Evaluation Methods

Joshua Bernhard

Evaluation Methods

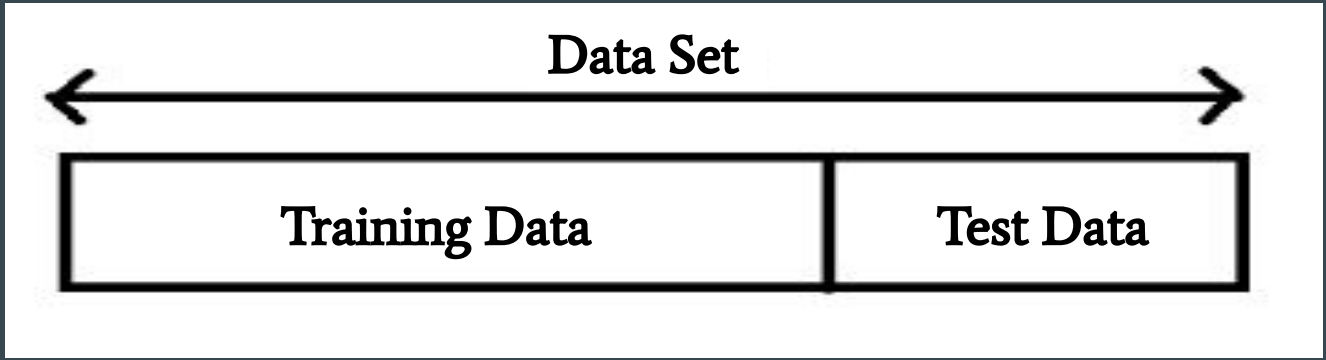
- Train/test split
- Classification Metrics
  - Precision
  - Recall
  - Accuracy
  - F1
  - AUC
- Regression Metrics
  - RMSE
- Cross Validation
- Offline vs. Online validation

Train/Test Split

Train/Test Split

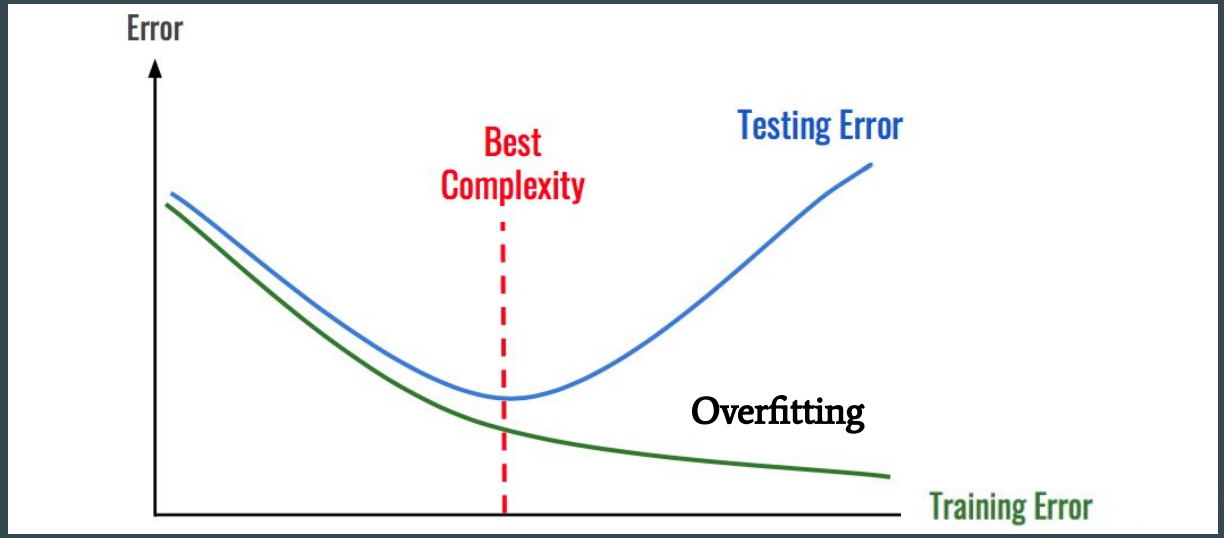
Measures of how well your recommender works on the training data (data it has already seen) won't tell you how well it will work once it is released into the real world.

One method used to better understand how well your recommender will work in the real world, is creating separate training and testing data.



Train/Test Split

Often when a recommender is created without validating on a test set, it will overfit.



The best model should have both low training and testing error.

Classification Metrics

Classification Metrics

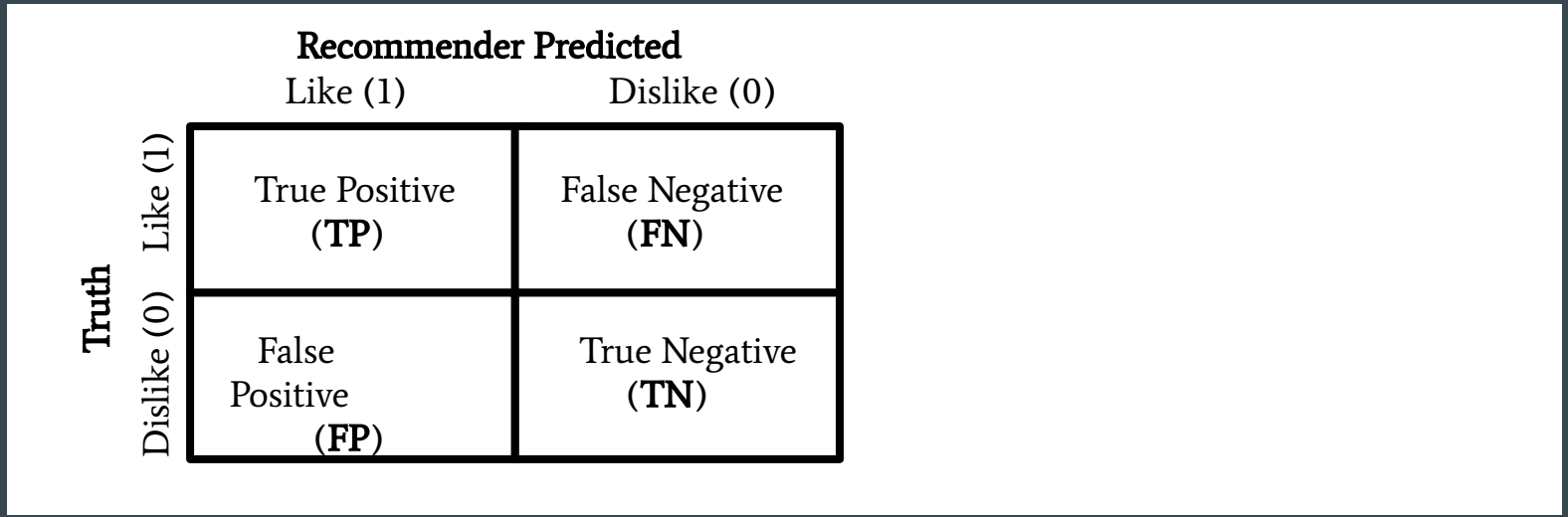
When your metric of interest is binary (click or not, watch or not, like or not, etc.), you will need to use classification metrics to measure the performance:

- Accuracy
- Precision
- Recall
- F1-Score

Understanding the context of your problem will assist in choosing an appropriate metric

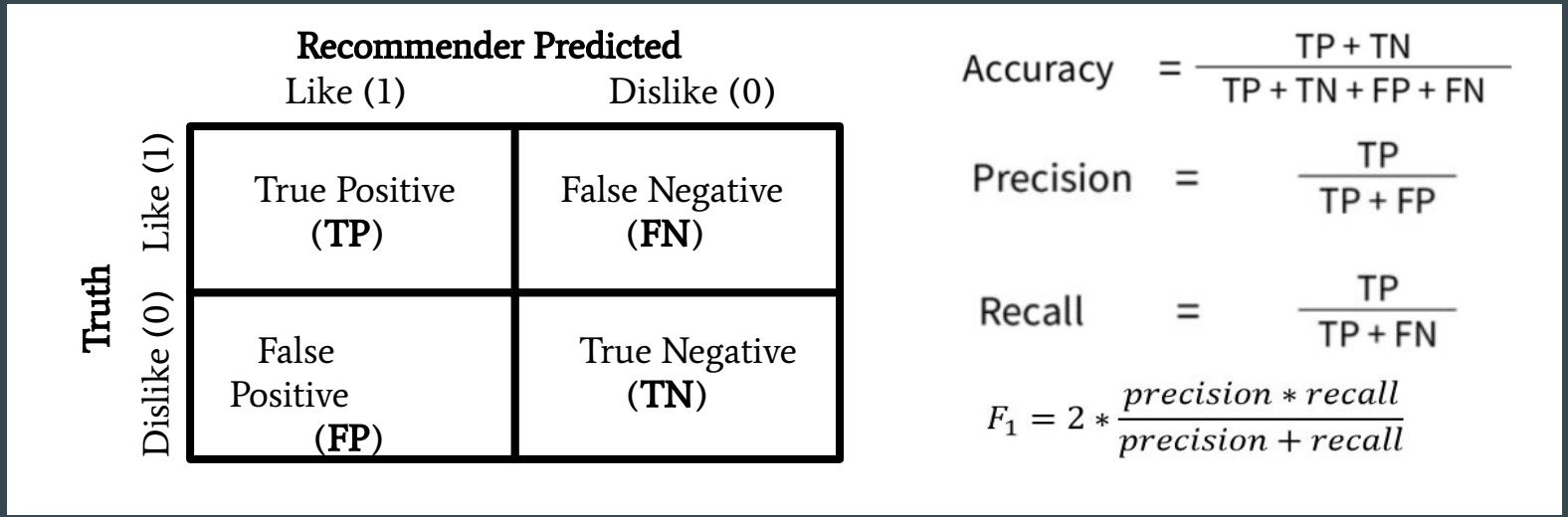
Classification Metrics

WIth any binary classification problem, there are four possible outcomes:



Classification Metrics

There are then a number of metrics you might choose to maximize

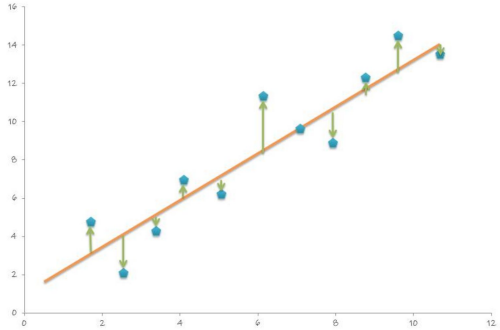


Regression Metrics

Regression Metrics

The most popular regression metric is RMSE

RMSE is a measure of how far on average predicted values are from actual values.



$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$\hat{y}_i$ = predicted recommendation value

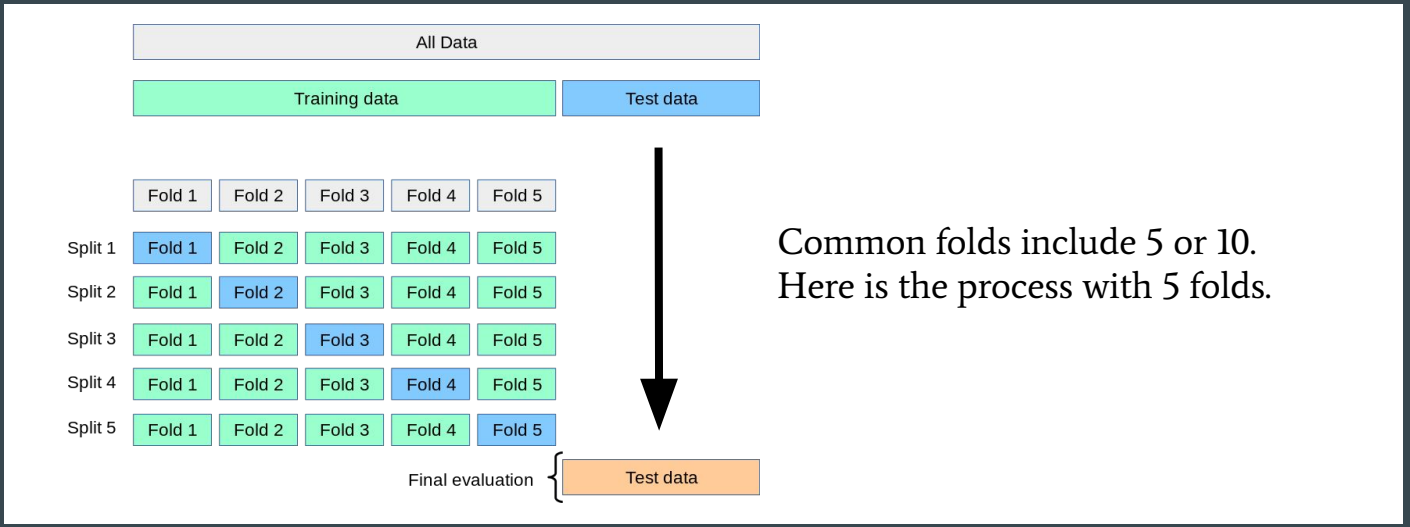
y_i = actual recommendation value

n = total number of recommendations

Cross Validation

Cross Validation

Another method for testing how well your model will perform rather than train-test splits is through multiple train test-splits. This method is known as **cross-validation**.



Offline vs. Online Validation

Offline vs. Online Validation

- Each of the techniques we looked at here are using “offline” validation, where we have all data available for both training and testing.

- Online validation uses live data. Even if a model works best by an offline evaluation, it is important that it works best once it is released in the real world.

- Online testing methods often involve A/B testing - an entire workshop could be devoted to this topic, and therefore, we won’t go into details here.

Recap

- Train-test datasets
- Classification Metrics
  - Why choose one over the other?
  - Class imbalance
- Regression Metrics
- Cross-Validation
- Offline vs. online evaluation methods