小红书销售渠道效果预测分析

一、问题界定

1. 背景介绍

小红书是目前非常热门的电商平台，和其他电商平台不同，小红书是从社区起家。在小红书社区，用户通过文字、图片、视频笔记的分享，记录了这个时代年轻人的正能量和美好生活。

2. 目标问题

（1）使用 Python 建立线性回归模型

（2）预测用户的消费金额变化

（3）找到对用户消费影响较大的因素

二、数据概览

1. 数据维度分析

维度	字段	含义	变量类型
用户自然属性	gender	性别	类别型
	age	年龄	数值型
用户行为	lifecycle	生命周期	类别型
	engaged_last_30	最近 30 天是否在 App 上参加重要活动	类别型
	days_since_last_order	最近一次下单距今天数	数值型
	previous_order_amount	用户以往累积购买金额	数值型
	3rd_party_stores	用户以往在第三方平台购买的数量	类别型/数值型
销售（预测目标）	revenue	销售额 or 用户消费金额	数值型

总计有 29452 行数据，8 个字段，3 大维度。

2. 缺失值分析

存在缺失值字段	缺失行数	占比
gender	17723	39.8%
age	17723	39.8%
engaged_last_30	17723	39.8%

缺失值占到近 40%，因此不能直接删除，需要后续处理。

3. 异常值和离群值分析

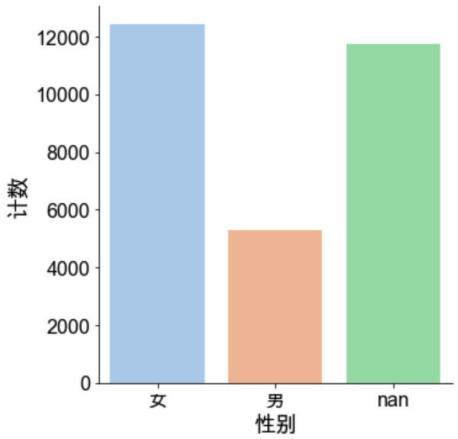
	revenue	age	days_since_last_order	previous_order_amount
count	29452.000000	17723.000000	29452.000000	29452.000000
mean	397.071515	29.419286	7.711348	2339.254020
std	959.755615	9.213604	6.489289	2361.572921
min	0.020000	14.000000	0.130000	0.000000
25%	74.970000	21.000000	2.190000	773.349500
50%	175.980000	29.000000	5.970000	1655.790000
75%	498.772500	37.000000	11.740000	3084.796500
max	103466.100000	45.000000	23.710000	11597.900000

查看所有数值型变量的统计学参数，大致判断出字段 revenue 和 previous_order_amount 中存在拖尾现象，存在离群值，需要处理。

三、单变量分析

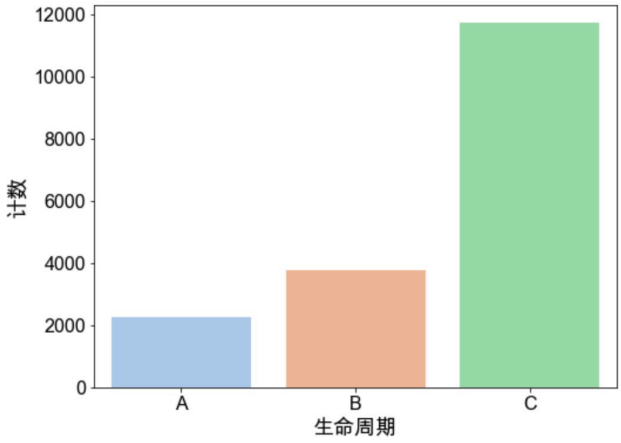
1. 类别型变量

（1）性别 gender



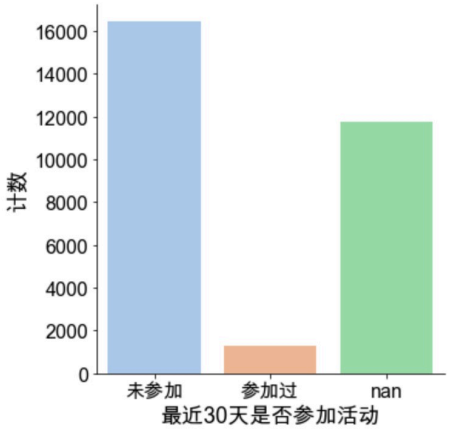
小红书用户中女性占多数，而对于缺失值 NaN 的填充只能用 Unknown。

（2）生命周期 lifecycle



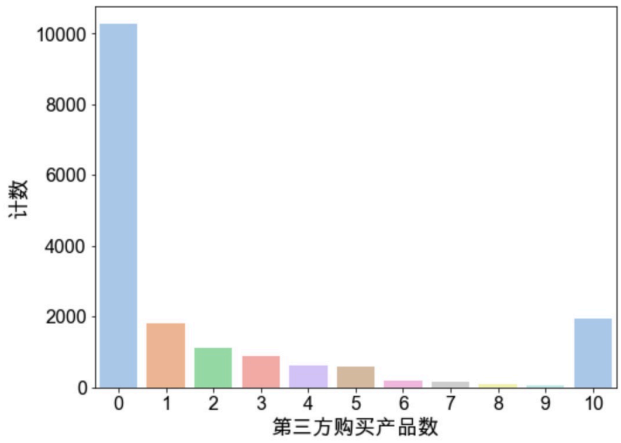
用户群体中，1-2 年的老用户占绝大多数，新用户数相对偏少。

（3）最近 30 天是否参加活动 engaged_last_30



一半以上的用户近 30 天内没有参加 APP 上的活动，而对于缺失值 NaN 的填充，可以填充为 0.0（即未参加）。

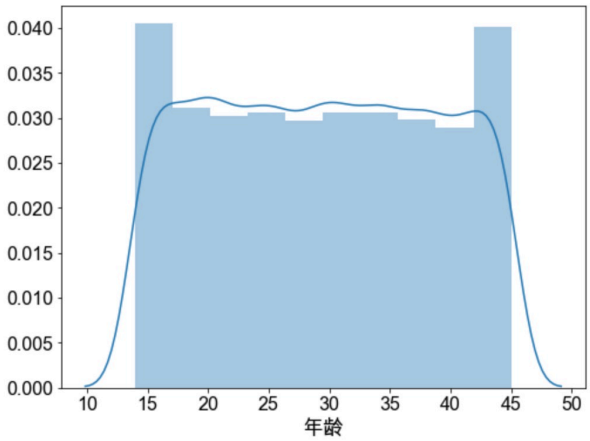
（4）第三方平台产品购买数 3rd_party_stores



小红书用户中在自营平台购买产品的人数居多，其次是很依赖第三方平台的用户。

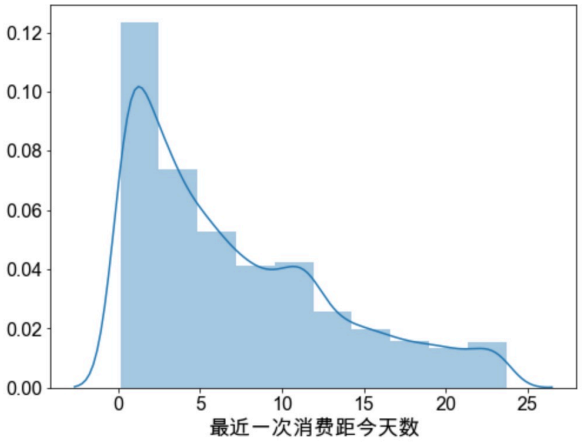
2. 数值型变量

（1）年龄 age



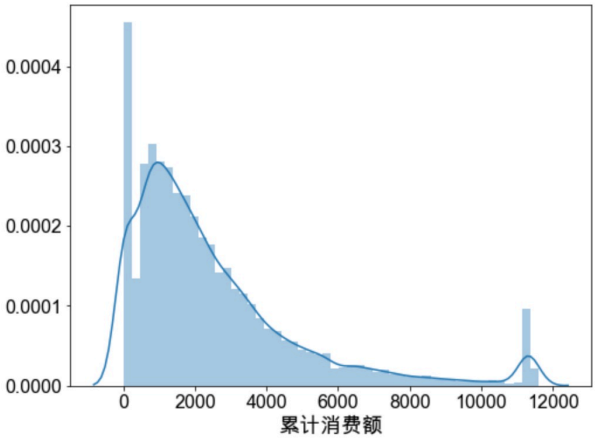
年龄层面上，没有呈现正态分布的趋势，青少年（15-20 岁）和中年（40-45 岁）的用户更多，后续可将年龄酌情分组。

（2）最近一次消费距今天数 days_since_last_order



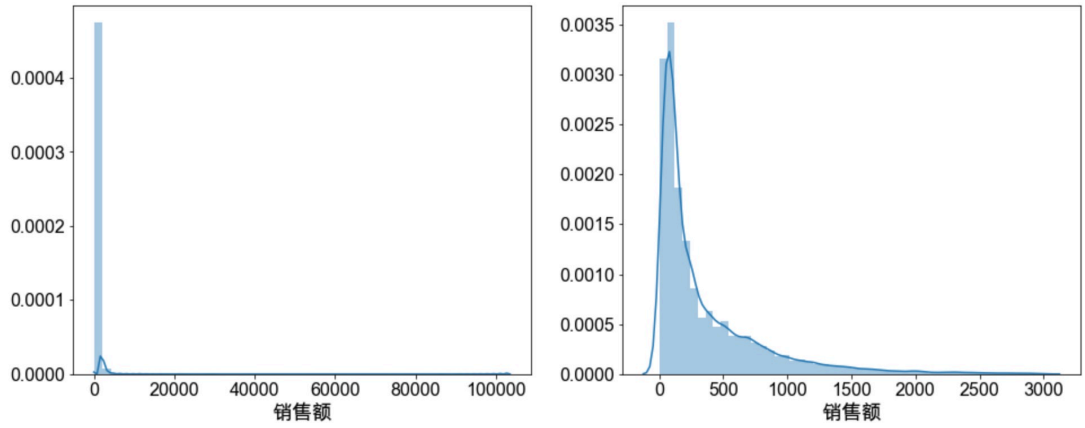
反映出活跃用户更多，若超过 30 天仍不在平台消费，说明用户已流失。小红书的销售额也主要依靠这些活跃的用户。

（3）用户以往累积消费金额 previous_order_amount



用户累计消费金额集中在0-4000左右，且几乎不消费的用户占很高的比例，而累计消费超过万元的也不在少数，说明用户消费情况存在两极分化。

（4）销售额 revenue



销售额拖尾严重，用户消费集中在 0-500 元之间，需要对离群值进行处理。

四、数据处理

1. 缺失值填充

（1）age 字段缺失值用平均值来填充

（2）gender 字段缺失值用 Unknown 来填充

（3）engaged_last_30 字段缺失值用 0.0 来填充

2. 字段重分组

（1）age 分组为<20、<30、<40、<50，命名为 age_range

（2）3rd_party_stores 分组为 0（代表没有在第三方平台购买过产品）和 1（代表在第三方平台购买过产品），命名为 3rd_new

3. 字段类别处理

将 gender、engaged_last_30、3rd_party_stores、lifecyle 定义为字符串类别，为后续哑变量处理做准备。

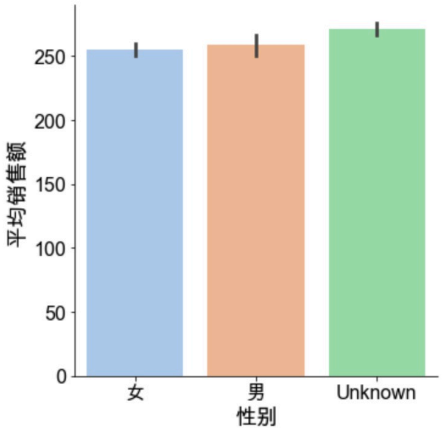
4. 离群值处理

将 revenue 和字段的离群值（此处定义为超过两个四分位点差值的 1.5 倍的值）去掉，总共去掉了 3814 行，占比 13%，可以接受。

五、多变量分析

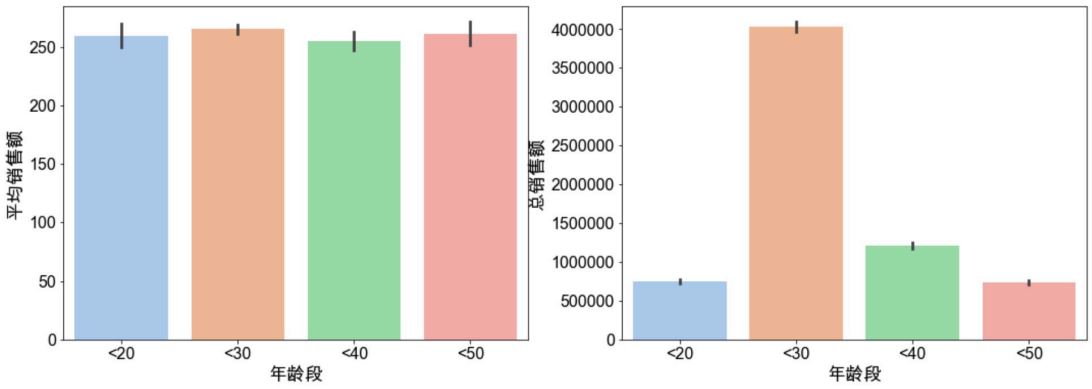
1. 各变量与销售额的情况

（1）性别与销售额



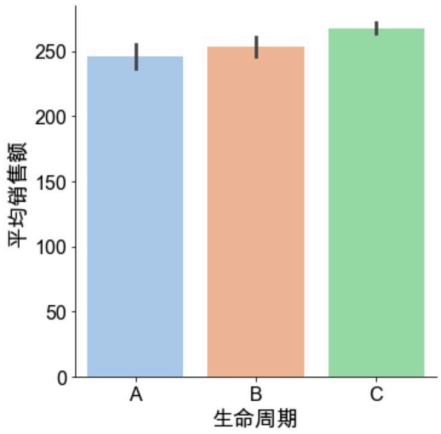
男性的消费均值高于女性，而不愿意透露性别的人群消费均值最高。

（2）年龄段与销售额



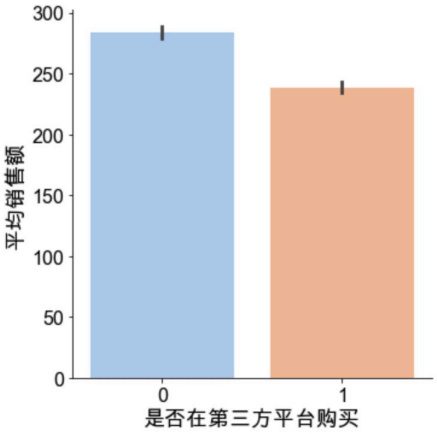
30-40 岁的用户消费均值最低，其余差别不大；20-30 岁群体是销售额主要贡献年龄层。

（3）生命周期与销售额



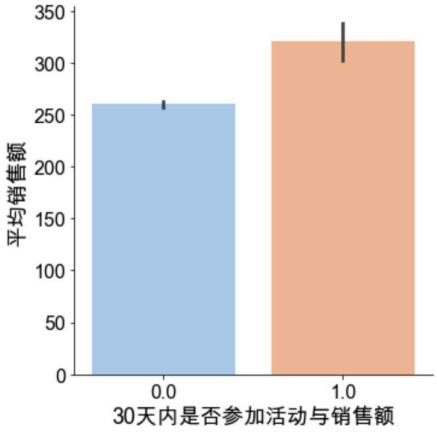
用户注册时间越长，消费均值越高。

（4）是否在第三方购买与销售额



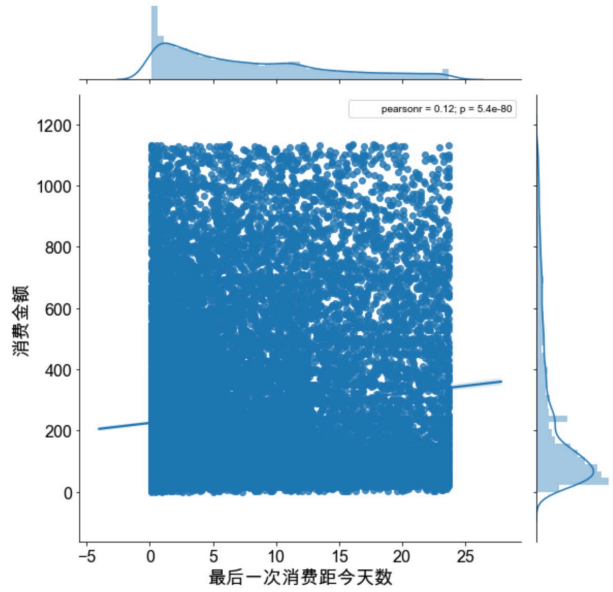
在自营平台消费的用户消费均值高于在第三方平台购买产品的用户，资金没有被分流。

（5）是否参加活动与销售额



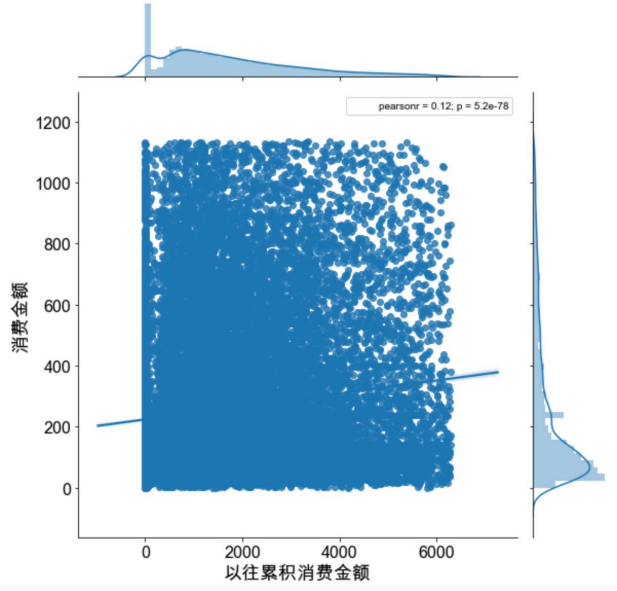
30 天内参加过 APP 上活动的用户在小红书消费均值更高，活动有刺激消费的作用。

（6）最后一次消费距今天数与销售额



没有看出明显的相关关系，数据主要集中在左下方，相关性系数为 0.12。

（7）以往累积消费金额与销售额



没有看出明显的相关关系，数据主要集中在左下方，相关性系数为 0.12。

2. 各变量与销售额的相关性系数

计算相关变量与销售额的相关性系数：

	revenue
revenue	1.000000
days_since_last_order	0.117874
previous_order_amount	0.116378
age	0.003108
3rd_party_stores	-0.048588

可以看出，销售额和最后一次消费距今天数、累计消费金额和年龄呈正相关，和在第三方平台购买产品数呈负相关；与销售额相关性最高的两个变量是最后一次消费距今天数和累计消费金额，但相关性很低，只有 0.12。

六、回归模型建立

1. 哑变量转化

将类别变量转化成哑变量，并去掉信息冗余变量。

2. 相关性系数重新计算和变量选择

	revenue
revenue	1.000000
days_since_last_order	0.117874
previous_order_amount	0.116378
engaged_last_30_1.0	0.041971
lifecycle_C	0.029354
gender_Unknown	0.028150
age_range_<30	0.013457
age	0.003108
age_range_<50	-0.001421
gender_1.0	-0.006155
age_range_<40	-0.013061
lifecycle_B	-0.015850
3rd_party_stores	-0.048588
3rd_new_1	-0.084859

根据相关性系数，全部放入模型中不太合适，我们选取与销售额 revenue 相关性相对较强的变量且保证信息不冗余，因此选择 days_since_last_order、previous_order_amount、engaged_last_30_1.0、lifecycle_C、gender_Unknown、age_range_<30 和 3rd_new_1 来作为自变量，而因变量为 revenue。

3. 分割数据集并标准化数据

按训练集和测试集为 7:3 的比例分割数据集，并且将训练集和测试集中的自变量标准化，数据绝对值量级差异带来的影响。

4. 训练模型并评估模型

（1）模型参数

参数名称	系数
days_since_last_order	35.735
previous_order_amount	29.078
engaged_last_30_1.0	11.147
lifecycle_C	11.984
gender_Unknown	11.054
age_range_<30	-3.103
3rd_new_1	-39.366
intercept	262.314

正负号表示与销售额正相关还是负相关，系数的绝对值反映了该变量对预测销售额的影响程度，可以看出，days_since_last_order、previous_order_amount 和 3rd_new_1（即上一次消费距今天数、累计消费金额和在第三方平台消费）对用户的消费影响最大，其中前两者是正向影响，后者为负向影响。

（2）评估模型

参数	结果
训练集打分	0.04630
测试集打分	0.04336
RMSE	260.435
MAE	203.630

可以看出，给数据集不太适合做线性回归模拟，或者需要进行优化。

（3）OLS 拟合

OLS 的报告也给出了详细的解读，皮尔森系数仍然很低。

Dep. Variable:	y	**R-squared:**	0.046
Model:	OLS	**Adj. R-squared:**	0.045
Method:	Least Squares	**F-statistic:**	174.7
Date:	Wed, 29 Apr 2020	**Prob (F-statistic):**	9.89e-254
Time:	19:13:23	**Log-Likelihood:**	-1.7879e+05
No. Observations:	25638	**AIC:**	3.576e+05
Df Residuals:	25630	**BIC:**	3.577e+05
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	192.1998	4.277	44.941	0.000	183.817	200.582
X[0]	5.9362	0.318	18.648	0.000	5.312	6.560
X[1]	0.0194	0.001	17.027	0.000	0.017	0.022
X[2]	58.4034	8.857	6.594	0.000	41.044	75.763
X[3]	23.7941	5.362	4.438	0.000	13.285	34.304
X[4]	23.2503	4.553	5.106	0.000	14.325	32.175
X[5]	-4.6329	4.499	-1.030	0.303	-13.451	4.185
X[6]	-75.8541	4.199	-18.066	0.000	-84.084	-67.624

Omnibus:	5119.995	**Durbin-Watson:**	1.995
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	8831.027
Skew:	1.333	**Prob(JB):**	0.00
Kurtosis:	4.079	**Cond. No.**	1.28e+04

5. 模型优化

（1）修改自变量

a. 放入全部变量

b. 筛选变量

选择 days_since_last_order、previous_order_amount、engaged_last_30_1.0、lifecycle_C、gender_Unknown 和 3rd_new_1 来作为自变量，是尝试了多种排列组合的最优。

参数	原方案	方案 a	方案 b
训练集打分	0.04630	0.04660	0.04623
测试集打分	0.04336	0.04221	0.04342
RMSE	260.435	260.592	260.428
MAE	203.630	203.711	203.654

由此可见，两种改进方案均没有显著的提升，因此舍弃。

（2）离群值操作

a. 数据处理时不对 revenue 和 previous_order_amount 的离群值进行操作

b. 数据处理时改变对 revenue 和 previous_order_amount 离群值的操作

将 revenue 和字段的离群值（此处定义为超过两个四分位点差值的 3 倍的值）去掉，总共去掉了 1450 行，占比 5%，可以接受。

参数	原方案	方案 a	方案 b
训练集打分	0.04630	0.03201	0.04921
测试集打分	0.04336	0.03988	0.04815
RMSE	260.435	690.445	342.371
MAE	203.630	347.239	257.292

可以看出，方案 a 的拟合效果更差，说明离群值的存在对模型具有一定影响，且残差显著提高；而方案 b 对离群值的判定调整稍微优化了模型，但是残差方面有所提升。

综合考虑来看，可以采用方案 b。