Outlier Detection: Through the Eyes of our Working Example

OUR STORY

Visualizations can help reveal the presence of outliers. Some outliers are representative of natural variations in the data, while others are the result of mislabeling or another unintended error. When they have been identified it is important to investigate their origin. This will generally require talking with the individuals that are close to the data production and data gathering. There is a good chance that it will require some understanding of the nature of the data itself.

THE DESIGN THINKING PROCESS

Understanding the nature of outliers will require extensive work with not only the data itself but also the experts in the enterprise who create and work with those data.

The design thinking exercise called *Data Understanding* is designed to allow a team to do a deep dive into each potential data source that might be used for machine learning models. As a data scientist your role during this exercise would be to identify all of the complications associated with using a particular data source, including the presence of outliers that could skew the model's results. It will be up to you to ask the domain experts about what sort of outliers might appear in the data, as well as how common they are.

If you've identified outliers in your data, it will be your responsibility to raise them up as an issue during your team's playback meetings. At that point you'll be ready to meet with the domain experts who can tell you if those outliers will be a problem for you as you build your models.