微额借款用户人品预测

https://github.com/Jennifer1996/Projects/tree/master/codes/%E5%BE%AE%E9%A2%9D%E5%80%9F%E6%AC%BE%E7%94%A8%E6%88%B7%E4%BA%BA%E5%93%81%E9%A2%84%E6%B5%8B

一、竞赛背景

互联网金融近年来异常火热，大量的资本和人才涌入这个领域发掘富藏价值。金融领域无论是投资理财还是借贷放款，风险控制永远是业务的核心基础。而在所有的互联网金融产品中，微额借款（借款金额 500 元~1000 元）因其主要服务对象的特殊性，被公认为是风险最高的细分领域。本次比赛的主题是通过数据挖掘来分析“小额微贷”申请借款用户的信用状况，以分析其是否逾期。

二、任务

1）构建优秀的分类器；

2）提交分类器对测试集中样本的评分（评分越高，表明样本人品好的可能性越大）。

三、数据

主办方为参赛队伍提供了大量非常宝贵的、来自微额借款行业第一线的实战数据。不仅有常规的带标签数据，还有无标签的数据供大家挑战 semi-supervised learning。数据集中共有 1138 个特征，以用户的多维度行为数据为主。既有数值型特征，也有类别型特征，且均经过脱敏处理。

数据主要包含以下几类：（编码均为 UTF-8）

1）训练集（带标签）：15,000 个样本

带标签的训练集中共有 15,000 个样本。train_x.csv 中存有样本的特征信息，uid 为样本的 id，x0、x1、x2...为特征。train_y.csv 中存有样本的标签信息，uid 为样本的 id，y 为样本的标签：1 为正样本（人品杠杠滴），0 为负样本（人品堪忧）；

2）测试集：5,000 个样本

test_x.csv 中存有测试集的特征信息，格式同 train_x.csv。参赛者的目标是尽可能准确地区分测试集中样本的标签。

3）训练集（无标签）：50,000 个样本

在微额借款的真实场景中，除了放款的客户（人品已知），还有相当一部分被拒绝的客户，他们的人品是未知的。为了提高本次比赛的趣味性与挑战性，我们从中挑选了 50,000 个样本，存在 train_unlabeled.csv 中，格式同 train_x.csv。供参赛者进行 semi-supervised learning 的探索。

4)特征描述：

features_type.csv 为本次比赛的 1138 个特征的类型资料；feature 为特征名：x1，x2，x3...type 为特征类型：numeric（数值型）或 category（类别型）。