阿里云天池牛年读书会

智能风控：
Python金融风险管理与评分卡建模

分享嘉宾：黄哲
个人简介：数据科学家（风控/反洗钱）

天池读书会

TIANCHI天池

机械工业出版社 华章公司

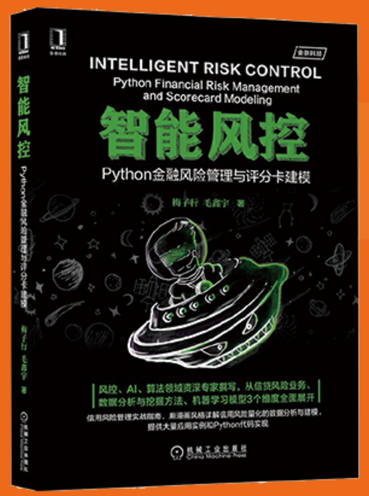
《智能风控：Python金融风险管理与评分卡建模》

基于Python讲解了信用风险管理和评分卡建模，

详细讲解了信用风险量化相关的数据分析与建模手段，并提供大量的应用实例。

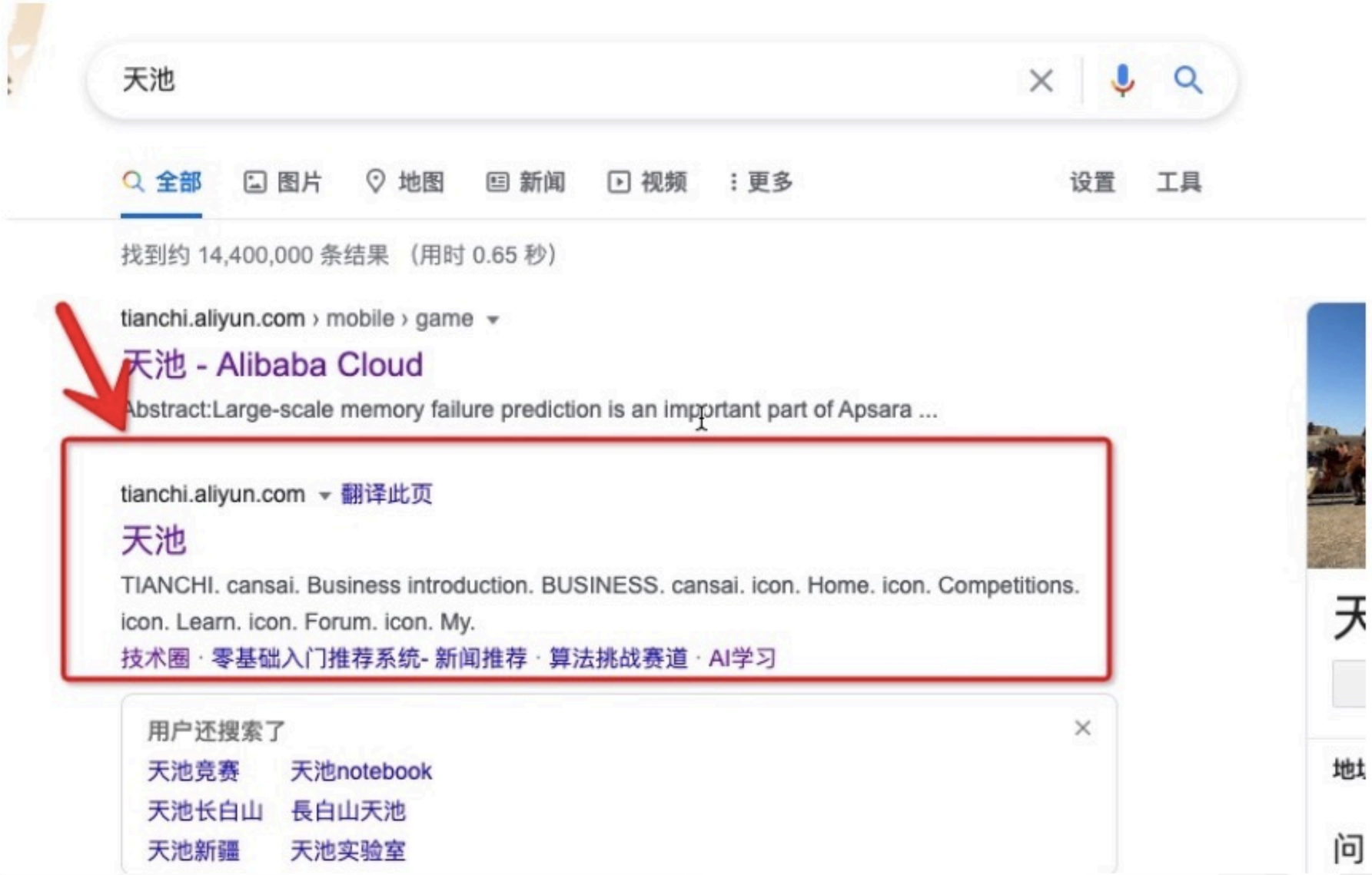
直播嘉宾：黄哲

直播时间：4月27日20:00



扫码领取读书会配套学习资源

1）首先需要进入天池官网，大家打开浏览器，搜索 天池 ，找到 tianchi.aliyun.com即可访问进入天池官网；



2）在天池官网，将鼠标移到 天池学习 ，即可出现下拉列表，点击 天池读书会 ，即可进入天池读书会的页面。

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

3）在天池读书会页面，你可以对对应的读书会图书进行提问，优秀的提问还有机会获得赠书，还可以点击配套的训练营或者课程资源进入学习，还有点击实践代码获取读书会的项目实践的代码，跟着我一起进行项目实践和代码学习，同时还有很多其他的读书会，大家也可以观看举办过的读书会的回放，或者预约还没开始的读书会。

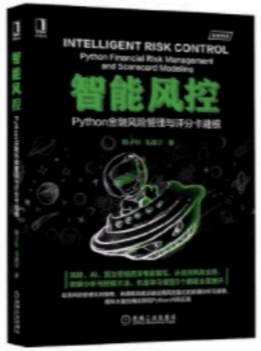
黄哲 南洋理工大学硕士、数据科学家

直播主题 《智能风控：Python金融风险管理与评分卡建模》

直播时间 2021年4月27日 20:00

学习资料 金融风控训练营

实践项目 风控评分卡全流程建模



提问 | 学习训练营 | 购买地址 | PPT下载 | 实践代码 | 预约直播

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

目录

TIANCHI天池

分享嘉宾简介

TIANCHI天池

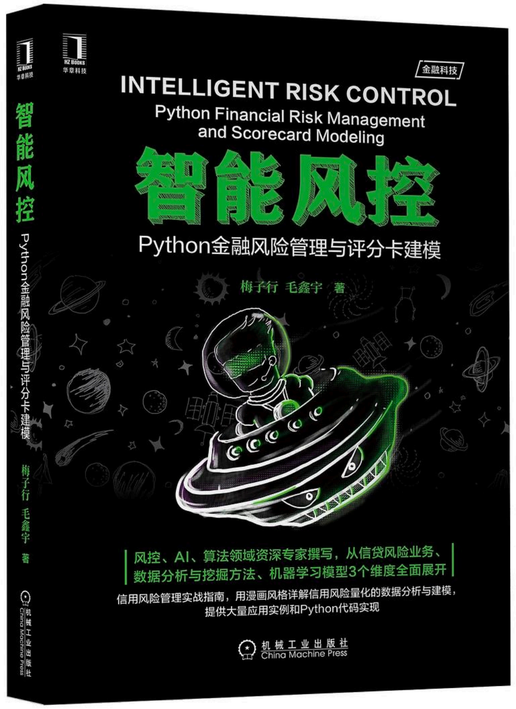
黄哲

南洋理工大学硕士

数据科学家 专注领域：智能风控，反洗钱

为多家跨国银行开发/部署了风控和反洗钱方案

图书简介



本次分享并不能覆盖每个细节，我会从风控从业者角度来挑选一些重点，以便初学者有的放矢

- 基于python讲解了信用风险管理和评分卡建模
- 从风险业务，统计方法分析，机器学习三个维度讲
- 为风控领域初学者提供了信贷业务的全局视图

书籍勘误：https://zhuanlan.zhihu.com/p/139907399

图书简介

TIANCHI天池

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

1/9 信用管理基础

重点：

- 术语解释
- 信贷风控架构

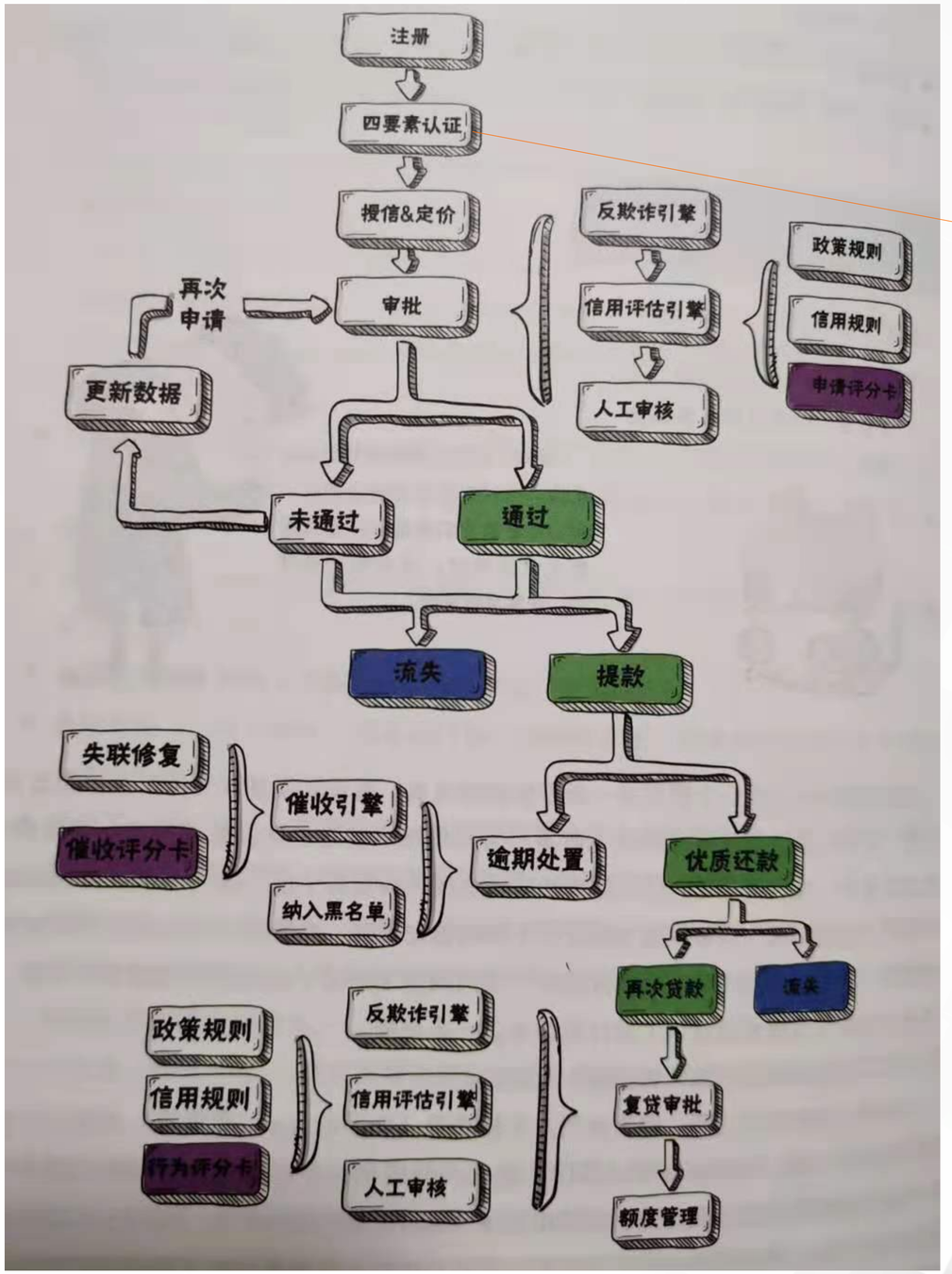
图书内容

常用术语：

- 账龄(MOB)
- 逾期天数（DPD）
- 不良率（Bad Rate）
- 负债比（Debt Burden Ratio, DBR）
- 迁徙率（Flow Rate）
- ……

图书内容

信贷风控架构



四要素：

- 身份证
- 姓名
- 手机号
- 银行卡号

2/9 评分卡

评分卡——重点：

- 评分卡概念，类型
- 建模流程

图书内容

TIANCHI天池

评分卡长啥样？

Characteristic	Attribute	Scorecard Points
AGE	<22	100
AGE	22<=AGE<26	120
AGE	26<=AGE<29	185
AGE	29<=AGE<32	200
AGE	32<=AGE<37	210
AGE	37<=AGE<42	225
AGE	>=42	250
HOME	OWN	225
HOME	RENT	110
INCOME	<10000	120
INCOME	10000<=INCOME<17000	140
INCOME	17000<=INCOME<26000	180
INCOME	26000<=INCOME<35000	200
INCOME	35000<=INCOME<42000	225
INCOME	42000<=INCOME<58000	230
INCOME	>=58000	260

Let cutoff=600

So, a new customer applies for credit.....

AGE	35	210 points
INCOME	$38K	225 points
HOME	OWN	225 points
Total		660 points
Decision:		GRANT CREDIT

Note: A scorecard is scaled with the **Odds**, **Scorecard Points** and **Points to Double the Odds** properties.

图书内容

评分卡——重点：

- 评分卡概念，类型
- 建模流程

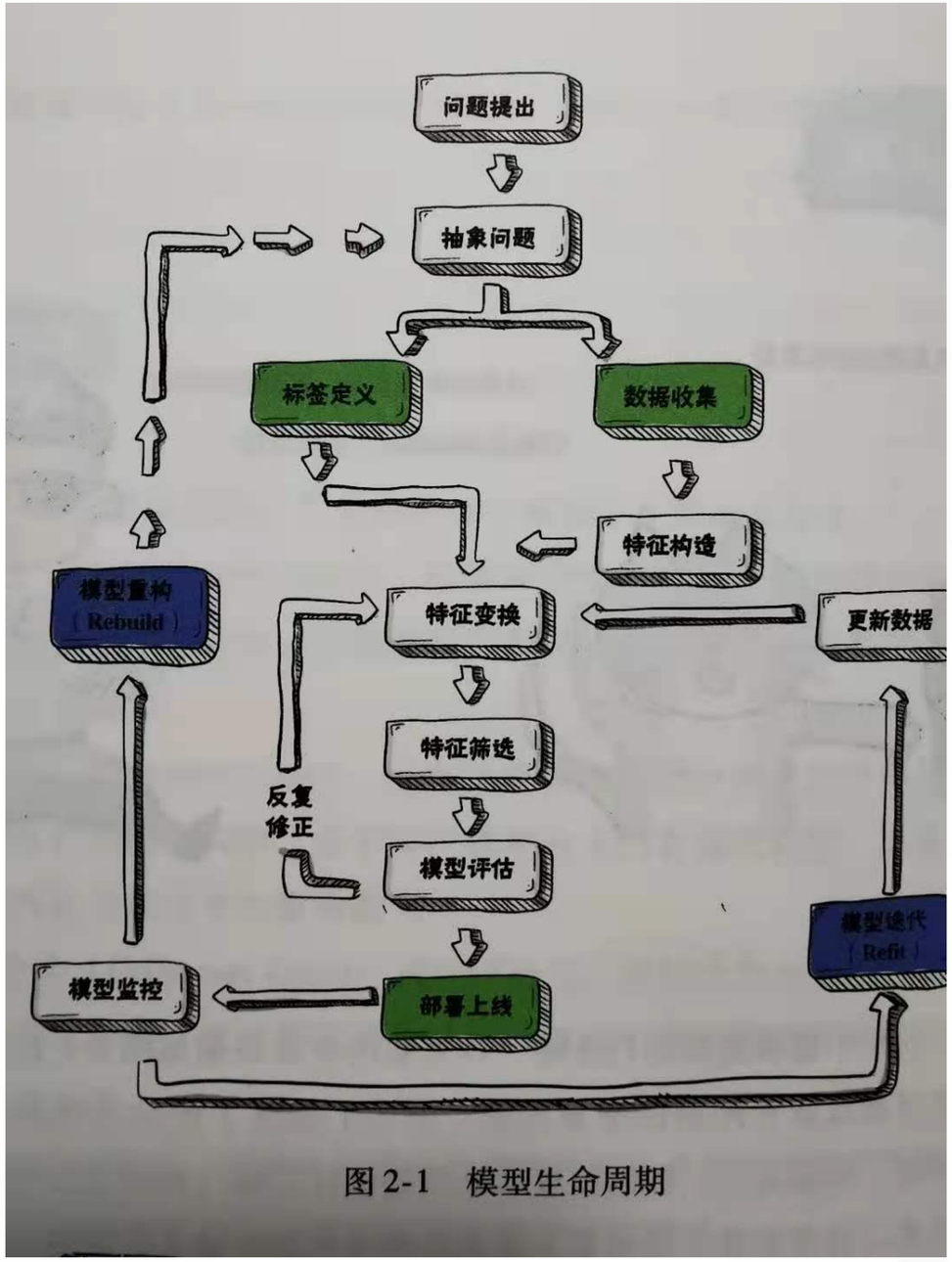


图 2-1　模型生命周期

TIANCHI天池

3/9 机器学习

机器学习——重点：

- 性能度量：AUC和KS曲线
- 业务评价：评分模型报告
- 模型解释性

朱明超 本书译者、复旦研究生

直播主题 《可解释机器学习：黑盒模型可解释性理解指南》

直播时间 2021年4月21日 20:00

学习资料 机器学习训练营

实践项目 对宫颈癌数据集的分类和解释

提问 | 学习训练营 | 购买地址 | PPT下载 | 实战代码 | 观看回放

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

图书内容

KS是什么？

KS=max(TPR-FPR)

代表了模型对正负样本的**区分能力**

KS曲线 = TPR曲线 - FPR曲线

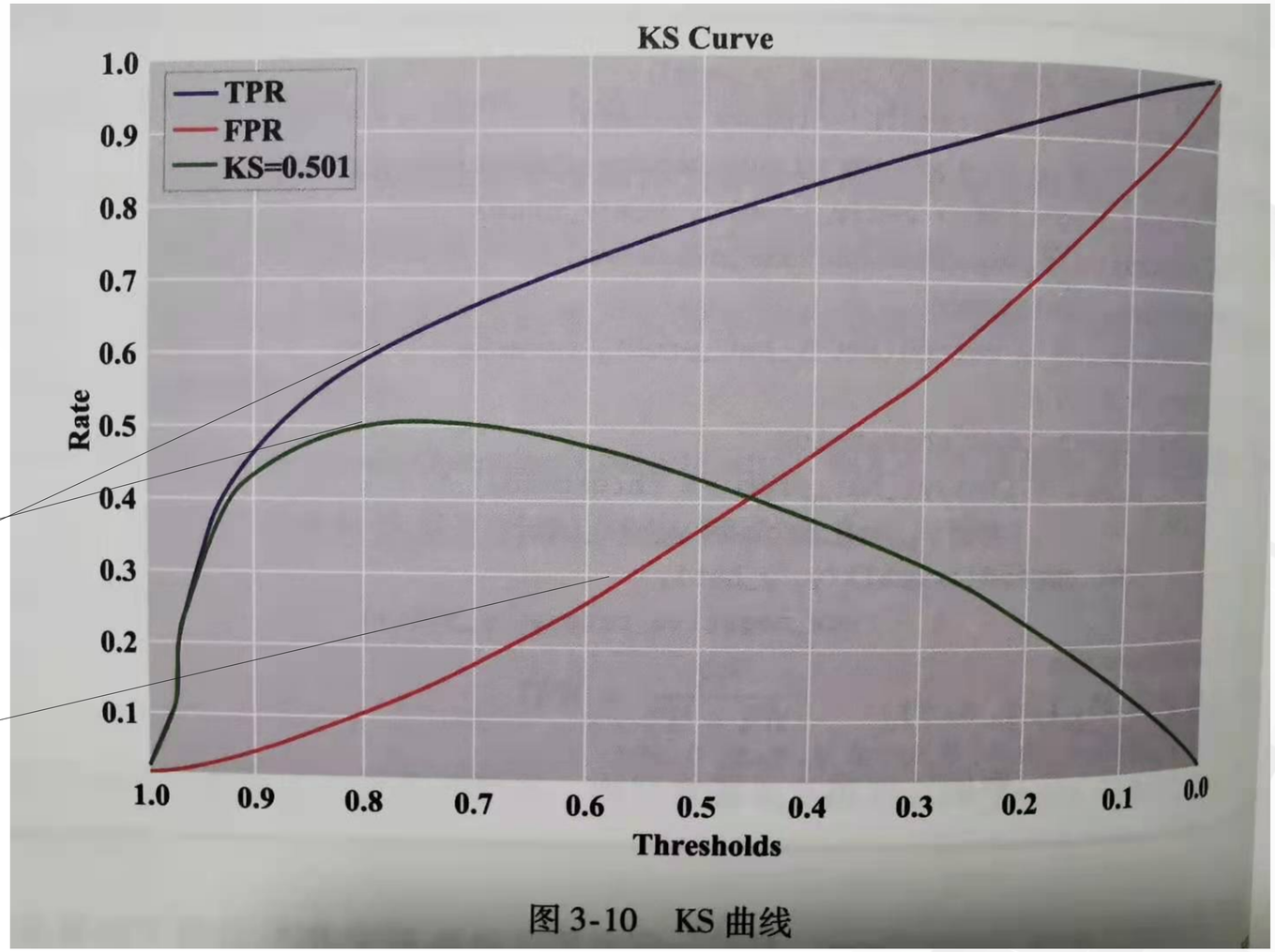


图 3-10 KS 曲线

图书内容

(什么是分箱？ -> p23)

业务评价：评分模型报告

等频分箱

	KS	负	个数	正样本个数	负	累计个数	正样本累计个数	捕获率	负样本占比
0	0.217		86	713		86	713	0.262	0.108
1	0.299		43	756		129	1469	0.393	0.054
2	0.339		29	770		158	2239	0.482	0.036
3	0.381		30	769		188	3008	0.573	0.038
4	0.398		22	777		210	3785	0.640	0.028
5	0.403		18	781		228	4566	0.695	0.023
6	0.408		18	781		246	5347	0.750	0.023
7	0.398		13	786		259	6133	0.790	0.016
8	0.396		16	783		275	6916	0.838	0.020
9	0.361		5	794		280	7710	0.854	0.006
10	0.332		7	792		287	8502	0.875	0.009
11	0.287		2	797		289	9299	0.881	0.003
12	0.258		7	792		296	10091	0.902	0.009
13	0.225		6	793		302	10884	0.921	0.008
14	0.208		11	788		313	11672	0.954	0.014
15	0.182		8	791		321	12463	0.979	0.010
16	0.137		2	797		323	13260	0.985	0.003
17	0.092		2	797		325	14057	0.991	0.003
18	0.045		1	798		326	14855	0.994	0.001
19	0.000		2	792		328	15647	1.000	0.003

图 3-13 评分模型报告

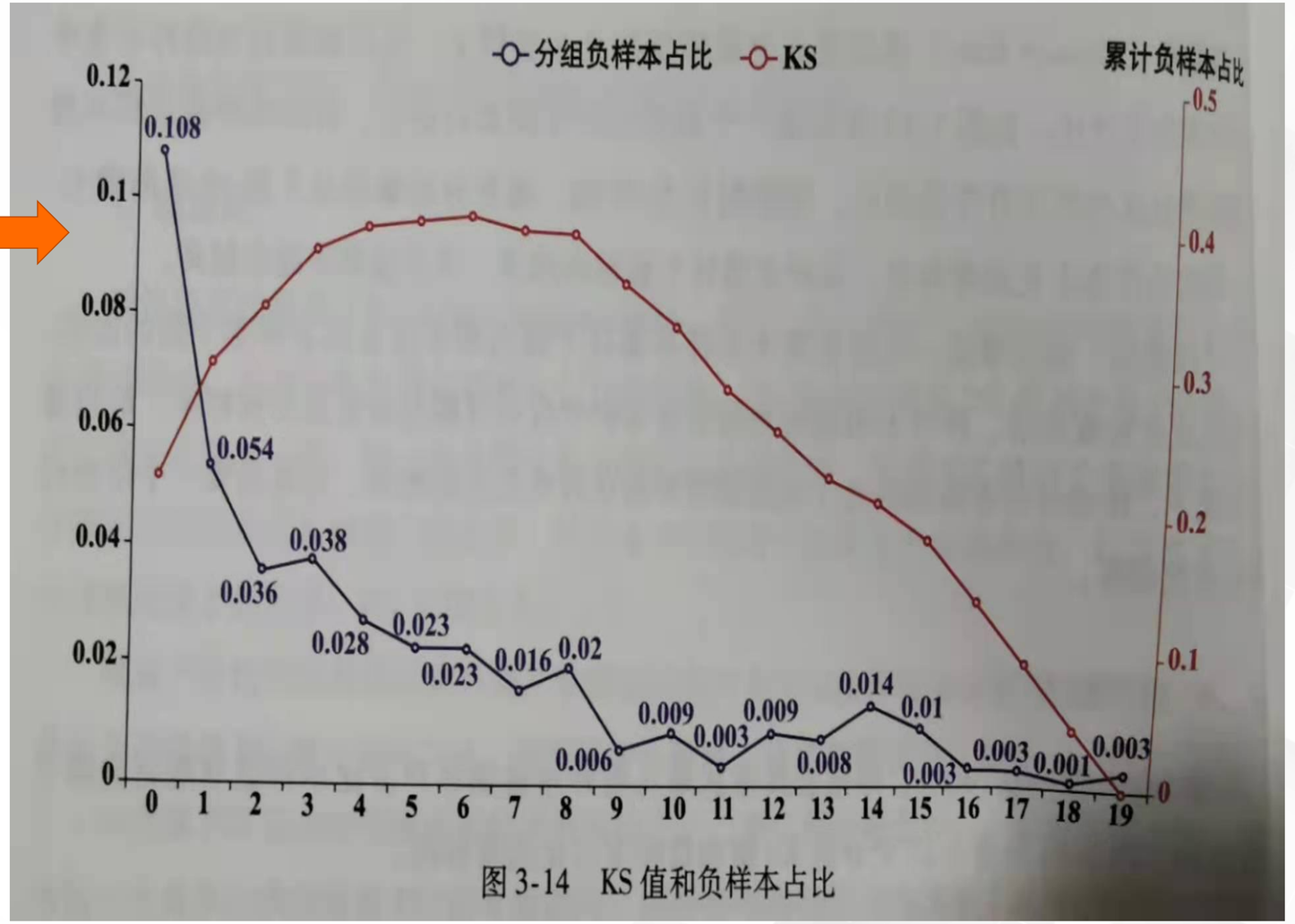


图 3-14 KS 值和负样本占比

直播相关资料获取及回放查看地址： https://tianchi.aliyun.com/specials/promotion/activity/bookclub

4/9 客户分群

重点

- 为什么要分群
- 分群方法（经验分群，技术分群）

图书内容

为什么要分群：

不同群体的趋势可能不一样

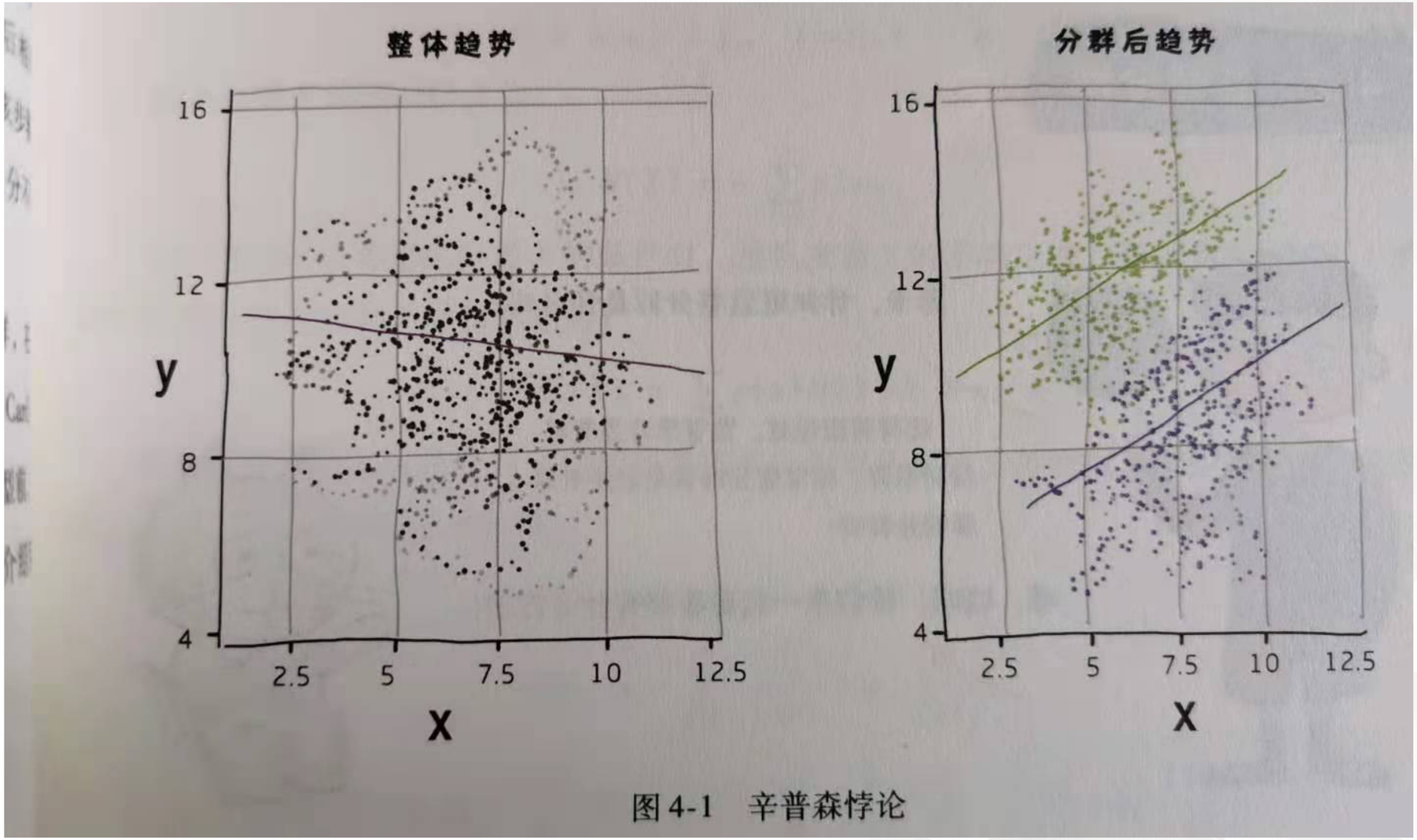


图 4-1　辛普森悖论

图书内容

经验分群例子：

- 不同国家一个群
- 根据职业分群
- 根据借贷场景分群

图书内容

技术分群：
有监督：例如决策树分群

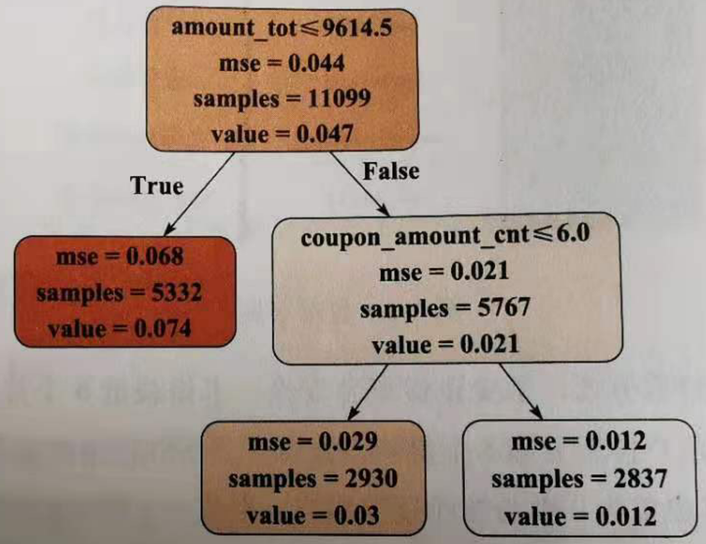


图 4-3　决策树运行结果

CART 回归树的节点预测属性 mse 表示当前子群中目标变量的均值。而当前案例中，目标变量的均值等价于标签为 1 的样本占当前子群样本的比例。从图 4-3 中可以看出，决策树将原始样本群划分为 3 个子群，其负样本占比依次为 0.074、0.03、0.012。

无监督：例如GMM分群

```
2. from sklearn.mixture import GaussianMixture as GMM
3. gmm = GMM(n_components=3, covariance_type='full').fit(x) # 指定聚类中心个数为3
4. labels = gmm.predict(x)
5. plt.scatter(x['coupon_amount_cnt'], x['amount_tot'], c=labels, s=5, cmap='viridis')
```

运行结果如图 4-7 所示。

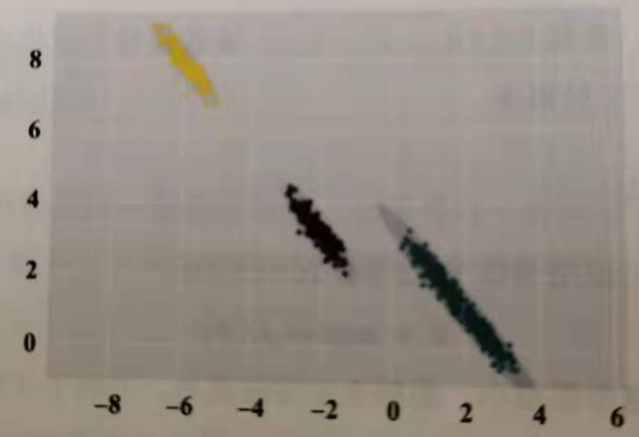


图 4-7　聚类结果

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

5/9 数据探索与特征工程

重点

最常用分箱：等频分箱，等距分箱

其他方法：卡方分箱，聚类分箱，决策树分箱

WOE（Weight of Evidence）：把非线性特征转成线性，对逻辑回归模型有用

图书内容

TIANCHI天池

等频分箱vs等距分箱

Pandas里：

pd.cut: 按照定义的区间来分箱（可等距，可不等距）

Pd.qcut:等频分箱，每箱的数量相同（偶尔会有一点不同，可能是除不尽，也可能是重复值多）

```
age_bins = [-math.inf, 25, 40, 50, 60, 70, math.inf]
df_train['bin_age'] = pd.cut(df_train['age'],bins=age_bins).astype(str)
dependent_bin = [-math.inf,2,4,6,8,10,math.inf]
df_train['bin_NumberOfDependents'] = pd.cut(df_train['NumberOfDependents'],bins=dependent_bin).astype(str)
dpd_bins = [-math.inf,1,2,3,4,5,6,7,8,9,math.inf]
df_train['bin_NumberOfTimes90DaysLate'] = pd.cut(df_train['NumberOfTimes90DaysLate'],bins=dpd_bins)
df_train['bin_NumberOfTime30-59DaysPastDueNotWorse'] = pd.cut(df_train['NumberOfTime30-59DaysPastDueNotWorse'], bins=dpd_bins)
df_train['bin_NumberOfTime60-89DaysPastDueNotWorse'] = pd.cut(df_train['NumberOfTime60-89DaysPastDueNotWorse'], bins=dpd_bins)


df_train['bin_RevolvingUtilizationOfUnsecuredLines'] = pd.qcut(df_train['RevolvingUtilizationOfUnsecuredLines'],q=5,duplicates='drop').astype(str)
df_train['bin_DebtRatio'] = pd.qcut(df_train['DebtRatio'],q=5,duplicates='drop').astype(str)
df_train['bin_MonthlyIncome'] = pd.qcut(df_train['MonthlyIncome'],q=5,duplicates='drop').astype(str)
df_train['bin_NumberOfOpenCreditLinesAndLoans'] = pd.qcut(df_train['NumberOfOpenCreditLinesAndLoans'],q=5,duplicates='drop').astype(str)
df_train['bin_NumberRealEstateLoansOrLines'] = pd.qcut(df_train['NumberRealEstateLoansOrLines'],q=5,duplicates='drop').astype(str)
```

图书内容

TIANCHI天池

卡方分箱

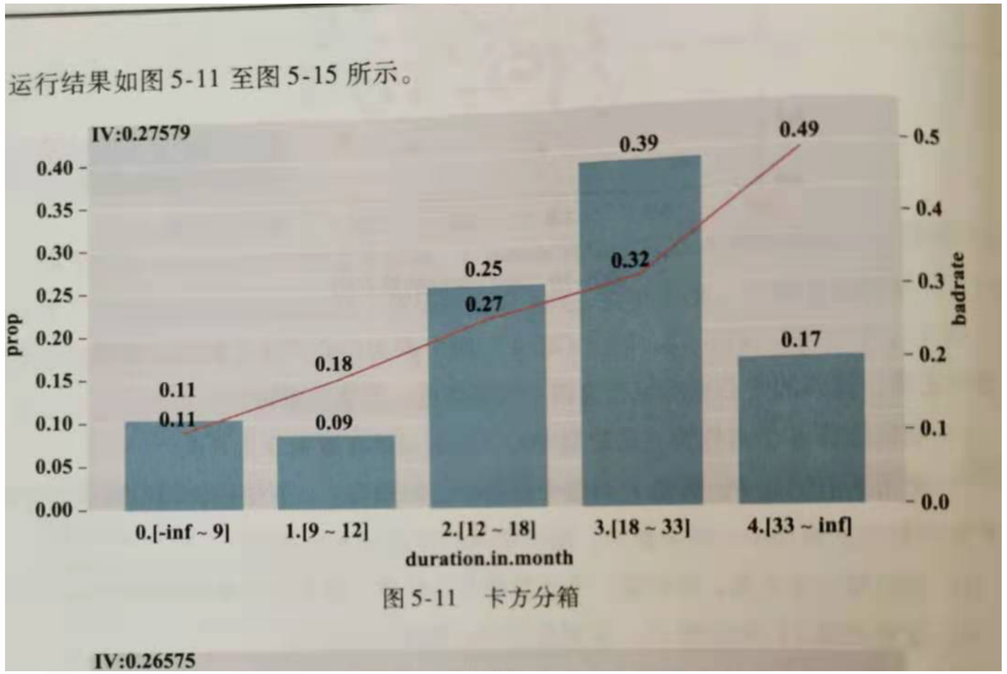
使用卡方检验确定最优分箱阈值

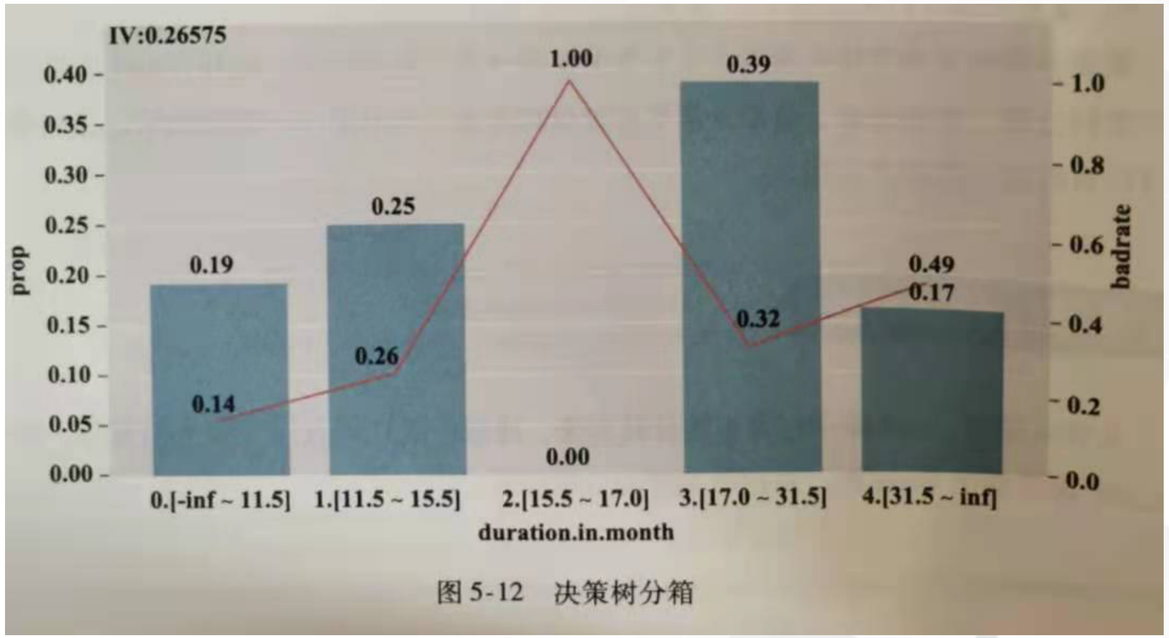
聚类分箱

聚类即分箱

决策树分箱

根据决策树阈值分箱



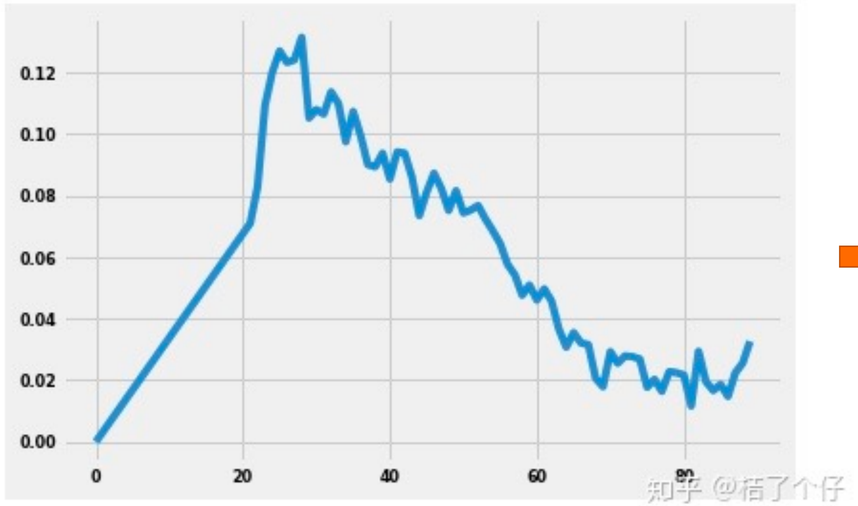


注：新手不推荐直接研究这块，建议对业务理解深入再研究技术分箱方法

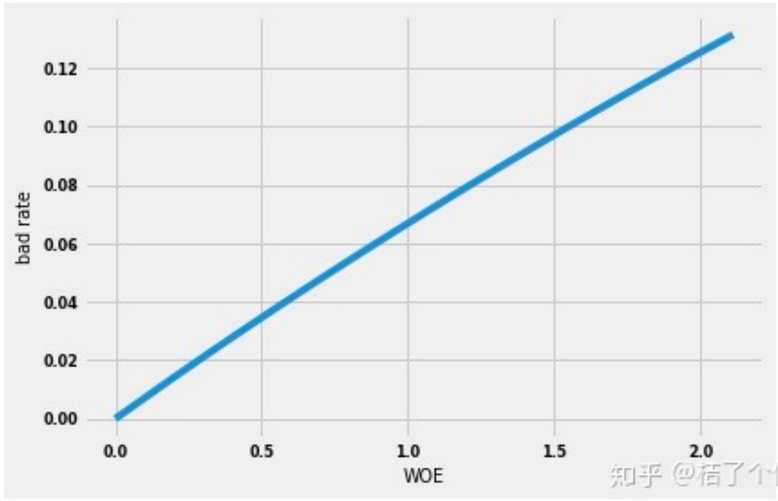
直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

WOE（Weight of Evidence），证据权重

$$WOE_i = ln(\frac{Bad_i}{Bad_T} / \frac{Good_i}{Good_T}) = ln(\frac{Bad_i}{Bad_T}) - ln(\frac{Good_i}{Good_T})$$



非线性特征
转成
线性特征



扩展阅读《 WOE编码为啥有效》： https://zhuanlan.zhihu.com/p/146476834

6/9 特征筛选与建模

重点：特征筛选

- 缺失率
- 信息量（IV）
- 相关性

图书内容

TIANCHI天池

缺失率

Null值超过70%：删除

Null值50%~70%：将是否为null作为特征

Null值50%以下：fillna，可用均值，中值等

信息量（IV）

什么是WOE？->p26

$$IV_i = (\frac{Bad_i}{Bad_T} - \frac{Good_i}{Good_T}) * WOE_i$$

$$= (\frac{Bad_i}{Bad_T} - \frac{Good_i}{Good_T}) * ln(\frac{Bad_i}{Bad_T} / \frac{Good_i}{Good_T})$$

$$IV = \sum_{i-1}^{n} IV_i$$

Information Value	Predictive Power
< 0.02	useless for prediction
0.02 to 0.1	Weak predictor
0.1 to 0.3	Medium predictor
0.3 to 0.5	Strong predictor
>0.5	Suspicious or too good to be true

相关性

可任意一种：

- 皮尔逊相关系数
- 斯皮曼相关系数
- 肯德尔相关系数

7/9 拒绝推断

重点：
什么是数据验证？目的？
什么是拒绝推断？常用方法？

图书内容

TIANCHI天池

数据验证：又称下探，即从拒绝样本中选取部分样本进行放款

目的：获得真实标签，进入评分卡模型学习

代价：收益损失

图书内容

拒绝推断：

通过数据分析的方法来修正模型的参数估计偏差

方法：

硬截断法（最常用）

模糊展开法

重新加权法

外推法

迭代再分类法

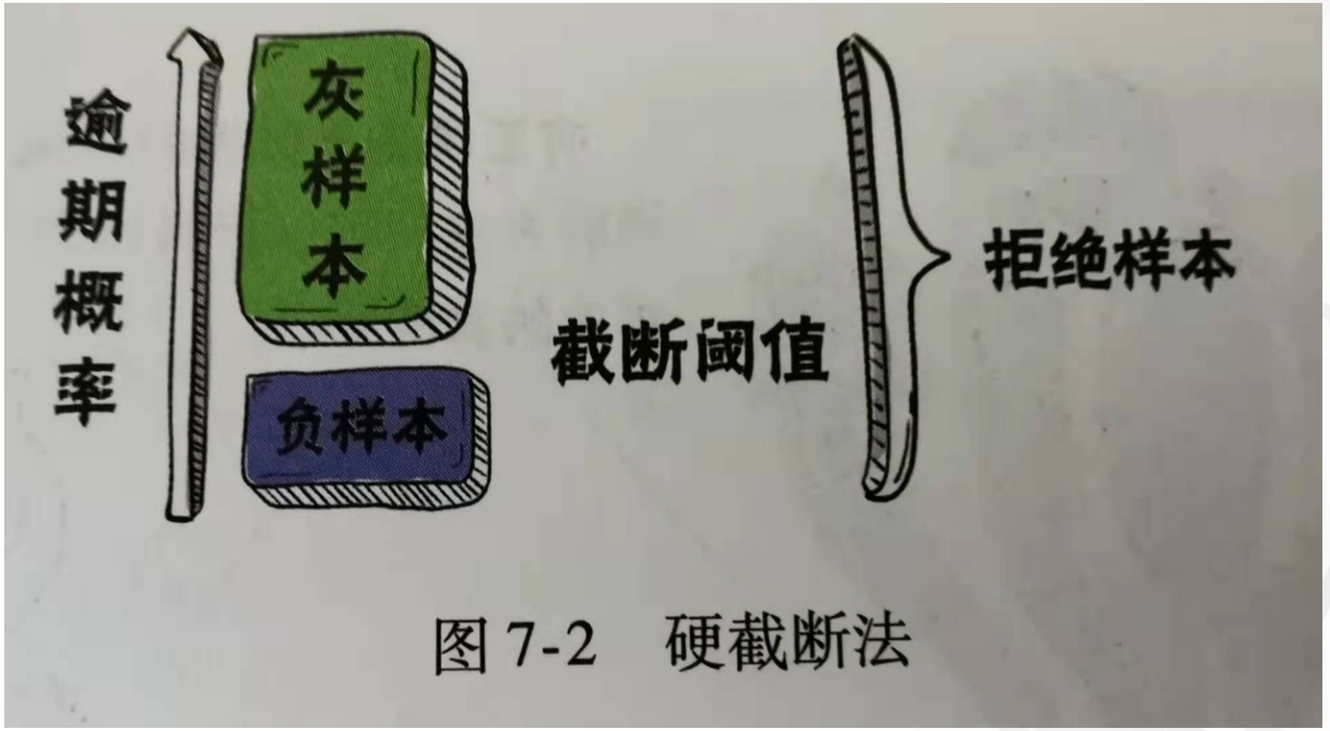


图 7-2　硬截断法

TIANCHI天池

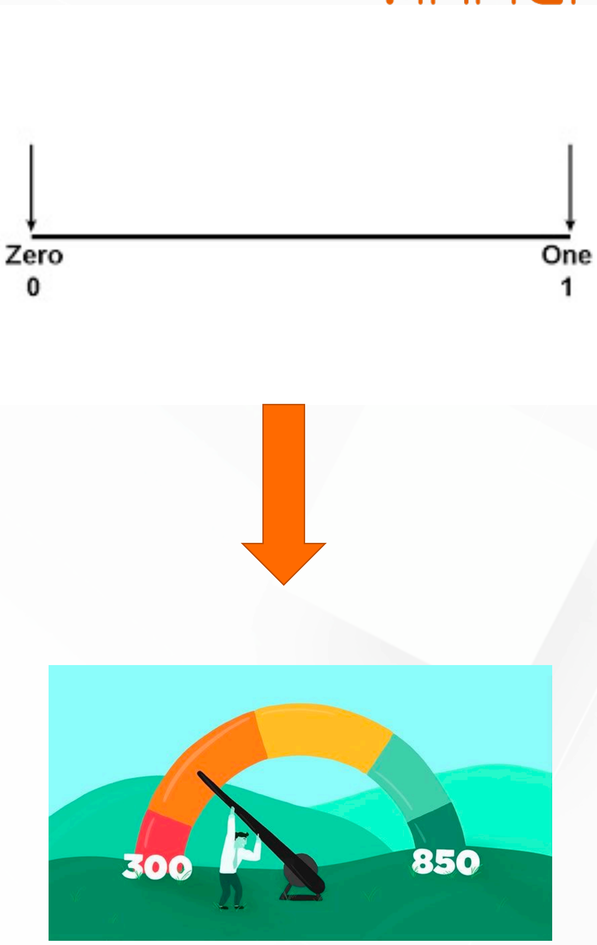
8/9 模型校准

重点：
概念？
常用方法？

图书内容

模型校准：

逻辑回归模型的输出为0~1之间的概率值，在评分卡建模流程中，通常会把逻辑回归模型输出的概率分数转成整数分数，这个过程称为评分卡分数校准



图书内容

TIANCHI天池

多模型校准：更好利用每个模型，防止单个数据源/模型出错带来的影响

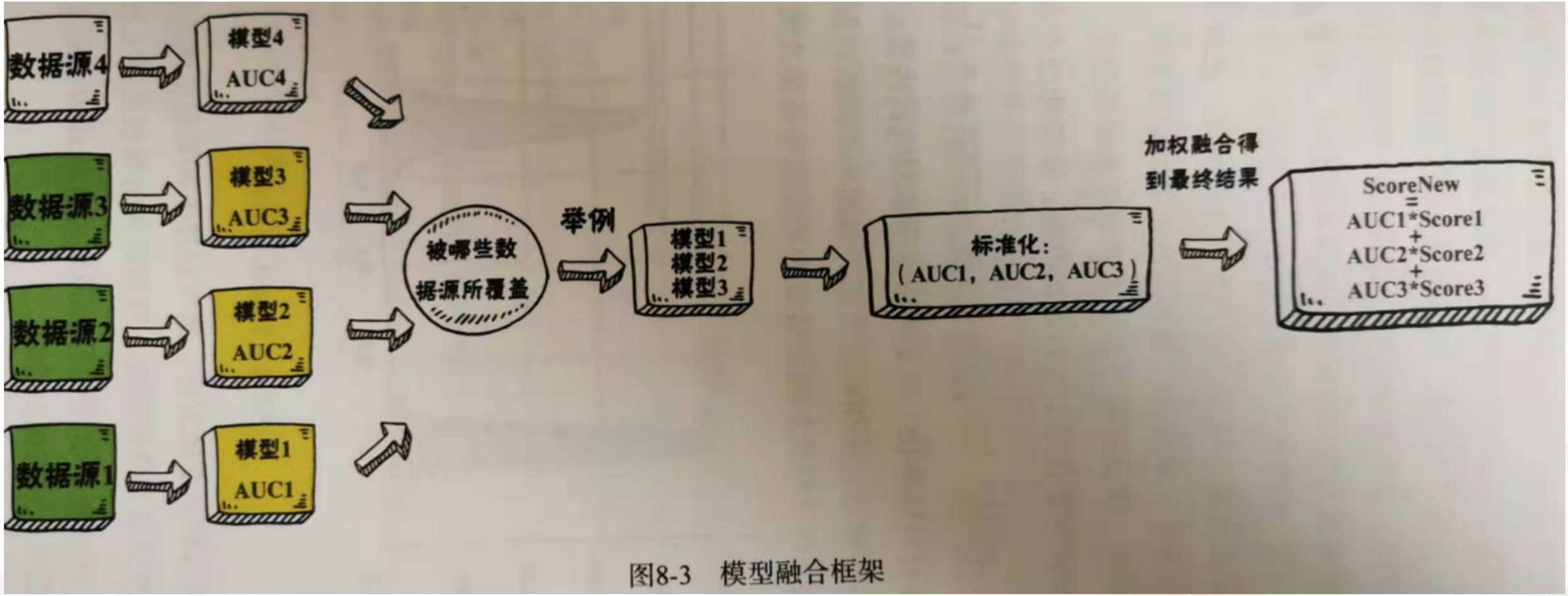


图8-3　模型融合框架

9/9 模型文档

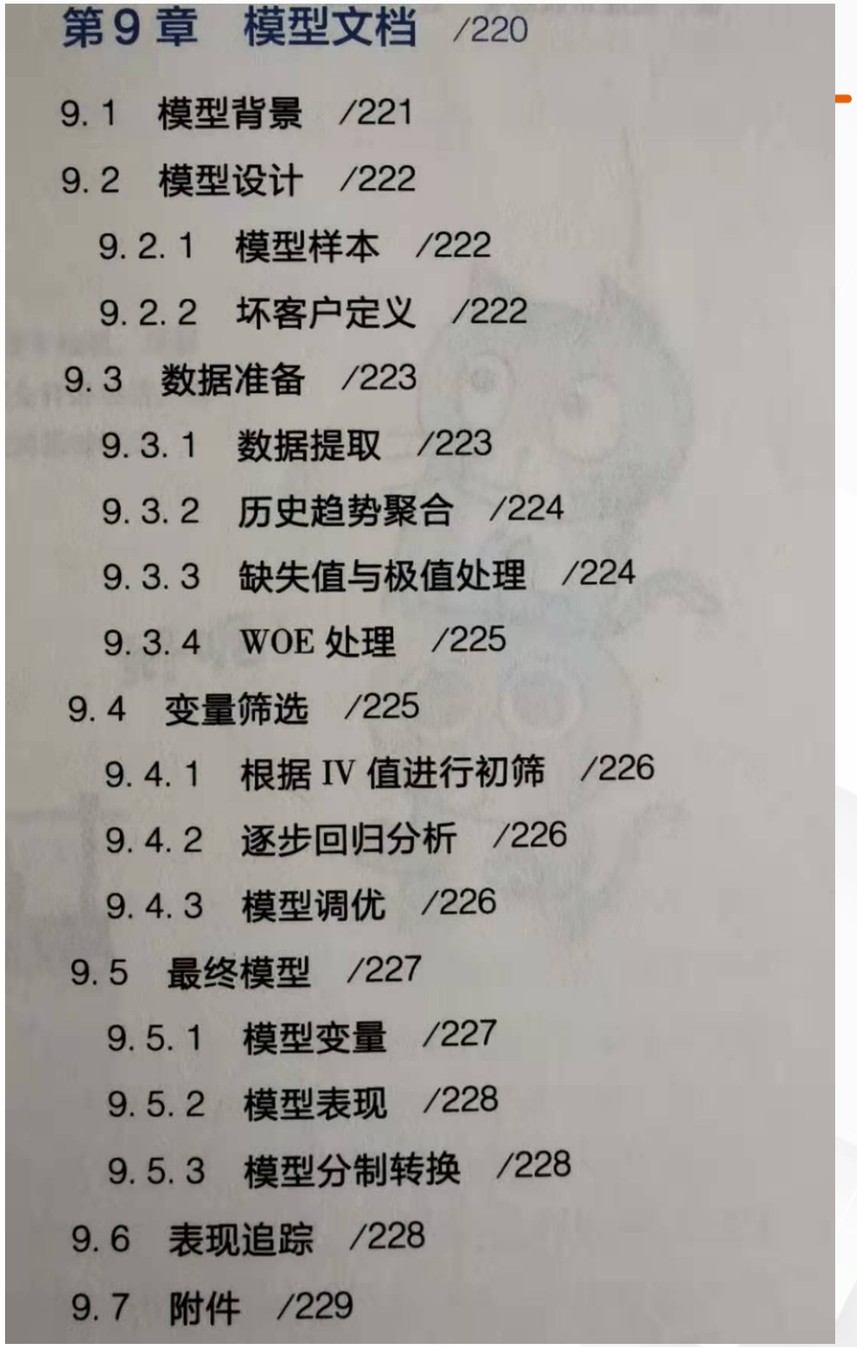
重点：
需要记录的内容

图书内容

模型文档的目的？

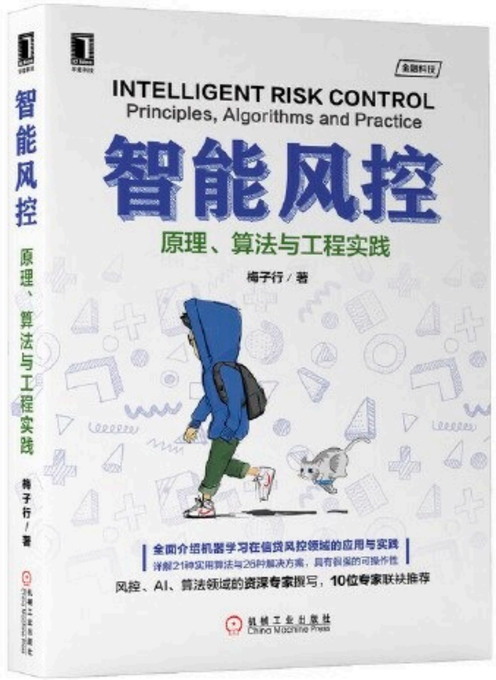
- 有效发现问题
- 方便工作交接

模型文档都要记录哪些内容？



图书特色

- 偏重理论和业务
- 主要是围绕评分卡



作者梅子行另一本书，更偏重技术和实践，4月28号陈旸博士将会分享

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

学习方法

TIANCHI天池

- 建立自己的练手项目（稍后可以fork我的代码作为练手的baseline）
- 把书中提到的方法应用到项目里
- 观察代码输出是否符合预期。若是模型，则对比效果
- 使用toad库。https://pypi.org/project/toad/

补充阅读：我写的风控技术系列文章
https://zhuanlan.zhihu.com/p/144732622

实战演示、互动交流

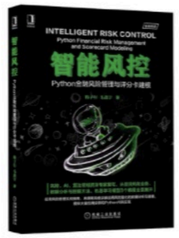
TIANCHI天池

大家可以使用手机扫左侧二维码，或者**电脑**访问下方地址进入天池读书会页面，点击今天读书会中的 **实践代码** 和我一起进行项目实践学习，天池为大家准备好了代码和运行环境，非常方便。

https://tianchi.aliyun.com/specials/promotion/activity/bookclub

黄哲 南洋理工大学硕士、数据科学家

直播主题 《智能风控：Python金融风险管理与评分卡建模》
直播时间 2021年4月27日 20:00
学习资料 金融风控训练营
实践项目 风控评分卡全流程建模



提问 | 学习训练营 | 购买地址 | PPT下载 | 实践代码 | 预约直播

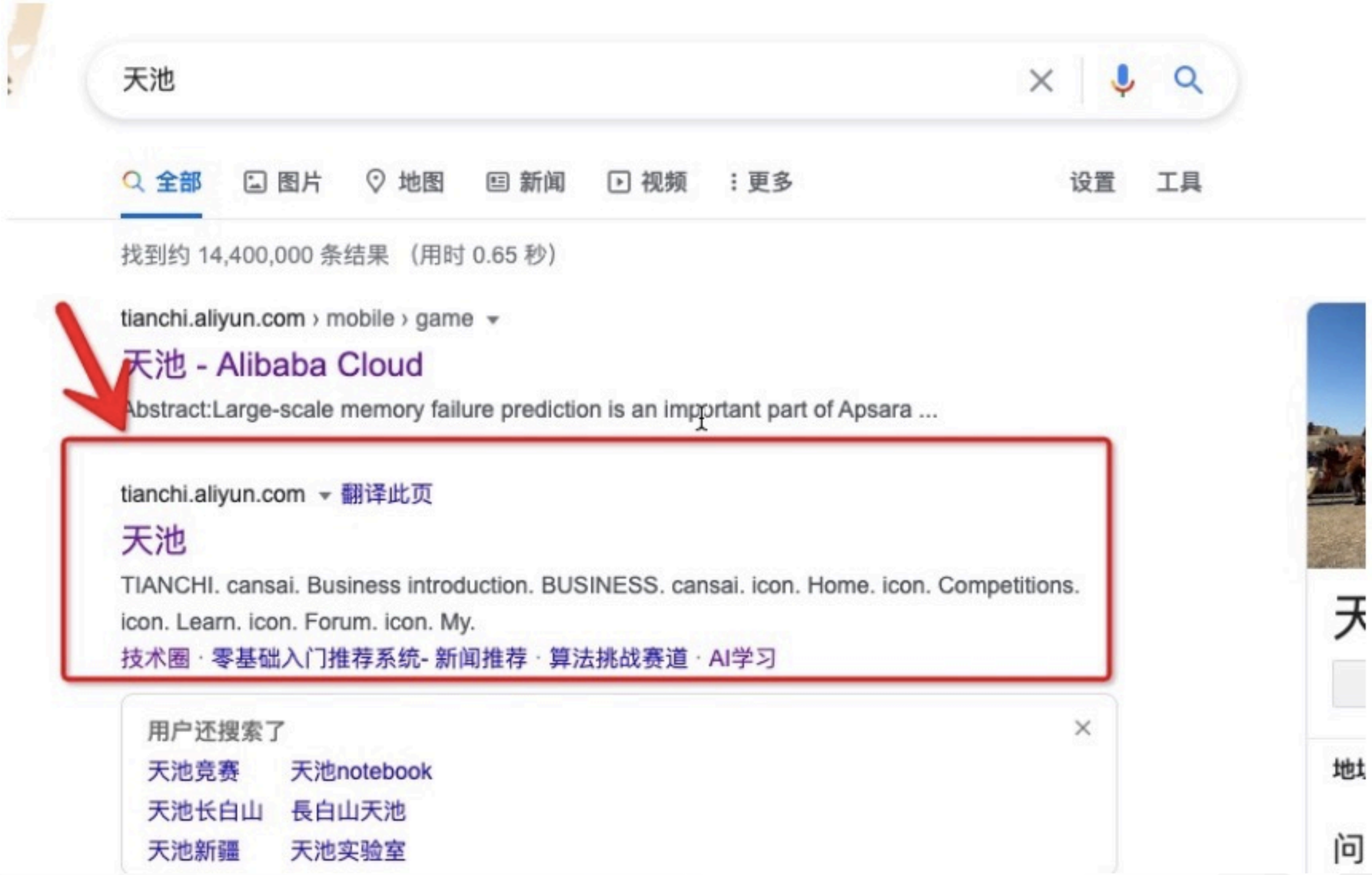
直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

TIANCHI天池

Q&A

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

1）首先需要进入天池官网，大家打开浏览器，搜索 天池 ，找到 tianchi.aliyun.com即可访问进入天池官网；



2）在天池官网，将鼠标移到 天池学习 ，即可出现下拉列表，点击 天池读书会 ，即可进入天池读书会的页面。

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

3）在天池读书会页面，你可以对对应的读书会图书进行提问，优秀的提问还有机会获得赠书，还可以点击配套的训练营或者课程资源进入学习，还有点击实践代码获取读书会的项目实践的代码，跟着我一起进行项目实践和代码学习，同时还有很多其他的读书会，大家也可以观看举办过的读书会的回放，或者预约还没开始的读书会。

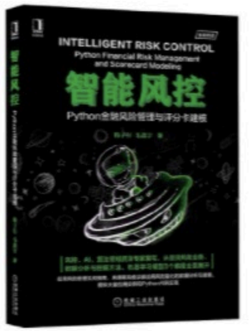
黄哲 南洋理工大学硕士、数据科学家

直播主题 《智能风控：Python金融风险管理与评分卡建模》

直播时间 2021年4月27日 20:00

学习资料 金融风控训练营

实践项目 风控评分卡全流程建模



提问 | 学习训练营 | 购买地址 | PPT下载 | 实践代码 | 预约直播

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub

谢谢观看

TIANCHI天池

直播相关资料获取及回放查看地址：https://tianchi.aliyun.com/specials/promotion/activity/bookclub