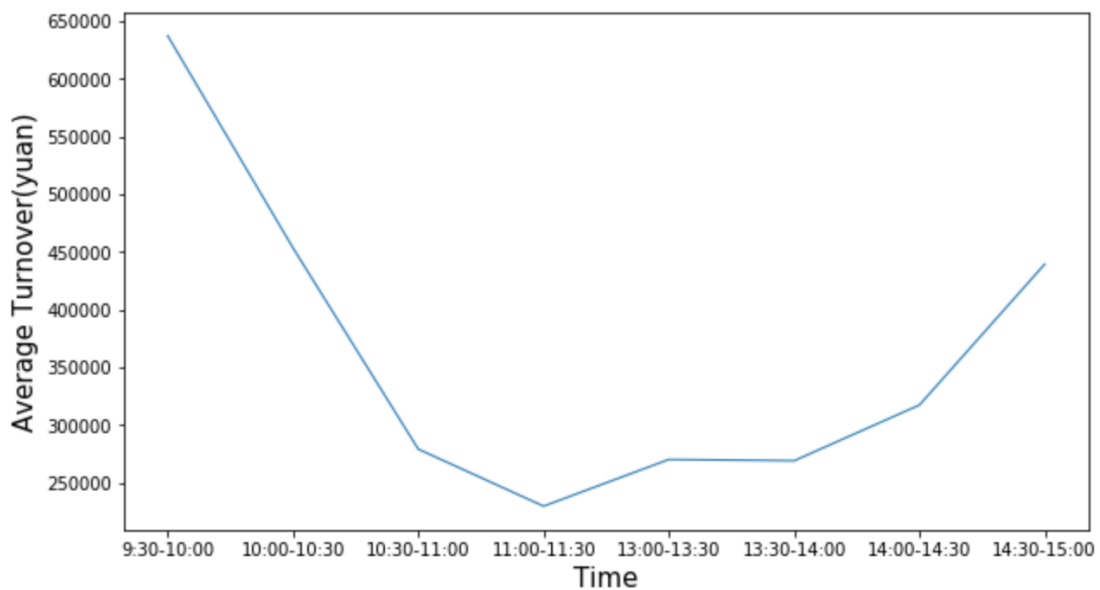


图 3.4 训练集上贵州茅台交易额分时段平均值变化图

使用章节 2.5.3 中介绍的日内平均法 (IAM)，将每个时段内训练数据上的整体平均值减去，得到去除日内效应后的数据。使用日内平均法去除日内效应前后交易额分布的对比如图 3.5 和图 3.6，虽然去除日内效应后依然是一个拖尾、不对称、均值不为 0 的分布，但是对比消除前已经有所改善。



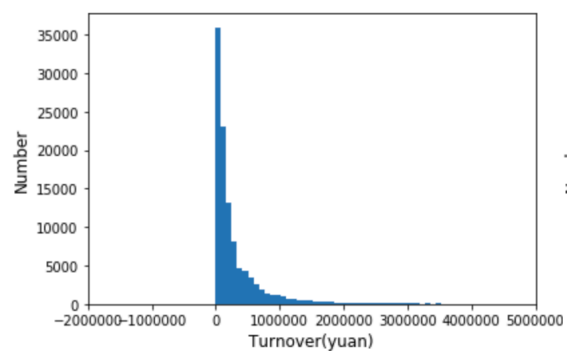


图 3.5 调整前交易量分布

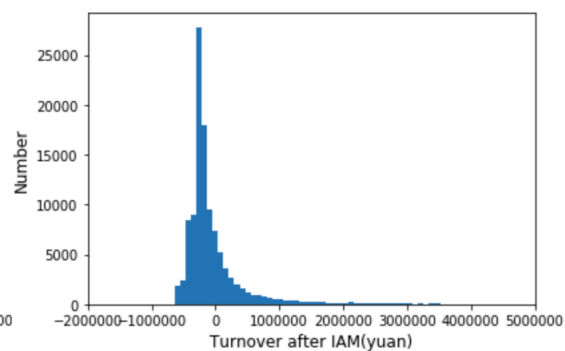


图 3.6 调整后交易量分布

交易量、交易笔数也具有相似的日内效应特征，本文同样采用日内平均法对他们进行了处理。

## 第四章 实验与结果分析

### 4.1 回测框架

本文采用固定持仓时间的交易策略（如图 4.1），每次固定买入价值当前资产净值 1/10 的股票，因此收益率主要受预测精度的影响。

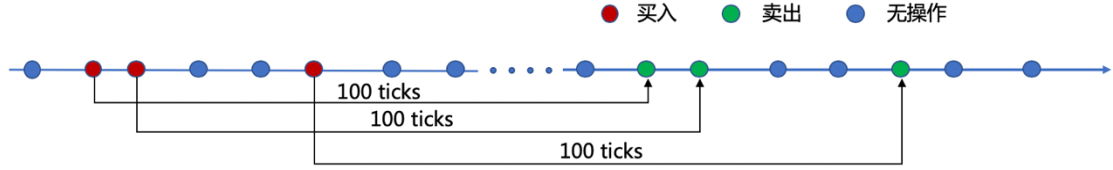


图 4.1 固定持仓时间（ticks）的交易策略示意图

首先计算出测试集上市场价格的收益率序列 $\{r_t^{test}\}$ ，初始时刻相对价格设为 1，通过计算复利得到相对价格序列 $\{v_t^{test}\}$ 作为比较的基准（benchmark）。在训练集上训练机器学习模型，计算模型在测试集上的预测信号。同样将交易策略的初始净值设为 1，之后对每个 tick 的预测信号进行记录，计算出每个 tick 对应的持仓情况。

对于每个 tick，资产净值由两部分构成：持有的现金和仓位的浮动净值。因此对于每个 tick，先计算仓位的浮动净值

$$v_{float}^t = v_{float}^{t-1} \times (1 + r_t) \quad (4.1)$$

然后考虑该 tick 内的调仓操作。本文只考虑交易的固定成本，大部分发生在卖出时，因此近似认为只在卖出操作时扣除交易成本。如果预测到买多信号，就在持有现金中减去买入成本，然后在浮动净值中加入新仓位的净值；如果预测到买空信号，则买入成本取负值；如果检测到保持仓位的信号，则买入成本取 0。

$$\begin{cases} v_{float}^t += new_{buy} \\ v_{currency}^t += -new_{buy} \end{cases} \quad (4.2)$$

除了买入操作，还要考虑该 tick 内持仓到期的卖出操作。如果卖出的是多头，就从浮动净值中减去卖出收益，在现金中加入扣除交易成本后的收益；如果卖出的是空头，就将卖出收益取为负值。

$$\begin{cases} v_{float}^t += -sell \\ v_{currency}^t += sell - |sell| \times cost_{rate} \end{cases} \quad (4.3)$$

最后，该 tick 的净值就等于两部分净值相加：

$$v^t = v_{float}^t + v_{currency}^t \quad (4.4)$$

通过迭代可以获得测试集上每一个 tick 对应的资产净值，因为初始净值设为 1，此时得到的就是相对净值序列。

策略在测试集上的收益率还与入场时机有关，仅凭单个测试集上的收益评价策略具有偶然性。因此本文还设计了其他指标来评价模型的预测能力：

- 1) 整体预测准确率 (Accuracy)

$$\frac{\text{预测信号与实际信号匹配个数}}{\text{实际信号个数}} \times 100\% \quad (4.5)$$

- 2) 多头预测准确率/空头预测准确率 (Precision)

$$\frac{\text{预测多(空)头信号正确个数}}{\text{预测多(空)头信号个数}} \times 100\% \quad (4.6)$$

- 3) 多头信号捕获率/空头信号捕获率，又称为召回率 (Recall)

$$\frac{\text{预测多(空)头信号正确个数}}{\text{实际上涨(下跌)信号个数}} \times 100\% \quad (4.7)$$

- 4) 夏普比率 (Sharpe Ratio)。因为测试时间较短，假设无风险利率为 0

$$\frac{\text{投资组合收益率}}{\text{投资组合标准差}} \times 100\% \quad (4.8)$$

- 5) 最大回撤 (Maximum Drawdown)。  $D_i$  代表  $i$  时刻价格

$$\text{MAX} \left( \frac{D_i - D_j}{D_i} \times 100\% \right)_{i < j} \quad (4.9)$$

本文将结合各个指标对结果进行综合评价和分析。

## 4.2 日内效应调整

日内效应调整采用章节 2.5.3 中介绍的日内平均法，比较调整前后逻辑回归模型和随机森林模型预测的效果。由于神经网络方法本身就具有挖掘隐藏特征的能力，因此不再单独进行日内效应调整，将在章节 4.3 中专门讨论。

当前 A 股市场对于买空限制很多，难以进行高频做空操作。如果只考虑做多，那么信号就只有买多和保持两种，因此模型采用二分类的输出，回测结果如图 4.2，评价指标如表 4.1。日内效应调整后，策略收益都有明显提升，逻辑回归模型策略的夏普比率从 2.56 提升至 2.85，随机森林模型策略的夏普比率从 -2.71 提升至 -1.51；多头信号预测的准确率大幅提升，逻辑回归模型从 50.39% 提升至 54.15%，随机森林模型从 27.77% 提升至 35.69%。对于逻辑回归模型策略，日内效应调整后多头信号的捕获率也从 0.50% 提升一倍至 1.05%，所有指标中仅有最大回撤略微升高，其他均得到改善。对于随机森林模型策略，

调整后多头信号的数量和捕获率略微降低，但预测准确率整体提高，最大回撤也得到降低。

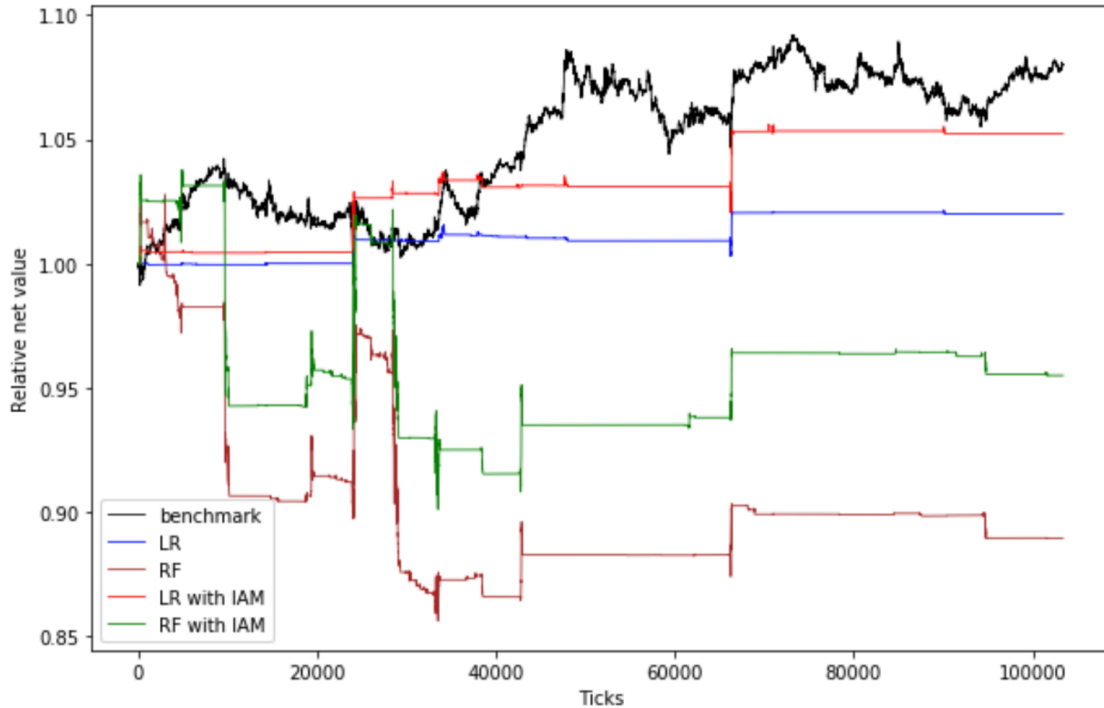


图 4.2 只开多头时各模型回测收益率曲线图

表 4.1 只开多头时各模型回测结果

指标	逻辑回归	逻辑回归+ 日内平均法	随机森林	随机森林+ 日内平均法
多头信号比	0.12%	<b>0.24%</b>	<b>2.61%</b>	2.01%
整体准确率	87.37%	<b>87.39%</b>	86.20%	<b>86.79%</b>
多头准确率	50.39%	<b>54.15%</b>	27.77%	<b>35.69%</b>
多头捕获率	0.50%	<b>1.05%</b>	<b>5.74%</b>	5.69%
夏普比率	2.56	<b>2.85</b>	-2.71	<b>-1.51</b>
最大回撤	<b>1.22%</b>	1.56%	18.30%	<b>13.16%</b>

如果单纯研究趋势预测，不考虑做空的操作难度，那么信号共有买多、买空和保持三种，模型调整为多分类输出，回测结果如图 4.3，评价指标如表 4.2。日内效应调整后，逻辑回归模型策略的夏普比率从 2.81 提升至 2.96，随机森林模型策略的夏普比率从 -0.79 提升至 0.49；多头信号预测的准确率提升，逻辑回归模型从 52.80% 提升至 68.97%，随机森林模型从 25.99% 提升至 31.32%；多头和空头的捕获率都上升。对于逻辑回归模型策略，调整后空头准确率从 19.74% 提升至 35.06%，多头和空头信号捕获率都大幅提升 4 倍左右，仅有最大回撤升高。对于随机森林模型策略，调整后多头方向的捕获率和准确率都有所改善，空头方向捕获率提高但准确率略微下降，最大回撤增大。

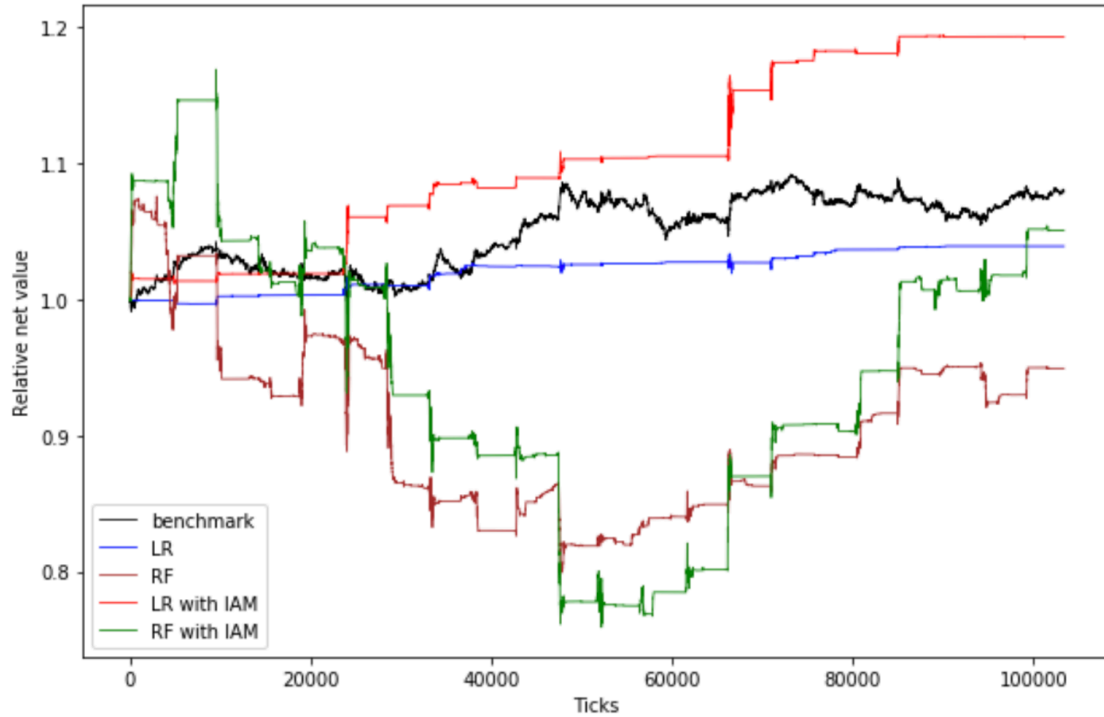


图 4.3 开多头和空头时各模型回测收益率曲线图

表 4.2 开多头和空头时各模型回测结果

指标	逻辑回归	逻辑回归+ 日内平均法	随机森林	随机森林+ 日内平均法
多头信号比	0.12%	<b>0.45%</b>	<b>3.46%</b>	3.42%
空头信号比	0.30%	<b>0.60%</b>	2.28%	<b>3.91%</b>
整体准确率	76.83%	<b>77.15%</b>	<b>75.06%</b>	74.91%
多头准确率	52.80%	<b>68.97%</b>	25.99%	<b>31.32%</b>
多头捕获率	0.51%	<b>2.45%</b>	7.11%	<b>8.48%</b>
空头准确率	19.74%	<b>35.06%</b>	<b>21.74%</b>	19.57%
空头捕获率	0.56%	<b>2.01%</b>	4.73%	<b>7.31%</b>
夏普比率	2.81	<b>2.96</b>	-0.79	<b>0.49</b>
最大回撤	<b>1.41%</b>	3.31%	<b>25.83%</b>	35.08%

无论是否考虑买空操作，利用日内平均法进行日内效应调整后，逻辑回归和随机森林模型策略的收益表现都得到改善，尤其是在多头方向；调整对于逻辑回归的改善效果更加显著。

### 4.3 神经网络方法

本文采用多神经元的 LSTM 层作为隐藏层，最后连接以 Sigmoid 函数为激活函数的全连接层作为输出层，输出结果是分布在 $[0,1]$ 之间的数，可以看作多头

信号出现的概率，通过设置不同的判断阈值，可以调节模型预测多头信号的数量和准确度。

本文使用了两种 LSTM 神经网络。第一种是单层 LSTM 神经网络，只有一个含 16 个单元的 LSTM 隐藏层；第二种是堆叠 LSTM 神经网络，包含三层 LSTM 隐藏层，单元数依次为 48 个、16 个和 2 个，三层结构旨在模拟特征重构、关系挖掘、二分类这个过程（如图 4.4），前两层设置 20% 随机失活（Dropout）防止过拟合。

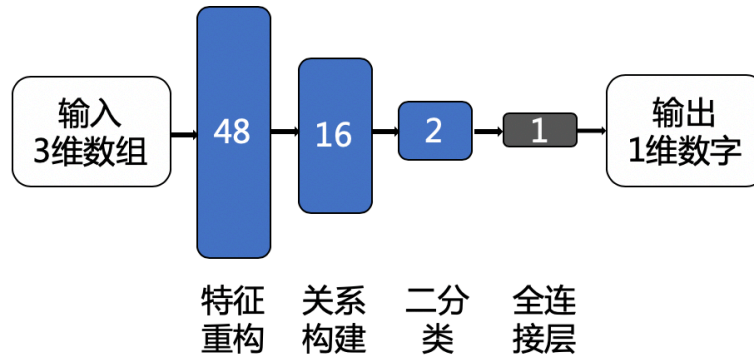


图 4.4 堆叠 LSTM 结构示意图

模型使用交叉熵损失函数，优化采用 Adam 优化器。实验发现训练批次为 2 时验证集上总损失最小，泛化能力最好（如图 4.5 和图 4.6）。实验中发现如果选取更多训练批次，非常容易造成过拟合现象；但如果继续增加模型复杂度，又会导致在验证集上泛化能力很差。本文认为这主要是由于训练集只有 87548 个样本，较少的样本量导致涵盖的情况比较少，因此可供挖掘的有预测作用的信息也不多，因此采用较简单的模型训练较少的批次已经可以最大化地挖掘出这些信息。

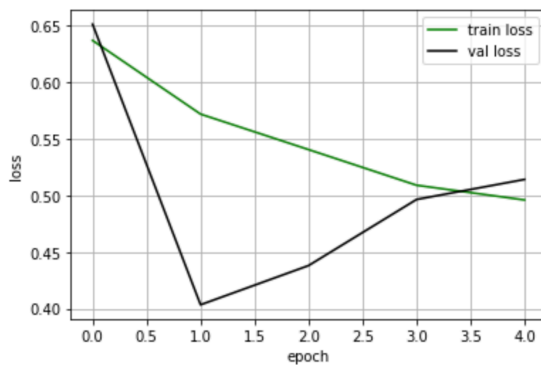


图 4.5 单层 LSTM 训练误差图

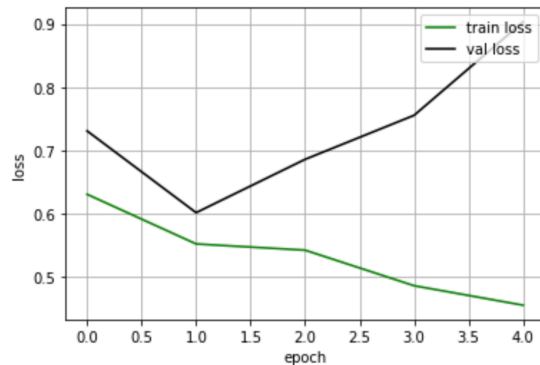


图 4.6 堆叠 LSTM 训练误差图

设定不同概率阈值，两种 LSTM 神经网络模型策略在验证集上的回测结果如表 4.3 所示。随着阈值升高，LSTM 神经网络的多头预测准确率单调升高，捕获率单调降低，说明模型输出与多头信号出现概率正相关。单层 LSTM 在预测准确率上表现更突出，阈值设定为 0.75 时可以在 2.02% 的捕获率下达到 55.11% 的预

测精度；堆叠 LSTM 在捕获率上表现更优，同样在 45%左右的准确率时拥有 9.39%的捕获率，超过单层 LSTM 的 4.75%，但整体准确率不如单层 LSTM 网络。

从实验结果可以看出，根据 LSTM 模型的输出数值，设定不同的阈值可以调节模型在测试集上准确率与捕获率的大小；而在统一阈值下，单层与多层神经网络由于结构不同，也会相应改变准确率与捕获率的大小。这说明通过调节神经网络模型的超参数，可以调整预测结果：随着阈值增大，结果准确率提升，但捕获率下降；随着层数增多，模型更加复杂，训练成本更高，但在结果的准确率和捕获率上可能做到同时提升。在本实验中，由于训练样本不多，无法有效训练层数和神经元更多的复杂网络，因此增加层数后虽然捕获率大幅上升，但准确率反而有少许下降，但捕获率上升的幅度远大于准确率下降的幅度。

表 4.3 不同阈值下 LSTM 预测结果

阈值	多头准确率		多头捕获率	
	单层 LSTM	堆叠 LSTM	单层 LSTM	堆叠 LSTM
0.50	24.95%	<b>31.46%</b>	18.74%	<b>24.10%</b>
0.55	27.78%	<b>32.21%</b>	14.16%	<b>22.73%</b>
0.60	<b>33.54%</b>	32.52%	9.21%	<b>21.26%</b>
0.65	<b>37.59%</b>	32.90%	6.87%	<b>19.71%</b>
0.70	<b>45.49%</b>	33.68%	4.75%	<b>17.21%</b>
0.75	<b>55.11%</b>	38.63%	2.02%	<b>15.28%</b>
0.80	<b>70.43%</b>	40.53%	0.62%	<b>13.79%</b>
0.85	<b>81.82%</b>	42.18%	0.14%	<b>12.29%</b>
0.90	<b>100.00%</b>	46.02%	0.03%	<b>9.39%</b>
0.95	—	54.69%	—	0.54%

基于 LSTM 模型在验证集上的实验结果，综合考虑准确率与捕获率，本文选择阈值为 0.75 的单层 LSTM 模型和阈值为 0.90 的多层 LSTM 模型在测试集上进行回测并与其他流派的模型进行比较。

#### 4.4 模型对比总结

在实际交易中我们可以设计更加复杂的交易策略来提高收益，例如设置止损、动态调整持仓时间和买卖仓位等，但能否盈利的决定性因素是模型的预测能力。因此本文主要关注各个模型策略在高频数据集上的预测能力，包括预测精度和预测数量两方面，可以用多头信号预测准确率和多头信号捕获率两个指标分别量化。

本文各模型在测试集上的预测能力对比如表 4.4。总体上准确率和捕获率存在权衡的问题，由于高频交易的目标之一是要获取绝对收益，因此需要保证一定的准确率，在准确率足够盈利的前提下再尽可能提高捕获率。

表 4.4 各模型预测能力对比

模型	多头准确率	多头捕获率
逻辑回归	<b>52.80%</b>	0.51%
逻辑回归+日内平均法	<b>68.97%</b>	2.45%
随机森林	25.99%	<b>7.11%</b>
随机森林+日内平均法	31.32%	<b>8.48%</b>
单层 LSTM（阈值 0.75）	<b>54.71%</b>	2.24%
堆叠 LSTM（阈值 0.90）	45.97%	<b>9.41%</b>

综合两个指标进行考虑，在本文数据集上逻辑回归+日内平均法模型和堆叠 LSTM 是综合表现最好的模型。从模型类型上来说，“统计学习”流派的回归模型在准确率上表现最好，“连接主义学习”流派的 LSTM 模型在准确率和捕获率上表现比较均衡，“符号主义学习”流派的随机森林模型虽然捕获率高但在准确率上表现太差。三种不同流派模型的实验结果表明，高频时间序列的噪声很大，基于逻辑决策过程的“符号主义学习”模型难以进行有效地判断；而基于简单线性统计假设的回归模型能够剥除噪音发掘趋势，但是简单的线性模型还不足以挖掘出足够多的信息，因此预测结果非常保守，错失许多机会；基于层状非线性结构的 LSTM 神经网络模型相较于线性回归模型更有效地挖掘了特征之间的非线性关系，从而能够更好地兼顾准确率和捕获率，但由于结构复杂调参困难，而且非常容易过拟合，因此需要更多的训练样本和更高的训练成本。



## 第五章 总结与展望

本文分析了沪市股票贵州茅台（600519.SH）的高频数据，指出其尖峰厚尾分布、日内效应等特征，并通过日内平均法对部分变量的日内效应进行消除处理。在实验部分，本文验证了使用日内平均法对于提升模型预测能力的显著效果，并通过构造简单的高频交易策略比较了不同流派的机器学习模型在该数据集上的预测能力。

本文发现“统计学习”流派的逻辑回归模型预测精度不俗，但发现复杂关系的能力弱，预测结果比较保守；“符号主义学习”流派的随机森林模型在金融高频时间序列上预测效果较弱；“连接主义学习”流派的LSTM神经网络模型最灵活，可以通过改变超参数来调整预测准确率和捕获率，但要求的训练样本更多、消耗的训练成本更大。

本文在实验部分的发现虽然可以通过日内平均法调整等方法提高预测准确率，但每个模型的信号捕获率普遍都偏低，这在实际交易中会导致错失许多交易机会，如何提高捕获率是一个重要的课题。本文同时指出长短期神经网络模型具有可塑性强的特点，但由于本文使用的训练样本较少，因此效果有限，但如果使用更大规模的训练样本来训练更复杂的网络结构，有望把准确率和捕获率同时提高。此外本文在处理日内效应时采用了简单的日内平均法，发现处理后的数据分布依然不够理想，因此如果采用一些更复杂的方法，比如核函数方法，可能会取得更好的效果，有待进一步研究。

## 第六章 参考文献

- [1] Black, Fischer, Scholes, Myron S. The Pricing of Options and Corporate Liabilities[J]. Journal of Political Economy, 1973, 81(3): 637-654.
- [2] Li D X. On Default Correlation: A Copula Function Approach[J]. SSRN Electronic Journal, 1999, 9(4).
- [3] Ross S A. The Arbitrage Theory of Capital Asset Pricing[J]. Journal of Economic Theory, 1976, 13(3): 341-360.
- [4] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 1-12.
- [5] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18: 1527-1554.
- [6] Lecun Y, Bengio Y, Hinton G E. Deep learning[J]. Nature, 2015, 521(7553): 436.
- [7] Krizhevsky A, Sutskever I, Hinton G. ImageNet Classification with Deep Convolutional Neural Networks[J]. Advances in neural information processing systems, 2012, 25(2).
- [8] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society, 2016.
- [9] Chen Q, Zhu X, Ling Z, et al. Enhanced LSTM for Natural Language Inference[C]. Meeting of the Association for Computational Linguistics, 2017: 1657-1668.
- [10] Devlin J, Chang M, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]. North American Chapter of the Association for Computational Linguistics, 2019: 4171-4186.
- [11] White H. Economic Prediction Using Neural Networks: the Case of IBM Daily Stock Returns[C]. IEEE International Conference on Neural Networks, 1988, 2(6):451-458.
- [12] Dunis C L, Nathani A. Quantitative Trading of Gold and Silver Using Nonlinear Models[J]. Neural Network World, 2007, 17(2): 93-111.

- [13] Persio L D, Honchar O. Analysis of Recurrent Neural Networks for Short-term Energy Load Forecasting[C]. International Conference of Computational Methods in Sciences & Engineering. American Institute of Physics Conference Series, 2017.
- [14] Choundhry R, Kumkum G. A Hybrid Machine Learning System for Stock Market Forecasting[J]. International Journal of Computer and Information Engineering, 2008, 2(3): 689-692.
- [15] Maknickiene, Maknickas. Application of Neural Network for Forecasting of Exchange Rates and Forex Trading[C]. The 7th International Scientific Conference “Business and Management 2012”, 2012: 122-127.
- [16] 袁祥枫. 基于 LSTM 的商品期货高频数据趋势预测模型的研究[D]. 北京邮电大学, 2019.
- [17] 张贵勇. 改进的卷积神经网络在金融预测中的应用研究[D]. 郑州: 郑州大学, 2016.
- [18] 魏瑾瑞, 朱建平, 谢邦昌. 金融高频数据仅仅是一个优质的时间序列吗:概念及统计特征的再考察[J]. 投资研究, 2014, 33(05): 11-21.
- [19] 王维国, 余宏俊. 超高频数据的日内效应调整方法[J]. 中国管理科学, 2015, 23(6): 49-56.
- [20] 孙达昌, 毕秀春. 基于深度学习算法的高频交易策略及其盈利能力研究[J]. 中国科学技术大学学报, 2018, 48(11): 923-932.
- [21] Sharpe W F. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk[J]. Journal of Finance, 1964, 19(3): 425-442.
- [22] Lintner, John. Security Prices, Risk, and Maximal Gains from Diversification[J]. The Journal of Finance, 1965, 20(4): 587-615.
- [23] Mossin J. Equilibrium in a Capital Asset Market [J]. Econometrica, 1966, 34(4).
- [24] Breiman L I. Random Forests[J]. Machine Learning, 2001, 45(1):5-32.
- [25] Breiman L I, Friedman J H, Olshen R A, et al. Classification and Regression Trees (CART)[J]. Biometrics, 1984, 40(3): 358.

- [26] Gers F A. Learning to forget: continual prediction with LSTM[C]. 9th International Conference on Artificial Neural Networks: ICANN' 99. IET, 1999.
- [27] Dyer C, Ballesteros M, Ling W, et al. Transition-Based Dependency Parsing with Stack Long Short-Term Memory[C]. International Joint Conference on Natural Language Processing, 2015: 334-343.
- [28] Kalchbrenner N, Danihelka I, Graves A. Grid long short-term memory[C]. 4th International Conference on Learning Representations: ICLR, 2016.
- [29] Wood R A, Mcinish T H, Ord J K. An Investigation of Transactions Data for NYSE Stocks[J]. Journal of Finance, 1985, 40(3): 723-739.

## 第七章 致谢

论文写到这里我已经很疲惫了，但是感激之情无法掩藏，仍有感而发写下诸多文字。

首先要感谢我的父母，他们宽容、勤勉，从小给我树立了良好的榜样，并一以贯之对我“放养”，尊重我的想法和几乎所有选择，在任何时候都给予了我莫大的支持。我从未当面对他们表达感情，所以在这里我想首先提到他们，即使他们可能无法看到；

第二位我想感恩我的导师李祥林教授。李教授从始至终都在为学生的长远发展殚精竭虑，不仅仅是在论文的指导上，更是在课程的安排、经验的分享和日常的答疑解惑等等方面，在导师的帮助下我得以更快地了解未来将要进入的这个领域，这对我的整个未来发展都意义非凡；

接下来想感谢的人很多，不分先后。我的好哥们儿们，就不一一点名了，不需要再多说什么了；所有教导过我的老师们，非常感恩你们无私的帮助，尤其是对我照顾颇多的张楠老师和朱雪宁老师，私下里觉得你们就像哥哥姐姐一样；我的室友们，一直以来跟我发疯跟我熬夜跟我肝作业，幸运的是我的每一位室友都非常棒；我的可爱，我还欠你好多零食呢，感谢你让我终于能把《复旦姑娘》唱出去了；书亭的各位朋友，感恩你们陪伴了我整个本科生涯，书我以无数美妙回忆；小白老师，我要特别提到你，在无数个最艰难的时刻是你给予我慰藉；实习时的老板们，各个社团、活动、夏令营结识的小伙伴们，同样感谢你们在我人生路上留下的色彩，是你们每个人的善意和帮助铸就了现在的我。

感恩复旦的四年，愿所有事顺利，祝所有人安好！

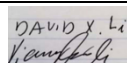
**指导教师对论文独立性的审查意见：**

指导教师应依据《学位论文作假行为处理办法》第五条的规定对毕业论文撰写人进行学术道德与学术规范教育，并在此基础上对其撰写论文进行指导。

✓ 本人经过尽职审查，未发现毕业论文撰写人有学位论文作假行为。本人认为毕业论文撰写人独立完成了本毕业论文。

□ 本人经过尽职审查，发现毕业论文撰写人有如下学位论文作假行为：

指导教师签名：

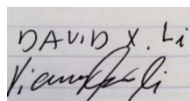


日期： 2020 年 6 月 9 日

**指导教师评语：**

This essay has applied machine learning techniques, logistic regression and LSTM to study high frequency trading strategies(tick data) for predicting stock price movement direction and has come up with interesting results. The essay is well written, and involves of dealing with large data set, and application and interpretation of machine learning results. It is excellent!

签名：



2020 年 6 月 9 日

**答辩委员会（小组）评语：**

签名：

20 年 月 日

学分：

成绩：

备注：