Data Wrangling in Python for Machine Learning Engineers

GETTING STARTED

Mike West
MACHINE LEARNING ENGINEER

Module Overview

What is supervised learning?

What is the machine learning process?

Why clean data is so important?

Skills for this course

Wrangle your data set in Python

Overview

The Pandas Dataframe

Titanic Dataset

Model our wrangled data set

Summary

Your Skills

Not Required	Required
Machine Learning	Basic Python
Deep Python knowledge	Tables, columns and rows
Statistics or advanced math	Basic math and statistics

Why Take this Course?

Required real-world skill

In the real-world, machine learning engineers spend most of their time wrangling data.

Why Take this Course?

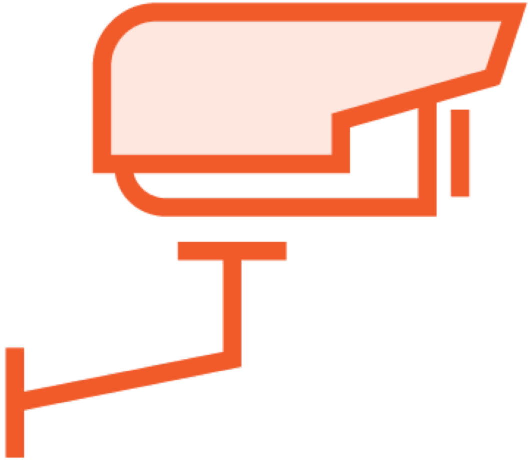
Required real-world skill

Applied machine learning is data wrangling

Models need well cleansed numerical data

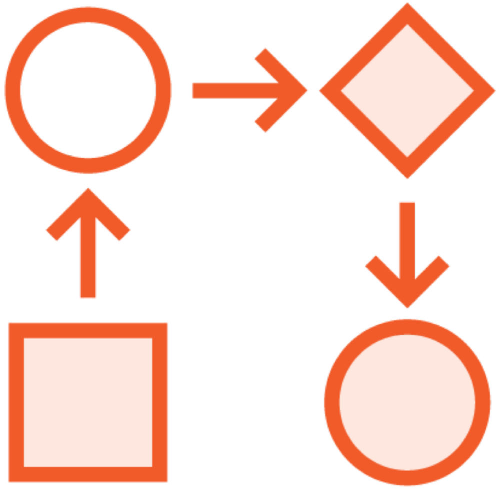
Improved model performance

Two Types of Machine Learning



Supervised

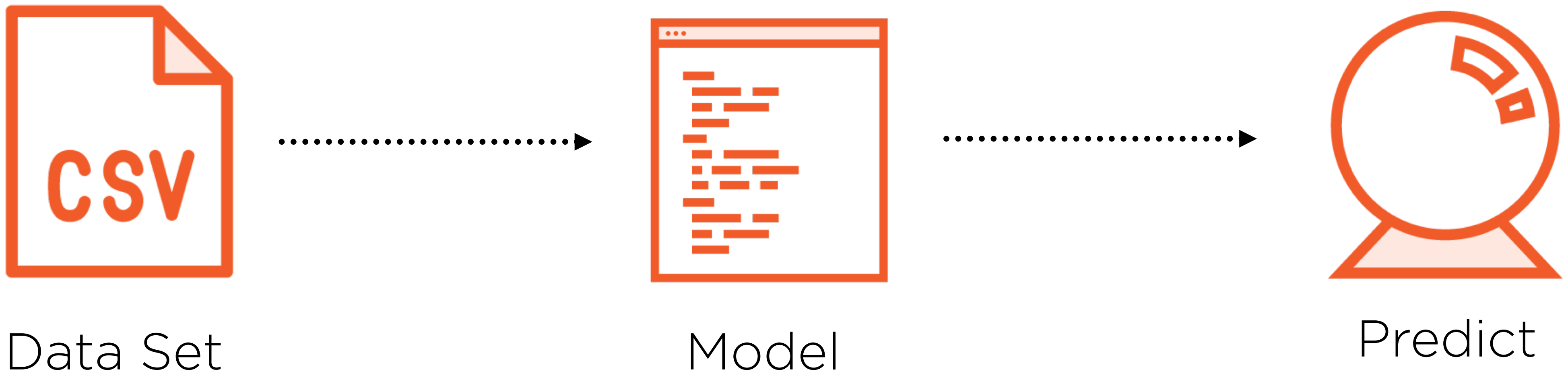
The models are fed clean labeled numerical data.



Unsupervised

The models find patterns and structure in unlabeled data.

Supervised Machine Learning



Data Set

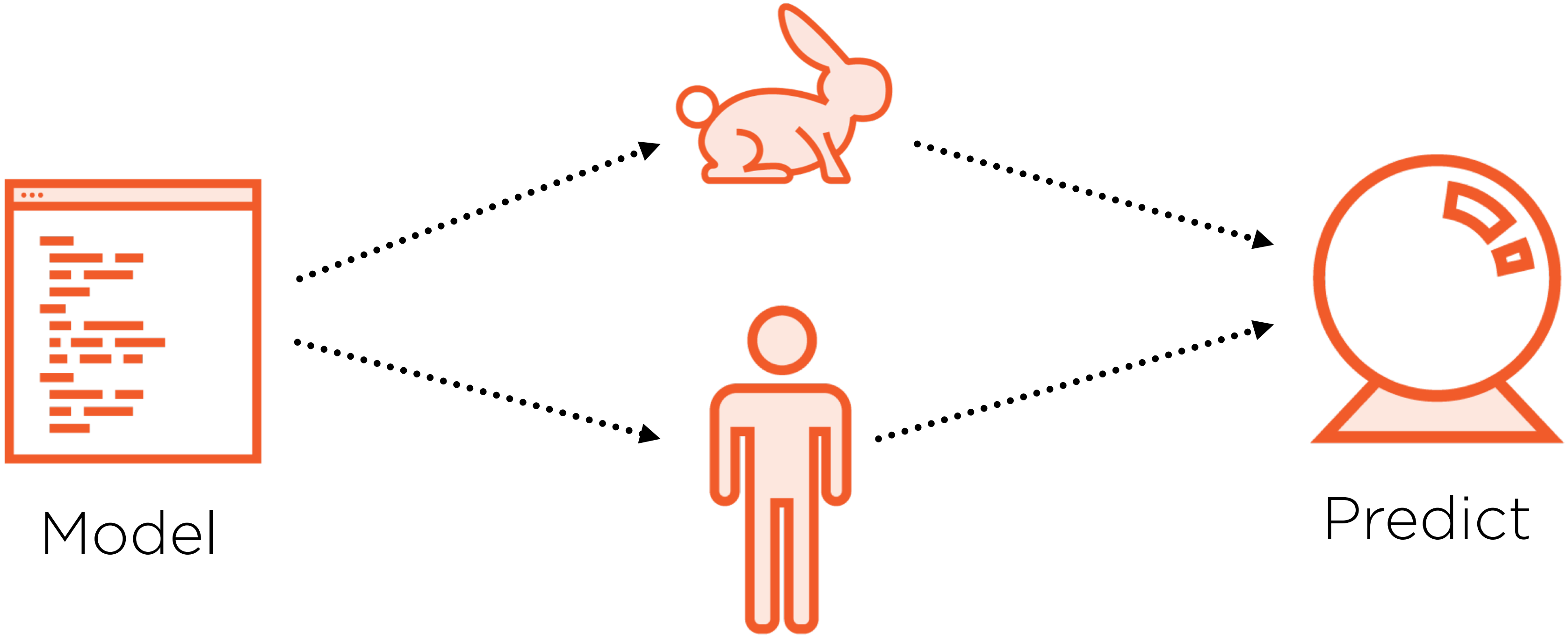
Model

Predict

	0	1	2	3	4
0	1	Braund, Owen	male	52	0
1	2	Cumings, John	male	38	0
2	3	Heikkinen, Laina	female	26	1
3	4	Futrelle, Jacques	male	35	1
4	5	Allen, Henry	male	35	0

	0	1	2	3
0	1	Braund, Owen	male	52
1	2	Cumings, John	male	38
2	3	Heikkinen, Laina	female	26
3	4	Futrelle, Jacques	male	35
4	5	Allen, Henry	male	35

Unsupervised Machine Learning



Model

Raw Images

Predict

The Machine Learning Process

Most data is sourced from relational databases. Data is also taken from big data platforms and comma separated files.

Data wrangling is time consuming and difficult. In Python, Pandas is the preferred tool.

Modeling involves feeding data to traditional algorithms and artificial neural networks.

Once the model has been tuned and tested it's ready for production. The models should perform well on data they've never seen.

Sourcing Our Data

Databases

Most models are currently sourced from relational data

Flat Files

Most open source data are comma separated files

Big Data

Approximately 90% of data collected globally is unstructured

The Core Algorithms

Traditional Models

All models that are not artificial neural networks

Neural Networks

Models that are artificial neural networks

Final Phase of Machine Learning

Completed Model

The model has been trained on our dataset. We've also tuned the model for best performance.

Production

The model is live. The model should be able to make accurate predictions on data it has never seen.

Demo

Python 3.6

Jupyter Notebooks

Import Pandas library

Remove an attribute

View wrangled dataset