1. 讲述一个你曾经做过的机器学习或者数据分析相关的项目。

我的硕士毕业论文是研究因果森林模型 (Causal Forest) 以及其与双重机器学习法，工具变量法结合之后的表现。获得帝国理工大学最佳毕业论文奖。链接：
https://github.com/PhillipLi-Hub/MasterDissertation/blob/main/Zhanghao_Li_Dissertation.pdf

传统的机器学习分类/回归模型往往注重预测精度而忽视了可解释性，比如随机森林/神经网络是“黑盒”模型。而在注重模型可解释性的计量经济学中，OLS 是经典方法，但其有缺陷，比如线性假设不成立，多重共线性等问题。因果森林可以视为随机森林的因果推断版本，学习了决策树的树结构和随机森林的 bagging、随机特征选取方法，解决了传统机器学习的可解释性不足以及传统计量经济学的假设难以完全成立问题。

我的研究目的是 1、探究因果森林在估计异质性处理效应（Heterogeneity）中的效果并将其与双重机器学习法、因果变量法结合来看估计表现 2、将其应用在现实数据集之中，观察和传统方法（线性回归，传统二次最小二乘以及逻辑回归法）相比的估计精度。

我所用的主要方法是使用 R 语言生成已知真实分布的具有因果关系的随机数据集，然后比较因果森林和传统方法（KNN 以及其他机器学习在因果推断的应用如 meta-learner 法）对于处理效应的估计精度。现实数据集我选取了 GSS（General Social Survey）和 OHIE（Oregon Health Insurance Experiment）数据集比较了因果森林和传统的线性回归/逻辑回归的估计精度并得出因果森林精确度显著高于传统计量经济学方法的结论。

2. 两个策略评估以及投资问题

1）两个策略净值走势以及评估指标

图 1：两个策略净值走势

我选取了年化收益，年化波动，最大回撤，夏普，克莱玛比率，索提诺比率来评估两个策略，无风险利率设为 2%，metric 如下表所示。

	Annual Return	Annual Volatility	Max Drawdown	Sharpe Ratio	Calmar Ratio	Sortino Ratio
策略 1	58.35%	15.96%	-5.27%	3.53	11.08	5.00
策略 2	14.01%	0.88%	-0.06%	13.60	239.80	41.58

表 1：两个策略各指标

2）阐述 metric 比较重要以及原因，两个策略比较

主要取决于投资者个人的风险偏好，对于风险偏好较高的投资者来说，年化收益更加重要，而对风险偏好低的投资者来说，波动率更重要。对于折中的投资者来说夏普/克莱玛/索提诺比率比较重要，因为同时考虑了风险和收益。我会倾向于综合风险与收益，主要以年化收益率和夏普作为衡量指标。

如图 1 和表 1 所示，策略 1 的收益更高，同时波动率也更高。策略 2 更加平滑，但收益较低。在夏普/克莱玛/索提诺比率上策略 2 显著好于策略 1。

3）如果有 100 万会如何投资这两个策略

可以将两个策略看成两个资产，我使用风险平价/等权分配/均值方差三个模型确定两个资产的权重。对于风险平价模型和均值方差模型，超参数为计算方差协方差矩阵和收益率向量的窗口期，可以使用滚动验证法来选取最优超参数，即对于每个窗口期 T,使用 T-1 个窗口期计算相关权重构建窗口期 T 的投资组合，选取最优的投资组合。对于第一个窗口期由于缺少之前的数据，使用等权分配。

所有组合的结果如下表所示：

Model	Window Size	Annual Return	Annual Volatility	Max Drawdown	Sharpe Ratio	Calmar Ratio	Sortino Ratio
Risk Parity	5	26.28%	11.00%	-4.51%	2.21	5.83	2.22
Mean-Variance	5	35.20%	8.02%	-2.45%	4.14	14.37	5.94
Risk Parity	10	27.66%	10.07%	-4.51%	2.55	6.13	2.14
Mean-Variance	10	35.20%	8.02%	-2.45%	4.14	14.37	5.94
Risk Parity	15	53.57%	10.79%	-4.51%	4.78	11.88	4.53
Mean-Variance	15	36.75%	8.05%	-2.45%	4.32	15.01	6.07
Risk Parity	20	54.97%	10.15%	-3.65%	5.22	15.07	5.93
Mean-Variance	20	36.75%	8.05%	-2.45%	4.32	15.01	6.07
Risk Parity	25	45.03%	9.54%	-3.65%	4.51	12.35	4.60
Mean-Variance	25	35.20%	8.02%	-2.45%	4.14	14.37	5.94
Risk Parity	30	49.87%	8.91%	-2.95%	5.37	16.92	6.05
Mean-Variance	30	36.75%	8.05%	-2.45%	4.32	15.01	6.07
Risk Parity	35	16.59%	2.07%	-0.92%	7.04	18.12	7.00
Mean-Variance	35	34.95%	8.01%	-2.45%	4.12	14.27	5.84
Risk Parity	40	26.92%	7.72%	-3.81%	3.23	7.06	3.85
Mean-Variance	40	36.75%	8.05%	-2.45%	4.32	15.01	6.07

Risk Parity	45	16.11%	2.98%	-2.12%	4.74	7.59	3.86
Mean-Variance	45	39.35%	8.18%	-2.45%	4.57	16.07	6.40
Risk Parity	50	17.21%	4.34%	-2.12%	3.51	8.11	3.80
Mean-Variance	50	35.20%	8.02%	-2.45%	4.14	14.37	5.94
Equal Weight (50/50)	N/A	34.37%	8.00%	-2.45%	4.05	14.03	5.82
策略 1	N/A	58.35%	15.96%	-5.27%	3.53	11.08	5.00
策略 2	N/A	14.01%	0.88%	-0.06%	13.60	239.80	41.58

表 2：均值方差/风险平价在各窗口期下表现，以及与基准等权分配和原始策略对比

标红的策略，即窗口期在 20-30 天的风险平价相对而言保留了策略 1 的高收益，同时较好地控制了风险，夏普相比于策略 1 的 3.53 提升至了 5.22/5.37，因此我会选取窗口期为 30 天根据风险平价模型确定权重后进行投资，计算得到分别向策略 1/2 投资 20173 及 979827 元。

20/30 天窗口期的风险平价，等权分配，以及两个基础策略的走势图如下图所示，可以看到风险平价模型主要是通过动态调整权重规避了震荡行情（在动荡期持有策略 2）。

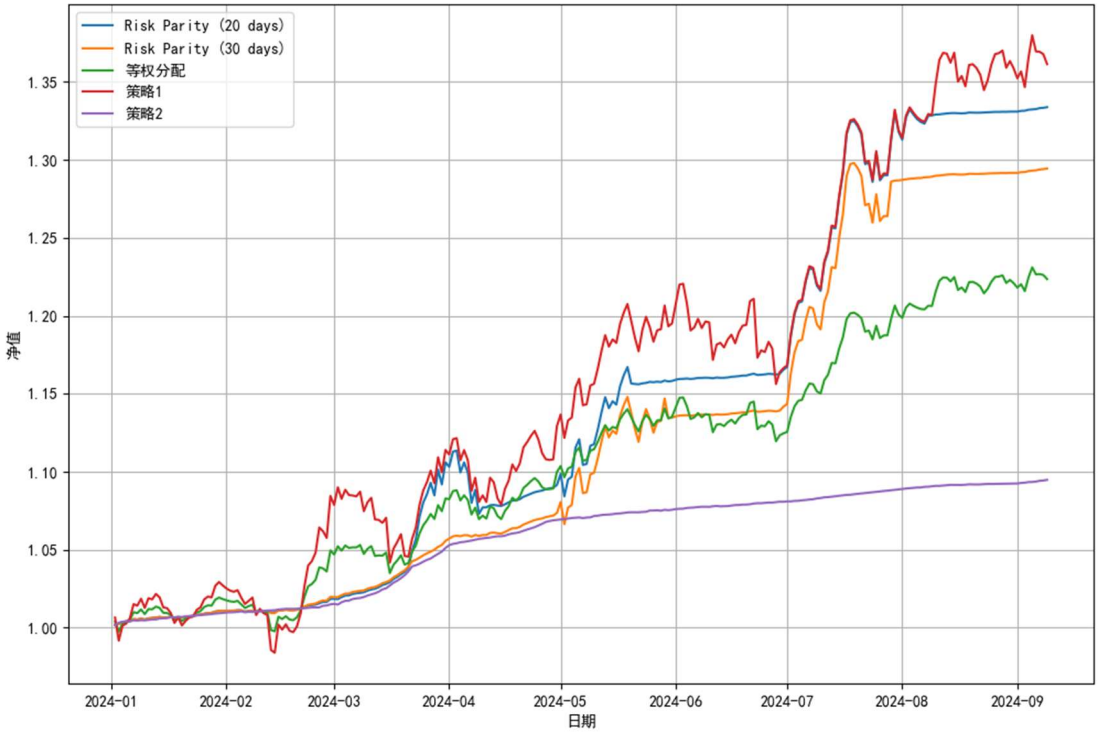

图 2：不同策略的净值走势