Selective editing of the occupation variable with random forests in the European Health Survey

Keywords: *Selective editing, categorical variable, random forest, European Health Survey*

1 Introduction

The use of machine learning techniques for the editing phase of the traditional statistical production process has received notable attention in the last years [see e.g. 1]. Neural networks as a long-standing machine learning technique were already proposed for data imputation some years ago [2, 3].

According to the recent Generic Statistical Data Editing Model (GSDEM) [4], three basic business function types can be identified in this production phase, namely, detection, selection, and correction functions. Imputation, in this terminology, would correspond to the third type of business functions. The extension of machine learning techniques beyond neural networks and to all business functions seems to offer a gain in cost-effectiveness, timeliness, and reduction of response burden for the production of official statistics.

Here we present a business case in ongoing production conditions with data from the Spanish branch of the European Health Survey. A two-fold objective is pursued, namely (i) to assess the use of random forest models to compute item score functions and (ii) the application of these selective editing techniques to categorical variables.

2 Methods

The starting point of our approach is the definition of an item score function as the conditional expectation of a given statistical model for measurement errors of target variables [5]:

$$s_k = d_k \cdot \mathbb{E}_m \left[\left| Y_k^{\text{raw}} - Y_k^0 \right| \mathbf{Z}^{\text{aux}} \right], \tag{1}$$

where d_k stands for the sampling design weight of unit k, m stands for the measurement error model, Y_k^{raw} denotes the raw value of variable Y for unit k, Y_k^0 denotes the true value of variable Y for unit k, and $\mathbf{Z}^{\text{aux}}$ stands for all available auxiliary variable at the moment of editing.

We shall apply definition (1) to the occupation variable in the Spanish branch of the European Health Survey since it stands as a crucial variable for multiple disaggregations of target health variables. Thus, $y_k = \delta_{kO}$ will be a binary indicator variable for each occupation value O. The measurement error will be thus a binary variable $e_k = 0, 1$ depending on both raw and true values being equal or not, respectively. As regressors, we make use of all sociodemographic variables collected for each unit (age, gender, economic activity, educational attainment, etc.).

As a working assumption true values are taken as the final validated values of each variable after the whole editing work has been conducted. In the past production system, the whole sample was manually edited through strict editing guidelines, thus rendering this assumption realistic and allowing us to use both raw and edited microdata sets to train algorithms. In the present edition, the production unit has partially conducted a parallel assessment of the traditional manual system (thus providing true values for all collected units) and of the selection of units to edit the occupation variable. It is important to point out that a stringent set of editing rules have already been applied during the collection phase before conducting this selection of units.

This parallel assessment has been implemented with an independent asynchronous concurrent model construction carried out weekly using actual microdata from the ongoing traditional production process. In this way, we can evaluate the performance of this approach in more realistic conditions for a future fully-fledged implementation of this selective editing technique.

3 Results

Equation (1) basically reduces to

$$s_k = d_k \cdot p_k, \tag{2}$$

where p_k stands for the measurement error probability, which will be predicted for each unit k with a classification random forest. Thus, we obtain the measurement error moment $s_k = d_k \cdot \hat{p}_k$, which is actually an item score for the occupation variable for each unit k.

We compare the three rankings provided by the design weights d_k, the predicted probabilities $\hat{p}_k$, and the error moments s_k (actually used in production). They are used to assign a correlative integer value $1, 2, 3, \ldots$ to each unit providing the prioritization of this variable. This procedure is applied to each weekly lot for which an updated random forest model is fitted using the latest values collected and edited during the running production process.

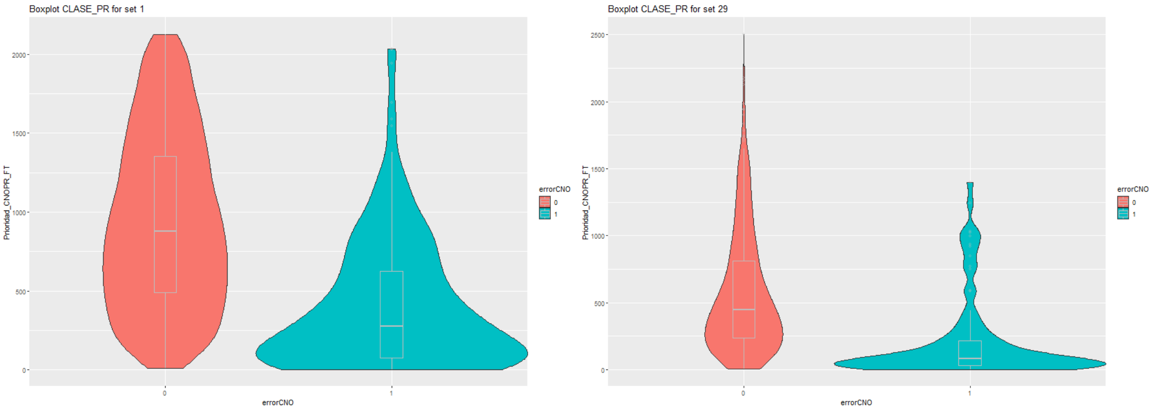
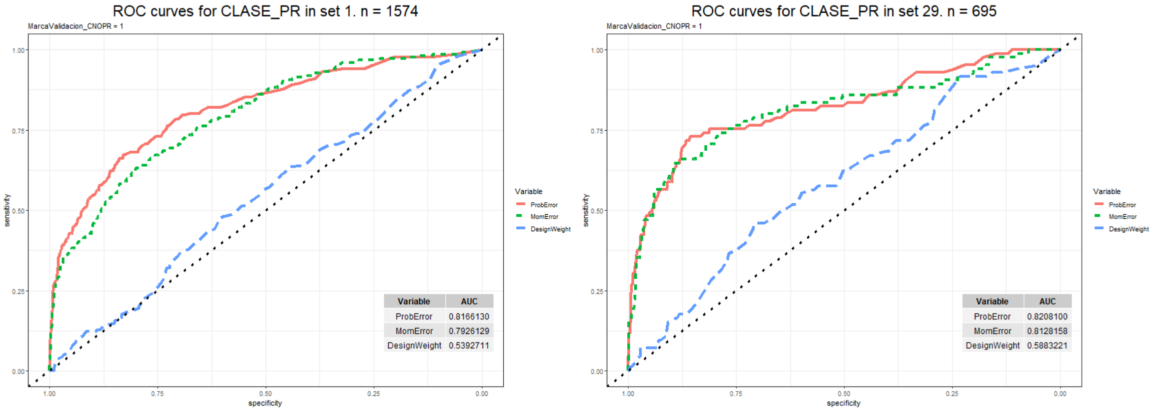


Figure 1: Violin plots for the priority number of each unit (household reference person) in lots 1 and 29 (latest as of this writing) grouped by their measurement error in the occupation variable.

In figure 1 we can observe how the model progressively learns to rank more efficiently those units with erroneous values (though some of them are not properly ranked



Variable	AUC
ProbError	0.8166130
MomError	0.7926129
DesignWeight	0.5392711

Variable	AUC
ProbError	0.8208100
MomError	0.8128158
DesignWeight	0.5883221

Figure 2: ROC curves for the three prioritizations according to d_k (design weight), $\hat{p}_k$ (error probability), and s_k (error moment) in lots 1 and 29 (latest as of this writing).

yet). In figure 2 we plot the ROC curves for the three rankings mentioned above. As expected, sorting out by the design weights is arbitrary whereas both the error probability and the error moment provide similar performance (AUC value around 0.70).

4 Conclusions

The production environment and the involvement of the production staff in the assessment of this business case allow us to reach important conclusions for a fully-fledged implementation of these ideas in the production system:

- Random forests provide a versatile statistical method both for categorical and continuous variables in the context of selective editing providing thus a unified framework for item score computation.

- The models should be enriched to achieve higher performance by seeking more regressors (e.g. paradata). Natural language processing techniques seem to offer an important possibility for incorporating valuable quantitative information from comments by respondents and data collection/editing clerks to both each item and the whole questionnaire.

- Production workflows must be taken into account when considering the application of these methods since statistical units are collected in lots thus providing a concurrency of model building, data collection, and data editing. These three business functions must be integrated for an optimal result.

- Once the prioritization of each lot is taken into account, the assumption of having true values for units below the editing threshold is only based on the model performance (thus possibly explaining those erroneous values not properly ranked). This raises the far-reaching issue of having true values for non-edited units and the algorithm training with a partial edited dataset. Further research is needed in this line.

References

[1] UNECE. Workshop on Statistical Data Editing, 2020. 31 Aug – 4 Sept, 2020.

[2] S. Norbotten. Editing statistical records by neural networks. *Journal of Official Statistics*, 11:391–411, 1995.

[3] S. Norbotten. Neural network imputation applied to the norwegian 1990 population census data. *Journal of Official Statistics*, 12(4):385–401, 1996.

[4] UNECE. *Generic Statistical Data Editing Model*, 2020. https://statswiki.unece.org/display/sde/GSDEM.

[5] I. Arbués, P. Revilla, and D. Salgado. An optimization approach to selective editing. *Journal of Official Statistics*, 29(4):489–510, 2013.