**背景**

信用记分卡是银行用来评估潜在借款人违约风险的统计工具。 记分卡是基于历史贷方数据的统计模型，可以参考以确定是否批准贷款。 这一过程有助于银行最大化贷款利润，并防止客户承担无法偿还的债务。

**过程**

我先研究了记分卡模型背后的理论。 一个主要的资源是Naeem Siddiqi的书“Credit Risk Scorecards”，其中详细介绍了信用记分卡背后的理论和发展过程。 本书讲了记分卡建模的所有基本理论基础。 另一篇中文的研究论文详细介绍了用逻辑回归的输出算得分的过程：https：//wenku.baidu.com/view/6a1ca267cc22bcd127ff0cdb.html。

我然后在kaggle.com上找到了样本数据（<https://www.kaggle.com/c/home-credit-default-risk/data>）。 主要用的是application\_train.csv。这个数据集有目标和很多特征。数据结构为行，每行代表一笔贷款。 第一列包含唯一的贷款ID。 第二列包含目标，表示该贷款是否违约（1 =违约，0 =没有违约）。 许多列的其余部分包含贷款接收者各方面的详细信息。 重要的是要注意，有的特徵是分类的，而其他功能是连续的。

然后我开始使用python编写记分卡。 我主要学习并使用'pandas'包进行数据处理，使用'sklearn'包进行统计分析。 每行代码都有comments，因此阅读它是理解实现的最佳方式。 还有一个PowerPoint详细说明了代码的各个部分。 到目前为止实施的功能是：

**编码**

scorecard\_zh.py

**接下目标**

* 做区分分类和连续特徵的功能，并适当地处理每个特徵。 我已经给连续的数据算了WOE。
* 把逻辑回归最好的factor和offset算出来
* 用confusion matrix, KS，ROC, AUC, GINI 等等测模型
* 把最好的特徵算出来