JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 30 No. 8, pp. 65-73, August 2025
https://doi.org/10.9708/jksci.2025.30.08.065

# Development of City Bus Passenger Prediction System Using Function Optimization

Min Kyu Jeong*, Ju Yeon Park**, Young-Tae Kwak***

*Student, Dept. of Physics, Jeonbuk National University, Jeonju, Korea
**Student, Dept. of IT & Engineering, Jeonbuk National University, Jeonju, Korea
***Professor, Dept. of Computer Science & AI, Jeonbuk National University, Jeonju, Korea

## [Abstract]

This study proposes a bus passenger prediction system that utilizes transportation card big data and reboarding passenger counts to overcome the limitations of missing alighting data in local cities. Focusing on three major routes in Jeonju (101, 165, and 970), the study applies and compares three machine learning algorithms (Random Forest, XGBoost, and LightGBM) while analyzing model performance by day of the week and time of day. April 2025 data was used for training and May 2025 for testing. Derived variables such as weekday/weekend indicators were created, and categorical features were label-encoded to fit the models. Performance was evaluated using RMSE, and LightGBM consistently showed the most stable and accurate results. The analysis revealed that prediction accuracy was higher on weekdays (Mon-Fri), whereas weekends (Sat-Sun) showed increased errors across all routes. By time of day, predictions were most accurate during morning commuting hours (06:00-08:00) and showed the largest errors during evening rush hours (17:00-19:00). The study demonstrates the feasibility of predicting demand even without alighting data and highlights the practicality and efficiency of the proposed system. The results can support transportation policy applications such as dynamic scheduling, route optimization, and efficient public resource allocation.

▸ **Key words**: City Bus, Prediction of Boarding Passengers, Transportation Card Big Data, Random Forest, XGBoost, LightGBM

• First Author: Min Kyu Jeong, Corresponding Author: Young-Tae Kwak
*Min Kyu Jeong (amaranth7@jbnu.ac.kr), Dept. of Physics, Jeonbuk National University
**Ju Yeon Park (okk1829@naver.com), Dept. of IT & Engineering, Jeonbuk National University
***Young-Tae Kwak (ytkwak@jbnu.ac.kr), Dept. of Computer Science & AI, Jeonbuk National University
• Received: 2025. 07. 31, Revised: 2025. 08. 07, Accepted: 2025. 08. 19.


http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

**[요 약]**

본 연구는 하차 태깅이 이뤄지지 않는 지방 도시의 교통 데이터 한계를 극복하고자, 교통카드 빅데이터와 재차인원 정보를 활용하여 정류소 단위의 승차 수요를 예측하는 시스템을 제안하였다. 전주시 주요 노선 3개(101, 165, 970번)를 대상으로 Random Forest, XGBoost, LightGBM 알고리즘을 비교 적용하고, 요일 및 시간대별 성능을 분석하였다. 실험에는 2025년 4월 데이터를 학습용으로, 5월 데이터를 시험용으로 활용하였다. 전처리 과정에서 요일, 주말 여부 등 파생 변수를 생성하고 Label Encoding을 통해 범주형 데이터를 정수값으로 변환하였다. 모델 성능 평가는 RMSE 지표를 사용하였으며, 전반적으로 LightGBM이 가장 낮은 예측 오차를 기록하여 안정적인 성능을 보였다. 특히 주중(월~금)에는 모든 모델이 안정적 성능을 보였으나, 주말과 퇴근 시간대(17~19시)에는 예측 오차가 급격히 증가하였다. 본 연구는 하차 정보 없이도 수요 예측이 가능하다는 실용성과, 데이터 전처리 및 학습 효율성 측면에서 높은 활용 가능성을 제시하며, 향후 탄력 배차, 노선 재설계, 예산 배분 등 정책적 활용에 기여할 수 있다.

▸ **주제어**: 시내버스, 승차인원 예측, 교통카드 빅데이터, Random Forest, XGBoost, LightGBM

## I. Introduction

최근 교통카드 데이터를 활용한 대중교통 수요 예측은 도시 교통 체계의 효율화를 위한 핵심 수단으로 각광받고 있다[1]. 특히 정류소 단위의 승·하차량 데이터를 정량적으로 분석함으로써, 교통 혼잡 완화, 노선 재설계, 운영 효율화 등에 실질적인 의사결정을 지원할 수 있다. 하지만 수도권 외 지방 중소도시의 경우 하차 시 태깅이 이뤄지지 않는 구조적인 한계가 있어, 정류소 단위의 하차 인원 파악은 물론 수요 예측도 어려운 상황이다. 이에 따라 별도의 장비 없이 기존의 승차 정보와 승객의 이동 패턴을 활용하여 수요를 추정하려는 연구가 활발히 이루어지고 있다.

기존의 관련 연구들은 크게 세 가지 방향으로 분류할 수 있다. 첫째, 정류소 추정 기반 접근법으로, 하차 데이터를 직접 확보하기 어려운 경우 다음 탑승 위치를 활용해 이전 하차 지점을 추정하는 방식이다. 이 방식은 카드 ID를 추적할 수 있는 데이터를 전제로 하며, 승·하차가 반복적으로 이뤄지는 정기 통근자의 행동 패턴에 기반해 상대적으로 정확한 추정이 가능하다[3-4]. 둘째, 센서 및 영상 기반 자동 승객 계수 시스템(APC: Automatic Passenger Counter) 을 활용한 방식이다. 적외선 센서, 열 감지기, CCTV 영상 등을 활용해 탑승·하차 인원을 자동으로 계수하지만, 고비용의 하드웨어가 필요하다는 한계가 있다[5-9]. 셋째, 머신러닝 또는 딥러닝 기반 예측 모델로, ARIMA, LSTM, GRU, GCN 등의 시계열 및 시공간 모델을 활용하여 승객 수요를 정량적으로 예측하는 방법이다[10-15]. 이들 기법은 예측 정확도가 높지만, 학습 데이터의 구조 복잡도 및 하이퍼파라미터 최적화의 어려움으로 인해 실사용 단계에서의 적용성이 낮은 경우도 많다.

본 연구에서는 이러한 복잡성과 한계점을 보완하기 위해, 교통카드빅데이터통합정보시스템의 재차인원 데이터를 기반으로 함수 최적화 방법을 이용하여 승객을 예측하는 시스템을 개발하였다. 모델링 단계에서는 Random Forest[16], XGBoost[17], LightGBM[18] 세 가지 대표적인 머신러닝 회귀 알고리즘을 비교 분석하였다. 모델 학습 이후 RMSE(Root Mean Square Error)를 성능 지표로 설정하였고, 학습 데이터와 테스트 데이터 모두에서 모델별 성능 차이, 노선별 성능 차이, 요일 및 시간대별 오차 패턴을 종합적으로 분석하였다. 그 결과, 전반적으로 LightGBM 모델이 가장 안정적인 예측 성능을 보였다.

기존 연구들이 하차지 추정 또는 정류소 패턴 분석에 집중한 데 반해, 본 연구는 실제 운영 데이터에서 추출한 '재차 인원'이라는 현실적 지표를 활용하여, 비교적 간단하면서도 실용적인 수요 추정 기반을 제공한다는 점에서 차별화된다. 실험 데이터는 2025년 4월을 학습용, 5월을 테스트용으로 구분하여 사용하였으며, 정류소명, 정류장 순번, 요일, 시간대 등의 범주형 변수를 Label Encoding을 통해 정수값으로 변환한 후, 예측 모델에 적용하였다.

본 연구의 장점은 다음과 같다. 첫째, 하차 태깅이 없는 지역에서도 교통카드 데이터와 재차 인원만으로 수요 예측이 가능하다는 실용성을 보여준다. 둘째, 기존 방식 대비 간단한 전처리와 비교적 낮은 계산 복잡도로도 충분한

예측 성능 확보가 가능하다. 셋째, 모델별·요일별·시간대별 성능 차이를 체계적으로 분석함으로써, 실제 대중교통 운영에서의 적용 가능성과 신뢰도를 높였다. 이러한 결과는 향후 지방도시에서의 버스 운영 최적화, 예산 재배치, 탄력적 배차 등에 유용한 기초자료로 활용될 수 있을 것으로 기대된다.

논문의 전개는 다음과 같다. 2장에서는 기존 교통카드 데이터를 활용하는 방법, 이미지분석, 통계적인 방법 등에 대하여 고찰하고, 각 방법의 한계와 문제점을 설명한다. 3장에서는 함수 최적화에 사용되는 Random Forest, XGBoost, LightGBM 등에 설명하며, 최적화 함수로 예측된 값을 이용하여 재차인원을 계산하는 방법 또한 제시한다. 4장 실험에서는 교통카드데이터 중에서 전주시에서 가장 많이 이용되는 시내버스 노선 3개를 선택하여 제시된 함수 최적화를 구현 및 실험하여 각 예측값을 비교하고 최적의 방법을 제시한다. 그리고 마지막으로 결론을 맺는다.

## II. Preliminaries

### 1. Related works

#### 1. Transportation Card Big Data

대중교통 이용 행태를 정량적으로 분석하기 위해 가장 널리 사용되는 자료는 교통카드(T-money, 후불카드 등) 빅데이터이다[1]. 이 데이터는 승객이 버스나 지하철 탑승 시 단말기에 교통카드를 태그함으로써 자동으로 수집되며, 탑승 일시, 정류소 ID, 노선 번호, 카드 유형, 차량 정보 등을 포함한다. 특히 시내버스의 경우, 이 데이터를 이용하면 정류소별 승차 인원을 정확하게 계산할 수 있다[2].

그러나 지방 중소도시 지역에서는 하차 시 교통카드 태그를 요구하지 않기 때문에, 하차 정류소에 대한 정보는 데이터에 존재하지 않으며, 이를 보완하기 위한 하차 정류소 추정 알고리즘이 필수적이다. 이를 위해 대표적으로 사용되는 방식은 다음 탑승 정보를 활용한 하차 추정법이다. 승객이 하차 후 일정 시간 이내에 다른 노선을 탑승한 경우, 두 정류소 간의 공간·시간 거리, 노선 연결성, 도보 이동 가능성 등을 분석하여 그 이전 버스의 하차 정류소를 추정한다[3].

또한 하차 패턴 기반 모델링 기법은 동일 승객의 과거 이용 패턴을 학습하거나, 다수의 유사 승객 패턴을 기반으로 특정 정류소에서의 하차 확률을 추정하는 방식으로 활용된다. 최근에는 이러한 방식에 통계적 추정 모델이나 기계학습 알고리즘을 적용하여 추정 정확도를 높이는 연구가 활발하다[4]. 더 나아가, 하차 추정 알고리즘을 통해 재구성된 승하차 데이터는 OD 행렬 생성, 수요 예측, 노선 개편 시뮬레이션 등 다양한 정책적 활용이 가능하다.

이처럼 교통카드 데이터 기반의 승하차 계산은 추가적인 장비 없이도 대규모 정류소 수준 데이터를 확보할 수 있는 경제적 방법이며, 실시간 수집과 분석이 가능하다는 점에서 스마트 교통 인프라의 핵심 요소로 간주되고 있다. 하지만 실제 하차 정보가 없어 정확도가 떨어질 수 있고, 다음 탑승이 없는 경우에는 추정이 어렵다. 또한 불규칙한 승객에게는 적용이 어렵다는 단점이 있다.

#### 2. Image Analysis

일반적으로는 교통카드 데이터를 통해 승차 인원을 추정하고 있으나, 이는 현금 승차자나 태그 누락 등의 문제로 인해 정확한 실승객 수를 반영하지 못하는 한계가 존재한다. 이러한 문제를 보완하기 위해 CCTV 영상 분석, 적외선 센서, 열 감지 카메라, AI 기반 영상 인식 기술 등을 이용한 자동 승객 계수 시스템(APC)이 주목받고 있다.

적외선 센서를 활용한 방식은 차량 출입문에 설치된 양방향 감지 센서를 통해 사람의 이동 방향을 감지하여 승차와 하차를 구분하고 계수하는 방식이다[5-6]. 이 방식은 교통카드 미사용자도 포함할 수 있다는 점에서 데이터의 포괄성이 높다. 그러나 혼잡 시 감지 정확도 저하나 두 명 이상이 동시에 탑승할 경우의 오류 등 기술적 한계도 보고되고 있다.

한편, CCTV 기반의 영상 분석 방식은 버스 내부 또는 정류소 외부에 설치된 카메라 영상을 인공지능 기반 영상 인식 알고리즘(예: YOLO, OpenPose 등)을 통해 분석하는 방식이다[7-8]. 이 방식은 승객의 이동 경로, 체류 시간, 행태 등을 종합적으로 파악할 수 있으며, 탑승 인원뿐만 아니라 혼잡도나 특정 시간대 밀집도 분석에도 활용 가능하다. 또한 열 감지 센서(thermal imaging)를 이용하면 조명, 그림자, 배경의 변화 등 일반 영상분석의 한계를 극복할 수 있어 야간 환경에서도 높은 정확도를 유지할 수 있다.

최근 연구에서는 이러한 자동 계수 기술을 교통카드 데이터와 융합하여 하차 추정 정확도를 향상시키는 시도도 이루어지고 있다. 특히, 일부 상용 APC 시스템은 영상 기반 계수 결과를 딥러닝 알고리즘과 결합하여 95% 이상의 계수 정확도를 달성하고 있다[9].

이처럼 영상 및 센서 기반 계수 시스템은 교통카드의 한계를 보완하고, 보다 정밀한 정류소별 승하차 데이터 수집이 가능하다는 점에서 향후 대중교통 데이터 인프라 확장

의 핵심 기술로 평가되지만 설치 비용이 높고, 혼잡한 상황에서는 감지 오류가 발생할 수 있다.

### 3. Statistic Model and Deep Learning

최근 대중교통 데이터의 수집과 분석 기술이 발전함에 따라, 통계 기반 기법과 딥러닝 기반 시공간 모델을 중심으로 연구가 활발히 진행되고 있다. 초기 연구에서는 주로 통계 기반 시계열 분석을 통해 특정 노선이나 정류장 단위의 이용 패턴을 예측하였다. 대표적으로 ARIMA (Autoregressive Integrated Moving Average) 모델이 널리 활용되었으며, 이는 과거의 승차량 데이터에 기반하여 미래 수요를 예측하는 데 효과적이었다. 특히 주중과 주말, 출퇴근 시간대의 이용 패턴 변화를 정량적으로 분석하는 데 강점을 보였다[10-11].

그러나 통계 기반 모델은 공간적 상호작용이나 비선형적 요소를 반영하기 어렵다는 한계가 있어, 최근에는 딥러닝 기반 시공간 예측모델이 대두되고 있다. 이들 모델은 버스 정류장 간의 위치 정보, 노선 구조, 시간대별 흐름 등을 통합적으로 고려할 수 있다는 장점이 있다. 대표적으로 LSTM(Long Short-Term Memory)은 시간적 패턴을 효과적으로 포착하는 데 활용되며, 이를 CNN(Convolutional Neural Network) 또는 GCN(Graph Convolutional Network)과 결합한 모델들은 공간적 연관성까지 반영한 고도화된 예측이 가능하다[12-13]. 또한 일부 연구에서는 실제 CCTV, Wi-Fi, GPS 센서 데이터를 활용하여 탑승객의 움직임을 실시간으로 탐지하고, 이를 자동 계수 시스템에 통합함으로써 기존 수작업 계수 방식의 한계를 극복하고 있다[14-15].

이처럼, 통계적 모델은 해석력이 뛰어나고 예측 구조가 단순한 반면, 딥러닝 기반 시공간 모델은 대규모 데이터를 활용해 복잡한 패턴을 학습할 수 있어 정확도 면에서 우수한 성과를 보이고 있다. 통계 기반 모델은 해석이 쉽고 구현이 간단하지만, 공간적 상호작용이나 복잡한 비선형 관계를 반영하기 어렵다. 반면 딥러닝 모델은 높은 정확도를 보이지만, 많은 학습 데이터와 연산 자원이 필요하며, 결과 해석이 어렵고 과적합 위험이 있다는 단점이 있다.

## III. The Proposed Scheme

### 1. Random Forest Algorithm

Random Forest는 Breiman[16]이 제안한 앙상블 학습 기법으로, 다수의 결정 트리(Decision Tree)를 학습시킨 뒤 이들의 예측을 집계(분류의 경우 다수결, 회귀의 경우 평균)하여 최종 결과를 도출하는 알고리즘이다. 개별 트리의 분산을 낮추고 과적합을 방지하기 위해 배깅(Bootstrap Aggregating)과 무작위 특성 선택(Random Feature Selection)을 결합한 방식으로 작동한다.

학습 데이터 집합 $D = \{(x_1, y_1), \cdots (x_n, y_n)\}$가 주어졌을 때, Random Forest는 다음과 같은 절차로 $T$개의 트리 $\{h_1, h_2, \cdots, h_T\}$를 생성한다.

1. 각 트리 $h_t$는 원본 데이터로부터 중복을 허용하는 부트스트랩 샘플링을 통해 학습 데이터 $D_t$를 생성한다.
2. 각 노드 분할 시, 전체 특성 중 무작위로 선택된 일부 특성 $m$개만을 고려하여 최적 분할 기준을 찾는다.
3. 트리는 최대 깊이까지 성장하거나 최소 샘플 수 기준에 도달할 때까지 분할된다.

최종 예측은 다음과 같이 집계된다.

분류 문제의 경우:

$$\hat{y} = majority - vote(h_1(x), \cdots, h_T(x)) - \text{식 (1)}$$

회귀 문제의 경우:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} h_t(x) - \text{식 (2)}$$

이러한 구조는 개별 트리의 예측 오차가 서로 상관이 낮을수록, 전체 앙상블의 예측 성능이 높아지는 성질을 활용한다. 또한, 무작위성을 도입하여 트리 간 상관성을 줄이고, 일반화 성능을 향상시킨다.

Random Forest는 별도의 검증 데이터 없이도 모델의 정확도를 추정할 수 있도록 각 트리 학습에 사용되지 않은 샘플(Out-of-Bag, OOB)을 이용한 오차 추정도 지원하며, 특성 중요도(feature importance)를 계산하는 데에도 활용 가능하다.

### 2. XGBoost Algorithm

XGBoost(Extreme Gradient Boosting)는 Gradient Boosting 기법을 확장한 고성능 머신러닝 알고리즘으로, 분류(Classification), 회귀(Regression), 순위 예측(Ranking) 등 다양한 문제에 적용이 가능하다[17]. 이 알고리즘은 부스팅(Boosting)을 기반으로 하며, 약한 학습기(보통 결정 트리)를 순차적으로 추가하면서 이전 모델의 오차를 보완하는 방식으로 작동한다.

XGBoost에서 모델의 예측값은 다음과 같이 표현된다:

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in F \; - \text{식 (3)}$$

여기서 $\hat{y_i}$는 $x_i$에 대한 예측값, $K$는 트리의 개수, $f_k$는 $k$-번째 결정 트리, $F$는 가능한 트리 함수의 집합이다. 전체 모델의 목적 함수(Objective Function)는 예측 오차를 나타내는 손실함수 $L$와 모델 복잡도를 제어하는 정규화 항 $\Omega$의 합으로 정의된다.

$$L(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad - \text{식 (4)}$$

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2$$

여기서 $T$는 트리의 리프 노드 개수, $w$는 리프 노드에 할당된 값, $\gamma$, $\lambda$는 정규화 계수이다. 모델은 반복적으로 각 단계 $t$에서 손실 함수의 2차 테일러 전개를 기반으로 새로운 트리를 학습한다. 이때 이차 근사 형태는 다음과 같다.

$$L^{(t)} \approx \sum_{i=1}^{n} \left[ g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad - \text{식 (5)}$$

$$g_i = \frac{\partial l(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}}, \qquad h_i = \frac{\partial^2 l(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^2}$$

여기서 $g_i$와 $h_i$는 각각 손실함수의 1차 및 2차 도함수로, 각 샘플에 대한 기울기와 곡률 정보를 나타낸다. XGBoost는 이러한 목적 함수 최적화를 통해 반복적으로 트리를 추가하면서 성능을 향상시키며, 정규화 항을 통해 모델의 과적합을 방지한다. 또한, 병렬 처리, 예측값 자동 처리, 조기 종료(Early Stopping) 기능 등을 통해 계산 효율성과 실용성을 높인다. 이로 인해 XGBoost는 함수 최적화에 널리 활용되고 있다.

### 3. LightGBM Algorithm

LightGBM(Light Gradient Boosting Machine)은 Microsoft에서 개발한 고성능 Gradient Boosting 프레임워크로, 대용량 데이터에서도 빠른 학습과 높은 예측 정확도를 보장하는 결정 트리 기반의 머신러닝 모델이다[18]. 이 알고리즘은 기존의 Gradient Boosting 방식을 따르면서도, 데이터 처리 및 트리 생성 과정에서 다양한 최적화를 적용하여 효율성과 성능을 동시에 향상시킨다.

LightGBM은 매 반복 단계 $t$에서 잔차를 보정하기 위한 약한 학습기 $f_t \in F$를 추가하며, 예측값은 다음과 같이 갱신된다.

$$\hat{y_i}^{(t)} = \hat{y_i}^{(t-1)} + f_t(x_i) \quad - \text{식 (6)}$$

모델의 전체 목적 함수는 손실 함수와 정규화 항의 합으로 구성되며, 다음과 같이 표현된다.

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \quad - \text{식 (7)}$$

여기서 $l$은 손실 함수이며, $\Omega(f_k)$는 트리 복잡도를 제어하는 정규화 항이다. 손실 함수는 2차 테일러 전개를 통해 근사되며, 이때 도함수를 활용하여 다음과 같은 형태로 최적화가 수행된다.

$$L^{(t)} \approx \sum_{i=1}^{n} \left[ g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad - \text{식 (8)}$$

여기서 $g_i$와 $h_i$는 각각 손실함수의 1차 및 2차 도함수로, gradient boosting 과정에서 오차 보정의 방향성과 크기를 결정하는 역할을 한다.

LightGBM은 이러한 기본 구조 위에 세 가지 주요한 개선 기법을 적용한다. 첫째, 트리 구조 생성에 있어 기존의 level-wise 방식이 아닌 leaf-wise 성장 전략을 채택함으로써, 손실 감소가 큰 리프 노드를 우선적으로 확장하여 동일한 트리 깊이에서 더 높은 정확도를 달성한다. 둘째, 데이터 샘플링 과정에서는 Gradient-based One-Side Sampling (GOSS) 기법을 도입하여, gradient가 큰 샘플은 모두 선택하고 작은 gradient를 가진 샘플 중 일부만을 무작위로 선택함으로써 계산량을 절감하면서도 정보 손실을 최소화한다. 셋째, Exclusive Feature Bundling (EFB) 기법을 통해 서로 상호배타적인 희소 특성들을 묶어 특성 차원을 효과적으로 축소함으로써 학습 속도를 크게 향상시킨다. 이러한 구조적 최적화 덕분에 LightGBM은 대규모 데이터셋에 적합하며, 분류, 회귀, 순위 예측 등 다양한 머신러닝 문제에서 널리 사용되고 있다.

### 4. Passenger Count Calculation

시내버스의 탑승 행태를 정량적으로 분석하기 위해, 본 연구에서는 각 정류소에서의 승차인원을 재차인원 및 하차인원 데이터를 이용하여 추정하는 방식을 제안한다. 일반적으로 버스는 각 정류소에 도착할 때마다 하차가 이루어지고, 그 후 승차가 발생한다는 운행 순서를 가정할 수 있다. 이에 따라, 정류소 통과 전후의 재차인원 변화를 통해 승차인원을 다음과 같은 수식으로 계산한다.

$$B_i = R_{i+1} - R_i + A_i \quad - \text{식 (9)}$$

여기서, $B_i$는 $i$번째 정류소에서의 승차인원, $R_i$는 $i$번째 정류소 도착 직전의 재차인원, $R_{i+1}$는 $i$번째 정류소 출발 직후의 재차인원, $A_i$는 $i$번째 정류소에서의 하차인원을 의미한다.

이 수식은 다음과 같은 논리적 기반을 갖는다. 정류소 도착 시점에서 버스에는 $R_i$명의 승객이 탑승해 있으며, $A_i$명이 하차하게 된다. 이후 $B_i$명이 승차하면, 정류소를 출발할 때의 재차인원은 $R_{i+1} = R_i - A_i + B_i$가 된다.

Table 1. Training Data Sample on April 1st

| 노선 | 기종점 | 정류장순번 | 정류장명 | 06시 | 07시 | ... | 11시 | 12시 | 13시 | ... | 16시 | 17시 | 18시 | 23시 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ... | ⋮ | ⋮ | ⋮ | ... | ⋮ | ⋮ | ⋮ | ⋮ |
| 101 | 전북대종점 - 전북대종점 | 17 | 호남제일문 | 0 | 0 | | 0 | 0 | 0 | | 6 | 3 | 0 | 0 |
| 101 | 전북대종점 - 전북대종점 | 18 | 월드컵경기장입구 | 0 | 0 | | 3 | 0 | 3 | | 7 | 0 | 4 | 0 |
| 101 | 전북대종점 - 전북대종점 | 19 | 월드컵경기장남문 | 0 | 0 | ... | 0 | 1 | 0 | ... | 0 | 0 | 5 | 0 |
| 101 | 전북대종점 - 전북대종점 | 20 | 장동에코르아파트 | 0 | 0 | | 1 | 2 | 0 | | 6 | 2 | 4 | 0 |
| 101 | 전북대종점 - 전북대종점 | 21 | 동재마을 | 0 | 0 | | 0 | 0 | 0 | | 0 | 0 | 0 | 0 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ... | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ | ⋮ |

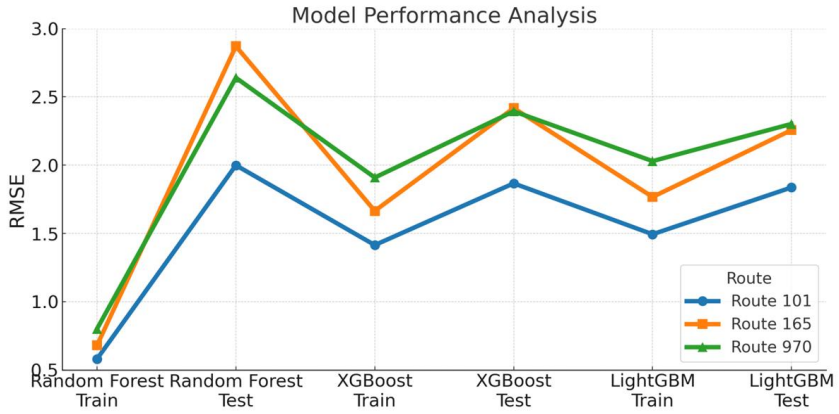



Fig. 1. Model Performance Analysis

이를 재정리하면 식 (9)와 같이 승차인원을 도출할 수 있다. 하지만, 서울과 수도권을 제외한 지방도시에서는 하차 시 교통카드를 태그하지 않기 때문에 식(9)을 적용할 수 없어, 관련 연구에서 제안한 카메라나 센서를 이용하여 하차인원을 추정할 수 있다면 각 정류소의 승차인원을 보다 더 정확히 예측할 수 있다.

## IV. Experimental Results

### 1. Definition of Experiment

본 논문의 실험은 교통카드빅데이터통합정보시스템의 전주 시내버스 데이터 중 가장 승차 인원이 많은 3개 노선(101번, 970번, 165번)을 대상으로 2025년 4월 한 달 동안의 재차 인원(승차하고 있는 인원)을 대상으로 학습데이터(213,840개)를 사용하였으며, 5월 한 달 데이터를 시험데이터(220,968개)로 사용하였다. 표 1은 101번 시내버스 4월 1일의 데이터 표본이다.

데이터 전처리 단계에서는 날짜 정보를 바탕으로 요일(0: 월요일 ~ 6: 일요일), 주말 여부(is_weekend), 공휴일 여부(is_holiday), 그리고 주말 또는 공휴일 여부(is_weekend_or_holiday)를 나타내는 파생 변수를 생성하였다. 이때 요일 값이 토요일(5) 또는 일요일(6)인 경우 is_weekend 값을 1로 설정하였다. 이후 분석에 불필요한 컬럼을 제거하고, 시간대별로 분리되어 있던 데이터를 melt 함수를 활용하여 Long 포맷으로 변환하였다.

주요 범주형 변수인 정류장순번, 정류장명, 시간, 요일, 날짜, 공휴일 유무 등은 Label Encoding을 통해 정수값으로 변환하였다. 학습 데이터와 시험 데이터에 동일한 변환을 적용함으로써, 데이터셋 간 일관성을 유지하였다. 전처리된 데이터을 바탕으로 파이썬으로 RandomForest, XGBoost, LightGBM 등을 구현하여 학습을 수행하였고, 예측 결과를 도출하였다. 모델 성능 평가는 RMSE(Root Mean Square Error)지표를 기준으로 수행하였다.

실험에서 사용된 머신러닝 모델은 모두 별도의 튜닝 없이 기본 하이퍼파라미터(default parameters)를 사용하였다. 즉, LightGBM의 num_leaves, max_depth, XGBoost의 max_depth, learning_rate 등 주요 파라미터는 모두 라이브러리 기본값을 따랐으며, 공통적으로 n_estimators=100, random_state=42로 설정하였다. 실험은 Google Colab Pro 환경(CPU 기반)에서 수행되었으며, 운영체제는 Ubuntu 22.04, Python 3.10.12 버전, 주요 라이브러리는 pandas(1.5.3), numpy(1.24.3), scikit-learn(1.2.2), xgboost(1.7.6), lightgbm(3.3.5), holidays(0.26)을 사용하였다. 하드웨어 사양은 Intel Xeon CPU @2.20GHz (2코어), RAM 13GB이며 GPU는 사용하지 않았다. 모델 학습 시간은 버스 1개 노선 단위 기준으로 Random Forest 약 2.5초, XGBoost 약 3.5초, LightGBM 약 2.8초가 소요되었다.

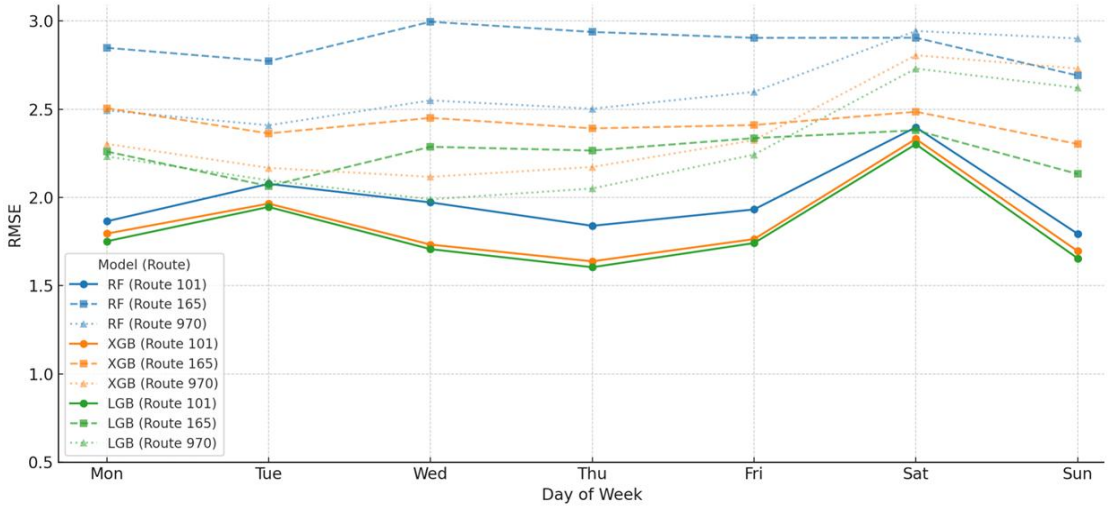



Fig. 2. Day-of-Week Performance Analysis

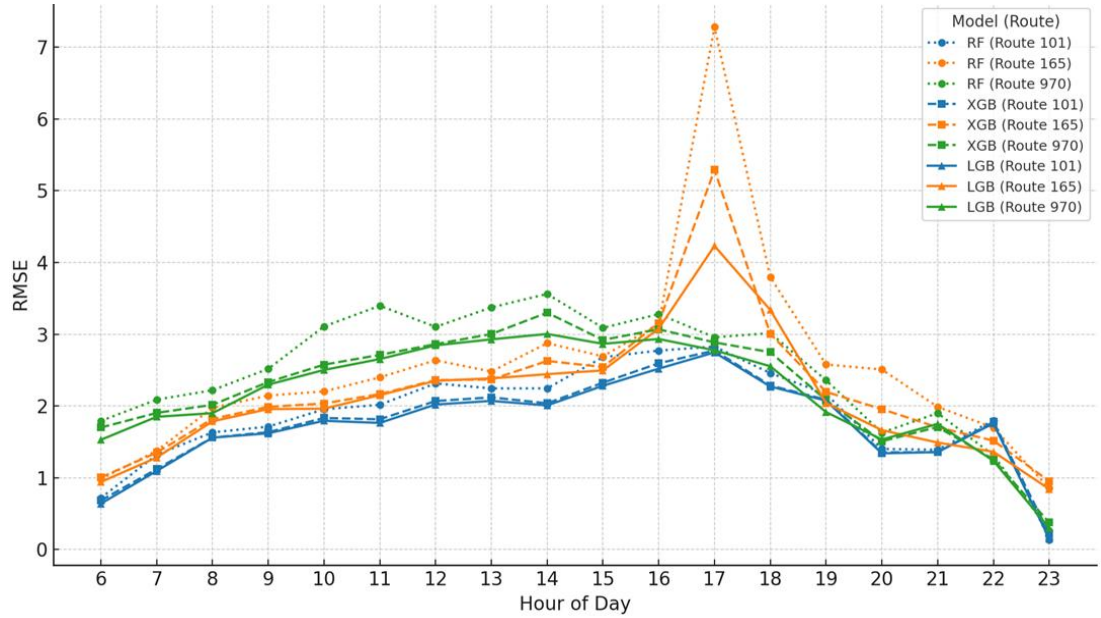



Fig. 3. Hourly Performance Analysis

## 2. Model Performance Comparison

그림 1은 세 가지 모델에 대해 각 노선의 학습 및 시험 데이터의 평균 RMSE 성능을 시각화한 것이다. RMSE는 예측값과 실제값 간 오차의 제곱 평균을 나타내는 지표로, 값이 낮을수록 예측 정확도가 높음을 의미한다.

Random Forest는 학습 데이터에서 가장 낮은 RMSE를 기록했지만, 시험 데이터에서는 모든 노선에서 RMSE가 급격히 증가하는 과적합(overfitting) 경향을 보였다. 이는 학습 데이터에 과하게 적응한 결과로, 실제 예측 상황에서는 일반화 성능이 저하된다는 점을 시사한다.

반면, XGBoost는 학습과 시험 데이터 간 성능 차이가 상대적으로 적으며, 전반적으로 가장 균형 잡힌 예측 성능을 보여주는 모델로 확인되었다. 특히 101과 165 노선에서 시험 RMSE가 안정적으로 유지되어, 일반화 능력이 우수한 것으로 평가된다.

LightGBM 역시 XGBoost와 유사한 수준의 예측 성능을 나타냈으며, 일부 노선에서는 XGBoost보다 더 낮은 RMSE를 기록했다. 특히 970번 노선에서의 시험 RMSE가 가장 낮게 측정되어, 대규모 데이터 처리나 복잡한 패턴 예측에서 LightGBM이 효과적일 수 있음을 의미한다.

그림 2는 모델별 그리고 요일을 기준으로 시험 데이터에 대한 RMSE 성능을 비교한 결과를 나타낸다. 분석 결과, 전반적으로 평일(월요일~금요일)에는 모든 모델과 노선에서 비교적 낮은 RMSE 값을 기록하며, 예측 정확도가 안정적으로 유지되는 경향을 보였다. 특히 LightGBM은 평일뿐만 아니라 전체 구간에서 가장 낮고 일관된 RMSE

값을 보이며, 전반적으로 가장 안정적인 예측 성능을 제공하는 것으로 나타났다.

반면, 주말(토요일과 일요일)에는 대부분의 모델과 노선에서 RMSE가 급격히 상승하는 패턴이 나타났으며, 이는 주말 수요의 불규칙성과 예측의 어려움을 반영한 결과로 해석된다. 특히 165번과 970번 노선은 주말에 RMSE가 두드러지게 상승하여, 해당 노선은 수요 패턴이 불안정하거나 예측에 민감한 특성을 가질 가능성이 있다.

한편, 세 모델 중 165 노선은 전 구간에서 가장 높은 RMSE를 기록하여 예측 난이도가 가장 높은 노선으로 확인되었다. 이러한 결과는 해당 노선이 다양한 승객 특성, 변동성 있는 이용 패턴 등을 포함하고 있을 가능성을 시사하며, 추가적인 피처 보강이나 모델링 전략의 개선이 필요할 수 있음을 암시한다.

그림 3은 시계열을 기반으로, 하루 24시간 중 주요 시간대별 RMSE(Root Mean Square Error) 값을 나타낸 것으로, 시험 데이터에 대한 노선(Route)별 및 모델별 예측 성능의 차이를 보여준다.

먼저, 970 노선은 전 시간대에 걸쳐 가장 높은 RMSE 값을 기록하였다. 이는 해당 노선의 수요 패턴이 타 노선에 비해 더욱 복잡하거나 불규칙하다는 점을 시사하며, 예측이 상대적으로 어려운 노선으로 판단된다. 반면, LightGBM 모델은 전반적인 시간대에서 가장 낮은 RMSE를 유지하여, 세 모델 중 가장 우수한 예측 성능을 보였다.

출근 시간대인 06~08시 구간에서는 전 노선에서 RMSE가 낮게 측정되었다. 이것은 이 시간대의 승객 탑승 패턴이 상대적으로 고정되어 있어 예측이 용이하다는 점을 보여준다. 반면, 17~19시 퇴근 시간대에는 모든 노선에서 예측 오차가 크게 증가하는 경향이 나타났다. 특히 970번 노선은 18시에 RMSE가 3.33까지 치솟아, 해당 시간대의 수요 예측이 매우 어려운 것으로 분석되었다. 21시 이후 심야 시간대에서는 RMSE가 다시 낮아지는 패턴을 보였으며, 이는 심야 시간의 탑승 수요 자체가 적고 패턴이 일정하여 모델의 예측 정확도가 다시 높아지는 원인으로 해석된다.

이상의 분석을 종합하면, LightGBM이 전반적으로 높은 예측 정확도와 안정성을 보여주었으며, Random Forest는 학습 데이터에는 적합하나 실제 적용에는 다소 불안정한 성능을 보였다. 이러한 분석 결과는 실제 운송 수요 예측 모델을 구축할 때 모델 선택 및 성능 검증 기준으로 활용될 수 있다. 그리고 모델 성능은 요일과 시간에 따라 차이를 보이며, 이러한 특성을 고려하여, 주말 및 특정 시간에 대해서는 별도의 예측 전략 혹은 보완적 모델 설계가 요구될 수 있다.

## V. Conclusions

본 연구는 하차 태깅이 없는 지방 도시의 현실적인 한계를 고려하여, 교통카드 빅데이터와 재차인원 데이터를 기반으로 시내버스 승차 수요를 예측하는 시스템을 제안하였다. 정류소 단위의 범주형 정보를 Label Encoding으로 정제하고, 대표적인 머신러닝 회귀 모델인 Random Forest, XGBoost, LightGBM을 적용하여 모델 성능을 비교하였다. 실험 결과, LightGBM은 모든 노선과 시간대에서 가장 안정적이고 우수한 성능을 보였으며, 특히 테스트 데이터에서의 예측 정확도가 높게 나타났다. 반면 Random Forest는 과적합 경향이 나타나 실운영에서는 주의가 필요함을 확인하였다. 또한, 요일 및 시간대별 성능 분석을 통해 평일 대비 주말, 특히 퇴근 시간대(17~19시)에 예측 오차가 크게 증가하는 경향을 확인하였으며, 이는 실제 운영상에서 고려해야 할 중요한 변수임을 시사한다.

이러한 연구 결과는 지방 도시 교통 운영의 효율화를 위한 의사결정(노선 재편, 탄력 배차, 혼잡 완화 전략 등)에 기여할 뿐만 아니라, 교통 빅데이터 기반 스마트 모빌리티 연구의 발전에도 긍정적인 영향을 줄 것으로 기대된다. 그럼에도 불구하고 본 연구에는 몇 가지 한계가 존재한다. 첫째, 하차 태깅 부재로 인해 재차인원 중심의 간접 추정 방식을 사용함에 따라 실제 승하차 행태를 완전히 반영하지 못할 수 있다. 둘째, 분석 변수가 교통카드 데이터와 요일·시간대에 국한되어 있어, 예측력이 특정 상황(주말, 퇴근 시간대 등)에서 제한적으로 나타났다.

향후 연구에서는 외부 변수를 적극적으로 통합하는 다변수 예측모델의 구체적 접근이 요구된다. 예를 들어, (1) 기상청 API를 통한 기상 데이터 연계, (2) 상권 및 유동인구 데이터와의 결합, (3) 지역 행사 일정이나 학사·근무 스케줄 등 이벤트성 요인의 반영, (4) 그래프 신경망(GNN)이나 시공간 딥러닝 모델을 통한 정류소 간 상호작용 학습 등을 고려할 수 있다. 이러한 확장은 수요 변동성을 더욱 정밀하게 반영함으로써, 실시간 배차 조정이나 맞춤형 서비스 설계가 가능한 고도화된 스마트 교통 운영 시스템으로 발전하는 기반이 될 것이다.

## REFERENCES

[1] Integrated Information System for Transportation Card Big Data https://stcis.go.kr/wps/main.do

[2] Kim, Seong-A, Lee, Jonghyeok, Park, Se Jin, An, Sehyun, Kim, Heungsoon, “Analyzing Urban Characteristics Affecting Public

Transportation Use of the Disabled in Seoul : Using Transportation Card Data," Journal of the Korean Regional Development Association, Vol. 35, No. 5, pp. 83-100, December 2023.

[3] Min Hyuck Lee, In Woo Jeon and Chulmin Jun, "Demand Estimation of Public Transport using Smart Card Big Data," Journal of Korean Society for Geospatial Information Science, Vol. 28, No. 3, pp. 3-10, September 2020. DOI: 10.7319/kogsis.2020.28.3.003

[4] Ho-Sung Lee, Jin-Woo Lee and Seong Baeg Kim, "Developing Public Transportation Bus Passenger Counting System Based on IoT and Deep Learning", Korean Institute of Information Scientists and Engineers, Vol. 27, No. 1, pp. 22-31, January 2021. DOI: 10.5626/KTCP.2021.27.1.22

[5] Radovan, A., Mršić, L., Đambić, G., and Mihaljević, B., "A Review of Passenger Counting in Public Transport Concepts with Solution Proposal Based on Image Processing and Machine Learning,". Eng, Vol. 5, No. 4., pp. 3284-3315, December 2024. DOI:10.3390/eng5040172

[6] Myoungbeom Chung, "A Disembarking Notification System in Public City Buses using Smart Device and High Frequency," Journal of The Korea Society of Computer and Information, Vol. 25, No. 8, pp. 55-63, August 2020. DOI:10.9708/jksci.2020.25.08.055

[7] Zhao, J., Li, C., Xu, Z., Jiao, L., Zhao, Z., and Wang, Z., "Detection of passenger flow on and off buses based on video images and YOLO algorithm," Multimedia Tools and Applications, Vol. 81, No. 4, pp. 4669-4692, 2022. DOI:10.1007/s11042-021-10747-w

[8] Dae-Hyun Kim, Hyuk-Jin Yoon, Junhyun Song and Chanho Park, "Study on CCTV-Based Seat Occupancy Recognition System for Public Transportation," Journal of The Korean Society for Railway, Vol. 28, No. 21, pp. 160-168, February 2025. DOI: 10.7782/JKSR.2025.28.2.160

[9] Pronello, C., and Garzón Ruiz, X. R., "Evaluating the performance of video-based automated passenger counting systems in real-world conditions: A comparative study." Sensors, Vol. 23, No. 18, pp. 7719. September 2023. DOI:/10.3390/s23187719

[10] Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M., Time series analysis: forecasting and control, John Wiley & Sons, 2015.

[11] Mendhe, R. R., Shingare, S., Shinde, D., Surve, A., and Kulkarni, S, "Forecasting Passenger Traffic in Metro Systems: ARIMA Analysis of Metro Ticket Reservation Data." 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI). IEEE, pp. 740-748, January 2025. DOI: 10.1109/ICMCSI64620.2025.10883401

[12] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting," International Conference on Learning Representations (ICLR), 2018.

[13] Drosouli, I., Voulodimos, A., Mastorocostas, P., Miaoulis, G., and Ghazanfarpour, D. "A spatial-temporal graph convolutional recurrent network for transportation flow estimation." Sensors, Vol. 23, No. 17, pp.7534 ,September 2023. DOI: 10.3390/s23177534

[14] Ya-Wen Hsu, Yen-Wei Chen and Jau-Woei Perng, "Estimation of the Number of Passengers in a Bus Using Deep Learning," April 2020.

[15] Baghbani, A., Rahmani, S., Bouguila, N., and Patterson, Z. "Predicting passenger flow using graph neural networks with scheduled sampling on bus networks," 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 3073-3078, September 2023. DOI: 10.1109/ITSC57777.2023.10422701

[16] Breiman, L.. "Random Forests," Machine Learning, Vol. 45, No. 1, pp. 5−32, October 2001. DOI:10.1023/A:1010933404324

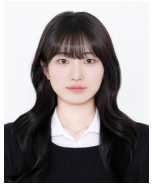
[17] Chen, T., and Guestrin, C., "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785-794, August 2016. DOI:10.1145/2939672.2939785

[18] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y., "LightGBM: A highly efficient gradient boosting decision tree." Advances in Neural Information Processing Systems, Vol. 30, 2017.

## Authors

Min Kyu Jeong is currently pursuing the B.S. degrees in Physics and IT Information Engineering at Jeonbuk National University, Republic of Korea, and is expected to graduate in February 2026.

Mr. Jeong is interested in machine learning, statistical learning, and mathematics.



Ju Yeon Park is an undergraduate student in the Dept. of Information Technology & Engineering at Jeonbuk National University, Republic of Korea. Her interests include machine learning, data science, and artificial intelligence.

Young-Tae Kwak received the B.S., M.S., and Ph.D. degrees in computer engineering from the Chungnam National University, Republic of Korea, in 1993, 1995, and 2001, respectively.

He joined the faculty of the Jeonbuk National University in 2002. His research interests include computer vision and neural networks.