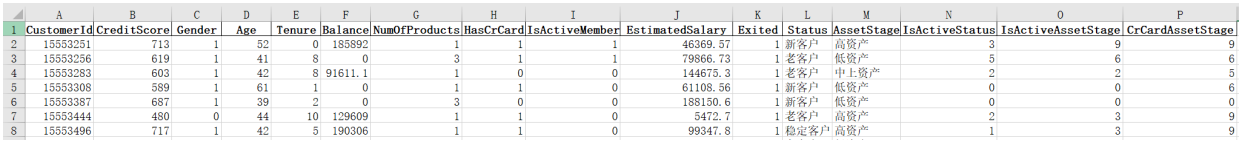
任务五的特征选择

经过任务一、任务三与任务四的数据预处理，我们最终得到的长期客户资源信息数据集如下：



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	CustomerId	CreditScore	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Status	AssetStage	IsActiveStatus	IsActiveAssetStage	CrCardAssetStage
2	15553251	713	1	52	0	185892	1	1	1	46369.57	1	新客户	高资产	3	9	9
3	15553256	619	1	41	8	0	3	1	1	79866.73	1	老客户	低资产	5	6	6
4	15553283	603	1	42	8	91611.1	1	0	0	144675.3	1	老客户	中上资产	2	2	5
5	15553308	589	1	61	1	0	1	1	0	61108.56	1	新客户	低资产	0	0	6
6	15553387	687	1	39	2	0	3	0	0	188150.6	1	新客户	低资产	0	0	0
7	15553444	480	0	44	10	129609	1	1	0	5472.7	1	老客户	高资产	2	3	9
8	15553496	717	1	42	5	190306	1	1	0	99347.8	1	稳定客户	高资产	1	3	9

其中，各个特征的含义与取值如下表所示：

名称	变量含义	变量类型	取值范围
CustomerId	客户ID	字符串	略
CreditScore	表示信用资格	整型变量	[350, 850]
Gender	客户性别	二值变量	{0, 1}
Age	客户年龄	整型变量	[18, 92]
Tenure	账号户龄	整型变量	[0, 10]
Balance	AUM，客户的金融资产	连续变量	[0, 250898.09]
NumOfProducts	客户购买产品数量	整型变量	{1, 2, 3, 4}
HasCrCard	客户持有信用卡状态	二值变量	{0, 1}
IsActiveMember	客户活动状态	二值变量	{0, 1}
EstimatedSalary	客户个人年收入	连续变量	[11.58, 199970.74]
Status	客户状态	字符串	{新客户, 稳定客户, 老客户}
AssetStage	客户资产阶段	字符串	{低资产, 中下资产, 中上资产, 高资产}
IsActiveStatus	新老客户活跃程度	整型变量	{0, 1, 2, 3, 4, 5}
IsActiveAssetStage	不同金融资产客户活跃程度	整型变量	{0, 1, 2, 3, 6, 7, 8, 9}
CrCardAssetStage	不同金融资产信用卡持有状态的特征	整型变量	{0, 2, 5, 6, 7, 9}
Exited	客户流失情况	整型连续变量	二值变量

首先，除去随机排列的用户ID与待预测的客户流失情况；其次，考虑到客户状态与客户资产阶段分别由账号户龄与客户的金融资产分箱得到，而且以这两个特征与为基础，新生成了新老客户活跃程度、不同金融资产客户活跃程度、不同金融资产信用卡持有状态的特征这三个特征，因此我们排除

Status与AssetStage这两个特征；此外，考虑到二值变量所含信息较少，且客户持有信用卡状态、客户活动状态的信息均包含在任务四生成的新特征中，因此我们排除Gender、HasCrCard与IsActiveMember这三个特征；最后，我们计算其余的整型变量、连续变量和客户流失情况之间的斯皮尔曼相关系数，如下表所示：

IsActiveStatus	-0.147899
NumOfProducts	-0.132828
IsActiveAssetStage	-0.084493
CreditScore	-0.026517
Tenure	-0.014312
EstimatedSalary	0.009994
CrCardAssetStage	0.062498
Balance	0.112880
Age	0.317239
Exited	1.000000

我们删去斯皮尔曼相关系数绝对值低于0.05的三个特征，最终筛选出的用于预测银行客户长期忠诚度的特征为：Age、Balance、NumOfProducts、IsActiveStatus、IsActiveAssetStage与CrCardAssetStage