CDA就业学院毕业设计

淘宝网店数据分析

第一组
程欣蕊
杨锋勃
马孟麒

案例背景

淘宝网共有约五亿注册用户，日活跃用户超1.2亿，在线商品数量达到10亿。该网站一服饰店主希望对其过去一年内的交易数据进行分析，挖掘出对其忠诚度较高的并且带来高额收益的用户群，从而在接下来的一年中最大限度的提高营销活动的效益。

淘宝店铺基本情况分析

订单状况分析

2013年1月1日——2013年12月31日，该淘宝店铺共发生120,757次订单，其中79.16%的订单交易成功，共计95,596次。

订单状态				
order_state	频数	百分比	累积频数	累积百分比
交易成功	95596	79.16	95596	79.16
交易关闭	25161	20.84	120757	100

淘宝店铺基本情况分析

运送方式分析

通过分析2013年全年该淘宝店铺的购买人员偏好运送方式可以发现，绝大多数人选择快递配送99.9%。其中，95.09%的人选择使用圆通速递，少数购买者选择申通和顺丰快递。

运送方式				
way_express	频数	百分比	累积频数	累积百分比
EMS	33	0.03	33	0.03
快递	120635	99.9	120668	99.93
卖家承担运费	29	0.02	120697	99.95
平邮	60	0.05	120757	100

物流公司				
express_company	频数	百分比	累积频数	累积百分比
EMS	3	0	3	0
百世汇通	1	0	4	0
龙邦速递	1	0	5	0.01
申通E物流	2170	2.24	2175	2.24
顺丰速运	2582	2.66	4757	4.9
圆通速递	92322	95.09	97079	99.99
韵达快运	3	0	97082	100
宅急送	1	0	97083	100
中通速递	2	0	97085	100
频数缺失 = 23672				

淘宝店铺基本情况分析

购买顾客分析

该淘宝店铺的购买人员遍布全国，订单最多的前5个的省市分别为北京（12.29%）、浙江（8.28%）、广东（8.26%）、江苏（8.25%）、上海（6.67%）。

收货地址				
收货地址	频数	百分比	累积频数	累积百分比
Свердловская Первоур	1	0	1	0
安徽省	3362	2.81	3363	2.78
北京	14841	12.29	18204	15.07
福建省	3374	3.12	21978	18.2
甘肃省	914	0.76	22892	18.96
广东省	9983	8.26	32875	27.22
广西	1612	1.32	34487	28.56
贵州省	2112	1.77	36599	30.31
海南省	300	0.23	36899	30.56
海外	1	0	36900	30.56
河北省	4087	3.38	40987	33.94
河南省	3732	3.1	44719	37.03
黑龙江省	1588	1.32	46307	38.35
湖北省	3421	2.83	49728	41.18
湖南省	5332	4.42	55060	45.6
吉打 哥打士打	1	0	55061	45.6
吉林省	1412	1.17	56473	46.77
江苏省	9977	8.25	66450	55.03
江西省	2124	1.78	68574	56.79
辽宁省	3638	3.01	72212	59.8
内蒙古	1799	1.5	74011	61.29
宁夏	644	0.53	74655	61.82
青海省	281	0.22	74936	62.06
山东省	5727	4.75	80663	66.8
山西省	2288	1.89	82951	68.69
陕西省	3515	2.91	86466	71.6
上海	8058	6.67	94524	78.28
四川省	4651	3.85	99175	82.13
台湾省	1904	1.57	101079	83.7
天津	3693	3.06	104772	86.76
西藏	62	0.04	104834	86.81
香港	86	0.07	104920	86.89
新疆	630	0.52	105550	87.41
云南省	2709	2.25	108259	89.65
浙江省	9990	8.28	118249	97.92
重庆	2508	2.08	120757	100

业务目标

该店铺店主准备于2014年农历年期间做一促销活动，希望挖掘出最有可能在活动期间购买产品，并且带来高额收入的用户群。

分析问题

- 挖掘出最有可能响应促销活动的用户。
- 假如响应用户购买该店铺产品，哪些人可能花费的金额更高。
- 获取最有可能在促销活动中带来高额购买的用户群。
- 对响应用户群进行分类，针对不同类用户实施成本不同的营销活动。

分析思路

- 定义客户流失情况。
- 检验变量的分布情况，做相应变换使数据尽量无偏。
- 分析变量间的相关性，判断出关键变量并建立模型。
- 使用线性回归对客户消费金额大小进行分类和预测。
- 使用logistics回归，判断客户是否响应活动，得到重要客户。
- 使用聚类分析，将响应活动客户分类，针对不同类客户采取相应营销手段。

原数据整理过程（将订单数据转换为客户数据）

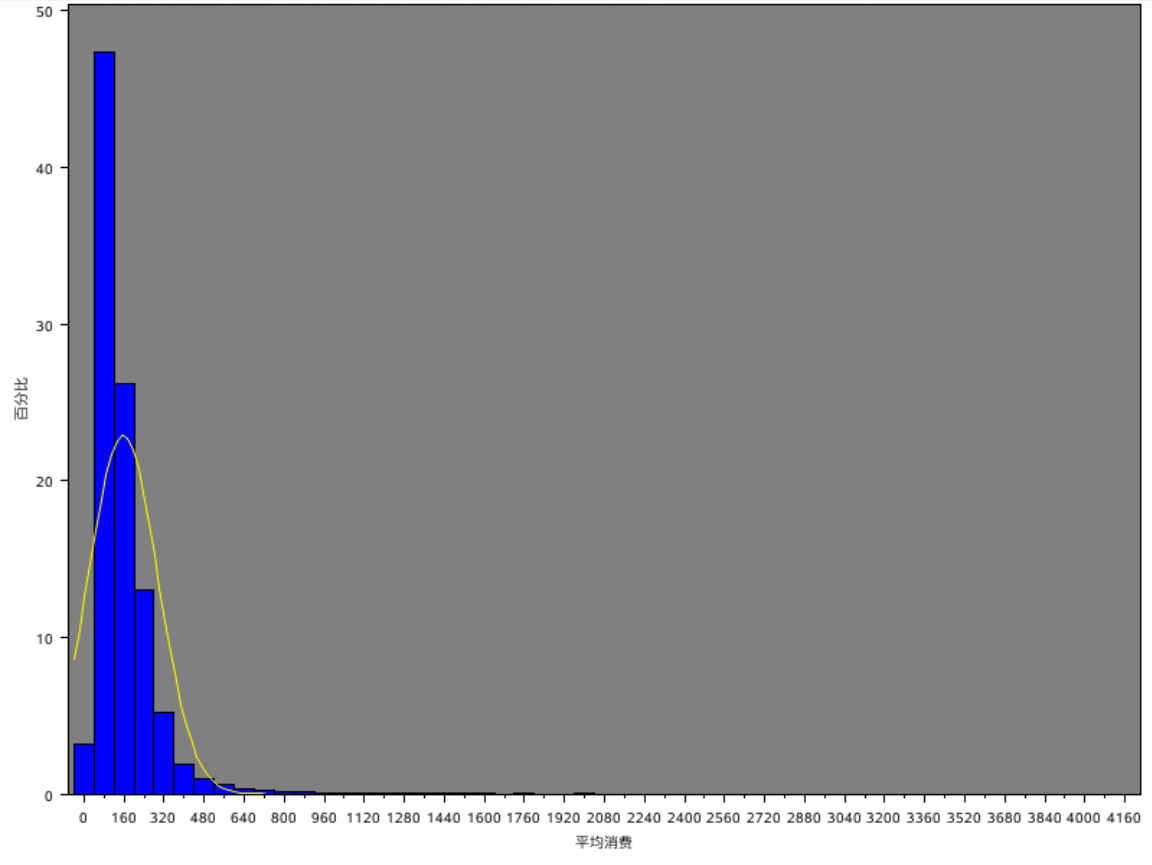
- 本案例选择交易成功的数据为研究对象
- 将购买次数、花费金额、购物量分季度拆分成新变量
- 买家留言字数无实际意义，可作为备选变量
- 不同买主可以用买家会员名或者买家支付宝账号识别，我们用买家会员名表示
- 将买家应付货款、买家应付邮费、总金额、实际支付金额这4个变量用买家实际支付金额代表

总结本案例可用变量名

变量名	含义
pay_id	买家支付宝账号
customer_id	买家会员名
order_remarks	订单备注
goods_kind	宝贝种类
order_state	订单状态
customer_says	买家留言
recent_buy	最近一次消费时间
freq_buy	窗口期内购买总次数
avg_money	平均消费
goods_total	总购物量
Response	客户是否响应
ln_max_customer_says	买家留言平均字数
avg_buy_gt2	平均购物数量
ln_max_order_remarks	订单是否备注
season_freq_lab	购物频次
buy_COUNT_QTR_1	第一季度购买次数
buy_COUNT_QTR_2	第二季度购买次数
buy_COUNT_QTR_3	第三季度购买次数
buy_COUNT_QTR_4	第四季度购买次数
season_expense_lab	总花费
expense_QTR_1	第一季度花费额
expense_QTR_2	第二季度花费额
expense_QTR_3	第三季度花费额
expense_QTR_4	第四季度花费额
season_goodscount_lab	宝贝总数量

检验变量数据是否无偏

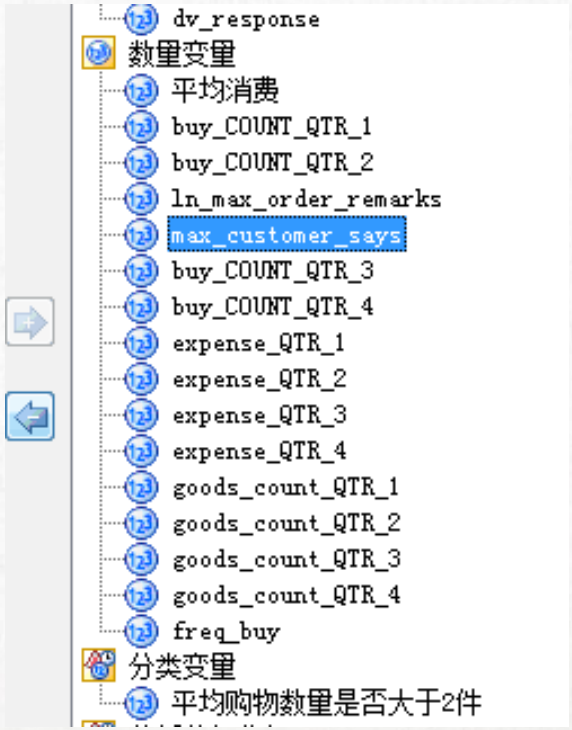
对自变量逐一进行分析，观察数据是否正态分布，如正态性不明显，对其采取对数变换。



变量选取

NAME	dv_response	avg_buy_gt	avg_money	buy_COUNT_	buy_COUNT_	buy_COUNT_
max_order_remarks	0.269632901	0.3017619	0.1590762	0.1178664	0.128107	0.127984
avg_buy_gt2	0.25690665	1	0.5101522	0.0272839	0.0563201	0.063993
avg_money	0.229576582	0.5101522	1	0.0134154	0.0196269	0.0534426
ln_avg_money	0.214012405	0.6227557	0.7889512	0.0315247	0.0274623	0.0369712
expense_QTR_2	0.171932663	0.158444	0.140141	0.4250786	0.7134613	0.5504923
goods_count_QTR_2	0.166763085	0.1684634	0.1182153	0.3907359	0.6616095	0.5182502
ln_max_order_remarks	0.166439909	0.2959274	0.1103821	0.0670735	0.0798951	0.0759284
expense_QTR_3	0.161729912	0.1452273	0.1661236	0.3354038	0.4806808	0.6741453
goods_count_QTR_3	0.153135968	0.1487048	0.138102	0.302456	0.4431257	0.630761
buy_COUNT_QTR_3	0.152040996	0.063993	0.0534426	0.5474249	0.8172924	1
buy_COUNT_QTR_2	0.131298212	0.0563201	0.0196269	0.6102852	1	0.8172924
freq_buy	0.117500993	0.0635694	0.0153862	0.6842927	0.9341199	0.9084241
buy_COUNT_QTR_4	0.107808009	0.1068885	0.1204326	0.5148573	0.7572338	0.7648272
expense_QTR_1	0.100368593	0.1185247	0.1304344	0.7895033	0.3477153	0.3046383
goods_count_QTR_4	0.098255536	0.1838982	0.2230393	0.2485479	0.3351327	0.3718897
buy_COUNT_QTR_1	0.092079466	0.0272839	0.0134154	1	0.6102852	0.5474249
expense_QTR_4	0.08732143	0.1678087	0.2530497	0.2700257	0.3601176	0.3941044
ln_max_customer_says	0.082407408	0.1158996	0.062921	0.038374	0.0445121	0.0661366
max_customer_says	0.066009491	0.0257036	0.0227337	0.037796	0.0267738	0.0265757
goods_count_QTR_1	0.015330896	0.0131275	0.0144788	0.1687042	0.0186976	0.0155608

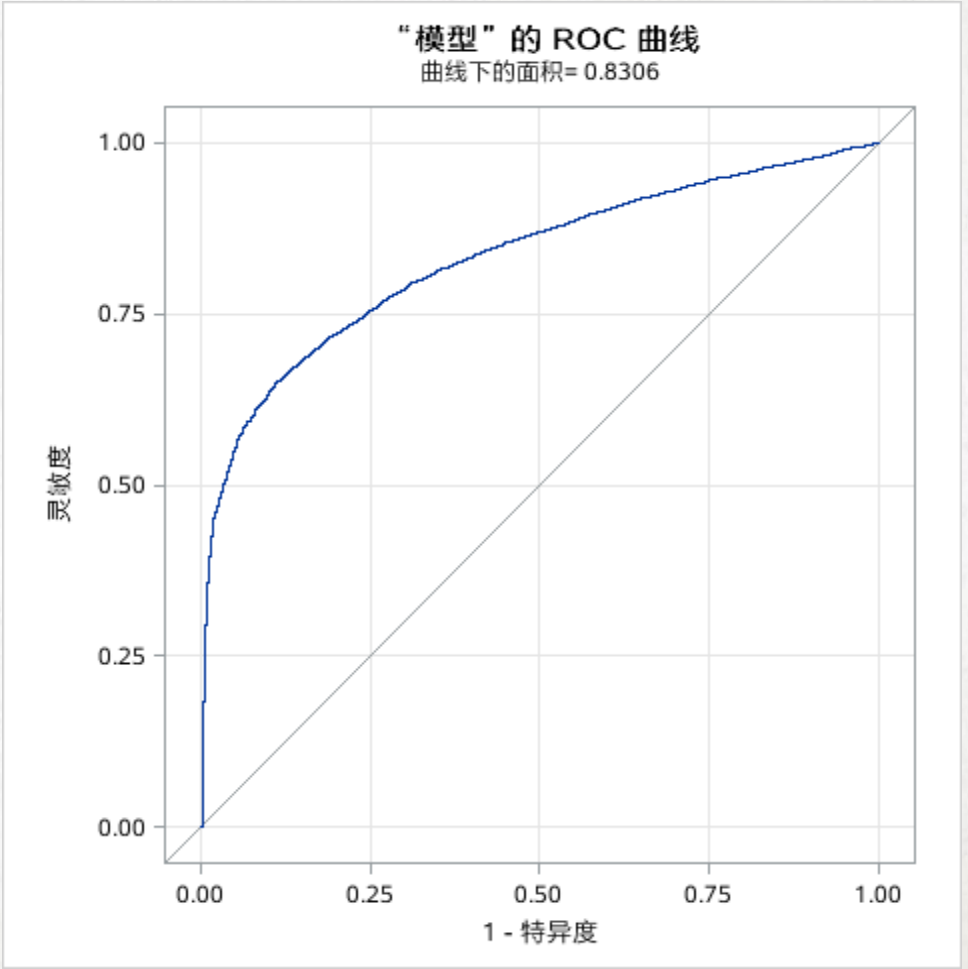
根据变量相关性选出关键变量建立模型，关键变量如下：



LOGISTICS回归过程

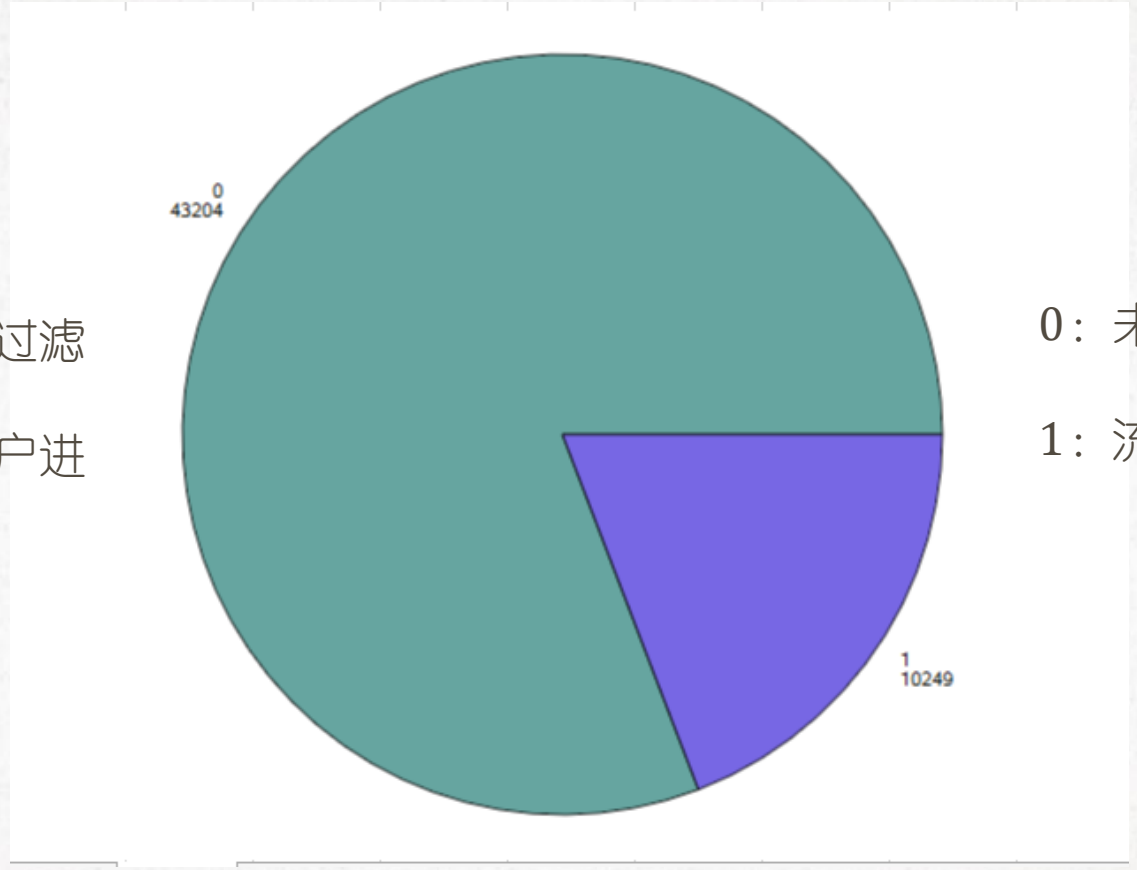
预测概率和观测响应的关联		
一致部分所占百分比	83.1Somers D	0.661
不一致部分所占百分比	16.9Gamma	0.661
结值百分比	0Tau-a	0.228
对	38755828c	0.831

根据logistics回归检验模型是否符合本案例的实际情况



客户流失情况

根据客户流失情况，筛选过滤掉流失用户，对未流失用户进一步进行logistics分析。



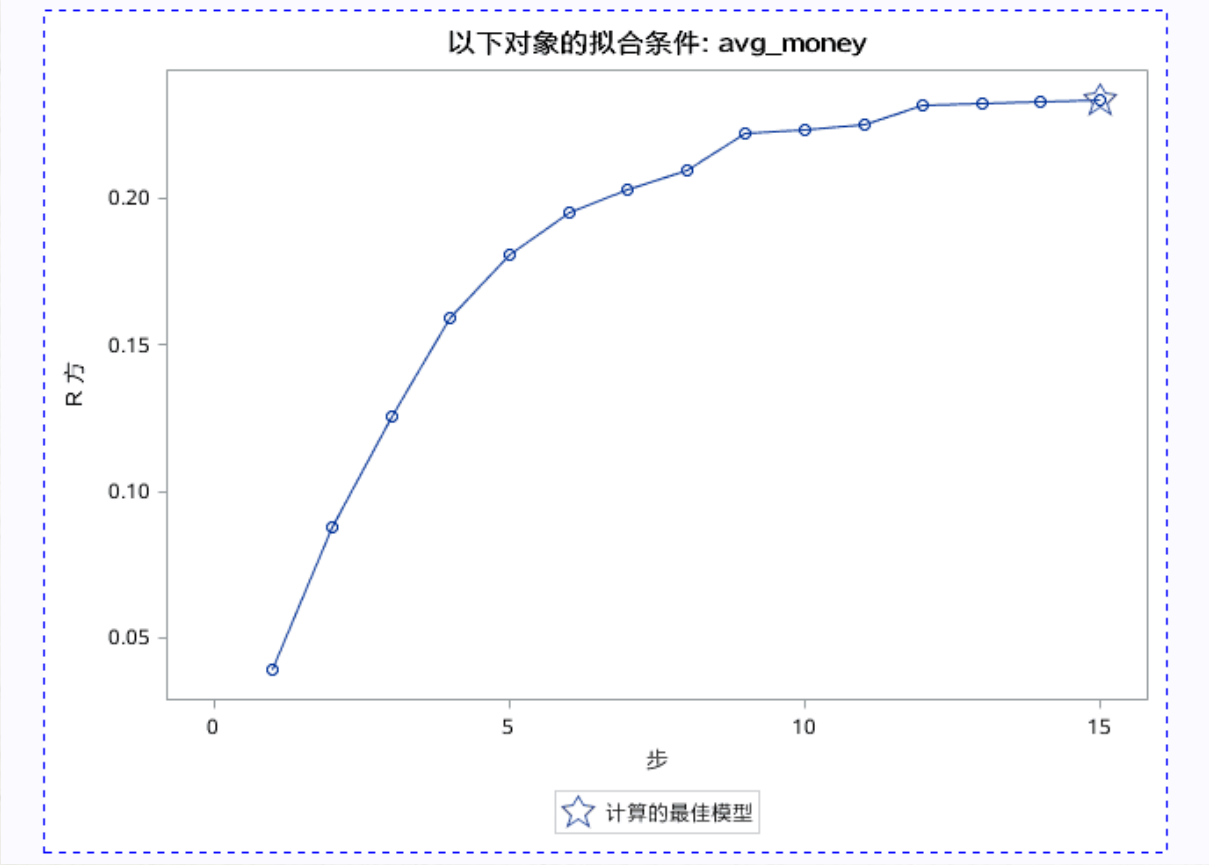
0：未流失用户

1：流失用户

线性回归变量选取

NAME	avg_money	DAYS_FROM_LAST_	avg_buy_gt2	avg_time	buy_COUNT_QTR_	buy_COUNT_QTR_2
avg_buy_gt2	0.51015223	-0.193463711	1	-0.30773	0.027283851	0.056320125
expense_QTR_4	0.253049667	-0.247753909	0.16780874	-0.21048	0.270025661	0.360117638
goods_count_QTR_4	0.223039315	-0.2431341	0.183898163	-0.21462	0.24854791	0.335132665
recent_buy	0.176324901	-0.999997888	0.193447076	-0.2495	-0.243432911	-0.037387373
DAYS_FROM_LAST_TO_END	-0.176321464	1	-0.193463711	0.249539	0.243436969	0.037385602
expense_QTR_3	0.16612364	-0.164728289	0.145227341	-0.20146	0.33540383	0.480680759
max_order_remarks	0.15907624	-0.139220147	0.301761864	-0.37437	0.117866381	0.128107003
expense_QTR_2	0.140140993	0.02039149	0.158443982	-0.23224	0.425078634	0.713461325
goods_count_QTR_3	0.138101955	-0.09310157	0.148704796	-0.18771	0.302456036	0.443125718
avg_time	-0.136583679	0.249538896	-0.307725308	1	-0.152683461	-0.165988881
expense_QTR_1	0.13043441	0.166879012	0.118524701	-0.17742	0.769503262	0.347715336
buy_COUNT_QTR_4	0.120432614	-0.131325867	0.106888453	-0.22448	0.514857342	0.757233819
goods_count_QTR_2	0.118215268	0.015458057	0.168463411	-0.22045	0.390735939	0.661609537
buy_COUNT_QTR_3	0.053442637	-0.089921155	0.063992988	-0.11726	0.547424862	0.817292362
max_customer_says	0.022733695	-0.0034452	0.025703647	-0.08219	0.03779605	0.026773797
buy_COUNT_QTR_2	0.0196269	0.037385602	0.056320125	-0.16599	0.610285156	1
freq_buy	0.015386194	-0.081940591	0.063569396	-0.20924	0.684292667	0.934119947
goods_count_QTR_1	0.014478802	0.014562619	0.01312752	-0.01889	0.168704241	0.018697641
buy_COUNT_QTR_1	0.013415446	0.243436969	0.027283851	-0.15268	1	0.610285156

线性回归拟合结果



	Label1	cValue1	Label2	cValue2
1	均方根误差	71.96475	R 方	0.4684
2	因变量均值	143.15093	调整 R 方	0.4682
3	变异系数	50.27194		

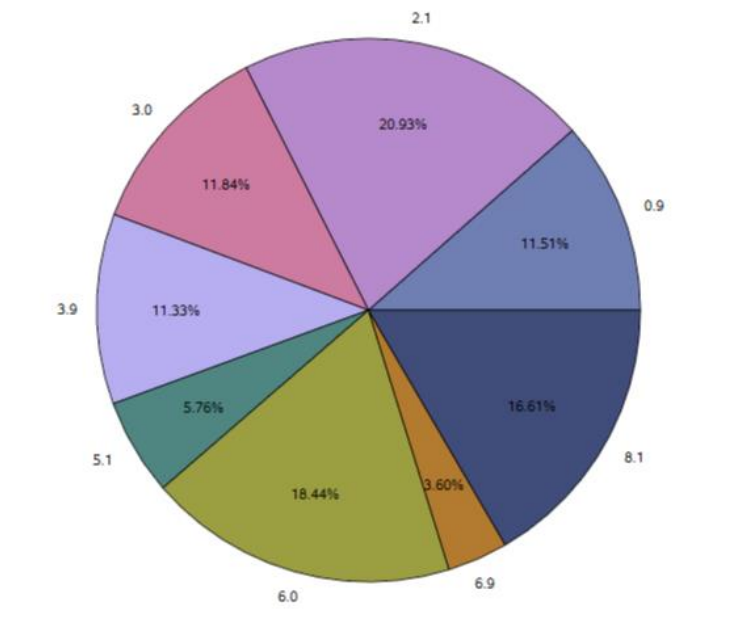
线性回归分组

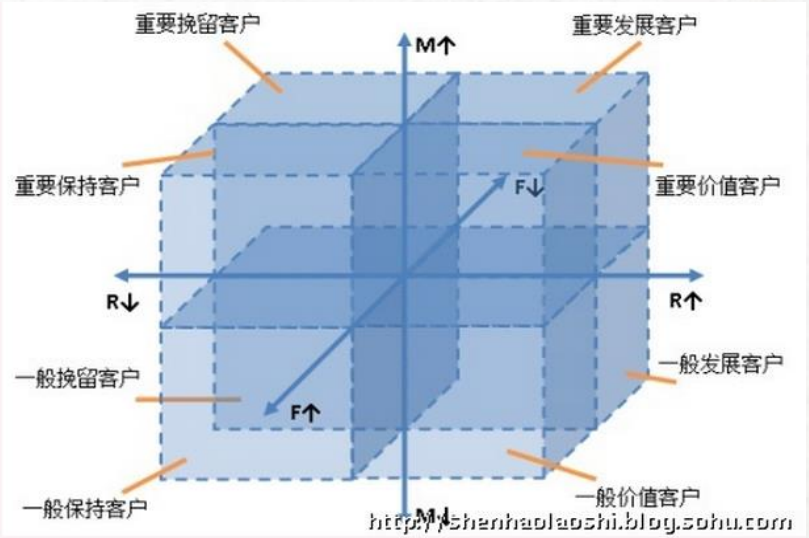
- 对于高价值目标客户挖掘项目，可以综合运用逻辑回归的响应预测模型和线性回归的购买金额预测模型
- 根据模型分数的高低，为客户挑选一定预算条件下的最优质用户，从而帮助其实现商业目标

响应 Decile \ 消费金额 Decile	1	2	3	4	5	6	7	8	9	10
1	19,624	13,217	10,274	8,042	6,449	4,694	2,939	1,184	1,185	1,186
2	12,543	10,565	9,115	8,237	7,671	6,818	5,965	5,112	4,259	3,406
3	10,070	9,240	8,336	7,737	7,349	7,416	7,483	7,550	7,617	7,684
4	8,043	8,514	8,018	7,745	7,484	7,287	7,090	6,893	6,696	6,499
5	6,195	7,686	7,599	7,514	7,515	7,516	7,517	7,518	7,519	7,520
6	4,296	6,390	7,283	7,492	7,465	7,297	7,297	7,297	7,297	7,297
7	3,048	4,865	6,407	7,166	7,233	8,143	9,054	9,964	10,874	11,784
8	2,087	3,604	5,011	6,130	7,074	8,536	8,536	9,510	10,241	10,972
9	1,408	2,481	3,650	4,788	5,785	8,928	12,072	15,215	18,359	21,502
10	1,272	2,014	2,940	3,728	4,555	9,321	14,087	18,853	23,619	28,385

聚类分析结果

对2013年交易客户进行聚类分析，大致将客户分为八类。





营销手段总结

- 建立旺旺群，对重要客户和一般客户做好区分，采取不同的操作方式。
- 根据重要客户的消费金额排序，对排序靠前的第一时间推送新款产品。
- 对重要价值客户进行关联产品推荐和商品绑定。
- 对于重要挽留客户和重要保持客户，实行会员优惠，增加用户口碑。
- 引导一般客户进行二次消费，采取买A送B、团购等营销手段。

谢 谢！

欢迎各位老师提出宝贵意见！