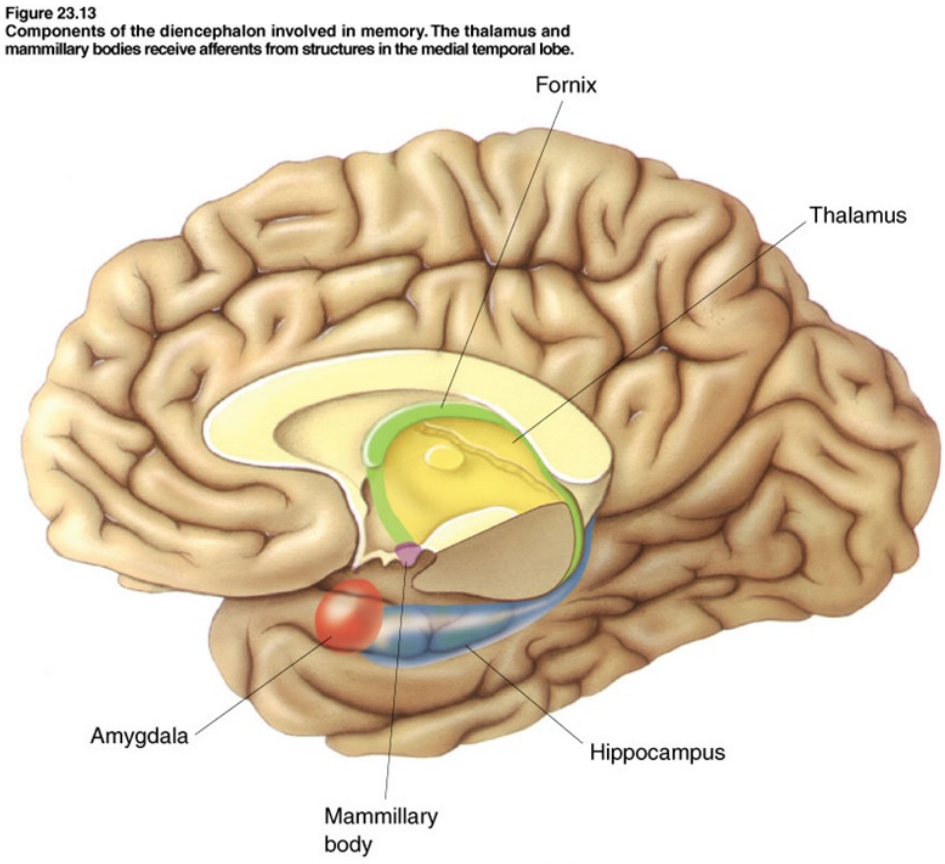
ML Modelling process

강사: 김남범(nbumkim@gmail.com)

머신러닝의 이란

학습이란?

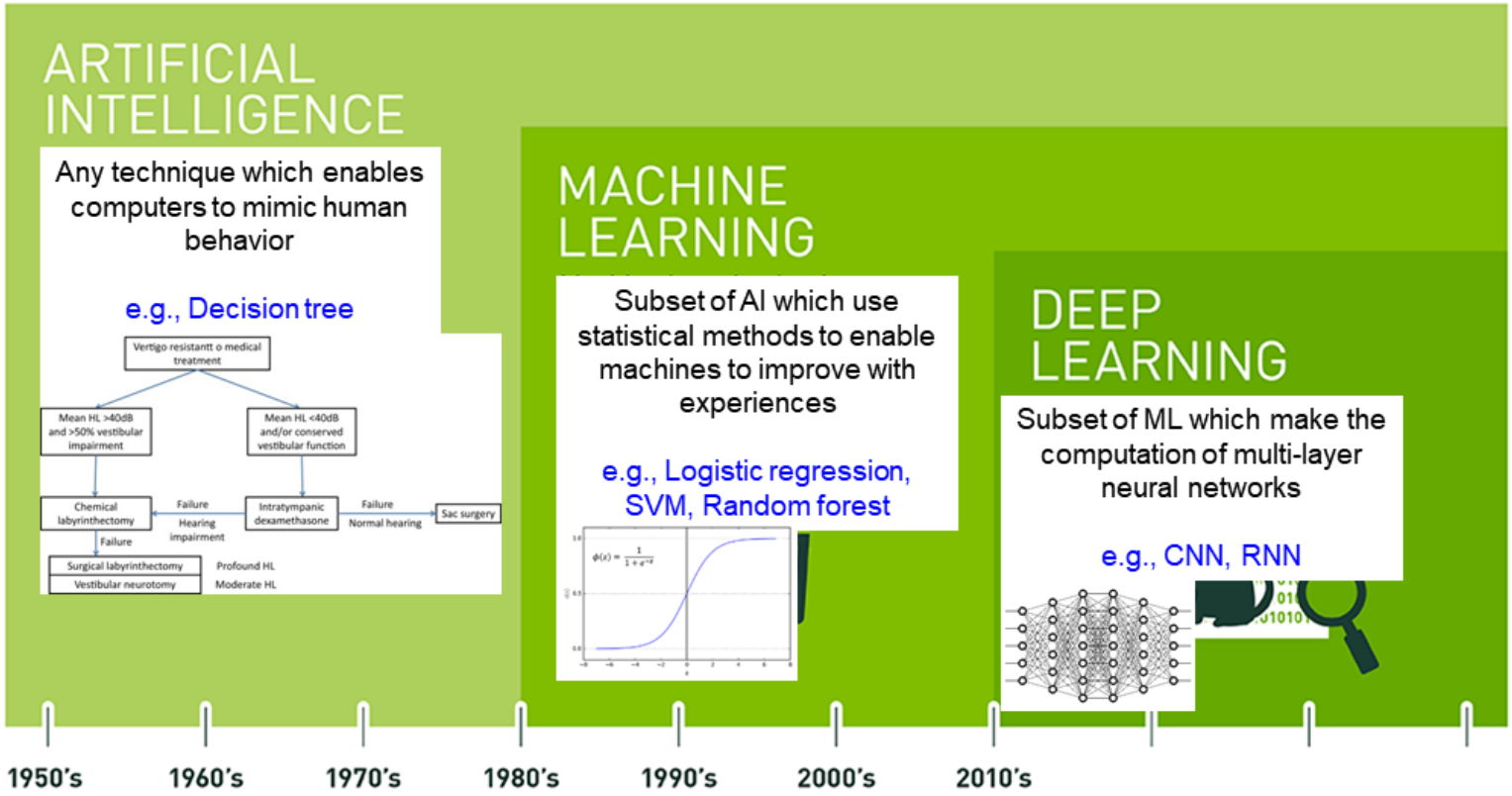
In human brain:



In machine:

패턴 인식

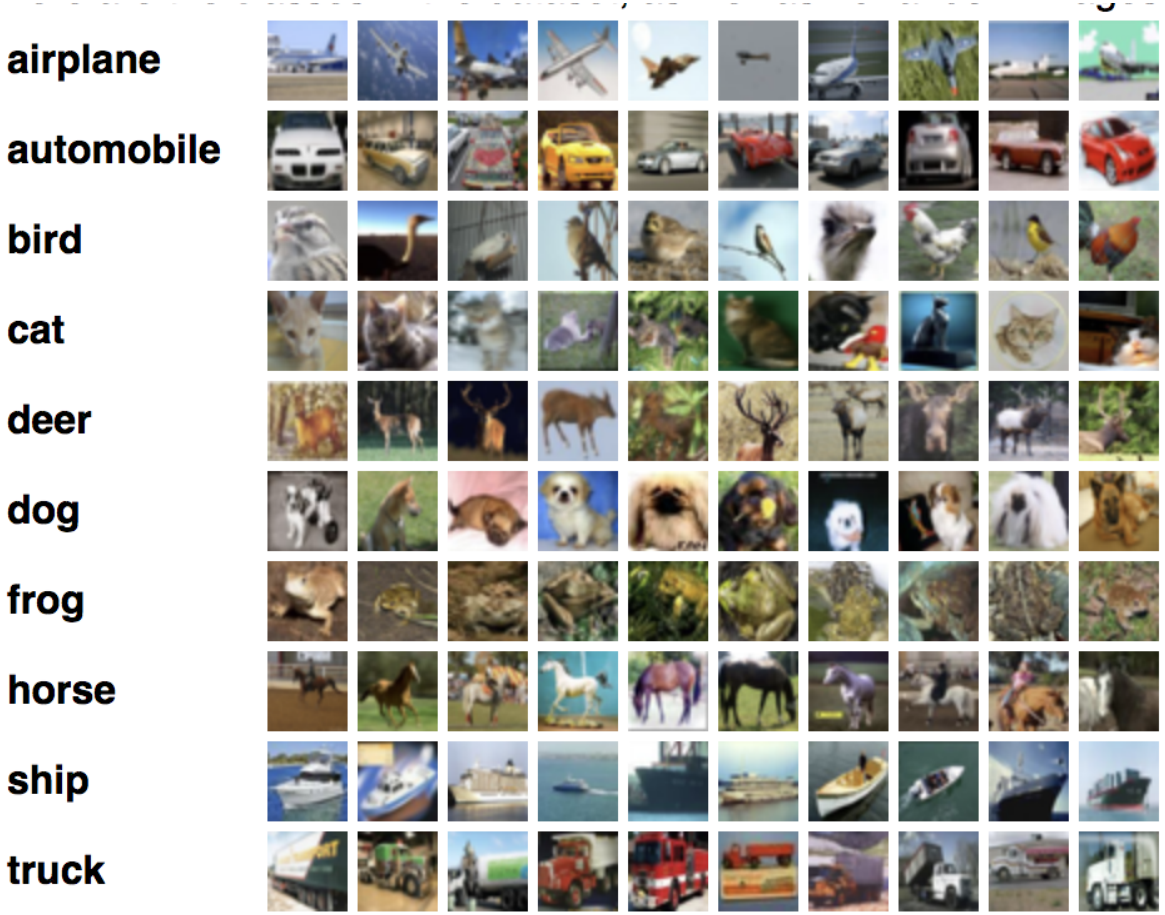
Difference between Artificial intelligence, Machine learning, and Deep learning



Since an early flush of optimism in the 1950s, smaller subsets of artificial intelligence – first machine learning, then deep learning, a subset of machine learning – have created ever larger disruptions.

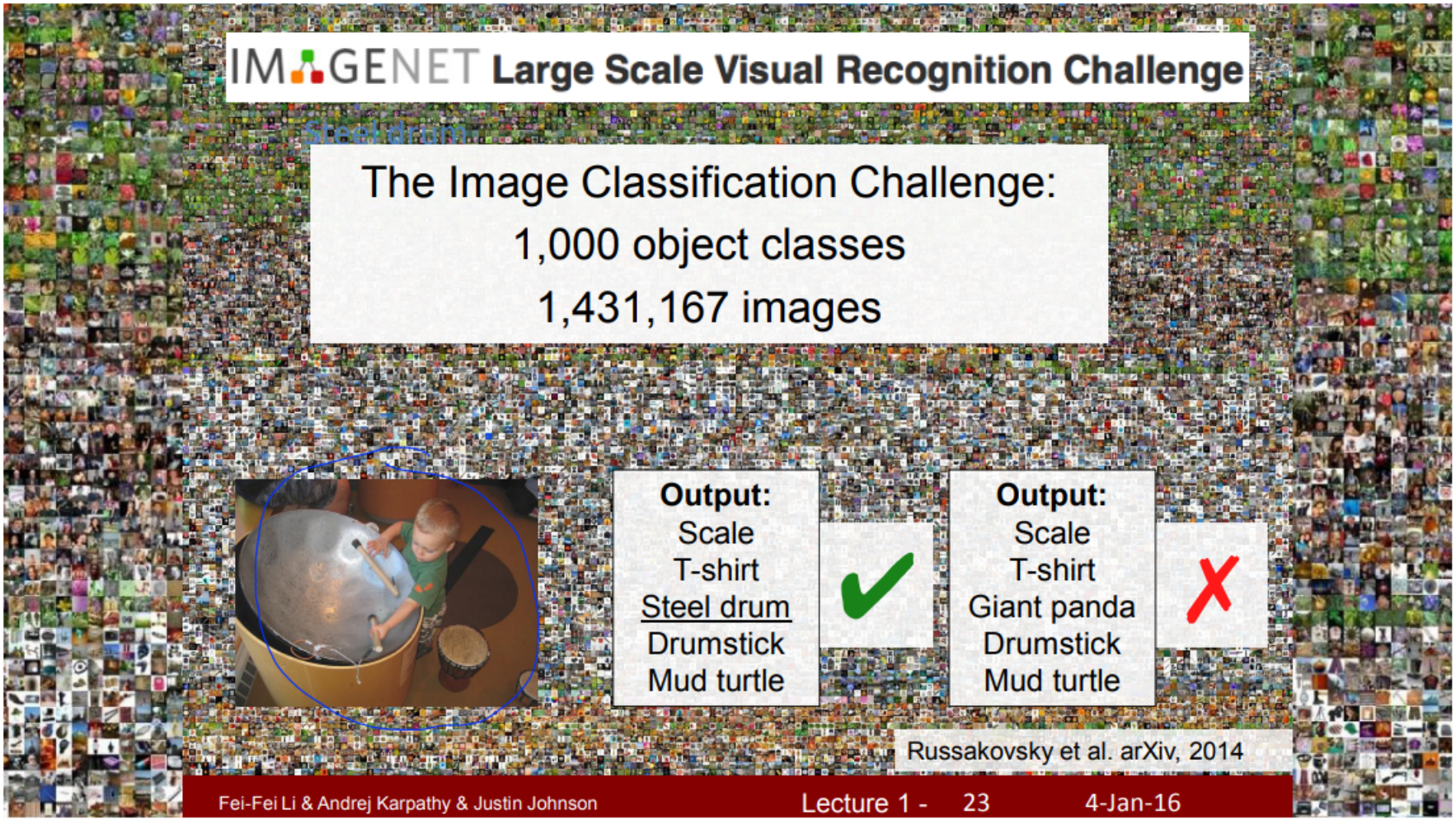
From NVIDIA, https://blogs.nvidia.com

Image classification



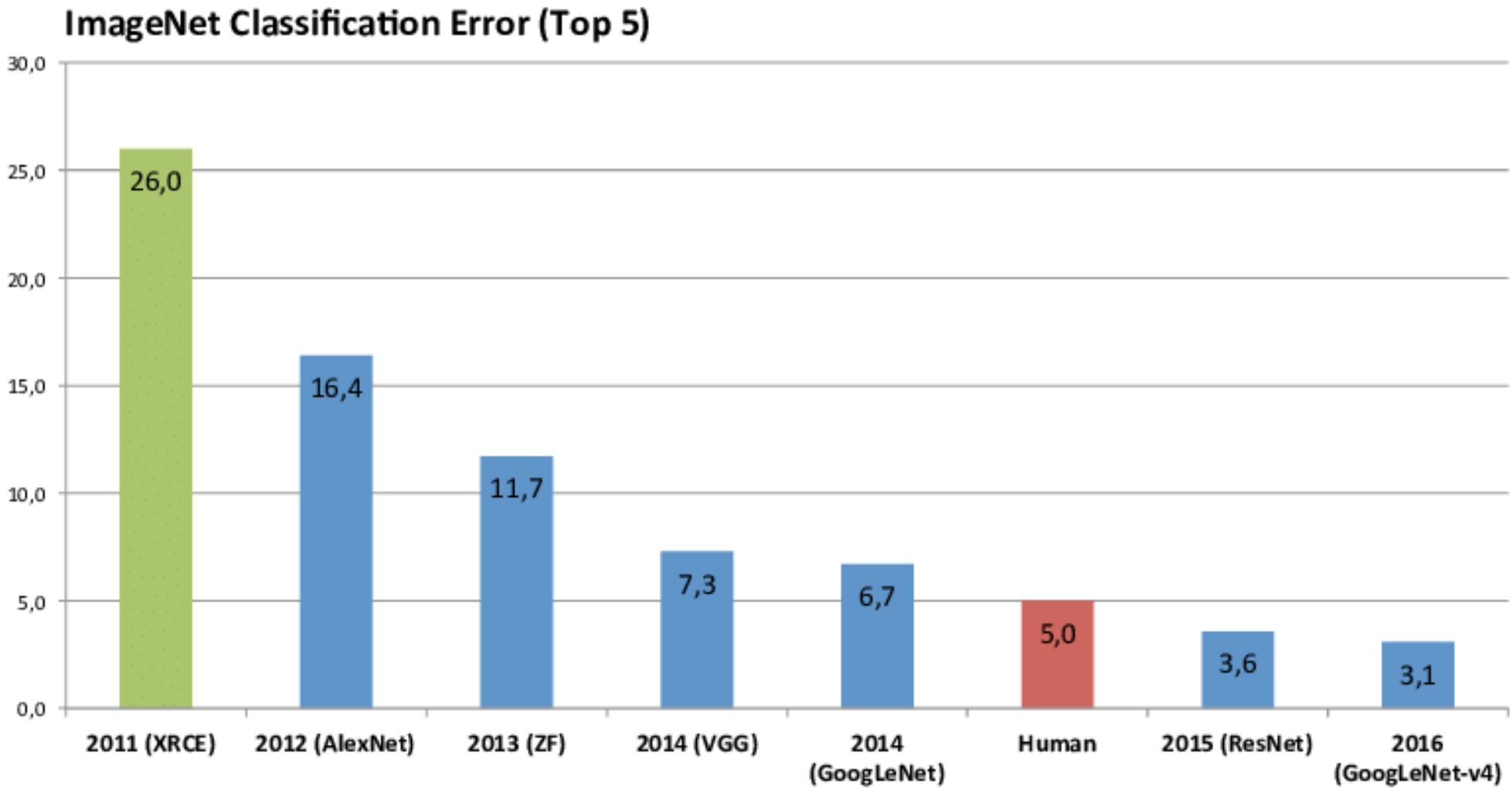
http://www.image-net.org/challenges/LSVRC/

Image classification



http://www.image-net.org/challenges/LSVRC/

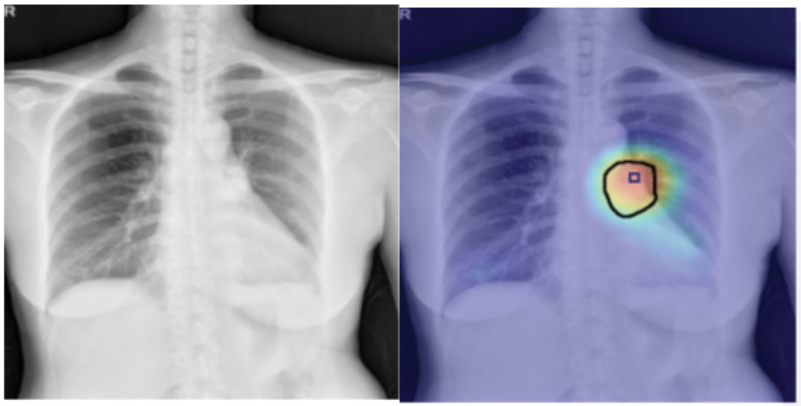
Image classification



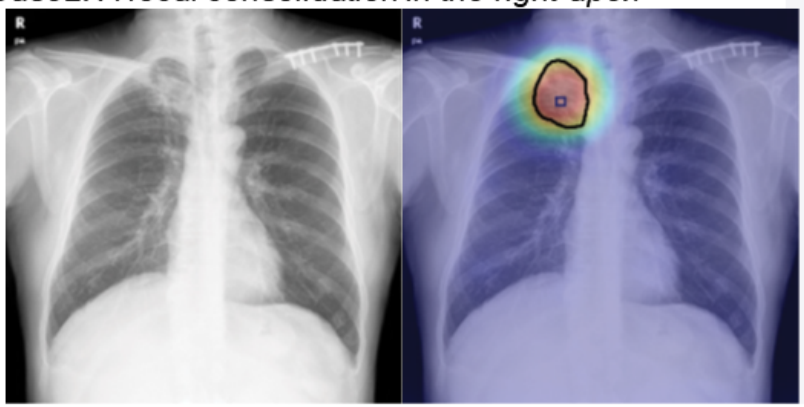
가슴 x-ray diagnosis

A new real-time imaging AI platform on the web at RSNA 2017

Case1. A lung cancer nodule located in the left hilar area



Case2. A focal consolidation in the right apex



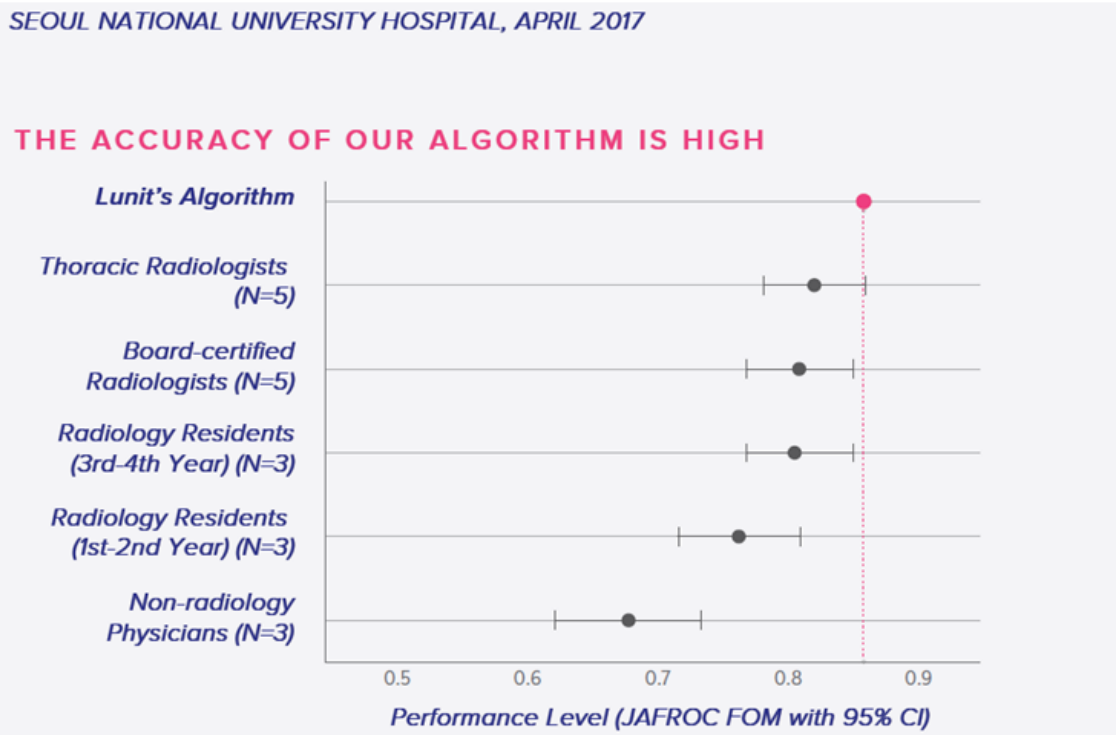
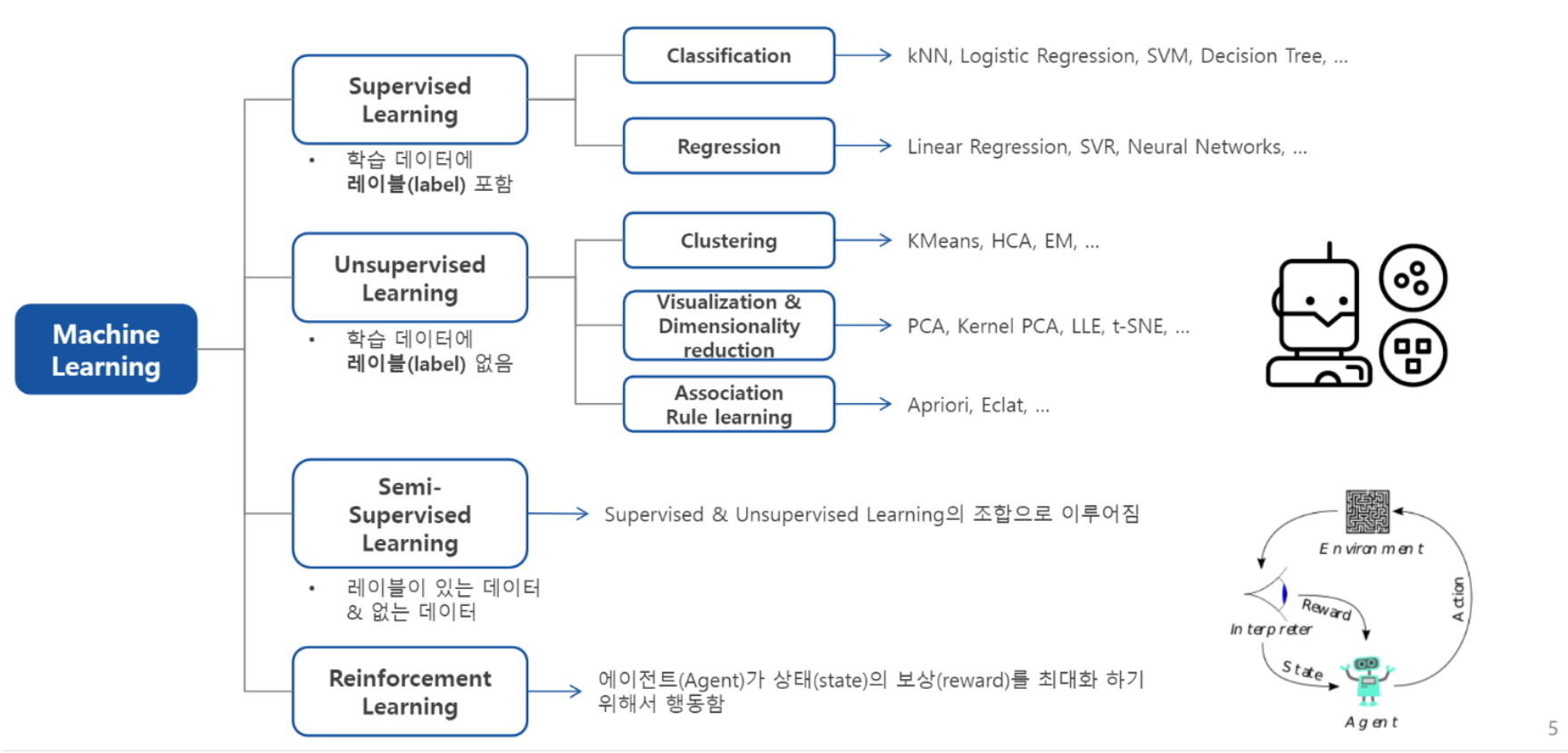


FIGURE 1. The accuracy of our algorithm was shown to be very high, comparable to that of thoracic radiologists (or likely even higher).

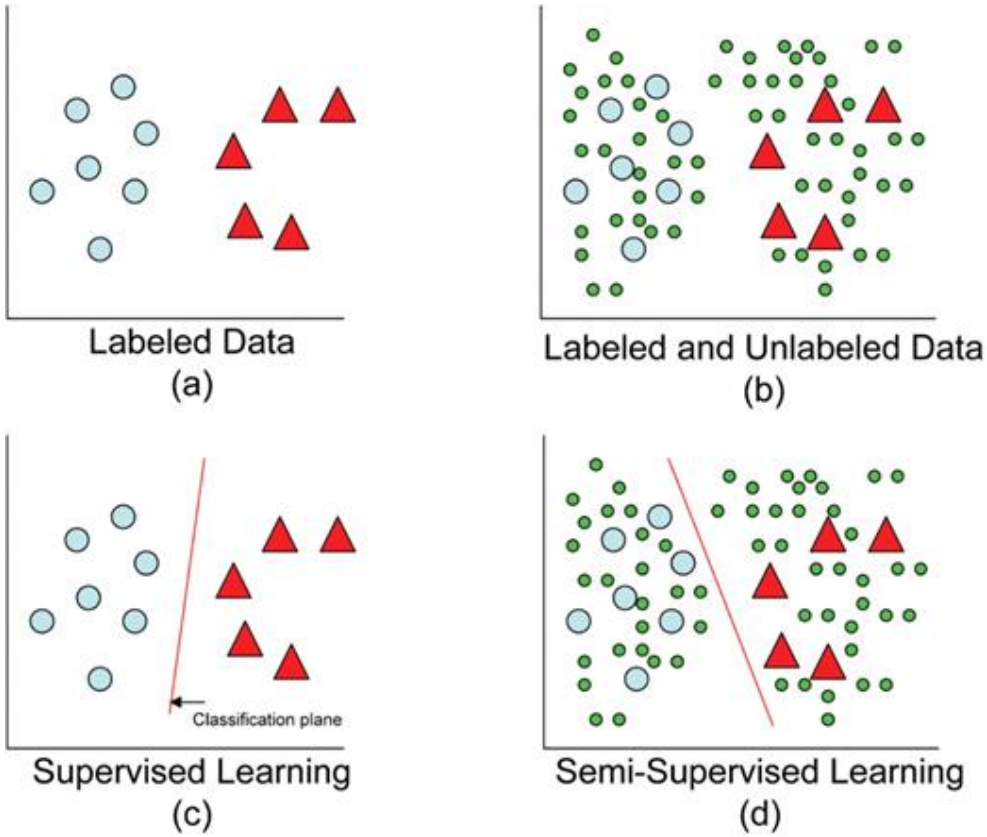
https://insight.lunit.io

기계학습의 분류



https://github.com/ExcelsiorCJH/Hands-On-ML/blob/master/Chap01-The_Machine_Learning_Landscape

준지도 학습 (Semi-Supervised learning)



http://bioinfomatic.oxfordjournals.org

머신러닝 수업 진행

ML 전처리

- Feature들에 대한 적절한 전처리 (pre-processing)
- 초매개변수 조절 (Tuning hyperparameters)
- 모델 성능 평가 (Assessing model performance)

ML 알고리즘

- ML 알고리즘 종류
- 모수(매개변수) 학습 (Parameter Learning)
- 파이썬을 이용한 실습

모델링 과정

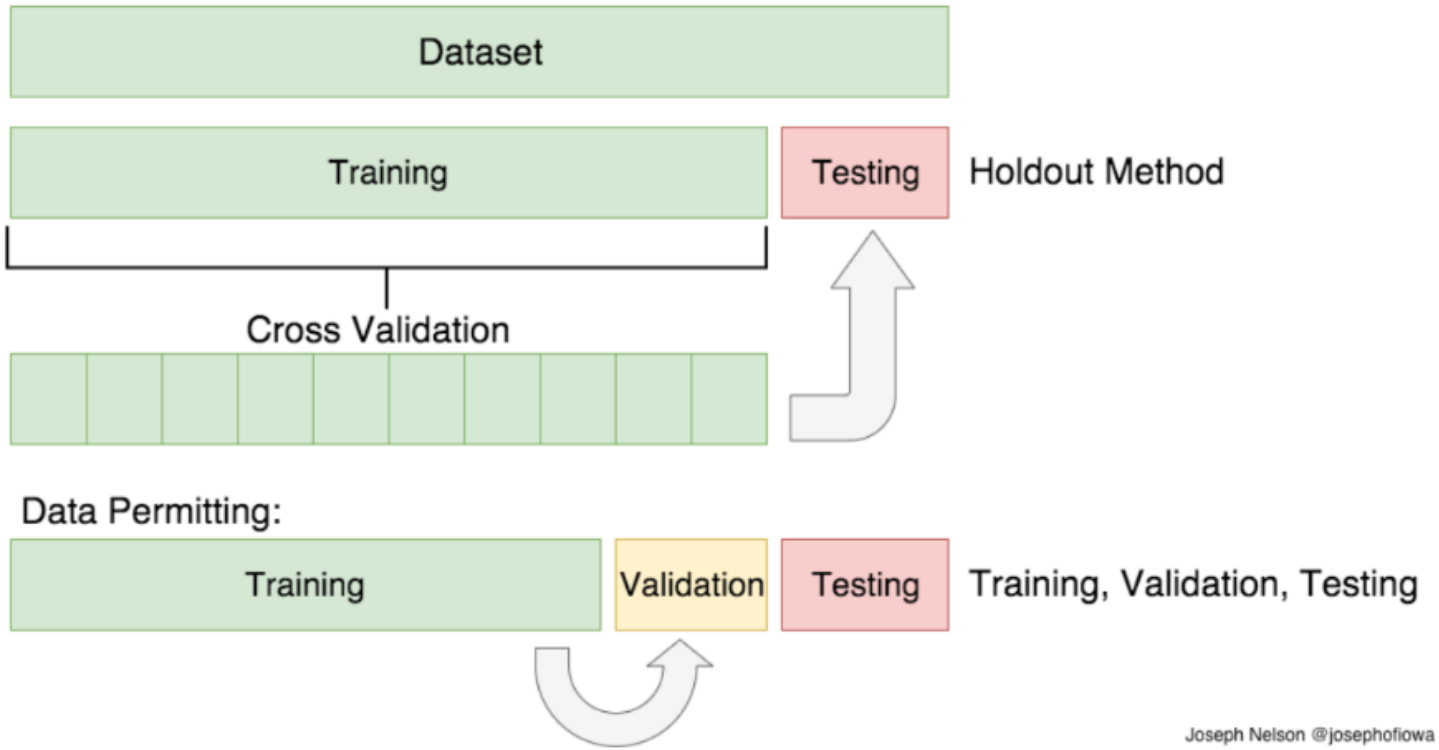
Training set and test set

- ML 모델의 성능평가를 위해서 자료를 분할

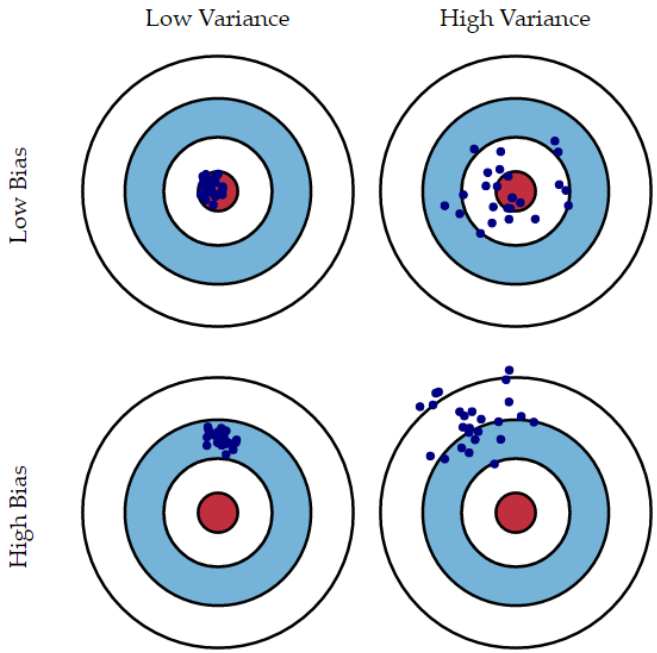
- **Training set**: 모델의 알고리즘 learning, 모델에 사용될 feature들을 결정, 초매개변수 조절 (약 전체 자료수의 70% 로 설정)
  - **Training set**: 모델의 알고리즘 learning
  - **Validation set**: 모델에 사용될 feature들을 결정, 초매개변수 조절, 과적합 (Over-fitting) 방지
- **Test set**: 최종 선택된 모델의 성능평가 (약 전체 자료수의 30% 로 설정), 자료의 수가 적을 경우 생략 가능

일반적인 ML 예측 과정

- 학습세트 (Training set): 머신러닝 모델을 학습할 때 사용
- 검증세트 (Validation set): 하이퍼 파라메터 결정할 때
- 학습세트 (Training set): 학습된 모델을 평가할 때

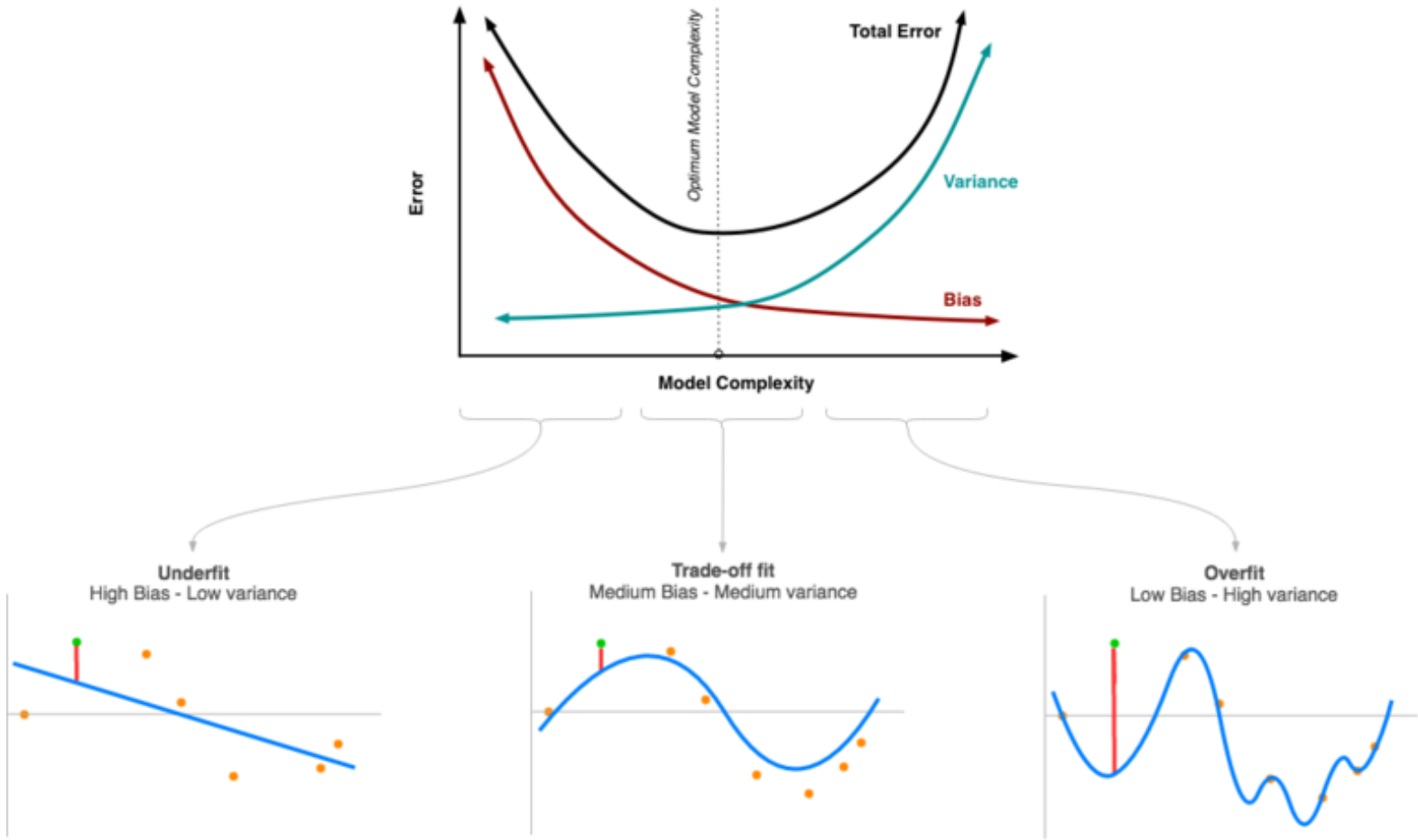


ML 모델의 치우침 (Bias) 과 분산(variance)



https://miro.medium.com

ML 모델의 치우침 (Bias) 과 분산(variance)



https://blog.naver.com/PostView.nhn?blogId=ckdgus1433&logNo=221594203319

치우침(Bias) – 분산(variance) trade-off

치우침(Bias)

- 치우침은 모델의 실제값 (또는 평균) 과 예측치 간의 차이를 의미.
- 과소적합 (underfitting)은 치우침이 높은 모델이 발생하기 쉬움.

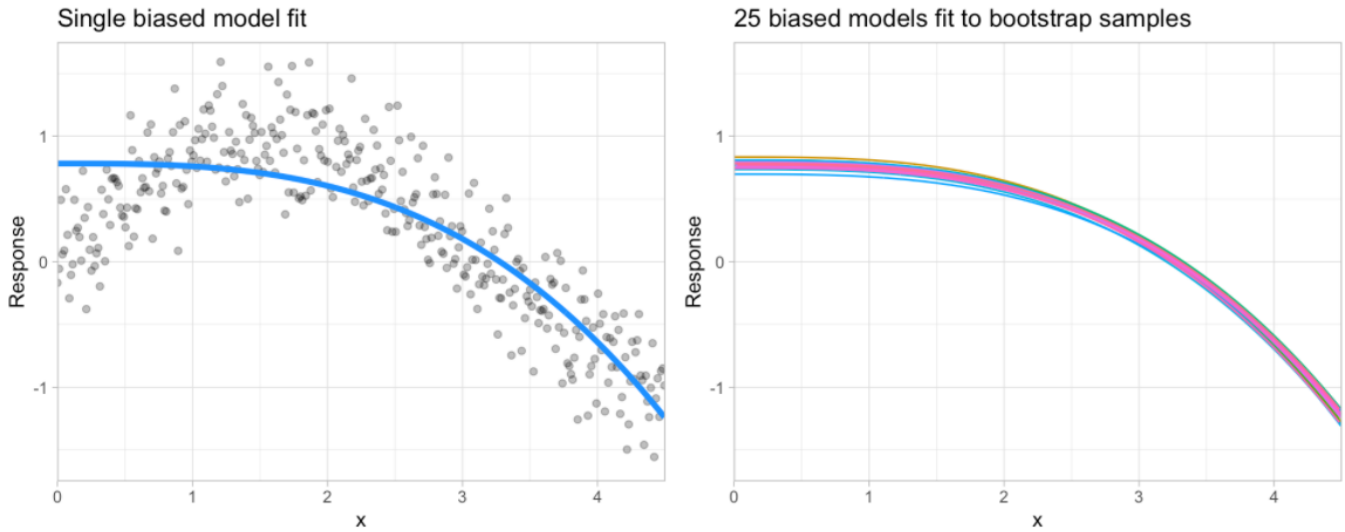


Figure 2.8: A biased polynomial model fit to a single data set does not capture the underlying non-linear, non-monotonic data structure (left). Models fit to 25 bootstrapped replicates of the data are underterred by the noise and generates similar, yet still biased, predictions (right).

https://bradleyboehmke.github.io/HOML

치우침(Bias) – 분산(variance) trade-off

분산 (Variance)

- 분산 (Variance)은 주어진 데이터에서 모델 예측은 변이로 정의.
- 과적합 (overfitting)은 분산이 높은 모델이 발생하기 쉬움.

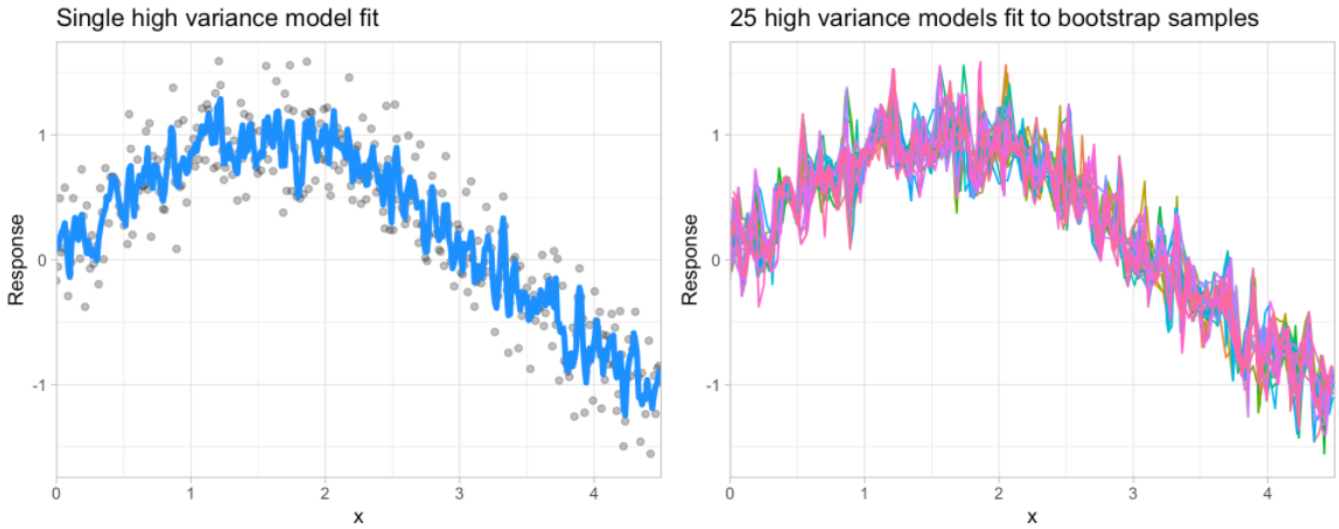
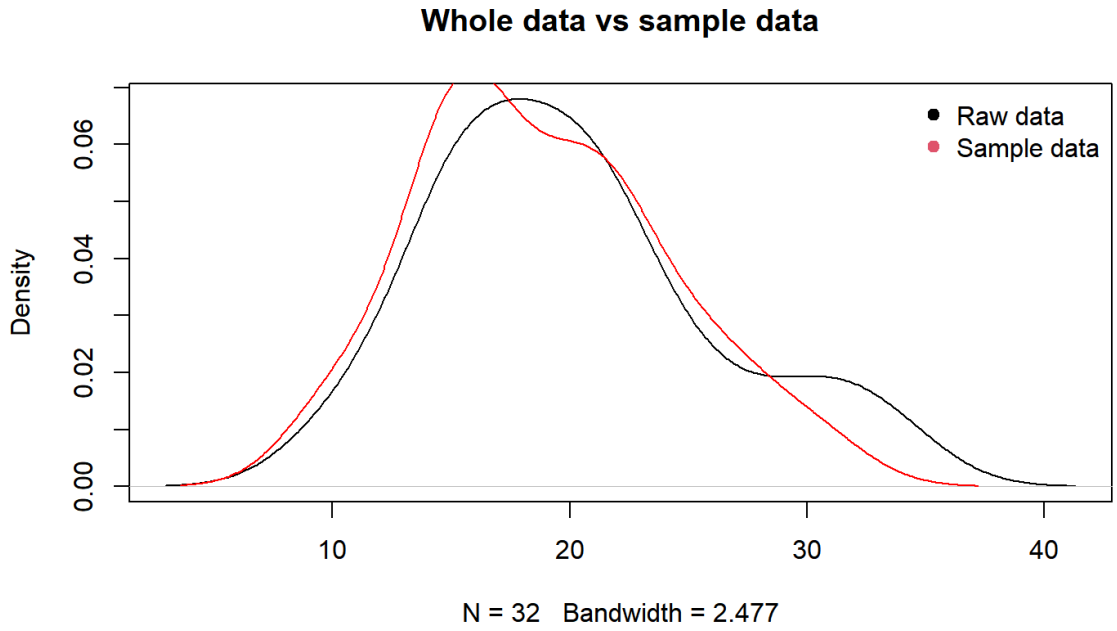


Figure 2.9: A high variance *k*-nearest neighbor model fit to a single data set captures the underlying non-linear, non-monotonic data structure well but also overfits to individual data points (left). Models fit to 25 bootstrapped replicates of the data are deterred by the noise and generate highly variable predictions (right).

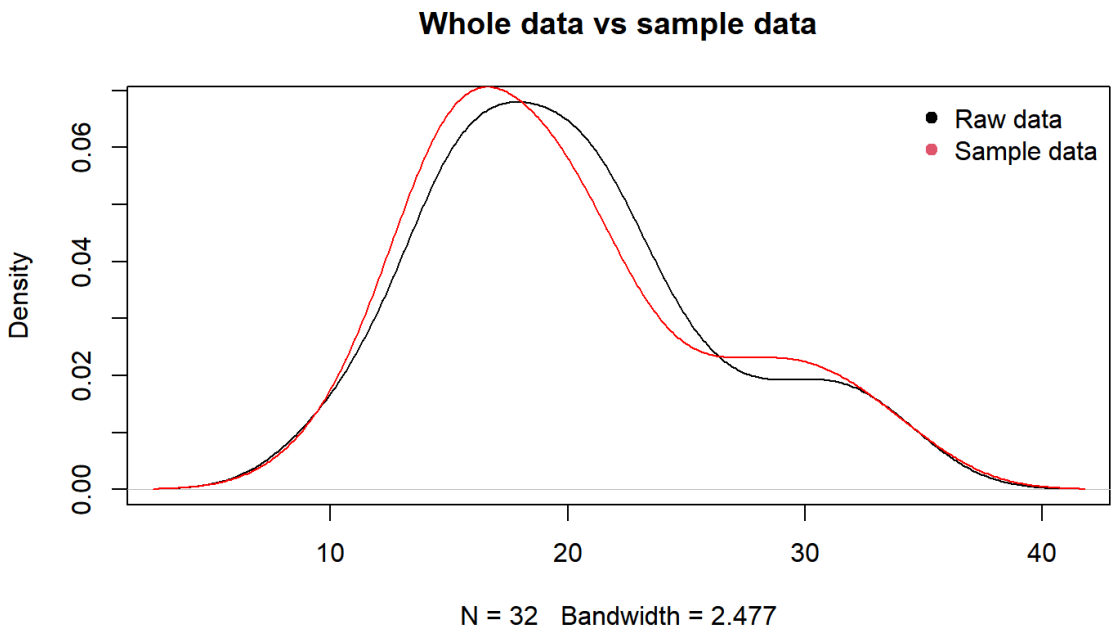
https://bradleyboehmke.github.io/HOML

Training set과 Testing set 분할

- 무작위 샘플링 (Random sampling)
- 계층화 샘플링 (Stratified random sampling)



q1	q2	q3	q4	q5	q6	q7	q8	q9	q10
4	4	2	3	4	3	2	3	3	4



기계학습 모델 평가

k-fold 교차검증 (k-fold cross validation (CV))

- k-fold 교차 검증 (일명 k-fold CV)은 훈련 데이터를 동일한 크기의 k 그룹 (k-fold)으로 무작위로 나누는 리샘플링 방법
- k-fold CV 추정치는 k 테스트 오류를 평균화하여 계산.

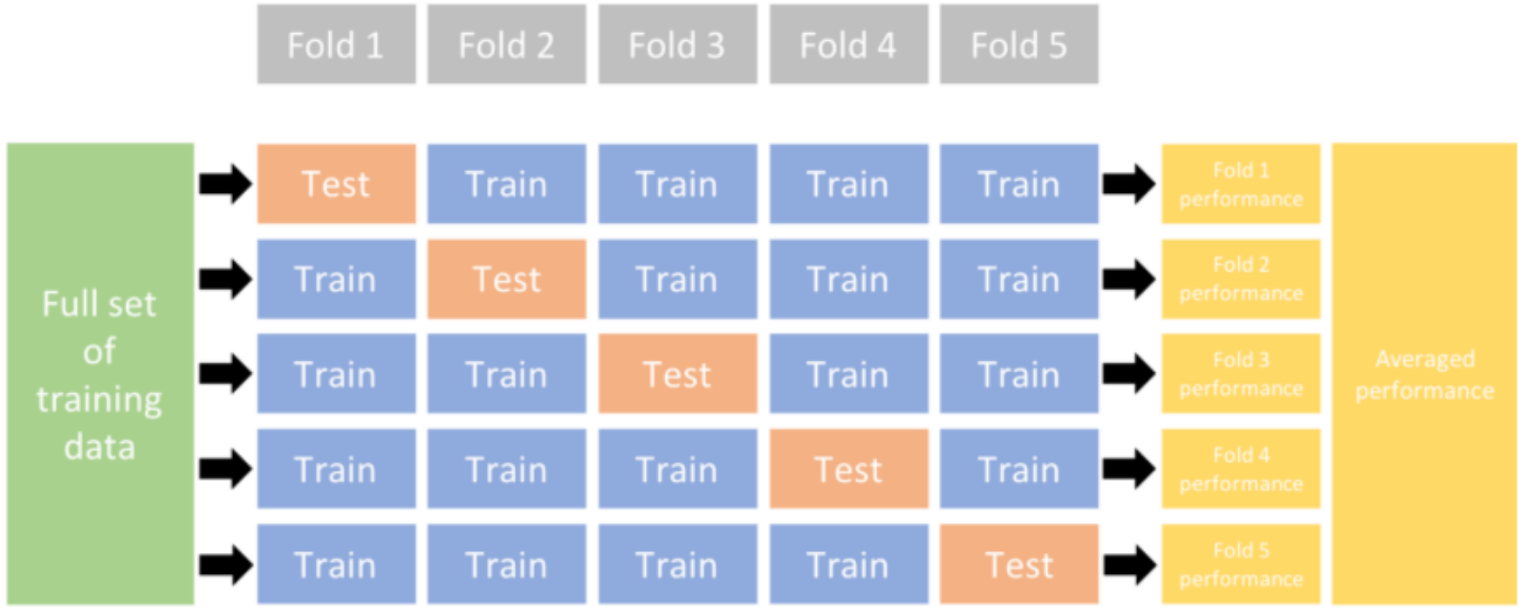


Figure 2.4: Illustration of the k-fold cross validation process.

https://bradleyboehmke.github.io/HOML

기계학습 모델 평가

Bootstrapping

- Bootstrapping 샘플은 복원추출을 이용한 데이터의 무작위 샘플.
- Bootstrapping은 선택한 샘플을 기반으로 모델을 구축하고 OOB (Out-of-Bag) 샘플을 이용하여 모델을 평가

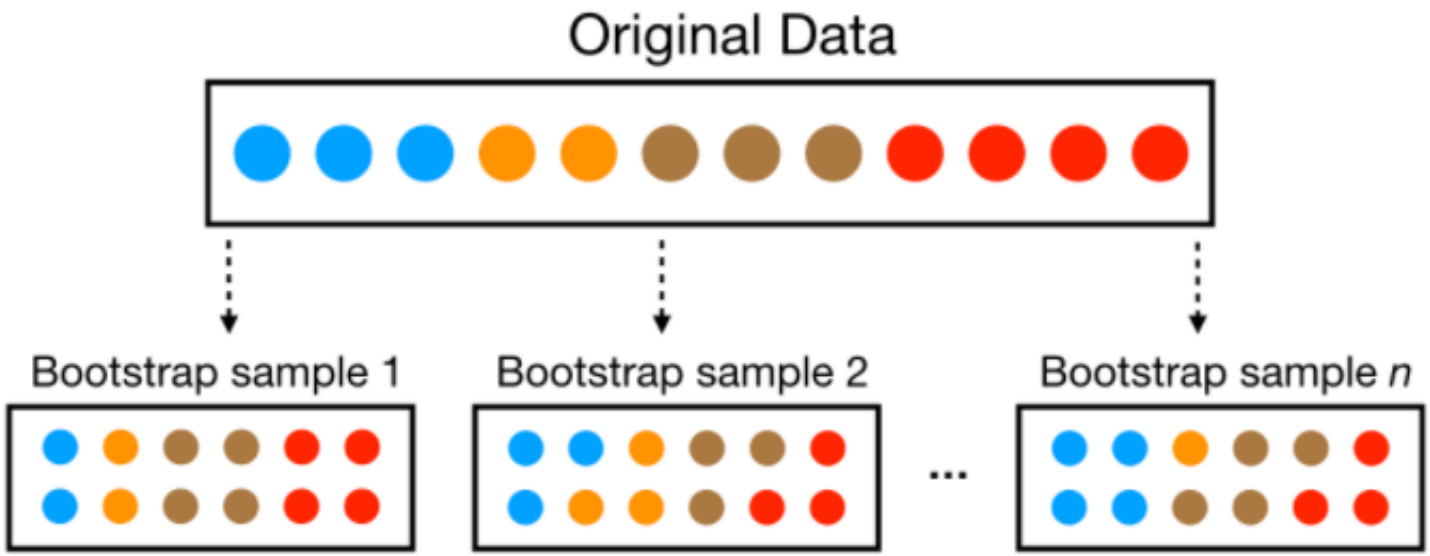


Figure 2.6: Illustration of the bootstrapping process.

https://bradleyboehmke.github.io/HOML

bootstrapping vs 10-fold CV (n = 32) 비교

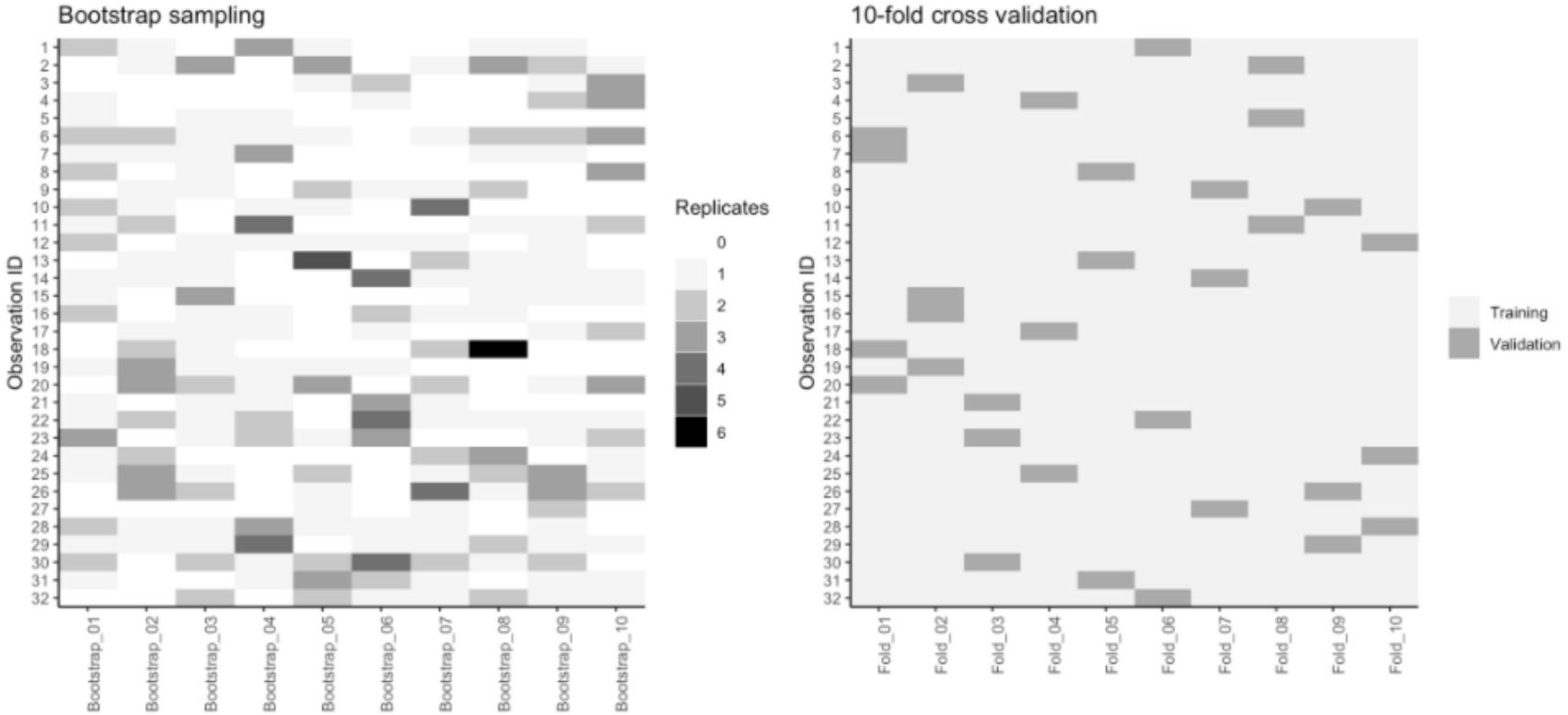


Figure 2.7: Bootstrap sampling (left) versus 10-fold cross validation (right) on 32 observations. For bootstrap sampling, the observations that have zero replications (white) are the out-of-bag observations used for validation.

https://bradleyboehmke.github.io/HOML

초매개변수 조절 (Hyperparameter tuning)

- 초매개변수는 학습 과정을 제어하는 데 사용되는 매개 변수를 의미
- 초매개변수는 모델 학습과정이 아닌 모델 개발자에 의해서 지정됨

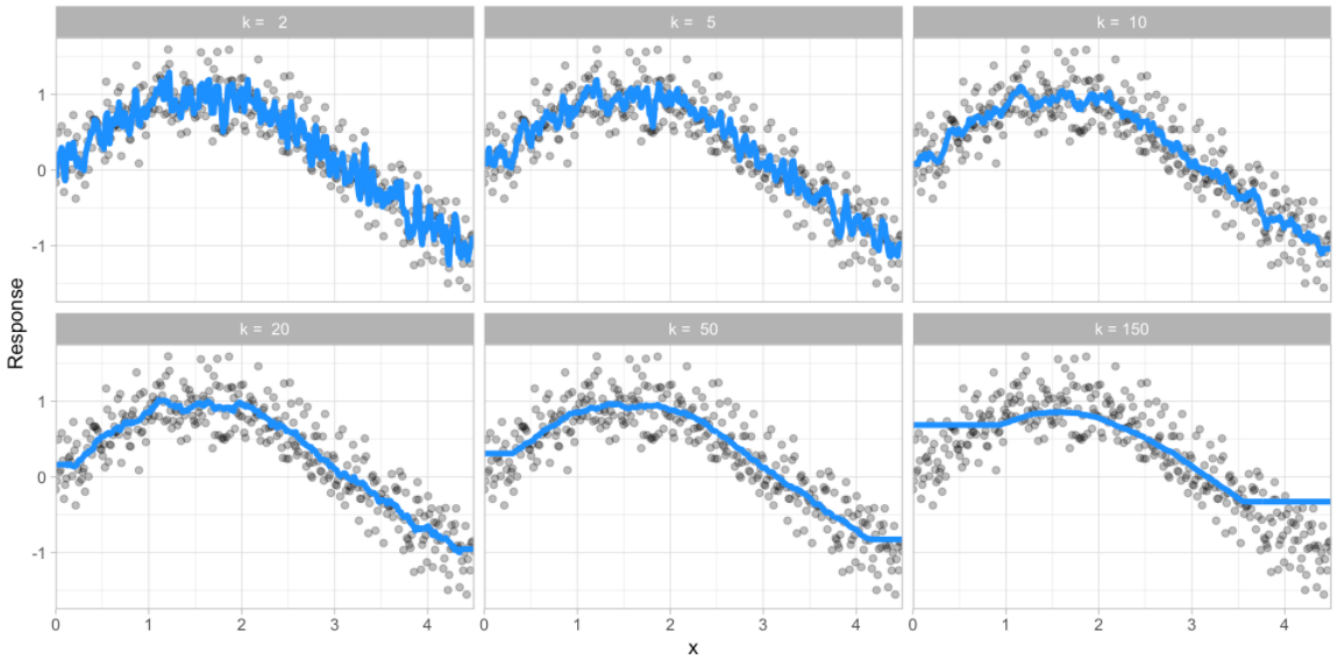
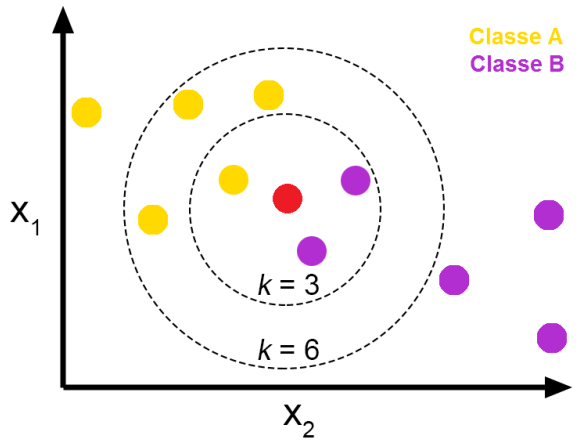


Figure 2.10: *k*-nearest neighbor model with differing values for *k*.

https://bradleyboehmke.github.io/HOML

K-nearest neighbors classification

- 지도학습으로서 분류(Classification) 나 회귀(Regression)에 사용되는 비모수적 방법
- 파라메터 학습을 위한 훈련과정이 없으나 훈련집합은 필요
- 각 데이터 간에 거리를 계산하기 위한 거리척도가 필요
- 초매개변수 k를 설정해야 함
- 거리에 대한 가중치



https://bkshin.tistory.com

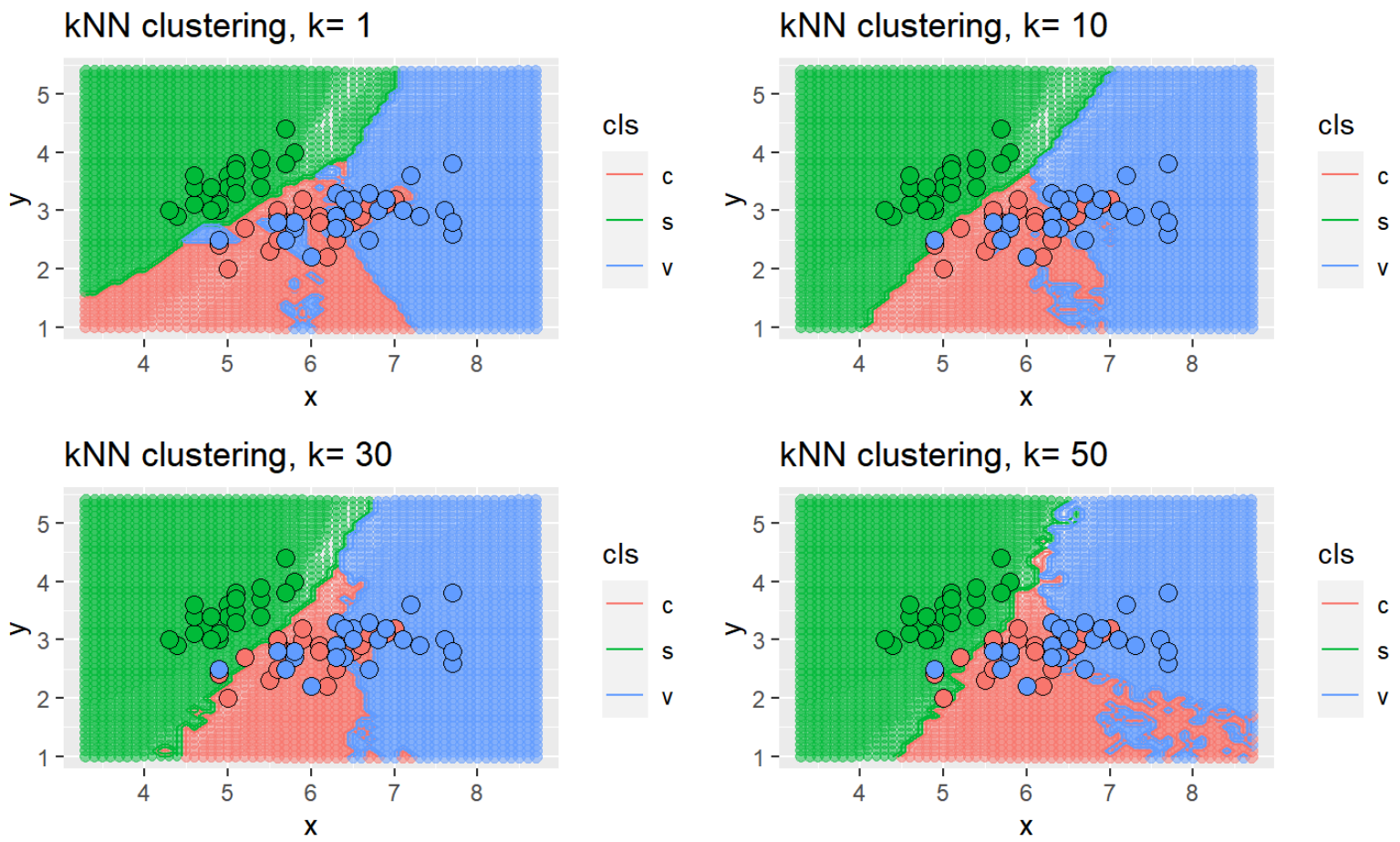
기하학적 거리 (Geometric distance measures)

- Euclidean: $d(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}, \quad \vec{x}, \vec{y} \in p$
- Manhattan: $d(\vec{x}, \vec{y}) = \sum_{i=1}^{p} |x_i - y_i|$
- Minkowski: $d(\vec{x}, \vec{y}) = (\sum_{i=1}^{p} |x_i - y_i|^q)^{\frac{1}{q}}$
- Gower: Manhattan(Continuous) + Dice coefficient(Nominal)

type	length	width
품종 A	10	3
품종 A	12	6
품종 B	14	5
품종 B	15	4
품종 B	16	8
new	12	4

```
# A tibble: 6 x 5
  type    length width sq_sum abs_sum
  <chr>    <dbl> <dbl>  <dbl>   <dbl>
1 품종 A      10     3   2.24       3
2 품종 A      12     6   2          2
3 품종 B      14     5   2.24       3
4 품종 B      15     4   3          3
5 품종 B      16     8   5.66       8
6 new         12     4   0          0
```

KNN clustering area vs k



초매개변수 조절을 위한 격자 탐색 (Grid search) 알고리즘

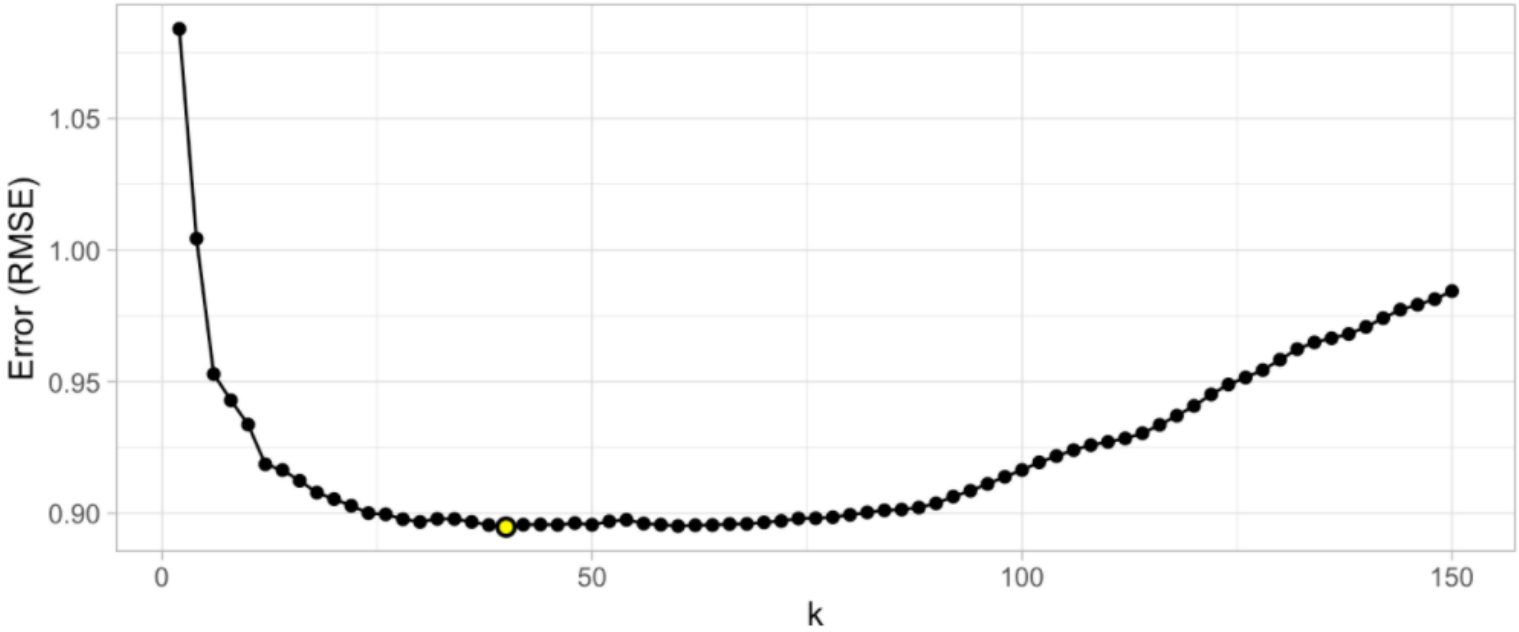


Figure 2.11: Results from a grid search for a *k*-nearest neighbor model assessing values for *k* ranging from 2-150. We see high error values due to high model variance when *k* is small and we also see high errors values due to high model bias when *k* is large. The optimal model is found at *k* = 46.

https://bradleyboehmke.github.io/HOML

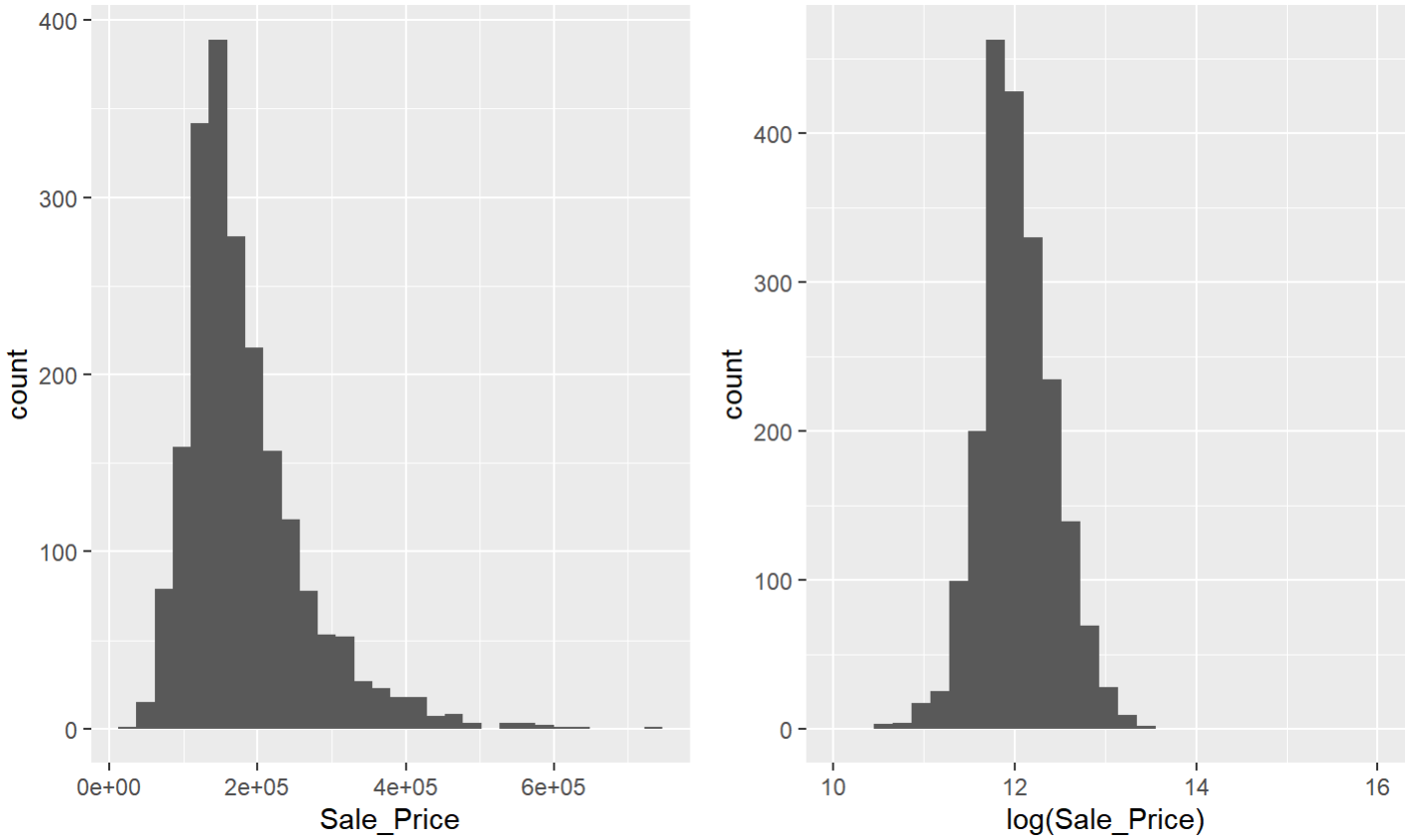
반응변수 전처리 (Target engineering)

- 주로 parametric model에서 예측 및 모델 적용을 위해서 사용
- e.g. Gaussian distribution, Ordinary linear regression

1. **Log transformation**
2. **Box-cox transformation**

Log transformation

- 오른쪽으로 치우친 분포 (Right skewed)가 정규 분포로 변환



Feature 표준화 (Standardization)

- 각 각 feature의 측정 단위에 대한 보정
  - 예) 아파트 값을 추정하기 위한 feature들 중 평수(30평)와 주변지역의 땅값 (3,000,000/평)
- Centering and scaling을 통해서 평균이 0, 표준편차가 1이 되도록 변환 해 줌

Feature 표준화 결과

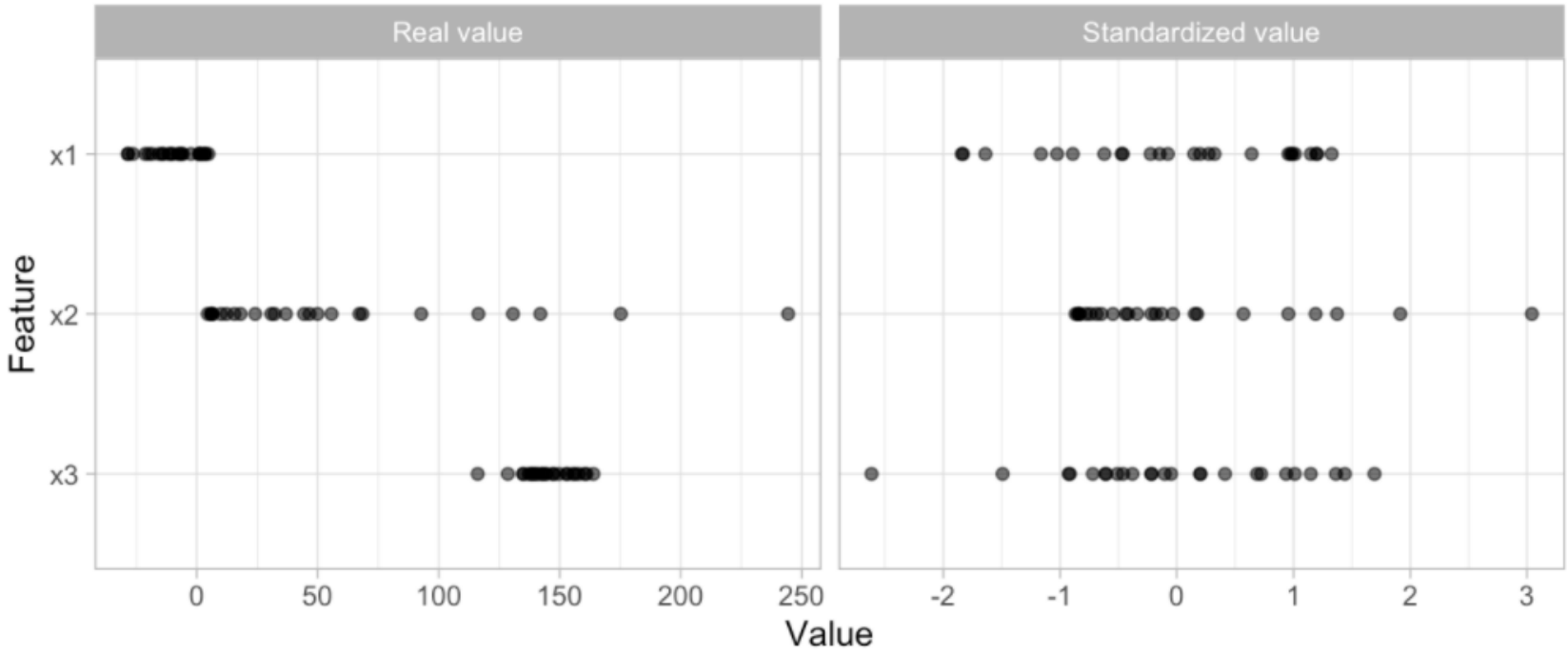


Figure 3.8: Standardizing features allows all features to be compared on a common value scale regardless of their real value differences.

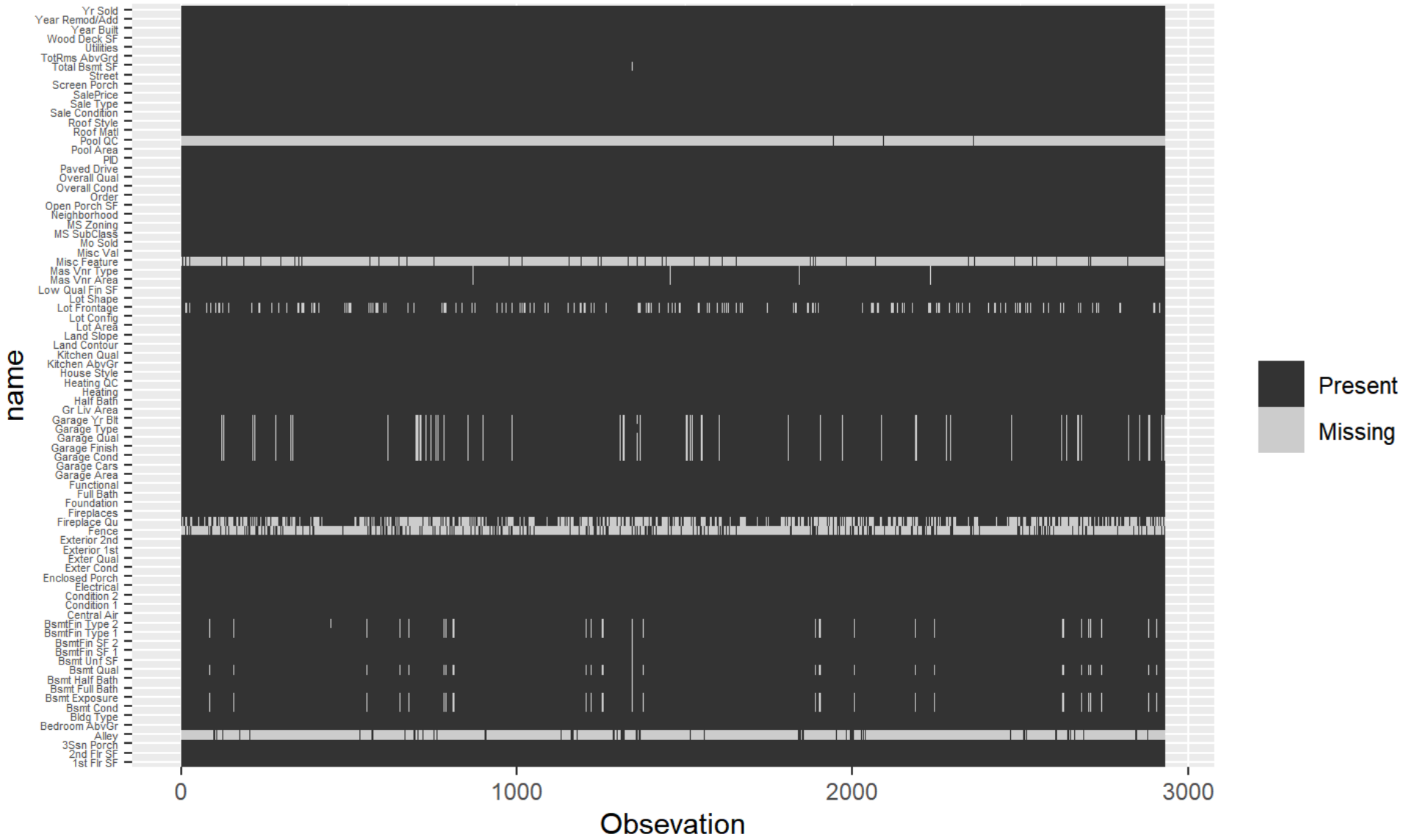
https://bradleyboehmke.github.io/HOML

결측치의 처리 (missing data)

결측치 종류

- 무작위 결측치 (Random missing value)
  - 완전무작위 결측치 (MCAR: Missing Completely At Random)
    - 예) 단순한 결측치
  - 무작위 결측치 (MAR: Missing At Random)
    - 예) 여성(X1)의 경우 체중(X2)에 대한 답이 없음
  - 비무작위 결측치 (NMAR: Not Missing At Random)
    - 체중(X2) 무거운 사람은 체중(X2)에 대한 답이 없음

Missing values plot



결측치 대체 (Imputation)

- 결측치를 “최상의 추측”값으로 대체
- Estimated statistic (e.g., Mean, Median, Mode, Regression)
- K-nearest neighbor
- Tree-based

결측치 대체 방법에 따른 비교

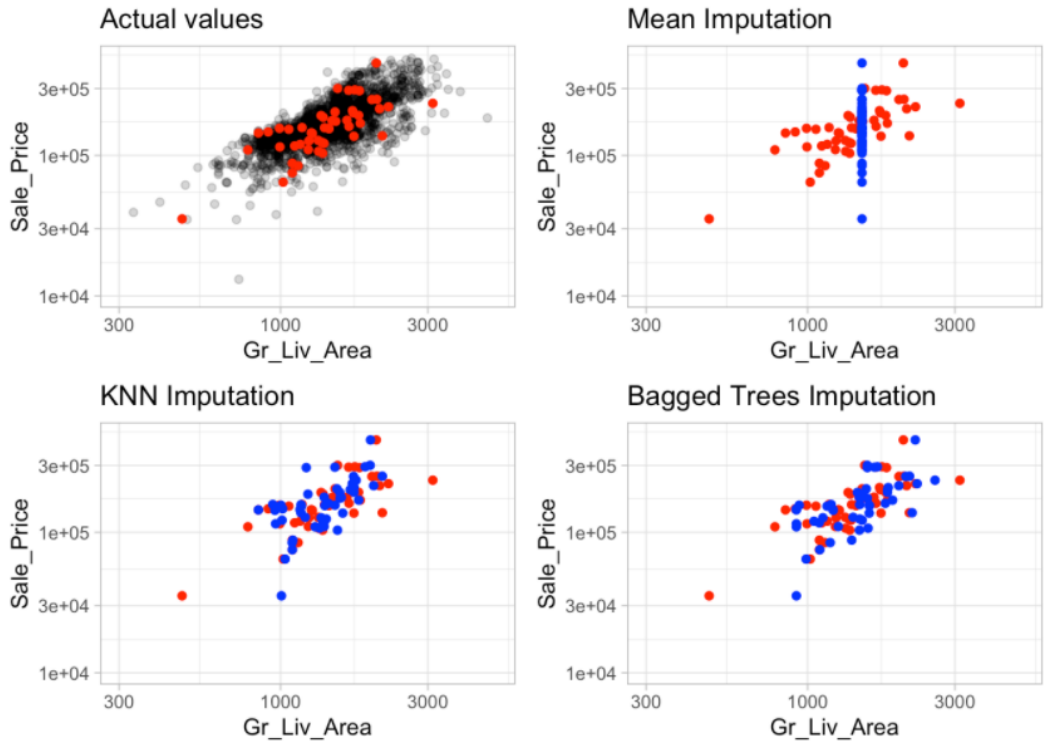


Figure 3.5: Comparison of three different imputation methods. The red points represent actual values which were removed and made missing and the blue points represent the imputed values. Estimated statistic imputation methods (i.e. mean, median) merely predict the same value for each observation and can reduce the signal between a feature and the response; whereas KNN and tree-based procedures tend to maintain the feature distribution and relationship.

https://bradleyboehmke.github.io/HOML

중요하지 않은 Feature 제거 (filtering)

의미없는 변수들 (non-informative predictors) 을 포함했을 때 RMSE 의 변화

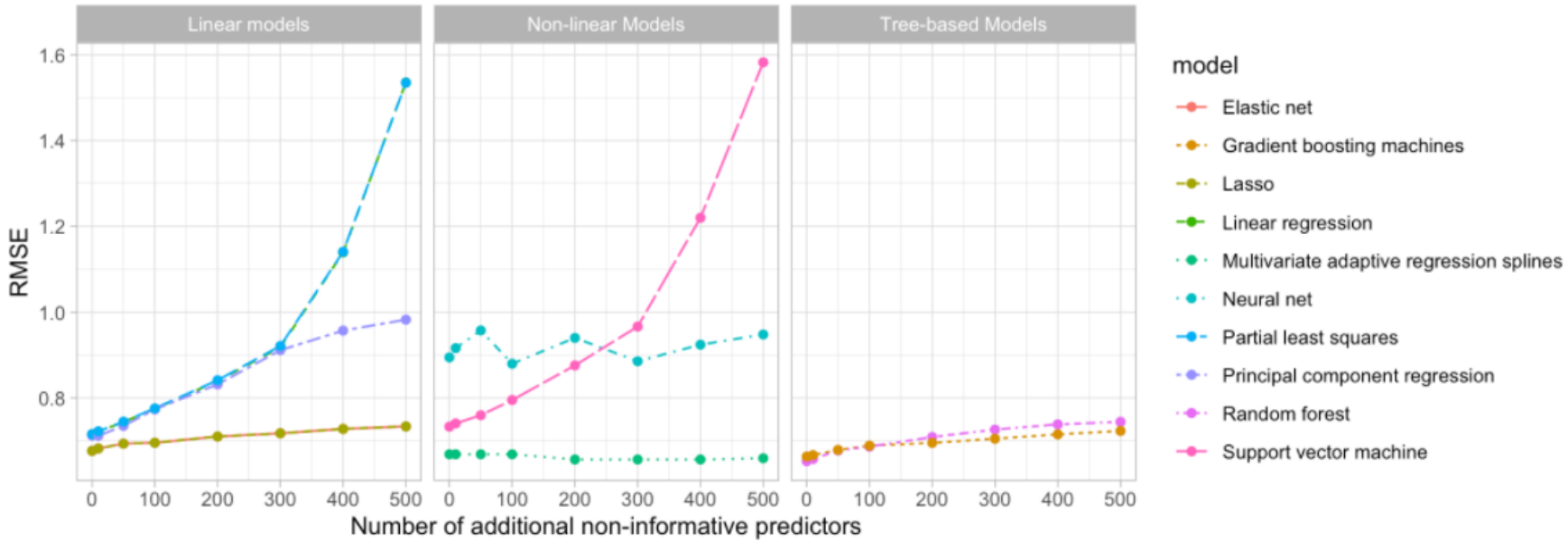


Figure 3.6: Test set RMSE profiles when non-informative predictors are added.

https://bradleyboehmke.github.io/HOML

중요하지 않은 Feature filtering

의미없는 변수들 (non-informative predictors) 을 포함했을 때 학습시간변화

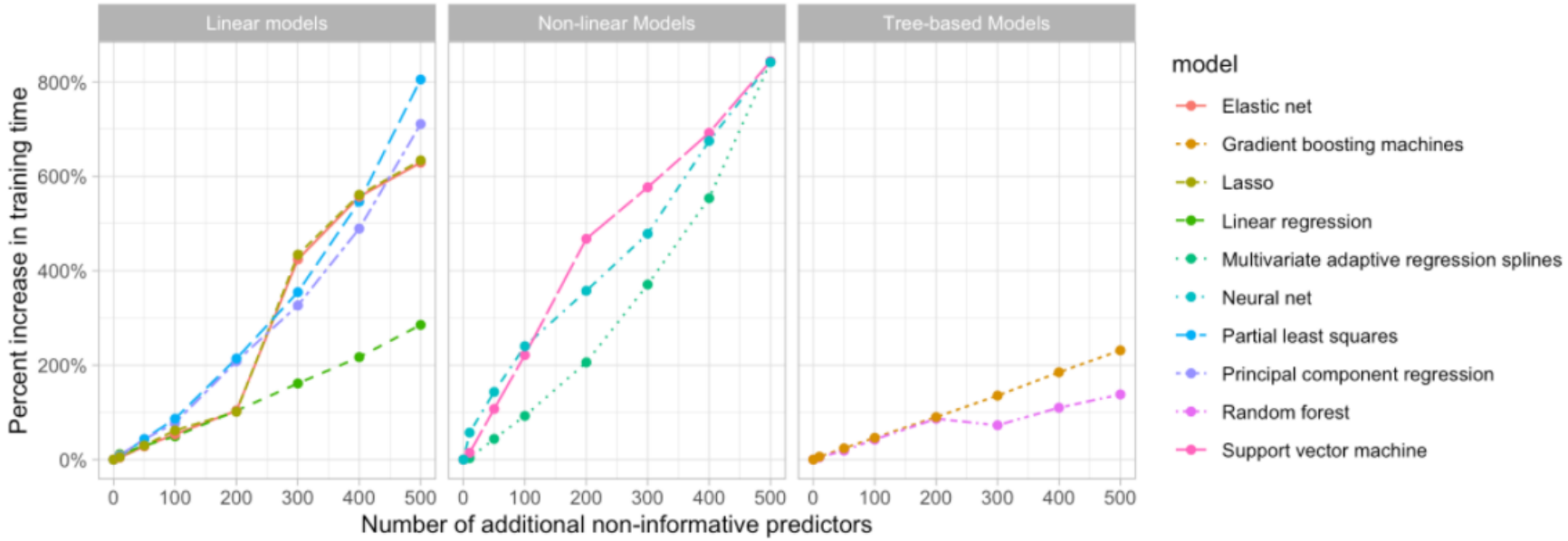


Figure 3.7: Impact in model training time as non-informative predictors are added.

https://bradleyboehmke.github.io/HOML

제로 분산 feature (Zero variance features)

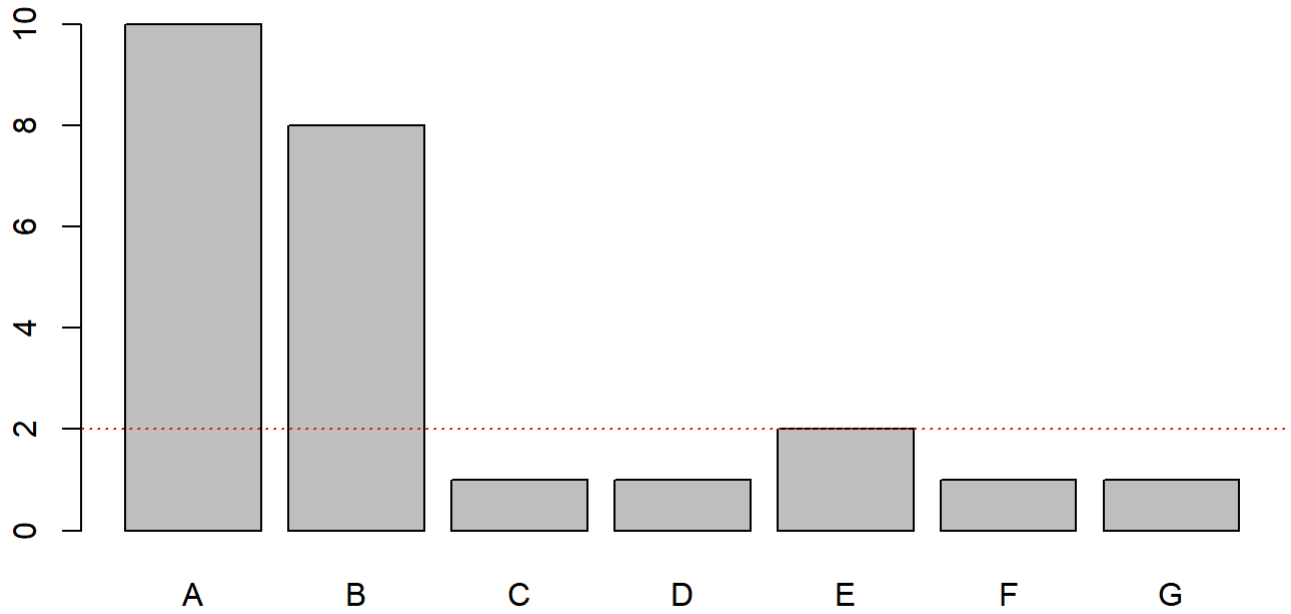
제로 분산 features를 판단하는 일반적 기준
- 전체 샘플중에 서로 다른 관측값의 비율이 낮은 경우 (약 ≤ 10%)
- 가장 빈도가 높은 관측값과 두 번째로 높은 관측값 과의 비가 높은 경우 (약 ≥ 20배)

범주형 데이터 (Categorical feature) engineering

- 재범주화 (Lumping)
- One-hot & dummy encoding
- Label encoding
- Replacing with the mean or proportion

Lumping

- 매우 작은 빈도를 갖는 범주들을 모아서 하나의 범주로 재범주화



One-hot & dummy encoding

- 각 범주를 1 또는 0 (True or False) 로 표시

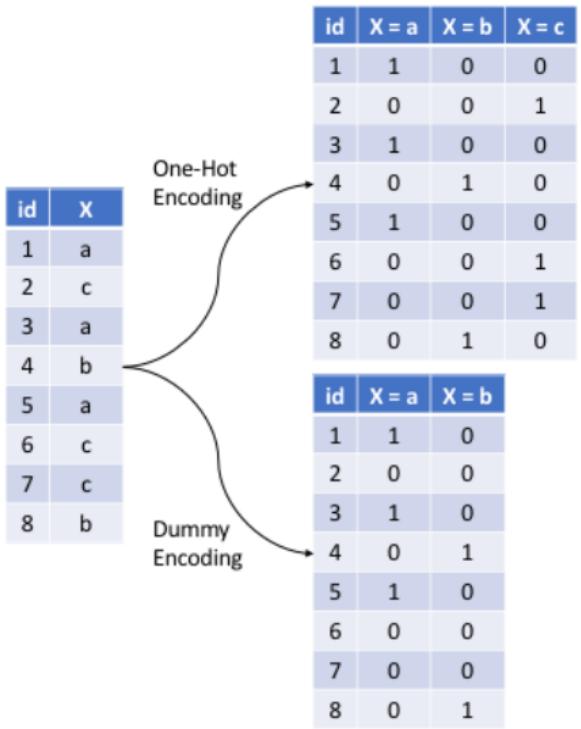


Figure 3.9: Eight observations containing a categorical feature X and the difference in how one-hot and dummy encoding transforms this feature.

https://bradleyboehmke.github.io/HOML

Label encoding

- 각 범주 자료를 연속형 변수로 바꾸어 표현 (순서형 자료의 경우)
  - e.g.) Very high (=5) , high (=4), moderate (=3), low (=2), very low (=1)

Replacing with the mean or proportion

- Target encoding is the process of replacing a categorical value with the mean (regression) or proportion (classification) of the target variable.

차원 축소 (Dimension reduction)

- 여러 개의 feature에서 불필요한 feature들 제거하는 방법
- 예) 주성분 분석 (PCA, principal components analysis)

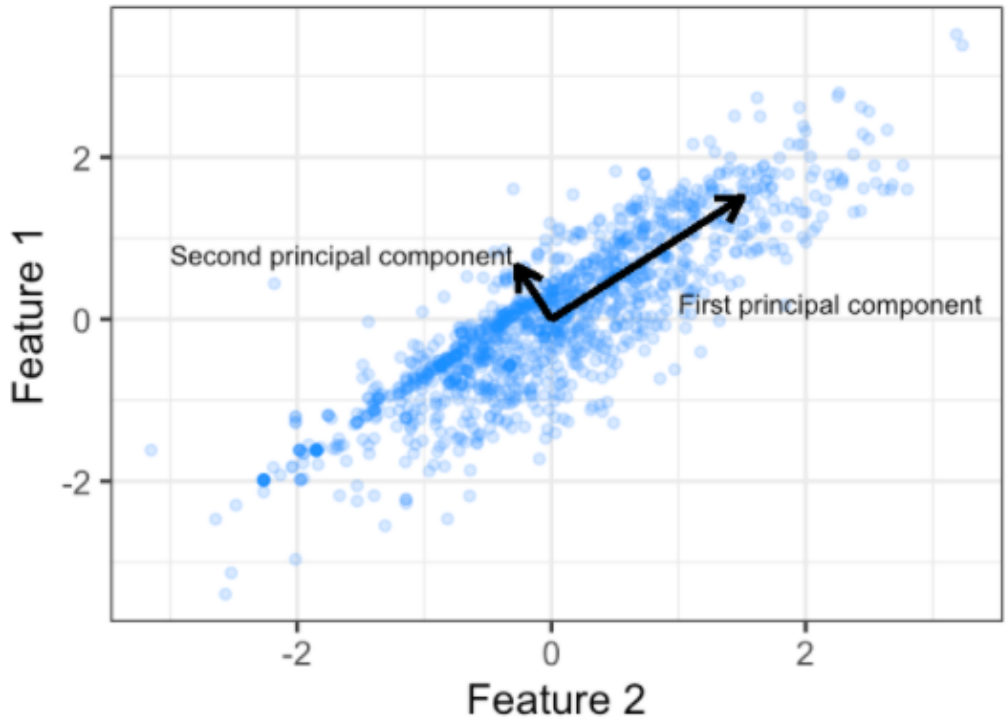


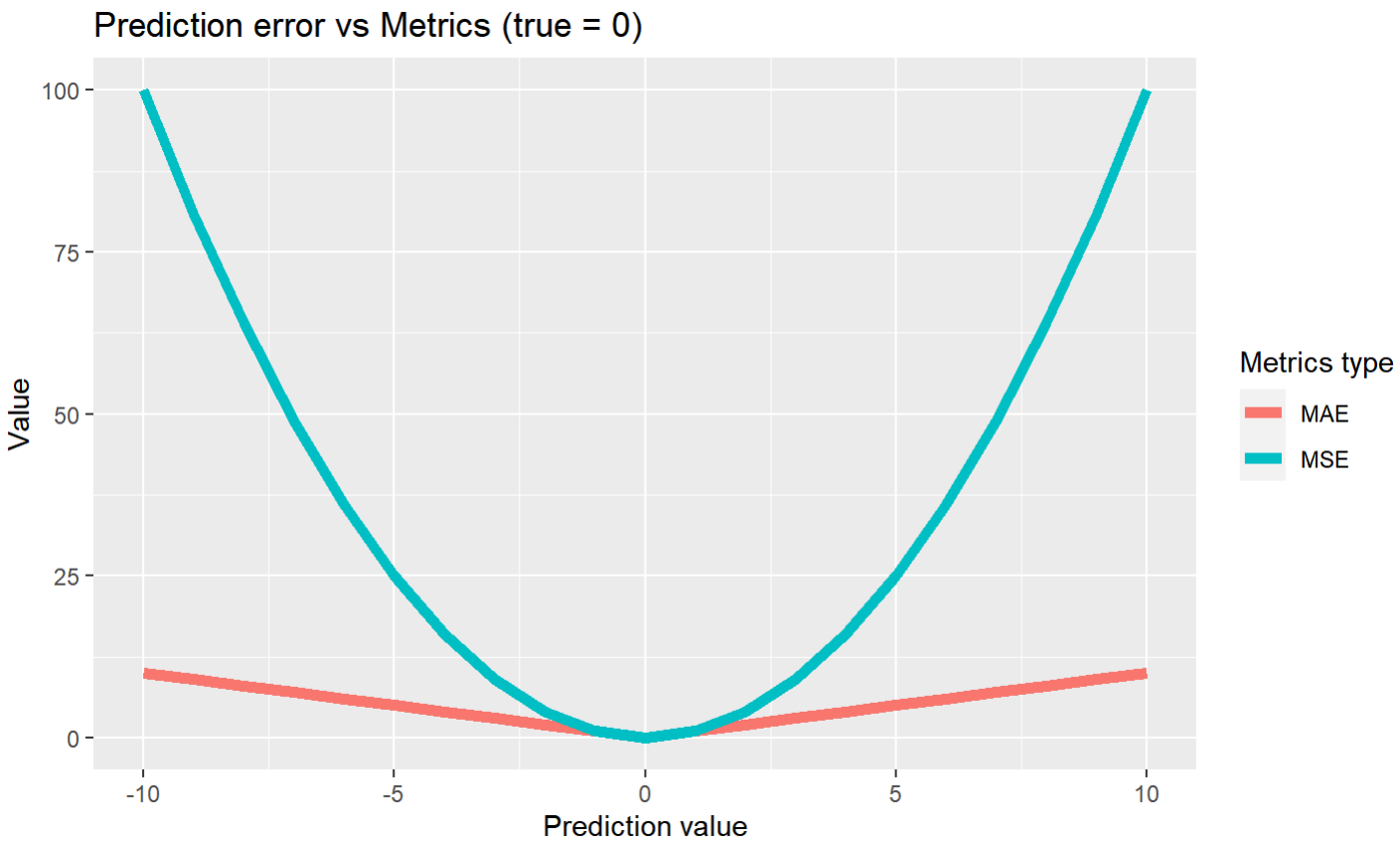
Figure 17.1: Principal components of two features that have 0.56 correlation.

모델평가 지표 (Model evaluation metrics)

회귀분석 모델 (Regression models)

- MSE (Mean squared error) = $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- RMSE (Root mean squared error) = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- MAE (Mean absolute error) = $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

Plot of MSE and MAE



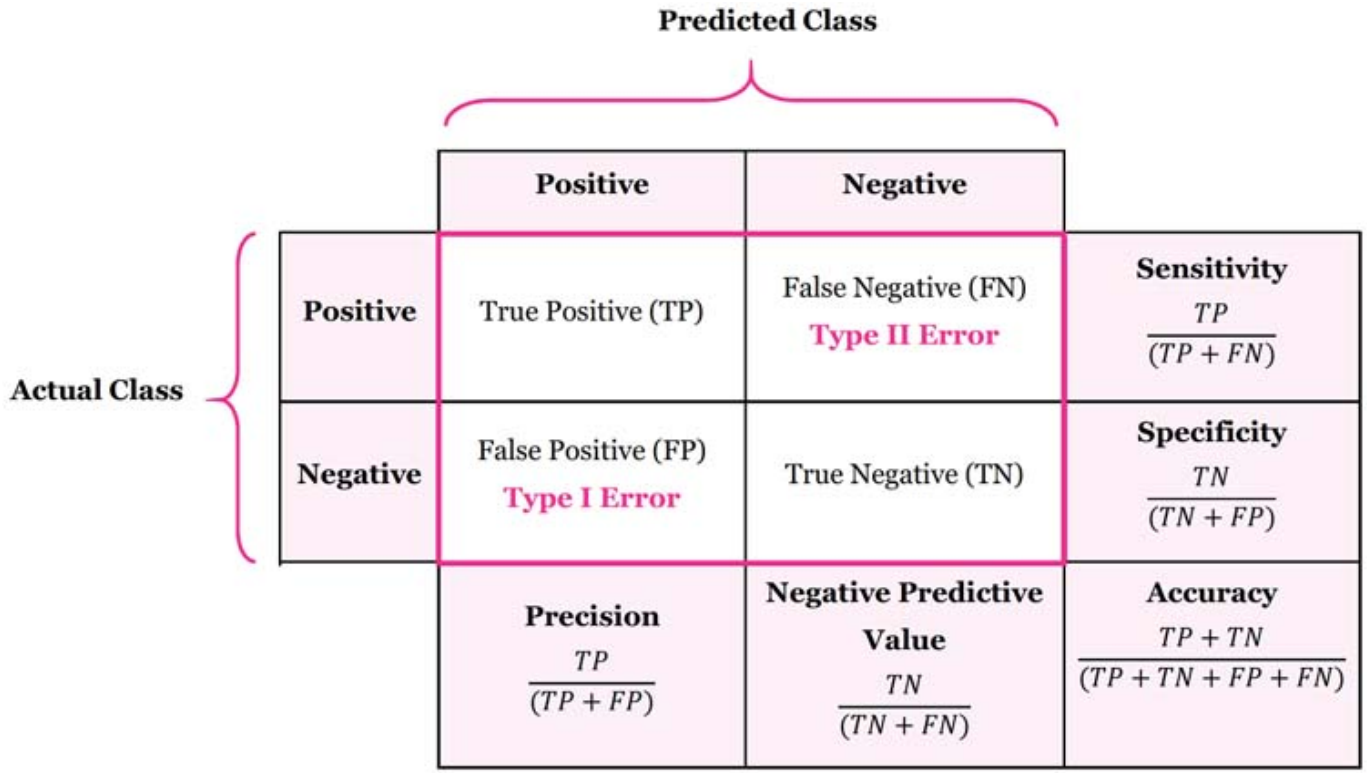
모델평가 지표 (Model evaluation metrics)

분류 모델 (Classification models)

- Misclassification
- Mean per class error
- MSE
- Cross entropy
- Gini Index

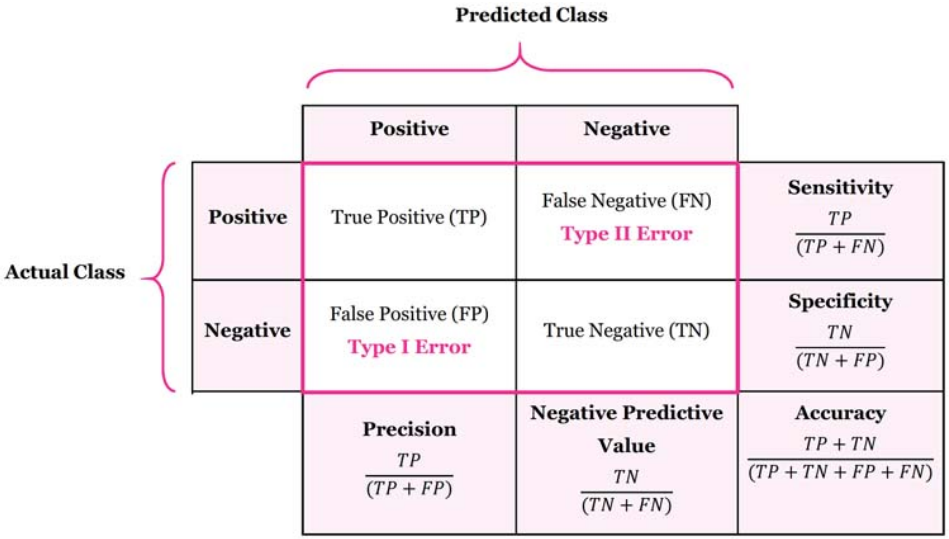
모델평가 지표 (Model evaluation metrics)

Confusion matrix (혼동행렬, 분류결과표)



https://manisha-sirsat.blogspot.com/2019/04/confusion-matrix.html

Confusion matrix 예제



True	Test_yes	Test_No
Yes	8	2
No	1	5

모델평가 지표 (Model evaluation metrics)

ROC (Receiver Operating Characteristic curve)와 AUC (Area under the curve)

- 좋은 분류모델은 높은 정밀도와 감도 가지게 되고 오분류율 (위양성 또는 위음성)을 최소화 함

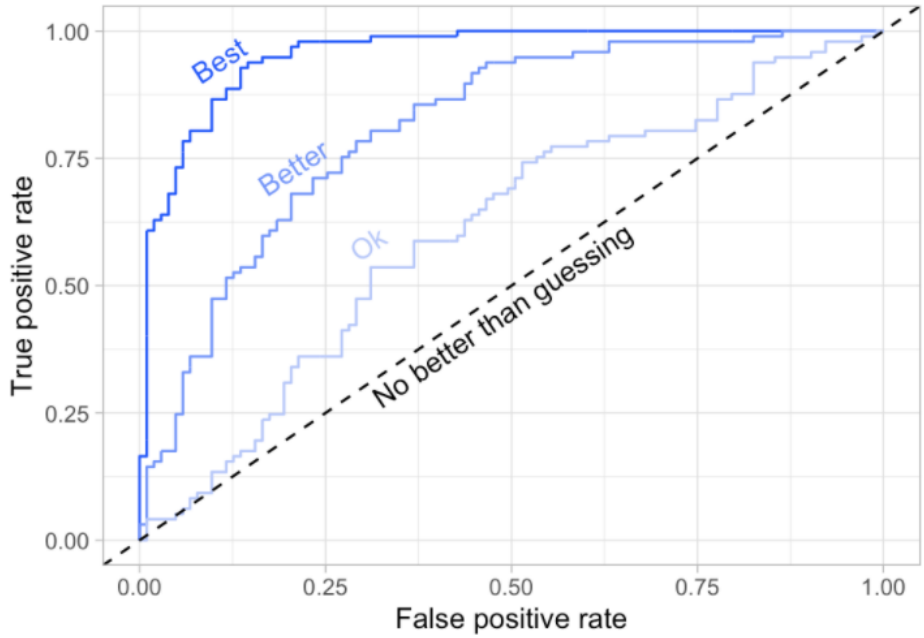


Figure 2.14: ROC curve.

https://bradleyboehmke.github.io/HOML

ROC (Receiver Operating Characteristic curve) 예제

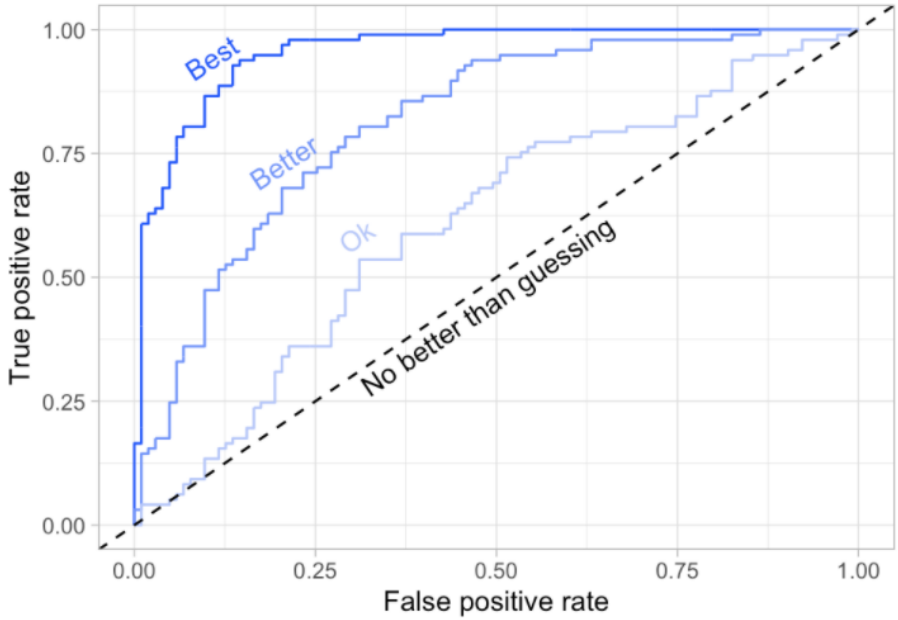


Figure 2.14: ROC curve.

https://bradleyboehmke.github.io/HOML