中国农业银行 2019 年“雅典娜杯”数据挖掘大赛

赛题介绍

报名并登录大赛平台后，可在数据源中查看赛题数据，具体位置为项目目录 problems 文件夹下，此为共享数据集，建议拷贝到本地，请勿修改、删除或重命名，请勿在 problems 目录下建文件！部分文件较大，请勿直接打开，可用程序载入预览。比赛严禁抄袭，违者经组委会核准，将取消比赛资格，其最终解释权归大赛组委会所有。

赛题 1：线上贷款违约预测

“网捷贷”是一款向符合特定条件的农行个人客户发放的，由客户自助申请、自助审批、自助用信得小额消费贷款产品，为客户提供了便捷的线上贷款渠道。为了降低信贷风险，请根据客户基本信息、交易流水、个人贷款、信用卡等信息，对客户发生逾期的历史数据进行分析并建模，预测线上贷款客户违约可能性。

一、训练数据部分

训练数据集共提供 11 张表：

1) 目标客户列表(YDN1_TARGET)
2) 客户基本信息（YDN1_CUST_INFO）
3) 定期存款账户信息（YDN1_IDV_TD）
4) 金融资产信息（YDN1_ASSET）
5) 理财账户信息（YDN1_FNCG）
6) 国债账户（YDN1_BOND）
7) 基金账户（YDN1_FUND）
8) 金融性交易信息（YDN1_TR）
9) 个贷账户信息（YDN1_LOAN）
10) 信用卡账户余额信息（YDN1_CC_ACCT_BAL）
11) 信用卡客户状态信息（YDN1_CC_CUST_STS）

其中，目标客户列表(YDN1_TARGET)包含 11 个月积累的客户信息，金融性交

易信息（YDN1_TR）提供 1 年数据（包含预测月份），其余表均提供半年数据（包含预测月份）。

二、预测数据部分

预测结果上传文件名：upload.csv

文件格式：utf-8

分隔符：半角逗号

内容格式：客户号，违约概率

违约概率取值区间为[0，1]

示例：（第一行无列名）

1642527539731111,0.4

1610117132571111,0.8

1633519943751111,0.1

三、评价指标

Kolmogorov-Smirnov(KS)是风险评分领域常用的评价指标，反映模型对正负样本的辨识能力，KS 越高表明模型对正负样本的区分能力越强。其计算方法为：

假设 $\mathrm{f}(s|P)$ 为正样本预测值的累计分布函数（cdf），$\mathrm{f}(s|\mathrm{N})$ 为负样本在预测值上的累计分布函数，则 KS 计算方法如下：

$$\mathrm{KS}=\max_{\mathrm{s}}\left\{\left|\mathrm{f}(s|\mathrm{P})-\mathrm{f}(s|\mathrm{N})\right|\right\}$$

四、表结构介绍

为保护用户的隐私和数据安全，所有数据均已进行了采样和脱敏。数据中部分列存在空值或 NULL，请参赛队伍自行处理。

1） 目标客户列表（YDN1_TARGET）

序号	英文字段名	中文字段名	备注
1	CUST_NO	客户编号	
2	DATA_END	数据日期	

3	LABEL	违约标记	LABEL='0'正常 LABEL='1'违约

2) 客户基本信息（YDN1_CUST_INFO）

序号	英文字段名	中文字段名
1	DATA_DAT	数据日期
2	CUST_NO	客户编号
3	PROV_CD	省市代码
4	REG_ORG_NO	开户机构编号
5	CUST_SEX_CD	客户性别代码
6	IPT_EXT_CERT_TYP_CD	参与人外部证件类型代码
7	CUST_EXT_OFFL_NO	客户外部官方编号
8	FST_ARG_CRT_DAT	首份合约建立日期
9	MRGE_STS_CD	婚姻状态代码
10	CULT_DGR_CD	文化程度代码
11	DGR_CD	学位代码
12	GC_BRTH	公历生日
13	YEAR_INCM	年收入
14	FMLY_YEAR_INCM	家庭年收入

3) 定期存款账户信息（YDN1_IDV_TD）

序号	英文字段名	中文字段名
1	DATA_DAT	数据日期
2	PROV_CD	省市代码
3	ORG_NO	机构编号
4	CUST_NO	客户编号
5	ARG_CRT_DAT	合约建立日期
6	CLS_ACCT_DAT	销户日期
7	MATU_DAT	到期日期
8	REG_CAP	开户本金
9	EXEC_RATE	执行利率
10	CRBAL	贷方余额
11	MOTH_CR_ACCM	月内贷方积数
12	MTH_ACT_DAYS_TOT	本月实际天数累计

4) 金融资产信息（YDN1_ASSET）

序号	英文字段名	中文字段名
1	DATA_DAT	数据日期

2	CUST_NO	客户编号
3	ORG_NO	机构编号
4	CCY_CD	币种代码
5	DAY_FA_BAL	当日金融资产余额
6	YAVER_FA_BAL	年日均金融资产余额
7	DAY_AUM_BAL	当日 AUM
8	YAVER_AUM_BAL	年日均 AUM
9	TOT_IVST_BAL	总投资余额
10	MAVER_TOT_IVST_BAL	月日均总投资余额
11	SAVER_TOT_IVST_BAL	季日均总投资余额
12	YAVER_TOT_IVST_BAL	年日均总投资余额
13	FA_BAL_MAX	金融资产余额最大值
14	FA_BAL_MAX_DATE	金融资产余额最大值日期
15	AUM_BAL_MAX	AUM 最大值
16	AUM_BAL_MAX_DATE	AUM 最大值日期

5) 理财账户信息（YDN1_FNCG）

序号	英文字段名	中文字段名
1	DATA_DT	数据日期
2	ARG_NO	合约编号
3	CUST_NO	客户编号
4	PROD_CLS_CD	产品分类代码
5	ARG_CRT_DAT	合约建立日期
6	MATU_DAT	到期日期
7	CLS_ACCT_DAT	销户日期
8	CHANL_CD	渠道代码
9	PROD_RSK_RANK_CD	产品风险等级代码
10	PROD_PFT_TYP_CD	产品收益类型代码
11	EXIT_SHR	退出份额
12	CUST_IVST_CST	客户投资成本

6) 国债账户（YDN1_BOND）

序号	英文字段名	中文字段名
1	DATA_DT	数据日期
2	ARG_NO	合约编号
3	CUST_NO	客户编号
4	ARG_CRT_DAT	合约建立日期
5	MATU_DAT	到期日期

6	ARG_LIF_CYC_STA_CD	合约生命周期状态代码
7	PTPN_AMT	参与金额
8	NET_VAL_TOT_AMT	国债净值额

7） 基金账户（YDN1_FUND）

序号	英文字段名	中文字段名
1	DATA_DT	数据日期
2	ARG_NO	合约编号
3	CUST_NO	客户编号
4	ARG_CRT_DAT	合约建立日期
5	CHANL_CD	渠道代码
6	FUD_PROD_TYP_CD	基金产品类型代码
7	RSK_RANK_CD	风险等级代码
8	FUND_BAL	基金余额

8） 金融性交易信息（YDN1_TR）

序号	英文字段名	中文字段名
1	CUST_NO	客户号
2	TR_CD	交易代码
3	TR_DAT	交易日期
4	TR_AMT	交易金额

9） 个贷账户信息（YDN1_LOAN）

序号	英文字段名	中文字段名
1	DATA_DT	数据日期
2	ARG_NO	合约编号
3	CUST_NO	客户编号
4	ACTG_ORG_NO	核算机构编号
5	PROV_CD	省市代码
6	MATU_DAT	到期日期
7	LN_TERM	贷款期限
8	PROC_STS_CD	核销状态代码
9	ARG_CRT_DAT	合约建立日期
10	BUS_BREED_CD	业务品种代码
11	GRTE_CTRT_NO	担保合同编号
12	RSK_CLS_CD	贷款十二级分类代码
13	FORM_STS_CD	贷款形态代码

14	NML_CAP_BAL	正常本金余额
15	TOT_PRVD_AMT	累计发放金额
16	TOT_REVK_AMT	累计收回金额
17	MTH_NML_CAP_ACCM	月内正常本金积数
18	MTH_ACT_DAYS_TOT	本月实际天数累计

10) 信用卡账户余额信息（YDN1_CC_ACCT_BAL）

序号	英文字段名	中文字段名
1	DATA_DAT	数据日期
2	SOR_CUST_NO	源客户编号
3	CUST_NO	客户号
4	GENR_AC_QUOT	一般授信额度
5	TOT_AC_QUOT	总授信额度(含专项)
6	CAN_AMT	核销金额
7	CAN_RETURN_AMT	核销再回收金额

11) 信用卡客户状态信息（YDN1_CC_CUST_STS）

序号	英文字段名	中文字段名
1	DATA_DAT	数据日期
2	CUST_NO	客户编号
3	CUST_CYC_CR_IND	客户循环信用标识
4	CUST_CD_VLU	客户级 CD 值