# Home Credit Default Risk 项目报告

撰写时间：2023年5月19日

撰写人：李林

# ****一、问题提出****

随着金融行业的发展，银行贷款逐渐变成了一件普遍存在的现象，而随之而来的是大量的贷款还款违约。违约会严重影响银行的正常商业经营活动，同时也会影响贷款人的征信记录。为了使贷款更加有效地发放给确实需要且负担得起的人，从而降低银行的经营风险以及保护贷款人的信用评级，有必要采取高效的措施来估计执行贷款发放风险。

目前，银行贷款的借出风险通常由银行的工作人员来评估，工作人员通过对借款人的各种信息的调查，特别是以往借款记录来分析其中的风险。而这个任务可以通过使用软件技术，来更加高效地解决。

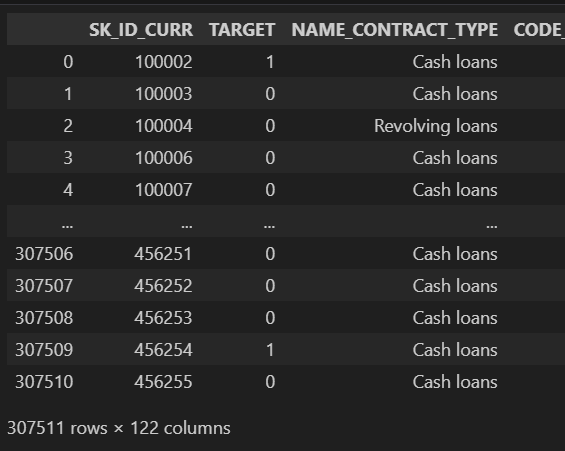
整个问题输入多维度客户信息，输出一个借款风险指标。我们将通过机器学习的方法，来探究客户的各种属性是如何影响最终的结果，最终实现正确预测借款风险指标。

# ****二、数据分析和可视化****

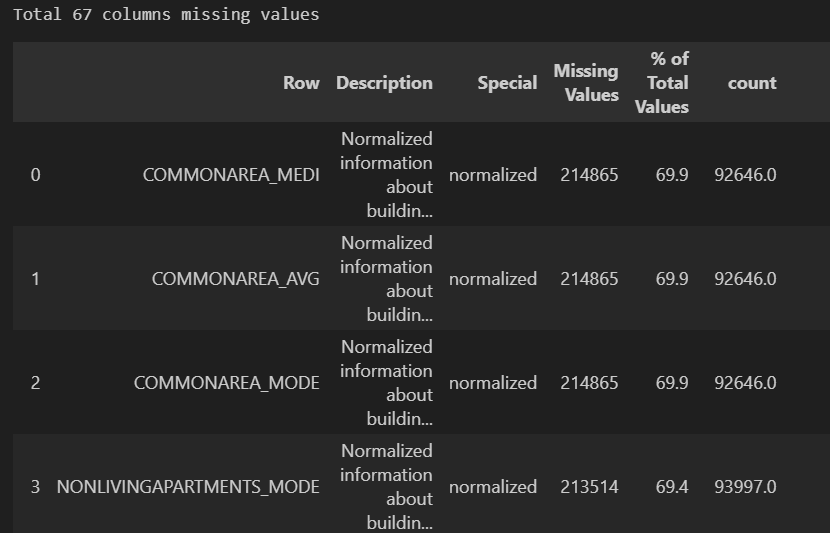
首先分析数据集，Home Credit Default Risk项目中含有的数据集，总共9个文件：

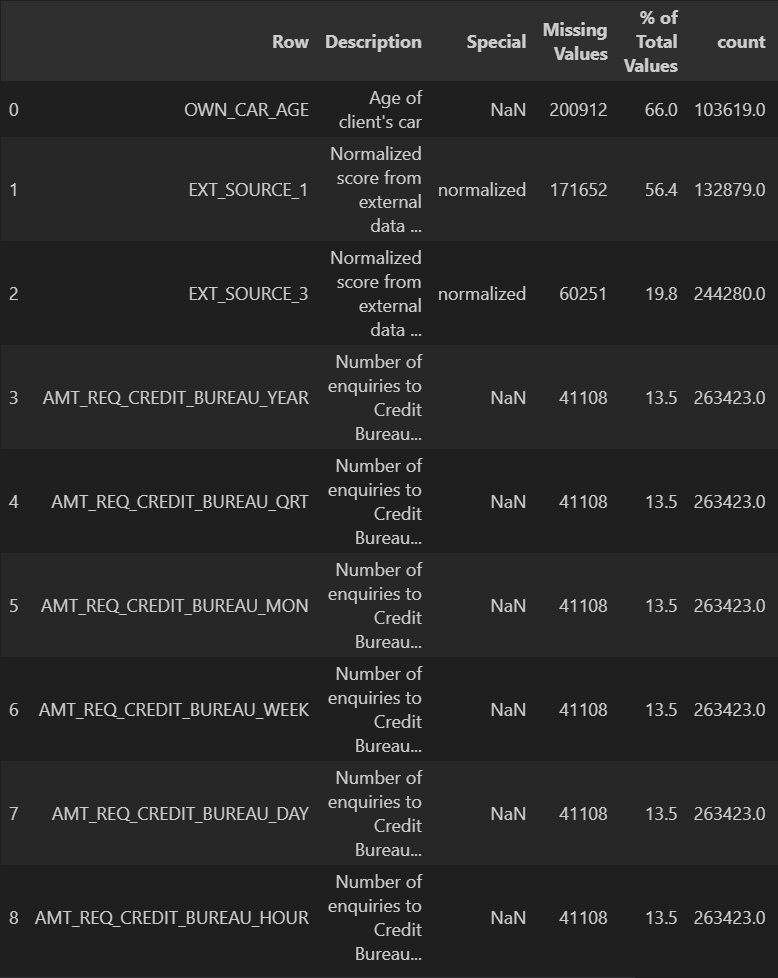
* application\_{train|test}.csv:客户申请表
* bureau.csv/bureau\_balance.csv:客户历史借款记录
* POS\_CASH\_balance.csv:客户POS和现金贷款历史
* credit\_card\_balance.csv:客户信用卡的snapshot历史
* previous\_application.csv:客户历史申请记录
* installments\_payments.csv:客户信用卡还款记录
* HomeCredit\_columns\_description.csv:对各表的列的描述

其中，application\_{train|test}.csv:客户申请表是模型训练与测试的主要依据。

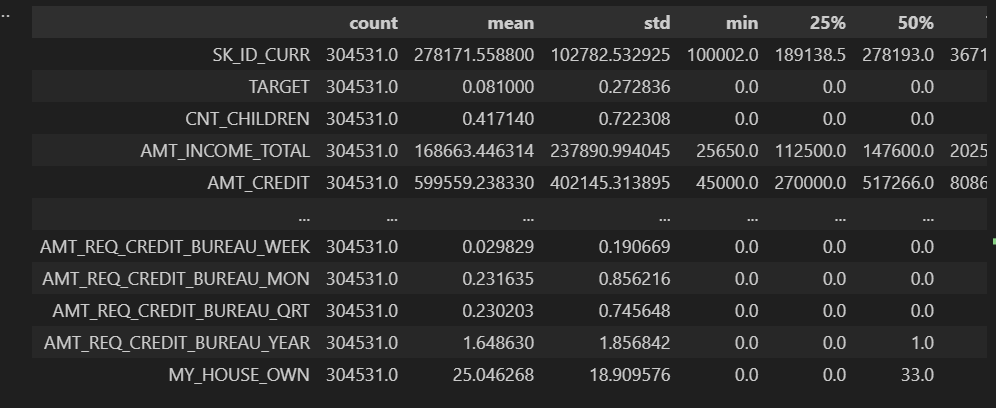
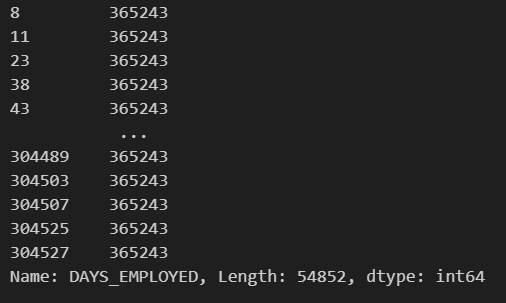
首先查看其基本数据特征：

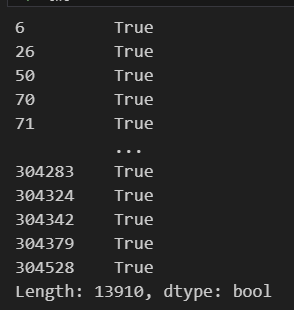
共307511个数据以及122个属性，然后查看缺失值：

缺失值，共67列出现缺失值，最高的一个有70%的缺失率。不过，许多缺失率较大属性都是和住房相关的。此外，也有许多小的缺失值，对于小的缺失率直接删除缺失数据即可。

最后，剩下8个属性，对于年限数据，缺失值可以置零。可以看到EXT\_SOURCE\_1和EXT\_SOURCE\_3比较平滑，直接用平均值代替。AMT\_REQ\_CREDIT\_BUREAU的一列数据用0填补，缺失值可能说明这些客户并未有对应的enquiries。

接下来。查看异常值。

查看到DAYS\_EMPLOYED异常

然后箱线图查看到几乎所有属性均出现了限界外的点，True代表应该删除的点。

经过整个数据可视化以及分析，我们对数据有了更近一步的了解。

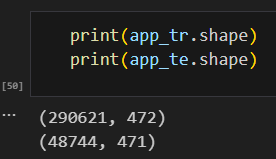
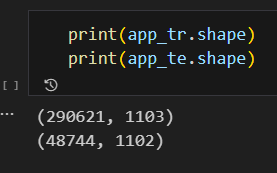
**三、模型探索**

首先本任务是一个二分类概率任务，同时，数据集是表格型数据。所以，应该使用传统机器学习算法来做。这里选择了GBDT来处理。

GBDT（Gradient Boosting Decision Tree）是一种基于决策树的集成学习方法，它通过不断地训练加入新的弱学习器，并在每一轮迭代中加强原模型对于错误样本的学习，从而提高整个模型的准确率。相比于其他集成学习方法，如随机森林和AdaBoost。处理表格型数据时，GBDT具有更好的性能和更高的准确率。这是因为GBDT在处理表格型数据时，能够有效地处理高维稀疏数据和非线性关系，并且能够处理缺失值和异常值。

# ****四、实验过程及选择****

在经过缺失值和异常值处理后，首先通过融合其他表的信息来加强模型对有效属性的理解。然后使用相关性以及特征重要性剔除无关属性：（融合后的属性数量，与选择后的属性数量）

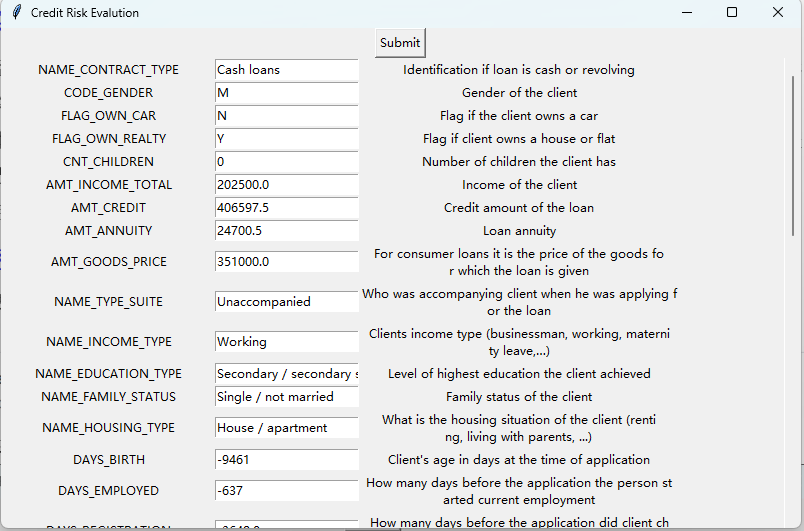


具体而言，将相关性大于0.75的属性剔除，然后将特征重要性为0的属性剔除。剩下的属性就是最终模型选择的属性。

最后，采用xgboost对train数据集进行拟合，并对test集进行测试，得到最终的模型与结果。

# ****五、挖掘结果和展示****

最终结果如下，相比于中期提升了0.01左右。

此外，为了表明本项目的使用价值，还实现了简易的系统UI：

# ****六、结论****

总的来说，本次项目还有很多不足以及可以改进的地方：

1. 数据集的清理仅包含主表，不包含其他表，这可能是添加额外其他特征表效果不明显的主要原因。
2. 数据的选择仍然需要更多的手段来缩减特征数量。
3. UI实现仍比较基础，还无法进行提交，仅仅只是形成了表单。