机器学习大作业补充说明

1 数据介绍

数据集中的特征数据分为以下两类：

连续型 (Numerical):Age, Number of nights/visits in CITY, Total expenditures …

类别型 (Categorical):Nationality, Immigration airport, Companion, Transportation, Activity, Destination, Activity_Destination, Trigger, Attractions …

类别型特征数据对应的实际内容在"FeatureDescription.xlsx" 中的各工作表中给出。部分类别型特征把所有选项列出，并通过数字表示各选项的重要程度。

本数据集来源于真实的未经处理的调查问卷结果，部分缺失值是调查统计流程中的缺失导致的。请根据实际情况结合课程内容自行处理异常数据，并分析数据处理方法对实验结果的影响。

2 滚动预测策略

滚动预测策略的目的是将数据划分为多个训练集和测试集，其中每次的训练集包含了当前时间点之前的数据，而测试集则包含了之后的数据，这样能够有效模拟实际的预测场景，即利用过去的数据预测未来的趋势。

请使用 `sklearn.model_selection.TimeSeriesSplit`（$n\_splits = 5$）来进行时间序列的交叉验证，根据第一列数据变量 **Survey date** 对数据进行切片。