节目播放量预测

整体解决方案：

数据预处理：

节目主演 leader_t、节目类型 kind_t 按‘,’分词取第一个。

特征工程：

查看各节目的走势图发现大部分样本具有明显的周特征循环，所以特征工程主要按 7 期做聚合，短期或长期时间窗口聚合特征没有明显提升模型效果。

（1） 滑窗特征：取近 7 期的均值、最大值、最小值、标准差、中位数、偏度、峰度；

（2） 指数加权：取近 7 期指数加权的均值、标准差；

（3） 滞后特征：将节目播放数 cid_day_vv_t 滞后 7 期；

（4） 差分特征：当前节目播放数 cid_day_vv_t 减去 7 期之前节目播放数 cid_day_vv_t；

（5） 标准化特征：按日期 nth_day、节目所属 IP 标识 seriesId_t、节目所属频道标识 channelId_t、节目主演 leader_t、节目类型 kind_t 分组对节目播放数 cid_day_vv_t 进行 MinMaxScaler 数据归一化处理；

（6） 日期特征：当天是都是节假日 is_holiday，是否是周末 weekday；

（7） 视频更新特征：当天是否有更新 date_has_update，近 1 期、近 7 期播放视频个数 vv_vid_cnt 差值、上线视频个数 online_vid_cnt 差值；

构建样本：

将当天节目播放数 cid_day_vv_t 提前 1~7 期作为预测值，分别取最后 35 天合并作为模型样本，cid_day_vv_t 为 0 为测试样本 x_test，cid_day_vv_t 非 0 的后 7 期作为验证样本 x_valid，剩余样本作为训练样本 x_train。

模型训练：

使用 LightGBM 回归模型，重要参数如下：

'objective': 'mape',
'lambda_l1': 0.1,
'lambda_l2': 0.0,
'max_depth': 0,
'num_leaves': 402,
'min_data_in_leaf': 130,
num_boost_round = 30000,
early_stopping_rounds = 300

后处理规则：

因为评估指标 mape 对小数值更加敏感，所以增加了模型预测值后处理规则，针对验证集 mape 大于 0.4 的节目 cid_t 模型预测值使用近七期的最小值取代，阈值 0.4 由线下验证集遍历搜索最佳阈值。