GENERAL ASSEMBLY

CHOOSING A SUPERVISED MODEL

Mason Gallo, Data Scientist

AGENDA

- Crash course on outliers
- Feature engineering soapbox
- Metrics for Evaluating our Models
- Choosing Algorithms

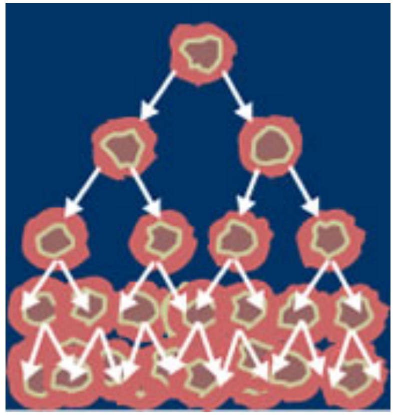
OBJECTIVES

- Understand outliers and how to address them
- Understand why feature engineering is important and difficult
- Understand the thought process for choosing the best model
- Choose the best model in Python

MOTIVATING EXAMPLES: PREDICTING CANCER AND FIRES

TWO PROBLEMS TO SOLVE

Can we predict fires based on weather data?

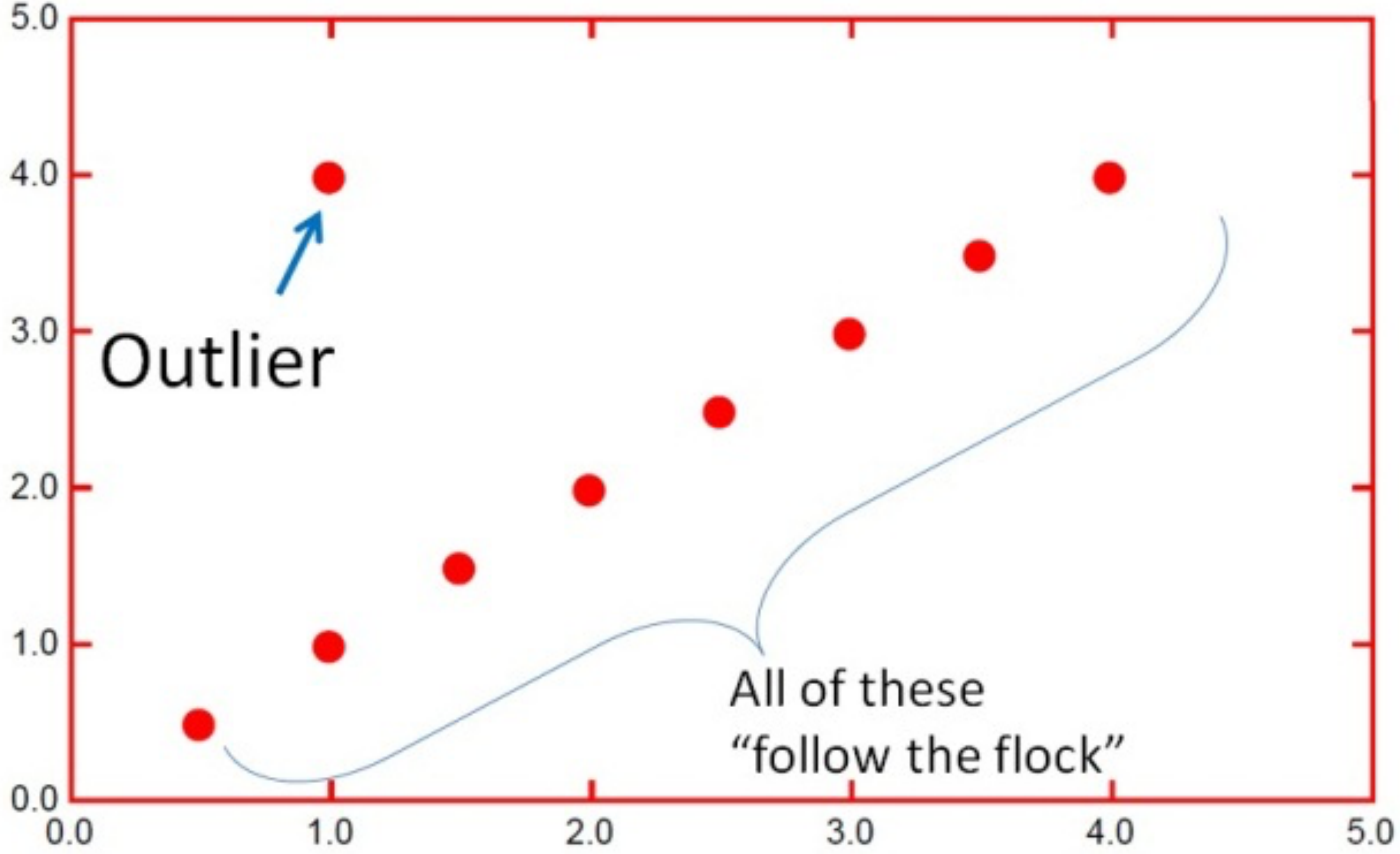


Can we predict breast cancer?

YOU WILL FIGURE OUT THE BEST MODELS FOR EACH OF THESE TASKS

HOW SHOULD WE DEAL WITH OUTLIERS?

OUTLIER VISUALLY



Never mind what the axes mean…

OUTLIER IN WORDS

"Outlier" is a scientific term to describe things or phenomena that lie outside normal experience. In the summer, in Paris, we expect most days to be somewhere between warm and very hot. But imagine if you had a day in the middle of August where the temperature fell below freezing. That day would be outlier. And while we have a very good understanding of why summer days in Paris are warm or hot, we know a good deal less about why a summer day in Paris might be freezing cold. In this book I'm interested in people who are outliers—in men and women who, for one reason or another, are so accomplished and so extraordinary and so outside of ordinary experience that they are as puzzling to the rest of us as a cold day in August."

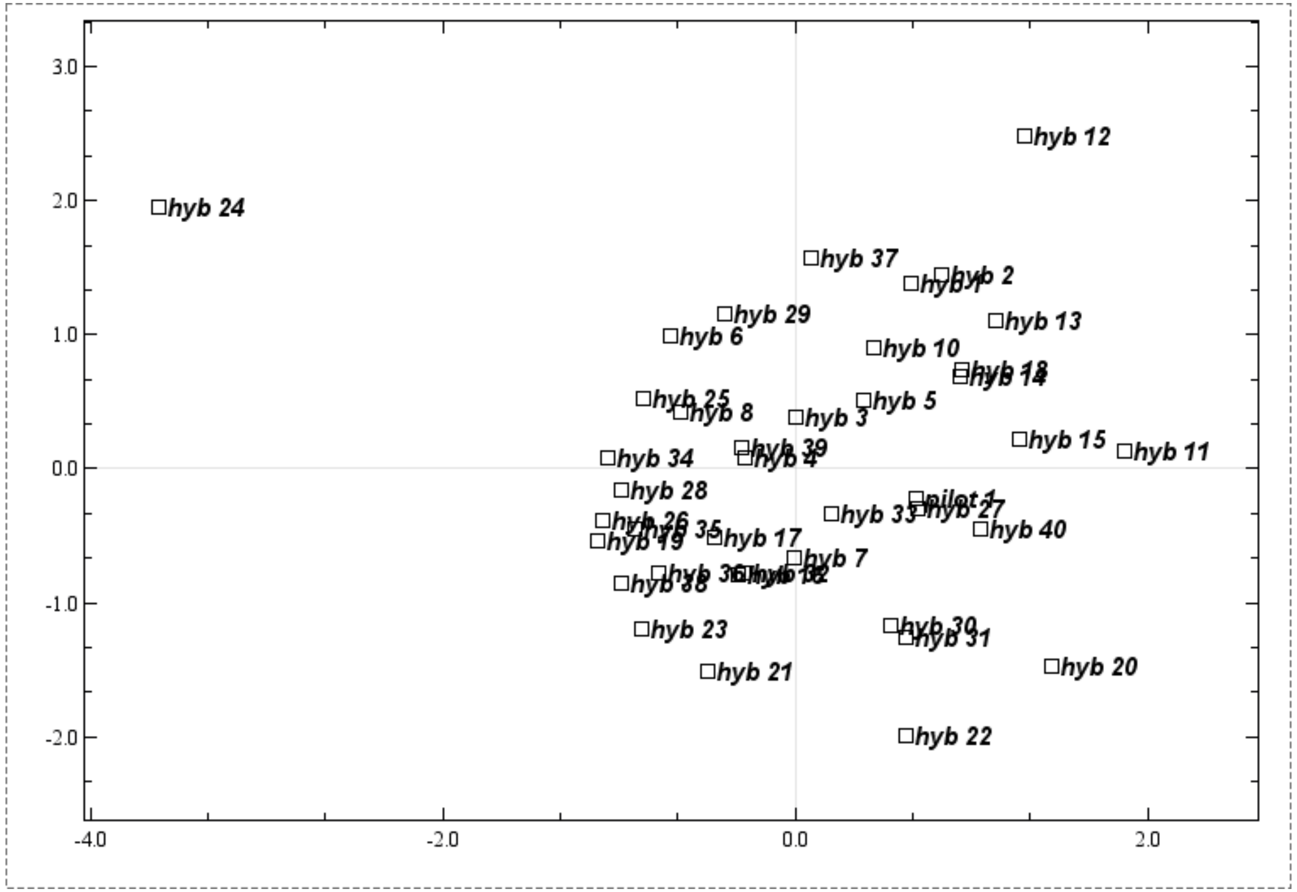
- Malcolm Gladwell from book “Outliers”

INTUITIVE DEFINITION

An unexpected or hugely different PREDICTED value (y)

INTUITIVE DEFINITION

An unexpected or hugely different PREDICTED value (y)



INTUITIVE DEFINITION

An unexpected or hugely different PREDICTED value (y)

Usually these unexpected y-values will have large errors or residuals

WE DON’T LIKE ERROR!!!!

WHAT CAUSES OUTLIERS?

- Sensor Error
- Data Entry Error
- Black Swans (ultra rare events)

WHAT CAUSES OUTLIERS?

- Sensor Error - SAFE TO IGNORE/DROP
- Data Entry Error - SAFE TO IGNORE/DROP
- Black Swans (ultra rare events) - DEPENDS ON YOUR OBJECTIVE!

WHAT CAUSES OUTLIERS?

- Sensor Error - SAFE TO IGNORE/DROP
- Data Entry Error - SAFE TO IGNORE/DROP
- Black Swans (ultra rare events) - DEPENDS ON YOUR OBJECTIVE!

WHAT IF YOUR GOAL IS TO FIND FRAUD? WHAT IF IT'S PREDICTING DAILY SALES?

EASY METHOD FOR DEALING WITH OUTLIERS IN REGRESSION

- Train your model
- Run a prediction on your labeled data
- Drop the rows that have the largest prediction error
- Retrain your model
- Repeat the above as necessary…

EASY METHOD FOR DEALING WITH OUTLIERS IN REGRESSION

- Train your model
- Run a prediction on your labeled data
- Drop the rows that have the largest prediction error
- Retrain your model
- Repeat the above as necessary…

YOU SHOULD BE IN CONSTANT DISCUSSION WITH STAKEHOLDERS WHEN DROPPING

WHAT ABOUT OUTLIERS WITH CLASSIFICATION?

NO ESTABLISHED METHOD BECAUSE IT DEPENDS ON YOUR OBJECTIVE

- Start with looking at counts of your y-categories
- Examine rare y-categories
- For logistic regression, look at probabilities that are very far away from actual
- Look at counts of your categorical features

GENERAL OPTIONS FOR DEALING WITH OUTLIERS

- Use a model that's robust to outliers, like tree-based models or SVMs
- Use Mean Absolute Error instead of MSE if you want to reduce their effect
- Set artificial boundaries for your data based on domain knowledge
- Remove the outliers as we mentioned previously
- Transform your data (coming up next)

FEATURE ENGINEERING

QUOTES ON FEATURE ENGINEERING

"The most important and most difficult part of building models"
- Me

"Art and science"
- Me

"Manually creating what the features should be"
- Me

SUPPOSE YOU HAVE A FEATURE CALLED URL

“http://mashable.com/some_article/2013/01/02”

Extracting 2013 from the above is feature engineering!

FEATURE ENGINEERING IS REALLY HARD

THIS IS MOST PEOPLE’S SECRET SAUCE

- I couldn’t find any definitive literature
- Most people won’t share their strategies
- Much like stock trading, be weary of people sharing their strategies
- HIGHLY dependent on your data and objectives
- No “best practices”

FEATURE ENGINEERING IS REALLY HARD

MANY DEPENDENCIES:

- Your performance metric (RMSE? F1?)
- Your model (Lasso? Logistic?)
- The raw data (properly cleaned?)

FEATURE ENGINEERING STEPS

BIG PICTURE

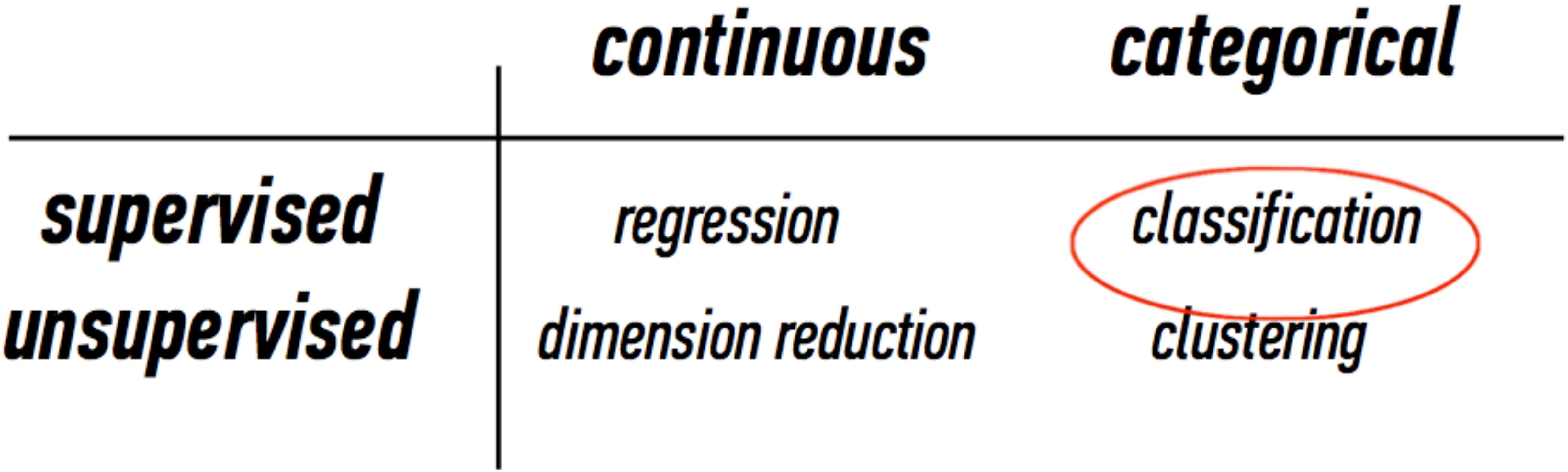
- Remove unnecessary features
- Remove highly correlated / redundant features
- Create new features
- Modify feature data types if necessary (numeric -> binary)
- Modify feature values if necessary (artificial floors or ceilings)

HOW DO WE MEASURE CLASSIFICATION SUCCESS?

MODEL TYPES

	continuous	*categorical*
supervised	???	???
unsupervised	???	???

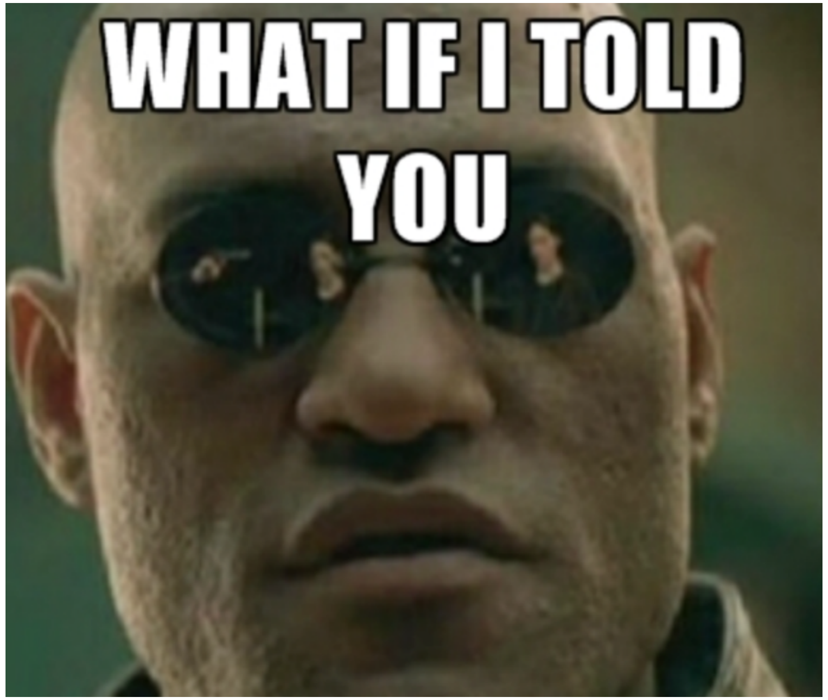
MODEL TYPES



	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

UP UNTIL NOW…

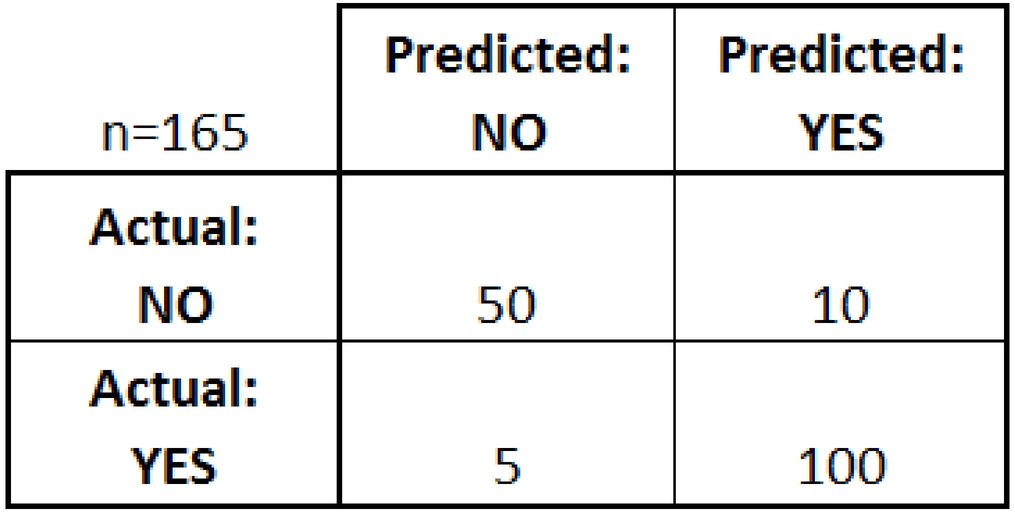
We have only focused on accuracy for classification



What if I told you there are more ways to measure classification?

CONFUSION MATRIX

Confusion Matrix: table to describe the performance of a classifier



n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

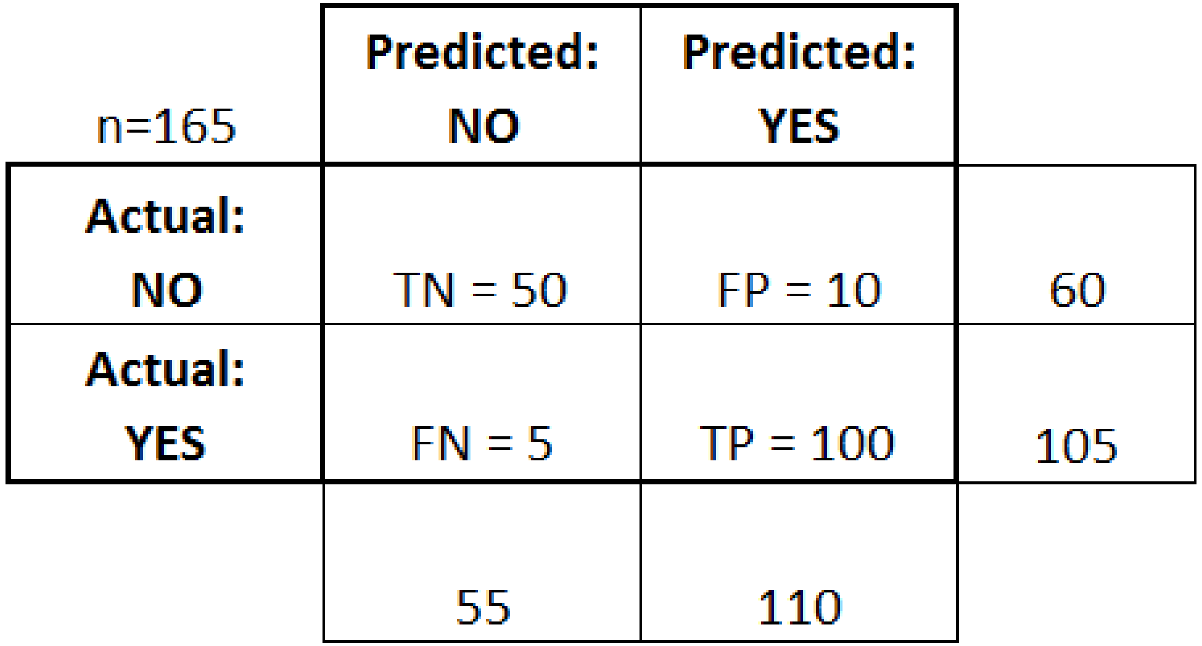
Example: Test for presence of disease

NO = negative test = False = 0

YES = positive test = True = 1

- *How many classes are there?*
- *How many patients?*
- *How many times is disease predicted?*
- *How many patients actually have the disease?*

CONFUSION MATRIX



n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Basic Terminology:

- *True Positives (TP)*
- *True Negatives (TN)*
- *False Positives (FP)*
- *False Negatives (FN)*

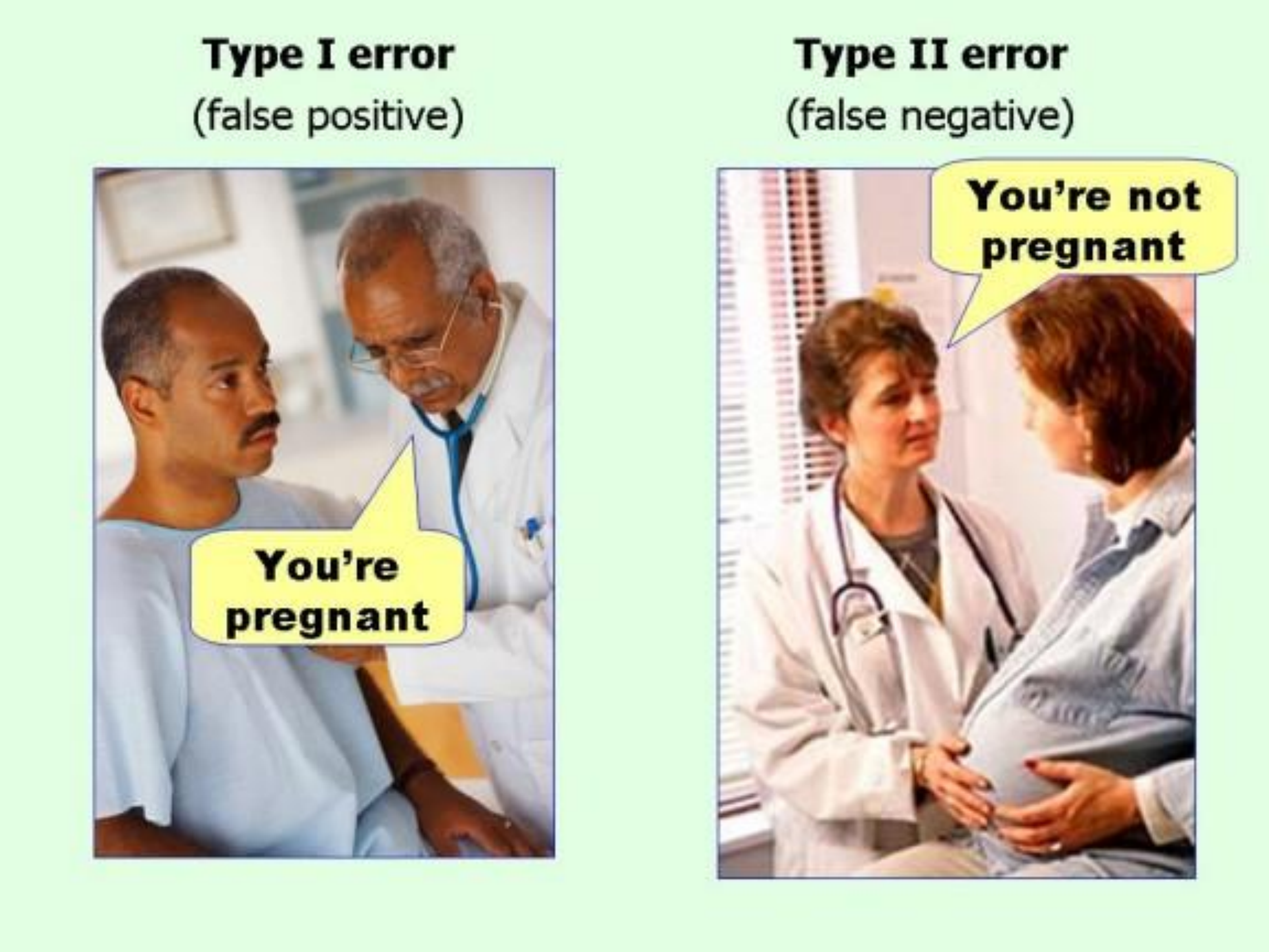
Accuracy:

- *Overall, how often is it **correct**?*
- *(TP + TN) / total = 150/165 = 0.91*

Misclassification Rate (Error Rate):

- *Overall, how often is it **wrong**?*
- *(FP + FN) / total = 15/165 = 0.09*

CONFUSION MATRIX



MORE TRADEOFFS

We need to balance:

Correctly detecting if someone is pregnant

vs.

Finding all pregnant people

MORE TRADEOFFS

We need to balance:

Correctly identifying if someone is pregnant

vs.

Finding all pregnant people

Thought exercise:

What if we just assume everyone is pregnant?
What if we only say a woman is pregnant when she’s going into labor?

FORMAL DEFINITIONS

PRECISION: when we say someone is pregnant, they really are

FORMAL DEFINITIONS

PRECISION: when we say someone is pregnant, they really are

RECALL: we find all the people that are pregnant

FORMAL DEFINITIONS

PRECISION: when we say someone is pregnant, they really are
“Innocent until proven guilty”

PRECISION = TP / (TP + FP)

RECALL: we find all the people that are pregnant
“We don’t want to let any criminals free, so we risk condemning an innocent
by calling everyone guilty”

RECALL = TP / (TP + FN)

SO WHAT DOES THIS MEAN FOR ME?

YOU need to balance based on your objectives

Ex:
If you’re trying to detect cancer, you don’t want to risk not detecting it, so you prioritize RECALL

If you’re trying to block spam, you don’t want to call something spam when it isn’t, so you prioritize PRECISION

BUT MASON, I WANT TO PRIORITIZE BOTH RECALL AND PRECISION

If you really want to maximize both, statisticians created:

F1 SCORE

(it’s just the mean of precision and recall)

BUT MASON, I WANT TO PRIORITIZE BOTH RECALL AND PRECISION

If you really want to maximize both, statisticians created:

F1 SCORE

(it’s just the mean of precision and recall)

Of course, if you don’t care either way, just use ACCURACY

SO WHY DO WE NEED ALL THIS FANCY STUFF?

In other words: why can’t I just use accuracy?

If we’re predicting breast cancer, we could just assume everyone doesn’t have cancer since breast cancer is rare

We would have the highest accuracy! Obviously this would be a bad idea…

This is called the accuracy paradox!

HOW DO WE MEASURE REGRESSION SUCCESS?

REGRESSION PROBLEMS

	continuous	*categorical*
supervised	???	???
unsupervised	???	???

REGRESSION PROBLEMS

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

HOW DO WE MEASURE HOW WELL OUR REGRESSION PERFORMS?

Any ideas?

HOW DO WE MEASURE HOW WELL OUR REGRESSION PERFORMS?

Root Mean Squared Error

Mean Squared Error

Mean Absolute Error

WHAT’S THE INTUITION

Psst…you might know error as residual, which is the same thing

HOW DO WE MEASURE HOW WELL OUR REGRESSION PERFORMS?

Root Mean Squared Error

Mean Squared Error

Mean Absolute Error

When do we use each?

THOUGHT EXAMPLE

You work as the Data Scientist for a company that manufactures screws.

The screws must be 8mm +/- 0.1mm in order to fit.

This means that any error greater than 0.1mm renders the screw useless

Which type of error should we use?

THOUGHT EXAMPLE

You work as the Data Scientist for a company that manufactures screws.

The screws must be 8mm +/- 0.1mm in order to fit.

This means that any error greater than 0.1mm renders the screw useless

Which type of error should we use?

Mean Squared Error or Root Mean Squared Error

THOUGHT EXAMPLE

Use MSE or RMSE when we care about outliers

Rule: if an error of 2x is > 2x worse than an error of 1x, use MSE or RMSE

Otherwise, use MAE

WHAT’S THE INTUITION

MSE OR RMSE penalize large errors MORE THAN small errors

MAE treats all errors the same

WHICH TO USE - MSE OR MAE?

It depends on your objective!

RECAPPING MODEL SUCCESS

CLASSIFICATION METRICS

- PRECISION if we want to be sure that it really is spam when we say it is
- RECALL if we want to make sure we catch all the spam
- F1 SCORE if we want a balance of PRECISION and RECALL
- ACCURACY if misclassifying spam is the same as ham (usually it isn’t)

CLASSIFICATION METRICS

- What if class labels are wildly unbalanced?
- Ex: only 1% of emails are actually spam

CLASSIFICATION METRICS

- What if class labels are wildly unbalanced?
- Ex: only 1% of emails are actually spam

- If we just predict that all emails are ham, we'll have accuracy of 99%!

- This is where PRECISION, RECALL and F1 SCORE shine

REGRESSION METRICS

- MSE when we care about outliers or want to penalize large errors
- Note: we take the square root of MSE to get in our target units
- Ex: errors above 0.2 mean our product isn't profitable

REGRESSION METRICS

- MSE when we care about outliers or want to penalize large errors
- Note: we take the square root of MSE to get in our target units
- Ex: errors above 0.2 mean our product isn't profitable

- MAE when we don't care about penalizing large errors
- Ex: predicting rainfall errors of +/- 2cm is exactly twice as bad as +/- 1cm

CHOOSING A SUPERVISED MODEL CHEATSHEET

WHEN TO USE NEAREST NEIGHBORS

- You’re predicting a class or a number
- You have no idea about the actual distribution of your data
- You don’t care about individual contributions of features
- You don’t have a large amount of features
- You don’t have many correlated features
- Note: shown to work well with imputing missing data

WHEN TO USE NAIVE BAYES

- You're predicting a class or a number
- Problem domain is text-based i.e. predicting sentiment or spam
- You are comfortable with assuming all features are independent
- You have small amount of training data (otherwise consider SVM)
- You don't care about individual contributions of features
- You don't have correlated features

WHEN TO USE LASSO

- You’re predicting a number
- You believe that only a few of the features should be relevant
- You’re comfortable trading off some predictive accuracy for interpretability
- You need to know how individual features contribute
- You don’t have correlated features

WHEN TO USE RIDGE

- You’re predicting a number
- You believe that all of the features should be relevant
- You’re comfortable trading off interpretability for some predictive accuracy
- You need to know how individual features contribute
- You don’t have correlated features

WHEN TO USE LOGISTIC

- You’re predicting a class
- You believe that all or some of the features should be relevant
- You need to know how individual features contribute
- You need probabilities rather than just raw class predictions

WHEN TO USE SVM

- You’re predicting a class or a number
- You need to fit a highly complex relationship that may not be linear
- You don’t have a huge dataset
- Your features are correlated
- You’re comfortable with a “black box”

WHEN TO USE RANDOM FOREST

- You’re predicting a class or a number
- You need to fit a highly complex relationship that may not be linear
- You need feature importance
- Your features are correlated
- You care about feature interactions
- Your dataset has outliers
- You don’t need individual feature contributions
- Note: shown to work well with imputing missing data

CLOSING WORDS

- There is no free lunch!
- The preceding slides are just rules of thumb
- You should always test and defend

CHOOSING A SUPERVISED MODEL

LET’S CODE!