金融工程

证券研究报告
2020 年 11 月 25 日

用树模型提取分析师预期数据中的非线性 alpha 信息

作者

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

祗飞跃 联系人
difeiyue@tfzq.com

相关报告

1 《金融工程：金融工程-市场情绪一览 2020-11-24》 2020-11-24

2 《金融工程：金融工程-指数样本股交叉调整背后的综合冲击效应》 2020-11-24

3 《金融工程：金融工程-市场情绪一览 2020-11-23》 2020-11-23

研究目的

用分析师预期数据设计的因子，结构复杂、数据缺失多、与盈利因子、市值类因子相关性高，线性模型未必能够充分提炼其中独有的 alpha 信息。

提升树模型是一种被广泛使用的机器学习方法，模型可以拟合非线性关系，可以自动处理数据缺失问题，使用方法灵活。

报告将提升树模型应用于对分析师预期数据因子和股票收益率建模之中，在确保与盈利类因子、市值类因子低相关的前提下，尝试提取因子中或有的非线性 alpha 信息。

测试方法

报告采用滚动建模的方法，同时使用线性模型和提升树模型，用分析师预期数据因子构建股票收益预测模型。在模型构建过程中，采用统计技术手段，降低模型预测值与盈利类因子和市值类因子之间的相关性，提高预测结果的独立性。

另外，为验证预测结果的有效性，报告用盈利类因子和市值类因子构建了基础股票收益率预测模型，分析预测结果对基础模型的增量效果。

测试结果

测试的数据周期为 2007 年初至 2020 年 10 月。通过测试发现，提升树模型可稳定有效地预测股票收益率，预测值与基础模型预测值相关性均值为 10%。另外，提升树模型与线性模型预测值相关性均值为 50%，提升树模型预测值更适合中小市值股票、而线性模型更适合大市值股票。等权使用两个模型构建的股票组合，其收益表现显著优于单独使用线性模型构建的股票组合，增量收益部分独立，且增量收益部分自 2016 年以来大幅提升。

另外，报告测试了等权模型对基础模型的增量效果。结果显示，在沪深 300 指数增强和中证 500 指数增强策略中，增量收益的年化收益率分别为 1.7%和 2.8%、夏普率分别为 1.65 和 2.20，增量收益与原策略收益相关性分别为-7%和 16%，增量部分独立、稳定、有效。

其他内容

报告构建了完整的滚动训练股票收益率预测模型的流程，介绍和尝试了几种去除相关性的技术方法，丰富了一些数据建模的技术细节，分析了预测模型的风险来源和评估、调整方法。其中提出的一些技术方法和评价指标，适用于一般性的股票收益率预测模型。

风险提示：模型基于历史数据，模型失效风险，因子失效风险，市场环境变动风险

内容目录

图表目录

1. 引言

分析师预期数据是一类结构化数据，具有如下特点：

- 数据零散、结构复杂、对股票的覆盖不全；
- 因子设计方法多、因子之间相关性高，alpha 有非线性特征，且特征难以参数化；
- 和财务数据、市值等因素有一定相关性。

很多研究在多因子框架下对分析师预期数据的预测性进行了探索，证明了数据中有独立的 alpha 信息。但用线性模型研究分析师预期数据时面临如下挑战：

- 线性模型能否充分提炼 alpha 信息；
- 利用分析师预期数据构建的因子与其他 alpha 因子相关性高，如何提取其中的独有 alpha。

我们尝试使用提升树模型、用分析师预期数据因子对股票收益率构建预测模型，提取数据中的 alpha 信息；同时采用统计技术手段、确保预测结果与财务、市值等因子保持低相关性。与此同时，我们用线性模型作为对比。经过测试发现：

1. 提升树模型可以提取数据中的 alpha 信息，用模型预测值构建的策略在全时段样本中有效、稳定；
2. 提升树模型的预测值和线性模型预测值历史相关性均值为 50%。两个模型呈现出不同的特点，相比较而言，线性模型更适合大市值股票，提升树模型则更均衡。
3. 等权使用提升树模型和线性模型，比单独使用线性模型效果更好，对多种组合策略的年化收益率和夏普率均有一定提升。尤其 2016 年以来，加入提升树模型的预测结果明显好于单独使用线性模型；
4. 去相关性的统计技术有效，线性模型和提升树模型与基础模型的预测值之间保持了低相关性，平均相关性低于 10%；
5. 等权分析师数据模型对财务类模型在不同组合策略上均有增量效果，增量部分与原策略收益时序相关性低。

除此之外，报告还做了如下几方面工作：

1. 完善了用滚动方法构建收益率预测模型的流程；
2. 丰富了用提升树模型进行建模的技术细节；
3. 提出了几种实用的、用于降低模型预测值与已有因子之间相关性的方法；
4. 分析了预测模型的风险来源，提出了对应的评估和调整方法。

2. 因子介绍及结构分析

在本节中，我们介绍模型中所使用的分析师预期数据因子，并分析这些因子的统计特征以及它们与某些财务因子和风格因子之间的相关性。

2.1. 设计和分类

我们从分析师预测利润收入的一致预期、分析师对股票评级、分析师预期目标价和分析师报告标题关键字四个角度来设计因子，并根据因子的数值类型将它们分为连续值因子和计数因子两类。

2.1.1. 一致预期类

我们把分析师对上市公司的净利润或营业收入的一致预期，定义为过去 90 个自然日内，所有分析师对该数据最新预测的平均值。计算季度预测值时，用年度预测值减掉已公布值（财报、快报、预告），除以剩余年内季度数。

表 1：一致预期类指标

英文简称	名称	数值类型
consroeoverntc_ew_0	季度一致预期 ROE	连续型
substract[consroeoverntc_ew_0,ntcroe_q_3]	季度一致预期 ROE 与上年同期真实 ROE 之差	连续型
substract[consroeoverntc_ew_0,consroeoverntc_ew_30]	30 天内季度一致预期 ROE 的变动	连续型
constooverstm_ew_0	季度一致预期周转率	连续型
substract[constooverstm_ew_0,assetturnover_q_3]	季度一致预期周转率与上年同期真实周转率之差	连续型
substract[constooverstm_ew_0,constooverstm_ew_30]	30 天内季度一致预期周转率的变动	连续型
consroedifoverntc_ew	下一年同季度一致预期 ROE 和当年季度一致预期 ROE 之差	连续型
conscounting_upadj_30	过去 30 天净利润预测上调家数	计数型
conscounting_upadj_60	过去 60 天净利润预测上调家数	计数型
conscounting_upadj_90	过去 90 天净利润预测上调家数	计数型
conscounting_cvrg_30	过去 30 天净利润预测覆盖家数	计数型
conscounting_cvrg_60	过去 60 天净利润预测覆盖家数	计数型
conscounting_cvrg_90	过去 90 天净利润预测覆盖家数	计数型
divide[conscounting_upadj_90,conscounting_cvrg_90]	过去 90 天上调家数和覆盖家数之比	连续型

资料来源：天风证券研究所

2.1.2. 评级类

我们对不同证券公司评级进行标准化处理，将评级分为五档（也可采用朝阳永续的评级标准化方法）。在此基础上我们设计如下因子：

表 2：评级类指标

英文简称	名称	数值类型
avgrating_ew_30	过去 30 天评级的平均值	连续型
avgrating_ew_60	过去 60 天评级的平均值	连续型
avgrating_ew_90	过去 90 天评级的平均值	连续型
ratecounting_add_30	过去 30 天买入和增持评级家数	计数型
ratecounting_add_60	过去 60 天买入和增持评级家数	计数型
ratecounting_add_90	过去 90 天买入和增持评级家数	计数型
ratecounting_buy_30	过去 30 天买入评级家数	计数型
ratecounting_buy_60	过去 60 天买入评级家数	计数型
ratecounting_buy_90	过去 90 天买入评级家数	计数型
upadjcounting_30	过去 30 天上调评级家数	计数型
upadjcounting_60	过去 60 天上调评级家数	计数型
upadjcounting_90	过去 90 天上调评级家数	计数型

资料来源：天风证券研究所

2.1.3. 目标价类

表 3：目标价类指标

英文简称	名称	数值类型
upadjestprccnt_60	过去 60 天股价上调家数	计数型
upadjestprccnt_90	过去 90 天股价上调家数	计数型

资料来源：天风证券研究所

2.1.4. 报告标题超预期类

我们根据分析师报告的标题，通过过滤关键字的方法，对于有上市公司业绩评价的报告标题，定义了超预期、一般、低于预期三种评级，以此构建如下因子：

表 4：报告标题超预期类指标

英文简称	名称	数值类型
avgexpass_ew_60	过去 60 天分析师标题业绩预期平均分	连续型
avgexpass_ew_90	过去 90 天分析师标题业绩预期平均分	连续型
expasscounting_excess_60	过去 60 天分析师在公司业绩发布 7 天内认为超预期的个数	计数类
expasscounting_excess_90	过去 90 天分析师在公司业绩发布 7 天内认为超预期的个数	计数类
expasscounting_allex_60	过去 60 天分析师认为业绩超预期的个数	计数类
expasscounting_allex_90	过去 90 天分析师认为业绩超预期的个数	计数类

资料来源：天风证券研究所

2.1.5. 常见但未使用的因子

有两类因子没有纳入到建模中，分别为：

1. 用分析师目标价构建的预期收益类因子。

此类因子与反转因子有相关性，不属于慢因子。

2. 分析师预期数据构建的估值类因子。

相关估值因子中主要 alpha 信息为一致预期数据，已经被第一类因子包含。

有兴趣的读者可以将以上因子加入模型中，以本文所介绍的流程进行测试。

2.2. 结构分析

我们从因子覆盖率、分布特征、与其他类别因子相关性及平稳性四个维度，对分析师类因子的结构进行分析。统计周期为 2007 年至 2020 年。

2.2.1. 覆盖率

我们统计了以上因子的覆盖率，其中，对于连续型因子我们以空值为数据缺失，对于计数型因子我们以 0 值为数据缺失。

表 5：历史平均覆盖率统计

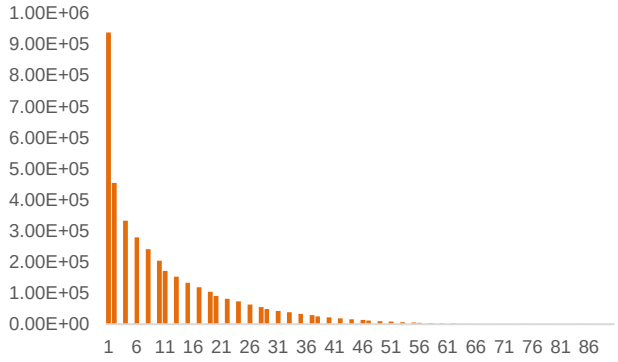
	因子平均覆盖度	因子最低覆盖度	因子最高覆盖度
一致预期财务指标	62%	36%	75%
分析师评级	52%	26%	72%
分析师目标价	21%	19%	23%
分析师报告标题	13%	6%	28%

资料来源：朝阳永续，wind，天风证券研究所

2.2.2. 统计分布

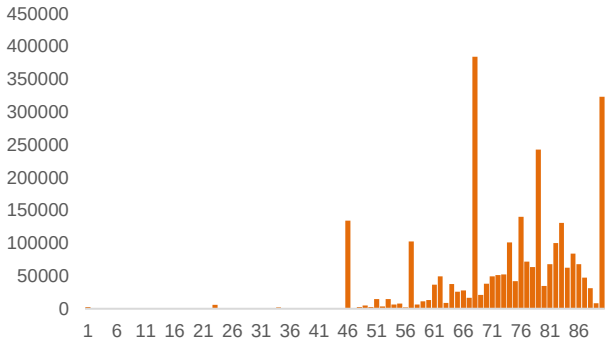
原始因子的分布并不规范，例如下面两个因子：

图 1：conscounting_cvrg_90 直方图



资料来源：朝阳永续，wind，天风证券研究所

图 2：avgrating_ew_90 直方图



资料来源：朝阳永续，wind，天风证券研究所

2.2.3. 与其他类别因子相关性

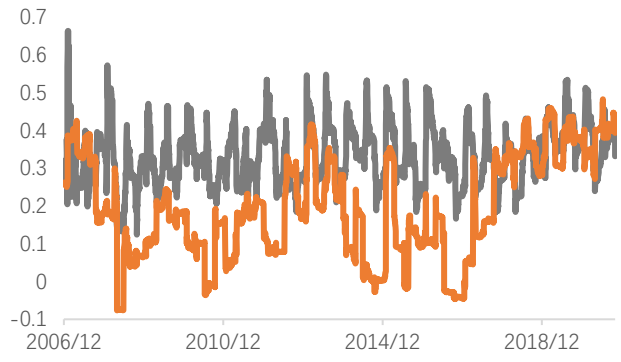
分析师倾向于覆盖、推荐基本面好的股票，也倾向于覆盖市值大的股票，这使得分析师数据因子与 ROE、ROE 同比、市值因子和中性化后市值平方因子有一定的相关性。我们用去掉缺失值的 Pearson 相关系数来描述相关性，展示了如下例子：

1. 因子 avgexpass_ew_90 和因子 substract[ntcroe_q_0,consroeonntc_ew_0]与中性化后的单季 ROE 同比因子的相关性；平均相关性分别为 33%和 20%。

2. 因子 substract[ntcroe_q_0,consroeonntc_ew_0]和因子 consroeoverntc_ew_0 与中性化后的单季 ROE 因子的相关性；平均相关性分别为 25%和 16%。

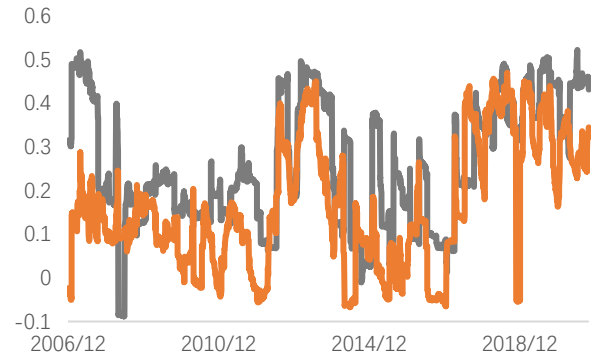
3. 因子 ratecounting_add_60 和因子 conscounting_upadj_60 与市值因子的相关性；

图 3：与单季 ROE 同比因子的相关性



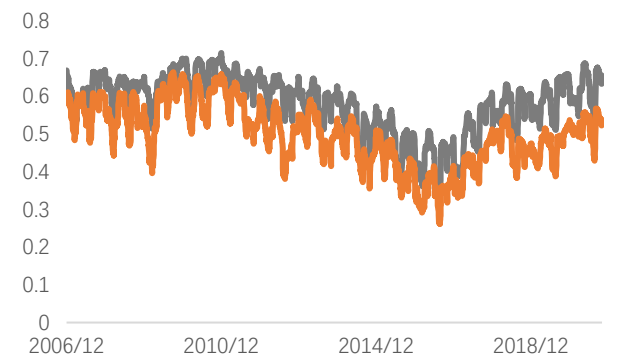
资料来源：朝阳永续，wind，天风证券研究所

图 4：与单季 ROE 因子的相关性



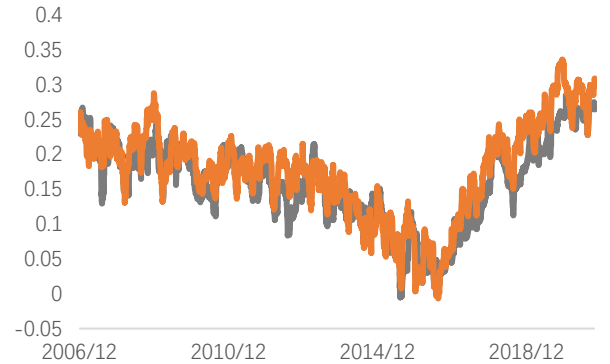
资料来源：朝阳永续，wind，天风证券研究所

图 5：与市值因子的相关性



资料来源：朝阳永续，wind，天风证券研究所

图 6：与中性化后市值平方因子相关性



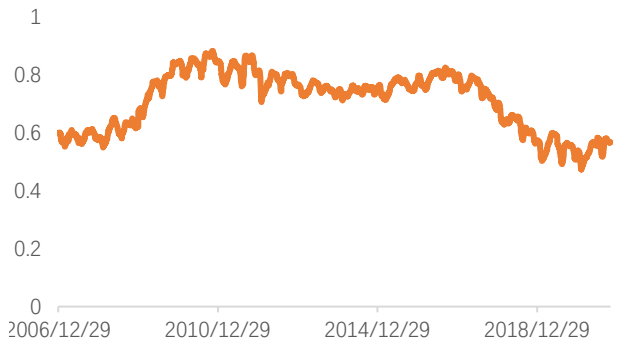
资料来源：朝阳永续，wind，天风证券研究所

4. 因子conscounting_upadj_90和因子ratecounting_add_60与中性化后的市值平方因子的相关性；平均相关性为16%和18%。

2.2.4. 时序平稳性

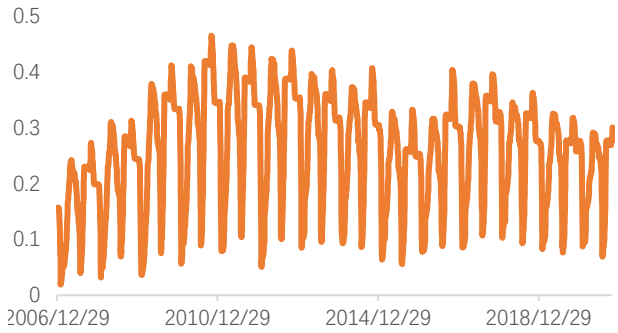
我们从覆盖率、因子的25%分位值和因子的75%分位值三个变量随时间序列的变动情况，来举例展示分析师数据类因子的时序是否平稳：

图7：consroeoverntc_ew_0的历史覆盖率



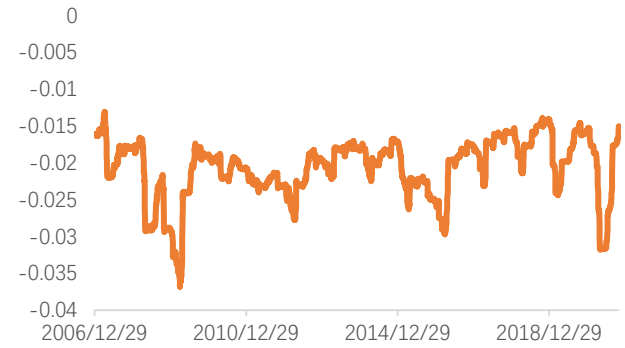
资料来源：朝阳永续，wind，天风证券研究所

图8：avgexpass_ew_90的历史覆盖率



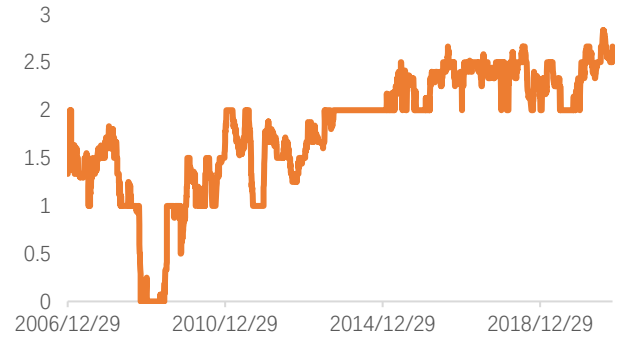
资料来源：朝阳永续，wind，天风证券研究所

图9：substract[ntcroe_q_0,consroeonntc_ew_0]的25分位值



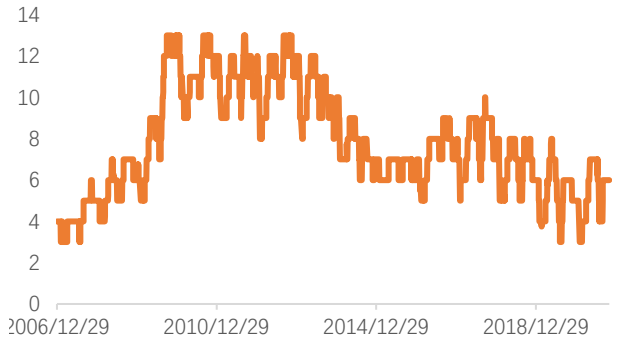
资料来源：朝阳永续，wind，天风证券研究所

图10：avgrating_ew_60的25分位值



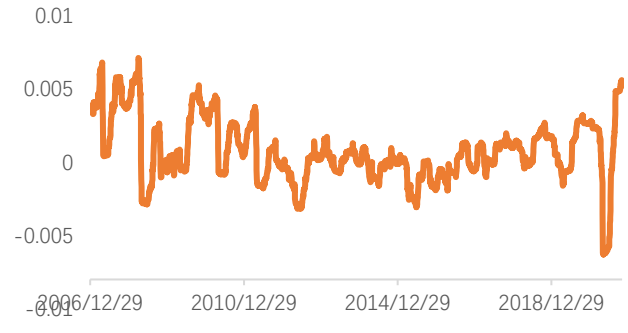
资料来源：朝阳永续，wind，天风证券研究所

图11：conscounting_cvrg_90的75分位值



资料来源：朝阳永续，wind，天风证券研究所

图12：substract[ntcroe_q_0,consroeonntc_ew_0]的75分位值



资料来源：朝阳永续，wind，天风证券研究所

2.2.5. 结论

从以上结构分析中我们可以看出，分析师预期数据因子的结构很差：

首先，它们的覆盖率低、数据缺失严重；

其次，因子分布不规范；

再次，与一些常用的因子有较高的相关性；

最后，时序上它们不平稳。

因此在建模的时候，我们要根据模型的特性，对因子进行相应的处理。

3. 建模目标和分析流程

本报告采用滚动构建股票收益率预测模型的方式来提取数据中的 alpha 信息，这与多因子方法不同之处在于，多因子方法把因子直接当做 alpha 使用，存在逻辑单一、线性假设过强的问题；用统计建模的方法，多了一层 alpha 提炼的过程。这样做的好处在于统计方法更为丰富、可以从金融逻辑和数据两个方向相向而行寻找 alpha；其次，大部分统计模型都能处理非线性关系；最后，以股票收益率作为因变量，统计模型的预测结果是预期收益率，在不考虑相关性的前提下，模型预测值之间可加。

在构建模型时我们使用技术手段，确保预测结果与财务因子和市值类因子保持低相关性，这样做的目的在于：

1. alpha 之间相关性是造成股票预测体系不稳定的因素之一，不在预测阶段解决相关性问题，就必须在整合阶段解决相关性问题；

2. 新 alpha 和老 alpha 之间有相关性，会给投资管理造成困难。分析师预测数据和财务数据来源不同，这两类 alpha 更适合分开管理。

值得指出的是，解决 alpha 之间相关性是非常难的统计问题，因此建议设计 alpha 模型的时候要考虑与已有 alpha 模型之间的相关性，尽量把处理 alpha 之间相关性的问题前置。遗憾的是，现有的一些报告并没有遵循这种做法，很多报告使用分析师预测数据和财务数据共同填充的方法来设计因子，这样会在分析师预测数据 alpha 和财务数据 alpha 之间引入相关性，为之后的因子整合带来隐患。

建模和分析大致步骤如下：

1. 构建基础模型作为“已有的”alpha 模型：

用中性化后的季度 ROE、中性化后的季度 ROE 同比以及中性化后的市值平方因子构建线性模型，作为考察分析师数据模型增量情况的基础模型，我们称以上三因子为基础因子。

2. 滚动构建目标模型

用分析师预期数据因子构建线性模型和提升树模型，评估和分析两种模型的历史表现及与基础模型之间的相关性。

3. 分析两种目标模型之间的相关关系，分析它们的预测特点，分析等权目标模型的表现。

4. 分析目标模型对基础模型的增强效果。

4. 基础模型介绍

本节我们简要介绍基础模型的构建方法和表现

4.1. 构建方法

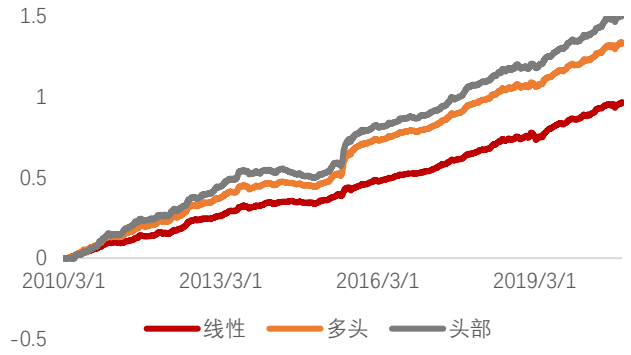
我们采用滚动建模的方式、用 IC-IR 方法对基础因子分配权重，并用加权后的因子值对因变量 y 进行一元线性回归，以此来构建基础模型。同时，我们用因子值对股票的日风险调整后的收益进行一元线性回归，用回归模型预测股票的日预期收益。

基础模型的建模方法与线性目标模型的建模方法相同，可参见 5.4.2，滚动回归、因变量、日预期收益的含义参见 5.2.3 和 5.3。

4.2. 模型表现

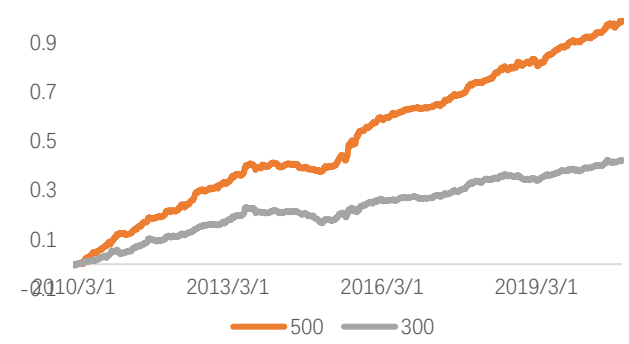
我们采用第 9 节中定义的模拟组合策略，来评估基础模型表现：

图 13：基础模型策略表现



资料来源：朝阳永续，wind，天风证券研究所

图 14：基础模型指数增强表现



资料来源：朝阳永续，wind，天风证券研究所

从上面的图中可以看出，基础模型是有效的。

5. 目标模型构建

本节我们介绍目标模型构建方法和部分细节。

5.1. 训练流程简介

模型训练周期为 2010 年 1 月至 2020 年 10 月。

5.1.1. 股票池

我们按照如下规则设定股票池：

1. 股票池更新日：沪深 300 指数成分股调整的下一个月份的第一个交易日
2. 入选规则：
   - 非 ST 股票；
   - 且上市日期超过 500 个交易日；
   - 且过去一年交易日超过 150 天；
   - 且过去 100 个交易日的股票收盘价最大值大于 3、最小值大于 2；
   - 且过去一年平均市值排名前 80%；
   - 且过去一年成交金额排名前 80%；
   - 或为最近一期沪深 300 指数、中证 500 指数或创业板指的成分股

5.1.2. 数据采样方法

我们在日级别采样。由于因变量的跨度周期为 20 个交易日，数据样本会跟前后的部分数据样本有信息上的重叠。日级别采样的好处是可以规避日历效应（Calendar Effect），但由于数据重叠，在进行假设检验时，例如 t-检验的时候，要调整 p 值计算方法。

对于任意给定交易日 t，我们定义 t 日的自变量 x，为根据 t 日早上开盘前的分析师预期数据所计算的因子值；而在计算 t 日所对应的因变量或股票收益时，以 t 日 vwap 价为买进或做空价格。这样的设计方法从时间的角度，确保模型的预测值是可交易的，同时也保证模型尽量使用了最新的因子信息——例如与财务有关的隔夜信息。

5.1.3. 滚动训练

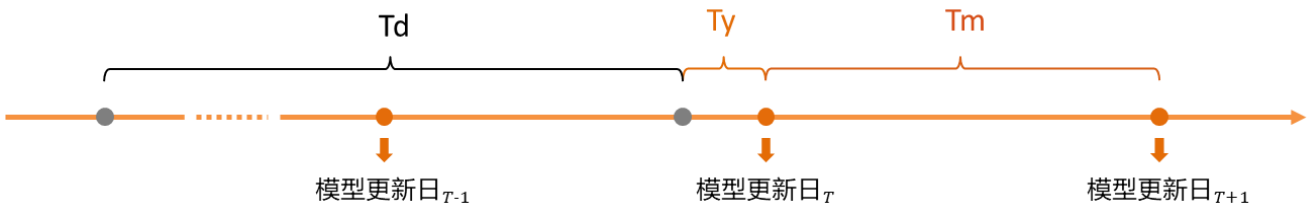
我们定期更新模型，使用模型直到下一个更新日。这里有三个与时间周期有关的参数，

分别为：

1. 投资周期 Ty。设定为 20 天，这是模型因变量的时间跨度长度（参见 5.2.3）；
2. 模型更新周期 Tm。设定为半年，以每年的 5 月底和 11 月底作为模型重新训练的日期。我们称参数周期 Tm 所覆盖的数据，为最近一个更新过的模型的测试集或应用集；
3. 数据采样周期 Td。设定为三年，这些数据我们称为对应模型的学习集。学习集在采样的时候，避开周期末尾最后一个投资周期长度内的所有样本。避免学习集和应用集之间的信息重叠。

周期参数 Tm 和 Td 对模型的影响可参见 6.2 中的讨论。如下为示意图。

图 15：滚动测试周期定义



资料来源：天风证券研究所

5.2. 因变量和自变量

5.2.1. 目标模型中的因变量

由于我们要构建长周期、中性化选股的股票预测模型，所以选取**经过截面标准化的连续 20 个交易日的中性化后股票收益率之和**作为模型的因变量；为了确保模型预测值可交易，训练模型时剔除了停牌股票和连续涨停的股票。具体步骤如下：

步骤一：每日开盘前，以 VWAP 价格、计算从当日开始（含）之后 20 个交易日之间的股票收益率；

步骤二：剔除连续停牌的股票，并根据涨跌停情况对股票收益率做调整；

步骤三：用市场因子、行业因子和市值因子对股票收益率进行中性化处理，取残差；

步骤四：在截面上对残差进行标准化处理；

所得结果为因变量。

5.2.2. 股票预期收益

因变量不是收益率，所以最终要把预测值统一到股票收益率这一量纲上。因此，模型除给出预测值之外，还要同时给出股票的预期收益。我们在构建完对因变量的模型之后，对预测值和股票日收益构建线性回归模型，以模型预测值为股票日预期收益。

5.2.3. 自变量

在线性模型里，我们对于分析师预期数据因子以中位数填充缺失值、排序并作因子中性化处理，将处理之后的因子作为线性模型的自变量。

在提升树模型中，我们对分析师预期数据因子只做了排序处理，以确保因子在时序上是较为平稳的，但未填充缺失值、也未进行中性化处理。**可以自动处理缺失值是树模型的一大特点。**

5.3. 降低与基础模型相关性的方法

下面我们介绍降低与基础模型相关性的方法。首先我们介绍降低相关性的两个主要考量角度：时间周期和技术选择

5.3.1. 时间周期

以因子举例，假设因子 A 和因子 B 之间有一定的相关性，若我们希望让 B 对 A 保持正交。那么我们到底需要两个因子在多长的周期数据中保持正交呢?

一般可选目标有两类：

截面（cross-section）正交：保证两个因子每天正交。

面板（panel data）正交：保证一段时间内（例如每个月），两个因子平均正交。

常见的因子正交化处理方法和中性化处理方法，是通过线性回归取残差的方式，来达到截面正交的目的；而面板正交要比截面正交宽松很多。实际使用中，要实现截面正交这一目标会面临很多困难：

1. 维度灾难。

将几个因子之间保证正交还容易，如果因子多了则正交变得不稳定，且容易丧失因子的经济意义。

2. 不能处理数据丢失情况。

截面数据量受股票数量限制，若数据丢失比例高，则无法构建可信的回归模型。

3. 无法处理因子之间的非线性关系。

线性相关性为零，并不表示两个因子之间独立。如果两个因子之间非线性关系较强，则难以应用此方法。

因此，在实战的时候，我们优先选择截面正交，若情况不允许则选择面板正交。

5.3.2. 技术选择

为了使得目标模型和基础模型相关性低，我们通常有如下三种技术方法：

- 自变量预处理

在训练模型前，对自变量因子 x 进行预处理，使其与基础因子之间无关。

- 因变量预处理

在训练模型前，对因变量 y 进行预处理，使其与基础因子之间无关。这种思路本质上是拟合残差的思路。

- 预测值后处理

在训练模型后，对预测值进行处理，使其与基础因子或模型预测值之间无关。

这三种方法各有优劣，见下表：

表 6：去除相关性的技术比较

	优势	劣势
自变量预处理	1.方法简单直接 2.结果相关性低 3.目标模型因变量与基础因子不相关	自变量的经济含义变弱 自变量不得有缺失值、对分布要求高
因变量预处理	1.方法简单 2.自变量可以有缺失值、对其分布无要求 3.强制目标模型学习残差信息	去相关性的效果不如另外两个方法
预测值后处理	1.方法最直接、效果最好	需要管理预测和正交化两个模型

资料来源：天风证券研究所

5.3.3. 去除低相关性时的一个陷阱

还是以因子 A 和 B 为例，为了达到让因子 B 与 A 保持低相关性的目的，我们通常用 A 来回归 B，取模型残差作为新的 B 因子。但是由于回归的不稳定性、A 和 B 之间的非线性关系、或者在我们最关心的尾部 A 和 B 之间或有的复杂关系，使得回归残差所包含的信息更为复杂。我们举例如下：

假设因子 A 和 B 之间的实际关系为：

$$B = f(A) + \varepsilon$$

我们希望通过正交化处理，获得样本残差ε。做正交化处理时，我们假设 A 和 B 之间是线性关系，并估计回归系数为β，那么此处所获得的实际残差因子为：

$$B - \beta A = f(A) - \beta A + \varepsilon$$

可见，尽管残差因子与 A 的线性相关性为零，但残差因子中掺杂了因子 A 的高阶信息，扭曲了我们原本的目的。

上面是对自变量进行预处理的例子，对于因变量进行预处理也有可能带来问题。假设 r 是收益率、A 是基础因子、X 是目标模型的因子，我们希望能够用 X 构建对 r 的预测，且使得预测值与 A 保持低相关。为此我们首先用 A 来回归 r，获取残差收益率，然后用 X 来拟合残差收益率。那么实际操作中，在第一步有可能事与愿违的得到了残差：

$$r - \beta A = f(A) - \beta A + \varepsilon$$

这种情况下，该残差收益率强化了 A 中的高阶信息。若此时用 X 来回归残差，算法有可能实际上在拟合$f(A) - \beta A$，而非我们的真实目标ε。

有一些技术方法可以缓解以上问题，例如采用非参数化法提取残差。比如使用树模型来进行回归，这样可以尽量保证残差的提取过程是自适应的。在构建回归模型时，可以让回归模型稍微过拟合一点，以确保能够充分剥离拟剔除的信息。我们在构建提升树模型时尝试了该方法。

5.3.4. 本报告采用的去相关性方法

从周期维度来说，我们对于线性回归模型和提升树模型都采用了面板数据相关性低这一原则，即保证目标模型预测值和基础因子在训练周期内保持低相关性。

技术方面，我们对线性回归和树模型采取了不同的技术方法。对于线性模型，我们使用了预测值后处理法，具体步骤如下：

1. 使用分析师预期数据因子回归因变量 y，计算模型预测值；
2. 使用基础因子回归 1 中的预测值；
3. 最终模型的预测值为 1 和 2 中两个线性模型预测值之差。

对于提升树模型，我们使用因变量预处理法，具体步骤如下：

1. 使用基础因子，构建与因变量 y 之间的回归模型，提取残差；
2. 构建提升树模型，使用因子和基础因子共同对 1 中的残差进行回归；
3. 最终，模型预测值为 2 中提升树模型的预测值。

这里我们解释一下为何在在步骤 2 回归时放入基础因子。为了方便，我们以线性回归为例：

仍旧假设 r 是收益率、A 是基础因子、X 是目标模型因子，e 为用 A 对 r 回归后的残差，此时我们用 X 回归 e 可得模型：

$$e = \beta X$$

可以看出该模型的预期收益为βX，它仍旧是一个与 A 相关的预测值。若将 A 放入，回归结果变成如下形式：

$$e = \beta X - \alpha A$$

可以证明该预期收益与 A 不相关。

5.4. 训练单期模型

5.4.1. 回归模型特点比较

线性回归和树回归都是回归分析的经典模型，它们的出发点不同，但均在各自假设下发展出成熟的理论。下表总结了两种模型的主要特点：

表 7：线性模型和提升树模型的特点比较

	线性模型	提升树模型
是否可以处理数据丢失	否	是
是否可以拟合非线性	否	是
是否可以处理离群点	否	是
是否容易过拟合	否	是
是否容易对模型进行解释	是	否
是否依赖因子量纲	是	否
是否容易训练	是	否
常见类别	OLS，Lasso，Ridge，Stepwise	Random Forest，Boosting Tree
Python 包	Scikit-learn，Statsmodels	Scikit-learn，Xgboost，Lightgbm

资料来源：天风证券研究所

5.4.2. 线性模型训练方法

线性模型的训练比较常规，步骤如下：

1. 在单期样本中，使用前向逐步回归（Forward Stepwise）进行因子选择；

2. 针对因变量 y，使用 ICIR 方法对第一步中选好的因子分配权重，计算加权后的因子值；

3. 用加权后的因子对因变量 y 进行回归，以获得调整系数。

那么，分析师预期数据因子的线性系数就是第 2 步的权重和第 3 步的调整系数之积，最后再用 5.3.4 中介绍的去相关性方法，获得最终的预测模型。

5.4.3. 提升树模型训练方法

首先我们根据 5.3.4 中的方法，对因变量 y 进行预处理，提取残差。然后用提升树模型对残差进行拟合。提升树模型的实现方法有很多种，我们使用了 Light GBM。训练过程中，我们采用了 3-fold 交叉验证法，步骤如下：

1. 设定超参数网格（grid）：

树的深度（max_depth）：3 层、5 层；

学习率（learning_rate）：0.05、0.1；

树的个数（n_estimators）：1-200。

提升树模型的超参数很多，我们仅尝试了如上三个。

2. 采用交叉验证的方法，在超参数网格上进行搜索（grid search），确定最优超参数。交叉验证的流程请见 5.5.4。

值得一提的是，对于树的个数这一超参数来说，常规做法是用交叉检验做法来确定合适的个数。但这样做的话训练时间过长，因此我们采用了提升树训练中的提前中止机制（Early Stop），并以 Early Stop 后 3 个 fold 树的个数的最大值作为最终参数，这样提高了交叉验证的效率。具体内容请见 Light GBM 的官方文档。

3. 用 2 中选出的最优超参数确定模型。

这一步我们有两种做法，第一种是直接将 3-fold 检验中的训练模型求平均，以此为预测模型；第二种是使用最优超参数在全部当期数据集上重新训练，将训练结果作为预测模型。此处，我们选择第二种方法。

4. 利用交叉验证过程中，模型在验证集上的预测值，估计模型的过拟合倍数。简单来说，将在数据集 fold-1 和 fold-2 上训练的最优超参数模型、应用于 fold-3，以获取一组预测值；如此三次，可以获得三组验证集上的预测值，结合对应的因变量，可以构建一元线性回归，我们以回归系数和数值 1 的最小值为模型的过拟合倍数。

最终，我们将

$$\text{最优超参数模型的预测值} \times \text{过拟合倍数}$$

作为当期模型的预测值 $\bar{y}$。

5.4.4. 在时间序列上进行交叉验证的方法

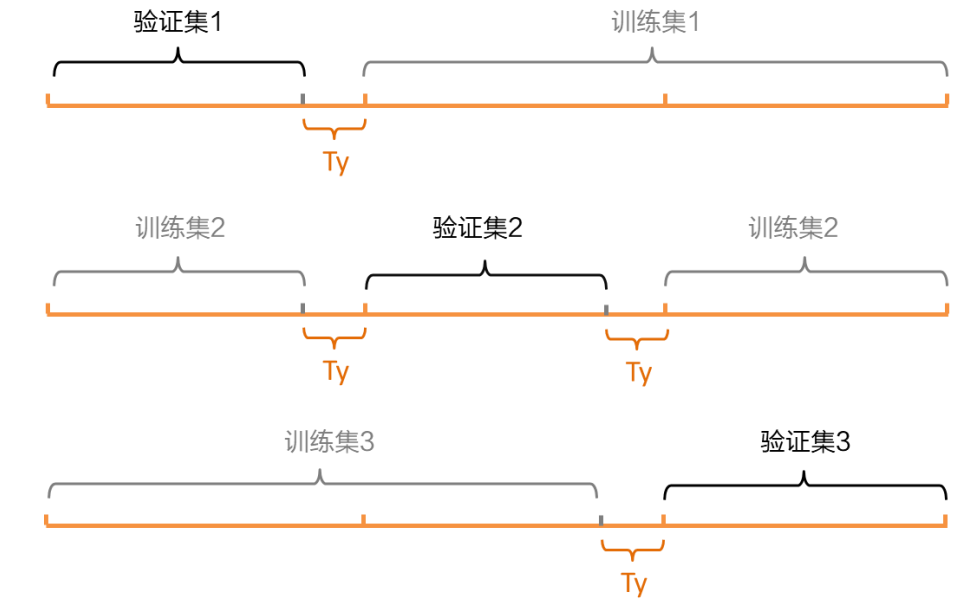
交叉检验方法是选择超参数、避免过拟合的有效方法，但传统的验证方法不是专门用于处理时间序列数据的，直接使用会出现以下两个隐患：

1. 训练集向验证集泄漏信息。由于我们的因变量的时间跨度为 20 个交易日，因此采样不当，会造成验证集和训练集之间时间重叠。

2. 无法估计不同年份的模型预测性差异。从时序的角度来说，我们希望交叉验证过程能捕捉不同年份模型预测性的差异，若采样不当，则无法对其进行评估。

因此，我们在模型训练时，遵循如下训练集和验证集的选取方法：

图 16：3-Fold 交叉检验示意图



资料来源：天风证券研究所

5.4.5. 估计股票日预期收益

如前所述，我们用回归方法来获得股票日预期收益。这一方法也可以理解为：使用模型从分析师预期数据因子中提炼出 alpha 因子，即因变量的预测值，然后计算因子的日因

子收益的预期值。

由此，对于给定 t 日，我们有两组预测值，分别为因变量 y 的预测值和股票的日预期收益。

5.5. 其它相关技术要点

5.5.1. 样本权重的选取

大多数回归方法都可以设置样本权重，线性模型和提升树均在此列。Light GBM 在训练模型时需提供样本权重，其默认值为等权重。然而，在训练集内赋予所有样本相同的权重，并不是一种合理的方法。例如，小市值股票的数量要远多于大中市值的股票，那么赋予样本相同权重，会令模型更加适应小市值股票，而非大中市值股票。这样的方式，会使得模型在大中市值股票上的模型偏差（Bias）和方差（Variance）都大于在小市值上。

但为了避免样本权重选取对结果产生额外影响，本报告仍旧采用等权重。

5.5.2. 自变量平稳性与调整预测值分布

根据我们在第二部分的讨论，分析师预期数据因子在时序上是非平稳的。这使得历史上因子数据的分布和未来因子数据的分布不一致，这种不一致性通常是有害的，会影响到预测模型在样本外的使用。这是因为，应用集自变量 x 分布发生变化，会导致预测值分布也发生变化，那么预测值的准确度会变得不可信。

以分析师的覆盖率因子为例，在历史上分析师覆盖率低，在有 4-5 个分析师覆盖的时候，模型倾向于给高预期收益；到前几年，整体分析师覆盖率提高了，用历史数据训练的模型，就倾向于给整体股票提高预期收益，但这显然不合理。尽管在 5.3.4 中，我们通过截面排序等处理手段，一定程度上将因子变得更加平稳，但是并不能使其完全平稳，因此因子分布的不平稳性带来的模型预测风险是存在的。

我们不讨论如何度量因子分布的偏差，而只给出两种根据训练集上的预测值分布函数、来修正测试集上的预测值方法：

- 调整截面均值

利用平移操作，使得截面均值与训练集上的预测值分布的均值一致

- 调整截面标准差

利用放缩操作，使得截面标准差与训练集上预测值的标准差一致

在本报告中，我们对提升树模型的预测值采用了调整均值的做法；而对于线性模型，由于因子本身已经经过正交化的处理，不存在均值不一致问题。

5.5.3. 交叉检验的效用函数

在进行交叉检验时，我们首先选取用于检验的效用函数，并计算验证集上模型的效用函数值，以此为依据来选择超参数。一般机器学习方法中有多种推荐的效用函数，但设计目的通常并不是针对面板数据。例如在回归问题中常用的残差平方和函数，并没有任何与时序有关的信息。因此我们可以根据模型需求，针对性的设计效用函数。

在本次测试中，我们采用复合效用函数，以加权残差平方和作为效用函数，同时要求在验证集上的日多空收益率的时序 t-检验 p 值不高于 0.025。

6. 模型风险的来源、评估与模型赋权

几乎所有的模型风险都是拟合带来的。下面我们对第 5 节中设计的模型构建流程、所隐含的拟合风险进行讨论。

6.1. 风险来源

建模流程中包含但不限于以下三种风险：

1. 训练单个模型时，提升树模型通过交叉检验来选择最优超参数，这一阶段会出现风险；线性模型优化了线性系数，也有拟合风险。

该阶段中，样本是固定的，我们可以认为这种风险主要由模型和训练方法的运用产生，与其它因素无关。

2. 训练单个模型时，选择与训练周期有关的超参数（例如：5.2.2 中的 Td、Tm）会出现过拟合。

在本报告中，我们选择训练样本的周期长度为三年、单个模型的使用周期长度为半年。但实际上也可以把这两个周期设定为参数，通过在样本内进行测试来选择最优的周期。

我们从信号处理的角度，对这里面蕴含的数学思想进行分析。假设数据中包含噪音（irreducible noise）和高中低频三种信息，这里所谓高中低频是相对于投资期限长度而言的。

首先，从历史中学东西，那么学的一定是历史中某种不变的东西，那么它应属于低频信息，而噪音、高频信息和中频信息都会我们的学习效果产生影响。其次，训练模型相当于对训练样本中的数据进行了某种“求平均”的操作；而滚动建模，相当于移动平均。

那么类似于信号处理中移动平均算子的分析逻辑，我们有如下推断：

- 模型学习是在降噪和过滤高频信息，学习结果为数据窗口内低频和中频信息的“平均”信息之和；
- 窗口越宽，降噪效果越好，但同时在低频和中频上学到的“平均”信息也越滞后；
- 低频的信息占比越高，学到的信息外延性越好；否则外延性越差。

从以上分析可知，我们可以把模型风险大致分为三个部分，分别为：

第一，噪音风险。

第二，高频部分的风险。学习的时候滤掉了，但实际上是存在的，若策略投资周期为长周期，那么高频部分风险可以接受。

第三，中频信息的外延风险。

因此，噪音、高频、中频和低频在整个数据中的方差占比，是最终影响模型外延有效性的关键。最理想的数据模式是，低频占比高，其他占比低。

可见，训练数据周期长度（Td）和模型应用周期长度（Tm）的选择实际上是在高、中、低频之间寻找平衡点。该平衡点是由高中低频 alpha 以及噪音在整体信息中的方差占比决定的。不幸的是，这种比例也不是一个常量，那么在样本内学到的最优的 Td、Tm，在样本外或实盘阶段可能会有变化，这一步产生了拟合风险。

3. 通过样本内数据构建因子、模型和交易策略同样会产生拟合风险。我们总是倾向于在样本内数据中寻找好的因子、确定最优的模型、采纳结果更好的交易策略。这种选择的过程会产生拟合风险。

所以，我们需要比较模型在样本内和样本外的表现，并进行必要的调整。

6.2. 风险评估

我们用模型在应用集上的表现来评估模型风险，一种方法是评估模型对因变量预测的稳定性，另一种是直接评估模型对股票收益率预测的稳定性。本报告选用第二种评估方法。我们构建股票组合策略，通过评估策略表现来评估模型风险。

假设向量w_t是一种股票组合策略在 t 日的权重，即：

$$w_t = w(\bar{y}_t, \bar{y}_{t-1}, \dots)$$

此处，$\bar{y}_t$是模型在 t 日的预测值。假设r_t是 t 日对应的股票收益率，$\bar{r}_t$是模型给出的预测值，那么我们预期：

$$\mathrm{E}[r_t] = \bar{r}_t$$

那么，策略在 t 日的预期收益为$w_t \cdot \bar{r}_t$，真实收益为$w_t \cdot r_t$，分别记为R_t和 $\bar{R}_t$。我们关心如下两个统计量：

1. 真实收益R_t的时序方差。

我们要证明，R_t的时序方差可以度量 6.1 中第 2 点里所提到的三种风险。启发式的，根据 6.1 中第 2 点，我们假设股票收益率满足：

$$r_t = \bar{r}_t^l + \bar{r}_t^h + \varepsilon_t$$

其中，$\bar{r}_t^l$和 $\bar{r}_t^h$分别代表低频和高频信息对股票收益的影响。那么，结合我们的预测模型：

$$r_t = \bar{r}_t + (\bar{r}_t - \bar{r}_t^l) + \bar{r}_t^h + \varepsilon_t$$

进一步假设以上四部分在时序上是独立的，那么：

$$\mathrm{Var}(R_t) = \mathrm{Var}(w_t \cdot \bar{r}_t) + \mathrm{Var}(w_t \cdot (\bar{r}_t - \bar{r}_t^l)) + \mathrm{Var}(w_t \cdot \bar{r}_t^h) + |w_t|^2\sigma_\varepsilon^2$$

若令

$$w_t = \frac{\bar{r}_t}{|\bar{r}_t|^2}$$

那么右端第一项为

$$\mathrm{Var}(w_t \cdot \bar{r}_t) = \mathrm{Var}\left(\frac{\bar{r}_t}{|\bar{r}_t|^2} \cdot \bar{r}_t\right) = \mathrm{Var}(1) = 0$$

于是，

$$\mathrm{Var}(R_t) = \mathrm{Var}(w_t \cdot (\bar{r}_t - \bar{r}_t^l)) + \mathrm{Var}(w_t \cdot \bar{r}_t^h) + |w_t|^2\sigma_\varepsilon^2$$

这正好与 6.1 中第 2 点里所提到的三种风险对应。

2. 真实收益R_t的时序均值与预期收益 $\bar{R}_t$的时序均值之比

我们当然希望这个统计量整好为 1，如此的话说明模型的估计是准确的。我们启发式的分析一下这个统计量的含义。假设股票收益率满足：

$$r_t = \beta_t \bar{r}_t + \varepsilon_t$$

那么R_t和 $\bar{R}_t$分别为，$\beta_t w_t \cdot \bar{r}_t$和$w_t \cdot \bar{r}_t$，即：

$$R_t = \beta_t \bar{R}_t$$

进一步，假设策略的日预期收益 $\bar{R}_t$是常数 $\bar{R}$，那么预期收益 $\bar{R}_t$的时序均值即为 $\bar{R}$，而真实收益R_t的时序均值为：

$$avg(R_t) = avg(\beta_t \bar{R}) = \bar{R}avg(\beta_t)$$

于是两个时序均值之比为$avg(\beta_t)$。

通过β_t大于 1 或者小于 1，可以诊断模型是欠拟合还是过拟合，亦或有偏差。继续上面的例子，我们来看一下 t 日r_t的截面方差：

$$\mathrm{Var}(r_t) = \beta_t^2 \mathrm{Var}(\bar{r_t}) + \mathrm{Var}(\varepsilon_t)$$

那么 $\bar{r_t}$这一信息对r_t的真实解释度R^2是：

$$\beta_t^2 \frac{\mathrm{Var}(\bar{r_t})}{\mathrm{Var}(r_t)}$$

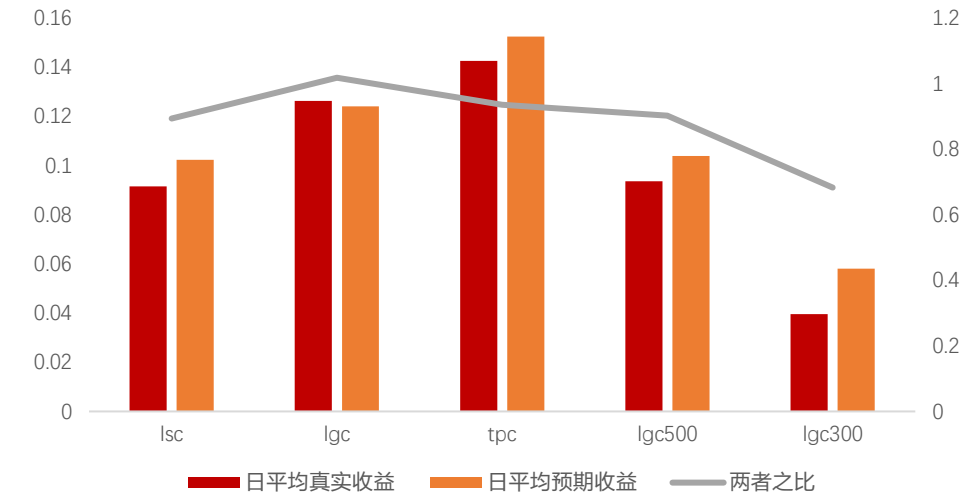
而我们对模型的预期解释度 $\bar{R^2}$（对应模型在学习集上的解释度）为

$$\frac{\mathrm{Var}(\bar{r_t})}{\mathrm{Var}(r_t)}$$

可以看出，β_t^2是真实解释度和预期解释度之比。如果模型过拟合，那么预期解释度会大于真实解释度，这样的话，β_t^2必然是一个小于 1 的数；反之大于 1。因此可以通过观察给定策略收益的β_t均值，来判断模型在策略反应的模型方面是否需要调整。

我们将基础模型五种组合策略的**真实收益R_t的时序均值与预期收益 $\bar{R_t}$的时序均值之比**展示如下：

图 17：基础模型模型风险



资料来源：朝阳永续，wind，天风证券研究所

从图中我们可以看出，除了沪深 300 指数对冲组合，其他四个组合的真实收益和预期收益之比大约为 0.9；沪深 300 指数对冲组合的真实收益和预期收益之比大约为 0.7，大幅低于其他组合，说明要么基础模型在该风格上有较大的方差，要么有较大的偏差，具体原因我们不在此分析。

6.3. 模型赋权

我们综合考虑 6.2 中的两个统计量来对模型赋权。由于模型赋权不是本报告的重点，这里我们只简单讨论赋权逻辑的出发点。

1. 用**真实收益R_t的时序均值与预期收益 $\bar{R_t}$的时序均值之比**来调整模型的预测值。

若两个收益之比小于 1，说明我们过高估计了模型的预测性。那么给模型赋权时要考虑给与惩罚。

2. 用**真实收益R_t的时序方差**$\mathrm{Var}(R_t)$做惩罚

这一点跟因子模型中的逻辑相同，从优化策略时序夏普率的角度出发，时序风险大的策略要给与惩罚。需要注意的是，根据我们的讨论，R_t的时序方差为：

$$\mathrm{Var}\left(w_t \cdot (\bar{r_t} - \bar{r_t^l})\right) + \mathrm{Var}\left(w_t \cdot \bar{r_t^h}\right) + |w_t|^2 \sigma_\varepsilon^2$$

其中一旦$|w_t|^2$是一个常数，那么$|w_t|^2\sigma_\varepsilon^2$这一部分对于任意预测模型都是相等的。所以

$$\mathrm{Var}(R_t) - |w_t|^2\sigma_\varepsilon^2 = \mathrm{Var}\left(w_t \cdot (\bar{r_t} - \bar{r_t^l})\right) + \mathrm{Var}\left(w_t \cdot \bar{r_t^h}\right)$$

才是对模型惩罚时的关键。注意σ_ε^2是未知的，但是我们可以给它一个估计值。

3. 综合考虑不同类别的策略组合

6.2 中介绍的两种统计量在构建的时候，依赖于组合策略w_t的选取。可以综合考虑几种不同的策略所计算的统计量作为参考。

4. 要放入到预测系统中进行印证

设计 alpha 模型时，我们尽量保证新模型与已有模型之间保持独立，但这一点是不可能完全做到的。所以要将新模型和已有模型放到一起验证新模型权重的合理性。

根据 6.2 中基础模型的真实收益与预期收益比，我们在下面模型整合时，为基础模型赋权 0.9。

7. 模型实证分析

下面我们分析根据第 5 节方法所建模型的效果，首先简要介绍模型效果分析方法。

7.1. 分析方法

我们根据如下步骤分析模型：

- 分析预测值分布和时序性质
- 衡量模型对因变量 y 的预测性

我们定义预测值和因变量的日截面 Pearson 相关系数为模型的 IC 值。

- 诊断模型

1. 线性关系

根据预测值$\bar{y_t}$将样本分成 10 组，计算每组$\bar{y_t}$的平均值$Avg(\bar{y_t})$和y_t的平均值$Avg(y_t)$。理想情况下，两者成线性关系。

2. 分行业表现

分行业计算预测值和因变量的截面 Pearson 相关系数，对比不同行业之间的差异。

3. 分市值表现

根据市值大小将股票分成 5 组，分组计算预测值和因变量的截面 Pearson 相关系数，对比不同市值的差异。

- 衡量模型之间相关性

分析基础模型和目标模型的预测值之间截面 Pearson 相关系数的时序特征。

- 分析组合策略表现

依据第 9 节所述方法构建组合策略，分析它们的表现。

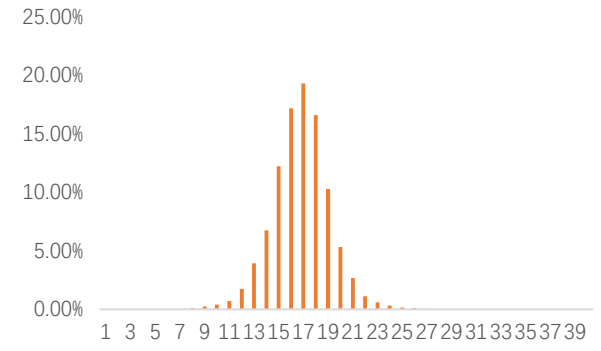
- 用线性模型和提升树模型构建等权模型，对比其表现及与线性模型的差异
- 分析目标模型对基础模型的增量效果

同时使用等权目标模型和基础模型，构建组合策略，对比与单独使用目标模型的组合收益率的增量效果，并分析增量收益和基础模型策略收益之间的相关性。

7.2. 预测值分布及时序表现

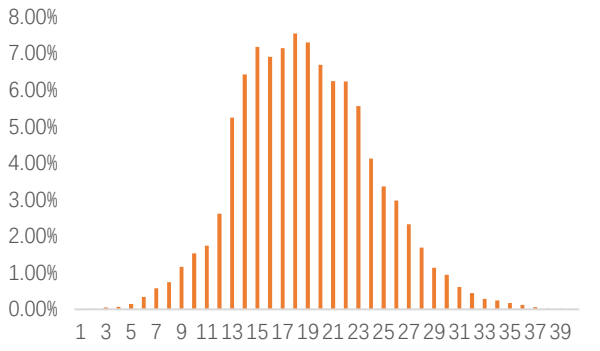
我们从预测值的分布、时序上的自相关性和截面标准差平稳性三个角度，对预测值进行分析。下面是线性模型和提升树模型在 2010 年和 2014 年预测值的分布图

图 18：线性模型 2010 年预测值分布



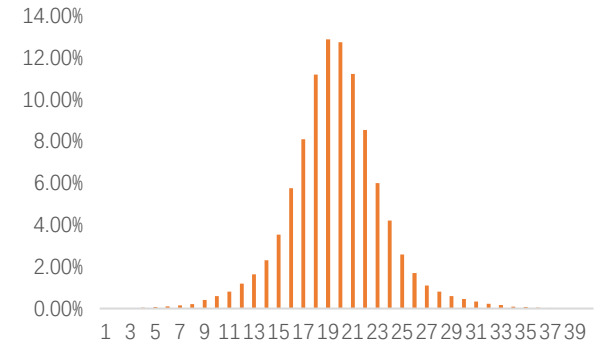
资料来源：朝阳永续，wind，天风证券研究所

图 19：提升树模型 2010 年预测值分布



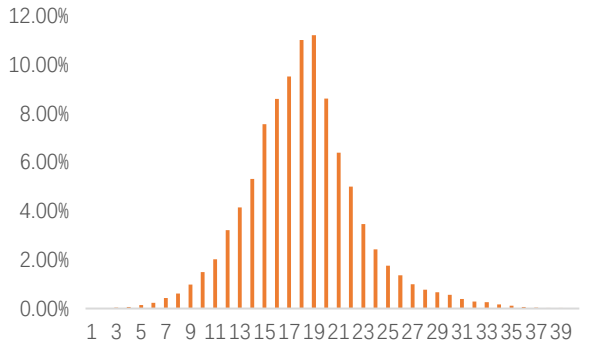
资料来源：朝阳永续，wind，天风证券研究所

图 20：线性模型 2014 年预测值分布



资料来源：朝阳永续，wind，天风证券研究所

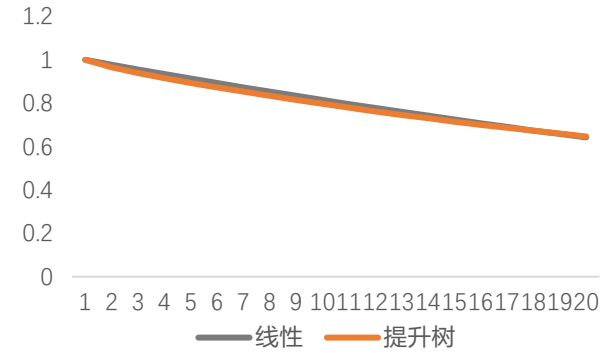
图 21：提升树模型 2014 年预测值分布



资料来源：朝阳永续，wind，天风证券研究所

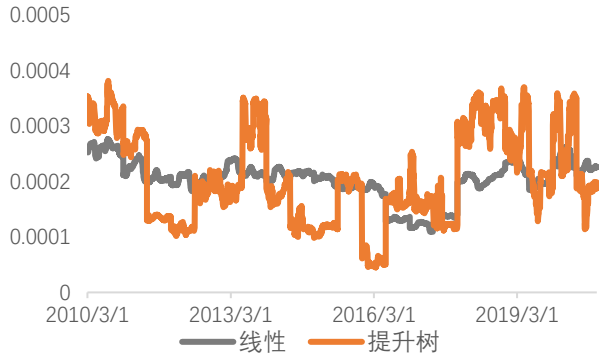
下面是预测值的截面自相关性和日预期收益截面标准差：

图 22：预测值时序自相关性



资料来源：朝阳永续，wind，天风证券研究所

图 23：日预期收益截面标准差



资料来源：朝阳永续，wind，天风证券研究所

从以上图中可以看出：

1. 线性模型的预测值基本上呈现尖峰肥尾的分布形态，同时两个年份的分布比较相似；而提升树模型预测值的分布形态并不相同，会随着时间发生变化；

2. 两模型预测值自相关性曲线相似，自相关性强；

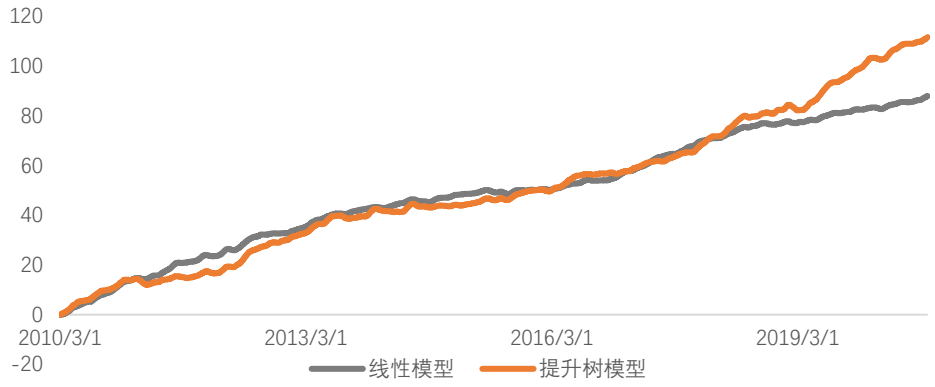
3. 提升树模型预测值的截面标准差的稳定性要比线性模型的差，波动更大。

7.3. 对因变量的预测性

两模型 IC 值的平均值分别为 3.4%和 4.3%，标准差分别为 4.1%和 6.1%。从下面的 IC 值的时序累加图和统计量中可以看出，

1. 两种模型的预测性有效；
2. 树模型预测性好于线性模型，但稳定性较差。
3. 树模型自 2018 年，预测性比线性模型好。

图 24：目标模型 IC 值时序累加图

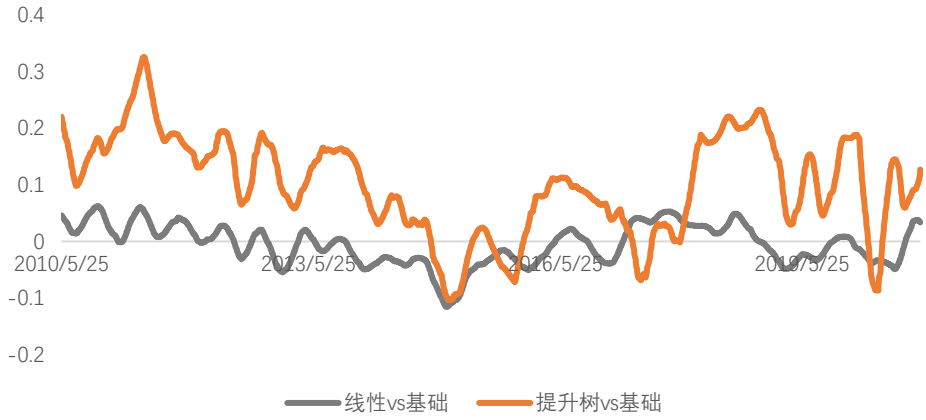


资料来源：朝阳永续，wind，天风证券研究所

7.4. 与基础模型相关性

下图是线性模型与提升树模型和基础模型日相关性的 60 天均值时序图，

图 25：目标模型预测值与基础模型预测值截面相关性 60 天均值



资料来源：朝阳永续，wind，天风证券研究所

从图中可以看出，线性模型与基础模型的相关性几乎为 0，而提升树模型与基础模型之间的相关性波动较大、均值约为 10%。值得一提的是，用提升树模型直接拟合分析师数据因子，预测值与基础模型预测值的相关性均值为 40%。

7.5. 模型诊断

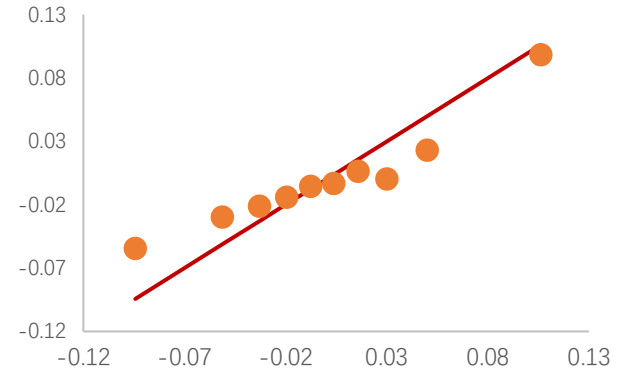
7.5.1. 线性关系

下图是两个模型的分组预测情况。横轴为预测值分组平均值的时序平均，纵轴为因变量分组平均值的时序平均；红线为{x=y}的参考线。

从图中可以看出，提升树模型的线性表现要好于线性模型，这符合我们的预期。一方面，提升树模型本身就可以做非线性拟合；另一方面，提升树模型在拟合的时候对自变量

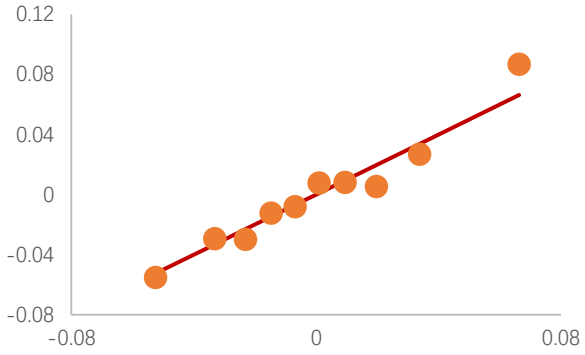
的分布不敏感，不会受到离群点的影响，鲁棒性更好。

图 26：线性模型预测值与自变量线性关系



资料来源：朝阳永续，wind，天风证券研究所

图 27：提升树模型预测值与自变量线性关系

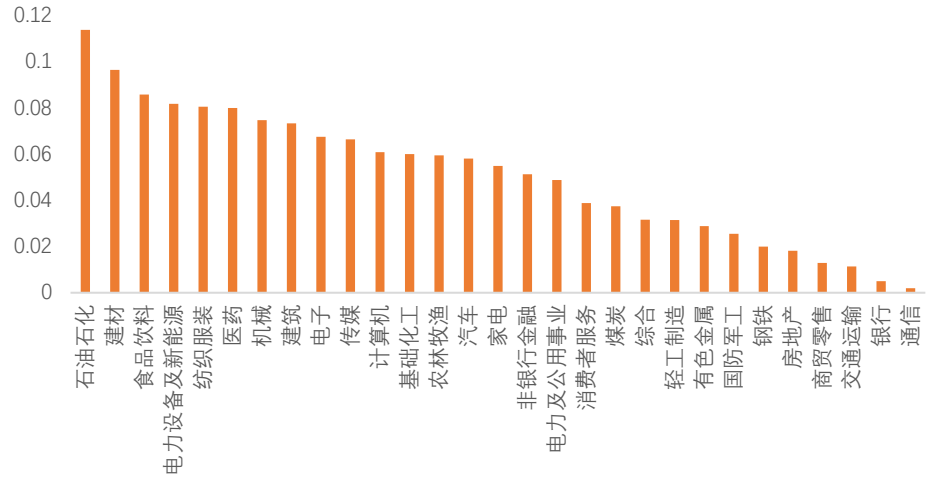


资料来源：朝阳永续，wind，天风证券研究所

7.5.2. 分行业表现

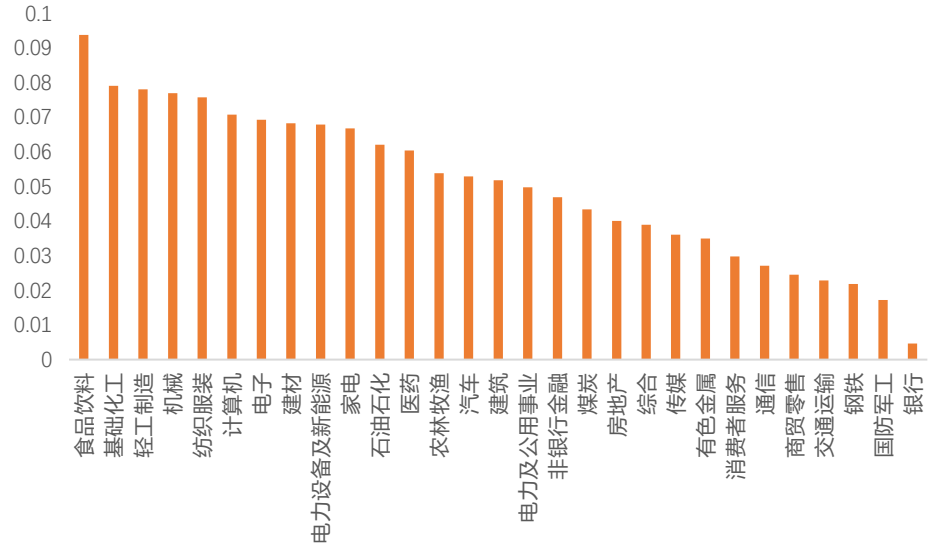
下面是两个模型的分行业平均 IC 值。

图 28：线性模型分行业表现



资料来源：朝阳永续，wind，天风证券研究所

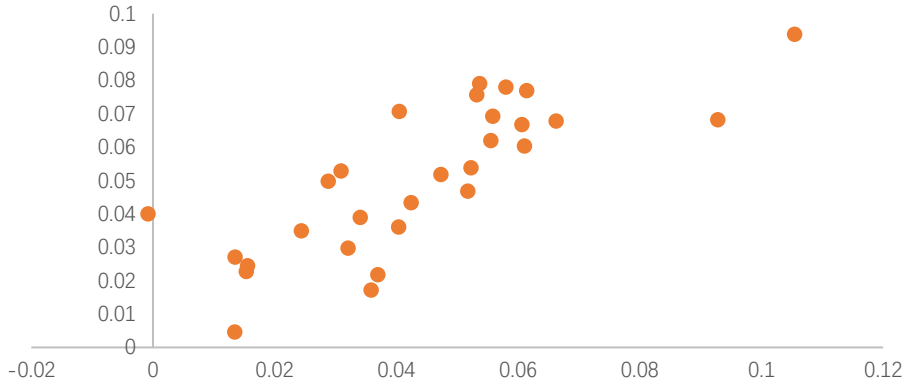
图 29：提升树模型分行业表现



资料来源：朝阳永续，wind，天风证券研究所

下面是两个模型在不同行业上 IC 均值的对比图：

图 30：线性模型与提升树模型分行业表现对比



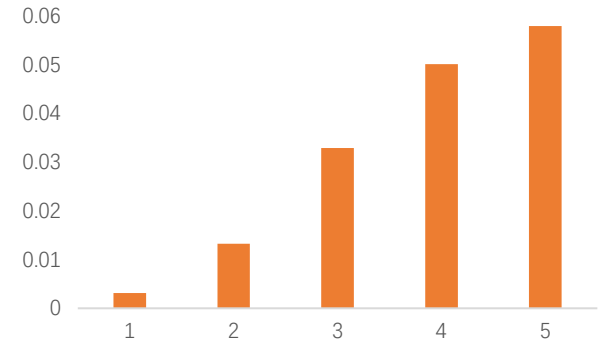
资料来源：朝阳永续，wind，天风证券研究所

从图中可以看出，两种模型在大部分行业上都表现出了预测性，且两者分行业的预测性有较高的相关性，其中提升树模型在不同行业上的预测性更平衡。

7.5.3. 分市值表现

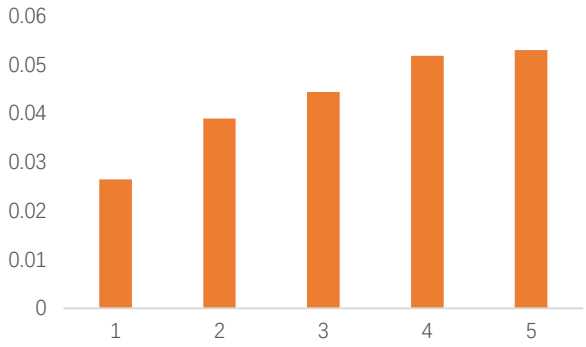
下面是两个模型的分市值表现，最左一组对应小市值股票：

图 31：线性模型分市值表现



资料来源：朝阳永续，wind，天风证券研究所

图 32：提升树模型分市值表现



资料来源：朝阳永续，wind，天风证券研究所

从图中可以看出：

首先，两个模型都随市值增大，预测性依次变好。这或与数据在小市值上覆盖率低有关；

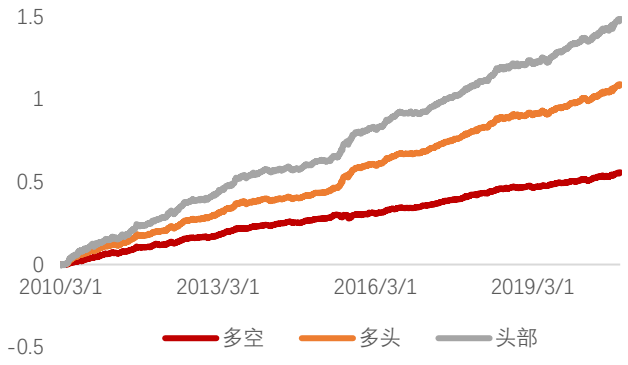
其次，线性模型变化更加陡峭；

最后，线性模型在小市值上表现不如提升树模型，而提升树模型在大市值上不如线性模型。

7.6. 组合策略表现

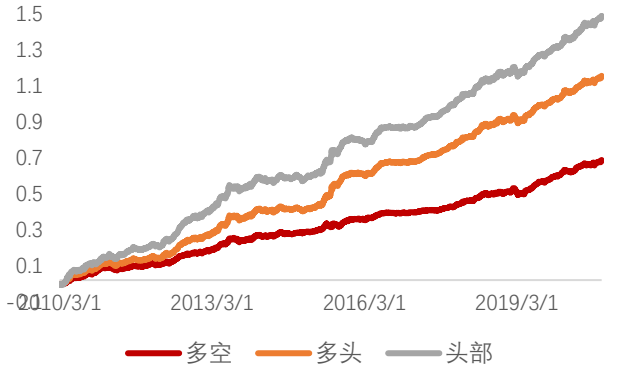
下面是策略的日累计收益图：

图 33：线性模型组合策略表现



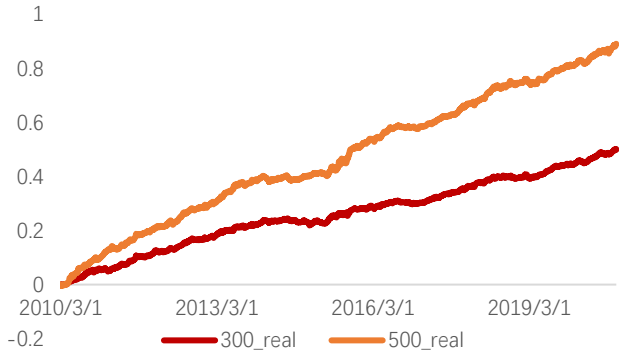
资料来源：朝阳永续，wind，天风证券研究所

图 34：树模型组合策略表现



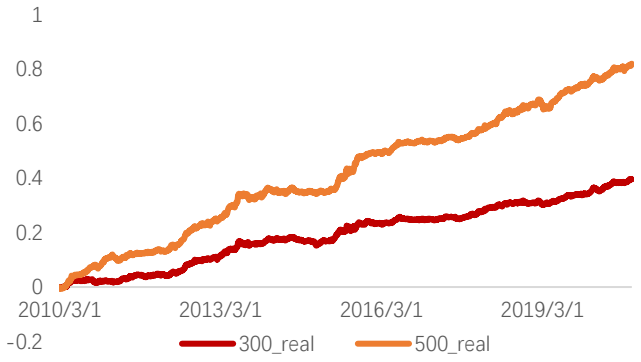
资料来源：朝阳永续，wind，天风证券研究所

图 35：线性模型指数增强表现



资料来源：朝阳永续，wind，天风证券研究所

图 36：提升树模型指数增强表现



资料来源：朝阳永续，wind，天风证券研究所

下表是各策略统计量：

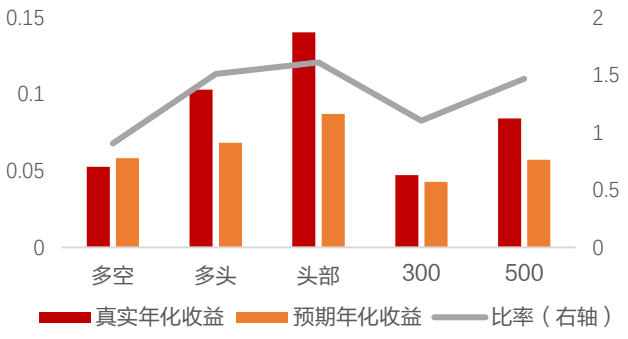
表 8：目标模型组合策略收益统计量

	线性模型					提升树模型				
	多空	多头	头部	300	500	多空	多头	头部	300	500
年化收益	5.3%	10.3%	14.0%	4.7%	8.4%	6.5%	10.9%	14.1%	3.7%	7.7%
年化波动率	1.8%	2.9%	4.0%	2.0%	2.7%	2.4%	3.4%	4.5%	2.1%	2.9%
夏普率	2.90	3.57	3.53	2.36	3.17	2.75	3.22	3.14	1.79	2.66

资料来源：朝阳永续，wind，天风证券研究所

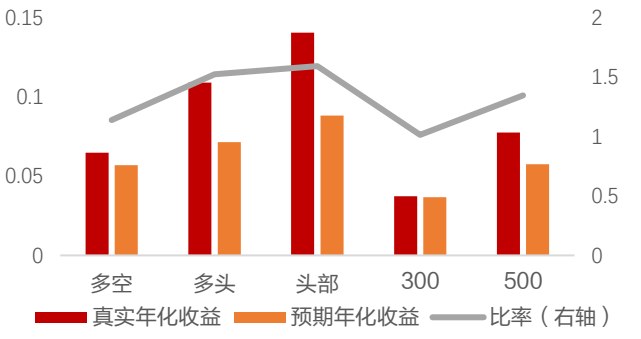
最后，下图中展示了组合策略的真实收益和预期收益之比：

图 37：线性模型策略真实与预期收益对比



资料来源：朝阳永续，wind，天风证券研究所

图 38：提升树模型策略真实与预期收益对比



资料来源：朝阳永续，wind，天风证券研究所

从图中我们可以看出：

1. 两个模型的在 5 种策略组合上都有较好表现；

2. 线性模型的组合策略稳定性好于提升树模型；

3. 提升树模型的多空组合收益优于线性模型，在多头和头部组合收益相仿；

4. 线性模型在指数增强上表现更好，这与该模型在大市值风格上表现更好有关；

5. 两个模型的多空策略、300 增强策略组合的真实收益和预期收益之比均在 1 附近，其他 3 个策略组合的真实收益和预期收益之比均大于 1，这与线性诊断结果一致。

7.7. 分析和总结

根据以上分析可知，

1. 两种模型都能稳定提取分析师预期数据中的 alpha 信息；

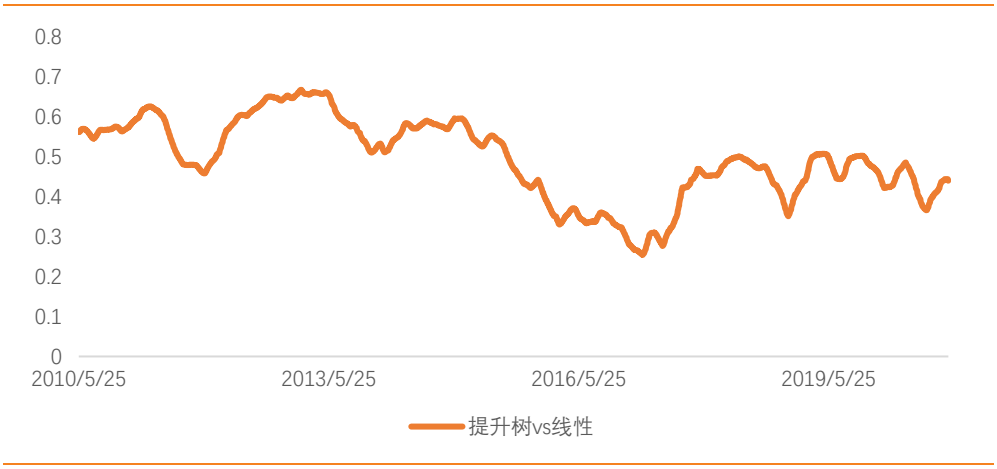
2. 线性模型的整体稳定性要好于提升树模型；

3. 线性模型预测性在不同市值上的差异较大，而提升树模型则较小，说明两者在预测性上有互补性。

7.8. 模型相关性和模型融合

下图是线性模型和提升树模型预测值之间的日截面相关系数的 60 天均值。

图 39：目标模型之间预测值截面相关系数 60 天均值

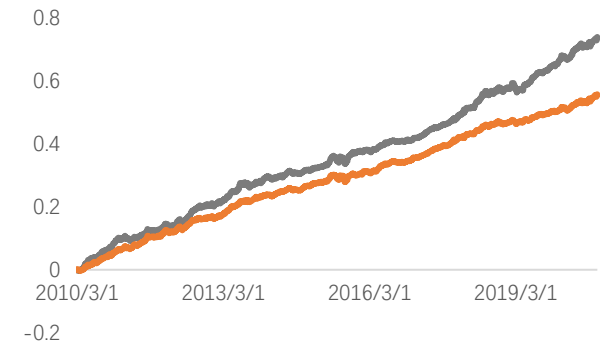


资料来源：朝阳永续，wind，天风证券研究所

从图中可以看出，两个模型的 60 天截面相关系数均值大约在 50%附近，且自 2015 年以来有整体下降的趋势，最低达到过 25%，这进一步说明两者之间有互补性。

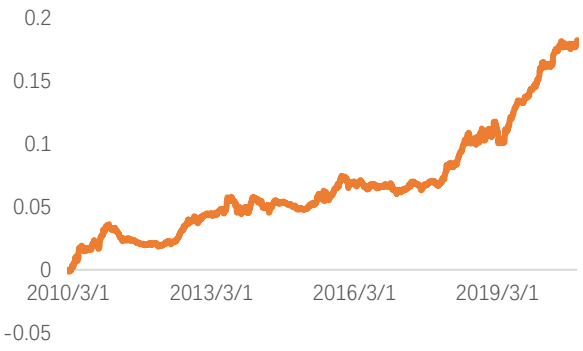
我们将两个模型赋予等权、称其为等权模型，考察等权模型的组合策略收益、以及与线性模型组合策略收益的差异，下面是组合策略的对比图和增量收益图：

图 40：等权模型与线性模型多空收益对比



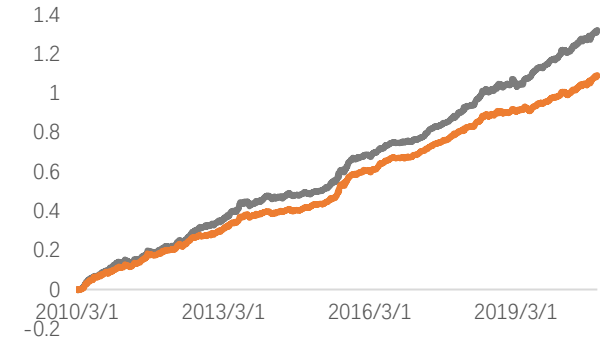
资料来源：朝阳永续，wind，天风证券研究所

图 41：等权模型与线性模型多空收益差



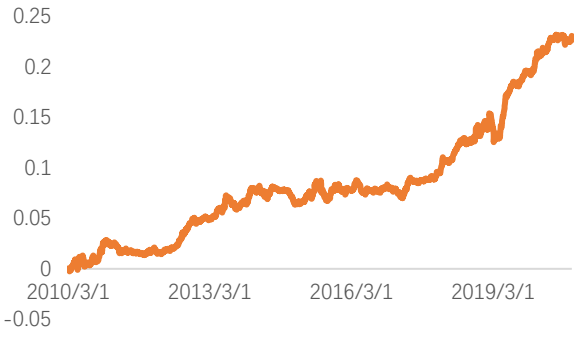
资料来源：朝阳永续，wind，天风证券研究所

图 42：等权模型与线性模型多头收益对比



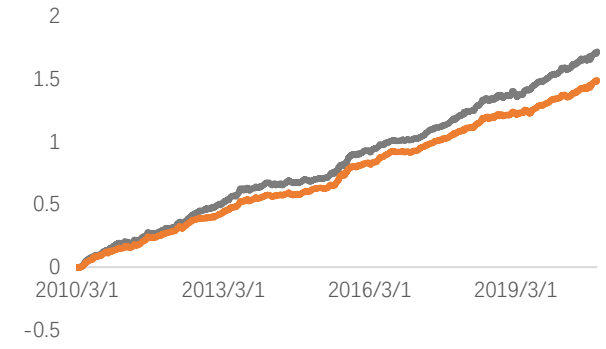
资料来源：朝阳永续，wind，天风证券研究所

图 43：等权模型与线性模型多头收益差



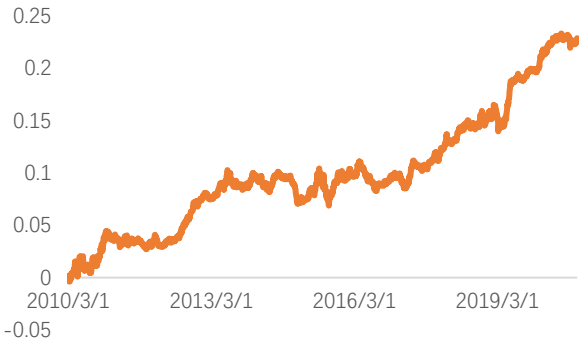
资料来源：朝阳永续，wind，天风证券研究所

图 44：等权模型与线性模型头部多头收益对比



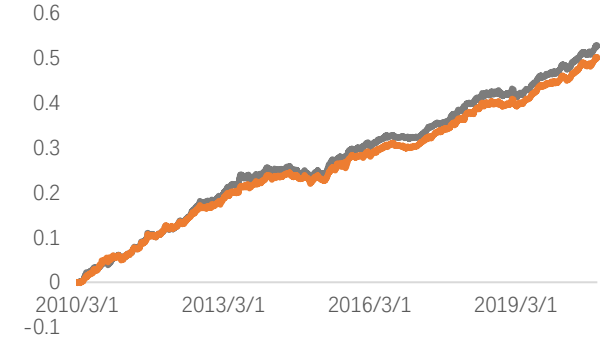
资料来源：朝阳永续，wind，天风证券研究所

图 45：等权模型与线性模型头部多头收益差



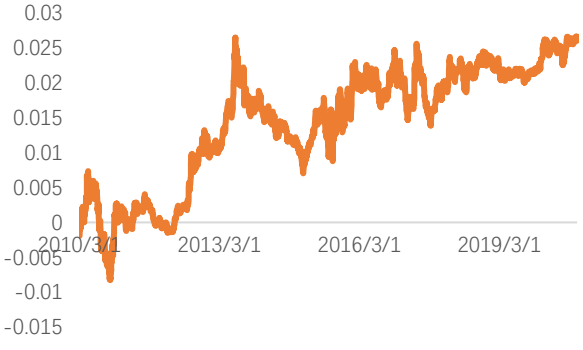
资料来源：朝阳永续，wind，天风证券研究所

图 46：等权模型与线性模型 300 增强收益对比



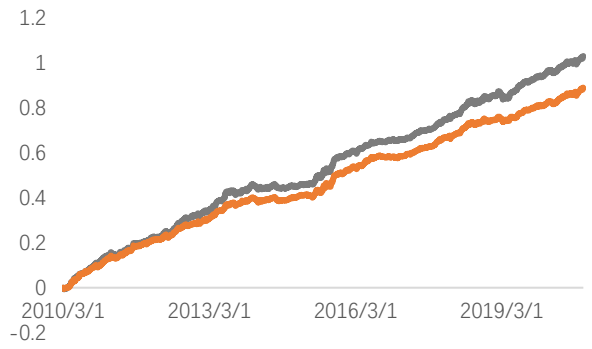
资料来源：朝阳永续，wind，天风证券研究所

图 47：等权模型与线性模型 300 增强收益差



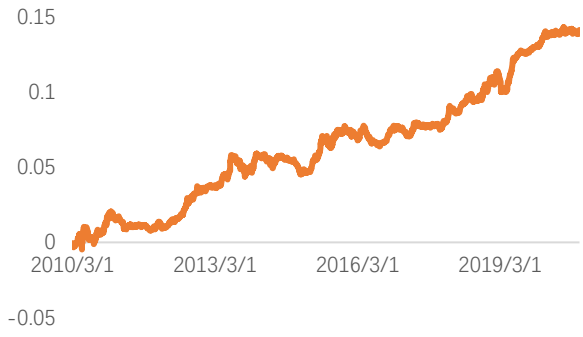
资料来源：朝阳永续，wind，天风证券研究所

图 48：等权模型与线性模型 500 增强收益对比



资料来源：朝阳永续，wind，天风证券研究所

图 49：等权模型与线性模型 500 增强收益差



资料来源：朝阳永续，wind，天风证券研究所

以下是等权模型组合策略的统计结果

表 9：等权模型策略表现

	多空	多头	头部	300	500
年化收益率	7.0%	12.5%	16.2%	5.0%	9.7%
年化波动率	2.1%	3.2%	4.1%	2.1%	2.8%
夏普率	3.33	3.91	3.91	2.36	3.52
增量与线性模型策略相关性	0.03	-0.01	0.03	0.03	0.02

资料来源：朝阳永续，wind，天风证券研究所

从上面的结果可以看出：

1. 等权模型的表现优于线性模型，在组合策略上，无论年化收益和夏普率都有改善；
2. 增量收益与线性模型组合策略收益相关性很低，说明增量收益是独立增量；
3. 在 2016 年中后，增量部分增幅变大；
4. 在沪深 300 指数增强组合上，等权策略提高不大。

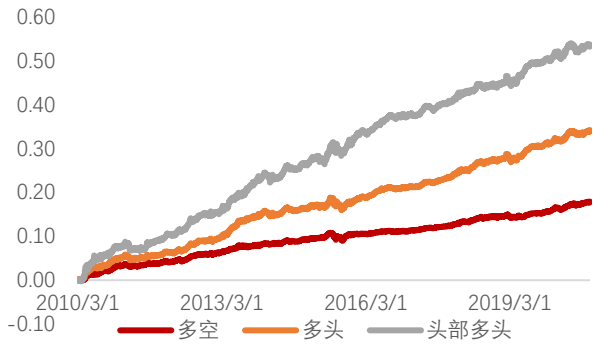
7.9. 相对基础模型的增量收益分析

最后我们把等权目标模型和基础模型整合，构建综合模型，根据 6.2 和 7.6 中真实收益与预期收益比的分析，设定模型权重为 1.0 和 0.9。我们对于整合后的模型计算组合策略收益，验证目标模型能否对基础模型达到独立增强效果。具体方法如下：

1. 用综合模型计算策略收益，计算与基础模型策略收益之差，记作增量部分；
2. 计算增量部分的年化收益率、年化波动率和夏普率；
3. 计算增量部分和基础模型策略收益率的相关性；
4. 计算真实增量收益和预期增量收益之比。

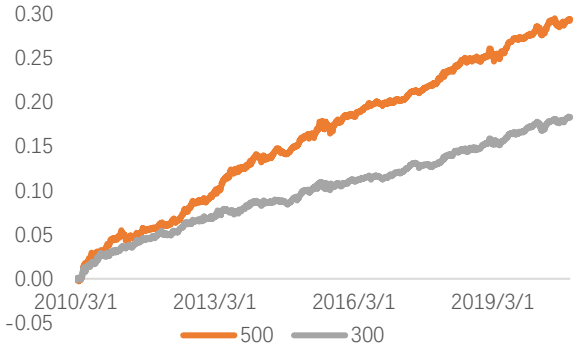
下面是增量部分累计图

图 50：多空、多头与头部多头策略增量收益



资料来源：朝阳永续，wind，天风证券研究所

图 51：指数增强策略增量收益



资料来源：朝阳永续，wind，天风证券研究所

下表是增量部分的统计指标：

表 10：增量收益统计量

	多空	多头	头部多头	300	500
增量年化收益率	1.7%	3.2%	5.1%	1.7%	2.8%
增量年化波动率	0.8%	1.6%	2.6%	1.0%	1.3%
增量夏普率	2.05	2.07	1.96	1.65	2.20
增量与基础策略相关性	35%	20%	1%	-7%	16%
真实增量与预期之比	1.16	1.44	1.84	0.99	1.35

资料来源：朝阳永续，wind，天风证券研究所

从结果中可以看出，等权目标模型可以稳定增强基础模型组合策略；除多空策略之外，增量部分与原策略收益率均保持了低相关性；真实增量与预期增量基本保持一致。这里值得指出的是，目标模型和基础模型截面低相关，并不保证它们组合策略的收益率时序相关性也低，两种策略可能受到同样的宏观或中观变量影响，使得策略收益率相关性变高。所以增量收益与原收益率保持低相关性是十分难得的。

8. 总结

本报告从统计模型方法出发，用定期滚动建模的方法，利用分析师数据构建股票收益率预测模型。通过采用线性模型和提升树模型，在保持预测值与财务类指标和市值类指标保持低相关性的前提下，提取分析师数据中“独有”信息。

针对线性模型和提升树模型的特点，报告采用了不同的技术方法来训练模型，均达到了事先设定的建模要求。值得一提的是，我们并没有对提升树模型进行过多的优化和设计，只是采用了最普通的做法。通过分析发现，两个模型的预测效果相似、预测值不完全相关，且在行业和市值风格上呈现不同的统计特征，说明两个模型有互补性。

线性模型和提升树模型等权配置后，组合策略的表现有所提升，均优于两个模型的组合策略表现。等权模型组合策略相对于线性模型的增量收益，与线性模型策略收益本身相对独立。

最后，等权模型可以有效增强基础模型的策略收益，增强部分与基础模型组合策略收益基本保持独立，增强有效、稳定。

9. 附：组合策略回测

本节我们介绍用于评估模型的组合策略的构建方法。组合策略构建要注意如下几个方面：

1. 要构建多种组合策略，对模型进行综合评价；

2. 策略设计要与模型的目标投资期限匹配；

3. 用策略考察模型预测性时，**不要**在策略中加入换手率限制；

4. 构建组合时，在满足约束的前提下，所持股票尽量分散，例如大于 250 只。

我们解释一下第 3 和第 4 点。

对于第 3 点，有些报告在对模型或因子进行分析和选择时，将它们的换手率也作为评价它们表现的一个因素。但实际上我们并不知道因子将被如何使用，换手率应该是股票组合构建策略索要考虑的，而非预测系统来考虑。最终，应由股票组合策略综合考虑股票收益预测结果和交易成本，来确定实际交易时的换手率。但是大家会质疑，若不统一换手率则策略不完备，策略和策略之间不可比。事实上，我们在回测一个模型时会通过投资周期来控制策略的交易频率；同时，对比两个模型的收益率的时候，使用同样的投资周期参数，以保证两者之间可比。

对于第 4 点，持股尽量分散的原因在于，这样才能考察模型的统计特征，如果持股数量太少，会有太多偶然性，不适合考察单个模型。

9.1. 组合设计

我们设计了五种组合策略，分别为：

1. 带约束的多空组合

组合在市场、行业、风格上的暴露度为零，在控制组合权重平方和的前提下，最大化组合的预期收益。以组合收益率为策略收益率。

2. 以等权组合为基准的正预期收益组合

以等权组合的行业分布、风格暴露为约束，选取预期收益为正的股票，在控制组合权重平方和的前提下，最大化组合的预期收益，构建多头组合。以组合的超额收益为策略收益率。

3. 以等权组合为基准的头部正预期收益组合

以等权组合的行业分布、风格暴露为约束，选取预期收益为正的股票，在控制组合权重平方和的前提下，最大化组合的预期收益，构建多头组合。以组合的超额收益为策略收益率。

与 2 不同之处在于，权重平方和的上限值取相应的两倍。

4. 沪深 300 增强组合

以沪深 300 组合的行业分布、风格暴露为约束，在控制组合权重平方和的前提下，最大化组合的预期收益，构建多头组合。以组合的超额收益为策略收益率。

5. 中证 500 增强组合

以中证 500 组合的行业分布、风格暴露为约束，在控制组合权重平方和的前提下，最大化组合的预期收益，构建多头组合。考虑到策略的投资周期为 20 个交易日，我们以每日过去 20 个交易日的模型预测值的平均值，为当日的预期收益。

我们将以上 5 个策略简称为多空、多头、头部多头、300 增强和 500 增强。

9.2. 回测流程

回测流程的目的是计算组合策略的收益率，同时要考虑组合的可交易性，即要根据市场情况模拟撮合结果，以避免策略收益可见但不可得。

以多头组合为例，流程步骤如下：

步骤一：开盘前，计算多头组合的当日日末目标权重；

步骤二：以开盘前组合市值计算当日日末目标组合，并根据开盘前所持组合，计算当日股票交易列表；此处，初始组合设定为 100 亿现金。

步骤三：模拟撮合。以 vwap 价格模拟成交，总成交量不超过当日真实成交量 10%，并返回模拟成交列表。

步骤四：根据成交列表和盘前所持组合，计算实际日末组合、收盘组合市值和收盘组合权重。

步骤五：根据上一日和当日收盘组合市值，计算当日组合收益率和超额收益率；

步骤六：重复以上环节。

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