2017 서울시 빅데이터 캠퍼스 상시 공모전

지하철 사업 수익성 개선을 위한 승객 수 예측

팀명 : KUBIG
팀원 : 목충협, 이성길, 천우진

KUBIG

목차

KUBIG

지하철 현황 조사

지하철 이용 현황

지하철 이용객 수 (서울시)

*출처-서울특별시 교통통계참고

16년도 평균 일간 지하철 이용객 수는 799만 9000명으로
서울시의 총 인구인 993만 1000명의 86%에 달할 정도로
매우 높은 수치를 보인다.

지하철 사업 현황

각 호선별 순이익(손실)

출처-20151102연합뉴스

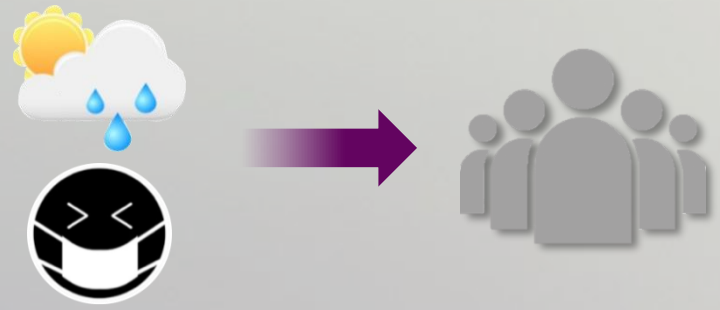
2호선(365억), 9호선(31억)을 제외하고
모든 호선이 총 적자 4215억원을 기록하였다. (2014년 기준)
특히 3호선(-1118억), 5호선(-913억)은 매우 높은 적자를
기록하였다.

KUBIG

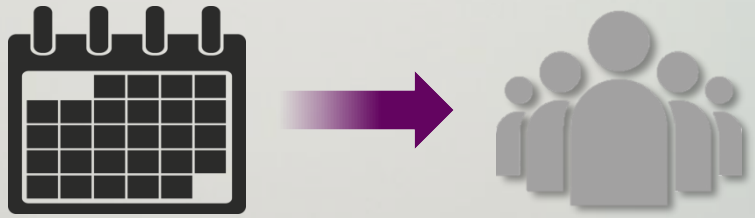
분석 목적

지하철 사업의 현황을 통하여 수익성 개선의 필요성을 느꼈고
여러 모델을 사용하여 승객 수 예측을 통하여 이에 기여하고자 함

대기환경, 기상관측정보를 활용한 승객수 분석



시계열 데이터를 통한 승객수 분석



사용 데이터

서울특별시 빅데이터 캠퍼스
-빅데이터 공유활용플랫폼 데이터셋-

대중교통 이용통계

운행일자	호선ID	호선	역ID	역	승차시간구분	30분시간구간ID	승차총승객수	하차총승객수
2014-02-01	003	3호선	0340	가락시장	00:00:00~00:59:59	00	2	19
2014-02-01	003	3호선	0340	가락시장	00:00:00~00:59:59	30	2	18
2014-02-01	003	3호선	0340	가락시장	01:00:00~01:59:59	00	0	4
2014-02-01	003	3호선	0340	가락시장	01:00:00~01:59:59	30	0	1

시간범위 : 2014.01 ~ 2015.10
내용 : 서울철도 일별/시간대별 승하차 인원을 호선,역별로 집계한 데이터.

기상관측정보

SAWS_OBS_TM	STN_ID	STN_NM	SAWS_TA_	SAWS_HD	CODE	NAME	SAWS_WS	SAWS_RN	SAWS_SOI	SAWS_SHI
2009030821	1154	도봉	4.7	34.3	9	남	0.5	0		
2009030821	1155	마포	4.3	58.1	14	서북서	0.7	0		
2009030821	1156	구로	5.1	49.9	12	서남서	1.8	0		
2009030821	1158	서초	6.3	37.7	11	남서	1.1	0	15.2	7.6

시간범위 : 2009.01.14 ~ 2017.09.11
내용 : 서울시 기상관측소별 실시간 기온, 습도, 풍향, 풍속, 강수, 일사, 일조 정보

대기환경정보

MSRDATE	MSRADM(	GRADE	MAXINDE	POLLUTAN	NITROGEN	NITROGEN	OZONE	OZONEIN(	CARBON	CARBONIN	SULFUROU	SULFUROU	PM10	PM10INDE	PM24	PM24INDE	MSRRGNC	MSRRGNN	MSRSTEN/
201401010000	111121	보통	79	PM-10	0.047	78	0.003	5	1.2	30	0.006	15	46	46	59	79	100	도심권	중구
201401010000	111262	보통	89	PM-10	0.048	80	0.002	3	0.8	20	0.006	15	47	47	69	89	104	동남권	서초구
201401010000	111131	보통	76	PM-10	0.042	70	0.004	7	0.5	13	0.005	13	30	30	56	76	100	도심권	용산구
201401010000	111221	보통	79	PM-10	0.044	73	0.006	10	0.6	15	0.004	10	35	35	59	79	103	서남권	구로구

시간범위 : 2014.01 ~ 2015.10
내용 : 서울시 25개 자치구의 실시간 대기환경정보. 오존, 이산화질소, 일산화탄소, 미세먼지(PM-10)등의 정보.

KUBIG

데이터 전처리

**세 가지 데이터셋을 통합하여 각 지하철역의 대기환경정보,
기상관측정보, 총 승객 수를 포함한 데이터셋으로 전처리를 진행.**

데이터 통합

대중교통 이용통계의 시간범위인 2014.01 ~ 2015.10의 데이터 통합

승객수
시간별 -> 일별 통합

기상관측
시간별 -> 일별 통합

대기환경
시간별 -> 일별 통합

KUBIG

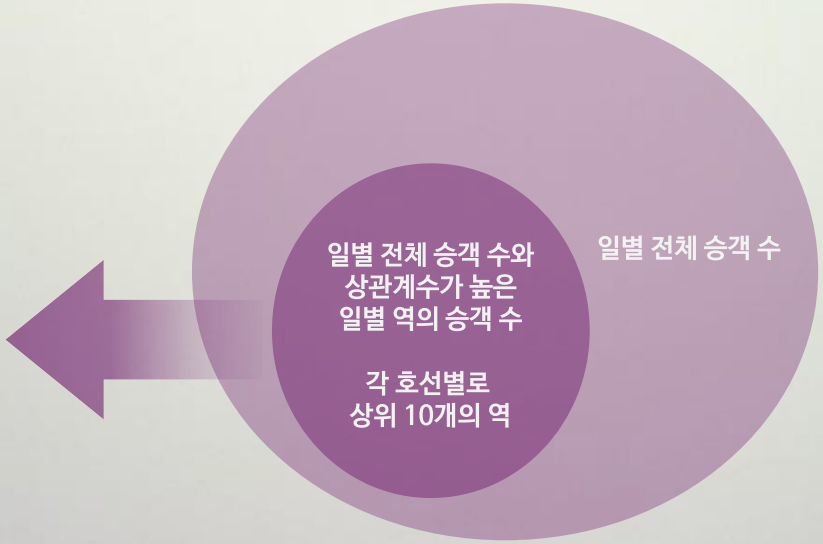
데이터 전처리

**세 가지 데이터셋을 통합하여 각 지하철역의 대기환경정보,
기상관측정보, 총 승객 수를 포함한 데이터셋으로 전처리를 진행.**

데이터 선별

각 호선별로 전체 승객 수와 상관계수가 높은 상위 10개의 역을 선별

전체 승객 수를 예측하는 것은
비효율적이므로
전체 승객 수와 상관계수가
높은 역을 선별하여
관객수 예측에 사용.



데이터 전처리

세 가지 데이터셋을 통합하여 각 지하철역의 대기환경정보, 기상관측정보, 총 승객 수를 포함한 데이터셋으로 전처리를 진행.

데이터 선별

각 호선별로 전체 승객 수와 상관계수가 높은 상위 10개의 역을 선별

	1호선	2호선	3호선	4호선	5호선	6호선	7호선	8호선	9호선
1위	종로5가	신천	연신내	총신대입구	까치산	합정	이수	잠실	노량진
2위	종로3가	신도림	독립문	창동	영등포시장	태릉입구	상봉	산성	신논현
3위	신설동	강남	홍제	수유(강북구청)	발산	응암	사가정	천호	당산
4위	종각	신림	구파발	당고개	오목교	신당	신중동	암사	공항시장
5위	동대문	서울대입구	옥수	성신여대입구	군자	약수	먹골	송파	신방화
6위	서울	왕십리(성동구청)	신사	미아사거리	왕십리(성동구청)	증산	반포	신흥	노들
7위	시청	합정	금호	노원	동대문역사문화공원	석계	노원	가락시장	구반포
8위	청량리(지하)	아현	종로3가	회현	종로3가	구산	군자	석촌	가양
9위	제기동	잠실	대치	동대문역사문화공원	화곡	삼각지	부천시청	문정	등촌
10위	동묘앞	동대문역사문화공원	잠원	미아	우장산	망원	철산	몽촌토성	염창

데이터 전처리

**세 가지 데이터셋을 통합하여 각 지하철역의 대기환경정보,
기상관측정보, 총 승객 수를 포함한 데이터셋으로 전처리를 진행.**

파생변수 생성

날짜별로 요일변수를 추가

DATE	ADDRESS	STATION	users	호선	temp	humid	wind	rain	pm10	mon	tue	wen	thu	fri	sat	sun
2014-01-08	종로	종로5가	60838	1호선	1.488463	63.636	1.388	0	211.625	0	0	1	0	0	0	0
2014-01-09	종로	종로5가	55705	1호선	9.136082	34.044	1.656	0	101.7917	0	0	0	1	0	0	0
2014-01-10	종로	종로5가	59952	1호선	4.885385	53.33636	0.836364	0	152.2083	0	0	0	0	1	0	0
2014-01-11	종로	종로5가	52634	1호선	0.08045	47.845	0.765	0	218.0833	0	0	0	0	0	1	0
2014-01-12	종로	종로5가	23590	1호선	1.557712	46.3	1.711111	0	223.375	0	0	0	0	0	0	1

시간 범위 : 2014.01 ~ 2015.10
일별 승객 수, 온도, 습도, 풍속, 강수량, 미세먼지(PM-10)
+ 요일변수(해당요일이면 1, 아니면0)를
전체 승객 수와 상관계수가 높은 역별로 집계한 데이터셋

대기환경, 기상관측정보를 활용한 승객 수 분석
-Multiple Linear Regression-

MLR 모델 구성

독립변수 → 종속변수

기상,대기, 요일변수 → 승객수

user = temp+humid+wind+rain+pm10
+mon+tue+wed+thu+fri+sat+sun

변수선택 - Stepwise Selection

변수선택 결과

user = temp+humid+~~wind+rain+pm10~~
+mon+tue+wed+thu+fri+sat+sun

최종 MLR 모델

user = temp+humid+mon+tue+wed+thu+fri+sat+sun

KUBIG

대기환경, 기상관측정보를 활용한 승객 수 분석
-Multiple Linear Regression by Tensorflow-

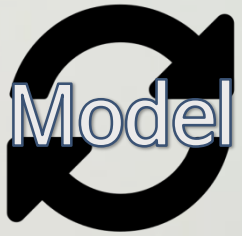
Tensorflow - 비용함수 설정

〈최종 MLR모델〉

user = temp+humid+mon+tue+wed+thu+fri+sat+sun

각 변수의 Coefficient에 따라
변하는 MSE를 비용함수로 설정

Tensorflow - 비용함수 최소화



학습률을 바꿔가며 비용함수를
최소화하는 Coefficient 탐색

Tensorflow - Coefficient 결정

40만번 반복결과 22만번째 이후부터 수렴하는 모습을 보임.
수렴한 값으로 Coefficient 결정

KUBIG

대기환경, 기상관측정보를 활용한 승객 수 분석
-최종 MLR 모델 결정-

MSE값 비교

MLR

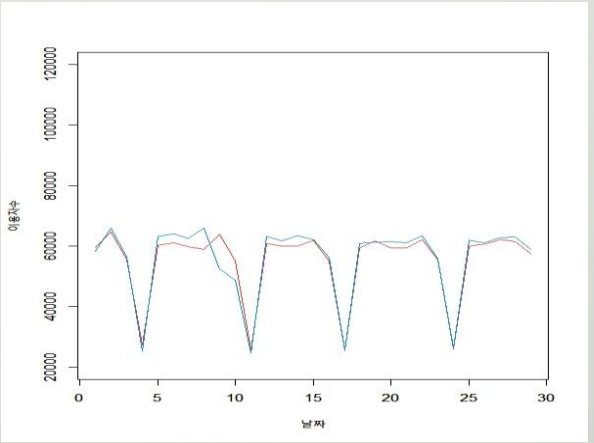
MLR by Tensorflow

종로 5가역의 2015년 10월 한달 간의 이용자 수 예측

10336346 선택 < 45115132

실제값과 예측값의 비교

〈종로 5가역 2015.10 이용자 수〉



KUBIG

시계열 데이터를 통한 승객 수 분석

시계열 분석

승객수의 시계열 데이터 → 승객수

시계열에 따른 승객 수의 패턴 분석을 통해 승객 수를 예측

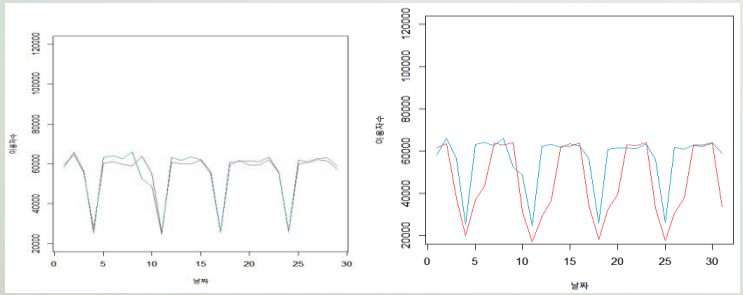
시계열 분석 방법

- 시계열 데이터를 정상시계열로 바꾸기 위해 데이터에 log를 씌우고 차분 과정을 거침
- auto.arima 함수를 이용해 arima모델을 최적화하는 변수를 탐색
- arima모델을 만들고 주기를 일주일로 입력해서 31일치를 예측
- 예측한 값을 다시 자연로그의 밑인 e의 제곱으로 계산하여 승객 수를 예측

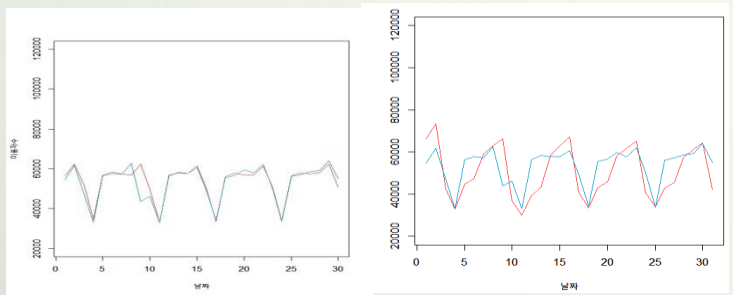
KUBIG

최종 모델 결정
- MLR vs 시계열 예측 결과 비교 (MSE) -

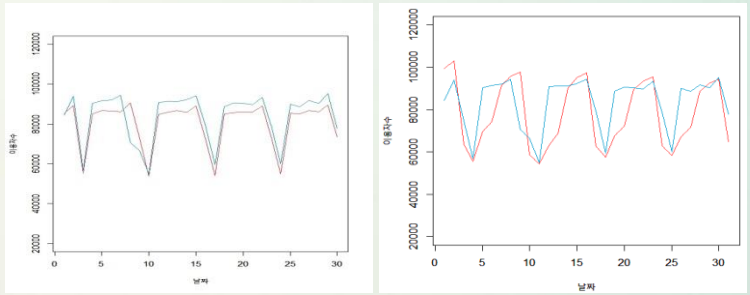
종로5가10336346 vs 271681692



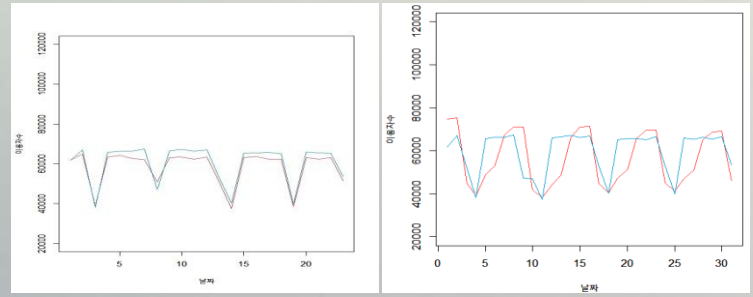
신천 15273158 vs 86227644



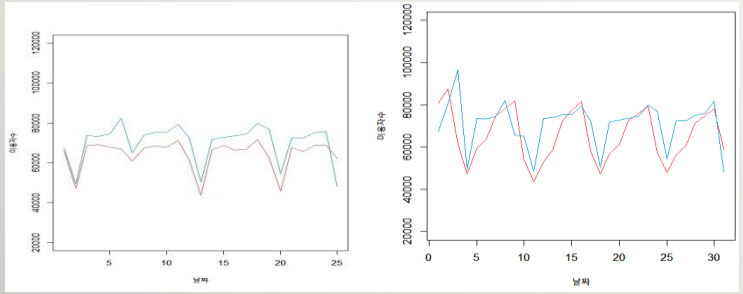
연신내 37073758 vs 178873453



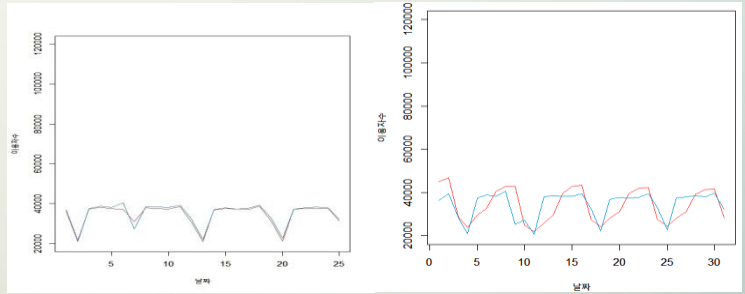
까치산 8356211 vs 113493422



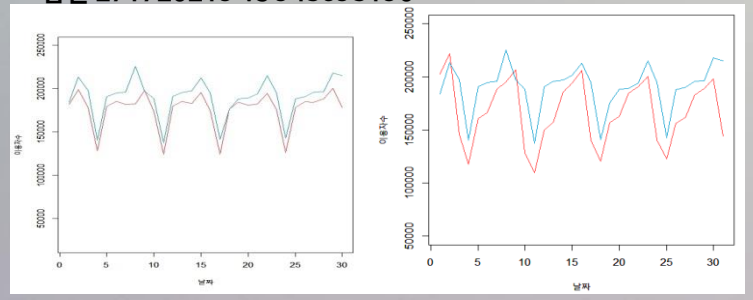
노량진 63441488 vs 141148174



이수 1592286 vs 39177521



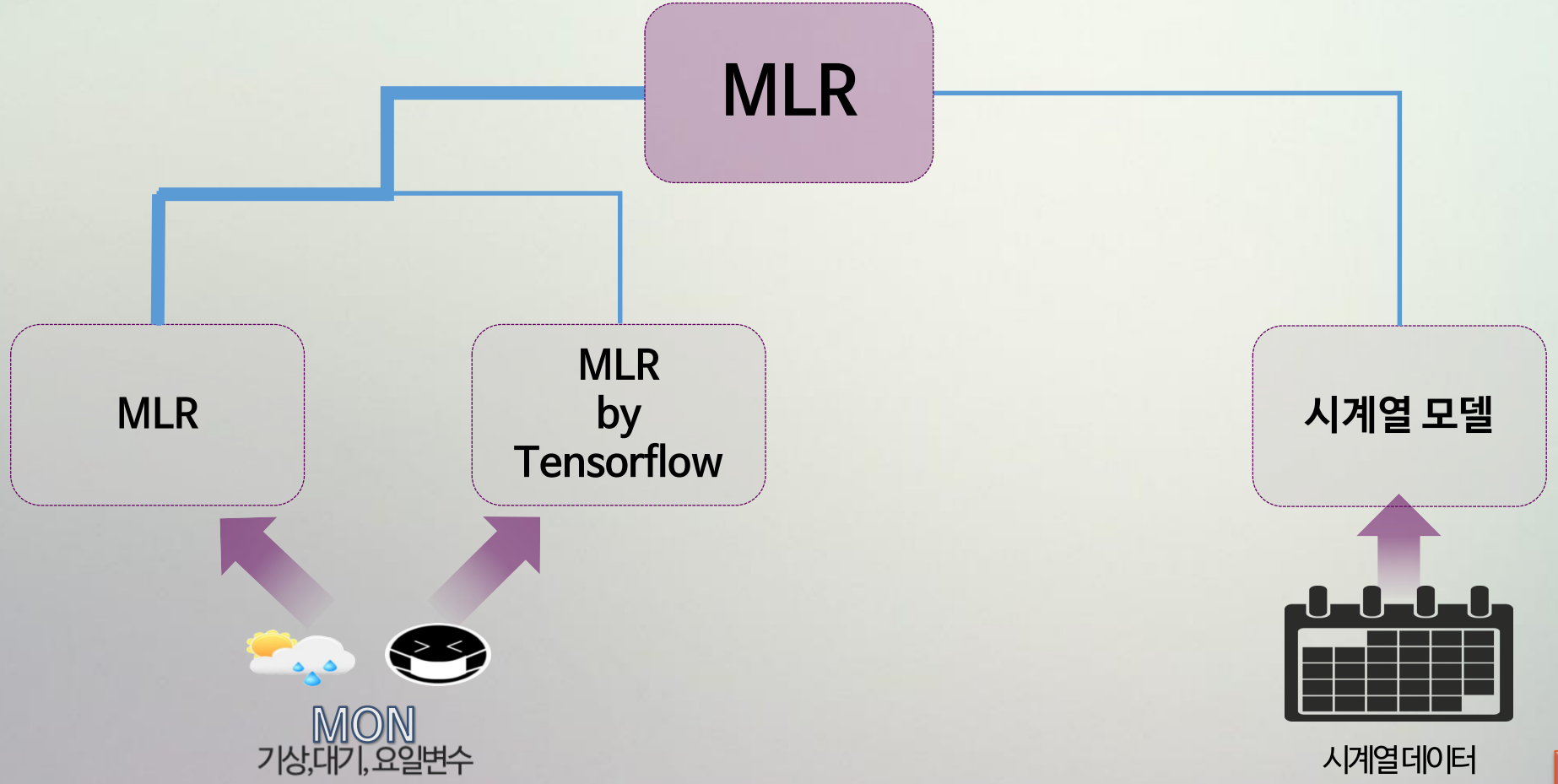
잠실 271720219 vs 943693190



전체 승객 수와 상관계수가 높은 상위 7개 역에 대해
14.01~15.09의 데이터로 training 한 뒤
15.10 한달 치 예측결과를 비교한 결과
모든 경우에 MLR의 MSE가 더 적다.

좌측 : MLR
우측 : 시계열
-관측치
-예측치

최종 모델 결정
-최종 모델-

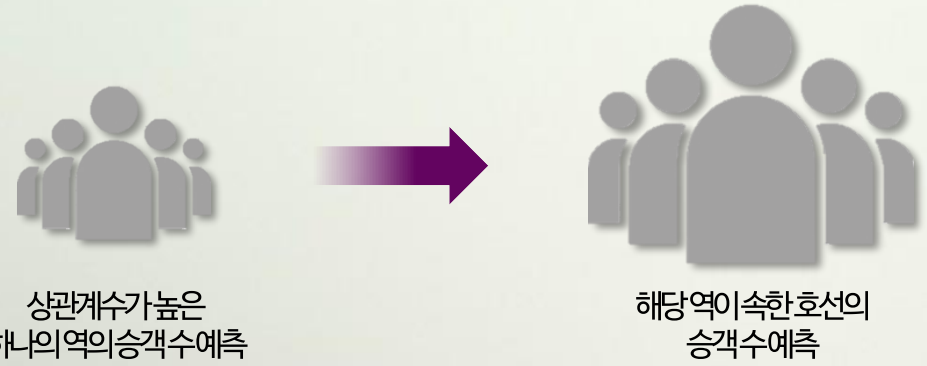


예측 결과
-최종 모델을 이용한 예측결과-

날짜	10/23	10/24	10/25	10/26	10/27	10/28	10/29
예측(명)	38766	31320	21290	37256	37582	37844	37679
실제(명)	39283	32912	22602	37272	37915	38405	38065
오차율(%)	1.31	4.84	5.80	0.04	0.88	1.46	1.01

최종 선정된 MLR모델을 사용하여 3호선 연신내 역의
2015년 10월 마지막 주의 이용객을 예측하였다.
5% 내외의 오차율을 가진다.
연신내 역은 3호선 전체 승객 수와 상관관계가
높은 역중에 하나이므로 이를 통하여
3호선의 전체 승객 수 경향을 파악할 수 있다.

기대 효과 및 활용 방안



- **하나의 역에 대한 승객 수 예측**으로 해당 호선의 승객 수 경향을 예측할 수 있으므로 **효율적**이다.

- 승객 수 예측에 따라 **배차간격을 조절**하여 **승객들에게 만족감**을 줄 수 있으며 **불필요한 배차**를 줄여 **수익성 개선**에 기여할 수 있다.

- **버스 승객 수**와의 **연계분석**을 통하여 버스의 배차 간격 또한 조절할 수 있을 것이다.

한계점

상관계수가 높은
하나의 역의 승객수 예측

해당 역이 속한 호선의
승객수 예측

- 요일변수에서 **공휴일**을 반영하지 못하였다.
- 온도변수가 띄는 **계절성**을 반영하지 못하였다.

- **출퇴근 인원**은 다른 변수에 영향 받지 않고 **고정적**이므로 이를 고려한 모델링이 필요하다.

- 승객 수 데이터가 **14~15년도**이므로 **현재(18년도)**를 예측하는 시계열 모델을 만들기 위해서는 **최신 데이터**가 필요하다.

활용 정보

01 활용 데이터

서울특별시 빅데이터 캠퍼스
-빅데이터 공유활용플랫폼 데이터셋-

대중교통 이용통계

기상관측정보(기온, 강수등)

대기환경정보

02 분석툴

R & Python

03 참고문헌

None

KUBIG