

# CLASSIFICATION EVALUATION

BEN ELYSE KENNETH

---

# CLASSIFICATION

---

UTILIZED BY DATA SCIENTISTS TO MAKE PREDICTIONS ABOUT DATA WHICH CAN BE SORTED INTO TWO OR MORE CATEGORIES

TRAINED ON DATA WITH KNOWN CATEGORIES AND USED TO PREDICT TO WHICH CATEGORY NEW, UNKNOWN DATA POINTS BELONG



# ACCURACY

First to evaluate classification models, a lot of people might start with accuracy; however, it is important to recognize that using only this metric cannot effectively evaluate classification models, as accuracy is not adequate or can be misleading.

tions

ons

- ▶ Can classification models be effectively evaluated and compared using only their accuracy?
- ▶ Accuracy is not adequate and can be misleading

For example when there are more than 2 classes in a data, the accuracy value received may be high, but it is unknown if some classes are being unattended to instead of all classes having been equally predicted.

# CONFUSION MATRIX

# What it is

# What it does

# Usefulness

- ▶ Summarizes the performance of a classification algorithm
  - ▶ Gives the amount of correct and incorrect predictions
  - ▶ Measures the effectiveness of models: Sensitivity, Specificity, Precision, F1Score



# CONFUSION MATRIX

# TABLE

		Predicted Values	
		Negative	Positive
Actual Values	Negative	TN	FP
	Positive	FN	TP

Shows different combinations of predicted and actual values and gives the counts

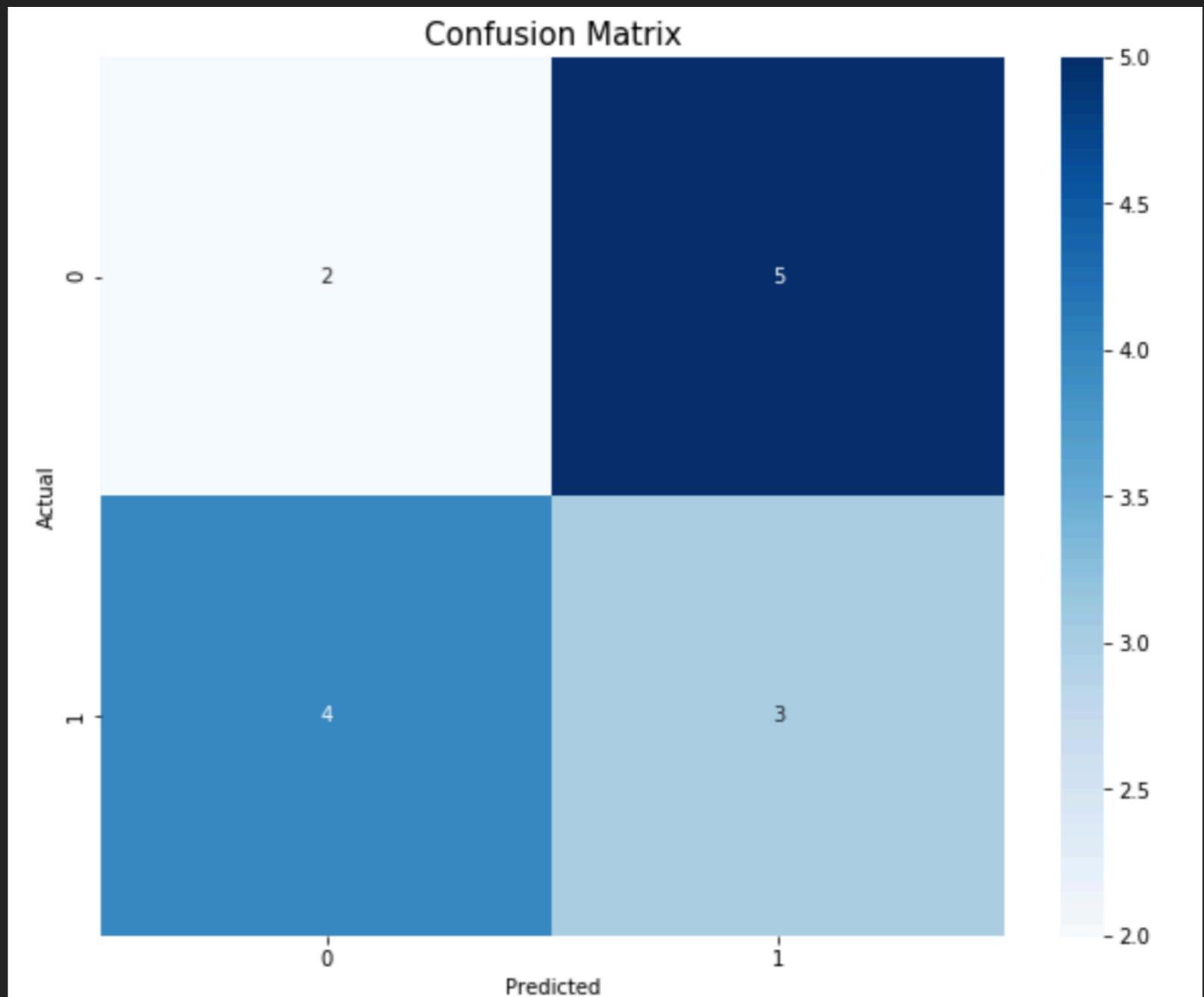
- ▶ Predicted Values: Positive or Negative
- ▶ Actual Values: True or False
- ▶ TP: Our prediction is yes, and it's true
- ▶ TN: Our prediction is no, and it's true
- ▶ FP: Our prediction is yes, and it's false  
(Type I Error)
- ▶ FN: Our prediction is no, and it's false  
(Type II Error)



Disease Example:

0 = Disease Free

1 = Has Disease



Sensitivity is when the actual value is yes, how often does it predict yes, so the equation is True Positives/ (actual values of yes which are the sum of true positives and false negatives)

# SENSITIVITY & SPECIFICITY

$$\frac{TP}{TP+FN}$$

$$\frac{TN}{TN+FP}$$

		Predicted Values	
		Negative	Positive
Actual Values	Negative	TN	FP
	Positive	FN	TP

Sensitivity: Proportion of positive predictions when it's actual value is true

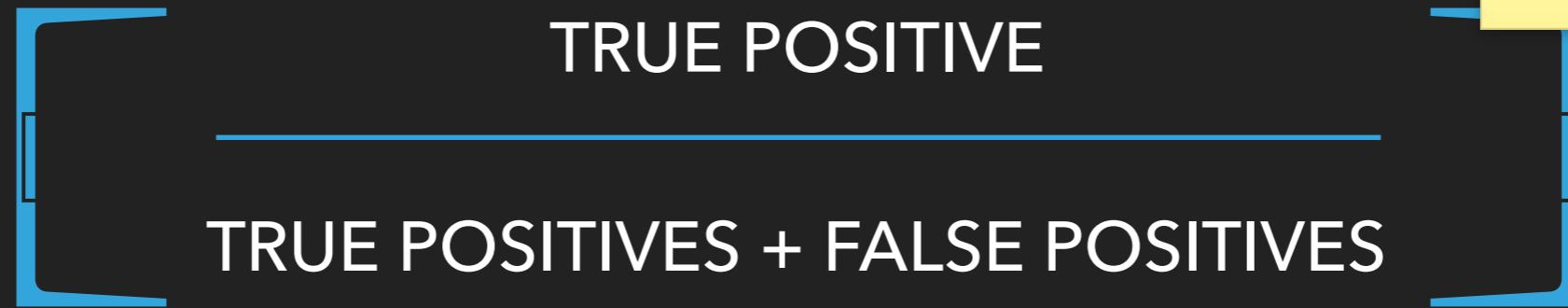
- ▶ A test with high sensitivity is useful for identifying sick patients with positive test results
- ▶ A 60% sensitivity test will recognize 60% of patients with disease, but will miss 40% of those with disease

Specificity: Proportion of negative predictions when it's actual value is false

- ▶ A test with high specificity is useful for correctly identifying healthy patients who don't have a disease
- ▶ A 60% specificity test will recognize 60% of patients without disease, but will miss 40% of those without disease

# PRECISION

When it predicts yes, how often is it correct?  
True positives/ predicted yes



- ▶ The rate by which the model is correct when it predicts positive
- ▶ Useful metric for determining when there are high costs of false positives.
  
- ▶ A model with low precision will tell lots of healthy patients that they do have a disease which will result in misdiagnoses
  
- ▶ So for models with low precision, thus high false positives, the doctors should not rely on the results of those tests as it can give them many false alarms.

# F1SCORE

$$2 \times (\text{PRECISION} \times \text{RECALL})$$
$$\text{PRECISION} + \text{RECALL}$$

For a balance between precision and sensitivity, combines the two to give a overall measure of accuracy

Good F1 Score: Low false positive and low false negatives, meaning no false alarms

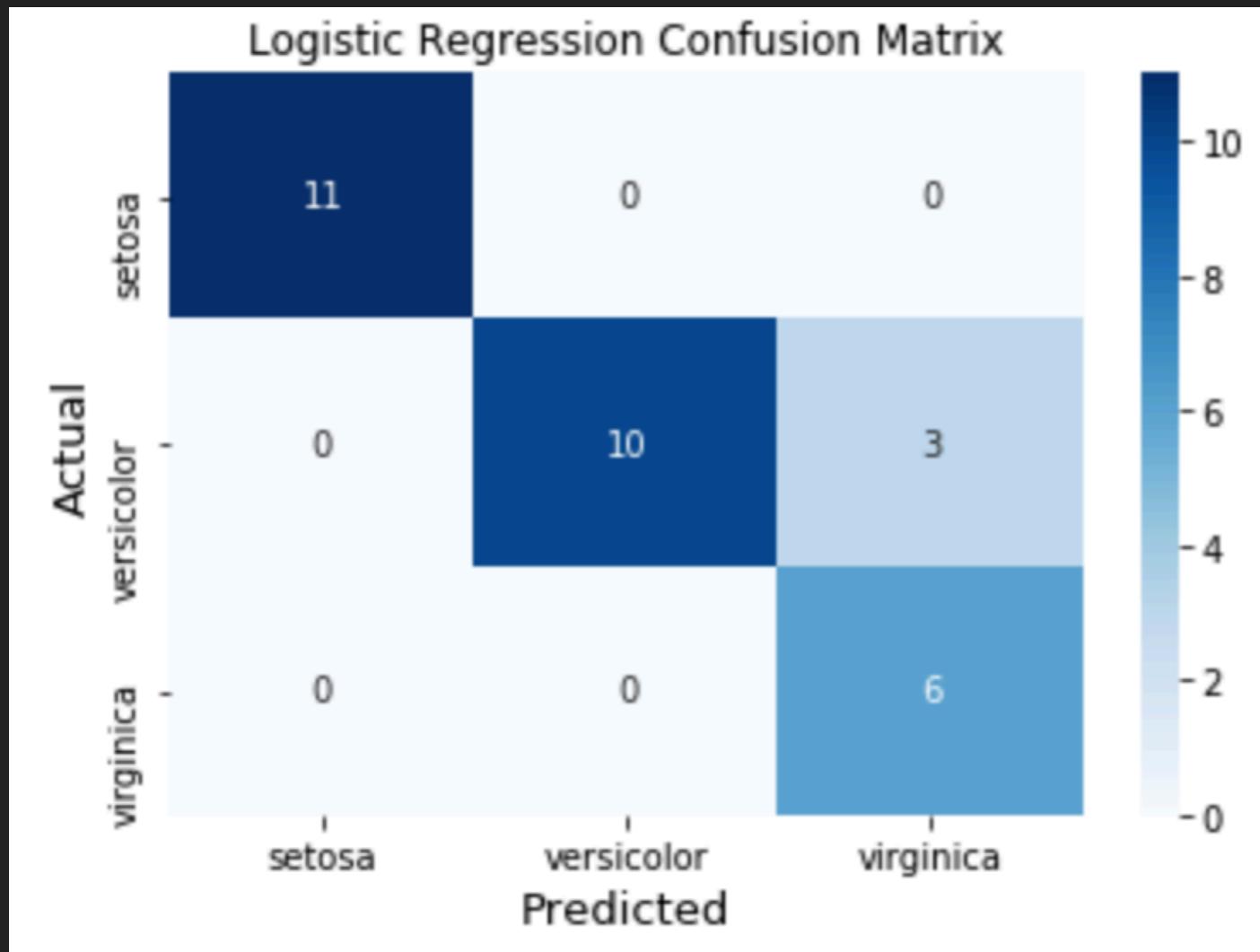
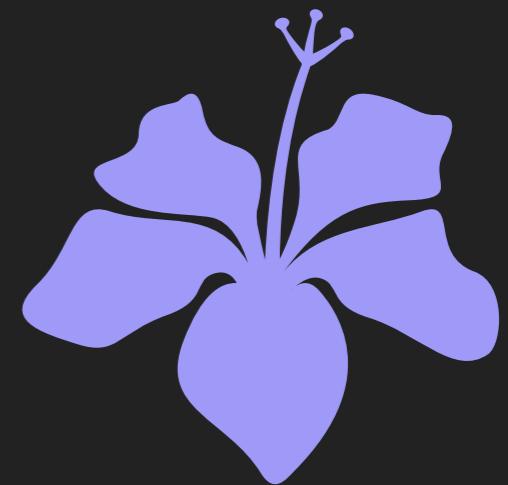
F1 score is perfect when it's 1, and model has failed when it's 0

Can be a better measurement when we need a balance between the two for a harmonic measure

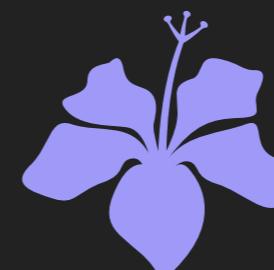


# CONFUSION MATRIX EXAMPLE

## DETECTING PLANT IRIS



Setosa and Virginica is predicted perfectly, while there are problems with Versicolor



## Setosa



## Versicolor



## Virginica



# STATS:

0      1      2

TPR: (Sensitivity, hit rate, recall)

1    0.769231                1

TNR=SPC: (Specificity)

1                        1    0.875

PPV: Pos Pred Value (Precision)

1                        1    0.666667

F1 score

1    0.869565                0.8

## Virginica

The sensitivity test recognizes 100% of flowers as Virginica. 87.5% of the time, the model predicts not Virginica when the flower is not Virginica, but misses 12.50% of those that are not Virginica. When it predicts positive: (Yes, it is Virginica), the model is correct at a rate of 66.67%. The F1 score is 0.8, which is on the high end of the range 0 to 1, meaning that the model is good.

# ROC CURVE AKA RECEIVER OPERATING CHARACTERISTIC CURVE

- ▶ Shows how well a classifier performs at all possible values of its discrimination threshold
  
- ▶ Discrimination threshold
  - ▶ Classifier outputs scores
  - ▶ How do you assign scores to classes?
    - I.e.,  $< 0.5 = 0$  &  $> 0.5 = 1$

# ROC CONT.

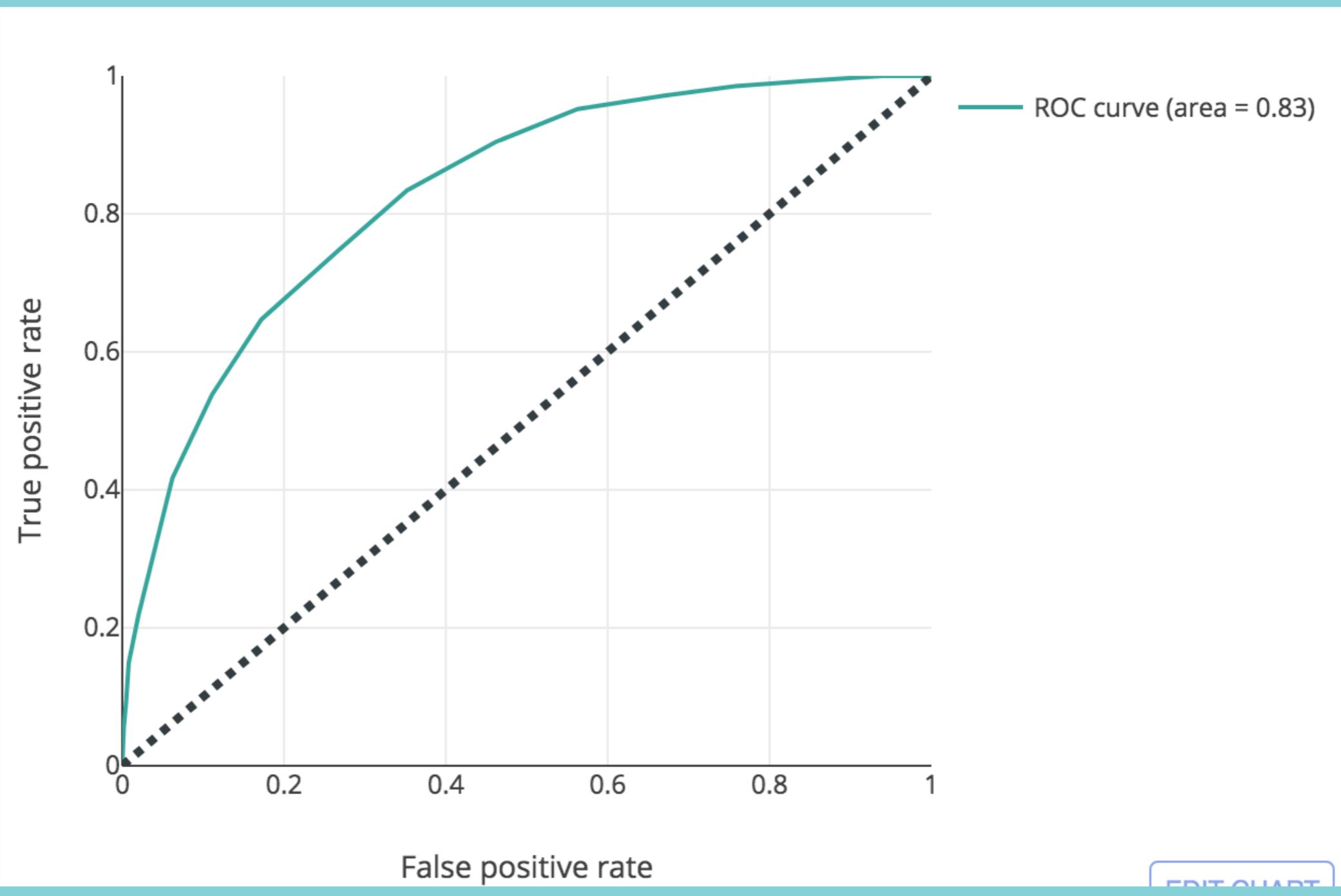
Four potential outcomes (from confusion matrix)

- ° ROC: x-axis = false positive rate

- ° y-axis = true positive rate

Purpose

- ° Model selection
- ° Threshold selection



EDIT CHART

# ROC RECAP

## Interpretation

- ▶ Top-left corner: optimal, perfect model
- ▶ Diagonal: line of no discrimination
- ▶ AUC: probability that a randomly selected 1 has a higher score than a randomly selected 0

## Use

- ▶ Choosing a mode: higher AUC & bowed towards top left
- ▶ Choosing a threshold based on desired sensitivity and specificity

# PRC CURVE:

AKA PRECISION-RECALL CURVE

Y

PRECISION VS. RECALL (SENSITIVITY)

Y-AXIS: PRECISION ( $TP/(TP+FP)$ )

X-AXIS: RECALL ( $TP/(TP+FN)$ )

X

# ROC & PRC EXAMPLE

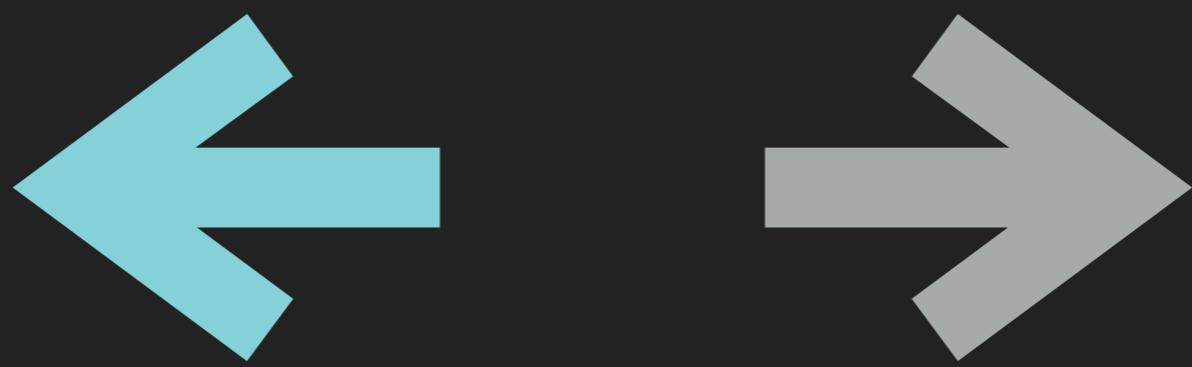
SIMULATED DATA: PREDICT GENDER USING ONLY HEIGHT

CASE 1: IMAGINE ALL MEN ARE TALLER THAN ALL WOMEN

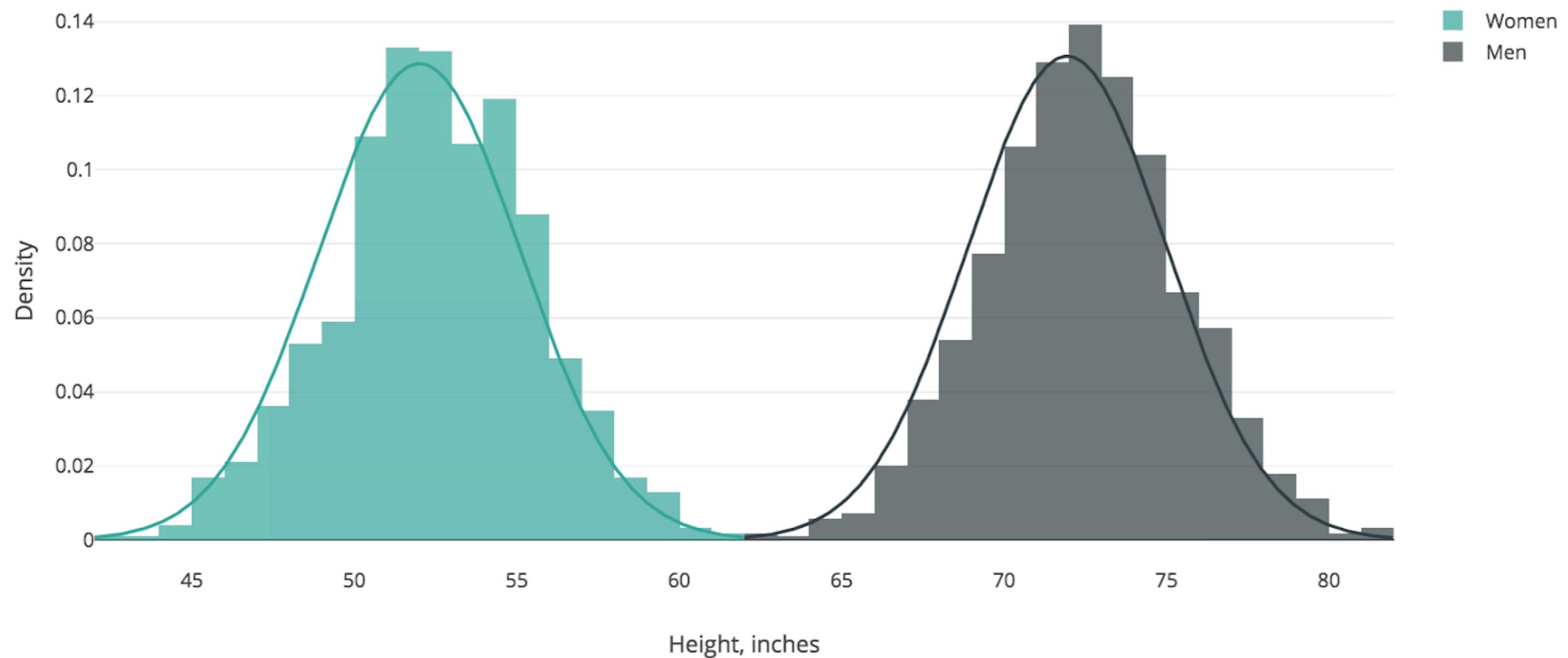
LOGISTIC REGRESSION, PREDICTING PROBABILITY OF BEING A MAN USING WEIGHT

EASY TO SEPARATE

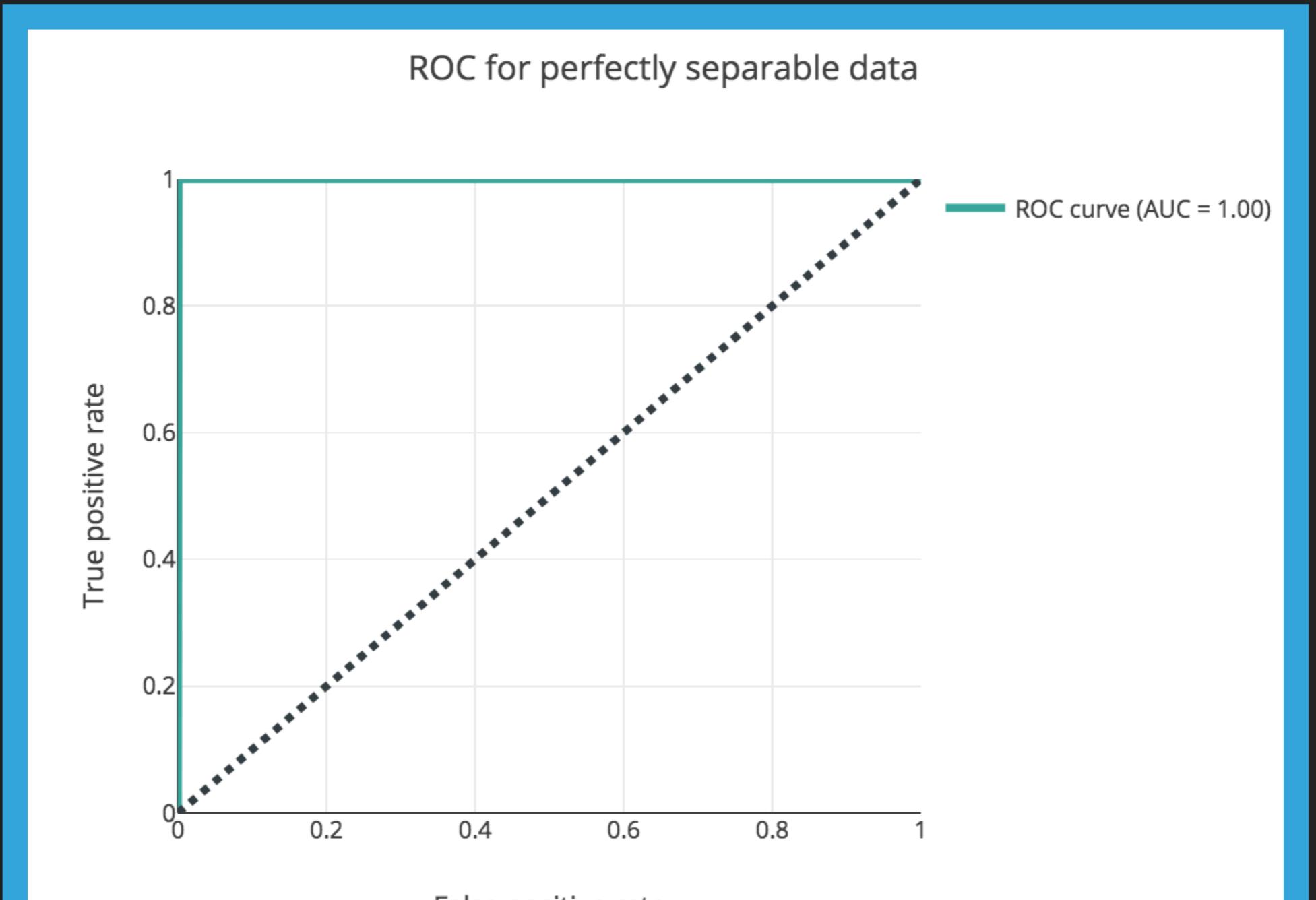




Men and Women Completely Distinct



## CASE 1



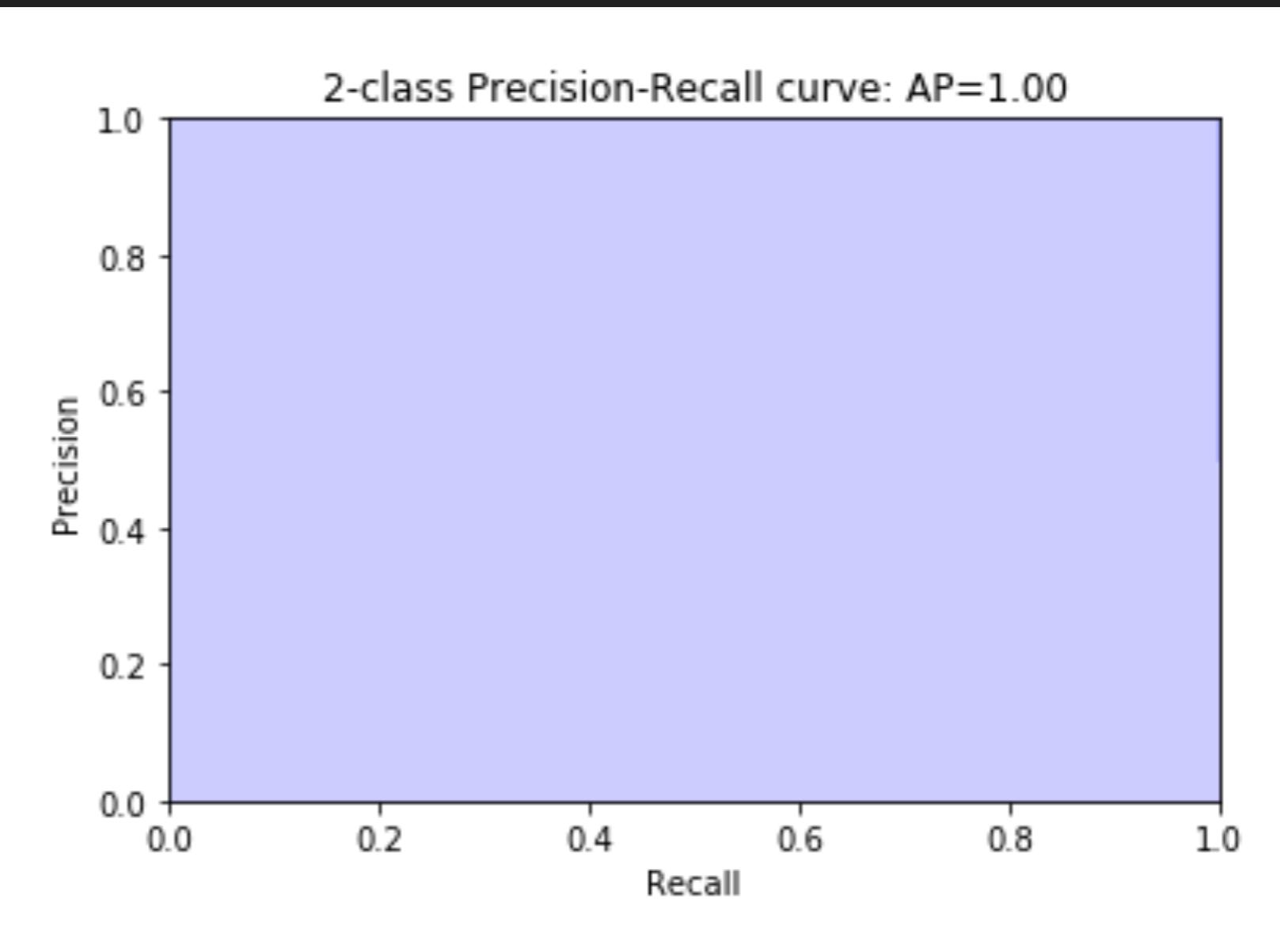
	height	is_man	lr_pred
0	74.0	1	0.973436
1	73.0	1	0.964881
2	68.0	1	0.866831

# PRC EXAMPLE WITH HEIGHT DATA

**PERFECT CURVE:**

**NO OVERLAP**

**100% ACCURACY**



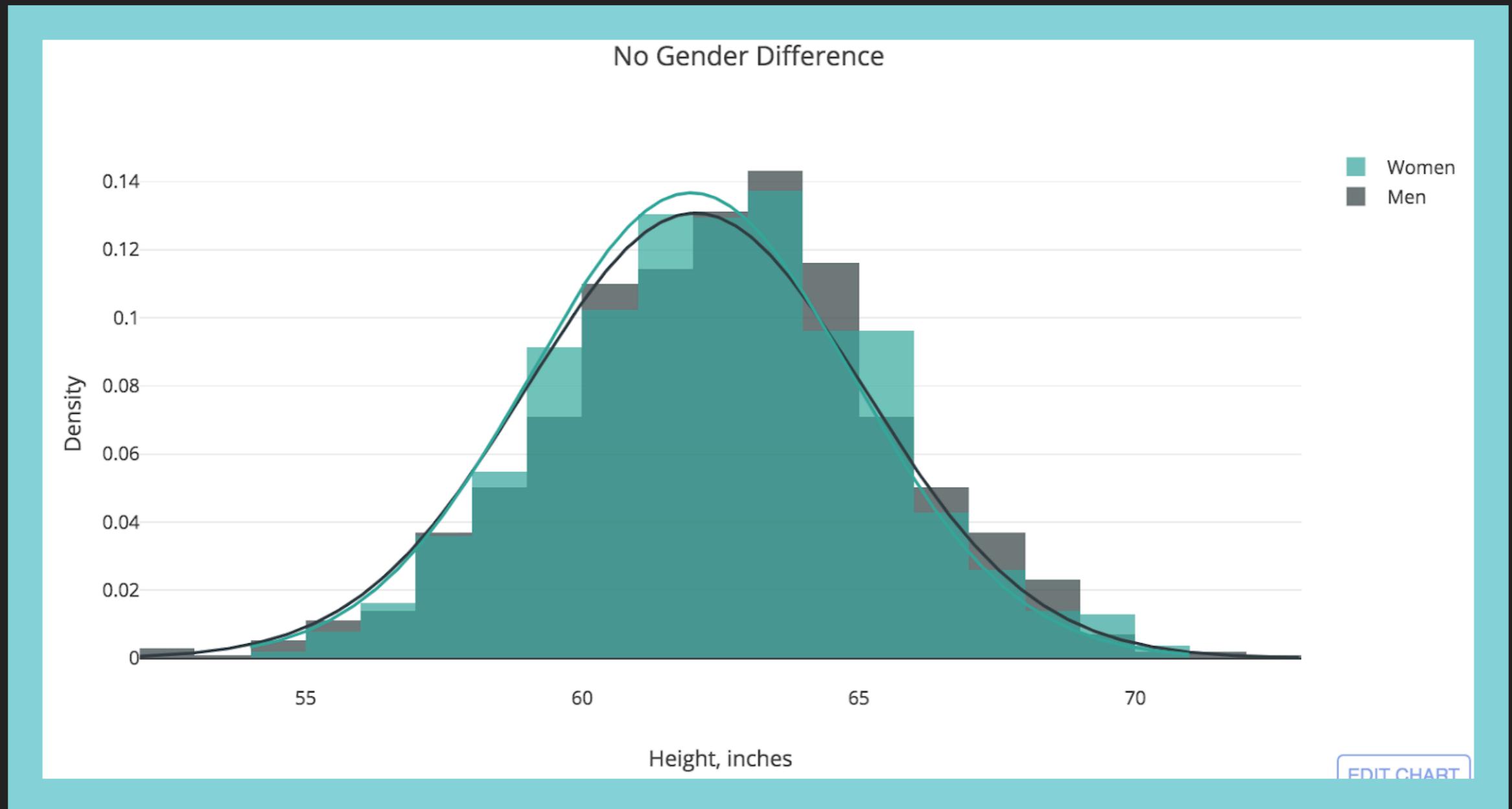
# ROC EXAMPLE

SIMULATED DATA: PREDICT GENDER USING ONLY HEIGHT

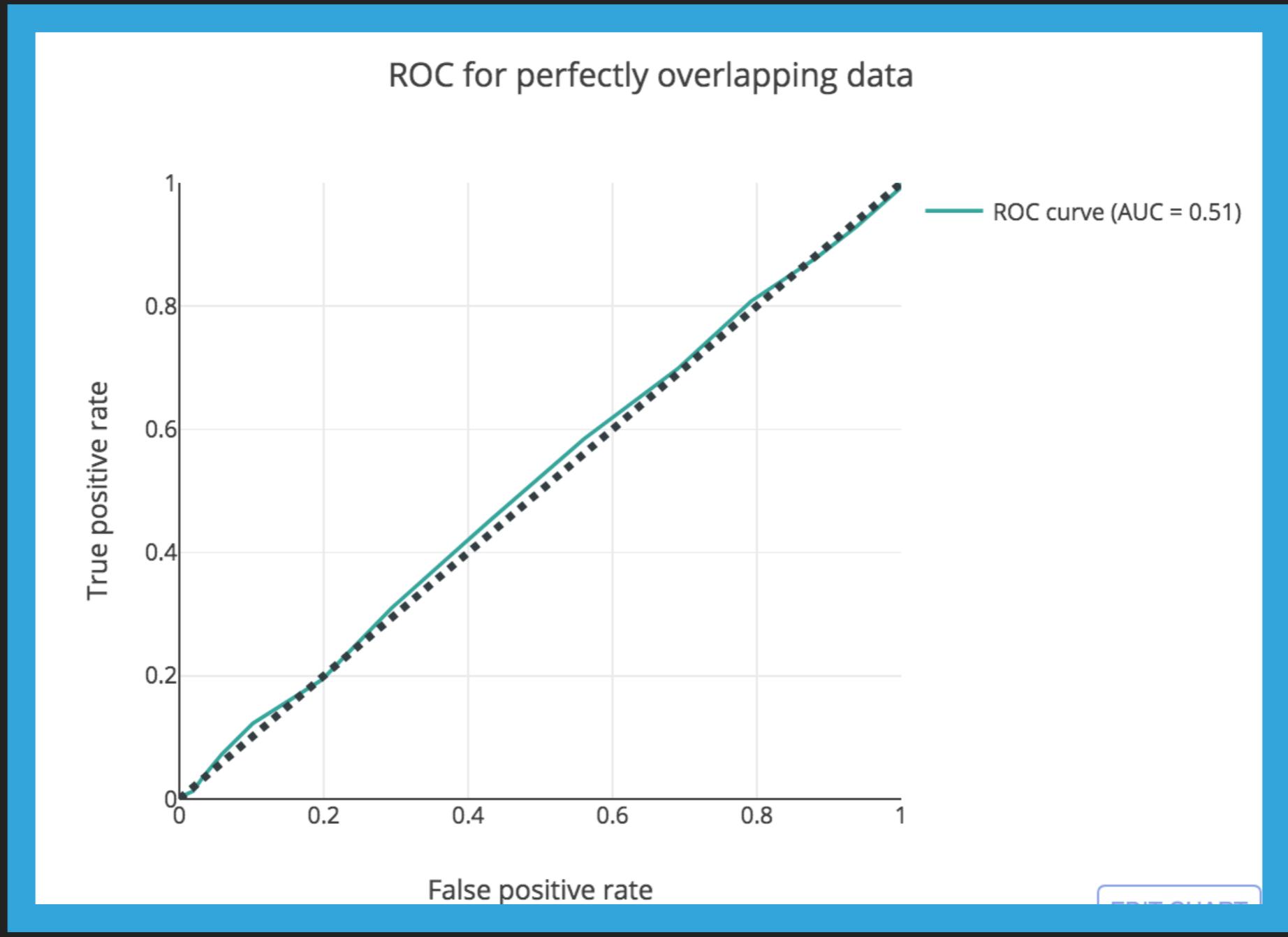
CASE 1: IMAGINE NO SYSTEMATIC HEIGHT DIFFERENCE BY GENDER

LOGISTIC REGRESSION FINDS IT IMPOSSIBLE TO PREDICT ANY BETTER THAN CHANCE





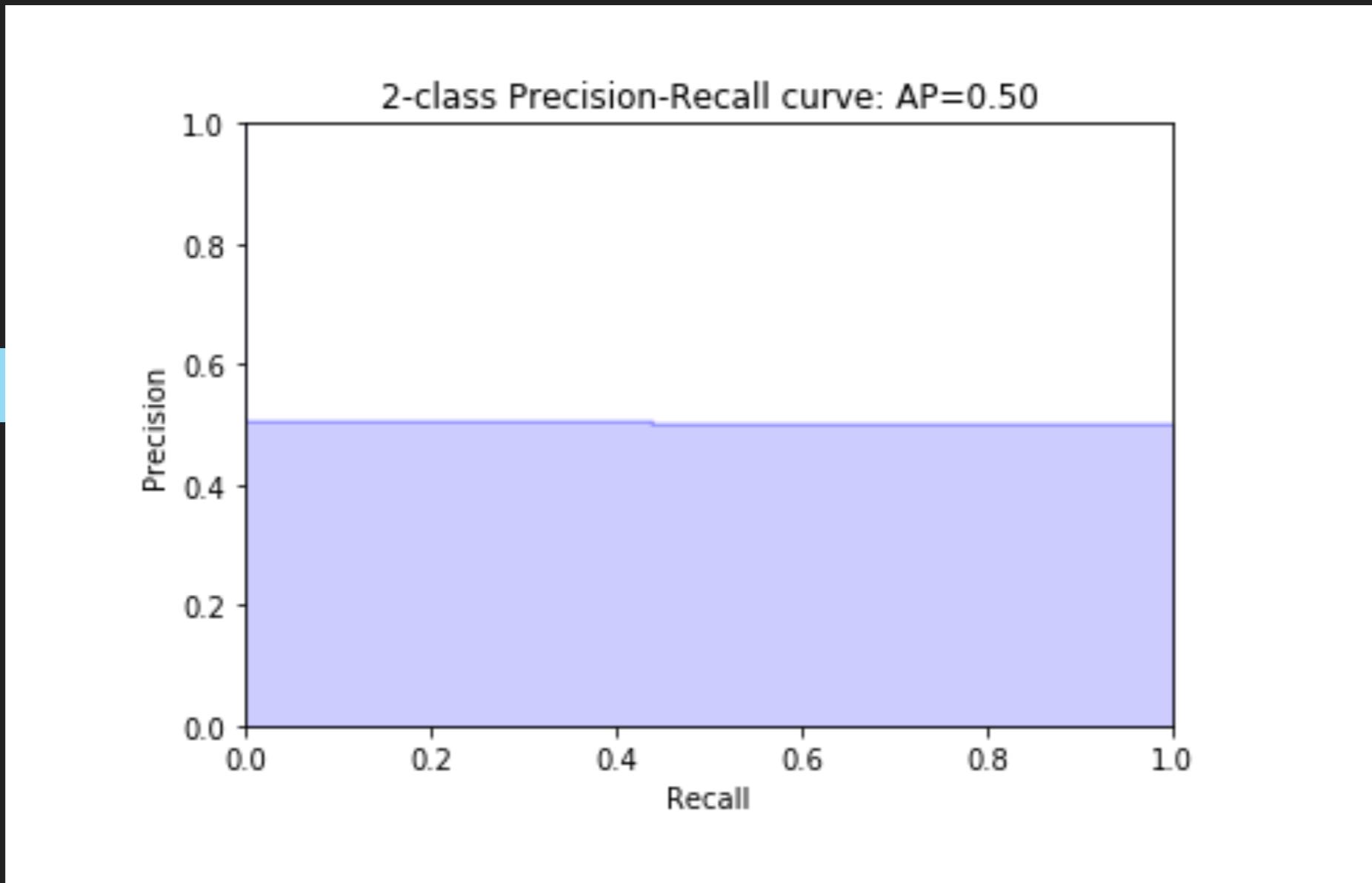
## CASE 2



	height	is_man	lr_pred
0	66.0	1	0.504591
1	64.0	1	0.502361
2	59.0	1	0.496786

# PRC CURVE

SINCE THE DATA OVERLAP AND ARE INDISTINGUISHABLE,  
THE PREDICTION IS ESSENTIALLY A 50/50 GUESS

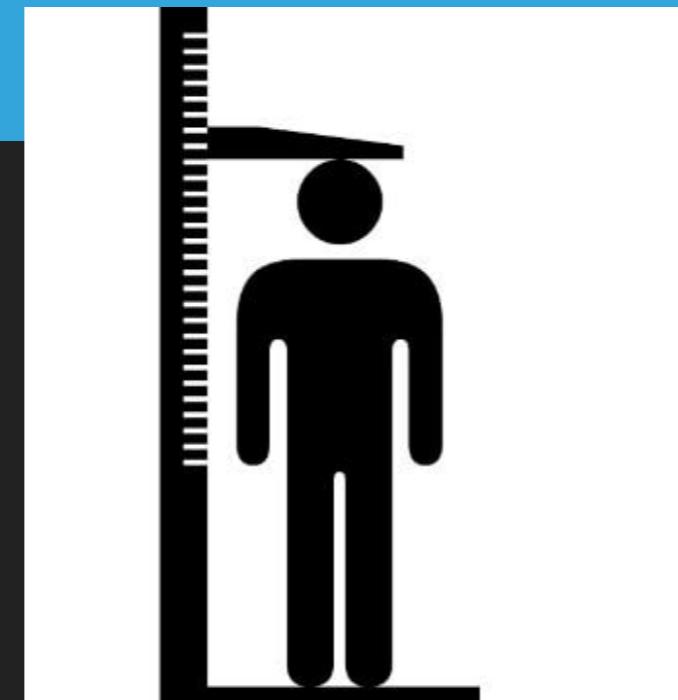
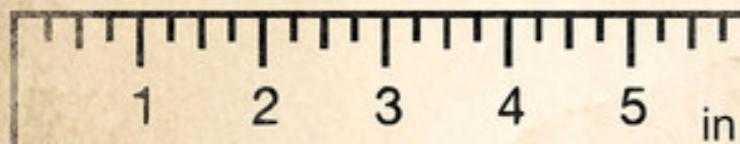


# ROC EXAMPLE

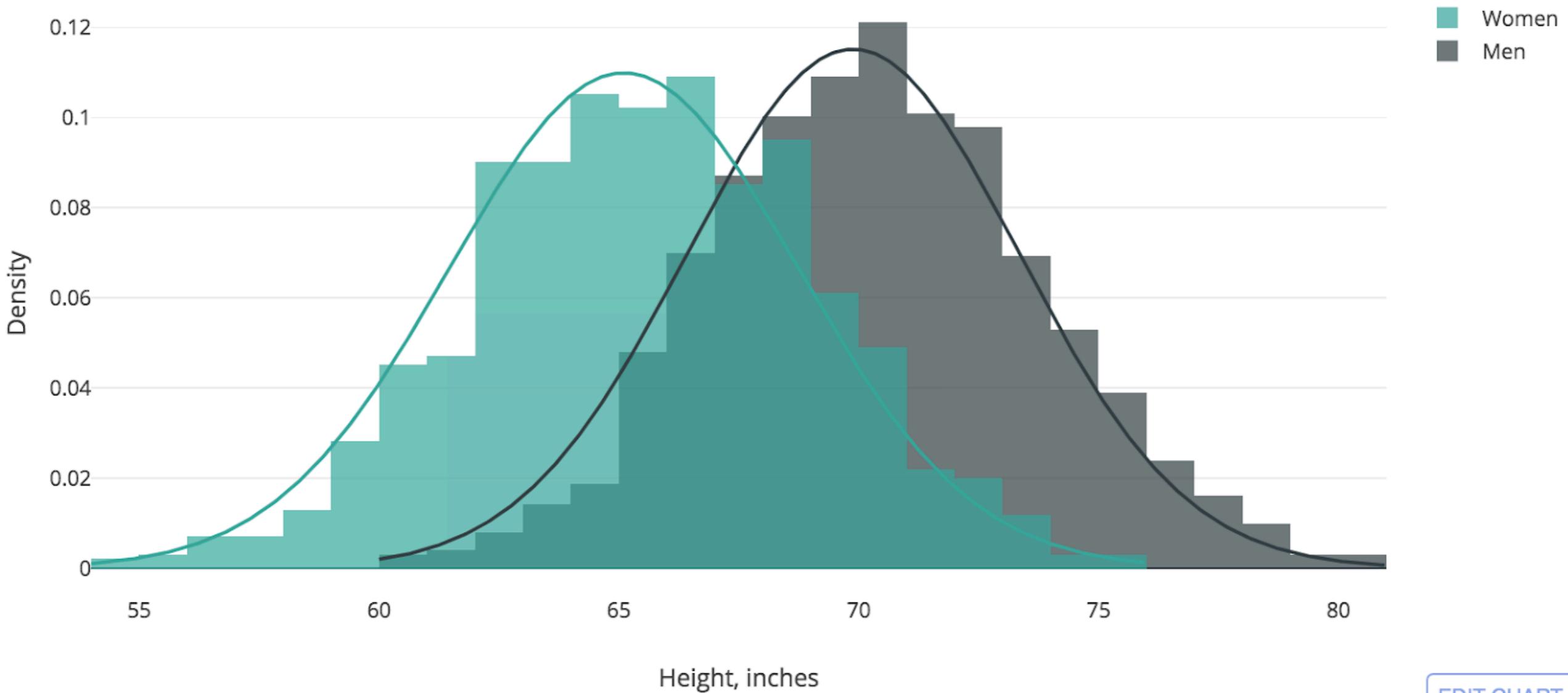
CASE 3: REAL-WORLD GENDER HEIGHT DIFFERENCES

MEN APPROX. 5 INCHES TALLER, SLIGHTLY LARGER STANDARD DEVIATION

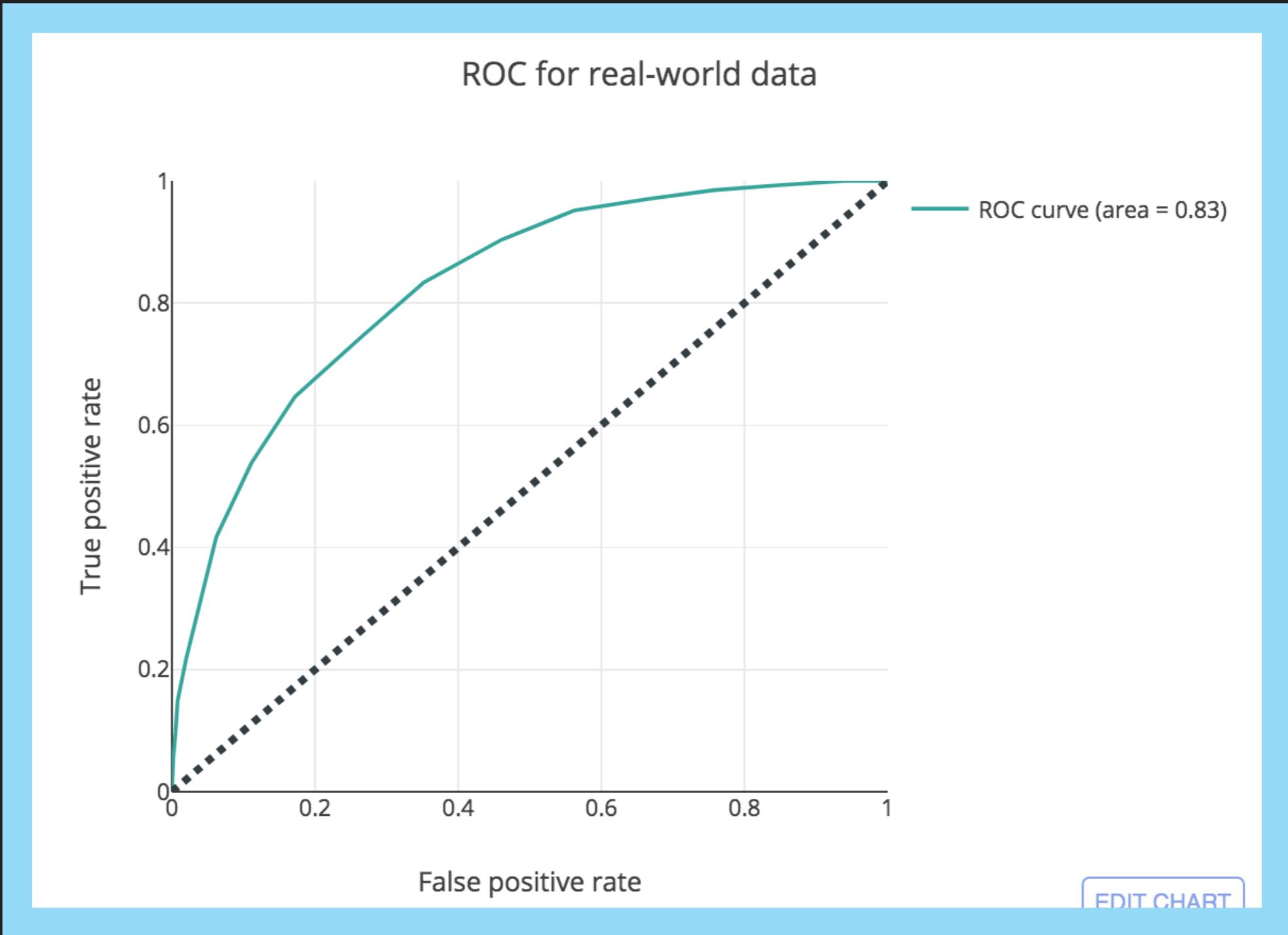
LOGISTIC REGRESSION HAS SOME INFORMATION, BUT THERE IS STILL NOISE



## Real Heights

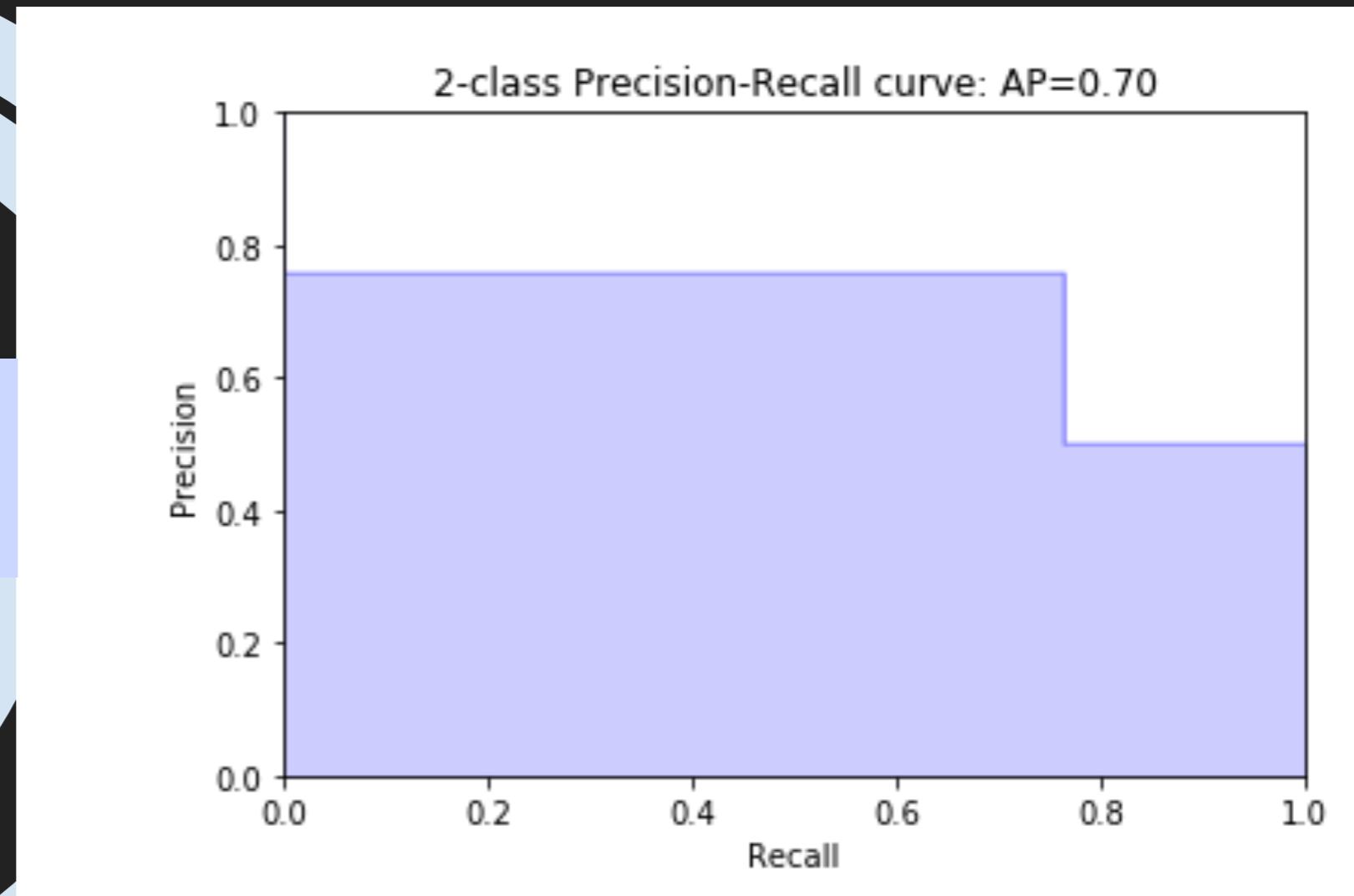
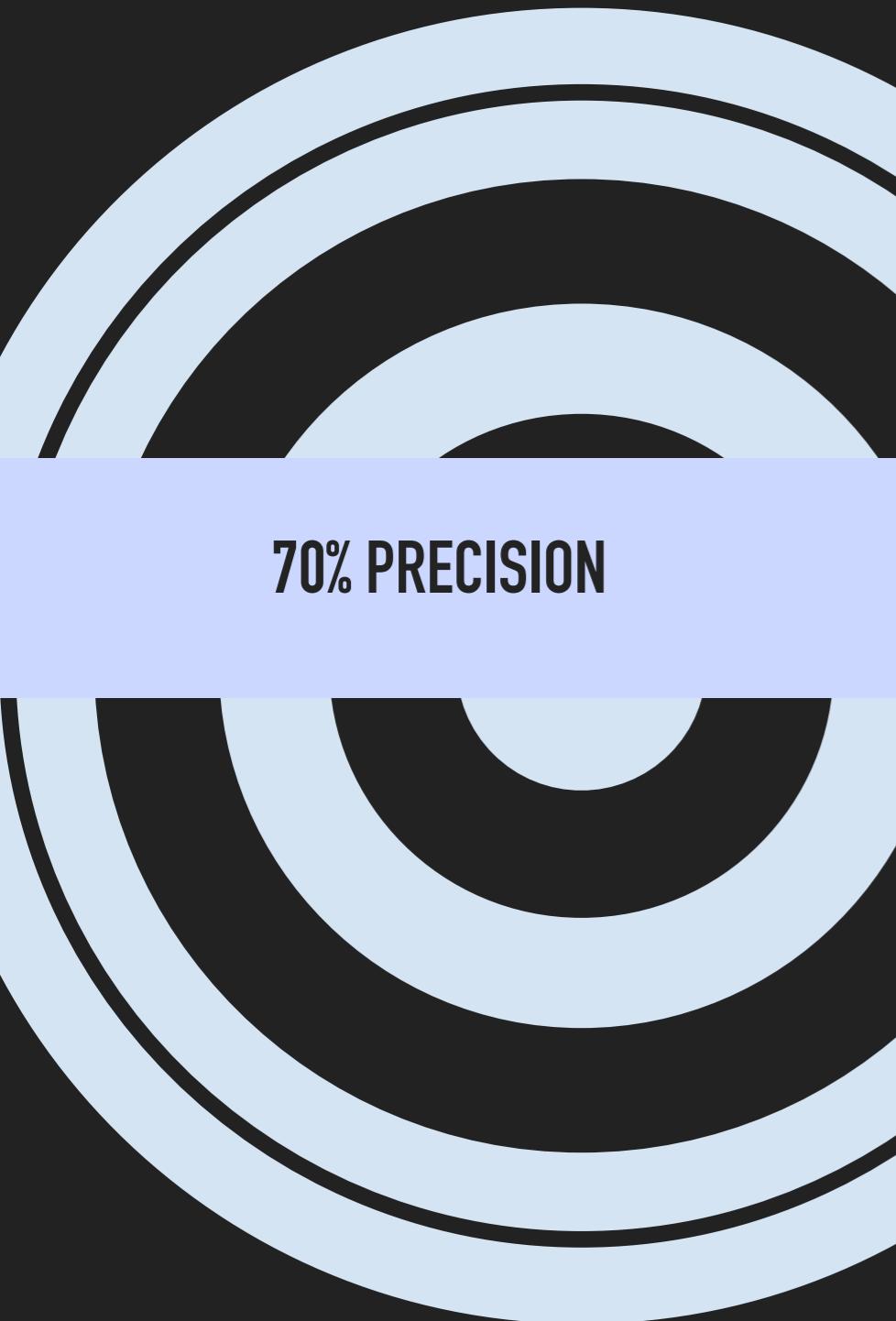


## CASE 3



	height	is_man	lr_pred
3	71.0	1	0.66723
4	74.0	1	0.78015
5	74.0	1	0.78015

# PRC CURVE



# WRAP-UP

We looked over:

- ▶ Accuracy and why not to rely solely on this measurement
- ▶ Confusion Matrix and it's Measurements:
  - ▶ Sensitivity & Specificity
  - ▶ Precision
  - ▶ F1 Score
- ▶ ROC Curve
  - ▶ Case 1: No systematic difference by class
  - ▶ Case 2: Perfectly overlapping data
  - ▶ Case 3: Real-World data
- ▶ PRC Curve
  - ▶ Case 1: No systematic difference by class
  - ▶ Case 2: Perfectly overlapping data
  - ▶ Case 3: Real-World data



Importance

With this wide array of performance metrics, modelers can choose those that benefit their specific classification scenario and objectives

They are a substantial improvement over using simple accuracy as a one-and-done evaluator

# QUESTIONS?

# RESOURCES

## Images

<https://www.inc.com/lisa-calhoun/4-predictions-for-artificial-intelligence-in-2017.html>

<https://becominghuman.ai/machine-learning-for-dummies-explained-in-2-mins-e83fb55ac6d>

<https://www.hypehartford.com/what-we-do/blog/hype-blog/2015/07/13/taking-the-mystery-out-of-metrics>

[https://www.plant-world-seeds.com/store/view\\_seed\\_item/4062](https://www.plant-world-seeds.com/store/view_seed_item/4062)

[https://www.plant-world-seeds.com/store/view\\_seed\\_item/4062](https://www.plant-world-seeds.com/store/view_seed_item/4062)

[https://www.plant-world-seeds.com/store/view\\_seed\\_item/3664](https://www.plant-world-seeds.com/store/view_seed_item/3664)

[https://www.etsy.com/uk/market/5\\_inch\\_ruler](https://www.etsy.com/uk/market/5_inch_ruler)

<http://jpberry.com/blog/genderrolls>

<http://www.wetpaint.com/celebrity-couple-height-difference-photos-1486077/>

<https://pixels.com/featured/iris-flowers-artwork-purple-irises-9-botanical-garden-floral-art-baslee-troutman-baslee-troutman-art-prints-giclee.html>