

机器学习

房屋信贷违约风险预测算法

主 研 人：董乐诗

参 研 人：董乐诗

审 核 人：

方 向：工业算法案例研究

版 本 号：A

声明：本作品权益属中冶赛迪。所含信息、专有技术应予保密。未经本公司书面许可，不得修改、复制、提供或泄露给任何第三方。

CLAIM: This work belongs to CISDI, MCC. All information and know-how shall not be copied, duplicated, altered, submitted and disclosed to any third party without the prior written permission of CISDI.

大数据及人工智能部®

中冶赛迪信息技术有限公司

二〇一八年九月

版本更新

日期	版本	更新描述	作者
2019/4/10	A	初稿	董乐诗

选题表格

时间	竞赛名	竞赛背景描述 (50 字以内)	类型 (分类/回归)
2019/4/1	Home credit default risk (房屋信贷违约风险)	通过各种替代数据帮助信用记录不足或根本不存在的人们预测其可信度从而帮助其获得贷款机会	回归

目录

1. 背景描述.....	4
1.1 竞赛赛题描述.....	4
1.2 评估指标描述.....	4
2. 数据来源及描述性统计分析.....	5
2.1 大赛数据来源.....	5
2.2 数据的描述性统计.....	5
2.2.1 数据基本情况描述:	5
2.2.2 数据字段介绍:	5
2.2.3 数据描述性统计.....	6
3. 优秀算法思路.....	6
3.1 方案一.....	6
3.1.1 方案一数据预处理及特征工程部分方案.....	7
3.1.2 方案一模型设计、建立部分方案.....	7
3.1.3 方案一结果、排名等.....	8
3.1.4 方案一算法流程图.....	8
4. 算法比较.....	11
表 4-1 算法比较.....	11
5. 总结与展望.....	11
5.1 总结.....	11
5.2 建模思路.....	11

1. 背景描述

由于信用记录不足或根本不存在，许多人难以获得贷款。而且，不幸的是，这些人口往往被不值得信赖的贷款人利用。Home Credit 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口具有积极的贷款经验，Home Credit 准备利用各种替代数据（包括电信和交易信息）来预测其客户的还款能力。

1.1 竞赛赛题描述

虽然 Home Credit 目前正在使用各种统计和机器学习方法来做出这些预测，但他们希望 Kagglers 帮助他们释放数据的全部潜力。这样做将确保有偿还能力的客户不会被拒绝，给他们贷款的动力和权益。

1.2 评估指标描述

提交结果是根据 ROC 曲线下方的面积评估的。

对于测试集中的每个 SK_ID_CURR，参赛者必须预测 TARGET 变量的概率。提交文件应包含标头，并具有以下格式：

```
SK_ID_CURR,TARGET
100001,0.1
100005,0.9
100013,0.2
etc.
```

2. 数据来源及描述性统计分析

2.1 大赛数据来源

数据来源：本次大赛的数据来源于 Home Credit 公司

本次比赛官方提供的数据如下：

[Home-credit-default-risk](#)

官方数据下载地址如下：

<https://www.kaggle.com/c/home-credit-default-risk/data>

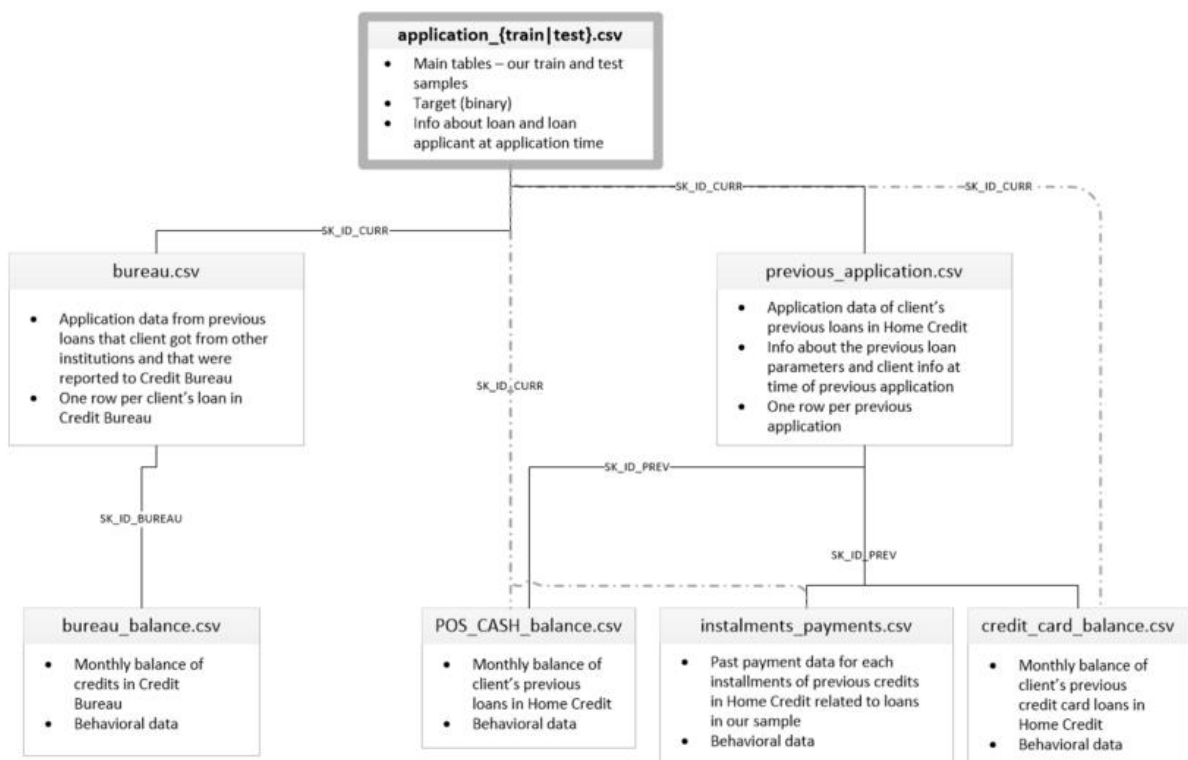
2.2 数据的描述性统计

2.2.1 数据基本情况描述：

官方提供的数据集包含如下文件：

- application_{train|test}.csv
- bureau.csv
- bureau_balance.csv
- POS_CASH_balance.csv
- credit_card_balance.csv
- previous_application.csv
- installments_payments.csv
- HomeCredit_columns_description.csv

所有数据文件之间的关系如下所示：



2.2.2 数据字段介绍

- application_{train|test}.csv

这个是主要的数据表，分为 Train（有 TARGET）和 Test（没有 TARGET）两个 CSV 文件。其包含所有应用程序的静态数据，每一行代表我们数据样本中的一笔贷款。

- bureau.csv

这个表格包含所有客户以前由其他金融机构提供给信用局的信用报告（包含表格中所有贷款的客户）。对于样本中的每笔贷款，客户在申请日期之前在信用局拥有的信用数量与行数一样多。

- bureau_balance.csv

信贷局以往学分的每月余额。该表每个月向历史局报告的每个信用证的历史记录都有一行

- POS_CASH_balance.csv

申请人拥有房屋信贷的以前 POS（销售点）和现金贷款的月度余额快照。

此表在我们的样本中与贷款相关的家庭信贷（消费者信贷和现金贷款）中每个先前信用的每个月的历史记录中有一行 - 即表格（样本中的#loans *相对于先前信用的数量*# 个月） 其中 Home Credit 有一些历史可观察到以前的信用）行。

- credit_card_balance.csv

申请人拥有 Home Credit 的以前信用卡的每月余额快照。此表在样本中与贷款相关的房屋信贷（消费信贷和现金贷款）中每个先前信贷的每个月历史记录中都有一行 - 即该表格（样本中的#loans *以前相对信用卡的数量*# of 几个月 Home Credit 有一些历史可以观察到以前的信用卡行。

- previous_application.csv

这个表格包含了所有先前申请的房屋信贷贷款的客户在的样本贷款。在 Home Credit 的数据样本中，每个先前的应用程序都有一行与贷款相关。

- installments_payments.csv

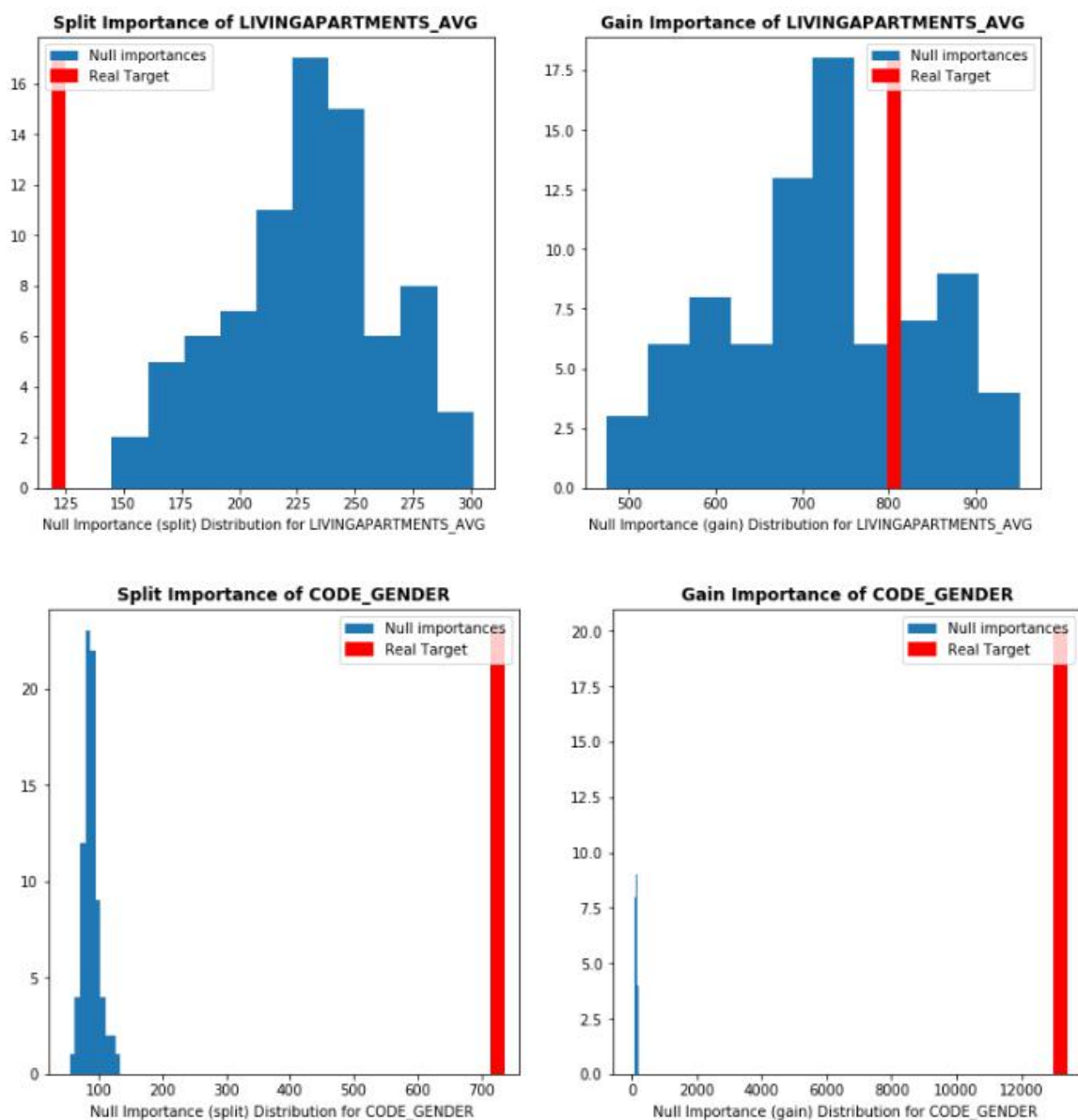
以前支付的房屋信贷信用的还款历史与 Home Credit 样本中的贷款有关。每一笔付款都有一行 a 加上 b) 每行付款一行。一行相当于一个分期付款或一个分期付款，相当于 Home Credit 样本中与贷款相关的一个先前房屋信贷信用的一次付款。

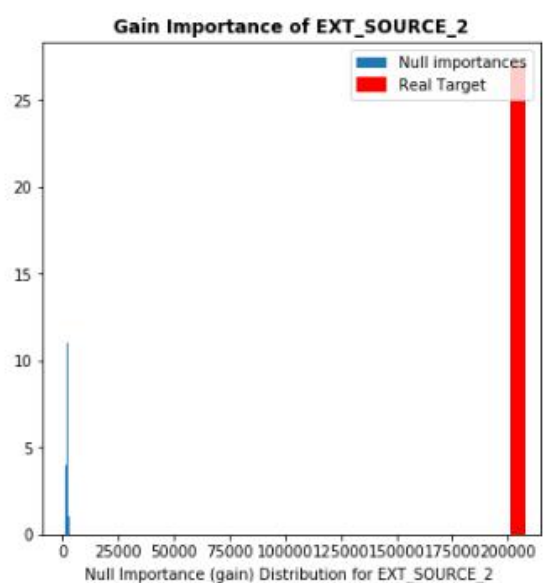
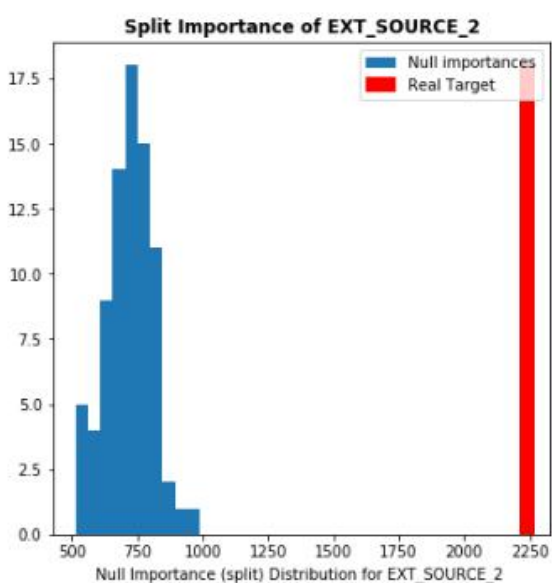
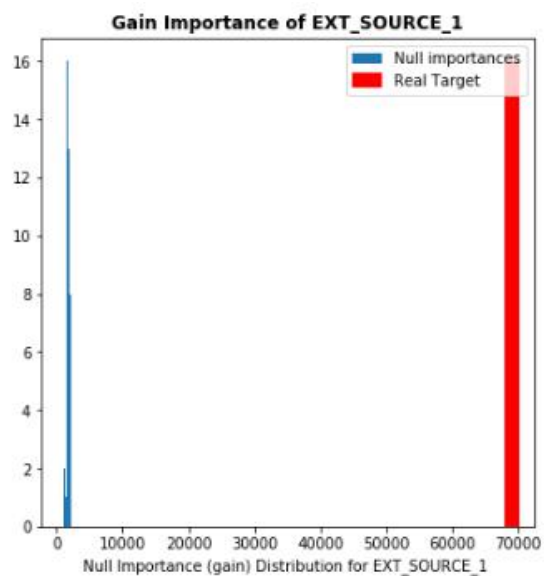
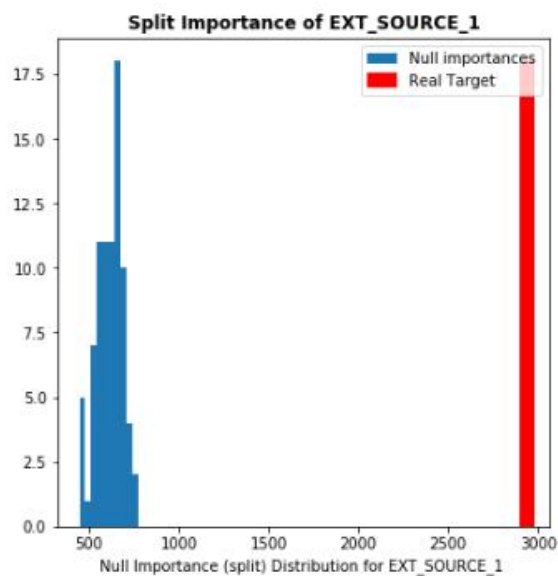
- HomeCredit_columns_description.csv

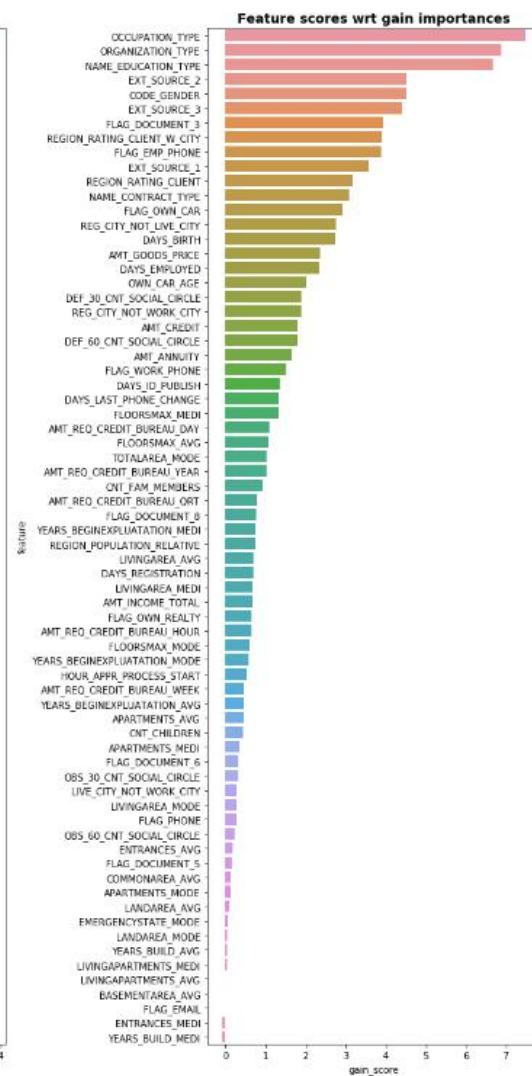
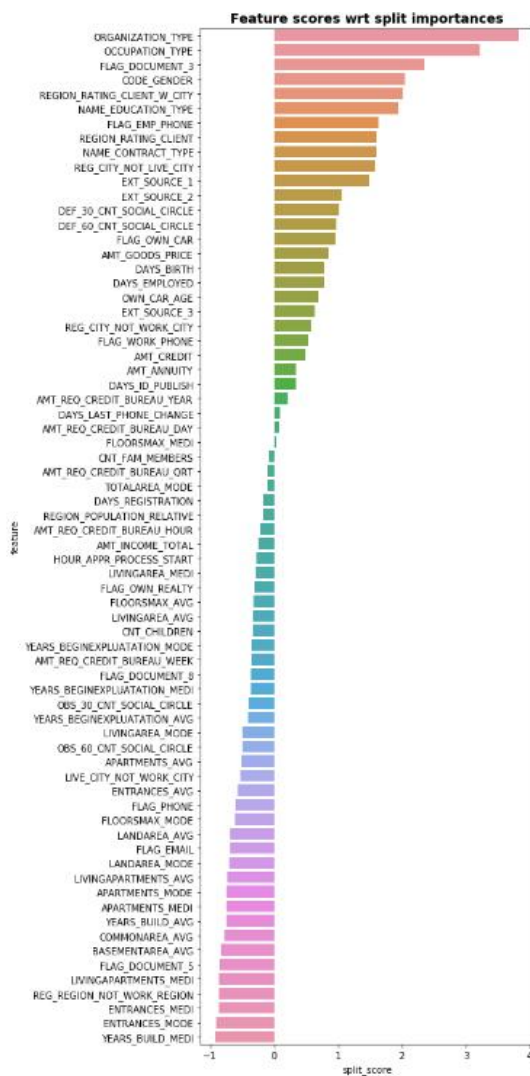
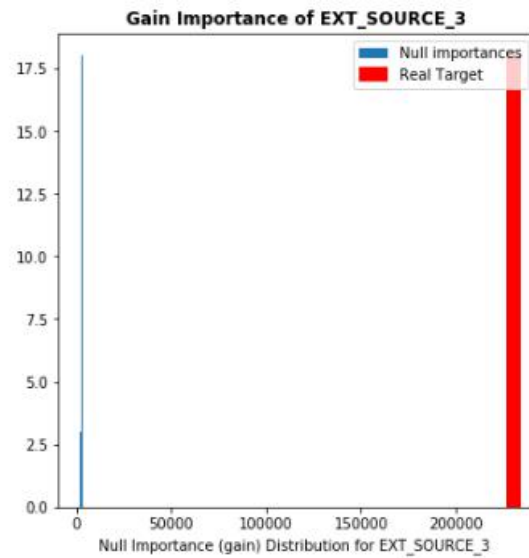
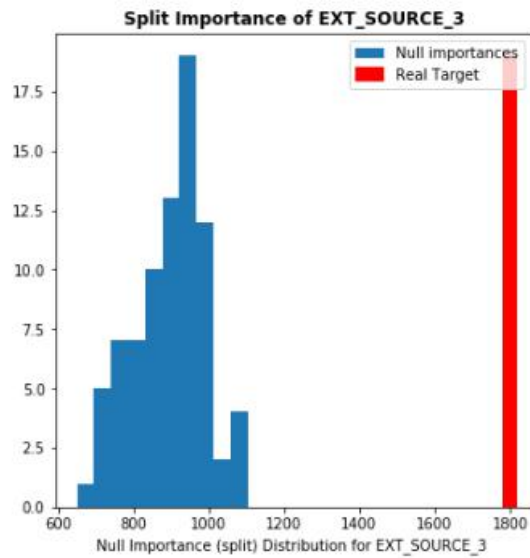
此文件包含各种数据文件中列的说明。

2.2.3 数据描述性统计

数据集中的一些关键的数据分布如下所示：







更多详细信息请参阅：<https://www.kaggle.com/ogrellier/feature-selection-with-null-importances>

3. 优秀算法思路

3.1 方案一

第一名的方案注重特征的模型的多样性，子模型用的是 XGB, LightGBM, CatBoost。本次命题应该是机器学习领域最复杂的问题之一。该领域的数据包往往是非常异构的，在不同的时间范围内收集，并且来自许多不同的来源，这些来源可能在数据收集过程中发生变化和改变。由于第一名得主的团队有很多个人，并且每个人都各有分工，所以下面的方案描述，是按照每名参赛选手所作的工作进行介绍的。

3.1.1 方案一数据预处理及特征工程部分方案

将 SKIDPREV 和 SKIDBUREAU 两个特征进行了简单的聚合，还基于 applicationtrain.csv 文件的除法和减法运算创建了许多功能。对不同的数据片段进行了聚合。对 Previousapplication.csv 文件的最后 3 个，5 个和前 2,4 个应用程序的聚合。对 Installmentpayments.csv 文件的最后 2,3,5 付款的汇总。在 DAYSINSTALMENT 过滤了过去 60,90,180 和 365 天的聚合。还对过期的所有分期付款进行了汇总。如果 DAYSEENTRYPAYMENT 和 DAYSINSTALMENT 之间的差异为正，则将逾期定义为 1，否则为 0。

对于 POSCASHbalance.csv 和 creditcardbalance.csv 文件进行了和 installmentpayments.csv 文件类似的操作，我在每个 SKIDCURR 的最后 5 个应用程序中使用了滞后特征。

- Olivier

计算年度利率（模型中得分最高的特征之一）。除此之外主要负责堆叠。

- Phil

改名参赛选手设计的最重要的资格特征，按重要性从大到小排列（通过

LGBM 模型中的增益来衡量），如下所示：

neighborstargetmean500：每行 500 个最近邻居的平均 TARGET 值，其中每个邻域由三个外部来源和信用/年金比率定义。

regionid：REGIONIDPOPULATION 字段被视为分类而非数字特征。

debtcreditratioNone：按 SKIDCURR 分组，所有信贷债务（AMTCREDITSUMDEBT）的总和超过所有信贷的总和（AMCREDITSUM）。

creditannuityratio：AMTCREDIT / AMTANNUITY

prevPRODUCTCOMBINATION：PRODUCTCOMBINATION 值来自最近的应用程序。**DAYS CREDITmean**：按 SKIDCURR 分组，来自局表的平均 CREDITDAYS 值。

creditgoodspriceratio：AMTCREDIT / AMTGOODSPRICE

lastactiveDAYS CREDIT：来自局内的活跃贷款，最近的 DAYS CREDIT 值，按 SKIDCURR 分组。

creditdownpayment：AMTGOODPRICE - AMTCREDIT
AGEINT：int（DAYS BIRTH / -365）

installmentpaymentratio1000meanmean：仅查看 DAYSINSTALLMENT > -1000 的分期付款，取 AMTPAYMENT - AMTINSTALLMENT 的平均值，首先由 SKIDPREV 分组，然后由 SKIDCURR 分组。

annuitytomaxinstallmentratio：AMTANNUITY /（installmentpayments 表中的最大分期付款，按 SKID_CURR 分组）。

- Yang

改名参赛选手对于特征处理的想法如下：一些来自开放式解决方案的最后 3,5,10 信用卡，分期付款和 pos。但我修改了时间段以包含更多的功能差异。此外，我应用加权平均值（使用时间段作为权重）来创建与年金，信用和付款相关的一些功能。我认为这些功能对于提取个人信用习惯非常有用。我认为最后一部分是有趣和有用的：我通过收入，付款和时间构建一些 KPI，这对于模型的得分有了显著的提高

- Bojan

主要工作是特征选择和缩减，下面是他对于自己工作的总结：

事实证明，对于本次比赛非常有用的事情之一是限制特征集的数量。聚合特征的各种方式通常会导致编号为数千的特征集，并且很可能很多（如果不是大多数）这些特征是冗余的，嘈杂的或两者兼而有之。我尝试了一些使用数字特征的频率编码分类特征来减少特征数量的非常简单的方法，然后使用岭回归运行非常简单的前向特征选择。我过去曾使用过这种技术进行 boosting，但这是我第一次尝试使用原始功能。令人惊讶的是，它的效果非常好。我能够将编号超过 1600 个功能的“大”原始功能集减少到仅 240 个功能。后来当 Olivier 在“基础集”中添加了更多功能时，我直接将这些功能添加到我的 240 中，结果是 287。这 287 个功能能够为我们提供 CV 为 0.7985 左右的模型，LB 分数为 0.802-0.803。当我们向团队添加更多成员时，尝试将他们的功能与我们的功能结合起来变得非常重要。Olivier 花了一些英勇的努力来辨别 Phil 的哪一个特征是对我们自己的特征的补充，以及大约 700 多个特征的组合。当 Ryan 和 Yang 加入团队时，无法重复大部分努力。我们试着粗略地看一下我们的哪些功能不同，并将它们添加到他们的。

在所有人的工作中，Yang 的特征工程是最有用的。

3.1.2 方案一模型设计、建立部分方案

①基模型

第一名的团队所有的基础模型都在 StratifiedKFold 上进行了 5 次训练。

- Olivier

我使用了 LightGBM, FastRGF 并尝试了 FFM, 但 CV 结果低于预期 (0.76 AUC)

- Bojan

我使用了 XGBoost, LightGBM, CatBoost 和简单的线性回归。我基本上只使用了一组用于 XGB 的超级参数（一个非常“标准”的参数，你可以在大多数内核中找到它，以及 LightGBMs 的三个不同的集合 - 一个你可以在内核中获得的，一个是海王星人使用的，我的大多数 XGB 模型都是在 GPU 上使用 gpu_hist 进行训练，而 LightGBM 则是在 CPU 上进行训练。CatBoost 模型并不是那么好，并且需要永远训练

，但我认为它们是对元特征的多样性有所帮助。

- Ryan 和 Yang

在他们的工程数据集上训练了几个 LightGBM 模型，以及他们的数据与其他数据集的某些组合。

- Michael Jahrer（神经网络等）

正如许多人在讨论论坛中所读到的那样，神经网络模型在 AUC 方面落后于增强树（lgbm, xgb）。当涉及到一定程度的准确性时，NN 一直是混合的热门候选者。

所有数据集的结果相似，DAE + NN 始终优于普通 NN（在 AUC 中可能+0.005），直到最后我从未尝试过单独的 NN。

DAE 表示将自动编码器预处理去噪为 NN 的输入。用于更好数据表示的无监督技术。与我在 porto seguro 中描述的技术相同。原始功能的数量在 100 到 5000 范围内，这意味着 DAE 需要非常大以避免信息丢失并保持过度完整/放松的表示。第一个 DAE 模型总是具有 10000-10000-10000 拓扑，这意味着我将功能数量扩大到 30k。受监督的 nn 总是 1000-1000 拓扑，这是我的标准推荐，也适用于此。尝试了更多的神经元，AUC 下降了。

DAE: swapnoise = 0.2, DAE 中的高启动学习（接近发散），1000 个时期。监督 nn: lRate = 2.5e-5, dropout = 0.5, 约 50 个时期（直到过度拟合），logloss 优化。所有隐藏单位都是 ReLU, Optimizer SGD, minibatchsize = 128, 小 lRateDecay。在 GTX1080Ti 上，5CV 的一次完整运行大约是 1 天，DAE 在运行时占主导地位。

原始数据规范化仍然是 rankGauss, 缺失值被替换为 0. 我将此规范化的参考代码添加到我的 github repo。

在竞赛的后期，我在 DAE 中尝试了更少+更大的隐藏层，我想这里更好。最好的 AUC 我用 DAE 只有 1 个隐藏层和 50k 神经元，然后是 1000-1000 监督 nn（CV = 0.794961 public = 0.80059 private = 0.79433）来自我们所有 6 个人的联合特征集。

除了所有的神经网络优化之外，一个带有小学习率的简单 lgbm 击败了它们。CV 上的最佳 lgbm 为 0.8039, 比 nn 高出近 0.01 的 AUC。

无论如何，我认为神经网络在这里起着次要的作用。我的猜测是，与 lgbm 相比，数据规范化仍然是他们未达到相同 AUC 水平的最大问题。

但是他们最终需要争取最后的 0.0001 助推，这就是 kaggle 的意思。

② 模型融合

产生 L1 密集矩阵并输入到第一级堆叠器：NN，XGBoost，LightGBM 和 Hill Climber 线性模型。L2 层包含 NN，ExtraTree 和 Hill Climber。L3 层是添加了一些原始特征进行了重新堆叠。

最终预测结果是这 3 个预测的等权重混合。

小组成员的经验总结如下所示：

- Michael

我们有一个神经网络也是成功的堆叠器。在这里，我使用了带有一个隐藏层 500 ReLU 单元的普通监督 nn。这里的技巧是找到正确的启动 lRate = 1e-3 和良好的 lRateDecay = 0.96（每个纪元后的 lRate 乘数）值以使平稳运行。此外，辍学=0.3 也很重要。

- 菲尔

ExtraTrees L3 模型是一个非常浅（maxdepth = 4），高度正则化（minsamplesleaf = 1000）模型，仅使用了 7 个 L2 模型加上一个原始特征 AMTINCOME_TOTAL。得分为 CV: 0.80665, LB: 80842, PB: 80565。

我们尝试过的其他事情，一些有趣的见解以及一些建议

正如上面简要提到的，作为一个团队，我们并没有在超级调整中投入太多时间。我尝试在 XGB 参数和 LightGBM 上运行 Olivier 准备的一个优化脚本，但局部结果有点令人沮丧。到那时，我们在很大程度上依赖于我们的服务，所以在我们的脑海中，有几个不同的模型在不同的超参数上训练，这些模型不具备单一的，高度优化的模型。

我们尝试的一件事是开发一个预测模型来估算 EXT_* 特征。估算特征的 AUC 为 0.78 倍，但它们对 CV 或公共 LB 的基础模型没有帮助。根据我们的分析，我们认为可能是因为这些特征是列车和测试集之间最一致的特征，因此对公共 LB 对本地简历的“提升”没有多大贡献。

第一名团队的更多经验总结和方法概括，详情请参考：
<https://www.kaggle.com/c/home-credit-default-risk/discussion/64510>

3.1.3 方案一结果、排名等

结果：得分 1514014.62

排名：最终排名 top1

详情见下表 3-1：



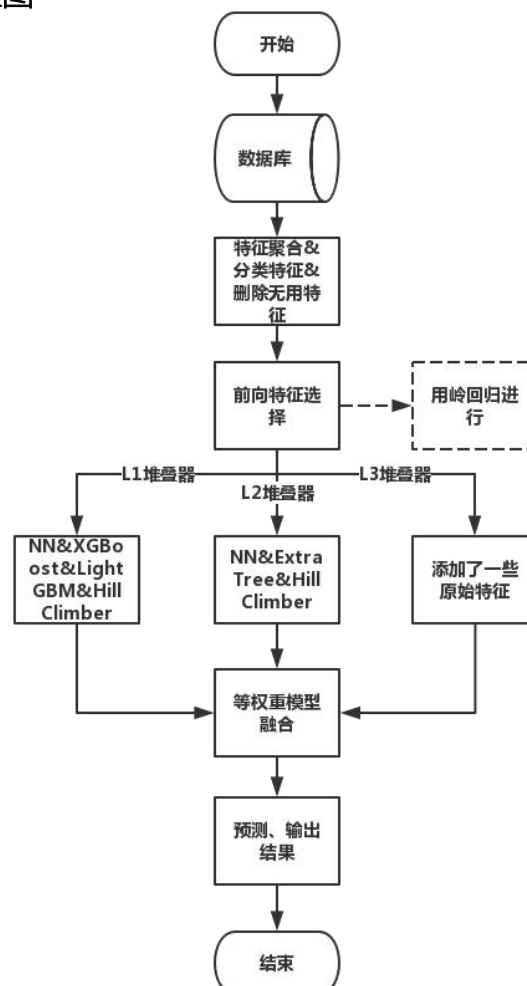
#	+pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	▲10	Home Aloan		 +3	0.80570	499	8mo
3		NighTurs		 1514014.62...	13	3mo	

表 3-1 第一名队伍比赛结果

3.1.4 方案一算法流程图



4. 算法比较

表 4-1 算法比较

评估指标数值		特征工程	基础算法	基本库	语言
算法一	1514014.62	特征聚合 特征分类 特征选择	NN、XGBoost、 LightGBM、Hill Climber、 ExtraTree 等	Numpy、pandas 等	Python

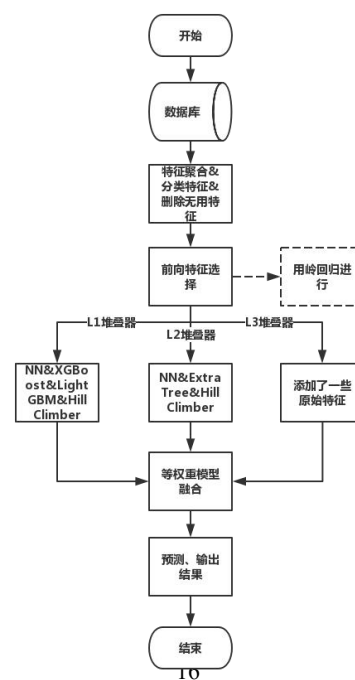
5. 总结与展望

5.1 总结

本次命题是机器学习领域最复杂的问题之一。整个建模过程比较复杂，所以参赛队伍的成员数量都较多，该领域的数据包往往是非常异构的，在不同的时间范围内收集，并且来自许多不同的来源，这些来源可能在数据收集过程中发生变化和改变。因此非常考验参赛者处理特征的能力。

5.2 建模思路

方案一参赛选手解题思路的流程图如下所示：



通过本次比赛的方案整理，再次认识到了数据处理的难度所在，尤其是针对较为复杂的数据集，可以学习的东西还很多。