

머신러닝 문제 해결 체크리스트

데이터 둘러보기(구조 탐색)

- 파일별 용도 파악
- 데이터 양(레코드수, 피쳐수, 전체용량등)
- 피쳐이해 (이름, 의미, 데이터 타입, 결측값 개수, 고윳값 개수, 실제값, 데이터 종류 등)
- 훈련 데이터와 테스트 데이터 차이
- 타깃값: 제출 해야하는 값

데이터 시각화

- 각종 시각화



- 피쳐파악 → 추가할 피쳐:
 - 제거할 피쳐:
 - 피쳐별 인코딩 전략:
- 이상치 파악 → 해당 피쳐별 처리 방법(결과물: 추가/제거 피쳐목록, 인코딩 전략, 이상치 처리 전략)

준비하기

- 데이터 불러오기
- (필요시) 기본적인 피쳐 엔지니어링
- 평가 지표 계산 함수 준비 → 결과물: 데이터, 평가지표 계산 함수

모델훈련

- 모델 생성
- 훈련 -> 결과: 훈련된 베이스라인 모델
- 피처이해 (이름, 의미, 데이터 타입, 결측값 개수, 고윳값 개수, 실제값, 데이터 종류 등)

성능검증

- 예측(검증 데이터 사용)
- 평가

예측 및 결과 제출

- 최종 예측(테스트 데이터 사용)
- 제출 파일 생성
- 제출 -> 결과물: 제출파일, 기준 **private/public** 점수

피처 엔지니어링

- | | |
|-----------|------------------------|
| □ 이상치 제거 | □ 시차 피처 생성(시계열 데이터 한정) |
| □ 결측값 처리 | □ 피처 스케일링 |
| □ 데이터 인코딩 | □ 피처명 한글화 |
| □ 타입 변경 | □ 데이터 다운캐스팅 |
| □ 파생피처 생성 | □ 데이터 조합 생성 |
| | □ 필요없는 피처 제거 |

모델 훈련 with 하이퍼파라미터 최적화

- 하이퍼파라미터 종류와 의미 파악
- 선별
 - 최적화할 하이퍼파라미터:
 - 값을 고정할 하이퍼파라미터:
- 값 범위 설정
- 최적화 기법: (그리드 서치, 베이지안 서치, OOF 예측 등)
- 모델 생성 및 훈련(최적화) -> 결과물: 최적하이퍼파라미터, 훈련된 모델

성능 검증

- 예측(검증데이터 사용)
- 성능 평가-> 결과물: 예측결과, 검증 평가 점수

*** 만족스러운 결과가 나올때까지 피처 엔지니어링, 훈련(다른 모델로 교체 포함), 성능 검증 반복**