# Outliers

There are times when the class imbalance is so extreme that we should consider outlier detection algorithms in place of more traditional supervised learning techniques. Visualization is often used to help determine which algorithm(s) makes the most sense for the data. For high-dimensional data it is sometimes necessary to pipe the data through a dimension reduction algorithm before applying the outlier detection algorithm. In addition to how you scale the data, there is the choice of dimension reduction algorithm and the choice of outlier detection algorithms. Moreover, there are generally parameters that modify these outlier detection algorithms, like an assumed level of contamination. Given the number of tuneable components it can take some time through grid-searching and iteratively comparing pipelines to settle on an outlier detection pipeline that is optimized for your data and business opportunity. Outlier detection can also be included as a step in the overall modeling pipeline which can improve performance of the core prediction algorithm, but the outlier observations should still be accounted for in some way.

***Important:*** Outlier detection algorithms can be useful in the case of extreme imbalance among classes. After all, they are intended for use in situations where one class completely overwhelms the other. However, there is no set rule about a given proportion that acts as a switch-point from re-sampling techniques to outlier detection. Iterative comparison based on model performance is generally accepted as the way of choosing a direction.

## Additional Resources

- Compare the effect of different scalers on data with outliers
- Outlier detection on the Boston housing data
- Customized sampler to implement an outlier rejections estimator