第四组：高频交易预测

徐宗杰：数据挖掘

陈梦玄：特征工程

邓杰：模型构建

1. 赛题理解：

沪深 300 股指期货合约，由于每天的数据是打乱的，考虑到时间序列未来数据的问题，并且业内期货 CTA 策略一般使用日内数据，所以使用日内数据建模，日间模型进行 boosting 处理。四个 y 指的是四个期货的交易方向：卖出开仓、卖出平仓、买入开仓、买入平仓。经过数据验证，y1y3 和 y2y4 所有缺失值在同一时刻出现，且所有时间点的和为零，验证了我们的想法说明 13 和 24 是对手方交易。

本次项目我们希望将科技带入金融，有效结合金融和科技进行建模预测，在模型效果好的基础上，同样满足实际业务。

2. 数据挖掘：

2.1 数据融合

本次数据的业务背景是高频交易，在同一时刻记录的交易记录，会导致在同一时间有多条交易记录，但最终的最优交易 action 只会有一个。这就产生了特征数据多条对应单条 label 的情况。其中的处理方式有两种：1.充分利用多条数据进行训练，用 label 去匹配每一个时刻的特征数据。2.根据某些重要的指标（例如成交量）来给同一时刻的交易记录赋予权重，进而数据融合。

2.2 缺失处理

本次数据中特征数据和 label 数据并不能完整的匹配上，并且其中存在些许缺失值；在实

际业务中切实会存在返回指标值为空的情况。在缺失值的处理上，指标或 label 全为空表示对照组 action0 不作操作。对于 label 的缺失，直接填补 0 值。若是 label 部分缺失，会选到其他非负的 action 上；若是 label 全体缺失，任意选项的 action 或者不操作都是满足业务要求的。

对于特征数据的缺失情况，全体缺失表明无操作数据，其业务中对应这 action0，选以 0 值填补；部分缺失，由于本次模型主要选择树模型，以 0 值填补不影响整体模型的训练。因为不希望破坏日内数据的连续性，没有选择删减缺失值的数据，在满足业务场景的情况下去进行填补。

2.3 数据探索

本次的 5 种 action（[0,1,2,3,4]），1 和 3 分布相同互为相反数，2 和 4 亦然，且 1-4action 求和为零；对应业务上也正是四个期货的交易方向。0 作为误操作的对照组保证了本次 return 非负。根据这个有趣的现象可以构造简单的二分类规则模型，来达到 return 收益不减的目的。

每次 action 根据两种对手方交易，我们只需要预测 1 和 2 的 return 值的正负形（亦或是 3 和 4），若 return 为正值，则选取 action1 或者 2 来保证 return 收益不减；若 return 为负值则选取对手方 action3 和 4。对于具体选取 action1 还是 action2,取决于预测的效果。选取预测效果最好的 action 来保证 return 收益不减。

在具体建模过程中，action1 的准确性较高，二分类准确性如下图所示：

```
xgb_auc_score: 0.6094425664652248
xgb_pre_score: 0.6879770485416261
```

3.特征工程：

3.1 期货交易金融业务含义上的特征构造

原始数据给出的是高频交易的 tick 级十档买卖盘口数据，可以根据这些数据构造相关的高频交易指标进行处理。从而构建了业务相关的商业性特征和数据本身的统计量特征，共 214 个。

- 盘口数据及其长短趋势项和波动率项，以 5 为起始，100 为终止，25 为步长针对卖一二三买一二三档价量数据构造移动平均指标
- 最大回撤：日内历史数据最高价格到最低价格时间的价格差异（所有的成交价、买卖价的变时长滑动窗口）
- 最大亏损：日内历史数据初始价格到最低价格之间的距离（所有的成交价、买卖价的变时长滑动窗口）
- 最大收益：初始价格到最高价格之间的距离，最高点和开盘价的差异（所有的成交价、买卖价的变时长滑动窗口）
- 当前点和前高点差异
- 买卖价差（bid-ask spread）
- 深度比：(买一价+买二价+买三价)/(卖一价+卖二价+卖三价)(ref：https://mp.weixin.qq.com/s/nTMjumvIJlFJ97yiwqQt3Q)
- 加权的深度比：(买一价*买一量+买二价*买二量+买三价*买三量)/(卖一价*卖一量+卖二价*卖二量+卖三价*卖三量)
- 上升比 rise_ask:买一价与当日最高的买一价之间的比值（ref：https://mp.weixin.qq.com/s/nTMjumvIJlFJ97yiwqQt3Q）
- 当天开盘价、当天开盘成交量
- 内盘/主动性卖盘，即成交价在买入挂单价的累积成交量;也就是说，期货交易在买入价成交，成交价为申买价，说明抛盘比较踊跃。当卖方主动成交时（买价成交），t1 最新价（价位）=t0 某一买价，这一买价为大于卖单限价的最低价，视为卖方主动成交；
- 外盘/主动性买盘，即成交价在卖出挂单价的累积成交量;也就是说，期货交易在卖出价成交，成交价为申卖价，说明买盘比较踊跃。当买方主动成交时（卖价成交），t1 最新价（价位）=t0 某一卖价，这一卖价为小于买单限价的最高价，视为买方主动成交。
- 交易意愿指标及中间价指标：根据交易订单的不均衡性（order imbalance）构造 bid 与 offer 的 volume 变动量之差衡量交易意愿强弱，以及标准化的交易意愿指标 rho，交易价与中间价差指标 mid，tp

$$OI_t = \delta V_t^B - \delta V_t^A, \delta V_t^B = \begin{cases} 0, & P_t^B < P_{t-1}^B \\ V_t^B - V_{t-1}^B, & P_t^B = P_{t-1}^B \\ V_t^B, & P_t^B > P_{t-1}^B \end{cases}, \delta V_t^A = \begin{cases} 0, & P_t^A > P_{t-1}^A \\ V_t^A - V_{t-1}^A, & P_t^A = P_{t-1}^A \\ V_t^A, & P_t^A < P_{t-1}^A \end{cases}$$

$$\rho_t = \frac{V_t^B - V_t^A}{V_t^B + V_t^A}$$

$$TP_t = \begin{cases} \frac{1}{300}\frac{T_t - T_{t-1}}{V_t - V_{t-1}}, & V_t > V_{t-1} \\ TP_{t-1}, & V_t = V_{t-1} \end{cases}, MPB = TP_t - \frac{M_t + M_{t-1}}{2}, M_t = \frac{P_t^B + P_t^A}{2}$$

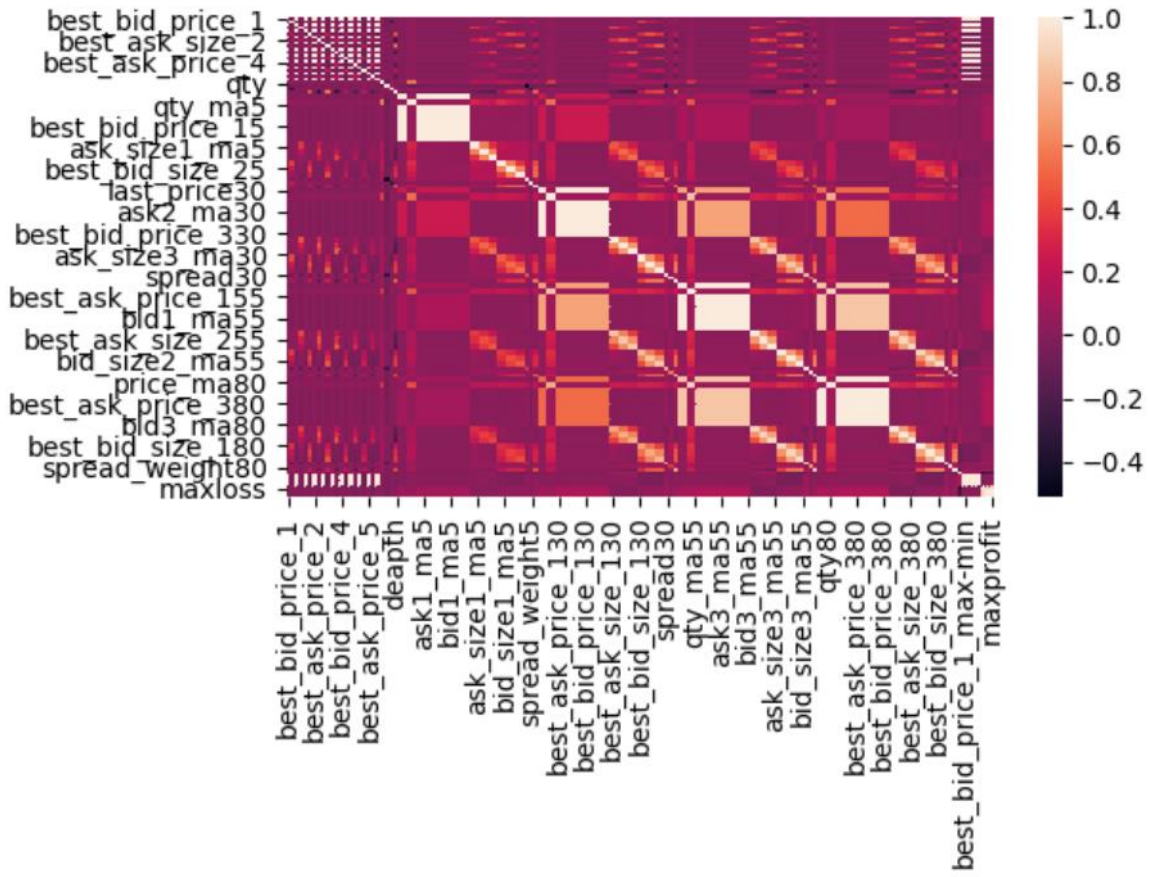
（ref：https://zhuanlan.zhihu.com/p/296523150）

3.2 统计意义以及机器学习等时间序列特征构造

我们使用五折样本数据的统计指标构造特征，还尝试使用 tsfresh（一个时间序列特征构造工具）+sklearn 的多项式特征构造+feature_tools（一个基于深度学习的特征构造工具）来扩展特征。构建了一共 400 多个特征。由于金融业务场景上的特征效果显著，而自动化特征工程运算效率更低，所以我们最终没有采用这个方案。

3.3 特征的检验

首先，对特征之间的相关性进行检验，其中可以看出构建的特征之间具有较为明显的聚类特定，也符合商业逻辑；也是后续特征重要性判断和特征降维的前提。



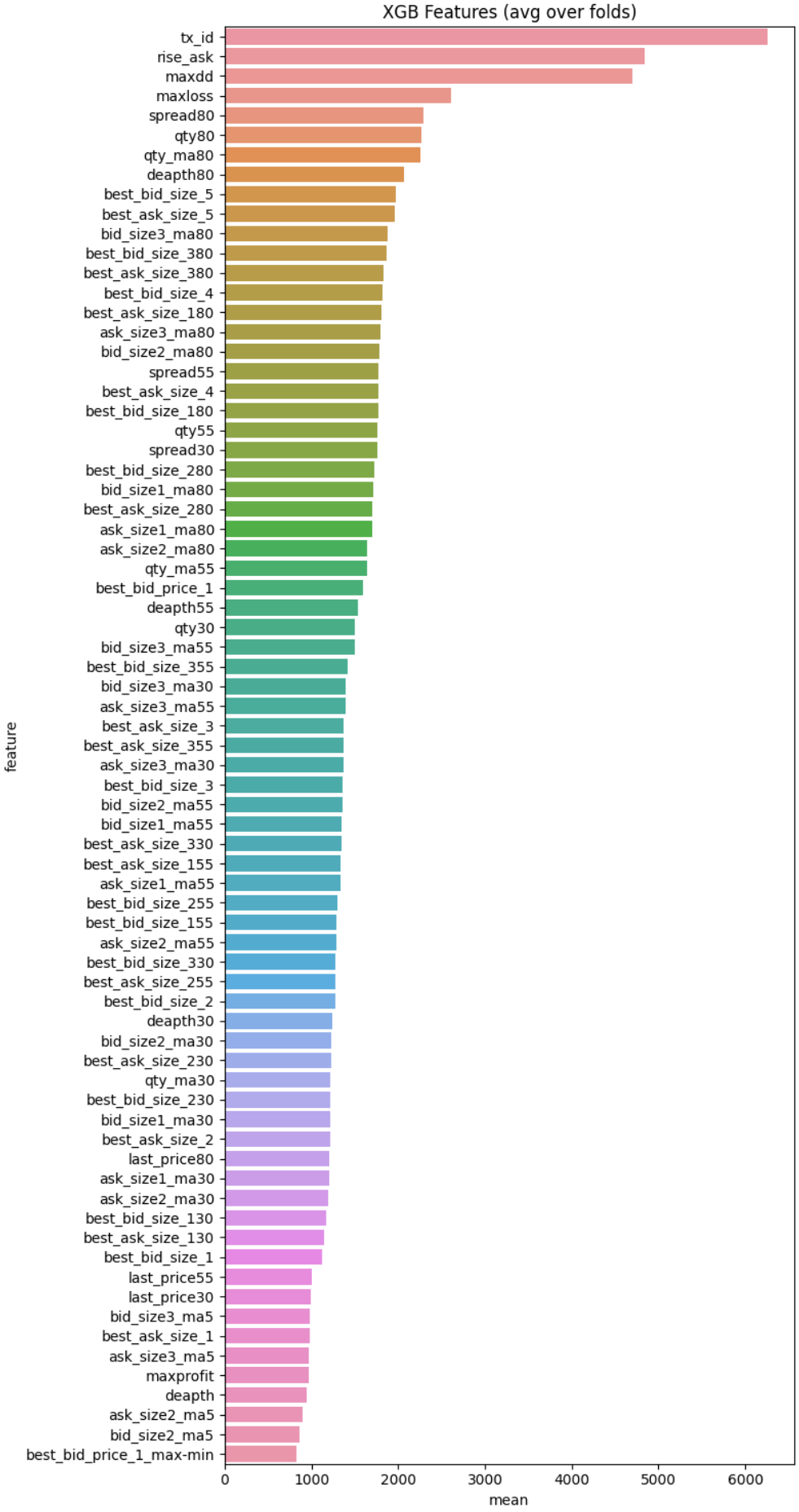
特征的缺失值检验：除了 maxprofit、maxloss 和 maxdd 其他均没有缺失值。缺失值占比 11.24%，由于本项目主要使用树模型进行回归预测，树模型对缺失值优秀的处理能力使得我们可以保持些许特征缺失的原样。

maxprofit	98890
maxloss	98890
maxdd	98890
best_bid_size_1	0
best_bid_size_25	0
last_price30	0
deapth5	0
spread_weight5	0

3.4 根据特征重要性进行特征筛选

上述根据统计量和商业背景所构建的特征，具有较为明显的聚类现象，我们希望通过适当可解释性的降维或是特征筛选来得到真正重要的特征。我们根据构建 LGB 和 XGB 模型对特征重要性进行筛选，交叉验证 feature_importance 进行平均，得到重要性的描述性统计分析。重要性的均值为 963，一共有 214 个特征，重要性大于 800 的特征一共 73 个，选择这 73 个特征进行后续的建模。

```
count        214.000000
mean         715.703271
std          837.495907
min            5.250000
25%          124.250000
50%          385.875000
75%         1220.750000
max         6251.250000
Name: mean, dtype: float64
```


XGB Features (avg over folds)
tx_id
rise_ask
maxdd
maxloss
spread80
qty80
qty_ma80
deapth80
best_bid_size_5
best_ask_size_5
bid_size3_ma80
best_bid_size_380
best_ask_size_380
best_bid_size_4
best_ask_size_180
ask_size3_ma80
bid_size2_ma80
spread55
best_ask_size_4
best_bid_size_180
qty55
spread30
best_bid_size_280
bid_size1_ma80
best_ask_size_280
ask_size1_ma80
ask_size2_ma80
qty_ma55
best_bid_price_1
deapth55
qty30
bid_size3_ma55
best_bid_size_355
bid_size3_ma30
ask_size3_ma55
best_ask_size_3
best_ask_size_355
ask_size3_ma30
best_bid_size_3
bid_size2_ma55
bid_size1_ma55
best_ask_size_330
best_ask_size_155
ask_size1_ma55
best_bid_size_255
best_bid_size_155
ask_size2_ma55
best_bid_size_330
best_ask_size_255
best_bid_size_2
deapth30
bid_size2_ma30
best_ask_size_230
qty_ma30
best_bid_size_230
bid_size1_ma30
best_ask_size_2
last_price80
ask_size1_ma30
ask_size2_ma30
best_bid_size_130
best_ask_size_130
best_bid_size_1
last_price55
last_price30
bid_size3_ma5
best_ask_size_1
ask_size3_ma5
maxprofit
deapth
ask_size2_ma5
bid_size2_ma5
best_bid_price_1_max-min
feature
0
1000
2000
3000
4000
5000
6000
mean

4.模型构建：

本次项目的主要目标是通过预测 action 来获得最大收益。除去上述过的保证 return 不减的二分类模型之外，由于本次高频交易数据属于典型时空高维度预测问题，同时本组具备一定的金融特征构造知识，因此主要考虑使用集成树模型进行了多分类和回归拟合。

4.1 分类模型

（1）基本思想：

基于 action 的选择，最初的模型构建想法考虑对 action 进行分类，即根据每个时刻收益最高的 action 作为训练 label，创建 4 分类模型。

（2）所采取的措施与方法：

- StratifiedKFold---采用分层分组的形式（有点类似分层抽样）,使每个分组中各类别的比例同整体数据中各类别的比例尽可能的相同。由于考虑分类情况下的样本 y4 的不均衡分布情况。
- 贝叶斯优化调参+optuna 自动化调参工具--通过对目标函数形状的学习，并找到使结果向全局最大提升的参数，同时为了快捷使用，考虑 optuna 此类自动化工具的导入。
- 逻辑回归、SVM 其他方法的尝试---较为简单的回归模型并不能全面反映多因素的高维特征；由于基于距离的运算量，我们以天为单位采样，并挑选重要性前十的特征来训练 SVM，由于数据量的限制，SVM 效果依然欠缺。

（3）不足与下一步想法：

经过数据集划分验证线下测试，发现该方法的预测效果并不理想；同时，由于单独的多分类问题本身就具备较大的困难，难以达到合适的收敛，另一方面，结合本题中的“高频交易预测”，若只是简单考虑分类，而忽略 label 本身的收益值。换句话说，并不是越准确的预测动作就一定能得到最高的收益，也不是不准确的预测动作就得不到总的高收益。

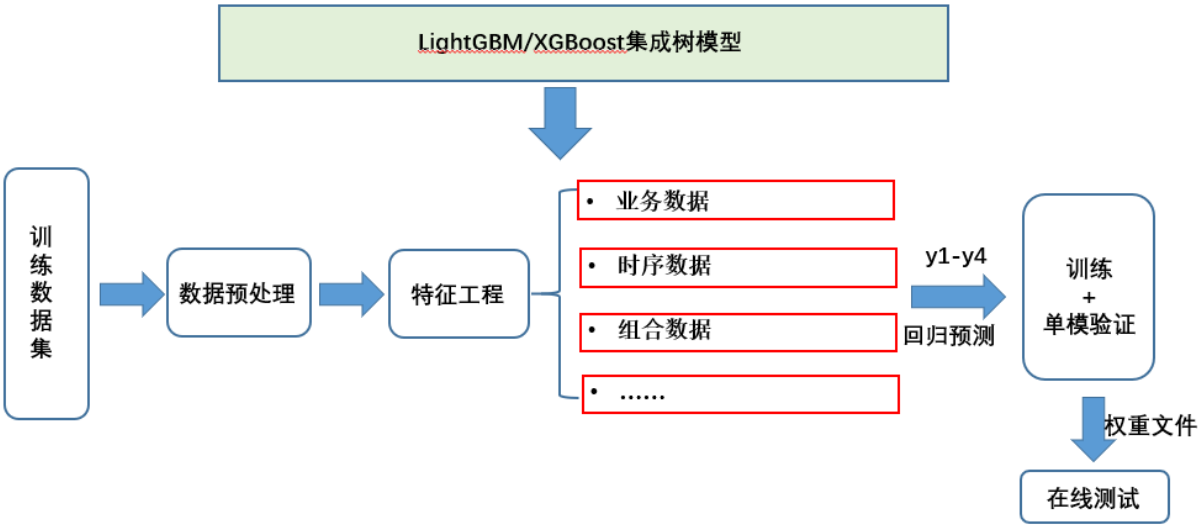
4.2 回归模型

（1）基本思想：

为了分别体现 4 个 action 的 return 值的意义，以及“+”“-”的实际特性体现，我们进一步考虑分别对 y1-y4 四个 action 的 return 值进行集成模型的预测，构建四种数据标签的预测架构，最终分别预测不同 action 的 return 值，从而选取预测出的最大 return 值作为最终的 action 预测值。

（2）所采取的措施与方法：

- KFold---由于 StratifiedKFold 等变体无法支持回归问题，因此考虑回归问题的简单交叉验证 KFold 进行解决。
- 贝叶斯优化调参+optuna 自动化调参工具--同样的，通过对目标函数形状的学习，并找到使结果向全局最大提升的参数，同时为了快捷使用，考虑 optuna 此类自动化工具的导入。
- 特征重要性 Weight+Gain 双重验证--打算结合特征重要性的两种评价 Weight+Gain 进行验证和排序，但是由于本组所构建的特征基本上都是数值类特征，Gain 大多数对于类别特征具有较好效果，因此采用 Weight 进行评价足够。将原本的 400 时间特征、金融业务特征、消费特征等进行排序和筛选，最终得到 73 个关键特征。
- XGBoost+LightGBM 的双思路--分别基于两种常用集成树框架进行分析与使用，考虑 GPU 加速的 XGBoost 训练优势，以及最终的线上测试时间效率因素，最终选取单模单验证的 XGBoost-GPU 模型进行训练和预测，并结合特征筛选进行特征降维。
- 树模型以其对高维特征的筛选效率和庞大数据的运行速度优于其他模型，并且根据滑窗回滚的方式生成的新特征难免会产生缺失值，树模型对缺失值的不敏感也是此项目中脱颖而出的优势。



（3）不足与下一步想法：

由于时间限制和前期的一些错误方向，导致许多方法未进行尝试，包括 Bagging 类型的集成模型、RNN 和 LSTM 相关时序网络。

对于特征来说，可以进一步考虑 CTR 的传统模式方法，即利用某种树模型（GBDT、RF）进行特征自主筛选，得到对应的叶子节点特征，从而实现特征的自动构造和筛选，基于这些编码特征送入分类/回归网络进行进一步预测。