Supervised Learning

Data Boot Camp

Lesson 20.1

Class Objectives

By the end of this lesson, you will be able to:

- Model and fit several supervised learning linear models (linear regression, logistic regression) using scikit-learn.
- Conceptualize and build training and test windows for supervised learning analysis.
- Define classification in the context of machine learning.
- Evaluate classification algorithms using a confusion matrix and classification report.


WELCOME

Introduction to Supervised Learning & Classification

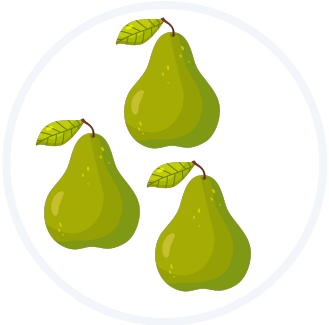
Introduction to Supervised Learning

In supervised learning, we take a set of known answers called **labels** and fit a model with a set of **features** (inputs) that corresponds to the labels.

These models are called **supervised learners**.

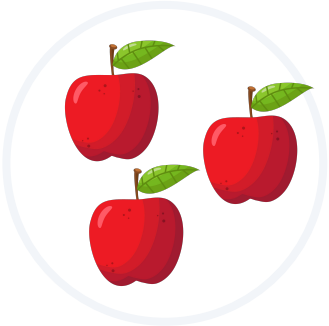
Fruit

Supervised Learning

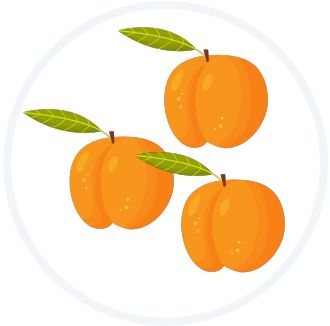


Pears

Bananas

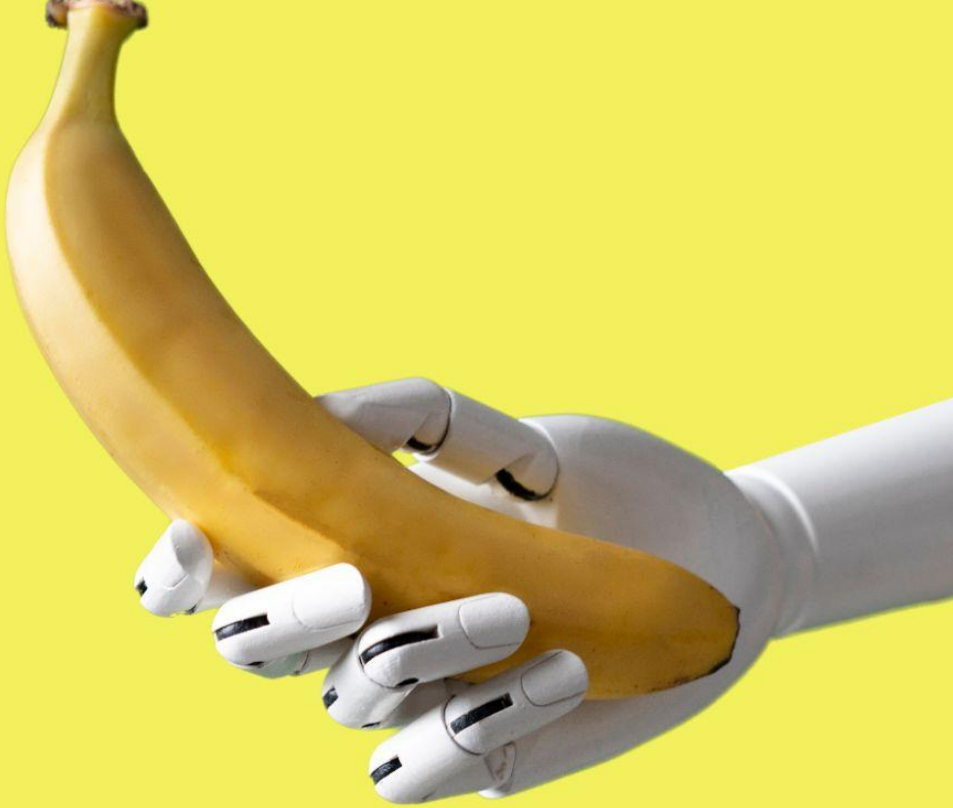


Apples



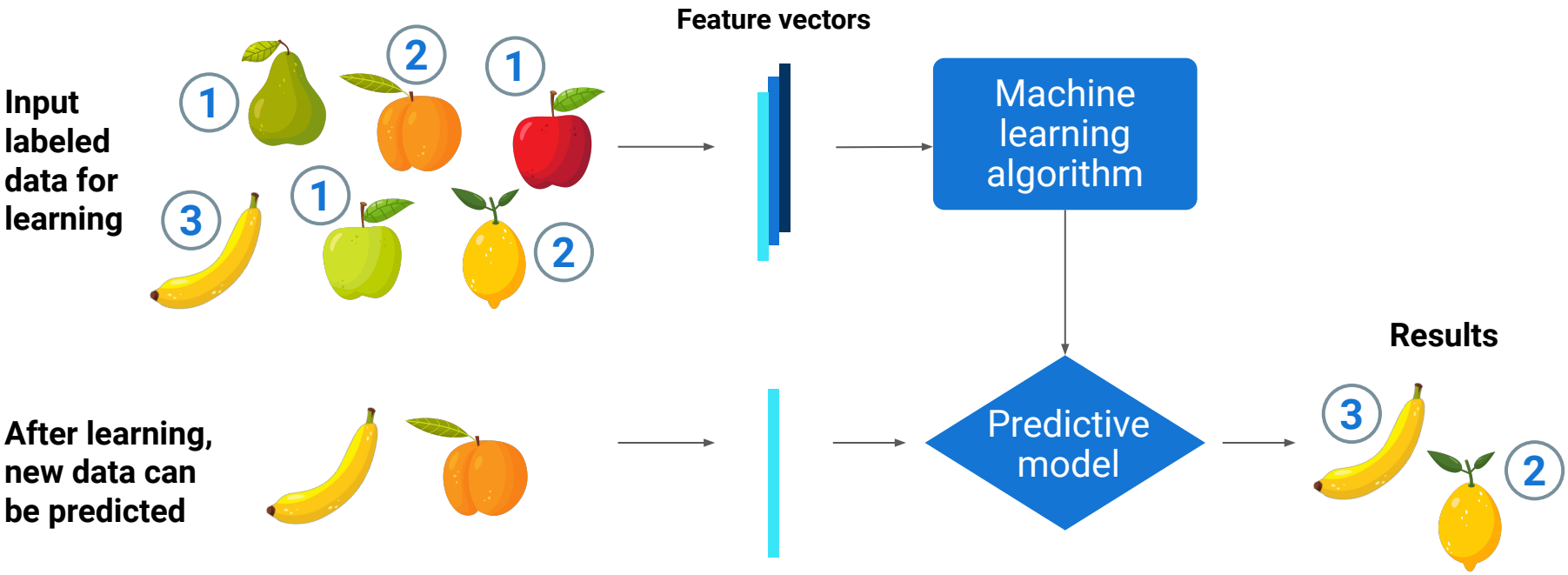
Oranges

Supervised learning requires us to feed the correct answers to the model.



Introduction to Supervised Learning

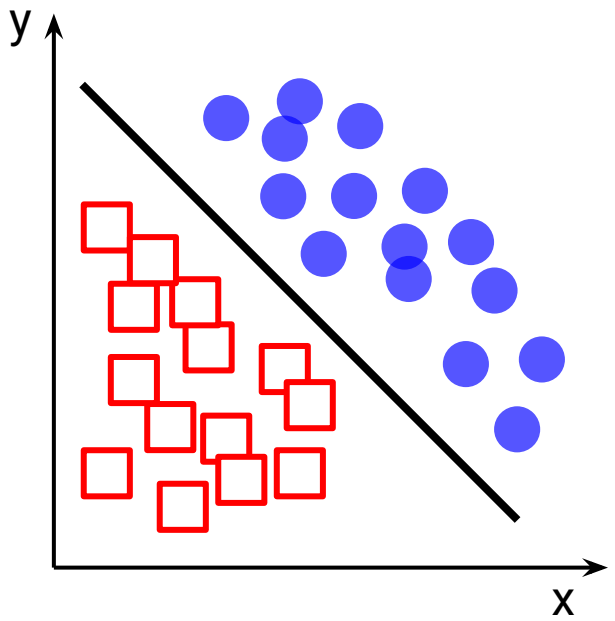
The model learns from the data and the answers. It becomes better at predicting the correct answer as we provide more data.



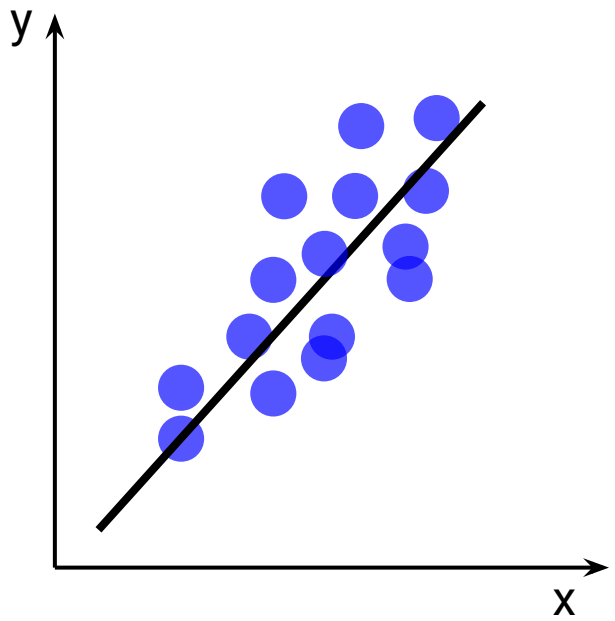
Introduction to Supervised Learning

Supervised learners generally fall into one of two broad categories:

Classification



Regression



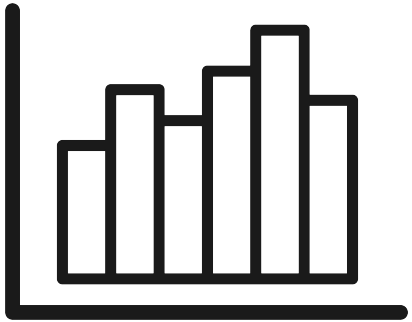
What is regression?

Regression

Regression is a method for predicting **continuous** valued variables.

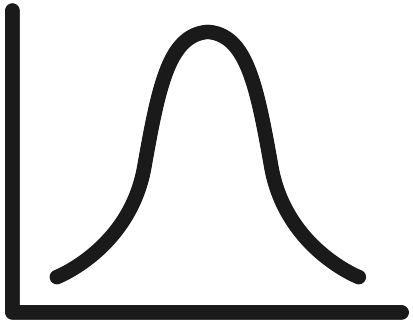
Continuous values

Those that can always be divided into smaller pieces.



Continuous variables

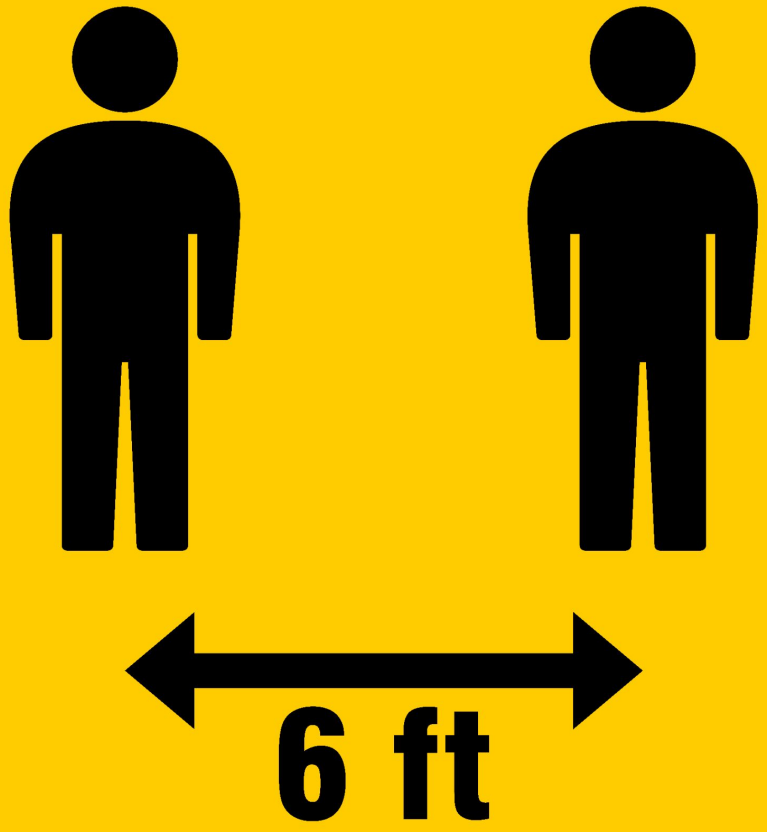
No matter how small, these will have a middle that we can find.



Regression

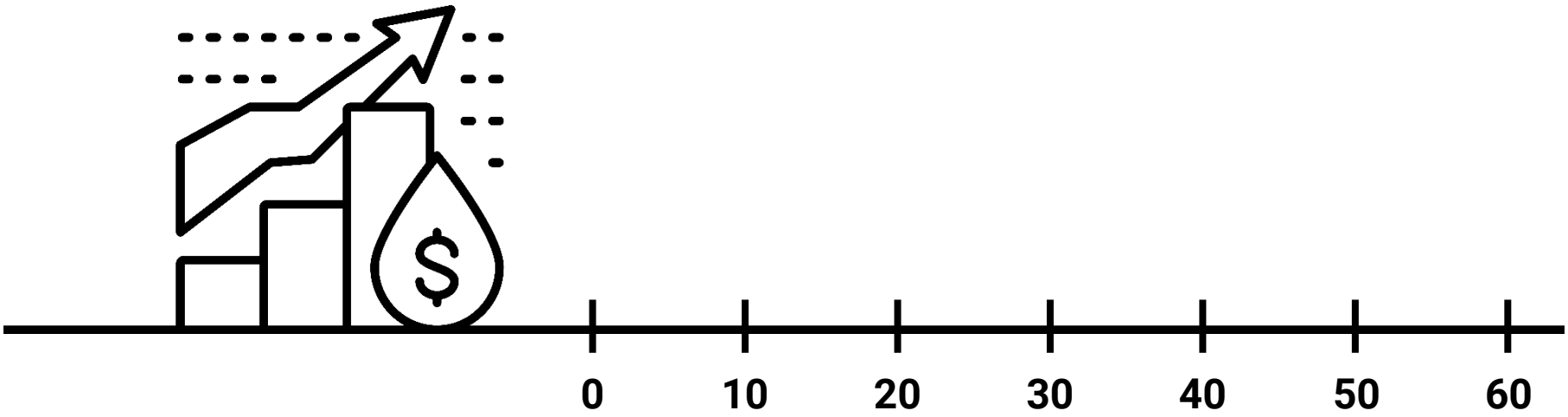
A variable of distance is **continuous**.

We can always find a smaller distance by dividing the current distance by half.



Regression

In finance, prices and rates are usually continuous.



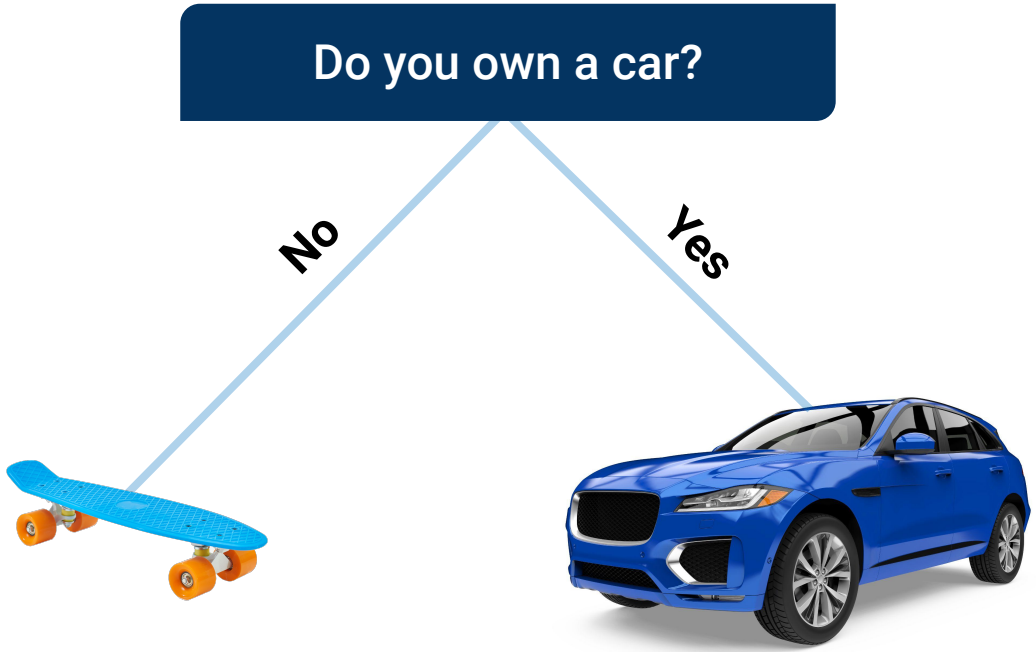
What is classification?

Classification is a method to predict discrete valued variables.

A discrete variable has no middle, and its values cannot be divided.

Classification

Consider a loan application that asks:



- The possible answers are yes or no.
- You either own a car or you don't.
- There is no middle value, so this type of variable, such as `car_ownership`, would be discrete.

Classification

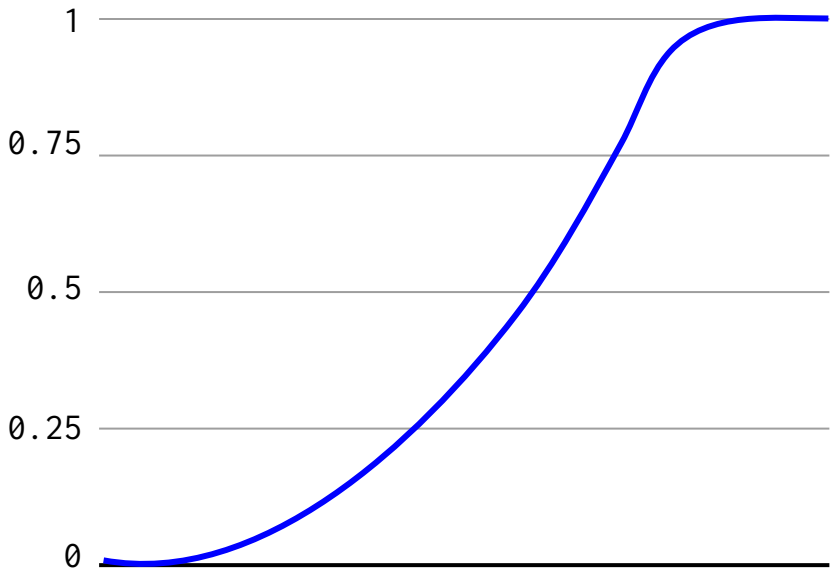
Data analysts use classification to draw categorical conclusions about data.

Instead of forecasting quantitative numbers, classification uses a binary (true-positive / true-negative) approach to predict membership in a category (i.e., will the outcome be of type A or type B).

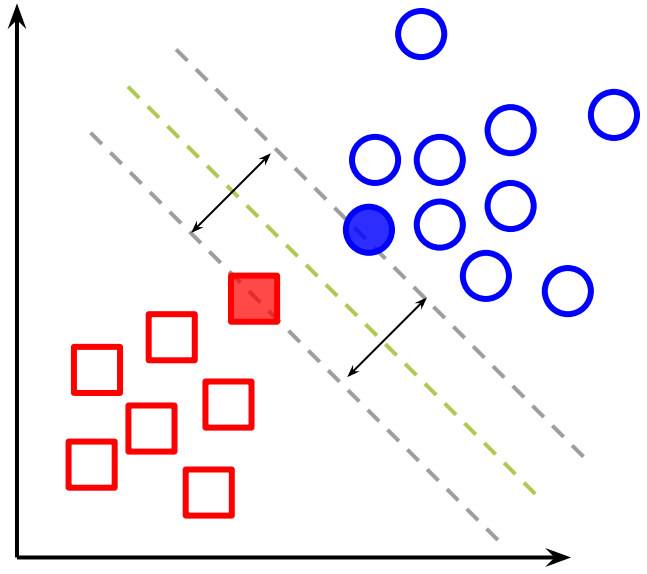
Qualified loan applicant	VS.	Unqualified loan applicant
Creditworthy		Credit risk
0		1

Classification

Today you'll learn to perform classification using **logistic regression**.



In the next class, you'll discover other tools for classification including **support vector machines**, decision trees, and random forests.



Classification

Classification models have greatly improved the ability for organizations to properly classify applicants, predict market decline, and classify fraudulent transactions or suspicious activity.

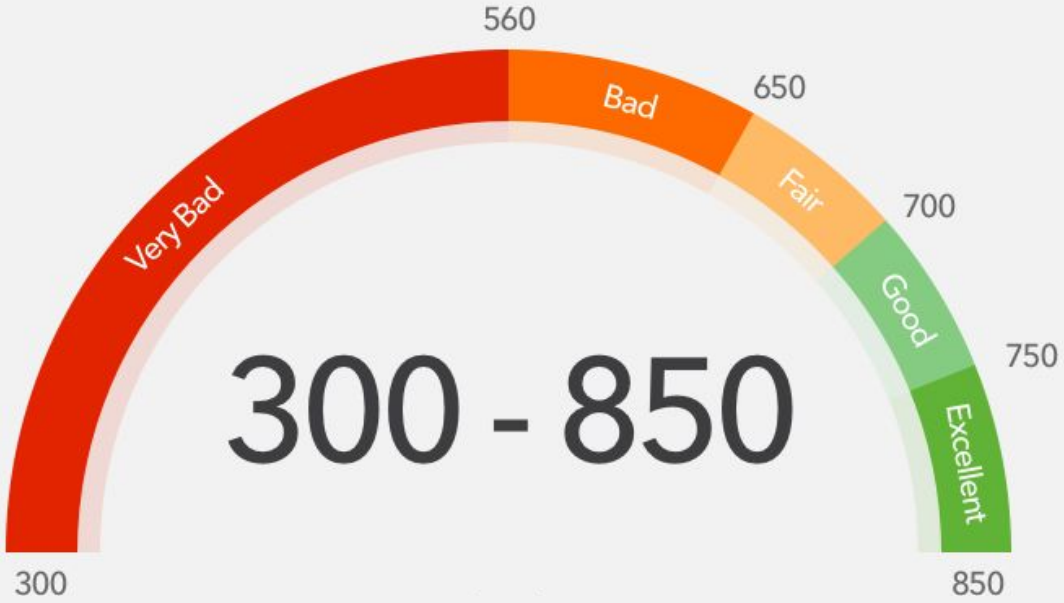
- Most large financial institutions use some form of machine learning to monitor and predict fraudulent activities.
- This is how banks know when to flag and decline transactions due to suspicion of fraud.

Classification

FICO credit scoring currently uses a classification model for their cognitive fraud analytics platform.

Classification models have allowed the financial industry to become more proactive.

Supervised learning algorithms can predict outcomes with a high degree of accuracy, which allows for more effective and efficient mitigation.



Homework Overview

This week’s homework focuses on creating a classification model for predicting and categorizing credit risk.

Homework Overview

You will use multiple models to complete the homework, including linear regression and decision trees.

These models are available in the `scikit-learn` package.

The homework comprises two goals:

01 Create a classification model that will categorize credit risk.

02 Compare and contrast the various machine learning models for their ability to classify credit risk.

Questions?

Linear Regression

Linear regression is a model for describing the relationship between a numerical response and one or more variables that explains the response.

Linear Regression

In statistics and machine learning:

Dependent variable

The numerical response is known as the **dependent variable** because its value depends on other variables.

We can use the term **target variable** for the dependent variable.

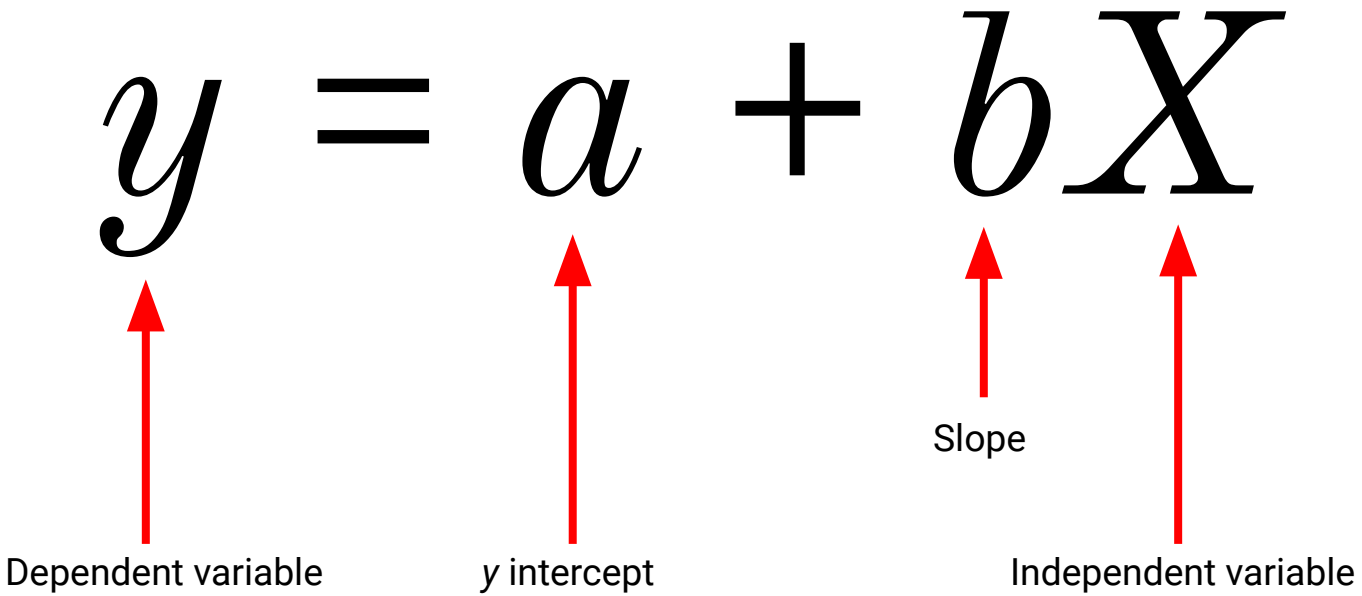
Independent variable

These other variables that explain the dependent variable are known as **independent variables**.

We can use the term **features** for independent variables.

Linear Regression

We will explore a simple linear regression with one independent variable.
This type of linear regression is represented by the following formula:



$$y = a + bX$$

This linear relationship implies the following:

- As *X* increases, *y* increases.
- How fast *y* increases in relation to *X* is called the slope.
- The slope is represented by the letter *b* in the formula.
- The value of *y* when *X* is 0 is called the *y*-intercept. It is represented by the letter *a*.

We consider a linear regression to be a supervised learning model because it can predict the value of *y* based on historical data.

Linear Regression

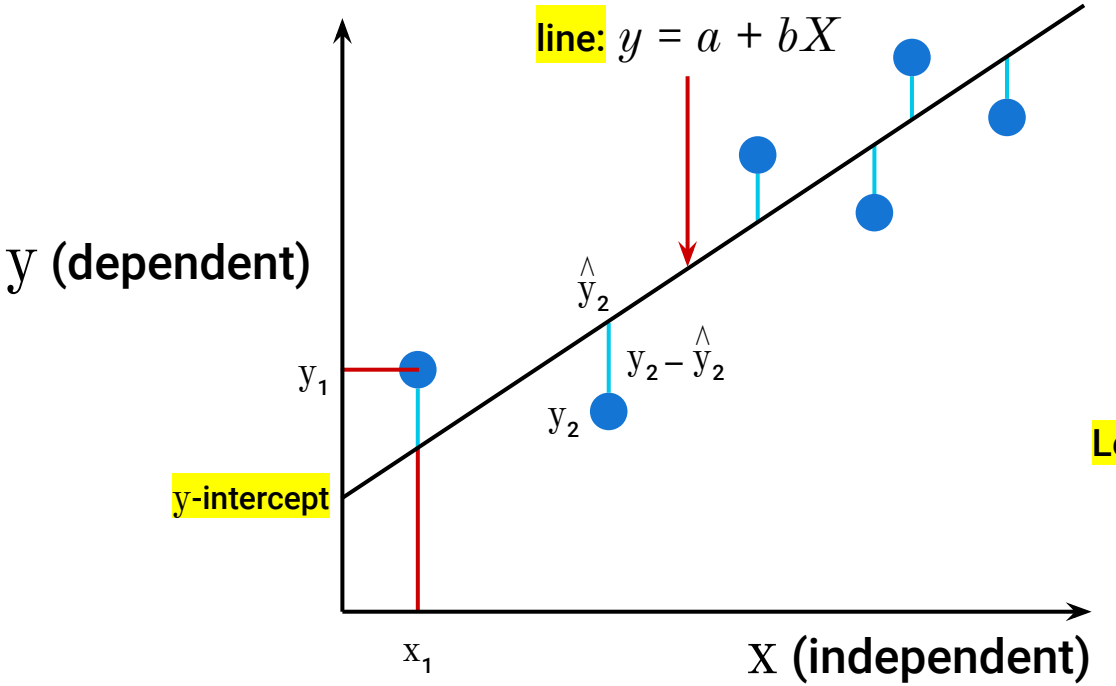
Let’s implement a simple linear regression model by using scikit-learn.

Instructor Demonstration

Linear Regression

Linear Regression Model

The linear regression model is mathematically constructed to minimize the sum of all the errors after they have been squared.



Minimize: $\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$

Least squares method

One way to assess the accuracy of a linear regression model is to observe the errors.

Linear Regression Model

The linear regression model is mathematically constructed to minimize the sum of all the errors after they have been squared, as the following image shows.

mean squared error (MSE)	The average of the square of the errors of the dataset. It is the variance of the errors in the dataset.
root mean square error (RMSE)	The square root of the MSE. It is therefore the standard deviation of the errors in the dataset.
correlation coefficient	A numerical description of the extent to which the two variables move together. It ranges from -1 to 1.
R2 or r-square value	The square of the correlation coefficient. It describes the extent to which a change in one variable is associated with the change in the other variable. It ranges from 0 to 1.

Low MSE and RMSE scores indicate a more accurate model.

Summary

Key points of linear regression:

It models data with a linear trend. It is not useful when the data does not follow a linear trend, e.g., exponential trends.

Based on the *X* values, it predicts *y* values.

It does not do a good job of describing nonlinear patterns.

We will cover techniques to model nonlinear data later in the course.

Questions?

Activity: Predicting Sales with Linear Regression

In this activity, you will learn apply linear regression to predict sales based on historical data.

Suggested Time:

15 Minutes

Time's Up! Let's Review.

Questions?

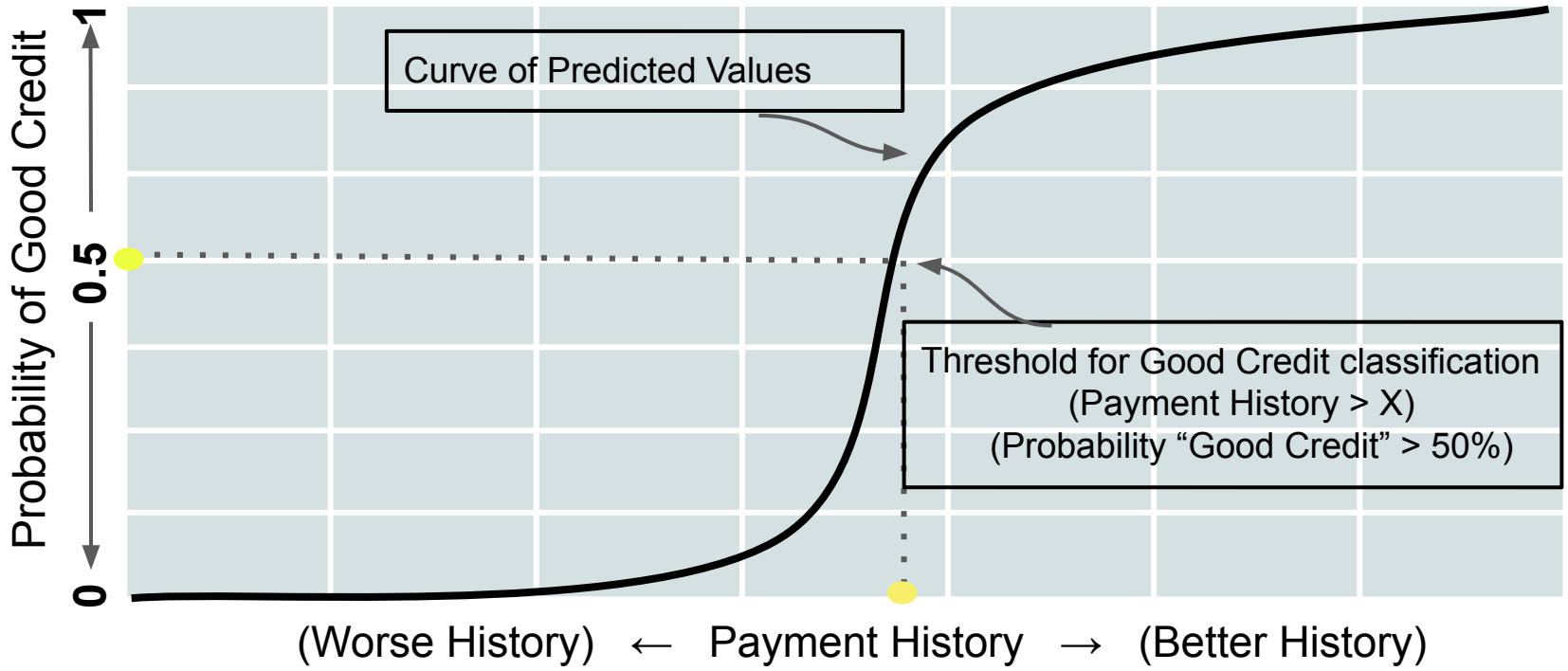

Break

Making Predictions with Logistic Regression

Logistic regression is a statistical method for predicting binary outcomes from data.

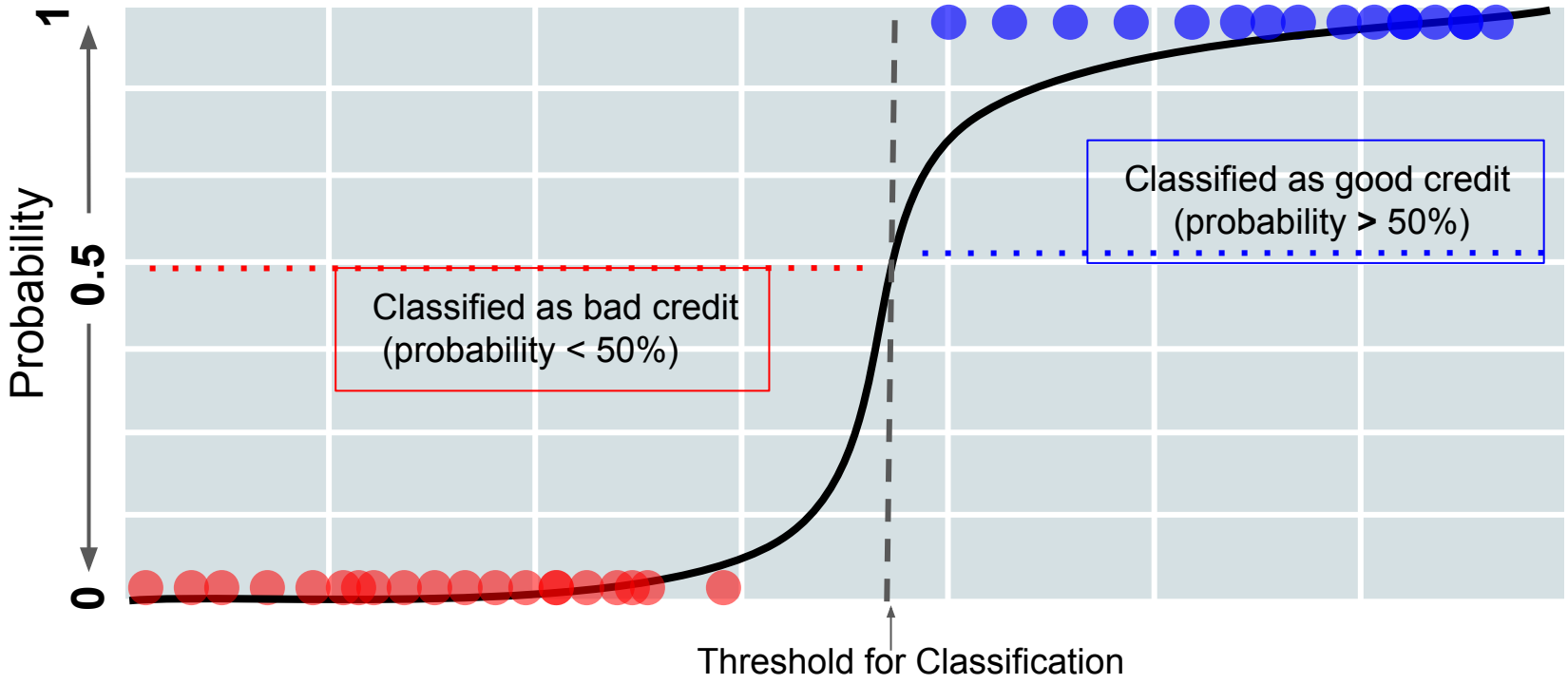
Making Predictions with Logistic Regression

Each data point receives a probability of being in the **1** category (e.g. “good credit”).



Making Predictions with Logistic Regression

If the probability is above a certain threshold, that data point is estimated to be a **1** (“good credit”). Below that threshold, the data point is a (**0**).



The Sigmoid Function

Logistic Regression converts using a **sigmoid (**or **squashing)** function:

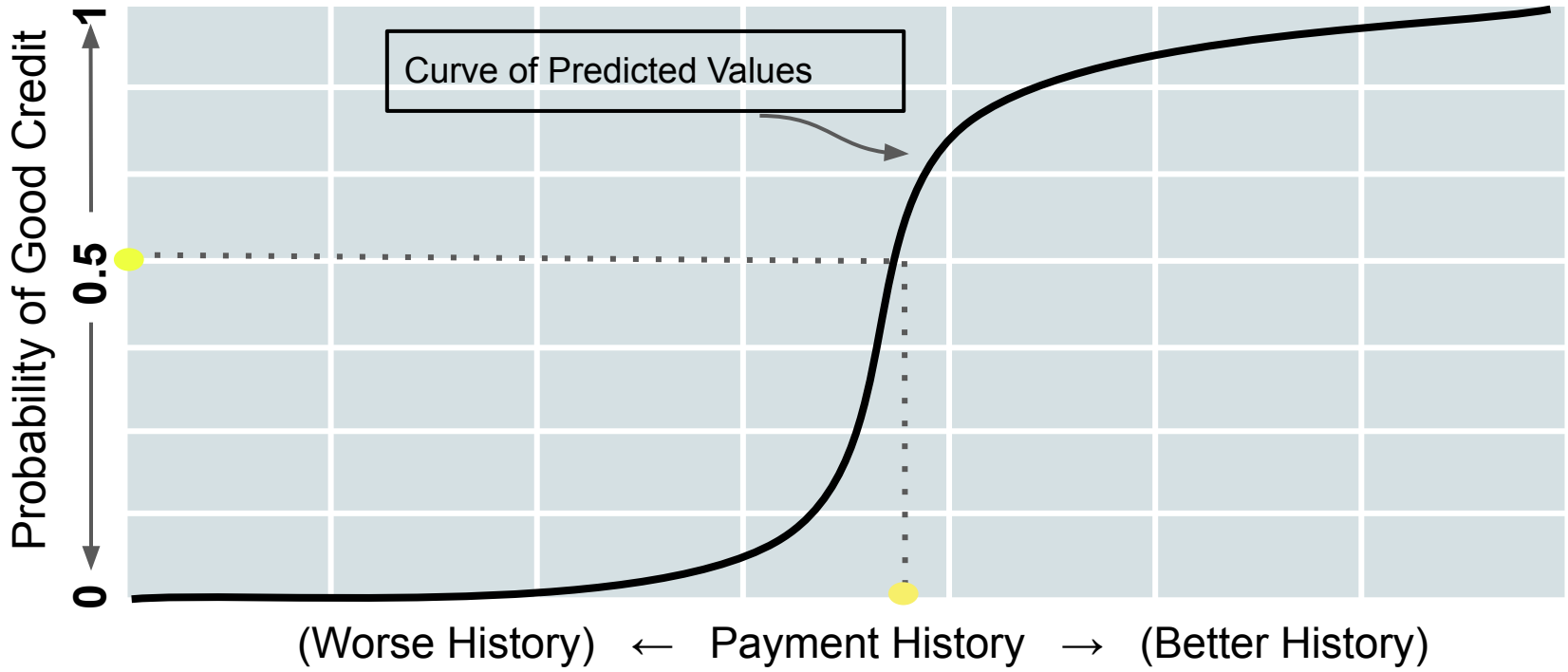
$$Probability\ of\ Good\ Credit = \frac{1}{1 + e^{-Payment\ History}}$$

Performing behind the scenes, this function converts continuous data on the borrower (e.g., number of months without a delinquent payment) to a percentage probability of being a “good credit” borrower.

A good logistic regression model will use more information than just “payment history,” but the **sigmoid** function can still convert all this information into a probability.

The Sigmoid Function

How do we translate continuous data like **payment history** into a probability of "good credit" ranging from 0 to 1?



Steps in Logistic Regression Modeling

We can use logistic regression to predict which category or class a new data point should go into.

01	02	03	04
Preprocess	**Train**	**Validate**	**Predict**
This step consists of cleaning the data (such as removing rows with missing values) and splitting it into subsets for training and testing the model.	Training is when we use a large subset of our labeled data to teach the model to recognize classification patterns.	In the validation step, we use a small subset of our labeled data to test how well the model is able to predict labels.	Finally, we use our model to predict labels for unclassified data.

Instructor Demonstration

Making Predictions with Logistic Regression

Hold-Out Validation

Hold-out validation is the concept applied with the function `train_test_split`.

We create two datasets where we hold out a subset of data that we refer to as the testing dataset.

This allows us to measure the performance of our model and parameter selections with a subset of data not used to train the model.

Questions?

Activity: Predicting Diabetes with Logistic Regression

In this activity, you will see how logistic regression works on the real-world problem of fraud detection.

Suggested Time:

20 Minutes

Time's Up! Let's Review.

Questions?

Evaluating Logistic Regression Predictions

Evaluating Logistic Regression Predictions

In this scenario, we used a logistic regression model to predict if an individual has diabetes, based on a set of diagnostic metrics provided as a dataset.

- We evaluated the logistic regression model by using a scoring feature.
- This revealed that the model is somewhat accurate.
- However, can you trust that its predictions are correct?

How sure are you that your models can actually predict diabetes?

78%

sure, as described by the scored accuracy.

Would you feel comfortable giving the diagnosis of diabetes based on the predictions of the model?

Answer

No.

The prediction is not 100% accurate.
There is room for error, and there are false positives.

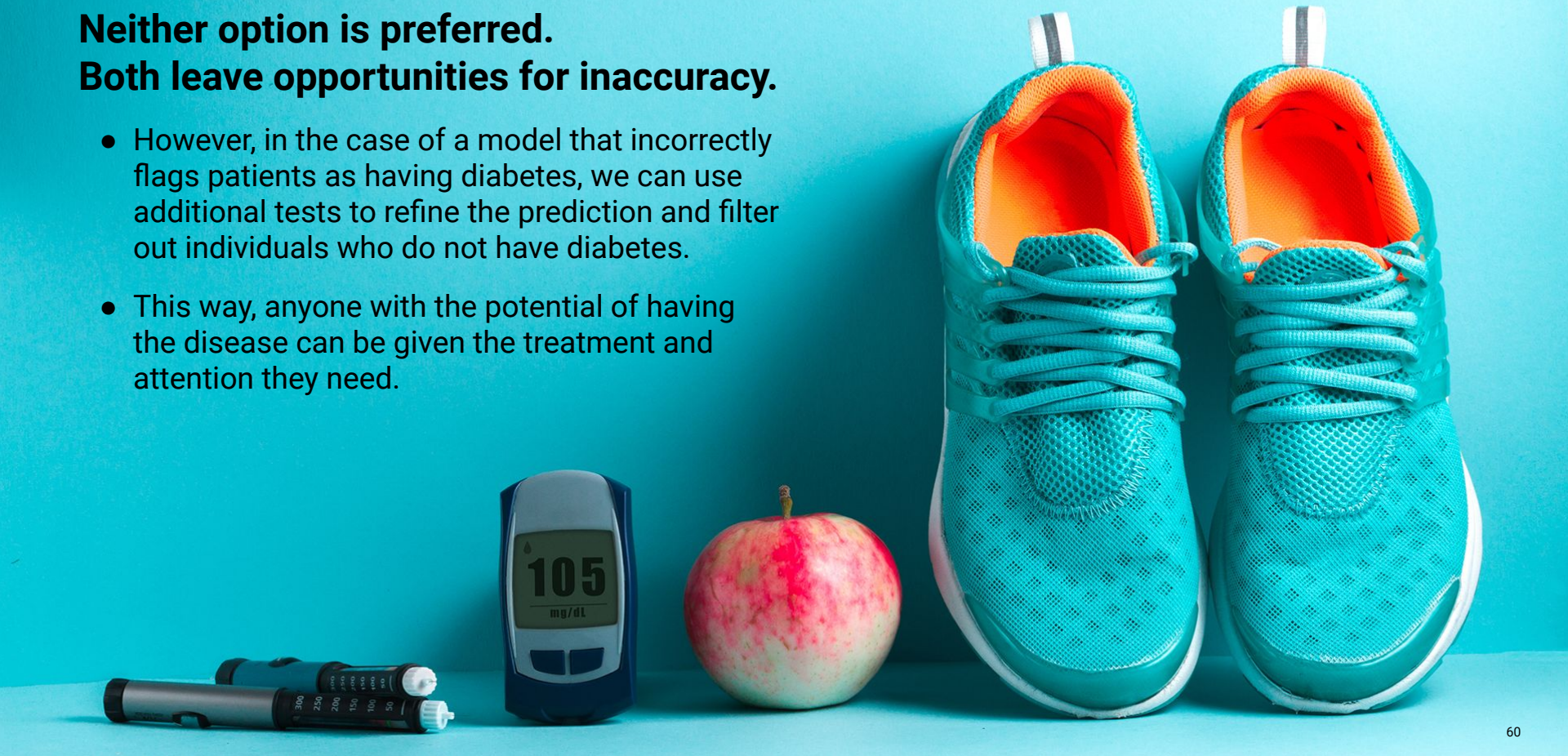
	Test says you don't have it	Test says you do have it
You don't really have it	True Negative	False Positive
You do really have it	False Negative	True Positive

What is better:
the false-positive or false-negative?

Answer

Neither option is preferred.
Both leave opportunities for inaccuracy.

- However, in the case of a model that incorrectly flags patients as having diabetes, we can use additional tests to refine the prediction and filter out individuals who do not have diabetes.
- This way, anyone with the potential of having the disease can be given the treatment and attention they need.



The process of evaluating a model requires more than simply scoring/measuring the model for accuracy.

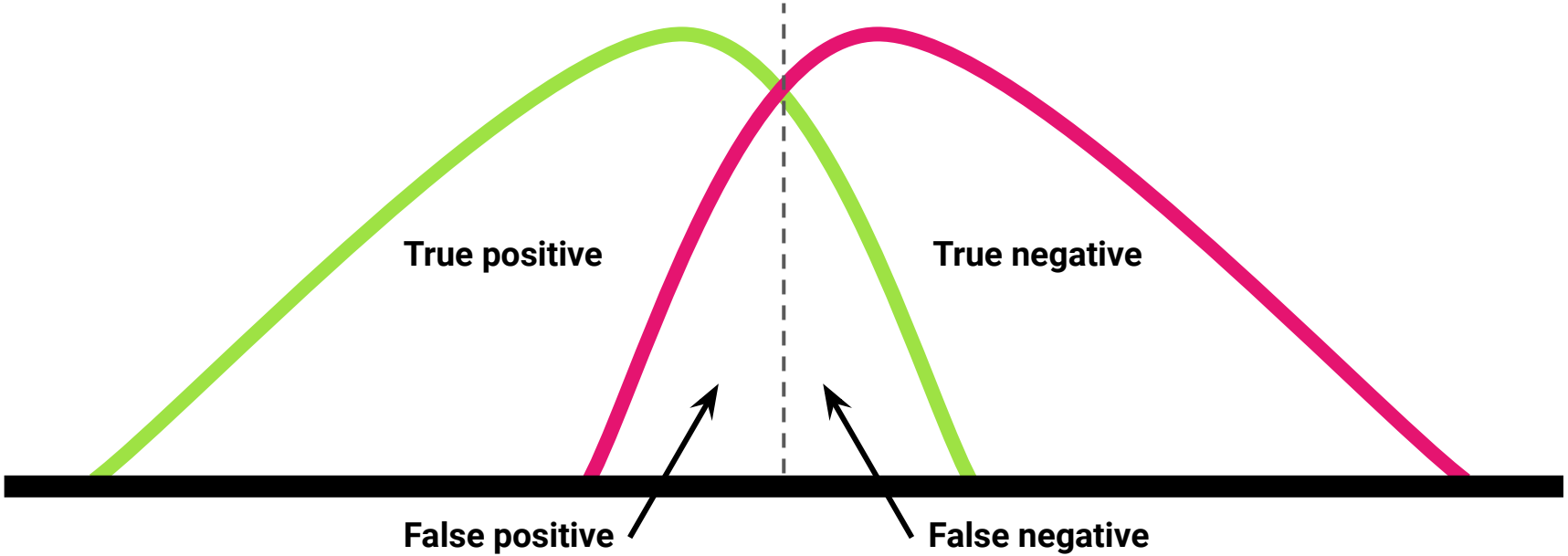

In addition to accuracy, we must measure for precision and recall, both of which can be used to eliminate false positives and false negatives.

Questions?

Accuracy, Precision, Recall

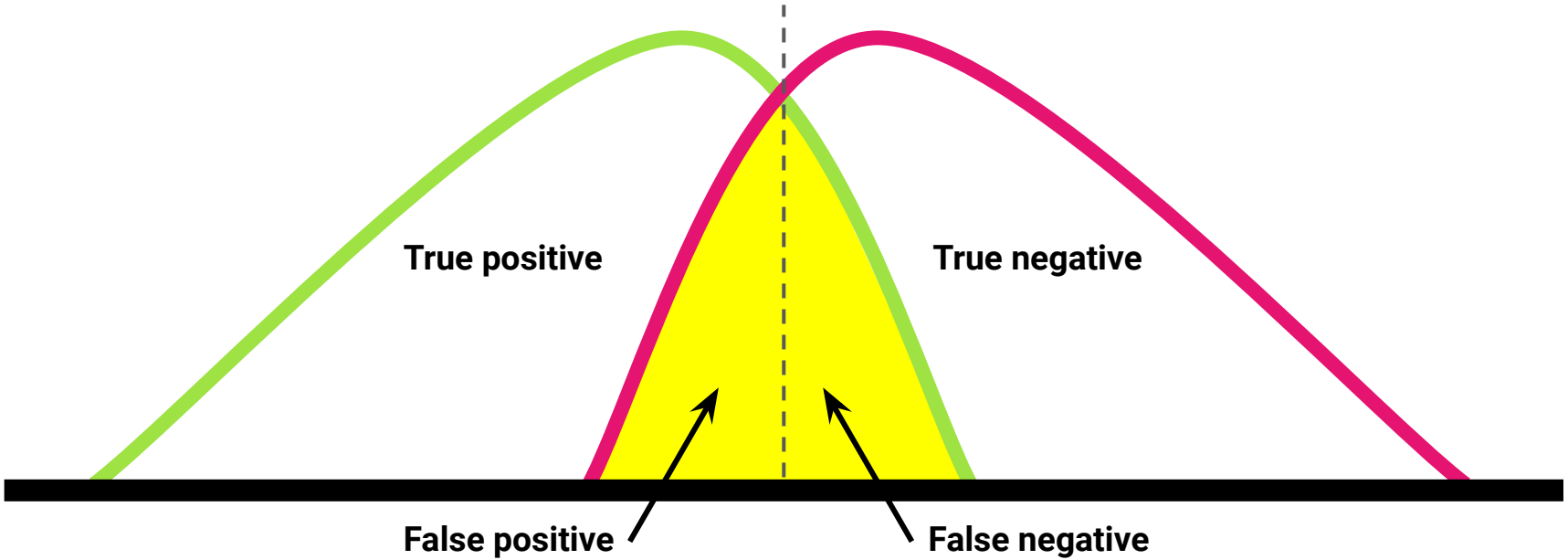
Accuracy, Precision, and Recall

Accuracy, precision, and recall are especially important for classification models that involve a binary decision problem. Binary decision problems have two possible correct answers: **True Positive** and **True Negative**.



Accuracy, Precision, and Recall

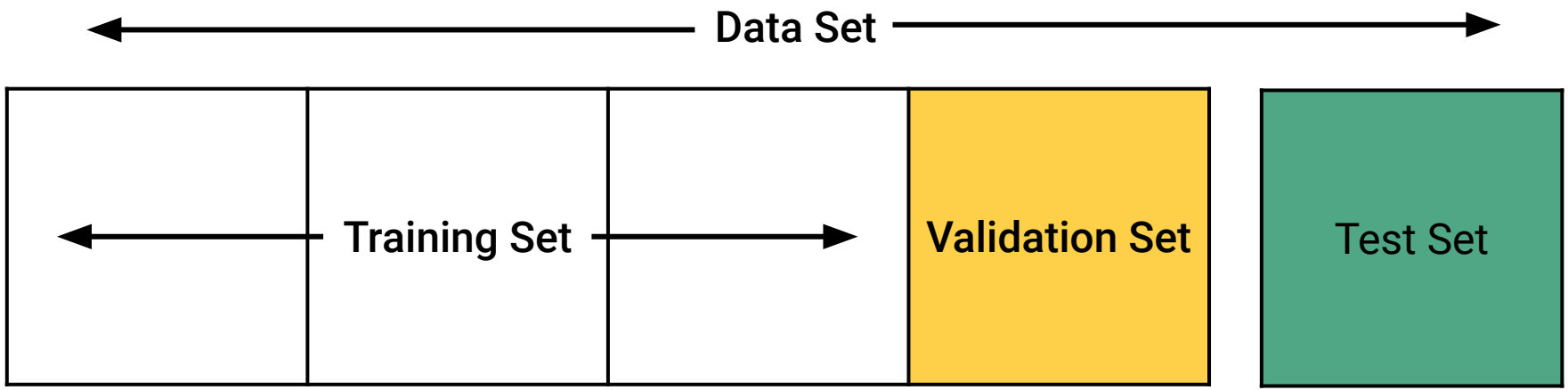
Inaccurate and imprecise models will return false positives and false negatives.



Accuracy is how often the model is correct—the ratio of correctly predicted observations to the total number of observations.

Accuracy

Scoring will reveal how accurate the model is. However, it does not communicate how precise it is.



Accuracy

Accuracy can be very susceptible to imbalanced classes. In the case of identifying poor loan candidates, the number of good loans greatly outweighs the number of at-risk loans. In this case, it can be really easy for the model to only care about the good loans, because these have the biggest impact on accuracy. However, we also care about the at-risk loans, so we need a metric that can help us evaluate each class prediction.

Calculation:

```
(TP + TN) / (TP + TN + FP + FN)
```

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

(i.e., of all the samples classified as having diabetes, how many actually have diabetes?)

Precision

Another example: For all the individuals who were classified by the model as being a credit risk, how many actually were a credit risk?

Did we classify them comprehensively and correctly?

Precision

High precision relates to a low false positive rate.

Calculation:

$$TP / (TP + FP)$$

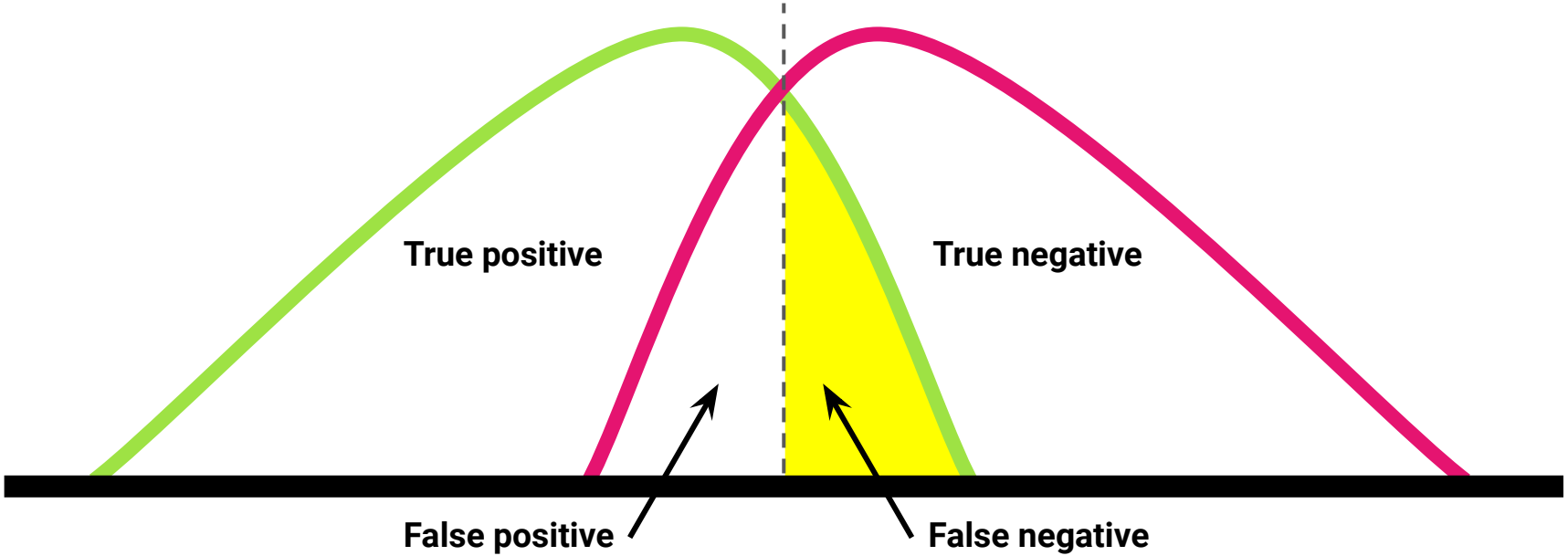
Recall is the ratio of correctly predicted positive observations to all predicted observations for that class.

(i.e., of all of the actual diabetes samples, how many were correctly classified as having diabetes?)

Recall

Of all of the actual diabetes samples, how many were correctly classified as having diabetes?

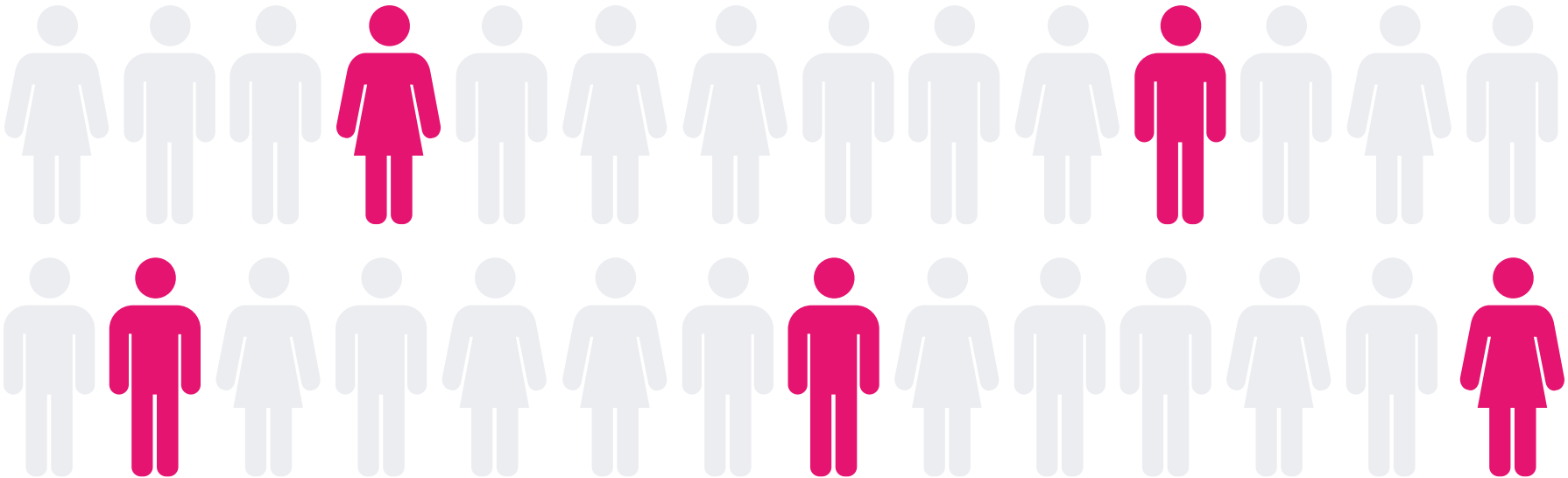
Did we classify all samples correctly, leaving little room for false negatives?



Recall

Another example:

For all the individuals who are a credit risk, how many were classified by the model as being a credit risk?



Recall

High recall correlates to a more comprehensive output and a low false negative rate.

Calculation:

$$TP / (TP + FN)$$

Questions?

Instructor Demonstration

Confusion Matrix & Classification Report

We use a **confusion matrix** to measure and gauge the success of a model.

Confusion Matrix

Confusion matrices reveal the number of true negatives and true positives (actuals) for each categorical class and compare them to the number of predicted values for each class.

n=165	**Predicted: No**	**Predicted: Yes**
Actual=No	50	10
Actual=Yes	5	100

Confusion Matrix

These values are then individually summed by column and row. The aggregate sums are then compared to assess accuracy and precision. If the aggregates match, we can consider the model to be accurate and precise.

n=165	**Predicted: No**	**Predicted: Yes**	
Actual=No	50	10	**=60**
Actual=Yes	5	100	**=105**
	=55	**=110**	

Confusion Matrix

Confusion matrixes are great because they describe the performance of the classification model.

By looking at the confusion matrix, we can determine whether the model has been correctly trained to produce comprehensive, accurate, and precise predictions.

For binary classifiers like the logistic regression classifier, a confusion matrix will measure the number of positive and negative predictions.

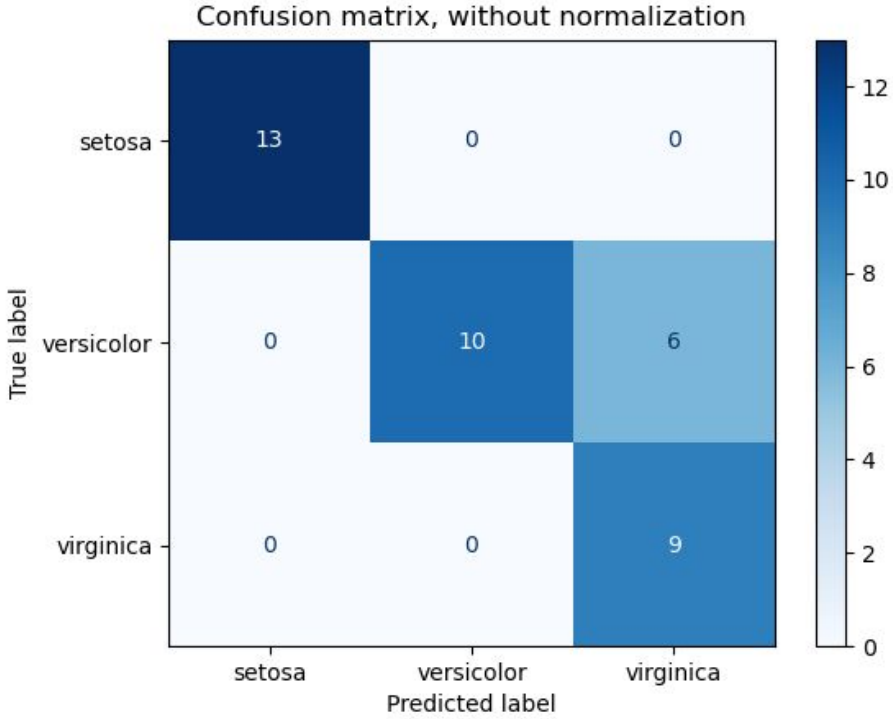
These will then be compared to the actuals.

How do we get data loaded into a confusion matrix for model evaluation?

Confusion Matrix

Most models will come equipped with a confusion matrix module, such as the [sklearn metrics.confusion_matrix] module.

The function accepts two arguments, one dataset containing the predicted values and another containing the actual data points.



You can also use a classification report to evaluate a model.

A **classification report** holds the test results so we can assess and evaluate the number of predicted occurrences for each class.

Classification Report

Classification report identifies the **precision**, **recall**, and **accuracy** of a model for each given class.

	precision	recall	fl-score	support
No Diabetes	0.77	0.90	0.83	125
Diabetes	0.72	0.49	0.58	67
accuracy			0.76	192
macro avg	0.74	0.69	0.71	192
weighted avg	0.75	0.76	0.74	192

Questions?

Confusion Matrix

Activity: Evaluating Classification Models

In this activity, you will practice building and interpreting classification reports and confusion matrixes on consumer usage data for a financial app.

Suggested Time:

20 Minutes

Time's Up! Let's Review.

Questions?


The End