데이터 분석가 이중훈 입니다.

통계와 데이터분석에 관심이 많은 신입 데이터 분석가가 되고싶은 이중훈입니다.

깃허브 주소 : https://github.com/JungHunL22

이중훈

1995.09.23

Tel. 010-2041-6211
Email. duddlswnddb@naver.com

학력

2011~2014 과천중앙고등학교 졸업
2015~2021 한신대학교 응용통계학과 (컴퓨터공학과 부전공)

경력

2022.06~2022.07 **통계청(사회조사과) 초중고 사교육비 조사**
. 조사표의 데이터 점검
. 데이터 품질점검
- 유사항이나 이상치 보완
- 조사항목 간 교차확인을 통한 오류 수정

프로젝트

2022 생체 지표데이터를 활용한 흡연자 예측
2022 날씨에 따른 역별 관광지 및 맛집 추천
2023 E-commerce 데이터 인사이트 도출

보유 스킬

- SQL
  1.subquery, join 사용 가능,
  2.로그데이터에서 원하는 지표 조회 가능
- Python
  1. 데이터 전처리와 시각화에 사용하는 라이브러리 사용 가능 (pandas,numpy / matplotlib, plotly)
  2. ML/DL에 사용하는 라이브러리 (scikit-learn, tensorflow, keras) 사용한 정형/비정형 데이터 분석 경험

- slack 및 github를 활용한 프로젝트 협업 경험
- css,html 웹 개발 기초에 대한 지식

PROJECT.1

흡연자 데이터 분석

01

ABOUT PROJECT

실제로 흡연 여부를 예측하는 데에 어떤 생체 지표가 영향을 미치는지 분석하고,
흡연자를 추적하기 위해 흡연 여부를 예측하는 모델링

Project Link

- 발표 자료
- 진행 코드

흡연자 예측

의료데이터를 활용한 흡연자 예측

맡은 역할

- 흡연 여부별 생체 지표 시각화
- 서포트벡터머신 모델링

기여도

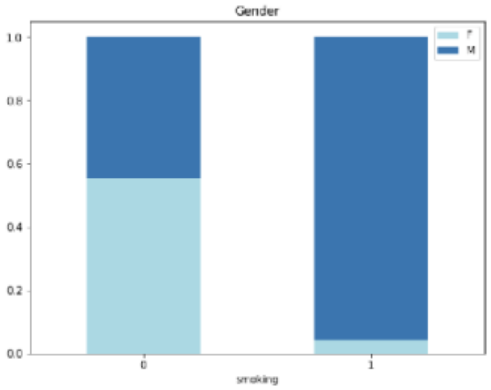
데이터 시각화 80

모델링 20

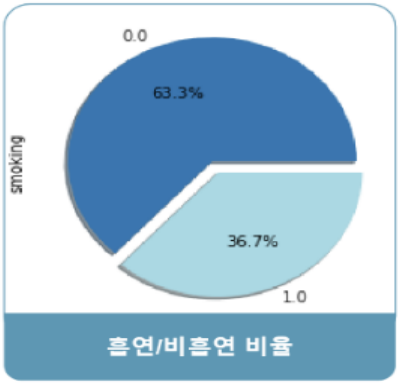
ETC

총 5명에서 프로젝트를 진행하였고, 저는 변수별 흡연여부에 따른 시각화와 서포트벡터머신 모델링 역할을 맡았습니다.
정확도,재현율을 고려해 가장 성능이 좋은 모델을 선정하였고, 피처 중요도를 통해 흡연 여부에 많은 영향을 미치는 유의미한 변수를 분석했습니다.

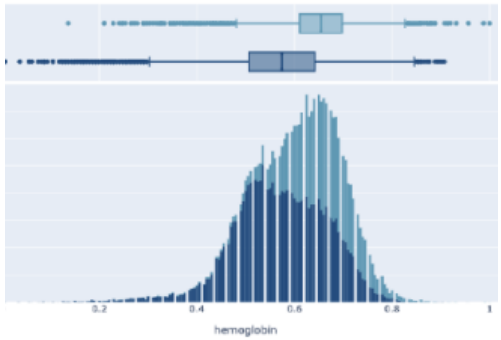
흡연 여부별 시각화



- 흡연자의 성비는 남성 96%/여성 4%
- 비흡연자의 성비는 남성 45%/여성55%



- 흡연자 36.7%, 비흡연자 63.3%



smoking
0.0
1.0

	AST	ALT	Gtp	smoking
AST	1.000000	0.740726	0.379959	0.059253
ALT	0.740726	1.000000	0.343934	0.097338
Gtp	0.379959	0.343934	1.000000	0.236619
smoking	0.059253	0.097338	0.236619	1.000000

- 간수치 지표 중 Gtp가 흡연과의 상관성이 있음

모델 성능 개선

	기본 모델	최종모델
정확도	73.6%	79.2%(5.6%↑)
정밀도	67.5%	70.3%(2.8%↑)
재현율	54.5%	75.6%(21.1%↑)
F1 Score	60.3%	72.8%(12.5%↑)
ROC AUC	69.6%	78.5%(8.9%↑)

최종모델 : Standard-Scaling(C: 3, gamma: auto, kernel: rbf) 파라미터 적용
모든 평가지표가 상승.
재현율이 21.1%상승해 가장 많이 상승함.

PROJECT.2

02

날씨에 따른 역별 관광지 및 맛집 추천 서비스

ABOUT PROJECT

"코로나19여파로 정체되어있던
서울시내 관광지 및 맛집 활성화"
"날씨 예보로 지하철 역별 혼잡도를
예측하여 소비자에게 선택지 제공"

PROJECT LINK

- 발표자료
- 진행 코드
- 웹페이지 시연 영상

날씨에 따른 역별 관광지 맛집 추천

데이터 엔지니어와 협업 프로젝트

맡은 역할

데이터 결측치 및 파생변수 생성 전처리.
모델링 성능개선.
백엔드 기능 구현.

기여도

모델링	80
백엔드	40
전처리	30

ETC

총 6명에서 프로젝트를 진행하였고,저는 데이터에 결측치를 전날과 다음날의 평균치로 대체하는 작업을 하였고, DB에 적재된 데이터를 불러와 학습시킨 모델을 이용해 유동인구를 예측할 수 있도록하는 기능을 웹페이지에 배포하는 역할을 맡았습니다.

서비스 기획 의도

개인의 성향, 상황, 날씨에 따라 약속장소 선호도 차이가 존재

날씨에 따라 유동인구를 예측하여 해당 지역의 관광지와 맛집을 추천

약속장소를 잡을 때 참고할 수 있는 웹서비스

성향에 따라
사람들이 많이 몰리는 지역인지 유추할 수 있다

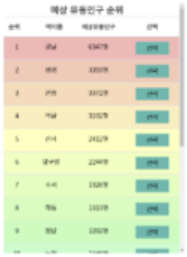
상황에 따라
거리두기 해제!
하지만 아직 사람많은 곳은 꺼려지는 경우

날씨에 따라
날씨는 관광지 선정에 중요한 요소
기후 변화와 기상요인이 관광전체에 미치는 영향이 매우 크다

주요 서비스 및 기능

날짜, 시간, 자치구 선택

지하철역 선택



관광지 및 맛집 저장

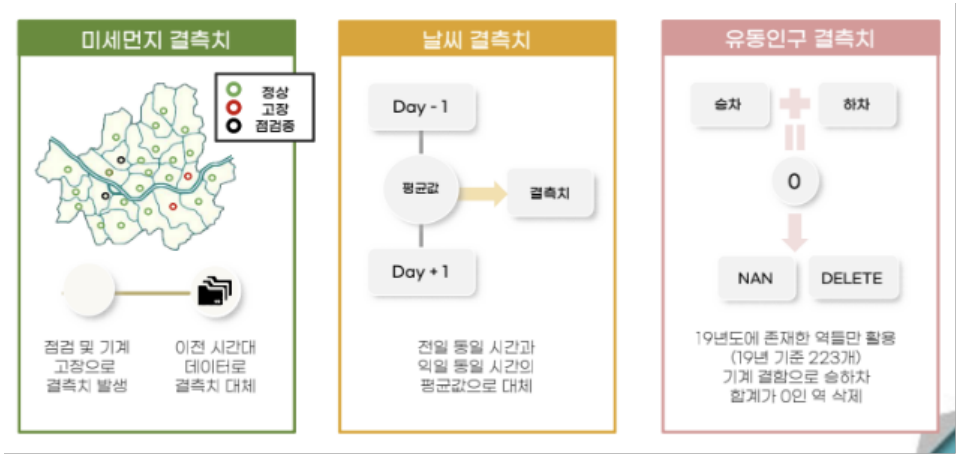
사용 데이터

날씨 및 대기환경 데이터
1. 기상청 종관 기상관측(ASOS)
2. 서울시 권역별 대기환경현황
3. 에어코리아 대기환경정보

지하철 유동인구 데이터
1. 서울교통공사 1~8호선
- 날짜시간대별 데이터
2. 공공데이터포털 역주소 및 좌표
3. 공공데이터포털 행정구 정보

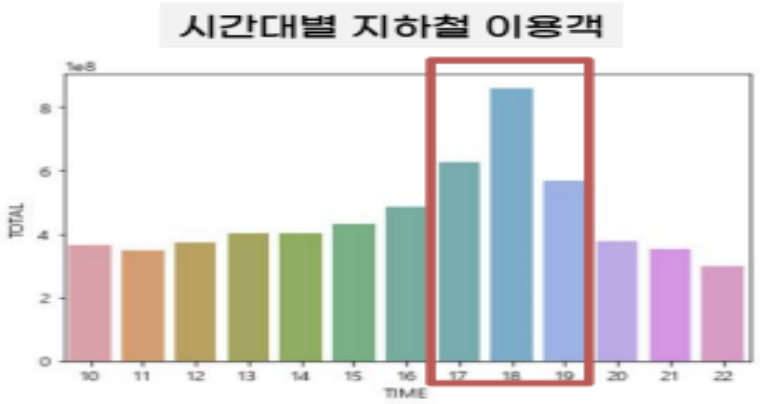
관광지 및 맛집 데이터
1. 한국관광데이터랩
- 관광지 및 맛집 정보
2. 네이버지도 관광지 근처 역정보
3. 카카오 api 좌표정보

결측치 처리 방법



사용 기술스택

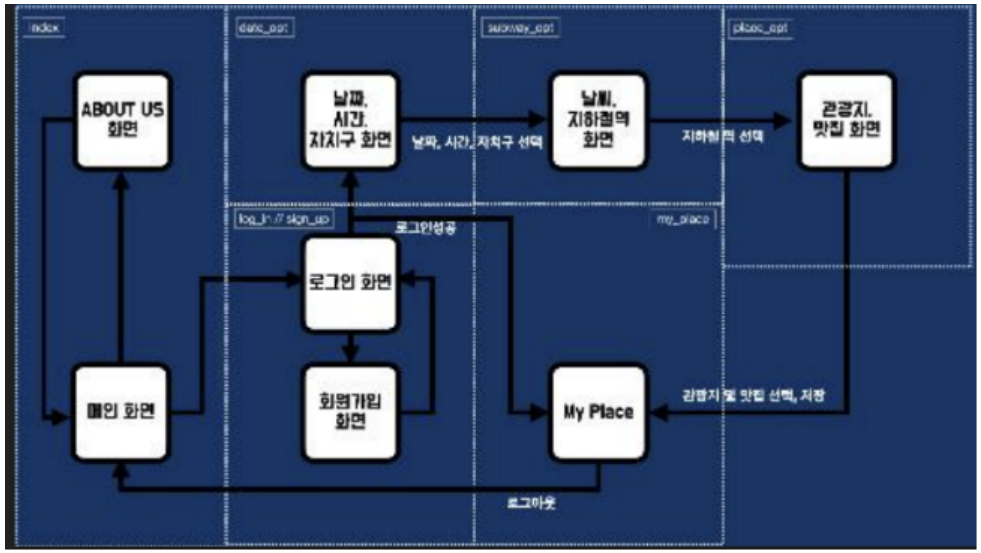
파생변수 생성



시간대별 지하철이용객 합계 중
17~19시의 이용객수 多

퇴근시간대가 아닌 시간과
퇴근시간대(17~19시)로 컬럼 생성

페이지 흐름도



PROJECT.3

E-COMMERCE 인사이트 도출

온라인 구매 기록 데이터 분석

03

ABOUT PROJECT

국가별 매출 시각화를 통한 인사이트 도출
고객별 구매지표에 따른 고객 세그먼트 분석

PROJECT LINK

- 분석 자료
- 진행 코드

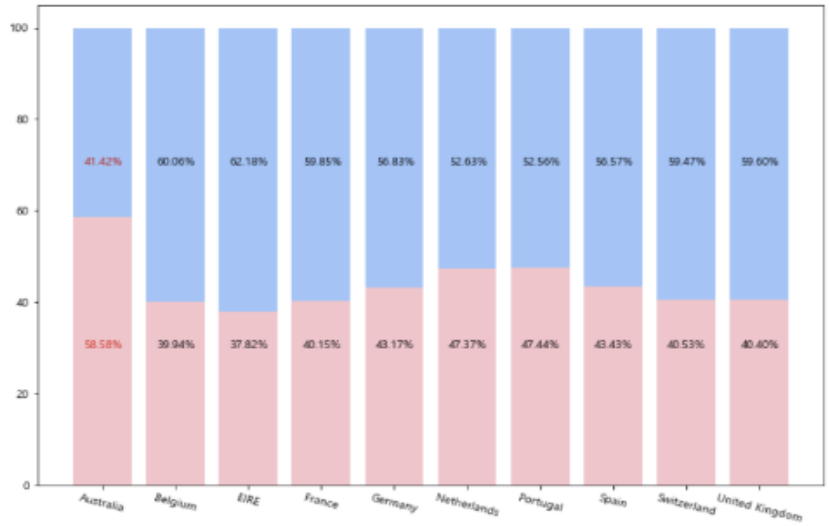
E-COMMERCE
인사이트 도출

SQL과 Python을 활용한 지표 분석

ETC

개인프로젝트로 진행하였고, 파이썬으로 데이터 전처리 후 SQL을 이용해 DB를 적재하고, 분석에 필요한 지표, 수치 데이터 테이블을 조회해 파이썬에서 시각화하는 방식으로 분석 진행.

국가별 상/하반기 매출 비율



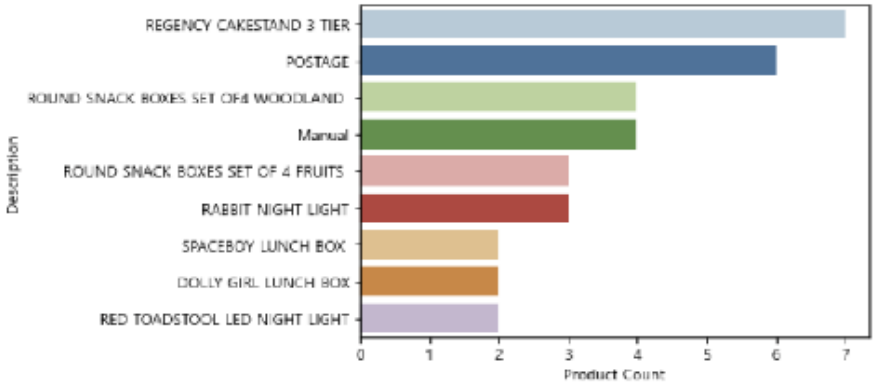
월별 매출추이를 볼 때, 대부분 하반기로 갈수록 매출이 높은 것을 확인해 상/하반기별 매출 비중을 시각화한 결과, 오스트레일리아를 제외하고 모든 국가가 하반기 매출이 높았음.

국가별 매출추이 히트맵

Month	01	02	03	04	05	06	07	08	09	10	11	12
Country												
Australia	9017.71	14695.42	17223.99	771.60	13638.41	25187.77	4964.38	22489.20	5106.73	17150.53	7242.72	nan
Belgium	1200.20	2181.07	3351.98	1989.48	2732.40	4274.82	2475.57	3554.02	4208.02	5685.38	6315.76	1417.73
EIRE	21904.19	10126.52	21674.36	7570.50	15982.19	19835.99	40905.15	16967.38	40995.49	24317.92	29473.41	6978.92
France	17740.12	8515.96	14589.55	5529.61	17614.88	16078.97	10000.19	13810.96	23428.04	33485.45	31337.09	7276.92
Germany	16910.84	9581.05	14392.69	12315.54	25751.20	13274.10	16440.98	19229.77	18091.22	31638.42	28025.02	7984.17
Netherlands	26611.10	23011.91	22416.49	2976.56	29185.88	26858.09	26.07	40327.81	26937.26	40708.65	25874.01	11728.02
Portugal	4055.71	1213.90	2660.85	1687.75	4202.23	884.46	2287.85	1221.40	1433.22	5899.46	2644.90	2808.19
Spain	10086.09	2114.50	5363.15	1785.65	3257.60	3333.21	7624.92	3346.91	5189.24	8636.94	8678.96	316.21
Switzerland	4231.23	2654.92	1870.23	2076.94	3610.01	7904.15	3762.65	4969.89	8284.86	7655.19	8118.96	nan

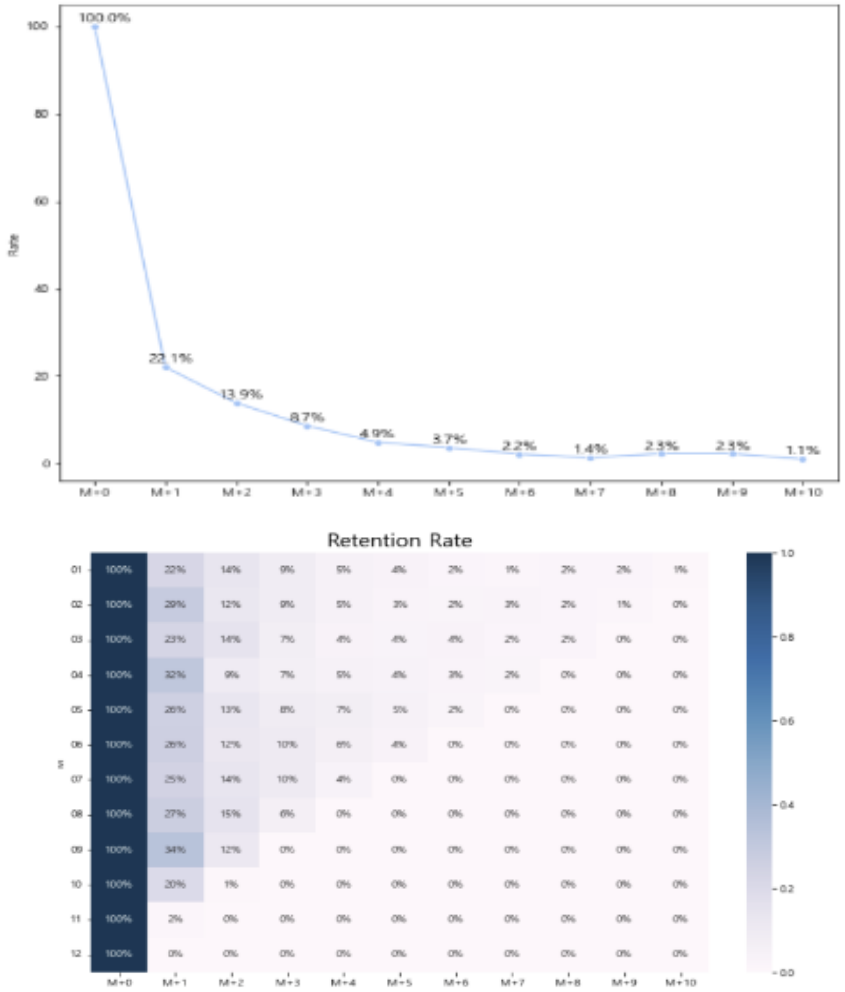
네덜란드가 구매 건수 대비 대체로 매출이 높은 것을 확인.
(구매건수는 6위인데 매출은 대체로 영국을 제외하고 가장 높았음을 알 수 있었음.)

국가별 상위 매출 상품 중 가장 많이 겹친 상품



전체 국가에서 상위매출을 차지하는 상품을 분석한 결과, REGENCY CAKESTAND 3 TIER는 총 10개국 중 7개국가인 독일 영국,오스트레일리아,스위스,아일랜드,프랑스,스페인이 겹칠정도로 판매가 잘되는 것을 알 수 있음.

고객 유지율 코호트 분석



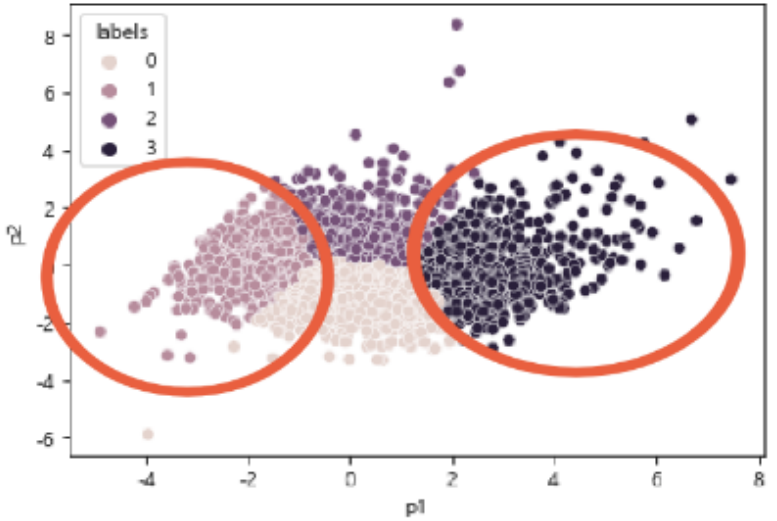
첫 구매일자가 1월인 고객 기준으로 마지막 달까지 유지된 고객의 비율은
약 1.1%. 2월부터 약 78%가 빠져나감.

M+1을 기준으로 월별로 유지율이 가장 높은 달은 9월이며 유지율은 34%
M+1을 기준으로 월별로 유지율이 가장 낮은 달은 10월이며 유지율은 20%

고객 군집별 시각화 및 변수 분석

	cluster0	cluster1	cluster2	cluster3
Recency	75.334906	180.741512	23.395492	8.760417
Frequency	3.815094	1.360340	2.220287	11.522135
Monetary	1817.486050	300.744445	486.598064	6885.006276
ATV	144.664211	23.359390	14.672518	114.078278

PCA차원 축소 후 군집별 시각화 결과
(4차원 변수를 2차원 평면에 그리기 위해 PCA를 이용해 2차원으로 축소)



0번 군집은 건당 구매금액은 높으나, 구매이력이 오래된 고객
1번 군집은 구매이력도 오래되고,자주구매하지 않으며,지출도 적은 고객
2번 군집은 구매이력이 오래되지 않았지만, 지출이 낮은 고객
3번 군집은 자주 구매하고, 총지출도 높은 고객으로 나타남.

Monetary와 Recency,Frequency변수에서 구매 성향이 가장 큰 차이를 보여
시각화 결과에서 1,3번군집이 가장 군집 간 거리가 먼것으로 예상됨.

감사합니다!
잘 부탁드립니다!

CONTACT
duddlswnddb@naver.com
010 2041 6211