血压预测新模型方案

一 数据源:

1 去年公司内部采集的数据(高血压用户数据较少);

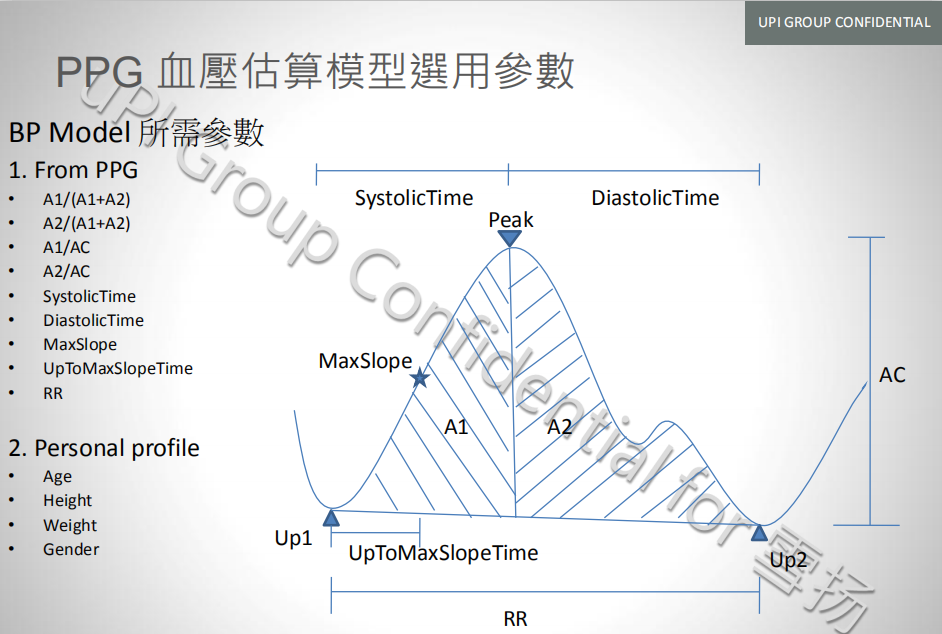
2 最近新下发的康康血压计采集的数据(高血压用户较多);

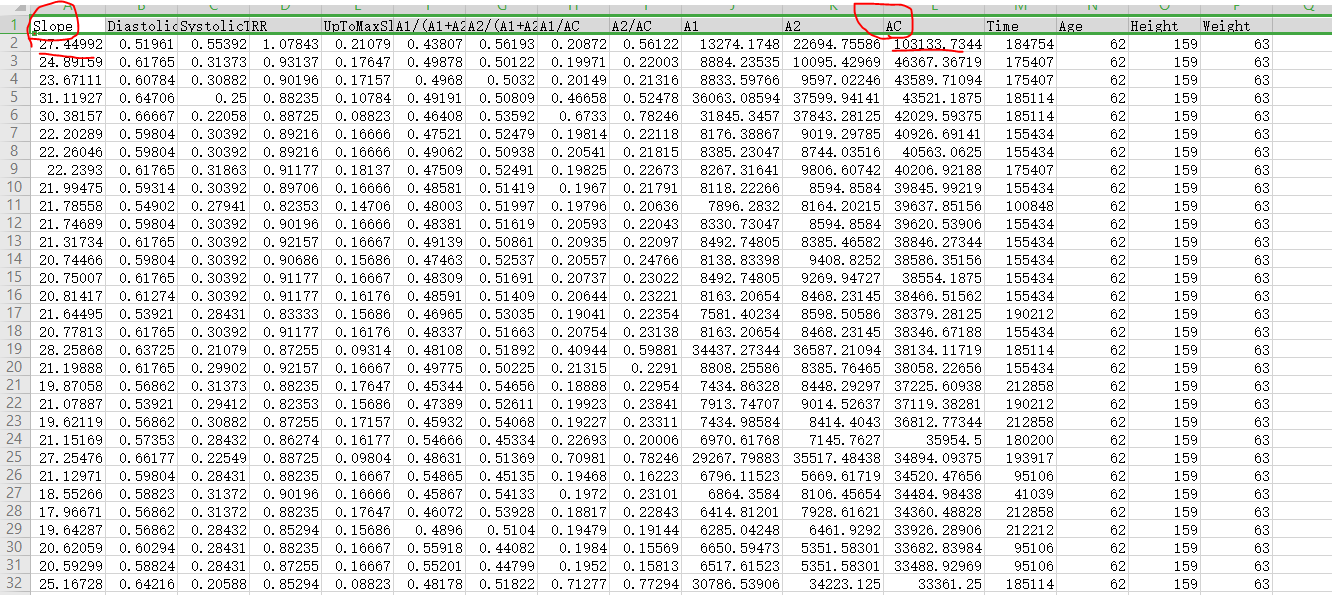
3 客服记录的部分用户的血压数据(大部分为高血压用户);

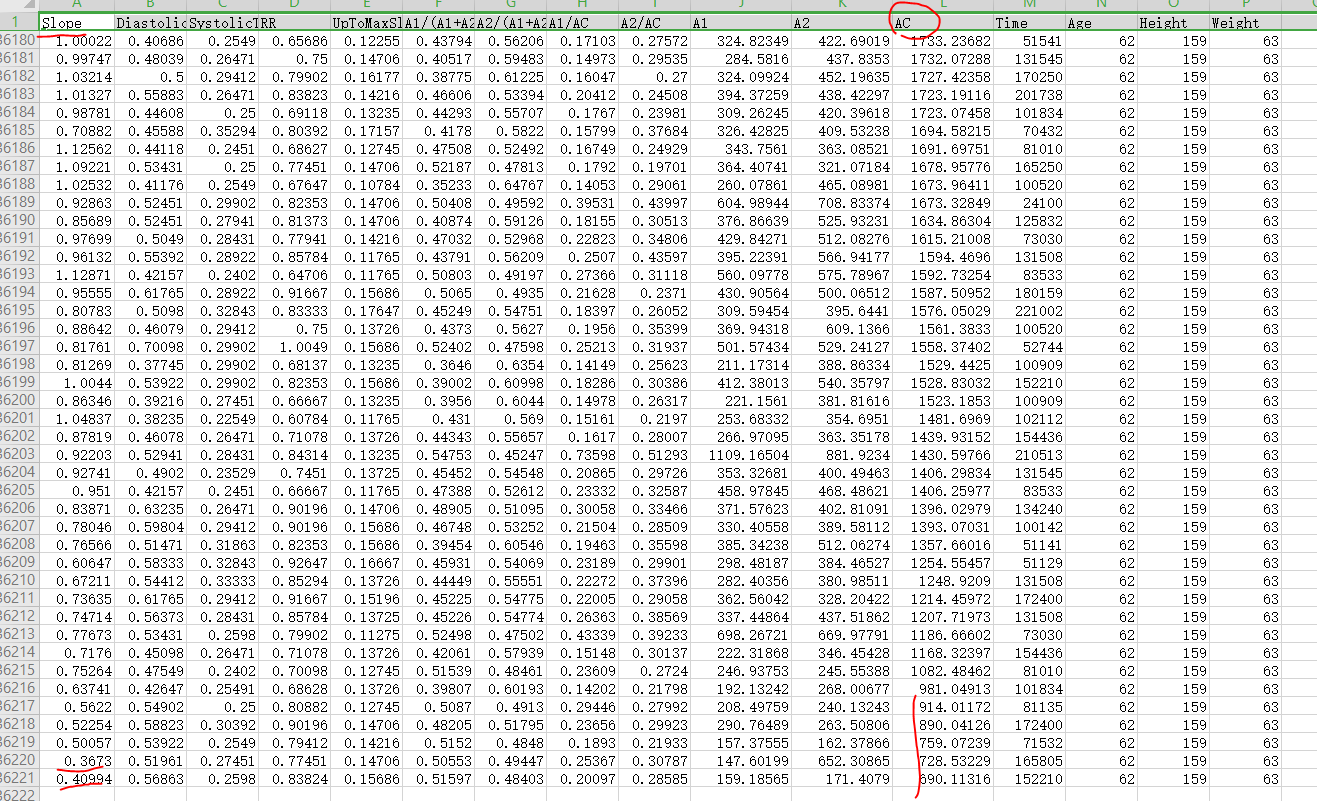
二 数据清洗方案:

1 异常血压值去除: 高压85<= SBP <=200, 低压: 45<= DBP <=140;

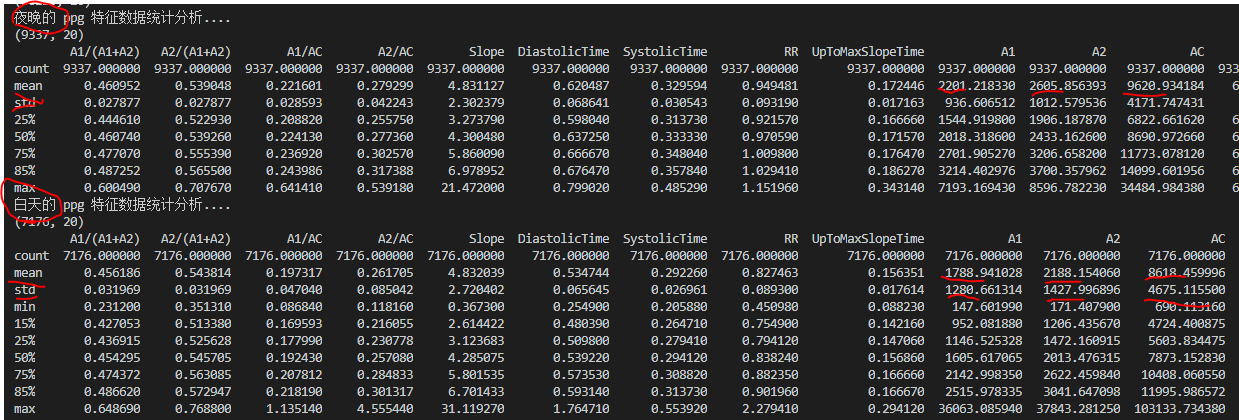
2 通过最近对一些ppg样本数据的分析,发现ppg特征数据的每个特征值, 变化范围很大,比如AC值,从最高100000,到最低700, AC代表的是 脉动波形的最高点的幅值, 目前推测这些异常的极值是由于手表PPG测量时受到了干扰, 关于此问题也向台湾UPI方面进行了咨询,但是并没有得到明确的回复;



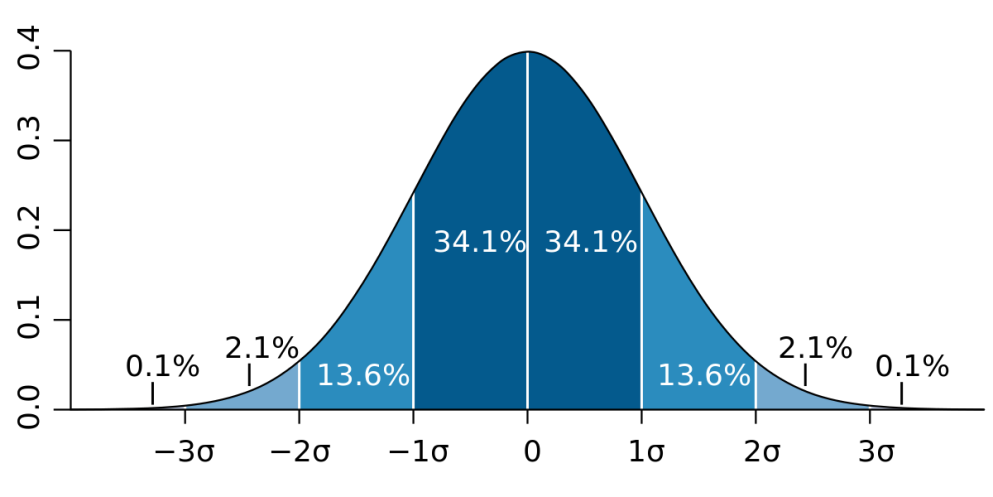




3 目前我们想到的针对于ppg特征数据的异常值的清洗方案: 每份样本数据利用统计分析, 计算每个ppg特征的均值mean,方差std, 而且在计算的时候发现,夜晚(2点-5点)的ppg数据,和白天(上午9点至晚上8点)的ppg数据,统计结果还是有很大差别的,其实这也符合实际情况,白天活动多一些,导致ppg受干扰较大.所以,综合考虑,采用夜晚的统计数据均值和方差,对数据进行过滤,去除异常值;



4 拉依达准则法（3δ准则）：如果实验数据值的总体x是服从正态分布的，则异常值（outlier）的判别与剔除，μ与σ分别表示正态总体的数学期望和标准差。此时，在实验数据值中出现大于μ＋3σ或小于μ—3σ数据值的概率是很小的。因此，根据上式对于大于μ＋3σ或小于μ—3σ的实验数据值作为异常值，予以剔除.



三 模型

1 去年公司内部采集的数据,和最近使用康康血压计采集的数据,训练一个base model,默认所有用户都使用此模型;

2 使用客服记录的数据训练一个special model,客服记录的数据里面,大部分是高血压用户,而且有的用户优于吃降压药,一天之内可能血压变化比较大,利用此部分数据单独训练一个模型,此模型主要用于反映血压不准的用户

3 对于反映血压不准的用户,或者说由于吃降压药导致血压波动范围比较大的用户,会单独对此用户训练一个或者多个personal model,目的是拟合此用户的血压波动;训练此模型的前提是客户需要配合采集不同血压区间的多组血压数据;按照目前的情况,数据采集方法就是用户自己测量血压,然后通过客服记录血压值及采集时间;

4 预测结果由上述的模型进行模型融合决定,举例如下:

pred\_sbp\_bm, pred\_sbp\_sm, pred\_sbp\_pm 分别由base model, special\_model,personal\_model的预测结果,最终的sbp预测结果计算如下:

pred\_sbp = W1\*pred\_sbp\_bm + W2\*pred\_sbp\_sm + W3\*pred\_sbp\_pm

W1,W2,W3分别是每个模型对应的权重,取值范围是[0,1], 而且三者之和为1, 权重值的大小需要根据不同用户血压值,血压波动范围等做调整.

5 每种model可以是机器学习模型(如随机森林, 线性回归,GBDT等),或者是深度学习模型(DNN), 每种模型的训练数据都是采用清洗,过滤后的相同的数据;

后台会考虑多种不同模型的组合方式,以便于部署之后进行校验和测试;

四 部署

1 现有的血压模型预测保持不变, 后台数据库需要建一个新表,此表保存需要使用新血压模型的用户,当java端需要调用血压预测服务时,首先进行判断,此用户是否需要使用新模型,如果是的话,就调用新服务接口,否则还是调用旧血压模型的接口;

2 新模型的接口参数(输入请求, 返回结果)和旧模型一致,这样java端改动工作较少;