전자책 독서에 영향을 주는 요인분석

과목 : 다변량및빅데이터분석(캡스톤디자인)
팀 : 1조
팀장:20200697권소연
팀원:20201227왕유진
20190813 조슬기
20190798 서지혜

요약

본 분석에서는 국민 독서실태 설문조사 결과를 이용하여 전자책 독서에 영향을 미치는 요인들을 분석하였다. 이를 위해 데이터의 기초통계량을 살펴보고 알맞은 형태로 변환하였으며 교차검정과 상관분석을 통해 변수 간의 관계를 알아보았다. 이후 로지스틱 회귀 모형, KNN, 서포트 벡터 머신, 의사결정나무, 인공신경망을 이용하여 각 모형의 성능을 평가하고 가장 우수한 성능을 보인 인공신경망을 최종모형으로 결정하였다. 분석 결과를 통해 '나이'와 '도서 선택 정보', '최종학력', '독서 장소' 변수가 전자책 독서에 큰 영향을 미치는 변수임을 확인하였고 이를 바탕으로 향후 독서 진흥을 위한 효과적인 정책 수립에 기여할 수 있는 방향성을 제시하고자 한다.

1. 서론

우리가 살고 있는 21세기를 디지털 미디어 시대로 불리며, 정보화 시대로서의 특징을 갖추고 있다. 이러한 시대에 종이책뿐 아니라 전자책은 디지털 기술과 스마트폰 등의 환경과 함께 성장하고 적응하면서 새로운 변화를 불러오고 있다. 전자책의 등장으로 인해 많은 사람이 종이책이 사라지고 쇠퇴할 것으로 예측했지만 독자들은 여전히 종이책을 전자책보다 선호하는 모습이 두드러진다.

그런데도 전자책을 활용한 독서 활성화가 필요한 이유는 전자책 독서율이 꾸준히 증가하고 있기 때문이다. 지난해 성인 독서 인구는 47.5%. 이들의 독서량은 평균 4.5권. 2년 전보다 각각 8.2%, 3권이 줄었다. 이 통계만 보면 독서 인구는 꾸준히 감소하는 추세다. 특이한 점은 성인들이 책을 읽기 어려운 이유로 '다른 매체 콘텐츠 이용(26.2%)'이 증가했다는 점이다. 이런 영향 때문인지 전자책 독서율은 오히려 2.5%P가 증가했다고 한다. 일례로 전자책을 서비스하는 '밀리의 서재'는 어느덧 500만 회원을 돌파했고 매출은 꾸준히 상승해 올해는 기업공개(IPO)에 도전하고 있다. 또한, 공공도서관들도 시대 변화에 맞춰 종이책만 고집하지 않고 전자책을 꾸준히 늘리고 있다.

전자책은 급속히 발전하는 전자매체 기술의 발전과 함께 우리에게 새로운 독서 문화를 형성하고 있다. 갈수록 전자책의 수요가 증가하고 있고, 태블릿PC 보급으로 최근 전자책 시장이 급성장한 것으로 나타난다. 태블릿PC의 등장은 가벼우면서도 큰 양의 콘텐츠를 수용할 수 있어 어디서든 쉽게 책을 읽을 수 있게 해주고, 여러 책을 한 번에 가지고 다니기 편리하게 한다. 또, 인터넷을 통해 손쉽게 전자책을 다운로드하고, 다양한 콘텐츠에 빠르게 접근할 수 있다.

따라서 본 분석에서는 국민독서실태조사를 통해 전자책 독서실태를 알아보고 전자책 독서에 영향을 미치는 요인을 다각도로 탐구하였다. 이를 바탕으로 독서환경 변화와 전자책 독서 활성화 방안을 제시하였다. 이를 통해 출판업계와 독자 간의 상호작용을 심층적으로 이해하고 전자책 독서를 활성화를 위한 방안을 연구하여 전자책이 앞으로 우리의 다음 세대들에게 유익한 영향을 미치며 또한 기성세대들에게도 전자책을 활용해 독서가 활성화되어 향후 독서 진흥을 위한 효과적인 정책 수립에 기여하고자 한다.

2. 본론

2.1 데이터 설명

본 분석에 쓰인 데이터의 자료 분석 기준과 정보이다.

변수	설명
종이책	인쇄물 형태의 일반도서를 총칭하는 것으로 본 조사에서는 교과서, 학습참고서, 수험서를 제외한 "일반도서"만 해당됨
전자책	컴퓨터나 스마트폰, 스마트패드, 전자책 전용단말기 등을 이용하여 화면으로 읽을 수 있는 각종 디지털도서를 총칭하는 것으로 본 조사에서는 교과서, 학습참고서, 수험서를 제외한 "일반도서와 웹소설"만 해당됨
독서율	조사 대상자 중 지난 1년간('20년 09월~'21년 08월) 교과서, 참고서, 수험서를 제외한 '일반도서' 및 웹소설(장르소설), 오디오북을 1권 이상 읽은 사람의 비율, 종이책, 전자책, 오디오북 독서율로 구분됨
독서시간	하루 중 교과서, 참고서, 문제집을 제외하고 종이책, 전자책, 오디오북을 포함하여 '일반도서'를 읽은 시간을 말함, 독서시간은 평일 독서시간과 주말/공휴일 독서시간으로 구분됨
독서량	지난 1년간('20년 09월 ~ '21년 08월) 교과서, 참고서, 수험서를 제외한 '일반도서와 웹소설'을 읽은 권수. (인터넷에서 웹소설/웹툰을 읽은 경우 연재 회차가 아닌 연재 작품을 1권으로 간주)
도서관 이용률	조사 대상자 중 지난 1년간('20년09월~'21년08월) 모든 관종의 도서관을 한 번 이상 이용한 응답자 비율
독서활동 /독서활동참여율	학교, 기업, 지역사회, 도서관 등에서 제공하는 독서 관련 활동(독후 활동, 독서 캠페인, 독서 행사, 독서교육, 독서치유 등) /지난1년간('20년09월~'21년08월)학교,기업,지역사회,도서관등에서제공하는독서관련 활동에참여해본응답자비율
독서 동아리(모임) /독서동아리 참여율	학교, 지역 사회, 도서관 등에서 구성원들끼리 자발적 의사에 의해 활동하는 독서 모임 /지난 1년간('20년09월~'21년08월) 학교, 공공도서관, 동네, 인터넷 등의 독서 동아리(독서 모임)에 참여해본 응답자 비율

2.2 데이터 탐색

본 분석에서 사용된 데이터는 마이크로데이터 통합서비스(MDIS)에서 제공하는 2021년 국민 독서 실태 조사(성인)이며 원 데이터는 전국의 만 19세 이상 성인남녀 6,000명에 대한 53개의 조사 항목과 변수로 구성되어 있다. 최종적으로 선택한 변수에 대한 세부 사항은 <표 2.1.1>에서 확인할 수 있다.

자료유형	변수명	설명		
연속형	freetime_total	여가시간	60-1680 (단위: 분)	
	read.preference	독서 선호도	1-5	
	age	연령	19-95	
	household.income	가구 소득	1-8 (단위: 100만원)	
범주형	ebookmore	전자책을 더 선호하는 경우	0 : 종이책을 선호 1 : 전자책을 선호	
	read.frequency	독서 빈도	1: 전혀 이용 안함 2: 몇 달에 한 번 3: 한 달에 한 번	4: 일주일에 한두 번 5: 매일
	gender	성별	1: 남	2: 여
	read.place	독서장소	1: 집 2: 직장(학교) 3: 이동 4: 도서관	5: 서점 6: 카페 7: 어디서든 8: 기타
	choice_information	도서선택 이용정보	1: 책 직접보고 2: 신문/잡지 소개 3:텔레비전/라디오 소개 4: 인터넷 소개 5: 지인 추천	7: 베스트셀러 목록 8: 추천도서 목록 9: 드라마/영화 원작 10: SNS 소개 11: 유튜브 소개

			6: 유명인 추천	12: 기타
	job	직업	1: 관리자 2: 전문가 3: 사무종사자 4: 서비스종사자 5: 판매종사자 6: 농림어업 종사자 7: 기능종사자 8: 장치종사자	9: 단순 노무 종사자 10: 군인 11: 자영업종사자 12: 학생 13: 전업주부 14: 무직 15: 기타
	library	도서관 이용 유무	0: 이용한 적이 없다	1: 이용한 적이 있다
	area	지역	1: 대도시 2:중소도시	3: 읍면
	reading.activity	독서활동 여부	0: 참여한 적이 없다	1: 참여한 적이 있다
	reading.club	독서동아리 참여여부	0 참여한 적이 없다	1: 참여한 적이 있다
	최종.학력	최종 학력	1: 교육을 안 받았음 2: 초등학교 3: 중학교 4: 고등학교	5: 대학(4년제 미만) 6: 대학(4년제 이상) 7: 대학원 석사과정 8: 대학원 박사과정

<표 2.2.1>

2.3 전처리 / EDA

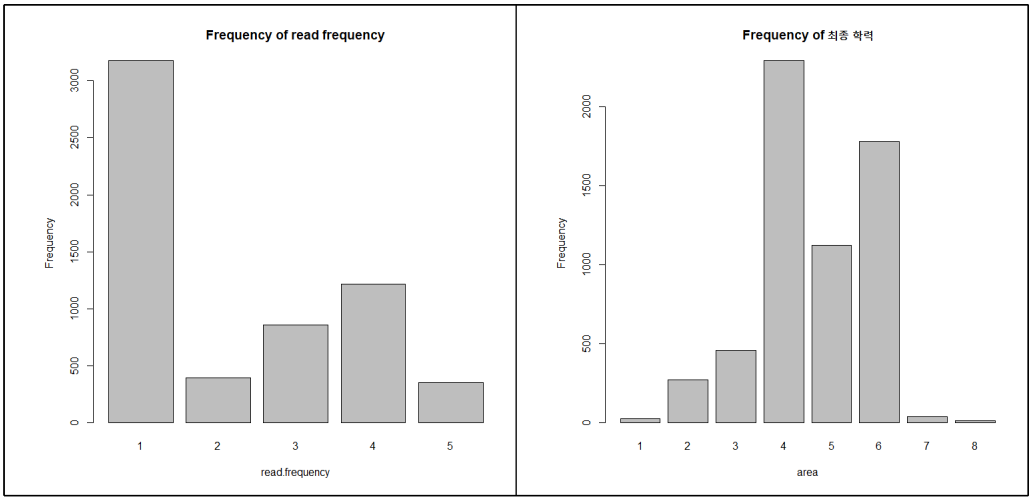
2.3.1 범주형 변수

변수명	범주		빈도		비율	
전자책 선호	1=종이책보다 전자책을 더 선호		0=5289	1=711	0=0.8815	1=0.1185
독서빈도	1=전혀이용안함 2=몇 달에 한 번 3=한 달에 한 번	4=주에 한두 번 5=매일	1=3176 2=393 3=858	4=1217 5=356	1=0.5293 2=0.0655 3=0.143	4=0.2029 5=0.0593
성별	1= 남	2=여	1=2786	2=3214	1=0.4643	2=0.5357
독서장소	1=집 2=직장(학교) 3=이동시 4=도서관	5=서점 6=카페 7=어디서든 8=기타	1=5385 2=203 3=165 4=37	5=13 6=48 7=146 8=3	1=0.8975 2=0.0338 3=0.0275 4=0.0062	5=0.0022 6=0.008 7=0.0243 8=0.0005
도서선택 이용정보	1=책 직접보고 2=신문/잡지소개 3=텔레비전/라디오소개 4=인터넷소개 5=지인추천 6=유명인추천	7=베스트셀러 목록 8=추천도서 목록 9=드라마/영화원작 10=SNS 소개 11=유튜브 소개	1=1796 2=565 3=888 4=1214 5=630 6=201	7=393 8=117 9=33 10=64 11=99	1=0.2993 2=0.0942 3=0.148 4=0.2023 5=0.105 6=0.0335	7=0.0655 8=0.0195 9=0.0055 10=0.0107 11=0.0165
직업	3=사무종사자 4=서비스종사자 5=판매종사자 6=농림어업종사자 7=기능종사자 8=장치종사자	9=단순 노무 종사자 11=자영업종사자 12=학생 13=전업주부 14=무직 15=기타	3=1029 4=823 5=682 6=316 7=372 8=192	9=242 11=565 12=319 13=1010 14=328 15=122	3=0.1715 4=0.1372 5=0.1137 6=0.0527 7=0.062 8=0.032	9=0.0403 11=0.0942 12=0.0532 13=0.1683 14=0.5467 15=0.0203
도서관 이용유무	1= 이용한 적이 있다		1=932	0=5086	1=0.1553	0=0.8477
지역	1=대도시 2=중소도시	3=읍면	1=2620 2=2780	3=600	1=0.4367 2=0.4633	3=0.1
독서활동 여부	1=참여한 적이 있다		1=96	0=5904	1=0.016	0=0.984
독서동아리 참여여부	1= 참여한 적이 있다		1=52	0=5948	1=0.0087	0=0.9913
최종 학력	1=교육을안받음 2=초등학교 3=중학교 4=고등학교	5=대학(4년제 미만) 6=대학(4년제 이상) 7=대학원 석사과정 8=대학원 박사과정	1=25 2=270 3=459 4=2293	5=1122 6=1782 7=37 8=12	1=0.0042 2=0.045 3=0.0765 4=0.3822	5=0.187 6=0.297 7=0.0062 8=0.002

<표 2.3.1.1 범주형 변수의 범주 비율>

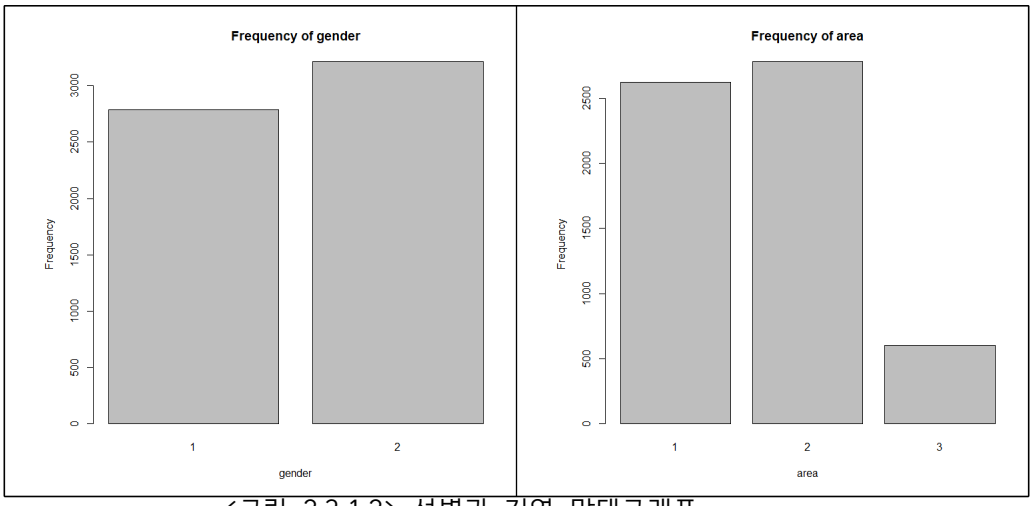
<그림 2.3.1.1>은 독서 빈도와 최종 학력의 빈도에 대한 막대그래프이다. 두 변수는 순서형

변수이다. 독서 빈도 변수의 경우 '전혀 이용하지 않는다'라는 요인이 압도적으로 많기는 하지만 나머지 요인들에 대해서는 정규분포의 형태를 띠는 것을 확인할 수 있다. 마찬가지로 최종학력 변수에 대한 막대그래프 또한 연속형 변수처럼 정규분포의 형태를 띠고 있다.



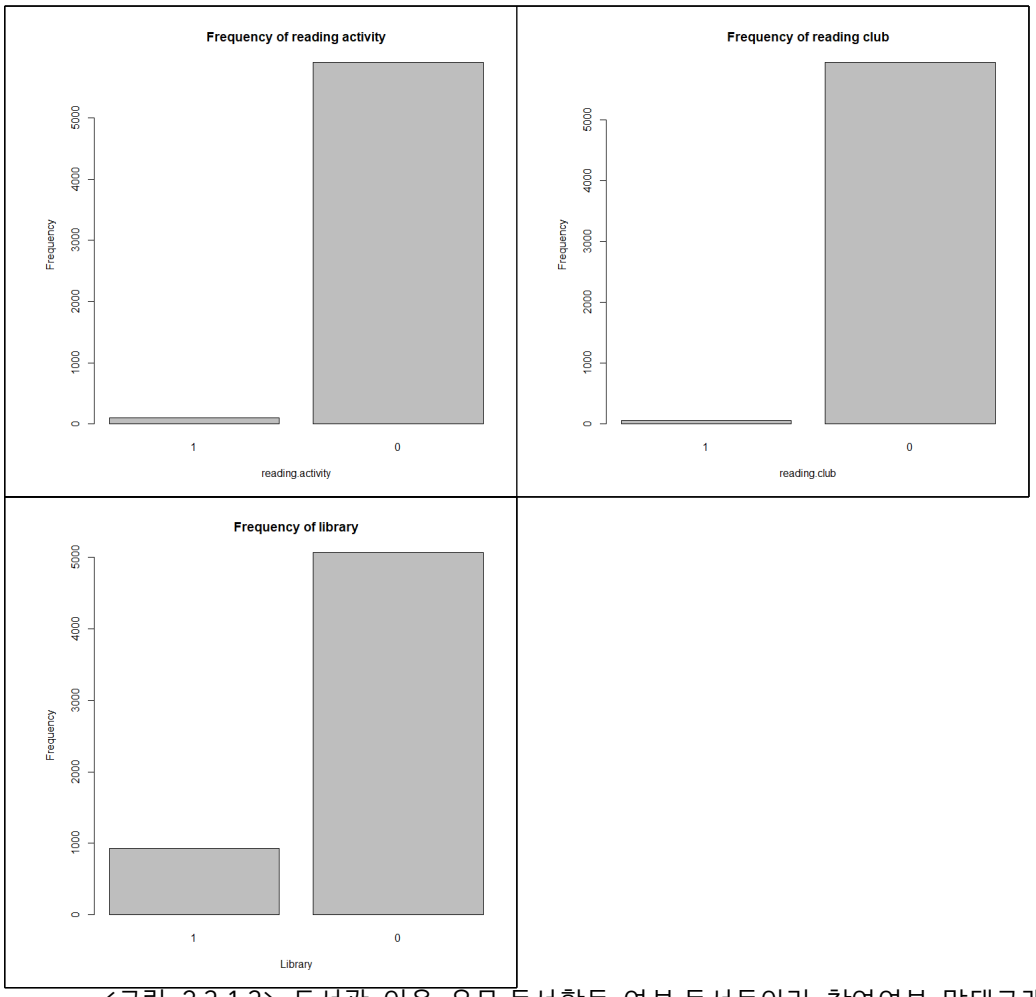
<그림 2.3.1.1> 독서 빈도와 최종 학력 막대그래프

<그림 2.3.1.2>는 성별과 지역의 빈도에 대한 막대그래프이다. 성별의 경우 남자와 여자가 비슷하게 조사되었다. 지역변수의 경우에도 3=읍면 요인은 비교적 적지만 1=대도시, 2=중소도시 요인은 비슷하게 조사되었다.



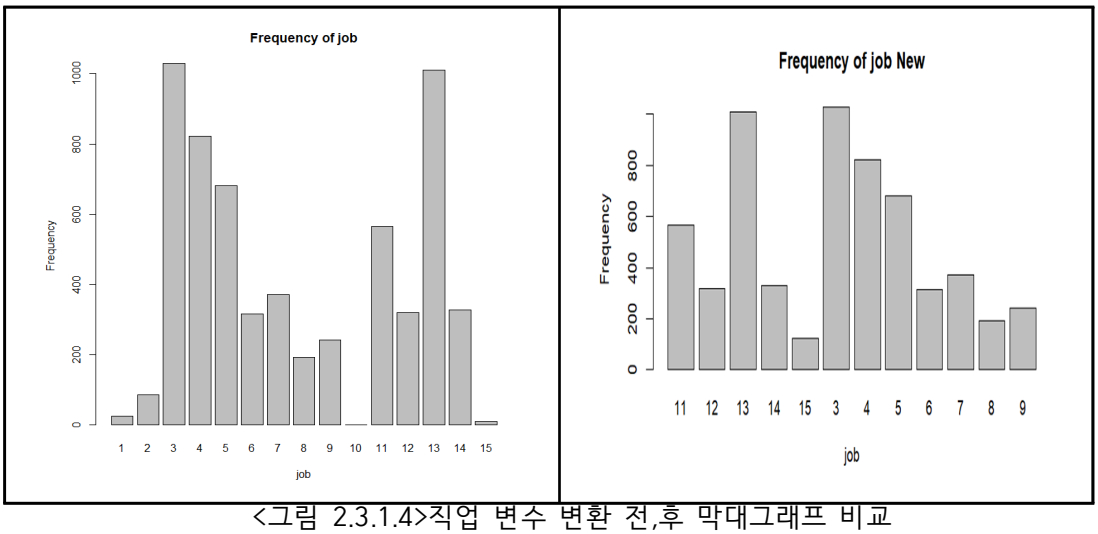
<그림 2.3.1.2> 성별과 지역 막대그래프

<그림 2.3.1.3>은 도서관 이용 유무, 독서 활동 여부, 독서동아리 참여 여부의 빈도에 대한 막대그래프이다. 세 변수의 경우'이용 안 함', '참여 안 함'을 나타내는 '0'요인이 압도적으로 많이 나타났다.



<그림 2.3.1.3> 도서관 이용 유무,독서활동 여부,독서동아리 참여여부 막대그래프

<그림 2.3.1.4>는 직업의 빈도에 대한 막대그래프이다. 범주형 변수 job의 경우 특정 요인들의 빈도가 너무 낮아서 후에 모델링 하는 과정에서 오류를 일으킬 수 있기 때문에 이러한 요인(1:관리자, 2:전문가, 10:군인) 들은 15: 기타에 포함시켰다.



<그림 2.3.1.4>직업 변수 변환 전,후 막대그래프 비교

실제 연구에서는 K-NN으로 결측치를 채운 후에 완전한 변수들을 이용하여 연관성 검정을 진행하였으나 보고서의 가독성을 위하여 다른 EDA 과정과 함께 연관성 검정을 서술하였다. 카이제곱 검정과 피셔의 정확 검정을 사용하면 각 범주형 변수와 종속변수 사이의 연관성을 확인할 수 있다.

H0: 각 범주형 변수와 종속변수인 전자책의 선호 사이에 연관성이 없다.

H1:각범주형변수와종속변수인전자책의선호사이에연관성이있다.

모든 변수에 대해서 p-value가 0.05보다 작게 나왔기 때문에 유의수준 0.05에서 귀무가설을 기각한다. 따라서 모든 범주형 변수들이 종속변수와 연관성을 띠고 있다고 할 수 있다.

변수명	카이제곱검정 p-value	피셔의 정확 검정 p-value
area	6.660373e-07	
최종.학력	1.146679e-111	
job	9.566297e-109	
read.place	3.6752e-93	
choice_information	1.739546e-107	
read.frequency	8.607765e-225	
gender	3.898e-05	3.629e-05
library	< 2.2e-16	< 2.2e-16
reading.activity	5e-08	1e-06
reading.club	0.001566	0.003616

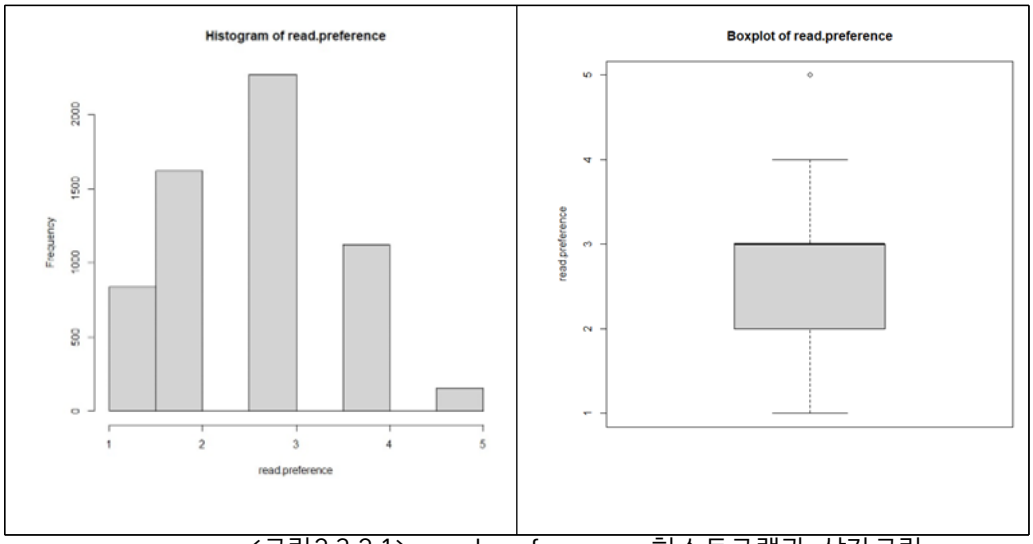
<표2.3.1.1> 범주형 변수 카이제곱 분석

2.3.2 연속형 변수

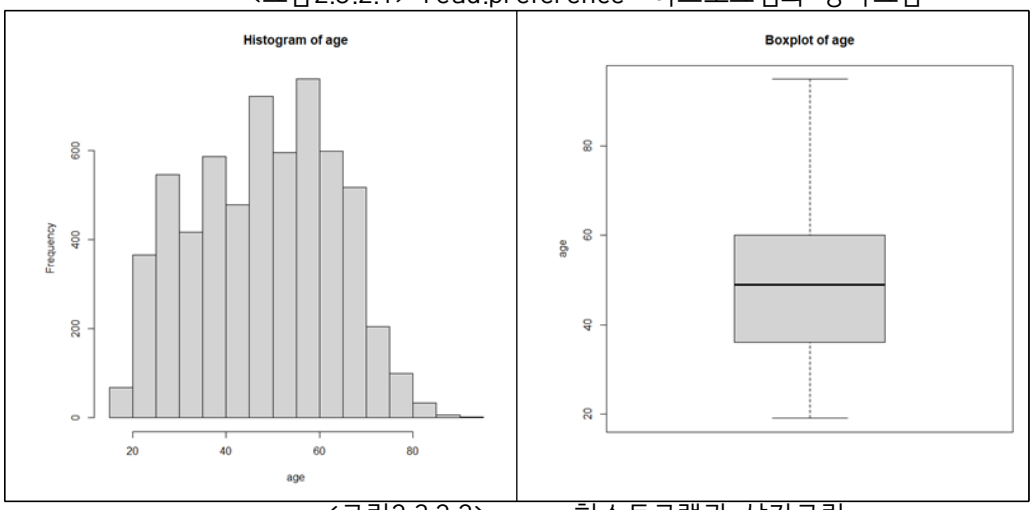
변수명	최솟값	최댓값	평균	표준편차	왜도	첨도	제1사분위수	중앙값	제3사분위수
read.preference	1	5	2.69	1.01	-0.01	-0.62	2	3	3
age	19	95	48.46	15.11	-0.05	-0.9	36	49	60
household.income	1	8	4.44	1.51	0.1	-0.23	3	4	5
freetime_total	60	1680	519.2	211.78	0.83	1.14	360	480	600

<표2.3.2.1 연속형 변수의 기술통계량>

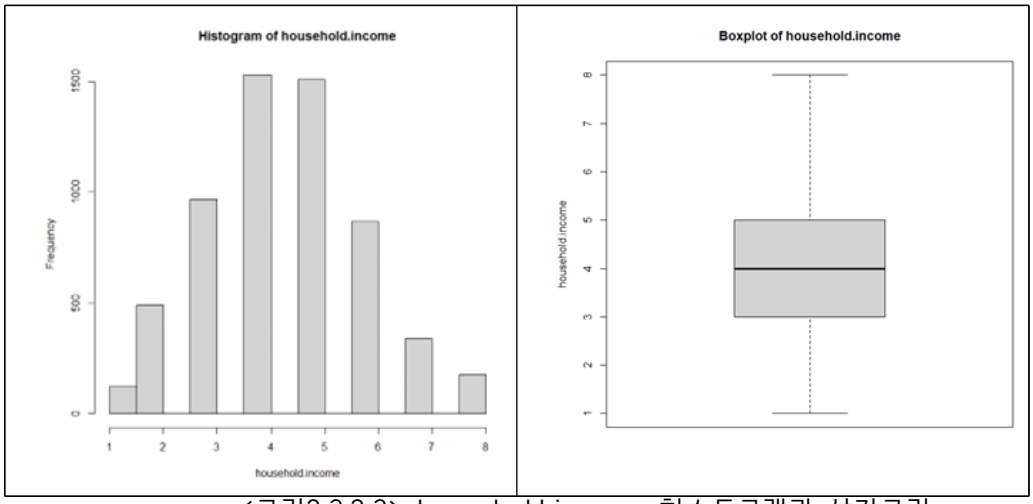
연속형 변수들에 대해서 히스토그램과 상자 그림을 그려보았을 때, 한쪽으로 치우져지지 않고 전반적으로 대칭성을 보이면서 정규분포의 형태를 띠기 때문에 추가적인 변환이 필요 없다고 판단되었다.



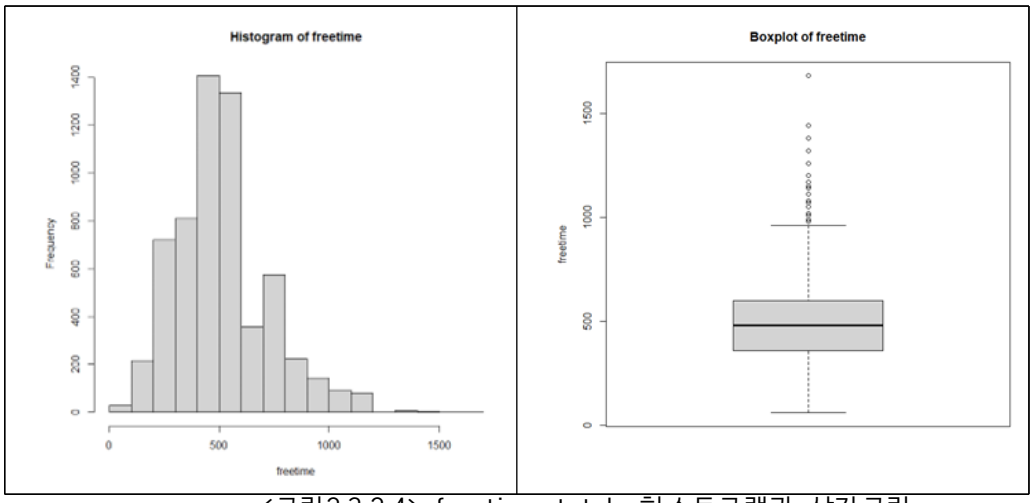
<그림2.3.2.1> read.preference 히스토그램과 상자그림



<그림2.3.2.2> age 히스토그램과 상자그림

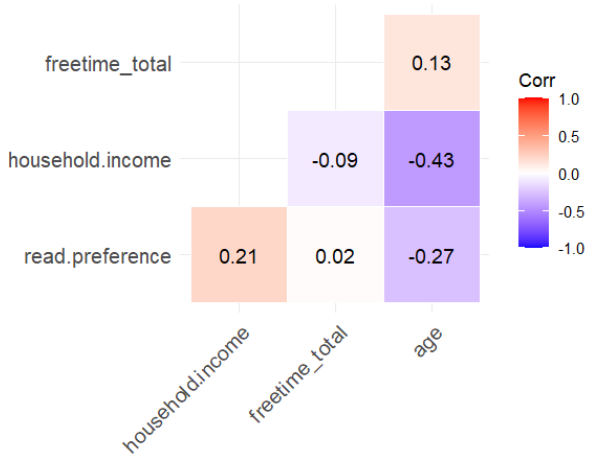


<그림2.3.2.3> household.income 히스토그램과 상자그림



<그림2.3.2.4> freetime_total 히스토그램과 상자그림

3가지 연속형 변수 간의 상관관계를 나타낸 것으로, 상관관계가 전반적으로 크지 않음을 확인할 수 있다.



<그림2.2.2.5> 연속형 변수 상관관계

연속형 독립변수들이 종속변수에 영향을 주는지 확인하기 위하여 전자책을 더 많이 읽는 집단과 종이책을 더 많이 읽는 집단으로 분리하여 각 집단 연속변수의 평균을 비교하였다.

H0: 두 집단의 연속형 변수의 평균은 동일하다.

H1: 두 집단의 연속형 변수의 평균은 차이가 있다.

t-test 결과 모든 변수에 대해서 p-value가 유의수준 0.05보다 작게 나왔기 때문에 귀무가설을 기각하였으며 두 집단의 평균이 다르다는 대립가설을 채택하였다. 따라서 사용된 연속형 변수들이 모두 종속변수와 관계가 있음을 확인할 수 있었다.

변수명	t	p-value
read.preference	25.202	< 2.2e-16
age	-34.804	< 2.2e-16
household.income	11.481	< 2.2e-16
freetime_total	-4.2658	2.195e-05

<표2.3.2.1> t-test 결과

2.4 결측치 제거

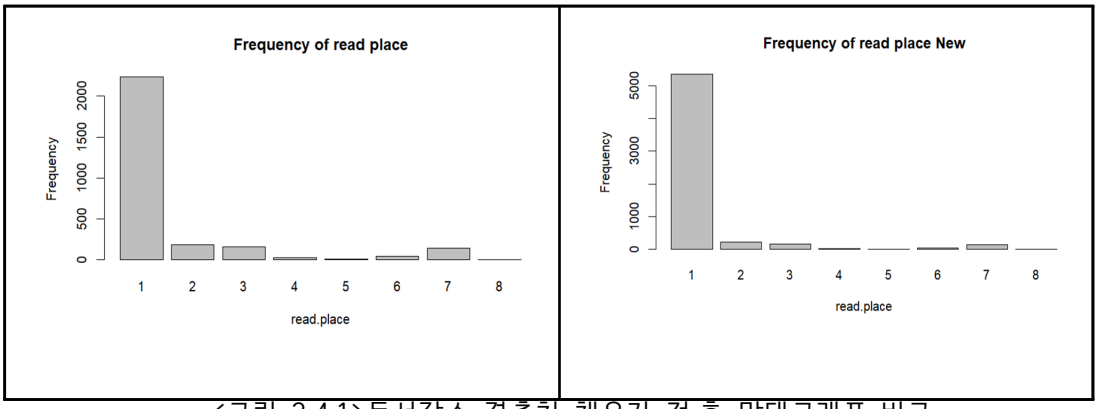
K-NN은 분류되지 않은 데이터점(개체)을 가장 비슷한 점들의 클래스 또는 그룹으로 할당하여 분류하는 과정이다. K-NN의 장점은 데이터에 대한 통계적 가정이 불필요한 비모수적 방법이므로 과정이 단순하고 모형을 학습하는 시간이 필요 없다는 것이다. KNN을 활용하여 결측치를 채우기 위해서는 적절한 K값을 사용해야 한다. K값이 커질수록 분류에서 잡음의 영향이 줄어들지만, 항목경계가 불분명해진다.

K값을 찾기 위해서 결측치가 있는 행을 제외한 행들만을 사용하여서 새로운 데이터 세트를 만들었고, 해당 데이터를 7:3의 비율로 train set과 test set으로 나눠주었다. K값을 바꿔가면서 train set을 활용하여 모형을 적합 시켰으며 이후 test set을 활용하여 얼마나 정확하게 분류하는지 확인하였다.

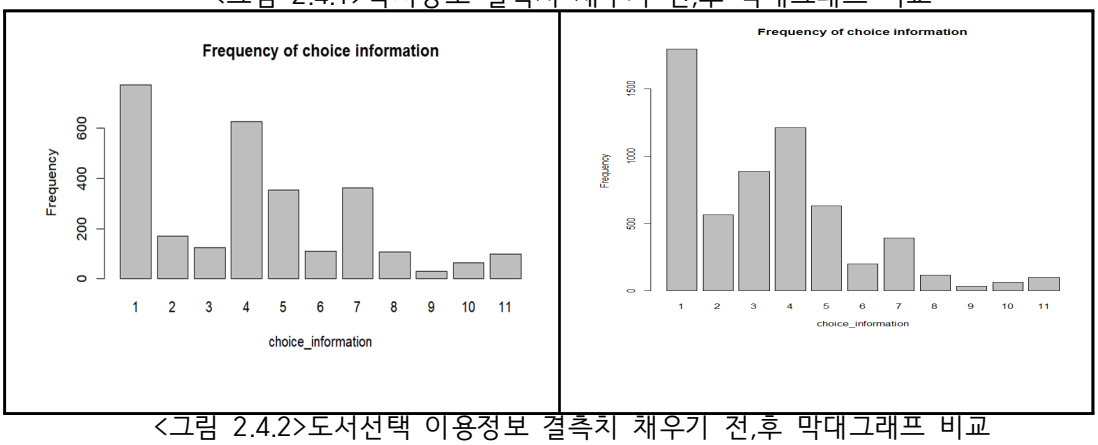
정분류율을 비교한 결과, k 값이 8일 때 독서 장소와 도서 선택 이용 정보의 분류를 가장 잘하였다고 판단하였다. 최종적으로 결측치가 있던 기존 데이터 세트를 KNN(k=8)을 사용하여 결측치가 없는 데이터로 만들었다.

K값	독서장소 정분류율	도서선택 이용정보 정분류율
5	0.7722420	0.3629893
6	0.7746145	0.3463820
7	0.7781732	0.3321471
8	0.7793594	0.3404508
9	0.7793594	0.3297746
10	0.7769870	0.3297746
11	0.7746145	0.3321471

<표 2.4.1> K값에 따른 정분류율



<그림 2.4.1>독서장소 결측치 채우기 전,후 막대그래프 비교



<그림 2.4.2>도서선택 이용정보 결측치 채우기 전,후 막대그래프 비교

3. 머신러닝 모델링

3.1 평가방법

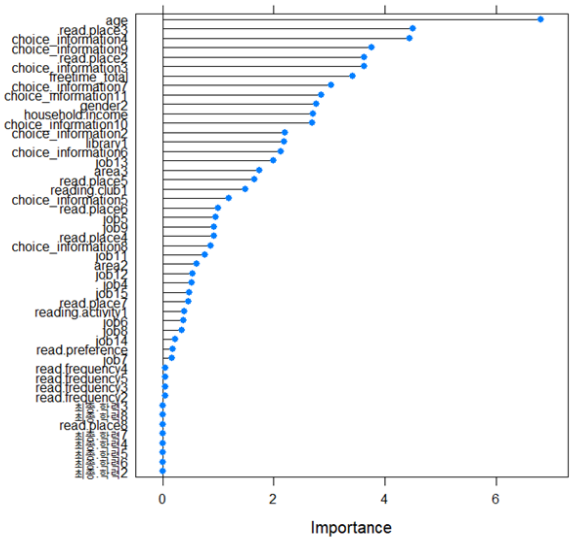
모델의 성능을 판단하기 위해서 기존의 데이터를 6 : 4의 비율로 학습용 데이터 세트와 테스트 데이터 세트로 분할하였다. 4개의 모델(로지스틱 회귀분석, 서포트 벡터 머신, 의사결정 나무, 인공신경망)을 사용하여 데이터를 분석하고 그 결과를 비교하여 하나의 최적 모형을 선택하여 결과를 도출했다.

분류 모델의 성능 평가 방법으로 각 모델의 혼동행렬의 결과를 사용하여 정확도(Accuracy), 민감도(Sensitivity), 특이도(Specificity), 정밀도(pos pred value), AUC(Area Under the Curve)를 기준으로 하여 모델을 비교했다.

3.2 로지스틱 회귀분석

로지스틱 회귀분석은 종속변수가 2개 값일 때의 중회귀분석이다. 독립변수와 종속변수 사이에 직선이 아닌 S자형 곡선(로지스틱 곡선)에 의해 분석한다. 계산 과정에서 종속변수는 어떤 사실이 일어나는 자체에서, 사실이 일어날 확률을 이용한 특수한 값(대수 오즈)으로 변환된다. 로지스틱 변환, 로짓변환이라고 부르는 이러한 변환에 의해서 독립변수가 극단적인 값을 가져도 확률이 1이 넘는다든지 마이너스가 되는 일이 없게 된다.

로지스틱 회귀 모형 기법을 통해 14개의 변수의 중요도를 계산하고, 이를 내림차순으로 정리하여 그래프로 나타낸 것이 <그림 3.2.1>이다. 중요도가 높게 나온 변수일수록 종속변수를 구분 짓는 데 큰 역할을 한 것이다. 중요도가 가장 높은 변수 6개는 'age', 'choice_information', 'read.place', 'household.income','freetime_total', 'gender'인 것을 확인할 수 있다. 해당하는 상위 6개 변수로 모형을 적합하여 새로운 축소 모형을 만들었다.



<그림 3.2.1> 로지스틱 회귀모형 중요도 그래프

분류 알고리즘의 수행 능력을 평가하기 위해서 분류 결과를 시각적으로 표로 정리한 것이 아래 <표 3.2.1>과 <표 3.2.2>의 혼동행렬이다. 각각은 전체 모형을 사용하였을 때의 결과와 축소 모형을 사용하였을 때의 결과이다. 혼동행렬의 결과를 이용해서 모델 평가 기준인 정확도(Accuracy), 민감도(Sensitivity), 특이도(Specificity), 정밀도(pos pred value), AUC(Area Under the Curve) 등을 계산할 수 있다.

		Reference	
		0	1
Prediction	0	2035	207
	1	73	85

<표 3.2.1> 전체 모형 혼동행렬

		Reference	
		0	1
Prediction	0	2064	238
	1	44	54

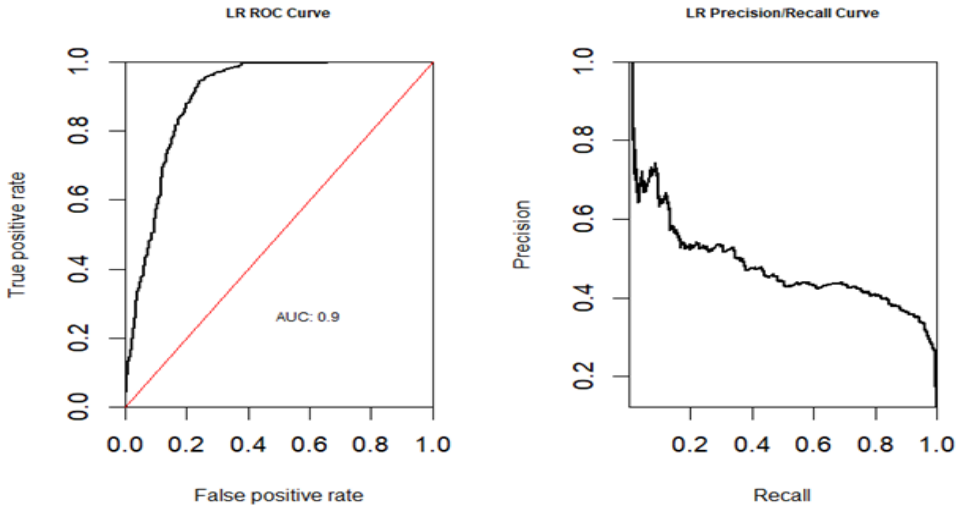
<표 3.2.2> 축소 모형 혼동행렬

정확도, 특이도, 정밀도 모두 0.02가 채 차이가 나지 않으며 차이가 있어 보이는 지표는 민감도로 약 0.1로 차이가 있다. 민감도는 실제 값이 1인 데이터에 대해서 얼마나 정확하게 예측하였는지는 나타내는 값인데 전체 모형과 축소 모형 모두에서 0으로 예측을 더 많이 하였기 때문에 민감도가 낮게 나왔다.

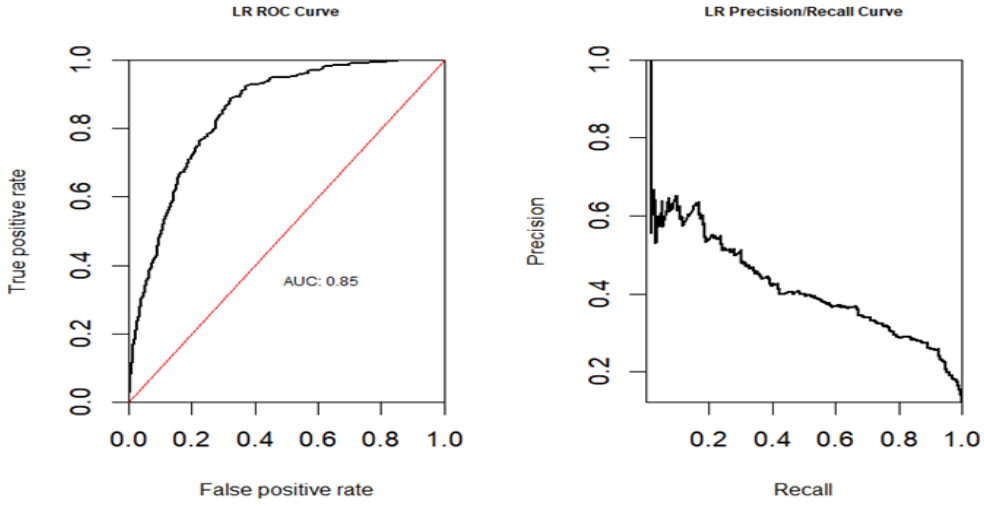
	전체모형	축소모형
정확도(Accuracy)	0.8833	0.8825
민감도(Sensitivity)	0.2911	0.1849
특이도(Specificity)	0.9654	0.9791
정밀도(pos pred value)	0.5380	0.5510
AUC(Area Under the Curve)	0.9	0.85

<표 3.2.3> 로지스틱 회귀분석 분류 모형 평가 방법

<그림 3.2.2>는 전체 모형의 ROC 곡선과 PR 곡선이고 <그림 3.2.3>은 축소 모형의 ROC 곡선과 PR 곡선이다. ROC 곡선이 왼쪽 모서리에 가까울수록 모형의 성능이 좋은 것을 의미하며, 이런 ROC 곡선을 수치상으로 나타낸 것이 AUC 값으로 ROC 곡선의 아랫부분의 넓이를 나타낸 것이다. 그렇기 때문에 값이 1에 가까울수록 모형의 성능이 좋은 것을 의미한다. 두 그림을 비교하면 전체 모형의 AUC 값은 0.9이고 축소 모형의 AUC 값은 0.85로 전체 모형이 더 높은 수치를 보인다.



<그림 3.2.2> 전체 모형의 ROC 곡선과 PR 곡선



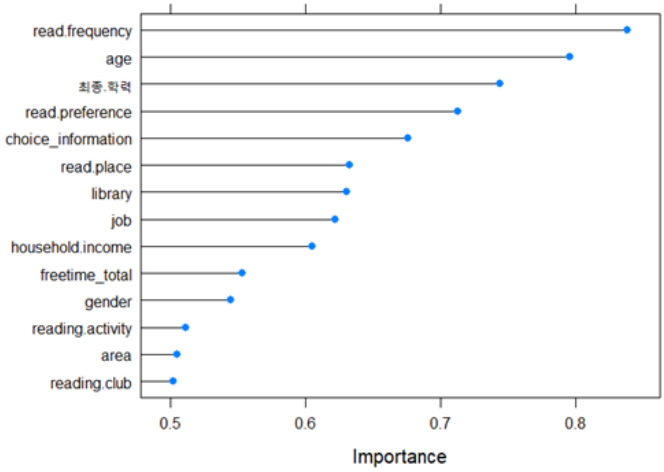
<그림 3.2.3> 축소 모형의 ROC 곡선과 PR 곡선

정확도, 민감도, AUC는 전체 모형이 더 좋게 나오고 특이도와 정밀도는 축소 모형이 더 좋게 나왔다. 하지만 앞에서 언급한 바와 같이 그 차이가 크지 않았기 때문에 더 간단한 모형인 축소 모형을 로지스틱 회귀분석의 최종 모형으로 선택하였다.

3.3 서포트 벡터 머신

서포트 벡터 머신(Support Vector Machine: SVM)은 근접 이웃(nearest neighbors)과 회귀 방법(regression method)의 개념을 결합한 모형으로, 초평면(hyperplane)이라는 공간을 만들고 결정 경계를 만들어 어떠한 공간에서 동질성을 갖는 그룹으로 분류한다.SVM의목표는한그룹의데이터포인트를다른그룹의데이터포인트와가장잘구분해내는초평면을찾는것이다.두변수사이의가장큰마진을갖는초평면일수록잘구분해내는초평면이다.마진은결정경계와서포트벡터사이의거리를의미한다.최적의결정경계는마진을최대화한것이다.

선형 분류가 가능한 경우 서포트 벡터 분류기를 사용하지만, 이 데이터 세트의 경우에는 단순한 선형 함수로 분리될 수 없어 커널(kernel) 함수를 사용하여 분류하였다. 서포트 벡터 머신도 앞서 로지스틱 회귀 분석 모형과 동일한 training set과 test set을 사용하였다.



<그림 3.3.1> 전체 모형 변수 중요도 그래프

서포트 벡터 머신을 통해 14개의 변수의 중요도를 계산하고, 이를 내림차순으로 정리한 그래프이다. 중요도가 높게 나온 변수일수록 종속변수를 구분 짓는 데 유의미한 변수이다. 중요도 상위 6개 변수는 'read.frequency', 'age', '최종.학력', 'read.preference', 'choice_information', 'read.place'이다.

혼동 행렬에서 '0'은 '독서를 하지 않음'이고 '1'은 '독서를 함'을 의미한다. 예를 들어 <표 3.3.1>의 전체 모형 혼동 행렬에서 '17'은 독서를 한다고 예측하였지만 실제로는 독서하지 않은 것이다.

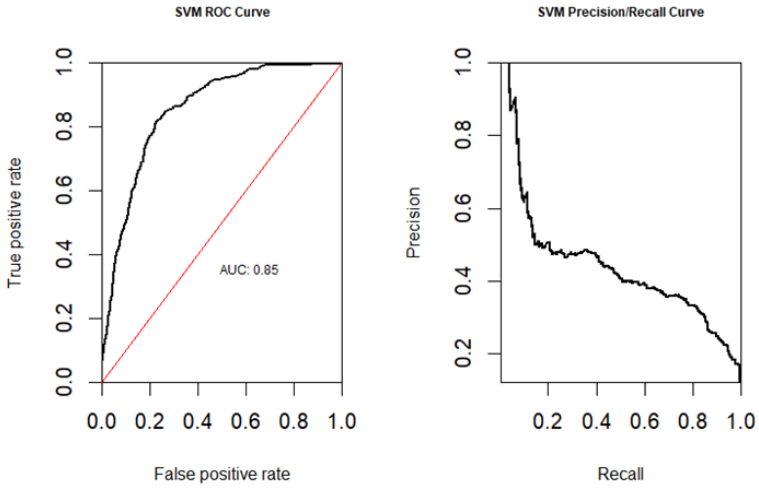
수신기 작동 특성(receiver operating characteristic : ROC) 곡선은 가능한 모든 기준값에 대하여 예측의 민감도와 특이도를 보여 주는 그림이다. ROC 곡선은 수직축에 민감도, 수평축에 (1-특이도)를 나타낸다. 특이도를 특정한 값으로 고정하면, 더 나은 예측력을 가질수록 더 높은 민감도 값을 가진다. 따라서 ROC 곡선이 더 높게 그려져 ROC 곡선 아래의 넓이가 넓을수록 예측력이 높다.

		reference	
		0	1
prediction	0	2091	263
	1	17	29

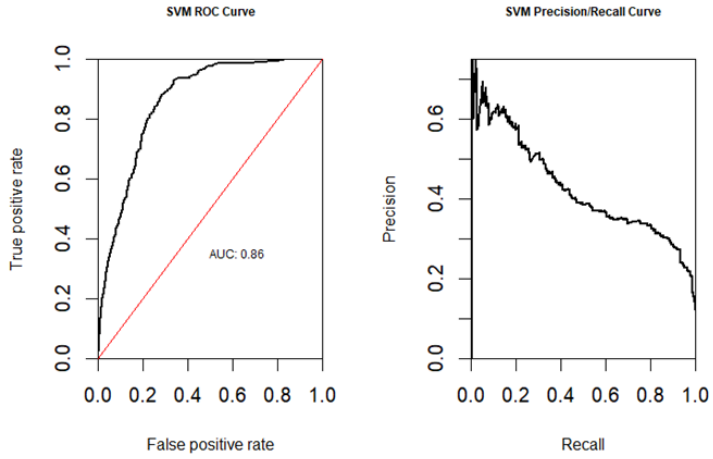
<표 3.3.1> 전체 모형 혼동 행렬

		reference	
		0	1
prediction	0	2028	209
	1	80	83

<표 3.3.2> 축소 모형 혼동 행렬

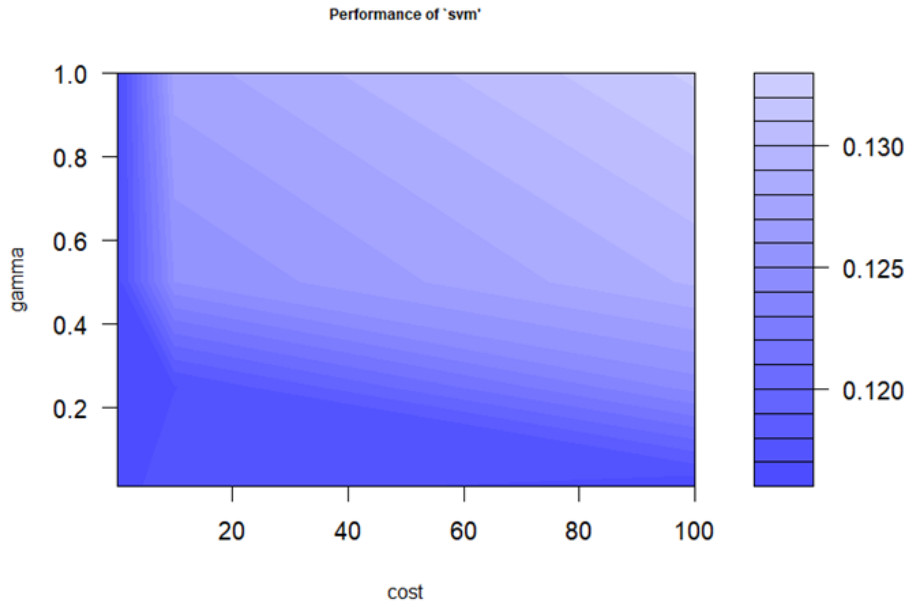


<그림 3.3.2> 전체 모형 ROC 곡선과 PR 곡선



<그림 3.3.3> 축소 모형 ROC 곡선과 PR 곡선

tune()함수를 이용하여 각 모형에 대한 Gamma를 구하고, 이를 plot() 함수를 이용하여 시각화하였다. Gamma가 작다면 reach가 멀다는 뜻이고, Gamma가 크다면 reach가 좁다는 의미이다. 여기서 reach는 결정경계의 굴곡에 영향을 주는 데이터의 범위이다. 색이 짙을수록 0에 가깝고, 이는 모형의 적합성이 높다는 의미이다. cost의 값이 20이하이거나 gamma의 값이 낮을수록 더 나은 모형임을 알 수 있다.



<그림 3.3.4> 모형별 gamma 히트맵

위의 과정을 거쳐 최적화 모델을 적합하여 혼동행렬을 도출하였다.

		reference	
		0	1
prediction	0	2078	258
	1	30	34

<표 3.3.3> 최적화모델의 혼동행렬

<표3.3.4>는 전체 모형, 축소 모형, 최적화 모형의 정확도(Accuracy), 민감도(Sensitivity), 특이도(Specificity), 정밀도(pos pred value), AUC(Area Under the Curve) 비교표이다. 정확도와 특이도, 정밀도, AUC 모두 비슷하지만, 민감도의 경우 최적화 모형의 값이 두 번째로 높았다. 모든 모델 평가 기준을 고려해보았을 때 최적화 모형이 가장 적합하다는 것을 확인할 수 있다.

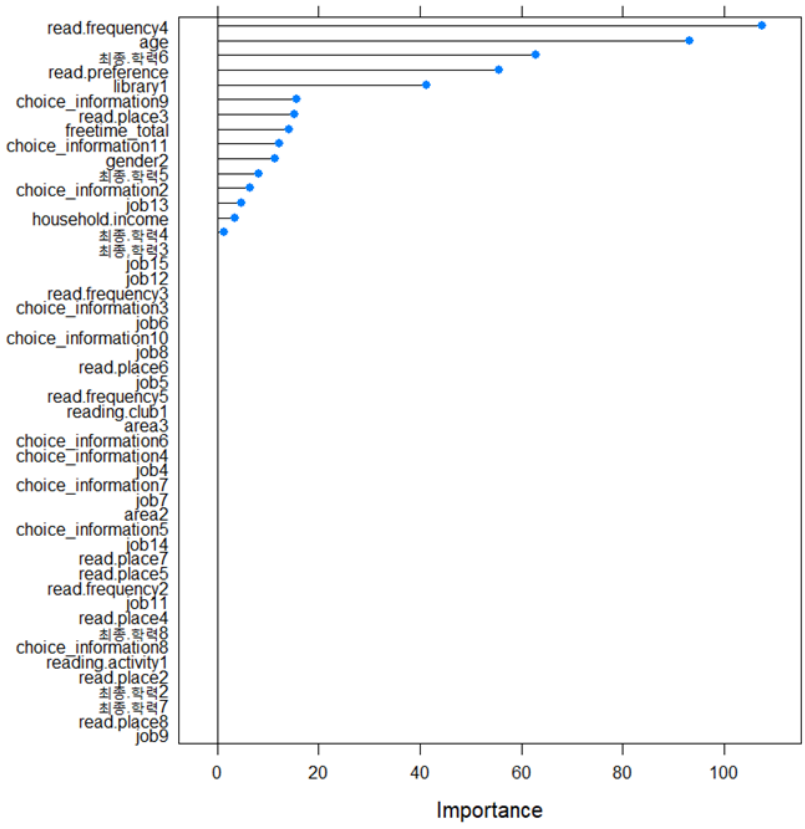
	전체모형	축소모형	최적화모형
정확도(Accuracy)	0.8833	0.8796	0.88
민감도(Sensitivity)	0.09932	0.28425	0.11644
특이도(Specificity)	0.99194	0.96205	0.98577
정밀도(pos pred value)	0.63043	0.50920	0.53125
AUC(Area Under the Curve)	0.85	0.86	0.86

<표 3.3.4> SVM 모형 평가 방법

3.4 의사결정나무

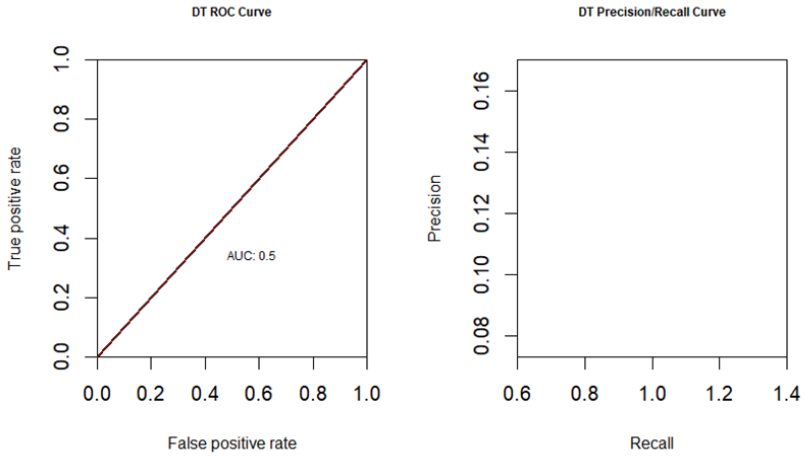
의사결정나무(decision tree)는 이름 그대로 나무 구조를 따르는 분류 모형으로, 나무를 거꾸로 세운 모양이다. 뿌리노드(root node)에서 출발하여 결정 노드(decision node)를 거치면서 각각의 가지(branch)로 나뉘고, 잎 노드(leaf node) 또는 끝 노드(terminal node)에서 의사결정은 마무리된다. 의사결정 모형의 시작은 맨 위에 있는 뿌리노드이며, 잎까지 분기가 되는 구조이다. 초기 지정은 뿌리노드이고 분기가 거듭될수록 그에 해당하는 데이터의 개수는 줄어든다. 각 끝 노드에 속하는 데이터의 개수를 합하면 뿌리노드의 데이터 개수와 일치한다. 즉, 끝 노드 간 교집합이 없도록 구성된다. 한편 끝 노드 개수는 분리된 집합의 개수가 된다.

의사결정나무 기법을 통해 전체 모형의 변수의 중요도를 계산하고, 이를 내림차순으로 정리하여 그래프로 나타낸 것이 <그림 3.4.1>이다. 중요도가 높게 나온 변수일수록 종속변수를 구분 짓는 데 유의미한 변수이다. 중요도가 가장 높은 변수 6개는 'read.frequency', 'age', '최종.학력', 'read.preference', 'library', 'choice_information'인 것을 확인할 수 있다. 해당하는 상위 6개의 변수로 모형을 적합하여 새로운 축소 모형을 만들었다.

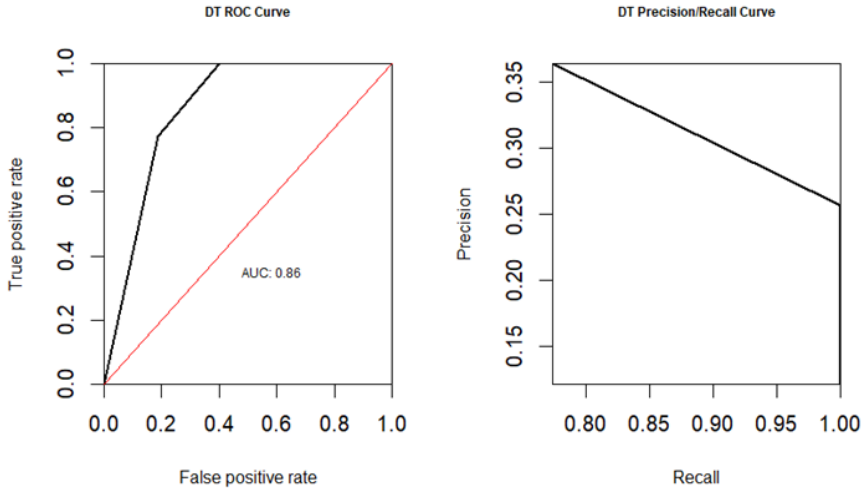


<그림 3.4.1> 의사결정나무 중요도 그래프

<그림 3.4.2>는 전체 모형의 ROC 곡선과 PR 곡선이고 <그림 3.4.3>은 축소 모형의 ROC 곡선과 PR 곡선이다. 둘을 비교하면 전체모형의 AUC 값은 0.5이고 축소 모형의 AUC 값은 0.86으로 축소 모형만이 높은 적합률을 보인다.



<그림 3.4.2> 전체 모형의 ROC 곡선과 PR 곡선



<그림 3.4.3> 축소 모형의 ROC 곡선과 PR 곡선

<표 3.4.3>은 전체 모형과 축소 모형의 정확도, 민감도, 특이도, 정밀도, AUC를 나타낸 것이다. 축소 모형이 정확도 측면에서는 전체모형보다 떨어지지만, 전반적으로 더 뛰어난 것을 확인할 수 있다. 따라서 축소 모형을 최종 모형으로 선택한다.

		Reference	
		0	1
Prediction	0	2108	292
	1	0	0

<표 3.4.1> 전체 모형 혼동행렬

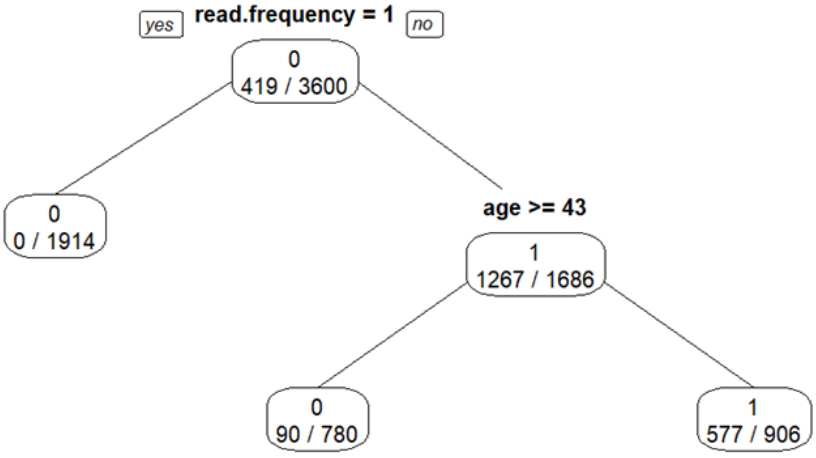
		Reference	
		0	1
Prediction	0	1713	66
	1	395	226

<표 3.4.2> 축소 모형 혼동행렬

	전체모형	축소모형
정확도(Accuracy)	0.8783	0.8079
민감도(Sensitivity)	0.0000	0.7740
특이도(Specificity)	1.0000	0.8126
정밀도(pos pred value)	NaN	0.3639
AUC(Area Under the Curve)	0.5	0.86

<표 3.4.3> 의사결정나무 분류 모형 평가 방법

<그림 3.4.4>는 최종 모형의 의사결정나무를 시각화한 것이다. 뿌리노드를 보면 '책을 전혀 읽지 않는' 경우 왼쪽 노드로, '책을 읽는 경우' 오른쪽 노드로 분기되는데, 이후 '나이'가 43세 이상이면 왼쪽 노드로 43세 미만이라면 오른쪽 노드로 분기된다. 따라서 총 3개의 끝마디로 의사결정이 마무리되었다.

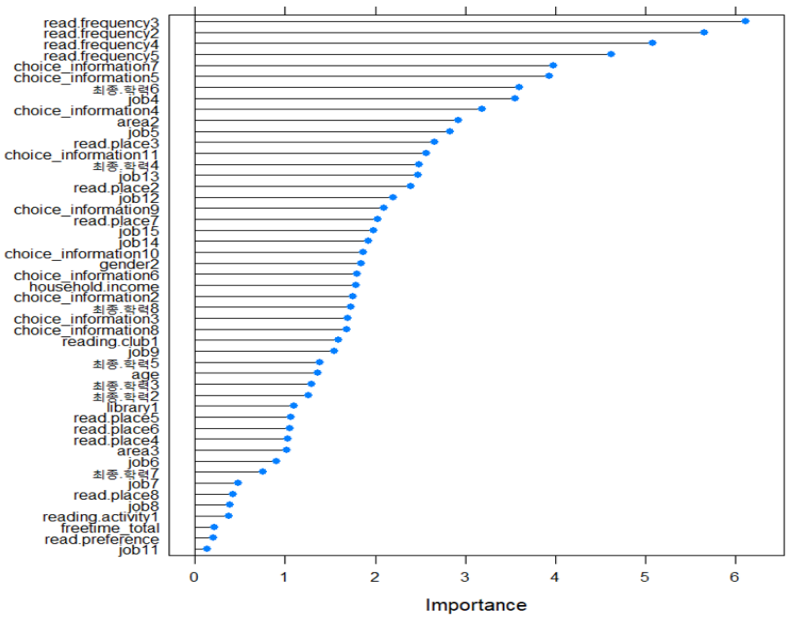


<그림 3.4.4> 의사결정나무 시각화

3.5 인공신경망

인공 신경망(ANN)은 인간의 뇌와 유사한 계층 구조로 이루어져 있는 노드나 뉴런을 사용하여 학습하는 적응형 시스템이다. 신경망은 1개의 입력 계층, 1개 이상의 은닉 계층, 1개의 출력 계층으로 구성된다. 각 계층의 노드는 이전 계층의 모든 노드에서 전달되는 출력을 입력으로써 사용하므로 모든 뉴런은 서로 다른 계층을 통해 상호연결된다. 각 뉴런에는 일반적으로 가중치가 할당되며, 이 가중치는 인공 신경망이 원하는 작업을 올바로 수행할 때까지 훈련 중 자동으로 조정된다. 가중치가 증가하거나 감소하면 해당 뉴런의 신호 강도도 변경되고, 신경망의 동작은 연결 강도나 가중치에 의해 정의된다.

먼저 모든 변수를 넣은 전체모형을 적합한 후, ANN 모델에서의 설명변수 중요도를 나타내는 <그림 3.5.1>에서 중요도 4를 기준으로 최종.학력, 지역(area), 독서빈도(read.frequency), 직업(job), 도서 선택 정보(choice_information), 독서장소(read.place) 요소 총 6가지의 변수만을 고려하여 축소 모형을 적합하였다.



<그림 3.5.1> 중요도 변수 그래프

전체 모형과 축소 모형의 혼동행렬을 각각 나타내는 <표 3.5.1>과 <표 3.5.2>를 참고하여 계산한 <표 3.5.3>은 전체 모형과 6개의 변수만을 활용하여 만든 모형을 각 모델 평가 기준으로 비교하여 정리한 것이다. 두 모형의 ROC curve를 각각 보여주는 <그림 3.5.2>와 <그림 3.5.3>을 보면 전체 모형의 ROC 커브 아래 면적이 더 넓다.

		reference	
		0	1
prediction	0	2039	215
	1	69	77

<표 3.5.1> 전체 모형 혼동 행렬

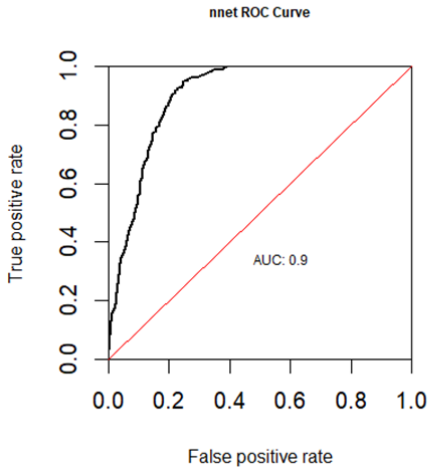
		reference	
		0	1
prediction	0	2073	254
	1	35	38

<표 3.5.2> 축소 모형 혼동 행렬

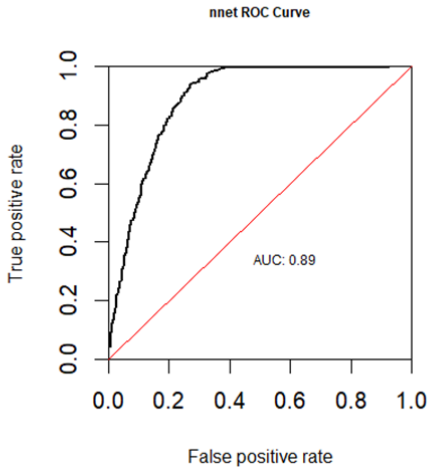
	전체모형	축소모형
정확도(Accuracy)	0.8817	0.8796
민감도(Sensitivity)	0.26370	0.13014
특이도(Specificity)	0.96727	0.98340
정밀도(pos pred value)	0.52740	0.52055
AUC(Area Under the Curve)	0.9	0.89

<표 3.5.3> 인공신경망 모형 평가 방법

정확도, 정밀도, AUC에서 전체 모형이 축소 모형보다 높은 수치를 가진다. 하지만 전체 모형의 변수는 다소 복잡하고 축소 모형보다 무겁다고 판단하여 인공신경망의 최종모형으로 축소 모형을 선정하였다.



<그림 3.5.2> 전체모형 ROC curve



<그림 3.5.3> 축소모형 ROC curve

4. 연구결과

4.1결과및최종모델선정

4개의 알고리즘으로 모델링하여 모델 성능 평가 방법을 통해 구한 값을 표 <4.1.1>에 정리했다. 앞에서 모델별로 선택한 5가지 모형들의 성능을 평가해 본 결과, 정확도 0.8796, 민감도 0.13014, 특이도 0.98340, 정밀도 0.52055, AUC 0.89로 인공신경망이 종합적으로 가장 좋게 평가되었다.

	LR	SVM	DT	ANN
정확도(Accuracy)	0.8825	0.8796	0.8079	0.8796
민감도(Sensitivity)	0.18493	0.28425	0.7740	0.13014
특이도(Specificity)	0.97913	0.96205	0.8126	0.98340
정밀도(pos pred value)	0.55102	0.50920	0.3639	0.52055
AUC (Area Under the Curve)	0.85	0.86	0.86	0.89

<표 4.1.1> 탐색 모형의 성능 지표

각 모형에서 중요한 변수를 확인한 결과, 공통으로 나온 변수는 도서 선택정보, 나이, 독서 빈도, 최종학력, 독서 장소, 독서 선호도, 도서관 방문이었다. 반응변수와의 상관계수를 확인했을 때 '독서 빈도와 '독서 선호도 모두 작은 값을 가졌기 때문에 설명변수로써 포함하였으나 결과적으로 '독서 빈도와 '독서 선호도 두 변수가 결과에서 중요도가 크게 나왔다.

4.2 결론 및 기대효과

본 연구는 여러 요인을 종합적으로 고려하여 전자책 선호에 영향을 주는 요소를 분석하기 위해 실시되었다. 설명변수는 범주형 변수와 연속형 변수로 나누어 각 변수 간의 연관성에 대해 알아보았다. 먼저, 범주형 변수 간의 연관성을 확인하기 위해 카이제곱 검정, 피셔의 정확 검정을 실시하였다. 그 결과 모든 변수의 p 값이 유의수준 0.05보다 작아 범주형 변수 사이의 연관성이 있는 것을 확인할 수 있었다. 연속형 변수의 경우 T 검정을 하여 연관성을 확인했다. 그 결과 모든 변수의 p 값이 유의수준 0.05보다 작게 나와 연속형 변수 사이의 연관성이 있는 것을 확인할 수 있었다. 연속형 변수의 경우 추가로 독립변수들 사이의 상관계수를 통해서 연관성이 있는지 확인해 봤으나 크게 관련 있어 보이는 변수는 없었기 때문에 추가적인 변환은 하지 않았다.

모든 변수를 포함한 전체모형과 변수의 중요도가 높은 상위 6개의 변수를 선택하여 만든 축소모형을 모두 고려하여 모델 성능 척도에 따라 성능을 비교하였다. 총 4개의 모형을 후보로 선택하였고, 이 모형들을 비교하기 위해 정확도, 민감도, 특이도, 정밀도, AUC를 모형 평가 기준으로 삼았다. 결과적으로 변수의 중요도를 고려한 인공신경망 축소모형을 최종 모형으로 선정했다. 인공신경망 복잡한 비선형 관계를 학습하고 대량의 데이터를 처리하고 학습하는 데 이점이 있어 본 데이터를 설명하기 적합하다고 할 수 있다. 이를 통해 전자책 독서에 큰 영향을 끼치는 요인은 '독서 선호도', '독서 빈도', '나이', '도서 선택 정보', '최종학력', '독서 장소'임을 알 수 있다. 이 중 독서 선호도와 독서 빈도의 경우 조절이 불가하여 활용하기 어렵지만 나이와 도서 선택 정보, 최종학력, 독서 장소를 통하여 전자책 이용량을 파악할 수 있을 것으로 전망한다.

이 모델을 활용하여 독서환경 변화와 전자책 독서 활성화 방안에 대해 다음과 같이 생각해 볼 수 있었다. 서점과 도서관의 온라인 서비스에서 이용자의 특성에 따라 전자책을 검색 결과 상위에 도출될 수 있게 하여 이용 시 편리함을 제공할 수 있다. 출판업계는 연령과 같은 독자의 특성에 맞춰 출판하고 도서 선택 정보를 이용해 홍보 방식에 대해 고려해 볼 수 있을 것이다. 이러한 방안을 통해 독서량 상승과 독서 문화 증진을 위한 정책에 활용할 수 있을 것이다.

참고문헌

박은영."전자책독서현황과활성화방안연구."국내석사학위논문韓國外國語大學校敎育大學院,2019.서울

김재희,"R을이용한다변량및빅데이터통계분석",르네싸이,2023.

3.2 로지스틱 회귀분석
Wataru Koyano , 김명수 역, 수학이 거북한 사람을 위한 다변량해석가이드,2003

3.3 SVM

김재희, "R을 이용한 다변량 및 빅데이터 통계분석", 르네싸이, 2023. p128

MathWorks.com, "Support Vector Machine", 2023.
(https://kr.mathworks.com/discovery/support-vector-machine.html?s_tid=srchtitle_site_search_1_svm)

Alan Agresti, "범주형 자료분석 개론", 자유아카데미, 2021. p147

나종화, "데이터마이닝 및 분석", 충북대학교, 2017.

3.5 인공신경망

MathWorks.com, "신경망이란?", 2023
(https://kr.mathworks.com/discovery/neural-network.html)

박은영. "전자책 독서 현황과 활성화 방안 연구." 국내석사학위논문 韓國外國語大學校 敎育大學院, 2019. 서울
김재희, R을 이용한 다변량 및 빅데이터 통계분석, 2023
Wataru Koyano , 김명수 역, 수학이 거북한 사람을 위한 다변량해석가이드,2003
김재희, R을 이용한 다변량 및 빅데이터 통계분석, 2023
MathWorks.com, Support Vector Machine, 2023
Alan Agresti, 범주형 자료분석 개론, 2021
나종화, 데이터마이닝 및 분석, 충북대학교 kocw, 2017
김재희, "R을 이용한 다변량 및 빅데이터 통계분석", 르네싸이, 2023.
MathWorks.com, "신경망이란?", 2023