



안전운전자 예측

대회 소개 및 목표

- ▷ 포르토 세구로의 안전 운전자 예측
- ▷ 보험료 최적화 및 고객 서비스 개선
- ▷ 브라질의 자동차 보험 시장에 대한 이해

데이터 이해

- ▷ 훈련 데이터와 테스트 데이터
- ▷ 변수 설명 (운전자의 나이, 성별, 차량 정보 등)
- ▷ 결측치 처리 방법
- ▷ 데이터 탐색 및 시각화

모델링 및 평가 지표

- ▷ 데이터 전처리 및 피쳐 엔지니어링
- ▷ 분류 알고리즘의 선택 (LightGBM, XGBoost)
- ▷ 베이지안 최적화를 활용한 하이퍼파라미터 튜닝
- ▷ OOF 검증 및 앙상블
 - 평가 지표: Normalized Gini Coefficient (정규화 지니 계수)

베이지안 최적화란?

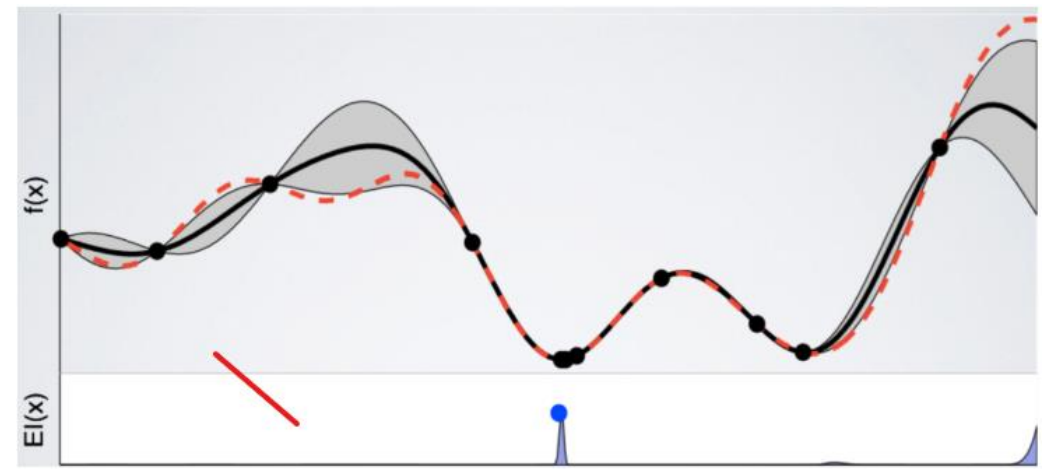
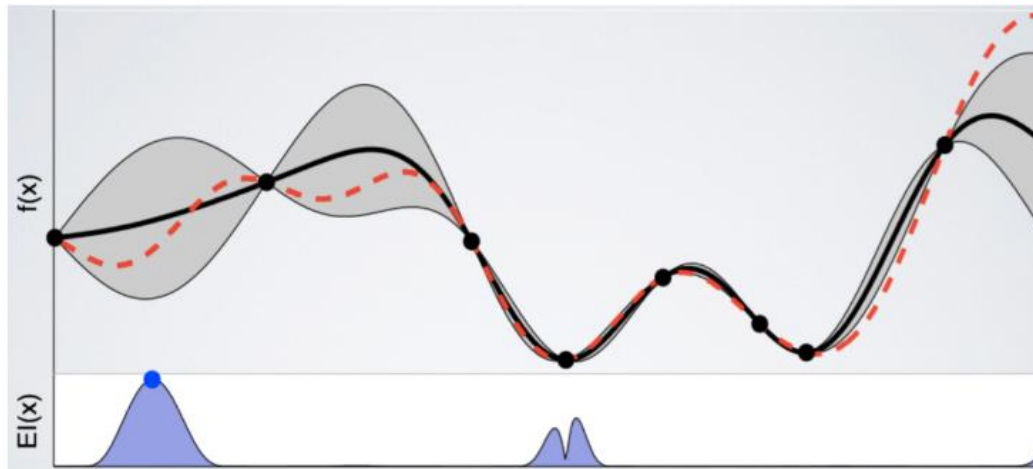
- ▶ 베이지안 최적화(Bayesian Optimization)는 비선형 함수의 최적화 문제를 효율적으로 해결하는 글로벌 최적화 알고리즘입니다. 베이지안 최적화는 머신러닝 모델의 하이퍼파라미터 튜닝, 실험 설계, 조합 최적화 등 다양한 분야에서 사용됩니다. 이 방법은 베이지안 추론과 가우시안 프로세스를 기반으로 하며, 목적 함수의 불확실성을 고려하여 최적의 해를 찾습니다.

베이지안 최적화의 원리

- ▶ 가우시안 프로세스(Gaussian Process) 사용: 가우시안 프로세스는 연속형 변수에 대한 목적 함수를 모델링하는 데 사용됩니다. 가우시안 프로세스는 함수의 평균과 분산을 추정하므로, 불확실성을 고려한 최적화가 가능합니다.

베이지안 최적화의 원리

- ▷ Acquisition Function: 베이지안 최적화에서는 Acquisition Function이라는 보상 함수를 사용하여 다음 탐색 지점을 선택합니다. 이 함수는 탐색과 이용의 균형을 맞추는 역할을 하며, 대표적으로 Probability of Improvement (PI), Expected Improvement (EI), Upper Confidence Bound (UCB) 등이 있습니다.

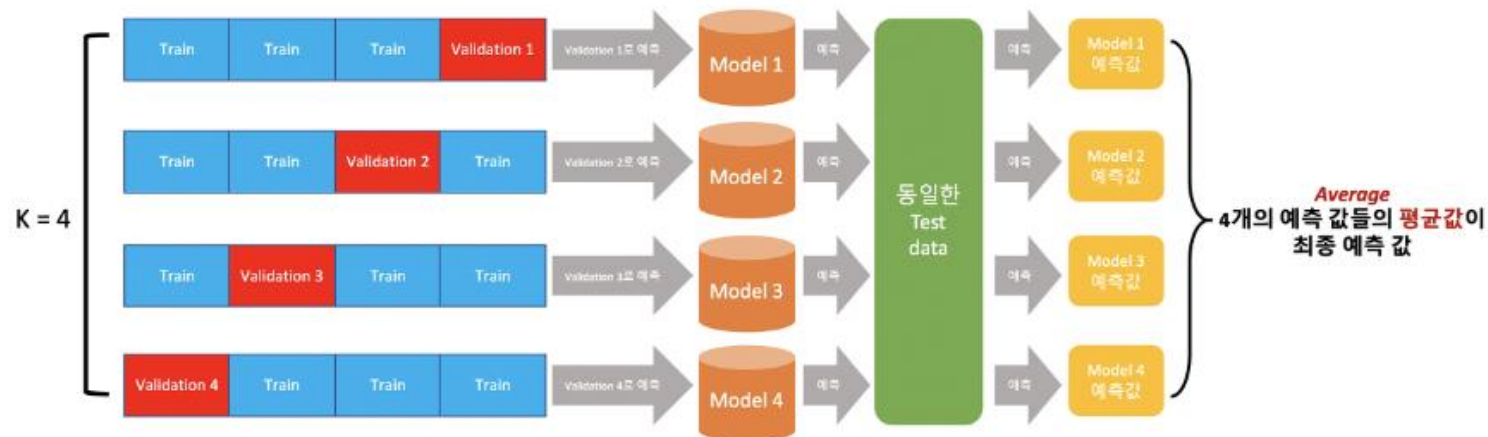


베이지안 최적화의 장점

- ▷ 목적 함수의 불확실성을 고려하여 효율적인 탐색이 가능합니다.
- ▷ 소수의 샘플을 사용하여 글로벌 최적해를 찾을 수 있으므로, 계산 비용이 상대적으로 낮습니다.
- ▷ 노이즈가 있는 목적 함수에 대해서도 안정적인 최적화 결과를 얻을 수 있습니다.

OOF(Out-of-Fold) 검증이란?

- ▶ OOF 검증은 교차 검증에서 각각의 fold에 대한 예측 결과를 모아 하나의 완전한 예측 결과를 만드는 방법입니다. OOF 검증을 사용하면, 각 fold에서 검증용 데이터에 대한 예측 결과를 모아 최종적으로 모델의 성능을 평가할 수 있습니다. 이를 통해 모델의 일반화 성능을 더 정확하게 추정할 수 있습니다.



OOF 방식으로 최종 예측하는 방법

OOF 검증 과정

- ▶ 데이터를 여러 개의 fold로 나눕니다. 일반적으로 5-fold 또는 10-fold 교차 검증을 사용합니다.
- ▶ 각 fold에 대해 모델을 학습하고 검증을 수행합니다. 이때, 각 fold에서 검증용 데이터에 대한 예측 결과를 저장합니다.
- ▶ 모든 fold에 대한 학습과 검증이 끝나면, 저장한 예측 결과를 모아 하나의 완전한 예측 결과를 만듭니다.
- ▶ 이렇게 얻은 예측 결과를 사용하여 모델의 성능을 평가합니다. 이를 통해 모델의 일반화 성능을 추정할 수 있습니다.

OOF 검증의 장점

- ▶ 교차 검증을 통해 얻은 예측 결과를 활용하므로, 모델의 일반화 성능을 더 정확하게 추정할 수 있습니다.
- ▶ 각 fold의 검증 결과를 종합적으로 분석할 수 있어, 모델의 성능 변동이나 특정 fold에서의 문제를 파악하기 쉽습니다.
- ▶ 여러 모델의 결과를 결합하여 앙상블(Ensemble) 기법을 사용할 때, OOF 검증을 활용하여 기여도에 따른 가중치를 부여할 수 있습니다.