易明 YIMING

清数易明

（润乾参股企业）出品

智能建模案例集合

案例列表 1/2

易明 YIMING

	专家人工建模	易明智能建模
案例1 个人用户分期违约预测-银行	AUC: 0.924 前10%lift: 8.0	AUC: 0.911 前10%lift: 7.95
	模型稳定性（模型在测试集和训练集得衰减率）：3.85%	模型稳定性：0.76%
	建模时间：几周	建模时间：13分钟（数据预处理+建模）
案例2 小微企业信贷客户违约预测-银行	AUC：0.972 前10%lift: 9.8	AUC：0.987 前10%lift: 9.6
	模型稳定性：2.6%	模型稳定性：0.9%
	建模时间：几周	建模时间：17分钟（数据预处理+建模）
案例3 多产品精准营销-银行	电话营销验证预测清单 SAS EM 模型有意向比例： 39.31%	电话营销验证预测清单 易明模型有意向比例： 38.27%
案例4 车险定价-保险	GINI：0.608	GINI：0.663
	重要衍生变量：无	重要衍生变量：3个
	建模时间：1-2个月	建模时间：60分钟（数据预处理+建模）
案例5 反欺诈-保险	AUC: 0.8542	AUC: 0.8534
	建模时间：1-2个月	建模时间：30分钟（数据预处理+建模）
	建模人员：专业建模团队	建模人员：无建模经验

案例列表 2/2

易明 YIMING

	专家人工建模	易明智能建模
案例6 续保预测-保险	AUC: 0.7435	AUC: 0.7442
	建模时间：1-2个月	建模时间：50分钟（数据预处理+建模）
	建模人员：专业建模团队	建模人员：无建模经验
案例7 健康风险预测案例-保险	前10%提升度：2.26	前10%提升度提高50%
	重要变量筛选困难	重要变量筛选711->20
	建模时间：2个月	建模时间：50分钟
案例8 个人信贷违约预测-征信	AUC: 0.957	AUC: 0.965
	模型稳定性：未知	模型稳定性：0.8%
	建模时间：2个月	建模时间：5分钟（数据预处理+建模）
案例9 资金流管理-金融信贷业务		Auc: 0.69~0.81 比赛排名6%

案例1 个人用户分期违约预测-银行

易明 YIMING

业务需求

智能建模与手工建模效果对比

	训练集 AUC	测试集 AUC	模型衰减	测试集前10%lift
模型1	1	0.973	0.027	9.22
模型2	0.999	0.971	0.028	9.18
模型3	0.999	0.968	0.031	9.09
模型4	0.998	0.922	0.076	7.9
模型5	0.996	0.965	0.031	8.63
模型6	0.995	0.959	0.036	8.77
模型7	0.993	0.927	0.066	7.99
模型8	0.988	0.956	0.032	8.63
模型9	0.982	0.928	0.054	7.99
模型10	0.976	0.914	0.062	7.76
模型11	0.969	0.919	0.05	7.85
模型12	0.961	0.924	0.037	7.95
易明智能建模	**0.918**	**0.911**	**0.007**	**8.0**

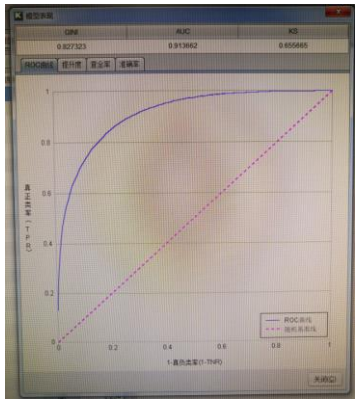
效果实测

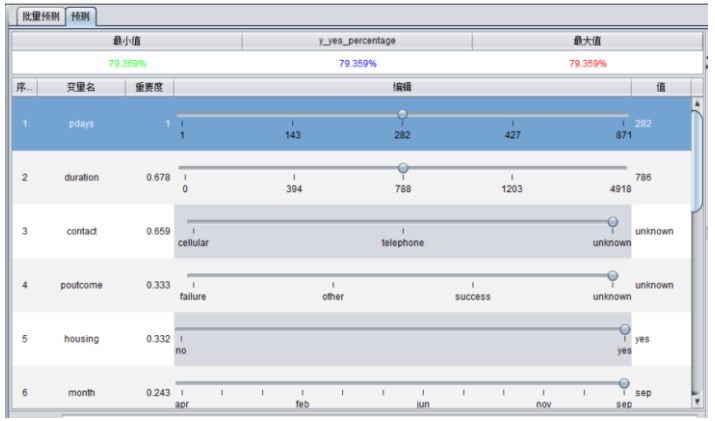
效率高：
13分 VS 1个月

无需手动调模型：
1次 VS 12次

模型衰减度低：
0.007 VS 0.037

为个性化分期策略提供建议





案例2 小微企业信贷客户违约预测-银行

易明 YIMING

智能建模与手工建模效果对比

	智能建模	传统建模
建模时间	17分钟 (数据预处理+建模)	-
建模数量	1	1
训练集auc	0.996	0.998
测试集auc	0.987	0.972
前10% lift	9.8	9.6
原数据维度	5500+	5500+
建模数据	36000+ / 453MB	300000+/ 4.5GB

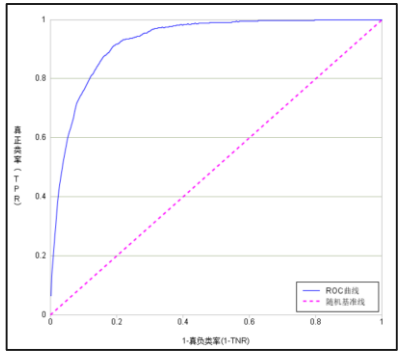
效率高：
17分 VS 1~2月

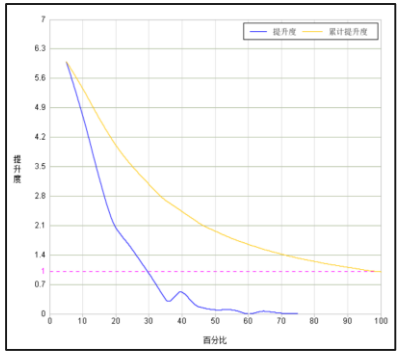
模型效果好：
Auc：0.987
前10% lift：9.8

时间变量信息自动提取

业务需求

效果实测





案例3 多产品精准营销-银行

易明 YIMING

银行需要挖掘潜力客户清单，推荐金融产品组合包，对目标客户进行多产品组合同步营销，夯实以客户为中心的经营体系，突出重点客群的综合经营

案例3 实测-建模效率

易明 YIMING

智能建模	客群1	客群2	客群3	客群4
建模人数	1	1	1	1
模型数量	13	13	13	13
建模时间	1.5小时/个	1.5小时/个	1分钟/个	2分钟/个
数据量	百万级别	百万级别	几千	几万

智能建模 VS 手工建模：

	模型数量	时间	项目参与人数
智能建模	50-60个	2周	1人
手工建模	不适合大量建模	1周~2个月/个 （实际取决于模型复杂程度和建模人员水平，时间不可控）	数人

效率高：
2周可建50~60个模型

人力成本低：
1名普通人员即可完成

批量生产，模型质量稳定

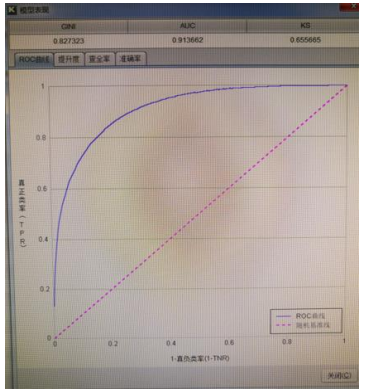
案例3 实测-模型效果

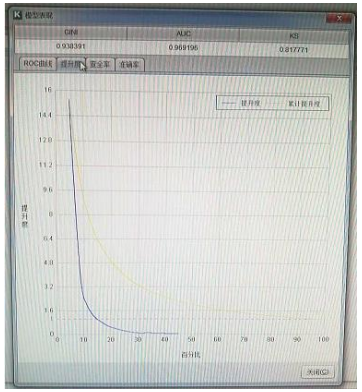
易明 YIMING

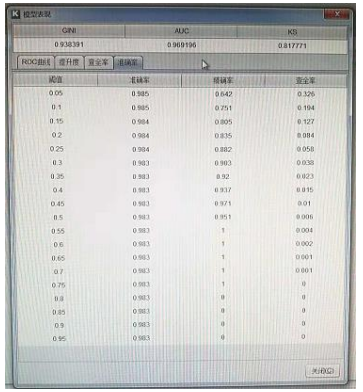
以客群1某产品为例：

基于累计提升度和累计捕获率的性能

	累计提升度	累计捕获率	AUC值
前5%	14.4	72%	>0.9
前10%	9.4	94%	
前15%	6.3	94.5%	
前20%	4.8	96%	
客群1当前该产品购买率为1.72%			







模型提升度高：

前5%的数据购买率提升14.4倍

客户捕获率高：

前5%数据可铺货72%的目标客户

案例3 实测-模型收益

易明 YIMING

生成营销清单如下：

顾客	金融产品组合包	1	2	3	4
大明	组合3	产品1	产品3	产品2	产品4
小张	组合1	产品a	产品c	产品d	产品b
小王	组合2	产品B	产品C	产品A	产品D
大刘	组合3	产品3	产品2	产品4	产品1
…	…	…	…	…	…

千人千面，个性化的营销清单

不同客户，不同的营销内容，不同的营销顺序

使用模型后，日均增长额提高了70多倍！

案例4 车险定价-保险

车险市场竞争激烈，保险公司希望建立更精确的定价模型，帮助公司更精准的定位客户。一方面利用价格弹性以较低的溢价带来更多低风险客户，另一方面以更高的溢价阻止更多高风险客户，从而提高利润率。

type	字符	车种
class	字符	车辆类别
seat	数值	座位数
Ton	数值	吨位
carage	数值	车龄
newvalue	数值	车价
series	字符	车系
age	数值	年龄
gender	字符	性别
seat_vlp	数值	车上人员投保座位数
glass	数值	玻璃（国产/进口）
si_tp	数值	三者限额
si_vld	数值	车上司机限额
si_vlp	数值	车上乘客限额
nprem	数值	净保费
eprem	数值	已赚保费
num_com_p	数值	续保客户上年出险次数
lr_com_p	数值	续保客户上年赔付率
py	字符	保单年份
claim_no_sal	数值	交强险平台上年出险次数
claim_amount_sal	数值	交强险平台上年出险金额
level_sal	字符	交强险平台NCD
claim_no	数值	商业险平台上年出险次数
claim_amount	数值	商业险平台上年出险金额
commercial_claim_record	字符	商业险平台NCD
renew	字符	新/续/转保

案例4 现有方案不足

原方案：

数据分析师通过手动编程的方式，分析和探索数据，进行数据预处理、建模等操作，有以下几点不足：

建模周期长

建模周期长

模型更新慢

采用线性模型

GLM广义线性模型，

模型精度低

数据质量差

有大量的缺失值和具有高基数的分类变量，与时间相关的特征难利用

案例4 实测-高基数变量处理

易明 YIMING

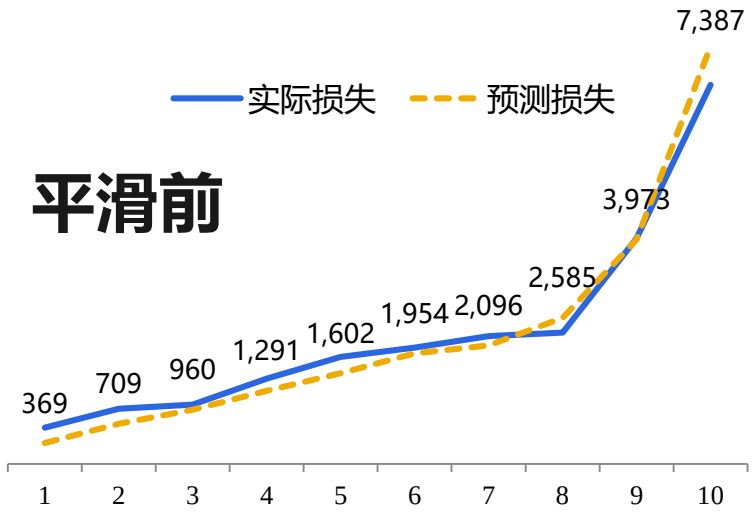
以**车型变量**为例，对承保标的数较少（小于1000台）的车型，对x平滑前后分别预测y，10分段的提升曲线效果如下。**平滑后两端业务过拟合的现象改善。**

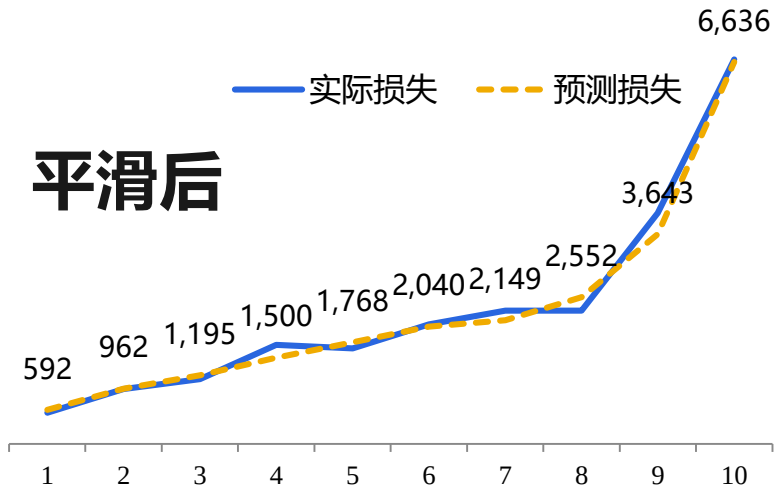
	模型表现
平滑前	86.44%
平滑后	86.64%
效果提升	0.23%

平滑处理后，GINI提升0.23%

注：

表中GINI是对承保1000台以下标的排序测算得到。若基于全体测试集，平滑前是15.71%，平滑后是15.73%。其中，承保1000台以下占比14.5%。



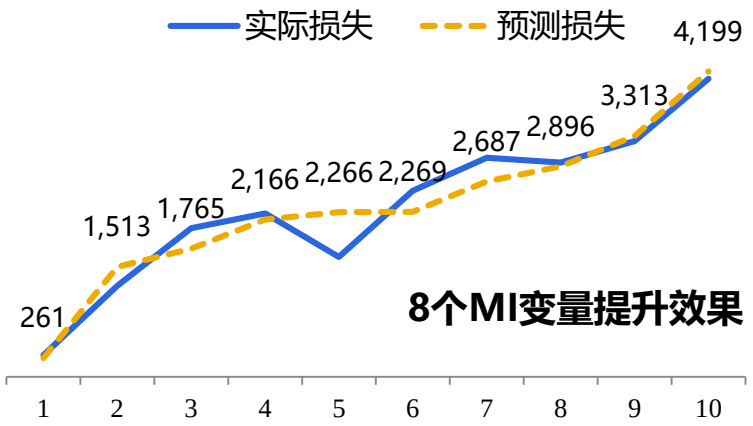


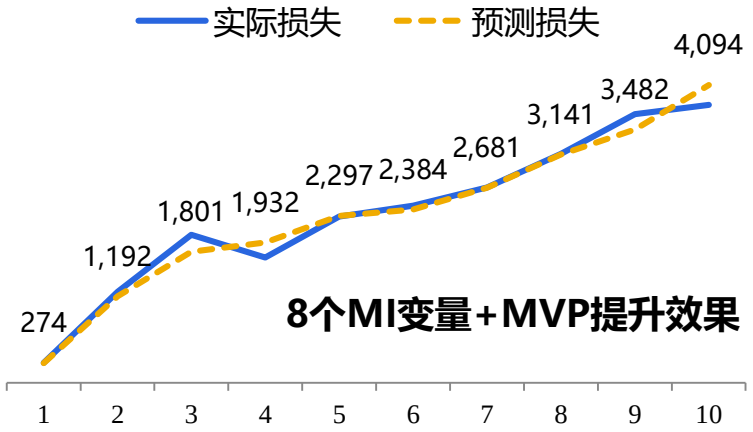
案例4 实测-缺失值处理

投保组合、上年出险记录等共计8个变量的缺失组合有 $2^8 = 256$ 种，分析缺失组合可以对风险细分。为了分析变量缺失的预测效果，对8个变量分别构造是否缺失的MI变量；此外生成由它们组合生的MVP变量。**对比MVP加入前后的效果改善**

	模型表现
MVP之前	74.04%
MVP 平滑之后	74.98%
效果提升	1.27%

加入MVP之后，GINI效果提升1.27%





案例4 实测-建模效率和效果

易明 YIMING

采用GLM和神经网络相结合的方法制定出新的定价模型，比原定价模型基于赔付的GINI表现提高了**12%**。

显著提升了定价模型的表现，使得保险机构的收益最大化

建模效率大幅度提升

建模结果对比

	智能建模	传统建模
建模时间	**60分钟**/模型 (数据预处理+建模)	1-2个月
建模数量	**2** 组合模型，模型效果更好	1
模型表现	GINI 0.683	GINI 0.608
重要衍生变量	3 自动生成衍生变量	-
建模数据量	1380000+ / 4G+	

非专家也能建模，模型质量稳定可靠

案例5 反欺诈-保险

易明 YIMING

建模结果对比

	智能建模	传统建模
人数	1	1
建模时间	30分钟/模型 (数据预处理+建模)	1-2个月
建模数量	1	1
模型表现	0.8542	0.8532
重要衍生变量	3	-
建模数量	1300000+ / 1.6G+	1380000+/ 1.6G+

效率高：
30分 VS 1-2个月

自动生成重要衍生变量

成功挑战阳性样本比率低的数据

业务需求

效果实测

案例6 续保预测-保险

易明 YIMING

传统续保预测建模的问题：

- 地域差异性明显，特殊区域预测偏差大
- 数据维度大，变量筛选难
- 建模周期长，模型生命短
- 阳性样本低，容易过拟合

使用易明智能建模工具：

- 分区域精准建立各自的续保预测模型
- 智能筛选重要变量
- 建模时间短，模型更新快
- 提升准确度及稳定度

潜在意义：模型更具有地区针对性和适应性，快速有效提升了续保预测的能力，从而提升客户续保率，使保险机构收益得到提升。

- 目标：保险公司希望建立模型预测客户续保概率，制定续保策略，从而提升客户续保率，防止顾客流失。
- 智能建模工具综合各地区差异性分别进行快速建模，建模用时仅50分钟，模型综合表现高于传统建模。

建模结果对比

	智能建模	传统建模
建模时间	50分钟/模型 (数据预处理+建模)	1-2个月
建模数量	1	1
模型表现	0.7442	0.7435
建模数量	1300000+ / 4G+	1380000+/ 4G+

案例7 健康风险预测案例

易明 YIMING

保险公司为拓展其产品线，新增一个关于老年人低健康风险的保险产品。此次拓展对于该公司来说将面临一个全新的领域，因此希望制定“如何高效获取到低健康风险客户的市场策略”。

目标变量	ID 变量	Character变量	指标变量 (IND Prefix)	健康描述变量	二值型指标变量	无法使用变量
MED_SUPP_IND	IND_ID_NBR and SBL_CONSUMER_ID	名义变量	二值变量	名义变量	二值变量	
		29	27			

保险公司数据

第一群疾病 (AIL1 Prefix)	第二疾病群 (AIL2 Prefix)	第三疾病群 (AIL	牙齿疾病 (DENTAL	疾病治疗 (TRE1
二值变量	二值变量	二值变		
71	53	21		

健康与疾病数据

家庭生活数据（HH Prefix)		生活方式数据（LIFE	地区计数数据（DISCNT
区间变量	名义变量	区间变	
59	322	59	

专业供应商数据

项目难点

- **数据来源多，质量难以评估：** 公司内部数据、健康与疾病相关数据、专业供应商数据
- **数据维度多：** 七百多个维度，难以筛选可以识别潜在客户的重要因素
- **地域特点强：** 健康风险会有地域差异性，需要针对重点区域单独建模

案例7 现有方案不足

易明 YIMING

原方案：

数据分析师们用R/Python通过手动编程的方式，进行数据预处理、建模等操作，不足之处如下：

工作量大

需要手动查看和探索711个变量的数据质量和重要信息；

项目周期至少2个月

结果不稳定

数据探索结果取决于分析师的水平，如果分析师没有探索足够广泛的模型或程序结构，可能会丢失数据的重要方面

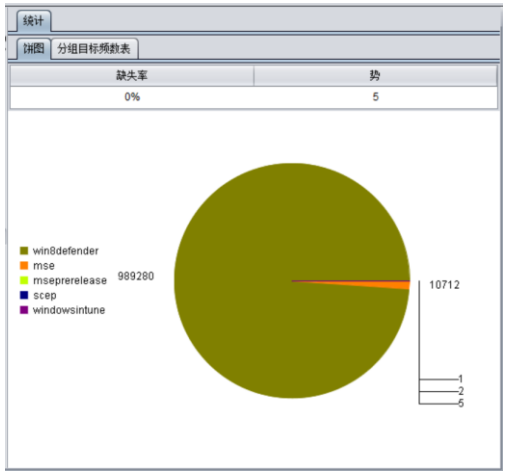
一个模型用全国

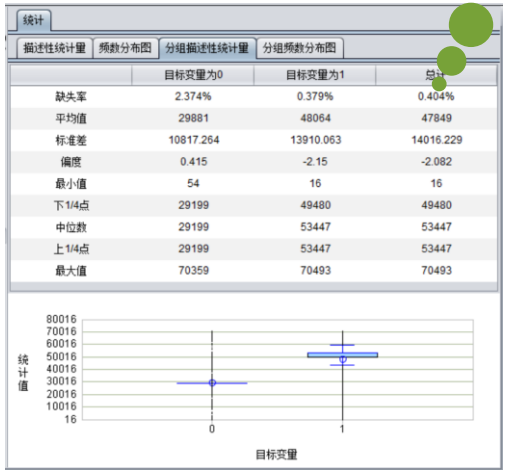
数据分析师需要大量的建模专业知识，高水平人才少；

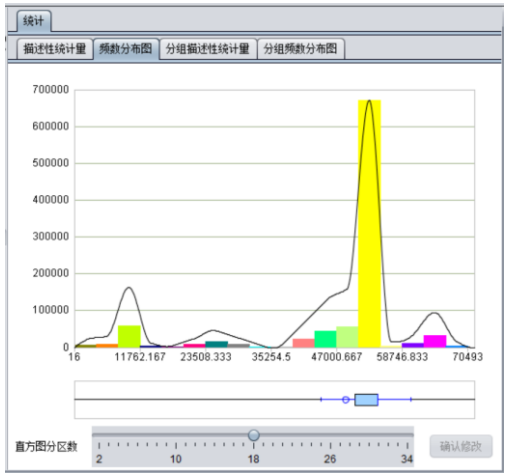
建模能力有限，全国只能通用一个模型，模型效果大打折扣

案例7 实测-数据检测

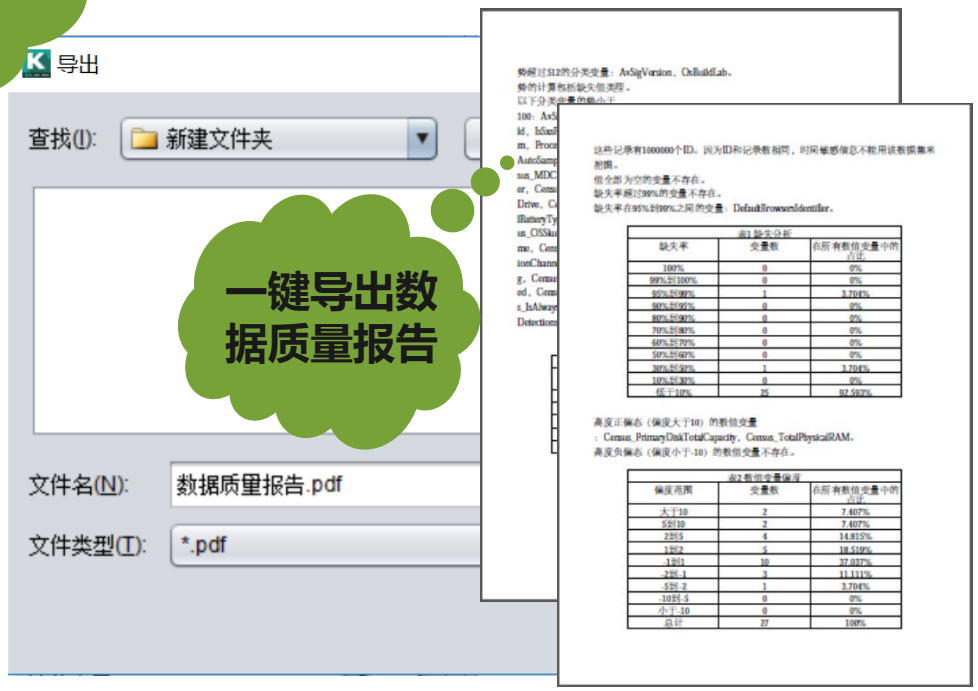
15分钟700+变量自动检测完毕







一键导出数据质量报告



数据导入后，可自动检测缺失情况，识别数据类型，计算各种统计量，自动剔除质量较差的变量等，检测完毕后可导出质量分析报告

案例7 实测-建模效率和效果

易明 YIMING

Table 1 Model Performance (Overall Health Risk)

Decile	Frequency	Predicted: High Risk Perc	
1			
2			
3			
4	122,642		
5	122,641		
6	122,642		
7	122,642		
8	122,642		
9	122,642		
10	122,642		

模型1

Table 2 Model Performance (Type I Health Risk)

Decile	Frequency	Predicted: High Risk P		Model Lift
1				2.39
2				1.89
3				1.46
4	122,642		14.4%	1.12
5	122,641		10.5%	0.81
6	122,642		7.77%	0.60
7	122,642		6.70%	0.52
8	122,642		5.99%	0.46
9	122,642		5.22%	0.40
10	122,642		4.33%	0.34

模型2

Variable Name	Ranking
FIPS_COUNTY_CD	1
AGE_IN_MONTHS_0_TO_11 and AGE_IN_YRS	2
HH_CNSS_EDUCATION_LEVEL	3
HH_FAM_POSITION_CD	4
BCBS_LANGUAGE_PREF_CD	5
HH_CNSS_PCT_HOMEOWNER	6
BCBS_ETHNICITY_CD	7
HH_ESTIMATED_HH_INCOME	8
GENDER	9
HH_DWELLING_TYPE	10
HH_TRAVEL_PERSONAL_	11
HH_CNSS_PCT_HOMEOWNER	12
HH_CONGRESSIONAL_DISTRICT	13
HH_BUYER_RETAIL_DOLLARS	14
HH_CNSS_AVG_NBR_AUTOMOBILE	15
MAILABILITY_SCORE	16
HH_LENGTH_OF_RESIDENCE	17
HH_HEALTH_INSURANCE_RESPND	18
HH_BUYER_TOTAL_DOLLARS	19
HH_BUYER_ORDERS_GENERAL_MERCHDSE	20

Table 3 Model Performance (Ty

Decile	Frequency	Predicted Risk Perc	
1	122,641		
2	122,642		
3			
4			
5			
6	122,642		
7	122,642	9.32%	0.64
8	122,642	8.48%	0.58
9	122,642	7.57%	0.52
10	122,642	6.47%	0.44

模型3

rmance (Type III Health Risk)

Decile	Frequency	Predicted: High Risk Percentage	Model Lift
1		24.3%	2.92
2		17.1%	2.05
3			1.50
4			1.06
5			0.71
6		4.04%	0.49
7	122,642	3.30%	0.40
8		2.82%	0.34
		2.48%	0.30
		1.95%	0.23

模型4

轻松筛选出20个重要变量

建模效率大幅度提升

建模结果对比

	智能建模	传统建模
建模时间	**50分钟**/模型 (数据预处理+建模)	2个月
建模数量	分地区建模，突出地域特色 **4**	1
模型表现	不同地域模型的前10% lift表现均比传统建模方式全国的提高**50%**	2.26
重要变量筛选	711->**20**	筛选困难

非专家也能建模，模型质量稳定可靠

案例8 个人信贷违约预测

易明 YIMING

目标

- 建立模型，给出用户信贷违约概率
- 给出用户合理的信贷额度
- 让业务人员根据经验选择数据建模，帮助业务人员接受模型的应用与普及
- 提高违约客户捕获率

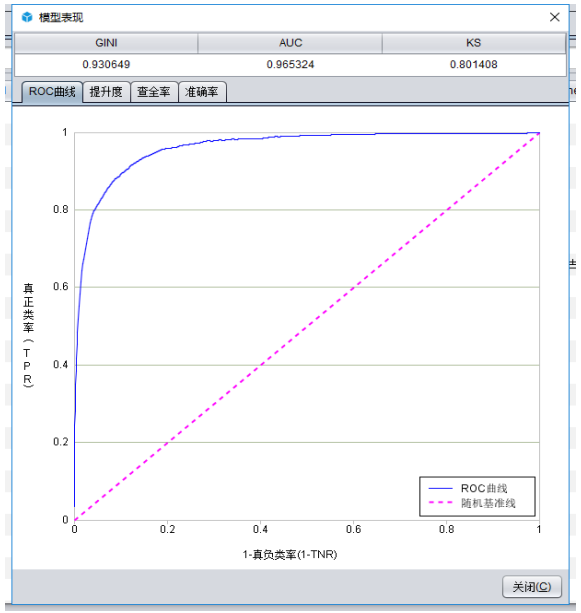
痛点

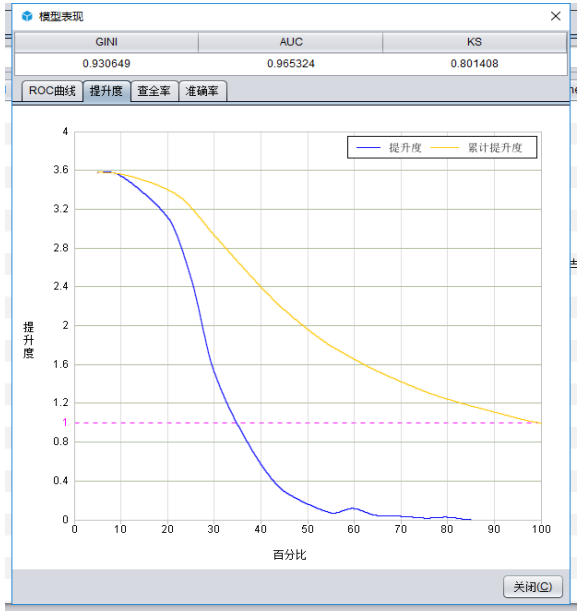
- 找不到合理的数据维度!
- 高基数与非线性问题对模型的影响大!
- 挑选合理的模型或者模型组合难!
- 阳性样本少，容易模型过拟合!

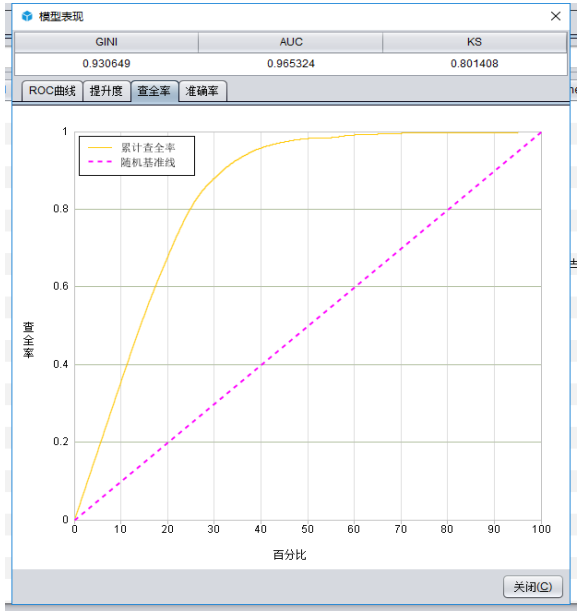
建模结果对比

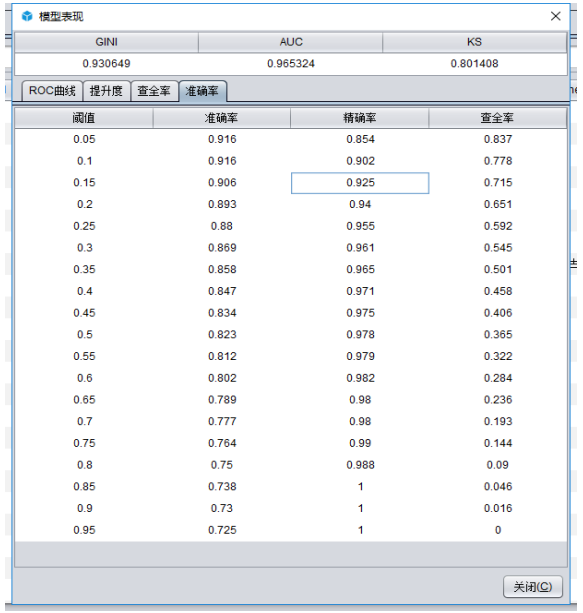
	智能建模	传统建模
人数	1	1
建模时间	**5分钟**（数据预处理+建模）	2个月
建模数量	1	1
数据规模	100000+ / 28MB	100000+/ 28MB
模型表现	0.9728（测试集0.965）	0.957

模型表现（测试集）









案例9 资金流管理-金融信贷业务

业务需求

数据包含用户基本信息，标的信息，日志信息，还款行为信息等

字段名	类型	说明	示例
user_id	bigint	用户ID	498924
listing_id	bigint	标的ID	1873205
auditing_date	string	标的成交日期	2018-01-01
due_date	string	标的第一期应还款日期	2018-02-01
due_amt	decimal(38,4)	标的第一期应还款金额	226.6526
repay_date	stri	标的第一期实际还款日期：	
repay_amt	dec		

字段名	类型	说明	示例
user_id	bigint	用户ID	498924
listing_id	bigint	标的ID	1873205
auditing_date	string	标的成交日期	2018-01-01
term	int	标的期限	9
			7.2
			1980.0000

字段名	类型	说明	示例
user_id	bigint	用户ID	498924
listing_id	bigint	标的ID	1529307
order_id	int	标的应还款期数序号	1
due_date	string	标的本期应还款日期	2017-12-22
due_amt	decimal(38,4)	标的本期应还款金额	464.3723
repay_date	string	标的本期实际还款日期： (1) 若还款时本期未逾期，值为还款日期 (2) 若还款时本期已逾期，值为"2200-01-01"	2017-12-15
repay_amt	decimal(38,4)	标的本期实际还款金额	464.3723

说明	示例
)	498924
像标签列表：各标签以"|"分隔	3161|676|244|... ...
入日期	2017-12-31

字段名			
user_id			
reg_mon			
gender			
age			
cell_province	string	用户手机号归属省份	c06
id_province	string	用户身份证归属省份	c06
id_city	string	用户身份证归属城市：前三位为省ID	c06266
insertdate	string	数据插入日期	2017-12-31

说明	示例
用户ID	498924
用户行为发生时间	2017-11-07 19:45:43
用户行为类别编号	1

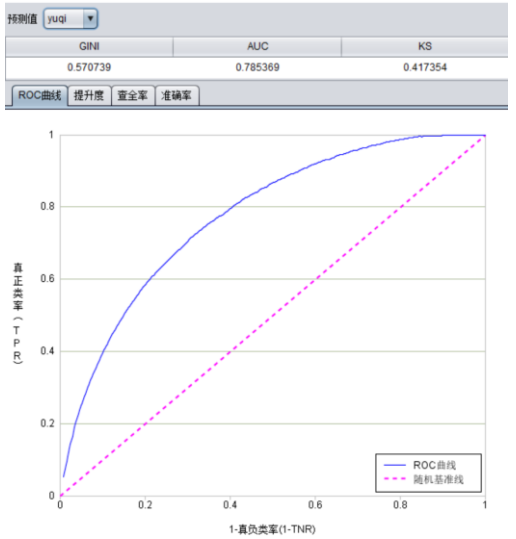
案例9 资金流管理-金融信贷业务

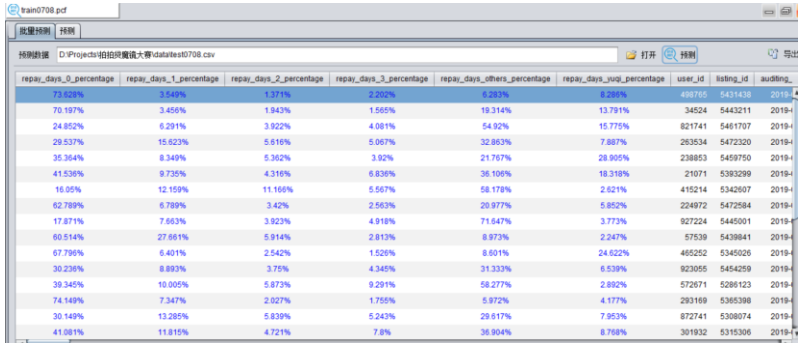
易明 YIMING

实施步骤和使用工具：

- 整理宽表，提取特征，定义目标变量 —— **集算器**
- 将业务目标转化为多分类问题，预测出用户每天还款的概率以及逾期的概率 —— **智能建模工具**
- 用户还款概率*应还金额，得出用户每天的还款金额，所有用户还款金额的总和即为当天的资金回款金额
- 使用2018年全年第一期的数据作为训练集，建模预测2019年第一季度的资金回款情况

模型效果	
人数	1
项目时间	5天
Rmse	6797
模型表现	Auc: 0.69~0.81
建模数据量	100万
模型排名	72/1215 前6%



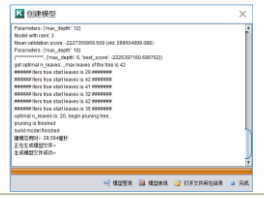
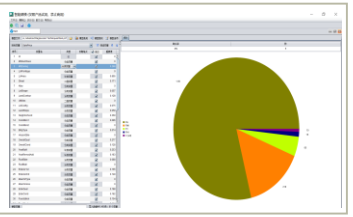
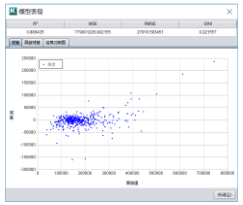
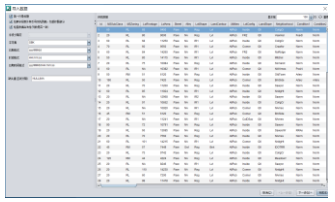


易明智能建模

产品优势

易明智能建模优势

易明 YIMING

人工建模弱点		易明智能建模产品优势	
建模周期长	一般建模项目，占整个计划20%的时间。	**建模效率高**	可以缩短大约80%的建模时间，使整个计划的时程缩短16% 
建模人员要求高	只有专业的建模人员，才能够建立良好的模型。由于专业人员的供给不足，价格极高，限制了模型的应用范围。	**使用门槛低**	无需专业的建模人员，公司内部的工作人员包括行销，业务，及行政人员，均可使用，大大拓展了模型的应用范围。 
模型质量不稳定	模型的品质，由建模人员的能力决定，参差不齐的个人建模能力，决定了模型质量良莠不齐。	**模型质量稳定**	智能建模，不受建模人员能力影响模型质量稳定。 
模型更新困难	模型成本很高，所以不能经常更新，经常过度使用，在模型的有效期末端，模型的精确度显著降低。	**模型更新方便**	模型可以随时更新，永远处于模型有效期的前端。 

易明智能建模产品特点

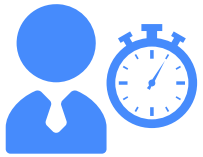
易明 YIMING

智能化

大幅减少人工工作量

效率高

有效提升建模速度，缩短模型生产周期



运行稳定

经过大量的业务场景验证，易明产品的模型质量精度高，泛化能力强，运行稳定

成本低

为企业节省人力成本和时间成本

提升数据驱动业务决策精准度及效率

易明

挖掘数据价值

THANKS