서울 시민카드 데이터 분석을 통한 사업 현황 파악 및 어플 개선방안 도출

이다현

이화여자대학교 통계학과
00dahyun00@ewhain.net

요 약

본 연구는 서울 시민카드 어플 이용자 증대를 위해 일별 가입자 현황 데이터와 어플 리뷰 데이터, 연계시설 정보 데이터에 대한 분석을 진행하여 사업 현황을 파악하고 관련 개선 방안 아이디어를 제시한다. 먼저 일별 가입자 현황 데이터로부터 ARIMA, Prophet, LSTM, GRU 시계열 모형 예측을 진행하여 서비스 이용률을 가장 잘 예측할 수 있는 모델로 다변량 LSTM 을 활용해 볼 것을 제안한다. 또한 어플 리뷰 분석을 통해 사용자가 불편함을 호소하는 앱의 기능적 결함 요인과 만족감을 표시하는 어플 내 주요 기능을 살펴본다. 마지막으로 연계 시설 정보 데이터를 시각화 하여 얻은 인사이트를 기반으로 지역별 공공 시설 정보 제공의 편차가 발생하지 않도록 하는 것이 중요하며 도서관 이외의 다양한 문화 시설 정보를 추가하는 것을 제안한다.

Keywords : 서울 시민카드, 시계열 예측, 텍스트 마이닝, Prophet, LSTM, Word Cloud, Topic modeling

1. 서 론

1.1 연구 배경 및 목적

서울 시민카드는 도서관, 미술관, 체육 센터 등의 공공 시설을 하나의 모바일 카드로 이용할 수 있게 하는 취지로 만들어진 서울시 공공 앱이다. [1] 공공 시설 이용의 편리성을 위해 회원 인증, 시설 예약 등의 과정을 모바일 화 했다는 점, 여러 시설의 회원 카드를 하나의 어플로 통합시키어 놓았다는 점, 그리고 모바일 쿠폰과 민간 제휴 할인 혜택을 누릴 수 있다는 점에서 공공 시설 이용 활성화와 서울 시민의 문화생활 향유에 기여한다. 최근에는 제로페이 등의 결제 수단 기능이 추가되어 전자지갑의 역할이 가능해 진 바, 시민들의 문화 생활 전 과정에 있어 이용의 편리성을 제공한다.

이처럼 서울 시민 카드는 서울 시민들의 편리한 문화생활을 지원하는 잠재성이 있는 국가 사업이다. 하지만 서울 시민 인구 (약 948만명)에 비하여 총 가입 회원은 약 35만명에 불과하다는 것을 통해 어플의 존재성을 잘 모르고 있는 사람들이 많다는 것을 알 수 있고 기존 사용자들 조차도 어플의 기능적 결함으로 인해 서비스를 이용하는 데 불편함을 겪는 경우가 다수 발생하고 있는 상황이다.

따라서 본 연구는 서울 시민카드 어플 이용자 증대를 위해 일별 가입자 현황 데이터와 어플 리뷰 데이터, 연계시설 정보 데이터 분석을 바탕으로 서울 시민카드의 사업 현황을 파악하고 관련 개선 방안 아이디어를 제시하고자 한다.

1.2 연구 범위 및 방법

본 연구는 크게 시계열 분석과 텍스트 분석으로 이루어져 있다. 시계열 분석은 '서울 시민카드 어플 일별 가입자 현황' 데이터셋을 바탕으로 진행했다. 선행 연구에서 발전 방향으로 제시한 '계절과 사회 현상을 반영한 분석' 을 위해 Prophet 시계열 모델을 적용하여 추세, 계절성, 이벤트 요인을 반영한 '서비스 이용률' 변수 분석을 진행하였다. [1] 또한 향후 서울 시민카드 기능 및 서비스 개선의 효과성을 입증하기 위한 모델링 가이드라인을 제시하기 위해 '서비스 이용률' 변수와 관

련하여 Prophet, ARIMA, LSTM, GRU 예측 모델링을 진행해 최적의 모델을 도출하였다.

텍스트 분석은 iOS 환경과 안드로이드 환경에서의 어플 내 기능적 결함 요인을 도출하기 위해, 앱 스토어 리뷰와 구글 플레이 스토어 리뷰를 각각 크롤링 하여 분석을 진행하였다. 별점을 기준으로 리뷰 텍스트를 긍정 (4~5점) 과 부정(1~3점) 으로 라벨링 하여 워드 클라우드 시각화를 진행하였다. 이를 통해 어플 기능 내 만족, 불만족 요인들을 도출하였으며 토픽 모델링을 통해 사용자가 구체적으로 어떠한 상황에서 만족 혹은 불만족을 표출하는지 살펴보았다.

더불어 서울시 연계 공공 및 문화 시설에 대한 정보가 다양하게 제공되고 있는지, 지역별로 정보의 편차가 존재하는지 등을 파악하기 위해 '연계시설 정보 데이터' 를 바탕으로 시각화를 진행하였다.

2. 선행 연구

2.1 시계열 이상치 탐지 분석

서울 시민카드 측에서 제공하는 데이터 중, '서울시민카드 일별 성별_연령별 통계정보' 데이터를 활용하여 분석을 진행한 선행 연구 '서울 시민카드 사용자 유입과 이탈에 대한 시계열 이상치 탐지와 DTW 클러스터링' 는 , ARIMA 이상치 탐지 모형과 DTW 클러스터링 기법을 적용하여 다음과 같은 연구 성과를 제시한다. [1]

먼저 ARIMA 시계열 이상치 탐지 모형을 적용해 어플의 '단기 및 중장기 이벤트'와 개인 인증, 제로페이 결제, 관광 할인패스 등의 새로운 서비스 도입이 사용자 이탈과 유입에 영향을 미친다는 관계성을 분석했다. 이후 DTW 클러스터링 기법을 적용하여 비슷한 시계열 양상을 보이는 사용자 집단의 군집화를 통해, 이벤트와 업데이트에 반응하는 양상을 분석했다. 3~40 대 여성은 꾸준한 접속을 보이는 군집, 2~50대 남성은 이탈 가능성을 보이는 군집, 20대 여성과 40대 남성 그리고 50대 여성은 이벤트와 업데이트에 반응하는 군집으로 분류하여 사용자의 이용 패턴과 관련된 군집 정보를 제시했다.

2.2 어플 리뷰 분석

어플 리뷰 분석에 대한 연구는 리뷰 데이터로부터 서비스에 관한 사용자의 만족, 불만족 요인을 도출할 수 있고 [8] 해당 어플만의 고유한 기능들을 파악할 수 있다고 제안한다.[9] 또한 최근 어플 리뷰 데이터 분석에 관한 연구는 공통적으로 토픽 모델링, 감성분석, 소셜 네트워크 기법을 적용하고 있다. [10]

3. 분석 방법

3.1 데이터 수집

'서울 시민카드 일별 가입자 현황' 데이터와 '연계 시설 정보' 데이터는 서울 열린 데이터 광장을 통해 2022년 5월 31일 까지의 데이터를 수집하였다. 분석에 사용된 날짜 범위는 2018년 1월 1일부터 2022년 5월 31일까지 해당한다.

리뷰 데이터의 경우 크롤링을 통해 iOS 환경을 제공하는 앱 스토어 리뷰와 안드로이드 환경을 제공하는 구글 플레이 스토어 리뷰를 수집하였다. 앱 스토어는 50 개의 사용자 리뷰, 플레이 스토어는 119 개의 사용자 리뷰와 103개의 관리자 답변 데이터를 수집하였다.

3.2 시계열 분석

일별 가입자 현황 데이터는 일별로 '총 가입회원 수, 회원 가입자 수, 회원 탈퇴자 수, 앱 설치자 수, 접속자 수' 각각을 집계한 값으로 이루어져 있다. EDA 를 통해 각 변수 별 추이 및 변동률을 살펴봄으로써 추가 분석이 필요한 지점들을 살펴보고 변수 별 관계성을 도출하기 위해 상관관계, 공적분 분석을 진행하여 사업 개선에 대한 인사이트를 도출했다.

'접속자 수' 칼럼으로부터 회원 가입 및 탈퇴를 위한 접속을 제외하고 앱에서 제공하는 서비스를 이용할 목적으로 접속한 접속자 수를 고려하기 위해 '서비스이용접속' 변수를 새로 생성하였으며 증가하는 총 가입 회원 수 경향성을 보정하기 위해 총 가입 회원 수 대비 앱 서비스 이용을 목적으로 접속한 접속자 수 비율을 계산해 '서비스 이용률' 칼럼을 생성하였다.

위와 같이 생성한 '서비스 이용률' 칼럼을 기준으로 추세에 영향을 미치는 시기적 요인들까지 고려할 수 있는 시계열 모형 Prophet, 전통적인 시계열 모형인 ARIMA, 딥러닝 시계열 모형인 LSTM과 GRU 로 예측 모델링을 수행하였다. 2018년부터 2021년까지의 데이터를 훈련 데이터로, 2022년 1월부터 5월까지 데이터 셋을 테스트 데이터로 활용했다. 예측 모델의 평가지표로는 RMSE 를 사용했다.

3.3 텍스트 분석

앱 스토어 리뷰와 구글 플레이 스토어 리뷰 데이터 각각에 대해 워드 클라우드 분석을 진행하여 iOS 환경과 안드로이드 환경에서 발생하는 사용자 만족, 불만족 요인을 도출하였다. 별점을 기준으로 리뷰 텍스트를 긍정 (4~5점) 과 부정(1~3점) 으로 라벨링 하여 긍정 리뷰로부터 만족 요인을, 부정 리뷰로부터 불만족 요인을 도출했다. 텍스트 전처리의 경우 네이버 맞춤법 검사기를 이용한 파이썬 한글 맞춤법 검사 라이브러리 Hanspell 을 사용하여 리뷰 글의 맞춤법과 띄어쓰기를 보정하고 한국어 텍스트 전처리 라이브러리인 konlpy 의 okt 태깅 기법으로 토큰화를 진행하였다. 이후 빈도분석을 기반으로 워드 클라우드 시각화를 진행했다.

서울 시민카드 앱은 공공 기관에서 운영하는 어플이기 때문에 관리자 혹은 개발자의 사용자 불만족 요인에 대한 대처방식이 향후 공공 사업의 지속 가능성을 좌우할 것이라는 판단하에, 관리자 답변까지 워드 클라우드 빈도 분석을 진행하여 사업 담당자의 대처 방식을 파악해 보았다.

이후 토픽 모델링 분석 진행했다. Coherence 값을 기준으로 최적의 토픽 개수 4를 설정하여 LDA 모델링을 진행하였다. 리뷰 데이터의 개수가 현저히 적었기 때문에 유의미한 토픽을 도출하기 어려웠으나 LDA 시각화를 통해 파악한 가장 관련성이 높은 키워드들을 중심으로 사용자가 구체적으로 어떠한 상황에서 만족 및 불만족을 표출하는지 정리했다.

4. 분석 결과

4.1 일별 가입자 데이터 분석

a. Domain knowledge

분석 결과를 타당성 있게 해석하기 위해 서울 시민카드 어플에 관한 기능적 특징과 이벤트 및 서비스 도입 시기를 정리하였다. 서울 시민카드 정책 소개 글을 통해 발견하거나, 직접 사용해보고 발견한 기능적 특징은 다음과 같다.

1. 공공시설 통합 모바일 회원 카드 사용
2. 시설 정보 , 문화행사 검색 : 시설 이용에 관한 기본 정보, 시설에서 제공하는 교육/강좌 목록, 문화 행사에 대한 기본 정보 제공, 서울 문화 포털과 연동
3. 민간 제휴 혜택 및 쿠폰 제공 : GS25 할인쿠폰, CGV 할인쿠폰, 백화점 무료 주차권, 세종 문화회관 및 대학로 티켓 예매
4. 공공 서비스 예약 : 튕김 현상이 자주 존재함
5. 내 위치를 기준으로 이용 가능한 주변 공공 시설 정보 제공 : 지도로 위치 확인이 가능하며, 클릭하면 정보 제공 화면으로 넘어감
6. 회원 가입 과정 : 휴대폰 인증, 약관동의, 이메일 입력 순 , 회원 탈퇴 과정 : 휴대폰 번호 입력 순

표 1 . 서울시민카드 서비스에 관한 사전 지식

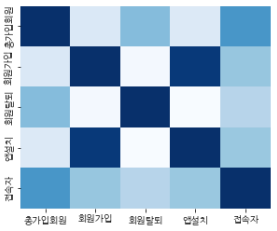
시기별 신규 이벤트 및 서비스 도입 시기를 정리한 표는 다음과 같다.

2018	◦ 12월 : 풋볼 입장권 할인 이벤트
2019	◦ 전체 기간 : 연극, 영화, 공연, 전시 할인 이벤트 ◦ 4월~12월 : 현대 아울렛 동대문점 할인 이벤트 ◦ 5월~6월 : 플라스틱 사용 줄이기 이벤트, 공정무역 관련 서비스 혜택, 동대문 기획전 전시할인, 심청각 카페 음료 할인
2020	◦ 전체 기간 : 민간문화 제휴혜택 및 할인 이벤트 ◦ 내 주변 제휴시설 지도 기능 도입 ◦ 9월 : 서울패스 도입 ◦ 제로페이 결제 기능 추가
2021	◦ 7월 : 신규 앱 가입 이벤트 ◦ 10월 : 서울 관광 할인패스 도입

표 2 . 서울시민카드 신규 이벤트 및 서비스 도입 시기

b. 변수 별 관계성 분석

시계열 데이터 사이에 동시성을 측정하기 위해 DTW, TLCC 등 다양한 방법을 사용하여 분석을 진행한다. [5] 본 연구에서는 상관관계 분석과 공적분 분석을 진행하여 특징적인 변수 별 관계성을 도출하였다. 먼저 피어슨 상관관계 분석 결과는 다음과 같다.



	총가입회원	회원가입	회원탈퇴	앱설치	접속자
총가입회원	1.000000	0.328011	0.555534	0.323400	0.692189
회원가입	0.328011	1.000000	0.231327	0.969848	0.521967
회원탈퇴	0.555534	0.231327	1.000000	0.214265	0.450224
앱설치	0.323400	0.969848	0.214265	1.000000	0.515914
접속자	0.692189	0.521967	0.450224	0.515914	1.000000

그림 1. (a) 히트맵 시각화 (b) 상관분석 결과 값

앱 설치자 수와 회원 가입자 수의 상관관계가 0.96 으로 매우 높은 것을 확인해볼 수 있었다. 앱 설치자 수와 회원 가입자 수를 원 데이터에서 확인해본 바, 값이 정확히 일치하진 않았다. 따라서 앱 설치 이후 회원가입을 바로 하는 경우가 많다는 의미로 해석할 수 있으며, 앱 설치 과정과 회원가입 과정이 유기적으로 연결될 수 있는 원인으로 다음과 같은 사실을 도출할 수 있었다. 먼저 회원가입을 반드시 해야만 기타 어플 서비스를 이용할 수 있도록 했다는 점과 회원 가입 절차도 휴대폰 인증 → 약관 동의 → 이메일 입력 순으로 직관적이고 간단하며, 추가로 기입할 여타 정보가 없다는 점을 원인으로 볼 수 있었다.

공적분은 두 시계열 데이터에서 어떠한 관계가 있다는 것을 통계적 지표로 보여주기 위해 사용하는 분석 기법으로, 공적분 관계에 있다는 것은 장기적으로 서로 일정한 관계에 있다는 것을 의미한다. '시계열 데이터는 서로 공적분 관계가 있다' 라는 귀무 가설을 기반으로 가설검증을 진행한다. 접속자 수 변수를 기준으로 회원 가입과 회원 탈퇴 각각에 대해 공적분 관계를 살펴봤을 때, 회원 가입자 수와는 공적분 관계에 없었고 회원 탈퇴자 수와 공적분 관계에 있음을 파악했다. 즉, 장기적인 관점에서 회원 가입의 목적보다 탈퇴의 목적으로 접속한 사람이 많다는 뜻으로 해석해볼 수 있다. 서울 시민카드 어플의 회원 탈퇴 절차는 휴대폰 번호 입력 한 차례로만 진행되어 탈퇴가 쉽기 때문에, 사용자가 어플에 접속했을 시 사용에 불편함을 겪어 회원 탈퇴로 이어지지 않도록 주기적인 앱 기능 개선이 필요할 것으로 보인다.

c. 서비스 이용률 칼럼 생성

사용자가 어플을 접속하는 다양한 이유가 존재할 것이다. 주어진 데이터 셋으로부터 개별 사용자의 로그 기록을 확인할 수 없기 때문에 구체적으로 어떤 서비스를 주로 이용했는지에 대한 정보를 파악할 수는 없었으나, 회원 가입이나 탈퇴 요인 외에 가입자가 '서비스 이용' 을 목적으로 어플을 접속한 빈도수는 파악할 수 있음을 확인했다. 따라서 아래와 같이 서비스 이용 접속에 관한 피처를 생성하였다.

```
data['서비스이용접속'] = data['접속자'] - (data['회원가입'] + data['회원탈퇴'])
```

총 회원가입자 수가 증가 추세에 있음을 고려하여 최종적으로 총 가입 회원 수 대비 앱 서비스 이용을 목적으로 접속한 접속자 수 비율을 나타내는 '서비스 이용률' 피처를 생성했다.

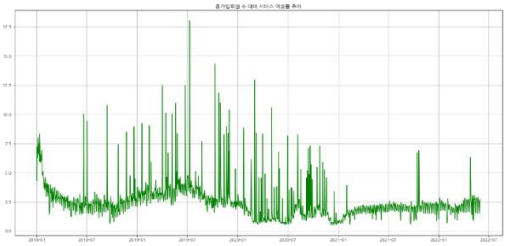


그림 2. 총 가입회원 수 대비 서비스 이용률 추이

총 가입 회원 수 대비 서비스 이용을 목적으로 접속한 사용자 수인 서비스 이용률을 시각화 한 결과이다. 평균값과 중앙값은 2%, 최대값은 18%로 나타났다. 전반적으로 2%~5% 사이의 값에 분포하고 있으나, 특징점들이 종종 관측되므로 기준점을 중심으로 이용률이 높은 시기의 공통점을 찾아볼 필요성이 있다. 더불어 2021년 이후로 서비스 이용률이 이전 시기보다 눈에 띄게 감소하므로 이에 대한 요인도 파악해볼 필요가 있다.

d. 서비스 이용률 예측 모델링

선행 연구 논문에서 언급한 바와 같이 서울 시민카드 어플 내에서 진행하는 할인 이벤트나, 제로페이 등 서울시만 제공할 수 있는 특징적인 사업과의 연동 기능을 통해 이용자 수 유입을 지속적으로 이끌어낼 수 있다. 그러나 '지속성이 얼마만큼 유지 되는지' 에 관하여 서울 시민카드의 기능 및 서비스 개선의 효과성을 보다 객관적으로 입증하기 위한 지표가 필요하다. 따라서 위에서 생성한 '서비스 이용률' 변수를 기준으로 하여 최적 예측 모델을 도출해보고자 한다. 본 모델링 과정을 통해, 가령 특정 기능을 도입했을 때 발생하는 서비스 이용률의 시계열 패턴을 예측 모델링으로부터 학습하게 된다

면 추후 이벤트 및 서비스 도입의 적절성을 평가하기 위한 항목으로 활용할 수 있을 것이다.

Prophet, ARIMA, LSTM, GRU 모델을 통해 예측을 진행하였으며 2018년~2021년 데이터를 훈련용으로 2022년 01월~05월 데이터를 테스트용으로 설정하여 분석을 진행하였다. 모델 성능 평가 지표로는 RMSE 를 사용했다.

① Prophet

Prophet 은 페이스북에서 만든 시계열 분석 라이브러리로 ARIMA 와 같은 전통적인 시계열 모형과 달리 curve fitting 방법을 도입하여 시간에 종속적인 구조를 갖지 않도록 하는 것이 특징인 모형이다. 또한 Growth (추세) , Seasonality (연간/주간 계절 성분), Holidays (휴일 등의 이벤트) 를 고려한 분석이 가능하기 때문에 비즈니스 분야에 많이 적용되는 시계열 모형이기도 하다. [2]

본 연구에서는 Prophet 을 활용하여 '서비스 이용률' 변수를 기준으로 Growth, Seasonality, Holiday 패턴을 시각화 하였고 예측 모델링과 더불어 이상치 검출을 통해 주요하게 살펴볼 이상치 기간을 선정하였다. Prophet은 다양한 파라미터를 통해 모델의 성능을 높일 수 있는데, 여러 실험과정을 통해 추세 경향을 얼마만큼 유연하게 반영할지 결정하는 change_prior_scale 파라미터가 0.3 일 때 RMSE 값이 가장 낮음을 확인하였다. 휴일과 계절 유연성에 관한 파라미터 설정과 add_country_holidays 를 'KR'로 설정하여 한국 기준 휴일을 고려하여 분석을 진행할 수 있도록 하였다. [3]

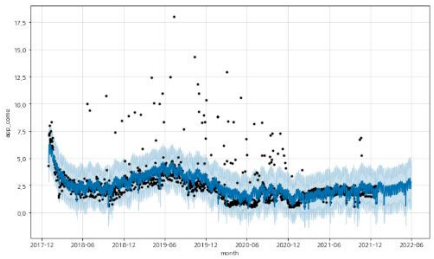


그림 3. Prophet fitting 결과

예측 결과를 시각화 한 그래프를 통해 원래 데이터와 예측한 값 (파란색 선), 예측 상한 및 하한 값 (하늘색 테두리) 을 확인해볼 수 있다. Growth, Seasonality, Holiday 패턴을 시각화한 결과는 다음과 같다.

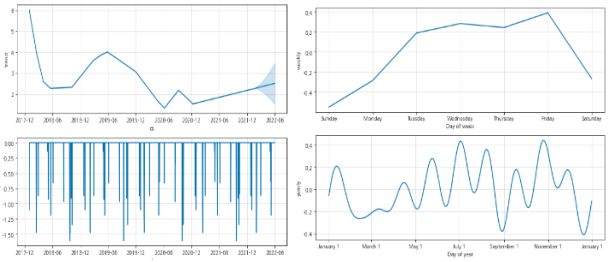


그림 4. Prophet 패턴 시각화 결과 (왼쪽 위 – 오른쪽 순)

순서대로 시각화 결과를 살펴보면 먼저 trend 그래프를 통해 연도별 서비스 이용률 추이를 파악할 수 있다. 2020년도 하반기에 서울패스와 제로페이 기능이 도입되어 이용률이 소폭 증가했음을 유추해볼 수 있고 2021년도 10월 서울 관광 할인 패스 기능의 도입으로 상승 추세를 이어 나가고 있음을 유추해볼 수 있다. 다음으로 Weekly 그래프를 통해 주말 이용보단 수요일부터 금요일 사이의 이용률이 높음을 알 수 있다. 주말에 휴관이거나 단축 운영을 하는 공공 기관의 특성 때문인 것으로 파악해볼 수 있으며 정확한 요인을 파악하기 위해선 연계시설 정보 데이터셋에서 이용 시간에 관한 변수를 생성해 추가로 분석할 필요가 있을 것으로 보인다. Holidays 그래프의 경우 서울 시민카드에서 제공하는 연계 시설이 대부분 공공 시설이기 때문에 공휴일에 휴관하는 곳이 많아 대부분 음수 값을 띄고 있음을 파악할 수 있다.

모델을 통해 test data를 예측한 결과는 다음과 같다.

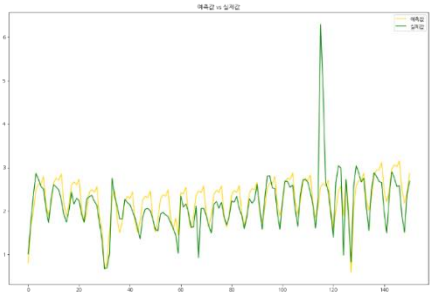


그림 5. Prophet 예측 결과

RMSE 값은 0.47 이며 시각화 한 결과를 보면 특정 구간을 제외하고 대부분의 구간에서 잘 예측한 것으로 보인다.

이상치 검출은 모델링을 통해 얻은 예측 상한 값과 하한 값을 기준으로, 관측치가 상한 값보다 큰 경우와 하한 값보다 작은 경우를 이상치로 간주하였다. 관측치가 너무 큰 경우는 73건, 너무 작은 경우는 4건이 존재하였다. 이상치들 중에 심

한 정도를 계산하여 이상치 중요도 점수를 산출하였고 0.5를 기준으로 0.5보다 큰 값을 도출한 결과는 다음과 같다.

	ds	trend	yhat	yhat_lower	yhat_upper	y	anomaly	importance	score
1029	2020-10-26	1.848867	2.004839	0.366399	3.640121	7.281745	1	0.500103	0.500103
994	2020-09-21	2.092391	1.945828	0.352311	3.619864	7.289507	1	0.503414	0.503414
912	2020-07-01	1.567501	2.278394	0.621950	3.789396	8.146814	1	0.534862	0.534862
854	2020-05-04	1.595166	1.133246	-0.487876	2.715261	10.545291	1	0.742514	0.742514
821	2020-04-01	1.888948	2.041156	0.310713	3.658046	8.316931	1	0.560169	0.560169
798	2020-03-09	2.112877	1.645729	-0.001221	3.155513	8.393993	1	0.624075	0.624075
792	2020-03-03	2.171293	2.155390	0.498656	3.831706	12.918799	1	0.703401	0.703401
700	2019-12-02	3.067009	2.924286	1.264946	4.521449	10.363177	1	0.563700	0.563700
662	2019-10-25	3.260377	4.091829	2.513817	5.745713	11.812519	1	0.513591	0.513591
648	2019-10-11	3.331618	3.676321	1.994362	5.296016	14.312448	1	0.629971	0.629971
556	2019-07-11	3.799767	4.298765	2.754602	5.935975	18.023230	1	0.670649	0.670649
540	2019-06-25	3.881170	4.302596	2.673386	5.902904	12.480908	1	0.527045	0.527045
457	2019-04-03	3.801544	3.974160	2.271144	5.566248	12.442082	1	0.552627	0.552627
256	2018-09-14	2.323802	2.844399	1.343031	4.380964	10.765044	1	0.593038	0.593038
183	2018-07-03	2.294177	2.917021	1.382840	4.542281	9.405412	1	0.517057	0.517057
171	2018-06-21	2.289676	2.597851	0.970443	4.225914	9.997322	1	0.577295	0.577295

그림 6. Prophet 이상치 결과

결과를 살펴봤을 때 특히 2020년 5월 4일은 0.74로 이상치 중요도 값이 가장 크다. 가령 어린이날 전날 이기 때문에 관련 공공기관 행사를 이용하기 위한 접속이 평상시보다 많았을 것이라 해석해볼 수 있다. 나머지 요일들에 대해서도 어떠한 요인에 의해 이상치로 간주되었는지 추가적인 분석이 필요할 것으로 보인다.

이처럼 Prophet 은 모델 피팅을 통해 예측 수행 뿐 아니라 이상치 검출과 추세, 계절성, 휴일 이벤트 등을 고려한 구체적인 분석이 가능하므로 일별 회원 탈퇴자 수와 회원 가입자 수 변수에 대해 적용하면 더욱 풍부한 인사이트를 도출할 수 있을 것으로 보인다. 추후 연구에서는 공공 시설의 운영 시간 및 휴일을 고려하여 이를 파라미터 설정에 반영해 보다 발전된 분석을 진행해 볼 것을 제안한다.

② ARIMA

ARIMA는 전통적인 시계열 예측 모형이다. AR 모델은 자기 자신의 과거 정보를 사용하고, MA 모델은 이전 항에서의 오차를 이용해 현재 항의 상태를 추론하는데 이 둘을 합친 것이 ARMA 모델이고 ARMA 모델에 추세 변동의 경향성까지 반영한 분석 모형이 ARIMA 에 해당한다. ARIMA 는 order 파라미터 설정에 따라 예측 성능이 좌우되는데AR 모형의 Lag (시차) p 와 차분 횟수 d, MA 모형의 Lag (시차) q 값을 차례로 설정해주면 된다. [13]

0부터 4까지 해당하는 정수 값으로 (p,d,q) 쌍을 고려하여 arima 모델을 적용하였으며 aic 값이 가장 낮은 (2,1,2) 를 최종적인 order 값으로 설정하였다.

ARIMA Model Results

Dep. Variable:	D.y	No. Observations:	1460
Model:	ARIMA(2, 1, 2)	Log Likelihood	-2526.565
Method:	css-mle	S.D. of innovations	1.365
Date:	Wed, 01 Jun 2022	AIC	5063.129
Time:	05:45:45	BIC	5089.560
Sample:	1	HQIC	5072.989

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1.D.y	0.6099	0.169	3.613	0.000	0.279	0.941
ar.L2.D.y	-0.1297	0.032	-4.083	0.000	-0.192	-0.067
ma.L1.D.y	-1.4145	0.169	-8.395	0.000	-1.745	-1.084
ma.L2.D.y	0.4585	0.157	2.922	0.004	0.151	0.766

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	2.3519	-1.4768j	2.7771	-0.0892
AR.2	2.3519	+1.4768j	2.7771	0.0892
MA.1	1.0970	+0.0000j	1.0970	0.0000
MA.2	1.9883	+0.0000j	1.9883	0.0000

그림 7. ARIMA fitting 결과

위는 ARIMA fitting 의 결과로 도출된 모델의 식은 아래와 같다.

$y't = 0.6099\ y't\text{-}1 - 0.1297\ y't\text{-}2 + et - 1.4145 et\text{-}1 + 0.4585 et\text{-}2$

상수항을 제외한 모든 계수의 p-value 가 0.05 이하이기 때문에 모형의 각 항이 유의하다고 볼 수 있다.

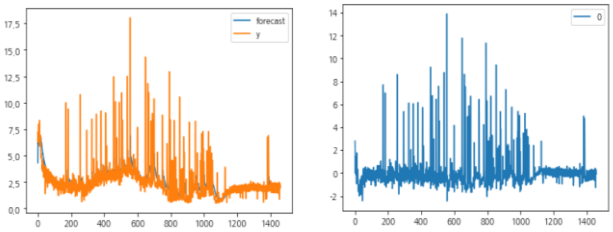


그림 8. (a) ARIMA fitting 결과 (b) 잔차 변동 시각화

위의 그래프는 왼쪽은 학습 데이터에 대한 arima 예측 결과 오른쪽은 잔차의 변동을 시각화 한 값이다. 오차 변동이 매우 불안정함을 살펴볼 수 있다.

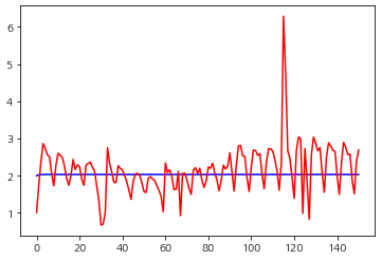


그림 9. ARIMA 예측 결과

테스트 셋에 대한 예측 결과를 살펴보면 다음과 같다. 파란색 실선이 예측 값인데 올바르게 예측해내지 못하고 있음을 살펴볼 수 있다. RMSE 값도 약 0.645 가장 성능이 좋지 않았다. ARIMA 는 시계열 데이터의 정상성을 가정하고 있기 때문에

평균과 분산이 시간에 따라 일정하지 않은 데이터의 경우, 즉 시간의 흐름에 따라 특성이 변하는 데이터인 경우 잘 예측해 내지 못한다. 따라서 복잡한 패턴을 모델링 하는 것에 적합하지 않은 모형이기 때문에 결과가 일률적으로 등장한 것이다. 서울 시민카드 일별 접속자 수에 관한 모형 예측 수행 시 ARIMA 모델의 적용은 최대한 지양하는 것을 제안한다.

③ LSTM

다음은 딥러닝 모형인 LSTM 으로 다변량 분석과 단변량 분석을 각각 진행해보았다. 다변량 분석의 경우 설명변수로 '회원 가입자 수, 회원 탈퇴자 수, 앱설치자 수' 칼럼을 사용하여 분석을 진행하였고 아래와 같은 예측 결과를 얻었다. RMSE 값은 0.125 로 앞선 Prophet, ARIMA 모형보다 성능이 많이 향상되었음을 확인해볼 수 있다.

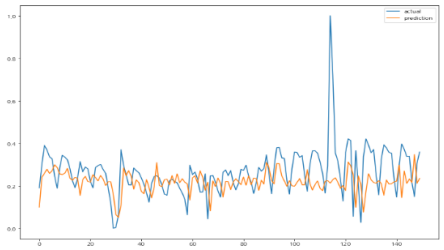


그림 10. 다변량 LSTM 예측 결과

단변량 분석은 서비스 이용률 변수 만을 활용하여 얻은 패턴을 기반으로 예측하는 방법으로 시차가 1 차이나는 시계열 데이터를 입력 값으로 넣어 그 다음 timestep 의 값을 예측하도록 훈련을 진행하였다.

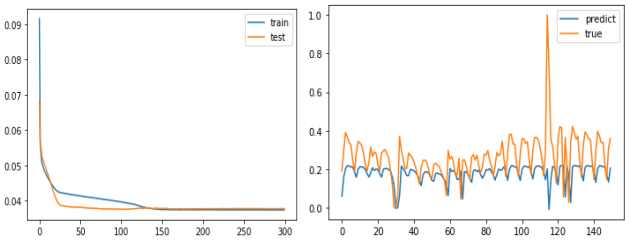


그림 11. 단변량 LSTM 예측 결과

Loss 관점에서 오버피팅이나 언더피팅이 발생했는지 살펴보기 위해 시각화 해본 결과, 두 loss 값이 벌어지지 않고 낮은 값으로 수렴함을 통해 훈련과 예측이 잘 이루어 졌음을 확인해볼 수 있다. RMSE 값은 0.141로 다변량 예측 모델보다는 성능이 낮음을 알 수 있었다.

④ GRU

GRU 역시 시계열 딥러닝 예측 모형에 해당하지만 LSTM 의 과적합 문제를 보완하여 등장한 모형이며 보다 빠른 속도로 훈련을 수행한다는 장점이 있다. 모델링 방식은 LSTM 과 동일하며 먼저 다변량 예측 결과는 다음과 같다.

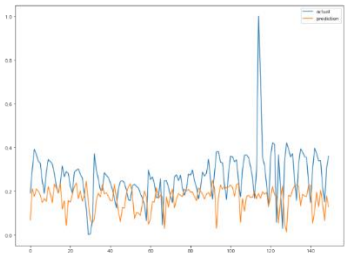


그림 12. 다변량 GRU 예측 결과

RMSE 값은 0.158로 앞선 LSTM 모델보다 성능이 낮음을 확인할 수 있었다. 단변량 분석의 결과는 아래와 같다. 왼쪽은 train, test loss 그래프이고 오른쪽은 예측 값과 실제 값을 시각화한 결과이다. RMSE 값은 0.147로 LSTM 모델 성능과 유사한 결과를 보였다.

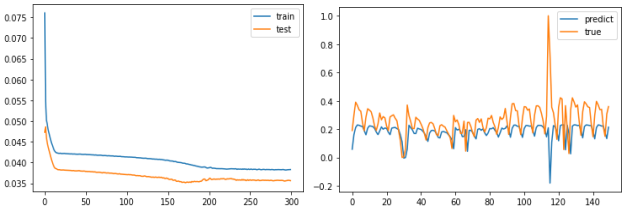


그림 13. 단변량 GRU 예측 결과

여러 모형을 적용시켜본 결과 성능 비교표는 아래와 같다. 결론적으로 LSTM 을 모형을 '회원 가입자 수, 회원 탈퇴자 수, 앱설치자 수' 를 설명변수로 적합 시켰을 때 '서비스 이용률' 예측 모델 성능이 가장 좋았다. 따라서 추후 이벤트 및 서비스 도입의 적절성을 위한 평가항목으로 활용할 모델 중 다변량 LSTM 모형을 적용해볼 것을 제안한다.

Prophet	0.4733	
ARIMA	0.6445	
LSTM	**다변량 : 0.125**	단변량 : 0.141
GRU	다변량 : 0.158	단변량 : 0.147

표3. 모델 성능 비교

4.2 앱 리뷰 분석

워드 클라우드와 토픽 모델링을 활용한 어플 리뷰 분석을 통해 이벤트나 새로운 서비스 기능 도입 요인 외에, 사용자의 만족 혹은 불만족을 일으키는 요인들을 추가로 파악할 수 있었다.

a. 워드 클라우드

앱 스토어 리뷰와 구글 플레이 스토어 리뷰 데이터를 수집하여, 각각에 대해 워드 클라우드 분석을 진행했다. 별점을 기준으로 4~5점은 긍정, 1~3점은 부정 감정으로 라벨링하여 긍정 리뷰로 부터 만족 요인을 부정 리뷰로 부터 불만족 요인을 도출하였다. 앱 스토어 리뷰는 50개, 구글 플레이 스토어 리뷰는 119 개로 리뷰 데이터가 적어 분석 결과를 일반화 하기는 어려우나, 추후 서울 시민카드 어플 이용자 수가 증가하여 더 많은 사용자가 리뷰를 남기게 된다면 아래의 분석 방향을 더욱 발전 시킬 수 있을 것이다.

① iOS 모바일 환경에서 발생하는 사용자 만족, 불만족 요인

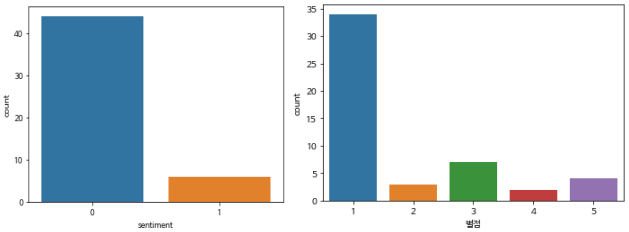


그림 14. (a) 긍부정 감정 분포 (b) 별점 분포

iOS 환경에서 서울 시민카드 어플을 이용한 사용자는 대부분 부정적인 리뷰를 남긴 것으로 파악하였으며 별점 분포 또한 1점에 가장 많이 분포하고 있음을 알 수 있었다.

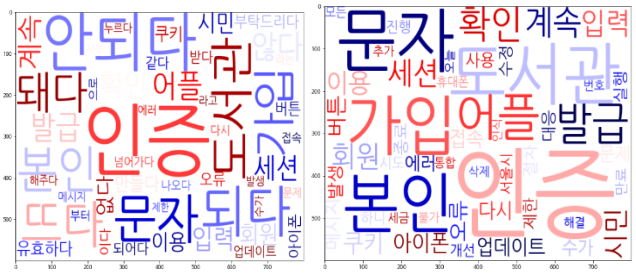


그림 15. 워드클라우드 (a) 형태소 기반 (b) 명사 기반

왼쪽은 형태소 기반 워드 클라우드 시각화 결과이고 오른쪽은 명사 기반 워드 클라우드 시각화 결과이다. 도출된 불만족 요인으로는 '인증' 과정 오류로 인한 것이 가장 많았으며 이외에도 쿠키 제한으로 인한 문자 인증 불가능, 세션 오류, 버튼 활성화 오류 등이 존재하였다. 긍정 요인으로는 도서관 및 체육관 카드 통합 지원 서비스와 할인 쿠폰 제공 등이 존재했다. 따라서 iOS 앱에 관한 업데이트 시 '인증'에 관한 기능적 결함을 개선할 필요가 있으며 이는 서울 시민 카드 통합 지원 서비스에서 가장 중요하고 기초적인 단계에 해당하기 때문에 사용자가 불편함을 겪지 않도록 즉각 대응하는 것이 필요할 것으로 보인다.

② 안드로이드 모바일 환경에서 발생하는 사용자 만족, 불만족 요인

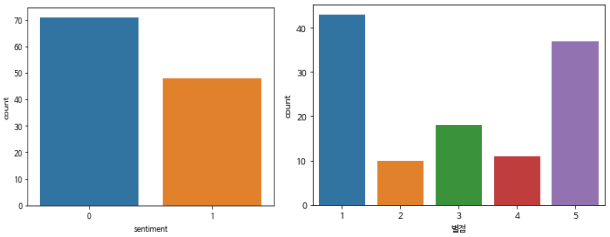


그림 16. (a) 긍부정 감정 분포 (b) 별점 분포

안드로이드 환경에서 서울 시민카드 어플을 이용한 사용자는 iOS 환경과 달리 긍정적인 리뷰를 다수 남겼음을 확인해 볼 수 있으며 별점 분포를 살펴보았을 때에 1점과 5점의 빈도 차이가 크지 않음을 확인해볼 수 있었다. 따라서 플레이스토어 리뷰의 경우엔 긍정, 부정 리뷰 별 워드 클라우드를 진행해보았다.

그림 17. 워드클라우드 (a)긍정리뷰 (b)부정리뷰 (c)명사 기반

안드로이드 리뷰의 경우 긍정과 부정 리뷰의 분포 차이가 크지 않기 때문에 라벨별로 시각화를 진행했다. 위의 시각화 그림 차례로 긍정 리뷰, 부정리뷰, 명사 워드 클라우드 시각화를 진행한 결과이다. 긍정 리뷰 워드 클라우드로부터 살펴볼 수

있는 사용자 만족 요인은 커피나 영화 할인 쿠폰 같은 혜택, 시설 카드 지갑 등록의 편리함, 여러 도서관 이용이 하나의 어플로 가능하다는 점 등이 존재했다. 부정 리뷰 워드 클라우드로부터 살펴볼 수 있는 사용자 불만족 요인에는 본인 인증 과정 오류, 통합 시설 연동 지연, 시설 예약 오류 등이 존재했다. iOS 환경과 마찬가지로 통합 카드 이용에서 가장 중요한 회원 인증과 관련된 불만족 요인이 가장 많았다. 또한 공공시설 예약과 관련된 불편사항을 확인해볼 수 있었는데, 실제로 어플을 통해 시설 예약을 진행했을 때 지속적으로 튕김 현상이 발생함을 확인해볼 수 있었다.

③ 관리자 응답

서울 시민 카드는 공공 기관에서 운영하는 어플이기 때문에 민간 기업에서 수익성을 목적으로 어플 기능 개발 및 업데이트가 이루어지는 양상과 다를 것이라 가정하여, 관리자 답변 글을 크롤링 한 후 어떠한 방식으로 사용자 불만족 요인에 대처하고 있는지 살펴보았다.

그림 18. 관리자 답변 워드 클라우드

워드 클라우드 시각화를 통해 '확인 후 빠른 조치를 취하겠다', '문의 게시판을 통해 문의해 달라' 는 대응이 가장 많았음을 확인해볼 수 있었고 회원 카드 연동이 되지 않는다는 불만족 리뷰에는 '연동이 되지 않는 시설에 대해 문의를 주면 확인 후 조치를 취하겠다' 는 대응을 보였다. 사용자 각자의 모바일 및 네트워크 환경에 따라 오류가 발생하는 경우에 대비해, 문의 게시판 혹은 질의 응답 게시판으로 유도하여 사용자 불만족 요인에 대처하고 있음을 확인할 수 있었다.

그러나 사용자 만족도를 높이기 위해선 보다 본질적이고 주기적인 대응 방식이 필요하다. 공공 사업의 지속 가능성 위해, 일정 기간 간격으로 사용자 불만족 요인을 도출한 후, 그에 따른 어플 기능 개선 및 업데이트를 실행하고 다시 사용자 평가를 기준으로 목표하고자 했던 부분이 개선이 되었는지 살펴보는 과정이 필요할 것으로 보인다.

b. 토픽 모델링

워드 클라우드 시각화를 통해 iOS 와 안드로이드 모바일 환경 각각에서 발생하는 사용자 만족 및 불만족 요인을 살펴보았다면, 토픽 모델링을 통해 두 리뷰를 종합적으로 분석하여 구체적으로 사용자가 어떤 상황에서 만족 혹은 불만족을 표출하는지 4가지로 정리하였다.

문서에서 추출된 단어를 통해 숨겨진 단어들을 추론하여 주제를 발견할 수 있는 LDA 모형으로 모델링을 진행하였다. 주제의 일관성을 측정하는 coherence 값을 기준으로 최적의 토픽을 4개로 설정하여 각 군집별로 파악한 가장 관련성이 높은 키워드들을 도출했다. 데이터 개수가 현저히 적었기 때문에 유의미한 해석이 어려웠으나, LDA 시각화를 통해 선정된 키워드들을 기준으로 관련된 내용의 리뷰들을 자세히 살펴봄으로써 4가지 인사이트를 도출할 수 있었다.

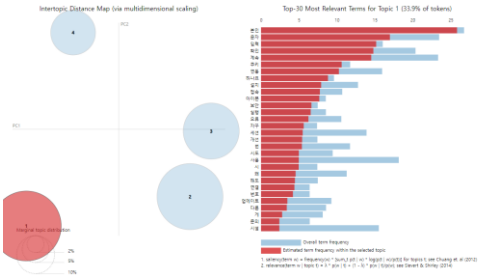


그림 19. 토픽 모델링 결과 시각화 (대표 : 토픽1)

첫 번째 토픽의 키워드로부터 '본인 인증' 이라는 주제를 도출할 수 있었고, 구체적인 리뷰를 확인해본 바 '휴대폰 문자 인증 불가능, 쿠키 설정 관련 오류' 등의 문제가 본인 인증 과정에서 발생함을 파악할 수 있었다. 두 번째 토픽의 키워드로부터 '회원 가입' 이라는 주제를 도출할 수 있었고 구체적인 리뷰를 확인해본 바 '도서관 서비스 이용 시 정회원 가입 신청이 불가능 하다' 는 문제가 존재함을 파악할 수 있었다. 여러 본인 인증 수단으로 가입 신청을 시도해도 계속 실패했다는 사용자 의견도 다수 보였다. 세 번째 토픽의 키워드로부터 '카드 발급' 이라는 주제를 도출할 수 있었고 구체적인 리뷰를 확인해본 바 '공공 시설 연계 카드' 와 '도서관 카드' 발급 시 오류가 발생함을 확인할 수 있었다. LDA 시각화 결과에서도 확인해볼 수 있듯 세 번째 토픽은 두 번째 토픽과 밀접하게

관련되어 있음을 확인해볼 수 있다. 마지막으로 네 번째 토픽의 키워드로부터 '핵심 기능 사용성에 관한 평가' 라는 주제를 도출할 수 있었고 구체적인 리뷰를 확인해본 바, 통합 카드 지원으로 사용성이 편리하다는 의견이 존재하는 반면 이와 반대로 사업의 목적성과 달리 통합 카드 사용이 오히려 불편하다는 의견도 존재하였다.

4.3 연계시설 정보 분석

마지막으로 '연계시설 정보 데이터' 를 바탕으로 서울시 연계 공공 시설, 문화 시설에 대한 정보가 다양하게 제공되고 있는지, 지역별로 편차가 존재하는지 등을 파악하기 위해 시각화를 진행하였다.

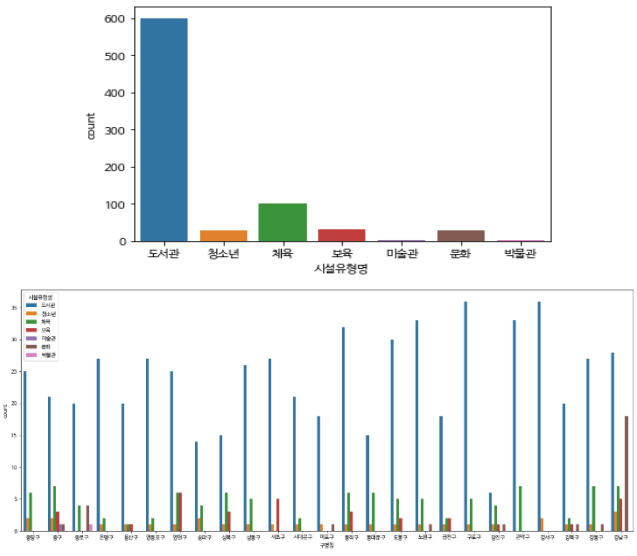


그림 20. (a) 시설 유형별 분포 (b) 지역별 시설 분포

시설 유형 별 분포를 살펴보았을 때, 연계된 시설 중 도서관이 가장 많았고 체육관, 보육 시설, 청소년 시설, 문화 시설 순으로 많았으며 미술관이나 박물관의 경우 각각 1개씩만 존재했다. 추가적으로 서울시에서 운영하는 공공 시설 정보에 관한 데이터를 병합해 연계 시설에 대한 정보를 충분히 반영하고 있는지 확인하는 과정을 거치는 것이 필요하겠으나, 도서관 시설 외에 보다 다양한 문화 시설에 관한 정보를 추가할 필요성이 보인다. 지역별로 시설 분포를 살펴보았을 땐, 강남구에 문화 시설이 집중적으로 분포하고 있음을 파악했고, 중구와 강남구 지역의 시설 분포가 가장 다양했음을 확인할 수 있었다. 지역별 서비스 제공의 편차를 줄이기 위해, 각 자치구별로 도서관 외의 문화시설 지원 여부를 확인해보아 관련 정보를 추가할 필요성이 있을 것으로 보인다.

5. 결론 및 한계점

본 연구는 서울 시민카드 이용자 증대를 위해 일별 가입자 현황 데이터와 어플 리뷰 데이터, 연계시설 정보 데이터에 대한 분석을 진행하여 어플리케이션 사업 현황을 파악하고 관련 개선 방안 아이디어를 제시하였다. 일별 가입자 현황 데이터로부터 다양한 시계열 모형을 적용하여 서비스 이용률을 가장 잘 예측할 수 있는 모델로 다변량 LSTM 을 사용해볼 것을 제안하였으며, 어플 리뷰 데이터 분석을 통해 사용자가 불편함을 호소하는 앱의 기능적 결함 요인과 만족감을 표시하는 어플 내 주요 기능을 살펴보았다. 더불어 연계 시설 정보 데이터를 시각화 하여 얻은 인사이트를 바탕으로, 지역별로 공공 시설 정보 제공의 편차가 발생하지 않도록 하는 것이 중요하며 도서관 이외의 다양한 문화 시설 정보를 추가하는 것을 제안하였다.

일별 가입자 현황 데이터에서 제공하는 정보가 제한적이고 사용자에 대한 구체적인 정보를 파악하기 어려워 분석 결과를 제시할 때, 방향성과 필요성만 제시했다는 한계점이 존재한다. 또한 접속자 수 변수에서 중복 접속이 포함되는지 여부를 파악하기 어려워 서비스 이용률 변수를 정의할 때 중복 접속은 고려하지 않았다는 한계점이 존재한다. 그러나 데이터에 대해 다양한 접근의 분석 방법을 적용하였다는 점과 공공 사업이라는 특수성에 기반하여 분석 결과를 제안한 것에 있어, 본 보고서는 추후 발전된 서울 시민카드 데이터 분석 연구의 기초 자료로 활용될 수 있을 것으로 기대한다.

참 고 문 헌

[1] 김경희 (2022), "서울시민카드 사용자 유입과 이탈에 대한 시계열 이상치 탐지와 DTW 클러스터링" , 성균관대학교 응용통계연구소, 통계연구 22권 0호, pp63-76

[2] 김준석, 강재환, 김성의, 윤주상 (2020), "Prophet 을 사용한 일변량의 시계열 예측", 한국정보통신학회, 춘계 종합학술대회 논문집, pp329-331

[3] 김준기, 류동근, 남형식 (2022), "Prophet 모형을 활용한 국내 중소형 컨테이너항만 물동량 예측에 관한 연구 : 인

천, 평택•당진, 울산항을 중심으로” , 아시아문화학술원, 인문사회21 제31권1호, pp561-575

[4]오승원, 임남희, 이상현, 김민수 (2020), “Prophet 모형을 이용한 마늘 가격의 장기 예측 및 트렌드 분석”, 한국자료분석학회

[5] 이지훈, 한혜림, 윤상후 (2020), “시계열 모형을 이용한 인천공항 이용객 수요예측”, 한국디지털정책학회, pp87-95

[6] 최애선, 김상수, 이현숙, 김중수, 염세경 (2022), “모바일 앱 서비스 업데이트가 사용자 만족도 및 유지율에 미치는 영향”, 한국통신학회, pp574-575

[7] 이새미, 이태원 (2021), “지역화폐 앱 사용자 리뷰 분석을 통한 마케팅 전략 수립 – 동백전과 인천 e음을 중심으로”, 한국콘텐츠학회, pp114-121

[8] 정지훈, 정혜인, 이준기 (2021), “텍스트 마이닝 기법과 ARIMA 모형을 활용한 배달의 민족 앱 리뷰 분석”, 디지털콘텐츠학회논문지, pp291-299

[9] 홍정림, 유미림, 최보름 (2019), “토픽 모델링을 활용한 모바일 증강현실 앱 사용자 리뷰 분석”, 디지털콘텐츠학회논문지, pp1417-1427

[10] 강성안, 김동연, 류민호 (2021), “텍스트 마이닝을 이용한 부동산 서비스 앱 리뷰 분석”, 한국정보시스템학회, pp227-245

[11] 김선영, 김명호 (2021), “다변량 시계열 데이터에서 이상 탐지를 위한 지식 증류” , 한국컴퓨터종합학술대회, pp992-994

[12] 변준형, 김지호 (2020), "딥러닝을 이용한 영화 흥행 예측과 주요 변수의 선택 연구 : 다변량 시계열 데이터를 중심으로", 한국컴퓨터정보학회, pp35-47

[13] 김규범, 최명락, 황찬익 (2020), "다변량 ARIMA, MLP 및 LSTM 을 활용한 지하수위 결측처리 비교", 지질학회지, pp561-569

[14] 김은지, 이택기 (2020), "LSTM 기반 다변량의 기상 데이터를 이용한 풍력 발전량 예측", 대한전기학회, pp245-246