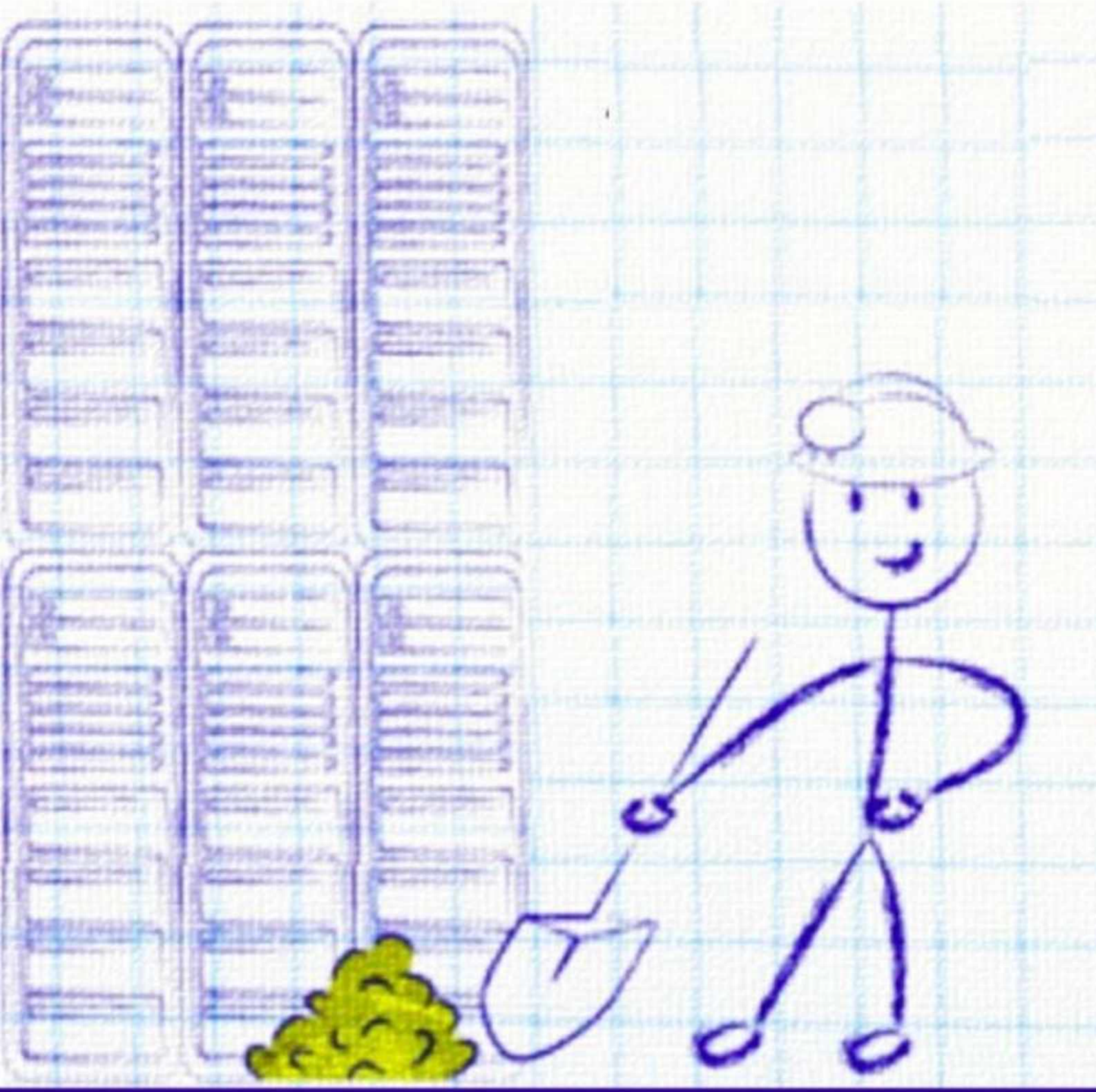
Machine Learning

For Absolute Beginners



Oliver Theobald

Machine Learning for Absolute Beginners

Oliver Theobald

First Edition

Contents Page

Introduction

It's a Friday night at home and you've just ordered a pizza from Joe's Pizzeria to be delivered to your house. The squeaky voice teen over the phone tells you that your pizza will arrive within 30 minutes.

But after hanging up the phone, you receive a message from your girlfriend (or boyfriend) asking if she/he can come over tonight.

Your girlfriend doesn't have a car, so you will have to drive over to her house and pick her up. While of course you want her to come over, you also don't want to wait until after the pizza has been delivered before you collect her - as the pizza will just sit there and get cold. You also don't want to pick her up after eating your pizza because then you'll miss the football game live on TV.

You need to make a quick decision. The first question you need to ask yourself, is do you have enough time to pick up your girlfriend before the pizza arrives?

Remember that the pizza is estimated to arrive within 30 minutes. If you leave now, you should be back within 30-40 minutes. As you know the route to your girlfriend's house, you can safely predict the journey time with a high degree of accuracy.

But just as you're about to walk out the door you realize there's another variable you haven't considered. You realize that what you also need to predict, in addition to the journey time to pick up your girlfriend, is the timing of the pizza being delivered. This too is something you have less control over.

Joe's Pizza is a popular pizzeria, and tonight also happens to be a Friday night. There's thus a range of factors that could affect your pizza delivery, including how many other people are ordering pizza, and the navigation ability of the delivery guy.

These two variables both have the potential to delay the delivery time of your pizza. However, this is your first time ordering a pizza on a Friday night. Perhaps unaware to you, Joe's Pizza has more delivery staff on call on Friday than say on a normal weeknight.

There are three potential methods to tackle this problem:

The first option is to apply existing knowledge. However you have no previous experience of ordering a pizza on a Friday night. Unfortunately there's also no app to calculate the average wait time on a Friday night for a pizza delivery in your area.

The second option is to ask someone else. You have exhausted this option already. The teenager on the other end of the phone at Joe's Pizzeria has already told you that your pizza will arrive "within 30 minutes".

The third option is to apply statistical modelling.

Given you've picked up this model on machine learning, let's go with the third option.

You think back to your previous experiences of ordering home delivery from Joe's Pizzeria. You then apply this information to predict the likelihood of the pizza arriving at your house on time. If the expected time of delivery exceeds 30 minutes then you can justify your decision to collect your girlfriend and return home in time for the delivery guy to arrive with your pizza.

Let's assume you have previously ordered pizza on 8 occasions, and the delivery time was late by greater than 10 minutes on four occasions. This means that the pizza arrived on time, or was early to arrive 50 percent of the time. This also means that there is roughly a 50% chance that the pizza delivery will be late again tonight.

Your mental decision-making progress is not comfortable with anything less than 70% (that the pizza delivery will be late). You thus remain at home to receive the pizza and make up an excuse not to see your girlfriend tonight.

Using existing data to base your decision is known as the empirical method. The concept of empirical data-backed decision-making is integral to what is known as machine learning.

Machine learning concentrates on prediction based on already known properties learned from the data.

In this example of the pizza delivery, we only considered the attribute of "frequency," the frequency of previous late deliveries. Machine learning models though consider at least two factors.

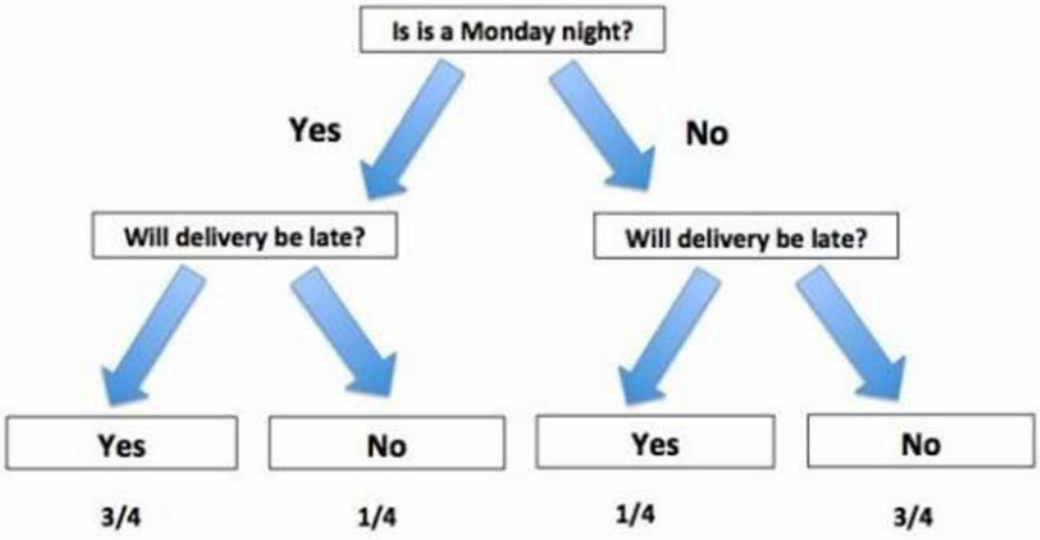
One factor is the result you wish to predict, known as the dependent variable. In this example, the dependent variable is whether the pizza delivery will be significantly late (more than 10 minutes). The second factor is the independent variable, which again predicts whether the pizza will be late but on a different independent variable. Day of the week, for example, could be an independent variable.

It could be a case that in the past, when the pizza was delivered on a Monday night the delivery time qualified as 'late'. This could be explained by the fact that Joe's Pizza has less delivery drivers on call on Monday nights.

Based on your previous experience, and not withstanding the three late deliveries that occurred on Monday night, pizza deliveries from Joe's Pizzeria typically arrive within the estimated time period.

This being the case, you could establish a model to simulate the probability that the pizza will arrive late based on whether or not it is a 'Monday night'.

A decision tree can be used to map out this particular example.



We now see that under this modelling there is only a 25% chance of the pizza delivery being late.

The process is relatively simple when considering a single independent variable. It does however become more complicated to calculate once a second or third independent variable are added to the equation.

Let's now add ‘rain’ as a third variable that could affect the pizza delivery time. A rainy night could of course slow down the delivery time due to safety precautions and extra traffic on the road.

This new variable is then added to the decision-making process. The new model now includes two independent variables in addition to one dependent variable.

We now need to predict the number of minutes the pizza will be late based on the level of rain (light = 2 minutes, moderate = 5 minutes, heavy = 15 minutes) and the day of the week. The predictions produced by this model will give us an idea on how late the pizza will be on any given day of the week. In this case though, a decision tree is of very little use as it can only predict discrete values (yes/no).

However, with the help of machine learning techniques you can apply the method of linear regression to predict the result.

It’s now time to sit down at your computer. For the sake of the story let’s forget the fact that your girlfriend is waiting for you to reply to her message.

Let’s also turn our attention to discuss machines learning.

For decades, machines operated on the basis of responding to user commands. In other words, the computer would perform a task as a result of the user directly entering a command.

But as you may know, that has all changed.

The manner in which computers are now able to mimic human thinking to process information is rapidly exceeding human capabilities in everything from chess to picking the winner of a song contest. This leads us into the realm of artificial intelligence and machine learning.

In the modern age of machine learning, computers do not strictly need to receive an ‘input command’ to perform a task, but rather ‘input data’. From the input of data they are able to form their own decisions and take actions virtually as a human would – but of course within the confines set by the machine’s operator.

In machine learning, a computer creates a model to analyze the scenario based

on existing data (experiences). The model in this case is predicting whether the pizza delivery will be late in future cases.

From here the computer treats the data very similar to normal human thinking. But given it is a machine, it can consider many more scenarios and execute far more complicated calculations to solve complex problems.

This is the element that excites data scientists and machine learning engineers the most. The ability to solve complex problems never before attempted. This is also perhaps one reason why you have picked up this book, to gain an introduction to machine learning, and techniques such as linear regression.

In the following sections we will first dive in and consider machine learning from an aerial view and discern the relationship between our topic and the larger field of data science.

Overview of Data Science