XGBoost结论

集成学习算法通过建立几个模型组合的来解决单一的信用逾期预测问题，无论是 Adaboost 还是随机森林的预测表现均不错。在本文针对大赛提供信贷数据的分类预测中，三种集成分类器预测准确率均在93%左右；特别是在XGBoost模型表现无论是在预测精度上还是在运行速度上都是逼近了数据构造的极限。在Recall和F值上都达到了90%左右，AUC值达到了0.93。针对信贷数据对真阳性率比较敏感的特点，比较注重对召回率的把控，所以构建的XGBoost模型是比较适应于金融以及信贷机构的。

从XGBoost模型所得到的结果中，根据分类节点的信息增益大小来判断出对于逾期有着重要印象的十个变量，总共分为三类变量，第一类，用户的信贷历史行为。包括欺诈行为，贷款最高逾期，信用卡延期月份，最大延期贷款等指标。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等 第三类，包含着其他贷款笔数和个人住房贷款笔数。金融信贷机构面对申请人是除了应该利用XGBoost的最终结果判断申请人是否有资格，更应该着重的观察申请人在这三类指标的表现情况。

利用构建的XGBoost模型预测大赛所提供的contest\_basic\_test.csv数据集，对于10000个申请贷款者进行了逾期与否的判断，并利用（1-XGBoost得到的违约概率）作为个人信用的评分，得到的结果以score和pred两个变量储存于表test\_rs中。

改进方向

1、从指标的构造角度上讲缺乏对交互项以及交叉项升阶的尝试，想要在logit模型中加入更高阶的交叉项，增加显示模型的预测性能，并提高对逾期变量的解释性。

2、对申请贷款的个人评分方面利用的是，XGBoost模型的逾期概率距离1的距离，是比较粗糙的一种评分方法。可以利用在信贷领域比较常见的评分卡模型进行进一步的个人信用评分。

3、对申请贷款的信用额度问题。因为本次大赛所提供的数据的因变量只是逾期与否，并不是欺诈行为和坏账行为，所以在逾期概率并不是很高的情况下，也可以批准借贷，只是批准的金额，也就是引用额度问题要根据信贷机构和申请人一般逾期时间

4、判断下一次借贷是否有逾期行为，还应该与贷款申请人当次所申请的贷款金额有关，所以如果能够收集更详细的信息，可以更加准确地判断贷款申请人的逾期行为是否存在。

5、由于本次的数据分析任务是基于网络借贷平台的，势必存在网络平台固有的弊端，平台不能够强迫申请人谈些所有信息，与之相对，处于对于网络平台信息谢落的担忧，用户大多省略或错填部分重要信息。又或者由于个人对价值衡量标准的不同而填入相异的信息，造成数据的缺失和错误。在本文的数据中，有着极高的数据缺失比例，粗略的插补数值会改变变量的变异，失去分析的价值，导致错误的分析结果。因此，增加数据的真实性和完整性也是急需要解决的问题。

因此，我们希望在决赛中将这些改进加入到论文中，使得分析更加全面，对信贷发放和审批更加科学，达到经济效益的最大化