Multidimensional Poverty Predictor

Introduction

This project applies machine learning to the study of poverty. Using the Eurostat multidimensional material deprivation indicator as target variable it builds a prediction model based on relevant sociodemographic variables.

The code written for this project is present on the repository. All the notebooks are numerically ordered by stage. Reference to the code notebooks is given throughout the report.

The Multidimensional Porverty Predictor is the streamlit app developed by this project. Link available at the end of this document.

Framework

Poverty is a complex phenomenon to measure and it can be approached in many different ways.

The traditional way to study poverty is through the definition of a poverty line. An indicator, usually income or expenditure, is chosen and people failing to pass a certain threshold are classified as poor. A typical example to measure relative poverty in this way is to draw a line at 60% of the median household income. People falling below that income value would be then regarded as poor in relative terms.

Poverty lines are useful for their simplicity and objectivity. However, poverty is a multidimensional phenomenon and monetary poverty is often not enough to capture it completely. Households with the same income may have different standards of living. It does not neatly adjust to certain cases (ex. Freelance workers). Individuals have other resources not reflected in monetary poverty that can be used to avoid poverty like education, a support net, access to credit, etc.

Researchers generally study multidimensional poverty by looking at indicators that directly measure living conditions. They focus on access to basic consumption elements or evidence of economic difficulties. Measurements built with multidimensional indicators are better at reflecting the depravation suffered by households. They can also be used to make fairer comparisons.

This project uses the Eurostat multidimensional material deprivation indicator as target variable. The methodology for this variables is explained below.

Methodology

The methodological process consist on four main stages:

Decoding source data → Data preparation → Data preprocessing → Feature selection → Modeling

Decoding source data

The European Union Income and Living Conditions survey (EU-SILC) is the data source for this project. The Eurostat conducts this survey every year through the official agencies on each of the EU countries. Its aim it so collect timely and comparable micro-data on income and living conditions.

The EU-SILC collects data at household and individual level. Households provide information on income, social inclusion and housing conditions. Individuals in turn reveal information on labor, education and health. The reference population consist in all private households and their current members. This can somewhat limits the scope of the survey since it excludes homeless individuals and people living in collective households and institutions. Any analysis or training model built with the EU-SILC data would need to be aware of this limitation.

This project focuses exclusively on the most recent survey carried out in Spain (2019 as of present date). Access to the micro-data can be obtained freely from the National Spanish Statistics Institute (*Instituto Nacional de Estadística*, INE) through this link.

The source data is stored in four CSV files. Two of them belong to the household level survey and the other two to the individual level survey. The values and column names are encoded but reference is available in the Eurostat methodological guidelines (A pdf copy is provided in the repo). I manually decoded every column following these indications in notebooks 1.1-1.4. In notebook 2 I merged all CSVs into one single dataframe using the 'Household ID' column and the 'Person ID' column as keys.

Data preparation

A set of variables are chosen and transformed from the source data. All the variables need to be relevant to the study. I comment in this document on their importance to the subject at hand in each case.

The EU-SILC is a rich survey with more than 200 questions. The variable creation process is sometimes straightforward as in the case of 'Sex'. Other times it requires some playing around with different questions as is the case with the 'Working status' variable. In the end, 11 categorical variables and 5 numerical variables are chosen. The code for this stage is stored in notebook 3.

A description of each variable follows down below:

-Material deprivation

Multidimensional material deprivation is the target variable in this project. The Eurostat methodology for severe material depravation is a complex variable made up of nine indicators. These are:

- Inability to afford paying for one week annual holiday away from home.

- Inability to afford a meal with meat, chicken or fish (or vegetarian equivalent) every second day.

- Inability to keep home adequately warm.

- Inability to face unexpected expenses.

- Arrears on utility bills, mortgage or rental payments, or hire purchases or other loan payments.

- Inability to afford a car.

- Inability to afford a telephone.

- Inability to afford a TV.

- Inability to afford a washing machine.

Households with positive values on at least four elements of this list are classified with material deprivation. This variable is available in the source data.

-Sex

This project uses 'Sex' and 'Gender' interchangeably. This is because the EU-SILC only considers the variable 'Sex', seemly referring to the physical traits of the respondent in opposition to the social role associated with the sex or 'Gender'. Leaving out gender identity results in the loss of potentially valuable information on the economic marginalization of non-normative gender people such as the trans-gender community.

Gender is variable commonly associated with poverty. Usual explanatory factors are women weaker attachments to the labor market, pay discrimination and concentration on lower paid occupations.

-Age

Age in the EU-SILC survey brings some limitation. First underage respondents get asked a different set of questions, which leaves information lacking on several of the aspects considered. For this reason, this project focuses exclusively on the adult population. Additionally, the survey asks for the year of birth but anyone born before 1932 is assigned with '1932' as his or her birth year regardless if he or she was born earlier. The possibility of excluding the 86 plus years old was pondered but eventually discarded. Removing the oldest people in the sample could result in a distorting effect on the potential relationships between old age and multidimensional poverty.

Growing old is sometimes associated with an additional risk of remaining or becoming poor. Old people reduce their working hours or stop working all together because of retirement or health issues. Sometimes pension systems are not adequate.

-Civil Status

Marital status consist on the status of each individual in relation to the marriage laws of their country. Therefore it does not necessarily correspond with the actual situation of the household in terms of co-habitatation or other arrengements. The EU-SILC accounts for legal statuses as well as non-married co-habitants partners through two separate questions. By combining both an additional category is added to account for this 'de facto' partnership.

The civil status variable ends up with the following categories: 'Married', 'Married *de facto*', 'Never married', Separated, 'Divorced' and 'Widowed'.

Wealth and marital status are often linked together due the economies of scale of being a single unit. Married couples can share expenses or be more likely to make long-term investments such as buying homes.

-Familial status

Famililal status means that underage individuals are part of a household. The EU-SILC considers different combinations of adults and children in the household composition. By means of simplification household composition is reduced to whether the household has children or not.

In high-income countries having children is does not necessarily imply a higher liklihood of being poor. Birth control allows families to decide whether to have children based on their economic expectations for the future. However, economic circumstances do change and the pressure of having one or more dependent children can arguably stress financial security when accompanied with a sudden loss of wages or employment.

-Region

Region accounts for the different Autonomous Communities of Spain as well as the african enclaves of Ceuta and Melilla. This information is present in the source data.

-Population density

The EU-SILC classifies population density in three ordinal categories. Densely populated area requires a minimum population of 50000 and at at least 1 500 inhabitants per km2 nearby. Intermediate area requires a minimum population of 5000 and a nearby population of 300 inhabitants per km2. The rest are classified as thinly-populated.

Both rural and urban environments have concerning relationships with poverty. Rural environments suffer from remoteness and isolation, which can lead to limited access to basic services and weaker labor market. Urban environments have strong segregation dynamics that tend to cluster poverty in certain neighborhoods. Which ones weights more depends on the particular characteristics of a country or region.

-Citizenship

The EU-SILC only considers the Spanish citizenship specifically. The rest are categorized on wether they come from the EU or outside the EU. The same is done with the country of birth. Combining both questions an additional category can be built for those who were born elsewhere but became Spanish citizens by naturalization.

Poverty is on itself an important driver of migration. The incidence of poverty once economic migrants arrive at their destinations can vary according to the position of migrants in the labor market, coupled with their legal status.

-Tenure status

The EU-SILC only contamplates owenership when the ownership is held over the household accomodation. The owner is considered as 'outright owner' when he/she has no more mortgage to pay and an owner is considered as an 'owner paying mortgage' when he/she still has to pay for the mortgage. ‘Tenancy at market rate’ makes a distinction from those renting social housing, renting at a reduced rate from an employer or live in an accomodation where the actual rent is fixed by law, whcih are categoraized as ‘Tenancy at reduced rate’. Finally ‘Free tenancy’ applies only when there is no rent to be paid, such as when the accommodation comes with the job or is provided rent-free from a private source.

Increases in income inequality over the years have been associated with a growing disparity in the affordability of housing between lower and middle-income homeowners and renters. The different housing options that the rich and poor can afford contributes to economic segregation, which exacerbates the effects of poverty.

-Education level

The educational attainment level of an individual is measured by his o her highest ISCED (International Standard Classification of Education) level successfully completed and validated by a recognised certification.

The contamplated categories are ‘Pre primary education’, ‘Primary education’, ‘Lower secondary education’, ‘Upper secondary education’ and ‘Higher education’.

Education can open access to better paying occupations and more opportunities in life which can help to avoid poverty.

-Working status

This variable considers different activity statuses. Categories are drawn from two different questions in the survey regarding the respondent current main activity status. This are: ‘Employed’, ‘Unemployed’, ‘Retired’, ‘Student’, ‘Disabled/Unfit to work’, ‘Unpaid carer/domestic worker’.

The loss of income caused by unemployement can be a cause of poverty, specially if perpetuated in time.

-Occupation

The EU SILC uses the International Standard Classification of Occupations (ISCO-08) to account for the respondents job positions. The ISCO has major, submajor and minor groups. The survey reaches down at the submajor layer which is commpossed of 43 distinct groups. I reagroup them into the major layer which is compossed of 9 groups. This are: ‘Managers’, ‘Professionals’, ‘Technicians and Associate Professionals’, ‘Clerical Support Workers’, ‘Services and Sales Workers’, ‘Skilled Agricultural, Forestry and Fishery Workers’, ‘Craft and Related Trades Worker’, ‘Plant and Machine Operators and Assemblers’ and ‘Elementary Occupations’. Those without a clearly defined occupation or without one all together are categorized under ‘Non defined’.

While employment reduces considerably the poverty risk in-work poverty can still exist. Low-skill jobs tend to have much lower wages and be less secure at times of economic distress.

-Years worked

This indicator provides a numerical total of the years of work experienced, since the respondent started their first regular job, whether as an employee or self-employed. For the working poor, underemployment can also be a major problem. This variable controls for intermitent periods of unemployement in the working history of an individual.

-Hours a week worked

This indicator provides a numerical total of the hours worked every week in any job held by the respondent. Another face of underemployement is not being able to work full time. Having more than one job can also be an indicator of economic necessity.

-Bad health

The EU-SILC asks four different questions on matters of health. They are subjective in nature and aimed at measuring different dimessions of health, not necessairly phisical but also social or emotional. The first one refers to a general state of health as perceived by the respondent. The second one inquires on the existance of any chronic or longstanding illness or condition. The third asks about limitations in activities caused by health problems. The last one asks about unmet needs for medical treatment.

'Bad health' consist on a yes or no variable. A negative awnser (in health terms) in any of the previously mentioned health questions grants a yes in this variable.

Poor individuals may have higher rates of physical limitation and of heart disease, diabetes, stroke, and other chronic conditions. Less access to fresh food or a built environment less conducive to physical activity (ex open spaces) can be attributed to poverty.

-Adjusted income

Even though monetary indicators not always make for suitable definitions poverty they can be a powerfull predictor for multidimensional poverty. There is an obvious relationship between poverty and income. However income can be tricky to measure since the benefits of income can be enjoyed by all household members. Someone who does not work can be well off it someone else in the household brings in enough income for both. Encomies of scale also apply in households. This project adjusts annual household income including imputed rent (imputed rent is the value of the rent the owner would pay if they were the tenant of their property) by dividing it by the OECD consumption unit scale. This unit assigns 1 plus 0.5 times the number of other household members older than 13, plus 0.3 times the number of other household members younger than 13. The resulting value is a fairer income variable.

-Proportion of social welfare

The EU-SILC provides datailed information on income sources. The proportion of social welfare is computed by dividing the annual social welfare received by a household by its total annual income. Those in need of benefits can be expected to be those more vulnerable to poverty. At the same time if the welfare received is not enough the liklihood of experiencing multidimensional poverty increases.

Data preprocessing

We are left with the resulting dataframe.

```
Data columns (total 18 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   weight                     32946 non-null  float64
 1   material_deprivation       32946 non-null  object
 2   sex                        32946 non-null  object
 3   age                        32946 non-null  int64
 4   civil_status               32946 non-null  object
 5   familial_status            32946 non-null  object
 6   region                     32946 non-null  object
 7   population_density         32946 non-null  object
 8   citizenship                32946 non-null  object
```

```
 9   tenure_status               32946 non-null  object
 10  education_level             32946 non-null  object
 11  working_status              32946 non-null  object
 12  occupation                  32946 non-null  object
 13  years_worked                32946 non-null  int64
 14  hours_week_worked           32946 non-null  int64
 15  adjusted_income             32946 non-null  float64
 16  proportion_social_welfare   32946 non-null  float64
 17  bad_health                  32946 non-null  object
dtypes: float64(3), int64(3), object(12)
```

In the preprocessing stage the categorical variables are encoded using the sklearn OneHotEncoder function. In turn the numerical variables are scaled using the sklearn MinMaxScaler function. I decide on the MinMax scaling method so all my variables would have values from 0 to 1. The code for both data preprocessing and feature selection can be found in the notebook 4.

Feature selection

I use univariate statistical tests for feature selection. I set the mark at a pvalue of 0.001. For the numerical variables an f test is performed. For the categorical variables a chi square test is performed.

Selected variable by p value rank:

```
proportion_social_welfare: 0.0
adjusted_income: 5.033285458156601e-275
working_status_unemployed: 8.051437484018846e-173
citizenship_other_(outside_eu): 1.8055583701040246e-133
tenure_status_tenancy_at_market_rate: 3.177962486690465e-128
tenure_status_tenancy_at_reduced_rate: 4.1693597553099486e-83
hours_week_worked: 3.3585584128110176e-66
years_worked: 3.692732521983624e-48
tenure_status_outright_owner: 5.482118720960313e-48
education_level_higher_education: 8.808748417405019e-43
civil_status_separated: 7.174860522449456e-40
occupation_elementary_occupations: 2.6537138848278216e-39
citizenship_spain_(naturalized): 3.7972457170917797e-35
age: 3.4710774381473146e-30
working_status_employed: 5.545308308897578e-28
occupation_professionals: 4.2033038139028187e-26
citizenship_spain: 1.5622364301290388e-24
education_level_pre-primary_education: 1.981190305131728e-24
working_status_disabled/unfit_to_work: 8.662988471131772e-22
working_status_retired: 4.851535652756932e-21
occupation_clerical_support_workers: 1.2747739477208076e-18
bad_health_yes: 8.460000529663107e-17
civil_status_married: 1.132149574742828e-15
region_andalusia: 6.922515083275146e-12
civil_status_never_married: 3.303162243846116e-11
education_level_lower_secondary_education: 4.042518702474464e-11
tenure_status_owner_paying_mortgage: 1.2560132369437168e-10
occupation_technicians_and_associate_professionals: 3.4070146060969716e-10
tenure_status_free_tenancy: 5.520409848378553e-10
```

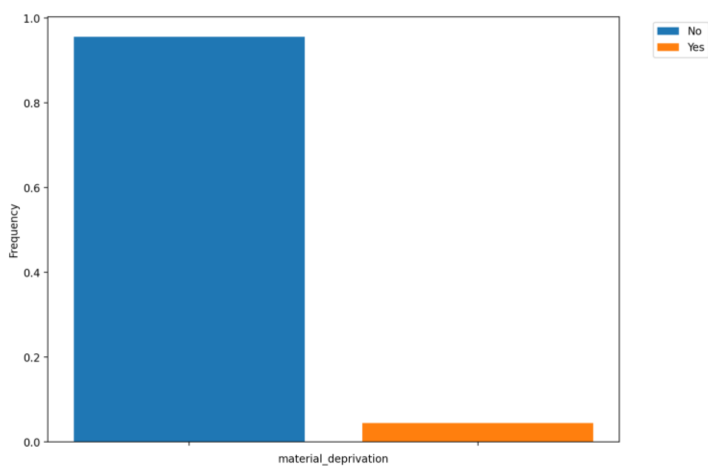
```
bad_health_no: 1.18294168521748e-08
occupation_non-defined: 1.675549797761655e-07
citizenship_other_(eu): 5.11003541616413e-07
occupation_managers: 5.0581778417534155e-05
region_castile_and_leon: 5.536542613771125e-05
population_density_thinly-populated_area: 9.534205087468083e-05
region_cantabria: 0.00020867557824668701
education_level_primary_education: 0.0003714450882445782
region_castile-la_mancha: 0.0004786865958479371
region_basque_country: 0.0005986628009323659
```

Sex and familial status categories are left behind. This does not necessarily imply women or children are not more likely to experience poverty in Spain. Intra-household disparities in the allocation for work and resources may for example account for higher degrees of inequality among the female members. However, the EU-SILC gathers data on social at household level, which makes impossible to account for potential differences taking place inside the households at individual level. For all intends and purposes this project assumes poverty is shared equeally among the household members.

At the top of the rank income, housing, unemployement and inmigration seem to be the areas with the strongest univariate relationship with material deprivation.

Modeling

Just 4.46% of the Spanish population is categorized with material deprivation. Imbalanced data can result in poor performance in most machine learning algorythms. I apply undersampling and oversampling techniques along with the imbalanced data to control for this risk.



The machine learning techniques used are: Logistic Regression (notebook 5.1), Decision Tree (notebook 5.2), K-Neighbors (notebook 5.3), Naïve Bayes (notebook 5.4), Support Vector Machines (notebook 5.5) and Stochastic Gradient Descent (notebook 5.6).

The resampling techiniques used with each of the algorightms are RandomUnderSampler, RandomOverSampler and SMOTE from the imlearn library. SMOTE is an oversampling technique were new synthetic samples are created from the minority labels and their k neighbors through a line segment in the feature space.

Given the imbalanced nature of the data I was less concerned with acquaracy and more concern with precission and recall. Bellow is a table with the F1 scores of all atempted models. Overall most models performed poorly. Resampling improved the results of the logistic regression and the support vector machine. Undersampling actually made performance worse with the decision tree and k-neighbors. Random oversampling performed just as good as no resampling for the decission tree but did not improved it.

F1 Scores	LogReg	Decision Tr	K Neighbor	Naïve Bay	SVM	SGD
Imbalanced	0.09	0.76	0.68	0.22	0.12	0.00
Undersampled	0.23	0.28	0.27	0.20	0.26	0.27
Oversampled	0.23	0.74	0.68	0.19	0.41	0.26
SMOTE	0.22	0.66	0.64	0.17	0.46	0.24

In the end the best performing model was the decision tree with the imbalanced training data.

```
              precision    recall  f1-score

           0       0.99      0.99      0.99
           1       0.75      0.77      0.76
    accuracy                           0.98
```

It results in the following feature importance rank:

```
adjusted_income: 37.0%
proportion_social_welfare: 8.0%
age: 6.0%
years_worked: 4.0%
population_density_thinly-populated_area: 3.0%
occupation_elementary_occupations: 3.0%
hours_week_worked: 3.0%
region_andalusia: 3.0%
civil_status_married: 2.0%
working_status_unemployed: 2.0%
bad_health_yes: 2.0%
education_level_lower_secondary_education: 2.0%
tenure_status_outright_owner: 2.0%
education_level_pre-primary_education: 2.0%
tenure_status_tenancy_at_market_rate: 1.0%
tenure_status_owner_paying_mortgage: 1.0%
civil_status_never_married: 1.0%
citizenship_other_(eu): 1.0%
occupation_technicians_and_associate_professionals: 1.0%
education_level_higher_education: 1.0%
education_level_primary_education: 1.0%
bad_health_no: 1.0%
working_status_employed: 1.0%
working_status_retired: 1.0%
```

```
occupation_non-defined: 1.0%
tenure_status_free_tenancy: 1.0%
civil_status_separated: 1.0%
region_basque_country: 1.0%
citizenship_spain: 1.0%
region_castile-la_mancha: 1.0%
occupation_clerical_support_workers: 1.0%
citizenship_other_(outside_eu): 1.0%
occupation_professionals: 1.0%
citizenship_spain_(naturalized): 1.0%
region_castile_and_leon: 0.0%
tenure_status_tenancy_at_reduced_rate: 0.0%
working_status_disabled/unfit_to_work: 0.0%
occupation_managers: 0.0%
region_cantabria: 0.0%
```

Conclusions

The application of machine learning to the study of poverty can benefit both researchers and public managers. It can help researchers uncover combined statistical dependencies. It can also help society to deal with terrible social phenomenon. The model achieved certain degree of success. It has virtually no false negatives and was able to identify three out of four positives. Improved sets of variables and implementation of techniques could surely take it even further. The potential for models like this to help allocate resources for those in need is certanily promising.

Annexes

Link to the streamlit app of this project:

https://share.streamlit.io/deividvalerius/multidimensional-poverty-predictor/poverty_predictor.py