|  |
| --- |
| **机器学习** |
| **桑坦德银行产品推荐** |
| 主 研 人：李琦  参 研 人：李琦 |
| 审 核 人：  方 向：工业算法案例研究  版 本 号：A |
|  |

**版本更新**

|  |  |  |  |
| --- | --- | --- | --- |
| **日期** | **版本** | **更新描述** | **作者** |
| 2018/10/31 | A | 初稿 | 李琦 |

**选题表格**

|  |  |  |  |
| --- | --- | --- | --- |
| 时间 | 竞赛名 | 竞赛背景描述（50字以内） | 类型（分类/回归） |
| 2019/03/05 | Sandanter 产品推荐（kaggle） | 桑坦德银行想通过更有效的推荐系统，更好地满足所有客户的个性化需求。比赛根据客户过去的行为来预测现有客户将在下个月使用哪些产品。 | 分类 |

目录

[1. 背景描述](#_Toc19702_WPSOffice_Level1) [4](#_Toc19702_WPSOffice_Level1)

[1.1 竞赛赛题描述](#_Toc20566_WPSOffice_Level2) [4](#_Toc20566_WPSOffice_Level2)

[1.2 评估指标描述](#_Toc32164_WPSOffice_Level2) [4](#_Toc32164_WPSOffice_Level2)

[2. 数据来源及描述性统计分析](#_Toc20566_WPSOffice_Level1) [4](#_Toc20566_WPSOffice_Level1)

[2.1 大赛数据来源](#_Toc28485_WPSOffice_Level2) [4](#_Toc28485_WPSOffice_Level2)

[2.2 数据的描述性统计](#_Toc17269_WPSOffice_Level2) [5](#_Toc17269_WPSOffice_Level2)

[2.2.1 数据基本情况描述：](#_Toc20566_WPSOffice_Level3) [5](#_Toc20566_WPSOffice_Level3)

[2.2.2 数据字段介绍：](#_Toc32164_WPSOffice_Level3) [5](#_Toc32164_WPSOffice_Level3)

[2.2.3 数据描述性统计](#_Toc28485_WPSOffice_Level3) [8](#_Toc28485_WPSOffice_Level3)

[3. 优秀算法思路](#_Toc32164_WPSOffice_Level1) [17](#_Toc32164_WPSOffice_Level1)

[3.1 方案一](#_Toc19099_WPSOffice_Level2) [17](#_Toc19099_WPSOffice_Level2)

[3.1.1 方案一数据预处理及特征工程部分方案](#_Toc17269_WPSOffice_Level3) [17](#_Toc17269_WPSOffice_Level3)

[3.1.2 方案一模型设计、建立部分方案](#_Toc19099_WPSOffice_Level3) [20](#_Toc19099_WPSOffice_Level3)

[3.1.3 方案一结果、排名等](#_Toc27957_WPSOffice_Level3) [28](#_Toc27957_WPSOffice_Level3)

[3.1.4 方案一算法流程图](#_Toc21429_WPSOffice_Level3) [28](#_Toc21429_WPSOffice_Level3)

[3.2 方案二](#_Toc27957_WPSOffice_Level2) [29](#_Toc27957_WPSOffice_Level2)

[3.2.1 方案二数据预处理及特征工程部分方案](#_Toc4600_WPSOffice_Level3) [29](#_Toc4600_WPSOffice_Level3)

[3.2.2 方案二模型设计、建立部分方案](#_Toc11226_WPSOffice_Level3) [29](#_Toc11226_WPSOffice_Level3)

[3.2.3 方案二结果、排名等](#_Toc8802_WPSOffice_Level3) [30](#_Toc8802_WPSOffice_Level3)

[3.2.4 方案二算法流程图](#_Toc27565_WPSOffice_Level3) [30](#_Toc27565_WPSOffice_Level3)

[3.2 方案三](#_Toc21429_WPSOffice_Level2) [31](#_Toc21429_WPSOffice_Level2)

[3.3.1 方案三数据预处理及特征工程部分方案](#_Toc23770_WPSOffice_Level3) [31](#_Toc23770_WPSOffice_Level3)

[3.3.2 方案三模型设计、建立部分方案](#_Toc26397_WPSOffice_Level3) [31](#_Toc26397_WPSOffice_Level3)

[3.3.3 方案三结果、排名等](#_Toc8227_WPSOffice_Level3) [32](#_Toc8227_WPSOffice_Level3)

[3.3.4 方案三算法流程图](#_Toc23057_WPSOffice_Level3) [32](#_Toc23057_WPSOffice_Level3)

[4. 算法比较](#_Toc28485_WPSOffice_Level1) [33](#_Toc28485_WPSOffice_Level1)

[表4-1 算法比较](#_Toc4600_WPSOffice_Level2) [33](#_Toc4600_WPSOffice_Level2)

[5. 总结与展望](#_Toc17269_WPSOffice_Level1) [34](#_Toc17269_WPSOffice_Level1)

[5.1 总结](#_Toc11226_WPSOffice_Level2) [34](#_Toc11226_WPSOffice_Level2)

[5.2 建模思路](#_Toc8802_WPSOffice_Level2) [34](#_Toc8802_WPSOffice_Level2)

1. 背景描述

准备在你的第一所房子上支付首期款项？或者希望利用您拥有的房屋的资产？为了支持一系列财务决策的需求，[桑坦德银行](https://www.santanderbank.com/us/personal" \t "https://www.kaggle.com/c/_blank)通过个性化的产品推荐为其客户提供贷款。

**1.1 竞赛赛题描述**

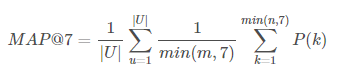
在他们目前的系统下，少数桑坦德银行的客户会收到很多建议，而其他许多客户很少看到建议，这导致客户体验不平衡。Santander要求挑战者根据客户过去的行为和相似客户的行为来预测他们现有客户将在下个月使用哪些产品。

通过更有效的推荐系统，桑坦德银行可以更好地满足所有客户的个性化需求，无论他们身在何处，都能确保满意。

**免责声明：**此数据集不包括任何真正的桑坦德西班牙客户，因此它不代表西班牙的客户群。

**1.2 评估指标描述**

平均精度均值@7 (MAP@7):



其中|U|为行数(在两个时间点的用户)，P(k)为截止k处的精度，n为预测产品的数量，m为给定用户在该时间点上添加的产品数量。如果m = 0，则精度定义为0。

1. 数据来源及描述性统计分析

**2.1 大赛数据来源**

在本次竞赛中，由Santander银行提供1.5年的客户行为数据，以预测客户将购买哪些新产品

**注意：此样本不包括任何真正的桑坦德西班牙客户，因此它不代表西班牙的客户群。**

[这里是数据的超链接。](https://www.kaggle.com/c/santander-product-recommendation/data)

**2.2 数据的描述性统计**

**2.2.1 数据基本情况描述：**

数据从2015年1月28日开始，包含客户拥有的产品的月度记录，例如“信用卡”，“储蓄账户”等。您将预测客户在2016年5月28日已经拥有的产品之外，在2016年6月6日-28日一个月还将获得哪些额外的产品。这些产品是名为:ind\_(xyz)\_ult1的列，这是训练数据中的#25 - #48列。您将预测客户在2016-05-28已经拥有的基础上还将购买什么。

**2.2.2 数据字段介绍：**

训练数据包括近100万用户，每月有2015年1月至2016年5月的历史用户和产品数据。用户数据包括24个预测变量，包括用户的年龄和收入。产品数据包含所有24种产品的布尔标志，并指示用户是否在相应月份拥有该产品。目标是预测哪个****新的**** 产品是929,615测试用户最有可能在2016年6月购买的。如果产品在2016年6月拥有但2016年5月没有，则该产品被认为是新产品。

**表2-1 train数据表字段介绍**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 变量含义 | 变量类型 | 缺失率 |
| fecha\_dato | 该表已为此列分区 | 日期 | 0% |
| ncodpers | 客户代码 | 离散 | 0% |
| ind\_empleado | 员工指数：A活跃，B前聘用，F孝顺，N不雇员，P 被动的 | 分类（字符） | 0.3962 %  003962 |
| pais\_residencia | 客户的国家居住地 | 字符 | 0.3962 % |
| sexo | 顾客的性别 | 分类（字符） | 0.3967 % |
| age | 年龄 | 离散 | 2.78% |
| fecha\_alta | 客户成为银行合同的第一持有人的日期 | 日期 | 0.3962 % |
| ind\_nuevo | 新客户指数。如果客户在过去6个月内注册，则为1。 | 分类（数字） | 0.3962 % |
| antiguedad | 客户资历（以月计） | 离散 | 2.78% |
| indrel | 1（第一/主要），99（本月的主要客户，但不是月末） | 分类（数字） | 0.3962 % |
| ult\_fec\_cli\_1t | 作为主要客户的最后日期（如果他不是在月末） | 日期 | 99.8436 % |
| indrel\_1mes | 月初的客户类型，1（第一/主要客户），2（共同所有者），P（潜在），3（前主要），4（前共同所有者） | 分类（数字+字符） | 1.4972 % |
| tiprel\_1mes | 月初的客户关系类型，A（有效），I（无效），P（前客户），R（潜在） | 分类（字符） | 1.4972 % |
| indresi | 住宅指数（如果居住国家与银行国家相同，则为S（是）或N（否）） | 分类（字符） | 0.3962 % |
| indext | 外国人指数（如果客户的出生国家与银行国家不同，则为S（是）或N（否）） | 分类（字符） | 0.3962 % |
| conyuemp | |  | | --- | | 配偶指数。1如果客户是雇员  的配偶（'N' 'S'） | | 分类（字符） | 99.9857 % |
| canal\_entrada | 客户用来加入的渠道 | 字符 | 1.7036 % |
| indfall | 已故指数。N / S | 分类（字符） | 0.3962 % |
| tipodom | |  | | --- | | 地址类型。1，主要地址 | | 分类（数字） | 0.3962 % |
| cod\_prov | 省代码（客户地址） | 离散 | 0.9156 % |
| nomprov | 省名称 | 字符 | 0.9156 % |
| ind\_actividad\_cliente | 活动指数（1，活跃客户; 0，非活跃客户） | 分类（数字） | 0.3962 % |
| renta | 家庭总收入 | 离散 | 17.8894 % |
| segmento | 细分：01 - VIP，02 - 个人03 - 大学毕业 | 分类 | 1.7189 % |
| ind\_ahor\_fin\_ult1 | 储蓄账户 | 分类（0/1） | 0 % |
| ind\_aval\_fin\_ult1 | 担保 | 分类（0/1） | 0 % |
| ind\_cco\_fin\_ult1 | |  | | --- | | 当前账户 | | 分类（0/1） | 0 % |
| ind\_cder\_fin\_ult1 | Derivada帐户 | 分类（0/1） | 0 % |
| ind\_cno\_fin\_ult1 | 工资账户 | 分类（0/1） | 0 % |
| ind\_ctju\_fin\_ult1 | 初级账户 | 分类（0/1） | 0 % |
| ind\_ctma\_fin\_ult1 | Más特别帐户 | 分类（0/1） | 0 % |
| ind\_ctop\_fin\_ult1 | 特别帐户 | 分类（0/1） | 0 % |
| ind\_ctpp\_fin\_ult1 | 特别加号帐户 | 分类（0/1） | 0 % |
| ind\_deco\_fin\_ult1 | 短期存款 | 分类（0/1） | 0 % |
| ind\_deme\_fin\_ult1 | 中期存款 | 分类（0/1） | 0 % |
| ind\_dela\_fin\_ult1 | 长期存款 | 分类（0/1） | 0 % |
| ind\_ecue\_fin\_ult1 | 电子邮件帐户 | 分类（0/1） | 0 % |
| ind\_fond\_fin\_ult1 | 资金 | 分类（0/1） | 0 % |
| ind\_hip\_fin\_ult1 | 抵押 | 分类（0/1） | 0 % |
| ind\_plan\_fin\_ult1 | 养老金 | 分类（0/1） | 0 % |
| ind\_pres\_fin\_ult1 | 贷款 | 分类（0/1） | 0 % |
| ind\_reca\_fin\_ult1 | 税 | 分类（0/1） | 0 % |
| ind\_tjcr\_fin\_ult1 | 信用卡 | 分类（0/1） | 0 % |
| ind\_valo\_fin\_ult1 | 证券 | 分类（0/1） | 0 % |
| ind\_viv\_fin\_ult1 | 家庭帐户 | 分类（0/1） | 0 % |
| ind\_nomina\_ult1 | 工资表 | 分类（0/1） | 0.2295 % |
| ind\_nom\_pens\_ult1 | 养老金 | 分类（0/1） | 0.2295 % |
| ind\_recibo\_ult1 | 直接付款 | 分类（0/1） | 0 % |

**表2-2 test数据表字段介绍**

|  |  |  |  |
| --- | --- | --- | --- |
|  | 变量含义 | 变量类型 | 缺失率 |
| fecha\_dato | 该表已为此列分区 | 日期 | 0% |
| ncodpers | 客户代码 | 离散 | 0% |
| ind\_empleado | 员工指数：A活跃，B前聘用，F孝顺，N不雇员，P 被动的 | 分类（字符） | 0 %  003962 |
| pais\_residencia | 客户的国家居住地 | 字符 | 0 % |
| sexo | 顾客的性别 | 分类（字符） | 0.00538 % |
| age | 年龄 | 离散 | 2.78 % |
| fecha\_alta | 客户成为银行合同的第一持有人的日期 | 日期 | 0 % |
| ind\_nuevo | 新客户指数。如果客户在过去6个月内注册，则为1。 | 分类（数字） | 0 % |
| antiguedad | 客户资历（以月计） | 离散 | 2.78 % |
| indrel | 1（第一/主要），99（本月的主要客户，但不是月末） | 分类（数字） | 0 % |
| ult\_fec\_cli\_1t | 作为主要客户的最后日期（如果他不是在月末） | 日期 | 99.8190 % |
| indrel\_1mes | 月初的客户类型，1（第一/主要客户），2（共同所有者），P（潜在），3（前主要），4（前共同所有者） | 分类（数字+字符） | 0.0025 % |
| tiprel\_1mes | 月初的客户关系类型，A（有效），I（无效），P（前客户），R（潜在） | 分类（字符） | 0.0025 % |
| indresi | 住宅指数（如果居住国家与银行国家相同，则为S（是）或N（否）） | 分类（字符） | 0 % |
| indext | 外国人指数（如果客户的出生国家与银行国家不同，则为S（是）或N（否）） | 分类（字符） | 0 % |
| conyuemp | |  | | --- | | 配偶指数。1如果客户是雇员  的配偶（'N' 'S'） | | 分类（字符） | 99.9888 % |
| canal\_entrada | 客户用来加入的渠道 | 字符 | 0.2239 % |
| indfall | 已故指数。N / S | 分类（字符） | 0 % |
| tipodom | |  | | --- | | 地址类型。1，主要地址 | | 分类（数字） | 0 % |
| cod\_prov | 省代码（客户地址） | 离散 | 0.4299 % |
| nomprov | 省名称 | 字符 | 0.4299 % |
| ind\_actividad\_cliente | 活动指数（1，活跃客户; 0，非活跃客户） | 分类（数字） | 0 % |
| renta | 家庭总收入 | 离散 | 0 % |
| segmento | 细分：01 - VIP，02 - 个人03 - 大学毕业 | 分类 | 0.2412 % |

**2.2.3 数据描述性统计**

Number of rows in train : 13647309

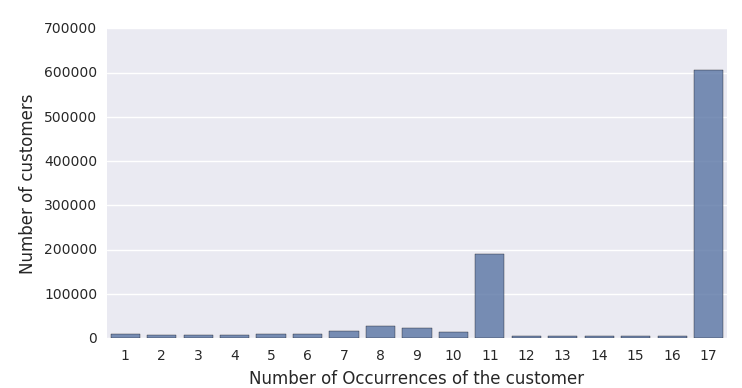
Number of rows in test : 929615

Number of customers in train : 956645

Number of customers in test : 929615

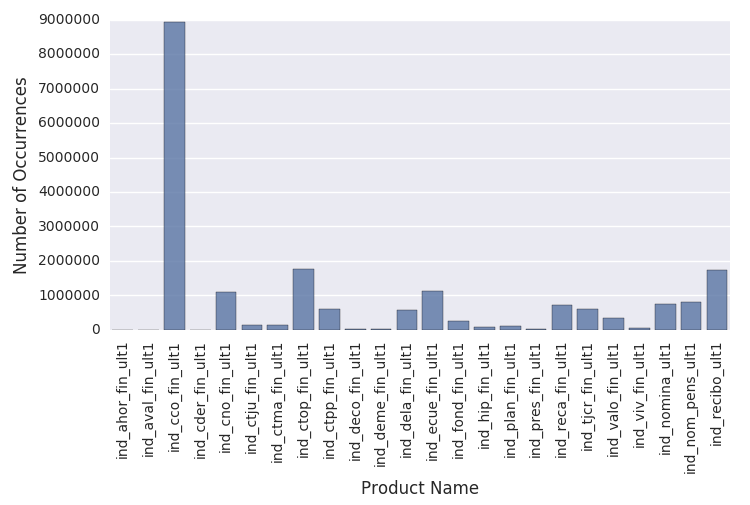
Number of common customers : 929615

**下图为客户出现次数的统计：**



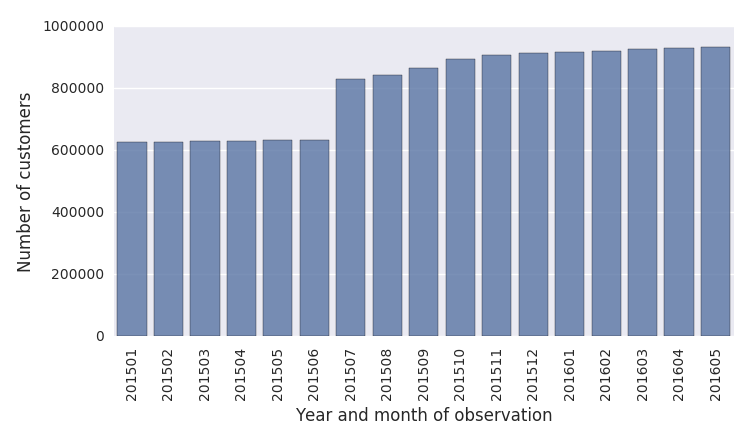
训练集有17个月的数据，由图可知，大多数客户都存在的所有17个月。与其他月份相比11个月也是一个小高峰。

**下图为给定产品在训练集中应用的次数：**



产品“ind\_cco\_fin\_ult1”是应用最多的，“ind\_aval\_fin\_ult1”是应用最少的。

**下图为对fecha\_dato（观察日期）的统计**

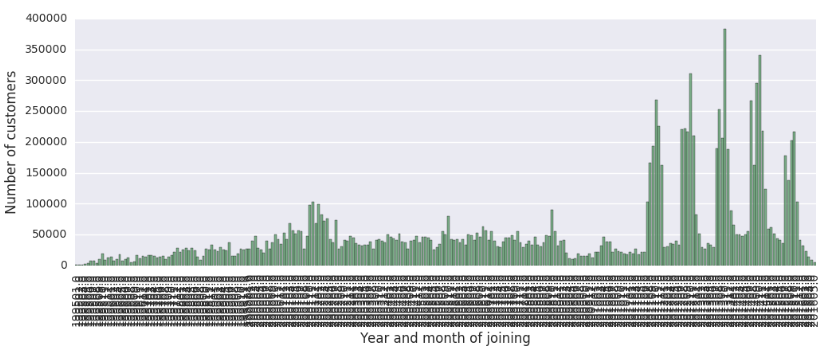


在给定的训练数据的前六个月，客户/观测的数量几乎保持不变，然后在2015年7月，客户/观测的数量突然增加。

**下图为对fecha\_alta（客户首次成为银行合同持有人的日期）的统计：**

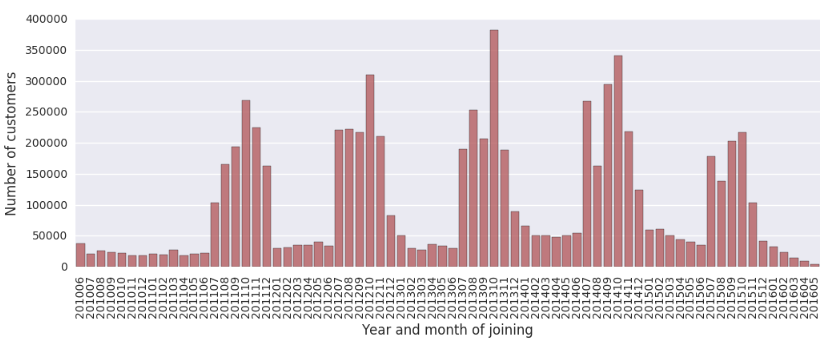
最早为 : 199501.0

最晚为 : 201605.0



第一个持股日从1995年1月开始。但是，这个数字在最近几年是很高的。此外，这些数据中似乎也有一些季节性高峰。

**进一步对fecha\_alta（客户首次成为银行合同持有人的日期）的分析:**

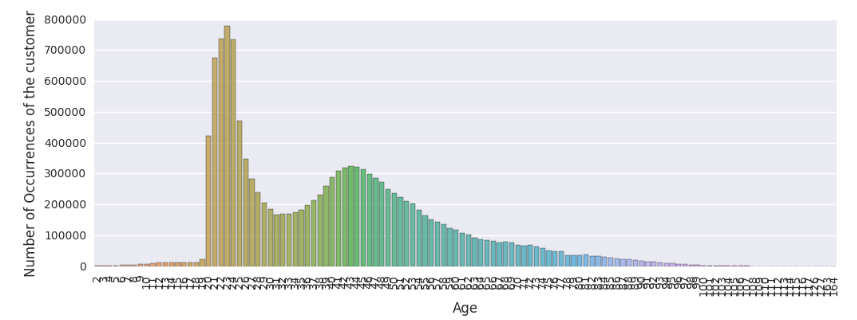


从2011年开始，首次成为合同持有人的客户数量在下半年（后6个月）远远高于上半年，而且在那之后的所有年份都是如此。从商业的角度来看，很有趣。

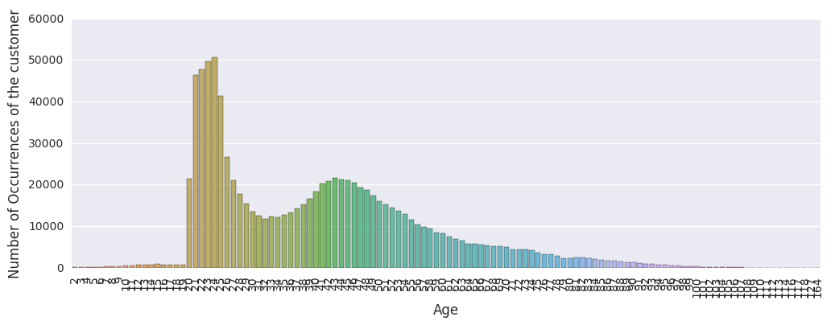
**数值变量：**

**Age：**

**Train:**



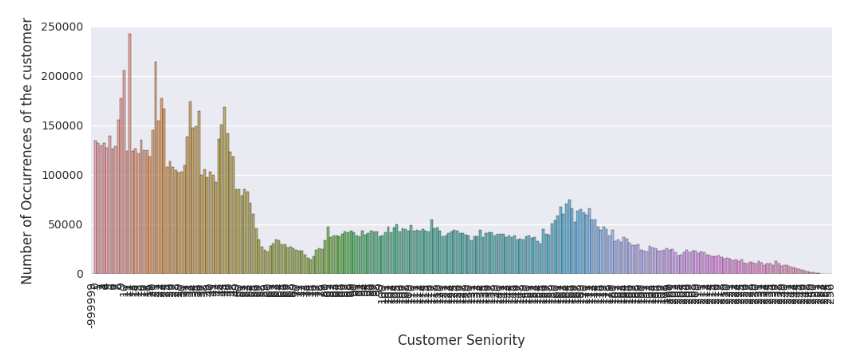
**Test:**



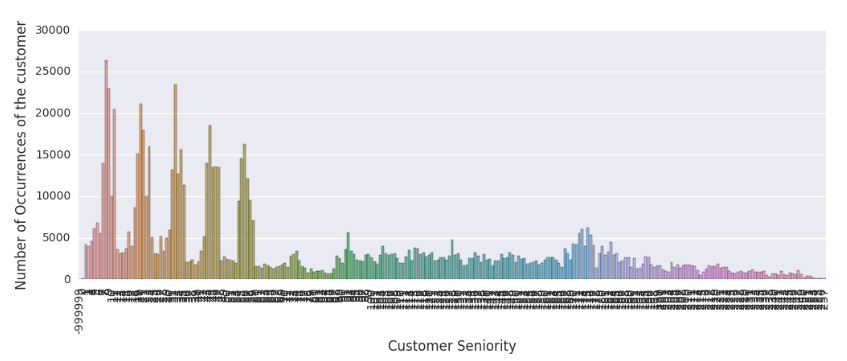
训练集与测试集年龄分布相似

****Antiguedad(客户资历-以月计):****

****Train：****

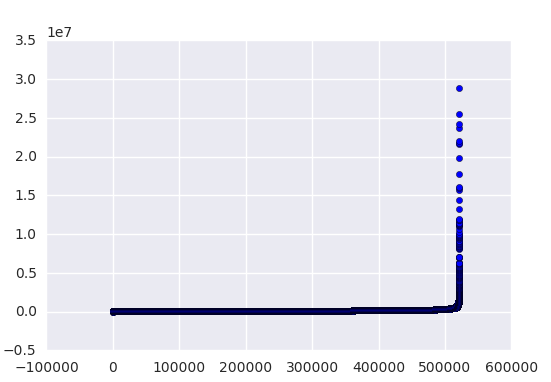


**Test：**



分布存在有规律，存在异常值 -999999。

****Renta（家庭总收入）:****

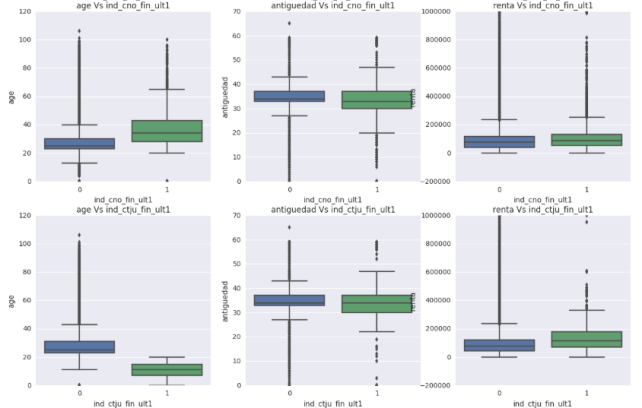
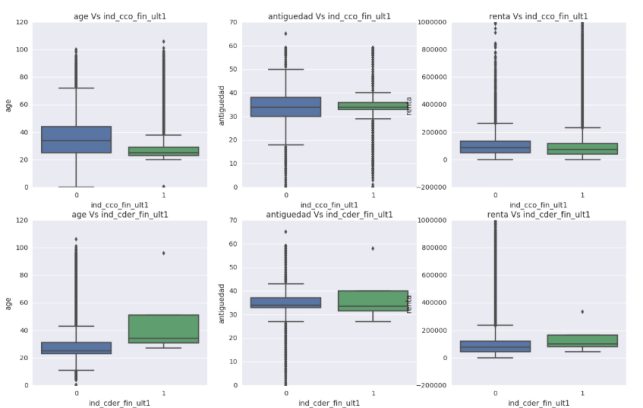


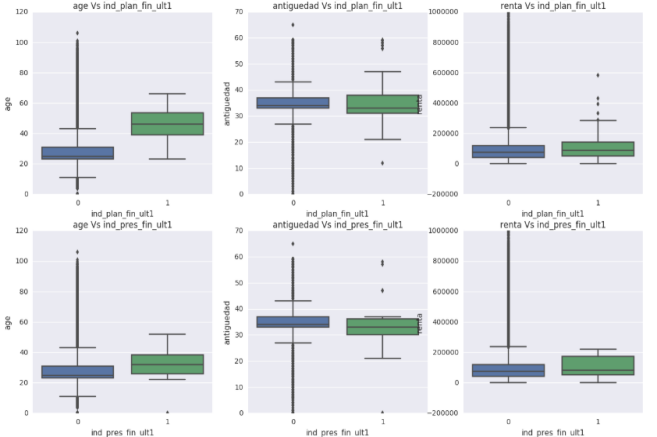
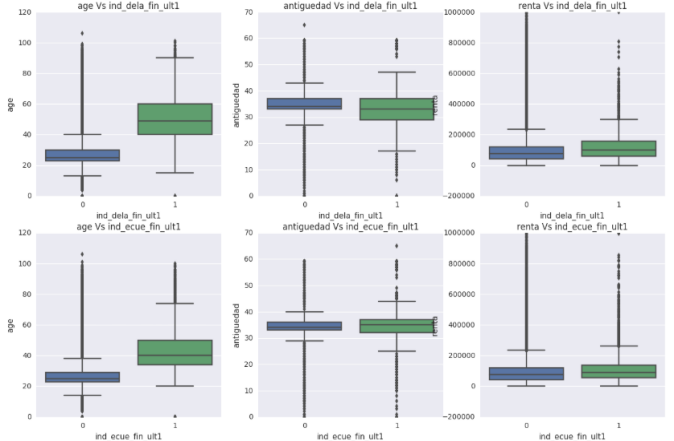
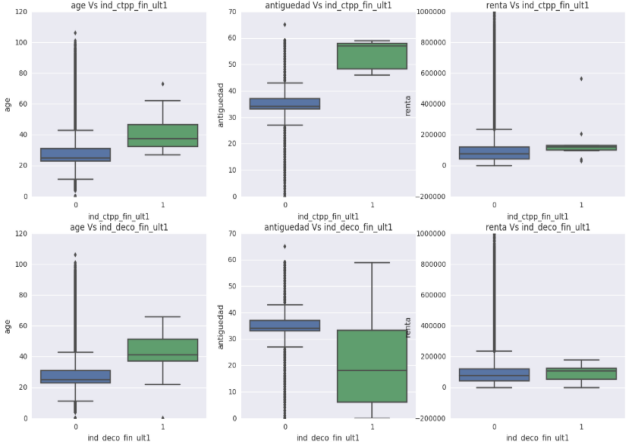
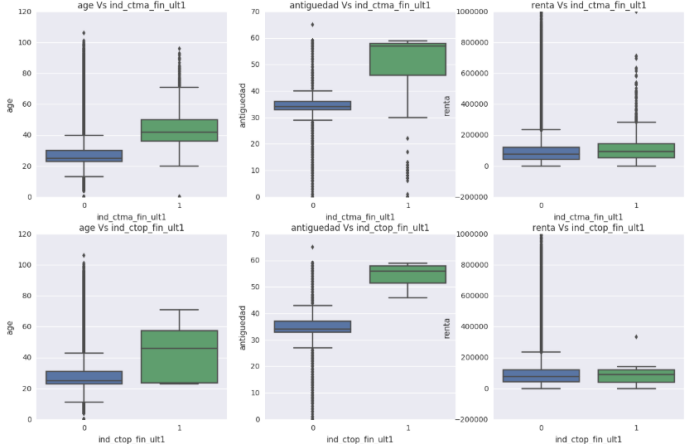
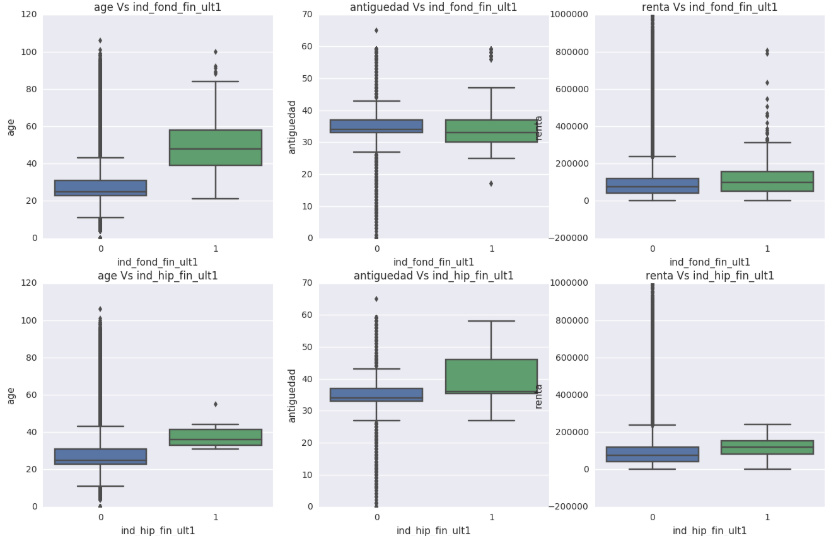
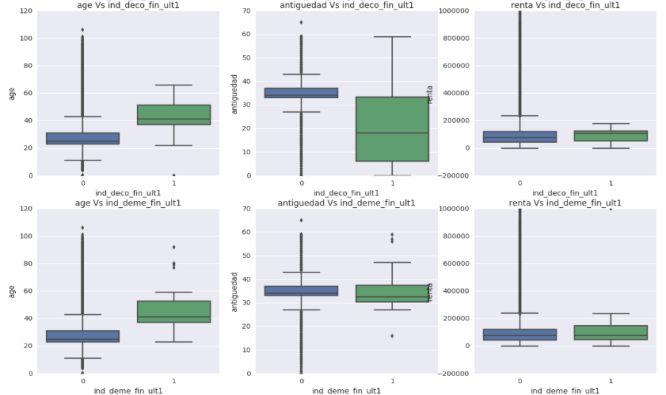
金额的分配似乎极不平衡。数据中很少有非常高价值的客户。

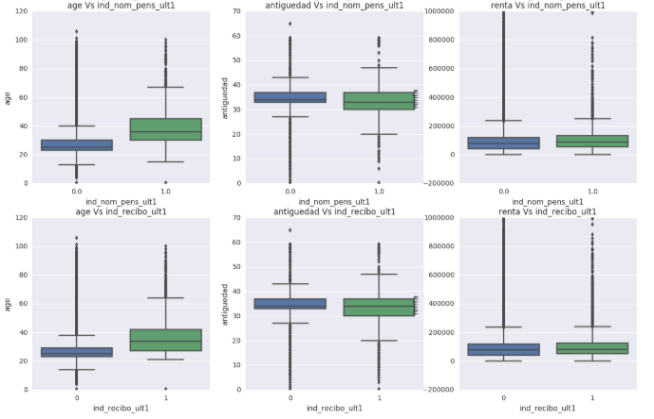
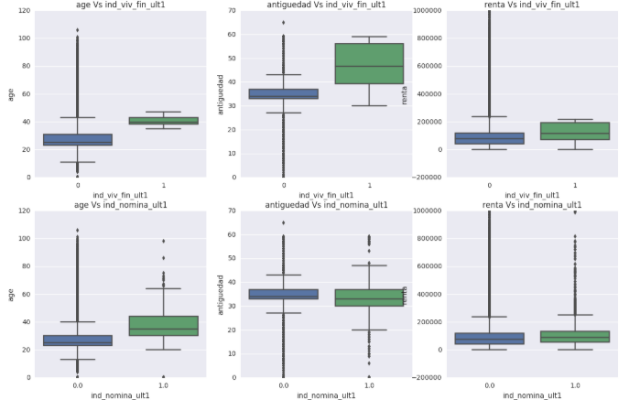
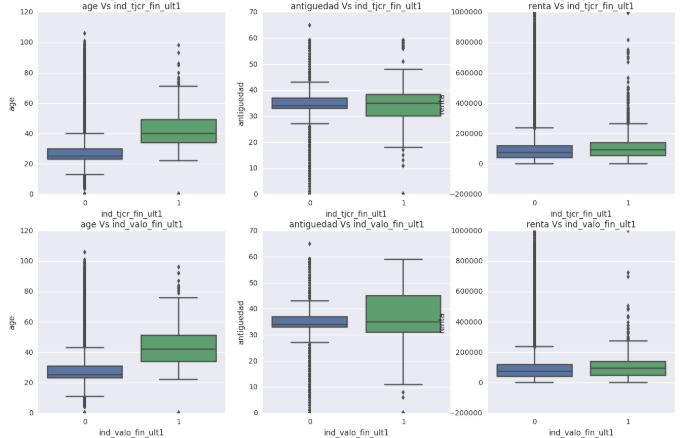
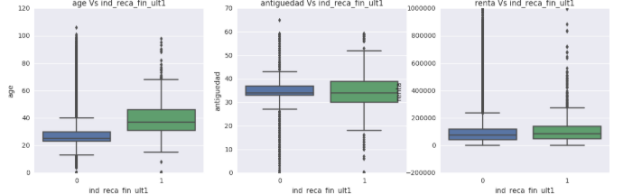
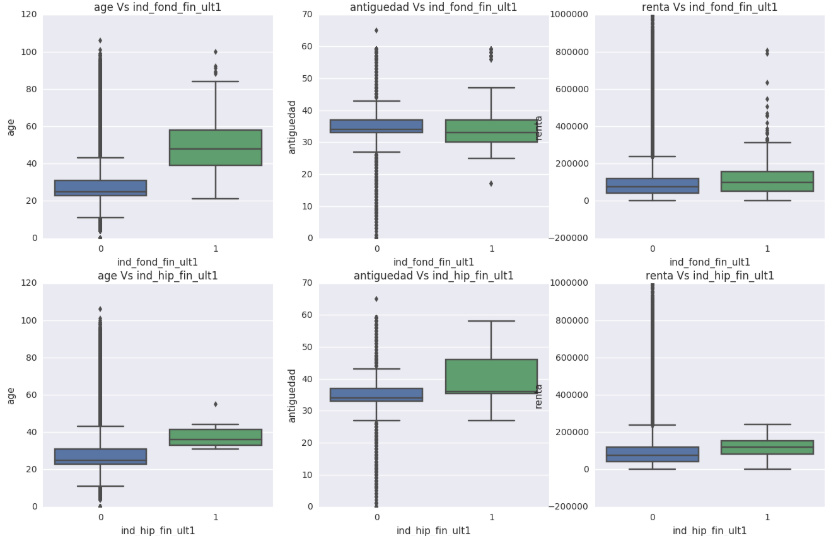
均值：134254.3182380867

中值：101850.0

**数值变量（age,a**ntiguedad,renta**）对目标值的影响：**

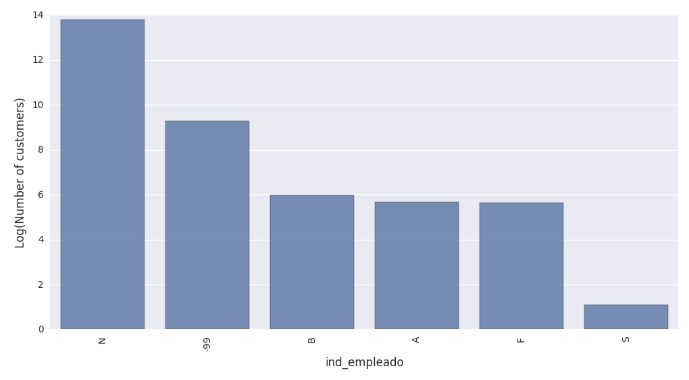




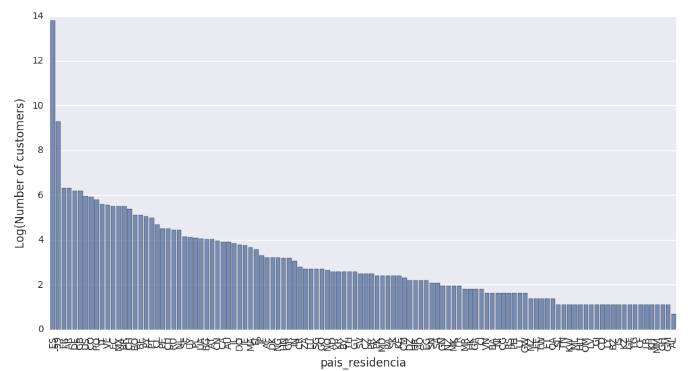


**分类变量：**

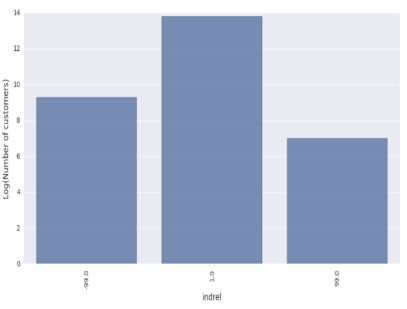
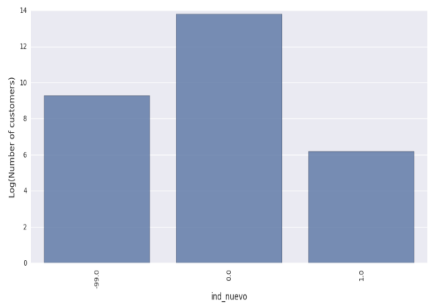
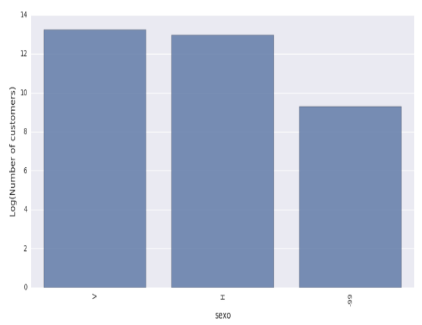
ind\_empleado：



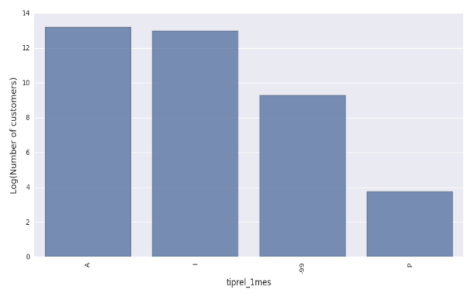
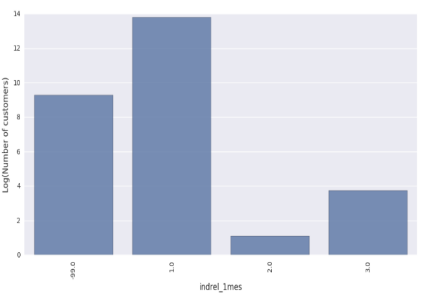
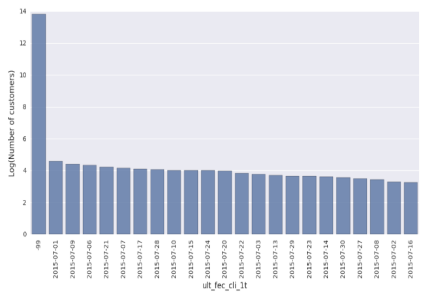
pais\_residencia : 114



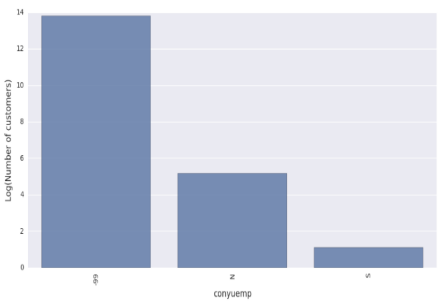
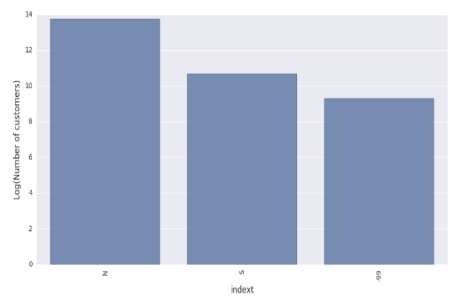
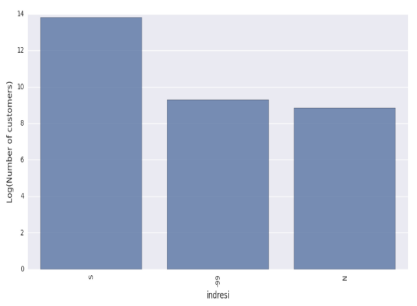
Sexo： ind\_nuevo： Indrel：



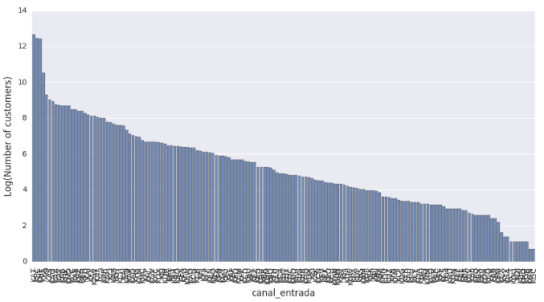
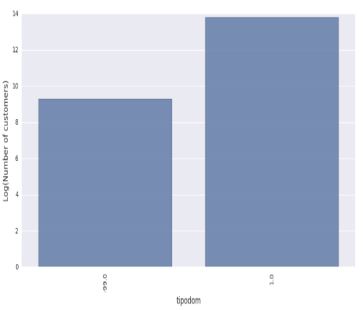
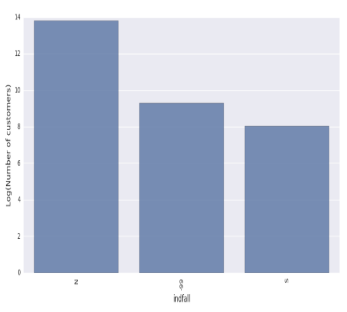
ult\_fec\_cli\_1t : 23 indrel\_1mes： tiprel\_1mes



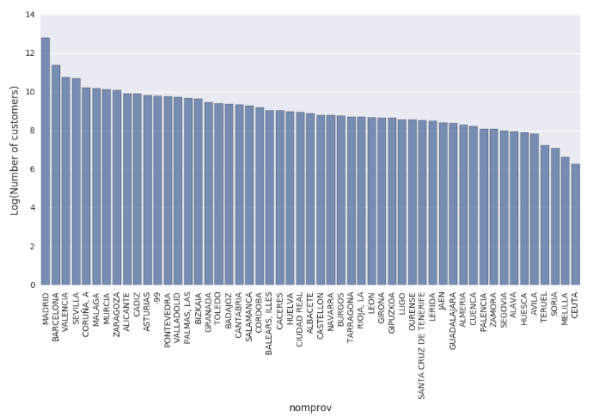
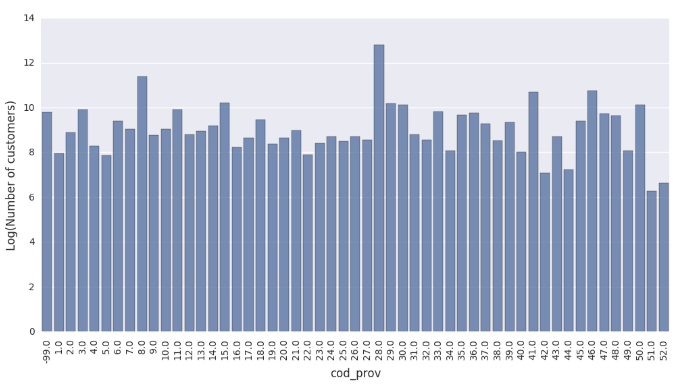
Indresi indext Conyuemp



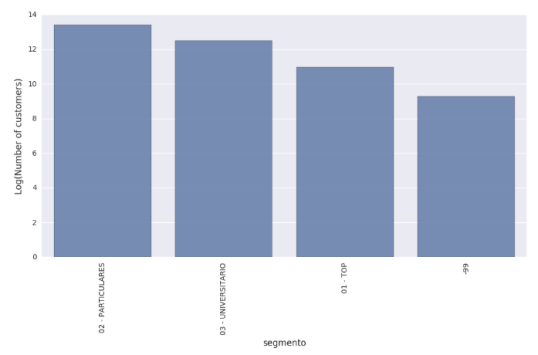
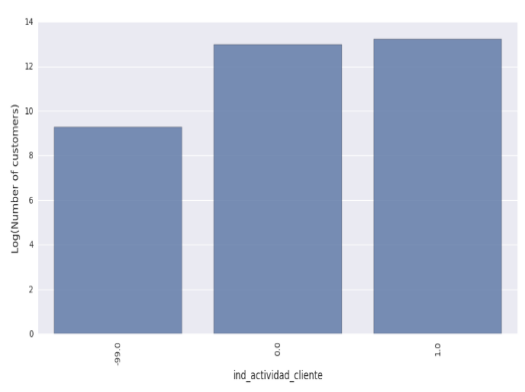
canal\_entrada : 157 indfall Tipodom

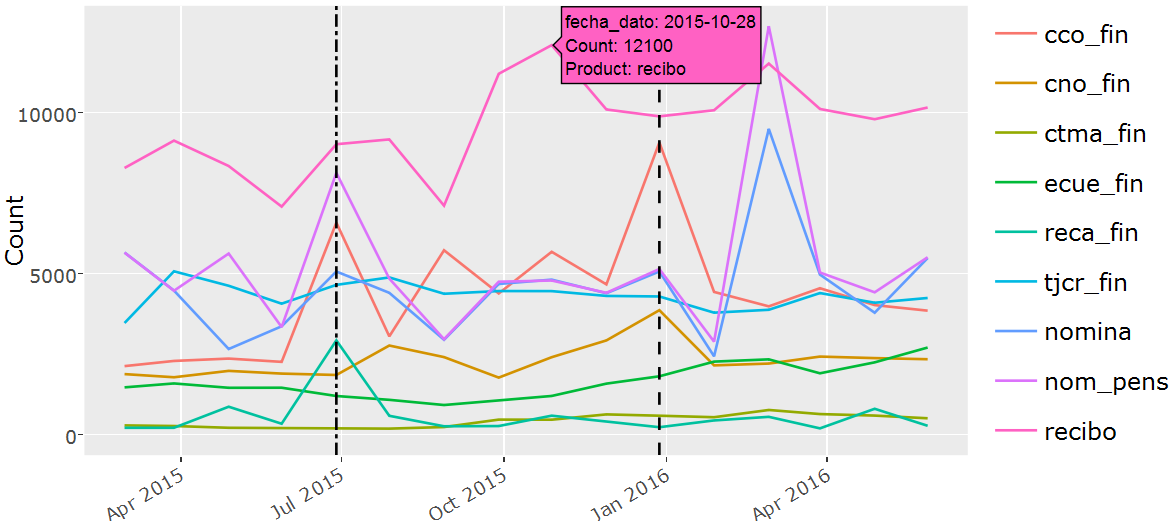
cod\_prov : 53 nomprov : 53



ind\_actividad\_cliente segmento



使用前9名产品的作为新产品使用的数量



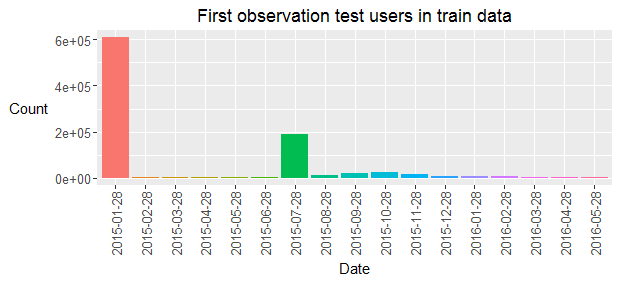
1. 优秀算法思路

代码链接：<https://github.com/ttvand/Santander-Product-Recommendation>

**3.1 方案一**

**3.1.1** 方案一数据预处理及特征工程部分方案

下一个图表显示，测试数据集中的大多数用户已经出现在了训练数据的第一个月，并且有相当大一部分的测试用户包含了2015年7月的第一个训练信息。几乎所有测试用户都包含他们首次出现在训练数据和数据结束期间（2016年5月）之间的月度数据。

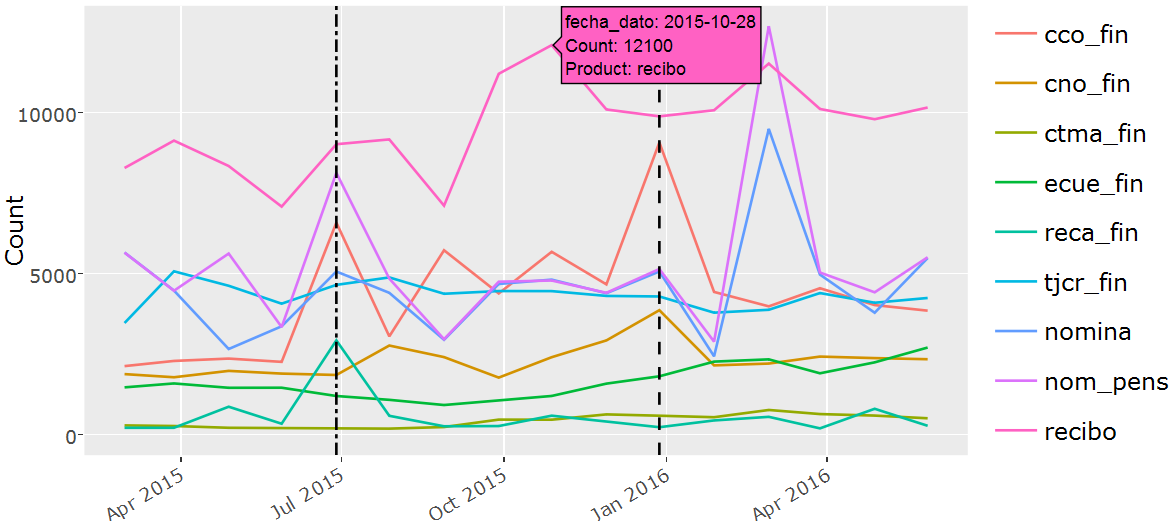


## **探索性分析**

我写了一个交互式的Shiny应用程序来研究原始数据。可以随意探索数据！这种交互式分析揭示了许多有趣的模式，并且是许多基本模型特征的主要动机。

<https://tvdwiele.shinyapps.io/Santander-Product-Recommendation/>

下一个图显示了训练月份中使用排名前9名产品的训练数据中的作为新产品使用数量。



前9名产品的新产品数量

随着时间的推移，产品的受欢迎程度不断发展，但每年的季节性因素也会影响产品新使用的数量。2015年6月（上图中的左虚线）特别有趣，因为与其他月份相比，它包含了一个完全不同的产品新使用分布（特别是Cco\_fin和Reca\_fin），可能是因为6月标志着西班牙纳税年度的结束。稍后将分析2015年6月新产品信息是2016年6月新产品的最佳指标，特别是由于税务产品（Reca\_fin）和支票账户（cco\_fin）的不同行为。在论坛上[热门的帖](https://www.kaggle.com/c/santander-product-recommendation/forums/t/25579/when-less-is-more" \t "https://ttvand.github.io/Second-place-in-the-Santander-product-Recommendation-Kaggle-competition/_blank)子建议将建模工作限制在2015年6月的新产品记录中，以预测2016年6月。

帖子内容（[BreakfastPirate](https://www.kaggle.com/breakfastpirate" \o "BreakfastPirate)第五名）：

我只对大约37,000个账户进行培训 - 仅限于2015年6月增加新产品的账户。我发现每月增加的产品分布差异很大，6月似乎是不寻常的月份。由于我们预计2016年6月，我仅在2015年6月接受培训。

我只使用了在2015年6月添加了一个新产品的账号，我们并不试图确定谁将添加新产品，我们只是试图预测如果他们添加了新产品，他们将添加哪些产品。所以排除了所有没有添加产品的帐户。

我将多个二进制因变量折叠成一个多值因变量。如果一个帐户在2015年6月添加了多个产品，那么我只需在我的训练集中多次使用该行数据 - 每添加一个产品一次。所以我的训练集中最终得到了大约46,000行。然后使用多级分类（multi：softprob）来提供每个产品被添加的概率，并且选择了得分最高的7个（并且也不是前一个月用户就拥有的产品，因为根据定义它们不是新添加的）。

交互式应用程序还表明，新产品概率与上个月拥有的产品之间存在重要关系。Nomina产品是一个极端的例子:它只在Nom\_pens是在前一个月拥有的情况下购买，或者是在同一个月与Nom\_pens一起购买的情况下购买。交互式应用程序的另一个有趣的观点与经常一起购买的产品有关。Cno\_fin经常与Nomina和Nom\_pens一起购买。其他大多数新产品的购买似乎相当独立。交互应用程序的最后一个应用程序显示了在特定月份购买新产品的用户的连续和分类用户预测因子的分布。

## **特征工程**

特征工程文件使用不同的**滞后**计算。例如，2015年6月培训的模型将根据2015年6月****之前和之后的****所有24个用户数据预测器****以及****2015年6月****之前的****产品信息进行特征培训。此方法模拟测试数据，其中还包含2016年6月的用户数据。测试特征使用最近几个月生成并基于滞后数据以便具有类似的特征解释。因此，2015年6月训练的使用5个滞后月的模型在仅根据2016年1月开始的滞后数据计算的测试特征上进行评估。

特征在几次迭代中添加。基于那些在基础模型中具有很强预测价值的特征添加了类似的特征。大多数有价值的特征存在于先前拥有的产品的滞后信息中。在月份滞后1到6和12处添加了所有产品的滞后特征，并包括自（第二个）上一个正面（新产品）和负面（下降产品）侧面以来的月数的特征。在整个滞后期间，正面和负面侧面的数量也被添加为所有产品的特征以及所有产品组合的正/负侧面数量，分别为1到6和12。一个有趣的观察结果是事实上，收入（renta）对于用户群的约30％是非独特的，其中大多数重复发生成对和大小<10的组。我认为这些代表来自同一家庭的人，并且这些信息可能会产生有价值的特征，因为同一家庭中的人可能会显示相关模式。但是，我尝试的所有与家庭相关的特征都没有什么价值。

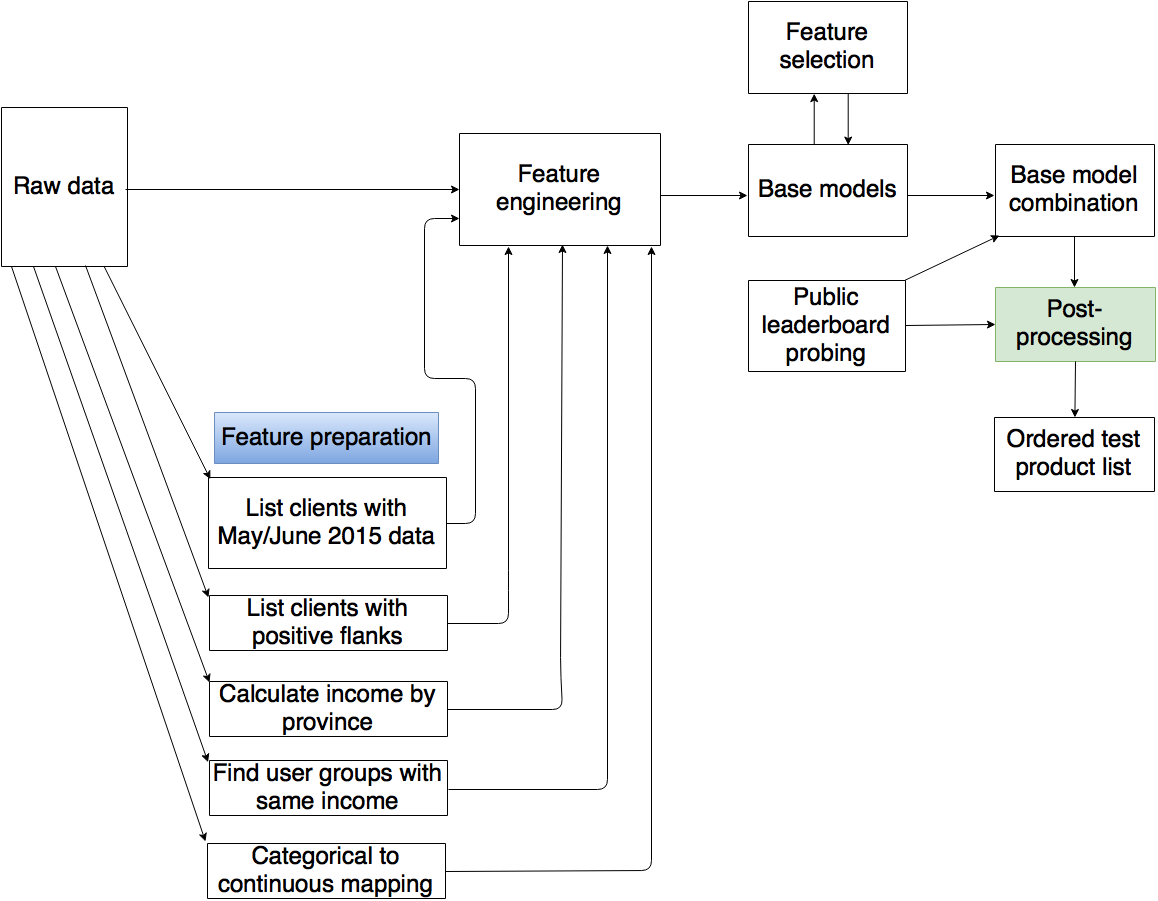
我为2015年5月和2015年6月有数据的用户添加了一个布尔标志，因为在2015年7月之后添加的用户显示了不同的购买行为。然而，这些特征几乎没有增加价值，因为基本模型已经能够使用其他特征捕获这种不同的行为。除收入特征外，原始数据始终以原始形式使用。在这里，我使用省的中位数，如果它丢失了，我还添加了一个标志，表明该值被估算。使用对序数数据的直观手动重新排序和标称数据的虚拟排序，将分类特征映射到数字特征。

添加了其他特征，以在24个用户数据预测器的滞后期内合并动态信息。然而，这些预测器中的许多是静态的并且对整体性能的附加值有限。研究收入变化对产品购买行为的影响会很好，但鉴于给定数据集中的静态收入值，这是不可能的。我没有在最重要的特征之间包含相互作用，并希望在阅读其他几个顶级竞争对手的方法之后能有这种相互作用。

**3.1.2** 方案一模型设计、建立部分方案

## **战略**

下面显示了生成单个提交的整体策略的简化。最后两个提交是多个单一提交的集合，基本模型组合和后处理逻辑的变化很小。



单一提交策略

我的方法的核心要素是基础模型。这些都是针对所有24种产品的一个月的数据进行训练。每个基本模型由一个新产品概率的xgboost模型组成，条件是用户前一个月没有该产品的使用。使用所有可用的历史信息训练基础模型。这只能通过计算2015年2月到2016年5月之间所有月份的单独特征文件来实现。2015年2月训练的模型仅使用一个滞后月，而2016年5月训练的模型使用16个滞后月。在生成特征文件之前，需要执行几个特征准备步骤。限制基本模型仅使用每个滞后产品对的顶部特征来加快建模和评估过程。在使用所有特征训练的基础模型上，结合5倍交叉验证的特征增益等级，得到特征的排序列表。对测试数据的基本模型预测进行线性组合。这些权重是利用2016年5月的公共排行榜信息和当地验证，以及对基本模型预测的相关性研究得出的。在为所有测试用户生成最有可能2016年6月的新产品的排序列表之前，对加权产品预测应用了几个后处理步骤。

## **基础模型**

基础模型是二进制xgboost模型，适用于所有24个产品和所有16个月(2015年2月- 2016年5月)的正侧翼（正侧翼 = 新应用的产品, 负侧翼 = 取消应用的产品）。我的主要见解是使用所有可用的数据。这意味着这些模型针对的是前几个月没有包含特定产品的所有用户。最初，我使用“边际”模型来计算任何正侧翼的概率，使用“条件”模型来计算给定至少一个新产品被应用的概率。由于只有3%到4%的用户在一个月内购买了至少一种产品，所以这种方法的拟合速度更快，但我发现，在使用所有数据(“联合”模型)建模时，我得到了稍微好一点的结果。

超参数是根据训练正侧翼的数量来确定的，我在训练数据中观察到的正侧翼越多，树越深。

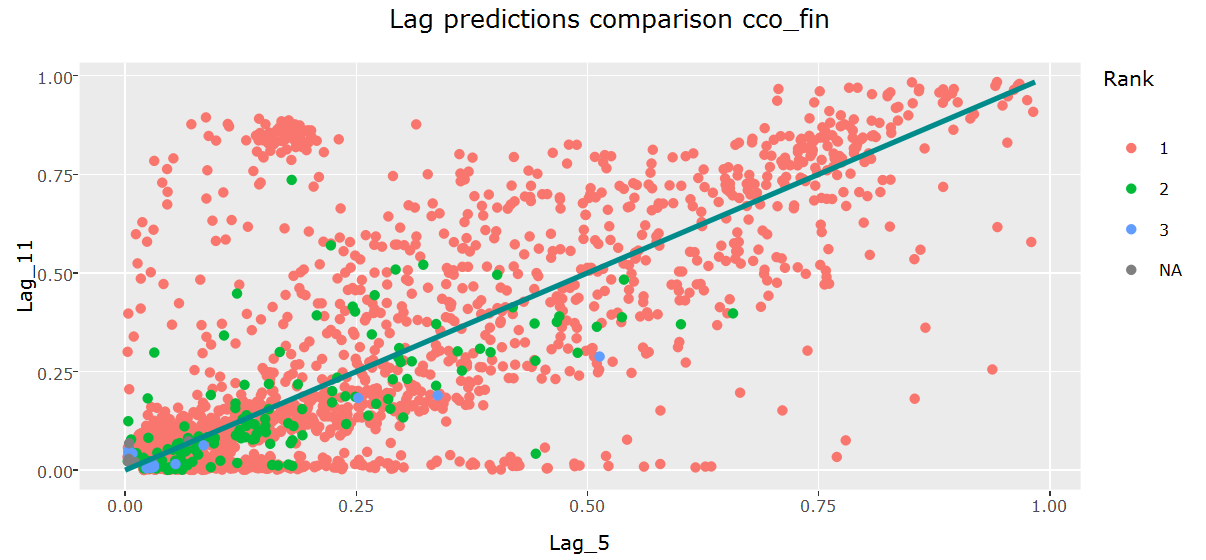
所有模型都是使用所有训练数据以及10个随机折叠建造的。

我尝试了几种叠加基本模型预测的方法，但似乎随着时间的推移，模式差异太大，叠加的值无法与基本模型预测的加权平均值相比。

我还尝试引导基本模型，但结果总是更糟。据我目前所知，没有一个顶尖的竞争对手能够在这个问题设置中获得成功。

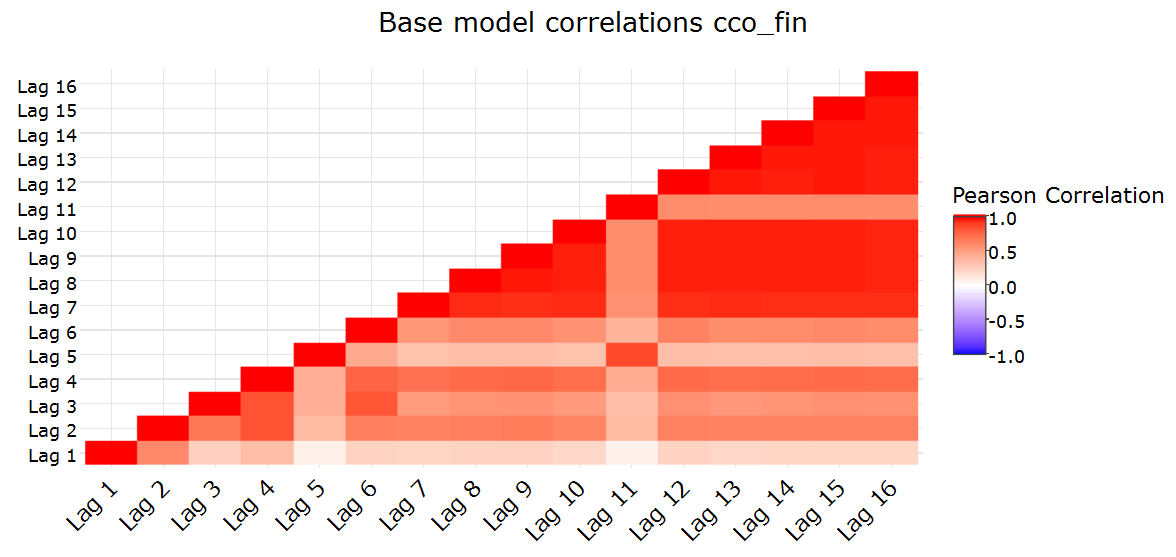
## **基础模型组合**

将前一节的基本模型与试验数据进行拟合，其中使用的试验滞后量设为训练数据中的滞后量。大多数权重是在6月15日给出的，但其他月份也都包含有价值的信息，尽管我有时将模型随时间变化的产品的权重设置为零(本节结束)。为了找到一种组合基本模型的好方法，交互式地查看数据是一个好主意。第二个交互式应用程序比较了测试集中最重要产品的基本模型预测，并显示了其他与基本模型相关的信息，如预测的置信度。下面的两张图给人展示应用程序，推荐以交互的方式查看数据。

<https://tvdwiele.shinyapps.io/Santander-Product-Recommendations-MonthModels/>

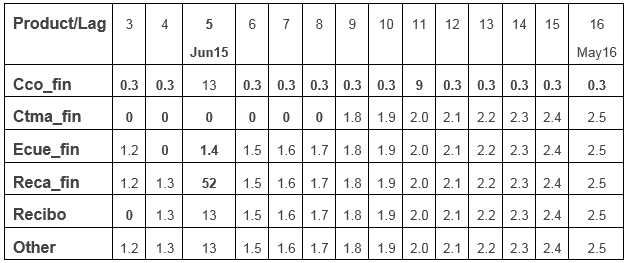
cco\_fin的基本模型相关性。2015年6月（滞后5）和2015年12月（滞后11）是

特殊月份。



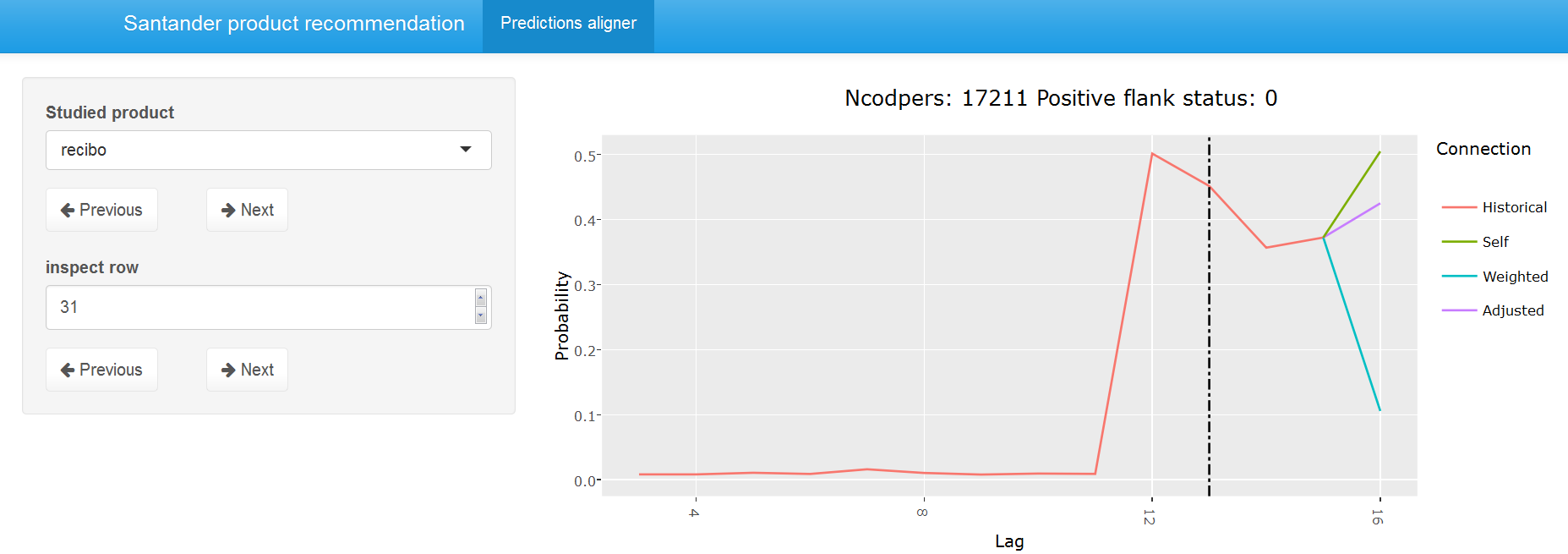
使用Lag 5（2015年6月）和Lag 11（2015年12月）对cco\_fin进行基本模型预测比较。预测的Pearson相关系数为0.86。

交交互式的基本模型预测应用程序让我想到了各种组合基本模型预测的方法。我尝试了几个加权变换，但是没有找到一个比标准的加权平均的基本模型预测更符合目标准则的变换。基于交互分析和公众排行榜反馈，对不同的产品使用不同的权重。例如，ctma\_fin的权重在2015年10月之前设置为0，因为在2015年10月之前，购买行为似乎遵循不同的规则。与其他月份相比，Cco\_fin在2015年6月和12月表现出了特别不同的行为，这几个月的典型Cco\_fin模式和年末(tax)特定模式的混合分布。下表显示了所有产品的相对滞后权重。这些都是标准化的，因此对于每个产品，它们的和为1，但更容易在下面的原始形状中解释。最近几个月通常贡献更大，因为它们可以模拟更丰富的动态，而且大部分权重都集中在2015年6月。所有产品的时滞1(2015年2月)和时滞2(2015年3月)的权重均为0。

****

按产品相对的基础模型权重

一些产品的正侧翼如nomina和nom\_pens主要依赖于之前滞后的信息，但是当有更多滞后信息时，像recibo这样的产品正侧翼会更有信心。例如，假设recibo在2015年10月被取消并在2015年11月被应用。2015年6月的模型将无法使用此信息，因为仅使用5个测试滞后月来评估测试集上的模型（2016年1月 - 2016年5月）。在这些情况下，根据在一些提交中使用更多数据的模型的概率，我调整了概率。事后看来，我希望我已将它应用于所有提交内容。下一个图使用上表中的权重与仅包含滞后子集的调整后预测来比较2016年5月的加权预测。



用户17211的基本模型recibo预测比较

## **后期处理**

****产品概率归一化****

将测试概率提升到指数，使产品概率的和与外推的公共排行榜计数相匹配，从而转换测试概率。指数变换比线性变换的优点是，它主要影响低概率。在这里，重要的是要认识到，调查的公众排行榜得分不能直接转化为积极的排行榜计数。因此，经常与nom\_pens和cno\_fin一起购买的nomina等产品，在一定程度上比它们对MAP的贡献更有可能。

****置信度****

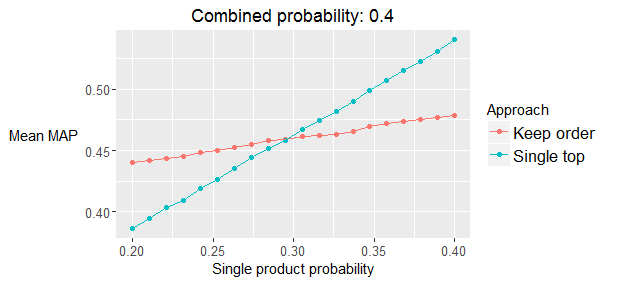
通过模拟研究，我可以证实我怀疑，为了优化预期的MAP，对于更自信的预测,应该缩小不太自信的预测。我在我的一些提交中应用了这个，这为最终的集成增加了有限但重要的价值。我将置信度计算为均值（预测给定的实际正侧翼）/均值（预测没有给出正侧翼），其中预测是在基础模型的10倍交叉验证中的折叠记录上计算的。

****Nomina Nom\_pens重新订购****

Nomina如果没有nom\_pens则永远不会购买，如果nom\_pens不是在上个月拥有的话。如果nomina的排名高于nom\_pens，并且两者都不属于上个月，我只是交换了他们预测的概率。

****MAP优化****

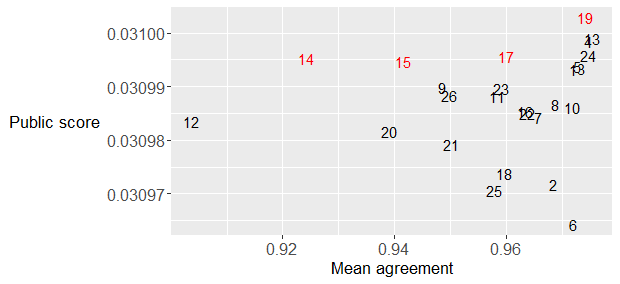
想象一下这种情况：cco\_fin的正侧翼概率为0.3，而nomina和nom\_pens的概率均为0.4，但它们总是具有相同的值。假设所有其他产品概率为零。下图显示了预期MAP的模拟研究，其中组合（nomina和nom\_pens）概率设置为0.4，单个（cco\_fin）概率在0.2和0.4之间变化。该图与[数学推导一致](https://www.kaggle.com/c/santander-product-recommendation/forums/t/26824/2nd-place-solution?forumMessageId=151987" \l "post151987" \t "https://ttvand.github.io/Second-place-in-the-Santander-product-Recommendation-Kaggle-competition/_blank)得出结论，如果cco\_fin的概率高于0.294，则应将cco\_fin列为最大化预期MAP的最佳产品。当相对概率差异有限时，我还关闭了nomina和nom\_pens之间的“差距” 。这种MAP优化在局部验证中具有很大的效果（~0.2％提升）但在公共排行榜上的价值有限。事实证明，对私人排行榜的影响具有相似的正值，但我过度拟合，导致我错误地认为MAP优化总体上价值有限。



MAP @ 7的平均值，使用两种可能排序的模拟研究，组合概率设置为0.4。单个概率在0.2和0.4之间变化。每个数据点使用200,000个独立模拟。

## **Ensembling**

我提交了两个合奏：一个使用我最近的26个提交，其中加权概率是根据与其他提交和公共排行榜得分的相关性计算的。第二个集合包括手动选择这26个提交中的4个，这些提交再次被选中并使用它们与其他提交和公共排行榜反馈的相关性进行加权。提交的详细信息可在[GitHub上获得](https://github.com/ttvand/Santander-Product-Recommendation/blob/master/Ensembling/Submission Inventory.xlsx" \t "https://ttvand.github.io/Second-place-in-the-Santander-product-Recommendation-Kaggle-competition/_blank)。下一个图显示了所有最终26个提交的公共排行榜得分与平均相互等级相关性。我使用此图表和提交的等级相关性选择了手动子集，以使提交尽可能不相关，同时仍获得良好的公共排行榜反馈。



公共排行榜得分与最终26份提交的平均相互排名相关性。在两份最终提交的第一份中使用了红色的四份提交材料。第二次最终提交使用了所有26份提交。

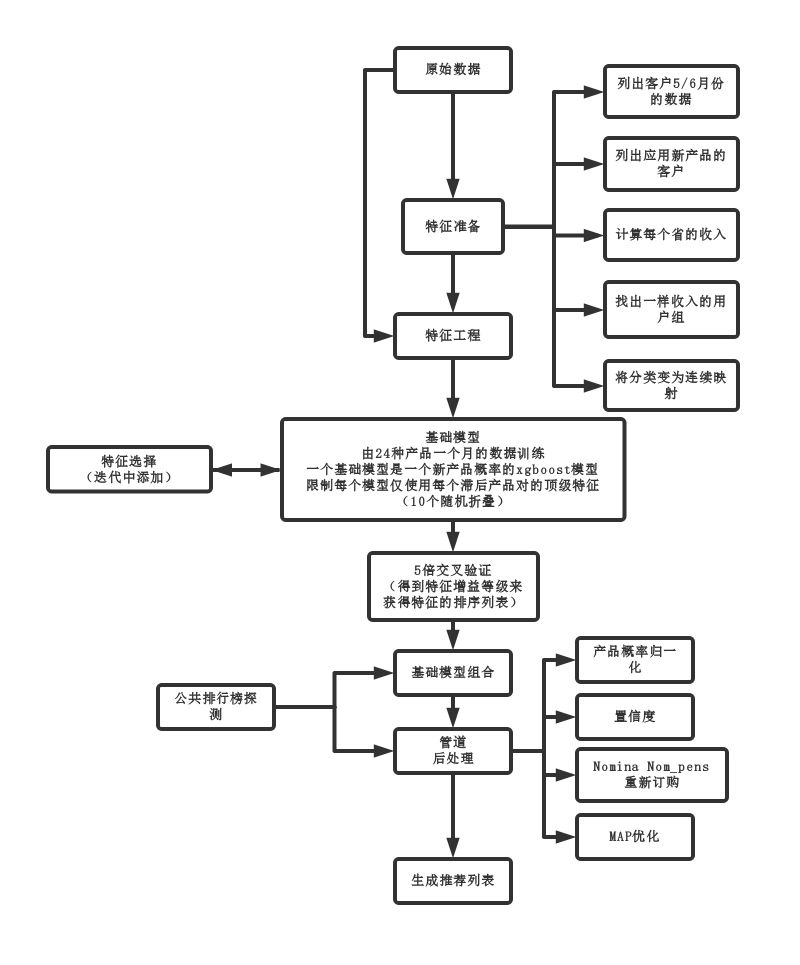
我只是在比赛的最后一天进行了整合，并且仅在截止日期之后才发现排名平均通常比概率平均效果更好。主要原因可能是我在最终提交内容中包含的变体的性质，因为大多数变体都在后处理步骤中。

**3.1.3** 方案一结果、排名等

结果：0.03131

排名：第二名

**3.1.4** 方案一算法流程图

**3.2 方案二**

这个解决方案是12个神经网络和8个GBM的集合，具有数百个特征。

**3.2.1** 方案二数据预处理及特征工程部分方案

**特征**

在大多数情况下，这些与人们在论坛上提到的相似：产品滞后，产品存在的时间，产品的平均值，自上次购买产品以来的时间等。

我在其他地方没有提到过的一些特征是时间的变化和一些非产品属性的滞后：segmento，ind *actividad*cliente，cod *prov，canal* entrada，indrel *1mes，tiprel* 1mes。

这些特征似乎伤害了个别的子模型，但无论如何都进入了整体，是产品的历史平均分段：*canalentrada，segmento，codprov。*

**3.2.2** 方案二模型设计、建立部分方案

**GBM模型**

这些子模型彼此相似，但与论坛上的流行方法略有不同。它们是17类多项式模型，针对16种最受欢迎的产品添加。剩余的第17类表示没有添加或添加剩余的8个产品之一。

如其他地方所述，通过添加具有不同目标的重复行来处理多个产品添加。这些重复的行由添加的产品数量的倒数加权。

添加第17个“无添加”类会大大扩展训练数据量。最大的子模型涵盖了6月15日至5月16日，约为1000万行。LightGBM是快速构建这些模型的关键。

子模型之间的差异主要是由于在不同的时间框架上进行训练和使用不同的特征组合。

**NN模型**

这些子模型不是针对产品的添加目标，而是针对给定月份中产品的存在。他们对产品是客户新使用的还是客户一直使用的感到矛盾。

这些模型是多目标而不是多项的。它们针对更受欢迎的16个产品和培训所有的客户，无论客户是否添加了产品。

所有网的结构都是一样的。它们有一个输入层，两个包含512个节点的隐藏层和一个16个节点输出层。这里最大的训练集也是大约1000万行，但Keras使这些网易于设置并且构建起来相对较快。

这些模型之间的差异再次是时间框架和特征以及具有不同种子的多次运行。

**后期处理**

在适用的情况下，每个子模型得分一次为6月16日，一次为6月15日，一次为12月15日。所谓“score as”，是通过将fechadato转换为数值(1到18)，将其合并到模型中。在测试集中修改fechadato会导致我们将“score as”设置为另一个月。

一般来说，每个子模型的Jun-16分数都被保留，除了reca分数被Jun-15分数代替，cco分数被Dec-15分数代替。

当客户在前一个月拥有一个产品时，子模型得分被设置为零。然后，将单个产品的得分与公共排行榜的水平挂钩。这是通过乘以平均产品得分与仅提交该产品在公共排行榜上获得的价值之比来实现的。

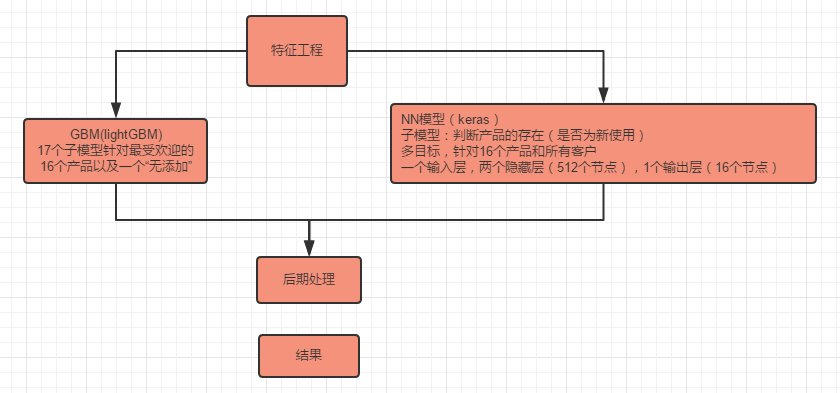
最终的集成是子模型的加权平均，权重来自于排行榜反馈。

**3.2.3** 方案二结果、排名等

结果：0.03140

排名：第一名

**3.2.4** 方案二算法流程图



**3.2 方案三**

**3.3.1** 方案三数据预处理及特征工程部分方案

代码：



**特征**

使用的特征数量为142.详情如下。

1. 除了'fecha dato *'，'ncodpers'，'fecha* alta'，'ult *fec* cli *1t'，'tipodom'和'cod* prov' 之外的原始特征（18个特征）
2. 'ind *actividad* cliente'和前一个月的值（1个特征）的串联
3. 'tiprel\_1mes'和前一个月的值（1个特征）的串联
4. 20件商品前一个月的值（20个特点）
5. 将上述20个特征串联为一个字符（1个特征）
6. 上个月购买的产品数量（1个功能）
7. 索引更改模式的计数（0到0,0到1,1到0和1到1）直到前一个月，20个产品（80个功能）
8. 连续0索引的长度直到前一个月，20个产品（20个特征）

在每个建模过程中，字符（因子）变量被目标均值替换。

**3.3.2** 方案三模型设计、建立部分方案

**模型**

1. 通过xgboost模型计算新购买每种产品的概率（由Pr表示）（除了ahor *fin，aval* fin，deco *fin和deme* fin 之外的20种产品）。
2. cco\_fin的Pr仅在2015-12-28数据中预测。
3. reca\_fin的Pr仅在2015-06-28数据预测。
4. 其他18种产品的Pr分别在2016-05-28,2016-04-28,2016-03-28,2016-02-28,2016-01-28和2015-12-28预测。
5. 其他18种产品的Pr也通过2015-12-28至2016-04-28的数据预测，仅使用具有新购买产品的记录。
6. ahor *fin，aval* fin，deco *fin和deme* fin的Pr 固定为1 ^ -10。

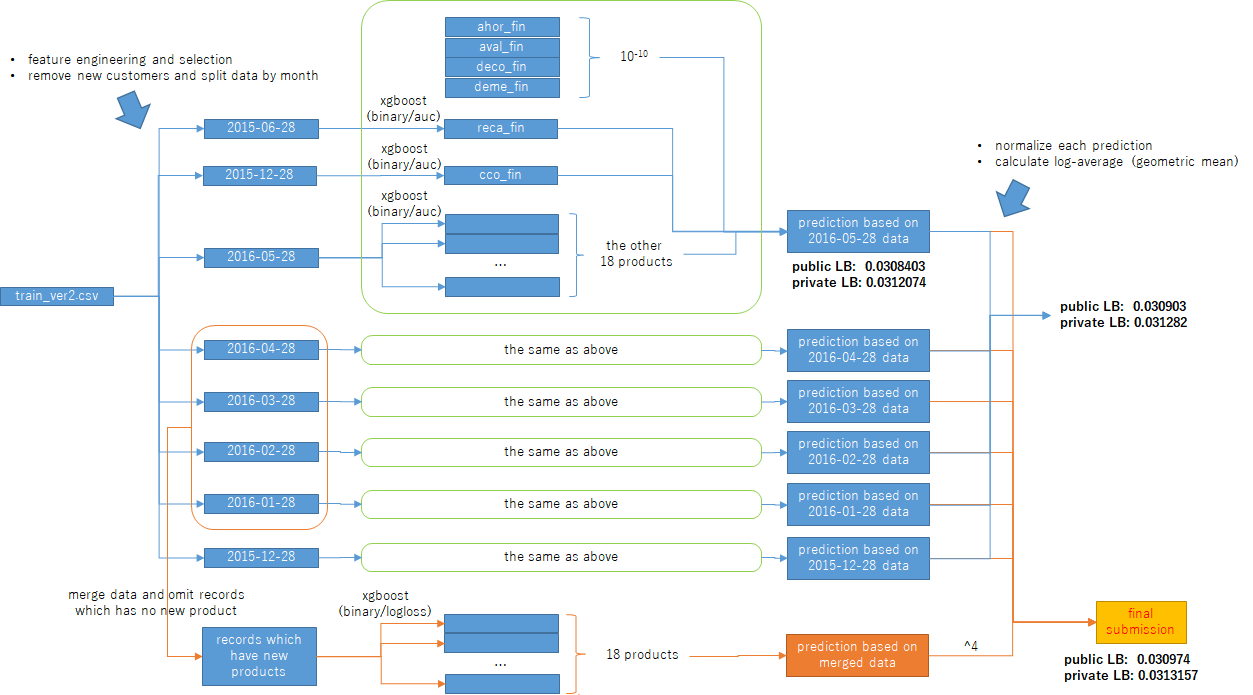
我基本上使用'2016-05-28'数据作为验证数据集，但当我使用'2016-05-28'数据作为训练数据集时，我使用'2016-04-28'数据作为验证数据集。

**3.3.3** 方案三结果、排名等

结果：0.03131

排名：第三名

**3.3.4** 方案三算法流程图



4. 算法比较

描述几种算法的基本情况及效果对比，以表格形式：

例如： 特征工程和建模部分可以描述建模用到的基础算法名称（PCA、LSTM、时间衍生等）基本库指用到的基本库（sklearn等）

**表4-1 算法比较**

|  | **评估指标**  **（**平均精度均值@7**）** | **特征工程** | **基础算法** | **基本库** |
| --- | --- | --- | --- | --- |
| **算法1** | 0.03131 | 时序特征：滞后 | xgboost | Sklearn  Pandas  XGBoost |
| **算法2** | 0.03140 | 产品滞后，产品存在的时间，  产品的平均值，自上次购买产品以来的时间，时间的变化和一些非产品属性的滞后 | GBM(lightGBM)  NN模型 | Sklearn  Pandas |
| **算法3** | 0.03131 | 18个原始数据，串联，  替换字符变量 | xgboost | Sklearn  Pandas  XGBoost |

**想法描述:**

此竞赛目的是为客户生成下月的推荐产品。

**建模思路相同**

通过每个客户的24个特征对24种产品进行推荐排序。

所有方案都是为每个产品建立子模型，通过每个产品选出的特征进行建模。

**子模型种类不同**

方案一和三都是建立的xgboost模型；

方案二运用GBM与NN模型。

5. 总结与展望

**5.1 总结**

**特征工程方面：**

主要是基于时间序列分析，主要目的是基于历史数据来预测未来信息。

对于时间序列，关注的是长期的变动趋势、周期性的变动（如季节性变动）以及不规则的变动。对于时间序列信息，当前时间点之前的信息通常很重要，例如滞后特征（也称为lag特征）使用非常广泛。滞后特征是时间序列预测问题转化为监督学习问题的一种经典方法。

本次多数参赛者都用到了滞后特征。

**建模方面：**

对每个产品建模，每个产品的模型由其对应的不同的特征集决定。

将子模型组合，计算出客户对每个产品新使用的概率，排序，生成推荐表。

**5.2 建模思路**

排除使用率极低的产品

对剩下的产品分别进行建模;

分析每个特征对产品新使用的影响

选择每个产品建模要用的特征

建立产品子模型xgboost

将子模型组合，为每个产品算出客户新使用的概率。

选择排名前面的产品推荐给客户