联通政企数据运营体系建设（评分卡开发）

小组成员：孙奇 吴习溱 刘彦妍 司腾

项目导师：黄云（硕恩网络）

一、背景

在联通的政企客户领域，普遍存在欠费问题，逾期对现金流的影响日益严重。根据 2022 年 9 月的数据，逾期客户数高达 97.9 万，占比达到 43.3%，逾期金额为 318.7 亿，占比 21.9%。此外，还有 8.8%的客户逾期超过两个月，显示出客户欠费问题的严重性。由于缺乏对逾期客户的预警机制、逾期风险、营收规模和客户自身经营风险（失信、受罚）等综合评估以及系统的客户评级，企业面临着许多问题。这些问题导致了客户服务力量和资源分配的不公平，现金流健康度的不清晰，以及无法量化的财务风险。

联通已有的信息化建设为解决这些问题提供了基础，但仍存在一些问题。目前，信息化建设已进入到运营和赋能阶段。在这个阶段，数据标准化和集约化已经有了基础，为联通内部的管理工作提供了更坚实的基础。然而，仍然存在以下问题：首先，企业的决策分析主要基于 Excel 等工具进行手工操作，这种方法容易出错，数据容易遗漏或错误，从而影响决策结果的准确性。而且手工操作效率低下，数据处理和分析的速度缓慢，耗费大量时间和人力成本。其次，目前企业的数据分析仅仅是对现状进行简单的描述，缺乏对数据背后原因的分析和预测。这种情况导致企业只能被动地应对问题，无法预测未来可能出现的问题，从而降低企业的决策效果。此外，目前的数据分析大多采用通用的算法库，缺乏和业务结合的具体模型，导致企业无法将数据分析结果直接应用于业务决策中，进一步降低了决策效果。再次，现有模型的复用性和通用性较低，各省份和各条线上传的模型无法被他人使用，模型商店虚设，导致各个部门之间缺乏数据共享和模型复用，无法发挥数据分析的最大效益。最后，联通缺乏内部分析建模团队，导致数据分析和建模工作都需要外部专业公司

来完成。这种情况既增加了企业的成本，又降低了数据分析的效果。

二、需求

联通需要搭建可解释的模型，类似金融风控中的催收评分模型，用于预测进入催收阶段后未来一定时间内还款的概率，并尽可能提高模型准确率，减少对用户的打扰。具体需求可分为以下几个部分：

1、客户监测

综合分析客户的信誉、经营状况等关键信息。通过外部数据监控客户的不良事件，如违规、工商行政处罚、法人变更、失信被执行、被起诉等。搭建可视化看板，方便筛选异常行为的客户。

2、欠费预测

分析客户已有逾期、费用拖欠和异常行为，掌握政府和企业在不同行业、地域和账龄下的目标风险事件发生频率。对于存在欠费情况的客户，企业可以及时采取措施加强催收，从而有效降低逾期风险。

3、客户评级模型

结合数据探索，确定区分好坏客户的标准，明确分类定义。针对上述目标中提到的逾期类风险进行建模，预测风险水平；对异常行为类风险进行建模和一般性预警。

将上述风险水平指标化，构建评分卡。建立系统的客户评级机制，对客户的信誉和经营状况进行评估，以科学调配服务力量和资源，提高客户服务质量和现金流的健康度。对每个企业评估其风险总分，并拆分为细项得分，为每个企业生成风险评分报告看板。

4、可视化平台促进业务使用

使用 DWF 平台搭建可视化用户评级体系，通过象限分析和组合指标，实现客户质量的监控。

总体而言，联通需要建立有效的客户预警机制，以及系统的客户评级机制。同时，加快数字化转型，优化业务流程，提高运营效率，以更好地适应市场的变化。

三、数据处理

在数据处理方面，主要采用 SQL 进行数据提取，以及使用 R 语言进行数据清理和表链接。联通的相关数据主要包括工商数据和过往欠费数据两类。以下是数据处理的主要步骤：

1、工商数据宽表整理

从 54 张工商信息中提取有效字段，根据独立的企业 ID（entid）进行链接和整合。

2、过往欠费数据宽表整理

a. 选择目标变量：

- 逾期可能性：根据业务沟通，将连续两个月以上出现逾期行为的客户定义为坏客户。

- 逾期严重程度：为了归一化企业自身的季节性消费波动，计算单月逾期金额占年度出账金额的比例。

b. 计算目标变量的逻辑：

- 逾期可能性：根据表现期（2023 年 1 月至 3 月账期）是否出现连续两个月及以上的逾期行为，将客户分为好（=0）和坏（=1）两类。

- 逾期严重程度：计算表现期逾期金额与年度出账金额之间的比例。

c. 使用 SQL 计算目标变量，形成宽表。

3、本地表组合

- 使用企业 ID（entid）和自然人客户 ID（NATURAL_CUST_ID）关联工商数据宽表和过往欠费数据宽表，形成使用表格。

通过以上的数据处理步骤，联通将得到清洗和整理后的数据，以便后续建立客户预警机制和评级模型，并利用可视化平台促进业务使用。

四、样本定义与分布

根据业务目标，我们分别从“变坏的可能性”与“变坏的严重性”两个维度来定义“坏样本”

或正样本，因此本建模项目是双目标驱动的。一方面，从变坏的可能性来看，我们根据甲方的业务经验，将“表现期内（2023 年 1 月至 3 月）连续 2 个月及以上处于逾期欠费状态”的企业客户定义为 M2+正样本（n=2857），剩余样本为 M2+负样本（n=22609），M2+目标下正负样本比例约为 8:1。

另一方面，从变坏的严重程度来看，基于逾期金额占年化出账金额的比例（后简称“逾期金占比”）为依据，进行二分类切割——大于或等于某逾期金占比阈值的为正样本、否则为负样本。因此，下一个问题是如何确定二分类切割的阈值。考虑到本项目最终的目标是输出对企业客户在不同维度上的可解释评分、且维度之间应当尽量保持独立，所以合理划分逾期金占比的二分类阈值，应该综合考虑 M2+的分布。为了分析逾期金额和 M2+的关联关系，我们绘制了在每一个逾期金占比取值下的 M2+召回率（召回率越低说明该取值下，可能会离开 APP）走势情况。根据“肘部法”原则，我们发现在逾期金占比介于 24%~31%之间的时候，M2+召回率下降得最快，逾期金占比超过 30%以后，M2+召回率基本收敛，如图 1 所示。因此，我们取 25%作为逾期金占比二值化的分类阈值，将“2022 年全年账期内，企业客户逾期总额占年化出账金额比例大于或等于 25%”的样本定义为正样本（n=1129）“小于 25%”的企业客户则为负样本（n=21337），正负样本比例约等于 1:5 左右。

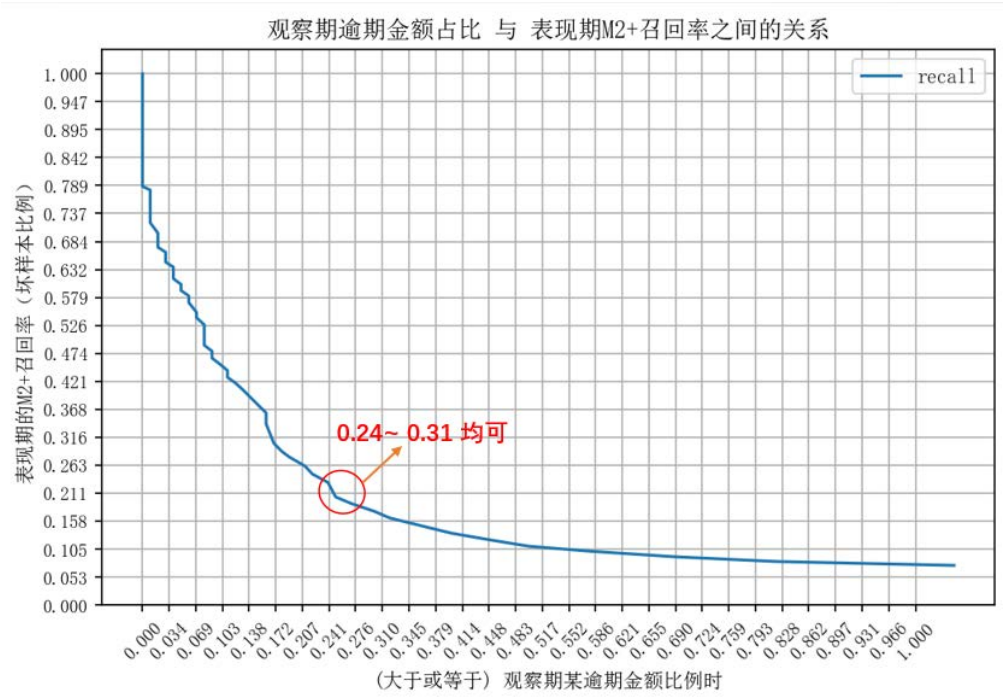

图 1：观察期逾期金额占比与表现期 M2+召回率间关系

五、特征粗筛与模型选择

在评分卡建模开始之前，我们需要平行对比不同的特征策略、模型算法、和随机性，才能决定正式的平法卡模型采用何种模型。以下三个因素是本项目模型实验重点关心的：

特征策略（x2）：同时构造两批预测变量，一组预测变量全部采用原始数据，另一组则全部采用WOE（weight of evidence）做线性变换；

$$WOE(x_i) = \ln\frac{count(x = x_i|Y = 1)/count(Y = 1)}{count(x = x_i|Y = 0)/count(Y = 0)}$$

模型池子（x3）：对比 Logistic Regression（LR）、LightGBDT 和 RandomForest（RF）三种算法的效果，其中 LR 属于线性模型、LightsGBDT 是集成学习里的 boosting 路线，而 RandomForest 是集成学习里面的 bagging 路线。三者各有所长；

随机性控制（x10）：所有实验都用十折交叉验证重复验证，最后聚合出十折平均后的效果评估指标（i.e. 测试集上的 KS 平均值、AUC 平均值、标准差等等）。

六种“特征-模型组合”、在十折交叉验证、以及两个目标的实验下，主要结论如下：

泛化能力：WOE 变换后的逻辑回归，其泛化能力超越 LightGBDT 和随机森林；但是不做 WOE 变换则毫无用处（因为特征原始值不是线性的）；

稳定性：LightGBDT 随机性太强（std>0.2），稳定性不如逻辑回归和随机森林；

可解释性：逻辑回归本质是线性模型，容易被人类经验理解、适合后期评分卡建模；

训练成本：随机森林和 LightGBDT 的参数都比逻辑回归更多、训练时间更长。

	test_KS				test_AUC			
	mean	max	min	std	mean	max	min	std
model								
LR	0.000000	0.000000	0.000000	0.000000	0.499978	0.500000	0.499779	0.000070
LR_WOE	0.503107	0.556327	0.442751	0.038237	0.751553	0.778164	0.721375	0.019118
LightGBDT	0.245142	0.547851	0.000000	0.260919	0.622571	0.773926	0.500000	0.130459
LightGBDT_WOE	0.357658	0.557062	0.000000	0.231403	0.678829	0.778531	0.500000	0.115702
RF	0.492219	0.538411	0.449772	0.027759	0.746110	0.769205	0.724886	0.013879
RF_WOE	0.488621	0.541670	0.464248	0.023900	0.744310	0.770835	0.732124	0.011950

	test_KS				test_AUC			
	mean	max	min	std	mean	max	min	std
model								
LR	0.001810	0.004739	0.000000	0.001856	0.500879	0.502370	0.499766	0.000959
LR_WOE	0.653516	0.684220	0.593559	0.034599	0.826758	0.842110	0.796779	0.017299
LightGBDT	0.391961	0.744703	0.000000	0.372615	0.695981	0.872351	0.500000	0.186307
LightGBDT_WOE	0.369748	0.710835	0.000000	0.322781	0.684874	0.855417	0.500000	0.161391
RF	0.668368	0.708448	0.629575	0.021046	0.834184	0.854224	0.814787	0.010523
RF_WOE	0.641897	0.684094	0.562820	0.040156	0.820948	0.842047	0.781410	0.020078

图 2：对 M2+（上）、逾期金占比二分类（下）的预测

因此，选择逻辑回归是一个相对准确/泛化能力好、稳定、迅速的模型。

在特征粗筛方面，主要综合比对 LR 和 Random Forest 的特征重要性排序，根据模型间的共识和异议，进行初步筛查。由图 3、4 两个目标分别绘制的特征重要性排序图可知，随机森林和逻辑回归都认为重要（importance> 0.02）的特征，后面会直接输入评分卡模型进行拟合评估；相反，随机森林认为不重要（<= 0.02），但逻辑回归认为重要（> 0.02）的需要后面在评分卡建模中通过特征精筛确认。

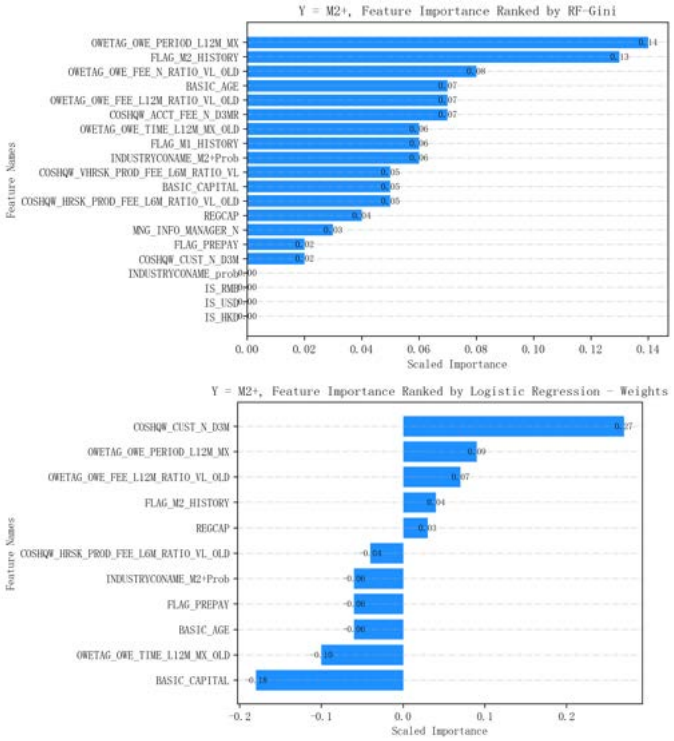

图 3：目标为预测 M2+，随机森林（上）与逻辑回归（下）的特征重要程度排序

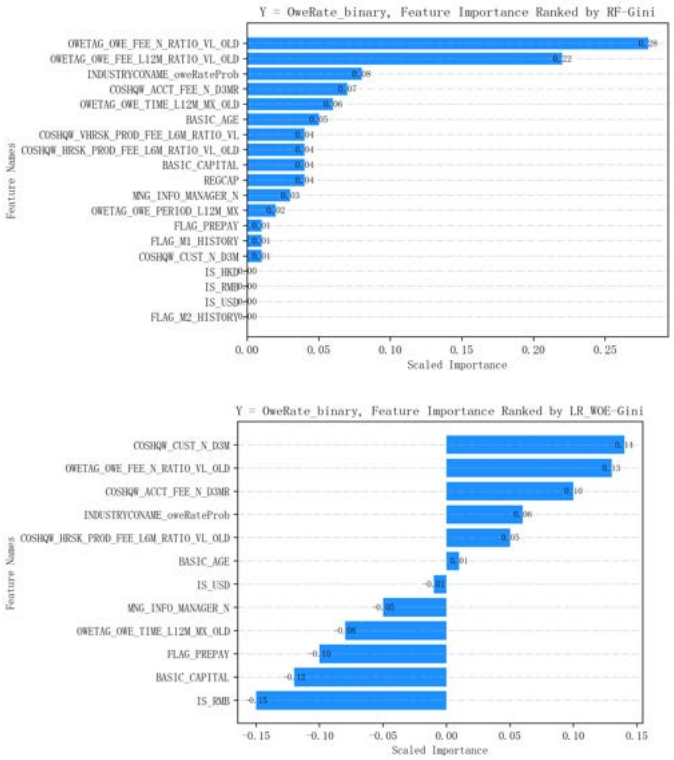

图 4；目标为预测逾期金额占比，随机森林（上）与逻辑回归（下）的特征重要程度排序

五、Toad 评分卡开发

本环节主要采用 Python 的 Toad 包进行评分卡建模。Toad 是由厚本金融风控团队内部孵化开发的标准评分卡库，覆盖信贷评分卡模型开发过程中的特征选择、WOE 分箱、模型性能评测等基本功能。

由前述不同算法初建模型及变量初步筛取的工作，得到用以进行评分卡建模的数据宽表 widetable_cleaned.csv，通过 pandas 读取为数据表 df_wide，包含以自然客户 ID 为主键的 25466 行、25 个字段。去除对评分卡建模无增益的自然客户 ID 列、类别多达 1165 的分类变量国民经济行业名称列、不作为评分卡目标变量列的逾期金额占比列，且清洗后的表内无缺失值。

1、特征选择——基于特征信息量

特征选择即选取对训练数据具有分类能力的特征，以提高模型的学习效率。如果在评分卡建模中，利用一个特征进行分类的结果与随机分类的结果差别不大，则认为该特征不具备分类能力，筛去该特征对模型学习的影响不大。

首先通过 Toad 库中的 quality()函数考察数据表各变量对于两个目标变量（target_M2plus, target_oweRate_binary）的信息增益，所得结果如图 5 所示。其中 IV（infomation value）与变量的正负样本比例之差与对数比相关，衡量变量的信息价值与预测能力，IV 值越高、特征的预测能力越强，信息贡献程度越高；gini（基尼指数）为变量中一个随机选中的样本被分错的概率，指数越小表示被选中的样本被分错的概率越小。本环节主要考察 IV 值，实际操作中，一般将 IV 值大于 0.02 的变量均筛选出来，进入后一步的分析；对于 IV 值在 0.5 到 1 之间的变量，一般结合业务背景来分析，并考虑过拟合可能；IV 值在 1 以上的变量有信息泄露的可能性，一般不会直接进入评分卡建模，而是结合业务实际作为规则或得分调整。

take target_M2plus as target, quality:

	iv	gini	entropy	unique
OWETAG_OWE_PERIOD_L12M_MX	3.985040	0.164667	0.349246	13.0
FLAG_M2_HISTORY	1.855806	0.132636	0.253199	2.0
FLAG_M1_HISTORY	0.794073	0.179066	0.309659	2.0
OWETAG_OWE_TIME_L12M_MX_OLD	0.481045	0.198985	0.350785	12.0
FLAG_PREPAY	0.480477	0.189447	0.327404	2.0
COSHQW_ACCT_FEE_N_D3MR	0.318278	0.193730	0.351067	21488.0
INDUSTRYCONAME_M2+Prob	0.242995	0.196518	0.348625	150.0
BASIC_AGE	0.149022	0.196923	0.351066	1318.0
OWETAG_OWE_FEE_L12M_RATIO_VL_OLD	0.148364	0.198407	0.351067	19204.0
OWETAG_OWE_FEE_N_RATIO_VL_OLD	0.133587	0.198103	0.351067	15108.0
COSHQW_CUST_N_D3M	0.110474	0.199199	0.351059	45.0
COSHQW_HRSK_PROD_FEE_L6M_RATIO_VL_OLD	0.107145	0.198399	0.351067	16989.0
COSHQW_VHRSK_PROD_FEE_L6M_RATIO_VL	0.099807	0.197732	0.351067	14892.0
BASIC_CAPITAL	0.091071	0.198586	0.351058	6772.0
REGCAP	0.085917	0.198573	0.351062	7002.0
MNG_INFO_MANAGER_N	0.083039	0.198930	0.351064	39.0
IS_RMB	0.002044	0.199163	0.350964	2.0
IS_HKD	0.000746	0.199188	0.351028	2.0
IS_USD	0.000152	0.199202	0.351060	2.0

take target_oweRate_binary as target, quality:

	iv	gini	entropy	unique
OWETAG_OWE_FEE_N_RATIO_VL_OLD	3.774740	0.133337	0.443150	15108.0
OWETAG_OWE_FEE_L12M_RATIO_VL_OLD	3.097230	0.163588	0.443102	19204.0
INDUSTRYCONAME_oweRateProb	0.804087	0.256424	0.441881	192.0
OWETAG_OWE_TIME_L12M_MX_OLD	0.774261	0.245286	0.434054	12.0
COSHQW_ACCT_FEE_N_D3MR	0.441829	0.261955	0.443198	21488.0
OWETAG_OWE_PERIOD_L12M_MX	0.211772	0.267488	0.438356	13.0
COSHQW_HRSK_PROD_FEE_L6M_RATIO_VL_OLD	0.115213	0.268940	0.443198	16989.0
FLAG_M1_HISTORY	0.087701	0.268987	0.437631	2.0
FLAG_PREPAY	0.080433	0.268925	0.437876	2.0
MNG_INFO_MANAGER_N	0.078331	0.271593	0.443198	39.0
REGCAP	0.071162	0.270285	0.443198	7002.0
BASIC_AGE	0.070961	0.269902	0.443198	1318.0
BASIC_CAPITAL	0.068312	0.270215	0.443198	6772.0
COSHQW_VHRSK_PROD_FEE_L6M_RATIO_VL	0.064272	0.269329	0.443194	14892.0
IS_USD	0.059407	0.270163	0.439702	2.0
COSHQW_CUST_N_D3M	0.043623	0.271636	0.443198	45.0
IS_RMB	0.036552	0.270545	0.440855	2.0
IS_HKD	0.011158	0.271498	0.442654	2.0
FLAG_M2_HISTORY	0.006979	0.271417	0.442711	2.0

图 5：各变量对目标变量的信息增益。左：目标变量 target_M2plus，右：目标变量 target_oweRate_binary

筛选通过 Toad 库的 selection.select()进行，指定筛去 IV 值小于 0.02 的变量，若两个变量间相关系数在 0.7 以上，则保留 IV 值高的变量。随机抽样取数据中 90%的数据为初步的训练集，通过 transform.Combiner()及 Combiner 对象的 fit()进行变量卡方分箱，规定每箱至少有 5%数据。WOE（weight of evidence，证据权重）表示分组中坏/好样本占所有坏/好样本的比例的差异，WOE 转换可以将 logistic 回归模型转变为标准评分卡格式；分箱的好坏直接影响 WOE 结果，坏样本率不单调、不进行 WOE 编码直接进入逻辑回归模型，一般很难 WOE 求解、难以找到线性公式描述关系。因而在得到推荐分箱的基础上，结合业务可解释性需求及 WOE 变换的坏样本率单调性需求对分箱进

行手动调整，以输入后续的逻辑回归模型。评分卡在模型拟合时使用 WOE 转换后的数据计算最终的分数，计算完成后便无需 WOE 值，可直接使用原始数据进行评分。

分箱结果可视化由 Toad 的 plot.bin_plot()完成。部分变量的分箱结果（含手动调整分箱后）如图 2 所示，各分箱图中红色折线为变量不同分箱段的坏样本率（badrate），蓝色条形为各分箱段的占比。部分变量分箱（如：图 6 中 M2+评分卡的 COSHQW_HRSK_PROD_FEE_L6M_RATIO_VL_OLD，M2+评分卡的 MNG_INFO_MANAGER_N，等）中最左列为取值全为 0，与对变量坏样本率的分箱单调性要求关系不大。

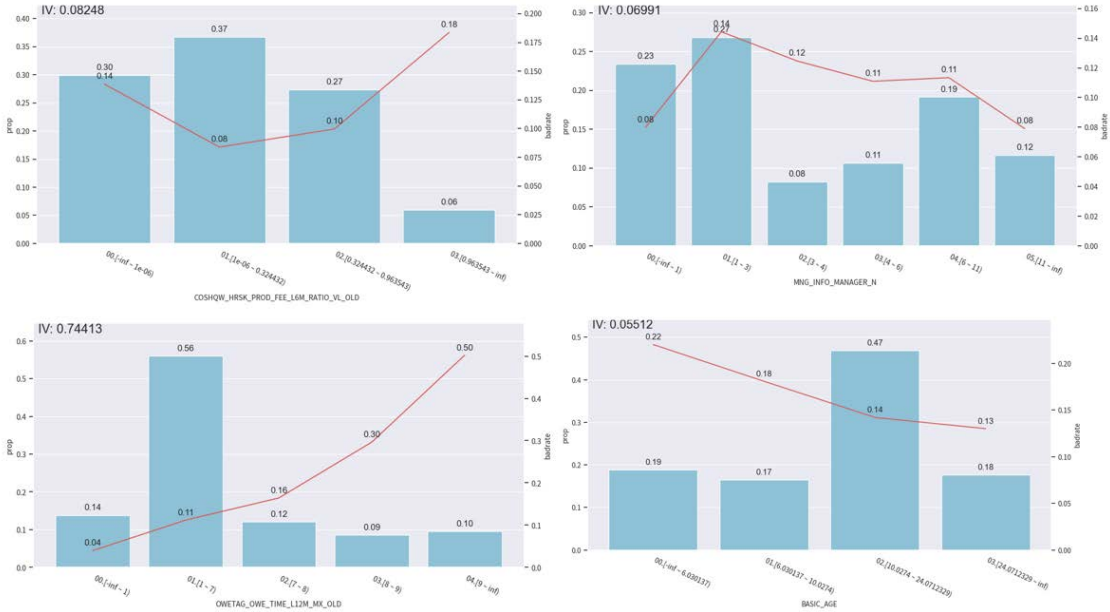

图 6：部分变量的分箱结果。上方：M2+评分卡；下方：rate 评分卡

2、特征选择——基于逐步回归算法

对评分卡模型开发而言，特征数量一般建议在 8~15 之间，特征太多可能使得模型稳定性欠佳、预测能力不足，特征太少则有过拟合的风险，且可迁移性有限。经过信息量的筛选后，下一步是基于算法，选择一个特征的子集，使最终模型对训练集有较好的拟合能力。

本项目中，主要采用逐步回归方法对特征进一步筛选。具体而言，对上一部分分箱挑选的特征输入 selection.stepwise()中，以 AIC 为指标，综合增加最优变量的正向选择与移除最差变量的逆向选择。

3、特征选择——逻辑回归模型拟合

本阶段将结合十折交叉验证方法，分别以两个目标变量构建评分卡，考察各自十次建模下的模型表现，选取最优模型。对于信贷评分卡而言，可解释性的重要性高于模型的预测能力，因而主流采用直观上可解释变量的预测方向的逻辑回归进行建模，并结合前述随机森林（RF）+逻辑回归（LR）共有特征选取最优模型。

本项目采用 sklearn.linear_model 的 LogisticRegression() 进行模型拟合。分别以 target_M2plus 与 target_oweRate_binary 为目标变量，以前述两步特征选择后筛选得到的测试集为输入。十折交叉得到的指标如图 7 所示。其中，KS 值与好坏客户的分数距离相关，衡量模型的区分力，一般以 0.4~0.75 为佳。在两个评分卡的十折交叉验证中，各模型的 KS 值均处于该水平。模型中变量的稳定性一般通过稳定性指数（PSI 值）衡量，计算实际的和预期的分值分布之间的差异，一般以 0.1 以下的 PSI 值为佳。两次十折交叉验证中，各模型的特征变量的 PSI 值均小于 0.1，部分模型的 PSI 计算结果如图 8 所示。

	F1	KS	AUC
0	0.199120	0.561861	0.869883
1	0.201805	0.559562	0.869507
2	0.201664	0.592482	0.879530
3	0.203707	0.587297	0.876525
4	0.202447	0.585779	0.871616
5	0.202447	0.552877	0.867023
6	0.201373	0.582211	0.873611
7	0.201020	0.560890	0.870356
8	0.201169	0.560532	0.868865
9	0.201585	0.561235	0.869773
avg	0.201634	0.570472	0.871669
max	0.203707	0.592482	0.879530
min	0.199120	0.552877	0.867023

	F1	KS	AUC
0	0.277554	0.536301	0.846896
1	0.281171	0.538031	0.848177
2	0.278783	0.538735	0.847257
3	0.277943	0.533618	0.847204
4	0.278589	0.539559	0.848790
5	0.280333	0.543609	0.851669
6	0.277619	0.530603	0.843648
7	0.279806	0.532023	0.846204
8	0.279031	0.536963	0.847327
9	0.278848	0.532490	0.845179
avg	0.278968	0.536193	0.847235
max	0.281171	0.543609	0.851669
min	0.277554	0.530603	0.843648

图 7：十折交叉验证模型指标。左：M2+评分卡；右：rate 评分卡

OWETAG_OWE_FEE_L12M_RATIO_VL_OLD	0.005411
OWETAG_OWE_TIME_L12M_MX_OLD	0.003095
COSHQW_CUST_N_D3M	0.000935
FLAG_M1_HISTORY	0.000328
COSHQW_HRSK_PROD_FEE_L6M_RATIO_VL_OLD	0.000365
BASIC_AGE	0.000790
FLAG_PREPAY	0.000903
BASIC_CAPITAL	0.002045
MNG_INFO_MANAGER_N	0.002349
REGCAP	0.001186
INDUSTRYCONAME_M2+Prob	0.002246
dtype: float64,	

OWETAG_OWE_TIME_L12M_MX_OLD	1.022095e-03
COSHQW_ACCT_FEE_N_D3MR	1.214442e-03
COSHQW_CUST_N_D3M	2.671986e-04
COSHQW_HRSK_PROD_FEE_L6M_RATIO_VL_OLD	7.839204e-04
BASIC_AGE	6.877586e-04
FLAG_PREPAY	7.145271e-04
BASIC_CAPITAL	1.598056e-05
MNG_INFO_MANAGER_N	7.644418e-04
IS_USD	4.780227e-08
INDUSTRYCONAME_oweRateProb	5.266203e-03
dtype: float64,	

图 8：十折交叉验证部分模型变量 PSI。左：M2+评分卡模型 2；右：rate 评分卡模型 6

比较评分卡的另一个工具是 KS bucket。一般对数据集按数据量进行十分位，统计各分位区间内的好/坏样本率、比率（odds）等，若坏样本率没有递增或相邻区间内坏样本率差距太小，认为评分卡的质量不够高。经过各模型间对比，决定选择 M2+评分卡模型 2（共 11 个特征，含 7 个 RF+LR 共有特征）及 rate 评分卡模型 6（共 10 个特征，含 6 个 RF+LR 共有特征），其 KS bucket 如图 9 所示。

	min	max	bads	goods	total	bad_rate	good_rate	odds
0	0.001679	0.012606	11	2280	2291	0.004801	0.995199	0.004825
1	0.012615	0.017311	17	2275	2292	0.007417	0.992583	0.007473
2	0.017312	0.023534	23	2270	2293	0.010031	0.989969	0.010132
3	0.023545	0.032303	58	2234	2292	0.025305	0.974695	0.025962
4	0.032306	0.043297	95	2195	2290	0.041485	0.958515	0.043280
5	0.043322	0.060289	147	2146	2293	0.064108	0.935892	0.068500
6	0.060303	0.087858	201	2079	2280	0.088158	0.911842	0.096681
7	0.087887	0.132128	238	2066	2304	0.103299	0.896701	0.115198
8	0.132248	0.307096	404	1888	2292	0.176265	0.823735	0.213983
9	0.307219	0.980307	1378	914	2292	0.601222	0.398778	1.507659

	min	max	bads	goods	total	bad_rate	good_rate	odds
0	0.000052	0.011344	3	2289	2292	0.001309	0.998691	0.001311
1	0.011369	0.026827	33	2240	2273	0.014518	0.985482	0.014732
2	0.026833	0.042252	64	2243	2307	0.027742	0.972258	0.028533
3	0.042253	0.060335	92	2203	2295	0.040087	0.959913	0.041761
4	0.060361	0.085273	158	2134	2292	0.068935	0.931065	0.074039
5	0.085302	0.119445	224	2068	2292	0.097731	0.902269	0.108317
6	0.119450	0.173370	345	1947	2292	0.150524	0.849476	0.177196
7	0.173379	0.271521	560	1732	2292	0.244328	0.755672	0.323326
8	0.271603	0.457338	837	1455	2292	0.365183	0.634817	0.575258
9	0.457351	0.983327	1421	871	2292	0.619983	0.380017	1.631458

图 9：最终评分卡模型的 KS bucket。左：M2+评分卡模型 2；右：rate 评分卡模型 6

4、评分卡赋分

最后，通过 toad.ScoreCard()为评分卡模型赋分，以投入实际应用。评分卡的赋分基于坏客户与好客户的比率，与信贷应用需求紧密相关，一般由决策层、管理层基于业务需要决定。本项目中对两个评分卡，设置基准评分为 600 分，基准比率为 1/10，若客户的特征对应分箱的比率为基准比率的 2 倍时，得分比基准评分减少 50 分。评分越低的客户，其逾期欠费的趋势（M2+评分卡）或程度（rate 评分卡）预计就越高。通过 export()将评分卡的评分细则输出，两个评分卡的部分评分区间如图 6 所示。以该评分细则对原表 df_wide 的所有客户一一赋分，最后输出 df_output 数据框（xlsx 文件），用于最后的 DWF 系统界面的输入。

	name	value	score
0	OWETAG_OWE_FEE_L12M_RATIO_VL_OLD	[-inf ~ 1e-06)	172.66
1	OWETAG_OWE_FEE_L12M_RATIO_VL_OLD	[1e-06 ~ 0.000428)	-93.29
2	OWETAG_OWE_FEE_L12M_RATIO_VL_OLD	[0.000428 ~ 0.150577)	25.47
3	OWETAG_OWE_FEE_L12M_RATIO_VL_OLD	[0.150577 ~ inf)	39.96
4	OWETAG_OWE_TIME_L12M_MX_OLD	[-inf ~ 1)	88.91
5	OWETAG_OWE_TIME_L12M_MX_OLD	[1 ~ 4)	-3.53
6	OWETAG_OWE_TIME_L12M_MX_OLD	[4 ~ 5)	4.60
7	OWETAG_OWE_TIME_L12M_MX_OLD	[5 ~ 6)	25.03

	name	value	score
0	OWETAG_OWE_TIME_L12M_MX_OLD	[-inf ~ 1)	169.54
1	OWETAG_OWE_TIME_L12M_MX_OLD	[1 ~ 7)	80.87
2	OWETAG_OWE_TIME_L12M_MX_OLD	[7 ~ 8)	39.04
3	OWETAG_OWE_TIME_L12M_MX_OLD	[8 ~ 9)	-22.24
4	OWETAG_OWE_TIME_L12M_MX_OLD	[9 ~ inf)	-94.78
5	COSHQW_ACCT_FEE_N_D3MR	[-inf ~ -0.168269)	-2.85
6	COSHQW_ACCT_FEE_N_D3MR	[-0.168269 ~ -2e-06)	81.57
7	COSHQW_ACCT_FEE_N_D3MR	[-2e-06 ~ 0.368811)	55.30

图 6：部分评分细则。上：M2+评分卡；下：rate 评分卡

5、评分卡特点及业界实际工作

对于信贷评分卡的使用人员，容易解释不同客户的不同得分，也容易理解得分的原因以及不同特征的提高得分方向。评分卡的赋分与特定的逾期欠费比率的对应，也使得业务人员方便设定信贷政策，控制预期的逾期账户比例及其对应的成本。在业界实际的评分卡开发与应用中，还需要进行评分卡实施前、实施后报告，考虑如何将评分卡与企业的总体经营战略相结合、如何衡量客户行为的变化并在信贷策略中对其原因进行说明等问题；并通过要素分析、拒绝演绎等环节，方能以多模型并行、双重矩阵审批、第三方数据联合建模等形式实现落地。限于人力、时间、业务可触达等资源的限制，本项目不展开以上工作。

五、DWF 评分卡用户界面制作

1、数据表格适配

根据评分卡和机器学习的结果，我们对最终 DWF 需要呈现的表格进行设计。

前期，我们主要设计表格时更注重表单的结构-宽表，以变量指标为列，样本企业为行。因为项目最终目标是展示企业的工商指标特征，以及两个评分卡 M2+和 rate 的评分情况、各指标分数。所以我们做了三个表格进行导入清华数为大数据应用系统，但是发现制作表单时设计文本框时比较复

杂；于是，进一步，我们将三个表格进行合并，做一个终表导入数据系统。终表的部分显示如图 7 所示。

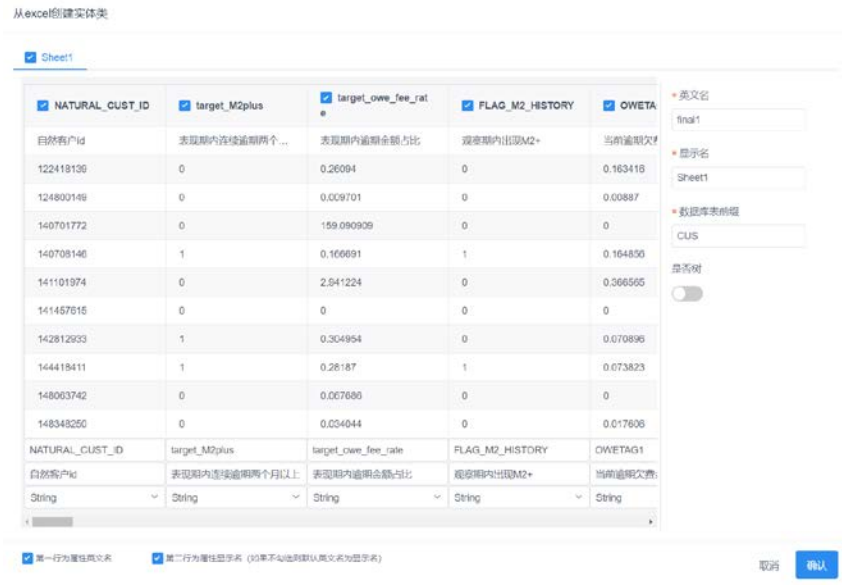

图 7：终表输入 DWF 显示（部分）

导入终表之后，两个问题导致录入失败：一个是标量名称过长且含有非法字符，另外一个是导入时对数据格式应该选择字符型（因为表格中数据有字符有数字，DWF 导入时必须格式统一）。于是我们批量修改了变量名称并设置字符形式数据，成功导入终表数据。DWF 虽然可以自动识别导入数据格式，但是修改格式无法批量修改。

2、表单制作

a. 多对象表单

首要的设计目标是在网页中实现查询功能，即使用者可以在网页上根据自己的兴趣来查询对应公司的工商情况、M2+和 rate 评分卡情况。所以我们在设计时考虑到现实情况和使用习惯，选取六个指标作为可查询的条件，分别是客户 ID, 企业成立时长，企业注册资本，国民经济行业名称，M2+评分卡总分，rate 评分卡总分。首先，客户名称是最便捷的查询方式；其次，企业成长时长和注册资本属于关键大类的工商数据，便于用户查询相关分类的企业群体情况；国民经济行业中分国企，私企等属性，数据样本中行业种类也比较多，便于查询行业情况；最后，可以查询两个评分卡某个

分数段的情况，这样有利于用户了解所有客户中整体的风险和收益情况，做出综合研判。

图 8：多对象表单显示

b. 单对象表单

和项目导师沟通中，我们认为该决策系统的使用主要是联通公司。所以我们尝试在查询功能之后，能够直观对公司的评分卡进行展示；且能够反映出评分卡的标准和区间。在此基础上，我们尝试在单对象表单中建立两个部分，一个是 M2+评分卡，一个是 rate 评分卡。同时，两个评分卡不仅要展示总得分，还需要展示三个维度：具体分指标，分指标对标评分卡的分箱情况，以及最终的分项得分。设计页面情况如图 9 所示。最左侧反映一个总分，然后右侧三列分别是指标，对应分箱以及得分。

图 9：M2+评分卡（上）与 rate 评分卡（下）的查询显示

3、应用操作

在设置应用上，有两个步骤需要实现，首先是绑定表单，其次是单对象表单和多对象表单之间的串联。在多对象表单中查询按钮上设置事件，绑定跳转到单对象列表，实现串联和点击功能。最后的 DWF 成品见网页：

http://i-b30u0hos.cloud.nelbds.cn:8180/dwf/appweb/PCQiYeChaXunFqu

账户为 admin，密码为 8ebe51af。

附 小组分工

孙奇（小组组长） - 评分卡建模，协调小组、导师、课程组间事务沟通

吴习漆 - 项目整体路线设计，ML 模型比较评估

刘彦妍 - 数据格式整理，小组展示 PPT 整合

司腾 - DWF 开发，组会场地与事务协调