**请结合附件Excel表中的数据完成下列问题：**

模型

1）以投资金额为目标变量建立一个线性模型。

2）以投资金额高低区分高投资与低投资用户，以此为目标变量建立逻辑回归模型。

3）以投资金额高低区分高投资与低投资用户，以此为目标变量建立一至两个机器学习模型（GBM, Random Forrest, Neural Network, SVM 等等）

**分析说明：**

**一、数据预处理（processing.py）**

1.人工删除了部分特征：

手机省份,注册时间,用户注册终端,用户注册渠道,最近一次登录省份,最近一次登录终端,最近一次登录城市,最近一次登录ip,最近一次登录时间,是否开通托管,开通托管日期,首次充值日期,首投时间,是否注册7天内投资,是否托管7天内充值,是否托管7天内投资,是否充值7天内投资，投资等级

部分依据：如首投时间可以由首投距今时间（天）来体现，故删除；再如，投资等级和投资金额成等比关系，显然无意义，故删除

2.根据手机型号的品牌出现频数，通过字符串匹配将手机分为：Xiaomi,iPhone,Samsung,OPPO,iPad,HUAWEI和其他手机

3.将年龄按照0-14、15-35、36-60、61-∞进行划分

4.将城市按照超一线，一线，二线，三线，四线，五线进行划分

5.对所有非连续值特征进行onehot编码

6.预处理数据导出为processed data.txt

**二、线性模型（线性模型.py）**

1.数据标准化

2.将数据集按4:1划分为训练集与测试集

3.基于梯度提升树GBDT的Tree importance和SBS方法筛选特征

4.尝试了支持向量机SVR，线性回归和GBDT回归算法，均无法得到有效模型

**二、分类模型（二分类模型.py）**

1.数据标准化

2.将数据集按4:1划分为训练集与测试集

3.以100000为界（权衡了中位数与数量级），给高投资与低投资用户打标签

4.基于梯度提升树GBDT的Tree importance和SBS方法筛选特征

5.通过网格搜索、交叉验证，建立了逻辑回归，梯度提升树和SVM模型，准确率分别为68.7%、68.3%和67.7%