



**本科课程报告**

|  |  |
| --- | --- |
| 课程名称： | 金融科技导论 |
| 小组成员： | 王俊 3170100186 |
| 小组成员： | 黎睿翔 3180106171 |
| 报告名称： | Santander银行用户个性化产品推荐 |
| 指导教师： | 郑小林 |

**Santander 银行用户个性化产品推荐**

王俊 黎睿翔

# 研究背景及意义

## 研究背景

在大数据背景下，精准营销代表了数据的高度利用与效果的可量化性(阳冀，2016)。精准营销与传统营销不同的是，在大数据背景下，银行愿意关注更贴近自己需求的客户，而客户愿意关注更贴近自己需求的产品。传统商业银行进行客户细分的主要依据是客户的静态属性等维度，以此维度的细分进行产品营销与设计导致营销效果成功率较低，造成营销资源的浪费与营销成本的上升。传统的数据库用于存储、处理和管理结构化数据，结构化数据只关注于整个数据集的特定细节，无法处理此类数据的数量、种类和速度。对于大数据来说，大部分数据

都是基于非结构化和庞大的在数量上，倾向于用来进行比较和对比，从而在大尺度上建立整个数据集的规范和趋势。以往的理财顾问需要不断地询问客户需求，根据客户描述，找到相应的理财产品，这样的营销方式缺乏精准与效率。实施精准营销后通过对客户进行画像细分，并使用差异化产品组合和营销策略进行个性化推荐，提高营销的精准度。

本次大作业基于Kaggle 2016年Santander Product Recommendation竞赛。Santander Bank是北美最大的储蓄零售银行之一，也是世界上最受尊敬的银行之一。它的母公司Santander Group为来自英国、拉丁美洲和欧洲的超过一亿用户提供服务。Santander银行会为其客户提供个性化的产品推荐，然而在他们的当前系统下，少量的客户会收到大量推荐，而大多数用户很少收到推荐，导致用户体验不均衡。Kaggle给出了2015年至2016年5月部分现有用户的相关信息和银行产品的拥有情况，来预测2016年6月他们将使用哪些产品。

## 研究意义

有助于银行制定合理的指导政策，找出已完成的、未完成的、正在进行的任务，以期提出的先进理念对银行业破解发展瓶颈，为实现创新与发展带来保障。有效地融合大数据技术与精准营销理念，在技术与理论可行性结合的基础上，有可验证的购买银行理财产品案例，以落地效果评价模型与算法，丰富了精准营销的研究。

本次大作业研究是一个智能营销的案例，旨在为用户提供个性化的产品推荐，提高用户体验，从而得到顾客认可，获得销量和利润的快速增长。一个更好的智能营销模型可以在同类型产品的竞争种脱颖而出，从而达到最大限度地占有市场的意义。

# 国内外研究现状与存在问题

## 2.1 关于精准营销的国外研究

基于市场细分的角度，大多数是 Kotler Philip（1967）引申出来丰富后续学者的研究，选定一个或几个目标进行精准定位市场。基于产品生命周期理论角度，雷蒙德（1966）提出产品的引入、成长、繁荣、退出的周期理论进行相对应的营销策略。基于大数据的角度，Gouth R（1999）认为数据挖掘技术为精准营销提供高精度的预测功能和模式识别（William G. Zikmund 等，2010）。基于客户生命周期理论角度，从客户的考察、形成、成熟、退出的阶段理论进行相对应的营销。基于精准营销决策框架的角度，Tadić、ZečevićKrstić(2014)开发了一种新的混合指标模型，结合决策和评价模型来为决策者提供支持；Ghasab 等（2015）提出人工智能的决策支持框架。基于营销策略来看，Francine（2018）等认为遵循客户个性特征，推荐客户“感觉正确”的产品增强其心理体验，银行从积极的客户判断中获益；Jonne Guyt 等（2018）根据客户异质性，推荐品牌特色、模式特色、预算特色等对于销售预算的有效分分配影响销售收益；Sümeyra Atmaca 等（2017）

发现银行的忠诚度取决于一级及二级家庭成员的影响。

## 2.2关于精准营销的国内研究

基于市场细分的角度，陈豪（2011 年）对金融品牌与差异化客户细分市场；周晓梅（2012）对产品细分开发强化品牌特色。基于客户行为特征角度，刘畅（2015）与蔚赵春（2013）认为客户行为与偏好是精准营销最常用的预测，沈金波等（2015）认为客户的网络、手机、电话等客户的行为特征属性带来营销的精准定向；基于传统银行转型精准营销角度，陈剑光（2017）与高彤（2015）对零售银行的渠道受限、客户洞察不足、营销转化率低等方面不足建议建立交叉销售、客户细分、关联分析、客户关系、流失预警等分析，提出相应的对策与案例研究。基于服务目标角度，陈磊（2018）提出做好面向实体经济的精准营销；基于大数据技术角度，吕浩（2016）用大数据研究银行的精准营销、高彤（2015）用大数据技术从策略、渠道、成本、风险角度对提供银行精准营销； 基于渠道角度，伍万裕（2015）、张林达等（2014）、苑鑫斌（2014）、毛宇舟（2018）从渠道建设、开发、拓展以及微博渠道、电子渠道、智能客服运营渠道等提升营销的精准。

## 2.3国内外学术研究的不足

国内外对精准营销的理论研究成果颇多，很多都是基于市场细分理论、行为特征、转型策略比较、大数据技术运用、渠道选择、流程等方面进行了理论方面的研究，精准营销定量的研究颇为匮乏，模型评估与指标评估缺乏数据效用，在运用客户洞察、画像、系统、算法等技术方面支持的研究比较少。在实体商品购买、电商商品购买、通讯消费、医疗诊断等就用到客户画像，基于客户画像的个性化产品推荐提出可验证性的银行产品案例几乎没有。大多数营销文献对客户大致分为新客户、潜在客户、老客户，没有细化价值分类及提出相应的价值提升策略，这些仍需继续进行研究。

# 研究目标与研究内容

## 3.1 研究目标

根据Santander银行提供的用户信息和产品选择情况（2015.1-2016.5），预测2016年6月用户的产品选择，为其提供个性化产品推荐。

## 3.2 研究内容

结合金融科技导论理论课的学习内容，对原始数据进行数据分析，恰当合适的数据处理，填充空值，并且建立特征工程，包括对部分特征的分箱处理，one-hot编码，对分类特征的处理，滞后特征的添加，相关特征的删除以及特征组合等内容，结合不同产品的实际情况进行处理。划分训练集和验证集，选择训练模型。根据模型的实际结果对模型进行修改，以达到最优化模型的目的。

# 研究方法与模型思路

## 4.1 研究方法

# 数据集分析

## 5.1 原始数据分析

实验提供的原始数据有48个维度。包括用户信息（第1到24列），银行金融产品信息（第25到48列），分别为

**用户信息：**

fecha\_dato --标识日期，用于分类，取值都是一个月的28日。

fecha\_alta --客户成为银行第一个合同的第一个持有人的日期(有缺失)

ncodpers --客户代码

ind\_empleado --职员属性: 代表该用户是否为银行职员、前职员等

pais\_residencia --客户所在国

sexo --客户性别(有缺失)

age --年龄

ind\_nuevo --新客户指数。如果客户在过去6个月注册为1。(有缺失)

antiguedad --客户资历（几个月）

indrel -- 1第一/主客户；99当月主客户，月底除外(有缺失)

ult\_fec\_cli\_1t --作为主要客户（如果他不是在月底）的最后日期(有缺失)

indrel\_1mes --月初的客户类型，1（第一/主要客户），2（合伙人），P（潜在），3（原主要用户），4（前合伙人）(有缺失)

tiprel\_1mes --月初客户关系类型，A（活动），I（不活动），P（前客户），R（潜在）(有缺失)

indresi --如果居住国与银行所在国相同，居住指数（S是或N否）（有缺失)

indext --外国人指数（S（是=出生国家不同于银行国家）或N 否）(有缺失)

conyuemp --配偶指数。 如果客户是雇员的配偶为1(有缺失)

canal\_entrada --客户加入的渠道(有缺失)

indfall --已故指数。 N / S (有缺失)

tipodom --地址类型。 1，主地址(有缺失)

cod\_prov --省代码（客户地址）(有缺失)

nomprov --省名称(有缺失)

ind\_actividad -活动索引（1，活动客户; 0，不活动客户）(有缺失)

\_cliente

renta --家庭总收入(有缺失)

segmento --细分：01 - VIP，02 - 个人03 - 大学毕业生(有缺失)

**--------------------------------------------------------**

**产品信息：**

ind\_ahor\_fin\_ult1 --储蓄账户 ind\_aval\_fin\_ult1 --担保

ind\_cco\_fin\_ult1 --活期账户 ind\_cder\_fin\_ult1 -- derivada 帐户

ind\_cno\_fin\_ult1 --工资帐户 ind\_ctju\_fin\_ult1 --初级帐户

ind\_ctma\_fin\_ult1 --Más的特别账户 ind\_ctop\_fin\_ult1 --特定帐户

ind\_ctpp\_fin\_ult1 --特别账户 ind\_deco\_fin\_ult1 --短期存款

ind\_deme\_fin\_ult1 --中期存款 ind\_dela\_fin\_ult1 --长期存款

ind\_ecue\_fin\_ult1 --电子账户 ind\_fond\_fin\_ult1 --基金

ind\_hip\_fin\_ult1 --抵押 ind\_plan\_fin\_ult1 --养老金

ind\_pres\_fin\_ult1 --贷款 ind\_reca\_fin\_ult1 --税

ind\_tjcr\_fin\_ult1 --信用卡 ind\_valo\_fin\_ult1 --证券

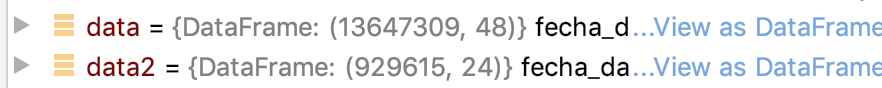
ind\_viv\_fin\_ult1 --家庭账户 ind\_nomina\_ult1 --工资(有缺失)

ind\_nom\_pens\_ult1--养老金(有缺失) ind\_recibo\_ult1 --直接借记

查看训练集和测试集的维度：







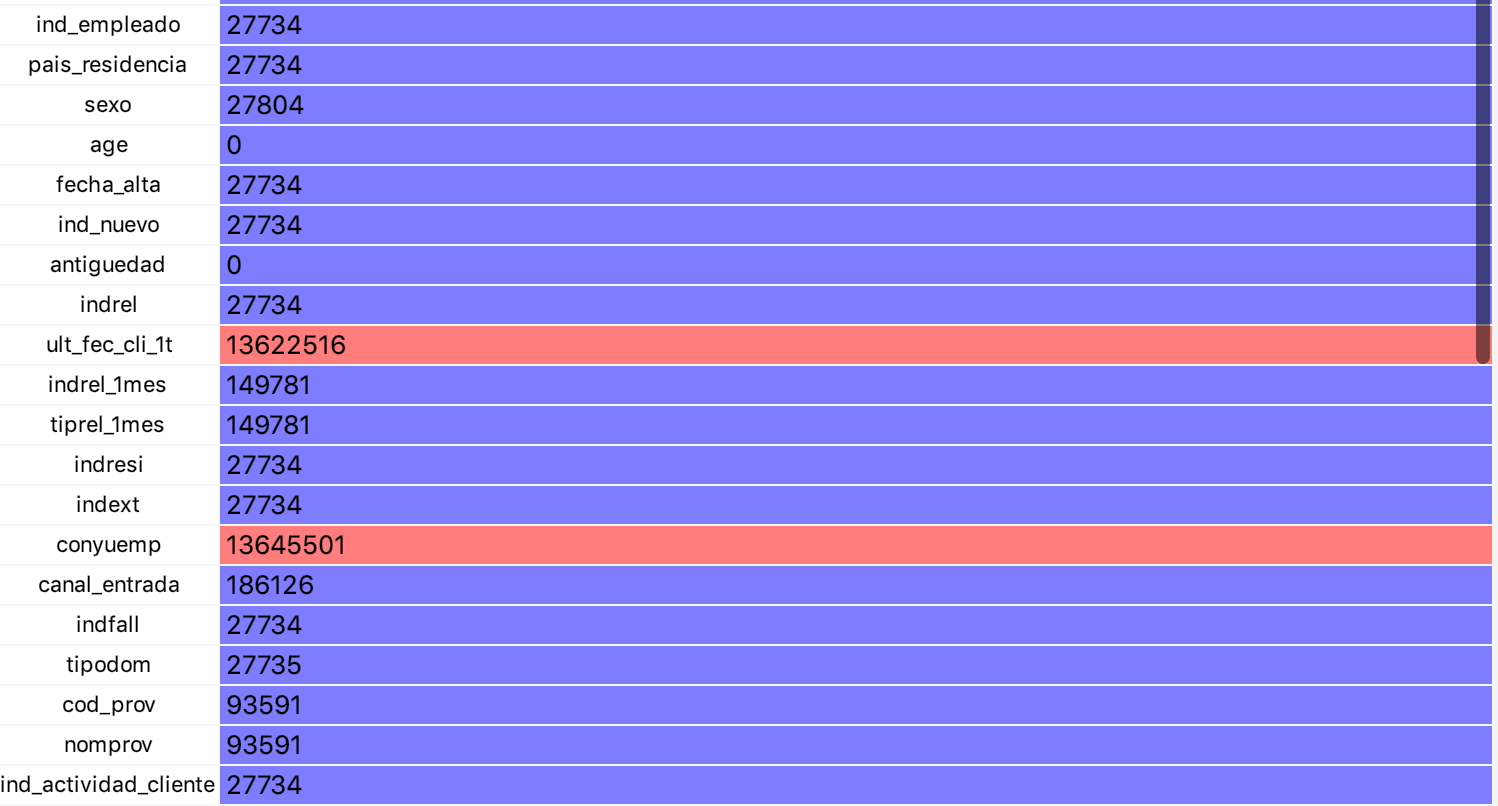
由于数据量较大，训练的时候避免消耗过大的内存，所以必须把float转换成int：



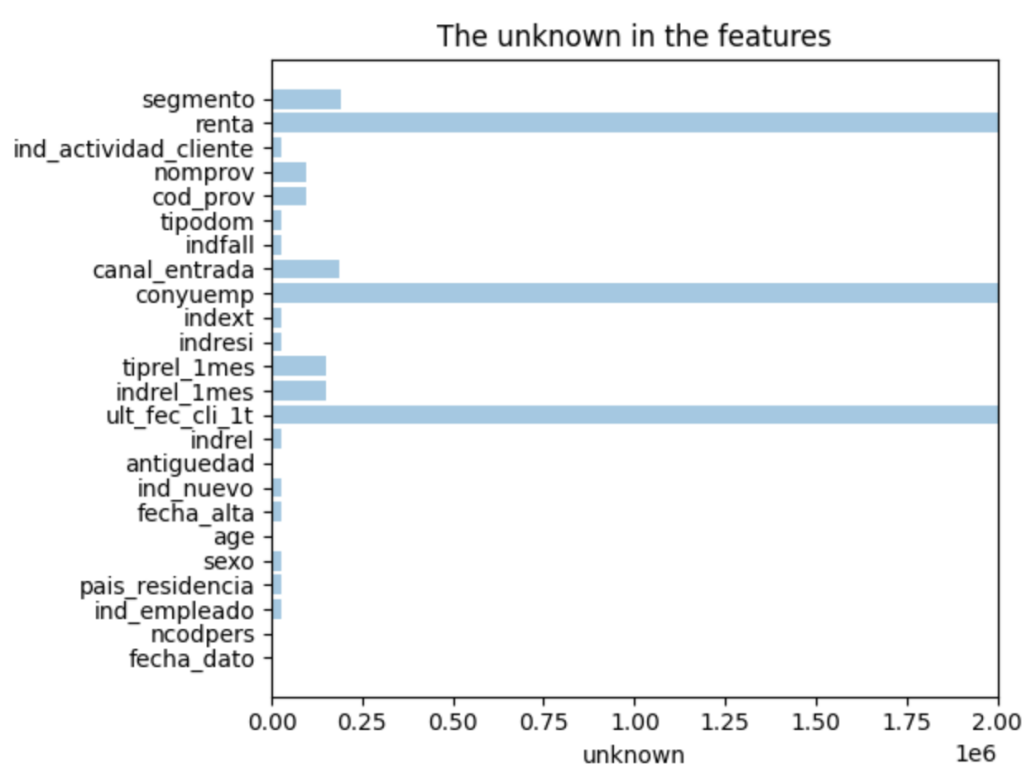
## 5.2 空值分析

* 参数部分的缺失

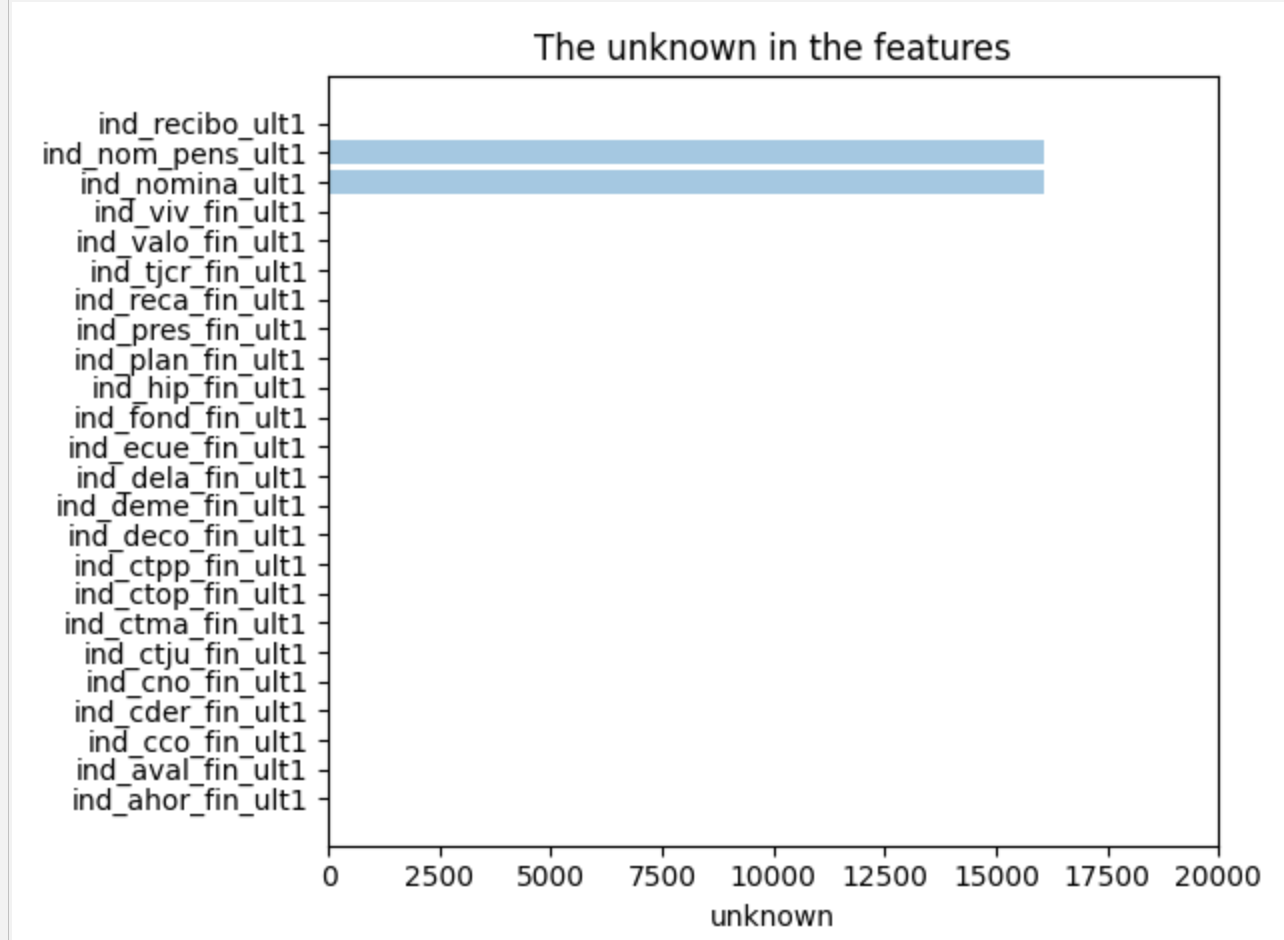




画图得：



* 目标部分的缺失

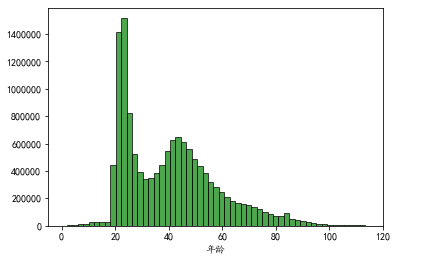


## 5.3 用户数据分析

在下面的数据分析中，采用未经处理的train\_ver2.csv进行分析。

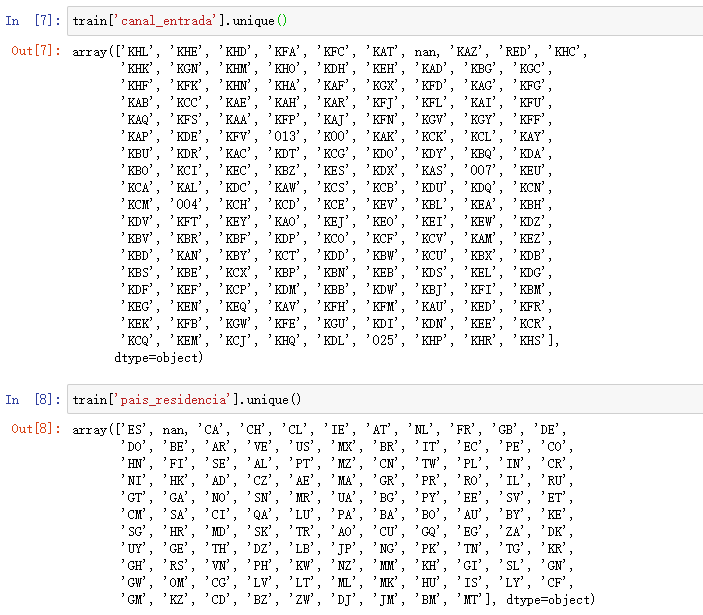
首先是用户年龄，采用下面的代码

|  |
| --- |
| plt.hist(pd.to\_numeric(train['age'],errors='coerce'),bins=80,facecolor='green',edgecolor='black',alpha=0.7)  plt.xlabel('年龄')  plt.xlim(-5,120)  ax=plt.gca()  ax.xaxis.set\_major\_locator(plt.MultipleLocator(20))  plt.show() |



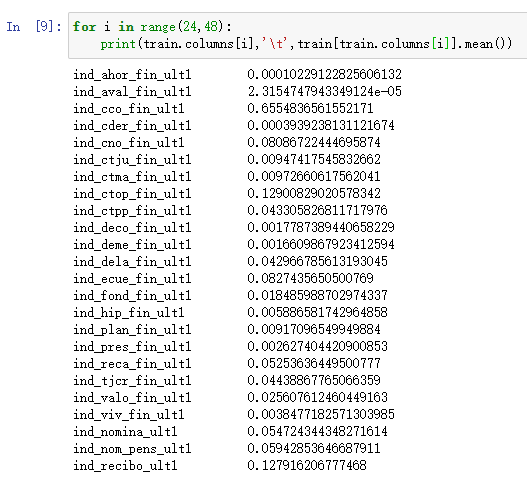
分析年龄分布，可以发现年龄基本呈现双峰分布。

然后是对用户加入渠道和用户所在国进行hash处理，下面unique观察对应特征的取值：



## 5.4 产品数据分析

下面分析对应每种产品选择人数情况，采用求每列平均值的方式来衡量，

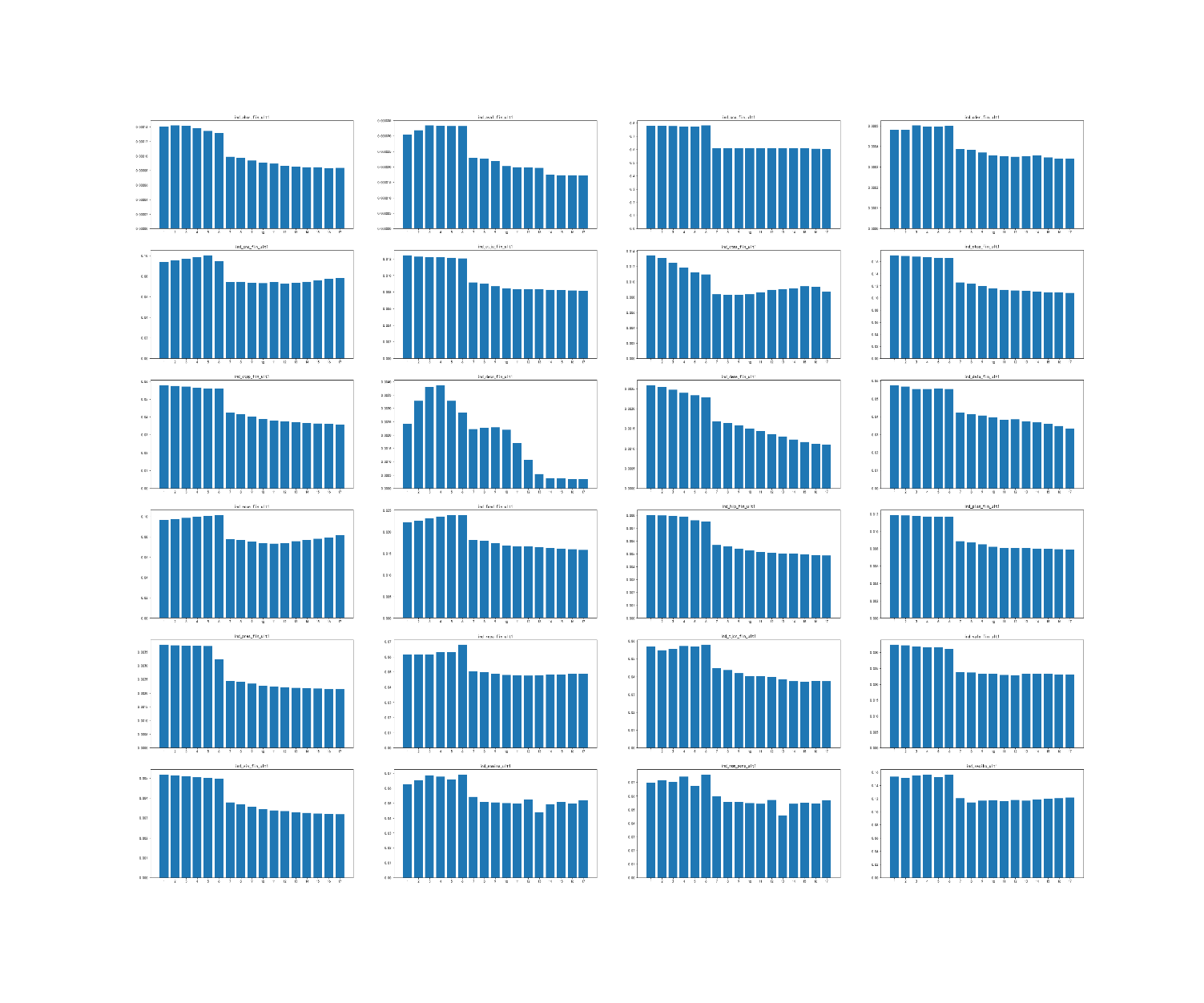


**可以发现ahor、aval、cder的平均值远低于0.005，属于选择率非常低的产品。viv、pres、deme、deco的平均值低于0.005，属于选择率低的产品。hip、plan的平均值低于0.01，的呼吁选择率一般的产品。**

此外，观察各产品逐月的受欢迎程度，采用逐月的平均值进行比较，代码如下：

|  |
| --- |
| months = ['1','2','3','4','5','6','7','8','9','10','11','12','13','14','15','16','17']  plt.figure(figsize=(48,40))  for i in range(24,48):  plt.subplot(6,4,i-23)  col = train.columns[i]  data = train[['fecha\_dato',col]].groupby('fecha\_dato').mean()[col].values  plt.title(col)  plt.bar(months,data)  plt.savefig('./1.png')  plt.show() |

得到的结果如下，具体可以到根目录下看。

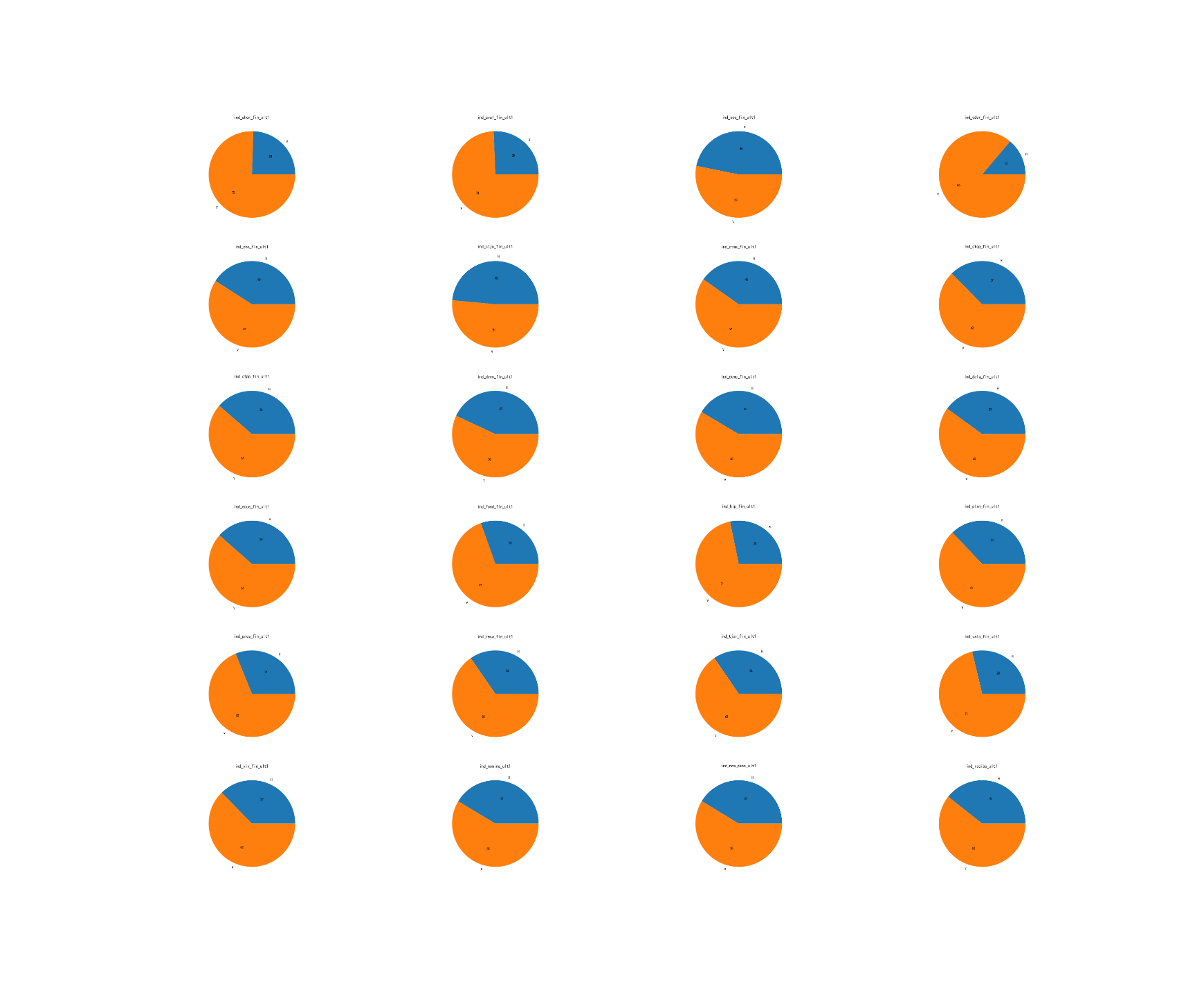


分析一下，绝大多数产品的受欢迎程度随时间的变化而逐渐变少，除了ecue和ctma以外。另外几乎所有的产品在前6个月的受欢迎情况与后面其他的月份有明显差异，因此需要分开处理。

关于性别对产品的影响，采用下面代码进行分析：

|  |
| --- |
| label\_list = ['H','V']  plt.figure(figsize=(48,40))  for i in range(24,48):  plt.subplot(6,4,i-23)  col = train.columns[i]  sum\_H = train[col][train['sexo']=='H'].sum()  sum\_V = train[col][train['sexo']=='V'].sum()  data = [sum\_H,sum\_V]  plt.pie(data,labels=label\_list,autopct='%d')  plt.title(col)  plt.savefig('./2.png')  plt.show() |

得到下面的结果，具体可以参见目录中的2.png。



可以发现不同产品对性别的吸引程度是完全不同的，普遍来说是V的数量大于H的个数，但是具体程度不相同。

一个特例：通过对产品的观察可以发现，拥有nomina产品的用户一定拥有nom\_pens，但是拥有nom\_pens的用户不一定拥有nomina产品。

下面是验证的代码：

|  |
| --- |
| print('Number of person who has nomina:',int(train['ind\_nomina\_ult1'].sum()))  print('Number of person who has nom\_pens:',int(train['ind\_nom\_pens\_ult1'].sum()))  nomina\_and\_nom = train['ind\_nomina\_ult1'].values+train['ind\_nom\_pens\_ult1'].values  cnt = 0  for node in nomina\_and\_nom:  if(node==2):  cnt+=1  print('Number of person who has nomina and nom\_pens:',cnt) |



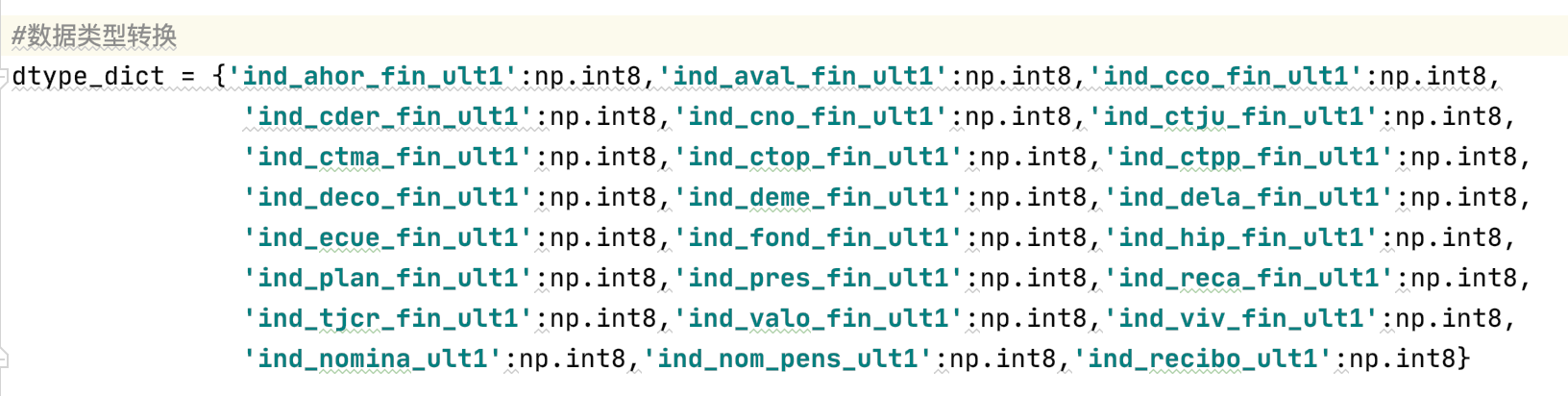
拥有nomina和共同拥有nomina、nom\_pens的用户是相同的。

# 实验与分析

## 6.1 数据预处理

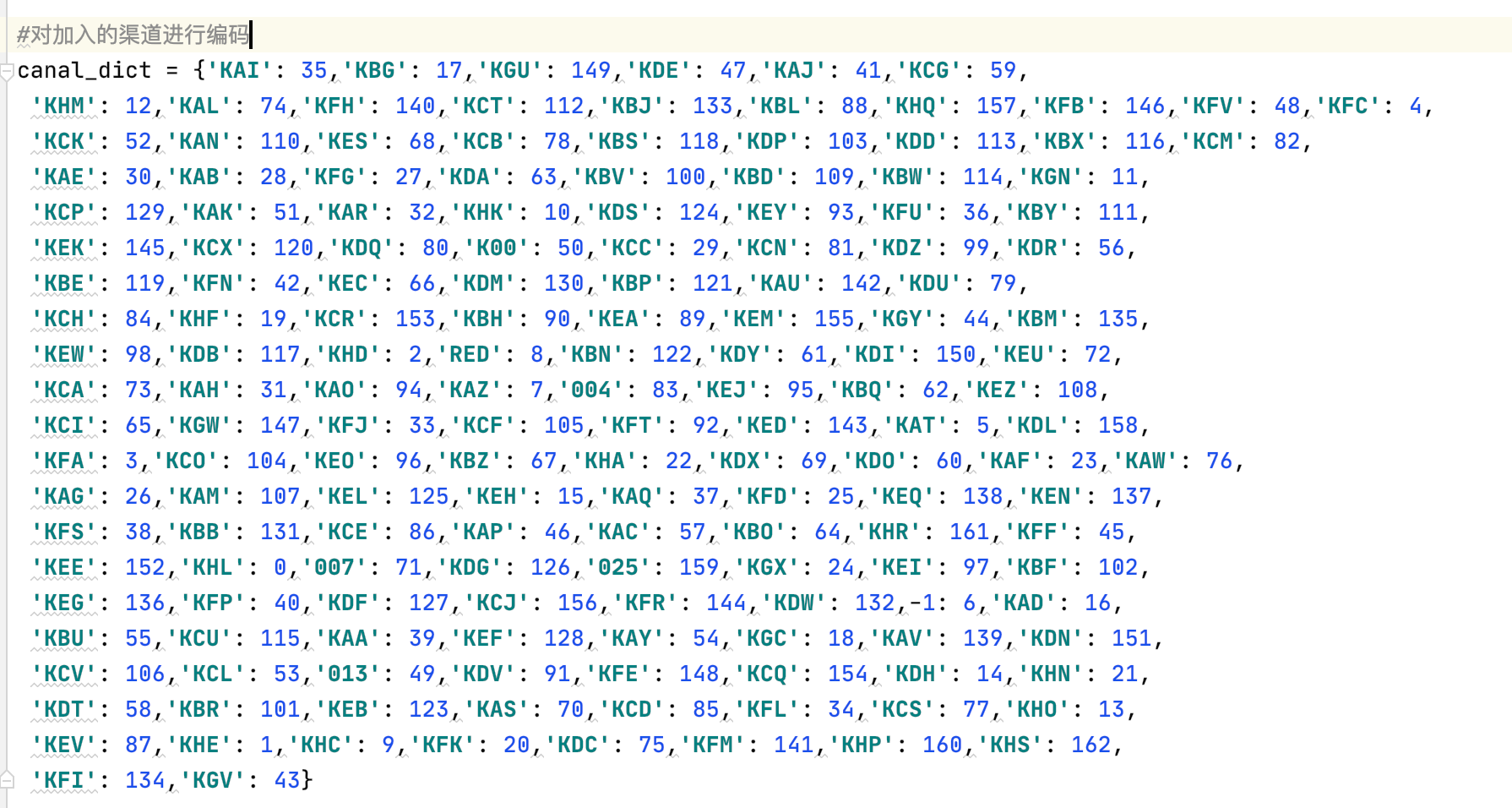
* 首先将目标列和信息列分开，并进行数据类型转换：

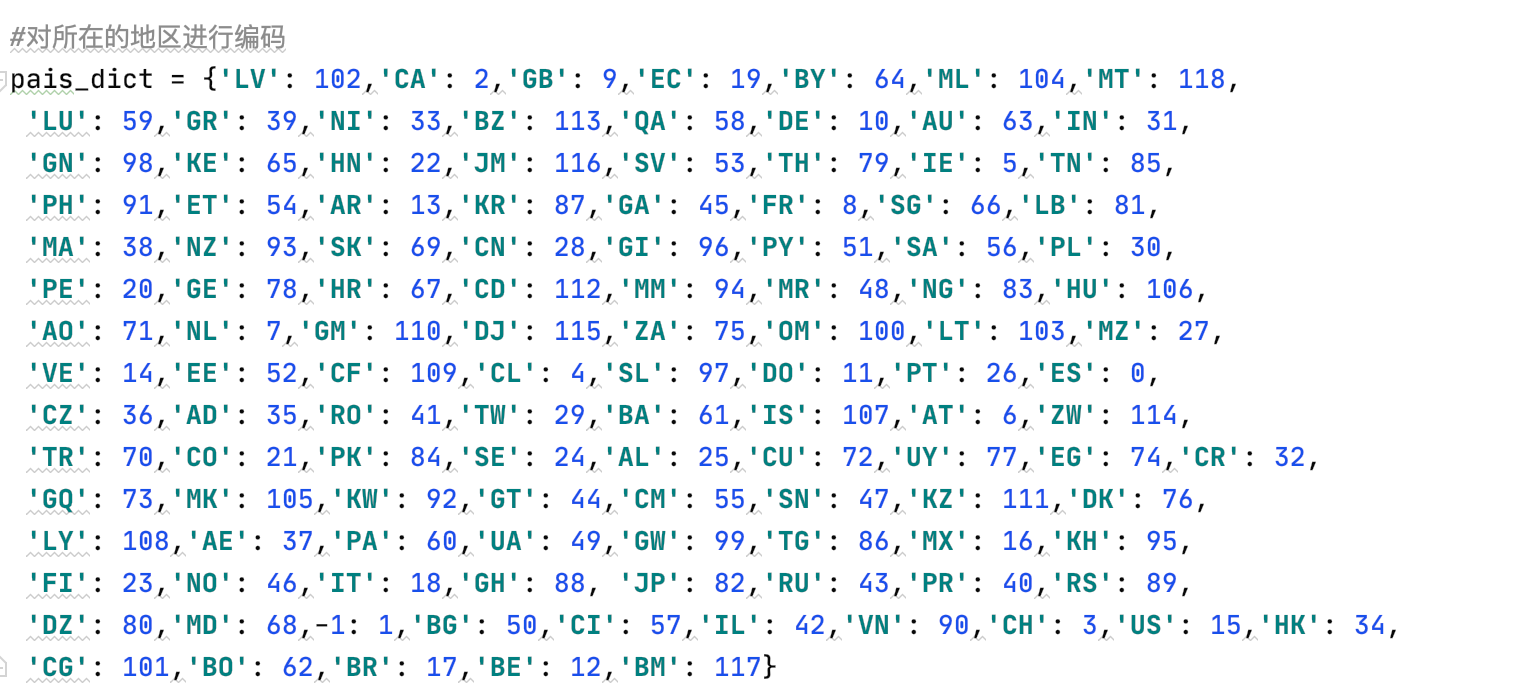




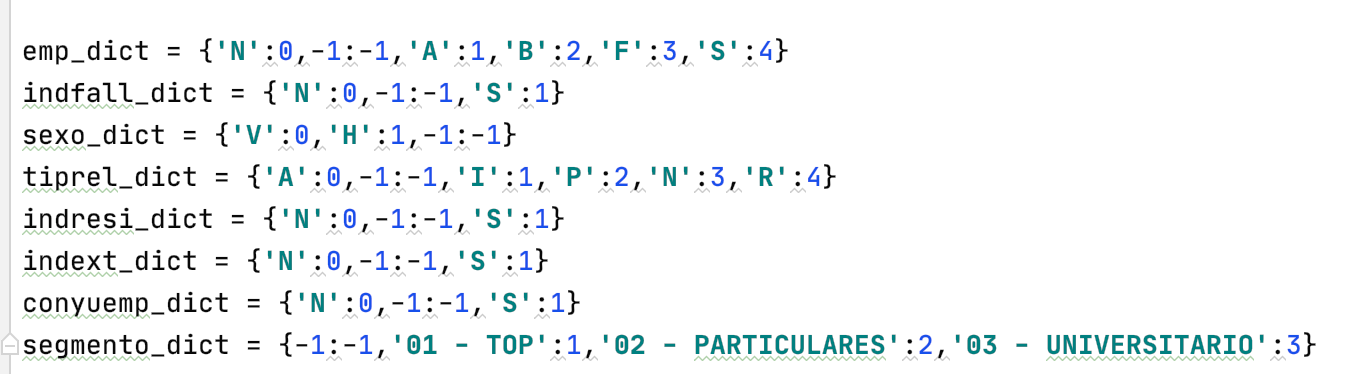
* 观察数据含义，找出示性信息进行编码

最难处理的是所在地区和加入的渠道，信息的值比较多，如下：



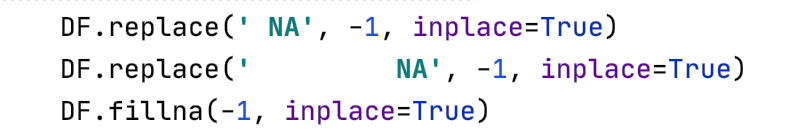


对于其他的示性数据也一样

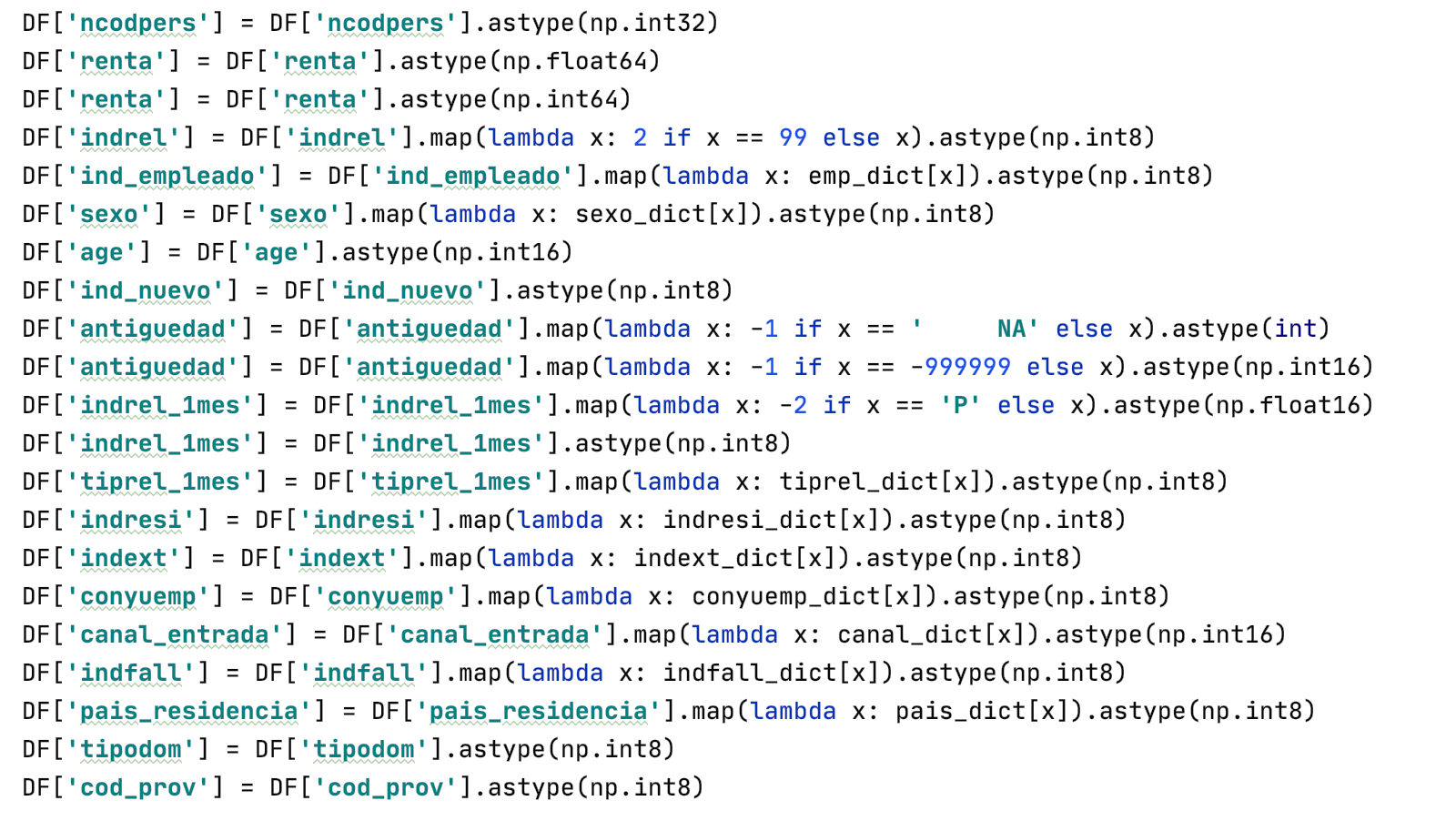


* 数据预处理：

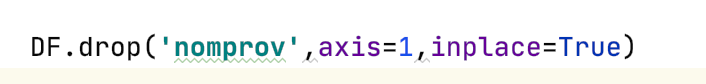
先填补空值：



然后对其他数据进行处理，示性数据根据字典转换，为了方便计算，最后都转换成int的格式：

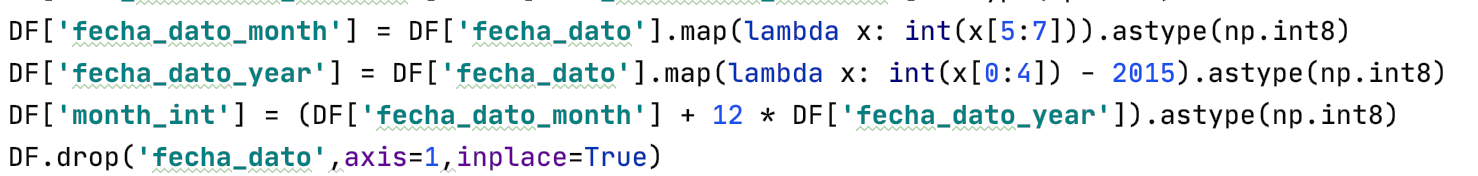


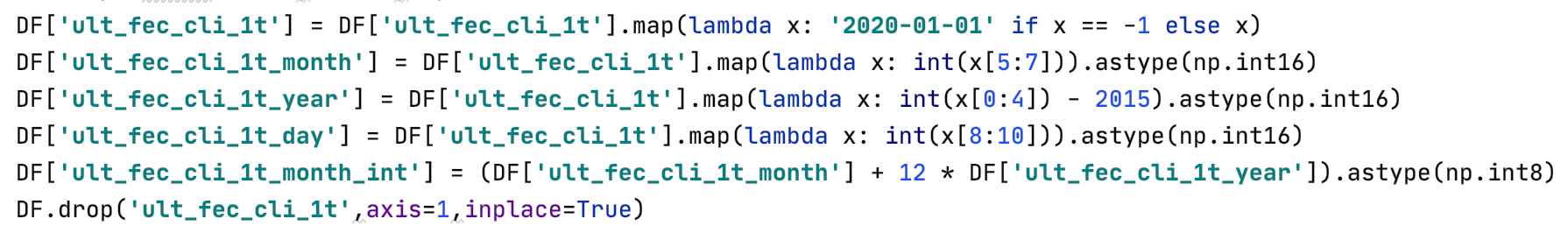
* **特别注意一下：cod\_prov是省的编号，nomprov是省的名字，只需要留下一个就可以了；**



* **对日期信息进行处理**

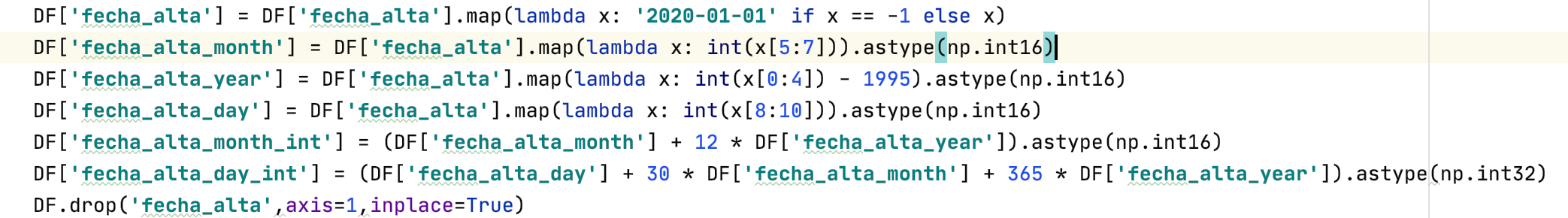
--从fecha\_dato和ult\_fec\_cli\_1t中抓取年月的信息后删除





由于数据从2015-2016，所以year还是进行了一个对2015的减法；

--从fecha\_alta进行同样的处理，同时把day天数也进行了运算



最后生成了train\_hash和test\_hash两个文件：

由于数据较大，load比较长时间；

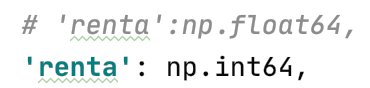




相比于原来多了日期拆分后的维度；

## 6.2 数据预处理

本来想renta当成float处理，但由于数据太大内存跟不上还是放弃了

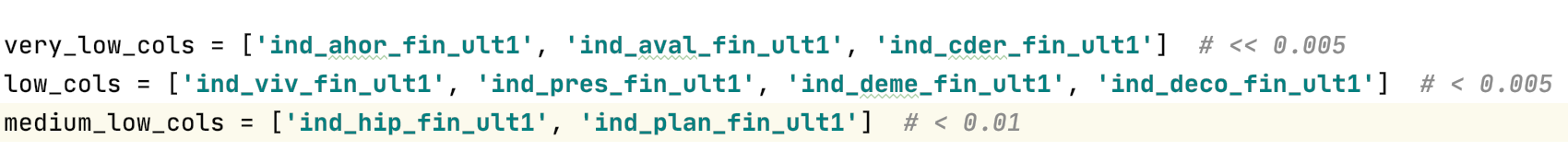




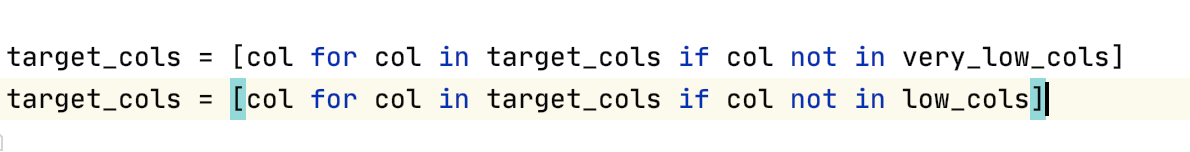
* 处理renta的缺失值



* 根据前面的数据分析将产品分为很不受欢迎，不受欢迎，中等受欢迎：

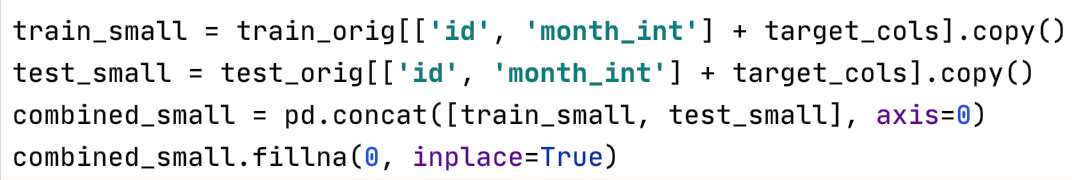


然后去掉很不受欢迎，不受欢迎的来进行预测：

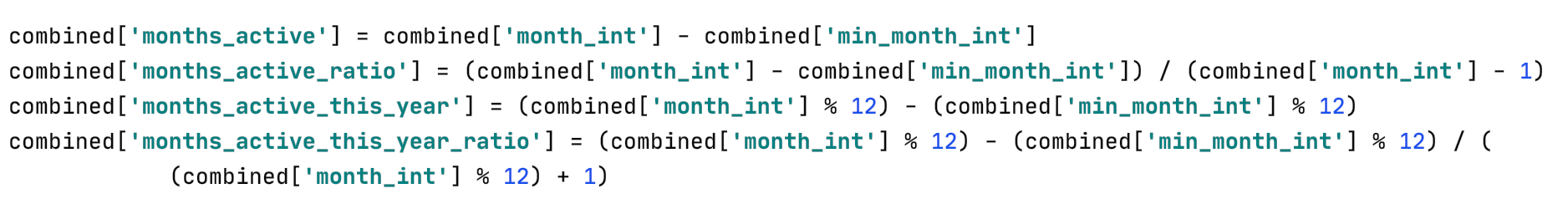


* 把月份和id值与要求的产品以及上面去掉的low和very low的产品信息combine在一起

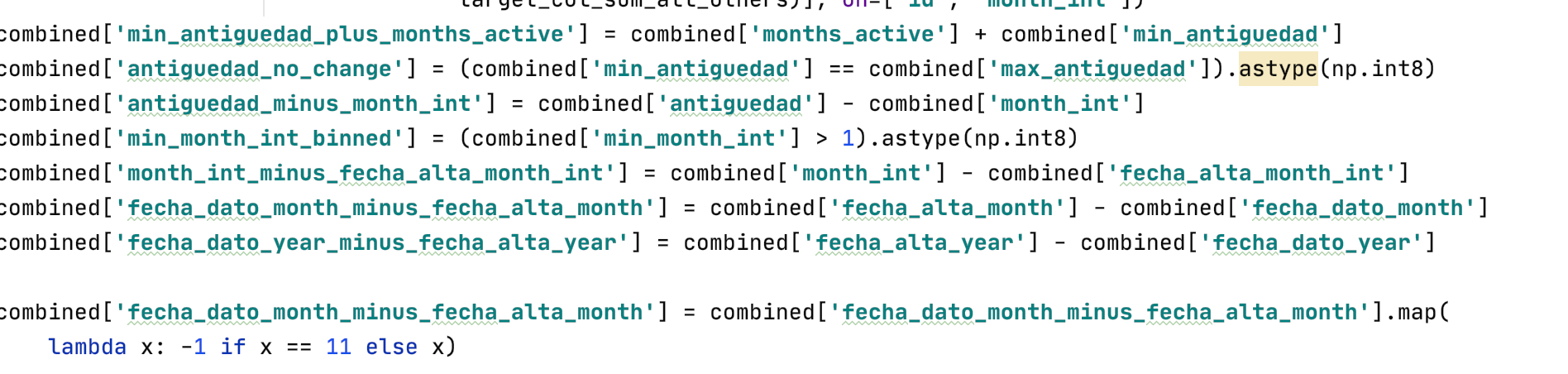




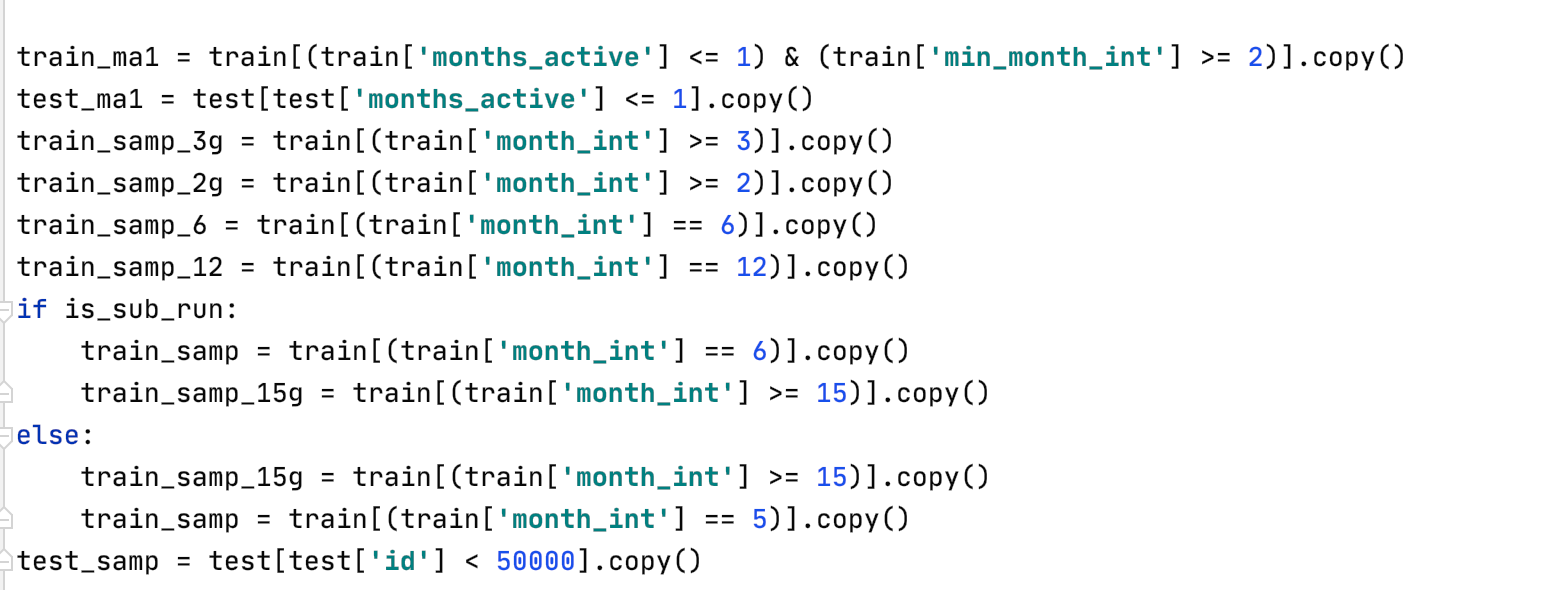
* **然后对数据型变量进行离散分箱处理，特别注意这里找出来活跃的月份数因为它才是真正有意义的数据：**

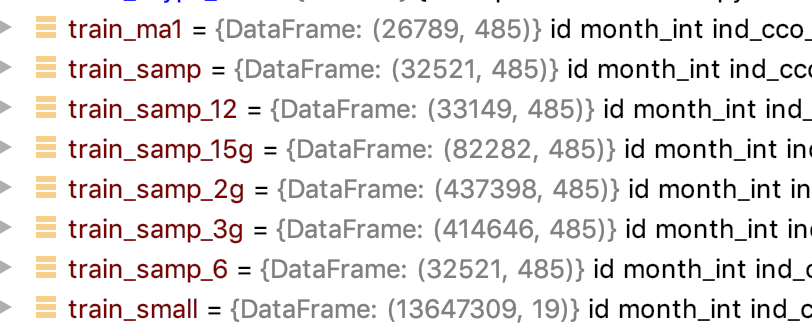


根据客户资历和之前的月份进行运算得到新的特征值：



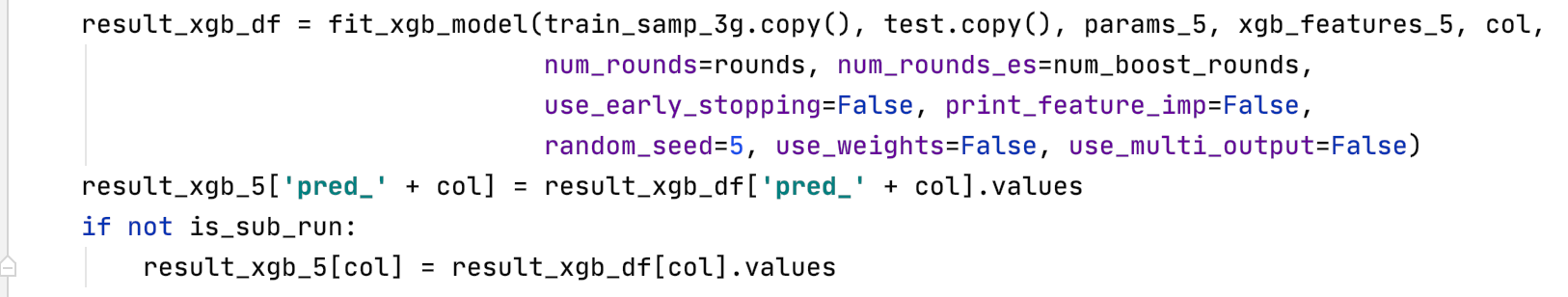
* **根据月份数进行分组进行模型的调参测试：**



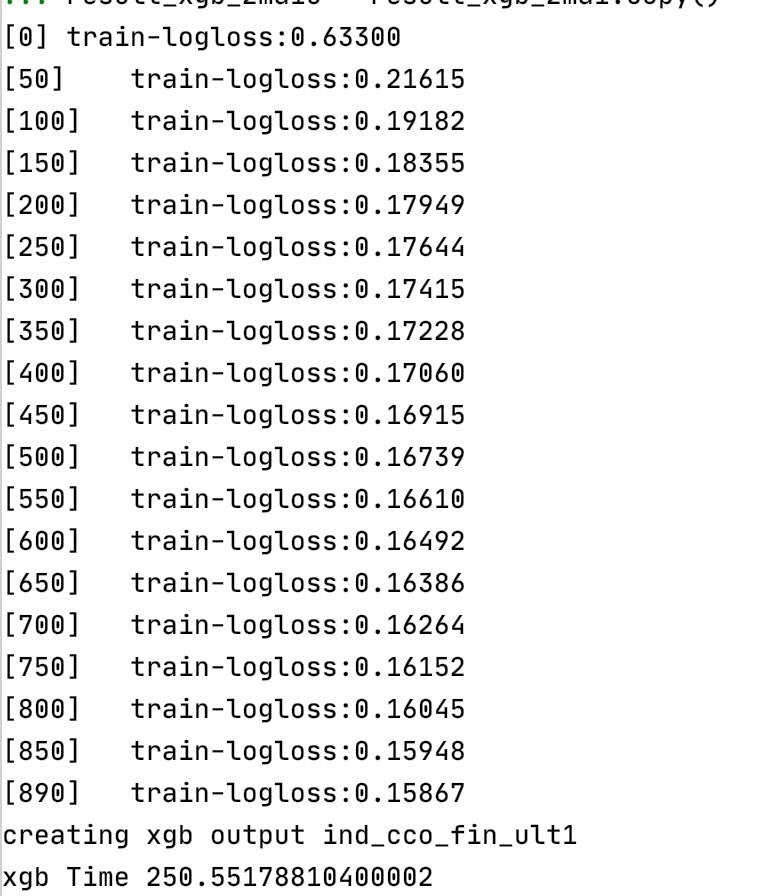


这里取了5个月，6个月，12个月为一季度的用户进行测试，因为原数据量比较大，直接拿来对模型进行测试不好调参，所以选取一定月数的用户作为样本

* **用xgb模型对不同的特征和样本进行训练：**

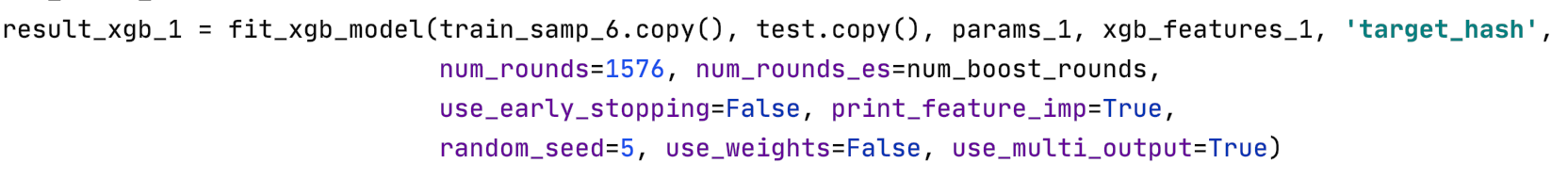


Result\_xgb\_5是用3g即使用超过3个月的数据对第一个产品进行预测，而特征是将日期，月份数去掉，增加了活跃月份，总收入等特征



可以看出这个损失值下降的比较快，效果还是可以的；

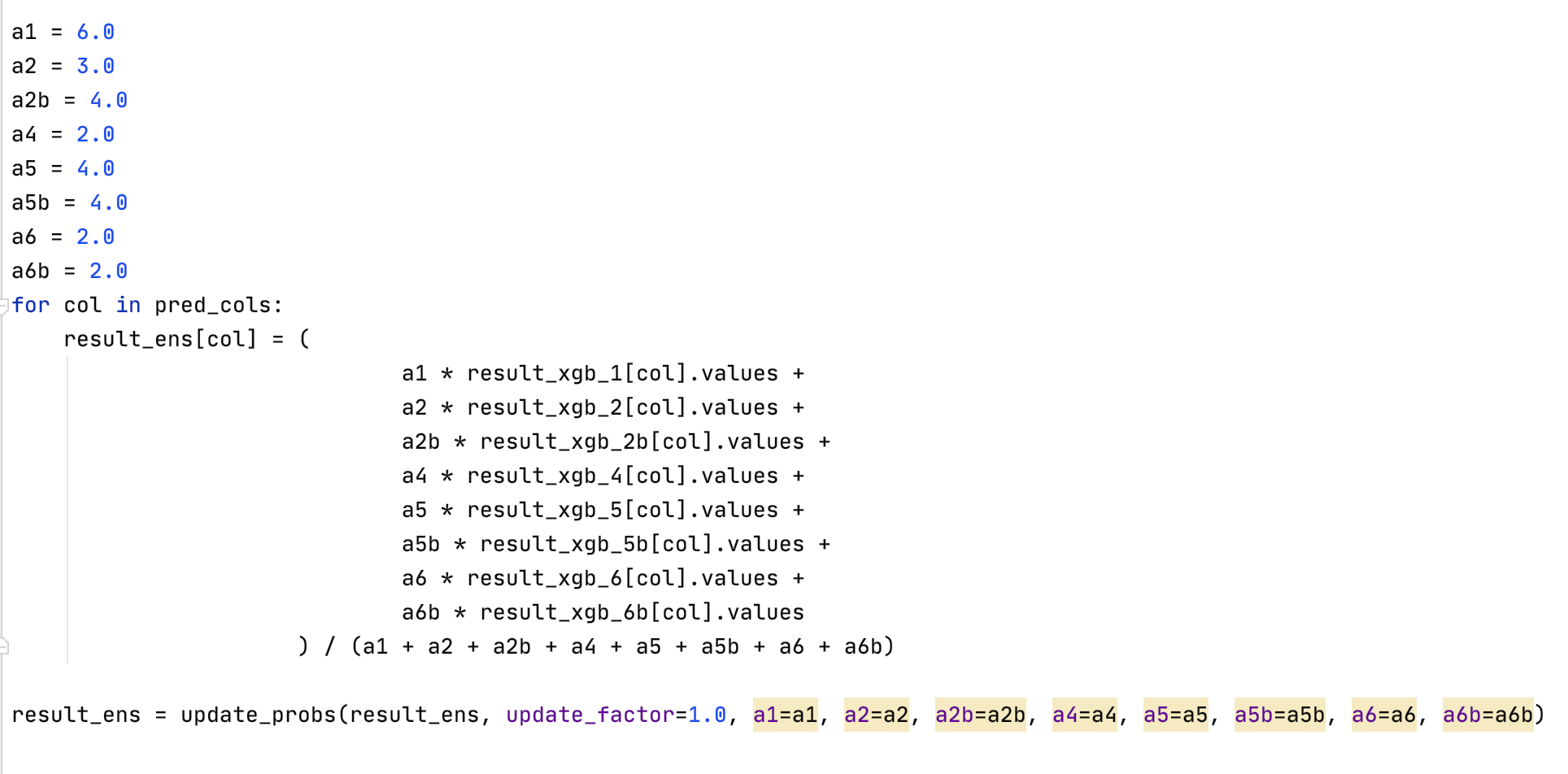
Result\_xgb\_1是用6个月的作为训练集进行训练



**整体的思想就是尽可能地划分数据集，不直接对大型的数据进行训练，根据不同的特征和选取的样本最后生成了以下的result：**



最后，再分别对他们进行加权处理：



比如由于result\_xgb\_5的数据是用了month>=3，肯定比只使用一个季度的结果更加准确，加权的值自然也越大。

# 结论与展望

进一步的改进：

有时似乎多个客户属于同一个家庭，因为他们的房租完全相同，有时他们的ncodpers甚至彼此紧挨着（例如1108430和1108431）。在上周，我花了很多时间试图从这些帐户关系中榨取信息，但我无法将其转化为得分的提高。我很想知道是否有人这样做。