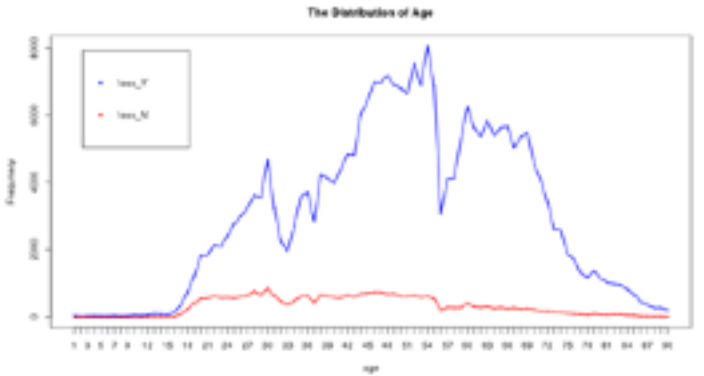
中高端客户流失预警模型

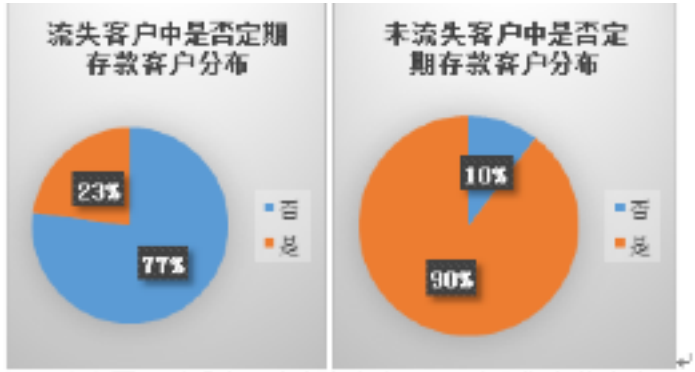
自变量分析与处理

- ✓ 查看分布，对于分类变量，缺失，或某一类别占比>=90%，予以剔除；对于连续变量，0值占比>=90%，予以剔除
- ✓ 剔除统计意义上的弱变量。通过单变量卡方检验，剔除在正负样本上无显著差异的变量
- ✓ 异常值处理: 主要是针对连续变量，超过3倍标准差的数据用邻近正常值替换。
- ✓ 标准化（（变量-均值）/标准差 ）。主要是针对连续变量，以达到去量纲的目的，使得各变量之间具有可比性

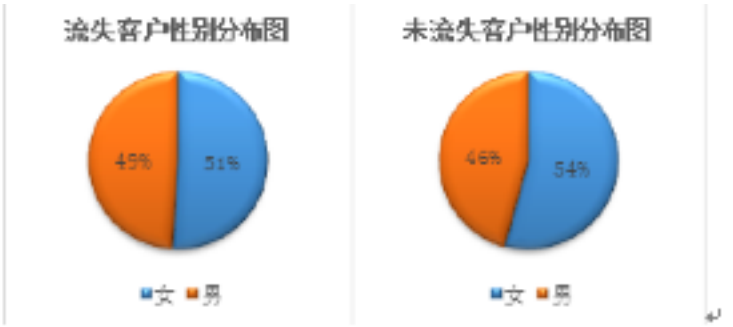


在非流失的客户中，年龄出现多个高峰，主要集中在 42-57 岁，相对流失客户，两者的波动情况相差无几。因此，年龄变里在是否流失客户类别中无显著差异。

- ✓



由上图可以看出，流失客户中，23%为定期存款客户，77%为非定期存款客户，而在未流失的客户中，90%为定期存款客户，10%为非定期存款客户。分布相差较大，说明该字段对是否流失有影响。



性别分布在流失与非流失客户中都比较均匀，在非流失客户中女性的比例相对而言略高，可能是因为现实生活中女性的资产流动较少。

但总体而言，该字段在是否流失客户类别中并无显著差异。

由上图可以看出，流失客户的近三月活期存款日均余额与未流失的分布情况类似。说明该字段对是否流失的影响不大。

数据建模

对剩余的**73个自变量**，使用逻辑回归模型、决策树、随机森林和支持向量机这4种不同的模型，在训练集上进行统计建模，并使用验证集进行模型检验。

模型	AUC	召回率	精确率	F-测度	K-S值
随机森林	0.879	0.888	0.874	0.881	0.761
逻辑回归	0.763	0.676	0.749	0.71	0.449
决策树	0.778	0.695	0.795	0.741	0.7118

逻辑回归、随机森林和决策树这三个模型都可以得到较好的预测效果。其中随机森林效果最优，最终选择**随机森林**用于构建中高端客户流失预测模型。