AI Bootcamp

Introduction to Supervised Learning and Linear Regression

Module 12 Day 1

Class Objectives

By the end of class, you will be able to:

1. Explain the differences between unsupervised and supervised learning.
2. Explain the differences between regression and classification.
3. Understand and explain key terminology, such as features and labels, model-fit-predict, model evaluation, and training and testing data.
4. Apply the model-fit-predict process to single and multiple linear regression models.
5. Evaluate a linear regression model.
6. Preprocess data by encoding categorical data and splitting it into training and testing sets.

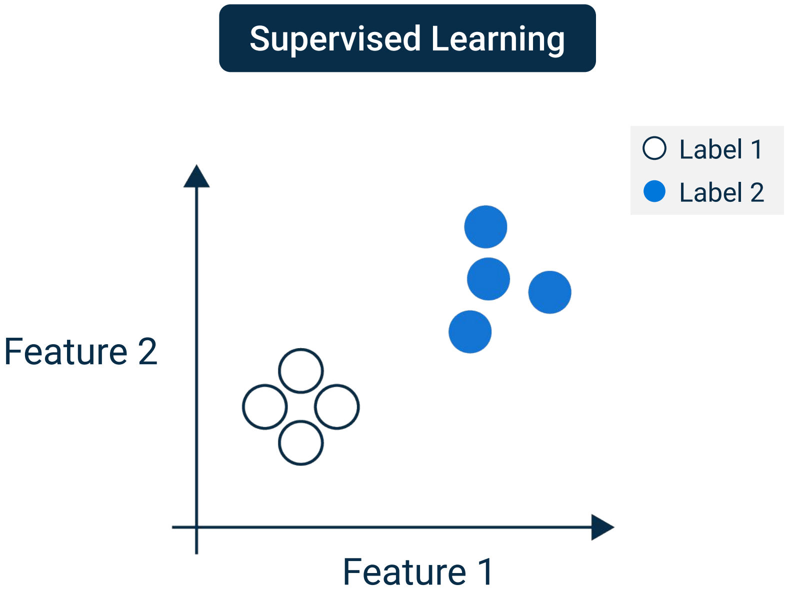
Welcome

Instructor **Demonstration**

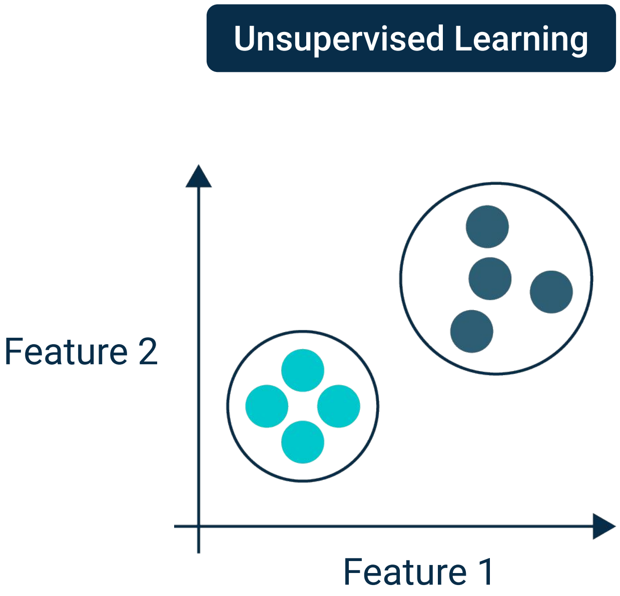
Supervised Learning Concepts

Supervised vs. Unsupervised Learning

These charts compare how supervised and unsupervised learning respectively label data or cluster patterns on a graph.



- Input data is labeled
- Uses training datasets
- Goal: Predict a class or value

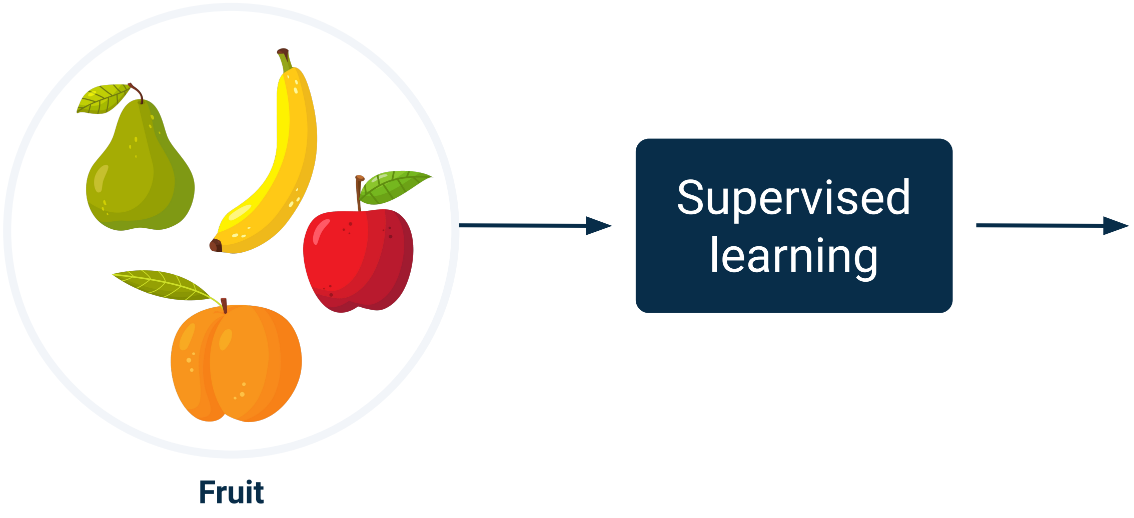


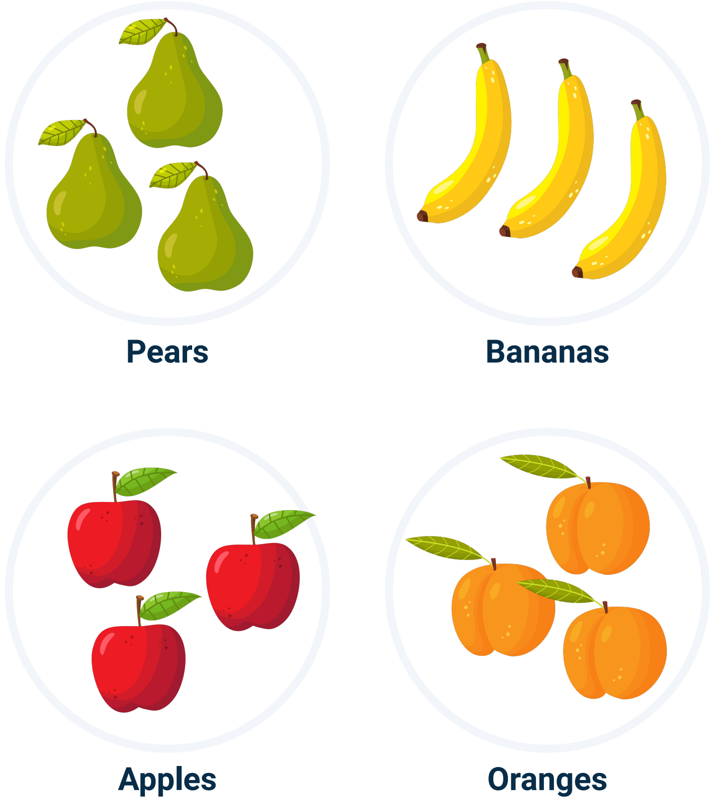
- Input data is unlabeled
- Uses just input datasets
- Goal: Determine patterns or grouping data

Introduction to Supervised Learning

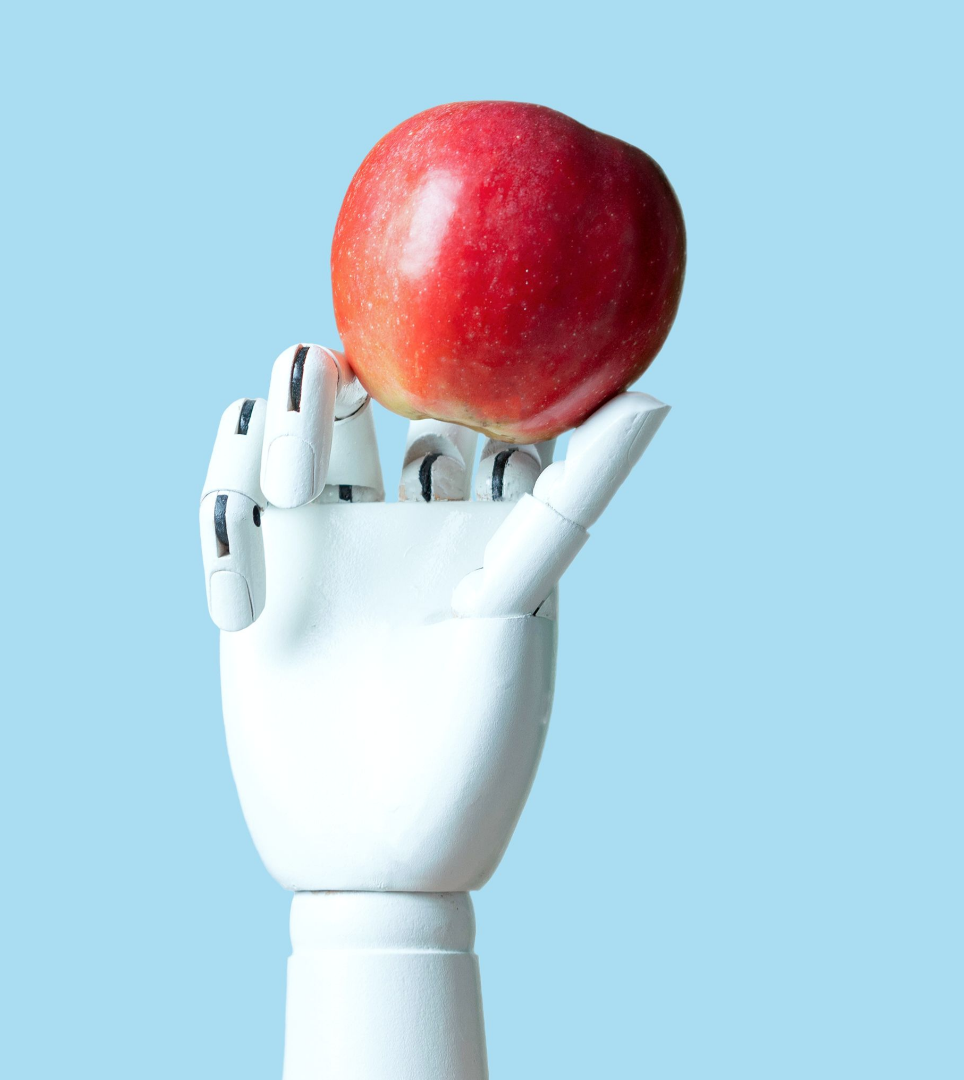
In supervised learning, we take a set of known answers called **labels** and fit a model with a set of **features** (inputs) that corresponds to the labels.

These models are called **supervised learners**.



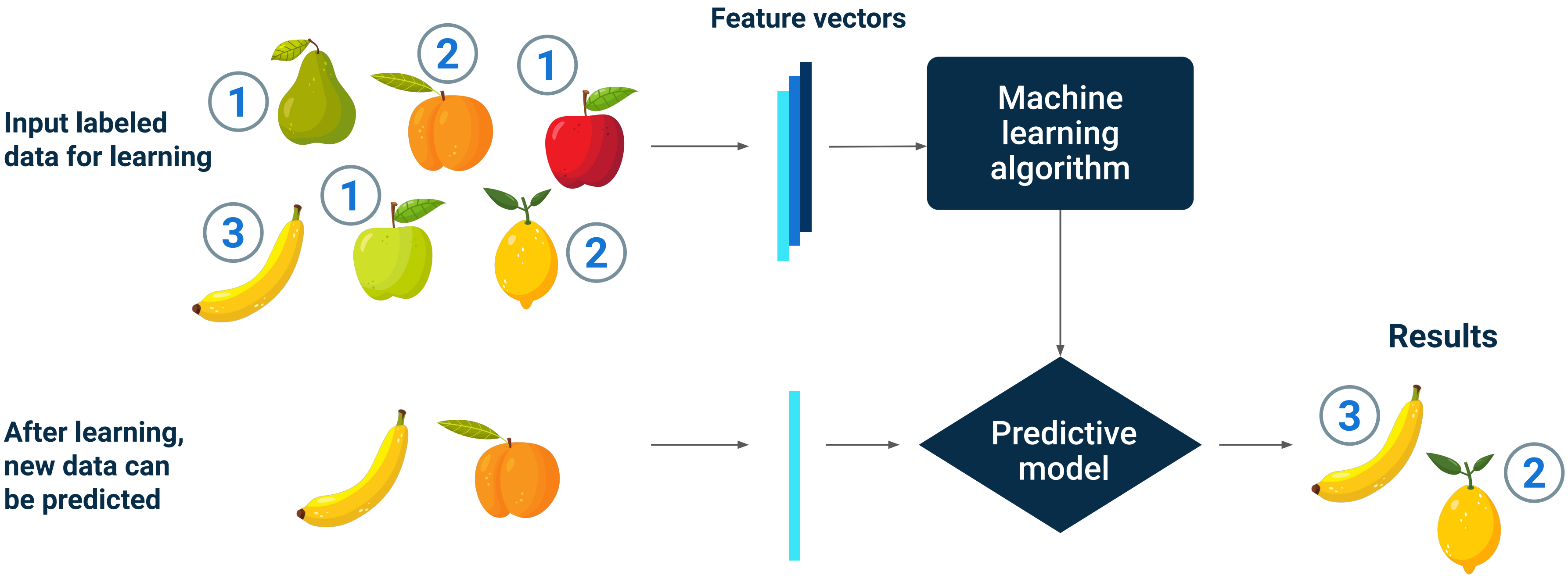


Supervised learning requires us to feed the correct answers to the model.



Introduction to Supervised Learning

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



Supervised vs Unsupervised Learning

The main differences between supervised and unsupervised learning:

	Supervised Learning	Unsupervised Learning
Human Intervention	Needed	Limited
Data	Labeled	Unlabeled
Tools	Linear and logistic regression, decision trees, and support vector machines	Clustering and dimensionality reduction
Models	Regression and classification	Clustering and association
Action	Model is trained to use known results and update the model's parameters.	Model segments the data into groups or reduces the dimensionality of the data.
Goal	Forecast and predict a predefined output.	Find hidden patterns in the data.
Results	Results include metrics that reflect the model's ability to generalize the data.	Exploratory results that help identify groups or anomalies within the data
Disadvantages	Labeled data is always required. Models may suffer from overfitting or underfitting.	May not produce meaningful results. The results are always subject to interpretation.

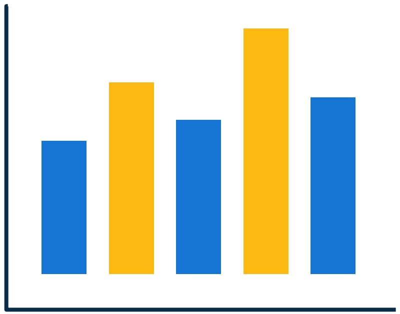
What is **regression?**

Regression

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

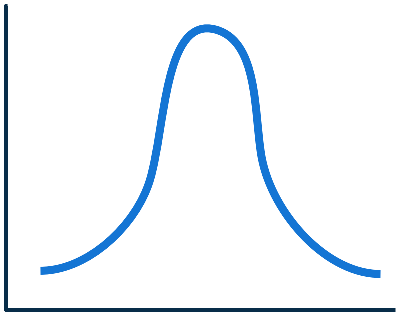
Continuous values

Can always be divided into smaller pieces



Continuous variables

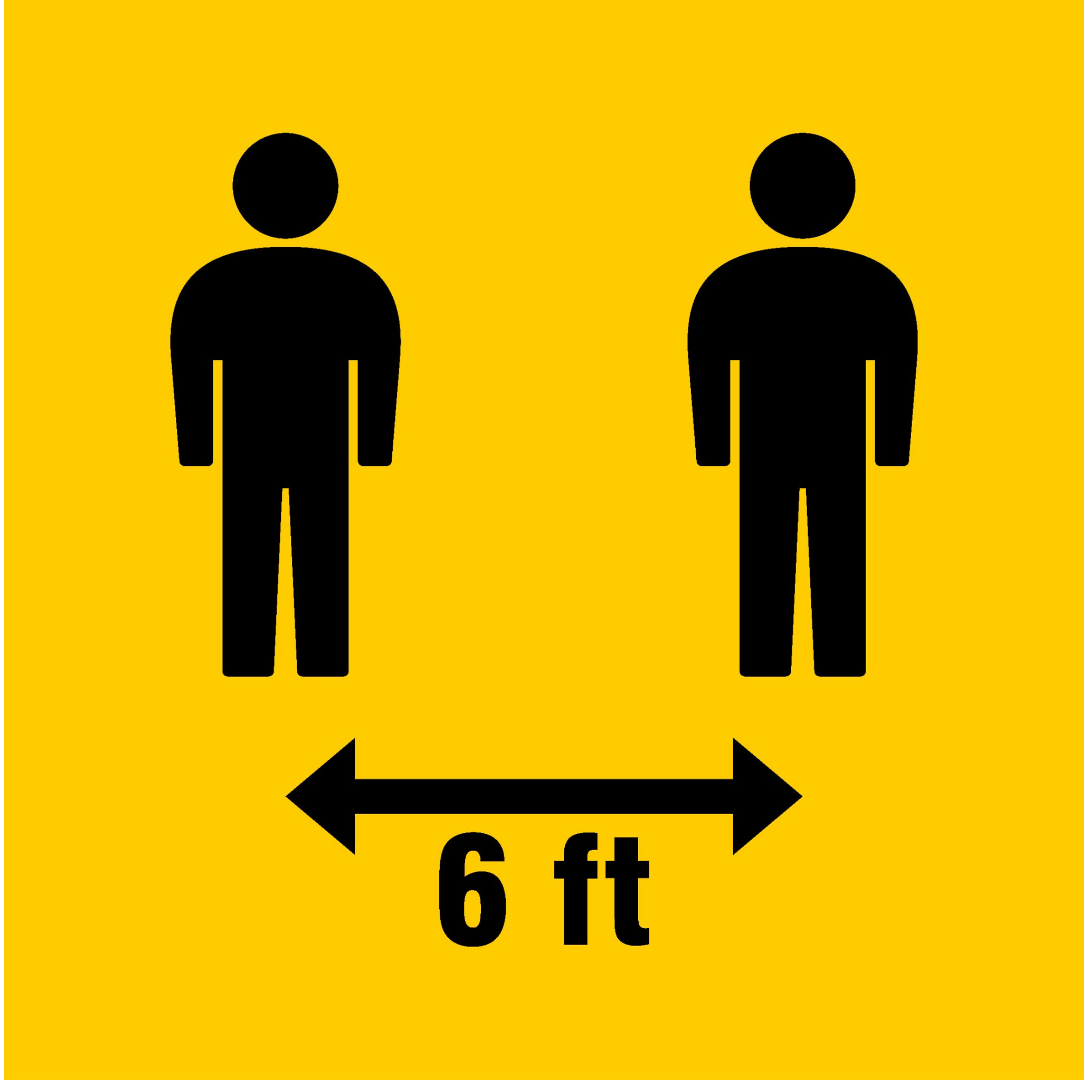
No matter how small, these will have a middle that we can find



Regression

A variable of distance is **continuous**.

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



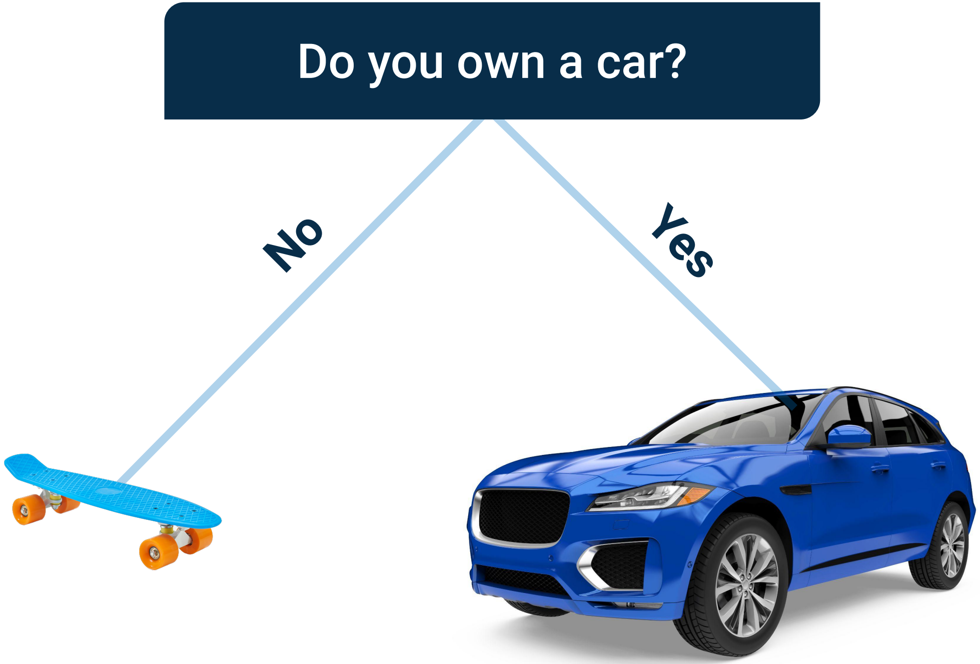
What is **classification?**

Classification is a method to predict discrete valued variables.

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

Classification

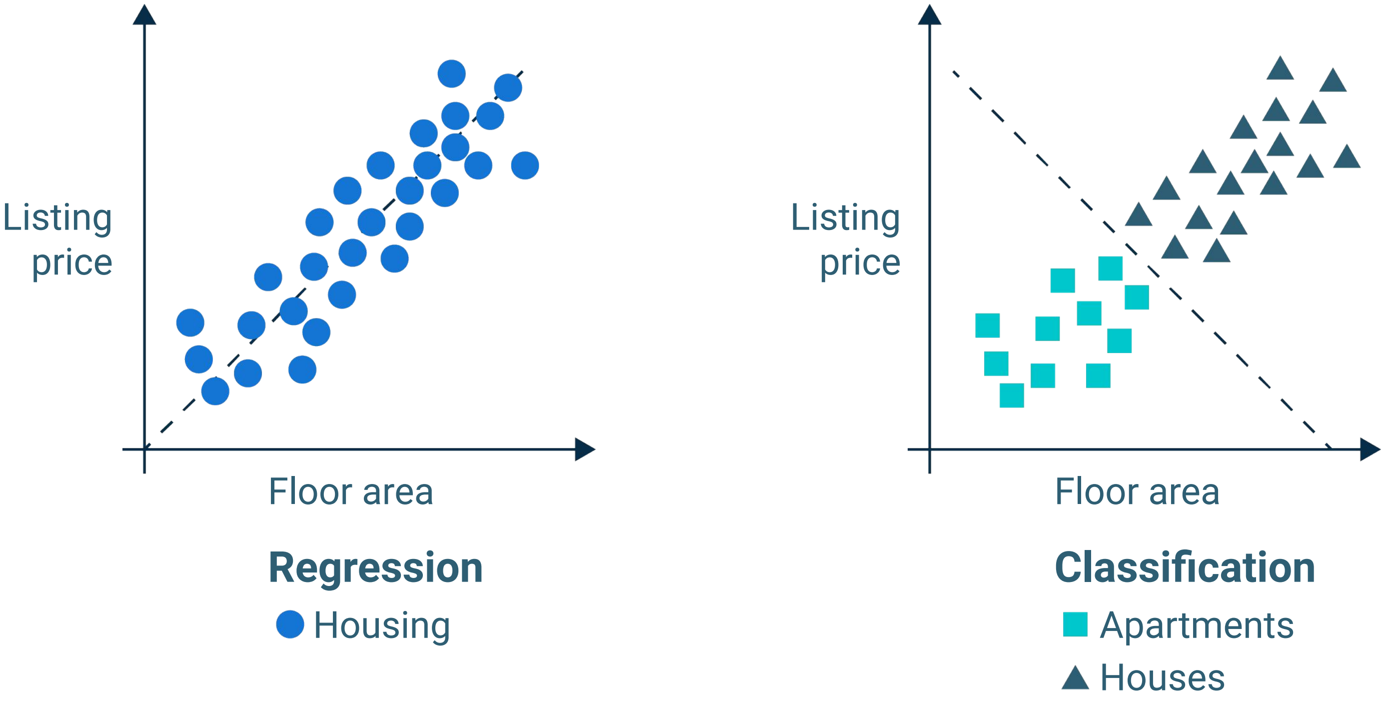
Consider a loan application that asks:



- 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`**, is discrete.

Classification

The difference between regression and classification plots for housing data:



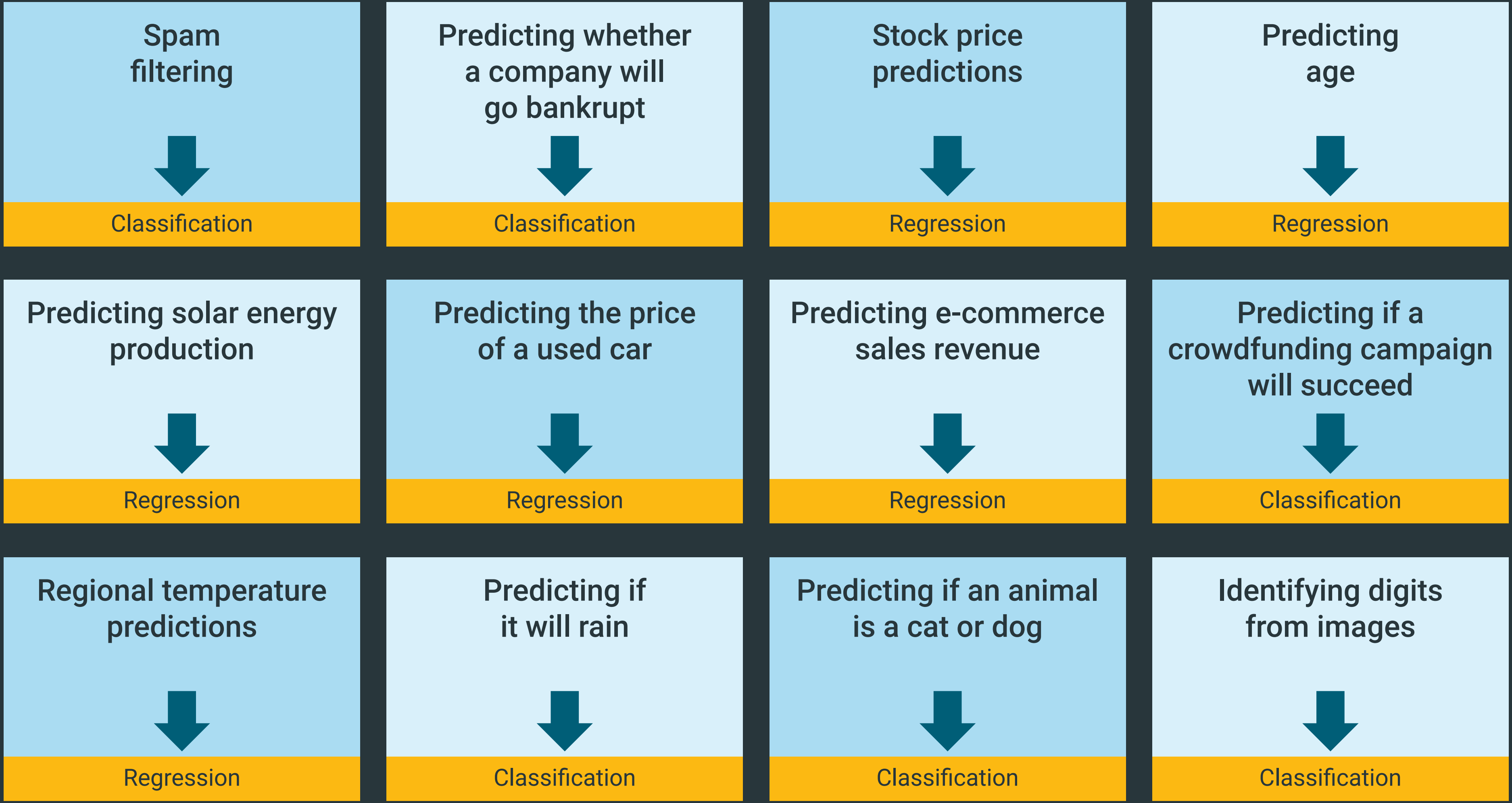
Activity:

Regression vs. Classification Scenarios

In this activity, you will determine when a regression or classification model is appropriate.

Suggested Time:

10 Minutes


Spam filtering
Classification
Predicting whether a company will go bankrupt
Classification
Stock price predictions
Regression
Predicting age
Regression
Predicting solar energy production
Regression
Predicting the price of a used car
Regression
Predicting e-commerce sales revenue
Regression
Predicting if a crowdfunding campaign will succeed
Classification
Regional temperature predictions
Regression
Predicting if it will rain
Classification
Predicting if an animal is a cat or dog
Classification
Identifying digits from images
Classification

Time’s up!

Let’s review

Instructor **Demonstration**

Supervised Learning Concepts *(continued)*

Features and Labels

In supervised learning, programmers categorize data into features and labels.

Features

- Also known as **independent variables**
- Can be used to **predict** changes in the target variable, but aren't necessarily influenced by other variables themselves
- Some features may have little to no impact on the label, but remain attributes of the label

Labels

- Also known as **dependent variables**
- **Depend** on the features
- The outcome that you want to predict
- Sometimes called **target variable**

Features and Labels

Features

	goal	pledged	backers_count	country	staff_pick	spotlight	category	days_active	outcome
0	100	0	0	3	0	0	0	17	0
1	1400	14560	158	0	0	1	1	27	1
2	108400	142523	1425	4	0	0	2	20	1
3	4200	2477	24	0	0	0	1	40	0
4	7600	5265	53	0	0	0	3	4	0
5	7600	13195	174	5	0	0	3	19	1
6	5200	1090	18	1	0	0	4	1	0
7	4500	14741	227	5	0	0	3	24	1
8	6200	3208	44	0	0	0	5	49	0
9	5200	13838	220	0	0	0	6	48	1

Labels

Source

The Pattern of Machine Learning

Model
Choose a model appropriate for the data

Fit
Let the model learn based on existing data

Predict
Have the model make predictions for new data

Instructor **Demonstration**

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

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

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

Linear Regression

A simple linear regression has one independent variable.
This type of linear regression is represented by the following formula:

$$\underset{\substack{\uparrow \\ \text{Dependent variable}}}{y} = \underset{\substack{\uparrow \\ y\text{-intercept}}}{a} + \underset{\substack{\uparrow \\ \text{Slope}}}{b}\,\underset{\substack{\uparrow \\ \text{Independent variable}}}{X}$$

$$y = a + bX$$

This linear relationship implies the following:

- In a positive relationship, as *X* increases, *y* increases.
  In a negative relationship, as *X* increases, *y* decreases.
- 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.

Key Points of Linear Regression Model

- 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 this course.

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

Activity:

Predicting Sales Linear Regression

In this activity, you will train a linear regression model.

Suggested Time:

20 Minutes

Time’s up!

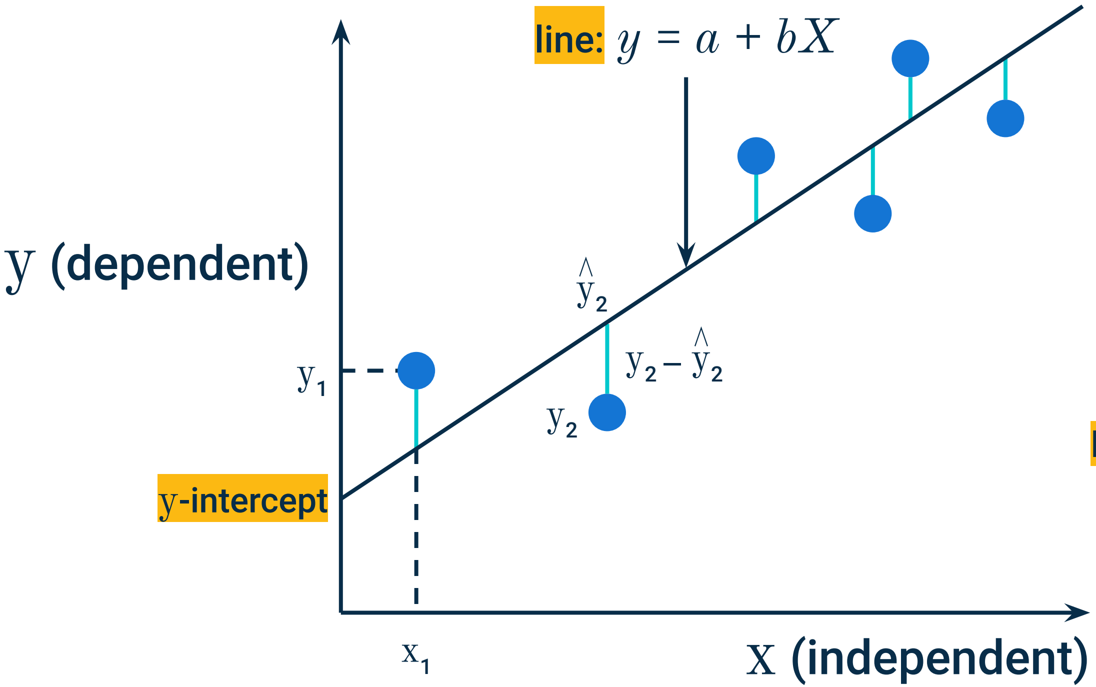
Let’s review

Instructor **Demonstration**

Model Evaluation

Linear Regression Model

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



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

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.

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 squared error (RMSE)	The square root of the MSE. It is the standard deviation of the errors in the dataset.
R^2 or R-squared 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.

Activity:

Evaluating Sales Prediction Model

In this activity, you will evaluate a linear regression model by checking its R-squared, MSE, and RMSE scores.

Suggested Time:

10 Minutes

Time’s up!

Let’s review

Break

15 mins

Activity:

Predict Electricity Generation

In this activity, you will use a linear regression model with real-world data.

Suggested Time:

15 Minutes

Time's up!

Let's review

Instructor **Demonstration**

Preprocessing Data

Real-world data **almost always needs to be processed** before it can be used in a machine learning algorithm.

Preprocessing Data

Two major preprocessing steps are converting categorical data and scaling:

Converting categorical data

Categorical data that is non-numeric needs to be converted to numeric data.

Scaling

Some machine learning algorithms are sensitive to large data values, so features need to be scaled to standardized ranges.

One-Hot Encoding
and Label Encoding

Label Encoding

Label encoding turns categorical variables into a series of integers.

It can cause problems though, because the difference between Saturday and Sunday is -6, but the difference for other consecutive days is +1.

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
0	1	2	3	4	5	6

One-Hot Encoding

One-hot encoding creates new “dummy” features for each category with 0 and 1 as Boolean values.

So the Weekday feature becomes 6 new features:

`isSunday`	`isMonday`	`isTuesday`	`isWednesday`	`isThursday`	`isFriday`	

Why is there no `isSaturday`?

One-Hot Encoding

`isSaturday` can be reconstructed from the other six. This means `isSaturday` is **collinear** with the other dummy features. Including `isSaturday` is an example of "the dummy trap," which can cause errors in the machine learning model.

isSunday	isMonday	isTuesday	isWednesday	isThursday	isFriday	

For each observation, only one of the new features will have a value of 1. This is why we call it "**one-hot encoding**."

One-Hot Encoding with Pandas

This requires the `get_dummies()` function.

- In Pandas, the `get_dummies()` function performs one-hot encoding. It can be applied to an entire DataFrame at once and returns a DataFrame of dummy-coded data.
- Setting the `drop_first` argument to `True` will automatically avoid the dummy trap.

Label Encoding with Pandas

This means using the `df.astype("category").cat.codes` method.

- In pandas, a column can be label encoded using the `astype()` function, converting the column to the `category` data type, and then using `.cat.codes` to create the numerical labels.

- This is a nonbinary method of encoding data and more useful for encoding target variables for classification models rather than features.

Instructor **Demonstration**

Training and Testing Data

Training and testing data can be used for **evaluating the performance** of any supervised learning algorithm.

Training and Testing Data

Splitting the data involves dividing the features and labels into two subsets.

Training dataset

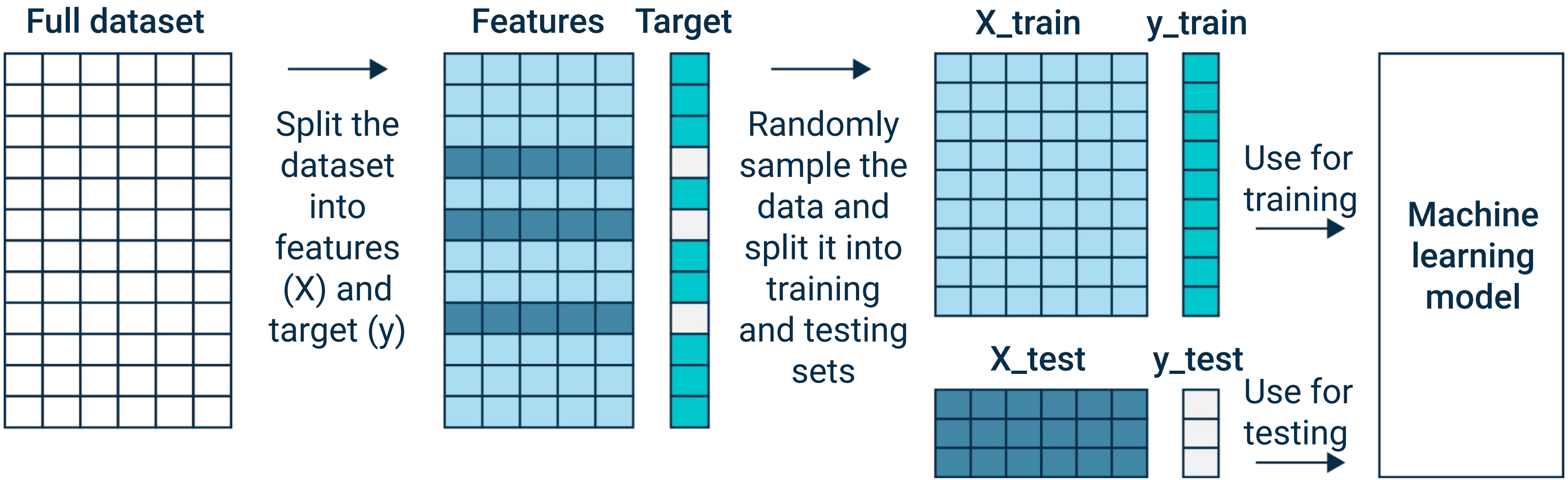
This first subset is used to fit the machine learning model to the data. The model learns from this dataset.

Testing dataset

This second subset is used to evaluate how well the machine learning model performs. It is used to test, rather than train, the model.

Splitting Training and Testing Data

The process of using the scikit-learn method `train_test_split()`.



- Manually split data into features (X) and target (y)
- `train_test_split(X, y)` randomly samples the data to split it into training and testing sets.
- Default split is 75% training data and 25% testing data.
- `X_train` and `y_train` are used for training the model.
- `X_test` and `y_test` are used for evaluating the model.

Instructor **Demonstration**

Multiple Linear Regression

Activity:

Multiple Linear Regression

In this activity, you will train a linear regression model with multiple features and use training data to train a model and testing data to evaluate a model.

Suggested Time:

10 Minutes

Time’s up!

Let’s review

Let’s **recap**

Review the Class Objective

In this lesson you learned how to:

1. Explain the differences between unsupervised and supervised learning.
2. Explain the differences between regression and classification.
3. Understand and explain key terminology such as features and labels, model-fit-predict, model evaluation, and training and testing data.
4. Apply the model-fit-predict process to single and multiple linear regression models.
5. Evaluate a linear regression model.
6. Preprocess data by encoding categorical data and splitting it into training and testing sets.

Next

In the next lesson, you will learn advanced regression techniques to improve model performance.

Questions?

The End