DIRTY DATA YOU HAVE
CLEAN IT UP YOU MUST
memegenerator.net

Welcome to Week 8 Lecture 1!

Data Science in Python & Machine Learning

Announcements

Final Items to Move onto Stack 3

- Passing Belt Exam
- Submitted 90% of all assignments including resubmits by Friday March 18th at 9am PST.
- Attended at least 80% of the live lectures

Grading Week 3 Assignments

- Still Grading week 3 assignments.

Learning Goals

After this class you will be able to:

1. Know when to use a logistic regression model

2. Use a logistic regression model with a pipeline

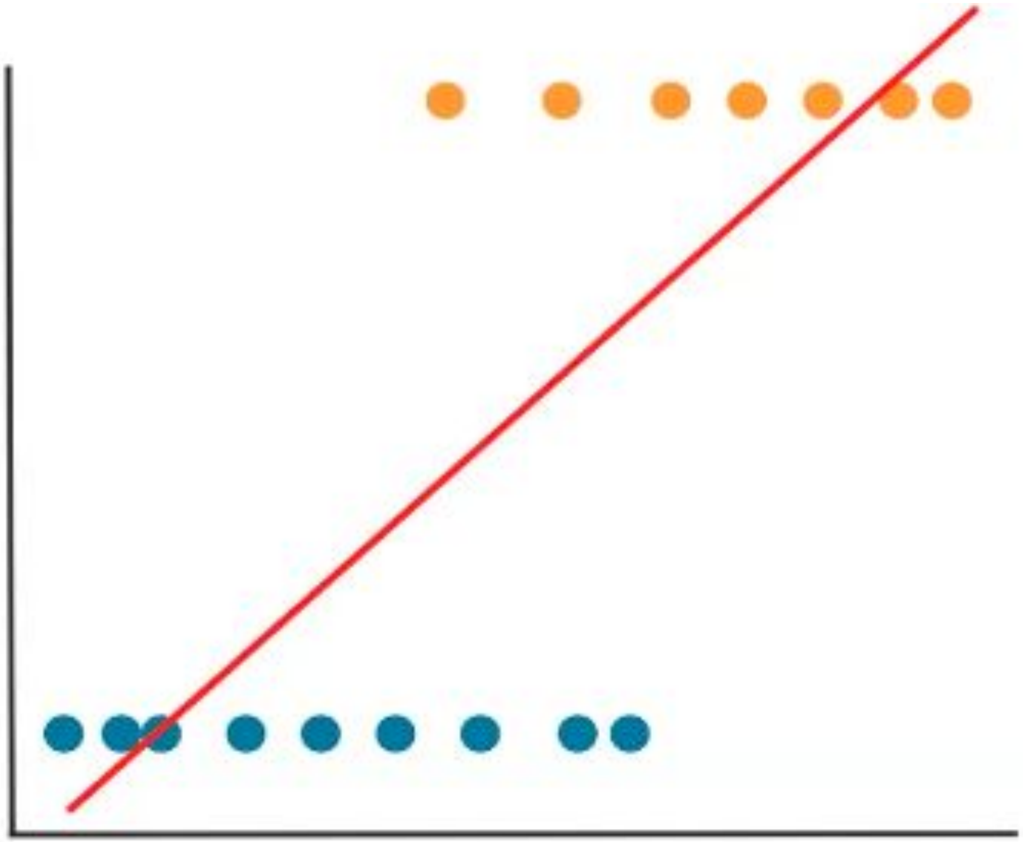
3. Know how to visualize the performance of a classification model using Receiver Operating Characteristics (ROC)

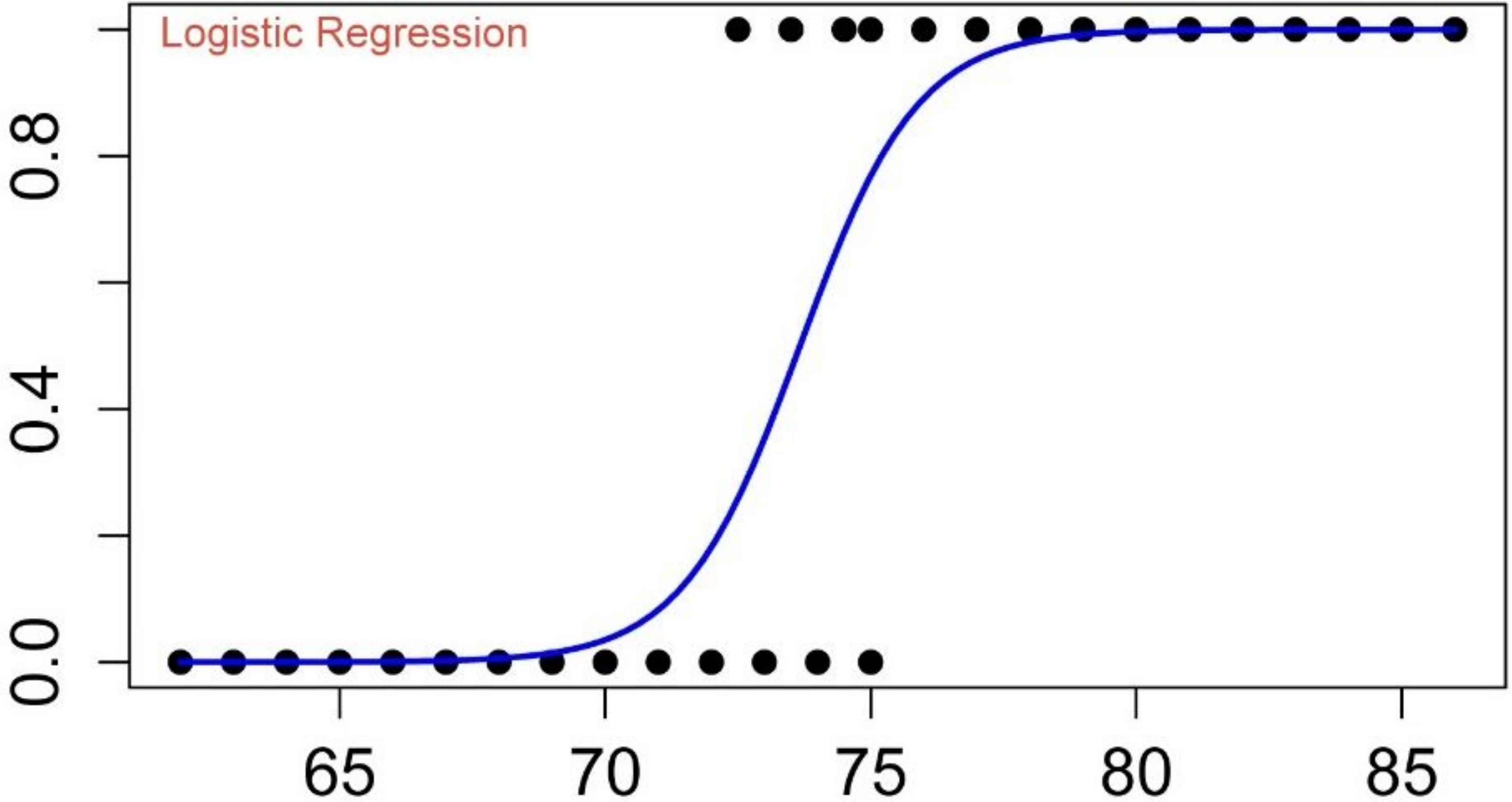
4. Evaluate classification models using Area Under the Curve (AUC)

5. Implement different methods when dealing with Class Imbalances

Logistic Regression

- Commonly used algorithm used to model binary classification problems
- Uses a sigmoid function which bounds the output between 0 and 1.
- Logistic Regression equation: $log(\frac{P(Y = 1)}{1 - P(Y = 1)}) = \beta_1 x + \beta_0$




Logistic Regression
0.0
0.4
0.8
65
70
75
80
85

Logistic Regression

Advantages

- Able to interpret how the model makes predictions
- Model training and predictions are relatively fast
- Can perform well with a small number of observations
- No tuning is usually needed for the model unless you want to regularize the model

Disadvantages

- Requires feature scaling
- Binary classification algorithm does not work for multiclass problems

Age	Exposure	Has Disease
37	Yes	No
52	No	No
48	No	No
24	Yes	No
13	No	No
78	Yes	Yes
28	No	No
5	No	No
18	No	No
63	No	No

Dealing w/Class Imbalance

Class Imbalance Options

- Assign each observation a weight
  - Use class_weight parameter

- Oversampling the minority class
  - Smote (Synthetic Minority Oversampling Technique)

- Undersampling the majority class

- Combine oversampling and undersampling to get a balanced dataset

ROC AUC

ROC AUC - Receiver Operating Characteristics - Area Under the Curve

- A way to visualize performance of a classification model at varying thresholds.
- A plot of the true positive rate (TPR aka Recall) vs. the false positive rate (FPR which is (1 - Specificity)).
- The higher the AUC, the better the classifier is doing at predicting the difference between the classes.

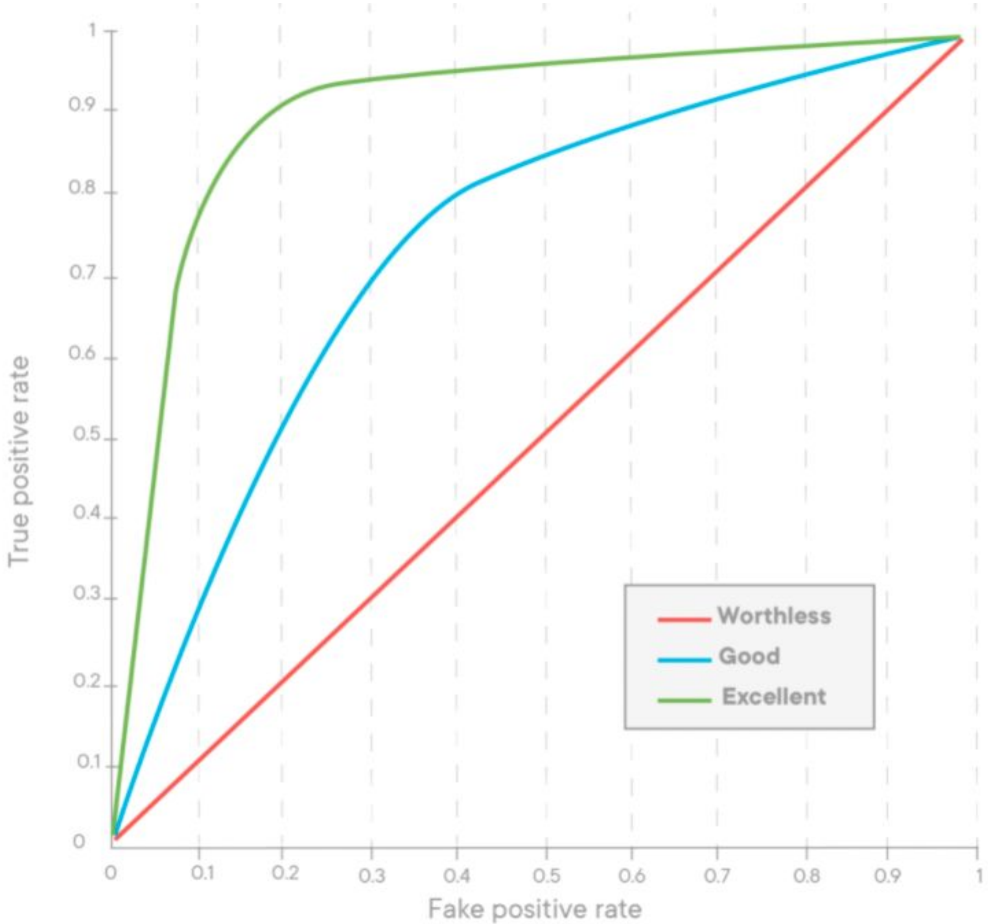
Recall

$$\frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

Specificity

$$\frac{\text{True Negatives}}{\text{True Negatives + False Positives}}$$

Comparing ROC Curves



CodeAlong Notebook

Challenge Notebook

Challenge Data