特征工程

main.py
47.4KB

main.py是【企业倒闭预测】的参考代码，加了些注释

0 关于模型！

- 首先明确，我们的输入样本是由（sku_id, 日期）唯一标识的。
- 输入的样本的 **特征向量** 可以说包括两部分：
  - （1）sku本身的**一直不变**的特征，比如 品牌，类别，**上架日期**等。
  - （2）**因“日期”不同而变化**的特征，比如 ~~点击量、~~是否应季、活动类型、**已售天数**。
  - 为什么要把点击量划掉，请看下文
- 输入的样本的 label是 该skuid 在该日期的 销量。
- 预测时，遍历所有要预测的（sku_id，data_date日期）元组，遍历的同时，根据一个已经计算好的**映射表**，将元组映射到相应的**特征向量（对于同一个sku，不同的特征是那些随日期变化的特征），**然后输入到训练好的模型中，模型就会预测出 该sku在那一天的销量
  - 但是，这里的日期是20180501开始的30天，这些天的“**随日期变化的特征**”只知道是否应季、活动类型、已售天数。
  - 因此，对于点击量这种，我觉得只能把点击量进行统计之后，当作“**sku本身的一直不变的特征**”，比如 点击量的总和、**应季条件下**点击量和以及占总和的比例、某个月份/季度/双十一阶段的点击量和、刚上架一个月内的点击量，等等。当然**直接算加和**或许不大好，可以**算日均的**。
- ***预测***时，如何做到，从（sku_id，data_date日期）元组，到特征向量 ，的映射，这个映射过程需要：
  - 一张表A，记录某个sku某一天的 **因“日期”不同而变化**的特征，我估计有且仅有：
    - 日期
    - 是否应季
    - 已售天数
    - 这一天的年份、月份、几号、周几、所在季节。
    - 活动类型
    - 该sku是否**曾经**参与过该活动类型的活动（注意比对时间先后）
    - 该活动类型的时间段内，该sku的日均销量
      - 当然这个特征也能换个角度，算作**“sku本身的一直不变的特征”**，即 某sku在活动类型1,2...5的时间段内的一般日均销量。
  - **其它的特征都**是记录在：一张记录sku本身的一直不变的特征的表B
  - ***其实也可以说是在一张表里，因为咱们已经知道要预测的元组都是哪些了。***
  - 【注】以下用 A类型 B类型 来 区分这两大类特征
- 刚刚说的是预测时应该怎么做，那是在训练完模型之后。那么如何训练模型：
  - 这时我们针对的就不仅仅是要预测的那些 sku_id和data_date

- 而是包括 给的所有数据 中的 （sku_id,data_date）主码 代表的 特征向量
- 即，需要一张表格，表格以（sku_id,data_date）为主码，***特征包括 A类特征，也包括 B类特征***

1 daily

- 关于data_date（原始为字符串形式），***A类***
  - 1. 直接作为一个特征（转为int类型？）
  - 2. 提取 哪一年、哪个月、该月的哪一号、甚至把星期几也提取出来作为特征
    - 这个计算，可以借助pd.datetime，像这样：先转换为datetime类型，这个类型的属性.year .month等等能直接获得上述信息。

```
import datetime
df.loc[:,'date_time'] = pd.to_datetime(df.data_date.apply(str))
```

- 关于在售天数
  - 直接用，***A类***
  - 算出上架日期，***B类***。同时也有利于下面的计算
- 我们把 goods_click, cart_click, favorites_click, sales_uv 即 点击、加购、收藏、购买，作为4个***指标***
  - 首先计算比例关系，得到另外6个***指标***，一共10个指标
    - 购买/收藏
    - 购买/加购
    - 购买/点击
    - 收藏/加购
    - 收藏/点击
    - 加购/点击
    - 计算6个指标的和
  - 对于上述10个指标，参照我在前面**【关于模型！】章节**中对于“点击量”的叙述，推广到对这10个指标进行各种操作，***B类***。下面以点击量为例：
    - 点击量的总和
    - 刚上架一个月内的点击量
    - **应季的时间段内** 点击量和 以及占总和的比例
    - 每个月份的点击量和。
      - 将这个特征理解为“该sku在5月份的一般销量”，诸如此类。
      - 考虑到不同sku的上架日期不同，上面的这个特征或许可以调整为，“该sku在上架一个月内的销量”，以此类推，在2个月内、在半年内。。？
    - 每个季度的点击量和
    - 双十一阶段（具体时间段自己掐）的点击量和
    - 活动类型1的时间段下的点击量和
    - 活动类型2的时间段下的点击量和
    - …
    - 上述的“点击量和”，或许做成“日均点击量”会更好？
  - 【核心思想】把有关于 ***销量、价格、点击量*** 的，这种 跟日期有关、但是20180501开始的这五周的数据不知道的，处理为 ***B类特征***

2 info

- 关于季节属性：
  - 直接作为特征，***B类***
  - 分为（春，夏，秋，冬）四个特征，每个特征填入0或1，***B类***
  - 根据data_date计算是否应季，***A类***
- 关于品牌ID，***B类***
  - 直接作为特征
  - 统计相同ID的goods的个数，作为一个特征
- 关于类别，***B类***
  - level1 level2作为两个特征输入
  - 相同level1的goods的个数
  - 相同（level1, brand_id）的goods的个数
  - 相同（level1, level2）的goods的个数
  - 相同（level1, level2, brand_id）的goods的个数
  - 【注】现在对于 这种类别型的特征，都是直接输入，先不进行“独热 + 降维”
  - 【奇思异想】level1 + level2， level1+level2+brand_id
  - 【甚至】还能再考虑level3

3 sale

- [x] ~~每个goods的记录的条数、总销量、日均销量，***B类***~~
- [ ] 630247 个GS
- goods_num销量，作为label，同时，把它（1）看做***一个指标***，（2）参照daily中对于点击量这一指标的处理，统计出类似的 ***B类*** 特征 如下：
  - 可以先总的mean sum一下
  - 某个[time1,time2]时间段内的日均销量：
  - 1.根据 data_date是否 大于等于time1且小于等于time2（可以直接字符串比较），筛选出这个时间段的sale记录，以GS为分组键，对goods_num求和，作为一个特征。
  - 2. 然后除以记录条数？还是这个时间段的总天数？日均销量，作为一个特征
    - 【尴尬】这归结于：缺失的记录 与 销量为0的记录 是否真的是一样的。
    - 不管了直接.mean，按记录条数。结果必然不是nan
  - 3. 所以下面这些特征的不同，就归结于 [t1,t2]时间段的取值不同
    - 开始日期 和 结束日期 是要 字符串类型~~，还是datetime类型~~
  - 4. 都套在一个循环里面就成
    - 每个**月份**的销量，或日均销量
      - 12*2
      - 【尴尬】20170301-201803016，怎么手动划分？最后多半个月？
      - 或：不放到这个循环里面去算，直接根据goods,sku,datetime.month进行分组，然后对goods_num进行sum和mean。
    - 每个**季节**的销量，或日均销量
      - 4*2
      - 同样有个问题，最后多半个月
      - 那还是根据Month来映射，然后分组，然后sum/mean。
    - **应季时间段**下的销量和，或日均销量
      - 2
      - 与季节类似

        - 某个类型的**活动 / 活动+节奏的时间段**内 的销量和，或日均销量
            - 活动6*， 活动+plan 12*2
            - no..：21*2 36*2
            - 不对！！！ 实现起来比较难
            - 先把 M MP 的分段时间表拿到，
            - 然后丢到循环里面去算，得出一堆特征，然后再把同一活动类型的弄出来取个平均
        - **双十一阶段（就取[11.6-11.16]）**的销量，或。。。。
            - 2
        - **刚上架30天/10天/60天**内的销量
            - =》需要上架日期-timedelta(30days)
            - 上架日期，有些异常，取个众数
- 关于goods_price均价与original_shop_price吊牌价 这两个指标
    - 计算 折扣率：（吊牌价-均价）/ 吊牌价
    - 对于以上这***三个指标，***同于上面***对于指标的处理***
        - 每个月份/季节的均价/吊牌价/折扣率 均值
        - 应季条件下的。。。。。
        - 。。。。
-

4 promotion

- 对于 shop price 标价 和 promote促销价 的处理，同于上面对于 goods_price均价与original_shop_price吊牌价 的处理。
- 但是考虑到 促销活动 不像sale daily那么连续，或许要其它的处理办法?
- 如何处理促销始末时间?
    - 始末时间相减，得到这个**商品的促销天数**，天数多的可能有问题。

5 marketing

- 对于每个类型的活动，计算：
    - ~~活动存在的总天数（不考虑活动是否连续）~~
    - ~~活动次数/频率~~
    - ~~总天数/活动次数 = 每次举办此类活动时的一般天数（考虑活动是否连续）~~
    - ~~***A类***：该活动持续了多久~~
    - ~~有多少种plan。***A B类两种理解，做A类吧***~~
- 再对于每个（活动类型marketing，活动节奏plan）的元组，计算：
    - ☐ 元组存在的总天数（不考虑活动是否连续）
    - ~~元组次数/频率~~
    - ~~总天数/活动次数 = 每次举办此类元组时的一般天数（考虑活动是否连续）~~
    - ~~***A类***：该活动节奏元祖持续了多久~~
    - ~~该（marketing，plan）的天数 占 活动类型marketing的天数 的比例，上面有两个特征除一下即可。***A类***~~
- 总体思路有两条：把每个类型的活动作为一个活动单位；每个（活动类型marketing，活动节奏plan）的元组作为一个活动单位