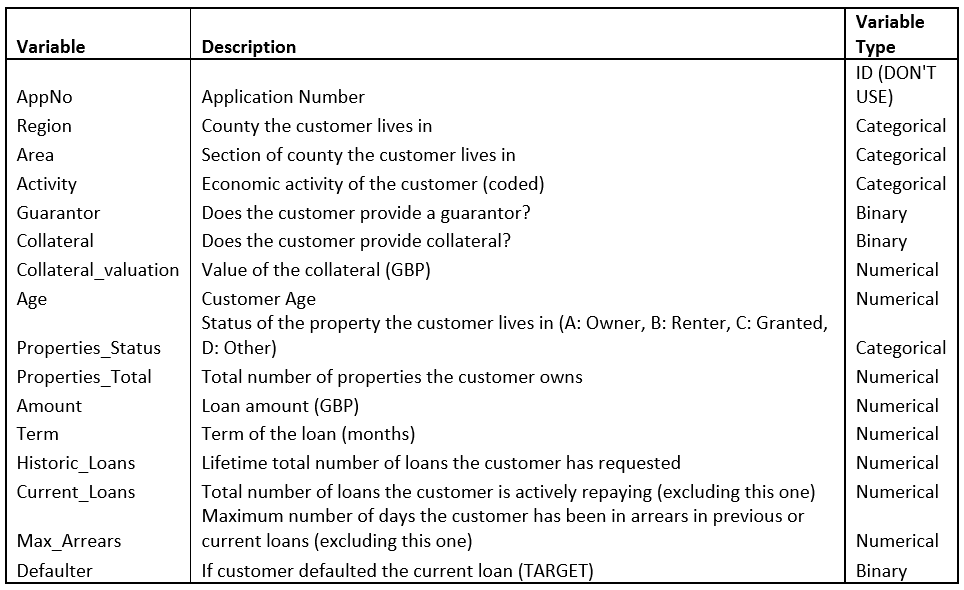
Question 1

MicroSyn数据集”。csv”包括英国一家公司提供给小企业的5万笔贷款信息。您可以使用以下变量:



1.1仔细预处理数据集，考虑以下:

•探索性数据分析

•缺少值处理(如果有的话)。只要用一个值替换掉分数就会打折扣，因此有必要对缺失的值进行正确的研究。

•异常值检测和治疗(如果有的话)。分数会被一个没有理由的值所抵消或取代，因此有必要对异常值进行正确的研究。

•连续变量的分类(如果认为有用)

•使用证据权重对名义变量进行编码(注意可能需要一些额外的粗分类)。

•将数据集分解为训练集和测试集。每个学生都应该以随机的方式单独做这件事，尽可能使用他们的学生证作为随机种子。因此，学生给出相同的参数估计是非常不可信的!对提出相同参数估计的学生将给予特别考虑。

1.2使用**logistic回归**分类器估计计分卡，并报告如下:

•最重要的变量

•变量对目标的影响

•模型性能。使用各种性能指标，并讨论它们之间的关系。

将这个记分卡与随机森林运行数据的结果进行比较，其中没有使用sad和sad转换。讨论你的结果。**为什么银行必须使用逻辑回归作为基本分类器?银行这样做的得失是什么?**