Homework 3: Trains

100 Possible Points

| 12/1/2022

Attempt 1

IN PROGRESS
Next Up: Submit Assignment

Add Comment

Unlimited Attempts Allowed

Details

Homework 3 involves a logistic regression on data recording information about freight train accidents (including mostly minor incidents, such as driving into an occupied block, but with no collisions). You will be making a Jupyter Notebook, and including the data file FRAFirm.csv, which is located in the module section of Canvas.

Each record in the file contains information about one particular shift that an engineer or conductor worked. Clock-in and clock-out information, plus many statistics, are provided. One column, named 'class', is actually the target. This column contains an integer that can have one of three values:

0: No accident occurred during this shift

1: An accident of type '1' occurred during this shift

2. An accident of type '2' occurred during this shift

A number of people have analyzed this data, and there are two schools of thought:

(A) The two types of accident, 1 and 2, really are different types of accident, caused by different situations, so these should be distinguishable by the various statistics in the file. The classifier should find 3 distinct groups (0, 1, and 2), and there should be different criteria used by the regressor to determine which group the sample belongs in.

(B) The two types of accident are indistinguishable, there is no real difference. So the regressor should only categorize into two groups: No Accident and Accident.

The first question we would like answered is this: Are the two types of accident distinguishable? Is there a fairly reliable way to tell these apart?

The second question (which depends upon the answer to the first!) is this: How accurately does the regressor classify the samples? Graphs and numbers may be interesting here, but a confusion matrix would be helpful. If you determine that the two classes are indistinguishable, that there is only one type of accident, then the simple confusion matrix suffices. But if you determine that the types

Submit Assignment

are distinguishable, then we either need 3 confusion matrices (No accident vs type 1, No accident vs type 2, and type 1 vs type 2), or if you can, a 3-way confusion matrix.

The third question is this: Which of the features (input columns) are significant in performing the classification, and which can/should be ignored?

There is also an opportunity for some extra credit. After you have completed the above, remove the column named 'FIRM' from the input data, then perform the regression again. How does removing this column affect the accuracy of the results, and how does this affect the significance of the other features.

In your Jupyter notebook, in addition to the cells which contain the code, include markdown cells explaining what you are doing, or highlighting conclusions that you can draw from the analysis.

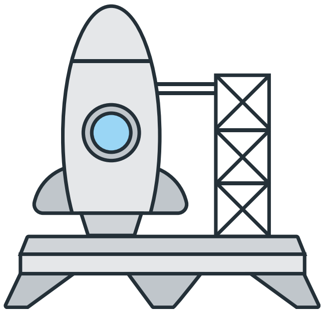
It is helpful if you do NOT clear the cells before turning in your results, because otherwise I have to run all of your results rather than just reading all of your results!

Check your notebook in to Canvas to submit your homework!

Webcam Photo

Canvas Files

or



Choose a file to upload

Submit Assignment