机器学习从入门到参加竞赛

——以JDD信贷需求预测为例

王乐 ML/NLP学习交流群 189105362

目 录

1、Python基础

基本数据结构操作：

list , tuple , dict ,

函数编程：

```
def function( * ):
    pass
```

条件、循环语句：

if… else…　　for i in range(N)

csv文件的读写：

pandas.read_csv()

Python机器学习常用库：

numpy, pandas , matplotlib, …

- 这些基础知识随便找一本Python教材或者网络教程就可以学会
- 我曾经开的一个知乎LIVE也基本讲了这些基础内容

2、数据分析

对数据进行一些探索分析有助于我们构造特征以及建立必要的数据敏感度

t_user.csv

	uid	age	sex	active_date	limit
0	26308	30	1	2016-02-16	5.974677
1	78209	40	1	2016-02-21	5.292154
2	51930	35	1	2016-04-19	6.292055
3	10113	25	1	2016-03-12	6.292055
4	17067	35	1	2016-02-16	5.974677

t_click.csv

	uid	click_time	pid	param
0	12177	2016-10-04 12:22:30	1	19
1	29226	2016-10-04 12:18:42	6	1
2	37351	2016-10-04 12:18:41	10	16
3	82053	2016-10-04 12:18:16	1	40
4	82053	2016-10-04 12:18:46	1	40

t_loan_sum.csv

	uid	month	loan_sum
0	34939	2016-11	6.316423
1	80338	2016-11	6.212631
2	5018	2016-11	6.153414
3	58005	2016-11	6.793132
4	52453	2016-11	4.292651

t_loan.csv

	uid	loan_time	loan_amount	plannum
0	12135	2016-08-03 00:05:26	3.862595	1
1	41403	2016-08-03 00:13:25	5.584137	3
2	74458	2016-08-03 00:13:58	4.723017	1
3	12959	2016-08-03 00:19:33	3.862595	1
4	89641	2016-08-03 00:23:13	4.292651	1

t_order.csv

	uid	buy_time	price	qty	cate_id	discount
0	45370	2016-11-23	3.995009	1	22	0.0
1	66975	2016-11-23	3.269410	1	26	0.0
2	75358	2016-11-23	2.255235	1	14	0.0
3	40597	2016-11-23	1.635284	1	20	0.0
4	83886	2016-11-23	1.920573	2	22	0.0

3、初步特征工程

☑特征工程是什么?

☑如何构建特征?

比如我们需要用机器通过一些信息来区分东方人和西方人，以此为例我们讲一讲什么是特征。

人：头发（长度，颜色，发型），身高，体重，肤色，着装，母语，声音，瞳孔颜色，五官特征（鼻子形状，嘴唇厚薄，耳朵形状），身份地位，职业，学历，收入。。。。。。

对于我们的任务（区分东方人和西方人），上面的特征是否都需要?上述特征的取值取连续的还是离散的？从上述特征还不能构建有助于提升区分准确率的特征?

所谓特征，就是某个样本实例的独一无二的特性，但是为了分类或者回归、排序，我们需要找到属于同一类样本的共有特性，和其他类别的样本有区分度的特性。

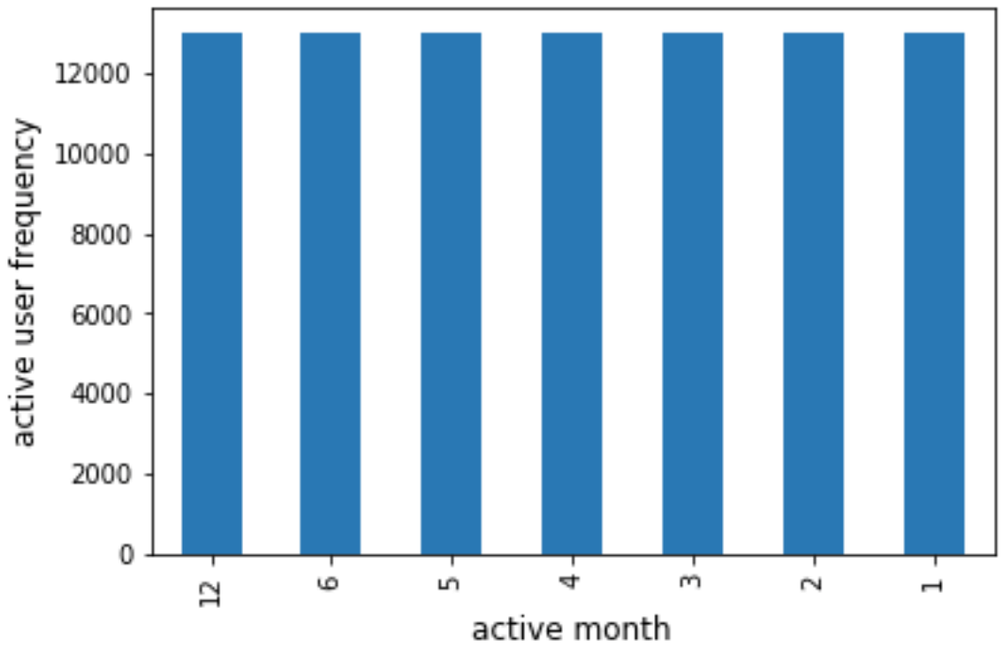
3.1、user表的特征构建

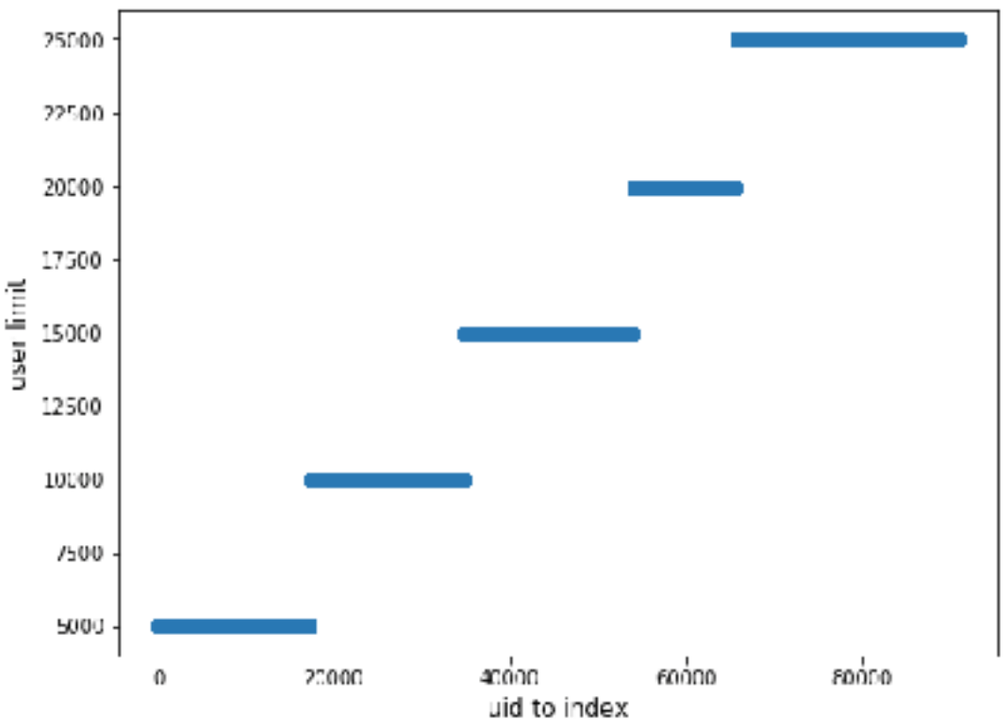
激活日期： 离现在激活多久

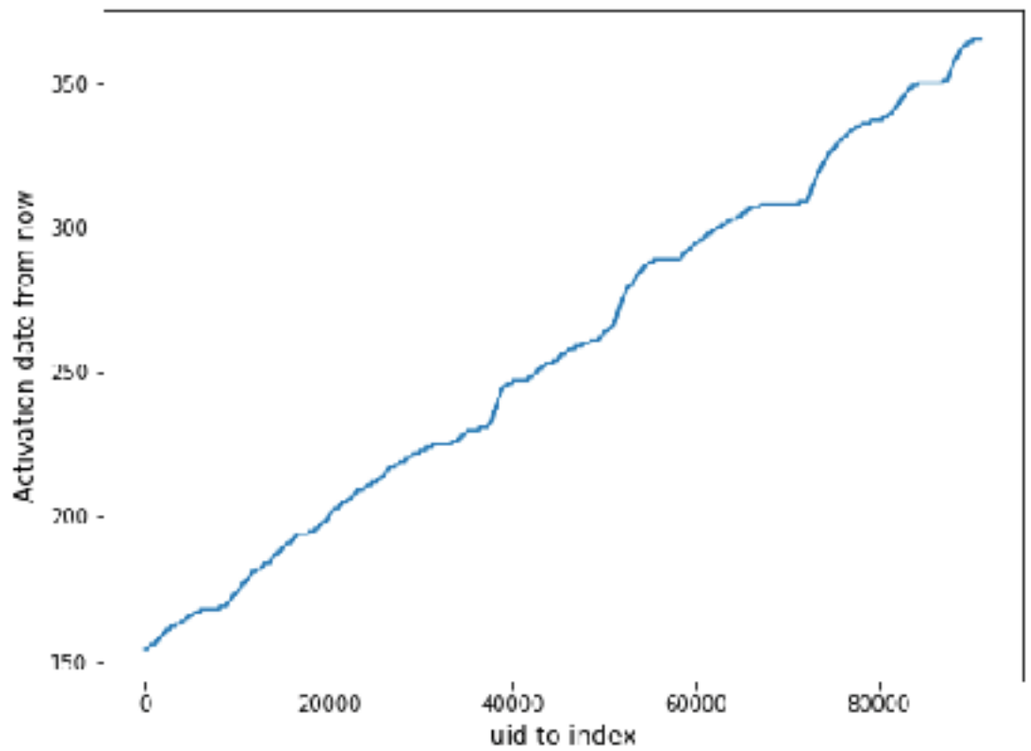
性别：男、女

年龄：20~50

初始额度：5个等级

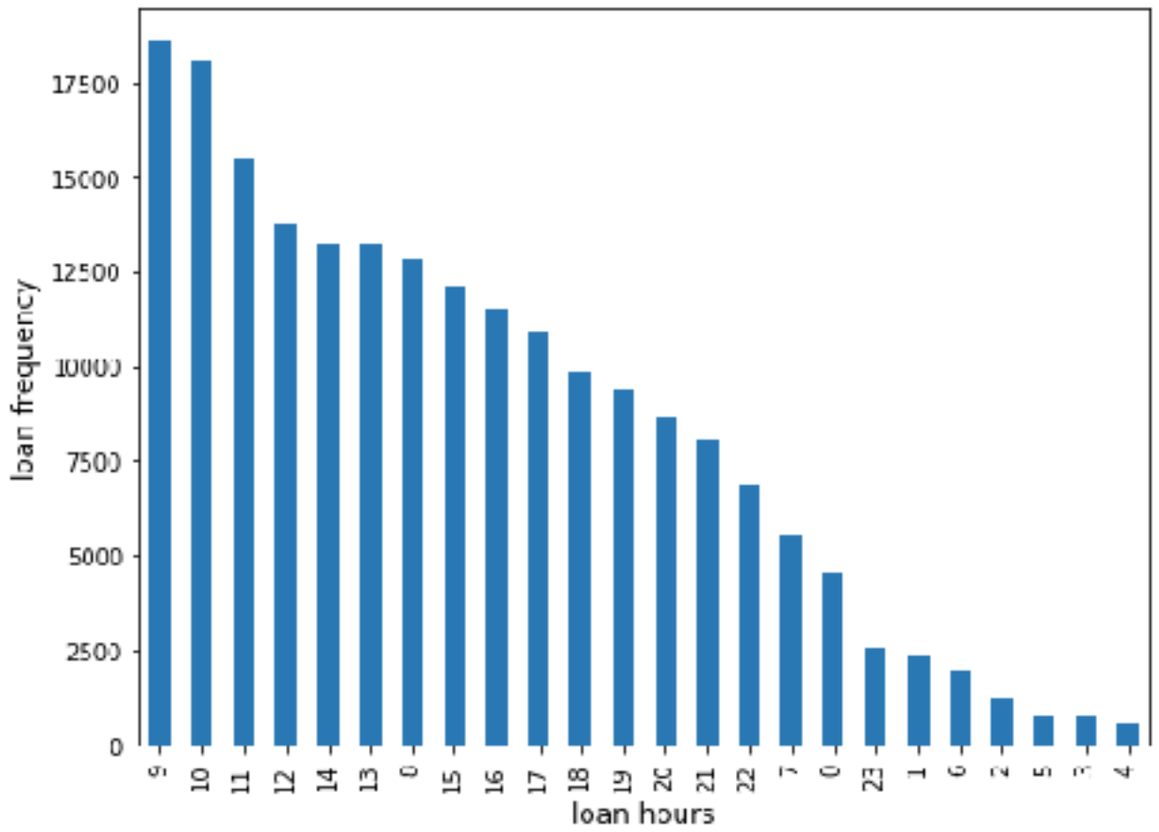


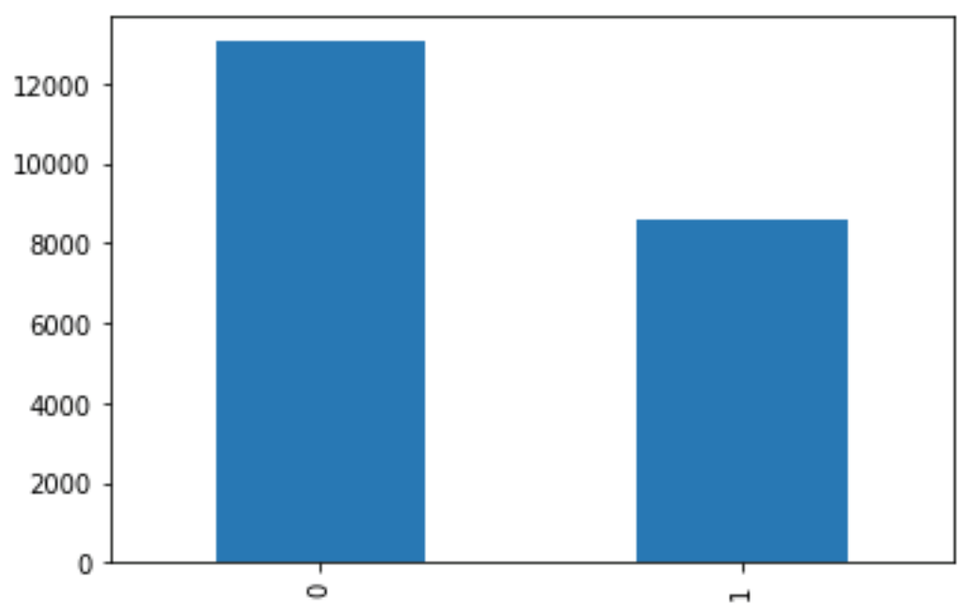




3.2、loan表的特征构建

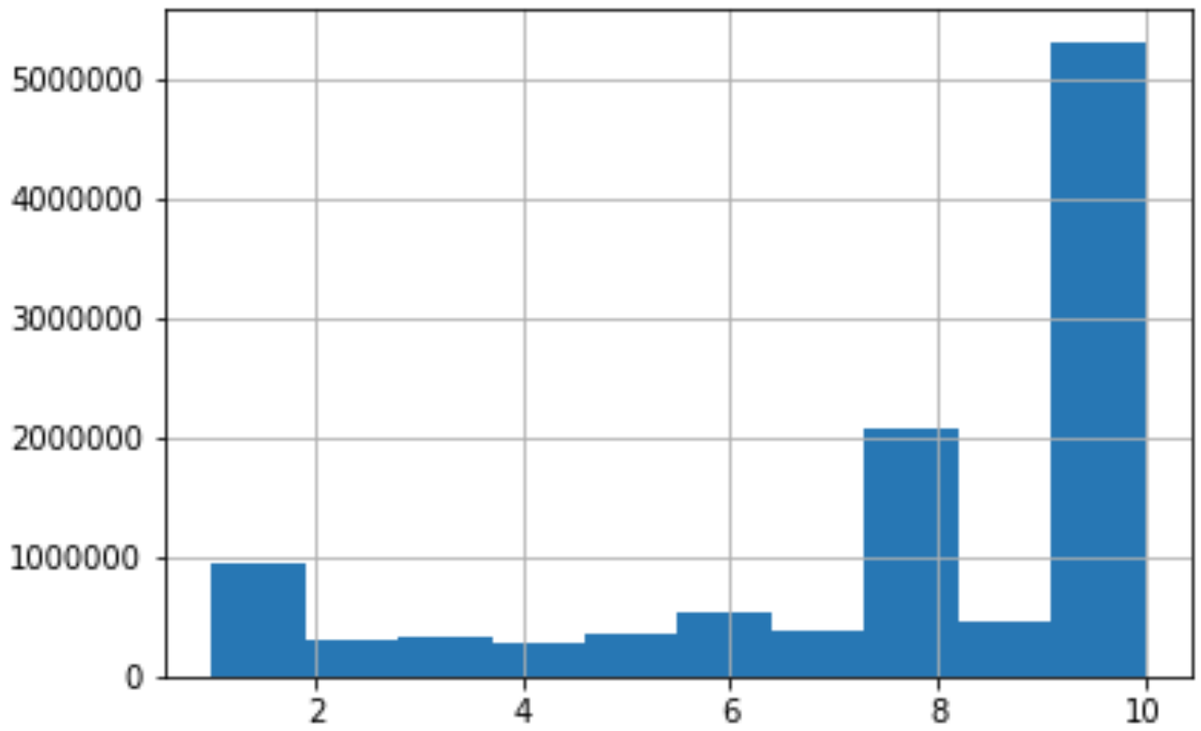
每月贷款次数
每月贷款金额
贷款时间特点
上次贷款离现在多久
每月贷款金额统计特征
（均值、方差、频次等）
是否连续贷款
贷款分期特点
每月的还款金额
某月的贷款金额是否超过初始金额





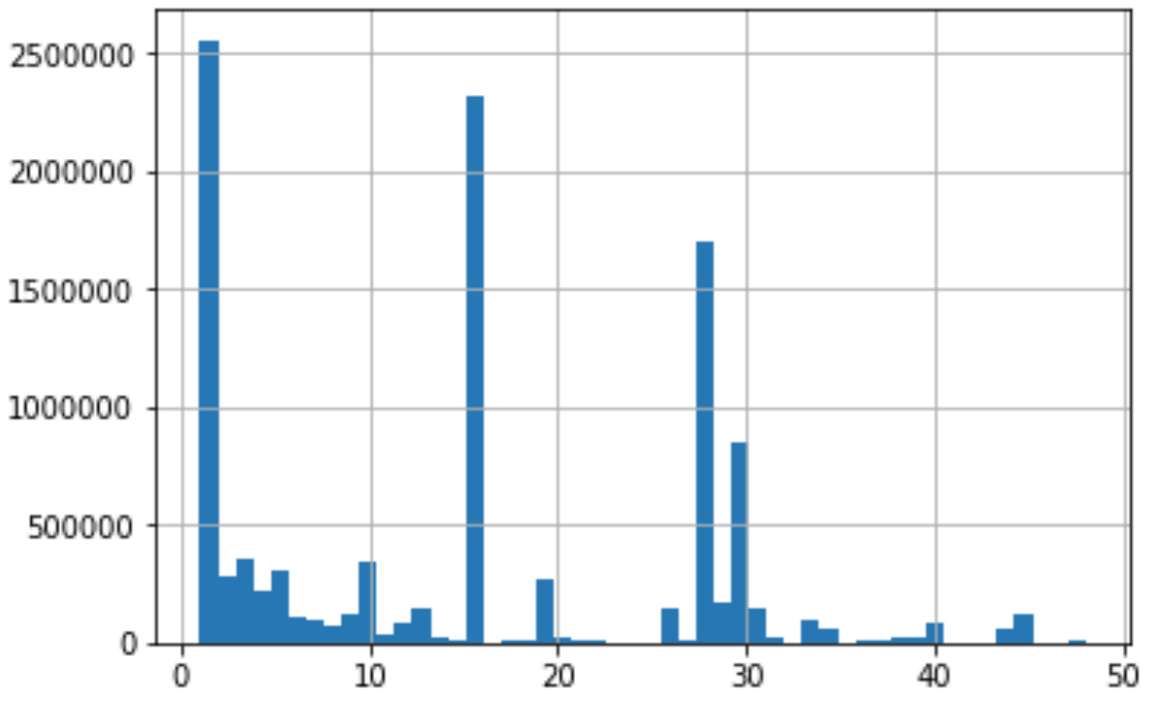
3.3、click表的特征构建

点击次数
点击权重
点击时间特点
每月点击统计特征
（均值、方差、频次等）



点击页面、参数哑编码

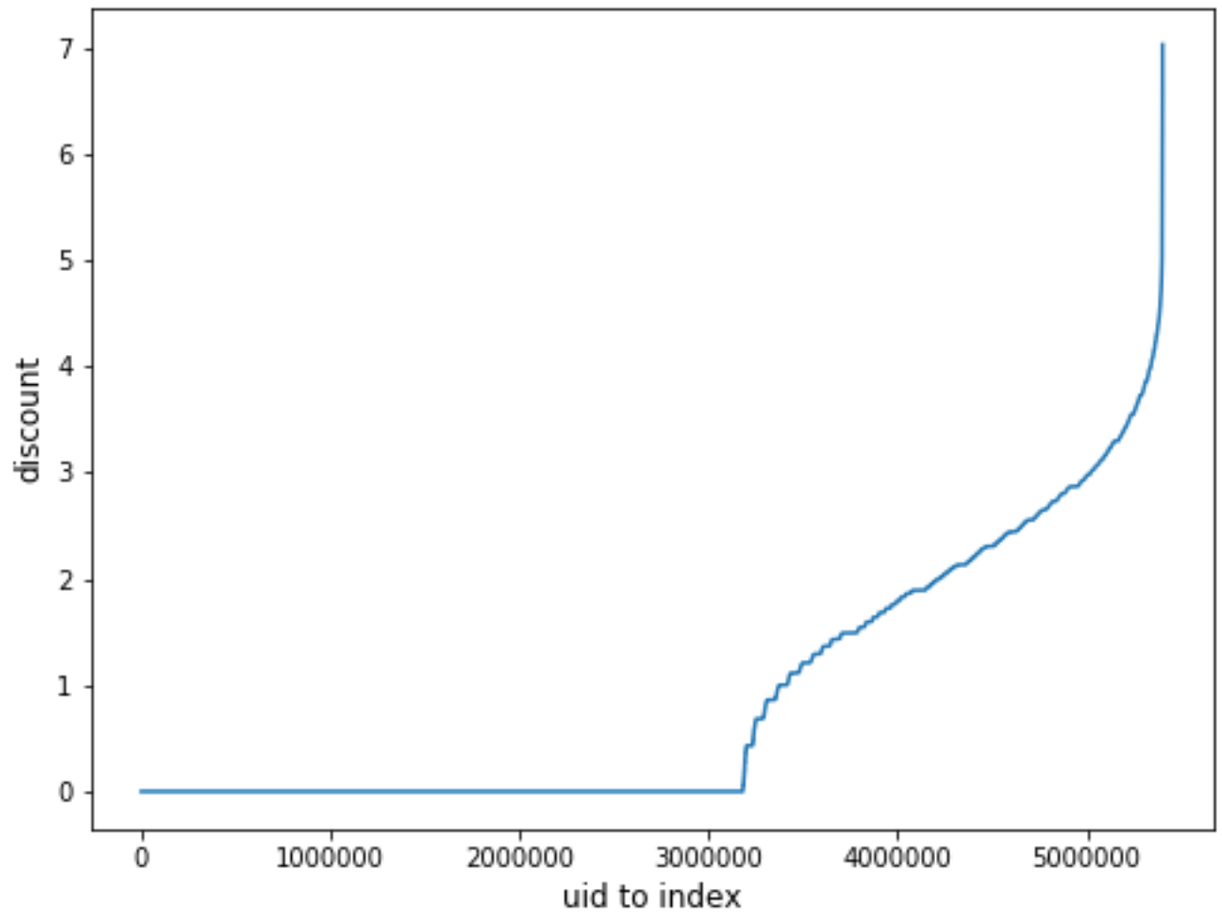
贷款的时候的点击特点
贷款次数与点击次数之比



3.4、order表的特征构建

每月消费情况
用户消费属性
消费频率
最后一次消费离11月多久
平均消费时间间隔

折扣率（是否爱买打折商品）
购买时间权重
时间窗口内消费金额统计特征
……



4、构建训练集和测试集

①训练集

训练集、验证集

11月的贷款金额作为线下训练集的标签，使用8，9，10三个月的历史数据构造特征；

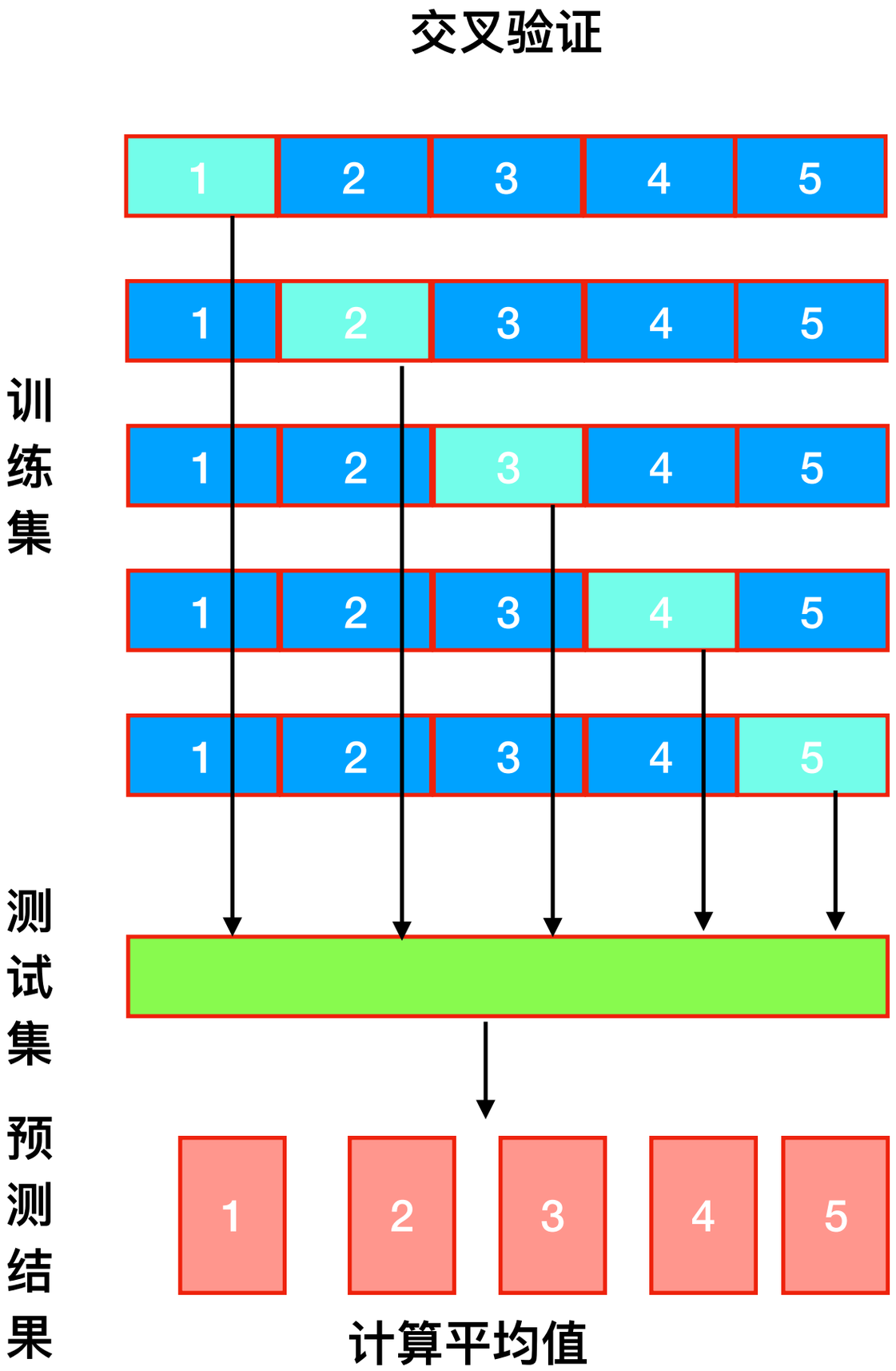
80%的样本作为训练集

剩下20%的样本为验证集

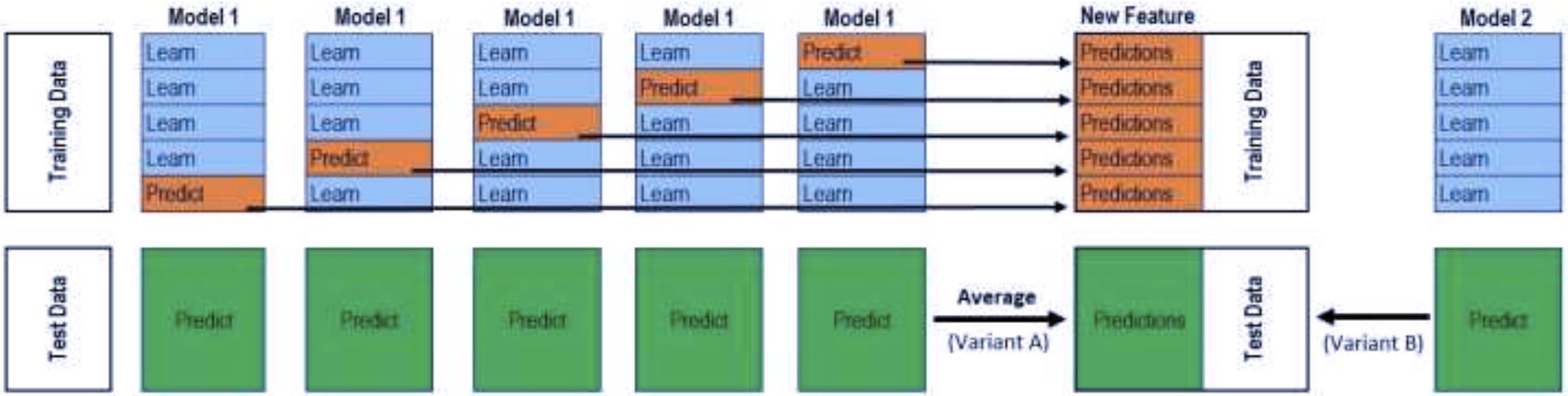
②测试集

使用9，10，11三个月的历史数据构造测试集

使用训练集训练的模型在测试集上进行预测，将结果作为12月的预测结果



7、模型融合


Model 1
Model 1
Model 1
Model 1
Model 1
New Feature
Model 2
Training Data
Learn
Predict
Predictions
Test Data
Average
(Variant A)
(Variant B)

model1	model2	model3	model4	....	...	predict
4.5	4.3					a
4.6	4.7					b
7.5	7.8					c
5.6	5.2					d