Data Scientist 이영훈

- 📂 Portfolio : 🔗 https://github.com/ANGHOOO
- 📲 Mobile : 010-6449-6063
- ✉️ Email : leeanghoo2@gmail.com
- 🪧 Blog : 🔗 https://anghoo-ai.tistory.com/

Core Competency

[체계적 접근을 통한 모델 성능 고도화 능력]

- Upstage가 보유한 실제 문서 데이터를 바탕으로 17개의 문서를 분류 하는 모델을 개발하였습니다. 머신러닝 프로세스를 반복적으로 수행하며 모델 성능 향상을 위한 다양한 가설을 수립하고, 원인을 분석하였습니다. 해당 경험을 바탕으로 Kaggle의 Harmful Brain Activity Competition에서 동메달을 수상하였습니다.

[데이터 기반 문제해결 및 분석능력]

- 구미시 상징 조형물 건설을 위해 텍스트 마이닝을 통하여 구미시에 대한 사람들의 인식을 조사한 경험이 있습니다. 문제 해결을 위한 데이터 수집 방법부터 모델 선택까지 고민 하였습니다. 데이터를 기반으로 구미시에 대한 사람들의 인식을 조사 하였고, 17개의 토픽 클러스터를 도출 하며 실제 문제 해결에 기여 하였습니다.

[문제 해결력, 집념, 분석적 사고력]

- 대구시 공공 데이터를 활용한 교통안전 프로젝트에서 보안상의 이유로 활용하지 못하는 데이터가 있음을 확인하고, 직접 18곳의 현장을 답사하여 데이터를 수집해 팀에 기여한 경험이 있습니다. 팀원들과 지속적으로 소통하며 4가지 개선안을 도출하며 프로젝트를 마무리 하였습니다.

[성실성, 성취 지향성]

- 관심있고 하고자 하는 일에 대해 끈기가 강합니다. 공학수학, 머신러닝, 딥러닝 과목에서 수석을 차지하여 이를 인정받아 후배들에게 해당 교과목의 멘토링을 진행한 경험이 있습니다. 어려운 내용을 다른 사람들에게 효과적으로 쉽게 전달할 수 있습니다.

Projects

[Kaggle] Harmful Brain Activity Classification (개인 프로젝트)

중환자의 뇌파 신호로부터 발작 및 유해한 뇌 활동을 감지하고 분류하는 모델 개발

2024.01.08 - 2024.04.08

- **신호 처리를 위한 딥러닝 모델 개발 및 성능 고도화**
  - **[데이터 EDA 및 전처리]** EDA 과정에서 진단 결과의 두 그룹 분리 및 각 그룹 내 클래스 불균형 문제 발견, KL-Divergence 평가 지표 고려 시 중요성 인지
  - **[모델링 방향성 결정]** CatBoost 모델을 활용하여 CNN 모델에서 중요한 변수 파악, 이를 기반으로 뇌파 신호를 스펙트로그램으로 변환하여 2D 이미지로 취급
  - **[모델 학습 및 성능 고도화]** 사전 학습된 2D backbone 모델 활용, 2-stage training, pseudo labeling, knowledge distillation 기법 적용
  - **[모델 선정]** Local KL-Divergence, Public KL-Divergence에서 좋은 성능을 보인 모델 3개를 각각 선정하여 Ensemble한 모델을 최종 제출
- **프로젝트 결과**
  - KL-Divergence : 0.363146 | PRIVATE Leaderboard 220/2767

[AI Stages] Dialouge Summarization (팀 프로젝트)

학교 생활, 직장, 치료, 쇼핑, 여가, 여행 등 광범위한 일상 생활 중 하는 대화들에 대해 요약하는 모델 개발

[담당 업무] KoBART 및 다양한 사전학습 모델들을 활용한 베이스라인 구축 및 하이퍼파라미터 튜닝, 외부 데이터를 활용한 데이터 증강

2024.03.08 - 2024.03.20 | 🔗 [Github]

- **Hugging Face에 등록된 언어모델 활용 및 성능 고도화**
  - **[복수의 사전학습 모델 실험]** kobart-base-v1, kobart-summary-v1, kobart-summary-v2, bart-r3f 등 한국어 요약 태스크에 특화된 모델들을 학습 데이터셋에 대해 파인튜닝하여 각 모델의 성능 비교
  - **[데이터 증강 기법 적용]** Back Translation(한-영, 영-한), 외부 데이터(AI Hub 한국어 대화 요약 데이터) 활용 등 다양한 증강 기법 적용으로 학습 데이터 확장
  - **[성능 고도화]** 정규표현식을 통한 특수문자, 이모티콘 제거 및 문장 분절, Komoran, KKma, Mecab 형태소 분석기를 활용한 토큰화 및 불용어 제거
  - **[모델 선정]** Local Final Result, Public LB Final Result가 가장 높은 두 개의 모델을 최종 모델로 선정 후 제출
- **프로젝트 기획서와 발표자료 취합**
- **프로젝트 결과**
  - Final Result : 39.1409 | Rouge1 : 0.4941 | Rouge2 : 0.2850 | Rouge L : 0.3951 | PRIVATE Leaderboard 8/8

[AI Stages] Document Type Classification (팀 프로젝트)

금융, 보험, 물류, 의료 등 다양한 도메인의 문서 이미지를 분류하는 모델 개발

[담당 업무] ResNet 및 EfficientNet 모델을 활용한 베이스라인 구축 및 하이퍼 파라미터 튜닝, TTA를 이용한 성능 고도화

2024.02.05 - 2024.02.19 | 🔗 [Github]

- **사전학습 모델을 활용한 딥러닝 모델 제작 및 성능 고도화**
  - **[복수의 사전훈련 모델 선정]** ResNet50을 베이스라인 모델로 선정 후, EfficientNet, ViT 계열 모델을 기반으로, 파라미터 개수 대비 분류 성능이 좋은 모델 선정
  - **[EDA 및 데이터 증강]** Train/Test 데이터 분포 불균형 확인 및 Test 데이터에 적용된 노이즈 분석, Albumentation, Augraphy 라이브러리를 이용한 오프라인 증강 적용
  - **[모델 학습 및 성능 고도화]** Learning Rate Scheduler 적용 및 WandB를 활용한 하이퍼파라미터 튜닝 수행, Ensemble, TTA 적용으로 성능 향상
  - **[모델 선정]** Local CV F1-Score, Public LB F1-Score가 가장 높은 두 개의 모델을 최종 모델로 선정 후 제출
- **프로젝트 기획서와 발표자료 취합**
- **프로젝트 결과**
  - F1-Score : 0.9400 | PRIVATE Leaderboard 5/9

[AI Stages] House Price Prediction (팀 프로젝트)

서울시 아파트 실거래가 매매 데이터를 기반으로 아파트 가격을 예측하는 모델 개발

[담당 업무] EDA, Feature Engineering, 데이터 전처리, LightGBM 모델을 이용한 베이스라인 코드 작성, 고가 아파트 모델 성능 고도화

2024.01.15 - 2024.01.29 | 🔗 [Github]

- **시계열적 특성을 반영한 예측 모델 제작 및 성능 고도화**
  - **[데이터 EDA 및 전처리]** 서울시 건축물 대장 데이터를 활용한 결측치 처리, Box plot을 통한 데이터 이상치 처리
  - **[Feature Engineering]** 한강과의 거리, 1인당 구별 평균 급여, 전용면적 타입 비율, 버스 정류장 개수, 가장 가까운 학교와의 거리 등 새로운 피처 생성
  - **[모델 학습 및 성능 고도화]** 저가(30억원 미만) 아파트 모델과 고가(30억원 초과) 아파트 모델 분리하여 학습 및 예측 수행, 시계열적 특성을 반영한 Time-Series Split
  - **[모델 선정]** Local RMSE, Public RMSE가 가장 낮은 두 개의 모델을 최종 모델로 선정 후 제출
- **프로젝트 기획서와 발표자료 취합**

- **프로젝트 결과**
  - RMSE : 86148.2275 | PRIVATE Leaderboard 2/9

[SoDA Lab] 구미IC 상징 조형물 리서치 프로젝트 (개인 프로젝트)

텍스트 마이닝을 통하여 구미시에 대한 사람들의 인식을 조사하는 프로젝트

2022.10 - 2022.11 | 🔗 [Github]

- **사람들의 인식 조사를 위한 감정 분류 모델 제작 및 클러스터링**
  - **[데이터 수집]** 일반인들이 가장 많이 접하는 매체가 뉴스 기사이므로, 이러한 뉴스기사가 일반인들의 인식을 형성하는 데 큰 영향을 미친다는 가정하에 크롤링과 빅카인즈를 통한 뉴스 기사 데이터 수집
  - **[데이터 전처리 및 기초통계량 분석]** 뉴스 기사의 본문 글만 나타내는 컬럼을 추출, 언어 모델 학습을 위해 KoNLPy를 사용하여 형태소 토큰화 수행, RANKS NL에 등록된 한국어 불용어 리스트를 활용하여 불용어 제거
  - **[뉴스 기사의 주요 토픽 파악]** BERT 기반의 토픽 모델링 기법인 BERTopic을 활용하여 구미시 관련 뉴스 기사의 토픽을 파악
  - **[뉴스 기사의 중립/긍정/부정 분류]** 기사의 논조를 살펴보는 것이 중요하다고 판단하여 KLUE-BERT를 사용하여 뉴스 기사 감정 분석 수행
- **결과 보고서 작성 및 다른 방법론을 사용하여 분석한 보고서와 취합**
- **프로젝트 결과**
  - 토픽으로 분류된 뉴스 기사 중 73%에 해당하는 뉴스 기사가 지역산업 및 정치에 해당하는 뉴스
  - 산업 투자 및 기업 육성 관련 뉴스 토픽 클러스터의 긍적적인 기사가 매우 많은 것을 확인
  - 이를 통해, 구미시에 관한 인식은 역동적인 산업도시, 투자가 필요한 도시, 대기업과 상생하는 도시 등의 이미지라 추측

Awards

Kaggle - Harmful Brain Activity Classification Competition

2024.04

- 개인으로 참가하여 동메달 수상 (220/2767)

Training

패스트캠퍼스 Upstage AI Lab Bootcamp(1기)

2023.10 - 2024.05 | 🔗 커리큘럼

- 회귀모델, 의사결정나무, KNN, XGBoost 등 기본적인 머신러닝 알고리즘을 적용하여 예측 모델을 구축할 수 있습니다.
- CNN, Image Classification, Object Detection, Segmentation 등 컴퓨터 비전 기본 모델을 구현하고 활용할 수 있습니다.
- Transformer, DETR, SegFormer, ViT 등 Transformer 기반 컴퓨터 비전 모델을 이해하고 적용할 수 있습니다.
- 언어학 기초 지식을 바탕으로 텍스트 전처리를 수행하고, 자연어 이해와 생성 작업을 할 수 있습니다.
- BERT, GPT 등 NLP 모델을 파인튜닝하여 자연어 처리 태스크의 성능을 개선할 수 있습니다.

네이버 부스트코스 PY4E 2022

2022.07 - 2022.08

- 파이썬 언어의 기본 문법과 핵심 개념을 이해하고 활용할 수 있습니다.
- 반복문, 조건문, 함수, 클래스 등 프로그래밍의 기본 구조를 파악하고 프로그램을 설계할 수 있습니다.
- 파일 읽기/쓰기, 데이터 구조, 데이터베이스 등을 활용하여 데이터를 처리하고 관리할 수 있습니다.
- 프로젝트 수행을 통해 문제 해결 능력을 기르고, 팀 협업과 코드 리뷰 경험을 쌓을 수 있습니다.

Studies

글또(9기)

2023.12 - 2024.04

- 지식 공유를 위해 학습한 내용과 논문 리뷰 내용을 독자 대상에 맞추어 작성할 수 있습니다.
- 주제 선정, 독자 대상 설정, 목차 구성, 시각화 자료 수집의 글쓰기 프로세스에 따라 글을 작성할 수 있습니다.

Education

금오공과대학교 산업경영공학 전공

2018.03 - 2024.02

- SoDA(Service Oriented Data Analysis) Lab (2022.03 ~ 2023.03) | 🔗 연구실 홈페이지

Certificates

2024.06	**빅데이터 분석기사**
2022.03	**ADsP(데이터분석 준전문가)**
2020.12	**컴퓨터활용능력 1급**

Languages

English	TOEIC 870점 (2024.06.30 ~ 2026.06.30)

Others

2021.05 - 2021.07	대구 공익데이터실험실 교통안전 분야 참가

Skills

[Language]

- Python

[Library / Framework]

- PyTorch
- HuggingFace
- OpenCV
- Scikit-Learn
- Numpy
- Pandas
- Matplotlib / Seaborn

[DB]

- MySQL

[Tool]

- Git, GitHub, Slack
- Visual Studio Code