Bochra CHEMAM

Data science

MOHAMMED VI POLYTECHNIC UNIVERSITY

"The best way to learn data science is to apply data science."

Activity 7- Assignement

21/04/2021-27/04/2021

For this Activity session we will:

- Work through a classification
- Visualize the data
- Evaluate the algorithms

a. Logistic Regression Assumptions

* Binary logistic regression requires the dependent variable to be binary.
* For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
* Only the meaningful variables should be included.
* The independent variables should be independent of each other. That is, the model should have little or no multicollinearity.
* The independent variables are linearly related to the log odds.
* Logistic regression requires quite large sample sizes.

Keeping the above assumptions in mind

I. Import different libraries and Load dataset (diabetes.csv)

II. VARIABLE DESCRIPTIONS

1. read in the dataset using the Pandas' read_csv() function.
2. Use .info() , head() , hist() to familiarise with data
3. Use .describe() , hist() and interpret results
4. Checking that your target variable is binary and plot it

III. Data preprocessing

5. check for missing values by calling the isnull() method, and the sum() method off of that, to return a tally of all the True values that are returned by the isnull() method.

6. **Taking care of missing values:**

   - replace the np.nan values with the median values of the attributes.

7. Check again missing value using .isnull().sum()

8. **Checking for independence between features : .corr()**

   Drop dependent variables if any

9. **Preparing data for models:**
   - Scaling of data using standardscaler

10. **Use .iloc to split your data x and y**

11. **Split to train and test set (75% / 25%) , (80%/20%) or (90%/10%)**

IV. **Deploying and evaluating the model**

- **Deploying**
  1. LogReg = LogisticRegression()
  2. Fit x_train and y_train LogReg.fit(X_train, y_train)
  3. Predict y_pred using y_pred=LogReg.predict(X_test)
- **Evaluation**
  - **Use confusion matrix :**
  - Calculate different metrics and interpret results

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

Model Performance

Accuracy	= (TN+TP)/(TN+FP+FN+TP)
Precision	= TP/(FP+TP)
Sensitivity	= TP/(TP+FN)
Specificity	= TN/(TN+FP)