TEAM PROJECT

딥러닝 & 머신러닝 팀프로젝트

팀 오스트랄로코딩쿠스　팀장: 이승열 | 서기 : 정재석 | 팀원 : 오태우, 서승화

목차

프로젝트의 목적

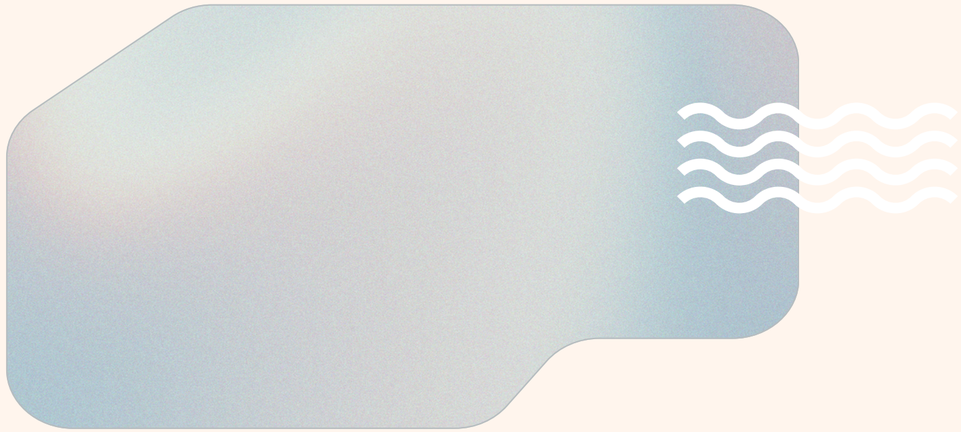
프로젝트의 목적은 주어진 데이터셋 (타이타닉 데이터셋) 을 활용하여
그 데이터를 분석하고 전처리 과정을 거쳐 결과적으로 생존자를 예측할 수 있게
모델을 학습시키는 것!

프로젝트의 조건

프로젝트의 조건은

1. seaborn 라이브러리를 사용해서 데이터셋을 불러올줄 알아야 한다.
2. 데이터의 feature를 분석 할 줄 알아야 한다.
3. 분석한 데이터를 전처리 할 줄 알아야 한다.
4. 각각 모델의 특성에 맞게 올바르게 학습 시킬 줄 알아야 한다.

#feature #engineering #model



팀장 이승열

맡은 역할

- 팀장 👍
- SA Document 관리 👀
- (필수과제) 데이터셋 불러오기 & 데이터 전처리 & 모델 학습 준비
- (도전과제) 리뷰 예측 모델 학습시키기
- (발표) 발표 ppt 준비 및 발표 🤦

서기 정재석

맡은 역할

- 서기 📃
- 팀 미팅 문서 관리
- (필수과제) 로지스틱 회귀 모델
- (도전과제) NLP
- (발표) 자료준비 💻

팀원 서승화

맡은 역할

- 팀원 👷
- README.md
- (필수과제) XGBoost
- (도전 과제) 워드 클라우드
- (발표) 발표 ppt 준비 및 자료정리 💻

팀원 오태우

맡은 역할

- 팀원 👷
- (필수 과제) 랜덤 포레스트
- (도전 과제) 데이터 전처리 와 feature분석
- (발표) 자료준비 💻

팀 과제를 할때의 우리 팀의 순서

팀과제의 가장 중요한 점은 팀워크 입니다.
우리 오스트랄로코딩쿠스팀은 이렇게 진행을 하였습니다.

팀 미팅 진행하기

팀 미팅을 통하여 각자의 의견을
공유할 수 있는 자리 마련

의견 나누기

하나의 주제 또는 미리 설정된 주제에
대해서 다른 사람의 의견을 들어보기

결정 하기

가장 합리적이고 효율적인 방면으로
의견을 수렴하여 진행하기

피드백 나누고 반복하기

진행 해보고 피드백 해보며
이과정을 반복하여 좀더 성장하기

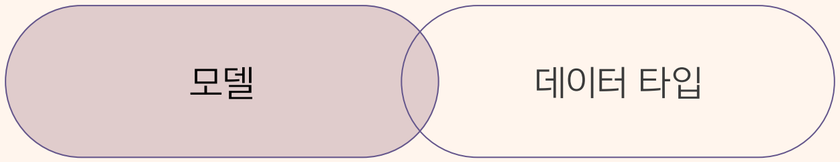
모든 미팅 미닛은 SA Document에 있습니다.

우리가 집중을 한 것

이번 팀프로젝트는 모든 팀이 같은 과제를 진행함으로

저희는 체크리스트로 간단히 기능 확인을 한 후,

데이터 타입과 모델간의 관계에 대해서 집중을 하기로 했습니다.



필수 과제

체크리스트		완료
seaborn 라이브러리를 사용하여 데이터셋 불러오기	→	완료
head와 describe 를 사용해보고 주석으로 간단히 설명하기	→	완료
isnull()과 sum()함수를 이용해서 결측치 갯수를 확인하기	→	완료
확인한 결측치를 중앙값과 최빈값으로 대체하기	→	완료
카테고리 데이터를 수치형 데이터로 인코딩 하기	→	완료
기존의 feature를 가지고 새로운 feature 생성하기	→	완료
feature와 target을 분리하여 학습 준비 하기	→	완료
로지스틱 회귀 함수 모델을 학습 시키기	→	완료
랜덤 포레스트 모델을 학습 시키기	→	완료
XGBoost 모델을 학습 시키기	→	완료

03 필수 과제 - 모델의 구조

모델의 구조

데이터 불러오기

데이터를 불러오고 그 데이터셋 안의 데이터타입과 형태를 읽는다.
1.문자열 데이터
2.카테고리 데이터

↓

전처리 하기

문자열 데이터처리
카테고리 데이터 처리

↓

모델에 학습시키기

각 모델에 맞게 데이터 준비시키기
모델에 맞는 함수를 사용하여 모델을 학습시키고
예측값을 도출 시키기

필수과제 - 모델 평가

각 모델의 특징과 학습시킨 결과를 보도록 하자.

Logistic Regression Model

특징이 명확한 feature 와 label이 존재해야 한다.
이진 분류 문제를 해결하는데 주로 사용된다. (생존 or 불가)

로지스틱 회귀 모델은 구조가 간단하여 이해가 쉽고 기본적인 통계학 지식으로도 해석 하기가 용이하다. 타이타닉 데이터셋 과 같이 어느정도의 큰 사이즈의 데이터 학습도 효율적으로 할 수 있고 L1, L2 정규화를 통해 과적합을 방지 할 수 있다.

```
Accuracy: 0.8044692737430168
Classification Report:
              precision    recall  f1-score   support

           0       0.82      0.86      0.84       105
           1       0.78      0.73      0.76        74

    accuracy                           0.80       179
   macro avg       0.80      0.79      0.80       179
weighted avg       0.80      0.80      0.80       179
```

필수과제 - 모델 평가

Random Forest Model

많은 특성을 가진 고차원 데이터를 잘 처리할 수 있다.
숫자형, 범주형 과 같은 다양한 데이터 타입도 활용 할 수 있다.

랜덤 포레스트 모델은 여러개의 결정 트리를 조합하여 예측하는 앙상블 모델이다. 각 트리는 데이터의 서브셋을 사용하여 훈련되고, 최종 예측은 모든 트리의 예측 결과를 평균하여 결정된다.

```
Accuracy: 0.8212290502793296
Classification Report:
              precision    recall  f1-score   support

           0       0.85      0.85      0.85       105
           1       0.78      0.78      0.78        74

    accuracy                           0.82       179
   macro avg       0.82      0.82      0.82       179
weighted avg       0.82      0.82      0.82       179
```

XGBoost Model

숫자형, 범주형 번수로 구성된 데이터셋을 사용하여 학습한다.
자동으로 중요한 특성을 선택하고, 덜 중요한 특성을 무시한다.

자동으로 결측치를 처리하여 사용하기가 편하다.
특성이 많은 복잡한 데이터셋에도 아주 효과적으로 학습할 수 있다.
이상치가 있는 데이터에도 강인하게 적용이 가능하여 매우 높은 정확도를 자랑한다.

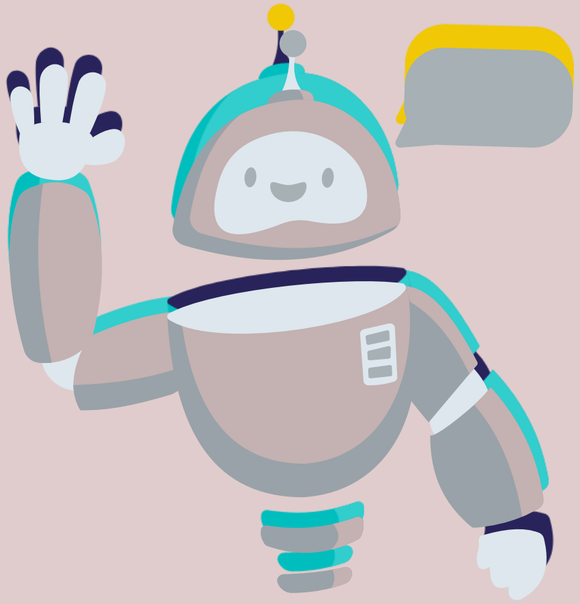
```
XGBoost 모델의 MSE: 0.12981004899201257
Accuracy: 0.8101
              precision    recall  f1-score   support

           0       0.80      0.90      0.85       105
           1       0.82      0.69      0.75        74

    accuracy                           0.81       179
   macro avg       0.81      0.79      0.80       179
weighted avg       0.81      0.81      0.81       179
```

필수과제 - 트러블 슈팅

에러가 발생한 부분	해결방법
ValueError : Input contains NaN	오버슈팅으로 인한 에러 > learning_rate 조정하기
family size : NaN	family size를 계산할때, 본인은 계산에서 제외시켜 생긴 문제 > 생성 조건을 달고 +1을 함
ConvergenceWarning: Liblinear failed to converge	max_iter피라미터를 증가, 데이터 스케일링 추가 및 정규화 조정 등의 방법으로 해결
failed to push some refs to 'origin	git pull origin main, git merge -X theris branch, git push origin main

↓

각자 많은 에러 코드와 의도에 맞지 않는 코드가 있었지만,
시간 관계상 각자 1개의 트러블 슈팅케이스를 소개 하기로 하였음.



도전 과제 살펴보기

팀 프로젝트를 끝내며...

오태우

머신러닝, 딥러닝 모두 처음 해보는 거라 강의 내용을 쫓아가기 어려웠다. 과제 가이드를 보고 막막했는데 협업을 통해 진행하니 다행히 마칠 수 있었고, 도전 과제도 건드려볼 수 있었다.
깃허브를 이용한 협업도 처음이다 보니 우여곡절이 있었는데 팀장님과 튜터님 지도 하에 꽤 효율적으로 파트를 나눌 수 있었고 다음 팀 프로젝트에는 이러한 경험을 살려 더 체계적인 진행을 할 수 있을 것 같다.

이승열

강의가 어려워 진도도 따라가기 어려웠지만,
과제를 하면서 팀원들과 의견을 나누면서 새로운 지식을 알게되고 좀더 발전할 수 있는 기회여서 좋았던 것 같다.
다들 열심히 했고, 끝가지 포기하지않고 도전과제까지 무사히 마쳐서 대단하다고 생각한다.
내배캠 끝날때 까지 모두 생존하기를!

정재석

비전공자라서 모두 생소한 강의 내용이었는데, 강의를 들으면 생각보다 요약된 내용이 많다 보니 팀원들과 같이 의논하거나, 튜터님에게 질문하여 배워 나가는 것이 좀 더 많았다. 그리고 과제를 통해 직접 코드를 작성해나가다 보니 모르는 코드들을 직접 활용하면서 뜻을 이해하니 기억에 남았다. github도 어려웠지만 팀원이나 튜터님에게 도움을 받아 어느 정도 활용할 줄 알게 되었다. 이를 통해 다음 팀 프로젝트에는 좀 더 발전된 모습으로 나아갈 수 있을 것 같다.

서승화

github를 통해서 팀프로젝트를 처음 진행 해보았고,
conflict도 많이 나서 고생도 좀 했지만 알찬 경험이었다

이것으로 마치겠습니다

Q&A

감사합니다

팀 오스트랄로코딩쿠스　팀장: 이승열 | 서기 : 정재석 | 팀원 : 오태우, 서승화