太平洋证券 PACIFIC SECURITIES

2019-10-18

量化研究报告

打开 AI 黑箱，探索可解释性

——“见微”系列之一

量化研究报告

证券分析师：徐玉宁

电话：021-58502206-8028

E-MAIL：xuyuning@tpyzq.com

执业资格证书编码：S1190519090003

报告摘要

可解释人工智能（XAI）能有效改善“准确性 VS 可解释性”的选择困境。在选择模型时，预测效果一直是首要考量，但同时又会担心模型过于复杂而难以掌控，往往要在准确性和可解释性之间做出抉择。XAI 能在不损失准确性的同时，一定程度上提高模型的可解释性，成为改善两难困境的一种重要手段。

XAI 可作为独立模块加入到现有的 AI 系统。XAI 存在三种常用的分类方法：按复杂程度分为事前与事后、按解释范围分为全局与局部、按模型相关度分为模型依赖与模型独立。从目前应用的情况来看，“事后&模型独立”的方法可作为独立模块加入到用户现有的 AI 系统中，颇受工业界的青睐。

XAI 的四类常用方法：Permutation Importance、SHAP、Partial Dependence 和 LIME。Permutation Importance 通过打乱特征值顺序来衡量特征重要性；SHAP 以合作博弈论为理论基础，衡量特征对预测结果的贡献；Partial Dependence 可以观察特征变化对预测结果产生的边际效应；LIME 利用白箱代理模型对给定实例附近的样本做局部近似，将白箱的解释结果作为该实例的决策依据。

XAI 可应用于多种策略模型。以多因子选股模型为例，XAI 从因子重要性、因子预测贡献、因子间交互、因子边际效应等多个视角展现选股模型的特性。

风险提示：对选股模型的结果展示仅用于工具介绍，不构成任何投资建议。XAI 相关方法存在一定局限性，不保证对所有场景和模型均得到理想的解释结果。

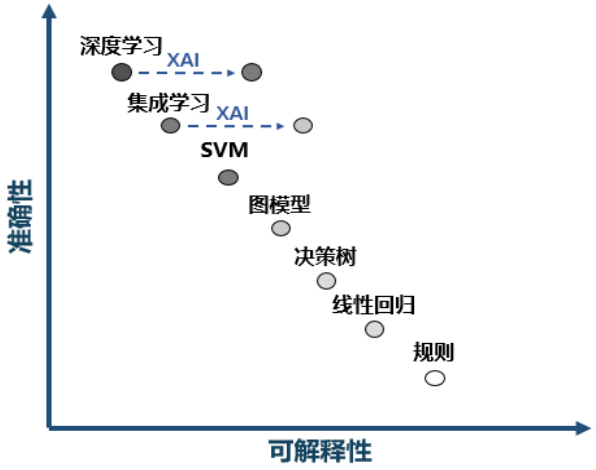
目录

图表目录

一、引言

近年来，AI 技术在多个领域都取得了骄人战果，某些场景的表现甚至超过人类，比如 AI 人脸识别的准确率超过人眼准确率、AlphaGo 战胜人类世界冠军。然而，新兴技术是把“双刃剑”，效果提升的同时也引发对模型掌控力下降。特别是自动驾驶、医疗检测等高风险低、容错率的应用场景，唯“效果论”可能会引发严重的后果。投资领域亦是如此，**我们不得不在可解释性与准确性之间做出抉择：通常会牺牲部分准确性，而去选择可解释性更高的模型**。

图表 1　AI 模型准确性与可解释性的示意图



资料来源：太平洋证券研究院整理

一般而言，将深度学习模型、复杂集成模型等归为“黑箱模型”(Black-box model)，庞大的参数量、特征间的深层交互和复杂的模型结构使它们具备强大拟合能力；而决策树和线性回归则因透明运作机制被归为“白箱模型”(White-box model)。**可解释人工智能（EXplainable Artificial Intelligence，简称 XAI）能在不损失准确性的同时，一定程度上提高模型的可解释性。**类似于对黑箱注入“白色”元素，使我们从“灰箱”中部分地了解模型运作方式。

本文是一篇关于 XAI 的工具类报告，主要涉及以下三个方面：

1、梳理 XAI 领域相关研究；

2、详细介绍几类常用方法的算法原理与特点；

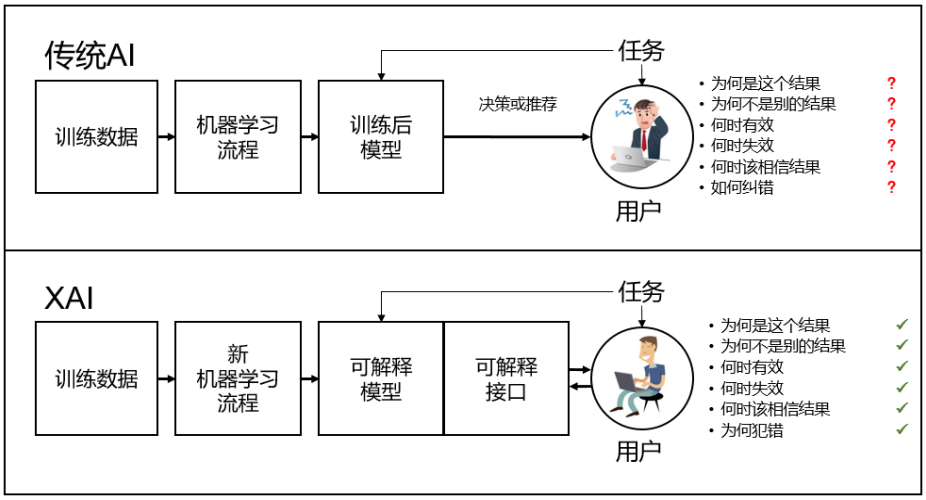
3、展示在多因子选股案例中的实现与应用。

二、初探 XAI

(一)、何为 XAI ？

传统 AI 的主要目的是输出最终决策结果，而 XAI 则在流程中融入可解释性要素，以能被人所理解的形式反馈给用户决策依据。用户可以从任务的视角尝试理解输出的结果，这不单提供心理上的信赖感，更给予管控和改进模型的可能，甚至能帮助我们加深理解数据和问题本身。以信用评估的场景为例，如果训练目标是最小化违约概率，那么模型很可能对某些人群存在“偏见”，XAI 或许能探测出这些人群的显著特征，帮助我们深入理解数据并协助设定更合理的目标函数。

图表 2　传统 AI 与 XAI 的流程对比图



资料来源：太平洋证券研究院整理

但客观来讲，目前的 XAI 仍存在较多局限，离“黑箱变白箱”的目标还有不小距离：1）**可解释性定义不清**：需要联系使用者的需求和主观感受，导致仍未有被广泛认可的阐述；2）**有时难以被解释**：比如解释难度会随模型复杂度的提升而同步提升，又比如模型多样性（Multipilicty of good models）的存在导致难以得出统一的解释；3）**可解释性与准确性的两难问题会有所削弱，但依然存在**。

(二)、分类框架

XAI 领域中存在数量众多、种类庞杂的各类方法，为了更系统化地介绍，我们引入学术界公认的三种划分维度：1）按复杂程度分为“事前”与“事后”；2）按解释范围分为“全局”与“局部”；3）按模型相关度分为“模型依赖”与“模型独立”。

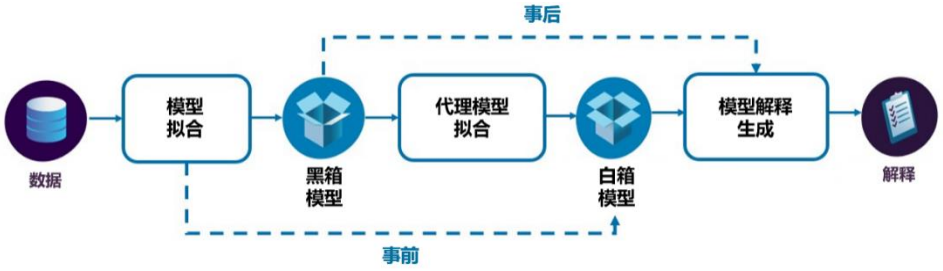
1、模型依赖与模型独立

模型依赖与模型独立的主要区别在于是否针对特定模型。对于“模型依赖”方法，低复杂度模型（线性模型）的诊断、检验、评估工具比较完善；而高复杂度模型（神经网络模型）或优化算法的研究成本较高、相关工具的研发仍在不断推进中，比较知名的有 Deep Lift、SLIM（Supersparse Linear Integer Models）等；也有用特定的决策树模型来解释模型内部结构。值得注意的是：**从效果层面看，“模型依赖”方法能更深入挖掘特定模型的内部机制，在解释能力上更具优势；从应用层面看，“模型独立”方法对模型本身无特定要求，还能对不同模型进行横向比较，具备更强的适用性和灵活性。**

2、事前与事后

事前（Ante-hoc）可解释性主要借助模型本身的可解释能力，通常是直接选择结构简单、可解释性好的白箱模型来训练。事后（Post-hoc）可解释性则采用类似逆向工程手段，先用高准确率的黑箱模型拟合，再用 XAI 方法去解释黑箱。

图表 3　基于事前与事后的机器学习流程图

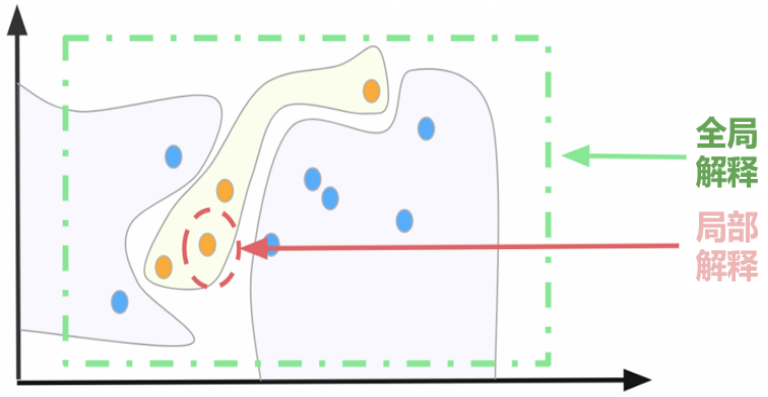


资料来源：太平洋证券研究院整理

3、全局与局部

全局可解释性（Global Interpretability）与局部可解释性（Local Interpretability）的差异在于解释的对象不同：前者是对于整个训练数据集，旨在帮助人们理解复杂模型背后的整体逻辑以及内部的工作机制；后者是针对特定样例，旨在帮助人们理解模型对特定输入样本的决策依据。另外，局部数据的决策边界更倾向于线性、单调，这使得局部可解释性在某种程度上会比全局可解释性更精确。

图表 4 全局解释与局部解释的示意图



资料来源：太平洋证券研究院整理

（三）、方法梳理

通过翻阅 XAI 的相关资料，发现该领域的研究数量近三年出现井喷，特别是 Google、IBM、微软等几家科技巨头都在大力开展研发，毫无疑问 XAI 已经成为学术界和工业界的研究热点。由于方法众多，我们仅在下表中列举了部分经典算法。**从论文数量上来看，目前的绝大部分研究都基于“事后”方法；而从目前应用的情况来看，“事后&模型独立”的方法可作为独立模块加入到用户现有的 AI 系统中，颇受工业界的青睐。**

XAI 方法得到的实际结果大多为：1）特征对预测值的影响程度；2）预测值的置信度；3）模型内部的近似结构，表现形式包括数值、二维曲线图、树状结构图等。**出于稳健性的考虑，我们建议读者对同一模型使用多种 XAI 方法，观察是否能得到较为一致的结论。**

图表 5 XAI 方法列举图表

名称	年份	作者	全局/局部	事前/事后	模型 独立/依赖
LIME	2016	Ribeiro et al.	局部	事后	独立
Permutation Importance	2018	Fisher et al.	全局	事后	独立
SHAP	2016	Lundberg et al.	局部	事后	独立
Deep Lift	2017	Shrikumar et al.	全局	事后	依赖
PDP	2001	Friedman	全局	事后	独立
ICE	2015	Goldstein et al.	局部	事后	独立
ALE	2016	Apley et al.	全局	事后	独立
TreeView	2016	Thiagarajan et al.	局部	事后	依赖
Rule Set	2016	Wang et al.	全局	事前	独立
FRL	2015	Wang et al.	全局	事前	独立

资料来源：太平洋证券研究院整理

三、方法介绍

我们挑选了四种常用方法做详细介绍，以相对的通俗语言阐述核心思路并列出主要算法流程，同时还梳理各方法的特点，期望能让读者有直观而不失深度的理解。

(一)、Permutation Importance

1、算法介绍

Permutation Importance（简称 PI）是 Feature Importance 中的一种方法，用来衡量每个特征对于提升模型预测能力的贡献程度。具体做法：**打乱数据集内目标特征数值的顺序，考察打乱前后模型预测效果的差异。**如果打乱后模型预测效果出现明显下降，说明目标特征对于预测任务的贡献程度较高；反之，则说明目标特征的重要性较弱。

图表 6 PI 做法示意图

特征A	特征B（目标特征）	特征C	标签Y
XA1	XB1	XC1	Y1
XA2	XB2	XC2	Y2
XA3	XB3	XC3	Y3
XA4	XB4	XC4	Y4
XA5	XB5	XC5	Y5
XA6	XB6	XC6	Y6
XA7	XB7	XC7	Y7

资料来源：太平洋证券研究院整理

图表 7 PI 算法的伪代码

- Require：训练好的模型f，特征矩阵X，目标向量v，误差函数$L(v,f)$
- $e^{origin} \leftarrow L(Y, f(X))$
- *For Feature* j *in* $\{1,\cdots,p\}$ *do*:
  - 随机重排X中特征j生成特征矩阵X^{perm}
  - $e^{perm} \leftarrow L(Y, f(X^{perm}))$
  - $FI^j \leftarrow e^{perm}/e^{origin}$ 或 $e^{perm} - e^{origin}$
- *return* FI

资料来源：太平洋证券研究院整理

2、方法特点

1. **不同模型间存在可比性。**在比例计算方法下，同一特征在不同模型中的重要性可以进行比较。

2. **无需重新训练模型。**对比 Drop Columns Importance 方法（删除目标特征后重新训练模型，再计算误差差异），PI 仅仅是改变特征值的顺序而非输入数据的结构，避免重新训练而节省了计算时间。

3. **考虑了特征间交互作用。**特征重新洗牌不仅打乱目标特征，也打乱目标特征与其它特征之间的原始关联，所以 PI 也反映了特征间交互作用的效果。

4. **易受多重共线性的影响。**假设高相关的特征 A 与特征 B 均对预测均有重要作用，但分别计算特征 A、特征 B 的 PI 会发现，两者都显示为非重要特征，由此掩盖了特征重要性的真实情况。

5. **易产生脱离真实的数据。**如果对数据集中存在高相关的特征重新洗牌，可能会产生现实中不存在的虚假数据。例如：特征中包含身高与体重，打乱其中一个特征后，或许会出现 2 米身高对应 30 公斤体重的数据。

(二)、SHAP

1、算法介绍

SHAP（全称 SHapley Additive exPlanation）是一种局部的特征重要性方法，显示了每次预测中目标特征对预测结果产生影响的大小和正负性。通俗理解，其效果就类似于对线性回归的预测值做拆分，将目标特征的特征值与对应回归系数的乘积作为特征重要性的度量。

核心的 Shapley Value 理论源自于“合作博弈论”，研究的是如何将参与者合作产生的收益进行分配。以纸牌游戏场景为例，涉及三位玩家 A、B、C。首先考虑三位玩家在场或不在场时产生的价值（共 8 种组合情况），由此可以推算出每位玩家以不同次序加入牌局时产生的边际价值（共 6 种排序情况），最后将每位玩家在所有排序情况下产生的边际价值求平均即可得到各自的 Shapley Value。

图表 8　Shapley Value 核心思路的示意图

*资料来源：*https://clearcode.cc/blog/game-theory-attribution/，*太平洋证券研究院整理*

根据 Lundberg et al.(2016)的描述，特征 i 贡献的计算公式如下：

$$\psi_i = \sum_{S \subset F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

其中，F 代表特征全集，S 表示排在特征 i 之前的特征子集，$|F| - |S| - 1$表示排列在特征 i 之后的特征数量，$|F|!$表示所有特征的排序数量，$\frac{|S|!(|F|-|S|-1)!}{|F|!}$ 类似于当前特征子集 S 情况下的权重，$f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$ 表示加入与不加入特征 i 的预测差异（边际贡献）。

不过，上述公式的计算复杂度过高，实际应用中会采用如下近似算法：

图表 9 SHAP 算法的伪代码

- Require：迭代次数M，模型f，数据集X，考察特征序号j，考察实例x
- $For\ m\ in\ \{1,2,3,\cdots M\}\ do$:
  - 从X中随机抽取实例z，生成特征的随机排序o
  - $x_o \leftarrow (x_{(1)}, x_{(2)}, \cdots, x_{(j)}, \cdots, x_{(p)})$
  - $z_o \leftarrow (z_{(1)}, z_{(2)}, \cdots, z_{(j)}, \cdots, z_{(p)})$
  - $x_{+j} \leftarrow (x_{(1)}, x_{(2)}, \cdots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \cdots, z_{(p)})$
  - $x_{-j} \leftarrow (x_{(1)}, x_{(2)}, \cdots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \cdots, z_{(p)})$
  - $\psi_j^m = f(x_{+j}) - f(x_{-j})$
- $return\ \psi_j(x) = \frac{1}{M}\sum_{m=1}^{M}\psi_j^m$

资料来源：太平洋证券研究院整理

2、方法特点

1. 考虑了特征间交互作用。

2. 允许对比解释。可以将目标样本的预测与任意数据集的预测进行比较，得到对应的特征贡献值。

3. 计算效率低。若需要精确计算 XGBoost 模型的 SHAP，外层循环遍历所有可能的特征组合，问题的复杂度为 O(TL2N)（T 是树的数量，L 是最大叶子节点数，N 是特征数）。

4. 易产生脱离真实的数据。

（三）、Partial Dependence

1、算法介绍

Partial Dependence（简称 PD）是一类可解释性方法的统称，其思路是控制其他变量不变的情况下改变目标特征的值，以图像形式呈现模型预测结果的变化。此外，根据处理问题的不同衍生出 2 种具体的方法：Partial Dependence Plot（简称 PDP）和 Individual Conditional Expectation（简称 ICE）Plot。

PD 计算公式如等式 1 所示，反映了预测平均结果与目标特征值之间的变化关系，其中 x_S 表示待考察特征集合（通常不超过 2 个），x_C 表示其余特征的集合。而在实际计算中一般会采用求均值方法，如等式 2 所示，$x_C^{(i)}$ 表示其余特征在样本 i 的取值。

图表 10　PD 计算公式

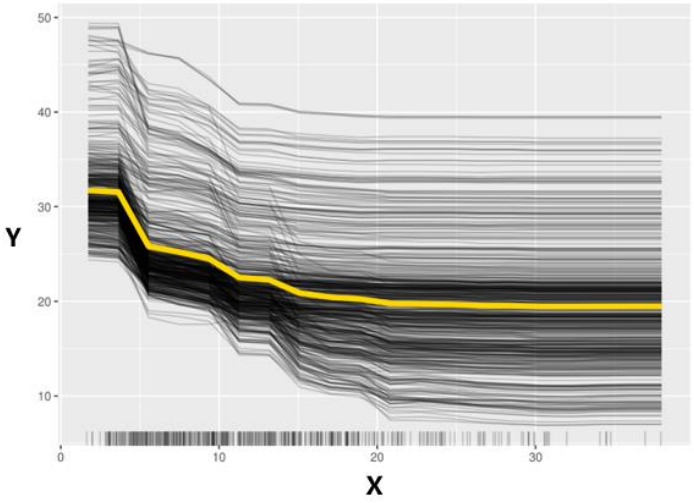
$$\hat{f}_{x_S}(x_S) = E_{x_C}[\hat{f}(x_S, x_C)] = \int \hat{f}(x_S, x_C) dP(x_C) \qquad \text{(等式1)}$$

$$\hat{f}_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} f(x_S, x_C^{(i)}) \qquad \text{(等式2)}$$

资料来源：太平洋证券研究院整理

相较于 PDP 展示全局变化情况，ICE 则类似于“局部版 PDP”，即展示各个样本单独的变化情况。两者效果如下图所示，黑线表示 ICE、黄线表示 PDP，**PDP 实际是所有 ICE 汇总后的结果。**

图表 11　PDP 与 ICE 效果示意图



资料来源：太平洋证券研究院整理

2、方法特点

1. PDP 展示总体，ICE 展示个体。

2. 特征数量受到限制。受制于可视化的原因，PDP 最多展示 2 个特征（三维图像或二维图像），ICE 只适用于 1 个特征（二维图像）。

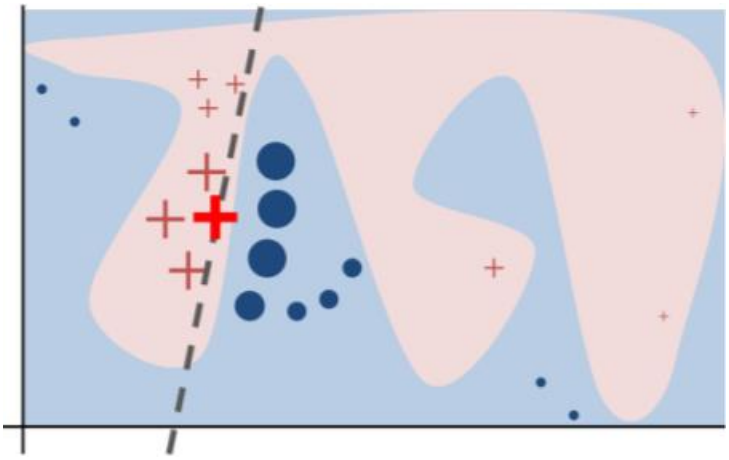
3. 当特征之间明显相关时，结果不可信。原因在于引入了现实中不存在的数据，导致估计结果存在偏差，ALE（Accumulated Local Effects）方法对此做了改进。

（四）、LIME

1、算法介绍

LIME（Local Interpretable Model-Agnostic Explanations）核心思想是用白箱模型对给定实例附近的样本做局部近似，将白箱的解释结果作为该实例的决策依据。虽然模型对全局的决策边界会非常复杂，但对单一样本的决策边界通常会比较简单，甚至可以近似为线性。如下图所示，粉红/蓝色背景为模型预测结果，“+”和“• ”代表标签类别，红色加粗“+”为给定实例，其它样本通过采样生成，灰色虚线为 LIME 用线性模型对该实例给出的决策边界。

图表 12 LIME 决策边界示意图



资料来源： https://github.com/marcotcr/lime，*太平洋证券研究院整理*

一般形式如下式所示，其中 x 为给定实例，f 为待解释的黑箱模型，G 为白箱模型集合，π_x 表示考察数据与 x 的相近程度，L 衡量白箱模型的解释误差。LIME 期望能得到对 x 解释误差最小的白箱模型。

$$explanation(x) = argmin_{g \in G} L(f, g, \pi_x)$$

在实际情况中，通常会给定白箱的模型类型（例如 Lasso）和误差函数（例如平方差函数），π_x 则体现为权重（相近样本的权重更大），具体如下式所示。

$$L(f, g, \pi_x) = \sum_{z \in X} \pi_x(z)(f(z) - g(z))^2$$

图表 13 LIME 算法的伪代码

- Require：黑箱模型f，抽样数量N，给定实例x，近似程度函数π_x，LASSO系数个数K
- $Z \leftarrow \{\}$
- $for\ i \in \{1,2,3,\cdots,N\}\ do$
  - $z \leftarrow sample\_around(x)$
  - $Z \leftarrow Z \cup \langle z_i, f(z_i), \pi_x(z_i)\rangle$
- $end\ for$
- $w \leftarrow K\text{-}LASSO(Z,K)$
- $return\ w$

资料来源：太平洋证券研究院整理

2、方法特点

1. 适用于非结构化数据，如文本、图像。

2. 需要选择合适的相近度函数。

四、应用案例

我们以多因子选股为具体的应用场景，使用机器学习模型在每月进行训练、测试，再将 XAI 方法用于解释模型预测的结果。需要说明的是，本文不会过多纠结于因子选择、预处理、模型选择、训练方式等环节的处理细节，也不刻意追求预测效果和回测表现，主要关注点在XAI方法使用的介绍和展示。此外，实验中以调用Python相关包为实现手段，使用其它语言（如 R 语言）的读者也能从网上找到类似资源。

(一)、数据介绍与整体流程

1．股票数据：全市场股票，剔除新上市、停牌股票

2．因子数据：选择估值、盈利、成长等 8 个维度共 31 个因子；

3．因子预处理：包括异常值缩尾、缺失值填补、截面标准化、行业市值中性；

4．标签预处理：1）分类 将下月收益排名前 30%标注为+1，后 30%标注为-1，中间 40%标注为 0；2）回归 下月相对全市场超额收益；

5．模型选择：选择拟合能力较强的 XGBoost，对其做一定程度参数调优；

6．训练方式：滚动训练，采用过去 2 年数据；

7．测试区间：2010 年 1 月 31 日至 2019 年 8 月 30 日。

图表 14 因子数据

类别	因子名称	因子描述
估值	BP	股东权益合计/总市值
估值	EP	净利润（TTM）/总市值
估值	SP	营业收入（TTM）/总市值
估值	OCFP	经营性现金流（TTM）/总市值
估值	DCFP	净现金流（TTM）/总市值
盈利	ROE_q	ROE（单季度）
盈利	ROE_ttm	ROE（TTM）
盈利	ROA_q	ROA（单季度）
盈利	ROA_ttm	ROA（TTM）
盈利	GrossMargin_q	毛利率（单季度）
盈利	GrossMargin_ttm	毛利率（TTM）
成长	ROE_YoY	ROE同比增速
成长	ROE_q_YoY	单季度ROE同比增速
成长	NetProfit_YoY	净利润同比增速
成长	NetProfit_q_YoY	单季度净利润同比增速
成长	OR_YoY	营业收入同比增速
成长	OR_q_YoY	单季度营业收入同比增速
成长	OCF_YoY	经营现金流同比增速
成长	OCF_q_YoY	单季度经营现金流同比增速
杠杆	CurrentRatio	流动资产/流动负债
杠杆	DebtAssetRatio	总负债/总资产
规模	MV_ln	总市值取对数
规模	FMV_ln	流通市值取对数
反转	Ret_1M	过去1月收益
反转	Ret_3M	过去3月收益
反转	MaxRet_1M	过去1月最大收益
流动性	Turnover_avg_1M	过去1月日均换手率
流动性	Volume_std_1M	过去1月成交量标准差
波动率	RetStd_1M	过去1个月日收益序列标准差
波动率	RetStd_3M	过去3个月日收益序列标准差
波动率	RetStd_FF3_1M	过去1个月特质波动率

资料来源：太平洋证券研究院整理

（二）、Permutation Importance 应用

虽然 PI 算法不难实现，但 Eli5 简洁的调用方式和对 sklearn 框架的支持，值得做一回“拿来主义”。（Github 链接：https://github.com/TeamHG-Memex/eli5）

具体使用上，1）需要输入经过训练的模型（也支持 Cross-Validation 方式下重新训练模型）；2）可设定迭代次数和随机种子，加大迭代次数能提高精度但也会拖慢运行效率（从实验情况看，不宜设置过高的迭代次数）；3）评估数据可选择样本内训练集或样本外测试集；4）支持自定义损失函数，默认为模型训练时的损失函数。返回的原始结果是每个特征每次迭代的得分矩阵，最终结果是统计每个特征在所有迭代的均值和标准差。

图表 15 Eli5 调用示例

```
import eli5
from eli5.sklearn import PermutationImportance

perm = PermutationImportance(model, n_iter=5, random_state=1).fit(X_train, Y_train)
eli5.explain_weights(perm, feature_names=factor_list)
```

资料来源：太平洋证券研究院整理

当然 XGBoost 属于树模型，模型自身就能依据特征作为划分属性的次数、样本覆盖量或信息增益来度量特征重要性。但 PI 并不会毫无价值，**它能提供额外的度量方式且不受模型类型限制，还能观察模型对样本外特征数据的泛化能力。**

以某一期训练模型为例，同时计算五种特征重要性结果（PI 对样本外测试集、PI 对样本内训练集、以及 XGBoost 自带的 weight、gain 和 cover 方法），再按照 PI 对样本外测试集的得分降序排列。从下图结果看，1）测试集上的特征亦会存在“负向贡献”；2）样本内外特征表现并非一致；3）PI 与 XGBoost 自带方法的结果存在些许出入，可以考虑作为特征重要性的补充。

图表 16 PI 特征重要性结果图表

因子名称	PI(测试集)		PI(训练集)		XGB(weight)		XGB(gain)		XGB(cover)	
	排名百分位	得分	排名百分位	得分	排名百分位	得分	排名百分位	得分	排名百分位	得分
RetStd_FF3_1M	1.00	0.0139	0.84	0.0050	0.77	0.0441	1.00	0.1631	1.00	0.0533
Ret_1M	0.97	0.0101	1.00	0.0100	1.00	0.0905	0.87	0.0476	0.81	0.0393
RetStd_3M	0.94	0.0092	0.90	0.0059	0.87	0.0517	0.97	0.1049	0.94	0.0435
Turnover_avg_1M	0.90	0.0063	0.97	0.0086	0.90	0.0608	0.94	0.0957	0.90	0.0414
OCFP	0.87	0.0042	0.61	0.0011	0.50	0.0254	0.58	0.021	0.58	0.0360
Volume_std_1M	0.82	0.0035	0.65	0.0012	0.81	0.0445	0.65	0.0225	0.68	0.0380
Ret_3M	0.82	0.0035	0.87	0.0056	0.94	0.0733	0.81	0.0404	0.74	0.0386
MV_ln	0.77	0.0022	0.94	0.0074	0.97	0.0891	0.74	0.0281	0.77	0.0389
OR_q_YoY	0.74	0.0019	0.77	0.0013	0.74	0.0354	0.55	0.0198	0.52	0.0347
GrossMargin_ttm	0.71	0.0017	0.13	0.0003	0.03	0.0105	0.19	0.0153	0.39	0.0301
SP	0.68	0.0017	0.55	0.0010	0.50	0.0254	0.23	0.0154	0.71	0.0384
NetProfit_q_YoY	0.65	0.0010	0.32	0.0005	0.40	0.0220	0.48	0.0188	0.29	0.0273
DCFP	0.60	0.0009	0.29	0.0005	0.61	0.0287	0.26	0.0157	0.35	0.0286
CurrentRatio	0.60	0.0009	0.19	0.0005	0.29	0.0172	0.32	0.0161	0.32	0.0276
FMV_ln	0.55	0.0006	0.74	0.0013	0.71	0.0326	0.39	0.018	0.45	0.0307
ROA_q	0.52	0.0005	0.58	0.0010	0.40	0.0220	0.77	0.0288	0.97	0.0439
OCF_YoY	0.48	0.0005	0.26	0.0005	0.26	0.0163	0.16	0.015	0.06	0.0172
NetProfit_YoY	0.45	0.0003	0.10	0.0002	0.06	0.0115	0.13	0.0148	0.19	0.0214
OCF_q_YoY	0.40	0.0002	0.42	0.0006	0.65	0.0292	0.35	0.0164	0.55	0.0356
RetStd_1M	0.40	0.0002	0.45	0.0008	0.68	0.0311	0.61	0.0216	0.48	0.0316
OR_YoY	0.35	0.0001	0.23	0.0005	0.32	0.0182	0.10	0.0148	0.26	0.0264
ROA_ttm	0.31	0.0001	0.16	0.0004	0.10	0.0120	0.45	0.0184	0.10	0.0194
MaxRet_1M	0.31	0.0001	0.52	0.0009	0.55	0.0268	0.42	0.0181	0.84	0.0404
ROE_q_YoY	0.26	-0.0001	0.35	0.0006	0.19	0.0144	0.68	0.0231	0.23	0.0258
DebtAssetRatio	0.23	-0.0002	0.03	0.0001	0.16	0.0134	0.06	0.0146	0.03	0.0060
ROE_ttm	0.19	-0.0003	0.48	0.0009	0.45	0.0249	0.52	0.0193	0.61	0.0362
ROE_YoY	0.16	-0.0003	0.06	0.0001	0.13	0.0129	0.29	0.016	0.13	0.0202
GrossMargin_q	0.13	-0.0005	0.39	0.0006	0.23	0.0158	0.03	0.0145	0.16	0.0202
EP	0.10	-0.0007	0.68	0.0012	0.58	0.0283	0.71	0.0255	0.42	0.0305
ROE_q	0.06	-0.0013	0.71	0.0012	0.35	0.0211	0.84	0.0447	0.87	0.0411
BP	0.03	-0.0025	0.81	0.0036	0.84	0.0508	0.90	0.0519	0.65	0.0377

资料来源：太平洋证券研究院整理

（三）、SHAP 应用

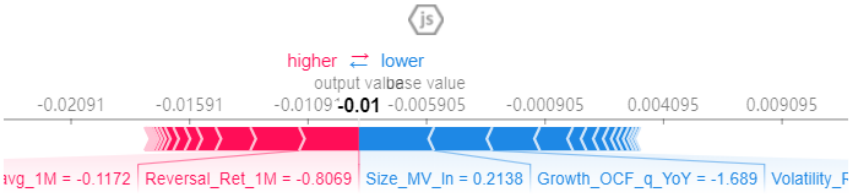
SHAP 的实现可参考同名包 shap，在基础功能之外还提供特征交互作用和类似 PDP 的展示。（Github 链接：https://github.com/slundberg/shap）

具体使用上，1）加载 JS 可视化代码；2）调用解释器，树模型可使用特定解释器 TreeExplainer 或通用解释器 KernelExplainer；3）计算训练数据预测均值（expected_value）和样本的 shap value。

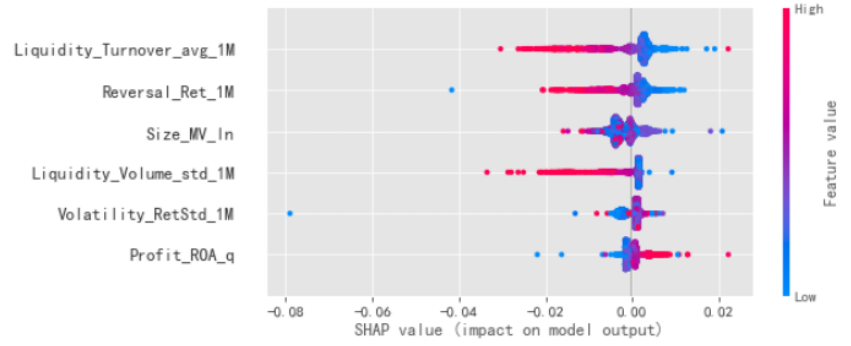
以回归问题为例，图中上半部分为单一样本分析，模型对训练数据总体的预测均值为-0.0059，对该样本预测值为-0.01，红蓝条状分别表示特征对预测值正向、负向贡献的大小，可见 MV_ln 因子和 OCF_q_YoY 因子是推动负向偏离的中坚力量。图中下半部分为多个样本整体分析，散点对应单个样本、颜色表示特征值大小，可见 Turnover_avg_1M 和 Ret_1M 因子值较高时对预测值有明显拉低作用。

图表 17　shap 调用示例与结果显示（回归）

```
import shap

### 回归模型
shap.initjs()
explainer = shap.TreeExplainer(model_reg)
shap_values = explainer.shap_values(X_test)
shap.force_plot(explainer.expected_value, shap_values[0], X_test.iloc[0])
```




```
shap.summary_plot(shap_values, X_test, max_display=6)
```

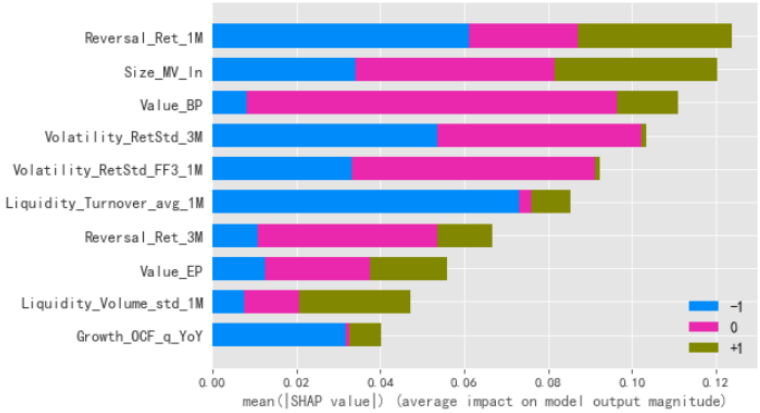


资料来源：太平洋证券研究院整理

分类问题则稍有不同，下图显示了每个特征对三种类别概率的影响大小，发现 Turnover_avg_1M 和 Ret_1M 因子对该样本分为-1 类别存在显著影响。

图表 18 shap 调用示例与结果显示（分类）

```
import shap

### 分类模型
shap.initjs()
shap_values = shap.TreeExplainer(model_clf).shap_values(X_test.iloc[0])
shap.summary_plot(shap_values, X_test.iloc[0:1], class_names=['-1','0','+1'],max_display=10)
```



资料来源：太平洋证券研究院整理

（四）、Partial Dependence 应用

目前有不少工具包含 PDP 算法和可视化功能（包括 sklearn 自带的 API），由于 PDPBox 在可视化的效果比较突出，我们将其选为具体介绍对象。（Github 链接：https://github.com/SauceCat/PDPbox）

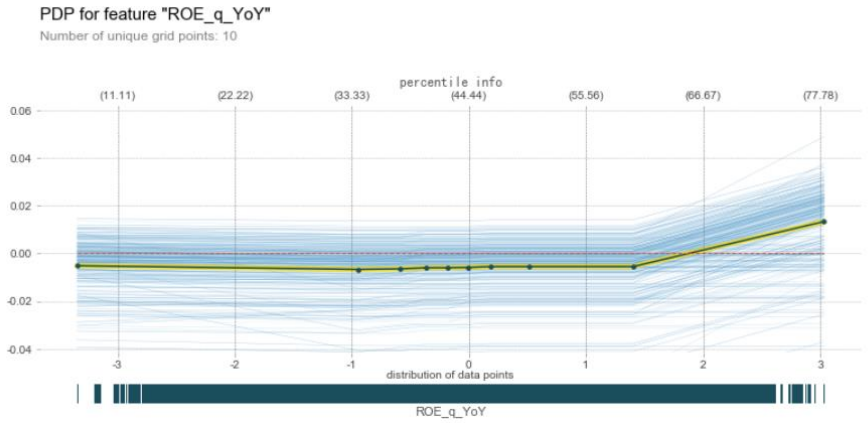
具体使用上，1）需要输入经训练的模型、数据集、特征列表以及选定的特征名称；2）调用绘图 API，涉及不少绘图参数，包括基本的图像大小、位置等，也包括 ICE 的显示、样本特征值分布、中心化显示等。

以 ROE_q_YoY 因子为例，从 PDP 结果看，大部分区间的偏效应变化比较平缓，但在因子值顶端区域出现明显上升，说明该特征值较大时会对预测产生较强的正向作用；从 ICE 结果看，黄线以上部分样本更集中、黄线以下部分的样本更扩散。

此外，PDPBox 提供了两个特征交互作用的展示。以 Volume_std_1M 和 Ret_1M 因子为例，在 Ret_1M 顶部和 Volume_std_1M 两端的交集区域，存在比较明显的负向偏效应。

图表 19　PDPBox 调用示例与结果显示（单特征）

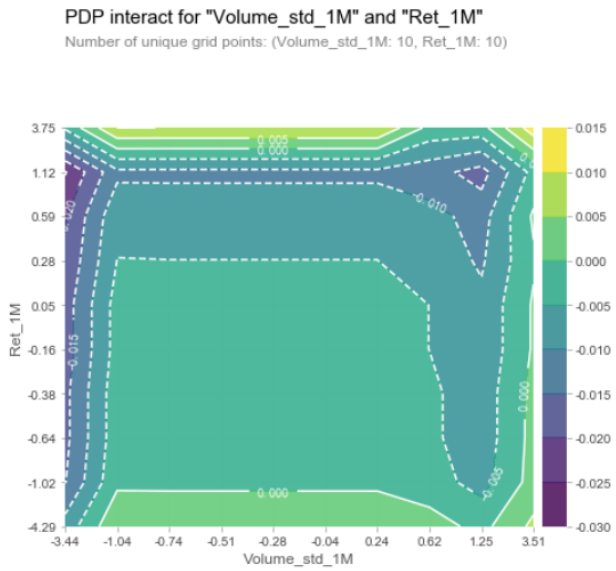
```
from pdpbox import pdp

pdp_factor = pdp.pdp_isolate(model=model_reg, dataset=X_train,
                             model_features=factorList,feature='Growth_ROE_q_YoY')
fig, axes = pdp.pdp_plot(pdp_factor,'ROE_q_YoY',figsize=(13,7), plot_lines=True, center=False,
                         frac_to_plot=300, plot_pts_dist=True, show_percentile=True)
```



资料来源：太平洋证券研究院整理

图表 20　PDPBox 调用示例与结果显示（双特征）

```
from pdpbox import pdp,info_plots
inter_rf = pdp.pdp_interact(
    model=model_reg, dataset=X_train, model_features=factorList,
    features=['Liquidity_Volume_std_1M', 'Reversal_Ret_1M']
)
fig, axes = pdp.pdp_interact_plot(
    inter_rf, ['Volume_std_1M', 'Ret_1M'], figsize=(7,8), x_quantile=True,
    plot_type='contour', plot_pdp=False
)
```



资料来源：太平洋证券研究院整理

（五）、LIME 应用

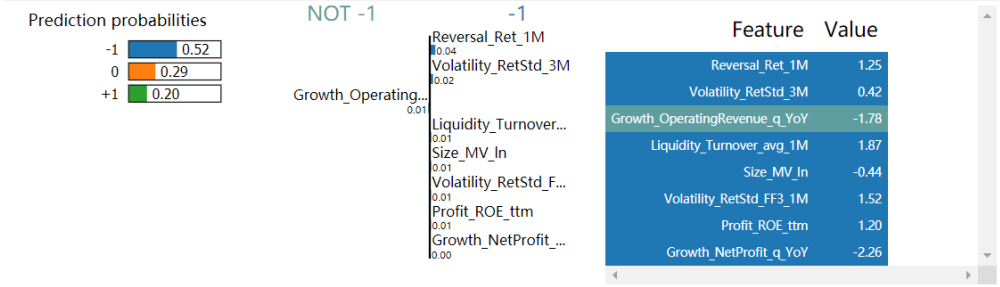
LIME 的实现可参考同名包 lime，由提出该方法的作者 Marco Tulio Ribeiro 作为主要开发人员，支持 sklearn 框架，属于使用频率较高的一款工具包。（Github 链接：https://github.com/marcotcr/lime）

由于 LIME 在算法上存在一些不明确的“自由空间”，所以我们参考 lime 源代码和原文《“Why Should I Trust You?”Explaining the Predictions of Any Classifier》对 lime 包所采用的具体做法做简要介绍。模型方面主要采用稀疏线性模型，1）白箱模型：默认为岭回归，也可调用 sklearn 中其它线性回归模型；2）损失函数：使用加权平方损失函数；3）特征筛选：默认是以岭回归为基模型做前向特征选择，也可直接选择高权重特征或 LASSO 筛选法。相近样本采样方面，1）样本来源：一种是在给定样例基础上添加随机扰动来生成相近样本，另一种是在训练数据中心生成附近样本；2）相近度度量：默认采用欧氏距离，也可参照 sklearn.metrics.pairwise_distances 支持的其它函数，然后利用核函数将距离转化为值域在 0 至 1 的权重系数。

具体使用上，1）需要输入训练数据、特征名称、类别名称、选定样本和模型预测函数；2）可设定参数包括：特征数量、采样数量、随机种子等。以某只个股为例，左侧显示黑箱模型的预测概率，中间显示将其分为-1 或非-1 影响最大的前八个特征及对应的影响系数，右侧显示这八个特征的特征值。

图表 21 lime 调用示例与结果显示

```
import lime
import lime.lime_tabular


explainer = lime.lime_tabular.LimeTabularExplainer(
    X_train.values, feature_names=factorList, class_names=['-1','0','+1'], sample_around_instance=True,
    discretize_continuous=False, random_state=1
)
exp = explainer.explain_instance(
    X_test.iloc[10].values, model_clf.predict_proba, top_labels=1, num_features=8, num_samples=10000
)
exp.show_in_notebook(show_table=True, show_all=False)
```



资料来源：太平洋证券研究院整理

五、总结与展望

本文对 XAI 的背景、理论、方法及应用做了全面介绍。报告开篇阐述了 AI 技术会引发“准确性 VS 可解释性”两难困境，由此凸显出 XAI 的重要性。随后，对 XAI 领域的相关研究做了大致梳理，尤其是包含三种维度的分类框架。第三部分则对四种常用方法的理论进行介绍，从核心思路到具体算法，再到特点总结。最后，以多因子选股为场景，从实际应用角度展示了各方法的工具资源、调用示例及效果展示。

XAI 研究正当时。随着 AI 技术不断渗透深入到各行各业，黑箱模型的弊端也愈发显现，引起了业界和学术界高度关注，相关研究也迎来爆发式增长。本文涉及的内容仅仅是管中窥豹，在梳理的过程中我们也领略到其它前沿的研究成果，尤其对深度学习的可解释性更是巨大挑战，有待今后密切跟踪和深入探索。

理性看待 XAI。尽管 XAI 方法众多，但大多在实际应用有诸多限制，离大众意识中的看透黑箱存在不小距离。但瑕不掩瑜，XAI 仍是提供模型内在机理信息的有效手段。

六、参考文献

[1] Lundberg, Scott M., and Su-In Lee. “A unified approach to interpreting model predictions,” Advances in Neural Information Processing Systems. 2017.

[2] Alvarez-Melis, David, and Tommi S. Jaakkola. “On the robustness of interpretability methods,” arXiv preprint arXiv:1806.08049 (2018).

[3] Friedman, Jerome H. “Greedy function approximation: A gradient boosting machine,” Annals of statistics (2001): 1189-1232.

[4] Goldstein, Alex, et al. “Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation,” Journal of Computational and Graphical Statistics 24.1 (2015): 44-65.

[5] Apley, Daniel W. “Visualizing the effects of predictor variables in black box supervised learning models,” arXiv preprint arXiv:1612.08468 (2016).

[6] Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. “Model Class Reliance: Variable importance measures for any machine learning model class, from the ‘Rashomon’ perspective,” (2018).

[7] Deng, Houtao. “Interpreting Tree Ensembles with inTrees,” arXiv:1408.5456. 10.1007/s41060-018-0144-8 (2014).

[8] Thiagarajan, Jayaraman J., et al. "TreeView: Peeking into deep neural networks via feature-space partitioning," arXiv preprint arXiv:1611.07429 (2016).

[9] Wang, Fulton, and Cynthia Rudin. "Falling rule lists," Artificial Intelligence and Statistics. 2015.

[10] Letham, Benjamin, et al. "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," The Annals of Applied Statistics 9.3 (2015): 1350-1371.

[11] Lakkaraju, Himabindu, Stephen H. Bach, and Jure Leskovec. "Interpretable decision sets: A joint framework for description and prediction," Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.

[12] Li, Zairan, et al. "Rule-based back propagation neural networks for various precision rough set presented KANSEI knowledge prediction: a case study on shoe product form features extraction," Neural Computing and Applications 28.3 (2017): 613-630.

[13] Amina Adadi, Mohammed Berrada. "Peeking inside the black-box: A Survey on Explainable Artificial Intelligence(XAI) ," IEEE Access(2018).

[14] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," arXiv preprint arXiv:1802.01933, 2018.

[15] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. "Machine learning interpretability: A survey on methods and metrics," Electronics, 8(8):832, 2019.

[16] Molnar, C., Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/, 2019.

[17] Ribeiro, M. T., Singh, S., Guestrin, C., " why should I trust you?" : Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016. pp. 1135 − 1144.

销 售 团 队

职务	姓名	手机	邮箱
华北销售总监	王均丽	13910596682	wangjl@tpyzq.com
华北销售	成小勇	18519233712	chengxy@tpyzq.com
华北销售	孟超	13581759033	mengchao@tpyzq.com
华北销售	付禹璇	18515222902	fuyx@tpyzq.com
华北销售	韦珂嘉	13701050353	weikj@tpyzq.com
华东销售副总监	陈辉弥	13564966111	chenhm@tpyzq.com
华东销售	李洋洋	18616341722	liyangyang@tpyzq.com
华东销售	杨海萍	17717461796	yanghp@tpyzq.com
华东销售	梁金萍	15999569845	liangjp@tpyzq.com
华东销售	杨晶	18616086730	yangjinga@tpyzq.com
华东销售	秦娟娟	18717767929	qinjj@tpyzq.com
华东销售	王玉琪	17321189545	wangyq@tpyzq.com
华东销售	慈晓聪	18621268712	cixc@tpyzq.com
华南销售总监	张茜萍	13923766888	zhangqp@tpyzq.com
华南销售	查方龙	18520786811	zhafl@tpyzq.com
华南销售	胡博涵	18566223256	hubh@tpyzq.com
华南销售	张卓粤	13554982912	zhangzy@tpyzq.com
华南销售	张文婷	18820150251	zhangwt@tpyzq.com

太平洋证券
PACIFIC SECURITIES

研究院

中国北京 100044

北京市西城区北展北街九号

华远·企业号 D 座

电话： (8610)88321761

传真： (8610) 88321566

重要声明

太平洋证券股份有限公司具有证券投资咨询业务资格，经营证券业务许可证编号 13480000。

本报告信息均来源于公开资料，我公司对这些信息的准确性和完整性不作任何保证。负责准备本报告以及撰写本报告的所有研究分析师或工作人员在此保证，本研究报告中关于任何发行商或证券所发表的观点均如实反映分析人员的个人观点。报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价或询价。我公司及其雇员对使用本报告及其内容所引发的任何直接或间接损失概不负责。我公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。