강원도 공공 도서관 도서대출 수요 예측 - 랜덤포레스트 변수 중요도 기반의 머신러닝 기법 중심으로 -

윤미리*, 황지애*, 김정우*
*강릉원주대학교 경제학과
e-mail:alfl13799@naver.com, hjia14159@naver.com, kurtkim@gwnu.ac.kr

Predicting the Number of Book Loans in Public Libraries in Gangwon-do - Focusing on Machine Learning Techniques Based on Random Forest Feature Importance -

Mi-Ri Yoon*, Ji-Ae Hwang*, Jeong-Woo Kim*
*Dept. of Economic, Gangneung-Wonju National University

요 약

본 연구는 변해가는 사회적 흐름에서 우리나라의 강원도 공공도서관의 도서 대출 권수를 다양한 머신러닝 기법으로 정확히 예측하고자 하였다. 기존 연구들은 대체적으로 도서관 내부 데이터들을 중심으로 이루어진 데 반해, 본 연구는 다양한 머신러닝 기법으로 지역적 측면도 고려하여 강원도 공공 도서관의 도서 대출 권수를 예측하였으며, 특히 랜덤포레스트 변수 중요도에 기반하여 오차율을 줄여 예측의 안정성을 도모하였다. 강원도 내 도서관 대출권수를 종속변수로 하고 도서관의 이용자수 , 연령층, 총인구수 등을 설명변수로 설정하여 서로 다른 특성을 지닌 5가지의 머신러닝 기법을 적용하여 예측결과들을 비교하였다. 분석 결과, 예측구간별로 예측성능이 높은 머신러닝 기법은 상이하였으나, 예측의 안정성 측면에서는 LASSO가 상대적으로 우수한 것으로 나타났다. 이에 따라, 본 연구는 도서관 대출에 미치는 영향요소가 무엇인지 다양한 시각에서 바라본 것에 시사점이 있다.

1. 서론

독서는 다양한 지식과 정보를 사람들에게 전달하며 사고력을 기르는 데 도움이 되는 매우 유용한 활동이다. 독서와 관련하여 공공도서관의 서비스가 사람들의 삶의 질에 긍정적인 영향을 준다는 연구 결과도 있다[1]. 하지만 최근 우리나라 사람들의 독서율은 점점 떨어지고 있다. 만 19세 이상 성인 중 지난 1년간(2020.9～2021.8) 교과서·학습참고서·수험서를 제외한 일반도서를 한 권 이상 읽은 연간종합 독서율은 성인 47.5%, 학생 91.4%이다. 2년 전(2019년) 대비 성인은 8.2%p 감소하고, 학생은0.7%p 감소하였다. 이를 자세히 살펴보면 2021년 조사 결과 2019년 대비 성인 전체 연령대의 종이책 독서율 감소가 10%p 전후로 큰 반면, 20～30대에서 전자책과 오디오북 독서율 증가가 뚜렷하게 나타났다[2]. 종이책에서 전자책으로의 관심은 전자도서관의 개관에 영향을 미치는 등 큰 변화를 가져왔다.

아울러, 현대사회에서 문제가 되고 있는 고령화·저출산 현상은 인구 외적인 규모의 변화, 사회적으로 큰 변화를 불러온다. 도서 대출 측면에서 노년 대출자는 다른 연령층과 비교되는 특이점을 보였다는 연구 결과가 있다[1,4]. 또한 감소하는 아이들의 수는 한 명당 대출하는 도서의 수에 한계가 있기 때문에 도서관 도서 대출권수에 영향을 미칠 것이다. 도서 대출수요를 예측한 연구는 주로 각 도서관의 자료 수와 총 예산, 대출자 연령과 같은 기관 내부 데이터를 이용하여 실행되었다[1,3,4]. 하지만 그 외에도 교육수준과 경제적 여건[4], 학교의 수[3] 등 지역의 특성을 반영하기도 했다.

본 연구는 선행연구와 비교하여 지역의 인구·복지적 특성에 중점을 두어 분석하고자 한다. 연구 대상은 강원도 지역으로 한정하며, 강원도가 다른 지역과 비교하여 평균연령이 높다는 특성을 이용하여 예측하고자 한다. 또한 여러 가지 머신러닝 기법을 활용하여 분석 기법간의 비교를 함으로써 여러 설명변수와 분석 대상간의 인과 관계에 대하여 풍부한 결과를 얻고자 한다. 본 연구 결과를 통해 지역 환경과 사람들의 도서 대출의 연관성을 파악하고 향후 대출수요를 보다 정확하게 예측 가능할 것을 기대한다.

2. 연구방법

본 연구는 머신러닝 기법을 활용하여 도서대출수요를 예측하는 것을 최종 목적으로 한다. 이를 위하여 국가도서관통계시스템의 2018년 도서관별 도서 대출권수 통계를 도서대출 수요지표로 활용한다. 인구수와 같은 데이터는 KOSIS를 이용하였다. 본 연구에서 사용된 데이터는 설명변수의 개수가 많은 편이므로, 랜덤 포레스트 변수 중요도(Random Forest Feature Importance, RFFI)를 사용하여 대출권수에 영향을 미치는 중요도에 따라 설명변수들을 선별하였다. 또한 여러 개의 머신러닝 기법을 사용하여 비교 및 분석하는 결과 값을 얻어 예측의 정확성을 보다 높이고자 하였다.

본 연구방법으로는 k-최근접 이웃법(k-Nearest Neighbor, kNN)과 K-평균 군집화(K-means clustering, kmeans), LASSO(Least absolute shrinkage and selection operator), 의사결정나무(Decision tree, tree), 서포트백터회귀(Support Vector Regression, SVR)의 총 5가지 머신러닝 기법을 사용하였다. 연구 데이터가 지역뿐만 아니라 도서관마다 대출권수와 그 설명변수 값이 매우 상이하게 나타나 회귀분석을 통한 예측보다는 분류를 통한 예측이 더 많다.

[표 1] 기초통계량

(단위 : 권, 명, ₩)

변수	MEAN	STD	MIN	MEDIAN	MAX
성인L	35,277	50,448.7	0	19,755	321,392
청소년L	3,991.1	6,090.0	0	1,539	33,339
어린이L	13,310	21,192.3	0	6,119	123,810
평균연령	45.6	3.0	41.3	44.5	50.9
총인구수	120,612	110,337	23,408	68,326	344,070
노년S	29.8	7.2	18.2	29.7	44.7
N.C.P	24.1	31.8	0	11	129
C.P.P	4,565	6,895.7	0	2,264	38,950
N.R.P	11.1	10.5	0	9	48
R.P.P	4,224.6	5,441.3	0	2,402	26,593
어린이M	974.3	1,715.3	0	278	8,516
청소년M	1,087.3	2,176.6	0	408	10,450
성인M	7,536	18,197	0	2,324	104,982
교육건수	35.2	59.6	0	12	356
교육시간	72.1	186.3	0	10	1,320
N.T.P	658	1,270	0	200	8,405

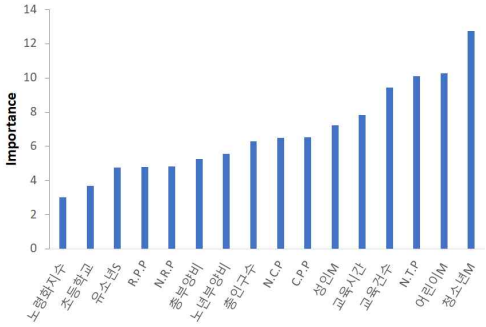
* L(대출권수), S(부양비), M(회원), N.C.P(Number of Cultural Programs), C.P.P(Cultural Program Participants), N.R.P(Number of Reading Programs), R.P.P(Reading Program Participants), N.T.P(Number of Training Participants)

3. 분석결과

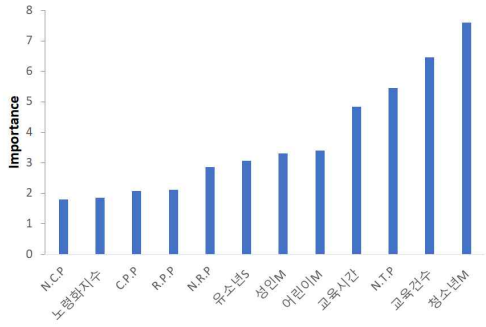
[그림 1]은 랜덤포레스트 변수 중요도에 따라 선별된 각 설명변수들과 중요도 점수를 나타내고 있다. 우선 어린이 대출권수를 보면, 어린이 회원수와 청소년 회원수가 어린이 대출권수를 예측하는데 큰 영향을 준다는 것을 알 수 있다. 청소년 회원수의 중요도가 다소 높게 나오는 것은 형제/자매 등의 요소가 작용한 것으로 보인다. 그 다음으로 교육관련 변수들의 영향력이 두드러지며, 상대적으로 고령층 관련 변수들은 어린이대출권수 예측에 큰 영향을 주지 못하는 것으로 나타났다.

다음으로 청소년 대출권수를 보면, 예측하는데 가장 높게 영향을 끼친 변수는 청소년회원수라는 사실을 알 수 있다. 그 다음으로는 교육 건수, 교육 참가자수, 교육시간 등의 교육관련 변수들의 중요도가 높게 나타났다. 이는 청소년이 이용자 교육에 대하여 상대적으로 더 큰 영향을 받은 것으로 보인다. 이에 반해 문화프로그램과 고령층 변수는 청소년대출권수 예측에 크게 작용하지 못한 것으로 나타났다.

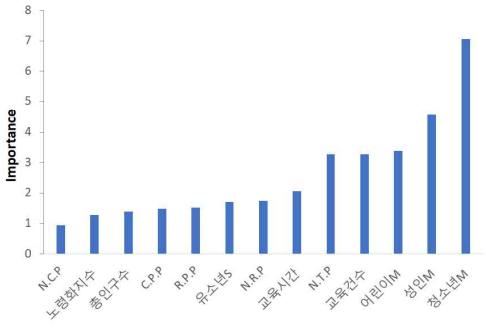
마지막으로 성인 대출권수는 청소년회원, 성인회원, 어린이 회원의 수가 예측을 하는데 크게 작용한 것으로 나타났다. 성인 연령대는 부모님의 나이대도 포함하고 있으므로 어린이회원과 청소년회원의 변수 중요도가 크게 작용한 것으로 보인다. 상대적으로 중요도가 낮은 고령층 관련 변수는 고령의 도서 대출자수가 적어서 성인대출권수에는 중요한 의미를 차지하지 못하는 것으로 보인다[1].



(a) 어린이 대출권수



(b) 청소년 대출권수



(c) 성인 대출권수

[그림 1] Random Forest Variable importance

[표 2]는 각 예측기법들의 도서 대출권수에 대한 예측성능을 나타낸 것이다. 예측성능은 평균절대비오차(Mean Absolute Percentage Error, MAPE)로 측정하였다. 기존의 모든 변수를 이용한 예측과 RFFI로 선별된 변수를 통한 예측을 구분하여 시행하였다. MAPE 수식은 (1)과 같다.

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - f_i|}{y_i} \times 100 \qquad (1)$$

여기서, N은 예측 수, y_i는 실제값, f_i는 각 기법의 예측값이며, $\frac{|y_i - f_i|}{y_i} \times 100$ 은 절대비오차(Absolute Percentage Error, APE)이다.

모든 변수의 경우를 살펴보면, LASSO가 가장 낮은 MAPE값을 보여주어 가장 높은 예측성능을 나타내었다. 그 다음으로 tree와 SVR이 높은 예측성능을 나타내었다. RFFI의 경우, SVR이 가장 높은 예측성능을 보여주었고 그 다음으로 LASSO와 tree가 높은 예측성능을 나타내었다.

각 예측기법들의 예측성능을 비교한 결과, LASSO, tree, SVR이 전반적으로 높은 예측성능을 나타낸 것을 알 수 있다. 또한 모든 변수를 고려한 것 보다 RFFI를 고려한 경우에 대체적으로 더 낮은 MAPE 수치가 관찰되어 선별 변수의 더 높은 예측 성능을 확인할 수 있다.

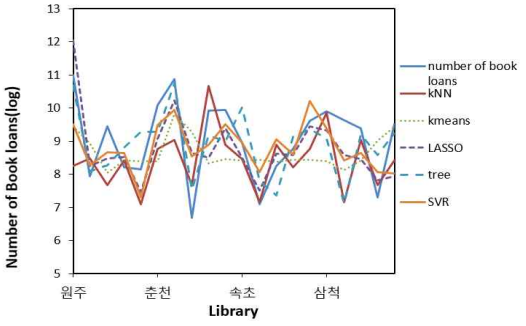
[표 2] MAPE 비교

(단위 : %)

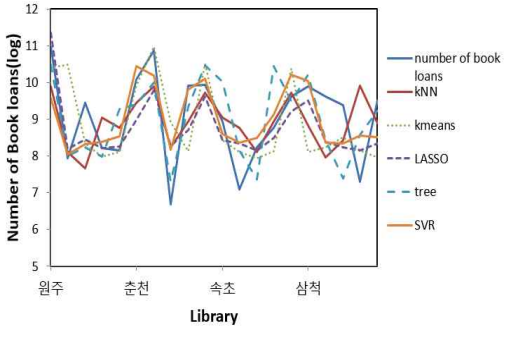
구분	kNN	kmeans	LASSO	tree	SVR
모든변수	10.19	12.53	8.51	8.61	9.13
RFFI	10.45	11.03	8.18	9.24	7.82

[그림 2]는 각 예측기법들의 2가지 예측값을 그래프상에서 비교하고 있다. 좌측부터 2개의 값은 원주, 5개는 춘천, 3개는 강릉, 2개는 속초, 3개는 동해, 각각 삼척과 횡성, 홍천, 태백, 양구지역의 20개 지역을 나타내었다. 모든 변수를 고려한 경우 실제 도서 대출 권수(파란색)에 LASSO(보라색 쇄선)와 tree(하늘색 쇄선)가 근접하게 나타났다. Table 2에서 가장 높은 MAPE를 보여준 kmeans가 실제 도서 대출 권수와는 거리가 먼 추세를 나타내었다. kNN(붉은색)의 경우에는 부분적으로는 실제 도서 대출 권수에 근접한 모습을 보이고 있다. 그러나 일부 지역의 도서 대출권수 예측은 다른 기법들보다 더 낮게 추정하고 있는 것으로 나타났다. 이는 kNN의 근접한 데이터들에 영향을 받아 새로운 데이터의 범주가 정해지는 기법 특성 상 나타나는 모습으로 보인다.

RFFI 항목은 랜덤 포레스트 기법을 이용하여 중요한 변수들로 실제 도서 대출 권수를 예측한 것이다. 그렇기 때문에 모든 변수를 고려하여 예측한 경우보다 대체적으로 MAPE 수치가 작은 것으로 나타났다. 다만, kNN의 경우 모든 변수를 고려한 경우와 마찬가지의 이유로 높은 예측성능을 보이지 못하고 있다. [표 2]의 MAPE상 가장 높은 예측성능을 보인 SVR(주황색)의 경우 특히 원주의 2번째 도서관과 거의 근접한 모습을 보인다. tree(하늘색 쇄선)도 일부 구간에서 실제 도서 대출 권수에 근접한 예측값을 보였으나, 다른 구간에서는 실제 도서 대출 권수와 거리가 먼 예측값을 나타내었다.

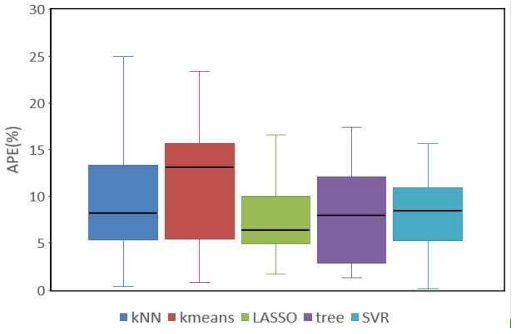


(a) 모든 변수

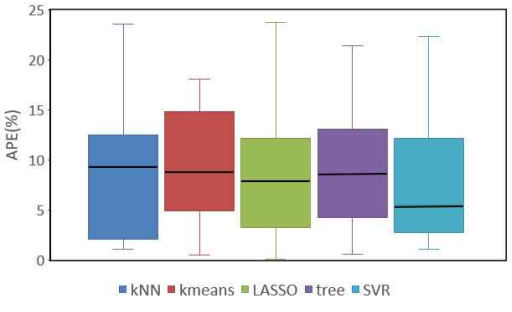


(b) RFFI

[그림 2] 예측값 비교



(a) 모든 변수



(b) RFFI

[그림 3] 상자그림 비교

[그림 3]은 실제값과 APE를 각 예측기법별로 상자그림(Boxplot)을 그린 것이다. [그림 2]와 마찬가지로 모든 변수와 RFFI를 통한 선별변수로 나누어 나타내었다. 모든 변수의 경우 LASSO의 중앙값이 가장 낮은 것으로 나타났다. 다른 기법과 비교하여 가장 낮은 사분위수 범위(InterQuartile range, IQR)를 보여 예측의 정확성뿐만 아니라 안정성도 높은 기법으로 확인된다. 이외에 kmeans를 제외한 나머지 세 가지 기법이 낮은 중앙값을 보이는데 IQR를 고려하면 SVR이 예측에 더 용이한 것으로 관찰된다.

RFFI의 경우 SVR이 가장 낮은 중앙값을 보이고 나머지 기법들은 모두 낮은 수준에서 비슷한 중앙값을 나타내었다. kmeans는 [표 2]에서 모두 비교적 높은 값을 보였는데 RFFI 상자그림에서 중앙값이 평균보다 낮은 모습을 보였다. 이는 관측치들이 평균이하에 많이 분포하고 있다는 뜻으로 변수 선별을 통해 kmeans의 예측 안정성이 높아졌으며 IQR로도 확인할 수 있다. 전체적으로 RFFI의 중앙값 평균은 모든 변수보다 낮지만 IQR은 크게 나타났다.

4. 결론

본 논문에서는 강원도 도서관 도서 대출수요를 예측하기 위하여 머신러닝 기법을 적용하기 전, 랜덤포레스트 중요도를 통한 변수 선별 후 머신러닝 기법을 적용하는 방식을 활용하였다. 이와 같은 예측 알고리즘은 모든 변수를 사용한 경우보다 전반적으로 낮은 수준의 예측 오차를 보여주어 예측정확성이 높은 것으로 나타났다. 서론에서 언급한 바와 같이 도서 대출수요는 사회의 다양한 요소에 의해서 영향을 받으며 이에 따라 시간에 따라서도 변동이 있을 것이므로, 향후에는 도서 대출수요의 시계열 데이터에 랜덤포레스트 중요도 기반의 머신러닝 기법 예측기법을 적용하는 연구도 의미가 있을 것이다.

참고문헌

[1] 권나현, “공공도서관 서비스평가-일상생활에서의 공공도서관 서비스 혜택에 대한 전국 성인들의 인식을 중심으로-”, 한국문헌정보학회지, 제49권 2호, pp.169-194, 5월, 2015년

[2] 백원근, “2021년 국민독서실태조사”, 문화체육관광부, pp.1-490, 12월, 2021년

[3] 오민기, 김경래, 정원웅, 김건욱, “문화적 특성을 고려한 공공도서관 도서 대출수요 분석: 대구광역시 시립도서관을 사례로”, 디지털융복합연구, 제19권 3호, pp. 55-64, 3월, 2021년

[4] 허선, 정연경, “대출기록을 통해 본 공공도서관 이용자 연구-강서·양천지역을 중심으로-”, 한국비블리아학회지, 제25권 4호, pp.187-207, 12월, 2014년