数据科学导论 结题报告

队长：段逸凡 PB16110818

队员：聂雷海 PB16080377

队员：杨　逍 PB16110256

队员：邹卫其 PB16061470

一、团队介绍

2018 BDCI 大赛

队名：来啊快活啊

- ID：61509
- 队长：段逸凡
- 其他成员：聂雷海 杨逍 邹卫其

队伍分工

- 段逸凡 & 杨逍：
  - 数据预处理
  - 编写代码

    - 背景研究
    - 特征提取
- 聂雷海 & 邹卫其：
    - 调研开源经验
    - 模型分析
    - 论文研究
    - 特征提取

二、项目背景

赛题简介

> 本次赛题核心内容是解决电信套餐的的个性化推荐问题。
>
> 随着信息产业的蓬勃发展，用户的套餐需求越来越多元化，电信产业作为国家基础产业之一，覆盖广、用户多，在遇到更多的发展机遇的同时，也面临着巨大挑战。只有满足用户的差异化需求，提供更好、更合适的服务，提升自身竞争力，才能在众多运营商中留住用户。
>
> 而其中一种重要的手段就是套餐的个性化推荐。面对种类繁多的套餐，如何选择最合适的一款对于运营商和用户来说都至关重要。合适的套餐推荐一方面可以提升用户的消费体验，另一方面可以带动用户需求，提升用户价值，从而实现利益最大化。
>
> 针对电信套餐的个性化推荐问题，通过数据挖掘技术来构建基于用户消费行为的电信套餐个性化推荐模型，能够在信息过载的环境中帮助用户发现合适套餐，也能将合适套餐信息推送给用户。

算法需求

> 在给定的训练数据的基础上，为用户群体画像、分类，建立匹配模型，依据测试集中的用户信息为用户匹配最合适的套餐

数据说明

字段	中文名	数据类型	说明
USERID	用户ID	VARCHAR2(50)	用户编码，标识用户的唯一字段
current_type	套餐	VARCHAR2(500)	/
service_type	套餐类型	VARCHAR2(10)	0：23G融合，1：2I2C，2：2G，3：3G，4：4G
is_mix_service	是否固移融合套餐	VARCHAR2(10)	1.是 0.否
online_time	在网时长	VARCHAR2(50)	/
1_total_fee	当月总出账金额_月	NUMBER	单位：元
2_total_fee	当月前1月总出账金额_月	NUMBER	单位：元
3_total_fee	当月前2月总出账金额_月	NUMBER 单位：元	
4_total_fee	当月前3月总出账金额_月	NUMBER	单位：元
month_traffic	当月累计-流量	NUMBER	单位：MB
many_over_bill	连续超套	VARCHAR2(500)	1-是，0-否
contract_type	合约类型	VARCHAR2(500)	ZBG_DIM.DIM_CBSS_ACTIVITY_TYPE
contract_time	合约时长	VARCHAR2(500)	/
is_promise_low_consume	是否承诺低消用户	VARCHAR2(500)	1.是 0.否
net_service	网络口径用户	VARCHAR2(500)	20AAAAAA-2G
pay_times	交费次数	NUMBER	单位：次
pay_num	交费金额	NUMBER	单位：元
last_month_traffic	上月结转流量	NUMBER	单位：MB
local_trafffic_month	月累计-本地数据流量	NUMBER	单位：MB
local_caller_time	本地语音主叫通话时长	NUMBER	单位：分钟
service1_caller_time	套外主叫通话时长	NUMBER	单位：分钟
service2_caller_time	Service2_caller_time	NUMBER	单位：分钟
gender	性别	varchar2(100)	01.男 02女
age	年龄	varchar2(100)	/
complaint_level	投诉重要性	VARCHAR2(1000)	1：普通，2：重要，3：重大
former_complaint_num	交费金历史投诉总量	NUMBER	单位：次
former_complaint_fee	历史执行补救费用交费金额	NUMBER	单位：分

三、初赛方案

模型分析

模型需求

在分析数据模型之前，我们应当结合任务背景挖掘任务需求，然后根据需求才能提炼更好的数据模型，本次试题核心需求可以概括成解决两个问题：

- 信息过载问题
  - 能够在信息过载的环境中帮助用户发现合适套餐，各种套餐满足了用户有明确目的时的主动查找需求。
- 用户无目的搜索问题
  - 能将合适套餐信息推送给用户，个性化推荐能够在用户没有明确目的的时候帮助他们发现感兴趣的新内容。

而任务需求具体来讲就是：利用已有的用户属性(如个人基本信息、用户画像信息等)、终端属性(如终端品牌等)、业务属性、消费习惯及偏好匹配用户最合适的套餐，对用户进行推送，完成后续个性化服务。

因此，建立模型的需求是：

- 多场景
  - 用户个性化需求种类繁多
- 多选择
  - 与用户需求对应的套餐种类可选性多
- 多对多匹配
  - 需要给出多对多的预测和匹配

评分方式

采用宏平均 F1-score 进行评价。

1. 针对每个用户套餐类别，分别统计 TP（预测答案正确），FP（错将其他类预测为本类），FN（本类标签预测为其他类标）。
2. 通过第一步的统计值计算每个类别下的 precision 和 recall，计算公式如下：

$$precision_k = \frac{TP}{TP+FP}$$
$$recall_k = \frac{TP}{TP+FN}$$

3. 通过第二步计算结果计算每个类别下的 F1-score，计算方式如下：

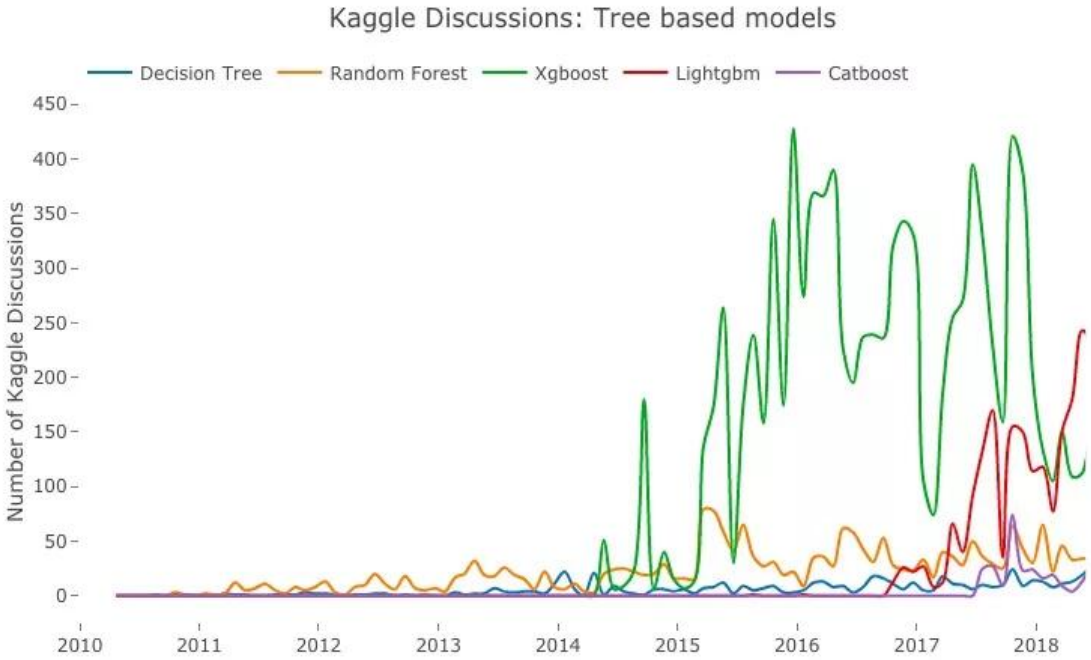
$$f1_k = \frac{2 * precision_k * recall_k}{precision_k + recall_k}$$

4. 通过第三步求得的各个类别下的 F1-score 求均值，得到最后的评测结果，计算方式如下：

$$score = (\frac{1}{n}\sum_{k=0}^{n} f1_k)^2$$

XGB 模型

GBDT 是一种常用的 Ensemble Learning 算法，全称为 Gradient Boosting Decision Tree。XGBoost 算法基于传统的 GBDT 算法，近年来在各种大型数据科学比赛中占据重要地位，经过小组的调研与讨论之后我们也决定用 XGBoost 算法，我们认真研究了 XGB 的论文以及应用，这里简述一下其原理。



kaggle 历年各大算法模型活跃度

首先从监督学习的角度来看 Boosting Tree 的模型与参数（Model and Parameters）和目标函数（Objective）：

- *Model*：$\hat{y}_i(x) = \sum_{k=1}^{K} f_k(x_i)$，$f_k \in \mathcal{F}$，其中，$\mathcal{F}$ 是包含所有回归树的函数的空间；
- *Parameters*：$\Theta = \{f_1, f_2, \ldots, f_K\}$，其中训练的参数是各个函数，这些函数刻画了树的结构；
- *Objective*：$Obj(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$，其中 Ω 的定义可以是树的节点数、深度，L2 正则项，等等；

基于上述模型、参数和损失函数，接下来面临的问题就是如何训练，论文中提到了 ***Additive Training*** 的方法：

- $\hat{y_i}^{(0)} = 0$
- $\hat{y_i}^{(1)} = f_1(x_i) = \hat{y_i}^{(0)} + f_1(x_i)$
- $\hat{y_i}^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y_i}^{(1)} + f_2(x_i)$
- $\ldots$
- $\hat{y_i}^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y_i}^{(t-1)} + f_t(x_i)$

其中，$\hat{y_i}^{(t)}$ 是第 t 轮训练的模型，$\hat{y_i}^{(t-1)}$ 保留了前面 $t-1$ 轮训练的结果，$f_t(x_i)$ 是新的函数。而我们需要做的就是最优化目标函数也即在每一轮最优化 $f_t(x_i)$，即找到 $f_t(x_i)$ 最小化目标函数。

整合式子之后，可得目标函数值为：

$$
\begin{aligned}
Obj^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} l\left(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t) + const \\
&= \sum_{i=1}^{n} \left(y_i - (\hat{y_i}^{(t-1)} + f_t(x_i))\right)^2 + \Omega(f_t) + const \\
&= \sum_{i=1}^{n} \left[2(\hat{y_i}^{(t-1)} - y_i) f_t(x_i) + (f_t(x_i))^2\right] + \Omega(f_t) + const
\end{aligned}
$$

> 其中，$2(\hat{y_i}^{(t-1)} - y_i)f_t(x_i)$ 常被称为残差（前几轮训练留下的）。

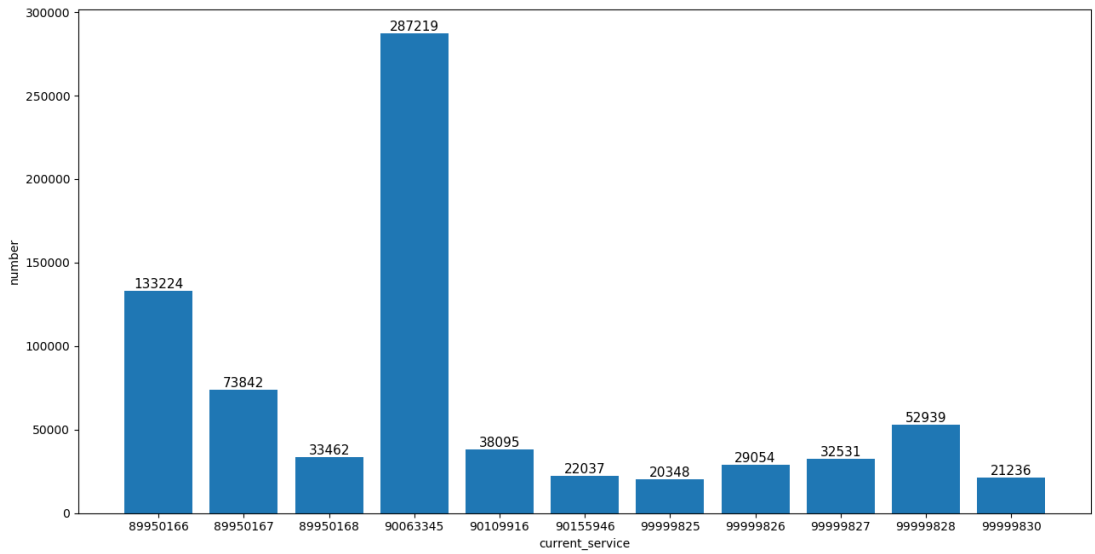
在上述基础之上，XGboost 核心原理就是提出了一组 $\Theta = \{f_1, f_2, \ldots, f_K\}$ 表示树的结构，并提出函数 $\Omega(f_t)$ 刻画树的复杂度，将其带入上述目标函数化简，对于最后的最优化目标函数问题，采用**泰勒**展开求解梯度，然后结合**贪心算法**，迭代达到最优化的结果，具体细节参见论文 « XGBoost: A Scalable Tree Boosting System »。

与传统的 GBDT 对比，XGBoost 具有很多优点，这也是我们采用 XGBoost 的原因：

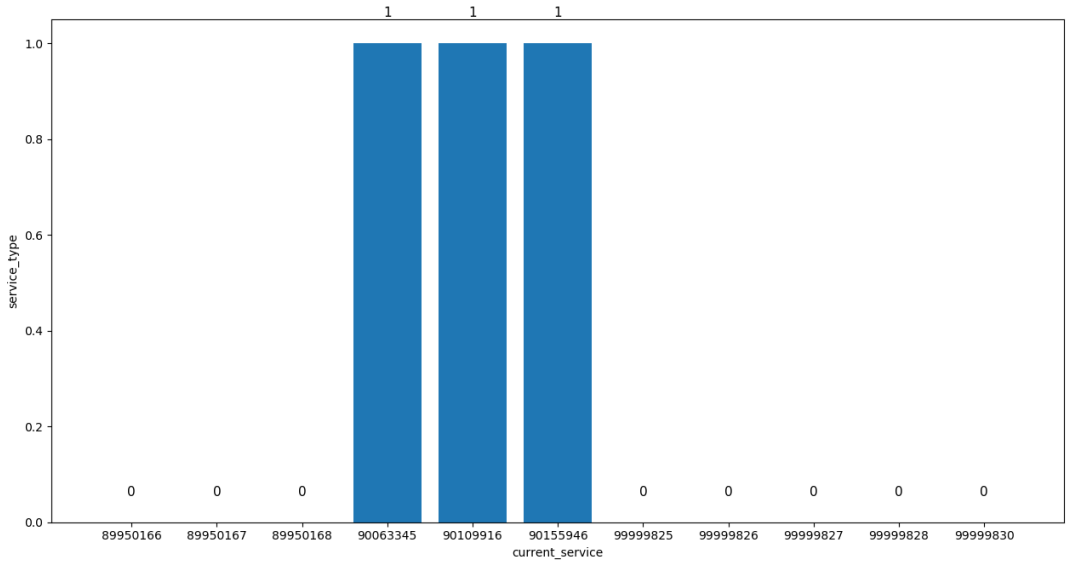
- 传统 GBDT 在优化时只用到一阶导数信息，XGBoost 则对代价函数进行了二阶泰勒展开，同时用到了一阶和二阶导数，其实计算起来并不难，这对损失函数做出来极大的优化。
- XGBoost 在代价函数里加入了正则项，用于控制模型的复杂度。正则项里包含了树的叶子节点个数、每个叶子节点上的 L2 正则项（模的平方和）。正则项使学习出来的模型更加简单，防止过拟合，这也是 XGBoost 优于传统 GBDT 的一个特性。
- XGBoost 支持并行化，在对叶子寻找最优分割的时候恰好可以并行，贪心算法在每一步寻找最优分割的时候其实还是枚举，这个时候并行计算大大减少了程序运行的时间。
- XGBoost 还特别设计了针对稀疏数据的算法，也提供了多种切分点查找算法等优化的奇技淫巧，改善了性能。

数据观察与预处理

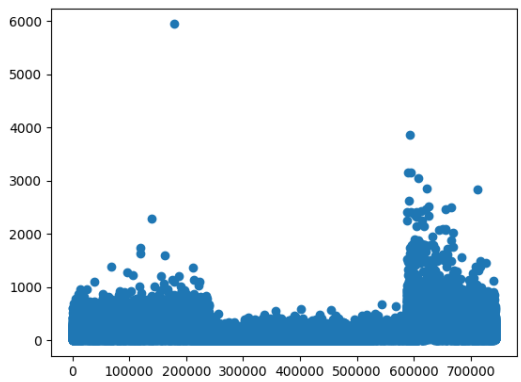
在拿到数据之后，我们使用图表的方式对数据先进行了直观的观察。部分图表如下：



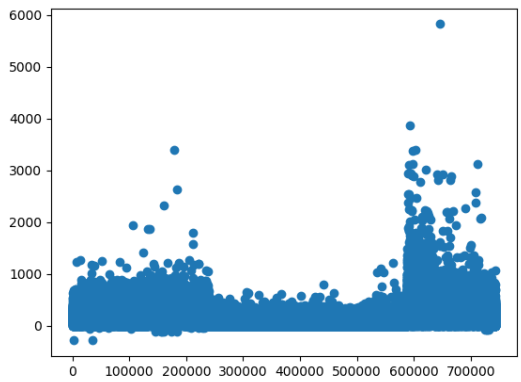
各大套餐的数量分布



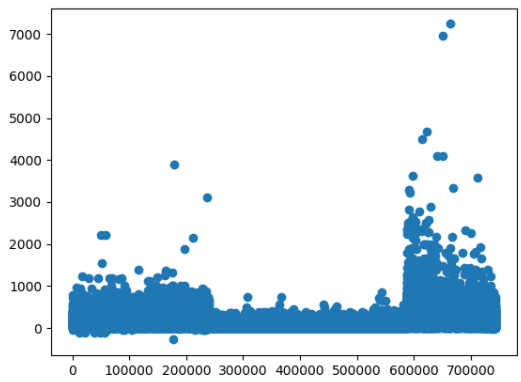
套餐与service_type对应



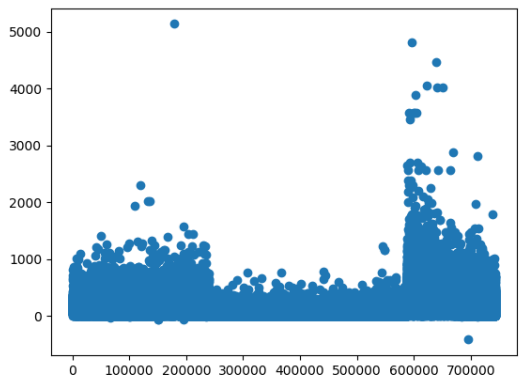
1_total_fee



2_total_fee

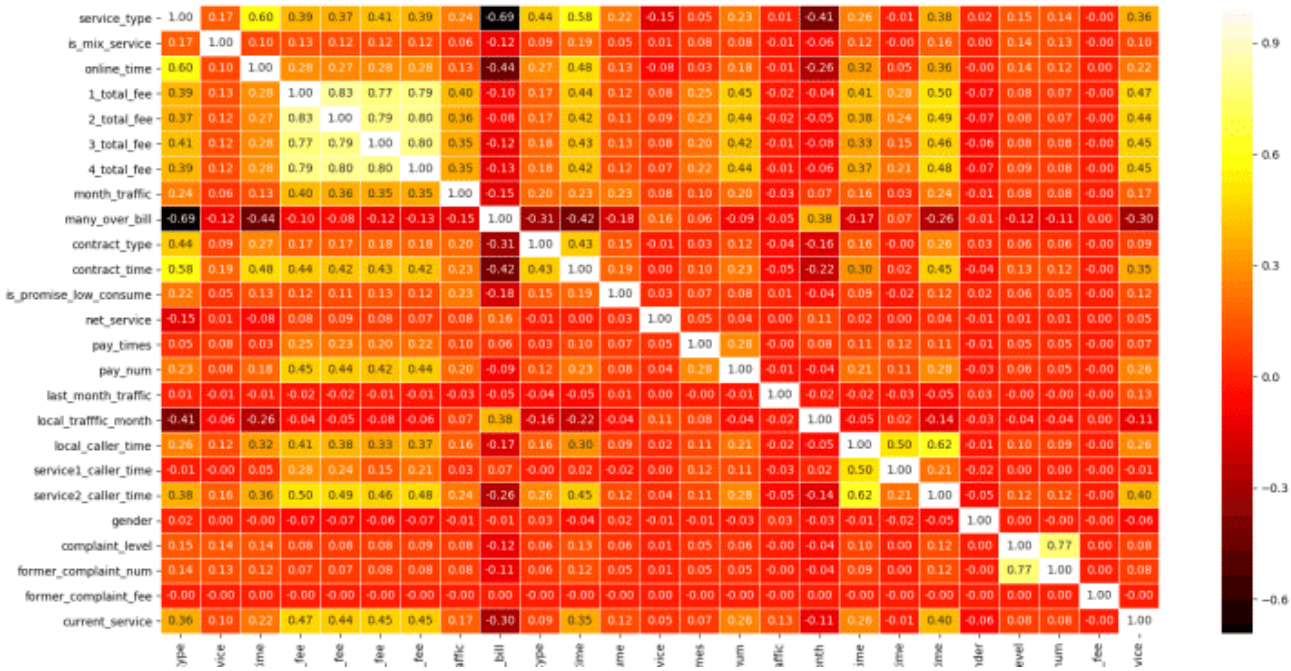


3_total_fee

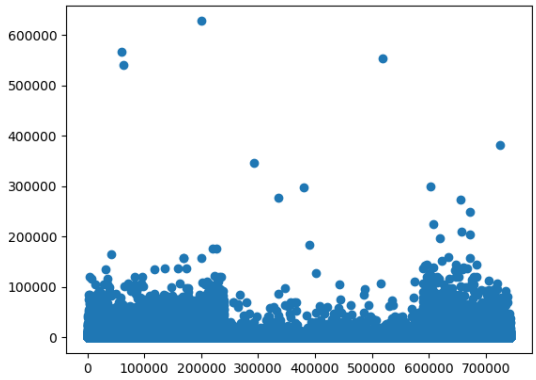


4_total_fee

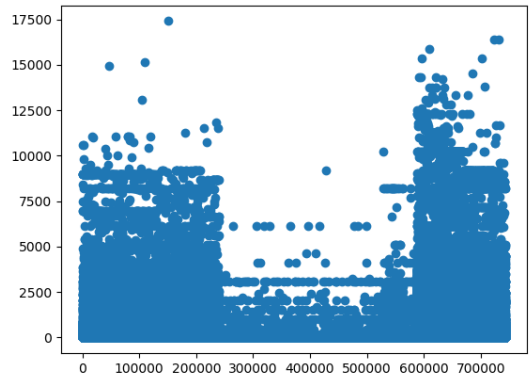
连续四个月的出账金额



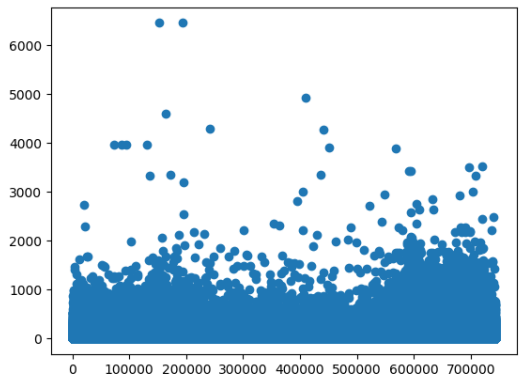
各特征的相关度



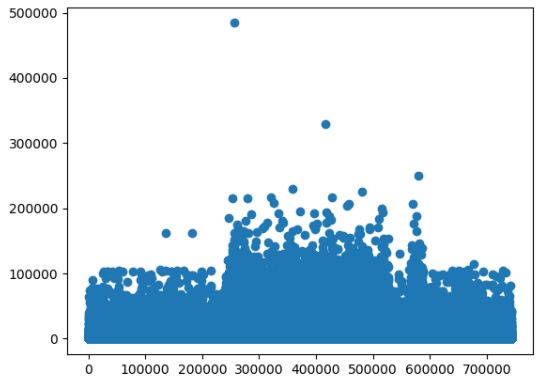
former_complaint_fee



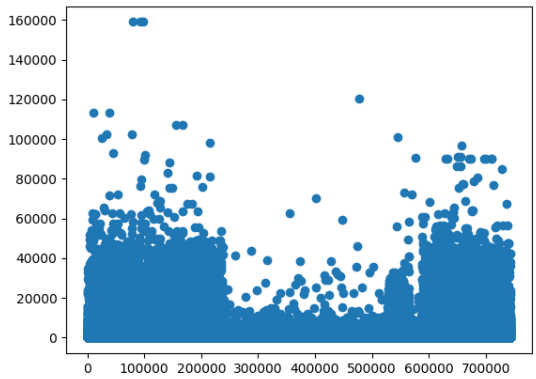
last_month_traffic



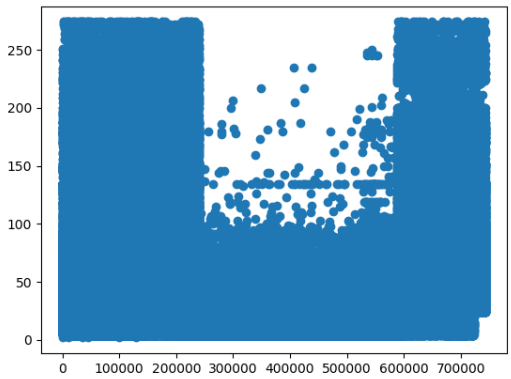
local_caller_time



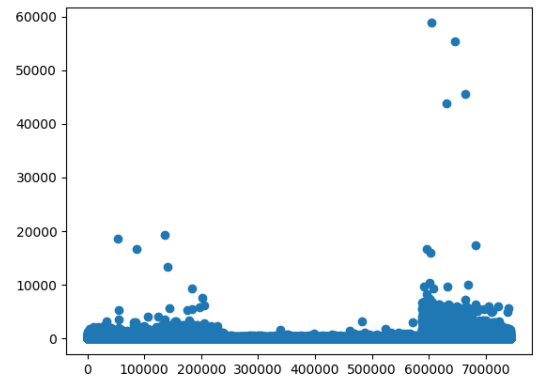
local_trafffic_month



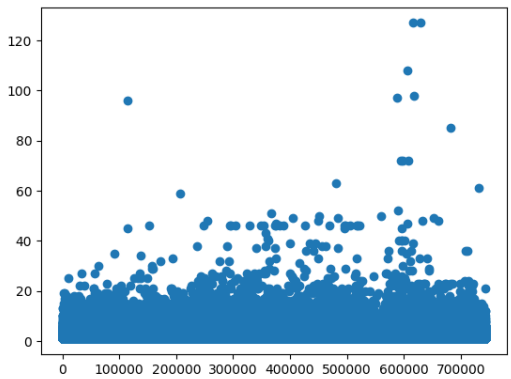
month_traffic



online_time



pay_num



pay_times

一些与套餐相关度较大的特征

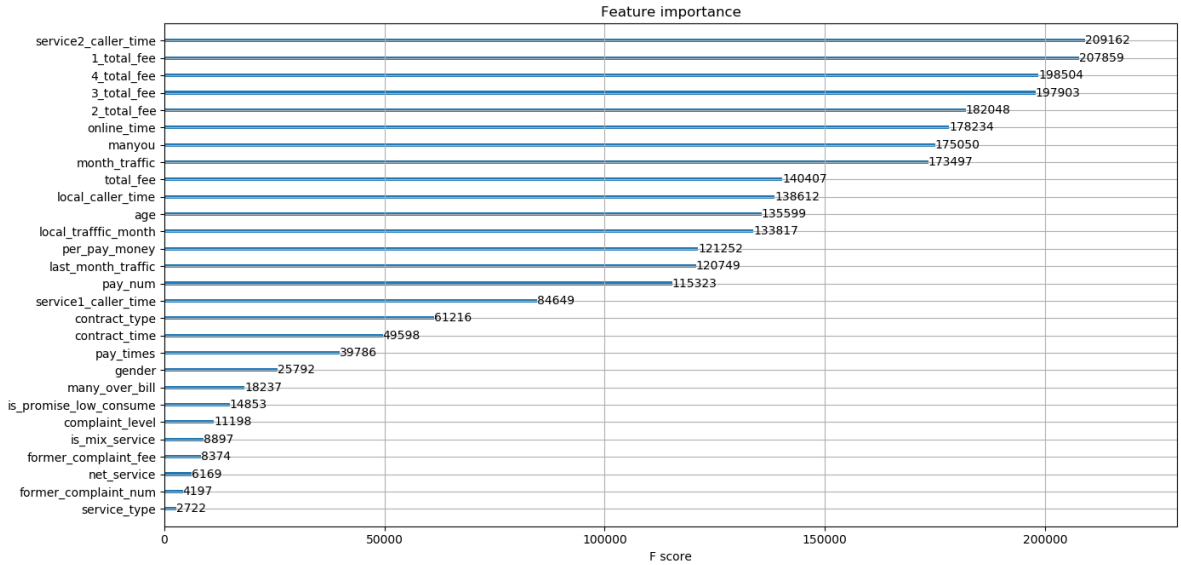
在数据预处理部分，对于缺省值的处理，我们采用改为该特征的众数的方法；对于离群点的处理，我们采用了直接删除的方法。

特征工程

在对数据的观察过程中我们发现，有关于费用，流量，通话的特征是区分不同套餐的良好特征。 我们进行了如下操作：

1. 对前四个月的出账金额进行了组合，为了描述该费用的变化过程，我们采用了加权的方法，时间离当前越近，权重越高。
2. 每个月的交费次数 pay_times 与缴费金额 pay_num，进行组合，计算每次缴费的平均金额
3. 对于每月的本地流量使用与每个月的总流量使用，计算漫游以及异地的流量使用情况
4. 对年龄，性别，投诉等级，网络类型，合约类型等进行了独热码的改造

利用 xgb 的特征重要性排序，我们得到了如下结果：



随后我们又尝试了删去不重要的特征，或将重要特征进行多项式组合等操作，得到了我们最终的结果。最好成绩为0.74144018000。 不过遗憾的是，这个成绩与我们直接运行 baseline 的结果，还有微小的差距。直接运行 baseline 的结果为 0.74175596000，这也是我们最终的榜上成绩。

四、训练赛方案

开源思路

[1. Top1 思路]

1. service_type 之间不存在交叉
2. 引导学习：在机器学习中，针对不同阶段的同源数据，通常有迁移学习和直接拼接数据的方法来利用数据，可以利用不同分布的数据提升结果。迁移学习直接对预测概率编码为被迁移模型的特征，可以保证同分布，但是不能充分利用数据；拼接充分利用数据，但是出来的结果不能同分布；为此，提出一种称为“**引导学习**”的方法。
3. 嫁接学习：结构化数据的迁移学习方法：用不同分布的数据训练得到模型，预测出来的概率作为新任务的特征编码。与一般迁移学习不同的是，嫁接学习强调利用因时间演化或者采样造成的分布产生变化的同源数据。因为被一名植物分类学家在比赛中最先使用，因此也被戏称为“**嫁接学习**”。

[2. Top2 思路]

Word2Vec 无监督学习聚类算法

1. Word2Vec 基于分布假设，将语义上相似的字词映射到几何图形上邻近的嵌套矢量。
2. 构建四维账单数据，将账单金额当作文本标签。为每个月话费各自生成10维向量，共40个特征。

特征工程

1. 业务特征：
    1. 话费减去 16 元是否是整数

    2. 流量的有效数字是否是 27 的整数倍
    3. 话费的有效数字能否被 15 乘除
    4. 话费是否是整数（用户可能未超套餐）
    5. 连续两个月套餐的差值能否被 5、10、15、27、30 等计费单元整除
    6. 四个月话费的最小值
    7. 计算流量的平均单价
    8. 计算通话时间的平均单价
2. 数据特征：
    1. 构造差值特征: 计算各个月之间费用差值(各种组合的差值)。比如，1 月和 2 月之间所缴费用差值，本月支付总额和所缴费用差值。
    2. 构造平均特征: 计算单个值除以总值的特征(比如某月消耗除以所有月份消耗)。由此可将数据处理到 [0, 1] 之间。
    3. 年龄中有异常值 0 (替换 **service_typpe** 字段原始值)
    4. gender 字段训练集与测试集不同分布(归一化处理)
3. Top6 思路：　（仅提供源代码）
    1. 参考 联通套餐
    2. 基于联通套餐，从源码看出: 对于连续值(比如费用、流量)，均作分段处理。每段取同一值。(此为特殊之处)

五、总结体会

参加这次比赛是一个收获学习的过程，虽然最后成绩不怎么样，但这两个月花费在上面的时间并没有白费。

比赛初期，我们队成员都是没有接触过数据科学的小白们，拿到数据以后一头雾水不知从何下手。还好比赛的交流群里有大佬好心，提供了一份成绩较好的 baseline。在 baseline 的基础上我们也逐渐掌握了一些门道，尝试着利用已学到的知识一点一点的改变 baseline。哪里不会点哪里，从最开始的数据预处理，到特征的提取，再到后面的模型选择。不知不觉中也学到了不少知识，更关键的是有现成的数据可以直接拿来练手。虽然我们所有的尝试最终仍然没有干过 baseline。。但我相信如果还有下次，一定会比这次做得更好。。

训练赛期间，阅读第 $1^{st}, 2^{st}, 6^{st}$ 的开源模型后，虽然处理方法各有千秋，但确实都对数据具有更深的认识。对于数据，我们未考虑到流量、话费是否整除某个数，对于联通手机套餐的咀嚼确实不够细致。同时，我们没有考虑到训练集与测试集的分布是否有所不同，而仅仅局限于训练集的部分分析(没有意识到这个)。对于训练模型的算法，我们尝试了 xgb 以及 lgb 算法，相较高手来说，机器学习的算法懂的太少，后来从源码中也可看出，领头羊都各自用了一些黑科技，而不是用一个算法来训练。

这次比赛让我们敲开了数据科学的一扇门，这个课题中，我们完整地经历了对背景的分析，需求的挖掘，特征的提取，算法的调研，模型的选择，不仅从中获益匪浅，更对以后的学习和科研提高了积极性和热情。