Evaluation of Classification Models

datasciencedojo
unleash the data scientist in you

Limitation of Accuracy

- Consider a 2-class problem:
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10
- If the model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading!

Measuring model performance

- Problem domain and business needs will decide what metric to use for measuring model performance
- Do you always want your model to be accurate?

Classifier Evaluation

- Metrics for Performance Evaluation

How to evaluate the performance of a model?

- Methods for Performance Evaluation

How to obtain reliable estimates?

- Methods for Model Comparison

How to compare the relative performance among competing models?

Model Evaluation

- **Metrics for Performance Evaluation**

How to evaluate the performance of a model?

- Methods for Performance Evaluation

How to obtain reliable estimates?

- Methods for Model Comparison

How to compare the relative performance among competing models?

Metrics for Performance Evaluation

- Focus on the predictive capability of a model
  - Rather than how fast it takes to classify or build models, scalability, etc.
- Confusion Matrix:

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

a: TP (true positive)

b: FN (false negative)

c: FP (false positive)

d: TN (true negative)

Metrics for Performance Evaluation

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a (TP)	b (FN)
	Class=No	c (FP)	d (TN)

Most widely-used metric:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

datasciencedojo
unleash the data scientist in you

Cost-Sensitive Measures

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

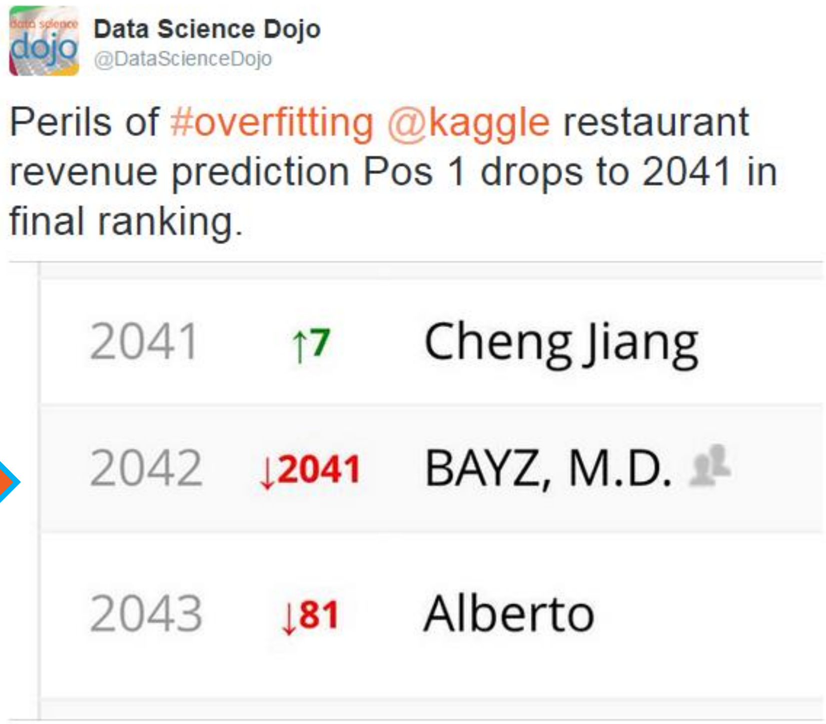
$$\text{F - measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

- Precision is biased towards C(Yes|Yes) & C(Yes|No)
- Recall is biased towards C(Yes|Yes) & C(No|Yes)
- F-measure is biased towards all except C(No|No)

$$\text{Weighted Accuracy} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

WILL MY MODEL BETRAY ME?

Perils of Overfitting

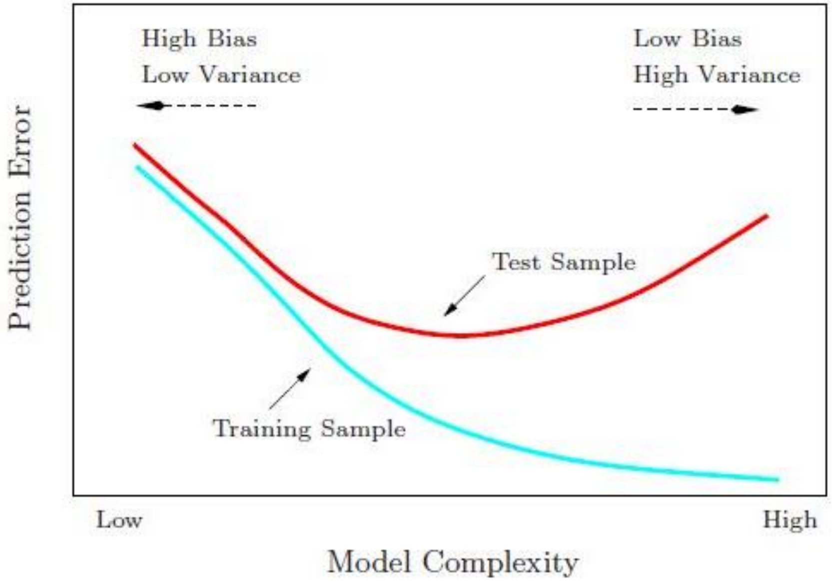


Overfitting

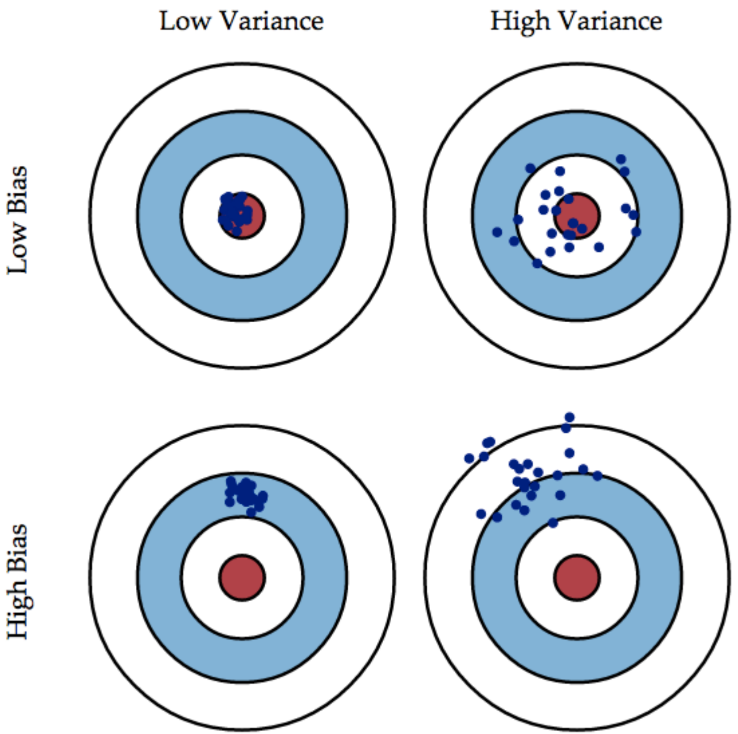
- The gravest and most common sins of machine learning
- Overfitting is when you try to learn so much from data that you memorize it.
  - You do well on training data
  - But don't do well (or even fail miserably) on test data

Bias/Variance Tradeoff

- You can beat your data to confess anything



Bias/Variance Tradeoff



Methods of Estimation

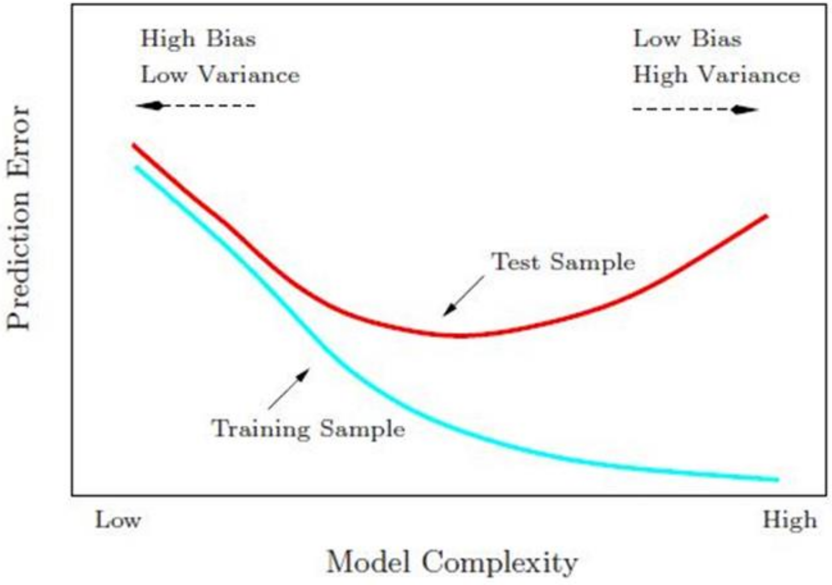
- Holdout
  - Reserve 2/3 for training and 1/3 for testing
- Cross validation
  - Partition data into k disjoint subsets
  - k-fold: train on k-1 partitions, test on the remaining one
  - Leave-one-out: k = n
- Random subsampling
  - Repeated holdout
- Stratified sampling
  - Oversampling vs undersampling
- Bootstrap
  - Sampling with replacement

You have done everything

- Data is clean
- Missing values, noise etc. are dealt with
- Features engineered
- Right metric has been chosen
- Model is trained.
- What is the next step?

Now tune the parameters

- You will tune the parameters until you get the right trade-off between bias and variance



Model Evaluation

- Metrics for Performance Evaluation

How to evaluate the performance of a model?

- Methods for Performance Evaluation

How to obtain reliable estimates?

- **Methods for Model Comparison**

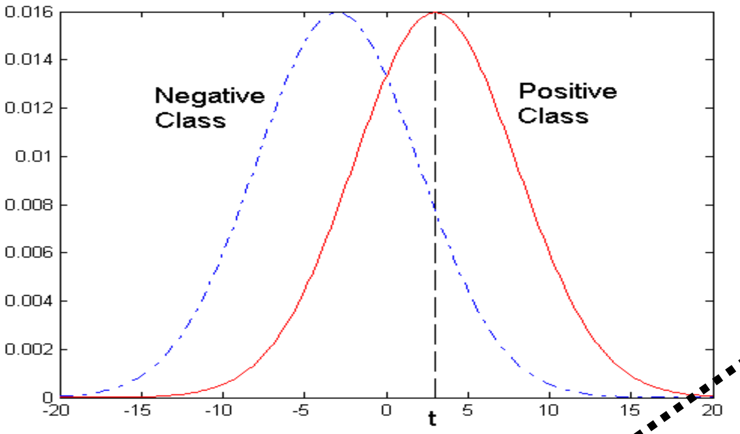
How to compare the relative performance among competing models?

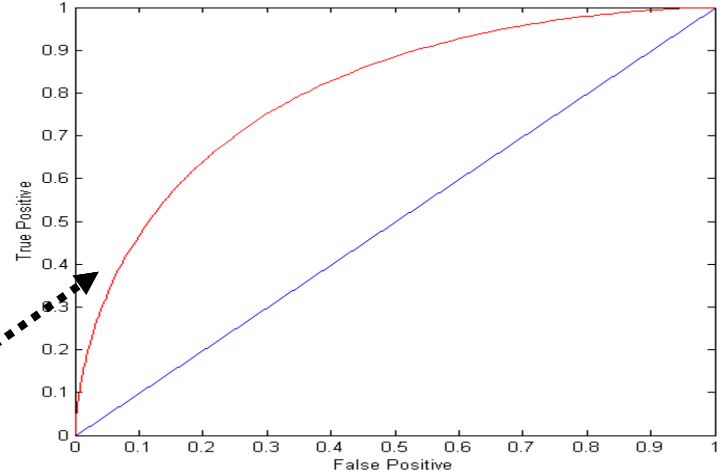
ROC (Receiver Operating Characteristic)

- Developed in 1950s for signal detection theory to analyze noisy signals
  - Characterize the trade-off between positive hits and false alarms
- ROC curve plots TP (on the y-axis) against FP (on the x-axis)
- Performance of each classifier represented as a point on the ROC curve
  - Changing the threshold of the algorithm, sample distribution, or cost matrix changes the location of the point

ROC Curve

- 1-dimensional data set containing 2 classes (positive and negative)
- Any points located at x > t are classified as positive





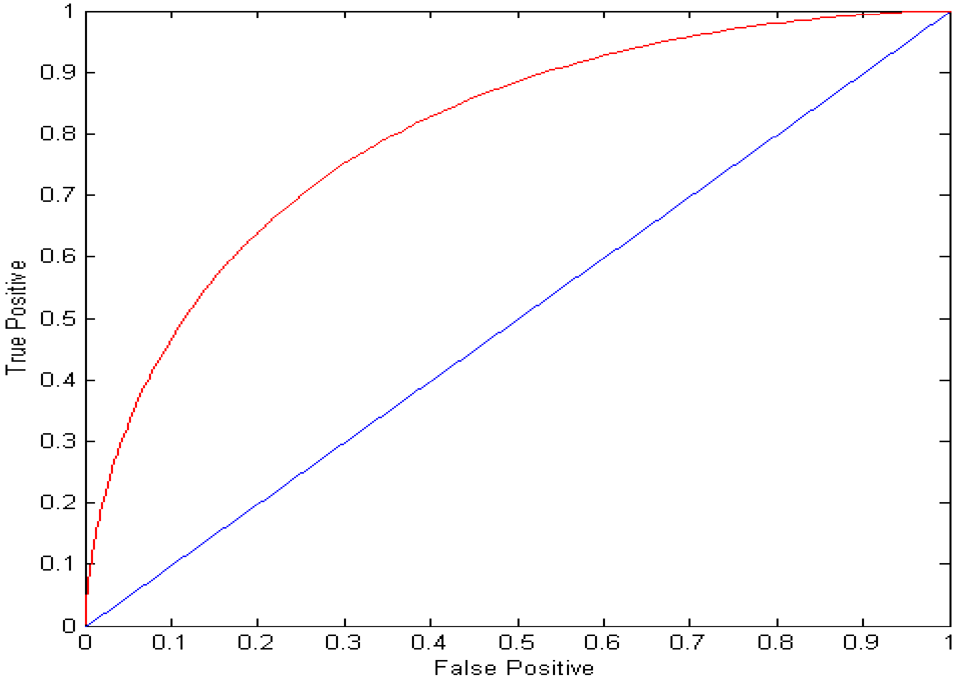
At threshold t:

TP=0.5, FN=0.5, FP=0.12, FN=0.88

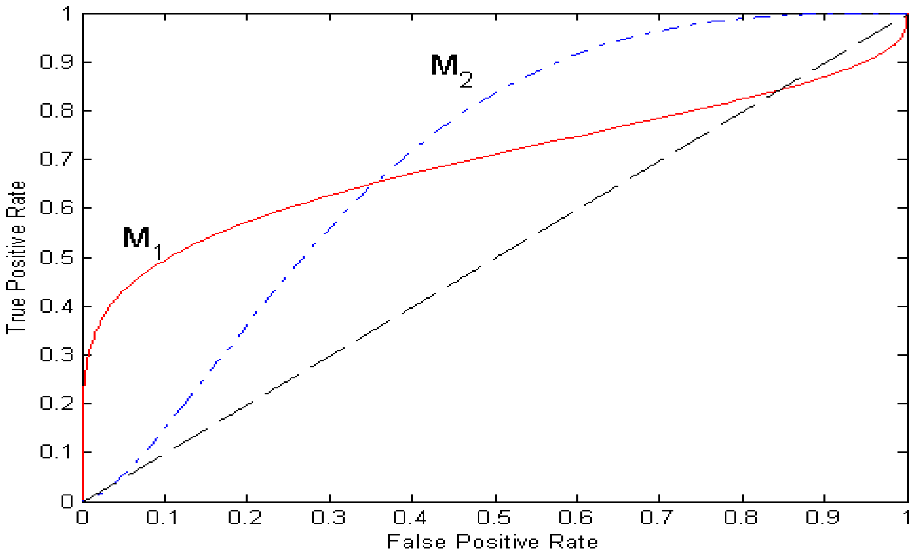
ROC Curve

- (TP,FP):
- (0,0): declare everything
  to be negative class
- (1,1): declare everything
  to be positive class
- (1,0): ideal

- Diagonal line:
  - Random guessing
  - Below diagonal line:
    - Prediction is opposite of the



Using ROC for Model Comparison



- No model consistently outperforms the other
  - M_1 is better for small FPR
  - M_2 is better for large FPR
- Area under the ROC curve
  - Ideal:
    - Area = 1
  - Random guess:
    - Area = 0.5

QUESTIONS