- After model deliver, it's better to build a evaluation metrics for online data to **detect distribution drift**
- Training metrics and evaluation metrics can be different in practice, try to avoid this. It is better to train the model to directly optimize for the metric it will be evaluated on.
- Data Skew: when 1 kind of data is much more smaller than others, such as data imbalance, outliers. Robust metrics may not be affected by outlier, but they will still influence the training model. Careful data cleaning and reformulating the tasks may be needed

- Classification Metrics: accuracy, confusion matrix, pre-class accuracy, logloss, AUC/ROC
- **Ranking Metrics**: precision, recall, precision-recall curve, NDCG
- **Regression Metrics**: RMSE/RMSD, MAPE
- **Model Validation**: hold-out, cross-validation, bootstrap
- **Hyper parameter Tuning**: grid search, random search, smart hyper parameter tuning
- **A/B Testing (online data evaluation method)**
- More about A/B Testing: http://www.evanmiller.org/

— confusion matrix - when the consequence of each class can be different
— pre-class accuracy - when classes can be imbalanced
— logloss - output is numerical probability not discrete classes
— AUC summarizes ROC from a curve to a number

— precision: predicted truly relevant/predicted to be relevant
— recall: predicted relevant/all real relevant
— k: number of answers returned by the ranker
— precision-recall curve: plot precision versus recall over a range of k values
— F1 score = 2*precision*recall/(precision+recall)

- When there are multiple queries, average precision and recall, check average precision@k, average recall@k
- F1 score is the harmonic mean of precision and recall, it tends toward the smaller of the 2 numbers

— NDCG (normalized discounted cumulative gain)

- CG (cumulative gain): sums up the relevance of top k items
- DCG, NDCG are important in information retrieval where the positioning of the returned items is important

— RMSE/RMSD: The square root of the average squared distance between the actual score and predicted score. Since it is an average score, it is sensitive to large outliers
— MAPE: median absolute percentage, and use it to compute 90th percentile of the absolute percent error

— Hold Out: use it only when the hold out subset is big enough to ensure reliable statistical estimates
— Leave one out is one of k-fold cross validation, but the k here equals to the number of data points. Use cross validation when the dataset is small
— Bootstrap: resamples with replacement, picks a data point uniformly at random

— Grid Search: a grid of hyper parameter values, evaluate each one of them and returns the winner
— Random Search: only evaluates a random sample of points on the grid, random search with 60 trails will find the region high probability
— Smart Hyper Parameter Tuning: pick a few hyper parameter settings, evaluate quality and decide where to sample next
— Nested parameter tuning: multiple models to be trained, for each model, turn the parameters, and for each set of parameters, do cross validation when necessary