基于拒绝推断的信用评分模型

吕海燕,郭欣冉
(中国人民大学,北京 100000)

摘 要:随着中国经济的发展,金融科技得到迅速发展。不同于西方国家,我国的征信体系并不完善,互联网金融更多是利用用户授权的信息来构建模型,从而推测用户是否具有借贷资格。因此信用评分模型具有非常重要的实用价值。由于建立信用评分模型的样本来源于放款数样本被拒绝的用户并没有被纳入到评分卡建模的数据集中,导致样本数据是非随机性的,建模数据存在偏差。为了减少这种偏差,使用拒绝推断来推测拒绝用户的贷后行为。本文详细介绍了拒绝推断的原因、适用场景、方法等。同时,利用统计学、机器学习的相关知识构建了基于拒绝推断的信用评分模型。

关键词:贷款;拒绝推断;信用评分模型
【DOI】10.12231/j.issn.1000-8772.2020.30.298

1 拒绝推断的背景

幸存者偏差(Survivorship Bias)是一个广泛存在的人类认知思考的逻辑谬误。我们在进行申请评分卡建模时,使用的是有贷后表现的样本,而样本具有贷后表现就要求其申请的借款被通过,且还款行为被观测。这种情况下就是用通过的样本来代替申请的样本,对于申请评分卡的建模样本而言并非是随机的也不是全面,即对总体样本的描述是有误的,得到的结论也会出现错误。

在申请评分卡建模中,由于无法获得未通用样本的贷后表现只使用了申请通过的用户信息,出现样本偏差(Sample Bias),从而导致模型参数估计误差较大,影响决策者对实际风险的判断。为了解决这一问题,在申请评分卡建模时,将拒绝的样本与通过的样本混合一起建模,就需要推断拒绝样本的贷后表现,从而出现了拒绝推断(Reject Inference)。

2 拒绝推断的原因

在整个信贷全链路中,典型的贷前风控流程一般为:客户申请->业务准入->反欺诈->信用评估->客户审批->信用政策。通过一层层的筛选过滤,相对优质的客户审核通过完成放贷;相对劣质的客户则被拒绝未得到放款。因此,全部的申请客户通过筛选之后,实际通过客户的占比一般不会高于10%,有贷后表现的客户占比甚至更低。

建立新模型或者优化已有模型都需要根据客户贷后表现积累的业务数据进行好坏标签的确定,这就只能使用具有贷后表现的样本。而贷前风控流程决定了具有贷后表现的样本只能是通过的客户,被拒的客户不在其中。实际上,通过的客户和被拒绝的客户群体在好坏分布上的表现是不一样的,如果仅使用通过并且有贷后表现的样本进行建模,那么新建模型基于的样本分布将不同于审核阶段的样本分布。当使用通过并且有贷后表现的样本构建的模型部署后,模型实际上是对全部申请的样本进行预测,这样就会出现估计偏差。

当使用通过样本建立的申请评分卡模型应用到信贷审核环节,一般会产生一定的偏差。对于不同的产品业务线,偏差造成的影响可能较大也可能较小。在大多数情况下,申请评分卡都是信贷风控流程其中的一部分,因此我们必须要对申请评分卡模型可能出现的偏差进行分析。一般情况下,根据全部申请的样本建立的模型和根据通过的样本建立的模型在违约率估计上会存在区别,尤其是在高风险端,审核通过样本建立的模型会极大的低估这部分客群的风险。低估高风险端的风险,会影响所有业务的收益,造成不必要的损失。为了解决这一问题,就引出了拒绝推断,即假设被拒绝的客户被放贷,推断该客户的贷后表现是好还是坏,进而将推断的好坏样本加入到建模样本中,从而减少模型建立和使用过程中的差异。

3 拒绝推断适用场景

当全部的申请样本通过率极高时,例如白名单客户,则没有必要做拒绝推断。因为在通过率高的情况下,通过的样本与全部申请的样本偏差不大,可以使用通过的样本代替全部申请的样本。

当全部的申请样本通过率不高时,一般要考虑新建申请评分卡模型存在偏差的大小。如果新的申请评分卡模型在实际应用中偏差较小,那么新的申请评分卡模型对被拒样本的风险预测会比较合理,例如,拒绝样本的风险预测程度会比所有样本或者通过样本的风险预测程度高,比如拒绝样本的风险预测程度是所有样本的三倍或者更多。如果达到这一水平,那么模型使用拒绝推断的意义不大;如果达不到这一水平,那么模型就需要使用拒绝推断,以减少建模阶段和应用阶段的偏差。

4 拒绝推断的方法

讨论了拒绝推断的原因和使用场景后,以下介绍在实际生产环境中,几种常用的拒绝推断的方法。

4.1 小批量测试

在一定时间内,准入所有用户,待用户具有确定的贷后表现行为时,用这部分样本数据直接建模。这是最准确的方法,拒绝客群的好坏标签不再需要推断,而是直接得到用户最真实的贷后表现。但是此方法需要承受较大的坏账损失。

4.2 外部状态赋值

通过第三方数据源,为拒绝用户群体打上是否会违约的0-1标签。例如通过数据发现用户在其他金融机构有类似的违约记录,那么这样的用户群体可以直接标记为1。

值得注意的是,其他机构的业务必须要与拒绝推断样本的产品相同或者相似。因为信用评分模型是衡量用户的还款意愿和还款能力。用户在大额产品上还款困难,并不意味着在小额产品上也没有还款能力。但是,该方法需要花费一定的数据成本,并且有些外部数据较难获取。

4.3 模型推断法

表 1 拒绝推断过程示意表

分组	总人数	拒绝人数	放款人数	逾期人数	通过率	逾期率	推断逾期率
1	2478	2003	475	119	19%	25.00%	75.00%
2	2407	2006	401	66	17%	16.45%	49.35%
3	2489	1856	633	89	25%	14.00%	42.00%
4	2456	1663	793	64	32%	8.11%	20.28%
5	2465	1573	892	45	36%	5.02%	12.55%
6	2400	1453	947	34	39%	3.56%	7.12%
7	2420	1452	968	27	40%	2.78%	5.56%
8	2502	1502	1000	20	40%	2.01%	4.02%
9	2393	1403	990	11	41%	1.12%	2.24%
10	2396	1452	944	3	39%	0.31%	0.62%
总计	24406	16363	8043	477	33%	5.93%	25%

用通过样本训练模型，应用在拒绝样本上，推测出被婉拒的客户的好/坏标签，后用这两部分样本重新训练模型。这种方法有一个要求，就是用通过样本开发的模型是一个具有一定区分度的模型。如果模型的区分度不够高，推测的好/坏标签的可靠性大大降低。如何使用通过样本建立的模型，去推断拒绝样本的好/坏，常用的有硬截断法、分组扩展法、加权扩展法。

4.3.1 硬截断法

硬截断法，同时也称为简单扩展法。是用准入用户群体的各维度信息建立相关模型，从而为拒绝用户打分，分值记为 P(bad 的概率)。通过数据分析和一定业务经验，设定一个固定阈值 P_base，小于 P_base 的用户群体认为是好用户，大于 P_base 的用户群体认为是坏用户。选择的阈值 P_base 要使拒绝用户群体的违约率(bad 的概率)比接受用户群体的违约率高，一般实践经验设置：拒绝用户群体违约率(bad 的概率)是接受用户群体违约率的 2-5 倍[4]。

硬截断法通俗易懂，简单明了，方便实施。但是阈值的设置，依赖风控建模师的经验和选择，风控建模师要结合当前业务现状、客群特征以及风控准入率等因素，去设定一个合理的 P_base。

4.3.2 分组扩展法

分组扩展是对简单扩展的进一步改进。用准入样本建立的模型为拒绝样本打分，这时准入客群与拒绝客群都有一个模型分数。把准入客户的分值由低到高排列，之后分成 10 组。拒绝客户也按此分值分组。以每一个分组的通过率和准入客户的坏账率，推断拒绝用户。每一组拒绝用户群体坏账率是准入用户群体坏账率的 2-5 倍，在实际业务中扩大的倍数，要结合相关数据分析、相关业务经验、每一组风控准入率等因素综合考虑。例如：假设拒绝用户群体的坏账率为相同分数分组接受用户群体坏账率 3 倍。以某一分数段为例，假定准入用户群体的坏账率为 2%，设置拒绝用户群体的坏账率为 6%，之后按照此坏账率，将此分数段内的用户群体随机的设置为好用户和坏用户。也可以依照分数高低排序，前 top n%的用户设定为好，其他设定为坏用户，如同简单扩展法的扩展推断方法。

4.3.3 加权扩展法

加权扩展法：这一方法不仅仅是将一个拒绝样本简单直接地赋值为 1 或者 0，而是将一个样本依照违约概率拆分成一个赋有权重 1-P 的好样本(p 为违约概率)，以及一个赋有权重 p 的坏样本。而每个准入的样本，由于贷后表现确定，我们将其权重设定为 1。

例如：在 300-400 的分数之间，推断拒绝用户群体的违约率为 33%，当一个拒绝用户分值为 350 分时，我们把这个样本拆分成一个权重为 0.33 坏样本和一个权重为 0.67 好样本(即我们认为，这个用户有 0.33 坏的概率以及 0.67 好的概率。而不是将其绝对直接的赋值为 1 或 0)。

5 实证分析

5.1 拒绝推断

用通过样本训练模型，并把此模型应用在拒绝样本上，给拒绝样本打分。最后，所有的申请样本都会有一个分数，这个分数即为推测模型分数。根据推测模型分数，推测被婉拒客户的好/坏标签。一般，将推测模型分数从小到大排序，并分成 10 组。准入客户和拒绝客户均按此分组，以每一个分组的通过率和准入客户的逾期率，推断拒绝用户逾期率，一般设定每一分组拒绝用户逾期是准入客户的 2-5 倍。

如表 1 所示，1-3 组为低分数组，通过率为 17%-25%，结合业务经验、客群特征、分组通过率等因素综合考虑，设定拒绝客群的逾期率为通过客群的 3 倍。4-5 组为中间分数段客群，设定拒绝客群的逾期率为通过客群的 2.5 倍。6-10 组的通过率为 39%-41%，设定拒绝客群的逾期率为通过客群的 2 倍。

拒绝客群每一分组的逾期率推测完成后，需要推测每个用户的好坏标识，在这里选用加权扩展法，不是将拒绝样本直接标记为好或坏，而是根据推测的逾期率将一个拒绝样本拆分成一个好样本和一个坏样本，并给两个样本赋于权重。例如，第 3 组，推断拒绝客群的逾期率为 42%，当一个拒绝用户分值落在第三组的区间时，我们把这个样本拆分成一个坏样本和一个好样本，坏样本的权重为 0.42，好样本的权重为 0.58。

5.2 预测模型

拒绝客群的好/坏标签推测完成后，需要把放款样本和拒绝样本综合起来，重新训练模型。

5.2.1 特征工程

特征工程是建模环节中最重要的一步，因为数据和变量决定了机器学习的上限，而模型和算法只是逼近这个上限。此次特征工程主要从以下几个方面展开：

表 2 混淆矩阵示例图

真实情况	预测结果		统计量
	正例	反例	
正例	TP	FN	召回率/灵敏度=TP/(TP+FN)
反例	FP	TN	特异度=TN/(FP+TN)
统计量	精确率 = TP/(TP+FP)		正确率=(TP+TN)/(TP+TN+FP+FN)

(1)数据预处理。本次分析使用的数据维度主要包含:①交易支付信息,例如用户收入、支出的现金流信息;②多头共债信息,例如用户在其他信贷机构的申请、被拒信息;③互联网行为信息,例如用户的上网时长、上网时间等信息;④电商消费行为信息,例如用户的订单数量、订单金额、订单类型等信息;⑤用户提供的其它类信息,比如年龄、籍贯、教育背景等信息。

本次数据预处理的过程主要包含不同维度的数据整合以及特征衍生。上述 5 个维度的数据存放在不同的数据表中,数据整合主要是根据各个数据表相关的主键完成不同维度数据之间的关联,最终合并在一张数据表中;特征衍生包含两部分,一部分是同一纬度数据之间的特征衍生,一部分是不同纬度数据之间的衍生。主要方法有:基于业务经验、基于时间戳处理等最终形成一张可以建模使用的宽表。特征衍生后共生成 300+变量,例如:近一年最大逾期金额、近一个月多头借贷笔数、年龄、性别等等。

(2)特征选择。当数据预处理和探索性分析(EDA)完成后,需要根据 EDA 的结果,选择有意义的特征变量进行机器学习算法和模型的训练。一般需要从两个方面分析完成特征的选择:①特征信息的有效性,比如,当某一特征的方差接近于 0 时(某一样本的性别变量全部为男性),那么这个特征在样本上的表现没有出现差异,也就达不到样本分类的目的,此时我们可以认定该特征没有有效信息;②特征与目标变量是否具有相关性:与目标变量相关性高的特征,其对目标变量的预测能力较强,可以优选选择;与目标变量相关性低的特征,其对目标变量的预测能力较差,可以考虑删除。

本次特征选择步骤如下:①根据业务经验,去掉不符合业务逻辑的特征;②删除缺失率大于 50%的特征变量;③在单变量跨时间验证时,PSI>=0.1 的变量存在一定程度的偏移,删除 PSI>=0.1 的变量;④此次单变量分析过程中,删除了 KS 在 10 以下的特征;⑤在信用建模领域中,更加注重特征的单调性、可解释性。所以,在此过程中删去了趋势波动的一些特征;⑥计算相关系数,如果特征变量之间的相关系数超过 0.8,那么在两个变量之间删除与目标变量相关系数较小的变量。

(3)降维。当特征选择完成后,就可以进行模型训练了。当特征矩阵过大,会导致计算量过大,训练模型的时间过长,因此当特征矩阵过大时,一般都需要考虑降低特征矩阵的维度,以便提高运算效率。本次特征选择完成后剩余 98 个变量,没有进行降维处理。

5.2.2 建立模型

用筛选后的特征,进行模型的拟合。本次建模使用 xgboost 算法,该算法是一种梯度提升算法,训练过程即对每一轮训练结果的残差进行拟合[2]。

5.2.3 模型评估

(1)模型评估指标。

构建机器学习模型的想法需要基于建设性的反馈原则。构建一个模型,通过模型的评价指标判断模型的优劣,当评价指标未达到预期时会对模型进行优化,直到达到预期为止。不同的评估指标可以从不同的角度评价模型的优劣。以下是模型评估指标的举例:

a.混淆矩阵(Confusion Matrix)。

混淆矩阵是一个 N*N 矩阵,其中 N 是预测的类数。混淆矩阵一些定义有:①准确率(Accuracy):分类模型中所有判断正确的结果占总观测值得比重;②精确率 (Precision):在模型预测是正例的所有结果中,模型预测对的比重;③真负率:在模型预测是负例的所有结果中,模型预测对的比重;④召回率、灵敏度(Sensitivity):在真实值是正例所有结果中,模型预测对的比重;⑤特异度(Specificity):在真实值是负例的所有结果中,模型预测对的比重。当 N=2 时,混淆矩阵可以表示为表 2 的形式。

b.AUC 曲线(AUC-ROC)。

参考表 2 混淆矩阵示例图,存在两个重要的指标:灵敏度和特异度。ROC 曲线的横坐标是(1-特异度),纵坐标是灵敏度。使用不同的阈值可以计算相应的灵敏度和特异度,收集这些点的数据就可以完成 ROC 曲线的绘制。一般 ROC 曲线主要用于对二分类问题的模型进行评估。

AUC 表示 ROC 曲线下与坐标轴围成的面积,用于衡量模型的泛化性能,对于分类模型而言就是模型分类效果的好坏。AUC 越大,表示模型分类效果越好。AUC 的取值范围在 0.5 和 1 之间。AUC>0.7 表示模型有很强区分度;AUC 在 0.6-0.7 之间,表示模型有一定区分度;AUC 在 0.5-0.6 之间,表示模型区分度较弱;AUC=0.5 表示区分度近似随机猜测[3]。

c.增益图和提升图(Gain and Lift charts)。

提升图(lift chart)和增益图(gain chart)是评估一个模型预测结果是否有效的一个指标,其作用主要是检验概率的排序;这两个值由使用模型所得到的结果和不使用模型所得到的结果对比计算而来。其中:

增益=(使用预测模型的期望响应)/(随机发送的期望响应)

提升=(使用预测模型的前 n 个用户的期望响应)/(随机发送的前 n 个用户的期望响应)

d.KS 图(Kolomogorov Smirnov chart)。

K-S 曲线的数据来源和本质与 ROC 曲线是一致的,ROC 曲线的横坐标是(1-特异度),纵坐标是灵敏度;而 K-S 曲线是将这两个值都当作是纵轴,横轴则由选定的阈值来充当。

KS 值可以表示为 TPR-FPR 绝对值的最大值,衡量的是好样本和坏样本累计部分之间的差值,用于模型区分能力的评估。好坏样

本累计的差异越大,KS 值越大,那么模型区分好坏的能力越强。

在大多数分类模型中,KS 将介于 0 和 1 之间,并且值越高,模型在区分正负例情况时越好。在实际生产环境中,KS>0.4 模型预测能力较好;KS 在 0.3-0.4 之间模型可以使用;KS 在 0.2-0.3 之间模型预测能力一般;KS<0.2 模型预测能力较差[3]。

KS 的计算步骤如下:

步骤一:按照模型预测返回的概率升序排列并分组,计算每个组内的好坏样本数。

步骤二:计算每个组内的累计好样本数占总好样本数的比率和累计坏样本数占总坏样本数的比率。

步骤三:计算每个组内累计坏账户占比与累计好账户占比差的绝对值,然后取这些绝对值的最大值,即得到 KS 值。

e.PSI。

稳定度指标(PSI)用于衡量模型上线后的应用客群与模型开发样本所对应客群的特征分布的差异,是比较常用的评估模型稳定性指标。PSI 表示的是:变量按照一定规则分组后,不同时间用户或者不同群体用户,在对应分组的分布情况,分析各分组内人数占总人数的比例是否有显著变化。一般认为稳定度指标(PSI)<=0.1,稳定性很高;0.1<PSI<0.25,稳定性一般;PSI>=0.25,认为客群有所偏移,模型稳定性差,要考虑模型是否迭代[3]。

∑(A%-E%)*LN(A%/E%)(5.2.4)

式中:E%-开发样本各个分数级占比;A%-测试样本各个分数级占比。

(2)模型评估。

①模型有效性评估。由于模型是利用一段时间内的用户群体开发,此模型是不是同样适用于其他客群,必须经过样本外数据的验证,即跨时间验证。在选取跨时间验证的模型样本时,必须关注时间的序列性,即要包含到所有时间。例如,用月中时间段的数据建模,跨时间的时间段,至少要包括月初、月末的时间段。特别需要注意的是,跨时间样本的选取也要避开一些市场产品运营策略变化的时间段,例如:市场环境、大型运营获客活动、产品变化,国家节假日以及前端相对应的风控策略变化等。

表 3 模型的有效性评估结果,从样本内验证(测试集)、样本外验证(跨时间验证集)都可以看出,模型具有一定的区分能力。

表 3 模型有效性评估结果

	GINI	AUC	KS
训练集	41.29	70.64	30.23
测试集	36.32	68.16	27.82
跨时间验证集	35.93	67.96	28.38

②模型稳定性评估。比较模型分数在建模样本与样本外时间的分布情况。计算 PSI。从表 4 模型分数 PSI 可知,PSI=0.003247,模型保持稳定。

6 对比分析

使用同时间段的放款样本构建模型,与以上使用拒绝推断的模型进行对比。发现直接使用放款样本构建模型 KS 为 19,比使用拒绝推断后模型的 KS 下降 10 个点。使用近期申请借款的跨时间样本,对比建模用户与近期申请借款用户的分布情况,PSI=0.26。可知,建模客群与模型上线后应用客群,有很大偏移。建模客群(即准入客群)不能完全代表申请借款客群。

7 结束语

在信用风险模型构建时,由于只有放款样本有贷后表现行为,被拒绝的用户没有确定的贷后表现没有加入到建模样本中,导致建模客群与实际应用的客群存在偏差。准入部分坏客户,又会给公司带来一定的经济损失,所以使用扩展法推测拒绝用户的贷后行为,具有一定实际意义。经过模型的对比分析也可以得出,使用拒绝推断后,模型的有效性与稳定性都得到一定程度的提升。另:值得思考的是,在特征维度基本一致的情况下,迭代模型时拒绝推断具有一定意义。若迭代模型的特征与准入样本使用的线上模型特征维度完全不一样,推断模型是否同样比仅仅使用准入样本建模的模型优,还需要更进一步的验证。

参考文献

[1]Refaat, Mamdouh.Credit Risk Scorecard: Development and Implementation Using SAS. Lulu. com,2011.

[2]Chen,Tianqi,et al."Xgboost: extreme gradient boosting."R package version 0.4-2(2015):1-4.

[3]单良,茆小林.互联网金融时代消费信贷评分建模与应用[M].北京:电子工业出版社,2015,3.

表 4 模型分数 PSI

group	E%	A%	A%-E%	LN(A%/E%)	(A%-E%)*LN(A%/E%)
1	10.00%	8.98%	-0.0102	-0.10759	0.001097
2	10.00%	9.49%	-0.0051	-0.05235	0.000267
3	10.00%	9.45%	-0.0055	-0.05657	0.000311
4	10.00%	9.96%	-0.0004	-0.00401	0.000002
5	10.00%	10.26%	0.0026	0.025668	0.000067
6	10.00%	9.79%	-0.0021	-0.02122	0.000045
7	10.00%	10.34%	0.0034	0.033435	0.000114
8	10.00%	10.44%	0.0044	0.043059	0.000189
9	10.00%	10.22%	0.0022	0.021761	0.000048
10	10.00%	11.08%	0.0108	0.102557	0.001108
总计	100.00%	100.00%			0.003247