

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  **Group C members:**  José Caloca 110558  Erick Moreno  Marek Wrucha 75664  Gayda Alsharabi 108560  Diego Rodriguez  Barbara Śmiech | **Professor:**  Dr. Adam Korczyński |

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):

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

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.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Figure 1: Main source of household income distribution

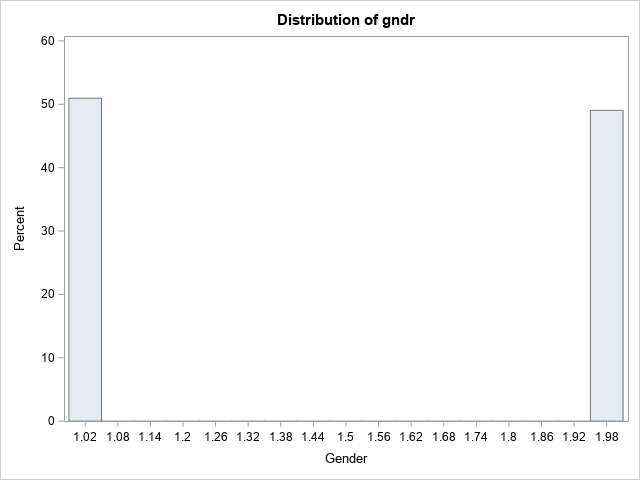
Now we can proceed with our discriminatory performance analysis. Each of the features will be analysed separately.

**Gender:**

Gráfico, Gráfico de barras

Descripción generada automáticamenteFirst 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:**

Gráfico, Gráfico de barras

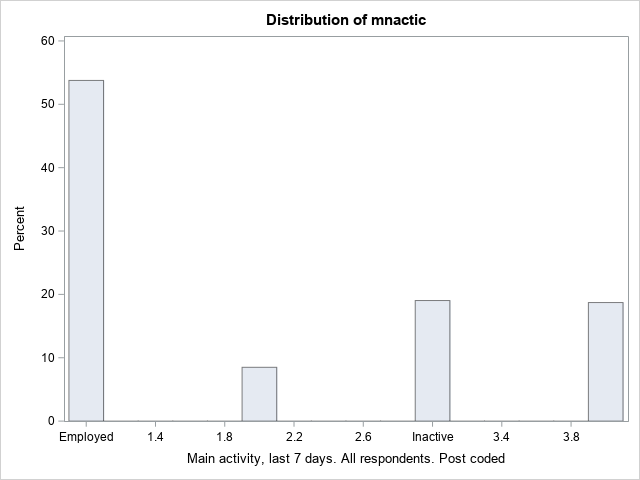
Descripción generada automáticamenteRegarding the second variable (Main activity, last 7 days) we can observe that the classes are not balanced and the “employed” class has more than 50% of the observations. The frequency of labour income is high in that which main source of income is: “Employed”, “Other”, 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:**

Gráfico, Gráfico de barras

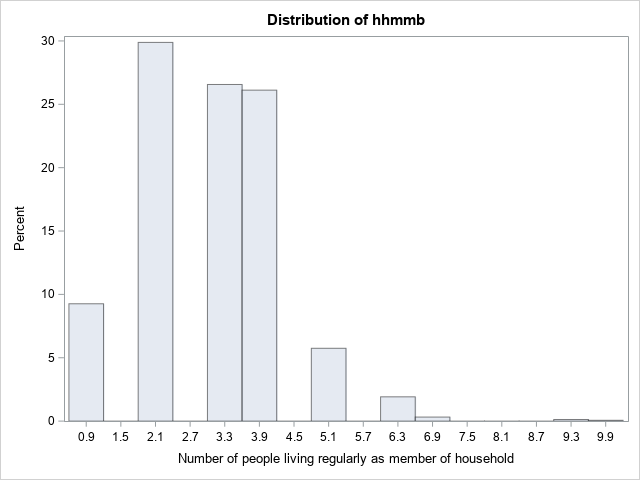
Descripción generada automáticamenteWhen it comes to our third variable, we find that the ordinal distribution right skewed and imbalanced. Labour income is the category with higher frequency among all the classes. Also, Small families tend to have more than bigger families. Additionally, most of the 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:**

Gráfico, Histograma

Descripción generada automáticamenteFor 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.

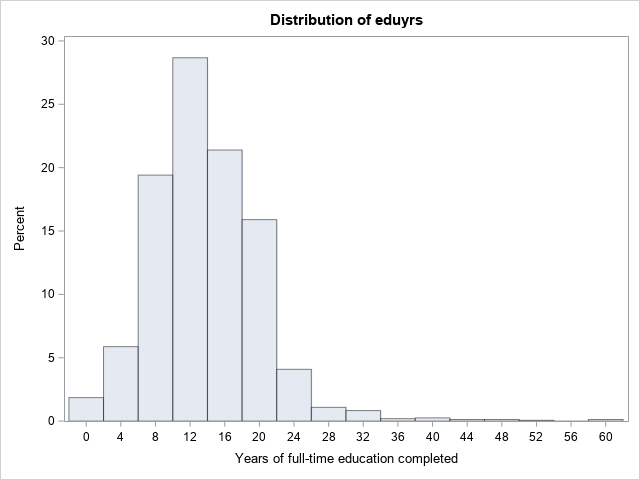


Figure 5:Sources of income distribution by years of full-time education completed.

**Age of respondent, calculated.**

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamenteOur 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.

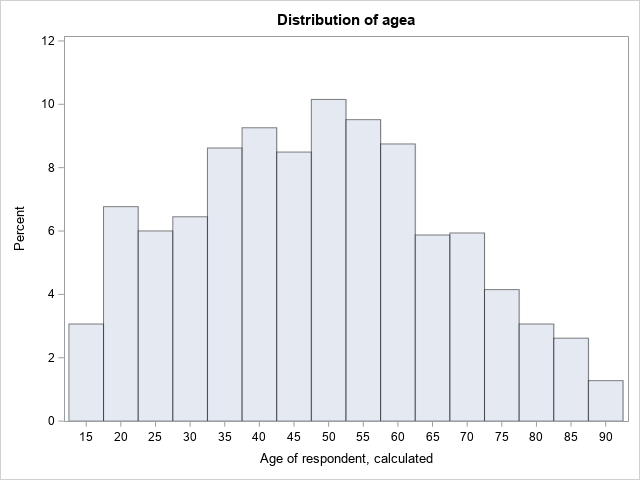
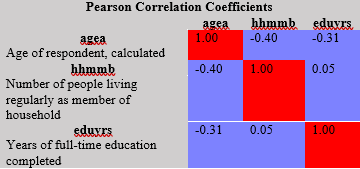


Figure 6: Sources of income distribution by age.

# **Substantive Analysis**

**Multicollianity assessment**

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

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.

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”.

Table 3: Model Information

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.

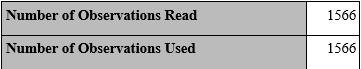
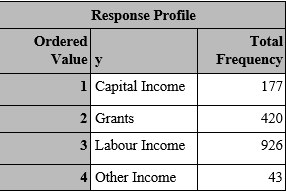
 Table 4:Number of Observations

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

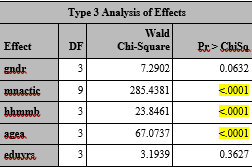
The table 8 describes the hypothesis tests for all variable in our model individually. The chi-square tests statistics concluding that highlighted variables are statistically significant since p-vales are lower than 0.05. By rejecting the accepting the null hypothesis in the case of the variable “years 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).

Table 8: Analysis of Effects

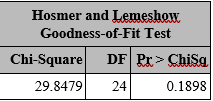
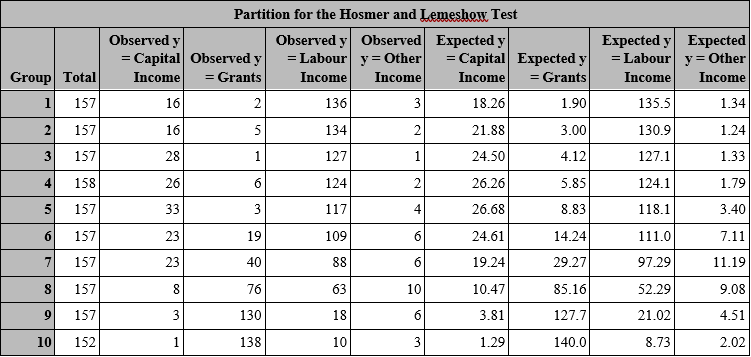
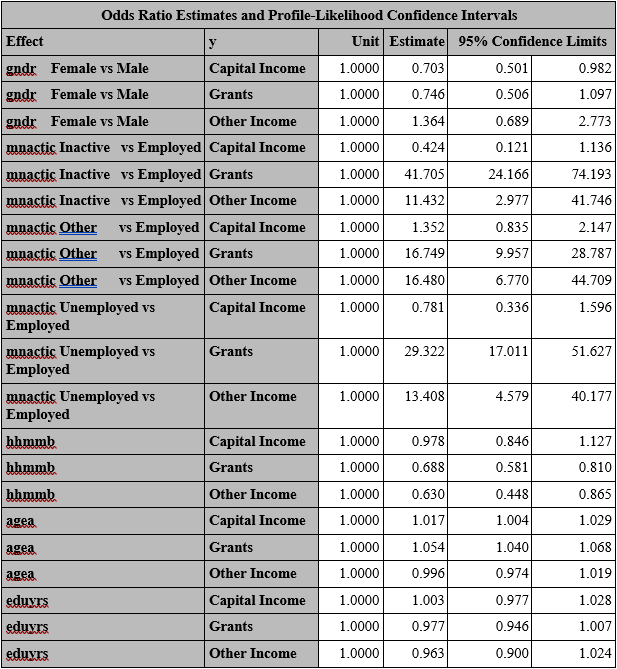
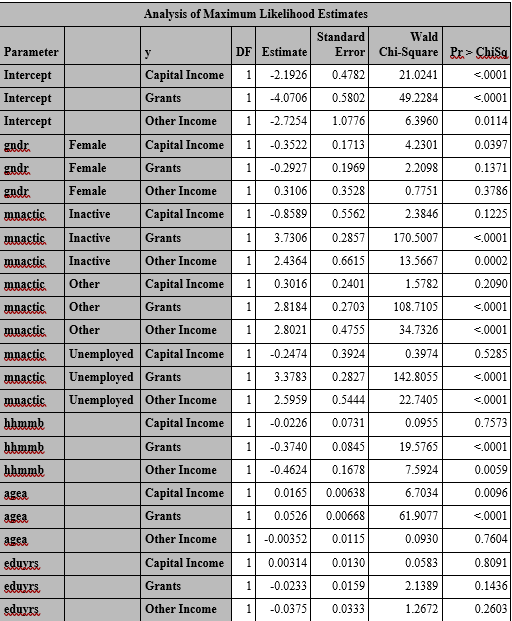
Hosmer and Lemeshow test describes the observed event rates that match expected event by subgroups. Also, we can observe by HL goodness fit test results where the p-value is 0.1898, failing to reject the null hypothesis meaning that there are not enough evidence to infer that the model is poor (low predictive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

Table 9: Hosmer and Lemeshow Test



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.

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)**

Table 10: Odds Ratio Estimates

Table 11: Analysis of Maximum Likehood Estimates

Chart, line chart

Description automatically generated 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.

Figure 7: Predicted Probabilities.

# **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.
* Chart, bar chart

  Description automatically generatedPeople 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**

[Table 1: Dataset Missing Values 3](#_Toc73537611)

[Table 2: Target variable frequency 4](#_Toc73537612)

[Table 3: Model Information **¡Error! Marcador no definido.**](#_Toc73537613)

[Table 4:Number of Observations 8](#_Toc73537614)

[Table 5: Response Profile 8](#_Toc73537615)

[Table 6: Model fit Statistics 8](#_Toc73537616)

[Table 7: Testing Global Null Hypothesis 9](#_Toc73537617)

[Table 8: Analysis of Effects 9](#_Toc73537618)

[Table 9: Analysis of Maximun Likehood Estimates **¡Error! Marcador no definido.**](#_Toc73537619)

[Table 10: Hosmer and Lemeshow test **¡Error! Marcador no definido.**](#_Toc73537620)

[Table 11: Odds Ratio Estimates Analysis **¡Error! Marcador no definido.**](#_Toc73537621)

# **List of Figures**

[Figure 1: Main source of household income distribution 4](#_Toc73651503)

[Figure 2: Sources of income distribution by gender 5](file:///C:\Users\titom\Desktop\projecrt.docx#_Toc73651504)

[Figure 3: Sources of income distribution by main activity, last 7 days 5](file:///C:\Users\titom\Desktop\projecrt.docx#_Toc73651505)

[Figure 4: Sources of income distribution by number of people in a household 6](file:///C:\Users\titom\Desktop\projecrt.docx#_Toc73651506)

[Figure 5:Sources of income distribution by years of full-time education completed. 6](file:///C:\Users\titom\Desktop\projecrt.docx#_Toc73651507)

[Figure 6: Sources of income distribution by age. 7](file:///C:\Users\titom\Desktop\projecrt.docx#_Toc73651508)

[Figure 7: Predicted Probabilities. 10](file:///C:\Users\titom\Desktop\projecrt.docx#_Toc73651509)

# **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' º  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() |  |