**2017京东金融信贷需求预测 赛题总结**

**赛题数据**

京东金融提供8到11月的用户借款信息，订单购买信息，页面点击信息、用户个人信息和贷款初始额度信息

要求预测12月每个用户的贷款信息，评价方式采用RMSE函数

**数据格式**

* 用户个人信息：用户ID，年龄，性别，激活日期，初始化金额
* 点击信息：用户ID，点击时间，点击pid、param参数（页面信息）
* 贷款信息：用户ID，贷款时间，贷款金额，贷款分期数
* 购买信息：用户ID，购买时间，价格，数量，种类ID，折扣

**数据分析**

题目给定任务是一个回归预测，也就是训练模型预测每一个用户在12月份的贷款金额

* 用户贷款与不贷款：经过分析，总共有90000个用户，而每个月大概只有20000左右的用户会贷款，分别画出90000个用户和20000个用户贷款金额密度分布图；发现对于有贷款行为的用户，他们的贷款金额大致服从正态分布，对于所有用户，用户贷款金额分布明显分为两段。根据此分析，可以先对用户分布做一个二分类，先将会贷款和不会贷款分出来，然后对会贷款的用户做一个回归预测。事实是根据已有数据二分类的准确率一直很低，所以此流程不适合，最终得到的分类结果可以作为一个特征。
* 每个用户的行为都是和时间有关，每个时间段他的行为都有所区别，导致有些用户在这个月借款了在下个月可能又不会借款，即随着时间移动，用户的借贷需求是在变化的。即不同时间段不同行为使得不同用户具有唯一性，但是每个用户存在唯一性之外，用户与用户之间还存在共性，只有这样，机器学习的模型才可以通过这些共性去预测其他用户样本。
* 业务逻辑：例如借款业务，有借款就有还款，通过借款金额和还款期数，我们可以构造每月的还款特征；每个用户都有借款限额，可以通过判断每月用户还款金额与限额之间的关系去构造特征。
* 异常数据：在数据中存在很多异常数据

**构建特征**

对于购买、点击、贷款：首先根据已有信息构造衍生字段，然后根据第一次、最后一次、金额最大一次、时间窗内计数、比值、时间间隔特征

交叉特征：三个表之间互相交叉构造特征

业务特征：根据贷款金额、分期数、限额等构造分期特征

分类特征：构造多种二分类，生成二分类特征

**借贷信息特征**

| **基础字段** |  |  |
| --- | --- | --- |
| uid | 用户ID |  |
| loan | 贷款时间 |  |
| loan\_amount | 贷款金额 |  |
| plannum | 分期数 |  |
| 衍生字段 |  |  |
| amt+per+plan | 每期金额 |  |
| interval\_between\_loan | 贷款周期 |  |

| **最近一次贷款特征** | **时间划分** | **特征计算公式** |
| --- | --- | --- |
| 用户A最近一次贷款的金额 | 最近一次 |  |
| 用户A最近一次贷款的分期数 | 最近一次 |  |
| 用户A最近一次贷款的每期金额 | 最近一次 |  |
| 用户A最近一次贷款的贷款周期 | 最近一次 | 最后一次贷款距离第二次贷款时间距离 |

| **三个月内贷款行为特征** | **时间划分** | **数值处理** |
| --- | --- | --- |
| 总贷款金额 | 1、3天、1、2、3周、1、2、3个月 | 减去均值除以方差 |
| 总贷款次数 |  |  |
| 总贷款均值 |  |  |
| 贷款金额max、min |  |  |
| 总贷款分期数 |  |  |
| 总贷款分期数均值 |  |  |
| 总贷款分期数max、min |  |  |
| 每期贷款均值 |  |  |
| 每期贷款金max、min |  |  |
| 平均贷款周期 |  |  |
| 贷款周期max、min |  |  |
| 最近一次贷款行为离现在的天数 |  |  |
| 用户各次贷款离现在的天数的平均 |  |  |

| **第一次贷款** |  |  |
| --- | --- | --- |
| 第一次贷款离现在的天数 |  |  |
| 第一次贷款金额与限额（limit）的比 |  |  |
| 第一次贷款日期与激活日期（激活日期）相差的天数 |  |  |

**购买行为特征**

| **基础字段** |  |  |
| --- | --- | --- |
| uid |  |  |
| buy\_time |  |  |
| price |  |  |
| qty |  |  |
| cate\_id |  |  |
| discount |  |  |
| 衍生字段 |  |  |
| bllamt | 交易金额 | 购买金额-折扣金额 |
| ratio\_discount | 折扣占比 | 折扣金额/购买金额 |
| 购买间隔 |  |  |

特征构造方式

| **特征构造** |  |  |
| --- | --- | --- |
| 第一次购买 |  |  |
| 最近一次购买 |  |  |
| 时间窗内 |  |  |
| 单价最贵一次 |  |  |
| 总价最贵一次 |  |  |
| 数量最多一次 |  |  |
|  |  |  |

滑动的时间窗口，从用户角度出发，默认用户的某种属性不随着时间变化，挖掘用户的股友属性

| **特征名（用户最近一次购买）** |  |  |
| --- | --- | --- |
| 购买商品种类 |  |  |
| 购买商品单价 |  |  |
| 购买商品总金额 |  |  |
| 购买商品数量 |  |  |
| 购买商品折扣 |  |  |
| 购买折扣占比 |  |  |

|  |  | **时间划分** |  |
| --- | --- | --- | --- |
|  | 用户A三个月内的购买总额/ | 1、3d 1、2、3w 1、2、3m |  |
|  | 用户A三个月内的购买次数 |  |  |
|  | 用户A三个月内的购买均值1 |  |  |
|  | 用户A三个月内的购买max min |  |  |
|  | 用户A三个月内的购买数量 |  |  |
|  | 用户A三个月内的购买均值2 |  |  |
|  | 用户A三个月内的购买产品价格的分位数 |  |  |
|  | 用户A三个月内使用的discount的次数 |  |  |
|  | 用户A三个月内使用的discount的占比 |  |  |
|  | 用户A三个月内使用的ratio\_discount的分位数 |  |  |
|  | 用户A三个月内购买的产品类别数 |  |  |
|  | 用户A三个月内购买的产品的数量 |  |  |
|  | 用户A三个月内购买单价最贵的产品的花费 |  |  |
|  | 用户A三个月内购买总价最多的一次的花费 |  |  |
|  | 用户A三个月内购买数量最多的一次花费 |  |  |
|  | 用户A三个月内的平均购买周期 |  |  |
|  | 用户A三个月内单价最贵的产品的花费占三个月内花费的比 |  |  |
|  | 用户A三个月内单价最贵的产品的销量占三个月内销量的比 |  |  |
|  | 用户A三个月内单价最贵的产品的次数占该产品被购买次数的比 |  |  |
|  | 用户A对单价最贵的产品的购买次数 |  |  |
|  | 用户A最近一次购买离现在的时间 |  |  |
|  | 用户A单价最贵一次购买离现在的时间的平均天数 |  |  |
|  | 用户A单价最贵一次产品最近一次购买离现在的天数 |  |  |
|  | 用户A总价最多一次购买离现在的时间的平均天数 |  |  |
|  | 用户A总价最贵一次产品最近一次购买离现在的天数 |  |  |
|  | 用户各次购买离现在的天数的平均 |  |  |

**点击信息特征**

| **基础字段** |  |  |  |
| --- | --- | --- | --- |
| uid |  |  |  |
| click\_time |  |  |  |
| pid |  |  |  |
| param |  |  |  |
| 衍生字段 |  |  |  |
| interval\_betweet\_click | 点击时间间隔 |  |  |

特征时间划分

| **特征时间范围** |  |
| --- | --- |
| 第一次 |  |
| 最后一次 |  |
| 现在 |  |
| 最近一次之前 |  |
| 点击最多一次 |  |
|  |  |

分别构造计数特征和比值特征

**交叉特征**

| **交叉特征** |  |  |
| --- | --- | --- |
| 贷款与购买交叉特征 | 贷款总金额与购买总金额 |  |
|  | 第一次贷款时间与第一次购买时间 |  |
|  | 最近一次贷款时间与最近一次购买时间 |  |
|  | 贷款次数与购买次数 |  |
|  | 贷款时间间隔与购买时间间隔 |  |
|  |  |  |
|  |  |  |
| 贷款与点击交叉特征 | 第一次贷款时间与第一次点击时间 |  |
|  | 最近一次贷款时间与最近一次点击时间 |  |
|  | 贷款次数与点击次数 |  |
|  | 贷款时间间隔与点击时间间隔 |  |
|  |  |  |
| 购买与点击交叉特征 | 第一次购买时间与第一次点击时间 |  |
|  | 最近一次购买时间与最近一次点击时间 |  |
|  | 购买次数与点击次数 |  |
|  | 购买时间间隔与点击时间间隔 |  |
|  |  |  |
|  |  |  |
|  |  |  |

**生成模型**

采用线性模型LR、RF以及sklearn GBDT、xgboost进行学习，得到多个分类器，然后进行stacking融合。

**模型融合**

stacking：

* 用数据集D来训练h1、h2、h3...
* 用这些训练出来的初级学习器在数据集D上面进行预测得到次级训练集
* 用次级训练集来训练次级学习器

这样实现室友很大缺陷，用这些模型在D上面再进行预测得到的次级训练集肯定是非常好的，会出现过拟合现象。（新的特征之间差异性很小）

采用交叉验证的思想来实现stacking模型

次级训练集的构成不是直接由模型在训练集D上面预测得到，而是使用交叉验证的方法，将训练集分成K份，对于每一份，用剩下的训练集取训练模型，然后预测这一份的结果。重复上面步骤，直至所有都预测出来，这样就不会出现上面的过拟合情况。

1. 将数据集为k各部分，有n个分类器，即n个模型
2. for i in range(n):

​ for j in range(k):

​ 对于训练集来说，用第j部分作为预测，其余部分作为训练集，获得预测的输出作为

1 将数据集为k各部分，有n个分类器，即n个模型

for i in range(n):

for j in range(k):

对于训练集来说，用第j部分作为预测，其余部分作为训练集，获得预测的输出作为第i部分 的新特征，

对于测试集来说，使用训练得到的模型对测试集进行预测

获得使用模型i在全部训练集上输出的一个新特征。（拼接方式）

获得使用模型i在全部测试集上输出的一个新特征。（平均方式）

在训练集上获得n个新特征

在测试集上或者n个新特征

使用新的训练集和测试集生成第二层模型，对结果进行预测

**blending**

Blending直接用不相交的数据集用于不同层的训练。

以两层的Blending为例，训练集划分为两部分（d1，d2），测试集为test。

第一层：用d1训练多个模型，讲其对d2和test的预测结果作为第二层的New Features。

第二程：用d2的New Features和标签训练新的分类器，然后把test的New Features输入作为最终的预测值。

**存在的问题**

初次比赛，还不懂套路，拿到赛题和数据后，我们没有做数据分析，直接巴拉巴拉暴力提取特征，导致很多特征都没有什么实际用处。另外在数据清洗上也没有进行操作，直接选取了全部数据；在一些缺失值，异常值（比如购买金额，贷款金额为负值）操作上，直接进行填零。不知道是否因为这些原因导致结果比较差。另外，在比赛初期，我们新增特征时，对结果有较大提高，特别是一些强特征（GBDT特征重要性排序），但在比赛后期，新增特征后，对结果几乎没有提高，有时甚至下降了。在本次比赛中，没有找到什么骚操作，也没有发掘出某个神奇的东西，对结果有大幅提高。但是，TPOT是个不错的东西，在特征确定的情况下，通过该工具，对模型调整有着突出的效果，缺点是运行速度太慢。过段时间，看看各位大佬的分享，再好好总结总结。