**<17-2 1조 데이터 마이닝 보고서>**

**1조(김단아, 김준회, 이재호, 이현준)**

**김단아 – 자료 수집 / 탐색 / 분석 / PPT**

**김준회 – 자료 수집 / 탐색 / 분석 / PPT**

**이재호 – 자료 수집 / 보고서 / PPT**

**이현준 – 자료 수집 / 분석 / 보고서 / PPT**

1. **연구의 목적**

현재 서울시에서는 택시 수요 발생시 인접 지역에 미리 대기하고 있는 택시를 보내는 ‘대기식 콜 택시’ 제도를 택하고 있다. 우리는 데이터 마이닝을 이용하여 콜택시 통화건수에 유의하게 영향을 미치는 요소들이 무엇인지 파악하고 구별 콜택시 수요를 미리 예측해보고자 한다. 이를 바탕으로 콜택시 사업체의 비용을 줄이는 동시에 효율적인 경영을 목표로 한다.

1. **데이터 수집**

택시 수요에 영향을 미칠 것이라 생각한 다양한 변수들을 수집하였다. 하루의 날씨가 영향을 많이 미칠 것으로 기대되었다.

1. 전화연결 콜택시 발신 건수[[1]](#footnote-1)

SKT 빅데이터 허브에서 2016년, SKT회선을 통한 서울시의 전화연결 콜택시 서비스의 발신 건수를 얻었다. 데이터는 날짜, 요일, 시간, 발신지 별 발신 건수로 구성되었다.

1. 기상 데이터[[2]](#footnote-2)

기상자료 개방 포털에서 2016년의 기상 데이터를 다운 받았다. 데이터는 날짜 별 평균 기온, 평균 강수량, 평균 습도, 평균 풍속, 평균 미세먼지 수치 등의 자료로 구성되었다.

1. 인구 통계학적 자료[[3]](#footnote-3)

서울 열린 데이터 광장에서 서울시 주민등록인구 데이터를 얻었다. 데이터는 구별 총 주민등록인구수로 구성되었다.

1. 대학가/유흥가/명소/호텔/사업체 종사자 수[[4]](#footnote-4)

서울특별시 지도 및 서울통계정보를 참조하여 각 구마다 대학가/유흥가/명소/호텔/사업체 종사자 수를 입력하였다.

1. 위도 & 경도 정보[[5]](#footnote-5)

시각화 자료를 위한 정보로, 서울 열린 데이터 광장에서 구별 좌표 데이터를 얻었다.

**데이터 요약**

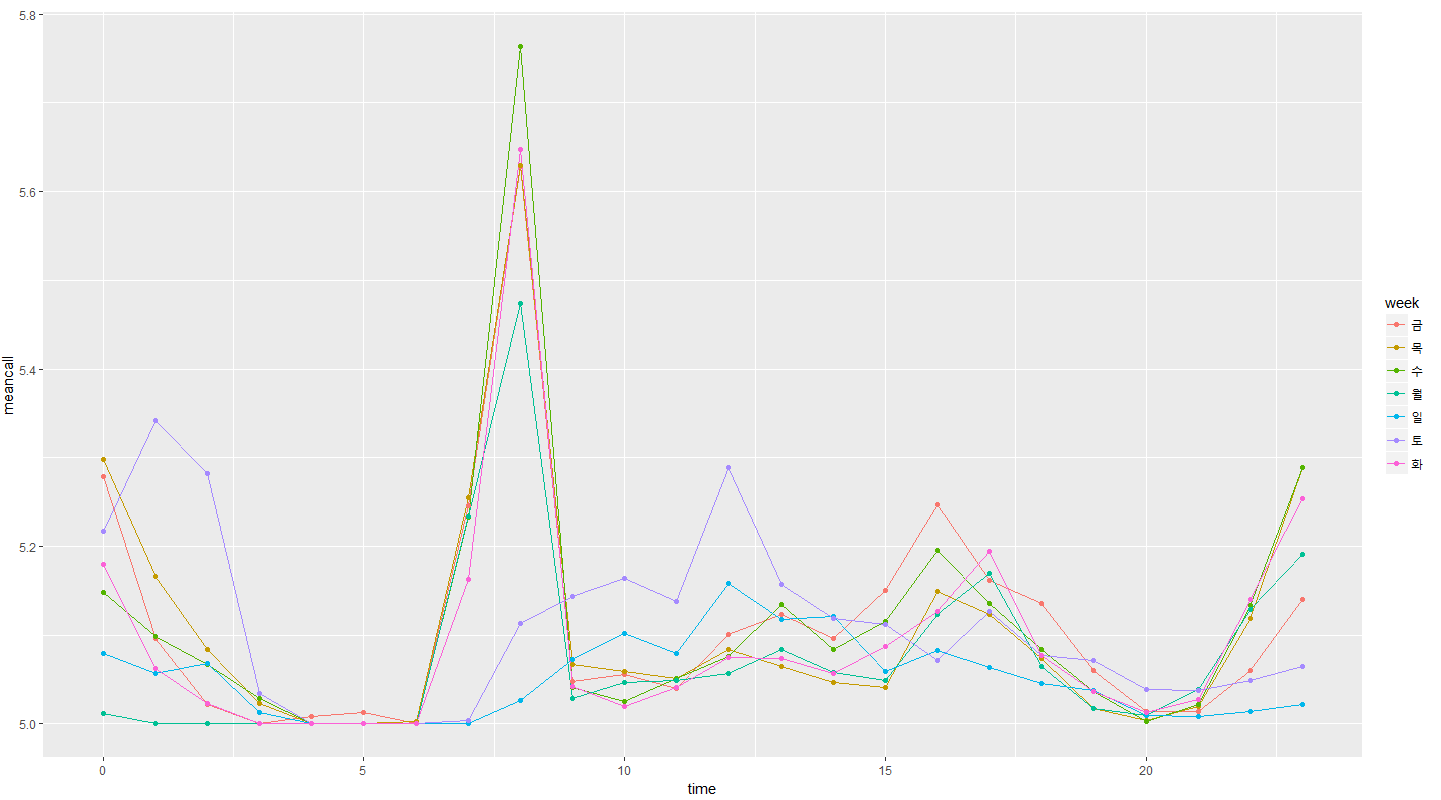
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **변수** | **정보** | **변수** | **정보** | **변수** | **정보** |
| Call | 발신건수 | Avg\_temp | 일 평균 기온 | Goo | 지역구 |
| Date | 날짜 | Min\_temp | 일 최저 기온 | Population | 구별 인구 수 |
| Dayofweek | 요일 | Max\_temp | 일 최고 기온 | Univ | 구별 대학 수 |
| Is\_holiday | 휴일여부 | Day\_rainfall | 일일 강수량 | Tour | 구별 관광명소 수 |
| Time | 시간대  morning/daytime/night | Avg\_wind | 일 평균 풍속 | Worker | 구별 산업 종사자 수 |
| Avg\_dust | 일 평균 미세먼지량 | Pub | 구별 유흥업소 수 |
| Lat/Lon | 구별 위도/경도 | Avg\_humid | 일 평균 습도 | Hotel | 구별 관광호텔 수 |

1. **데이터 탐색 및 전처리**
2. 시간대 별 통화량

24시간 동안의 통화량 분포를 확인한 결과 3개의 구간에서 특징이 발견되었다. 이에 따라

06시~09시 / 10시~21시 / 22시~05시를 각각 morning / daytime / night로 분할하여 진행하였다.

구간의 길이가 다르므로 target 변수 발신건수(call)를 1시간당 통화 건수(meancall)로 변경하였다.



1. 상관관계 분석

연속형 변수 간의 상관관계 도표를 그려 변수 변환, 군집 분석을 시도하였다.

1. **데이터 모델링**
2. 초기 모델 선정



선정 기준

* 1. 연속형 타겟 변수 분석 모델
  2. 예측력이 높은 모델
  3. 비선형 회귀 모델

🡪 위 기준을 만족하는 모델로 **신경망** 모델이 가장 적합하다. 신경망 분석의 특성 상 해석의 어려움이 있지만, 의사결정나무 등을 이용하여 간접적으로 해석한다.

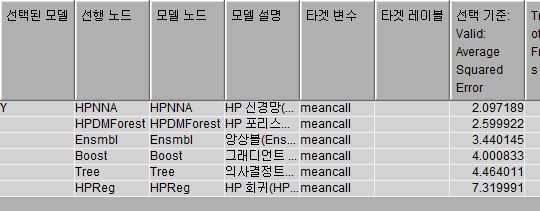
1. 신경망 모델 빌딩
2. 방향

- 자료의 분산 축소 🡪 z-score 이용 표준화

- 너무 많거나 적지 않은 변수 필요 🡪 의미 있는 변수 선택 필요

1. 변수 선택 과정
2. 회귀분석으로 선택된 변수를 이용한 모형 비교

선택된 변수 : univ, time, population1, day of week, is\_holiday, day\_rainfall



🡪 초기 모델에 비해 Average Squared Error 하락

1. 랜덤 포레스트 모형의 중요 변수를 이용한 모형 비교

중요한 변수 : worker, pub, time, tour, population1, univ, hotel, day of week, is\_holiday, day\_rainfall



🡪 (ㄱ)에 비해 Average Squared Error 하락

1. 둘을 결합한 변수 선택

공통된 변수 : univ, time, population1, day of week, is\_holiday, day\_rainfall

유의하다고 판단되는 변수 : tour, worker, pub



🡪 (ㄴ)에 비해 Average Squared Error 0.04가량 상승

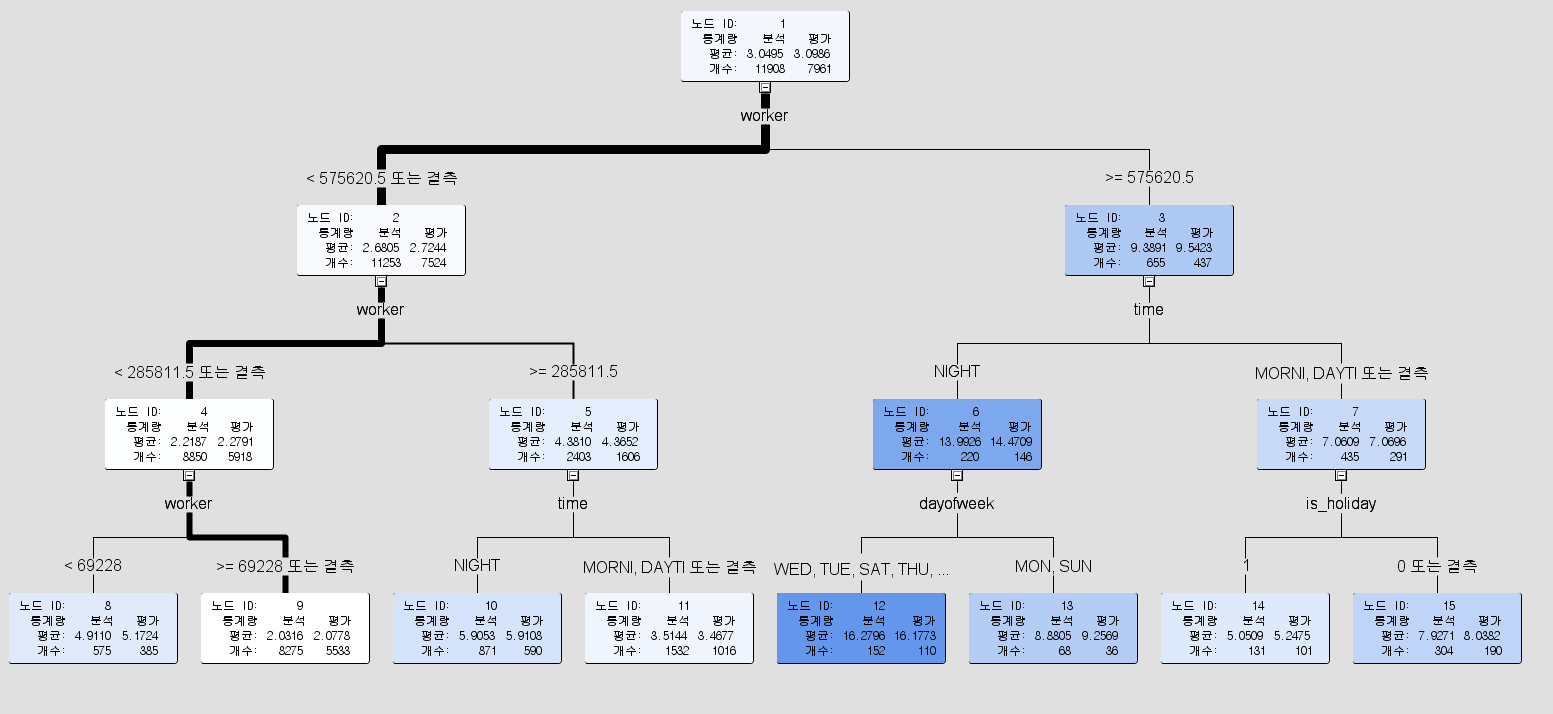
1. 모델 선택

앞의 변수 선택에 따라 모델의 error가 점차 감소하는 것을 확인할 수 있었으며, 마지막으로 선택한 변수의 분석 결과 error가 0.04가량 상승하였으나 변수 1개를 제거하는 것이 더 경제적일 것이라 판단하였다.

이에 따라 <univ, time, population1, day of week, is\_holiday, day\_rainfall, tour, worker, pub>을 포함한 신경망 모델을 최종 모델로 선택하였다.

1. **모델 해석**

최종 모델에 선택된 변수를 이용한 의사결정나무를 다음과 같이 그릴 수 있었다.



1) 첫번째로 worker가 제일 중요변수이다. worker가 많으면 시간당 통화량이 증가하는 경향을 보인다.

2) 두번째로 time이 중요하다. night에 시간당 통화량이 증가하는 경향을 보인다.

3) 그 다음으로 요일(dayofweek)과 공휴일(is\_holiday)이 중요하다.

1. **결론과 보완 방향**
2. 결론

* 선택된 변수에 관하여

1. 날씨는 예상과 달리 상대적으로 유의미하지 못한 정보였다.
2. 산업종사자 수와 시간대가 콜 택시 수요에 가장 많은 영향력을 가지고 있었다.

- 모델 활용 방안

우리가 위에서 분석한 콜 택시 수요예측모델을 이용하여 미래의 변수를 입력하면 해당 시간대의 지역별 콜 택시 수요를 예측할 수 있다. 이를 이용하여 택시 회사에서는 예상되는 택시 수요에 따라 배차를 결정한다면 택시 사업의 비용을 줄일 수 있을 것이다.

1. 한계점과 보완 방향

* 일반택시 승하차 수 등의 교통정보를 반영하면 콜 택시 수요 예측에 더 많은 도움이 될 것으로 기대한다.
* 예측 콜 수와 실제 콜 수의 차이가 큰 경우는 강남구 야간, 공휴일일 때가 많았다.

설날과 같은 명절에는 서울의 인구가 다른 지역으로 이동하기 때문에 평균적으로 야간 콜 수가 매우 높은 강남구는 다른 구에 비해 예측력이 더욱 떨어진 것이라 추측한다. 휴일 정보를 좀더 세분화하여 반영한다면 이러한 점을 해결할 수 있을 것으로 기대한다.

* 도봉구의 경우도 마찬가지로 예측 콜 수와 실제 콜 수의 차이를 보였다. 조사 결과 창동의 경우 장애인 및 노인 시설이 집중되어 장애인 콜 택시 건수가 많이 포함되어 있었다. 따라서 장애인 콜 택시 건수를 반영할 수 있는 변수를 추가한다면 도봉구까지 정확한 예측을 할 수 있을 것이라 기대한다.

\* 더 자세한 과정은 PPT 확인 부탁드립니다.

1. https://www.bigdatahub.co.kr/product/view.do?pid=1001683 [↑](#footnote-ref-1)
2. https://data.kma.go.kr/data/grnd/selectAsosRltmList.do?pgmNo=36 [↑](#footnote-ref-2)
3. https://data.seoul.go.kr/openinf/linkview.jsp?infId=OA-876 [↑](#footnote-ref-3)
4. https://data.seoul.go.kr/openinf [↑](#footnote-ref-4)
5. http://data.seoul.go.kr/openinf/sheetview.jsp?infId=OA-11677 [↑](#footnote-ref-5)