통계분석방법론 기말 프로젝트

은행 고객의 정기 예금 상품 가입 분석

2020020344 통계학과 정해성

은행 고객의 정기 예금 가입 분석

2020020344 통계학과 정해성

Ⅰ. 개요

기업이 마케팅에 데이터 분석을 사용하는 목적 중 하나는 적은 비용으로 마케팅을 진행하여 최대한의 수익을 창출하기 위함이다. 이는 현재 진행 중이거나 완료한 마케팅의 성과와 원인을 이해하는 것으로부터 시작한다. 그 후 최적화, 고객 세분화. 예측 및 평가 단계를 거쳐 실질적으로 수익 창출을 달성할 수 있다.
이번 프로젝트에서는 통계 분석 방법론을 수강하며 익힌 기법들을 마케팅 관련 데이터에 적용해보고자 고객의 정기 예금 가입 여부에 대한 분석을 프로젝트의 주제로 선정하였다. 데이터는 UCI Machine Learning Repository의 Bank Marketing Data Set을 사용했으며 총 20개의 features와 1개의 binary target variable이 존재한다. 샘플은 총 41,188개이다. 모형의 성능을 평가하기 위하여 EDA 및 전처리 전 층화추출 방법으로 약 10%가량의 4119개의 테스트 데이터를 별도로 준비하였다. 그 후 나머지 37069개의 훈련용 데이터로 EDA 및 전처리, 그리고 모형 적합을 진행하였으며, 테스트 데이터를 사용해 분류 결과의 일반화에 대한 평가를 진행하였다.

Ⅱ. 데이터의 형태

데이터는 포르투갈 은행 기관의 전화 통화를 통해 이루어진 마케팅 데이터이다. 이번 과제의 목표는 마케팅을 통해 수집한 데이터를 이용하여 고객이 정기 예금에 가입할 것인지 여부를 예측하는 모형을 만드는 것이다. 데이터는 크게 4가지의 정보가 들어있으며 각 범주마다의 특성 변수들은 다음과 같다.

<표 1> Bank Marketing Data Set의 변수 설명

범주	변수명	설명
Bank client data	Age	나이
	Job	직업
	Marital	결혼 여부
	Education	교육 수준
	Default	신용 불이행 여부
	Housing	주택 담보 대출 여부
	Loan	개인 융자 여부
Related with the last contact of the current campaign	Contact	연락 유형
	Month	마지막으로 연락한 월
	Day_of_week	마지막으로 연락한 요일
	Duration	마지막 통화 시간(단위:초); **마케팅 종료 전에는 알 수 없는 정보**

Other attributes	Campaign	이번 마케팅 기간동안 연락한 횟수
	Pdays	이전 마케팅에서 마지막 연락 후 경과 일수 (999:이전 연락이 없음)
	Previous	이전 마케팅에서 연락한 횟수
	Poutcome	이전 마케팅의 결과
Social and economic context attributes	Emp.var.rate	고용 변동률-분기별
	Cons.price.idx	소비자 물가지수-월간
	Cons.conf.idx	소비자 신뢰지수-월간
	Euribor3m	3개월 만기 유리보율-일간
	Nr.employed	고용자 수-분기별

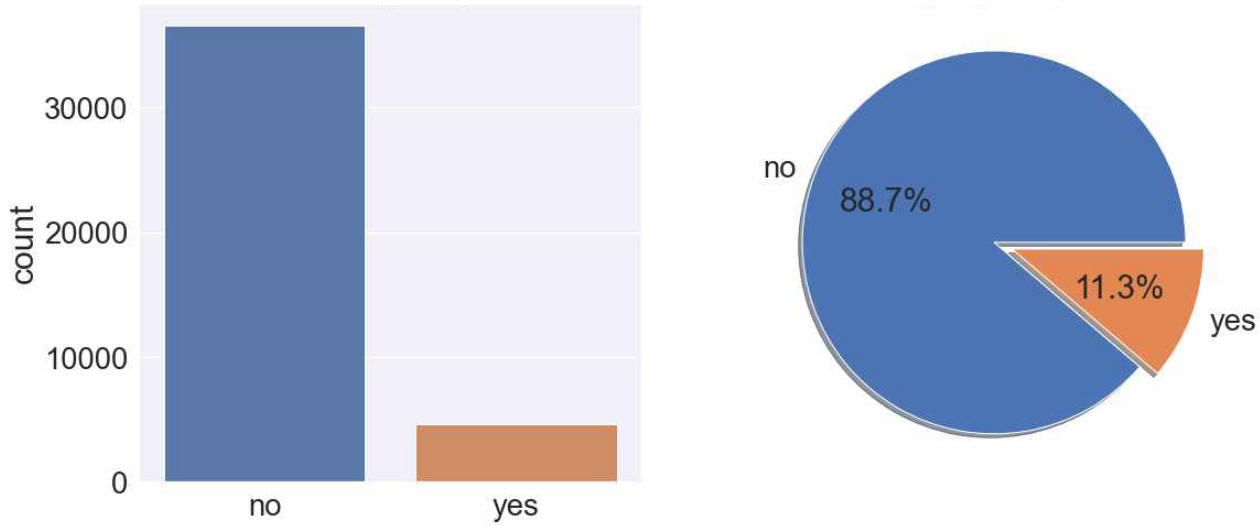
*유리보: 유료화를 단일통화로 하는 유럽연합 12개국의 시중은행 간 단기차입 금리

범주	변수명	설명
종속	y	고객의 정기 예금 가입 여부 ('yes', 'no')

대부분의 독립 변수는 범주형 또는 순서형으로 이루어져 있으며 종속 변수 y는 이항변수이므로 주어진 문제는 이항 분류 문제라고 정의 내릴 수 있다.

주목할 점은 이번 과제의 주요 관심사인 y의 값이 <그림 1>에 나타난 것처럼 'no'가 88.7%, 'yes'가 11.3%로 분포가 매우 불균형하다는 점이다. 그러므로 각 모형의 분류 결과를 평가할 때에도 이러한 불균형을 고려하여야 한다.

<그림 1> 종속 변수 y의 분포



이에 따라 우선 'no'를 0으로 'yes'를 1로 코딩한 뒤 훈련 데이터와 시험 데이터로 분리할 때, y의 0과 1의 비율에 맞게 층화추출을 진행했다. 즉, Train set의 차원은 (37069, 21), Test set의 차원은 (4119, 20)으로 분리했다. 따라서 이후의 분석에서 train set만을 사용해 EDA 및 모형 적합을 시도하였으며, test set은 모형 평가에만 사용하였다.

Ⅲ. EDA 및 전처리

훈련 데이터의 각 범주에 속한 변수들의 분포와 특징을 살펴보고 간단한 전처리를 시행한다. 각 변수를 살펴보기 전, 주어진 데이터의 가장 특이한 점은 <표 2>에서 보듯 몇 개의 변수가 unknown이라는 값을 가진다는 것이다. 즉, 주어진 데이터는 결측값이 'unknown'으로 코딩되어 있다.

<표 2> unknown을 포함하는 변수와 unknown의 개수, 비율 (train)

변수명	unknown의 개수	비율
job	296	0.80%
marital	69	0.19%
education	1564	4.22%
default	7773	20.97%
housing	893	2.41%
loan	893	2.41%

총 6개의 변수가 1개 이상의 'unknown' 값을 가지고 있다. 추후 각 변수에 대해 이 값을 어떻게 처리할지 EDA를 통해 결정하기로 한다. 특히, default의 경우 7773개로 전체 데이터의 약 21%에 달하는 값이 'unknown'이기 때문에 단순히 제거하는 방식은 지양한다.

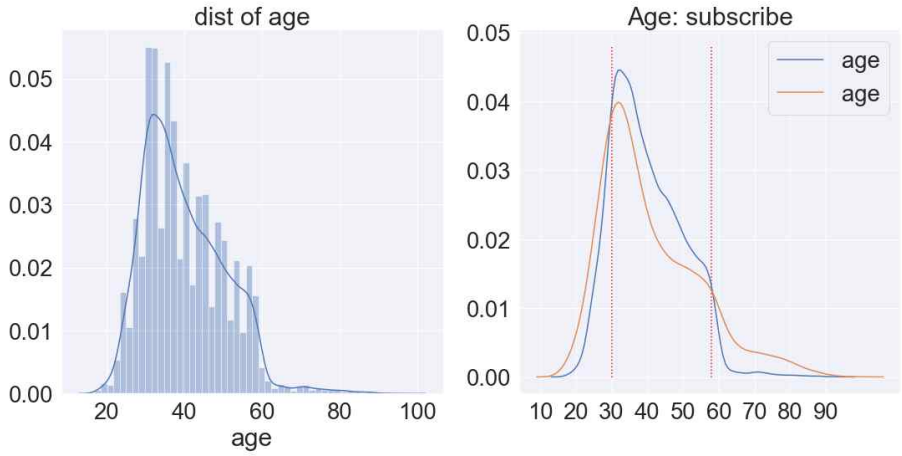
또한, 대부분의 독립 변수가 범주형이기 때문에 범주형 변수인 독립 변수에 대해서는 피어슨의 카이제곱 검정을 통해 종속변수 y와의 연관성을 검정하기로 한다. 특히, unknown이 큰 영향을 미치거나, 독립 변수의 수준에 따라 종속 변수 y와 연관성이 보이지 않는 변수는 피어슨의 카이제곱 통계량을 이용해 변수 선택의 일환으로 제거하기로 한다. 이때 검정을 시행하는 변수 A에 대하여 귀무가설 H_0와 대립가설 H_1은 다음과 같다.

H_0 : A와 y는 독립이다.
H_1 : A와 y는 독립이 아니다.

1. Bank Client Data

1.1 age

<그림 2> age 변수의 분포와 y별 분포



age 변수를 살펴보면 30세 미만과 58세 이상에서 정기 예금에 가입한 사람이 더 많다. 이처럼 나이 분포에 따른 정기 예금 가입 여부가 차이가 나기 때문에 age는 y를 예측하는 데 중요한 변수라고 생각할 수 있으며, age를 30세 이하, 30 ~ 58세 이하, 59세 이상으로 이루어진 age_cat이라는 범주형 변수를 생성했다. age_cat의 각 수준에 따른 정기 예금 상품 가입 여부 비율은 <표 3>과 같다.

<표 3> age_cat과 y의 교차분석표 및 y 값별 비율

age_cat	y		
	no	yes	
0	5640 (84.6%)	1026 (15.4%)	6666
1	26251 (90.8%)	2660 (9.2%)	28911
2	1002 (67.2%)	490 (32.8%)	1492
	32893	4176	37069

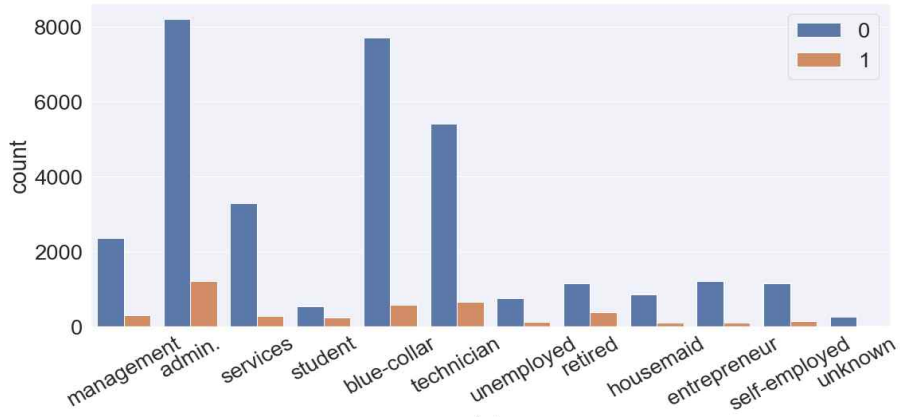
$\chi^2 = 931.7,\ df = 2,$ p<0.001

59세 이상인 사람의 33%가 정기 예금에 가입했으며, 30세 이하는 15%, 31세~58세는 9%로 범주별로 정기 예금 가입률에 차이가 크게 난다. 또한, age_cat과 y의 피어슨 카이제곱 검정의 결과는 p 값이 유의수준 0.05 보다 매우 작으므로 'H_0: age_cat과 y는 독립이다.'를 기각한다.

1.2 job

<그림 3> 직업별 정기 예금 가입률과 그래프

job	student	retired	unemployed	admin.	unknown	management	technician	self-employed	housemaid	entrepreneur	services	blue-collar
y	0.317164	0.253255	0.145856	0.129966	0.118243	0.111867	0.107979	0.105723	0.098517	0.085431	0.080785	0.068187



학생(student)과 은퇴자(retired)의 정기 예금 가입률은 각각 31.7%와 25.3%로 가장 높았으며, 그 외 직업들은 6.8%(blue-collar) ~ 14.6%(unemployed) 사이에 머물러있다. 또한, 'job'은 296개의 'unknown' 값을 갖는데, 총 37069개의 샘플 중 약 0.8%에 불과하므로 큰 영향을 미치지 않고 있다. 그러므로, job 변수의 값이 unknown인 296개의 관측치를 제거하기로 한다.

1.3 marital

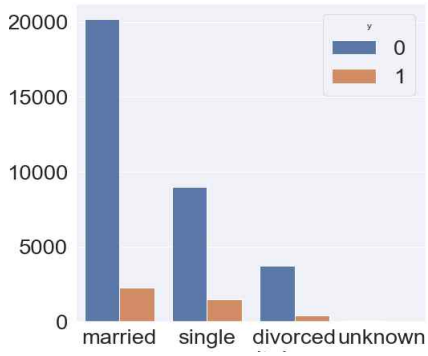
- marital 변수는 총 69개의 unknown을 갖고 있다.

<표 4> marital과 y의 교차분석표 및 y 값별 비율

marital	y		
	no	yes	
divorced	3712 (89.7%)	425 (10.3%)	4137
married	20143 (89.8%)	2281 (10.2%)	22424
single	8980 (86.0%)	1459 (14.0%)	10439
unknown	58 (84.1%)	11 (15.9%)	69
	32893	4176	37069

$\chi^2 = 109.1,\ df = 3,$ p-value<0.001

<그림 4> y별 marital의 분포



또한, 'yes'의 비율 역시 다른 수준과 큰 차이를 보이지 않으므로 'unknown'은 y에 큰 영향을 미치지 않는 것으로 보이므로 marital이 unknown인 관측치를 제거하기로 한다. 또한, marital이 'single'일 때 14.1%로 divorced 혹은 married보다 예금 가입률이 소폭 높다. 또한, marital과 y의 피어슨 카이제곱 검정의 결과는 p 값이 유의수준 0.05 보다 매우 작으므로 'H_0: marital과 y는 독립이다.'를 기각한다. 즉, marital과 y는 연관성이 있다.

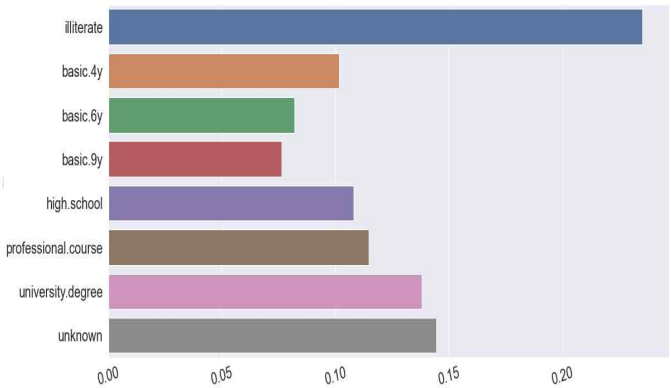
1.4 education

<표 5> education과 y의 교차분석표

education	y		
	no	yes	
illiterate	13	4	17
basic.4y	3367	381	3748
basic.6y	1901	170	2071
basic.9y	5021	415	5436
high.school	7625	924	8549
professional.course	4176	541	4717
university.degree	9452	1515	10967
unknown	1338	226	1564

$\chi^2 = 187.3,\ df = 7,$ p-value<0.001

<그림 5> education 별 정기 예금 가입 비율



순서형 변수인 교육 수준과 정기 예금 가입률은 감소 후 증가하는 특정 패턴이 보인다. 또한, unknown은 전체 훈련 데이터의 약 2%에 해당하는 69개의 값이지만, 정기 예금 가입률이 약 15.9%로 매우 높은 특징을 보이기 있어 쉽게 제거할 수 없다. 따라서, 가장 비슷한 비율을 갖는 'university.degree'로 다시 코딩하는 것이 합리적이라고 판단했다.

1.5 default, housing & loan

- default 변수는 yes가 오직 3개의 값만 존재하며 이는 전체 훈련 데이터의 0.009%에 불과하고 모두 y=0의 값을 갖고 있다. 또한, unknown 값이 7773개나 존재하기 때문에 실질적으로 이 변수는 y에 대한 정보를 주지 않고 있다고 볼 수 있다. 따라서, 모형을 적합할 때 이 변수를 사용하지 않기로 한다.

- 또한, 주택 담보 대출 여부(housing)와 개인 융자 여부(loan)은 각각 893개의 unknown을 갖고 있다. 또한, 'yes', no, unknown 별 y=1의 비율의 차이가 없는 것처럼 보이기 때문에 피어슨의 카이제곱 검정을 진행했다.

<표 6> housing과 y의 교차분석표 및 카이제곱검정 결과

housing	y		
	no	yes	
no	14919	1825	16744
unknown	794	99	893
yes	17180	2252	19432

$\chi^2 = 4.31,\ df = 2,\ p-value = 0.12$

<표 7> loan과 y의 교차분석표 및 카이제곱검정 결과

loan	y		
	no	yes	
no	27065	3453	30518
unknown	794	99	893
yes	5034	624	5658

$\chi^2 = 0.42,\ df = 2,\ p-value = 0.81$

카이제곱 검정결과 housing과 y, loan과 y의 검정 통계량은 각각 4.31, 0.42로 유의수준 0.05 하에서 'H_0 : housing과 y는 독립이다.'와 'H_0 : loan과 y는 독립이다.'라는 귀무가설을 기각하지 못한다. 그러므로, housing과 loan 역시 모델링에 사용하지 않는다.

2. Related with the last contact of the current campaign

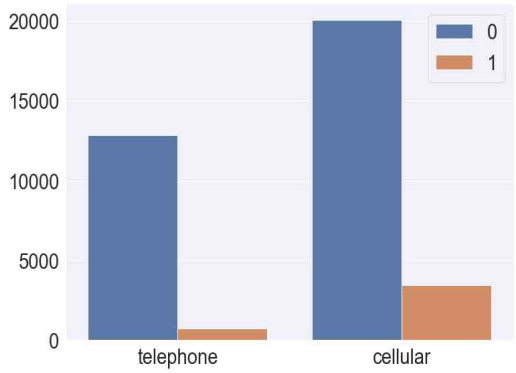
2.1 contact

<표 8> contact와 y의 교차분석표 및 y 값별 비율

contact	y		
	no	yes	
cellular	20053 (85.3%)	3462 (14.7%)	23315
telephone	12840 (94.7%)	714 (5.3%)	13554

$\chi^2 = 767.9,\ df = 1,$ p-value<0.001

<그림 6> y 수준별 contact의 분포



<표 8>과 <그림 6>을 통해 Contact의 수준에 따라 정기 예금 가입률이 크게 달라지는 것을 확인할 수 있다. 이는 카이제곱 검정의 결과와도 일치한다. 즉, 카이제곱 검정 통계량이 767.9로 유의수준 0.05 하에서 기각역인 3.841보다 매우 크므로 'H_0 : contact와 y는 독립이다.'라는 귀무가설을 기각한다. 즉, contact와 y는 연관성이 있다.

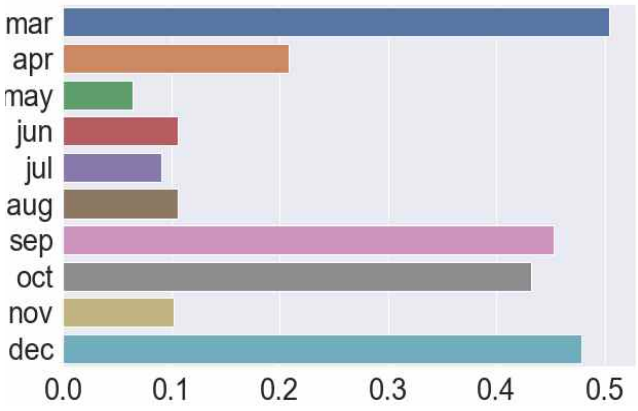
2.2 month

<표 9> month와 y의 교차분석표

month	y		
	no	yes	
mar	246	250	496
apr	1851	486	2337
may	11632	792	12424
jun	4268	503	4771
jul	5909	591	6500
aug	4957	586	5543
sep	288	238	526
oct	366	278	644
nov	3291	374	3665
dec	85	78	163

$\chi^2 = 2791.2,\ df = 9,$ p-value<0.001

<그림 7> month 별 정기 예금 가입 비율



month 변수에서 주목할 점은 1월과 2월에는 데이터가 없다는 것이다. 즉, 마케팅은 모두 3월~12월에 진행되었다. 또한, 대부분의 연락이 may(5월)에 몰려있다. 하지만 가장 적은 비율 가입률을 보이며, 오히려 mar(3월), sep(9월), oct(10월), dec(12월)은 상대적으로 적은 통화량이나 마케팅이 성공(y=1)일 확률이 훨씬 높다. 또한, month와 y의 피어슨 카이제곱 검정의 결과는 p 값이 유의수준 0.05 보다 매우 작으므로 'H_0 : month와 y는 독립이다.'를 기각한다. 즉, month와 y는 연관성이 있다.

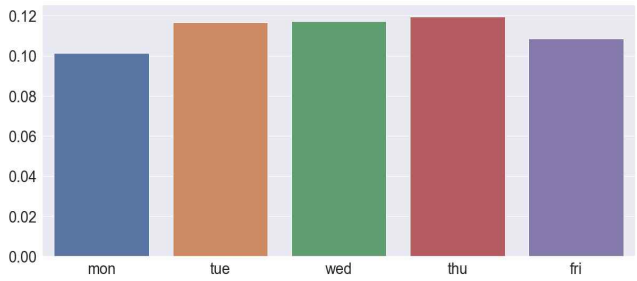
2.3 day_of_week

<표 10> day_of_week와 y의 교차분석표

day_of_week	y		
	no	yes	
mon	6882	777	7659
tue	6411	847	7258
wed	6482	861	7343
thu	6811	924	7735
fri	6307	767	7074

$\chi^2 = 17.21,\ df = 4,$ p-value=0.002

<그림 8> day_of_week 별 정기 예금 가입 비율



day_of_week 변수의 주목할만한 점은 주말에는 마케팅을 진행하지 않았다는 것이다. 마케팅을 목요일에 진행할 때 정기 예금 가입률이 가장 높았다. 또한, day_of_week와 y의 피어슨 카이제곱 검정의 결과는 p 값이 유의수준 0.05 보다 작으므로 'H_0 : day_of_week와 y는 독립이다.'를 기각한다.

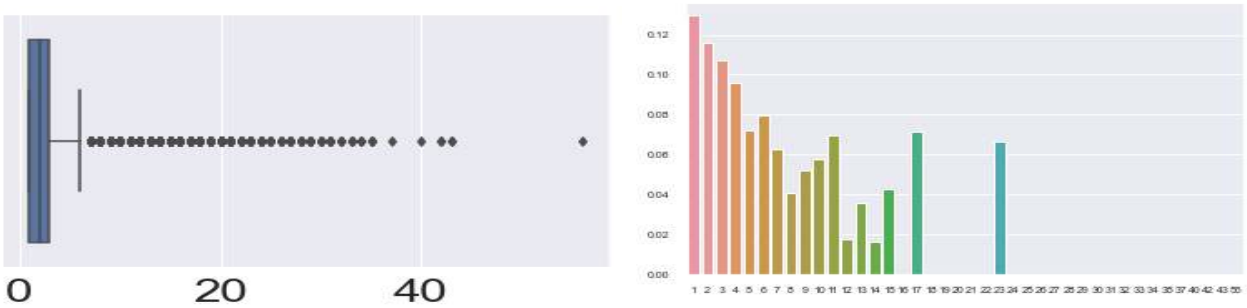
2.4 duration

이번 과제의 목표는 정기 예금 상품에 가입할 확률이 높은 고객을 찾는 것이다. 그러나 duration은 마케팅 통화가 이루어지기 전에는 알 수 없는 변수이므로 제거 후 예측 모형을 만들어야 한다. 예를 들어, durtaion이 0인 고객은 이번 마케팅 기간동안 연락을 취하지 않은 고객이므로 정기 예금 가입을 하지 않았을 것이다. 이처럼 UCI Machine Learning Repository의 데이터 명세서에도 마케팅 수행 전에는 알 수 없는 정보이므로 제거해야 한다고 명시하고 있다. 그러므로 duration 변수는 제거한다.

3. Other Attributes

3.1 campaign

<그림 9> campaign의 상자 그림과 campaign 횟수별 정기 예금 가입 비율



이번 마케팅 기간 동안 고객에게 연락한 횟수를 뜻하는 campaign은 대부분이 10보다 작은 값을 갖는 매우 편향된 분포를 갖는다. 또한, y=1인 비율 역시 campaign이 증가할수록 감소하는 추세가 보이는데 상식적으로 한 번의 마케팅에서 수차례 전화를 하는 것은 상품 가입률을 떨어뜨릴 것이기 때문이다. 따라서 boxplot의 수염(1.5*IQR)을 벗어나는 데이터(8 초과) 1239개를 이상값으로 여기고 제거한다.

3.2 pdays

pdays는 이전 마케팅에서 고객에게 마지막 연락 이후의 경과 일수지만, 대부분의 값이 999로 다수는 이전 마케팅 대상 고객이 아니다. 그러므로 999면 0 아니면 1의 값을 갖는 이항 변수 pdays_cat으로 변환한다. 변환 후의 교차분석표는 <표 11>에 나와있다. pdays_cat과 y의 피어슨 카이제곱 검정의 결과는 p값이 유의수준 0.05 보다 작으므로 'H_0: pdays_cat과 y는 독립이다.'를 기각한다. 즉, pdays_cat과 y는 연관성이 있다.

<표 11> pdays_cat과 y의 교차분석표

pdays_cat	y		
	no	yes	
0	31152	3241	34393
1	499	875	1374

$\chi^2 = 3814.33$, $df = 1$, p-value<0.001

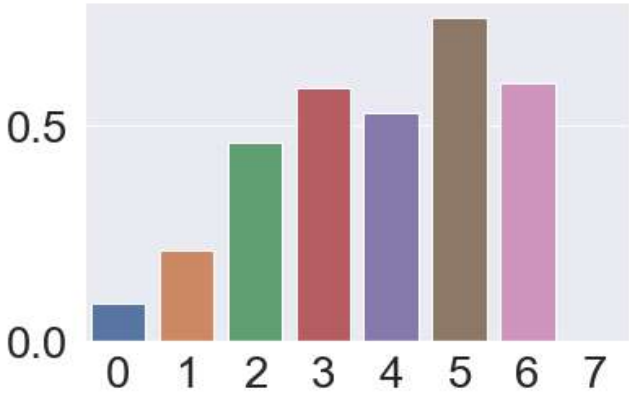
3.3 previous

<표 12> precious와 y의 교차분석표

previous	y		
	no	yes	
0	29165	2824	31989
1	3242	871	4113
2	367	316	683
3	81	115	196
4	31	35	66
5	4	12	16
6	2	3	5
7	1	0	1

$\chi^2 = 1987.5$, $df = 7$, p-value<0.001

<그림 10> previous 별 정기 예금 가입 비율



previous는 순서형 변수로 이전 마케팅에서 고객에게 연락한 횟수를 뜻하며 이전에 한 번이라도 연락을 취했던 고객이라면 정기 예금 가입률이 높아지는 경향이 보인다. 하지만, 피어슨의 카이제곱 검정 통계량은 5보다 작은 cell이 25% 이상이므로 믿을 만하지 못하다. 또한, 3번 이상의 연락을 한 고객의 수가 급격하게 감소하므로 previous가 2 이상이면 2, 그리고 1과 0을 갖는 변수 previous_cat을 생성하여 추후 분석에 사용하기로 결정했다.

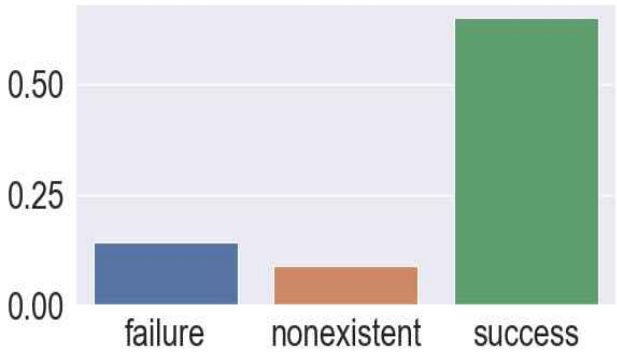
3.4 poutcome

<표 13> poutcome과 y의 교차분석표

poutcome	y		
	no	yes	
failure	3291	543	3834
nonexistent	29165	2824	31989
success	437	809	1246

$\chi^2 = 3690.9$, $df = 2$, p-value<0.001

<그림 11> poutcome 별 정기 예금 가입 비율



이전 마케팅의 결과를 나타내는 변수로, success의 값에서 y=1의 비율이 매우 높다. 즉, 이전 마케팅이 성공적으로 끝났다면, 이번 마케팅도 성공했을 가능성이 높은 모습을 보인다. 뿐만 아니라, 이전에 실패했던 고객이라도 이전에 마케팅을 하지 않은 고객에 비해 정기 예금 상품에 가입하는 경향성을 보인다. 그러므로 이전 마케팅을 실패한 고객에게 다시 연락하는 것이 중요해 보인다. 또한, poutcome과 y의 피어슨 카이제곱 검정의 결과는 p 값이 유의수준 0.05 보다 작으므로 'H_0: poutcome과 y는 독립이다.'를 기각한다. 즉, poutcome과 y는 연관성이 있다.

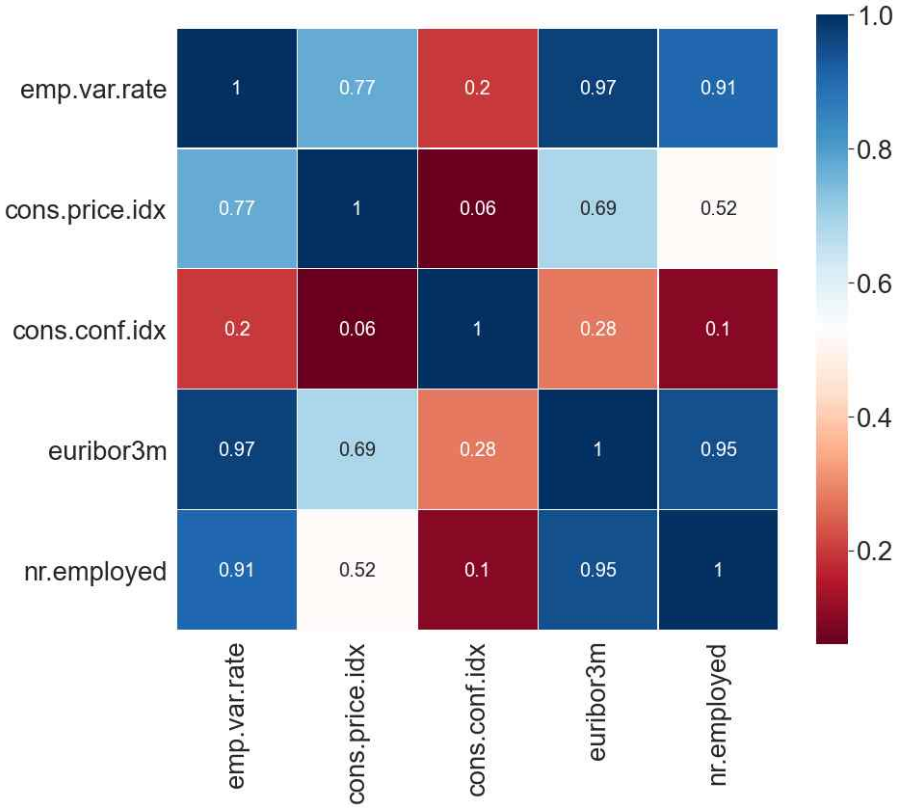
4. Social and Economic Context Attributes

이 범주에 속하는 총 5개의 변수들은 모두 연속형 변수들이며 사회 및 경제 지표를 나타내는 변수들이다. 이러한 변수들의 피어슨 상관계수는 <그림 12>와 같다.

현실적으로 은행이 마케팅을 진행할 때 사회 및 경제 지표를 나타내는 변수들을 통제하는 것은 불가능한 일이다. 그러므로 우선, 상관계수가 각각 0.97, 0.91, 0.95로 매우 높은 [emp.var.rate, euribor3m], [emp.var.rate, nr.employed], [euribor3m, nr.employed]의 처리를 고려하자. 즉, 다중공선성의 문제를 예방하기 위해 이 세 가지 변수 중 emp.var.rate와 nr.employed를 제거한다. 그 후 모형을 적합할 때마다 모형에 맞추어 사회 및 경제 지표를 나타내는 변수의 처리에 대해 논하도록 한다.

<표 14>에는 여태까지 논의한 독립 변수들에 대한 전처리 결과가 정리되어 있다. 'age', 'pdays', 'previous' 변수는 각각 'age_cat', 'pdays_cat', 'previous_cat'으로 범주형 변수로 변환하였다. 또한, 데이터 명세서와 카이제곱 검정의 결과를 참고하여 'default', 'housing', 'loan' 및 'duration' 변수를 제거하였다. 또한 피어슨의 상관계수를 활용해 상관계수가 매우 높게 나온 [emp.var.rate, euribor3m], [emp.var.rate, nr.employed], [euribor3m, nr.employed] 3가지 변수 중 emp.var.rate와 nr.employed를 제거했다. 마지막으로 'campaign' 변수는 상자그림을 활용해 이상치를 제거하였다.

<그림 12> 사회 및 경제 지표 변수들의 피어슨 상관계수



<표 14> 변수별 전처리 결과

범주	특성 변수	전처리
Bank client data	Age	age_cat(0: x≤30, 1: 30<x<58, 2: 60≤x)
	Job	
	Marital	
	Education	
	Default	제거
	Housing	제거
	Loan	제거
Related with the last contact of the current campaign	Contact	
	Month	
	Day_of_week	
	Duration	제거
Other attributes	Campaign	8 이상인 이상치 제거
	Pdays	pdays_cat(0: 999, 1: o/w)
	Previous	previous_cat(0: x=0, 1: x=1, 2: 2≤x)
	Poutcome	
Social and economic context attributes	Emp.var.rate	emp.var.rate와 nr.employed를 제거
	Cons.price.idx	
	Cons.conf.idx	
	Euribor3m	
	Nr.employed	

Ⅳ. 모형의 적합

앞서 보았듯이 훈련 데이터의 y의 값은 <그림 1>과 같이 'no'가 88.7%, 'yes'가 11.3%로 분포가 매우 불균형하다. 이러한 경우 모형 평가 시 accuracy를 사용한다면, 모형 간 올바른 비교가 이루어지기 힘들다. 단적인 예로 훈련 데이터의 모든 샘플에 대해서 타겟을 0('no')으로 예측한다면 Accuracy는 88.7%를 기록할 것이다. 그러므로 모형 평가의 기준이 Accuracy만으로는 부족하다.

이번 과제의 목표는 고객이 정기 예금에 가입할지 여부를 예측하는 것이다. 이러한 목표 하에서 다음과 같은 2가지의 오류가 발생할 수 있다.

1. 고객이 정기 예금에 가입하지 않았는데 가입했다고 예측하는 False Positive
2. 고객이 정기 예금에 가입했는데 가입하지 않았다고 예측하는 False Negative

주어진 상황에서 더 줄여야 하는 오류를 골라야 한다면, False Negative라고 할 수 있다. y가 'no'인 경우보다 y가 'yes'인 것을 잘 분류해내는 것이 중요하기 때문이다. 즉, 실제 가입한 고객들을 제대로 분류하는 것이 더 중요하다. 따라서, 모형 평가 시 이를 더 잘 반영할 수 있는 Recall과 Accuracy를 함께 비교하기로 한다. 여기서 $(\text{Recall}) = \text{TP}/(\text{TP}+\text{FN})$이다. 즉 FN가 작다는 말은 Recall이 크다는 것과 동일하다.

1. 로짓 분석

로짓 분석은 반응 값이 범주형일 때 사용하며 성공 확률에 유의한 영향을 미치는 설명변수를 식별하는데 유용하다. 따라서, 정기 예금 상품 가입을 뜻하는 y=1을 성공 확률로 가정하고 로짓 모형을 적합한다. 그러나, 모든 독립 변수를 사용한 로짓 모형은 유의하지 않은 변수가 매우 많으며 previous_cat의 회귀계수는 NA가 나온다. 그러므로 유의하지 않은 변수들을 제거하고 다시 모형을 적합한 결과는 다음과 같다.

<그림 13> 로짓 모형 적합 결과

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -64.382542   3.709220 -17.357  < 2e-16 ***
contacttelephone -0.823486   0.054935 -14.990  < 2e-16 ***
campaign         -0.056188   0.013512  -4.158 3.20e-05 ***
poutcomenonexistent  0.603673   0.059436  10.157  < 2e-16 ***
poutcomesuccess   0.792672   0.192497   4.118 3.82e-05 ***
cons.price.idx    0.711100   0.040446  17.582  < 2e-16 ***
cons.conf.idx     0.062536   0.003579  17.471  < 2e-16 ***
euribor3m        -0.566579   0.014229 -39.819  < 2e-16 ***
age_cat1         -0.246542   0.045134  -5.462 4.70e-08 ***
age_cat2          0.266041   0.078197   3.402 0.000668 ***
pdays_cat1        1.136466   0.189985   5.982 2.21e-09 ***
```

- 추정된 로짓 모형은 다음과 같다.

$$\widehat{\text{logit}(y=1)} = -64.38 - 0.82Contact_{tel} - 0.06Campaign + 0.60Poutcome_{non} + 0.79Poutcome_{suc} + 0.71cons.price + 0.06cons.conf - 0.57euribor3m - 0.25age_{cat_1} + 0.27age_{cat_2} + 1.14pdays_{cat1}$$

가장 먼저, 로짓 모형이 적합한지 적합도 검정을 수행한다. 귀무가설과 대립가설은 다음과 같다.

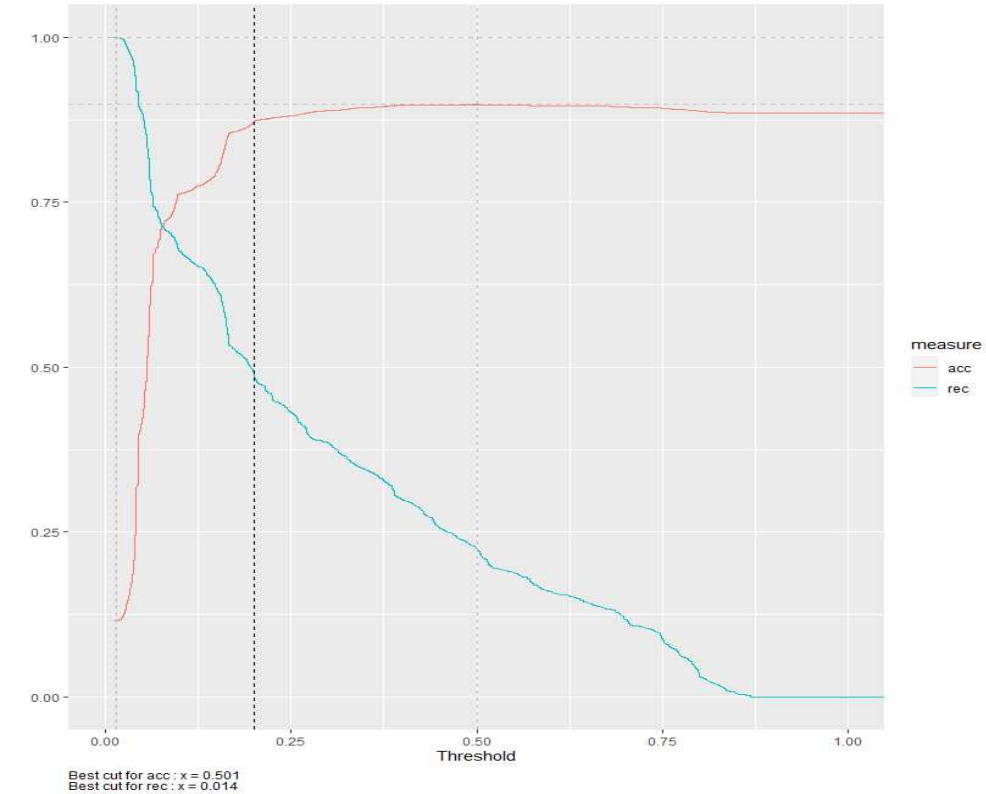
H_0 : 로짓 모형이 적합하다.
H_1 : 로짓모형이 적합하지 않다.

이때, 검정통계량은 값은 20446이며, 귀무가설 하에서 자유도가 35481인 카이제곱분포를 따른다. 따라서, 유의수준 5% 하에서 기각역에 해당하는 35920.3보다 작으므로 귀무가설을 기각하지 못한다. 즉, 로짓 모형이 적합하다고 판단할 수 있다.

모형의 해석은 다음과 같다. contact의 경우 다른 독립변수들을 고정시킨 상태에서, telephone을 통해 마케팅을 진행할 경우 고객이 정기 예금 상품에 가입할 확률의 추정 오즈는 cellphone을 통해 마케팅을 진행할 때의 $\exp(-0.82)$ 즉, 0.44배 낮다. 또한, 마케팅 기간에 연락(campaign)이 1회 증가할수록 고객이 정기 예금에 가입할 확률의 추정 오즈는 $\exp(-0.06)$배 즉, 6%씩 낮아진다. 이처럼 로짓 모형은 앞으로 적합할 SVM과 부스팅, 그리고 랜덤 포레스트 모형에 비해 해석력이 좋다는 장점이 있다. 다른 변수들도 마찬가지로 오즈비로 해석이 가능하다.

훈련 데이터와 시험 데이터에 대한 로짓 모형으로 예측값과 y의 confusion matrix는 <표 15>와 <표 16>과 같다. 이때 <그림 13>의 cut-off 별 accuracy와 recall의 변화를 참고하였다. accuracy를 가장 높게 하는 임계값은 0.501이며, recall을 가장 높게 하는 임계값은 0.014였다. 그러므로, 적절한 균형을 만드는 cutoff=0.2로 설정하였다(점선).

<그림 13> 로짓 모형: cut-off 별 accuracy와 recall의 변화



<표 15> 로짓 모형의 예측 결과: train

		True		
		0	1	
Pred	0	28960	2091	31051
	1	2447	1994	4441
		31407	4085	35492

accuracy : 0.872, recall : 0.488

<표 16> 로짓 모형의 예측 결과: test

		True		
		0	1	
Pred	0	3221	223	3444
	1	263	232	495
		3484	455	3939

accuracy : 0.877, recall : 0.510

2. 비선형 SVM

기계학습 분야에서 지도 학습의 한 종류인 SVM은 서포트 벡터의 역할을 하는 소수의 관측 개체들로 분류 및 회귀에 모두 사용이 가능하다. 특히 이번 과제와 같이 설명변수가 많은 경우에도 잘 작동한다는 장점이 있다. 그러므로 앞서 로짓 분석에서는 몇 개의 독립 변수만을 선택해서 모형을 적합했지만, SVM을 적합할 때에는 모든 독립 변수를 사용해서 y를 예측하고자 한다.

훈련 데이터에 대하여 10-fold cross-validation을 수행한 결과 가우스 커널과 C=100이 최적의 조합으로 선정되었다. 이때 서포트 벡터의 개수는 총 8247개이며 훈련 데이터와 시험 데이터에 각각에 대한 SVM의 예측과 실제 y 간의 confusion matrix는 <표 17>과 <표 18>과 같다.

<표 17> SVM의 예측 결과: train

		True		
		0	1	
Pred	0	30920	3217	34137
	1	487	868	1355
		31407	4085	35492

accuracy : 0.896, recall : 0.641

<표 18> SVM의 예측 결과: test

		True		
		0	1	
Pred	0	3440	366	3806
	1	44	89	133
		3484	455	3939

accuracy : 0.896, recall : 0.669

SVM은 로짓 모형에 비해 종속 변수 y와 독립 변수들 간 연관성에 대한 해석력은 떨어지지만, accuracy와 recall이 모두 높은 것을 확인할 수 있다.

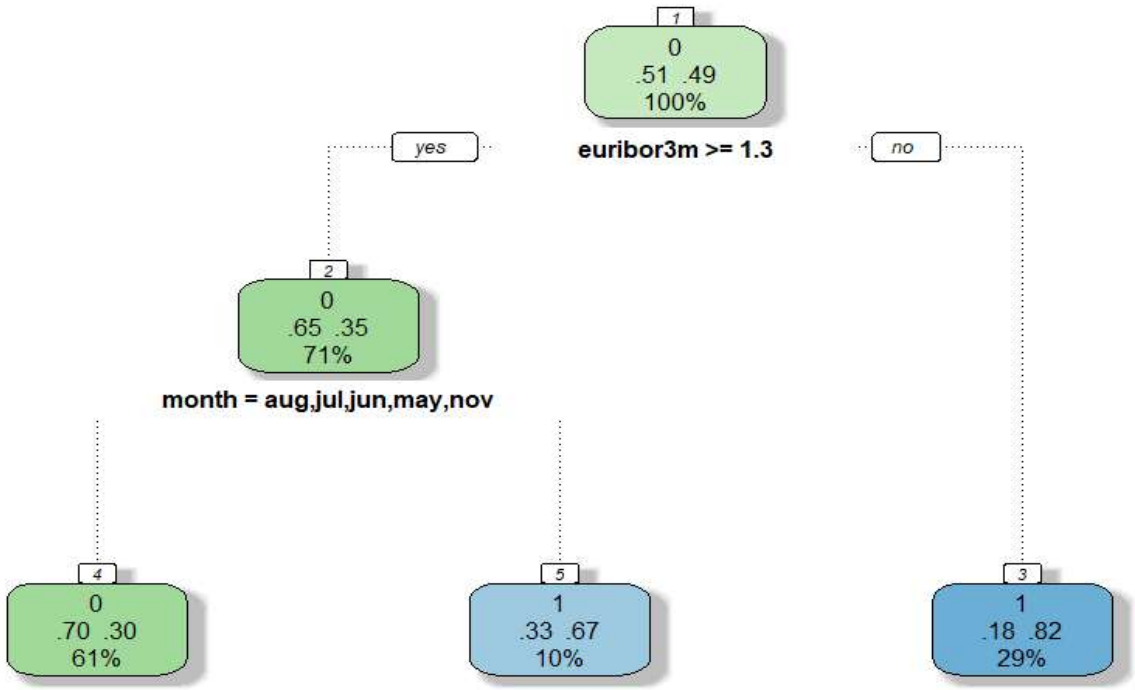
3. CART

CART는 분류에 회귀에 모두 사용 가능한 나무 알고리즘이다. 데이터의 분포에 특별한 가정이 필요 없으며 시각화가 용이하고, 해석력 또한 우수하다. 주의할 점은 훈련 데이터의 종속 변수인 y의 불균형이 모형의 생성과 예측에 모두 부정적인 영향을 준다는 점이다. 이번 과제에서 사용하는 훈련 데이터의 종속 변수 y의 비율은 <표 19>와 같이 y=0이 88%, y=1이 12%로 매우 불균형하다. 이런 경우 종속 변수 범주에 따른 가중치 없이 나무 모형을 단순히 적합하고 예측한다면 모형의 y=1에 대한 예측 수가 실제 y=1에 비해 적을 것이다. 따라서, y=0('no')에는 가중치 0.12를 주고, y=1('yes')에는 가중치 0.88을 준 뒤, 엔트로피를 분리 기준으로 하여 나무 모형을 적합했다. 그 결과는 <그림 14>에 나와있다.

<표 19> 훈련 데이터의 종속 변수 y의 비율

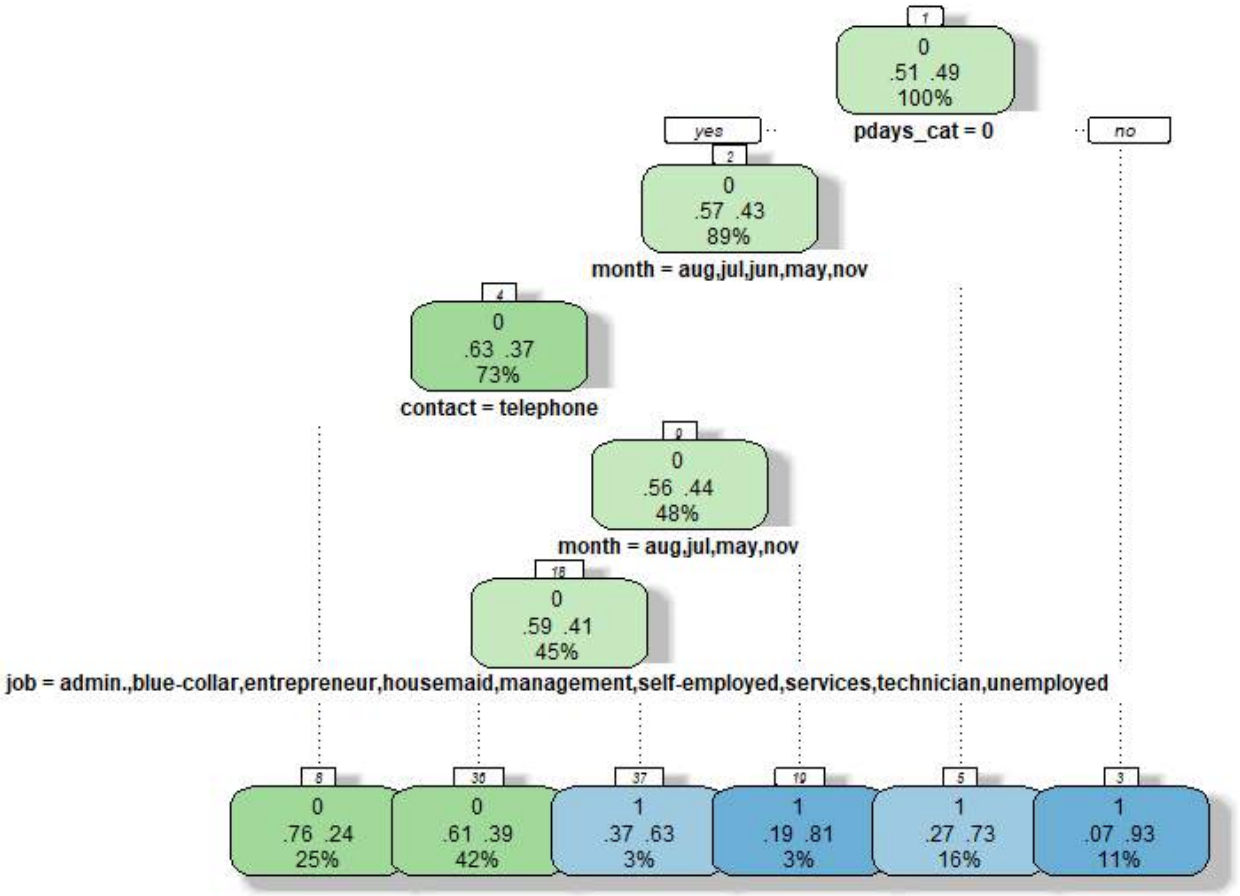
	y		
	0	**1**	
개수	31407	4085	35492
(비율)	(88%)	(12%)	(100%)

<그림 14> 훈련 데이터에 대한 나무 모형 1 (사회 및 경제 지표 변수 제거 전)



엔트로피를 분리 기준으로 사용한 나무 모형(<그림 14>)은 사회 및 경제 지표를 나타내는 변수들(Social and Economic Context Attributes) 중 하나인 'euribor3m'으로 첫 노드의 분할을 시작한다는 특징이 있다. 그러나 은행이 마케팅을 진행할 때 사회 및 경제 지표를 나타내는 변수들을 통제하는 것은 불가능한 일이다. 그러므로, 'euribor3m'를 비롯한 사회 및 경제 지표 변수를 제거하고 나무 모형을 적합해 보기로 하였다. 이 나무의 결과는 <그림 15>에 나와있다. 이 결과를 통해 나무 모형의 높은 분산(variance)을 확인할 수 있다. 또한, CART의 훈련 데이터와 시험 데이터에 각각에 대한 예측과 실제 y 간의 confusion matrix는 <표 20>과 <표 21>에 정리했다.

<그림 15> 훈련 데이터에 대한 나무 모형 2 (사회 및 경제 지표 변수 제거 후)



나무 모형은 이전 마케팅에 연락한 적이 있던 고객(pdays_cat=1)이라면 정기 예금에 가입할 고객으로 분류한다. 또한, 이전 마케팅에서 연락한 적이 없고(pdays_cat=0) 고객에게 연락하는 달이 3,4,9,12월이라면 이 고객 역시 정기 예금에 가입할 고객으로 분류한다. 반면, 이전 마케팅에서 연락한 적이 없는 고객에게 이번 마케팅에서 연락하는 달이 5,6,7,8,11월이며, telephone을 통해 고객에게 연락했다면 이 고객은 정기 예금에 가입하지 않을 고객으로 분류한다. 이러한 이진 분할 방식으로 나무 모형이 훈련 데이터와 시험 데이터에 대해 예측한 결과는 <표 20>과 <표 21>에 나와있다.

<표 20> CART의 예측 결과: train

		True		
		0	**1**	
Pred	**0**	27191	1875	29066
	1	4216	2210	6426
		31407	4085	35492

accuracy : 0.828, recall : 0.541

<표 21> CART의 예측 결과: test

		True		
		0	**1**	
Pred	**0**	2997	206	3203
	1	487	249	376
		3484	455	3939

accuracy : 0.824, recall : 0.547

4. Bagging

나무 모형은 bias가 작지만, variance가 크므로 안정성이 떨어진다는 단점이 있다. 즉, 앞서 적합한 모형에서 보았듯이 사회 및 경제 지표 변수와 같이 큰 영향을 미치는 몇 개의 변수에 의해 노드의 분리 값이 쉽게 바뀔 수 있다. 따라서 훈련 데이터의 샘플이 달라지면 모형도 쉽게 변한다. 이러한 문제점을 variance를 낮춰 해결하는 배깅 모형을 사용하고자 한다. 배깅은 훈련 데이터에서 중복을 허용하여 같은 크기의 붓스트랩 샘플을 추출한 후 각 붓스트랩 샘플에 대해 독립적으로 나무 모형을 적합한다. 이를 m번 반복해 통합 모형을 만드는 방식으로 분산을 낮출 수 있다. 이번 과제는 분류 문제이므로 투표를 통해 통합 모형을 만든다.

훈련 자료에서는 목표 변수 y의 2개 범주 간 빈도의 불균형이 매우 심하다. <표 19>에서 보았듯 y의 값이 0이 88%, 1이 12%이므로 0과 1의 비율이 약 7:1이다. 이러한 불균형 문제를 해결하기 위해 y가 1인 개체를 7배로 만든 후에 배깅 모형을 만들었다.

<그림 16> 불균형 문제 해결을 위한 훈련 데이터의 목표 변수 균형화

```
train.0 <- train[train$y == 0,]
train.1 <- train[train$y == 1,]
train.balanced <- rbind(train.1,train.1,train.1,train.1
,train.1,train.1,train.1,train.0)
```

따라서 train.balanced에는 y=0이 31407개, y=1이 28595개가 있으며 약 1.1 : 1의 비율을 갖는다. 이제 이렇게 생성한 train.balanced 데이터에 M=100 즉, 반복 횟수가 100인 배깅 모형을 적합했다. 이때 Out-of-bag error는 0.0825였으며, 훈련 데이터와 시험 데이터 각각에 대한 분류 결과는 다음과 같다.

<표 22> 배깅 모형의 예측 결과: train

		True		
		0	**1**	
Pred	**0**	29244	100	29344
	1	2163	3985	6148
		31407	4085	35492

accuracy : 0.936, recall : 0.976

<표 23> 배깅 모형의 예측 결과: test

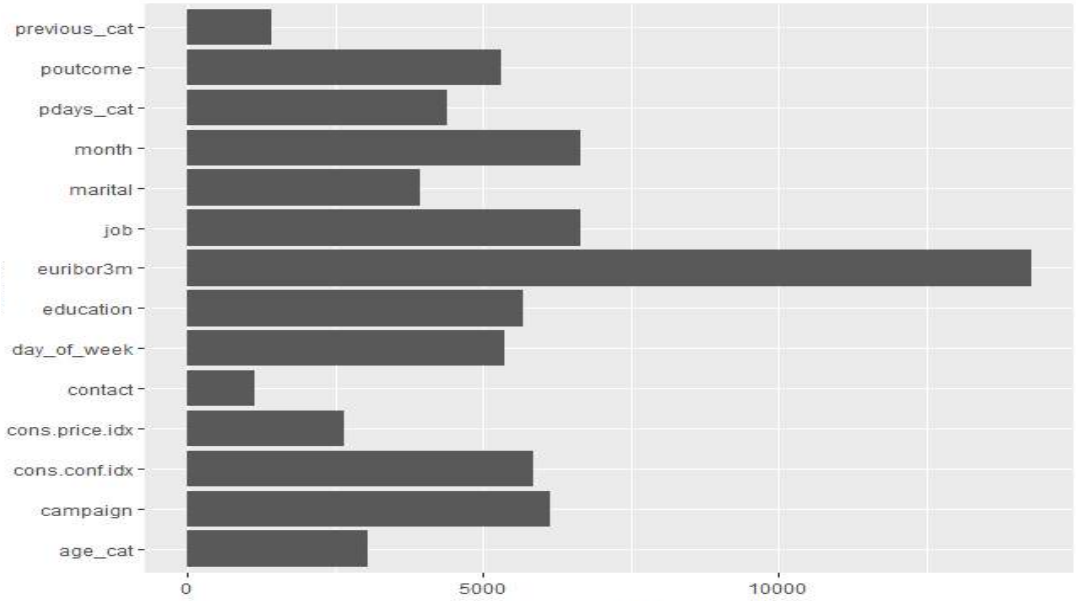
		True		
		0	**1**	
Pred	**0**	3024	246	3270
	1	460	209	669
		3484	455	3939

accuracy : 0.821, recall : 0.459

분류 결과를 보면 훈련 데이터에 비해 시험 데이터에 대해서 accuracy와 recall이 모두 현저히 낮은 경향을 보인다. 즉, 훈련 데이터에 과대 적합되어 일반화 가능성이 떨어지는 모습이 보였다.

배깅은 몇 가지 한계점을 가지고 있다. 우선, 각 붓스트랩 샘플에 대하여 나무 모형을 적합할 시 변수 전체를 사용하기 때문에 특정 변수가 분류에 있어 중요한 역할을 하는 경우 각 나무 모형 간 독립성이 떨어지는 문제가 있다. 뿐만 아니라, 변수 간 상관관계가 높은 경우 특정 변수만 선택되는 문제점도 존재한다. 이러한 관점에서 <그림 17>의 변수 중요도를 토대로 배깅 모형의 성능이 좋지 않은 이유를 추론해볼 수 있다.

<그림 17> 배깅 모형: 변수 중요도



'euribor3m'이 압도적으로 높은 변수 중요도를 갖는 것을 보아, 100개의 붓스트랩 샘플 각각에 생성한 나무 모형은 가장 먼저 'euribor3m'을 분할 변수로 선택했을 것이다. 그러므로 나무 모형 간 연관성이 높을 것이며 비슷한 나무 모형들끼리의 분류 결과를 투표하기 때문에 단일 나무 모형과 큰 차이가 없는 것처럼 보인다.

또한, 앞서 적합한 CART 나무 모형처럼 사회 및 경제 관련 변수를 제거한 뒤 배깅 모형을 적합한 결과도 <표 22>, <표 23>과 크게 다르지 않았다.

5. 랜덤 포레스트

앞서 사용한 배깅 모형은 몇 가지 문제점을 가지고 있다. 우선 훈련 데이터에 과대 적합되었으며 좋은 성능을 발휘하지 못했다. 이는 상관관계가 있는 나무들 간의 통합 모형이라는 배깅의 한계점에서 비롯되었다. 따라서 각 노드에서 변수를 비복원 임의 추출하는 랜덤 포레스트를 사용해 배깅의 tree correlation이라는 한계점을 보완해보고자 한다.

배깅 모형을 적합할 때와 마찬가지로 목표 변수 y의 범주 간 불균형이 해소된 train.balanced 데이터를 이용해 랜덤 포레스트 모형을 만들었다. 각 나무 모형을 적합 시, 분류 문제이기 때문에 각 노드에서 $q = \sqrt{14} \approx 4$ 개의 변수를 비복원 임의 추출하였다. 또한, ntree=500을 설정해 500개의 붓스트랩 샘플에서 생성한 나무 모형들의 통합 모형을 생성하였고 이때 랜덤 포레스트 모형의 OOB 오류율은 12.48%이다.

```
Call:
 randomForest(formula = y ~ ., data = train.balanced, mtry = 4,      importance
 = T)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 12.48%
```

<표 24> 랜덤포레스트 모형 예측 결과: train

		True		
		0	1	
Pred	0	29905	567	30472
	1	1502	3518	5020
		31407	4085	35492

accuracy : 0.942, recall : 0.861

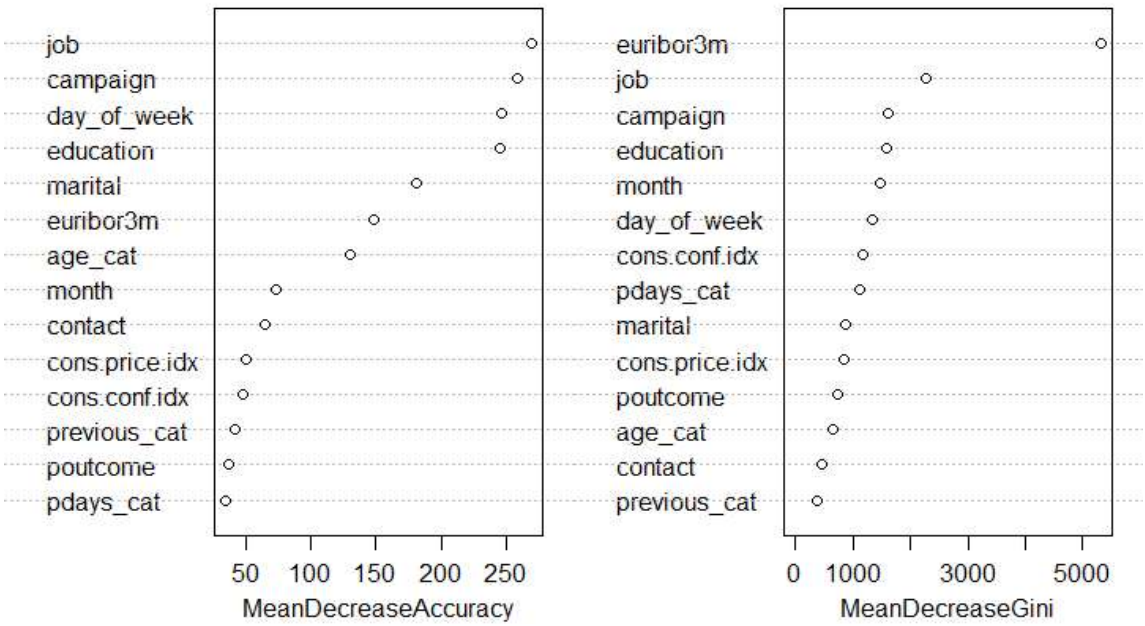
<표 25> 랜덤포레스트 모형 예측 결과: test

		True		
		0	1	
Pred	0	3154	211	3365
	1	330	244	574
		3484	455	3939

accuracy : 0.863, recall : 0.536

<표 24>와 <표 25>는 랜덤 포레스트 모형의 분류 결과를 나타내고 있다. 배깅 모형에 비해 과대 적합의 양상이 줄어들었다. 또한, <그림 18>은 랜덤 포레스트 모형을 활용한 예측 변수의 중요도를 보여준다. accuracy 감소량을 기준으로 했을 때, 변수 중요도의 순서는 job > campaign > day_of_week > education > marital > euribor3m > age_cat > month > contact > cons.price.idx > cons.conf.idx > previous_cat > poutcome > pdays_cat이고, 지니 지수의 감소량을 기준을 따르면 변수 중요도의 순서는 euribor3m > job > campaign > education > month > day_of_week > cons.conf.idx > pdays_cat > marital > cons.price.idx > poutcome > age_cat > contact > previous_cat이다. 따라서, 고객이 정기 예금 상품에 가입할지를 예측하는 문제에서 고객의 직업(job), 연락 횟수(campaign), 교육 수준(education) 및 경제 상황(euribor3m) 등이 중요하게 작용한다는 것을 알 수 있다.

<그림 18> 랜덤 포레스트 모형: 변수 중요도



6. XGBoost

부스팅 방법은 accuracy가 0.5를 상회하는 weak leaner라 불리는 모형들을 연결하는 앙상블 방법이다. 즉, 앞서 적합한 모형을 보완해가는 방식이다. 일반적으로 널리 쓰이는 그래디언트 부스팅 모형은 전체 훈련 데이터에 대해 나무 모형을 적합한 뒤, 이전까지의 오차를 보정하도록 나무 모형을 순차적으로 추가해 나간다. 즉, 매 반복마다 이전 나무 모형으로부터의 오차에 새로운 나무 모형을 학습하고 학습된 나무 모형들을 더해나가는 방식으로 작동한다.

그러나, 이러한 그래디언트 부스팅 모형은 오차를 계속 줄여나가므로 훈련 데이터에 과적합 되는 경향과 학습시간이 매우 느리다는 단점을 갖고 있다. XGBoost는 이러한 그래디언트 부스팅에 수식적으로는 규제항을 추가하고, 알고리즘적으로는 병렬 연산을 가능하게 하여 연산 속도를 높인 모형이다. 여기서 규제항은 weak learner로 사용하는 CART 모형의 복잡도(터미널 노드)가 증가할수록 손실함수에 패널티를 주는 방식으로 작동한다.

본 과제에서는 훈련 데이터에 이러한 XGBoost를 적합하고 결과를 평가 및 다른 모형들과 비교해보기로 한다. 배깅 모형 및 랜덤 포레스트 모형를 적합할 때와 마찬가지로 목표 변수 y의 범주 간 불균형이 해소된 train.balanced 데이터를 사용했다. XGBoost에는 많은 초모수(hyper-parameter)들이 있는데, 다음과 같은 모형을 10-fold cross-validation을 통해 최적의 모형으로 선택하였다.

<그림 19> 훈련 데이터에 적합한 XGBoost의 초모수

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0, gpu_id=-1,
              importance_type='gain', interaction_constraints='',
              learning_rate=0.1, max_delta_step=0, max_depth=9,
              min_child_weight=1, missing=nan, monotone_constraints='()',
              n_estimators=100, n_jobs=0, num_parallel_tree=1, random_state=0,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
              tree_method='exact', validate_parameters=1, verbosity=None)
```

<표 26> XGBoost 모형 예측 결과: train

		True		
		0	1	
Pred	0	28806	1132	29937
	1	2602	2953	5555
		31407	4085	35492

accuracy : 0.895, recall : 0.531

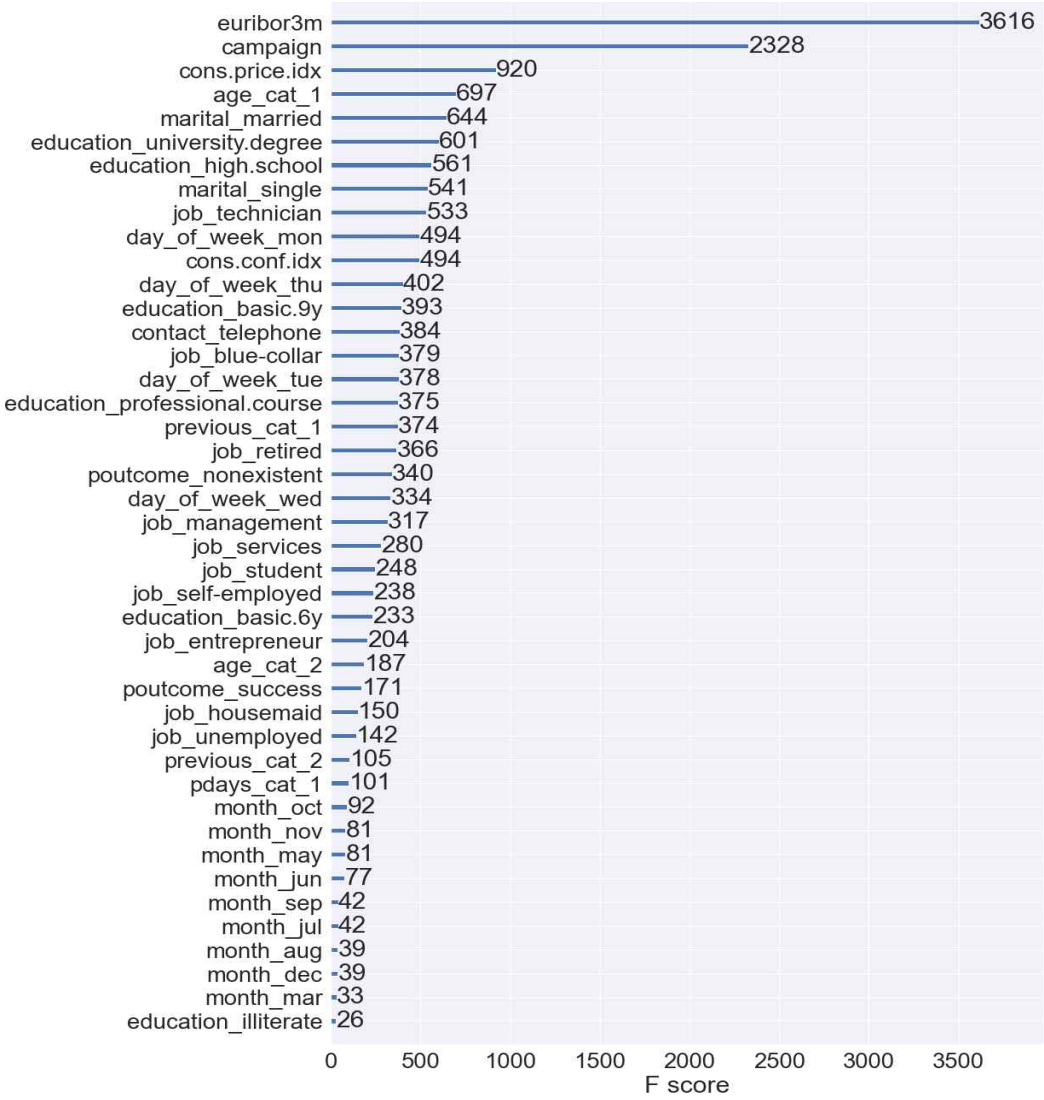
<표 27> XGBoost 모형 예측 결과: test

		True		
		0	1	
Pred	0	3133	185	3318
	1	351	270	621
		3484	455	3939

accuracy : 0.864, recall : 0.435

또한, XGBoost도 나무 기반 앙상블 모형이기 때문에 배깅 모형, 랜덤 포레스트 모형과 같이 변수 중요도를 산출할 수 있다. <그림 20>을 보면, 고객이 정기 예금 상품에 가입할지를 예측하는 문제에서 고객의 경제 상황(euribor3m), 연락 횟수(campaign), 교육 수준(education), 나이(age_cat) 및 등이 중요하게 작용한다는 것을 알 수 있다.

<그림 20> XGBoost 모형: 변수 중요도



- 앞서 적합한 6개 모형들의 최종 성능은 다음과 같다.

<표 28> 훈련 데이터와 시험 데이터에 대한 각 모형의 accuray와 recall

모형	훈련 데이터		시험 데이터	
	Accuracy	Recall	Accuracy	Recall
로짓 모형	0.872	0.488	0.877	0.510
비선형 SVM (radial)	0.896	0.641	0.896	0.669
CART	0.828	0.541	0.824	0.547
Bagging	0.936	0.976	0.821	0.459
랜덤 포레스트	0.942	0.861	0.863	0.536
XGBoost	0.895	0.531	0.864	0.435

Ⅴ. 결론

<표 28>은 앞서 적합한 로짓 모형, 비선형 SVM, CART, Bagging, 랜덤 포레스트 및 XGBoost의 시험 데이터에 대한 분류 결과를 보여주고 있다. 과제의 목표는 고객이 정기 예금에 가입할지 여부를 예측하는 것이다. 즉, 모형의 전체적인 정확도를 나타내는 accuracy도 충분히 높아야 하며, 동시에 y가 'yes'(1)인 고객을 잘 분류해내는 것이 중요하다. 또는, 만약 예측 결과를 통해 독립 변수들과 종속 변수들의 관계를 해석하고 이를 이용해 마케팅의 성과를 향상시키는 것이 목적이라면 해석력과 시각화가 중요할 수 있다. 이러한 관점에서 다음과 같은 2가지 기준을 적용할 수 있을 것이다.

<모형 비교의 기준>

1. 성능 기준
   1.1 accuracy는 모든 관측치를 y=0으로 예측했을 때의 값인 0.88의 값을 상회해야 한다.
   1.2 1.1을 만족하며 이와 함께 recall이 높아야 한다.
2. 해석력 및 시각화 기준
   2.1 종속 변수 y와 독립 변수들 간의 관계를 설명할 수 있어야 한다.
   2.2 시각화가 가능해야 한다.

첫 번째 기준인 성능 기준을 적용하면, 1.1에서 CART 모형과 Bagging을 가장 먼저 제외해야 한다. 또한, 로짓 모형, 비선형 SVM, 랜덤 포레스트, XGBoost 중에서 recall을 기준으로 가장 좋은 성능을 보이는 모형은 비선형 SVM 모형이다. 그러므로 은행의 입장에서 정기 예금 상품에 가입할 가능성이 높은 고객을 중심으로 마케팅을 진행하기 위해 가우스 radial 커널을 이용한 비선형 SVM 분류 모형을 선택할 수 있다. 또는 위 모형들 중 몇 가지를 선택하여 투표 방법을 이용한 앙상블 방법을 이용할 수 있을 것이다. 이러한 방법의 장점은 각 모형들이 독립적일수록 분산을 낮추는 효과가 있다.

반면, 해석력 및 시각화가 더 중요한 기준이라면 로짓 모형 또는 CART 모형이 이러한 장점을 지니고 있으므로 이들을 선택할 수 있을 것이다. 로짓 모형은 특정 독립 변수의 변화에 따라 고객이 정기 예금 상품에 가입할 오즈비가 어떻게 변화하는지 설명할 수 있으며, CART 모형은 특정 변수의 특정 기준에 따라 고객이 정기 예금 상품에 가입하고 가입하지 않는지 이를 시각화할 수 있다는 장점이 있다.

- 한계점

데이터가 목표 변수 y의 범주에 따라 매우 불균형한 데이터이다 보니 모형 훈련 시 필연적으로 편향이 생긴다. 이를 트리 기반 모형들에서는 과대 표집(oversampling) 방법을 이용해 해결했으나, SVM 모형에서는 이를 제대로 해결하지 못했다. 추후 과대 표집 기법이며 과대적합을 방지하는 효과가 있는 SMOTE 및 ADASYN 등을 이용해 이러한 문제를 해결해야 할 필요성이 있다.