本次比赛最终使用了5个模型结果做average，分别做如下说明：

1、其中两个文件gbdt\_xiaoyu\_16000.ipynb、6gbdt\_modify.py结果都为gbdt模型预测结果，但是所作特征处理不同，最后的average结果能有很大提升。（注：只预测pay\_price大于0的用户，并且训练集也只包括pay\_price大于0的用户），两个文件分别对应线上rmse为56.47和57.00。

2、其中knn\_lasso.ipynb文件先对部分pay\_price为特定值的用户用knn模型计算对应pay\_price的预测值，如下所示的部分pay\_price值：

flist = [5.98, 6.97, 10.98, 16.96, 11.97, 9.99, 26.95, 21.95, 3.96, 11.96]，剩余的用户利用线性模型lasso计算。（注：只预测pay\_price大于0的用户，并且训练集也只包括pay\_price大于0的用户），线上对应rmse为57.82

3、其中nn.ipynb使用神经网络模型来进行预测，此时训练集不光包括pay\_price大于0的用户，对pay\_price等于0的用户先进行筛选，因为训练集存在噪声点，筛选完训练样本增多4000多个，然后进行神经网络10折交叉验证训练并用训练好的模型预测。（注：由于nn使用sgd梯度下降，每次预测结果都是不同的）。线上对应成绩为56.47.

4、其中lightgbm.ipynb文件使用lightgbm来对结果进行预测，这个模型是1个月前的结果，目前参数不能调到最优，跑出来的结果不好，暂使用之前最好的结果文件进行融合，代码也附带在压缩文件中。（注：只预测pay\_price大于0的用户），此模型运行时间很长，线上rmse 57.65。

这个比赛是为了预测高氪玩家，统计训练集发现pay\_price为0的用户在后45天付款的钱数占比非常小，所以5个模型预测的都是pay\_price不为0的用户，也是因为没有考虑pay\_price为0的用户，造成B榜变化很大，同时，为了保证训练集和测试集分布大致相同，会对测试集中pay\_price不为0的用户的prediction\_pay\_price的均值适当放缩，保证其和训练集的均值大致相同，这个trick非常管用。