어느 분야든 빠르게 시작할 수 있는 준비된 Meta-learner 최영제입니다.

최영제 1995년 (28세/만 26세) | 남 | 구직 준비중

c73801477@gmail.com

010-7380-1477

010-7380-1477

(03726) 서울 서대문구 연희로8길

학력사항	경력사항	희망연봉	희망근무지/근무형태	포트폴리오
대학원(석사) 졸업예정	신입	회사내규에 따름	서울전체 정규직, 인턴직	-

학력 최종학력 | 대학원 석사졸업예정

재학기간	구분	학교명(소재지)	전공	학점
2020.03 ~ 2022.02	졸업예정	연세대학교(서울) 대학원(석사) (서울)	산업공학	3.91 / 4.3
		논문&졸업작품 SAFE: Unsupervised Image Feature Extraction Using a Self-Attention-Based Feature Extraction Network		
2014.03 ~ 2020.02	졸업	국민대학교 (서울)	빅데이터 경영통계전공	3.93 / 4.5
		논문&졸업작품 트랜잭션 기반 추천 시스템에서 워드 임베딩을 통한 도메인 지식 반영. 지식경영연구, 21(1), 117-136.		
2011.03 ~ 2014.02	졸업	광양백운고등학교	문과계열	-

경력 신입

자격증/어학/수상내역

취득일/수상일	구분	자격/어학/수상명	발행처/기관/언어	합격/점수
2018.07	자격증/면허증	데이터분석준전문가(ADsP)	한국데이터베이스진흥원	최종합격
2018.08	수상내역/공모전	제 7회 UNIST 빅데이터 분석 경진대회 (우수상)	UNIST	-
2018.12	수상내역/공모전	2018 빅데이터 해커톤 (우수상)	우정사업본부	-
2021.11	수상내역/공모전	SKT AI Fellowship 3기 (대상)	SKT	-

대외활동

기간	구분	기관/장소	내용
2018.01 ~ 2019.01	동아리	빅데이터 연합 동아리 ToBIg's	- 제 6회, 7회 투빅스 빅데이터 컨퍼런스 참여 / Yolo&CNN 을 이용한 한국어 수화번역 프로젝트 / 강화학습을 이용한 재난구조 에이전트 프로젝트 / - 기타 데이터 분석 및 머신러닝 딥러닝 교육 진행 및 이수
2017.03 ~ 2019.12	교내활동	빅데이터 분석 학회 D&A	- 부학회장 역임 / R 기초 교육, 머신러닝 교육 진행 - '인터넷 로그 데이터 기반 이용자 성별 연령 예측 프로젝트' 머신러닝 Competition 참여 - 제 1회 D&A 빅데이터 컨퍼런스 참여 / Automatic music transcription with machine learning - 제 2회 D&A 빅데이터 컨퍼런스 참여 / ML & DL 기반 KOSPI 200 선물 예측 (https://ds.kookmin.ac.kr/community/?board_name=Community&order_by=fn_pid&order_type=desc&board_page=2&vid=1)
2019.07 ~ 2019.08	인턴	클리오(CLIO)	- 화장품 제조업 클리오 마케팅 팀 하계 인턴 진행 - 데이터 구축: Python Selenium 이용, 화장품 리뷰 사이트 및 Instagram 피드 동적 크롤링 - 텍스트 마이닝: Word2Vec 이용, 경쟁사-팔로워 & 페리페라-팔로워 간 해시태그 상호 관계 추론
2021.06 ~ 2021.11	수행과제	SKT AI Fellowship 3기	- Unsupervised time-series anomaly detection 알고리즘 개발 - 사용 모델: DAGMM, USAD (Autoencoder based) / GANomaly (GANs based) - Unsupervised anomaly detection 성능 고도화를 위한 학습 framework 개발

포트폴리오/기타문서

파일 구분	파일명
기타	http://yjchoi-95.gitbook.io/paper-review/ http://www.yes24.com/Product/Goods/102416641
경력기술서	PDF 참여 프로젝트 상세 기술서.pdf | 1.2MB

경력기술서

프로젝트 및 경진대회 - 학부
1) (D&A 학회, 2017) 인터넷 이용 기록 분석을 통한 이용자 성별 연령 예측 프로젝트
2) (학부 수업, 2018) 서울시 부동산 가격 예측 프로젝트
3) (ToBig's 연합 동아리, 2018) Yolo와 CNN을 활용한 한국어 수화 번역 프로젝트
4) (경진대회 우수상, 2018) 제 7회 UNIST 빅데이터 분석 경진대회
5) (학부 수업, 2018) 노래 가사와 앨범커버를 이용한 힙합/비힙합 분류
6) (Kaggle, 2018) Human protein atlas image classification
7) (D&A 학회, 2018) 제 1회 D&A 빅데이터 컨퍼런스 - Automatic music transcription with machine learning
8) (경진대회 최우수상, 2018) 2018 우정사업본부 빅데이터 해커톤
9) (ToBig's 연합 동아리, 2018) 강화학습을 이용한 재난구조 에이전트 프로젝트
10) (학부 수업, 2019) CLIO - 마케팅 애널리틱스 산학 협동, 하계 인턴 연계
11) (학부 수업, 2019) 트랜잭션 기반 추천 시스템에서 워드 임베딩을 통한 도메인 지식 반영
12) (D&A 학회, 2019) 제 2회 D&A 빅데이터 컨퍼런스 - ML & DL 기반 KOSPI 200 선물 예측

프로젝트 및 산학과제 - 대학원

13) (산학과제, 2019) SK hynix - 객체 탐지 활용 반도체 불량 탐지 모델 개발
14) (산학과제, 2020) 삼성전자 - 제품별 수요 예측 모델 개발
15) (산학과제, 2020) 국방과학연구소(ADD) - 강화학습을 통한 AI 조종사 개발 및 기종 변경을 위한 전이 학습기법 연구
16) (산학과제, 2020) LG innotek - 이미지 & 정형 데이터 활용 수율 예측 모델 개발
17) (프로젝트, 2021) SKT AI Fellowship 3기 - Unsupervised anomaly detection 알고리즘 개발

1) 프로젝트명 : 인터넷 이용 기록 분석을 통한 이용자 성별 연령 예측 프로젝트
- 사용 툴 : R
- 프로젝트 목적 : Classification
- 프로젝트 기간 : 2017.10 ~ 2017.12
- 평가 척도 : Logloss
- 업무 성과 :
[프로젝트 개요]
해당 프로젝트는 학과 내 빅데이터 분석 학회 'D&A'에서 진행하였습니다. 전공 교수님 논문 '인구통계 특성 기반 디지털 마케팅을 위한 클릭스트림 빅데이터 마이닝(2016)' 을 R을 사용해 모델 구현 및 분석을 진행하였습니다. 인터넷 로그 기록을 바탕으로 이용자의 성별과 연령을 예측하였습니다.

[예측 모델링 및 파생변수 생성]
예측 모형으로써 SVM과 같은 분류 모델, Random Forest, XGBoost 등 앙상블 모델을 사용하였습니다. 특정 성별, 연령의 선호가 높은 사이트와 검색어 데이터는 word2vec 모델을 통해 단어를 임베딩한 후 코사인 유사도를 변수로 추가하여 성능을 높였습니다. 생성했던 변수 중 임베딩을 통해 만든 유사도 및 위치 벡터값이 가장 성능이 좋았습니다. 최종적으로 단일 모델의 결과값에 기하평균을 적용한 앙상블을 적용하여 모델의 성능을 높였습니다.

[프로젝트 의의]
머신러닝을 이용하여 데이터 분석을 해본 첫 경험이었으며 데이터 전처리, EDA, 모델링 등 데이터 마이닝의 필수적인 요소를 깊게 공부하였습니다. 정제되지 않은 인터넷 이용기록을 이용하였기에 집계를 내고 변수로 추가하는 과정을 통해 데이터 전처리 과정을 체화시켰습니다. 또한, 텍스트 마이닝의 일부분인 임베딩을 모델링에 적용하는 과정에서 비정형 데이터 분석의 중요성을 느끼게 되어 빅데이터 연합동아리 ToBig's에 지원하는 계기가 되었습니다.

키워드 : R, Classfication, SVM, Random Forest, XGBoost, Ensemble Prediction, Word2Vec

2) 프로젝트명 : 서울시 부동산 가격 예측 프로젝트
- 사용 툴 : R
- 프로젝트 목적 : Regression
- 프로젝트 기간 : 2018.04 ~ 2018.06
- 평가 척도 : RMSE
- 업무 성과 :
[프로젝트 개요]
'데이터 마이닝' 수업에서 진행하였으며 서울시 특정 기간동안 아파트 실거래 데이터를 기반으로 부동산 가격의 추이를 예측하였습니다.

[예측 모델링 / EDA 기반 파생변수 생성]
예측 모형으로써 회귀 모형과 의사결정 나무를 결합하여 잎 노드를 회귀 모델로 대체한 RWeka패키지의 '모델트리', 분류와 회귀문제에서 높은 성능을 보이는 Random Forest, XGBoost를 사용하였습니다. 이후 성능 향상을 위해 Random Forest의 예측결과를 변수에 넣어 XGBoost로 최종 예측을 하는 Stacking을 사용하였습니다.
EDA를 통해 아파트 거래가와 강남, 광화문과의 거리가 높은 상관이 있음을 발견하였습니다. 웹 크롤링을 통해 가장 가까운 지하철 역까지의 거리, 강남과 광화문까지의 거리 변수를 추가하였고 타팀과의 가장 큰 차이점이였습니다.

[프로젝트 의의]
학회에서 진행한 프로젝트와 유사한 방향으로 진행되었으나 크롤링을 통해 웹상의 정보를 끌어와 변수를 추가하는 방법을 익혔습니다. 또한 분석 데이터가 정제가 되어있지 않았기에 데이터 오류로 인해 생긴 이상값 및 결측값를 처리하는 여러 기법(Mice, Miss Forest)을 익혔습니다.

키워드 : R, Regression, RWeka::모델트리, Random Forest,, XGBoost, Mice, Miss Forest, Web Crawling, Stacking

3) 프로젝트명 : Yolo와 CNN을 활용한 한국어 수화 번역 프로젝트
- 사용 툴 : Python, Keras, Darknet
- 프로젝트 목적 : Object detection
- 프로젝트 기간 : 2018.05 ~ 2018.07
- 업무 성과 :
[프로젝트 개요]
전국 빅데이터 연합 동아리 '투빅스'에서 진행한 프로젝트입니다.

[Object detection]
실시간 영상 처리가 가능한 yolo 모델과, 이미지 분류에 사용되는 CNN 모델을 결합하여 지문자(자음, 모음)영상을 input으로 넣으면 해당 글자가 화면에 출력되는 통합 모델을 구축하였습니다.
해당 프로젝트에서 CNN 모델링 및 UI구축부분을 맡았으며 프로젝트 종료 후 Yolo 부분을 공부해 프로젝트 모든 과정을 채워 넣었습니다. 제6회 ToBig's

컨퍼런스 참가 주제이며 2018 양재혁신허브 AI 스쿨 인공지능 R&D 실무자 양성과정에 초청되어 프로젝트 소개를 하였습니다.

[프로젝트 의의]
좋은 분석가는 어떠한 데이터가 주어져도 분석을 해내야 한다고 생각합니다. 해당 프로젝트를 통해 그간 정형 데이터 분석에만 머물렀던 한계를 깨고 이미지 데이터를 처리 할 수 있는 역량을 갖출 수 있었고 딥러닝을 처음 활용한 프로젝트였습니다.

제 6회 투빅스 데이터 분석 컨퍼런스 - DeepKSL (YOLO를 이용한 한국 수화 번역)
- http://www.datamarket.kr/xe/board_pdzw77/44513
인공지능 프로젝트 8. 한국어 수화 번역
- https://www.youtube.com/watch?v=ymCZTgFfv64&feature=youtu.be

키워드 : Python, Keras, Object Detection - Yolo, CNN,

4) 프로젝트명 : 제 7회 UNIST 빅데이터 분석 경진대회 - 우수상
- 사용 툴 : R
- 프로젝트 목적 : Regression, 물동량 예측을 통한 울산항만공사 선제적 대응방안 수립
- 프로젝트 기간 : 2018.08.01 ~ 2018.08.03
- 업무 성과 :
[프로젝트 개요]
제7회 UNIST 빅데이터 분석 경진대회의 목적은 '2017년 울산항만 물동량 예측'이었습니다.

[데이터 소개 및 문제점]
주어진 데이터는 2014, 2015, 2016, 2017년 울산항만 물동량 현황이었으며 대회의 목적에 맞게 당일이 2017년 1월 1일이라 가정하고 분석을 진행하였습니다. 시계열 모델인 ARIMA와 앙상블 모델 중 하나인 XGBoost를 사용하였는데 후자인 부스팅 기법을 사용할 때 문제가 발생하였습니다. 가용할 수 있는 데이터는 14~16년 데이터였기에 17년도 물동량을 예측할 때 17년도의 독립변수들을 사용할 수 없는 상황이었습니다.

[문제 해결 방안 및 결과]
문제를 해결하기 위해서 두 가지 방법으로 접근하였습니다. 17년도의 물동량을 1년씩 앞당겨서(dplyr::lead) 학습하는 방법과 17년도의 독립변수들을 예측해서 채워 넣은 후 예측된 x들을 이용하여 물동량을 예측하는 방법을 제시하였습니다. 두 가지 방법 중 1년씩 앞당겨서 학습한 방법의 RMSE가 더 낮아 전자의 방법을 사용하였습니다. 또한, 17년의 물동량을 예측할 때 연도로 접근하는 것이 아니라 분기별로 예측한 후 합산하는 과정으로 변형하여 진행하였고 학습데이터의 양을 증가시키는 효과가 있어 더 정확한 결과를 보여주어 높은 성능을 보였습니다.

[프로젝트 의의]
다른 분석팀들은 고려하지 못했던 17년도의 독립변수들을 처리하는 방안과 분기별로 예측 후 합산하여 1년 전체의 물동량을 계산하는 방법의 참신성은 관계자들에게 참신하게 다가와 대회 우수상을 받을 수 있었습니다. 앞서 진행한 머신러닝 프로젝트들을 통해 습득된 빠르고 정확한 전처리 실력과 모델링 구축 능력이 짧은 시간 내에 결과물을 만들어야 하는 대회에서 큰 도움을 주었습니다.

키워드 : R, Regression, XGBoost, ARIMA, 독립변수 채워넣기

5) 프로젝트명 : 노래 가사와 앨범커버를 이용한 힙합/비힙합 분류
- 사용 툴 : Python, Keras
- 프로젝트 목적 : Classification, 텍스트 데이터(가사)와 이미지 데이터(앨범 커버)를 이용하여 힙합/비힙합 장르 분류
- 프로젝트 기간 : 2018.10 ~ 2018.12
- 업무 성과 :
[프로젝트 개요]
'비정형데이터 분석' 수업 중 진행한 프로젝트로 가사 및 앨범 커버사진을 통한 힙합/비힙합 분류 프로젝트를 진행하였습니다.

[데이터 수집 및 전처리]
Bucks 뮤직에서 크롤링을 통해 직접 데이터를 구축하였으며 노래 가사는 TF-IDF를 이용하여 TDM 구축, LDA를 통한 주제분석, LSTM을 활용한 분류 모델링이 주요 내용이고 앨범커버는 CNN 모델링을 적용하였습니다.

[예측 모델링]
당시 프로젝트는 가사와 앨범커버를 별개로 진행하였으나 CNN을 통과한 추출된 특징 벡터를 LSTM의 마지막 분류 layer 를 통과하기 직전에 LSTM을 거친 벡터와 concat하여 딥러닝 모델을 자유롭게 활용해 보았습니다.

[프로젝트 의의]
해당 프로젝트를 통해서 크롤링을 통해 웹상의 데이터 구축하는 방법부터 주제 분석 및 분류 등 텍스트 데이터를 다루는 법을 구체적으로 배웠습니다. 비정형 데이터의 중요한 축인 이미지와 텍스트를 모두 다룰 수 있게 되는 계기였습니다.

Github 링크
- https://github.com/YeoungJe/hiphop_nonhiphop-classification

키워드 : Python, Keras, Classification, Web Crawling, CNN, LSTM, text mining

6) 프로젝트명 : Kaggle - Human protein atlas image classification
- 사용 툴 : Python, Keras, Pytorch
- 프로젝트 목적 : Multi label classification, 단백질의 혼합된 패턴을 분류할 수 있는 모델을 개발하는 것
- 프로젝트 기간 : 2018.10. ~ 2018.12.
- 업무 성과 :
[프로젝트 개요]
해당 프로젝트는 3-2학기 '인공지능' 수업에서 진행한 기말 프로젝트입니다. Kaggle에서 개최한 인간의 단백질 세포 패턴을 분류하는 대회로써 약 2달간 진행되었으며 대용량 이미지 처리가 필요한 대회였기에 주로 Colab과 kaggle 커널을 사용하여 모델링을 실시하였습니다.

[데이터 소개 및 고려한 전처리 방안]
데이터는 1개의 이미지가 4채널 (R,G,B,Y)로 구성되어 128,244장(32,061개의 세포)이 Train data로 주어졌습니다. 상위 커널을 분석한 결과 직접 모델 구조를 구성하기 보다는 transfer learning 통해 진행한 것을 확인하였습니다. 이를 참고하여 전이학습을 시도할 때 고려해야할 점이 생겼습니다. 주어진 데이터의 형태는 512*512*4의 형태였고 Resnet의 인풋 layer의 채널은 3이였기에 그대로 적용할 수가 없었습니다. 그래서 저희는 크게 데이터를 수정하는 것과, 모델을 수정하는 것 두가지 방법으로 진행하였습니다.
첫번째, 데이터를 수정하는 방안으로는, R,G,B,Y 중 1개의 채널을 사용하지 않고 학습시키는 것과 데이터의 R,G,B,Y를 각각 섞어 3개의 채널로 만드는 방법(ex, a = R * 0.5 + G * 0.5, b = B * 0.5 + Y * 0.5 ...)이 있었습니다. 저희는 프로젝트를 진행하며 주어진 데이터를 모두 활용하지 않으면 높은 성적을 얻을 수 없다는 것을 깨달아 후자의 방법을 사용한 모델을 구축하였습니다.
두번째, 모델을 수정하는 방안은 Resnet의 output layer를 제외한 layer들은 freezing 시킨 후 Resnet input layer앞단에 14층의 Convolution layer를 덧붙여 224*224*3의 output을 출력하도록 설계 후 이를 Resnet과 연결시켜서 학습을 진행하였습니다. train 이미지에서 feature를 추출하여 224*224*3으로 보내는 것이 더 좋은 방법이라고 기대했으나 실제로 진행해보니 앞서 말한 4채널의 데이터를 섞어서 3채널로 만든 후 학습한 것이 더 좋은 성능을 보였습니다.

[Multi label classification]
이 프로젝트는 예측 결과물 형태가 한 이미지가 여러 class를 가질 수 있는 Multi label classification 문제였습니다. 따라서 output layer의 활성함수를 softmax가 아닌 sigmoid로 설정하여 학습 하였고 나온 예측값과 train label을 토대로 thresholding을 진행하여 최종 제출파일을 만들었습니다.

[프로젝트 의의]
해당 프로젝트를 통해서 노트북 등의 로컬자원 이외의 Kaggle, Colab 사용법을 익혔고 일반 분류문제가 아닌 Multi label classification 문제를 처음 접하게 되었는데 이를 다루는 방법을 알게되었습니다. 머신의 한계가 있어 많은 실험을 해보지 못한게 아쉬웠지만 상위 커널들을 보면서 이미지 처리에 관한 다양한 생각들을 접할 수 있어 뜻깊은 프로젝트였습니다.

키워드 : Python, Pytorch, Kaggle, Colab, Multil label classification, transfer learning, ResNet

7) 프로젝트명 : 제 1회 D&A 빅데이터 컨퍼런스 - 음악을 악보로 변환하는 프로젝트
- 사용 툴 : Python, Keras,
- 프로젝트 목적 : Automatic music transcription with machine learning
- 프로젝트 기간 : 2018.09~2018.11
- 업무 성과 :
[프로젝트 개요]
음악을 쓰다는 국민대 빅데이터 분석학회 'D&A'에서 진행한 프로젝트입니다. 당시 부학회장을 맡고 있었던 저는 '알고리즘에 대한 깊은 이해도를 바탕으로, 일상에서 흥미로운 주제를 발표하여 학회원들의 동기부여를 하자!'라는 생각에서 주제를 선정하였습니다.

[데이터 전처리 및 구축]
전체적인 개요는 librosa 라이브러리를 활용하여 음을 Spectrogram으로 변환하여 음이 변하는 지점을 추출 및 저장합니다. 이미지로 변환된 데이터는 KNN-PCA-DBSCAN의 과정을 거쳐 Label 축소, Outlier 제거의 과정을 거칩니다. 이후 정제된 데이터를 바탕으로 CNN을 적용하였으며 Multi-label Classification을 푸는 문제였기에 마지막 layer의 활성함수를 softmax 대신 sigmoid로 변경하여 진행하였습니다.

[변환 모델링]
이 프로젝트를 통해 강조하고싶은 점은 알고리즘(KNN, PCA, DBSCAN)들의 상황에 맞는 적용과정입니다. 피아노 음은 무수히 많은 조합의 수가 존재하기 때문에 학습에 장애를 줄 수 있다 판단하여 KNN을 적용하였습니다. KNN을 통해 많이 눌리지 않은 특수한 음 조합은 가장 유사하며 대표성을 띄는 음에 할당하였습니다. 이상치 탐지하기 위해 사람이 하는 과정(눈으로 확인 후 이상치 제거)을 PCA를 이용하여 2차원 축소 후 DBSCAN을 적용하여 최대 군집만을 남겨 자동으로 제거하는 프로세스를 구축하였습니다. 이는 각 알고리즘에 대한 깊이 있는 이해가 있었기에 가능한 작업이라고 생각합니다.

[프로젝트 의의]
음악을 쓰다 프로젝트를 통해 비정형 데이터 중 음성 데이터를 처리하는 방법을 터득했고 이상치를 제거하는 방법론에 대해 심도있는 고민을 진행하였습니다. 코드와 간략한 진행 과정은 아래 Github 링크를 첨부하였습니다!

제 1회 D&A Conference - 음악을 쓰다
- https://github.com/YeoungJe/DnA_Conference

키워드 : Python, Keras, Spectrogram, KNN, PCA, DBSCAN, Outlier 제거, CNN

8) 프로젝트명 : 2018 우정사업본부 빅데이터 해커톤 - 최우수상
- 사용 툴 : R, Python

- 프로젝트 목적 : Regression, 2017년 하반기, 2018년 하반기 연휴기간(연휴일 전,후 3일) 택배 물량 예측
- 프로젝트 기간 : 2018.12.20 ~ 2018.12.21
- 업무 성과 :
[데이터 수집]
우정사업본부에서 제공한 데이터는 17,18년도 상반기 구별로 집계된 정보였습니다. 저희는 추가로 구글맵 크롤링, 구별 인구 census 정보, 주변시설(아파트, 병원, 산업단지 유무 등) 등 공간정보를 포함한 내용과 네이버 검색어 트렌드, Twiiter API를 활용한 SNS 정보, 기상특보 등 시간 데이터를 수집하여 분석을 진행하였습니다.

[EDA, 탐색적 데이터 분석]
해커톤에서 수행해야하는 과업은 공휴일 인근의 택배량을 예측하는 것이었습니다. 주어진 데이터를 EDA한 결과 실제로 공휴일 당일을 제외한 주변 3일간 택배량은 평소와 다름을 확인하였습니다. 또한 30대~40대 여성 거주 인구가 높을 수록 택배량이 많아지는 양의 상관관계를 도출하였습니다.
내부 데이터 외, 네이버 검색어 키워드와 함께 EDA를 한 결과 택배 배송량과 "택배", "택배 지연", "이사", "군입대" 등 특정 키워드와 높은 상관관계를 갖는 것을 발견하였습니다. 이에 저희는 택배량을 간접적으로 유추할 수 있는 대리지표로 설정하고 이를 독립변수로 삼았습니다.

[텍스트 마이닝]
SNS에서 확인한 트위터의 내용 중 우정사업본부의 이미지에 영향을 미칠 수 있는 내용을 파악하기 위해 긍정 트윗과 부정 트윗을 LSTM을 사용하여 진행하였습니다. 해당 모델의 정확도는 94%로 높은 수준을 보여주었습니다. 추가로 모델의 예측결과의 이유를 설명해주는 LIME 알고리즘을 통해 긍, 부정으로 분류된 핵심 키워드를 도출하였고 단순히 결과를 제시한 것 보다 높은 설득력을 보여주었습니다.

[택배 배송량 예측 모델링]
앞서 언급한 데이터 및 독립 변수들을 이용하여 예측을 진행하였습니다. 가장 흥미로웠던 변수는 "우체국 택배 ', '우체국 택배 배송조회' 검색어 수였으며 시각화하여 확인한 결과 택배 배송량과 놀라울 정도로 일치하는 모습을 보여주었습니다. SNS 변수를 추가했을 때 RMSE가 2407.33에서 1940.92으로 좋은 성능을 보여주었습니다. 예측 모델로는 XGBoost 모델 예측값의 분산을 낮추기 위해 Bagging 개념을 추가로 도입하여 진행했습니다. 이는 ToBig's 동아리에서 배웠던 내용으로 Bagging의 base 모델인 의사결정나무 대신에 XGBoost를 base모델로 삼도록 함수를 구성하여 사용하였습니다. 평균적으로 동일한 변수로 진행을 할 때보다 약 80~100 정도의 RMSE를 줄이는 효과를 가져왔습니다.

[프로젝트 의의]
대회를 통해 비정형 데이터의 영향력을 다시 한 번 느꼈고 XGBoost에 Bagging 개념을 도입하여 '앙상블 위의 앙상블'이 실제로 단일 XGBoost보다 효과가 좋음을 확인할 수 있었습니다.

키워드 : R, Python, Regression, Web Crawling, textmining, XGBoost based Bagging

9) 프로젝트명 : 강화학습을 이용한 재난구조 에이전트 프로젝트
- 사용 툴 : Python
- 프로젝트 목적 : Reinforcement learning, 선박 침몰로 인한 재난 구조 에이전트 구축
- 프로젝트 기간 : 2018.09.01 ~ 2019.01.14
- 업무 성과 :
[프로젝트 개요]
빅데이터 연합 동아리 ToBig's에서 진행한 제 7회 컨퍼런스로 재난이 발생했을 때 인명 구조를 진행하는 agent를 구축하는 것이 목표였습니다. Agent의 목표는 두 가지로 기름 제거와 인명 구조입니다. 목표로 한 agent의 행동은 인명을 구조를 최우선으로 하되 이동 경로에서 빈 바다를 지나는 대신 기름 제거를 최대로 할 수 있는 경로를 찾아가는 것이었습니다.

[환경 구축]
강화학습의 경우 환경 구축이 매우 중요합니다. 환경은 10*10 grid world를 변형하여 구축하였으며 state, action, reward는 다음과 같습니다.
State: 사람, 기름과의 상대 좌표 x,y
Action: 상, 하, 좌, 우
Reward: 인명 구조시 + 10, 기름 제거시 +1, 정해진 time-step 내 임무 완수 시 +100, 실패시 -100

[강화학습 알고리즘]
사용한 알고리즘은 크게 3가지로 deep q-newtorks (DQN, value based), REINFORCE (policy based), advantage actor-critic (A2C, value-policy based)를 사용하였습니다. 각각 가치 기반, 정책 기반, 가치-정책 기반 알고리즘을 사용하여 병렬적으로 수행하여 알고리즘별 차이를 확인하고자 하였습니다.

[학습 시 문제점 및 개선 사항]
앞서 소개한 state의 경우 사람, 기름별 x, y 상대 좌표를 펼쳐서 dense layer의 input으로 삼았습니다. 위 경우 3가지 알고리즘 모두 수렴하지 않았는데, 그 이유로 모든 사람, 기름별 좌표 정보를 펼쳐서 받을 경우 개별 정보가 손실될 수 있다는 의견이 존재하였습니다. 따라서 사람, 기름별로 독립된 정보를 달리 반영하고자 conv 1d를 사용하여 모델에 넘겨주는 방식으로 변경하였고, 이에 DQN, A2C는 수렴에 성공하였습니다.

[프로젝트 의의]
기존에 정형, 비정형(이미지, 텍스트, 음성) 데이터를 통한 supervised, unsupervised learning은 많이 다루어 보았으나 강화학습은 처음이었습니다. 아이디어 설계, 환경 구축, 강화학습 모델링 학습 과정을 통해서 깊이있게 공부할 수 있었고 network 구조 변화가 학습에 큰 영향을 끼친다는 사실 또한 알게되었습니다.

- 제 7회 투빅스 컨퍼런스: http://www.datamarket.kr/xe/index.php?mid=board_pdzw77&page=2&document_srl=50423

키워드 : Reinforcement learning, DQN, REINFORCE, A2C, Conv 1d

10) (학부 수업, 2019) CLIO - 마케팅 애널리틱스 산학 협동, 하계 인턴 연계
- 사용 툴 : Python, Selenium
- 프로젝트 목적 : AI, 빅데이터 분석을 기반으로 한 개인 맞춤형 제품 추천 서비스, 인스타그램 크롤링을 통한 잠재 고객 탐지
- 프로젝트 기간 : 2019.03 ~ 2019.08
- 업무 성과 :
[프로젝트 개요]
화장품 업체 CLIO와 연계하여 빅데이터 프로젝트: 캡스톤 디자인에서 진행한 프로젝트 입니다. 클리오가 저희에게 의뢰한 내용은 '클리오의 타겟층은 누구인가?', "클리오 타겟층의 관심사는 무엇인가?", "그들에게 효율적으로 접근하는 방안은 무엇인가?" 입니다.

[데이터 구축]
주어진 과제를 해결하기 위해 Youtube, 인스타그램, 페이스북 등 SNS 분야와 대표적인 여초 카페인 '여성시대', 화장품 리뷰 사이트인 '파우더룸', '글로우픽'을 크롤링한 후 텍스트 마이닝을 진행하였습니다. 분석 결과 주 타겟층의 연령은 18-24세이며 주된 관심사는 특정 단어로 정의할 수 없으나 크게는 "취업", "여가" 등으로 분류할 수 있었습니다. 분석과정 중 소비자들의 여론을 파악하기 위해 설문조사를 진행하였으며(표본 약 30명) 그 결과 '구매 시 탐색 비용을 절감해주는 추천 시스템'에 관한 Needs를 포착하여 주제를 추천 서비스로 잡았습니다.
이후 클리오 인턴을 진행하며 클리오, 페리페라 등 자사 브랜드를 팔로우 하고 있는 인스타그래머들의 피드와 경쟁 브랜드를 팔로우 하고 있는 피드들의 차이는 무엇인지 파악하기 위하여 Selenium을 활용한 동적 크롤링을 진행하였습니다.

[상황을 고려한 추천시스템 모델링]
추천 시스템은 '기본적인 추천 시스템'과 '상황을 고려한 추천 시스템' 두가지로 나뉩니다. 기본적인 추천 시스템으로는 클리오 고객들의 구매내역을 기반으로 한 Association rule과 협업필터링(CF)을 진행하였습니다.
추천시스템 구축 시 핵심 아이디어로는 상황을 고려한 추천 시스템인데 이는 국내 어느 화장품 업체에서도 적용하고 있지 않은 내용으로 독창적인 아이디어입니다. 예를 들어 '결혼식 하객 메이크업'이나 '물놀이 메이크업' 등의 검색어를 입력하더라도 그에 맞는 클리오 제품이 추천되는 방식입니다. 현재 이러한 검색기능은 어떠한 화장품 브랜드에서도 결과가 나오지 않습니다.
해당 방법은 간단하게 Youtube 크롤링을 통해서 구축할 수 있었습니다. 일반적으로 메이크업 유튜버의 경우 '물놀이 메이크업' 등의 영상을 올리고 해당 영상에서 사용된 화장품을 기록해 놓습니다. 이를 크롤링하여 DB를 구축 후 이용자가 '물놀이 메이크업', '하계 메이크업' 등의 검색을 하면 위 제품을 추천해주는 방식인데, 여기서 한 가지 문제는 유튜버들이 반드시 클리오 제품만을 사용하지는 않는다는 것입니다.
이를 해결하기 위해 존재하는 모든 화장품들의 성분을 크롤링하고, 영상에서 활용된 타 브랜드 제품과 가장 성분이 비슷한 클리오 제품을 추천하는 방식으로 진행하였습니다. 이를 통해 '하객 메이크업', '물놀이 메이크업', '새내기 메이크업' 등의 검색시에도 제품을 추천할 수 있게 됩니다.

[텍스트 마이닝]
클리오 인턴을 진행하며 구축한 인스타그램 피드 DB에서 게시글, 이미지 태그, 해시태그등을 추출하였고 Word2Vec을 사용하여 경쟁사-팔로워, 자사-팔로워들 간의 관심사 차이를 파악하였습니다. 목적은 자사 브랜드는 팔로우하지 않으나 타사 브랜드를 팔로우 하고 있는 사람을 잠재고객으로 특정하여 그들의 특징은 무엇인지 파악하였고, 이후 클리오, 페리페라에서 인스타 업로드를 할 때 어떤 유형의 게시글, 해시태그를 사용해야하는지 적용하기 위함이었습니다.

[프로젝트 의의]
지금까지 모델링 위주, 성능 개선 위주의 프로젝트만 진행하였는데, 이와는 달리 철저히 현업 관점에서 진행된 프로젝트입니다. 실제로 현업 마케팅에서 적용하기에 활용 가능한 머신러닝 및 딥러닝 모델이 많지 않음을 깨닫게 된 경험이었습니다.

Keyword: Selenium, text mining, recommendation system

11) (학부 수업, 2019) 트랜잭션 기반 추천 시스템에서 워드 임베딩을 통한 도메인 지식 반영
12) (D&A 학회, 2019) 제 2회 D&A 빅데이터 컨퍼런스 - ML & DL 기반 KOSPI 200 선물 예측

13) (산학과제, 2019) SK hynix - 객체 탐지 활용 반도체 불량 탐지 모델 개발
14) (산학과제, 2020) 삼성전자 - 제품별 수요 예측 모델 개발
15) (산학과제, 2020) 국방과학연구소(ADD) - 강화학습을 통한 AI 조종사 개발 및 기종 변경을 위한 전이 학습기법 연구
16) (산학과제, 2020) LG innotek - 이미지 & 정형 데이터 활용 수율 예측 모델 개발
17) (프로젝트, 2021) SKT AI Fellowship 3기 - Unsupervised anomaly detection 알고리즘 개발