

第八届“泰迪杯”数据挖掘挑战赛——

A 题：基于数据挖掘的上市公司高送转预测

近年来，我国证券市场的高速发展催生了一批题材股，根据重大事件的不同分类，可以分为资产重组板块、粤港澳板块、新能源板块等等。在这些题材中间，高送转这一题材无疑是中小投资者强烈追捧的对象。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“未来 6 个月是否存在解禁”、“是否存在定增方案”等因子需要挖掘。

问题 1 针对附件给出的因子数据，根据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。

问题 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，并对附件提供的数据，用所建立的模型来预测第 8 年上市公司实施高送转的情况。

名词解释

送股：上市公司分红的一种形式，把上市公司的盈余公积金或未分配利润转化为股份，以股票的形式派发给投资者，把资金以股票的形式留存在公司。由于是利润分配，因此股民需要缴税。

转股：上市公司分红的一种形式，把上市公司的资本公积金转化成股票，再向投资者派发。转股和送股一样是把资金以股票的形式留存在公司，由于资本公积金是公司投入资本本身所带来的资金增值，并不是公司经营带来的利润，所以转股并不属于公司的利润分配。通过转股实现分红只需借贷公司的会计科目，不需要缴税，且操作简单，成本低廉，为上市公司所青睐。

除权：根据上市公司总市值（总股本 × 每股单价），送股、转股都不影响公司总市值，但公司的总股本增加了，使得每股股票所代表的企业实际价值（每股净资产）有所减少，因此需要在发生该事实后从股票市场价格中剔除这部分因素，该行为称为除权。

填权：在某一股票除权后的一段时间里，如果多数人对该股看好，该股票的交易市价将高于除权的基准价，这种行情称为填权行情。

高送转：指送红股或者转增股票的比例较大。

附件

附件数据为上市公司数据，包含年数据、日数据、基础数据三张表。

附录

请仔细阅读以下说明。

1. 关于赛题数据

- (1) 示例数据：2020年3月1日 9:00:00 随赛题公布。
- (2) 全部数据：2020年4月11日 9:00:00 公布。
- (3) 测试数据：2020年4月25日 9:00:00 公布。

2. 提交作品

(1) 命名方式：论文命名为“A题”，附件命名为“作品附件”，测试结果命名为“作品测试结果”。

(2) 论文及附件内请勿出现队号、学校、学院、队员以及指导老师相关任何信息，否则该作品视为无效作品。

(3) 请参赛队于2020年4月24日 16:00:00之前在竞赛官网“提交作品”处提交论文（PDF版，大小不超过50M）及附件（论文正文（Word版）、程序源代码、代码实现的流程图以及输入、输出说明的压缩包，大小不超过200M）。

3. 公布测试数据，提交测试结果

2020年4月25日 9:00:00准时公布测试数据，请在“赛题与数据”页面对应的题目右下方下载测试数据，并于2020年4月26日 9:00:00前在“提交测试结果”页面提交测试结果。