A SCORE FUNCTION TO PRIORITIZE EDITING IN HOUSEHOLD SURVEY DATA: A MACHINE LEARNING APPROACH

2023

BANCO DE ESPAÑA
Eurosistema

Documentos de Trabajo
N.º 2330

Nicolás Forteza and Sandra García-Uribe

A SCORE FUNCTION TO PRIORITIZE EDITING IN HOUSEHOLD SURVEY DATA: A MACHINE LEARNING APPROACH

A SCORE FUNCTION TO PRIORITIZE EDITING IN HOUSEHOLD SURVEY DATA: A MACHINE LEARNING APPROACH [*]

Nicolás Forteza

BANCO DE ESPAÑA

Sandra García-Uribe

BANCO DE ESPAÑA

(*) Corresponding authors: Nicolás Forteza (nicolas.forteza@bde.es) and Sandra García-Uribe (sandra.garcia.uribe@bde.es). We thank the incredible help of Paloma Urcelay and Younes Aberkan El Hajui, as well as the numerous helpful comments and suggestions made by an anonymous referee, Cristina Barceló, Olympia Bover, Laura Crespo, Carlos Gento and Ernesto Villanueva, and for the work of the Spanish Survey of Household Finance team, the one producing the data we use. We also thank seminar participants at the 2021 Big Data & Data Science online conference by the Centro de Estudios Monetarios Latinoamericanos (CEMLA), Banco de España and the 2023 Symposium on Data Science and Statistics by the American Statistical Association at St. Louis (Missouri) for their invaluable comments. The views expressed in this paper are our own and do not necessarily reflect the views of the Banco de España or the European System of Central Banks.

Documentos de Trabajo. N.º 2330

October 2023

https://doi.org/10.53479/34613

ISSN: 1579-8666 (on line)

Abstract

Errors in the collection of household finance survey data may proliferate in population estimates, especially when there is oversampling of some population groups. Manual case-by-case revision has been commonly applied in order to identify and correct potential errors and omissions such as omitted or misreported assets, income and debts. We derive a machine learning approach for the purpose of classifying survey data affected by severe errors and omissions in the revision phase. Using data from the Spanish Survey of Household Finances we provide the best-performing supervised classification algorithm for the task of prioritizing cases with substantial errors and omissions. Our results show that a Gradient Boosting Trees classifier outperforms several competing classifiers. We also provide a framework that takes into account the trade-off between precision and recall in the survey agency in order to select the optimal classification threshold.

Keywords: machine learning, predictive models, selective editing, survey data.

JEL classification: C81, C83, C88.

Resumen

Los errores en la recopilación de datos de las encuestas financieras de los hogares podrían propagarse y afectar a las estimaciones poblacionales, sobre todo cuando existe un sobremuestreo de algunos grupos de población. Hasta ahora se han realizado revisiones manuales de cada entrevista para identificar y corregir los posibles errores y omisiones, como es el caso de los activos, ingresos o deudas omitidos o recogidos con información errónea. En este trabajo se ofrece un enfoque de aprendizaje automático para clasificar aquellos datos de encuestas que presentan errores y omisiones importantes durante la fase de revisión. Utilizando datos de la Encuesta Financiera de las Familias, se muestra el mejor algoritmo de clasificación supervisado con el fin de priorizar tales casos. Asimismo, se demuestra que con un modelo *Gradient Boosting Trees* (árboles de potenciación del gradiente) se obtienen mejores resultados que con otros clasificadores. Finalmente, se proporciona un marco que tiene en cuenta la disyuntiva entre precisión y exhaustividad *(recall)* en la entidad encuestadora para escoger el umbral óptimo de clasificación.

Palabras clave: aprendizaje automático, modelos de predicción, edición selectiva, datos de encuestas.

Códigos JEL: C81, C83, C88.

1 Introduction

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). As Kennickell (2017) documents the production of household finance data is complex. Survey data editing is time-consuming and costly and takes a substantial part of the production process. Automatic data editing methodologies alleviate some of the costs, through the identification of demographic inconsistencies which are easy to programme and identify. However, the detection of other type of data erros such as omissions, implausible values or inconsistencies is much harder to program exhaustively and requires manual editing intervention. This represents a challenge for the data production process because measurement errors can propagate and affect several variables along the interview given the complexity of the questionnaire. For example Kennickell (2006) argues that not editing this kind of data may have important consequences for what is the resulting wealth distribution whereas Vermeulen (2018) poits out that measurement error might induce important biases, specially when estimating the wealth distribution, which is typically very asymmetric. Thus, manual case-by-case revision has been so far applied in order to identify and correct potential errors and omissions.

In the particular case of the EFF, if important errors and omissions are detected during the revision of a case and those cannot be solved with the available information (audio records for some questions or interviewer comments), such interview is classified to be eligible for the recontacting of the household. This implies that the household respondent receives a new telephone call where she would be re-asked about the parts of the questionnaire that are affected by errors and any new information would be incorporated in the corresponding revision of the case. This task has been done in every wave since 2002 because has been proved to be crucial to increase the accuracy and quality of the data. However, as it was mentioned above, this manual classification is extremely time consuming so that improving the automatization of the process is highly desirable.

Pursuing that goal, in this paper we find the best-performing machine learning algorithm that classifies interviews with such substantial errors and omissions in this survey editing set up by learning from the manual classification of cases made in previous waves. We decide to compare over a set of algorithms which comprises classical machine learning models (Logistic Regression, K-Nearest Neighbors, Support Vector Machines) and the well-known tree-based algorithms (Random Forests and Gradient Boosting Trees) given that there is no prior on which of these models would be better in this set up. The best performing algorithm, a Gradient Boosting Trees, outputs a score function assigning a probability of being a case with substantial errors (or to be recontacted) to each questionnaire given a large set of covariates. We show that the algorithm predicts with high accuracy the manual classification using different test sets. The modelprovides a AUC-ROC above 75%, which is in-between a random classifier and a perfect classifier. We believe that this score is not low given the complex data generation process behind the phenomenon we are studying. First, the data from the outcome variable is imbalanced. Second, our setting is subject to a relatively larger number of variables than observational units, which is due to the fact that the questionnaire logical is unique for each household. In addition, our trained classifier is able to achieve an average precision of 23% and an average recall of 71% for multiple test sets. Using new data from an incoming EFF wave, we observe that the performance of the selected model is stable on new data and robust to a refinement of the target variable. We also show that a considerable part of the prediction errors are related to unobservable variables at the time of the prediction. Lastly, since the best model is not directly interpretable, as might be the coefficients of a regression, we use the SHAP

interpretability framework to interpret the rationale behind the estimated model; it helps a reviewer to interpret and explain the score asigned to each case, which also builds trust in the results of the model.

This methodology also allows us to characterize what is the desirable probability threshold that classifies questionnaires into the positive or negative class. To do so, it takes into account the acceptable amount of false negatives relative to false positives that the statistical office (or the responsibles 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. In particular, it provides an automatically-generated score that increases the efficiency of the manual case-by-case revision process giving priority to those cases more likely or prone to contain errors. In this sense, the use of this type of score might be especially useful in cases where not massive manual revision can be performed because of limited funding since just a small classification exercise to train and test the model is needed. Furthermore, the score can also be informative about over and under-editing, which is a crucial to monitor and discipline the data editing process.

We contribute to the survey methodology literature with an empirical tool to automatically identify cases with substantial errors, saving manual revision time. In this sense, our paper speaks to previous literature on selective editing (De Waal, 2013; Arbués et al., 2013) by providing a framework within the machine learning literature where the research team can set the acceptable amount of false negatives given the trade-off with false positives. Our work also speaks to the literature on edit prioritization using score functions (Latouche and Berthelot, 1992; Allard et al., 2001; Hedlin, 2003; Gismondi, 2007).

Recently, the application of machine learning techniques in survey methods research has spread. The are works in forecasting panel attrition (Kern et al., 2021), in modelling unit non-response (Toth and Phipps, 2014; Kern et al., 2021, 2019), to find errors in textual data (He and Schonlau, 2021), to classify coding errors (Schierholz and Schonlau, 2020) and in classical imputation methods (Dagdoug et al., 2021). In general, these techniques are becoming increasingly important in the survey data production process. In this sense, we contribute to this literature by providing an empirical approach and the optimal model given the data to produce scores that allow to prioritize revision. We find the best-performing model in our setup is a Gradient Boosting Trees.

The paper is organised as follows. In section 2 we discuss some research done in this area and the use of machine learning techniques in data editing. Section 3 describes the data and methodology used to approximate the recontact score function, and in section 5 we disentangle the results for the in-sample and out of the sample data. We also briefly discuss about the main advantages and caveats of this methodology in the last section.

2 Background

The Spanish Survey of Household Finances (EFF) is a survey conducted by the Banco de España since 2002 that provides detailed information on income, assets, debts and spending of Spanish households. The majority of questions refer to the household as a whole except for labor and related income that refer to each particular household member over the age of 16. Most of the information makes reference to the time of the interview, although information on all pre-tax income sources is also referring to the previous calendar year. The information is collected through personal interviews with households, conducted by interviewers with specific training and computer-assited (CAPI).[1] The EFF fieldwork lasts around 9 months starting at October of the corresponding wave year.

Throughout the data production process, which starts immediately after the beginning of the fieldwork, a number of data quality control and validation tasks are carried out; see Barceló et al. (2020) for

[1] In 2020, due to the pandemic context, interviews were conducted by telephone (CATI).

a detailed description of the EFF methodology. In addition to many consistency (hard and soft) checks that are programmed in the CAPI instrument to minimise different types of errors (values out or range, implausible values and inconsistencies), BdE together with the field company (FC) conduct an extensive manual revision process of all completed interviews. On top of this, interviewer's work is closely supervised not only regarding response rates but also in terms of data quality.

The revision process is iterative between the FC and BdE, Figure B.1 summarizes it. In the first part of the revision, the revision team at the FC reads all completed questionnaires and flags errors, e.g., implausible values, coding errors, inconsistencies, monetary errors, and omitted information. Comments and clarifications entered during the interview by interviewers, in addition to audio records, are also useful sources for reviewing and checking collected data. Given the strong impact that major errors and omissions in the data can have on the properties of the measures collected, the FC also tags cases affected by those - priority cases versus no-priority cases- so that the BdE team can perform a revision of the priority cases[2]. If in the second revision of a priority case, errors or omissions that cannot be solved either with the available information, the BdE requests the survey agency to recontact the household to clarify responses and collect important omitted information. The type of omissions that usually lead to a recontact are, for example, unreported labor status, omitted income, omitted real state assets or debts, incorrect valuation of business, mistakes in household composition. The application presented in this paper aims at achieving a good-performance ML classifier that automatically tags a case requiring a recontact without the need of the previous manual revision of both BdE and the FC.[3]

The main reason why we do not target the first manual revision made by the FC is that, so far, the coding made at this first round is not available. In addition, one could think that the manual classification of recontacts may have measurement error as it is the result of human work. To avoid the possibility of coding mistakes in the first revision made by the FC, BdE makes also random revisions of no-priority cases in order to monitor and guide the work of the FC. There are also posterior tabulations and checks which provide further filters to correct for substantial errors and omissions that may have not been detected in the manual revision stage. In this part of the data production, the team at BdE did not find a significant portion of cases requiring a recontact that were not identified at the revision phase. Thus, we can consider the manual coding of recontact as a reliable variable to learn from.

Manual editing of survey data is an important but time-consuming task which influences the timeliness of the publication of the data. In addition, in the case of the EFF, it also involves recontacting respondents to correct substantial errors and omissions which further increases the editing burden. Selective editing comprises a set of techniques which are applied to intermediate data in order to prioritize editing to those cases with influential errors. To our knowledge this strand of the literature has focused in prioritizing cases based on their influence in some expected result (Latouche and Berthelot, 1992; Allard et al., 2001; Hedlin, 2003; Gismondi, 2007). Most of its applications are for establishment and census survey data. In household finance surveys, it is not trivial to define what an influential error is beforehand. Our approach prioritizes based on the likelihood that there are substantial errors and omissions that lead to recontact respondents, as opposed to with respect to certain expected result from the data. In this sense, this work falls into the set of micro-selection approach, in particular, it is a prediction model approach, for a review of this literature see De Waal et al. (2011) and Granquist and Kovar (1997). So far, this tool is unlikely to substitute all manual revision given that cases with no priority still need editing and there is no automatic editing tool covering the correction of that part of the data. However, our tool

[2]BdE also revises many non-priority interviews coming from interviewers that require close monitoring at the beginning of the field in order to give feedback and correct interviewer protocol or conceptual mistakes

[3]The recontact consists of a phone call to the household respondent to made a shorter questionnaire that is focused on those aspects that need to be revised or corrected. The EFF has given careful consideration to the trade-off between obtaining additional information and bothering households on a case-by-case basis. One of the advantages of recontacting households is that the overall measurement error of the survey is considerably reduced. In addition, the representativeness of the sample is better ensured, as fewer cases/questionnaires have to be discarded.

will prioritize the revision of cases with substantial errors, leading to a reduction in the time to recontact a respondent if needed, it will also be a guide for new editors to identify a priority case. Further research should complement this work with the study of a way to further reduce manual revision.

3 Data

The outcome to be predicted takes value 1 if the case needs a recontact i.e. it contains substantial errors/omissions that require to recontact the household, and 0 otherwise, that is, if it does not have substantial errors and omissions in the first revision by the FC or the FC flagged it with priority but the second revision at BdE did not consider necessary to recontact the household. We use data from the manual revision process of two previous survey waves, namely EFF2017 and EFF2020 [4]. Table 1 presents the distribution of cases among the recontact indicator in each wave.

Table 1: Distribution of Recontacts

	EFF17	EFF20
0	5049	5577
1	1380	746

Explanatory variables come from multiple datasets, mainly, questionnaire responses, paradata, and metadata. Table 2 presents a description of each set of inputs. From reviewers experience, household financial characteristics tend to be informative in identifying problems in the data, e.g. households with complex financial structures are more likely to require follow-up contact. Additionally, it is known that interviewers with less experience and training in conducting complex surveys tend to generate lower quality interviews. Bristle et al. (2019) found that interviewer characteristics, such as education level and experience, may serve as good predictors of panel co-operation. Interviewer effects on household answers have been previously studied (Durrant et al. (2010), Flores-Macias and Lawson (2008)). Moreover, paradata such as the time taken to answer a question, can provide insights into the cause of recontact and can be extremely useful in the production of survey statistics (Groves and Heeringa, 2006). Thus, the set of predictors used in this application includes both household and interviewer generated data and characteristics. Additionally, we exploit text data from interviewers' comments and clarifications introduced during the interview with the CAPI software, a novel source of data that, to our knowledge, has not been exploited in the literature. Such comments are very useful in the editing process as they help to decide if the question has been answered well and/or if the interviewer has made a mistake in asking the question. Appendix C explains the details on how we exploit the data. We also generate other set of predictors, such as total number of comments and mean comment length. Lastly, we incorporated a set of automated indicators for errors and inconsistencies that are used to correct data, after the manual revision, in a final data editing stage, see Table A.1 of the Appendix for details. Overall, our set of predictors consists of approximately 275 variables.

[4] It should be noted that the EFF2017 survey saw several significant methodological changes, which resulted in the modification of the revision process in comparison to previous survey waves. One notable change involved the recording of audio data during certain sections of the interviews, among other methodological changes introduced from EFF17 onward. Additionally, the storage of the to-be-recontacted flag was included as part of these changes

Table 2: Description of variables used by source

Source	Variables
Household answers	Acceptance of being audio-recorded in certain parts of the interview, whether the household is a panel unit or not, use of a proxy person to respond at the interview, number of household members, educational level of reference person, main residence ownership regime, holdings of unlisted shares, holdings of listed shares, holdings of investment funds, holdings of fixed income investments, total number of pension funds, 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.
Paradata	Number of Euros (closed and interval) questions, 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.
Other paradata filled by interviewer	Dummies indicating if: the household was mistrustful before and after the interview, the household was showing some interest during the interview, number of people present when the interview was held, the household consulted external documents during the interview, 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 (Survey of Financial Competences by the Bank of Spain), educational level.
Error indicators and Inconsistencies	Tabulation, inconsistency and information content automated checks. See Table A.1 for details.

4 Empirical Strategy

4.1 Training and Evaluation

We employ a supervised machine learning approach and train several classifiers. This set comprises classical machine learning models - Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Machines (SVM) - and the well-known tree-based algorithms - Random Forests and Gradient Boosting

Trees (XGBoost) - given that there is no prior on which of these models would be better in this set up. We evaluated the performance of the different classifiers and compare their results. Appendix D provides a brief description of each classifier. The algorithmic capability increases in the aforementioned list, from simple linear models to non-linear and flexible algorithms that exhibit improved performance in higher dimensional settings. The use of bagging techniques with decision trees, also known as random forests, has recently gained popularity in the survey research community. Buskirk (2018) provide a detailed explanation of the approach. Although neural networks may be a viable option, the literature suggests that boosting and bagging techniques outperform neural net algorithms in predicting tabular data (Borisov et al., 2021).

We design a three-step process of training and evaluation in order to compare the models. This is motivated by the relatively small sample size, which implies a relatively small test set. To reduce any possible bias stemming from seed initialization and consequent data splitting, we compute the following three steps on ten different seeds:

1. We stratified randomly split the dataset into 70% train and 30% test sets.

2. We fit the model with a 5-fold stratified cross-validation strategy for hyperparameter tuning, using the 70% training data. The cross-validation aims to minimize 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. We evaluate the performance of the model by computing evaluation metrics on the test set.

In the process of training a machine learning algorithm, it is necessary to split the dataset to evaluate the classifier performance (step 1). Step 2 involves fitting the given classifier to the data. To accomplish this, a cross-validation strategy is used to optimize the classifier's configuration by tuning its hyperparameters. The computational cost of this task is high, particularly when ensemble methods are utilized, due to the curse of dimensionality (Bellman, 1966). Therefore, we employed a grid search optimization algorithm for the first three classifiers (i.e. the Logistic Regression Classifier, SVMs, and K-NN), and a random search algorithm with a maximum of 2500 iterations for the Random Forest and Gradient Boosting Classifier. The hyperparameter strategies, along with their hyperparameter space and search method, are provided in Table A.3. We explore a wide range of hyperparameters but, after numerous experiments, we determined that the hyperparameters listed in Table A.3 were a suitable representation of the suboptimal and optimal spaces for each model. It has been established that in a high-dimensional hyperparameter space, random search is a valid approach (Bergstra and Bengio (2012)).

In step 3, we assess the performance of the classifiers using two evaluation metrics averaged across all ten random seeds. The Receiver Operating Characteristic Area Under the Curve (ROC AUC) serves to evaluate the performance of binary classification models. It measures the model's ability to distinguish between positive and negative classes by plotting the true positive (TP) rate against the false positive (FP) rate at different classification thresholds and computing the area under the resulting curve. The score ranges from 1, which indicates perfect classification, to 0 with a score of 0.5 indicating that the model is no better than random. The ROC AUC score is insensitive to imbalanced datasets, which happens in this application. We also use a second metric, the area under the curve (AUC) of the precision-recall curve. The precision-recall curve plots the proportion of true positive classifications among all positive classifications (precision) against the proportion of true positive classifications among all actual positives (recall). Again, the AUC of the precision-recall curve is the integral of the curve, ranging from

0 to 1, with a higher value indicating better performance. The precision-recall curve focuses on the positive class and is more informative in cases where recall is more important than precision. This is the case of this application, since the rise of false negative cases can lead to a higher measurement error in the final data while the rise in false positive cases increases the revision time. Generally, a model with a higher ROC AUC is better at differentiating between the two classes, whereas a model with a higher PR AUC is better at identifying positive cases.

4.2 Optimal Threshold

Once we choose the best fitted classifier, we select the optimal threshold that classifies cases as a function of estimated test set probabilities. Given the data imbalance, a 50% threshold would not make sense since the predicted probability distribution is left-skewed. On the other side, increasing the threshold returns a lower FP rate, and an increasing FN rate. In the context of the survey, a FN occurrence implies that a case is not flagged but it should have been since it contains errors or inconsistencies, while a FP 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. Let the weighted harmonic mean of x_i for $i = 1, ..., n$ is:

$$F_l = \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{x_i}} \tag{1}$$

The weighted harmonic mean for recall and precision is then:

$$F_l = \frac{\beta + 1}{\frac{\beta}{recall} + \frac{1}{precision}} = (1 + \beta) \cdot \frac{precision \cdot recall}{(\beta \cdot precision) + recall} \tag{2}$$

where β stands for the differential weight of recall with respect to precision. This is a *linear* F-Beta score, a variation of the generally used F-Beta score. 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.

5 Results

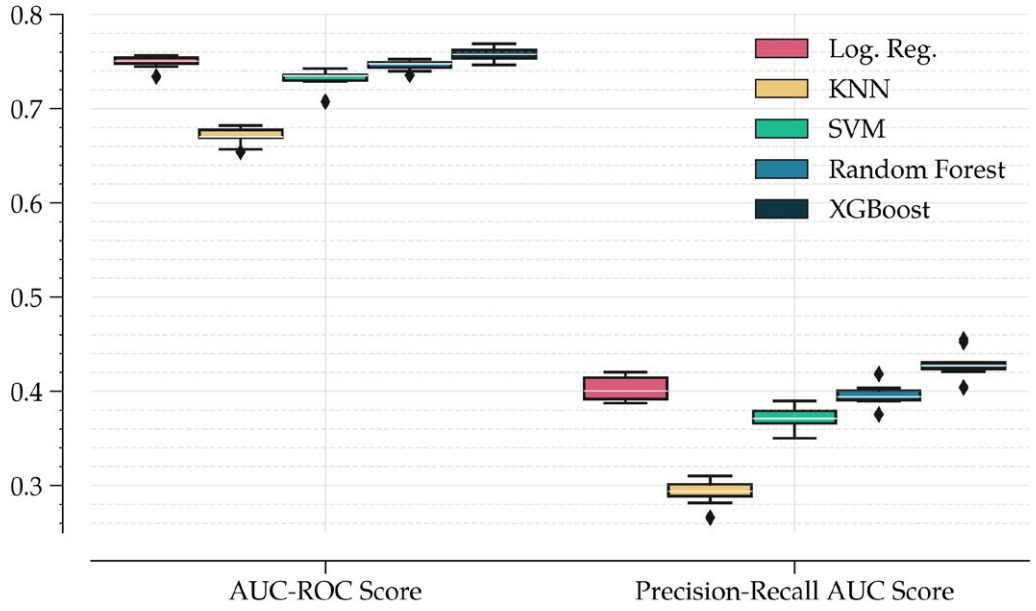
5.1 Optimal Model

As shown in Figure 1, the KNN classifier and SVM classifier are the worst performers among the competing algorithms, with XGBoost, random forest, and logistic classifier being the best performers, both in terms of AUC-ROC and PR-AUC scores. This result supports the use of ensemble and tree-based algorithms in tabular data, in line with Kern et al. (2019). XGBoost, with its boosting feature, outperforms all other algorithms.

The fact that the metric for PR-AUC is lower than for AUC-ROC means that algorithms ability to detect positives (questionnaires masking multiple omissions, errors, etc) is lower than algorithms ability to differentiate between the two classes. Figure A.4 of the Appendix contains the precise scores of each metric with additional metrics and the results are consistent.

In general, achieving high levels across performance metrics is challenging, the closest reference in

Figure 1: 10 random seeds Area Under the Curve (AUC) of the ROC and Precision-Recall curve



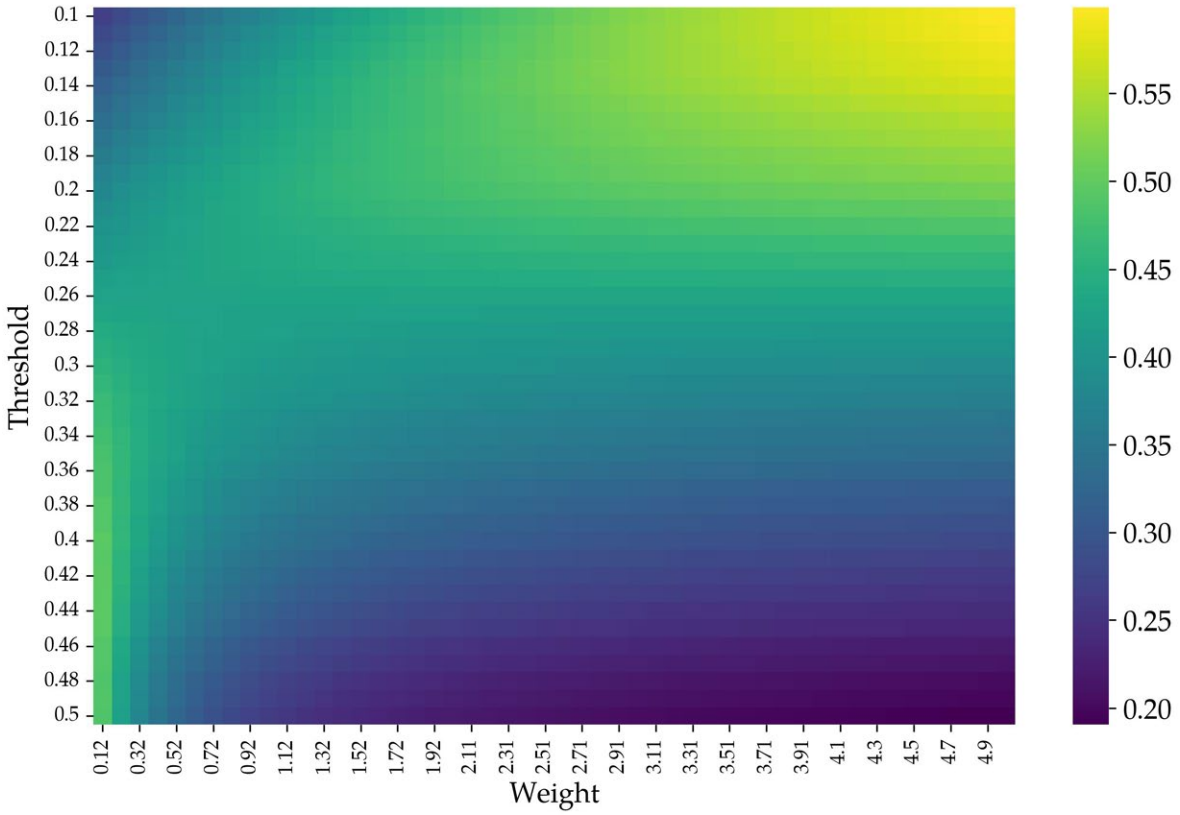
the literature Kern et al. (2021) and they face a different classification task and data. In our application, the data generation process (DGP) is complex. The DGP depends on the extensive depth of the logical tree of the questionnaire. This implies that there is huge heterogeneity and variability in the data, and therefore, it is very difficult to generalize on multiple test sets. In other words, no two surveys are the same in the whole sample. In fact, the number of variables collected throughout the survey amounts to more than 7500, more than observations. In addition, interviewers and editors provide a sort of heterogeneity in the DGP. In Section 5.4.2 we account for unobservable heterogeneity in the prediction errors of the model. The fact that the ROC curve shows a level of 0.75 already implies that the prediction of the model is 50% better than that of a fully random model.

Following the discussion in Section 4.2, Figure 2 presents the derived linear F-Beta values for different combinations of thresholds and βs for the best classifier fitted in the test samples. Setting the threshold at a low level significantly enhances the linear F-Beta Score only when β is higher than 2. In a setup where the optimization of recall is twice as important as the optimization of precision, the resulting optimal threshold is estimated to be approximately between 0.1-0.14. This threshold is also consistent with the more recent empirical share of cases with substantial errors in the data (16%).[5] If recall and precision are given the same weight, we would be optimizing F1 score and the resulting optimal threshold is in the range of 0.2 to 0.25. However, this is a larger proportion than the observed share of cases with substantial errors.

These results speak to the literature on selective editing. By assigning a score to each interview, the editing team can prioritize cases. For example, Figure B.2 presents a histogram with the distribution of scores allotted to each household for $\beta = 2$, where the relative importance of recall to precision is assumed to be 2. Based on this, households with a score lower than the optimal threshold, 0.14 for $\beta = 2$, will not be revised, those in yellow-green. Conversely, cases with scores above the threshold are deemed worthy of review, those in red. In Table 3 we observe that for a set of pre-determined betas, the optimal threshold varies and so does the recall and precision. As expected, a higher beta means a

[5]The share is 21.5% in EFF2017 and 11.8% in EFF2020. The latter is closer to the ongoing rate in EFF2022.

Figure 2: Linear F-Beta Score - weighting scheme for Gradient Boosting (XGBoost) Classifier



lower rate of false negatives, which translates into a higher recall score, at the expense of lower precision (higher false positives).

Table 3: Selected Betas, and correspondent thresholds, recall and precision.

Beta	Optimal Threshold[a]	Recall	Precision
0.5	0.30	0.37	0.46
1	0.20	0.55	0.37
1.5	0.19	0.58	0.35
2	0.14	0.71	0.23

[a]Calculated as the maximum F-Beta score for the selected Beta, as seen in figure 2

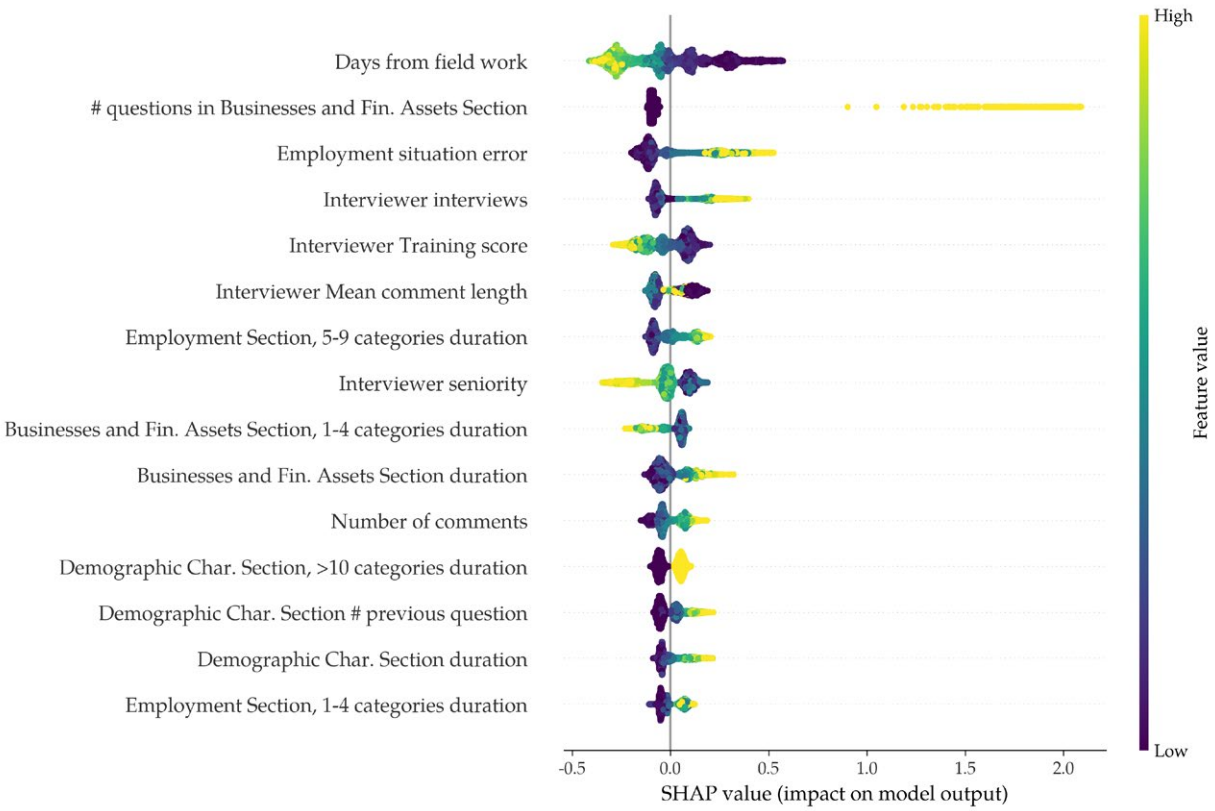
In addition, by choosing a different threshold the revision team can efficiently reallocate resources. Ideally, one would like to evaluate the impact of selecting one or another threshold in the resulting wealth distribution (De Waal, 2013). We believe that this is a worth exercise that due to the its own complexity 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.

5.2 Interpretability

One of the main drawbacks of using ensemble and tree-based algorithms is their interpretability. According to Miller (2019), interpretability of a machine learning model refers to the degree to which a

human can understand the cause of a decision made by the model. We use the SHAP (SHapley Additive exPlanations) framework developed by Lundberg and Lee (2017) to look at the most determinant features of the best-performing model that impact on the prediction. We first calculate the SHAP values

Figure 3: SHAP Values for XGBoost



Note: Based on the selected trained XGBoost algorithm. Random seed is 10; hyperparameters are: `"n_estimators": 300, "reg_alpha": 2, "reg_lambda": 4.3, "base_score": 0.5, "subsample": 0.95, "colsample_bytree": 0.5, "learning_rate": 0.05, "gamma": 0.14, "max_depth": 6`

for each feature of the input data. The SHAP values measure the impact of each feature on the model's output for a specific instance. The sum of the SHAP values for all features equals the difference between the model's output for the specific instance and the expected output for the population. Using SHAP values for interpreting tree-based classifiers can provide insight into how the model makes predictions and help identify areas where the model can be improved. For example, if a feature has a high SHAP value, it suggests that it has a strong impact on the model's prediction and should be carefully considered when making decisions based on the model's output. Additionally, if a feature has a low SHAP value, it suggests that it has little impact on the model's prediction and could potentially be removed from the model without affecting its performance. In Figure 3 SHAP values are plotted for the top 15 variables that have the strongest impact in determining the probability output of the trained model. For each variable, it shows the distribution of the SHAP values in the sample, each dot is mapped with its feature's value; the lighter the feature value, the higher the value is (and viceversa). Examining Figure 3, we observe that the number of days from the beginning of the field work is the most significant feature and it negatively affects the recontact score. Interviewers are more prone to making mistakes at this stage but also the reviewing team has historically identified more problematic cases to be recontacted

during this time. The number of questions asked in financial assets and businesses section is the second most significant feature for the predicted outcomes. In other words, the more complexity in this section, the higher the probability of being recontacted. Another predictive feature is errors in the working status which positively affects the score. This inconsistency may suggest that the member is omitting some labor information, resulting in a higher probability of being recontacted. The number of interviews conducted by the interviewer prior to the first household contact is the fourth most significant feature in determining the model output.

5.3 Validation with EFF2022 Ongoing Field Data

We also provide an out-of-sample evaluation of the model using data from the ongoing EFF2022 wave.[6] This allows us to compare the predictions of the best-performance trained model[7] with the most recent manual classification of the data reviewers. Figure B.3 presents the cumulative reviewed cases (the number of cases reviewed up to each date), the cumulative recontact rate (the percentage of detected recontacted cases at each point in time) and the cumulative ROC AUC score for the gradient boosting trees algorithm that is achieved at each point in time. For each day, we know how many cases are manually reviewed and recontacted, so we can make predictions and compare the on-going manual classification. Table 4 presents the corresponding evaluation metrics.

Table 4: 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

Table 4 demonstrates that the model generalizes well based on the observed test sample metrics. The ROC AUC score remains consistent at 0.725, indicating that the fitted classifier does not overfit and can consistently make predictions in new generated data (out-of-sample). It is important to note that each survey wave implements improvements on data reviewers and editing techniques. Thus, one could expect different in-sample evaluation and out-of-sample scores. However, the status of the survey is well-developed due to more than 20 years of experience and documentation efforts and this is fundamental to be able to implement this algorithmic procedure. The results indicate that the prediction tool is reliable across waves.

5.4 Additional Robustness

In order to clarify potential concerns regarding the relevance of the outcome we present an analysis based on an alternative of the outcome variable which incorporates expost information regarding the success or failure of a recontact. In addition, we also explore the unexplained variance of the model errors.

5.4.1 Successful Recontacts

Recontacts can fail if the household respondent does not want to answer any more questions or when she is out of reach. Thus, we can obtain an alternative measure of realized or successful recontacts. In

[6] This test sample is composed of 3430 observations, which represent approximately 55% of the final sample that will be available at the end of the fieldwork.

[7] We calculate the score or predicted probability of a household to be recontacted as the median of the 10 random seed fitted classifiers

practice, 89% of recontacts were successful between 2017 and 2020. We re-train and evaluate the steps in Section 3 using this measure as alternative target variable Y'. Table 5 shows the out-of-sample metrics where we do not find important differences from baseline results. There is a decrease in all performance metrics but the model still performs well and similar than in the original target variable.

Table 5: Ouf of sample metrics, using confirmed recontacts as target variable

	ROC AUC	PR AUC
Gradient Boosting Trees	0.718	0.252
Logistic Classifier	0.705	0.253
Random Forest	0.699	0.257

5.4.2 External Factors

Although we include hundreds of explanatory variables in the prediction model, there is a part of the prediction errors that cannot be explained. These errors are explained in an extent by the *reviewer* and *wave effects*. These factors are varying from wave to wave and unobservables at the prediction stage. In Table 6 we show that reviewer FE and wave FE account for 20% of the log-loss error variation.

Table 6: Exploratory analysis of unexplained errors of the optimal model

	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: Clustered standard errors in parenthesis. *p<0.1; **p<0.05; ***p<0.01

6 Conclusion

This study proposes a novel application of machine learning in survey methodology, specifically in the editing process of the Spanish Survey of Households Finances. The objective is to find the best-predicting models for detecting substantial errors in the questionnaires collected during the fieldwork process. Exploiting the revised data from previous waves, we show that tree-based ensemble models outperform other models in predicting substantial errors in the data. The algorithm outputs a score function assigning a probability of being a case with substantial errors (or probability of recontact) to each questionnaire given a large set of covariates. We show that the algorithm predict or matches well the manual classification in different test sets. This methodology also allows us to characterize what is the desirable probability threshold that classifies questionnaires into the positive or negative class. To

do so, it takes into account the acceptable amount of false negatives relative to false positives that the statistical office (or the responsibles 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. In particular, it provides an automatically-generated indicator that increases the efficiency of the manual case-by-case revision process giving priority to those cases more likely or prone to contain errors. In this sense, the use of this algorithm might be especially useful in cases where not massive manual revision can be performed because of limited funding since just a small classification exercise to train and test the model is needed. Furthermore, the indicator can also detect over and under-editing, which is a crucial to monitor and discipline the data editing process. In this sense, we provide a set of tools and results that speak to previous literature on edit prioritization using score functions and selective editing.

References

Allard, Mary Dorinda, Gordon Mikkelson and Linda I. Unger. (2001). "Implementing a Score Function to Prioritize Business Survey Edit Failures at BLS". *Proceedings of the Annual Meeting of the American Statistical Association*. http://www.asasrms.org/Proceedings/y2001/Proceed/00300.pdf

Arbués, Ignacio, Pedro Revilla y David Salgado. (2013). "An Optimization Approach to Selective Editing". *Journal of Official Statistics*, 29-(4), pp. 489-510. https://doi.org/10.2478/jos-2013-0037

Barceló, Cristina, Laura Crespo, Sandra García-Uribe, Carlos Gento, Marina Gómez and Alicia de Quinto. (2020). "The Spanish Survey of Household Finances (EFF): description and methods of the 2017 wave". Documentos Ocasionales, 2033, Banco de España. https://repositorio.bde.es/handle/123456789/14531

Bellman, Richard. (1966). "Dynamic Programming". *Science*, 153(3731), pp. 34-37. https://doi.org/10.1126/science.153.3731.34

Bergstra, James, and Yoshua Bengio. (2012). "Random Search for Hyper-Parameter Optimization". *Journal of Machine Learning Research*, 13(10), pp. 281-305. http://jmlr.org/papers/v13/bergstra12a.html

Borisov, Vadim, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk and Gjergji Kasneci. (2022). "Deep Neural Networks and Tabular Data: A Survey". *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1-21. https://doi.org/10.1109/TNNLS.2022.3229161

Breiman, Leo. (2001). "Random Forests". *Machine Learning*, 45, pp. 5-32. https://doi.org/10.1023/A:1010933404324

Bristle, Johanna, Martina Celidoni, Chiara Dal Bianco and Guglielmo Weber. (2019). "The contributions of paradata and features of respondents, interviewers and survey agencies to panel co-operation in the Survey of Health, Ageing and Retirement in Europe". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(1), pp. 3-35. https://doi.org/10.1111/rssa.12391

Buskirk, Trent 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). https://doi.org/10.29115/SP-2018-0003

Chen, Tianqi, and Carlos Guestrin. (2016). "XGBoost: A Scalable Tree Boosting System". *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Cortes, Corinna, and Vladimir Vapnik. (1995). "Support-vector networks". *Machine learning*, 20 (3), pp. 273-297. https://doi.org/10.1007/BF00994018

Dagdoug, Mehdi, Camelia Goga and David Haziza. (2021). "Imputation Procedures in Surveys Using Nonparametric and Machine Learning Methods: An Empirical Comparison". *Journal of Survey Statistics and Methodology*, 11(1), pp. 141-188. https://doi.org/10.1093/jssam/smab004

De Waal, Ton. (2013). "Selective Editing: A Quest for Efficiency and Data Quality". *Journal of official statistics*, 29(4), pp. 473-488. https://doi.org/10.2478/jos-2013-0036

De Waal, Ton, Jeroen Pannekoek and Sander Scholtus. (2011). "Selective Editing". In *Handbook of Statistical Data Editing and Imputation*, chap. 6. John Wiley & Sons, pp. 191-221. https://doi.org/10.1002/9780470904848.ch6

Durrant, Gabriele B., Robert M. Groves, Laura Staetsky and Fiona Steele. (2010). "Effects of Interviewer Attitudes and Behaviors on Refusal in Household Surveys". *Public Opinion Quarterly*, 74(1), pp. 1-36. https://doi.org/10.1093/poq/nfp098

Flores-Macias, Francisco, and Chappell Lawson. (2008). "Effects of Interviewer Gender on Survey Responses: Findings from a Household Survey in Mexico". *International Journal of Public Opinion Research*, 20(1), pp. 100-110. https://doi.org/10.1093/ijpor/edn007

Gismondi, Roberto. (2007). "Score Functions and Statistical Criteria to Manage Intensive Follow Up in Business Surveys". *Statistica*, 67(1), p. 27-54. https://doi.org/10.6092/issn.1973-2201/3496

Granquist, Leopold, and John G. Kovar. (1997). "Editing of Survey Data: How Much Is Enough?" In *Survey Measurement and Process Quality*, chap. 18. John Wiley & Sons, pp. 415-435. https://doi.org/10.1002/9781118490013.ch18

Groves, Robert M., and Steven G. Heeringa. (2006). "Responsive design for household surveys: tools for actively controlling survey errors and costs". *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3), pp. 439-457. https://doi.org/10.1111/j.1467-985X.2006.00423.x

He, Zhoushanyue, and Matthias Schonlau. (2021). "A Model-Assisted Approach for Finding Coding Errors in Manual Coding of Open-Ended Questions". *Journal of Survey Statistics and Methodology*, 10(2),pp. 365-376. https://doi.org/10.1093/jssam/smab022

Hedlin, Dan. (2003). "Score Functions to Reduce Business Survey Editing at the U.K. Office for National Statistics". *Journal of Official Statistics*, 19(2), pp. 177-199. https://www.proquest.com/docview/1266794939?pq-origsite=gscholar&fromopenview=true

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

Kennickell, Arthur B. (2006). "How Do We Know if We Aren't Looking? An Investigation of Data Quality in the 2004 SCF". Working Paper, Federal Reserve Board. https://www.federalreserve.gov/econresdata/scf/files/asa20063.pdf

Kennickell, Arthur B. (2017). "Look again: Editing and imputation of SCF panel data". *Statistical Journal of the IAOS*, 33(1), pp. 195-202. https://doi.org/10.3233/SJI-160268

Kern, Christoph, Thomas Klausch and Frauke Kreuter. (2019). "Tree-based Machine Learning Methods for Survey Research". *Survey Research Methods*, 13(1), pp. 73-93. https://doi.org/10.18148/srm/2019.v1i1.7395

Kern, Christoph, Bernd Weiß and Jan-Philipp Kolb. (2021). "Predicting Nonresponse in Future Waves of A Probability-Based Mixed-Mode Panel With Machine Learning". *Journal of Survey Statistics and Methodology*,11(1), pp. 100-123. https://doi.org/10.1093/jssam/smab009

Latouche, Michel, and Jean-Marie Berthelot. (1992). "Use of a Score Function to Prioritize and Limit Recontacts in Editing Business Surveys". *Journal of Official Statistics*, 8(3), pp. 389-400. https://www.proquest.com/scholarly-journals/use-score-function-prioritize-limit-recontacts/docview/1266807065/se-2

Lundberg, Scott M., and Su-In Lee. (2017). "A Unified Approach to Interpreting Model Predictions". In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna Wallach, Rob Fergus, S. V. N. Vishwanathan and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30.* Curran Associates, Inc., pp. 4765-4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

Miller, Tim. (2019). "Explanation in artificial intelligence: Insights from the social sciences". *Artificial Intelligence,* 267, pp. 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Porter, Martin F. (2001). "Snowball: A language for stemming algorithms". Published online. Accessed 11.03.2008, 15.00h. http://snowball.tartarus.org/texts/introduction.html

Sammut, Claude, and Geoffrey I. Webb (eds.) (2010). "TF-IDF". In *Encyclopedia of Machine Learning.* Springer US, pp. 986-987. https://doi.org/10.1007/978-0-387-30164-8_832

Schierholz, Malte, and Matthias Schonlau. (2020). "Machine Learning for Occupation Coding—A Comparison Study". *Journal of Survey Statistics and Methodology,* 9(5), pp. 1013-1034. https://doi.org/10.1093/jssam/smaa023

Toth, Daniell, and Polly Phipps. (2014). "Regression Tree Models for Analyzing Survey Response". In *Proceedings of the Government Statistics Section.* American Statistical Association, pp. 339-351. https://www.bls.gov/osmr/research-papers/2014/pdf/st140160.pdf

Vermeulen, Philip. (2018). "How Fat is the Top Tail of the Wealth Distribution?" *The Review of Income and Wealth,* 64(2), pp. 357-387. https://doi.org/10.1111/roiw.12279

A Tables

Table A.1: Error indicators and inconsistencies (cont.)

Name	Error indicator description (whether the household or any member...)
Panel Error	Panel households that don't have any panel member.
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.
Rent Revenue	Declare that the rent revenue is higher than the property value.
Other Properties Inconsistency	Declare that doesn't have any other properties but declares to have possessed other property in the past 12 months.
Jewels Inconsistency	Declare that doesn't have any jewels or art but declares to have possessed jewels or art in the past 12 months.
Squared Meters Indicator	Price of squared meter of property is too high.
Loan Monthly Payments	Monthly loan payments is higher than all pending loans value.
Loan Inconsistency	Pending amount in loan is higher than initial value of loan.
Loan Term Inconsistency	Pending term is higher than solicited loan term.
Business Member Inconsistency	Number of members that work on the family business is higher than total number of family business employees.
Stocks Inconsistency	Owns stocks of the firm that he or she works at, but the portfolio is not composed at 100% by these stocks.
Dividends Inconsistency	Yearly dividend yield is higher than whole portfolio value.
Accounts	Declares to have a financial account, but none in particular.
Accounts 2	Declares that the number of accounts is lower than the sum of the particular accounts.
Interest	The interest rate of an account is higher than the balance.
Investment Funds	The value of the investment funds is not equal to the sum of individual investment funds value.
Fixed Income Earnings	The fixed income earnings is higher than the fixed income portfolio.
Insurance Premium	The insurance premium is higher than the insurance value.
Insurance Valuation	The insurance valuation is the same as the insurance hedge for mixed insurances.
Revenue Growth	The revenue growth in income is higher than current regular income.

Table A.2: Error indicators and inconsistencies (cont.)

Name	Error indicator description (whether the household or any member...)
Employment history 2	Working years is higher than years with minimum legal working age.
Employment History	Declared to have worked the year prior to the interview, but worked less than 12 months.
Pension Young	Declared to receive the pension from a very young age.
Family Subsidy	The household does not receive any family subsidy but declared in other parts of the interview that they were receiving help.
Monthly Income 1	Declared that the monthly labor income (employed workers) is higher than the 50% of the previous year labor income.
Monthly Income 2	Declared that the monthly labor income in kind is higher than the 50% of the previous year labor income in kind.
Monthly Income 3	Declared that the monthly unemployment benefit is higher than the 50% of the previous year income from unemployment benefits.
Monthly Income 4	Declared that the monthly labor income (own account workers) is higher than the 50% of the previous year labor income.
Monthly Income 5	Declared that the monthly pension (retirement or inability) income is higher than the 50% of the previous year pension income.
Monthly Income 5	Declared that the monthly pension (retirement or inability) income is higher than the 50% of the previous year pension income.
Monthly Income 6	Declared that the monthly pension (widowhood/orphanhood) income is higher than the 50% of the previous year pension income.
Monthly Income 7	Declared that the monthly income from grants and scholarships is higher than the 50% of the previous year pension income from grants or scholarships.
Business Profit Inconsistency	Declared that the business profit is the same as the perceived salary.
Full Time Employment Years	The worked years full time are too high.
Never Worked	Never worked full time but in other parts of the questionnaire he or she did so.
Part Time Employment Years	The worked years part time are too high.
Worked Years Employer 1	The years working and contributing to social security are too high.
Worked Years Employer 2	Working years are too high.
Worked Years	Declared that work or worked, but 0 years in part and full time worked.
Retirement Age	The retirement age is too low.
Duplicated Payment	Declared that an external person from the household, help in the payment of a declared debt (duplicated in different sections of the questionnaire).
Credit Cards 1	Use more cards than they declared to possess.
Credit Cards 2	Use credit cards but any member has any financial account.
Credit Cards 3	Use credit cards but any member has account to make payments.
Banck Checks	Issue checks but do not have any account.
Accounts	Receive regular income but do not own any financial account.
Debit Payments	Debit payments but do not own any account.
Internet Banking 1	Use financial services (retail banking) through internet, but do not own any account.
Internet Banking 2	Are clients of a digital bank, but do not own any account.
Expenditure	Declare that food expenditure is higher than total expenditure.

Table A.3: Algorithms and selected Hyperparameters

Algorithm	Search Method	Hyperparameter Space
Logistic Regression Classifier	Grid	`"C": np.logspace(-1.5, 3, 10),` `"penalty": ["l1", "l2"]`
K Neighbors	Grid	`"K": [3, 5, 7, 10, 15, 20, 30, 50],`
Support Vector Machine	Grid	`"C": np.logspace(-1.5, 3, 10),` `"kernel": ["poly", "rbf"]`

Random Forest	Random	`"min_samples_leaf": [2, 4, 8, 16, 32, 64],`
		`"n_estimators": [25, 50, 70, 100],`
		`"max_features": ["sqrt", "log2", "auto"],`
		`"max_samples": [0.6, 0.7, 0.8, 0.9, None],`
		`"max_depth": [2, 4, 6, 8 ,16, 32, None]`
		`"min_samples_split": [2, 4, 6, 8 ,16, 32]`

Extreme Gradient Boosting	Random Search	`"gamma": np.linspace(0.05, 1.5, 10),`
		`"n_estimators": [100, 300, 500],`
		`"reg_alpha": np.linspace(1, 11, 20),`
		`"reg_lambda": np.linspace(1, 11, 25),`
		`"base_score": np.linspace(0.1, 0.6, 10),`
		`"subsample": np.arange(0.5, 1, 0.05),`
		`"colsample_bytree": np.arange(0.5, 1, 0.05),`
		`"learning_rate": [0.1, 0.05],`
		`"max_depth": [2, 3, 4, 5, 6]`

Table A.4: Main metrics (mean values of 10 random data splitting initializations)

	AUC-ROC Score	Average Precision Score	Mathew's Corr. Coefficient	Precision-Recall AUC Score
K Neighbors	0.670	0.286	0.069	0.293
Logistic Clf.	0.749	0.403	0.237	0.403
Random Forest	0.746	0.397	0.150	0.396
SVM	0.732	0.372	0.173	0.371
XGBoost	0.758	0.430	0.271	0.429

B Figures

Figure B.1: Recontact Process in the Spanish Survey of Household Finances (EFF)

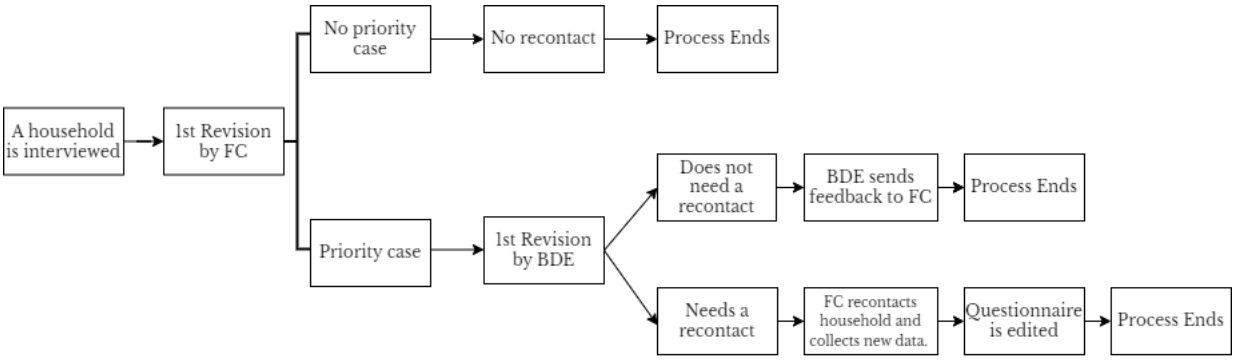
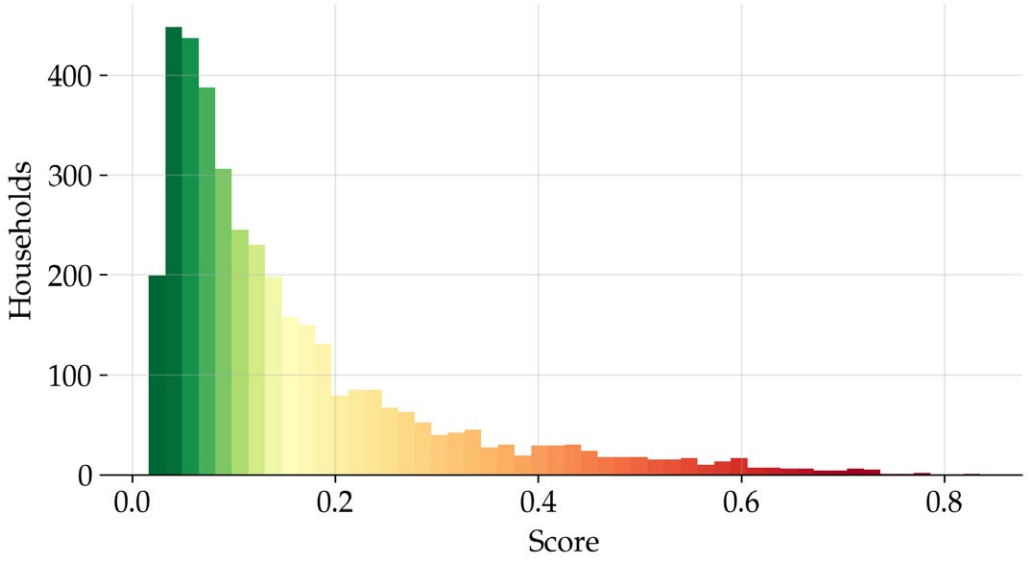


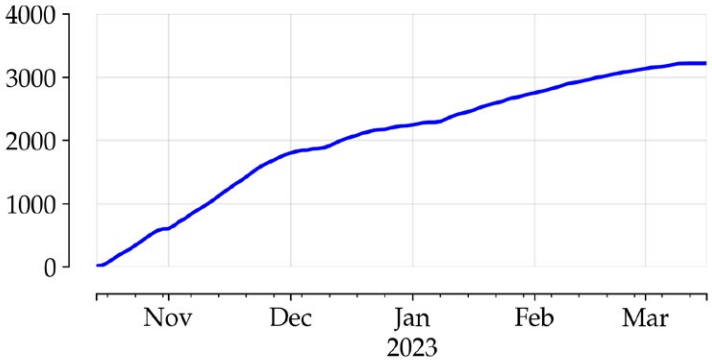
Figure B.2: Score Histogram



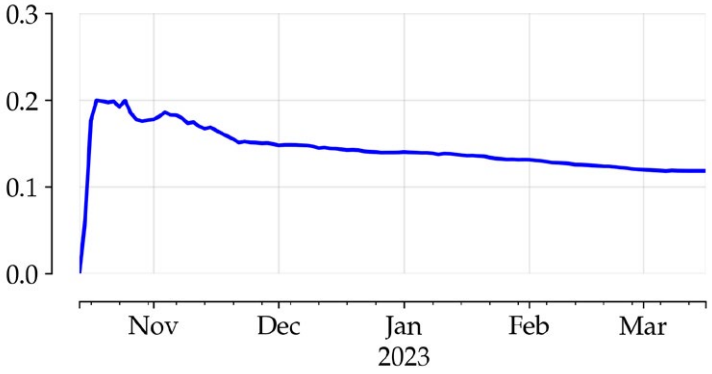
Note: Based on the selected trained XGBoost algorithm. Random seed is 10; hyperparameters are: `"n_estimators": 300, "reg_alpha": 2, "reg_lambda": 4.3, "base_score": 0.5, "subsample": 0.95, "colsample_bytree": 0.5, "learning_rate": 0.05, "gamma": 0.14, "max_depth": 6`

Figure B.3: EFF2022 Evaluation

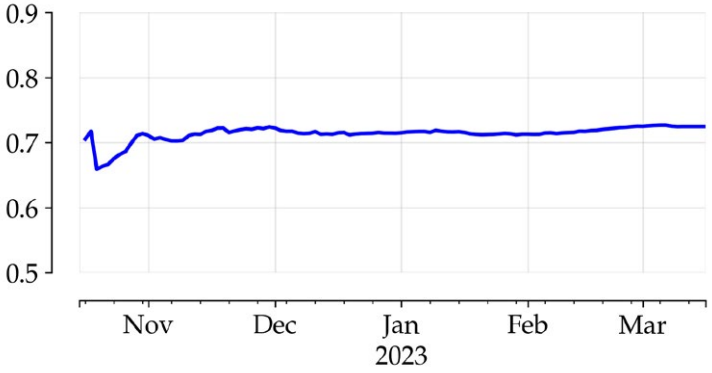
(a) Reviewed Cases



(b) Cummulative Rate



(c) Cummulative ROC AUC Score



Note: Time series evolution of (a) reviewed cases (absolute terms), (b) cummulative recontact rate and (c) cummulative estimated AUC-ROC score from the beginning of the field until the last available date. The x-axis indicates dates of the field.

C NLP Pipeline

We parse the data with pre-trained models from Honnibal and Montani (2017), removing stopwords, punctuation signs, alpha numeric characters, and others. Then, we apply Porter (2001) stemming and produce word counts under a bag of words approach. After cleaning the data, we select those words that are more present in each class (top 20 words with highest relative importance within that class). After running several experiments, we found that the extra complexity of other models based on TF-IDF counts (Sammut and Webb, 2010), were not adding improvements to the final stage.

D Classifiers

- Logistic regression: a statistical model that is commonly used for binary classification problems. This model is used to estimate the probability of a binary response variable (e.g., success/failure, true/false) based on one or more predictor variables. In addition to the standard logistic regression model, which uses maximum likelihood estimation to fit the model parameters, there are a number of variations that can be used to improve model performance. One of these variations is the use of an L1 penalty. We include this type of variation in order to improve the model performance. An L1 penalty, also known as a Lasso penalty, is a type of regularization technique that is used to prevent overfitting in the model. The penalty is applied to the model's coefficients, which are the values that are estimated for each predictor variable in the model. The penalty adds a term to the loss function of the model that penalizes coefficients for being too large, effectively shrinking them towards zero. This can help to reduce the impact of noisy or irrelevant predictor variables in the model, and can improve the model's ability to generalize to new data.

- K-Nearest neighbors (K-NN): this algorithm assumes that similar data points are closer to each other in the feature space. Given a new data point, the algorithm finds the K nearest neighbors to that point and assigns the class label of the majority of those neighbors to the new point.

  In the case of a classification problem, the K-NN algorithm determines the class label of a new data point by taking into account the K nearest data points in the training set. The number of neighbors, K, is a hyperparameter that needs to be set by the user before the algorithm is applied to the data. The K-NN algorithm works as follows. First, the algorithm takes the training data and stores the feature vectors and class labels. When a new data point is presented to the algorithm, the algorithm calculates the distances between the new data point and all the training data points. The distance between two data points can be calculated using various distance metrics such as Euclidean distance, Manhattan distance, or Cosine similarity. The K nearest neighbors to the new data point are selected based on the calculated distances. The algorithm assigns the class label of the majority of the K nearest neighbors to the new data point. If K is an odd number, there won't be any ties in the voting process. If K is an even number, there might be a tie in the voting process. In such cases, the algorithm can either choose the class label of the nearest neighbor or use a weighted voting scheme, where the closer neighbors have a greater impact on the final class label.

- Support vector machines (SVM): proposed by (Cortes and Vapnik, 1995), this algorithm relies on the basic idea of finding the optimal hyperplane that separates the data into different classes. SVMs can be used to solve linear as well as nonlinear classification problems. In the case of nonlinear classification problems, SVMs use a technique called the kernel trick to transform the data into a higher-dimensional space where it can be separated by a hyperplane. The kernel function is used

to measure the similarity between two data points in the transformed space. Common kernel functions include linear, polynomial, and radial basis function (RBF) kernels.

SVMs have several advantages and disadvantages. One of the main advantages is that they are effective in high-dimensional spaces, where other algorithms may struggle. They also perform well in cases where the number of dimensions is greater than the number of training data points. Additionally, SVMs have a strong mathematical foundation, which makes them easy to interpret and explain. However, one of the main disadvantages of SVMs is that they can be computationally expensive, especially for large data sets. They also require a good choice of kernel function and hyperparameters to be effective, which can be challenging. Additionally, SVMs can be sensitive to the presence of outliers in the data.

- Random forests: proposed by Breiman (2001), Random Forest is an ensemble learning algorithm. It combines multiple decision trees to create a more robust and accurate model. Each decision tree in a Random Forest is built using a random subset of the training data and a random subset of the features, which helps to reduce overfitting and improve generalization. For classification problems, the algorithm calculates the class label that is predicted by each decision tree and chooses the class label with the highest frequency. Random Forests have several advantages and disadvantages. The Random Forest algorithm works as follows. First, the algorithm takes the training data and randomly selects a subset of the data points with replacement. This is known as bootstrapping. Next, the algorithm constructs a decision tree using the selected subset of data points and a random subset of the features. The decision tree is built using a recursive process, where the algorithm selects the feature that best separates the data into different classes. The process of bootstrapping and building a decision tree is repeated multiple times, typically hundreds or even thousands of times. Once all the decision trees are built, the algorithm uses them to make predictions. For classification problems, the algorithm calculates the class label that is predicted by each decision tree and chooses the class label with the highest frequency.

- Gradient Boosting Classifier: this classifier is an ensemble learning algorithm that combines multiple weak learning models to create a more robust and accurate model. In contrast to Random Forests, which construct decision trees independently, Gradient Boosting Classifier trains decision trees in a sequential manner, with each tree correcting the errors made by the previous one. Specifically, we decide to use the XGBoost version, proposed in Chen and Guestrin (2016). XGBoost (Extreme Gradient Boosting) is a popular implementation of gradient boosting that is optimized for performance and scalability. It is an extension of the standard gradient boosting algorithm that includes additional features such as regularization and tree pruning, which help to reduce overfitting and improve generalization.

BANCO DE ESPAÑA PUBLICATIONS

WORKING PAPERS

2215 JOSÉ MANUEL CARBÓ and SERGIO GORJÓN: Application of machine learning models and interpretability techniques to identify the determinants of the price of bitcoin.
2216 LUIS GUIROLA and MARÍA SÁNCHEZ-DOMÍNGUEZ: Childcare constraints on immigrant integration.
2217 ADRIÁN CARRO, MARC HINTERSCHWEIGER, ARZU ULUC and J. DOYNE FARMER: Heterogeneous effects and spillovers of macroprudential policy in an agent-based model of the UK housing market.
2218 STÉPHANE DUPRAZ, HERVÉ LE BIHAN and JULIEN MATHERON: Make-up strategies with finite planning horizons but forward-looking asset prices.
2219 LAURA ÁLVAREZ, MIGUEL GARCÍA-POSADA and SERGIO MAYORDOMO: Distressed firms, zombie firms and zombie lending: a taxonomy.
2220 BLANCA JIMÉNEZ-GARCÍA and JULIO RODRÍGUEZ: A quantification of the evolution of bilateral trade flows once bilateral RTAs are implemented.
2221 SALOMÓN GARCÍA: Mortgage securitization and information frictions in general equilibrium.
2222 ANDRÉS ALONSO and JOSÉ MANUEL CARBÓ: Accuracy of explanations of machine learning models for credit decisions.
2223 JAMES COSTAIN, GALO NUÑO and CARLOS THOMAS: The term structure of interest rates in a heterogeneous monetary union.
2224 ANTOINE BERTHEAU, EDOARDO MARIA ACABBI, CRISTINA BARCELÓ, ANDREAS GULYAS, STEFANO LOMBARDI and RAFFAELE SAGGIO: The Unequal Consequences of Job Loss across Countries.
2225 ERWAN GAUTIER, CRISTINA CONFLITTI, RIEMER P. FABER, BRIAN FABO, LUDMILA FADEJEVA, VALENTIN JOUVANCEAU, JAN-OLIVER MENZ, TERESA MESSNER, PAVLOS PETROULAS, PAU ROLDAN-BLANCO, FABIO RUMLER, SERGIO SANTORO, ELISABETH WIELAND and HÉLÈNE ZIMMER. New facts on consumer price rigidity in the euro area.
2226 MARIO BAJO and EMILIO RODRÍGUEZ: Integrating the carbon footprint into the construction of corporate bond portfolios.
2227 FEDERICO CARRIL-CACCIA, JORDI PANIAGUA and MARTA SUÁREZ-VARELA: Forced migration and food crises.
2228 CARLOS MORENO PÉREZ and MARCO MINOZZO: Natural Language Processing and Financial Markets: Semi-supervised Modelling of Coronavirus and Economic News.
2229 CARLOS MORENO PÉREZ and MARCO MINOZZO: Monetary Policy Uncertainty in Mexico: An Unsupervised Approach.
2230 ADRIAN CARRO: Could Spain be less different? Exploring the effects of macroprudential policy on the house price cycle.
2231 DANIEL SANTABÁRBARA and MARTA SUÁREZ-VARELA: Carbon pricing and inflation volatility.
2232 MARINA DIAKONOVA, LUIS MOLINA, HANNES MUELLER, JAVIER J. PÉREZ and CRISTOPHER RAUH: The information content of conflict, social unrest and policy uncertainty measures for macroeconomic forecasting.
2233 JULIAN DI GIOVANNI, MANUEL GARCÍA-SANTANA, PRIIT JEENAS, ENRIQUE MORAL-BENITO and JOSEP PIJOAN-MAS: Government Procurement and Access to Credit: Firm Dynamics and Aggregate Implications.
2234 PETER PAZ: Bank capitalization heterogeneity and monetary policy.
2235 ERIK ANDRES-ESCAYOLA, CORINNA GHIRELLI, LUIS MOLINA, JAVIER J. PÉREZ and ELENA VIDAL: Using newspapers for textual indicators: which and how many?
2236 MARÍA ALEJANDRA AMADO: Macroprudential FX regulations: sacrificing small firms for stability?
2237 LUIS GUIROLA and GONZALO RIVERO: Polarization contaminates the link with partisan and independent institutions: evidence from 138 cabinet shifts.
2238 MIGUEL DURO, GERMÁN LÓPEZ-ESPINOSA, SERGIO MAYORDOMO, GAIZKA ORMAZABAL and MARÍA RODRÍGUEZ-MORENO: Enforcing mandatory reporting on private firms: the role of banks.
2239 LUIS J. ÁLVAREZ and FLORENS ODENDAHL: Data outliers and Bayesian VARs in the Euro Area.
2240 CARLOS MORENO PÉREZ and MARCO MINOZZO: "Making text talk": The minutes of the Central Bank of Brazil and the real economy.
2241 JULIO GÁLVEZ and GONZALO PAZ-PARDO: Richer earnings dynamics, consumption and portfolio choice over the life cycle.
2242 MARINA DIAKONOVA, CORINNA GHIRELLI, LUIS MOLINA and JAVIER J. PÉREZ: The economic impact of conflict-related and policy uncertainty shocks: the case of Russia.
2243 CARMEN BROTO, LUIS FERNÁNDEZ LAFUERZA and MARIYA MELNYCHUK: Do buffer requirements for European systemically important banks make them less systemic?
2244 GERGELY GANICS and MARÍA RODRÍGUEZ-MORENO: A house price-at-risk model to monitor the downside risk for the Spanish housing market.

2245 JOSÉ E. GUTIÉRREZ and LUIS FERNÁNDEZ LAFUERZA: Credit line runs and bank risk management: evidence from the disclosure of stress test results.

2301 MARÍA BRU MUÑOZ: The forgotten lender: the role of multilateral lenders in sovereign debt and default.

2302 SILVIA ALBRIZIO, BEATRIZ GONZÁLEZ and DMITRY KHAMETSHIN: A tale of two margins: monetary policy and capital misallocation.

2303 JUAN EQUIZA, RICARDO GIMENO, ANTONIO MORENO and CARLOS THOMAS: Evaluating central bank asset purchases in a term structure model with a forward-looking supply factor.

2304 PABLO BURRIEL, IVÁN KATARYNIUK, CARLOS MORENO PÉREZ and FRANCESCA VIANI: New supply bottlenecks index based on newspaper data.

2305 ALEJANDRO FERNÁNDEZ-CEREZO, ENRIQUE MORAL-BENITO and JAVIER QUINTANA: A production network model for the Spanish economy with an application to the impact of NGEU funds.

2306 MONICA MARTINEZ-BRAVO and CARLOS SANZ: Trust and accountability in times of pandemic.

2307 NATALIA FABRA, EDUARDO GUTIÉRREZ, AITOR LACUESTA and ROBERTO RAMOS: Do Renewables Create Local Jobs?

2308 ISABEL ARGIMÓN and IRENE ROIBÁS: Debt overhang, credit demand and financial conditions.

2309 JOSÉ-ELÍAS GALLEGOS: Inflation persistence, noisy information and the Phillips curve.

2310 ANDRÉS ALONSO-ROBISCO, JOSÉ MANUEL CARBÓ and JOSÉ MANUEL MARQUÉS: Machine Learning methods in climate finance: a systematic review.

2311 ALESSANDRO PERI, OMAR RACHEDI and IACOPO VAROTTO: The public investment multiplier in a production network.

2312 JUAN S. MORA-SANGUINETTI, JAVIER QUINTANA, ISABEL SOLER and ROK SPRUK: Sector-level economic effects of regulatory complexity: evidence from Spain.

2313 CORINNA GHIRELLI, ENKELEJDA HAVARI, ELENA MERONI and STEFANO VERZILLO: The long-term causal effects of winning an ERC grant.

2314 ALFREDO GARCÍA-HIERNAUX, MARÍA T. GONZÁLEZ-PÉREZ and DAVID E. GUERRERO: How to measure inflation volatility. A note.

2315 NICOLÁS ABBATE, INÉS BERNIELL, JOAQUÍN COLEFF, LUIS LAGUINGE, MARGARITA MACHELETT, MARIANA MARCHIONNI, JULIÁN PEDRAZZI and MARÍA FLORENCIA PINTO: Discrimination against gay and transgender people in Latin America: a correspondence study in the rental housing market.

2316 SALOMÓN GARCÍA: The amplification effects of adverse selection in mortgage credit suply.

2317 METTE EJRNÆS, ESTEBAN GARCÍA-MIRALLES, METTE GØRTZ and PETTER LUNDBORG: When death was postponed: the effect of HIV medication on work, savings and marriage.

2318 GABRIEL JIMÉNEZ, LUC LAEVEN, DAVID MARTÍNEZ-MIERA and JOSÉ-LUIS PEYDRÓ: Public guarantees and private banks' incentives: evidence from the COVID-19 crisis.

2319 HERVÉ LE BIHAN, DANILO LEIVA-LEÓN and MATÍAS PACCE: Underlying inflation and asymmetric risks.

2320 JUAN S. MORA-SANGUINETTI, LAURA HOSPIDO and ANDRÉS ATIENZA-MAESO: Los números de la regulación sobre igualdad. Cuantificación de la actividad normativa sobre no discriminación en España y su relación con las brechas de género en el mercado de trabajo.

2321 ANDRES ALONSO-ROBISCO and JOSÉ MANUEL CARBÓ: Analysis of CBDC Narrative of Central Banks using Large Language Models.

2322 STEFANIA ALBANESI, ANTÓNIO DIAS DA SILVA, JUAN F. JIMENO, ANA LAMO and ALENA WABITSCH: New technologies and jobs in Europe.

2323 JOSÉ E. GUTIÉRREZ: Optimal regulation of credit lines.

2324 MERCEDES DE LUIS, EMILIO RODRÍGUEZ and DIEGO TORRES: Machine learning applied to active fixed-income portfolio management: a Lasso logit approach.

2325 SELVA BAHAR BAZIKI, MARÍA J. NIETO and RIMA TURK-ARISS: Sovereign portfolio composition and bank risk: the case of European banks.

2326 ÁNGEL IVÁN MORENO and TERESA CAMINERO: Assessing the data challenges of climate-related disclosures in european banks. A text mining study.

2327 JULIO GÁLVEZ: Household portfolio choices under (non-)linear income risk: an empirical framework.

2328 NATASCHA HINTERLANG: Effects of Carbon Pricing in Germany and Spain: An Assessment with EMuSe.

2329 RODOLFO CAMPOS, SAMUEL PIENKNAGURA and JACOPO TIMINI: How far has globalization gone? A tale of two regions.

2330 NICOLÁS FORTEZA and SANDRA GARCÍA-URIBE: A Score Function to Prioritize Editing in Household Survey Data: A Machine Learning Approach.