

포트폴리오



김형준(Gorden)

- 학교 : 선문대학교 졸업(2020)
- 전공 : 스페인어중남미학 전공
상담·산업 심리학 전공

1

Python 활용 빅데이터 분석가 양성과정

- 약 4300시간의 데이터 분석 및 실전을 통한 역량 발전
- '20 9월 ~ '21 3월 / 고용노동부&한국고용정보원 주관

2

데이터 분석 & ML 프로젝트 참여

- Kaggle, Dacon 공모전 및 사이드 프로젝트 수행
- '20 12 ~ 현재

3

다양한 분석 기법 및 상용툴 경험

- Tableau, R, SQL, PostgreSQL 등 분석에 필요한 툴 학습
- 대시보드 제작 및 통계기법을 활용한 분석에 대한 다양한 서적을 통한 주도적 학습

4

풍부한 대내외 활동

- 다양한 프로그램 및 사회활동을 통한 커뮤니케이션 능력 발달
- 다양한 아르바이트를 통한 성과 달성 경험

Python을 활용한 빅데이터 분석가 양성과정을 통해
데이터 분석가에 필요한 분석 및 활용 능력을 갖추었습니다.

'21 9월 ~ '21 3월 / 약 4300 시간 교육 / 고용노동부&한국고용정보원 주관

데이터 수집 및 적재

- Python 기반 데이터 수집
 - Open API를 활용한 데이터 수집
 - 워크넷 오픈 API, 공공데이터포털, 고속도로 공공 데이터 포털 등
 - 채용공고, 통행량, 등 5여종의 데이터 수집
- 수집 데이터 DB 적재
 - PostgreSQL 활용

데이터 분석 및 시각화

- 다양한 빅데이터 분석 방법 적용
 - 변수의 형태에 따른 적합한 통계 기법 활용
Ex) 상관, 분산, 회귀 등
 - 자연어 처리 기반 감정분석
 - 채용공고 분석을 통한 인사이트 발굴
- Python을 활용한 시각화
 - Pyechart, plotly 라이브러리를 이용한 동적 시각화
 - Python dash를 이용한 대시보드 제작&배포

미니 프로젝트 진행

- 네이버 쇼핑 리뷰 감성 분류
 - 네이버 쇼핑 리뷰 데이터 크롤링
 - 데이터 전처리 & 토큰화 진행
 - 긍정과 부정 리뷰에 대한 길이 분포 확인
 - 단어집합(vocabulary)생성
 - 단어 빈도수에 따라 주요 키워드 확인
 - GRU를 이용한 리뷰 감성 분류
- Kaggle 미니 프로젝트
 - 보스턴 집값 예측
 - 악성 댓글 심각도 분석
 - 트위터 가짜뉴스 판독

데이터 분석가는 문제를 바라보는 관찰력과 다양한 시각이 필요하다고 생각합니다.
다양한 분석&시각화 경험을 데이터 분석 직무에 활용하고자 합니다.

Python을 활용한 빅데이터 분석가 양성과정을 통해 데이터 분석가에 필요한 분석 및 활용 능력을 갖추었습니다.

'21 9월 ~ '21 3월 / 약 4300 시간 교육 / 고용노동부&한국고용정보원 주관

데이터 수집 및 적재

데이터 분석 및 시각화

미니 프로젝트 진행

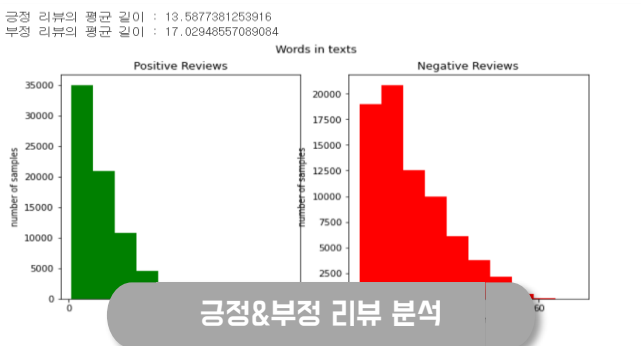


필요한 데이터 선정

API / 크롤링을 사용하여 수집

PostgreSQL 사용 적재

목적에 따라 정확한 데이터를 선정하고, 효율적으로 수집 및 적재 하는 것이
뛰어난 데이터 분석가로 성장하는 길이라는 것을 배웠습니다.



분석하고자 하는 방향에 따라 다양한 분석 방법들을 사용하여 이슈를 도출 하였으며,
한 눈에 직관적으로 표현할 수 있는 시각화 방법을 학습했습니다.

Python을 활용한 빅데이터 분석가 양성과정을 통해 데이터 분석가에 필요한 분석 및 활용 능력을 갖추었습니다.

'21 9월 ~ '21 3월 / 약 4300 시간 교육 / 고용노동부&한국고용정보원 주관

데이터 수집 및 적재

데이터 분석 및 시각화

미니 프로젝트 진행

모형 정의 및 검증평가

• cv_rmse() 함수

```
1 from sklearn.metrics import mean_squared_error
2 from sklearn.model_selection import KFold, cross_val_score # 교차검증 시 평가 메트릭 함수
3
4 def cv_rmse(model, n_folds = 5): # 5번 교차검증 실행한다
5     cv = KFold(n_splits = n_folds, random_state=42, shuffle=True)
6     rmse_list = np.sqrt(-cross_val_score(model, X, y, scoring = "neg_mean_squared_error", cv=cv))
7     print("CV RMSE value list:", np.round(rmse_list, 4))
8     print("CV RMSE mean value:", np.round(np.mean(rmse_list), 4))
9     return (rmse_list)
```

```
1 from sklearn.linear_model import LinearRegression
2
3 n_folds = 10
4 rmse_scores = {}
5 lr_model = LinearRegression()
6
7 score = cv_rmse(lr_model, n_folds)
8
9 CV RMSE value list: [0.1489 0.1016 0.1089 0.1165 0.1559 0.1392 0.1165 0.1154 0.121 0.1128]
10 CV RMSE mean value: 0.1342
```

```
1 from sklearn.model_selection import cross_val_predict
2
3 X = all_df.iloc[:, 1:]
4 X_test = all_df.iloc[:, 1:]
5 X.shape, y.shape, X_test.shape
6
7 lr_model.fit = lr_model.fit(X, y)
8 # y 로그변환, y값= 로그값, 그러므로 y값을 반로그값으로 변환해준다
9 final_preds = np.floor(np.exp(lr_model.predict(X_test)))
10 print(final_preds)
```

[11704, 15154, 1874]

주택가격 예측 모형 검증 평가

```
24 encoded = tokenizer.texts_to_sequences([new_sentence]) # 정수 인코딩
25 pad_new = pad_sequences(encoded, maxlen = max_len) # 패딩
26
27 score = float(model.predict(pad_new)) # 예측
28 if(score > 0.5):
29     print('!:.21% 확률로 긍정 리뷰입니다.', format(score * 100))
30 else:
31     print('!:.21% 확률로 부정 리뷰입니다.', format((1 - score) * 100))
```

```
Epoch 1/15
1875/1875: ===== - ETA: 0s - loss: 0.2734 - acc: 0.8884
Epoch 0001: val_acc improved from -inf to 0.9127, saving model to best_model.h5
1875/1875: ===== - 47s 25ms/step - loss: 0.2734 - acc: 0.8884 - val_loss: 0.2338 - val_acc: 0.9127
Epoch 2/15
1875/1875: ===== - ETA: 0s - loss: 0.2156 - acc: 0.9222
Epoch 0002: val_acc improved from 0.9127 to 0.9293, saving model to best_model.h5
1875/1875: ===== - 47s 25ms/step - loss: 0.2156 - acc: 0.9222 - val_loss: 0.2178 - val_acc: 0.9293
Epoch 3/15
1875/1875: ===== - ETA: 0s - loss: 0.1962 - acc: 0.9296
Epoch 0003: val_acc improved from 0.9293 to 0.9274, saving model to best_model.h5
1875/1875: ===== - 50s 27ms/step - loss: 0.1962 - acc: 0.9296 - val_loss: 0.2034 - val_acc: 0.9273
Epoch 4/15
1875/1875: ===== - ETA: 0s - loss: 0.1877 - acc: 0.9329
Epoch 0004: val_acc did not improve from 0.9274
1875/1875: ===== - 47s 25ms/step - loss: 0.1877 - acc: 0.9329 - val_loss: 0.2034 - val_acc: 0.9266
Epoch 5/15
1875/1875: ===== - ETA: 0s - loss: 0.1787 - acc: 0.9387
Epoch 0005: val_acc did not improve from 0.9274
1875/1875: ===== - 47s 25ms/step - loss: 0.1787 - acc: 0.9386 - val_loss: 0.2010 - val_acc: 0.9273
Epoch 6/15
1874/1875: ===== - ETA: 0s - loss: 0.1690 - acc: 0.9403
Epoch 0006: val_acc did not improve from 0.9274
1875/1875: ===== - 46s 24ms/step - loss: 0.1690 - acc: 0.9403 - val_loss: 0.2075 - val_acc: 0.9240
Epoch 7/15
1874/1875: ===== - ETA: 0s - loss: 0.1611 - acc: 0.9430
Epoch 0007: val_acc did not improve from 0.9274
1875/1875: ===== - 47s 25ms/step - loss: 0.1611 - acc: 0.9430 - val_loss: 0.2179 - val_acc: 0.9204
Epoch 8/15
1873/1875: ===== - ETA: 0s - loss: 0.1532 - acc: 0.9471
Epoch 0008: val_acc did not improve from 0.9274
1875/1875: ===== - 46s 24ms/step - loss: 0.1533 - acc: 0.9471 - val_loss: 0.2150 - val_acc: 0.9213
Epoch 9/15
1873/1875: ===== - ETA: 0s - loss: 0.1437 - acc: 0.9502
Epoch 0009: val_acc did not improve from 0.9274
1875/1875: ===== - 47s 25ms/step - loss: 0.1437 - acc: 0.9503 - val_loss: 0.2384 - val_acc: 0.9176
Epoch 0009: early stopping
```

1 sentiment_predict('이 상품 진짜 싫어요... 교환해주세요')

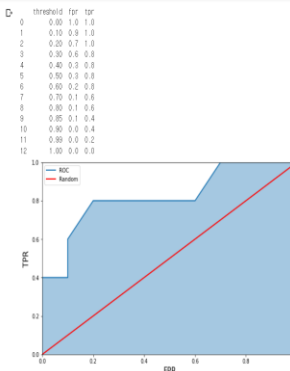
99.99% 확률로 부정 리뷰입니다.

1 sentiment_predict('')

99.99% 확률로 긍정 리뷰입니다.

감정분류 정확도 평가

```
29 ax.set_xlim(0, 1.0)
30 ax.set_ylim(0, 1.0)
31 ax.set_xlabel("FPR", fontsize=15)
32 ax.set_ylabel("TPR", fontsize=15)
33 plt.legend()
34 plt.show()
```



```
1 from sklearn.metrics import roc_auc_score
2
3 y_true = [0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1]
4 y_pred = [0.1, 0.3, 0.2, 0.6, 0.8, 0.05, 0.9, 0.5, 0.3, 0.66, 0.3, 0.2, 0.85, 0.15, 0.99]
5
6 print("AUC:", roc_auc_score(y_true, y_pred))
```

AUC: 0.8300000000000001

• AUC에 관한 결과 해석

= AUC = 1.0: 가장 좋은

= AUC = 0.5: 무작위

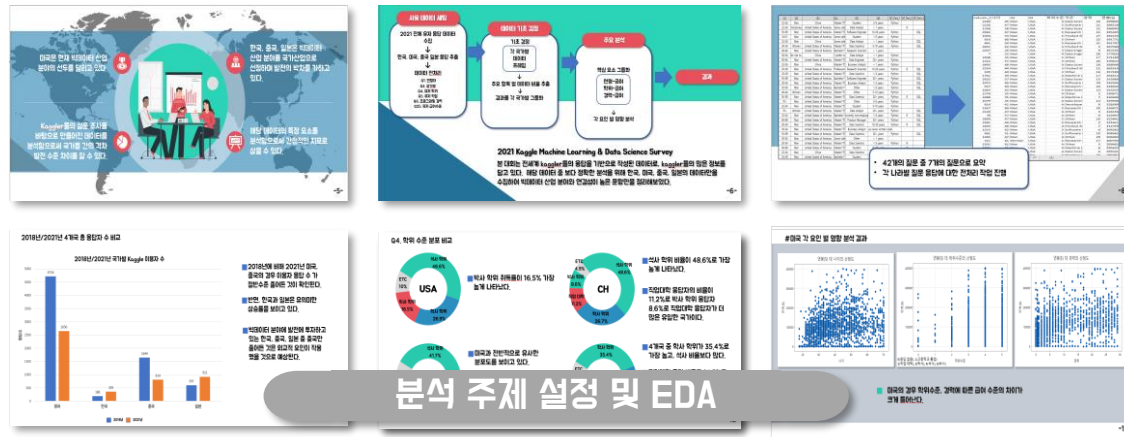
= AUC = 0.0: 최악의

가짜뉴스 AUC 모델 검증

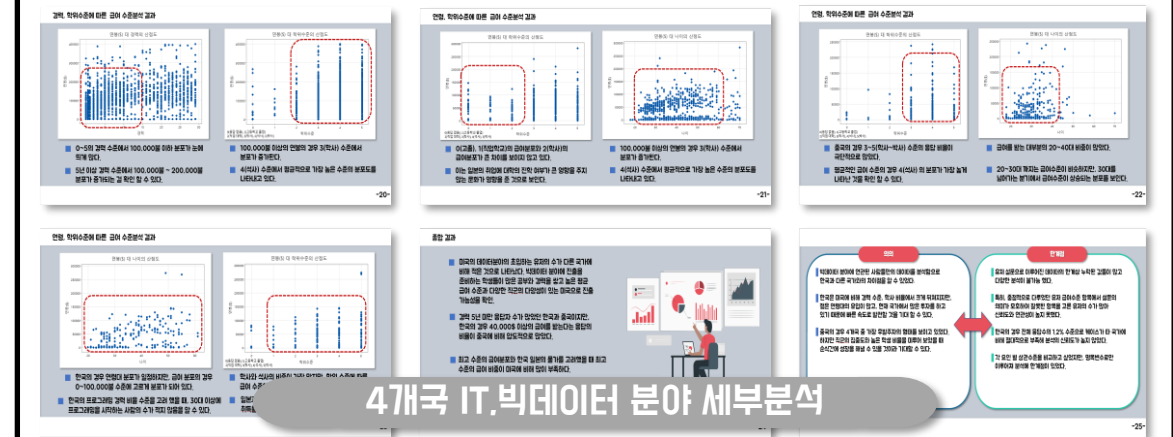
학습한 데이터 수집 및 분석 능력을 바탕으로 다양한 미니 프로젝트를 수행하였으며,
검증과 평가를 통해 보완점 및 방향성을 제시했습니다.

Kaggle 유저들의 설문 응답을 분석 및 시각화 대회를 통해 분석 기획 능력과

Kaggle 유저 설문 분석&시각화 경진 대회



- 유저 설문 응답의 특성을 파악하여 목적에 필요한 응답 추출
 - 학위, 직업, 경력, 급여수준 등 핵심 응답 추출
- 데이터 분석 및 1차 시각화
 - 국적 데이터를 통해 국가별 유저 그룹화
 - 전체 데이터 분석을 통해 국가별 특성 파악
 - 학력, 경력, 학위수준과 급여 수준 간의 관계 분석

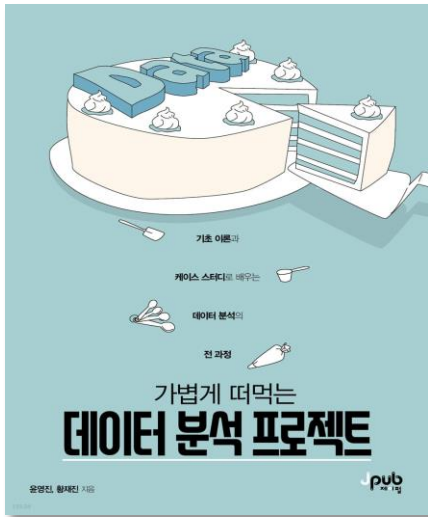


- EDA를 통해 분석된 결과를 국가별 세부 분석
 - 각 국가별 학력, 경력, 학위수준, 연령 별 특성을 대입
 - 급여 수준의 차이를 통해 발전도 비교
- 분석 결과에 대한 의의와 한계점 파악 및 피드백 작성

분석의 방향성을 확실하게 잡지 못했지만, 다양한 데이터를 핸들링 할 수 있었습니다.
이러한 저의 데이터 핸들링 능력을 데이터 분석 직무에 활용하고자 합니다.

데이터 분석가는 발전을 위해 끊임없는 학습하는 자세가 필요하다 생각합니다.
다양한 관련 서적을 통해 전문적인 데이터 분석가가 되기 위한 자질을 갖추었습니다.

데이터 분석 관련 서적을 통한 데이터 분석 및 대시보드 작성 능력 함양



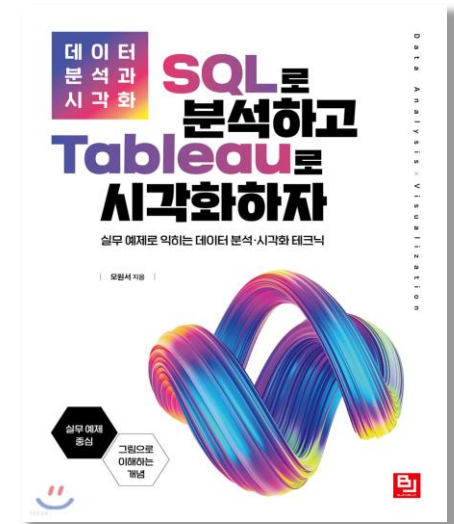
Python 기반 분석



R기반 통계 분석



대시보드 실전 예제



SQL & Tableau

- Python을 메인으로 데이터 핸들링 및 시각화의 기본을 다졌습니다.
- 6단계의 프로젝트 구성, 데이터 발굴, 검증, 전처리 및 Python에서 제공하는 다양한 분석&시각화 툴을 학습했습니다.
- R을 통해 변수에 따른 다양한 함수와 패키지를 학습했습니다.
- Lubridate, geom, DT, plotly, googlesheets4 등

- 대시보드의 구성과 실무 적용 사례를 공부하였습니다.
- 다양한 시나리오에 대한 효과적인 대시보드 작성법을 학습하고 실질적인 실무의 사용 예시들에 대해 공부하였습니다.
- SQL과 Tableau와 같은 상용툴을 공부하였습니다.
- SQL과 Tableau를 연계한 사용을 예제를 통해 학습했습니다.

다양한 대내외 활동을 통해
소통과 협업 능력을 발휘해 다양한 성과를 이루었습니다.

교육 봉사 동아리 활동



결손 가정 대상으로 학습 지도

- 교육 봉사 동아리 '나누리' ('14년 ~ '15년)
 - 결손 가정, 가출 청소년 등 학습에 어려움을 겪는 학생들을 대상으로 학업 공부와 다양한 체험학습과 진로상담을 진행했습니다.
 - 학생들과 많은 대화와 상담을 하기 위해 다양한 커뮤니케이션 기법을 탐색했습니다. 이를 통해 상대방의 상황에 따라 필요한 것이 무엇인지 분석하는 능력을 키웠습니다.

다문화 가정 자녀 멘토링



- 다문화·탈북 가정 자녀 멘토링 ('14년)
 - 다문화 가정 아이를 대상으로 1:1 멘토링을 진행했습니다.
 - 처음 안내 받은 언어 학습 부진 문제점에만 집중하는 것이 아닌 아이와 많은 대화를 통해 교우관계의 문제점을 발견했습니다. 이후, 프로그램을 자체적으로 재구성 하여 또래 아이들과의 다양한 경험을 할 수 있도록 프로그램을 재구성 하였습니다.

파란사다리 그룹 스터디



- 파란사다리 해외연수 사전 준비 ('19년)
 - 베트남 해외연수 전 언어 학습을 위한 그룹 스터디. 사전 문화 탐방을 진행했습니다.
 - 연수 기간 필요한 언어에 대한 원어민과의 그룹 스터디를 진행하고 베트남의 문화에 대한 이해를 위해 조별로 문화를 탐색하고 발표하는 시간을 가졌습니다.

Communication 능력을 향상 시킬 수 있었던 대표적인 경험!

다양한 아르바이트 경험을 하면서
커뮤니케이션과 분석을 이용해 다양한 성과를 이루었습니다.

Linc사업 조교



- 성균관대학 LINC사업단 ('16.03 ~ 17.02)
 - LINC사업단 조교를 통해 사업 운영에 대한 전반적인 지식과 사업보고서 작성에 대한 기반을 다졌습니다.
 - 캡스톤디자인 강의 관리, 가족기업 관리, 현장실습 신청 및 학생 관리, Linc+ 사업 보고서 작성에 참가해 사업단 연장이라는 성과를 이루었습니다.

이마트 아르바이트



- 이마트 서수원점 판촉&재고관리 ('17.06 ~ 10)
 - 수박 및 파인애플 판촉 업무와 청과 품목에 대한 재고관리 업무를 통해 마케팅 능력의 기반을 다졌습니다.
 - 판촉 업무를 수행하며 적극적인 홍보와 고객과의 커뮤니케이션을 통해 수박과 파인애플 품목을 계약된 일시보다 빠르게 전량 판매하며 직접적인 매출 상승과 연결시켰습니다.

백화점 사은행사장 아르바이트



- 롯데백화점 사은행사장 (18.05 ~ 19.01)
 - 사은행사 홍보 및 참여 유도, 사은행 지급 및 재고관리, VIP응대 업무를 수행하며 CRM마케팅의 기반을 다졌습니다.
 - 고객과 적극적으로 소통하여 민원을 해결하고 서비스 어플의 문제점을 발견, 본사에 직접 보고하여 문제를 즉시 보완하고 이를 통해 우수직원상을 받았습니다.

직접적인 성과를 달성 시킬 수 있었던 대표적인 경험!

1. 다양한 문화체험과 현지 학생들과 교류

- 하노이 HUST(하노이 과학 기술 대학교)의 학생들과 베트남 언어와 문화에 대해 공부하고 전통음식과 의상 체험. Saint-Paul Hospital에서의 봉사활동에 참가했습니다.
- 베트남 현지의 역사적인 장소 (Ex. 베트남 군사 역사 박물관, 하노이 문묘, 호찌민 묘소, 주석궁) 에 방문에 베트남의 역사에 대해 공부하며 다양한 문화를 이해하는 방법을 배울 수 있었습니다.

2. 연수 이후 성과보고 및 멘토링

- 베트남 연수 이후의 경험들을 바탕으로 YouTube 영상 제작 및 성과 보고 발표를 주도했습니다.
- 성과보고회 이후 다음 해외연수 참가를 희망하는 학생들을 대상으로 멘토 역할을 수행했습니다.



End of Document

수많은 데이터에서 가치를 도출해내는 데이터 분석가에게 가장 필요한
덕목은 '끈기' 와 '소통' 이라고 생각합니다.

문제 해결을 위해 다각적 시각으로 접근하며 동시에,
구성원들과의 지속적 소통을 통해 다채로운 해결책을 생각해내겠습니다.
매사에 적극적이고 책임감을 지닌 자세로 꼭 함께 하고 싶습니다.
감사합니다.