ML cheat sheet

A model that exhibits small variance and high bias will underfit the target, while a model with high variance and little bias will overfit the target. A model with high variance may represent the data set accurately but could lead to overfitting to noisy or otherwise unrepresentative training data.

Data scientists must do this while keeping underfitting and overfitting open\_in\_new in mind. A model that exhibits small variance and high bias will underfit the target, while a model with high variance and little bias will overfit the target.

Variance comes from highly complex models with a large number of features. Models with high bias will have low variance. Models with high variance will have a low bias.

In statistics, variance **measures variability from the average or mean**.

Model **bias** refers to the presence of systematic errors in a model that can cause it to consistently make incorrect predictions. These errors can arise from many sources, including the selection of the training data, the choice of features used to build the model, or the algorithm used to train the model.

Statistical **bias** is a term used to describe statistics that don't provide an accurate representation of the population. Some data is flawed because the sample of people it surveys doesn't accurately represent the population.

Great link to visit for detailed understanding of bias and variance and trade-off between them: <https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229>

Other reasons for overfitting

- the data used for training is not cleaned contains noise (garbage values) in it

- size of the training data is not enough

- the model is too complex

Techniques to overcome with/handle imbalanced datasets:

<https://www.analyticsvidhya.com/blog/2021/06/5-techniques-to-handle-imbalanced-data-for-a-classification-problem/>

understanding ROC:

<https://towardsdatascience.com/understanding-the-roc-curve-in-three-visual-steps-795b1399481c#:~:text=To%20plot%20the%20ROC%20curve,That's%20it>!

Feature engineering (feature creation, extraction, etc.) in ML

<https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10#:~:text=Feature%20Extraction%3A%20Feature%20extraction%20is,quantities%20for%20algorithms%20to%20process>.

Feature extraction from text data (transform text data in numeric, so that machine can understand.

<https://www.analyticsvidhya.com/blog/2022/05/a-complete-guide-on-feature-extraction-techniques/#:~:text=Feature%20Extraction%20is%20also%20called,time%20let's%20start%20our%20article>.

Difference between feature selection and feature extraction, clearly explained.

<https://vitalflux.com/machine-learning-feature-selection-feature-extraction/?utm_content=cmp-true>

Features selection

<https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/>

Encoding is converting categorical and Boolean data into numeric for processing. Imputing is the process of filling NaN (missing/null values)

<https://towardsdatascience.com/preprocessing-encode-and-knn-impute-all-categorical-features-fast-b05f50b4dfaa>