How to build a model that’s more robust to outliers?

This can be done either by getting the model more robust or the data more robust.

Getting model more robust:

1. It’s possible to choose tree-based models that are resistant to outliers.
2. Another alternative is to choose robust metrics like mean absolute error or something like Huber Loss instead of mean squared error.

Getting data more robust:

1. Winsorizing the data: The extreme values found above 95% can be replaced by the maximum value in the 95% zone and the ones found below 5% can be replaced by the minimum value in the 5% zone instead of dropping them.
2. Log transforming the extreme values in case of highly right skewed or exponential data will reduce the variance due to the extreme values.
3. If one is very clear about outliers, that are not worth predicting, then those values can be removed.

Simple methods to detect outliers:

* Visualizations like Box plot
* Cook’s Distance – only usable in regression situation
* Z-score: to detect extreme data points beyond threshold value

Some of the advanced methods to detect outliers are:

* PCA
* LOF (Local Outlier Factor)
* HiCS: High Contrast Subspaces for Density-Based Outlier Ranking.