

PORTFOLIO

전창민

INDEX



1. Intro

- 1) About me
- 2) University
- 3) Education
- 4) Certificate
- 5) Experiences
- 6) Awards



2. Skills

Hard Skills



3. Projects

Team Projects

CONTENTS

1. Intro



"고객이 만족할 수 있는 유의미한 인사이트를 도출할 때까지 노력하는 전문가가 되고 싶습니다."

About me

성명 전창민 **생년월일** 1997.08.21 (만 26세)

연락처 010-9412-4298 **Github** <https://github.com/jcm821>

E-mail chuncm@naver.com

주소 서울시 관악구 관악로 285

University

2016.03 ~ 2022.02	경기대학교	응용통계학전공 (경제부전공)	졸업(3.87/4.5) 전공학점 (4.09/4.5)
2013.03 ~ 2016.02	구암고등학교	인문계	졸업

Education

2023.02 ~ 2023.07	데이터분석 & 엔지니어 취업캠프	멀티캠퍼스	데이터 분석/시각화부터 SQL 머신러닝등을 중점적으로 경험하여 데이터분석 역량을 강화
-------------------	-------------------	-------	---

Certificate

2022.12	SQLd(SQL 개발자)	한국데이터산업진흥원
2021.12	컴퓨터활용능력 1급	대한상공회의소
2021.11	사회조사분석사 2급	한국산업인력공단
2020.09	ADsP(데이터분석 준전문가)	한국데이터산업진흥원

Experiences

2019.05 ~ 2019.05	2019년 이민자 체류실태 조사 및 고용조사 도급조사원	경인지방통계청 서울사무소	- 데이터 수집의 전반적인 과정과 실제 통계청 설문조사에 필요한 문항과 프로세스를 간접적으로 경험
-------------------	--------------------------------	---------------	--

Awards

2021.05	2021 통계데이터분석 경진대회 장려상	경기대학교 응용통계학과	- PM과 프로젝트 PPT작성 - 실제 데이터를 활용하여 시계열분석의 과정을 경험했다는 점에 의
---------	-----------------------	--------------	--

CONTENTS

2. Skills

언어	python	pandas, numpy, matplotlib, seaborn, sklearn, folium, tensorflow, keras등의 라이브러리를 활용해서 데이터분석 및 머신러닝/딥러닝을 했습니다.
	R	통계분석에 사용되는 라이브러리 외에 dplyr, ggplot2, konlp등의 라이브러리를 사용해봤습니다.
DB	SQL Server	CRUD 스키마와 테이블생성, 데이터 추가, 수정, 삭제 등의 과정을 경험했습니다.
	Oracle	
	MySQL	
Visualization	Tableau	시각화 구현과 매개변수 및 계산필드 사용, 데이터추출을 통한 Tableau Public업로드, 대시보드와 스토리 기능을 사용해봤습니다.
Framework	Django	클론코딩으로 게시판 앱을 만들어 댓글, 로그인 및 로그아웃 기능 구현, 게시판을 배포하는 과정을 경험했습니다.

3. Projects

교량 데이터분석을 통한 안전등급 안전점검체계 보완

기간 2023.05.03 ~ 2023.05.24

구성 총 5명(데이터분석 및 머신러닝)

주요업무

- 교통량 관련 시각화 및 다른 요소들과 연관 지어 추가 시각화 진행
- 랜덤포레스트 모델 적용하여 안전등급 분류

기술스택

Python(pandas, numpy, sklearn, matplotlib, seaborn)

목적

분류모델을 통해 기존의 안전점검체계의 한계점을 보완하여 안전검사 품질을 개선하고자 했음

링크

[Project/Semi-Project/\[D25\]문제해결 프로젝트 4조 전창민 포트폴리오.pdf at master · jcm821/Project \(github.com\)](#)



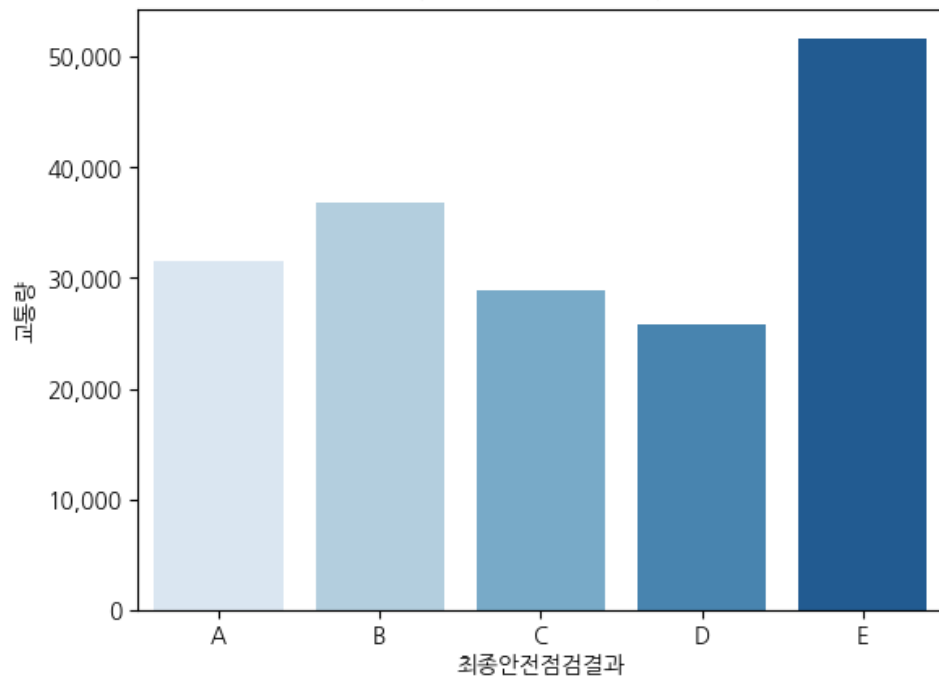
"전국교량표준데이터를 활용하여 데이터 시각화와 머신러닝 위주의 분석을 진행했습니다."

3. Projects

데이터 분석

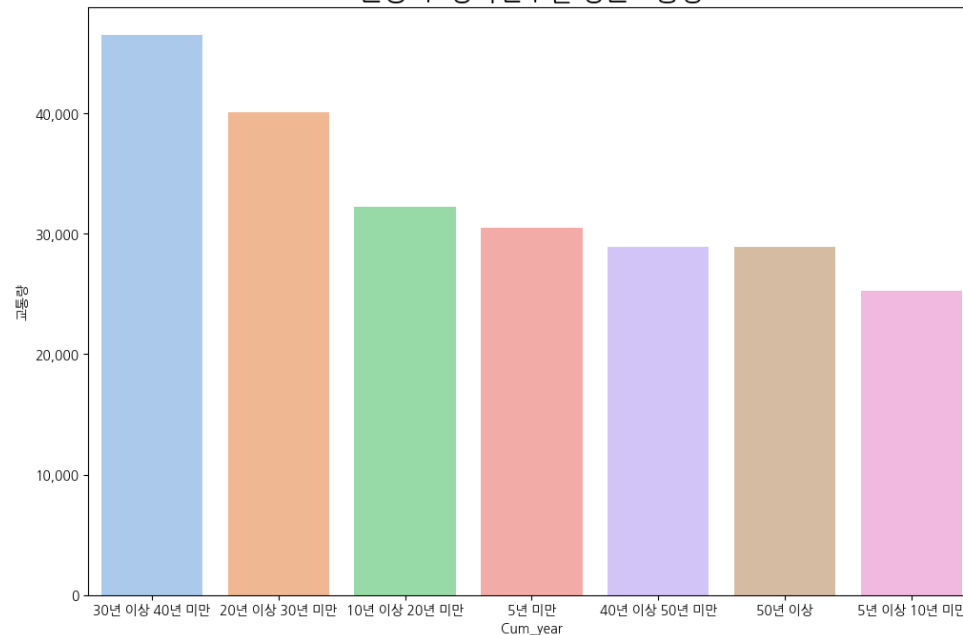
“안전등급에 영향을 주는 요인 중 하나로, 교통량을 선정했습니다.”

최종안전등급별 평균교통량



E등급 교량의 평균 교통량이 높음
교통량이 많을수록 낮은 등급 가능성 ↑

준공 후 경과연수별 평균교통량

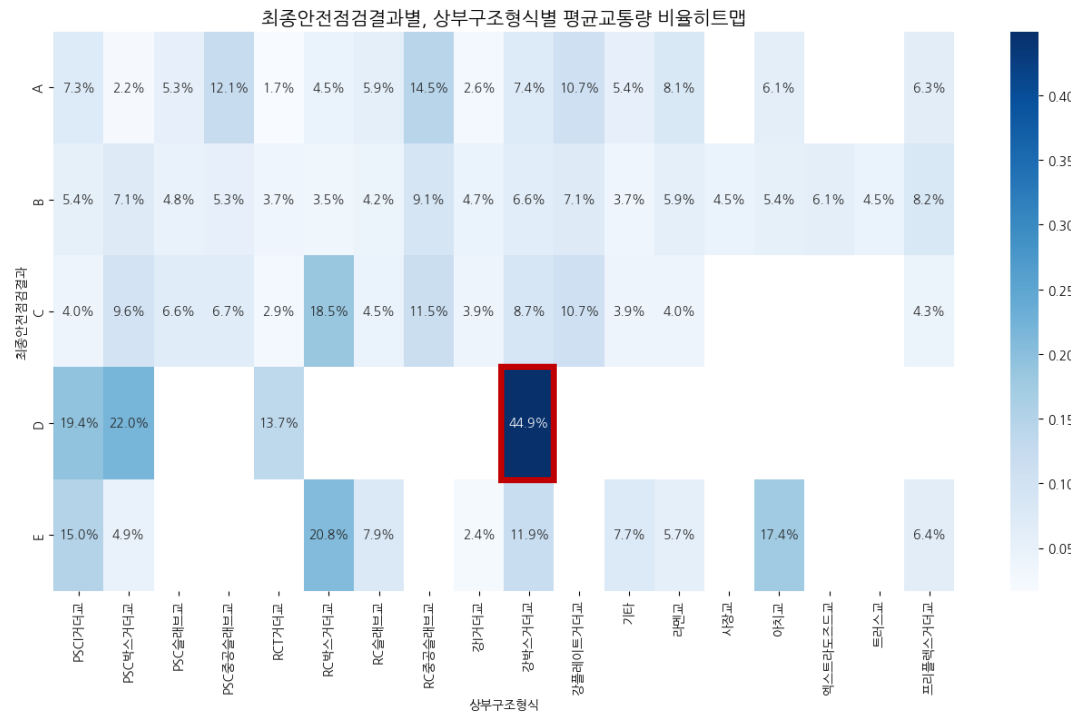


경과연수가 꽤 오래된 교량들의 평균교통량이 많음
오래된 교량일수록 마모, 손상이 심할 것으로 예상

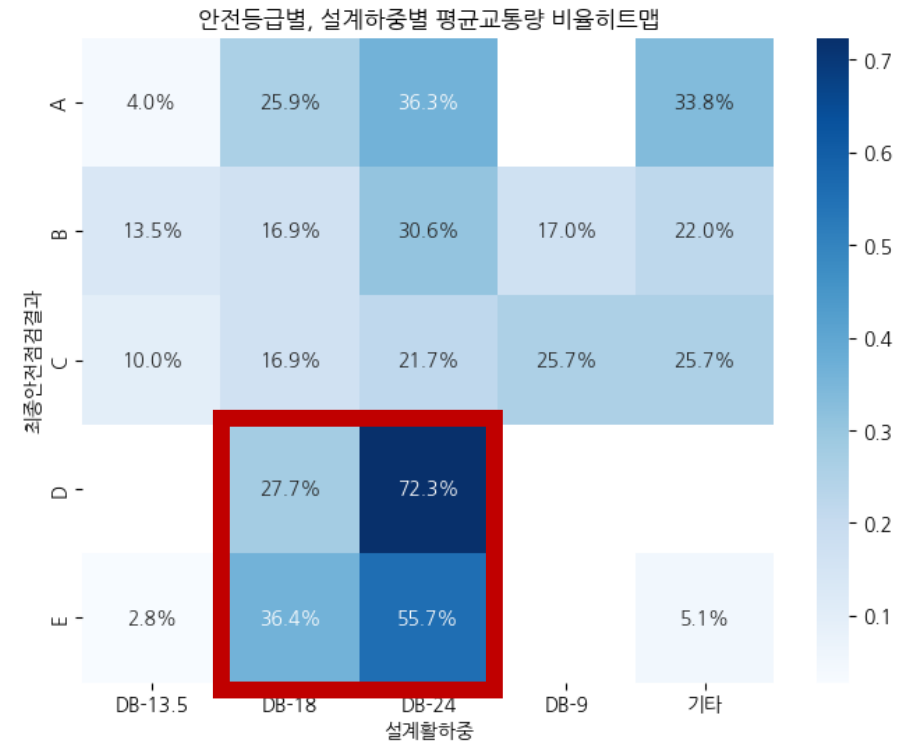
3. Projects

데이터 분석

"추가적으로 교통량과 다른 요소들 간의 상관관계를 파악하고자 히트맵 시각화를 시도했습니다."



안전등급이 D이고 상부구조형식이 강박스거더교인 교량의 평균교통량 비율이 높음



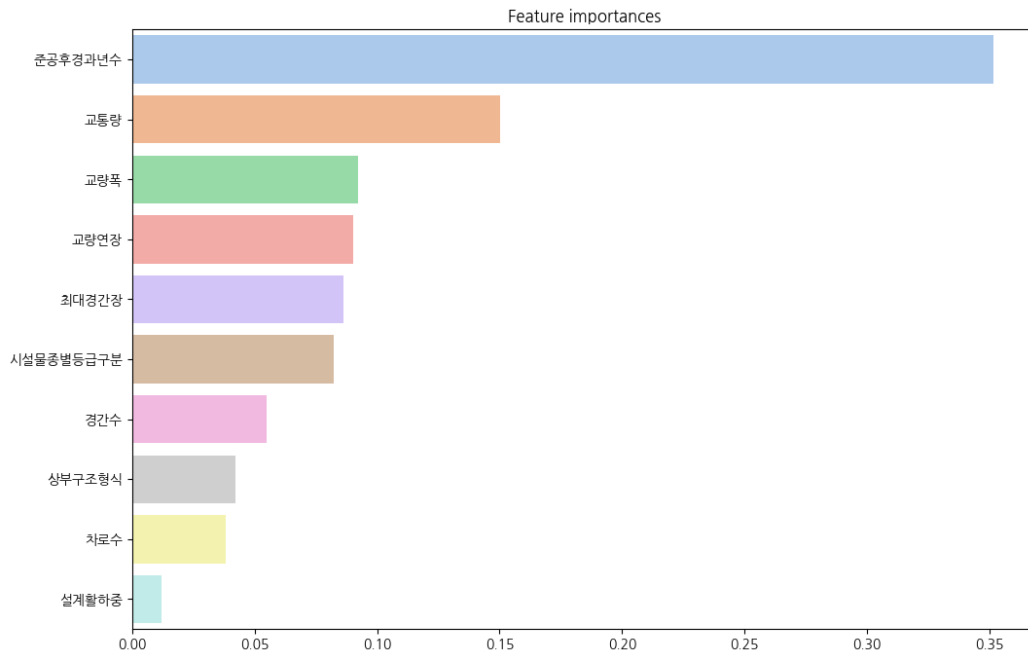
최종안전점검결과가 D,E로 분류된 교량 중 설계하중 규모가 큰 DB-24와 DB-18가 차지하는 비중이 높음

3. Projects

머신러닝

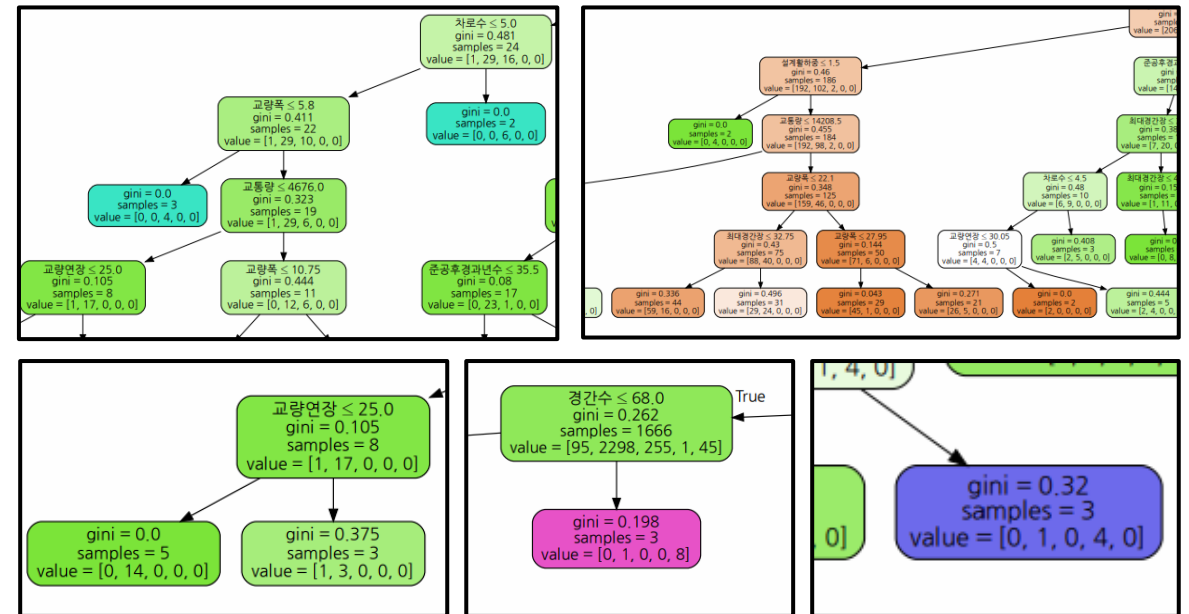
“랜덤 포레스트 모델적용을 통해 중요변수 시각화와, 트리맵 시각화를 통해 데이터적으로 안전등급의 분류를 확인했습니다.”

중요요인변수 시각화



최적모델을 토대로 중요요인을 시각화 했을 때, 준공 후 경과연수, 교통량, 교량폭, 교량연장, 최대경간장, 시설물종별등급구분, 경간수, 상부구조형식, 차로수, 설계활하중 순으로 도출

랜덤포레스트 트리맵 시각화



- 결정트리가 어떻게 구성되고, 각각의 규칙노드를 통해 안전등급이 분류되는지 시각적으로 파악
- 머신러닝 모델링을 통해 데이터적으로 안전등급을 분류할 수 있음

3. Projects

각 변수와 최종안전점검결과 간 관계 시각화 결과 선정된 중요 변수의 타당성 확인

전체 변수 모델링 결과 중요도 순위

순위	변수	중요도
1	준공후경과년수	0.352
2	교통량	0.150
3	교량폭	0.092
4	교량연장	0.090
5	최대경간장	0.086
6	시설물종별등급구분	0.082
7	경간수	0.055
8	상부구조형식	0.042
9	차로수	0.038
10	설계활하중	0.012

각 변수 별 최종안전점검결과 관계 시각화 결과
연관성이 높다고 판단한 중요 변수가
대체적으로 상위권 분포

구현사항

- Python을 활용하여 데이터프레임에서 유의미한 인사이트 도출을 위해 다양한 시각화 시도
- 중요변수들을 통해 머신러닝 모델을 활용하여 안전등급을 재분류

선정기술

- 교량점검체계를 통해 안전등급을 재분류하는 것이 목적
- 회귀모델보다 분류모델을 활용하여 정확도를 조절하는 것이 적합하다 판단되어 분류모델 사용
- 다양한 모델을 시도하여 랜덤포레스트와 Xgboost 모델 선정, 정확도, 트리맵의 시안성 등을 고려하여 랜덤포레스트 모델 최종 선정

성과

- 데이터 시각화를 통해 유추한 중요 요인들과 중요요인을 포함하여 머신러닝 모델에 적용 후, 중요요인을 시각화 했을때 유추했던 중요 요인들이 대체적으로 상위권에 분포
- 이를 통해 팀원들과 시각화를 했던 방향의 타당성 입증

리뷰

- 데이터 결측치를 대체하기 위해 여러 방면에서 고려해야 한다는 것을 알게 되었음
- 분류모델 선정에 있어서 정확도만이 아니라 시각화, 주제의 목표와 관련하여 고민해야 함을 경험하게 되었음
- 무심코 지나가던 교량에 대해서 경각심을 가지게 된 계기가 되었음

3. Projects

온라인쇼핑 트렌드 분석을 통한 마케팅 활용방안

기간 2023.06.20 ~ 2023.07.31

구성 총 5명(맡은 파트: 데이터분석 및 머신러닝)

주요업무

- API데이터를 이용하여 카테고리 별 분석 및 시각화
- 시계열 SARIMA모형을 사용하여 PC부품 클릭량 예측 시도

기술스택

Python(pandas, numpy, statsmodel, sklearn, pmdarima, matplotlib, seaborn, geopandas)
Tableau

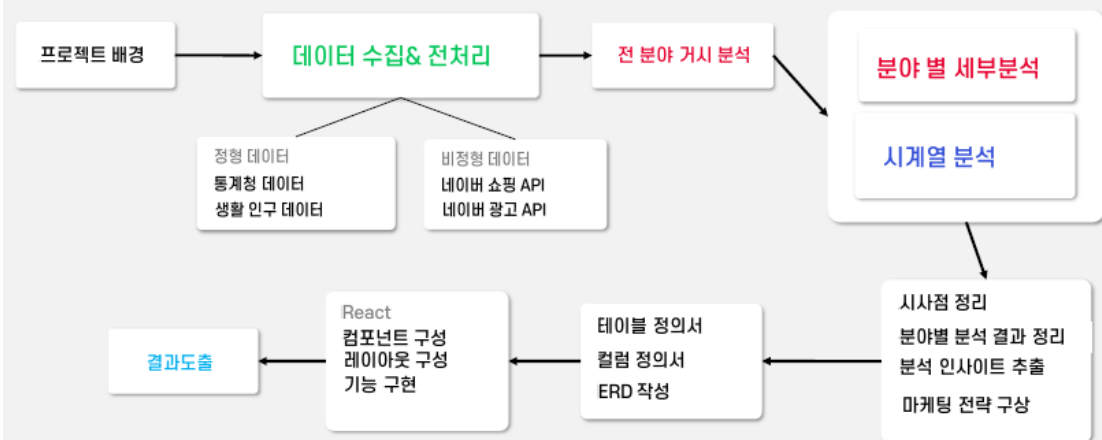
목적

카테고리 별 세부분석을 통해 항목들의 트렌드를 파악 후 웹 사이트를 통해 마케터가 필요로 할 정보를 제공하는 것을 목표로 했음

링크

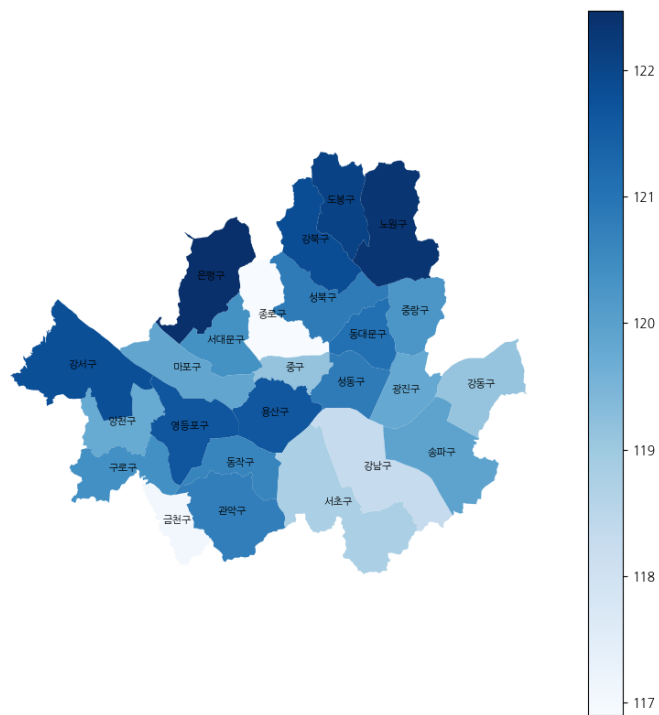
[Project/Final-Project/\[D25\]1조 포트폴리오 전창민.pdf at master · jcm821/Project \(github.com\)](#)

전체 개요



“정형 데이터인 통계청데이터와 반정형 데이터인 네이버 API데이터를 사용하여 분석과 머신러닝, 분석결과에 대한 웹사이트를 구현했습니다.”

데이터 분석



“먼저 파이썬 시각화를 통해 데이터를 파악 후, 보다 효과적이고 가시성을 위해 태블로를 사용하여 시각화자료를 구성했습니다.”

3. Projects

데이터 분석

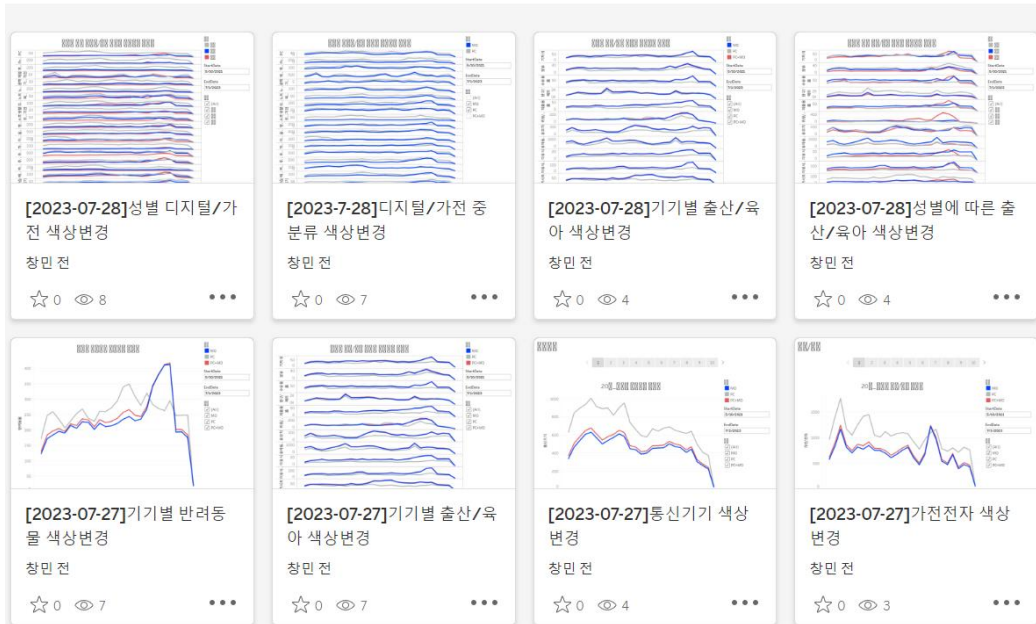
Tableau Public에 업로드 된 시각화 자료들

창민 전

프로필 편집

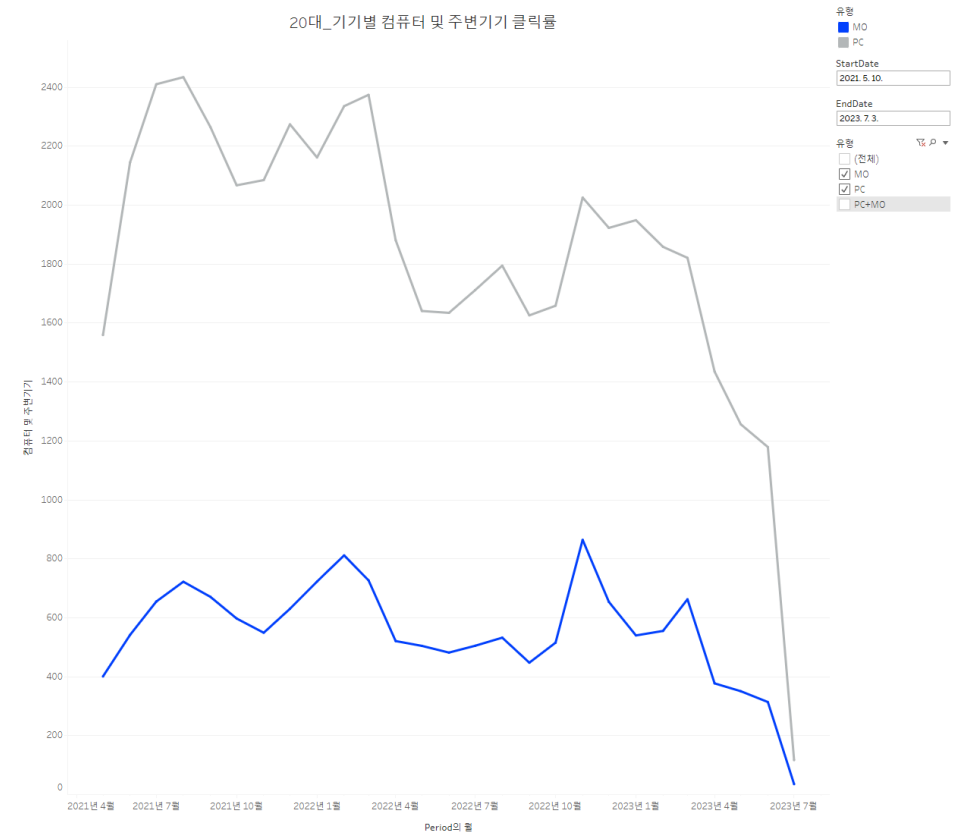
비주얼리제이션 50 즐겨찾기 0 팔로우 0 팔로워 0

비주얼리제이션 만들기



"카테고리 시각화 결과물은 Tableau public에 업로드 하여 관리했습니다."

20대_사용기기별 컴퓨터 및 주변기기 클릭량



"시각화 자료들은 대시보드화 하여 스토리를 구성하여 한 페이지에서 숫자 버튼 클릭을 통해 한눈에 보기 쉽게 배치했습니다."

CONTENTS

3. Projects

머신러닝

SARIMA모델 적용

20대 여성의 특정 기간에 유의성이 보였던 항목들을 비교

기준이 되는 모델 -> 임의로 PC부품으로 선정

auto_arma 메서드를 사용 -> 최적의 모델 탐색

최적의 시계열 모델
SARIMA(1,0,2)(0,0,2,12)

```
SARIMAX Results
=====
Dep. Variable:          y      No. Observations:      785
Model:      SARIMAX(1, 0, 2)x(0, 0, 2, 12)  Log Likelihood      1201.638
Date:              Mon, 24 Jul 2023      AIC      -2389.275
Time:              14:51:13      BIC      -2356.615
Sample:      05-10-2021      HQIC      -2376.717
              - 07-03-2023

Covariance Type:      opg
=====
              coef      std err      z      P>|z|      [0.025      0.975]
-----
intercept      0.0074      0.003      2.782      0.005      0.002      0.013
ar.L1      0.9662      0.006     151.682      0.000      0.954      0.979
ma.L1     -0.0220      0.013     -1.660      0.097     -0.048      0.004
ma.L2     -0.2101      0.016    -12.825      0.000     -0.242     -0.178
ma.S.L12    -0.2229      0.024     -9.222      0.000     -0.270     -0.176
ma.S.L24    -0.0976      0.018     -5.495      0.000     -0.132     -0.063
sigma2      0.0027      4.82e-05     56.451      0.000      0.003      0.003
=====
Ljung-Box (L1) (Q):              0.09      Jarque-Bera (JB):      28442.64
Prob(Q):              0.76      Prob(JB):              0.00
Heteroskedasticity (H):          6.90      Skew:              2.51
Prob(H) (two-sided):          0.00      Kurtosis:             32.06
=====

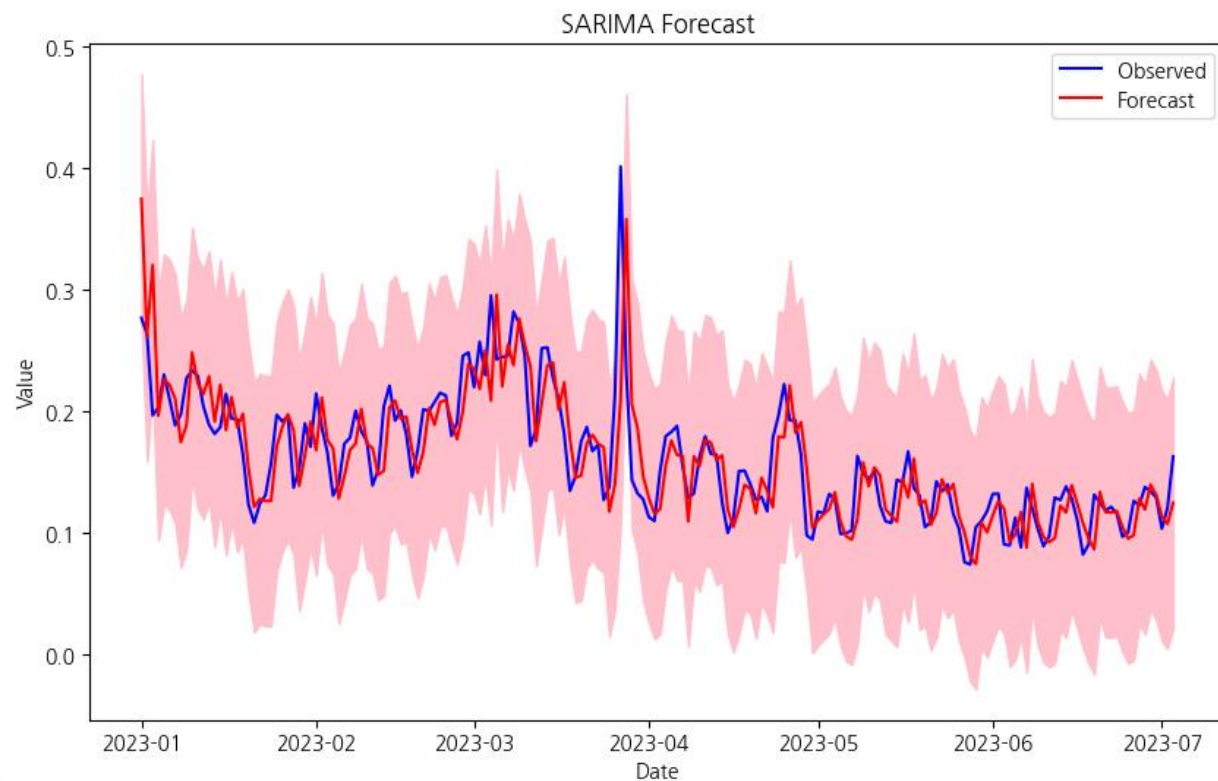
Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

"auto_arma 메서드를 사용하여 오른쪽의 결과와 같이 해당 카테고리 품목에 대한 최적의 SARIMA모델을 도출했습니다."

3. Projects

머신러닝

SARIMA 시계열모델 구간예측 결과



MSE

0.0010423

RMSE

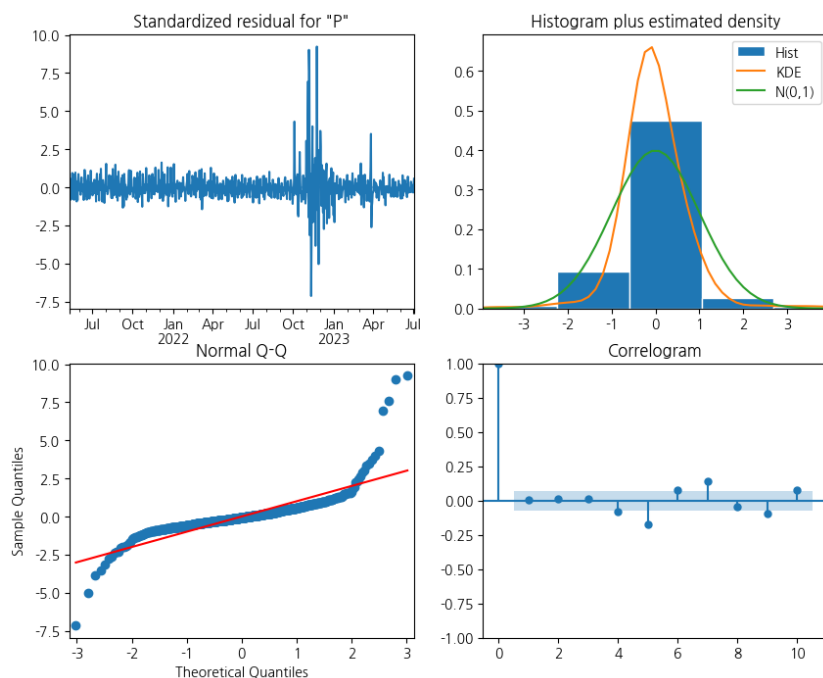
0.0322855

"23년 1월 1일부터 데이터의 마지막인 7월3일
까지 구간을 설정하여 시계열예측을 시행했습
니다."

3. Projects

머신러닝

잔차분석과 단위근 검정



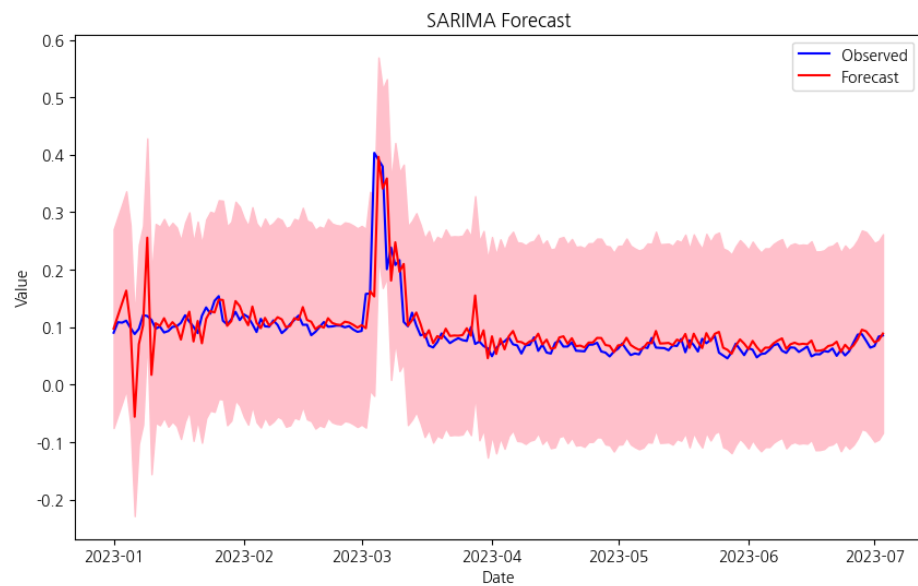
```
ADF Statistic: -5.379224968334608
p-value: 3.7446759617040443e-06
Critical Values:
  1%: -3.438938229437747
  5%: -2.8653304587462944
 10%: -2.568788425002056
```

“잔차분석과 단위근 검정을 통해 해당모델의 정상성을 확인했습니다.”

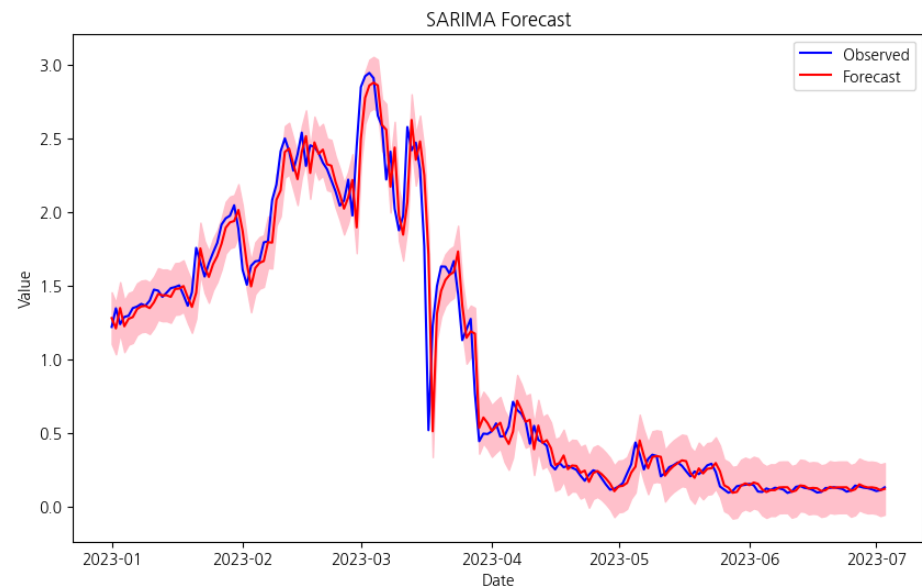
3. Projects

머신러닝

(동일 데이터) PC 카테고리 모델에 적용

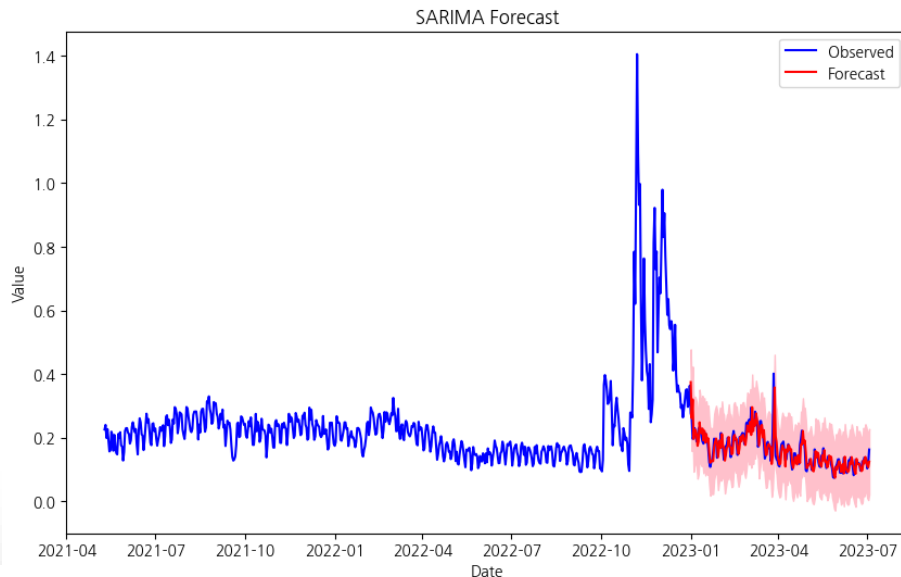
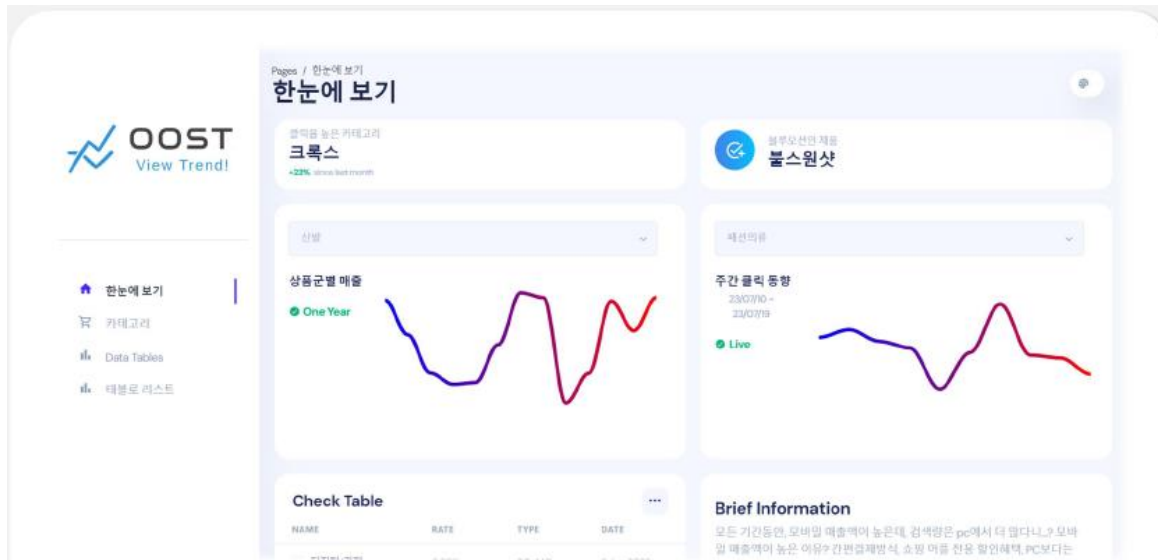


(다른 데이터) 여성 전체 연령대_모니터 카테고리 모델에 적용



"다른 카테고리에 적용하여 모델의 확장성을
검증했습니다."

3. Projects



구현사항

선택기술

성과(의의)

리뷰

- Python과 Tableau를 활용하여 카테고리 별 시각화 결과를 도출하고 웹서비스에 시각화 결과를 적재
- 마케터가 보고자 하는 카테고리에 대한 매출액, 클릭량, 경쟁강도 등의 정보가 제공됨
- 추가적으로 특이점을 보인 항목의 클릭량 예측 시도
- 카테고리 분석을 통해 특정 항목의 클릭량 변화에 대한 이유 파악
- 사용자에게 필요한 시각화 자료를 태블로 대시보드/스토리 기능을 통해 웹사이트에 구현
- 시각화 결과를 마케팅 전략에 활용
- API데이터를 사용하여 예측모델을 구현하려 시도
- 데이터 분석과 머신러닝 과정을 통해 보고자 하는 관점에 대해 세부적으로 고민해보게 되는 계기가 되었음
- 시계열 모델을 적용하여 결과를 마케팅 전략에 활용하기에는 많은 문제점이 존재한다는 것을 인식
- 데이터분석과 머신러닝을 통해 특이점이 보이는 데이터에 대해서 세부적으로 고민해보게 된 좋은 계기가 되었음
- 머신러닝을 구현하기에 부족한 데이터 수 -> 계절성분, train-test데이터셋 분할 어려움
- 시계열모델을 일반적으로 확장시키기 위해서는 추가적으로 사회적 영향요인을 고려하여 그에 따른 자료수집과 데이터 보완이 필요하다는 사실을 느끼게 되었음