UNBALANCED DATA

THE PROBLEM OF UNBALANCED DATA

UNBALANCED DATA

- You have a binary outcome
- One outcome (high risk) is much less frequent than the other (low risk)
- Say only 5% of people are high risk
- I want to improve my prediction of high risk individuals using ML

UNBALANCED DATA

- I'll give you a "stupid" prediction rule

  Call everyone low risk

  - This predictor is 95% accurate!!!

HOW TO APPROACH THIS PROBLEM

DATA MANIPULATION

- Classifiers work better if classes are balanced
- You can downsample the majority group to match the minority group
- Fit a classifier
- Repeat and aggregate predictions to get average prediction for the minority class

ARE WE MEASURING PERFORMANCE APPROPRIATELY?

AN EXPLORATION OF DIFFERENT METRICS

1. Accuracy : What proportion of predictions are correct
2. Precision: What proportion of positives are true
3. Recall: What proportion of true positives are called positive
4. AUC : Area under the receiver operating characteristic (ROC) curve

- ROC maps (1-specificity) against sensitivity for different cutoffs
- Sensitivity = Recall
- Specificity = What proportion of true negatives are called negative

5. F1 score: harmonic mean of Precision and Recall
6. Brier score: Mean squared error of probability

predictions

EVALUATING THE CONFUSION MATRIX

	Predict -ve	Predict +ve
True -ve	TN	FP
True +ve	FN	TP

- Precision = TP/(TP+FP)
- Recall = TP/(TP+FN) = Sensitivity
- Specificity = TN/(TN+FP)
- F1 = 2(precision * recall)/(precision + recall)