머신러닝 모델

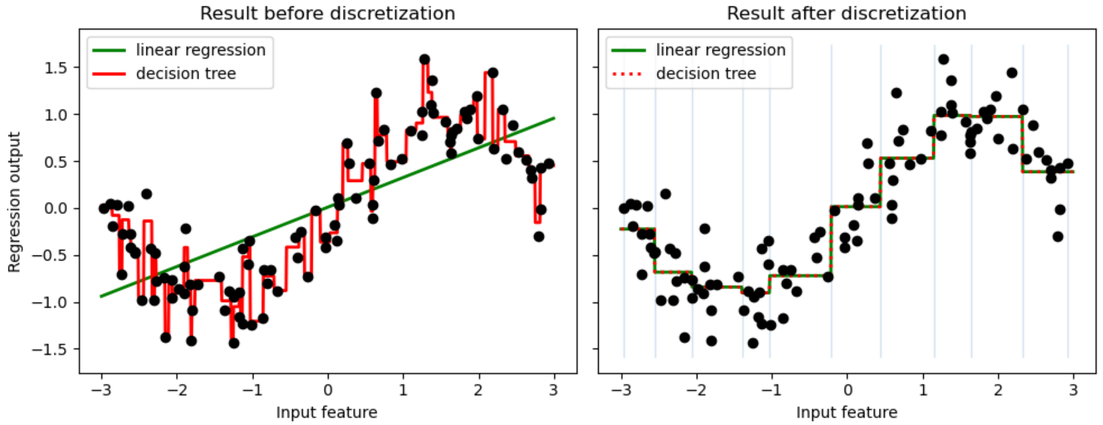
회귀(Regression)

김용담 강사

Contents

이번에 배울 내용은요

1. 회귀(Regression)란 어떤 일을 하는 건가요
2. 회귀 모델에는 어떤 것들이 있나요
3. 회귀 모델은 어떻게 평가를 하나요



1

회귀란 어떤 일을 하는 것인가요

회귀(Regression)

데이터의 경향성을 파악해봅시다

- 회귀의 (비교적) 엄밀한 정의 (Formal Definition)

  "In statistical modeling, **regression analysis** is a set of statistical processes for estimating the relationships between a ***dependent variable(y)*** and one or more ***independent variables(X)***."

회귀(Regression)

데이터를 경향성을 파악해봅시다

- 회귀의 직관적인 의미

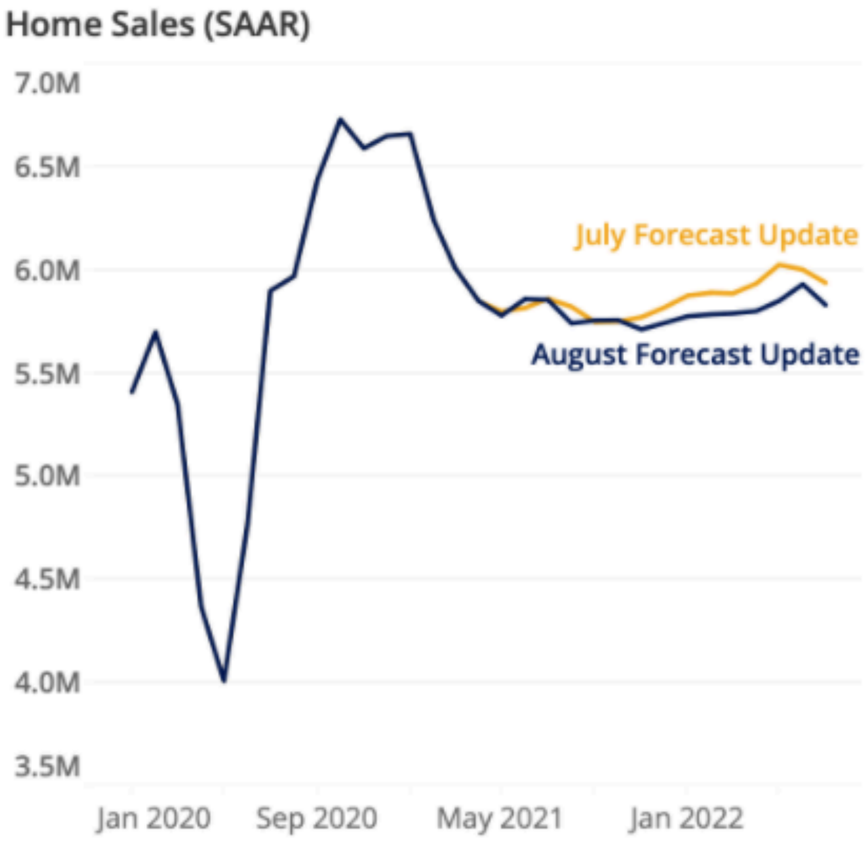
  - 주어진 데이터(***X***)와 원하는 값(***y***) 사이의 관계를 찾는 방법

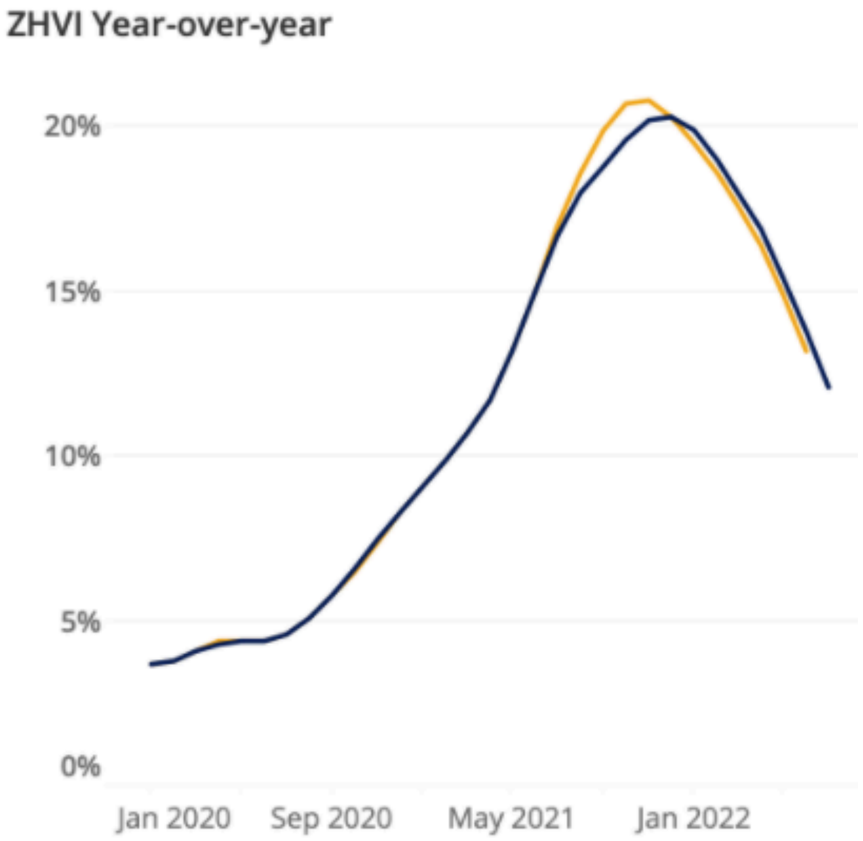
  - 주어진 데이터(***X***)를 통해서 원하는 값(***y = target value***)을 예측하는 방법

  e.g. 부동산 매물 관련된 여러 가지 데이터(***X***)가 주어졌을 때, 집값(***y***)을 예측하는 작업

집값 예측(House Price Prediction)

부동산 정보를 토대로 집값 예측하기





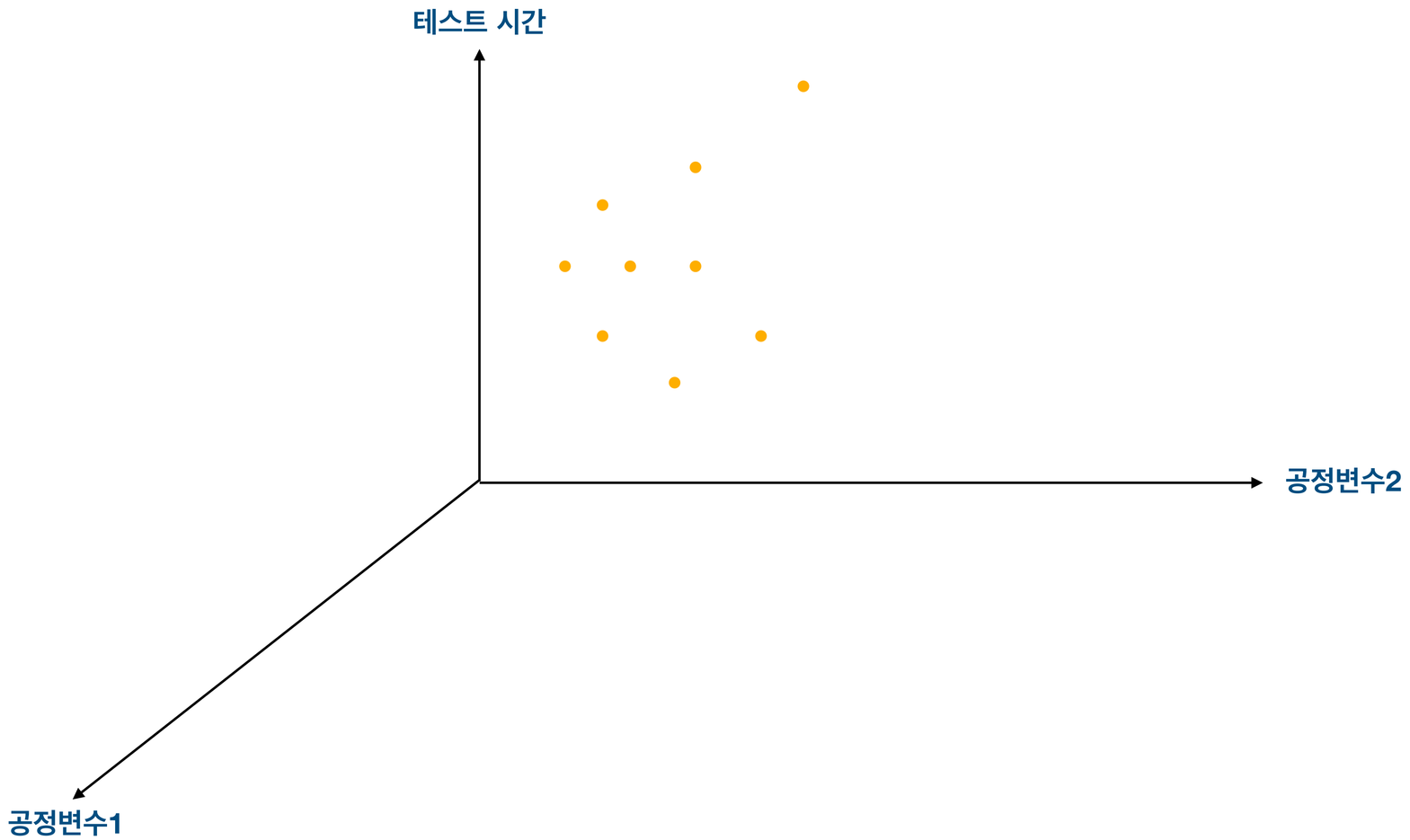
Source : https://www.zillow.com/research/home-values-sales-forecast-july-2021-29941/

2

회귀 모델에는 어떤 것들이 있나요

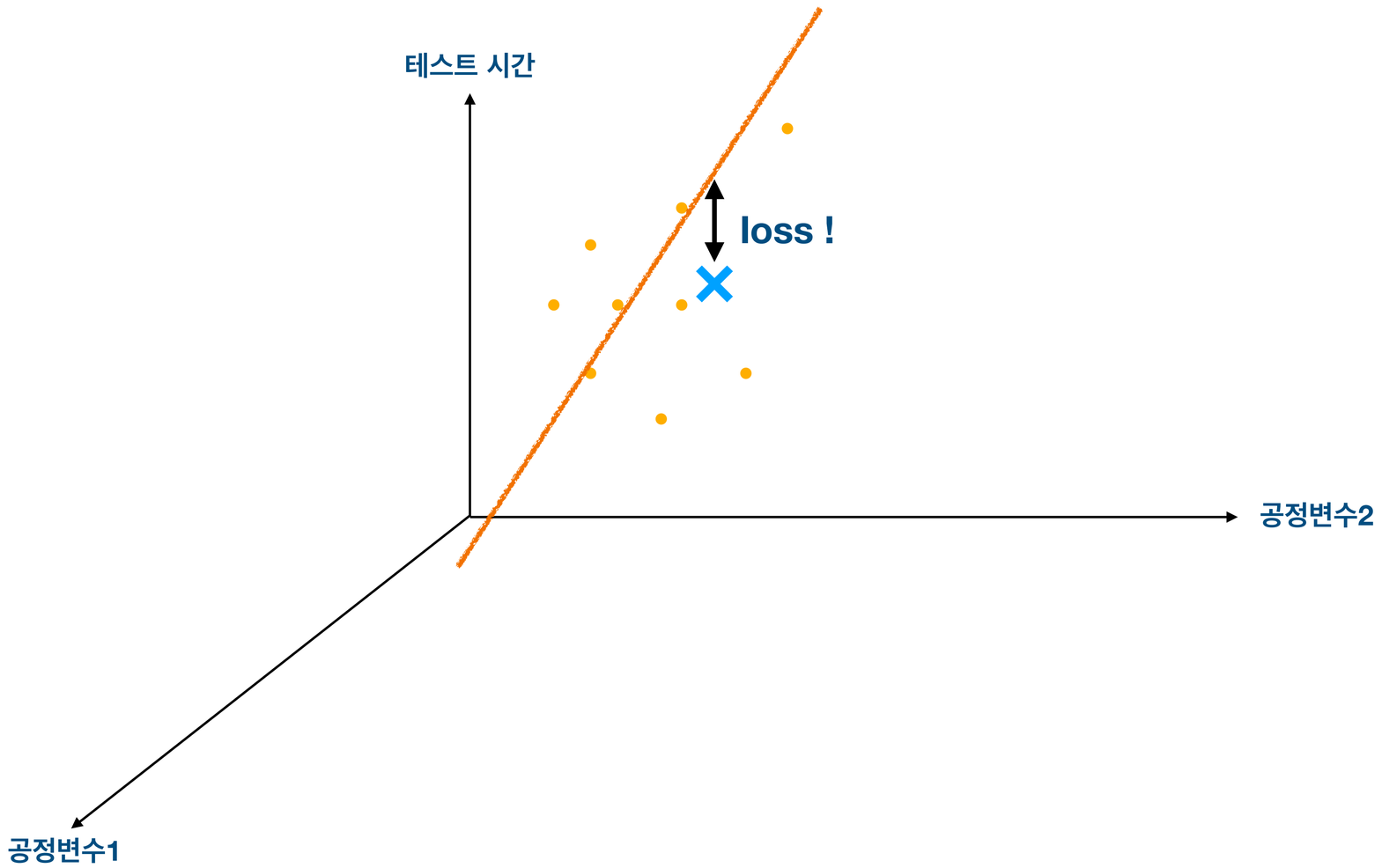
1) Linear Regression

가장 직관적이고 많이 사용되는 선형 회귀 모델



1) Linear Regression

어떻게 직선을 찾을 것인가?

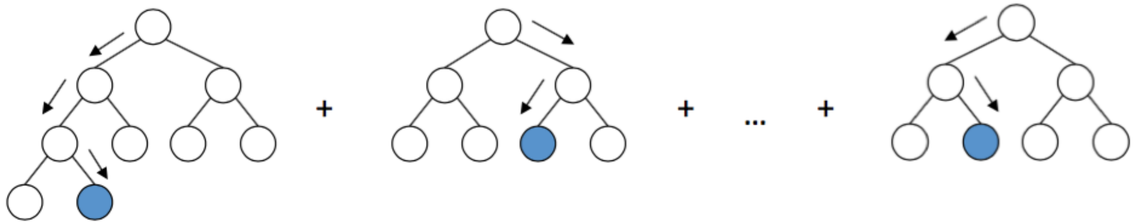


1) Linear Regression

안 좋은 예측 결과라면?

2) LightGBM Regressor

실제 데이터 분석 대회에서 가장 많이 사용하는 효과적인 회귀 모델



- kaggle 같은 실전 데이터 분석 대회에서 가장 많이 사용하는 회귀 모델
- 여러 DecisionTree중에 target value를 잘 찾는 tree들만 찾아서 그 방향으로 트리를 확장해 나갑니다.
- 대용량 데이터에 대해서 적은 메모리로도 빠르게 성능이 좋은 회귀 모델을 만들 수 있습니다.

Source : https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

2) LightGBM Regressor

실제 데이터 분석 대회에서 가장 많이 사용하는 효과적인 회귀 모델

파라미터 명 (파이썬 래퍼)	파라미터명 (사이킷런 래퍼)	설명
num_iterations (100)	**n_estimators** (100)	- 반복 수행 트리개수 지정 - 너무 크면 과적합 발생
learning_rate (0.1)	**learning_rate** (0.1)	- 학습률
max_depth (-1)	**max_depth** (-1)	- 최대 깊이 - default → 깊이에 제한이 없음
min_data_in_leaf (20)	**min_child_samples** (20)	- 최종 리프 노드가 되기 위한 레코드수 - 과적합 제어용
num_leaves (31)	**num_leaves** (31)	- 하나의 트리가 가지는 최대 리프 개수
boosting ('gbdt')	**boosting_type** ('gbdt')	- gbdt : 일반적인 그래디언트부스팅 트리 - rf : 랜덤포레스트
bagging_fraction (1.0)	**subsample** (1.0)	- 데이터 샘플링 비율 - 과적합 제어용
feature_fraction (1.0)	**colsample_bytree** (1.0)	- 개별트리 학습시 선택되는 피처 비율 - 과적합 제어용
lambda_l2 (0)	**reg_lambda** (0)	- L2 Regularization 적용 값 - 피처 개수가 많을 때 적용을 검토 - 클수록 과적합 감소 효과
lambda_l1 (0)	**reg_alpha** (0)	- L1 Regularization 적용 값 - 피처 개수가 많을 때 적용을 검토 - 클수록 과적합 감소 효과
objective	**objective**	- 'reg:linear' : 회귀 - binary:logistic : 이진분류 - multi:softmax : 다중분류, 클래스 반환 - multi:softprob : 다중분류, 확률반환

- hyper-parameter에 영향을 많이 받기 때문에 parameter tuning이 중요합니다.
- 기존에 많이 쓰는 파라미터 세팅을 기억해두고, 필요에 따라 다양한 조합을 테스트해봅니다.
- 우리는 오픈소스 라이브러리에게 맡깁니다.
- 이러한 방식을 "AutoML"이라고 합니다.

Source : https://injo.tistory.com/48

3

회귀 모델 평가 방법

회귀 모델 평가

머신러닝의 평가 기준은 다양합니다

- 주어진 데이터로 모델을 학습시키는 것은 지정한 성능 평가 지표를 향상시키는 과정입니다.
- 성능 평가 지표의 값은 “예측 성능”을 기준으로 합니다.
- 정량적 기준을 설정하고, 달성할 때까지 모델을 학습시키고 성능을 개선합니다.
- 목표한 성능에 도달한 모델을 실제 서비스에 적용합니다.

성능 평가

대표적인 회귀 모델 평가 지표

1. MSE(Mean Squared Error)
2. RMSLE(Root Mean Squared Log Error)
3. MAE(Mean Absolute Error)
4. R^2 Score(Coefficient of Determination)

성능 평가

대표적인 회귀 모델 평가 지표

1. MSE(Mean Squared Error)
2. RMSLE(Root Mean Squared Log Error)
3. MAE(Mean Absolute Error)
4. R^2 Score(Coefficient of Determination)

성능 평가

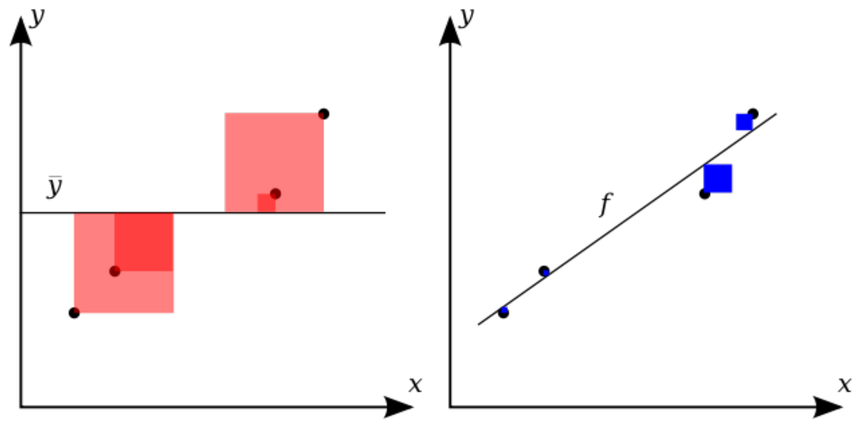
대표적인 회귀 모델 평가 지표

1. MSE(Mean Squared Error)
2. RMSLE(Root Mean Squared Log Error)
3. MAE(Mean Absolute Error)
4. R^2 Score(Coefficient of Determination)

성능 평가

대표적인 회귀 모델 평가 지표

1. MSE(Mean Squared Error)
2. RMSLE(Root Mean Squared Log Error)
3. MAE(Mean Absolute Error)
4. R^2 Score(Coefficient of Determination)



End of Slides