날씨 데이터를 이용한 기계 학습 알고리즘 기반 시간별 지하철 이용량 예측

권 산[1], 이예진[2], 김현정[3], 김조은[1], 조석헌*

[1]국립안동대학교, [2]계명대학교, [3]건국대학교, *University of California, San Diego (UCSD)

[1]jmt30269@gmail.com, [2]ofxxcrax@gmail.com, [3]b.bhj0817@gmail.com, [1]joenkim200212@gmail.com,

*justinshcho@gmail.com

Prediction of Hourly Subway Usage based on Machine Learning Algorithms Using Weather Information

[1]San Gwon, [2]Ye-Jin Lee, [3]Hyeon-Jeong Kim, [1]Jo-eun Kim, Seokheon Cho*

[1]Andong National Univ., [2]Keimyung Univ., [3]Konkuk Univ., *University of California, San Diego (UCSD)

요 약

본 논문은 워크플로 기반 데이터 분석 도구인 KNIME을 활용하여 날씨에 따른 시간별 지하철 이용량 예측을 위해 다양한 기계 학습 알고리즘들을 적용하여 성능을 비교 및 분석하였다. 이때 여러 알고리즘들 중, Multiple Linear Regression과 Random Forest를 이용하여 예측하였다. 지역적 특성이 다른 을지로입구역과 뚝섬유원지역의 독립변수의 상관관계를 기반으로 전체 독립변수를 사용하여 예측한 결과와 기준에 따라 선정한 8개의 독립변수만을 사용하여 예측한 결과를 비교하였다. 고정적인 수요가 강한 을지로입구역에서는 기상과 관련된 변수의 영향을 거의 받지 않았지만, 뚝섬유원지역은 여가 활동을 즐기는 사람으로 인해 기상 변수의 영향을 받는 것으로 나타났다. 또한, 모든 데이터 셋에서 Multiple Linear Regression보다 Random Forest가 더 뛰어난 성능을 보였다.

Ⅰ. 서 론

1.1 연구 배경

현대사회에서 지하철은 시민들에게 중요한 대중교통 수단이다. 국토교통부에 따르면 서울시의 경우, 지하철 이용량 (67.9%)은 다른 이동 수단의 이용량 (32.1%)보다 두 배 이상 높은 수치를 보인다 [1]. 수요에 따른 원활한 배차 간격을 제공하기 위해서는 정확한 이용량 파악이 중요하다. 대중교통 이용량에 영향을 미치는 지표로는 여러 가지 요인들이 존재한다. 그 중 날씨는 사람들의 행동 패턴에 많은 영향을 준다. 이러한 행동 패턴은 지하철 이용량에 영향을 미칠 수 있다. 그러나 기상 데이터와 지하철 이용량의 상관관계를 분석하는 연구는 현저히 적다. 본 논문은 월별, 일별 데이터를 사용한 선행연구와 달리, 시간별 기상 데이터와 지하철 이용량 데이터를 이용하여 이용량을 예측하고자 한다.

1.2 선행 연구 분석

Hee-Jin Kim *et al.*은 월별 서울시의 대중교통 이용량과 기상 데이터를 이용하여 이용량 예측하기 위해 Gaussian Process, Bagging, Random Forest 알고리즘을 활용하였다. 해당 논문에서는 기상 조건으로 기온, 강수량, 미세먼지를 선정하였다. 3개의 알고리즘 중 Gaussian Process가 테스트 세트에서 평균 15.49%로 가장 낮은 오차율을 보여주었다 [2]. 또한, SangGi Choi *et al.*은 기상 조건이 대중교통 수요에 미치는 영향을 알아보기 위해 서울시의 일별 기상 데이터와 대중교통 이용량을 이용하였다. 연구를 위해 Least Square Simple Regression을 사용하였고, R^2를 통하여 적합도를 검증하였다. 기상 조건은 강수량, 불쾌지수, 적설량, 체감온도를 선정하였다. 분석 결과, 강수량과 체감온도에서 R^2가 각각 0.77, 0.75로 좋은 모델임을 보여주었다 [3].

Ⅱ. 데이터 수집 및 전처리

2.1 데이터 수집

본 논문에서는 '공공데이터 포털 [4]'과 '기상자료 개방 포털 [5]'에서 제공하는 데이터를 이용하였다. 날씨 데이터는 '종관기상관측 (ASOS)'를 이용하였다. 지하철 데이터는 '역별 일별 시간대별 승하차 인원 정보'를 이용하였다. 모든 데이터는 2016년 1월 1일부터 2019년 12월 31일까지 4년간의 일별 시간별 데이터이다.

2.2. 데이터 처리

본 논문은 일관성 있는 데이터를 사용하기 위하여 주말과 공휴일을 제외한 주중의 데이터를 이용하였다. 또한, 지하철 데이터에서 제공되는 시간의 범위가 6시부터 23시이기 때문에 기상 데이터도 해당 시간을 제외한 데이터는 제거하였다. 시계열 데이터를 예측하기 위해 표 1과 같이 독립변수의 한 시간 전, 일주일 전의 데이터를 동일 행에 추가하여 독립변수로 사용하였으며 날짜 데이터는 제거하였다. 휴일이 아닌 날의 데이터만 사용하고 있으므로 일주일 전의 데이터가 휴일이라면 2주일 전의 데이터를 사용하였다. 그리고 2주일 전또한 휴일이라면 해당 날짜의 데이터를 모두 삭제하였다.

표 1. 사용 독립변수

독립변수
Rain
Humidity (%)
Sensible_temp (℃)

Wind (m/s)
Hour (h)
Rain(hour)
Humidity(hour) (%)
Sensible_temp(hour) (℃)
Wind(hour) (m/s)
Usage(hour)
Rain(week)
Humidity(week) (%)
Sensible_temp(week) (℃)
Wind(week) (m/s)
Usage(week)

2.2.1 기상 데이터 전처리

본 논문에서는 여러 가지 기상 조건 중 이용량에 영향을 미치는 요인으로 체감온도 (Sensible_temp), 강수 여부 (Rain), 풍속 (wind), 습도 (Humidity) 4가지를 선정했다. 선정된 요인 중 체감온도는 제공되는 데이터셋에 존재하지 않는다. 따라서 기상청에서 제공하는 식에 따라 계산하였다. 체감온도는 여름철 (5~9월)과 겨울철 (10월~익년 4월)에 따라 계산식이 나뉜다. 식 (1)은 여름철 체감온도의 수학적 정의를 나타낸다.

$$sensible\_temp = -0.2442 + 0.55399\,Tw + 0.45535\,Ta - 0.0022\,Tw^2 + 0.00278\,Tw\,Ta + 3.0 \quad (1)$$

이 식에서 Tw는 습구온도, Ta는 기온을 의미한다. 습구온도란 수은주의 끝을 물에 적신 솜으로 감싸 측정한 온도를 말한다. 습구온도는 Stull의 추정식 (2)를 이용하여 계산한다.

$$Ta*ATAN(0.151977*(RH+8.313658)^{0.5}) + ATAN(Ta+RH) - ATAN(RH-1.67633) + 0.00391838*RH^{\frac{2}{3}}*ATAN(0.023101*RH) - 4.686035 \quad (2)$$

여기서 RH는 상대습도 (%)를 의미한다. 여름철과 달리 겨울철은 습구온도 대신 10분 평균 풍속 (km/h)을 이용하여 식 (3)과 같이 계산한다 [6].

$$sensible\_temp = 13.12 + 0.6215\,Ta - 11.37\,V^{0.16} + 0.3965\,V^{0.16}\,Ta \quad (3)$$

지하철은 역 안에서 승·하차를 하는 특성으로 인해 다른 교통수단에 비해서 강수량 및 적설량으로 인한 영향이 적다. 또한, 선행연구와 달리 해당 연구는 시간별 데이터를 이용하였기 때문에 강수량 및 적설량과 이용량의 관계성이 낮았다. 그렇기에 강수량 및 적설량이 아닌 해당 시간에 0.1mm라도 강수 혹은 적설이 되었다면 해당 시간에 강수가 있었다고 판단하였다.

2.2.2 지하철 데이터 전처리

서울의 지하철역 중 본 논문에서는 을지로입구역과 뚝섬유원지역을 선택하였다. 을지로입구역은 주변에 많은 회사가 집중되어 있어 출·퇴근 시간에 이용량이 급증한다는 특징을 가지고 있다. 그와 반대로 뚝섬유원지역은 주택가와 한강공원이 연결되어 있어 주중에는 이용량이 거의 일정하며 날씨가 이용량에 영향을 미친다는 특징을 가지고 있다. 본 논문은 앞서 말한 두 역의 6시부터 23시까지 지하철 이용량의 승·하차 인원수를 각각 분석에 이용하였다.

Ⅲ. 분석 방법

3-1. Multiple Linear Regression

Multiple Linear Regression (이하 'MLR'이라고 한다)은 여러 개의 독립변수와 하나의 종속 변수의 선형 관계를 모델링한다. 단순 선형 회귀와 달리 여러 개의 독립변수를 사용하여 변수 간 상호작용을 반영할 수 있다.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_i x_i + \epsilon \quad (4)$$

식 (4)는 MLR의 일반식으로 ε와 β_0는 각각 오차와 y 절편이고 β_i은 독립변수 x_i에 대한 종속 변수의 증분을 나타내는 값으로 회귀계수에 해당한다.

3-2. Random Forest Regression

Random Forest Regression은 여러 개의 의사 결정 트리를 결합하여 회귀 작업에 대해 정확한 예측을 수행하는 앙상블 기계 학습 알고리즘이다. 다양한 데이터에 대해 안정적인 예측 성능을 제공하며 비선형 관계를 갖는 복잡한 데이터 셋에서 유용하게 활용되는 특징이 있다. 또한, 일부 속성을 선택해서 학습에 이용하기 때문에 다른 알고리즘에 비해 과적합이 발생할 확률이 낮다. 본 논문에서는 tree depth를 10으로 고정하였고 number of model을 1500으로 고정하였다.

3-3. 성능 평가 지표

본 논문은 각 알고리즘에 대한 성능 평가 지표로 R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) 총 4개를 사용한다. R^2는 독립변수들의 종속변수에 대한 설명력을 보여주는 결정계수이다. R^2의 범위는 0에서 1이며 값이 1에 가까울수록 적합도가 높다고 할 수 있다. MAE는 알고리즘의 예측값과 실제값의 차이의 절대값의 평균으로, 절대값을 취하기 때문에 가장 직관적으로 알 수 있는 지표이다. 하지만 절대값을 취하기 때문에 알고리즘에 overfitting인지 underfitting인지 알 수 없다는 단점이 있다. MSE는 MAE와 다르게 절대값 대신 제곱을 한 값의 합이고 제곱을 하기 때문에 이상치를 감지하기 쉽다는 장점이 있다. RMSE는 MSE에 루트를 씌운 값으로 지표가 예측변수와 같기 때문에 직관적이라는 장점이 있다. MSE와 RMSE 모두 예측변수의 크기에 영향을 크게 받는다는 단점이 있다.

Ⅳ. 본론

을지로입구역과 뚝섬유원지역의 주요 독립변수를 알아보기 위해 역별 승·하차 이용량 (Usage)과 나머지 독립변수 간의 Linear Correlation과 Random Forest의 Feature Importance를 표 2와 같이 구했다.

표 2. Usage와 독립변수 간의 상관관계와 및 특성 중요도

Linear Correlation	Euljiro 1(il)-ga		Ttukseom Park	
Features	Boarding	Exiting	Boarding	Exiting
Rain+B34:J65	0.028	0.011	0.040	0.073
Wind	0.183	0.119	0.141	0.153
Humidity	0.141	0.122	0.156	0.114
Sensible_temp	0.048	0.079	0.056	0.179
Hour	0.612	0.58	0.459	0.487
Rain(hour)	0.033	0.014	0.041	0.074
Wind(hour)	0.269	0.169	0.160	0.228
Humidity(hour)	0.243	0.198	0.192	0.199
Sensible_temp(hour)	0.086	0.099	0.052	0.214
Usage(hour)	0.644	0.51	0.641	0.332
Rain(week)	0.030	0.016	0.012	0.043
Wind(week)	0.193	0.13	0.138	0.181
Humidity(week)	0.154	0.142	0.224	0.077
Sensible_temp(week)	0.053	0.083	0.030	0.154
Usage(week)	0.975	0.993	0.889	0.849
Random Forest	tree depth: 10 number of model: 1500			
Features	Boarding	Exiting	Boarding	Exiting
Rain	0	1	4	4
Wind	89	68	53	65

Humidity	62	51	114	22
Sensible_temp	11	8	34	63
Hour	252	215	256	283
Rain(hour)	5	1	1	3
Wind(hour)	177	150	93	190
Humidity(hour)	128	175	140	64
Sensible_temp(hour)	35	36	21	150
Usage(hour)	213	278	248	211
Rain(week)	3	2	0	0
Wind(week)	124	103	68	119
Humidity(week)	71	105	175	11
Sensible_temp(week)	17	11	7	24
Usage(week)	313	296	286	291

일주일 전 이용량과 시간의 경우, 같은 요일의 데이터이기 때문에 고정적 이용객이 존재하는 대중교통의 특성상 상관관계가 높게 나온다. 또한, 전체 데이터 중 비가 오지 않는 시간의 비율이 97%이기 때문에 강수 여부, 한 시간 전 강수 여부, 일주일 전 강수 여부의 상관관계가 매우 낮게 나왔다. 각 역에서 상관관계가 높은 상위 8개를 선정하여 분석한 결과, 각 역의 승·하차 시 고려되는 변수가 변하는 것을 알 수 있다. Linear Correlation의 경우에 을지로입구역과 뚝섬유원지역 모두 1 ~ 3개의 독립변수가 달라지고 Feature Importance는 뚝섬유원지역이 2개가 달라진다. 이를 통해 이용객들이 승·하차 시 고려하는 조건이 다르다는 것을 알 수 있다. 을지로입구역의 승·하차와 뚝섬유원지역의 승차는 전체적으로 비슷한 상관관계 양상을 보이는 반면, 뚝섬유원지역의 하차는 현재 체감온도와 한 시간 전 체감온도가 상대적으로 높은 상관관계를 보인다. 뚝섬유원지역에 하차하는 경우의 상당수가 여가 활동이며 서울 내 역 대부분에서 한 시간 내로 뚝섬유원지역에 도착할 수 있다. 따라서 뚝섬유원지역의 하차 이용객 수는 출발할 때의 기상 조건을 영향을 받게 되기 때문에, 한 시간 전 체감온도의 상관관계가 높게 나온다. 역에 도착했을 시에도 기상에 따라 다른 역으로 다시 이동할 가능성이 높기 때문에 현재 체감온도와 높은 상관관계를 가지게 된다. 이러한 이유로 본 논문은 두 역의 승·하차 예측 결과를 분석하기 위해 8개의 독립변수를 새로 정했다. 선정 기준은 공통적으로 높은 상관관계가 나타난 Hour, Usage(hour), Usage(week), Wind(hour), Wind(week), Humidity(hour)와 대부분의 경우에 포함되는 Wind, Humidity(week)를 선정하여 총 8개의 독립변수를 사용하였다.

표 3. 독립변수의 개수에 따른 예측 결과

	Euljiro 1(il)-ga Boarding		Euljiro 1(il)-ga Exiting	
	15 variables			
	MLR	Random Forest	MLR	Random Forest
R^2	0.949	0.968	0.984	0.986
MAE	256.465	271.369	188.545	245.759
MSE	463428.246	288197.650	193202.858	170837.378
RMSE	680.756	536.840	439.548	413.325
	8 variables			
R^2	0.948	0.976	0.984	0.989
MAE	250.907	230.990	188.821	204.479
MSE	471987.413	211695.463	193339.998	128374.148
RMSE	687.013	460.104	439.704	358.293
	Ttukseom Park Boarding		Ttukseom Park Exiting	
	15 variables			
	MLR	Random Forest	MLR	Random Forest
R^2	0.836	0.927	0.798	0.899
MAE	82.976	53.481	83.666	60.393
MSE	17360.718	7703.532	21294.975	10677.527
RMSE	131.760	87.770	145.928	103.332
	8 variables			
R^2	0.827	0.931	0.822	0.899
MAE	84.815	51.595	79.238	57.304
MSE	18224.725	7313.890	17471.379	9888.176
RMSE	134.999	85.521	132.179	99.439

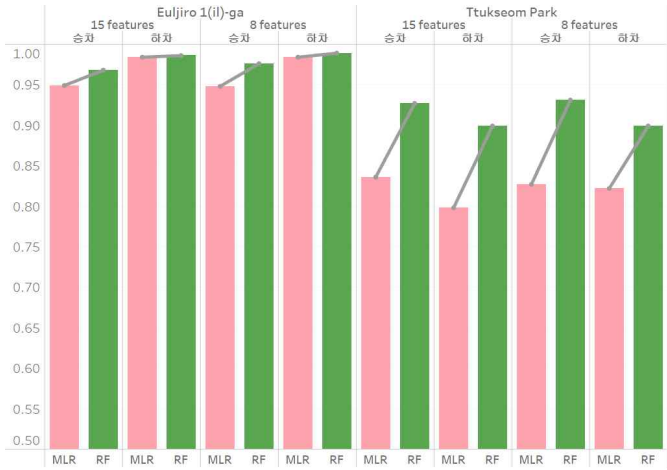

그림 1. 독립변수의 개수에 따른 두 역의 알고리즘별 R^2 차이

15개의 독립변수와 선정된 8개의 독립변수를 이용하여 MLR과 Random Forest 모델을 통한 시간에 따른 기상데이터와 사용량의 분석 결과를 표 3과 그림 1로 표시하였다. 승하차와 관계없이 역끼리 비교 분석하였을 때, 을지로입구역은 뚝섬유원지역에 비해 MLR의 성능이 더 높았다. 이를 통해서 상대적으로 뚝섬유원지역의 데이터들이 선형성이 낮은 것을 알 수 있다. 이러한 원인은 역 별 주 이용객의 특징을 들어 설명할 수 있다. 을지로입구역의 경우에는 주 이용객의 특성으로 인해 날씨의 영향을 덜 받기에 독립변수 간의 상관관계가 높지 않다. 또한 교통 혼잡시간에 의해, 시간에 따라 사용량이 특정한 패턴으로 변화한다. 반면, 뚝섬유원지역의 주 이용객은 공원에서 여가시간을 보내기 위한 사람들이므로 역 이용량은 날씨에 영향을 더 받기 때문에 독립변수 간의 상관관계가 높다. 이는 오히려 MLR의 성능을 떨어트리는 요인이 될 수 있기에 MLR보다 Random Forest가 더욱 유의미한 예측값을 도출한 것이 설명된다. 그림 1에서 볼 수 있듯이, 상위 8개의 독립변수만을 이용한 MLR 결과는 뚝섬유원지역의 하차를 제외한 나머지는 모두 R^2값이 같거나 소폭 하락하였다. 일주일 전 이용량을 포함한 상위 3개의 변수가 너무 높은 상관관계를 가지고 있어 다른 변수가 전혀 영향을 주지 못한다. 하지만 Random Forest의 경우에는 모두 결과가 같거나 향상되었다. MAE, MSE, RMSE 값이 전체적으로 줄어들어 성능이 향상되었음을 보여준다. 이를 통해 전체적으로 MLR보다 Random Forest의 성능이 높다는 것을 알 수 있다.

Ⅴ. 결론

본 논문은 날씨에 따른 시간별 지하철 이용량 예측을 위해 Multiple Linear Regression, Random Forest 알고리즘을 사용하였다. 기상에 따른 시간별 이용량을 예측하기 위해 서로 다른 특징을 가지고 있는 두 역을 선정하여 독립변수별 상관관계를 그렸다. 을지로입구역은 고정된 수요가 강하기 때문에 기상 데이터의 영향이 적게 나타났다. 반면에 뚝섬유원지역의 경우에는 고정적인 수요층이 적어 기상 데이터의 영향이 크게 나타났다. 해당 연구를 통해 지역별 특성에 따라 고려해야 할 변수가 다르다는 것을 알 수 있다. 향후에는 다양한 지역 및 기상 변수를 추가하여 지역별 특성에 따른 이용량 예측을 연구할 예정이다.

ACKNOWLEDGMENT

본 과제 (결과물)은 교육부와 한국연구재단의 재원으로 지원받아 수행된 3단계 산학연협력 선도대학 육성사업 (LINC 3.0과 2023년도 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구 결과입니다. 또한, Qualcomm Institute, University of California San

Diego 소속이신 한창희 박사님의 멘토링에 감사드립니다.

참 고 문 헌

[1] Ministry of Land Infrastructure and Transport, Public transportation usage status at a glance, Retrieved July. 16, 2023, https://www.molit.go.kr/USR/NEWS/m_35045/dtl.jsp?lcmspage=1&id=95085966

[2] Hee-Jin Kim, Sujin OH, and Ung-Mo Kim, "A Study on the Prediction of Public Transportation Consumption in Seoul by Weather," in *KIPS 2017 Autumn Academic Presentation*, pp. 656-659, Seoul, Korea, Nov. 2017.

[3] Sang Gi Choi, Jong Ho Rhee, and Seung Hwoon Oh, "The Effect of Weather Conditions on Transit Ridership," Journal of the Korean Civil society, vol. 33, no. 6, pp. 2447-2453, Nov. 2013.

[4] Seoul Transportation Corporation, Seoul Transportation Corporation_Information on the number of people getting on and off by daily time zone by station (2023), Retrieved July., 15, 2023, from https://www.data.go.kr/data/15048032/fileData.do#layer_data_infomation

[5] Korea Meteorological Administration, Longitudinal Meteorological Observation (ASOS) (2023), Retrieved July., 16, 2023, from https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36

[6] Korea Meteorological Administration, Sensible Temperature , Retrieved July. 16, 2023, from https://data.kma.go.kr/climate/windChill/selectWindChillChart.do?pgmNo=111