Evaluation/Performance Metrics

Rayid Ghani and Kit Rodolfa

Carnegie Mellon University

ML
MACHINE LEARNING
DEPARTMENT

Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Reminders

Coming up next week:

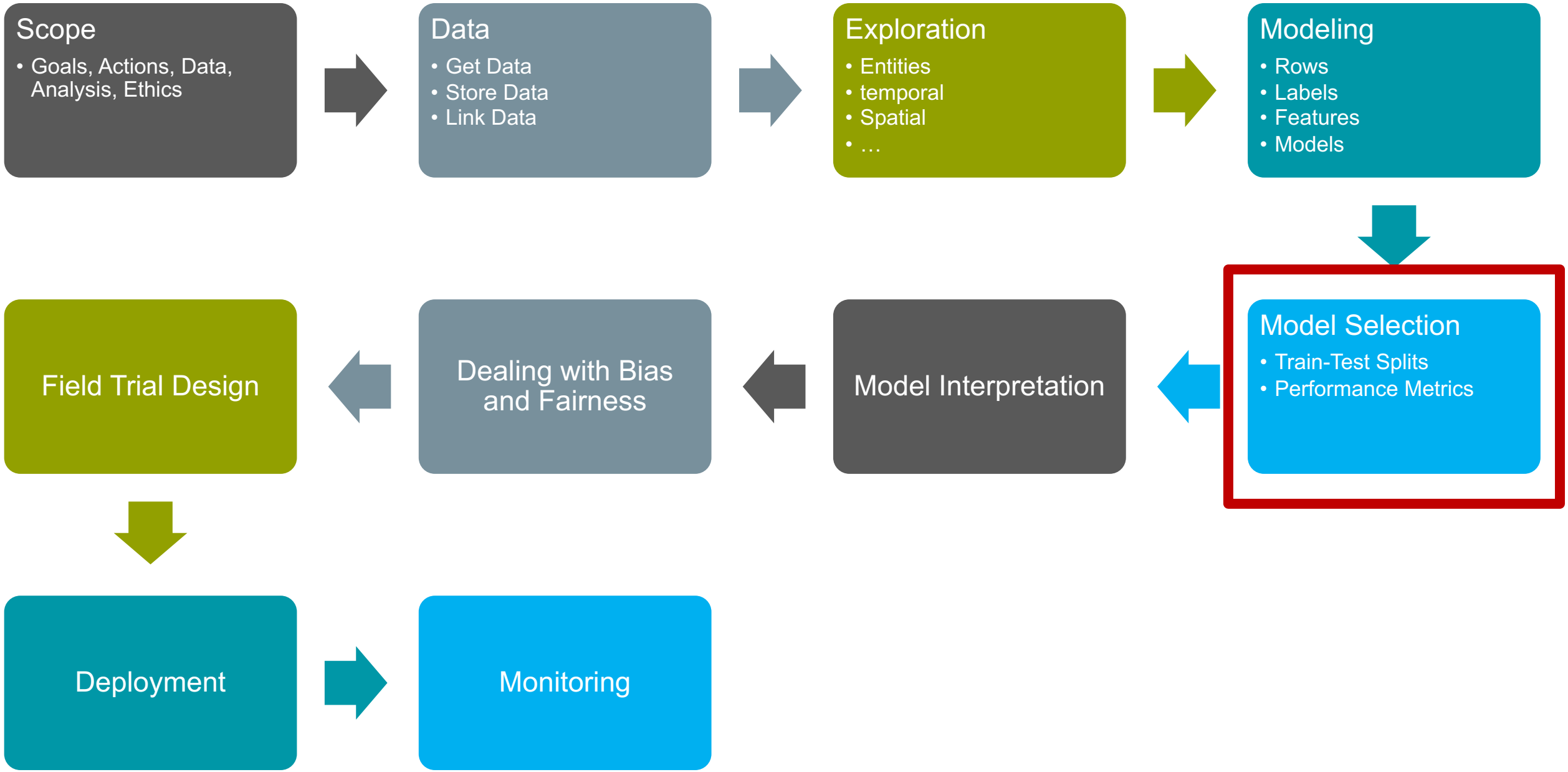
- Monday: Project Update 2
- Tuesday: Weekly Feedback Form
- Wednesday: Deep Dive Session on Modeling and Validation Plans

Project Update 1 Review

- Formulation – first of every month?
- Base rate – definition and keeping parallel with label
- Commonsense Baselines – what makes sense here?

Discussion Question

What validation strategy is your group using for the Donors Choose project?


Scope
• Goals, Actions, Data, Analysis, Ethics
Data
• Get Data
• Store Data
• Link Data
Exploration
• Entities
• temporal
• Spatial
• …
Modeling
• Rows
• Labels
• Features
• Models
Model Selection
• Train-Test Splits
• Performance Metrics
Model Interpretation
Dealing with Bias and Fairness
Field Trial Design
Deployment
Monitoring

How to solve a prediction problem

- Define and Create Rows (unit of prediction)
- Define and Create Label (outcome/target variable – what event and when?)
- Define and Create Features (features/predictors)
- Create Training and Validation/Test Sets
- Train model(s) on Training Set(s)
- Validate model(s) on Validation/Test Set(s)
- Select “best” model

What is the goal of model selection?

What is the goal of model selection?

- You've run a large number of different types of models varying …
- You need to understand what types of models are effective under what circumstances, **and**
- You need to decide which one(s) to use in the **future**

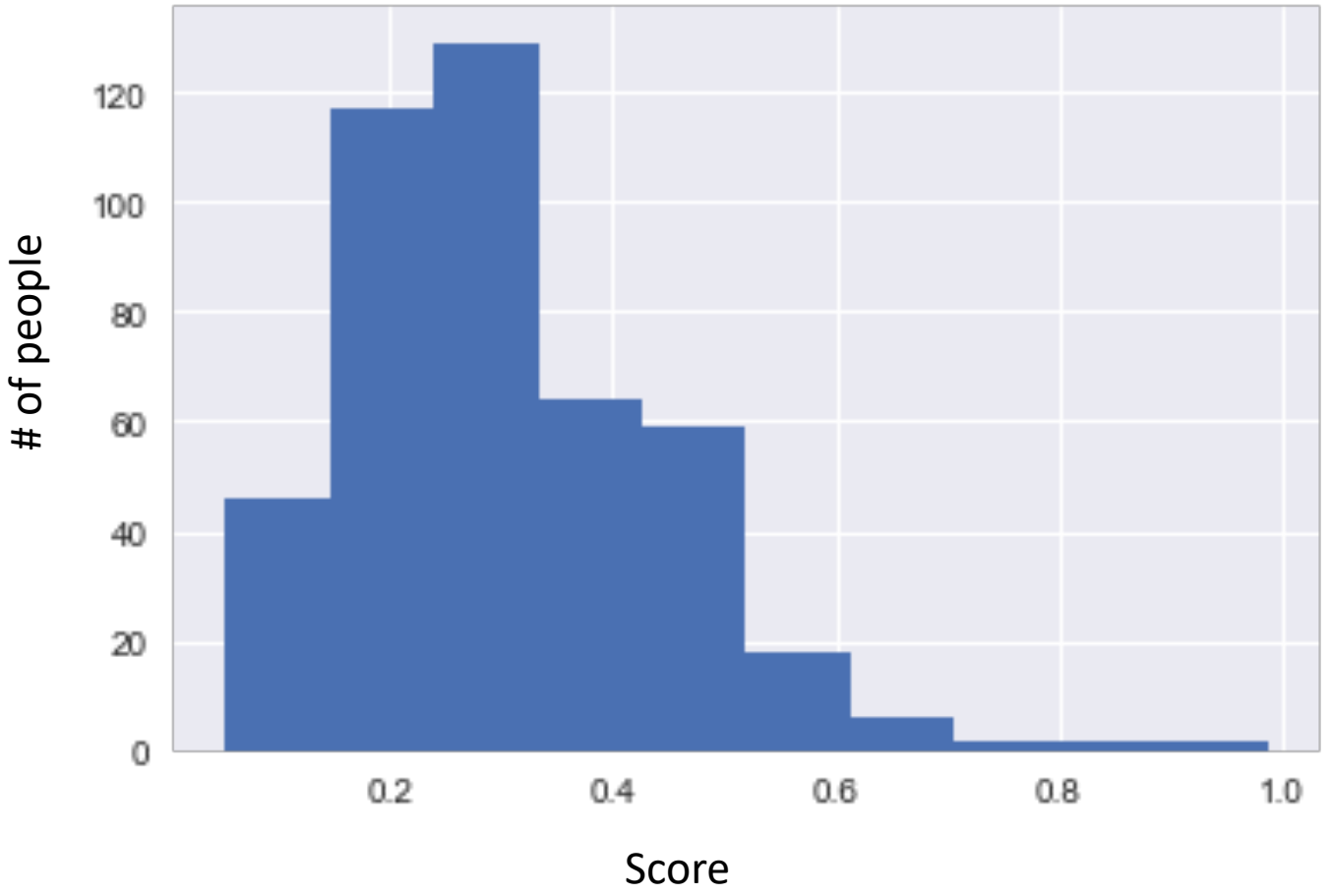
What do we need our selected model to do?

What do we need our selected model to do?

- Perform well
  - What metric?
  - Compared to what?
- Generalize
  - To what?

Performance Metrics

Score Distribution on the Test Set



Evaluation - Metrics

- Predictions are often **scores** between 0 and 1
- We need to first turn them into 0 or 1 by selecting a threshold

Predicted Class

Actual Class

	Yes	No
Yes	**True Positives (TP)**	**False Negatives (FN)**
No	**False Positives (FP)**	**True Negatives (TN)**

Evaluation – Metrics (at a threshold k)

- Accuracy = **(TP + TN) / (TP + TN + FP + FN)**
- Precision (aka PPV: Positive Predictive Value) = **TP / (TP + FP)**
- Recall (aka Sensitivity, TPR) = **TP / (TP + FN)**
- Specificity = **TNR**

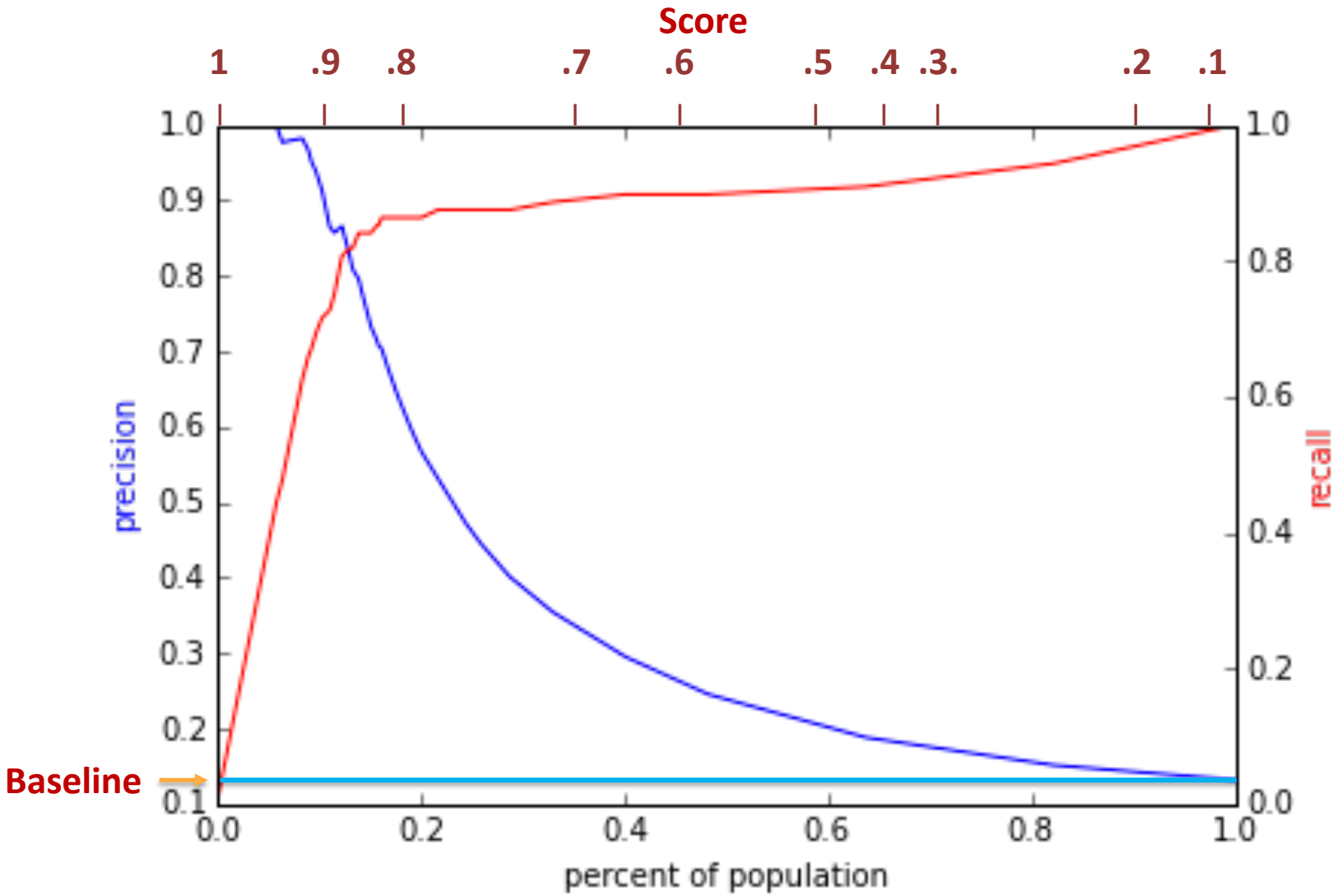
		Predicted: Yes	Predicted: No
Actual	Yes	**True Positives (TP)**	**False Negatives (FN)**
Actual	No	**False Positives (FP)**	**True Negatives (TN)**

Confusion Matrix-based Metrics Cheatsheet

		True condition				
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$	F_1 score = $\frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$		

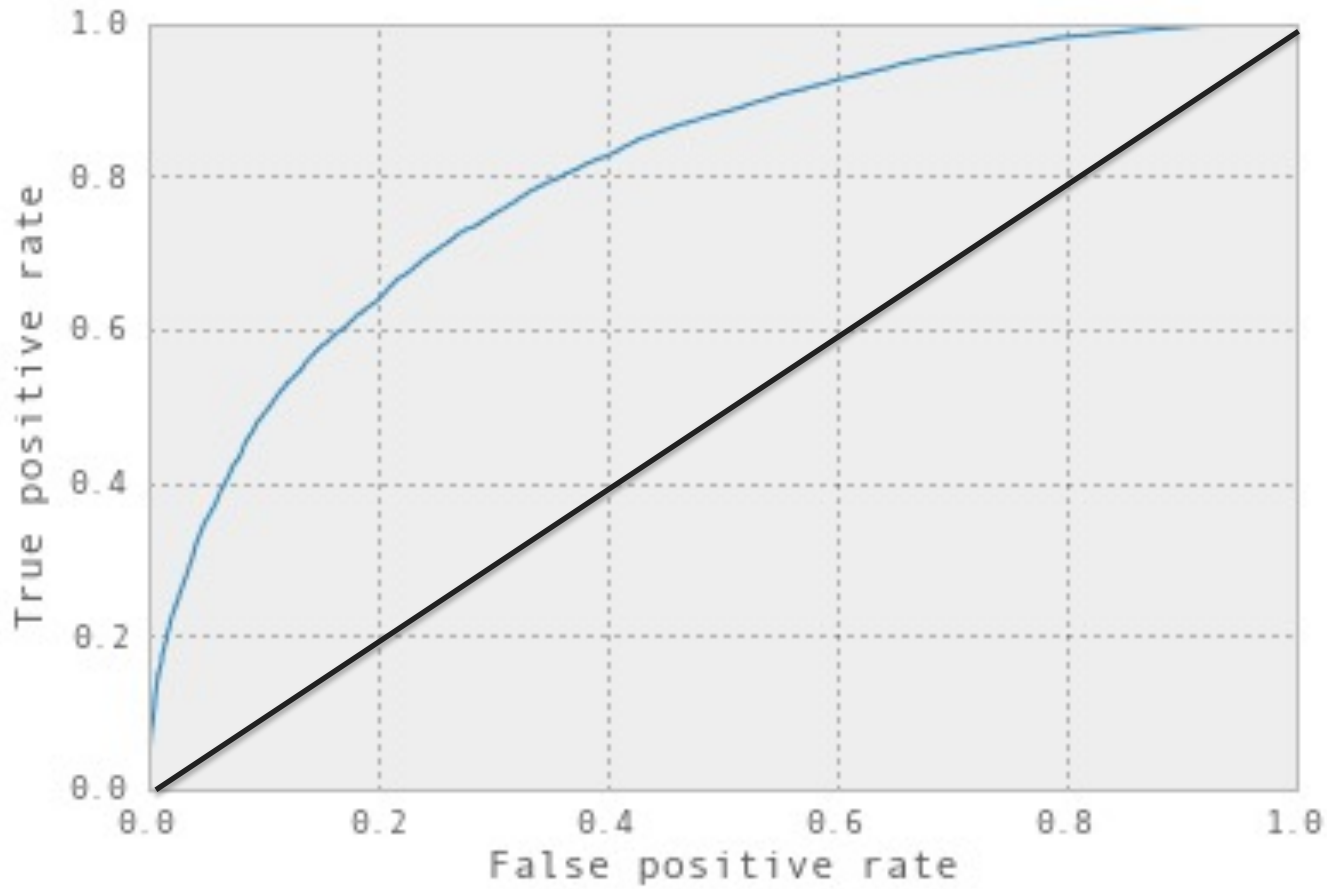
Source: https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Varying the Threshold



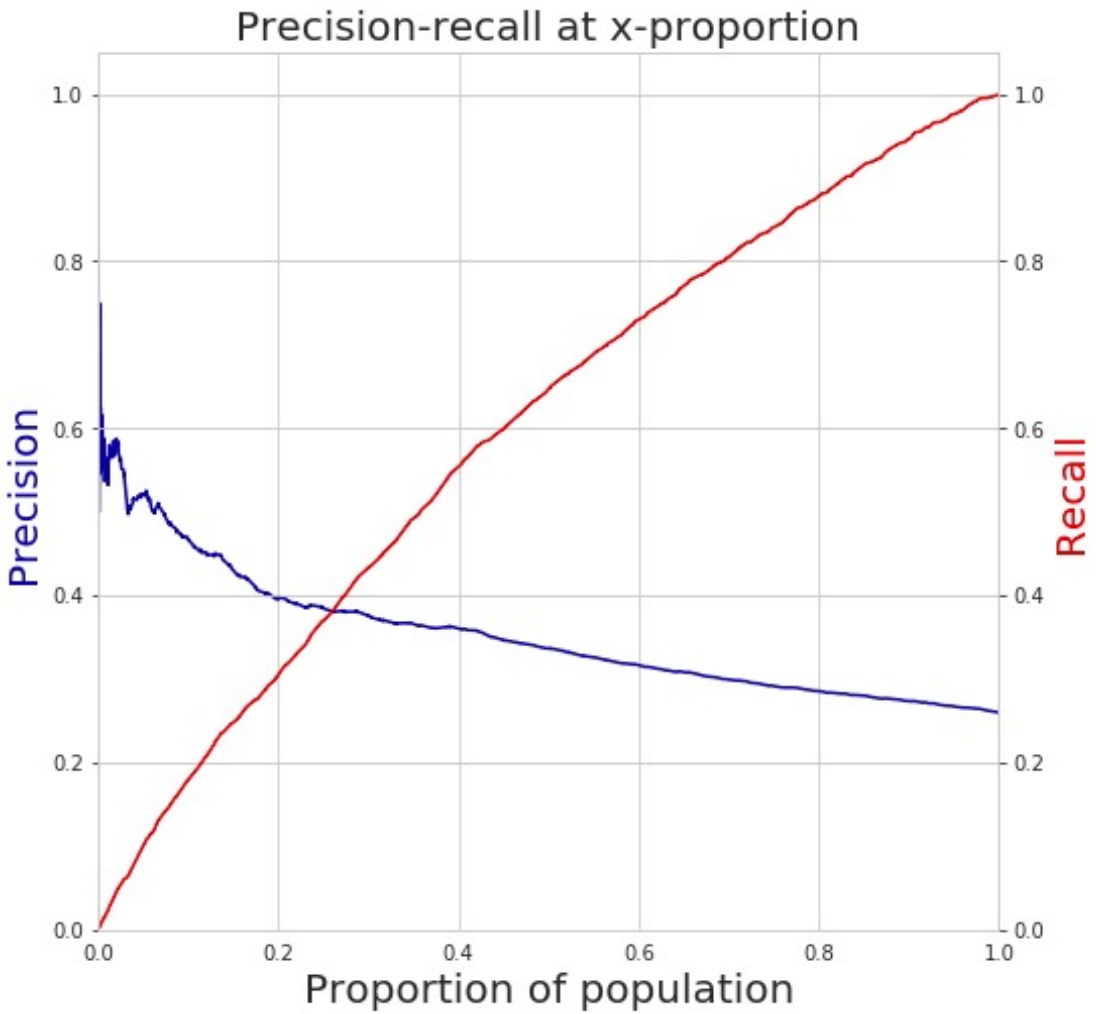
ROC Curve

Receiver Operator Characteristic Curve



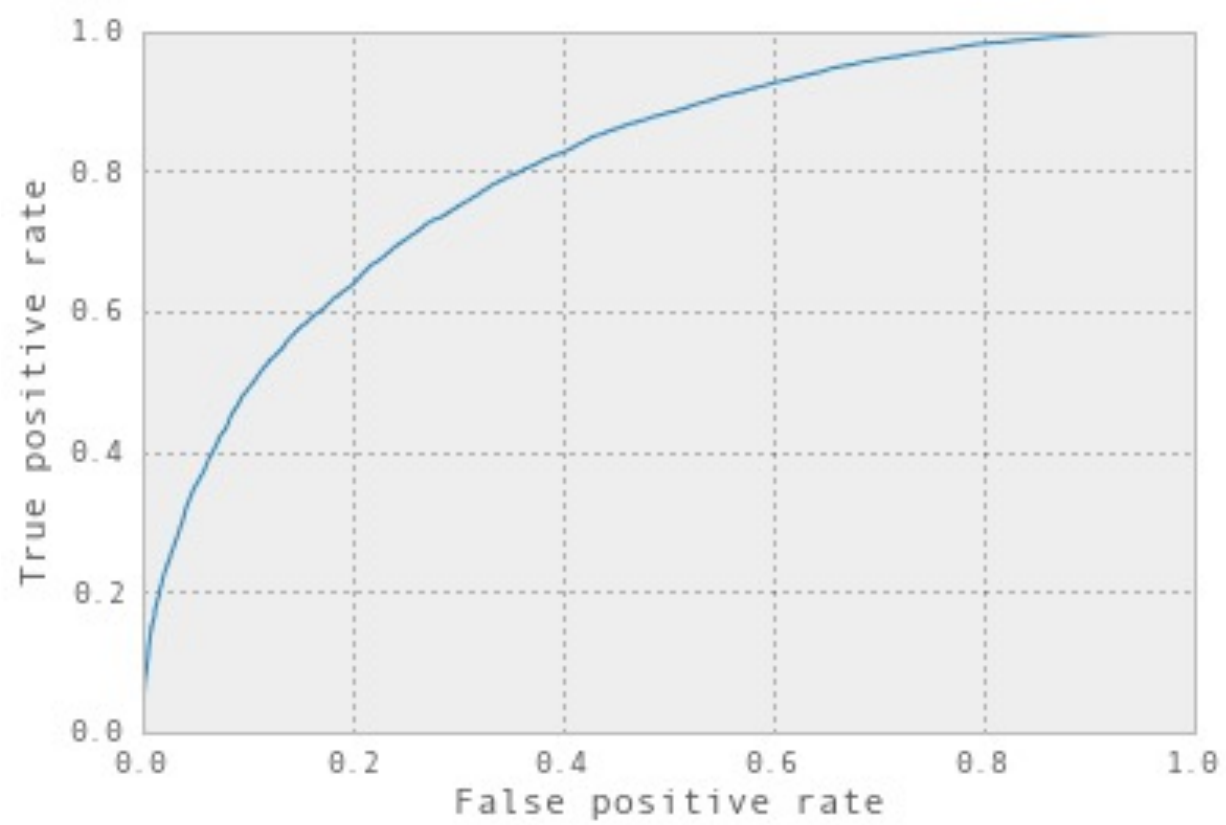
AUC (Area Under Curve)

- Overall measure of performance
  - 1 if all 1s are ranked above all 0s
  - 0 if all 0s are above all 1s


Precision-recall at x-proportion
Precision
Recall
Proportion of population
0.0
0.2
0.4
0.6
0.8
1.0

AUC – Area Under Curve

- If you care about the entire space



Evaluation - Baselines

- Random according to the base rate/class prior
- What they do today
- What they could easily do today (without much if any ML)

Discussion Question

What would be some potential strategies for integrating considerations around fairness in the process of model evaluation and selection?

Reminders

Coming up next week:

- Monday: Project Update 2
- Tuesday: Weekly Feedback Form
- Wednesday: Deep Dive Session on Modeling and Validation Plans