homework_module_1

October 10, 2024

1 .

. ,
, , . .
, . code kaggle.
, : https://www.kaggle.com/hacktech33/titanic-solution-xgboost ( )
https://www.kaggle.com/shrutijhaa/in-top-3-titanic-machine-learning-from-disaster (,
,)
https://www.kaggle.com/blackhurt/top-3-using-voting-classifier (,
)

[1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
```

1.1

Titanic . , – ,
.
Google Colab, Colab Google
Drive. Google Drive , ,
. Google Drive + Google Colab

[3]:
```
#from google.colab import drive
#drive.mount('/content/gdrive/')
#data = pd.read_csv('/content/gdrive/My Drive/titanic_data.csv,␣
↪index_col='PassengerId')

#                    :
data = pd.read_csv('titanic_data.csv', index_col='PassengerId')
print(data)
```

```
             Pclass                                               Name  \
PassengerId
1                 3                            Braund, Mr. Owen Harris
```

```
2                 1  Cumings, Mrs. John Bradley (Florence Briggs Th…
3                 3                             Heikkinen, Miss. Laina
4                 1       Futrelle, Mrs. Jacques Heath (Lily May Peel)
5                 3                           Allen, Mr. William Henry
…               …                                                   …
887               2                              Montvila, Rev. Juozas
888               1                       Graham, Miss. Margaret Edith
889               3           Johnston, Miss. Catherine Helen "Carrie"
890               1                              Behr, Mr. Karl Howell
891               3                                Dooley, Mr. Patrick

                Sex   Age  SibSp  Parch            Ticket     Fare Cabin  \
PassengerId
1              male  22.0      1      0         A/5 21171   7.2500   NaN
2            female  38.0      1      0          PC 17599  71.2833   C85
3            female  26.0      0      0  STON/O2. 3101282   7.9250   NaN
4            female  35.0      1      0            113803  53.1000  C123
5              male  35.0      0      0            373450   8.0500   NaN
…               …   …    …    …                 …       …   …
887            male  27.0      0      0            211536  13.0000   NaN
888          female  19.0      0      0            112053  30.0000   B42
889          female   NaN      1      2        W./C. 6607  23.4500   NaN
890            male  26.0      0      0            111369  30.0000  C148
891            male  32.0      0      0            370376   7.7500   NaN

            Embarked
PassengerId
1                  S
2                  C
3                  S
4                  S
5                  S
…                …
887                S
888                S
889                S
890                C
891                Q

[891 rows x 10 columns]
```

[4]:
```
basic_features = data.columns

y = pd.read_csv('titanic_surv.csv')
y.index = data.index

print(f'    {len(data)}            ')
```

```
data.head()
```

```
   891
```

[4]:
```
             Pclass                                               Name  \
PassengerId
1                 3                            Braund, Mr. Owen Harris
2                 1  Cumings, Mrs. John Bradley (Florence Briggs Th…
3                 3                             Heikkinen, Miss. Laina
4                 1       Futrelle, Mrs. Jacques Heath (Lily May Peel)
5                 3                           Allen, Mr. William Henry

                Sex   Age  SibSp  Parch            Ticket     Fare Cabin  \
PassengerId
1              male  22.0      1      0         A/5 21171   7.2500   NaN
2            female  38.0      1      0          PC 17599  71.2833   C85
3            female  26.0      0      0  STON/O2. 3101282   7.9250   NaN
4            female  35.0      1      0            113803  53.1000  C123
5              male  35.0      0      0            373450   8.0500   NaN

            Embarked
PassengerId
1                  S
2                  C
3                  S
4                  S
5                  S
```

[ ]:
```
y.head()
```

[ ]:
```
             Survived
PassengerId
1                   0
2                   1
3                   1
4                   1
5                   0
```

[ ]:
```
data = data.join(y)
data.head()
```

[ ]:
```
             Pclass                                               Name  \
PassengerId
1                 3                            Braund, Mr. Owen Harris
2                 1  Cumings, Mrs. John Bradley (Florence Briggs Th…
3                 3                             Heikkinen, Miss. Laina
4                 1       Futrelle, Mrs. Jacques Heath (Lily May Peel)
5                 3                           Allen, Mr. William Henry
```

```
                Sex   Age  SibSp  Parch            Ticket     Fare Cabin  \
PassengerId
1              male  22.0      1      0         A/5 21171   7.2500   NaN
2            female  38.0      1      0          PC 17599  71.2833   C85
3            female  26.0      0      0  STON/O2. 3101282   7.9250   NaN
4            female  35.0      1      0            113803  53.1000  C123
5              male  35.0      0      0            373450   8.0500   NaN

            Embarked  Survived
PassengerId
1                  S         0
2                  C         1
3                  S         1
4                  S         1
5                  S         0
```

1.2

1.2.1 1 (1)

. , ? .
pd.Series.value_counts.

, . .
.

plt.hist().

[10]:
```
#<YOUR CODE>
print(pd.Series.value_counts)
```

```
<function IndexOpsMixin.value_counts at 0x00000194C6E76A70>
```

1.3

:

[11]:
```
data.columns[data.isna().any()].tolist()
```

[11]:
```
['Age', 'Cabin', 'Embarked']
```

“0”, Age — .

[12]:
```
data.loc[:, ['Cabin', 'Embarked']] = data.loc[:, ['Cabin', 'Embarked']].
↪fillna('0')
data['Age'] = data['Age'].fillna(data['Age'].median())
```

1.4

1.4.1 2. (0)

70:30.

[13]:
```
from sklearn.model_selection import train_test_split
#data_train, data_test =  #<YOUR CODE>
data_train, data_test = train_test_split(data, test_size=0.3, random_state=42)

#
print(f'              : {data_train.shape[0]}')
print(f'             : {data_test.shape[0]}')
```

```
             : 623
            : 268
```

1.5

. . ,

?

1.5.1 3 (1)

. .

: - . - .

. ( ) - , .

- kaggle.com.

[ ]:
```
data
```

[ ]:
```
             Pclass                                               Name  \
PassengerId
1                 3                            Braund, Mr. Owen Harris
2                 1  Cumings, Mrs. John Bradley (Florence Briggs Th…
3                 3                             Heikkinen, Miss. Laina
4                 1       Futrelle, Mrs. Jacques Heath (Lily May Peel)
5                 3                           Allen, Mr. William Henry
…               …                                                  …
887               2                              Montvila, Rev. Juozas
888               1                       Graham, Miss. Margaret Edith
889               3           Johnston, Miss. Catherine Helen "Carrie"
890               1                              Behr, Mr. Karl Howell
891               3                                Dooley, Mr. Patrick

                Sex   Age  SibSp  Parch            Ticket     Fare Cabin  \
PassengerId
1              male  22.0      1      0         A/5 21171   7.2500   NaN
2            female  38.0      1      0          PC 17599  71.2833   C85
```

```
3            female  26.0      0      0  STON/O2. 3101282   7.9250   NaN
4            female  35.0      1      0            113803  53.1000  C123
5              male  35.0      0      0            373450   8.0500   NaN
…               …    …    …     …                  …       …    …
887            male  27.0      0      0            211536  13.0000   NaN
888          female  19.0      0      0            112053  30.0000   B42
889          female   NaN      1      2        W./C. 6607  23.4500   NaN
890            male  26.0      0      0            111369  30.0000  C148
891            male  32.0      0      0            370376   7.7500   NaN

             Embarked  Survived
PassengerId
1                   S         0
2                   C         1
3                   S         1
4                   S         1
5                   S         0
…                 …         …
887                 S         0
888                 S         1
889                 S         0
890                 C         1
891                 Q         0

[891 rows x 11 columns]
```

[14]:
```
data['Cabin']
```

[14]:
```
PassengerId
1          0
2        C85
3          0
4       C123
5          0
        …
887        0
888     B42
889        0
890    C148
891        0
Name: Cabin, Length: 891, dtype: object
```

[17]:
```
def get_cabin_letter(row):
    return row['Cabin'][0]

data_train['cabin_type'] = data.apply(get_cabin_letter, axis=1)
print(data_train)
```

```
             Pclass                                Name     Sex   Age  SibSp  \
PassengerId
446               1         Dodge, Master. Washington    male   4.0      0
651               3                 Mitkoff, Mr. Mito    male  28.0      0
173               3      Johnson, Miss. Eleanor Ileen  female   1.0      1
451               2             West, Mr. Edwy Arthur    male  36.0      1
315               2                Hart, Mr. Benjamin    male  43.0      1
…               …                                 …     …   …    …
107               3  Salkjelsvik, Miss. Anna Kristine  female  21.0      0
271               1             Cairns, Mr. Alexander    male  28.0      0
861               3           Hansen, Mr. Claus Peter    male  41.0      2
436               1         Carter, Miss. Lucile Polk  female  14.0      1
103               1         White, Mr. Richard Frasar    male  21.0      0

             Parch        Ticket      Fare    Cabin Embarked cabin_type
PassengerId
446              2         33638   81.8583      A34        S          A
651              0        349221    7.8958        0        S          0
173              1        347742   11.1333        0        S          0
451              2    C.A. 34651   27.7500        0        S          0
315              1  F.C.C. 13529   26.2500        0        S          0
…              …             …       …       …      …        …
107              0        343120    7.6500        0        S          0
271              0        113798   31.0000        0        S          0
861              0        350026   14.1083        0        S          0
436              2        113760  120.0000  B96 B98        S          B
103              1         35281   77.2875      D26        S          D

[623 rows x 11 columns]
```

1.6

, , . ,
: , .

1.6.1 4 (0)

, pd.get_dummies. ,
, .

[18]:
```
#<YOUR CODE>
pd.get_dummies
```

[18]:
```
<function pandas.core.reshape.encoding.get_dummies(data, prefix=None,
prefix_sep: 'str | Iterable[str] | dict[str, str]' = '_', dummy_na: 'bool' =
False, columns=None, sparse: 'bool' = False, drop_first: 'bool' = False, dtype:
'NpDtype | None' = None) -> 'DataFrame'>
```

1.7 baseline-

1.7.1 5 (1)

. - (!)

[19]:
```
#<YOUR CODE>
pd.get_dummies(data)
```

[19]:

PassengerId	Pclass	Age	SibSp	Parch	Fare	Name_Abbing, Mr. Anthony	Name_Abbott, Mr. Rossmore Edward	Name_Abbott, Mrs. Stanton (Rosa Hunt)	Name_Abelson, Mr. Samuel
1	3	22.0	1	0	7.2500	False	False	False	False
2	1	38.0	1	0	71.2833	False	False	False	False
3	3	26.0	0	0	7.9250	False	False	False	False
4	1	35.0	1	0	53.1000	False	False	False	False
5	3	35.0	0	0	8.0500	False	False	False	False
…	…	…	…	…	…	…	…	…	…
887	2	27.0	0	0	13.0000	False	False	False	False
888	1	19.0	0	0	30.0000	False	False	False	False
889	3	28.0	1	2	23.4500	False	False	False	False
890	1	26.0	0	0	30.0000	False	False	False	False
891	3	32.0	0	0	7.7500	False	False	False	False

```
890                                          False                    False
891                                          False                    False

             Name_Abelson, Mrs. Samuel (Hannah Wizosky)  …  Cabin_F2  \
PassengerId                                              …
1                                                False  …     False
2                                                False  …     False
3                                                False  …     False
4                                                False  …     False
5                                                False  …     False
…                                                    …  …       …
887                                              False  …     False
888                                              False  …     False
889                                              False  …     False
890                                              False  …     False
891                                              False  …     False

             Cabin_F33  Cabin_F38  Cabin_F4  Cabin_G6  Cabin_T  Embarked_0  \
PassengerId
1                False      False     False     False    False       False
2                False      False     False     False    False       False
3                False      False     False     False    False       False
4                False      False     False     False    False       False
5                False      False     False     False    False       False
…                    …          …         …         …        …           …
887              False      False     False     False    False       False
888              False      False     False     False    False       False
889              False      False     False     False    False       False
890              False      False     False     False    False       False
891              False      False     False     False    False       False

             Embarked_C  Embarked_Q  Embarked_S
PassengerId
1                 False       False        True
2                  True       False       False
3                 False       False        True
4                 False       False        True
5                 False       False        True
…                     …           …           …
887               False       False        True
888               False       False        True
889               False       False        True
890                True       False       False
891               False        True       False

[891 rows x 1731 columns]
```

1.7.2 6 (1)

. , , ohe-hot .
. - ?

[27]:
```
#<YOUR CODE>
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import OneHotEncoder

label_encoder = LabelEncoder()
onehot_encoder = OneHotEncoder(sparse=False)
```

1.8

1.8.1 7 (2)

. , .
. .

[28]:
```
#<YOUR CODE>
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
from sklearn.datasets import load_breast_cancer

data = load_breast_cancer()
X = data.data
y = data.target

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,␣
 ↪random_state=42)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

logreg = LogisticRegression()

param_grid = {
    'C': [0.01, 0.1, 1, 10, 100],
    'penalty': ['l1', 'l2'],
    'solver': ['liblinear']
}

grid_search = GridSearchCV(logreg, param_grid, cv=5, scoring='accuracy')
grid_search.fit(X_train_scaled, y_train)
```

```
print("         :", grid_search.best_params_)
print("           -     :", grid_search.best_score_)

best_logreg = grid_search.best_estimator_
y_pred = best_logreg.predict(X_test_scaled)

print("                :", accuracy_score(y_test, y_pred))
print("             :\n", classification_report(y_test, y_pred))
```

```
         : {'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}
           -     : 0.977373417721519
                : 0.9941520467836257
             :
                precision    recall  f1-score   support

           0       1.00      0.98      0.99        63
           1       0.99      1.00      1.00       108

    accuracy                           0.99       171
   macro avg       1.00      0.99      0.99       171
weighted avg       0.99      0.99      0.99       171
```

1.9

1.9.1 8 (1)

- . y_test.
.

[29]:
```
#<YOUR CODE>
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split, cross_val_score,␣
 ↪GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
from sklearn.datasets import load_breast_cancer

data = load_breast_cancer()
X = data.data
y = data.target

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,␣
 ↪random_state=42)
```

```
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

logreg = LogisticRegression()

param_grid = {
    'C': [0.01, 0.1, 1, 10, 100],
    'penalty': ['l1', 'l2'],
    'solver': ['liblinear']
}

grid_search = GridSearchCV(logreg, param_grid, cv=5, scoring='accuracy')
grid_search.fit(X_train_scaled, y_train)

print("         :", grid_search.best_params_)
print("             -     :", grid_search.best_score_)

best_logreg = grid_search.best_estimator_

#                -
cv_scores = cross_val_score(best_logreg, X_train_scaled, y_train, cv=5,␣
 ↪scoring='accuracy')
print("         -                :", cv_scores)
print("              -     :", np.mean(cv_scores))

#
y_test_pred = best_logreg.predict(X_test_scaled)

#
test_accuracy = accuracy_score(y_test, y_test_pred)
print("                  :", test_accuracy)

print("                          :\n", classification_report(y_test,␣
 ↪y_test_pred))
```

```
         : {'C': 0.1, 'penalty': 'l2', 'solver': 'liblinear'}
             -     : 0.977373417721519
         -                : [0.975      0.975      0.9875
0.97468354 0.97468354]
              -     : 0.977373417721519
                  : 0.9941520467836257
                          :
               precision    recall  f1-score   support

           0       1.00      0.98      0.99        63
           1       0.99      1.00      1.00       108
```

```
    accuracy                           0.99       171
   macro avg       1.00      0.99      0.99       171
weighted avg       0.99      0.99      0.99       171
```

1.10

1.10.1 9 (3)

. ? ?