프리온보딩

PORTFOLIO

최지선

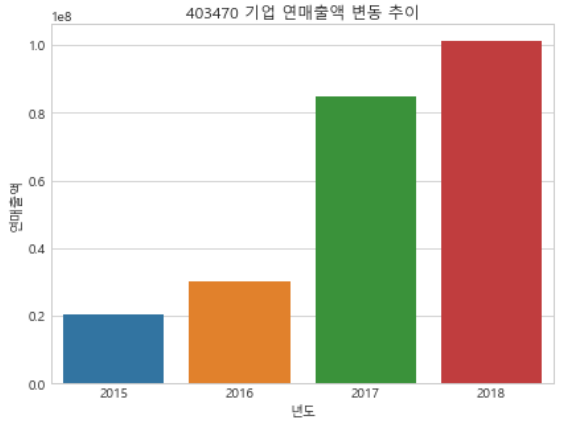
프리온보딩 프로젝트1 : 국민연금 DB 이용하여 유니콘 기업 발굴

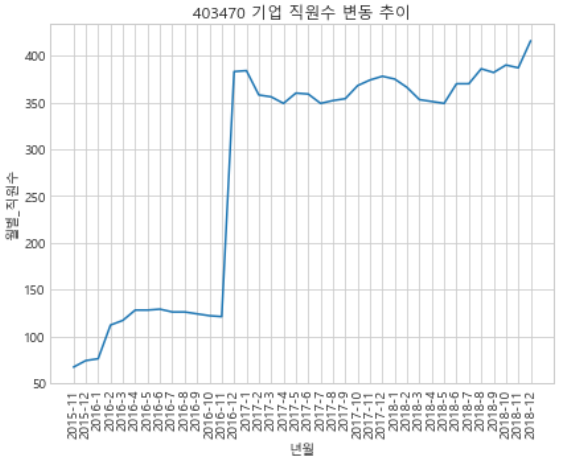
01 프로젝트 1 개요 및 결과

- **프 로 젝 트 기 간** : 2020.05.01 ~ 2020.05.07 (7일)
- **프 로 젝 트 인 원** : 6명
- **언어 및 라이브러리** : python / pandas, numpy, matplotlib
- **데이터 출처 및 형식** : 원티드랩 기업 제공 raw data / csv

프로젝트 요약

연매출액과 월별 연금보험료와 직원수에 따른 팀이 정한 유니콘 기업 기준에 부합하는 회사ID 최소 5개 추출





프리온보딩 프로젝트1 : 국민연금 DB 이용하여 유니콘 기업 발굴

02 프로젝트 1 기획 및 결과물

프로젝트 기획 순서

데이터 EDA 요약

- 데이터 내 직원수가 0명인 회사 제거
- 데이터 개수가 24개 이하인 회사 제거 (업력 2년 이상)
- 연매출액 400억 이하인 회사ID 제거 (캐치 참고)
- 직원수가 100명 이하인 회사ID 제거 (캐치 참고)
- 연매출액_변화량, 월별_직원수_변화량 컬럼 생성
- 연매출액 유지 및 증가한 회사 ID만 추출
- 직원수 유지 및 증가한 회사 ID만 추출
- 시각화 진행

프로젝트 진행 및 결과

1. 1차 기준은 캐치가 선정한 매출액 top10 스타트업을 참고하여 매출 400억 이상, 직원수 100명 이상 기준 선정
2. 2차 기준은 유니콘 기업 기준인 '최근 3개년 매출성장률이 연평균 20%이상인 기업'을 적용하여 매출액 증가율이 20% 이상인 회사 선정

최종 유니콘 선정 기업 ID : 294337 / 127366 / 469677 / 403470 / 440094

프리온보딩 프로젝트1 : 국민연금 DB 이용하여 유니콘 기업 발굴

03 프로젝트 1 마무리

진행한 업무

1. 기초 EDA
2. 스타트업 매출액 수집
3. 시각화

프로젝트 진행하며 어려웠던 점

raw data는 정제되어 있어 크게 전처리를 할 일은 없었지만, 기초 EDA에서 연매출액과 국민연금 보험료, 직원수와의 상관관계가 크지 않아 길을 잡기 어려웠지만, 유니콘 기준과 팀원들과 상의 끝에 이미 유니콘 기업으로 선정된 스타트업의 정보를 이용해 부합하는 기업을 찾기로 분석 방향을 확실히 잡으니 해결 방향이 보였다.

프로젝트 후, 깨달은 점

데이터 안에서만 문제를 해결하려고 하지 않고, 외부 데이터를 사용하여 해결해 볼 수 있는 넓은 시선을 가져야한다고 느꼈다.

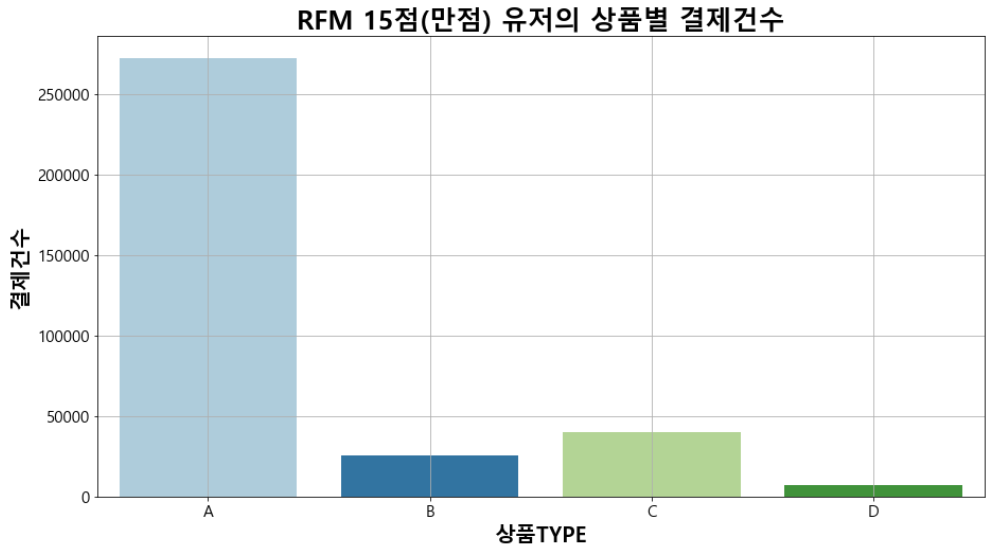
프리온보딩 프로젝트2 : 주차장 앱 이용자별 향후 이용건수 예측

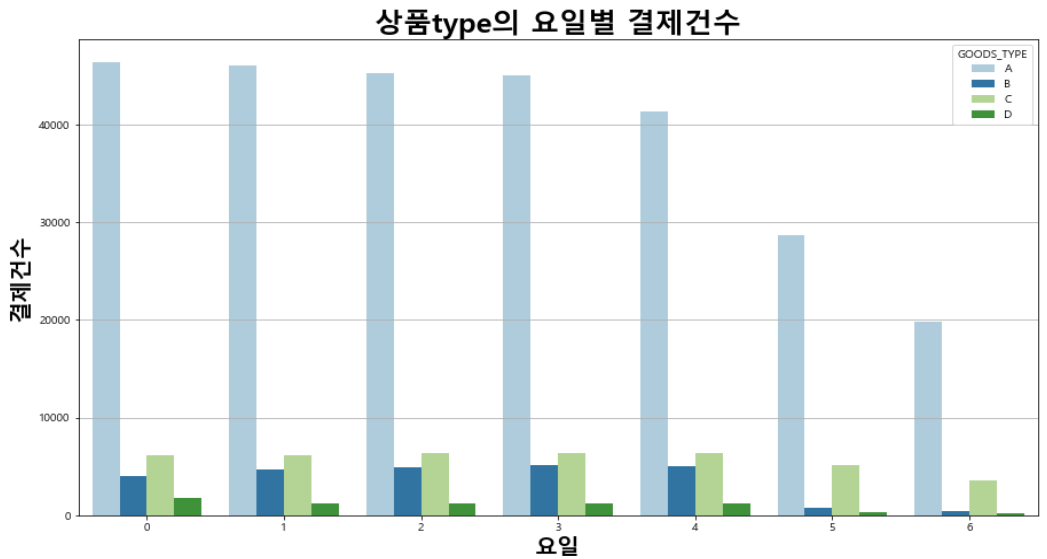
01 프로젝트 2 개요 및 결과

- **프 로 젝 트 기 간** : 2020.05.08 ~ 2020.05.14 (7일)
- **프 로 젝 트 인 원** : 6명
- **언어 및 라이브러리** : python / pandas, numpy, matplotlib, scikit-learn
- **데이터 출처 및 형식** : 모두의주차장 기업 제공 raw data, 서울시 기상데이터 / csv

- **프로젝트 요약** :

Cohort 분석과 RFM 분석을 통하여 결론과 action plan을 제시하고, 머신러닝을 통해 이용자별 10월~12월 이용건수를 예측하고, mse값이 낮은 모델과 하이퍼 파라미터를 찾음





프리온보딩 프로젝트2 : 주차장 앱 이용자별 향후 이용건수 예측

02 프로젝트 2 진행 및 결과

프로젝트 진행 및 결과

1. 코호트 분석

- 문제점 : product life Cycle 좋지 않음
- 개선방향 : 현 상품에 집중, 고객에게 긍정적 경험을 주려는 노력 필요
- action plan : 앱 환경 개선, 불법주차 신고제 운영(보상제 및 캠페인)

2. RFM 분석

- 문제점 : 고객점수 15점 만점에 4점, 5점, 7점, 6점 고객 분포가 다수.
  특히 4점 고객은 앱 1회 방문인 것으로 판단
- 개선방향 : 1회 사용 유저에 대한 복귀 장려 위해 높은 점수 그룹 확인 필요

3. 머신러닝 : 일별 분석, 월별 분석 진행

- 일일 카운트 예측에 영향력이 있는 컬럼만 이용(11개 컬럼)
- 한 달 단위 예측에 영향력이 있는 컬럼만 이용(15개 컬럼)
- 4개 모델 사용
  : GradientBoostingRegressor, XGBRegressor, LGBMRegressor, RandomForestRegressor
- **최종결론(LGBMRegressor) : mse : 0.26355187894719134 / mae : 0.10392207522470769**

프리온보딩 프로젝트2 : 주차장 앱 이용자별 향후 이용건수 예측

03

프로젝트 2

마무리

진행한 업무

1. 기초 EDA
2. 교통데이터 및 기상데이터 정보 확인 및 전처리
3. 모두의주차장 앱 이용 및 후기
4. RFM분석
5. 시각화 및 인사이트 도출

프로젝트 진행하며 어려웠던 점

데이터 안에 있는 AD1, D_TYPE, 등 변수의 의미를 알기가 쉽지 않아 기초 EDA를 통해 변수간의 관계와 의미파악을 하는데 시간이 필요했다. 변수들의 상관관계를 살펴보면서 데이터프레임을 잘 처리하기 위해 많은 공부를 하며 코딩에 더욱 자신감이 붙었다.

프로젝트 후, 깨달은 점

항상 데이터의 변수들의 의미를 정확히 알 수 없으므로, 데이터 전처리와 변수간의 상관관계를 파악하고 찾아내는 기술이 필요하다고 느꼈다. 또한, Cohort 분석과 RFM 분석과 같은 마케팅 분석 기법을 이용해 고객 행동 분석을 다양하게 시도할 수 있다는 점이다.

프리온보딩 프로젝트3 : 패션 공유 플랫폼 고객 데이터 분석

01 프로젝트 3 개요 및 결과

👉 **프 로 젝 트　기 간** : 2020.05.15 ~ 2020.05.21 (7일)
👉 **프 로 젝 트　인 원** : 6명
👉 **언어 및 라이브러리** : python / pandas, numpy, matplotlib
👉 **데이터 출처 및 형식** : 클로젯셰어 기업 제공 raw data / csv

👉 **프로젝트 요약** :

최근 3개월 간 신규 구매 고객 데이터를 분석하여 결과 도식화하는 프로젝트.
구매 소요 시간 분석, 구매 상품 트렌드 분석, 분석 후 인사이트 도출하여 마케팅 등 제안사항 도출

---

02 프로젝트 3 결과

프로젝트 진행 및 결과

1. AARRR분석에서 3가지 분석 방향을 정함

- step1. 잠재고객 확보에 주력하라
- step2. 잠수고객을 구매고객으로 연결하라
- step3. 우량고객의 파이를 늘려라
- jumping step. 고객의 시야를 확장하라

2. 분석 인사이트 결과

1) action plan

- SNS 바이럴 마케팅 포인트
  : 판매가 아닌 공유라는 BM을 고려, 인기 제품의 특성을 찾아 재고상품으로 유인
- 제품 상세 필터 기능 추가
- 장기간 미사용 고객 대상 1:1 의류할인 프로모션 제공
- CLOSET SHARE 만의 Mixed 브랜드 런칭

2) 제안사항

- 제안1 : 의류 상품 활성화에 주력할 것!
  * 전체 상품 중 의류의 비율 95% but! 2021 1분기 구매 중 가방 비율 47%
- 제안2 : Sharer에게 안심을 줄 것!
  * 제품 훼손시 합리적인 보상을 받을 수 있다는 인식 제고 필요
- 제안3 : 베스트 상품 재고 확보 채널을 다각화 할 것!
  * Sharer는 재화적 수익은 물론 환경을 생각하는 CLOSET SHARE의 가치를 공유

프리온보딩 프로젝트3 : 패션 공유 플랫폼 고객 데이터 분석

03 프로젝트 3 마무리

진행한 업무

1. 클로젯셰어 서비스 조사
2. 기초 EDA
3. 우량고객 분석
4. 인사이트 기반 아이디어 도출

프로젝트 진행하며 어려웠던 점

공유 사업과 패션 업계에 대한 도메인 지식이 부족하여 기초 EDA 후에도 방향성을 잡기가 어려웠다. 클로젯셰어의 수익 모델과 업계 현황을 충분히 조사하고 팀원들과 많은 회의를 걸쳐 원하는 목표를 설정할 수 있었다.

프로젝트 후, 깨달은 점

도메인 지식이 부족할 경우에는, 데이터 속에서 인사이트를 찾기 애쓰기 보다는 충분한 조사와 팀원과의 소통이 중요하다고 느꼈다. 특히, 이번 프로젝트에서는 팀원 모두 마케팅과 패션 공유 플랫폼에 대한 이해가 처음에는 부족했지만 집단 지성의 힘과 각자의 아이디어를 모아 분석 방향을 확실히 잡을 수 있었다. 처음에는 갈피가 잡히지 않는 것 같았으나, 프로젝트 마무리 후에는 팀원 모두 만족했던 기억에 남는 프로젝트였다.

기업연계 프로젝트 : 이커머스 기업의 매출액 향상 프로젝트

01 프로젝트 4 개요 및 결과

👉 **프 로 젝 트 기 간** : 2020.04.21 ~ 2020.05.07 (17일)
👉 **프 로 젝 트 인 원** : 6명
👉 **언어 및 라이브러리** : python / pandas, numpy, matplotlib, scikit-learn
👉 **데이터 출처 및 형식** : 기업 제공 raw data / csv

👉 **프로젝트 요약** :

식품 포장 용기 쇼핑몰 데이터를 이용하여 품목별 판매량 예측을 통한 재고관리와 고객 특성을 반영한 판매전략으로 **매출액 향상을 목표**로 진행

프로젝트 결과

(개선안)

1. 재고 관리를 위한 판매량 예측 시스템 도입 검토
2. 고객 분석에 기반한 회원 관리 및 광고 마케팅 전략 수립
3. 고객 분석에 기반한 홈페이지 UI 개선
4. 주문 데이터를 활용하여 연관분석을 하여 제품 추천 서비스 제공

(추가 개선안)

1. 판매페이지 내 사이즈별 상품 분류도를 게시, 구매 편의성 제고
2. 기존 자체제작 상품 및 서비스를 확대, 브랜드화를 통해 판매량 증가

기업연계 프로젝트 : 이커머스 기업의 매출액 향상 프로젝트

02 프로젝트 4 기획 및 결과물

☞ 프로젝트 기획 순서

1) **현황 조사** : 포장 용기 시장 조사 & 경쟁업체 서비스 조사
2) **문제 도출** : 매출 증가율 감소, 신규 고객 유입 마케팅 부재
3) **분석 계획** : 주요변수 상관 관계 분석 & 영향인자 및 주요변수 도출
4) **데이터 전처리** : 분석을 위한 유용한 새로운 데이터 셋 생성
5) **시각화 및 데이터 분석** : 시계열분석, 연관분석, 구매력과 구매여부 예측 모델링
6) **마케팅 방안 도출** : 적정 재고 관리, 상품 추천 서비스 도입

---

3. 분석 결과 - 재고관리

용기 카테고리마다 월별 판매량의 차이가 발생되고 있음.

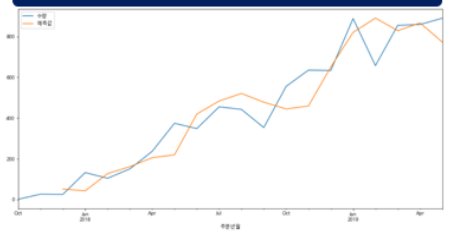
정확한 수요예측을 통하여 용기 카테고리 별 적정 재고관리를 할 수 있을 것으로 판단

[탕 용기에 대한 ARIMA 모델 결과]

Dep. Variable: D2.수량 No. Observations: 15
Model: ARIMA(2, 2, 0) Log Likelihood -91.944
Method: css-mle S.D. of innovations 103.206
Date: Fri, 01 May 2020 AIC 189.887
Time: 20:22:17 BIC 192.011
Sample: 12-01-2017 HQIC 189.865
- 02-01-2019

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1.D2.수량	-1.2158	0.213	-5.719	0.000	-1.632	-0.799
ar.L2.D2.수량	-0.7327	0.193	-3.793	0.002	-1.111	-0.354

[탕 용기에 대한 수요예측 그래프]

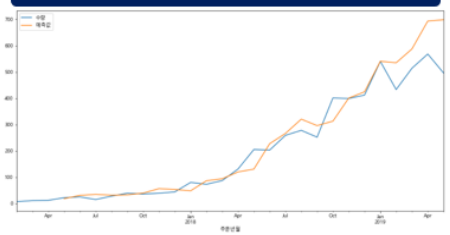


[다용도 컵에 대한 ARIMA 모델 결과]

Dep. Variable: D2.수량 No. Observations: 22
Model: ARIMA(2, 2, 0) Log Likelihood -108.035
Method: css-mle S.D. of innovations 30.825
Date: Sat, 02 May 2020 AIC 222.069
Time: 20:57:19 BIC 225.342
Sample: 04-01-2017 HQIC 222.840
- 01-01-2019

	coef	std err	z	P>|z|	0.025	0.975]
ar.L1.D2.수량	-1.1023	0.134	-8.234	0.000	-1.365	-0.840
ar.L2.D2.수량	-0.8304	0.112	-7.385	0.000	-1.051	-0.610

[다용도 컵 용기에 대한 수요예측 그래프]



3. 분석 결과 - 구매력 분석

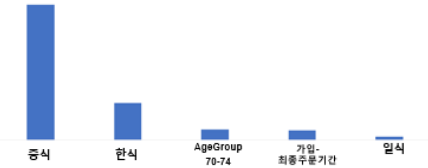
고객에 대한 마케팅 전략의 부족, 고객 데이터를 활용하여 고객의 구매력 분석

중식 제품 구매율에 따라 구매력이 증가하고, 나이 70-74 구간에 높은 구매력을 보임 → 업종 및 나이별 마케팅 전략 필요

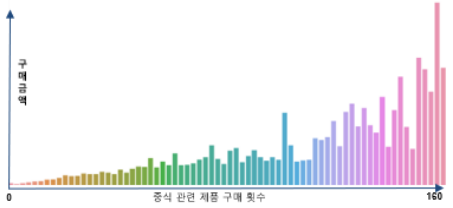
[구매력 예측 모델링 결과]

모델링 기법	정확도	Rmse	mae
다중선형회귀	70%	735671	296842
의사결정나무	78.4%	708024	269738
랜덤포레스트	77.5%	726989	271194
그래디언트 부스팅	78.7%	703328	279691

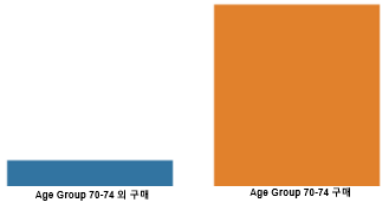
[구매력 예측 중요 변수]



[중식 관련 제품 구매에 따른 평균 구매 금액]



['Age Group 70-74' 구간 평균 구매 금액]



기업연계 프로젝트 : 이커머스 기업의 매출액 향상 프로젝트

03 프로젝트 4 마무리

맡은 업무

1. 매출 증가율 상승을 위한 비용적 측면 접근 조사
2. 데이터 결측치 & 이상치 처리 및 새로운 변수 생성
3. 새로운 데이터 셋 시각화
4. 구매력 예측 모델링
5. 고객 분석에 기반한 마케팅 전략 도출(다양한 업종 용기 마케팅 & 홈페이지 개선)

프로젝트 진행하며 어려웠던 점

로그데이터의 많은 데이터 양을 컨트롤 하기 힘들었고, 그에 따른 결측치와 이상치의 적절한 판단을 하기 어려워서 팀원들과 많은 대화가 필요했다.

프로젝트 후, 깨달은 점

데이터 전처리 및 모델 코드 작성 시, 팀원과의 원활한 코드 공유를 위해선 진행상황과 변수 설명 주석을 꼼꼼히 적어야 되겠다.

AI 프로젝트 : 저 해상도 이미지 복원을 통한 차량 판독 프로젝트

01 프로젝트 5 개요 및 결과물

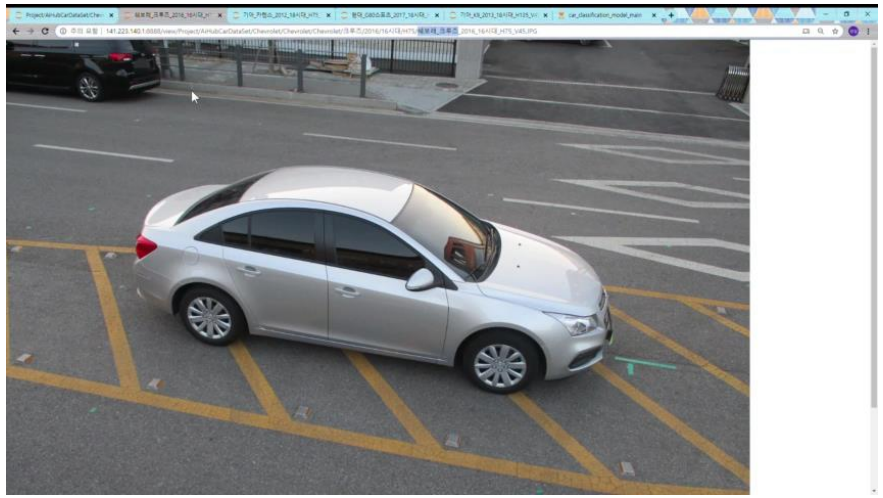
- **프 로 젝 트 기 간** : 2020.05.26 ~ 2020.06.04 (10일)
- **프 로 젝 트 인 원** : 5명
- **언어 및 라이브러리** : python / keras / Linux
- **프로젝트 선정 이유** : 사회적 문제를 인공지능을 통해 해결
- **데이터 출처 및 형식** : Aihub (https://aihub.or.kr/) / image, csv

---

- **프로젝트 요약** :

super resolution의 이미지 전처리 기법을 통한 선명한 CCTV와 블랙박스 영상 확보가 가능하고 딥러닝을 사용해서 차종 판독 모델의 정확도를 기대해 볼 수 있으며 차종과 번호판을 파악하는 시간과 비용을 줄이는 효과를 기대

- **시연 영상**



AI 프로젝트 : 저 해상도 이미지 복원을 통한 차량 판독 프로젝트

02 프로젝트 5 기획 및 결과물

논문 분석

1) Object Detection - YOLO v4 최신 연구 동향

우리가 원하는 차종을 맞추기 위한 object detection에 가장 성능이 좋은 모델로 판단

2) Super Resolution GAN 최신 연구 동향

SRGAN을 이용한 CCTV 영상 화질 복원 등 화질 개선을 위한 연구가 현재까지 지속

3) EfficientNet 최신 연구 동향

ResNet 모델 대신 정확도는 크게 차이 나지 않으면서 학습 속도는 더 빠른 EfficientNet을 이용하기로 결정

프로젝트의 기능

1) 저해상도 이미지를 고해상도 이미지로 변환

- downsampling된 이미지를 SRGAN이용 고해상도로 변환

2) Detect된 차량 이미지를 분류(차종 판단)

- YOLO v4를 이용하여 이미지에서 차량은 detect하여 crop한 이미지를 분류기(EfficientNet)에 넣어 차종 판단.

3) OCR로 번호판 문자 인식

- 번호판 detect하여 문자 인식(기존 구현된 시스템 사용)

AI 프로젝트 : 저 해상도 이미지 복원을 통한 차량 판독 프로젝트

03 프로젝트 5 마무리

맡은 업무

1. 데이터 사용 요청 (AI hub 차량 이미지 데이터(kcar))
2. 프로젝트 진행 관련 기술 논문 요약
3. 데이터셋 생성 위한 이미지 downsampling 진행
4. 기술 설명 보고서 작성

프로젝트 진행하며 어려웠던 점

1. 비전공자들로 이루어진 팀이라 기술에 관한 충분한 논의가 어려웠고, 시스템의 정확도를 위한 많은 양의 데이터 셋 생성이 버거웠다.
2. 온라인 진행으로 linux 환경에 원격접속으로 detiection 라이브러리들을 설치했지만, 오류로 인해 결국 핵심 기술인 YOLO 등을 직접 코드를 만져보지 못해 아쉬웠다.

프로젝트 후, 깨달은 점

1. 시스템 환경에 따른 라이브러리 설치와 운영에 관해 확인이 필요하다.
   (설치가 제대로 이루어지지 않으면 프로젝트 진행 과정에서 어려움이 많이 따름)
2. 현재까지 이루어진 기술 개발력을 알아보고, 새로운 AI 기술을 접목하기 위해서는 AI 분야의 깊은 관심과 논문을 통해 최신 기술 동향에 대해 파악이 굉장히 중요하다.

공공데이터 프로젝트 : 서울시 공공도서관 이용적합도 분석

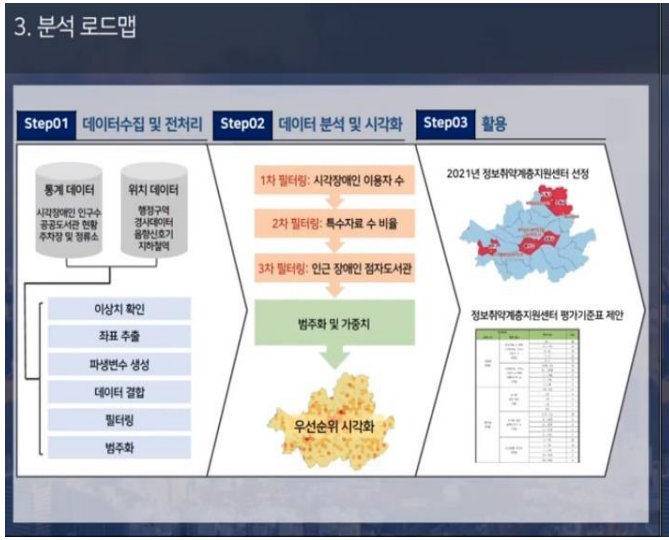
01 프로젝트 6 개요 및 결과물

- **프 로 젝 트 기 간** : 2020.09.13 ~ 2020.09.24 (12일)
- **프 로 젝 트 인 원** : 8명
- **언어 및 라이브러리** : python / pandas, numpy / R / Q-GIS
- **데이터 출처 및 형식** : 공공데이터(https://www.data.go.kr/), 통계청 / csv

---

- **프로젝트 요약** :

시각장애인의 독서권 보장을 위해 서울시 공공도서관 이용적합도 분석을 통해
시각장애인 공공도서관 지정 정책 개선을 위한 프로젝트



공공데이터 프로젝트 : 서울시 공공도서관 이용적합도 분석

02 프로젝트 6 기획 및 결과물

프로젝트 기획 순서

1) 분석 개요
- 취약계층도서관 선정 방법 & 공공도서관 내 특수 자료수와 독서보조기기 보유 수 등

2) 문제 도출
- 시각장애인 공공도서관 선정 기준에 대한 객관적 지표 부재
- 지원내용에 대한 가이드라인 부재로 특수자료 보충 지연

3) 데이터 수집
- 공공데이터포탈, 통계청

4) 데이터 전처리
- 접근성 변수들의 전처리를 위해 Q-GIS이용 위.경도 변환

5) 데이터 분석
- 변수별 가중치 산정하여 점수화를 통한 취약계층 지정 도서관 선정

6) 개선 방안 도출
- 데이터기반의 취약계층을 위한 공공도서관 선정
- 공공도서관별로 평가기준표를 제안하여 부족 항목 보완

공공데이터 프로젝트 : 서울시 공공도서관 이용적합도 분석

03

프로젝트 6

마무리

맡은 업무

1. 공공도서관 취약계층 담당자 전화 인터뷰
2. 공공데이터포털, 통계청 등 데이터 수집
3. 각 공공도서관 반경 500m, 1km 등 접근 인프라 데이터 전처리
4. Q-GIS 사용하여 위.경도 변경
5. 보고서 및 PPT 작성

프로젝트 진행하며 어려웠던 점

1. 현재 진행되고 있는 정책을 평가하고 개선하기 위한 새로운 방안책을 찾기 위해선 또 다른 정책과의 연계성까지 파악해야 했다.
2. 정량적 데이터로 접근하는 것 외에 실제 시각장애인들의 사용 편의성에 대한 민원 데이터를 얻고 싶었지만, 취약계층의 공공도서관 이용에 관한 민원 온라인 게시판이 없어서 시각장애인이 직접 느끼는 불만사항을 파악하기 어려웠다.

프로젝트 후, 깨달은 점

1. 진행하고자 하는 프로젝트를 위한 단순한 데이터 수집이 아닌 도메인 지식을 얻기 위한 논문 서치, 담당자와의 인터뷰, 실제 이용자 추이 등의 파악이 필히 필요하다.
2. 목표를 포괄적이지 않게 구체적으로 정하여 접근해야 원하는 결과값을 얻을 수 있다.

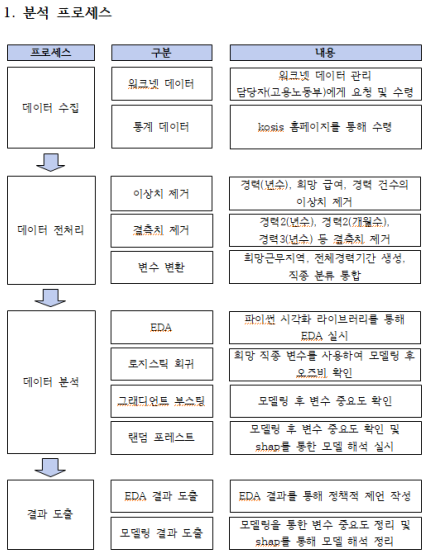
인턴십 프로젝트 : 인천시 일자리 미스매치 해소를 위한 구직 현황 분석

01 프로젝트 7 개요 및 결과물

- **프 로 젝 트 기 간** : 2020.11.02 ~ 2020.12.31 (60일)
- **프 로 젝 트 인 원** : 2명
- **언어 및 라이브러리** : python / pandas, numpy, matplotlib, scikit-learn
- **데이터 출처 및 형식** : 워크넷 / csv

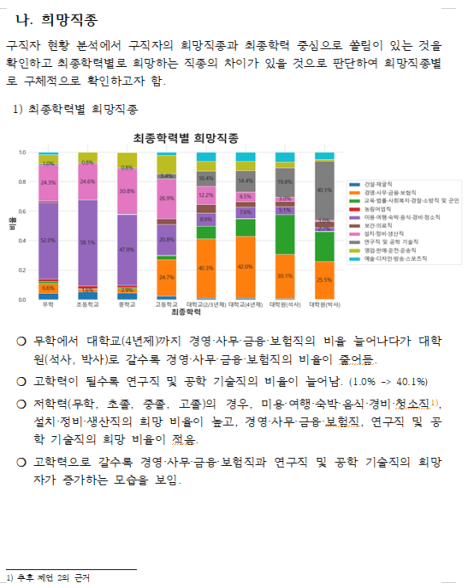
- **프로젝트 요약** :

인천광역시 일자리 미스 매칭 해결을 위해 일자리 표준분석모델을 활용하여 인천광역시 구직자의 현황 분석을 통해 워크넷 데이터의 문제점과 한계점 및 정책 제언을 위한 프로젝트



1. 분석 프로세스

프로세스	구분	내용
데이터 수집	워크넷 데이터	워크넷 데이터 관리 담당자(고용노동부)에게 요청 및 수령
	통계 데이터	kosis 홈페이지를 통해 수령
데이터 전처리	이상치 제거	경력(년수), 희망 급여, 경력 건수의 이상치 제거
	결측치 제거	경력2(년수), 경력2(개월수), 경력3(년수) 등 결측치 제거
	변수 변환	희망근무지역, 전체경력기간 생성, 직종 분류 통합
데이터 분석	EDA	파이썬 시각화 라이브러리를 통해 EDA 실시
	로지스틱 회귀	희망 직종 변수를 사용하여 모델링 후 오즈비 확인
	그래디언트 부스팅	모델링 후 변수 중요도 확인
	랜덤 포레스트	모델링 후 변수 중요도 확인 및 shap를 통한 모델 해석 실시
결과 도출	EDA 결과 도출	EDA 결과를 통해 정책적 제언 작성
	모델링 결과 도출	모델링을 통한 변수 중요도 정리 및 shap를 통해 모델 해석 정리



나. 희망직종

구직자 현황 분석에서 구직자의 희망직종과 최종학력 중심으로 쏠림이 있는 것을 확인하고 최종학력별로 희망하는 직종의 차이가 있을 것으로 판단하여 희망직종별로 구체적으로 확인하고자 함.

1) 최종학력별 희망직종

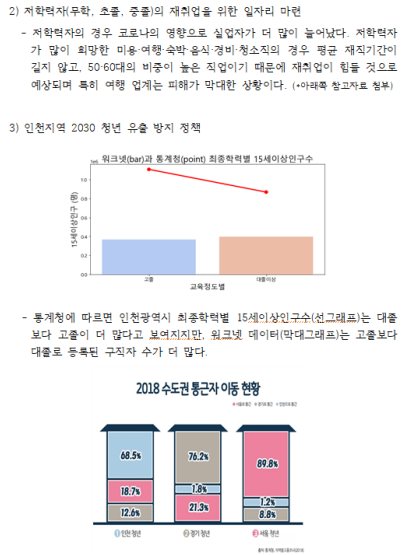
최종학력별 희망직종

○ 무학에서 대학교(4년제)까지 경영·사무·금융·보험직의 비를 늘어나다가 대학원(석사, 박사)로 갈수록 경영·사무·금융·보험직의 비율이 줄어듦.

○ 고학력이 될수록 연구직 및 공학 기술직의 비율이 늘어남. (1.0% -> 40.1%)

○ 저학력(무학, 초졸, 중졸, 고졸)의 경우, 미용·여행·숙박·음식·경비·청소직[1], 설치·정비·생산직의 희망 비율이 높고, 경영·사무·금융·보험직, 연구직 및 공학 기술직의 희망 비율이 적음.

○ 고학력으로 갈수록 경영·사무·금융·보험직과 연구직 및 공학 기술직의 희망자가 증가하는 모습을 보임.



2) 저학력자(무학, 초졸, 중졸)의 재취업을 위한 일자리 마련

- 저학력자의 경우 코로나의 영향으로 실업자가 더 많이 늘어났다. 저학력자가 많이 희망한 미용·여행·숙박·음식·경비·청소직의 경우 평균 재직기간이 길지 않고, 50·60대의 비중이 높은 직업이기 때문에 재취업이 힘들 것으로 예상되며 특히 여행 업계는 피해가 막대한 상황이다. (•아래쪽 참고자료 첨부)

3) 인천지역 2030 청년 유출 방지 정책

워크넷(bar)과 통계청(point) 최종학력별 15세이상인구수

- 통계청에 따르면 인천광역시 최종학력별 15세이상인구수(선그래프)는 대졸보다 고졸이 더 많다고 보여지지만, 워크넷 데이터(막대그래프)는 고졸보다 대졸로 등록된 구직자 수가 더 많다.

2018 수도권 통근자 이동 현황

인턴십 프로젝트 : 인천시 일자리 미스매치 해소를 위한 구직 현황 분석

프로젝트 기획 순서

1) 분석 개요
– 일자리 미스 매칭 해결 위한 분석 접근 방법 & 워크넷과 인천 인구통계 데이터 수집

2) 문제 도출
– 워크넷 데이터 내에 결측치와 오류 다수 발견
– 인천 지역 구직자 특성에 대한 현황 파악 부재

3) 데이터 수집
– 워크넷, 통계청

4) 데이터 전처리
– 구직자의 특성을 나타내지 못하는 변수 삭제 및 수치데이터 표준화

5) 데이터 분석
– 구직자의 특성을 시각화로 나타내고, 취업유무에 따른 중요 변수 선정

6) 한계점 및 개선 방안 도출
– 워크넷 데이터 부실과 많은 오류를 줄이기 위한 표준화 작업과 선택지 적용 필요
– 여성을 위한 다양한 직업 교육과 코로나로 인해 영향을 많이 받은 저학력자 재취업을 위한 일자리 마련 및 인천지역 2030 청년 유출 방지 정책을 제안하여 일자리 다양화 보완

03

프로젝트 7

마무리

맡은 업무

1. 워크넷 및 통계청 데이터 수집
2. 연봉 등 수치데이터 표준화 작업
3. 학력, 연봉, 성별 데이터 현황 분석 및 시각화
4. 인천 일자리 정책 및 문제점 도출
5. 보고서 작성

프로젝트 진행하며 어려웠던 점

1. 워크넷 원본 데이터의 범주화 및 수치 표준화 등이 되어 있지 않아 변수들의 의미 파악이 어려웠다. 실제로 주소는 우편번호와 주소가 다르게 기재되어 있는 경우도 많아 인천 지역 구직자를 분석하는데 있어 정확성이 떨어질 가능성을 안고 분석했다.
2. 취업유무 변수도 어떤 방식으로 입력되었는지 확인이 불가하여 취업유무에 영향을 준 변수를 선정하는 데 있어 역시 정확성이 떨어진다고 생각했다. 워크넷 구직자들이 워크넷 내의 구인 공고에 지원을 했는지 알 수 있는 데이터도 없어 구직·구인 미스 매칭에 대한 원인분석 및 예측 분석을 진행하기 어려웠다.

03 프로젝트 7 마무리

프로젝트 후, 깨달은 점

1. 분석에 진행하기 앞서, 해당 주제의 문제 및 시행하고 있는 정책의 문제점을 파악하는 것이 중요하고, 이러한 문제점을 해결할 수 있는 데이터를 충분히 수집해야 한다.
2. 단순히 문제점을 파악하여 기존 정책 보충 및 제언만 하는 것이 아니라, 구체적으로 어떻게 해결할 지, 해결하기 위한 보충 분석이 필요할 지 남기는 것이 필요하다.

감 사 합 니 다