1

行为评分卡模型的开发

目录

行为评分卡模型的基本概念

行为评分卡的基本概念

基本定义

根据贷款人放贷后的表现行为，预测未来逾期／违约风险概率的模型。

使用场景

和申请评分卡不同，行为评分卡用在贷款发放之后、到期之前的时间段，即“贷中”环节。

使用目的

监控贷款人在贷款结束之前的逾期／违约风险

行为评分卡模型的基本概念

- 适用的信贷产品
  - 还款周期长的信贷产品
    - 太短的还款周期难以构造有效的特征
    - 长周期的产品有房贷、车贷、某些信用现金贷
  - 循环授信类的信贷产品

信用卡、某些信用贷

注：

不宜用在按月还利息、本金一次付清的产品；

每一期的风险不同。最后一期的风险远高于之前的账单期

行为评分卡模型的基本概念

- 表现期与观察期

行为评分卡预测的是条件概率：

$$P(\frac{\text{未来一段时间内发生违约}}{\text{当前没有违约}})$$

两个时间段

- 当前以及过去一段时间
- 未来一段时间

行为评分卡模型的基本概念

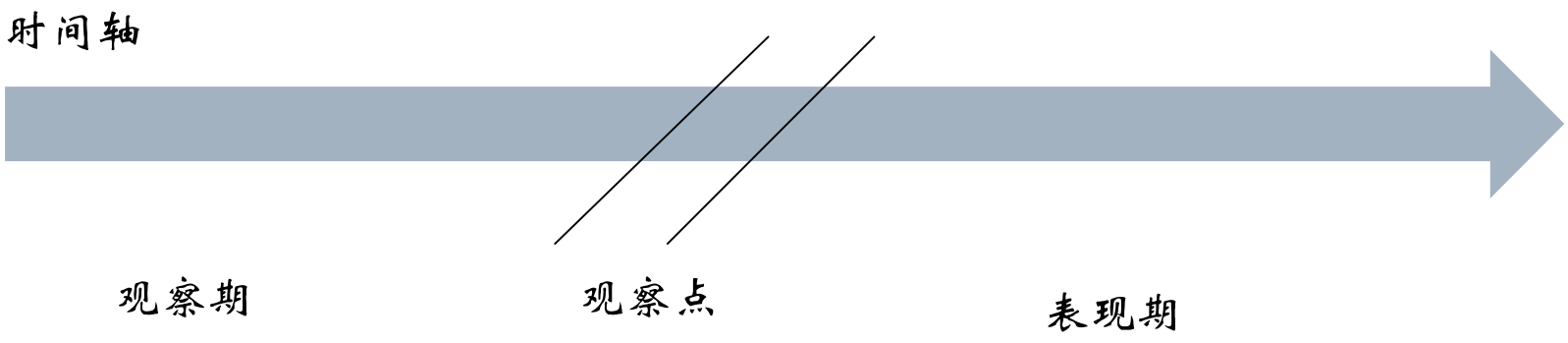
观察期与表现期

观察期

- 搜集变量、特征的时间窗口，通常3年以内
- 带时间切片的变量

表现期

- 搜集是否出发坏样本定义的时间窗口，通常6个月～1年



行为评分卡模型的基本概念

- 观察期与表现期(续)
  - 观察点的设定
    - Month on Book(MoB) = 观察点 – 贷款发放日
    - MoB不宜太短，否则无法构建出合适的行为变量，建议在6个月以上
  - 表现期的设定

不宜太短，否则

- 失去预测的意义
- 时间长度无法保证
- 概率难以预测

行为评分卡模型的基本概念

- 观察期与表现期(续)
  - 观察期的设定
    - 不宜太长，否则MoB过长，大量客户无法进入模型
    - 不宜太短，否则构建的变量，有效性不够

目录

行为评分卡的特征构造

时间切片

定义：两个时刻间的跨度

例：　观察日期之前30天内信用卡帐户的总消费额

基于时间切片的衍生

- 观察日期之前180天内，平均每月(30天)的逾期次数

常用的时间切片

- (1、2个)月，(1、2个)季度，半年，1年，1年半，2年

时间切片的选择

- 不能太长：保证大多数样本都能覆盖到
- 不能太短：丢失信息

行为评分卡的特征构造

还款率类型特征

定义

与还款行为有关的变量。还款行为由用户的还款能力与还款意愿决定。还款能力强、还款意愿高的客户，发生违约的可能性较小。通常情况下还受到(上)月末欠款余额有关。因此在定义还款行为时，需要将还款额转换成还款率：

$$\text{本月还款率} = \frac{\text{本月总还款额}}{\text{上月末总欠款额}}$$

行为评分卡的特征构造

- 还款率类型特征(续)

常用的还款率类型特征

过去半年内，最大(小)的月还款率

$$\text{maxPaymentL6}=\max\{\frac{PaymentAmount_i}{Outstanding_{i-1}}, \text{i}=1,2,\dots,6\}$$

过去半年内，平均月还款率

$$avgPaymentL6 = \frac{\sum_{i=1}^{6} PaymentAmount_i}{\sum_{j=0}^{5} Outstanding_j}$$

行为评分卡的特征构造

额度使用率类型特征

定义

关于授信额度使用情况的特征。使用额度较多的帐户，未来还款压力较大，相对容易引发违约。同时使用额度也收到授信总额的影响，需要将使用额度转换成使用率：

$$\text{额度使用率} = \frac{\text{本月使用额度}}{\text{授信总额度}}$$

注：

分母是授信总额度而非当前可以使用余额

行为评分卡的特征构造

- 额度使用率类型特征(续)

常用的额度使用率类型特征

过去6个月内，平均额度使用率：

$$avgUrateL6M = \frac{1}{6}\sum_{i=1}^{6}\frac{Spending_i}{Limit}$$

过去6个月内，月额度使用率升高的月份数

$$\text{increase}URateL6M = \#\{\frac{Spending_{i+1}}{Limit} > \frac{Spending_i}{Limit}, i = 1,2,\dots,6\}$$

行为评分卡的特征构造

逾期类型特征

定义

关于M0，M1，M2状态的特征。较高的逾期状态易导致较大的违约概率。

注：用在违约预测模型而非逾期预测模型

常用的逾期类型特征

当前的逾期状态

- 过去6个月的最大逾期状态
- 过去6个月M1、M2、M3的次数

行为评分卡的特征构造

- 消费类型特征

定义

反应用户消费行为的特征。在信用卡客户中，可以建立：

- “国外使用”类型特征
- “提现”类型特征
- “线上消费”类型特征

等等

目录

行为评分卡模型的开发

- 行为评分卡模型的开发步骤

a. 数据预处理

b. 特征衍生

c. 特征处理与筛选

d. 模型的参数估计、特征挑选

e. 模型的性能测试

行为评分卡模型的开发

特征构造

构造了时间窗口为1、3、6、12个月的观察期

每种观察期内包含的行为变量类型：

- 逾期类：*最大逾期状态，M0/M1/M2的次数*
- 额度使用类：*最大月额度使用率，平均月额度使用率，月额度使用率增加的月份*
- 还款类：*最大月还款率，最小月还款率，平均月还款率*

行为评分卡模型的开发

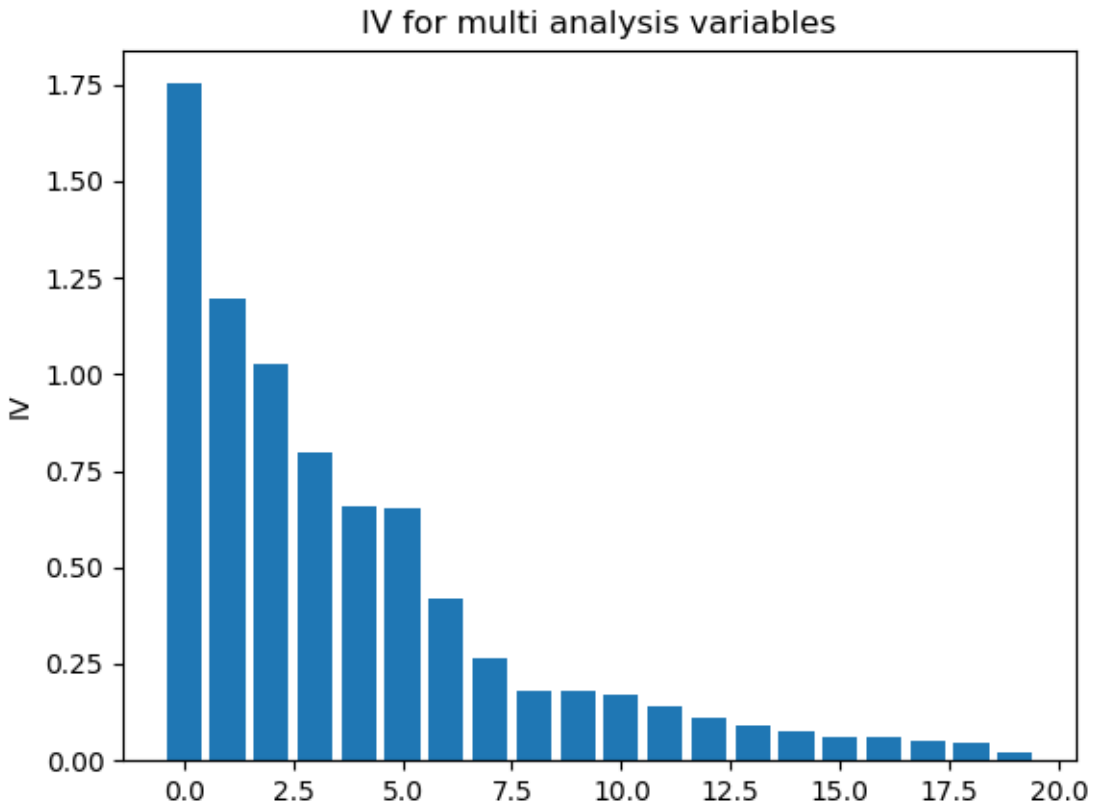
特征挑选

要求

- IV>0.02
- WOE编码后，两两线性相关性低于0.7
- WOE编码后，共线性VIF<10

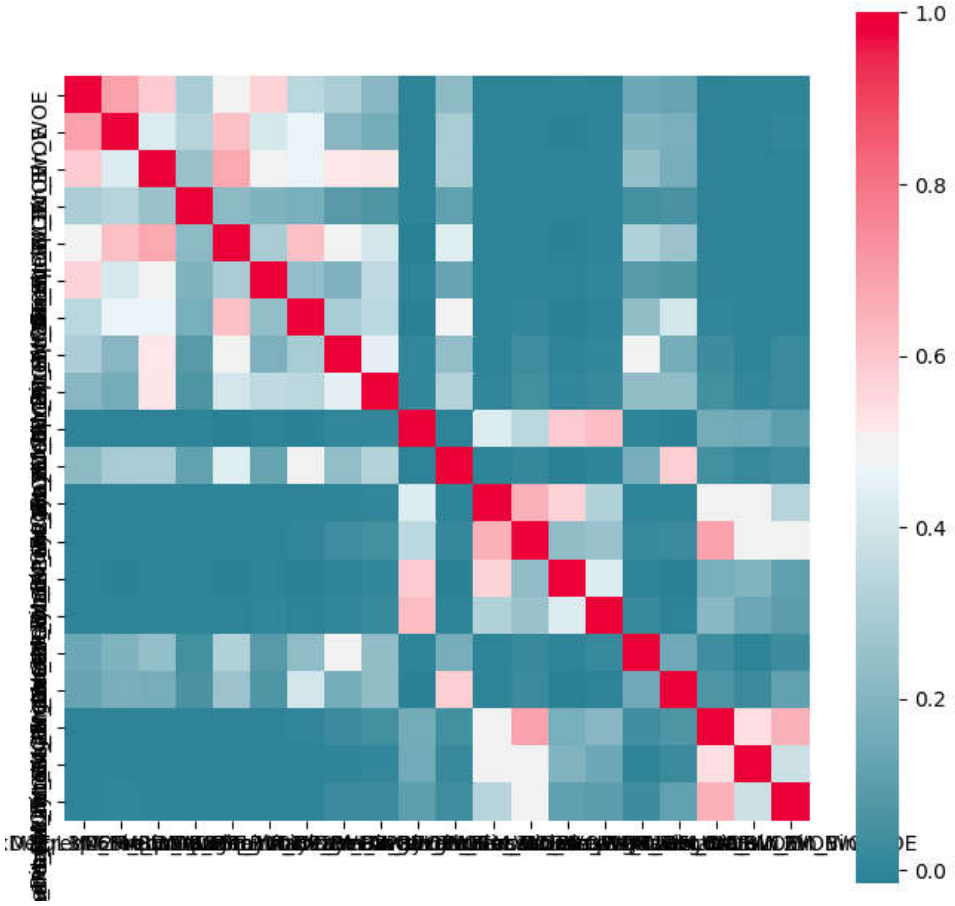
行为评分卡模型的开发

- 特征挑选后的IV分布



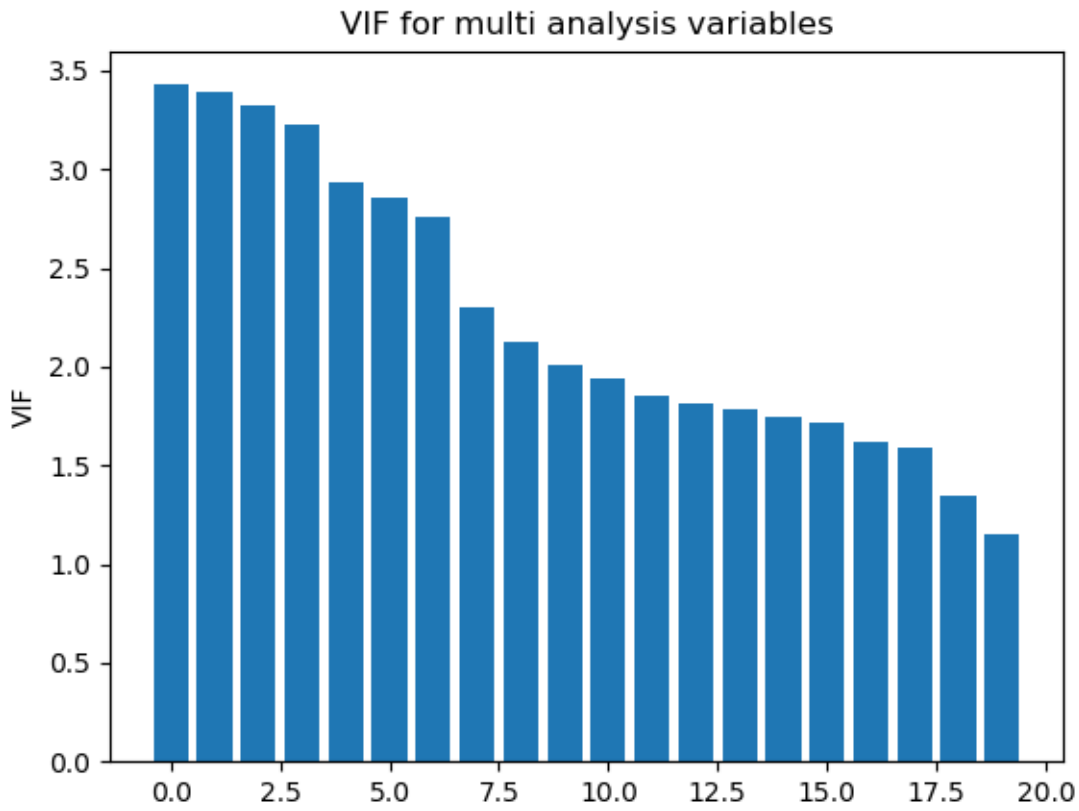
行为评分卡模型的开发

- 特征挑选后的两两线性相关性



行为评分卡模型的开发

- 特征挑选后的多重共线性(VIF分布)



行为评分卡模型的开发

逻辑回归

第一次逻辑回归

Logit Regression Results

Dep. Variable:	label	No. Observations:	28099
Model:	Logit	Df Residuals:	28078
Method:	MLE	Df Model:	20
Date:	Tue, 26 Dec 2017	Pseudo R-squ.:	0.3636
Time:	22:57:12	Log-Likelihood:	-7237.1
converged:	True	LL-Null:	-11372.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
maxDelqL3M_Bin_WOE	-0.7434	0.027	-27.605	0.000	-0.796	-0.691
M1FreqL6M_Bin_WOE	-0.1530	0.036	-4.225	0.000	-0.224	-0.082
M0FreqL3M_WOE	-0.4570	0.035	-13.071	0.000	-0.526	-0.388
M2FreqL3M_Bin_WOE	-0.6817	0.040	-16.876	0.000	-0.761	-0.603
M0FreqL6M_Bin_WOE	0.1532	0.047	3.288	0.001	0.062	0.245
maxDelqL1M_Bin_WOE	-0.1395	0.030	-4.631	0.000	-0.199	-0.080
avgPayL12M_Bin_WOE	-0.0856	0.047	-1.815	0.070	-0.178	0.007
minPayL3M_Bin_WOE	0.2029	0.061	3.319	0.001	0.083	0.323
minPayL1M_Bin_WOE	0.1366	0.072	1.888	0.059	-0.005	0.278
increaseUrateL6M_Bin_WOE	-1.1740	0.078	-15.089	0.000	-1.326	-1.022
maxPayL6M_Bin_WOE	-0.1369	0.069	-1.990	0.047	-0.272	-0.002
avgUrateL1M_Bin_WOE	-0.6618	0.097	-6.789	0.000	-0.853	-0.471
avgUrateL3M_Bin_WOE	-0.5263	0.111	-4.733	0.000	-0.744	-0.308
increaseUrateL3M_WOE	0.0017	0.105	0.016	0.987	-0.204	0.207
increaseUrateL12M_Bin_WOE	-0.0386	0.102	-0.378	0.706	-0.239	0.162
minPayL6M_Bin_WOE	0.1076	0.106	1.013	0.311	-0.101	0.316
maxPayL12M_Bin_WOE	-0.0237	0.111	-0.214	0.830	-0.240	0.193
avgUrateL6M_Bin_WOE	-0.1948	0.166	-1.171	0.242	-0.521	0.131
maxUrateL6M_Bin_WOE	-0.1239	0.131	-0.944	0.345	-0.381	0.133
avgUrateL12M_Bin_WOE	0.1284	0.196	0.655	0.512	-0.256	0.512
intercept	-1.8163	0.024	-76.333	0.000	-1.863	-1.770

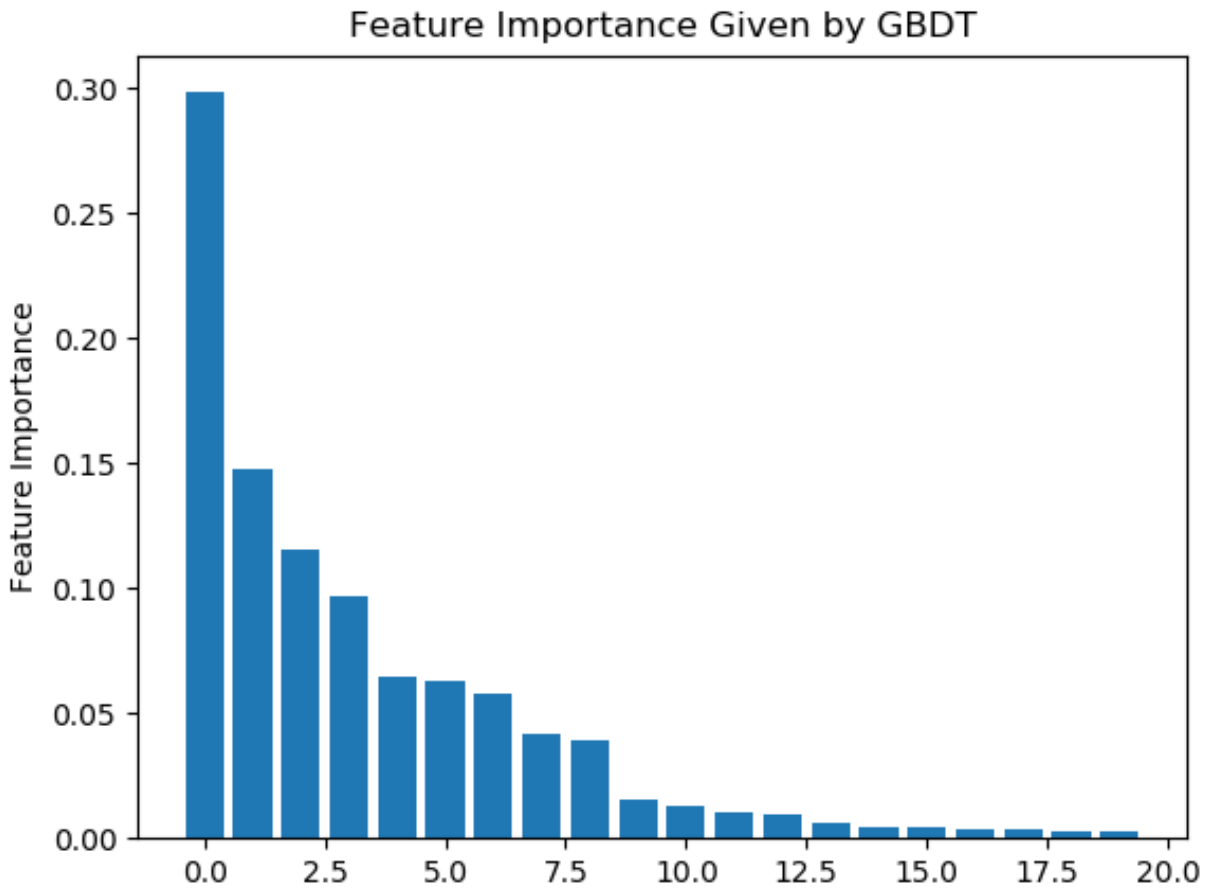
存在符号为正的系数

存在不显著的变量

变量筛选

行为评分卡模型的开发

- 基于GBDT模型的变量挑选



行为评分卡模型的开发

- 基于GBDT模型的变量挑选

步骤一

1，从GBDT的结果中，挑出4个最重要的变量

2，在1的基础上，按重要性逐渐添加新的变量

3，当新添入的变量的符号为正时，舍弃该变量。否则保留

4，直到添加(或剔除)最后一个变量

行为评分卡模型的开发

第二次逻辑回归

Logit Regression Results

Dep. Variable:	label	No. Observations:	28099
Model:	Logit	Df Residuals:	28084
Method:	MLE	Df Model:	14
Date:	Tue, 26 Dec 2017	Pseudo R-squ.:	0.3616
Time:	23:13:17	Log-Likelihood:	-7260.2
converged:	True	LL-Null:	-11372.
		LLR p-value:	0.000

	coef	std err	z	P>|z|	[0.025	0.975]
maxDelqL3M_Bin_WOE	-0.7442	0.025	-29.335	0.000	-0.794	-0.694
increaseUrateL6M_Bin_WOE	-1.1690	0.072	-16.340	0.000	-1.309	-1.029
M0FreqL3M_WOE	-0.3407	0.028	-11.975	0.000	-0.396	-0.285
avgUrateL1M_Bin_WOE	-0.6600	0.086	-7.699	0.000	-0.828	-0.492
avgUrateL3M_Bin_WOE	-0.5108	0.107	-4.780	0.000	-0.720	-0.301
M2FreqL3M_Bin_WOE	-0.6830	0.040	-16.979	0.000	-0.762	-0.604
M1FreqL6M_Bin_WOE	-0.1131	0.033	-3.469	0.001	-0.177	-0.049
maxDelqL1M_Bin_WOE	-0.1364	0.027	-5.063	0.000	-0.189	-0.084
maxUrateL6M_Bin_WOE	-0.1321	0.131	-1.008	0.313	-0.389	0.125
increaseUrateL12M_Bin_WOE	-0.0327	0.101	-0.322	0.747	-0.231	0.166
maxPayL6M_Bin_WOE	-0.0725	0.067	-1.086	0.278	-0.203	0.058
avgUrateL6M_Bin_WOE	-0.1310	0.145	-0.906	0.365	-0.414	0.152
avgPayL12M_Bin_WOE	-0.0256	0.045	-0.571	0.568	-0.114	0.062
maxPayL12M_Bin_WOE	-0.0003	0.109	-0.002	0.998	-0.214	0.213
intercept	-1.8136	0.024	-76.540	0.000	-1.860	-1.767

有不显著的变量

行为评分卡模型的开发

变量显著性检验

对于每一个不显著的变量，单独建立逻辑回归模型，检验显著性

变量	p值
maxPayL6M_Bin_WOE	≈ 0
maxUrateL6M_Bin_WOE	≈ 0
avgUrateL6M_Bin_WOE	≈ 0
avgPayL12M_Bin_WOE	≈ 0
increaseUrateL12M_Bin_WOE	≈ 0
maxPayL12M_Bin_WOE	≈ 0

每个变量都显著。需要再次进行挑选

行为评分卡模型的开发

- 基于带L1约束的逻辑回归模型

a) 对逻辑回归模型中加入L1约束，挑选变量

b) 寻找最优的惩罚因子，使得模型尽可能多地包含变量，且每个变量都显著！

c) 惩罚因子越大，变量越稀疏

d) 经计算，最优的惩罚因子是54，共计有8个变量入选

行为评分卡模型的开发

逻辑回归模型的检验

计算逻辑回归模型在训练集、测试集上的AUC和KS

	AUC	KS
训练集	83.64%	59.82%
测试集	84.43%	64.94%

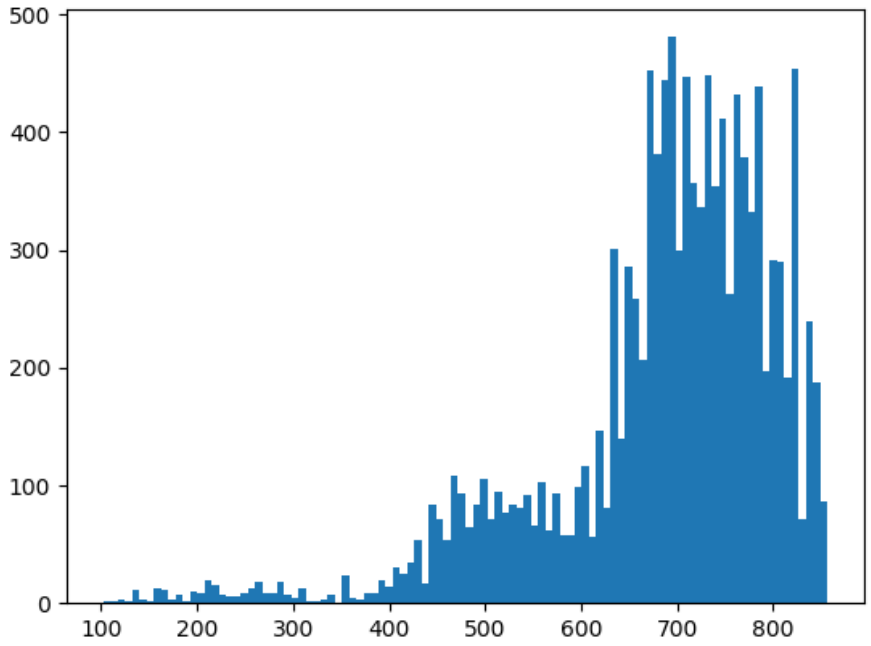
两组的KS、AUC都满足标准，且相差不大。

行为评分卡模型的开发

- 分数计算

$$score = Base\ Point + \frac{PDO}{\ln(2)}(-y)$$

选取Base Point = 500， PDO = 50，分数在测试集上的分布



疑问

- 小象问答官网
  - http://wenda.chinahadoop.cn

课程视频资源扫码：

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象学院
- 新浪微博：小象AI学院