第八届“泰迪杯”
数据挖掘挑战赛

优秀作品

作品名称：基于数据挖掘的上市公司高送转预测

荣获奖项：特等奖并获泰迪杯

作品单位：南京师范大学

作品成员：杨陆 吉小为 屠怡婷

指导老师：解锋昌

基于数据挖掘的上市公司高送转预测

摘要

“高送转”是高比例送红股或转增股本的简称，它是市场的常发事件，而且在预案公告日前一段时间有显著正的超额收益。因此投资者若能在公告前识别“高送转”事件股票，则能获得较好的投资收益。

针对问题一， 采用了 LR、RFC、SVM、XGBoost、Lightgbm 和 Catboost 六种机器学习算法去求出对上市公司实施高送转方案有较大影响的因子，并基于 AUC 指标与 k 折交叉验证和网格搜索给机器学习算法参数调优，通过评价指标 AUC 对比对效果较好的 XGBoost、LGBoost、CatBoost 三个模型，选择算法特征重要性排名前 20 个重要特征，挑选出三个模型共同确定的重要因子，从而得出问题一的中对上市公司实施高送转方案有较大影响的因子。

这 20 个特征因子为基本因子：上市年限、总资产净利率、投资支出/折旧和摊销、息税折旧摊销前利润/负债合计、最低价、最高价、收盘价；成长因子：基本每股收益、每股净资产、稀释每股收益、每股资本公积、每股收益(期末摊薄)；时序因子：基本每股收益同比增长、总资产相对年初增长、最高价下半年变异系数、收盘价上半年变异系数、收盘价下半年变异系数。

针对问题二，在预测模型的选择上本文选择了基于 Stacking 集成学习的融合分类模型，第 1 层基学习器选择 LR、RFC、SVM、XGBoost、Lightgbm 和 Catboost，第 2 层元学习器选择了 Lightgbm，从而确定了最优的 Stacking 集成学习预测模型。Stacking 集成模型在测试集上的 AUC 得分为 85.71%，高于所有基础分类器，可见建立的模型较为稳定，不存在严重过拟合且效果较好。并求出第 8 年预测结果：10.01%的上市公司选择高送转高送转，即有 347 个上市公司决定高送转，3119 个上市公司不会选择高送转。

本文利用机器学习算法，充分使用上市公司历史数据，融合了多种算法，且建立的 Stacking 集成学习预测模型较为稳定，具有较大的参考价值和现实意义。

关键词：机器学习 AUC 指标 网格搜索 高送转 Stacking 集成学习分类模型

目 录

第1章　绪论

1.1 问题背景

近年来，我国上市公司频繁实施“高送转”股利分配政策，市场反应强烈，虽然＂高送转＂概念往往与市场炒作联系，但机构、投资者以及广大散户对此趋之若鹜并且逐渐成为我国股市市场在股利分配政策方面的一种特色。

因为实施高送转后股价将做除权处理，投资者可以通过填权行情从二级市场的股票增值中获利。很多股票在公布派送预案的第二天直接涨停，而等除权后再买入可能会面临很大的回撤风险。如果我们能准确用某一年的股票相关数据预测下一年可能实施高送转的上市公司并提前买入，这对我们投资的安全性具有很大的现实意义。

经过研究，影响上市公司实施高送转的因子主要有两类：一是基本因子，包括股价、总股本、上市年限等；二是成长因子，包括每股未分配利润、每股资本公积、每股现金流、每股收益等。除此之外，还有其他因子需要研究者去挖掘。

1.2 问题重述

（1）针对 3466 支股票年数据、日数据和基础数据中给出的因子数据，根据因子自身经济学意义以及数理统计方法，筛选出对上市公司实施高送转方案有较大影响的因子。

（2）利用问题 1 中确定的因子建立模型来预测哪些上市公司可能会实施高送转，并对提供的数据，用所建立模型来预测第 8 年上市公司实施年高送转的情况。

1.3 本文主要工作与创新点

（1）对数据的预处理

对年数据而言，添加了重要特征因子：本年是否进行高送转；设定因变量为：下一年是否高送转。

对日数据而言，日数据中某些因子的变化趋势会对上市公司是否会实施高送转有影响。因此先对日数据按股票编号和年份分组求因子数据的年平均值，并计算数据中“开盘价”，“最高价”，“最低价”，“收盘价”，“成交价”，“成交量”这 6 个特征因子的上下半年分别的变异系数，表示其变化趋势。

对基础数据而言，大量资料表现股票是否是小盘、是否为次新股、是否为国

企等特征对上市公司是否会实施高送转有重要影响，在已给特征的基础上将特征因子“所属概念板块”转换为“所属概念板块个数”、“是否为次新股”、“是否为国企”、“是否为小盘”。

（2）通过数据分析筛选对上市公司实施高送转方案有较大影响的因子

本文将特征工程筛选后的因子数据，根据机器学习算法 XGBoost、CATBoost 和 lightGBM 算法中特征重要性的数值得出特征因子的重要性为前 20 的因子。

（3）机器学习算法分类预测下一年是否高送转

本文对问题一中特征工程后确定的因子，使用六种不同类型的机器学习算法去预测下一年哪些上市公司可能会实施高送转，基于AUC指标与 k 折交叉验证或网格搜索给机器学习算法参数调优，再用模型融合提高。

1.4 模型假设

（1）假设所获得的数据是真实可靠的。

（2）假设第 8 年未发生重大事件和灾难或国家未推行重要政策影响证券市场。

1.5 本文研究意义

从量化投资的角度来看，我国资本市场目前发展已经步入稳定，这就使得量化投资成为了资本投资的主要方向，不过由于经济体制以及市场化程度的发展深度，我国资本市场量化投资仍旧存在许多风险要素，必须要通过一些切实的模型预测对有可能出现的风险进行防控。

从投资风险的角度来看，高送转预测可以有效规避风险。如果我们能在上市公司正式公告“高送转”预案前，重点关注好“高送转”的真实目的，警惕上市公司出于配合二级市场炒作，或者配合大股东和高管出售股票等情况，就能获得一定的超额收益。因此，研究上市公司“高送转”行情，尤其是准确预测下一年将要实施相关方案的上市公司，对保护中小投资者的利益以及维护资本市场的稳定，具有一定的现实意义。

第2章　相关理论

2.1 高送转相关知识介绍

2.1.1 高送转的实质

“高送转”顾名思义是指高比例送股或高比例转股，例如每 10 股送股 5 股，每 10 股转增 10 股等。送股和转股比例达到多少才能称之为“高送转”，目前专家学者还未达成统一意见。本文将以每 10 股送转 5 股及以上作为界定“高送转”的选择标准。

“高送转”是通过降低每股价格从而扩大股票发行的一种行为，也相当于是一种送股行为。是通过送股并没有任何意义，其目的仅仅是来稀释股本降低股价的，也就是降低股价方便购买。将一个公司的现在股本总额和将来会收得的盈利相结合起来，来告知股东公司相关信息，同时可以让股份交易活跃起来。这个方式实质上就是企业所有者权益内部的调整，将资本从留存收益转入了其他股东权益。

2.1.2 预测下一年上市公司高送转的一些其他条件

(1) 是否真正具备“高送转”能力：本文使用送股能力公式进行运算：负债合计/资产总计。

(2) 总股本的结构在 1.5 亿以下，使用公式：未分配利润/每股未分配利润。

(3) 本年年中报表或者三季度报表披露业绩持续增长

(4) 股票是否是小盘、是否为次新股、是否为国企

(5) 上市公司本年度有没有已经高送转

2.2 机器学习算法介绍

2.2.1 LogisticRegressor

逻辑回归是机器学习中一个应用非常广泛的分类模型，它是一种分类方法，主要用于两分类问题（即输出只有两种，分别代表两个类别）。

它将数据拟合到 sigmiod 函数，其函数形式为：$g(z)=\frac{1}{1+e^{z}}$ ，寻找预测函数：$h_{\theta}(x)=g(\theta^{T}x)=\frac{1}{1+e^{\theta^{T}x}}$ ，函数的 $h_{\theta}(x)$ 值有特殊的含义，它表示结果取 1

的概率，因此对于输入 x 分类结果为类别 1 和类别 0 的概率分别为：

$$P(y=1|x;\theta)=h_\theta(x), P(y=0|x;\theta)=1-h_\theta(x)$$

构造损失函数：

$$J(\theta)=\frac{1}{m}\sum_{i=1}^{n}\mathrm{Cos}\,t(h_\theta(x^{(i)}),y^{(i)})$$

其中 $\mathrm{Cos}\,t(h_\theta(x),y)=\begin{cases}-\log(h_\theta(x)) & if \quad y=1\\ -\log(1-h_\theta(x)) & if \quad y=0\end{cases}$。

求解使得 $J(\theta)$ 函数最小并求得回归参数。

对于逻辑回归的损失函数构成的模型，可能会有些权重很大，有些权重很小，导致过拟合（就是过分拟合了训练数据），使得模型的复杂度提高，泛化能力较差（对未知数据的预测能力），正则化是结构风险最小化策略的实现，是在经验风险上加一个正则化项或惩罚项，正则项可以取不同的形式，在回归问题中取平方损失，就是参数的 L2 范数，也可以取 L1 范数。取平方损失时，模型的损失函数变为：

$$J(\theta)=\frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x_i)-y_i)^2-\lambda\sum_{i=1}^{m}{\theta_j}^2$$

其中 λ 为正则项系数，正则化后的梯度下降算法的更新变为：

$$\theta_j:=\theta_j-\frac{\alpha}{m}\sum_{i=1}^{m}(h_\theta(x_i)-y_i)x_i^j-\frac{\lambda}{m}\theta_j$$

从而完成对事件发生概率的预测。

2.2.2 RandomForestClassifier

在机器学习中，随机森林由许多的决策树组成，因为这些决策树的形成采用了随机的方法，因此也叫做随机决策树。随机森林中的树之间是没有关联的。当测试数据进入随机森林时，其实就是让每一颗决策树进行分类，最后取所有决策树中分类结果最多的那类为最终的结果。因此随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定。随机森林可以既可以处理属性为离散值的量，比如 ID3 算法，也可以处理属性为连续值的量，比

如 C4.5 算法。另外，随机森林还可以用来进行无监督学习聚类和异常点检测。

随机森林由决策树组成，决策树实际上是将空间用超平面进行划分的一种方法，每次分割的时候，都将当前的空间一分为二，比如说下面的决策树（其属性的值都是连续的实数）：

随机深林的优点：比较适合做多分类问题；训练和预测速度快；对训练数据的容错能力，是一种有效地估计缺失数据的一种方法，当数据集中有大比例的数据缺失时仍然可以保持精度不变；能够有效地处理大的数据集；可以处理没有删减的成千上万的变量；能够在分类的过程中可以生成一个泛化误差的内部无偏估计；能够检测到特征之间的相互影响以及重要性程度；不过出现过度拟合；实现简单容易并行化。当可以生成好决策树后，就比较容易生成随机森林了。下面是随机森林的构造过程：

1. 假如有 N 个样本，则有放回的随机选择 N 个样本(每次随机选择一个样本，然后返回继续选择)。这选择好了的 N 个样本用来训练一个决策树，作为决策树根节点处的样本。

2. 当每个样本有 M 个属性时，在决策树的每个节点需要分裂时，随机从这 M 个属性中选取出 m 个属性，满足条件 m << M。然后从这 m 个属性中采用某种策略（比如说信息增益）来选择 1 个属性作为该节点的分裂属性。

3. 决策树形成过程中每个节点都要按照步骤 2 来分裂（很容易理解，如果下一次该节点选出来的那一个属性是刚刚其父节点分裂时用过的属性，则该节点已经达到了叶子节点，无须继续分裂了）。一直到不能够再分裂为止。注意整个决策树形成过程中没有进行剪枝。

4. 按照步骤 1~3 建立大量的决策树，这样就构成了随机森林了。

从上面的步骤可以看出，随机森林的随机性体现在每颗数的训练样本是随机的，树中每个节点的分类属性也是随机选择的。有了这 2 个随机的保证，随机森林就不会产生过拟合的现象了。

2.2.3 SVM

SVM（支持向量机）是一种分类算法，根据输入的数据不同可做不同的模型（输入标签为分类值则用 SVC()做分类）。通过寻求结构化风险最小来提高学习机泛化

能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。通俗来讲，支持向量机是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，即支持向量机的学习策略便是间隔最大化，最终可转化为一个凸二次规划问题的求解。

支持向量机的具体求解过程如下：

(1) 设已知样本训练集：

$$T=\{(x_1,y_1),\cdots(x_n,y_n)\}\in(X\times Y)^n$$

其中，$x_i\in X=R^n, y_i\in Y=\{-1,+1\}(i=1,2,\cdots,n)$，$x_i$ 为特征向量。

(2) 选择适当核函数 $K(x_i,x_j)$ 以及参数 C，解决优化问题：

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i y_i\alpha_j y_j K(x_i,x_j)-\sum_{j=1}^{n}\alpha_j$$

$$s.t.\ \sum_{i=1}^{n}\alpha_i y_i=0,\ 0\le\alpha_i\le C,\ i=1,\cdots,n$$

得最优解：$\alpha^*=(\alpha_1^*,...,\alpha_n^*)^T$。

(3) 选取 α^* 的正分量，计算样本分类阈值：$b^*=y_i-\sum_{i=1}^{l}y_i\alpha_i^*K(x_i,x_j)$。

(4) 构造最优判别函数：

$$f(x)=\mathrm{sgn}\left[\sum_{i=1}^{n}y_i a_i^*K(x_i,x_j)+b^*\right]。$$

支持向量机内积核核函数 K 的主要种类有：

① 线性内核函数　$K(x_i,x_j)=(x_i,x_j)$

② 多项式核函数　$K(x_i,x_j)=[(x_i,x_j)+1]^q$

③ 高斯径向基核函数 (RBF)　$K(x_i,x_j)=\exp\{-\frac{\|x_i-x_j\|^2}{\sigma^2}\}$

④ 双曲正切核函数 (Sigmoid核函数)　$K(x_i,x_j)=\tanh(v(x_i\cdot x_j)+c)$

一般地，用 SVM 做分类预测时必须调整相关参数(特别是惩罚参数 c 和核函数参数 g)，这样才可以获得比较满意的预测分类精度，采用 Cross Validation 的思想可以获取最优的参数，并且有效防止过学习和欠学习状态的产生，从而能够对于测试集合的预测得到较佳的精度。

2.2.4 XGBoost

XGBoost 模型是典型 boosting 算法，是对 GBDT 模型的算法和工程改进。区别 Bagging 模型，基学习器可以并行，boosting 模型的基学习器间存在先后依赖。GBDT 是一种提升树模型，第 m 轮用一棵 acrt 回归树拟合前m－1 轮损失的负梯度，降低模型的 bias。

XGBoost 是在 GBDT 等提升算法基础上进行优化的算法，引入二阶导数信息，并加入正则项控制模型的复杂度；此外，虽然基模型的训练存在先后顺序，但每个基学习器内部的树节点分裂可以并行，XGBoost 对此进行了并行优化，实现优化目标函数以达到误差和复杂度综合最优。其原理如下：

目标函数 $L(x)$ 由误差函数 $F(x)$ 和复杂度函数 $\Omega(x)$ 组成：

$$L(x) = F(x) + \Omega(x)$$

$$L(x) = \sum_i l(y_i, \hat{y}_i) + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \left\| W_j \right\|^2$$

其中 l 是用来衡量 $\hat{y}$ 与 y 的相近程度的可导且凸的损失函数，通过每一步增加一个基分类器，贪婪地去优化目标函数，使得每次增加都使得损失变小。然后让后一次迭代的基分类器去学习前一次遗留下来的误差。这样可以得到用于评价当前分类器性能的评价函数,如下：

$$L_m(x) = \sum_i l[y_i, \hat{y}_i^{m-1} + f_m(x_i)] + \Omega(f_m)$$

这个算法又可以成为前向分步优化。为了更好更快的优化此函数，可以在 $f\_m = 0$ 附近二阶泰勒展开，泰勒展开的形式为公式。

$$f(x + \Delta x) = f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2$$

令 $g_i = \frac{\partial l(y_i, \hat{y}^{(t-1)})}{\partial \hat{y}^{(t-1)}}, h_i = \frac{\partial^2 l(y_i, \hat{y}^{(t-1)})}{\partial \left(\hat{y}^{(t-1)}\right)^2}$，最后可得到目标函数，在剔除常数项后可以得到最终的表达式，如公式所示：

$$L_m(x) = \sum_{i=1}^{n} [g_i + f_m(x_i) + \frac{1}{2} h_i f_m^{\,2}(x_i)] + \Omega(f_m)$$

2.2.5 LightGBM

LightGBM 相较于 XGBoost，提出 Histogrm 算法，对特征进行分桶，减少查询分裂节点的事件复杂度；此外，提出 Goss 算法减少小梯度数据；同时，提出 EFB 算法捆绑互斥特征，降低特征维度，减少模型复杂度。

Lightgbm 使用了如下两种解决办法：一是 GOSS，不是使用所用的样本点来计算梯度，而是对样本进行采样来计算梯度；二是 EFB，不是使用所有的特征来进行扫描获得最佳的切分点，而是将某些特征进行捆绑在一起来降低特征的维度，是寻找最佳切分点的消耗减少。这样大大的降低的处理样本的时间复杂度，但在精度上，通过大量的实验证明，在某些数据集上使用 Lightgbm 并不损失精度，甚至有时还会提升精度。下面就主要介绍这两种方法。

1、GOSS 算法描述

输入：训练数据，迭代步数 d，大梯度数据的采样率 a，小梯度数据的采样率 b，损失函数和若学习器的类型（一般为决策树）；

输出：训练好的强学习器；

（1）根据样本点的梯度的绝对值对它们进行降序排序；

（2）对排序后的结果选取前 a*100%的样本生成一个大梯度样本点的子集；

（3）对剩下的样本集合（1-a）*100%的样本，随机的选取 b*（1-a）*100%个样本点，生成一个小梯度样本点的集合；

（4）将大梯度样本和采样的小梯度样本合并；

（5）将小梯度样本乘上一个权重系数；

（6）使用上述的采样的样本，学习一个新的弱学习器；

（7）不断地重复（1）~（6）步骤直到达到规定的迭代次数或者收敛为止。

通过上面的算法可以在不改变数据分布的前提下不损失学习器精度的同时大大的减少模型学习的速率。

从上面的描述可知，当 a=0 时，GOSS 算法退化为随机采样算法；当 a=1 时，GOSS 算法变为采取整个样本的算法。在许多情况下，GOSS 算法训练出的模型精确度要高于随机采样算法。另一方面，采样也将会增加若学习器的多样性，从而潜在的提升了训练出的模型泛化能力。

2、EFB 算法描述

输入：特征 F，最大冲突数 K，图 G；

输出：特征捆绑集合 bundles；

（1）构造一个边带有权重的图，其权值对应于特征之间的总冲突；

（2）通过特征在图中的度来降序排序特征；

（3）检查有序列表中的每个特征，并将其分配给具有小冲突的现有 bundling（由控制），或创建新 bundling。

上述算法的时间复杂度为并且在模型训练之前仅仅被处理一次即可。在特征维度不是很大时，这样的复杂度是可以接受的。但是当样本维度较高时，这种方法就会特别的低效。所以对于此，作者又提出的另外一种更加高效的算法：按非零值计数排序，这类似于按度数排序，因为更多的非零值通常会导致更高的冲突概率。 这仅仅改变了上述算法的排序策略，所以只是针对上述算法将按度数排序改为按非 0 值数量排序，其他不变。

2.2.6 CATBoost

CatBoost 是一种基于对称决策树为基学习器实现的参数较少、支持类别型变量和高准确性的 GBDT 框架，主要解决的痛点是高效合理地处理类别型特征，CatBoost 是由 Categorical 和 Boosting 组成。此外，CatBoost 还解决了梯度偏差（Gradient Bias）及预测偏移（Prediction shift）的问题，从而减少过拟合的发生，进而提高算法的准确性和泛化能力。

与 XGBoost、LightGBM 相比，CatBoost 的创新点有：嵌入了自动将类别型特征处理为数值型特征的创新算法。首先对 categorical features 做一些统计，计算某个类别特征出现的频率，之后加上超参数，生成新的数值型特征。Catboost 还使用了组合类别特征，可以利用到特征之间的联系，这极大的丰富了特征维度。并采用 ordered boost 的方法避免梯度估计的偏差，进而解决预测偏移的问题，还采用了完全对称树作为基模型。

CatBoost 算法可以更好的处理 GBDT 特征中的 categorical features，基于改进 Greedy TS，它是添加先验分布项，这样可以减少噪声和低频率类别型数据对于数据分布的影响。其公式为：

$$x_{i,k}=\frac{\sum_{j=1}^{p-1}[x_{\sigma j,k}=x_{\sigma p,k}]\cdot Y_j+a\cdot p}{\sum_{j=1}^{p-1}[x_{\sigma j,k}=x_{\sigma p,k}]+a}$$

其中 p 是添加的先验项，a 通常是大于 0 的权重系数。添加先验项可以减少噪声数据。对于二分类，先验项是正例的先验概率。

CatBoost 的另外一项重要实现是将不同类别型特征的组合作为新的特征，以获得高阶依赖，为当前树构造新的分割点时，CatBoost 会采用贪婪的策略考虑组合。对于树的第一次分割，不考虑任何组合。对于下一个分割，CatBoost 将当前树的所有组合、类别型特征与数据集中的所有类别型特征相结合，并将新的组合类别型特征动态地转换为数值型特征。

第 3 章 数据预处理及因子筛选

3.1 数据的选取

本文通过 Jupyter 的 Python 接口读取本题提供的数据，对提供的三类数据本文都将采用。

对于年数据而言，读取数据获得了 3466 只股票 7 年的 360 个特征因子，其中‘每股送转’，与‘是否高送转’两个与高送转方案直接相关的数据，将特征因子‘是否高送转’改名为‘本年是否高送转’，用机器学习算法解决二分类问题。再根据特征类型删去特征类型为‘object’的五个与‘是否高送转’无关的特征：'会计准则','货币代码','高转送预案公告日','高转送股权登记日','高转送除权日'，并删去所有特征因子的数据均为空值的对应股票的某年的数据。

对于日数据而言，去除日数据删除整列全为缺失值的特征因子“120 日信息比率(InformationRatio120)”。

对于基础数据，将全部读取进行下一步的特征创造与转换。

3.2 特征创造及转换

在年数据中，由于本题要求是预测上市公司下一年是否实施高送转方案，因此创造因变量‘下一年是否高送转’，创造新的特征因子“总股本”：未分配利润/每股未分配利润、新的特征因子“送股能力”：负债合计/资产总计和特征因子“本年是否高转送”，从而探索特征因子对上市公司下一年是否实施高送转方案的影响。

在日数据中，在第一步数据选取的日数据参与建模，为了将日数据转换为年数据，探究日数据中特征因子每年对上市公司实施高送转决策的影响，因此计算每只股票各指标的年平均值作为年指标值。且查阅资料得知，数据中“开盘价”，“最高价”，“最低价”，“收盘价”，“成交价”，“成交量”这6个特征因子的变化趋势会对上市公司是否会实施高送转有影响。因此我们并按股票编号和年份分组求这6个特征因子的上、下半年的变异系数，表示其变化趋势。

在基础数据中，大量资料表现股票是否是小盘、是否为次新股、是否为国企等特征对上市公司是否会实施高送转有重要影响，在已给特征的基础上将特征因子“所属概念板块”转换为“所属概念板块个数”、“是否为次新股”、“是否为国

企”、“是否为小盘”。

3.3 特殊数据的处理

3.3.1 异常值的处理

本文中的数据来自于 3466 只股票，数据分布较为离散，使用箱式图异常值处理不合理，以特征因子“固定资产合计”和“基本每股收益”为例做箱式图，如图 1 所示，无法判断异常值。因此不采用箱式图来查看异常值，而是将部分异常值 0 更改为缺失值，之后再填充缺失值。

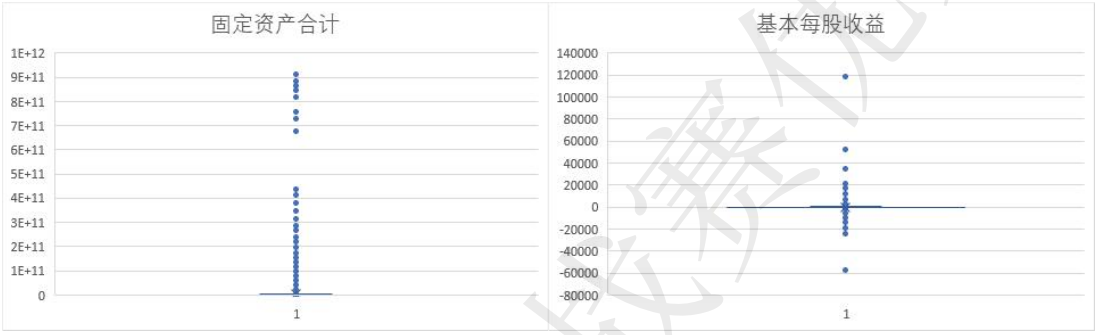

图 1 异常值分布

3.3.2 缺失值的处理

3.3.2.1 处理缺失值的几种方法

1、如果某个特征因子缺失的样本占总数极大，我们可能就直接舍弃了；因为如果作为特征加入的话，可能反倒带入噪音，影响最后的结果。

2、如果某个特征因子缺失的样本适中，且为数值型特征属性，我们可以用 0、平均数或众数进行缺失值填补。

3、如果某个特征因子缺失的样本较少，使用随机森林填补缺失值，随机森林（Random Forest）作为一种比较新的机器学习方法，是一个包含多个决策树的分类器， 并且其输出的类别是由个别树输出的类别的众数而定，在运算量没有显著增加的前提下提高了预测精度，同时其运算结果对缺失数据和非平衡的数据能达到相当稳健的水平。随机森林填补缺失值基本步骤：

1) 将数据中所有有缺失值的列提取建立模型，并按列缺失值个数从小到大排序。

2) 将缺失值个数最小的一列提取，其他列的缺失值用 0 填补。

3) 使用随机森林回归模型填补这一列的缺失值。

4) 重复 2、3 步骤，直到所有缺失值被填补，得到完整数据。

3.3.2.2 本文处理缺失值的过程

通过探索数据发现，日数据与年数据均含有缺失值。由于日数据的样本总体数据相对较少，我们直接使用随机森林填补上一步特征创造与转换后的缺失值。

但年数据样本总数较大，因此将缺失值占比大于 50%的特征因子所在的股票数据全部删除；将缺失值占比在 20%至 50%的特征因子所在的股票数据全部使用均值填补；将缺失值占比小于 20%特征因子所在的股票数据用随机森林填补。

这样不仅能降低噪音的影响而且降低运算复杂度提高正确性，且缺失值得到了较为准确的数值填补。

3.3.2 分类型特征的处理

通过探索数据发现，基础数据中的特征因子“所属行业”的取值之间没有联系且没有大小之分，不能互相计算，需要使用哑变量的方式处理。这里使用独热编码，将“所属行业”中的 18 个特征：房地产业， 制造业，批发和零售业， 租赁和商务服务业，综合，信息传输、软件和信息技术服务业，文化、体育和娱乐业， 建筑业，电力、热力、燃气及水生产和供应业，卫生和社会工作， 采矿业，科学研究和技术服务业，交通运输、仓储和邮政业，农、林、牧、渔业，水利、环境和公共设施管理业，金融业，住宿和餐饮业和教育都转换为哑变量。

3.4 数据合并

最后将数据处理后的基础数据、日数据、年数据按股票编号、年份‘Groupby’合并，进行下一步特征选择。

3.5 特征选择

从所有特征中，选择出有意义、对模型有帮助的特征，以避免必须将所有特征都导入模型去训练的情况。

3.5.1 Filter 过滤法

过滤方法通常用作预处理步骤，特征选择完全独立于任何机器学习算法。是根据各种统计检验中的分数以及相关性的各项指标来选择特征。

1.方差过滤

方差过滤通过特征本身的方差来筛选特征的类。比如一个特征本身的方差很小，就表示样本在这个特征上基本没有差异，可能特征中的大多数值都一样，甚至整个特征的取值都相同，那这个特征对于样本区分没有什么作用。所以本文先消除方差为 0 的特征。

2.相关性分析：互信息法

方差挑选之后考虑相关性，为选出与标签相关且有意义的特征，因为这样的特征能够为我们建立模型提供大量信息。如果特征与标签无关，可能会给模型带来噪音。本文中使用互信息法评判特征与标签之间的相关性。

互信息法是用来捕捉每个特征与标签之间的任意关系（包括线性和非线性关系）的过滤方法。互信息法可以找出任意关系。互信息法不返回 p 值或 F 值类似的统计量，它返回“每个特征与目标之间的互信息量的估计”，这个估计量在[0,1]之间取值，为 0 则表示两个变量独立，为 1 则表示两个变量完全相关。

3.5.2 基于 LinearSVC 算法的嵌入法

嵌入法是一种让算法自己决定使用哪些特征的方法，即特征选择和算法训练同时进行。在使用嵌入法时，先使用某些机器学习的算法和模型进行训练，得到各个特征的权值系数，根据权值系数从大到小选择特征。这些权值系数往往代表了特征对于模型的某种贡献或某种重要性，比如决策树和树的集成模型中的 feature_importances_属性，可以列出各个特征对树的建立的贡献，我们就可以基于这种贡献的评估，找出对模型建立最有用的特征。

因此相比于过滤法，嵌入法的结果会更加精确到模型的效用本身，对于提高模型效力有更好的效果。并且，由于考虑特征对模型的贡献，因此无关的特征（需要相关性过滤的特征）和无区分度的特征（需要方差过滤的特征）都会因为缺乏对模型的贡献而被删除掉。

本文使用 LinearSVC 算法的嵌入法，LinearSVC 算法的 L1 正则化将系数 w 的 l1 范数作为惩罚项加到损失函数上，由于正则项非零，这就迫使那些弱的特征所对应的系数变成 0。因此 L1 正则化往往会使学到的模型很稀疏（系数 w 经常为 0），这个特性使得 L1 正则化成为一种很好的特征选择方法。

第 4 章 基于机器学习模型的问题一研究

4.1 模型的构建

4.1.1 测试集、训练集的划分

将特征选择后的特征因子和因变量‘下一年是否高送转’，通过 Jupyter 的 Python 接口读入数据。数据包括三个部分：年数据、日数据、基础数据：

表 1 数据结构

	特征数量	观测行	备注
年数据	362	24262	3346 只股票 7 年年度数据
日数据	61	5899132	3346 只股票 7 年日度数据
基础数据	4	3466	3346 只股票

使用 Sklearn 中 model_selection 的 train_test_split 模型人为划分训练集与测试集，确定训练集与测试集的比例为 7：3。

4.1.2 数据标准化

在机器学习算法中，再将有着将不同规格的数据转换到同一规格，或不同分布的数据转换到某个特定分布的需求，即将数据“无量纲化”。在逻辑回归，支持向量机等中，无量纲化可以加快求解速度。

本文使用数据标准化（Standardization）将数据“无量纲化”，即将数据按均值中心化后，再按标准差缩放，数据就会服从为均值为 0，方差为 1 的正态分布（即标准正态分布）。其计算公式为：

$$x^{*}=\frac{x-\mu}{\sigma}$$

4.1.3 模型评价指标

机器学习算法评价指标有很多种，这里首先观察准确率(accuracy)、召回率(recall)和曲线下面积（AUC）的值来评价本文使用的六种机器学习算法。

准确率(accuracy)是正确预测数占总样本数的比值，只有当属于每个类的样本数量相等时才有效，在正负样本不平衡的情况下，准确率这个评价指标有很大的缺陷。

召回率(recall)是正确分类的正例个数占实际正例个数的比例也称查全率。

AUC(Area Under Curve)是机器学习二分类模型中非常常用的评估指标，分类器的 AUC 等价于随机选择正样本高于随机选择负样本的概率，相比于 F1-Score 对项目的不平衡有更大的容忍性，目前常见的机器学习库中(比如 scikit-learn)一般也都是集成该指标的计算。

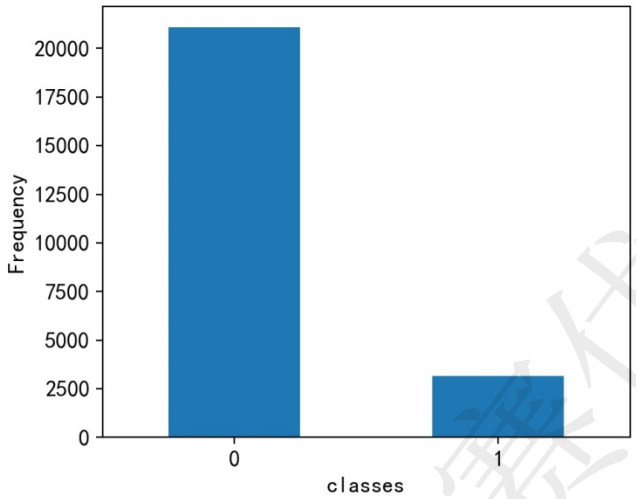

图 2 样本分布

通过对样本特征因子“本年是否高送转”的分布去观察因变量的分布，如图 2 所示，直观可见正负样本不平衡，因此本文主要根据和曲线下面积（AUC）的值评价模型。AUC 越接近 1.0，检测方法真实性越高。

4.2 模型参数调优与模型重要特征

4.2.1 参数调优概念及方法

机器学习模型参数众多，参数选择不恰当，就会出现欠拟合或者过拟合的问题。为了提高模型的精度，同时提升模型的泛化能力，调参过程不可缺少。而在选择超参数的时候，有两个途径，一个是凭经验微调，另一个就是选择不同大小的参数，带入模型中，挑选表现最好的参数。

本文主要选择网格搜索，网格搜索是一种重要调参手段，即穷举搜索：在所有候选的参数选择中，通过循环遍历，尝试每一种可能性，表现最好的参数就是最终的结果。其原理就像是在数组里找到最大值. 网格搜索可以保证在指定的参数范围内找到精度最高的参数。

4.2.2 各个模型参数调优

建模时先固定每个参数的初始值，再设定其调参范围，进行网格搜索和交叉验证寻找最优化结果。其中设置的初始值、范围和调参结果见各算法框架参数结果详情表，本文模型优化评价指标设为和曲线下面积（AUC）。

表 2　LR 调参过程

参数名称	初始值	调参范围	调参结果	调参后 AUC
penalty	l1	l1、l2	l1	81.84%
C	1	[2.6,5]	5	82.50%

逻辑回归模型需要调整的参数有两个，分别为 penalty 和 C，penalty 表示正则化的方式，C 表示正则化强度的倒数，其默认值为 1，即默认正则项与损失函数的比值是 1：1。C 越小，损失函数会越小，模型对损失函数的惩罚越重，正则化的效果越强。

从 AUC 的结果 82.50%不难看出模型拟合的效果一般，难以将其作为最优算法预测，继续探索其他模型。

表 3　RFC 调参过程

参数名称	初始值	调参范围	结果	调参后 AUC
n_estinators	500	[400,500,600,700,800,900,1000]	650	82.55%
max_depth	8	[2,7,9,11,13,15,17]	15	83.10%
min_samples_split	100	[10,30,50,70,90,110,130]	70	83.10%
min_samples_leaf	20	[10,20,30,40,50,60,70,80,90,100]	20	83.10%

随机森林调整的参数有四个，分别为 n_estinators：随机森林中树模型的数量、max_depth：树的最大深度、min_samples_split：中间节点要分枝所需要的最小样本数和 min_samples_split：叶节点要存在所需要的最小样本数。

该模型的 AUC 结果有 83.10%较 LR 算法高 3%，提升较高，继续探索下列模型。

表 4　SVM 调参过程

参数名称	初始值	调参范围	调参结果	调参后 AUC
kernel	linear	'linear','poly','sigmoid','rbf'	rbf	83.19%
C	0	Np.linespace(0.744898,1.389796,8)	1.205539	83.46%

SVM 需要调整的参数也有两个，分别为 kernel 和 C，kernel 表示算法中采用的核函数类型，可选参数有：‘linear’:线性核函数，‘poly’：多项式核函数，‘rbf’：径像核函数/高斯‘sigmod’：核函数。C 表示错误项的惩罚系数。C 越大，即对分错样本的惩罚程度越大，因此在训练样本中准确率越高，但是泛化能力降低，也就是对测试数据的分类准确率降低。相反，减小 C 的话，容许训练样本中有一些误分类错误样本，泛化能力强。SVM 调参后 AUC 有 83.46%较上述两个模型效果较好，但拟合程度依旧不高，因而下面选用复杂程度更高的集成算法建模调参。

表 5　XGBoost 调参过程

参数名称	初始值	调参范围	结果	调参后 AUC
n_estinators	500	[700,725,750,775,800,825]	750	84.35%
Min_child_weight	1	[1, 2, 3, 4, 5]	2	84.91%
Max_depth	5	[3, 4, 5, 6, 7, 8]	8	84.91%
gamma	0.6	[0.2, 0.3, 0.4, 0.5, 0.6,0.7]	0.7	85.04%
subsample	0.8	[0.6, 0.7, 0.8, 0.9]	0.8	85.04%
Colsample_bytree	0.8	[0.6, 0.7, 0.8, 0.9]	0.8	85.04%
Reg_alpha	1	[0,0.03,0.05, 0.1,1,2]	0	85.20%
Reg_lambda	0	[0.05,0.1,1,2,3]	2	85.20%
Learning_rate	0.1	[0.01,0.03, 0.05, 0.07, 0.1,0.15,0.2]	0.01	85.36%

XGBoost 算法调整第一个参数是 n_estimators，这个参数非常强大，该参数越大，模型的学习能力就会越强；下面只介绍该模型中几个相对重要参数：参数

subsample 表示随机抽样的时候抽取的样本比例，范围是(0,1]；参数 Learning_rate 表示集成中的学习率，又称为步长以控制迭代速率，常用于防止过拟合。默认是 0.1，取值范围[0,1]。

XGBoost 算法调参后 AUC 有 85.36%，AUC 很接近 1.0，检测方法真实性高，表明模型有较好的拟合效果。

为得出对上市公司实施高送转方案有较大影响的因子，我们求出在XGBoost算法中特征重要性为前20的特征因子探究问题1的结论。将特征因子对算法的重要性从大到小排列，其特征因子和其重要性数值如下图所示。

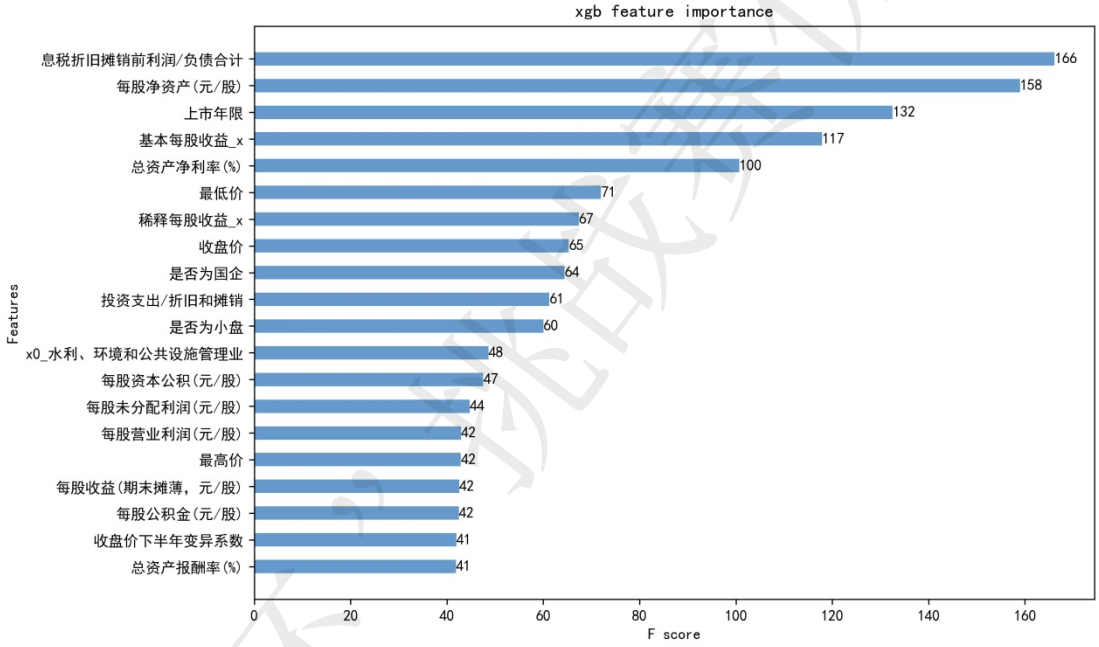

图3 xgboost前20重要特征

分别为： 息税折旧摊销前利润/负债合计、每股净资产、上市年限、基本每股收益_x、总资产净利率、最低价、稀释每股收益_x、收盘价、是否为国企、投资支出/折旧和摊销、是否为小盘、x0_水利、环境和公共设施管理业、每股资本公积、每股未分配利润、每股营业利润(元/股)、最高价、每股收益(期末摊薄、每股公积金、收盘价下半年变异系数、总资产报酬率。

表 6 LightGBM 调参过程

参数名称	初始值	调参范围	结果	调参后 AUC
max_depth	500	[10,11,12,13,14,15,16,17]	11	85.40%
num_leaves	30	[20,30,40,50,60,70,80,90,100]	30	85.40%
min_data_in_leaf	1	[1, 16, 31, 46, 61, 76, 91]	61	85.78%
max_bin	255	range(5,255,10)	195	85.78%
feature_fraction	1	[0.6,0.7,0.8,0.9,1.0]	0.8	85.78%
bagging_fraction	1	[0.6,0.7,0.8,0.9,1.0]	0.6	85.78%
lambda_l1	0	[1e-5,1e-3,1e-1,0.0,0.1,0.3,0.5,0.7,0.9,1.0]	1e-5	85.78%
lambda_l2	0	[1e-5,1e-3,1e-1,0.0,0.1,0.3,0.5,0.7,0.9,1.0]	1e-5	85.78%
Learning_rate	0.1	[0.01,0.02,0.03,0.04,0.05,0.06,0.07,0.08,0.09,0.1]	0.01	85.82%
n_estimators	67	[450,500,550,600,650,700,750]	650	85.82%

LightGBM 的基本调参过程如下：首先选择较高的学习率，大概 0.1 附近，这样是为了加快收敛的速度。这对于调参是很有必要的。其次是对决策树基本参数调参，最后是正则化参数调参。因此，第一步先确定学习率和迭代次数，第二步，确定 max_depth 和 num_leaves，这是提高精确度的最重要的参数。第三步，确定 min_data_in_leaf 和 max_bin。第四步，确定 feature_fraction、bagging_fraction、bagging_freq。第五步，确定 lambda_l1 和 lambda_l2。第六步，确定 min_split_gain。第七步，降低学习率，增加迭代次数，验证模型。

Lightgbm 算法调参后 AUC 有 85.82%，AUC 很接近 1.0，检测方法真实性高，表明模型有较好的拟合效果。

为得出对上市公司实施高送转方案有较大影响的因子，我们求出在 Lightgbm 算法中特征重要性为前 20 的特征因子探究问题 1 的结论。将特征因子对算法的重要性从大到小排列，其特征因子和其重要性数值如下图所示。

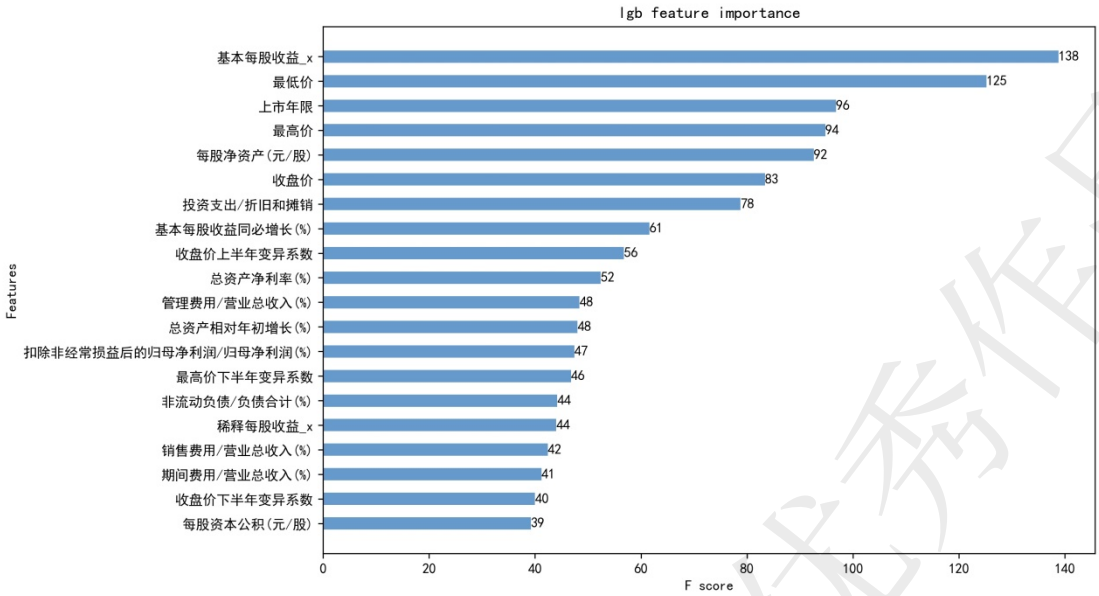

图 4 lightgbm 前 20 重要特征

分别为： 基本每股收益_x、最低价、上市年限、最高价、每股净资产、收盘价、投资支出/折旧和摊销、基本每股收益同比增长、收盘价上半年变异系数、总资产净利率、管理费用/营业总收入、总资产相对年初增长、扣除非经常损益后的归母净利润/归母净利润、最高价下半年变异系数、非流动负债/负债合计、稀释每股收益_x、销售费用/营业总收入、期间费用/营业总收入、收盘价下半年变异系数、每股资本公积。

表 7　CatBoost 调参过程

参数名称	初始值	调参范围	结果	调参后 AUC
Learning_rate	0.03	[0.01, 0.03, 0.05, 0.07, 1]	0.05	85.45%
Depth	6	[4, 5, 6, 7, 8, 9]	6	85.61%
Bagging_temperature	0.83	[0.6,0.7,0.83,0.9,1]	0.6	85.65%

CatBoost 算法调整参数 Learning_rate、Depth 和 Bagging_temperature。Learning_rate 不再阐述，参数 Depth 表示树的深度，bagging_temperature 表示贝叶斯套袋控制强度，区间[0, 1]，默认为 1。

CatBoost 算法调参后 AUC 有 85.65%，也很接近 1.0，算法真实性高，表明模型有较好的拟合效果。

这里也求出在CatBoost算法中 特征重要性为前20的特征因子探究问题1的结论。将特征因子对算法的重要性从大到小排列，其特征因子和其重要性数值如下图所示。

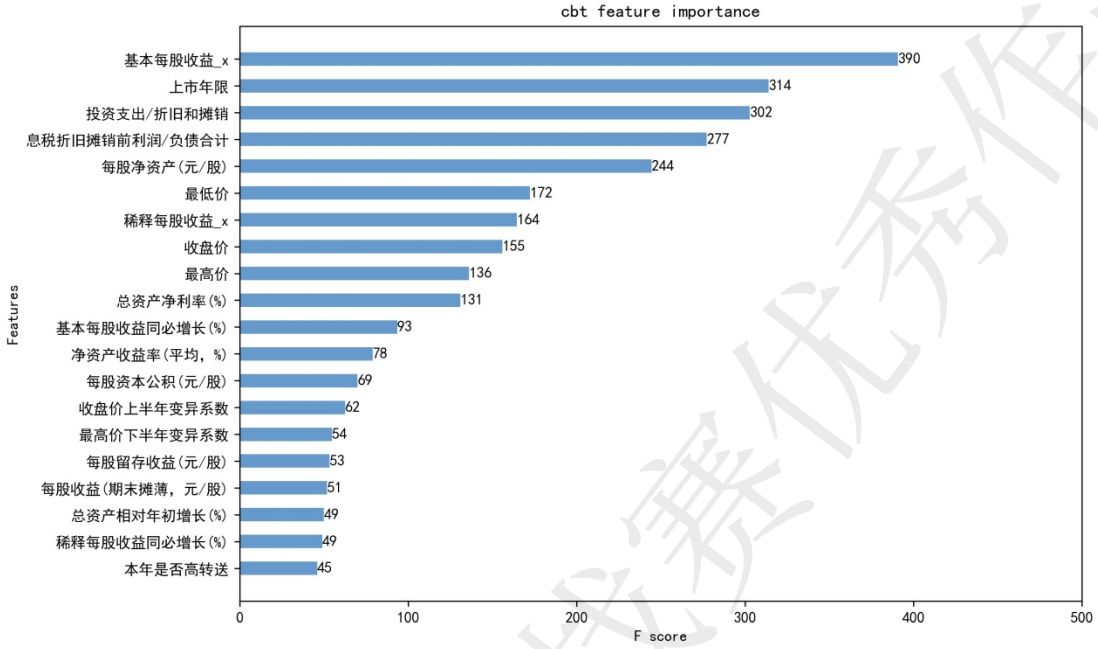

图 5 catboost 前 20 重要特征

分别是：基本每股收益_x、上市年限、投资支出/折旧和摊销、息税折旧摊销前利润/负债合计、每股净资产、最低价、稀释每股收益_x、收盘价、最高价、总资产净利率、基本每股收益同比增长、净资产收益率(平均、每股资本公积、收盘价上半年变异系数、最高价下半年变异系数、每股留存收益、每股收益(期末摊薄、总资产相对年初增长、稀释每股收益同比增长、本年是否高转送。

4.3 确定对决策影响较大的因子

本文在 7 个模型确定最优参数之后，在测试集上进行预测，模型训练结束后，选择效果较好的 XGBoost、LGBoost、CatBoost 三个模型所得特征重要性，由于特征变量较多，我们已经选择了这三个模型排名前 20 个重要特征。在此基础上，挑选出三个模型共同确定的重要因子，并进行分类，得到结果如下：

重要影响因子

基本因子	上市年限
	总资产净利率
	投资支出/折旧和摊销
	息税折旧摊销前利润/负债合计
	最低价
	最高价
	收盘价
成长因子	基本每股收益
	每股净资产
	稀释每股收益
	每股资本公积
	每股收益(期末摊薄)
时序因子	基本每股收益同比增长
	总资产相对年初增长
	最高价下半年变异系数
	收盘价上半年变异系数
	收盘价下半年变异系数

可以看到，影响因子分为基本因子、成长因子、时序因子三个部分。其中基本因子包括上市年限、总资产净利率、投资支出/折旧和摊销、息税折旧摊销前利润/负债合计、最低价、最高价、收盘价。一般而言，市场投资者对于上市年限长的公司关注程度高于上市时间较短的公司。另一方面，上市年限长的公司相对资金充裕、经营稳定、核心竞争力强。总资产净利率是衡量上市公司盈利能力的重要指标。净利率越高，说明投资带来的收益越高。

成长因子中包含的特征较多，包括基本每股收益_x、每股净资产、稀释每股收益_x、每股资本公积和每股收益(期末摊薄)是公司未来可扩大再生产或是可分配的重要物质指标。这些特征数值越大，表明该公司盈利能力越强，意味着该公

司未来高转送的概率越大。每股收益_x、每股净资产、稀释每股收益_x、反应企业的经营成果、盈利能力和成长潜力，投资者可以通过这些指标来衡量获利水平以及投资风险。每股净资产反应了每股股票代表公司净资产价值，为支撑股票市场价格的重要基础。每股净资产值越大，表明企业财富越雄厚，盈利能力和抵御外来因素影响能力越强。

时序因子包含基本每股收益同比增长、总资产相对年初增长、最高价下半年变异系数、收盘价上半年变异系数、收盘价下半年变异系数。高送转不会增加上市公司的价值，但是高转送前因为投资者的规避风险等操作，会有投资者大量买入卖出，因此成交量变异系数也是预测高转送的重要指标。

第 5 章 基于多种算法问题二的研究

5.1 基于模型融合的预测模型构造

5.1.1 模型选择

本文主要选择在测试集上得出的 AUC 作为评价各分类算法的指标，因此下面将六个算法调参前后各模型的测试集上得出的 AUC 对比，如表 8 所示。

表 8　各模型 AUC 数值

算法	调参前	调参后
LR	81.84%	82.07%
RFC	81.20%	83.10%
SVM	83.19%	83.46%
XGBOOST	83.40%	85.25%
LIGHTGBM	85.25%	85.57%
CATBOOST	85.18%	85.65%

将调参后 AUC 的值从小到大排序，依次是 LR 82.07%、RFC 83.10%、SVM 83.46%、XGBoost 85.25%、Lightgbm85.57% 和 Catboost 85.65%、，可见发现 Xgboost、Catboost、Lightgbm 这三中对 GBDT 的优化实现的算法效果远好于其他算法。

为增加对比效果，再使用 ROC 曲线观察各模型之间的优劣，ROC 曲线图是反映敏感性与特异性之间关系的曲线。横坐标 X 轴为 1-特异性，也称为假阳性率（误报率），X 轴越接近零准确率越高；纵坐标 Y 轴称为敏感度，也称为真阳性率（敏感度），Y 轴越大代表准确率越好。根据曲线位置，把整个图划分成了两部分，曲线下方部分的面积被称为 AUC（Area Under Curve），用来表示预测准确性，AUC 值越高，也就是曲线下方面积越大，说明预测准确率越高。曲线越接近左上角（X 越小，Y 越大），预测准确率越高。

ROC 曲线如图 6 所示。

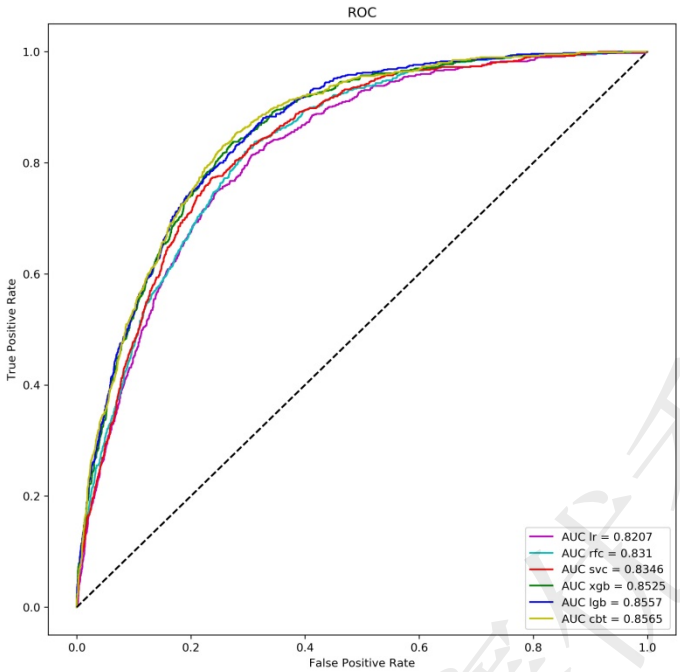

图 6 ROC 曲线

不难发现 Xgboost、Catboost、Lightgbm 这三中对 GBDT 的优化实现的算法效果远好于其他算法，其中 Catboost 比 XGBoost、LightGBM 表现的更为优秀。本文为了更好的提升模型的预测准确率和泛化能力，使用 Stacking 模型融合，将这几个学习器的预测结果作为新的训练集，去学习一个新的学习器。

5.1.2 模型融合的介绍

本文主要使用 Stacking 方法进行模型融合，在介绍 Stacking 模型融合方法前，首先需要了解集成学习这个概念。集成学习(ensemble learning)指的并非是某一个特定的机器学习算法，而是结合多个机器学习算法来构建模型去完成学习目标，也就是常说的“博采众长”。

集成学习主要包括： Bagging、Boosting、Stacking，它们分别使用并行、串行、树行的计算方法。本文用到的 Stacking 模型融合方法，即为使用树行计算方法的集成学习方法。

由于人解决问题的思维是树形的，将模型树行化符合问题本身的逻辑，精确率和召回率呈稳态正相关。因此采用树行计算方法的 Stacking 方法可以整合不同模型的最好表现，使模型融合更加科学化，用以提升模型的预测准确率和泛化

能力。

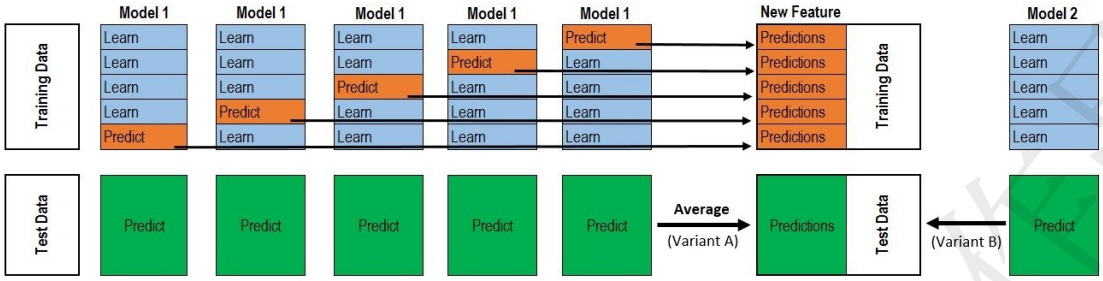

图 7 Stacking 融合模型过程图示

Stacking 融合模型一般分为 2 层内容。第 1 层模型主要用于产生第 2 层模型的训练集数据(Training Data)。产生过程如下：首先，训练集数据内容依图？所示，是用一个基础模型进行 k 折交叉验证的结果。k 折交叉验证，就是先拿出 k-1 折作为训练数据，另外一折作为测试数据(Testing Data)。每一个交叉验证产生的预测结果组合起来，作为第 2 层模型的训练集数据。另外，还要对数据集原来的整个训练数据进行预测，这个过程会生成 k 个预测数据集，对于这部分数据，本文将数据集各部分相加取平均作为下一层模型的测试集数据。第 2 层学习模型采用非线性模型，通过将第 1 层模型输出的结果作为训练数据训练模型，得到新的预测结果。通过将新的预测结果和第 2 层模型的测试数据集进行对比，观察预测准确度。

Stacking 集成学习算法的效果好坏取决于两个方面: 一个是基分类器的预测效果, 通常 基分类器的预测效果越好, 集成学习模型的预测效果越好; 另一个是基分类器之间需要有一定的差异性, 因为每个模型的主要关注点不同, 这样集成才能使每个基学习器充分发挥其优点。

本文基于经典的 Stacking 模型融合方法进行改进：

1. 将其每一步验证中使用的单个相同模型，改为 6 个不同的机器学习模型进行预测。
2. 使用第 1 层所有基分类器所产生的类别概率值作为第 2 层模型的输入。

5.1.3 模型融合过程

本文充分考虑了决定 Stacking 集成学习模型效果好坏的两个方面: 一是选择

学习能力较强的基学习器；二是充分考虑基学习器之间的差异性。LR 是解决工业规模问题最流行的算法，对于数据中小噪声的鲁棒性很好，并且不会受到轻微的多重共线性的特别影响。SVM 在解决非线性数据集的分类和回归中具有非常好的效果。XGBoost、Lightgbm、Catboost 是集成学习 Boosting 中泛化能力和学习能力较强的算法。6 种算法不仅有充分的理论支撑，而且在科学研究中正扮演着重要的角色。第 2 层元学习器同样选择学习能力较强的 Lightgbm 算法，用于对第 1 层基学习器的集成，并且使用 7 折交叉验证划分数据的方式防止过拟合的发生。

综上所述，本文基于 Stacking 集成学习的分类模型第 1 层基学习器选择 LR、RFC、SVM、XGBoost、Lightgbm、Catboost，第 2 层元学习器选择 lightgbm，模型结构如图所示。

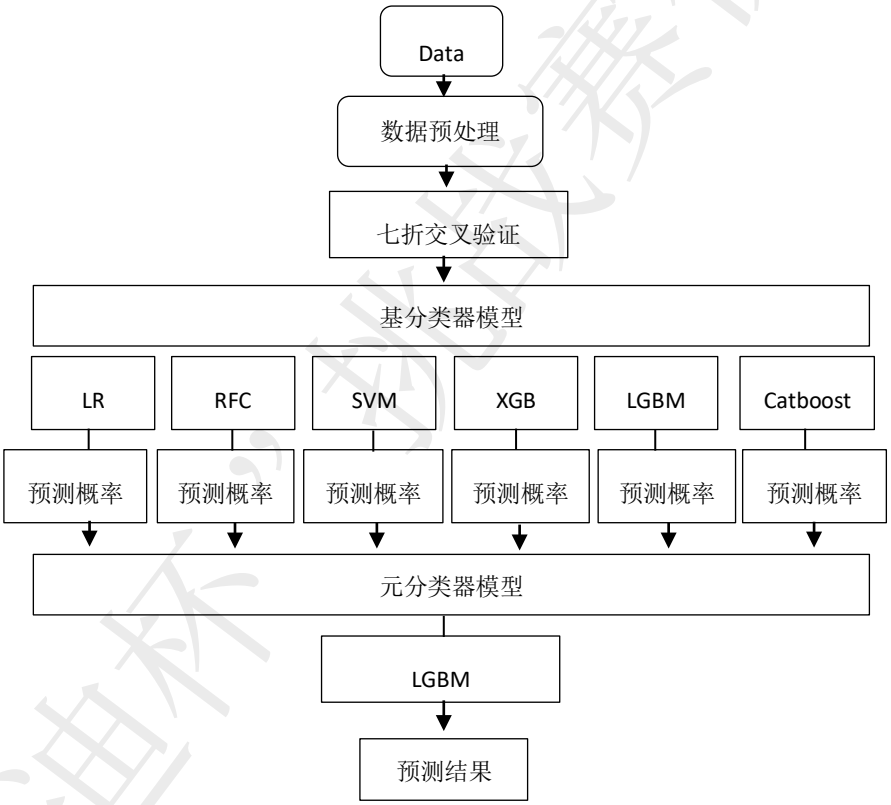

图 8 基于 Stacking 集成学习的预测模型

最终结果如上图所示：stacking 集成模型在测试集上的得分高于所有基础分类器。并且 Stacking 集成模型在测试集上的 auc 为 85.71%，说明模型较为稳定，不存在严重过拟合。

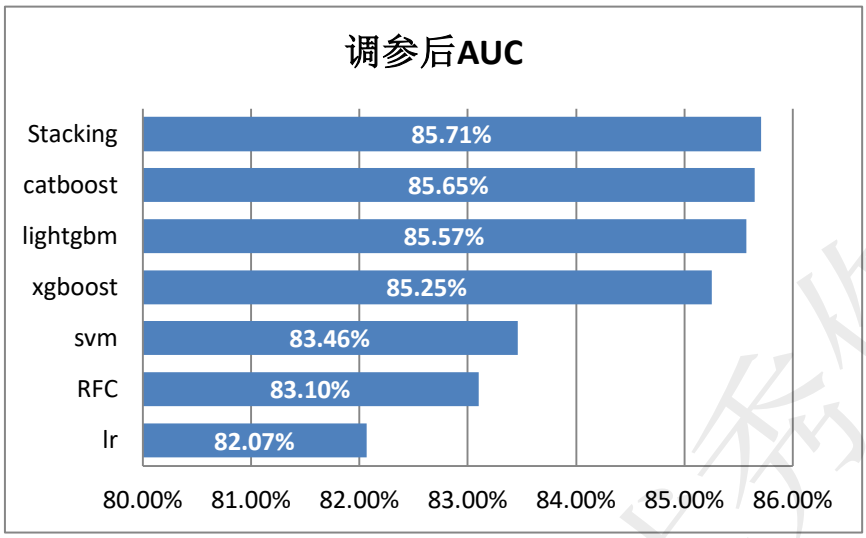

图 9 Stacking 集成学习模型与其他模型 AUC 对比

由此可见，基于 Stacking 集成学习模型综合了各个基分类器的优点后，并充分发挥了集成模型的性能，因此具有更强的泛化能力和更好的预测效果。

5.2 基于融合模型的预测第八年的决策结果

通过对下一年上市公司是否高送转的 Stacking 集成学习融合模型的建立，使用第 7 年年数据、日数据和基础数据数据预处理并特征选择后的合并数据中的特征因子去预测 3466 只股票第 8 年上市公司是否决定高送转。

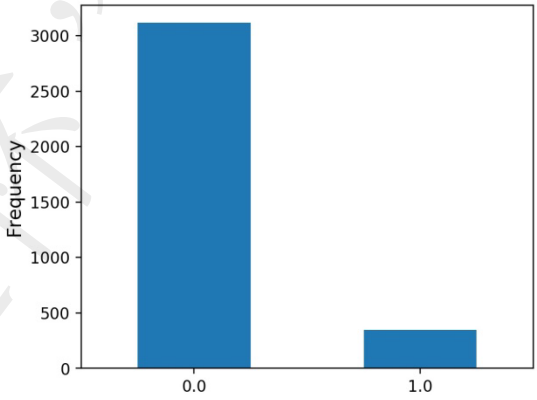

图 10 第 8 年上市公司是否高送转直方图

得到预测结果：存在 10.01%只股票高送转，即有 347 个上市公司决定高送转，3119 个上市公司不会高送转。为方便观察，绘制第 8 年上市公司是否决定高送转的 0-1 决策直方图，0 表示不执行高送转，1 表示执行高送转，如图 10 所示。

第 6 章 总结

本文通过结合影响上市公司下一年是否高送转决策实际情况与机器学习算法进行建模预测，主要目的是构建最优的分类模型预测上市公司下一年是否实施高送转。最终得到以下结论：

一、本文将年数据、日数据、基础数据按股票编号和年分组合并，将合并后特征处理的整体数据放入机器学习算法中预测上市公司下一年是否实施高送转。通过对模型训练集 AUC 的计算，模型得到了较好的预测能力，可见将三类数据按本文的方法合并的方法有一定的合理性与正确性。

二、在使用 LR、RFC、SVM、XGBoost、Lightgbm 和 Catboost 六种机器学习模型的评价指标训练集 AUC 对比，不难发现 Xgboost、Catboost、Lightgbm 这三种算法效果远好于其他算法。对于效果较好的 XGBoost、LGBoost、CatBoost 这三个模型，选择特征重要性排名前 20 个重要特征，挑选出三个模型共同确定的重要因子作为问题一的结果。

三、问题一的特征因子为基本因子：上市年限、总资产净利率、投资支出/折旧和摊销、息税折旧摊销前利润/负债合计、最低价、最高价、收盘价；成长因子：基本每股收益、每股净资产、稀释每股收益、每股资本公积、每股收益(期末摊薄)；时序因子：基本每股收益同比增长、总资产相对年初增长、最高价下半年变异系数、收盘价上半年变异系数、收盘价下半年变异系数。

四、在预测模型的选择上，本文基于 Stacking 集成学习的分类模型，第 1 层基学习器选择 LR、RFC、SVM、XGBoost、Lightgbm、Catboost, 第 2 层元学习器选择了 lightgbm，从而确定了最终最优的预测模型。Stacking 集成模型在测试集上的得分高于所有基础分类器，Stacking 集成模型在测试集上的得分高于所有基础分类器，其 auc 得分为 85.71%，可见建立的模型较为稳定，不存在严重过拟合且效果较好。

五、预测第 8 年上市公司是否高送转的结果：10.01%的上市公司选择高送转高送转，即有 347 个上市公司决定高送转，3119 个上市公司不会高送转。

综上所述，本文采用的策略与方法有一定的准确性和现实意义，可以作为预测上市公司下一年是否实施高送转方案的有效模型。

参考文献

[1]邢小艳.基于模式识别的“高送转”投资策略研究[D].广东:华南理工大学.2016:8-10.

[2]关怡婕.上市公司高送转的动因与财务效省略研究[J].Value Engineering. 2020:13-15.

[3]赵爱玲.上市公司高派现高送转股利政策影响因素研究[J].财会月刊.2019(17):36-38.

[4]王悦.上市公司高送转的影响因素分析[J].中外企业家. 2019(29):15.

[5]陈健,宋文达.量化投资的特点、策略和发展研究[J].时代金融. 2016(29):245-247.

[6]白一池.资本市场量化投资策略和风控措施探析[J].金融视线. 2020(05):118.

[7]周志华.机器学习 [M].清华大学出版社.2016.

[8]徐慧丽.基于随机森林的多阶段集成学习方法[J].高师理科学刊,2018(2).

[9]曹正凤.随机森林算法优化研究[D].北京:首都经济贸易大学,2014.

[10]林晓明.华泰人工智能系列之三:人工智能选股之支持向量机模型[R].广东:华泰证券研究所,2017.

[11]林晓明.华泰人工智能系列之六:人工智能选股之 Boosting 模型[R].广东:华泰证券研究所,2017.

[12]Guolin Ke.LightGBM: A Highly Efficient Gradient BoostingDecision Tree[J].2018.

[13]Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system[J]. In Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,2016:785 - 794.

[14] Lightgbm基本原理介绍[DB/OL].https://blog.csdn.net/qq_24519677/article/Details/82811215.

[15] chencas.catboost 原理[DB/OL].https://blog.csdn.net/chencas/java/article/details/104418476.

[16]齐常青.面向不平衡样本分类的过采样集成学习算法研究[D].哈尔滨：哈尔滨工业大学.2019:15-16.

[17]贾文慧.基于 XGBoost 算法的骨科辅助诊断模型研究[D].太原:太原理工大学.2018:12-13.

[18]沈磊.基于 Spark 的微博舆论监控系统的设计与实现[D].成都:电子科技大学.2018:8-10.

附 录

三种算法前 20 个重要特征

	cbt_features	lgb_features	xgb_features
1	基本每股收益_x	基本每股收益_x	息税折旧摊销前利润/负债合计
2	上市年限	最低价	每股净资产
3	投资支出/折旧和摊销	上市年限	上市年限
4	息税折旧摊销前利润/负债合计	最高价	基本每股收益_x
5	每股净资产	每股净资产	总资产净利率
6	最低价	收盘价	最低价
7	稀释每股收益_x	投资支出/折旧和摊销	稀释每股收益_x
8	收盘价	基本每股收益同比增长	收盘价
9	最高价	收盘价上半年变异系数	是否为国企
10	总资产净利率	总资产净利率	投资支出/折旧和摊销
11	基本每股收益同比增长	管理费用/营业总收入	是否为小盘
12	净资产收益率(平均	总资产相对年初增长	x0_水利、环境和公共设施管理业
13	每股资本公积	扣除非经常损益后的归母净利润/归母净利润	每股资本公积
14	收盘价上半年变异系数	最高价下半年变异系数	每股未分配利润
15	最高价下半年变异系数	非流动负债/负债合计	每股营业利润(元/股)
16	每股留存收益	稀释每股收益_x	最高价
17	每股收益(期末摊薄	销售费用/营业总收入	每股收益(期末摊薄
18	总资产相对年初增长	期间费用/营业总收入	每股公积金
19	稀释每股收益同比增长	收盘价下半年变异系数	收盘价下半年变异系数
20	本年是否高转送	每股资本公积	总资产报酬率

（注：蓝色标记为同时被两种或三种算法选上的特征）