风控数据分析建模

目录

[描述性分析 1](#_Toc42540452)

[建立模型 3](#_Toc42540453)

[数据预处理 3](#_Toc42540454)

[模型评估方法 4](#_Toc42540455)

[随机森林 5](#_Toc42540456)

[逻辑回归 8](#_Toc42540457)

[支持向量分类 10](#_Toc42540458)

[GBDT 11](#_Toc42540459)

[Xgboost 12](#_Toc42540460)

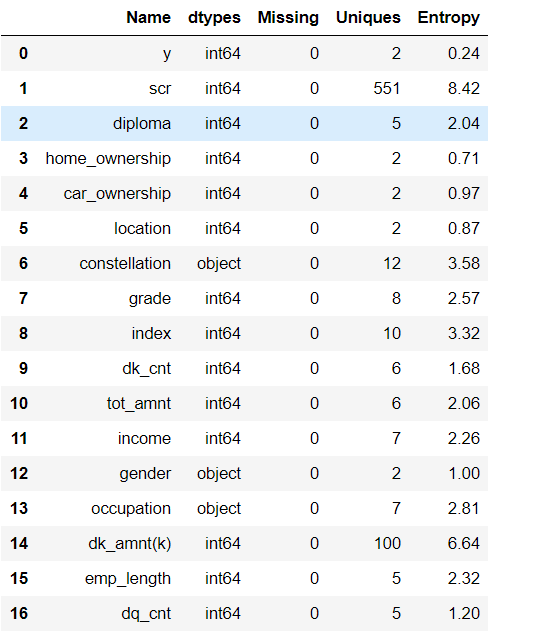
[神经网络 13](#_Toc42540461)

[总结 15](#_Toc42540462)

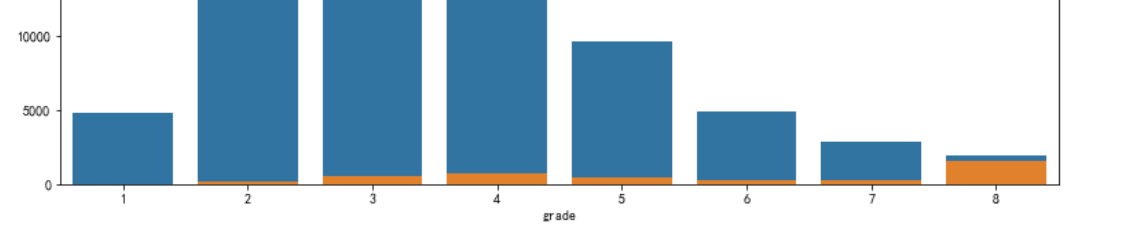
## 描述性分析

通过对数据集各个变量进行统计可以看到其中大部分变量均为分类变量，其中src与dk\_amnt(k)可以看做为数值型变量，其他的变量均分为了2-12不等的类别。其中gender与constellation为字符串形式的分类变量。

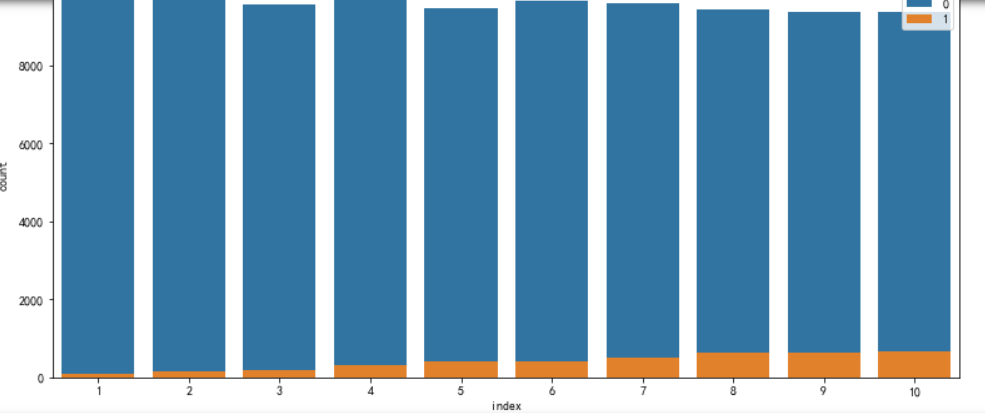
我们的目标变量为Y，统计看到Y中有约4000个违约，占4%。因此可以预见的是我们最后的结果预测准确率并不能做很好的解释与区分，因为即便不做判断把所有的y预测为0也能达到96的准确率。



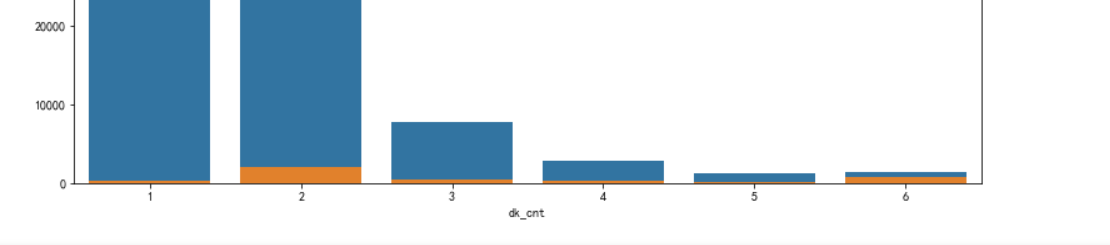
对其中所有变量做柱形图可以看到，在变量取不同分类时，出现y 0-1比例巨大变化的有grade中取值为8的类



Index对于违约率也有一定的影响：



类似的还有dk\_cnt, tot\_amnt, dq\_cnt,home\_owner以及income变量，例如



此外，通过条形图比对可以看到性别、星座、职业、emp\_length(雇佣时间)对于违约率没有明显影响。

描述性分析对于数据展示暂不作更详细探索。如有需要也可以利用条形图、小提琴图进行展示。

## 建立模型

### 数据预处理

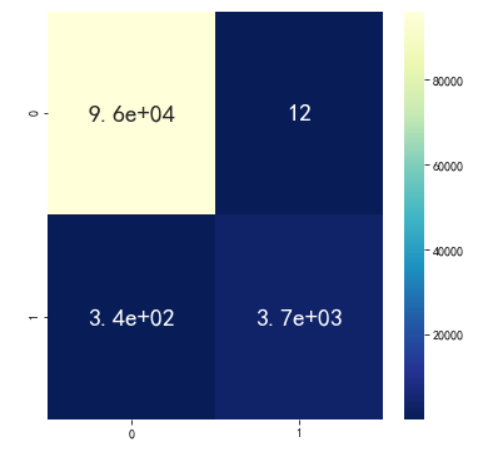
首先查看缺失值，发现并没有确实，其次筛选出需要的数值型变量，这里认为取值较多的scr与dk\_amnt可以看做离散数值变量。其他所有变量分为数值型分类变量与字符型分类变量。

首先对字符型分类变量进行编码。这里用到了两种编码方式，分别是one-hot key用于神经网络与生成dummy变量用于机器学习模型。最后对所有的变量进行规模化到0-1之间。并且分割为x与y。

由于任务要求模型使用全样本，因此本文不对样本进行预测集与测验集的区分。

### 模型评估方法

本文采用分类器的混淆矩阵与ROC曲线作为主要的评估指标。并且考虑到本课题的意义，我们对于y=1(违约用户)的查全率以及犯第二类错误的概率(相当于假正例率FPR)更加看重，前者表示模型筛选出违约用户的能力，我们希望能够尽力筛选出所有违约的用户，此外我们不希望拒绝过多不违约的客户。以混淆矩阵为例：(横坐标表示判断类别，纵坐标表示真实类别)

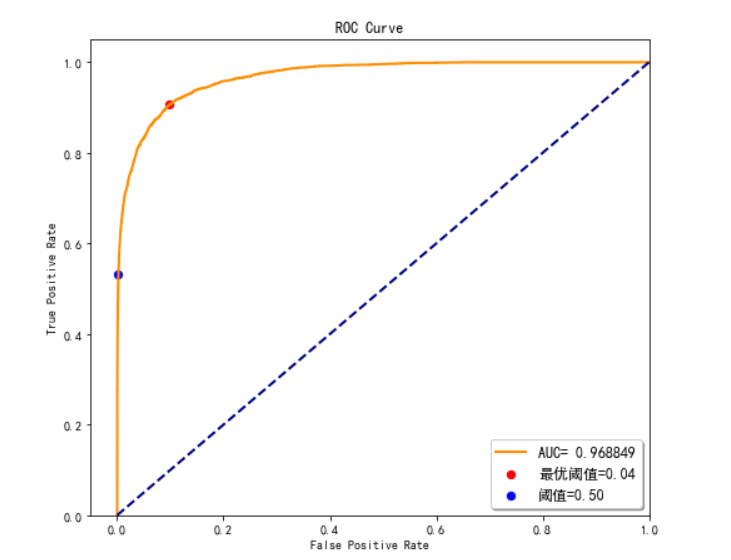


我们希望右下角的TP尽量多，右上角的FP尽量少。体现在ROC曲线上时我们的纵坐标为我们比较看重的召回率，越大越好，而横坐标为我们犯第二类错误的概率越小越好。

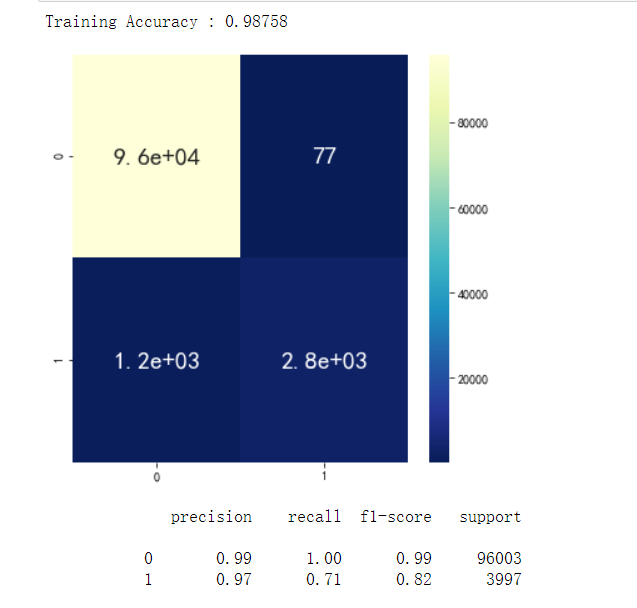
### 随机森林

#### 参数调整

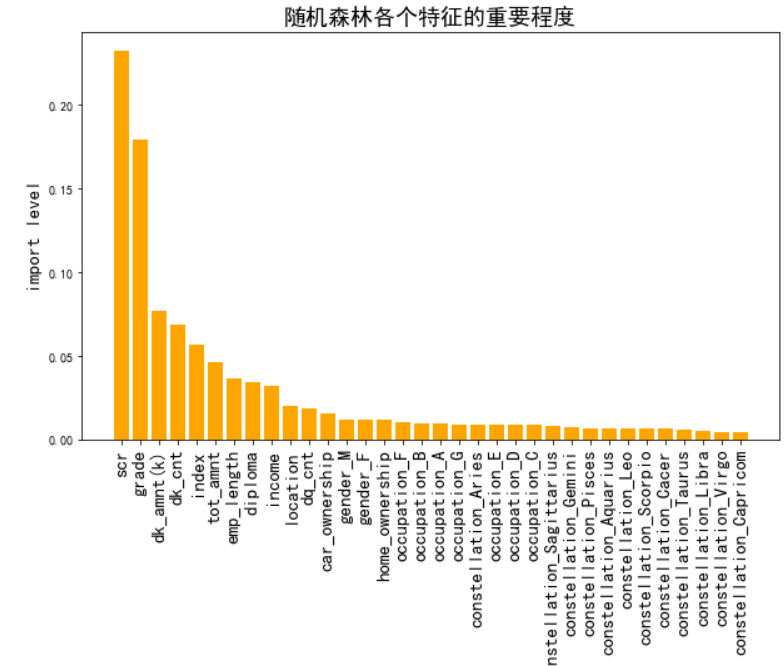
经过实验发现，在默认参数下样本内预测精度达到了1，对袋外样本预测的准确率达到了0.97419。显然出现了过拟合现象。这是因为默认参数对于弱学习器个数没有限制并且对于决策树最大深度没有限制。默认100的学习器个数会产生过拟合的现象，因此当把学习器个数调节到一个较小的10时，依然有非常不错的结果，当我们继续限制最大深度时，当最大深度为10时尽管预测准确率依然很高0.97893，但是此时模型违约用户(y=1)的召回率已经到了0.53。如下图，尽管我们可以通过改变判断概率阈值来提升模型性能，但是代价是犯第二类错误的概率上升，意味着我们会拒绝很多有价值的用户。



因此继续对决策树深度进行调节，当最大深度为15时对于违约用户的召回率为0.71，



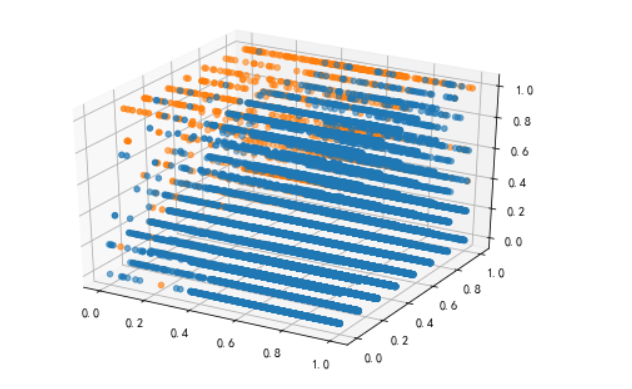
#### 变量重要性分析：



可以看到scr与grade是最重要的两个变量，其他的相对重要的变量还有index,dk\_cnt等，而不重要的包括了星座、性别和职业，这也与我们描述性分析中的特征一致。

#### 3D展示：

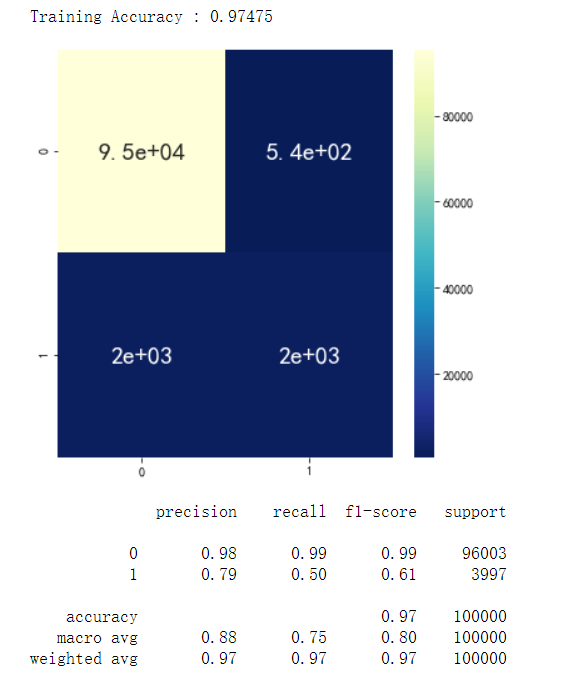
对三个最重要变量进行3D展示，另外还做了主成分分析的3D展示，两者相差不大



可以看到进行了相对有效的分离

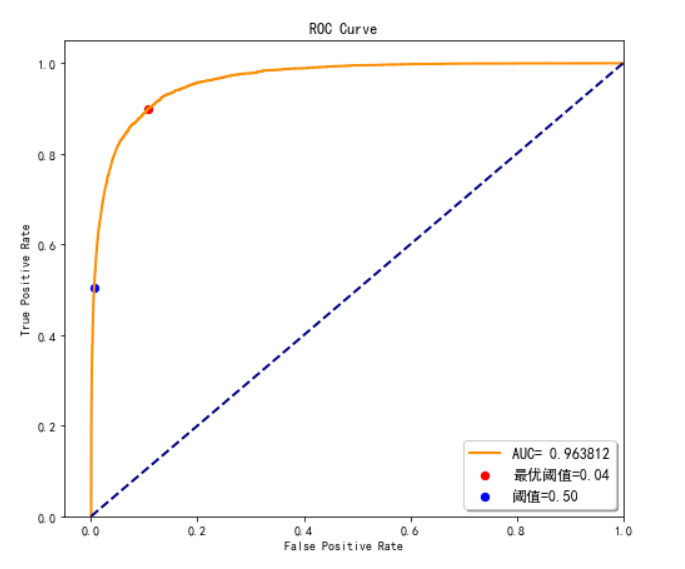
### 逻辑回归

对样本进行逻辑回归：



#### 模型评估：

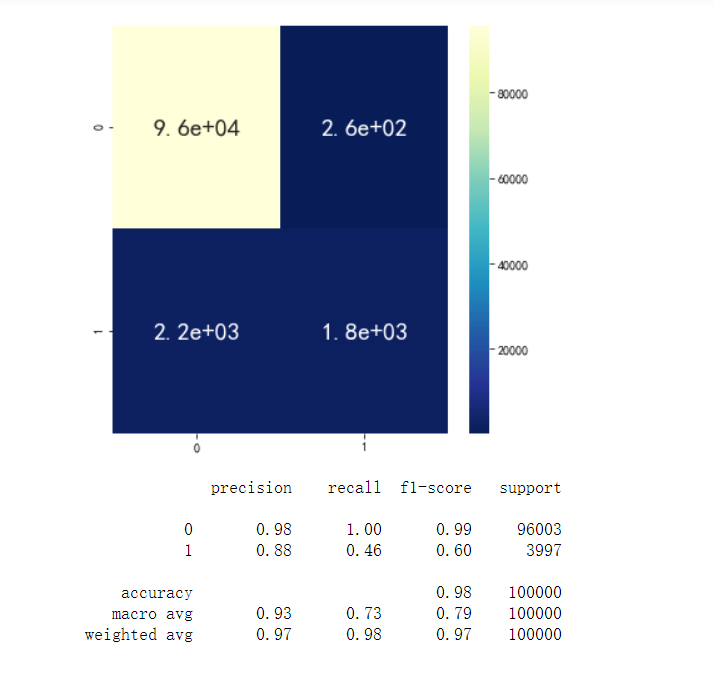
可以看到虽然预测准确率比较高，但是对于违约用户的召回率只有0.5相当于只筛选出了一半的违约用户。



进行阈值调整后，虽然召回率大幅上升，但是考虑到未违约用户的大量人数，FPR(假正例率)的上升幅度是不可接受的。

### 支持向量分类

类似的



同样在高精确率的同时较低的召回率

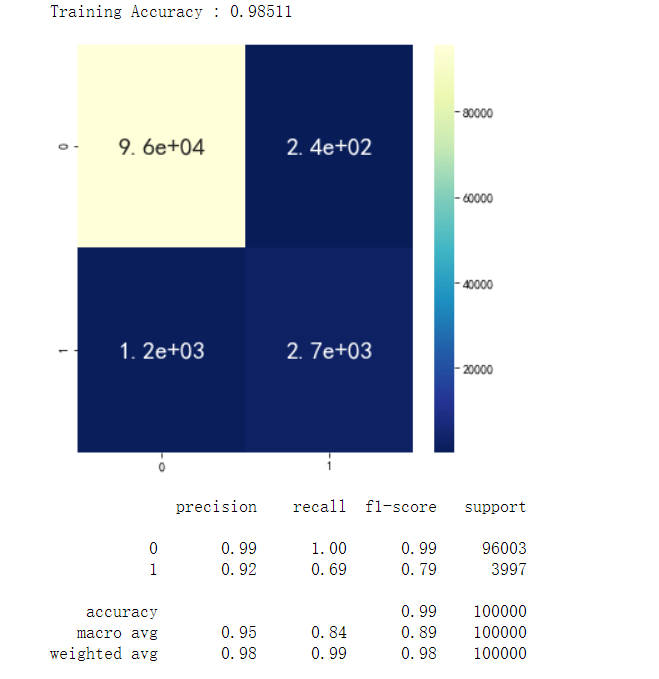
### GBDT

基于boosting 的决策树模型，根据结果相比决策树要较差一些，体现在虽然准确率很高， 但是在与对未违约用户的准确预测，犯第一类错误的概率较高。这是因为，未违约用户占比很大使得划分未违约用户更准确的决策树的权重更大。

由于初始模型表现一般因此不做进一步的调参。

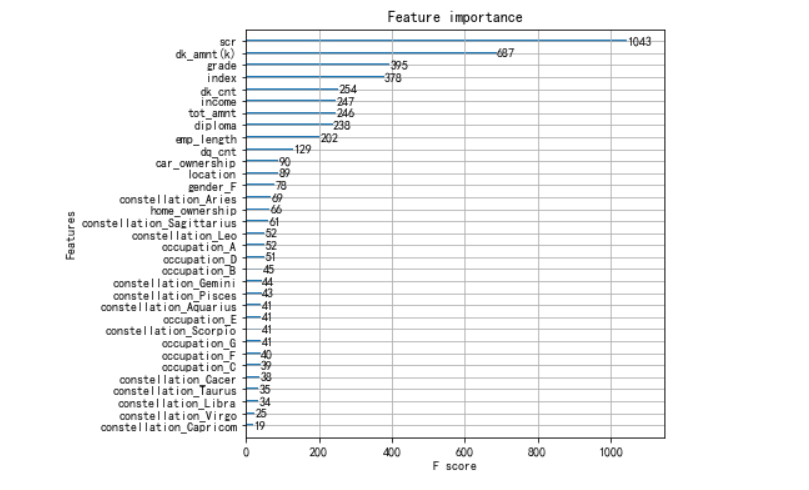
### Xgboost

该模型原理与boosting类似，改模型的准确度接近于随机森林。



达到了很高的准确率与召回率。但是同样的也存在了过拟合的嫌疑。

#### 重要性：



在重要性上得到的结论也与随机森林重要性基本一致。

### 神经网络

#### 数据处理：

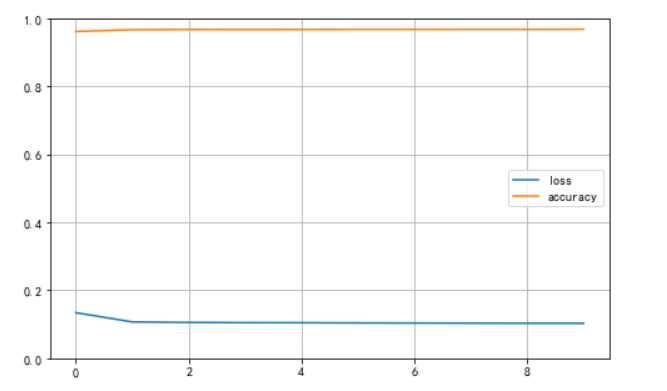
神经网络的数据处理要稍微复杂一些，对于两个数值型变量进行标准化，对其余分类变量进行one-hot映射为特征列。

#### 模型构建：

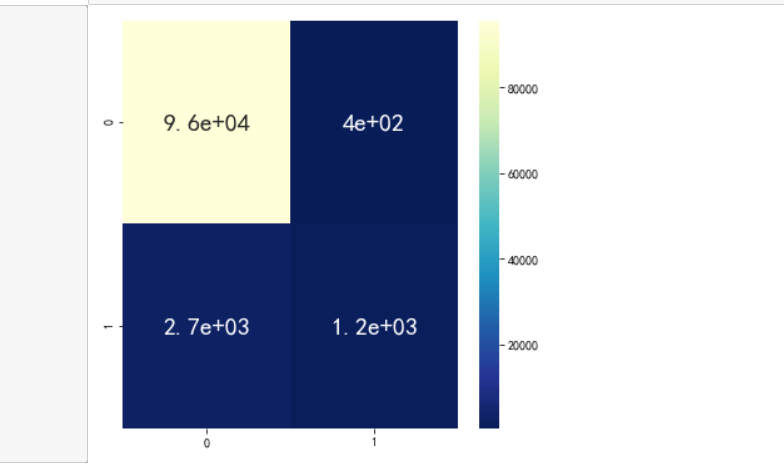
采取的神经网络为输入层-隐藏层-隐藏层-输出层的结构，隐藏成初步为32节点-16层节点，因为我们的变量15个，因此按照2d选择32节点。最后的输出层为1个神经节通过sigmoid激活函数输出一个0-1变量。

#### 训练结果：

通过10批次的迭代后损失函数与准确率趋于稳定为



利用神经网络对原数据集进行预测得到混淆矩阵



可以看到对于Y=1的违约用户筛选能力依然不够好。

## 总结

通过模型比对可以看到，由于数据集中Y=1的违约用户占比很小，这使得许多模型以总体的loss为优化目标进行优化时无法保证对于正例的准确率保证。因此可以选用基于特征的决策树模型，尤其是随机森林模型与Xgboost模型，准确率均在0.98以上.

后续可以考虑提升y=1数据在总样本中的权重来尝试模型性能提升，或者获取更多的违约样本数据来增加模型的预测性能。

后续还可以考虑加入成本考虑或者代价考虑等因素在模型中。

### 拓展(引入权重)：

对于分类y=0与y=1尝试引入不同的权重尝试，赋予了y=1 即违约用户20倍的权重赋值。

在其他参数不变的情况下，随机森林出现了更多的样本被划分为了y=1，对于违约用户的查全率也没有显著的改善，这与我们的初衷不符。因此认为当决策树模型足够好的时候引入权重没有帮助。

对于逻辑回归和支持向量机，确实可以提升违约用户的查全率，但是代价是预测准确率的下降以及犯第二类错误的概率上升，因此需要结合实际进行取舍，是多发现违约用户的代价大还是失去一些用户的代价大。