PART. III
연구과제

서울대학교 도서관 데이터 분석

연구배경
연구내용
연구결과
향후계획

서울대학교 도서관 데이터 분석

KAIST 인문사회융합과학대학
김란우

연구배경

본 연구는 학생들이 도서관을 이용하면서 느끼는 두 가지 불편한 점(pain point)을 개선하고자 하는데 있다. 관악 캠퍼스는 너무 넓고, 접근성이 좋지 않다. 그리고 장서들이 다양한 단과대학교 도서관에 흩어져있다. 도서관 이용자들이 다양한 학내도서관을 사용할 때는 크게 두 가지 패턴이 있을 것이다.

1) 단과대학 도서관에서만 대여할 수 있는 책이 있다.
2) 관악 캠퍼스가 넓기 때문에, 멀리 위치한 도서관에 방문하기는 어렵다. 책을 빌리겠다는 의지는 상대적으로 commitment가 낮기 때문에, 지리적 이유로 포기하기 쉬울 수 있다.

두 가지 패턴을 종합하면, 책을 대여하여 읽음으로써 생겨날 수 있는 특정 분야에 대한 관심이 도서관의 지리적인 위치가 가지는 한계로 인하여 없어질 수 있다. 이는 서울대학교의 단과대학 체제를 공고히 하고 융합학문의 가능성을 줄일 수 있다는 한계점을 내포한다.

도서관 서비스는 기본적으로 내가 빌리는 책을 매개로 그 책을 대여한 다른 누군가와 관계가 생기는 2-mode network의 성격을 띤다. 따라서 책을 대출하려할 때 해당 책이 이미 대출중인 경우, 상대방에 대한 정보를 사용하지 않으면 해당 도서가 언제 반납될지 예측할 수 없다. 대기 신청을 하였을 때 알람 문자가 오지만, 이는 도서관을 이용할지 여부를 결정하는데는 크게 효율적이지 않다.

도서 반납 기간에 영향을 주는 요인은 다음과 같다.

○ 개인의 특성이 도서 반납 일자를 정하는데 영향을 미친다 : 책을 오래 읽는 사람과 그렇지 않은 사람 간에 차이가 있음

○ 학교에서의 소속, 지위 등에 따라 도서 대여 기간이 다르다. 학부생의 도서 대여 기간이 교수에 비해 짧다

○ 책의 특성이 도서를 반납하는데 영향을 미친다. 책의 언어, 페이지 수, 분야 등의 특성에 따라 책이 얼마나 빨리 읽히는지 예측할 수 있다

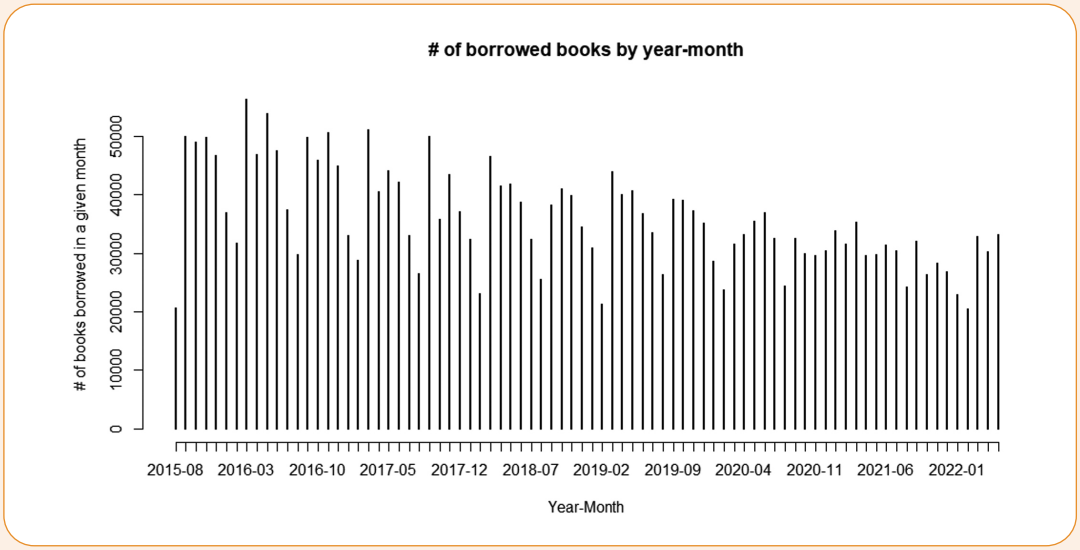
○ 시기의 특성이 도서를 반납하는데 영향을 미친다. 학기가 끝나는 시기에 대규모로 도서들이 반납될 것을 예측할 수 있다

이와 같은 변수를 고려하여, prediction machine learning model을 만들고, 원하는 책이 대출중일 때, 예상 반납시간을 제시할 수 있다.

연구내용

도서관 접근성은 필요한가?

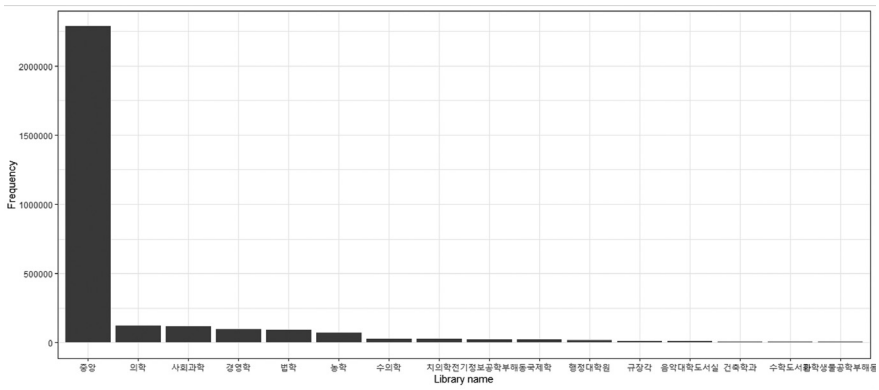
○ 본격적인 분석에 앞선 데이터 트렌드 및 문제점



위의 그림은 대출 데이터만을 통해 월별 대출 트렌드를 살펴본 것이다. 2015년부터 데이터가 시작하며, 전반적으

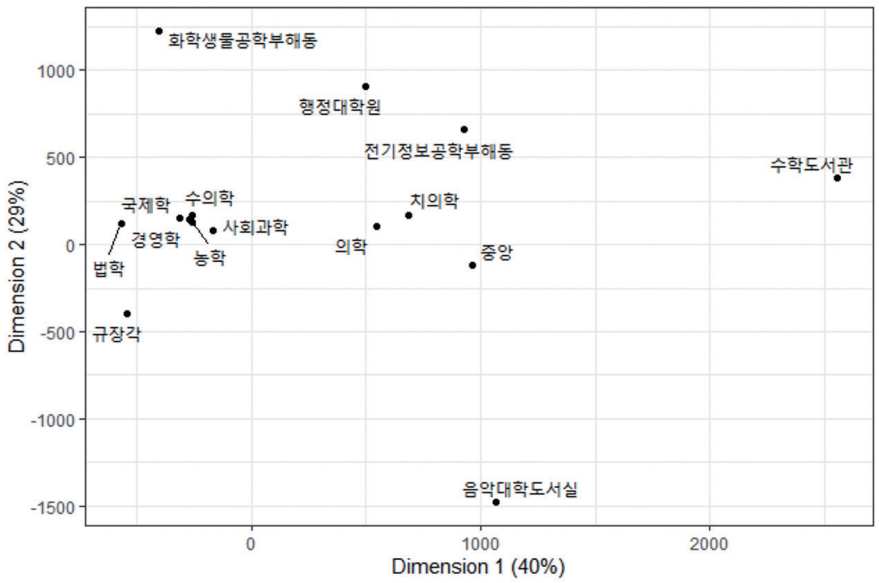
로 대출 횟수는 꾸준히 감소하는 것으로 나타난다. 그리고 뚜렷한 Seasonality가 있는 점을 찾아볼 수 있다.(학기 시작할 때 대량 대출). Seasonality도 Covid 기간 동안에는 희미해진 점을 찾아볼 수 있다. 하지만 같은 데이터를 대출자 정보(borrower)와 연동하여 데이터베이스에서 query를 진행했을 경우, 위의 그림과 다른 트렌드를 보여주고 있는 점을 찾아볼 수 있다. 즉, 대출 데이터의 초반 부분 (2015년~ 2018년)의 대출자 정보에 어딘가 구조적인 missing이 있는 점을 찾아볼 수 있다.

○ 중앙도서관에 대출 집중



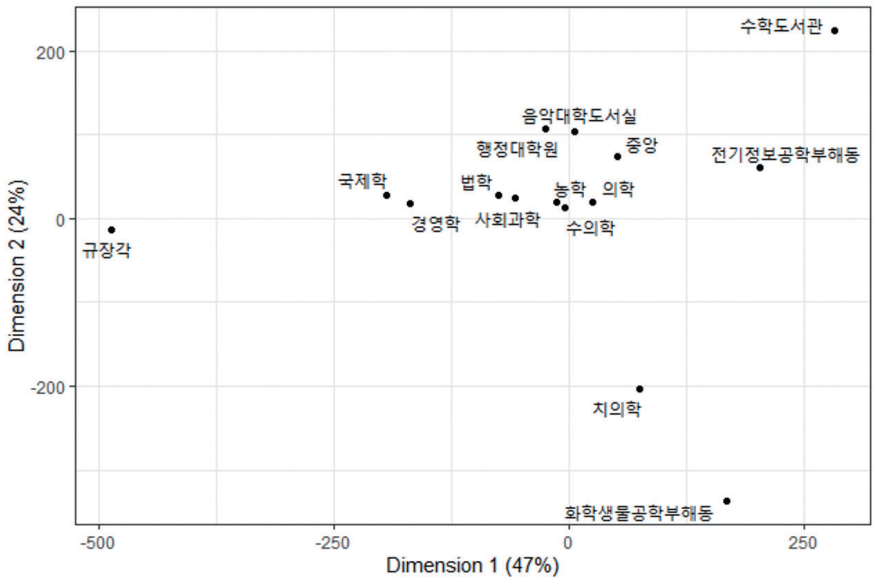
중앙도서관에 대부분의 도서 대출이 집중되어 있는 점을 찾아볼 수 있다.

○ 도서관-책의 2-mode network를 사용한 distance 분석



가장 이상적인 분석은 각 도서관이 소장하고 있는 장서 정보를 정확히 수집하여, 얼마나 장서가 겹치는지를 분석하는 것이겠으나, 해당 데이터가 현재는 없기 때문에 이용자의 도서 대출 데이터를 토대로 분석하였다. 두 도서관 사이가 가깝다는 것은 두 도서관에서 대출된 책들이 겹치는 경우가 많다는 것을 뜻한다. MDS 분석에서 2 dimension이 도서관들 사이의 distance matrix의 69%를 설명하는 것으로 나타났다. 특히, 중앙 도서관은 이 2 dimension에서도 중앙에 위치하는 것으로 나타났다. 국제학, 경영학, 수의학, 농학, 사회과학 도서관은 가깝게 cluster 되어 있는 것으로 나타났다. 음악대학 도서실, 수학도서관, 행정대학원 도서관의 대출 기록은 타 도서관들과 차이가 큰 것으로 보인다. 하지만 해당 분석은 2권 이상 겹치는 책들만을 대상으로 했다는 한계점이 있다. 즉, 예를 들어 규장각에서 100권의 책은 규장각에만 있는 책들이었다면, 해당 분석에서 이 데이터는 반영되지 않았다.

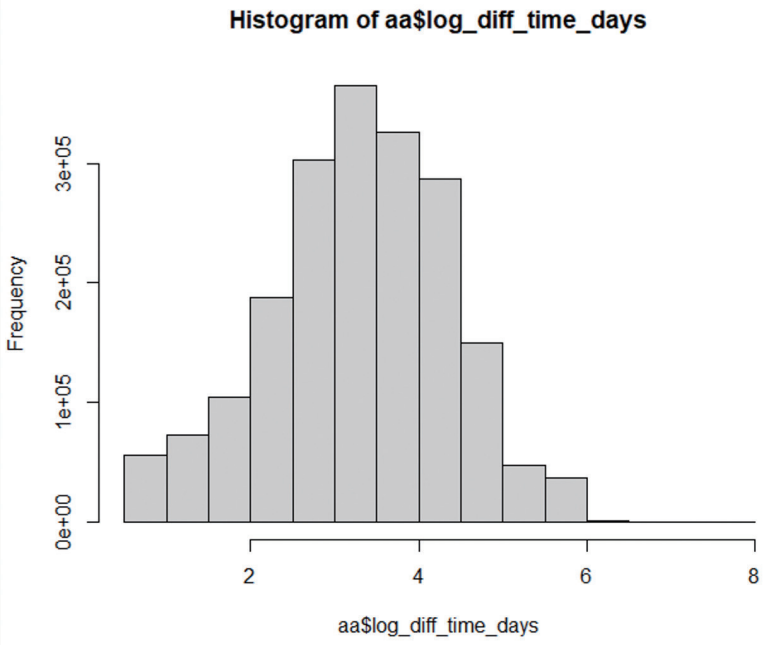
○ 대출자-책의 2-mode network를 사용한 distance 분석



이번 분석에서는 책이 아닌 대출자를 중심으로 도서관 사이의 distance를 계산하였다. MDS 분석에서 2 dimension이 도서관들 사이의 distance matrix의 71%를 설명하는 것으로 나타났다. 첫 번째 dimension에서 가장 거리가 먼 곳은 규장각과 전기정보공학부해동/수학도서관인 것으로 나타났다. 두 번째 dimension에서 가장 거리가 먼 곳은 수학도서관과 화학생물공학부해동이었다. 그 외의 대부분 도서관들은 중앙에 모여있는 것을 확인할 수 있다. 도서관-책의 클러스터링 결과와 비교하여 보았을 때, 대출자를 기반으로 한 분석이 책을 기반으로 한 분석보다 더 중앙에 밀집해 있는 모습을 확인할 수 있다. 즉, 장서는 도서관에 따라 구비하고 있는 책에 특성이 있고 차이가 있지만, 사람들은 필요하다면 어느 도서관이든 방문한다는 점을 보여준다.

도서 대출 예측 모델

- 종속변수: 도서 대출 기간. Right-skewed 분포이기 때문에 log값을 취함. log를 취한 이후에는 정규 분포와 매우 흡사함.



- 독립변수: 출생연도, 대출자 그룹(강사, 교수, 학부생 등), 소속 단과대학, 학과, 성별, 학위과정, 입학연도, 대출연도, 현재 재적 상황(수료, 수학가능, 재적, 졸업 등), 대출 도서관. NA가 많기 때문에 이들을 삭제하기 보다는 하나의 그룹으로 생성하여, 모든 변수는 범주형 변수로 변형되었다.
- 모델: 종속변수가 정규분포를 따르기 때문에 가장 간단한 OLS 회귀를 사용하였다.
- 결과: Adjusted R-squared가 23%로 약 23%의 variance가 위의 변수로 설명되는 것으로 나타났다. 몇몇 결과는 아래와 같다.

```
> summary(m1)

Call:
lm(formula = log_diff_time_days ~ birth_year_category + user_group +
    college + gender + degree.y + admission_year_category + loan_year +
    department + grade_detail + library_code, data = aa)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7617 -0.5436  0.0970  0.6971  5.0422

Coefficients: (119 not defined because of singularities)
                                                                          Estimate Std. Error  t value Pr(>|t|)
(Intercept)                                                             -1.126e+02  8.933e-01 -126.042  < 2e-16 ***
birth_year_categoryUnder1980                                            -1.446e-01  4.568e-03  -31.659  < 2e-16 ***
birth_year_categoryUnder1990                                            -1.985e-01  5.282e-03  -37.582  < 2e-16 ***
birth_year_categoryUnder2000                                            -3.315e-01  5.643e-03  -58.744  < 2e-16 ***
birth_year_categoryUnder2010                                            -5.158e-01  6.914e-03  -74.611  < 2e-16 ***
user_group강사                                                           3.803e-01  1.116e-01    3.408 0.000654 ***
user_group교수                                                           9.002e-01  1.116e-01    8.063 7.43e-16 ***
user_group교환학생                                                       7.935e-02  3.582e-01    0.222 0.824676
user_group기금교수                                                       1.351e+00  1.160e-01   11.641  < 2e-16 ***
user_group도서관출입증(졸업생)                                           1.037e-01  1.116e-01    0.929 0.352838
user_group동문회원(대출)                                                 1.279e-01  1.116e-01    1.147 0.251524
user_group동문회원(회원제)                                               2.681e-01  1.311e-01    2.044 0.040909 *
user_group명예교수                                                       6.190e-01  1.130e-01    5.480 4.26e-08 ***
user_group명예회원(교수) (구 일용직)                                     1.086e+00  1.169e-01    9.296  < 2e-16 ***
user_group박사과정                                                       4.026e-01  1.115e-01    3.610 0.000306 ***
user_group법인직원                                                       3.689e-01  1.129e-01    3.266 0.001090 **
user_group병원직원                                                       6.610e-01  1.515e-01    4.362 1.29e-05 ***
user_group부교수                                                         1.142e+00  1.117e-01   10.223  < 2e-16 ***
user_group부속학교 교사                                                  1.473e+00  5.564e-01    2.648 0.008091 **
user_group비전임교수                                                     7.797e-01  1.120e-01    6.962 3.35e-12 ***
user_group석박사통합과정                                                 6.466e-01  1.474e-01    4.388 1.15e-05 ***
user_group석사과정                                                       2.814e-01  1.115e-01    2.524 0.011617 *
user_group연구생(대학원생)                                               3.055e-01  1.115e-01    2.739 0.006171 **
user_group연구원(특,교,보)                                               2.823e-01  1.118e-01    2.526 0.011525 *
user_group인턴                                                           2.251e-01  1.111e-01    2.026 0.042764 *
user_group임상강사                                                       4.116e-03  1.490e-01    0.028 0.977963
user_group임상교수                                                       2.580e-01  2.569e-01    1.004 0.315339
user_group자체직원                                                       4.357e-01  1.117e-01    3.902 9.53e-05 ***
user_group전공의                                                         1.625e-01  1.140e-01    1.425 0.154106
user_group조교                                                           3.197e-01  1.118e-01    2.859 0.004251 **
user_group조교수                                                         8.165e-01  1.119e-01    7.298 2.93e-13 ***
user_group조빙교원                                                       1.189e+00  1.254e-01    9.483  < 2e-16 ***
user_group특별회원(회원제)                                              -2.455e-01  1.448e-01   -1.695 0.090079 .
user_group학부                                                           5.177e-02  1.115e-01    0.464 0.642476
user_group휴학생(대학원)                                                 3.019e-01  1.117e-01    2.704 0.006861 **
user_group휴학생(학부)                                                   1.425e-02  1.116e-01    0.128 0.898404
collegeRNA 연구단                                                       -2.329e-01  1.263e-01   -1.844 0.065185 .
college간호과학연구소                                                   -3.186e-02  8.412e-02   -0.379 0.704890
college간호대학                                                         -2.357e-01  9.616e-02   -2.452 0.014225 *
college강상관계물질연구단                                                4.261e-01  1.643e-01    2.593 0.009506 **
college건강사회정책실                                                    -6.650e-01  5.660e-02  -11.751  < 2e-16 ***
college건설환경종합연구소                                                2.204e-02  3.896e-02    0.566 0.571554
college경력개발센터                                                      -7.935e-01  1.254e-01   -6.328 2.48e-10 ***
college경영대학                                                          -5.652e-02  1.440e-02   -3.925 8.66e-05 ***
```

```
college자동화시스템공동연구소                                             2.900e-01  9.633e-02    3.011 0.002606 **
college자연과학대학                                                      -1.089e-01  1.874e-02   -5.812 6.18e-09 ***
college자유전공학부                                                       2.602e-02  1.446e-02    1.799 0.071977 .
college재단법인기초전력공학공동연구소                                    -3.241e-01  1.668e-01   -1.943 0.052015 .
college재단법인서울대학교발전기금                                        -1.426e-01  2.107e-01   -0.677 0.498566
college재정전략실                                                        -2.506e-01  8.146e-02   -3.076 0.002096 **
college전력연구소                                                        -5.314e-02  9.688e-02   -0.549 0.583340
college정밀기계설계공동연구소                                             6.806e-02  3.999e-02    1.702 0.088783 .
college정보화본부                                                         4.616e-01  2.981e-01    1.549 0.121427
college조형연구소                                                         5.837e-01  1.643e-01    3.554 0.000380 ***
college종합약학연구소                                                    -5.201e-01  3.917e-02  -13.277  < 2e-16 ***
college중앙도서관                                                        -3.830e-01  8.399e-02   -4.560 5.12e-06 ***
college중권·금융연구소                                                    6.640e-02  3.556e-01    0.187 0.851865
college자세대우주주진연구센터                                            -2.235e-01  9.402e-01   -0.238 0.812132
college자세대융합기술연구원                                              -4.634e-01  2.220e-01   -2.087 0.036851 *
college천연물과학연구소                                                   8.371e-03  2.284e-01    0.037 0.970765
college철학사상연구소                                                    -1.715e+00  2.082e-02  -82.361  < 2e-16 ***
college치과대학                                                           9.617e-02  5.852e-02    1.643 0.100315
college치의학대학원                                                       6.921e-02  5.186e-02    1.334 0.182066
college치학연구소                                                         6.588e-03  4.619e-02    0.143 0.886577
college컴퓨터연구소                                                      -3.360e-01  8.647e-02   -3.886 0.000102 ***
college통일평화연구원                                                     3.681e-02  4.241e-02    0.868 0.385366
college평생교육원                                                         1.360e-01  7.229e-02    1.882 0.059829 .
college학생생활관                                                         2.659e-01  9.724e-02    2.734 0.006253 **
college학생처                                                            -2.352e-02  3.441e-02   -0.684 0.494282
college한국우주신기술연구소                                              -4.718e-01  1.012e-01   -4.663 3.12e-06 ***
college한국행정연구소                                                     4.538e-01  5.202e-02    8.725  < 2e-16 ***
college해양시스템공학연구소                                               8.981e-02  1.964e-01    0.457 0.647557
college해양연구소                                                        -3.383e-01  1.011e-01   -3.346 0.000821 ***
college행정대학원                                                        -3.373e-02  1.482e-02   -2.275 0.022878 *
college화학공정신기술연구소                                               2.051e-01  9.108e-02    2.252 0.024348 *
college환경계획연구소                                                     3.740e-01  2.511e-02   14.895  < 2e-16 ***
college환경대학원                                                         1.137e-01  7.045e-02    1.614 0.106550
college환경안전원                                                         1.359e-01  2.616e-01    0.519 0.603509
gender여자                                                                7.582e-02  1.627e-03   46.599  < 2e-16 ***
degree.y박사                                                              2.249e-01  6.241e-03   36.033  < 2e-16 ***
degree.y불명                                                              1.002e-02  2.210e-02    0.453 0.650309
degree.y석사                                                              1.942e-01  3.056e-03   63.563  < 2e-16 ***
degree.y학사                                                             -1.055e-01  2.106e-03  -50.097  < 2e-16 ***
admission_year_categoryUnder 2020                                        -1.340e-01  3.285e-03  -40.772  < 2e-16 ***
admission_year_categoryUnder1980                                         -5.266e-02  4.329e-02   -1.217 0.223749
admission_year_categoryUnder1990                                         -2.413e-01  7.407e-03  -32.576  < 2e-16 ***
 [ reached getOption("max.print") -- omitted 439 rows ]
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9401 on 1724803 degrees of freedom
  (결측으로 인하여 210559개의 관측치가 삭제되었습니다 )
Multiple R-squared:  0.2268,	Adjusted R-squared:  0.2266
F-statistic:   975 on 519 and 1724803 DF,  p-value: < 2.2e-16
```

연구결과

제안 분석 모델 1 - 지리적인 이유로 대출되지 않는 건수 측정

중앙도서관에 있으며 사람 A에 의해 대여될 확률이 높은 책들 중 실제 대여되는 확률

분관에 있으며 사람 A에 의해 대여될 확률이 높은 책들 중 실제 대여되는 확률

첫 번째 데이터 분석 결과는 책은 비교적 분관별로 특성화되어 있으나, 대출자는 이에 크게 구애받지 않고 이동하는 것을 확인할 수 있다. 하지만 위의 데이터 분석 결과는 두 군데 이상의 도서관을 방문한 대출자에 한하여 분석한 결과이며, 대부분의 대출자는 중앙도서관 한 군데만 이용하는 것으로 추정된다. 서울대학교 관악캠퍼스의 까다로운 지리적 특성을 고려할 때, 이는 개인의 선호라기보다 강제된 측면이 있다. 분관의 책이 중앙도서관에 진열되어 있었다면, 얼마나 더 다양한 책이 대출될 수 있었을지에 대한 분석 모델이 필요하다. 다른 연구자들 및 기업으로부터 도출된 도서대여 추천 시스템을 사용하여, 분관의 책과 중앙도서관의 책 사이에 propensity score matching을 진행할 필요가 있다. 그래서 propensity score matching이 된 책들이 중앙도서관에 있지 않다는 이유로 얼마나 덜 대출되는지를 분석한다면, 지리적인 이유로 줄어드는 대출의 정도를 파악할 수 있을 것이다.

제안 분석 모델 2 - 더 구체화된 대출 반납 시기 prediction algorithm 분석

두 번째 분석 결과에서는 오직 학교에서 제공하는 해당 대여자에 관한 정보만을 사용하여 분석하였으며, 개인의 특성 / 책의 특성 / 시기적 특성 등을 고려하지 않았음에도 반납할 날짜를 약 22% 예측할 수 있었다. 해당 정보를 더 추가하고 기계학습 모델을 더 견고하게 구성한다면, 책의 반납 날짜를 더 정확하게 예측할 수 있을 것이다.

단, 데이터상의 문제를 해결해야할 필요가 있다. 책 대출 데이터는 각 row가 1건의 대출 기록으로 구성되어 있다. 하지만 책을 대출한 대출자 정보의 경우 각 row가 1명의 대출자로 구성되어 있다. 문제는 대출자의 정보는 시간이 지나며 계속 업데이트 되지만, 현재 도서관 데이터의 정보는 2022년 현재 최종 업데이트 된 정보만 포함하고 있기 때문에, 정확히 책을 대출한 시점에 대출자가 어떤 상황이었는지 알기는 어렵다. 예를 들어, 2003년 입학하여 2007년 졸업한 학부생의 정보는 대출자 정보에서는 '졸업' 상태로 최종 기록되어 있다. 하지만 이 학부생이 2003~2007년 사이에 책을 대출하였다면, 이 학부생은 졸업생 신분으로 대출한 것이 아니라 재학생 신분으로 대출한 것이다. 그렇기 때문에 두 데이터베이스를 단순히 결합하여서는 안되며, 입학년도와 졸업년도를 사용하여 해당 시기의 상태를 파악할

필요가 있다. 그렇다고 하더라도 구체적으로 휴학생 신분이었는지 등의 여부는 알기 어렵다. 수업 데이터를 사용하여 수업을 수강한 학기였는지를 파악하면 휴학 중이었는지의 여부도 더 구체적으로 추정할 수 있을 것이다.

향후계획

도서관 circulation desk 운영

만약 분석모델 1의 결과에서 지리적인 이유로 대출되지 않는 건수가 상당량 이상으로 밝혀진다면, 도서관 차원에서 circulation desk 운영을 고려해볼 필요가 있다. 현재 반납은 어디에서나 가능하지만 대출은 어디에서나 가능하지는 않다. 미리 도서관 시스템을 통해 학교 어느 도서관에서나 대출하고자 하는 책을 신청하고, 정해진 한 군데의 장소에서 신청한 책을 받을 수 있는 circulation desk를 운영한다면 지리적인 이유로 대출을 포기했던 책을 대출받을 수 있도록 도와줄 수 있을 것이다.

대출된 도서의 반납 일자 예측 서비스 제공

분석모델 2의 예측 알고리즘 정확도가 상당 수준 이상이 된다면 도서관 서비스 개선에 활용할 수 있다. 도서관 홈페이지에서 검색한 책이 대출중인 경우, 사용자가 원할 때에 한하여 반납 날짜를 예측하는 서비스를 제공하는 방식(예: “이 책은 80%의 확률로 11월 17일에서 21일 사이에 반납될 예정입니다.”)이 유용할 것으로 생각된다.

본 데이터 중심 접근의 한계점

이용자 중심의 도서관 서비스를 만드는 것도 중요한 일이나, 이용자에게 필요한 모든 서비스를 제공하는 것이 도서관의 역할이 되어야 할까? 예를 들어, 해리포터 시리즈 대출을 10명이 예약해놓았다면, 도서관은 해리포터 시리즈를 10권 더 사는 것이 이용자중심의 도서관일까? 어떤 이용자 중심의 필요를 우선시해야할지 순위를 정하고, 도서관이 추구하고자 하는 원칙과 어긋나지 않는지 고민할 필요가 있다.