数据清洗预测实操练习题

**一、问题描述：**

本数据集是某在线购物网站真实购物数据，包括用户的浏览历史和购物记录。数据集分为训练集ads\_train.csv，和测试集ads\_test.csv。对原始数据进行数据清洗等预处理后，使用训练集ads\_train.csv对用户是否在网站购物建立预测模型，然后在测试集上预测用户是否购物以及购物概率。

**二、数据字段说明：**

isbuyer–是否曾在本网站购物

buy\_freq–购买频率

visit\_freq–访问频率

buy\_interval–两次购物的时间间隔

sv\_interval – 两次访问网站的时间间隔

expected\_time\_buy–预期购买时间

expected\_time\_visit–预期访问时间

last\_buy - 距离上次购买的天数

last\_visit- 距离上次访问网站的天数

multiple\_buy- 是否多次购买

multiple\_visit- 是否多次访问

uniq\_url - 浏览器登录时使用过的IP数

num\_checkins - 浏览器登录次数。

*y\_buy–目标变量，关注期内是否有购物行为，0：没有；1：有*

**三、任务描述：**

1. 数据观察. 对各字段进行统计，最大值、最小值、均值、方差、标准差、缺失率等。请将数据观察的结果存为：ads\_train\_observation.csv
2. 数据清洗。根据观察结果对数据进行处理，包括：缺失值处理，如填充、插值等；异常值处理，如均值修正、删除等。请将如何进行数据清洗的、以及原因整理成说明文档。
3. 数据处理。根据现有字段，设计并提取新的特征。对于正样本过少的原始数据，采取合适的方法，如过采样/欠采样的方法，生成训练数据，并说明处理方法。处理后的数据命名为ads\_train\_sorted.csv。
4. 以y\_buy为目标，使用至少一种不同的预测模型（如逻辑回归、随机森林、SVM等任何分类器），对y\_buy建立分类模型，预测用户未来是购买者，还是不购买者。解释您对模型的选择、如何进行模型选择、如何验证模型的质量以及哪些特征是最有效的特征。
5. 使用您选择的分类模型对测试数据集ads\_test.csv进行预测，并报告未来购买的预测概率（y\_buy\_prob）和未来购买的预测标签(y\_buy)。预测结果请存为： ads\_test\_predicted.csv

**四、提交方式：**

请将代码、数据观察文档、数据清洗说明文档、数据处理结果文件、预测集上的预测结果，压缩为“电商数据清洗预测实操-姓名-日期.zip”，于4月26日前，提交至部门钉钉群：”电商数据清洗预测实操“目录下。