날씨에 따른 서울 지하철 2호선 혼잡도 예측

김은총, 김주원, 이지현

1. 분석 배경

지하철은 버스와 더불어 서울시 시민들이 많이 이용하는 대중 교통 중 하나이다. 그러나 지하철이 혼잡할수록, 승객들의 스트레스를 유발하고(연합뉴스, 2017), 범죄율 또한 높아진다는 기사가 있었다(이데일리, 2019). 지하철의 혼잡도를 미리 알 수 있다면, 승객들의 스트레스 수준을 미리 예측하고 승객들의 만족도를 높이는 방법을 찾고, 범죄또한 예방할 수 있을 것이라고 생각이 들었다. 또한 기상청 공식 블로그에 따르면 날씨에 따라 대중교통의 선택 비율이 달라짐을 알 수 있었다. 이러한 배경지식을 통해 기상 상태에 따른 지하철 혼잡도를 분석해 보기로 하였다.

2. 분석을 위한 데이터 준비

2.1. 사용한 데이터

날씨에 따른 지하철 2호선의 혼잡도 정도를 알아보기 위해 다음과 같은 데이터를 사용하였다.

- 서울시 지하철 호선별 역별 시간대별 승하차 인원 정보(2017, 2018) (서울 열린 데이터광장 제공)
- 서울시 일별 최고기온, 최저기온, 평균기온 (2017, 2018) (기상청 날씨누리 제공)
- 서울시 일별 일 강수량 (2017, 2018, 기상청 날씨누리 제공)
- 서울시 일별 평균 풍속 (2017, 2018 ,기상청 날씨누리 제공)
- 서울시 일별 평균 상대습도 (2017, 2018, 기상청 날씨누리 제공)
- 서울시 일별 일 최심적설 (2017, 2018, 기상청 날씨누리 제공)

2.2. 데이터 전처리

서울시 지하철 호선별 역별 시간대별 승하차 인원 정보를 통해 일별 출퇴근 시간 이용객

과, 출퇴근이 아닌시간의 이용객을 구분하여 구하였다. 출퇴근 시간은 오전6시부터 10시, 오후 5시부터 9시로 두었다.

일별로 구해진 출퇴근, 비 출퇴근 인원을 각 시간대에 운행하는 지하철 배차 수로 나누어 출퇴근 시간의 평균 탑승객 수와, 비 출퇴근 시간의 평균 탑승객 수를 구하였다.

평균 탑승객 수를 이용하여 혼잡도 변수를 추가해 주었다. 혼잡도는 서울 교통 공사 웹진을 참조하였다.

혼잡도 구간은 다음과 같다.

구간	800 미만	800이상 1600미만	1600 이상 2000 미만	2000이상 2400 미만	2400 이상 2800 미만	2800 이상 3200 미만	3200 이상
혼잡도	1	2	3	4	5	6	7

혼잡도가 구해진 지하철 데이터에 서울시의 일별 기상 데이터를 합쳐 주었다.

3. kNN방법을 이용한 데이터 분석

앞서 데이터 전처리 과정을 통해 분석에 필요한 데이터를 추출하였다. 이렇게 가공된 데이터를 통해 크게 holiday(휴일; 토, 일요일 및 공휴일 category)와 notholiday(평일; 월~금요일)의 범주 안에서 바쁘게 움직이는 user(rushuser)와 바쁘지 않은 user(notrushuser) data를 골자로 표 1과 같이 분석해 연관성을 파악하고자 한다. 단, 휴일에 바쁜 user의 경우는 유의미한 결과값을 가지지 않을 것으로 생각되어 바쁨의 정도를 0으로 처리했다.

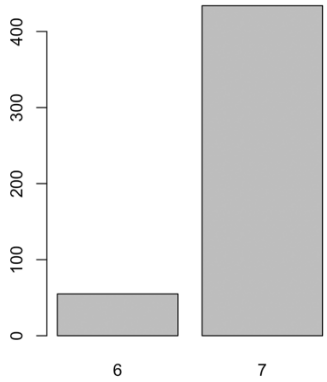
표 1

구분	
평일, 바쁘지 않은 user (notholiday_notrushuser)	평일, 바쁜 user (notholiday_rushuser)

휴일, 바쁘지 않은 user Holiday_notrushuser	휴일, 바쁜 user Holiday_rushuser

3.1. 평일에 바쁘지 않은 user(notholiday_notrushuser)의 데이터 분석

평일에 바쁘지 않은 시간대의 지하철 혼잡도 레벨 분포는 다음과 같다.



Train set과 test set을 8:2의 배율로 나눠주는 과정을 우선적으로 거친다. 그 후, 입력변수와 목적변수 및 na값을 str 함수와 unique함수를 통해 확인한다.

```
# na값 확인.
str(train)
str(test)

sum(is.na(train))
sum(is.na(test))

# 목적변수 확인.
unique(train$notrush_busylevel)
```

다음 과정으로, 앞서 나눠준 train set과 test set의 데이터를 normalization을 통해 각각의 데이터가 같은 범위 내에서 서로 비교될 수 있게끔 만들어주는 과정을 거친다. 또한, 보다 변수의 구분을 용이하게 하기 위하여 각각의 변수 별 이름을 정해 새롭게 변수에 저장한다.

```
# normalization을 해준다.
minmax_norm <- function(x){
  (x-min(x))/(max(x)-min(x))
}

train_norm <- sapply(train[, -9], minmax_norm)
#diagnosis를 제외한 변수들에 minmax_norm 적용.
test_norm <- sapply(test[,-9],minmax_norm)

train_label <- train[, 9]
test_label <- test[, 9]
```

이제 kNN에 사용될 k값을 정해줄 필요가 있다. k값은 일반적으로 데이터셋의 전체 행의 개수에 루트(√) 를 적용한 값으로 다음과 같이 표현될 수 있다.

$$\mathrm{k}_{\mathrm{value}} = \sqrt{nrow}$$

(각각의 변수는 $\mathrm{k}_{\mathrm{value}} = \mathrm{k}$값, nrow는 데이터 셋의 전체 행의 개수)

이를 통해 필요한 k값을 계산해보면,

```
# determine k
sqrt(nrow(train))
```

그 값이 대략적으로 $\mathrm{k}_{\mathrm{value}} = 20.0494$ 로 계산되며, 따라서 대략적으로 20을 우리의 k값으로 선택하였다. 또한, 우리가 정한 k값을 통해 knn을 실행하고 accuracy와 precision 및 recall 값을 계산해보면 표2와 같은 matrix를 얻을 수 있다.

표 2

	Level "6"	Level "7"
Level "6"	13	6
Level "7"	19	48

마지막으로, 표2를 통해 accuracy와 precision, recall을 표3과 같이 계산할 수 있다.

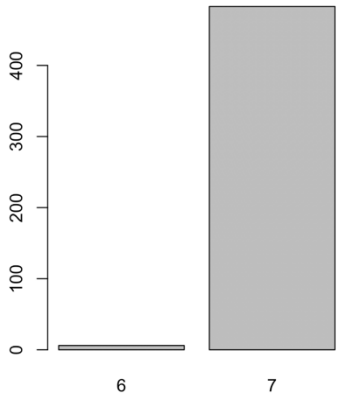
표 3

Overall Accuracy	$\frac{(13+48)}{13+6+19+48} = 0.709$
Precision "6"	$\frac{13}{(13+13+19)} = 0.288$
Precision "7"	$\frac{48}{6+48+48} = 0.470$
Recall "6"	$\frac{13}{(13+13+6)} = 0.406$
Recall "7"	$\frac{48}{(19+48+48)} = 0.417$

표3을 통해 확인할 수 있듯이, 전체적으로 지표가 매우 낮게 나타나고 있다. 이는 날씨와 지하철의 붐비는 정도가 실제로 크게 연관성을 가지지 못함을 보여주며, 이렇게 분석 결과가 나타난 이유로는 2호선이 원래 붐비는 호선인 점이 가장 클 것으로 보인다.

3.2. 평일에 바쁜 user(notholiday_rushuser)의 데이터 분석

평일, 바쁜시간대의 지하철 혼잡도 레벨 분포는 다음과 같다.



평일, 출퇴근시간대의 데이터를 trainset과 testset 으로 8:2로 나눈 뒤, knn분석을 위해 normalization을 거쳤다. 또한 K의 값을

$$k_{value} = \sqrt{nrow}$$

(각각의 변수는 $k_{value} = k_{값}$, nrow 는 데이터 셋의 전체 행의 개수) 의 방법으로 위와 동일하게 구했다.

K의 값은 대략 20으로 구해졌다.

knn을 실행하고 accuracy와 precision 및 recall 값을 계산해보면 표와 같은 matrix를 얻을 수 있다

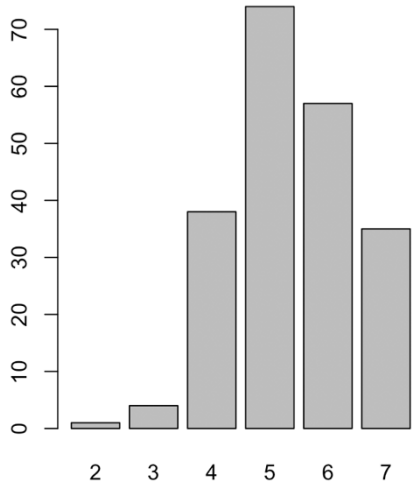
	Level "6"	Level "7"
Level "6"	0	1
Level "7"	0	91

위의 표를 통해 Accuracy를 계산 할 수 있다.

Overall Accuracy	91/92 = 0.99

3.3. 휴일에 바쁘지 않은 user(holiday_notrushuser)의 데이터 분석

휴일, 바쁘지 않은 시간대(전체 시간대)의 지하철 혼잡도 레벨 분포는 다음과 같다.



위 두개와 같은 방식으로, trainset과 testset을 8:2 비율로 나누고, 변수 들을 normalize해준뒤, k를 구한다.

여기서 k의 값은 약 13이 나온다. knn의 결과를 표로 나타내면 아래와 같다.

	Level "1"	Level "2"	Level "3"	Level "4"	Level "5"	Level "6"	Level "7"
Level "1"	0	0	0	0	0	0	0
Level "2"	0	0	0	0	0	0	0
Level "3"	0	0	0	0	0	0	0
Level "4"	0	0	0	3	2	1	0
Level "5"	0	0	0	2	4	6	1
Level "6"	0	0	0	2	4	4	0
Level "7"	0	0	0	0	8	0	0

위의 표를 통해 accuracy를 계산 할 수 있다.

Overall Accuracy	11/37 = 0.30

4. 활용방안

위의 분석만으로는, 서울 지하철 2호선의 승객수와 기상 상태가 상관관계가 있는지 알기 힘들다. 그 이 유는 지하철 데이터셋을 2018년과 2017년 밖에 확보할 수 없었기 때문이라고 생각된다. 따라서 더 많은 데이터셋을 확보하거나, 서울시 일별 지하철 혼잡도 데이터를 확보 할 수 있다면(현재는 월별 혼잡도 밖에 나와있지 않다) 지하철을 이용하는 시민들에게 날씨에 따른 지하철 혼잡도를 예측하여 알려줄 수 있는 지하철 혼잡 지수 모델을 만들 수 있을

것이다.

혼잡지수를 알 수 있다면 지하철 역사 내 관리 인원을 조절하여 역사 내 범죄 예방에 사용 될 수 있을 것이며, 또한 지하철 역사 내의 상인들에게 상품 수요를 예측 하는데 유용하게 사용될 것이다. 또한 시민들에게는 미리 혼잡도를 알고 대중교통이나 자가용 등 다른 교통수단을 유연하게 선택하는것에 도움이 될 것이다.

5. 출처

韓출퇴근 시간 OECD 최대...[웹사이트]. (연합뉴스, 2017-11-24). Retrieved from https://www.yna.co.kr/view/AKR20171116148800797

서울 지하철 범죄 절반이 성범죄...2·9호선에서 많이 발생 [웹사이트]. (이데일리, 2019-01-20). Retrieved from https://www.edaily.co.kr/news/read?newsId=01538326622359688&mediaCodeNo=257

[날씨와 대중교통] 기상조건이 대중교통 이용에 영향을 미친다? [웹사이트]. (기상청 공식 블로그, 2017-05-02). Retrieved from http://blog.naver.com/PostView.nhn?blogId=kma_131&logNo=220996892740

서울시 지하철 호선별 역별 시간대별 승하차 인원 정보[웹사이트]. (서울시 열린 데이터 광장). Retrieved from.(https://data.seoul.go.kr/dataList/datasetView.do?infId=OA-12252&srvType=S&serviceKind=1&currentPageNo=1&searchValue=&searchKey=null)

기상데이터[웹사이트]. (기상자료개방포털). Retrived from(https://data.kma.go.kr/)

서울 지하철 혼잡도 구간 [웹사이트]. (서울시 교통공사 웹진). Retrieved from.(http://webzine.seoulmetro.co.kr/enewspaper/articleview.php?master=&aid=1771&ssid=73&mvid=684)