# STOCKPRICEPREDICTIONUSINGLSTM,RNNANDCNN-SLIDINGWINDOWMODEL

# 摘要：

股票市场对今天的经济产生深远的影响。股价的涨跌对决定投资者的收益有着重要的作用。现有的预测方法既使用线性（AR,MA,ARIMA）又使用非线性算法（ARCH,GARCH,NeuralNetworks）,但他们专注于使用每日收盘预测单个公司的股指指数变动或价格预测。所提出的方法是独立于模型的方法。在这里,我们不是将数据拟合到特定模型,而是使用深度学习架构来识别数据中存在的潜在动态。在这项工作中,我们使用三种不同的深度学习架构来对NSE上市公司的价格进行预测,并比较它们的表现。我们正在应用滑动窗口方法来短期预测未来值。模型的性能使用百分比误差进行量化

关键词：时间序列,股票市场,RNN,LSTM,CNN

# 简介

预测可以被定义为通过分析历史数据来预测未来的一些事件或事件。它涵盖了许多领域,包括商业和工业,经济学,环境科学和金融。预测问题可以分类为

•短期预测（预测几秒钟,几分钟,几天,几周或几个月）

•中期预测（预测1至2年）

•长期预测（预测超过2年）

**许多预测问题涉及时间分析。时间序列数据可以被定义为选定变量的时间顺序观测值。在我们的情况下,变量是股票价格。它可以是单变量也可以是多变量的。单变量数据包括仅关于一种特定股票的信息,而多变量数据包括不同时间的多个公司的股价。时间序列数据分析有助于识别数据中存在的模式,趋势和周期或周期。就股市而言,早期了解看涨或看跌模式有助于明智地投资资金。同时,模式分析有助于确定在特定时期内表现最佳的公司。这使得时间序列分析和预测成为重要的研究领域。**

现有的股票价格预测方法可分为以下几类[1]

•基本分析

•技术分析

•时间序列预测

基本分析是一种投资分析,通过分析其销售额,收益,利润和其他经济因素来估计公司的股票价值。该方法最适合长期预测。技术分析使用股票的历史价格来识别未来价格。滑动平均法是技术分析的常用算法。它可以被认为是过去n个数据点的未加权平均数。这种方法适用于短期预测。第三种方法是分析时间序列数据。它基本上涉及两类算法,它们是

•线性模型

•非线性模型

不同的线性模型是AR,ARMA,ARIMA及其变体[2][3][4]。这些模型使用一些预定义的公式来将数学模型拟合为单变量时间序列。这些模型的主要缺点是,它们没有考虑到数据中存在的潜在动态。由于它们仅考虑单变量时间序列,因此这些模型不能识别各种股票之间的相互依赖关系。另外,针对一个系列确定的模型将不适合另一个系列。由于这些原因,无法确定整体数据中存在的模式或动态。

非线性模型涉及ARCH,GARCH,[3]TAR,深度学习算法[5]等方法。[6]对NIFTY50中列出的29家选定公司的股票价格和股票量之间的相互依赖关系进行了分析。该项研究的重点是深度学习算法在股票价格预测中的应用[7][8]。深度神经网络可以被认为是能够映射非线性函数的非线性函数逼近器。根据应用的类型,使用各种类型的深度神经网络架构。它们包括多层感知器（MLP）,递归神经网络（RNN）,长期短期记忆（LSTM）,CNN（卷积神经网络）等[9]。它们已经应用于图像处理,自然语言处理,时间序列分析等深度学习算法能够通过自学习过程识别数据中的隐藏模式和潜在动态。

在股市的情况下,所产生的数据是巨大的并且非常非线性。为了模拟这种动态数据,我们需要能够分析隐藏模式和潜在动态的模型。深度学习算法能够通过自学习过程来识别和利用数据中存在的交互和模式。与其他算法不同,深度学习模型可以有效地对这些类型的数据建模,并且可以通过分析数据中的交互和隐藏模式来提供良好的预测。在[5]中,我们可以看到各种深度学习模型在多元时间序列分析中的应用。[10]中介绍了使用神经网络模型对金融时间序列进行建模的第一次尝试。**这项工作试图建立一个神经网络模型来解释IBM的资产价格变动的非线性规律。然而,这项工作的范围有限,但它有助于建立针对EMH的证据[11]。**

在使用NN模型的金融时间序列分析领域的研究使用不同的输入变量来预测股票收益。在一些工作中,来自单个时间序列的数据被用作输入[10],[8]。某些工作被认为包含了异质的市场信息和宏观经济变量。在[12]中,金融时间序列分析和NLP的组合已经被引入。在[13]和[7]中,深度学习架构已被用于多元金融时间序列的建模。在[14]中,使用技术分析变量的NN模型已经被应用于上海股票市场的预测。该工作比较了两种学习算法和两种权重初始化方法的性能。结果表明反向传播效率可以通过多重线性回归权重初始化的共轭梯度学习来提高

1996年,[15]使用反向传播和RNN模型来预测五种不同股票市场的股指。在[16]中,引入时间延迟,递归和概率神经网络模型用于每日股票预测。在[17]中,PSO和LS-SVM等机器学习算法的应用已被用于标准普尔500指数股票市场的预测。[18]中介绍了遗传算法和神经网络模型的实现。这项工作结合了遗传算法和人工神经网络在预测中的应用。在这项工作中,神经网络的权重是从遗传算法中获得的,然而,这个模型的预测精度很低。小波变换在预测中的应用被引入[19]。这项工作使用小波变换来描述股票趋势的短期特征。随着LSTM[20]的引入,对时间相关数据的分析变得更加高效。这些类型的网络具有保存过去信息的能力。他们已被用于股票价格预测[8],[7]。

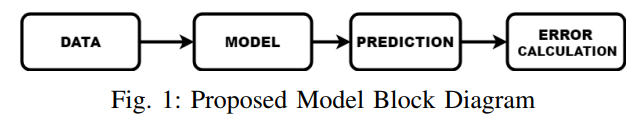
所提出的方法侧重于预测NSE（国家证券交易所）上市公司的股票价格。我们采用的方法是数据重叠的滑动窗口方法。数据集包含NSE上市公司的分析数据。在这里,我们试图获得一个用于预测目的的广义模型,它可以使用分钟级别数据作为输入。这种建模在高频交易发生的算法交易中有应用。

本文结构如下[II]部分解释了所提出的方法论。结果和讨论可以在第[III]节中找到,第[IV]节包含结论。

# 方法

该数据集包括2014年7月至2015年6月期间1721家NSE上市公司的分钟股票价格。其中包括日期戳记,时间戳,交易编号,股票价格和每分钟出售股票数量等信息。对于这项工作,我们选择了两个不同的部门,IT部门和制药部门。来自IT部门的两家公司和制药部门的一家公司被带入了这项研究。这些公司通过NIFTYIT指数和NIFTY-Pharma指数得到了确认。这三家公司的数据是从现有数据中提取的,并进行预处理以获取股票价格。

该工作基于滑动窗口方法进行短期未来预测。窗口大小被固定为100分钟,重叠了90分钟的信息,并且将来预测10分钟。通过计算各种窗口大小的误差来确定最佳窗口长度。列车数据包括2014年7月1日至2014年10月14日期间Infosys的股票价格,测试数据包括Infosys,TCS和CIPLA在2014年10月至2014年11月28日期间的股票价格。



**Infosys和TCS的数据在2000到4000的范围内变化,Cipla的数据在400到700之间变化。为了统一数据范围,将其进行归一化并映射到0到1的范围。将该归一化后的数据提供给神经网络用于训练。通过改变图层大小进行微调,所有模型都训练了1000个时期。如果当前时期的损失（均方误差）小于前一时期获得的值,则存储该时期的权重矩阵。在训练过程之后,对这些模型中的每一个进行测试,并将具有最小RMSE（均方根误差）的模型作为预测的最终模型。**

**我们已经使用了三种不同的深度学习架构,RNN,LSTM和CNN来完成这项工作。RNN是一类神经网络,其中计算单元之间的连接形成定向圆。与前馈网络不同,RNN可以使用其内部存储器来处理任意输入序列。RNN中的每个计算单元都具有随时间变化的实值激活和可修改的权重。RNN是通过在图形结构上递归地应用相同的一组权重而创建的。许多RNN使用（1）来定义其隐藏单元的值。**



在RNN的情况下,学习模型始终具有相同的输入大小,因为它是根据从一个状态到另一个状态的转换来指定的。此外,体系结构在每个时间步使用具有相同参数的相同转换函数。LSTM是Hochreiter和Schmidhuber于1997年推出的一种特殊的RNN[20]。在LSTM架构的情况下,通常的隐藏层被LSTM单元取代。单元由各种可以控制输入流量的门组成。LSTM单元由输入门,单元状态,忘记门和输出门组成。它也由S形层,tanh层和点乘法操作组成。各种门和它们的功能如下

•输入门：输入门由输入组成。

•细胞状态：通过整个网络运行,并且有能力通过大门帮助添加或删除信息。

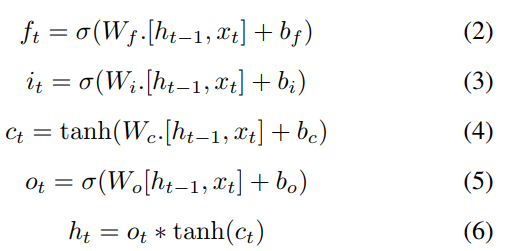
•忘记门层：决定允许的信息部分。

•输出门：它由LSTM生成的输出组成。

•Sigmoid图层生成0到1之间的数字,描述每个组件应该通过多少。

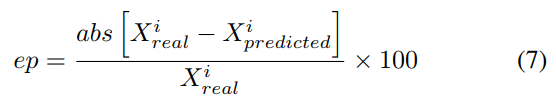
•Tanh层生成一个新的向量,它将被添加到状态中

单元状态根据门的输出进行更新。在数学上,我们可以用下面的等式来表示它。



xt：输入矢量,ht：输出矢量,ct：单元状态矢量,ft：忘记门矢量,it：输入门矢量,ot：输出门矢量,W,b是参数矩阵和矢量。卷积神经网络或CNN是用于处理具有已知网格状拓扑的数据的专用神经网络。这包括时间序列数据,可以将其视为1D和图像数据,可以将其视为二维像素网格。网络采用称为卷积的数学运算,因此称为卷积神经网络。它是一种专门的线性操作。卷积网络至少在其一个层中使用卷积而不是一般的矩阵乘法。使用这三种模型的动机是确定给定数据中是否存在长期依赖关系。这可以从模型的性能中确定。RNN和LSTM体系结构能够识别长期相关性并将其用于未来预测。然而,CNN体系结构主要关注给定的输入序列,并且在学习过程中不使用任何以前的历史或信息。使用其他公司数据测试模型的动机是检查公司之间的相互依赖性并了解市场动态。

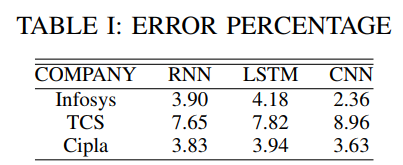
训练数据归一化。测试数据也经历了相同的归一化。在获得预测输出后,应用非归一化并使用可用的真实标签计算百分比误差。使用（7）计算误差百分比,



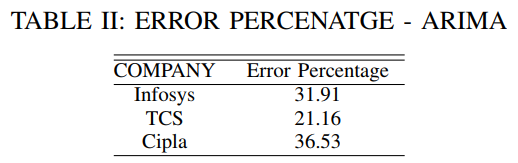
其中ep是误差百分比,Xireal是第i个实际值,Xipredict是第i个预测值。错误百分比给出输出中存在的错误的大小

# 结果和讨论

该实验针对三种不同的深度学习模型完成。表1给出了每个模型获得的误差百分比的最大值。从表格中可以清楚地看出,CNN比其他两种模式的结果更准确。这是由于CNN不依赖任何以前的信息进行预测的原因。它仅使用当前窗口进行预测。这使模型能够理解当前窗口中发生的动态变化和模式。但是,在RNN和LSTM的情况下,它使用先前滞后的信息来预测未来的情况。由于股票市场是一个高度动态的系统,系统中存在的模式和动态并不总是相同的。这导致LSTM和RNN架构的学习问题,因此模型无法准确捕获动态变化。

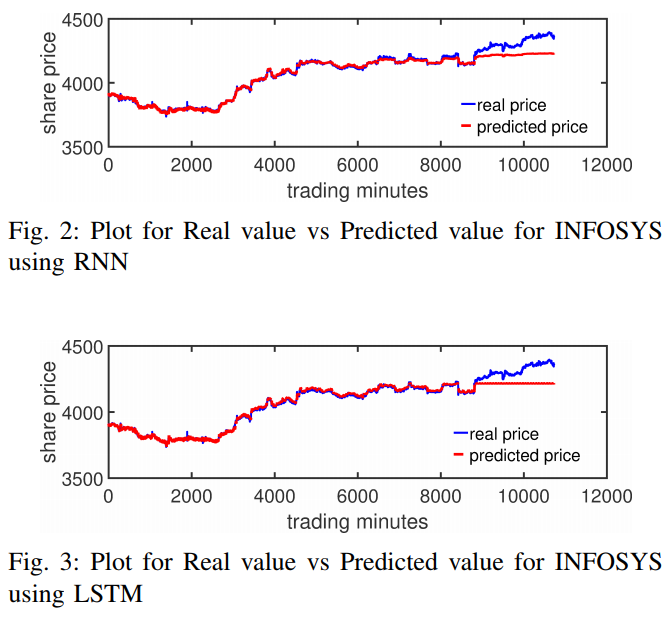


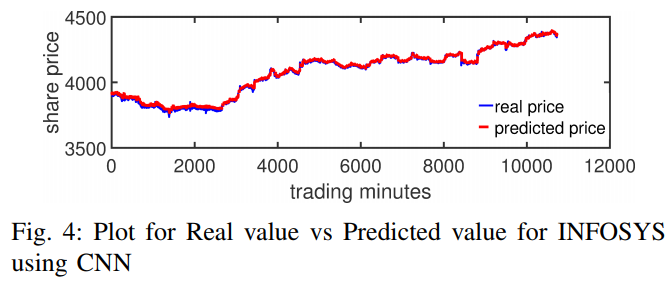
为了比较,我们使用了ARIMA,这是一种用于预测的线性模型。三家公司获得的误差百分比如下



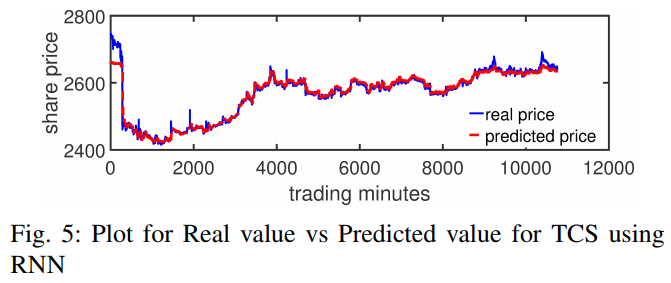
从表[I]和表[II]可以看出,深度学习模式的表现优于ARIMA。

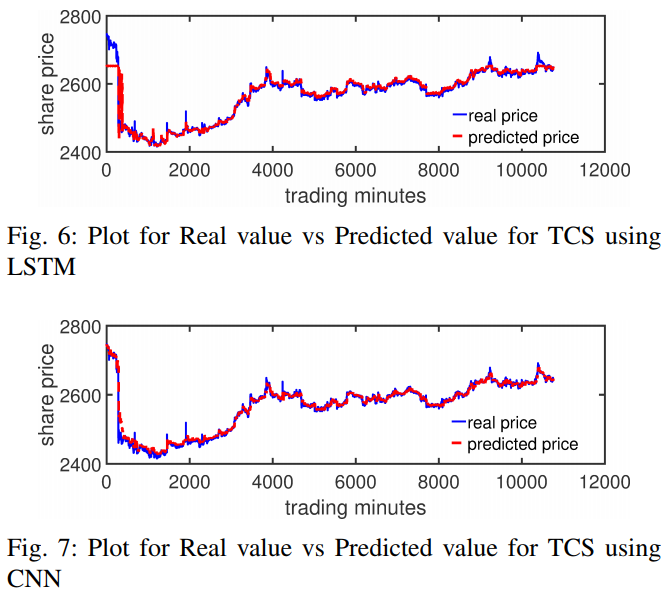
从图（2）和图（3）可以看出,RNN和LSTM都未能捕捉趋势（最终在9000到11000之间）与之前的窗口相比,该时间窗口中股票模式的行为发生了变化。在CNN的情况下,从图（4）可以看出,网络能够捕获9000到11000年的趋势变化

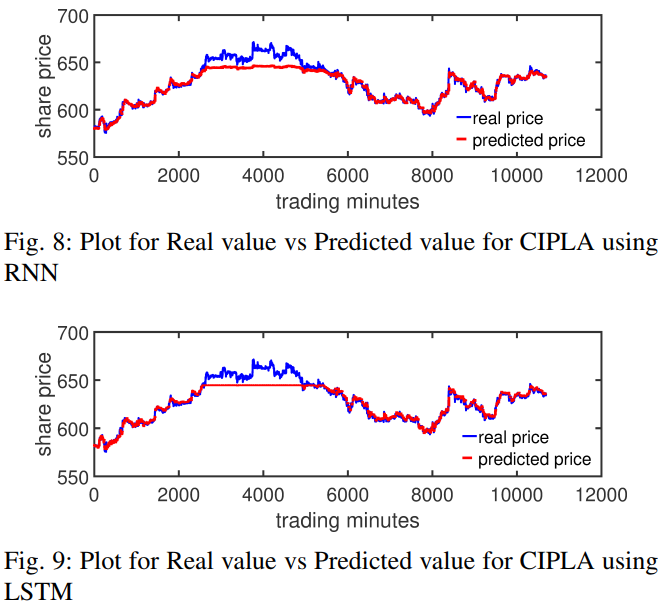


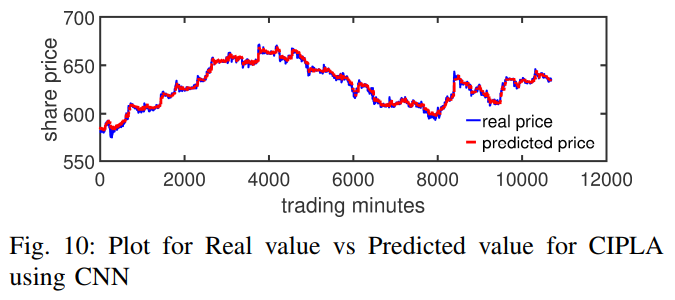


对于TCS,图（5）和图（6）,RNN和LSTM网络不能识别窗口开始的模式（在前1000分钟内）。TCS在此期间的趋势发生了变化。这使得预测不太准确,而在图（7）中,我们可以看到CNN比其他两个网络更准确地捕获这些变化。









在Cipla，Fig（8）和Fig（9）的情况下，在2000和6000之间，显然RNN和LSTM的预测值与原始数据的模式不匹配。 这可以被认为是系统行为的改变。 在Cipla的情况下（图10），我们可以观察到CNN能够捕捉特定时期股票价格行为的变化。

可以看出，与其他两种模式相比，CNN网络几乎能够捕捉到趋势并给出准确的预测结果。 CNN能够分析Infosys，TCS和Cipla的趋势变化。 还应该注意的是，我们使用Infosys的数据在2014年7月1日至2014年10月14日期间对网络进行了培训，即使与其他公司相比，Infosys的测试准确性也较低。这表明无论趋势 Infosys在7月至10月14日期间的展品没有出现在测试数据中（从2014年10月16日到2014年11月28日） 动态变化。这解释了Infosys与其他模型相比误差百分比的差异。 该模型还可以预测Infosys以外的公司的股价。这表明，该模型确定的模式或动态对其他公司也是共同的。

# 总结

我们提出了一个基于深度学习的形式化股票价格预测。可以看出，深度神经网络架构能够捕捉隐藏的动态并能够进行预测。我们使用Infosys的数据对该模型进行了培训，并能够预测Infosys，TCS和Cipla的股价。这表明，所提出的系统能够识别数据中的一些相互关系。而且，从结果中可以明显看出，CNN架构能够识别趋势的变化。对于所提出的方法，CNN被确定为最佳模型。它使用在特定时刻提供的信息进行预测。尽管其他两种模型在许多其他时间依赖数据分析中使用，但在这种情况下，它并没有执行CNN架构。这是由于股市发生的突然变化。股市中发生的变化可能并不总是规律的，也可能并不总是遵循相同的周期。基于公司和部门，趋势的存在和存在的时期将有所不同。对这些趋势和周期的分析将为投资者带来更多利润。为了分析这些信息，我们必须使用像CNN这样的网络，因为它们依赖于当前的信息。