04-main

2023 年 5 月 21 日

1 住户信息预测房屋是否屋主所有案例

地产公司在做房屋的租售业务之余，也进行住户与房屋相关数据的调查，在数据库中，存在如图所示的调研数据。

1. 完成数据集读取；
2. 数据预处理（删除缺失值）；
3. onehot 编码；
   - onehot 特征列
   - 构建独热编码器
   - 训练独热编码器，得到转换规则
   - 独热数据转换
   - 构建数值特征列
   - 合并独热特征与数值特征
4. 构建逻辑回归模型并训练；
5. 完成 K 折交叉检验
6. 完成模型预测。

在 Github 中查看

[29]:
```python
import pandas as pd
```

[30]:
```python
data = pd.read_csv('使用住户信息预测房屋是否屋主所有.csv')
data
```

[30]:
```
       Age     Education Level Gender Home Ownership Internet Connection
0     33.0           Doctorate   Male            Own             Dial-Up  \
1     47.0           Doctorate   Male            Own                 DSL
2      NaN           Doctorate   Male            Own                 DSL
3     35.0   Bachelor's Degree   Male            Own         Cable Modem
```

```
4     32.0   Bachelor's Degree   Male            Own        Cable Modem
…    …                 …    …            …                 …
3182  27.0     Master's Degree   Male           Rent                DSL
3183  45.0  Associate's Degree   Male            Own            Dial-Up
3184  38.0     Master's Degree   Male            Own               IDSN
3185  31.0     Master's Degree   Male           Rent               IDSN
3186  39.0     Master's Degree   Male           Rent            Dial-Up

     Marital Status  Movie Selector  Num Bathrooms  Num Bedrooms  Num Cars
0           Married  Spouse/Partner            2.5             3       1.0  \
1           Married  Spouse/Partner            2.0             2       2.0
2           Married  Spouse/Partner            2.5             4       2.0
3           Married              Me            2.5             4       2.0
4           Married              Me            3.5             5       2.0
…             …             …            …           …       …
3182  Never Married              Me            1.0             1       2.0
3183        Married  Spouse/Partner            1.0             1       2.0
3184        Married              Me            1.5             3       2.0
3185        Married              Me            1.0             1       2.0
3186  Never Married              Me            1.0             1       1.0

      …  Num TVs  PPV Freq Prerec Buying Freq Prerec Format
0     …      2.0    Rarely            Monthly           DVD  \
1     …      1.0     Never            Monthly           DVD
2     …      2.0     Never             Rarely           DVD
3     …      2.0    Rarely             Rarely           DVD
4     …      3.0     Never             Rarely           DVD
…   …      …       …                …           …
3182  …      2.0     Never            Monthly           DVD
3183  …      1.0    Rarely             Rarely           DVD
3184  …      4.0     Never             Rarely           DVD
3185  …      2.0     Never             Rarely           DVD
3186  …      1.0     Never             Rarely           DVD

     Prerec Renting Freq Prerec Viewing Freq CustomerID  Theater Freq
0                 Rarely             Monthly     877687       Monthly  \
1                Monthly              Weekly     877723        Rarely
2                 Weekly              Weekly     877757        Rarely
3                Monthly             Monthly     877792        Rarely
4                Monthly             Monthly     877840       Monthly
…                  …                 …        …           …
3182             Monthly              Weekly     927084       Monthly
3183               Never              Rarely     927147        Rarely
3184             Monthly              Weekly     927197        Rarely
3185              Weekly              Weekly     927390       Monthly
3186              Weekly              Weekly     927818        Rarely

     TV Movie Freq          TV Signal
0          Monthly              Cable
1           Weekly              Cable
2           Weekly              Cable
3            Daily              Cable
4           Weekly              Cable
…            …                …
3182        Rarely              Cable
3183        Weekly              Cable
3184         Never              Cable
3185         Daily  Digital Satellite
3186        Rarely              Cable

[3187 rows x 21 columns]
```

[31]:

```
data = data.dropna()
data
```

[31]:

```
       Age     Education Level Gender Home Ownership     Internet Connection
0     33.0           Doctorate   Male            Own                 Dial-Up  \
1     47.0           Doctorate   Male            Own                     DSL
3     35.0   Bachelor's Degree   Male            Own             Cable Modem
4     32.0   Bachelor's Degree   Male            Own             Cable Modem
5     32.0   Bachelor's Degree   Male            Own  No Internet Connection
…    …                 …    …            …                     …
3182  27.0     Master's Degree   Male           Rent                     DSL
```

```
3183  45.0  Associate's Degree   Male            Own                 Dial-Up
3184  38.0     Master's Degree   Male            Own                    IDSN
3185  31.0     Master's Degree   Male           Rent                    IDSN
3186  39.0     Master's Degree   Male           Rent                 Dial-Up

     Marital Status  Movie Selector  Num Bathrooms  Num Bedrooms  Num Cars
0           Married  Spouse/Partner            2.5             3       1.0  \
1           Married  Spouse/Partner            2.0             2       2.0
3           Married              Me            2.5             4       2.0
4           Married              Me            3.5             5       2.0
5           Married              Me            2.5             4       2.0
…             …             …             …            …       …
3182  Never Married              Me            1.0             1       2.0
3183        Married  Spouse/Partner            1.0             1       2.0
3184        Married              Me            1.5             3       2.0
3185        Married              Me            1.0             1       2.0
3186  Never Married              Me            1.0             1       1.0

      …  Num TVs  PPV Freq Prerec Buying Freq Prerec Format
0     …      2.0    Rarely            Monthly           DVD  \
1     …      1.0     Never            Monthly           DVD
3     …      2.0    Rarely             Rarely           DVD
4     …      3.0     Never             Rarely           DVD
5     …      1.0    Rarely             Rarely           DVD
…   …      …       …                …           …
3182  …      2.0     Never            Monthly           DVD
3183  …      1.0    Rarely             Rarely           DVD
3184  …      4.0     Never             Rarely           DVD
3185  …      2.0     Never             Rarely           DVD
3186  …      1.0     Never             Rarely           DVD

     Prerec Renting Freq Prerec Viewing Freq CustomerID  Theater Freq
0                 Rarely             Monthly     877687       Monthly  \
1                Monthly              Weekly     877723        Rarely
3                Monthly             Monthly     877792        Rarely
4                Monthly             Monthly     877840       Monthly
```

```
5                 Weekly              Weekly     877988        Weekly
…                  …                 …        …           …
3182             Monthly              Weekly     927084       Monthly
3183               Never              Rarely     927147        Rarely
3184             Monthly              Weekly     927197        Rarely
3185              Weekly              Weekly     927390       Monthly
3186              Weekly              Weekly     927818        Rarely

     TV Movie Freq          TV Signal
0          Monthly              Cable
1           Weekly              Cable
3            Daily              Cable
4           Weekly              Cable
5           Weekly  Digital Satellite
…             …                …
3182        Rarely              Cable
3183        Weekly              Cable
3184         Never              Cable
3185         Daily  Digital Satellite
3186        Rarely              Cable

[3085 rows x 21 columns]
```

[32]:
```
one_hot_cols = ['Gender', 'Internet Connection', 'Marital Status',
                'Movie Selector', 'Prerec Format', 'TV Signal',
                'Education Level', 'PPV Freq', 'Theater Freq',
                'TV Movie Freq', 'Prerec Buying Freq', 'Prerec Renting Freq',
                'Prerec Viewing Freq']
```

[33]:
```
from sklearn.preprocessing import OneHotEncoder
```

[34]:
```
one_hot_encoder = OneHotEncoder()
one_hot_encoder.fit(data[one_hot_cols])
one_hot_data = one_hot_encoder.transform(data[one_hot_cols])
```

[35]:
```
numeric_cols = ['Age', 'Num Bathrooms', 'Num Bedrooms',
                'Num Cars', 'Num Children', 'Num TVs']
```

[36]:

```
from scipy.sparse import hstack
```

[37]:

```
x = hstack([
    one_hot_data,
    data[numeric_cols].astype(float).values
])
y = data['Home Ownership']
```

[38]:

```
from sklearn.linear_model import LogisticRegression
```

[39]:

```
lrModel = LogisticRegression()
```

[40]:

```
from sklearn.model_selection import cross_val_score
```

[41]:

```
cvs = cross_val_score(
    lrModel,
    x,
    y,
    cv=10
)
cvs.mean()
```

```
/Users/liang/anaconda3/envs/python-course/lib/python3.9/site-
packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/Users/liang/anaconda3/envs/python-course/lib/python3.9/site-
packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/Users/liang/anaconda3/envs/python-course/lib/python3.9/site-
packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/Users/liang/anaconda3/envs/python-course/lib/python3.9/site-
packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/Users/liang/anaconda3/envs/python-course/lib/python3.9/site-
packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
```

```
  n_iter_i = _check_optimize_result(
/Users/liang/anaconda3/envs/python-course/lib/python3.9/site-
packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/Users/liang/anaconda3/envs/python-course/lib/python3.9/site-
packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/Users/liang/anaconda3/envs/python-course/lib/python3.9/site-
packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/Users/liang/anaconda3/envs/python-course/lib/python3.9/site-
packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
```

```
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
/Users/liang/anaconda3/envs/python-course/lib/python3.9/site-
packages/sklearn/linear_model/_logistic.py:458: ConvergenceWarning: lbfgs failed
to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
    https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
    https://scikit-learn.org/stable/modules/linear_model.html#logistic-
regression
  n_iter_i = _check_optimize_result(
```

```
[41]: 0.8359832723910392
```