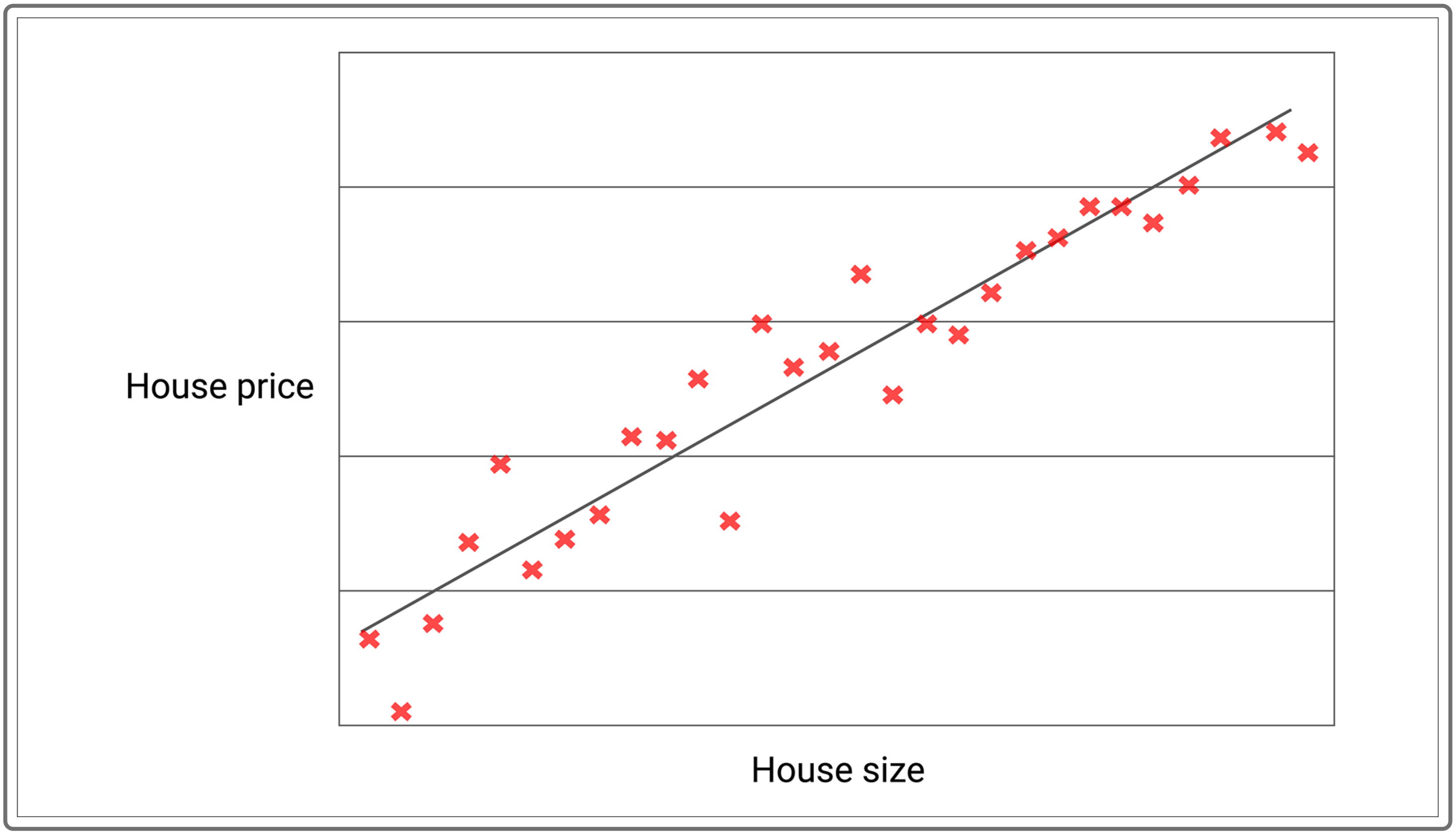
17.3.3

How Logistic Regression Works

Jill informs you that a good data scientist not only understands the hows but the whys. She explains that understanding how a model works helps a data scientist assess a machine learning model's strengths, weaknesses, and how best to use it. She asks you to look into how a logistic regression model works.

When you protest that you haven't taken a math class in years, she reassures you that while math is indeed helpful to know, many basic underlying ideas in machine learning can be grasped without a graduate degree in math.

Before launching into a discussion of logistic regression, let's quickly review linear regression. The image below shows a scatter plot, through which a best fit line is drawn. In this case, the chart depicts the size and price of houses in a particular district:



As the size of a house increases, it will generally fetch a higher price on the market. Each point represents a specific house in that district. How would you best describe the data values in the graph?

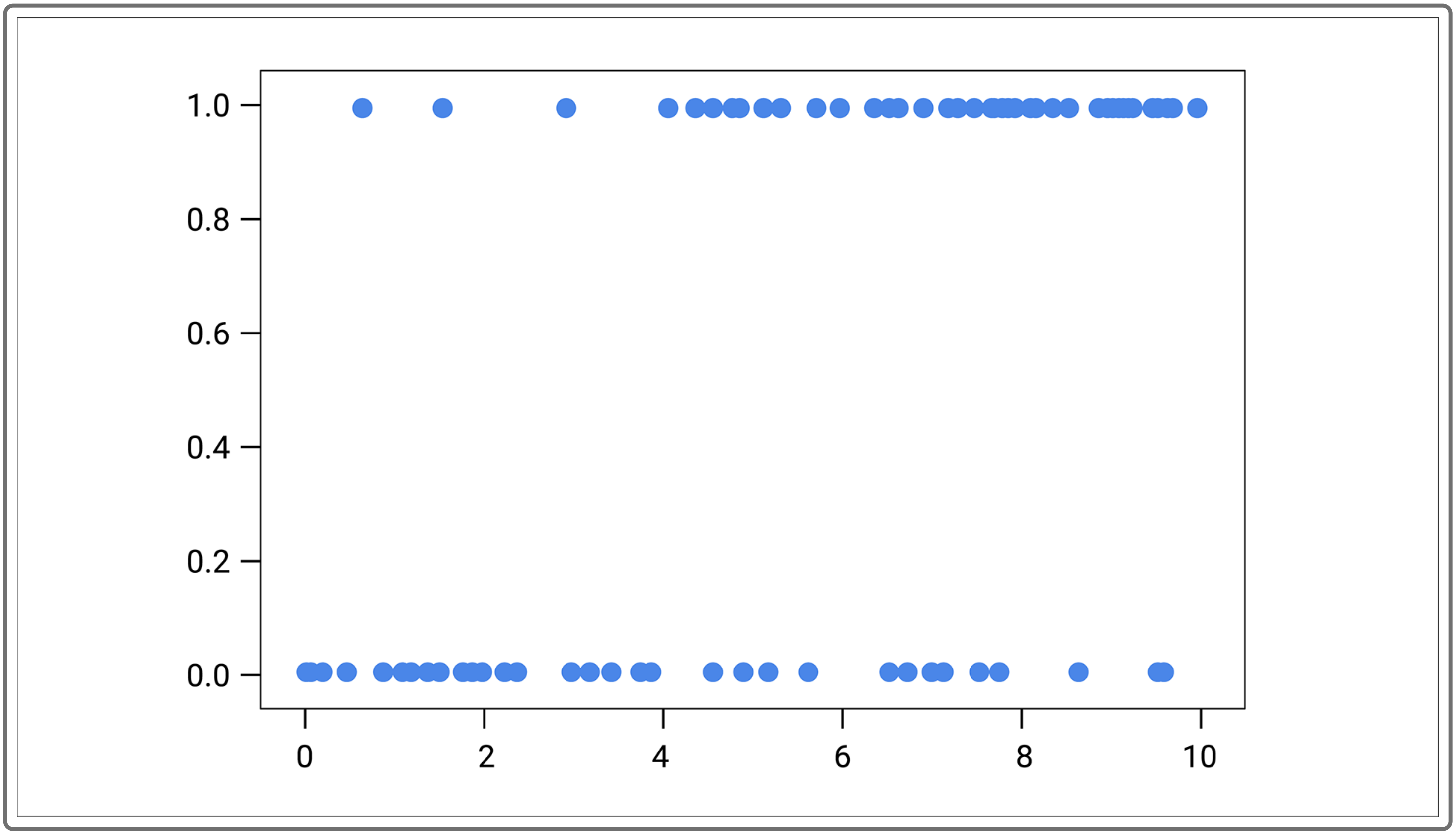
○ The data values are continuous.

○ The data values are discrete.

Check Answer

Finish ▶

But what happens when the outcome variable is binary, meaning that only two outcomes are possible? For example, let's say that the x-axis on the scatter plot below represents a college applicant's score on an entrance exam, and that the y-axis represents acceptance to a particular college. There is a wide range of test scores, but there are only two possible outcomes: acceptance (1) or rejection (0):

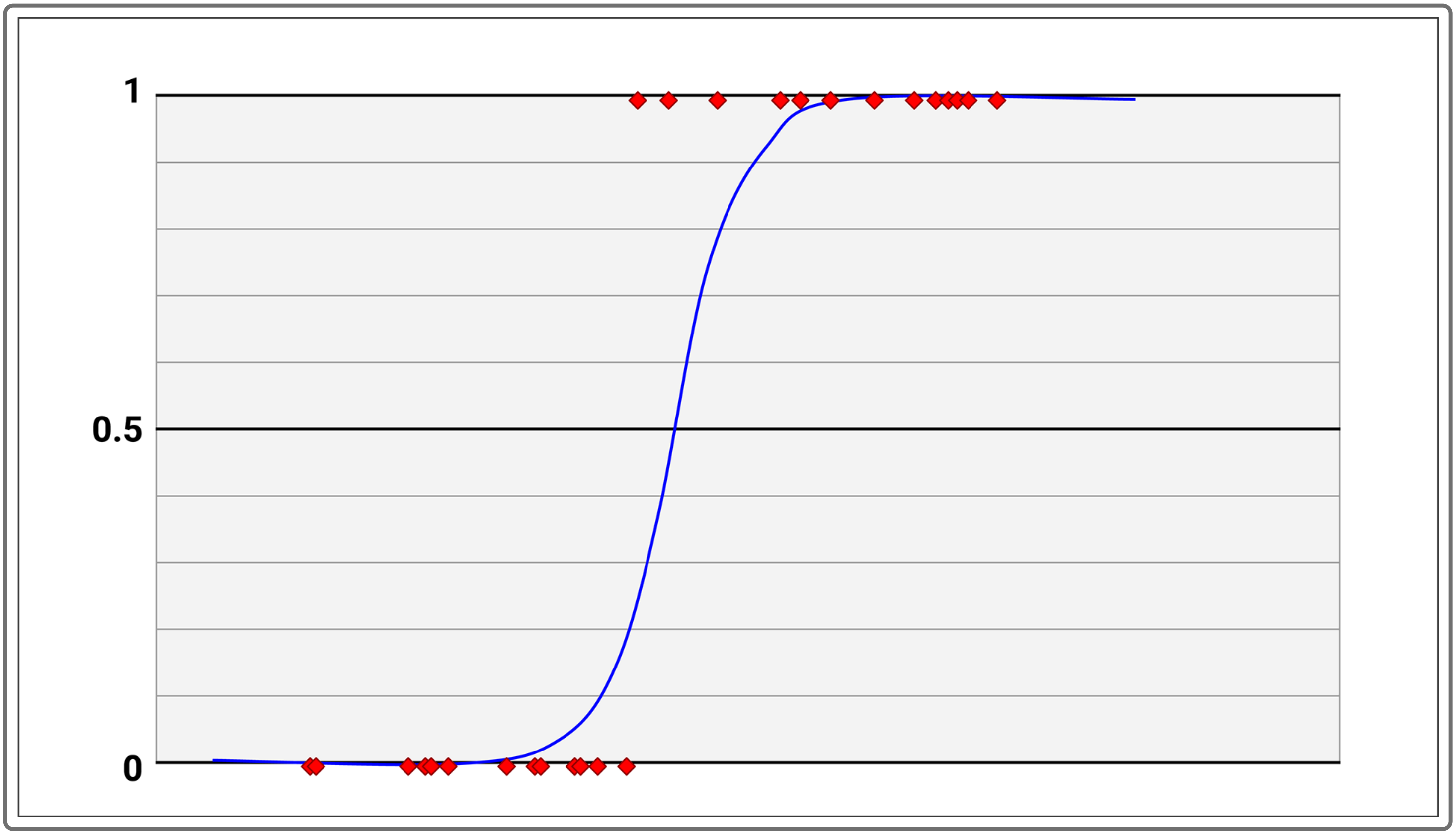


Linear regression clearly would not work in this case. Try drawing a best fit line here! A best fit line drawn through this scatter plot would be neither descriptive nor meaningful.

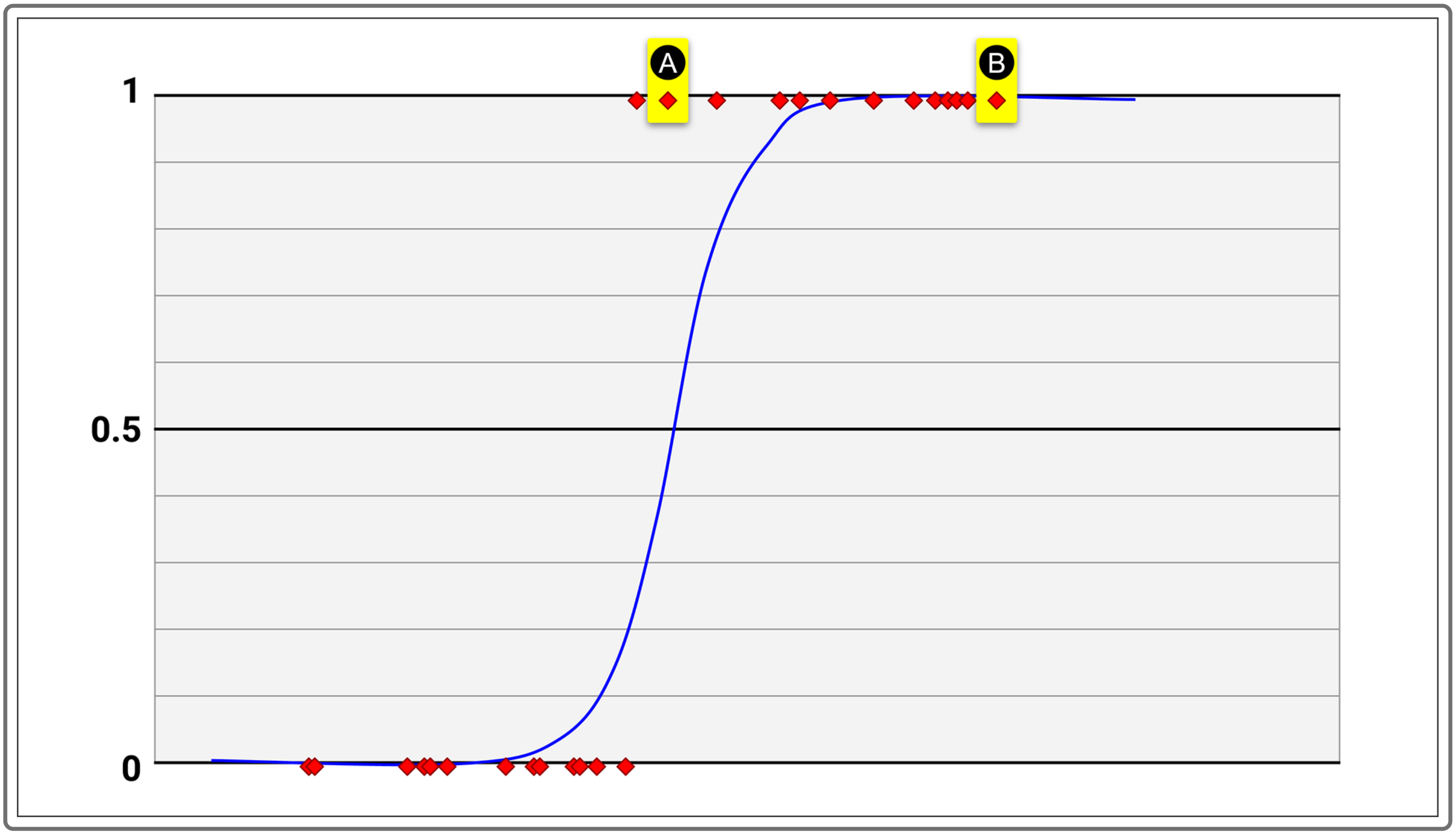
Instead, the probability of an outcome is represented with the following equation:

```
log(probability of admission/(1 - probability of admission))
```

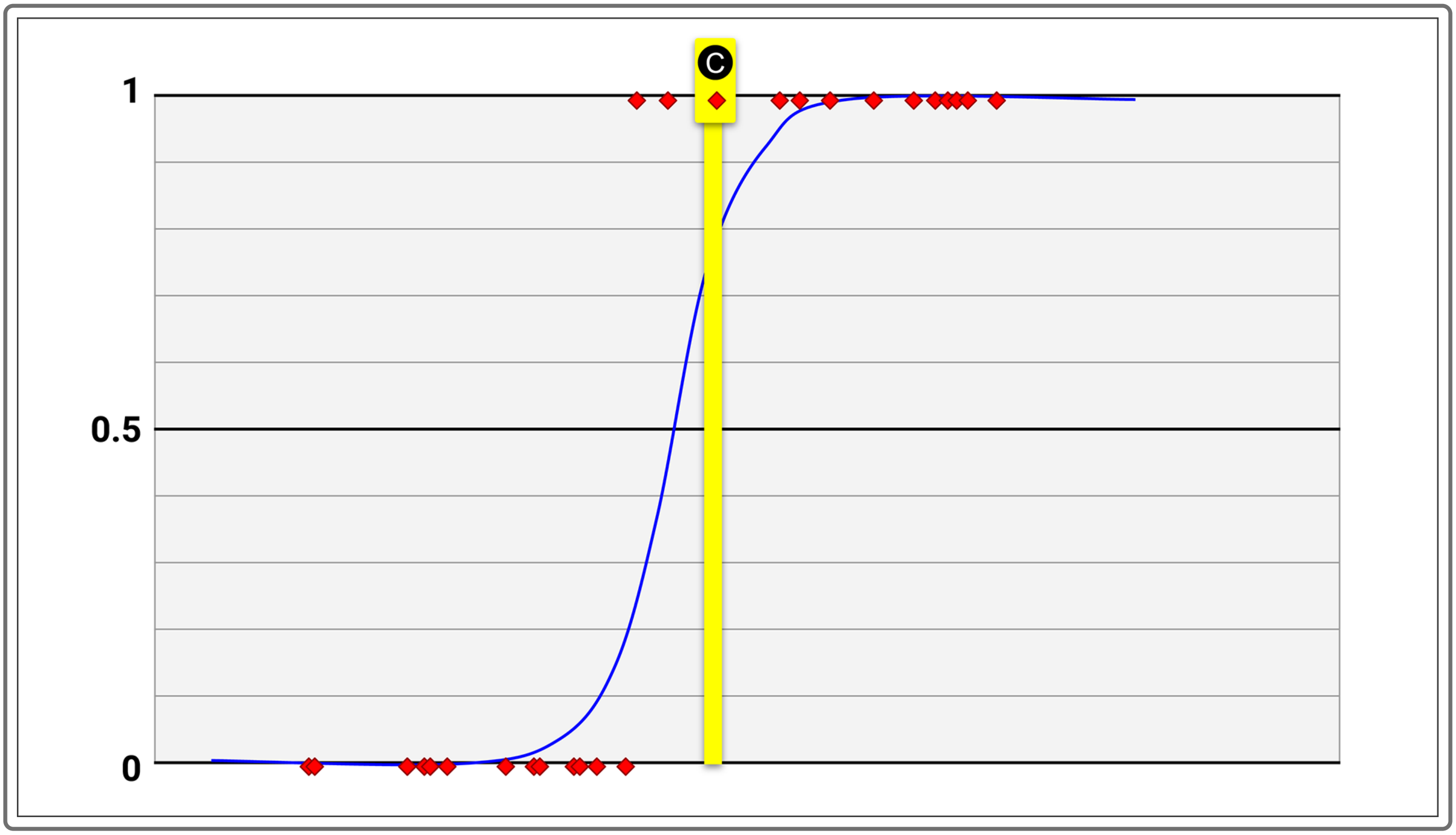
It's important to note that the results of this equation ultimately generate an S-shaped curve that represents the probability of being admitted at a given test score:



Each test score along the x-axis is associated with a probability of acceptance. In the following example, for a student with a score of A, the probability of acceptance is somewhat higher than 50%, whereas a student with a score of B has a nearly 100% chance of being admitted:



This **S-shaped curve,** also called a sigmoid curve, can then be used to predict acceptance for new applicants. The score at which the vertical line is drawn has approximately 80% probability of acceptance. Because this value exceeds the cutoff point, which in this case is defined as 50%, it's predicted that candidates with this score will be accepted:



The company you are working for wants to filter spam emails based on certain criteria. What type of regression analysis is this?

○ Linear

○ Logistic

Check Answer

What type of variable would describe whether an email is spam or not?

○ Continuous

○ Binary

Check Answer

Finish ▸

Finally, the results are made linear with a little more mathematical manipulation. The final product is a linear equation, like the one seen in linear regression. It is for this reason that both linear regression and logistic regression are considered linear models.

The Logit Function

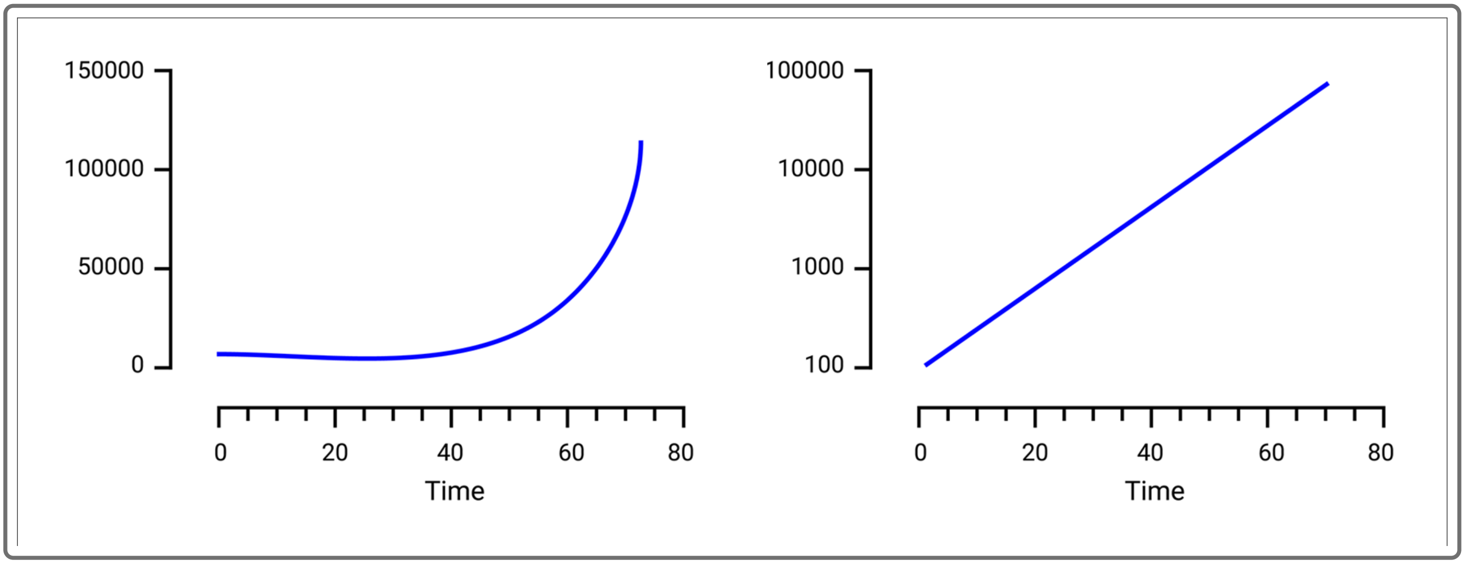
This brief mathematical discussion of logistic regression is optional. Feel free to skip this section, but keep reading if you'd like a bit more explanation behind the math.

We saw earlier that the sigmoid probability curve is generated based on the following equation, which is also called the logit function:

```
log(probability of admission/(1 - probability of admission))
```

The fraction seen here is the ratio of the probability of an occurrence and nonoccurrence. For example, let's say that applicants with a given score have a 90% probability of acceptance. Therefore, they will have a 10% probability of rejection. The logarithm of the ratio of the two probabilities is expressed as log(0.9/0.1). Based on this equation (which undergoes some rearrangement), the S-shaped curve can be created from the existing data points.

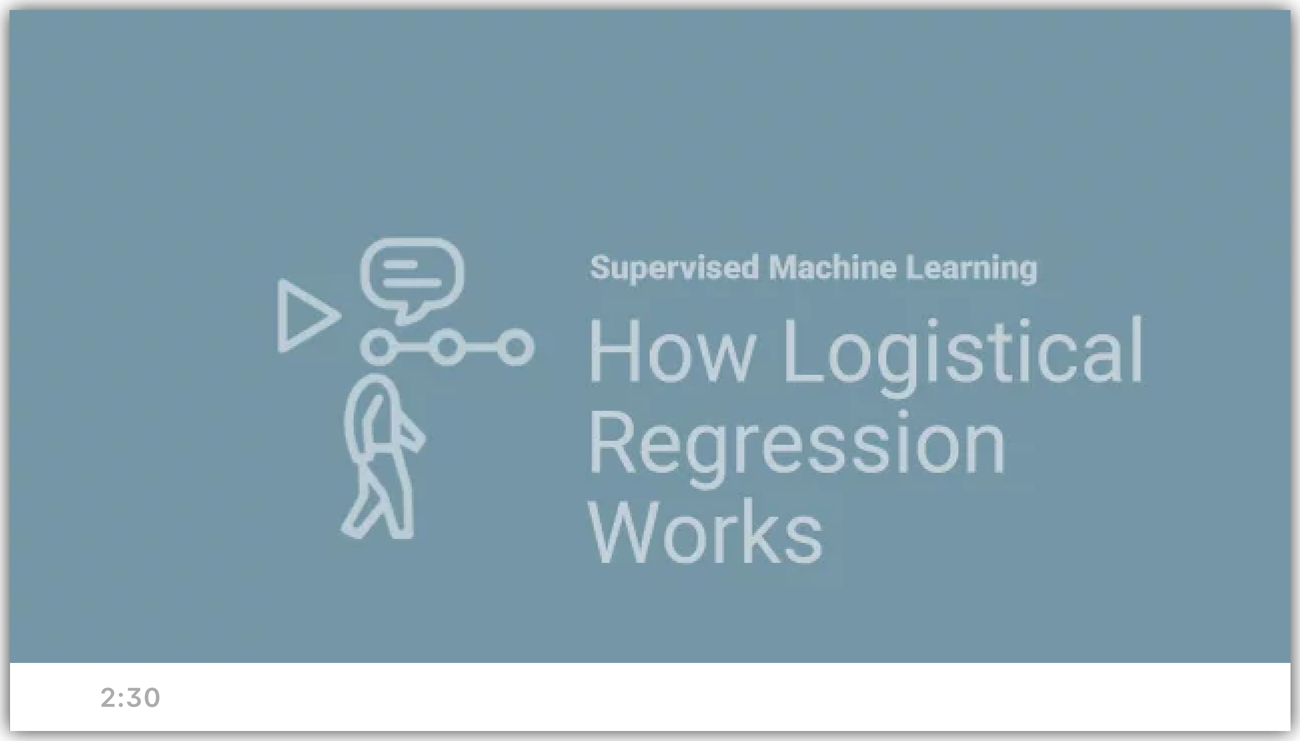
So why is a logarithm of the ratio obtained? Logarithms are useful when dealing with ratios. If you are unfamiliar or rusty with logarithms, a short description of them is that they are the opposite of exponents, just as subtractions undo additions, and divisions undo multiplications. In the following illustration, to the left is an exponential curve. In this case, as the value on the x-axis increases, its y-axis value increases rapidly. The illustration on the right shows that the curve is been straightened into a line after plotting the logarithms of the values, since logarithms undo exponents:



To understand why logarithms might be useful, consider two ratios: an even ratio and an extremely lopsided ratio. A score with an even chance (50%) of acceptance also has an even chance (50%) of being rejected. The ratio of the two probabilities is 0.5/0.5, or 1. On the other hand, an application that has 99.999% chance of being accepted has 0.001% chance of being rejected. The ratio of the two probabilities is 0.99999/0.00001 or 99999. The discrepancy between the two ratios is almost 100,000-fold! The use of logarithms smoothens out this asymmetry by scaling the numbers.

NOTE

To learn more, consult online resources such as Khan Academy's **introduction to logarithms (https://www.khanacademy.org/math/algebra2/x2ec2f6f830c9fb89:logs/x2ec2f6f830c9fb89:log-intro/v/logarithms)** .



Supervised Machine Learning

How Logistical Regression Works