2021

Multidimensional Poverty Predictor

MEMORY

DAVID VALERO SANCHO

Multidimensional Poverty Predictor

Introduction

This project applies machine learning to the study of poverty. Using the EU-SILC 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 sorted by stage. Reference to the code notebooks is given throughout the report.

The following sections decribe in detail the work done in this project. The framework section explains the theoretical approach at the study of poverty taken here. This section is important to understand what exactly I am trying to measure and predict. The data process section is the bulk of the project and explains the work done in the notebooks, from the source data to the modeling. The app development sections talks about the poverty predictor app, which is the culmination of the project. I finish the document with some concluding thoughts.

Link to the streamlit app of this project here:

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

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.

Firstly, poverty lines tend to measure relative poverty in oppostion to absulute poverty. Relative poverty exist in context with its surroundings. Falling below a certain threshold might mean more in some regions than others. Purchasing power varies accodingly with the economic contrext of a given location even whithin the same country. Access to affordable housing for example is not the same in a big city than in a small village. When we generalise using variable that measure relative poverty we run into the risk of making unaccurate estimations. Absoulte poverty conversely remains the same anywhere.

Secondly households with the same income can have drastically different standards of living. Households may be larger and include economically dependent individuals like children or disabled people. They may have outstanding debts to pay. Also income stability can be fragile if employement in temporary. Or it might just be inconsistent as in the case of freelance workers. Lastly 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.

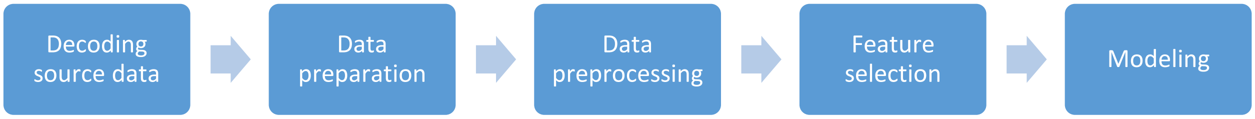
Researchers generally look at the multidimensional nature of poverty by looking at indicators that directly measure living conditions. They focus on access to basic consumption elements or evidence of economic difficulties. Measurements taken with multidimensional indicators are better at reflecting the depravation suffered by households. This is the reason why material deprivation or severe material deprivation are usually the umbrella names for this type of indicators. They can also be used to make fairer comparisons as they are more succesful at measuring absolute poverty than poverty lines.

Income is still however very important. It is a crucial variable affecting poverty, maybe the most important one. However it should be taken into consideration along with many other potentially explanatory variables when predicting poverty. That is the aim of this project.

This project uses the Eurostat multidimensional material deprivation indicator as the target variable for poverty. It is a complex indicator built from other indicators in order to capture poverty as accurately as possible. The methodology for this variable is explained below.

Data process

The methodological data process follows five main stages:



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.

Another very important characteristic of the source data is the ponderation weights assigned to each household. The sample is stratified and any statistical information obtained through this data needs to take weighting into account. This added some challenge to the code.

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 EU-SILC 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 explain 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, starting with the target variable.

-Material deprivation

Material deprivation is the target variable in this project. The Eurostat methodology for its material depravation variable is a multidimensional combination of another 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. Fourtunately this variable is available in the source data so I did not need to write the code to compute it.

-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 limitations. 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 assigns anyone born before 1932 with '1932' as their birth year regardless if they were born earlier. The possibility of excluding the 86 plus years old for this reason 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 material deprivation.

Growing old is indeed 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. At the same 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 multidimensional yes or no health 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.

Weigthning

As mentioned above the EU-SILC assigns ponderation weights to every household. They need to be taken into account otherwise any statistical measure obtained with the data will be biased. I resolve this problem by pre-defining the following functions:

- weighted_freq(df, cat_column): This function takes a dataframe and a categorical column inside the dataframe. It multiplies each row by a factor proportional to its ponderation weight. It returns frequency dictionary.

- weighted_cat(df, cat_column): This function takes a dataframe and a categorical column inside the dataframe. It multiplies each row by a factor proportional to its ponderation weight. It returns a weighted transformation of the categorical column.

- weighted_num(df, num_column): This function takes a dataframe and a numerical column inside the dataframe. It multiplies each row by a factor proportional to its ponderation weight. It returns a weighted transformation of the numerical column.

- weighted_df(df): This function takes a dataframe. It multiplies each row by a factor proportional to its ponderation weight. It returns a weighted transformation of the dataframe.

To see the code in detail refer to the notebook my_functions. These weigthing functions allow me to perform accuarate visualizations of the data, statistical tests and apply machine learning techniques.

Visualizations

I pre-define specific functions to aid me with my data exploration and preparation process. They use matplotlib.pyplot and seaborn plot functions in combination with the above weighting functions and other features. In notebook 3 I make extensive use of these visualization functions.

- barplot (x, data, legend): Takes a column name and a dataframe. It returns a weighted barplot with the frequency for each category on the column. The legend functionality allows to add a small legend on the side of the plot. The purpose of this function is to analyse the distribution of categorical variables.

- bardiagram (x, y, data, legend): Takes two column names and a dataframe. It returns as many weighted barplots as categories in column 'x' with the frequency for each category in the column 'y'. The legend functionality allows to add a small legend right on the side of the last plotted subplot. The ytick range is the same for every subplot to make comparations easier. It adjust automatically to the highest category variable in any of the subplots. It also adjusts its length to the number of subplots it needs to plot. The purpose of this function is to compare the relationship between two categorical variables.

- histplot (x, data, bins): Takes a column name and a dataframe. It returns a weighted histogram with the density for each specified number of bins in the value range of the column. The number of bins can be controled with the bins funcionality. The purpose of this function is to analize the distribution of numerical variables.

- scatterplot (x, y, data, bins): Takes two column names and a dataframe. It returns a weighted scatterplot of the two columns. The purpose of this function is to analize the the relationship between two numerical variables. It is not used in the repo notebooks but is availabe if anyone replicating this work wants to use it. It was helpful during the data praparation trials.

- boxplot (x, y, data, bins): Takes two column names and a dataframe. It returns a weighted boxplot of the 'y' column distribuiton for each of the 'x' column categories. The purpose of this function is to analyze the relationship between categorical and numerical variables.

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)
```

Before modeling there is some preprocessing to be done. The code for both data preprocessing and feature selection can be found in the notebook 4.

First I deal with the numerical columns. Scaling numerical variables into a similar range is a common transformation in machine learning preprocessing. Unscaled variables give worse test results than scaled variables. I decide on the sklearn MinMaxScaler method for scaling. This method compresses the range of a variable between 0 and 1.

All columns need to have integers for modeling and at the moment this dataset is mostly made of string columns. The strategy for the categorical variables is encoding them using the sklearn OneHotEncoder function. This function creates numerical 1 or 0 columns for each column category.

In the end I am left with 70+ variables with values between 0 and 1. Having all the values in the same range is beneficial for machine learning. However it would be desireble to reduce the number of variables, specially to get rid of the least important ones.

Feature selection

I use univariate statistical tests for feature selection. This method examines each feature individually to determine the strength of the relationship of the feature with the target variable. 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.

I end up with the following 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 equally among 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	0.75	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.

App deployement

The aim of the app is actually two fold.

This project gathered a lot of data relevant to the study of povery in general and in Spain in particular. This data can be useful to researchers and public managers interested in the variables that affect poverty. The app that I have built can serve as a quick tool for statistical queries. It allows the user to interact with the data in a fast way, exploring the relationships of the variables with each other, filtering, etc. Therefore its first aim is to serve as an useful statistical tool for visualization and statistics.

The user can also interact with the model. Changing the parameters at will to see is a certain combination results in a prediction of poverty or not.

The app was developed with the streamlit library. The script is availabe at poverty_predictor.py. It can be run from the command line with ‘streamlit run poverty_predictor.py’.

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.