Cross-Validation Evaluation

|  |  |  |
| --- | --- | --- |
| Type of error | Cause | Example |
| Bias | Model doesn’t capture the necessary variables | Linear Model Trying to model exponential growth |
| Variance | Over-fitting: the model responds to noise and outliers | Outlier that makes a linear model look exponential |
| Irreducible | There is randomness in the data |  |

Trade-off between bias and variance.

Training error

Feature importance analysis?

Do a better evaluation by doing multiple testing and get the average. K-fold cross-valuation to train all k portions – use all data

Automate searching for the best algorithm set by sklearn.grid\_search.GridSearchCV

Gasussian distribution

Easy way to mark categorical data

Label\_enc.fit

Fit\_transform

Turning categorical data into binary matrix ( set)

Fit\_transform

One Hot Encoder

Set ( finding

Ensemble models to predict a value ( example, predicting salary from phone case)

Try to reduce the number of features by setting max\_feature -> 2

n\_jobs=-1 and grid\_search this (

feature\_importance\_ within the decision tree model to show component

random forests to improve bag decision trees - best one so far

AdaBoost to improve

GradientBoostingRegressor to boost RandomFroestRegressor

Exercise to distinguish random data and real data