모델 성능 해석 및 모델 설명력 QA

Q1. SHAP와 LIME은 어떻게 다르고 어떤 상황에서 각각 쓰는 것이 적합한가요?
A1.
1. SHAP(Shapley Additive exPlanations)와 LIME(Local Interpretable Model-agnostic Explanations)은 모델의 예측 결과를 설명하는 데 사용되는 대표적인 해석 기법입니다.
2. 공통점: 둘 다 블랙박스 모델의 예측 결과를 특성 단위로 해석합니다.
3. 차이점:
- SHAP: 게임 이론 기반. 전체 모델을 설명하며, 모든 가능한 특성 조합의 마진 기여도를 평균하여 계산합니다. 전역 및 국부 해석 모두 가능.
- LIME: 국소 모델 기반. 예측 대상 주변의 데이터만 샘플링하여 선형 모델로 근사합니다. 지역적 해석에 특화됨.
4. 사용 가이드:
- 모델 전반의 변수 중요도 파악 → SHAP 적합
- 개별 예측 건에 대한 빠른 설명 필요 → LIME 적합
5. SHAP은 계산량이 크지만 이론적으로 더 정교하고 안정적입니다. 반면, LIME은 속도는 빠르지만 결과가 불안정할 수 있습니다.

Q2. Permutation Importance는 어떤 개념이며 언제 사용하면 좋나요?
A2.
1. Permutation Importance는 학습된 모델의 각 특성이 예측 성능에 얼마나 기여하는지를 평가하는 방법입니다.
2. 방법:
- 개별 특성의 값을 무작위로 섞어서 모델 예측에 주는 영향을 측정합니다.
- 성능 감소가 클수록 해당 특성이 중요한 것으로 간주합니다.
3. 장점:
- 모델에 독립적 (모델 agnostic)
- 직관적이고 해석이 쉬움
4. 사용 조건:
- 데이터가 준비된 상태에서 학습된 모델의 검증 또는 테스트 성능에 대해 사용
- overfitting 없이 모델이 잘 학습되었을 때 효과적
5. 한계:
- 특성 간 상관관계가 높을 경우 중요도 결과 왜곡 가능
- 계산 시간이 오래 걸릴 수 있음

Q3. 예측 확률과 클래스 예측은 어떻게 다르고, 각각을 어떤 상황에서 사용하는 것이 좋은가요?
A3.
1. 예측 확률: 모델이 각 클래스에 대해 계산한 확률 값 (예: 클래스 A일 확률 0.72)
2. 클래스 예측: 예측 확률 중 가장 높은 값을 가진 클래스를 선택한 결과 (예: A 클래스)
3. 사용 목적:
- 예측 확률: 리스크 기반 의사결정, threshold 조정, 우선순위 정렬 등에 유용
- 클래스 예측: 단순 이진 판단, 최종 라벨링이 필요한 경우
4. 예시:
- 보험 사기 탐지에서 확률이 0.4~0.6이면 수작업 검토, 0.9 이상이면 자동 차단
- 이메일 스팸 필터처럼 확실한 클래스 판단이 필요한 경우에는 클래스 예측 사용
5. 주의사항:
- 불균형 데이터의 경우 확률 해석이 왜곡될 수 있음 → calibration 필요
- 확률 값은 모델 종류에 따라 해석 방식이 다를 수 있음 (예: 로지스틱 회귀 vs 랜덤 포레스트)

Q4. 모델 예측 결과를 도메인 지식과 연결하려면 어떤 점을 고려해야 하나요?
A4.
1. 목적: 모델이 특정한 결과를 예측한 이유를 업무나 현업의 언어로 해석하는 것
2. 방법:
- 모델 설명 결과 (ex. SHAP, Permutation)와 비즈니스 변수의 의미를 연결
- 데이터 수집 방식, 변수 정의, 스케일 등에 대한 도메인 전문가의 의견 반영
3. 적용 예시:
- 통신사 해지 예측 모델에서 '3개월간 데이터 사용량 감소'가 주요 변수일 때 → "데이터 사용 감소 = 관심 저하"로 해석
- 금융에서 '최근 입금 건수 감소'가 주요 변수일 경우 → "현금 흐름 악화"로 해석
4. 고려사항:
- 도메인 변수의 실제 의미와 알고리즘 상 변수 중요도가 불일치할 수 있음
- 도메인 전문가와 협업하여 변수 생성/해석 과정에 피드백을 주고받는 것이 중요
5. 실무 팁:
- 모델 결과를 업무용 용어로 번역해서 설명해야 현업에서 수용됨
- "왜 그런 예측이 나왔는지"를 설명하는 데 도움이 되는 시각화(Tooltip, SHAP summary plot 등)를 함께 제공하면 좋음

Q5. SHAP summary plot은 어떻게 해석하나요?
A5.
1. SHAP summary plot은 각 특성이 모델 예측에 미친 영향력을 시각화한 것입니다.
2. 구성:
- Y축: 주요 feature들 (중요도 순)
- X축: SHAP value (해당 feature가 예측값에 기여한 정도)
- 색상: 각 샘플의 feature 값 (보통 파란색=낮음, 빨간색=높음)
3. 해석법:
- 오른쪽(+)에 위치한 SHAP값 → 예측값을 증가시키는 특성값
- 왼쪽(-)에 위치한 SHAP값 → 예측값을 감소시키는 특성값
- 붉은 점이 오른쪽에 몰려 있으면, 값이 클수록 예측값 증가
4. 사용 목적:
- 전반적인 feature 영향력 파악
- 특정 특성값이 예측에 어떤 방식으로 작용하는지 확인
5. 한계:
- 해석에 도메인 지식 필요
- 고차원일수록 해석 복잡

Q6. SHAP decision plot은 언제 쓰고 어떻게 해석하나요?
A6.
SHAP decision plot은 예측값이 어떤 경로를 거쳐 결정되었는지를 누적적으로 보여주는 시각화입니다.
X축: 특성 추가 순서
Y축: 누적된 예측값
색상: 각 특성 (feature)
사용 예:
개별 예측값이 어떻게 baseline(평균 예측값)에서 최종값으로 이동했는지를 설명
어떤 특성이 예측값을 끌어올렸고, 어떤 특성이 끌어내렸는지를 확인
적합한 상황:
개별 사례에 대한 예측 근거 설명이 필요한 경우
이탈 고객 예측, 이상거래 탐지 등
한계:
여러 특성이 얽힌 복잡한 경우에는 시각화가 난해해질 수 있음

Q7. SHAP의 baseline 값은 무엇이며 왜 중요한가요?
A7.
1. SHAP의 baseline은 모든 특성이 없다고 가정했을 때의 모델 예측값입니다. (보통 평균값)
2. SHAP value는 baseline으로부터 얼마만큼 각 특성이 예측값을 이동시켰는지를 계산합니다.
3. 중요성:
- baseline이 기준이 되어, 모든 SHAP값의 합은 예측값과 정확히 일치
- 예측값 = baseline + Σ(SHAP values)
4. 적용:
- 회귀: 평균 예측값
- 분류: 로짓(logit) 기반의 확률 변환값
5. 실무 팁:
- baseline 값이 의미 있는 기준이 되도록 전처리를 신중히 해야 해석력이 높아짐

Q8. 모델의 feature importance와 SHAP 값은 어떻게 다른가요?
A8.
1. Feature importance: 모델 학습 과정에서 계산된 변수 중요도 (e.g., Gini gain, split count)
2. SHAP 값: 예측값에 대해 각 변수의 실제 기여도를 해석적으로 계산
3. 차이점:
- Feature importance는 전역적, SHAP은 전역 + 국부
- Feature importance는 상관된 변수끼리 중요도를 나누는 데 약함
- SHAP은 feature 간 상호작용을 고려해 더 정밀한 해석 제공
4. 사용 팁:
- 빠른 전체 중요도 파악 → feature importance
- 신뢰도 높은 설명 필요 시 → SHAP

Q9. LIME의 해석 결과는 얼마나 신뢰할 수 있나요?
A9.
1. LIME은 예측값 근방의 샘플을 생성하여 국소 모델(보통 선형 회귀)로 해석합니다.
2. 신뢰도 요인:
- 국소 근사이므로 예측 포인트 주변에서만 신뢰 가능
- 샘플링 및 random seed에 따라 결과 변동 가능
3. 한계:
- 복잡한 경계 근처에서는 근사가 부정확할 수 있음
- 샘플 수, kernel width 등에 따라 결과 민감
4. 개선 팁:
- 반복 실행 후 평균값 비교
- 비슷한 결과를 다른 해석기법(SHAP 등)과 교차 검증
5. 적합한 경우:
- 실시간 또는 빠른 해석이 필요한 업무

Q10. 모델 해석 결과를 시각화할 때 어떤 점을 주의해야 하나요?
A10.
1. 목적: 모델 해석 결과를 현업이 쉽게 이해할 수 있도록 직관적으로 전달하는 것
2. 주의점:
- feature 명칭은 도메인 용어로 번역하여 사용
- 중요도 순서 강조, 양/음의 방향성 명확히 구분
- 시각화에서 단위, 범위, 기준선을 명확히 표시
3. 도구별 팁:
- SHAP summary plot → 색상 해석 가이드 필요
- waterfall/decision plot → 누적 효과 강조
4. 실무 활용:
- 리포트, 프레젠테이션에 삽입 시, 표제어(예: “주요 해지 요인 TOP5”)로 설명 추가
5. 궁극적 목표는 ‘그래서 왜 그런 예측이 나왔는지’를 한눈에 이해시키는 것

Q11. 왜 SHAP이 최근 모델 해석 분야에서 표준처럼 쓰이나요?
A11.
1. 이론적 기반: 게임이론(Shapley value) 기반으로 공정성(fairness) 보장
2. 강점:
- 모델 불문하고 사용 가능 (model-agnostic)
- 전역 및 국부 해석 모두 가능
- SHAP값 합이 예측값과 정확히 일치
3. 실제 적용성:
- 금융, 의료, 제조 등 고위험 영역에서 신뢰성 확보 수단으로 채택
- 규제 대응 (모델 투명성 확보)
4. 단점:
- 계산량 큼 → 대규모 데이터엔 속도 병목
5. 요약:
- 해석력 + 정량성 + 신뢰성을 모두 만족시키는 유일한 프레임워크