代码分为5个部分：

Pre\_process 对数据集进行预处理：处理原始数据集，合并规模较小的3个日志（注册，登录，创建视频），分析4个日志中的重要特征随时间的变化趋势，掌握数据的大概情况；

Dataset\_split 使用滑窗法分割数据集。将没有分割的原始的数据集，分割成训练集，验证集和测试集。

feature\_engineering\_for\_split\_dataset 对分离后的数据集进行特征提取

final\_dataset\_process 对特征提取之后的数据集进行简单处理，便于后续算法使用

xgb\_model\_for\_new\_feature 使用xgb模型训练数据集，使用验证集观察模型效果，适当调参，在测试集上使用最终训练调整好的模型进行预测。

【pre\_process】

原始数据是txt形式，并且部分日志含有重复数据。

将重复数据转化为计数特征，表达为行为的重复次数。

将前3个日志合理合并为1个数据量较大的日志，便于联合观察特征的分布情况。

同时在合并处理过程中，处理数据原有的问题——极少量用户有创建视频的行为记录，但是没有那一天对应的登录记录，这显然是不符合常理的。需要为这些行为数据补充对应上的登录数据。

观察每个日志的关键特征随天数的宏观变化趋势，比如每天的总注册人数随日期的变化趋势。分析这种变化趋势可以初步对比观察测试集A和测试集B的数据分布情况。

【Dataset\_split】

提取特征的日期1-16 用户注册的日期1-16 预测17-23某用户是否活跃

提取特征的日期8-23 用户注册的日期1-23 预测24-30某用户是否活跃

提取特征的日期15-30 用户注册的日期1-30 预测31-37某用户是否活跃

【feature\_engineering\_for\_split\_dataset】

对分离后的数据集进行特征提取。

提取完特征后，将从合并后的3个日志中提取出的特征，和第4个日志（行为日志）中提取的特征，合并到一个表格中。

【final\_dataset\_process】

特征提取时没有进一步处理登录日志，这里先加入登录日志中的特征。

特诊提取时的辅助提取用的特征去除掉，这些特征不用于后续的建模过程中。