Logistic Regression for Predicting Binary Responses

Dr Rebecca Barter

Labeled data

Outcome of interest: Amount spent on healthcare in next year

Labeled data

Age	Sex	Weight	Diabetes	Healthare expenses
54	M	132	N	4,300
76	F	155	Y	3,467
49	M	166	Y	103
39	F	129	N	844
47	M	177	N	6,591
70	F	192	N	8,089

Labeled data

Outcome of interest: ~~Amount~~ More than 1K spent on healthcare in next year

Age	Sex	Weight	Diabetes	Healthare expenses
54	M	132	N	4,300
76	F	155	Y	3,467
49	M	166	Y	103
39	F	129	N	844
47	M	177	N	6,591
70	F	192	N	8,089

Labeled data

Outcome of interest: More than 1K spent on healthcare in next year

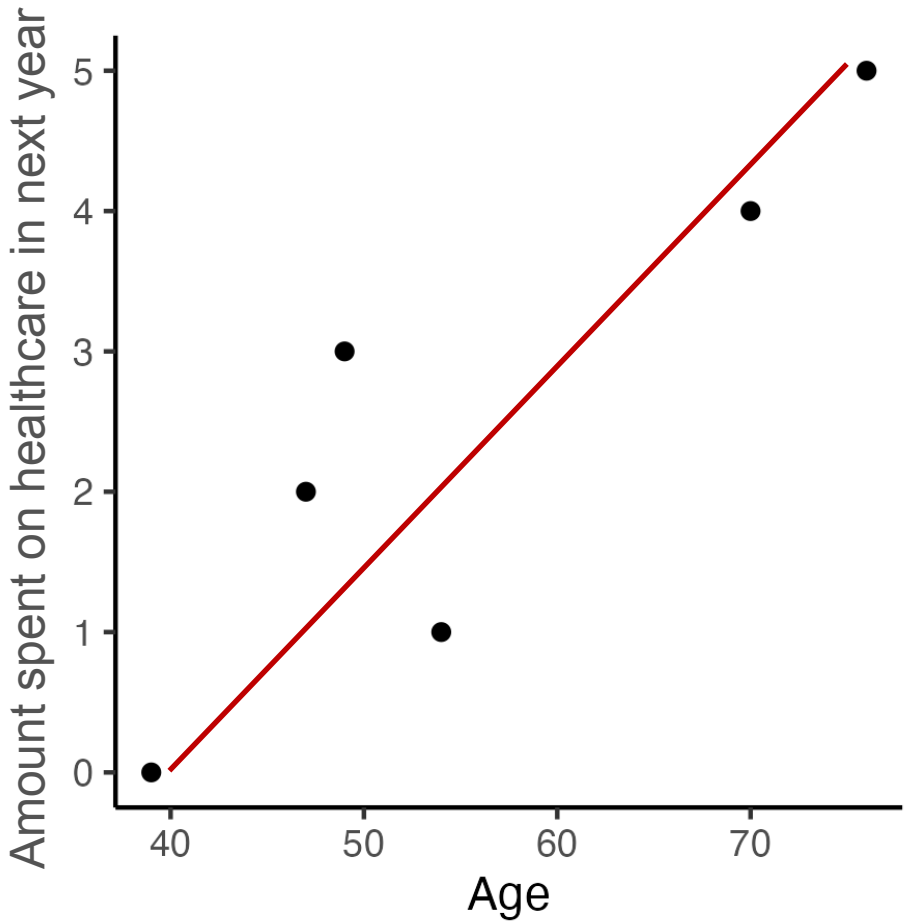
Age	Sex	Weight	Diabetes	Healthcare expenses
54	M	132	N	1
76	F	155	Y	1
49	M	166	Y	0
39	F	129	N	0
47	M	177	N	1
70	F	192	N	1

(New) Unlabeled data

Outcome of interest: More than 1K spent on healthcare in next year

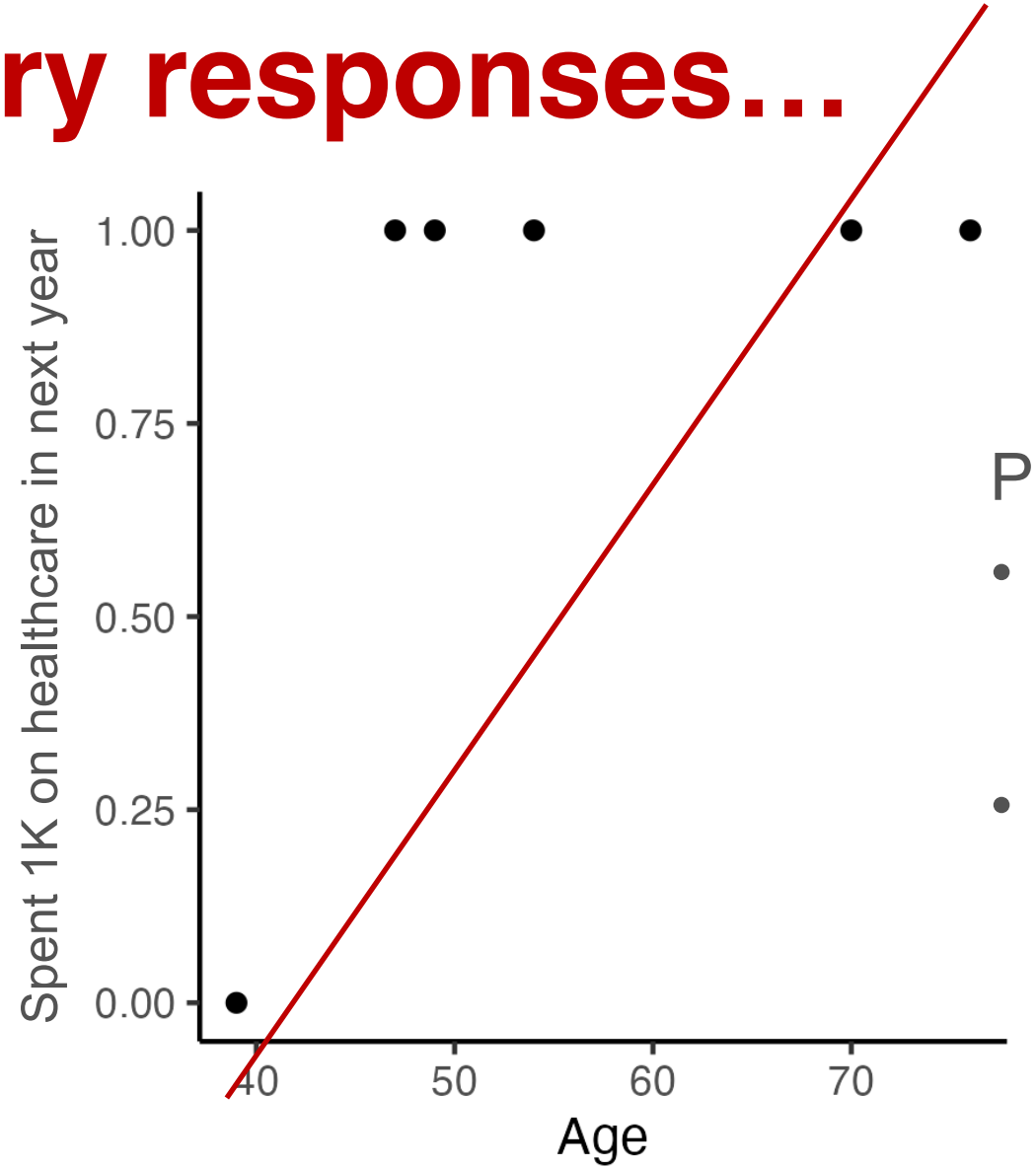
Age	Sex	Weight	Diabetes	Healthcare expenses
44	M	165	N	?
69	F	161	Y	?
78	M	170	N	?
66	M	191	N	?

For continuous responses…



$$\widehat{\text{hlthx}} = b_0 + b_1 \text{ age}$$

For binary responses…

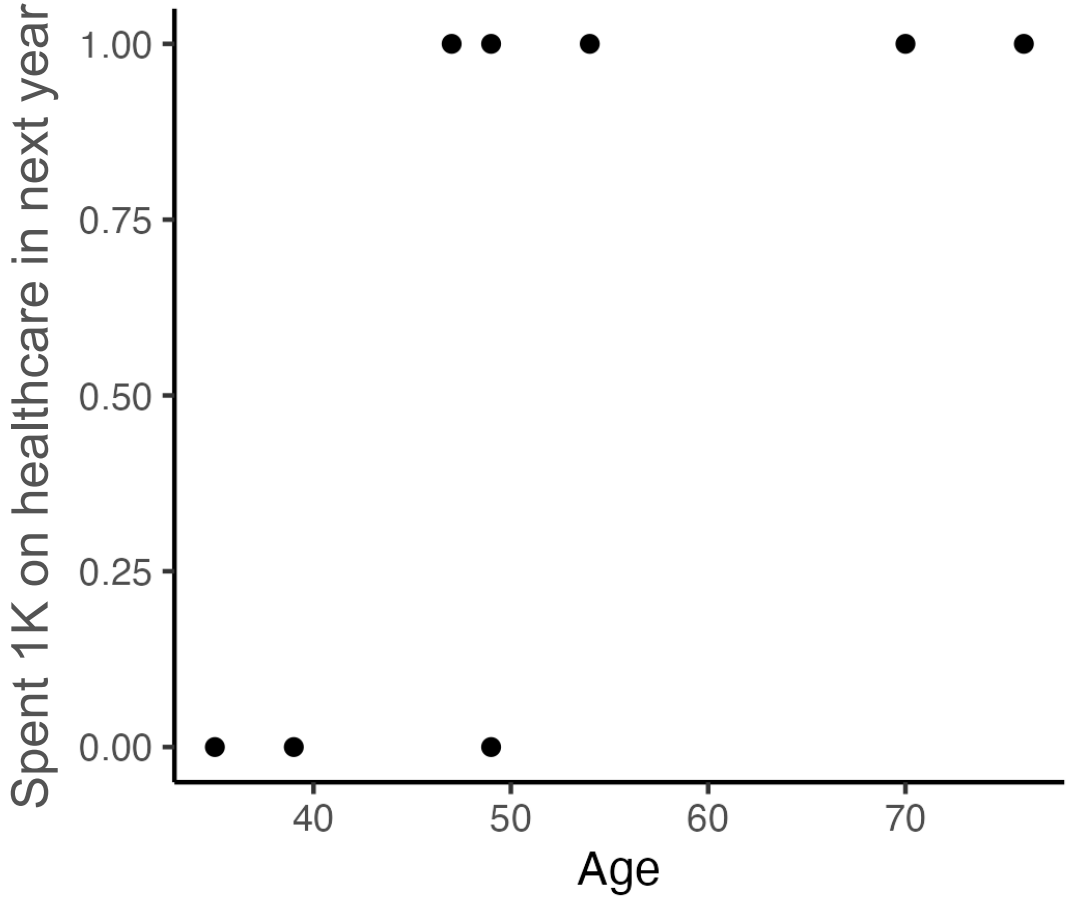


$$\widehat{\text{hlthx1k}} = b_0 + b_1 \text{ age}$$

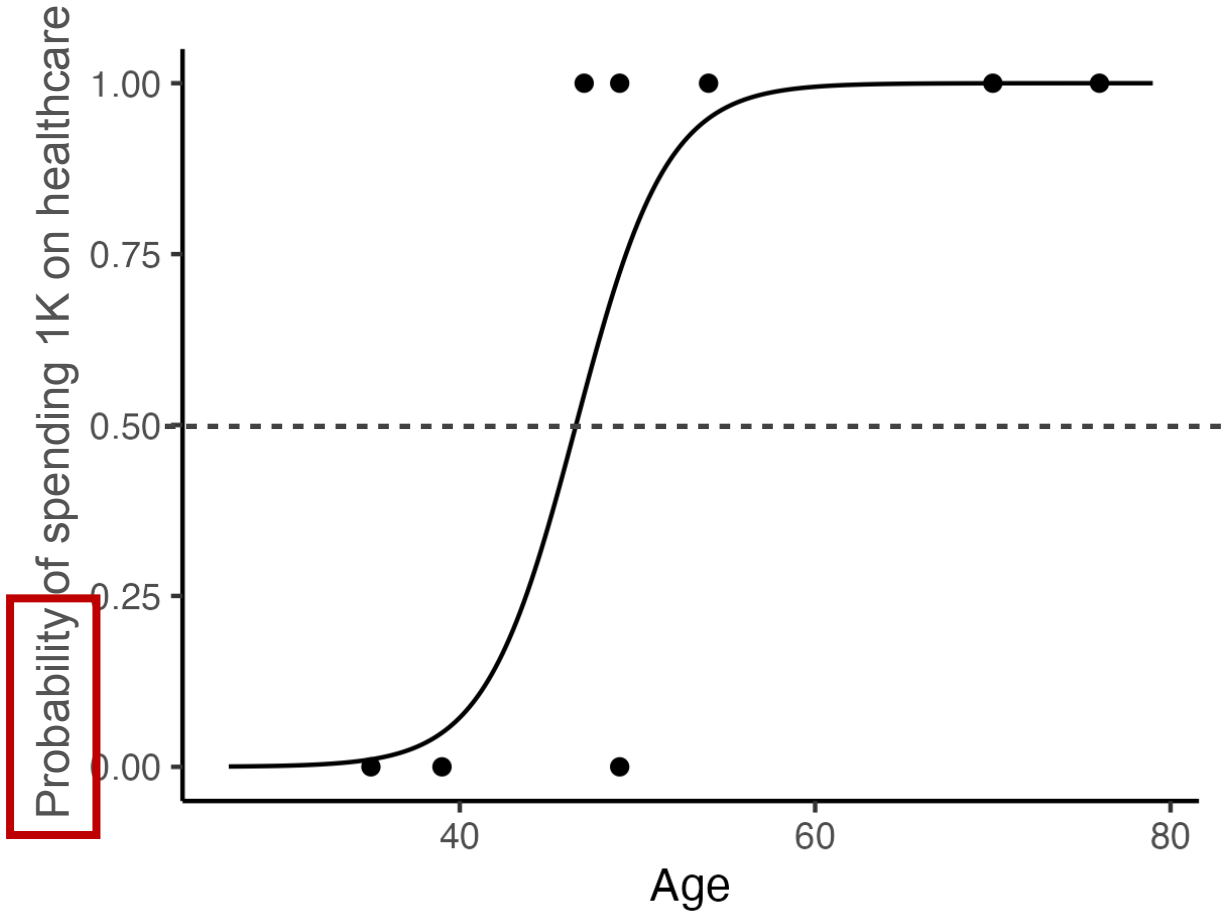
Problems:

- Predictions less than 0 and greater than 1 can be produced
- What do predicted responses between 0 and 1 mean?

For binary responses…



Logistic regression



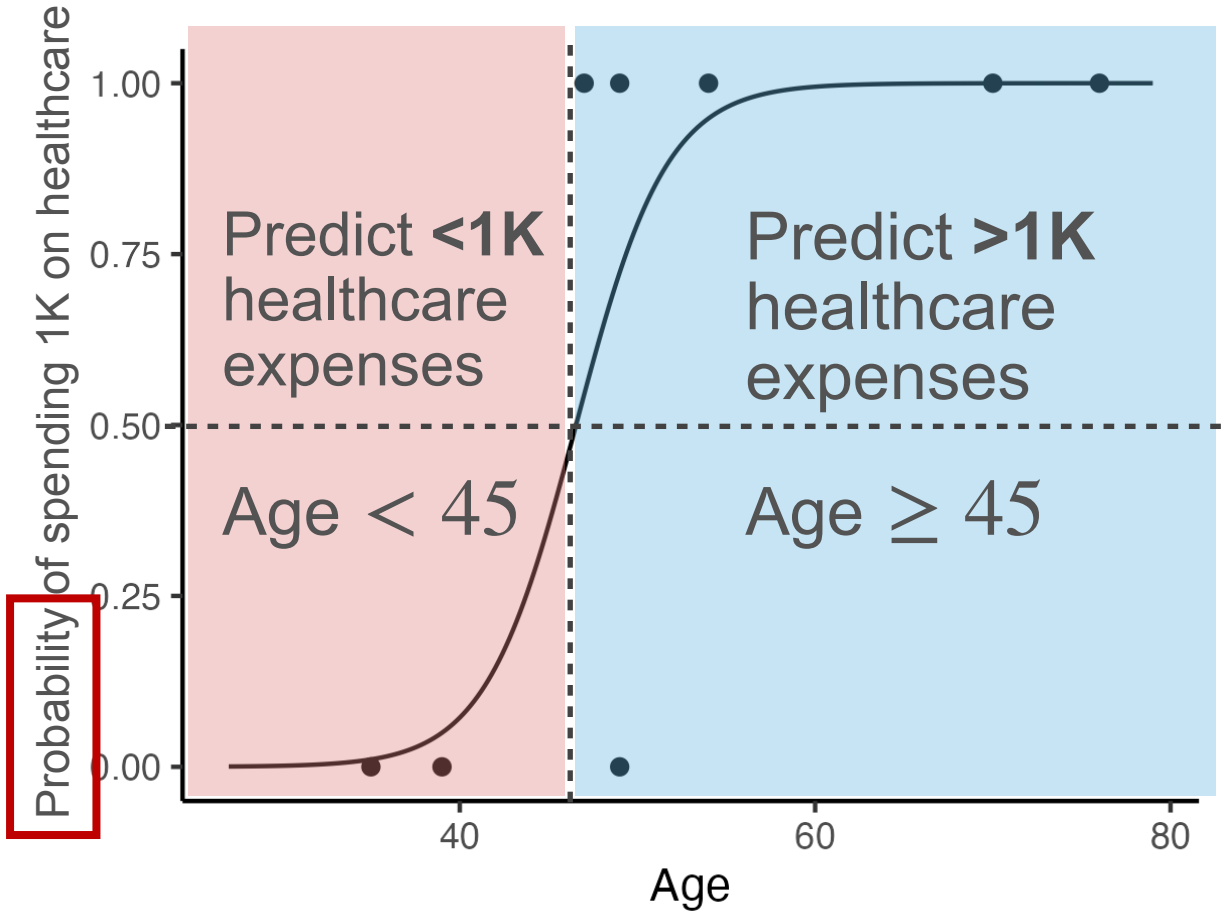
Y-axis

$$\widehat{P(\text{hlthx1k})} = \frac{e^{b_0+b_1 age}}{1+e^{b_0+b_1 age}}$$

This is the **logistic function**

- It is always between 0 and 1
- Values between 0 and 1 are interpreted as the *probability* of >1K healthcare expenditure

Logistic regression



Y-axis

$$\widehat{P(\text{hlthx1k})} = \frac{e^{b_0+b_1 age}}{1+e^{b_0+b_1 age}}$$

This is the **logistic function**

- It is always between 0 and 1
- Values between 0 and 1 are interpreted as the *probability* of >1K healthcare expenditure

Training an algorithm and generating a prediction

Labeled data

Age	Sex	Wt	Diab	Hlthx 1k
54	M	132	N	1
76	F	155	Y	1
49	M	166	Y	0
39	F	129	N	0
47	M	177	N	1
70	F	192	N	1

We **know** the **observed** healthcare expenses for the labeled data

Unlabeled data

Age	Sex	Wt	Diab
44	M	165	N
69	F	161	Y
78	M	170	N
66	M	191	N

We **do not know** the **observed** healthcare expenses for the unlabeled data

Prediction

Pred Hlthx 1k
0.95
1.0
0.8
0.7

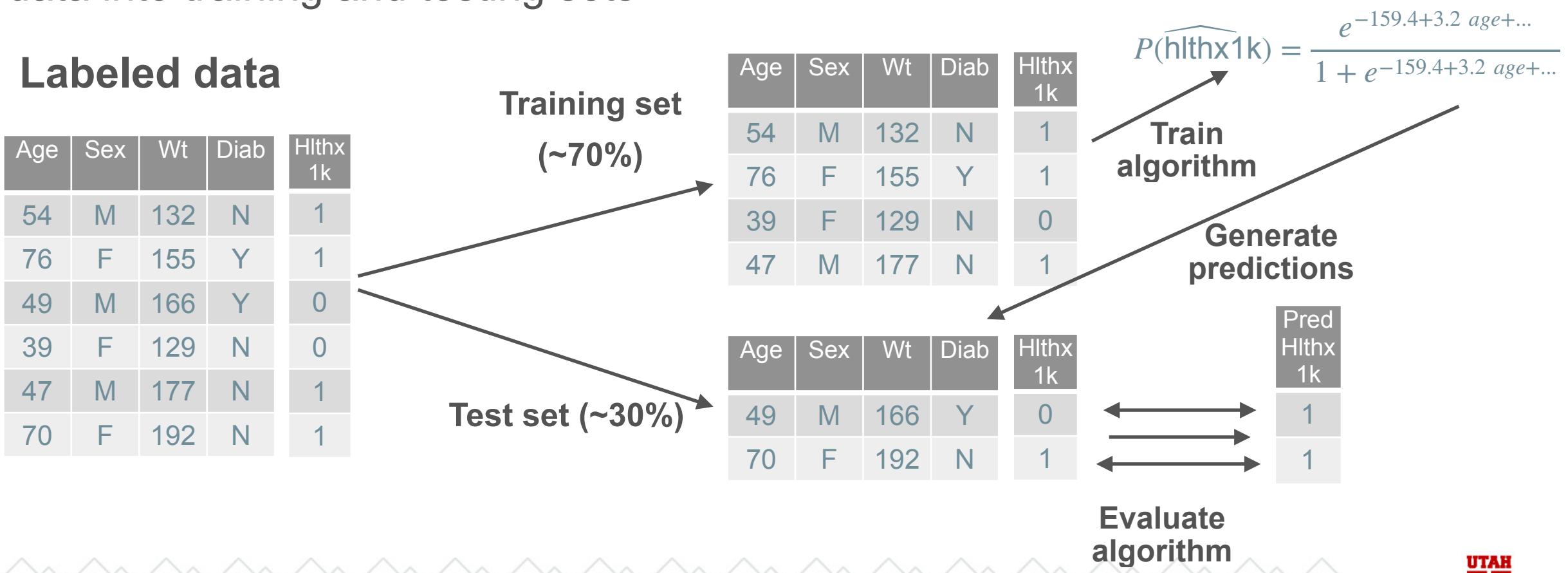
$$\widehat{P(\text{hlthx1k})} = \frac{e^{-159.4+3.2\ age+71.3\ sex-0.22\ weight-10.6\ diab}}{1+e^{-159.4+3.2\ age+71.3\ sex-0.22\ weight-10.6\ diab}}$$

Evaluating binary response predictions

Dr Rebecca Barter

Training and testing sets

Since the only labeled data is usually the data we have, we need to **split** our data into training and testing sets



Evaluating binary response predictions

When evaluating binary response predictions, we are comparing 0's and 1's.

(Observed) Hlthx1k		Predicted Hlthx1k
1	⟷	1
1	⟷	1
0	⟷	0
0	⟷	1
1	⟷	1
1	⟷	1

Accuracy

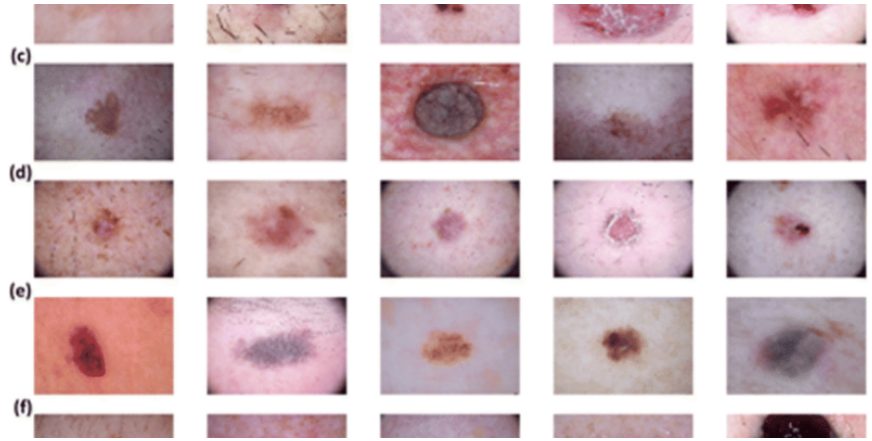
Proportion of responses where observed == predicted

Accuracy = 5/6 = 88.3%

Accuracy is not everything

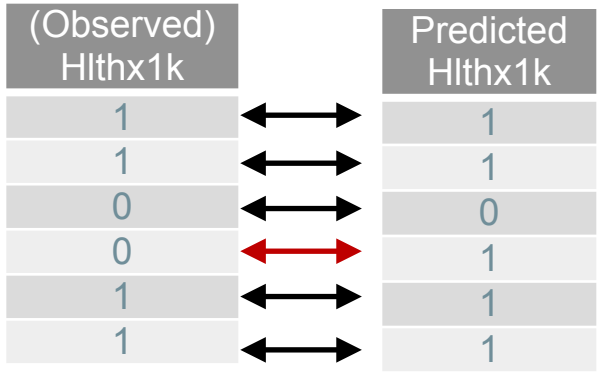
A skin cancer image dataset contains images for which:

5% are **melanoma** and **95%** are **non-melanoma**



What is the **accuracy** of an algorithm that *always predicts "**non-melanoma**"?*	**95%**
How accurate is the algorithm **for the non-melanoma patients**?	100%
How accurate is the algorithm **for the melanoma patients**?	0%

Confusion Matrix



	Observed Yes hlthx1k (1)	Observed No hlthx1k (0)
Predicted Yes Hlthx1k (1)	4	1
Predicted No Hlthx1k (0)	0	1

$$\text{Accuracy} = \frac{A + D}{A + B + C + D} = 5/6 = 0.83$$

	Observed Yes Hlthx1k (1)	Observed No Hlthx1k (0)
Predicted Yes Hlthx1k (1)	A	B
Predicted No Hlthx1k (0)	C	D

$$\text{Sensitivity/Recall (True Positive)} = \frac{A}{A + C} = 4/4 = 1$$

$$\text{Specificity (True Negative)} = \frac{D}{B + D} = 1/2 = 0.5$$

Confusion Matrix

$$\text{Accuracy} = \frac{A + D}{A + B + C + D} = \frac{23 + 45}{23 + 10 + 6 + 45} = 0.81$$

$$\text{Sensitivity (TP)} = \frac{A}{A + C} = \frac{23}{23 + 6} = 0.79$$

$$\text{Specificity (TN)} = \frac{D}{B + D} = \frac{45}{45 + 10} = 0.82$$

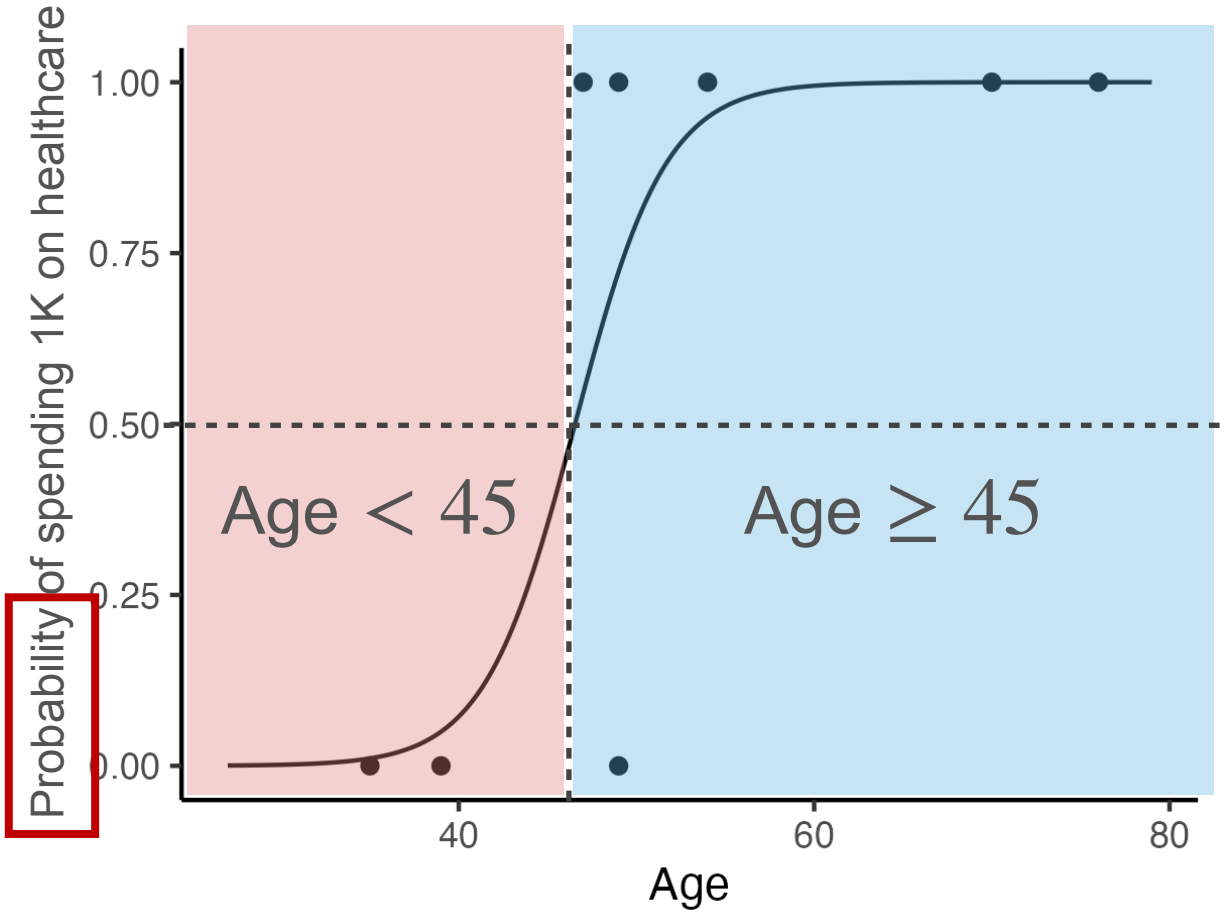
	Observed Yes hlthx1k (1)	Observed No hlthx1k (0)
Predicted Yes hlthx1k (1)	23	10
Predicted No hlthx1k (0)	6	45

	Observed Yes hlthx1k (1)	Observed No hlthx1k (0)
Predicted Yes hlthx1k (1)	A	B
Predicted No hlthx1k (0)	C	D

ROC curves for binary predictions

Dr Rebecca Barter

Converting probability predictions to binary predictions

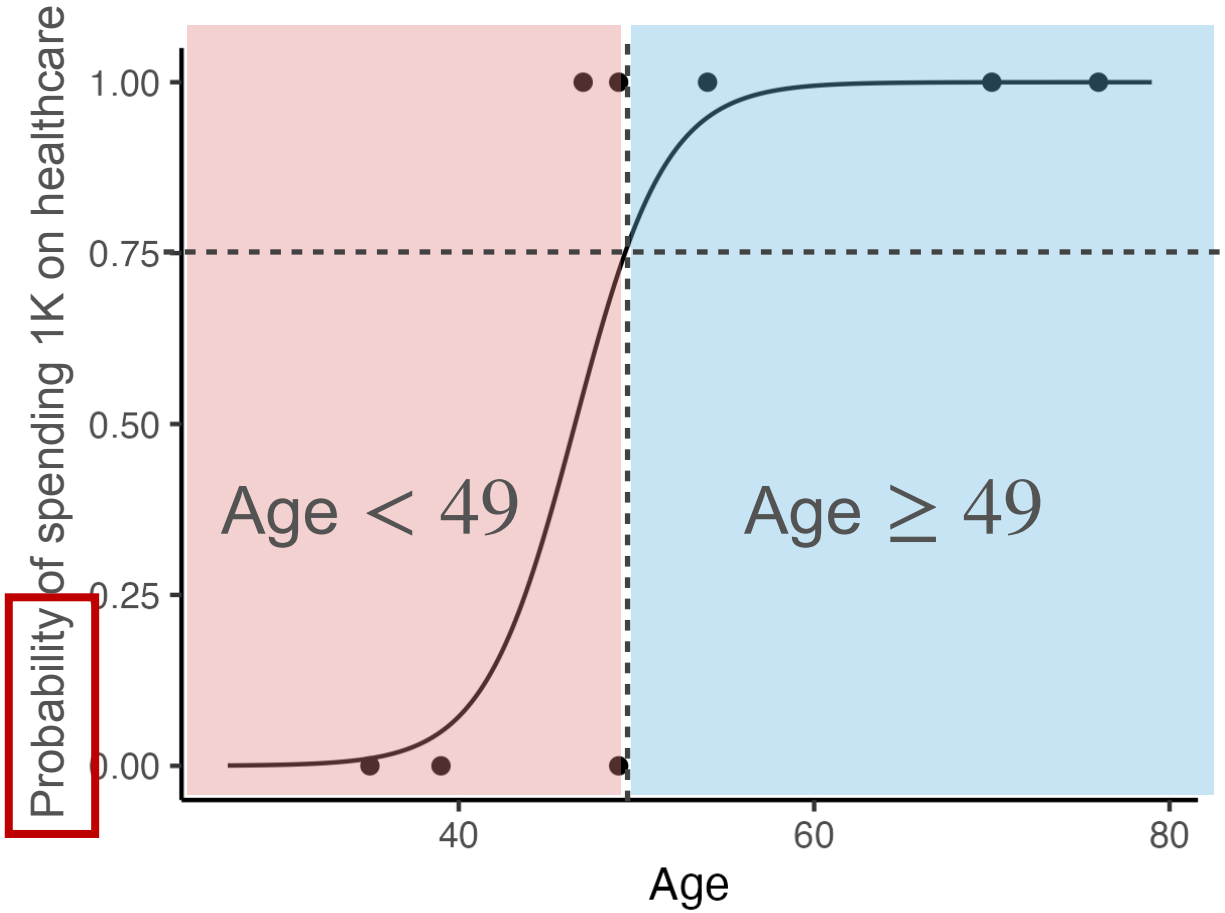


We can choose the probability threshold.

Predict “Spent 1K on healthcare” if:

Threshold: P(hlthx1k) > 0.5

Converting probability predictions to binary predictions

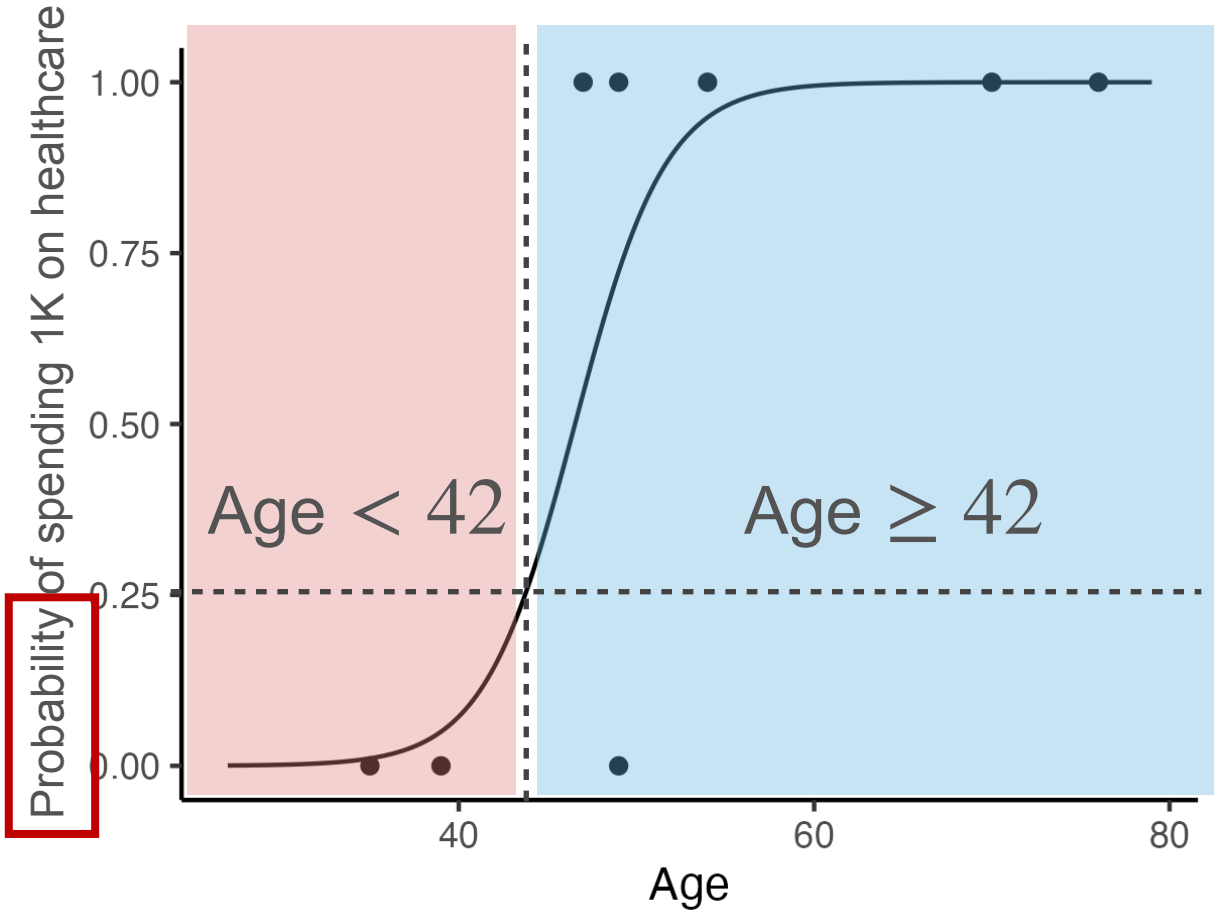


We can choose the probability threshold.

Predict “Spent 1K on healthcare” if:

Threshold: P(hlthx1k) > 0.75

Converting probability predictions to binary predictions



We can choose the probability threshold.

Predict “Spent 1K on healthcare” if:

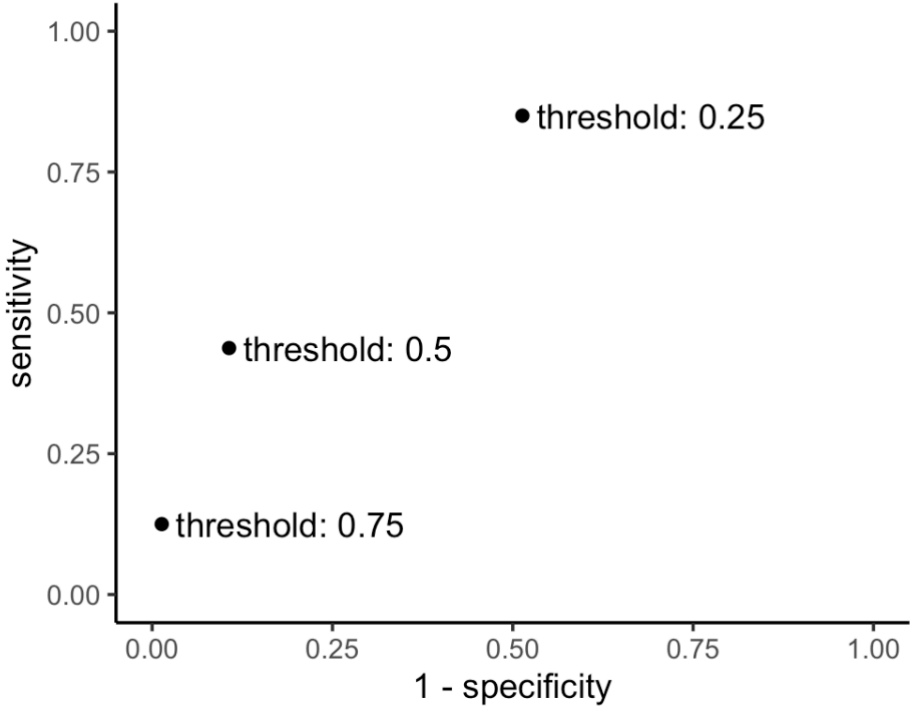
Threshold: P(hlthx1k) > 0.25

Each choice will have different accuracy/sensitivity/specificity

ROC curve

Predict “healthcare expenses over 1K” if predicted probability is greater than a threshold value

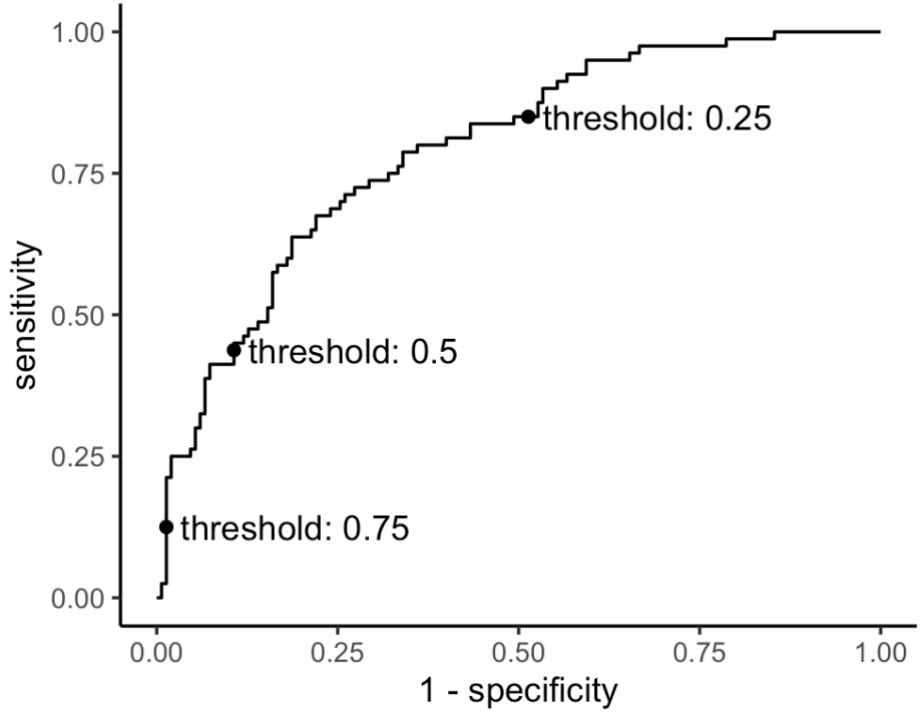
Threshold (predict “hlthx1k” if)	Accuracy	Sensitivity (TP)	Specificity (TN)
P(hlthx1k) > 0.25	0.62	**0.85**	0.49
P(hlthx1k) > 0.5	**0.73**	0.44	0.89
P(hlthx1k) > 0.75	0.69	0.13	**0.99**



ROC curve

Predict “healthcare expenses over 1K” if predicted probability is greater than a threshold value

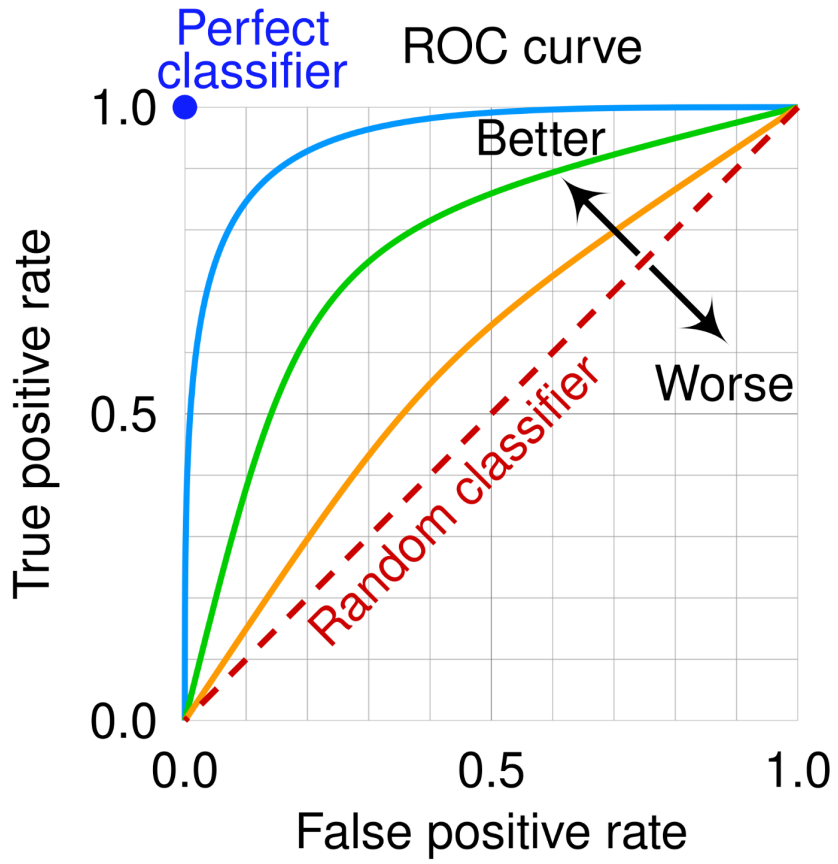
Threshold (predict “hlthx1k” if)	Accuracy	Sensitivity (TP)	Specificity (TN)
P(hlthx1k) > 0.25	0.62	**0.85**	0.49
P(hlthx1k) > 0.5	**0.73**	0.44	0.89
P(hlthx1k) > 0.75	0.69	0.13	**0.99**



Often the “best” threshold value is the **proportion of positive class obs**

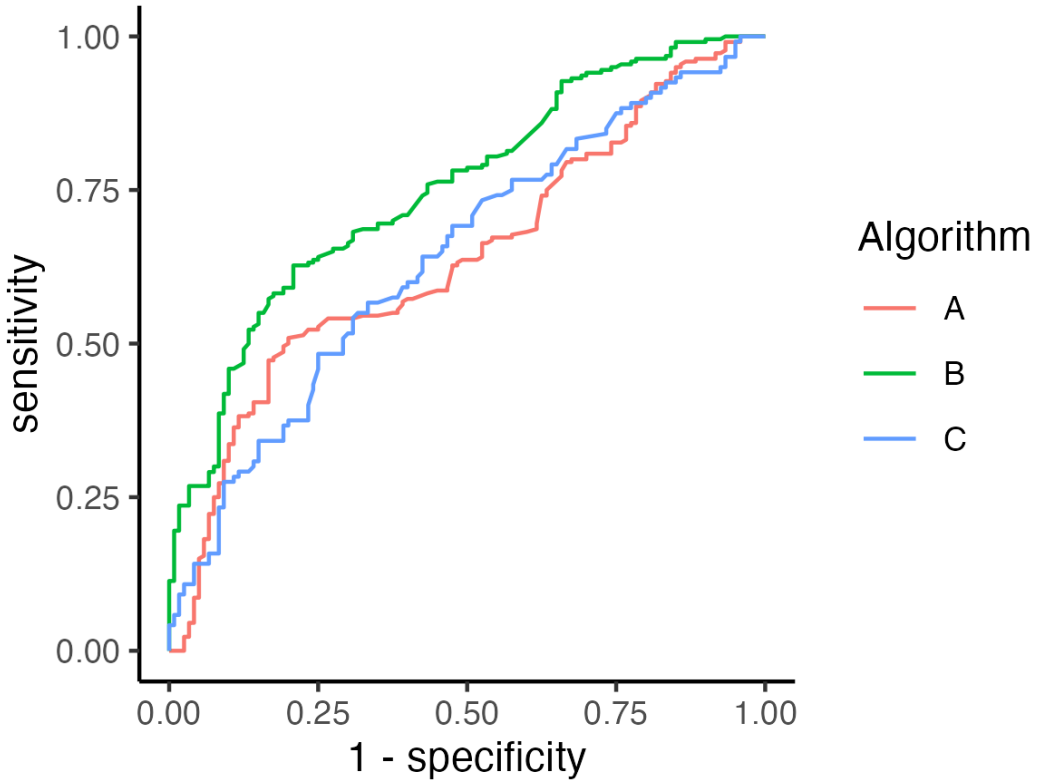
ROC curve

The “best” ROC curve is the one that gets the closest to the top-left corner



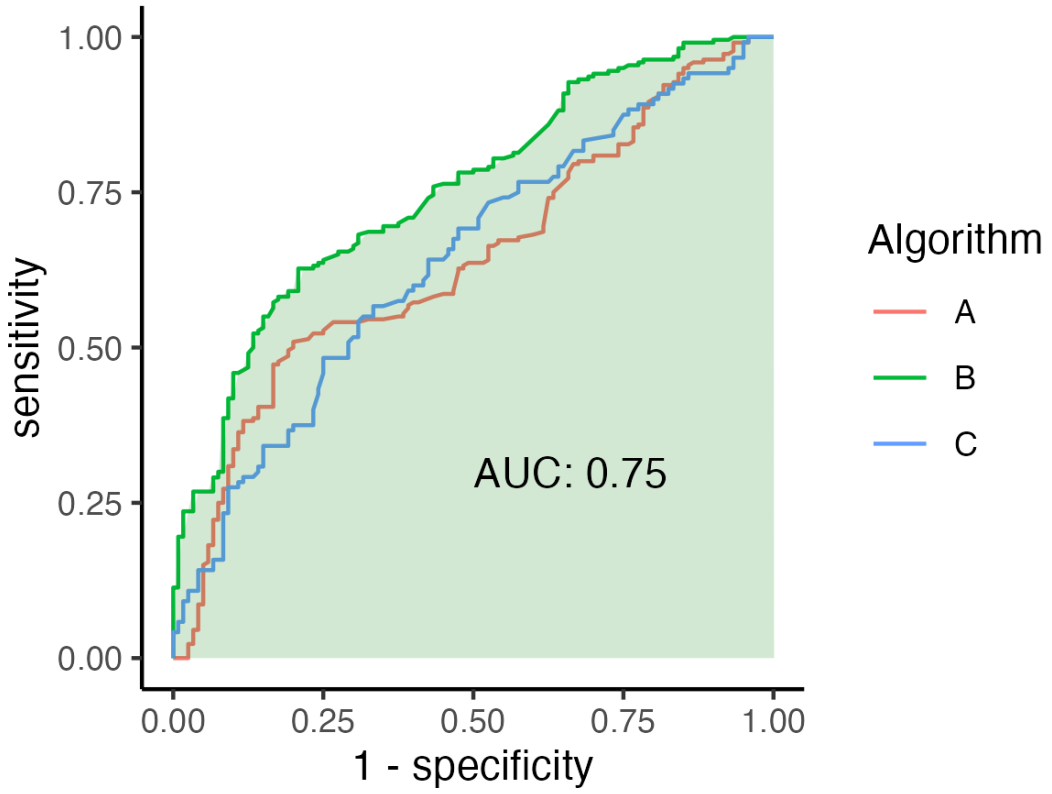
Area under the ROC curve (AUC)

The “best” model is the one that has the highest Area Under the Curve (AUC)



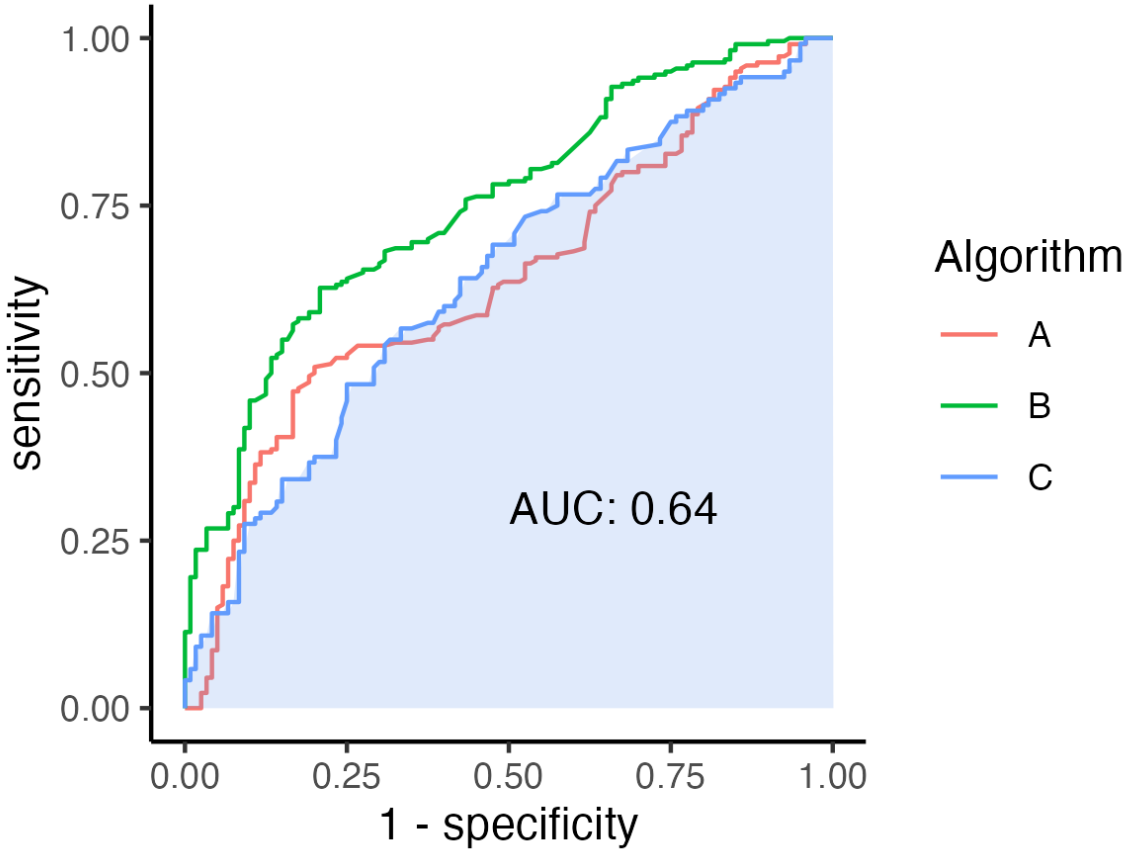
Area under the ROC curve (AUC)

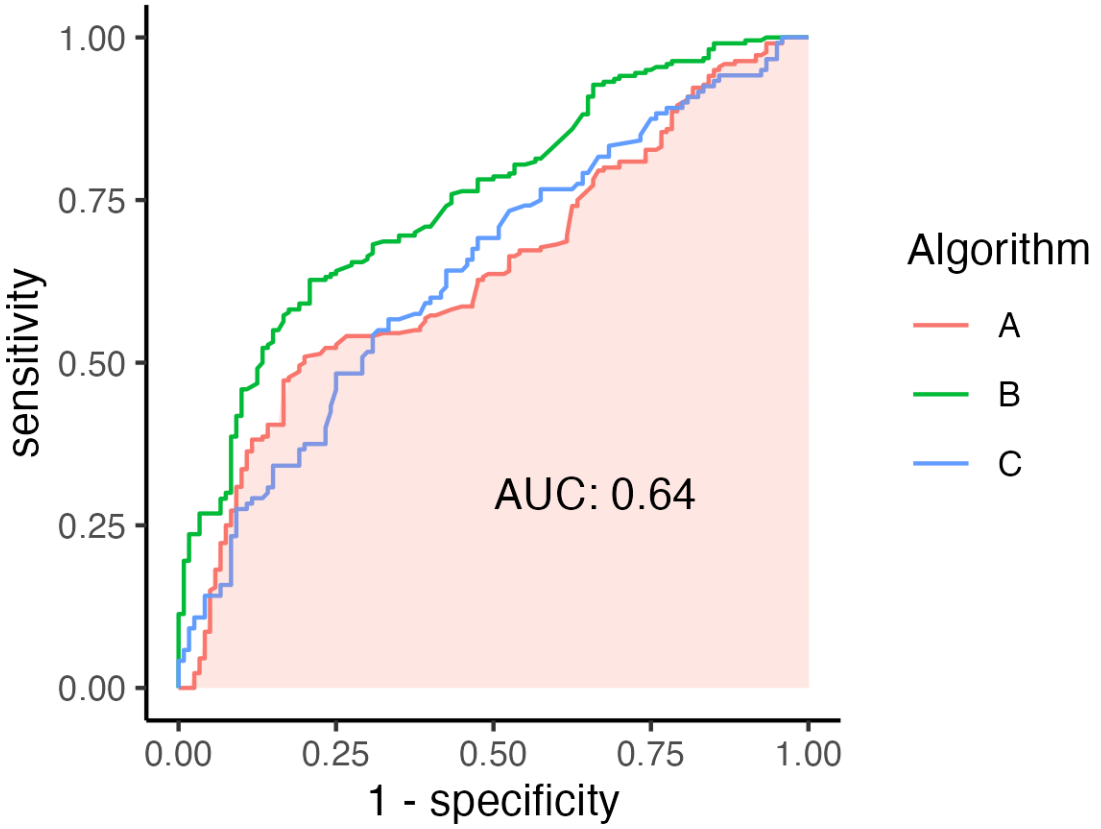
The “best” model is the one that has the highest Area Under the Curve (AUC)



Area under the ROC curve (AUC)

The “best” model is the one that has the highest Area Under the Curve (AUC)





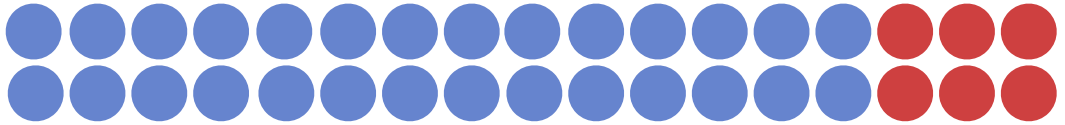
Class imbalance in binary prediction problems

Dr Rebecca Barter

Class imbalance

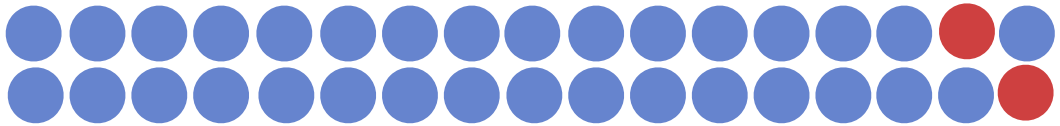
Class imbalance occurs when you have many more observations in one class than the other

Actual "observed" class



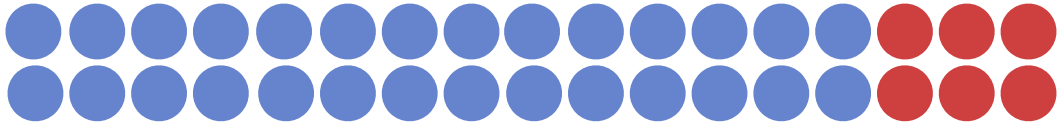
Problem: algorithms will tend to "over predict" the "majority" class

Predicted class



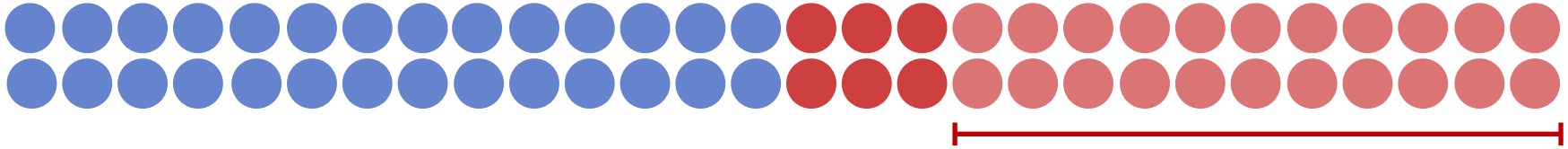
Sampling methods for class imbalance

Original imbalanced data



Upsampling

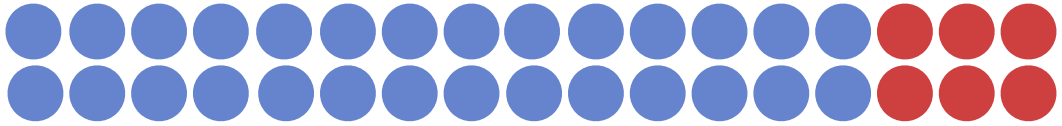
When training predictive algorithms, use a version of the training data where the **minority** class observations are randomly **repeated** so their number matches the majority class



Repeated data observations

Sampling methods for class imbalance

Original imbalanced data

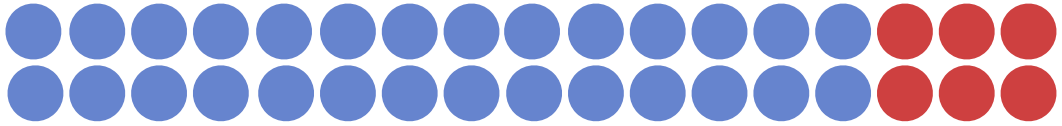


Downsampling

When training predictive algorithms, use a version of the training data where only a random **subsample** of the **majority** class is used so their number matches the minority class

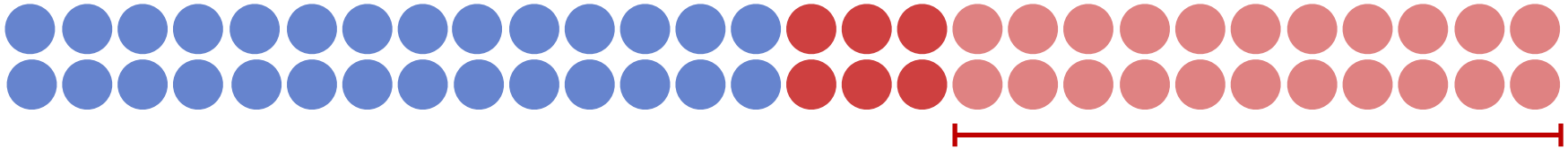
Sampling methods for class imbalance

Original imbalanced data



SMOTE (Synthetic Minority Over-sampling Technique)

When training predictive algorithms, use a version of the training data where ***new*** **"synthetic" minority** class data points are created based on "interpolating" existing minority class observations



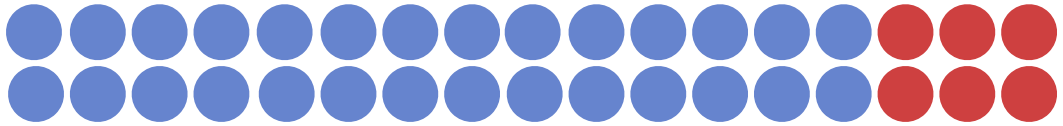
"Synthetic" (made-up) data observations

Choosing your threshold for class imbalance

When converting a binary probability prediction to a binary 0/1 value, we use a **threshold**.

The default threshold is predict positive class when **P > 0.5**

Improve predictive performance by setting the **threshold** to be the **proportion** of positive (minority) class observations in training data



Proportional threshold = $\frac{6}{34} = 0.176$

Predict positive class when **P > 0.176**