摘要：

近日，我国对外开放金融市场，而同时大数据行业也正蓬勃发展。不难发现，金融与大数据的结合已是未来大势所趋，势必会为金融行业添加新型力量。金融市场积淀多年的海量数据，涵盖了大量金融市场的潜在规律。而如何挖掘这些数据的潜在价值，提取其中普适性，精炼性的信息就显得十分重要了。利用这些信息，我们可以更充分地了解金融市场，及时把握市场走向，磨砺完善我国金融手段，从而有更大底气与信心，面对外来投资者的挑战。

本项目意在从股票资讯平台，搜集股票的历史数据，以作为模拟股票操作的战场，检验与调整短线抢反弹，中线区间法，机器模型预测法三个股票量化策略，并横向比对，以凸显各自优劣，再在此基础上加以完善。

关键词：量化策略，机器学习，交易信号

绪论

1. 项目背景

近日，我国对外开放金融市场，而同时大数据行业也正蓬勃发展。不难发现，金融与大数据的结合已是未来大势所趋，势必会为金融行业添加新型力量。金融市场积淀多年的海量数据，涵盖了大量金融市场的潜在规律。而如何挖掘这些数据的潜在价值，提取其中普适性，精炼性的信息就显得十分重要了。利用这些信息，我们可以更充分地了解金融市场，及时把握市场走向，磨砺完善我国金融手段，从而有更大底气与信心，面对外来投资者的挑战。

本项目意在从互联网金融平台，搜集股票的历史数据，以作为模拟股票操作的战场，检验与调整短线抢反弹，中线区间法，机器模型预测法三个股票量化策略，并横向比对，以凸显各自优劣，再在此基础上加以完善。

1. 项目内容

本项目基于历史股票数据，对股票市场的交易策略进行研究。检验两个经典的股票策略和一个以机器学习为核心的新型量化策略的有效性，并比对各自优劣，分析总结策略适用场景。

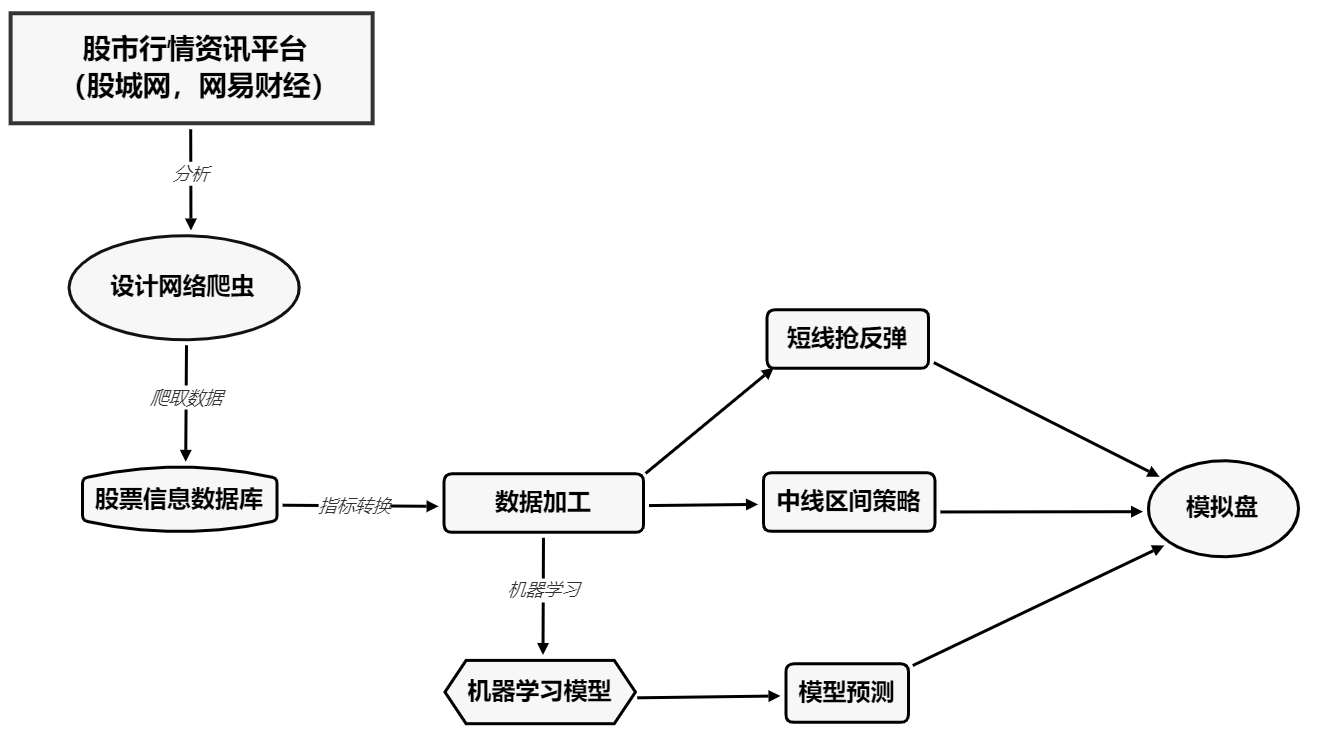
三个股票量化策略分别为：

短线抢反弹：当股价出现阶段性低价或暴跌时买入，在股价回涨的过程中逐步抛出；

区间策略：回望股票过去的历史数据，总结高估，偏高，正常，偏低，低估几个评级的股价区间，并以此作为买卖依据；

模型预测：将股票数据转换为股市分析的专用指标，运用机器学习提炼股市涨跌规律，并以此预测股票走向。

1. 项目组织框架



本报告内容将按如下顺序进行安排：

第一章为绪论，介绍本项目的项目背景以及研究内容；

第二章为数据采集，详述本项目所使用数据集的采集过程与部分爬虫代码示

例；

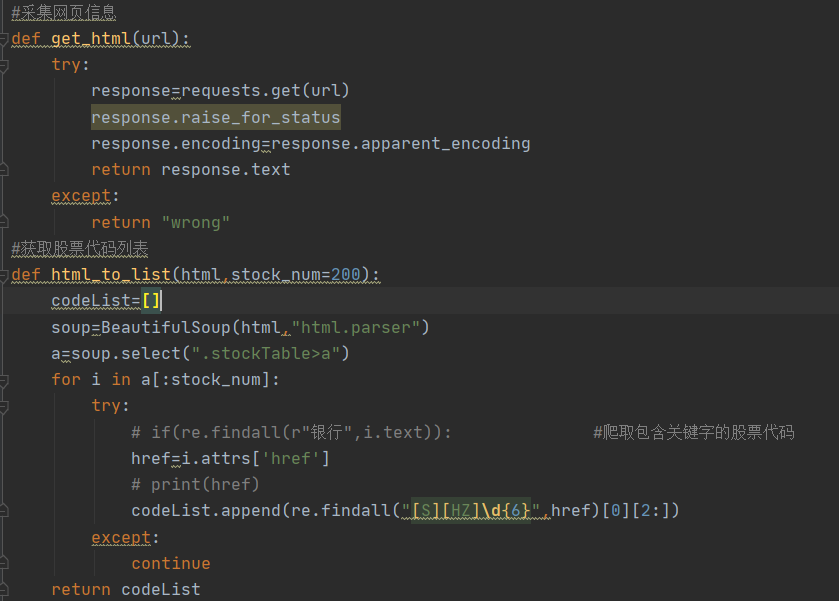
第三章，对本项目使用的数据集进行介绍，描述数据的预处理过程；

第四章，详细介绍本项目的实验过程，并对实验结果进行解释；

第五章，总结项目。

二，数据采集

查看各大股票咨讯平台，都未能发现直接显示股票历史数据的页面（多为当日股票数据和以k线图形式展示的股票历史数据），故常规爬取难度大。经过网上查阅，获得网易财经股票历史数据的接口网址。于是将爬取策略转变为，从股城网爬取股票代码，而后通过网易接口，得到股票历史数据的csv文件。



通过get\_html()爬取股城网的html数据，在html\_to\_list()的函数中，运用bs4对html进行解析，并采集得到所需股票代码的列表。



获得目标股票代码后，通过网易财经的接口，获得股票历史数据的csv文件，通过urlretrieve（）将文件下载保存至规定路径。

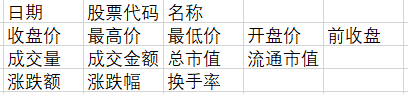


通过pymysql建立数据库，根据csv文件内属性建表，并以此插入数据。

三，数据集介绍与预处理

1. 数据集介绍

采集了前200个股票2006年以后至今的数据，信息包括以下内容：



数据预处理

1. 短线抢反弹

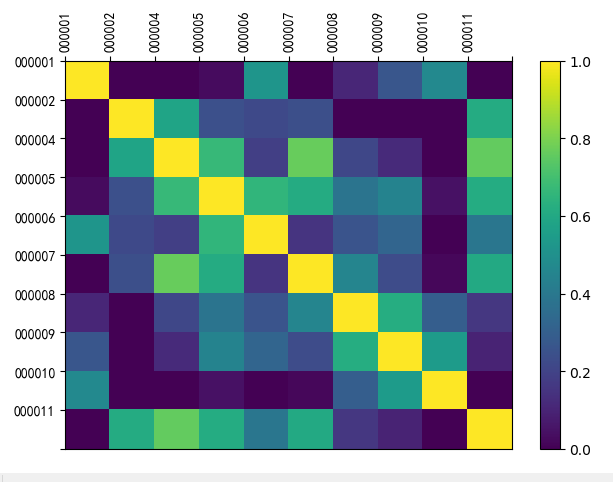
仅包括模拟交易时间内的收盘价信息。

1. 中线区间法

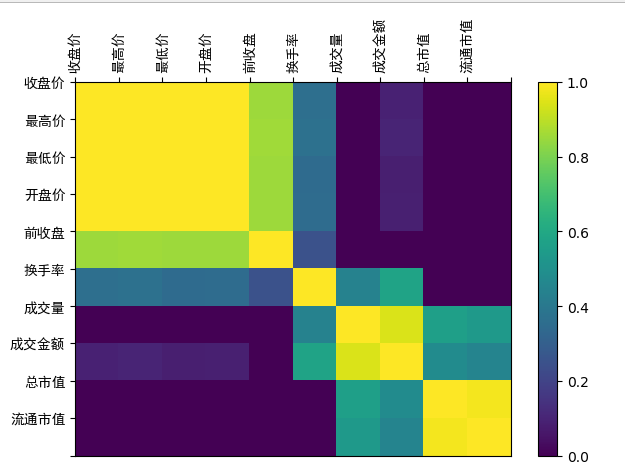
包括模拟交易时间内以及，模拟交易起始日期前250个交易日的收盘价信息。

1. 机器模型预测法

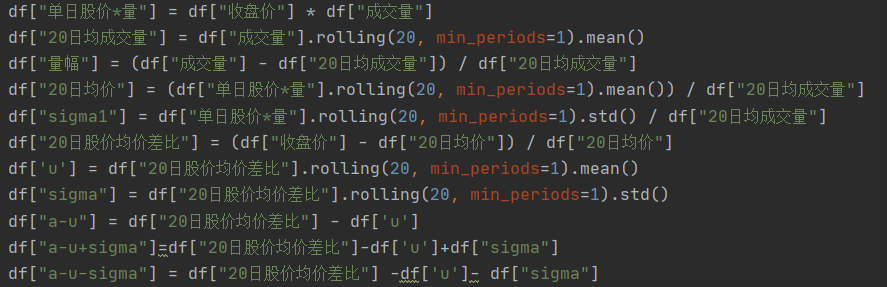
对部分股票进行了相关系数图的绘制，发现不同股票之间的相关性参差不齐，故难以用一个机器学习成果统一总结所有股票起伏规律。且经过实验发现，将机器学习的样本集扩大至多只股票时，不管如何调整机器学习的参数与样本属性，结果都差强人意。因此，决定采取定制化的机器模型。（即对在不同时间段的不同股票定制专属机器模型，从而提高预测准确率）



对于机器学习的样本数据选择，我将存在高度相关的一组属性中只取一个用于机器学习，以防止属性过多对机器学习产生的干扰，并提高学习效率。因此我选择收盘价，成交量，涨跌幅三个属性作为机器学习的母属性。



将母属性通过一定的处理，将之转换为下图中一系列的更具股市预测意义的属性，作为机器学习的样本属性选择。（并不是都要用，而是作为待选属性，以便对模型进行调整）



四、实验设计与实验结果

本章将设计实验，在股票历史数据的数据集上，对三种量化策略进行代码实现。并对机器学习的预测结果，以及三种量化策略的表现进行分析

（一）实验设计

1.模拟盘，模拟账户设计，以便于对三种量化策略进行评估。

定义一个账户Account的类，内部包含账户的股票市值，账户金额，股票成本，持股数量等一系列信息。同时支持对股票进行买入，卖出的操作，并实时更新股价波动对帐户金额的影响。

2.短线策略的实现

当股票价格跌至近三十个交易日最低时，即买入1成仓位，越跌越买，而后随着股价回涨获利过程中，逐步抛股。

3.中线区间策略的实现

以模拟交易时间段的起点为锚，向前统计300个交易日的价格变动。并划分股价的高估区，偏高区，正常区，偏低区，低估区。当股价到达偏低区时买入，到达低估区则加速买入；当股价到达偏高区时卖出，到达高估区则加速卖出。（当模拟交易时间段大于300个交易日时，为保证区间的有效性，会重新划分前300个交易日的价格区间。）

4.机器模型生成

目标：让机器模型预测未来七日的涨跌，若涨幅超过3%，则发出买入信号(1),当跌幅超过-2%，则发出卖出信号(-1),若在-2%与3%之间，则发出持有信号（0）。

样本集：对某一股票一年的股票数据，进行处理作为样本集；同时便历这一年的股价变动情况，采集股票七日涨跌的信息，从而得到买卖信号集，作为答案，对机器学习进行监督。

训练模型：将上述样本集与监督集随机分为训练集与测试集，比例为80%：20%，代入使用随机森林的机器学习模型进行训练。在实验结果的评估环节，本项目选用了分类问题常用的confidence（可信度）指标对预测结果进行评估。

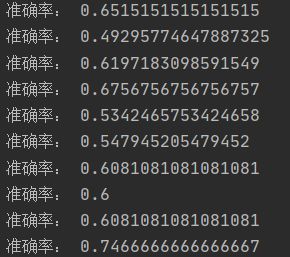
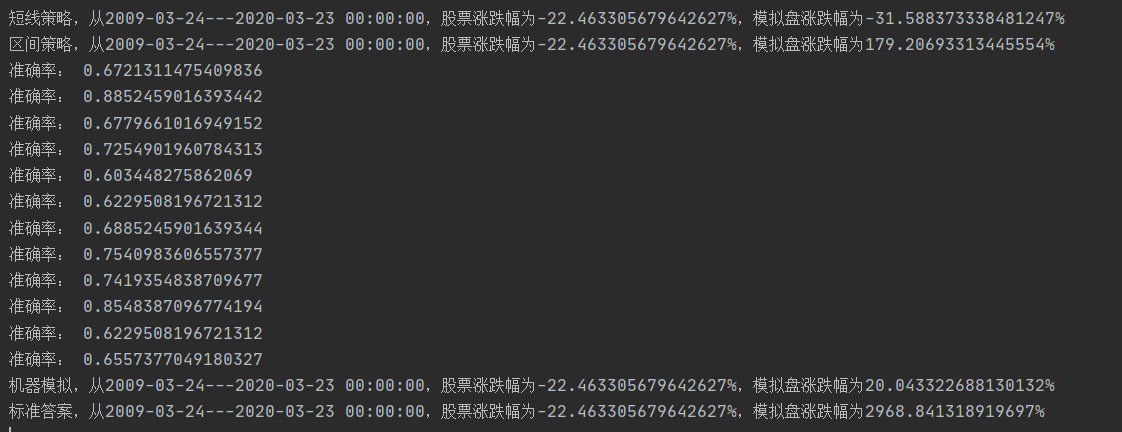
5.模型预测策略的实现

以模拟交易时间段的起点为锚，将这之前一年的股票数据作为样本集进行机器学习，生成机器模型（考虑到模型的时效性，将每半年更新一次模型），从而对之后行情走向进行预测，给出包含买入（1），持有（0），卖出（-1）三个指令的集合，账户根据指令进行操作。为了避免频繁交易，还设置了quota，作为每两次买入卖出操作的最小间隔，以进行限制。

（二）结果评估

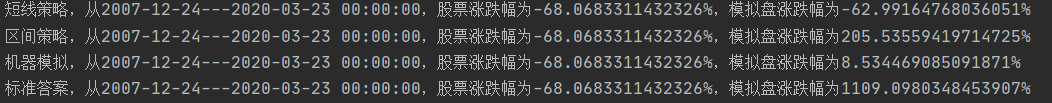
1.机器学习成果

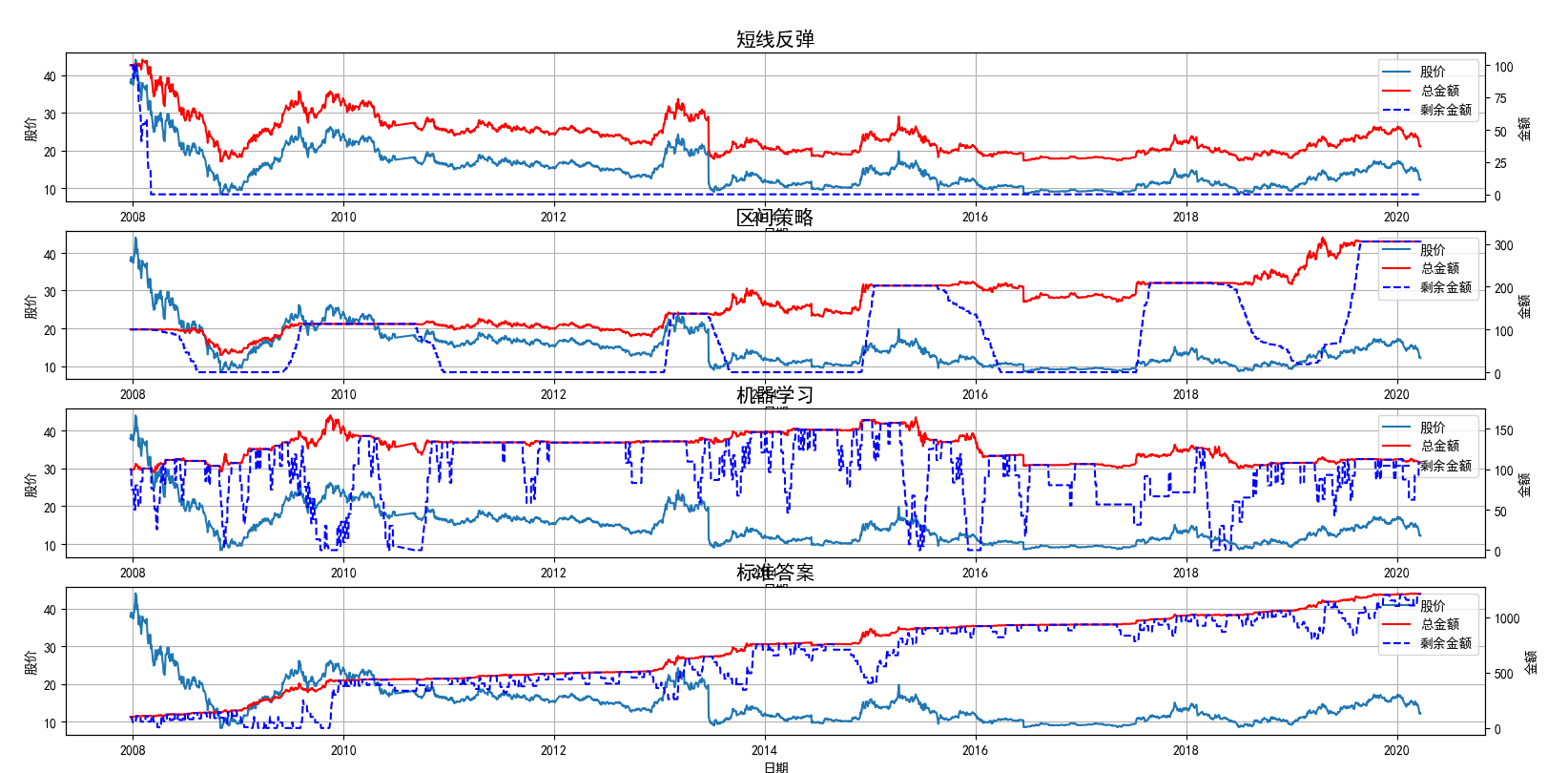
该实验是将预测的实验信号分为买入（1），持有（0），卖出（-1）三个类别。对于不同股票不同时间段而制作的模型，从准确率上看，平均都在65%左右，可以说是有着相当良好的表现。这也证明了利用股票历史数据预测股市走向是有可能实现的。

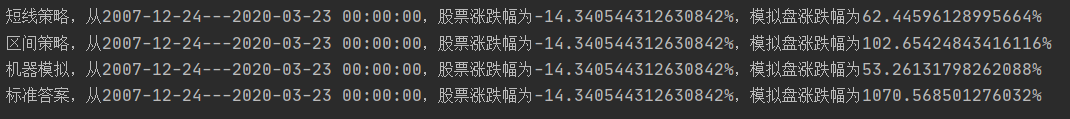


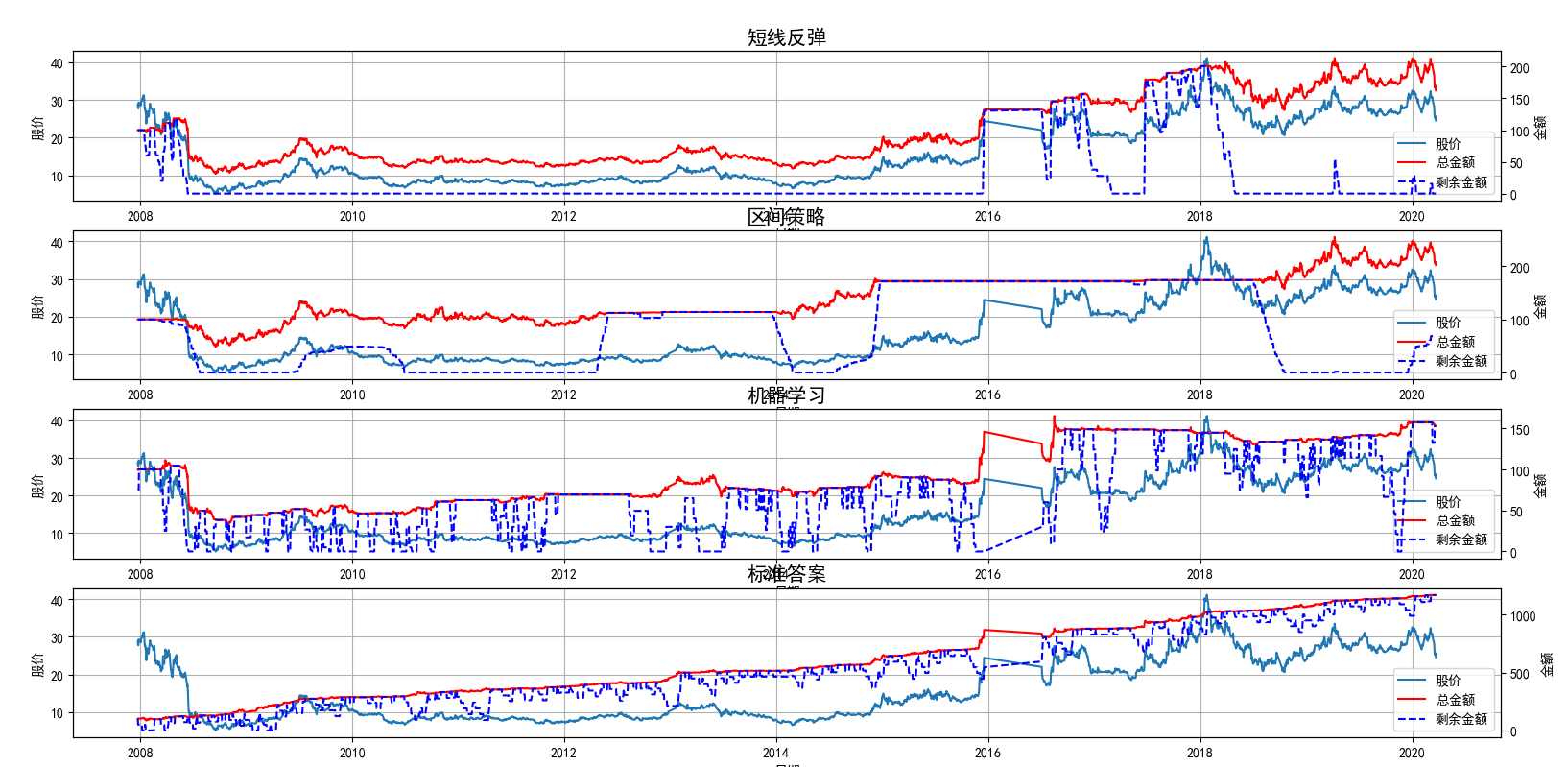
2.量化策略模拟盘成果

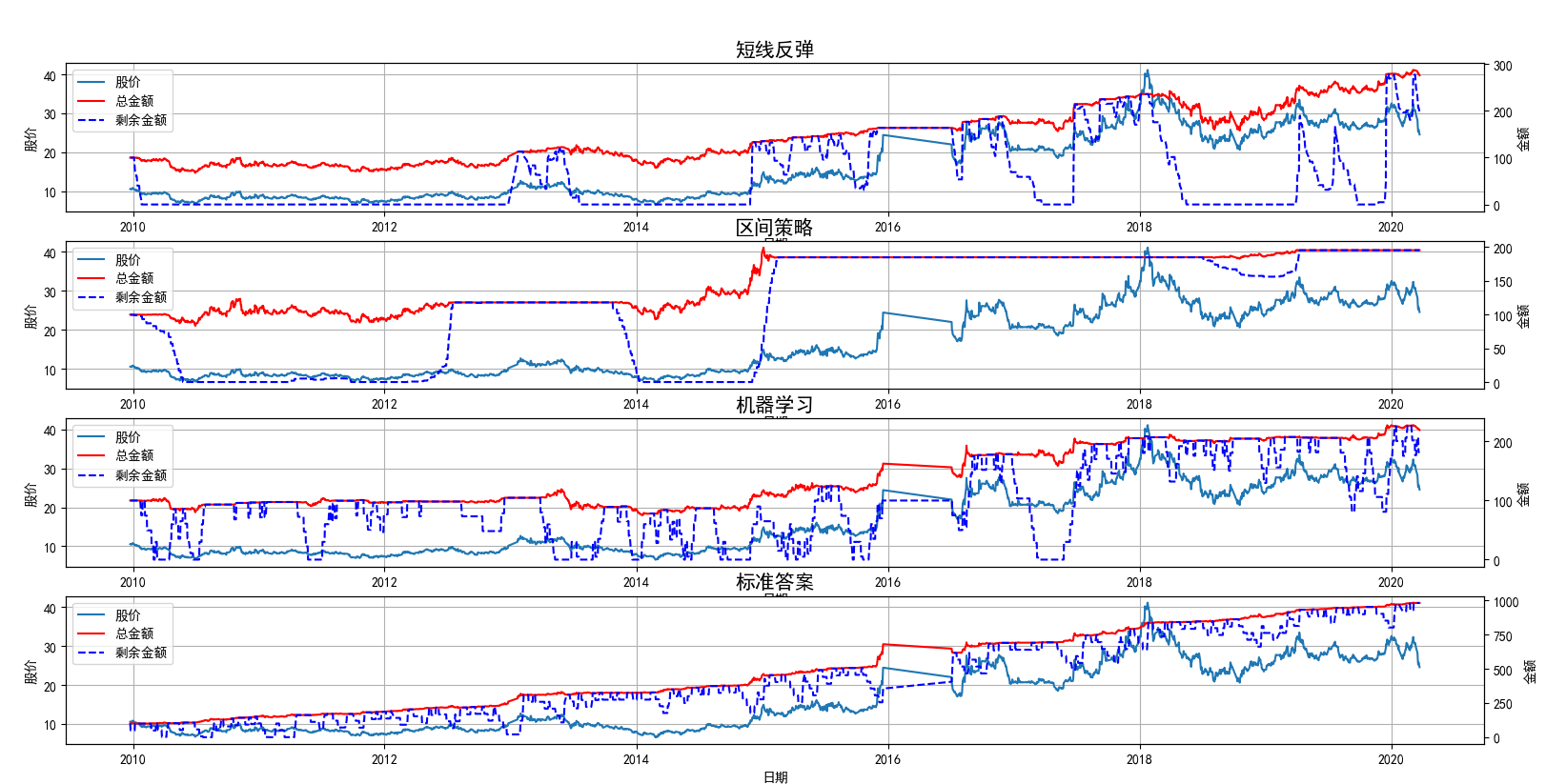
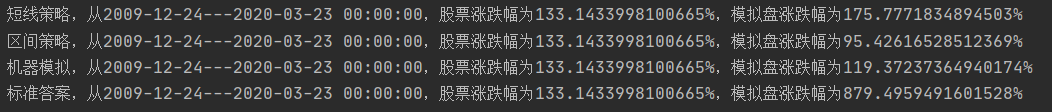
将三个量化策略的模拟盘成果与已知未来涨跌的完美操作成果进行横向比对（下列比对图可能不太清晰，故已保存至当前文件夹下，以供参考）：











在大周期（10年，以减少偶然性对策略评估的影响）的投资中，三种量化策略大都跑赢了股票基准，显示了其有效性。故以上三种策略作为股票交易的参考，可见是有其可取之处的。

下面对每一种策略进行针对性的分析与建议：

短线反弹：

分析：该策略的核心是，当股价达到阶段性低价，进行买入抄底，以期后续回涨，从中套利。该策略，在多数情况下，都有着良好回馈。但一旦遇到股价出现持续暴跌情况时，往往会不断接盘，在股价到达真正底部前就已全仓买入，从而造成不必要的亏损。比如在第一张模拟图中，我们发现该策略在2008年股票暴跌时，贸然抄底，在下跌动能仍充足，股价尚未见底时就已打光子弹，全仓买入了。而其后续也因为这一举动，产生了大量亏损。

建议：策略执行时，结合股价，量能，市场周期与相关资讯进行分析，从而大概估计下跌幅度，以作策略参考。

区间策略：

分析：该策略适用于股价稳定波动的股票。以历史数据划分价格区间，可以有效实

现高抛低吸，从而套利。但过于关注股票价位，而忽视对波动趋势分析，将可能导致错过牛市红利，在熊市阴跌中贸然接盘。比如第三张模拟图中，在2015年时，该策略就开始了持续空仓，从而错过了15-18年股价暴涨接近150%的红利，这就是不注重趋势分析所导致的。

建议：通过k线图与当前金融大背景的分析，了解股市趋势，对股市走向有大局观上的了解，不拘泥于股票价位一隅。

模型预测：

分析：通过机器学习观测量能，股价与一些趋势指标对股市波动影响，形成模型，并借此预测7日波动，发出交易指令。但不难发现，尽管模型测试时的准确率较高，但该策略与我们期望得到的结果，“标准答案”模拟图的盈利情况相比，却是相去甚远。收益情况明显与模型的高准确率不契合。

究因：模型生成时，通过train\_test\_split()随机拆分成了训练集与测试集。两个集合是在样本集内部随机交叉排列的。而我们又知道，一个时间段内的股价波动是相似的。这就意味着测试集所使用的信息不仅仅是自身所带信息，模型提取规律，还有其前后相邻时间点的股价波动情况。而我们上模拟盘实操时，是没有这一信息的。这也解释了，为什么模型准确率高，但模拟盘情况却不尽人意。后来，我将模拟盘实操时，模型预测的交易指令集与标准答案进行比对，发现准确率骤降至40%-45%。可见，该模型有一定参考价值，但并不建议作为一个核心交易策略使用。

五，项目总结

本项目基于股票历史数据，对两个经典量化策略，与一个以机器学习为核心的模型预测策略进行了检验与分析。首先，从股城网爬取了股票代码，然后通过网易财经的接口获得股票历史数据；其次，将历史数据针对不同策略需要，进行了个性化处理；接着，构建模拟盘，将三种策略放入模拟盘中进行演练，比对总结各自优劣，并对此提出完善建议。最后，本项目对实验结果进行总结，发现短线抢反弹，与区间策略，虽在大多模拟盘中的表现不错，但当股票市场出现大波动时，抗风险能力与盈利能力是不足的。建议在实盘中交易时，应结合更多元的指标与金融形势进行分析。而模型预测，就目前该实验的模型而言，尚有较大进步空间，但也初步证实了机器学习对辅助股市决策是有一定的可行性。

参考文献：