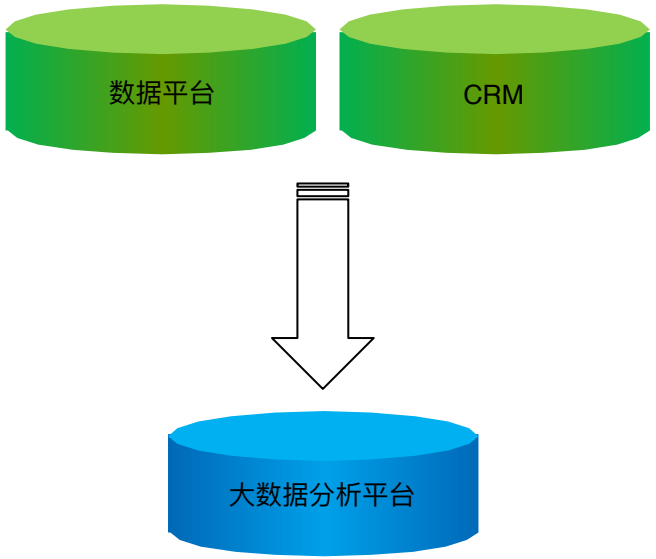
◘ 数据理解与准备

从存款、贷款、信用卡、理财、借记卡方面准备了客户基本信息、产品持有信息、客户交易信息、客户统计信息等，**总计123个字段特征。**

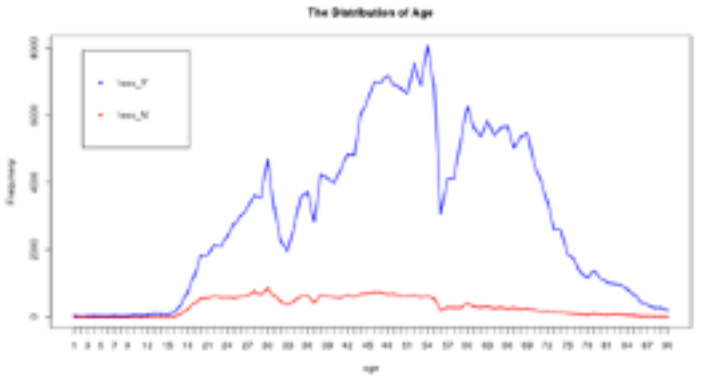


字段英文名	字段属性	字段中文名
UNI_CUST_ID	VARCHAR(32)	客户内码
IDV_BRTH_DT	BIGINT	年龄
IDV_GND_CD	BIGINT	性别
END_CST_DT	DECIMAL(20,2)	本行行龄
CIT_F	VARCHAR(1)	居民标志
MTH_DMD_AVG_BAL_AMT	DECIMAL(20,2)	月存款日均余额
M3_DMD_AVG_BAL_AMT	DECIMAL(20,2)	近3月存款日均余额
M6_DMD_AVG_BAL_AMT	DECIMAL(20,2)	近6月存款日均余额
MTH_DEP_DB_TXN_AMT	DECIMAL(20,2)	月累计借方发生额
MTH_DEP_DB_TXN_NBR	BIGINT	月累计借方交易笔数
MTH_FT_DB_TXN_AMT	DECIMAL(20,2)	月累计借方发生额
MTH_FT_DB_TXN_NBR	BIGINT	月累计借方交易笔数
M3_ACML_FT_DB_TXN_AMT	DECIMAL(20,2)	近3月累计借方交易金额
M6_ACML_FT_CR_TXN_NBR	DECIMAL(20,2)	近6月累计贷方交易笔数
……	……	……
M3_ACML_DEP_DB_TXN_AMT_RT	DECIMAL(22,4)	月累计借方交易金额（近3月相对前3月变化率）
M3_ACML_DEP_DB_TXN_NBR_RT	DECIMAL(22,4)	月累计借方交易笔数（近3月相对前3月变化率）

中高端客户流失预警模型

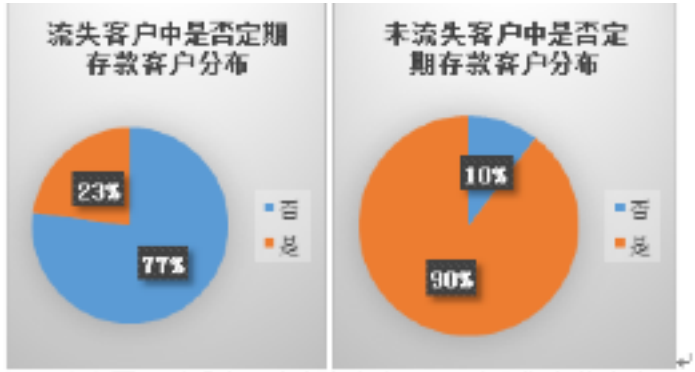
自变量分析与处理

- ✓ 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

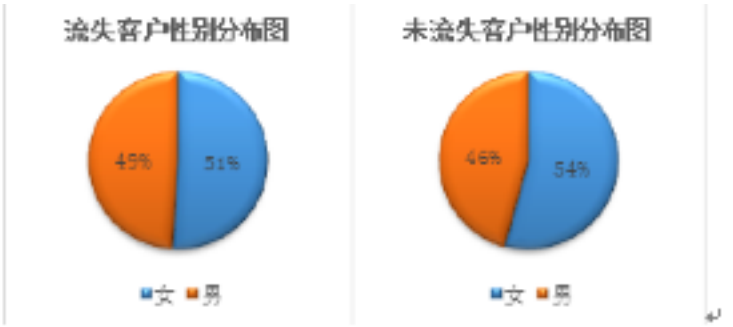


在非流失的客户中，年龄出现多个高峰，主要集中在 42-57 岁，相对流失客户，两者的波动情况相差无几。因此，年龄变里在是否流失客户类别中无显著差异。

- ✓



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。