实战案例4：员工离职预测

作者： Robin
日期： 2018/09
数据集来源： kaggle

1. 案例描述

该项目的目的主要是基于员工的各项信息预测该员工是否会离职。

2. 数据集描述

- Kaggle提供的数据集包含40698条员工记录，每条记录包含31项信息。
- 数据字典
  - **Age**: 年龄，整型
  - **Attrition**: 是否离职，整型，1表示已经离职，2表示未离职，这是目标预测值
  - **BusinessTravel**: 商务差旅频率，字符串，Non-Travel表示不出差，Travel_Rarely表示不经常出差，Travel_Frequently表示经常出差
  - **Department**: 员工所在部门，字符串，Sales表示销售部，Research & Development表示研发部，Human Resources表示人力资源部
  - **DistanceFromHome**: 公司跟家庭住址的距离，整型，从1到29，1表示最近，29表示最远
  - **Education**: 员工的教育程度，整型，从1到5，5表示教育程度最高
  - **EducationField**: 员工所学习的专业领域，字符串，Life Sciences表示生命科学，Medical表示医疗，Marketing表示市场营销，Technical Degree表示技术学位，Human Resources表示人力资源，Other表示其他
  - **EmployeeNumber**: 员工编号，整型
  - **EnvironmentSatisfaction**: 员工对于工作环境的满意程度，整型，从1到4，1的满意程度最低，4的满意程度最高
  - **Gender**: 性别，字符串，Male表示男性，Female表示女性
  - **JobInvolvement**: 员工工作投入度，整型，从1到4，1为投入度最低，4为投入度最高
  - **JobLevel**: 职业级别，整型，从1到5，1为最低级别，5为最高级别
  - **JobRole**: 工作角色，字符串，Sales Executive是销售主管，Research Scientist是科学研究员，Laboratory Technician实验室技术员，Manufacturing Director是制造总监，Healthcare Representative是医疗代表，Manager是经理，Sales Representative是销售代表，Research Director是研究总监，Human Resources是人力资源
  - **JobSatisfaction**: 工作满意程度，整型，从1到4，1代表满意程度最低，4代表满意程度最高
  - **MaritalStatus**: 婚姻状态，字符串，Single代表单身，Married代表已婚，Divorced代表离婚
  - **MonthlyIncome**: 月收入，整型，范围在1009到19999之间
  - **NumCompaniesWorked**: 员工曾经工作过的公司数，整型
  - **Over18**: 年龄是否超过18岁，字符串
  - **OverTime**: 是否加班，字符串，Yes表示加班，No表示不加班
  - **PercentSalaryHike**: 涨薪百分比，整型
  - **PerformanceRating**: 绩效评估，整型

- **RelationshipSatisfaction**: 关系满意度，整型，从1到4，1表示满意度最低，4表示满意度最高
- **StandardHours**: 标准工作时间，整型
- **StockOptionLevel**: 股票期权水平，整型
- **TotalWorkingYears**: 总工龄，整型
- **TrainingTimesLastYear**: 上一年的培训时长，整型，从0到6，0表示没有培训，6表示培训时间最长
- **WorkLifeBalance**: 工作生活平衡程度，整型，从1到4，1表示平衡程度最低，4表示平衡程度最高
- **YearsAtCompany**: 在目前公司工作年数，整型
- **YearsInCurrentRole**: 在目前工作职责的工作年数，整型
- **YearsSinceLastPromotion**: 距离上次升职时长，整型
- **YearsWithCurrManager**: 跟目前的管理者共事年数，整型

3. 任务描述

- 使用scikit-learn建立不同的机器学习模型进行员工离职预测

4. 主要代码解释

- 代码结构

```
lect04_proj
├── data
    ├── employee.csv    # CSV数据文件
├── output
    ├── model_comparison.csv    # 模型比较结果的CSV文件(程序的输出)
    ├── pred_results.png    # 模型比较结果的png文件(程序的输出)
├── config.py     # 配置文件
├── utils.py     # 工具类文件
├── main.py      # 主程序
├── lect04_proj_readme.pdf  # 案例讲解文档
```

- **main.py**

使用的模型及相关参数配置。该项目中使用了4个机器学习模型，并为不同的学习模型指定了参数空间。 如：kNN，指定了3个k值用于比较对结果的影响：5, 11, 15。

```
def main():
    ...
    model_name_param_dict = {'kNN':     [5, 11, 15],
                             'LR':      [0.01, 1, 100],
                             'DT':      [5, 10, 15],
                             'SVM':     [0.01, 1, 100]}
    ...
```

- **utils.py**

将类别型数据转换为“哑变量”

```
def process_data(raw_data):
    ...
    # 对类别型特征进行编码
    enc_feat_df = pd.get_dummies(feat_df, columns=config.cat_cols)
    ...
```

5. 案例总结

- 该项目通过员工是否离职，巩固并实践了使用scikit-learn搭建简单的预测模型:
    - 机器学习流程
    - scikit-learn中常用预测模型的使用，如：kNN, logistic regression，decision tree及SVM
    - 参数的选择

6. 课后练习

- 参考随堂代码，试着使用不同的参数空间观察对结果的影响。
- 思考：对于非平衡数据集，使用“准确率”是否为合适的评价指标？

参考资料

1. scikit-learn官方教程
2. 通过scikit-learn理解机器学习