此次零售个贷行为评分模型开发采用分箱型模型的方式，因此需要对变量转换为WOE变量，即对每个变量，按照其各个分组的WOE值，将变量转换为连续变量，其变量值根据不同的分组取各个分组的WOE值，此转换为WOE转换。该转换在下述变量处理细分组/粗分组章节中均有体现。

WOE的计算公式如下所示：

$$f_G(j) = Good\%, f_B(j) = Bad\% \quad \text{that fall into bin} \quad j$$

$$WOE(j) = \log\left[\frac{f_G(j)}{f_B(j)}\right]$$

其中WOE的值为正数表示细分组坏账率好于平均，为负数表示细分组坏账率高于平均。

如开篇方法论所述，具体的变量处理过程分为细分组/粗分组两步骤环节。

细分组（Finebin）

对于变量取值为连续型的变量，其细分组基本方法为：将变量按照其取值排序，细分为不超过10-20组，每组取值约占10%-5%的样本数，分析每个分组的好、坏、不确定账户数以及占比，从而可以了解每个分组的WOE。

对于变量取值为字符型的变量，其细分组基本方法为：将变量的所有取值列举，分析每个分组的好、坏、不确定账户数以及占比，从而可以了解每个分组的WOE。

在获得备选变量的细分组后，可根据IV（Information Value）值进行初步的筛选。IV值是判定变量好坏预测能力的重要参数，经过细分组后的变量按IV值排列，通常IV值低于0.05的变量说明几乎对好坏没有预测能力，这些变量可以直接排除。由于本次行为评分模型候选变量的IV值普遍较低，所以为了考虑尽可能充分多样的变量类型，我们适当放松了IV值的门槛，仅排除了IV值低于0.02的变量。IV值的具体计算公式如下：

$$f_G(j) = Good\%, f_B(j) = Bad\% \quad \text{that fall into bin} \quad j$$

粗分组（Coarsebin）

在获得备选变量的细分组后，可观察变量WOE在细分组情况下的变化趋势是否单调，这意味着该变量随着分组结果的单调变化，其好坏比率趋势也将有单调趋势变化，从而表明该变量具有良好的区分能力。为更好的获得这种结果，可通过将细分组进一步粗分合并来实现，基本方法为：可将WOE接近的相邻多个细分组进行合并，以消除波动趋势，转化为单调趋势的粗分组结果，同时WOE非常接近也意味着其好坏区分能力相同，保留细分没有实质意义，利于最终变量被选入模型后的评分卡使用。在粗分组时，可进一步手动调整切分节点，获得更有利于业务解释

的分界点。粗分组除了尽量使得趋势获得单调趋势之外，需注意尽可能保证每个粗分组中样本总数不少于5%。

需要注意的是，对于变量的趋势波动剧烈，明显不具备合理解释趋势的变量，不需要勉强进行粗分组，以刻意获得单调趋势，对此类变量不建议纳入后续模型分析，可予以剔除。

备选变量经过粗分组之后，重新计算IV值，IV值小于0.02的变量不建议纳入后续模型分析，可予以剔除。同时，单一分箱的占比大于90%的变量（不适用于逾期类字段），也不建议纳入后续模型分析，可予以剔除。

(三)模型优化标准及最终模型变量选择

对模型的表现定义分配数值，好账户为0，坏账户为1，同时在SAS逻辑回归中指定预测事件为“账户成为坏账户”，则运用逻辑回归进行回归时相关的模型变量系数为负数，则表明在此模型中，随着此变量各个分组的坏账率趋势和粗分组的坏账率趋势一致，这样才符合业务逻辑。

从上一步骤粗分组的结果中选择模型的最终变量集，然后运用逻辑回归来确定一个账户是坏的可能性的评分权重。本次使用逐步判别方式实施逻辑回归，变量选择过程中逐一引入变量，每增加一个变量后都要检查去除无助于模型预测能力的变量。

逻辑回归的因变量可以是二分类的，也可以是多分类的，但是二分类的更为常用，也更加容易解释，此次模型开发的因变量就是二分类的（好/坏账户）。逻辑回归对自变量类型一般不做规定，但它要求自变量与因变量的逻辑转换之间应符合线性关系。当自变量为分类变量时，可不必考虑线性关系，但当自变量为连续型变量时，则需要检验二者之间的线性关系是否成立。由于此次开发采用分箱型模型的方法，所有的自变量属于分类变量，所以不需要检验线性关系。

回归结果中的每个变量的估计值，即逻辑回归的系数，就是变量转换后所对应的评分权重，变量的评分权重乘以WOE，即为该变量分箱对应的评分权重，所有分箱的评分权重与常数项之和为账户的评分权重，即行为评分结果。

分箱的评分权重越大，表示风险较低，对应模型评分也越高；反之，分箱的评分权重越低，表示风险较高，对应模型评分越低。同样，账户的评分越高，表明账户变坏账户的风险越低。

模型最终变量需要对其进行检测，包括如下所述的回归结果检查、评估方差膨胀系数和二维变量报告检查。上面描述的几种统计测试便是判断在最终模型中，变量是否被包括或被排除的基础。很多变量需要人工对其进行测试，直至完成每个评分模型的最终结果。

根据Wald卡方检验，要求引入模型的每个变量达到0.95的置信度，即检验的P值小于0.05。对每个变量的估计值都要从经验上，业务上和统计的合理性方面进行检查。这一检查包括模型每个自变量Wald卡方检验的P值是否都小于0.05，这是一般意义上的常规设置，对于样本量比较少的情况，也可以适度放宽这一限制，但是要尽量保持P值越小越好。同时，需要特别检查每个变

量的回归系数，确保其为负数，否则需从最初变量集中去除该变量并进行进一步的逻辑回归分析，如此反复，直至完成每个评分模型的最终结果。

同时，需要特别检查每个引入模型变量的回归系数，确保其为负数，否则需从最初变量集中去除该变量并进行进一步的逻辑回归分析。

另外，对于引入模型的每个变量还需要特别检查两两变量之间的线性相关程度，最常用的是皮尔森（Pearson）相关系数，通常相关系数超过0.6时，认为两个变量之间存在强相关，建议从最初的变量集中排除其中一个变量，并进行进一步的逻辑回归分析。

坚持如此反复，直至完成每个评分模型的最终结果。最终回归结果如下：

最大似然估计分析					
变量	标签	估计	评分卡方	Pr > 卡方	VIF
Intercept	Intercept	-3.8681	-	<.0001	-
PAY_DELQ_PRIN_max_pct_12m	12个月内逾期本金还款占贷款余额的最大百分比	-0.8903	885.1398	<.0001	1.07564
Delq0_cnt_6m	6个月内有过逾期大于0的月份数	-0.683	457.9908	<.0001	1.04648
LNBAL_pct_1to_max_12m	观察月余额占12个月内最大余额的百分比	-1.9478	311.8017	<.0001	1.09861
v_cust_EDUEXPERIENCE	最高学历	-0.8449	271.4502	<.0001	1.02005
PAY_NOR_max_pct_12	12个月内正常还款占余额的最大百分比	-0.6906	91.1741	<.0001	1.34954
v_cust_sex	性别(优先身份证)	-0.9656	59.5435	<.0001	1.00584
v_cust_MARRIAGE	婚姻状况	-0.3831	7.174	0.0074	1.01137
PAY_02to_BAL_13	观察点3个月内还款总额除以前3期余额总和	-0.2245	6.5051	0.0108	1.39035

相关系数	Delq0_cnt_6m	PAY_DELQ_PRIN_max_pct_12m	PAY_NOR_max_pct_12	PAY_02to_BAL_13	LNBAL_pct_1to_max_12m	v_cust_sex	v_cust_EDUEXPERIENCE	v_cust_MARRIAGE
Delq0_cnt_6m	1	0.37466	0.08886	0.09948	-0.08612	0.02503	0.10448	0.04826
PAY_DELQ_PRIN_max_pct_12m	0.37466	1	0.19442	0.15478	-0.08177	0.03609	0.14738	0.06355
PAY_NOR_max_pct_12	0.08886	0.19442	1	0.47482	-0.05735	0.0273	0.08362	0.05401
PAY_02to_BAL_13	0.09948	0.15478	0.47482	1	0.17827	0.04368	0.00027	0.05238
LNBAL_pct_1to_max_12m	-0.08612	-0.08177	-0.05735	0.17827	1	0.04373	0.03622	0.02589
v_cust_sex	0.02503	0.03609	0.0273	0.04368	0.04373	1	0.03193	-0.04347
v_cust_EDUEXPERIENCE	0.10448	0.14738	0.08362	0.00027	0.03622	0.03193	1	0.07595

(六)评分的校准

评分的校准是指通过线性转换将最终模型评分与好/坏比（odds）建立一定对应关系的过程。评分的校准共需要3个参数：标准评分、标准odds和PDO（odds翻倍所需增加的分值），标准odds是指在标准评分时所对应的odds，PDO是指要将评分提高多少分才能使其所对应的odds成为原来的两倍。通过与标准评分作比较，可直观的得到校准后评分所对应的好/坏比（odds），如校准后评分比标准评分高PDO分，则其对应的odds为标准odds的两倍。

评分校准公式为：

模型评分 = 标准评分 + PDO * (ln(odds) - ln(标准odds)) / ln(2)

由于逻辑回归模型的预测结果为ln(odds)，因此上式中的ln(odds)可替换为评分权重(所有变量) 与常数项之和，替换后公式如下：

模型评分 = 标准评分 + PDO * (常数项 + ∑评分权重(所有变量) - ln(标准odds))/ln(2)

= 标准评分 + PDO * (常数项-ln(标准odds))/ ln(2) + PDO * (∑评分权重(所有变量))/ln(2)

即评分校准后，新的常数项 = 标准评分- (PDO/ln(2)) * ln(标准odds) + (PDO/ln(2)) * 常数项

变量的模型评分= (PDO/ln(2)) * 变量评分权重

由于之前提到对模型的表现定义分配数值，好账户为0，坏账户为1，同时在SAS逻辑回归中指定预测事件为“账户成为坏账户”，则运用逻辑回归进行回归时相关的模型变量系数为负。因此，在运用上述公式时，需要先把逻辑回归系数乘以-1，从而得到上述公式中的常数项和变量评分权重。同时，广发银行本次申请评分模型校准参数设定为：标准评分为600分，标准odds为20:1，PDO等于20。根据上述公式，逻辑回归结果与最终模型评分的转换关系为：

模型评分常数项=600-(20/ln(2))*ln(20)+(20/ln(2))*(-1)*逻辑回归结果的常数项

变量的模型评分=(20/ln(2))*(-1)*逻辑回归结果的各变量估计值

变量分组的模型评分=变量的模型评分*分组的WOE

账户的评分是各模型变量分箱校准后评分与常数项之和。

此时，评分卡变量的评分存在负分，如果客户需要所有变量的评分都为正分且最低分相同，我们将会对常数项进行拆分。首先，每个变量的评分加上变量最低分的绝对值，这样每个变量最低分均为0分，剩余的常数项=原常数项-所有变量原最低分的绝对值之和；然后将剩余的常数项等分到每个变量（取整数值），最终模型结果每个变量的最低分相同，账户的评分是各模型变量评分之和。

(七)模型表现

1.区分能力