Cement to save lives ?

Hugo Birch, Elia Escoffier, Pierre Vuillecard
Applied Data Analysis, EPFL, Switzerland

***Abstract*—Piso Firme is a large-scale Mexican program that replaces dirt floors with cement floors. This study aims to identify main causes of children's diseases with the help of machine learning methods and estimate the impact of Piso Firme on exposure to disease by developing an observational study. Using a set of indicators from a survey conducted in 2005 about a cohort of 6 693 individuals living in both cities targeted or not by the program, a logistic regression model was trained to estimate identify the most relevant and interpretable combination of features. Considering results from the logistic regression, an observational study was conducted and results were then compared with those from a naive study. Said results can help identify targets for the improvement of children's development and health.**

I. Introduction

In the latest classification of countries issued by the United Nations, Mexico is categorized as a developing economy. Mexico is at an advanced stage in its demographic transition, the population is expected to stabilize in the upcoming times. In the meantime, the government has to deal with improving quality of life for its population. This is the exact scope of large-scale Mexican social program - Piso Firme, which replaces dirt floors with cements floors, in the state of Coahuila. The urban area of La Lagune is located on the border of Coahuila and Durengo, two neighboring states. As Dureng did not implement Piso Firme in its state, La Laguna is the perfect location to compare the effect of such program on its population. A first study [1], published in 2009 tried to identify the impact on health of young children as well as the mental health and happiness of their mothers in two cities of La Laguna : Torreón and Gómez Palcio. Our goal is to extend their work, and further the relationship between housing and children's health. We will investigate about the main causes of diseases and determine the impact of Piso Firme on these factors.

II. Data exploration

The data provided information on individual. It is an aggregation of the 2000 Mexican census [2] and of a survey performed in partnership with the Mexican National Institute of Public Health in 2005[3] . The dataset is given in a dta format and more information can be found in the *README.pdf* file in the Data directory.

A. Importance of children

The dataset includes 6 693 individuals, each have 89 different characteristics. As this study focuses on children only, it seems important to visualize what proportion of children there is in the whole set. The median age is 5 years old. There is a clear distinction between adults and children, as seen on Figure 1: considering all individuals under 15 years old seems to be a good threshold to include all children, this represents 4 093 individuals.

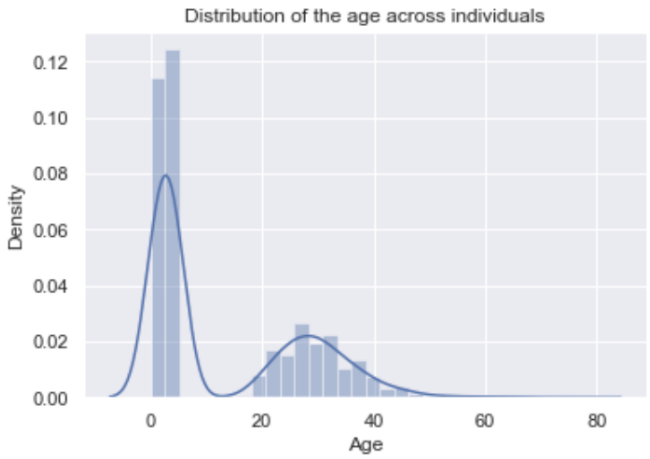

Fig. 1. Distribution of individuals according to their age.

B. Missing values

Originally, without selecting any group of individual, there were 62 333 missing values, a distribution according to each feature is presented in Figure 2. By only looking at children, the number of missing values is reduced by 80%, reaching 16 867. As shown in Figure 2, some features contain more than 60% of missing content. It turns out that 2 features are completely empty for children : *S_malincom* and *S_palincom*. These empty features and all features with over 60% missing value were removed.

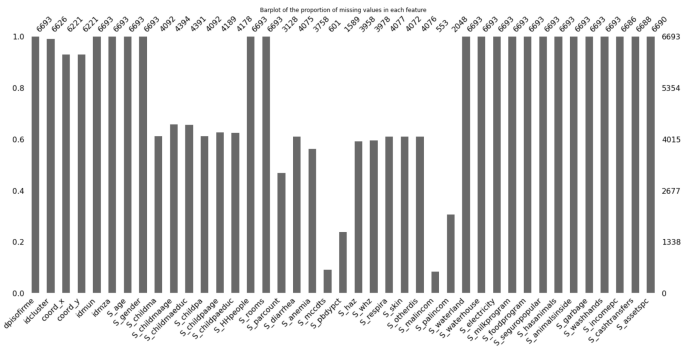

Fig. 2. Proportion of missing values in each feature.

The features about children cognitive development tests, also contain a significant amount of missing values, as these tests were only performed on a very precise subpopulation of children. For this reason, they also had to be removed. This brought the number of missing content down to 704. It represents less than a quarter of the entire population of children, it seems reasonable to impute them with a K-nearest neighbors algorithm. It uses the mean value of the 5 nearest

neighbors. It was ensured that this imputation stays consistent with the original distributions of concerned features.

C. Feature engineering

Before starting the analysis, each feature's distribution was plotted. It appears that 2 variables, *S_incomepc* and *S_cashtransfers*, look like power laws : one can see a peak on the distribution around 0, and then the distribution is right long-tailed. To reduce skewness from these 2 distributions, a log transform was applied to *S_incomepc* and square root transform was applied to *S_cashtransfers*. All continuous features were standardized following these changes.

D. Construction of a target for feature selection

In the original list of features, 6 of them can help to determine the global health status of children : *S_parcount*, *S_diarrhea*, *S_anemia*, *S_respira*, *S_skin*, *S_otherdis*. Except for *S_parcount*, the other features are binary variables (either 0 or 1), that's why this feature was changed to a binary one : 1 if you have at least one parasite, 0 otherwise. From these variables, it was possible to set up a target for the next step of the analysis: creating a variable which reflects the general health status of children. If at least one of the aforementioned variables is positive then the patient is classified as not being healthy and the variable is set to (1). Otherwise he is classified as healthy and the variable is set to (0). Out of the population of children, 71% of them were identified as not healthy, and 29% as healthy. After establishing the target, these variables were removed from the feature set.

III. Identifying the main causes of diseases

This section focuses on identifying the main causes of why the children in this study contract diseases. First, a selection of features will be chosen after analyzing which are the most meaningful. Then, the importance of each feature will be studied and quantified.

A. Feature selection

Seven different methods of feature selection were applied to the data. These methods helped to identify which variables have the most impact on the diseases. Five of these methods (Pearson's correlation, Chi-2 test, LightGBM, RFE and Random forest) can have a maximum number of feature applied. The selection is either based on a ranking according to correlation score or Chi-2 score, or on tree-based learning and it's own definition of feature importance. The 2 others (Lasso regularization and RFECV) identify informative features either by imposing a condition on the sum of the absolute value of regression coefficients which sets some coefficients to 0, or by ranking them according to their coefficients and eliminating recursively the weakest. Both of these methods attempt to eliminate dependencies and collinearity that may exist in the model.

By setting the number of feature to select to $\frac{15}{26}$ this allows to capture a large range of possibility. Figure 3, shows the nuber of times each feature was selected by the different tests, it appears that 5 most selected variables are : *S_whz*, *S_washhands*, *S_haz*, *S_childpaeduc*, *S_assetspc*. *S_whz* is a measure of the nutritional status of a children compared to references, the same principle applies to S_haz, those features are indicators of growth and help identify prolonged malnutrition or chronic diseases. *S_washhands* helps to predict disease as it's a common hygienic gesture. *S_assets* is a measure of wealth per capita, it allows to estimate living conditions of an individual that reflect on its health. Using *S_childpaeduc*, it is possible to make assumptions linking education and income, but the relation between this indicator and the target is not as clear as for the others.

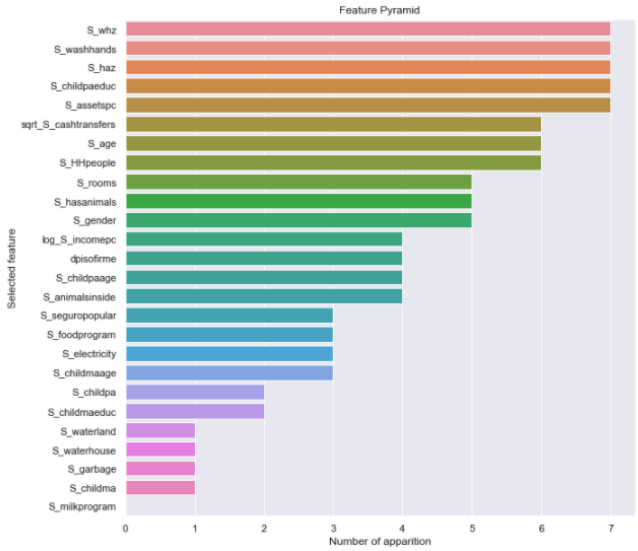

Fig. 3. Number of times each feature was selected by the one of the seven method mentioned in Section III-A

B. Feature importance

Now using the most selected features (top 15), it is possible to try to quantify the true impact on the diseases. After fitting a logistic regression to predict the health condition with the top selected features, one can see that having animals is associated with a 14% percentage-point increase of not being healthy. Gender is also associated with 17% percentage-point increase of contracting diseases. Then, height score is associated with 10% percentage-point decrease of being not healthy. Finally, father education and age are associated with respectively 10% and 12% percentage-point decrease of being ill. However having an animals is not statically important in terms of p-value for the coefficient. A closer look was given at each of these features.

Gender : By using an observational study to see the effect without confounders variables. The study reveals an average treatment effect (ATE) of 0.036 with a 95% confidence interval of $[0.013, 0.062]$. This means that, on average, boys are 3.6% more likely to be unhealthy.

Height score : A t-test to assess if there is a significant difference when the children are healthy or not was performed. The test reveals the presence of statistical evidence that a lower height score is prone to not be healthy.

Age : With the exact same reasoning, the age is also an important factor, as baby is more likely to not be healthy.

Father education : There is no statistical evidence that a higher level of education of the father impacts the health of his child.

Animals : An observational study showed that on average having an animal increases the chance to be unhealthy by 3.4% and one can be statistically confident about this result.

Washing hands : However for washing hands, it is a bit surprising that it doesn't seem to impact the children's health.

Finally, we have seen that having an animal, the age of the children, the height score and the gender of the children seems to be important in determining children's health. It's important to notice that abnormality in height score could be consequences of illnesses.

IV. Impact of Piso Firme

This section allows to have clearer picture of Pis Firme and it's potential impact on health. This is done by performing an observational study using pair matching.

A. Study design

The second research question, focused on the effect on Piso Firme on contracting illnesses, was first studied by performing a logistic regression. This regression used the features most relevant according to Section III-A. According to said regression, Piso Firme can reduce the contraction of illnesses by 9.03 percentage-point, but the coefficient was not statisticaly significant since the p-value is about 0.2. That why it was decided to go further in the investigation and to perform a pair matching to isolate the effect Piso Firme and eliminate the possible confounders. The First step is to compute the propensity score. It is computed using a logistic regression where the target is the Piso Firme feature. Then the goal is to construct a pair matching based on a similarity. The similarity between two observations is define as the propensity score difference :

$$\text{similarity(x,y)} = |\text{propensity\_score}(x) - \text{propensity\_score}(y)|$$

In fact, all subjects (treated and control) with equal propensity score have equal distribution of observed covariates. The matching was then performed by computing the similarity and only matching those below a certain ϵ. Also, for the matching to be valid, the features with biggest coefficient from the logistical regression had to be equal. This insures that any feature that could have a significant impact is removed from the comparison, allowing to better estimate the true impact of Piso Firme. After, the data was matched, the distribution between of features was observed to see if there was any significant differences between the control and treatment group. The causal effect by Rubin [4], is defined as the difference between the outcome from the treatment and control group. On an individual-level this is unobservable, as an individual cannot be in both group at the same time. But by looking at it on a population-level and taking the average of the outcome, this gives the average treatment effect which can show the causal effect.

B. Results

The data possessed 1981 samples in the treatment group, 2112 samples in the control group, of those 1906 pairs were matched and 75 samples were not matched. Identical distributions showed the data was well balanced across the both groups.

The ATE is of -0.009 with a confidence interval at 95% of $[-0.039; 0.02]$. Thus Piso Firme is responsible of a decrease about 0.9% of disease. This is an indicator that Piso Firme doesn't have a significant impact on contracting diseases. One reason for the treatment to not have a significant impact could be, that the treatment effect is heterogeneous. Treatment effect heterogeneity is when a treatment might affect different experimental subject in different ways [5]. By checking the impact of the treatment on subgroups it is then possible to verify to analyze the impact of the treatment itself.

From previous part it was discovered that some features had a bigger impact on contracting diseases. The subgroups were chosen from these more meaningful features:

- Infants: the infants group were selected as the first quartile of the age feature.
- Gender: two subgroups were created, one for male and one for female gender.
- Low income: this group was created from the 50^{th} percentile of the income per capita feature.
- Animals: if there is at least one animal on the land.

To check for this heterogeneity, the ATE was computed for each subgroup, this is called a conditional average treatment effect (CATE) as it is conditioned on the membership of a subgroup. The results of the CATE and its confidence interval are presented in Table I. Piso Firme seems to impact the gender differently. These findings are confirmed by fitting a logistic regression on the target using the important feature and add an interaction between Piso Firme and gender. It shows that in fact being a boy with the program impacts the health differently. However, coefficient is just above the threshold to be statistically significant.

Subgroups	CATE	Lower CI	Upper CI
Infants	-0.024	-0.010	0.057
Gender (Female)	0.017	-0.019	0.056
Gender (Male)	-0.034	-0.072	0.005
Animals	-0.015	-0.058	0.022
Low Income	0.012	-0.028	0.058

Table I. Results of conditional average treatment effect for each subgroup

V. Conclusion

Using multiple statistical tools, it was possible to identify the main features that seem to impact the children's health, which are the age, having an animals and the gender. Surprisingly, washing hands, the father's education and transfers per capita from government programs don't seem to have an important effect on the children's health, according to statistical tests run in this study.

A logistic regression performed on Piso firme, showed a decrease of 9.03 percentage-point of contraction of illnesses thanks to the program, however this is not statistically significant. An observational study was performed to further analyze this impact, showed that the implication of the Mexican program Piso Firme doesn't improve the children's health. However, by conducting observational studies on subgroups, it was discovered that it had a impact on young boys.

However these results came from a naive model. Thus further investigation with a sensitivity analysis could add weight to our findings.

REFERENCES

[1] M. D. Cattaneo, S. Galiani, P. J. Gertler, S. Martinez, and R. Titiunik, "Housing, health, and happiness," *American Economic Journal: Economic Policy, vol. 1, no. 1, pp. 75–105*, 2009.

[2] Global Health Data Exchange. (2000) Mexico population and housing census 2000. [Online]. Available: http://ghdx.healthdata.org/record/mexico-population-and-housing-census-2000

[3] ——. (2006) Mexico national survey of health and nutrition 2005-2006. [Online]. Available: http://ghdx.healthdata.org/record/mexico-national-survey-health-and-nutrition-2005-2006

[4] G. W. Imbens and D. B. Rubin, "Rubin causal model," in *Microeconometrics*. Springer, 2010, pp. 229–241.

[5] S. Wager and S. Athey, "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1228–1242, 2018.

[6] R. M. Hinojosa, "The future population of mexico. 123 million by the year 2010," 1988.