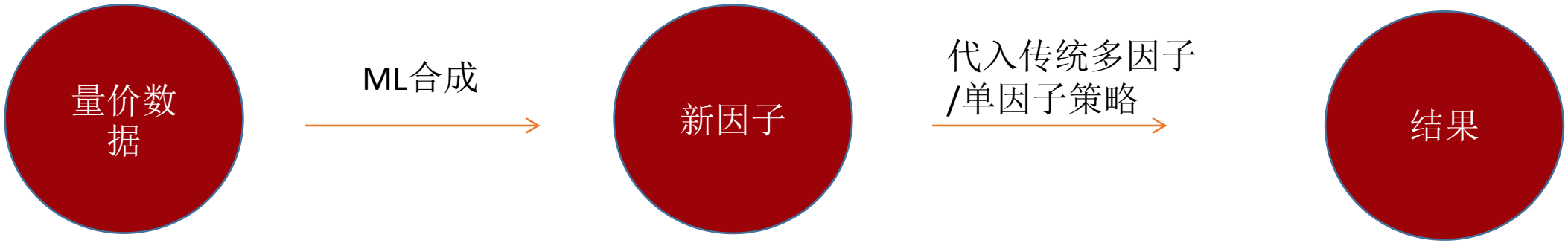
北京大学量化交易协会2020

Is ML an efficient way to construct factors?

QTA ML组

Content

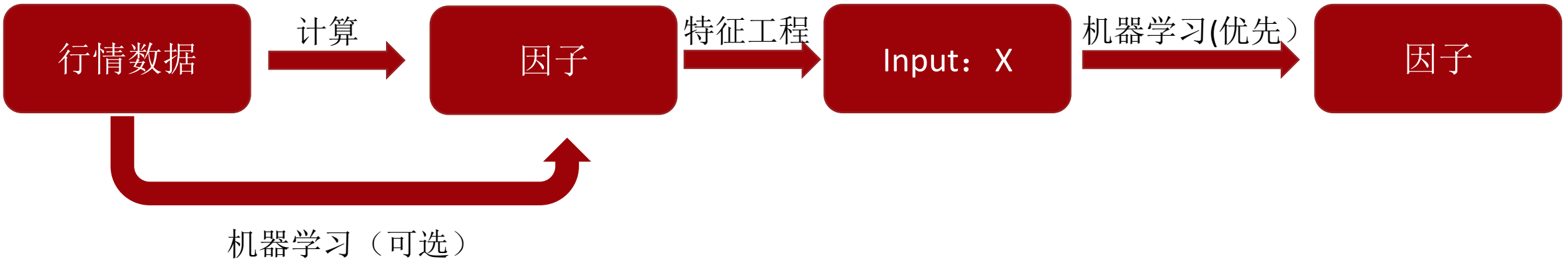
Overview



根据量价数据，使用tree-based或者线性回归等机器学习模型合成新因子，通过传统因子投资的流程回测检验新因子有效性。

Content

因子合成流程



从行情数据计算因子库

从研报或论文获得有经济意义的量价因子（优先）

- **多周期：**月频到日频再到分钟频；Tick级别行情待定，取决于能否取得数据以及计算机资源是否充足。
- **因子不可以太多：**防过拟合。目前发现海通证券的《选股因子系列研究》可以参考。

运用机器学习直接从行情数据得到量价因子（可选）

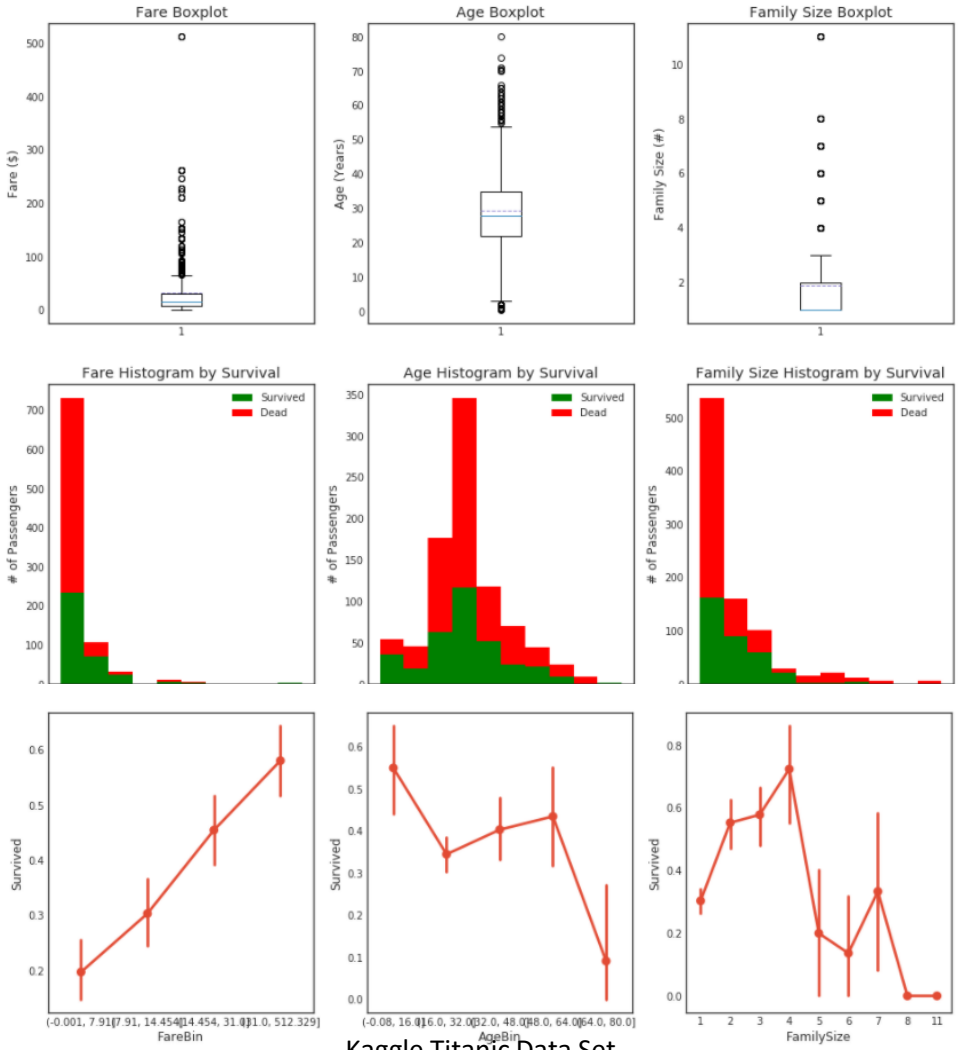
- **只考虑经济意义强的：**防过拟合

特征工程

几乎80% 的工作量会在特征工程步骤。特征工程不是纯粹的数据挖掘，更是了解模型为什么work的途径。【1】

- **预处理**：无量纲化、特征编码。
- **填补缺失值**：根据数据之间的关系得出缺失数据最可能的值。
- **生成新特征**：研究模型输入特征与输出（一般是股票收益率）之间的关系，从而得出更有效的新输入特征。
- **特征选择**：Filter、Wrapper、Embedded
- …

Eg：



Kaggle Titanic Data Set

参考【1】THE 10 REASONS MOST MACHINE LEARNING FUNDS FAIL， Marcos López de Prado

机器学习

预测目标

- **股票下期收益率：**简单粗暴，回归问题。
- **股票下期收益率分组：**即预测t+1期股票qcut的分组序号。是分类问题。
- **因子IC：**预测因子库因子的下期IC，根据所预测的IC决定当期用哪一个因子。
- **关于时间差：**如果要考虑得更精细，可以考虑下期与今期的时间差，比如预测自然日第二天的收益率和预测国庆假期后第一个交易日的收益率理应分开对待，因为两者的“上期与下期的时间差”不一致。

模型选择

- **线性模型最优先：**OLS、Lasso与Ridge Regression处理回归问题，Logistic Regression或Perceptron处理分类问题，可对建模结果做统计检验以了解模型可信度、输入feature的有效性。
- **树模型次之：**决策树（单用须剪枝），GBDT、XGBoost、LightGBM，Random Forest。可以查看特征重要性。
- **Learning Curve：**线性模型作为baseline，通过观察Learning Curve决定是否要提升模型复杂度。

训练策略

- N fold CV：不考虑时间序列性。
- 滑动窗口：用过去N期的数据训练模型，每隔一定的步长就重新训练一次。

机器学习

机器学习起什么作用？

- 高维模型寻找量价数据中的非线性逻辑，因子间的相互作用
- Lasso、Tree-based model的feature importance可用于因子筛选

Content

Validation test

因子检验

- **分组回测：**按因子值分组（应该要控制行业和市值），对比每个分组的收益率。以及每个分组各自收益率的稳定性。
- **回归：**加入到经典的因子模型中，查看模型解释力是否有所提升，及提升是否显著与稳定。
- **因子IC：**因子值与下期因子的秩相关系数。
- **因子与经典因子的相关性：**如与市值因子的相关性。我们希望得到好而不同的因子。
- **因子特性（可选）：**换手率、平均持仓周期等。

Validation test—does our model outperform the original?

Model comparison

- Follow the skeleton of literatures
  - Group portfolios
    - ✓ Anomaly variables (Hou, Xue and Zhang, 2015)
    - ✓ Frim characters (Fama and French, 1993& Fama and French，2015)

      eg: Size, BM, OP, etc.
  - Regression with/without our ML factor
    - ✓ α：The intercept
    - ✓ GRS result：whether we can regard the alpha of the model as 0
    - ✓ |α|/|r|：To what extent the dispersion of average excess returns can be explained
    - ✓ ……

Validation Test——另一些可能的方向

- GMM检验：用于估计股票的风险暴露；
- Fama-MacBeth 两步截面回归检验：排除残差在截面上的相关性对标准误的影响；
- BARRA提出的因子有效性检验标准：从因子对收益率影响的显著程度、稳定性以及因子之间的共线性；

Content

Timeline

Week1（10.5~10.12）：1. 阅读文献，包括机器学习在量化投资中应用的常见误区、机器学习处理高频量价数据、传统量价因子的构造、传统单因子检验与多因子模型，用ppt或markdown形式写成读书笔记；

Week2（10.12~10.19）:与Barra组讨论合作开发事宜并撰写需求文档、确定分工。

按分工编写程序，准备数据库，汇丰的同学尝试获取学校提供的高频数据；

Week3（10.19~10.26）：编写程序，完成与Barra组合作的部分（行情数据库、单因子检验、多因子模型、回测框架）；

Week4（10.26~11.2）：完成机器学习建模有关的程序。

Week5（11.2~11.9）：实现一套基于传统量价因子的多因子模型，作为之后的实验的Baseline。后续跟踪模型模拟盘表现。

Week6（11.9~11.16）：复现研报或论文中已有的机器学习因子，检验效果，攥写实验报告。如果有因子表现出好的效果，后续跟踪此因子模拟盘表现。

Week7（11.16~11.23）：基于上周实验结果，头脑风暴，提出并实现自己的想法，检验效果，攥写实验报告。如果有因子表现出好的效果，后续跟踪此因子模拟盘表现。

Week8（11.23~11.30）：整理所有项目资料，准备FinalPre。

北京大学量化交易协会