|  |  |  |
| --- | --- | --- |
| **Problem Chosen** C | **2021 MCM/ICM Summary Sheet** | **Team Control Number** 2226434 |

Title (Add your paper title)

Summary

Market traders buy and sell volatile assets frequently, with a goal to maximize their total return. volatile assets的价格数据就是一个典型的时间序列数据，由于现实中交易者常常交易多种资产进行投资组合，一般的一元时间序列模型不够起作用，以本题为例，交易者要同时投资黄金和比特币。为此我们开发了一个模型，仅根据当天的黄金和比特币价格即可提供最佳的每日交易策略。

我们的模型基于黄金和比特币从2016-2021的五年交易数据，对原始数据进行滑动窗口处理，使其更适合在机器学习中被运用。然后，我们选用了随机森林回归模型来完成时间序列预测问题，并达到了很好的预测结果。进一步地，根据预测出的第二日黄金和比特币价格来求解最佳的当日交易策略，这是一个典型的目标规划问题，且其目标函数就是要求最大化总回报，这里我们采用遗传算法来求解这一问题，从而给出每日最佳交易策略。依照我们的策略，在5年后最初的1000美元投资价值约270000美元。最后我们给关于交易成本对策略结果的影响进行了一些灵敏度分析。

Market traders buy and sell volatile assets frequently, with a goal to maximize their total return. The price data of volatile assets is a typical time series data. In reality, traders often trade a variety of assets for portfolio, so the general unitary time series model cannot play a great role. In this case, traders want to invest in gold and bitcoin at the same time To do this, we developed a model that provides the best daily trading strategy based solely on the day's gold and Bitcoin prices.

Our model is based on the five-year transaction data of gold and bitcoin from 2016 to 2021. The sliding window processing of the original data makes it more suitable for application in machine learning. Then, we choose the random forest regression model to complete the time series prediction, and achieve good prediction results Furthermore, the optimal daily trading strategy is solved according to the predicted price of gold and bitcoin on the second day. This is a typical goal programming problem, and its objective function is to maximize the total return. Here, we use genetic algorithm to solve this problem, so as to give the optimal daily trading strategy According to our strategy, the initial $1000 investment is worth about $270,000 after 5 years. Finally, we do some sensitivity analysis on the effect of transaction costs on the outcome of the strategy.

**Keywords:** keyword1; keyword2; keyword3; (list three to ten pertinent keywords specific to the article) 时间序列预测；随机森林回归；量化交易；遗传算法；

Content

[1 Introductions 2](#_Toc26864723)

[1.1 Background 2](#_Toc26864724)

[1.2 Restatement of Problems 2](#_Toc26864725)

[1.3 Our work 2](#_Toc26864726)

[2 Assumptions 3](#_Toc26864727)

[3 Notations 3](#_Toc26864728)

[4 The Model 3](#_Toc26864729)

[4.1 3](#_Toc26864730)

[4.2 XXXX 4](#_Toc26864731)

[4.3 XXXX 4](#_Toc26864732)

[5 Model Analysis 4](#_Toc26864733)

[5.1 Sensitivity Analysis 4](#_Toc26864734)

[5.1.1 XXXX 4](#_Toc26864735)

[5.1.2 XXXX 4](#_Toc26864736)

[5.2 Strengths and Weaknesses 4](#_Toc26864737)

[6 Conclusions 4](#_Toc26864738)

1. Introductions
   1. Background

货币是衡量一切商品的其他商品, 其中纸质货币是目前世界上最重要的一种货币形式。黄金是存在几个世纪的十分稳定的货币类型；比特币是在08年经济危机后出现的虚拟货币，与黄金不同，它是全天候的，24小时波动，不受涨跌幅限制，也没有第三方监督。两者都是存储量有限的稀缺物资。黄金、白银等贵金属具备内在价值，所以不需要承担信用风险；而比特币更加便于存储，而且不受各种物理条件限制。

计算机技术的发展使我们能够获得前所未有的数据和计算能力来分析黄金与比特币的市场选择，这也将有助于帮助我们更好地了解两者不同的市场特性。

* 1. Restatement of Problem

在这个问题中，我们需要实现的目标如下：

* 开发一个模型，该模型仅根据当天的价格数量提供最佳的每日交易策略，使最后一天所得的资产总值最大。
* 提供证据证明此模型提供了最佳策略。
* 确定策略对交易成本的敏感程度，并得到交易成本对策略和结果的影响能力。

1. Assumptions

We make the following assumptions to complete our model through this paper. Furthermore, improvements of these simplified assumptions will be achieved later.

* 假设数据是准确的，并且没有伪造的数据：这意味着我们的分析符合事实。
* 假设黄金与比特币的市场运行稳定，也没有出现经济危机等严重影响行情的情况。
* 假设市场价格只与前几日价格有直接关联，因为这是我们仅有的信息。

1. Notations

In this work, we use the nomenclature in Table 1 in the model construction. Other none-frequent-used symbols will be introduced once they are used.

Table 1. Parameter Settings

|  |  |
| --- | --- |
| Parameter | Description |
|  | The amount of gold held |
|  | The amount of bitcoin held |
|  | Gold transactions  Bitcoin transactions  Today's gold price  Today's bitcoin price  Tomorrow's gold price  Tomorrow's bitcoin price |

1. Data Processing

我们的模型所实现的功能是每日提供一个比特币和黄金的投资组合的交易策略，所以在建模之前要先对数据进行一步连接的处理，由于比特币每天都开市，而黄金不是每天都开市，因此这里采用比特币表格中的date来进行数据的连接。

连接完之后可以看到在比特币开市而话赶紧没有开市的那些日期上黄金的价格数据就为空。

再进一步查看缺失值，可以看到连接完后的表格中有571个缺失值，这表明了在我们的五年投资期间有571天黄金没有开市。

我们还可以对原始数据进行一个每日收益率的计算，并对其进行可视化，可以看到比特币的每日收益率波动较大，因此我们在比特币上的交易策略需要更加灵活并且是一些短线的操作。

在我们开始使用机器学习算法之前，时间序列数据必须重新构建为监督学习数据集。时间序列中没有输入和输出功能的概念。相反，我们必须选择要预测的变量，并使用时序数据滑窗平移来构建将用于对未来时间步骤进行预测的所有输入。

1. Model Construction
   1. 时间序列预测
      1. 时间序列数据滑窗操作

进行机器学习时，需要把数据划分为训练集和测试集。对于时间序列数据的预测，往往是建立由好几个历史数据预测下一时刻的未来数据，这时候为了充分利用全部数据，应该对原始数据集进行滑窗操作。滑动窗口就是能够根据指定的单位长度来框住时间序列，从而计算框内的统计指标。相当于一个长度指定的滑块正在刻度尺上面滑动，每滑动一个单位即可反馈滑块内的数据。

实验中，采用了时间序列数据的滑动操作，通过第一天的黄金和比特币的价格进行第二天的价格预测，通过第二天的价格进行第三天的价格预测，以此类推。

* + 1. 损失函数定义

均方误差（Mean Square Error）：

#### 平均绝对值误差(Mean Absolute Error):

#### 根均方误差（Root Mean Square Error）:

**R Squared:**

* + 1. 随机森林回归

随机森林回归是在生成众多决策树的过程中，是通过对建模数据集的样本观测和特征变量分别进行随机抽样，每次抽样结果均为一棵树，且每棵树都会生成符合自身属性的规则和判断值，而森林最终集成所有决策树的规则和判断值，实现随机森林算法的回归。其中的输入为自变量 X（X 为 1 个或 1 个以上的定类或定量变量），因变量 Y（Y 为一个定量变量），输出为模型输出的结果值及模型预测效果。

随机森林的算法步骤如下：  
a.从原始样本集中抽取训练集。每轮从原始样本集中使用 Bootstraping 的方法抽取 n 个训练样本（有放回的抽样）。共进行 k 轮抽取，得到 k 个训练集。（k 个训练集之间是相互独立的）  
b.每次使用一个训练集得到一个模型，k 个训练集共得到 k 个模型。  
c.对分类问题：将上步得到的 k 个模型采用投票的方式得到分类结果；对回归问题，计算上述模型的均值作为最后的结果。

本实验中测试集占整体样本的30%，为了使得实验结果能够固定，将random\_state设为5。

* 1. 目标规划

通过时间序列预测我们已经预测出来第二天的黄金和比特币的价格，这时我们就要制定一个最优的投资策略。

第二题是同时投资黄金和比特币的投资组合分析问题，一般针对长线的投资组合会采用Markovitz方法，然而考虑到这里的比特币市场波动很大，更适合频繁的短线交易，并且根据题意模型需要提供最佳的每日交易策略，因此我们决定采用目标规划来解决这个问题。

* + 1. 目标函数（黄金和比特币都开市）

我们目标规划的目的就是最大化金额总量。

此时我们已知的量是当前黄金持有数、当前比特币持有数、今日黄金价格、今日比特币价格、预测得到的明日黄金价格、明日比特币价格，变量是今日的黄金交易数和比特币交易数。

显然，按照预测值可以计算得出明日的持有金额总量Z应为

因此我们的目标函数就是使金额总量Z最大化，约束条件是交易数不得超过当前持有数的绝对值（无论买入还是卖出），且按预测结果交易后获利要不少于交易成本，也就是说黄金收益率要大于等于佣金0.01，且比特币收益率要大于等于佣金0.02，在这种情况下才进行交易才会真正获利。

由此，可以列出交易策略的目标函数：

* + 1. 目标函数（黄金未开市）

由于黄金不是每天都开市，在黄金不开市的日期我们无法对我们持有的黄金进行交易，因此持有的黄金金额按兵不动，而金额总量Z为

目标函数：

* + 1. 遗传算法

上述的目标规划采用遗传算法来实现。

遗传算法是用于解决最优化问题的一种搜索算法，算法的整体思路是建立在达尔文生物进化论“优胜劣汰”规律的基础上。它将生物学中的基因编码、染色体交叉、基因变异以及自然选择等概念引入最优化问题的求解过程中，通过不断的“种群进化”，最终得到问题的最优解。

假设函数值所对应的所有解是一个容量超级大的种群，而种群中的个体就是一个个解，接下去遗传算法的工作就是让这个种群中的部分个体去不断繁衍，在繁衍的过程中一方面会发生染色体交叉而产生新的个体。另一方面，基因变异也会有概率会发生并产生新的个体。接下去，只需要通过自然选择的方式，淘汰质量差的个体，保留质量好的个体，并且让这个繁衍的过程持续下去，那么最后就有可能进化出最优或者较优的个体。

按照5.2.1和5.2.1的两种情况分别设置目标函数，并假设初始的1000美元资金被平均地分配去投资了500黄金和500比特币。

遗传算法的一些参数设置：

定义变异概率为0.01，种群规模为100，迭代次数为500，采用默认交叉概率。

遗传算法的具体步骤如下：

Step 1 计算适应性分数：本题中适应性分数可以直接采用金额总量来进行评价，总金额数越大则该成员适应性越强。

Step 2 Selection: 轮盘赌选择法(Roulette wheel selection)。

Step 3 Crossover

Step 4 Mutation

Step 5 新一代种群产生，返回step 1.

按以上步骤不断地进行迭代，将最终每日遗传算法收敛至的结果保存并导入至excel表格中。

* 1. 运行结果

连接表格，通过计算可得比特币和黄金的日收益率。表一展示了比特币和黄金的日收益率的前五行，图一以图片的形式展示了比特币和黄金的日收益率，可以发现比特币的收益率波动幅度较大。

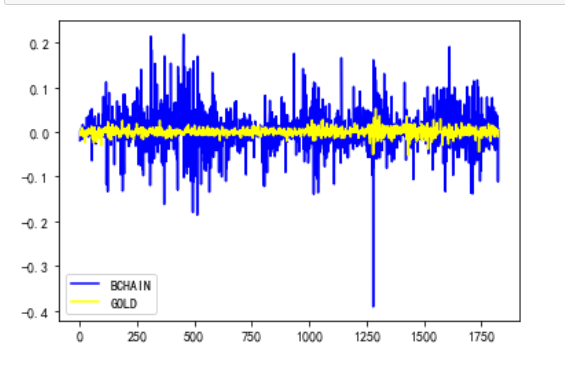
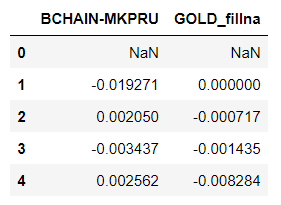


Table1：比特币和黄金的日收益率 Figure1:比特币和黄金的日收益率

对黄金的时间序列数据进行滑窗平移，再将数据集拆分为训练集和测试集，并完成随机森林回归，通过计算可以得到MSE、RMSE、MAE、的值，如图2所示。将计算结果和黄金当日收益进行对比，可以得到如图3所展示的数据。经计算，的值约为0.9960，拟合度较高。

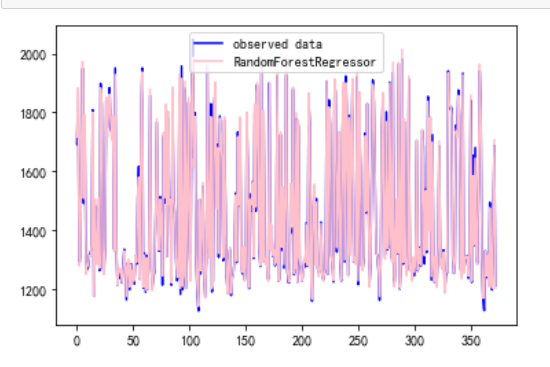
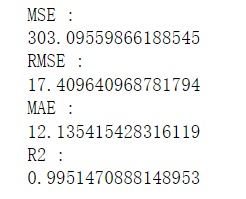


Figure3:模拟得到的黄金数据与真实数据对比

Figure2:对黄金的数据做随机森林回归得到的MSE、RMSE、MAE、的值

同理，对比特币的时间序列数据进行同样的操作，得到MSE、RMSE、MAE、的值，如图4所示。将计算结果和比特币当日收益进行对比，可以得到如图5所展示的数据。经计算，的值约为0.9948，拟合度较高。

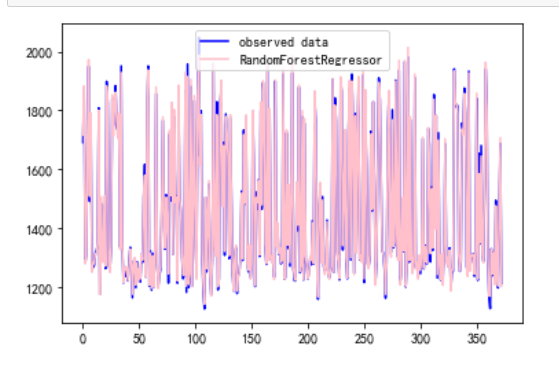
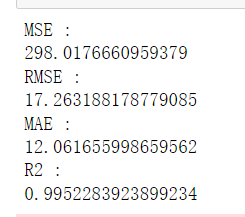


Figure5:模拟得到的比特币数据与真实数据对比

Figure4:对比特币的数据做随机森林回归得到的MSE、RMSE、MAE、的值

整合后的黄金与比特币的预测数据与真实数据如表2所示。根据上述比较可知，模型拟合度较高，提供了最佳策略。

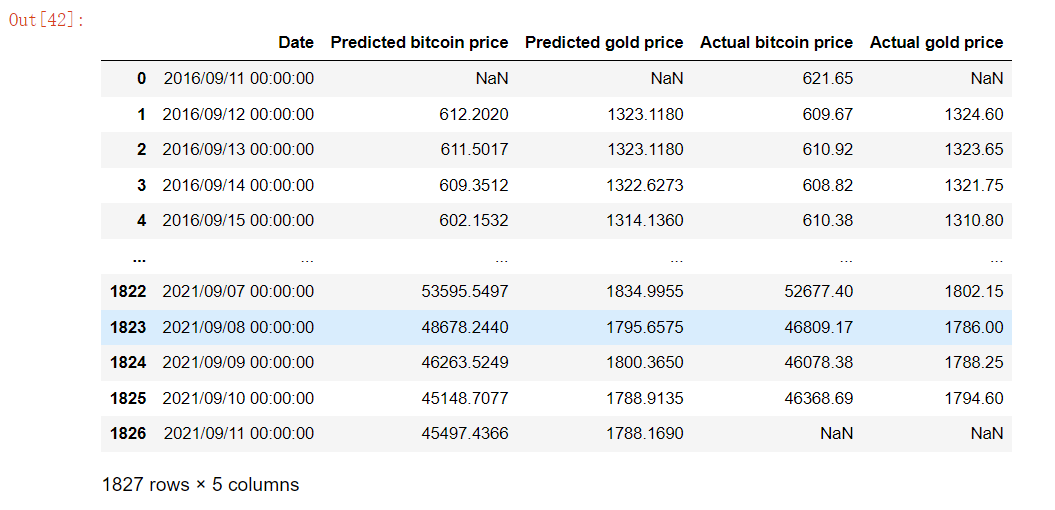
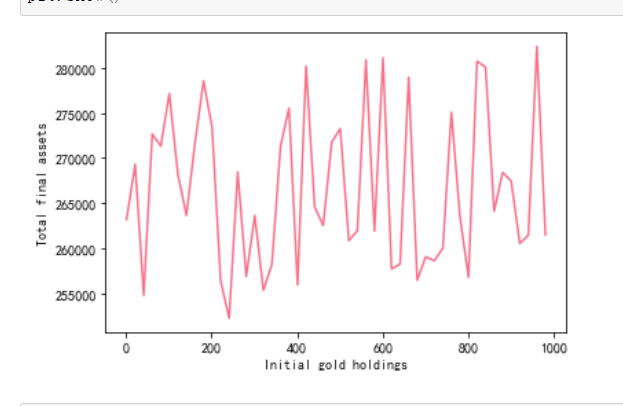


Table2：比特币和黄金的预测数据与真实数据

1. Model Analysis
   1. Sensitivity Analysis
      1. 交易初始值对投资结果的影响

在上述模型中，我们假设1000美元在一开始被平均分配给黄金和比特币，在该假设下进行投资策略的制定以及总资产的计算。然而，初始值不同一定会导致最终资产总额的不同，为了研究交易初始值对投资结果的影响，我们计算了从1-1000美元不同的黄金投资初始值，剩下的用于投资比特币的各种情况下所达到的最终资产总额。

可以看到资产总额总是在255000至280000之间，平稳度较高，交易初始值对投资结果的影响不大。



* + 1. 交易佣金对投资结果的影响

我们尝试增加了佣金，发现随着佣金的增加交易者的交易额有所下降，其投资结果也有一定下降。

* 1. Strengths and Weaknesses

优点：

* 我们使用随机森林算法构建模型，本身精度高于大多数单个算法，准确性高；由于随机性的引入，使其不易陷入过拟合；算法对数据集的适应情况良好，且很好地处理了缺失的值（黄金关闭交易的时候等）
* 我们采用时间序列预测，使两份数据能够完美地融合，并制造数据的时序滑动，使得回归之作变得较为容易。
* 我们对数据进行了可视化工作，以便于理解。
* 我们使用遗传算法，其通过变异机制避免算法陷入局部最优，搜索能力强；它引入了自然选择中的概率思想，个体的选择具有随机性。

缺点：

* 我们本可以对模型进行进一步的优化完善，但由于篇幅的限制，我们的工作暂时只能做到这。
* 由于缺乏实际的数值参考，我们只能从实际情况中观察到模型的准确性，而不能通过数值方法来测量模型的准确性。
* 遗传算法内包含的交叉率、变异率等参数的设定需要依靠经验确定，且其对于初始种群的优劣依赖性较强

1. Conclusions

Add your text here.

References

1. Zhou Zhihua. Machine learning [M]. First edition. Beijing: Tsinghua University Press, January 2016.
2. XXXX