金融工程

证券研究报告

2018 年 11 月 07 日

海外文献推荐 第 62 期

利用 CART 决策树选股

机器学习在金融领域有着非常广泛的应用，本文将 CART 决策树算法应用于选股模型之中。决策树模型相比于传统的线性模型或者判别分析其优势在于能解释模型中的非线性关系以及变量之间相互依赖的现象。本文以罗素 1000 指数中科技板块的选股为例，作者展示了 CART 决策树模型在于截面选股中的应用，动态 CART 决策树模型相比于简单的指标筛选方式表现出更高的多空收益以及夏普比率。

风险提示：本报告内容基于相关文献，不构成投资建议。

作者

吴先兴 分析师
SAC 执业证书编号：S1110516120001
wuxianxing@tfzq.com
18616029821

相关报告

1 《金融工程：金融工程-海外文献推荐 第 61 期》 2018-10-31

2 《金融工程：金融工程-海外文献推荐 第 60 期》 2018-10-24

3 《金融工程：金融工程-海外文献推荐 第 59 期》 2018-10-17

内容目录

图表目录

利用 CART 决策树选股

文献来源： Eric H. Sorensen, Keith L. Miller, and Chee K. Ooi，2000，The Decision Tree Approach to Stock Selection，The Journal of Portfolio Management，42-52

推荐原因：机器学习在金融领域有着非常广泛的应用，本文将 CART 决策树算法应用于选股模型之中。决策树模型相比于传统的线性模型或者判别分析其优势在于能解释模型中的非线性关系以及变量之间相互依赖的现象。本文以罗素 1000 指数中科技板块的选股为例，作者展示了 CART 决策树模型在于截面选股中的应用，动态 CART 决策树模型相比于简单的指标筛选方式表现出更高的多空收益以及夏普比率。

1.简介

量化投资的一种常见方式是将可投资的股票样本缩减为一组拥有特定特征的股票组合，投资经理通常用多重筛选的方式以达到其目的。虽然大多数投资经理不完全根据优化以及数学方法进行纯粹地量化选股，但是很多人都会借助数量化方式进行股票筛选。其中用以筛选的股票特征包含股票估值、盈利表现、流动性、动量以及投资风格等。

筛选的方式是有用的，然而它并不是一种完全科学的方式。举例而言，一些股票完全符合其它筛选特征但是它们会因为不满足某种筛选特征要求而被排除在组合之外。相反的，多变量打分的方式根据给不同的因子打分加权最终得到每只股票的排序，但是某些股票也可能因为一个指标具有非常高的权重而被纳入或者排除组合，而其它指标可能只被分配到很低的权重。因此多变量打分系统也并非完美。

在这篇文章中，作者将 CART 决策树算法应用于截面选股，并以此为基础构建了一个选股模型。相比较于传统的线性选股模型或者判别分析等，本文的 CART 决策树模型决定了筛选因子间的层次以及交互关系。估值因子是否要优先于动量因子使用，或者应该按照相反顺序进行？估值因子是如何与其它因子相互交互的？作者利用 CART 决策树在科技板块中选股并说明了这些深层次思考的重要性。

2.树和递归分类

CART 代表分类和回归树，这种统计方法是递归分类算法（RPA）的一种特定实现方式。顾名思义，分类技术将观测样本进行二分类或者多分类，最终的目的在于预测。举例而言，我们希望通过一辆车的外观预测它能跑多快，我们可以收集众多汽车的特征并按照它们的速度从高到低进行排序。这些数据中可能包括汽车的颜色、大小、轮胎的宽度等。首先，直觉上我们可能尝试根据汽车大小进行分类（小车通常比大车跑的快）；然而更好的模型可能先根据汽车轮胎的宽度进行分类，然后再考虑汽车的大小。拥有大轮胎的小汽车将被分类为速度最快的。

Breiman et al.在 1984 年提出了 CART 算法，最初的应用主要在于医药预测领域，其后该算法被应用于金融建模领域但一般用于解决时间序列问题。例如 Kao and Shumaker（1999）估计时间序列用以区分成长股和价值股的收益。

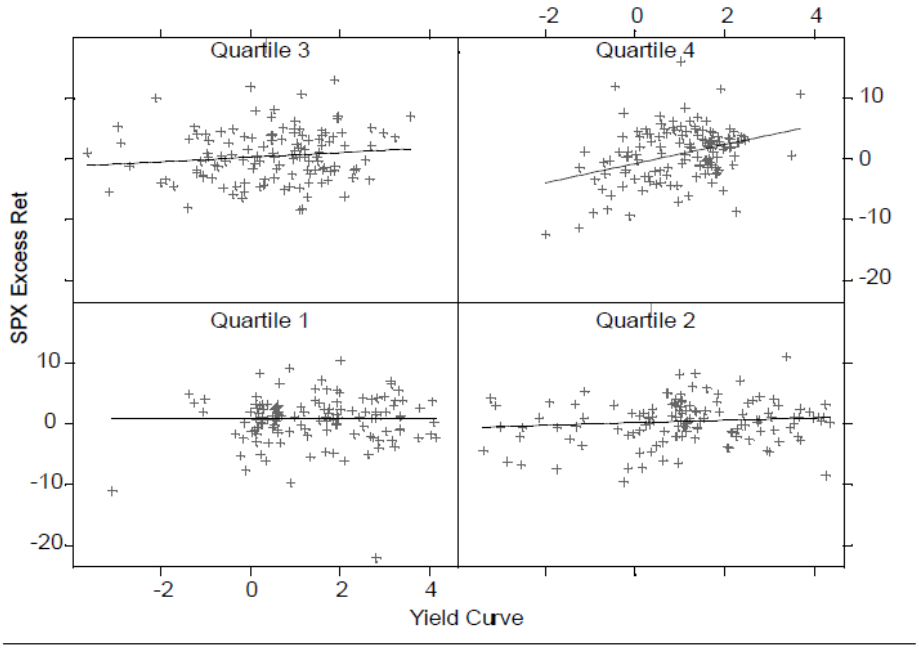
CART 决策树的优势在于它可以决定各因素之间的非线性层次最终最优化排序结构。层次关系通过分二分类树来估计，最终生成条件的组合用以减少数据的维度同时提高预测的准确性。更有效的，决策树通常根据一系列“如果-那么”的准则指引我们找到最优的决策，这是优于变量被允许承担更优先级的层次同时也被运行与其他变量交互，因此变量在不同的条件下可以用不同的影响。

CART 算法并非是一个黑箱，所有的输入变量和目标变量和我们在多元线性模型和判别分析中的变量是可比的。模型中变量的选择应该是有逻辑和理想的，这与传统的统计模型相一致。然而传统的线性模型中，最终方程要求所有的应变量独立、可加，并且在所有时期拥有一样的系数。

线性假设很明显是有局限的。举例而言，当首先以经济状态（以长期债券收益减去短期国债收益衡量）作为条件时，标普 500 指数的市场择时树表明股票市场的相对价值对于标普 500 指数未来的变现具有更显著的影响。一个首先考虑当前的宏观经济环境其次再考虑当前是相对价值的模型显著的更优于那些简单的将这两个变量看作独立的模型。

下图举例说明了一个简单的例子，作者在不同信用利差情境下展示了标普 500 指数收益率和债券收益率之间的线性关系。从下图可以很明显的看出，当信用利差最宽时（Quantile4，右上图），陡峭的债券收益率曲线一般会导致标普 500 指数更好的表现。在信用利差的其他情形下，二者之间的关系都是不明显的。分情境考虑下，我们发现了一个很容易被线性模型忽略的隐藏关系。

图 1：不同信用利差下标普 500 收益率-债券收益率关系



资料来源：The Journal of Portfolio Management，天风证券研究所

3.将数据分类

在估计的时候，使用离散的分类代表自变量和因变量是一种有效的初始步骤。分类树模型的输出是一个二分类树，它赋予因变量分类以不同的概率，因变量可以按照 10 分位或者 5 分位来取值。例如，我们需要为小市值溢价显著的情景建模，那么数据应该被按照市值溢价情形分位 3 组：1）表现相似；2）大市值表现好；3）小市值表现好。类似的方法在描述自变量上也很有用，例如可以按照市场情绪分成 3 种区间：1）高波动；2）正常波动；3）低波动。

在决定树的结构，CART 决策树利用数学算法决定变量和相应的用于分类的阈值大小。变量-阈值的选取将样本分成最同质的两组，这个决定了树的顶级层次，并给我们按照准则将样本分位两组的结果。

例如如果市场波动率是最重要的输入变量，我们首先需要找出波动率的阈值能最好的解释大市值和小市值股票的收益率差价。一旦在树的顶层做出第一个分割，后续的递归分割将保持树的高阶结构，同时提高分类的效果。

4.CART 算法简述

CART 决策树每次分割通过（变量-阈值）来决定分割的方式。顶层节点分割后，在后续的子节点继续通过选取（变量-阈值）来决定后续最优的分割。假设存在两个输入变量，则 CART 决策树算法可以描述如下。若在样本中存在 N 个观测，$U:(X_1,c_1),(X_2,c_2),\dots,(X_N,c_N)$，其中$X=(x_1,x_2)$，c 是目标变量，满足$c\in(X,O)$。定义函数 d()描述数据的无序程度（熵），其满足：

$$d(S)=\sum_{C}-\frac{N_C}{N}log_2\frac{N_C}{N}$$

其中 S 为观测集，N_C是类别 c 的样本数目，N 是样本总数。

对于每个分割，样本集将被分为两类，则分割后的数据平均熵为$d'(U)$满足：

$$d'(U)=\frac{N_{left}}{N}\times d(S_{left})+\frac{N_{right}}{N}\times d(S_{right})$$

其中N_{left}、N_{right}分别为左右字数的样本数。

5.CART 决策树在截面数据中应用

诸如 CART 算法等递归分类算法的效果来自于：

1. 树结构层次的直觉
2. 解释了数据的非线性
3. 解释了变量之间相互依赖的关系
4. 为结果给出条件概率产出。

CART 算法非常适合于诸如选股等截面问题的解决。将股票数据按照时间分成多个区间，那么我们估计的模型就是截面的。作者描述了一个树结构用以在科技板块中 赢者股票组合和败者股票组合。使用 1992 年以来罗素 1000 指数的数据，作者计算科技板块股票的收益率，其中每个时期科技板块的股票数量在 70 至 110 只之间。利用这些股票的月度收益率可以得到代表股票相对表现的因变量。而最终目的在于构建一个稳定的模型用以区分赢者和败者，其中这些用以区分的自变量来自于合理的股票或上市公司特征。

6.板块共性的探寻

将科技股作为一个整体分析的原因是什么？显然将表现出共性的股票分位一组将提高我们发现显著关系的能力，因子对于股票收益率的解释能力在不同板块之间存在差异，风格表现的分析表明不同组别股票之间相对收益的驱动因素存在显著的区别。例如盈余动量相比估值在科技板块中对股票表现具有更强的预测能力，而相反的估值指标在金融股中具有更重要的预测能力。

一种分组方式是按照收益率的相关性将股票分组，我们可以按照股票的历史收益率进行聚类；而另一种分组方式这是主观的主题分类，利润标普或罗素的行业板块分类。本文作者按照罗素科技板块提取科技股，其他的行业板块还有医疗健康、可选消费、必要消费、金融服务于、石油等等。

7.输入变量

首先本文计算了从 1992 年至 1997 年所有股票的月度收益率，其次将股票的收益率减

去该时期所有股票收益率的中位数以得到超额收益率，这使得在每个月对每只股票可以分类为高于平均水平和低于平均水平两组。

目标很简单，将表现高于平均水平的股票与低于平均水平的股票区分开，也就是说因变量是将样本等分的二元变量。每个观测对应于特定股票在特定月份的收益率，因此一半的样本被标记为高于平均水平，而另一半被标记为低于平均水平。

表 1：变量描述

变量	定义
SALES-PRICE	市销率倒数
CFLOW-PRICE	市现率倒数
EPS-PRICE	分析师一致预期 EPS
ROA%	净资产收益率同比变化
EPS-MOM	分析师一致预期 EPS 变化
PRICE MOM	过去一个月股票收益率

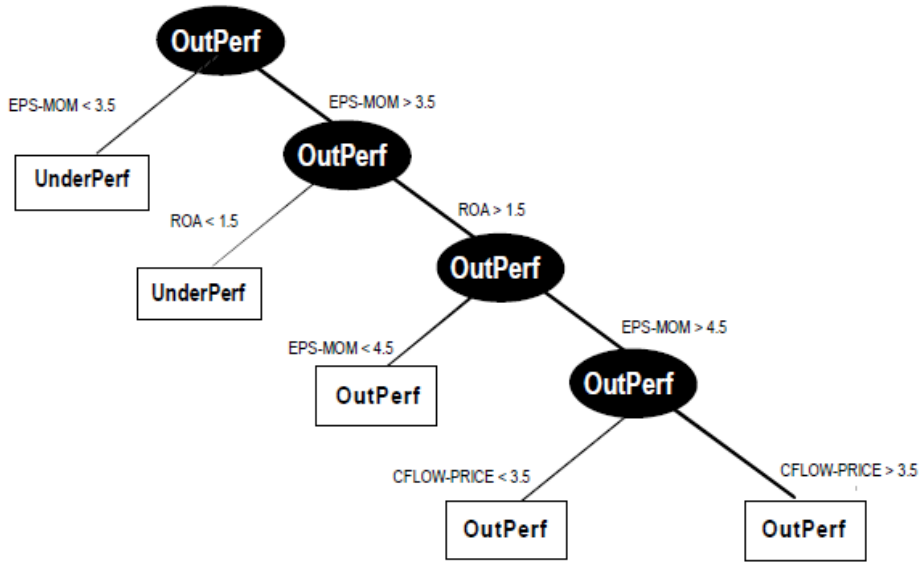
资料来源：The Journal of Portfolio Management，天风证券研究所

本文选取了一个自变量的小集合，这些变量来自于投资经理们所常用的因子集合包括估值、盈利、分析师预期、价格动量等。所选取的 6 个变量均对于股票收益率具有一定的解释能力。其中关键的考虑是每个因子的说明，在计算每个指标以及动量因子后，每个变量在每个月份被分为 5 组以提供更为稳定的估计结构。因为如果使用更精细的数据，例如连续变量，最终的树结构可能产生过拟合现象。过拟合的结果是虚假的，一方面它对于历史表现出更强的解释能力，另一方面由于其过度拟合到处缺失预测能力。此外过拟合也将导致对于树结构最终的逻辑解释性。

8.科技股选股模型：静态树

本文估计科技板块选股模型的第一种方法假设股票相对表现与输入变量之间的交互作用稳定的简化版本。在这个模型中作者水世界分为两组：1）1993 年至 1995 年；2）1996 年至 1999 年。第一组作为样本内的训练集用以估计模型，第二组作为样本外的测试集用以检验模型样本为表现。样本外的检验为模型最终的预测能力给出评估，同时由于整个样本中只有一个树模型，作者将这个模型定义为静态树方法。

图 2：静态树模型



资料来源：The Journal of Portfolio Management，天风证券研究所

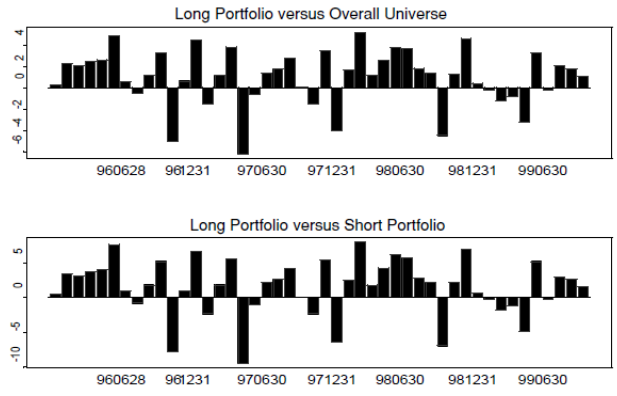
由于本文将每个变量分为 5 组，因此可能的分割有 4 种：1）在组 1 和组 2 间分割；2）在组 2 和组 3 间分割；3）在组 3 和组 4 间分割；4）在组 4 和组 5 间分割。如上图所示，在树顶端第一个变量是 EPS-MOM，模型首先将样本按照 EPS-MOM 分为两组：分析师盈利估计修正最高的 2 组 vs 分析师盈利估计修正最低的 3 组。

递归分类算法将在此分类的基础上在树的右支进行第二层次的分类。在树的右边，RPA 算法根据 ROA 动量指标以 1.5 作为阈值继续对样本进行分类。如果股票在盈利修正指标位于最高的两组，那么其可能在下个月有更好的表现。更进一步，如果股票在盈利修正指标位于最高的两组，同时其在 ROA 动量上位于最高的 4 组，那么其将更高的概率在下个月表现超过平均水平。

以上分析在逻辑上令人满意的结果，因为市场在分析师乐观预期的股票中分辨出那些有基本面改善的公司（ROA 增速高）。CART 决策树确认了我们的直觉，并且其相对于简单的线性筛选具有更丰富的特征。最终生成的树模型很简单，每只股票在每个月都存在和树节点相对于的特征，每只股票按照节点特征分类进入下一个层次的分类。

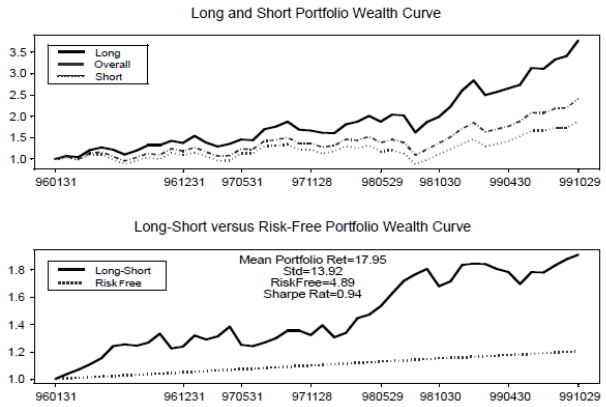
每个月将预测表现高于平均水平的记为多头组合并等权持有组合，将预测表现低于平均水平的记为空头组合并等权持有组合。下左图为多头组合以及多空组合的月度超额收益率，右图为多头、空头组合以及全样本的净值表现。可以发现多头组合能显著的战胜空头组合，多空组合年化收益达到 13.92%，T 检验和 WILCOXON 秩检验结果均表明多空超额收益在统计意义上显著。多空组合平均每月超额空头组合 1.40%，并且收益率差在 5%显著性水平上区别于 0。

图 3：静态模型月度收益



资料来源：The Journal of Portfolio Management，天风证券研究所

图 4：静态模型净值



资料来源：The Journal of Portfolio Management，天风证券研究所

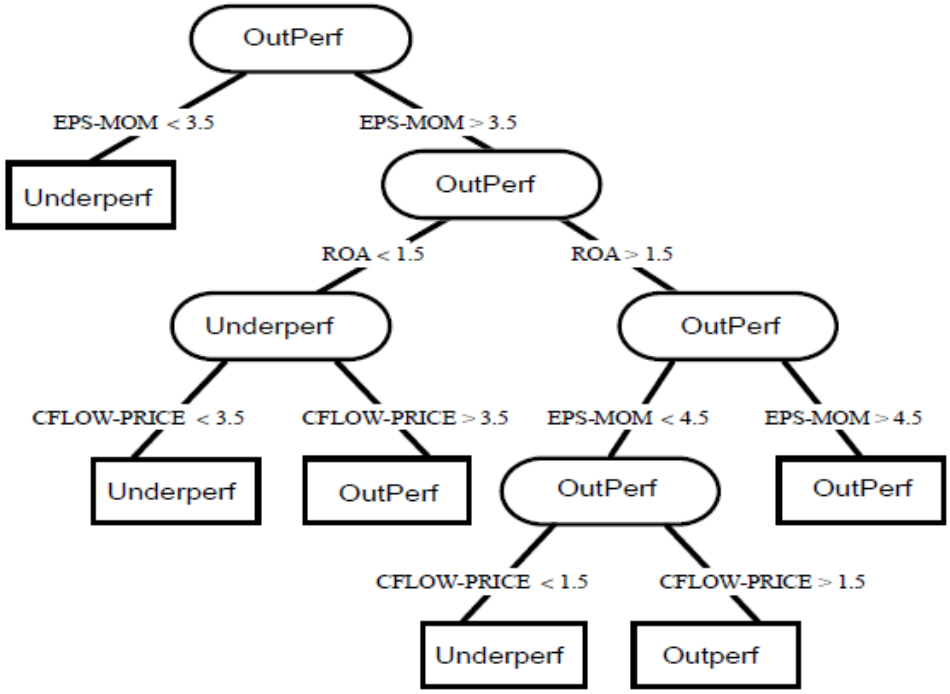
9.科技股选股模型：进化树

以上描述的静态树模型论证了树结构的稳定性，根据 1992 年至 1995 年估计的树结构对于随后的 1996 年至 1999 年仍然具有预测能力。相比于静态模型，作者发现在每个月重新估计树结构可以提供更高的样本外预测能力。

本文的第二种方法是在每个月利用之前的所有信息重新估计树结构。1995 年 12 月从 1993 年 1 月至 1995 年 12 月的数据被用以训练树结构，其后将树模型用于 1996 年 1 月的股票分类。对于每个随后来临的月份，最新的样本将加入到训练集中以估计树结构，最终的预测将根据最新的树模型得到。

显然这种动态方式在每个月都能得到不同的树模型，由于训练集中仅增加了最新的样本因此短期内树的结构是稳定的，在月与月之间仅有较为微小的变化；但是长期来看模型的树结构却有着显著的变化，1999 年 6 月的树结构肯定相比 1996 年发生了重大的变化。因此，作者将这种技术称为“进化树”。

图 5：动态树结构（1999 年 10 月）

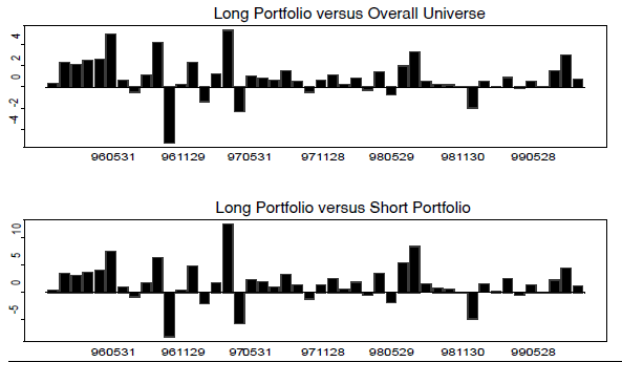


资料来源：The Journal of Portfolio Management，天风证券研究所

进化树方法有很多优点。首先，树的估计流程需要大量的数据以保证统计显著性，在进化树方式中更多的数据将被用于树结构估计。其次，逻辑上模型不断的进化方式也更有意义，它允许模型有渐进的改变并将市场和企业的变化引导进模型中。

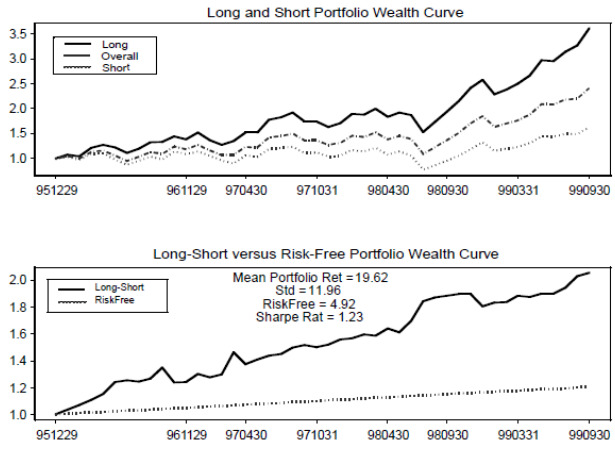
下图展示了动态模型的样本外绩效表现，多头组合相比于空头组合具有显著超额收益，T 检验 T 值为 3.25，WILCOXON 秩检验 Z 值为 4.10，p 均小于 0.01，组合月度多空收益为 1.47%，显著区别于 0。

图 6：动态模型月度收益



资料来源：The Journal of Portfolio Management，天风证券研究所

图 7：动态模型净值表现



资料来源：The Journal of Portfolio Management，天风证券研究所

10.绩效评估

决策树模型相比于简单的股票筛选或者排序表现如何？为了回答这个问题作者对比了多种筛选策略并评估了它们的绩效表现，下图中展示了不同模型的绩效表现。作者利用 EPS-MOM，ROA 以及 CFLOW-PRICE 构建了 3 个单因子模型，同时利用这三个指标的均值构建了一个多因子模型，作者对比本文的两个树模型与这几个模型的绩效。

由下图结果可知，两个 CART 决策树模型的夏普比率均显著高于单因子的排序模型，除了 EPS-MOM 指标外，其余利用单个指标筛选出的多头组合超额收益均不显著。在所有的模型中，进化决策树有最高的夏普比率以及 t 统计量。

图 8：不同模型继续对比

	Outperform Portfolio Excess Returns		Underperform Portfolio Excess Returns		Mean Difference of Excess Returns		Long-Short Strategy	
Model	Mean	T-Stat	Mean	T-Stat	Mean	T-Stat	Ann. Ret	Sharpe
Static Tree Model	0.91	2.49	-0.48	1.69	1.40	2.94	17.95	0.94
Evolving Tree Model	0.75	2.39	-0.72	2.21	1.47	3.25	19.22	1.15
EPS-MOM	0.62	2.01	-0.63	1.89	1.25	2.76	14.62	0.60
ROA	0.12	0.38	-0.12	0.37	0.24	0.53	1.85	0.27
CFLOW-PRICE	0.18	0.62	-0.19	0.55	0.37	0.82	6.00	0.07
Mean of EPS-MOM, ROA, and CFLOW-PRICE	0.28	0.93	-0.32	0.92	0.60	1.31	7.65	0.23

资料来源：The Journal of Portfolio Management，天风证券研究所

分析师声明

本报告署名分析师在此声明：我们具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，本报告所表述的所有观点均准确地反映了我们对标的证券和发行人的个人看法。我们所得报酬的任何部分不曾与，不与，也将不会与本报告中的具体投资建议或观点有直接或间接联系。

一般声明

特别声明

投资评级声明

类别	说明	评级	体系
股票投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	买入	预期股价相对收益 20%以上
		增持	预期股价相对收益 10%-20%
		持有	预期股价相对收益-10%-10%
		卖出	预期股价相对收益-10%以下
行业投资评级	自报告日后的 6 个月内，相对同期沪深 300 指数的涨跌幅	强于大市	预期行业指数涨幅 5%以上
		中性	预期行业指数涨幅-5%-5%
		弱于大市	预期行业指数涨幅-5%以下

天风证券研究

北京	武汉	上海	深圳
北京市西城区佟麟阁路 36 号	湖北武汉市武昌区中南路 99 号保利广场 A 座 37 楼	上海市浦东新区兰花路 333 号 333 世纪大厦 20 楼	深圳市福田区益田路 5033 号平安金融中心 71 楼
邮编：100031	邮编：430071	邮编：201204	邮编：518000
邮箱：research@tfzq.com	电话：(8627)-87618889	电话：(8621)-68815388	电话：(86755)-23915663
	传真：(8627)-87618863	传真：(8621)-68812910	传真：(86755)-82571995
	邮箱：research@tfzq.com	邮箱：research@tfzq.com	邮箱：research@tfzq.com