HUDK 4050: CORE METHODS IN EDM

11/12/19 3:40 PM

In the news

EDUCATION DIVE

To feed today's 'on-demand' students, colleges turn to robots and mobile apps

EdSurge

ISTE to Acquire EdSurge, in Move to Nonprofit

EDWEEK Market Brief

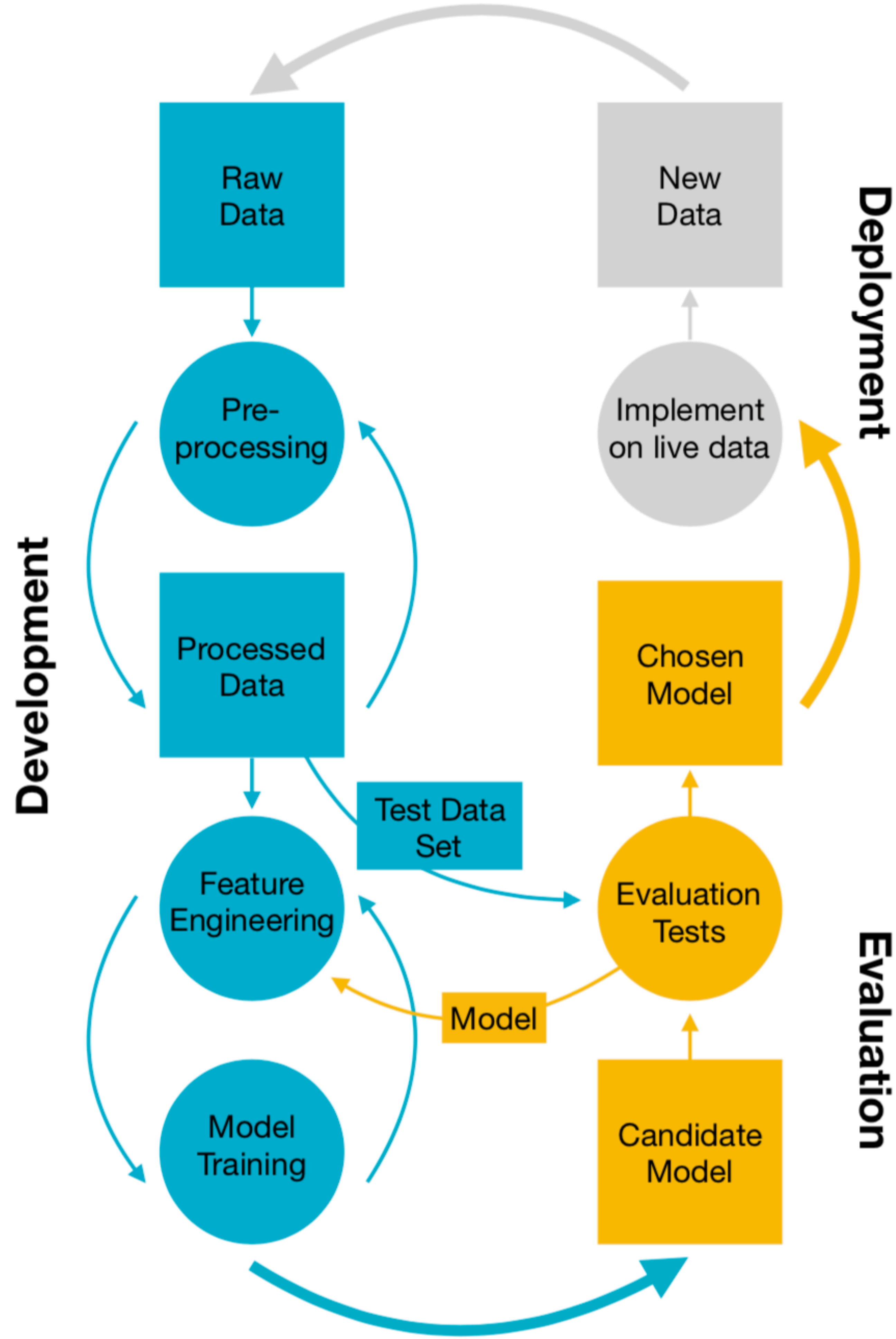
Amazon, Already a Presence in K-12, Expands Its Support for Ed-Tech Startups

INSIDE HIGHER ED

A Fresh Look at Blockchain in Higher Ed

Events

Title	Date - Time	Location
Why We Need Learning Engineers	11/13	Online
DSI Town Hall	11/15 - 10:00am	Davis Auditorium
LearnLaunch 2020 Volunteer	11/15	Online
The Global Education Conference	11/18-11/20	Online
Population Genetics in An Era of Precision Medicine	11/19 - 11:00am	Davis Auditorium
HUD Happy Hour	11/26 - 5:00pm	E’s Bar
Science Communication Workshop	11/20 - 9:30am	Low Library
Citizens and Technology Summit	11/25	Ford Foundation
Reinventing Privacy Workshop	12/11 - 9:00am	TBD


Raw Data
New Data
Deployment
Pre-processing
Implement on live data
Development
Processed Data
Chosen Model
Test Data Set
Feature Engineering
Evaluation Tests
Model
Evaluation
Model Training
Candidate Model

Chapter 6

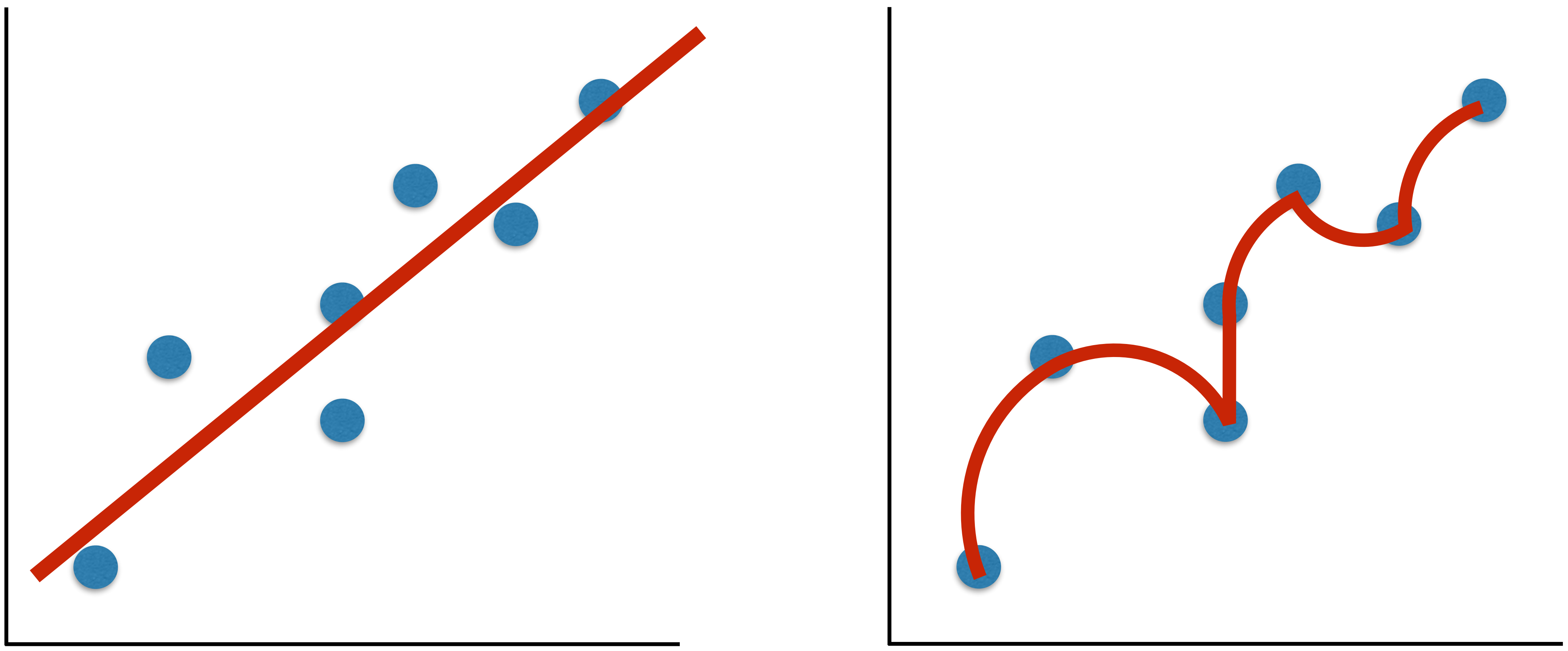
Handbook of Learning Analytics

First Edition

Going Beyond Better Data Prediction to Create Explanatory Models of Educational Data

Ran Liu & Kenneth R. Koedinger

Cross Validation

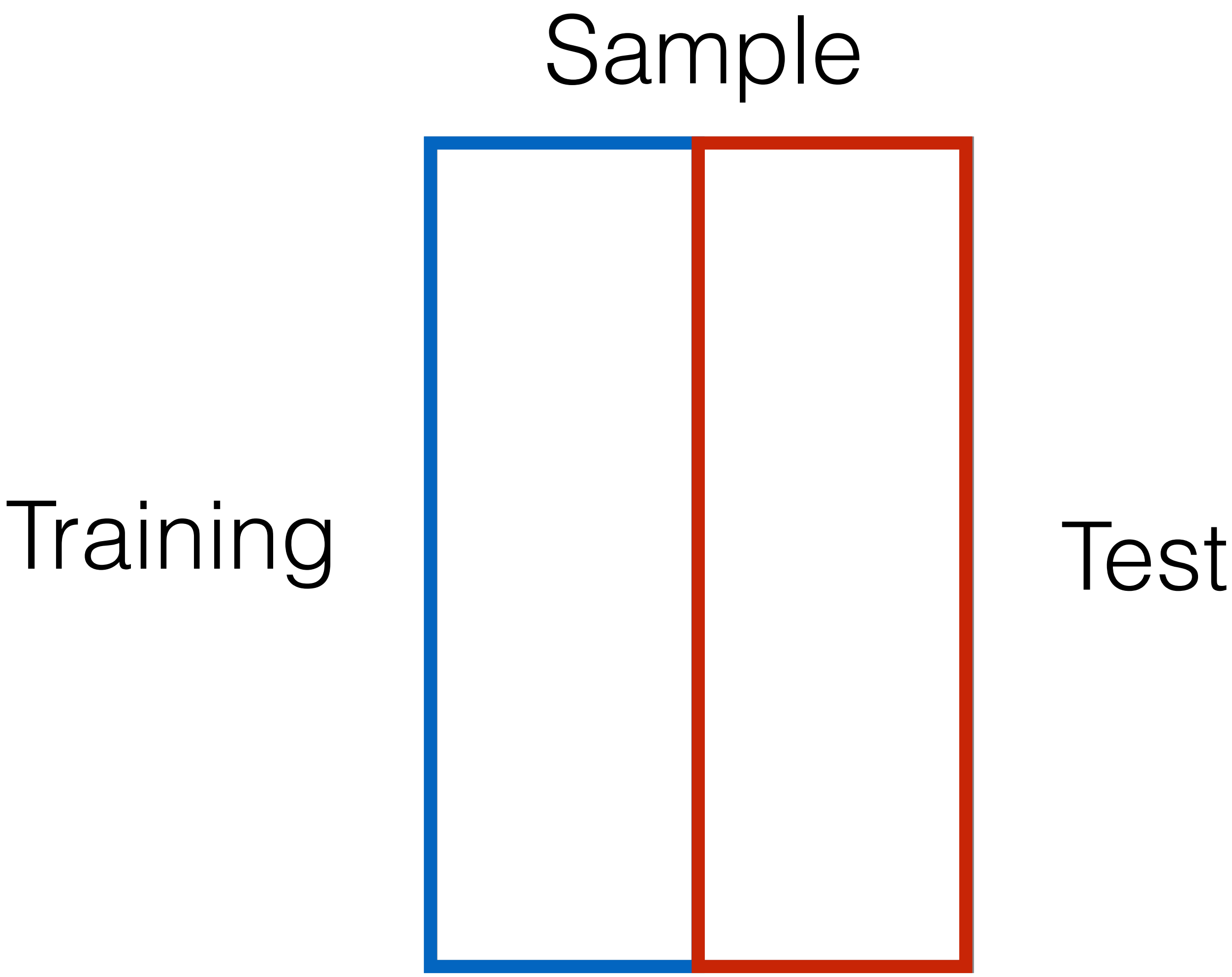


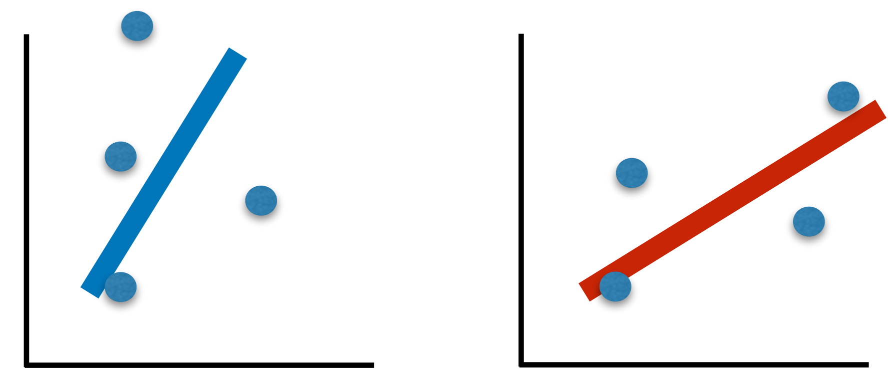
Cross Validation

- Estimate how accurately a predictive model will perform in practice

- Give an insight on how the model will generalize to an independent dataset

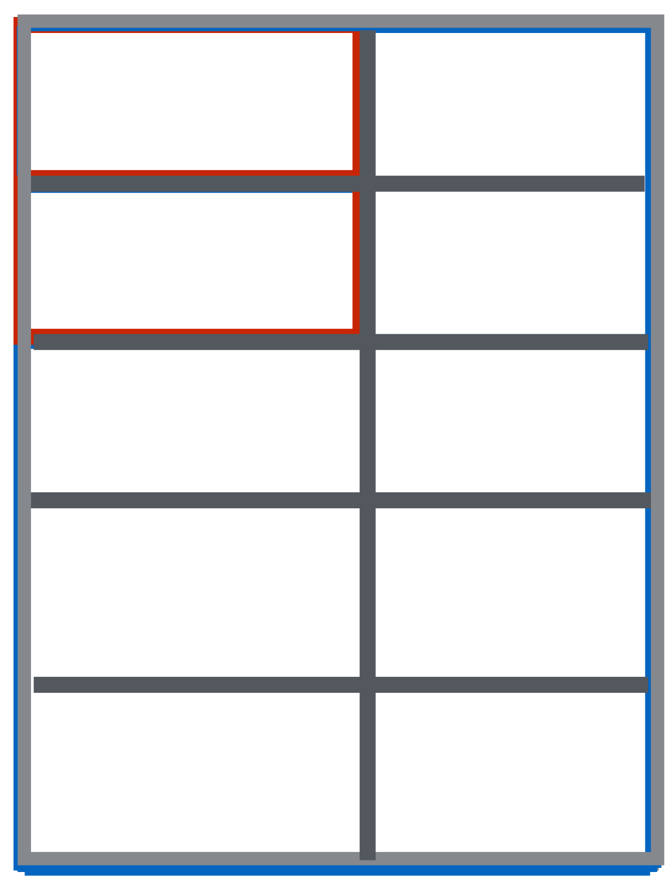
Hold-out Validation





Problem: very dependent on which data are in each group

K-Fold Cross Validation



M1 Error	M2 Error
5	2
4	2
3	1
5	4
4	2
4.2	2.2

Calculate how accurate we are in each “fold” and average the answer

Activity

Which make better pets? Dogs or cats?

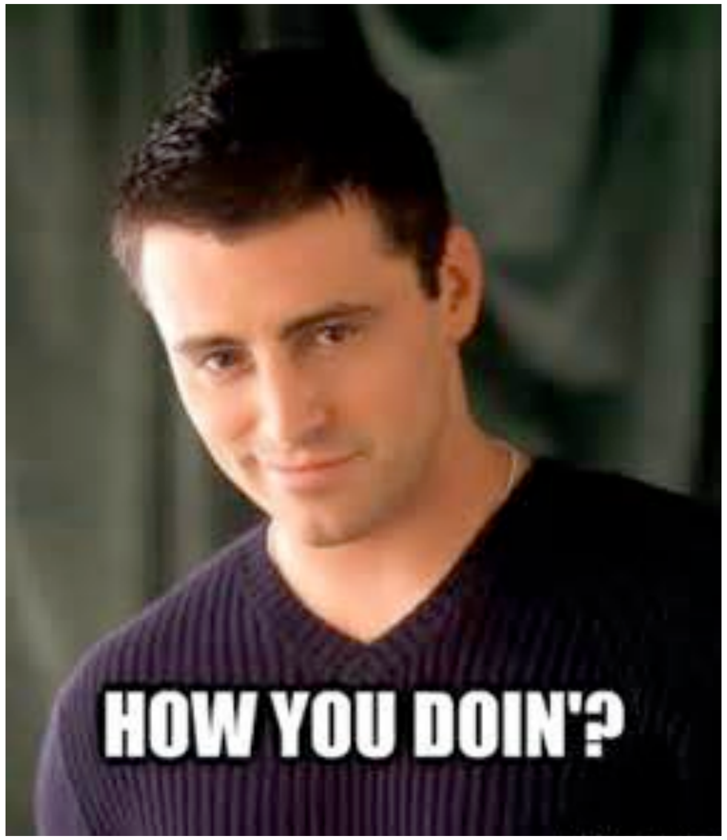
Model	1 Fold	2 Fold	3 Fold
Random	Correct/ Incorrect	Correct/Incorrect 1: Correct/Incorrect 2:	Correct/Incorrect 1: Correct/Incorrect 2: Correct/Incorrect 3:
Majority Class	Correct/ Incorrect	Correct/Incorrect 1: Correct/Incorrect 2:	Correct/Incorrect 1: Correct/Incorrect 2: Correct/Incorrect 3:

RPART

Complexity Parameter **SSE/RMSE** **Cross-Validation Error**

	CP	nsplit	rel error	xerror	xstd
1	0.052	0	1.000	1.072	0.035325
2	0.012	1	0.948	0.992	0.036941
3	0.010	2	0.936	1.004	0.036733

Diagnostic Metrics



How to determine how well your model is doing

Diagnostic Metrics

Classification

- Accuracy
- Cohen's Kappa
- ROC/AUC/A'
- Correlation

Regression

- MAE/RMSE
- Pearson's Correlation/R^2
- AIC/BIC

Terms

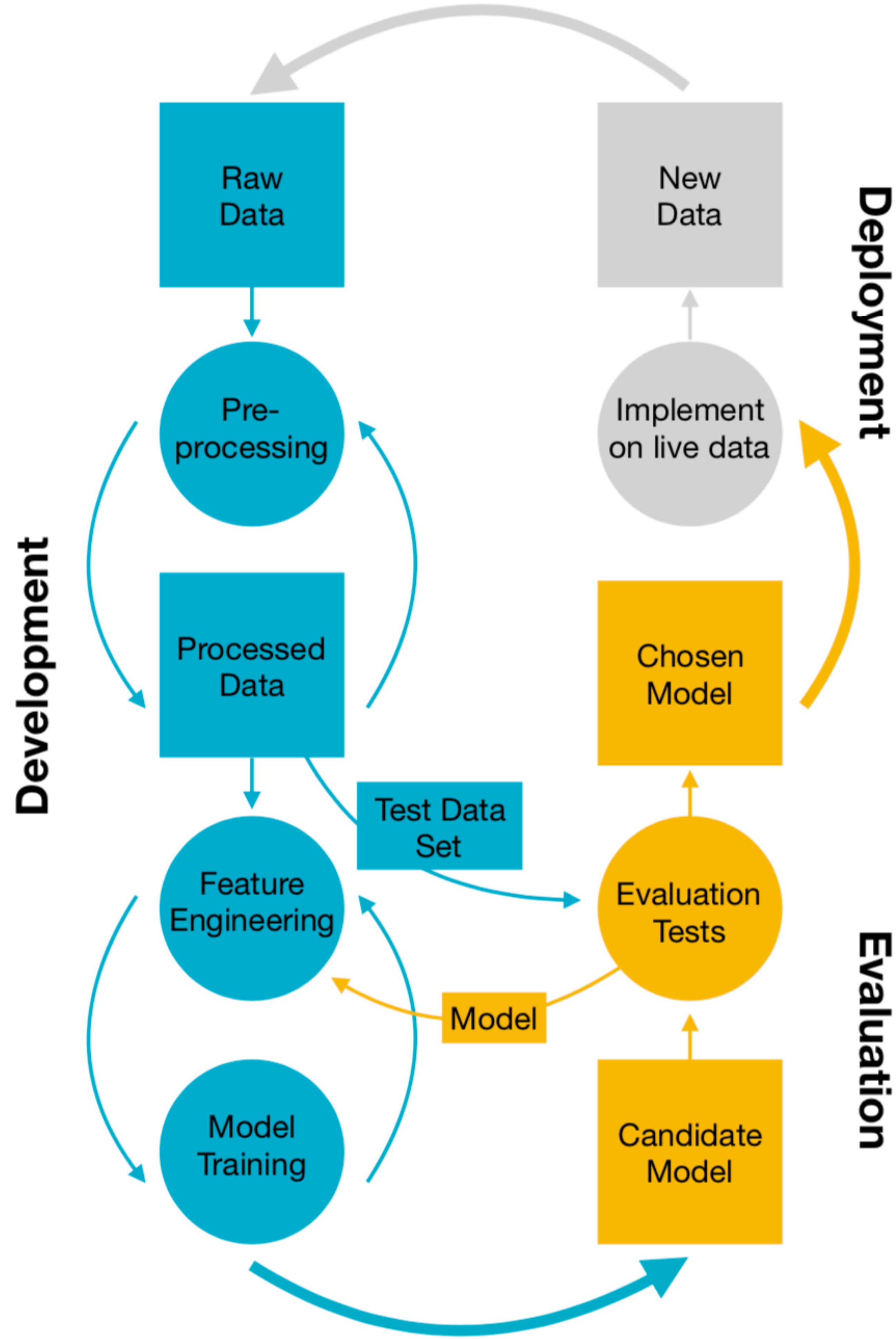
- **Ground truth**: data that is available, relevant, and most trustworthy to train your model

- **Baseline**: initial measurement

- **Gold standard**: (expensive) comparative measurement

- **Inference**: data that is inferred from logic + data

Diagnostics for Classifiers


Raw Data
New Data
Deployment
Pre-processing
Implement on live data
Development
Processed Data
Chosen Model
Test Data Set
Feature Engineering
Evaluation Tests
Model
Evaluation
Model Training
Candidate Model

Accuracy

- $\dfrac{\text{correct predictions}}{\text{total predictions}}$
- Gotcha: unequal categories
- EG - Predicting fraudulent credit card transactions
- False positives/negatives (over/under predict)

Precision & Recall/Sensitivity

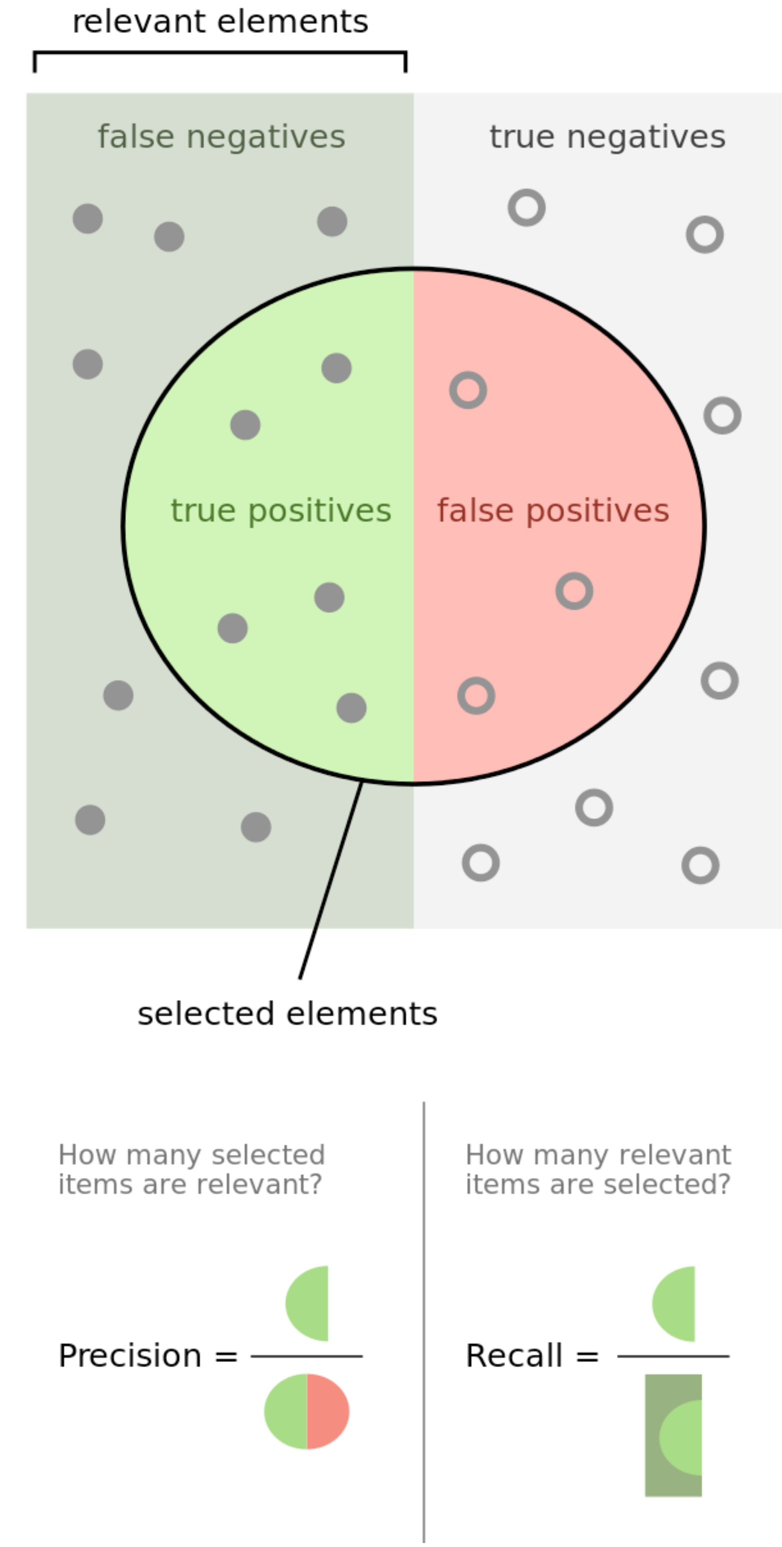
$$\textbf{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\textbf{Recall/Sensitivity} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

Precision & Recall

Precision
The fraction (probability) of predictions that are ***relevant***

Recall/Sensitivity
The fraction (probability) of relevant instances that are ***predicted***



Walber, 2014

F1-Score

$$\textbf{F1} = \frac{\text{Recall x Precision}}{(\text{Recall + Precision})}$$

*Harmonic mean of precision & recall

Specificity

$$\textbf{Specificity} = \frac{\text{True Negative}}{\text{True Negative + False Positive}}$$

Cohen's Kappa (κ)

- Traditionally used for inter-rater reliability
- We will use it to look at the reliability between the data and our model

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Observed Agreement

Expected Agreement

(hypothetical probability of chance agreement)

		Model		
		Yes	No	
Data	Yes	4	2	6
	No	3	3	6
		7	5	12

$$p_o = (4 + 3)/12 = 0.58$$

$$p_e = (7/12) \times (6/12) + (5/12) \times (6/12) = 0.5$$

$$\kappa = (0.58 - 0.5)/(1 - 0.5) = 0.16$$

Is this good? Depends on the context

Gotchas with Kappa

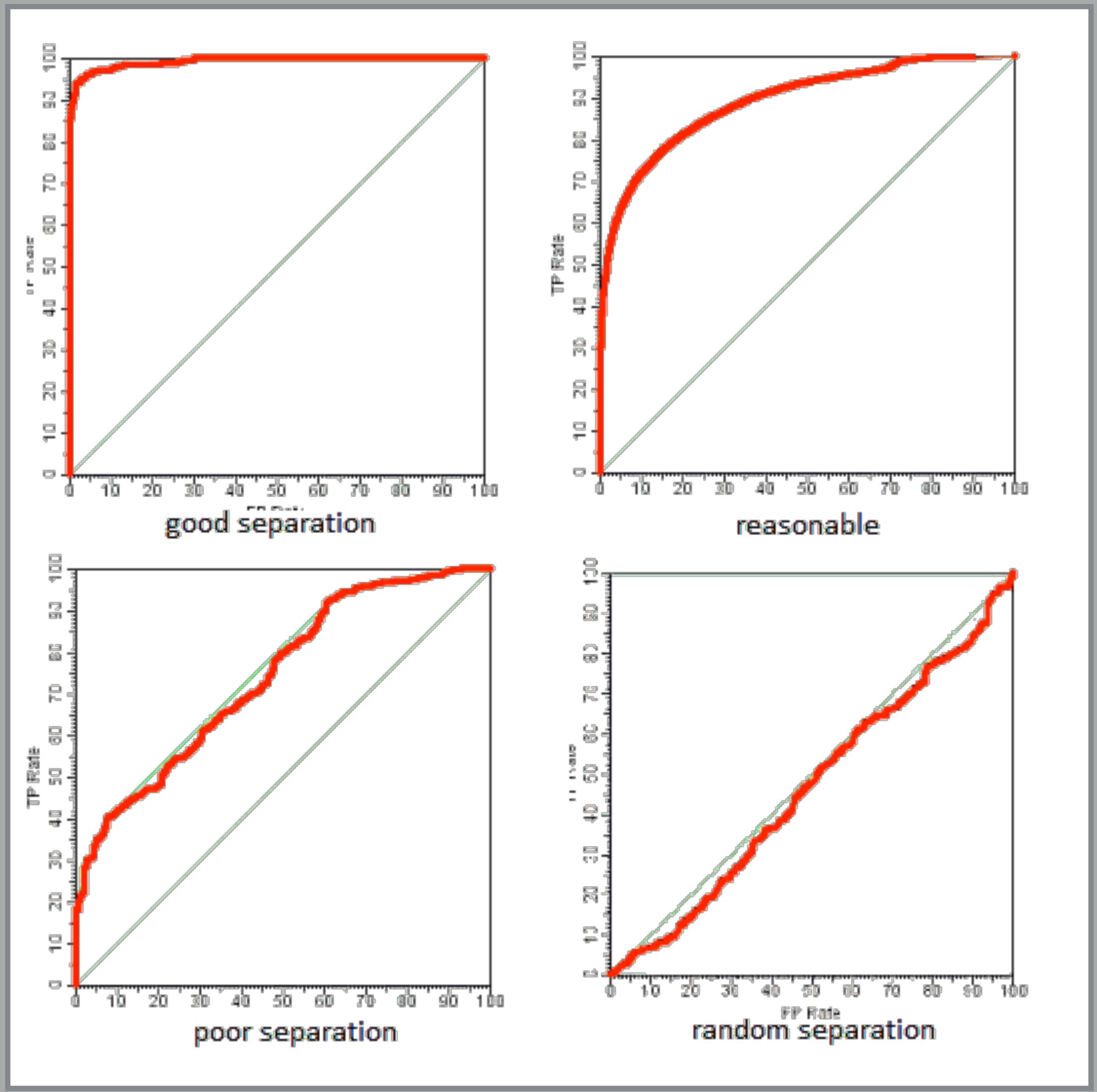
- Again, rare categories pose a problem and will incur a higher penalty than common categories
- Does the marginal probability represent “chance”?

Probabilities

- Model assigns a probability of belonging to a class, rather than a class directly
- Then choose a probability threshold to assign to a class
- Allows us to choose a preference based on the consequences of false positives/negatives
- http://www.navan.name/roc/

Receiver Operating Characteristic (ROC)

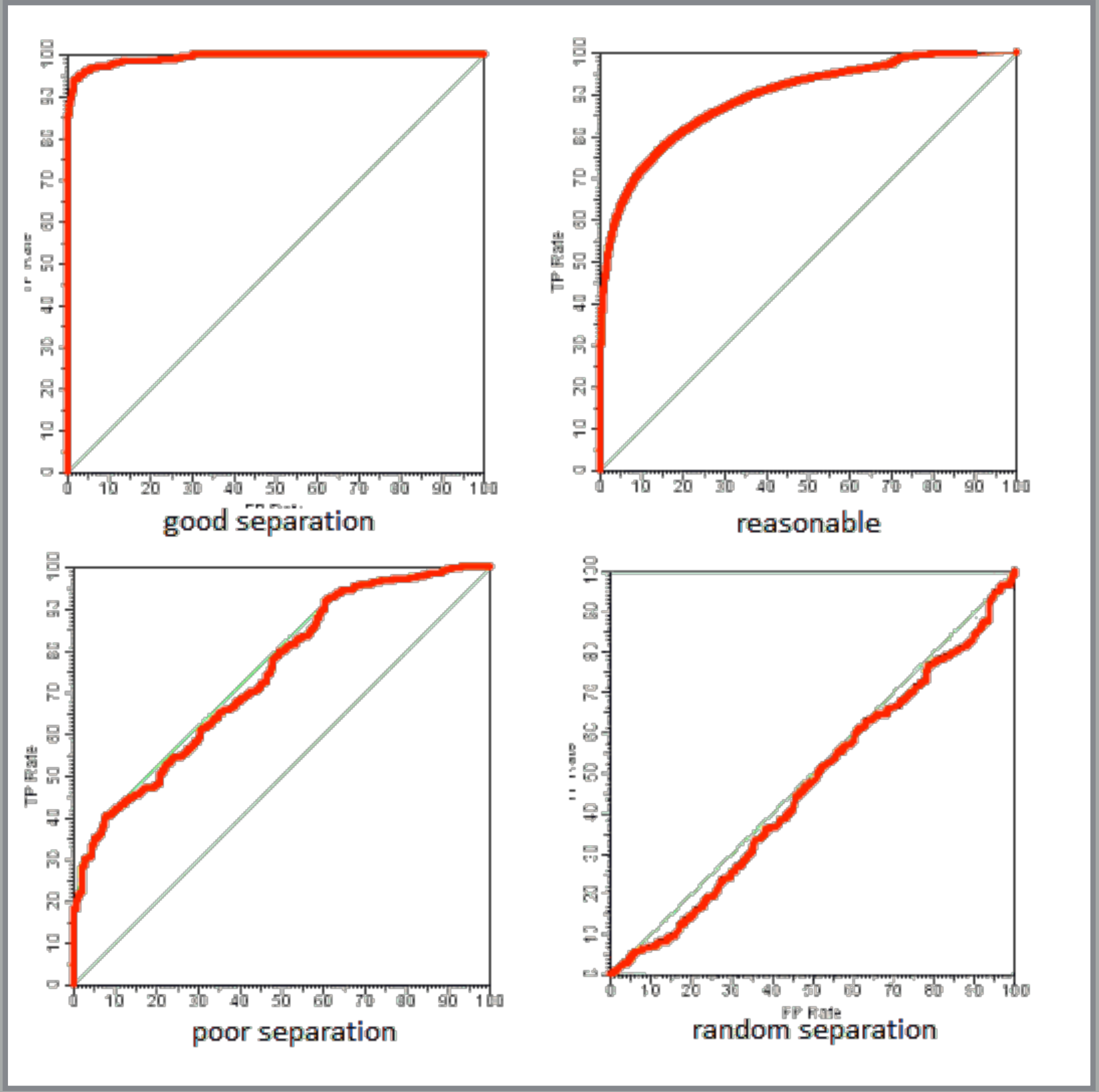
- Relationship between the false positive and the true positive rates
- World War II for detecting enemy objects on radar in response to Pearl Harbor
- Demonstrates the sensitivity vs specificity tradeoff



Receiver Operating Characteristic (ROC)

<u>Area Under the ROC Curve:</u>

- AUC is the collapsed metric to compare models
- AUC is the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one
- It is equivalent to the Wilcoxon Sum-Rank Test and can therefore generate a probability test
- Is sometimes called A' (A Prime) depending on how it is calculated



Diagnostics for Regressors

Mean Absolute Error

- Mean of observed values minus predicted values

$$\text{MAE} = \frac{\sum |x - \bar{x}|}{n}$$

Root Mean Squared Error

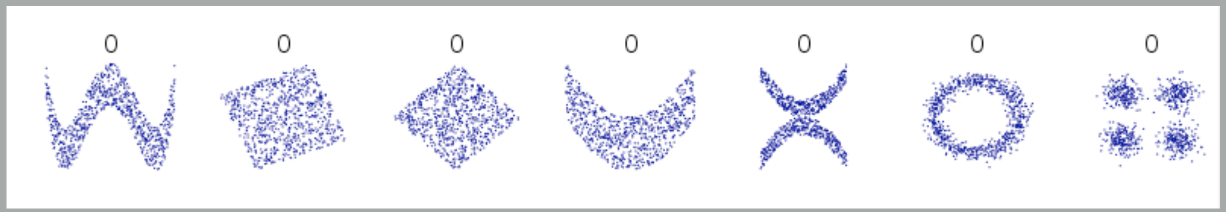
- Square root of the observed values minus predicted values squared

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (p_i - a_i)^2}{n}}$$

Pearson’s Correlation

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Measure of the linear dependence between two variables
- Covariance between two variables divided by the product of the standard deviation of those variables
- Development began ~ 1880s by Galton and then Pearson
- Gotcha: must be a linear relationship



Denis Boigelot, 2011

r^2

- The proportion of the variance in the dependent variable that is predicted from the independent variable
- There are several ways to calculate R^2
- If it involves two variables it is the square of the correlation (OLS classes will go more in depth)

Akaike Information Criterion (AIC)

AIC = number of parameters - goodness of fit

- Developed by Akaike in 1971 based on thermodynamics
- Relative estimate of the information lost when a given model is used to represent the process that generates the data
- Model with lowest AIC “wins”
- Represents the trade off between goodness-of-fit with model complexity
- It compares models, cannot give an estimate of model fit in an absolute sense
- Gotcha: Software implementation was not always reliable

赤池 弘次
Akaike Hirotsugu
(1927-2009)

Bayesian Information Criterion (BIC)

BIC = number of parameters x sample size - goodness of fit

- Developed by Schwarz in 1978
- Uses Bayes Theorem to penalize the addition of parameters
- Penalty for adding parameters is greater than in AIC
- Represents the trade off between goodness-of-fit with model complexity
- Lowest BIC “wins”
- Gotcha: Does poorly when dealing with many parameters

Assignment 6

core-methods-in-edm/
assignment6