大数据精准营销项目

实训报告

姓名	周一豪
学号	2016111257
专业	管理科学与工程
班级	上海财经大学
日期	2020年8月9日

目录

1　项目背景与目标

1.1　项目背景

随着移动互联网时代的到来，传统零售从业者逐渐转向电商平台，区别于传统区域化、实体化的营销方法，电商平台则是通过实时推广来吸引消费者，获取关注从而延长产品的生命周期。

精准定位消费者的需求和偏好是电商营销的重点难点，大数据分析技术日趋成熟、各种消费平台争相涌现，借助客户消费的历史数据，企业可以通过分析，构建用户画像，准确找到其产品的目标客户群，最大限度获得目标客户，实现精准营销。

1.2　项目目标

首先通过对海量结构化数据和非结构化文本数据的深度分析和挖掘，构建全方位的客户标签体系。其次，基于客户标签体系，从基本信息、消费能力、行为习惯等多个维度对客户进行精准画像。最后，计算用户商品兴趣度排行榜，从而支持精准目标客户筛选。

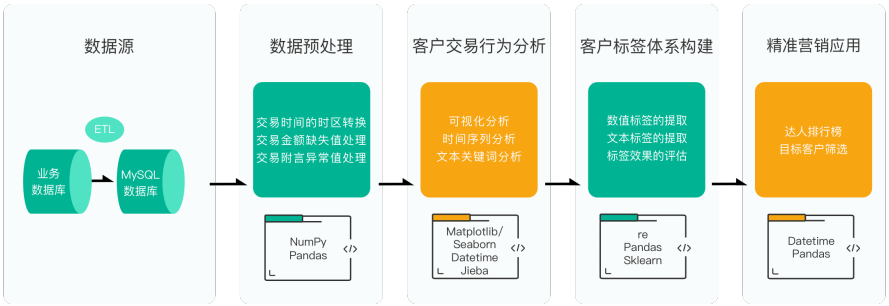

图1：项目流程

- 从数据源读取数据：使用Python接口从MySQL数据库中读取数据，并将数据转换为易于处理的DataFrame格式。
- 数据预处理：在数据读取并转化为DataFrame格式后，我们利用Pandas提供的便利工具和函数，对交易数据进行初步探索和数据预处理。数据预处理的目的是提高数据质量，便于后续的数据分析。
- 客户交易行为分析：在数据预处理后，我们分别从时间维度和交易行为维度对客户数据进行探索分析，利用Python中的绘图库(如Seaborn、Matplotlib)进行一些可视化操作。
- 客户标签体系构建：在进行客户交易行为分析之后，参照电商行业用户画像的建立方式，开始构建客户标签体系，包含事实类、规则类、预测类和文本类四大类标签。标签体系建立之后，我们对典型客户构建用户画像，进行了重点分析。
- 精准营销应用：通过分析前面构建的客户标签体系来个性化营销方案。在项目中我们采用了两种方式：构建商品兴趣度排行榜和目标客户的筛选。

2　客户数据预处理与客户交易行为分析

2.1 数据集介绍

本项目包含客户在某移动支付平台的367万脱敏交易流水数据，交易时间跨度为5年。每个客户有多条交易，每条交易记录了客户ID、交易时间、交易金额和交易附言四个字段。各字段的英文名称、中文名称和备注如表1所示：

表1：客户交易记录字段说明

英文名称	中文名称	备注
user_id	客户ID	客户的唯一标识
payment	交易金额	正为支出，负为收入
describe	交易附言	对此项交易的文字描述
unix_time	交易时间	unix时间戳

原始数据存放在MySQL数据库中，在进行数据读取之后，我们进行了数据格式的转换。

> 将数据结构转为了二DataFrame结果，这样便于我们利用其函数、属性对数据进行更加便捷的操作。

DataFrame为二维表结构，每一行代表一条交易记录，数据概览如表2所示：

表2：数据概览

user_id	payment	describe	unix_time
2099234	200000	支付宝（95188）	1487088000
22677084	-27500	北京京东三佰陆拾度电子商务公司	1510329600
17775403	5000000	银联POS消费	1503849600
8960118	6600	贷记卡转账还款	1496332800
14042200	-400000	转帐存入　厦门建行会计部清算	1473342861

2.2 数据预处理

在上一个环节，我们已经从MySQL数据库中读取了交易流水数据，并将数据转换成Pandas中的DataFrame格式。在本环节，我们将利用Pandas提供的便利工具和函数，对交易数据进行初步探索和数据预处理。数据预处理的目的是提高数据质量，便于后续的数据分析。具体地，对交易数据进行预处理的流程如图2所示。

图2：预处理流程

- 统计分析：对数据进行统计分析，初步了解数据特点。例如查看交易数据的行数和列数，以及数据类型和各字段的缺失值情况，统计交易数据中包含的客户数量等。
- 异常值处理：对交易时间等字段中出现的异常数据进行诊断，并确定异常值处理方法。
- 缺失值处理：对于存在缺失值的交易金额和交易附言字段，进一步诊断缺失值产生的原因，从而确定缺失值处理方法。
- 数据格式转换：为了便于后续分析，对于金额字段的量纲、交易时间字段的时间格式进行转换。
- 重复数据过滤：检测交易数据中存在的重复交易记录，并删除重复的记录。

Unix时间戳是指格林威治时间1970年01月01日00时00分00秒（北京时间1970年01月01日08时00分00秒）起至现在的总秒数。观察时间戳字段，我们发现有一些异常值，如表3所示。

表3：时间戳字段异常值

user_id	payment	describe	unix_time
6729440	-4540	NaN	0
27164939	-2580	支付宝网络还款	14 3264000

对时间为0的数据字段，直接删除；对于时间字段缺一位数的情况，我们首先根据正则表达式进行匹配到相应字段，对其填充缺失值。 这样处理的原因是，由于我们要基于时间进行观察，所以时间取值为0没有意义删除掉；而缺一位的数据可以通过上文填充

此外交易附言（describe）字段存在一些缺失，对应记录的交易金额和交易时间都是完整的，这些交易记录可以反映客户的一些消费习惯，所以我们对其进行保留。

为了便于后续处理与分析，在本项目中我们将时间戳转换成“年-月-日：时：分：秒”的格式。接着进行时区转换，将格林威治时间转换为北京时间，转换后的数据保存为交易记录中新的一列，列名为pay_time，如表4所示：

表4：时间戳字段转换

user_id	payment	describe	unix_time	pay_time
31096165	849600	提现 实时提现	1478425197	2016-11-06 17:39:57
31096165	1600	支付-乐和彩充值	1497970624	2017-06-20 22:57:04
31096165	-3000	支付-赚钱高手感谢您的使用	1501321472	2017-07-29 17:44:32
31096165	-811000	转账 余额宝支付	1502193874	2017-08-08 20:04:34
31096165	-179800	充值	1498702285	2017-06-29 10:11:25

交易金额（payment）字段以‘分’为单位，为了符合我们的观察习惯，我们将其量纲改为‘元’。

2.3 客户交易行为分析

在上一个环节，我们已经从对客户交易数据进行了异常值、缺失值和重复值的处理，同时也对时间和交易金额进行了相应的转换。在本环节，我们将利用Python中的绘图库（如Seaborn、Matplotlib），通过可视化的方式，对客户交易行为进行分析。对客户交易行为进行分析的目的是统计和观察在各个维度上的数据分布，得到一些关于客户行为的有价值的结论。具体地，客户交易行为分析的流程如图3所示。

图3：客户交易行为分析流程

下面我们对图3中的每一个流程进行简要说明。

- 时间维度的分析：对交易时间进行分析，探索交易随时间的分布规律。例如分析交易次数随时间的变化，不同时段的交易次数分布等。
- 交易属性的分析：对交易金额和次数进行分析，例如分析客户的交易次数分布和平均交易金额分布等。
- 文本数据预处理：为了便于后续分析，对交易附言的文本进行预处理，例如文本分词和去除停用词等。
- 文本数据的分析：对预处理后的文本进行分析，例如绘制词云分布图和提取关键词等。

从交易时间和交易次数两个维度进行分析，交易流水数据中不同时间的交易次数分布如图4所示：

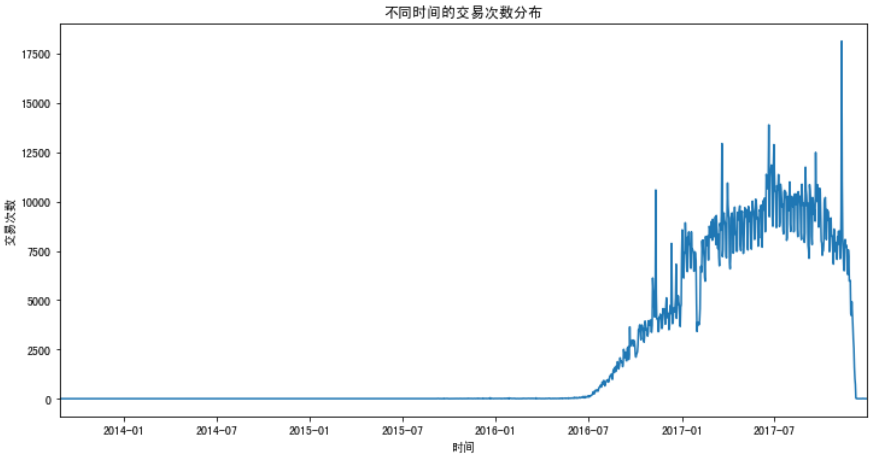

图4：不同时间的交易次数分布

从交易时间和交易金额两个维度进行分析，交易流水数据中不同时间的交易金额分布如图5所示：

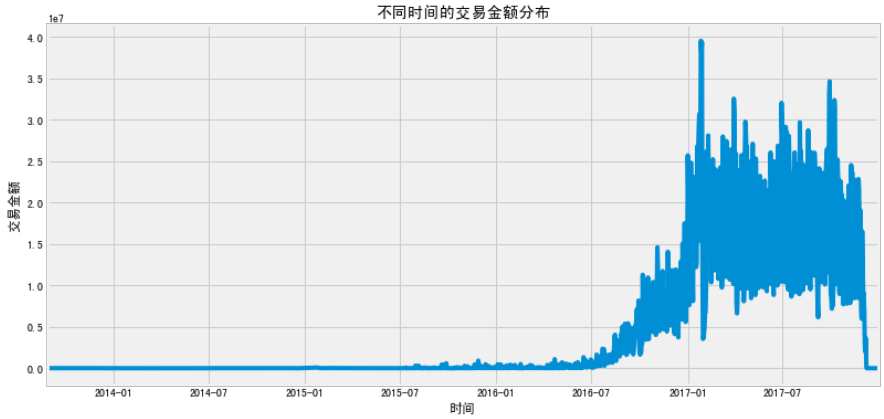

图5：不同时间的交易金额分布

交易次数和交易金额都是在2016年7月之后出现较大提升，说明这以后的交易活跃。在项目中，为了统计每日的数据，我们首先将时间戳转为pd.dt.day，然后通过groupby函数，将统计数据基于日期分类，求和计算。

统计每天各时段的客户交易数量，如图6所示：

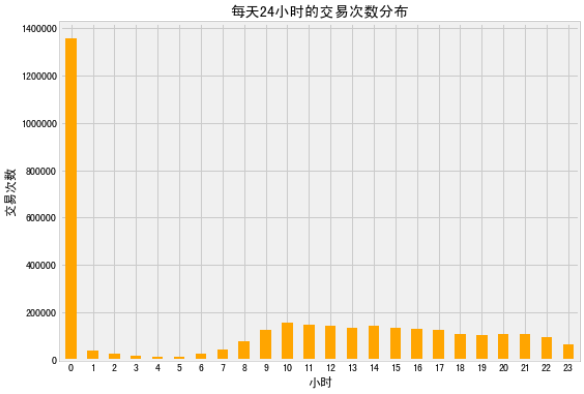

图6：每天24小时的交易次数分布

客户时间段在0时，机构主要用于交易清算；然后在一天中的工作时间段，早上七点开始到晚上，交易较为 均匀分布，说明除开睡眠时间，其他时间段交易分布较为均匀。

客户交易次数和平均交易金额的核密度曲线如图7和图8所示：

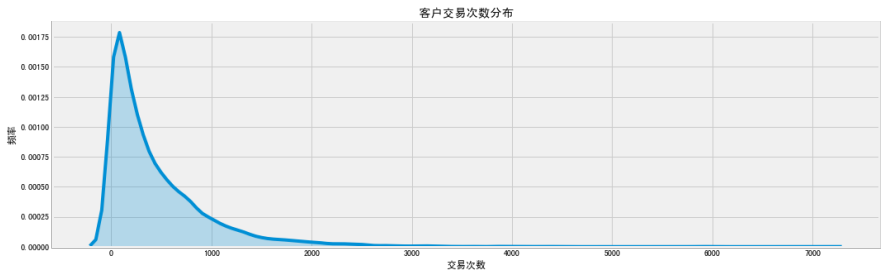

图7：客户交易次数分布

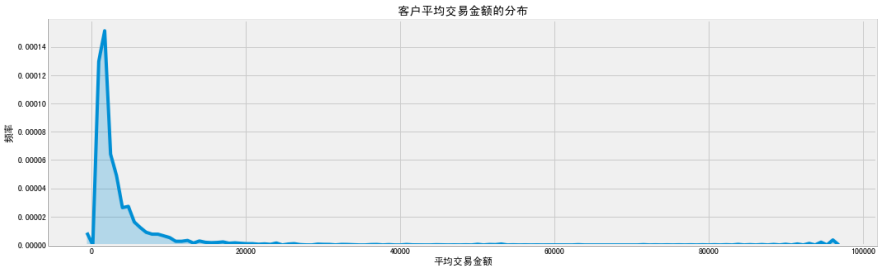

图8：客户平均交易金额的分布

客户交易次数主要集中在0-2000次，而交易金额主要分布在10000左右。

接下来我们对交易附言中的文本进行了可视化，设置停用词并画出词云，如图9所示：

图9：交易附言词云图

交易附言中，转账，充值，支付宝等关键字占据主导地位，说明消费者主要是基于这些方式在消费，相互转账，充值的需求也很高。之后可以基于不同的交易行为进行不同客户的标签建立。

3　客户标签体系构建

通过客户交易数据的深度挖掘，构建全面的客户标签体系，能够对客户进行精准画像，从而在产品营销中准确高效地定位和筛选目标客户，实现精准营销。

3.1　客户标签体系介绍

在本环节，我们将结合结构化数据和非结构化文本数据，从交易属性、消费偏好、行为特征三大维度构建客户标签体系。我们的标签体系一共包含150个标签，客户标签体系框架如图9所示：

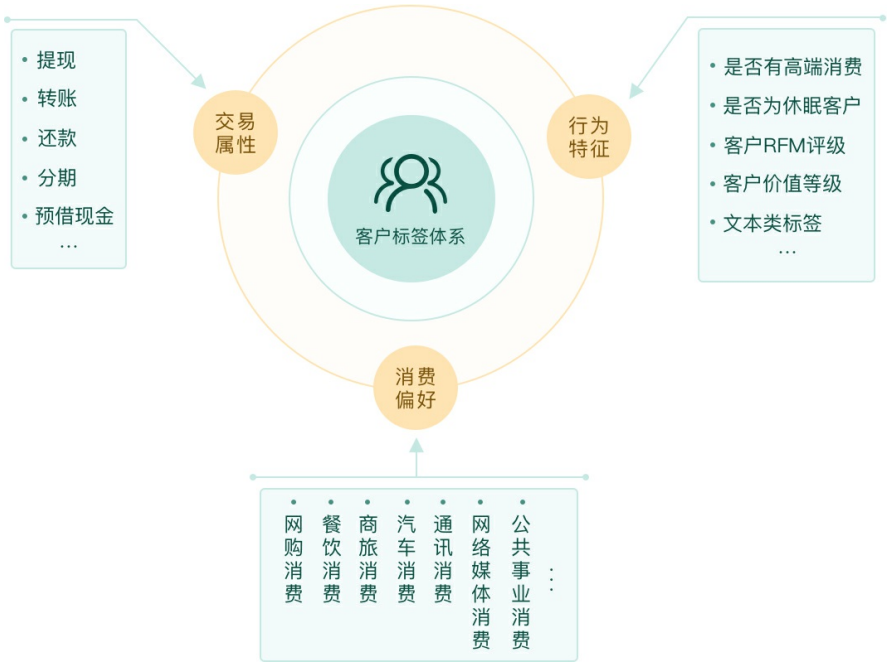

图10：客户标签体系

根据具体标签的构建方式，可以将客户标签分为以下四类：

- 事实类标签：可以直接从客户交易记录中进行统计和计算的标签。例如网购消费、餐饮消费、商旅消费等。
- 规则类标签：在事实类标签的基础上，结合人工经验，对客户的某项指标进行的计算或归类。例如RFM标签、是否休眠客户、是否有高端消费等。
- 预测类标签：原始数据中不能直接提取，需要借助模型进行预测的标签。例如客户价值等级。
- 文本类标签：从客户交易记录的文本中提取的关键词也可用于描述客户偏好，将此部分关键词作为文本类标签。例如彩票、儿童、孕妇、基金等。

具体的标签构建流程如图11所示：

图11：客户标签体系构建流程

在标签构建完成后，我们选取了两个典型客户，从交易时间、交易次数、交易金额和交易附言内容等几个维度进行了用户画像的刻画和分析。例如客户的词云图、交易流入流出的对比分析、交易描述的分布和不同客户行为的排序及分析等，进一步加深对典型客户的理解。

3.2 事实类标签构建

事实类标签可以直接从客户交易记录中进行统计和计算。本项目共提取40个事实类标签，包括交易流水类、转账类、消费类、商旅类、公共事业缴费和有无分期等，如表5所示：

表5：事实类标签

英文	中文	英文	中文
max_consume_amt	单次最大消费金额	return_cnt	退货订单数
consume_order_ratio	消费订单比例	public_pay_amt	公共事业缴费金额
mon_consume_frq	月均消费频度	internet_media_cnt	网络媒体类消费次数
consumption_channel	最常用支付工具	internet_media_amt	网络媒体类消费总金额
online_cnt	网购订单次数	phone_fee_cnt	话费通讯类消费次数
online_amt	网购订单总金额	phone_fee_amt	话费通讯类消费总金额
online_avg_amt	网购订单平均金额	is_installment	有无分期
mon_online_frq	月均网购频度	cash_advance_cnt	预借现金次数
online_buy_first_date	网购首单时间	cash_advance_amt	预借现金总金额
online_buy_last_date	网购尾单时间	total_transactions_amt	交易总金额
dining_cnt	餐饮订单次数	total_transactions_cnt	交易次数
dining_amt	餐饮订单总金额	withdraw_cnt	提现次数
dining_avg_amt	餐饮订单平均金额	withdraw_amt	提现总金额
business_travel_cnt	商旅次数	total_deposit	ATM存款总金额
business_travel_amt	商旅消费金额	total_withdraw	ATM取款总金额
business_travel_avg_amt	商旅消费平均金额	transfer_cnt	转账次数
mon_business_travel_frq	月均旅行频次	transfer_amt	转账总金额
car_cnt	汽车消费次数	transfer_mean	转账平均金额
car_amt	汽车消费总金额	credit_card_repay_cnt	信用卡还款次数
payroll	有无代发	credit_card_repay_amt	信用卡还款总金额

我们通过Series对象的str.contains()方法，来匹配不同类型的消费行为，从而对其进行分类；同时，我们也根据user_id进行groupby()操作，统计每位用户的单次最大消费金额等事实类标签。

3.3 规则类标签构建

规则类标签是在事实类标签的基础上，结合人工经验，对客户的某项指标进行的计算或归类，如定义高端消费需要人为规定阈值。在项目中，我们提取的规则类标签包括有无高端消费、是否休眠、RFM等，如表6所示：

表6：规则类标签

英文	中文
won_high_end	有无高端消费
sleep_customers	是否休眠客户
recency	近度
frequency	频度
monetary	值度
R_score	近度得分
F_score	频度得分
M_score	值度得分
Total_Score	RFM总得分

随着可支配水平的升高，客户的消费偏好会发生变化，其对某些中、高档商品的购买和消费量会增加，对低档消费品的需求减少。

> 我们首先对user_features这张表中，提取出每位客户的最大单笔交易金额，然后利用quantile(0.75)找到阈值，最终比阈值高的人群定义为有高端消费。 对于高端消费人群，可以推送一些定制化的，高档消费的服务。

精准营销的某一个目标是激活 “休眠” 客户。所谓 “休眠” 客户，是指那些已经了解企业和产品，却还在消费与不消费之间徘徊的客户。

> 统计用户总的消费数量，设定好阈值，消费总次数低于这个阈值的就判断为休眠客户。通过发送短信，推送，或者优惠卷等方式刺激他们重新消费

RFM模型是衡量客户价值和客户创利能力的重要工具和手段。该模型通过客户的近期消费行为、消费的总体频率以及消费总金额三项指标来描述该客户的价值状况。在项目中我们计算了RFM标签以及RFM总评分，相关分布如图11所示：

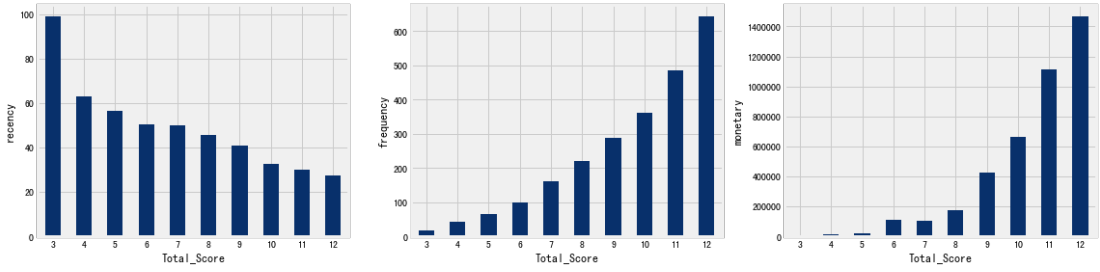

图12：RFM分布

> 可能的原因是随着时间周期的不断变化，消费者的行为发生变化，消费也有对应的大小周期。

3.4 预测类标签构建

原始数据中不能直接提取，需要借助模型进行预测的标签称之为预测类标签。在这个项目中，客户价值等级为预测类标签。我们根据提取出的事实类和规则类标签建立模型，对没有客户价值等级标签的客户进行预测。

在预测客户价值等级之前，需要对标签进行一些处理。如字符型标签进行One-Hot编码、连续型标签进行数值离散化等。标签处理完成后进行训练测试集划分，再训练模型对客户价值等级进行预测。流程示意如图13所示：

图13：标签预测流程

针对字符型标签，我们通常需要进行数值编码，以便模型能更好的处理数据。数值编码有多种方式，我们在项目中选择One-Hot编码。

> 对支付方式进行了one_hot编码，利用pd.get_dummies()对支付方式进行编码。

在处理连续型数据时，我们通常需要将数据进行离散化，使模型对异常数据有更强的鲁棒性，同时也能降低模型运算复杂度，提升模型运算速度。

> 对日期进行了等距离离散化，对剩余的连续型数据进行了等频离散化。

对标签进行处理后，我们按照8:2的比例将有客户价值等级的数据随机划分为训练集与测试集，接着建立逻辑回归模型对客户进行分类。

逻辑回归（logistic regression）采用回归分析的思想，解决分类问题。在线性回归的基础上，利用一个非线性函数（sigmoid）进行映射，将不同类别的样本区分开。模型结果的输出范围在（0，1）之间，可以解释为样本属于正类（高价值）的概率，常用于解决二分类问题。

我们将事实类、规则类标签共49个全部作为输入标签，将客户价值等级作为预测标签。使用Sklearn工具包构建逻辑回归模型，在划分好的训练集上进行训练，再将模型在测试集上进行评估。

对于二分类问题常用的评价指标有准确率(Accuracy)、精度(Precision)、召回率(Recall)和F1值(F1 Score)等。通常以关注的类(高价值等级)为正类，其他类(低价值等级)为负类。

四个评价指标的计算方法如下：

- $Accuracy = \frac{\text{所有正确预测的样本数}}{\text{总样本数}}$
- $Precision = \frac{\text{预测为正类且正确预测的样本数}}{\text{所有预测为正类的样本数}}$
- $Recall = \frac{\text{预测为正类且正确预测的样本数}}{\text{所有真实情况为正类的样本数}}$
- $F_1Score = \text{精确率和召回率的调和平均数} = 2\frac{Precision \times Recall}{(Precision+Recall)}$

在项目中，我们建立的逻辑回归模型准确率为 72.07% ，输出的分类报告如表7所示：

表7：模型在测试集上的分类效果

	precision	recall	f1-score	support
0	0.73	0.88	0.80	521
1	0.68	0.46	0.55	306
avg/total	0.71	0.72	0.71	827

- 其中列表左边的一列为分类的标签名，右边support列为每个标签的出现次数。
- avg / total行为各列的均值（support列为总和）。
- precision、recall和f1-score三列分别为各个类别的精确度、召回率及F1值。

> 分类结果的精确度为0.71，效果还不错，recall数据等还可以进一步提高。

接下来我们使用训练好的逻辑回归模型，预测其它客户价值等级。最终7630个客户价值等级分布如图14所示：

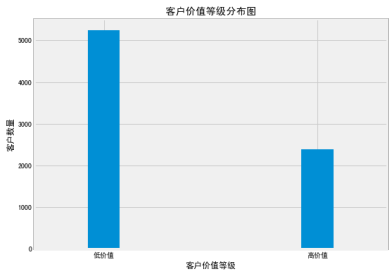

图14：客户价值等级分布

我们将客户价值等级与其他特征进行组合分析。画出客户价值等级与是否拥有高端消费、月均消费频度之间的关系图：

图15：客户价值等级与是否拥有高端消费、月均消费频度之间的关系

左图中，高端消费中，高价值客户所占的比例更大，右图中，随着消费频度的增加，高价值客户所占的比例也越来越高。 由于高端消费群体、高频度消费群体其购买能力更强，所以其更有可能是高价值客户，所以高价值客户在这些类别中占比更高。

3.5 文本类标签构建

对事实类、规则类和预测类标签进行构建后，我们还利用了交易附言中的文本数据建立非结构化标签体系，进一步完善整体的用户画像。对文本数据进行预处理后，使用统计词频的方法进行关键词抽取，构建文本标签。

我们采用了sklearn模块中的CountVectorizer()和TfidfVectorizer（）进行文本提取的两个方法，作用分别是统计词语出现的次数，出现的频度，并得出一些统计量如F1,Accuracy等

项目中提取出的100个文本标签如表8所示：

表8：文本类标签

1	2	3	4	5	6	7	8	9	10
iphone	taobao	tmall	t恤	三星	专柜	休闲	保暖	保湿	保险
修身	停彩	儿童	内衣	内裤	冬季	分期	加厚	加绒	卡通
双色球	可爱	司机	商城	圆领	基金	复古	夏装	外套	大乐透
大润发	大码	套装	女士	女童	女装	婴儿	孕妇	宝宝	家用
宽松	小米	彩票	情侣	打底	投资	新品	无线	时尚	易宝
春秋	春装	显瘦	服饰	欧美	汽车	淘粉	游戏	电子	电脑
男士	男女	男装	百货	真皮	短袖	短裤	福利彩票	秋装	积分
竞彩	纯棉	网易	美团网	腾讯	苹果	蕾丝	衬衫	裤子	证券
贷款	超市	超薄	足球	车费	运动	进口	连衣裙	迷你	追号
透气	酒店	金融服务	铁路	长袖	闪电	零食	韩版	食品	餐饮

从模型提取出的文本标签中，我们可以大致描绘出如下几类人群：

表9：文本标签的分类

文本特征	对应人群
停彩、大乐透、双色球、福利彩票、彩票、竞彩、追号	爱好购买彩票的人群
儿童、卡通、可爱、女童、零食	家庭中有孩子的人群
男士、衬衫、透气、真皮、足球、运动、小米、汽车、电子、汽车	广大的男士人群
修身、专柜、保湿、女士、女装、显瘦、打底、蕾丝、欧美、韩版	爱美的女士人群
婴儿、孕妇、宝宝	家庭中有孕妇/婴儿的人群
分期、贷款、金融服务	有借贷分期需求的人群
基金、投资、证券	有投资需求的人群

从文本挖掘中可以看到用户的消费倾向，从而预测用户属于哪一类消费者，可能有哪些消费需求。

3.6 典型客户画像分析

在客户标签体系建立完成后，我们对典型客户进行了重点的分析。如画出典型客户的词云分布图、分析各自的文本标签和数值标签等，进一步完善用户画像。

我们首先对交易流水最多的客户进行分析，通过对客户的交易附言信息进行整合，画出客户交易附言的词云图，如图16所示：

图16：客户词云图

该客户是一个男士客户，喜欢购买护肤的相关产品，同时也有使用信用卡的相关需求。所以，其更有可能购买护肤类产品。

接下来我们分析该客户的文本类标签，取出出现次数最多的前20个关键词，如表10所示：

表10：客户交易附言关键词

男士	0.926927
保湿	0.272997
专柜	0.250537
保险	0.0273075
电子	0.0231014
儿童	0.0226176
新品	0.0180205
休闲	0.012668
套装	0.0124769
短袖	0.0124374
汽车	0.0103755
分期	0.00996224
韩版	0.00978032
衬衫	0.00844436
纯棉	0.00706058
修身	0.00687432
女童	0.00656358
夏装	0.00645988
春装	0.00585317
短裤	0.00581557

取出该客户的所有数值标签（事实类、规则类和预测类），如表11所示：

表11：客户数字标签

max_consume_amt	单次最大消费金额	29800
consume_order_ratio	消费订单比例	0.328765
mon_consume_frq	月均消费频度	166.214
consumption_channel	最常用支付工具	CreditCardPay
car_cnt	汽车消费次数	0
car_amt	汽车消费总金额	0
credit_card_repay_cnt	信用卡还款次数	81
credit_card_repay_amt	信用卡还款总金额	217134
is_installment	有无分期	1
cash_advance_cnt	预借现金次数	0
cash_advance_amt	预借现金总金额	0
payroll	有无代发	0
total_transactions_amt	交易总金额	4.79013e+06
total_transactions_cnt	交易次数	7078
total_deposit	ATM存款总金额	0
total_withdraw	ATM取款总金额	36400
transfer_cnt	转账次数	1074
transfer_amt	转账总金额	325052
transfer_mean	转账平均金额	302.655
internet_media_cnt	网络媒体类消费次数	0
internet_media_amt	网络媒体类消费总金额	0
public_pay_amt	公共事业缴费金额	385

> 这是一个组建了家庭的男士，并且有孩子。他的职业可能是保险类，汽车相关类。他的收入较高因为取款消费记录都较高，并且信用卡的消费力度也很大。

我们再随机挑出一个客户进行分析，通过对该客户的交易附言信息进行整合，画出客户交易附言的词云图，如图17所示：

图17：客户词云图

> 该客户拥有一个小孩子，喜欢购买生活类产品，喜欢正品，家具、衣物类产品，其很可能是一个家庭主妇。

接下来我们分析该客户的文本类标签，取出出现次数最多的前20个关键词，如表12所示：

表12：客户交易附言关键词

tmall	0.374911
纯棉	0.283061
儿童	0.279741
宝宝	0.26742
加厚	0.265128
婴儿	0.257679
taobao	0.223027
韩版	0.194658
时尚	0.189276
女士	0.181985
内裤	0.165739
保湿	0.154425
套装	0.148029
男士	0.1392
t恤	0.138722
大码	0.133426
孕妇	0.1165
大润发	0.109986
保暖	0.10998
男女	0.107079

取出该客户的所有数值标签（事实类、规则类和预测类），如表13所示：

表13：客户数字标签

max_consume_amt	单次最大消费金额	2570
consume_order_ratio	消费订单比例	0.697025
mon_consume_frq	月均消费频度	117.154
consumption_channel	最常用支付工具	AliPay
car_cnt	汽车消费次数	0
car_amt	汽车消费总金额	0
credit_card_repay_cnt	信用卡还款次数	25
credit_card_repay_amt	信用卡还款总金额	68494.8
is_installment	有无分期	0
cash_advance_cnt	预借现金次数	0
cash_advance_amt	预借现金总金额	0
payroll	有无代发	1
total_transactions_amt	交易总金额	567586
total_transactions_cnt	交易次数	2185
total_deposit	ATM存款总金额	0
total_withdraw	ATM取款总金额	0
transfer_cnt	转账次数	170
transfer_amt	转账总金额	145968
transfer_mean	转账平均金额	858.635
internet_media_cnt	网络媒体类消费次数	1
internet_media_amt	网络媒体类消费总金额	58
public_pay_amt	公共事业缴费金额	0

该客户经常购买孩子的相关产品和家具，同时也购买一些潮流的女性衣物和男性衣物。所以我们可以推测出该客户是一位拥有孩子的女性客户，她的消费方向主要是家庭中常用的衣物类，家具类。

4　精准营销应用

精准营销建立在客户历史数据的基础之上，通过客户之前的消费习惯来推测之后的消费倾向。通过分析前面构建的客户标签体系来个性化营销方案，这里我们主要参考两种方式。

- 商品兴趣度排行榜的构建：即计算客户对某类商品的兴趣度，根据兴趣度的排名进行营销。
- 目标客户的筛选：根据客户标签体系，筛选满足固定特点的群体进行营销。

4.1　商品兴趣排行榜的构建

4.1 商品兴趣排行榜的构建

从时间衰减、文本权重和消费金额三个方面分别计算客户对商品的兴趣度。将三个方面进行综合，得出总的兴趣度排行榜。

首先我们设定一个观察的时间轴，然后计算用户最近一次消费记录和这个时间轴的距离，并且除以pd.Timedlta(days=30)。这样处理的原因是可以有效的衡量消费者衰变的程度。

我们首先设置匹配的字符串，然后筛选出对应的消费记录。然后我根据user_id进行groupby()分组，讲分组后的结果进行求和，算得每位用户对应的特征数据。

从如上三个方面进行综合计算，以彩票类商品的消费为例，得到的兴趣度排行榜top10如表14所示：

表14：彩票类消费客户兴趣度排行榜

user_id	tfidf_sum	time_penalty	payment_sum	final_score
544518	0.993610	1.000000	0.487399	2.481008
22321749	1.000000	0.521733	0.600132	2.121865
17488993	0.999996	0.310164	0.750119	2.060280
19199845	0.999999	0.009039	0.978040	1.987078
15001742	0.981358	0.004678	1.000000	1.986036
12947427	0.992627	0.026958	0.956306	1.975891
11957177	1.000000	0.080297	0.881162	1.961460
11012971	1.000000	0.069979	0.880869	1.950848
3382860	1.000000	0.293317	0.654995	1.948311
8878476	0.999948	0.011284	0.933472	1.944704

取出排名前三的客户，画出三名客户在彩票类消费中交易次数和交易金额随时间的变化曲线。

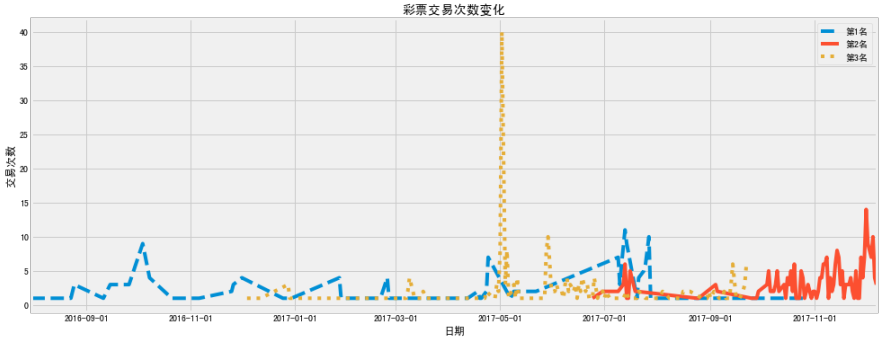

图18：彩票交易次数变化

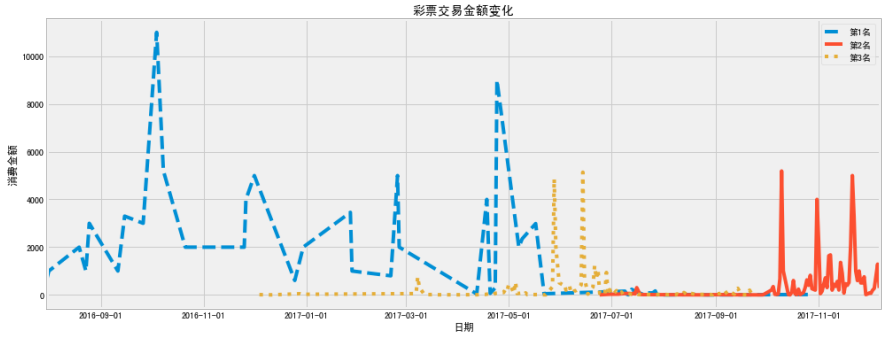

图19：彩票交易金额变化

当一位客户单次购买彩票消费金额较高，如第1名，或者其交易次数很多次，如第2类，又或者其交易金额和次数都相对较多，如第3名，那么该客户的总体得分就会靠前。

4.2 目标客户的筛选

在推销某类商品时，可以使用上一节的方法：计算平台客户在该类商品的兴趣度排行榜，参考排行榜选取排名靠前的客户。

另外一种方式为营销人员根据实际的业务需求，选取几个客户的关键特点，然后从客户数据库中查询符合关键特点的客户如图22所示：

图20：目标客户筛选

如果我们推销的是某类已有兴趣度排行的商品，那么我们可以基于方法一直接找到感兴趣的用户；如果我们要推荐的商品有很多的要求，需要分别筛选，那么第二种方法更合适。

5 项目总结与心得体会

5.1 项目总结

项目目的：解决了什么样的实际问题？ 本文基于用户以往的交易记录，来统计不同客户的特征，包括是否是高价值客户，其在不同商品的消费权重，然后可以根据我们产品的特点，如信用卡推广产品，找到有信用卡消费需求的产品。
项目流程：项目的总体执行流程是怎样的？ 首先对数据进行预处理，预处理缺失值，重复值，异常值等操作，然后我们得到事实类标签，根据一些经济学知识定义好我们的规则类标签，作图观察不同属性之间的联系，然后我们基于用户的文本标签构造其消费特点，通过词云等进行观察，画出用户画像。最后根据用户的标签和我们希望推广的产品的特点进行匹配，实现精准营销。
数据预处理：对客户交易数据进行了怎样的处理？ 对缺失值直接删除，对异常值如差一位的日期进行填充，对冗余值进行删除，同时也转换了日期的格式。
客户交易行为分析：从哪些维度对客户的交易数据进行了探索？有什么初步的结论？ 从客户的交易次数，交易金额，交易时间等进行统计，交易金额又可以细分为交易的平均额度，交易频率。结论：交易次数和交易时间主要集中在2016-2017的时间段，同时不同的客户我们建立他们的文本标签，也可以发现他们不同的交易倾向。
客户标签体系构建：使用了哪些标签构建客户标签体系？分析用户画像有什么结论？ 我们使用了事实类标签，规则类标签，文本类标签三种标签构建客户标签体系。事实类标签是已经存在的数据，规则类标签则是我们根据知识来计算的标签，文本标签则是从用户交易记录中包含文本信息的属性中提取有用的信息。分析用户画像，我们得到了不同的用户，其性别，消费领域，收入的大致信息。 精准营销应用：使用了哪些方法进行精准营销？有哪些应用场景？ 基于时间的商品兴趣度的计算，基于金额的商品兴趣度的计算，基于tf-idf兴趣度的计算，同时将这三个数据进行归一化处理得到总的评价score分数，按照分数高低进行排序。 应用场景：给用户推荐感兴趣的商品；筛选出符合信用卡消费习惯的客户，并进行银行卡的推广。
项目结论：通过本项目，你得到了哪些结论？ 基于对大数据的处理，我们可以有效的利用好数据，从数据中发现信息，将发现的信息和我们想要营销的商品信息相结合，两者匹配，找到我们的目标客户。

5.2 心得体会

知识，尤其是大数据的知识，不仅仅是在书本上，而更多的应该是从现实的实际案例中来学习。所以，要掌握好知识，就应该结合生活中的实际案例来学习。在本次实训项目的学习过程中，令我感受最深的是老师能将理论与实际知识相结合，再运用python语言，实现我们的想法。另一方面，我也感受到了Python第三方库的强大，众多的工具包能让我们很好的处理数据，从数据中挖掘信息。数据并不是信息，它需要我们利用工具去挖掘其潜在的知识，结合业务，达到我们的目的。