|  |  |  |
| --- | --- | --- |
| **AIN 380** | **Homework 10**  **Logistic Regression Classification** | **DUE:** |

Create a Logistic Regression Classifier machine learning model to serve as an automated rating system for a website’s community discussion forum. Posts to the forum are to be rated as “bad” denoted “B” or “good” denoted “G”. The feature data consists of the following:

|  |  |  |
| --- | --- | --- |
| ◊ | num\_words | Number of words in the post |
| ◊ | num\_characters | Number of characters in the post |
| ◊ | num\_misspelled | Number of misspelled words in the post |
| ◊ | bin\_end\_qmark | Whether or not the post ends with a question mark (1 = yes) |
| ◊ | num\_interrogative | Number of interrogative words in the post |
| ◊ | bin\_start\_small | Whether or not the post begins with a lowercase letter (1 = yes) |
| ◊ | num\_sentences | Number of sentences per post |
| ◊ | num\_punctuations | Number of punctuation symbols in the post |
| ◊ | label | Quality label **G**ood or **B**ad |

Use the data found in the ‘*quality.csv*’ file to create training and test data sets. Use a split of 70%-30% and be sure to replace “G” with 1 and “B” with 0.

Print the sklearn metrics accuracy score for your model as well as the confusion matrix. In addition, print the ROC curve and the Area Under the Curve (AUC) score.

What is the AUC score you obtained?

.722

Using your ML model, what is the post quality prediction for the input below? Print it as G or B.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| num\_words | num\_characters | num\_misspelled | bin\_end\_qmark | num\_interrogative | bin\_start\_small | num\_sentences | num\_punctuations |
| 18 | 95 | 2 | 0 | 2 | 0 | 2 | 3 |

What is the quality you obtained (G or B)?

G

Using your ML model, what is the post quality prediction for the input below? Print it as G or B.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| num\_words | num\_characters | num\_misspelled | bin\_end\_qmark | num\_interrogative | bin\_start\_small | num\_sentences | num\_punctuations |
| 18 | 95 | 2 | 0 | 2 | 0 | 2 | 3 |

What is the quality you obtained (G or B)?

G

*On the due date, submit your answers on this sheet along with a printed listing of your Python code and the output results.*