

10/26

1. 问题：用 random search cv, cross validation 会涉及到 data leakage 的问题

- 所以要用 pipeline 来解决，每一次 fold 都要重新进行一次完整的 pipeline（从 data preprocessing 到 model training）

- 学到了 column transformer, feature union

2. 问题：在用 PCA + Random Forest pipeline 的时候出很多次 bug

- 发现之前写的 custom estimator 里面用了很多 to_array() 把 sparse matrix 转成 dense matrix, 以及用了全局的 matrix 变量和把涉及到全局 matrix 的方程塞进了 custom estimator 导致 random search cross validation 的时候出问题

- custom estimator 尽量不要转 densmatrix, 并且整个 pipeline 就保持只用到输入的矩阵 X

3. 问题：用来 PCA 根本没有提升模型效果

- PCA 不适合降维文本 matrix, 用 truncated svd 更好

10/27

1. 问题：在写整个从 data preprocessing 到 model fitting 的 pipeline 的时候有 bug

- 注意 pipeline 里命名, '___' 的使用

2. 问题：写 custom transformer 的时候，还是会忘记 data leakage 的问题

- 如果在 custom transformer 里面要用 sklearn 里面预设的 transformer, fit() 写在 fit(), transform() 写在 transform() 里，避免 leakage

3. 问题：用了 truncated svd 模型效果也没有提升

- 回过头删掉了 $y = 0$ 的 row 之后，效果明显好多了，说明 outlier 影响很大

4. 问题：mse 有明显进步，但是 R^2 还是很差，只有 0. 几

10/28

争取把 test data text 清洗完

已经清洗完了，专门建了用来 text cleaning 的 notebook

dimension reduction 效果很好，之前把 R^2 的值看错了

还可以进一步用 model stacking, 看效果怎么样

10/29

之前 fit 错了

training set 用了 dimension reduction 提升明显，但是 testing set 的表现一坨

沟槽的 AI 把我之前的排序给消掉了，我才发现

我现在觉得原因可能是 data 里面有很歌重复出现了很多次，并且流行度变化很大，得去重一下

10/30

果然是 data 里面有很多歌重复的问题

自己加的 text matrix 和 categorical data 对于模型 generalization 没用

就用 **numerical feature** 了，最后再看 **model stacking** 还有提升没