A Score Function to Prioritize Editing in Household Survey Data: A Machine Learning Approach

Nicolás Forteza and Sandra García-Uribe

Bank of Spain

Introduction

Data

Empirical Strategy

Results

Conclusion

Any views or opinions expressed are solely those of the author and do not necessarily represent those of the Bank of Spain

Motivation

- The *Spanish Survey of Household Finances* (EFF) is a large-scale survey and a complex statistical operation.
- Main output of the survey are wealth and income related statistical distributions; very asymmetrical.
- Data editing is a major task in the production process of survey data.
- Due to asymmetry in wealth and income, every case must be reviewed. The revision team manually checks the consistency among questions while incorporates the information of interviewer comments and audio records to edit the data if necessary.
- However, the manual revision process enteails several costs, i.e. time and measurement error.

This Paper

- We find the best-performing machine learning algorithm that classifies interviews with substantial errors.
- We show that the scores of the algorithm predict with high accuracy the categories in different test sets.
- We contribute to the survey data production literature in two dimensions:
  - By providing a prediction engine that shortens revision times.
  - By providing a framework for statistical data editing for statistical agencies to edit data more efficiently.
- Our strategy is applicable to other surveys with existing previous manual classification data.

Related Literature

- The SCF (Survey of Consumer Finances by the US Federal Reserve) has made many efforts in quantifying the editing costs Kennickell (2006, 2007, 2017) in financial household surveys.
- However, there is scarce literature regarding score functions to identify interviews with substantial errors; only in business surveys and focusing in the error generation process (De Waal (2013); Arbués et al. (2013))
- Machine learning is starting to take a role in the production of survey statistics: to reduce panel attrition (Kern et al. (2021)), to generate data more efficiently (Schierholz and Schonlau (2020), to find errors in text data (He and Schonlau (2021)).
- Kern et al. (2019) documents that tree-based models are often used due to its flexibility in these set ups.

Background

- EFF Survey data revision is a long process which lasts through all the production span of a wave.
- The initial phase coincides with the beginning of the field where the revision team has to analize all interviews to detect potential mistakes or omissions.
- When there are a number of errors or omissions in an interview which cannot be solved with the available information, the interview is classified as one that requires to recontact the household.
- This implies that the household would be asked about the parts of the questionnaire that are affected by errors in a second telephone contact.
- The second telephone contact can only happpen once the classification has been confirmed by the team: it involves a double check by the both the field company and, in this case, Bank of Spain.

A (real) Recontact Example

- A certain household is interviewed. After running the inconsistency detection and tabulation programs, no errors are detected.
- In the last section (consumption and savings), the household is asked for other inheritance or received gifts they could've had. The household reports *"Doesn't know"*.
- However, the interviewer annotates: *"Inherited 2 properties and owns 20% of them, and the household rents both of them"*. This is an ommited income and asset! The household should've reported these properties in another previous section.
- The household is recontacted to know these properties' value and potential rental income.
- Also, during the recontact, the household reports that the income of that rent accounts for the whole main residence expenses!

Goal

- To find a prediction model that allows to classify raw and incoming survey data into a "potential recontact" class.
- We exploit previous manual classification work from other waves to train the algorithms.
- We exploit suvey data together with paradata, interviewers notes, consistency rules and response times to inform the prediction models.

Data

EFF 2017 and 2020 raw data and paradata.

- Dependent variable: binary outcome for each interview:
  - 0 = No recontacted household (majority class)
  - 1 = Recontacted household (minority class)

	EFF17	EFF20
0	5049	5577
1	1380	746

Table: Recontacts Distribution

Features I

Source	Variables
Household answers	Acceptance of being audio-recorded in certain parts of the interview, whether the household is a panel unit or not and the use of a proxy person to respond at the interview. Number of household members, educational level of reference person, sex of reference person, auto-perceived satisfaction with life of reference person, main residence ownership regime, number of other properties, type of these other properties, estimated value of these other properties, total number of contracted loans by household, number of bussinesses related to self-employment, holdings of unlisted shares, holdings of listed shares, holdings of investment funds, holdings of fixed income investments, total number of pension funds.

Features II

Source	Variables
Paradata	Number of Euros (closed and interval) questions, number of times a questions was asked, non response ratios, total seconds per section, total repeated number of questions per section, total seconds when numerous categories are asked, number of total interviews performed by the interviewer prior to the contact, whether is weekend or not, number of days since the start of the field work, time slot of the day.
Comments from the interviewer	Total number of opened comments by interviewer, mean length of comments, top words from NLP data pipeline.

Features III

Source	Variables
Other paradata filled by interviewer	Type of residential unit, condition of residential unit building, conditions of surrounding buildings, perceived wealth in surrounding area of residential unit, perceived level of questions understanding by the household, whether the household was mistrustful before and after the interview, number of people in the room when the interview was held, whether the household consulted external documents during the interview, where the interview was held, motives of acceptance of the interview.
Characteristics of the interviewer	Number of previous survey editions, seniority at field work company, normalised score at the training programme, participated in ECF Survey, educational level.
Error indicators and Inconsistencies	See Table 4 of the Annex for details.

Descriptive Statistics

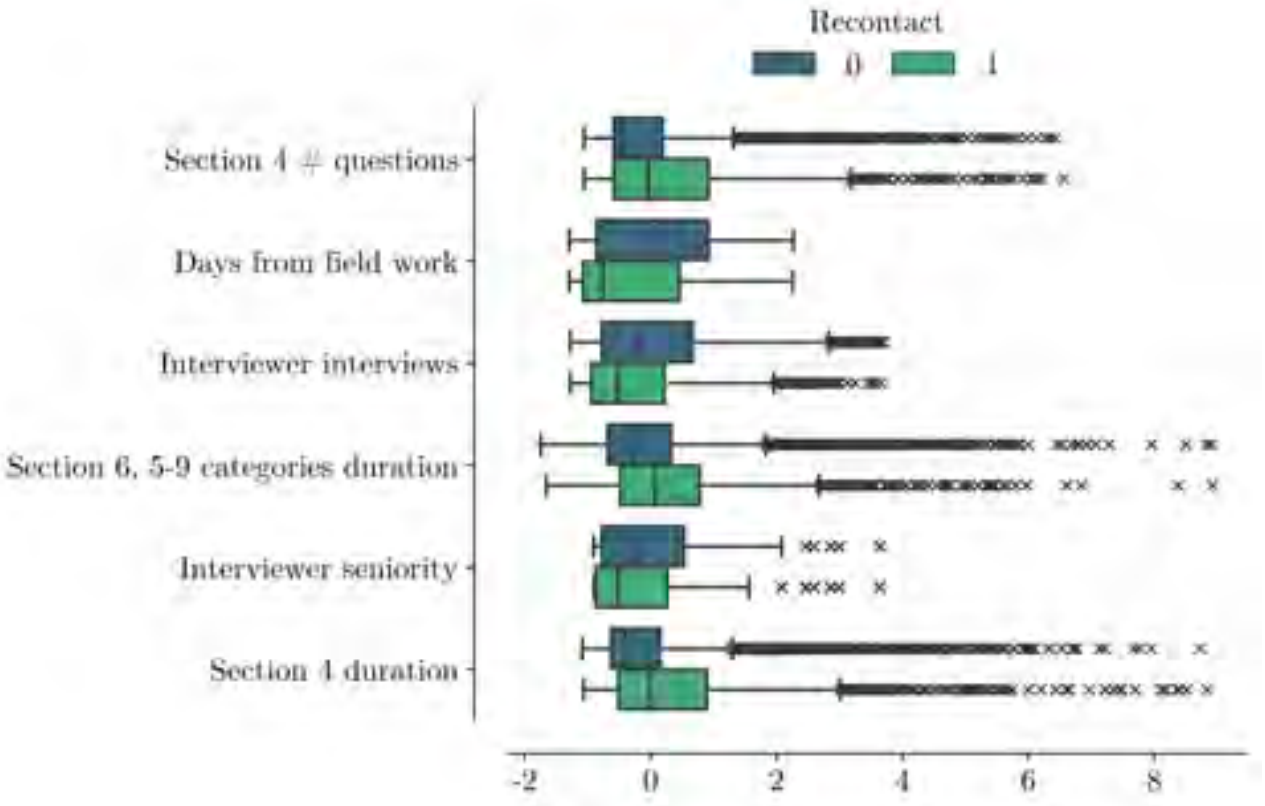


Figure: Some Numerical Variables

NLP Pipeline

1. Parse text data with pre-trained models from Honnibal and Montani (2017), removing stopwords, punctuation signs, alpha numeric characters, and others.
2. Apply Porter (2001) stemming and produce word counts.
3. Select words that are more present in each class (top 20 words with highest importance within that class).

TF-IDF or other techniques as an input (SVD) not showing any improvements.

Machine Learning Classifiers

We want to find the best-performing model among the more commonly used machine learning classifiers in the literature, these are, sorted by capacity:

- Logistic Classifier with L1 penalty.
- K-Nearest Neighbors Classifier.
- Support Vector Machines.
- Random Forests.
- Gradient Boosting Trees (XGBoost).

There's been a recent increment in the use of bagging techniques with decision trees in the survey research community Buskirk (2018).

Model Training

Due to the sample size, we cannot offer a large test set. Thus, to find the optimal hyperparameters and estimate the models we use the following algorith using 10 different random seeds:

1. Split the dataset into 70% train and 30% test using stratified random sampling.
2. 5-fold stratified cross validation for hyperparameter tunning. Cross validation aims to minimise the log-loss function:

$$L_{\log}(y, p) = -(y \log(p) + (1 - y) \log(1 - p))$$

   where p is fitted probability of being recontacted, y is observed (target variable).
3. Retrain the model with the best hyperparameters on the whole training data set.

To make predictions or inferences, we use median of all 10 fitted best performing model.

◂ Hyperparams optimisation strategy

Model Evaluation

To select the best performing model, we average over the 10 test-set resulting areas under the curves (AUC) for the following two trade-off curves:

- False positive Rate and True Positive Rate (Recall) i.e. *ROC or receiving operating characteristic*.
- True Positive Rate (Recall) and Precision; i.e., the *Precision-Recall curve*.

Due to the dataset imbalance, the precision-recall curve captures better the tradeoff between false positives and false negatives.

Optimal Threshold

Decision framework for statistical agencies:

- Maximizing recall (FN = errors and inconsistencies in data, distorting final survey figures) is relatively more important than maximizing precision (FP = cost and time in allocating resources)
- We use the weighted harmonic mean of precision and recall with a set of varying thresholds to look at the optimal decision boundary.

$$F_l = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}} = \frac{\beta + 1}{\frac{\beta}{recall} + \frac{1}{precision}} = (1 + \beta) \cdot \frac{precision \cdot recall}{(\beta \cdot precision) + recall} \quad (1)$$

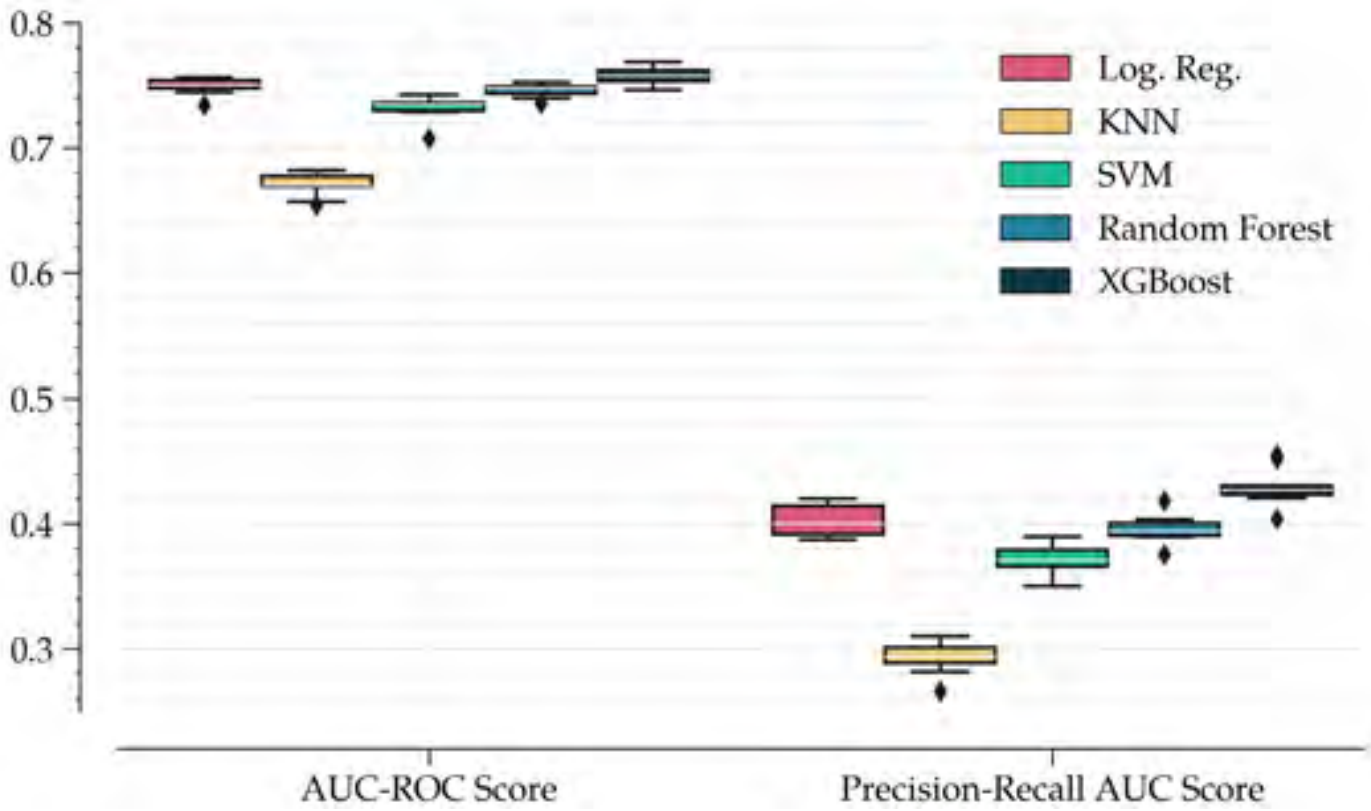
- Our modified linear version allows us to interpret the trade-off between precision and recall, as opposed to the general, non-linear, formula. In our modified version, β is the relative weight of recall with respect to precision.

Best Model I

Table: Main metrics (mean values of 10 random data splitting initializations)

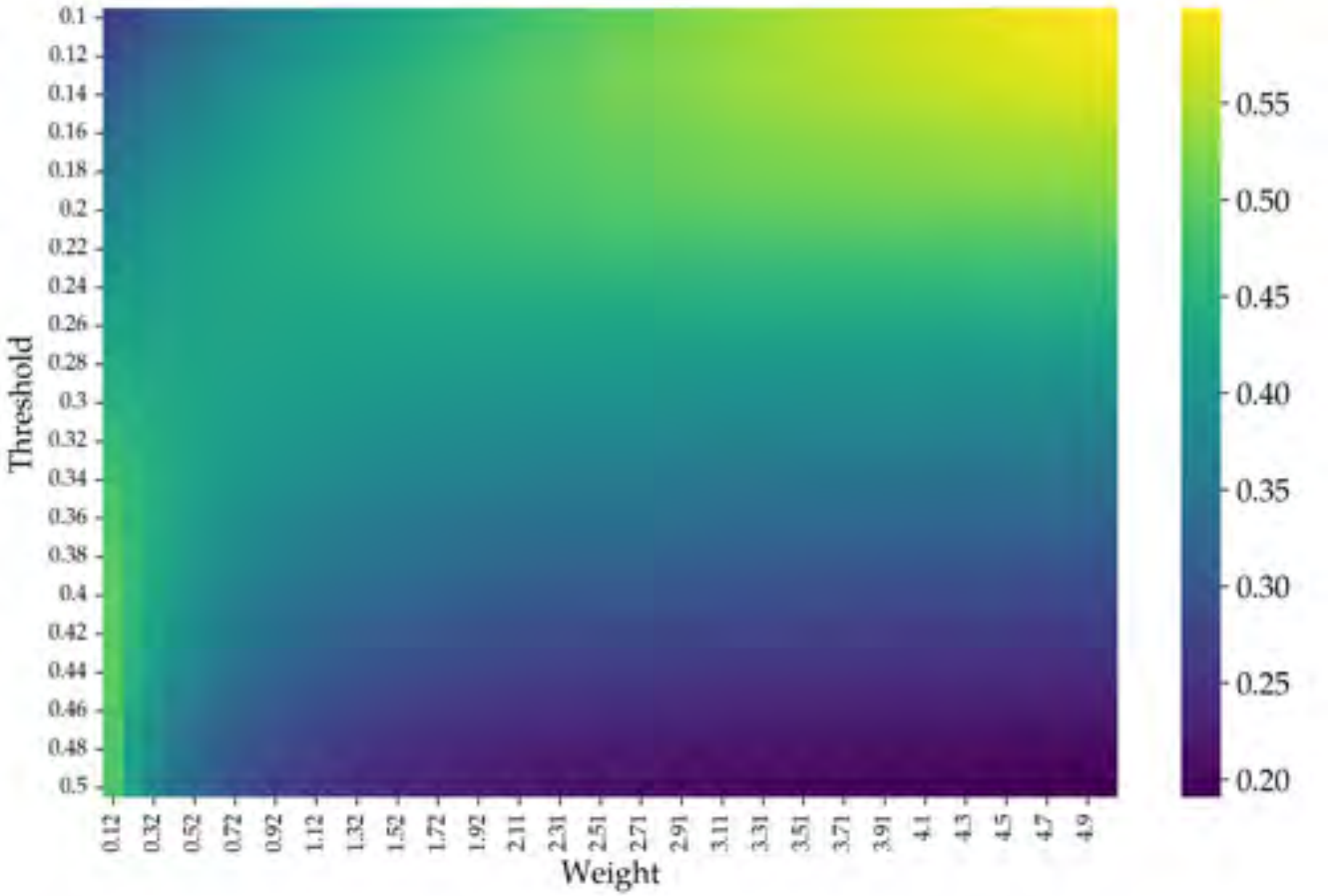
	AUC-ROC Score	Average Precision Score	Precision-Recall AUC Score
K Neighbors	0.670	0.286	0.293
Logistic Clf.	0.749	0.403	0.403
Random Forest	0.746	0.397	0.396
SVM	0.732	0.372	0.371
XGBoost	**0.758**	**0.430**	**0.429**

Best Model II



Optimal Threshold I

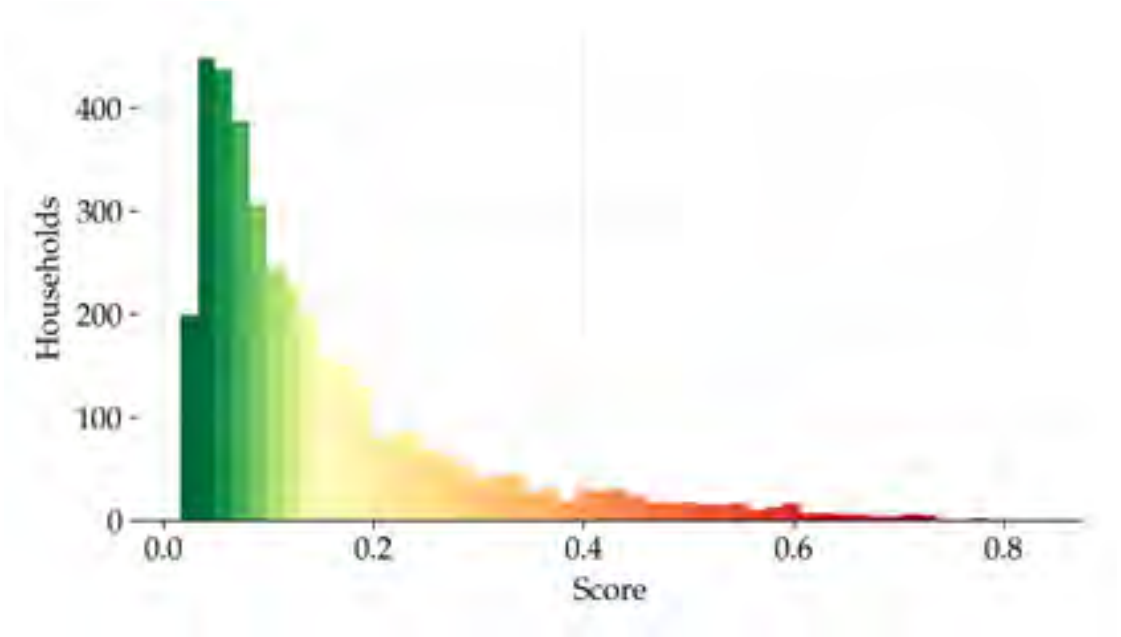
Figure: Linear F-Beta Score - weighting scheme for Gradient Boosting Classifier



Optimal Threshold II

In practice, the revision team will review data according to the score that masks problems and errors within the questionnaire:

Figure: Score Histogram (gradient boosting trees) when $\beta = 2$



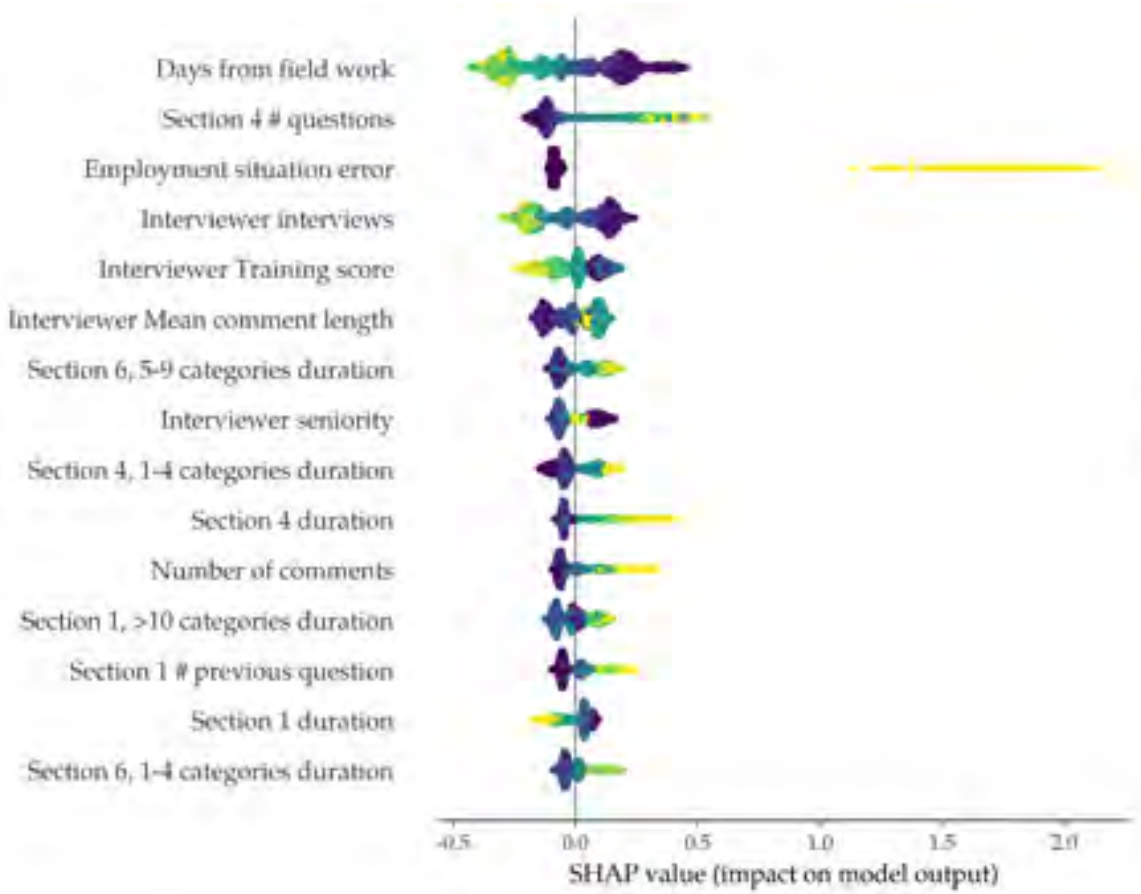
Main Results

Conclusions from test evaluation results:

- The best performing model is the gradient boosting trees (in line with literature)
- Data generation process depends on the extensive depth of the logical tree of the questionnaire. Huge heterogeneity and variability in the data.
- In our case, when $\beta = 1$, optimal threshold is set at 0.22, and 0.14 when $\beta = 2$.

Interpretability

We also measure interpretability with the framework developed by Lundberg and Lee (2017) to look at the most determinant features of the model that impact on the prediction.



- Each dot represents an observation.
- The brighter the color, the higher is the value for that observation-feature combination.
- The SHAP value indicates the log odds contribution of that feature value to the predicted probability.

Validation with EFF2022 Data

- The revision team manually classified EFF2022 incoming data and now we have a fraction of the data to validate the tool now.
- We compare the performance of the top-performing three models AUC metrics.
- The Gradient Boosting Trees algorithm (XGBoost) presents again a superior performance.
- This means that the model was properly calibrated and generalises well in an out of sample basis.

Table: Evaluation over the EFF2022 Field

	ROC AUC	PR AUC
Gradient Boosting Trees	0.723	0.257
Logistic Classifier	0.716	0.264
Random Forest	0.703	0.263

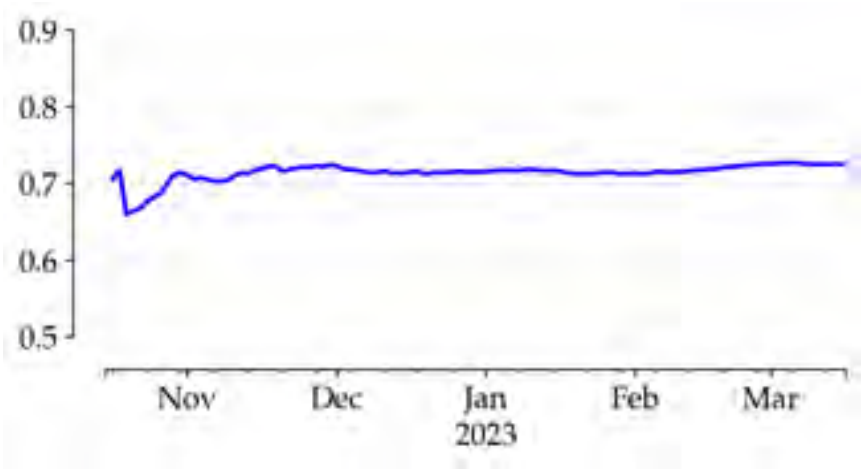


Figure: Cumulative AUC ROC Score for Gradient Boosting Trees

External Factors

There is a part of the prediction errors of the model that cannot be explained with any additional controls. These errors are very much explained by the *reviewer* and *wave effects*: unobservables at the training stage.

	Dependent variable:			
	Log-Loss Error			
	(1)	(2)	(3)	(4)
Coefficient	0.360***	0.377***	0.522***	0.591***
	(0.013)	(0.065)	(0.073)	(0.013)
Interviewer FE	Yes	Yes	Yes	Yes
Regional FE	No	Yes	Yes	Yes
Reviewer FE	No	No	Yes	Yes
Wave FE	No	No	No	Yes
Observations	12,573	12,573	12,573	12,573
Adjusted R^2	0.045	0.046	0.204	0.204
Note:				*p<0.1; **p<0.05; ***p<0.01
				Reportado con Clustered Standard Errors

Conclusions

- We provide a model that automatically classifies household survey data with substantial errors.
- The data production benefits mainly from paradata, interviewer characteristics and inconsistency indicators.
- We show that the gradient boosting trees algorithms outperform the rest of models, even when the data is imbalanced.
- We also show that the model is robust and stable out of sample, displaying similar performance results over the new EFF2022 field work data.
- By using *linear* F-Beta score, we provide a framework for statistical agencies for selective editing.

Future Research

- Deep Learning models do not outperform with tabular data, but Transformer embeddings (Vaswani et al. (2017)) could be used as to generate new (although not interpretable) features.
- Audio features could also improve model performance.
- Temporal cross validation could bring improvements in the model performance, as in Kern et al. (2021).
- Train a task-based AI with Large Language Model (GPT) to edit cases.

Thank you

Bibliography I

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

Buskirk, T. D. (2018). Surveying the forests and sampling the trees: An overview of classification and regression trees and random forests with applications in survey research. *Survey Practice*, 11(1).

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

He, Z. and Schonlau, M. (2021). A Model-Assisted Approach for Finding Coding Errors in Manual Coding of Open-Ended Questions. *Journal of Survey Statistics and Methodology*. smab022.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Kennickell, A. B. (2006). Who's asking? interviewers, their incentives, and data quality in field surveys. *Survey of Consumer Finances Working Paper, SCF Web Site: http://www. federalreserve. gov/pubs/oss/oss2/scfindex. html*.

Kennickell, A. B. (2007). Look and listen, but don't stop: Interviewers and data quality in the 2007 scf. *Proceedings of the Survey Research Methods Section. American Statistical Association. www. amstat. org/Sections/Srms/Proceedings/y2007/Files/JSM2007-000648. pdf*.

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

Kern, C., Klausch, T., and Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, 13:73–93.

Bibliography II

Kern, C., Weiß, B., and Kolb, J.-P. (2021). Predicting Nonresponse in Future Waves of A Probability-Based Mixed-Mode Panel With Machine Learning. *Journal of Survey Statistics and Methodology*. smab009.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Porter, M. F. (2001). Snowball: A language for stemming algorithms. Published online. Accessed 11.03.2008, 15.00h.

Schierholz, M. and Schonlau, M. (2020). Machine Learning for Occupation Coding—A Comparison Study. *Journal of Survey Statistics and Methodology*, 9(5):1013–1034.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Table: Some Error indicators and inconsistencies

Name	Error indicator description (whether the household or any member...)
Employment History	Declared to have worked the year prior to the interview, but worked less than 12 months
Panel Error	Do do not have any panel member and viceversa, when in fact should have.
House Mortgage	Declare that the mortgage amount is higher than main residence value
House Mortgage	Declare that the mortgage amount is higher than the initial mortgage amount
Other properties loan	Declare that the other properties pending loan is higher than the initial loan amount
Main Residence Loan Term	Declare that the remaining term is higher than the initial declared loan term
Other Properties Loan Term	Declare that the remaining term is higher than the initial declared loan term
Main Residence Monthly Amount	Declare that the monthly payment is higher than the pending amount
Other Properties Monthly Amount	Declare that the monthly payment is higher than the pending amount

Hyperparameter Tuning Strategies

Each classifier gets a different hyperparameter tuning strategy:

Classifier	Strategy
Logistic	Grid Search over L1 penalties
K-Nearest Neighbors	Grid Search over K values
Support Vector Machines	Grid Search over kernel type and penalties
Random Forest	Random Search (2500 iterations)
Gradient Boosting Trees	Random Search (2500 iterations)

Table: ML Classifiers and Hyperparameter Tuning Strategy