코스타리카 보조금 예측

목차

프로젝트 선정 배경

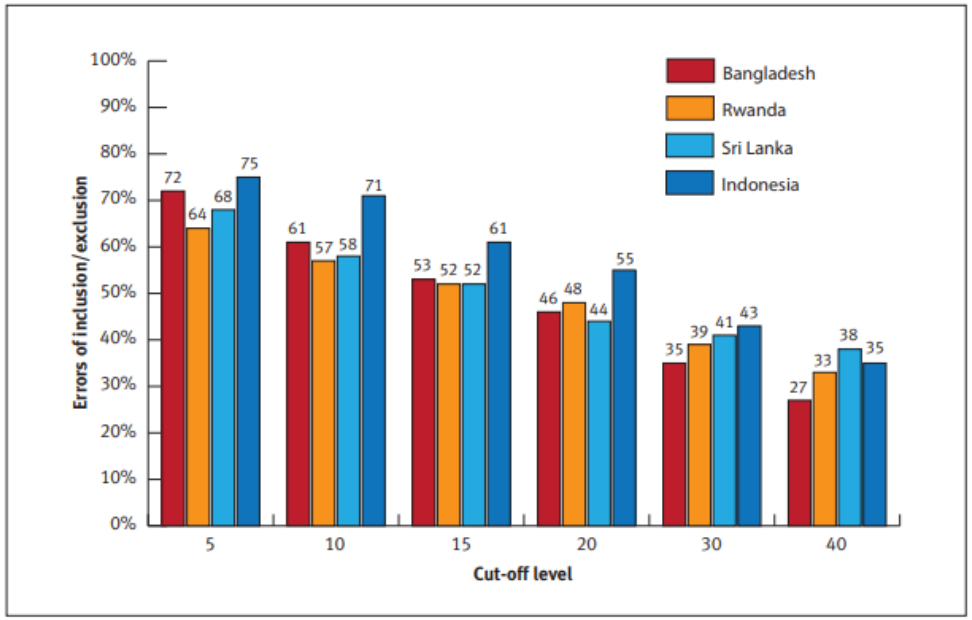
데이터셋은 <미주 개발 은행(Inter-America Development Bank)>에서 가구의 빈곤 수준을

측정하기 위해 수집한 자료

복지 재원을 분배할 때 복지 사각지대를 줄이기 위해 PMT(Proxy Mean Test) 기법을 주로 활용해옴

PMT 기법의 오차율을 줄이기 위하여, 은행이 보유한 데이터셋을 Kaggle Competition에 공개하여 전세계 데이터 분석가들의 참여를 유도함

Proxy Means Test?



오차율(%)

<빈곤선 threadhold>

- Proxy Mean Test란 가구의 소득 수준을 정확하게 파악하기 힘들 때, 대신 그 가구의 특성을 변수로 이용하여 복지 수급 대상 가구인지 아니인지를 분류하는 테스트 기법이다.(worldbank, 2018)

- PMT 기법은 소득 통계 확보가 미비한 저개발국의 소득 수준을 측정하기 위하여 여러 국가에서 활용되어 왔으나, 전통적인 회귀분석을 이용할 경우 빈곤선 설정 수준에 따라 40%~70%의 큰 오차율을 보이는 경우가 많았다.(UNICEF, 2011)

데이터 전처리

Agg_df

샘플들을 가구별 ID로 묶어준다.(Groupby)

다음의 기준으로 연산을 실시한다.

변수명	설명	연산
'estadocivil'	결혼여부(7점 척도)	평균,빈도
'parentesco’	가정내 지위(12점 척도)	평균,빈도
'instlevel’	최종학력(9점 척도)	평균,빈도
‘age’	나이	최대,최소,평균
‘escolari’	취학년도	최대,최소,평균

데이터 전처리

Pd.groupby("idhogar")

	idhogar	age	escolari	instlevel1	instlevel2	instlevel3	instlevel4	estadocivil1	estadocivil2	estadocivil3
3	513	17	9	0	0	0	1	0	0	0
4	513	37	11	0	0	0	0	0	1	0
5	513	38	11	0	0	0	0	0	1	0
6	513	8	2	0	1	0	0	1	0	0

<aggregate>

(min,max,mean) (mean,count) (mean,count)

aggAll_age_MIN	aggAll_age_MAX	aggAll_age_MEAN	aggAll_instlevel1_MEAN	aggAll_instlevel1_COUNT	aggAll_estadocivil3_MEAN	aggAll_estadocivil3_COUNT
8.0	38.0	25.0	0.0	4.0	0.0	4.0
8.0	38.0	25.0	0.0	4.0	0.0	4.0
8.0	38.0	25.0	0.0	4.0	0.0	4.0
8.0	38.0	25.0	0.0	4.0	0.0	4.0

이제 가정 내 가구원들은 가족 모두 동일한 특성을 공유한다.

데이터 전처리

Missing Value 표시기

결측값이 있는 컬럼과, 해당 컬럼의 결측값 비율을 보여준다.

```
# missing value 표시기
missing = pd.DataFrame(train.isnull().sum()).rename(columns = {0: 'total'})

missing['percent'] = missing['total'] / len(train)

missing.sort_values('percent', ascending = False).head(10)
```

	total	percent
rez_esc	7928	0.829549
v18q1	7342	0.768233
v2a1	6860	0.717798
SQBmeaned	5	0.000523
meaneduc	5	0.000523
Id	0	0.000000

데이터 전처리

Dependency

Dependency : (가정 내 18세 이하/64세 이상 피부양자 수) / (가정 내 18세 이상 64세 이하 노동인구)

SQBdependency가 이미 존재하므로, 이를 이용해 Dependency로 역연산을 한다.

<dependency-original>

dependency
0.0
8.0
8.0
1.0
1.0

x^2

<SQBdependency>

```
0     0.0
1    64.0
2    64.0
3     1.0
4     1.0
5     1.0
6     1.0
7     1.0
8     1.0
9     1.0
Name: SQBdependency, dtype: float64
```

$\sqrt{x}$

	dependency
0	no
1	8
2	8
3	yes
4	yes

데이터 전처리

Edjefe, edjefa

Yes값과 no값이 있는 변수들인 edjefa, edjefe 변수를 0 또는 1로 전부 처리해준다.

	edjefe	edjefa
0	10	no
1	12	no
2	no	11
3	11	no
4	11	no
5	11	no
6	11	no
7	9	no
8	9	no
9	9	no
10	9	no

Yes = 1

No = 0

	edjefe	edjefa
0	10	0
1	12	0
2	0	11
3	11	0
4	11	0
5	11	0
6	11	0
7	9	0
8	9	0
9	9	0
10	9	0

데이터 전처리

maxedu

남성 가장 최고학력과, 여성 가장 최고학력을 통합하여 maxedu 변수를 만들어준다.

	edjefa	edjefe
8481	0	9
8482	0	9
8483	0	9
8484	0	9

	maxedu
8481	9
8482	9
8483	9
8484	9

데이터 전처리

Rez_esc

전체 나이 중 미취학년도를 나타내는 변수

	rez_esc	age	escolari
0	NaN	43	10
1	NaN	67	12
2	NaN	92	11
4	NaN	37	11
5	NaN	38	11
8	NaN	30	9
9	NaN	28	11
11	NaN	18	12
12	NaN	34	11
13	NaN	79	4

(Age – Escolari)

	0
0	26
1	48
2	74
4	19
5	20
8	14
9	10
11	-1
12	16
13	68

If (Age-Escolari) > 5 : 5

If (Age-Escolari) = 5 : 0

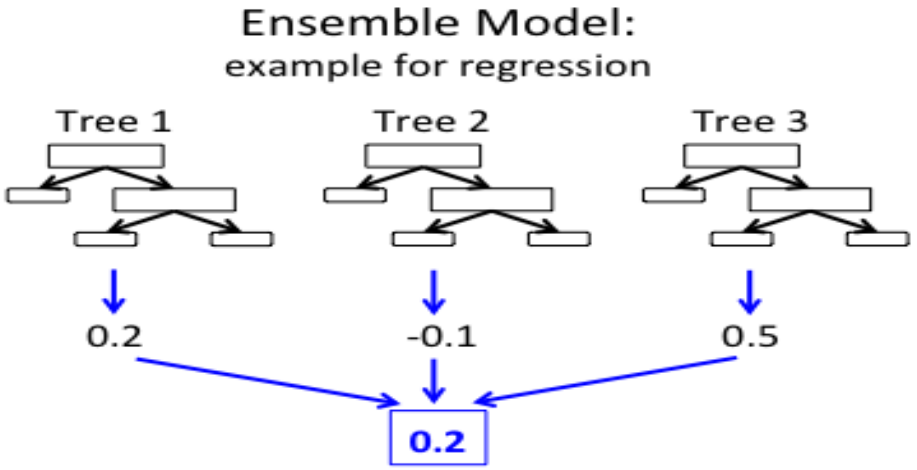
	0
0	5
1	5
2	5
4	5
5	5
8	5
9	5
11	0
12	5
13	5

데이터 전처리

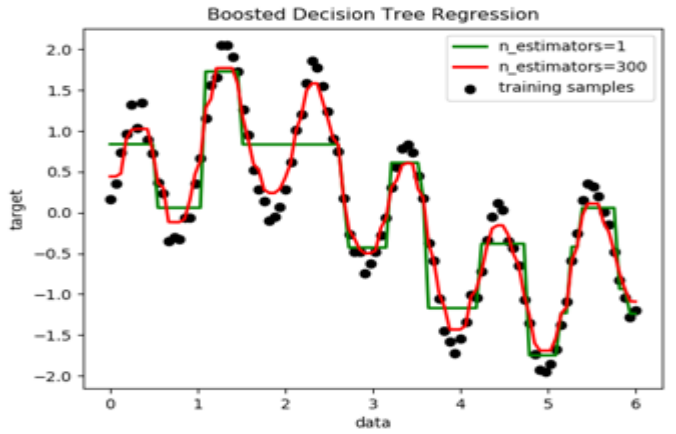
V2a1(임대료)

XtreeRegressor를 이용해 특성들이 비슷한 것으로 분류된 샘플들의 v2a1 평균값으로 v2a1의 결측값을 예측한다.

앙상블 기법에 기반한 트리모형의 회귀예측은 여러 약분류기(트리)들의 동일 그룹 평균값의 평균값을 토대로 예측치를 생산한다.



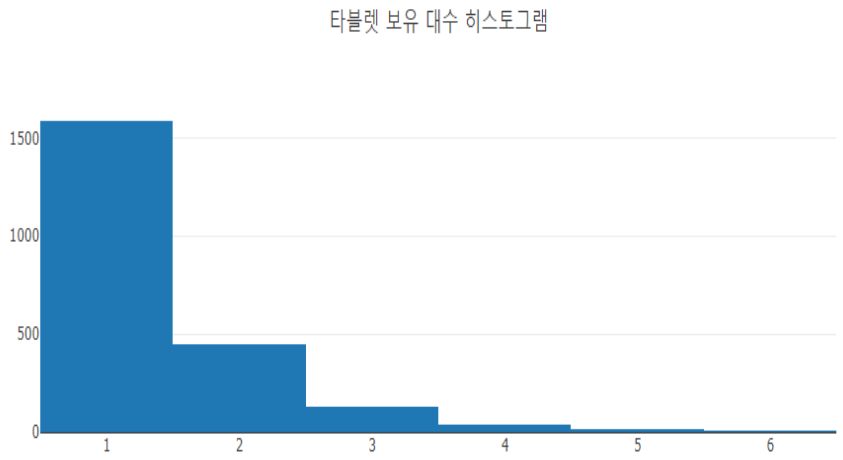
<앙상블 모형의 회귀 예시>



<회귀선>

데이터 전처리

V18q1



"0대"라고 응답한 경우가 없는 것으로 봐서,
결측치가 곧 0대임을 의심해볼 수 있다.

	결측	비결측
"보유했다"	0	2215
"보유하지 않았다"	7342	0

<결측-비결측 비교 행렬>

미보유 응답자 = 결측치

∴ 결측을 전부0으로 처리

데이터 전처리

그 외 변수들

```
train['meaneduc'].fillna(0, inplace=True)
test['meaneduc'].fillna(0, inplace=True)
train['SQBmeaned'].fillna(0, inplace=True)
test['SQBmeaned'].fillna(0, inplace=True)
```

정보성 결측치들을 대부분 전처리 하였으므로
의미가 없다고 판단된 나머지 결측치들은
전부 0으로 일괄 처리

데이터 전처리

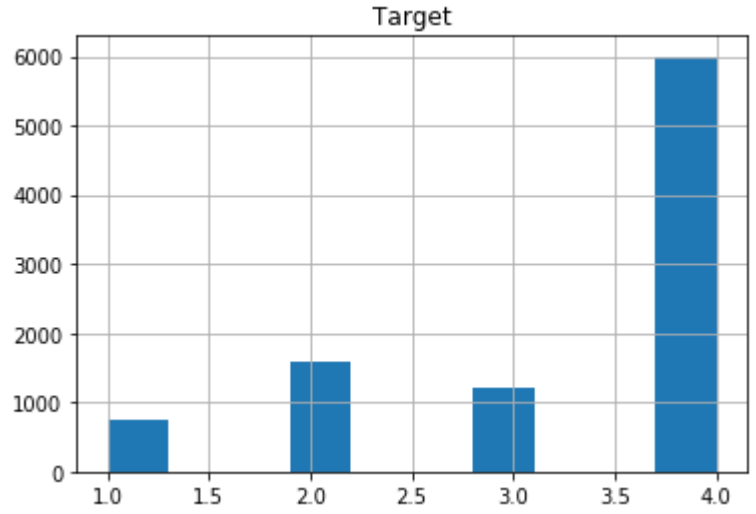
전처리 결과 (Missing Value 표시기)

```
# missing value 표시기
missing = pd.DataFrame(train.isnull().sum()).rename(columns = {0: 'total'})

missing['percent'] = missing['total'] / len(train)

missing.sort_values('percent', ascending = False).head(10)
```

	total	percent
v2a1	0	0.0
Target	0	0.0
age	0	0.0
SQBescolari	0	0.0
SQBage	0	0.0
SQBhogar_total	0	0.0
SQBedjefe	0	0.0
SQBhogar_nin	0	0.0
SQBovercrowding	0	0.0
SQBdependency	0	0.0

모델 적합

타겟 불균형의 처리



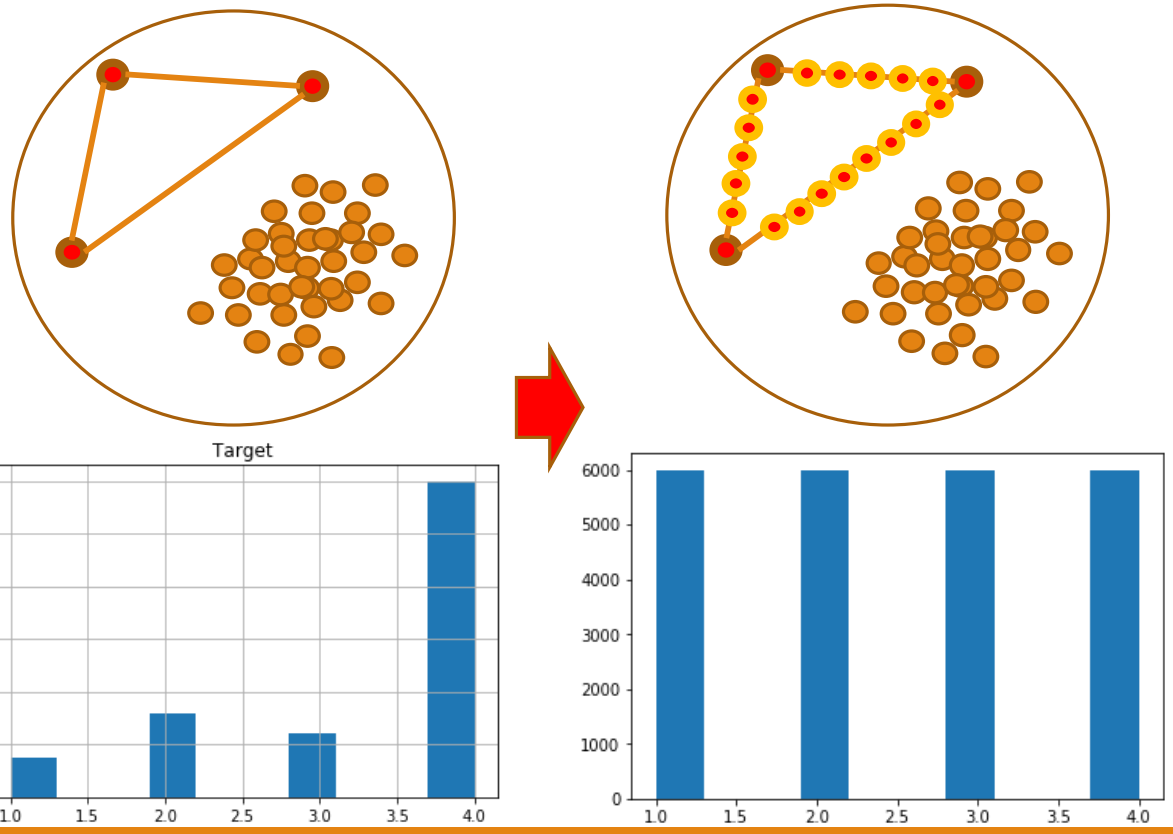
“보통(4)”이 전체의 62%를 차지

타겟수가 적은 타겟에 대해선 학습기회를 별로 갖지 못하고

F1 Score상 오분류된 샘플에 대한 평가치가 크게 상승함

모델 적합

타겟 불균형의 처리

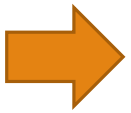


SMOTE 알고리즘을 통해 업샘플링(Up-sampling) 실시

모델 적합

파라미터 검색

파라미터명	설명	Max	Min
Learning_rate	경사하강법 학습률	수렴 속도 상승	수렴 속도 하락
		정확도 하락(수렴 실패)	정확도 상승
N_estimator	약분류기 수	과적합 방지	과적합
		학습 속도 하락	학습속도 상승
Num_leaves	분류기당 가지 수	정확도 상승	정확도 하락, 과소적합
		과적합,학습속도 하락	학습속도 상승
Colsample_bytree	무작위 변수 선택 확률	정확도 상승	정확도 하락
		과적합, 학습속도 하락	학습속도 상승
subsample	무작위 샘플 선택 확률	정확도 상승	정확도 하락
		과적합, 학습속도 하락	학습속도 상승



```
model4 = lgb.LGBMClassifier(boosting_type = "gbdt",
                            objective = "mutliclass",
                            max_depth = -1,
                            random_state = 0)
```

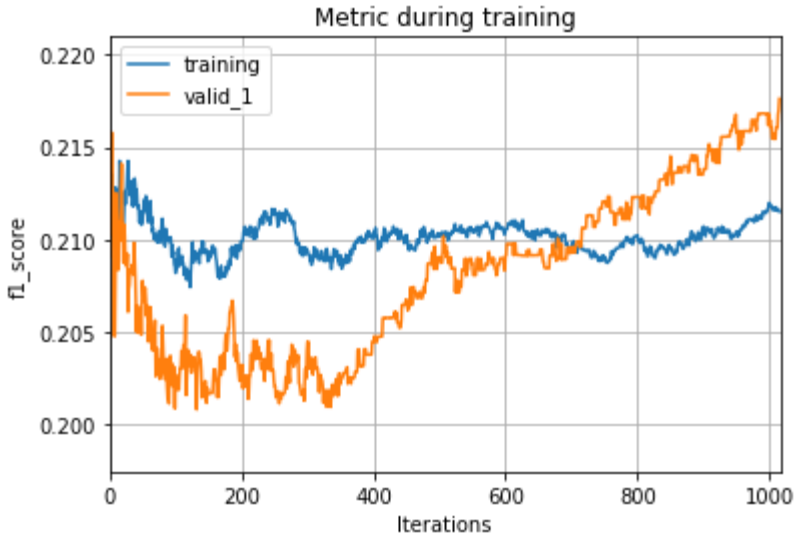
```
gridParams = {
    'learning_rate': [0.01, 0.005],
    'n_estimators': [40,80,100],
    'num_leaves': [10,20,30,40],
    'colsample_bytree' : [0.65, 0.66],
    'subsample' : [0.7,0.75],
    }


grid = GridSearchCV(model4,
                    gridParams,
                    verbose = 0,
                    cv=5,n_jobs=-1)
```

변수명	그리드 서치 추천값
colsample_bytree	0.66
learning_rate	0.01
n_estimators	100
Num_leaves	10
Subsample	0.7

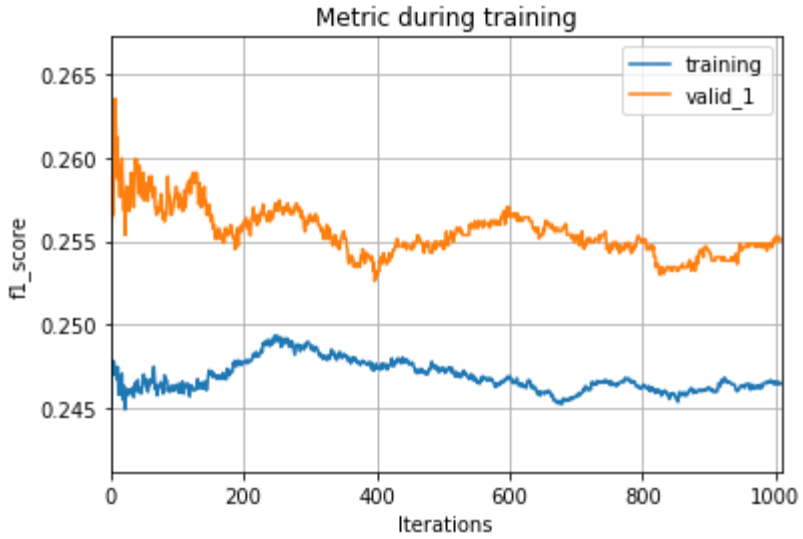
모델 적합

모델 적합

<원본 데이터셋>



<업샘플링 데이터셋>



f1_score 0.5점 상승 확인, Kaggle 최다점 0.410점에 근접

감사합니다

부록

변수표

개괄	가구특성(거주 형태)	가구특성 (어린이)	가구특성 (생산가능인구)	개별특성 (가구원 지위)	개별특성 (결혼)	개별특성 (교육)
ID : 개별 케이스(가구원)	V2a1 : 월간 임대료	R4h1 : 12세 이하 남자 가구원 수	Tamhog : 총 가구원 수	Parentesco1 : 가장	Estadocivil1 : 10세 미만6	Escolari : 취학년도
Target : 소득 수준	Tipovivi1 : 자가소유, 대출 없음	R4h2 : 12세 이상 남자 가구원 수	Hhsize : 총 가구원 수	Parentesco2 : 아내/남편/동거인	Estadocivil2 : 동거	Rez_esc: 미취학년도
IDhoger : 가구를 구분하는 구분자	Tipovivi2 : 자가소유, 대출 있음	R4h3 : 총 남자 가구원수	Tamviv : 가구와 함께 거주하는 사람의 수	Parentesco3 : 자녀	Estadocivil3 : 결혼	Instlevel1 : 무학
	Tipovivi3 : 임대	R4m1 : 12세 이하 여자 가구원 수	Hogar_nin : 0~19세 아동의 수	Parentesc4 : 의붓자녀	Estadocivil4 : 이혼	Instlevel2 : 초등학교 중퇴
	Tipovivi4 : 불확실	R4m2 : 12세 이상 여자 가구원 수	Hogar_adul : 성인의 수	Parentesco5 : 사위/며느리	Estadocivil5 : 별거	Instlevel3 : 초등학교 졸업
	Tipovivi5 : 기타	R4m3 : 총 여자 가구원 수	Hogar_mayor : 65세 이상 성인의 수	Parentesco6 : 손자녀	Estadocivil6 : 사별	Instlevel4 : 인문계 중등과정 중퇴
		R4t1 : 12세 이하 가구원 수	Hogar_total : 총 가구원 수	Parentesco7 : 부모	Estadocivil7 : 독신	Instlevel5 : 인문계 중등과정 졸업
		R4t2 : 12세 이상 가구원 수	Dependency : 의존도	Parentesco8 : 장인/장모		Instlevel6 : 실업계 중등과정 중퇴
				Parentesco9 : 형제자매		Instlevel7 : 실업계 중등과정 졸업
				Parentesco10 : 처남/매부/매형		Instlevel8 학부 졸업 혹은 석사과정 수료
				Parentesco11 : 그 외 가족구성		Instlevel9 : 박사 과정 수료
				Parentesco12 : 그 외 비가족구성		

가전제품 보유현황	벽의 형태	바닥의 형태	지붕의 형태	인프라 연결	위생 시설
V18q : 타블렛 보유 여부	Paredblolad : 벽의 주된 재질이 (시멘트)블록이거나 벽돌	Piscomoscer 바닥의 재질이 대리석 등	Techozinc 지붕의 재질이 금속	Abastaguadentro 수도시설이 집안 내부까지 연결됨	Sanitario1 집 내부에 화장실 없음
V18q1 : 가구에서 보유한 타블렛의 수	Paredzocalo: 벽의 주된 재질이 합판(socket)	Pisocemento 바닥의 재질이 시멘트	Techoentrepiso 지붕의 재질이 직물,시멘트 등	Abastaguafuera 수도시설이 외부하고만 연결됨	Sanitario2 화장실이 하수도와 직결됨
refrig : 냉장고 보유 여부	Paredpreb: 벽의 주된 재질이 조립식	Pisoother 바닥의 재질이 기타	Techocane 지붕의 재질이 자연물	Abastaguano 수도시설 연결 안됨	Senitario3 화장실이 정화조와 연결됨
mobilephone : 핸드폰 보유 여부	Pareddes : 벽의 주된 재질이 폐기물(waste0	Pisonatur 바닥의 재질이 자연물	Techootro 지붕의 재질이 기타	Public 전력 회사로부터 전력 공급	Senitario5 화장실이 단순 구덩이 혹은 공중변소
qmobilephone : 핸드폰 보유 대수	Paredmad: 벽의 주된 재질이 나무	Pisonotiene 바닥 없음	Cielorazo 집이 천장을 가지고 있는지 여부	Planpri 개별 발전	Senitario6 화장실이 다른 처리 시스템과 연결됨
computer : 컴퓨터 보유 여부	Paredzinc: 벽의 주된 지질이 아연	Pisomadera 바닥의 주된 재질이 나무	Etecho1 지붕 상태가 불량	Noelec 집안 내부까지 전력 공급 안됨	Elimbasu1 폐기물을 쓰레기차가 처리
television : TV 보유 여부	Paredfibras 벽의 주된 재질이 직물	Eviv1 바닥 상태가 불량함	Eteco2 지붕 상태가 보통	Coopele 전력을 공동 생산	Elimbasu3 폐기물을 땅에 묻음
	Paredother: 벽의 주된 재질이 기타	Eviv2 바닥 상태가 보통	Eteco3 지중 상태가 좋음		Elimbasu3 폐기물을 소각
	Epared1 벽 상태가 불량함	Eviv3 바닥 상태가 좋음			Elimbasu4 폐기물을 빈 공간에 무단폐기
	Epared2 벽 상태가 보통				Elimbasu5 폐기물을 강, 바다에 폐기
	Epared3: 벽 상태가 좋음				Elimbasu6 폐기물을 기타 방법으로 처리