## 前海征信“好信杯”大数据算法大赛

## 赛题描述：

本次赛事中主办方前海征信开放业务数据，设计国内首个迁移学习赛题：参赛选手需依据给定的4万条业务A数据及2千条业务B数据，建立业务B的信用评分模型。其中业务A为信用贷款, 其特征是债务人无需提供[抵押品](http://baike.baidu.com/item/%E6%8A%B5%E6%8A%BC%E5%93%81)，仅凭自己的信誉取得贷款，并以借款人信用程度作为还款保证；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。

由于业务A、B存在关联性，选手如何将业务A的知识迁移到业务B，以此增强业务B的信用评分模型，是本次比赛考察的重点

## 比赛数据

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

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。

B\_train.csv(feature+label)为2k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。

B\_test.csv(feature)为15k条业务B测试数据。

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。

## 提交要求

1. 选手提交需要提交由no和pred两列组成的UTF8无bom格式的csv文件，逗号分隔。其中no为测试数据里的no，pred为选手预测的对应no的flag值。

2. 提交样例文件，请见文件“submit\_sample.csv”

3. 文件命名规则为prediction\_playername\_date，date要求为yyyymmdd形式

4. 提交截止时间：2017-05-10

5. 评审使用AUC（Area Under ROC Curve）值来判断分类模型的好坏。[参考链接](http://scikit-learn.org/stable/modules/generated/sklearn.metrics.auc.html)

6. 得奖选手要求后续需提供实现代码及说明PPT

7. 要求代码为python/R/Java/C格式

**数据说明**

## 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" \t "_blank)，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。

pred是指概率，就是预测的flag为1的概率

**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 |