Evaluation and Prediction of Cell Phone Sales Based on Various Techniques

对手机销量的多方法预测及评估

曹凌微 钱成 田肇阳

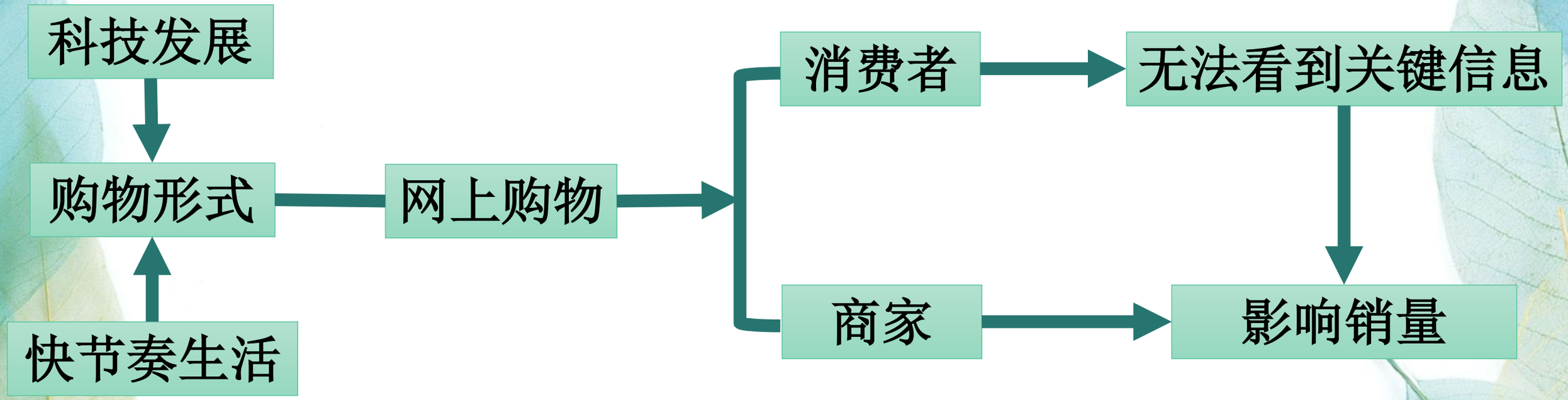
清华大学附属中学

指导教师：吴昊 王殿军

01 研究背景

研究目的与意义

研究背景



研究目的

以往成果

1 灰色关联度分析
——王法涛 2013

2 C2C模型
——薛有志 2012

3 BP神经网络拟合
——马艳丽 2014

深度探究

1 对销量的影响因素及大小

2 每个因素中提高销量的特征

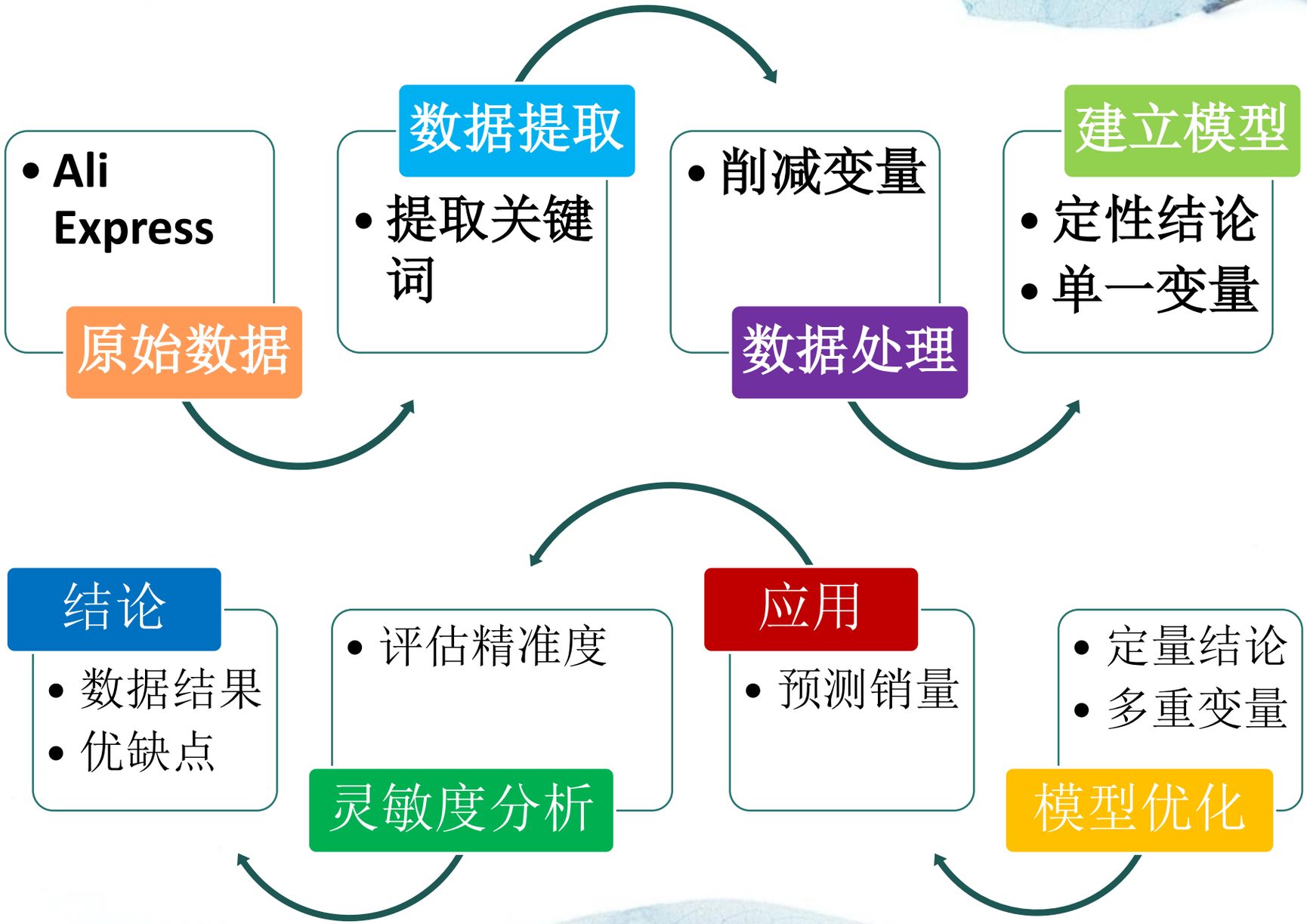
3 预测销量

发现不足

没有系统全面地分析影响因素

没有明确的得出高销量的特征

流程图


原始数据
• Ali Express
数据提取
• 提取关键词
数据处理
• 削减变量
建立模型
• 定性结论
• 单一变量
模型优化
• 定量结论
• 多重变量
应用
• 预测销量
灵敏度分析
• 评估精准度
结论
• 数据结果
• 优缺点

02 模型假设

以对问题作出适当的简化

假设

$$\left\{\begin{array}{l} 转化率 = \dfrac{成交量}{点击量} \\ 点击率 = \dfrac{点击量}{浏览量} \end{array}\right\} \quad 点击率 \times 转化率 = 成交率 \longrightarrow \textbf{成交量}$$

03 数据提取与处理

找出最关键独立变量

信息熵算法

利用信息熵算法计算各个独立变量相对于点击率与转化率的信息增益，以定量探究各个因素的重要性

$$E(X) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

以ROM为例，进行分类分组计算独立变量的信息熵

ROM(GB)	Group number in Category Click Rate	1	2	3	4	5	Information entropy
	2	0	3	2	1	0	1.459148
	4	5	6	13	7	0	1.89366
	8	64	38	63	54	8	2.122787
	16	110	89	132	143	36	2.205866
	32	64	44	82	63	29	2.242444
	64	83	38	62	49	16	2.160525
	128	3	4	5	7	0	1.931295
	256	0	0	1	0	0	0

信息熵算法

利用信息熵计算每个独立变量的信息增益进行横向比较，得出相对最重要参量。

$$IGain(Category\ Click\ Rate, Rom) = E(global) - \sum Information\ entropy \times Possibility$$

	Category Click Rate	Category Convert Rate
Global information entropy	2.200779	2.081891

	Sum of the products	IGain
ROM	2.174619842	0.026159369

相对于点击率独立变量重要性排名	Information Gain
Comment Count	0.732792417
Good Comment Count	0.680453664
Search Count	0.392386753
Score	0.173242295
Brand	0.124112475
Is Gallery Featured	0.060358001
Battery Capacity(mAh)	0.050232189
RAM(G)	0.031072544

相对于转化率独立变量重要性排名	Information Gain
Comment Count	0.950131659
Good Comment Count	0.910616696
Search Count	0.631528548
Score	0.288394004
Brand	0.261397755
Is Gallery Featured	0.220147548
Battery Capacity(mAh)	0.102065066
Highest camera resolution	0.067310002

主成分分析

主成分分析对数据再次进行处理，减少参量而尽量多保留原始数据信息

$$X=\begin{bmatrix} x_{11} & \cdots & x_{1l} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{ml} \end{bmatrix} \qquad \overline{x_j}=\sum_{t=1}^{i}\frac{x_{tj}}{i},\sigma_j=\sqrt{\sum_{i=1}^{n}\frac{\left(\overline{x_j}-x_{ij}\right)^2}{n-1}},x_{ij}{}^{*}=\frac{x_{ij}-\overline{x_j}}{\sigma_j}$$

$$\frac{\lambda_i}{\sum_{k=1}^{q}\lambda_k}\ (i=1,2,\ldots,p) \qquad \frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{q}\lambda_k}\ (i=1,2,\ldots,p)$$

利用上述公式对数据进行标准化，并计算得主成分回归特征向量，计算总贡献率，选取前14个主成分（其贡献率超过80%）

$$z_1=a_{11}x_1+a_{21}x_2+a_{31}x_3+a_{41}x_4+a_{51}x_5+\cdots+a_{26\,1}x_{26}$$

$$z_2=a_{12}x_1+a_{22}x_2+a_{32}x_3+a_{42}x_4+a_{52}x_5+\cdots+a_{26\,2}x_{26}$$

…

$$z_{14}=a_{1\,14}x_1+a_{2\,14}x_2+a_{3\,14}x_3+a_{4\,14}x_4+a_{5\,14}x_5+\cdots+a_{26\,14}x_{26}$$

04 模型建立

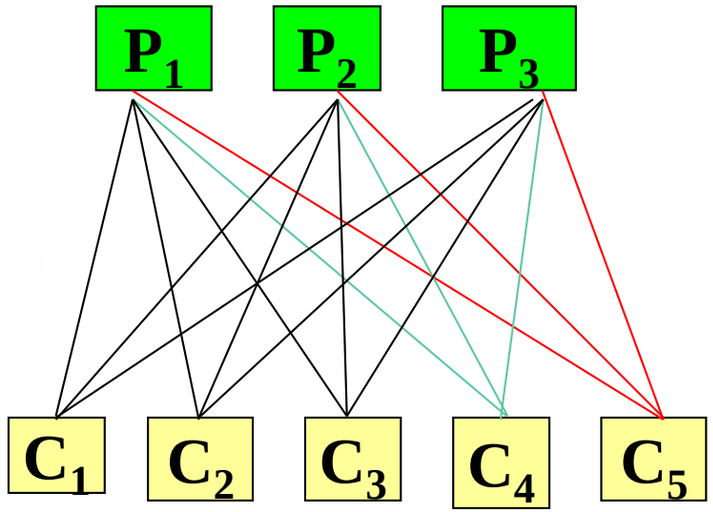
定性结论与模型深度分析

基础统计

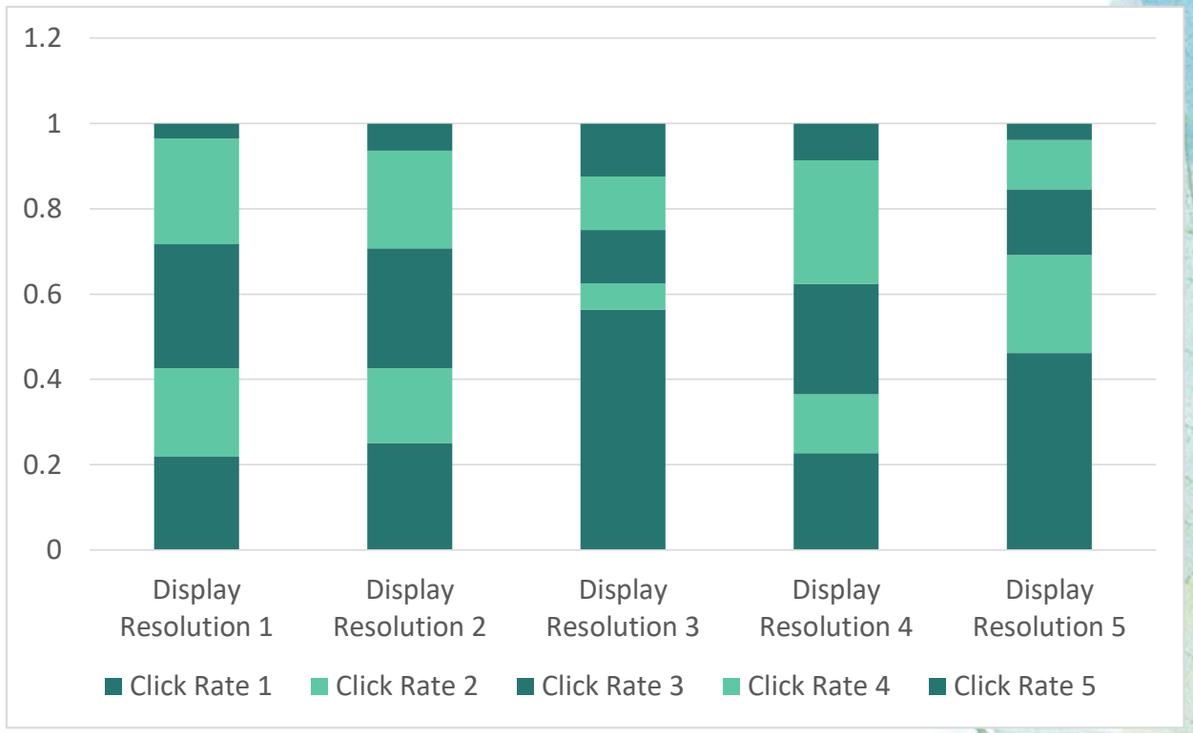
权重确定方法

$$Aw = \lambda w$$

Target Layer (RAM)



Scheme Layer(Click Rate)



中等的屏幕清晰度手机中销售量较高和较低的占比都较大

线性回归

Click Rate	Convert Rate
0.115124	0.162528

KNN 算法

$$d(x,y) = \sqrt{(x-y)\Sigma^{-1}(x-y)^T}$$

05 模型优化

优化算法与模型定量结论

主成分回归

$$y_n{}^* = \beta_1{}'z_1 + \beta_2{}'z_2 + \beta_3{}'z_3 + \cdots + \beta_{14}{}'z_{14}(n \in \{n \in N^* | n \leq 2\})$$

$$y_n = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_{26} x_{26}(n \in \{1, 2\})$$

Click Rate	Convert Rate
0.805032	0.826614

贝叶斯判别

	高销量	低销量
高清晰度	3	7
低清晰度	1	9

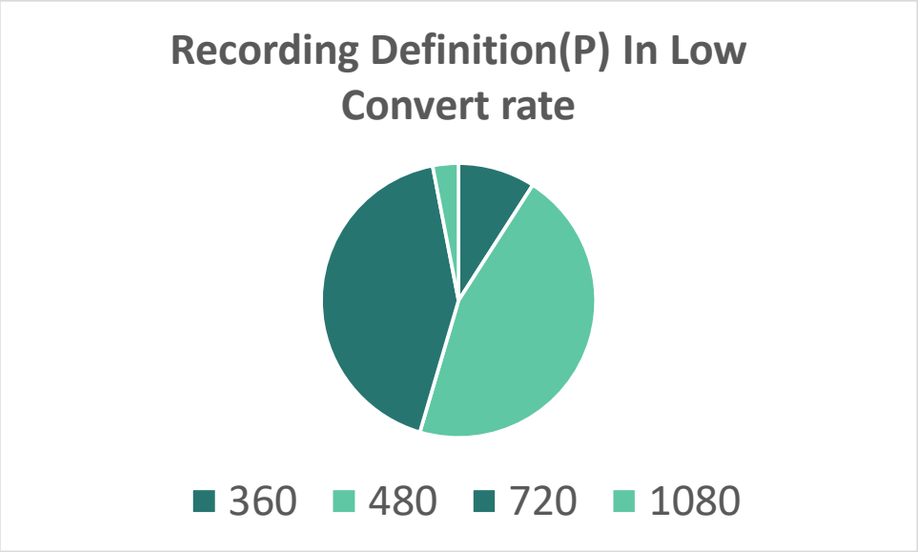
$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{\Sigma P(A \mid B_i)P(B_i)}$$

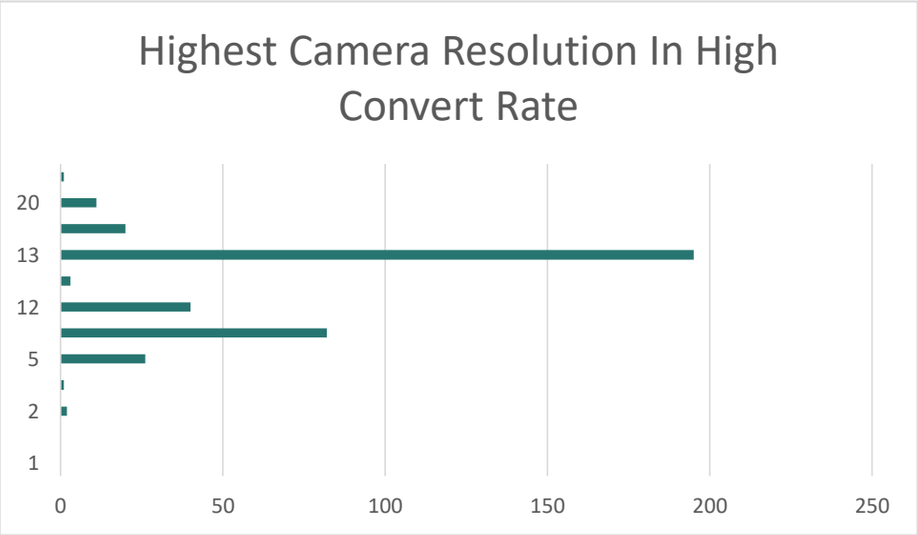
$$\boldsymbol{P(G_l \mid x_0) = \frac{p_l f_l(x_0)}{\sum p_j f_j(x_0)} = \max_{1 \le i \le k} \frac{p_i f_i(x_0)}{\sum p_j f_j(x_0)}}$$

$$\boldsymbol{ECM = \sum_{i=1}^{k} p_i \sum_{j \neq 1} C\left(\frac{j}{i}\right) P\left(\frac{j}{i}\right)}$$

$$p(j/i) = P(X \in D_j / G_i) = \int_{D_j} f_i(x)dx \qquad i \neq j$$

贝叶斯判别



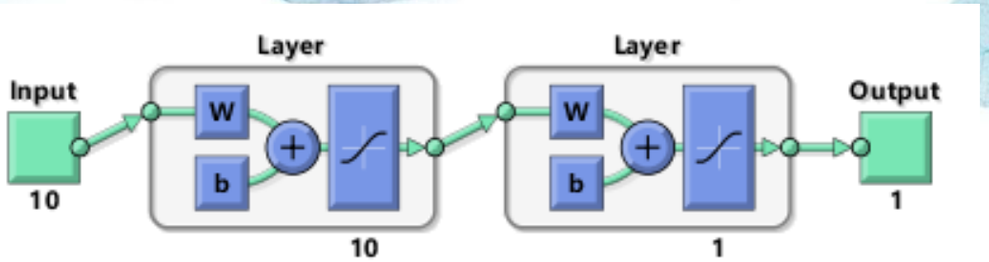


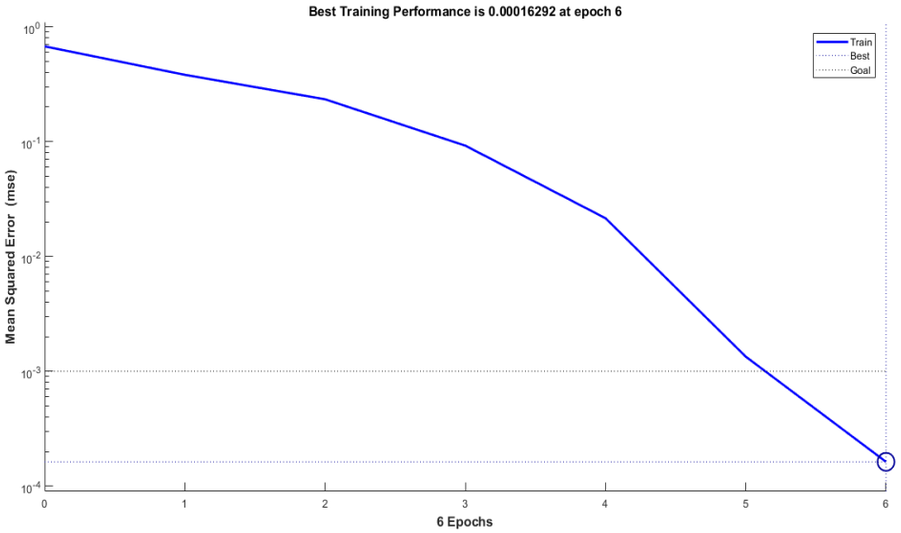
RAM to Click rate (the smaller the category is, the lower the click rate is)

	Category 1	Category 2	Category 3	Category 4
0.125	2	0	0	0
0.5	7	0	0	1
1	49	0	1	7
1.5	1	0	0	0
2	91	0	15	25
3	41	1	64	19
4	1	13	68	2
6	0	11	22	0
8	0	0	1	0

低清晰度有一个较低的转化率。高相机清晰度的手机展示出了更好的销售情况。

BP神经网络





O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
					ClickRate mean							ConvertRate	
0.02078	0.011384	0.008857	0.004261	0.021461	0.013348		0.014479	-0.00362	0.02533	0.022221	0.007262	0.013133	
0.020471	0.002053	0.005953	0.002532	-0.00187	0.005829		0.016294	0.001955	0.009609	0.013991	0.007917	0.009953	
0.021792	0.020835	0.013656	0.017043	0.010632	0.016791		0.016819	0.009979	0.020041	0.002321	0.021112	0.014054	
0.015907	0.013295	0.005368	0.002473	0.001955	0.0078		0.017374	0.007013	0.009318	0.010874	0.007137	0.010343	
0.012225	0.020101	0.01652	0.019162	0.021755	0.017953		0.016328	0.008709	0.013901	0.014945	0.017033	0.014183	
0.020453	0.010045	0.006676	0.002474	-0.00278	0.007373		0.013149	0.007797	0.007339	0.013637	0.008797	0.010144	
0.015864	0.010883	0.006219	0.002389	-0.00382	0.006307		0.015677	0.009632	0.007558	0.010519	0.007801	0.010238	
0.016053	0.003646	0.006422	0.0025	-0.00062	0.0056		0.016442	0.00814	0.008473	0.011153	0.007686	0.010379	
0.021603	0.021945	0.016347	0.019545	0.016197	0.019127		0.015357	0.019486	0.012833	0.01432	0.012956	0.01499	
0.018423	0.012065	0.007084	0.002474	-0.00151	0.007707		0.013997	0.007761	0.007401	0.013494	0.008178	0.010166	
0.01248	0.0216	0.012888	0.017519	0.017937	0.016485		0.015002	0.012329	0.016478	0.005335	0.012801	0.012389	
0.016136	0.010158	0.006237	0.002339	0.000183	0.007011		0.015353	0.009618	0.007716	0.011093	0.008007	0.010357	
0.015526	0.008238	0.006689	0.002239	0.002612	0.007061		0.014566	0.011249	0.006636	0.006922	0.004731	0.008821	
0.017688	0.010685	0.008449	0.005342	0.007715	0.009976		0.013769	0.008002	0.002354	0.021218	0.006657	0.0104	
0.024372	0.022045	0.016973	0.019874	0.018377	0.020328		0.014417	0.027674	0.015893	0.014049	0.012057	0.016818	
0.005814	0.020225	0.00924	0.00513	0.011588	0.010399		0.01433	0.010461	0.012536	0.009471	0.006633	0.010686	
0.023505	0.012874	0.016141	0.019691	0.01486	0.017414		0.01492	0.024921	0.017553	0.01721	0.012591	0.017439	
0.021031	0.003062	0.006779	0.002463	-0.00594	0.005479		0.016598	-0.00305	0.011454	0.017018	0.008545	0.010112	
0.007895	0.019178	0.005459	0.002725	0.006349	0.008321		0.014353	-0.00967	-0.00093	0.019134	0.006623	0.005903	
0.004709	0.011095	0.006211	0.002544	0.004508	0.005813		0.018337	0.007168	6.45E-05	0.010817	0.004856	0.008249	
0.020242	0.001131	0.006219	0.002481	-0.00251	0.005513		0.015004	0.007746	0.008241	0.012762	0.008476	0.010446	
0.015613	0.014638	0.006331	0.002231	0.00345	0.008452		0.014434	0.011731	0.006555	0.007063	0.005558	0.009068	
0.022037	0.009821	0.006919	0.001946	0.000596	0.008264		0.02166	0.003533	0.007823	0.010596	0.007032	0.010129	
0.016142	0.010266	0.017648	0.019628	0.022345	0.017206		0.015378	0.016143	0.022779	-0.00053	0.019856	0.014726	
-0.00863	0.017637	0.019993	0.017025	0.030315	0.015269		0.016931	0.016732	0.012842	-0.00257	0.009733	0.010734	

10层的层数用时较少，同时**结果效果良好**，训练表现也在逐渐提升。

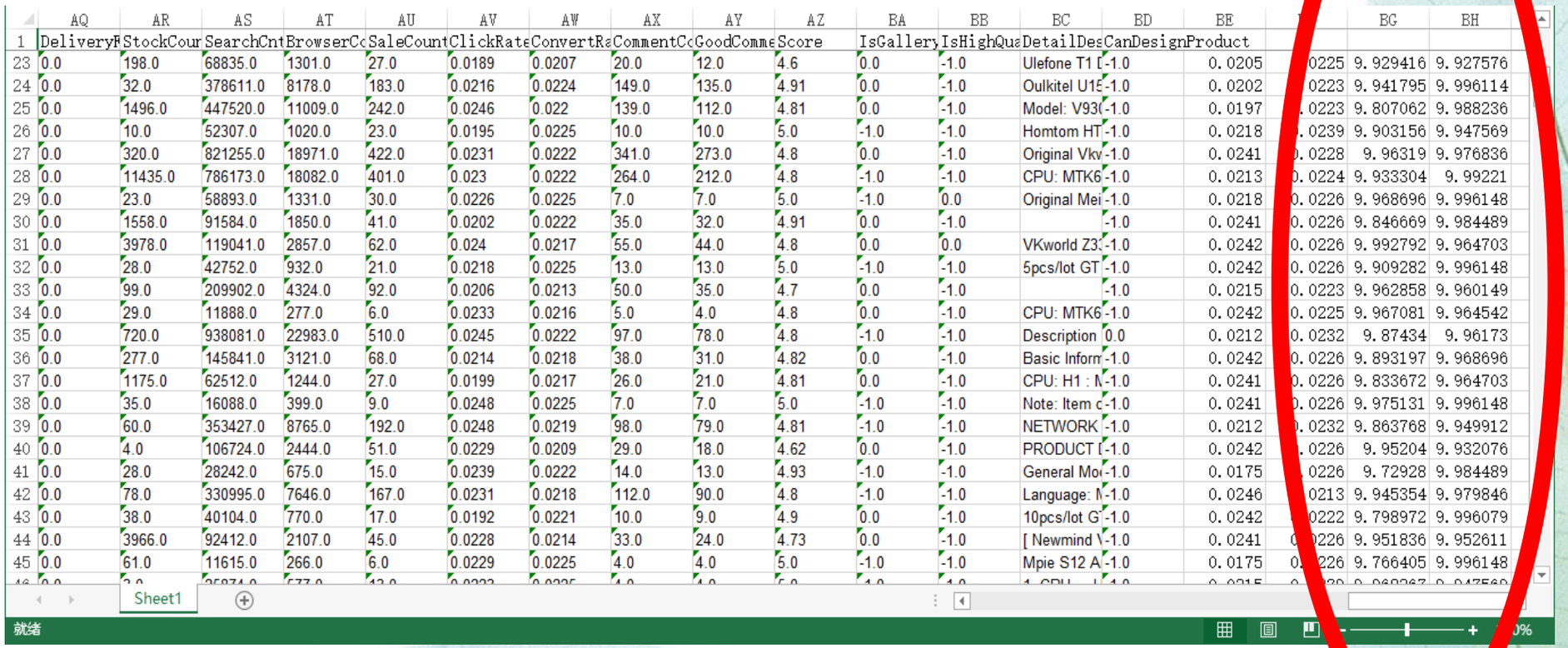
表中部分数据的**误差低于1%**

XGBoosting 算法

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i)$$

年龄	月购物金额	经常在百度知道提问还是回答
14岁	小于1000元	提问
16岁	小于1000元	回答
24岁	大于1000元	提问
26岁	大于1000元	回答

组合几个弱学习器成为强学习器，通过最小化损失函数和误差。



	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BG	BH
1	DeliveryR	StockCour	SearchCnt	BrowserCc	SaleCount	ClickRate	ConvertRa	CommentCc	GoodComme	Score	IsGallery	IsHighQua	DetailDes	CanDesign	Product		
23	0.0	198.0	68835.0	1301.0	27.0	0.0189	0.0207	20.0	12.0	4.6	0.0	-1.0	Ulefone T1 [	-1.0	0.0205	9.929416	9.927576
24	0.0	32.0	378611.0	8178.0	183.0	0.0216	0.0224	149.0	135.0	4.91	0.0	-1.0	Oulkitel U15	-1.0	0.0202	9.941795	9.996114
25	0.0	1496.0	447520.0	11009.0	242.0	0.0246	0.022	139.0	112.0	4.81	0.0	-1.0	Model: V93(	-1.0	0.0197	9.807062	9.988236
26	0.0	10.0	52307.0	1020.0	23.0	0.0195	0.0225	10.0	10.0	5.0	-1.0	-1.0	Homtom HT	-1.0	0.0218	9.903156	9.947569
27	0.0	320.0	821255.0	18971.0	422.0	0.0231	0.0222	341.0	273.0	4.8	0.0	-1.0	Original Vkv	-1.0	0.0241	9.96319	9.976836
28	0.0	11435.0	786173.0	18082.0	401.0	0.023	0.0222	264.0	212.0	4.8	-1.0	-1.0	CPU: MTK6	-1.0	0.0213	9.933304	9.99221
29	0.0	23.0	58893.0	1331.0	30.0	0.0226	0.0225	7.0	7.0	5.0	-1.0	0.0	Original Mei	-1.0	0.0218	9.968696	9.996148
30	0.0	1558.0	91584.0	1850.0	41.0	0.0202	0.0222	35.0	32.0	4.91	0.0	-1.0		-1.0	0.0241	9.846669	9.984489
31	0.0	3978.0	119041.0	2857.0	62.0	0.024	0.0217	55.0	44.0	4.8	0.0	0.0	VKworld Z3	-1.0	0.0242	9.992792	9.964703
32	0.0	28.0	42752.0	932.0	21.0	0.0218	0.0225	13.0	13.0	5.0	-1.0	-1.0	5pcs/lot GT	-1.0	0.0242	9.909282	9.996148
33	0.0	99.0	209902.0	4324.0	92.0	0.0206	0.0213	50.0	35.0	4.7	0.0	-1.0		-1.0	0.0215	9.962858	9.960149
34	0.0	29.0	11888.0	277.0	6.0	0.0233	0.0216	5.0	4.0	4.8	0.0	-1.0	CPU: MTK6	-1.0	0.0242	9.967081	9.964542
35	0.0	720.0	938081.0	22983.0	510.0	0.0245	0.0222	97.0	78.0	4.8	-1.0	-1.0	Description	0.0	0.0212	9.87434	9.96173
36	0.0	277.0	145841.0	3121.0	68.0	0.0214	0.0218	38.0	31.0	4.82	0.0	-1.0	Basic Inforn	-1.0	0.0242	9.893197	9.968696
37	0.0	1175.0	62512.0	1244.0	27.0	0.0199	0.0217	26.0	21.0	4.81	0.0	-1.0	CPU: H1 : M	-1.0	0.0241	9.833672	9.964703
38	0.0	35.0	16088.0	399.0	9.0	0.0248	0.0225	7.0	7.0	5.0	-1.0	-1.0	Note: Item c	-1.0	0.0241	9.975131	9.996148
39	0.0	60.0	353427.0	8765.0	192.0	0.0248	0.0219	98.0	79.0	4.81	-1.0	-1.0	NETWORK	-1.0	0.0212	9.863768	9.949912
40	0.0	4.0	106724.0	2444.0	51.0	0.0229	0.0209	29.0	18.0	4.62	0.0	-1.0	PRODUCT [	-1.0	0.0242	9.95204	9.932076
41	0.0	28.0	28242.0	675.0	15.0	0.0239	0.0222	14.0	13.0	4.93	-1.0	-1.0	General Mo	-1.0	0.0175	9.72928	9.984489
42	0.0	78.0	330995.0	7646.0	167.0	0.0231	0.0218	112.0	90.0	4.8	-1.0	-1.0	Language: M	-1.0	0.0246	9.945354	9.979846
43	0.0	38.0	40104.0	770.0	17.0	0.0192	0.0221	10.0	9.0	4.9	0.0	-1.0	10pcs/lot G	-1.0	0.0242	9.798972	9.996079
44	0.0	3966.0	92412.0	2107.0	45.0	0.0228	0.0214	33.0	24.0	4.73	0.0	-1.0	[Newmind \	-1.0	0.0241	9.951836	9.952611
45	0.0	61.0	11615.0	266.0	6.0	0.0229	0.0225	4.0	4.0	5.0	-1.0	-1.0	Mpie S12 A	-1.0	0.0175	9.766405	9.996148

Sheet1

就绪

06 模型应用

未知数据的模型预测

应用——预测手机销量

Params	Price	DeliveryF	StockCour	SearchCnt	BrowserCc	SaleCount	ClickRate	ConvertRa	CommentCc	GoodComme	Score	IsGallery	IsHighQua	DetailDes	CanDesign	Product	
Unlock Phoi	159.99	0.0	5948.0	1029954.0	22659.0	517.0	0.022	0.0228	379.0	342.0	4.9	0.0	-1.0	Real Stock,	-1.0	0.0251	0.023
Unlock Phoi	111.78	0.0	307.0	2186615.0	57508.0	1356.0	0.0263	0.0236	192.0	173.0	4.9	-1.0	-1.0	We will sen	-1.0	0.0251	0.0232
Unlock Phoi	160.99	0.0	214.0	1147155.0	25811.0	592.0	0.0225	0.0229	448.0	404.0	4.9	-1.0	-1.0	Main Featur	-1.0	0.0251	0.0229
Unlock Phoi	205.99	0.0	54.0	33974.0	812.0	18.0	0.0239	0.0222	12.0	11.0	4.92	0.0	-1.0	2K Display,	-1.0	0.0247	0.0226
Unlock Phoi	25.99	0.0	1595.0	423333.0	2540.0	11.0	0.006	0.0043	0.0	0.0	0.0	0.0	-1.0		-1.0	0.0247	0.0225
Unlock Phoi	374.99	0.0	197.0	38333.0	207.0	1.0	0.0054	0.0048	0.0	0.0	0.0	0.0	-1.0	CPU: MTK6	-1.0	0.0247	0.0225
Unlock Phoi	159.99	0.0	200.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-1.0	1. MTK6750	-1.0	0.0247	0.0225
Unlock Phoi	317.99	0.0	1000.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-1.0	Tips: 1, Plea	-1.0	0.0247	0.0225
Unlock Phoi	148.7	0.0	80.0	22083.0	159.0	1.0	0.0072	0.0063	0.0	0.0	0.0	0.0	-1.0	The real pho	-1.0	0.0247	0.0225
Unlock Phoi	439.83	0.0	387.0	8349.0	177.0	4.0	0.0212	0.0225	1.0	1.0	5.0	-1.0	-1.0	Original 5.2	-1.0	0.0247	0.0225
Unlock Phoi	28.49	0.0	2400.0	58061.0	1318.0	28.0	0.0227	0.0212	21.0	15.0	4.71	0.0	-1.0	Brand Name	-1.0	0.0247	0.0225
Unlock Phoi	38.99	0.0	23.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-1.0	Original E L	-1.0	0.0247	0.001
Unlock Phoi	113.13	0.0	400.0	9030.0	177.0	4.0	0.0196	0.0225	4.0	4.0	5.0	-1.0	0.0	NETWORK	-1.0	0.0247	0.0225
Unlock Phoi	45.99	0.0	99.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-1.0	Tip: Unlocke	-1.0	0.0247	0.0225
Unlock Phoi	28.49	0.0	294.0	25505.0	227.0	2.0	0.0089	0.0088	0.0	0.0	0.0	-1.0	-1.0	Model : XGC	-1.0	0.0247	0.001
Unlock Phoi	35.99	0.0	200.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-1.0	0.0	Highlights L	-1.0	0.0247	0.001
Unlock Phoi	19.76	0.0	19845.0	289282.0	6046.0	129.0	0.0209	0.0213	84.0	59.0	4.7	0.0	-1.0	[xlmodel]-[c	-1.0	0.0247	0.001
Unlock Phoi	185.39	0.0	40.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-1.0	2017 New E	-1.0	0.0247	0.0225
Unlock Phoi	71.99	0.0	146.0	45025.0	887.0	19.0	0.0197	0.0214	12.0	9.0	4.75	0.0	-1.0	CPU: MTK6	-1.0	0.0247	0.001
Unlock Phoi	129.99	0.0	400.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-1.0	-1.0	All of our m	-1.0	0.0247	0.0225
Unlock Phoi	39.99	0.0	392.0	8349.0	177.0	4.0	0.0212	0.0225	2.0	2.0	5.0	0.0	-1.0	Language S	-1.0	0.0247	0.0225
Unlock Phoi	164.82	0.0	2000.0	4625.0	74.0	1.0	0.016	0.0134	0.0	0.0	0.0	0.0	-1.0	Original vers	-1.0	0.0247	0.0225
Brand Name	197.99	0.0	1998.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-1.0		-1.0	0.0247	0.001
Unlock Phoi	139.99	0.0	300.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-1.0	-1.0	16.0MP Ou	-1.0	0.0247	0.001

07 灵敏度分析

模型稳定性检验

灵敏度分析

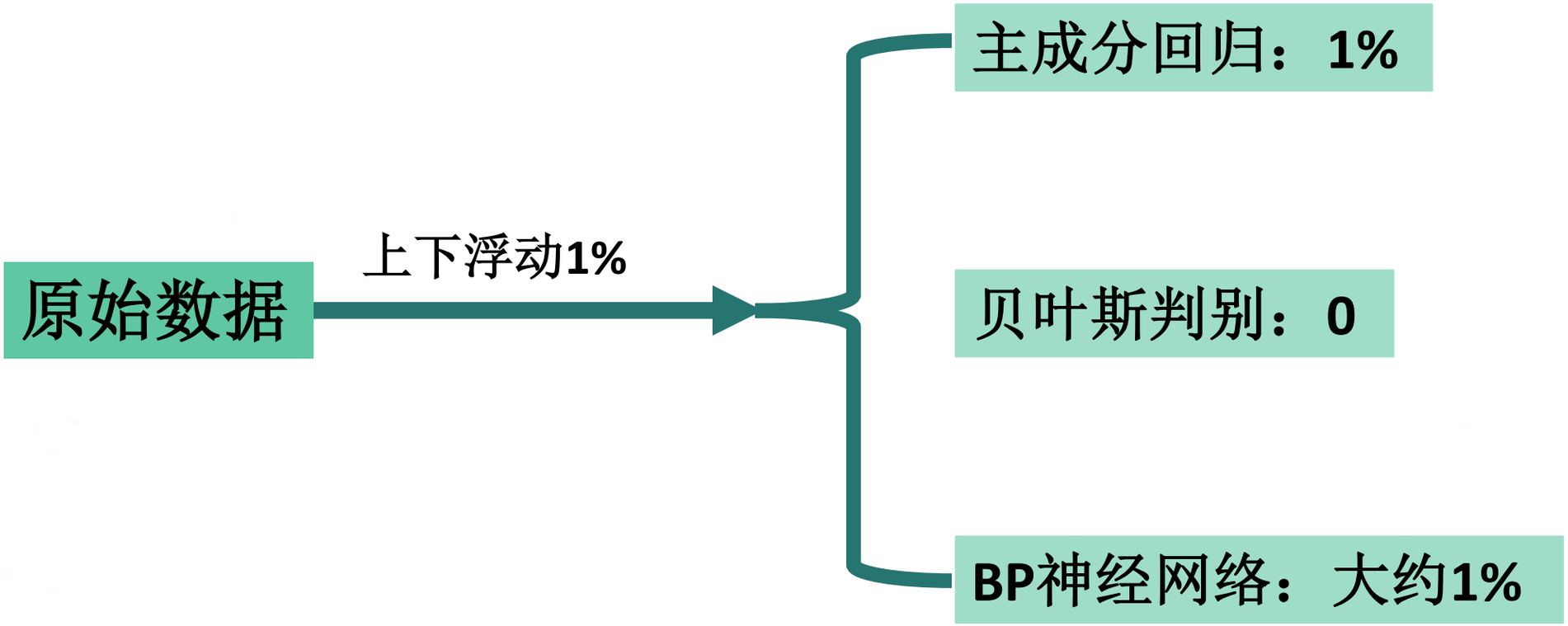
对XGBoosting算法的结果进行精准度评估打分：

$$S_k = max\left(0, 10 - 10 \times \left|\frac{log_{10}\left|\frac{x_{predict}}{x_{real}}\right|}{5}\right|\right)$$

点击率：9.81

转化率：9.74

灵敏度分析



08 结论综述

优势，应用与价值

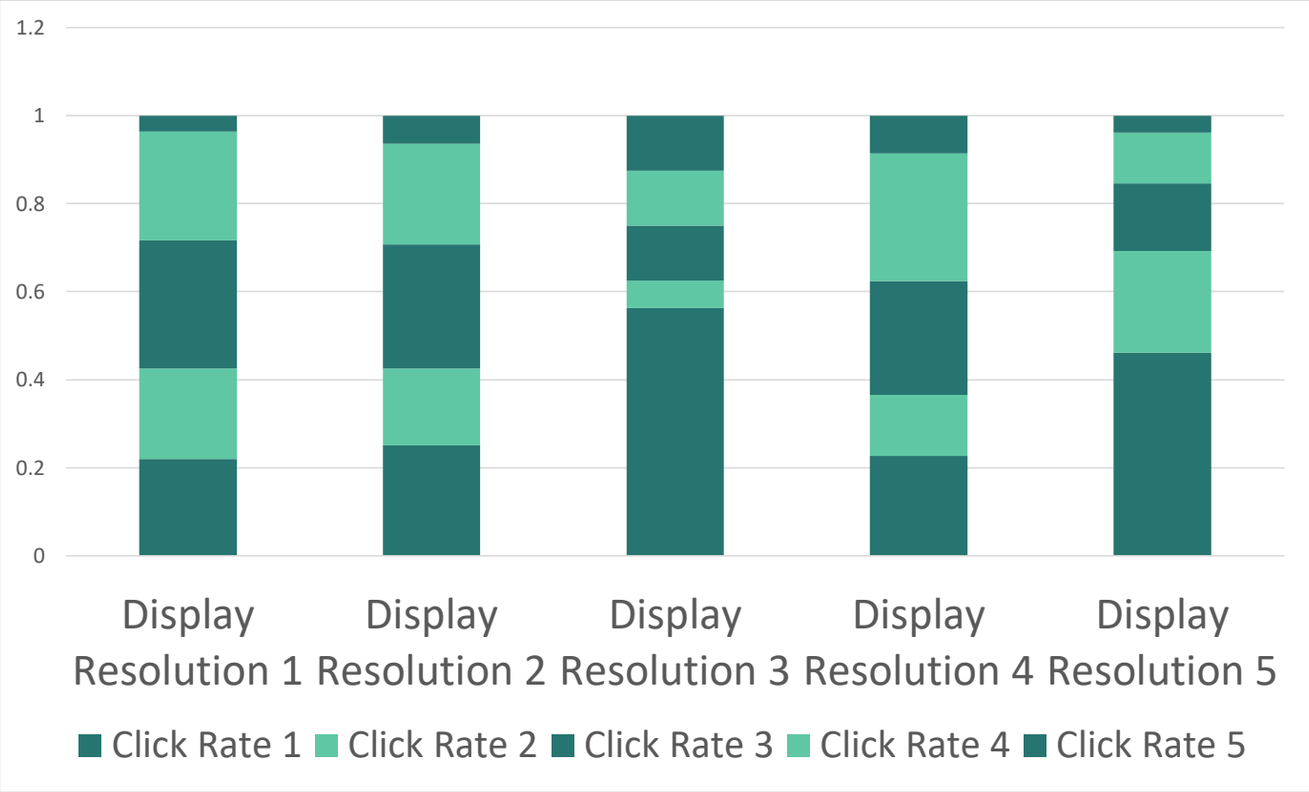
重要因素提取结论

独立变量削减（数据处理）结论展示

独立变量重要性排名	独立变量
1	Comment Count
2	Good Comment Count
3	Search Count
4	Score
5	Brand
6	Is Gallery Featured
7	Battery Capacity(mAh)
8	Highest camera resolution

模型建立（权重确定方法）与优化结论

中档分辨率，低档或高档视频率，高像素和中等价格手机更受消费者欢迎。相对于点击率与转化率，较低的，而较高的RAM，ROM，CPU点击率更高。



电池配置（容量）中等，金色与白色手机，内存大而摄像头清晰度高的手机为生产利润获取最大的手机配置

模型中的分析与发现

数据分析与处理优化发现：

首选因素

- RAM,ROM,CPU可成为生产时质量提高的首选

不敏感因素

- 公众对屏幕分辨率与像素敏感度不高，相较于价格，生产时应更注意成本降低

模型建立方法优势

定性分析

- 权重确定方法可对独立变量重要程度进行直观的定性展现，帮助生产商进行辅助分析

因素排名

- 主成分分析，神经网络等定量数据处理方法分析独立变量，探究大众需求，得出了重要因素的具体排名

定量优化

- 优化模型对手机具体特征对销量的影响进行了定量分析，其结论对具体产品制造进行了指导优化

应用价值

- 应用与优化模型可以成功高效的预测销量，有着重要的实际应用价值

感谢倾听
以下是提问环节

曹凌微 钱成田肇阳

清华大学附属中学

指导教师：吴昊 王殿军