Can machine learning algorithms predict if a person earns a living wage?[1]

Evidence from socioeconomic data of Chilean individuals

Joaquin Perez-Lapillo
City, University of London

1. DOMAIN DESCRIPTION, RESEARCH QUESTIONS AND STRATEGY

1.1. DOMAIN DESCRIPTION AND DATA SOURCE

1.1.1 Domain

The concept of a living wage has been around for more than fifteen years in the UK thanks to a movement started by citizens interested in ensuring that "everyone can earn a real living wage that meets the costs of living, not just the government minimum"[2]. Today, their movement has turned into a new benchmark for private companies in terms of corporate social responsibility (CSR).

On the other hand, the discussion regarding living wages is just beginning in developing countries. For the case of Latin-American economies, poverty reduction in the last decades has started shifting the attention to an enormous segment of the population that does not qualify as poor but neither do they earn enough to cover real costs of living.

Given this context, the motivation of this project is to find if machine learning algorithms can predict if a person belongs to that group given a set of socioeconomic attributes.

This study should be relevant for governmental agencies, NGOs, CSR specialists, and the public interested in incorporating living wage standards in developing countries.

1.1.2 Data

Data comes from the last release (2017) of a Chilean national survey called CASEN, owned by the Ministry of Social Development[3].

This survey is the official source to know about socioeconomic conditions of households and for poverty assessment policymaking.

1.2. RESEARCH QUESTIONS

The objective of this project is to answer to the following questions:

[1] Jupyter notebook available at: https://smcse.city.ac.uk/student/aczd100/LivingWageProject.html
[2] https://www.livingwage.org.uk/history
[3] http://observatorio.ministeriodesarrollosocial.gob.cl/casen-multidimensional/casen/basedatos.php

1. Can machine learning algorithms predict if an individual earns a living wage in Chile, given a set of socioeconomic attributes?
2. Which are the most relevant features that derive from the prediction?

1.3. STRATEGY AND PLANS

1.3.1 Strategy

Based on previous work in earning potential prediction using machine learning algorithms (Hoffman, 2011), the strategy will consist in comparing the performance of 4 algorithms in a binary classification problem using a set of socioeconomic attributes as predictors.

1.3.2 Plans

The analysis will be performed in 8 steps:

1. Import libraries and data
2. Filter the dataset to match the scope of the project
3. Apply transformations to variables
4. Deal with missing values and outliers
5. Perform exploratory data analysis (EDA)
6. Apply final transformations, train/test split and dimensionality reduction techniques
7. Build models and compare their performance
8. Analise feature importance

2. FINDINGS AND REFLECTIONS

2.1. EXPLORATORY DATA ANALYSIS (EDA)

2.1.1 Describing wages

The living wage in Chile has been set in 391,400 Chilean pesos (CLP) per month[4] by the Wage Indicator Foundation (WIF)[5], an international organisation that reports living wages for countries based on surveys. This estimation will be used as a threshold to separate the data into two target groups.

Figure 1 shows that the distribution of wages in Chile is right-skewed like in most countries. In terms of presence of both target classes, people classified as “earning a living wage” account for 43% of total (minority class) while 57% of Chileans are not earning a living wage.

[4] Upper bound living wage for a single-adult.
[5] https://wageindicator.org/salary/living-wage/chile-living-wages-2018-country-overview

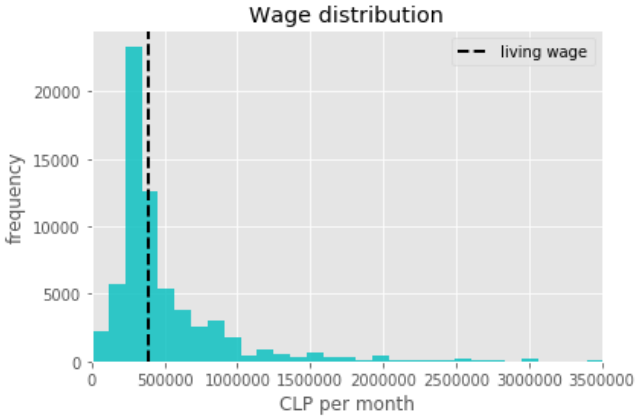

Figure 1: distribution of wages in Chile (2017)

2.1.2 Exploring continuous variables

The dataset includes 4 continuous variables: years of education, age, work experience, and hours worked per week. Figure 2 shows histograms of these features by the two target groups. The main observations that emerge from the figure are:

- Distributions of both groups are clearly different for the cases of education and age, having clear peaks at different points of the distribution. Hence, these two variables show great potential for prediction.
- Individuals with little work experience are more likely to not earn a living wage and there is not much difference in terms of the number of hours worked.

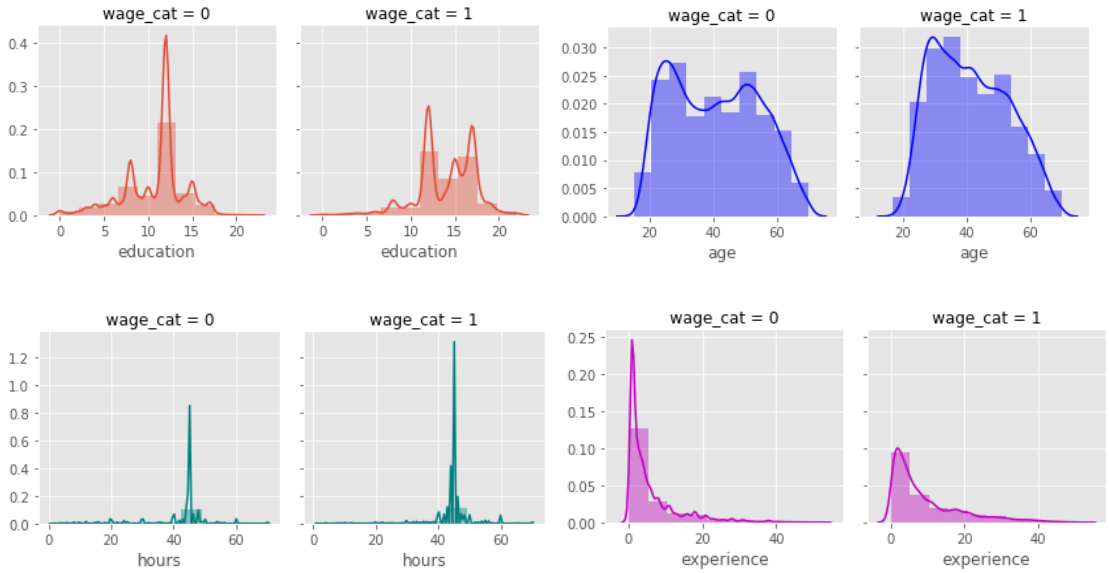

Figure 2: distribution of continuous variables by target groups

2.1.3 Exploring categorical variables

A total of 12 categorical attributes were selected from the dataset. The group could be separated into 3 subgroups:

- **Demographics:** sex, region, zone, marital status, and ethnic group
- **Job characteristics:** position, sector, type of contract, and industry
- **Other:** bank account holding, social security affiliation, and health condition

Figure 3 shows the probabilities of earning a living wage by 3 attributes (job position, industry, and region) as an example of what has been performed for all categorical variables.

A characterization of both groups can be made from the visual EDA:

Most likely to NOT earn a living wage (target=0)	**Most likely to earn a living wage** (target=1)
Women, single	Man, married
Living in centre regions and rural areas	Living in extreme regions and the capital
Having a part-time contract	Having a permanent full-time contract
Salespersons, operators, farmers, and having a job that requires no qualifications	Professionals, military, and politicians
Working in farming, restaurants and accommodation, and retail industries	Working in mining, financial companies, and government
Not having a bank account and social security affiliation	Having a bank account and social security affiliation

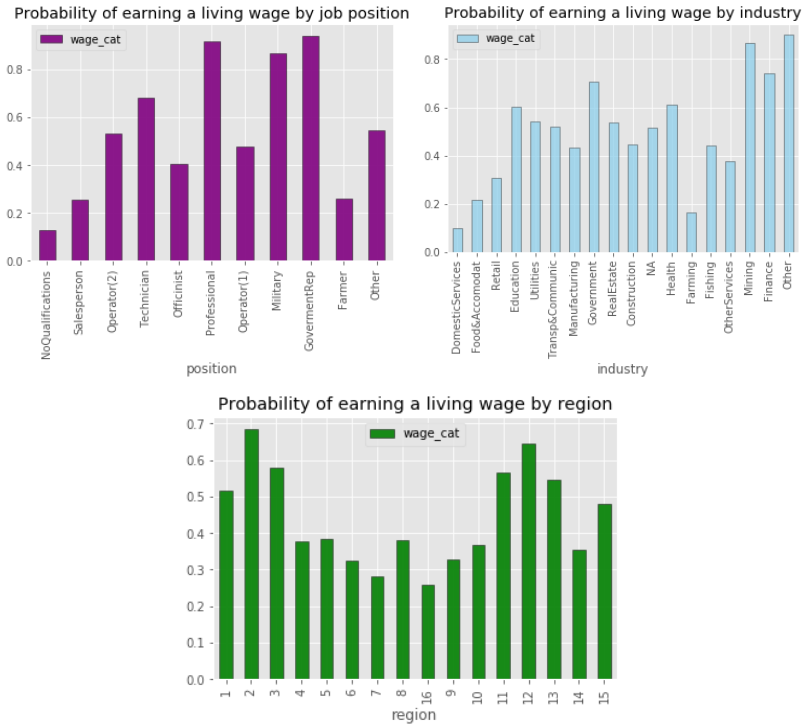

Figure 3: probability of earning a living wage by job position, industry and region

2.2. MACHINE LEARNING ALGORITHMS PERFORMANCE IN CLASSIFICATION

2.2.1 Building models

To answer the research questions of the project 4 models were built to compare their performance in a supervised classification task. The algorithms selected for study are well known for being adequate for classification: Naïve Bayes (NB) and Decision trees (DT) are relatively simple, fast algorithms while Random forest (RF) is more computationally intense but usually achieve higher results (M. Fernández-Delgado, 2014). A classic statistical model - Logistic regression (LR) - was included to test its efficiency in solving the problem.

A random split of the dataset into training and test sets was applied using the 70/30 rule. This process was done over two alternative datasets: one with the 16 original predictors (input for NB, DT, and RF) and another with categorical variables transformed into dummies (input for LR).

Due to the sensitivity and relevance of hyperparameter tuning for Random forest modelling (Braiman, 2001), a 5-fold cross-validated grid search was performed over a combination of trees, features per node, maximum leaf depth and evaluation criterion. The resulting optimal combination of hyperparameters[6] was then used to train a final model.

2.2.2 Performance evaluation

As the evaluation criteria was defined to maximize accuracy of both target classes and given that the dataset is relatively balanced, the algorithms performance was compared in terms of accuracy and area under the curve (AUC) scores.

The overall accuracy results in Figure 4 show a significantly better performance of RF and LR over NB and DT, which is confirmed by later AUC scores calculation (Figure 5). According to a commonly used scale[7], AUC scores of RF and LR can be classified as "good" (almost "excellent") while NB and DT results can be classified as "fair".

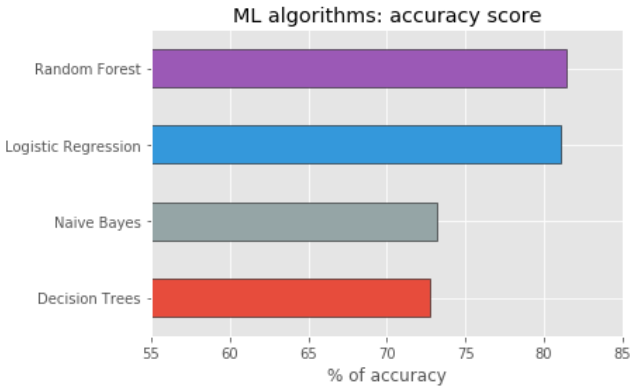

Figure 5: Accuracy scores in classification. RF (81.4), LR (81.1), NB (73.2) and DT (72.8)

[6] Optimal hyperparameters found: 300 trees, default for number of features, maximum depth of 15 and entropy as evaluation criterion.

[7] http://gim.unmc.edu/dxtests/roc3.htm

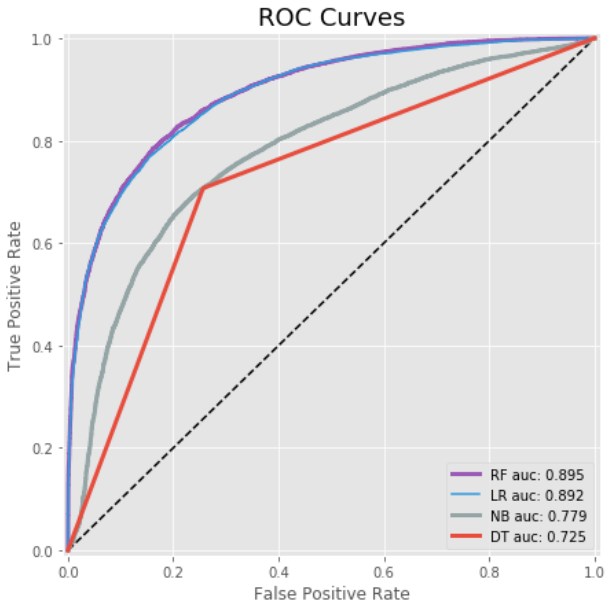

Figure 6: ROC Curves and AUC scores of ML algorithms

2.3. FEATURE IMPORTANCE ANALYSIS

Using RF rank of features by their importance in the prediction (Figure 7) and LR coefficients as a complement, we can derive the following:

- Job position and years of education were selected as the most important features. In the case of job position, the decision seems reasonable given that higher positions are commonly associated with higher salaries, meanwhile for education, the results confirm EDA findings.
- LR coefficients found that job positions more likely to make a living wage are professionals, politicians, and military as opposed to farmers, salespersons, and non-qualification jobs.
- Having a bank account is relevant in predicting living wage, which could mean that banks in Chile are unwilling to have low-wage workers as clients.
- Although it did not seem relevant in the EDA, hours worked per week was selected as part of the top 5.
- The fact that region is also an important feature shows the need for having different living wages depending on the geographical area.
- Experience is less relevant than education in defining if an individual earns a living wage. This could be related to the low qualifications of older workers who are the ones with more experience.

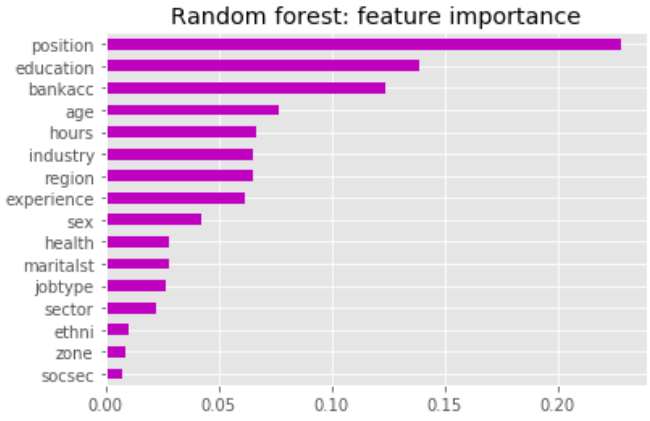

Figure 7: Random forest feature importance in prediction

2.4. FINAL REMARKS AND LIMITATIONS

Results showed that Random forest and Logistic regression achieve high performance with accuracy scores of 81.4% and 81.1% respectively. Hence, we say that these algorithms are well suited to predict if an individual earns a living wage in Chile given a set of socioeconomic variables.

This finding could be helpful to agencies looking to assist individuals without having to access to income information which is often restricted.

The most relevant features found in predicting living wage were job position and years of education. Other relevant features that were found are bank account holding, age, hours worked per week, and region.

The study showed the need for having different living wages for macrozones of the country (e. g. extreme regions, centre and capital) due to their relevance in the prediction.

Some limitations of the study relate to the characteristics of data. Given that the survey's main objective is to measure poverty, it could be biased in terms of not fully represent wealthier individuals. Also, the selection of variables could have omitted other relevant features.

REFERENCES

Braiman, L. (2001). Random Forest. *Machine Learning Journal*, 5-32.

Hoffman, M. (2011). *Predicting earing potential on Adult Dataset.* Dublin: Institute of Technology Blanchardstown.

M. Fernández-Delgado, E. C. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, 3133-3181.