

# 实战案例3-1：贷款违约预测 (1)

作者：Robin

日期：2018/05

数据集来源：[kaggle](#)

声明：[小象学院](#)拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利

## 1. 案例描述

该项目的目的主要是基于贷款记录的各项信息预测该笔贷款是否有风险。

## 2. 数据集描述

- Kaggle[提供的数据集](#)包含1000条贷款记录，每条记录包含9项信息。
- 数据字典
  - **Age:** 年龄，整型
  - **Sex:** 性别，字符串（male, female）
  - **Job:** 工作等级，整型，0-4表示从事“非技能”到“技能”的工种，值越大表示从事越高的技能工种
  - **Housing:** 住房状况，字符串，free: 无住房，rent: 租房，own: 拥有住房
  - **Saving accounts:** 储蓄账户，字符串，little: 很少, moderate: 中等, quite rich: 较多, rich: 很多
  - **Checking account:** 支票账户，字符串，little: 很少, moderate: 中等, rich: 很多
  - **Credit amount:** 信用卡可用额度，整型
  - **Duration:** 贷款期限，整型，单位（月）
  - **Purpose:** 贷款目的，字符串
  - **Risk:** 贷款质量（是否违约），字符串，good: 没有违约，bad: 违约

## 3. 任务描述

- 使用scikit-learn建立不同的机器学习模型进行贷款违约预测

## 4. 主要代码解释

- 代码结构

```
lect03_proj
├── data
│   ├── german_credit_data.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
└── lect03_proj_readme.pdf  # 案例讲解文档
```

- **utils.py**

类别型数据转换为整型数字，使用到了Pandas的map()函数，将对应的映射字典作为参数传入函数中。如：

```
def transform_data(data_df):
    ...
    trans_data_df['Sex'] = data_df['Sex'].map(config.sex_dict)
    ...
```

- **main.py**

使用的模型及相关参数配置。该项目中使用了2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]}
    ...
```

- **main.py**

选择最优模型，np.argmax(list)返回list中最大值的索引号，如np.argmax([5, 10, 2])的返回值是1，因为第1个位置上的值最大，为10。之后根据该索引号获取最优的准确率及模型。

```
def train_test_model(X_train, y_train, X_test, y_test, param_range, model_name='SVM'):
    ...
    # 记录最优模型
    best_idx = np.argmax(scores)
    best_acc = scores[best_idx]
    best_model = models[best_idx]
    ...
```

## 5. 案例总结

- 该项目通过预测贷款是否可能违约，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等
  - 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。

## 参考资料

1. [scikit-learn官方教程](#)

## 2. [通过scikit-learn理解机器学习](#)