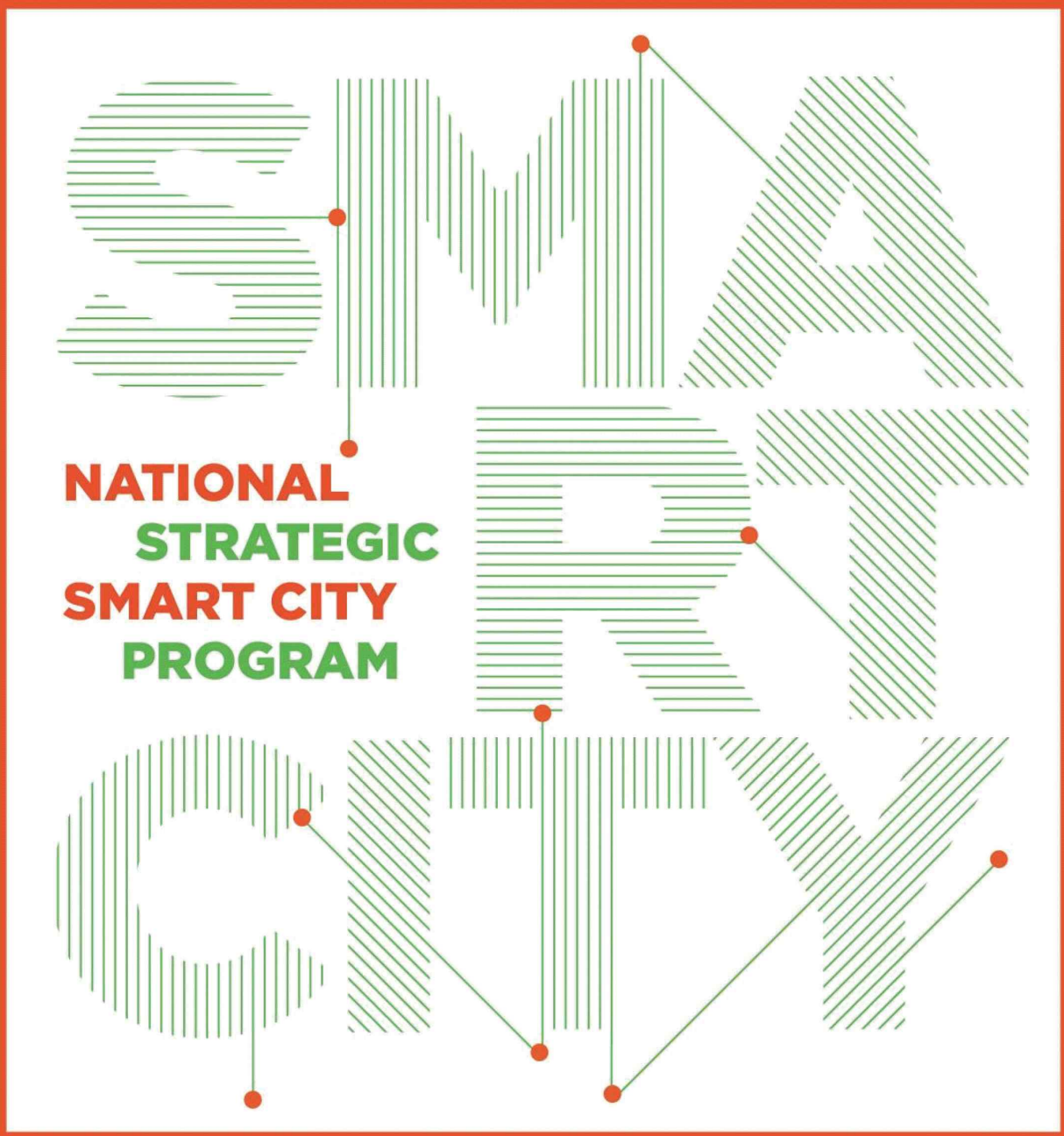


융복합 서비스 사례
– 버스노선 분석 서비스

Technical Report [1부-2권 별책1]

스마트시티
혁신성장동력 프로젝트

[2-3세부과제]
주관연구기관-SK텔레콤

국토교통부 과학기술정보통신부 KAIA 국토교통과학기술진흥원 DEAPCITY National Strategic Smart City Program

서비스 명	융복합 서비스
단위서비스 명	버스노선 분석 서비스

서비스 설명

- (배경) 대구광역시는 상대적으로 대중교통의 이용률이 낮아 수요에 맞는 효율적인 버스운영 필요성이 대두되었다. 도시는 시시각각 변화하기 때문에 버스노선의 수요와 효율적 이용에 대한 분석이 필요하다.
- (목적) 교통, 환경, 유동인구 데이터 등을 이용한 융복합 분석을 통해 대중교통 이용을 활성화하여 시민들의 승용차 이용이 감소하면 도로교통이 원활해지면서 에너지와 탄소 배출을 줄이고 도시 교통문제까지 해결하는 것을 목적으로 한다.

제공자·사용자 편익

- (지자체) 대구광역시 교통정책과 등
- (버스 운영사) 버스운영 비용효율성을 고려
- (시민) 시민의 버스 접근성, 정시성, 환승편의성 고려

운영방안 및 추정비용

- (운영주체) 지자체 교통정책과/ 산하 도시공사
- (운영방안) 데이터허브 관제센터에서 버스노선 분석
- (추정비용) 초기구축 비용 : 약 200백만원운영 · 유지보수 : 연간 약 30백만원

서비스 아키텍처

인프라 목록

구분	인프라명	수량
S/W	버스노선 분석 알고리즘	1식
서버	Data Lake(분석 서버)	5노드
클라이언트	버스노선 표출 서버	1식

데이터셋

연계 데이터	제공방식
대중교통이용율	API
대구수요지수	API
대수수요지수 시뮬레이션	API
버스노선 효율지수	API

As-is ⇒ To-be

As-is	To-be
대중교통 이용 수요 예측이 어려워 개선에 대한 가이드 부재	대기수요지수와 노선별 효율지수를 통한 대중교통 이용의 효율성 증대

활용데이터 및 시스템 연계 대상

대상기관	연계 데이터	유형
기상청	동네 날씨정보(온·습도)	API
SKT	유동인구, Tmap 데이터	Setup
LH	토지, 건축 데이터	Setup
DGB유페이	교통카드사용내역	DB2DB
대구광역시	버스, 주차장 데이터	DB2DB

Contents

• 목차 •

• 용어 정리 •

용어	정의
Exploratory Data Analysis	탐색적 분석/탐색적 데이터 분석, 데이터 사이언티스트가 데이터셋을 분석하고 조사하여 주요 특성을 파악하는 데에 사용하는 기법
K-Fold Cross Validation	모델 훈련과정에서 모든 데이터가 최소 한 번은 테스트 셋으로 쓰이도록 하는 기법(K겹 교차검증)
R-square	결정계수, 종속변수의 분산 중에서 독립변수로 설명되는 비율을 의미
Spearman	서열상관분석, Pearson 상관분석과 마찬가지로 두 변수 간의 상관관계를 분석하는 기법이나, Pearson 상관분석이 연속변수라면 Spearman은 서열척도로 측정된 순위형 변수임
결측값	데이터의 값이 없는 경우
다중공선성	통계학의 회귀분석에서 독립변수들 간에 강한 상관관계가 나타나는 문제로 독립변수들 간에 정확한 선형관계가 존재하는 완전공선성의 경우와 독립변수들 간에 높은 선형관계가 존재하는 다중공선성으로 구분하기도 함
대기수요지수	해당 지역에 대중교통의 이용 가능성이 있는 수요를 숫자로 표현한 지수
상관관계	상관관계는 2개 변수가 선형관계에 있는(상수 비율에서 함께 변경됨을 의미함) 범위를 표현하는 통계적 측도

• 표 목차 •

Contents

• 그림 목차 •

제1장 개요

1 | 배경 및 필요성

1-1 배경 및 수행 목표

◔ 배경

– 전국 대비 대구광역시는 대중교통 이용률이 낮고 승용차의 이용 비율이 높았다. 대중교통 이용이 활성화되면 승용차 이용이 감소되고 교통이 원활해진다. 또한 에너지와 탄소 배출을 줄일 수 있다는 점이 환경적으로 긍정적인 혜택으로 부각되었다.

– 대구광역시의 경우 대중교통 중 구축 비용이 높은 도시철도보다는 버스노선을 최적화하는 것이 효과적이라 볼 수 있다. 대구광역시는 2015년 도시철도 위주의 노선 개편으로 이용객이 증가하면서 도시철도와 시내버스의 환승률은 하락하였다. 버스 이용률의 감소 원인으로는 역사와 승강장 간 접근성 감소와 버스의 배차 간격 증가를 들 수 있다. 또 다른 원인으로 주 이용객층인 청소년 수의 감소로 인한 자연감소, 시민의 승용차 대수 증가로 대중교통을 덜 이용하는 현상을 들 수 있다.

– 결국 인구 대비 대중교통 이용 수요에 맞는 효율적인 버스 운영의 필요성이 대두되었다. 대규모 아파트 단지 개발로 인구가 증가하게 되면 다른 지역으로 출근하는 인구가 많아지고 기존 대중교통 노선은 불편하게 된다. 버스노선의 정확한 이용자의 수요인 '대기수요'를 파악하여 노선을 최적화하는 작업이 필요하다.

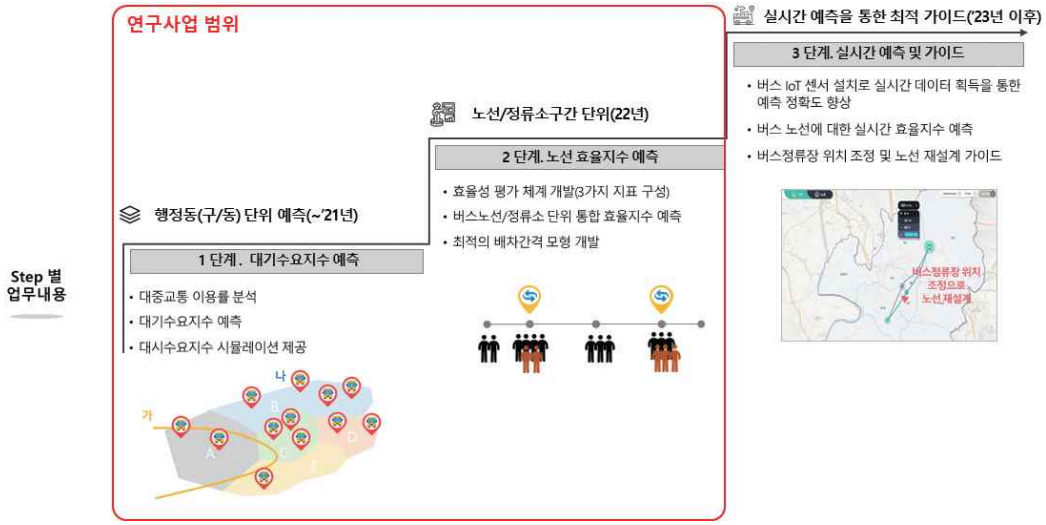
◔ 수행 목표

– 교통, 환경 데이터 등의 융복합 분석을 통해 대중교통 이용을 활성화하여 시민들의 승용차 이용이 감소하면 교통이 원활해지면서 에너지와 탄소 배출을 줄일 뿐만 아니라 복잡한 도시 교통문제 해결을 목적으로 한다.

- 대구광역시의 대중교통 이용 활성화를 위해 '대중교통 대기수요 예측', '시뮬레이션을 통한 버스노선 분석', '노선 효율지수 분석을 통한 개선 가이드'를 목표로 한다.
- 대구광역시 권역별, 구별, 행정동별 유동인구와 대중교통 이용인구 데이터에 기반해 잠재적 대중교통 수요 지표를 개발하는 것을 목표로 한다. 즉 머신러닝 모델 개발 방법론을 통해 대기수요지수를 예측한다.
- 대기수요지수 기반으로 버스노선을 최적화하기 위해 시뮬레이션을 개발하고 대기수요지수가 높은 주요 이슈 지역에 대해 버스노선 및 정류장의 최적화를 위한 기틀을 마련한다.
- 버스노선별, 구간(정류장과 정류장)별 효율지수를 예측하여 해당 노선 또는 구간에 대한 최적화 가이드를 마련한다.

1-2 수행 방안

- 버스노선 최적화를 위한 수행 방안은 융복합 분석 기획에서 도출된 데이터인 유동인구, 대중교통 이용인구, 대중교통인프라, 자동차 이용, 토지, 건축물, 카드 매출에 대한 데이터를 수집한다. 데이터 수집에 있어서 가능성을 모두 열어놓고 가능한 한 많은 데이터를 수집하는 것을 목표로 했다.
- 탐색적 데이터 분석을 통해 수집된 데이터의 현황을 파악한다. 통계적인 측면, 결측치 등 데이터의 특성을 파악하여 추가적으로 의미 있는 파생 변수를 도출한다. 또 탐색적 분석과정에서 발견되는 추가 데이터 수집도 진행할 수 있다. 탐색적 데이터 분석은 분석모델을 만들기 전에 데이터의 특성에 맞게 데이터 전처리를 위한 필수 단계이다.
- 버스노선 최적화 서비스의 목표인 '대기수요지수', '버스노선 효율지수'를 예측하기 위한 모델을 개발한다. 개발된 복수의 모델에 대해 모델 평가 방법에 의해 평가를 진행하여 최적의 정확도를 보이는 모델을 선택한다.
- 예측에 필요한 데이터 속성 기반으로 버스정류소 및 노선 증감에 따른 대기수요지수 변화를 분석하고 시뮬레이션한다. 개발된 모델의 결과 데이터를 이용하여 대중교통 이동정보, 대중교통 이동 동선, 대기수요지수, 리포트 화면 등 시각화 개발을 수행한다.



〈그림 1-1〉 버스노선 최적화 로드맵

– 버스노선 최적화의 로드맵을 살펴보면, 1단계 행정동 단위 예측은 대중교통 이용률을 분석, 대기수요지수 예측, 대기수요지수 시뮬레이션 모델을 만들며 4차년도(2021년)까지 진행되었다. 연구사업 마지막인 5차년도(2022년)에는 노선/정류소 단위로 효율성 평가 체계 개발, 노선 단위로 통합 효율지수 예측, 최적의 배차간격 모형 개발을 목표로 진행된다. 이번 연구사업의 목표는 2단계까지 진행이며, 마지막 3단계 실시간 예측을 통한 최적 가이드 사업은 대구광역시의 필요에 따라 추가로 진행할 수 있다.

2 | 서비스 특징

– 대중교통을 활성화하고자 하는 대구광역시의 정책과 반대로 현재 버스노선은 승용차와 비교했을 때(각 교통수단 속성 기준) 상대적으로 비효율적으로 운행되고 있어 문제 현황을 개선하고자 근본적인 원인이 되는 버스 운행에 대한 이슈 검토, 신규 BM 개발, 효율적인 운행을 위한 도시 교통 정책 수립 지원 목적으로 구현한다.
– 교통, 환경 데이터 등의 융복합 분석을 통해 대중교통 이용을 활성화하여 시민들의 승용차 이용이 감소하고 교통이 원활해지면 에너지와 탄소 배출이 줄 뿐 아니라 복잡한 도시 교통 문제를 해결할 수 있다.

– 버스노선 최적화 서비스는 버스 승하차, 유동인구 등 데이터를 활용하여 버스 이용률, 거점 간 이동시간 등 노선 효율성을 분석하고 의사 결정자에게 배차간격, 노선 통/폐합 등 다양한 솔루션을 제시한다.

3 | 기대효과

– 버스노선 분석은 다각도적인 대중교통 이동 흐름 분석이 가능하다. 예를 들어, 행정도시(다중 포함), 노선 등 설정 기능을 통해 다양한 형태로 대중교통 흐름을 이해하고 현황에 대한 이슈를 명확히 분석할 수 있어 추후 서비스 구현 시 활용이 가능하다.
– 현황 분석부터 Insight까지 제공되는 통합지원 서비스(시각화)이다. 익숙한 Map 구조 기반 솔루션을 통해 현황 파악이 빠르고 주변 변화 및 디테일한 이슈 변화도 쉽게 감지하여 추후 BM 상용 시 지원 서비스 활용이 가능하다.
– 효율성 있는 노선 시스템 기반으로 구축하였다. 지역별로 잠재적 이용 수요인 대기수요 예측 지수까지 추출하여 추후 상용서비스 구축 시 좀 더 정교화한 설계가 가능하도록 참고 및 활용이 가능하다.

제2장 연구 개발 성과

1 | 버스노선 최적화 시나리오

1-1 기능

◔ 대중교통(버스, 지하철)에 대한 구별, 행정동별 이용률 분석

– 구별, 행정동별로 시민들의 이동 중 대중교통을 이용한 비율을 예측하는 분석을 말한다. 전체 유동인구 중에 버스와 지하철 등 대중교통 이용한 비율로 계산한다. 현 시점에 시민들이 대중교통을 얼마나 이용하고 있는지를 알려주는 지표다.

◔ 구별, 행정동별 유동인구수, 이동시간 예측

– 구별, 행정동별 얼마나 많은 수의 인구가 이동하는지와 평균 이동시간을 예측하는 기능이다. 유동인구수와 이동시간을 통해 시민들의 대중교통 이용현황을 알려주는 지표다.

◔ 구별, 행정동별 대기수요지수 예측

– 구별, 행정동별 이동 시 대중교통(버스, 지하철 등)을 이용하지 않고 승용차를 이용하는 등 대중교통에 대한 대기수요를 예측하는 기능이다. 얼마나 많은 대중교통의 수요가 있는지를 나타내는 지표다.

◔ 버스노선별, 구간별 효율지수 예측

– 대구광역시 버스노선의 특징 및 문제점을 진단하고 버스노선별, 정류장과 정류장 사이 구간별 효율지수를 예측하는 기능이다. 효율지수는 이용자 관점, 운영자 관점, 사회적 관점에서 총 11개의 평가항목에 의해 추출된다.

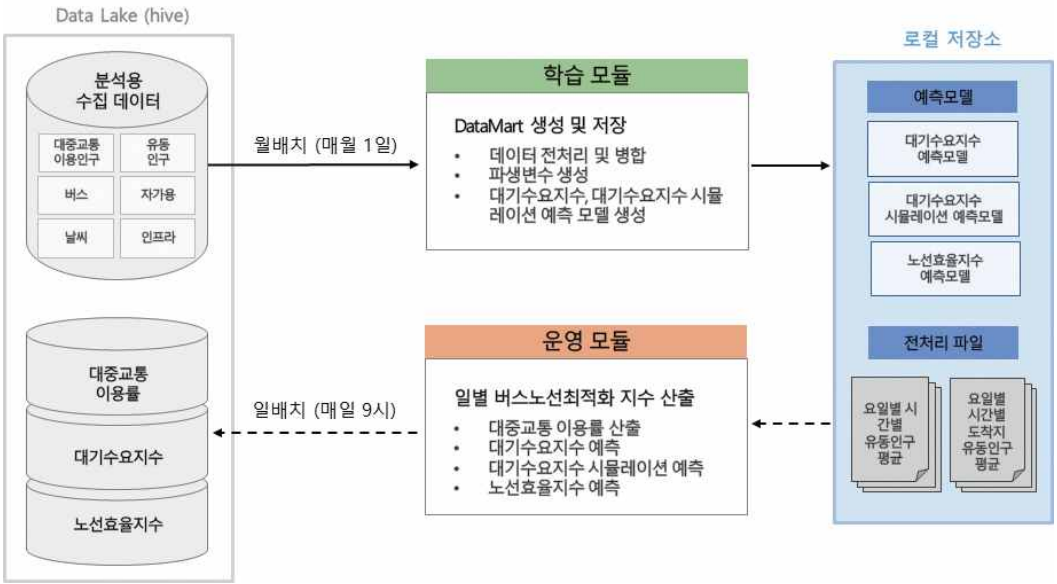
1-2 시나리오

- 대구광역시 교통에 대한 정책결정자가 적용지역을 대상으로 출발지, 도착지 권역(구별, 행정동별, 버스 권역)을 설정한다.
- 설정한 조건에 따른 대중교통 이용률, 이동시간, 유동인구 분석, 대기수요지수, 버스노선별 효율지수 결과 및 최적화 노선을 제안한다. 가능한 활용방안은 크게 다음의 세 가지를 들 수 있다. 첫째, 버스의 이동거리 분석을 기반으로 대구광역시 버스노선의 재설정 및 최적화를 위해 활용 가능하다. 둘째, 버스 이용률 분석 기반 중복노선에 대해 통/폐합 및 배차 간격 재조정에 대한 제안이 가능하다. 셋째, 대중교통 접근성 분석 기반 버스 정거장 위치 조정 등에 대한 정책 결정 가이드를 제공할 수 있다.
- 대구광역시 전역에 대해 지도 위에 대중교통 최적화를 위한 분석 결과를 제공한다. 또한 현황을 통해 개선 방향을 시각화해 표현할 수 있다.
- 시간, 지역별 조건에 따라 선택적 결과 조회가 가능하고 분석 데이터에 대해 생성·추출 기능을 제공한다.

〈그림 2-1〉 버스노선 최적화 시나리오

2 | 아키텍처(시스템 구성도)

– 버스노선 최적화 서비스는 빅데이터 프레임워크인 Data Lake에 구성되어 있다. 분석용 수집 데이터들은 Hive를 통해 접근이 가능하며, 대중교통 이용인구, 유동인구, 버스, 자가용, 날씨, 교통인프라 등 데이터를 보관하고 있다. 학습 모듈은 매월 1일에 수집된 데이터를 이용하여 알고리즘을 최신화하는 작업을 수행한다. 알고리즘은 대기수요지수 예측, 대기수요지수 시뮬레이션 예측, 노선효율지수 예측 모델이 있으며, 로컬 저장소에 보관된다. 이렇게 예측된 모델은 매일 일배치를 통해 운영 모듈이 수행되고, 전처리 파일을 이용하여 해당 날짜의 버스노선 분석 지수들을 산출한다.



〈그림 2-2〉 아키텍처 구성도

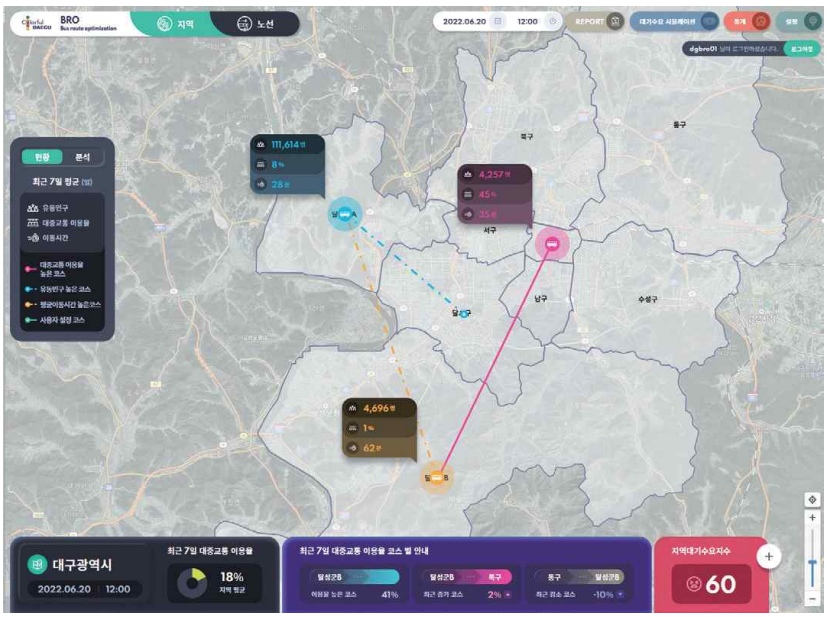
3 | 단위서비스별 시나리오

3-1 대중교통 이동정보(지역)

– 행정동별 버스노선 현황정보 확인이 가능하다. 아래 그림과 같이 최초 유동인구가 높은 코스, 대중교통 이용률이 높은 코스, 평균 이동시간이 높은 코스가 안내된다.

– 정보를 확인하고자 하는 행정 선택을 위해 지역 선택 클릭 후 대중교통 출발/도착지

역을 선택한다. 지역 선택은 구/동 단위로 선택할 수 있으며 선택된 행정 지역 정보가 MAP에 호출된다. 이를 통해 대중교통 이용률, 이동시간, 유동인구 기본 내용을 확인할 수 있다.



〈그림 2-3〉 서비스 메인 화면

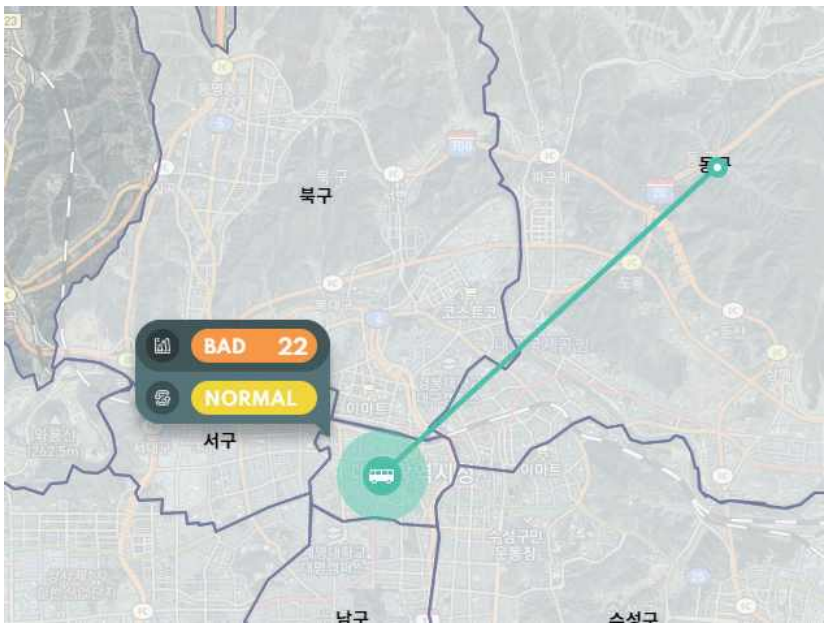
– 상세정보 확인을 위해 해당 코스를 누르면 대중교통 이용률, 이동시간, 유동인구 예측정보를 확인할 수 있으며, 대중교통 이용률 Tab에서는 최근 7일 평균 이용률 정보, 평균 이용률이 높은 동 단위 코스, 전체 주요 이동 코스 정보를 제공한다. 또 이동시간 Tab에서는 최근 7일 평균 이동시간 정보, 평균 이동시간이 높은 동 단위 코스, 전체 주요 이동 코스 이동시간 정보를 제공한다. 아울러 유동인구 Tab에서는 최근 7일 평균 유동인구정보, 유동인구 대비 이용인구 비율정보, 전체 주요 이동 코스 유동인구정보를 안내한다.

3-2 효율/대기수요 지수(지역)

○ 행정동별 효율지수, 대기수요 지수 분석정보 확인이 가능하다.

– 정보를 확인하고자 하는 행정 지역 선택을 위해 '지역 선택' 클릭 후 대중교통 출발/도착지역을 선택한다. 지역 선택은 구/동 단위로 선택할 수 있다. 선택된 행정 지역

정보가 MAP에 호출되며 간단한 내용을 확인할 수 있다.



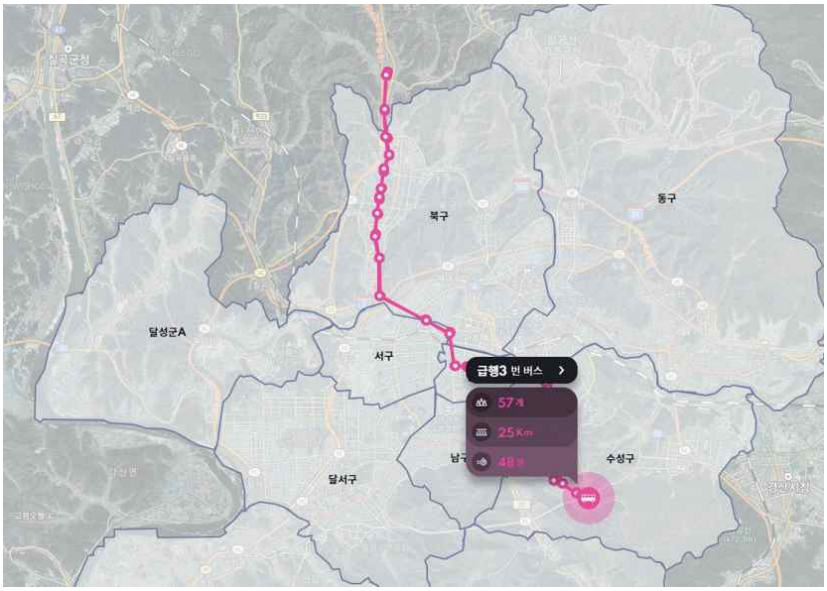
〈그림 2-4〉 지역 선택 후 효율/대기수요 지수 메인 화면

– 상세정보 확인을 위해 해당 코스를 누르면 효율지수/대기수요 지수정보를 확인할 수 있으며, 효율지수 Tab에서는 최근 7일 평균 효율지수 정보, 평균 효율지수가 높은 동 단위 코스, 전체 주요 이동 코스 효율지수정보를 안내한다. 대기수요 지수 Tab에서는 최근 7일 평균 대기수요 지수정보, 평균 대기수요 지수가 높은 동 단위 코스, 전체 주요 이동 코스 유동인구정보를 확인할 수 있다.

3-3 대중교통 이동정보(노선별)

○ 버스노선별 현황정보 확인이 가능하다.

– 버스노선별 총정거장 수, 총운행 거리, 평균 운행시간이 표시되며, 정보를 확인하려면 노선을 선택한다. 노선 선택은 버스노선 단위로 선택할 수 있다. 선택된 노선정보는 MAP에 호출되며 간단한 내용을 확인할 수 있다.



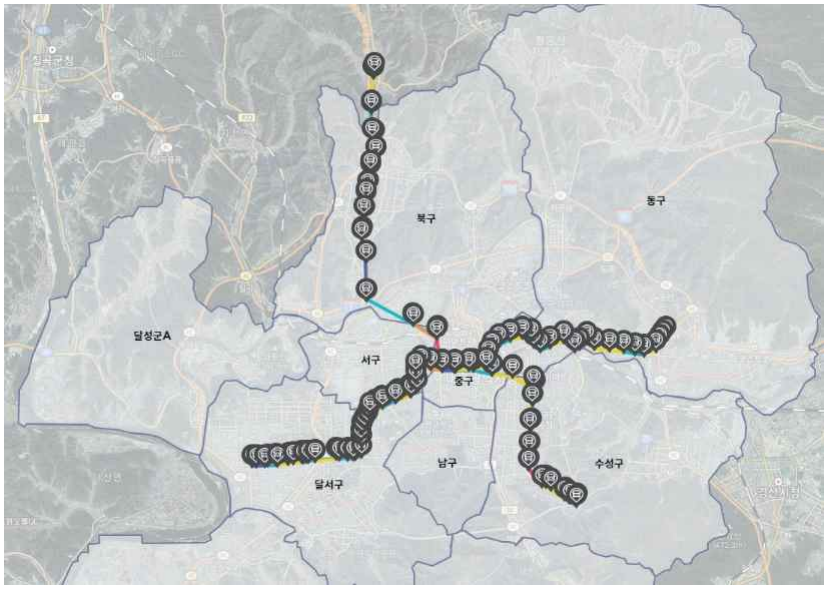
〈그림 2-5〉 노선 선택 후 버스현황정보 메인화면

– 상세정보 확인을 위해 해당 노선을 누르면 버스 이용률/운행시간 정보를 확인할 수 있다. 버스 이용률 Tab에서는 최근 7일 버스노선 평균 이용률 정보, 최근 7일 평균 이용률이 높은 운행구간, 주요 정류소 평균 이용률 정보를 안내한다. 운행시간 Tab에서는 최근 7일 평균 운행시간 정보, 최근 7일 평균 높은 운행시간의 운행구간, 정류소별 운행시간 정보를 확인할 수 있다.

3-4 효율/대기수요 지수(노선)

○ 버스노선별 효율지수, 대기수요 지수 분석정보 확인이 가능하다.

– 정보를 확인하고자 하는 노선 선택을 위해 노선 선택 클릭 후 버스노선을 선택한다. 노선 선택은 버스노선 단위로 선택할 수 있다. 선택된 노선정보가 Map에 호출되며 간단한 내용을 확인할 수 있다.



〈그림 2-6〉 노선 선택 후 버스노선 효율/대기수요 메인화면

– 상세정보 확인을 위해 해당 노선을 선택하면 정류장별 지수 및 현황을 확인할 수 있다. 정류장별 지수 Tab에서 제공하는 정보는 정류장 이동 간의 효율/이용/운영/사회지수를 확인할 수 있고 지수 현황 Tab에서는 최근 30일 지수정보, 최근 7일 평균 효율지수 Best 5, 최근 7일 평균 효율지수 Worst 5정보를 확인할 수 있다.

〈그림 2-7〉 버스분석정보 (정류장별 지수)

3-5 리포트 화면정보

○ 지역별/ 노선별 Report 정보 확인이 가능하다.

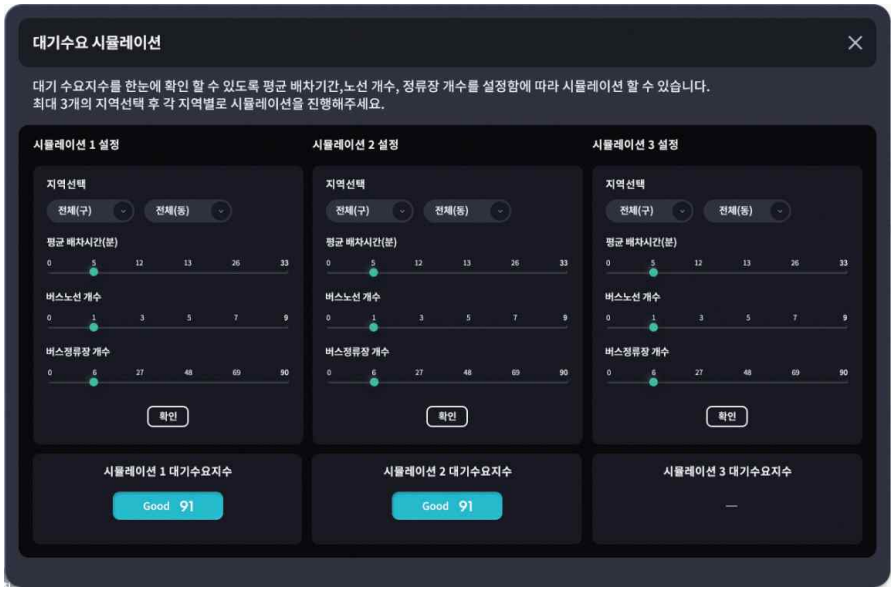
– 서비스 메인 확인 후 오른쪽 위에 배치된 Report 버튼 클릭 시 지역/노선별 Tab 구성으로 정보가 안내된다. Report 제공정보는 지역과 노선 2가지 Tab으로 구분하여 확인할 수 있다. 지역 Tag에서는 대중교통 활성화 필요 코스, 대중교통 조정 필요 코스, 유동인구 현황정보를 확인할 수 있다. 노선 Tab에서 제공하는 정보는 활성화 필요 버스노선, 조정 필요 버스노선, 이용률이 높은 운행구간 현황정보를 확인할 수 있다.

〈그림 2-8〉 Report (지역)

3-6 시뮬레이션 화면정보

○ 동 단위 지역별로 대기수요 지수 시뮬레이션이 가능하다.

– 서비스 메인 확인 후 오른쪽 위에 배치된 대기수요 시뮬레이션 버튼을 클릭하면 설정 화면이 안내된다. 최대 3개 지역까지 비교 시뮬레이션이 가능하며, 먼저 시뮬레이션을 진행할 지역을 선택 후 평균 배차 시간, 버스노선 개수, 버스 정류장 개수를 설정 후 확인 버튼을 선택하면 선택한 설정값에 따라 각각의 대기수요 지수 시뮬레이션 값이 안내된다.



〈그림 2-9〉 대기수요 시뮬레이션 선택

3-7 통계 화면정보

○ 버스노선 최적화 서비스에서 제공되는 수치에 대한 상세 통계를 확인할 수 있다.

- 제공되는 통계 항목에는 크게 지역과 노선에 관련된 각각의 특성에 따라 통계정보가 제공된다. 지역 데이터 기반으로는 대중교통, 이동시간, 유동인구, 효율지수, 대기수요 지수 통계를 확인할 수 있으며, 버스 데이터 기반으로는 노선정보 통계를 확인할 수 있다.
- 서비스 메인 확인 후 오른쪽 위의 배치된 통계 버튼 클릭 시 통계정보가 안내된다. 지역 대중교통 이용률 상세정보에서는 지역 설정(행정동)을 통해 출발지와 도착지 기준의 대중교통 평균 이용률을 일자별, 월별 평균 수치 확인이 가능하다. 또 대중교통 평균 이용 시간 상세정보에서는 지역 설정(행정동)을 통해 출발지와 도착지 기준의 대중교통 평균 이용 시간의 일자별, 월별 평균 수치 확인이 가능하다.
- 유동인구 상세정보에서는 지역 설정(행정동)을 통해 출발지와 도착지 기준의 평균 유동인구 수치를 시간대별로 확인할 수 있으며, 일자별 평균 유동인구 수치, 월별 평균 유동인구수치의 확인이 가능하다. 효율지수 상세정보에서는 지역 설정(행정동)을 통

해 출발지와 도착지 기준의 평균 대중교통 효율지수를 일자별, 월별 평균 수치로 확인할 수 있다.

– 대기수요 지수 상세정보에서는 지역 설정(행정동)을 통해 출발지와 도착지 기준의 평균 대중교통 대기수요 지수를 일자별, 월별 평균 수치로 확인할 수 있다. 노선정보 이용률 상세정보에서는 버스노선(전체 코스 기준)만 선택하면 해당 노선의 전체 구간 평균 운행시간, 평균 이용률을 확인할 수 있으며 일자별/월별 평균 수치를 확인할 수 있다. 노선정보 이용률의 경우 버스노선의 상/하행 코스를 직접 선택하여 해당 버스노선 코스의 정류소별 평균 이용률, 효율지수 확인이 가능하다.

〈그림 2-10〉 통계(대중교통)

3-8 기타 화면정보

○ 설정 기능을 통해 해당 서비스에서 제공되는 수치에 대해 기준치를 설정할 수 있다.

– 서비스 메인 확인 후 오른쪽 위에 배치된 설정 버튼 클릭 시 통계정보가 안내되며 변경하고자 하는 지수의 범위값을 숫자로 입력하여 값을 변경 후 확인 버튼 선택 시 변경 기준값에 따라 서비스의 모든 값이 변경되어 표시된다.

〈그림 2-11〉 설정화면

○ 빠르게 정보에 접근할 수 있도록 서비스 메인 하단에 주로 보는 '특정 현황정보 수치'에 대한 정보를 안내하며 클릭 시 확장되어 상세 내용을 확인할 수 있다.

– 서비스 메인 하단에는 주로 보는 정보가 안내되며 +, – 버튼을 통해 확장/축소하여 볼 수 있다. 특정 정보 현황은 대구광역시 전체 기준의 정보로 최근 7일 평균 유동인구, 평균 대중교통 이용률, 평균 이동시간 정보, 일별 이용률 변화 추이, 평균 이동시간 변화 추이, 최근 7일간 대중교통 이용률 높은 코스, 최근 7일간 주요 이슈 코스, 대구광역시 전체 대기수요 지수 및 행정구별 대기수요 지수정보를 안내한다.

〈그림 2-12〉 특정 정보 확장화면

4 | 요소기술

4-1 데이터 수집 및 전처리

분석 데이터 마트 구성

- 대기수요자 수 예측 모델 개발에 유의미한 데이터 분석 결과 도출을 위하여 대중교통 이용에 영향을 미치는 외부 환경 요인, 내부 환경 요인, 인구통계학적인 요인과 관련된 데이터 수집 등 데이터 풀을 구성한다.
- 수집된 데이터 풀을 대상으로 탐색적 분석(EDA)을 수행한다. 데이터를 분석에 맞게 전처리하는 과정을 거친다.
- 탐색적 데이터 분석은 수집한 데이터를 기반으로 데이터 현황 및 데이터 분포를 파악하는 작업을 말하며 산점도와 같은 그래프로 표현하여 확인한다.
- 파생 변수 생성은 분석 데이터의 시간/공간 기준에 따라 데이터 분할 및 통합을 의미 있게 할 수 있는 속성이 있는지 파악한다. 데이터의 특성을 반영하여 분석에 정확도를 높여줄 파생변수를 생성한다.
- 값의 품질이 불량한 데이터를 1차적으로 제고하고 타깃 분석 범위를 위한 데이터를 추출한다. 데이터가 빈값으로 수집된 결측값에 대해 이상치 처리를 한다.
- 유동인구 데이터를 기존으로 EDA 및 데이터 전처리 과정에서 생성한 속성들을 통합하여 분석 데이터 마트를 구성한다.
- 버스노선 분석에서 유동인구, 교통카드, 자동차, 버스, 날씨, 토지/건축, 카드매출, 교통인프라 데이터가 활용되었다.

데이터 수집

- 대구지방 경찰청, 대구광역시, LH, SK텔레콤의 협조를 통해 수집한 원천 데이터를 Data Lake에 적재하고, 대기수요지수 예측모델 개발을 위해 다양한 분야의 데이터를 수집한다.

〈표 2-1〉 데이터 수집 내역

구분	데이터	활용 컬럼	출처
인구통계 데이터	유동인구	행정동 단위 유입지 시간대별 유입인구	SKT
교통 데이터	대중교통 인프라	행정동 단위 버스정류소, 버스노선, 지하철역, 지하철노선 개수	대구광역시
	주차장 인프라	공용 / 민간 / 총합 주차면수	대구광역시
	버스	버스노선 개수, 일일 운행 수, 평균배차시간	대구광역시
	자가용 유입수	행정동 단위 자동차 대수, 평균 거리, 평균 소요시간	Tmap
금융상권 데이터	교통카드 승하차 정보	대중교통이용자수, 대중교통 평균 소요시간	대구광역시
	카드매출	일별 시간대별 매출	SKT
공간특성 데이터	토지	상업/주거/공업/기타 지역 비율	LH
	건축	세대수, 가구수, 건축수	LH
환경 데이터	날씨	강수확률, 습도값, 풍속값, 강수량, 기온값	기상청

◔ 파생변수 생성

– 데이터 분석을 위해 구성한 데이터셋을 기반으로 파생변수를 〈표 2-2〉와 같이 생성한다.

〈표 2-2〉 파생변수 내역

변수명		설명	선정 사유
날짜/시간	시간대	유동인구와 대중교통 이용인구의 시간별 특징에 따라 시간대 그룹	대기수요자 수가 시간대별로 차이가 있음
	요일	1주일의 각 날	대기수요자 수가 요일별로 차이가 있음
	월	12개월의 각 날	대기수요자 수가 월별로 차이가 있음
	계절	봄, 여름, 가을, 겨울에 대한 계절 그룹	대구수요자 수와 계절과의 상관관계를 확인하기 위해 선정
	공휴일	해당 날짜가 법정 공휴일인지 여부	대중교통 이용의 공휴일 여부에 의한 영향을 받는지 확인하기 위해 선정

인구통계	평균 유동 인구수	이전 달 요일별 시간대별 유동인구 평균값	이전 달 요일별 시간대별 평균 유동인구 수가 대기수요지수에 미치는 영향을 보기 위해 선정
	최대 유동 인구수	이전 달 요일별 시간대별 유동인구 최댓값	이전 달 요일별 시간대별 유동인구 최댓값이 대기수요지수에 미치는 영향을 보기 위해 선정
	도착지 유동인수	이전 달 도착지 요일별 시간별 유동인구 평균값	도착지의 평균 유동인구 수가 대기수요지수에 미치는 영향을 보기 위해 선정
시차변수	한 시간 전 대기 수요자수	한 시간 이전의 대기수요자 수	한 시간 이전의 대기수요자 수가 대기수요지수에 미치는 영향을 보기 위해 선정
	하루 전 대기 수요자수	하루 전 같은 시간의 대기수요자 수	하루 전 대기수요자 수가 대기수요지수에 미치는 영향을 보기 위해 선정
	대기 수요자수	대기수요자 수 = 유동인구 – 대중교통 이용인구	대기수요지수 예측모델 예측값 산출을 위해 선정
예측값	대기 수요자수	대기수요지 수 = 대기수요자 수/유동인구*100	대구수요지수 예측모델 예측값

○ 데이터 전처리

– 데이터 전처리 과정은 데이터 분석가의 업무 시간 중 60% 정도를 차지하는 만큼 중요하다. 하지만 여러 가지 측면을 고려해야 한다는 점에서 결코 쉽지 않은 과정이기도 하다. 가장 메인 데이터인 교통카드 데이터는 초기 3개월 샘플의 경우 원본 데이터의 건수는 1억5천만이고 27개 속성을 활용하였다.

– 1차 데이터 가공에서는 교통카드 데이터 중 버스노선과 상관이 없는 택시를 이용한 건수를 제외하였다. 2차 데이터 가공에서는 교통카드의 역정보가 없는 경우 등 이상 태그 내역을 제거하고 중복 데이터도 제거하였다. 3차 데이터 가공에서는 카드번호별로 승하차 정보로 가공한 이동 통계 데이터를 생성하였다. 이때 승차 후 하차가 이어지는 데이터가 아니면 제외하였다. 또한 이동시간이 0 이하인 데이터도 제외하였다. 행정동별 교통카드 이용인구 및 평균 소요시간을 집계한 행정동 통계 데이터 또한 생성하였다.

– 교통카드 데이터는 원본데이터(156,634,191건)에서 1차 가공(147,973,668건), 2차 가공(146,892,930건), 3차 가공(8,790,495건)으로 도출되었다.

〈그림 2-13〉 교통카드 전처리 데이터 건수 변화

– 출발지와 도착지가 같은 주거인구에 대해서는 유동인구가 아니므로 유동인구 데이터 전처리에서 제외하였다. 전체 30,505,361건에서 276,888건이 제외되어 30,228,473건을 도출하였다.
– 대중교통 데이터는 분석 범위인 대구광역시 외 지역 데이터를 제외하고 대중교통 타입을 통합한 전체 대중교통 이용인구를 산출하였다. 버스노선 데이터는 버스노선별 정류소 데이터를 기반으로 행정동 별 경유노선 개수 데이터를 생성하였다. 이는 행정동별로 대기수요지수를 예측에 활용하기 위한 속성으로 활용하기 위함이다.
– 기타 데이터로는 교통 인프라 및 건축/토지 데이터가 있으며 이 데이터를 활용하기 위해서 행정동 표준 주소 체계를 채택했다. 대기수요지수를 산출하는 유동인구 데이터와 대중교통 이용 데이터의 기준인 행정동으로 되어 있었기 때문에 동일한 체계로 통일했다.

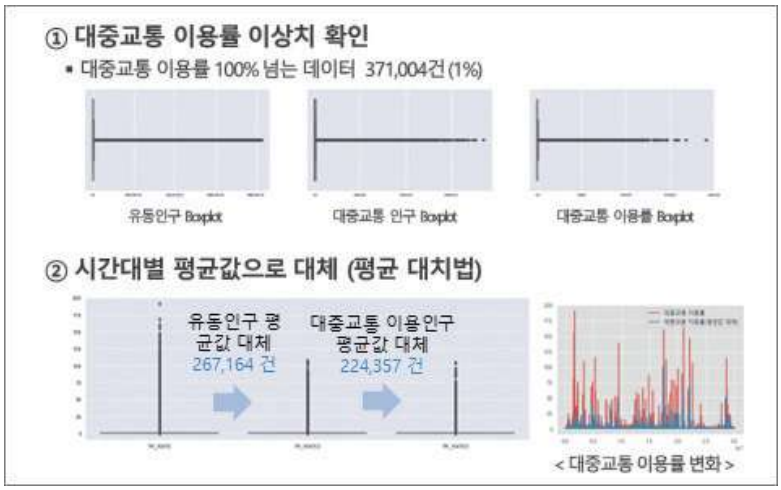
○ 데이터 결측 및 이상치 처리

– 앞서 수집한 데이터를 기반으로 데이터 현황을 파악하고 데이터 결측 및 이상치 처리를 수행하였다. 결측값을 확인하기 위해 품질점검도 하였다. 예를 들어, 교통카드 이용인가가 없거나 하차 태그가 없는 결측값 데이터에 대해 대체 처리하는 작업을 수행했다. 결측값을 보완하는 방법으로 출발지와 도착지의 시간대별 평균값을 구해

서 대체하는 방식을 사용했다. 하지만 평균값으로 대체할 수 없는 결측값도 존재했다. 새벽 시간대이거나 대중교통인프라가 마련되어 있지 않아 이용이 불가능한 지역의 경우 대중교통 이용 데이터 자체가 없었다. 이런 경우는 숫자 0으로 대체하는 방법을 사용했다.

〈그림 2-14〉 결측값 처리 방안

– '대중교통이용률'을 구하는 방식은 지하철과 버스를 이용한 '교통카드이용내역'을 유동인구 이동량으로 나눈 값이다. 하지만 모든 시민이 DGB유페이의 교통카드를 이용하여 이동하지는 않기 때문에 특정 시점에는 이상치가 발생할 수 있다. '대중교통이용률'이 100%를 넘는 데이터 1%(371,004건)는 이상치로 판단하였다. 이런 데이터의 경우 시간대별 평균값을 구해서 "평균 대치법"을 사용하여 처리하였다. 또 IQR 방식으로 이상치 탐색을 진행하였고 최댓값(0.41)이 넘는 이상치의 경우 Scaling(0.41~1)하여 보정하는 방법을 사용하였다.

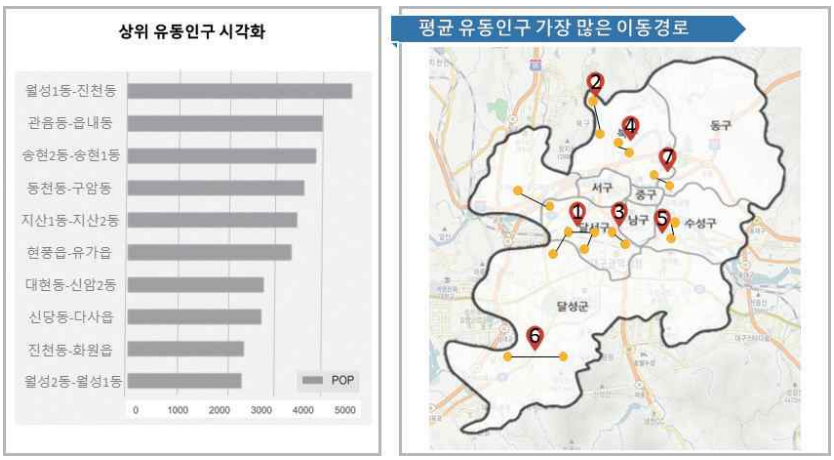


〈그림 2-15〉 이상치 처리 방안

4-2 탐색적 데이터 분석

유동인구 분석

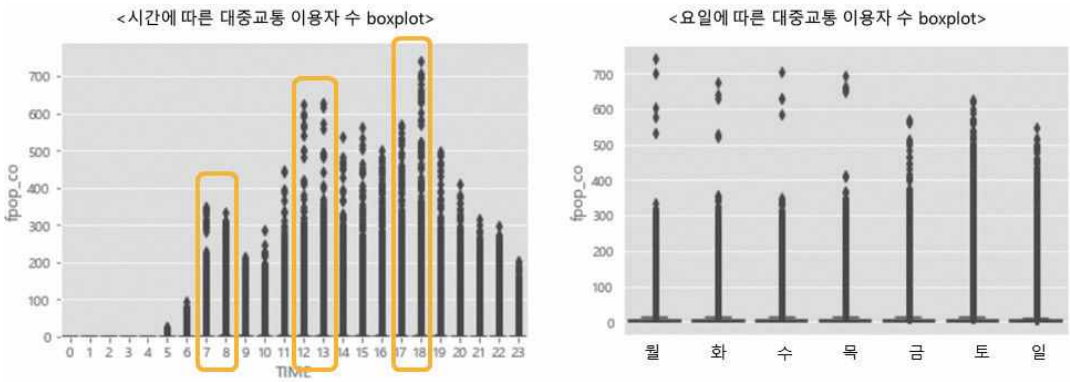
– 유동인구는 SK텔레콤에서 파일 형태로 수집한 데이터로 대구광역시민들이 이동한 전체적인 흐름을 알 수 있는 데이터이다. 평균 1시간 단위로 유동인구 이동량의 순위를 시각화해보면 달서구에서 가장 많은 유동인구가 있었음을 알 수 있다. 대구광역시의 특징으로는 수도권과 다르게 같은 구내에서의 이동이 대부분을 차지하는 경향성을 보이는 점이었다. 또 한 가지 특징으로는 대구 중심부인 달서구, 수성구 등에서의 이동이 많음을 알 수 있었다.



〈그림 2-16〉 유동인구 시각화 분석

대중교통 이용시간 분석

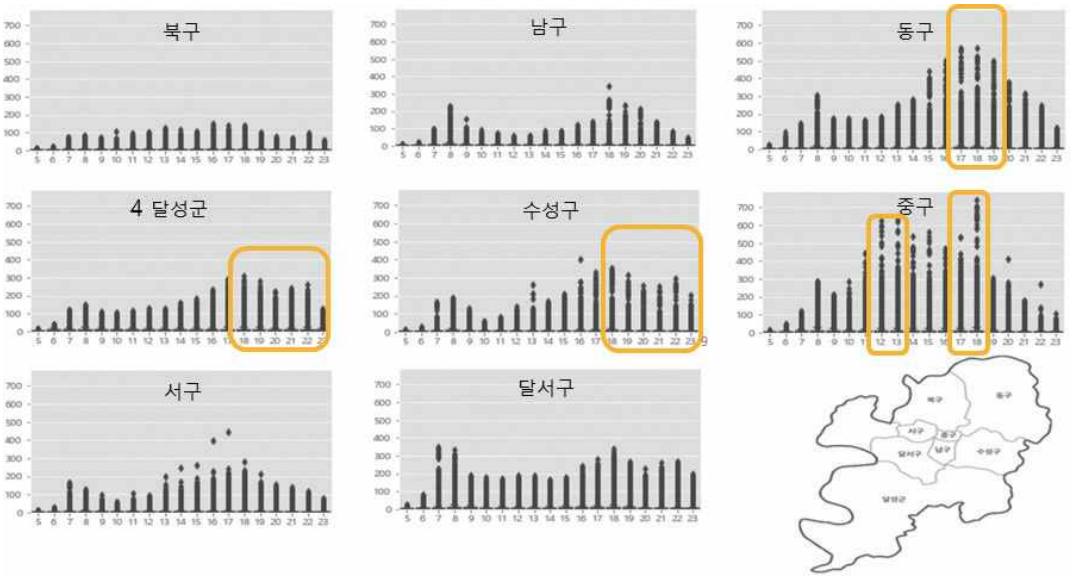
- 대중교통 이용자 수에 대해 시간대별로 박스플롯으로 분석해 보았다. 대부분 시간대에서 대중교통 이용자 수가 10 이하로 매우 작고, 최대 대중교통 이용자 수인 742명까지 골고루 퍼져 있음을 알 수 있었다. 버스 첫 이용시간인 새벽 5시부터 시작해 출근시간인 7~8시에 급증했다가 이용인구가 줄어들며, 점심시간인 12~13시에 다시 급증함을 확인할 수 있었다. 이 데이터로 보아 직장인이 출근을 위해 대중교통을 활용하는 숫자 이상으로 일반 시민들도 이용함을 알 수 있었다. 또한 하루 중 퇴근시간인 17~18시에 대중교통 이용자 수가 가장 많은 것으로 보아 퇴근 직장인 수와 저녁 약속을 위해 이동하는 수요가 더해져 나타나는 현상으로 해석할 수 있었다.
- 요일에 따른 대중교통 이용자 수요 분석에서 요일별 차이보다 주말과 평일의 대중교통 이용자 수의 차이가 컸다. 평일의 대중교통 이용자 수가 주말에 비해 많음을 알 수 있었다. 주말에 비해 평일에 대중교통 이용자 수가 급증하는 이유는 직장인 출퇴근의 영향이라고 추론할 수 있다. 또 출퇴근 시에는 자동차보다는 대중교통을 선호하는 것으로 보이며, 반대로 주말의 이동에서는 자가용을 이용하는 수요가 많음을 추론할 수 있다.



〈그림 2-17〉 대중교통 이용자 수 시각화 분석

- 출발 구별 대중교통 이용자 수 분석 측면에서 보면 북구와 남구는 대중교통을 이용해 이동하는 사람이 다른 구에 비해 매우 적음을 알 수 있었다. 또 유독 중구에서 12~13시인 점심시간대에 대중교통 이용자 수가 많으며, 실제로 중구에 음식점, 영화관, 서점 등의 문화시설이 밀집되어 있음을 확인할 수 있었다. 한편 동구와 중구에서는 퇴근시간에 대중교통 이용자 수가 월등히 높아 회사가 있는 업무지구일 것으

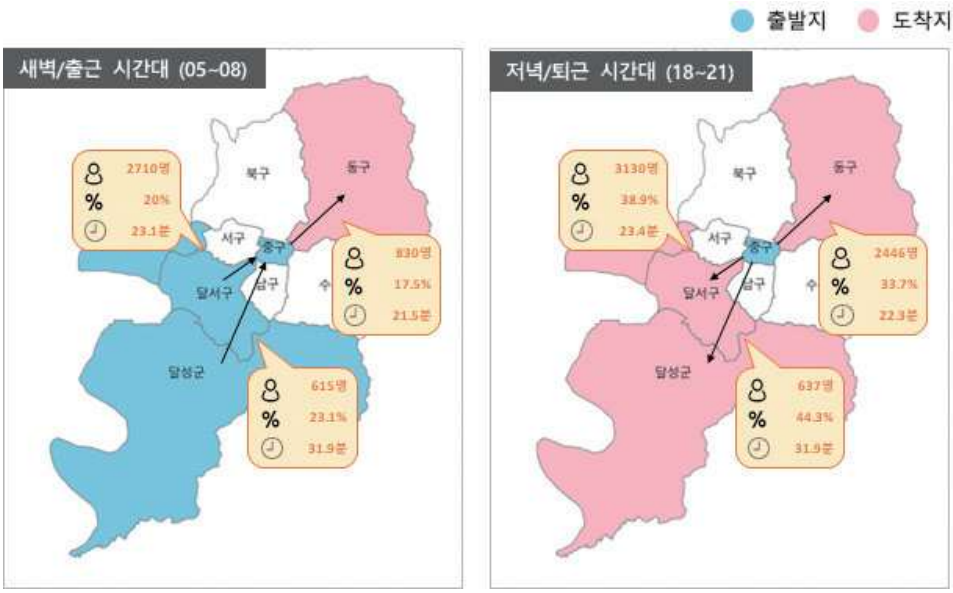
로 추정할 수 있다. 수성구와 달성군의 저녁 시간대 대중교통 이용자 수가 다른 시간보다 많은 것을 알 수 있고, 실제로 수성구에 고등학교와 학원가가 밀집해 있는 것 알 수 있다.



〈그림 2-18〉 구별 대중교통 이용자 수 시각화 분석

대중교통 이용률 분석

– 대중교통 이용률에 대해 구/군 단위 분석 결과는 중구에서 출발하는 경로가 높은 순위를 차지했다. '중구→달성군'이 22.2%, '중구→달서구'가 21.8%, '중구→동구'가 19.6%로 도심지인 중구에서 출발하는 경로의 대중교통에 대해 시민 이용률이 높게 나타났다. 주요 시간대별로 분석해 보니, 출근시간대 지하철 승하차 인원이 가장 많은 동대구역이 위치한 동구로의 이동경로에서 대중교통 이용이 활발하다. 또 달서구와 달성군 주거 단지에서 상업지역인 중구로 이동할 때 대중교통 이용률이 높게 나타났다. 주간 시간대, 저녁/퇴근 시간대, 야간 시간대 모두 중구에서 출발할 때 대중교통 이용률이 높았으며, 중구에서 주요 주거지인 달서구와 달성군, 동구로 대중교통을 이용하여 이동하는 것을 알 수 있었다.



〈그림 2-19〉 대중교통 이용률 시각화 분석

- 행정동 단위 대중교통 이용률 분석에서 평균 대중교통 이용이 가장 많은 코스는 중구 남산2동과 성내2동에서 출발하는 코스였다. 해당 동은 1호선, 2호선 환승역이면서 대구에서 일일 승하차 인원이 가장 많은 반월당역이 위치해 있는 교통 중심지임을 알 수 있었다. 행정동 단위로 시간대별 대중교통 이용률을 보면 평균적으로 5시~8시 사이에 대중교통 이용이 가장 활발하며 이것은 직장인들의 출근을 위한 이동으로 판단되며, 상대적으로 주간 시간대인 9시~17시에 가장 이용이 저조한 것을 알 수 있다.
- 일반적으로 교통 인프라가 좋고 유동인구가 많은 지역에서 대중교통을 이용하는 인원이 많아 이용률이 높게 나타난다. 이용률이 높은 지역은 승하차 인원이 많은 지하철역과 버스정류장이 위치해 있는 것을 알 수 있었다. 시간대별로 볼 때 출근시간대인 5~8시에는 상업시설이 발달되고 교통이 밀집된 중구로 이동하는 대중교통 이용률이 높으며, 계명대와 산업단지가 있는 달서구 신당동으로 이동할 때 대중교통 이용이 활발했다. 18~21시 시간대에는 아파트가 밀집되어 주거단지가 형성되어 있는 행정동으로 이동하는 현상이 나타났다.

〈그림 2-20〉 시간대별 대중교통 이용률(행정동 단위)

상관관계 분석

– 대기수요지수 상관관계 분석 기법으로 계량형 변수 또는 순서형 변수 사이의 단순 관계를 평가하는 방식인 Spearman을 사용하였다. 대기수요자 수와 다른 변수 간에 상관도를 보았을 때, 한 시간 전(Lag_1), 하루 전 대기수요자 수(Lag_1D)가 가장 관련성이 높게 나왔다. 자동차 수, 대중교통 이용자 수와 양의 상관관계를 가지는 것으로 보아 유동인구가 많은 지역일수록 대기수요자 수도 많음을 알 수 있다. 대중교통 이용자 수는 대중교통 평균 소요시간과 유동인구의 통계값 사이에 상관성이 있음을 알 수 있다. 지하철역 개수, 버스경유노선 개수와 같이 대중교통 편의성이 대중교통 이용자 수와 관계가 있음을 알 수 있었다.

<대기수요자 수와의 상관계수 상위11>

	corr	
lag_1	0.992158	한시간 전 대기수요 자 수
lag_1D	0.974219	하루 전 그 시간대에 대기수요자 수
POP_MEAN	0.772943	요일별 시간대별 유동인구 평균값
POP_MAX	0.756676	요일별 시간대별 유동인구 최댓값
car_num	0.362221	자동차 대수
ROUTE_CNT	0.297358	버스노선개수
HCODE_POP	0.174165	도착지 요일, 시간대별 유입인구 평균값
fpop_co	0.150288	대중교통 이용자 수
strt_area	0.065608	출발지권역
cd3	0.046835	공업지 비율
bus_stvr_rut_co	0.039400	버스경유노선개수

<대중교통 이용자수와의 상관계수 상위12>

	corr	
avrg_rqr_tm	0.242534	대중교통 평균 소요시간
POP_MEAN	0.226966	요일별 시간대별 유동인구 평균값
POP_MAX	0.224276	요일별 시간대별 유동인구 최댓값
car_num	0.186442	자동차 대수
ROUTE_CNT	0.160118	버스노선개수
WAIT_POP	0.150288	대기수요자 수
lag_1D	0.149173	하루 전 그 시간대에 대기수요자 수
lag_1	0.143912	한시간 전 대기수요 자 수
sbw_stvr_rut_co	0.138399	지하철 경유노선 개수
HCODE_POP	0.129213	도착지 요일, 시간대별 유입인구 평균값
sbw_sttn_co	0.107866	지하철 역 개수
bus_stvr_rut_co	0.104302	버스경유노선 개수

〈그림 2-21〉 변수상관도 분석 결과

– 독립변수 간에 강한 상관관계가 나타나는 경우 분석 결과가 편향될 수 있는 '다중공선성' 문제가 발생하며, 이를 해결하기 위해 변수 선택 및 가공 과정을 거친다. 변수들에 대해 VIF 점수를 이용하여 다중공선성을 확인할 수 있다. VIF가 10 이상인 변수들을 확인한 결과 대기수요자 수의 통곗값 변수들과 대중교통 인프라 변수들이 이에 해당된다. VIF가 높은 변수들의 상관도를 분석해 서로 종속되어 있는 변수인 요일별 시간대별 유동인구 평균값(POP_MAX), 한 시간 전 대기수요 자 수(Lag_1), 월(MONTH), 주거지 비율(Cd1), 지하철 경유노선 개수(Sbw_sttn_co), 버스정류소 개수(Bus_sttn_co) 속성을 제외하였다. 서로 종속된 변수들을 제거한 결과 다중공선성 문제가 해결됨을 알 수 있다.

VIF Factor	features		
154.198396	POP_MEAN	요일별 시간대별 유동인구 평균값	POP_MEAN과 POP_MAX는 서로 높은 상관성을 가지므로 POP_MAX변수를 제외
152.428857	POP_MAX	요일별 시간대별 유동인구 최댓값	
47.339944	MONTH	월	
25.662790	cd1	주거지비율	
18.516692	lag_1	한시간 전 대기수요 자 수	한시간 전, 하루전 같은시간의 대기수요자 수 사이에 높은 상관성을 가지므로 lag_1변수 제외
17.649048	lag_1D	하루 전 같은 시간 대기수요자 수	
17.587354	bus_stvr_rut_co	버스경유노선개수	버스와 지하철 등 대중교통 인프라 변수들이 서로 종속되므로 버스 지하철 경유노선개수만 사용
16.388166	bus_sttn_co	버스정류소개수	
13.319524	sbw_stvr_rut_co	지하철경유노선개수	
13.167865	tot_main_bldg_cnt	주건물수	
11.989911	hd_val	습도값	
10.343482	sbw_sttn_co	지하철역개수	

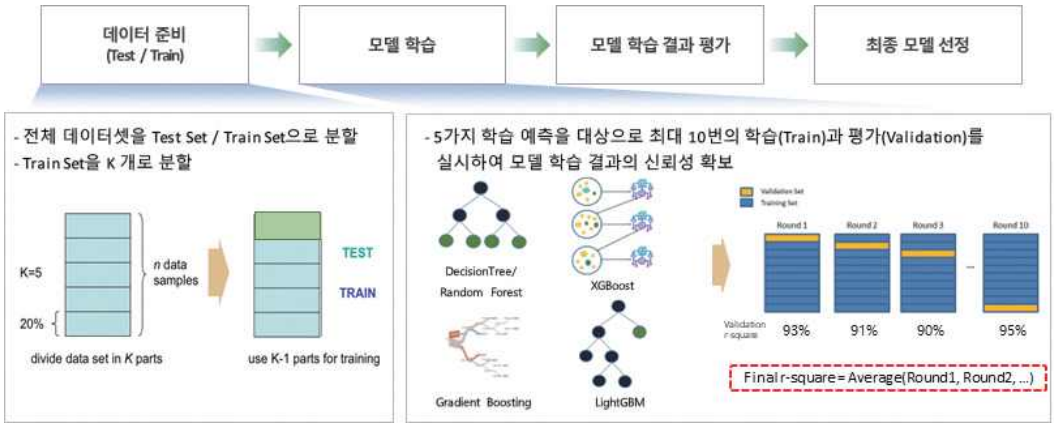
〈그림 2-22〉 다중공선성 분석

4-3 알고리즘 구현

알고리즘 평가 기준 및 수행 방법

– R-square(R)는 회귀모델의 정확도를 평가하는 방법으로 실제 모델의 변동량 중에서 근사모델로 설명 가능한 부분의 비율을 나타내는 값이다. R-square를 활용하여 모델 예측값과 실제값의 차이를 보여주는 방식으로 예측 분석모델의 정확도를 확인할 수 있다. 이 방법은 1에 가까운 값을 가질수록 정확한 근사 모델이다. K-Fold Cross Validation(교차검증)을 활용하여 예측 분석모델이 여러 번의 모델을 학습하여 결과를 도출하고, 산출된 각 모델의 평가 점수를 종합한다. 그리고 추가로 Test세트

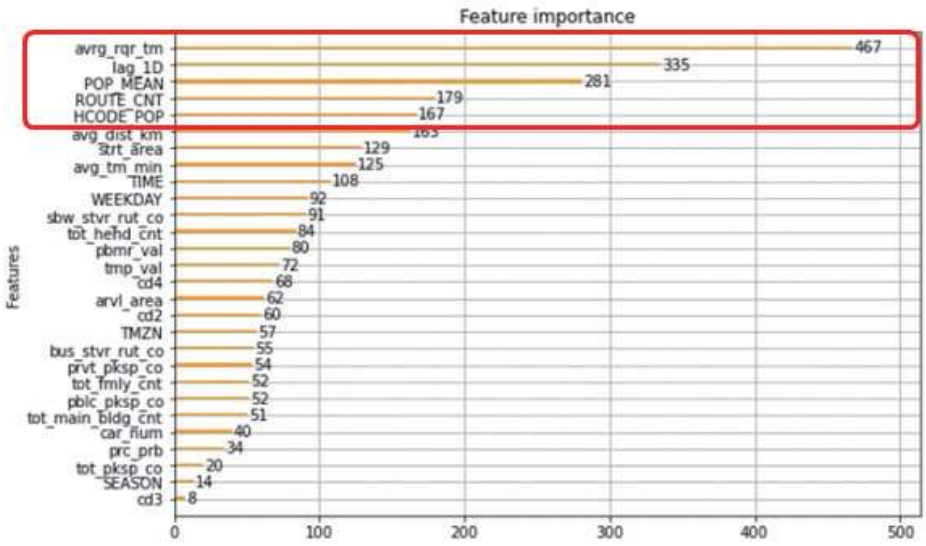
를 사용하여 점수를 도출한 뒤, 모델별로 비교하여 최적의 모델을 선정하는 방식으로 정확도를 높였다.



〈그림 2-23〉 알고리즘 평가 및 수행 방안

대기수요지수 예측 모델 구현

– 예측 변수 중요도를 분석하기 위해 트리 기반 모델이 제공하는 plot_importance 내장함수를 이용했다. 이 함수를 통해 관광객 수에 많은 영향을 끼치는 변수를 찾을 수 있다. 특정 Feature가 트리를 분할하는 데 얼마나 기여했는지, 엔트로피와 지니 계수의 변화량에 따라 중요도가 결정된다. 하지만 이 방법은 노드가 분기할 때의 이러한 계수만을 고려해 중요도를 부여하기 때문에 과적합에 대해 고려하지 못하는 단점이 있다. 그래프의 값이 큰 변수들은 대기수요지수에 대한 설명력이 높음을 나타낸다. 대기수요지수에 대해 영향도 기준 상위 5개 변수는 대중교통 평균 소요시간(Avrg_rqr_tm), 하루 전 대기수요자 수(Lag_1D), 요일별 시간대별 유동인구 평균값(POP_MEAN), 버스노선개수(ROUTE_CNT), 도착지 유입인구 평균값(HCODE_POP)으로 결과가 나왔다.



〈그림 2-24〉 대기수요지수- 변수 간 변수중요도

- 예측 모델의 대한 평가는 K-Fold Cross Validation을 활용해 학습용 데이터셋을 5개로 분할하여 모델별로 5번의 성능 평가 점수를 산출하였다. 모델별 점수 평균값을 통해 종합점수를 계산하는 방식 사용하였다. 보다 정확한 비교를 위해 추가적으로 테스트용 데이터셋을 이용하여 모델별로 테스트 평가점수를 산출하였다. 결과를 종합해 보니, 테스트 점수와 교차검증 종합점수가 가장 높고 표준편차가 낮아 모델의 신뢰성이 보장할 수 있는 LightGBM 모델을 최종 모델로 선정하였다.

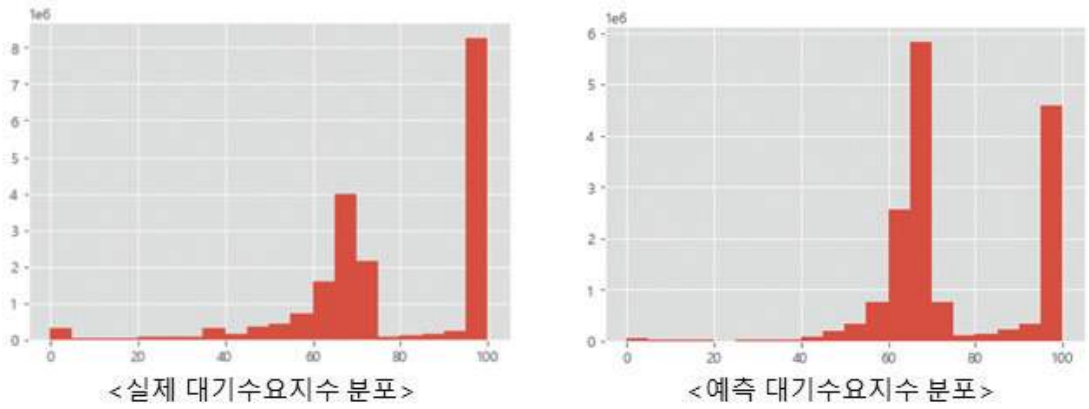
예측 모델 / 학습데이터셋	DecisionTree	LightGBM	Gradient Boost	Random Forest	XGBoost
CV 평균점수	0.238	0.643	0.624	0.633	0.631
CV 표준편차	0.019	0.007	0.008	0.007	0.007
Test평가점수	0.145	0.634	0.61	0.614	0.62

최종모델 채택

〈그림 2-25〉 대기수요지수 예측 모델 평가 결과

대기수요지수 예측 모델 구현 결과

– 모델링을 통해 산출된 결과를 토대로 대기수요지수 예측 결과 테이블을 구성하고 검증용 데이터 기준 대기수요지수 통계정보를 도출하였다. 일자별, 시간대별 각각의 이동에 대해 예측 대기수요지수와 함께 대중교통 편의와 관련된 변수들을 포함해 향후 시뮬레이션에 활용할 수 있도록 구성하였다. 버스노선이 없는 이동과 버스가 운행하지 않는 새벽 시간에 대해서는 대기수요지수가 100이며 이외의 대기수요지수는 60~70 사이에 집중되어 있다. 대기수요지수 예측지를 가지고 이동 출발지에 따른 평균 대기수요지수 현황을 시각화하였다. 각각의 행정동 단위의 대기수요지수 또한 화면상에 표현하였다. 출발지나 도착지가 달성군인 경우 대기수요지수가 높은 것으로 보아 대중교통이 발달되지 않아 유동인구에 비해 대중교통 이용자가 적은 것으로 추정된다. 대기수요지수가 낮은 경로는 대중교통 이용률이 높은 중구에서 출발하는 사례들이며, 이는 지하철 환승역과 버스 환승 정류장이 많아 대중교통 이용이 활발하기 때문으로 보인다.



〈그림 2-26〉 실제 대기수요지수와 예측 대기수요지수의 분포 비교

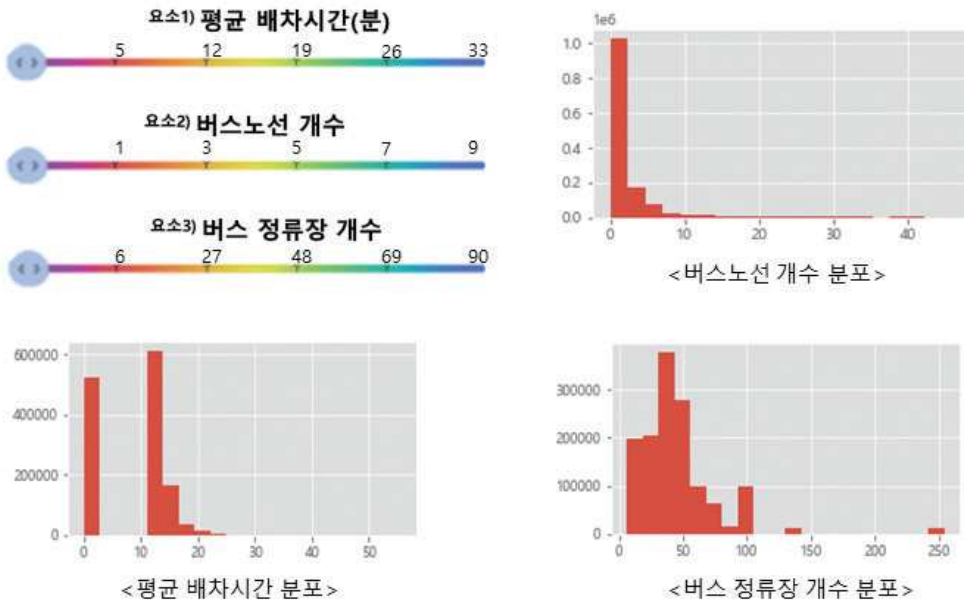
4-4 활용방안

시뮬레이션 분석

– 예측 대기수요지수를 파악하고, 평균 배차시간, 버스노선 개수, 버스 정류장 개수의 값에 따라 예측 대기수요지수의 변화를 보여주는 것으로 시뮬레이션이 기획됐다. 이동구간별 평균 배차시간, 버스노선 개수, 버스 정류장 개수, 이 3가지 요소에 따른 대기수요지수 변화를 보여주는 것으로 시뮬레이션을 할 수 있다. 3가지 시뮬레이션 요

소는 5구간의 값을 가지며 각각의 값에 따른 대기수요지수가 표현된다.

- 5가지 구간으로 미리 정한 것은 화면 표출상 정해지지 않는 값의 범위로 했을 경우 변수가 너무 많기 때문이다. 3가지 요소값을 조정하여 최소의 대기수요지수를 가지는 지점을 참고 현황 값으로 비교할 수 있다.
- 평균 배차 시간이 0인 데이터는 버스노선이 없는 이동구간이다. 배차시간이 5 이하였다가 5 이상으로 되면 대기수요지수가 감소하는 것이다. 대부분 버스 배차시간은 10분에서 20분 이하 범위를 가짐을 확인할 수 있다. 원래 버스노선이 있는 경우 평균 배차 시간이 12분을 초과하면 대기수요지수가 급격하게 증가한다. 이것으로 보아 배차 시간이 대중교통을 이용하는 데 있어 중요한 요소라고 할 수 있다. 한편 버스노선 개수의 분포가 오른쪽으로 긴꼬리 모양인 것으로 보아 노선의 불균형이 있는 것으로 확인된다.
- 버스노선이 없는 이동구간이 많으며, 특정 지역에 몰려있는 것으로 보인다. 버스노선이 없는 이동구간이 많으므로 결정트리 모형을 보면 버스노선의 작은 변동도 대기수요지수에 큰 변화를 줄 수 있다. 버스 정류장 개수의 평균은 45개이고, 중앙값은 39개이다. 대부분 100 이하의 값을 가지는 것을 알 수 있다. 버스정류장 개수는 다른 요소들과 달리 대기수요지수에 대한 영향이 낮고, 버스정류소 개수의 증가가 대기수요지수 값의 감소를 야기하지는 않는다.



〈그림 2-27〉 대기수요지수 각 요소에 대한 분포

대기수요지수 활용방안

- 버스노선 개수에 따라 예측 대기수요지수를 파악할 수 있는 시뮬레이션을 구현했다. 이러한 시뮬레이션 방법을 통해 각 요소인 배차시간, 버스노선 개수, 버스 정류장 개수 조정을 통해 대기수요가 얼마나 증가하거나 감소하는지 예측할 수 있다. 예를 들어, 대기수요지수 시뮬레이션을 통해 원래 예측 대기수요지수의 값이 90 이상이었던 데이터들이 버스노선의 증가에 따라 예측 대기수요지수가 80점대, 90점 초반으로 떨어짐을 확인할 수 있다.
- 구/행정동 단위 출발지에서 도착지 이동 경로에 해당하는 유동인구, 대중교통 이용률, 평균 이동시간을 확인할 수 있어, 도시 정책 결정자들이 시민들의 대중교통 이용에 대해 시각화된 화면으로 한눈에 파악할 수 있다. 또한 이동 코스 선택 시 해당 코드에 관한 시간대별 상세정보인 이동시간, 유동인구수 등을 확인할 수도 있다.
- 구/행정동 단위 출발지에서 도착지 이동 경로에 해당하는 대기수요지수, 효율지수를 확인할 수 있다. 역시 이동 코스 선택 시 해당 코스에 관한 시간대별 상세정보를 확인할 수 있다. 데이터의 Insight 도출을 위한 이슈 코스 정보인 대기수요지수 또는 효율지수가 높은 경로를 제공하여 보고한다.

제3장 실증 경과

1 | 실증체계

○ 버스 승/하차, 유동인구 등 데이터를 활용하여 버스 이용률, 이용 시간 등 노선 효율을 분석하고 의사 결정자에게 이동시간 최적노선, 효율적인 배차 간격과 같은 다양한 솔루션 제시를 위한 서비스를 의미한다.

○ 해당 서비스의 대표 핵심성과지표(KPI)는 융복합(버스노선 최적화) 표출 서비스 만족도 80% 이상으로 삼으며, 융복합(버스노선 최적화) 분석보고서로 결과 보고한다.

○ 실증을 위한 시나리오 구성은 구축된 서비스를 기반으로 정책결정자가 서비스를 이용하는 상황을 가정하며, 아래와 같이 진행한다.

- 1. 정책결정자가 적용지역을 대상으로 출발지, 도착지 권역(구별, 행정동별, 버스 권역)을 설정한다.
- 2. 설정한 조건에 따른 대중교통 이용률, 이동시간, 유동인구 분석 결과 및 최적화 방안을 제시하며, 상세한 최적화 방안으로는 버스 이동 거리 분석 기반 버스노선 재설정 및 최적화 진행, 버스 이용률 분석 기반 중복노선 통/폐합 및 배차 간격 재조정 제안, 대중교통 접근성 분석 기반 정거장 조정의 가이드를 제안한다.
- 3. 지도 위의 분석 결과를 기반으로 화면이 표출되며, 현황과 개선안을 시각화하여 표현한다.
- 4. 조건별(시간별, 지역별) 선택적 결과를 조회할 수 있으며, 해당 결과의 분석데이터를 생성하여 추출하는 기능을 제공한다.

2 | 실증대상

○ 실증대상 및 실증내용으로는 대구광역시 전역을 실증지역으로 설정하여 하위 행정동별 대중교통 이용률을 분석하고 버스노선별 효율지수를 제시한다. 또 수집 및 활용 데이터는 다섯 가지 측면으로 분류해 진행한다.

〈표 3-1〉 수집 활용 데이터 분류

인구통계 측면	유동인구, 행정동 단위 유입지의 시간대별 유입인구 데이터
교통 측면	대중교통인프라, 주차장 인프라, 버스노선, 자가용 유입 데이터
금융상권 측면	교통카드 승하차 정보, 일별/시간대별 카드 매출데이터
공간 특성 정보 측면	토지(상업, 주거, 공업, 기타 지역 비율), 건축(세대수, 가구 수, 건축 수) 데이터
환경 측면	날씨(강수확률, 습도, 풍속, 강수량, 기온 값) 데이터

– 수집 및 활용 데이터 기반으로 대중교통 및 인프라 상황을 고려한 대중교통 이용률 분석, 대중교통 대기수요 지수 SW 개발, 버스노선별 효율지수 등급을 제시할 수 있는 분석 방안을 도출한다. 이를 통해 대구광역시 전역의 대중교통을 대상으로 한 대구광역시 공무원 서비스로 실증 및 활용을 진행한다.

3 | 실증 경과

– 현재까지의 주요 성과 및 실증결과는 데이터 수집 및 분석, 서비스 구축, 설명회 진행, 대외 시연 등으로 실증을 진행하였다. 데이터 수집 및 분석 실증으로는 교통 플랫폼 및 유동인구, Tmap LH 공간 특성 정보 데이터 수집 및 분석을 진행하였다. 또한 버스노선 최적화 알고리즘 개발 및 표출 서비스의 구축을 완료하였고, 대구광역시 공무원(교통정책과, 스마트정책과 등 참석) 대상으로 버스노선 최적화 설명회를 진행하였다. 대외적으로는 안양시와 수자원공사 대상의 버스노선 최적화 서비스 시연을 완료하였다.

4 | 실증 결과

– 최종적으로 버스노선 최적화 서비스 구축을 완료하여 버스노선 최적화 알고리즘 기획 및 설계, 데이터허브에서의 융복합 알고리즘 개발 및 예측 환경 구축, 버스노선 최적화 표출 서버로 데이터 전송환경을 구축하였다. 그리고 지자체와 민간 데이터 수집을 통해 버스노선 분석을 진행하여 데이터허브에 수집된 융복합 데이터를 활용하여 대중교통 이용률, 대기수요 지수, 버스노선 효율지수를 도출할 수 있었다. 또한 다양한 웹 환경 및 모바일 환경에 최적화된 UI/UX 디자인 및 화면을 개발하였다. 이를 통해 대기수요 지수 기반의 시뮬레이션을 진행하여 대기수요 지수가 높은 주요 이슈 지역의 버스노선/정류소 및 배차 시간의 최적화를 진행할 수 있고, 대중교통 대기수요 예측을 통해 대기수요 지수가 높은 행정동, 시간대에 버스노선을 집중하여 버스 운영의 효율성을 증대할 예정이다.

제4장 확산 방안

1 | 대구광역시 추가 개발 사항

1-1 버스노선 효율 분석

◔ 데이터 수집 및 데이터 전처리

– SKT와 LH, DGB유페이, 대구광역시, 빅데이터플랫폼 등에서 효율지수를 산출하는 데 필요한 인구, 버스, 지하철, 공간 특성 정보, 교통카드 원천 데이터를 수집한다. 원천 데이터는 Data Lake에 적재되어 통계 테이블 및 효율지수 산출, 효율지수 예측모델에 활용된다.

〈표 4-1〉 효율지수 설계를 위한 데이터 수집 내역

구분	데이터	활용 컬럼	출처
인구	주거인구	행정동별 총 인구 수	SKT
버스	버스노선목록	노선별 총 운행시간, 총 운행거리	대구광역시
	버스노선경유정류소	노선별 경유 정류소	대구광역시
지하철	도시철도역사	지하철역 위치	대구광역시
공간특성	토지	주거, 상업, 공업지 면적	LH
	문화시설	문화시설 위치	빅데이터플랫폼
	도시공원	도시공원 위치	빅데이터플랫폼
	관공서	관공서 위치	빅데이터플랫폼
	체육시설	체육시설 위치	빅데이터플랫폼
	전통시장	전통시장 위치	빅데이터플랫폼
	고등학교	고등학교 위치	공공
	의료시설	의료시설 위치	공공
	대규모 점포	대규모 점포 위치	공공
금융	교통카드	승하차 태그 이력	대구광역시

○ 분석 및 통계 테이블 생성

– 구간별 효율지수 산출을 위해 구간 테이블을 생성하여 기준 구간을 만들고, 교통카드 데이터에서 구간별 승차인원과 속도, 소요시간을 산출한다.
– 서비스 운영 시 운영 모듈의 시간 단축 및 통계치 사용을 위해 행정동 단위 통계 테이블과 버스정류장 단위의 통계 테이블을 생성한다.

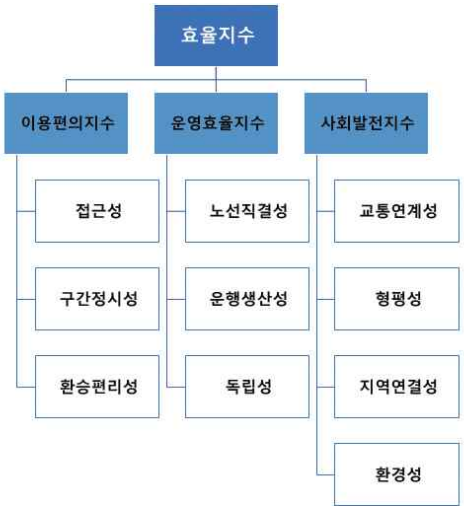
○ 데이터 전처리

– 구간 소요시간 결측값 채우기 : 교통카드 데이터를 기반으로 구간별 효율지수 산출 시 '정류소–다음 정류소' 단위로 소요시간이 추출되어야 한다. 그 과정에서 생긴 결측값을 보완하기 위해 겹치는 구간을 이용하여 최소 구간의 소요시간을 계산한다. 예를 들어, A 정류소에서 승차하여 D 정류소에 하차한 교통카드 데이터를 기반으로 A–B, B–C, C–D 구간별 소요시간을 각각 산출하기 위해 A–C 이동 교통카드 데이터를 이용하여 C–D 구간의 소요시간을 산출한다.
– 버스정류장 인근 주요시설 및 지하철역 개수 추출 : 버스정류장의 위치 좌표와 문화시설 및 도시공원, 관공서, 대규모 점포 등 주요 시설의 위치 좌표를 이용하여 버스정류장 인근 300m(약 도보 5분) 안에 위치한 주요시설의 개수 및 지하철역 개수를 추출한다. 지리 정보 데이터의 처리 및 기하학적 연산을 돕는 Geopandas와 shapely라는 패키지를 활용한다.
– 신규 버스정류장 행정동 정보 추출 : 버스노선경유정류소 데이터에서 업데이트되는 신규 정류소에 대해 행정동 경계 데이터를 이용하여 정류소가 위치한 행정동을 추출한다. Geopandas를 이용해 버스정류장의 위치 좌표(Point)가 어떤 행정동(Polygon)에 속해 있는지를 알 수 있다.

○ 효율지수 설계

– 노선의 효율지수는 이용자, 운영자, 사회적 관점을 모두 반영할 수 있도록 설계되어 종합적인 관점에서 노선의 효율지수를 평가할 수 있다.
– 각 지표의 중요도에 따라 가중치가 부여되며, 각 지표 점수와 가중치를 곱한 후 합하여 최종 효율지수를 산출한다.
– 이용자 관점의 지표 점수를 합해 '이용편의지수'로, 운영자 관점의 지표 점수를 합해 '운영효율지수'로, 사회적 관점의 지표 점수를 합해 '사회발전지수'로 통합하여 표현함

으로써 직관적인 이해를 돕는다.



〈그림 4-1〉 효율지수 구성

〈그림 4-2〉 이용편의지수 지표 설명

– (1) 이용편의지수 : 이용자의 편의를 나타낼 수 있는 지표 점수들의 합으로, 버스정류장이 가까운지 나타내는 접근성과 정해진 시간에 맞춰 버스가 운영되는지 나타내는 정시성, 환승이 많이 이루어지는지 나타내는 환승편리성으로 구성된다.

– 접근성 : 노선별 출발지에서 정류장까지 얼마나 가까운지를 나타내는 지표로 행정동의 버스정류장 공급도 기반으로 점수가 산출된다. 시가화면적은 토지 용도지역 중 주거지, 공업지, 상업를 합한 면적을 말한다. 정류장공급도가 클수록 버스정류장이 많이 설치되어 대중교통 접근성이 좋은 지역을 의미한다.

$$\text{접근성} = \frac{1}{n}\sum_{i=1}^{n}\frac{S_i}{A_i}$$

A_i : 행정동 i의 시가화면적 (km^2)
S_i : 행정동 i의 버스정류장수(개)
n : 노선을 지나는 행정동 수

– 구간정시성 : 구간별 통행시간에 맞게 버스가 운행되는지 나타내는 지표로 구간의 평균 통행시간과 변동성이 얼마나 있는지를 표현하는 구간 소요시간의 변동계수를 이용한다. 1년 운행 소요시간 평균 대비 표준편차로 구간의 변동계수를 구해 변동성을 나타내며, 변동성이 클수록 정시성이 감소한다. 운행 시마다 변동 없이 기준 소요시간에 맞추어 운행될 경우 정시성은 1의 값을 가진다.

$$\text{구간정시성} = \frac{1}{\text{구간변동성}+1}, \quad \text{구간변동성} = \frac{\sqrt{\frac{\sum(t-\bar{t})}{n}}}{\bar{t}}$$

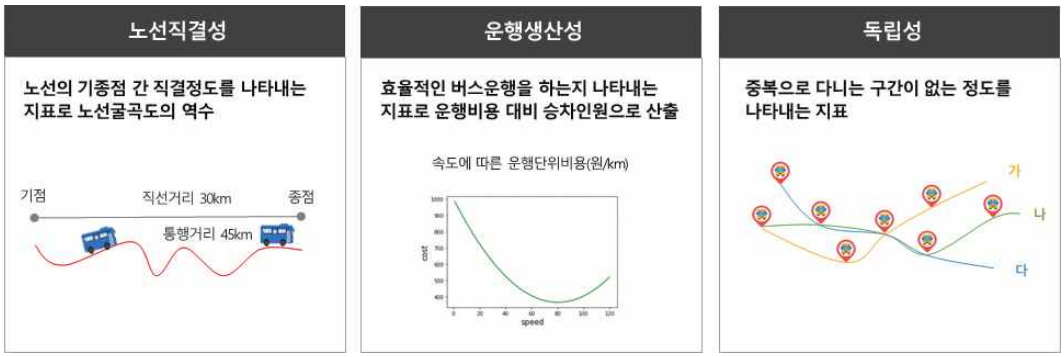
t : 소요시간(초)
$\bar{t}$: 소요시간평균(초)

– 환승편리성 : 노선에서 환승이 얼마나 활발하게 이루어지는지 나타내는 지표로, 총 탑승객 수 대비 환승탑승객 수로 표현한다. 환승이 많이 이루어지는 노선이 곧 환승이 편리한 노선이며, 버스 이용자 입장에서 타 노선 및 지하철로 이동 선택의 폭이 크다는 것을 의미한다.

$$\text{환승편리성} = \frac{TP}{P}$$

TP : 환승인원(명)
P : 승차인원(명)

– (2) 운영효율지수 : 운영의 효율성을 나타내는 지표 점수들의 합으로, 노선 굴곡도의 역수인 노선직결성과 운행비용 대비 승차인원으로 버스 운행의 효율성을 나타내는 운행생산성, 노선이 다른 노선과 중복되지 않는 정도를 나타내는 독립성으로 구성된다.



〈그림 4-3〉 운영효율지수 지표 설명

– 직결성 : 노선의 기종점 간 직결 정도를 나타내는 지표로 노선굴곡도의 역수로 표현한다. 노선굴곡도는 노선의 굴곡 정도를 말하며, 노선의 통행거리와 직선거리(기점-회차지)의 비율이다. 굴곡도가 1에 가까울수록 통행시간과 비용 등에 최적화된 노선으로 효율지수 산출 시 정(+)의 상관성을 갖기 위해 노선굴곡도의 역수로 노선직결성을 산출한다. 따라서, 직결도가 1에 가까울수록 통행시간, 노선거리, 통행비용 등에 최적화된 노선이며 굴곡이 없는 노선의 경우 최댓값인 1을 갖게 된다.

$$노선직결성 = \frac{1}{노선굴곡도}, \quad 노선굴국도 = \frac{D}{L}$$

$$D : 직선거리(km)$$
$$L : 운행거리(km)$$

– 운행생산성 : 효율적인 버스 운행에 대한 지표로, 운행비용 대비 승차인원으로 표현한다. 구간 운행 속도에 따라 운행비용원단위가 결정되며 속도가 높을수록 비용이 감소하지만 속도가 80km/h를 넘어가면 운행비용원단위가 증가한다. 운행비용원단위는 '교통시설투자평가지침'에 나와 있는 속도 10km/h 당 대형버스의 차량운행비

용 표를 바탕으로 속도 1km/h 당 비용을 산출할 수 있도록 회귀식을 만들어 적용한다. 구간운행거리에 구간운행속도에 따른 운행비용원단위를 곱해 구간운행비용이 산출되고 구간운행비용 대비 승차인원으로 구간운행생산성을 산출한다.

$$\text{운행생산성} = \frac{P_k}{OC_k}, \quad OC_k = d_k \times Operation\ Cost\ Unit\ per\ Speed(\text{원}/km)$$

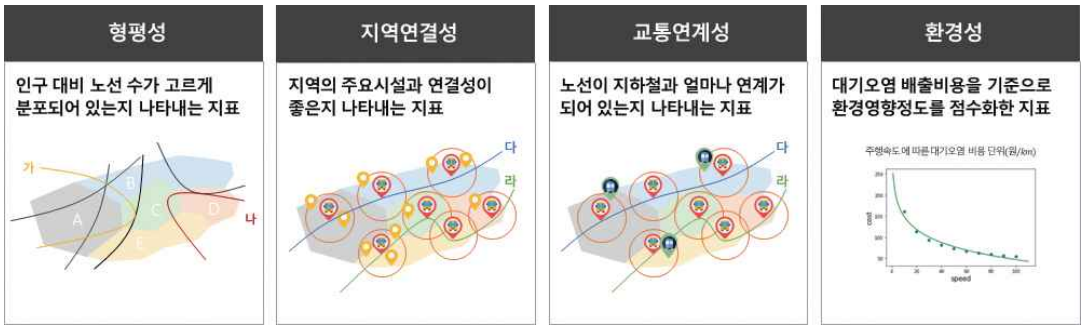
P_k : 구간 k의 승차인원(명)
OC_k : 구간 k의 차량운행비용(원)

– 독립성 : 중복으로 다니는 구간이 없는 정도를 나타내는 지표로, 버스노선이 특정 구간에 집중되는 정도를 나타내는 노선중복도의 역수이다. 독립성이 1이라는 것은 노선의 기점에서 종점까지 전체 구간 중 노선이 중복된 구간이 하나도 없다는 것을 의미하며, 중복 정도가 심할수록 독립성 점수가 감소함.

$$\text{노선독립성} = \frac{1}{\text{노선중복도}}, \quad \text{노선중복도} = \frac{\sum_{k=1}^{n}(d_k \times l_k)}{\sum_{k=1}^{n} d_k}$$

d_k : 구간 k의 길이(km)
l_k : 구간 k의 경유하는 노선 수(개)
n : 전체 구간 수

– (3) 사회발전지수 : 사회적 관점을 고려한 지표 점수들의 합으로, 인구 대비 노선이 적절하게 공급되는지 나타내는 형평성과 주요 시설과 연결되어 있는지 나타내는 지역연결성, 지하철역과 연계 정도를 나타내는 교통연계성, 대기오염비용을 기반으로 환경영향 정도를 나타내는 환경성으로 구성된다.



〈그림 4-4〉 사회발전지수 지표 설명

- 형평성 : 노선을 지나는 지역 인구 대비 노선 수가 고르게 분포되어 있는지를 나타내는 지표로 노선 보급률을 기반으로 노선 보급의 평균대비 차이가 얼마나 나타나는지 알 수 있다. 노선 보급률은 인구 대비 노선 수로 표현되며, 행정동별 노선보급률의 표준편차의 합으로 노선의 편향성이 산출된다. 편향성이 0일수록 인구대비 적절한 노선 수가 지나가고 있음을 의미하며, 높을수록 인구대비 노선 수가 과다하게 혹은 적게 분포하고 있음을 의미한다.

$$\text{노선형평성} = \frac{1}{\text{노선편향성}+1}, \quad \text{노선편향성} = \sqrt{\frac{\sum_{i=1}^{n}(RP_i - \overline{RP})^2}{n}}, \quad RP_i = \frac{\text{경유노선수(개)}}{\text{주거인구(천명)}}$$

RP_i : 행정동 i의 노선보급률(개/천명)
$\overline{RP}$: 전체 행정동의 노선보급률 평균
n : 노선을 지나는 행정동 수

- 지역연결성 : 노선이 지역의 주요시설과 가까운지 나타내는 지표로, 노선의 전체 버스정류장 수 대비 버스정류소의 인근 주요시설 수의 합으로 표현한다. 주요시설은 주요 문화시설, 교육시설, 의료시설, 학교, 관공서, 대형마트, 전통시장, 체육시설을 의미하며 약 도보 5분 거리 기준으로, 버스정류장 기준 300m 반경 내에 위치한 주요시설의 개수를 합한다. 지역연결성이 높을수록 노선이 지역의 주요 시설과의 근접성이 좋음을 의미한다.

$$지역연결성 = \frac{\sum_{j=1}^{S} F_j}{S}$$

F_j : 정류소 j의 인근 주요시설 수(개)
S : 전체 정류소 수(개)

– 교통연계성 : 노선이 지하철과 연계가 잘 되어 있는지 나타내는 지표로, 노선의 전체 버스정류소 수 대비 버스정류소 인근 지하철역 수의 합으로 표현한다. 인근 지하철역은 약 도보 5분 거리인 버스정류소 기준 반경 300m 안에 위치한 지하철을 의미한다. 교통연계성이 높은 노선은 지하철과 버스 간 환승이 용이하여 지하철 연계 정도가 높은 노선을 의미한다.

$$교통연계성 = \frac{\sum_{j=1}^{S} SB_j}{S}$$

SB_j : 정류소 j의 인근 지하철역 수(개)
S : 전체 정류소 수(개)

– 환경성 : 노선의 운행에 따라 배출하는 대기오염비용을 근거로 환경영향정도를 나타낸 지표이다. 10km/h 속도에 따라 달라지는 환경오염비용을 기준으로 1km/h 속도에 따른 환경오염비용 원단위 식을 생성하여 구간 주행속도의 km 당 대기오염비용에 구간 길이를 곱해 구간의 환경성 점수를 산출한다. 오염비용이 낮을수록 환경성이 높아지도록 환경오염비용의 역수로 환경성을 나타낸다.

$$환경성 = \frac{1}{PC_k}, \quad PC_k = d_k \times Air\ Pollution\ Cost\ Unit\ per\ speed(원/km)$$

d : 구간 k의 거리(km)
PC : 구간 k의 환경오염비용(원)

효율지수 예측 알고리즘

- 효율지수 예측은 일자별, 노선별, 시간별, 상행하행유형별 정류소의 승차인원을 예측한 후, 효율지수 산출식에 따라 다가올 미래의 효율지수를 산출한다.
- 유동인구, 교통카드, 공간 특성 및 자동차, 날씨 등 이종 데이터를 융복합하여 승차인원 예측 모델의 독립변수로 사용한다.
- 수집한 데이터셋을 기반으로 승차인원 예측에 영향을 줄 수 있는 통계값 및 시간변수 등을 파생변수로 생성한다.
- 시차 파생변수 : 시간의 흐름에 따라 승차인원이 변화하며 이전 승차인원 정보가 미래의 승차인원 예측에 영향을 줄 것으로 예상되어 전일 승차인원과 전주 승차인원, 전월의 승차인원 데이터를 파생변수로 생성한다.
- 시간 파생변수 : 교통카드의 일자 데이터 기반으로 월, 요일, 주말 여부, 공휴일 여부에 따라 승차인원에 차이가 있을 것으로 예상되어 승차인원과의 상관분석 후 독립변수로 사용한다.

○ 승차인원 예측모델 구현

- 2019년 교통카드 데이터 기반 분석 데이터셋에서 2019년 1월부터 9월까지 9개월의 데이터를 Train Data Set으로 10월부터 12월까지 3개월의 데이터를 Test Data Set으로 분할한다.
- 5가지 예측 알고리즘 DecisionTree, LightGBM, GradientBoost, RandomForest, XGBoost을 사용하여 승차인원 예측 모델을 생성한다.

○ 예측모델 평가

- 모델에 대한 성능 평가는 K-Fold Cross Validation(K겹 교차검증)을 활용해 학습용 데이터셋을 5개로 분할한 뒤 나온 Cross Validation 평가점수 평균값과 시험용 데이터셋으로 예측하여 산출된 Test 평가점수를 종합적으로 판단하여 최종 모델이 선정된다.

〈표 4-2〉 효율지수 설계를 위한 데이터 수집 내역

구분	데이터	활용 변수	출처
시간		월, 요일, 주말 여부, 공휴일 여부	
금융	교통카드	전일/전주/전월의 승차 수, 환승 승차 수, 구간 소요시간, 구간 속력	DGB페이
버스	버스노선 목록	노선별 총 운행시간, 총 운행거리	대구광역시
	버스노선 경유 정류소	노선별 경유 정류소	대구광역시
지하철	도시철도역사	인근 지하철역 개수	대구광역시
공간특성	토지	주거, 상업, 공업지 면적, 비율	LH
	문화시설	인근 문화시설 개수	빅데이터플랫폼
	도시공원	인근 도시공원 개수	빅데이터플랫폼
	관공서	인근 관공서 개수	빅데이터플랫폼
	체육시설	인근 체육시설 개수	빅데이터플랫폼
	전통시장	인근 전통시장 개수	빅데이터플랫폼
	고등학교	인근 고등학교 개수	공공
	의료시설	인근 병원 개수	공공
	대규모점포	인근 대규모점포 개수	공공
자동차	주차장	공공/민간/총 주차면수	대구광역시
	내비게이션	이동 차량 수	TMAP
인구	pcell 유동인구	정류소 pcell의 유동인구수	SKT
	행정동 유입인구	정류소 행정동의 평균 유입인구수	SKT
	주민등록인구	총 인구수, 남자/여자 인구수	공공
환경	날씨	기온값, 풍속값, 강수량	기상청

1-2 배차 간격 최적화

◔ 최적 배차 간격 설정 알고리즘 개발

– 총 교통비용 최소화하는 모형으로 요일별 버스 배차간격으로 최적의 배차간격을 구한다. 총 교통비용은 버스운행 비용과 전체 승객의 대기시간 비용, 승객 통행시간 비용을 합해 산출하여 운행생산성과 이용자의 편의를 고려하여 배차 간격이 설정된다.

– 버스운행비용은 시간당 운행대수에 대당 차량운행비를 곱해 시간별 노선의 버스운

행비용이 산출된다. 시간당 운행대수는 운행시간에서 배차 간격을 나눠 계산된다. 차량운행비는 '교통시설투자평가지침'의 차종별 속도별 차량운행비용을 참고하여 대형버스의 속도에 따른 차량운행비용식을 만들어 적용한다. 차량운행비용의 세부 구성은 속도별 유류 소모량과 엔진오일비, 타이어마모비, 유지관리비, 감가상각비로 구성된다.

- 승객 대기시간 비용은 승객의 대기시간가치(원/시-인)에 노선당 총 탑승객수(인)와 평균 대기시간의 곱으로 표현한다. 승객의 대기시간은 버스정류장에 도착 후 버스에 탑승하기까지의 시간을 말하며, 승객의 버스정류장 도착 시간에 대한 정보를 알 수 없어 배차 간격의 1/2로 평균 대기시간을 산정한다. 승객의 대기시간 가치는 통행시간 가치의 1.62배로 인당 5,011원의 1.62배인 8,117.82원으로 계산한다.
- 승객 통행시간 비용은 총 탑승객 수에 승객의 평균 통행시간과 통행시간가치를 곱해 산출한다. 승객의 평균 통행시간은 교통카드 태그 이력 데이터를 활용하여 요일별 노선의 통행시간 평균값을 이용한다. 통행시간 가치는 '교통시설투자평가지침'을 참고하여 인당 5,011원을 적용한다.
- 배차 간격 최적화 알고리즘은 버스노선의 총 교통비용을 최소화하면서 대중교통에 대한 수요를 고려한 최적의 배차 간격에 대한 정보를 제공한다. 운영자의 최소한의 이윤과 승객의 수요를 만족시킬 수 있는 배차 간격을 제약조건으로 설정한다.

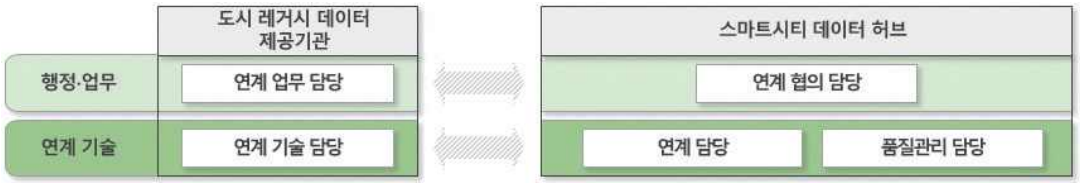
2 | 타 지자체 적용 시 확산 방안

2-1 데이터 확보 방안

- 타 지지체 확산 시 주요 데이터 수집이 가장 중요하다. 예를 들어, 교통 관련 데이터(버스 승하차, 버스예상도착이력 등), 유동인구, Tmap 데이터 등은 확보가 되어야 정상적인 분석이 가능하다. 해당 지자체에 교통관제센터와 협의해서 버스 승하차 및 버스예상도착 데이터를 연계할 수 있는지부터 진행되어야 한다. 데이터 확보가 완료되어야 그 후 분석 작업이 진행 가능하기 때문이다.
- 데이터가 database 형태의 정형 데이터라면 City Legacy I/F 시스템으로 연계가 가능하다. 우선 지자체와 연계 대상 시스템 및 대상 데이터 정의를 우선 진행한다.

이때 데이터의 활용도, 정보공개 가능 여부 등을 고려하여 연계 대상을 담당자와 협의한다.

– 대상 시스템 연계를 위한 인터페이스 표준을 정의한다. 송신기관, 수신기관 등 연동 정보를 정의하고 연동 유형에 따른 연계 방식을 설정한다. 연계 표준의 정의가 완료되면 연계 개발 및 테스트 작업을 수행한다. 이때 인터페이스 정의서 작성, 수집 테이블 설계, 테이블 생성을 수행한다. 다음으로 연계 프로그램 개발 및 연계 테스트를 하게 된다.

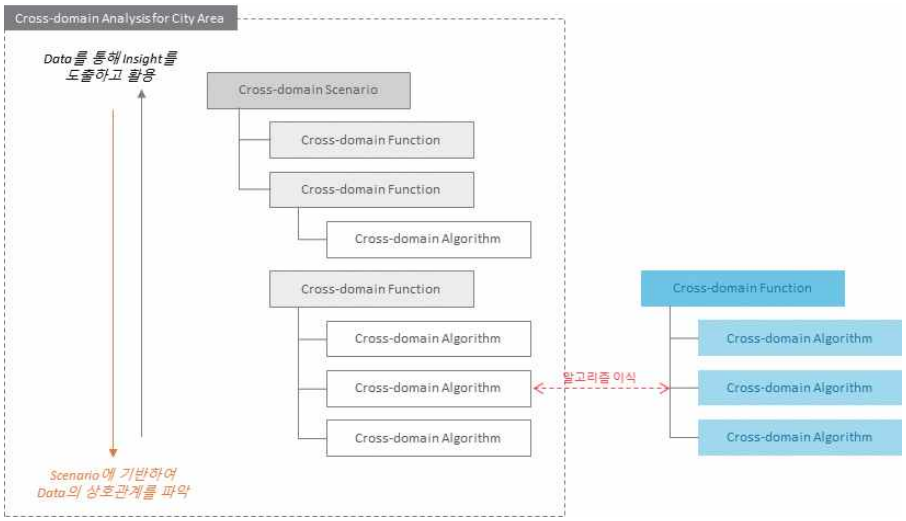


〈그림 4-5〉 데이터 확보 방안 (역할과 책임)

2-2 알고리즘 적용 방안

– 융복합 분석(Cross-domain Analysis) 알고리즘을 개별 지자체 단위가 아니라 전국적으로 제공하기 위해서는 분석 알고리즘의 표준 구조가 마련되어야 한다. 스마트시티의 융복합 분석이 일반적인 데이터 분석과 차별되는 핵심은 '도시 간 데이터의 공유와 연결'이라 할 수 있다. 스마트시티의 융복합 분석은 도시의 개별 서비스와 결합이 쉬워야 한다. 그리고 더 나아가 특정 도시의 모범사례를 다른 도시로 '서비스 적용'이 가능해야 한다. 이러한 조건을 만족시키기 위해 시나리오(Scenario), 기능(Function), 알고리즘(Algorithm)이라는 3단계 계층구조 만들었다.

– 이 내용을 더 설명하자면 하나의 시나리오에는 여러 개의 기능이 있을 수 있고 이 기능들은 독립적으로 관리된다. 그래서 다른 시나리오에서 활용이 가능한 체계이다. 예를 들어, '관광' 시나리오의 '유동인구예측' 기능을 만들었다면 그 기능을 '교통' 시나리오에서도 사용할 수 있다. 더 나아가 타 도시의 시나리오에도 활용할 수 있다는 점에서 '서비스 적용'이 쉽다고 할 수 있다. 이러한 3단계 계층구조는 알고리즘 모듈화를 통해 구현할 수 있다.



〈그림 4-6〉 스마트시티 융복합 알고리즘 구조

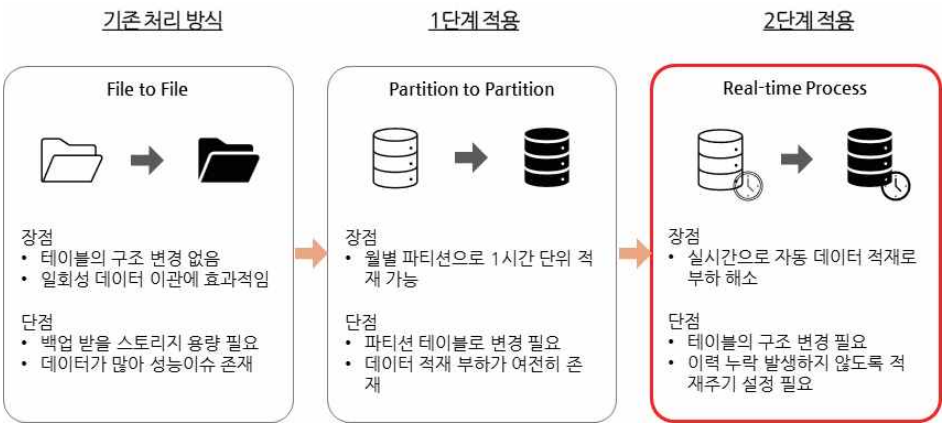
– 대구 데이터허브 시연, 안전2.0 워크숍 행사에서 타 도시의 실무자들이 다수 참석하여 의견을 주었다. 그중 많은 부분이 융복합 표출 화면의 분석 방법들에 대한 부분이다. 특정 도시에서 발생하는 문제들은 다른 도시에도 일어날 개연성이 아주 높다. 그래서 재활용성이 높은 분석 알고리즘의 표준 구조가 중요한 이유다. 동일한 도시 문제에 대해 타 도시에서 손쉽게 활용할 수 있도록 알고리즘 구조의 설계가 필요하다.

제5장 Lesson Learned

1 | 문제해결 사례

1-1 데이터 수집 사례

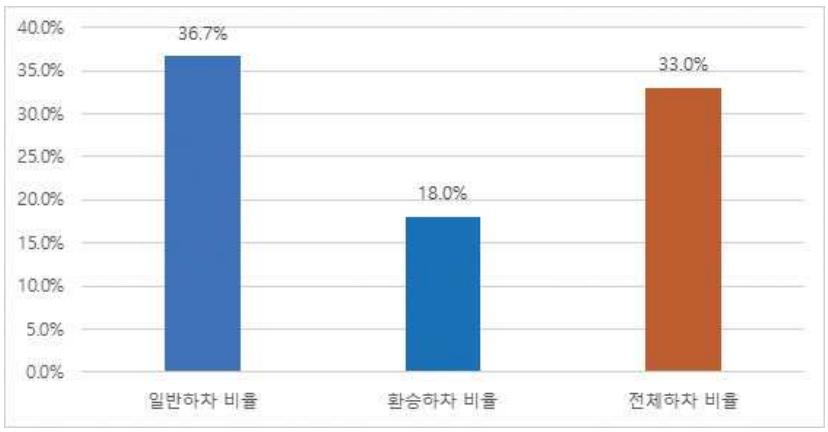
- 버스노선 최적화 서비스를 수행하면서 다양한 데이터의 한계를 보았다. 고객으로부터 데이터 제공이 어렵다는 통보를 받기도 하고, 데이터 수집 방식이 달라 추가 개발 이슈로 데이터를 받는 데까지 오랜 시간이 걸리기도 한다. 반면에 데이터 양이 너무 많아 수집이 어려울 거라 여겼지만 스마트한 처리 방법으로 활용 가능한 데이터로 바뀌기도 한다. 그렇게 수집한 데이터가 '버스노선최적화' 융복합 분석에서 중요하게 사용됐고, 서비스에 꼭 필요한 데이터였기에 무엇보다 의미 있는 경험이었다.
- 이 데이터는 너무 커서 레거시(Legacy) 시스템에서도 활용할 수 없는 상태였음을 담당자와 협의하면서 알았다. 아래 그림에서 알 수 있듯이 아이디어는 두 가지였다. 첫째, 빅(Big)데이터를 잘게 쪼개는 파티션(Partition) 기법을 적용하자. 둘째, 데이터를 실시간 처리함으로써 적재 부하를 최소화하자. 이 전략으로 데이터를 파티션화하여 '데이터 접근(Data Access) 부하'를 최소화하고, 실시간 처리를 이용해 '데이터 적재(Data Upload) 부하' 또한 낮춰 성능 이슈를 완전히 해결할 수 있었다. 이렇게 데이터 한계를 정확히 인지한다면 '데이터 양' 때문에 활용할 수 없었던 데이터도 스마트한 데이터로 변모할 수 있다.



〈그림 5-1〉 실시간 데이터 수집을 통한 문제해결 사례

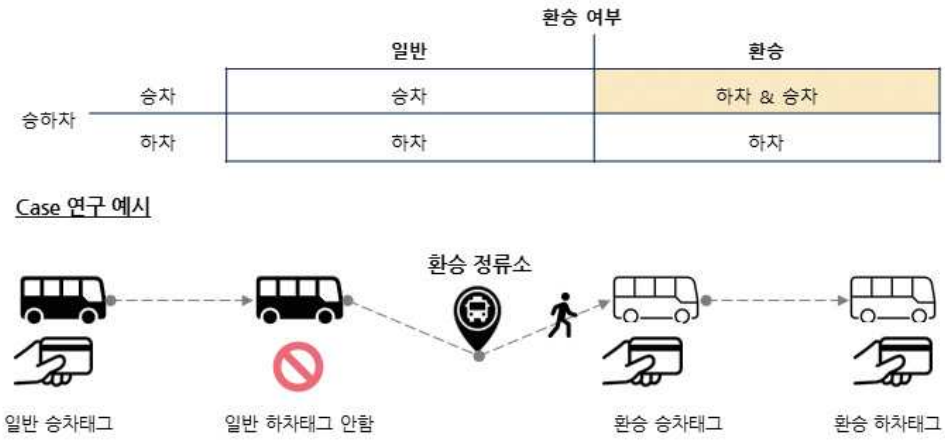
1-2 데이터 전처리 사례

– 데이터의 값이 누락된 데이터 처리를 1차 가공이라고 한다면 데이터를 분석에 활용할 수 있는 형태로 만드는 과정을 2차 가공이라고 한다. 예를 들어, 대구광역시의 '교통카드사용내역' 데이터를 가공한 사례가 좋은 예라 할 수 있다. 이 데이터는 시민들이 교통카드를 이용하면서 승하차 태그를 할 때 발생한다. 아래 그림에서 알 수 있듯이 대구광역시에서는 버스 하차 태그 데이터가 33% 밖에 존재하지 않았다. 하차 태그를 하지 않아도 추가 금액이 부가되지 않기 때문에 나타난 현상이었다. 그렇다고 분석가로서 67%나 되는 하차가 없는 승차 데이터를 제거하는 판단을 내릴 수는 없었다.



〈그림 5-2〉 대구광역시 버스 하차 태그 비율

– 시민들의 버스 이용 패턴을 분석하기 위해 없는 하차 데이터를 만들어 줘야 하는 상황이었다. 우리는 우선 데이터가 가진 속성들의 맥락을 연구하기 시작했다. 그리고 발생할 수 있는 Case별로 하나씩 하차 정보를 찾아 나가는 방식으로 가능한 많은 하차 데이터를 만드는 과정을 반복했다. 예를 들어, 아래 그림의 Case 연구 예시와 같이 버스의 환승 과정에서는 환승하기 전에 하차 태그가 없어 데이터가 빈 상태라도 맥락상 환승 승차 시점보다 몇 분 전으로 추정하여 값을 채울 수 있다. 이런 방법으로 모든 데이터의 하차 정보를 추정할 수는 없어도 분석의 정확도를 높일 수 있는 더 많은 데이터를 확보할 수 있었다.



〈그림 5-3〉 대구광역시 버스 승하차 맥락 연구

2 | 기술적 한계

2-1 버스노선 및 정거장 조정

– 버스노선 최적화 서비스의 초기 아이디어에서는 버스노선 재설정 및 정거장의 위치 조정 등이 활발하게 논의되었다. 하지만 버스노선 및 정류장을 조정한다는 것은 단순히 대기수요지수와 같이 시민들의 수요가 있다고 조정 가능한 것은 아니라는 결론을 낼 수 있었다. 공공적 측면, 사회적 측면, 그리고 버스회사의 운영비용 측면까지 고려하여 정의해야 하는 영역이다. 이러한 점에서 버스노선에 관련된 다양한 요소까지 데이터화되어 있지 않다는 한계점이 있었다. 예를 들어, 과거에 버스노선의

조정에 대한 구체적인 사유 및 내용에 관한 정보, 또 버스노선을 이용하는 데 시민들의 불편 요소는 무엇이었는지에 대한 조사 등 버스노선을 설계하고 정의하는 정책결정자들의 INPUT 데이터가 되는 모든 데이터가 충족되어야 실질적인 버스노선에 대한 조정 및 통폐합 가이드가 제공 가능하리라고 생각한다.

2-2 실시간 데이터 연계 기술의 한계

– 융복합 분석인 버스노선 최적화 서비스의 3대 주요 데이터는 버스 운행 데이터인 '버스정류소도착예상이력', 시민들의 이동현황을 나타내는 '유동인구 데이터', 자동차 이동 정보인 'Tmap 데이터'이다. 하지만 이 데이터는 현재 실시간 연계가 되고 있지는 않다. 우선 '버스정류소도착예상이력' 데이터의 경우 대구 교통서비스센터에서 직접 데이터 연계가 어려워 공공데이터 포털을 통해 API 방식으로 수집하고 있다. 이런 방식은 실시간 데이터 연계가 어렵고 10분 단위로 데이터 수집에 한계가 있다. 또 관련된 '버스정류장정보', '버스노선정류장목록'과 같은 데이터도 공공데이터포털에서 파일로 제공받고 있어 데이터의 시점에 대한 문제가 발생하고 있다. 또 유동인구와 Tmap 데이터의 경우 SK텔레콤에서 제공한 데이터로 특정 기간에 한정되어 제공했다. 이 데이터는 현재 시점의 데이터가 반영되지 않아 역시 예측 정확도를 높이는 데 기술적 한계가 있다고 볼 수 있다.

3 | 거버넌스 관련

3-1 데이터 제공 및 소유권의 정책 마련 필요

– 버스노선 분석 결과 데이터에 대한 제공 및 공개 가능한 범위를 정의하고 데이터 생애주기별 소유권 및 관리 책임 등에 대한 정책 마련이 필요하다. 대구광역시에서 어떤 데이터를 개방형 포털에서 제공할지 결정하여야 한다. 데이터 공유를 위한 담당자의 업무 및 책임소재를 분명히 하고, 데이터 사용자의 이용에 따른 책임을 보다 명확히 하여 담당자를 지정해야 한다. 그리고 데이터 활용 전문성 강화를 위한 교육과 정책 활용의 모범사례를 개발하여 홍보하고 전문기관과 협업을 통해 데이터의 지

속적인 활용체계를 구축해야 한다.

3-2 민간 데이터 활용 측면

– 현재 버스노선 분석에서 필요한 민간 데이터는 유동인구 데이터, Tmap 데이터로 SK텔레콤에서 제공하는 데이터이다. 이 데이터 현행화를 위해 공동구매 및 부서 내 공유방안 모색이 필요하다. 빅데이터 전담 지원조직을 통해 유동인구, 신용카드 매출데이터, Tmap 데이터 등 활용성과 수요도가 높은 데이터를 일괄적으로 확보하여 공유할 수 있는 기반을 마련해야 한다. 현재 활용하고 있는 민간데이터의 범위, 권한 등의 분석을 통해 적정 라이센스 비용을 산출하고 MOU, 협동연구 등을 통해 현실적인 구매 및 이용 방안을 마련해야 한다.

참고문헌

- 유병용, 양승태, 배상훈, 교통수단간 연계를 위한 최적 버스 배차간격 조정 알고리즘 개발, 2009. 대한토목학회지, p17~23
- 김수정, 신용은, 운행시간 및 수요 기반 버스 최적배차간격 산정에 관한 연구, 2018. Journal of the Korean Society of Civil Engineers, p167~174
- 이상용, 박경아, 시내버스노선체계 평가를 위한 정량적 지표의 설정 및 적용, 2003, 대한교통학회지, p29~44
- 신용은, 노면 대중교통노선 평가틀 구축에 관한 연구, 대한토목학회지, 2008, p477~483

스마트시티
혁신성장동력
프로젝트

SMART CITY