Social Welfare Model

Author: Erika Gomes

Context

- **This preliminary analysis** was made to show the value of building an effective model that allow us to **detect those households that are in conditions of poverty,** using data from surveys.
- In this report we will explain the steps that we followed, and we will suggest further analysis**.**

Objectives

- Present the descriptive analysis of the data.
- Describe the models used to solve the problem.
- Suggest next steps.

Phases of the analysis

The analysis and modelization of the data have several phases that it is necessary to consider.

This steps can vary depending of the problem

Conceptualisation phase

- Workshops with decision makers to determine priorities, share information and define success metrics.

- Review of academic sources, data available sources, propose ideas.

- Definition of deliverables.

Analysis and data preparation

- Selection of the databases to be used.

- Verification of the quality of the data.

- Aggregation of the data to the desired granularity.

- Analysis of the feasibility of the deliverables.

Modeling phase

Data modeling

- Definition of the methodology to be used.

- Modeling and visualization.

- Descriptive or/and predictive analysis according to the objectives of the project.

- Monitoring results

Diagnosis of the results

- Refinement of the model and definition of period of the monitoring and evaluation process.

- Validation of the model and. Application of the metrics defined in the conceptualization phase.

This is planned depending of the final product

Prototype development

- Definition of the prototype's functionality.

- Creation of dashboards, heatmaps and visualization of the results.

Knowledge transfer

- Delivery of the results, reports, apps or other output previously defined in the conceptualization phase.

- Transfer of knowledge and communications for its use.

Productization plan

Tools to use in the phase of exploration

- We use **Jupyter Notebooks because are easy to read for non-programmer's thanks to the possibility of include code, images and text.** It can be transformed in a script for its deployment in any cloud platforms as AWS, Google Cloud or Azure or even in your personal laptop or server. **It is used by companies as Netflix for their data analysis and experiments.**

- **We use Python and R for data treatment, getting the best of two great languages.** Jupyter Notebooks allows to use any language that we want as: scala, spark, javascript, between others. **Due to the size of the data we did not consider to use Spark for this case.**

Tools to use in the phase of production

- Now, that we have our first prototype, we can suggest paths to follow, in order to introduce new tools that help to get the best of our data. **The tools to choose would depend of the budget of our client and the use that will be given to the tool.**

- First, it is necessary to create a database to save the data, as the data come from surveys a relational database will be enough. Due to the possibility that the data will increase with time, we can consider include spark in our implementations to accelerate the calculations, and we can use some cloud computing as AWS (EC2 AMI) to execute our final producto.

- It is necessary to define when and where our final tool would be deployed. This how we indicate before depend of the budget, frequency of use and the utility of the tool.

Data preparation: scope & cleansing process

Target

- Develop a predictive model, that using data from surveys, allow us to **determine which households has more probability to be in conditions of poverty**. Then, suggest further analysis.

Datasets and Variables to use

- We have two datasets (one for training and one for testing) **of anonymized households**, with categorical and continuous variables. **These datasets describe the characteristics of the houses surveyed**. We also have a target variable that described the situation of each household (45% poor/55% non poor).

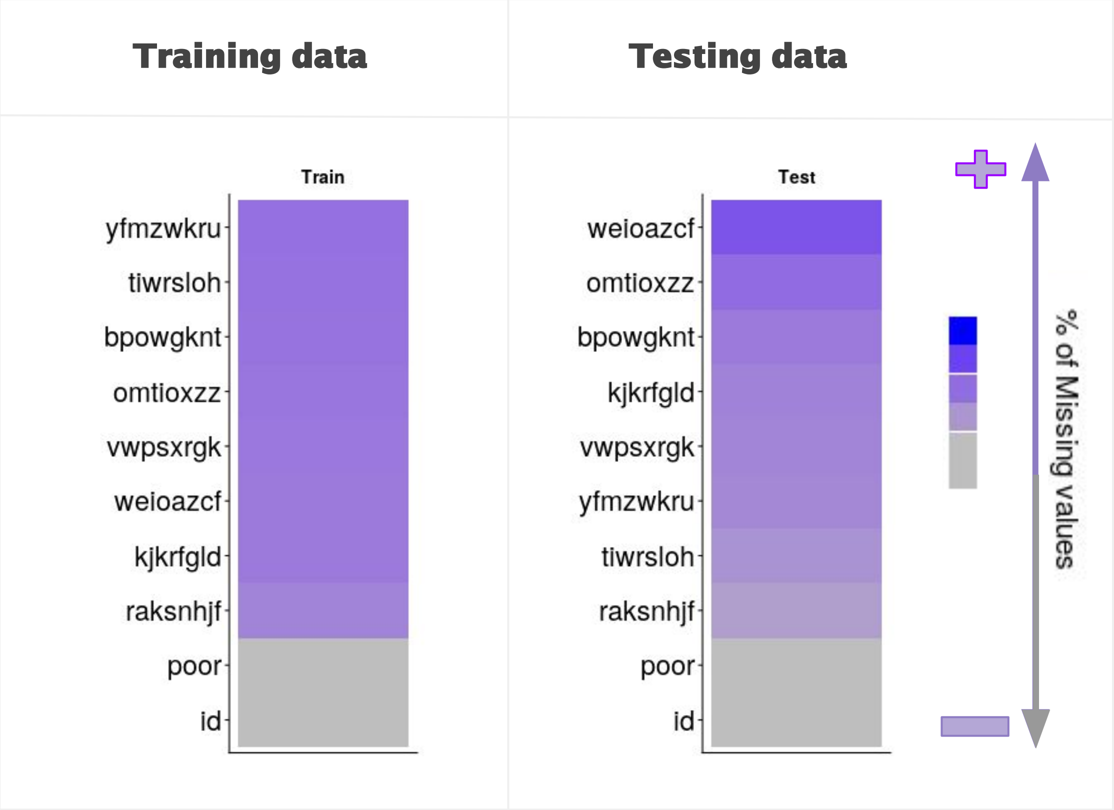
Cleansing process

- The acquired data has been prepared in the following way:

1. **Transform categorical data** (transform this variables to numeric/one hot encoding)**, treat numerical data** (eliminate outliers) and **complete missing data**.
2. **Cleaning inconsistencies:** Some variables were not considered in the model due to inconsistencies or due to poor predictive power or high correlations values (multicollinearity/noise).
3. **Creating new variables:** interactions between couple of variables can help to find more complex hidden patterns in the data that may improve the results (to explore in more detail in the future).

Missing Values (MVs)

- In the datasets we have several variables with MVs. **yfmzwkru** is the **most incomplete variable (3.18%)** in the training set**.**
- **Alternatives** to address this issue is to try to **estimate** the **MVs**, **eliminate them** or **use robust ML methods.**
- **Eliminate MVs** means **loss of information**. In order to **estimate them** we have several alternatives that are indicated in the Appendix.
- Methodologies as **boosted trees** and **Random forest** are robust methods that can **handle MVs**, these are two possibilities to explore.



Numerical treatments

There were necessary some transformations before to apply the methodologies.

→ Manage outliers

- **Understand and modify those values that can distort the final results.** Ex. Eliminate the values or apply log- transformation.

→ Create interactions

- **Create interactions between continuous variables.** In the future we can consider interactions between binary and continuous variables.

→ Fill Missing variables

- In the case of missing variables there are **several methodologies that can be used to complete the data** (see Appendix)

→ Scale data

- For some algorithms would be necessary **to scale the data to avoid that large values distort the results.** This is true for methodologies as dimensionality reduction, clustering, deep learning between others.

→ Avoid multicollinearity and redundant variables

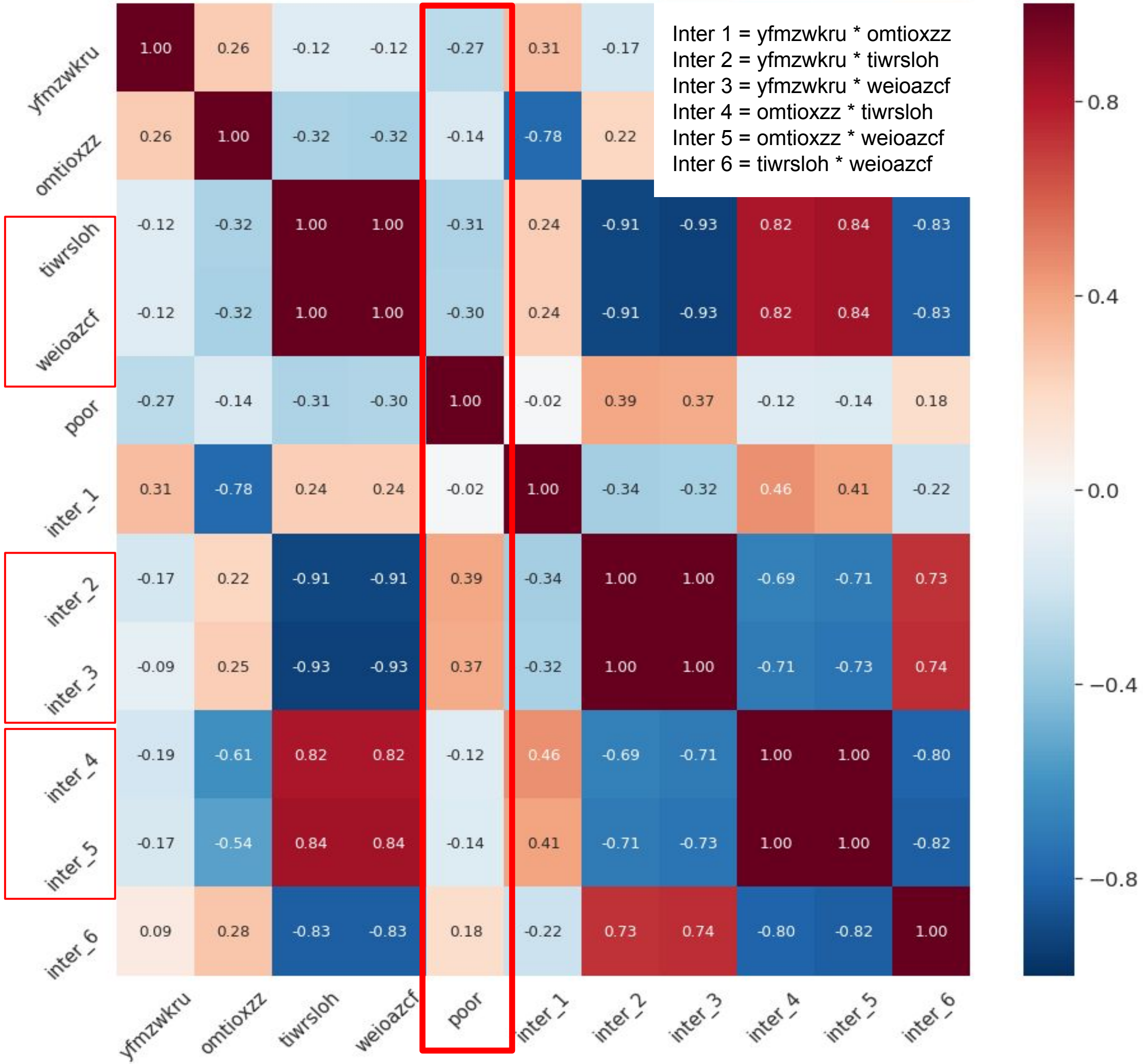
- **Take only those variables that gives information to the target variable** (poor/non poor) and **eliminate those explanatory variables that share the same information** (high correlation between them).
- Centering variables is an alternative to reduce collinearity.

Correlations

- Here we present the correlation of the numerical variables along with interactions between this variables.

- The variables more related with **poverty** are: the interaction 2 and 3, which are highly related between them. Later, tiwrsloh and weioazcf that also are higly related.

- There are variables that are highly correlated between them that can affect the results if they are all included. These are:
  - weioazcf vs tiwrsloh
  - Interactions between variables
- We **eliminate those variables that are highly related between them**. But we consider other criteria like mutual information to select between the best pair of variables.



Inter 1 = yfmzwkru * omtioxzz
Inter 2 = yfmzwkru * tiwrsloh
Inter 3 = yfmzwkru * weioazcf
Inter 4 = omtioxzz * tiwrsloh
Inter 5 = omtioxzz * weioazcf
Inter 6 = tiwrsloh * weioazcf

	yfmzwkru	omtioxzz	tiwrsloh	weioazcf	poor	inter_1	inter_2	inter_3	inter_4	inter_5	inter_6
yfmzwkru	1.00	0.26	-0.12	-0.12	-0.27	0.31	-0.17				
omtioxzz	0.26	1.00	-0.32	-0.32	-0.14	-0.78	0.22				
tiwrsloh	-0.12	-0.32	1.00	1.00	-0.31	0.24	-0.91	-0.93	0.82	0.84	-0.83
weioazcf	-0.12	-0.32	1.00	1.00	-0.30	0.24	-0.91	-0.93	0.82	0.84	-0.83
poor	-0.27	-0.14	-0.31	-0.30	1.00	-0.02	0.39	0.37	-0.12	-0.14	0.18
inter_1	0.31	-0.78	0.24	0.24	-0.02	1.00	-0.34	-0.32	0.46	0.41	-0.22
inter_2	-0.17	0.22	-0.91	-0.91	0.39	-0.34	1.00	1.00	-0.69	-0.71	0.73
inter_3	-0.09	0.25	-0.93	-0.93	0.37	-0.32	1.00	1.00	-0.71	-0.73	0.74
inter_4	-0.19	-0.61	0.82	0.82	-0.12	0.46	-0.69	-0.71	1.00	1.00	-0.80
inter_5	-0.17	-0.54	0.84	0.84	-0.14	0.41	-0.71	-0.73	1.00	1.00	-0.82
inter_6	0.09	0.28	-0.83	-0.83	0.18	-0.22	0.73	0.74	-0.80	-0.82	1.00

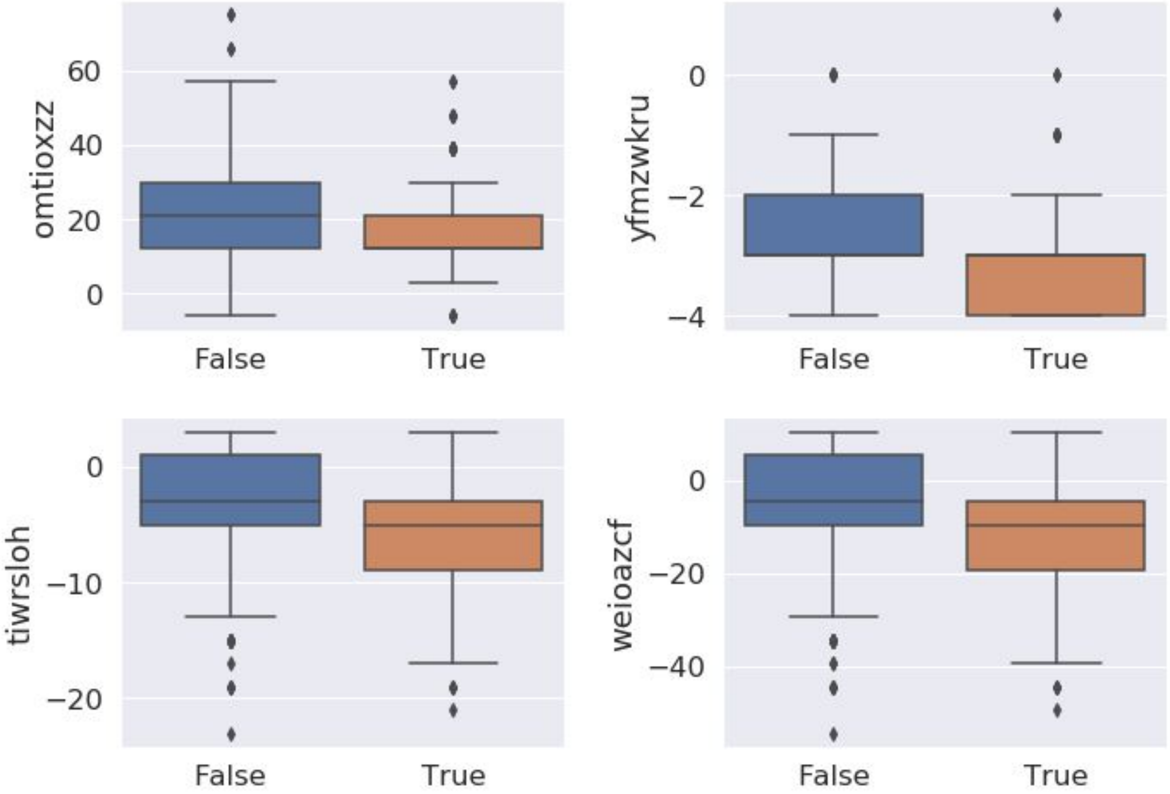
Outliers

- Here, we present the distribution of the numerical variables after modify the most extreme values.

- Clearly, it is necessary to keep treating the outliers, because they are distorting the distribution of the variables.

  It is necessary try to investigate the reason of the outliers, because change the distribution without a clear understanding of the variables could be problematic.

- Tiwrsloh and weioazcf seem to follow a similar distribution, would be necessary to eliminate one, because their contribution is the same.



Alternatives to detect anomalies:

- **Traditional statistics:** standard deviation, normal distribution, boxplots.
- **DBScan Clustering:** clustering method.
- **Robust Random Cut Forest:** Amazon’s unsupervised algorithm to detect anomalies.
- **Isolation Forest:** as the method above, the anomalies get an score, detecting the features that make them different from others.

Categorical treatments

The categorical variables have a different treatment in comparison with continuous variables*.

- One hot encoding

- We should **transform the categorical variables to 0-1 continuous variables to use them** in the modelling part.

- Combine levels/categories**
  - Using Business Logic
  - Using frequency on response rate

- In our case we could **combine categories considering the response/target rate of each level.** Using this rate we can combine levels with similar response rate into same group. Also, we can combine rare categories together, etc.
- Business logic cannot be applied in this case.

- Relation between the target variable and the
  - Categorical variables
  - Chi-squared test

- The Chi-Square test of independence is a statistical test to determine if there is a significant relationship between 2 categorical variables. We are interested in **determine if there is a relation between the target (poor/not poor) and other categorical variables.**

- Check similarity between variables
  - Cosine similarity/ Jaccard Similarities
  - Distance measures

- To avoid the problems associated with the redundant variables, we have to **determine if our categorical variables are similar or not to avoid any kind of noise in our model.**

* We also can create interactions between categorical variables or between numerical and continuous variables, this would be done in future analysis.

** In the samples of training and testing there could be different categories, so it is necessary to find the way of group this variables.

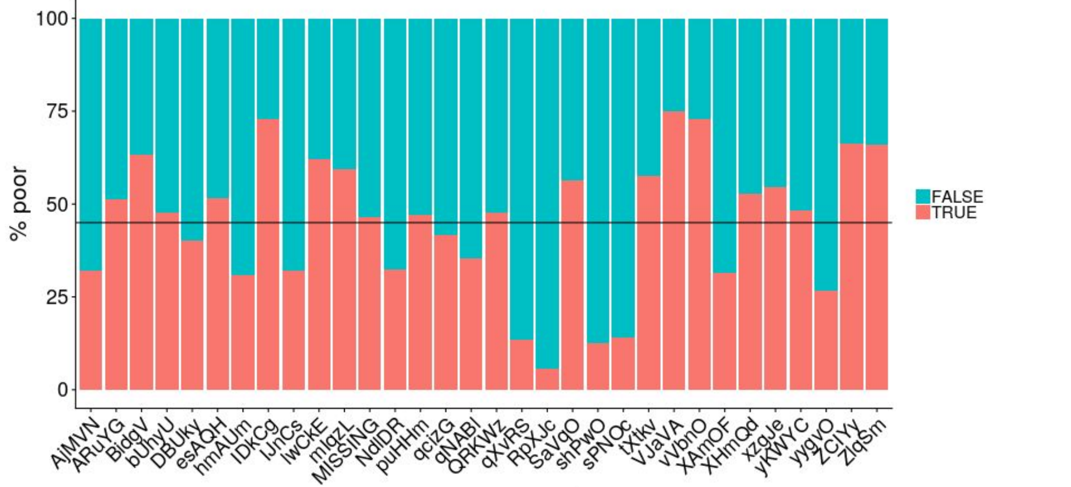
Levels

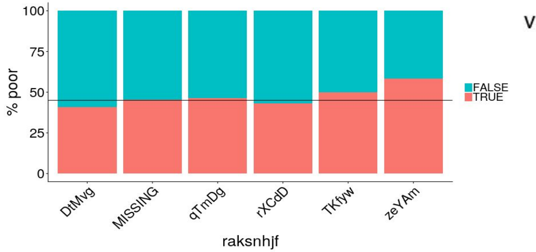
- The categories in the variables are between 5 to 31. This would mean an increase in the number of variables, when we transform them to binary.
  It is possible to combine categories observing the response rate for each level or some business logic.

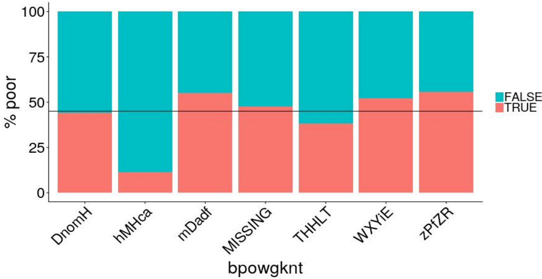
- Looking the variable ‘vwpsxrgk’, we can see some rare categories that can be combined, the same happen to ‘bpowgknt’.

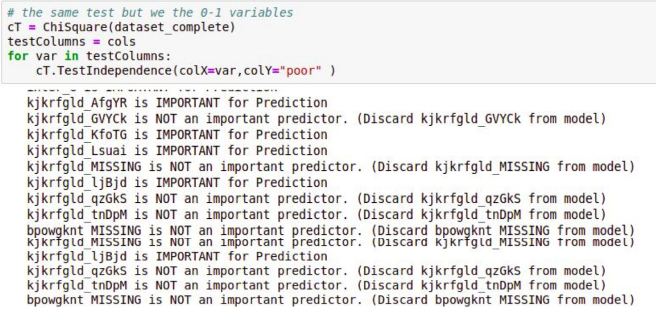
- We have to consider the possibility of other categories that are not present in the training data, but in the test data.

- The variables ‘raksnhjt’ and ‘kjkrgld’ do not show a pattern for aggregation. Alternatives based in data could be chi-squared test and/or clustering methodologies.



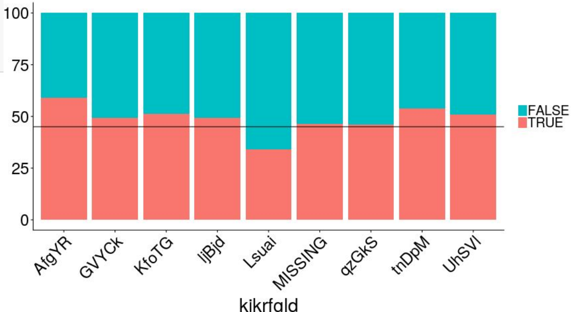






```
# the same test but we the 0-1 variables
cT = ChiSquare(dataset_complete)
testColumns = cols
for var in testColumns:
    cT.TestIndependence(colX=var,colY="poor" )
```

```
kjkrfgld_AfgYR is IMPORTANT for Prediction
kjkrfgld_GVYCk is NOT an important predictor. (Discard kjkrfgld_GVYCk from model)
kjkrfgld_KfoTG is IMPORTANT for Prediction
kjkrfgld_Lsuai is IMPORTANT for Prediction
kjkrfgld_MISSING is NOT an important predictor. (Discard kjkrfgld_MISSING from model)
kjkrfgld_ljBjd is IMPORTANT for Prediction
kjkrfgld_qzGkS is NOT an important predictor. (Discard kjkrfgld_qzGkS from model)
kjkrfgld_tnDpM is NOT an important predictor. (Discard kjkrfgld_tnDpM from model)
bpowgknt_MISSING is NOT an important predictor. (Discard bpowgknt_MISSING from model)
kjkrfgld_MISSING is NOT an important predictor. (Discard kjkrfgld_MISSING from model)
kjkrfgld_ljBjd is IMPORTANT for Prediction
kjkrfgld_qzGkS is NOT an important predictor. (Discard kjkrfgld_qzGkS from model)
kjkrfgld_tnDpM is NOT an important predictor. (Discard kjkrfgld_tnDpM from model)
bpowgknt_MISSING is NOT an important predictor. (Discard bpowgknt_MISSING from model)
```



Data Transformations

After some descriptive analysis we choose to make the following transformations.

Manage Outliers

Fill Missing categorical variables*

Fill Missing numerical variables*

Include numerical interactions

Apply one-hot encoding to categorical variables

Select variables based in correlations, chi.squared test, similarities and mutual information

We apply this same procedure to the train and test data.

NOTE: Normally, several datasets are considered in order to search the best combination of variables, but due to limitations of time we consider this procedure to generate one dataset for all the models.

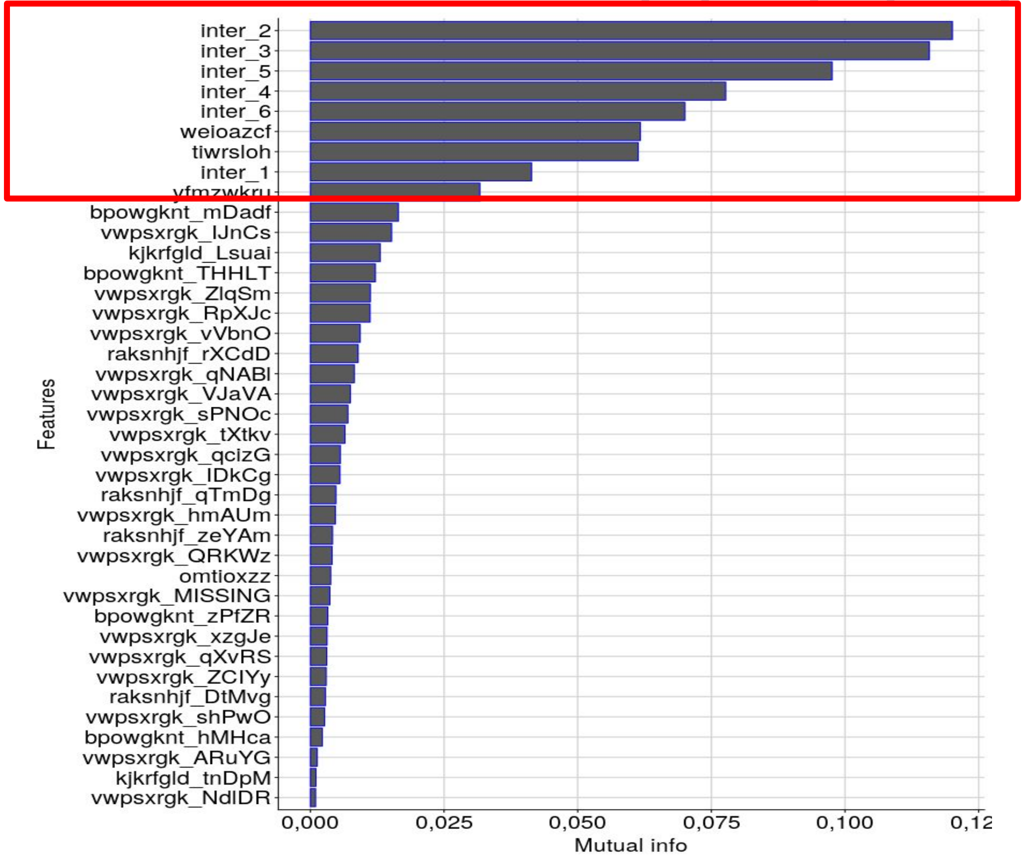
* See appendix

Dataset: Featuring Selection

Using the **Entropy measure*** we can observe, which are the **variables that determine the propensity of poverty**.

- Here we show those variables with mutual information greater than zero.

- Clearly, the interactions between the numerical variables contribute more than the single variables. We have to select those with higher MI, taking care of not to take pairs with high correlation.

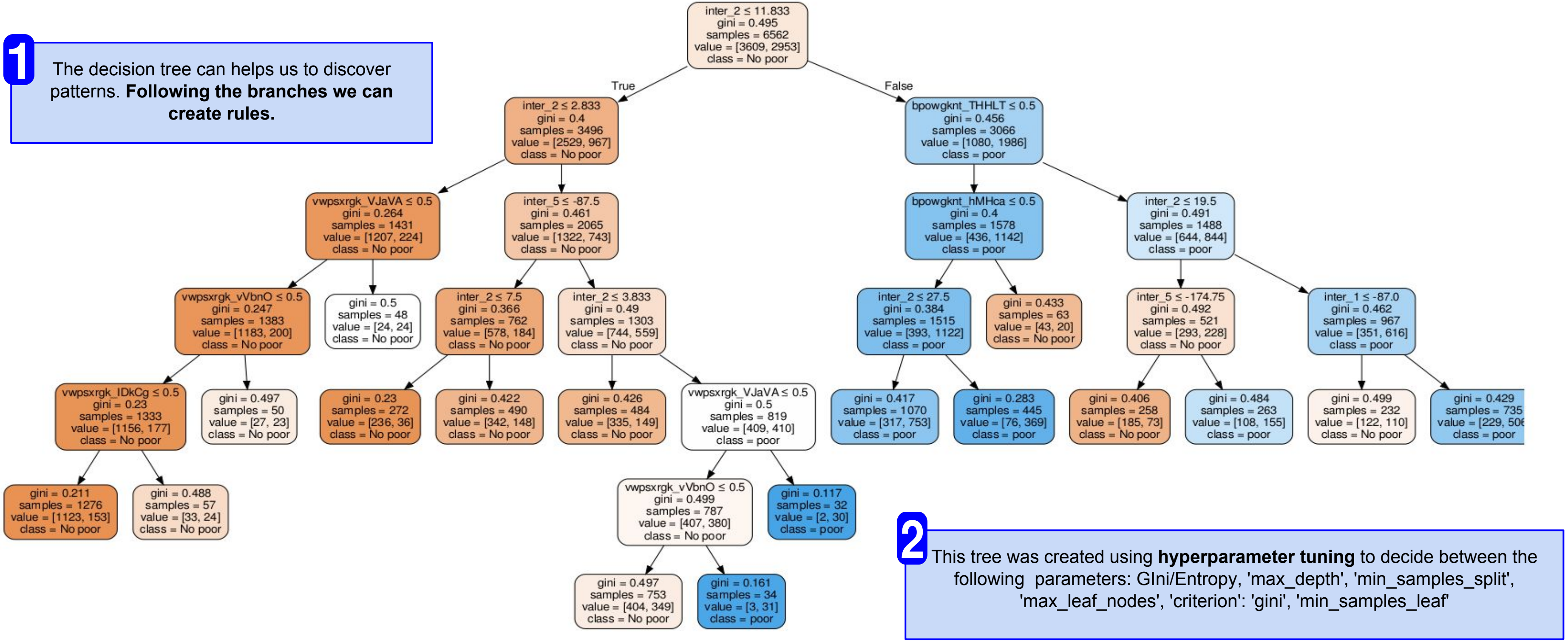


*Entropy measure** is used to calculate information gain that you obtain with each variable

Dataset: Featuring Selection

Decision trees is a technique that is mostly used as descriptive approach, through this we can get insights, that can help us to understand the variables and select those that seems interesting. .

1 The decision tree can helps us to discover patterns. **Following the branches we can create rules.**



2 This tree was created using **hyperparameter tuning** to decide between the following parameters: GIni/Entropy, 'max_depth', 'min_samples_split', 'max_leaf_nodes', 'criterion': 'gini', 'min_samples_leaf'

Methodology

- There are several methodologies that we can use to solve this classification problem. For simplicity and time, we will use **Logistic Regression, Boosted trees and Deep Learning** (Multilayer perceptron)

- **Logistic Regression** is used to understand the contribution of each variable in the model. It is not always the most accurate model but **it has high explicability.**

- **Boosted trees** is a methodology that works well in the majority of the cases, **it is robust to missing data and outliers.** It give us a set of important variables.

- **Deep learning**, used for image and text analytics mostly. **The hidden layers create interactions between variables,** that can give more information to the model.

BASELINE

It is very common to create a **baseline model, that represents the most simple model that you can develop with the information you have.** It is a common sense model, that's no involve modelling, just simple calculations. In this case, we do not have much information so will be consider the Random Assignments the baseline model.

Applied methodology

The process to create a powerful and robust predictive model relies on **the following steps:**

Handling imbalanced data	In this case we are treating with a balanced data, so in this case it would be not necessary to use methodologies to balance the data (ex. SMOTE*)	No poor aprox. 55% Poor aprox. 45%
Algorithms comparison & selection	We used **Logistic Regression (LR)**, **Gradient Boosted decision trees (GB)** and **Multilayer perceptron (MLP)** to predict the probability of a successful sale.	Value to minimize: **Classification Error**
Parameters optimization	We spent some time to optimize the parameters of the algorithms in order to obtain the best results.	**5-fold cross val** LR:34% GB: 28% MLP: 26%
Robust validation	First, we used **5-fold cross-validation** technique to train and test the models. Later, we **tested the model again** on an independent test set, not used for the training.	**Error** LR:36% GB: 30% MLP: 30%

*SMOTE (Synthetic Minority Oversampling Technique) consists of add elements of the minority class, based on those that already exist. It works randomly picking a point from the minority class and computing the k-nearest neighbors for this point. The synthetic points are added between the chosen point and its neighbors.

Results

With a **test set of 20%**, the model seems to **validate the patterns of the poverty present in the data** (confusion matrix). It is necessary to keep working with the data in order to get improvements.

Test Set (20%)

Confusion matrix (Threshold 0.5)

Logistic Regression (LR)

		Prediction		
		No	Si	**TOTAL**
Actual	No	623	268	891
Value	SI	325	425	750
TOTAL		948	693	1641

Accuracy	**0.64**
Precision	**0.61**
Recall	**0.56**
f1-score	**0.59**

Gradient Boosted Tree (GB)

		Prediction		
		No	Si	**TOTAL**
Actual	No	680	211	891
Value	Si	284	466	750
TOTAL		964	677	1641

Accuracy	**0.70**
Precision	**0.69**
Recall	**0.62**
f1-score	**0.65**

Deep Learning (MLP)

		Prediction		
		No	Si	**TOTAL**
Actual	No	566	325	891
Value	Si	138	612	750
TOTAL		704	937	1641

Accuracy	**0.72**
Precision	**0.65**
Recall	**0.82**
f1-score	**0.73**

Clearly, **Deep learning gives the best results,** the recall value is a good indicator that **we are minimizing the FN** with this methodology. These results can be improved:

- Reducing the number of variables, to those that are more relevant.
- Changing the parameters of the different models (tuning hyperparameters).
- Changing the thresholds for the probabilities of poor/non poor (now is 0.5).
- Considering other methodologies.
- Investigate the possibility of include new variables (external variables).

Conclusions & Recommendations*

- **Using external variables we can improve the results.** We did several treatments to the variables depending of its nature: numerical or categorical. **Although, more analysis are required to get better results.** A recommendation is to understand the source of the variables, in order to get better insights and get external macro variables, that can help to improve the solution.

- Due to limitations of time we only use 4 methodologies. Additional methodologies could be applied as **Catboost, that is a boosted model designed to consider categorical variables and get the most of them.**

- The **performance of the models could be improved** by incorporating additional features which include dynamic interactions (time), and more interactions between variables.

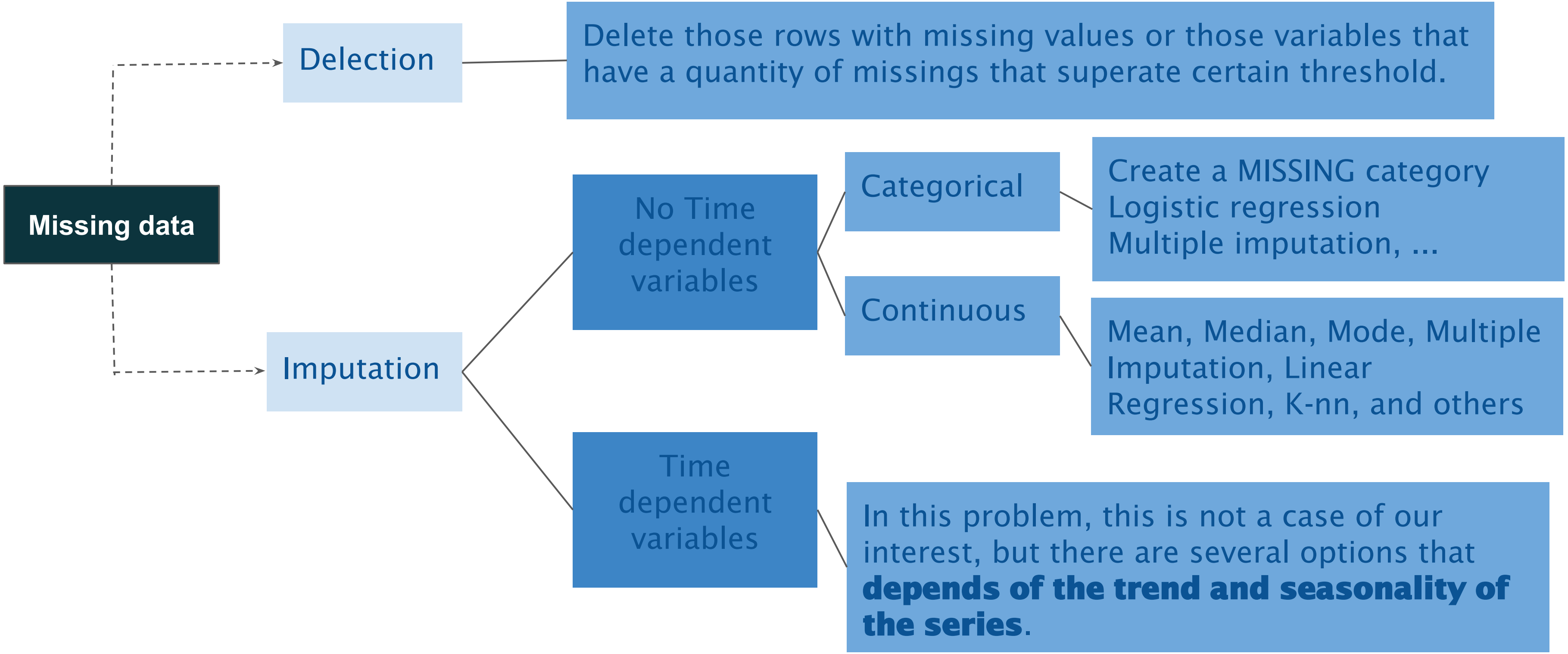
- Additional analysis as **hyperparameter tuning** (grid search/bayesian optimization) is required to improve analysis. Also, **ensemble methods** (see Appendix) can be useful as additional methodology.

- To understand better the categories inside of each variables (in order to combine the most similar categories together), would be interesting use the T-SNE method to see how the categories work inside of the generated groups.

Appendix

Treatment of missing data

Depending of the data and the feasibility of the solution we have several alternatives:



Missing data

Delection

Delete those rows with missing values or those variables that have a quantity of missings that superate certain threshold.

Imputation

No Time dependent variables

Categorical

Create a MISSING category
Logistic regression
Multiple imputation, ...

Continuous

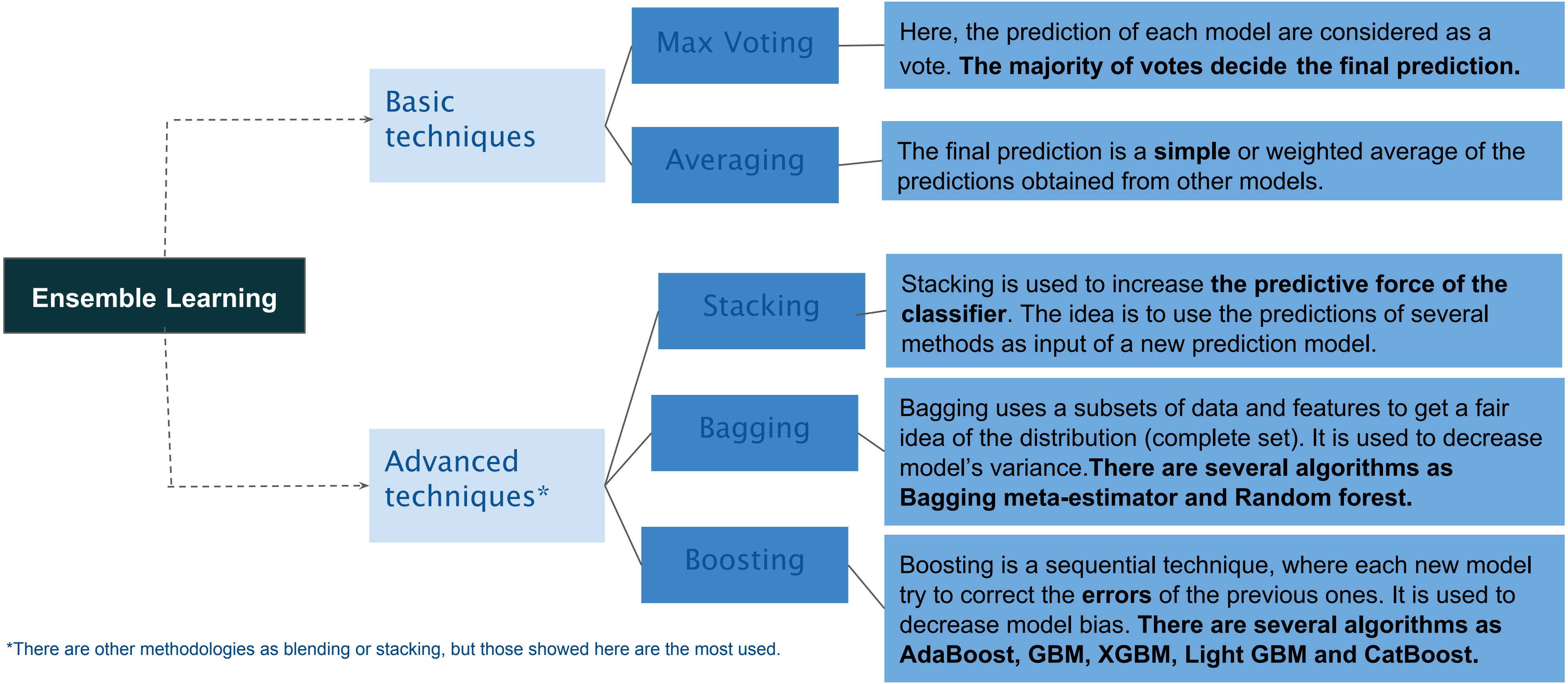
Mean, Median, Mode, Multiple Imputation, Linear Regression, K-nn, and others

Time dependent variables

In this problem, this is not a case of our interest, but there are several options that **depends of the trend and seasonality of the series**.

Ensemble Learning Techniques

A future modelling strategy would be use Ensemble techniques to improve our results.



Ensemble Learning

Basic techniques

Max Voting

Here, the prediction of each model are considered as a vote. **The majority of votes decide the final prediction.**

Averaging

The final prediction is a **simple** or weighted average of the predictions obtained from other models.

Advanced techniques*

Stacking

Stacking is used to increase **the predictive force of the classifier**. The idea is to use the predictions of several methods as input of a new prediction model.

Bagging

Bagging uses a subsets of data and features to get a fair idea of the distribution (complete set). It is used to decrease model's variance.**There are several algorithms as Bagging meta-estimator and Random forest.**

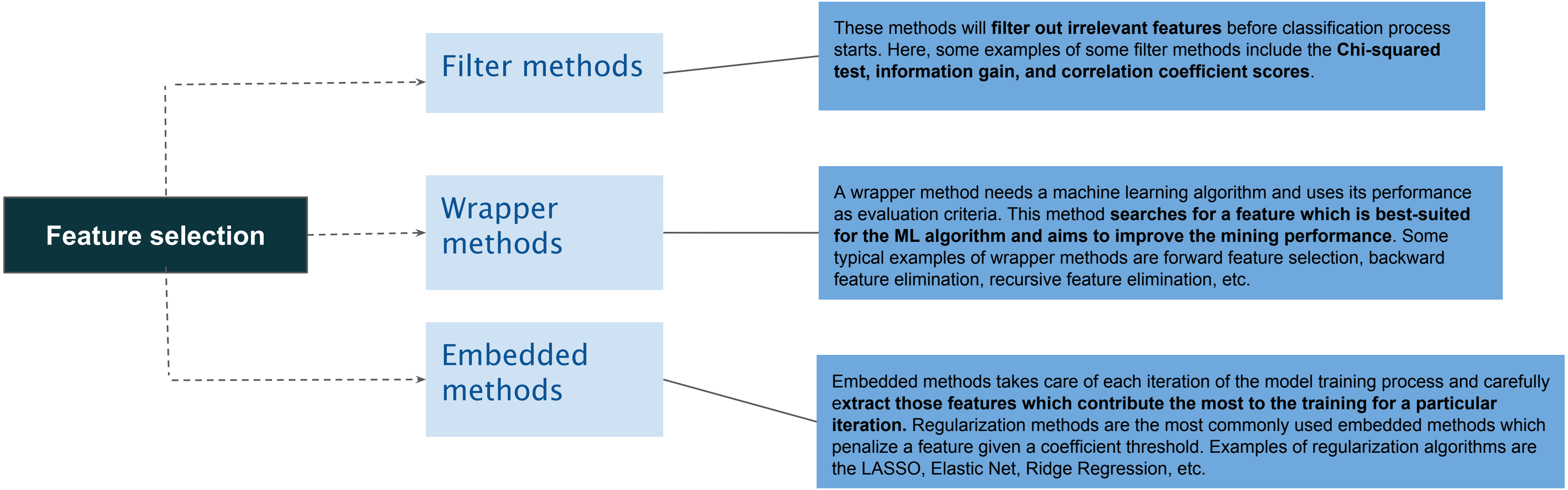
Boosting

Boosting is a sequential technique, where each new model try to correct the **errors** of the previous ones. It is used to decrease model bias. **There are several algorithms as AdaBoost, GBM, XGBM, Light GBM and CatBoost.**

*There are other methodologies as blending or stacking, but those showed here are the most used.

Featuring Selection

Unnecessary features act as a noise for which the machine learning model can perform terribly poorly.



Feature selection

Filter methods

These methods will **filter out irrelevant features** before classification process starts. Here, some examples of some filter methods include the **Chi-squared test, information gain, and correlation coefficient scores**.

Wrapper methods

A wrapper method needs a machine learning algorithm and uses its performance as evaluation criteria. This method **searches for a feature which is best-suited for the ML algorithm and aims to improve the mining performance**. Some typical examples of wrapper methods are forward feature selection, backward feature elimination, recursive feature elimination, etc.

Embedded methods

Embedded methods takes care of each iteration of the model training process and carefully **extract those features which contribute the most to the training for a particular iteration.** Regularization methods are the most commonly used embedded methods which penalize a feature given a coefficient threshold. Examples of regularization algorithms are the LASSO, Elastic Net, Ridge Regression, etc.

Hyperparameters Tuning

A good choice of hyperparameters can really make an algorithm shine. The process of finding the most optimal hyperparameters in machine learning is called hyperparameter optimization.

➔ Grid Search

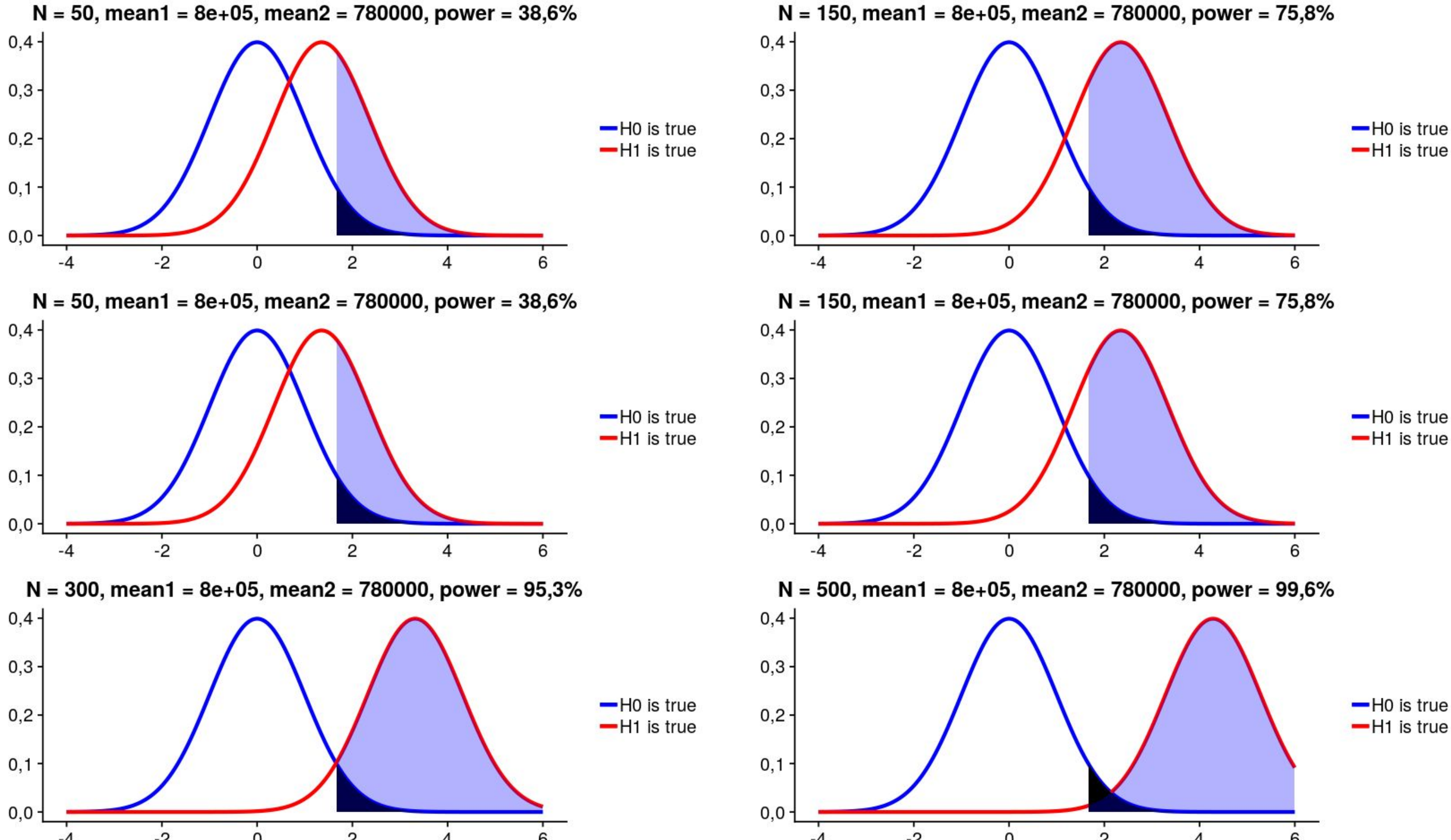
- Search for all possible combinations of parameters and measure its performance using cross-validation.

➔ Random search

- Randomly samples the search space and evaluates sets from a specified probability distribution.

➔ Bayesian Optimization

- Bayesian optimization typically works by assuming the unknown function was sampled from a Gaussian Process (GP) and maintains a posterior distribution for this function as observations are made.



N = 50, mean1 = 8e+05, mean2 = 780000, power = 38,6%

N = 150, mean1 = 8e+05, mean2 = 780000, power = 75,8%

N = 50, mean1 = 8e+05, mean2 = 780000, power = 38,6%

N = 150, mean1 = 8e+05, mean2 = 780000, power = 75,8%

N = 300, mean1 = 8e+05, mean2 = 780000, power = 95,3%

N = 500, mean1 = 8e+05, mean2 = 780000, power = 99,6%

Metodología

Impacto de las variables: Análisis de regresión multivariado

Es posible hacer un análisis más profundo de correlaciones e interacciones entre las variables a través de **análisis de regresiones**.

Cuando hemos descubierto una correlación entre dos variables queremos agregar tres o más variables para determinar:

1. Si la relación entre dos variables puede ser debida a otros factores que afectan indirectamente a la variable de interés,
2. Si las relaciones son las mismas para diferentes tipos de individuos,
3. Entender si hay diferencias entre distintos tipos de tratamientos (dar 5 ptos o 7ptos a un tipo de incentivos)
4. Controlar diferencias entre individuos, años, trimestres, entre otros.

Los efectos explicados por las regresiones son las medias ponderadas de múltiples comparaciones agrupadas para entender el impacto de la variable de interés.

Controlar diferencias entre individuos, grupos y/o periodos de tiempo

$$Y_i = \alpha + \beta P_i + \gamma A_i \; + \epsilon_i$$

Variación media conseguida con el tratamiento, cuando se mantienen constantes el resto de las variables

α: intercepto
β: tamaño del efecto del tratamiento(s).
γ: efecto de la variable(s) de control.

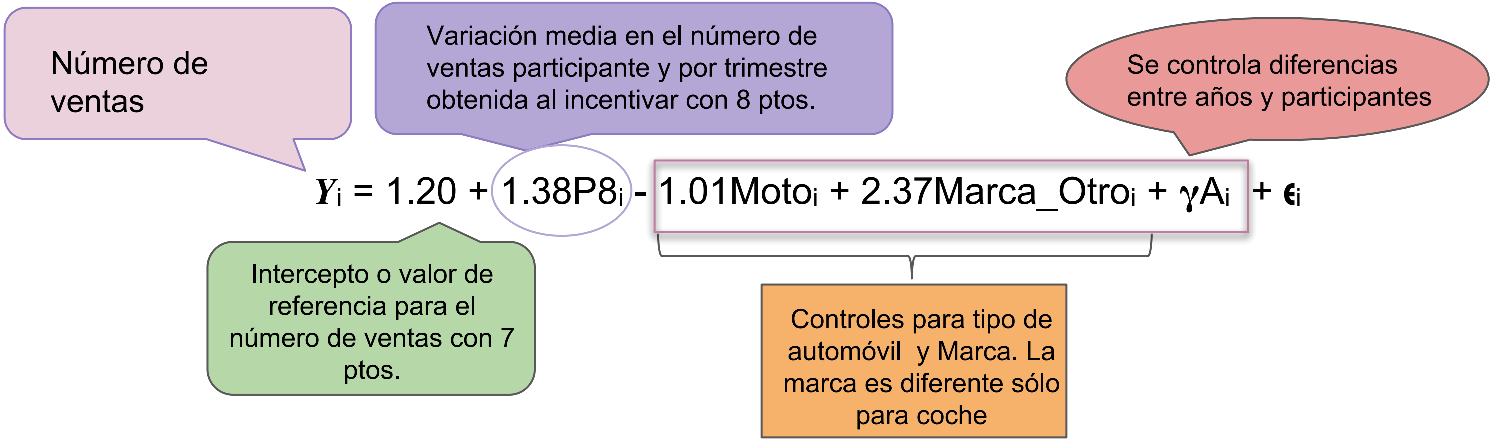
Ejemplos:

- Nivel de ingresos (Y) vs. tipo de universidad (P: pública o privada) tienen una alta correlación o alto efecto (β), siempre que no se consideren los resultados de los exámenes de admisión en la ecuación. Y y P tienen una correlación espuria que se desvanece al incluir una tercera variable.
- La edad tiene un efecto positivo (β) en el uso de facebook cuando controlas por la variable trabajo (0 no trabaja, 1 trabaja). De lo contrario la correlación es muy baja.

Metodology

Feature Impact: Multivariate Analysis

¿Existen diferencias entre otorgar incentivos de 7 y 8 ptos. en el número de ventas por participante y trimestre ?



Número de ventas

Variación media en el número de ventas participante y por trimestre obtenida al incentivar con 8 ptos.

Se controla diferencias entre años y participantes

$$Y_i = 1.20 + 1.38P8_i - \boxed{\underbrace{1.01Moto_i + 2.37Marca\_Otro_i}_{} + \gamma A_i} + \epsilon_i$$

Intercepto o valor de referencia para el número de ventas con 7 ptos.

Controles para tipo de automóvil y Marca. La marca es diferente sólo para coche

- Es claro que **los incentivos de 8ptos. generan 1.38 más de ventas en promedio en comparación con el incentivo de 7 ptos.**
- Estos resultados controlan las diferencias que pueden existir entre los vendedores y las diferencias macroeconómicas que pudieron ocurrir en diferentes años.

t-SNE: a nonlinear nondeterministic algorithm (T-distributed stochastic neighbor embedding) that tries to preserve local neighborhoods in the data, often at the expense of distorting global structure. You can choose whether to compute two- or three-dimensional projections.

Look for the parameters

..