下次保养时间与公里数预测模型

陆 琴

目录

数据说明与目标

1.静态数据

车主信息：姓名、性别、身份证、出生日期、车主性质等

车辆信息： 车型、车型代码、购车时间、购车经销商、发动机型号、变速箱型号、技术代码、排量、是否电动车等

会员卡信息：会员卡生成时间、会员卡类型、会员卡状态、会员升降级日期、是否绑定微信等

2.动态数据

维修主单信息：即历史用户进店的动态数据，包括了：进店日期、当下公里数、所去经销商、维修类型、消费金额、是否给了折扣等

3.目标

- 预测全量用户的下次进店保养时间和公里数

难点

1.进店日期脏数据多，比如昨天进店保养了，今天又进店保养了

用户id	进店日期	经销商代码	维修类型
Id1	2020-09-01	dealer1	常规保养
id1	2020-09-07	dealer1	首保;常规保养

2.人工录入的公里数脏数据更多
(1)比如后面多一个0
(2)比如前面少掉一个1
(3)比如同样一个进店日期，公里数可能是大相径庭的
(4)比如今天的公里数比半年后的公里数还要小

3.最重要的字段是购车时间，存在脏数据多、缺失值多的情况

4.多种训练方式，该如何选择?

业务宽表重构

(1) 原先的宽表形式：一个vin一个进店保养日期下有多条动态样本

vin	修理日期	维修站号	委托书号	结算单号	金额	维修类型
vin1	2020/1/1	站号1	1	1	300	常规保养
vin1	2020/1/1	站号1	1	2	300	一般维修
vin1	2020/1/1	站号1	2	1	0	常规保养

(2) 重构业务宽表形式：一个vin一个进店保养日期代表一条动态样本

vin	修理日期	维修站号	总金额	维修类型	常规保养	一般维修
vin1	2020/1/1	站号1	600	常规保养,一般维修	300	300

数据清洗

(1) 保养日期清洗

--经多次数据试验，对于非多频快用户设置14天限制，上次与本次进店保养如果小于14天的，默认为上次保养造成的未完待续事宜，非真实意愿下进店保养行为；

--对于多频快用户，比如出租车网约车用户，设置为7天，同时满足间隔必须不小于日常保养天数习惯的一半；

(2)公里数清洗

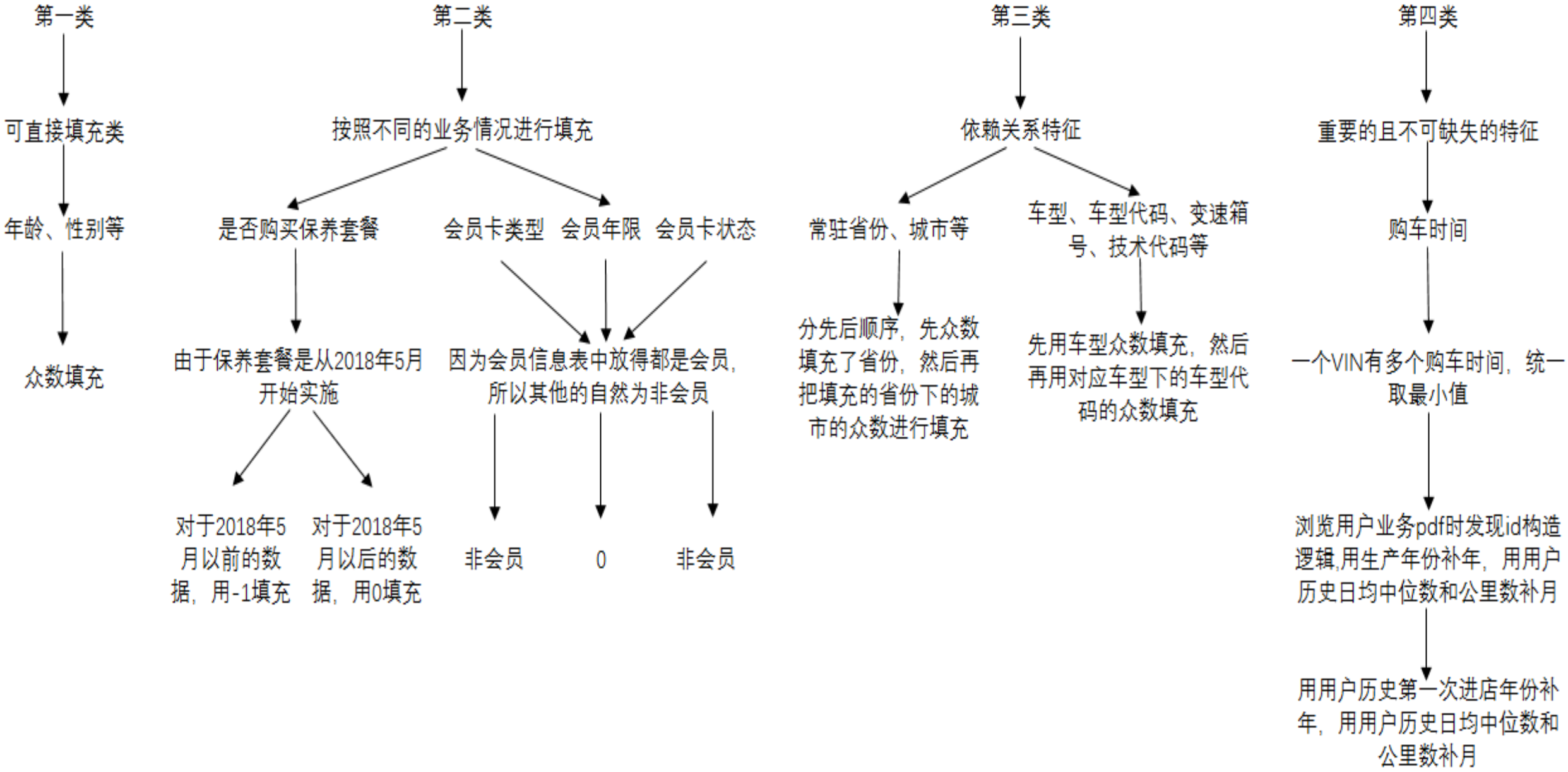
--对于一个进店日期下有多个不同的公里数的情况，经测试，直接取最大值可以保证得到的公里数脏数据的比例最少；

--对于公里数脏数据的情况，如果脏数据处于vin历史的一头一尾的情况，则会使用近期的日均与天数差进行纠正拟合；

--如果是处于vin历史数据中间的情况，曾经试过使用统计法、随机森林、插值法去拟合，但是出现动一发牵全身的情况，所以这类脏数据的情形，则直接舍弃这个VIN的数据，不再放入模型，而是放入规则中；

预测方式	筛选条件	占比
模型预测	被清洗干净的数据	69%
规则预测	没有保养记录的、新车用户、无首保用户、存在无法清洗掉公里数脏数据	31%

补充缺失值


第一类
可直接填充类
年龄、性别等
众数填充
第二类
按照不同的业务情况进行填充
是否购买保养套餐
会员卡类型
会员年限
会员卡状态
由于保养套餐是从2018年5月开始实施
因为会员信息表中放得都是会员，所以其他的自然为非会员
对于2018年5月以前的数据，用-1填充
对于2018年5月以后的数据，用0填充
非会员
0
非会员
第三类
依赖关系特征
常驻省份、城市等
车型、车型代码、变速箱号、技术代码等
分先后顺序，先众数填充了省份，然后再把填充的省份下的城市的众数进行填充
先用车型众数填充，然后再用对应车型下的车型代码的众数填充
第四类
重要的且不可缺失的特征
购车时间
一个VIN有多个购车时间，统一取最小值
浏览用户业务pdf时发现id构造逻辑,用生产年份补年，用用户历史日均中位数和公里数补月
用用户历史第一次进店年份补年，用用户历史日均中位数和公里数补月

主要特征工程

单维度统计特征群

-不同车型代码/性别/年龄/车龄等下的人数统计
-每次保养工时费和零件费的合计、占比的统计
-用户不同时间维度下的历史保养次数、有折扣的次数、占比等

群体统计特征群

-相同车型/车龄/地区等的群体的历史保养天数差/公里数差/日均的min,max,median,mean,std等
-以及交叉特征下的关于天数差/公里数差/日均的统计特征群

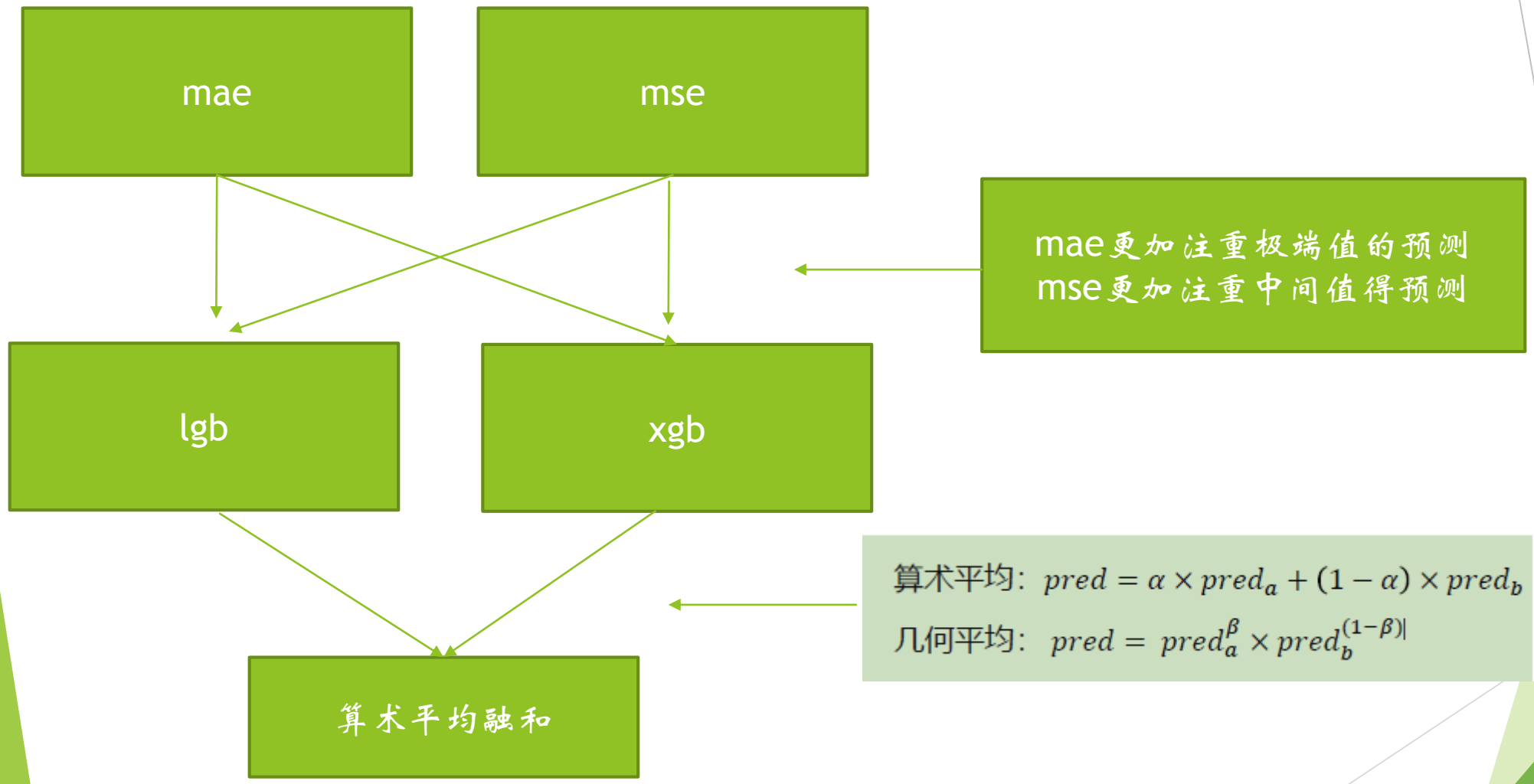
用户个人天数差/公里数差/日均统计特征群

-用户个人历史保养天数差/公里数差/日均的min,max,median等
-用户个人历史保养天数差/公里数差/日均的环比、间隔比、依次间隔比、环比增长率、间隔比增长率、依次间隔比增长率等趋势统计特征

模型

序号	训练法	重构y	优缺点	采用
训练方法1	每一天保养样本即为一个数据	这一个样本的下次保养时间/公里数就是y	提取特征的时候容易穿越	
训练方法2	窗格法，以1年为窗口，即转化为预测之后1年的第一次进店保养时间	以1年为窗口，即接下去1年的第一次保养时间/公里数就是y	大数据，单机测试版上提取特征跑起来太慢	
训练方法3	Batch训练法，即用历史保养n次的人去预测历史保养次数为n-1次的人的第n次保养时间和公里数	即用户的最后一次进店保养时间和公里数就是y	batch式训练，速度快了一倍多，精度也得到了提升	√

模型预测的框架


mae
mse
lgb
xgb
算术平均融和
mae更加注重极端值的预测
mse更加注重中间值得预测
算术平均：$pred = \alpha \times pred_a + (1-\alpha) \times pred_b$
几何平均：$pred = pred_a^{\beta} \times pred_b^{(1-\beta)}$

模型评价结果展示

天数模型

batch	mae_train	mae_vali	r2_train	r2_vali	mse_train	mse_vali
1	9324.07	10758.46	0.72	0.67	63.00	66.19
2	13536.83	15962.75	0.82	0.78	70.58	74.46
3	10837.43	13199.56	0.90	0.87	63.38	67.53
4	8524.76	10663.83	0.94	0.92	57.41	61.82
5	6539.60	8906.39	0.96	0.94	51.32	57.47
6	6063.85	8490.78	0.97	0.95	49.86	56.13
7	4835.22	7402.92	0.98	0.96	45.25	52.38
8	3615.69	6434.28	0.98	0.97	39.79	49.27
9	2480.78	5210.62	0.99	0.97	33.69	45.09
10	2464.51	5087.99	0.99	0.98	33.27	44.10
11	2807.38	7161.02	0.99	0.96	37.53	54.18
avg	6457.28	9025.33	0.93	0.91	49.55	57.15

公里数模型

batch	mae_train	mae_vali	r2_train	r2_vali	mse_train	mse_vali
1	19733381.93	22736271.62	0.37	0.27	2129.01	2183.54
2	27488291.09	33447261.43	0.67	0.60	2606.29	2720.89
3	24944614.58	31571774.99	0.79	0.74	2433.40	2535.78
4	23962478.75	31638790.55	0.85	0.80	2394.12	2522.46
5	22176259.00	29411655.76	0.89	0.86	2303.61	2452.57
6	19562577.57	27311136.53	0.92	0.89	2217.81	2394.14
7	14973739.29	23275357.98	0.95	0.92	2097.11	2375.34
8	23532704.21	36705699.02	0.94	0.91	2202.42	2447.17
9	19325650.83	30385285.03	0.96	0.93	2050.81	2363.60
10	16849638.61	27467175.20	0.97	0.94	2006.89	2388.58
11	20939097.85	39383057.73	0.99	0.98	2508.83	3110.07
avg	21226221.25	30303042.35	0.85	0.80	2268.21	2499.47

规则

序号	细分人群	规则
1	有首保,无常规,预测第1次常规	使用首保习惯相同的同车型群体的天数/公里数/日均的趋势比喂入
2	有首保,1常规,预测第2次常规	使用首保和常规保养习惯都相同的同车型群体的天数/公里数/日均的趋势比喂入
3	有首保,2次常规,预测第3次常规	根据用户个人最近2次常规保养的趋势系数,给予不同的权重融和得到下一次天数间隔/公里数间隔/日均
4	3次及以上常规,预测接下去常规	
5	无首保,1次常规,预测第2次常规(购车)	使用保养习惯相同的群体的天数/公里数/日均的趋势比喂入
6	无首保,2次常规,预测第3次常规	使用2次常规保养习惯都相同的同车型群体的天数/公里数/日均的趋势比喂入
7	无保养记录,预测首保	使用用户标签、车型相同的群体的第一次首保的天数间隔/公里数间隔/日均喂入
8	无保养习惯	100%属于已流失客户，所以直接用车型群体的趋势喂入

模型和规则评价结果展示

模型

```
1  tongji1(dt,'天数偏差分类1')
```

	天数偏差分类1	fenzi	fenmu	ratio
0	30-60天	68418	316906	0.215894
1	30天以内	126972	316906	0.400661
2	60-90天	40412	316906	0.127520
3	90天以上	81104	316906	0.255924

```
1  tongji1(dt,'公里数偏差分类1')
```

	公里数偏差分类1	fenzi	fenmu	ratio
0	2500-5000公里	56457	316906	0.178151
1	2500公里以内	221241	316906	0.698128
2	5000-7500公里	15433	316906	0.048699
3	7500公里以上	23775	316906	0.075022

规则全量

```
1  tongji1(rule,'天数偏差分类1')
```

	天数偏差分类1	fenzi	fenmu	ratio
0	30-60天	99440	480498	0.206952
1	30天以内	146486	480498	0.304863
2	60-90天	63439	480498	0.132028
3	90天以上	171133	480498	0.356158

```
1  tongji2(rule,'公里数偏差分类1')
```

	公里数偏差分类1	cnt	ratio
1	2500公里以内	412725	0.858953
0	2500-5000公里	35121	0.073093
3	7500公里以上	20658	0.042993
2	5000-7500公里	11994	0.024962

去掉模型VIN以后的规则人群的统计 ¶

```
1  tongji1(ruledf,'天数偏差分类1')
```

	天数偏差分类1	fenzi	fenmu	ratio
0	30-60天	31465	163592	0.192338
1	30天以内	44319	163592	0.270912
2	60-90天	21804	163592	0.133283
3	90天以上	66004	163592	0.403467

```
1  tongji1(ruledf,'公里数偏差分类1')
```

	公里数偏差分类1	fenzi	fenmu	ratio
0	2500-5000公里	12000	163592	0.073353
1	2500公里以内	136011	163592	0.831404
2	5000-7500公里	4798	163592	0.029329
3	7500公里以上	10783	163592	0.065914

主要创新

- 1.了解业务的时候挖掘出VIN隐含的业务逻辑；
- 2.业务宽表的重构，即让数据形式变成更加符合业务逻辑，一个VIN一个进店日期一个样本，也让后续的无论是业务统计还是KPI统计的结果更加精准更加符合实际情况；(如下附表)；
- 3.训练方式的创新；

业务宽表的转换方式的不同，几乎影响了所有统计！！！

1。原先的是不做任何处理与转换的

vin	修理日期	维修站号	委托书号	结算单号	金额	维修类型
vin1	2020/1/1	站号1	1	1	300	常规保养
vin1	2020/1/1	站号1	1	2	300	一般维修
vin1	2020/1/1	站号1	2	1	0	常规保养

2。优化2：涉及到整体统计的时候没问题，但是涉及到维修类型下的统计也会出现不精确的问题

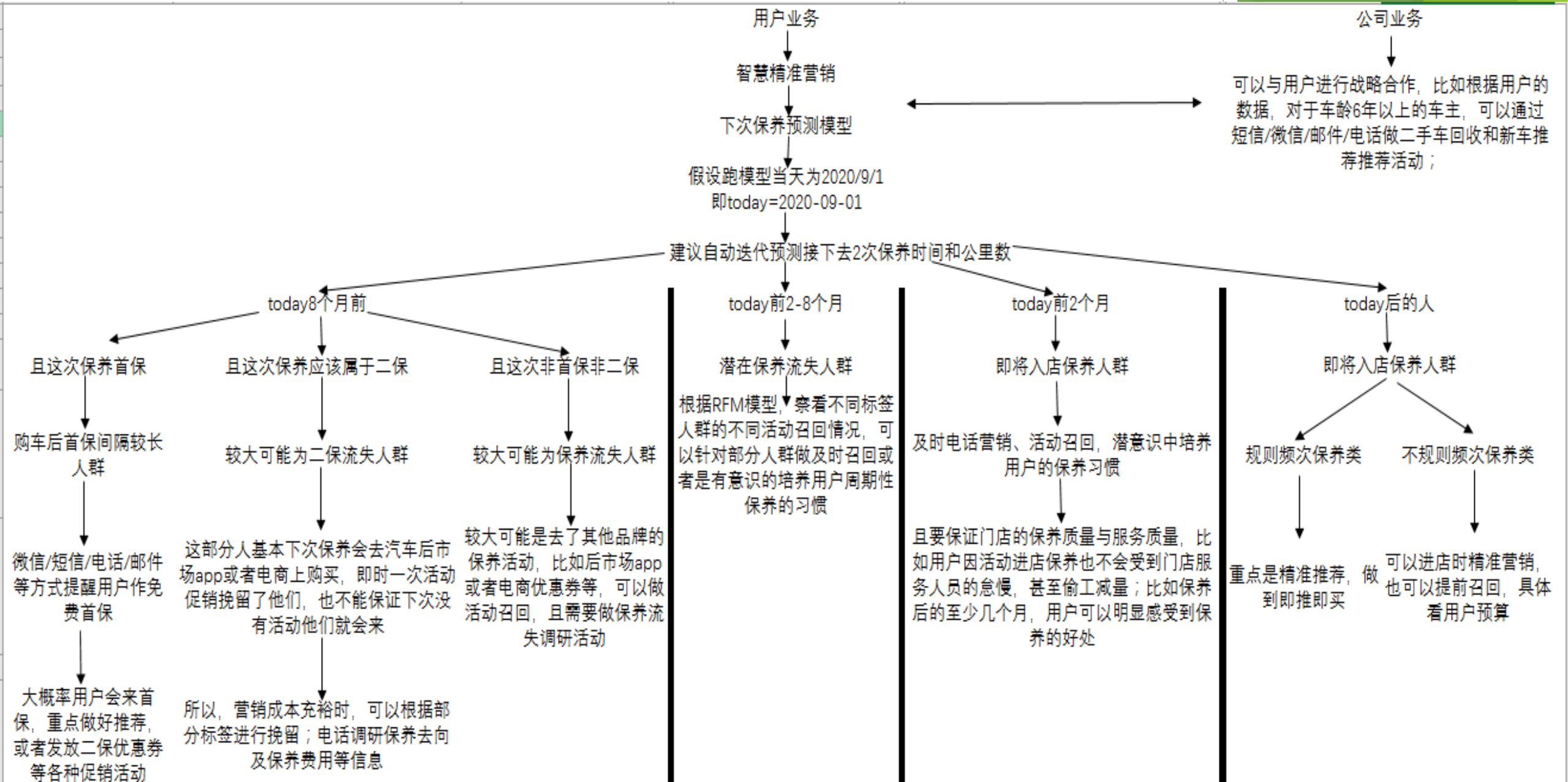
vin	修理日期	维修站号	总金额	维修类型	常规保养	一般维修
vin1	2020/1/1	站号1	600	常规保养,一般维	300	300

不同维修类型下的平均单车产值统计

vin	总金额	台次	平均单车产值	维修类型	
vin1	600	3	200	所有	低估
vin1	300	2	150	常规保养	低估
vin1	300	1	300	一般维修	正常

vin	总金额	台次	平均单车产值	维修类型	
vin1	600	1	600	所有	正常
vin1	300	1	300	常规保养	正常
vin1	300	1	300	一般维修	正常

模型落地方案简介


用户业务
智慧精准营销
下次保养预测模型
假设跑模型当天为2020/9/1
即today=2020-09-01
公司业务
可以与用户进行战略合作，比如根据用户的数据，对于车龄6年以上的车主，可以通过短信/微信/邮件/电话做二手车回收和新车推荐推荐活动；
建议自动迭代预测接下去2次保养时间和公里数
today8个月前
且这次保养首保
购车后首保间隔较长人群
微信/短信/电话/邮件等方式提醒用户作免费首保
大概率用户会来首保，重点做好推荐，或者发放二保优惠券等各种促销活动
且这次保养应该属于二保
较大可能为二保流失人群
这部分人基本下次保养会去汽车后市场app或者电商上购买，即时一次活动促销挽留了他们，也不能保证下次没有活动他们就会来
所以，营销成本充裕时，可以根据部分标签进行挽留；电话调研保养去向及保养费用等信息
且这次非首保非二保
较大可能为保养流失人群
较大可能是去了其他品牌的保养活动，比如后市场app或者电商优惠券等，可以做活动召回，且需要做保养流失调研活动
today前2-8个月
潜在保养流失人群
根据RFM模型，察看不同标签人群的不同活动召回情况，可以针对部分人群做及时召回或者是有意识的培养用户周期性保养的习惯
today前2个月
即将入店保养人群
及时电话营销、活动召回，潜意识中培养用户的保养习惯
且要保证门店的保养质量与服务质量，比如用户因活动进店保养也不会受到门店服务人员的怠慢，甚至偷工减量；比如保养后的至少几个月，用户可以明显感受到保养的好处
today后的人
即将入店保养人群
规则频次保养类
重点是精准推荐，做到即推即买
不规则频次保养类
可以进店时精准营销，也可以提前召回，具体看用户预算

谢 谢

Thank You