教师以前申请的项目数--lhy

* + 一个教师三个月、一个月、一周内分别申请的项目数目
  + 教师以前申请成功的项目数比例（通过率，经别人测试后特征权重比较大）
  + 同一个教师申请的资助类别是否以前申请过？
    - 曾经是否申请过
    - 过去半年内是否申请过
    - 过去3个月内是否申请过
    - 过去1个月内是否申请过
    - 过去1周内是否申请过
* teacher\_prefix--hee
  + 性别类别
  + 缺失值的填补
    - 用出现最多的特征代替
    - 根据概率分布生成一个值
  + 类别      Mr, Mis, Miss, teacher,   (类别的东西，在表示的时候不希望让机器觉得谁大谁小)
  + **one\_hot**方式还是数字表示！！   类别数据one\_hot is better， 有层次的数据，标量is better
* school\_state--lirui
  + ~~school\_state所在东部、西部、中部~~
  + ~~每个州有多少学校~~
    - ~~每个州好学校数量~~
  + ~~school\_state的外部贫困数据：贫困、富裕、一般~~
  + 每个州的通过率(11%-18%)
    - 带权重的one-hot  [0.18,0,0,0,0], 权重是该州的通过率，但是各个州的通过率差距不大，建议采用平方，立方的形式？
    - 53维，稀疏问题？
    - 两个特征：1)州的one\_hot？二进制的表示？,  2)通过率
* project\_submitted\_datetime--lhy
  + 转为时间戳（在别人的特征选择里权重还是比较高的）
  + 分为开学时间、假期时间、放假时间、学期中间
  + 选择一个还是两个特征？
* project\_grade\_category--lhy
  + project\_grade\_category，分为两类/四类 (PreK-2, 3-5, 6-8, and 9-12)
  + one\_hot is better.   (~~还是数值？（毕老师也疑惑）~~)
* project\_subject\_categories--lxw
  + 类别整理
  + 位向量编码  [1,1,0,0]
  + 类别数目（数值）
* project\_subject\_subcategories--lxw
  + 类别整理
    - 构建与project\_subject\_categories的树型结构（构建从project\_subject\_subcategories到project\_subject\_categories的映射关系，对映射后的结构使用one\_hot编码表示，可降维）
    - 每个父节点有几个孩子节点
  + 位向量编码？？文学-科学[1,0,0，*,  0,1,0,0,\**,0,0，]
    - 太稀疏？
  + 类别数目（数值）
* quantity--hee
  + sum, min, max, std, mean
* price:--hee
  + sum, min, max, std, mean
* quantity\***price:--hee**
  + sum, min, max, std, mean
* essay
  + 当前essay与approved\_essay的相似度大还是与not\_approved\_essay的相似度大(0/1)---lirui
    - approved\_essay: 所有通过的文本组合
    - not\_approved\_essay: 所有没有通过的文本组合
    - n-gram，Jaccard距离，向量空间相似度，海明距离等
      * knn    样本不平衡--> 负样本的距离×3   正样本距离×1
    - 是所有essay的合并还是分别与每个essay比较相似度
      * 新essay跟每个合并的essay比较-->knn
    - ~~连续值？~~
    - 0/1 is better
  + essay的情感极性---lxw
    - 极性
    - 主观性
  + tf-idf--
* **description**
  + **~~当前description~~**~~与approved\_description的相似度大还是与not\_approved\_description的相似度大(0/1)~~
    - ~~approved\_description: 所有通过的文本组合~~
    - ~~not\_approved\_description: 所有没有通过的文本组合~~
    - ~~是否需要去掉重复的部分？~~
    - ~~n-gram距离，Jaccard距离等~~
  + **~~description与~~**~~project\_subject\_categories和project\_subject\_subcategories相关联，是否可以得到概率分布？~~