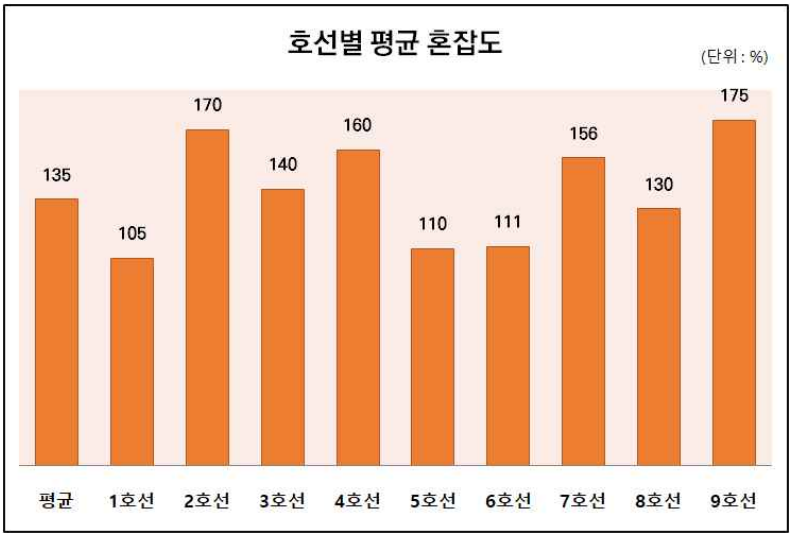
2019 Farm 경진대회 결과보고서 - 빅데이터팜

<table>
<tr><td>프로젝트명</td><td colspan="5">9호선 열량 대비 유동인구 수 분석 을 통한 9호선 추가 열량 예측</td></tr>
<tr><td>수행책임교수</td><td colspan="5">송윤석 교수님</td></tr>
<tr><td>프로젝트 기간</td><td colspan="5">2019.10.31. ~ 2019.11.26</td></tr>
<tr><td rowspan="7">참여인원</td><td>구분</td><td>성명</td><td colspan="3">정보</td><td>팀장/팀원</td></tr>
<tr><td></td><td></td><td>소속</td><td>연락처</td><td>이메일</td><td></td></tr>
<tr><td>교수</td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td rowspan="5">학생</td><td>이윤정</td><td>통계학과</td><td></td><td></td><td>팀장</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td colspan="2">프로젝트 개요</td><td colspan="4"></td></tr>
</table>

1) 추진 배경

서울지하철 9호선은 현재(2019.12. 기준) 개화~중앙보훈병원을 노선으로 가지며, 서울의 동서를 잇는, 중심부에 위치한 지하철이다. 따라서 다른 지하철보다 수요가 많으며 특히 출퇴근시간대에 매우 혼잡하여 '지옥철'이라는 별명을 가지고 있다. 서울지하철 9호선의 평균 혼잡도 175%(2018.12. 기준)는 서울시 지하철 중 가장 높은 혼잡도를 가지며, 출근시간대 혼잡도는 급행기준 172%로 1~8호선 중 혼잡도가 높은 2호선 170.3%, 4호선 159.7%, 7호선 155.9%와 비교했을 때도 가장 높은 수치를 가진다(2019.9. 기준). 서울메트로 9호선의 혼잡도 문제를 해결하기 위해 열차의 열량을 4량에서 6량으로 늘리거나 배차간격을 줄이는 등 노력을 해 왔으나, 2/3단계 노선 연장(언주역~종합운동장역/삼전역~보훈병원역) 후 수요예측에 실패해 혼잡도는 개선되지 않았다. 따라서 본 조는 열량증설, 배차간격 조정 등의 해결방안을 제시할 때 보다 정확한 수요예측을 통해 시간대별 혼잡도를 예측해야 할 필요성을 느꼈다.



[그림 1.] 호선별 평균 혼잡도

2) 추진 목표

본 프로젝트는 폭발적인 9호선의 유동인구와 아직 다른 호선에 비해 부족한 9호선의 열량 문제로 인해 9호선 이용에 불편함을 겪고 있는 탑승객들 에게 편리 제공에 목적이 있다. 이와 같은 목적을 달성하기 위하여 설정한 문제는 다음과 같다.

1. 9호선의 열량이 기존 4, 6량에서 8량으로 증가된다면 실제로 탑승객에게 편리를 제공하는가?
2. 열량 증설이 불가능하다면 9호선의 배차 간격을 조정함으로써 탑승객에 게 편리를 제공할 수 있는가?

이상의 과제를 수행하기 위해 열량수 변화에 따른 '시기별 유동인구 데이터분포를 확인한다. 또한 9호선의 열량이 증설되었을 때 9호선의 혼잡도를 예측하는 모델을 생성하여 6량에서 8량으로 추가증설 되었을 때의 9호선 유동인구 분포를 예측하고자 한다.

2. 최종결과물

본 조가 9호선 탑승객에게 편리를 제공하기 위해서 설정한 문제를 만족하기 위한 연구 가설은 다음과 같다.

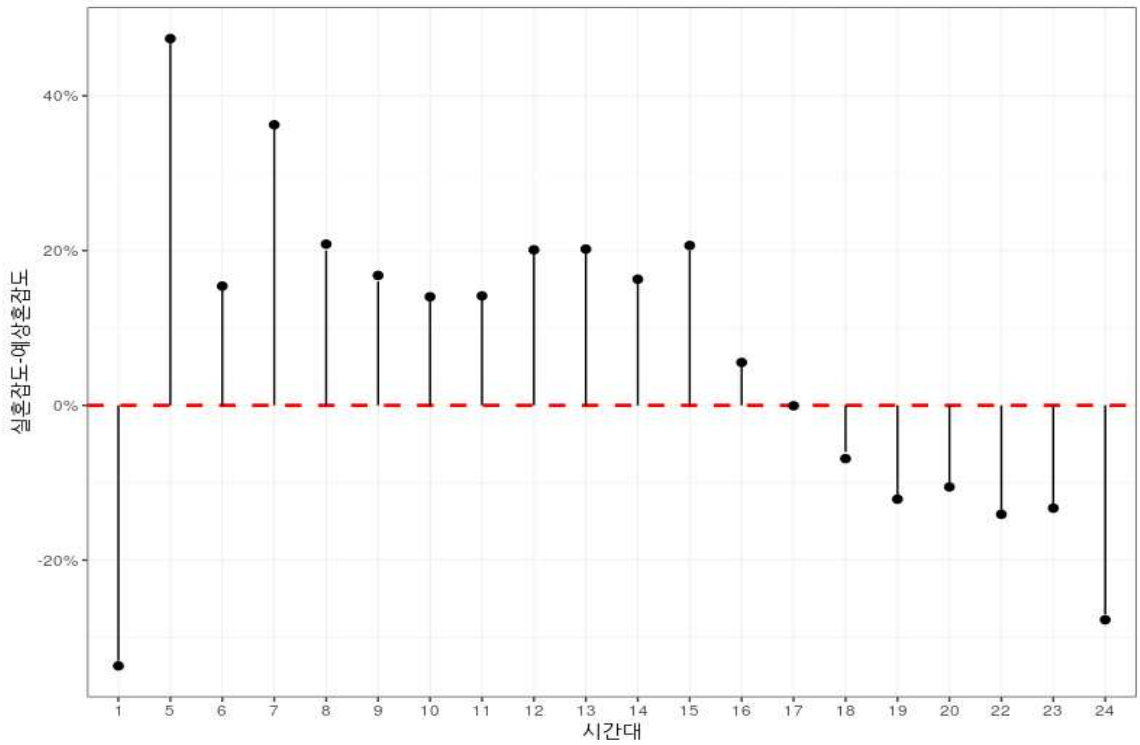
"9호선 열차의 열량이 증가한다면, 한 회 운행 시 수송할 수 있는 인원이 증가하므로 혼잡도가 감소할 것이다."

하지만, MLP Regressor 모델과 Liner Regression 모델을 이용하여 도출한 결과를 살펴보면, 이용자의 수가 적은 시간대에는 혼잡도가 감소하였으나 , 그와 대비되게 이용자의 수가 많은 시간대에는 오히려 혼잡도가 증가한 결과를 볼 수 있었다.

구할 수 있던 데이터가 한정적일 뿐만 아니라, 승객수와 혼잡도 간의 양의 상관관계가 높은 것을 보았으나 데이터 전처리 과정에서 두 변수의 상관관계를

고려하지 않고, 각각의 독립된 변수로 설정하였다. 이를 미루어보아 9호선의 혼잡도와 열차 증설은 상관이 없다는 결론을 내릴 수 있었고 추후 다른 분석 기법을 사용하여 9호선 승객들의 편의를 높일 수 있는 또 다른 방법을 고려 해보고자 한다.

본 프로젝트에서는 예상한 결과를 낼 수 없었지만 승객의 편의와 지하철의 증설이 관계없다는 결과 자체로도 충분한 분석이 되었다고 생각하며 추후에는 열차의 배차시간과 노선과 같은 데이터를 사용하여 분석을 해보려 하며 이에 따른 승객의 편의성, 즉 혼잡도에 직접적인 영향을 끼칠 수 있을지에 대하여 다시 한 번 연구할 시간을 가질 계획이다.



[그림 1.] 실 혼잡도와 분석을 통한 혼잡도의 평균 사이의 관계

3. 추진내용

1) 데이터 파악 및 수집

먼저, 혼잡도를 알기 위해 역별 승하차인원이 필요하여 '공공데이터포털(https://www.data.go.kr)'에서 **역별 시간대별 승·하차 인원(월별)** 데이터를 받아왔다. 그리고 **9호선 지하철 시간표**, **9호선 열량 변화 시기 및 시간별 운행 열량** 정보를 'metro9(https://www.metro9.co.kr)'에서 받아왔다.

각 데이터는 다음과 같다.

① 역별 시간대별 승·하차 인원(월별)

	사용월	호선명	지하철역	04시-05시 승차인원	04시-05시 하차인원	05시-06시 승차인원	05시-06시 하차인원	06시-07시 승차인원	06시-07시 하차인원	07시-08시 승차인원	...	23시-24시 하차인원	00시-01시 승차인원	00시-01시 하차인원	01시-02시 승차인원	01시-02시 하차인원	02시-03시 승차인원	02시-03시 하차인원	03시-04시 승차인원	03시-04시 하차인원	작업일자
15497	201909	9호선	양천향교	3	0	4310	856	9137	7314	26539	...	10094	426	3301	3	118	0	0	0	0	20191003
15498	201909	9호선	국회의사당	4	0	1038	1664	1895	9525	3697	...	2418	968	849	4	6	0	0	0	0	20191003
15499	201909	9호선	구반포	4	0	568	173	1606	1809	4775	...	1917	172	685	3	43	0	0	0	0	20191003
15500	201909	9호선	종미	0	0	2216	718	6877	2754	19027	...	5512	224	1777	2	35	0	0	0	0	20191003
15501	201909	9호선	흑석(중앙대입구)	6	1	3361	975	5936	4651	17856	...	10297	1462	3133	2	91	0	0	0	0	20191003

32711 rows × 52 columns

② 9호선 시간표

김포공항, 중앙보훈병원 방면	시각	
분 (일반/*급행)		분 (일반/*급행)
30 42 54	05	
06 18 29 40 51	06	
00 08 17 27 37 47 57	07	
07 18 29 40 51	08	
02 13 24 35 46 57	09	
08 19 30 41 52	10	
03 14 25 37 49	11	
00 11 22 33 44 55	12	
06 17 28 39 51	13	
03 14 25 36 47 58	14	
09 20 31 42 53	15	
00 04 15 26 33 37 48 57	16	
05 14 22 31 40 48 57	17	
06 15 24 33 42 51 59	18	
08 17 26 36 46 56	19	
06 16 26 36 46 56	20	
06 16 26 36 46 56	21	
07 18 29 40 51	22	
02 13 24 37 50(삼전)	23	
03(신논현) 16(동작) 29(당산) 42(가양)	24	
	01	

[그림 2.] 예시_개화역 지하철시간표(출처: metro9(https://www.metro9.co.kr))

2) 데이터 전처리

1. 결측치 처리

먼저, 역별 시간대별 데이터(월별)에서 '호선명'이 9호선인 데이터만 취하여 역별 시간대별·월별 승차인원수를 가져온다. 그 중 시간대가 04시~05시 열은 첫차 시간이 5시 이후이므로 결측치 처리하여 삭제하였다.

2. 행과 열 전치

앞서 결측치 처리한 데이터는 각 시간대가 열로 설정되어있다. 시간대를 하나의 변수로 설정하기 위해 역별 변수가 시간대, 승차인원수인 데이터셋을 생성하였다.

	time	month	people_num
0	1	6	0
1	1	1	1
2	1	8	0
3	1	3	0
4	1	10	0
…	…	…	…
29699	19	12	14830
29700	20	12	12724
29701	22	12	11621
29702	23	12	7194
29703	24	12	1426

3. 역별 시간대별 승차인원수 데이터와 열량 별 열차 수 데이터 join

9호선 열차의 열량 변화 시기를 나누면 다음과 같다.

날짜	일반열차	급행열차
201501~201712	4량	4량
201801~201811	4량	6량 3편성
201812~201910	4량	6량

[표 1.] 9호선 열차의 열량 변화 시기

열량 변화 시기를 기준으로 역별 시간대별·월별 승차인원수에 일반열차 수, 4량 급행열차수, 6량 급행열차수를 join한다.

	time	month	people_num	norm	fast_4	fast_6
0	1	6	0	2	1	0
1	1	1	1	2	1	0
2	1	8	0	2	1	0
3	1	3	0	2	1	0
4	1	10	0	2	1	0
…	…	…	…	…	…	…
29699	19	12	14830	13	0	0
29700	20	12	12724	13	0	0
29701	22	12	11621	12	0	0
29702	23	12	7194	12	0	0
29703	24	12	1426	11	0	0

4. 혼잡도 join

먼저 혼잡도에 대한 식은 다음과 같다.

$$Congestion = Passengers / Capacity \times 100 \ldots (1)$$

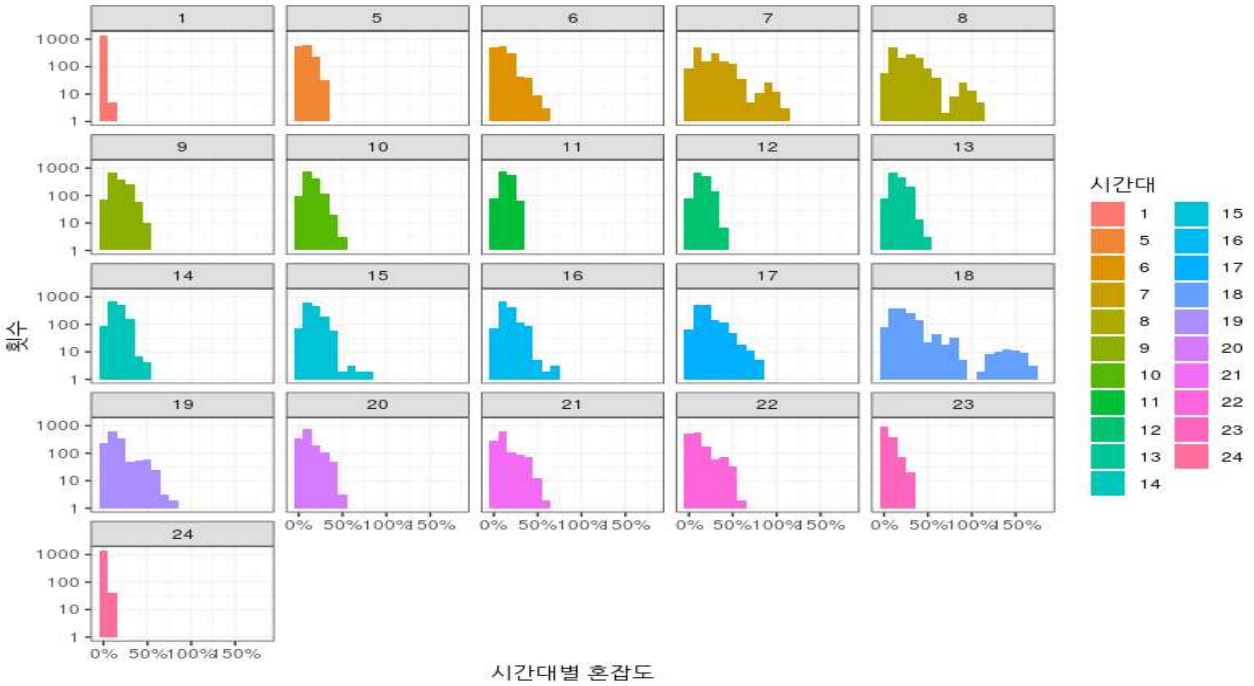
(Congestion : 혼잡도, Passengers : 재차인원수, Capacity : 차량정원)

여기서, 재차인원수는 좌석인원과 입석인원의 총합으로 정확히 계수하기 어렵다. 따라서 본 조는 승차인원수를 재차인원수로 대체하였다.

위의 혼잡도 식(1)을 이용해 혼잡도를 계산한 다음, 위의 데이터에 join한다.
join한 데이터는 다음과 같다.

	confuse_num	time	month	people_num	norm	fast_4	fast_6
0	0	1	6	0	2	1	0
1	0.001	1	1	1	2	1	0
2	0	1	8	0	2	1	0
3	0	1	3	0	2	1	0
4	0	1	10	0	2	1	0
…	…	…	…	…	…	…	…
29699	1.882	19	12	14830	13	0	0
29700	1.615	20	12	12724	13	0	0
29701	1.598	22	12	11621	12	0	0
29702	0.989	23	12	7194	12	0	0
29703	0.214	24	12	1426	11	0	0

시간대별 혼잡도 데이터는 아래와 같다.



[그림 3.] 시간대별 혼잡도

3) 모델 선정

1. 선형회귀모델(Linear Regression)

(1) 회귀모델(regression model) : 어떤 자료에 대해서 그 값에 영향을 주는 조건을 고려하여 구한 평균.

통계학적인 관점에서 보면 모든 데이터는 아래와 같은 수식으로 표현할 수 있다고 가정.

$$y = h(x_1, x_2, x_3, \ldots, x_k; \beta_1, \beta_2, \beta_3, \ldots, \beta_k) + \varepsilon \ldots (2)$$

위 수식에서 h()가 위에서 말한 조건에 따른 평균을 구하는 함수이며 이것을 보통 '회귀 모델'이라고 부른다. 이 함수는 어떤 조건(x1,x2,x3,...)이 주어지면 각 조건의 영향력($\beta_1, \beta_2, \beta_3, \ldots$)을 고려하여 해당 조건에서의 평균값을 계산해준다. 뒤에 붙는 ε 는 '오차항'을 의미한다. 이론적으로 ε 는 평균이 0이고 분산이 일정한 정규분포를 띄는 성질이 있다.

(2) 회귀분석(regression analysis) : 변수들 간의 함수적인 관련성을 규명하기 위하여 어떤 수학적 모형을 가정하고, 이 모형을 측정된 변수들의 자료로부터 추정하는 통계적 모델.

회귀분석을 한다는 것은 위의 식(2)에서 h()함수가 무엇인지를 찾는 과정을 의미한다. 추정한 회귀모델이 h()라고 정확히 알 방법은 없다. 다만 그럴 것이라고 어느 정도는 확신할 수 있는 방법이 있는데, 바로 우리가 만든 회귀 모델의 예측치와 실측치 사이의 차이인 '잔차(residual)'가 정말 우리가 가정한 오차항(ε) 의 조건을 충족하는지 확인하는 것이다. 이런 확인 작업을 '모델 검정'이라고 부른다.

회귀계수에 따라 회귀모델이 선형이냐 비선형이냐를 결정되는데, 대부분의 회귀모델

은 선형회귀모델이다.

(3) 선형회귀모델 종류

[그림 4.] 선형회귀모델 종류

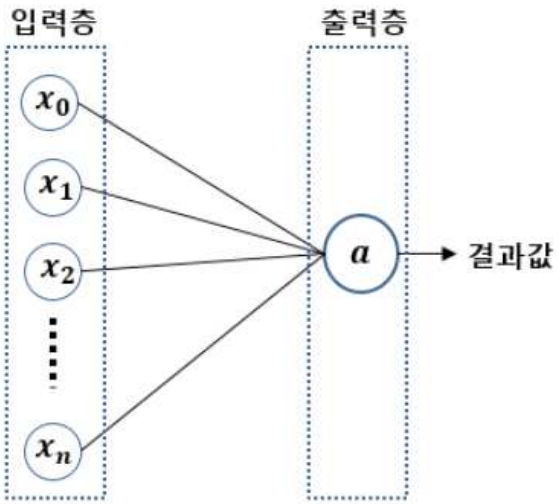
(4) 다중선형회귀모델

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_q x_q + \varepsilon$$

(β_0: 상수항 계수, β_j:j=1,...,p: 회귀계수: 다른 변수는 고정시키고 x_j를 1단위 증가 시켰을 때의 y의 증가량)

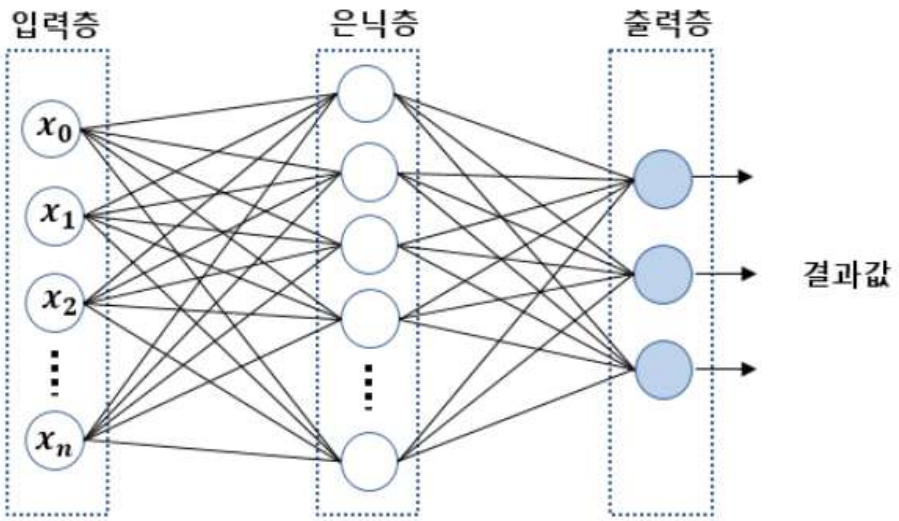
2. MLP(Multilayer Perceptron)

(1) 인공신경망인 단층 퍼셉트론은 비선형적으로 분리되는 데이터에 대해서는 제대로 된 학습이 불가능하다는 한계가 있다.



[그림 5.] 단층 퍼셉트론

이를 극복하기 위한 방안으로 입력층과 출력층 사이에 하나 이상의 중간층을 두어 비선형적으로 분리되는 데이터에 대해서도 학습이 가능하도록 다층 퍼셉트론이 고안되었다.

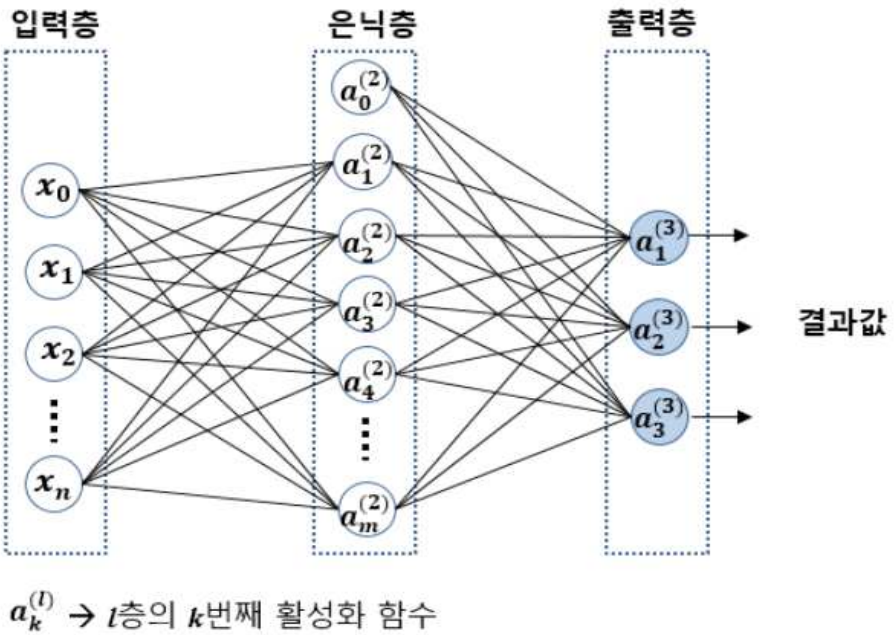


[그림 6.] 다중 퍼셉트론의 한 예

입력층과 출력층 사이에 존재하는 중간층을 숨어 있는 층이라고 해서 은닉충이라고 부른다. 입력층과 출력층 사이에 여러 개의 은닉층이 있는 인공 신경망을 심층 신경망(deep neural network)이라 부르며, 심층 신경망을 학습하기 위해 고안된 특별한 알고리즘들을 딥러닝(deep learning)이라고 부른다.

(2) 정의 : 딥러닝 방법 중 하나인 다층 지각(MLP)은 일련의 입력에서 일련의 출

력을 생성하는 인공 신경망이다.



[그림 7.] 입력층의 노드수가 n개, 은닉층의 노드수가 m개, 출력층의 노드수가 3개인 n-m-3 다중 퍼셉트론

(3) 순전파(feedfoward) : 입력층에서 전달되는 값이 은닉층의 모든 노드로 전달되며 은닉층의 모든 노드의 출력값도 출력층의 모든 노드로 전달된다. 이 형식으로 값이 전달되는 것을 순전파라고 한다.

(4) 다중퍼셉트론의 동작 원리

1. 각 층에서의 가중치를 임의의 값(보통 0)으로 설정한다. 각 층에서 bias 값은 1로 설정한다.
2. 하나의 트레이닝 데이터에 대해서 각 층에서의 순입력 함수값을 계산하고 최종적으로 활성 함수에 의한 출력값을 계산한다.
3. 출력층의 활성 함수에 의한 결과값과 실제값이 허용 오차 이내가 되도록 각층에서의 가중치를 업데이트한다.
4. 모든 트레이닝 데이터에 대해서 출력층의 활성 함수에 의한 결과값과 실제값이 허용 오차 이내가 되면 학습을 종료한다.

4) 모델 구현

본 프로젝트에서 모델은 위에 설명한 선형회귀모델과 MLPRegressor를 사용했다. 회귀를 바탕으로 하는 모델을 선정한 이유는 종속변수가 범주형 변수가 아닌 연속형 변수이기 때문이다. 모델의 input인 데이터는 R프로그램을 사용하여 전처리를 하였고,

2970×48 의 구조를 가진다. input데이터 분석은 파이썬에서 sklearn 패키지에서 제공하는 모델들을 사용하여 진행했다.

1) 선형회귀분석

sklearn패키지 내의 LinearRegression 모델을 사용하여 종속변수는 혼잡도, 나머지 7개의 변수들은 독립변수로 사용했다. 독립변수끼리의 상관관계를 확인하였으나, 양의 상관관계가 높은 승객 수 변수와 혼잡도 변수의 관계를 수정하지 못하였다. 따라서 다중공선성이 의심될 수 있다. 이 점은 개선점으로 남겨둬야 할 것 같다.

2) MLPRegressor(다중 퍼셉트론 회귀)

sklean패키지 내의 MLPRegressor 모델을 사용하여 선형회귀분석과 마찬가지로 종속변수는 혼잡도, 나머지 7개의 변수들은 독립변수로 사용했다. MLPRegressor는 선형회귀분석처럼 7개의 변수들 각각에 대한 선형관계를 파악하여 종속변수 값을 예측하는 모델이지만, 선형회귀분석과는 달리 중간에 활성함수를 두어 예측값에 비선형성을 부여하였고, 여러 개의 층을 사용하여 feature engineering을 모델 내부에서 수행하게 된다. 따라서 선형회귀분석보다 향상된 성능을 기대할 수 있다.

5) 보완점

본 프로젝트는 세 가지 한계점을 가진다. 첫 번째는 혼잡도 데이터 일부에서 수정치를 사용하였다는 점이며 두 번째는 선형회귀분석 모델에서 독립변수 간의 다중공선성 문제를 해결하지 못했다는 점, 그리고 세 번째는 다중 공선성을 해결하기 위해 제안한 MLPRegresser 모델이 추론된 결과에 대한 해석을 제공하기 어렵다는 점이다.

본 프로젝트에서 사용된 데이터는 해당 역의 이용자가 태그한 교통카드로부터 비롯된 데이터이므로 일반열차와 급행열차가 구분되지 않은 데이터이다. 하지만 급행열차는 유동인구 수가 많은 호선의 주요 역들만 거쳐가기 때문에 해당 프로젝트에서 목표를 달성하기 위해선 일반열차와 급행열차 간의 구분이 반드시 필요하였다. 때문에 일반열차와 급행열차를 구분 짓기 위해 역별 시간표를 통해 해당 시간대에 해당 역을 지나가는 일반열차와 급행열차의 운행 횟수를 사용하여 추정치로 대체하였다.

또한, 2018년 12월 1일 이전 데이터들은 9호선 3단계 노선 연장 이전 역별 시간표 데이터가 존재하지 않아, 현재 역별 시간표와 연장 이전 운행 횟수를 통해 추정치를 대체하였다.

선형회귀분석 모델에서의 다중공선성은 모델의 예측 정확도를 저하시킨다. 본 프로젝트에서 제안하는 모델에서는 독립변수 간의 양의 상관관계가 존재함을 확인하여 다중공선성 문제의 발생 가능성을 확인하였으나, 해석 가능한 모델을 제안하기 위하여 독립변수를 그대로 두고 사용하였다는 한계점이 있다.

다중 공선성 문제는 MLPRegresser를 사용할 경우 해결 가능하다. MLPRegresser는 비선형성과 다층 구조를 가지게 되어 모델 내재적인 feature engineering을 기대할 수 있고 이는 다중 공선성 문제를 해결할 수 있다. 그러나 복잡한 모델 구조로 인해 모델의 추론 결과에 대하여 설명이 불가능하다는 단점이 있다.

4. 기대효과

9호선은 구간 연장 시 수요 예측에 실패하여 서울지하철 최고의 혼잡도를 기록했다. 이를 해결하기 위해 차량을 늘렸지만 역시 잘못된 수요예측으로 혼잡도에는 큰 영향을 미치지 못했다. 본 프로젝트는 9호선 수요 및 혼잡도 예측을 목표로 모델을 생성하였으며, 이를 통해 '서울 메트로 9호선'에서 혼잡도 개선을 위한 열량 증설이나 배차 시간 간격 조정 대책을 마련할 때 혼잡도를 줄일 수 있는 최소 열량, 적정 배차 시간 간격을 예측하여 경제적 손실을 최소화한 대책을 마련할 수 있을 것이다.

5. 주요성과

본 프로젝트를 통하여 전반적인 데이터 수집과 전처리 과정에 대해서 알 수 있었고, 이를 통한 기계학습, 데이터 분석에 대해 보다 깊이 알 수 있는 시간 이였으나 중간과정에서 좀 더 구체적인 기법에 관한 지식이 부족하여 놓친 부분이 아쉬웠다.

5. 지출내역

해당사항 없음.