机器学习

桑坦德银行客户满意度

主　研　人：杜思君
参　研　人：
审　核　人：
方　　向：工业算法案例研究
版　本　号：A

大数据及人工智能部®

中冶赛迪信息技术有限公司

二〇一九年四月

版本更新

日期	版本	更新描述	作者
2019/04/15	A	初稿	杜思君

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）
2019/4/15	Santander Customer Satisfaction	预测银行客户对交易体验的感觉（好/坏）根据上百维未知的特征，来预测客户的体验，筛选特征的重要程度。	回归

目录

1. 背景描述

从一线支持团队做到最高管理层，客户满意度是衡量成功的关键因素。不满意的顾客几乎不会留下来。更为重要的是，不满意的顾客在离开之前很少表达他们的不满。

桑坦德银行希望参赛者们能通过对他们和客户的关系早期进行分析，帮助他们识别不满意的客户，使桑坦德银行能够及时采取积极措施，在为时已晚之前提高客户的体验和幸福感，挽留更多的客户。

1.1 竞赛赛题描述

在本次比赛中，参赛者们将使用数百种匿名功能来预测客户是否对其银行体验感到满意或不满意。

1.2 评估指标描述

竞赛中评估模型优劣的指标。

在预测类别和实际类别之间的 ROC 曲线下的面积上评估提交，即 AUC 值，其计算公式如下：

$$AUC = \frac{\sum_{i \in positiveClass} rank_i - \frac{M(1+M)}{2}}{M \times N}$$

2.1 大赛数据来源

赛题的数据来自桑坦德银行客户的真实数据。

以下是数据的超链接：

test.csv（测试集数据）

train.csv（训练集数据）

sample_submission.csv（提交样例）

2.2 数据的描述性统计

2.2.1 数据基本情况描述：

数据集中包含大量数字变量的匿名数据集。训练集中“TARGET”列是要预测的变量，对于不满意的客户，它等于 1，对满意的客户，它等于 0。任务是预测测试集中每个客户是不满意客户的概率。

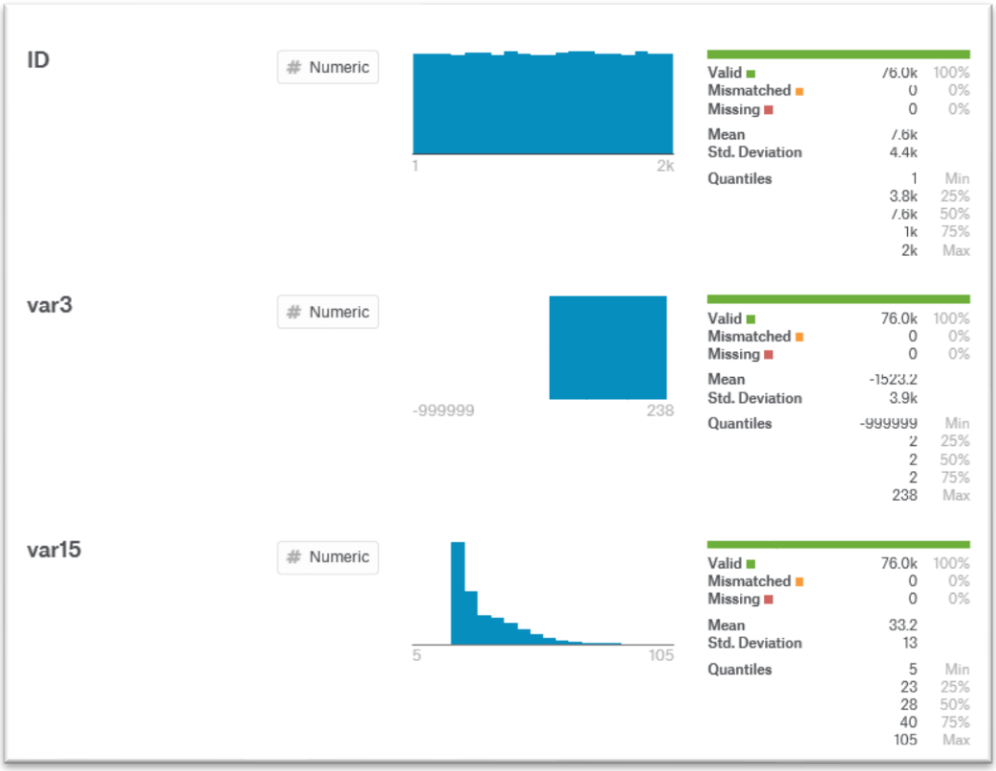
文件说明：s

- train.csv – 包括目标的训练集
- test.csv – 没有目标的测试集

➢ sample_submission.csv – 格式正确的示例提交文件

2.2.2 数据字段介绍:

由于训练集和测试集中有 370 维的匿名变量，难以分析其含义，由于篇幅有限在此也不对所有变量的数据类型和缺失率做统计。在 Kaggle 的比赛界面 https://www.kaggle.com/c/santander-customer-satisfaction/data 有此类介绍，如下图所示



2.2.3 数据描述性统计

1) 对训练集中目标值进行分析

	TARGET	Percentage
0	73012	96.043147
1	3008	3.956853

Figure 2.2.3-1： 目标变量训练集直方图统计图及占比和数量情况

从上图我们不难看出，大多数客户还是持满意态度的，不过我们更需要找出那些不满意的客户。

2） 通过数据分析确定匿名变量可能的含义

i) 经过分析 num_var4 是银行的产品数，下面是对其的一些统计分析

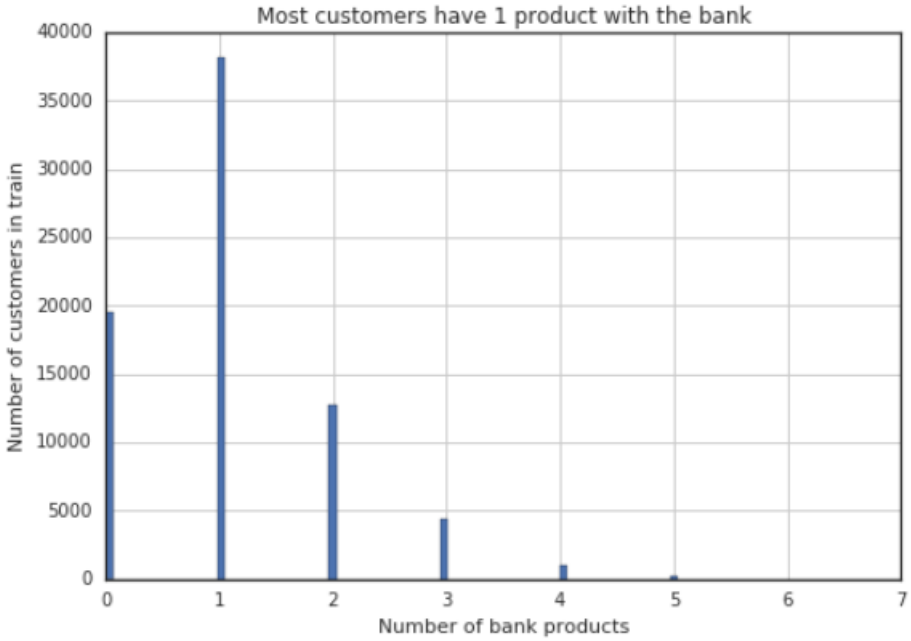


Figure 2.2.3-2: 银行的产品数与客户数量的分别关系

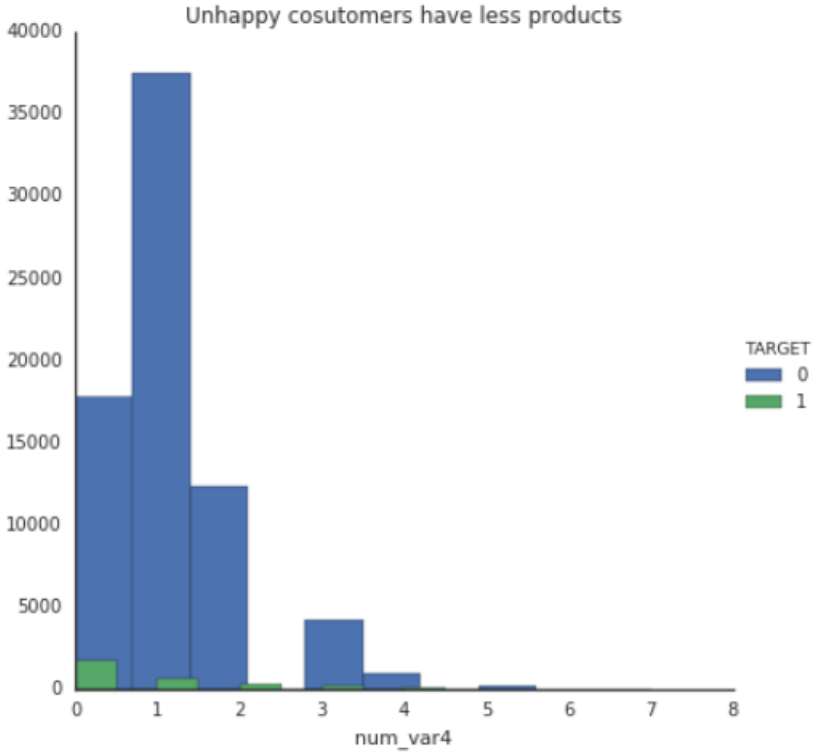


Figure 2.2.3-2: 产品数与用户满意度的统计

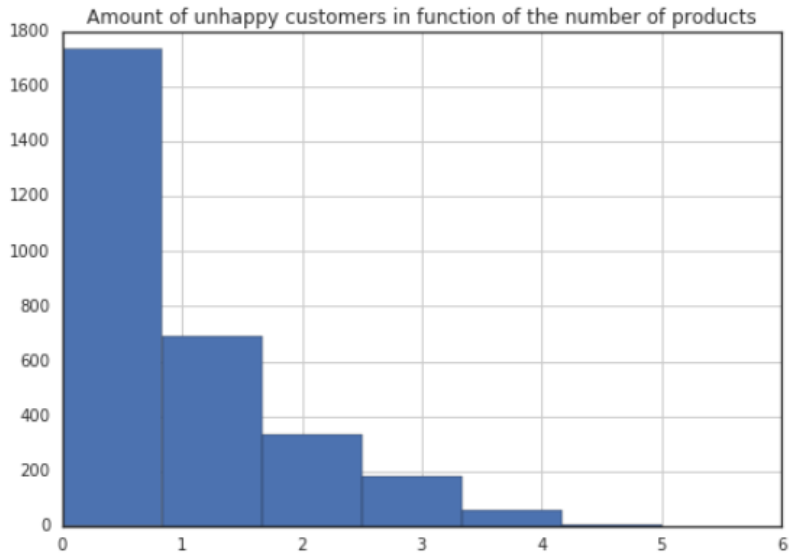


Figure 2.2.3-2: 产品数与用户不满意的分布

3） 重要变量分析

i）通过 XGBoost 得到的变量重要性：

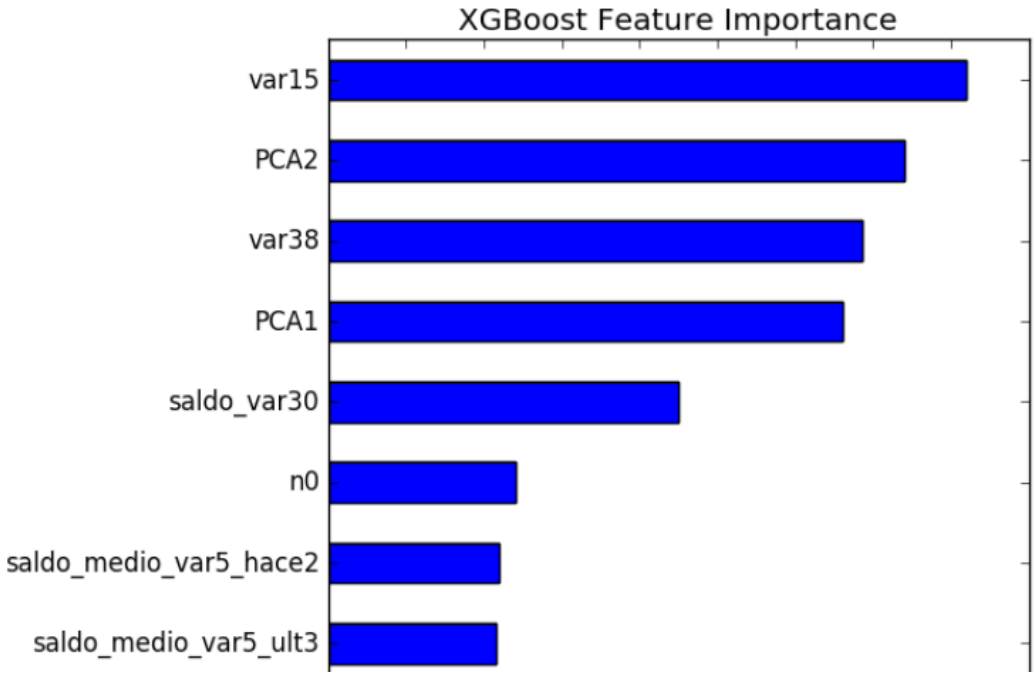


Figure 2.2.3-2: XGBoost 下变量的重要性排序（局部）

ii） 随机森林得到的变量重要性：

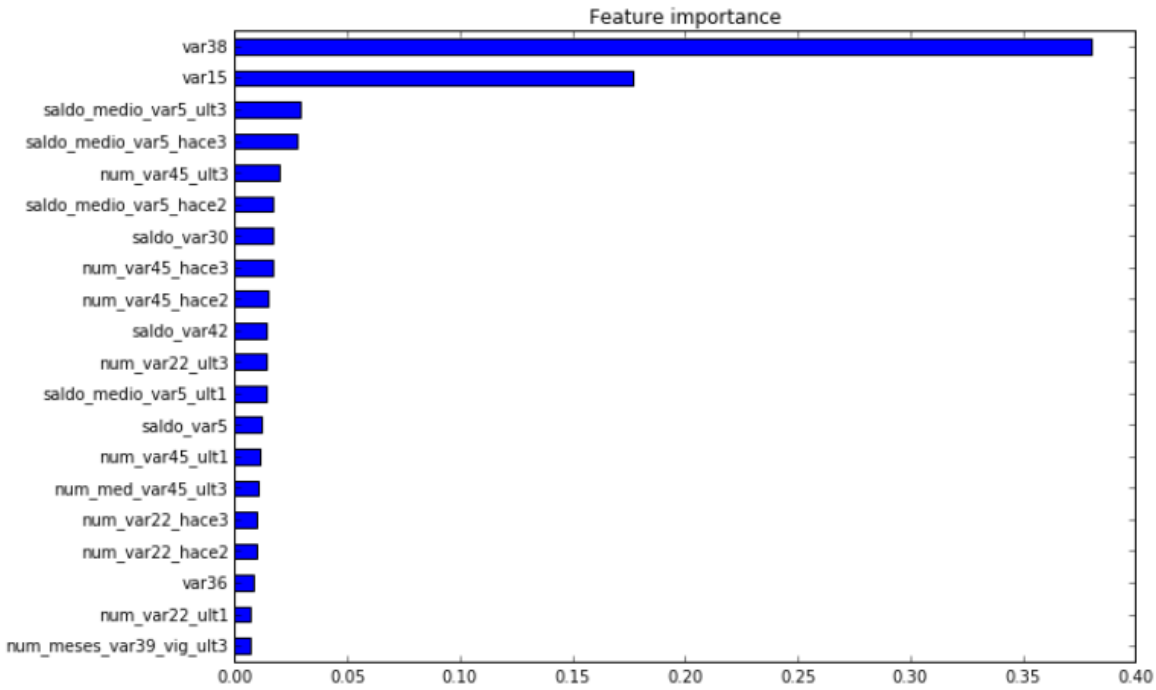


Figure 2.2.3-3:Random Forest 下的变量重要性排序图

由此我们不难看出 var38 和 var15 是两个非常重要的特征，下面我们针对这两个特征进行分析。

Var38

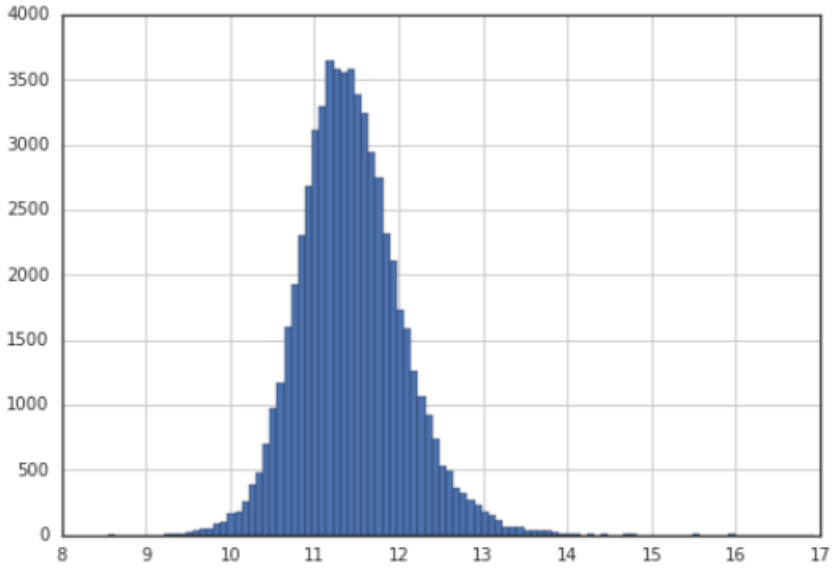


Figure 2.2.3-4:var38 的分布

Var15

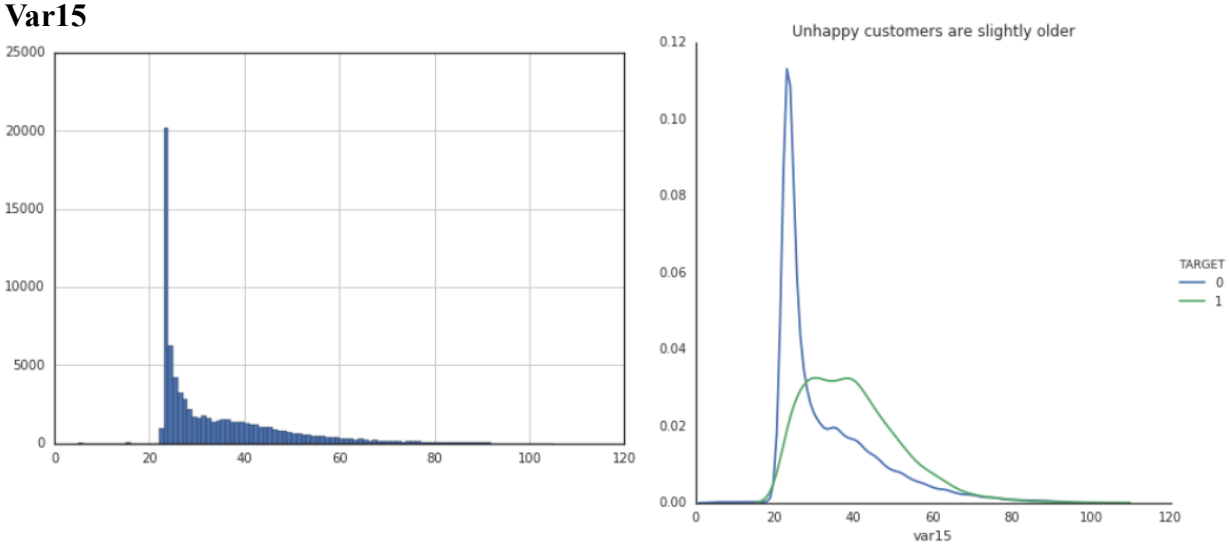


Figure 2.2.3-5:var15 的分布和对应目标值

Var38 和 Var15 的相互关系分析

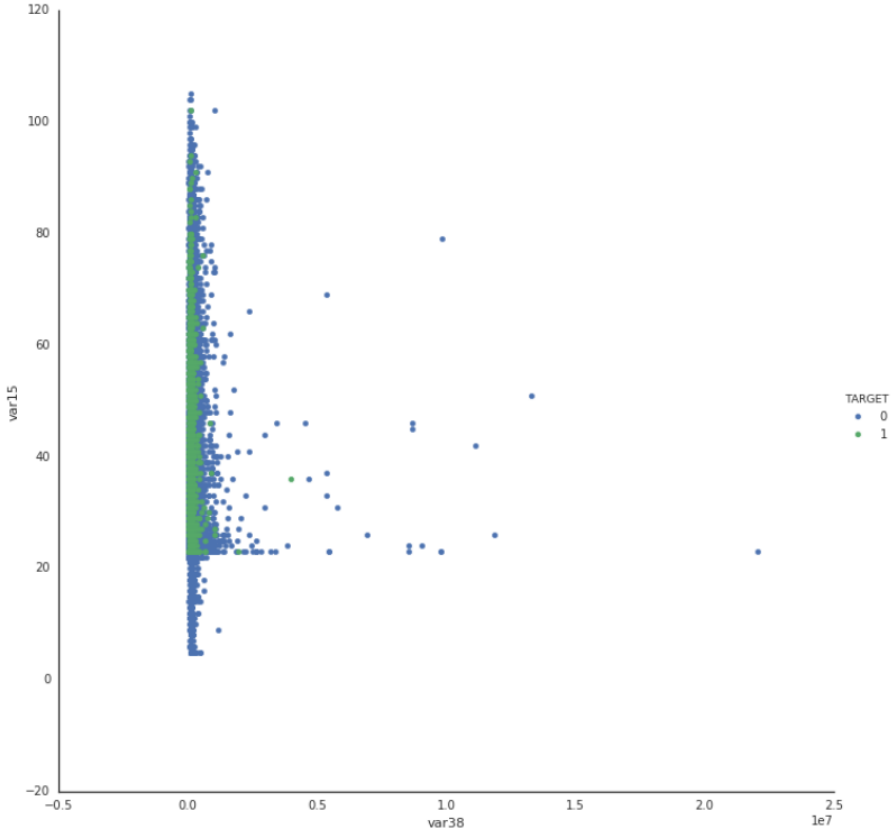


Figure 2.2.3-6:var15 和 var38 的关联和对应目标值

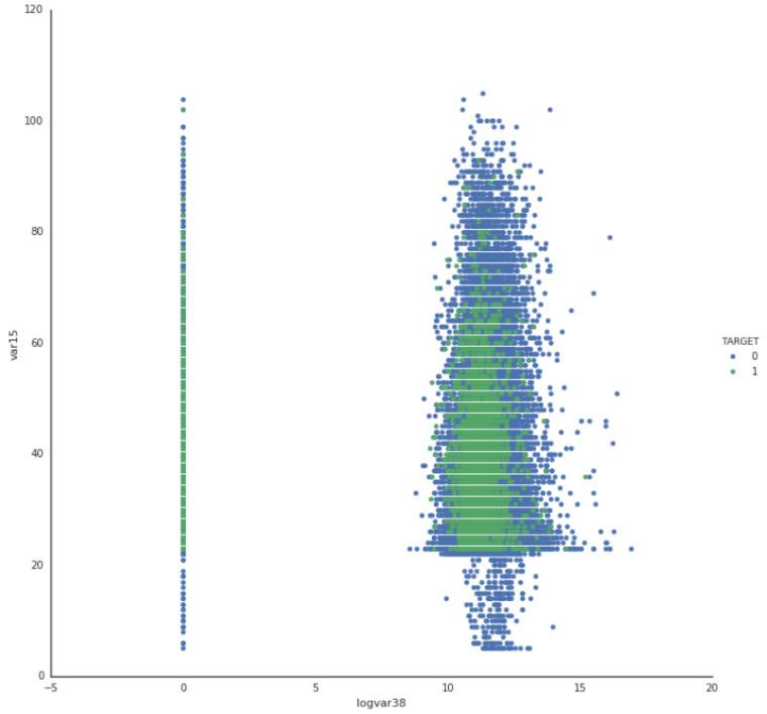


Figure 2.2.3-7:var15 和 var38 的对数分布和对应目标值

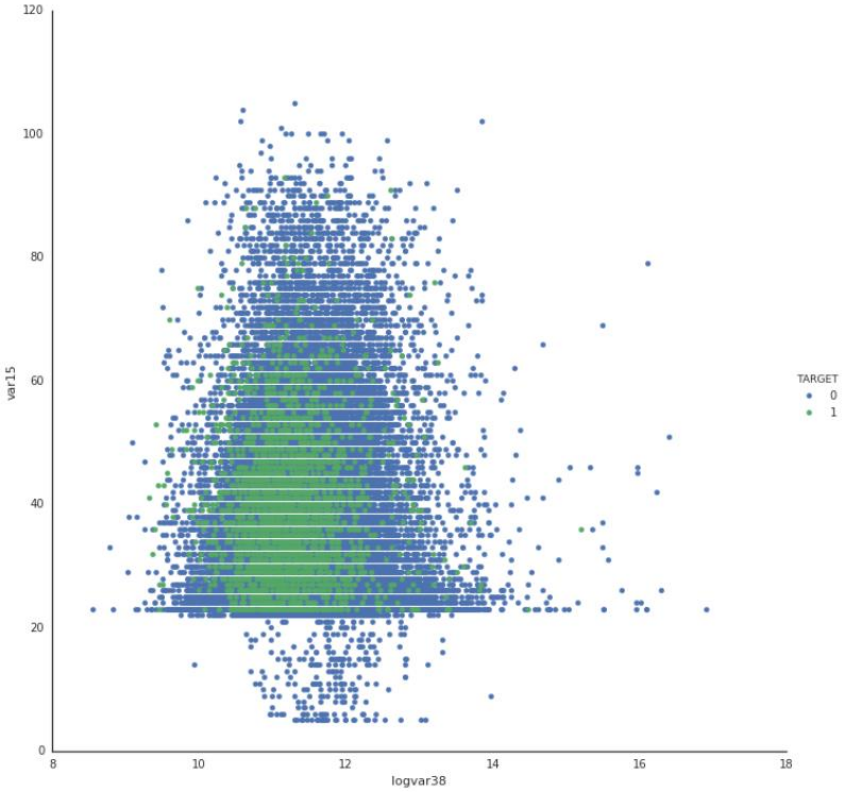


Figure 2.2.3-7:var38 和 var15 的对数分布和对应目标值

4） 部分变量与目标值的热力图

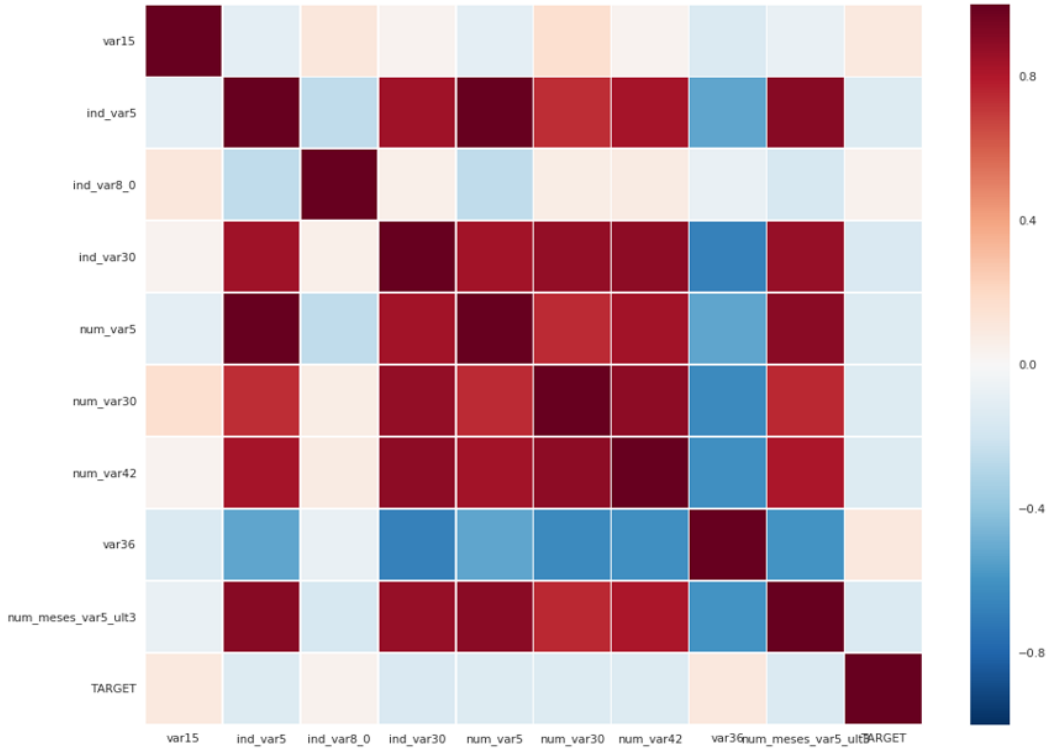


Figure 2.2.3-8:var15 和 var38 部分变量与目标值的热力图

3. 优秀算法思路

3.1 方案一

3.1.1 方案一数据预处理及特征工程部分方案

1） 数据预处理

该方案中应用了几个手动预处理步骤，例如用 NA 替换一些值，删除稀疏和重复的特征，标准化特征等。

2） 特征工程

在模型训练之前生成了以下类型的特征：

- **零和特征：**对每一个训练和测试的样例分别计算其零值数
- **t-SNE 特征：**t-Distributed 随机邻域嵌入是一种将空间映射到低维空间的降维技术。
- **PCA 特征：**主成分分析是一个降维的过程，可最大限度地减少到该子空间的距离。

- **K-Means 特征：**应用 K-Means 算法对数据进行聚类，并将分配给每个训练示例的聚类数量作为特征（算法中考虑使用 2 到 10 个聚类）。
- **似然特征：**似然特征是利用公式的 *out-offold* 预测来计算的，其公式如下：

$$LL = \frac{30 \cdot \bar{y} + \sum_{i \in G} y_i}{30 + |G|}$$

其中，G 是训练样例的索引集，其中选择以一些原始值（如特征 saldo_var13），$|G|$是 G 集合的大小，$\bar{y}$ 是训练样例输出的平均值。

3.1.2 方案一模型设计、建立部分方案

模型由以下几个算法构建的模型融合而成：

- FTRL2

这个想法依照 *The(Proximally) Regularized Leader (FTRL-Proximal)*模型的描述来实施的。

假设目标是尽量减少损失函数：

$$L = -\frac{1}{m}\sum_{i=1}^{m}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)) \qquad (1)$$

其中$\hat{y}_i$是对实例 i 的预测值，我们将设计矩阵 X 转换为稀疏二进制设计矩阵，使得一列对应于每个特征的一个值。

对每一个样例 i 的预测值被构造成 $\hat{y}_i = \sigma(w_i \cdot \chi_i)$，其中 w_i 是对每一个 i 的权重向量其大小为 n。

该算法根据之前的采样结果{1,…,i}更新了样本 i+1 的结果，其算法如下：

$$w_{i+1} = \arg\min_w \left(\sum_{r=1}^{i} g_r \cdot w + \frac{1}{2}\sum_{r=1}^{i}\tau_r||w - w_r||_2^2 + \lambda_1||w||_1\right) =$$
$$= \arg\min_w \left(w \cdot \sum_{r=1}^{i}(g_r - \tau_r w_r) + \frac{1}{2}||w||_2^2 \sum_{r=1}^{i}\tau_r + \lambda_1||w||_1 + \text{const}\right)$$

以及

$$\sum_{r=1}^{i}\tau_{rj} = \frac{\beta + \sqrt{\sum_{r=1}^{i}(g_{rj})^2}}{\alpha} + \lambda_2,\ j \in \{1, \ldots, N\}$$

其中λ_1, λ_2是正则化参数，α, β为学习率参数，$\tau_r = (\tau_{r1}, \ldots, \tau_{rN})$是对每一步 r 的学习率向量。$g_r = \left(\frac{\partial L}{\partial w_{r1}}, \ldots, \frac{\partial L}{\partial w_{rN}}\right)$是对学习的每一步 r 的对数损失的梯度向量。

在这个模型中使用了原始特征，零和特征以及似然特征。

- RGF3

这是正则化贪婪森林（*Regularized Greedy Forest*）算法。在这个模型中我使用了原始特征，零和特征，PCA 特征和似然特征。

- RGF5

在这个模型中使用了原始特征，零和特征，tSNE 特征和似然特征

- RGF6

在这个模型中使用了原始特征，零和特征，K-means 特征和似然特征

- AdaboostClassifier

  从 scikit-learn 库里调用了 AdaboostClassifier，并使用了原始特征，零和特征，PCA 特征和似然特征来训练。

- XGBoost

  XGBoost 是著名的梯度增强算法最有效的实现之一，在这个模型了使用了原始特征，零和特征，PCA 特征和似然特征。采用了 5 种不同种子的迭代套袋法对预测中的噪声进行平滑处理。

3.1.3 方案一结果、排名等

得分：*AUC* 值: 0.82853
排名：*rank*: 3/5123

3.1.4 方案一算法流程图

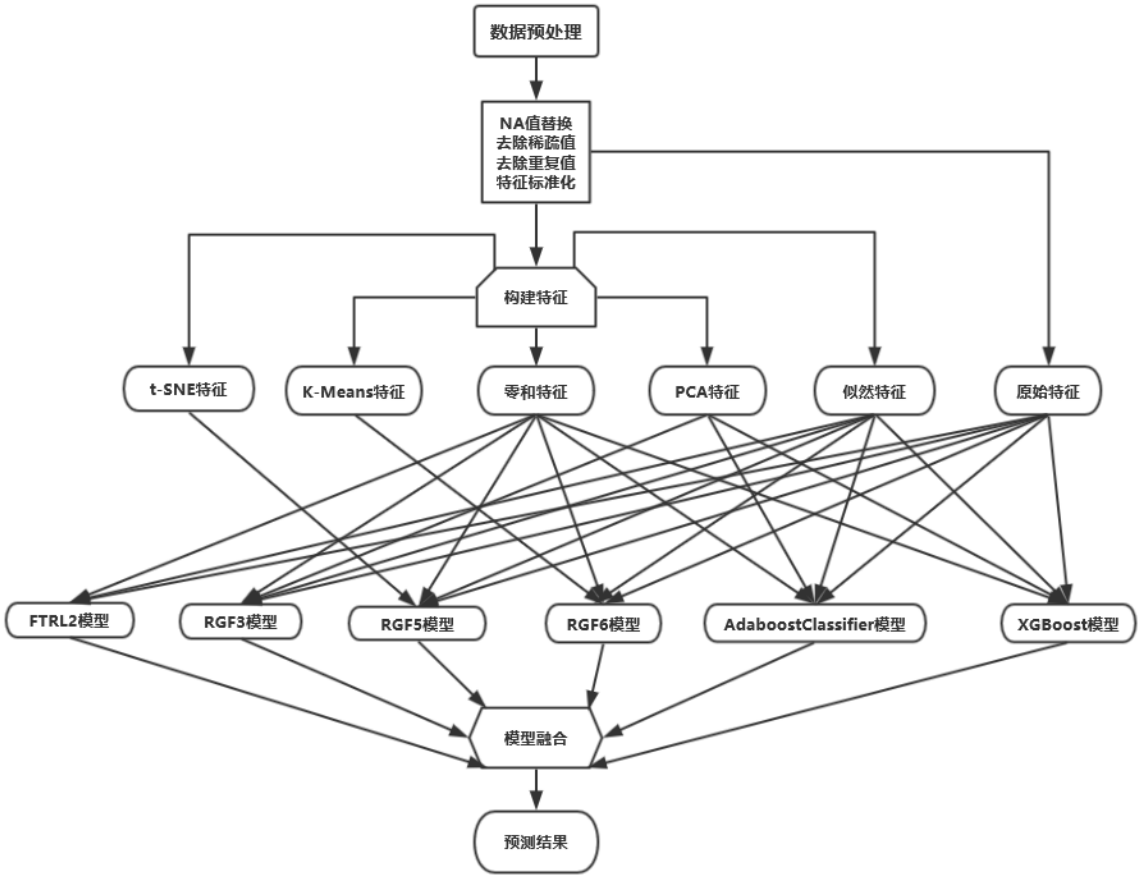


Figure3.1.4-1:算法流程图

4. 算法总结

	评估指标	特征工程	基础算法	基本库
算法 1	AUC 值	特征构造	RTRL RGF XGB 等	Sklearn Xgb

5. 总结与展望

5.1 总结

本次比赛面对大量匿名的数据，分析特征重要性和筛选、构建特征变得相当重要，并有针对性的构建模型。很多优秀的参赛者都注意到了这一点。

在本次收集中学习到了很多构建特征，强化特征方法。

5.2 建模思路

根据以前对回归问题的处理经验，首先还是选择了效果较好的 XGBoost 算法和 Random Forest 作为主要算法构建模型，在以 Lasso 和 Ridge 两种算法构建次级模型，融合之后得到最终的模型和预测结果，不过在融合时，会增加了 XGBoost 的结果权重。

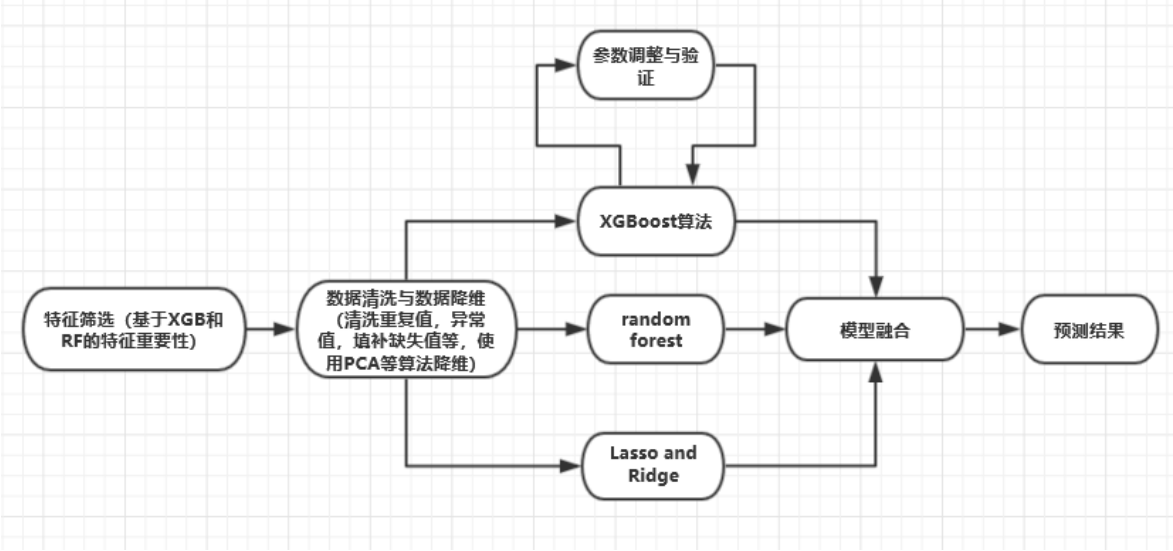


Figure5.2-1 大致的建模思路