Week 3

Concepts in Machine Learning

fredhutch.io

Fred Hutchinson Cancer Research Center

Week 3 Learning Objectives

- Quick CRISP-DM Review
- Supervised Learning vs Unsupervised Learning
  - Regression vs. Classification
    - Binary vs. Multiclass Classification
- Hard vs. Soft Classification
- Evaluating Classification Models
  - Soft vs Hard
- Hard Classification Metrics for Soft Classification Models
  - Confusion Matrix
  - ROC and AUC
- Discussion: Do You Really Need Hard Classification?

Definitions

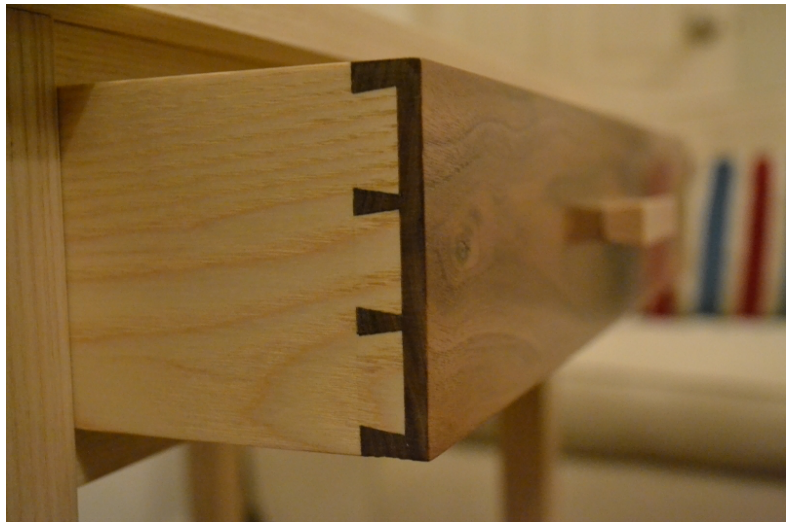
- Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. (https://expertsystem.com/machine-learning-definition/)
- Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead. (https://en.wikipedia.org/wiki/Machine_learning)
- Problem + Data + Algorithm(self-adjusting) + Compute ==> Insight

Experimental Design

- Difficult to master or even do well
- Close interplay between
  - Goals
  - Methods
  - Data
  - Execution
- Requires thoughtful approach and broad understanding

Capable Cabinet Maker

- Inspects and understands raw materials
- Uses the tools thoughtfully to shape and join materials
- Chooses approach and tools based on materials and goals
- Thoughtfulness born of experience



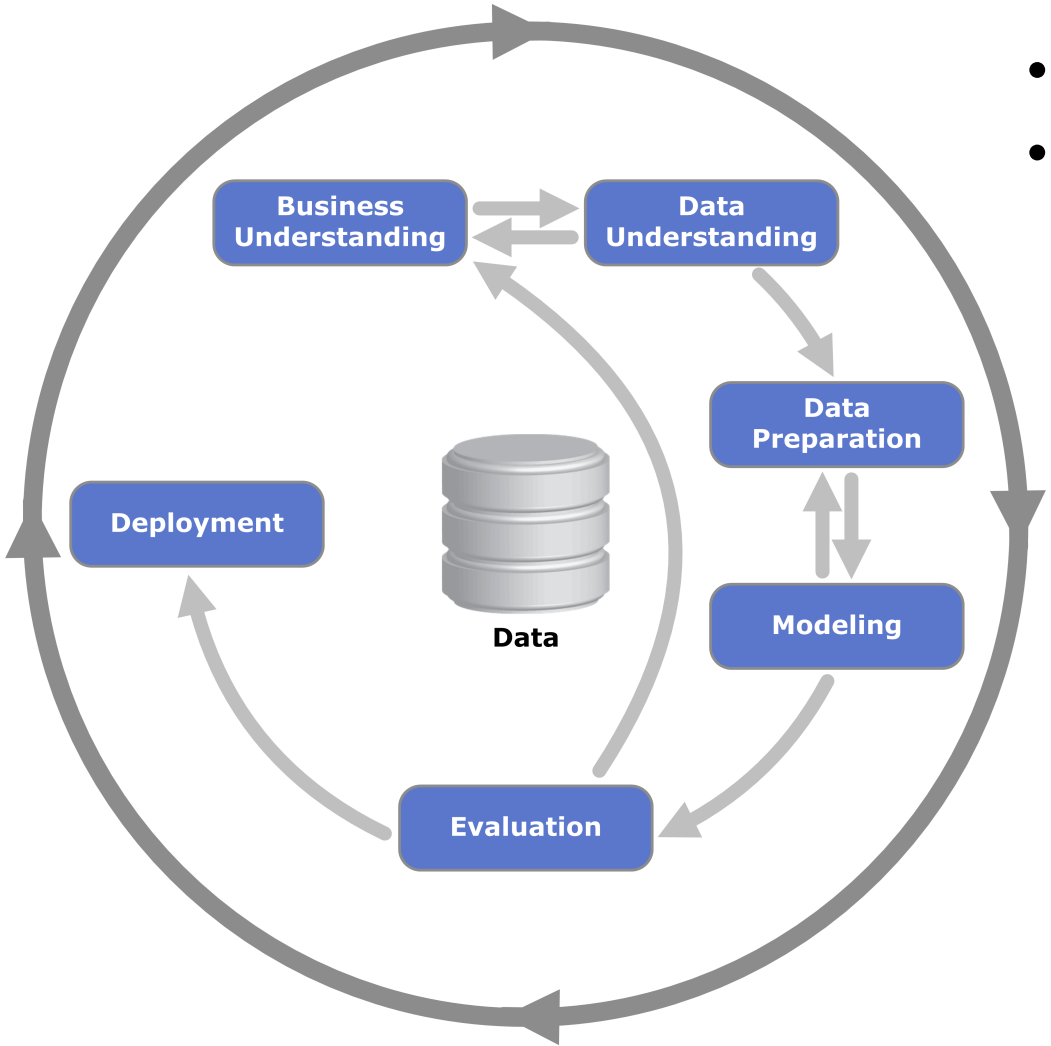
Cabinet Making?

- Storage Need + Raw Materials + Tools + Work ==> Cabinet

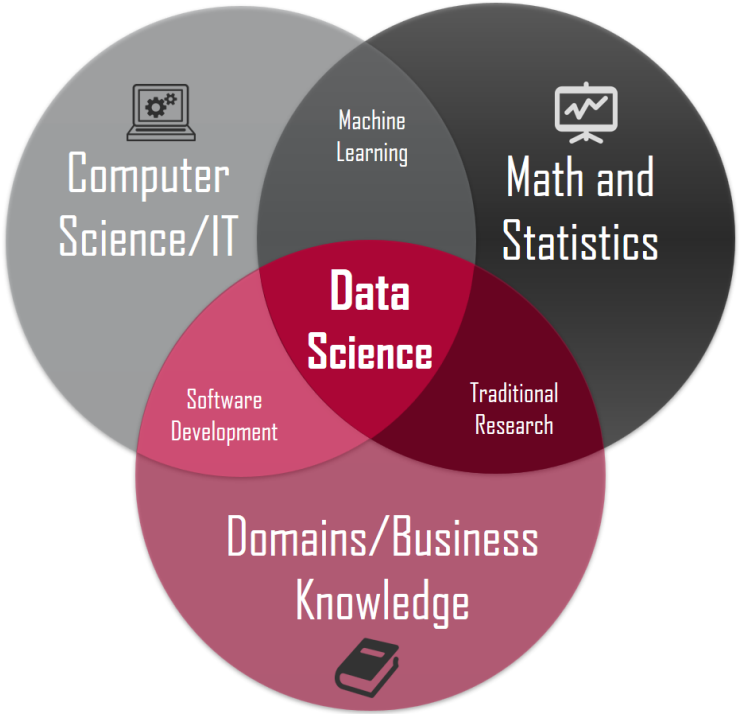
- SAT analogy format

  Tools : Cabinet Making as Algorithms : Machine Learning

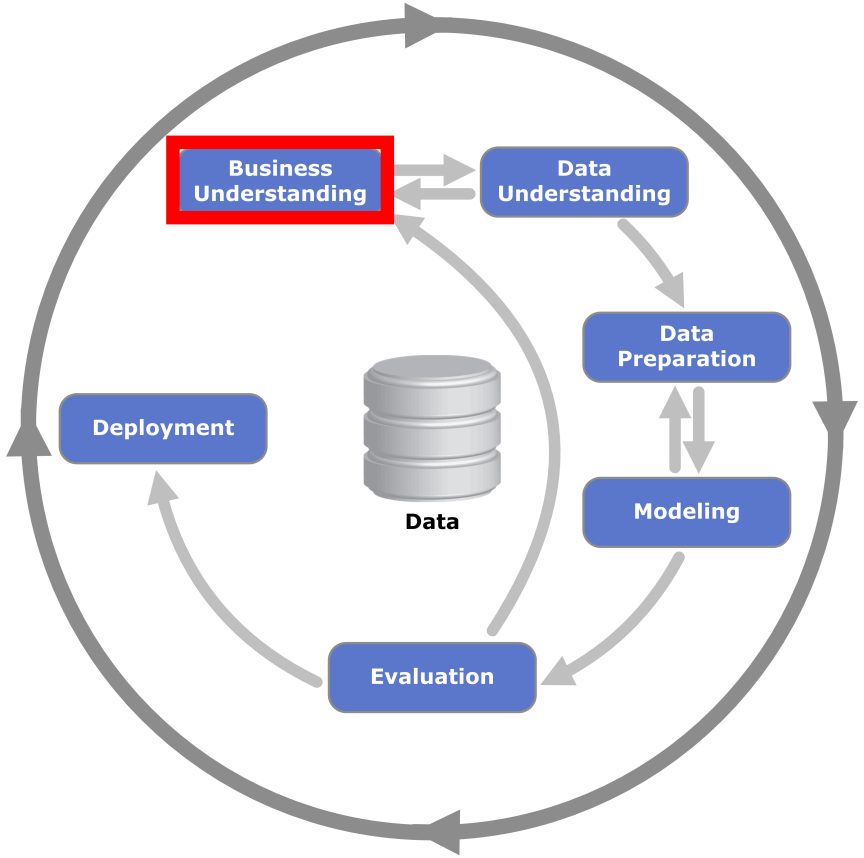
CRISP-DM
Cross-industry standard process for data mining



- Cyclical
- Iterative
- Connecting all 3 areas of the classic notion of “data science”

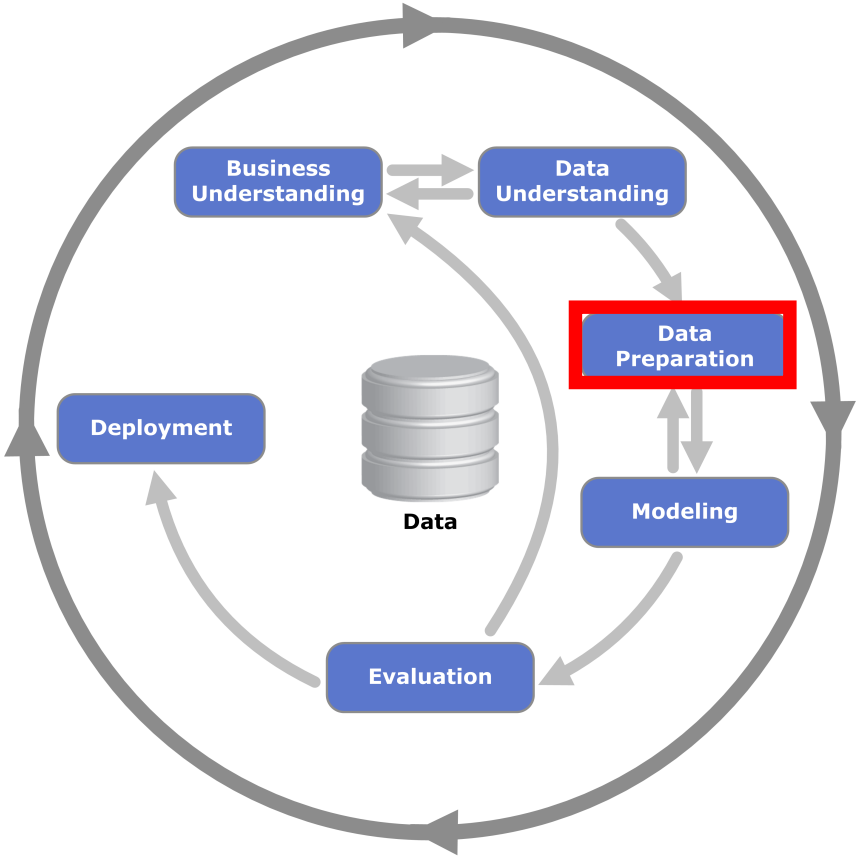


Business/Scientific Understanding



Data Understanding

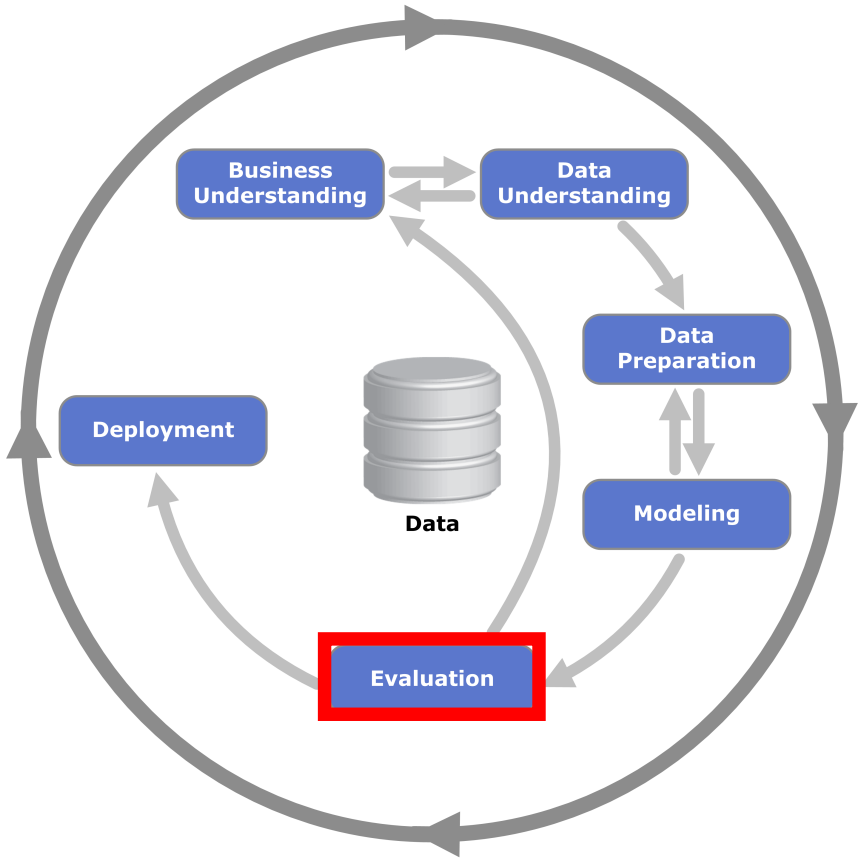
Data Preparation



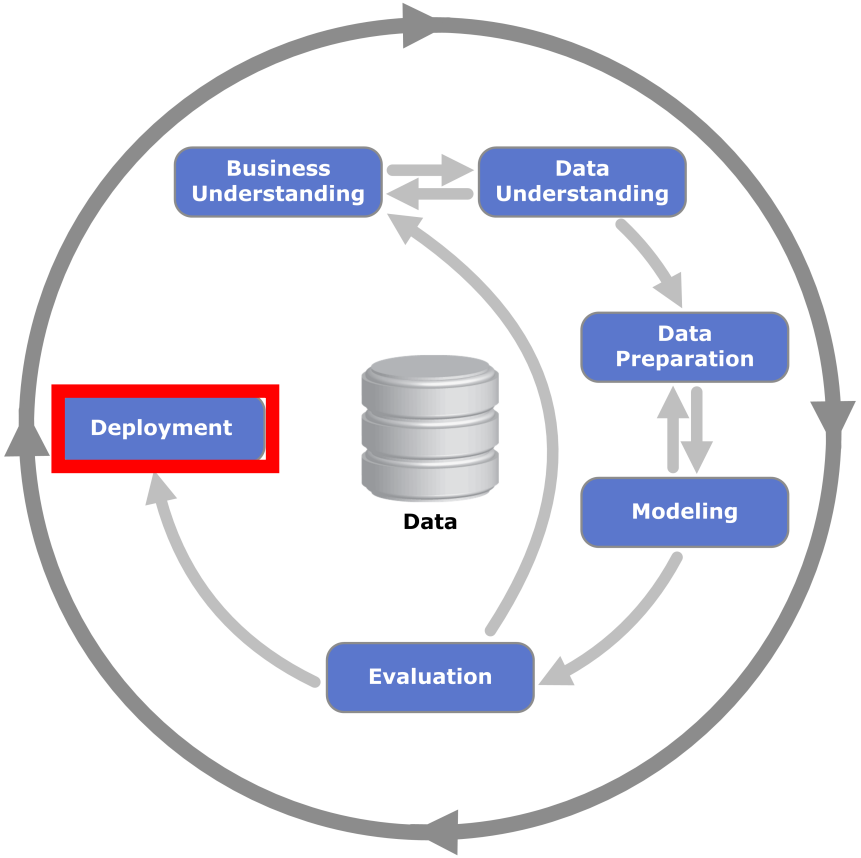
Modeling


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Evaluation

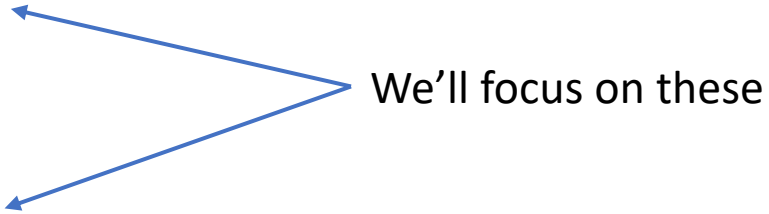

Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Deployment


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

3 or 4 Machine Learning Paradigms

- Supervised Learning
- Unsupervised Learning



- Reinforcement Learning
- Semi-Supervised Learning

3 or 4 Machine Learning Paradigms

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi-Supervised Learning

3 or 4 Machine Learning Paradigms

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning
- Semi-Supervised Learning

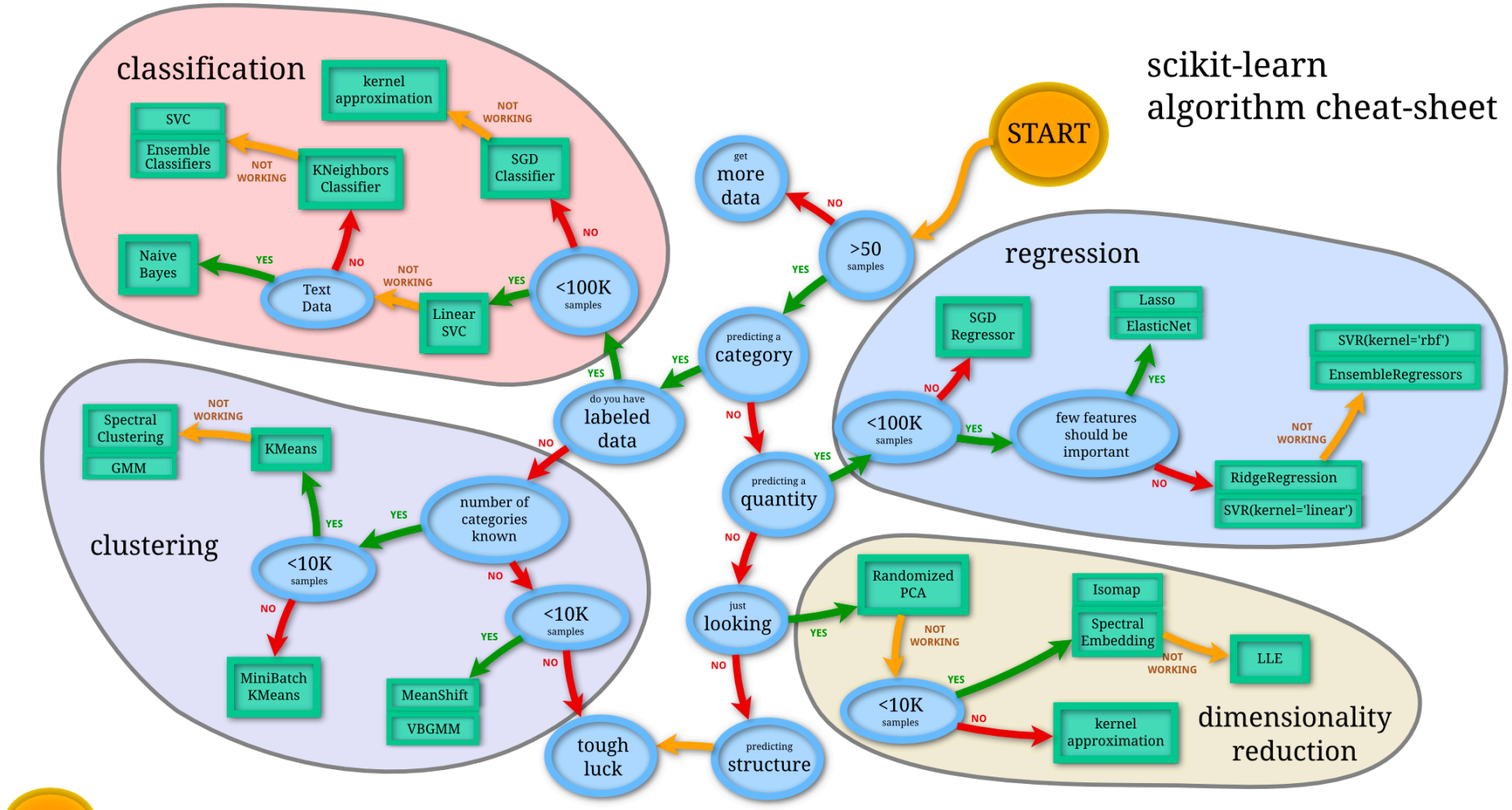
Supervised Problems: Categorical Problems

- Can we state our outcome as a choice between of A **vs.** B? (could be any number of categories)

Supervised Problems: Categorical Problems

- Can we state our outcome as a choice between of A **vs.** B? (could be any number of categories)
  - Yes, today we're interested in categories

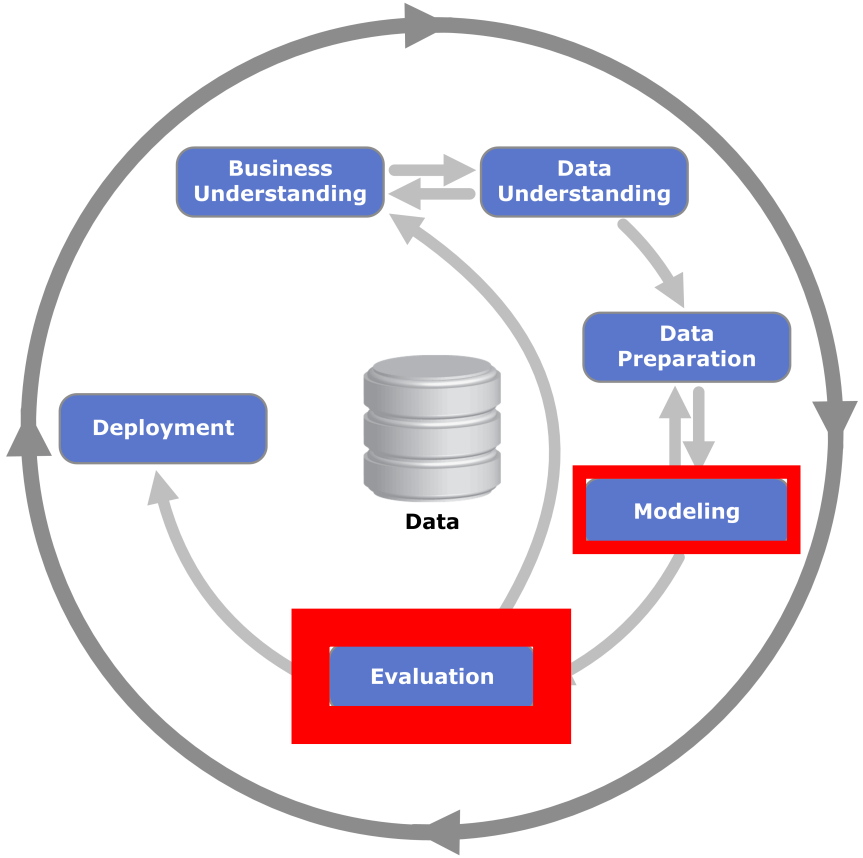
A Nifty Chart



A Nifty Chart

https://scikit-learn.org/stable/tutorial/machine_learning_map/

Modeling & Evaluation



Classification - Binary vs Multiclass

- Classification → Categorical Response
  - red, yellow, or green →Multiclass (class A or B or C or…)
  - dead or alive →Binary (A or B, A or –A)
  - HIV- or HIV+ →Binary (- or +, 0 or 1)

- Binary response → Only two values: 0 or 1
  - Most common case
  - Response 0 → negative class
  - Response 1 → positive class

Classification - Hard vs Soft

- Soft Classification:

  → Attempt to estimate probabilistic information

- Hard Classification:

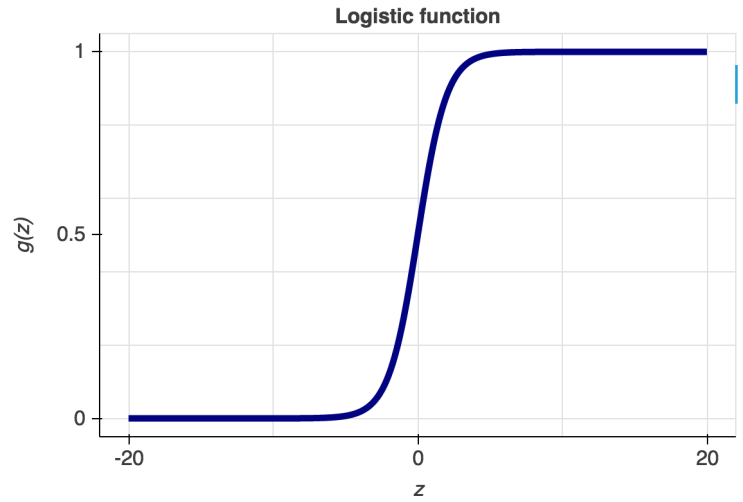
  → Only attempt to estimate class membership

Soft Classification

- Model attempts to estimate the probability that an observation belongs to a class.
  - For binary: $P(y = 1 \mid X)$
  - Conditional probability
    - If we change data X, then estimate of probability changes

- Basic example is Logistic Regression

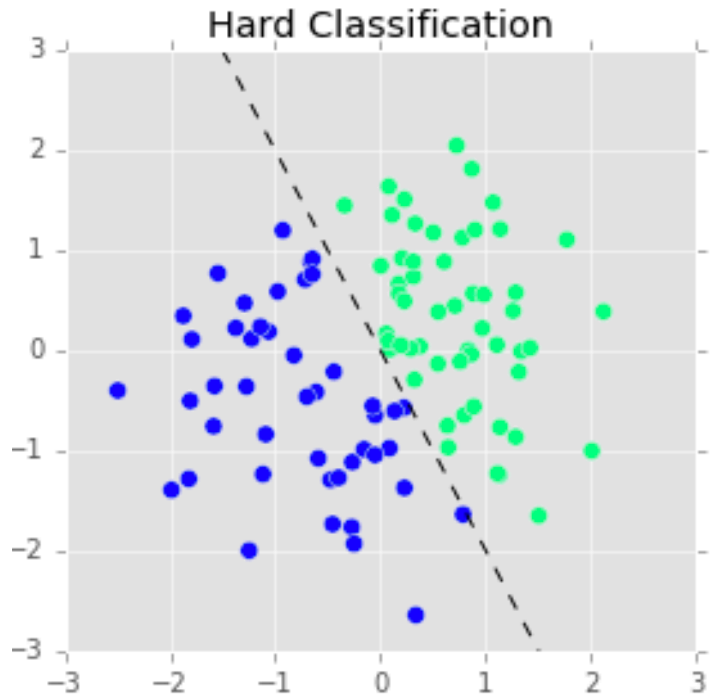
$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$

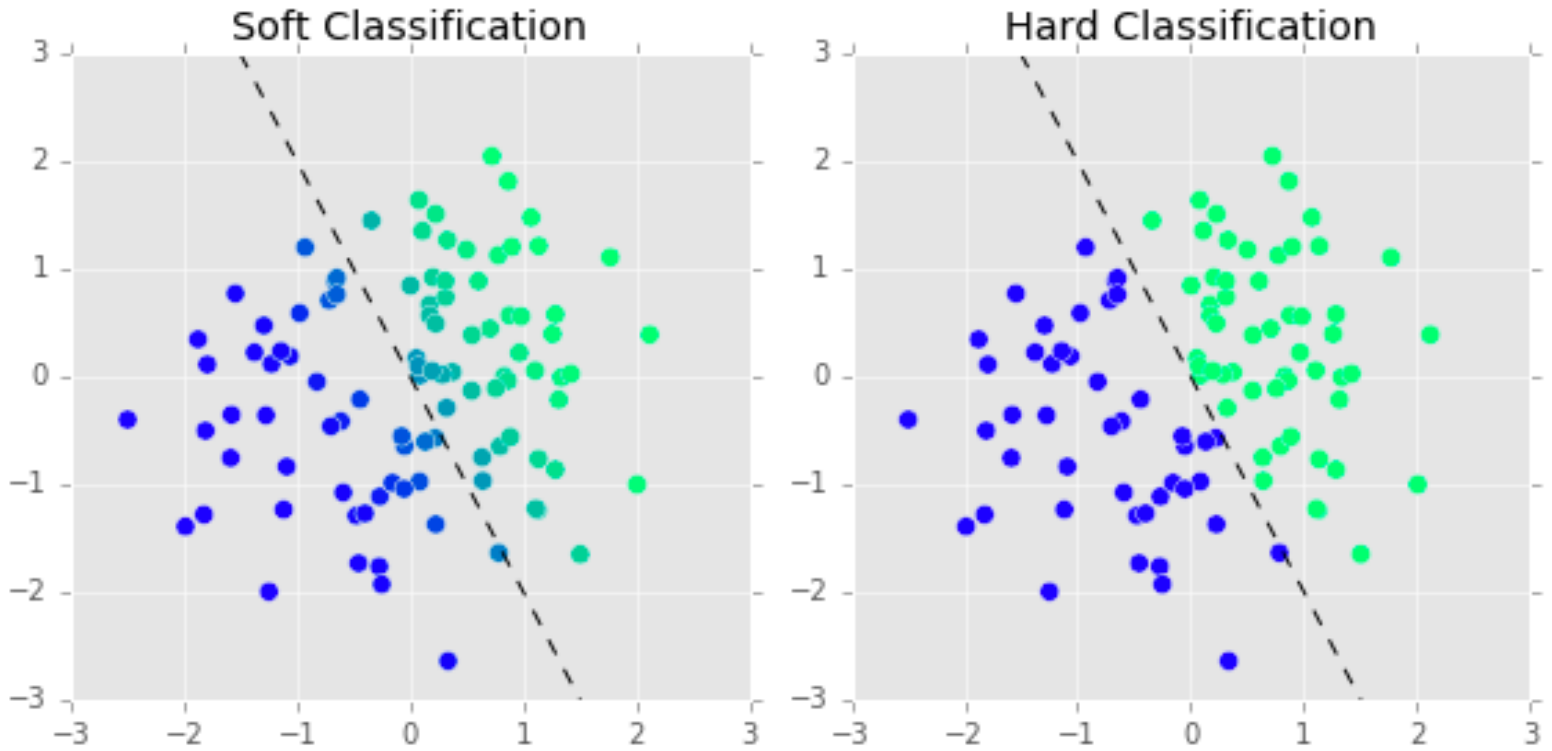


Note:
0 < y < 1

Hard Classification

- Model only attempts to estimate class membership
- Predictions are **not** probabilities
- Predictions are either zeros or ones





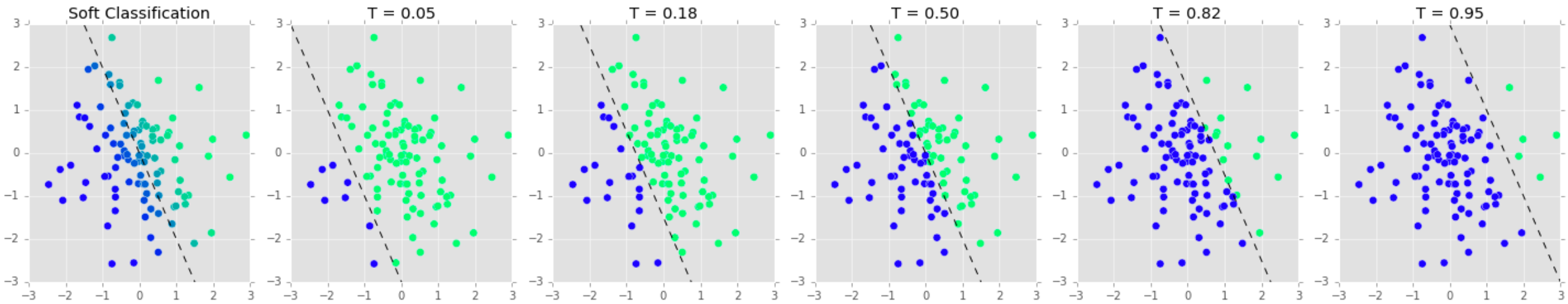
- Common misconception: **All** classification models are hard classifiers
  - **Not true!**
- There **are** some models strictly designed to be hard classifiers
  - For instance SVMs (Support Vector Machines)
- Often, soft classification is a better strategy!

Relationship: Soft and Hard Classifiers

- You can always turn a soft classifier into a hard classifier!
- For binary soft classifiers, this is called **Thresholding**
- Pick a **cutoff** or **threshold T**

$$\text{If } P(y = 1 \mid X) < T \text{ then classify } \hat{y} = 0$$

$$\text{If } P(y = 1 \mid X) \geq T \text{ then classify } \hat{y} = 1$$



Evaluating Classification Models: Two Basic Approaches

- Evaluate probabilities directly

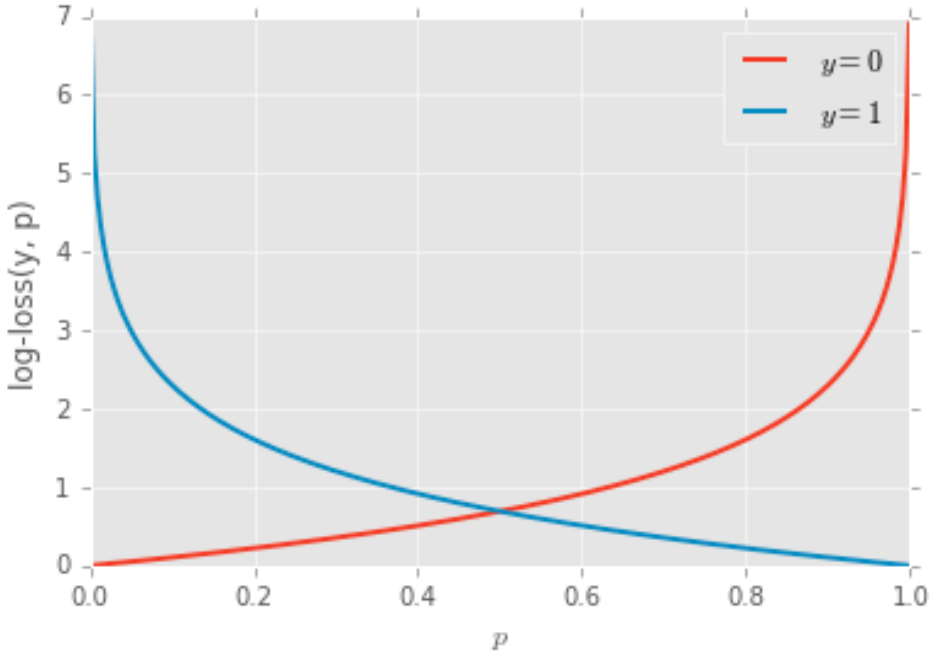
or

- Evaluate ordering of probabilities
  - i.e. something predicted as more probable than something else actually **is** more probable

Evaluating Classification Models: Soft

- Soft classification models are best evaluated and compared on the basis of their predicted probabilities (*not* by converting them to hard classifiers and evaluating the result)
- The gold standard metric for evaluating soft classification models is the log-loss

$$\text{log-loss}(y, p) = -\sum_i y_i \log(p_i) + (1 - y_i) \log(1 - p_i)$$



Evaluating Classification Models: Hard

- When evaluating hard classifications, there are only four possibilities for each data point:
  - **True Negative**: Actual 0, Predicted 0.
  - **False Positive**: Actual 0, Predicted 1.
  - **False Negative**: Actual 1, Predicted 0.
  - **True Positive**: Actual 1, Predicted 1.
- Arranged in a grid → Confusion Matrix
  - Many hard classification evaluation metrics can be derived from this table

	Predicted 1	**Predicted 0**
Actual 1	True Positive	False Negative
Actual 0	False Positive	True Negative

Accuracy

- The simplest and most natural thing to do is compute the proportion of predictions that we got correct
- Lots of problems here

$$\text{Accuracy} = \frac{\#\text{ True Positives} + \#\text{ True Negatives}}{\text{Total \# of Data Points}}$$

or

$$\text{Accuracy} = 1 - \frac{\#\text{ False Positives} + \#\text{ False Negatives}}{\text{Total \# of Data Points}}$$

False Positive and True Positive Rate

- While the accuracy measures how our predictions perform across all of our dataset, the **false positive rate** and **false negative rate** focus on the positive and negative classes individually.

$$\text{FP Rate} = \frac{\#\text{ False Positives}}{\#\text{ False Positives} + \#\text{ True Negatives}}$$

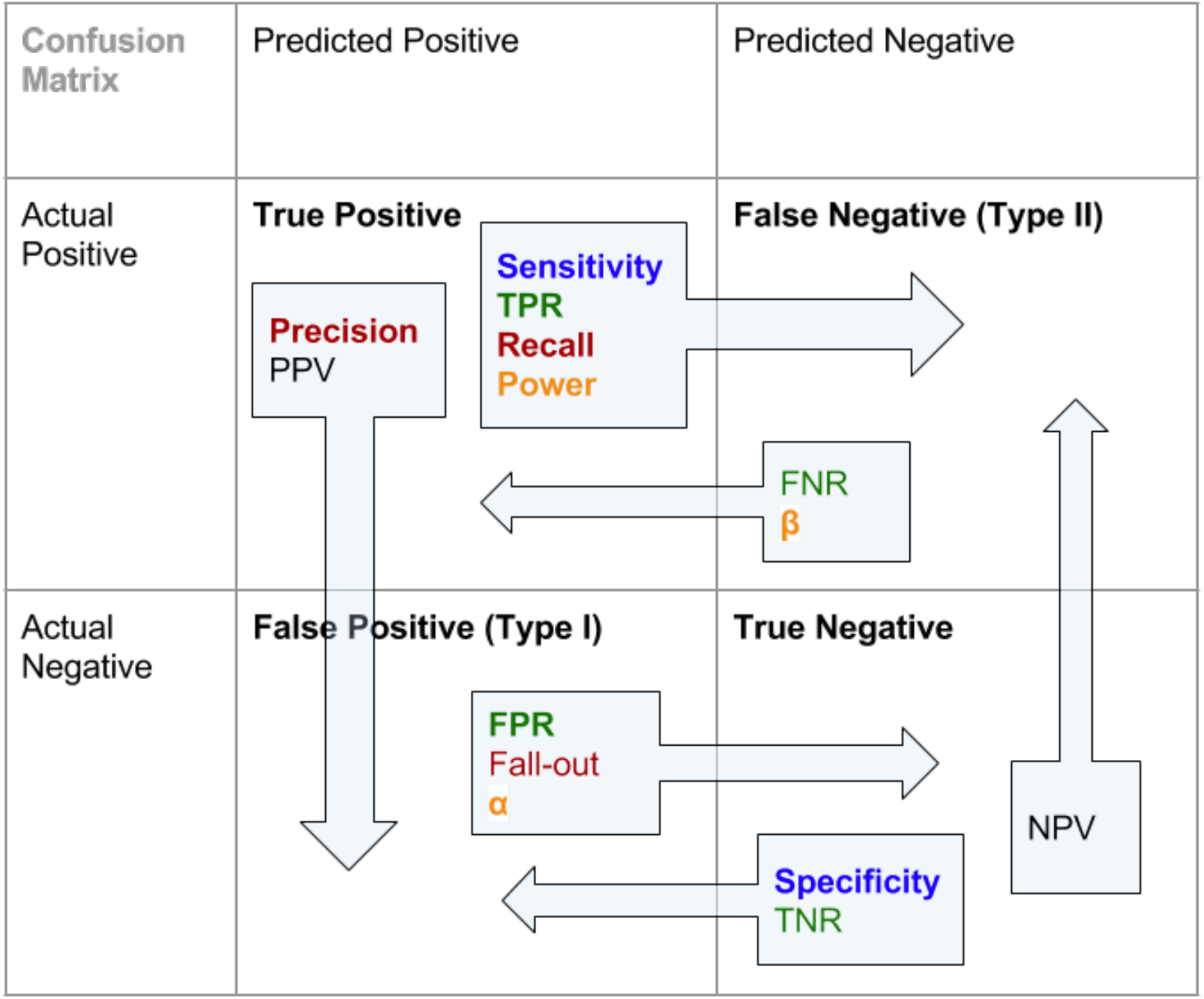
$$\text{TN Rate} = 1 - \frac{\#\text{ False Positives}}{\#\text{ False Positives} + \#\text{ True Negatives}} = \frac{\#\text{ True Negatives}}{\#\text{ False Positives} + \#\text{ True Negatives}}$$

$$\text{TP Rate} = \frac{\#\text{ True Positives}}{\#\text{ False Negatives} + \#\text{ True Positives}}$$

Grand Confusion Matrix

- Each is a ratio
- (TP, TN, FP, FN) /

  sum of 2 others
- Numerator given by label location
- Denominator given by locations spanned by arrow

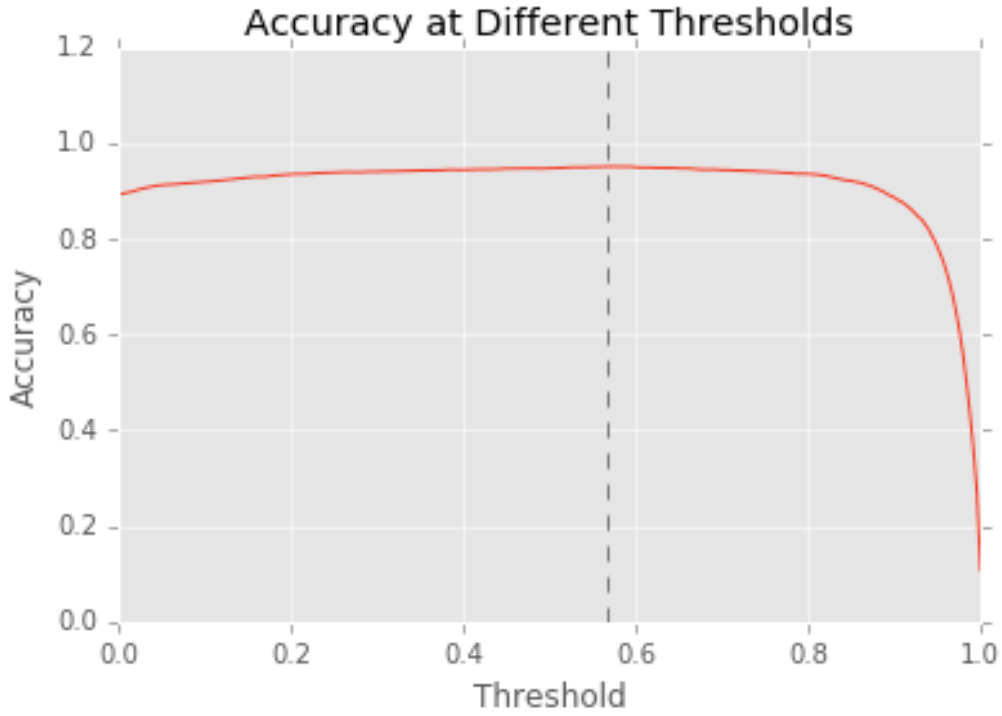


Hard Classification Metrics for Soft Classification Models

- Some metrics that can be used to evaluate both hard and soft classifiers.
  - Since a soft classifier can be made into a hard classifier by thresholding the predicted probabilities at a fixed level…
  - …any metric designed for hard classification models can, after thresholding, be applied to a soft classification model.
- Interesting to see how each of these metrics behaves as we vary the threshold.
- The following slides have plots of some of these measures for a logistic regression on a testing data set.

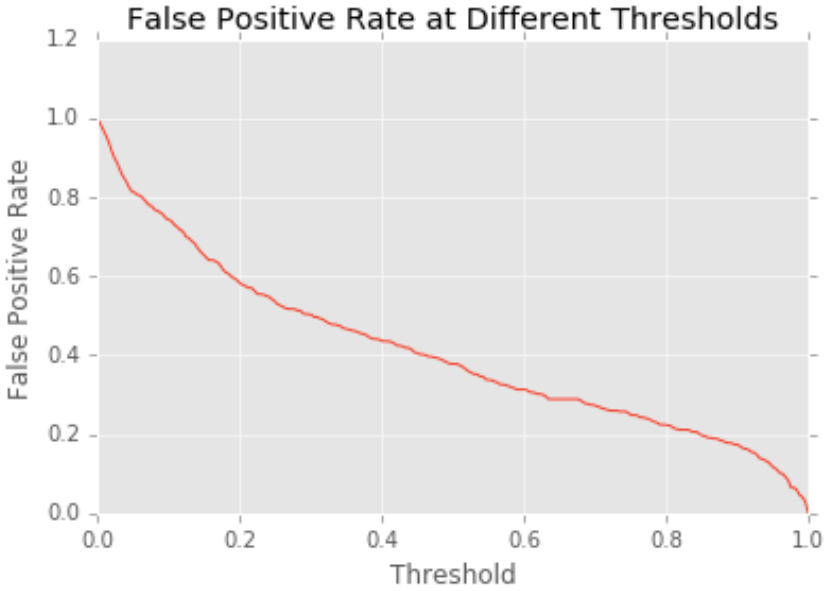
Thresholds - Accuracy

- Tons of things to talk about here, but the main takeaway is that accuracy is rarely a good metric to evaluate hard classifiers
- Example: HIV test with >99% accuracy?
  - Easy!
  - Classify everyone as HIV-
  - For most populations, that's >99% accurate
- Can be a decent metric in rare cases
  - Balanced dataset
  - Scaling minority class to balance data?
    - Not worth it just to use accuracy

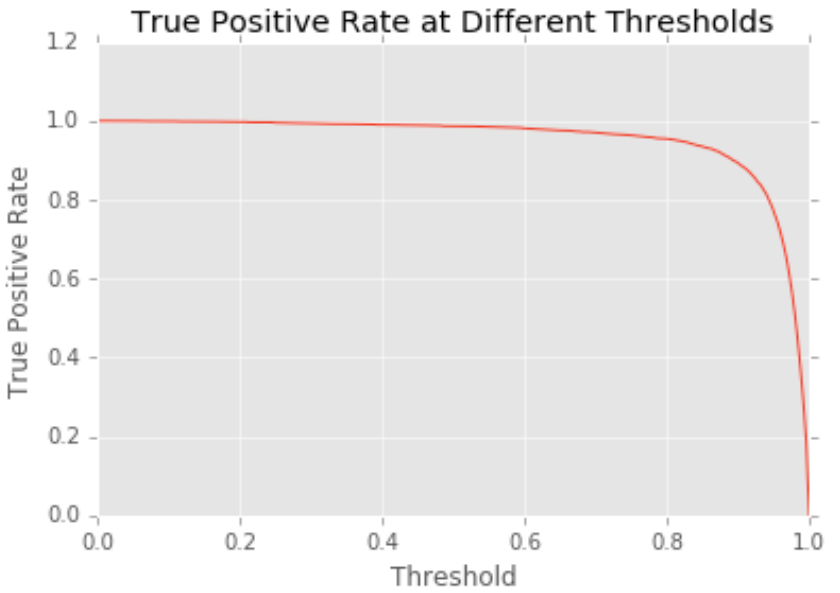


Thresholds - False Positive and True Positive Rates

- Intuitively, as threshold increases
  - FP rate decreases
- At the extremes
  - A threshold of zero classifies everything as positive, so all our negative classes are falsely classified
  - A threshold of one we classify nothing as positive, so all the negative classes are correctly classified.

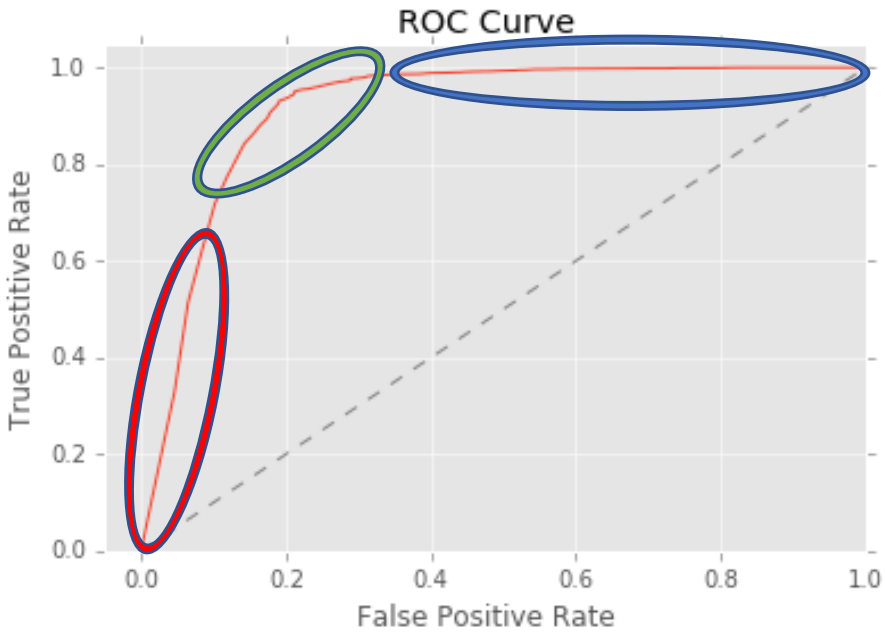


- We interpret behavior of TP rate curve similarly



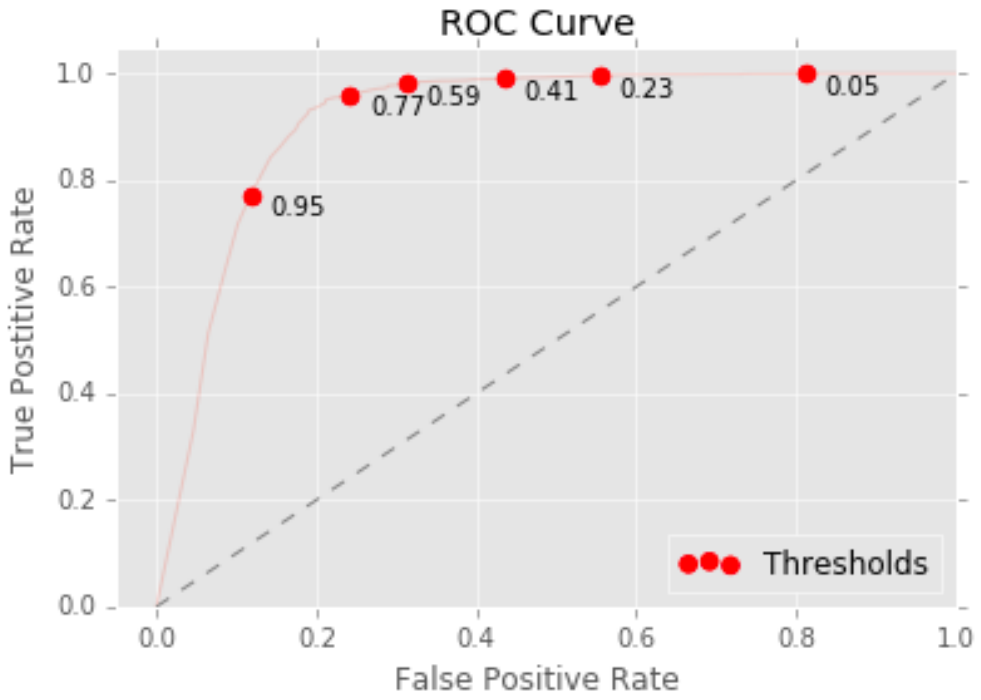
ROC Curves

- A trade-off when setting the threshold of a soft classifier if a hard classification is needed
  - If the threshold is set too low, then we will be very liberal in what observations we classify as positive, which drives up the false positive rate of the model.
  - If the threshold is set too high, then we will be very conservative in what observations we classify as positive, which will drive down the true positive rate.
  - We suspect that in the middle there is some sweet spot, where we have balanced a trade-off between the false positive and true positive rates.
- An **ROC** (receiver operating characteristic) **curve** expresses this relationship in one picture.



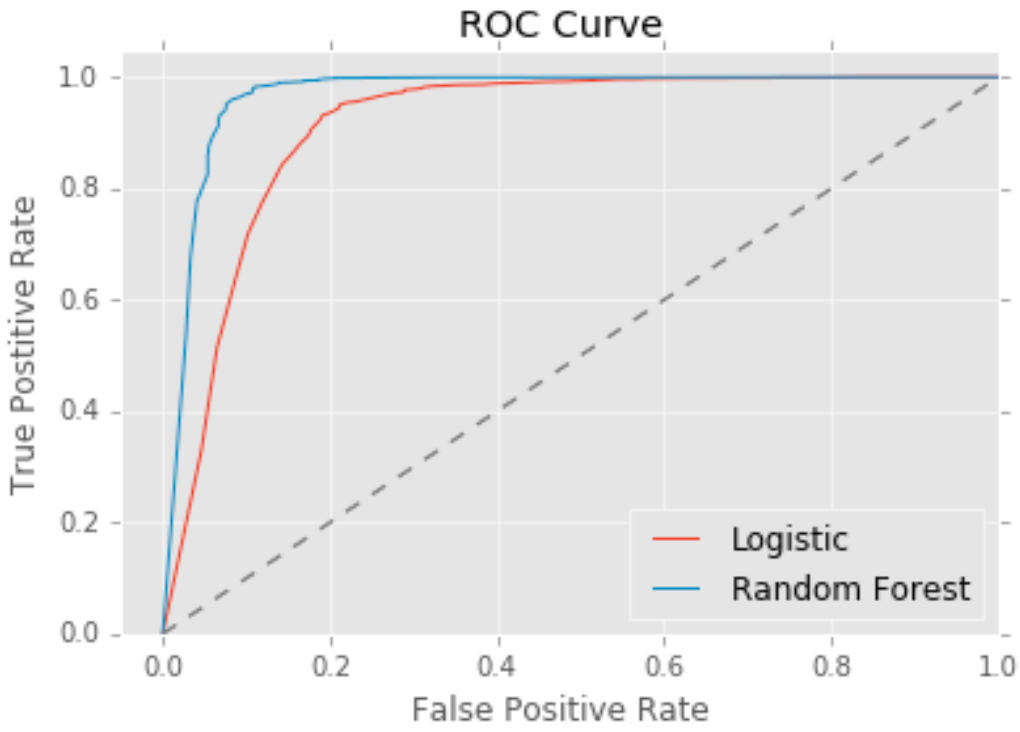
ROC Curves

- An ROC curve visualizes the trade-offs between false positives and true positives as we vary the threshold on a soft classification model.
- Each choice of the threshold results in **one** point on the ROC curve
- Dashed line between (0, 0) and (1, 1)
  - Traditional to include
  - Represents random class assignment
  - Curve above line
    - Predictive power better than random guess



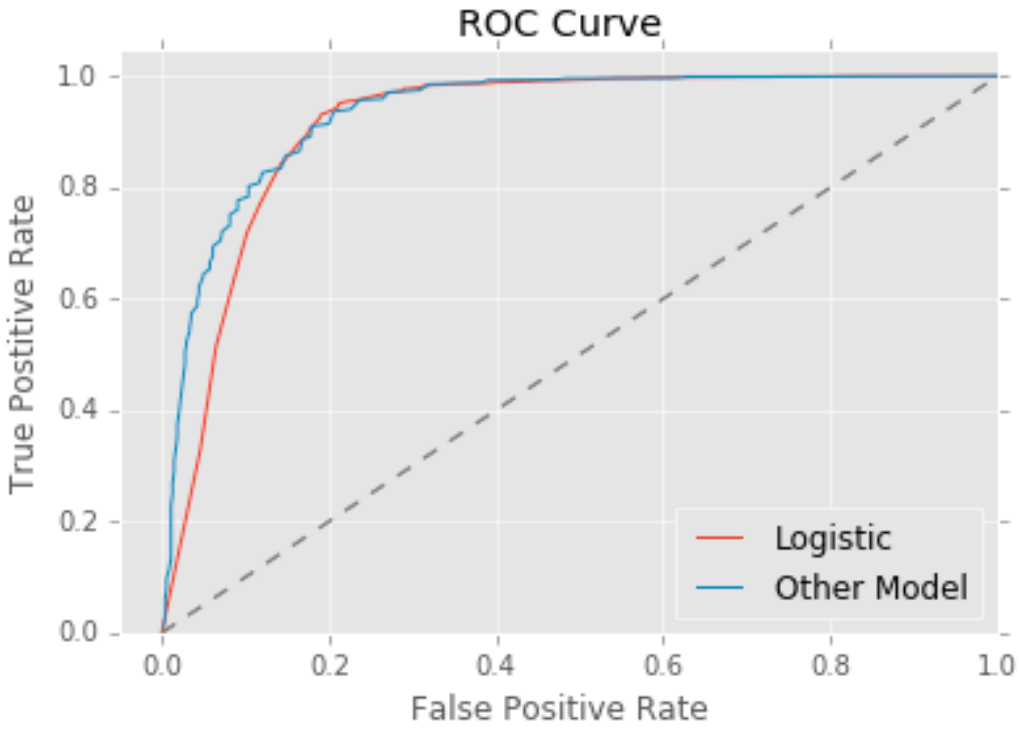
ROC Curves

- Common to inspect ROC curves of competing classifiers
  - Get a feel for how models make tradeoffs for FP and FN rates
- Here, random forest does a better job of distinguishing the positive from negative classes.



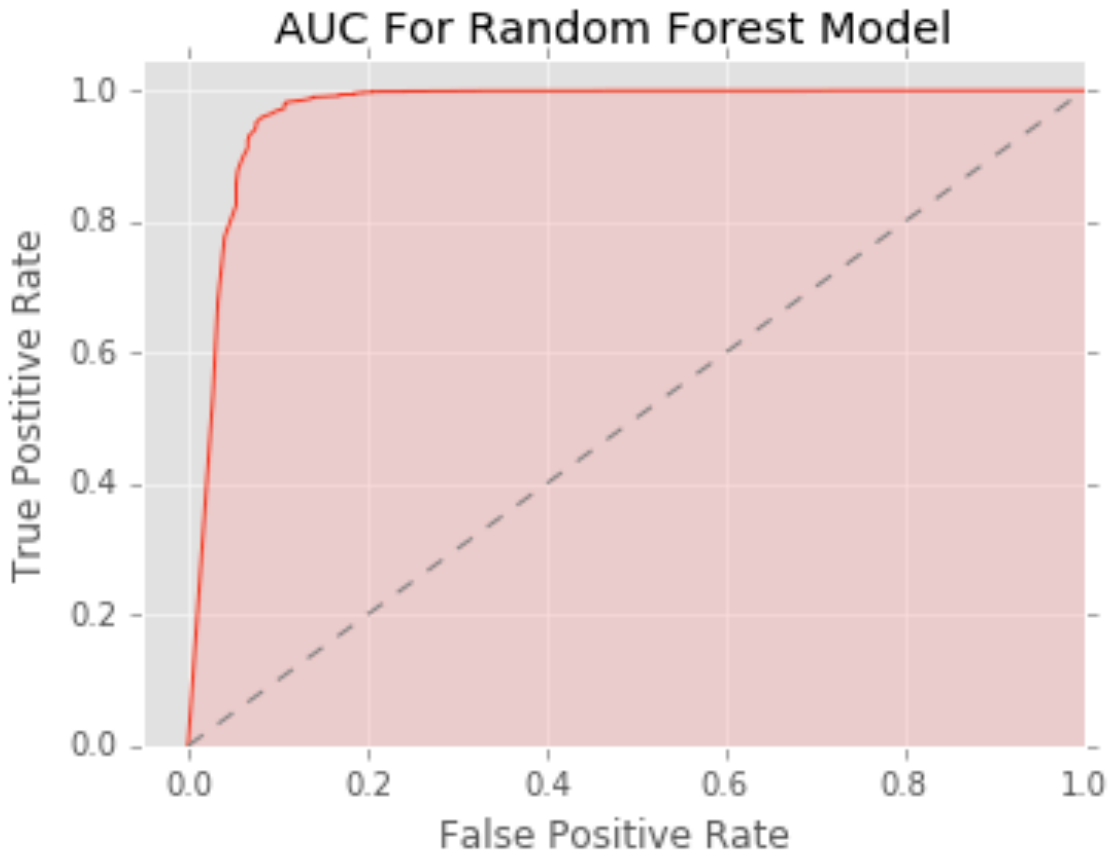
ROC Curves

- In some cases, the ROC curves for the models being compared will cross
- ROC curves give us no statistical reason to prefer one model over another
  - It depends on the costs of false positive and false negative errors in our decision problem.
  - Profit curves useful in business questions, but each problem space will need its own approach



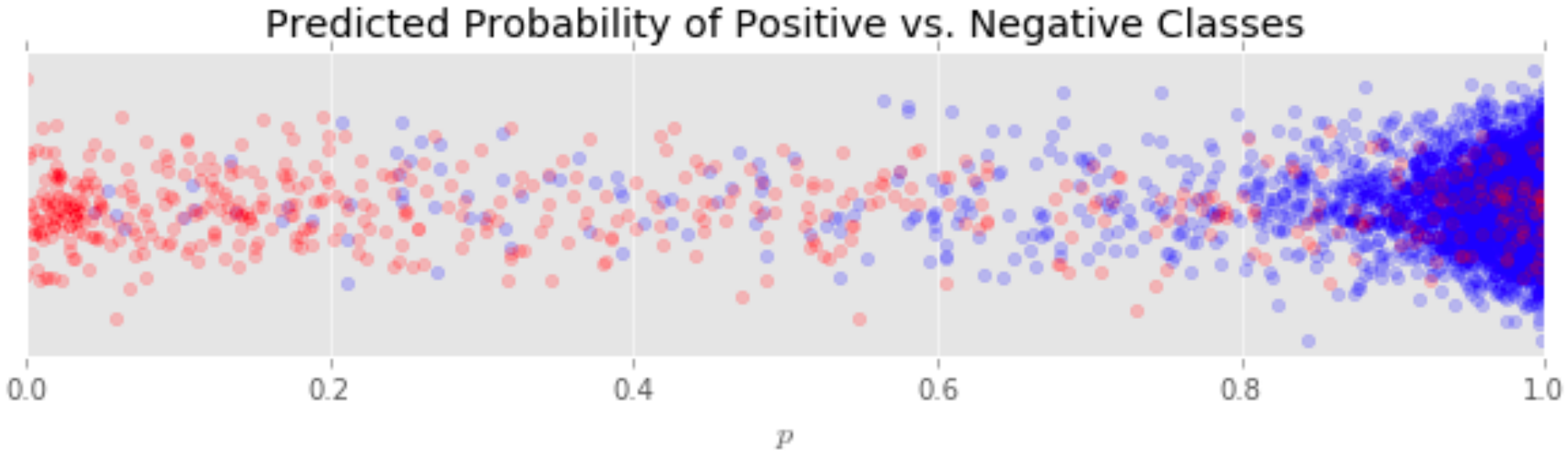
The AUC

- The **area** under the ROC curve is a numeric measure of a model's overall ability to distinguish between positive and negative classes.
- AUC is a nice measure of model fit, as it averages over **all** thresholds.



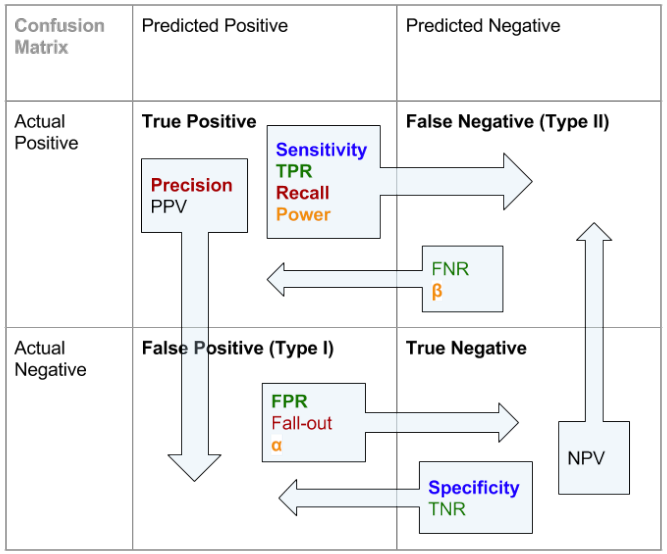
Interpretation of AUC

- The AUC has a very nice probabilistic interpretation:
  - The probability of our model predictions *ranking* a randomly chosen positive class higher than a random negative class
- The AUC statistic measures **the chance that a random pair of (red, blue) points is ranked in the correct order**

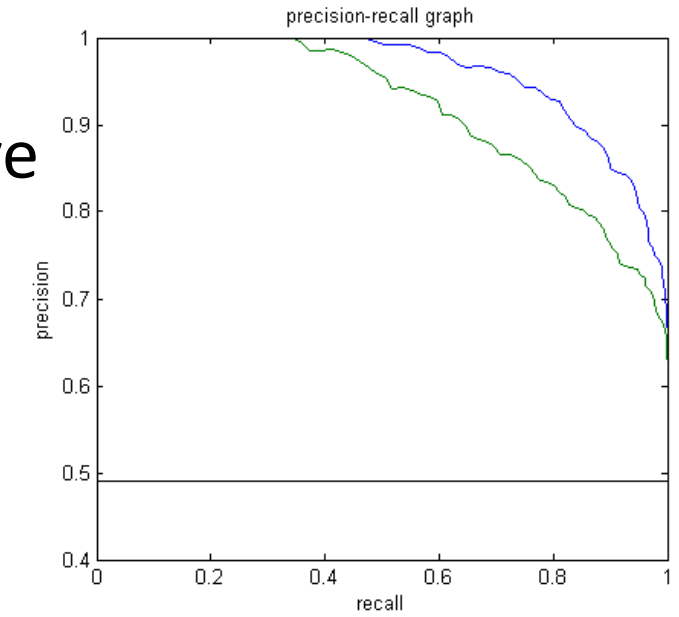


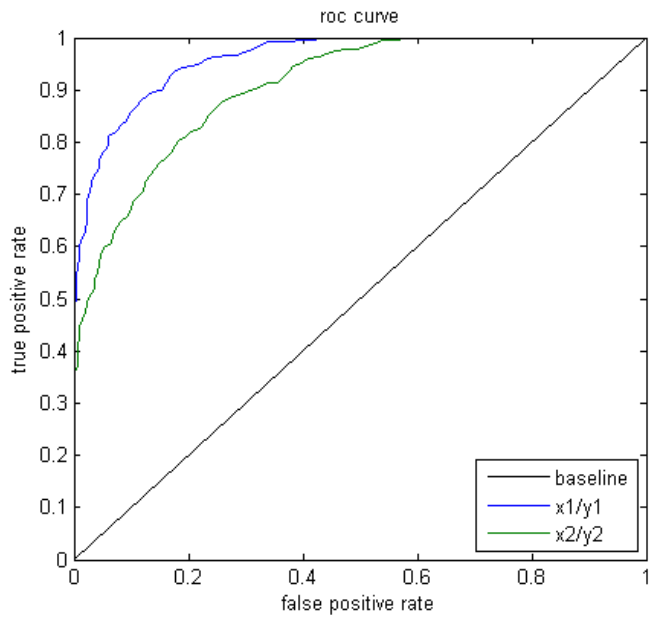
Precision-Recall Curve

- Another chart used to describe the trade-off at different thresholds
- Also parameterized by the threshold (like the ROC)
- But plots the precision (the ratio of true positives over predicted positives) against the recall (a.k.a. sensitivity a.k.a. TPR).



- Precision-recall curve
  - x is recall
  - y is precision





- ROC curve
  - x is FPR
  - y is TPR

Conclusion:
Discussion - Do You Really Need Hard Classification?

- The most persistent mistake in this area is the *overuse of hard classification models*. Let's be explicit: soft classification models are more flexible in practice, and they should generally be preferred.
- But many business or scientific problems **do** call for hard classification, especially when some decision must be made
  - e.g. we are going to do this or not, we are going to invest in this or that, etc.
- It is important to distinguish in our problem solving process between:
  - Estimating the probability of uncertain events.
  - Using those probabilities to make hard decisions or develop decision rules.
- The best general principle
  - Convert your probabilities into decisions as late into the process as possible.