Instructions
Deadline: 11.59 PM May 17, 2024
Submission File Format: A compressed zip file containing a report in PDF format – 3000 words and code implementation in the jupyter notebook file

Bio216 Artificial Intelligence for Life Science 2023-24SEM22
Coursework 2 (Individual Work)
Question: Regression - Heart disease prediction

Introduction

World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardiovascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using regression methods.

Source

The dataset (https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression?resource=download) is available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

● Variables

Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.

● Demographic:

- Sex: male or female (binary outcome (0, 1) corresponds to male and female individuals).
- Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

● Behavioral

- Current Smoker: whether or not the patient is a current smoker (Nominal)
- Cigs Per Day: the number of cigarettes that the person smoked on average in one day. (can be considered continuous as one can have any number of cigarettes, even half a cigarette.)

● Medical (history)

- BP Meds: whether or not the patient was on blood pressure medication (Nominal)
- Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
- Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
- Diabetes: whether or not the patient had diabetes (Nominal)

● Medical(current)

- Tot Chol: total cholesterol level (Continuous)
- Sys BP: systolic blood pressure (Continuous)
- Dia BP: diastolic blood pressure (Continuous)
- BMI: Body Mass Index (Continuous)

- Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
- Glucose: glucose level (Continuous)

Predict variable (desired target)

- 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

Study Objective and Outcome: You will be expected to develop a regression model based on a binary CHD outcome with certain accuracy (Like above 80%). Then, the proposed model can assess the impact of all risk factors (variables) on CHD susceptibility.

Marking Criteria

Data Splitting (10%):

Reasonable data splitting into training and testing sets.

Model Training (20%)

Successfully training the model on the training set.

Utilizing the selected regression algorithm and optimizing its parameters to best fit the training data.

Choose an appropriate regression algorithm based on the nature of the problem and the characteristics of the dataset. Explain the rationale behind the selection, considering factors such as linearity, complexity, and interpretability of the model.

Model Evaluation (20%):

Evaluating the model’s performance on the testing set.

Using appropriate metrics such as mean squared error, R-squared, or other relevant metrics.

Visualizing the model’s performance through appropriate plots or graphs.

Feature Importance Analysis (20%):

Analyzing the importance of each feature in predicting the outcome.

Using techniques like feature importance scores, coefficients, or permutation importance.

Providing clear explanations for the importance of key features.

Documentation and Reporting (30%):

Documenting the entire process in a well-organized Jupyter notebook.

Presenting the findings and results clearly in the report.

Submitting a compressed zip file containing the Jupyter notebook and a report in PDF format with the word limitation as 3000 words (exclude references, table of contents, appendix and tables) with a logical structure (introduction – main body – conclusion, use sub-title to divide main parts into several sections according to the needs).

Using headings for sections

Typeset properly, e.g., in Google Doc, Word, or LaTex (recommended); you could use an online LaTex environment such as Overleaf and use one of its Homework Templates

Academic writing style, no spelling or grammar mistakes.

Include Page numbers

Font (Times New Roman), Font size (10), Line spacing (Single)

Answer:

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
import scipy.stats as st
import matplotlib.pyplot as plt
import seaborn as sn
from sklearn.metrics import confusion_matrix
import matplotlib.mlab as mlab
%matplotlib inline

heart_df=pd.read_csv("../input/framingham.csv")
heart_df.drop(['education'],axis=1,inplace=True)
heart_df.head()

heart_df.rename(columns={'male':'Sex_male'},inplace=True)

heart_df.isnull().sum()

count=0
for i in heart_df.isnull().sum(axis=1):
    if i>0:
        count=count+1
print('Total number of rows with missing values is ', count)
print('since it is only',round((count/len(heart_df.index))*100), 'percent of the entire dataset the rows with missing values are excluded.')

heart_df.dropna(axis=0,inplace=True)

def draw_histograms(dataframe, features, rows, cols):
    fig=plt.figure(figsize=(20,20))
    for i, feature in enumerate(features):
        ax=fig.add_subplot(rows,cols,i+1)
        dataframe[feature].hist(bins=20,ax=ax,facecolor='midnightblue')
        ax.set_title(feature+" Distribution",color='DarkRed')

    fig.tight_layout()
    plt.show()
draw_histograms(heart_df,heart_df.columns,6,3)

heart_df.TenYearCHD.value_counts()

sn.pairplot(data=heart_df)
```

```
heart_df.describe()

from statsmodels.tools import add_constant as add_constant
heart_df_constant = add_constant(heart_df)
heart_df_constant.head()

st.chisqprob = lambda chisq, df: st.chi2.sf(chisq, df)
cols=heart_df_constant.columns[:-1]
model=sm.Logit(heart_df.TenYearCHD,heart_df_constant[cols])
result=model.fit()
result.summary()

def back_feature_elem (data_frame,dep_var,col_list):
    """ Takes in the dataframe, the dependent variable and a list of column names, runs the
regression repeatedly eleminating feature with the highest
    P-value above alpha one at a time and returns the regression summary with all p-values below
alpha"""

    while len(col_list)>0 :
        model=sm.Logit(dep_var,data_frame[col_list])
        result=model.fit(disp=0)
        largest_pvalue=round(result.pvalues,3).nlargest(1)
        if largest_pvalue[0]<(0.05):
            return result
            break
        else:
            col_list=col_list.drop(largest_pvalue.index)

result=back_feature_elem(heart_df_constant,heart_df.TenYearCHD,cols)

result.summary()

params = np.exp(result.params)
conf = np.exp(result.conf_int())
conf['OR'] = params
pvalue=round(result.pvalues,3)
conf['pvalue']=pvalue
conf.columns = ['CI 95%(2.5%)', 'CI 95%(97.5%)', 'Odds Ratio','pvalue']
print ((conf))

import sklearn
new_features=heart_df[['age','Sex_male','cigsPerDay','totChol','sysBP','glucose','TenYearCHD']]
x=new_features.iloc[:,:-1]
```

```
y=new_features.iloc[:,-1]
from sklearn.cross_validation import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=.20,random_state=5)

from sklearn.linear_model import LogisticRegression
logreg=LogisticRegression()
logreg.fit(x_train,y_train)
y_pred=logreg.predict(x_test)

sklearn.metrics.accuracy_score(y_test,y_pred)
```

```
from sklearn.metrics import confusion_matrix
cm=confusion_matrix(y_test,y_pred)
conf_matrix=pd.DataFrame(data=cm,columns=['Predicted:0','Predicted:1'],index=['Actual:0','Actual:1'])
plt.figure(figsize = (8,5))
sn.heatmap(conf_matrix, annot=True,fmt='d',cmap="YlGnBu")

TN=cm[0,0]
TP=cm[1,1]
FN=cm[1,0]
FP=cm[0,1]
sensitivity=TP/float(TP+FN)
specificity=TN/float(TN+FP)

print('The acuuracy of the model = TP+TN/(TP+TN+FP+FN) = ',(TP+TN)/float(TP+TN+FP+FN),'\n',

'The Missclassification = 1-Accuracy = ',1-((TP+TN)/float(TP+TN+FP+FN)),'\n',

'Sensitivity or True Positive Rate = TP/(TP+FN) = ',TP/float(TP+FN),'\n',

'Specificity or True Negative Rate = TN/(TN+FP) = ',TN/float(TN+FP),'\n',

'Positive Predictive value = TP/(TP+FP) = ',TP/float(TP+FP),'\n',

'Negative predictive Value = TN/(TN+FN) = ',TN/float(TN+FN),'\n',

'Positive Likelihood Ratio = Sensitivity/(1-Specificity) = ',sensitivity/(1-specificity),'\n',

'Negative likelihood Ratio = (1-Sensitivity)/Specificity = ',(1-sensitivity)/specificity)

y_pred_prob=logreg.predict_proba(x_test)[:,:]
y_pred_prob_df=pd.DataFrame(data=y_pred_prob, columns=['Prob of no heart disease (0)','Prob of Heart Disease (1)'])
y_pred_prob_df.head()

from sklearn.preprocessing import binarize
for i in range(1,5):
    cm2=0
    y_pred_prob_yes=logreg.predict_proba(x_test)
    y_pred2=binarize(y_pred_prob_yes,i/10)[:,1]
    cm2=confusion_matrix(y_test,y_pred2)
    print ('With',i/10,'threshold the Confusion Matrix is ','\n',cm2,'\n',
            'with',cm2[0,0]+cm2[1,1],'correct predictions and',cm2[1,0],'Type II errors( False
Negatives)','\n\n',
        'Sensitivity:    ',cm2[1,1]/(float(cm2[1,1]+cm2[1,0])),'Specificity:
',cm2[0,0]/(float(cm2[0,0]+cm2[0,1])),'\n\n\n')

from sklearn.metrics import roc_curve
fpr, tpr, thresholds = roc_curve(y_test, y_pred_prob_yes[:,1])
plt.plot(fpr,tpr)
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.0])
plt.title('ROC curve for Heart disease classifier')
plt.xlabel('False positive rate (1-Specificity)')
plt.ylabel('True positive rate (Sensitivity)')
plt.grid(True)

sklearn.metrics.roc_auc_score(y_test,y_pred_prob_yes[:,1])
```