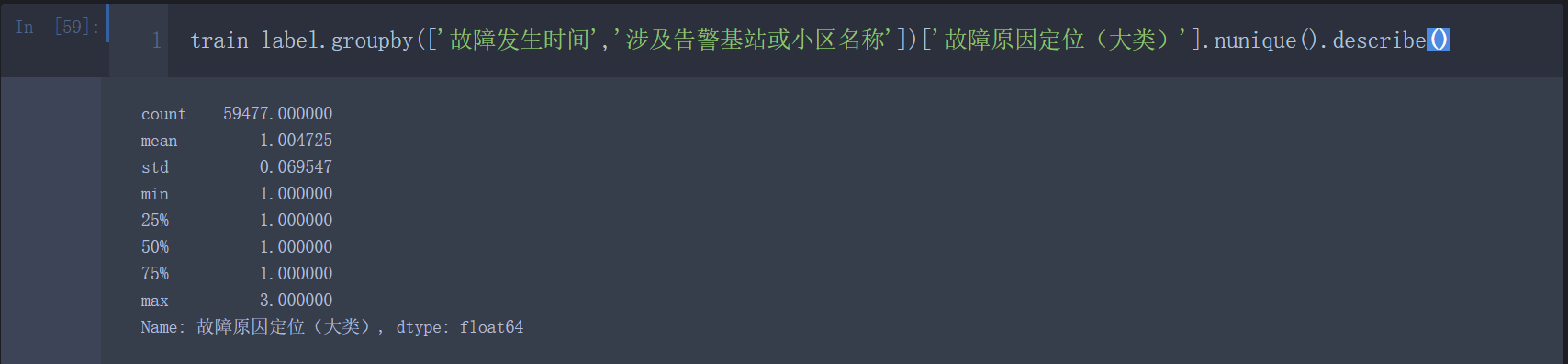
**太难了团队方案**

**1：数据清洗**

对train\_label信息完全相同的进行 去重处理，测试集中也存在完全重复信息，且部分与训练集完全重复，这部分进行进行规则直接给标签线上会减分，即存在信息相同，标签不同的数据，为不可消除误差。



**2：特征工程**

特征工程我们团队主要有两种思路，第一是nlp特征+时序特征；第二是历史统计特征加时许特征。本题重点构造时间相关特征，主要原因是告警信息只有130种，同一小区告警重复度很高，但是故障类型又显得不同，此时只能用时间先后来加以区分。

第一种方案：对告警信息采用word2cev词向量挖掘，找出告警信息潜在规律，同时考虑到word2cev周期性问题，按照周和小区生成文本去重，进行训练；时序信息主要有同一小区最早告警，最晚告警，以及按照时序的告警间隔时间。

第二种方案：对于历史数据的统计挖掘，主要包括各种针对各小区各类告警信息出现次数，同一天，同一周，告警计数，以及出现几种告警，以小区和时间为关键词，对告警数据和故障数据进行统计，并统计故障数据和告警数据的一个比率特征；时序信息主要有小时，周，天，同一小区最早告警，最晚告警，以及按照时序的告警间隔时间，故障时间的rank特征。

**3：模型lgb+xgb**

根据不同的特征集采取多维度建模，并进行一定的特征选择，模型训练时间短，12核，15min可以出结果，不同的特征工程分别对应着不用的lgb模型，同时我们团队发现一个有意思的现象，lgb不同评测指标mlogloss，merror进行融合会产生几个千的提升，分析得出mlogloss对于多分类提高auc具有很好的效果，而merror对于提升Accuracy更具优势。Xgb模型主要为了增加模型差异性，我们也针对xgb设计了一套特征，最后我们团队总共3个模型，2个lgb和一个xgb。

**4：融合方案**

采用2个lgb和一个xgb结果加权融合，对模型提升最大。最优线上分数93.0557。

**5：其他模型尝试**

我们在比赛中查阅了移动与某些高校合作的网络告警项目论文，对于其中的数据挖掘方案，进行了尝试，比如DNN，deepfm，catboost，但是由于单模徘徊在91左右，我们 并未放到最终方案中。

**6：致谢**

感谢中国移动通信集团有限公司提供的竞赛机会，真实场景下的数据让我们感受到了数据挖掘的魅力！感谢队友，团队合作的机会让我认识更多大佬，能力也越来越强！祝愿AIIA杯越办越精彩！