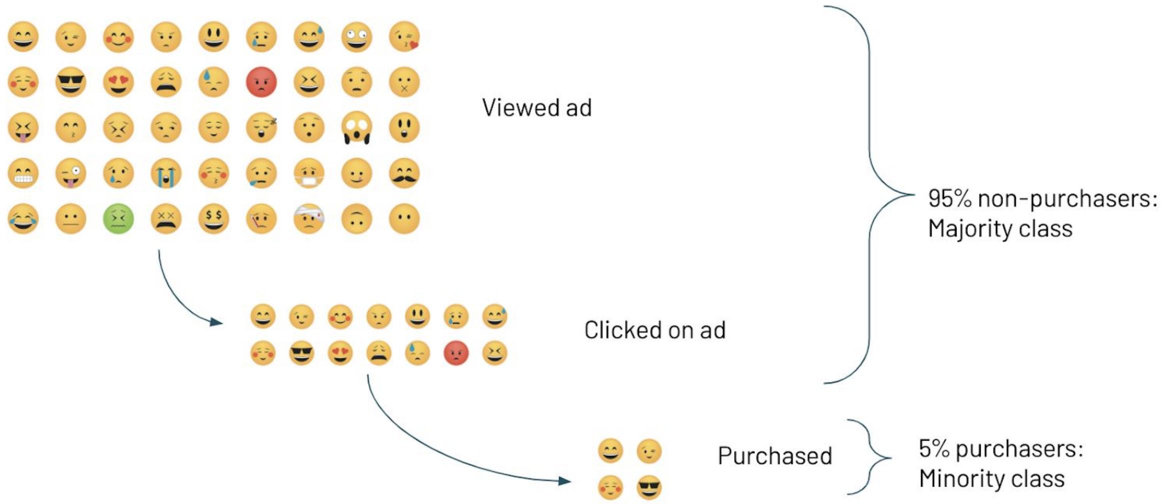
Problems with Class Imbalance

Estimated time: 5 minutes

Often, a dataset will have many more cases of one class label than the other. This is known as class imbalance, and is very common. An example of this is customers who complete a sale after viewing a social media ad. Most of the people who view the ad don't click on the link, and of the few who click, most don't end up actually purchasing the product. Therefore, if you are trying to build a model to classify which viewers will purchase the product, you will have to address this class imbalance issue.

To illustrate why this is a problem, imagine that 95% of viewers don't purchase the product. If you built a model that classifies viewers into non purchasers and purchasers, and got 95% accuracy, you and your boss might think that sounds like a good model. But as a savvy data scientist, you know that if you didn't even build a model and just predicted every viewer as a non purchaser, you would be accurate 95% of the time! Therefore your 95% accuracy just went from good to garbage.



Accuracy is clearly not an adequate metric to evaluate a model with an imbalanced data set. In fact, precision and recall give us a better understanding of how a model performs on an imbalance data set. Looking at the confusion matrix can also help to understand the model's predictions.

When you have a very imbalanced dataset, you may have very few samples of the minority class. This makes it hard for algorithms to learn patterns, or recognize and correctly classify new samples. It simply hasn't seen enough of the minority class to understand when to predict it. Fortunately, there are techniques that address the label imbalance problem. We will learn about these in the next videos.