**数据说明**

本次比赛提供不同业务的各种产品特征和用户特征数据，格式为UTF8编码格式的excel文件，以逗号分隔。

## 内容

数据集共包含：A.train.csv，B\_train.csv，B\_test.csv三个数据文件。

1.A.train.csv（feature+label）为40k条，贷款期限为1-3年，平均贷款金额为几千至几万的中等额度信用贷款业务A训练数据。

2.B\_train.csv(feature+label)为4k条，贷款期限为7-30天，平均贷款金额为一千的小额短期贷款业务B训练数据。

3.B\_test.csv(feature)为13k条业务B测试数据。

备注：1）提供的数据经过了脱敏，和产品的实际信息有一些差距，但是不会影响这个问题的可解性；2）null代表缺失值。

## 背景介绍

业务A为信用贷款, 其特征就是债务人无需提供[抵押品](http://baike.baidu.com/item/%E6%8A%B5%E6%8A%BC%E5%93%81)，仅凭自己的信誉就能取得贷款，并以借款人信用程度作为还款保证的。

业务B为现金贷，即发薪日贷款（payday loan），与一般的消费金融产品相比，现金贷主要具有以下五个特点：额度小、周期短、无抵押、流程快、利率高，这也是与其借贷门槛低的特征相适应的。

现金贷业务在初期采用了“闭着眼睛放贷”的高利率覆盖风险模式大量的盈利。但是这个模式并不具有可持续发展性，风控将开始成为现金贷业务中后期的关键。然而由于小额现金贷本身的客户由于收入和学历的普遍偏低，大部分人存在潜在的信用风险。

因而随着现金贷市场的成熟，建立行之有效的风控模型，为当前的重中之重。然而由于现金贷目前的产品积累时间不够，没有足够的时间来收集足够的目标变量（flag）建立模型判断来用户的信用风险。

为了帮助银行更好的发展新的无抵押业务，参赛选手可以通过迁移学习来建立新模型，帮助银行风控人员在没有足够时间积累数据的情况下开展新业务。而对于我们征信公司，我们希望通过采用迁移学习的方法，使用小数据样本，就能建立针对不同贷款场景的风控模型。

## 数据字典

训练集里的feature包括了no, ProductInfo\_1, ProductInfo\_2, ... , UserInfo\_1, UserInfo\_2, ... , WebInfo\_1, WebInfo\_2 ...，以及其他除flag字段外小写字母开头的字段。

其中no是每条记录的id，flag字段即是label。训练集A和训练集B的字段相同。

**A.csv：业务A数据字典**

|  |  |  |  |
| --- | --- | --- | --- |
| 变量名 | 类型 | 说明 | 示例 |
| ProductInfo\_1 | string | 用户是否办理了产品1 | 0 |
| Webinfo\_1 | string | 用户是否具有1的网络行为 | 1 |
| UserInfo\_1 | string | 用户特征 | 0 |
| Flag | string | 用户是否为违约用户 | 1 |

**B.csv：业务B数据字典**

|  |  |  |  |
| --- | --- | --- | --- |
| 变量名 | 类型 | 说明 | 示例 |
| ProductInfo\_1 | string | 用户是否办理了产品1 | 0 |
| Webinfo\_1 | string | 用户是否具有1的网络行为 | 1 |
| UserInfo\_1 | string | 用户特征 | 0 |
| Flag | string | 用户是否为违约用户 | 1 |