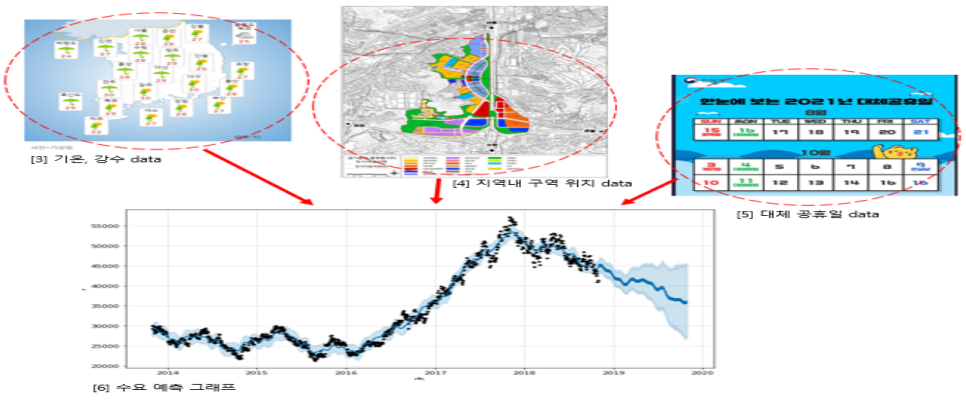
PoC Lab – 아이펠톤 프로젝트 계획서

개발아이템명	*Soctopus*		
소속	*Aiffel SOCAR* 캠퍼스 *3*기		
신청팀	중.꺾.마	담당퍼실	

□ 프로젝트 아이템 개요(요약)

아이템 소개	· 쏘카의 고객들은 다양한 목적으로 카셰어링 서비스를 이용한다. 성별, 연령대와 같은 기본적인 정보 외에도 서비스를 어떤 방식으로(언제, 얼마나 등) 이용했는지에 따른 데이터를 통해 고객의 특성과 트렌드를 파악할 수 있다. 이를 위해 차량 수요에 주요 영향을 미치는 독립변수(설명변수)를 미시적으로 접근, 발견하여 지역 및 구역별로 예측하고, 회귀 모델과 시계열 모델을 고려한 동적 회귀모델을 구축하여 예측 성능을 높인다.
아이템의 특징 및 차별성	· 지역별 독립변수(설명변수)와 종속변수(반응변수)간 상관관계가 다를 수 있다는 가정하에, 외부 데이터를 수집하거나 피처 엔지니어링 과정을 거쳐 독립변수를 생성 또는 발견하여 선형 회귀 분석과 시계열 모델을 활용하여 지역별로 차량 수요를 예측한다. · 대부분의 수요 예측은 재고 관리(Inventory Management) 관점을 중심으로 활용된다. 본 프로젝트에서는 수요 예측을 통해 우리 고객이 어떤 특성과 패턴을 가지고 있는지 그 이면의 인사이트(Customer Insight)까지 발견하고자 한다. 이를 통해서 단순 예측에만 그치는 것이 아니라, 발견된 인사이트를 바탕으로 고객 친화적인 서비스를 제공하고 수익을 더욱 창출할 수 있는 방향을 설정할 수 있다.
이미지	

1. 문제인식 (Problem)

1-1 프로젝트의 목표 및 목적**(**필요성**)**

- 정확한 수요예측

  - 카셰어링 산업에서는 수요예측을 통해 서비스 가격, 차량종류, 쏘카존의 위치 등이 결정된다. 또한 유휴차량을 최소화하거나 공급의 부족으로 감당하지 못한 수요를 수익으로 연결하여 영업이익에 크게 기여할 수 있다.
  - 수요를 정확하게 예측하는 것은 불가능에 가깝다. 하지만, 과거의 추이와 예측 가능한 설명변수가 있다면, 예측의 신뢰성을 높일 수 있을 것이라 예상된다.

- 정량적 기법을 적용한 수요예측

  - 과거 자료로부터 그 추세나 경향을 파악하여 미래의 수요를 예측하는 시계열 예측**[2]**과 수학적으로 인과관계를 나타내는 기법 중 독립변수(설명변수)의 변화에 따른 종속변수(반응변수)의 변화를 예측하는 인과형 예측**[1]**을 결합하여 많은 데이터를 종합하여 예측 가능성을 높인다.

1-2 아이템의 독창성

- Feature Engineering

  - 보다 정확한 예측을 구현하기 위해 변수들과의 상관관계를 분석한다.
  - 종속변수(반응변수)에 영향을 미치는 독립변수(설명변수)는 기본적으로 주어진 데이터셋의 features 보다 더 많을 수도 있고, 변형을 시켜 유의미하게 만들 수도 있다.
  - 따라서, 수요에 영향을 미칠 수 있는 주요 변수들을 탐색하고 검증하여 예측 정확성을 높일 수 있도록 한다.

- 수요예측을 세분화

  - 쏘카 서비스는 전국적으로 이용 가능하다. 이는 전국적으로 수요가 발생한다는 뜻인데, 수요는 지역별로 다를 수 있다. 그 이유는 다음의 예로 설명한다. 공장이 밀집한 도시는 주말의 개념이

일반 상업 도시와 다를 수 있다. 공장의 휴일은 교대근무로 time off 가 주말에 무조건적이지 않다. 또한, 한 지역내에서도 주거밀집지역과 업무밀집지역에 따라 카 셰어링 수요가 다를 수 있다.

- 위 내용을 검증하기 위해 지역별로 수요를 예측하고 지역별로 독립변수와 종속변수의 상관관계를 분석하여(독립변수와의 상관관계, R^2, p-value 값 등) 지역별 수요예측의 필요성을 검증한다.
- 지역 내 특정 구역을 구분하여 쏘카존의 위치를 시각화하고, 이를 구역별 예약률과 비교하여 쏘카존의 위치를 추가할 때 참고자료가 되도록 데이터를 이용한다.

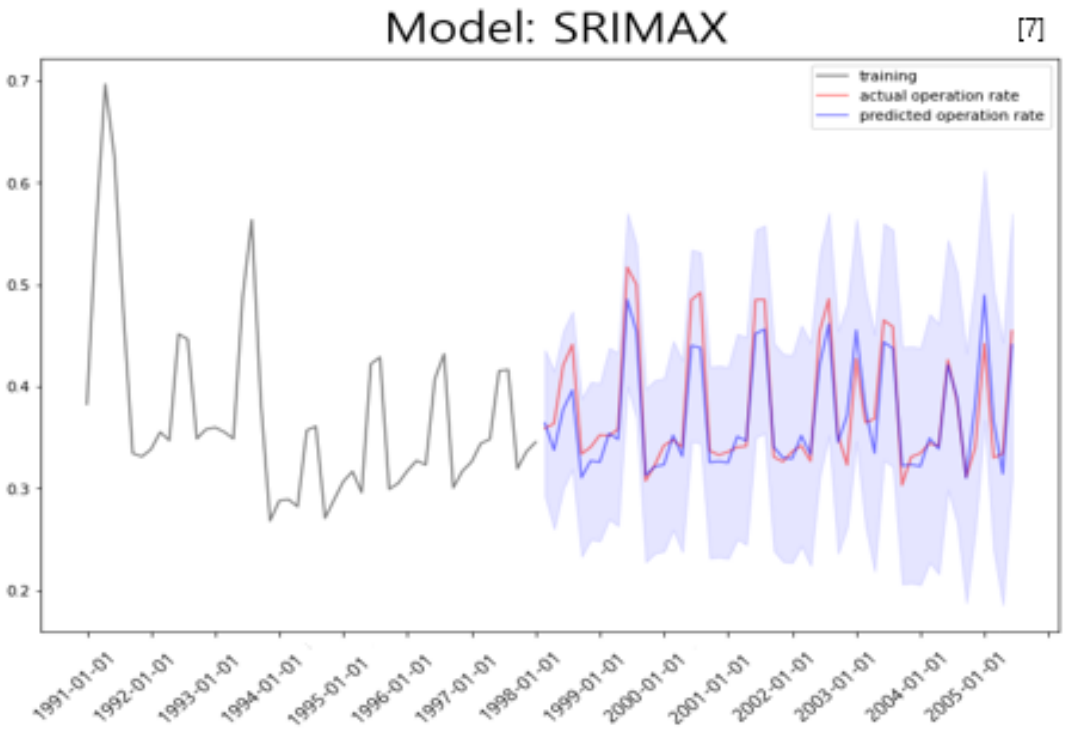
2. 개발 및 연구 내용

2-1. 구현 내용 상세

1. Feature Engineering
   A. 기존 데이터 외 온도, 강수여부, 대체 공휴일 여부 데이터를 추가한다.
   B. 데이터를 지역별로 데이터셋을 나누어 담고, 구역별로 위도, 경도 데이터를 추가한다.

2. Regression
   A. 프로젝트의 목적은 수요를 예측하는 Task 이다.

3. Time Series Analysis
   A. 시계열 자료와 외생 변수를 모두 다루는 SARIMAX 모델 적용[2]
   B. SARIMAX 는 시계열 모델 ARIMA 모형과 외생 변수를 모두 고려한 동적 회귀 모형(Regression with ARIMA error / Dynamic Regression Model) 이다.



4. Linear Regression Analysis

   A. OLS 선형 회귀 검정을 통해 설명변수와 반응변수간의 상관관계를 확인한다.

   B. 선형 회귀 분석의 기본 가정 4가지(잔차의 독립성, 잔차의 정규성, 잔차의 등분산성, 모형의 선형성)을 검증하여 회귀 모형의 예측 능력을 확인한다.[1]

OLS Regression Results

Dep. Variable:	price	**R-squared:**	0.678
Model:	OLS	**Adj. R-squared:**	0.678
Method:	Least Squares	**F-statistic:**	1.269e+04
Date:	Mon, 13 Nov 2017	**Prob (F-statistic):**	0.00
Time:	22:24:25	**Log-Likelihood:**	-84517.
No. Observations:	6028	**AIC:**	1.690e+05
Df Residuals:	6026	**BIC:**	1.691e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
intercept	9587.8878	7637.479	1.255	0.209	-5384.303	2.46e+04
area	348.4664	3.093	112.662	0.000	342.403	354.530

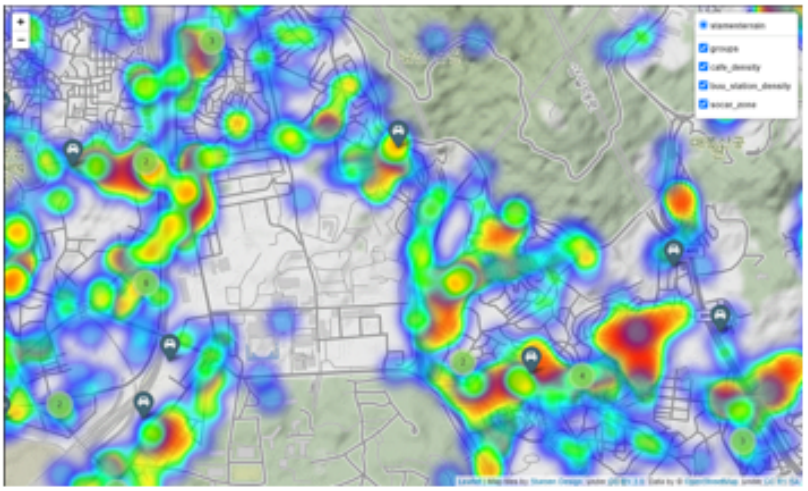
Omnibus:	368.609	**Durbin-Watson:**	2.007
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	349.279
Skew:	0.534	**Prob(JB):**	1.43e-76
Kurtosis:	2.499	**Cond. No.**	4.93e+03

[8] OLS 선형회귀 분석

5. GIS(Geographic Information System)

   A. 지리정보를 컴퓨터 데이터로 변환하여 활용하는 정보시스템

   B. 지리정보시스템을 이용하여 구역별 쏘카존의 위치를 적용하여, 구역별 카 셰어링 횟수를 파악하고 쏘카존의 증감 의사결정에 근거가 되도록 시각화 한다.



[9] GIS 시각화(Folium

2-2. 개발 아이템 기대효과

- 다양한 설명변수를 통해 지역별로 더 높은 수요예측을 할 수 있다.
  - 지역별 근로자의 휴일은 다를 수 있고, 한국의 대체 공휴일이 존재하기 때문에 이 점을 새로운 설명변수로 두어 차량 수요예측의 성능을 더 높일 수 있다.
  - 연령대별 선호하는 차량이 다를 수 있으므로, 주기적인 수요예측을 통해 쏘카존 별 고객이 선호하는 차종을 공급하여 수익을 극대화할 수 있다.

- 쏘카존 위치 설정의 기본값을 정의할 수 있다.
  - 기존 실습 중 쏘카존 위치설정과 관련된 노드에서 카페와 버스 정류장이 중복된 곳으로만 쏘카존의 위치를 가정하였다. 하지만, 경제력이 약한 대학생들이 이용할 수 있도록 대학가 근처에 쏘카존을 위치시키거나 쏘카 이용 시작부터 종료까지 서비스 이용에 불편함을 최대한 줄일 수 있도록 주거밀집지역에 쏘카존을 위치시킨다면 더 높은 수익을 창출할 수 있을 것이라는 가정을 해 보았다.
  - 현재 데이터에는 수집이 불가능하나 수요예측 데이터 셋에 이용된 쏘카존 데이터의 위치기 있다면, 더 정확한 예측이 가능 할 것이라 예상할 수 있다.

3. 실행 계획

3-1. 기간내 프로젝트 구현 완성을 위한 전략

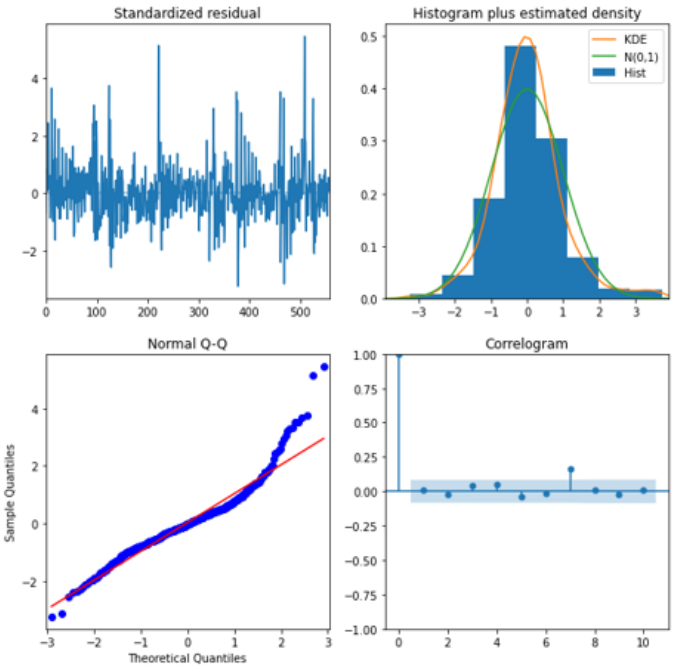
- 데이터 EDA
  - 데이터 info. 및 describe 등 탐색
  - 데이터 시각화

- 외부 데이터 수집
  - 대상 기간의 지역별 기온 및 기상 데이터 수집
  - 대상 기간 대체 공휴일 수집
  - 쏘카존 및 지역 아파트, 대학 위치 데이터 수집

- 데이터 전처리
  - 결측치 확인 및 처리
  - 이상치(outlier) 처리
  - Label Encoding(범주형 → 연속형)

- Feature Engineering
  - 외부 데이터를 기존 보유 데이터와 merge
  - 날짜 데이터를 월, 요일, 주차 별로 분리

- Linear Regression Analysis
    - OLS(Ordinary Least Squares) 회귀분석; 잔차의 독립성, 잔차의 정규성, 잔차의 등분산성, 모형의 선형성 검증[1]



[10] 선형 회귀의 기본 4 가지 가정

- 모델 구성 및 학습
    - SARIMAX 적용을 위한 모수 Grid Search
    - 모델 학습[2]

- GIS 를 통한 위치 시각화
    - Heatmap 형식으로 생성하기 위해 레이어 적용 그룹(아파트, 대학, 쏘카존) 만들기
    - 쏘카존 마커 클러스터 만들기
    - 그룹 레이어 컨트롤 박스 만들기
    - 각 그룹 밀도 map에 추가
    - folium 을 통한 map 시각화

- 모델 평가 및 결론
    - 모델의 예측 성능 확인을 위해 Epoch 별 accuracy, validation accuracy & loss 추이 시각화
    - 평가지표로 RMSE(Root Mean Square Error; 평균 제곱근 오차)를 이용하여 성능 확인
    - 지역별 수요예측의 차이를 확인하고 가설을 검증

3-2. 아이펠톤 기간 내 마일스톤

Task	목표기간	세부내용
데이터 ***EDA***	*2023.1.3. ~ 2023.1.5.*	*'3-1*
외부 데이터 수집	*2023.1.3. ~ 2023.1.5.*	*'3-1*
데이터 전처리	*2023.1.6. ~ 2023.1.10.*	*'3-1*
Feature Engineering	*2023.1.11. ~ 2023.1.13.*	*'3-1*
Linear Regression Analysis	*2023.1.13. ~ 2023.1.18.*	*'3-1*
모델 구성 및 학습	*2023.1.19. ~ 2023.1.27.*	*'3-1*
GIS 를 통한 위치 시각화	*2023.1.30. ~ 2023.1.31.*	*'3-1*
모델 평가 및 결론	*2023.2.1. ~ 2023.2.2.*	*'3-1*
발표 준비	*2023.2.3. ~ 2023.2.7.*	발표 요건에 따라 준비

3-3. 팀장 및 팀원의 역할 분배

순번	주요 담당업무	역할 상세	인원
1	데이터 ***EDA***, 데이터 전처리	전반적인 데이터 탐색 및 전처리, ***feature engineering***	*1*
2	외부 데이터 수집, ***Linear Regression Analysis***	외부 데이터 수집, ***feature engineering***, ***OLS*** 검정	*1*
3	모델 구성 및 학습	***Time Series model*** 구성 및 학습, 그 외 ***support***	*1*

4. Reference

[1] 통계적 회귀분석에 관한 연구

https://dcollection.yonsei.ac.kr/public_resource/pdf/000000104524_20221223143906.pdf

[2] 2-Step SARIMAX를 이용한 단기 주택용 전력수요 예측 기법

ttps://dcollection.korea.ac.kr/public_resource/pdf/000000257640_20221223144138.pdf

[3] 기상청

[4] “용인시”, 경기용인 플랫폼시티 도시개발사업 토지이용계획도

[5] 법제처

[6] “widisocs.net, 토닥토닥 Prophet – 시계열 회귀를 위한 딥러닝, 2022년 12월 23일 접속, https://wikidocs.net/141382

[7] Aiffel Going Deeper(Socar)_S03, “시간과 계절은 공유 차량 수요에 어떤 영향을 미칠까?”

[8] “귀퉁이 서재”, DATA - 17. 최소자승법(OLS)을 활용한 단순 선형 회귀 (Simple Linear Regression), https://bkshin.tistory.com/entry/DATA-17-Regression

[9] Aiffel Going Deeper(Socar)_S03, “쏘카존(Socar Zone), 최적의 위치를 찾아라!”

[10] Aiffel Going Deeper(Socar)_S03, “공유 차량의 소요 예측! 무엇을 고려해야 할까?”