머신러닝 문제해결 체크리스트 (beta)

▼ *문제(경진대회) 이해*

제목 :

미션 :

문제 유형 : 회귀 / 이진분류 / 다중분류 / (기타 :)

평가지표 :

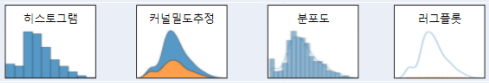
▼ *탐색적 데이터 분석*

데이터 둘러보기(구조 탐색)

- ☐ 파일별 용도 파악
- ☐ 데이터 양(레코드 수, 피처 수, 전체 용량 등)
- ☐ 피처 이해(이름, 의미, 데이터 타입, 결측값 개수, 고윳값 개수, 실제값, 데이터 종류 등)
- ☐ 훈련 데이터와 테스트 데이터 차이
- ☐ 타깃값 : 제출(예측)해야 하는 값

데이터 시각화

- ☐ (필요 시) 효과적인 시각화를 위한 피처 엔지니어링
- ☐ 각종 시각화
  - ☐ 수치형 데이터 시각화



  - ☐ 범주형 데이터 시각화

  - ☐ 데이터 관계 시각화

- ☐ 피처 파악
  - ☐ 추가할 피처 :
  - ☐ 제거할 피처 :
  - ☐ 피처별 인코딩 전략 :
- ☐ 이상치 파악
  - ☐ 해당 피처별 처리 방법

결과물 : 추가/제거 피처 목록, 인코딩 전략, 이상치 처리 전략

▼ *베이스라인 모델*

준비하기

- ☐ 데이터 불러오기
- ☐ (필요 시) 기본적인 피처 엔지니어링
- ☐ 평가지표 계산 함수 준비

결과물 : 데이터, 평가지표 계산 함수

모델 훈련

- ☐ 모델 생성
- ☐ 훈련

결과 : 훈련된 베이스라인 모델

성능 검증

- ☐ 예측(검증 데이터 사용)
- ☐ 평가

결과물 : 예측 결과, 검증 평가 점수

예측 및 결과 제출

- ☐ 최종 예측(테스트 데이터 사용)
- ☐ 제출 파일 생성
- ☐ 제출

결과물 : 제출 파일, 기준 private/public 점수

▼ *성능 개선*

피처 엔지니어링

- ☐ 이상치 제거
- ☐ 결측값 처리
- ☐ 데이터 인코딩
- ☐ 타입 변경
- ☐ 파생 피처 생성
- ☐ 시차 피처 생성(시계열 데이터 한정)
- ☐ 기타 :
- ☐ 피처 스케일링
- ☐ 피처명 한글화
- ☐ 데이터 다운캐스팅
- ☐ 데이터 조합 생성
- ☐ 필요 없는 피처 제거

결과물 : 피처 엔지니어링된 훈련 데이터와 검증 데이터

모델 훈련 with 하이퍼파라미터 최적화

- ☐ 하이퍼파라미터 종류와 의미 파악
- ☐ 선별
  - 최적화할 하이퍼파라미터 :
  - 값을 고정할 하이퍼파라미터 :
- ☐ 값 범위 설정
- ☐ 최적화 기법 : (그리드서치, 베이지안서치, OOF 예측 등)
- ☐ 모델 생성 및 훈련(최적화)

결과물 : 최적 하이퍼파라미터, 훈련된 모델

성능 검증

- ☐ 예측(검증 데이터 사용)
- ☐ 성능 평가

결과물 : 예측 결과, 검증 평가 점수

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

예측 및 결과 제출

- ☐ 최종 예측(테스트 데이터 사용)
- ☐ 제출 파일 생성
- ☐ 제출

결과물 : 제출 파일, 최종 private/public 점수

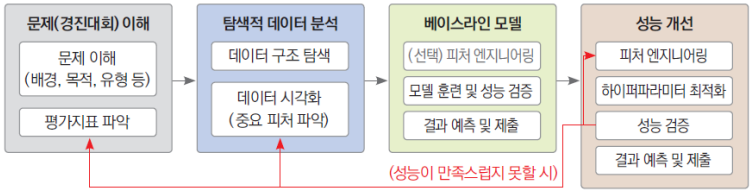
▼ *안내*

머신러닝 문제를 해결하는 과정에서 점검해야 할 사항을 단계별로 정리했습니다. 이를 기초로 여러분의 경험과 노하우를 녹여 더 풍성하고 유용한 체크리스트로 발전시켜 활용하시기 바랍니다.

* 이 체크리스트는 신백균 저자의 《머신러닝·딥러닝 문제해결 전략》 부록입니다.

▼ *문제해결 프로세스*

머신러닝 문제(캐글 머신러닝 경진대회)를 해결하는 프로세스는 다음과 같습니다. 문제에 따라 이상적인 세부 내용은 조금씩 다를 수 있지만 큰 흐름과 구조는 대부분 비슷합니다.



- **1단계 : 문제(경진대회) 이해** - 어떤 일이든 주어진 문제를 이해하는 데서 시작해야 합니다. 문제를 정확하게 이해해야 목표점을 정확히 설정할 수 있습니다.
- **2단계 : 탐색적 데이터 분석** - 주어진 데이터를 면밀히 분석합니다. 머신러닝은 결국 데이터를 다루는 기술이므로 데이터를 잘 알아야 다음 단계에서 가장 효과적인 모델을 찾고 최적화할 수 있습니다.
- **(데이터 전처리)** : 머신러닝에 이용되는 현실 세계 데이터에는 다양한 잡음이 섞여 있고 형태도 일정하지 않아서 솎아주거나 형태를 일치시키는 등의 전처리 작업을 해줘야 합니다. 다만 캐글 경진대회들은 대부분 전처리가 상당 수준 이루어진 데이터를 제공하므로 이 체크리스트에서는 전처리 단계를 따로 구분해 설명하지 않았습니다.
- **3단계 : 베이스라인 모델** - 기본 모델을 만들어봅니다. 첫 술에 배부를 수 없으니 시작부터 최고 성능의 모델을 시도해도 성공하기 어렵습니다. 또한 기본 모델이 있어야 최적화 기법 적용 후 얼마나 더 좋아졌는지 비교해볼 수 있습니다.
- **4단계 : 성능 개선** - 다양한 아이디어를 적용해 모델의 성능을 끌어올립니다. 창의력이 가장 필요한 단계입니다. 하나의 모델을 점진적으로 개선해볼 수도 있고, 여러 가지 모델을 시도해볼 수도 있습니다. 데이터 자체를 가공해도 성능이 좋아질 수 있습니다. 무언가 놓친 것 같다면 1단계나 2단계로 돌아가 문제와 데이터를 다시 살펴보는 것도 좋습니다.

체크리스트 개선 아이디어는 bylee@goldenrabbit.co.kr 로 부탁드립니다. (_ _)