A score function to prioritize editing in household survey data: A machine learning approach

NICOLÁS FORTEZA AND SANDRA GARCÍA-URIBE

Summary of Banco de España Working Paper no. 2330

Household finances surveys are major public sources of information which are available for research in many countries. The Spanish Survey of Household Finances (EFF by its Spanish acronym) was one of the first to be launched in Europe. The EFF is a longitudinal survey conducted by the Banco de España (BdE by its Spanish acronym) that since 2002 provides detailed information on households' assets, debt, income and spending (Barceló et al., 2020).

The production of data for surveys like EFF are complex, as Kennickell (2017) states. The detection of data errors as omissions, implausible values or inconsistencies requires manual and human editing intervention. In the particular case of the EFF, if important errors and omissions are detected during the interview, such interview is classified as to be eligible for the recontacting of the household. It is very important to minimize this kind of data errors, since measurement error might induce important biases, especially when estimating the wealth distribution, which is typically very asymmetric (Vermuelen, 2018). Within this context, the automatization of this process is highly desirable.

Pursuing that goal, in Forteza and García-Uribe (2023) we find the best-performing machine learning algorithm that automatically classifies interviews with such substantial errors and omissions, by learning from the manual classification of cases made in previous waves. We also develop a framework to find the desirable probability threshold that classifies questionnaires into the positive (recontact) or negative (no recontact) class. To do so, the framework takes into account the acceptable amount of false negatives relative to false positives that the statistical office (or those responsible of the study) previously sets. This empirical approach might be useful for other surveys as long as they can use or exploit information from the revision and editing process in previous waves.

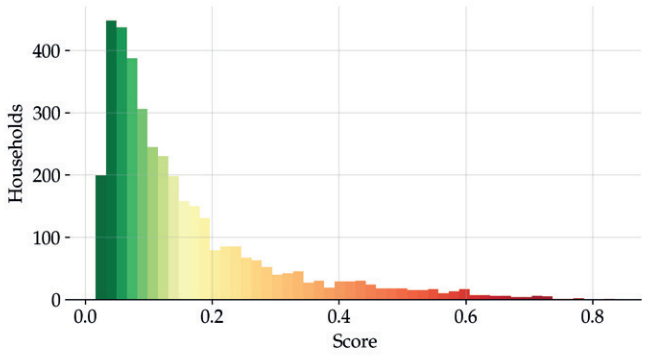
To automatically classify interviews with such kind of errors and omissions, we decide to compare over a set of algorithms which comprises classical machine learning models: Logistic Regression, K-Nearest Neighbors, Support Vector Machines, Random Forests and Gradient Boosting Trees. Using cases from previous waves, we use a target variable that indicates whether a household was recontacted or not. As an input for these methods, we use:

- Household reported answers: socioeconomic characteristics, such as number of household members, age, tenure of financial and real assets, main residence ownership regime, etc.
- Paradata: variables related to the duration, date and time of the interview.
- Comments and data filled from interviewers: using Natural Language Processing techniques, we leverage on interviewer's annotations to explore this additional dimension.
- Characteristics of the interviewer
- Error indicators and inconsistency checks that might indicate omission of assets, debts, etc.

We fine tune the models using cross validation. We find that the Gradient Boosting Trees classifier is the best performing algorithm among his competitors, achieving an average AUC ROC score of 75% and a 42% Precision-Recall AUC Score over ten test sets. An example of the output of this trained classifier can be seen in Figure 1. The graph shows the predicted probability (re-contact score) and it represents the likelihood of a household interview of having several important errors, omissions or inconsistencies.

For selecting which interviews may be subject to be followed-up with a second contact, we select households with a score above certain threshold. Given the data imbalance, a 50% probability threshold would not make sense since the predicted probability distribution is left-skewed (see again Figure 1). On the other side, increasing the threshold returns a lower false positive rate, and an

Figure 1
Household Recontact Score Histogram



Notes: Probabilities of being recontacted (recontact score) are plotted for a random subsample of households/questionnaires. The higher the score, the higher is the chance of being recontacted due to data errors and omissions along the interview/questionnaire.

increasing false negative rate. In the context of the survey, a false negative occurrence implies that a case is not flagged but it should have been since it contains errors or inconsistencies, while a false positive occurrence implies that the case was flagged but it should not be and thus, the review team would allocate additional time and resources to revise a case that does not contain substantial errors or inconsistencies. Thus, maximizing recall is relatively more important than maximizing precision. By relating the trade-off between precision and recall to the potential classification thresholds, we can explore the set of threshold values that lead to performance scores. We use the weighted harmonic mean of precision and recall with a set of varying thresholds to look at the optimal decision boundary. We call this metric the linear F-Beta score. This metric varies along a set of weights. The weight is measured as the beta parameter, accounting for the relative importance of false negatives with respect to false positives. If we believe a false negative is twice as costly as a false positive (beta = 2), we will obtain a recall of 0.71 and a precision of 0.23, which seems a reasonable gain in comparison with the current status for the EFF revision phase. In figure 1, the threshold is set at 0.14, so red areas indicate that the household should be re-contacted, where green areas indicate that the household interview is less likely to mask errors, omissions or inconsistencies.

This linear F-Beta score based framework can be extended to other statistical agencies where the data revision phase entails several timing costs. In the EFF case, one would like to evaluate the impact of selecting one or another threshold in the resulting wealth distribution. We believe that this is a worth but complex exercise that we left for future research. As De Waal (2013) also notes, at the end of the day, it is up to the statistical agency to decide the optimal threshold, assuming fixed costs (time, resources) throughout the fieldwork.

REFERENCES

Barceló, C., Crespo, L., García-Uribe, S., Gento, C., Gómez, M., and de Quinto, A. (2020).The Spanish Survey of Household Finances (EFF): Description and methods of the 2017 wave. Documento Ocasional – Banco de España N. 2033.

De Waal, T. (2013). Selective editing: A quest for efficiency and data quality. Journal of Oficial Statistics, 29(4):473–488.

Forteza, N. and García-Uribe, S. (2023). A Score Function to Prioritize Editing in Household Survey Data: A Machine Learning Approach. Documentos de Trabajo – Banco de España N. 2330.

Kennickell, A. B. (2017). Look again: Editing and imputation of scf panel data. Statistical Journal of the IAOS, 33(1):195–202.

Vermeulen, P. (2018). How fat is the top tail of the wealth distribution? Review of Income and Wealth, 64(2):357–387.