Warsaw School of Economics

Master's Degree: Advanced Analytics – Big Data

Logistic Regression with SAS 223481-1380

PROJECT 3: MULTINOMIAL LOGISTIC REGRESSION MODEL

ANALYSIS OF THE FACTORS INFLUENCING ON THE SOURCES OF THE GROSS NATIONAL INCOME IN SPAIN

Authors:

José Caloca 110558

Erick Moreno

Professor:

Dr. Adam Korczyński

Group C members:

José Caloca 110558

Erick Moreno

Marek Wrucha 75664

Gayda Alsharabi 108560

Diego Rodriguez

Barbara Śmiech

Warsaw, 06 June 2021.

Introduction

In macroeconomics, there is a fundamental principle that explains that production equals expenses, and income. The Gross Domestic Product (GDP) is a measure of all the goods and services produced within a country in a specific timespan. The GDP can be calculated following 3 different methods, one of these methods is the income approach. The income approach is based on the addition of all the incomes within a country. More formally, the Gross Domestic Income (GDI), also known as GDP based on the income approach is defined as follows (Landefeld et. al, 2008):

$$\boldsymbol{GDI = Total\ National\ Income + Sales\ Taxes + Depreciation + Net\ Foreign\ Income}$$

The Gross National Income (GNI) is explained by the remuneration of the factors of production. And it is defined as follows (Piana, 2010):

$$\boldsymbol{GNI = Labour\ income + Capital\ Income + State\ Revenue + Amortizacion}$$

During the past decades, the share of labour income has been decreasing while the capital income share is increasing at the same speed. On the other hand, it is well-known that Spain has one of the largest unemployment rates in Europe and this problem generates a fiscal burden on the public budget and debt. Due to this, many resources are dedicated to grants and subsidies for the unemployed. This problem is exacerbated in the young population (Garcia, 2011). The goal of this project is to analyse the factors influencing on the risk of having a type of source of income for a Spanish household based on the age of the individuals.

To this, our **research question** is the following: What is the chance that a person in Spain will have "Grants" as main source of income and not "labour income" in a group of young people.

In order to do that we will build a multinomial logistic regression model using data from the European Social Survey (ESS) Round 9. For the model, **our target variable** is the main source of household income. Our **explanatory variables** are selected based on similar features used in the literature when assessing behavioural effects on the gross national income (van den Bergh, 2008), we will then use: age, gender, number of people living regularly as member of household, years of full-time education completed, and main activity over the last 7 days.

Based on the model description, our **hypothesis** is the following: Young people are more likely to have Grants as a main source among all the possible sources.

As for the code, first we do some data cleaning, then we study the distribution of the features in our dataset, later we implement our multinomial logistic regression model, and finally, as an **innovative aspect** we will compare performance of the logistic regression model with a Random Forest model with tuned hyperparameters in Python.

Descriptive Statistics

It is important to remark once again that our data comes from the ESS, and as a behavioural survey to individuals we can expect observations such as: "not applicable", "refusal", "don't know" and "no answer". We realise that these values represent between 1 – 5 % of the frequency in each variable. Therefore, in order to show the distribution of our variables, we can drop these meaningless observations. This belongs to our pre-processing section in our code annex.

The following table provides a big picture of our dataset. We can observe that our dataset does not have missing values.

Variable	Label	N	N Miss
hincsrca	Main source of household income	1566	0
gndr	Gender	1566	0
mnactic	Main activity, last 7 days. All respondents. Post coded	1566	0
hhmmb	Number of people living regularly as member of household	1566	0
eduyrs	Years of full-time education completed	1566	0
agea	Age of respondent, calculated	1566	0

Table 1: Dataset Missing Values Analysis

We proceed to do a discriminatory performance analysis to all our variables.

Target variable: Main source of household income

Our target variable has 8 classes. These 8 classes are grouped into 4 categories following the following criteria:

1. Labour income: Income as a form of compensation to employees.
2. Capital income: Income resulting from the profits of any entrepreneurial or financial activity.
3. Grants: Grants or subsidies received from the state.

4. Other income: Any other source of income (payments done "under the table", gifts, remittances, illegal activities, etc.).

Based on the above criteria our target variable is transformed as follows:

Main source of household income				
Variable: hincsrca	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Wages or salaries	915	58.43	915	58.43
Income from self-employment (excluding farming)	168	10.73	1083	69.16
Income from farming º	11	0.70	1094	69.86
Pensions	351	22.41	1445	92.27
Unemployment/redundancy benefit	29	1.85	1474	94.13
Any other social benefits or grants	40	2.55	1514	96.68
Income from investments, savings etc.	9	0.57	1523	97.25
Income from other sources	43	2.75	1566	100.00
Variable: y	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Labour Income	926	59.13	926	59.13
Capital Income	177	11.30	1103	70.43
Grants	420	26.82	1523	97.25
Other Income	43	2.75	1566	100.00

Table 2: Target variable frequency

We can appreciate that most of the observations are concentrated on the labour income class, this represents around 60% of the observations, a common value for developed economies, in the case of the G7, the average is 65% (Vermeiren, 2017). On the other hand, the second largest category is "Grants" which holds around 27% of the observation. This is a high and alarming value since it represents the scale of the burden on the public budget in Spain. Other income class represents around 3% of the observation, we infer that the underground economy in Spain is representative.

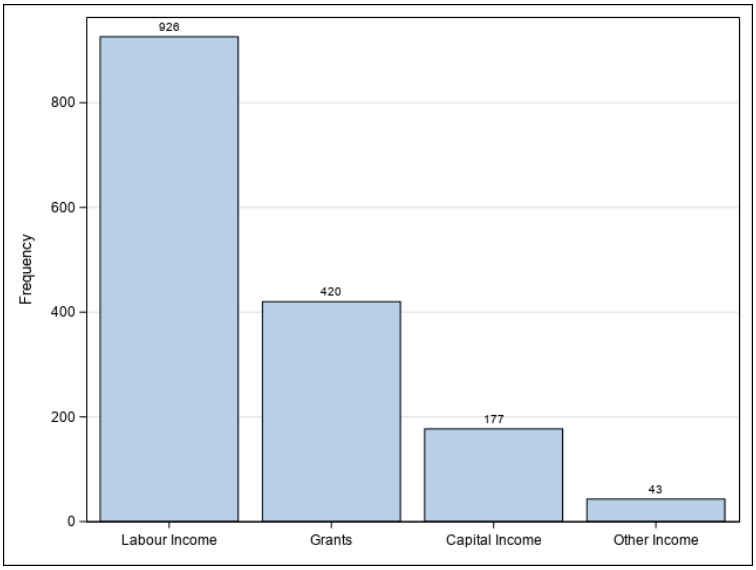


Figure 1: Main source of household income distribution

Now we can proceed with our discriminatory performance analysis. Each of the features will be analysed separately.

Gender:

First let us observe our first variable (Gender). The classes are balanced. Also, the distribution of the target variable by gender does not differ much per class, therefore, we can infer a priori that this a representative variable.

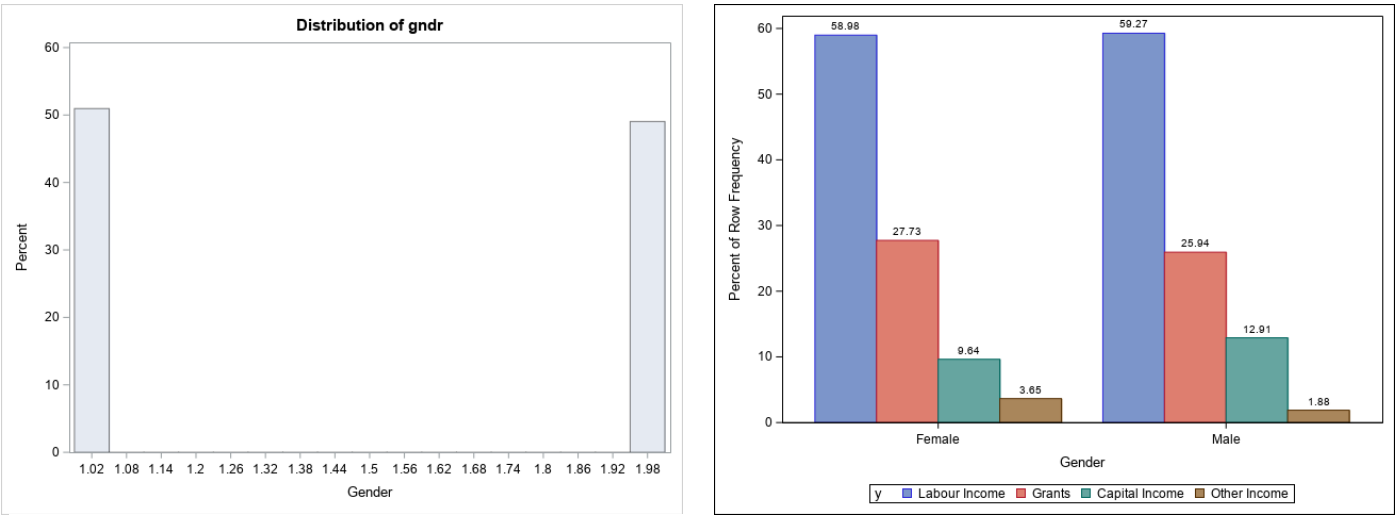


Figure 2: Sources of income distribution by gender

Main activity, last 7 days:

Regarding the second variable (Main activity, last 7 days) we can observe that the classes are not balanced and the "employed" class has more than 50% of the observations. The frequency of labour income is high in that which main source of income is: "Employed", "Other", and "Unemployed". Also, the frequency of Grants is the higher class when the main activity is "Inactive". We infer that this variable is not significant enough as it changes considerably among classes.

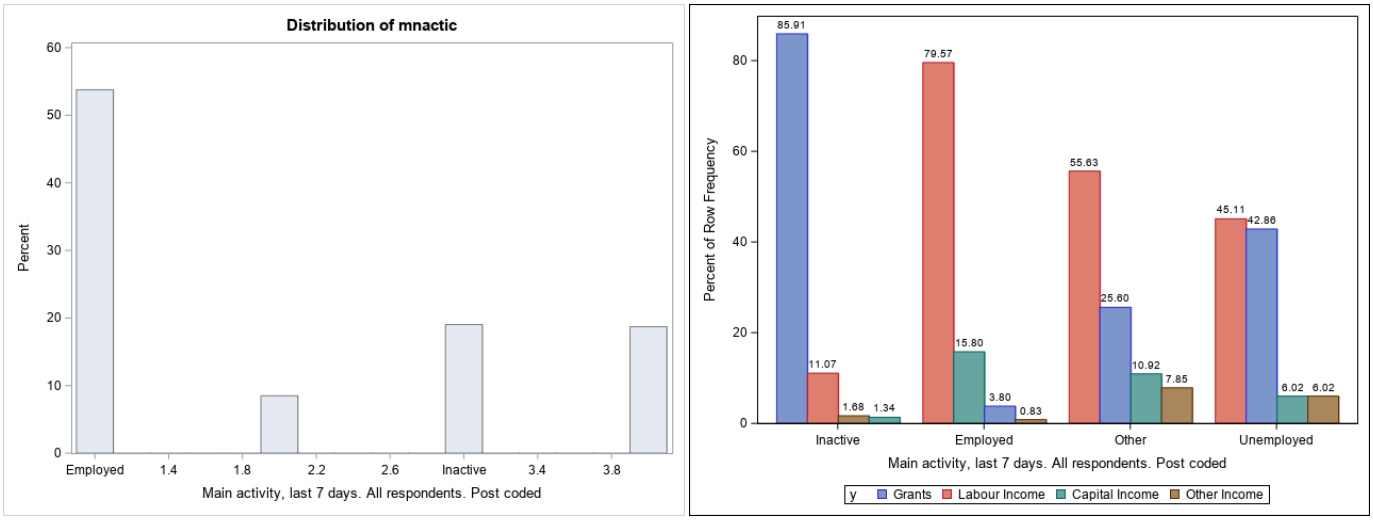


Figure 3: Sources of income distribution by main activity, last 7 days

Number of people living regularly as member of household:

When it comes to our third variable, we find that the ordinal distribution right skewed and imbalanced. Labour income is the category with higher frequency among all the classes. Also, Small families tend to have more than bigger families. Additionally, most of the families have between 1-4 members.

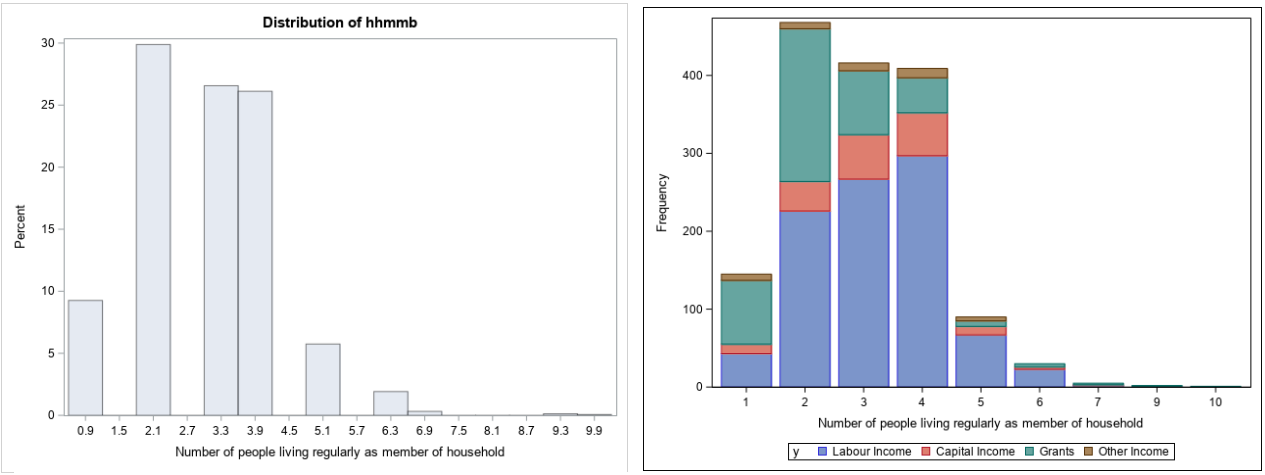


Figure 4: Sources of income distribution by number of people in a household

Years of full-time education completed:

For the fourth variable the distribution might look as normal, however it has many outliers that makes it right skewed. For those with lower education, mostly the main source of income are "grants". We find that capital income is present regardless the educational background. The main source of income for people with mid and high education is labour income. Lastly, people with higher education don't receive many grants and other groups.

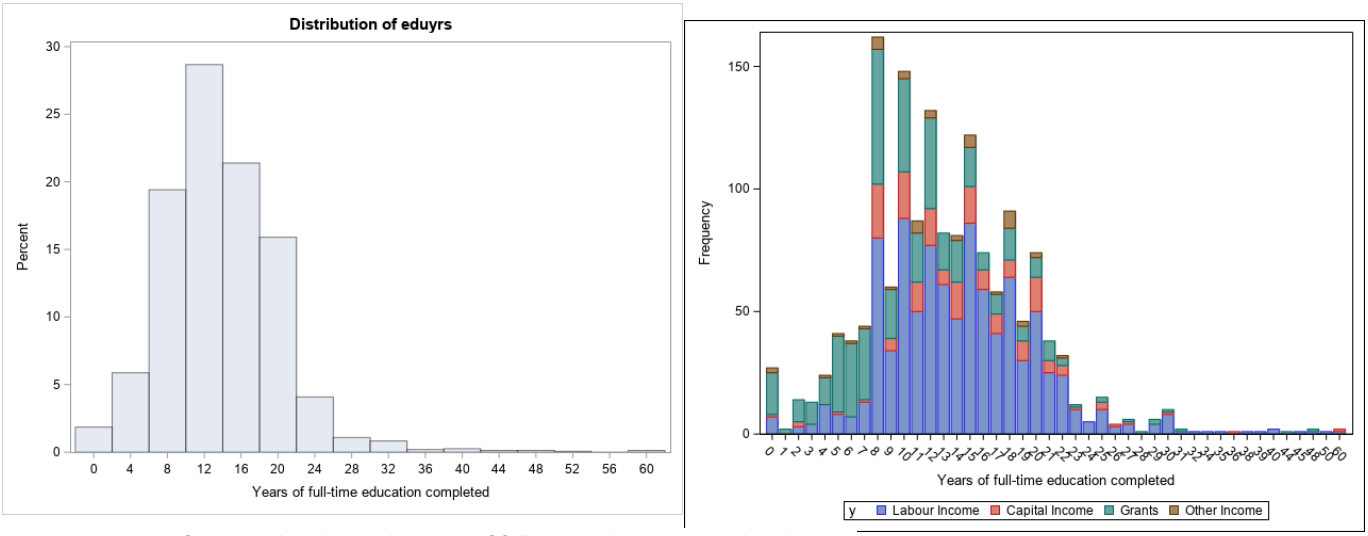
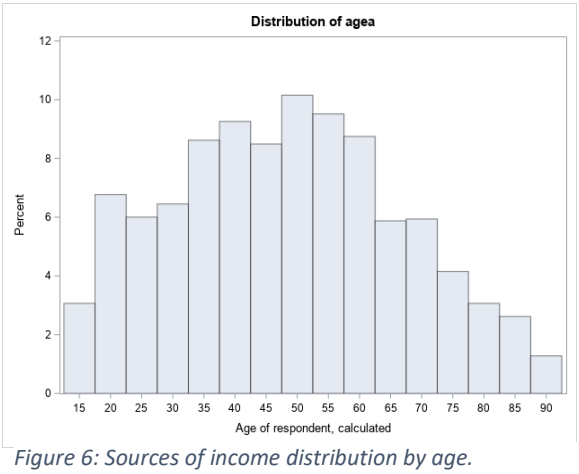


Figure 5:Sources of income distribution by years of full-time education completed.

Age of respondent, calculated.

Our last variable (Age) has values of Kurtosis and Skewness close to zero in the range of +/- 2, therefore we can assume normality (Gravetter, et. al 2014 and George, et. al. 2010). For people in the working age, labour income is the main source of income along with the capital income. Old people however tend to have more Grants and capital income.



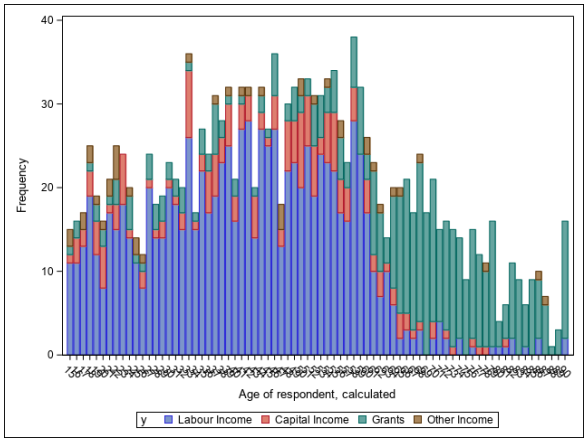


Figure 6: Sources of income distribution by age.

Substantive Analysis

Multicollianity assessment

In the above correlation matrix between the numerical variables, we don't find any alarming correlation between them.

Pearson Correlation Coefficients

	agea	hhmmb	eduyrs
agea Age of respondent, calculated	1.00	-0.40	-0.31
hhmmb Number of people living regularly as member of household	-0.40	1.00	0.05
eduyrs Years of full-time education completed	-0.31	0.05	1.00

When running a Chi-Square test between our categorical variables (gender, and main activity) we find that these variables are dependent. Which means that we have risk of collinearity between these two variables and may bias the estimations.

Statistic	DF	Value	Prob
Chi-Square	3	61.2201	<.0001

We run proc logistic to estimate a multinomial logistic regression model. To understand better, multinomial logistic regression is used to modelling nominal outcome variables where the log odds of the mentioned outcome are modelled as a liner combination of

predictor variables. We choose as reference class in the categorical variables gender and main activity, those with larger frequency in order to balance the dataset and avoid any unwanted bias which is considered as a good practice (Grace-Martin, 2008). The reference in our target variable is “labour income”, in gender “Male”, and main activity is “Employed”.

In the table below, we could observe the general information of the model with 4 response Level and for scoring, the optimisation technique used for obtaining the beta coefficients was Newton-Raphson. We specified the baseline category for **gndr** using ref='Male' and reference group **mantic** ref='Employed'. **Param=ref** tells SAS to use dummy coding rather than effect coding for variables.

We can observe below that all 1566 observations from our data were used for the analysis. Also, we could observe our response outcome variable Y values with respective frequency observations.

Table 3: Model Information

Number of Observations Read	1566
Number of Observations Used	1566

Table 4:Number of Observations

Response Profile		
Ordered Value	**y**	**Total Frequency**
1	Capital Income	177
2	Grants	420
3	Labour Income	926
4	Other Income	43

Table 5: Response Profile

The tables 6 describes and tests the overall fit of the model. Since the -2 Log L and other statistics decrease when augmenting with more variables, we infer that the model performs well with the current training features.

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	3165.455	2186.874
SC	3181.524	2315.425
-2 Log L	3159.455	2138.874

Table 6: Model fit Statistics

We can observe below that the likehood ratio chi-square of 1020.58 with p-value of 0.0001 explains that our model as whole fits significantly good with the more explanatory variables, than with a single one.

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	1020.5812	21	<.0001
Score	951.9919	21	<.0001
Wald	512.9724	21	<.0001

Table 7: Testing Global Null Hypothesis

The table 8 describes the hypothesis tests for all variable in our model individually. The chi-square tests statistics concluding that highlighted variables are statistically significant since p-vales are lower than 0.05. By rejecting the accepting the null hypothesis in the case of the variable "years of education, we infer that the coefficient is very small relative to its standard error and does not impact significantly on the source of income in Spain (target variable).

Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
gndr	3	7.2902	0.0632
mnactic	9	285.4381	<.0001
hhmmb	3	23.8461	<.0001
agea	3	67.0737	<.0001
eduyrs	3	3.1939	0.3627

Table 8: Analysis of Effects

Hosmer and Lemeshow test describes the observed event rates that match expected event by subgroups. Also, we can observe by HL goodness fit test results where the p-value is 0.1898, failing to reject the null hypothesis meaning that there are not enough evidence to infer that the model is poor (low predictive power).

Partition for the Hosmer and Lemeshow Test									
Group	**Total**	**Observed y = Capital Income**	**Observed y = Grants**	**Observed y = Labour Income**	**Observed y = Other Income**	**Expected y = Capital Income**	**Expected y = Grants**	**Expected y = Labour Income**	**Expected y = Other Income**
1	157	16	2	136	3	18.26	1.90	135.5	1.34
2	157	16	5	134	2	21.88	3.00	130.9	1.24
3	157	28	1	127	1	24.50	4.12	127.1	1.33
4	158	26	6	124	2	26.26	5.85	124.1	1.79
5	157	33	3	117	4	26.68	8.83	118.1	3.40
6	157	23	19	109	6	24.61	14.24	111.0	7.11
7	157	23	40	88	6	19.24	29.27	97.29	11.19
8	157	8	76	63	10	10.47	85.16	52.29	9.08
9	157	3	130	18	6	3.81	127.7	21.02	4.51
10	152	1	138	10	3	1.29	140.0	8.73	2.02

Table 9: Hosmer and Lemeshow Test

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
29.8479	24	0.1898

Analysis of Maximum Likelihood Estimates							
Parameter		y	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		Capital Income	1	-2.1926	0.4782	21.0241	<.0001
Intercept		Grants	1	-4.0706	0.5802	49.2284	<.0001
Intercept		Other Income	1	-2.7254	1.0776	6.3960	0.0114
gndr	Female	Capital Income	1	-0.3522	0.1713	4.2301	0.0397
gndr	Female	Grants	1	-0.2927	0.1969	2.2098	0.1371
gndr	Female	Other Income	1	0.3106	0.3528	0.7751	0.3786
mnactic	Inactive	Capital Income	1	-0.8589	0.5562	2.3846	0.1225
mnactic	Inactive	Grants	1	3.7306	0.2857	170.5007	<.0001
mnactic	Inactive	Other Income	1	2.4364	0.6615	13.5667	0.0002
mnactic	Other	Capital Income	1	0.3016	0.2401	1.5782	0.2090
mnactic	Other	Grants	1	2.8184	0.2703	108.7105	<.0001
mnactic	Other	Other Income	1	2.8021	0.4755	34.7326	<.0001
mnactic	Unemployed	Capital Income	1	-0.2474	0.3924	0.3974	0.5285
mnactic	Unemployed	Grants	1	3.3783	0.2827	142.8055	<.0001
mnactic	Unemployed	Other Income	1	2.5959	0.5444	22.7405	<.0001
hhmmb		Capital Income	1	-0.0226	0.0731	0.0955	0.7573
hhmmb		Grants	1	-0.3740	0.0845	19.5765	<.0001
hhmmb		Other Income	1	-0.4624	0.1678	7.5924	0.0059
agea		Capital Income	1	0.0165	0.00638	6.7034	0.0096
agea		Grants	1	0.0526	0.00668	61.9077	<.0001
agea		Other Income	1	-0.00352	0.0115	0.0930	0.7604
eduyrs		Capital Income	1	0.00314	0.0130	0.0583	0.8091
eduyrs		Grants	1	-0.0233	0.0159	2.1389	0.1436
eduyrs		Other Income	1	-0.0375	0.0333	1.2672	0.2603

Table 11: Analysis of Maximum Likehood Estimates

Analysis of Maximum Likelihood is used to check the significance of each category. As mentioned before if a particular category has a p-value < 0.05 is statistically significant. Categories which are statistically significant:

- **gndr (Female - Capital Income)**
- **mnactic (Inactive - Grants)**
- **mnactic (Inactive – Other Income)**
- **mnactic (Other – Grants)**
- **mnactic (Other – Other Income)**
- **mnactic (Unemployed - Grants)**
- **mnactic (Unemployed - Other Income)**
- **hhmmb (Grants)**
- **hhmmb (Other Income)**
- **agea (Capital Income)**
- **agea (Grants)**

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals					
Effect	y	Unit	Estimate	95% Confidence Limits	
gndr Female vs Male	Capital Income	1.0000	0.703	0.501	0.982
gndr Female vs Male	Grants	1.0000	0.746	0.506	1.097
gndr Female vs Male	Other Income	1.0000	1.364	0.689	2.773
mnactic Inactive vs Employed	Capital Income	1.0000	0.424	0.121	1.136
mnactic Inactive vs Employed	Grants	1.0000	41.705	24.166	74.193
mnactic Inactive vs Employed	Other Income	1.0000	11.432	2.977	41.746
mnactic Other vs Employed	Capital Income	1.0000	1.352	0.835	2.147
mnactic Other vs Employed	Grants	1.0000	16.749	9.957	28.787
mnactic Other vs Employed	Other Income	1.0000	16.480	6.770	44.709
mnactic Unemployed vs Employed	Capital Income	1.0000	0.781	0.336	1.596
mnactic Unemployed vs Employed	Grants	1.0000	29.322	17.011	51.627
mnactic Unemployed vs Employed	Other Income	1.0000	13.408	4.579	40.177
hhmmb	Capital Income	1.0000	0.978	0.846	1.127
hhmmb	Grants	1.0000	0.688	0.581	0.810
hhmmb	Other Income	1.0000	0.630	0.448	0.865
agea	Capital Income	1.0000	1.017	1.004	1.029
agea	Grants	1.0000	1.054	1.040	1.068
agea	Other Income	1.0000	0.996	0.974	1.019
eduyrs	Capital Income	1.0000	1.003	0.977	1.028
eduyrs	Grants	1.0000	0.977	0.946	1.007

Table 10: Odds Ratio Estimates

The odds ratio estimates analysis describes variables that have p-value < 0.05. In the table 11, the analysis describes the significance of our variables. The most relevant variables shows the following:

- gndr Female is 30% less likely to live on bases of Grants than Male.
- 42 times likely for mnactic Inactive to live on bases of Grants than Employed.
- 17 times likely for mnactic other to live on bases of Grants than Employed.
- 29 times likely for mnactic Unemployed to live on bases of Grants than Employed
- hhmmb is 30% less likely to live on bases of Grants and 37% less likely on Other income.
- Increasing age by 1 year rises the likelihood of obtaining Capital Income and Grants by around 2% and 5% respectively.

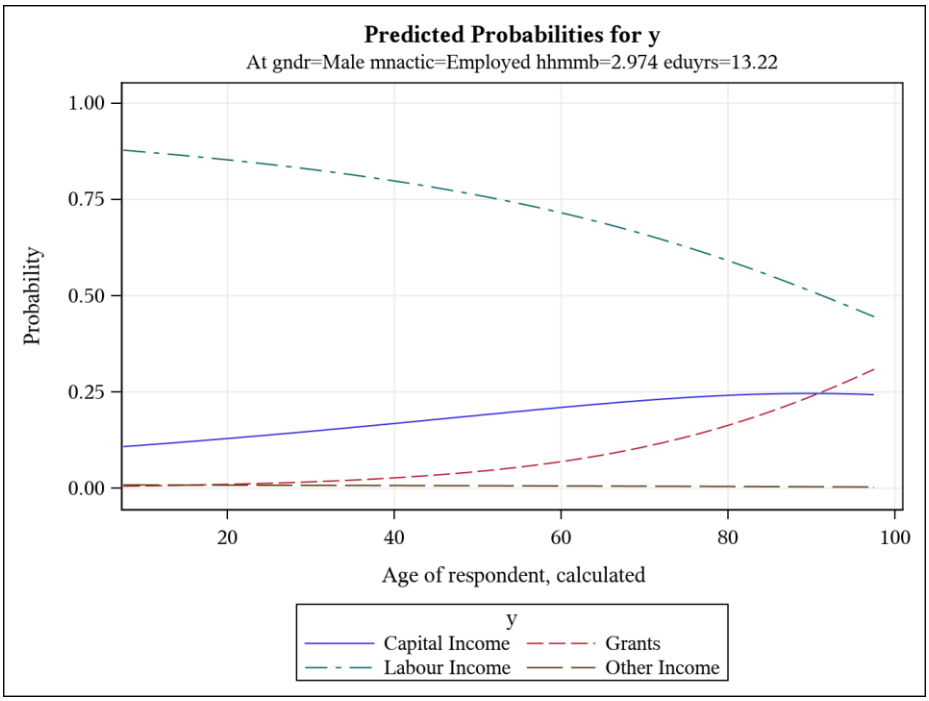


Figure 7: Predicted Probabilities.

From the above graph we infer that young and middle-aged people are more likely to have labour income and the main source. However, old people are more likely to have as main source of income grants and capital income.

Comparison Log Reg vs Random Forest Log Reg

Value of "mnactic" feature shows the biggest impact for both models:

- Unemployed people are 29.322 times more likely to have grants as their main source of income than employed.
- Inactive people are 41.705 times more likely to have grants as their main source of income than employed.
- People that are in a category of "Other" are around 16 times more likely than employed to have either other income or grants as a main source of income.

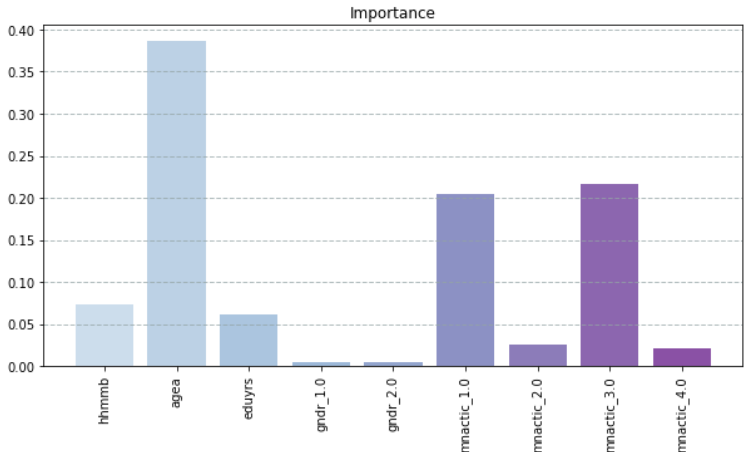


Figure 8: Importance RF

In the importance bar chart for Random Forest Classifier model, we can see that mnactic_1 (employed) and mnactic_3 (inactive) features have over 0.2 score. Taking into consideration that all four columns (mnactic_1, mnactic_2, mnactic_3, mnactic_4) are in fact one category column after one-hot encoding we can assume that it is overall importance score is above 0.4 which is the highest for all given features.

There is an interesting difference between Logistic Regression and Random Forest for this feature.

In table 8, analysis of effects also showed that age is highly statistically significant because it has p-value lower than 0.05 Also, the analysis of effects showed that for Logistic Regression hhmmb variable is one of the most significant ones (with p-value as low as mnactic variable and age). This is different for Random Forest Classifier where hhmmb importance value is low (similarly to years of finished education which importance is low for both models).

Conclusion

The aim of the project was to discover what is the chance that a person in Spain will have "Grants" as main source of income and not "labour income" in a group of young people. In the final analysis, we can observe in figure 7 that the probability of a person to live on bases of "labour income" decreases as the age of the person increases.

However, we found out that as age increases, the probability of a person to live on bases of "grants" and "capital income" as main source of income increases. In other hand, a person that lives on bases on other income does not varies as the age of persons increases.

Consequently, we rejected our hypothesis as we were not able to prove that young people are more likely to have Grants as a main source among all the possible sources as we were not able to prove it.

Furthermore, we found that age variable is significant for both models. For Random Forest Classifier in the importance chart, we can observe that age has the highest score. We were able find that from both "Logistic Regression Model and Random Forest" gender has the smallest impact on the results. It is visible in distribution, both woman and man have similar probability of having each option as a main source of income.

Bibliography

Landefeld, J., Seskin, E., Fraumeni, B. (2008). *Taking the Pulse of the Economy: Measuring GDP*. Journal of Economic Perspectives, 22 (2): 193-216.DOI: 10.1257/jep.22.2.193

Piana, V. (2001) Gross Domestic Product: a key concept in Economics. Economics Web Institute. Available online: http://www.economicswebinstitute.org/glossary/gdp.htm

Garcia, J. R. (2011). *Youth unemployment in Spain: causes and solutions.* Working Papers 1131, BBVA Bank, Economic Research Department.

Van den Bergh, J. (2009) *The GDP paradox.* Journal of Economic Psychology. Volume 30, Issue 2, Pages 117-135. DOI: https://doi.org/10.1016/j.joep.2008.12.001.

Vermeiren, Mattias. (2017). Global Macroeconomic Imbalances after the Crisis: From the Great Moderation to Secular Stagnation. The International Spectator. 52. 10.1080/03932729.2017.1383758.

George, D. and Mallery, M. (2010) SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 Update, 10th Edition, Pearson, Boston.

Gravetter, F., & Wallnau, L. (2014). Essentials of statistics for the behavioral sciences (8th ed.). Belmont, CA: Wadsworth.

Grace-Martin, K. (2008) *Strategies for Choosing the Reference Category in Dummy Coding.* The analysis factor. Available online: https://www.theanalysisfactor.com/strategies-dummy-coding/

List of Tables

List of Figures

Code Annex

```
*/////////////////////////////;
* LOAD DATASET
*/////////////////////////////;

/* Create a library - get access to
the data;*/
libname b "C:\Users\Jose
Caloca\Desktop";

/* set formats;*/

PROC FORMAT lib=work;
value sex
1 = 'Male'
2 = 'Female'
9 = 'No answer' .d = 'No answer';
value sourceinc
1 = 'Wages or salaries'
2 = 'Income from self-employment
(excluding farming)'
3 = 'Income from farming'  °
4 = 'Pensions'
5 = 'Unemployment/redundancy
benefit'
6 = 'Any other social benefits or
grants'
7 = 'Income from investments,
savings etc.'
8 = 'Income from other sources'
77 = 'Refusal' .b = 'Refusal'
88 = 'Don''t know' .c = 'Don''t
know'
99 = 'No answer' .d = 'No answer' ;
value mainactivity
1 = 'Employed'
2 = 'Unemployed'
3 = 'Inactive'
4 = 'Other';
value peoplelivinghouse
77 = 'Refusal' .b = 'Refusal'
88 = 'Don''t know' .c = 'Don''t
know'
99 = 'No answer' .d = 'No answer' ;
value yearsedu
77 = 'Refusal' .b = 'Refusal'
88 = 'Don''t know' .c = 'Don''t
know'
99 = 'No answer' .d = 'No answer' ;
value age
999 = 'Not available' .d = 'Not
available' ;
value national_income
1 = 'Labour Income'
2 = 'Capital Income'
3 = 'Grants'
4 = 'Other Income';
run;

/* load dataset;*/

data ess;
set b.ess9e03_1;
format hincsrca sourceinc. gndr sex.
mnactic  hhmmb peoplelivinghouse.
eduyrs yearsedu. agea age.;
keep hincsrca gndr mnactic hhmmb
eduyrs agea;
where cntry = 'ES';
run;


*/////////////////////////////;
* DROPPING: not applicable, refusal,
don't know
and no answer ;
*/////////////////////////////;

*Check the distribution of the
response variable PRIOR MODIFYING;

proc freq data=ess;
    table hincsrca;
run;

data ess_01;
set ess;
if hincsrca in (77,88,99) then delete;
if gndr = 9 then delete;
if mnactic in (66,77,88,99) then
delete;
if hhmmb in (77,88,99) then delete;
if eduyrs in (77,88,99) then delete;
if agea = 999 then delete;
run;

* Check missing values;

proc means data=ess_01 n nmiss;
var hincsrca gndr mnactic hhmmb eduyrs
agea;
run;

*Check the distribution of the response
variable AFTER MODIFYING;

proc freq data=ess_01;
    table hincsrca;
run;
*/////////////////////////////;
* TARGET VARIABLE PREPARATION ;
*/////////////////////////////;


/* relabel hincsrca, mnactic*/
data ess_02 (drop=hincsrca);
set ess_01;
format y national_income. mnactic
mainactivity.;
if hincsrca in (1, 3) then y=1;
else if hincsrca in (2, 7) then y=2;
else if hincsrca in (4, 5, 6) then y=3;
else if hincsrca=8 then y=4;
else y=.;
if mnactic in (1, 7) then mnactic=1;
else if mnactic in (3, 4) then
mnactic=2;
else if mnactic in (5, 6) then
mnactic=3;
else mnactic=4;
run;

*Check the distribution of the response
variable AFTER MODIFYING;

proc freq data=ess_02;
    table y;
run;

*/////////////////////////////;
* EXPLORATORY DATA ANALYSIS ;
*/////////////////////////////;
/*DISCRIMINATORY PERFORMANCE ANALYSIS;

/*Folder to save the plots*/
%let graphs = C:\Users\Jose
Caloca\Desktop;

/*Bar Plot of the hincsrca variable */
ods listing gpath="&graphs";
ods graphics /
imagename="hincsrca_barplot"
imagefmt=png;

proc SGPLOT data = ess_01;
vbar hincsrca / datalabel
categoryorder=respdesc;
xaxis display=(nolabel);
yaxis grid ;
run;
quit;
ods close;

/*Bar Plot of the Target variable */
ods listing gpath="&graphs";
ods graphics /
imagename="source_of_income_barplot"
imagefmt=png;

proc SGPLOT data = ess_02;
vbar y / datalabel
categoryorder=respdesc;
xaxis display=(nolabel);
yaxis grid ;
run;
quit;
ods close;

/* Categorical predictors;*/

%macro Frequency(Var);
    proc freq data=ess_02;
        tables &Var.*y;
        ods output CrossTabFreqs=pct01;
    run;
    ods listing gpath="&graphs";
    ods graphics /
    imagename="&Var._barplot"
    imagefmt=png;
    proc sgplot
data=pct01(where=(^missing(RowPercent)));
        vbar &Var. / group=y
groupdisplay=cluster response=RowPercent
datalabel categoryorder=respdesc;
    run;
%mend;
%Frequency(gndr);
%Frequency(mnactic);
/* Continuous predictors;*/

%macro Continuous(Var);
    ods listing gpath="&graphs";
    ods graphics /
    imagename="&Var._barplot"
    imagefmt=png;
    proc sgplot data=ess_02;
    vbar &Var. / group=y;
    run;
%mend;
%Continuous(hincsrca); *target variable;
%Continuous(hhmmb);
%Continuous(eduyrs);
%Continuous(agea);

/******* DISTRIBUTION ANALYSIS;

/* Statistical outputs for all varables
*/
proc univariate data=ess_02 plots;
var gndr mnactic hhmmb eduyrs agea;
histogram;
run;
*/////////////////////////////;
* COLLINEARITY ;
*/////////////////////////////;

*correlation matrix numerical variables;
proc corr data=ess_02;
 var agea hhmmb eduyrs;
run;
*chi-square test categorical variables;
proc freq data=ess_02;
tables gndr*mnactic/ chisq;
run;

*/////////////////////////////;
* MODELLING ;
*/////////////////////////////;

proc logistic data=ess_02;
class  gndr (param=ref ref='Male')
mnactic (param=ref ref='Employed');
model y (ref='Labour Income') = agea gndr
mnactic hhmmb eduyrs /
link=glogit expb rsquare aggregate
scale=none;
output out=out predicted=p;
run;

proc sgplot data=out;
scatter x=agea y=p / group=_LEVEL_;
run;
```

```
#PYTHON RF CODE

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from pprint import pprint
from sklearn.model_selection import RandomizedSearchCV
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import accuracy_score
from sklearn import metrics
import matplotlib.pyplot as plt

dataset = pd.read_sas("ess_02.sas7bdat")
dataset.info()

dataset['gndr'] = dataset.gndr.astype('object')
dataset['mnactic'] = dataset.mnactic.astype('object')

#First: we create two data sets for numeric and non-
numeric data
numerical = dataset.select_dtypes(exclude=['object'])
categorical =
dataset.select_dtypes(include=['object'])

#Second: One-hot encode the non-numeric columns
z = pd.get_dummies(categorical)

#Third: Union the one-hot encoded columns to the
numeric ones
df = pd.concat([numerical, onehot], axis=1)

# We create the X and y data sets
X = df.loc[ : , df.columns != 'y']
y = df[['y']]

# Create training, evaluation and test sets
X_train, test_X, y_train, test_y = train_test_split(X,
y, test_size=.3, random_state=123)

# percentage of the classes in the training set
round(y_train['y'].value_counts()*100/len(y_train['y']
), 2)

# Number of trees in random forest
n_estimators = [int(x) for x in np.linspace(start =
200, stop = 2000, num = 10)]
# Number of features to consider at every split
max_features = ['auto', 'sqrt']
# Maximum number of levels in tree
max_depth = [int(x) for x in np.linspace(2, 10, num =
8)]
max_depth.append(None)
# Minimum number of samples required to split a node
min_samples_split = [2, 5, 10]
# Minimum number of samples required at each leaf node
min_samples_leaf = [1, 2, 4]
# Method of selecting samples for training each tree
bootstrap = [True, False]
# Create the random grid
random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf,
               'bootstrap': bootstrap}
pprint(random_grid)

# Use the random grid to search for best
hyperparameters
# First create the base model to tune
rf = RandomForestRegressor()
# Random search of parameters, using 3 fold
cross validation,
# search across 100 different combinations, and
use all available cores
rf_random = RandomizedSearchCV(estimator = rf,
param_distributions = random_grid, n_iter =
100, cv = 3, verbose=2, random_state=42, n_jobs
= -1)
# Fit the random search model
rf_random.fit(X_train, y_train)
#We can view the best parameters from fitting
the random search:
rf_random.best_params_
# we make predictions
best_random = rf_random.best_estimator_
predictions =
pd.DataFrame(best_random.predict(test_X))
#We calculate the AUC
fpr, tpr, thresholds =
metrics.roc_curve(test_y, predictions,
pos_label=3)
metrics.auc(fpr, tpr)
# get importance
importance = best_random.feature_importances_
# summarize feature importance
var_importance = pd.DataFrame({'col_name':
best_random.feature_importances_},
index=X_train.columns).sort_values(by='col_name
', ascending=False)
# plot feature importance
importance = pd.DataFrame({'col_name':
best_random.feature_importances_})
index = np.array(X_train.columns)
#index= np.arange(len(X_train.columns))
plt.figure(figsize=(10, 5))
colors = plt.cm.BuPu(np.linspace(0.2, 0.7,
len(importance)))
plt.xticks(rotation=90)
plt.bar(index, importance['col_name'],
color=colors)
plt.grid(color='#95a5a6', linestyle='--',
linewidth=1, axis='y', alpha=0.7)
plt.title('Importance')
plt.show()
```