****

**卓越学院**

**《暑期实践1》课程报告1**

**基于GAN的股票未来趋势预测**

**学期： 2020-2021学年第 4 学期**

**班级： 20184111**

**姓名： 张孜远**

**学号： 20151521**

**基于GAN的股票未来趋势预测项目实验报告**

**1、项目需求分析**

**1.1 项目背景**

由于金融市场的信息具有滞后性的特点（理论上 在证券公司工作的从业人员，其直系亲属是不能进行金融投资的，为了防止业内一手信息传播），股民们会选择依靠一些投资公司开发的股票软件（大多年费是365人名币）与当代社会同步的股票业内新闻进行整合分析，从而进行金融投资的买卖。本项目正是基于此背景、出于此需求，结合时下热门计算机深度学习技术，对股票的未来趋势进行预测。

**1.2 项目概述**

本项目是基于GAN的股票未来趋势预测，采用时下较新的量化投资技术GAN，对股票的历史时间数据进行回测（机器学习），从而推出接下来该个股的未来涨跌幅趋势情况。

**1.3项目受众群体**

本项目的受众群体是所有参与金融投资的全体公民，由此可见其受众群体之广，数量之庞大。（当然：投资有风险，入市需谨慎。本项目是基于GAN的股票未来趋势预测，该模型具有一定的“保质期”，即使加入了适当的“情绪因子”，但由于股票市场日新月异的更迭，该模型也需要进行手动更新，以达到预期效果）

1. **项目内容**

本项目是基于GAN的股票未来趋势预测模型构建。首先我们采取一些非传统的指标同时结合一些传统的技术和财务指标对沪深300以及中证500的股票进行初步的筛选，然后利用GAN模型对股票池中的股票进行排名打分。

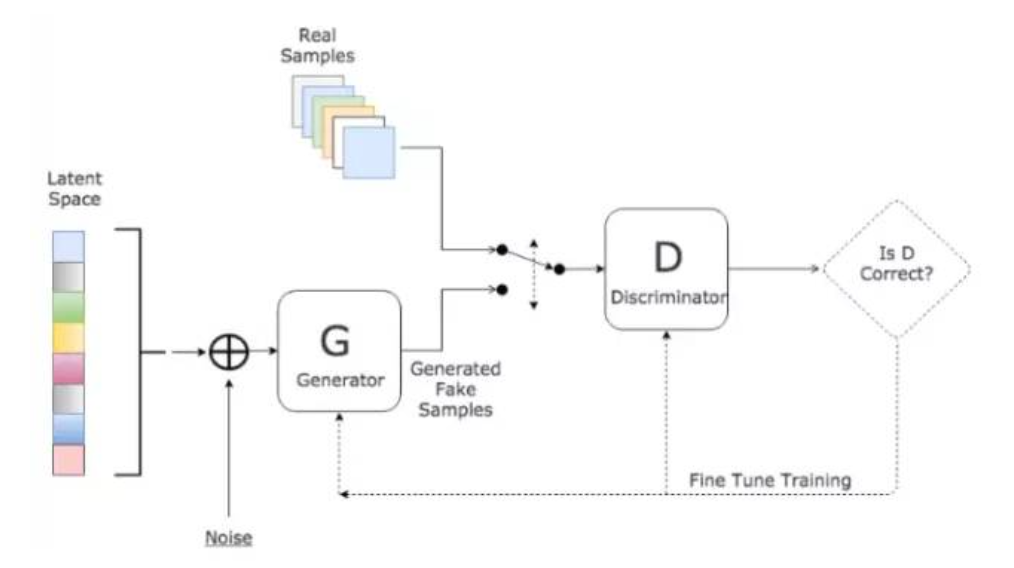
在GAN对于未来的趋势预测方面，我们将引入行业指标、综合指数、技术指标等参数，让模型进行自主学习，通过一些外在指标的变化因素，对股票池中的股票进行打分排名，最终拟合出股票的未来走势曲线。在模型的训练中，有两个重要的部分，它们分别是生成器和判别器。计算机通过对于现有真实数据的学习，找出股票价格走势的规律，然后对于我们给出的实验数据进行股票价格走势的模拟。在生成器生成未来股票价格预测趋势之后，数据会先进入判别器中，通过判别器识别数据是否是真实的，如果是真实的那就输出结果；如果识别出来数据是假的，那就返回生成器进行再一次预测。直到判别器无法判别出预测的数据到底是真实的数据或者是拟合出来的数据为止，从而最终实现对于个股为俩趋势的预测。

**3、项目编程实现**

本项目的编程是基于Python本地IDE进行的，依据已有的计算机深度学习模型嵌套，逐步实现项目的既定功能。

**4、项目结果和分析**

生成对抗网络（Generative adversarial network，简称GAN）。GAN由两个主要网络构成：一个是Generator Network,称为生成器;另一个是Discriminator Network, 称为判别器。GAN生成对抗性网络的核心逻辑是生成器与判别器之间的相互对抗和相互博弈，其主要表现在 当模型GAN迭代次数达到一定数量级时，此时生成器和判别器达到一个平衡，即判别器无法判断生成器产生的数据的真假性，从而达到模型GAN预测未来数据趋势的最终目的。



GAN的大致训练流程如下(以训练GAN生成股价未来趋势为例)。

第一步: 根据数据集的大小，设置合理的参数，初始化生成器G和判别器D。

第二步:在每一轮历史数据集的训练中，执行如下步骤。

(1)固定生成器，训练判别器的参数，使之收敛

a.因为此时生成器的参数没有收敛，所以生成器通过未收敛参数生成的未来股价不是特别真实、准确。

b.从股票历史数据集中有序选择一组历史数据。

c.通过上面两步操作，此时就有了两组数据:一组是生成器生成的股价趋势数据;另一组是真实股价历史数据。通过这两组数据训练判别器，让其对真实股价趋势赋予高分，给生成股价趋势赋予低分，最终达到收敛判别器参数的目的。

(2)固定判别器，训练生成器的参数，使之收敛

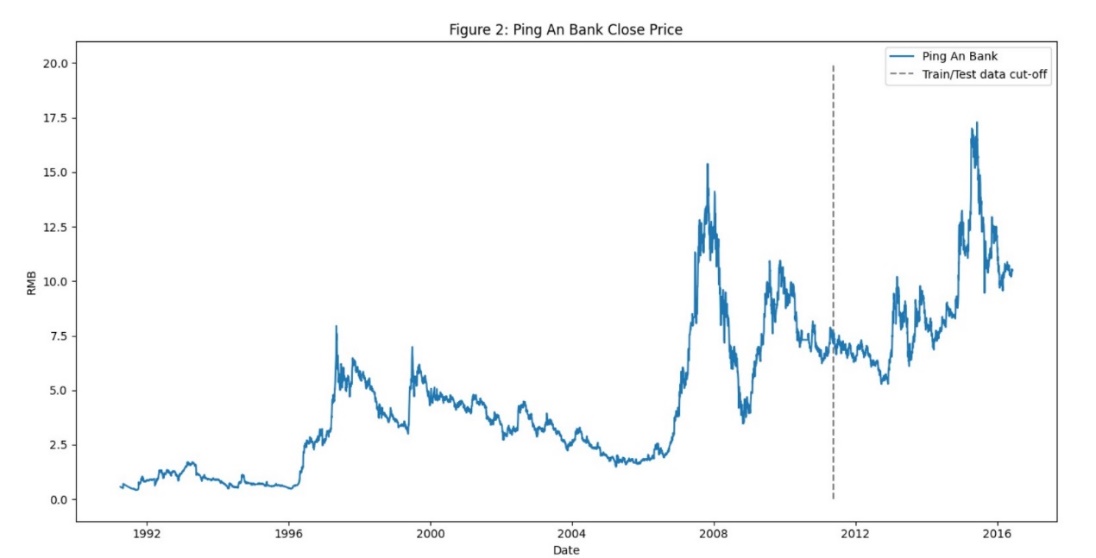
a.依托GAN中的模型LSTM（递归神经网络），给予生成器一组随机的噪声，使其预测未来股价趋势。

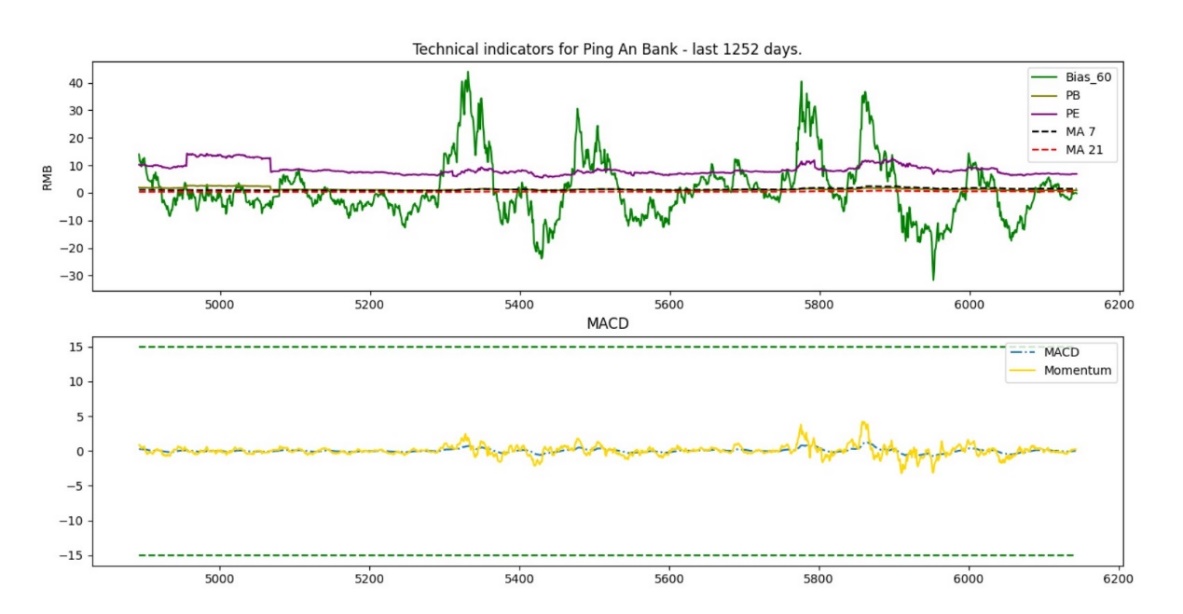
b.将生成的股价趋势传入判别器中，此时判别器会给该股价趋势一个分数（例如分数可能是0.5，其中分数范围在（0，1）的区间内，0代表该股价趋势完全错误，1代表该股价趋势完全正确），而生成器的目标就是使这个分数更高，生成出判别器可以赋予高分的股价趋势，此时也就代表了生成器生成的未来股价趋势在判别器的约束下越来越真实、准确，最终达到收敛生成器参数的目的。

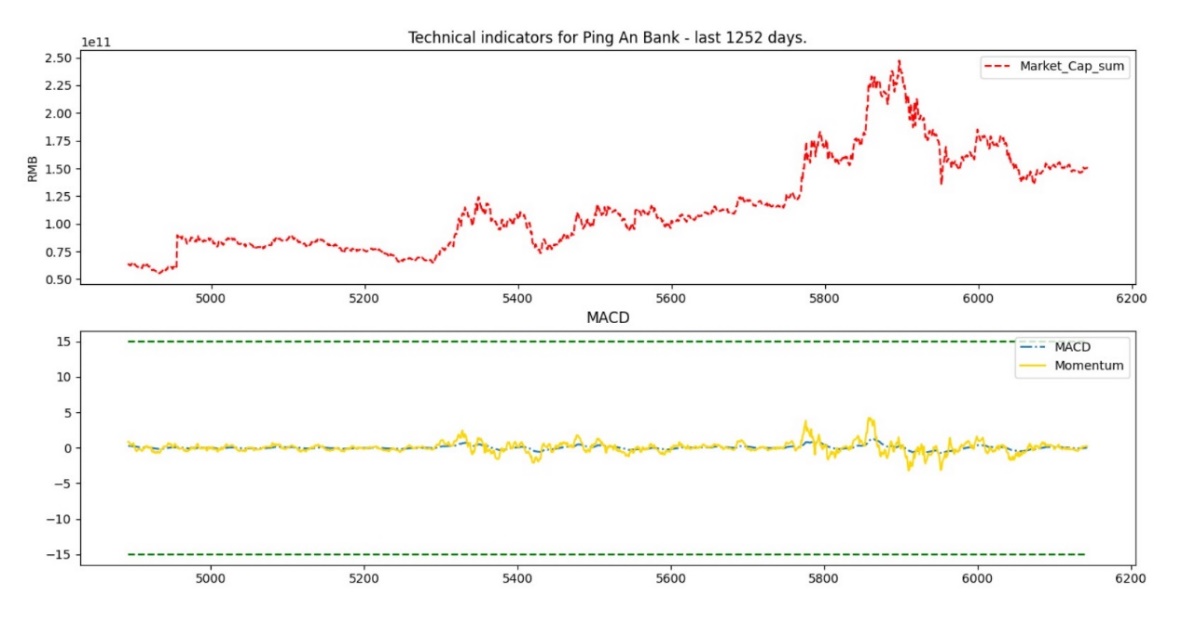
对于时序数据的未来趋势预测，传统方法是使用RNN循环神经网络。RNN循环神经网络相较于CNN与狭义DNN已经赋予神经网络一定的记忆效应，即当前新的记忆状态为下一个特征的输入产生决策作用，也就是输出依赖于输入的记忆。在进行股票未来趋势的预测时，我们需要将该股票历史25年的参数指标作为数据集，对该股票的未来趋势进行预测。但由于样本数据集过于庞大，RNN循环神经网络将面临遗漏数据集中重要信息以及梯度消失的问题，从而无法记住所有的输入数据、每次输出结果对下一次输入数据的影响，最后导致RNN循环神经网络对该时序数据的未来趋势预测产生偏差。所以当进行对股票的未来趋势预测的时候，我们将使用模型GAN，其中包含的模型LSTM（递归神经网络）内含的“门机制”能有效解决短时记忆问题以及梯度消失问题。在传统应用范围内，模型GAN主要用于图像生成、视频剪辑和数据增强等方面，很少有利用GAN对时序数据进行时间序列预测的项目。但是我们考虑到：股票的行为模式在历史时间序列上应是大致相同的，因此我们希望模型GAN根据已有的历史交易数据分布，生成与之类似的股票未来趋势的数据，同时我们也将尝试解决过拟合与对偏方差进行权衡，优化传统的时间序列模型。

数据集分割以及对个股经典指标参数呈现

我们以00001.SZ平安银行为例，爬取其历史25年的数据（1991.04.03-2016.06.08），并将数据集按照7：3的比例分割，其中7代表训练集（18年），3代表测试集（7年），后者用于验证模型GAN是否训练有素。同时，我们以个股000001.SZ 平安银行为例，对个股的经典指标参数加以呈现。







GAN预测的指标选取

为了了解有哪些因素影响了股票的未来涨跌幅情况，我们需要从不同的方面和角度来描述股票，同时利用GAN的预测结果进行排名打分，得到最终我们需要购买的股票。同时我们将使用长达18年的数据来训练模型中的各种算法，同时通过训练出来的模型对未来的股票情况进行预测。之后通过一次次模型GAN的训练最终使我们的预测与真实数据达到拟合，从而，生成基于历史时间序列的统计特性，且建立在假设未来不发生黑天鹅时间的未来股价预测趋势。为此，我们将选用下述指标作为特性（部分省略，只展示与机器学习相关部分）：

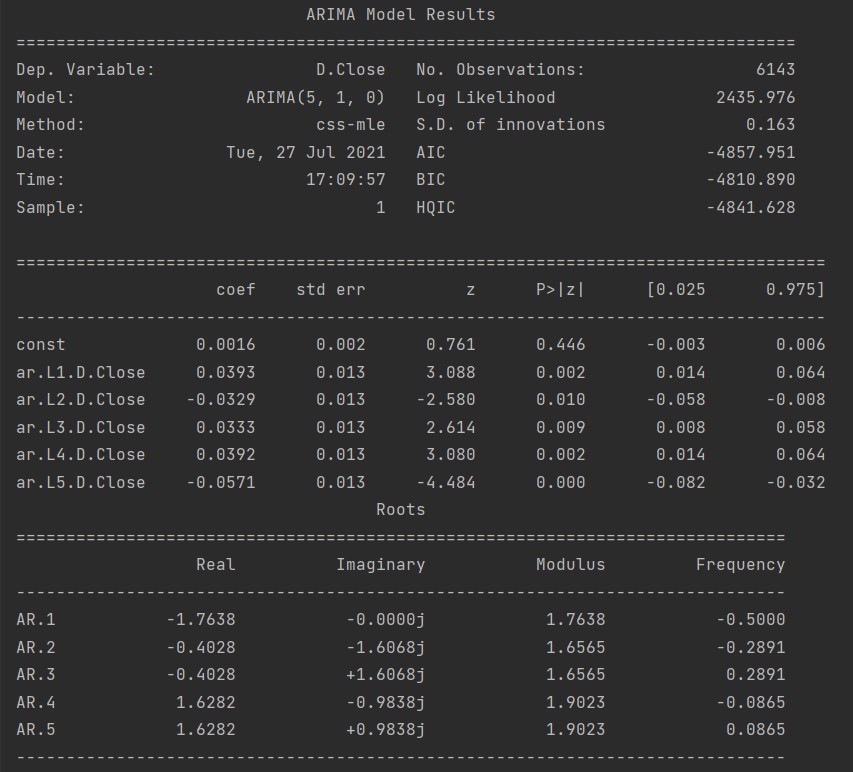
基本面分析

我们利用python通过同花顺财经爬取该股的相关新闻。与此同时我们会对每个特性进行重要性的判断。为了到达我们设想的目标，我们使用BERT，（一种预先训练的模型）。并且我们也将直接调用训练有素的BERT模型。

最后使用sigmoid函数，结果将在0到1之间。得分越接近0，负面消息就越多（越接近1表示正面情绪）。对于每一天，我们将创建平均每日分数，并将其添加为股票的其中一个特征。

ARIMA

ARIMA是一种时间序列预测数据的方法，我们希望通过ARIMA来降低库存的噪音，并提取一些新的模式与特性。



统计检查：

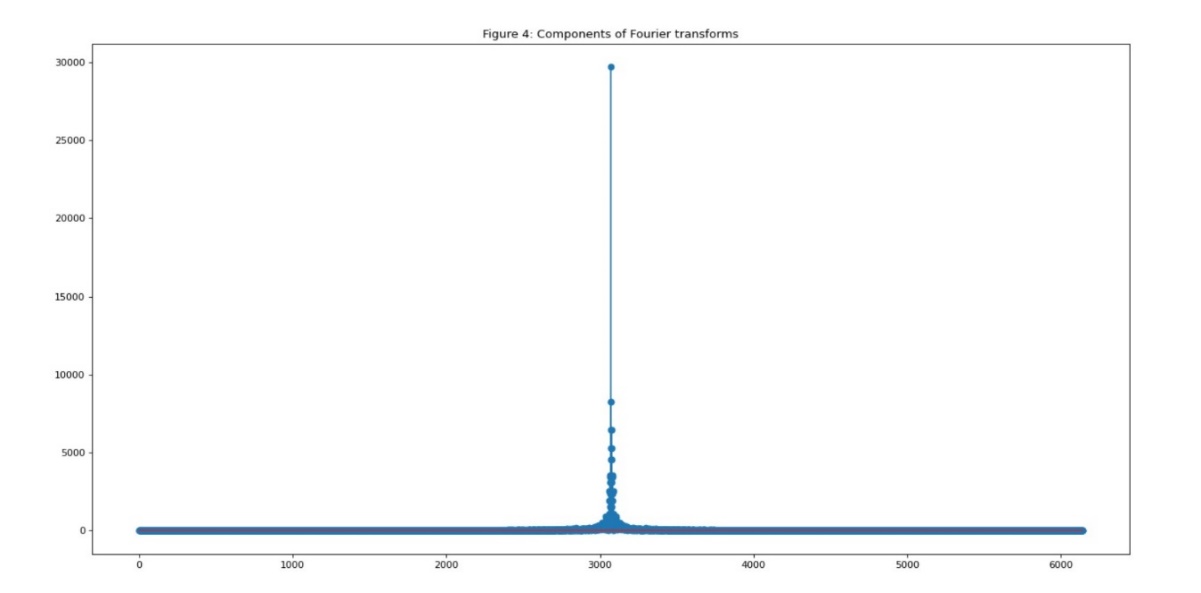
为了确保数据的准确性，我们将利用一些高等数学的知识去做检测：

1、条件异方差发生在误差项（通过回归得到的预测值与实际值之间的差）依赖于数据。

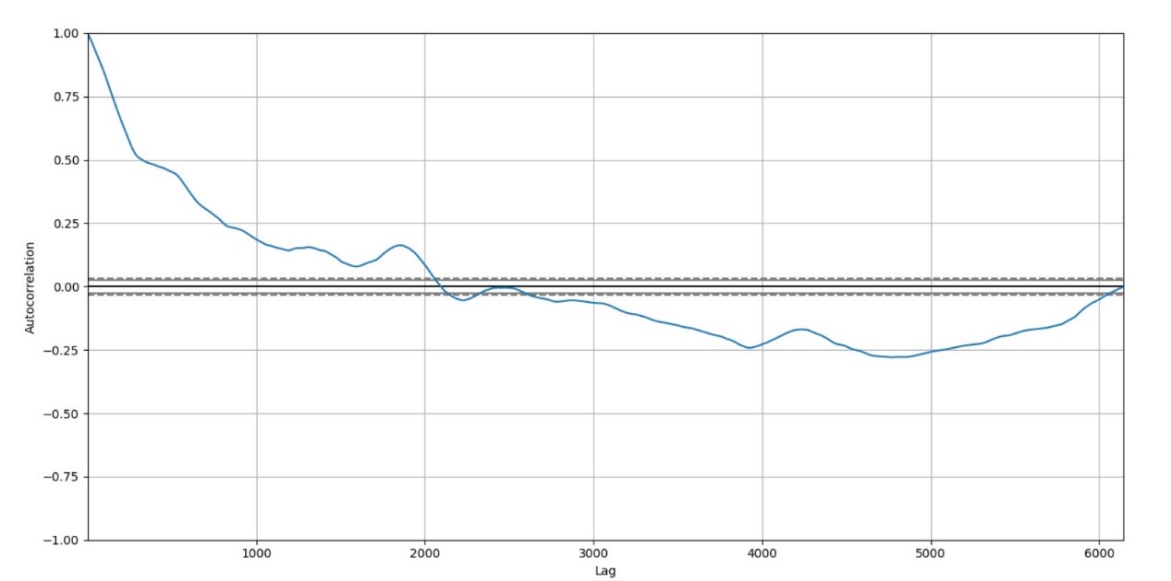
2、多重共线性是指错误项（也称为残差）相互依赖。

3、序列相关性是指一个数据（特征）是另一个特征的公式（或完全不相关）。

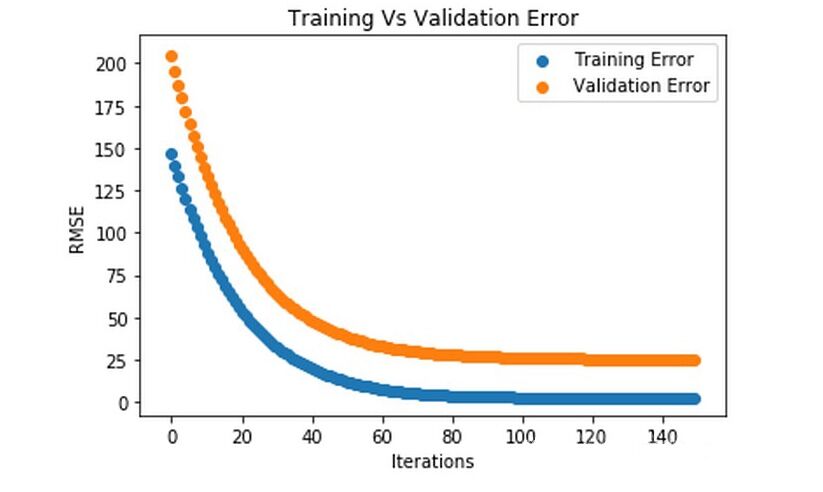
通过统计数据调查，我们确信数据是准确无误的。以下分别是数据降噪、数据自我相关性误差矫正与训练集模型收敛结果的示意：



数据降噪



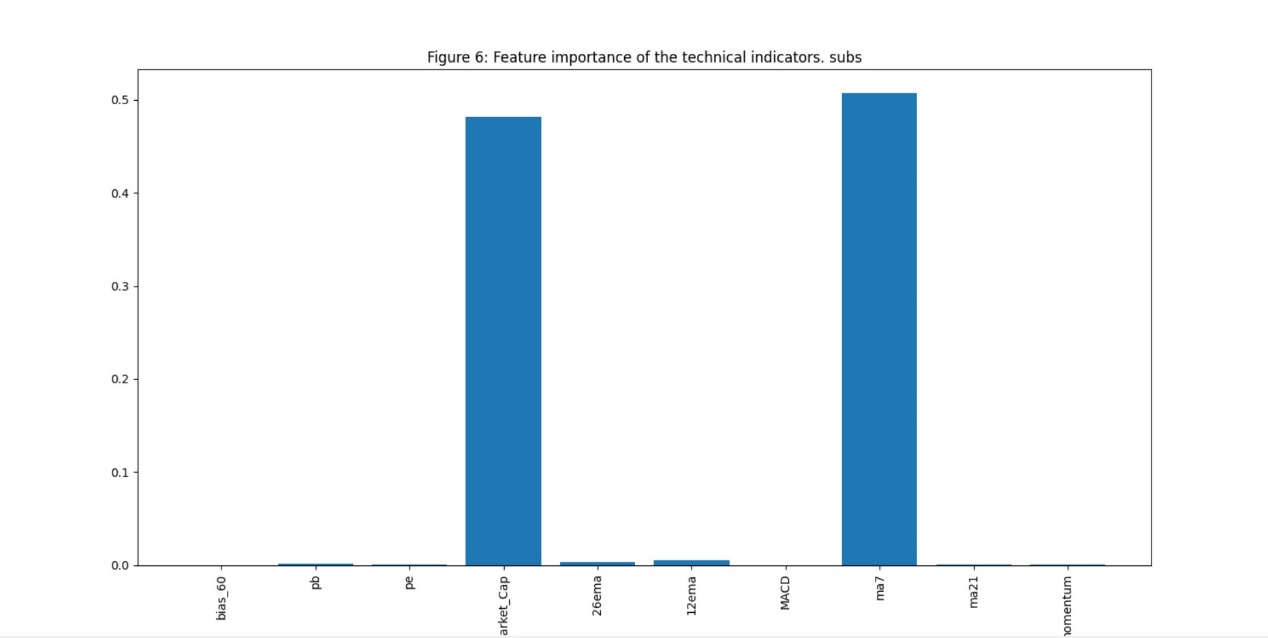
数据自我相关性误差矫正



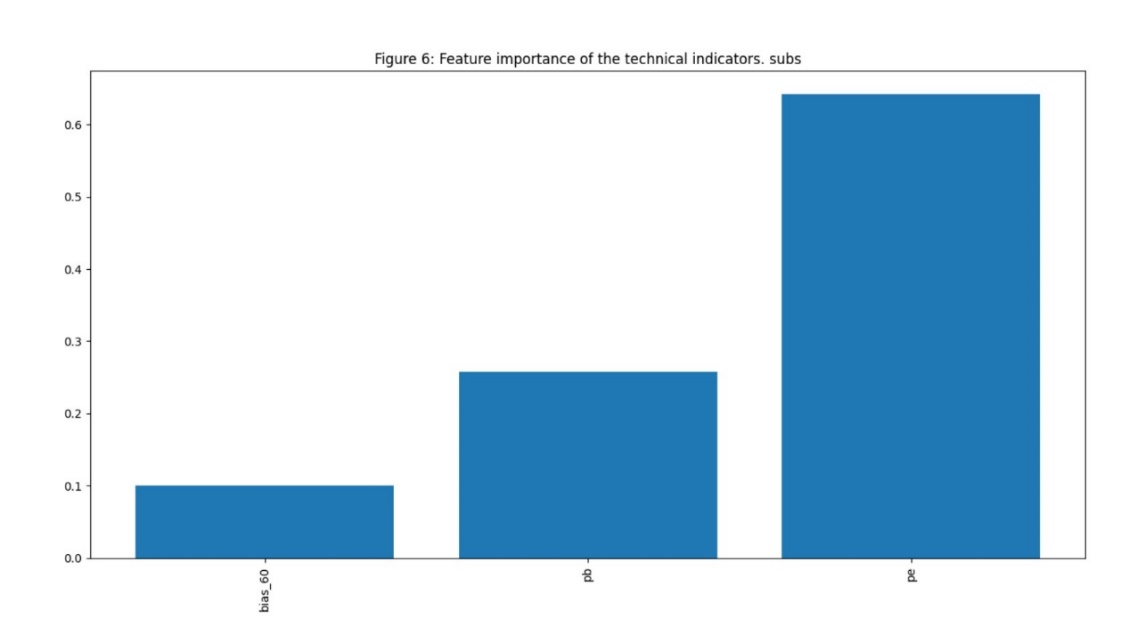
训练集模型收敛结果

特征工程

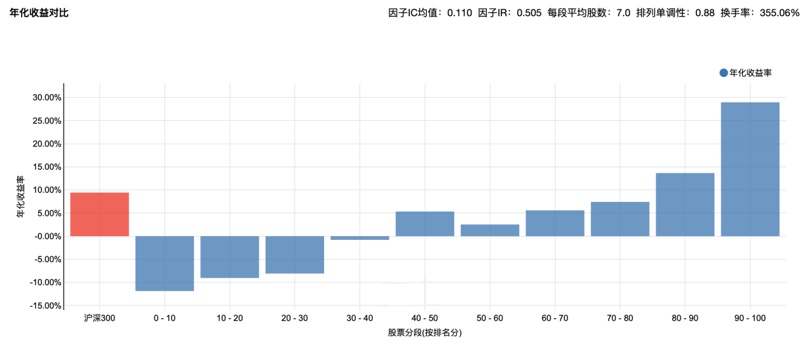
通过上述操作，我们添加了所有类型的数据后，我们发现在这3年中总共有97个特征，正如前文所述，训练时间是18年，而且我们也会由自动编码器中生成更多的特征。但是，显然不是所有的特征都能给我们正确的指向型，我们希望通过回归算法XGBOOST来测试特征重要性，但是这显然是一个计算量庞大的过程，为此，我们希望通过简单的梯度，来建立特征重要性的正确逻辑。



合理特征重要性示意图\_1



合理特征重要性示意图\_2

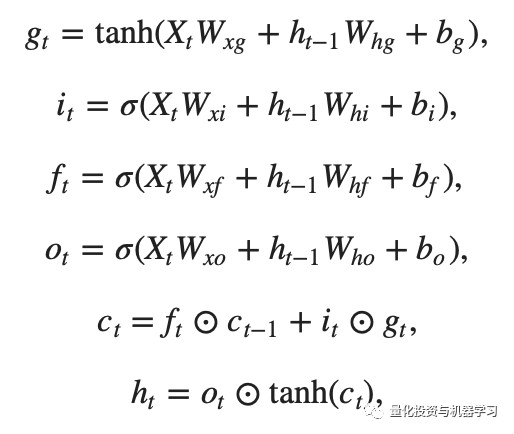


特征数据收益对比

生成器LSTM运作

GAN模型中的生成器运用的是LSTM网络，这是一种递归神经网络（RNN）。一般情况下RNN会用于跟踪时间序列的数据，同时了解该数据随时间发展变化的关系。我们希望于此，创建对股价的预测。但因为RNN循环神经网络将面临遗漏数据集中重要信息以及梯度消失的问题，最终的结果无法达到收敛。

目前在GAN中主流运用的解决上述问题的方法主要有两个，它们分别是门控循环单元（GRU）和长短期记忆（LSTM）。我们选择LSTM，LSTM单元背后的数学是：



抛开上述的表述，简单来说LSTM的整体逻辑是：我们将获取17天的数据（同样，这些数据是股票每天的股价+当天的所有其他特性——相关资产、情绪等），然后尝试预测第18天的数据。然后我们用一天移动17天窗口，再次预测第18天。我们这样迭代整个数据集。

但是因为LSTM中有很多特征和神经网络，我们需要确保我们避免过拟合，并注意总损失。于是我们采用了这两种方法：

1、确保数据质量。我们已经进行了统计检查，确保数据不受多重共线性或序列自相关的影响。进一步，我们对每个特性执行了特性重要性检查。最后，利用相关股票市场运作机制的领域知识进行初始特征选择。

2、正则化。最常用的两种正则化技术是LASSO (L1)和Ridge (L2)。在这两种情况下，L1和L2正则化模型的参数都“收缩”，但是在L1正则化的情况下，收缩直接影响模型的复杂性(参数的数量)。准确地说，岭回归在最小二乘估计方差较大的情况下最有效。通过查阅相关文献，我们不难发现L1对于有异常的数值更加的敏感，且当数据量不大的情况下，使用L1更加的有效，同时依据这个特性我们发现与我们的情况更加的契合，所以我们使用L1。

除了上述两种方法，在构建复杂神经网络时，我们需要考虑的另一个重要的因素是偏方差权衡。基本上，我们得到的错误当训练网是一个函数的偏差，方差和不可约错误-σ（由于噪声和随机误差）。我们使用了相对简单的权衡：

**Error=bias^2+variance+σ**

Bias偏差衡量的是一个训练过的（在训练数据集上）算法在不可见数据上的泛化程度。高偏差（欠拟合）意味着模型不能很好地处理不可见数据。

Variance方差度量模型对数据集中的更改的敏感性。高方差是过拟合。

判别器CNN运作

一般而言，我们通常使用CNNS进行与图像相关的工作。它们在从特征中提取特征方面非常强大。那么将同样的思路带入到股价趋势中，数据点形成小趋势，小趋势形成大趋势，趋势反过来形成模式。也就是说CNNS检测特征的能力可用于提取有关股价走势模式的信息。

此外，CNNS可以很好地处理空间数据——这意味着彼此距离较近的数据点之间的关系比分布在各处的数据点之间的关系更密切。基于认知，两天的时间越近，它们之间的关系就越密切，所以这对于时间序列数据整体上是正确的。

GAN训练200期后，它将记录（LSTM的误差函数,GG）并将其作为奖励价值传递给强化学习，将决定是否改变hyperparameters与同一组hyperparameters保持训练。

如果RL决定更新超参数，它将调用贝叶斯优化库，该库将给出下一个最佳预期超参数集。

在我们的实际训练中遇到了这几个问题：

一、梯度消失

当我们使用的判别器CNN越接近最优，最小化生成器LSTM损失就越近似最小化JS散度，但是在实际中，我们的生成数据分布与真实数据分布之间极大概率不存在不可忽略的重叠部分，因此，两个分布无论它们之间的距离究竟是多少，其JS的散度都表现为log2，这种现象最后也会导致梯度消失的结果，即生成器的梯度始终近似为0，无法进一步对于LSTM的生成器进行优化。

二、训练时梯度不稳定

三、模式崩溃

简单总结，最小化生成器的目标函数等价于，我们不仅仅要减少真实的数据与生成器生成的数据之间分布的距离，同时我们也要增大这个距离。这稍微有点不合理，导致了GAN训练时梯度不稳定，以及GAN训练时容易崩溃，导致生成数据多样性不足。

原始GAN与WGAN对比

|  |  |  |
| --- | --- | --- |
| 原始GAN | DCGAN | WGAN |
| 1、训练难度大  2、GAN中的生成器与判别器的loss难以表明训练的进程  3、生成的样本多样性不足 | 1、当失去特殊的架构设计模型就容易崩。  2、通过重复实验对判别器与生成器的架构进行罗列,由此能在其中找到一组相对较优的网络架构设置。 | 1、WGAN能够从根本上解决GAN在训练时不够稳定的问题,由此使得生成器与判别器在训练程度达到动态平衡。  2、它几乎解决了collapse mode问题,在很大程度上保证生成样本多样性  3、在训练过程中有了一个准确的的数值来衡量训练的进程,当这个数值越来越小时就代表GAN的训练成果越来越好,同时它也能指代生成器所产生的图像质量的高低。  4、我们只通过最简单的多层全连接网络即就可以达到上述的目的。 |

鉴于上述模型的对比，我们借鉴GAN研究的最新进展，Wasserstein GAN - WGAN进行问题的应对。

GAN背后的主要目标是让生成器开始将随机噪声转换成我们想要模拟的某些依据特定的历史时间序列的数据。因此，在GAN中，比较两个分布之间的相似性的是非常必要的。最广泛使用的两种指标是：

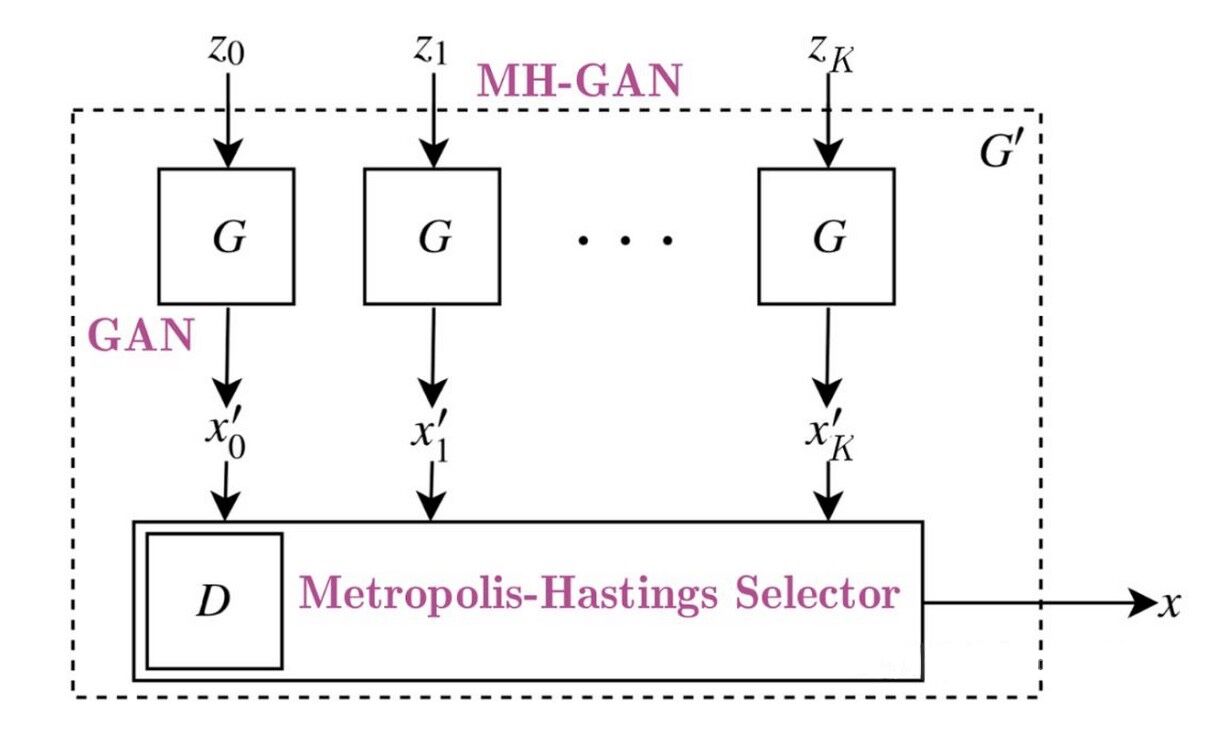
1、KL divergence

2、JS Divergence

JS Divergence以0和1为界，与KL divergence不同，它是对称的，更平滑。当损失从KL转移到JS散度时，GAN训练取得了显著的成功。但是，无论是JS散度还是KL散度都存在一些bug，因此在WGAN中，完全抛弃了这两个。

WGAN使用Wasserstein distance，W(pr,pg)=1Ksup‖f‖L≤K

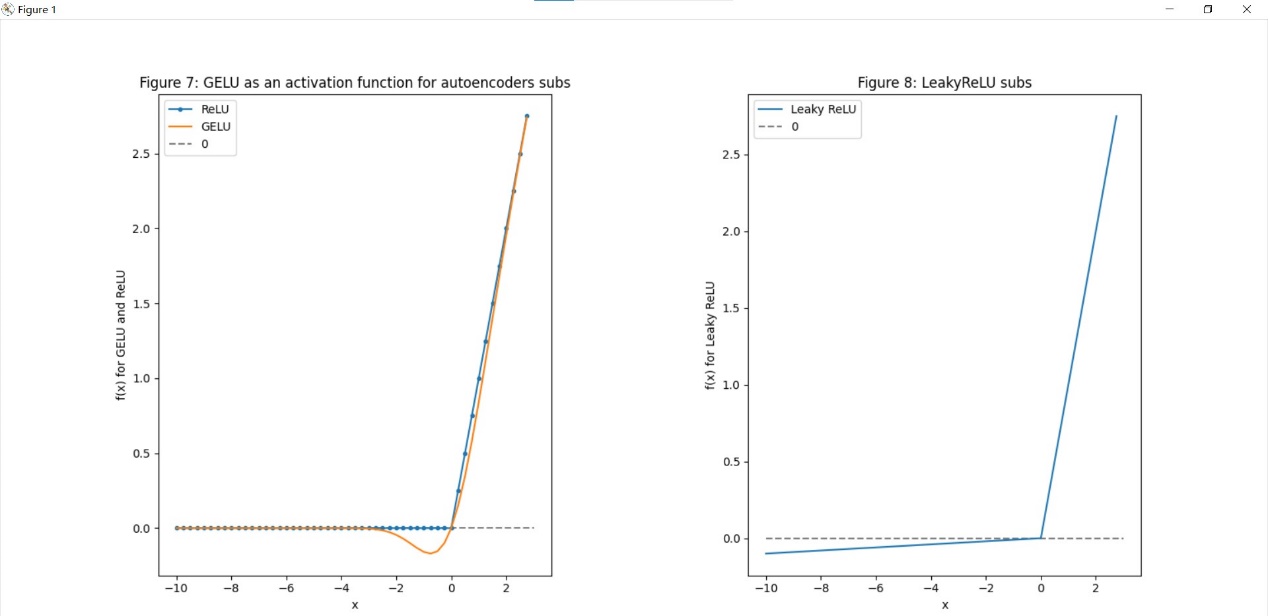
而且，与KL和JS相比，Wasserstein距离几乎在任何地方都是可微的。正如我们所知，在反向传播期间，我们对损失函数进行微分，以创建梯度，从而更新权重。



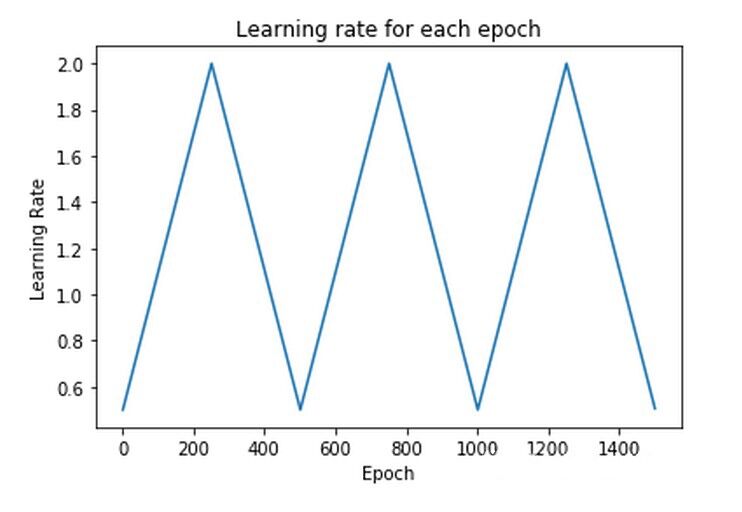
GAN模型的最终训练实例

通过傅里叶变换拟合函数、激活函数Relu过程，我们对模型GAN进行合理的训练，最终得到了较好的预测股票未来趋势的结果。

以000001.SZ平安银行为例，以下分别是激活函数Relu以及模型GAN的学习率在每次迭代中发生的变化示意图。

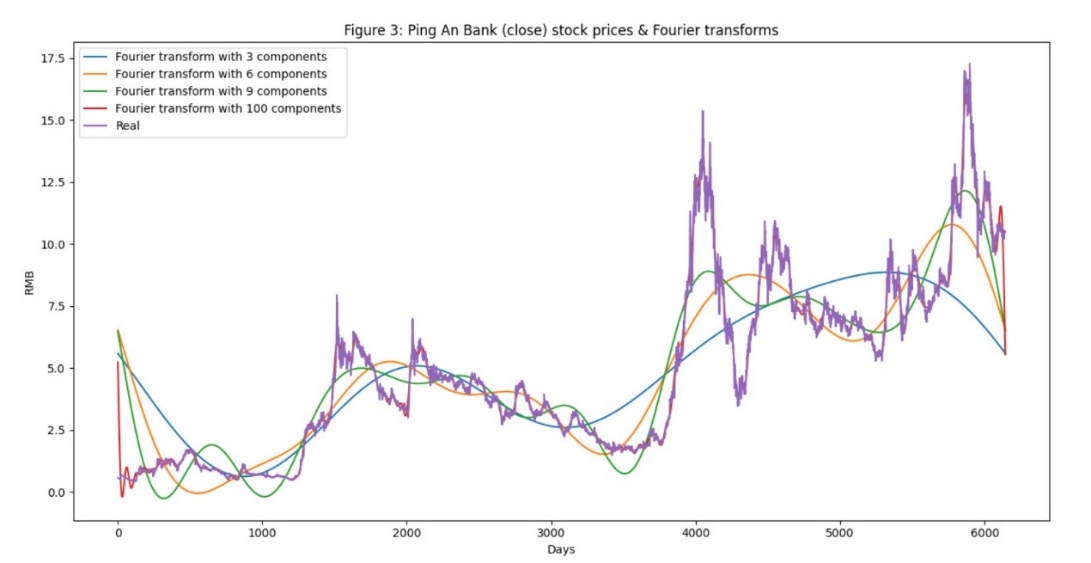


激活函数Relu

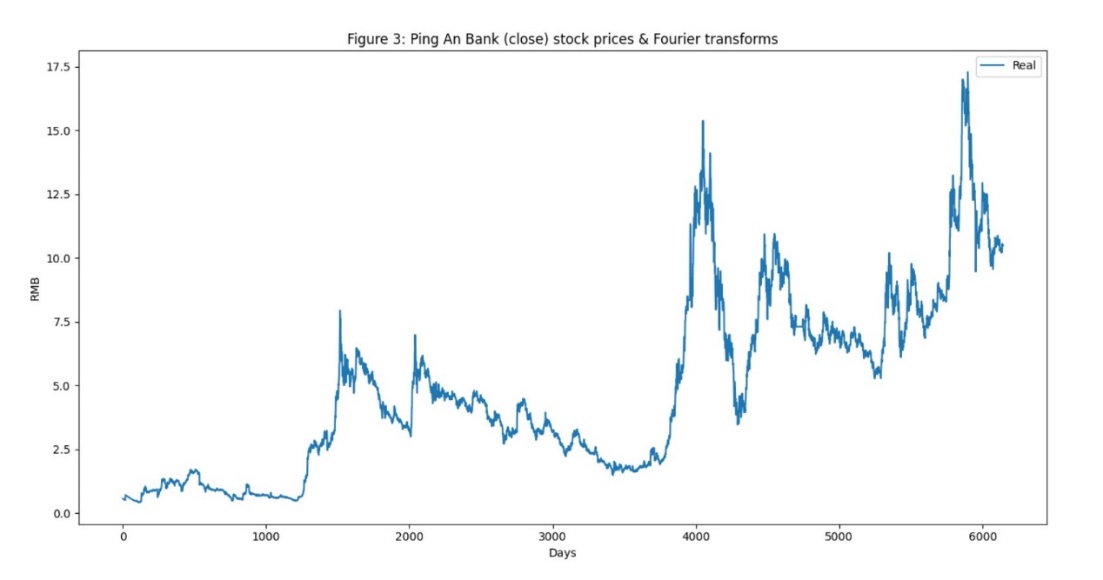


模型GAN的学习率的变化

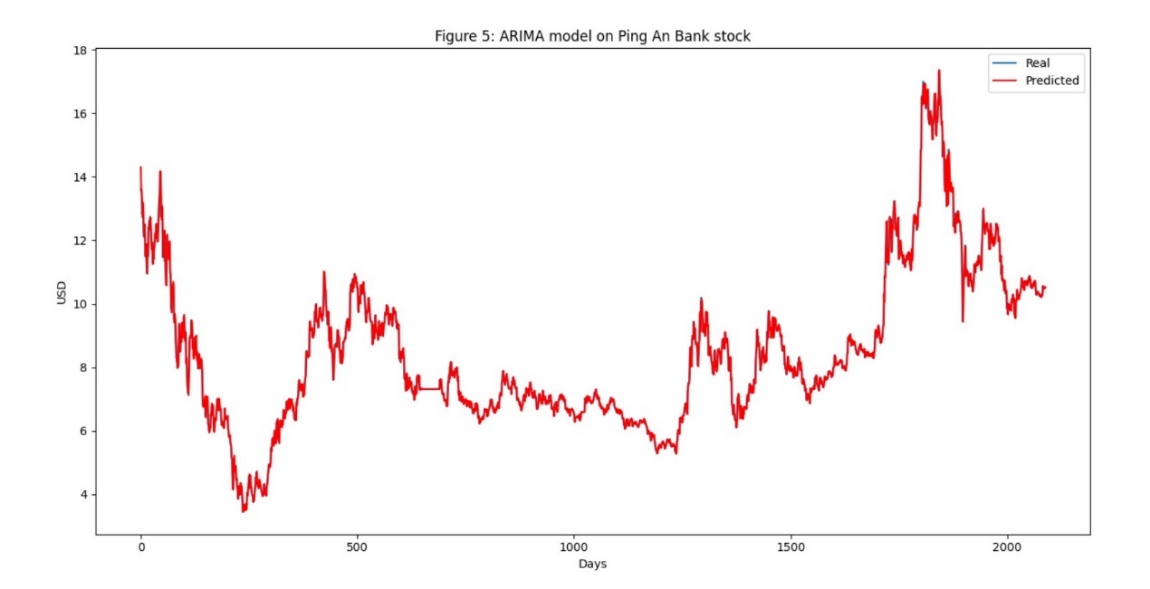
通过GAN对平安银行000001.SZ未来趋势预测的示意图如下



未来趋势预测的示意1



未来趋势预测的示意2

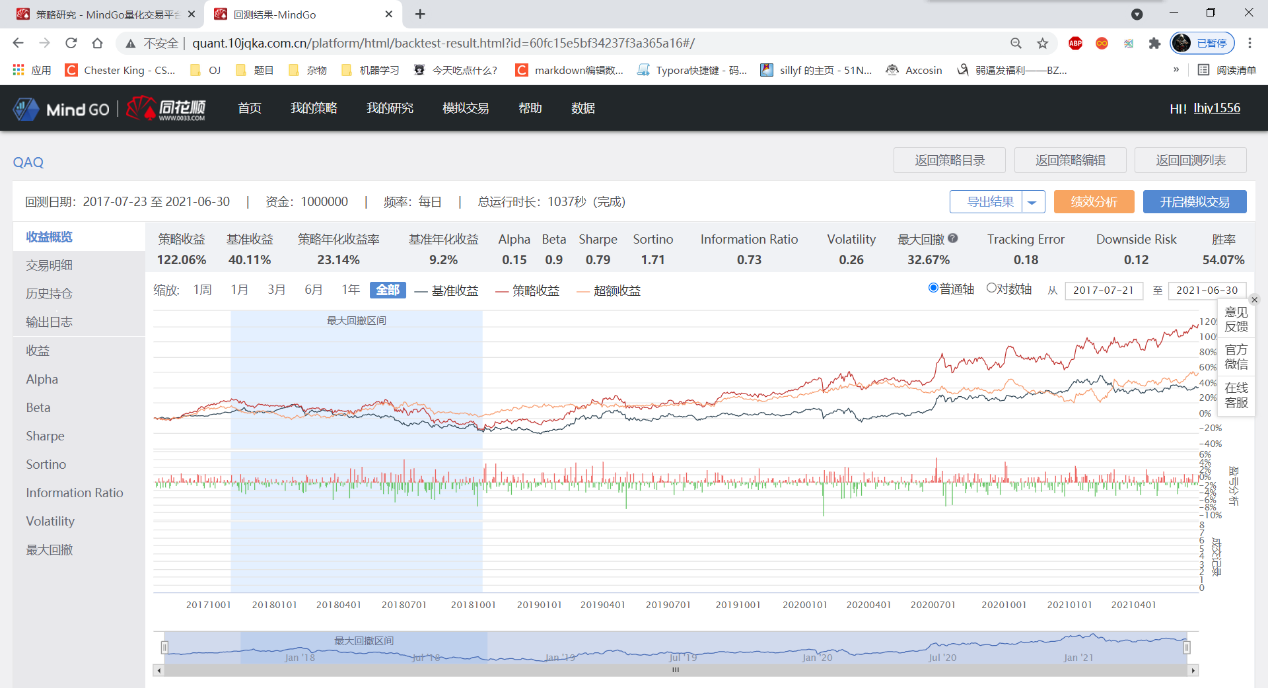


未来趋势预测的示意3

上图较好证明了模型GAN是训练有素的，其在历史时间序列以及现实股票市场内都展现出较好的训练结果。

回测结果展示：

回测时间：2017.6.30-2021.06.30 基于mindgo 平台半年期回测，策略收益122.06%，基准收益40.11%，超额收益81.95%



**5、项目总结**

通过此次基于GAN的股票未来趋势预测项目（该项目同时也参加了浙江省证券投资比赛，现在暂时在130支队伍中排名第29），对机器学习中时下的GAN模型进行了比较透彻的理解，同时接触了计算机的深度学习在实际生活中的应用，感受颇多，做完模型的成就感很大。

**6、自己的话**

概括模型GAN：计算机中的生成器模拟人类小时候学习的过程，判别器就像是考试，对你学习后的成果进行检验。训练集就像是平时的作业，测试集就像是正式考试；情绪因子就好比考试的时候突发状况（肚子疼、饿了、想上厕所等等），最终都会影响期末大考的结果。但其他一些大的环境因素（比如考试的时候有人影响你、监考老师对你“格外照顾”等等），这些“参数”需要对代码进行人为的调参，否则模型可能会失去效果。

此次比赛为了学习模型GAN还恶补了其他附着于模型GAN的东东（LSTM、ARIMA、Bert入门等等），就是可能当时时间有点紧张，现在回想起有些知识点有些模糊，需要后续再不断补充、学习。

显然，完全依赖于GAN对股票未来趋势的预测是不可能的（毕竟如果成真了那岂不是写个代码封装一下卖股民，大家互利互惠），这只是深度学习在实际生活中的应用，仅供参考（深度学习练练手、比赛包装）。

再接再厉。

**7、参考资料**

1、利用深度学习最新前沿预测股价走势

2、[20-Book-李沐]动手学深度学习