第二届风云杯大学生机器学习建模大赛解决方案

团队名称：phm
队员：胡聪、林智敏、周青松

在第二届风云杯大学生机器学习建模中，主办方共提供了场景 A 和场景 B 两类不同的数据，其中场景 A 的训练集包含标签信息，而场景 B 的训练集不包含标签信息，且场景 B 的数据较之于场景 A，少了 ccx.csv 这个文件信息，所以这是一个迁移学习的问题，通过本次赛题所给出的业务数据我们可以看出，A、B 为相似信贷业务。A 业务数据相对丰富，为源数据。B 业务数据较少，为目标数据。数据中包含用户信息，消费信息等，经过统计，A 榜的训练集中，一共有 21245 个样本，其中正样本 6982 条，负样本有 14263 条，因为在现实场景中，违约的人毕竟是少数，所以会存在正负样本不平衡的情况。本次任务需要我们分别去预测场景 A 和场景 B 两类场景的用户违约概率，这显然是一个二分类的问题。下面我将从数据清洗、特征提取、模型训练这几个方面展开阐述。

一. 数据清洗

1）在 behavior 表里面，前 18 个字段（var1-var18）是用户的基础信息属性，后面的（var19-var2270）是用户的行为数据，通过对 var1-var2270 的缺失值进行统计，我们发现在这 2270 维属性中，有 2163 维的缺失值都在 80%以上，其中缺失率在 90%以上的有 1934 条。说明很多属性缺失都非常严重，当一个属性具有太多的缺失值时，会在模型训练的时候引入太多的噪声数据，这样是不利于模型学习到数据的分布，所以对于缺失值大于 85%的属性，我们选择直接剔除掉，这样能够减少噪声。

var22	0.969028
var23	0.971852
var24	0.975947
var25	0.979854
var26	0.980560
var27	0.981831
var28	0.985220
var29	0.996658
var30	0.996940
...	...
var2241	0.942904
var2242	0.947705
var2243	0.955943

图 2. behavior 表中字段缺失情况

2）在 consume 表里面，我们统计了几个连续字段 V_5,V_6,V_13 等字段的分布情况。从图一中我们看出，在 V_5 和 V_6 这个属性上面，红色圈出来的数据属于异常数据，远远大于其他的值，所以剔除掉训练集中红色圈出来的数据。在 V_10 这个属性上面，红色圈出来的值也远远小于其他值，故也剔除掉。

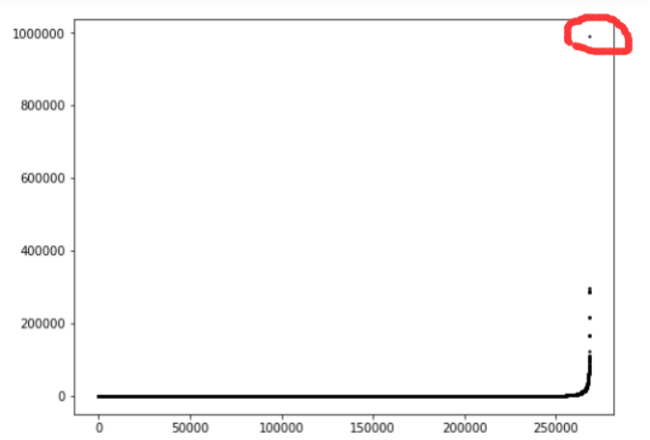

图 2. V_5 值的分布

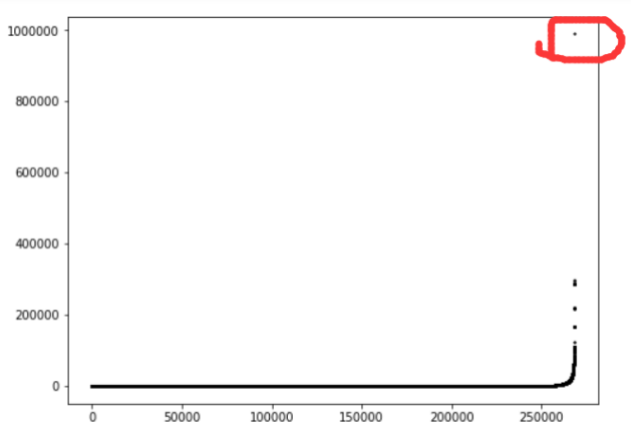

图 3. V_6 值的分布

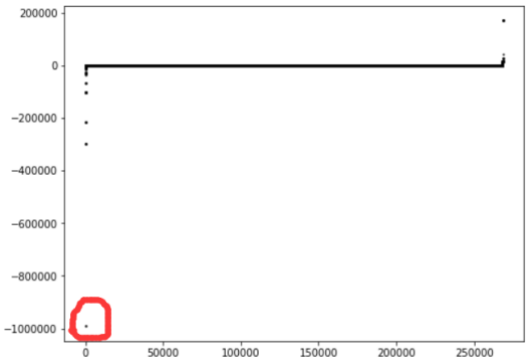

图 4. V_10 值的分布

3）观察 consume 表，我们还发现在 consume 里面有很多重复的记录，如图 5 所示，在 V_7 时间和 V_5,V_6 等属性上面，完全是一致的，因为在现

实中，在同一时间，每个人的行为只能存在一种。所以我们对 consume 表进行了去重复操作，去除掉了相同时间和相同 Id 的数据。

1	P0	C0	a0	10.0	38.00	48.00	2015-06-04 09:41:08	PL0	0.0	0.00	0000-00-00 00:00:00	1.0	38.0000	GN0
1	P0	C0	a0	10.0	38.00	48.00	2015-06-04 09:41:08	PL0	0.0	0.00	2015-06-03 23:04:21	1.0	38.0000	GN0
48996	P0	C113	a4	0.0	131.00	131.00	2014-11-29 15:16:35	PL4	0.0	0.00	2014-11-29 15:16:29	1.0	0.0000	GN27
48996	P0	C113	a4	0.0	131.00	131.00	2014-11-29 15:16:35	PL4	0.0	0.00	2014-11-29 15:16:29	1.0	131.0000	GN27

图 5.重复记录

二. 特征工程

特征工程决定了上限，一个有效的特征将极大的提高预测的准确度。由于本次比赛的数据脱敏，字段的含义未知，所以很难从业务的角度去提取特征，因此我们只能根据以往的比赛经验进行特征提取。

1. behavior 表

1） 在数据清洗的步骤，我们已经去除了缺失值大于 85%的属性，在剩余的属性当中，有很多是类别特征，也就是像 M0,MC0 这类的属性值，我们直接将前面的字母去除，保留后面的数字作为类别特征。

2） 在信用贷款领域，一个人的信息完善程度也是很重要的因素，因此我们统计了每个样本在 var1-var18 这几个基础信息属性的缺失率和 var19-var2270 这些行为属性的缺失率作为特征。

2. consume 表

1） 众数特征：与 behavior 表类似，consume 表里面也存在着很多的类似 P0,GN0 之类的属性值，处理的方法和 behavior 表一样，去除前面的字母，保留后面的数字作为特征值。每个样本有多条记录，取多条记录中类别的众数作为最后的特征值。

2） 聚合特征：统计连续属性的最大值 max、最小值 min、求和 sum、中位数 median、均值 mean、方差 std、偏度 skew。

3） 时间差特征：用 V_7 与 V_11 相减的时间，V_11 中缺失值用 1970-01-01 00:00:00 填充，对于相减后的时间，转换成秒数，继续采用聚合特征中的方法进行处理

4） 通过观察我们发现，V_12、V_13 和 V_5 变量存在着一些关系，对于很多样本，存在 V_12*V_13=V_5 这个规律，但是有些样本不满足，根据这个现象，我们提取了 V_12*V_13 是否等于 V_5 这个特征。

5） 最近消费的特征：直接将每个 ccx_id 中 V_7 最大的那一条消费记录直接当作特征。

6） 计数特征：统计每个 ccx_id 在 consume 表中出现的次数。

3. ccx 表

1） 众数特征：对于 C2、T0 这类的属性值，只保留后面的数字，然后对每个 ccx_id 的值取众数。

2） 计数特征：统计每个 ccx_id 在 ccx 表中出现的次数。

3） 时间特征：对于每个样本中的多条记录，选择 var_06 时间最大的记录与“2017-06-01”相减的秒数。

4. 交叉特征

现在很多的机器学习模型比如 XGBOOST，LightGBM 等树模型都可以在训练的过程中输出特征的重要性，以此给用户提供特征筛选的参考。交叉从理论上而言是为了引入特征之间的交互，也即为了引入非线性性，是有实际意义的。由于不知道数据字段的具体含义，所以我们通过选择特征重要性大的来进行交叉，我们的提取过程如下，具体示意图如图所示。

1） 对于得到的 293 维原始特征，采用 LightGBM 模型对 A 榜的训练集进行线下五折交叉验证，输出特征重要性。然后将特征重要性 top30 的特征进行两两相乘和相除的操作，得到 900 多维的交叉特征。

2） 将得到的 900 多维交叉特征输入到 LightGBM 模型中继续训练，同样得到这 900 多维交叉特征的重要性，取特征重要性 top100 的交叉特征加入到原始特征中。

3） 然后将特征重要性 top200 的原始特征加上 top100 的交叉特征，同样的用 LightGBM 进行训练，得到特征重要性。取这次训练中特征重要性 top30 的特征进行两两交叉相乘相除操作。得到 900 多维二阶的交叉特征。

4） 对 900 多维的二阶交叉特征同样的用 LightGBM 模型进行训练，选出特征重要性 top100 的二阶交叉特征加入到特征集里面。

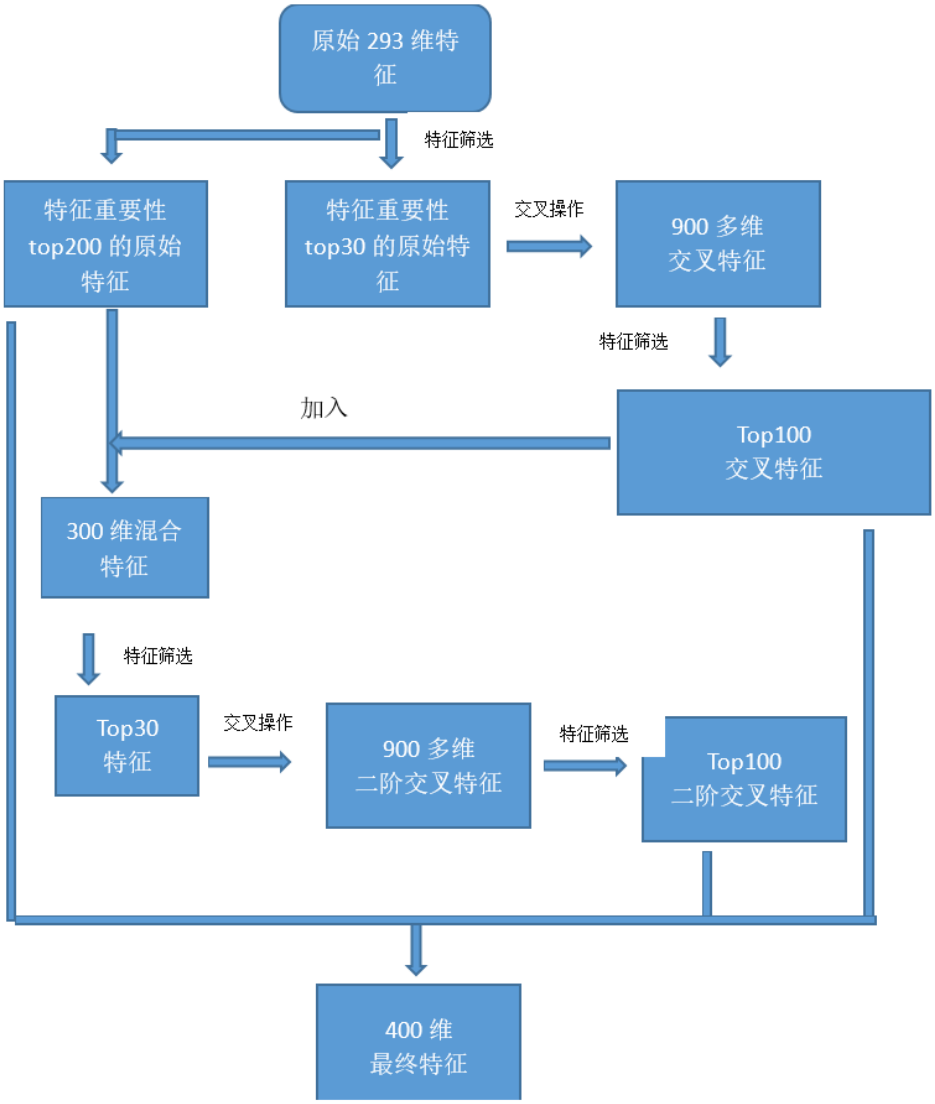

图 6.交叉特征示例

三. 场景 A 到场景 B 的迁移

场景 A 提供了 behavior 表、consume 表、ccx 表，而场景 B 只提供了了 behavior 表和 consume 表，其中场景 A 的训练集是包含标签，而场景 B 则不包含标签。场景 B 的属性是场景 A 属性的一个子集，因此我们选择用场景 A 全部的数据训练一个适用于场景 A 的模型，然后用场景 A 的 behavior 表和 consume 表去训练一个适用于场景 B 的模型，然后再分别对场景 A 和场景 B 的测试集进行预测。

我们还有考虑过在用场景 A 的数据训练适用于场景 B 的模型后，然后去预测场景 B 的训练集和测试集，采用半监督的方式，对于预测结果大于某个阈值的样本，就把它加入到训练集中，以此达到扩充训练集的目的，但是由于本次比赛提交次数有限，因此没有实现该想法。

四. 模型训练

在竞赛圈一般都是使用树模型，尤其以梯度提升树为典范，主流使用的是 XGBoost、LightGBM 和 CatBoost。由于 LightGBM 训练速度快，支持类别特征，且准确率也高，所，并且本次比赛给参赛者的运行时间只有 30 分钟，所以选择了 LightGBM 来进行模型训练。

在模型评估方面，线下我们采用的是 5 折交叉验证的方法，将训练集按照固定的正负样本比平均划分为五份，每份的数据量都和预测集的数据量接近，保证验证集合预测集有着相近的数据集分布，使评估结果更准确。

1. 单模型

本次比赛中，我们首先使用了 LightGBM 单模型，线上成绩 A 榜 0.64199，B 榜 61762。

2. 模型融合

在众多的数据挖掘比赛中，融合一直是很重要的提升手段，它能够防止单模型的过拟合，多数情况表现比单模型更好。根据以往的参赛经验，对于 A 榜，我们选择训练 4 个 LightGBM 模型，每个模型的参数都在最优参数附近，最后用 4 个模型去预测测试集，对得到的 4 个结果进行取平均操作，得到最后 A 榜的提交结果。

五. 创新点

1. 统计了用户缺失值个数，得到用户信息的完整度
2. 采用了交叉特征，引入了非线性的组合，提升了模型的效果