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

$$\frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

- 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

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

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

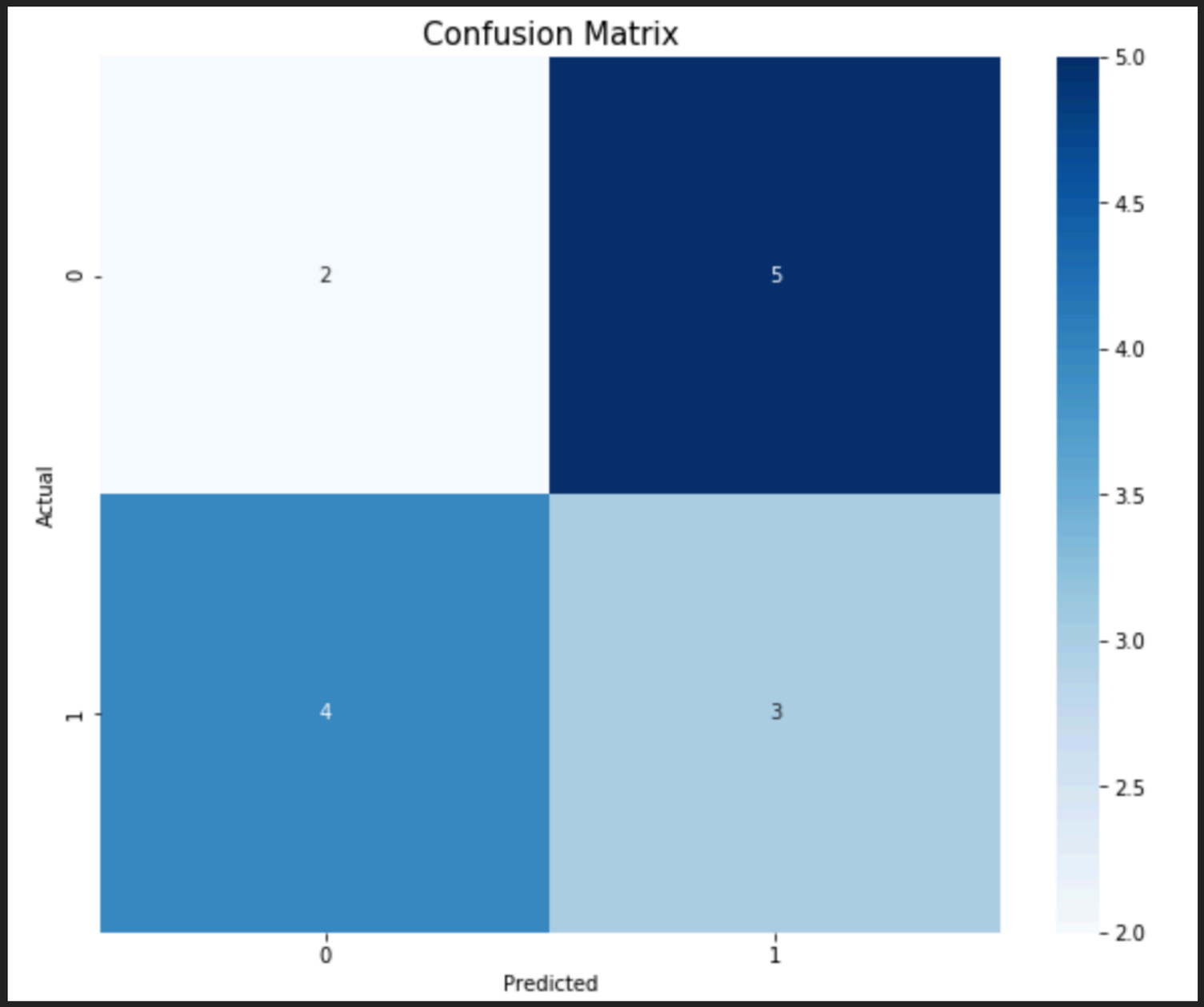
- Predicted Values: Positive or Negative
- Actual Values: True or False

- 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 1 Error)
- FN: Our prediction is no, and it's false (Type II Error)

Disease Example:

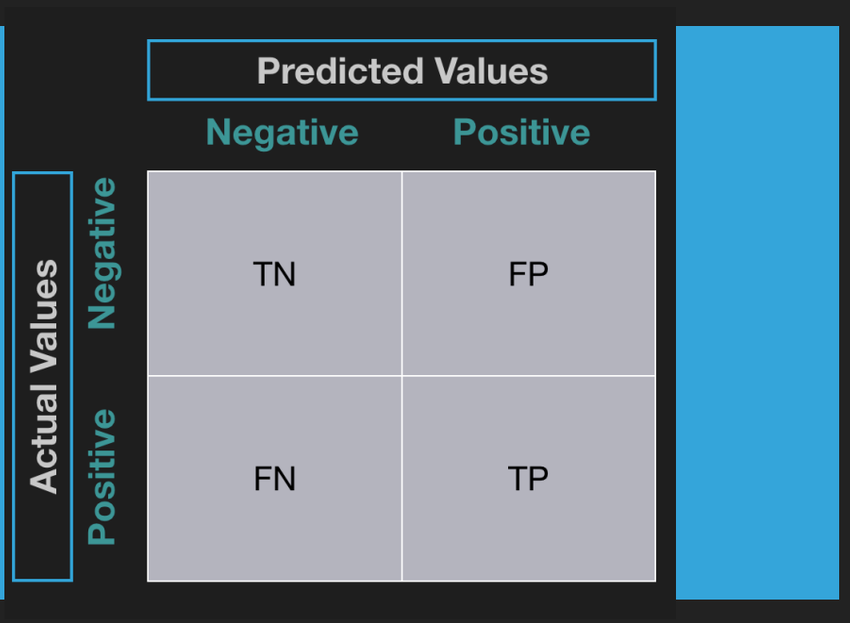
0 = Disease Free

1 = Has Disease



SENSITIVITY & SPECIFICITY

$$\frac{TP}{TP+FN} \qquad \frac{TN}{TN+FP}$$



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

$$\frac{\text{TRUE POSITIVE}}{\text{TRUE POSITIVES + FALSE POSITIVES}}$$

- 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

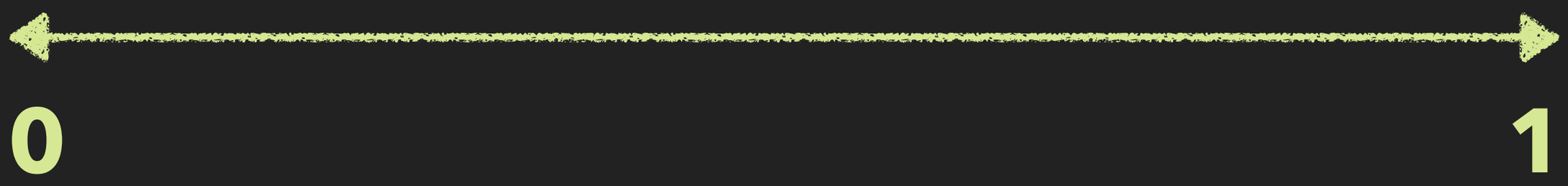
$$\frac{2 \times (\text{PRECISION} \times \text{RECALL})}{\text{PRECISION} + \text{RECALL}}$$

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

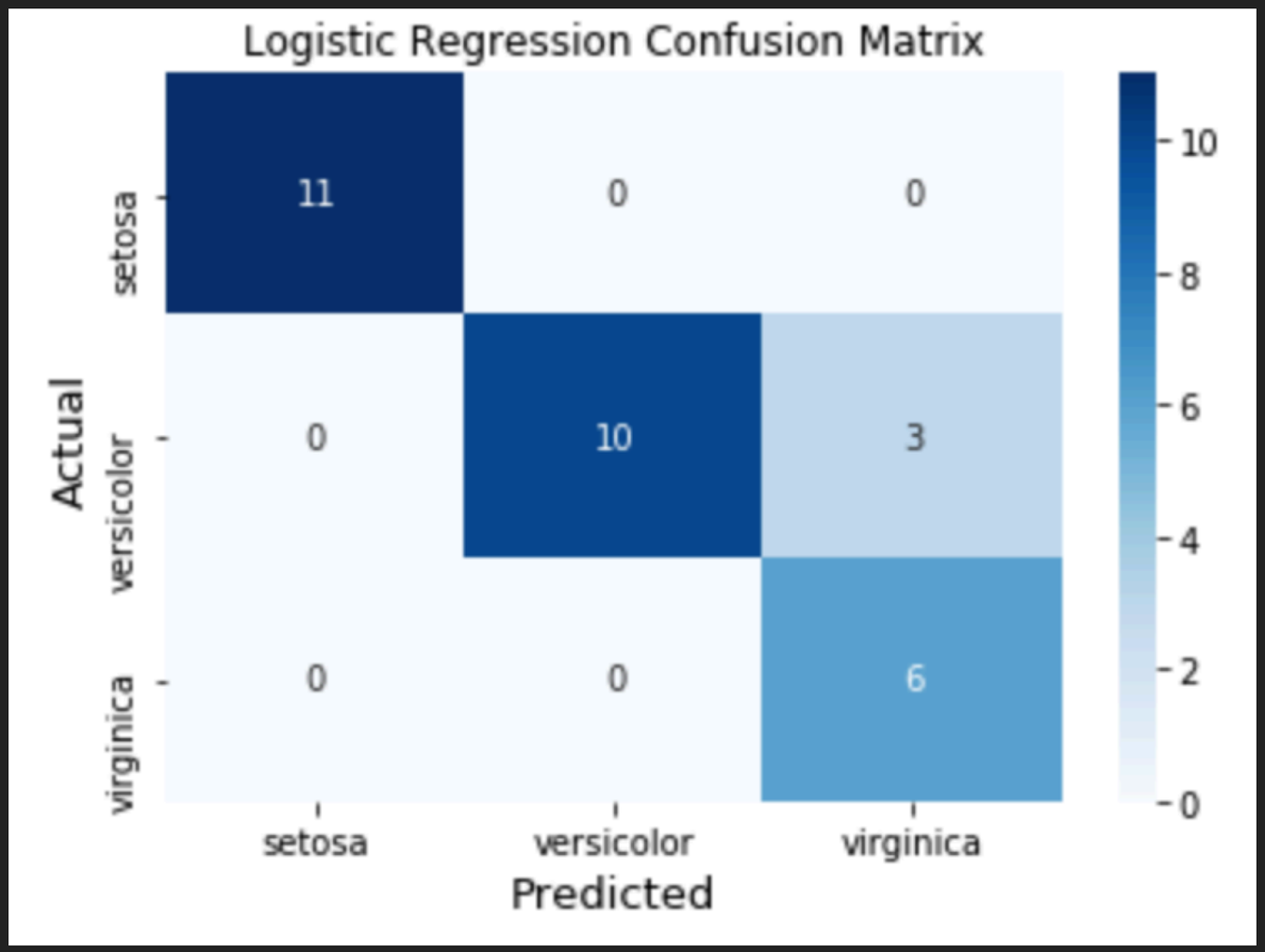
F1 score is perfect when it’s 1, and model has failed when it’s 0

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.

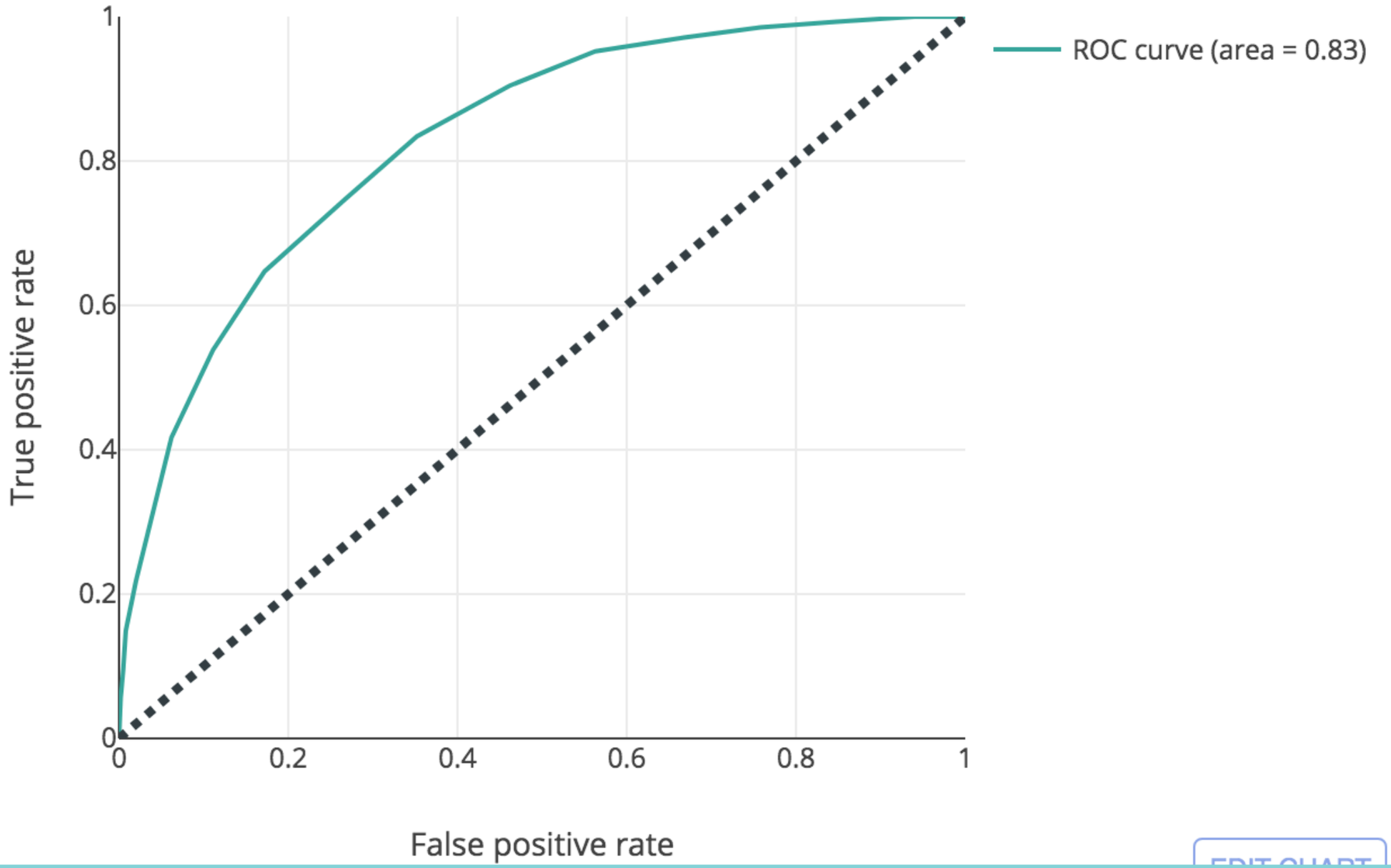
Four potential outcomes (from confusion matrix)

- ROC: x-axis = false positive rate

  y-axis = true positive rate

Purpose

- Model selection
- Threshold selection


1
0.8
0.6
0.4
0.2
0
True positive rate
0
0.2
0.4
0.6
0.8
1
False positive rate
ROC curve (area = 0.83)

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

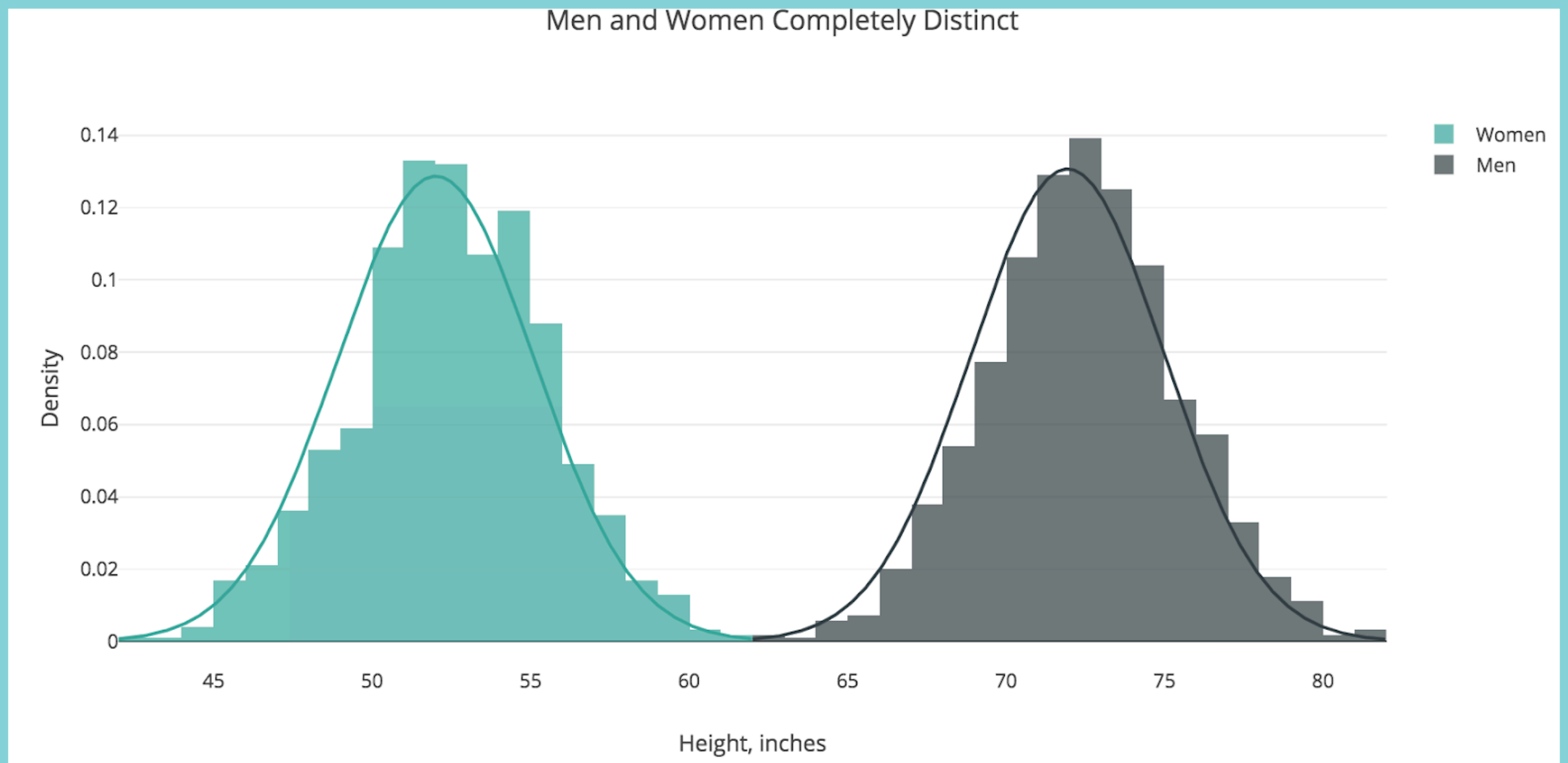
ROC & PRC EXAMPLE

SIMULATED DATA: PREDICT GENDER USING ONLY HEIGHT

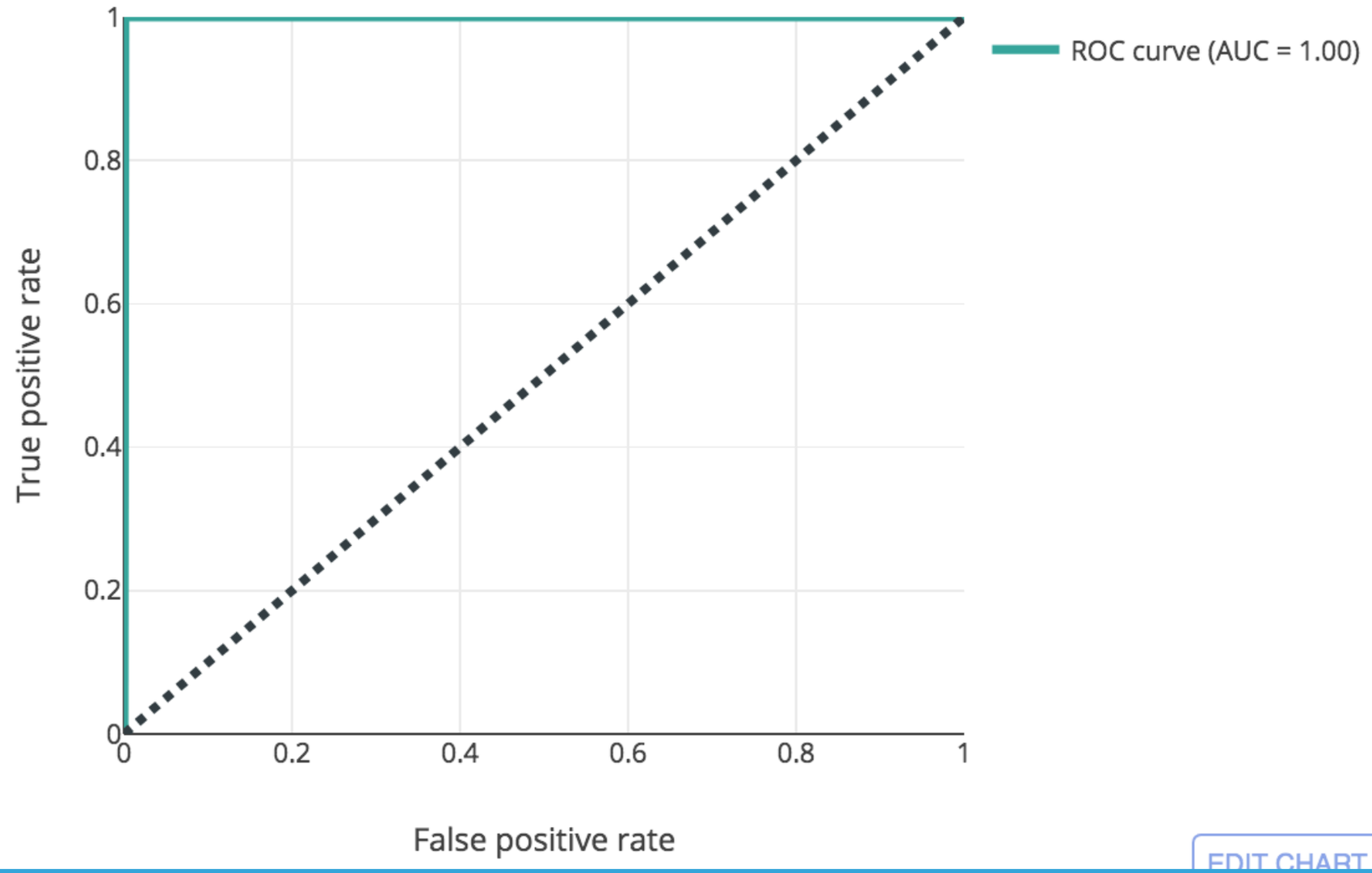
CASE 1: IMAGINE ALL MEN ARE TALLER THAN ALL WOMEN

LOGISTIC REGRESSION, PREDICTING PROBABILITY OF BEING A MAN USING WEIGHT

EASY TO SEPARATE


Men and Women Completely Distinct
Women
Men
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Density
45
50
55
60
65
70
75
80
Height, inches

CASE 1



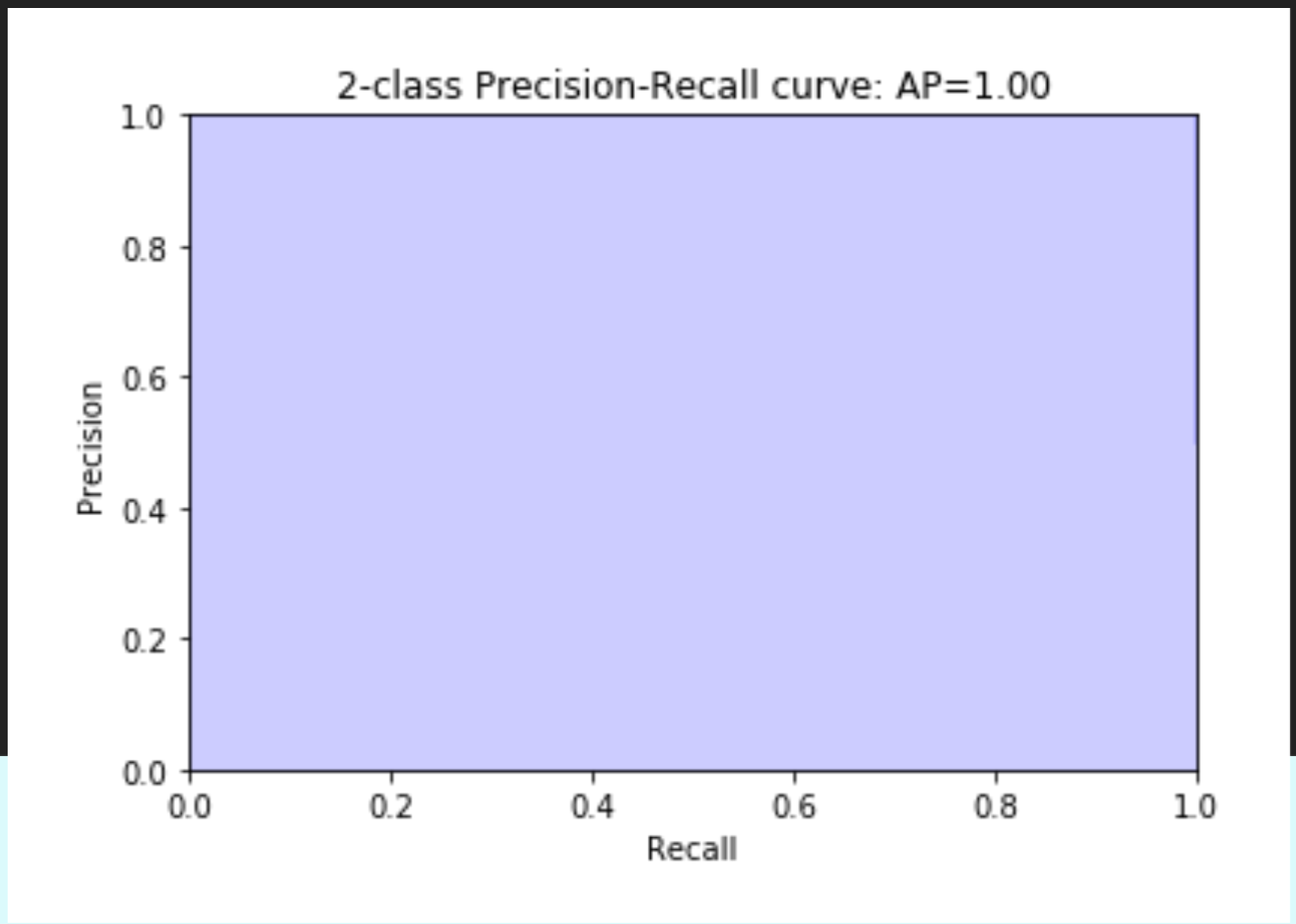
	height	is_man	lr_pred
0	74.0	1	0.973436
1	73.0	1	0.964881
2	68.0	1	0.866831

PRC EXAMPLE WITH HEIGHT DATA

PERFECT CURVE:

NO OVERLAP

100% ACCURACY

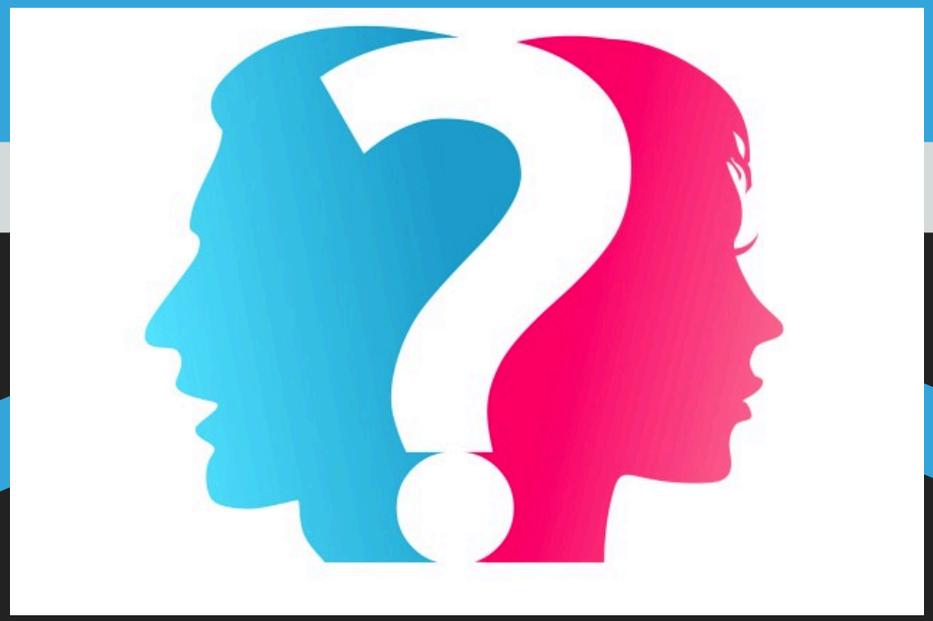


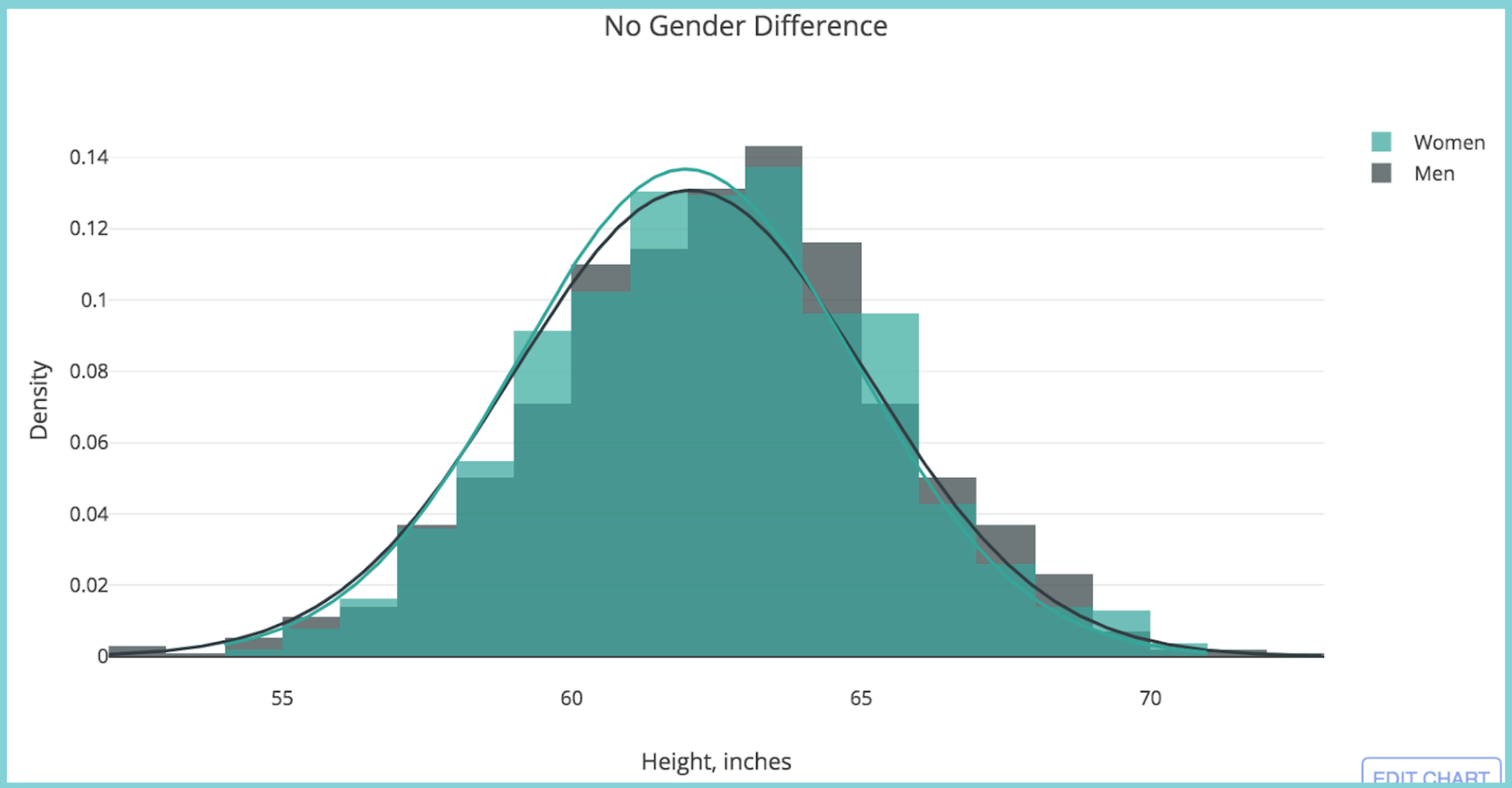
ROC EXAMPLE

SIMULATED DATA: PREDICT GENDER USING ONLY HEIGHT

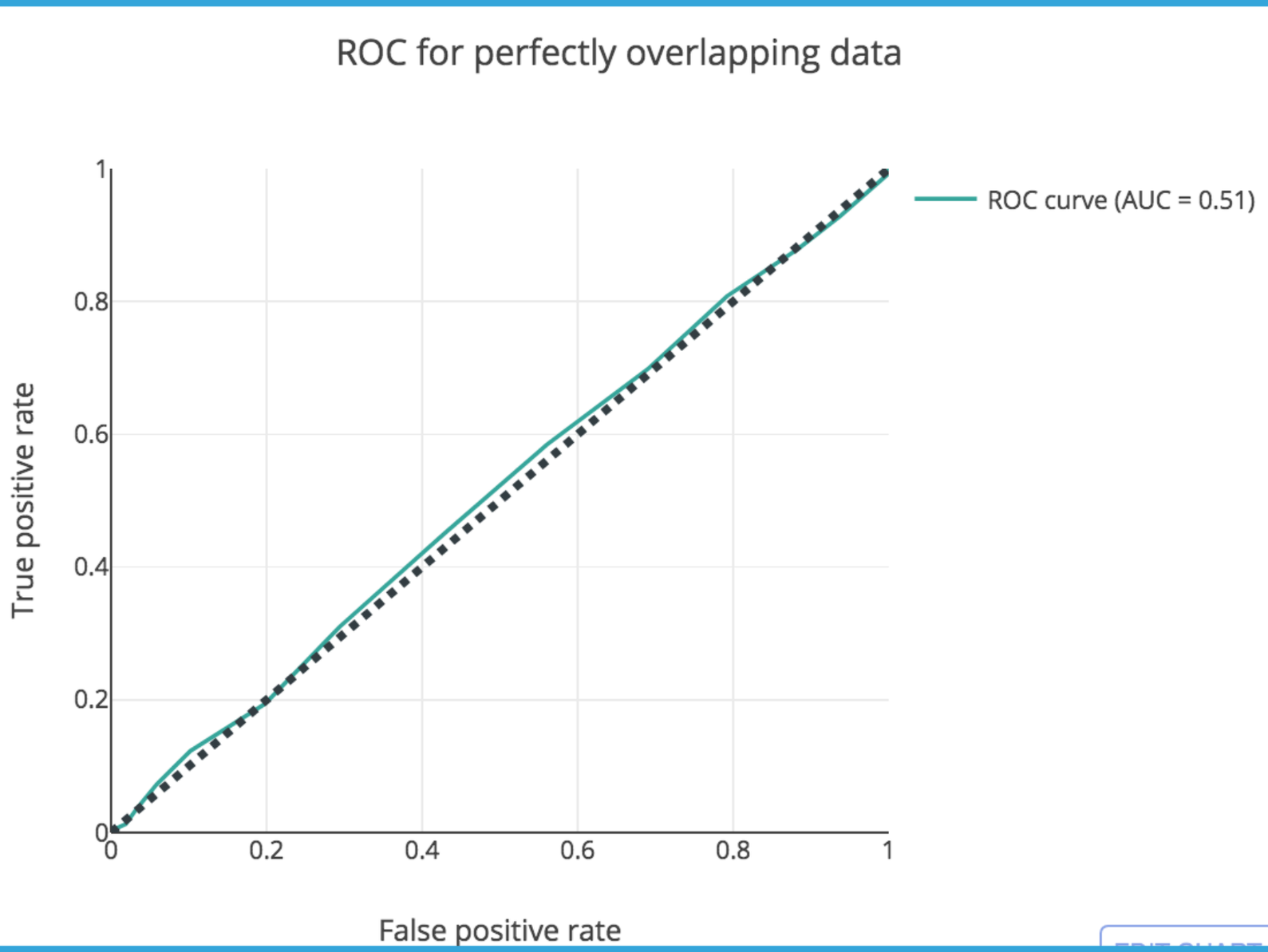
CASE 2: IMAGINE NO SYSTEMATIC HEIGHT DIFFERENCE BY GENDER

LOGISTIC REGRESSION FINDS IT IMPOSSIBLE TO PREDICT ANY BETTER THAN CHANCE




No Gender Difference
Women
Men
Density
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
55
60
65
70
Height, inches
EDIT CHART

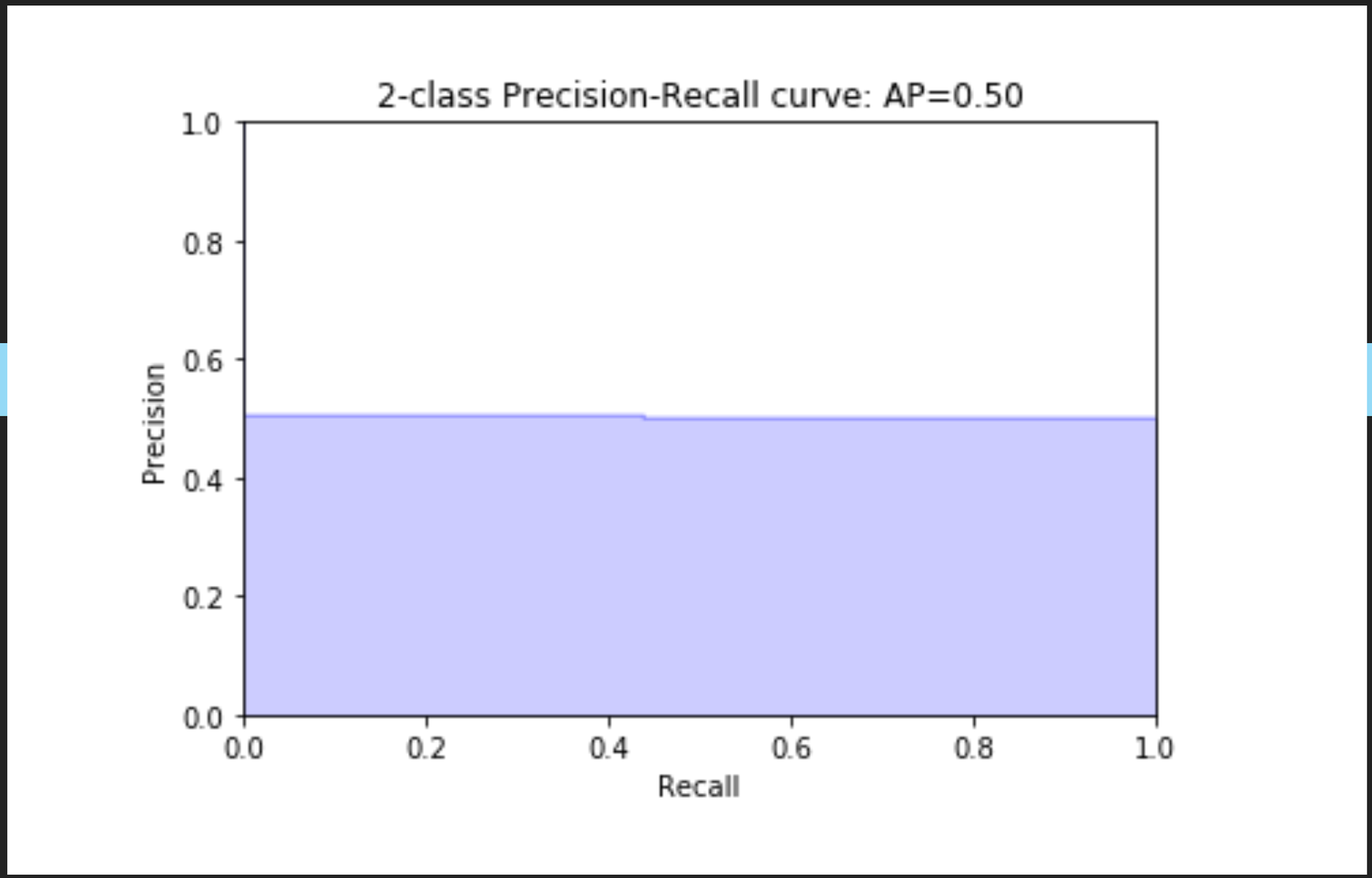
CASE 2



	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

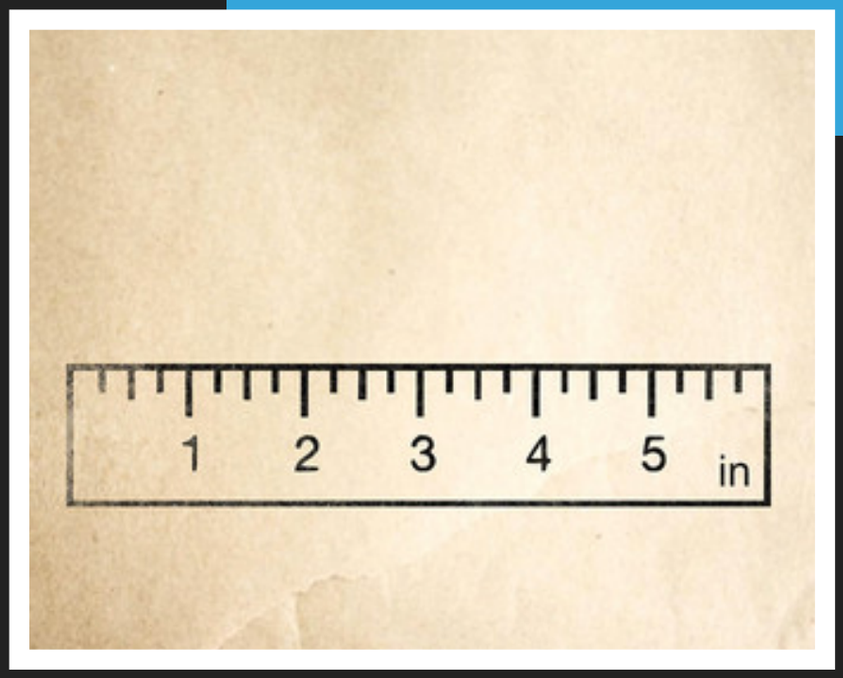


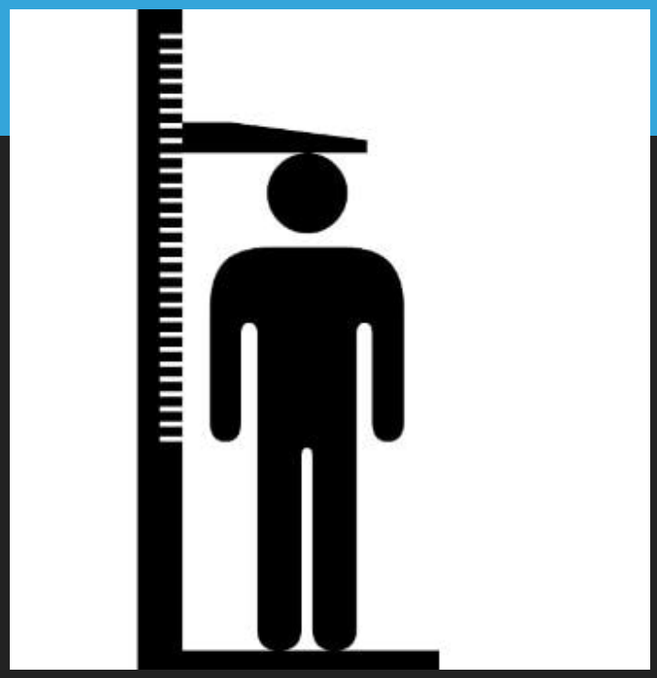
ROC EXAMPLE

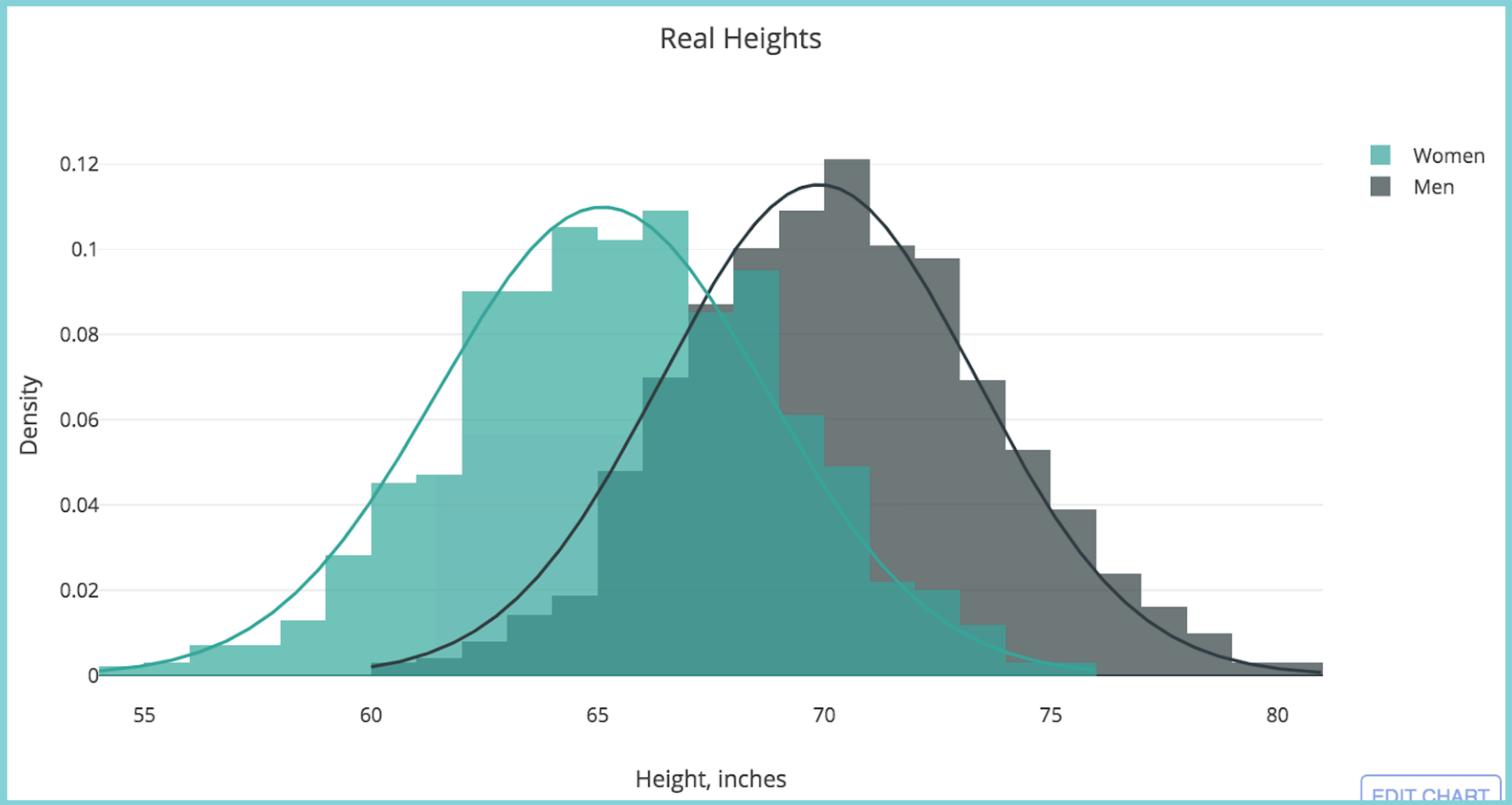
CASE 3: REAL-WORLD GENDER HEIGHT DIFFERENCES

MEN APPROX. 5 INCHES TALLER, SLIGHTLY LARGER STANDARD DEVIATION

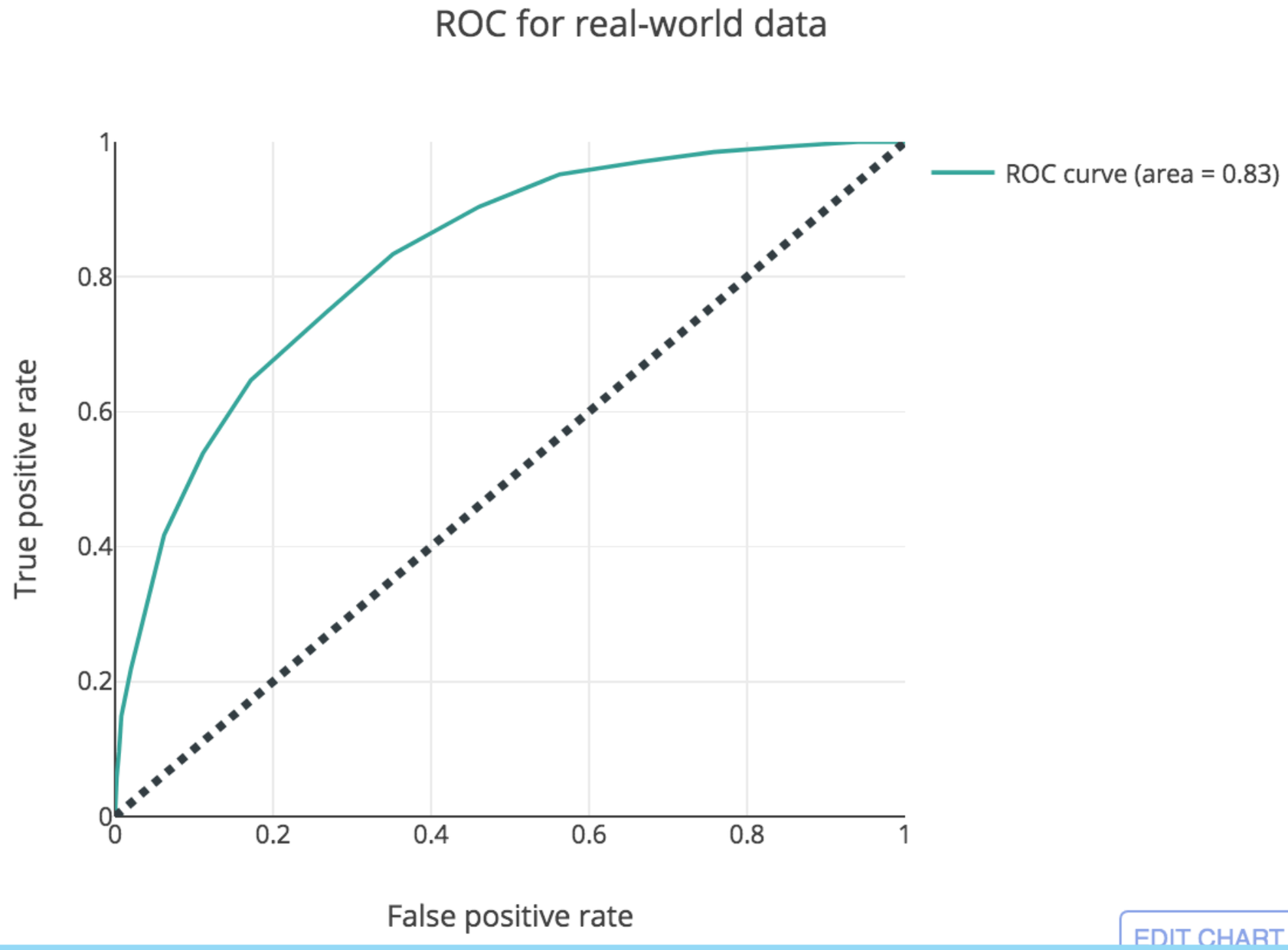
LOGISTIC REGRESSION HAS SOME INFORMATION, BUT THERE IS STILL NOISE





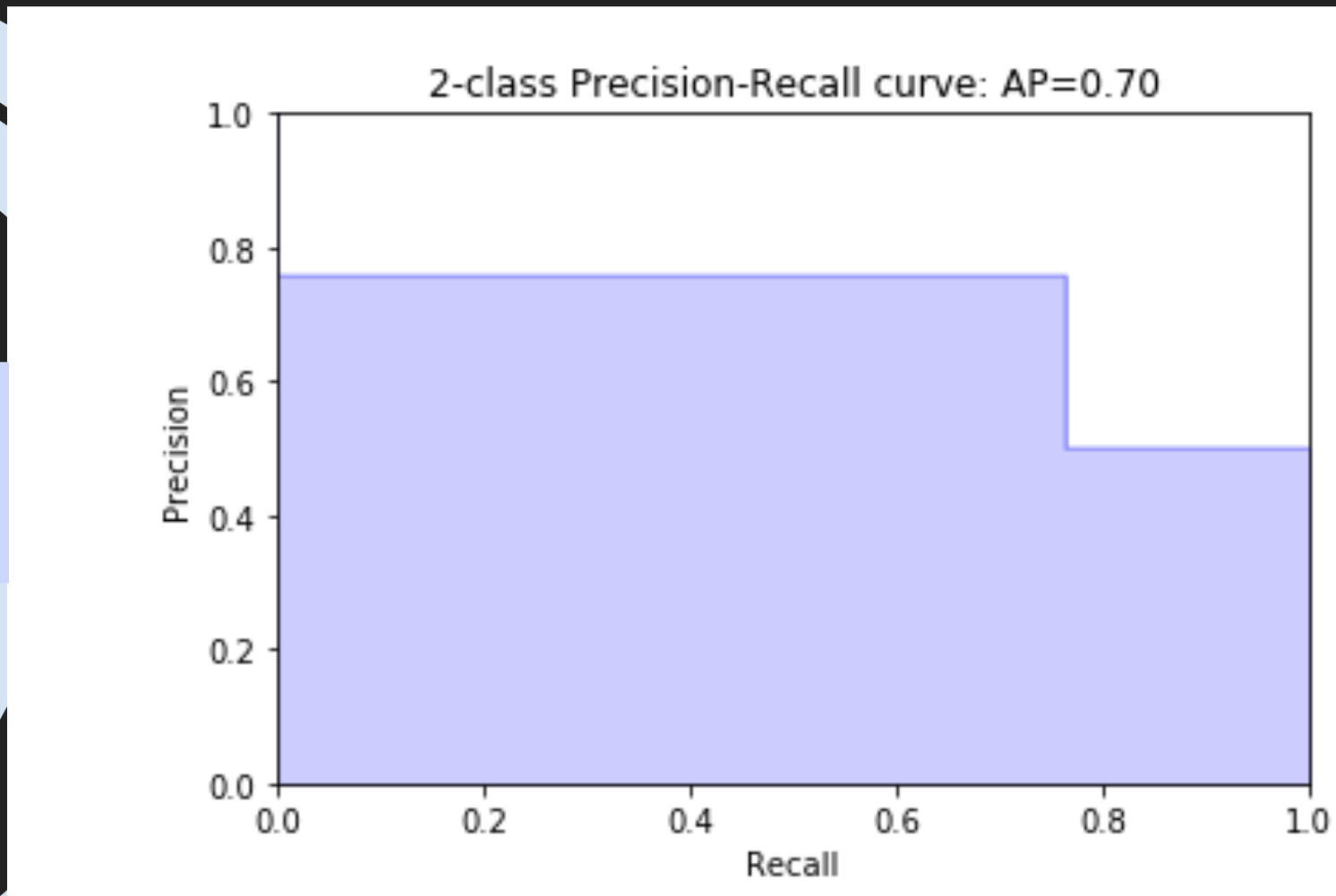

Real Heights
Women
Men
0.12
0.1
0.08
0.06
0.04
0.02
0
Density
55
60
65
70
75
80
Height, inches
EDIT CHART

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

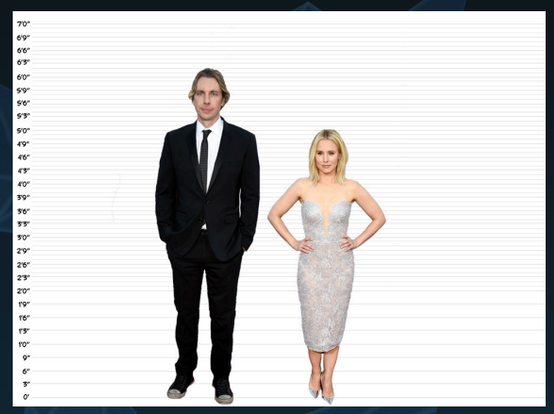


70% PRECISION

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-e83fbc55ac6d

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/3664

https://www.etsy.com/uk/market/5_inch_ruler

http://jpderryberry.com/blog/genderolls

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