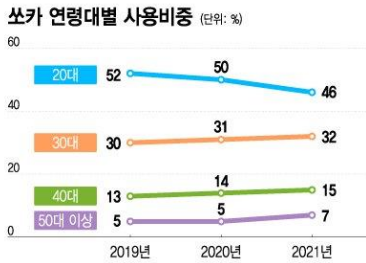



PoC Lab – 아이펠톤 프로젝트 계획서

| | | | |
|--------|---------------------|------|-----|
| 개발아이템명 | 카셰어링에서의 데이터 기반 수요예측 | | |
| 소속 | AIFTEL 쏘카캠퍼스 3 기 | | |
| 팀명 | 우리 쏘카 타 | 담당퍼실 | 문소정 |

□ 프로젝트 아이템 개요(요약)

| 아이템 소개 | ·경기도 지역의 쏘카 사용자 연령 중 20, 30 대의 연령층과 행복주택 주거지를 데이터 기반 수요예측 모델링 | | | | | | | | | | | | | | | | | | | | |
|---------------|---|-------|-------|-------|-------|-----|----|----|----|-----|----|----|----|-----|----|----|----|--------|---|---|---|
| 아이템의 특징 및 차별성 | <p>전개 방향은 신규수요지를 예측 할 수 있는 모델 구축을 목표로 하며 ,</p> <ol style="list-style-type: none">1) 각종 논문 및 카셰어링 비즈니스 분석을 통한 수요예측에 필요한 Feature 를 도출하고2) 도출 된 Feature 기반 학습에 필요한 내,외부 Data 를 수집한다3) 기존 Data 와 수집 Data 를 Classification 하여(K-means 등) 수요지에 대한 탐색 및 예측 가능한 모델을 만들어본다4) 탐색된 수요지에 대해 수요예측모델을 적용하여(딥러닝 모델등) 신규 쏘카존 설치와 기존 쏘카존의 조정과 변경을 실시한다. 시간대 및 차종, 요일 예측치등 제공가능한 예측치를 생성하여 카셰어링 비즈니스를 제안한다 | | | | | | | | | | | | | | | | | | | | |
| 이미지 | <div><div><p>쏘카 연령대별 사용비중 (단위: %)</p><table><thead><tr><th>연령대</th><th>2019년</th><th>2020년</th><th>2021년</th></tr></thead><tbody><tr><td>20대</td><td>52</td><td>50</td><td>46</td></tr><tr><td>30대</td><td>30</td><td>31</td><td>32</td></tr><tr><td>40대</td><td>13</td><td>14</td><td>15</td></tr><tr><td>50대 이상</td><td>5</td><td>5</td><td>7</td></tr></tbody></table><p>그래픽: 이지혜 디자인기자</p><p>머니투데이_08/16/22</p></div><div><p>■ 경기행복주택 추진 지구 ~ 경기도 내 17개 시·군 29개지구 10,409호</p><p>경기도_ 행복 주택현황</p></div></div> | 연령대 | 2019년 | 2020년 | 2021년 | 20대 | 52 | 50 | 46 | 30대 | 30 | 31 | 32 | 40대 | 13 | 14 | 15 | 50대 이상 | 5 | 5 | 7 |
| 연령대 | 2019년 | 2020년 | 2021년 | | | | | | | | | | | | | | | | | | |
| 20대 | 52 | 50 | 46 | | | | | | | | | | | | | | | | | | |
| 30대 | 30 | 31 | 32 | | | | | | | | | | | | | | | | | | |
| 40대 | 13 | 14 | 15 | | | | | | | | | | | | | | | | | | |
| 50대 이상 | 5 | 5 | 7 | | | | | | | | | | | | | | | | | | |

1. 문제인식 (Problem)

1-1 프로젝트의 목표 및 목적(필요성)

◦쏘카 사용자 연령대 중 20, 30 대 연령층을 공략하기 위한 연구

- 온디맨드모빌리티서비스가 익숙한 연령대로써 자차수요가 낮으며, 카셰어링 수요자가 가능한 지역을 연구
- 경기도 행복주거지역의 입주민 유형은 사회초년생, 신혼부부, 대학생으로 입주자의 유형이 골고루 분포
- 도심지이지만 대중교통과의 연결이 멀고 도시교통을 어떻게 이용하는지 패턴 연구
- 경기도 행복주거 지역주변에 전기차 충전소와 쏘카존과의 연결성 연구

◦ 쏘카 사업의 목표인 전기차 수요를 확보할 수 있는 지역과 쏘카존 사용자 확대 방안

- 경기도 행복주거 지역주변에 전기차 충전소와 쏘카존과의 연결성 연구
- MZ 세대의 소비가치를 결정하는 요건중의 하나인 사회가치를 연결하여, 쏘카 사용자는 사회적 가치 실현의 소비로 탄소발자국 마일리지 제공 연구

*탄소발자국: 자동차 생애에서 배출하는 이산화 탄소 총량

1-2 아이템의 독창성

◦ 다양한 기법들의 활용

- **(앙상블)** 수요예측 문제를 풀에 있어서, 이러한 다양한 기법들을 사용해보고, 성능을 비교 분석하는 것은 필수적이다. 다시 말해, 어떤 모델이 괜찮은 성능을 보이는 지를 확인할 것이다. 더 나아가서는, 서로 다른 여러 모델들을 합쳐서(ensemble) 새로운 모델을 만들어낼 수도 있다.
- **(지역별 수요예측)** 주어진 자료는 지역 정보와 (대여자의) 성별정보, 차종 정보 등이 기재되어 있다. 이 중에서 가장 중요하다고 생각되는 정보는 지역정보이다. 앞서 적었던 대로 특정 지역에 얼마나 차량들을 비치하느냐 하는 문제가 중요할 수 있다. 따라서, 기본적으로는 **시간에 따른 전체 수요를 예측하는 것을 골자로** 하겠지만, 부가적으로 **각 지역별로 시간에 따른 수요를 예측하는** 작업을 수행한다. 더 세부적으로 들어간다면, 차종별 수요를 분석하고 예측해볼 수 있다. 특정 지역에 특정 차종이 부족하거나 많지는 않은지, 예를들어, 부산 남구에 너무 경차만 많고 SUV 가 비치되지 않아

고객 불만을 초래하지는 않는지, 하는 점도 분석해볼 수 있다. 이것을 수요예측과도 관련지어서 생각해볼 여지도 있을 것이다.

2. 개발 및 연구 내용

2-1. 구현 내용 상세

- 45 만 개의 데이터 중 특정 조건과 만족하는 데이터를 추려낸다. I) 비수도권 지역이면서 20, 30 대 사용자 수가 가장 많은 경기도 지역일 것 II) 경기도에 위치한 행복주거지역 중 대중교통 패턴을 분석하여, 사용자를 확보할 수 있는 지역 일 것 III) 경기도에 위치한 쏘카존 위치일 것.
- 추려낸 데이터를 이용해 프로젝트에서 구현하고자 하는 바는, 주어진 데이터에 대하여 수요를 잘 예측하는 모델을 구축하는 것이다. 하지만 더 나아가서는 몇 십 만 개로 구성된 데이터를 입력값으로 넣었을 때, 이를 통해 타겟 고객층 즉 20, 30 대 고객의 미래의 수요 예측치를 예상하여 행복주거지역과 쏘카존의 수요예측을 비교하는 것이다.
- 주어져있는 데이터들은 45 만 여 건의 대여 이력이다. 이 원본 데이터를 그대로 사용하는 것은 적절치 못하다. 각 날짜별로 대여 횟수를 뽑아내고, 그 횟수에 대하여 시계열 분석을 진행하게 된다.
- 정리하면, 풀고자 하는 문제는 다음 두가지이다. **(1) 시간대별 전체 차량대여 수요치를 예측하는 것, (2) 각 지역별 차량대여 수요치를 예측하는 것, (3) 행복주거지역과 쏘카존의 수요 예측을 비교하는 것.** 이 과정에서 성별, 지역, 차종, 대여시간 등이 추가적인 feature 로서 들어갈 수 있다.
- 풀고자 하는 시계열에서 "계절성(seasonality)"은 중요한 고려요소가 될 것으로 보인다. 여기서 말하는 계절성이란, 우리나라의 사계절을 뜻하지는 않는다. 전체 데이터셋의 시간범위가 11 개월 정도 되기 때문에, 사계절로서의 계절성을 고려하는 것은 의미가 없다고 보여진다. 하지만, 아마도, 주중인지 주말인지, 정확하게는 월-목요일인지 금요일인지 토요일인지 일요일인지에 따라서 수요값이 상당히 많이 변할 것이다. 따라서 일주일 단위의 계절성을 세심하게 고려하여 예측에 활용할 필요가 있을 것이다.
- 필요하다면 여러 종류의 아키텍처 혹은 데이터작업을 사용할 수 있다. 머신러닝의 두 방법(Tree based method, clustering)과 딥러닝의 방법(LSTM, transformer), 전통적인 시계열 방법(ARIMA)을 고려할 수 있다. 또한 데이터를 더 잘 분석하기 위한 방편으로 주어진 시계열을 Fourier series 나 wavelet transform 으로 잘라서 분석하는 방법 또한 고려할 만하다고 보인다.
- 데이터셋은 기본적으로 8:2 정도로 train set 과 validation set 으로 나누도록 한다. (test set 은 상정하지 않는다.) 어떻게 나눌 것인가 하는 것은 중요한 문제가 될 듯하다. 하지만, 지금 그냥 생각하기로는, 1 월부터 11 월까지의 데이터들 중, 마지막 2 개월 데이터들을 validation set 으로 두고 나머지는 train set 으로 두어도 괜찮지 않을까 하는 생각이다. 차량 대여 수요는 해당 기간이 성수기인지 비수기인지에 큰 영향을 받을 것이다. 하지만 11 월이라고 하면, 아직 겨울 방학, 겨울 휴가등이 시작되지 않은 시기이기 때문에 일반적인 시기(비수기)라고 생각된다. 따라서 10, 11 월을 validation set 으로 잡아도 괜찮을 것 같다. 하지만 혹시라도, 기간을 특정해 데이터를 나누는 것이 마음에 걸릴 것 같으면 k-fold validation 을 사용하면 될 것 같다(k=11).

- 평가지표로는 MAPE 를 사용하는 것이 가장 적절할 것으로 보인다. MSE 를 사용하는 것도 (시간별로 scale 이 크게 변할 것 같지 않으므로) MAPE 를 사용하는 것과 거의 유사하지만, 직관적으로 받아들이기 쉬운 값으로서는 MAPE 가 더 적절하다고 보인다. 풀고자 하는 문제는 분류 (classification) 문제가 아닌 회귀 (regression) 문제이다. 더 정확하게는 (training set 의 크기를 N , validation set 의 크기를 M 이라고 하면)

$$\begin{aligned} D &= \{(t_i, y_i): t = 1, 2, \dots, M + N\} \\ D_T &= \{(t_i, y_i): t = 1, 2, \dots, M\} \\ D_V &= \{(t_i, y_i): t = M + 1, M + 2, \dots, M + N\} \\ D &= D_T \cup D_V \end{aligned}$$

이다. 여기에서 D_T 는 train set 이고 D_V 는 validation set 이며 D 는 전체 데이터셋이다. 해야 하는 것은 $y_i (i = M + 1, M + 2, \dots, M + N)$ 를 예측하는 것이다. 그 예측값을 $\hat{y}_i (i = M + 1, M + 2, \dots, M + N)$ 이라고 하면, MSE 는

$$MSE = \frac{1}{n} \sum_{i=M+1}^{M+N} (y_i - \hat{y}_i)^2$$

으로 주어진다. 한편 MAPE 는

$$MAPE = \frac{100}{n} \sum_{i=M+1}^{M+N} \left(\frac{y_i - \hat{y}_i}{y_i} \right)^2$$

로 주어질 것이다.

2-2. 개발 아이템 기대효과

이 서비스 아이템이 개발되고 나면, 앞서 1-1 에서 언급한 여러 가지 효과들을 기대해볼 수 있다.

- 먼저, 리스크 관리 차원에서의 기대효과이다. 만약, 갑작스럽게 어떤 지역의 차량대여 수요가 폭발적으로 증가한 경우, 만약 이 갑작스러운 수요 폭발을 예상하지 못했을 경우, 고객들이 차량을 대여할 수 없는 상황이 생길지도 모른다. 조금 더 정확하게 말하면, 많은 경우에 쏘카존에 대여할 차량이 아예 없는 경우는 별로 없을 것이다. 기본적으로 쏘카에서는 쏘카존에 많은 종류의 차량들을 비치해놓기 때문이다. 하지만, 특정한 차종을 원하는 고객들의 경우에는 원하는 차종을 대여하지 못하는 상황이 발생할 수 있다. 특히 서울, 경기권이 아닌 경우에는 쏘카존에 비치된 차량이 한 두 대에 불과한 경우도 존재한다. 이러한 위험성은, 괜찮은 정확도의 모델이 개발되었을 때 예방될 수 있는 것이다.
- 두 번째는, 자원관리 및 비용절감의 차원에서의 효과 또는 수요 증대 효과를 기대할 수 있다. 만약, 어떤 지역의 차량대여 수요가 낮을 것으로 예상되는 경우임에도 불구하고, 쏘카존에 차량들이 지나치게 많이 비치되어 있는 경우라면, 쏘카 입장에서는 불필요한 비용이 나가는 셈이다. 하지만, 반대로 행복주거지역의 수요예측을 이용해 근처 쏘카존에 추가 차량을 배치한다면 이득을 볼 수 있다는 말이다. 만약, 특정 시점에서 특정 지역의 수요를 정확히 예상할 수 있다면, 해당 지역에 위치한

쏘카존의 차량수를 조절함으로써 이득을 볼 수 있을 것이다.

- 마지막으로, 고객들의 차량 대여 동향을 이해하는 데 도움을 줄 수 있다. 미래의 차량 대여 수요를 예측할 수 있다면, 이는 현재의 차량 대여가 어떤 흐름으로 이루어지는 지 정확하게 예상할 수 있다는 말이다. 이러한 이해를 바탕으로 프로모션 전략을 짜거나, 정책을 결정할 수 있을 것이다.

3. 실행 계획

3-1. 기간내 프로젝트 구현 완성을 위한 전략

- 1. 기존의 데이터 분석:
 - 비수도권 경기지역의 쏘카존을 분석
 - 20,30 층의 사용자 시간과 요일을 분석
- 2. 외부 데이터 조사:
 - 경기도 행복주택지역을 지도에서 분석하여 쏘카존과 비교
 - 위의 데이터를 토대로 대중교통 패턴 분석
- 3. 데이터 통합, 전처리, 데이터 시각화
- 4. 수요예측에 적합한 모델과 검증
- 5. 프로젝트 PT 준비.

3-2. 아이펠톤 기간 내 마일스톤

- 정기 회의(목적: 진도관리) : 매주 월,금 (1 시간) , 비정기회의 : 화,수,목(30 분)
- 학습 및 이슈회의 : 수시

| Task | | 목표기간 | 세부내용 |
|--------------------|--|-------------------------|---|
| 팀빌딩 | - 팀빌딩 및 계획서 제출 · 제출 : 2022.12.26 18:00 | 2022.12.26 ~ 2022.12.26 | - 팀장 및 역할, 운영 정립 |
| Mini Aiffelthon | - 운영관리체계 정립 | 2022.12.27 ~ 2022.12.27 | - 개발관리체계 정립 · 프로그램 운영 체계 · Naming Rule 정립 · 운영체 소통관리(on,offline) |
| | - 미니 프로젝트 | 2022.12.27 ~ 2023.01.02 | |
| | · Data 선정 및 EDA | 2022.12.28 ~ 2023.12.30 | - 데이터 확보 및 관리체계 수립 - Data 분석 및 관계도 작성 |
| | · 모델 학습 및 선정 | 2022.12.28 ~ 2023.12.30 | - 2 ~3 모델 학습 · Hyper parameter 검토 |
| | · 프로젝트 계획서 수정 · 제출 : 2023.01.03 18:00 | 2022.12.27 ~ 2023.01.03 | - 프로젝트 계획서 작성 · 진도관리 체계 포함 |

| | | | |
|------------|---|-------------------------|--|
| Aiffelthon | - 카셰어링비즈니스 분석 | 2022.01.02 ~ 2023.01.04 | - 카셰어링 Flow 검토 |
| | - Feature 검토 및 재구성 | 2022.01.02 ~ 2023.01.06 | - Feature 정리표 |
| | - Data 확보 및 전처리 | 2022.01.04 ~ 2023.01.13 | - Data 분석 및 관계도 재검토 . 산출물 : Data set 구성 |
| | - 모델 검토 및 모델링 . Clustering 모델 . 수요예측 모델 | 2022.01.09 ~ 2023.01.20 | - 모델 학습 : 이해도 향상 - 실 Data 로 모델링 . hyper parameter 선정 |
| | - 모델 평가 및 재구성 . Clustering 모델 . 수요예측 모델 | 2022.01.16 ~ 2023.01.31 | - Optimizer , 손실함수 |
| | - 프로젝트 완료 보고서 | 2023.01.23 ~ 2023.02.03 | - 프로젝트 완료 보고서 - 소스코드 |
| | - 개발프로그램 취합 및 정리 | 2023.01.30 ~ 2023.02.03 | |
| | - 발표 자료 최종 검토 . 제출 : 2023.02.07 18:00 | 2023.02.06 ~ 2023.02.07 | |

3-3. 팀장 및 팀원의 역할 분배

| 순번 | 역할 | 역할 상세 | 인원 |
|----|-----|--|----|
| 1 | 차윤겸 | - 대외 업무 - 프로젝트 진행관리 - Data 분석 및 전처리 지도 - 프로젝트 문서 지도 | 1 |
| 2 | 김선중 | - 모델링 주관 및 지도 - Hyper parameter 튜닝 및 확정 - Optimizer 및 손실함수 토의 및 결정 - 프로젝트 문서 작성 지원 | 1 |
| 3 | 이승준 | - Data 전처리 주관 / 변수 통일 / 시각화 - Google Drive 관리 - 프로젝트 문서 작성 지원 | 1 |
| 4 | 배현우 | - GitHub 관리 | 1 |

| | | | |
|---|-----|---|---|
| | | <ul style="list-style-type: none"> - Data 전처리 지원 - 모델링 개발 지원 - 프로젝트 문서 작성 지원 | |
| 5 | 김연수 | <ul style="list-style-type: none"> - Data 전처리 지원 - 모델링 개발 지원 - 프로젝트 문서관리 | 1 |

4. Reference

- [1] Alencar, V. A., Pessamilio, L. R., Rooke, F., Bernardino, H. S., & Borges Vieira, A. (2021). Forecasting the carsharing service demand using uni and multivariable models. *Journal of Internet Services and Applications*, 12(1), 1-20.