ds-and-health

October 21, 2020

0.1 Vignette: asking and answering a predictive question

by Mike Gelbart, Assistant Professor of Teaching @ UBC Department of Computer Science

Presented for Data Science & Health 2020

[2]:
```python
import numpy as np
import pandas as pd
from sklearn.compose import ColumnTransformer
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression, LogisticRegressionCV
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import OneHotEncoder, StandardScaler
from sklearn.tree import DecisionTreeClassifier

from sklearn.dummy import DummyClassifier


# Preprocessing and pipeline
from sklearn.impute import SimpleImputer

# train test split and cross validation
from sklearn.model_selection import (
    GridSearchCV,
    cross_val_score,
    cross_validate,
    train_test_split,
)
from sklearn.pipeline import FeatureUnion, Pipeline, make_pipeline
from sklearn.preprocessing import (
    FunctionTransformer,
    Normalizer,
    OneHotEncoder,
    StandardScaler,
    normalize,
    scale,
)
from sklearn.svm import SVC
from sklearn.tree import DecisionTreeClassifier
from sklearn.datasets import load_breast_cancer
```

Task: Predicting the presence or absence of heart disease (the target) based on a set of 13 different biophysical measures (the features).

Dataset: from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[3]:
```
heart_df = pd.read_csv("heart_disease.csv", index_col=0)
```

[4]:
```
df_train, df_test = train_test_split(heart_df, random_state=123)
```

[5]:
```
df_train
```

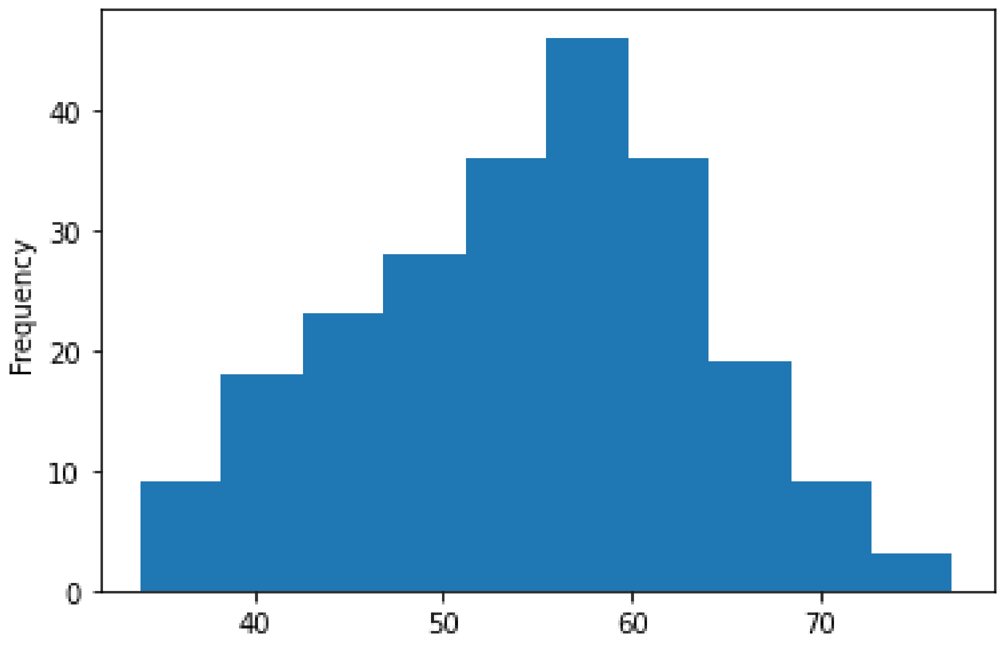
[5]:
```
            age     sex   chest_pain_type  resting_blood_pressure  \
patient_id
36           54  female  non-anginal pain                     135
148          44    male  non-anginal pain                     120
21           44    male  non-anginal pain                     130
187          54    male            angina                     124
161          55  female  non-anginal pain                     132
…           …     …                 …                      …
106          69    male  non-anginal pain                     160
83           52    male  non-anginal pain                     152
17           66  female  non-anginal pain                     150
230          47    male  non-anginal pain                     108
98           43    male  non-anginal pain                     130

            cholesterol    fasting_blood_sugar  rest_ecg  \
patient_id
36                  304  greater than 120mg/ml  abnormal
148                 226    lower than 120mg/ml  abnormal
21                  233    lower than 120mg/ml  abnormal
187                 266    lower than 120mg/ml    normal
161                 342    lower than 120mg/ml  abnormal
…                 …                      …       …
106                 234  greater than 120mg/ml    normal
83                  298  greater than 120mg/ml  abnormal
17                  226    lower than 120mg/ml  abnormal
230                 243    lower than 120mg/ml  abnormal
98                  315    lower than 120mg/ml  abnormal

            max_heart_rate_achieved exercise_induced_angina  st_depression  \
patient_id
36                              170                      no            0.0
148                             169                      no            0.0
21                              179                     yes            0.4
187                             109                     yes            2.2
```

```
161                              166                 no            1.2
…                              …                 …            …
106                              131                 no            0.1
83                               178                 no            1.2
17                               114                 no            2.6
230                              152                 no            0.0
98                               162                 no            1.9

               st_slope  num_major_vessels thalassemia  target
patient_id
36          downsloping                  0      normal       1
148         downsloping                  0      normal       1
21          downsloping                  0      normal       1
187                flat                  1    abnormal       0
161         downsloping                  0      normal       1
…                …                …         …     …
106                flat                  1      normal       1
83                 flat                  0    abnormal       1
17            upsloping                  0      normal       1
230         downsloping                  0      normal       0
98          downsloping                  1      normal       1

[227 rows x 14 columns]
```

[6]:
```
df_train['target'].value_counts()
```

[6]:
```
1    128
0     99
Name: target, dtype: int64
```

[7]:
```
df_train['age'].plot.hist();
```



[8]:
```
numeric_features = [
    "age",
    "resting_blood_pressure",
    "cholesterol",
    "max_heart_rate_achieved",
    "st_depression",
    "num_major_vessels",
]
categorical_features = [
    "sex",
    "chest_pain_type",
    "fasting_blood_sugar",
    "rest_ecg",
    "exercise_induced_angina",
    "st_slope",
    "thalassemia",
]
```

[9]:
```
X_train = df_train.drop(columns=["target"])
y_train = df_train["target"]

X_test = df_test.drop(columns=["target"])
y_test = df_test["target"]
```

[10]:
```
df_train["st_slope"]
```

```
[10]: patient_id
      36     downsloping
      148    downsloping
      21     downsloping
      187           flat
      161    downsloping
                 …
      106           flat
      83            flat
      17       upsloping
      230    downsloping
      98     downsloping
      Name: st_slope, Length: 227, dtype: object
```

```
[11]: preprocessor = ColumnTransformer([
          ("scale", StandardScaler(), numeric_features),
          ("ohe", OneHotEncoder(), categorical_features),
      ])
      preprocessor.fit(X_train)
      new_columns = numeric_features + list(preprocessor.named_transformers_["ohe"].
       ↪get_feature_names(categorical_features))
      X_train_enc = pd.DataFrame(preprocessor.transform(X_train),␣
       ↪columns=new_columns, index=X_train.index)
      X_train_enc
```

```
[11]:                  age  resting_blood_pressure  cholesterol  \
      patient_id
      36         -0.057019                0.213311     1.033292
      148        -1.163276               -0.706088    -0.414936
      21         -1.163276               -0.093155    -0.284967
      187        -0.057019               -0.460915     0.327745
      161         0.053607                0.029432     1.738839
      …                …                     …            …
      106         1.602368                1.745644    -0.266400
      83         -0.278270                1.255298     0.921890
      17          1.270490                1.132711    -0.414936
      230        -0.831399               -1.441608    -0.099297
      98         -1.273902               -0.093155     1.237529

                  max_heart_rate_achieved  st_depression  num_major_vessels  \
      patient_id
      36                         0.861754      -0.900071          -0.715888
      148                        0.818128      -0.900071          -0.715888
      21                         1.254386      -0.525523          -0.715888
      187                       -1.799420       1.159945           0.293470
      161                        0.687250       0.223574          -0.715888
      …                                …             …                  …
```

```
106                    -0.839652     -0.806434           0.293470
83                      1.210760      0.223574          -0.715888
17                     -1.581291      1.534493          -0.715888
230                     0.076489     -0.900071          -0.715888
98                      0.512747      0.879034           0.293470

            sex_female  sex_male  chest_pain_type_angina  \
patient_id
36                 1.0       0.0                     0.0
148                0.0       1.0                     0.0
21                 0.0       1.0                     0.0
187                0.0       1.0                     1.0
161                1.0       0.0                     0.0
…                  …         …                       …
106                0.0       1.0                     0.0
83                 0.0       1.0                     0.0
17                 1.0       0.0                     0.0
230                0.0       1.0                     0.0
98                 0.0       1.0                     0.0

            chest_pain_type_non-anginal pain  …  \
patient_id                                    …
36                                       1.0  …
148                                      1.0  …
21                                       1.0  …
187                                      0.0  …
161                                      1.0  …
…                                        …  …
106                                      1.0  …
83                                       1.0  …
17                                       1.0  …
230                                      1.0  …
98                                       1.0  …

            fasting_blood_sugar_lower than 120mg/ml  rest_ecg_abnormal  \
patient_id
36                                              0.0                1.0
148                                             1.0                1.0
21                                              1.0                1.0
187                                             1.0                0.0
161                                             1.0                1.0
…                                               …                  …
106                                             0.0                0.0
83                                              0.0                1.0
17                                              1.0                1.0
230                                             1.0                1.0
98                                              1.0                1.0
```

```
            rest_ecg_normal  exercise_induced_angina_no  \
patient_id
36                      0.0                         1.0
148                     0.0                         1.0
21                      0.0                         0.0
187                     1.0                         0.0
161                     0.0                         1.0
…                         …                           …
106                     1.0                         1.0
83                      0.0                         1.0
17                      0.0                         1.0
230                     0.0                         1.0
98                      0.0                         1.0

            exercise_induced_angina_yes  st_slope_downsloping  st_slope_flat  \
patient_id
36                                  0.0                   1.0            0.0
148                                 0.0                   1.0            0.0
21                                  1.0                   1.0            0.0
187                                 1.0                   0.0            1.0
161                                 0.0                   1.0            0.0
…                                     …                     …              …
106                                 0.0                   0.0            1.0
83                                  0.0                   0.0            1.0
17                                  0.0                   0.0            0.0
230                                 0.0                   1.0            0.0
98                                  0.0                   1.0            0.0

            st_slope_upsloping  thalassemia_abnormal  thalassemia_normal
patient_id
36                         0.0                   0.0                 1.0
148                        0.0                   0.0                 1.0
21                         0.0                   0.0                 1.0
187                        0.0                   1.0                 0.0
161                        0.0                   0.0                 1.0
…                            …                     …                   …
106                        0.0                   0.0                 1.0
83                         0.0                   1.0                 0.0
17                         1.0                   0.0                 1.0
230                        0.0                   0.0                 1.0
98                         0.0                   0.0                 1.0

[227 rows x 21 columns]
```

[12]:
```
pipe = make_pipeline(preprocessor, LogisticRegressionCV())
```

```
[13]: cross_val_score(pipe, X_train, y_train, cv=20).mean()
```

```
[13]: 0.868560606060606
```

```
[15]: pipe.fit(X_train, y_train);
```

```
[17]: X_test[:1]
```

```
[17]:             age     sex   chest_pain_type  resting_blood_pressure  \
      patient_id
      11           48  female  non-anginal pain                     130

                  cholesterol  fasting_blood_sugar  rest_ecg  \
      patient_id
      11                  275  lower than 120mg/ml  abnormal

                  max_heart_rate_achieved exercise_induced_angina  st_depression  \
      patient_id
      11                              139                      no            0.2

                     st_slope  num_major_vessels thalassemia
      patient_id
      11          downsloping                  0      normal
```

```
[18]: pipe.predict(X_test[:1])
```

```
[18]: array([1])
```

```
[19]: pipe.predict_proba(X_test[:1])
```

```
[19]: array([[0.10810309, 0.89189691]])
```

```
[20]: y_test[:1]
```

```
[20]: patient_id
      11    1
      Name: target, dtype: int64
```

```
[21]: pipe.score(X_test, y_test)
```

```
[21]: 0.8026315789473685
```

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```

0.2 Challenges

Type of data:

- Images
- Videos
- Text
- Time series
- Censored data (survival analysis)
- ...

Data quality:

- Missing data
- Incorrect data
- Outliers
- ...

Data quantity:

- Need lots of data to get this to work
- For this type of analysis, need **labeled** data

Computational issues:

- Code may take a long time to run?
- Do we need distributed / cloud computing?
- Are the tools actively maintained?
- ...

Error metrics:

- False positive vs. false negatives (sensitivity vs. specificity)
- ...

Ethical challenges:

- How confident are we really in our results?
- Can we trust it if we don't understand it?
- Relevant article: https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98
- Is our model biased?

0.3 Where to learn more data science at UBC?

- https://extendedlearning.ubc.ca/programs/key-capabilities-data-science and https://prog-learn.mds.ubc.ca/en
- https://masterdatascience.ubc.ca/
- DSCI 100, CPSC 330
- Upcoming undergraduate courses

```
[ ]:
```