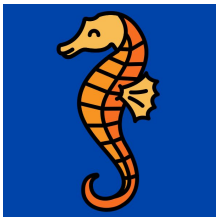


Error Analysis Checklist

Error Analysis follows the same concepts as EDA. There are some standard steps that one does before modeling and after modeling.

Before modeling

Before the modeling starts one wants to make sure that their train and validation and test datasets are following the iid principles.

- make sure that your target variable has similar distributions in all dataset splits
    - you need to find out what feature(s) is influencing this.. and do stratified sampling
- check the distribution of all the different features with the target variable in train and test
    - → if the features are important for the model and have differences.. then go for stratified sampling
- plot also categorical data distribution between train/test
- try PCA and see if you can see groups or not
- feature/target correlation → do not use features that are uncorrelated to target at all

After modeling

In this step your objective is to see if there are systematic errors, that is, are there groups of data for which your model predicts wrongly, under predicts, over predicts, or has large errors. The techniques go from high level to zooming in to particular data observations.

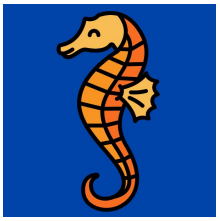
High level errors

Here you look at the error and variance, is your model overfitting or underfitting .. do you need regularization or do you need a more complex model

Data group level errors

Your errors can be due to particular features or for a group of features.

- Regression: plot residuals by predicted vs true
- Classification: confusion matrix for the errors: predicted vs true
- Is the model overfitting/underfitting for specific groups/classes?
- You can check again that the distribution of features for the wrongly predicted data points fits the train data distribution.. (subtract descriptive statistics) .. are there any trends? (features with large differences)
- For which group of data does your model perform better or worse?
    - check distribution of categorical variable (for the errors for example)
    - check bins of numerical variables



Error Analysis Checklist

- You can try PCA again on the errors

Zooming in
Here you want to look in detail at some of the examples that are misclassified or mispredicted. Pick 10-100 examples and look at the features.. You can here create a table and collect observations or ideas on why you think something was misclassified. These can in turn lead to better features.

Further reading:

Nice explanation
https://towardsdatascience.com/how-to-do-error-analysis-to-make-all-of-your-models-better-a13c4ca643a

Regression with mini error analysis
https://github.com/bot13956/ML_Model_for_Predicting_Ships_Crew_Size/blob/master/Ship_Crew_Size_ML_Model.ipynb

Classification with kernel of error analysis
https://www.kaggle.com/elitcohen/forest-cover-type-eda-modeling-error-analysis#Error-Analysis

Classification with error analysis and conclusion
https://www.kaggle.com/c/global-wheat-detection/discussion/157258
https://www.kaggle.com/pestipeti/error-analysis

Error analysis on text data
https://www.kaggle.com/jannen/model-error-analysis#Part-2.-Analysing-Classifier-errors---Manual-Error-Analysis