xgboost部分

数据预处理

利用经过拼接融合、缺失值处理、smote平衡数据处理后的信贷数据train数据集进行xgboost极端梯度上升法的建模。首先采用分层抽样的方法将train训练集分为按照比例划分为训练集train\_train和测试集train\_test。

1、合并变量

MARRY\_STATUS和EDU\_LEVEL是具有重合标签的变量，合并重合标签。MARRY\_STATUS婚姻状况最终被合并为5个类型：“离异，已婚，未婚，丧偶，其他”。EDU\_LEVEL教育状况也被分为5个类型：“研究生，本科，专科，专科以下，其他”。并将所有的名义变量的值利用数字符号表示。

2、one-hot独热编码

XGBoost仅适用于数值型向量。需要将所有其他形式的数据转换为数值型向量。一个简单的方法将类别变量转换成数值向量是一个"独热编码"，只有合法的值是0和1。

将多分类变量MARRY\_STATUS和EDU\_LEVEL进行独热编码，分别获得5个0-1变量，将变量合并到train\_train表和train\_test中。并将 one hot 数值变量代替原始分类变量。

3、XGBoost算法DMatrix矩阵处理

由于 XGBoost 包在进行节点的分裂时，需要计算每个特征的增益，最终选增益最大的那个特征去做分裂。而这又恰恰是各种 GBDT 算法中最耗时间的。所以 XGBoost 想到将各个特征的增益计算就进行多线程进行。那么在定义数据结构时， XGBoost 在 R 重定义了一个自己数据矩阵类 DMatrix。因此，预处理过程和随机森林和Adaboost 算法不同，XGBoost在训练之前，首先要进行数据结构的处理。

经过 XGBoost 数据处理的用于构建模型的数据结构一共产生4个，包括训练 数据集 train，训练标签 train\_label，测试数据集 test，测试标签 test\_label。都是DMatrix类型数据。

信贷数据机器学习算法比较

在实际的信用审批中，许多机器学习算法如随机森林，支持向量机，Adaboost等常被用来判断用户贷款资质。这些算法的优势在于可以通过快捷的计算实现相对理想的预测准确率，部分算法如随机森林等可以根据计算每个特征的增益Gain，来输出对变量特征的排名。利用整理后的个人信贷数据集进行多种机器学习模型的建模，得到的模型预测结果如下表所示：

表：各机器学习算法对比表

|  |  |  |
| --- | --- | --- |
| 机器学习方法 | 预测精度（%） | 运行时间（S） |
| 随机森林 | 93.16% | 44.46294 |
| Adaboost | 93.09% | 23.54862 |
| 支持向量机 | 88.81% | 56.30024 |
| XGBoost | 93.61% | 2.098691 |

先不考虑信贷的实际问题，单纯比较算法性能来讲，可以看出单从准确率上看，集成学习算法之间差距并不是很大，都在93%以上。说明在集成算法上可能都已经逼近了我们提取构造变量的信贷数据数据本身应有的特征，无法有大的改进，但值得注意的是，随机森林和Adaboost都需要进行多棵决策树的生成，迭代次数较多，成千上百次，而XGBoost只需要进行少量的迭代次数，在本次试验中仅需要25次，收敛速度快。而且运行时间上XGBoost的多线程并行运行，也发挥了极大作用，比Adaboost缩短时间达到了近10倍，而比随机森林缩短时间近20倍。在本次信贷的数据分析中，XGBoost更有调节阈值后可以更加符合贷款方的商业要求，更加准确的识别出违约逾期人。

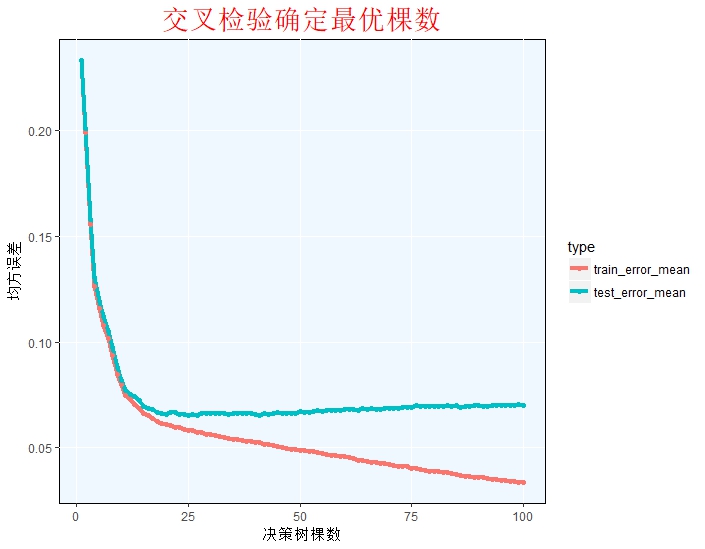
可以看出，四种机器学习算法的预测结果都比较优秀，皆维持在93%以上。其中XGBoost的机器学习算法表现更为优秀，所以在逾期预测模型中，采用XGBoost进行建模分析。

XGBoost模型建立

在R包中的XGBoost的建模函数中只需要进行简单的参数调整即可，学习率利用牛顿优化方方法会很大降低计算效率，因此在这里采用的是自己设定一个超参。

交叉验证确定迭代次数

利用训练集数据进行交叉验证分析。根据交叉验证所得到的测试均方误差和测 试残差标准差，来选取较为合理决策树棵数即最佳迭代次数。交叉验证采取10折交叉验证，基学习器的树深度规定为3，学习率规定为1，目标函数是基于二分类问题的logistic损失进行模型建立的，交叉验证结果如图所示：



从交叉验证图中都可以看出，训练均方误差在随着迭代次数的增加时一直在减少，而测试均方误差则是先减小，而后由于模型的复杂度的提升导致过拟合使测试均方误差增大。测试均方误差在迭代次数为25次时，XGBoost 模型在台湾信贷测试数据集上的表现达到最好。所以在XGBoost模型中选择25为迭代次数。

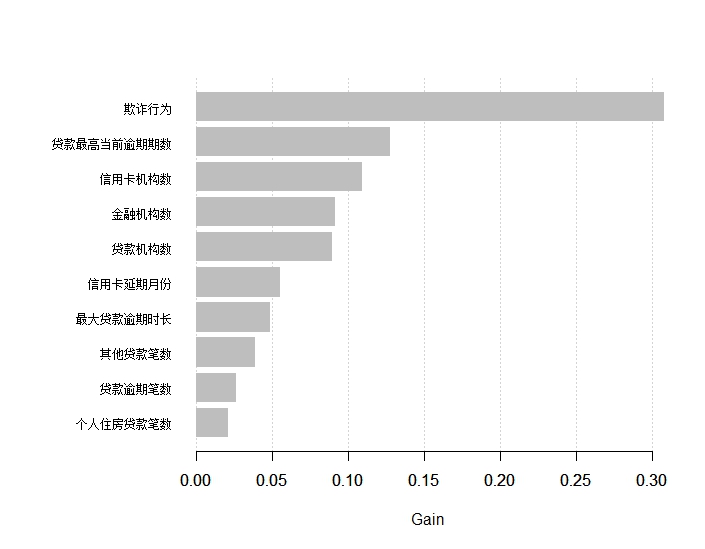
建立 XGBoost 模型

根据交叉验证结果，选择决策树的深度为 3，学习率为 1，迭代次数为25次的， 建立 XGBoost 模型，设置的判断阈值为 0.5。由是单机并行处理，XGBoost模型的训练时间很短，相对于随机森林和 Adaboost 进步很多。模型的训练过程如下表7所示：

表：XGBoost训练过程

|  |  |  |
| --- | --- | --- |
| 迭代次数 | 训练误差 | 测试误差 |
| 1 | 0.233032 | 0.233149 |
| 5 | 0.109234 | 0.110069 |
| 10 | 0.074127 | 0.075664 |
| 15 | 0.065408 | 0.069042 |
| 20 | 0.062009 | 0.066639 |
| 25 | 0.059254 | 0.066580 |

通过模型的迭代次数的增加，模型的表现也在增加，均方误差在减小，XGBoost模型的前十个变量重要性如下图所示：



图：特征变量

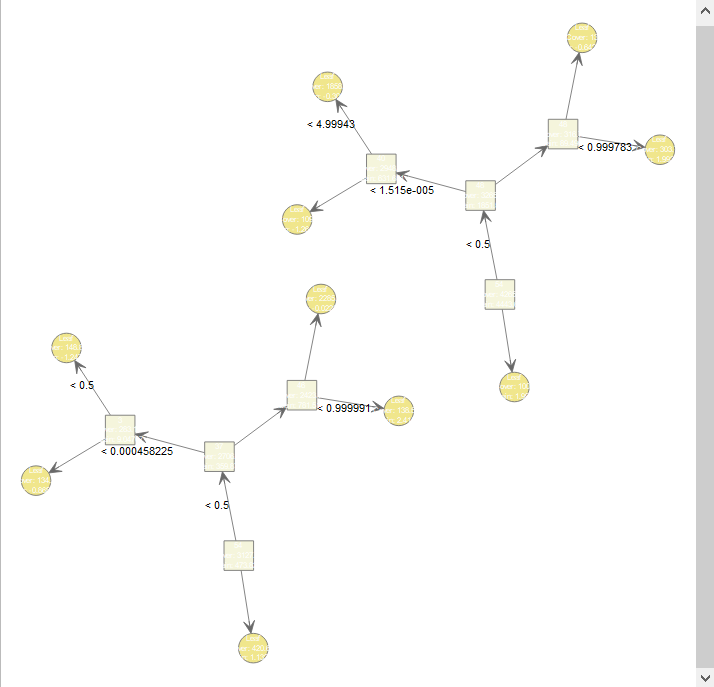
XGBoost模型通过计算自变量对模型的Gain增益的贡献程度评价变量重要性。本信贷案例分析中，Gain是增益，树分支的主要参考因素。模型Gain系数衡量在分类过程中节点的纯度，节点样本分布越平均，Gain系数值越小，节点的纯度越低。根据变量增益衡量的特征指标间的重要性衡量，可以看到，有影响力的变量大致分为了三类：

第一类，用户的信贷历史行为。“欺诈行为”，代表的是申请贷款人在本次申请贷款中是否在其他机构中有欺诈行为，对系统的信息增益极大。贷款最高逾期，信用卡延期月份，最大延期贷款，贷款逾期等信用贷款的违约行为对逾期的影响比较大。越容易产生逾期，也就更不容易获得贷款审批。

第二类，用户的贷款来源状况。包括信用卡机构数，贷款机构数和金融机构数，在越多的金融机构进行借款说明该申请人的负债情况和信用情况比较堪忧，在一家金融机构无法获得更高额度的贷款审批，所以信用状况比较危险，获得审批的机率会降低

第三类，用户目前经济状况。其余的变量对系统的信息增益到比较小，都别划分为第三类。其他贷款笔数和个人住房贷款笔数代表了目前该申请人的负债状况，越多的贷款笔数代表着更高的经济压力，会对申请新的贷款造成更大的影响，不容易获得审批。

从图中可以看出，欺诈行为、贷款最大逾期数和信用卡机构数等几个重要变量会对客户是否会对贷款逾期产生重要影响。信贷机构可以根据这十个重要的特征变量来进一步衡量申请贷款人的信用状况，在这是个变量上表现较好的，可以将贷款的额度提高，合理的安排借贷服务，根据信息增益较高的变量生成的分类树如图所示：



最终的 XGBoost 模型训练参数如下表所示：

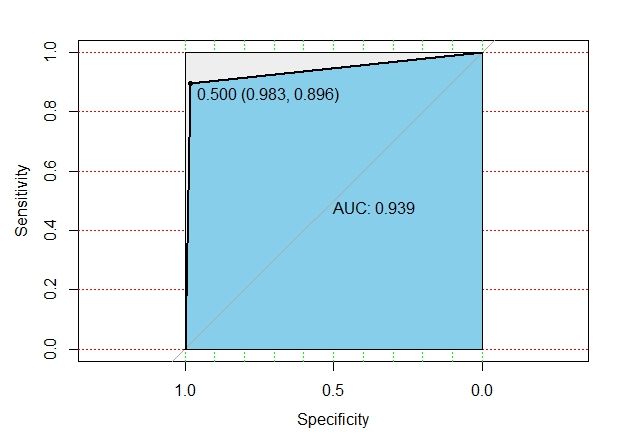
模型评价指标表

|  |  |  |  |
| --- | --- | --- | --- |
| Accuracy | Recall | Precision | F\_measure |
| 93.61 | 89.55 | 98.43 | 93.78 |

在表 3-2 可以看到，模型的训练结果是比较不错的，总体的正确率已经达到了 93.61%，而且召回率也到达了90%左右，准确率也达到 98.43%，但是F值也超过了90%。XGBoost模型的分类混淆矩阵，如表9所示：

|  |  |  |
| --- | --- | --- |
|  | Pred\_0 | Pred\_1 |
| True\_0 | 3323 （45.43%） | 56（0.765%） |
| True\_1 | 411（5.62%） | 3523 (48.17%) |

从表 3-3 中混淆矩阵可以看出XGBoost模型的正确率是很高的。已经可以将 正确率提高到 93.61%左右，而且XGBoost的运算效率很高。但是由于分类器是会犯两种错误的，所以XGBoost的7.39%的预测错误率是包含着将违约者判别到不违约，以及将不违约者分类到违约。由于数据是信贷数据，对于商业信贷机构来讲，两种错误都会给银行带来损失，但相比之下，将违约者判断到不违约给银行造成的损失是比较大的。所以针对本次的信贷数据来讲，分类器的重点应该放在正确识别违约者上，提高真阳性率。为此可以适当的提高假阳性率，也就是未违约者被分类到违约的错误率。根据以上的思想，画出台湾信贷数据的XGBoost分类器的ROC曲线。如图 3-2 所示，AUC取值为 0.939。ROC曲线向左上角的偏移程度很高，AUC数值来说也很高。因此金融机构可以通过XGBoost机器学习模型来获得高预测精度的审批模型，并可以通过Gain增益较高的重要变量进行对贷款申请人的甄别和筛选。



XGBoost模型应用

预测contest\_basic\_test.csv数据集

对于大赛所给的测试数据集，同样经过对contest\_basic\_train.csv数据集相同的数据处理，利用做构建的XGBoost模型，进行每个贷款申请人的是否逾期的预测，预测结果为在10000个用户中，利用XGBoost的预测结果为逾期人数大概为9503人，非逾期人数为497人，而将XGBoost模型对每一个人所判断的（1-逾期概率）作为每个人的信用评分。所得到的逾期预测结果和信用评分在附件表格test\_rs.csv中。