2.2.x知识点总结

- **数据加载**
  同1.1.x,2.1.x
- **数据预处理和数据集划分**
  - **数据预处理**
    - **将分类变量转变成数值变量**
      pd.get_dummies(待转换分类变量列)
    - **使用匿名函数处理**
      y = data['待转换列'].apply(lambda x: int(x.split(' ')[0]))
  - **数据集划分**
    同2.1.x
- **模型训练和测试**
  - **训练模型**
    - **逻辑回归模型**
      LogisticRegression(max_iter=最大迭代次数).fit(特征变量训练集，目标变量训练集)
    - **线性回归模型**
      LinearRegression().fit(特征变量训练集，目标变量训练集)
    - **随机森林回归模型**
      RandomForestRegressor(n_estimators=决策树数量, random_state=42).fit(特征变量训练集，目标变量训练集)
    - **XGBoost回归模型**
      xgboost.XGBRegressor(n_estimators=决策树数量,learning_rate=学习率,max_depth=每棵树最大深度, random_state=42).fit(特征变量训练集，目标变量训练集)
    - **决策树回归模型**
      DecisionTreeRegressor(random_state=42).fit(特征变量训练集, 目标变量训练集)
    - **使用管道串联数据预处理和模型训练**
      pipeline = Pipeline([('scaler', 数据预处理方法), ('linreg', 模型训练方法)])

      这个时候使用pipline.fit(特征变量训练集，目标变量训练集) 会自动先将训练集数据使用数据预处理方法处理好之后再进行模型训练
  - **保存模型为文件**
    - **使用pickle保存**
      with open('模型文件名','wb') as file:
      　　pickle.dump(训练好的模型, file)
    - **使用joblib保存**
      joblib.dump(训练好的模型, 模型文件名称)
  - **测试模型预测**
    目标变量预测集 = 模型对象.predict(特征变量测试集)
  - **使用模型的测试工具测试数据集**
    训练集得分 = 模型对象.score(特征变量训练集,目标变量训练集)

    测试集得分 = 模型对象.score(特征变量测试集,目标变量测试集)
- **分析模型预测结果**
  - **计算预测准确率**
    准确率 = (目标变量测试集==目标变量预测集).mean()
  - **计算预测结果均方误差**
    均方误差 = mean_squared_error(目标变量测试集,目标变量预测集)
  - **计算预测结果决定系数**
    决定系数 =r2_score(目标变量测试集,目标变量预测集)
- **调整数据集和模型**
  - **数据重采样**
    特征变量重采样训练集，目标变量重采样训练集 = 重采样方法(smote).fit_resample(特征变量训练集，目标变量训练集)
  - **换模型**
- **模型性能分析**
  - **混淆矩阵**
    - **TP真正例**
      模型预测为真，实际结果为真
      - **预测真**且预测**对了**
    - **TN真负例**
      模型预测为假，实际结果为假
      - **预测假**且预测**对了**
    - **FN假负例**
      模型预测为假，实际结果为真
      - **预测假**但预测**错了，即漏报**
    - **FP假正例**
      模型预测为真，实际结果为假
      - **预测真**但预测**错了，即误报**
  - **评价指标**
    - **精确率**
      TP/(TP+FP)
      - 所有**预测为真**的里面预测**对了多少**，误报越少值越高
    - **召回率**
      TP/(TP+FN)
      - 所有**实际为真**的里面预测**对了多少**，漏报越少值越高
    - **F1-Score**
      2 * (Precision * Recall) / (Precision + Recall)
      - 精确率和召回率的**调和平均**，分数越高，模型在精确率和召回率的表现就越**平衡**
    - **Support**
      测试集中正负样本的实际样本数量
    - **均方误差(MSE)（平均绝对误差同理）**
      0 到 +∞，越接近0说明模型性能越好
    - **决定系数(R^2 Score)**
      -∞到1，越接近1说明模型性能越好，0为均值预测性能分界，负数表示不如均值预测的性能