**泰迪杯B题**

数据分析+建立模型进行预测

数据集中的销售区域编码：表示华中、华南等的一个销售区域

如果是新品的话要根据同品类情况来进行预测

如果是预测月或周的销售量的话，则需要先提取出来数据

过程：

1. 数据读取（B题文件是CSV格式）
2. 数据预处理
3. 数据合并（merge、contact函数）

有的数据集是多个表格组成，需要先进行合并处理

B题不用

1. 数据清洗（前三个需要先检测）
2. 缺失值处理（删除法、填补法（均值插补、同类均值插补、建模预测、高维映射、多重插补、极大似然估计、压缩感知和矩阵补全）） 根据任务要求选择 也可选择多种插值看最后哪种模型效果好
3. 异常值处理（一是异常值检测：3σ原则、简单统计分析、箱线图（常用）；二是异常值处理：删除、视为缺失值、平均值修正）
4. 重复值处理（一般删除处理） （查找重复值是看按照一整行一样的数据进行删除，还是根据某一列有重复的数据进行删除，要看数据情况）
5. 时间数据处理（转换为时间戳格式，方便后续提取节假日、工作日这部分数据作准备）

3.数据可视化分析（看题目要求分析什么内容）

4.特征

（1）特征构建（0到有的过程）

当数据拿到手后，首先需要从现有数据中挑选或将现有数据进行变形，组合形成新特征，此过程即为特征构建

（2）特征抽取（降维处理）

当特征维度比较高，通过映射或变化的方式，用低维空间样本来表示样本，称为特征抽取

（3）特征选择（并不一定数据集的所有特征都要用上） （一般会用相关性系数来进行特征选择）

从一组特征中挑选出一些最有效的特征，以达到降低纬度和降低过拟合风险的目的，称为特征选择

5.数据标准化（量纲存在差异）

在回归预测中，标准化是为了让特征值具有均等权重

在神经网络中，通过数据标准化能够加速权重参数的收敛

在主成分分析中，进行标准化处理就是默认指标间权重相等，不考虑指标间的差异和相互影响

（1）min-max标准化（归一化）

（2）z-score标准化（规范化）

（3）小数定标标准化

6.模型构建（训练集）

（1）机器学习（常用sklearn库调用）

（2）时间序列模型（数据有很强的时序性时才考虑）

（3）深度学习

7.模型预测与评估（预测集）

看是分类问题还是回归问题

B题是回归问题

**从竞赛到论文编写**

**不准泄露个人信息、学校信息等**

**没有模板（按正常论文来写）**

**查重（泰迪自有库查重）**











