中泰证券融资业务压力测试
客户预判模型

模型名字：压力测试客户预判模型

模型开发人员：唐月月、张研、姜萌、

候盈男、陆军、沈伟

指导老师：山东大学林路教授

模型版本：

文档版本：

模型编号：

中泰证券股份有限公司风控合规总部

二〇一五年十二月

目录

一、 模型概述

（一）、模型目的

2015 年 6 月中旬以来，股市震荡加剧，给两融业务带来巨大压力，尤其在 6 月中旬的股灾中，两融业务客户的资金损失严重，部分客户甚至被强行平仓。这种压力不仅给客户带来巨大财产损失，也可能给证券公司带来流动性风险，并导致市场行情持续恶化。在市场行情极端下跌中，很容易产生流动性危机，所以加强两融风险管控，特别是对有可能出现被强制平仓甚至违约的客户进行预判变得尤为重要。

在正常市场行情下，通常的客户信用评级技术发挥着重要作用，可以很好的区分好坏客户。但是当市场行情出现极端下跌时，传统的评级方法已经不再适用，可能正常评级很好的客户也会在极端压力下变成定时炸弹，给证券公司带来巨大损失。而目前证券公司尚没开发出压力下的危险客户预判模型，因此我们尝试建立这样一种客户评价模型，主要目的就是从当前客户的持仓特点预判其在压力情形下成为危险客户（持仓市值迅速缩水）的概率。

（二）、应用的范围

本压力测试客户预判模型只适用于证券公司融资业务。

（三）、模型使用部门

中泰证券股份有限公司风控合规总部拥有该模型的使用权。

（四）、模型开发人员和运行版本

本模型主要由山东大学中泰证券金融研究院林路教授指导其硕士生——唐月月、张研、姜萌、候盈男、陆军、沈伟共同开发完成。在建模过程中，中泰证券风控合规总部程鹏经理以及林路教授的博士生刘永欣、柴海涛、董平给予了开

发人员很多有益建议。

（五）、模型的开发工具和运行环境

该模型主要基于 Matlab R2013a 软件，利用 Excel 软件和 wind 资讯的 Matlab 接口储存并读取数据。

该模型可在任何安装了 Matlab、Excel、wind 客户端的标准 PC 上运行，我们开发模型时用到的 Lenove windows7 PC 配置如下：Intel(R) Core(TM) i5 CPU 650 @ 3.20GHz; RAM 4.00GB(2.99 GB 可用); 系统类型：32 位操作系统。

（六）、模型与其他模型的相关性

暂未发现该模型与其他模型的相关性。

（七）、模型方法/流程的总体介绍和模型的预期结果

该模型包括两个子模型：夏普比预判模型以及收益-波动率判别分析模型。夏普比预判模型是根据客户当前持仓证券的加权夏普比率预判坏客户（维持担保比下降最快的 10%），我们发现坏客户的持仓证券加权夏普比率显著偏低，我们就能通过寻找夏普比率最低的客户以一定置信度把坏客户筛选出来；收益-波动率判别分析模型利用客户持仓证券加权收益率和波动率作为特征对客户待判集进行判别分类，找出坏客户。

夏普比预判模型流程：首先整理初始数据，下载股价数据，整理成客户资料库，获取每个交易日每个客户的收益率、波动率、维持担保比变化百分比等信息；计算每个客户当前持仓证券组合加权夏普比；取夏普比率的一个低区段（区段长度根据需要自己调整），将所有夏普比率落在该区段的客户视为需要重点关注的客户（有一定的误判率），这些客户在下一交易日维持担保比可能是跌幅最大的那一类客户。

收益-波动率判别分析模型流程如下：指定某一部分客户（客户持仓历史数据集）为训练集，训练集中的客户信息有客户的收益率、波动率，客户的类别标

签（“坏客户”或“正常客户”）。需要预测的客户作为待判集，待判集中的客户信息有客户的收益率、波动率，客户的类别待判定。

（八）、不适合该模型应用的情况

首先，本压力测试模型暂时不适用于除融资业务外的其他业务；其次，本模型不能量化坏客户将会带来的损失。

二、 模型的数据介绍

（一）、模型的开发数据来源

模型开发过程中用到的数据来源主要有两个：一是 wind 咨讯，二是中泰证券风控合规部。其中，wind 咨讯的 Matlab 接口可以直接将模型需要的股票价格等数据导入 Matlab 程序，中泰证券风控合规部提供的客户持仓数据是 Excel 格式。

（二）、数据检验和转换

建模过程需要对股票数据根据客户持仓转化加权收益率和波动率，后文模型理论部分会介绍计算过程。

（三）、模型的运行数据

与开发数据一致，模型运行数据用到两个数据源：证券公司当前融资业务客户持仓情况和 wind 提供的最近 8 个月内的股价数据。

运行模型时，不需要再预处理数据，开发的模型可以自动加载数据、预处理数据、计算出每个客户的夏普比，给出判别结果。

三、夏普比预判模型方法

（一）、模型总体介绍

3.1.1 总体思路

- 刻画客户的持仓特点

  所用指标：前一交易日收益率、波动率、夏普比率

- 刻画客户的维持担保比变化

  单个交易日、连续几个交易日维持担保比变化百分比

- 刻画维持担保比变化情况与持仓特点的关系

  举例：根据客户在 6 月 16 日的维持担保比变化百分比挑出最坏的一类客户，回到 6 月 15 日，观察最坏客户与正常客户、最坏客户与最好客户在 6 月 15 日的持仓特点有何差异。找出一个指标使得坏客户在这个指标上的表现与其他客户很不相同。

  比如说，我们得出的规律是，坏客户的夏普比率显著偏低，我们就能通过寻找夏普比率最低的客户以一定置信度把坏客户筛选出来。

- 流程图：

Step1：建立客户数据库

整理初始数据，下载股价数据，整理成客户资料库，获取每个交易日每个客户的收益率、波动率、维持担保比变化百分比等信息。

Step2：挑选指标，区分客户

找指标区分出坏客户与好客户

- 刻画最坏客户与最好客户的持仓差异

  单个交易日、连续 2 个交易日

- 刻画最坏客户与正常客户的持仓差异

  单个交易日、连续 2 个交易日

- 4 种情况

  （1）单个交易日内最坏客户与最好客户的指标差异

  （2）单个交易日内最坏客户与正常客户的指标差异

  （3）连续 2 个交易日内最坏客户与最好客户的指标差异

  （4）连续 2 个交易日内最坏客户与正常客户的指标差异

单个交易日：
- 找指标区分出坏客户与好客户
- 找指标区分出坏客户与正常客户

连续 2 个交易日：
- 找指标区分出坏客户与好客户
- 找指标区分出坏客户与正常客户

Step3：画图并统计区分结果

（二）、指标介绍

3.2.1 刻画客户的持仓特点

选用指标：收益率、波动率

3.2.1.1 单支股票的收益率、波动率

3.2.1.1.1 单支股票的收益率

第 j 支股票价格在第 i 个交易日的收益率为 $r_j^{(i)}=\frac{S_j^{(i)}-S_j^{(i-1)}}{S_j^{(i-1)}}$， $i=2,3,\cdots,7$

3.2.1.1.2 单支股票的波动率

第 j 支股票收益率在第 i 个交易日的波动率记为 $\sigma_j^{(i)}$，波动率用于度量股票所提供收益的不确定性，可以用收益率的标准差来作为其估计值。

在某个固定的时间区间内（如每天、每周或每个月），我们观测了 $n+1$ 次股票价格，记第 i 个时间区间结束时股价为 $S_i(i=0,1,2,\cdots,n)$，每个时间区间的长度为 τ（以年为单位），第 i 个时间区间结束时收益率为 $r_i(i=1,2,\cdots,n)$，则 $\{r_i\}_{i=1}^n$ 的标准差通常估计为 $s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(r_i-\bar{r})^2}$，其中 $\bar{r}$ 为 $\{r_i\}_{i=1}^n$ 的均值。则有这段时间内波动率的估计 $\hat{\sigma}=\frac{s}{\sqrt{\tau}}$。实际应用中，$r_i(i=1,2,\cdots,n)$ 既可以用简单收益率也可以用对数收益率。波动率的平方称为方差率。

波动率估计方法：指数加权移动平均模型（EWMA）

假设 σ_n 为第 $n-1$ 天估计的第 n 天的波动率，则用最近 m 个交易日的收益率数据所计算出的方差率 $\sigma_n^2=\frac{1}{m-1}\sum_{i=1}^{m}(r_{n-i}-\bar{r})^2$，其中 $\bar{r}$ 为 $\{r_i\}_{i=1}^n$ 的均值。在实际应用中，将 $\bar{r}$ 取为 0，将 $m-1$ 用 m 替代，这两个变化对计算结果影响不大，我们将公式简化为

$$\sigma_n^2 = \frac{1}{m}\sum_{i=1}^{m} r_{n-i}^2 \qquad (1)$$

公式（1）中所有项具有相同的权重，我们也可以对每一项赋予不同的权重，得到公式（2）：

$$\sigma_n^2 = \frac{1}{m}\sum_{i=1}^{m} \alpha_i r_{n-i}^2 \qquad (2)$$

其中α_i是为i天前的收益率所赋的权重，$\sum_{i=1}^{m}\alpha_i = 1$。

EWMA 是公式（2）的一种特殊形式，我们希望估计当前的波动率水平，因此考虑将较大的权重赋给最近的数据，可以假设α_i随着时间递减，具体地讲，假设$\alpha_{i+1} = \lambda\alpha_i$，其中$0 < \lambda < 1$，在这样的假设下，波动率公式可以简化为：

$$\sigma_n^2 = \lambda\sigma_{n-1}^2 + (1-\lambda)r_{n-1}^2 \qquad (3)$$

第n天的波动率估计值σ_n（在第$n-1$天估算）由第$n-1$天波动率估计值$\sigma_{n\text{-}1}$（在第$n-2$天估算）和最近一天的收益率r_{n-1}决定。

3.2.1.2 客户的收益率、波动率、夏普比率

第i个交易日，第k个客户对第j支股票的资金投入占比定义为$w_k^{(i,j)} = \frac{S_j^{(i)} * n_k^{(i,j)}}{\sum_j S_j^{(i)} * n_k^{(i,j)}}$，是第$j$支股票的市值在该客户所有股票持仓总市值中的比例。

3.2.1.2.1 客户收益率的计算

第i个交易日，第k个客户的总收益率$R_k^{(i)}$是该客户当日所持有股票的收益率的加权值：$R_k^{(i)} = \sum_j w_k^{(i,j)} * r_j^{(i)}$

3.2.1.2.2 客户波动率（组合波动率）的计算

客户持有多种股票，我们将客户的持股看作一个交易组合，设组合的价值为P，在交易组合中有n个不同资产，投资组合中资产i的数量为α_i，资产i的每天回报定义为Δx_i，投资α_i数量于资产i所产生的每天回报为$\alpha_i \Delta x_i$，并且$\Delta P = \sum_{i=1}^{n} \alpha_i \Delta x_i$，$\Delta P$为整个交易组合实际的价值变化。

为了计算$\Delta P / P$（组合的收益率）的方差σ_p^2，假定σ_i为第i项资产每天的波动率，ρ_{ij}为资产i及资产j的相关系数，ω_i为组合中第i个资产的权重，则得组合的方差为：

$$\sigma_p^2 = \sum_{i=1}^{n}\sum_{i=1}^{n} \rho_{ij}\omega_i\omega_j\sigma_i\sigma_j = \sum_{i=1}^{n}\sum_{i=1}^{n} \mathrm{cov}_{ij}\,\omega_i\omega_j$$

第i项资产每天的波动率σ_i，我们采用指数加权移动平均法计算协方差矩阵。指数加权移动平均法通过将近期的数据赋予较高的权重来反映波动性的动态特征，是对简单移动平均法之缺陷的一种有效弥补方法。按照指数移动平均法定义的收益率方差为：

$$\sigma_{t+1|t}^2 = \sum_{k=0}^{t} \lambda^k (1-\lambda)(\mathrm{r}_{t-k} - \bar{r})^2$$

其中λ决定了不同时期的历史数据在计算波动性时的相对权重，我们称其为衰减因子。指数加权移动平均法所确定的权重是随着期限临近而加大的，因此相对于简单平均而言能够较迅速地对市场冲击的变化做出反映，并且在冲击过后随着权重的减小，波动性会呈指数形式衰减。这样就能较为准确第刻画出时间序列波动的聚类性。

3.2.1.2.3 夏普比率

$sharp = \frac{r - r_f}{\sigma}$，其中$r_f$是无风险收益率，夏普比率度量每付出一单位风险，所获

得的超额收益率。在实际处理中，为了计算方便，我们取 $sharp = \frac{r}{\sigma}$。

当夏普比率为正值时，夏普比率越大，表明单位风险获得的超额收益率越多，所以此时夏普比率越大越好；当夏普比率为负值时，夏普比率越小，表明单位风险获得的亏损越多，所以此时夏普比率越小越差。在我们所关注的时间段 2015 年 6 月 12 日至 2015 年 6 月 19 日内，市场处于极端下跌行情，大部分股票的夏普比率均是负值。由前面讨论可知，无论夏普比率的取值正负，夏普比率数值越大单位风险的收益率越大，对股票持有者越有利。

3.2.2 刻画客户的维持担保比变化

3.2.2.1 维持担保比单个交易日变化幅度

我们定义第 k 个客户在第 i 个交易日的维持担保比变化百分比为 $\frac{ratio_k^{(i)} - ratio_k^{(i-1)}}{ratio_k^{(i-1)}}$，我们认为客户在第 i 个交易日的维持担保比变化幅度是由客户在第 $i-1$ 个交易日的持仓特点所决定的，这样便于我们做预测。

3.2.2.2 维持担保比连续几个交易日变化情况

我们最主要的目标是找出维持担保比下降幅度最大的那一批客户（正是这一批客户最有可能给证券公司带来最大的损失），捕捉这一批客户的持仓特点（通过客户资产的收益率和波动率等指标）。得出在压力情形下，怎样的持仓特点最有可能使客户成为一个坏客户。在下一次类似的极端行情发生时，我们给予这一类客户重点关注，及时采取措施避免这类客户造成损失。

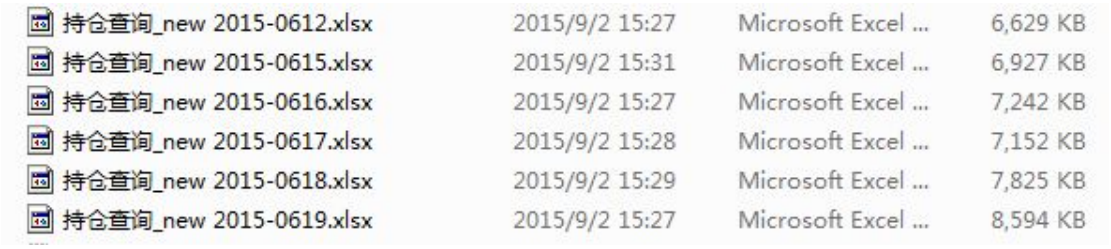
如果只观察客户在单个交易日的维持担保比变化，得到的结果可能不太稳健，我们希望通过观测客户在连续几个交易日内的维持担保比变化情况，筛选出表现最坏的那一类客户。如果某个客户的维持担保比在连续 2 个或是多个交易日内持续下降，并且保持较大的降幅，我们就认为该客户是表现最坏的一类客户。

（三）、模型详细步骤

3.3.1 建立客户数据库

3.3.1.1 原始数据：

6 张表：



每张表内容：

持仓查询_new

查询条件-> 业务日期：2015年6月12日；客户号：；维持担保率(%)<=：3；

客户号	客户名称	资金账号	股票代码	股票名称	数量	总融资负债	维持担保率(%)
611000026186	范立波	1010900015	002025	航天电器	197349	29303382.64	2.534842393
611000026186	范立波	1010900015	002023	海特高新	600	29303382.64	2.534842393
611000026186	范立波	1010900015	000063	中兴通讯	578244	29303382.64	2.534842393
611000026186	范立波	1010900015	600998	九州通	17800	29303382.64	2.534842393
611000026186	范立波	1010900015	600050	中国联通	40000	29303382.64	2.534842393
611000018463	关晓晨	1010900017	002341	新纶科技	10000	1379401.76	2.752031181
611000018463	关晓晨	1010900017	000507	珠海港	25300	1379401.76	2.752031181
611000018463	关晓晨	1010900017	601318	中国平安	8000	1379401.76	2.752031181
611000018463	关晓晨	1010900017	600755	厦门国贸	24600	1379401.76	2.752031181
611000018463	关晓晨	1010900017	600738	兰州民百	29800	1379401.76	2.752031181
611000018463	关晓晨	1010900017	600705	中航资本	6800	1379401.76	2.752031181
611000018463	关晓晨	1010900017	600698	湖南天雁	23700	1379401.76	2.752031181
611000018463	关晓晨	1010900017	600485	信威集团	8300	1379401.76	2.752031181
611000018463	关晓晨	1010900017	600067	冠城大通	30000	1379401.76	2.752031181
611000027712	连明	1010900019	300310	宜通世纪	237000	82123297.41	2.369675598

从初始数据中提取信息获得客户资料库：

6 个交易日总共涉及 36319 个客户，提取每个交易日每个客户的维持担保比，提取每笔记录的股票代码、数量，统计每个客户持股种类、数量，计算出每个客户在其持有的每支股票上的资金投入占比、客户每日维持担保比变化百分比，形成初始的客户资料库。概览：

日期	交易日序号（从 2015 年初开始）	涉及股票个数	记录条数
2015 年 6 月 15 日	109	2976	139777
2015 年 6 月 16 日	110	2994	146053
2015 年 6 月 17 日	111	3001	144745
2015 年 6 月 18 日	112	2990	160230
2015 年 6 月 19 日	113	3023	176403

3.3.1.2 下载股票相关指标

对指定的交易日，对前一步整理出的股票名单添加后缀，从 wind 上下载这支股票从 2015 年 1 月 1 日至目标交易日的收盘价数据。为了估算股票波动率，选择从 2015 年年初开始下载股价数据。为了使用股票的多种指标，还下载了股票的换手率、自由流通市值、净流入资金、资金流向占比、市盈率、融资余额占自由流通市值之比等等指标作为备选指标。

3.3.1.3 计算客户各项指标，形成客户信息库

对第二步下载的各项股票指标，按照第一步中每位客户的持股信息计算每个客户的相应指标。除了客户的波动率需要考虑各支股票之间的相关性外，其他指标一律使用按资金投入权重加权的方法算得每位客户的相应指标。

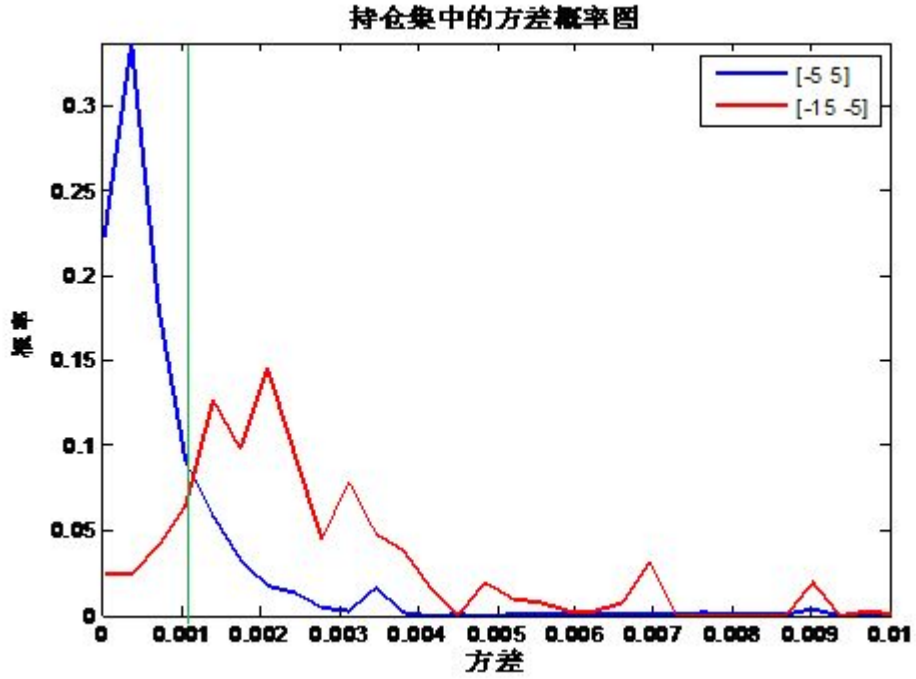
3.3.2 挑选指标区分客户

我们希望通过分析客户在前一个交易日的持仓特点，预测客户在下一个交易日将更可能成为一个坏客户还是一个正常客户。

为了达到这个目标，我们需要从结果倒推原因（发病的病征）去找到规律。比如

说，我们用 6 月 16 日所有客户的维持担保比变化百分比，找出表现最差的 10%的那部分客户，回到 6 月 15 日，考察这些客户与其他正常客户在 6 月 15 日的持仓特征上有何不同。所谓的持仓特征就是客户的收益率、波动率、持仓的流通市值等指标。

如果我们可以找到一个指标，在前一个交易日对潜在的坏客户具有较好的区分度，并且这个指标持续保持较稳定的区分度，我们就可以用这个指标去做预测。预期的效果如下图所示：



红线是下一个交易日将要变成坏客户的那些客户在当前交易日方差分布概率，蓝线代表下一个交易日为正常客户的那些客户在当前交易日方差分布概率，可以发现图像的右侧大部分被红线所占据，蓝线在右侧覆盖的面积非常小。

如果我们设置一条像图中绿线那样的垂直线，将绿线右侧全部预测为坏客户，绿线左侧全部预测为正常客户。我们能够在一定的置信度下区分出坏客户与正常客户。当然，绿线左侧红色曲线的面积是我们遗漏掉的坏客户的概率，绿线右侧的蓝色曲线的面积是被误判为坏客户的正常客户所占的比率。

实际的处理步骤（以 6 月）；

在实际操作过程中，我们有以下几种尝试：

（1） 试算的指标主要有收益率、波动率、换手率、自由流通市值、净流入资金、资金流向占比、市盈率、融资余额占自由流通市值之比，最后发现对坏客户区分度最大的是收益率、波动率两个指标，以及结合两者的夏普比率这一指标，这意味着传统的均值-方差指标正好是最有效的；

（2） 为了刻画坏客户与正常客户的持仓差异，首先试算坏客户与好客户的持仓差异，以便快速找到坏客户区别于一般客户的鲜明特征；

（3） 为了使得结果更加稳定，我们既找出了单个交易日表现极坏的客户，也找出了连续 2 个交易日持续表现极坏的客户。

实际的处理步骤如下：

（1） 单个交易日区分好坏客户（以 6 月 16 日为例）：

找出 6 月 16 日维持担保比变化百分比表现最差的 10%的客户，这部分客户标记为坏客户，找出 6 月 16 日维持担保比变化百分比表现最好的 10%的客户，这部分客户标记为好客户。分析 6 月 15 日坏客户与好客户的持仓差异。

（2） 单个交易日区分坏客户与正常客户（以 6 月 16 日为例）：

找出 6 月 16 日维持担保比变化百分比表现最差的 10%的客户，这部分客户标记为坏客户，其他客户标记为正常客户。分析 6 月 15 日坏客户与正常客户的持仓差异。

（3） 连续 2 个交易日区分好坏客户（以 6 月 16 日、6 月 17 日为例）：

分别找出 6 月 16 日、17 日维持担保比变化百分比表现最差的 10%的客户，这 2 部分客户取交集标记为坏客户。分别找出 6 月 16 日、17 日维持担保比变化百分比表现最好的 10%的客户，这 2 部分客户取交集标记为好客户。分析 6 月 16 日坏客户与好客户的持仓差异。

（4） 单个交易日区分坏客户与正常客户（以 6 月 16 日、6 月 17 日为例）：

分别找出 6 月 16 日、17 日维持担保比变化百分比表现最差的 10%的客户，这 2 部分客户取交集标记为坏客户。其他客户标记为正常客户。分析 6

月 16 日坏客户与正常客户的持仓差异。

3.3.3 画图并统计区分度

输出直观图线，计算两类误判率。

计算误判率的步骤：

Step1：主观指定目标百分比，为找出的坏客户占实际坏客户的百分比，对所取指标（假设该指标越小客户越容易成为坏客户）求的下侧分位点；

Step2：计算落在分位点左侧的正常客户比例，可以取左侧的正常客户占全部正常客户的比例（取左侧的正常客户占全部客户的比例也可以，最后可以换算）记作；

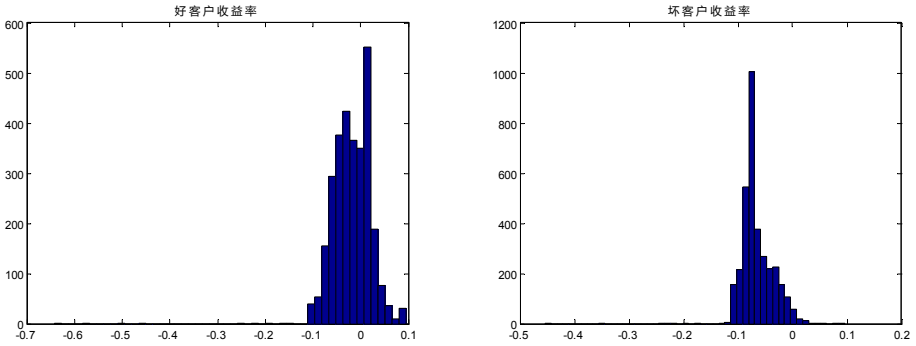
Step3：找出的坏客户占实际坏客户的比例为，遗漏 1-的坏客户，将正常客户中的比例误判为坏客户。

四、夏普比预判模型运行与测试

（一）、比较单个交易日最坏客户与最好客户前一交易日持仓特点

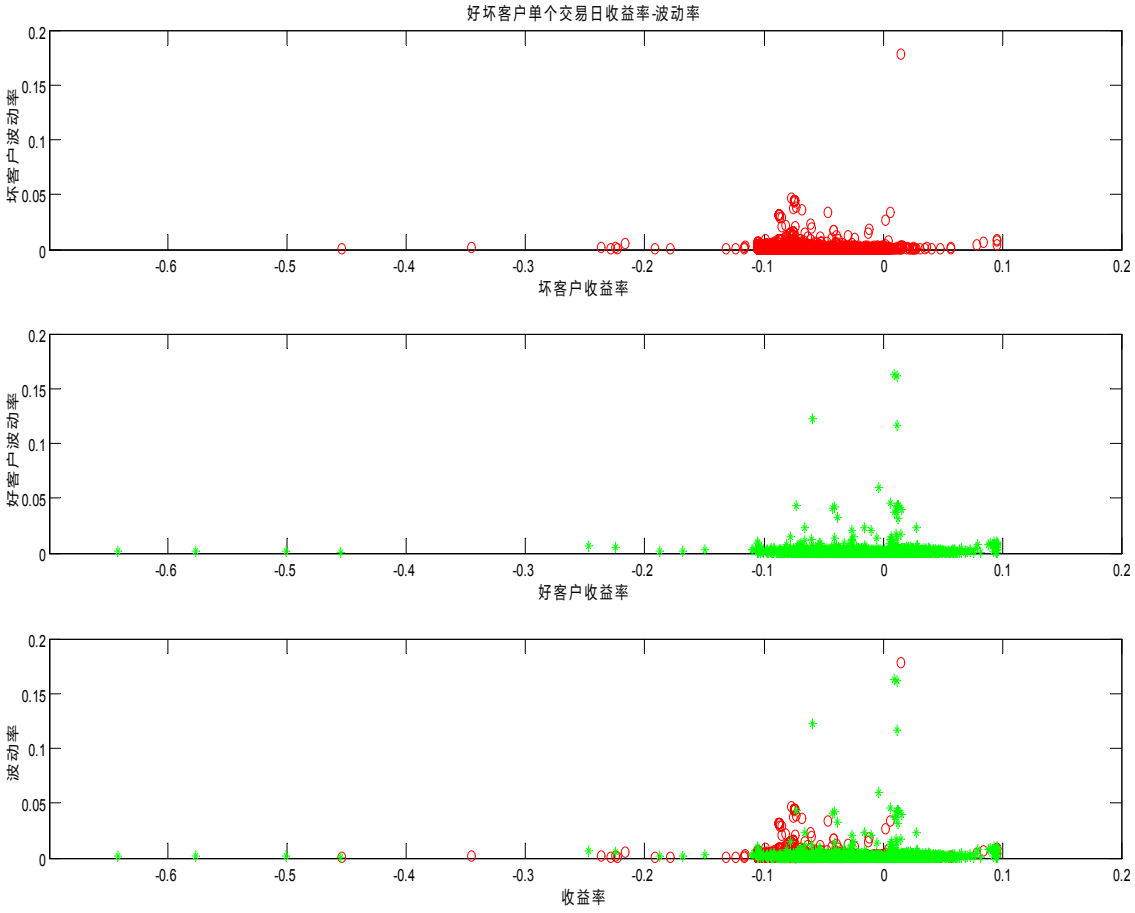
指定某个交易日（以 6 月 16 日为例），我们找出这个交易日维持担保比下降百分比最大的客户(最坏客户)与这个交易日维持担保比上升百分比最大的客户(最好客户），

	维持担保比变化百分比	客户个数
最好的一类客户	[0.37%,73%]：升幅大于最大升幅的 0.5%	2960
最坏的一类客户	[-45%,-6.8%]：降幅大于最大降幅的 15%	3393

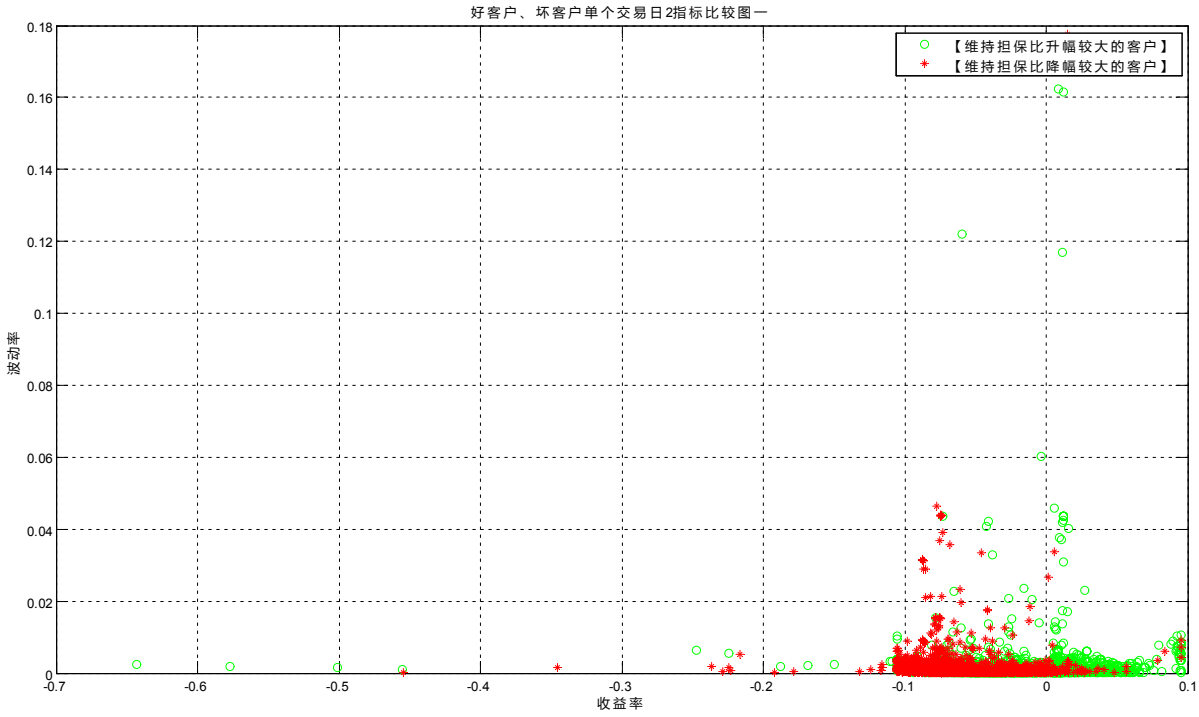


图一.单个交易日最好与最坏客户收益率直方图

由图一可知好客户收益率较均匀地分布在 0 周围[-10%,10%]内，坏客户相对于好客户左偏，大部分在[-10%,0]内。

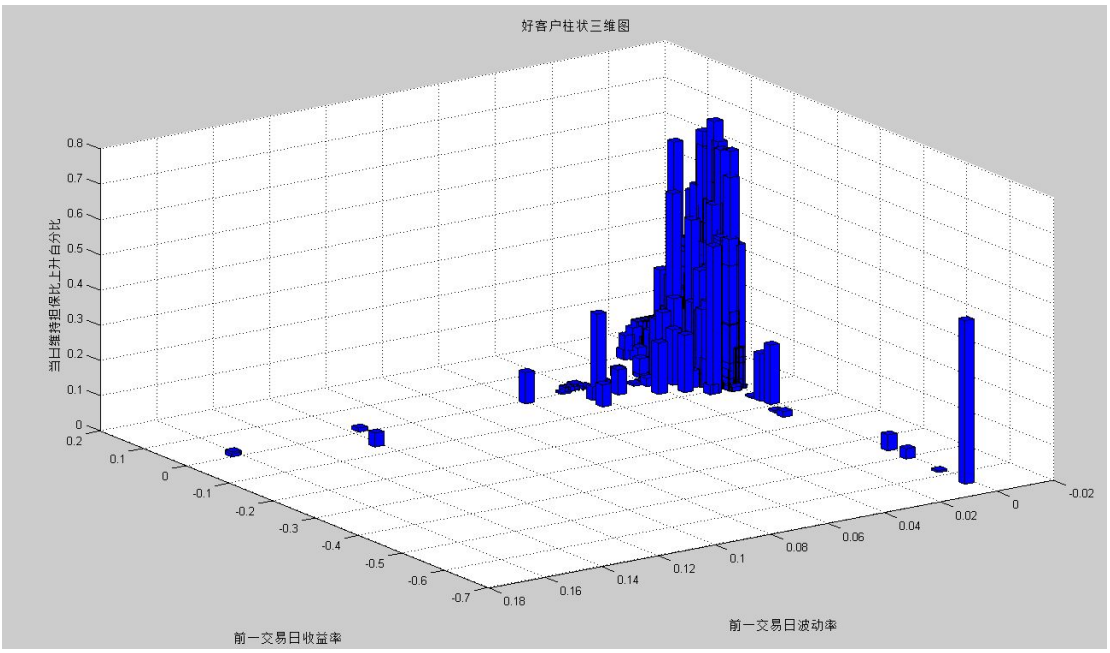


图二、单个交易日内好坏客户两个指标比较图



图三、单个交易日内好坏客户两个指标比较图

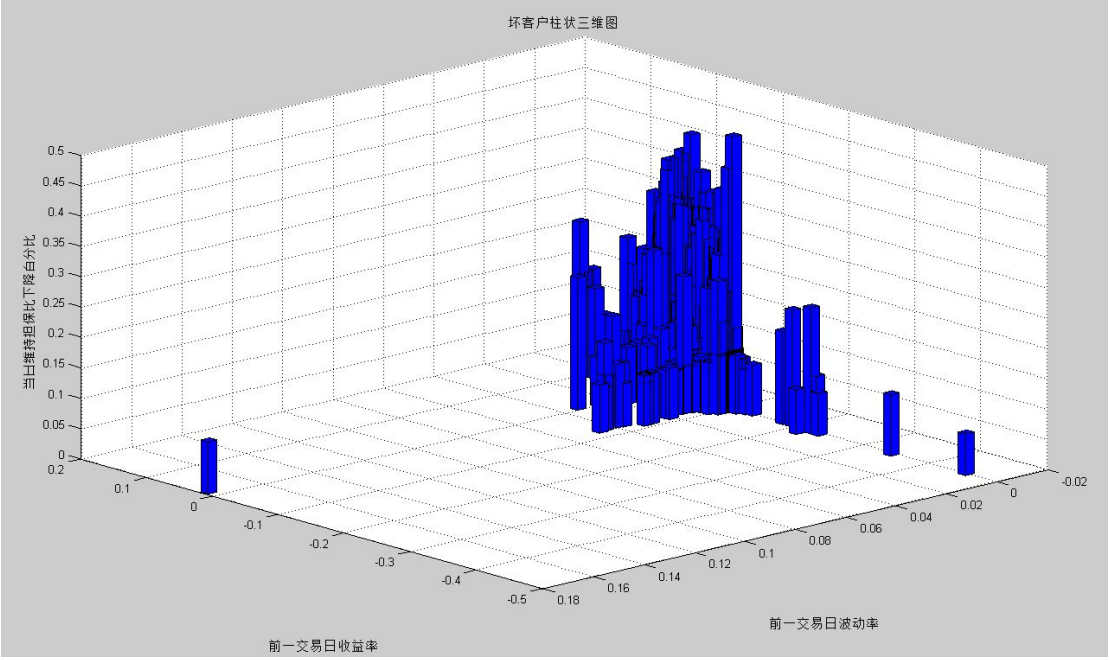
由图二和图三可知，好坏客户的波动率没有太大差别，只是收益率区别较大。



图四、好客户收益率-波动率-维持担保比变化百分比三维图

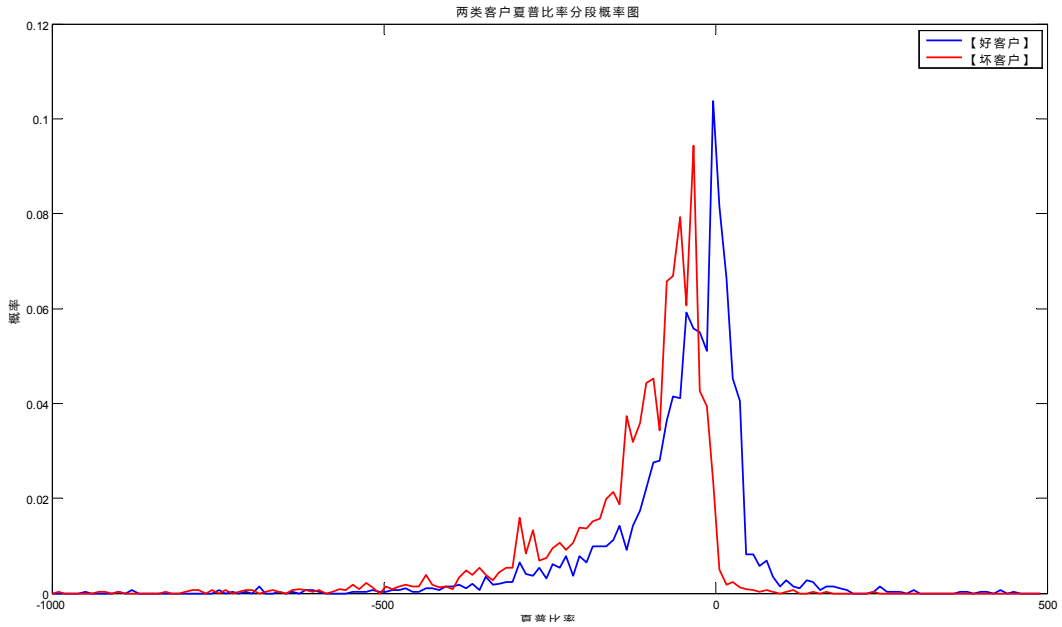
好客户维持担保比是上升的，图四柱形高度表示维持担保比上升百分比，可见维

持担保比表现较好的客户大多数聚集在低波动率、中等收益率区域。



图五、坏客户收益率-波动率-维持担保比变化百分比三维图

坏客户维持担保比是下降的，图四柱形高度表示维持担保比下降百分比，可见维持担保比表现最坏的客户大多数聚集在低波动率、低收益率区域，这可能是因为在股市大跌的极端情形下，收益率较低、收益率变动较小的股票大多数停盘所造成的。



图六、单个交易日好坏客户夏普比率分段概率图

在收益率大于 0 的部分，夏普比率也为正值，夏普比率越高，表示单位风险所获收益为正，夏普比率数值越大越理想；

在收益率小于 0 的部分，夏普比率为负数，夏普比率越低，表示单位风险带来更多的亏损，所以夏普比率数值越大越好。

在图六中，好客户的夏普比率比坏客户的夏普比率偏右，这与上面的分析是相符的。所以我们也可以尝试使用夏普比率这一指标来区分好坏客户。

红线代表坏客户，蓝线代表好客户。从对比图可以看出，坏客户的图比好客户的图偏左，说明坏客户投资组合的夏普比率小。

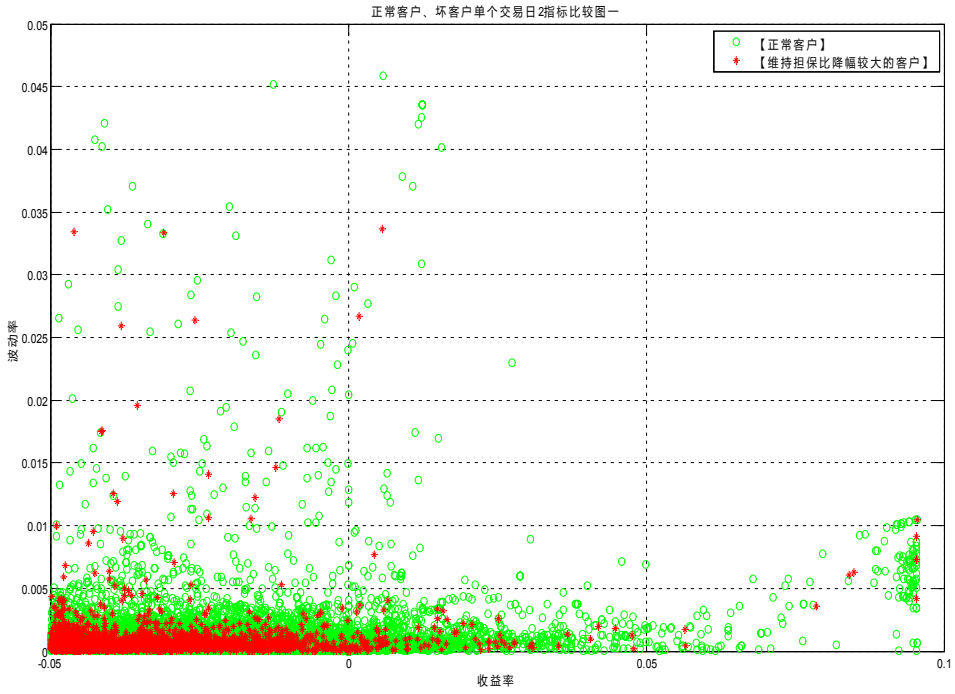
取 85%的下侧百分数，红线分位点为-28.4，如果我们取夏普比率的[-inf,-28.4]区段，这将挑选出 85%的坏客户。同时好客户有 47%的比重落入夏普比率的[-inf,-28.4]区段，这意味着这部分好客户被误判为坏客户。

我们忽略了 15%的坏客户。

（二）、比较单个交易日最坏客户与正常客户前一交易日持仓特点

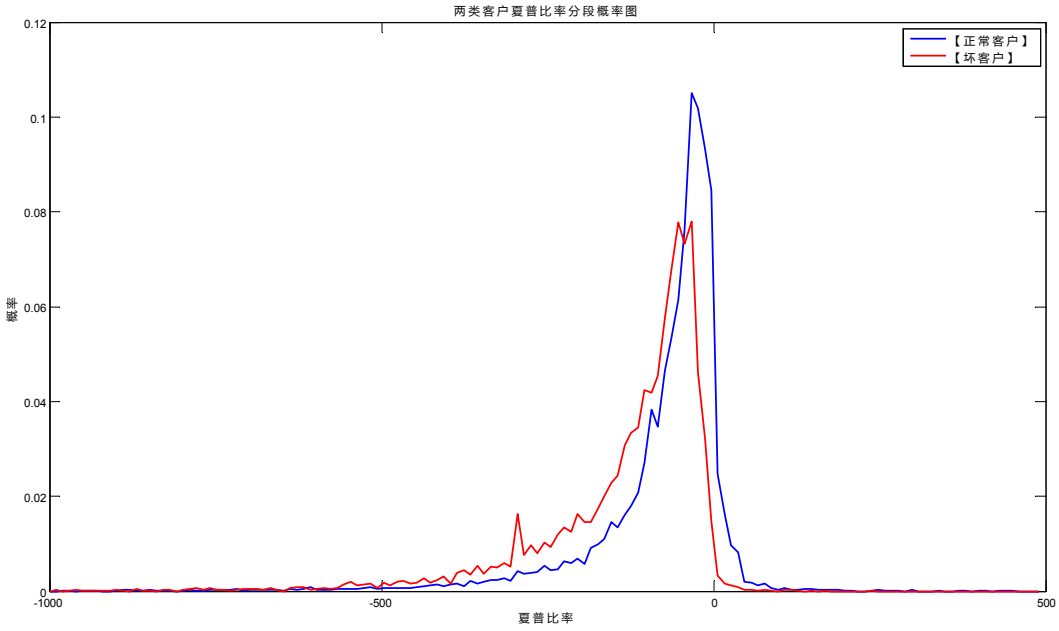
指定某个交易日（以 6 月 16 日为例），我们找出这个交易日维持担保比下降百分比最大的客户（最坏客户），其他客户称为正常客户。

	维持担保比变化百分比	客户个数
正常客户	其他客户:[-4.5%,73%]	15214
最坏客户	降幅大于最大降幅的 10%:[-45%,-4.5%]	8586



图七、正常客户、坏客户单个交易日收益率-波动率

从图七来看，正常客户和坏客户的波动率差别不大，坏客户的收益率集中在比较低的水平。正常客户两指标的范围相对坏客户更广，这可能是正常客户数量较多造成的。



图八、正常客户与坏客户夏普比率分段概率图

正常客户的夏普比率相对坏客户右偏。

坏客户的下侧 80%分位点为-36.44，我们如果取[-inf,-36.44]作为坏客户区段，可以挑选出 80%的坏客户，同时这一区段包含了约 52%的正常客户，这部分正常客户被误判为坏客户。在所有[-inf,-36.44]客户中，被误判的正常客户达到了 53%。

坏客户的下侧 70%分位点为-50.2，我们如果取[-inf,-50.2]作为坏客户区段，可以挑选出 70%的坏客户，同时这一区段包含了约 42%的正常客户，这部分正常客户被误判为坏客户。在所有[-inf,-50.2]客户中，被误判的正常客户达到了 51%。

（三）比较连续二个交易日最坏客户与最好客户

指定某个交易日（以 6 月 15、16 日为例），我们找出这 2 个交易日维持担保比连续下降百分比最大的客户（最坏客户），找出这 2 个交易日维持担保比连续上升百分比最大的客户（最好客户）。

	维持担保比变化情况	客户个数
最好客户	连续二天升幅大于最大升幅的 3%	245
最坏客户	连续二天降幅大于最大降幅的 15%	250

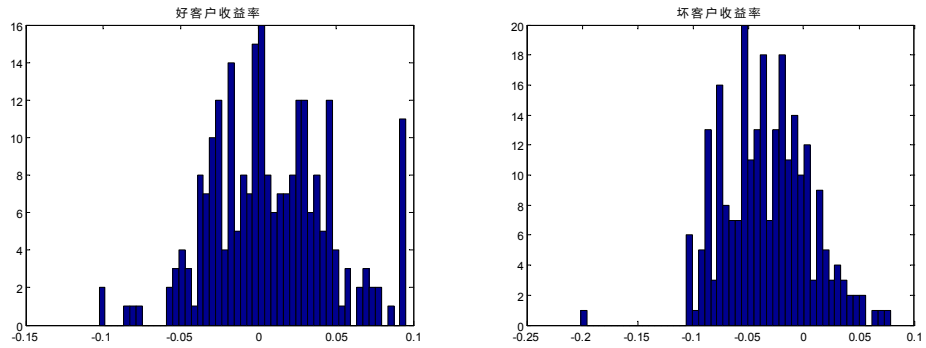
其中：

第一个交易日分类情况：

	当日维持担保比变化情况	客户个数
第一类客户	升幅[2.3%,75%]	2319
第二类客户	升幅[-43.6%,-6.5%]	2128

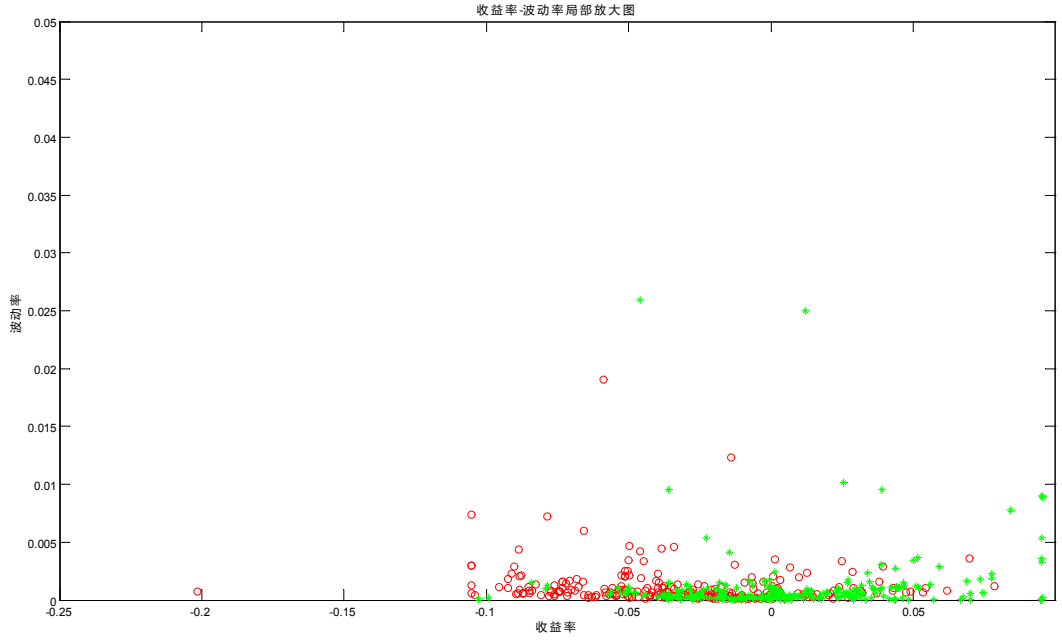
第二个交易日分类情况：

	当日维持担保比变化情况	客户个数
第一类客户	升幅[2.2%,73%]	1908
第二类客户	升幅[-45%,-6.8%]	3393



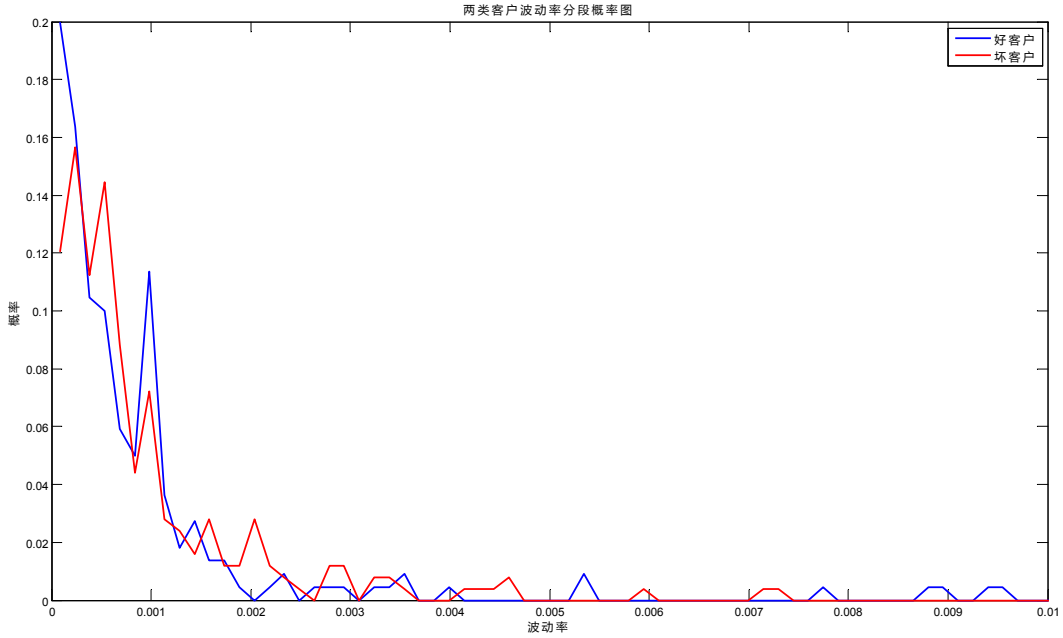
图九、连续 2 个交易日好、坏客户收益率分布直方图

好客户的收益率集中在[-0.05,0.05],坏客户的收益率集中在[-0.1,0.05],坏客户相对好客户收益率左偏。

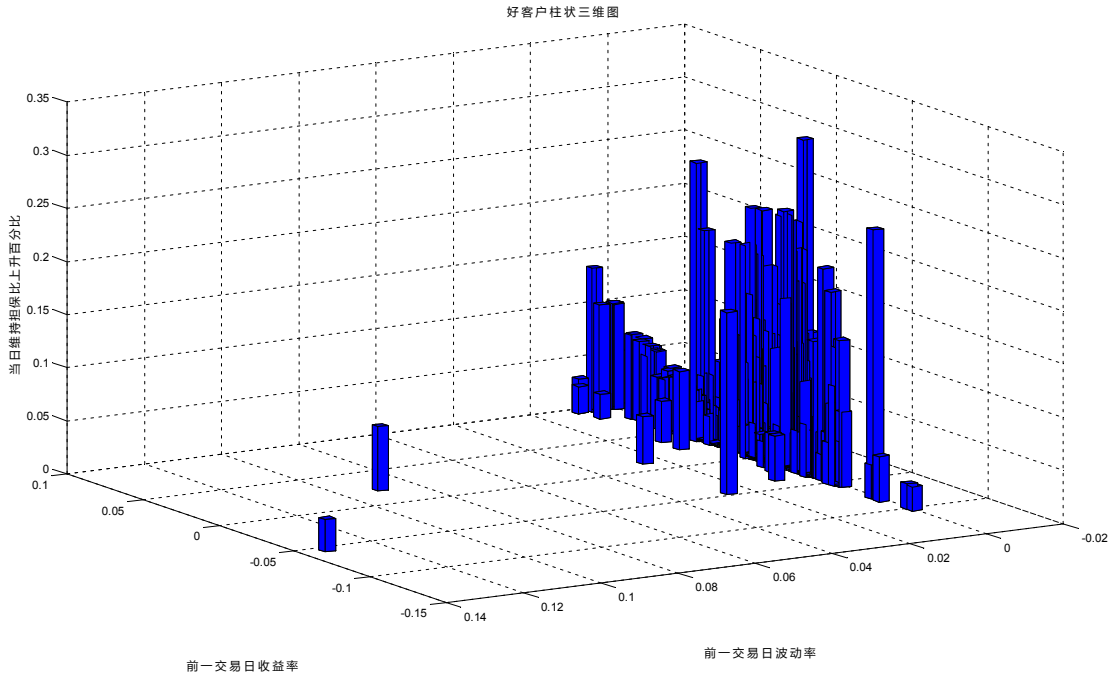


图十、好坏客户收益率-波动率

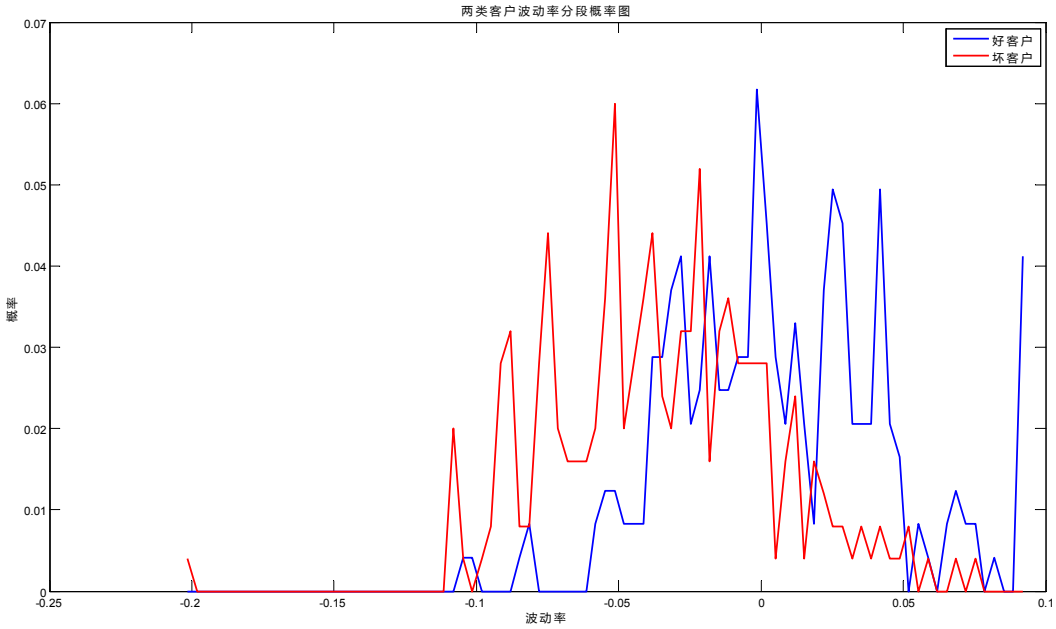
好客户比坏客户的收益率普遍要大。



图十一、好坏客户波动率分段概率图

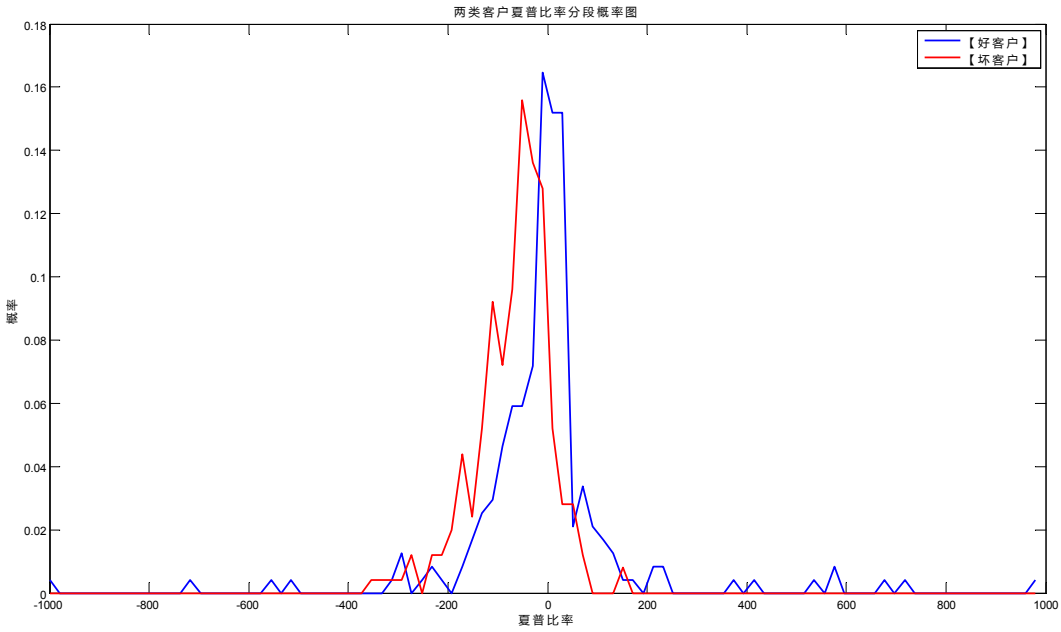


图十二、好客户柱状三维图



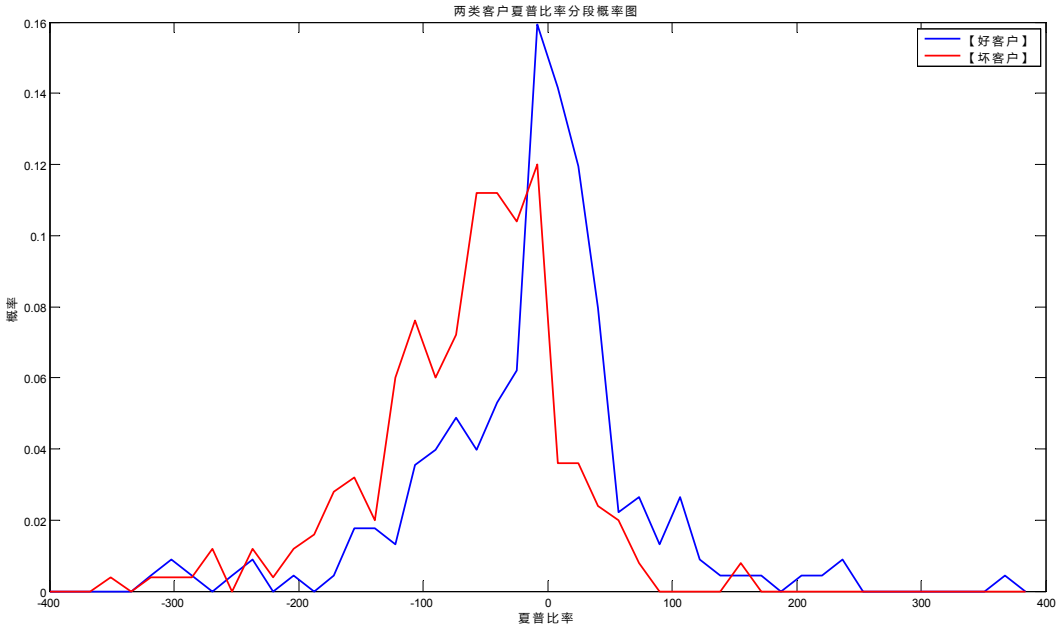
图十三、好坏客户收益率分段概率图

坏客户下侧 90%的收益率区段为[-inf,1.56%],此区段含有的好客户占好客户总数的 59.3%。



图十四、好坏客户夏普比率分段概率图

删去两侧异常值，只考虑[-400,400]之间的数，做出概率图如下：



图十五、好坏客户夏普比率分段概率图

坏客户下侧 85%的夏普比率区段为[-inf,5.11],此区段含有的好客户占好客户总数的 50.9%。

坏客户下侧 70%的夏普比率区段为[-inf,-18.27],此区段含有的好客户约占好客户总数的 34%。

（四）、比较连续二个交易日最坏客户与正常客户

指定某个交易日（以 6 月 15、16 日为例），我们找出这 2 个交易日维持担保比连续下降百分比最大的客户（最坏客户），其他客户称为正常客户。

	维持担保比变化情况	客户个数
最坏客户	连续二天降幅大于最大降幅的 10%	1615
正常客户	其他	11230

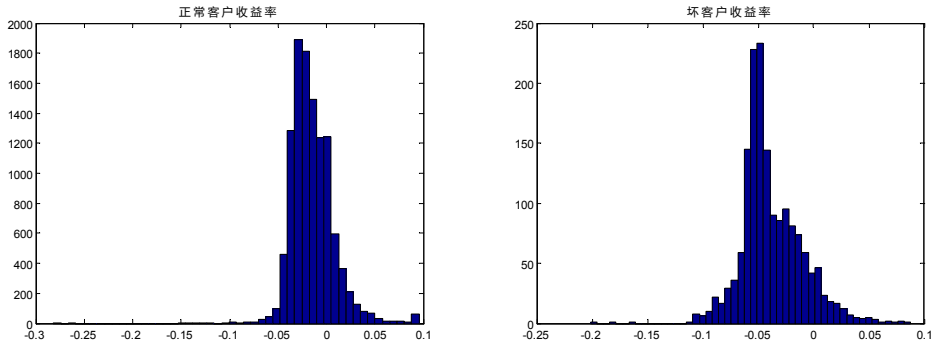
其中：

第一个交易日分类情况：

	当日维持担保比变化情况	客户个数
最坏客户	升幅[-44%,-4.4%]	4356
其他客户	升幅[-4.4%,75%]	18234

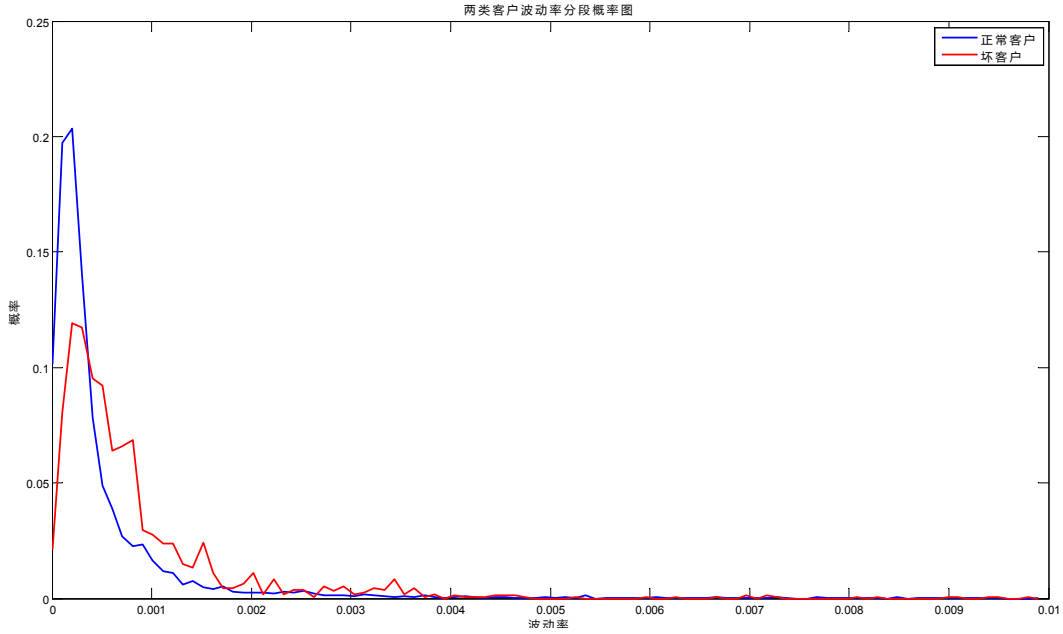
第二个交易日分类情况：

	当日维持担保比变化情况	客户个数
最坏客户	升幅[-45%，4.5]	8586
其他客户	升幅[-4.5%，73%]	15214



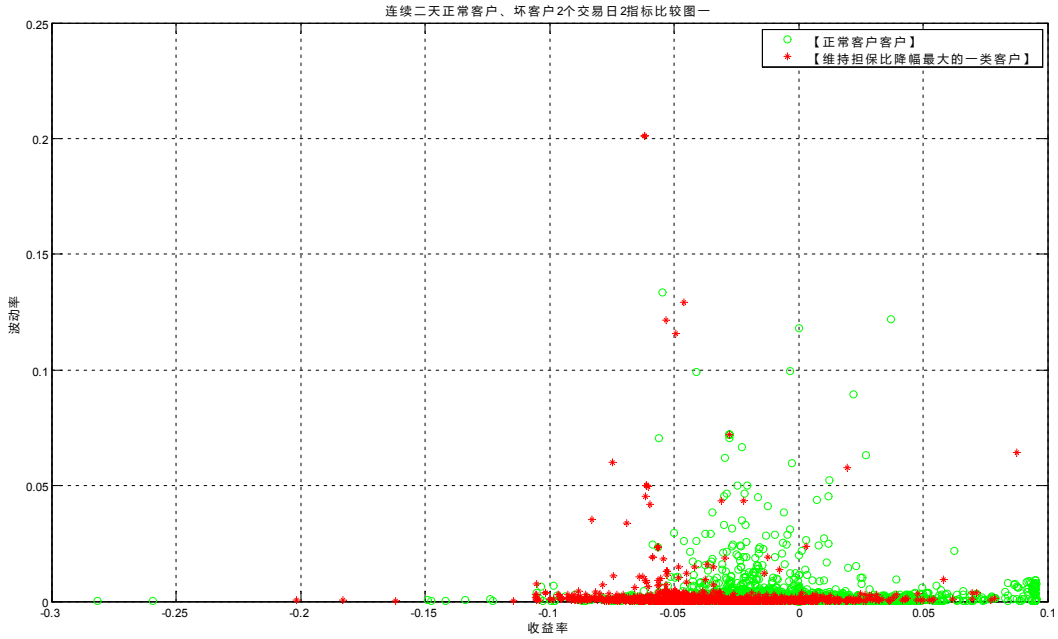
图十三、正常客户、坏客户收益率直方图

正常客户收益率集中的区间[-0.05,0.05],坏客户收益率集中区间[-0.1,0.05],坏客户收益率相对正常客户明显左偏。

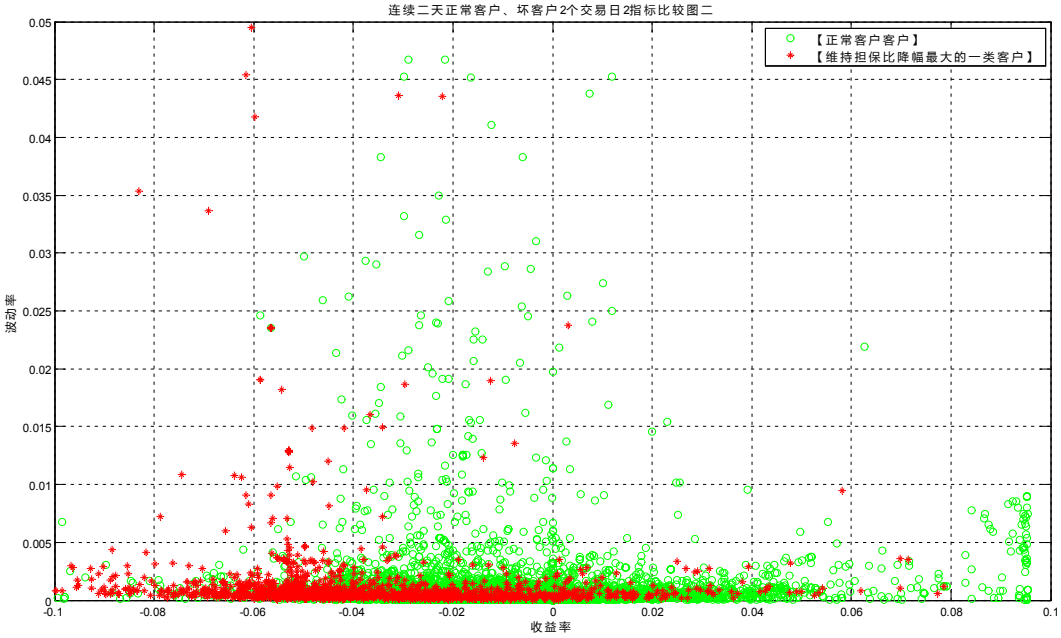


图十四、正常客户与坏客户波动率分段概率图

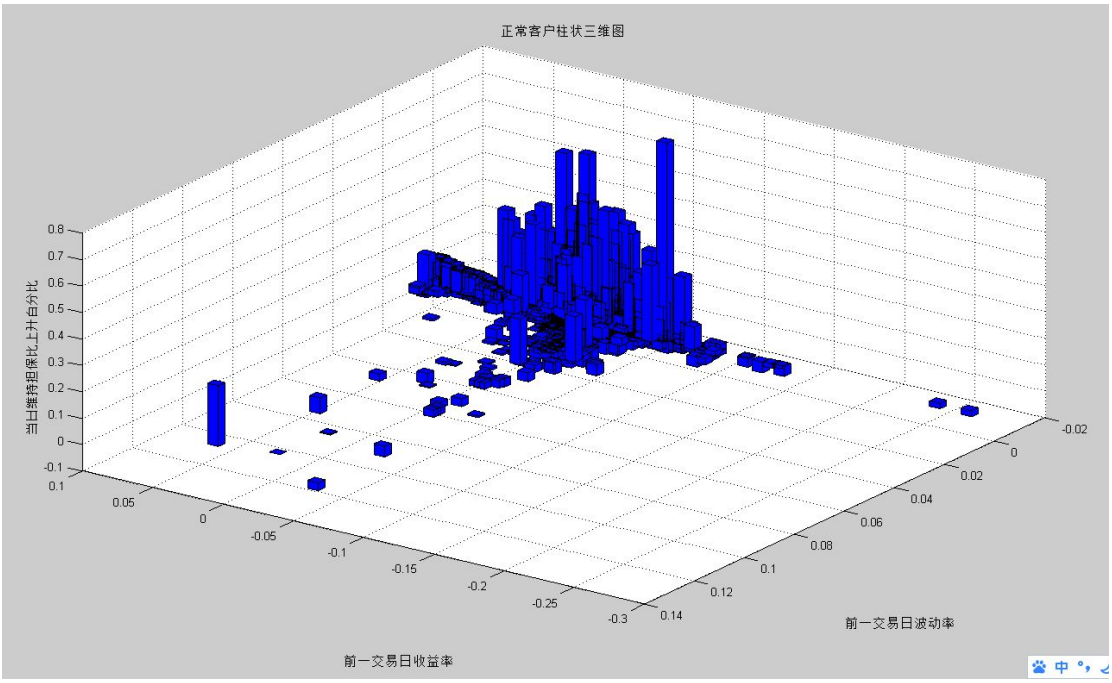
从图十四来看，正常客户在波动率区段[0,0.001]占比重比较多,坏客户在[0,0.002]占比比较多。



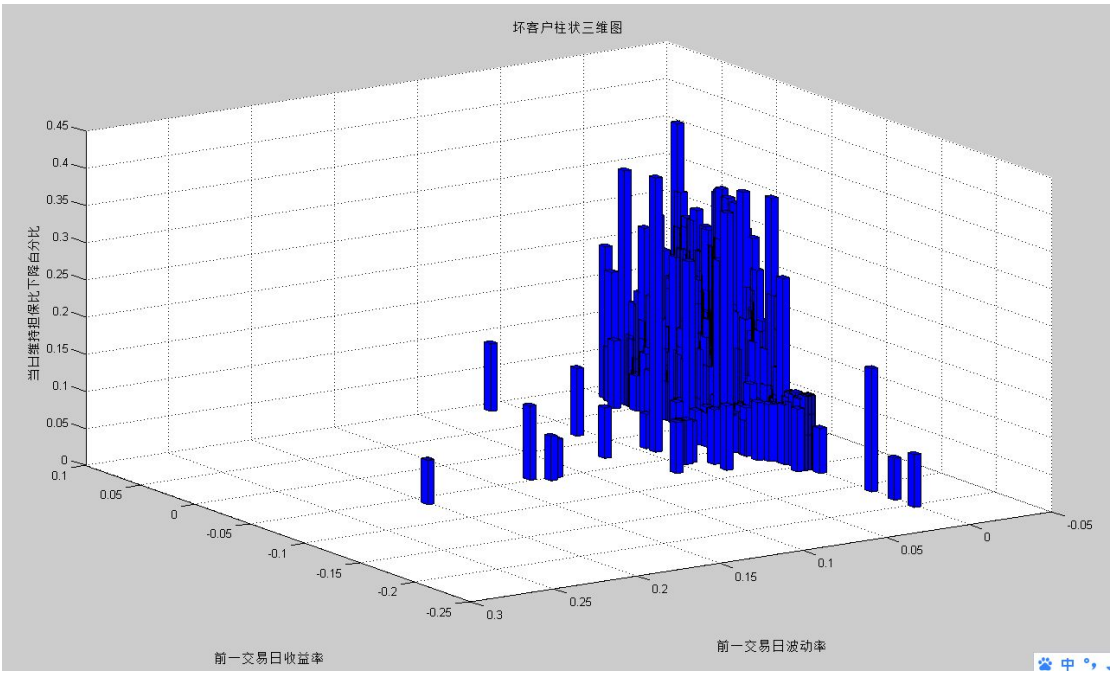
图十五、正常客户与坏客户收益率-波动率比较图



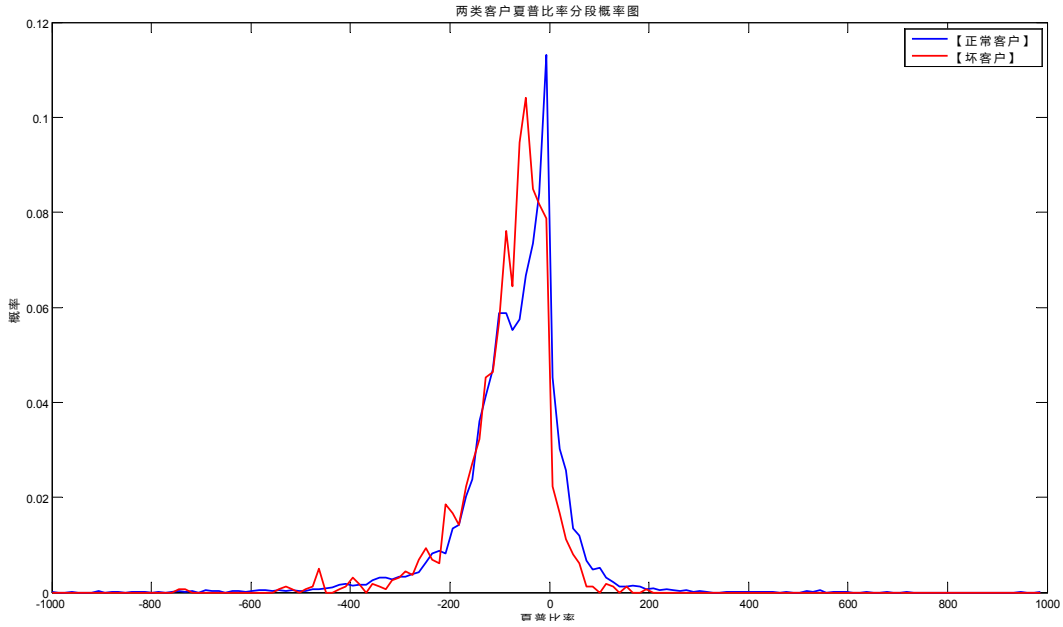
图十六、正常客户与坏客户收益率-波动率比较图局部放大



图十七、正常客户三维柱状图



图十八、坏客户三维柱状图



图十九、坏客户与正常客户夏普比率分段概率图

正常客户的夏普比率相对坏客户右偏。

坏客户的下侧 80%分位点为-15.20，我们如果取[-inf,-15.20]作为坏客户区段，可以挑选出 80%的坏客户，同时这一区段包含了约 67%的正常客户，这部分正常客户被误判为坏客户。在所有[-inf,-15.20]客户中，被误判的正常客户达到了 85%。

坏客户的下侧 60%分位点为-45.13，我们如果取[-inf,-45.13]作为坏客户区段，可以挑选出 60%的坏客户，同时这一区段包含了约 51%的正常客户，这部分正常客户被误判为坏客户。在所有[-inf,-45.13]客户中，被误判的正常客户也达到了 85%。

五、夏普比预判模型的风险和局限性

（1）好客户与坏客户在前一个交易日的持仓特点上，就已经体现出了差异；

（2）从实际计算来看，夏普比率这一指标具有最大的可操作性，在上面 4 种情况当中，用夏普比率基本上能得到两类客户的最大区分度；

（3）区分好坏客户、区分坏客户和正常客户，前者的区分度高，指标差异更明显；

（4）考虑单个交易日维持担保比变化幅度，比连续二个交易日区分度高；考虑连续 2 个交易日，取出来的坏客户数量少，误判率高。虽然能够锁定持续 2 天表现最坏的客户，但是反应不够灵敏，在预测方法上面也有一些问题。通过比较前面的实验结果，建议只考虑单个交易日的维持担保比变化幅度就可以了；

（5）我们现在的水平：在全部坏客户当中找出 70%~80%，会同时将 40%~60%的正常客户误判为坏客户;

（6）做预测：前一交易日通过所有的客户持仓信息计算夏普比率，取夏普比率的一个低区段（区段长度根据需要自己调整），将所有夏普比率落在该区段的客户视为需要重点关注的客户（有一定的误判率），这些客户在下一交易日维持担保比可能是跌幅最大的那一类客户。

六、收益-波动率判别分析模型方法

（一）、前期结果

6.1.1 使用夏普比预判模型的结论

6.1.1.1 夏普比预判模型的结论

（1）好客户与坏客户在前一个交易日的持仓特点上，就已经体现出了差异；

（2）从实际计算来看，夏普比率这一指标具有最大的可操作性，在上面 4 种情况当中，用夏普比率基本上能得到两类客户的最大区分度；

（3）区分好坏客户、区分坏客户和正常客户，前者的区分度高，指标差异更明显；

（4）考虑单个交易日维持担保比变化幅度，比连续二个交易日区分度高；考虑连续 2 个交易日，取出来的坏客户数量少，误判率高。虽然能够锁定持续 2 天表现最坏的客户，但是反应不够灵敏，在预测方法上面也有一些问题。通过比较前面的实验结果，建议只考虑单个交易日的维持担保比变化幅度就可以了；

（5）我们现在的水平：在全部坏客户当中找出 70%~80%，会同时将 40%~60%的正常客户误判为坏客户。

（6）做预测：前一交易日通过所有的客户持仓信息计算夏普比率，取夏普比率的一个低区段（区段长度根据需要自己调整），将所有夏普比率落在该区段的客户视为需要重点关注的客户（有一定的误判率），这些客户在下一交易日维持担保比可能是跌幅最大的那一类客户。

6.1.1.2 对收益-波动率判别分析模型做判别分析的规划

回顾 1.1.1 夏普比预判模型的 6 条结论，我们可以对收益-波动率判别分析模型做判别分析做出与上述 6 条结论分别相对应的 6 条规划：

（1） 做判别分析也需要选择指标，结论（1）中的“持仓特点”是指客户的收益率、波动率以及由这两个指标衍生出来的夏普比率指标，本质上最能体现坏客户与正常客户差异的正是收益率和波动率，所以判别分析所选用的指标仍然是收益率和波动率；

（2） 由于夏普比率是由收益率和波动率衍生出来的，做判别分析时直接列出收益率和波动率作为特征指标，收益-波动率判别分析模型不再使用夏普比率；

（3） 最终目的是要把未来的坏客户挑选出来，判别分析的类别标签只分两类：坏客户、正常客户；

（4） 在判别分析中需要给出客户的类别标签：在训练集中，客户信息包括客户的类别标签作为已知信息；为了在给出待判客户的预测结果后统计误判率，需要给出待判客户的实际类别标签。由结论（4）可知，以单个交易日中客户的维持担保比变化百分比来作为区分坏客户与正常客户的依据比连续 2 个交易日更加简便且有效，我们在做判别分析时也只考虑客户的单个交易日维持担保比变化百分比的表现作为计算类别标签的依据；

（5） 要将收益-波动率判别分析模型的误判率等信息与夏普比预判模型的预测结果作比较；

（6） 判别分析的目的也是做预测，只是预测方法与夏普比预判模型不一样。收益-波动率判别分析模型的预测方法是：指定历史数据作为训练集，不再像夏普比预判模型一样人为画出分界线，反复调试所设定的百分比来给出客户分类，而是使用 Matlab 的判别函数，给出待判客户在未来一段时间的分类状态。

6.1.2 使用夏普比预判模型的客户信息库

参考夏普比预判模型中“比较单个交易日最坏客户与正常客户前一交易日持仓特点”的操作方法，对 6 月 15 日至 19 日的 5 个交易日分别计算当日所有客户的收益率、波动率、客户分类标签、各类客户数量等信息，制作成客户资料库。客户信息概览如下面两张表：

日期	交易日序号	涉及股票个数	记录条数	带类别标签客户总数
2015年6月15日	109	2976	139777	22590
2015年6月16日	110	2994	146053	23800
2015年6月17日	111	3001	144745	24830
2015年6月18日	112	2990	160230	24775
2015年6月19日	113	3023	176403	25942

图表 1：客户数量概览

日期	客户分类	客户个数	维持担保比变化百分比范围	维持担保比变化百分比均值	收益率范围	收益率均值	波动率范围	波动率均值
2015年6月15日	正常客户信息	18234	[-0.04, 0.75]	0.00	[-0.67, 0.10]	-0.01	[0.00, 0.13]	0.0010
	坏客户信息	4356	[-0.44, -0.04]	-0.09	[-0.68, 0.10]	-0.03	[0.00, 0.20]	0.0017
2015年6月16日	正常客户信息	15214	[-0.05, 0.73]	0.00	[-0.70, 0.10]	-0.03	[0.00, 0.19]	0.0011
	坏客户信息	8586	[-0.45, -0.05]	-0.07	[-0.60, 0.10]	-0.06	[0.00, 0.18]	0.0014
2015年6月17日	正常客户信息	23837	[-0.04, 0.99]	0.03	[-0.18, 0.10]	0.02	[0.00, 0.20]	0.0020
	坏客户信息	993	[-0.44, -0.04]	-0.12	[-0.10, 0.10]	0.01	[0.00, 0.20]	0.0022
2015年6月18日	正常客户信息	14613	[-0.04, 0.81]	-0.01	[-0.69, 0.10]	-0.03	[0.00, 0.12]	0.0012
	坏客户信息	10162	[-0.43, -0.04]	-0.07	[-0.88, 0.09]	-0.06	[0.00, 0.35]	0.0017
2015年6月19日	正常客户信息	6123	[-0.04, 0.94]	0.01	[-0.43, 0.10]	-0.04	[0.00, 0.13]	0.0016
	坏客户信息	19819	[-0.42, -0.04]	-0.08	[-0.51, 0.06]	-0.08	[0.00, 0.31]	0.0024

图表 2：各类别客户信息概览

收益-波动率判别分析模型做判别分析的数据全部建立在这个客户资料库的基础上，对坏客户与正常客户的区分方法也沿用夏普比预判模型中的“维持担保比变化百分比处于最小 10%为坏客户，其他为好客户”。

（二）、总体思路

6.2.1 预测方法

运用夏普比预判模型创建客户信息库的方法，可以根据维持担保比变化百分比将每个交易日中的所有客户区分成“坏客户”和“正常客户”两大类，这是实际的分类。和夏普比预判模型目的一样，我们希望以一定置信度预测出后一段时间的坏客户。预测的方法，夏普比预判模型使用的是根据单一的指标（夏普比率）的某一条参考线，来将客户划分成两部分，一部分预测为坏客户，一部分预测成正常客户，在收益-波动率判别分析模型中，我们不再采用这一预测方法，改用判别分析，指定某一部分客户为训练集，训练集中的客户信息有客户的收益率、波动率，客户的类别标签（“坏客户”或“正常客户”）。需要预测的客户作为待判集，待判集中的客户信息有客户的收益率、波动率，客户的类别待判定。

判别分析是对未知类别的样品进行归类的一种方法。判别分析的研究对象有两

类：已知类别的样本称为训练集，未知类别的样本称为待判集。训练集中的样本已经有了分类，根据这些抽取的样本建立判别公式和判别准则，然后根据这些判别公式和判别准则，判别未知类别的样品所属的类别。从理论上来看，判别分析主要包括距离判别、贝叶斯（Bayes）判别和 Fisher 判别，我们直接调用 Matlab 的判别函数 classify 来做判别分析这一环节，对理论部分不再深究，我们处理的是“两总体的距离判别”，classify 函数的 type 参数使用“mahalanobis”：各组的协方差矩阵不全相等并未知时的距离判别，我们对判别公式和判别准则也不深究。

6.2.2 方案描述

根据训练集和待判集的不同取法，我们分别进行以下四种方案。这四种方案是并列的，具体选用哪一种需要参考已知信息的掌握情况。

方案名称	方案描述	举例	实际操作
方案一	待判集与训练集中的信息处于同一个交易日（同一交易日中预测新客户）	在 6 月 16 日所有客户中随机取出一部分作为训练集，其余作待判集	指定比例（35%），分别在 15、16、17、18、19 日做判别并统计效果
方案二	待判集中的信息是训练集的下一个交易日（用今天预测明天）	取 6 月 16 日的全部客户作为训练集，6 月 17 日的全部客户作为待判集	5 个交易日滚动做，共 4 轮，统计效果
方案三	待判集涉及多个交易日（用今天预测未来几天）	取 6 月 15 日的客户作为训练集，同时预测 16、17 日	取 15 日作为训练集，分别预测 16、17、18、19 日
方案四	训练集涉及多个交易日（用过去若干个交易日	取 15 日、16 日作为训练集，预测 17 日	取 15 日、16 日、17 日作为训练集，预测

	预测未来一天或几天）		18 日

图表 3：收益-波动率判别分析模型的四种方案说明

（三）、试验过程

6.3.1 方案一实现过程

方案一：待判集与训练集中的信息处于同一个交易日

6.3.1.1 处理步骤

第一步：取参数

取定每个交易日的训练集中客户占该交易日所有客户的百分比，试验中取为[35%，35%，35%，35%，35%]，这个参数可以调整；

第二步：读数据

对每个交易日，首先读入该交易日所有客户信息，包括收益率、波动率、客户类别标签；

第三步：构造训练集和待判集

打乱原始样本中的顺序，按比例随机抽取出训练集，剩下的作为待判集，整理训练集和待判集的数据格式，使它们符合做判别的统一格式；

第四步：做判别

删去训练集中的无效数据，调用 classify 函数给出待判集中的客户分类标签；

第五步：统计误判率

对每个交易日重复上述操作，将 classify 函数给出的分类结果与客户实际类别作对比，统计每个交易日的误判百分比。

6.3.1.2 部分结果展示

下面列出训练集占 35%时各个交易日的主要判别结果和判别效果：

每个交易日的判别结果：

日期	6月15日	6月16日	6月17日	6月18日	6月19日
当日总样本数	22590	23800	24830	24775	25942
训练集样本数	7906	8330	8690	8671	9079
训练集中无效样本数	0	1	2	0	2
训练集中正常客户数	6339	5321	8351	5115	2119
训练集中坏客户数	1567	3009	339	3556	6960
待判集中样本数	14684	15470	16140	16104	16863
待判集中实际正常客户数	11895	9893	15486	9498	4004
待判集中实际坏客户数	2789	5577	654	6606	12859
待判集中被判为正常客户数	7251	9961	4476	8839	5970
待判集中被判为坏客户数	7433	5509	11664	7265	10893
误判客户数	6238	2184	11352	3137	3966
正常判为坏	5441	1058	11181	1898	1000
坏判为正常	797	1126	171	1239	2966
遗漏坏客户	28.58%	**40.37%**	26.15%	18.76%	23.07%
误判正常客户	**45.74%**	8.89%	**72.20%**	19.98%	24.98%

图表 4：方案一判别效果统计（1）

由上表可以看出：

（1） 误判率大多数维持在 18%至 28%，加粗的 3 个数差一些；

（2） 6 月 17 日对正常客户的误判率太大了。

下表展示待判客户的实际类别以及被判类别各自数量：

日期	实际类别 \ 判定结果	正常客户	坏客户
6月15日	正常客户	6454	5441
	坏客户	797	1992
6月16日	正常客户	8835	1058
	坏客户	1126	4451
6月17日	正常客户	4305	11181
	坏客户	171	483
6月18日	正常客户	7600	1898
	坏客户	1239	5367
6月19日	正常客户	3004	1000
	坏客户	2966	9893

图表 5：方案一判别效果统计（2）

下表是判别的效果：

日期	待判客户数	错判数量	判别正确率	正常客户判别正确	坏客户判别正确率
6月15日	14684	6238	57.52%	54.26%	71.42%
6月16日	15470	2184	85.88%	89.31%	79.81%
6月17日	16140	11352	**29.67%**	**27.80%**	73.85%
6月18日	16104	3137	80.52%	80.02%	81.24%
6月19日	16863	3966	76.48%	75.02%	76.93%

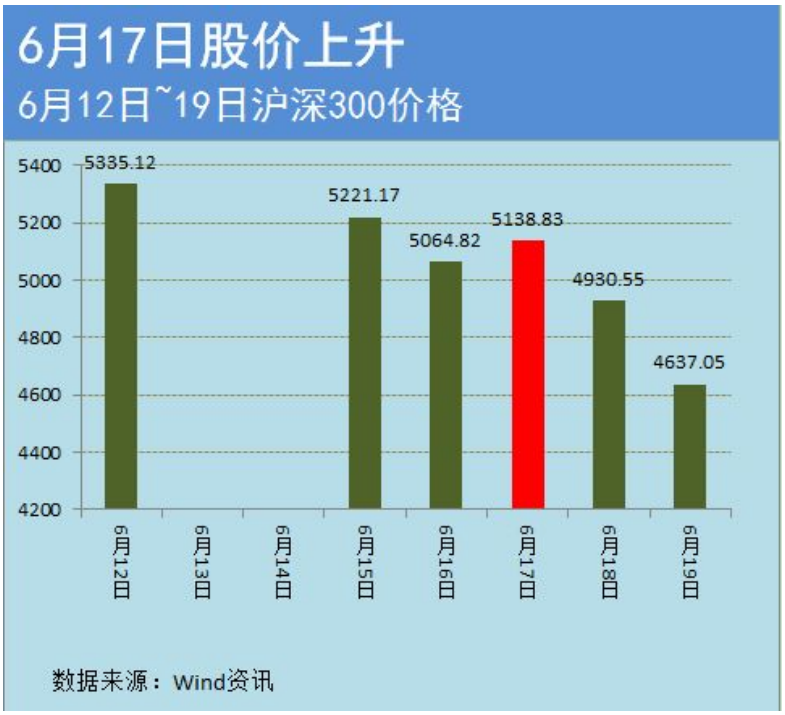
图表 6：方案一判别结果统计（3）

从上表来看，可以发现：

（1） 大部分判别正确率在 72%至 90%；

（2） 6 月 17 日的总体判别率为 29.67%，偏低，主要的是 17 日这一天对正常客户的判别正确率太低，对坏客户的判别率为 73.85%，还是比较好的。17 日判别正确率低是由于将太多正常客户误判为坏客户。

结合前面几张表可知 6 月 17 日这一天实际上坏客户占所有客户的比是比较小的，比其它几个交易日都要小，因为这一天股票市场不同于其它几日的下跌状况，股价在 17 日有所上升，所以 17 日坏客户量较小。17 日显著地区别于其他几日，从股价上也可以看出：



图表 7：股价详情

17 日的实际坏客户并不像其它交易日那样多，做判别时判出了许多坏客户，从而使得误判率较高。可能是判别准则的原因，但是具体为什么还不是很清楚。

6.3.2 方案二实现过程

方案二：待判集中的信息是训练集的下一个交易日

分别用 6 月 15 日判 16 日、16 日判 17 日、17 日判 18 日，18 日判 19 日。

6.3.2.1 处理步骤

第一步：指定目标日期，目标日期的客户作为待判集合，目标日期的前一个交易日的客户作为训练集合，类似方案一的处理方法，读入两个集合中的客户信息；

第二步：做判别并且统计判别结果和误判信息。

6.3.2.2 部分结果展示

下表是判别结果，其中的日期是目标日期，训练日期是目标日期的前一日。

日期与类别	6月16日	6月17日	6月18日	6月19日
遗漏坏客户	42.45%	**99.19%**	0.65%	19.49%
误判正常客户	**77.77%**	0.69%	**99.64%**	**95.93%**
训练集中正常客户数	18234	15214	23837	14613
训练集中坏客户数	4356	8586	993	10162
无效训练样本数	6	4	4	5
待判集中实际正常客户数	15214	23837	14613	6123
待判集中实际坏客户数	8586	993	10162	19819
误判客户数量	15477	1149	14626	9736
被判为正常客户数	7027	24658	119	4111
被判为坏客户数	16773	172	24656	21831
正常被判为坏	11832	164	14560	5874
坏被判为正常	3645	985	66	3862

图表 8：方案二判别结果统计

效果比方案一差很多。

6.3.3 方案三实现过程

方案三：待判集涉及多个交易日

6.3.3.1 处理步骤

第一步：指定多个目标日期，目标日期中的的客户都是待判别类别的，训练集合取为目标日期前一个交易日中的最靠前的那一个，读入所有相关交易日中的客户信息；

第二步：做判别并且统计判别结果和误判信息。

6.3.3.2 部分结果展示

用 6 月 15 日作为训练集，分别去判别 6 月 16 日、17 日、18 日、19 日的客户类别，结果如下表：

日期	6月16日	6月17日	6月18日	6月19日
遗漏坏客户	**42.45%**	**84.69%**	**33.05%**	9.06%
误判正常客户	**77.77%**	14.09%	**86.29%**	**98.78%**
训练集中正常客户数	18234	18234	18234	18234
训练集中坏客户数	4356	4356	4356	4356
无效训练样本数	6	6	6	6
待判集中实际正常客户数	15214	23837	14613	6123
待判集中实际坏客户数	8586	993	10162	19819
误判客户数量	15477	4200	15969	7843
被判为正常客户数	7027	21319	5362	1870
被判为坏客户数	16773	3511	19413	24072
正常被判为坏	11832	3359	12610	6048
坏被判为正常	3645	841	3359	1795

图表 9：方案三判别结果统计

对上表我们给出解释：

（1） 方案三的效果总体上比方案二更差，正确判别的比率更小了，这一点很好理解，方案三是用单个训练日期分别去预测往后若干个交易日，方案二是用训练日期的信息去预测紧随其后的下一个交易日，方案二的表现更好是可以理解的；

（2） 4 个被判日期横向比较，6 月 17 日明显是一个特例，这和前面的发现是同一个问题。可能是因为 15 日、16 日、18 日、19 日都是下跌行情，而 17 日股价是有上涨的，17 日市场行情不同于其它几日，只有 17 日的情境与其它交易日不同；

（3） 遗漏坏客户的百分比与误判正常客户的百分比呈现此消彼长的关系，可以类比于统计上的第一类错误和第二类错误。我们的目的是尽量多地准确地筛选出坏客户，兼顾误判正常客户的比率变化，即主要目的是控制坏客户的遗漏率在较低的水平，其次在同时希望对正常客户的误判率不要过高。常规来看，16 日、18 日、19 日的坏客户遗漏率都不太高，但是正常客户的误判率实在太高。

6.3.4 方案四实现过程

方案四：训练集涉及多个交易日

6.3.4.1 处理步骤

第一步：指定多个目标日期，目标日期中的的客户都是待判别类别的，同时也指定训练日期，读入所有相关交易日中的客户信息，类似前面的处理方法构造出训练集与待判集；

第二步：做判别并且统计判别结果和误判信息。

6.3.4.2 部分结果展示

下面展示用 6 月 15 日、16 日、17 日三个交易日作为训练集，18 日作为待判集的结果，其它的结果在第 4 部分展示。下表是主要结果：

训练日期：6月15、16、17日，待判日期：6月18日			
训练集中所有的客户	71206	待判集中的所有客户	24775
训练集中的正常客户	57271	待判集中的正常客户	14613
训练集中的坏客户	13935	待判集中的坏客户	10162
实际类别 \ 被判类别		正常客户	坏客户
正常客户		5519	9094
坏客户		633	9529
判别正确率统计			
待判集的总体判别正确率		60.74%	
正常客户的判别正确率		37.77%	
坏客户的判别正确率		93.77%	
误判率统计			
待判集的总体误判率		39.26%	
正常客户的误判率		62.23%	
坏客户的遗漏率		6.23%	

图表 10：方案四判别结果统计

由于方案四已知的信息比较多，上面例子中的训练集合比待判集合大得多，所以判别效果比较理想。可以预测出坏客户中的绝大多数，缺点就是对正常客户误判率也比较大。

七、收益-波动率判别分析模型运行与测试

（一）、方案一试验结果

7.1.1 方案一全部试验思路

换参数：训练集中客户数量占当日所有客户量之比

测试参数变动对判别效果的影响，参与测试的训练集中客户数量占当日所有客户量之比主要有 20%，25%，35%。

7.1.2 方案一主要试验结果

7.1.2.1 待判集大小不同

这部分先指定训练集中客户数量占当日所有客户量之比，剩下的全部作为待判集。

训练集占比20%					
待判日期	6月15日	6月16日	6月17日	6月18日	6月19日
误判概率估计值	38.13%	16.20%	44.90%	19.29%	24.14%
该日总样本数	22590	23800	24830	24775	25942
训练样本数	4518	4760	4966	4955	5188
训练集中正常客户数	3651	3070	4763	2905	1228
训练集中坏客户数	867	1690	203	2050	3960
待判样本数	18072	19040	19864	19820	20754
待判集中正常客户数	14583	12144	19074	11708	4895
待判集中坏客户数	3489	6896	790	8112	15859
无效训练样本数	1	0	1	0	1
判为正常客户数	8178	12302	8380	10978	7554
判为坏客户数	9894	6738	11484	8842	13200
误判客户数	8253	2764	11310	3938	5035
正常判为坏	7329	1303	11002	2334	1188
坏判为正常	924	1461	308	1604	3847
总体误判率	45.67%	14.52%	56.94%	19.87%	24.26%
遗漏坏客户	26.48%	21.19%	38.99%	19.77%	24.26%
误判正常客户	50.26%	10.73%	57.68%	19.94%	24.27%

图表 11：训练集占比 20%

训练集占比25%					
待判日期	6月15日	6月16日	6月17日	6月18日	6月19日
误判概率估计值	36.81%	15.55%	46.74%	20.46%	23.16%
该日总样本数	22590	23800	24830	24775	25942
训练样本数	5647	5950	6207	6193	6485
训练集中正常客户数	4548	3749	5979	3627	1522
训练集中坏客户数	1099	2201	228	2566	4963
待判样本数	16943	17850	18623	18582	19457
待判集中正常客户数	13686	11465	17858	10986	4601
待判集中坏客户数	3257	6385	765	7596	14856
无效训练样本数	2	1	2	2	2
判为正常客户数	8575	11475	4012	9545	7030
判为坏客户数	8368	6375	14611	9037	12427
误判客户数	7083	2596	14152	3955	4721
正常判为坏	6097	1293	13999	2698	1146
坏判为正常	986	1303	153	1257	3575
总体误判率	41.80%	14.54%	75.99%	21.28%	24.26%
遗漏坏客户	30.27%	20.41%	20.00%	16.55%	24.06%
误判正常客户	44.55%	11.28%	78.39%	24.56%	24.91%

图表 12：训练集占比 25%

训练集占比35%					
待判日期	6月15日	6月16日	6月17日	6月18日	6月19日
误判概率估计值	34.99%	16.29%	49.63%	20.83%	24.82%
该日总样本数	22590	23800	24830	24775	25942
训练样本数	7906	8330	8690	8671	9079
训练集中正常客户数	6361	5348	8334	5087	2100
训练集中坏客户数	1545	2982	356	3584	6979
待判样本数	14684	15470	16140	16104	16863
待判集中正常客户数	11873	9866	15503	9526	4023
待判集中坏客户数	2811	5604	637	6578	12840
无效训练样本数	1	1	3	3	0
判为正常客户数	7977	9443	1393	8257	6349
判为坏客户数	6707	6027	14747	7847	10514
误判客户数	5732	2413	14196	3455	4198
正常判为坏	4814	1418	14153	2362	936
坏判为正常	918	995	43	1093	3262
总体误判率	41.80%	14.54%	75.99%	21.28%	24.26%
遗漏坏客户	32.66%	17.76%	6.75%	16.62%	25.40%
误判正常客户	40.55%	14.37%	91.29%	24.80%	23.27%

图表 13：训练集占比 35%

从图表 11、12、13 来看，训练集占比的变化对误判率影响率太小，接着试了两个极端情况，训练集占比 5%和 95%：

训练集占比5%					
待判日期	6月15日	6月16日	6月17日	6月18日	6月19日
误判概率估计值	37.43%	14.37%	47.87%	22.77%	25.75%
该日总样本数	22590	23800	24830	24775	25942
训练样本数	1129	1190	1241	1238	1297
训练集中正常客户数	933	767	1184	759	308
训练集中坏客户数	196	423	57	479	989
待判样本数	21461	22610	23589	23537	24645
待判集中正常客户数	17301	14447	22653	13854	5815
待判集中坏客户数	4160	8163	936	9683	18830
无效训练样本数	0	1	0	2	0
判为正常客户数	11694	14881	1664	10587	9434
判为坏客户数	9767	7729	21925	12950	15211
误判客户数	8367	3280	21093	5941	6277
正常判为坏	6987	1423	21041	4604	1329
坏判为正常	1380	1857	52	1337	4948
总体误判率	38.99%	14.51%	89.42%	25.24%	25.47%
遗漏坏客户	33.17%	22.75%	5.56%	13.81%	26.28%
误判正常客户	40.38%	9.85%	92.88%	33.23%	22.85%

图表 14：训练集占比 5%

训练集占比90%					
待判日期	6月15日	6月16日	6月17日	6月18日	6月19日
误判概率估计值	25.75%	14.37%	47.87%	22.77%	24.16%
该日总样本数	25942	23800	24830	24775	25942
训练样本数	1297	1190	1241	1238	23347
训练集中正常客户数	308	767	1184	759	5522
训练集中坏客户数	989	423	57	479	17825
待判样本数	24645	22610	23589	23537	2595
待判集中正常客户数	5815	14447	22653	13854	601
待判集中坏客户数	18830	8163	936	9683	1994
无效训练样本数	0	1	0	2	5
判为正常客户数	9434	14881	1664	10587	909
判为坏客户数	15211	7729	21925	12950	1686
误判客户数	6277	3280	21093	5941	610
正常判为坏	1329	1423	21041	4604	151
坏判为正常	4948	1857	52	1337	459
总体误判率	25.47%	14.51%	89.42%	25.24%	23.51%
遗漏坏客户	26.28%	22.75%	5.56%	13.81%	23.02%
误判正常客户	22.85%	9.85%	92.88%	33.23%	25.12%

图表 15：训练集占比 90%

训练集占比分别取为 5%、20%、25%、35%、90%时，误判率总体上变化不大，体现不出差别，这是因为取定训练集后，待判集直接取为剩余的全部，这 5 种做法中的待判集大小不一致，所以判别正确率表现稳定。

7.1.2.2 待判集大小相同

为了比较判别正确率随训练集大小变化的变化情况，下面训练集占比分别取为当天客户总数的 20%，25%，30%，35%，待判集占比统一取为当天客户总数的 50%。

训练集占比20%					
待判日期	6月15日	6月16日	6月17日	6月18日	6月19日
误判概率估计值	36.51%	16.57%	46.71%	20.45%	24.63%
总体误判率	38.26%	13.65%	81.89%	20.47%	23.21%
遗漏坏客户	32.14%	19.49%	12.26%	17.80%	23.12%
误判正常客户	39.77%	10.41%	85.00%	22.32%	23.52%
该日总样本数	22590	23800	24830	24775	25942
训练样本数	4518	4760	4966	4955	5188
训练集中正常客户数	3668	3024	4793	2911	1205
训练集中坏客户数	850	1736	173	2044	3983
待判样本数	11296	11901	12416	12388	12972
待判集中正常客户数	9065	7653	11886	7331	3040
待判集中坏客户数	2231	4248	530	5057	9932
无效训练样本数	3	1	2	0	2
判为正常客户数	6177	7684	1848	6595	4621
判为坏客户数	5119	4217	10568	5793	8351
误判客户数	4322	1625	10168	2536	3011
正常判为坏	3605	797	10103	1636	715
坏判为正常	717	828	65	900	2296

图表 16：训练集占比 20%

训练集占比25%					
待判日期	6月15日	6月16日	6月17日	6月18日	6月19日
误判概率估计值	39.22%	16.01%	44.70%	21.44%	23.62%
总体误判率	45.18%	15.57%	39.90%	23.20%	22.66%
遗漏坏客户	27.19%	19.39%	51.23%	15.28%	22.01%
误判正常客户	49.43%	13.40%	39.44%	28.78%	24.73%
该日总样本数	22590	23800	24830	24775	25942
训练样本数	5647	5950	6207	6193	6485
训练集中正常客户数	4565	3840	5944	3652	1523
训练集中坏客户数	1082	2110	263	2541	4962
待判样本数	11296	11901	12416	12388	12972
待判集中正常客户数	9141	7590	11928	7269	3065
待判集中坏客户数	2155	4311	488	5119	9907
无效训练样本数	1	2	1	2	2
判为正常客户数	5209	7409	7474	5959	4488
判为坏客户数	6087	4492	4942	6429	8484
误判客户数	5104	1853	4954	2874	2939
正常判为坏	4518	1017	4704	2092	758
坏判为正常	586	836	250	782	2181

图表 17：训练集占比 25%

训练集占比30%					
待判日期	6月15日	6月16日	6月17日	6月18日	6月19日
误判概率估计值	35.79%	15.96%	47.79%	19.86%	24.14%
总体误判率	39.18%	15.13%	64.37%	22.07%	24.53%
遗漏坏客户	30.98%	20.82%	28.33%	16.96%	24.43%
误判正常客户	41.10%	11.89%	65.96%	25.64%	24.86%
该日总样本数	22590	23800	24830	24775	25942
训练样本数	6777	7140	7449	7432	7782
训练集中正常客户数	5454	4561	7164	4390	1859
训练集中坏客户数	1323	2579	285	3042	5923
待判样本数	11296	11901	12416	12388	12972
待判集中正常客户数	9156	7584	11890	7294	3005
待判集中坏客户数	2140	4317	526	5094	9967
无效训练样本数	0	1	2	1	2
判为正常客户数	6056	7581	4196	6288	4693
判为坏客户数	5240	4320	8220	6100	8279
误判客户数	4426	1801	7992	2734	3182
正常判为坏	3763	902	7843	1870	747
坏判为正常	663	899	149	864	2435

图表 18：训练集占比 30%

训练集占比35%					
待判日期	6月15日	6月16日	6月17日	6月18日	6月19日
误判概率估计值	37.60%	16.00%	48.58%	22.63%	23.61%
总体误判率	42.63%	14.31%	84.09%	24.34%	23.58%
遗漏坏客户	28.32%	20.70%	9.66%	14.67%	23.03%
误判正常客户	46.06%	10.73%	87.19%	30.94%	25.35%
该日总样本数	22590	23800	24830	24775	25942
训练样本数	7906	8330	8690	8671	9079
训练集中正常客户数	6392	5292	8321	5037	2166
训练集中坏客户数	1514	3038	369	3634	6913
待判样本数	11296	11901	12416	12388	12972
待判集中正常客户数	9117	7630	11919	7359	3069
待判集中坏客户数	2179	4271	497	5029	9903
无效训练样本数	3	1	1	3	1
判为正常客户数	5535	7695	1575	5820	4572
判为坏客户数	5761	4206	10841	6568	8400
误判客户数	4816	1703	10440	3015	3059
正常判为坏	4199	819	10392	2277	778
坏判为正常	617	884	48	738	2281

图表 19：训练集占比 35%

训练集占比在 20%至 35%范围内，误判率对训练集占比的变化反应不敏感，输出的各项结果都比较稳定。

7.1.3 方案一结论

方案一是用当天的部分客户作为训练集，同一天的另一部分客户作为待判集，通过上面各种对训练集占比参数的调试，我们可以得到下面结论：

（1） 在所有的结果中，6 月 17 日的判别结果是最差的，并且是最不稳定的，原因在第三部分有所讨论；

（2） 除去 6 月 17 日，其他交易日判别效果大致为：总体误判率维持在[15%,30%]范围内，坏客户遗漏率维持在[10%,30%]范围内，正常客户误判率维持在[15%,45%]范围内；

（3） 训练集占比在 20%至 35%范围内，误判率对训练集占比的变化反应不敏感，输出的各项结果都比较稳定。

（二）、方案二试验结果

方案二在第 3 部分中已经全部完成。所得的结论如下：

方案二判别效果很不理想，而且不稳定，这可能是各对两个连续交易日之间的相似程度不同造成的。

（三）方案三试验结果

7.3.1 方案三全部试验思路

除了第 3 部分所举例子，其它思路还有如下 2 种情形：

训练日期	待判日期
6月16日	6月17日、18日、19日
6月17日	6月18日、19日

图表 20：方案二规划

分别做一遍，与第 3 部分所举的例子比较效果

7.3.2 方案三主要试验结果

训练集：6月16日			
日期	6月17日	6月18日	6月19日
总体误判率	4.63%	68.94%	41.79%
遗漏坏客户	99.19%	91.00%	26.15%
误判正常客户	0.69%	53.61%	92.39%
训练集客户数	23800	23800	23800
训练集中正常客户数	15214	15214	15214
训练集中坏客户数	8586	8586	8586
无效训练样本数	4	4	4
待判集客户数	24830	24775	25942
待判集中实际正常客户数	23837	14613	6123
待判集中实际坏客户数	993	10162	19819
误判客户数量	1149	17081	10840
被判为正常客户数	24658	16026	5649
被判为坏客户数	172	8749	20293
正常被判为坏	164	7834	5657
坏被判为正常	985	9247	5183

我们比较关心的项目“坏客户遗漏率”数值偏大，效果不好。

图表 21：训练集 6 月 16 日

训练集：6月17日			
日期	6月18日	6月19日	比例
总体误判率	59.04%	23.71%	
遗漏坏客户	0.65%	0.15%	
误判正常客户	99.64%	100.00%	
训练集客户数	24830	24830	
训练集中正常客户数	23837	23837	训练集坏客户占比
训练集中坏客户数	993	993	3.9992%
无效训练样本数	4	4	
待判集客户数	24775	25942	
待判集中实际正常客户数	14613	6123	二日待判集坏客户实际占比
待判集中实际坏客户数	10162	19819	41.0172%
误判客户数量	14626	6152	76.3973%
被判为正常客户数	119	29	二日待判集被判坏客户占比
被判为坏客户数	24656	25913	99.5197%
正常被判为坏	14560	6123	99.8882%
坏被判为正常	66	29	

图表 22：训练集 6 月 17 日

坏客户遗漏率非常小，但是正常客户误判率太高了。从图表 22 的最后一列来看，训练日期（6 月 17 日）中坏客户占比较小，而待判日期（6 月 18 日、19 日）中坏客户占比较大，训练集和待判集中两个类别比例结构相差太大，这可能是判别效果不理想的原因之一。

7.3.3 方案三结论

方案三结果不理想，判别效果非常依赖判别日期和训练日期内客户的结构的相似性。

（四）、方案四试验结果

7.4.1 方案四全部试验思路

训练日期	待判日期	训练集与待判集所涉及天数
6月15日、16日	6月17日	2:1
16日、17日、18日	6月19日	3:1
6月15日、16日、17日、18日	6月19日	4:1
6月17日、18日	6月19日	2:1
6月15日、16日	6月17日、18日	2:2
……	……	……

图表 23：方案四规划

方案四有许多做法，可以从上表中挑一些有代表性的做一下。从第三部分的结果来看，方案四的判别效果并没有超过方案一。直接做预期判别效果最好的前 4 个交易日预期 6 月 19 日。

7.4.2 方案四主要试验结果

训练集：6月15、16、17、18日，待判集：6月19日		
待判日期	6月19日	
误判概率估计值	22.06%	
总体实际误判率	16.97%	
遗漏坏客户率	2.11%	
误判好客户率	65.07%	
训练客户数	95976	
训练集中的正常客户	71881	训练集中坏客户占比
训练集中的坏客户	24095	25.11%
待判集中客户数	25942	
待判集中的正常客户	6123	待判集中坏客户占比
待判集中的坏客户	19819	76.40%
被判为正常	2558	被判为坏客户的占待判集之比
被判为坏	23384	90.14%
被误判的客户数	4403	
正常被判为坏	3984	
坏被判为正常	419	

图表 24：4 个训练日

图表 24 使用了 4 个交易日去预测 1 个交易日，效果比第 3 部分的方案四效果更好一些：

图表 24 中实际的总体误判率为 16.97%，相比第 3 部分的实际误判率 39.26%下降了不少；坏客户遗漏率仅为 2.11%，相比第 3 部分的坏客户遗漏率 6.23%又下降了不少；对正常客户的误判率为 65.07%，相比第 3 部分的正常客户误判率 62.23%上升得也不太多。

7.4.3 方案四结论

使用待判集的前 4 个交易日客户信息作为训练集，比使用待判集的前 3 个交易日客户信息作为训练集，效果要更好。在所有 4 个方案里面，方案四总体效果最好，主要表现在总体误判率低、坏客户遗漏率低。同夏普比预判模型比起来，正常客户误判率相近，但是收益-波动率判别分析模型方案四的坏客户遗漏率相当低。

八、收益-波动率判别分析模型的风险和局限性

收益-波动率判别分析模型 4 个方案的评价与比较

方案名称	方案简述	总体误判率	坏客户遗漏率	正常客户误判率
方案一	预测同一个交易日	稳定，平均 33.93%	稳定，平均 23.79%	稳定，平均 35.13%
方案二	预测下一个交易日	不稳定	不稳定	不稳定且普遍偏高
方案三	多个待判交易日	不稳定	不稳定	不稳定
方案四	多个训练交易日	39.26%，16.97%	6.23%，2.11%	62.23%，65.07%

图表 25：各方案判别效果数据

（1）其中方案二已经全部试验完，结果不理想，主要是不稳定，如果下一个交易日的股市行情或者客户结构与训练交易日内非常相似，判别效果就会好一些，如果下一交易日与训练交易日的情境相差较大，判别结果就会很差，所以舍弃掉。

（2）方案三的表现与方案二类似，对于固定的训练日期，如果待判别交易日的情境恰巧与训练日期相似，得到的预测效果就会很理想，否则判别效果太差，表现不稳定，也舍弃掉。

（3）方案一由于训练集与待判集处于同一个交易日，使用到的训练信息很新，所以判别结果较好，三种误判率都还可以。表中只展示平均水平，实际上所有情况如第 4 部分总结的：除去 6 月 17 日，其他交易日判别效果大致为：总体误判率维持在[15%,30%]范围内，坏客户遗漏率维持在[10%,30%]范围内，正常客户误判率维持在[15%,45%]范围内。

（4）方案四的优点是对坏客户的遗漏率非常低。但是对正常客户的误判率稍高，只看数值，正常客户的误判率近似于夏普比预判模型的水平。在正常客户的误判率近似相等的情况下，收益-波动率判别分析模型方案四对坏客户的遗漏率比夏普比预判模型要小得多。

下面总结各个方案的特点：

方案名称	优点	缺点	适用范围
方案一	[20%,35%]以内，训练客户占比对预测效果影响不大；结果较稳定；简便；3 种误判率都控制在较好水平。	需要掌握当天[20%,35%]的客户信息，要求比较高	已知当天较多的信息
方案二	无	非常依赖于判别日期与训练日期是否相似，不稳定	舍弃
方案三	无	非常依赖于判别日期与训练日期是否相似，不稳定	舍弃
方案四	对坏客户的遗漏率非常低	需要较多历史数据构造训练集；对正常客户误判率较高	掌握较多历史数据；希望少漏判坏客户或者不介意误判正常客户

图表 26：各方案的优缺点总结

方案四和方案三各有优缺点，对历史信息的掌握度要求也是不一样的。

九、模型评价与对比

（一）、收益-波动率判别分析模型与夏普比预判模型在思想、方法上的差异

收益-波动率判别分析模型采用判别分析，给出历史数据作为训练集，不显式地写出两类客户的分类标准，而是通过现成的判别方法来划分客户；夏普比预判模型对客户的分类，是人为地指定一个坏客户捕捉率，再寻找夏普比率的某条参考线对客户分类。

收益-波动率判别分析模型现在来看，不能控制最终对坏客户的捕捉率，只能指定训练集和待判集，被动接受 classify 函数的分类结果。而夏普比预判模型在这一点上更加灵活，可以根据风险控制程度的要求，自主设定对坏客户的识别水平。

总的来讲，收益-波动率判别分析模型和夏普比预判模型的目的一致，就是为了预测客户在未来某一时段的类别（坏客户与正常客户），都更关注对坏客户的判别。夏普比预判模型在判定坏客户时比较主观，但是对坏客户的捕捉率可以自主掌控，收益-波动率判别分析模型不指定要找出坏客户某一比率，比较客观，但是也失去了一定自主性。

（二）、收益-波动率判别分析模型与夏普比预判模型在判别正确率上的差异

如果不管两个模型在思想和方法上的差别，只关注最终的判别正确率的话，误判率数据的比较如下：

模型名称	模型名称	总体误判率	坏客户遗漏率	正常客户误判率
夏普比预判模型	夏普比率客户分类模型	未计算，不必要算	20%~30%	40%~60%
收益-波动率判别分析模型方案一	收益率-方差判别分析模型： 预测同一个交易日	稳定，平均33.93%	稳定，平均23.79%	稳定，平均35.13%
收益-波动率判别分析模型方案四	收益率-方差判别分析模型： 多个训练交易日	39.26%，16.97%	6.23%，2.11%	62.23%，65.07%

图表 27：各模型的误判率

（1）从坏客户的遗漏率来看，收益-波动率判别分析模型方案四表现最好；

（2）收益-波动率判别分析模型方案一在三个误判率上都处于一个中间水平，但是上表只是展示一个均值。它的效果受到许多参数的影响：选取不同交易日，判别胜率存在差别；选取不同的训练集占比，判别胜率也存在差别。总体上表现还算稳定；

（3）单从判别胜率来看，在同等水平的正常客户误判率下，收益-波动率判别分析模型方案四的表现比夏普比预判模型要好。

（三）、收益-波动率判别分析模型与夏普比预判模型的预测效用

实际上两个模型使用的范围是不一样的，在预测时效和对坏客户的判别胜率的控制上也有差别，并不能像 5.2.2 那样用判别胜率去单一地比较好坏。

模型名称	模型名称	预测时效	对误判率的自主掌控程度
夏普比预判模型	夏普比率客户分类模型	往下预测一个交易日	可以自主调试对坏客户的捕捉率，被动接受正常客户误判率
收益-波动率判别分析模型方案一	收益率-方差判别分析模型：预测同一个交易日	预测同一个交易日的其他客户	无法控制任何一个误判率，被动接受判别分析对坏客户的遗漏率和对正常客户的误判率
收益-波动率判别分析模型方案四	收益率-方差判别分析模型：多个训练交易日	众多历史交易日数据用以预测下一个交易日	无法控制任何一个误判率，被动接受判别分析对坏客户的遗漏率和对正常客户的误判率

参考文献

[1].《MATLAB 统计分析与应用：40 个案例分析》，谢中华，北京航空航天大学出版社，2010.6

[2].《期权、期货及其他衍生产品（第 8 版）》，Hull，J.C.，王勇译，机械工业出版社，2011.9