大数据分析的一条入门途径

——以拍拍贷风控模型预测为例

范方达

Kesci“魔镜杯”风控算法大赛 涌泉队

2016.4

出发点

- 大数据是更多人可以理解的
- 大数据的方法也是更多人可以学会的
- 大数据没有祖传秘方——不要把曾经初学的我们拦在外面
- 这并不是唯一一个正确答案，而是恩典在面对每一个小小的困难中的累积

目的

- 为数据分析初学者提供一点点数据分析的思路
- 为Python初学者提供一点点Python处理数据的技巧
- 为机器学习过程遇到的难题提供一点点解决方案

大纲

预备

数据读取

数据摘要与清洗

模型选择

模型训练与评估

模型组合与预测

回顾

大纲

预备

数据读取

数据摘要与清洗

模型选择

模型训练与评估

模型组合与预测

回顾

数据与目标

“魔镜杯”风控算法大赛复赛数据

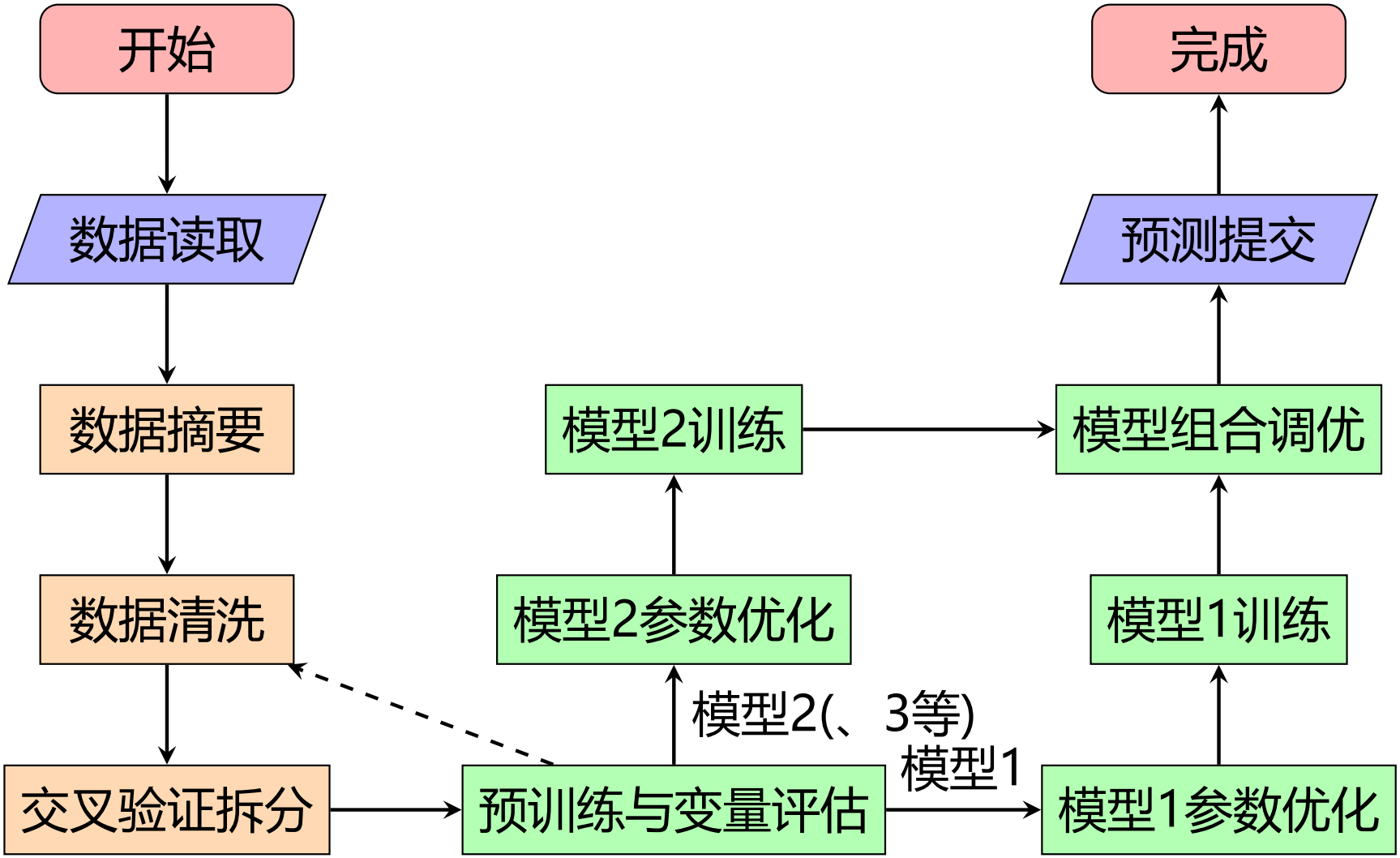
- 样本
  - 训练样本：初赛训练集+初赛预测集+复赛训练集 （8万）
  - 预测样本：复赛预测集 （1万）
- 自变量
  - 主表（226个）
  - 登录信息（4个，但每个index有多条）
  - 用户更新信息（3个，但每个index有多条）
- 预测变量Y: 每个index的6个月内贷款逾期情况（0-1）
- 优化目标：预测变量Y在预测样本的AUC得分

代码平台

Python 3.5

- Packages：
  - 代码笔记本： jupyter
  - 基础: numpy, scipy, pandas, matplotlib, time, re
  - 模型: sklearn, xgboost, keras (theano), hyperopt
- Windows下建议Anaconda，包含科学计算的众多常用包

流程预览


开始
数据读取
数据摘要
数据清洗
交叉验证拆分
预训练与变量评估
模型2(、3等)
模型1
模型2参数优化
模型2训练
模型组合调优
模型1训练
模型1参数优化
预测提交
完成

大纲

预备

数据读取

数据摘要与清洗

模型选择

模型训练与评估

模型组合与预测

回顾

数据读取

- Q: 数据集有一些不同的文件，怎样合成一个数据呢？
- A: 首先我们可以根据数据类型为它们重命名来分门别类
  - 项目名(PPD)可以做前缀，区分项目时一目了然
  - 主表(da)、历史记录(dah)、辅助(daa)、初赛预测列(day)
  - 训练集(t)、预测集(v)
  - 重复的可以通过字段和数字序号添加后缀标识
- 用pandas包批量读数据
  - pd.concat + map + pd.read_csv + 文件名的list
  - 记得读数据时将文件中表示空值的一些符号标记为空值
  - 通过主表DataFrame的fillna把初赛预测列填充好

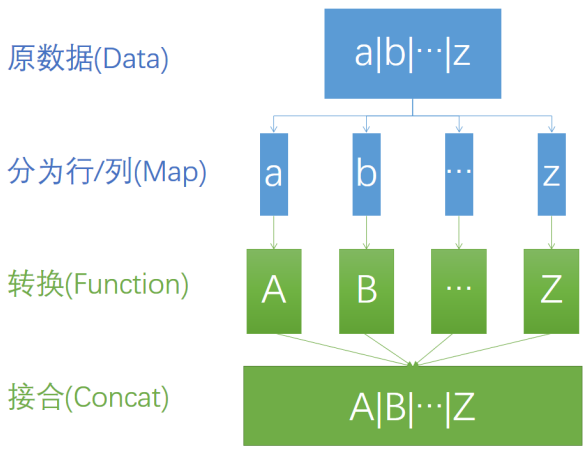
历史记录处理

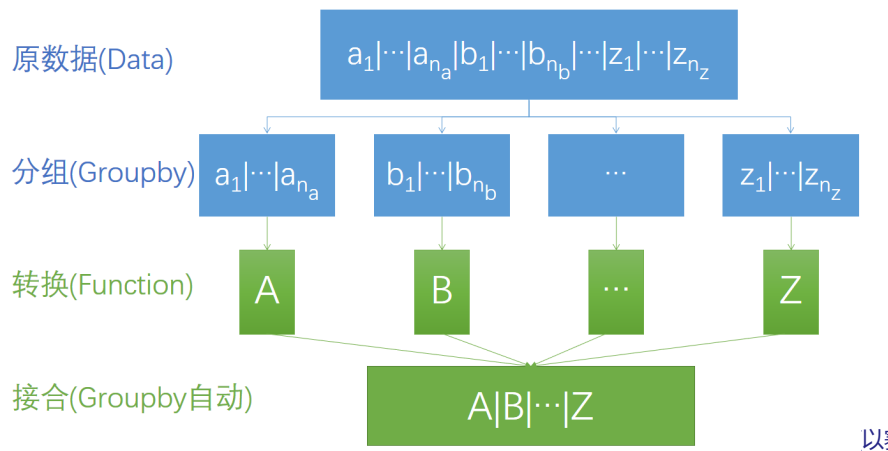
- Q: 历史记录的两个表LogInfo和UserUpdate怎样使用呢？
- A: 通常地说，历史记录相对于主表的核心差异是：对于每个index的各项信息，主表按列汇总，而历史记录按行堆叠
- 解决方案：将历史记录按index分组，把各行信息汇总到各列上，使各index对应唯一一行以与主表连接
- 风控中，对每笔贷款的历史记录，其起始时间和（登录/信息更新）总频率对衡量借款人的行为或许较重要，进一步，也统计每类子事件的频率
- 接下来，我会介绍Python的Pandas包（简称pd）中两个批量转换数据的重要组合

数据批处理实现

这两个批量转换数据的组合在我们整个数据处理阶段非常实用，并将数据处理简化为3个问题：1.决定如何分组，2.编制什么函数，3.怎样安排处理顺序

组合	pd.concat + map	pd.DataFrame + groupby
处理单元	单级行/列（分组）	多级行/列分组
应用函数	任意	简单统计量
按行应用	数据读取	历史记录处理
按列应用	数据摘要、变量清洗	概括系列变量





以赛亚书11:9)

大纲

预备

数据读取

数据摘要与清洗

模型选择

模型训练与评估

模型组合与预测

回顾

数据摘要

- Q: 当我读取整理好数据集之后，接下来要做什么呢？
- A: 首先我们要从大局出发，简化并理解数据特征
- 具体地，可以对各变量处理汇总成一个表格。各行是变量名（原数据的每一列），而各列的内容有：
  - 变量类型
  - 变量的空值/非空值数量
  - 变量出现频数前5大的值与数量，和其他值的数量（尾巴）
  - 数值变量的统计量：均值、方差、四分位数（含最值）
- 可以用批处理函数pd.concat + map + (function)实现

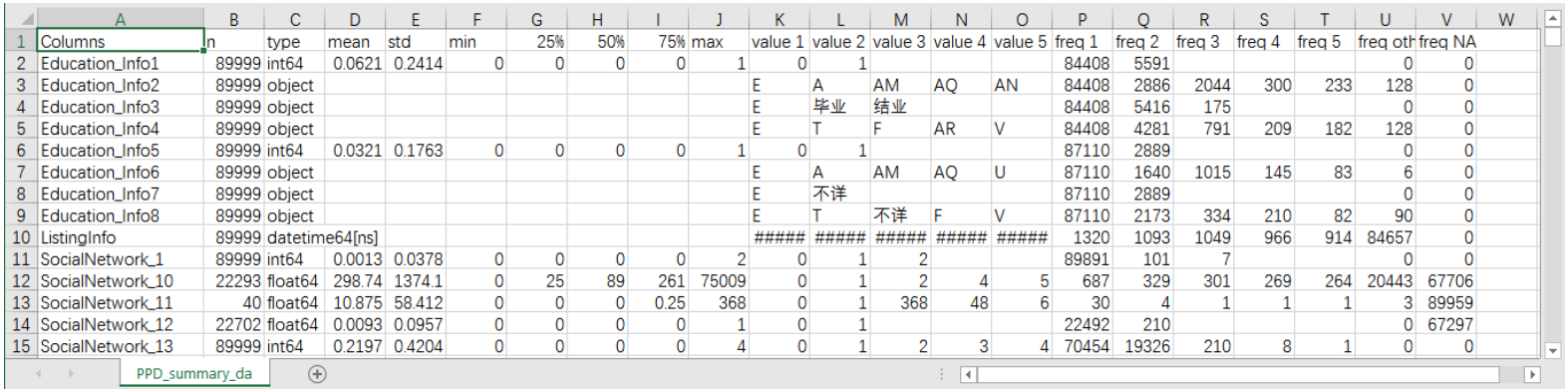
数据摘要展示

- 摘要前（90000行*354列）：

Idx	User	User	User	User	Wet	Wet	Wet	Wet	Wet	Wet	Wet	Wet	Wet	Wet	Wet	Wet	Wet	Wet	Wet	User	User	User	User	User	User	User	User	User	User	User	User	User	User	User	User	User	User	User	Edu	Edu	E
##	1	深圳	4	深圳	1		1	1	1	14	0	0	0	0	0	0	6	6	0	6	2	2	2	广东	深圳	中国	0				2	2	1	1	29	四川	南充	0	D	D	D	0	E	E
##	1	温州	4	温州	0		1	1	1	14	0	0	0	0	0	0	0	0	7	7	0	2	2	浙江	温州	中国	1	0	0	1	3	3	2	2	28	福建	不详	0	D	D	D	0	E	E
##	1	宜昌	3	宜昌	0		2	2	2	9	3	0	0	0	0	0	0	0	3	4	2	2	2	湖北	宜昌	中国	0	0	0	1	4	4	2	1	32	湖北	宜昌	0	D	D	D	0	E	E
##	4	南平	1	南平					2	0	0	0	0	0	0	0	0	0	2	0	2	2	2	福建	南平	中国	0	0	0	1	2	2	1	1	33	江西	不详	0	D	D	D	0	E	E
##	5	辽阳	1	辽阳	0		1	1	1	3	0	0	0	0	0	0	0	0	3	0	2	2	2	辽宁	辽阳	中国	0				3	3	1	1	43	辽宁	锦州	0	D	D	D	0	E	E
##	1	吴忠	5	银川	0		2	2	2	4	0	0	0	0	0	0	0	1	1	2	0	2	2	不详	不详	不详	0				4	4	2	1	24	山东	不详	0	D	D	D	0	E	E
##	1	绵阳	3	赤峰	0		13	1	13	15	0	0	0	0	0	0	0	2	2	10	1	2	2	内蒙	包头	中国	1	1	0	0	2	2	1	1	33	内蒙	赤峰	0	D	D	D	0	E	E
##	4	东莞	5	东莞	0		12	6	12	8	0	0	0	0	0	0	0	2	1	3	2	2	2	广东	广州	中国	1				5	5	1	1	29	湖北	荆州	0	D	D	D	0	E	E
##	1	赤峰	6	赤峰	0		10	6	10	19	0	0	0	0	1	0	1	9	7	2	1	2	2	内蒙	赤峰	中国	0				3	3	1	1	29	内蒙	赤峰	1	未婚	M	内蒙	1	AM	与
##	3	武汉	5	鄂州	0		1	1	1	6	0	0	0	0	0	0	0	1	3	2	1	1	1	湖北	鄂州	中国	0				4	4	1	1	34	湖北	鄂州	0	D	D	D	0	E	E
##	5	武汉	5	武汉	0		3	2	3	11	0	0	0	0	0	0	0	0	2	9	0	1	1	湖北	武汉	中国	0	0	1	0	2	2	1	2	28	湖北	武汉	0	D	D	D	0	E	E
##	5	长沙	5	长沙	1		1	1	1	4	0	0	0	0	0	0	0	0	1	3	0	1	1	湖南	长沙	中国	0				3	3	1	1	24	湖南	湘潭	0	D	D	D	0	E	E
##	3	漳州	3	漳州	1		11	3	11	31	0	0	0	0	0	0	0	2	3	25	1	2	2	福建	漳州	中国	0				6	4	3	1	27	福建	漳州	0	D	D	D	0	E	E
##	3	牡丹	6	牡丹江	0		1	1	1	4	0	0	0	0	0	0	0	0	0	4	0	1	1	黑龙	牡丹	中国	0				3	3	1	1	31	黑龙	牡丹	0	D	D	D	0	E	E

PPD_dat_1

- 摘要后（354行*22列）：



Columns	n	type	mean	std	min	25%	50%	75%	max	value 1	value 2	value 3	value 4	value 5	freq 1	freq 2	freq 3	freq 4	freq 5	freq oth	freq NA
Education_Info1	89999	int64	0.0621	0.2414	0	0	0	0	1	0	1				84408	5591				0	0
Education_Info2	89999	object								E	A	AM	AQ	AN	84408	2886	2044	300	233	128	0
Education_Info3	89999	object								E	毕业	结业			84408	5416	175			0	0
Education_Info4	89999	object								E	T	F	AR	V	84408	4281	791	209	182	128	0
Education_Info5	89999	int64	0.0321	0.1763	0	0	0	0	1	0	1				87110	2889				0	0
Education_Info6	89999	object								E	A	AM	AQ	U	87110	1640	1015	145	83	6	0
Education_Info7	89999	object								E	不详				87110	2889				0	0
Education_Info8	89999	object								E	T	不详	F	V	87110	2173	334	210	82	90	0
ListingInfo	89999	datetime64[ns]								#####	#####	#####	#####	#####	1320	1093	1049	966	914	84657	0
SocialNetwork_1	89999	int64	0.0013	0.0378	0	0	0	0	2	0	1	2			89891	101	7			0	0
SocialNetwork_10	22293	float64	298.74	1374.1	0	25	89	261	75009	0	1	2	4	5	687	329	301	269	264	20443	67706
SocialNetwork_11	40	float64	10.875	58.412	0	0	0	0.25	368	0	1	368	48	6	30	4	1	1	1	3	89959
SocialNetwork_12	22702	float64	0.0093	0.0957	0	0	0	0	1	0	1				22492	210				0	67297
SocialNetwork_13	89999	int64	0.2197	0.4204	0	0	0	0	4	0	1	2	3	4	70454	19326	210	8	1	0	0

PPD_summary_da

- 摘要信息帮助我们对变量特征一目了然，以开展清洗工作

“因为认识耶和华的知识要充满遍地，好像水充满洋海一般。”（以赛亚书11:9）

数据清洗：目的

- Q: 为什么我们要进行数据清洗？
- A: 模型向往的是分布良好的数值，数据却有着骨感的现实
  - 空缺、类别（字符串）……——模型陷进了Bug中
  - 稀疏性、共线性、极端值……——模型迷失在数学难题中
  - 时间、地理名称……——模型在人类知识面前踌躇不进
- 我们要为模型铺平数据的道路，使模型能在其上飞驰
- 整个数据分析流程的重中之重

数据清洗：思路

- Q: 这么多变量，真的需要我一个一个看来清洗吗？
- A: 不必的，我们要搭建通用的5步法依次批量完成清洗：
  1. 数值变量保留，**非数值变量**全部转为数值变量：
     - **有额外信息的非数值变量**转化为对应的数值：时间→年月日周、相对天数，地名→经纬度和城市等级，定序变量→序数
     - **其余非数值变量**全部转为0-1哑变量
  2. 选取统计量概括**一系列相似变量**：取中位数、方差、求和、最值、空值数等概括各时期第三方信息、几个城市变量等。统计量重精不重多，尽量互相独立
  3. 删除**稀疏变量**： 空值/同一值占绝大比例（如99.9%）的列
  4. 删除**共线变量**： 相关矩阵的严格下三角阵有接近± 1的列
  5. 用中位数填充**空值**，最后正态**标准化**：rank与正态分布的百分位函数复合

数据清洗：答疑

- Q: 为何选用中位数而不是平均数填充空值呢？
- A: 数据分布不对称时，中位数比平均数更能保持排序关系
- Q: 为何进行正态标准化，而不是中心归一标准化呢？
- A: 是为应对实际数据的大量有偏分布和极端值
  - 出发点：决策树集成类模型不依赖于数据分布，预测效果往往好；反而考虑分布信息的模型经常受分布偏差的负面影响
  - 正态标准化特点：只保留排序关系，彻底去除有偏分布和极端值，在大样本下直接满足众多模型假设
  - 在本数据集能明显提高逻辑回归和神经网络的效果
- 当我们完成了清洗的工作后，即将踏入建模阶段

大纲

预备

数据读取

数据摘要与清洗

模型选择

模型训练与评估

模型组合与预测

回顾

Logistic Regression

- Q: 如果我刚入门机器学习，应该从什么模型开始？
- A: Logistic Regression：最简洁、快速、稳健的做法，可解释性强，适于工业界。可使用sklearn包
- 但由于比赛以精度为标准，由于Logistic Regression对变量关系的线性限制，难以达到精度最优
- 但是我们在建模时可以充分发挥它的特性：
  - 通过增加L2罚函数减少过拟合
  - 作为基准，对数据清洗效果和模型表现进行快速评估
  - 与结构不同的模型加权组合预测，补充原模型精度和稳健性

XGBoost (Gradient Boosting Trees)

- Q: 如果我对机器学习已经有所了解，打算以精度为目标，用什么模型效果好？
- A: 考虑这是一个非线性的分类问题，变量成分较多元，样本和变量间无固定模式关联（图像、语音、时间序列等）。如果以精度为目标，综合考虑稳健性、速度、通用性等因素可以首选XGBoost
- Q: XGBoost的原理是什么？有哪些重要参数？
- A: XGBoost一种梯度提升树(Gradient Boosting Trees)。好比用大石头雕刻人像，每棵决策树都凿掉一些石头（残差），然后对剩下的石头继续雕刻，直到雕出人形
  - 步长(eta)雕刀：大斧子 vs. 小凿子
  - 变量抽样(colsample_bylevel)匠师：项羽 vs. 刘邦
  - 深度(depth)刀法：平推 vs. 直钻

Keras (Neural Network)

- Q: 如果我对XGBoost的精度仍不满足，想达到更好的预测效果，该如何做？
- A: 可以尝试神经网络包Keras，并把XGBoost与多模型组合
- XGBoost的出发点是各变量完全独立，而从决策树的二分关联叠加向真实关联趋近；而神经网络的出发点是各变量充满复杂的非线性关联，而不断去优化网络权重向真实关联趋近。两种模型结构具有较高的互补性
- 由于神经网络内部结构复杂，寻找最优解困难，需要详细了解并合理搭建网络结构并优化参数，建模难度较前两者高。

模型对比

- 精度以相同的10-folds交叉验证为准
- 训练样本8万，变量经清洗后共389个，正态分布标准化
- 计算平台：Intel Core i5 4300U 双核 2.5 GHz, 8 GB 内存

模型	LR	XGBoost	Keras
类别	逻辑回归	梯度提升树	神经网络
基本单元	线性	二分	网络
平均精度(AUC)	0.775	**0.787**	0.771
最差精度(AUC)	0.759	**0.768**	0.753
单模型时间(s)	**8**	350	400
调参个数	**2**	8	10+
可解释性	**好**	中	水平不行
功用	**快速建模**	**精度建模**	精度辅助
	评估辅助	**变量评估**	
	精度辅助		

大纲

预备

数据读取

数据摘要与清洗

模型选择

模型训练与评估

模型组合与预测

回顾

交叉验证与模型训练

- Q: 为什么要用交叉验证？怎样用？
- A: 较比用单训练预测集建模，交叉验证的优势主要有：
  - 更准确地估计模型预测精度：均值
  - 预估模型预测效果范围：标准差（置信区间）、箱线图
  - 减少过拟合风险
- 实现方法：以10-folds为例做交叉验证：
  1. 把样本行的index随机拆成10份保存起来
  2. 每次取1份作验证集index，其余9份粘起来作训练集index，取X和Y的训练和验证集训练模型，把模型保存起来
  3. 依次取10份不同的index，得到一组10个模型
  4. 预测时用10个模型预测结果取平均

变量评估

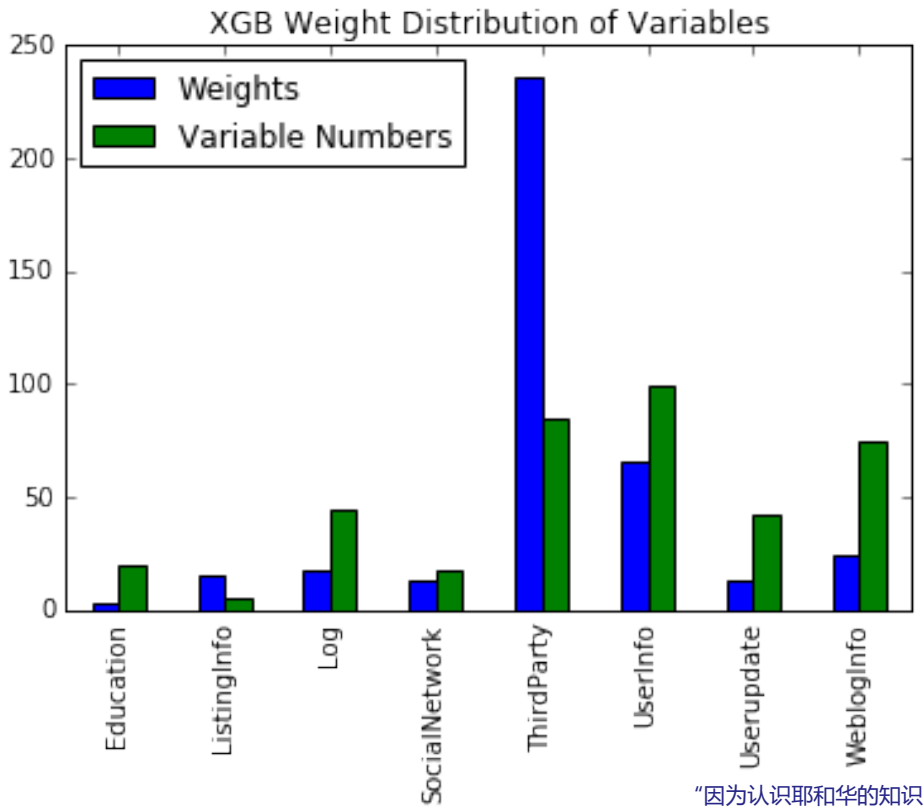
- Q：我们该如何了解Y受哪些变量影响呢？
- A：可以用XGBoost判断变量重要性，再用LR看影响方向

模型系数	LR线性权重	XGBoost分支相对频率
针对目标	变量线性影响	变量非线性影响
计算方式	(变量标准化后)权重	$\frac{\text{fscore}}{\text{Mean(fscore)}}$
判断标准	正负号：影响方向 绝对值：影响大小	大于或接近1：变量重要 远小于1：变量不重要
稳健性	差（共线性虚假相关）	好（共线性影响小）

变量评估展示

如图为XGB变量相对频率按组汇总，可据此改进我们的

- 数据收集：增加对重要变量的收集工作
- 变量清洗：针对重要变量进一步转换组合，及需要情况下对相对频率几乎为0的变量的清除



参数优化

- Q: 如何进行参数优化？怎么选取初始值？
- A: 模型调参是非常考验耐心和时间的过程
  1. 在调参前，首先要**理解模型和参数的含义**，这步非常关键
  2. 先用单数据集，从默认值开始，**手工逐个调参**熟悉模型:小范围用等差数列，大范围用等比数列，确定合理参数范围
  3. 确定大致范围后，可以用**交叉验证+自动搜索**来得到最优参数，如Python的HyperOpt包
- 自动搜索最优参数时，我们可以用更少folds做交叉验证，以及稍大的梯度步长训练模型
  - 节约调参时间
  - 数据集不同，减少对交叉验证结果的过拟合
- 在找到最优参数后，我们重新在原交叉验证集上用最优参数训练模型，至此模型训练阶段结束

大纲

预备

数据读取

数据摘要与清洗

模型选择

模型训练与评估

模型组合与预测

回顾

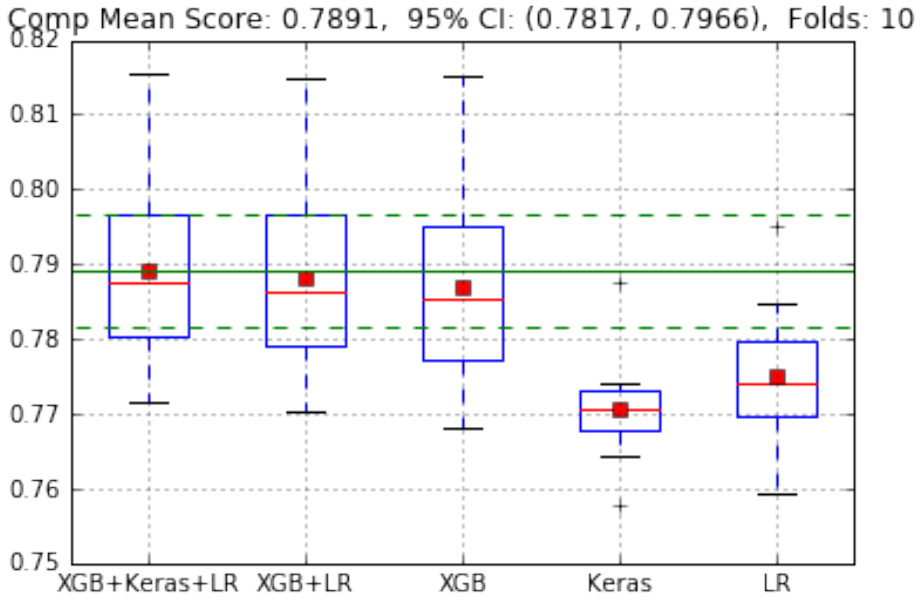
模型组合

当训练优化好各组交叉验证模型后，就可将各组模型**加权平均**预测了，比如我们这里使用XGBoost，Keras和LR三组模型加权平均，并使用HyperOpt取得最优加权权重

- Q: 为什么要使用模型加权平均而不是最优模型预测？
- A: 数学上，如果一个无偏模型的预测方差为V_1，当我们加入另一个无偏而完全独立的模型，该模型预测方差为V_2，当我们对两模型的预测结果加权平均取最优解时，预测方差会变成原V_1的$\frac{V_2}{V_1+V_2}$倍
- 当然因为实际数据集和模型结构所限，真实的模型往往是有偏的，而只有一小部分相互独立，因此改进效果并没有理论上那样明显，但至少是一种比较稳健的方法
- 预测提交时，我们先对三组模型的交叉验证预测Y分别算术平均，再把这三个Y照权重加权平均，就可以提交了

效果展示

- 不同模型组合在同一10-folds交叉验证集上的得分分布
- 训练样本8万，变量经清洗后共389个，正态分布标准化
- 最优权重：XGB+LR=90:10，XGB+Keras+LR=75:20:5
  - Keras虽然预测精度较低，但结构互补进一步改善模型效果
- 实际预测集分数：0.7887



预测反馈

- Q: 为什么排行榜上的结果要比交叉验证的结果要好/差？
- A: 通常来说，预测的结果稍可能比交叉验证略好，原因是在不同数据集的交叉验证模型取平均形成部分互补减小误差
- 当然，因为预测集数据分布有随机性，预测效果的区间大致可以通过交叉验证的均值$\pm \frac{2}{\sqrt{K}}$标准差来估算（K-folds）
- 我们也要在全程中注意避免过拟合，包括：
  - 避免将Y的真值/预测信息在数据清洗或建模时引入到X中
  - 模型优化时采用另外划分的交叉验证集
  - 尽量能说清所做每一步处理的必要性和通用性
  - 注意：反复尝试变量组合提高验证集分数时，可能造成过拟合

大纲

预备

数据读取

数据摘要与清洗

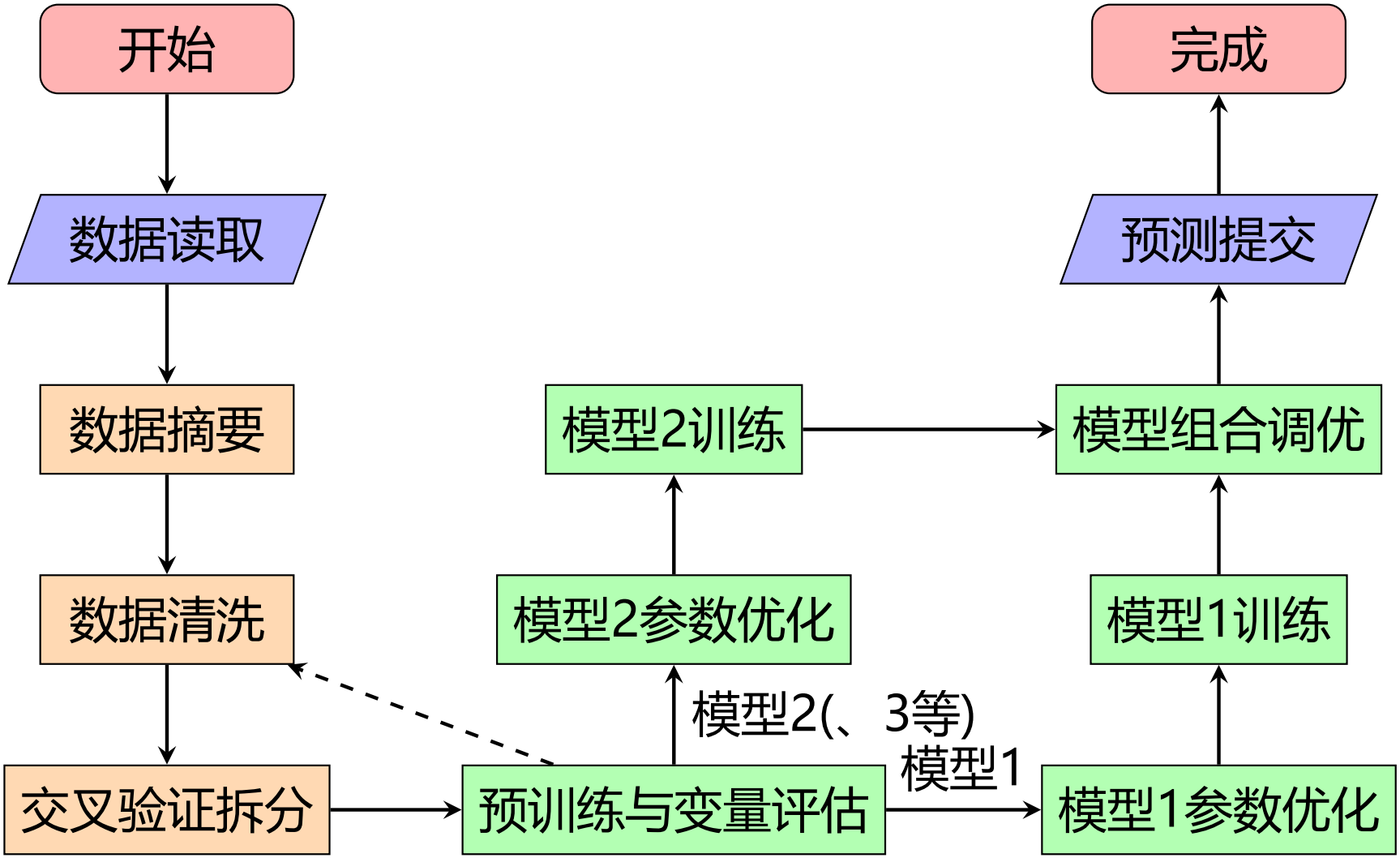
模型选择

模型训练与评估

模型组合与预测

回顾

流程回顾


开始
数据读取
数据摘要
数据清洗
交叉验证拆分
预训练与变量评估
模型2(、3等)
模型1
模型2训练
模型2参数优化
模型1训练
模型1参数优化
模型组合调优
预测提交
完成

流程思想要点

- 为数据建模逐步搭建通用的函数（清洗、拆分、训练、优化等），将整个流程尽量自动化、可重复、可移植
- 注意对数据建模的整个过程进行评估（时间、复杂性、过拟合等），减少不必要的中间环节
- 清洗数据时，构造的人工变量要少而精，在相互独立和完备覆盖之间取得平衡，从而为模型增加有效信息帮助预测

改进潜力

- 当我们完成所有数据建模的必要工作时，在有需要而且有足够资源的前提下，可以在当前预测精度上进一步改进
  - 数据清洗：对预测Y较重要的变量之间可尝试多种组合变换（四则运算、各种分布变换、系列变量的各类统计量等），增加模型可以发掘的有效信息
  - 模型选择：可引入更多种类的模型，如随机森林，不同结构（层数、激活函数等）的神经网络等，改善模型互补性
  - 参数优化：减小梯度步长，及在更多超参数中搜索最优参数
- 但是
  - 可能会继续指数级增加所需时间、精力、计算量
  - 精度提升和算法的通用性改进可能明显减少
  - 可能陷入为改进而改进的循环中
  - 直到机器学习界的AlphaGo取代人工劳作的数据分析师

“因为认识耶和华的知识要充满遍地，好像水充满洋海一般。”(以赛亚书11:9)

局限与反思

- 同时，我们目前所做的数据模型是很有限的：
  - 数据的预测局限：当试图穷尽数据处理、模型、调参等方法时，投入时间、复杂度与计算量会呈现指数级增长，然而往往仅能取得1%，甚至0.1%的提升，与真理相去仍然甚远
  - 模型的视角局限：模型只是指引决策的参考，却不能对它的决策造成的影响从人性上进行价值判断和承担责任（歧视、刷信用、校园贷、...）
  - 模型的反馈局限：模型在欠拟合的经济/数据体系中发挥正面作用，当经济/数据体系已经过拟合，模型和体系的系统性风险会成倍放大（金融危机、评级垄断、高频交易、...）
- 我们到目前所学习与创作的，只是浩瀚历史中一朵瞬间的水花，我们生命的盼望却不在这里
- 愿恩惠平安从主基督耶稣临到所见的人

“因为认识耶和华的知识要充满遍地，好像水充满洋海一般。”(以赛亚书11:9)

神爱世人，甚至将他的独生子赐给他们，叫一切信他的，不至灭亡，反得永生。——约翰福音3章16节