## **项目背景概述**

由于信用记录不足或不存在，许多人难以获得贷款。更不幸的是，这些人经常被不值得信任的放款人利用。

住房信贷通过提供积极和安全的借贷经验，努力扩大无银行账户人口的金融包容性。为了确保这些服务不足的人群拥有积极的贷款体验，家庭信贷利用各种替代数据（包括电信和交易信息）来预测他们客户的还款能力。

家庭信贷目前正在使用各种统计和机器学习方法来做出这些预测，以帮助他们释放数据的全部潜力。这样做将确保有能力还款的客户不会被拒绝，并且贷款的本金、到期日和还款日历将使他们的客户获得成功。

## **业务需求分析**

银行放贷审批核心要素为风险控制，因此对于申请人的审查关注的要点为违约可能性。而违约可能性通常由：申请人收入情况、稳定性、贷款数额和偿还年限决定的。

换句话说，对于任何一个通过了贷款审批的申请人，他可以贷款的上限一定是大于等于申请成功数额的。

因此，在处理异常值时，应遵循的逻辑为：

1. 高收入申请低额度贷款成功的不能算作异常。
2. 高收入申请低额度贷款失败的需要仔细研究原因，进行进一步判断。
3. 低收入申请高额度贷款成功的需要仔细研究原因，进行进一步判断。

至此目标理解完成，我们明白了贷款审批的工作原理以及重要影响因素。在数据理解阶段，就能避免出现不理解业务的可能。

探索性数据分析（EDA）是一个开放式的过程，在这个过程中，我们计算统计数据并绘制数据，以发现数据中的趋势、异常、模式或关系。EDA的目标是了解我们的数据可以告诉我们什么。它通常从一个高层次的概述开始，然后缩小到特定的领域，因为我们发现了有趣的数据领域。这些发现本身可能很有趣，或者可以用来通知我们的建模选择，例如帮助我们决定要使用哪些特性。

目标是我们需要预测的：要么贷款的0按时偿还，要么1表示客户有支付困难。我们可以首先检查每个类别的贷款数量。

## **数据需求分析**

这些数据由某家庭信贷企业提供，这是一项专门为无银行账户人群提供信贷（贷款）的服务。预测一个客户是否会偿还贷款或有困难是一个关键的业务需求，看看使用什么样的模型，机器学习可以帮助他们完成这项任务。

应用训练/应用测试：训练和测试数据，包括有关家庭信贷每个贷款申请的信息。每一笔贷款都有自己的记录行，共307511行，122个字段，并由字段sk\_id\_curr标识。申请数据带有目标，指示0:贷款已偿还，指示1:贷款未偿还，还包含了其他的一些指标例如：申请前一小时向信用局询问客户的情况、申请前一天向信贷局询问有关客户的数目（不包括申请前一小时）、申请前一周向信贷局询问有关客户的数目（不包括申请前一天）、申请前一个月向信用局询问客户的数量（不包括申请前一周）、客户子女数量、客户收入等指标。