Lab_assignment 3

2017010146 산업경영공학과 이시현

0. 데이터 설명 및 데이터 프레임

1) 데이터 수집

Kaggle competition 중 Costa Rican Household Poverty Level Prediction의 Data를 다운로드 하였다.

2) 데이터 설명

2-1) 데이터 셋의 행, 열 개수

각 행은 하나의 개인을 나타내며 각 열은 개인에 고유한 특징이거나 개인의 가구에 대한 특징이다.

Set 종류	행 개수	열 개수
Training set	9557	143
Testing set	23856	142

2-2) 데이터 타입 별 개수 및 설명

데이터 타입	개수	추가 설명
실수형	8	
정수형	130	Boolean 변수, 순서형 변수
객체형	5	모델에 직접 넣을 수 없기 때문에 변환 필요

2-3) 중요 열 설명

열이 143개로 꽤 많으므로 중요한 몇 개의 열에 대해서만 설명하고, 다른 설명이 필요한 열에 대해서는 해당 변수를 사용할 때 설명을 추가하였다.

열 명	의미
Id	각 개인의 고유 식별자
Idhogar	각 가구의 고유 식별자
Parentesco1	이 사람이 가장인지 아닌지 표시
Target	한 가구의 모든 구성원에 대해 같아야 하는 레벨

2-3) Target 변수 설명

빈곤 레벨	의미
1	extreme poverty
2	moderate poverty
3	vulnerable households
4	non vulnerable households

2-4) 데이터 분석 목표

가구 별 빈곤 레벨을 예측하는 것이다.

3) 데이터프레임으로 만들기

Read in Data and Look at Summary Information

```
pd.options.display.max_columns = 150

# Read in data
train = pd.read_csv('train.csv')
test = pd.read_csv('test.csv')
train.head()
```

	Id	v2a1	hacdor	rooms	hacapo	v14a	refrig	v18q	v18q1	r4h1	r4h2	r4h3	r4m1	r4m2	r4m3	r4t1	r4t2	r4t3	tamhog	tamviv	escolar
0	ID_279628684	190000.0	0	3	0	1	1	0	NaN	0	1	1	0	0	0	0	1	1	1	1	10
1	ID_f29eb3ddd	135000.0	0	4	0	1	1	1	1.0	0	1	1	0	0	0	0	1	1	1	1	12
2	ID_68de51c94	NaN	0	8	0	1	1	0	NaN	0	0	0	0	1	1	0	1	1	1	1	11
3	ID_d671db89c	180000.0	0	5	0	1	1	1	1.0	0	2	2	1	1	2	1	3	4	4	4	9
4	ID_d56d6f5f5	180000.0	0	5	0	1	1	1	1.0	0	2	2	1	1	2	1	3	4	4	4	11

1. 전처리한 내용 정리 및 전처리 이유 제시

1) 가구별로 target 값 통일

데이터 분석의 목표는 개인별이 아닌 가구별 빈곤도를 예측하는 것이다. 이를 위해 같은 가구의 구성원들이 서로 다른 target 값을 갖는다면 가구의 가장의 target 값으로 target 통일시켰다.

2) 결측 값 처리

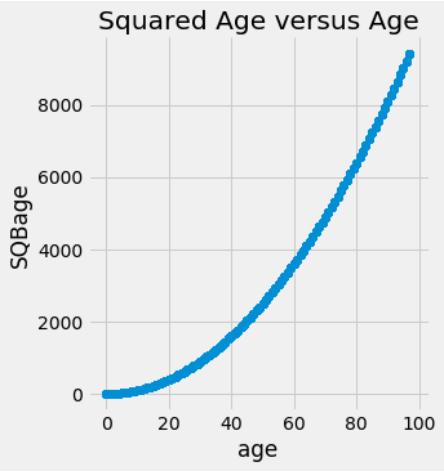
예측 모델을 만들기 위해서는 null값이 들어가면 안된다.

3) 변수 간 상관관계가 높은 변수 쌍 중 한 변수 제거

상관관계가 높으면 다중공선성 문제가 발생할 수 있다. 다중공선성은 독립변수들 간에 강한 상관관계가 나타나서 회귀변수의 전제 가정인 독립변수들 간에 상관관계가 높으면 안된다는 조건을 위배하는 경우를 의미한다. 다중공선성 문제를 해결하기 위해 상관관계 분석을 진행하였다.

3-1) squared variables 제거

제곱한 변수와 제곱하지 않은 변수 사이의 상관관계가 매우 높기 때문에 제곱한 변수 제거하였다. 하나의 예로, SQBage(제곱한 나이)와 Age(나이) 사이의 상관관계가 높아 SQBage를 제거하였다. 아래의 두 변수의 값 분포 그래프에서 볼 수 있듯이 두 변수는 매우 높은 상관관계를 보인다.



3-2) tamhog(size of the household), hogar_total(# of total individuals in the household), r4t3(Total persons in the household) 변수 제거

아래 히트 맵에서 볼 수 있듯이 tamhog와 상관계수가 0.9 이상인 변수는 4가지가 있다. 이 중 tamhog, hogar_total, r4t3를 제거하였다. Hhsize 변수는 tamviv과 항상 같진 않다. 집에 같이 살지

않은 가족이 있을 수 있고, 가족이 아니어도 같이 살 수 있기 때문이다. 대신, tamviv 값에서 Hhsize을 빼, hhsize-diff 라는 새로운 변수를 생성했다.

3-3) male 변수 제거

Female 과 상관관계가 커 male 변수를 제거해주었다.

3-4) area2(zona rural) 변수 제거

사는 집이 시골 지역에 있는지 없는지를 나타내는 변수인데, 사는 집이 도시 지역에 있는지를 나타내는 열인 area1가 있기 때문에 area2 열은 제거했다.

4) 순서형 변수 생성

의미가 같은 여러 개의 변수를 하나의 순서형 변수로 통합해 생성하였다.

4-1) elec 변수 생성

전기 수급 방식에 관한 변수 4개(public, planpri, noelec, coopele)를 하나의 순서형 변수로 생성하였다.

4-2) walls+roof+floor 변수 생성

집, 바닥, 지붕의 상태 3개로 나눠서 나타내는 9개의 변수를 이용해 하나의 순서형 변수로 생성하였다.

4-3) minus 변수 생성

화장실, 전기, 바닥, 수도, 천장의 유무를 나타내는 변수를 이용해 하나의 순서형 변수로 생성하였

다.

4-4) bonus 변수 생성

한 가정이 냉장고, 컴퓨터, 태블릿, TV를 소유하면 +1씩 얻게 되는 bonus 변수를 생성했다.

4-4) 1인당 특성 변수 생성

변수 명	설명
phones-per-capita	한 사람당 가지고 있는 폰의 개수
tablets-per-capita	한 사람당 가지고 있는 태블릿의 개수
rooms-per-capita	한 사람당 가지고 있는 방의 개수
rent-per-capita	한 사람당 내야하는 월세 비용

4-5) Inst 변수 생성

교육수준을 나타내는 instlevel1(교육을 받지 않음)~instlevel9(고등교육 이수) 변수를 0~8의 값으로 바꿔 Inst라는 순서형 변수를 생성하였다.

5) 그 외에 변수 생성

5-1) 교육 관련 변수 생성

교육과 관련된 변수들이 예측 모델에 도움을 줄 것 같아 교육관련 변수를 생성하였다.

escolari(교육년수)/age = escolari(교육년수)/age

inst(교육수준)/age = inst(교육수준)/age

5-2) Tech 변수 생성

Tech 변수는 v18q(소유한 태블릿수) 변수와 mobilephone(소유한 폰 수) 변수 의미 중복을 피하기 위해 생성하였다.

tech = v18q(소유한 태블릿수) + mobilephone(소유한 폰 수)

5-3) 개인에 고유한 특징을 나타내는 변수를 가구에 대한 특징을 나타내는 변수로 통합

개인에 고유한 특징을 나타내는 변수를 가구별로 묶어 변수 별 min, max, sum, count, std, range_ 값을 구해 변수로 추가해주었다.

	v18q						dis						female						estadocivil1				
	min	max	sum	count	std	range_	min	max	sum	count	std	range_	min	max	sum	count	std	range_	min	max	sum	count	
idhogar																							
000a08204	1	1	3	3	0.0	0	0	0	0	3	0.000000	0	0	1	1	3	0.577350	1	0	1	1	3	
000bce7c4	0	0	0	2	0.0	0	0	1	1	2	0.707107	1	0	1	1	2	0.707107	1	0	0	0	2	
001845fb0	0	0	0	4	0.0	0	0	0	0	4	0.000000	0	0	1	2	4	0.577350	1	0	0	0	4	
001ff74ca	1	1	2	2	0.0	0	0	0	0	2	0.000000	0	1	1	2	2	0.000000	0	0	1	1	2	
003123ec2	0	0	0	4	0.0	0	0	0	0	4	0.000000	0	0	1	1	4	0.500000	1	0	1	2	4	

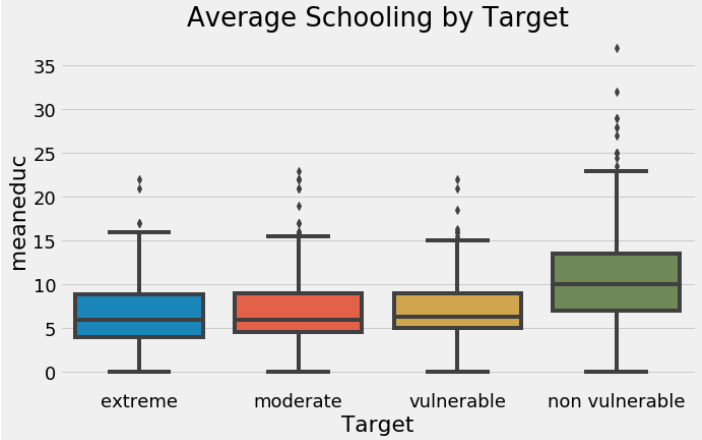
6) 개인에 고유한 특징을 나타내는 변수와 가구에 대한 특징을 나타내는 변수를 모두 통합한 열에서 111개의 열 제거

위와 같은 변수 생성 및 제거 과정을 거친 후, 마지막으로 변수 간 상관계수가 0.95 이상인 변수는 제거하였다.

2. 데이터 시각화 결과 및 이를 통해 유추해 볼 수 있는 점 제시

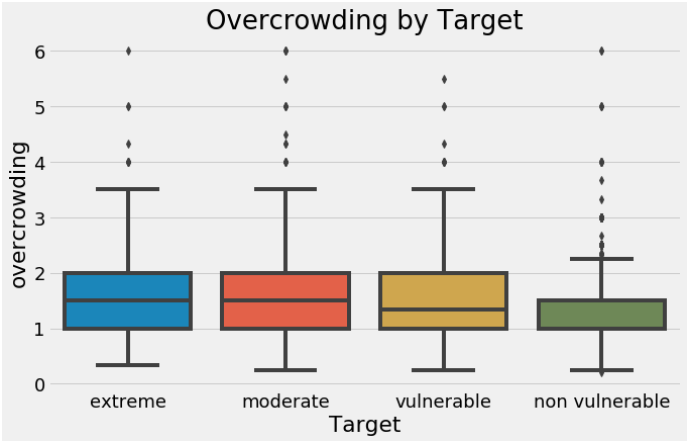
1) meanedu(average years of education for adults (18+)) 변수는 빈곤도를 예측하는데 유용할 것이다.

아래 박스 플랏을 통해 target 변수 label별로 meandedu의 분포가 유의미하게 다른 것을 알 수 있다. 이는 meanedu 변수가 빈곤도를 예측하는데 유용할 것임을 유추할 수 있다.



2) overcrowding(# persons per room) 변수는 빈곤도를 예측하는데 유용할 것이다.

label이 vulnerable, non vulnerable 일 때 중간 값이 나머지 label일 때의 값보다 유의미하게 낮기 때문에 overcrowding 변수가 빈곤도를 예측하는데 유용할 것임을 유추할 수 있다.



3) 새로 생성한 walls+roof+floor 변수는 빈곤도를 예측하는데 유용할 것이다.

아래 박스 플랏에서 확인할 수 있듯이 각 label별 walls+roof+floor 값의 분포가 현저히 차이나는

것을 확인할 수 있다. 그러므로 walls+roof+floor 변수는