**使用随机森林预测德国人信贷风险**

**1. 实验目的**

* 熟练掌握使用随机森林构建分类模型
* 熟练掌握随机森林模型的调参过程

**2. 实验内容**

银行在给客户贷款之前，会提前预测贷款者信用风险，以决定是否贷款给他们。

本实验基于德国信贷风险数据，使用随机森林模型预测信贷风险。预测结果为有风险/无风险。

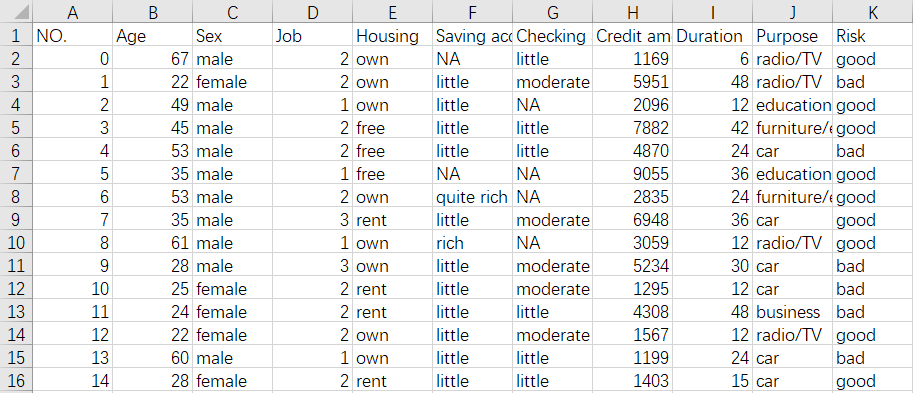
**3. 实验数据**

**数据集位置**

数据集存放在/home/dataset目录下。数据集名字是 german\_credit\_data.csv

**数据集格式**

德国人信贷风险数据集包含 1000 条德国人贷款数据。



**数据字段解释**

* NO.：数据序号
* Age：数值类型；年龄
* Sex：字符串类型；性别；有2个值，分别为male（男性）和female（女性）
* Job：数值类型；工作状态；有4个值，分别为0、1、2、3；0：unskilled and non-resident（无特长且非本地居民），1：unskilled and + + resident（无特长且是本地居民），2：skilled（有特长），3：highly skilled（技艺高超）
* Housing：字符串类型；有3个值，分别为own（有房子），rent（租房子），free（自己没房子也不租房，比如政府的救济房、比如住在别人家里）
* Saving accounts：字符串类型；活期储蓄账户存款多少（有固定利率）；有4个值，分别是little（几乎没有），moderate（中等），rich（富有），quite rich（非常富有）；如果有缺失数据，很大可能是这个人没有银行账户
* Checking account：字符串类型；普通支票账户存款多少（一般无息）；有3个值，分别是little（几乎没有），moderate（中等），rich（富有）；如果有缺失数据，很大可能是这个人没有银行账户
* Credit amount：数值类型；信贷额度（单位DM，即马克）
* Duration：数值类型；信贷时间（单位月）
* Purpose：字符串类型；贷款目的；有8个值，分别是car（买车），furniture/equipment（家具/装备），radio/TV（收音机/电视），+ + + + domestic appliances（家用电器），repairs（维修），education（教育），business（做生意），vacation/others（度假/其它）
* Risk：字符串类型；信用风险等级；有2个值，分别为good（无风险），bad（有风险）

**4. 实验知识点**

* 随机森林模型做分类
* 变量重要性排名

**5. 实验时长**

1学时

**6. 实验环境**

* Linux Ubuntu 操作系统
* Jupyter 代码编辑器
* Python 3.6.9
* numpy 1.18.5
* pandas 1.1.5
* scikit-learn 0.24.2

**7. 实验分析**

1. 导入包、导入数据、做数据预处理
2. 构建随机森林模型
3. 使用不同参数，比较随机森林模型性能
4. 得到变量重要性排名

**8. 实验过程**

In [1]:

*# -\*- coding: utf-8 -\*-*

*"""*

*使用随机森林预测德国人信贷风险*

*"""*

Out[1]:

'\n使用随机森林预测德国人信贷风险\n'

**8.1 导入包**

In [2]:

*# 导入包*

**import** **numpy** **as** **np**

**import** **pandas** **as** **pd**

**8.2 导入数据集**

In [3]:

*# 导入数据集*

data = pd.read\_csv("/home/dataset/german\_credit\_data.csv")

data

Out[3]:

|  | **NO.** | **Age** | **Sex** | **Job** | **Housing** | **Saving accounts** | **Checking account** | **Credit amount** | **Duration** | **Purpose** | **Risk** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 0 | 67 | male | 2 | own | NaN | little | 1169 | 6 | radio/TV | good |
| **1** | 1 | 22 | female | 2 | own | little | moderate | 5951 | 48 | radio/TV | bad |
| **2** | 2 | 49 | male | 1 | own | little | NaN | 2096 | 12 | education | good |
| **3** | 3 | 45 | male | 2 | free | little | little | 7882 | 42 | furniture/equipment | good |
| **4** | 4 | 53 | male | 2 | free | little | little | 4870 | 24 | car | bad |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **995** | 995 | 31 | female | 1 | own | little | NaN | 1736 | 12 | furniture/equipment | good |
| **996** | 996 | 40 | male | 3 | own | little | little | 3857 | 30 | car | good |
| **997** | 997 | 38 | male | 2 | own | little | NaN | 804 | 12 | radio/TV | good |
| **998** | 998 | 23 | male | 2 | free | little | little | 1845 | 45 | radio/TV | bad |
| **999** | 999 | 27 | male | 2 | own | moderate | moderate | 4576 | 45 | car | good |

1000 rows × 11 columns

去掉NO.列，因为该列无用处。

In [4]:

data = data.drop(['NO.'], axis = 1)

data

Out[4]:

|  | **Age** | **Sex** | **Job** | **Housing** | **Saving accounts** | **Checking account** | **Credit amount** | **Duration** | **Purpose** | **Risk** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 67 | male | 2 | own | NaN | little | 1169 | 6 | radio/TV | good |
| **1** | 22 | female | 2 | own | little | moderate | 5951 | 48 | radio/TV | bad |
| **2** | 49 | male | 1 | own | little | NaN | 2096 | 12 | education | good |
| **3** | 45 | male | 2 | free | little | little | 7882 | 42 | furniture/equipment | good |
| **4** | 53 | male | 2 | free | little | little | 4870 | 24 | car | bad |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **995** | 31 | female | 1 | own | little | NaN | 1736 | 12 | furniture/equipment | good |
| **996** | 40 | male | 3 | own | little | little | 3857 | 30 | car | good |
| **997** | 38 | male | 2 | own | little | NaN | 804 | 12 | radio/TV | good |
| **998** | 23 | male | 2 | free | little | little | 1845 | 45 | radio/TV | bad |
| **999** | 27 | male | 2 | own | moderate | moderate | 4576 | 45 | car | good |

1000 rows × 10 columns

**8.3 数据预处理**

**8.3.1 检测并处理缺失值**

In [5]:

*# 检测缺失值*

null\_df = data.isnull().sum() *# 检测缺失值*

null\_df

Out[5]:

Age 0

Sex 0

Job 0

Housing 0

Saving accounts 183

Checking account 394

Credit amount 0

Duration 0

Purpose 0

Risk 0

dtype: int64

Saving accounts和Checking account 这2个字段有缺失值，需要填补

In [6]:

*# 处理Saving accounts 和 Checking account 这2个字段*

**for** col **in** ['Saving accounts', 'Checking account']: *# 处理缺失值*

data[col].fillna('none', inplace=**True**) *# none说明这些人没有银行账户*

填补完毕后，需要再次检查缺失值，确保缺失值缺失填补了

In [7]:

*# 检测缺失值*

null\_df = data.isnull().sum()

null\_df

Out[7]:

Age 0

Sex 0

Job 0

Housing 0

Saving accounts 0

Checking account 0

Credit amount 0

Duration 0

Purpose 0

Risk 0

dtype: int64

由输出结果得出结论，无字段有缺失值

**8.3.2 处理因变量**

对因变量做字符编码。

In [8]:

data['Risk']

**from** **sklearn.preprocessing** **import** LabelEncoder

le = LabelEncoder()

data['Risk'] = le.fit\_transform(data['Risk'])

data['Risk']

Out[8]:

0 1

1 0

2 1

3 1

4 0

..

995 1

996 1

997 1

998 0

999 1

Name: Risk, Length: 1000, dtype: int64

In [9]:

le.classes\_

Out[9]:

array(['bad', 'good'], dtype=object)

bad被转换为了0，good被转换为了1。这和Risk字段的含义不符合。应该把bad转换为1，把good转换为0。

In [10]:

data['Risk'] = data['Risk'] - 1

data.loc[data['Risk']==-1, 'Risk'] = 1

data['Risk']

Out[10]:

0 0

1 1

2 0

3 0

4 1

..

995 0

996 0

997 0

998 1

999 0

Name: Risk, Length: 1000, dtype: int64

此时bad被转换为1，good被转换为0。符合要求。

**8.3.3 处理类别型变量**

In [11]:

*# 查看所有字段的类型*

print(data.dtypes)

Age int64

Sex object

Job int64

Housing object

Saving accounts object

Checking account object

Credit amount int64

Duration int64

Purpose object

Risk int64

dtype: object

结合输出的字段类型结果和字段解释，只有Job字段的类型有误。Job字段的类型应该是类别型，在数据集中却是整型。

In [12]:

*# 处理Job字段的类型*

data['Job'] = data['Job'].astype('object')

In [13]:

data[data['Sex']==1]

Out[13]:

|  | **Age** | **Sex** | **Job** | **Housing** | **Saving accounts** | **Checking account** | **Credit amount** | **Duration** | **Purpose** | **Risk** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

字段类型无误后，处理类别型变量。

In [14]:

*# 处理类别型变量*

data = pd.get\_dummies(data, drop\_first = **True**)

data

Out[14]:

|  | **Age** | **Credit amount** | **Duration** | **Risk** | **Sex\_male** | **Job\_1** | **Job\_2** | **Job\_3** | **Housing\_own** | **Housing\_rent** | **...** | **Checking account\_moderate** | **Checking account\_none** | **Checking account\_rich** | **Purpose\_car** | **Purpose\_domestic appliances** | **Purpose\_education** | **Purpose\_furniture/equipment** | **Purpose\_radio/TV** | **Purpose\_repairs** | **Purpose\_vacation/others** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 67 | 1169 | 6 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **1** | 22 | 5951 | 48 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | ... | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **2** | 49 | 2096 | 12 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | ... | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **3** | 45 | 7882 | 42 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **4** | 53 | 4870 | 24 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **995** | 31 | 1736 | 12 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **996** | 40 | 3857 | 30 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | ... | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **997** | 38 | 804 | 12 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | ... | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **998** | 23 | 1845 | 45 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | ... | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| **999** | 27 | 4576 | 45 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | ... | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |

1000 rows × 24 columns

**8.3.4 得到自变量和因变量**

In [15]:

*# 得到自变量和因变量*

y = data['Risk'].values

data = data.drop(['Risk'], axis = 1)

x = data.values

**8.3.5 拆分训练集和测试集**

In [16]:

*# 拆分训练集和测试集*

**from** **sklearn.model\_selection** **import** train\_test\_split

x\_train, x\_test, y\_train, y\_test = train\_test\_split(x, y, test\_size = 0.2, random\_state = 1)

print(x\_train.shape)

print(x\_test.shape)

print(y\_train.shape)

print(y\_test.shape)

(800, 23)

(200, 23)

(800,)

(200,)

**8.3.6 特征缩放**

In [17]:

*# 特征缩放*

**from** **sklearn.preprocessing** **import** StandardScaler

sc\_x = StandardScaler()

x\_train = sc\_x.fit\_transform(x\_train)

x\_test = sc\_x.transform(x\_test)

**8.4 使用不同的参数构建随机森林模型**

**8.4.1 模型1**

参数：max\_depth=9, max\_features='auto', min\_samples\_leaf=5, n\_estimators=50

**8.4.1.1 构建模型**

In [18]:

*# 使用不同的参数构建随机森林模型*

*# 模型1：构建随机森林模型*

**from** **sklearn.ensemble** **import** RandomForestClassifier

classifier = RandomForestClassifier(max\_depth=9, max\_features='auto', min\_samples\_leaf=5, n\_estimators=50, random\_state = 1)

classifier.fit(x\_train, y\_train)

Out[18]:

RandomForestClassifier(max\_depth=9, min\_samples\_leaf=5, n\_estimators=50,

random\_state=1)

**8.4.1.2 测试集做预测**

In [19]:

*# 在测试集做预测*

y\_pred = classifier.predict(x\_test)

**8.4.1.3 评估模型性能**

In [20]:

*# 评估模型性能*

**from** **sklearn.metrics** **import** accuracy\_score, confusion\_matrix, classification\_report

print(accuracy\_score(y\_test, y\_pred))

0.75

准确率是0.75，比较低，表明模型性能比较差。

In [21]:

print(confusion\_matrix(y\_test, y\_pred))

[[131 10]

[ 40 19]]

由混淆矩阵可知，真实值是0（实际没风险），预测值也是0（预测没风险），这样的样本有19条；真实是0，预测是1，这样的样本有40条；真实值是1，预测是0，这样的样本有10条；真实值是1，预测值也是1，这样的样本有131条。

In [22]:

print(classification\_report(y\_test, y\_pred))

precision recall f1-score support

0 0.77 0.93 0.84 141

1 0.66 0.32 0.43 59

accuracy 0.75 200

macro avg 0.71 0.63 0.64 200

weighted avg 0.73 0.75 0.72 200

1. 根据业务场景，银行想尽可能预测出来有风险的人，以避免贷出去的钱收不回来，所以Recall指标更加重要。
2. Recall指标的weighted avg只有0.75，相对比较低。

**8.4.1.4 变量重要性排名**

In [23]:

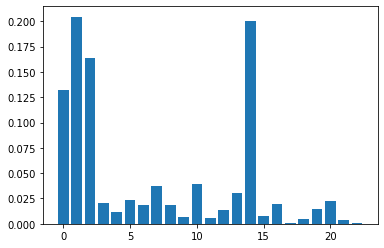
*# 得到变量重要性排名*

importance = classifier.feature\_importances\_

**import** **matplotlib.pyplot** **as** **plt**

plt.bar([x **for** x **in** range(len(importance))], importance)

plt.show()



通过变量重要性的柱状图可见，第0、1、2、3、15个自变量对因变量的影响较大。可以考虑做特征选择，进一步提升模型性能。特征选择在后面的章节会讲到。

查看第0、1、2、3、15个自变量的名字

In [24]:

print(data.columns[0])

print(data.columns[1])

print(data.columns[2])

print(data.columns[3])

print(data.columns[15])

Age

Credit amount

Duration

Sex\_male

Checking account\_rich

**8.4.2 模型2**

参数：max\_depth=3, max\_features='auto', min\_samples\_leaf=50, n\_estimators=100

In [25]:

*# 模型2：构建随机森林模型*

classifier = RandomForestClassifier(max\_depth=3, max\_features='auto', min\_samples\_leaf=50, n\_estimators=100, random\_state = 1)

classifier.fit(x\_train, y\_train)

Out[25]:

RandomForestClassifier(max\_depth=3, min\_samples\_leaf=50, random\_state=1)

In [26]:

*# 在测试集做预测*

y\_pred = classifier.predict(x\_test)

In [27]:

*# 评估模型性能*

print(accuracy\_score(y\_test, y\_pred))

0.705

In [28]:

print(confusion\_matrix(y\_test, y\_pred))

[[141 0]

[ 59 0]]

In [29]:

print(classification\_report(y\_test, y\_pred))

precision recall f1-score support

0 0.70 1.00 0.83 141

1 0.00 0.00 0.00 59

accuracy 0.70 200

macro avg 0.35 0.50 0.41 200

weighted avg 0.50 0.70 0.58 200

/root/miniconda3/lib/python3.9/site-packages/sklearn/metrics/\_classification.py:1248: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/root/miniconda3/lib/python3.9/site-packages/sklearn/metrics/\_classification.py:1248: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

/root/miniconda3/lib/python3.9/site-packages/sklearn/metrics/\_classification.py:1248: UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero\_division` parameter to control this behavior.

\_warn\_prf(average, modifier, msg\_start, len(result))

由分类报告可见，recall的值降低了。

In [30]:

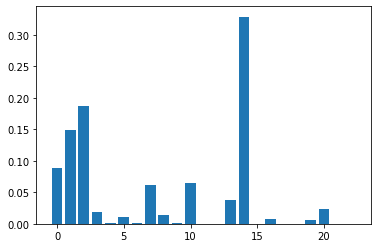
*# 得到变量重要性排名*

importance = classifier.feature\_importances\_

**import** **matplotlib.pyplot** **as** **plt**

plt.bar([x **for** x **in** range(len(importance))], importance)

plt.show()



由上图可见，变量重要性排名也发生了变化。

**8.4.3 模型3**

参数：max\_depth=9, max\_features=15, min\_samples\_leaf=5, n\_estimators=25

In [31]:

*# 模型3：构建随机森林模型*

classifier = RandomForestClassifier(max\_depth=9, max\_features=15, min\_samples\_leaf=5, n\_estimators=25, random\_state = 1)

classifier.fit(x\_train, y\_train)

Out[31]:

RandomForestClassifier(max\_depth=9, max\_features=15, min\_samples\_leaf=5,

n\_estimators=25, random\_state=1)

In [32]:

*# 在测试集做预测*

y\_pred = classifier.predict(x\_test)

In [33]:

*# 评估模型性能*

print(accuracy\_score(y\_test, y\_pred))

0.74

In [34]:

print(confusion\_matrix(y\_test, y\_pred))

[[124 17]

[ 35 24]]

In [35]:

print(classification\_report(y\_test, y\_pred))

precision recall f1-score support

0 0.78 0.88 0.83 141

1 0.59 0.41 0.48 59

accuracy 0.74 200

macro avg 0.68 0.64 0.65 200

weighted avg 0.72 0.74 0.72 200

In [36]:

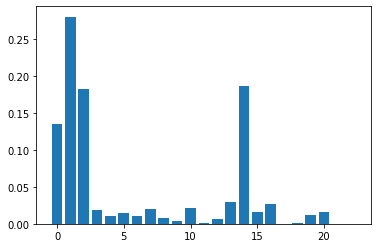
*# 得到变量重要性排名*

importance = classifier.feature\_importances\_

**import** **matplotlib.pyplot** **as** **plt**

plt.bar([x **for** x **in** range(len(importance))], importance)

plt.show()



**9. 实验结果（结论）**

1. 不同超参数对模型性能的影响不同
2. 通过随机森林模型可以得到变量重要性排名