Model Tuning

Training Data Tuning

Training set too small?
- Sample and label more data if possible

Training set biased against or missing some important scenarios?
- Sample and label more data for those scenarios if possible

Can't easily sample or label more?
- Consider creating synthetic data (duplication or techniques like SMOTE)

Important: Training data doesn't need to be exactly representative, but your test set does.

← Training: over or under sample as needed

SMOTE Synthetic minority over sampling technique

Feature Set Tuning

- Add features that help capture pattern for classes of error
- Try different transformations of the same feature
- Apply dimensionality reduction to reduce the impact of weak features

→ Also interaction features, for ex the product of two features

But... avoid overly complicated models which fit both the signal and the noise (overfitted)

Dimensionality Reduction

Motivation: Cause of overfitting: too many features for the amount of data
- Exacerbated if there are noisy / irrelevant features
- Also a problem for the curse of dimensionality

(Hard to group sparse points in high-dim space)

Dimensionality reduction: Reduce the (effective) dimension of the data with minimal loss of info