PIT NON-FILERS PREDICTIVE MODELS

Data Mining & Big Data Unit
IT Department

February 2023

ABSTRACT

Every year, the AEAT receives a lot of information from different sources related to personal income allocations. Based on this information, and in accordance with current regulations, the set of taxpayers who are required to submit a personal income tax return is determined. Unfortunately, every year, part of these taxpayers do not file the return.
In September, the non-filing campaign begins, the first step being sending a request to all those taxpayers who have not file a return and whose calculated tax amount exceeds a certain threshold. Followed by other compliance measures that we will not detail here.

In order to tackle this issue, the AEAT has decided to develop several predictive models to advance the non-filing campaign. The objective is to contact those taxpayers who are considered most likely not to submit their PIT return during the voluntary period.

OBJECTIVES

- To improve taxpayer **support**, through contact during the voluntary filing period, to achieve an increase in the percentage of personal income tax returns filed on time.
- To reduce the number of **sanctions and surcharges** applied to those taxpayers who do not file a return on time.
- To increase the AEAT **efficiency**. By contacting potential non-filers during the voluntary filing period, we expect the final amount of non-filers to be reduced, thus decreasing the number of request to be issued as well as subsequent audit actions.

STARTING POINT

In the 2021 income tax campaign, there were just over **100.000 non-filers of the Personal Income Tax** (PIT or IRPF in Spain), taxpayers who were obliged to file the model 100, but did not do so on time.

- Approximately half of them corresponded to taxpayers who had filed a declaration in the previous income tax campaign (which we will call **Former Taxpayers**).
- Meanwhile, the other half corresponded to taxpayers who had not filed a declaration in the previous income tax campaign (which we will call **New Taxpayers**).

In September, the **non-filer campaign begins**, loading the data of this group of taxpayers in a specific application for the processing of the corresponding files. A group is selected consisting of those taxpayers for whom, based on the tax data, the calculation of the settlement exceeds a certain amount. A requirement is sent to this group of taxpayers urging them to file the declaration. In cases where they still do not file it, further actions are taken which may end in a tax assessment, including the possibility of a penalty.

It would be desirable to **contact those taxpayers who are considered more likely not to file the declaration on time before the filing deadline ends**. Therefore, assuming that a large proportion of those contacted will choose to file the declaration, the workload of the subsequent non-filer campaign can be reduced.

TECHNICAL OBJECTIVE

Predict the non-filers who, being obligated to file their income tax, will not do so within the filing deadline, as well as **estimate the result of their declaration** applying **Machine Learning techniques**. (Supervised Learning)

PREDICTIVE MODELS

For the **prediction of potential non-filers** among the taxpayers two different models are used. Each model is applied **based on the past declaration of the taxpayer**. Both models are implemented as **Binary Classifiers**, making use of the **estimated probability** of belonging to the classes, which allows for adjusting the threshold to establish the classes:

- **Former taxpayers model**: Prediction about the probability of filing the declaration for taxpayers required to file in this income tax campaign who **DID** file in the previous campaign.
- **New taxpayers model**: Prediction about the probability of filing the declaration for taxpayers required to file in this income tax campaign who **DID NOT** file in the previous campaign.

In both cases, the prediction of the taxpayer filing the declaration is obtained along with a prediction of the outcome of their declaration, using a third model, in this case a **Regression model**:

- **Tax liability model**: Prediction of the tax liability resulting from the declaration

TRAINING OF THE PREDICTIVE MODELS

The model learns from past cases to predict the probability of new cases. The model is trained with cases of which we already know whether they were filed on time or not. Information from the previous four income tax campaigns is used (years 2018 to 2021)

There are two main data sources that will provide the explanatory variables:

- **Income imputation data** in the current and previous tax income campaigns
- **PIT declaration data** from previous tax income campaigns (if any)

MODEL PERFORMANCE EVALUATION

20% of the training set is reserved for **validating the performance of the different models** that are trained (**test set**). The best model is selected from among these. The test set is never used in the hyper-parameter tuning performed for each model, instead a Cross-Validation method is applied. However, once the best model is chosen, it is once again trained but with the complete set of data.

In order to **evaluate the performance of predictive models in the future**, **10% of the population, on which the predictions are made, is reserved for his purpose**. Each data instance reserved for evaluation is marked in the final output, indicating that those instances are part of the **Control Group**, and their predictions will not be used in subsequent actions that use the information from the models.

EXAMPLE OF DATA USED IN THE LEARNING:
(Where N the year of the tax income campaign)

New taxpayer model

Income imputation data N

Reasons for mandatory declaration and tax data

Income imputation data N-1

Reasons for mandatory declaration and tax data

Former taxpayer model

Income imputation data N

Reasons for mandatory declaration and tax data

Income imputation data N-1

Reasons for mandatory declaration

PIT declaration data N-1

Data from the model 100 declaration

Prediction
Probability of filing the declaration for the N-campaign

Tax liability model

Income imputation data N

Tax data

Prediction
Tax liability resulting from the declaration on the N-campaign

NEW TAXPAYERS MODEL

Obtaining the probability of not filing on time from those taxpayers who are obliged to do so, but did not file last income tax campaign, either because they were not obliged to or because they chose not to.

The model is applied at the beginning of the tax campaign using the following data: income imputation data in the current campaign (N) and previous campaign (N-1), together with the reasons for mandatory declaration in the current campaign.

TRAINING

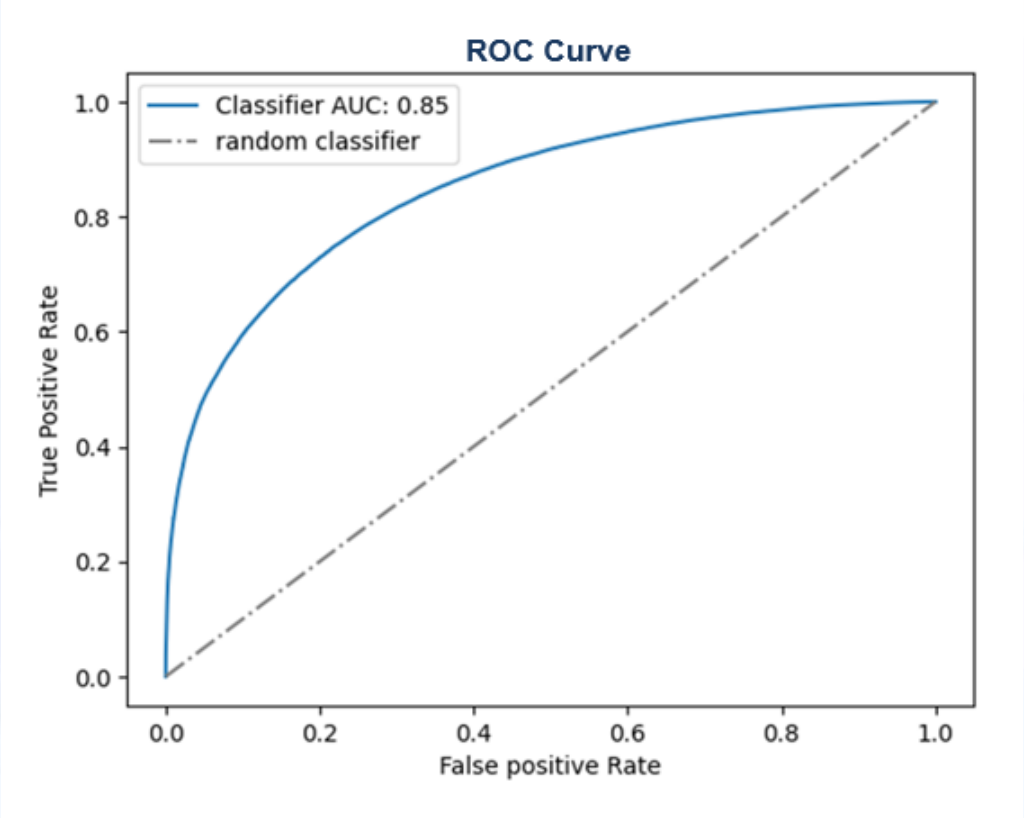
The best model (determined empirically) for this classification problem is a sequential ensemble of decision trees, XGBoost.

The optimization metric used is the area under the ROC curve. The proportion of both classes in the training dataset is fairly balanced, with 57% positive class (Non-filers).

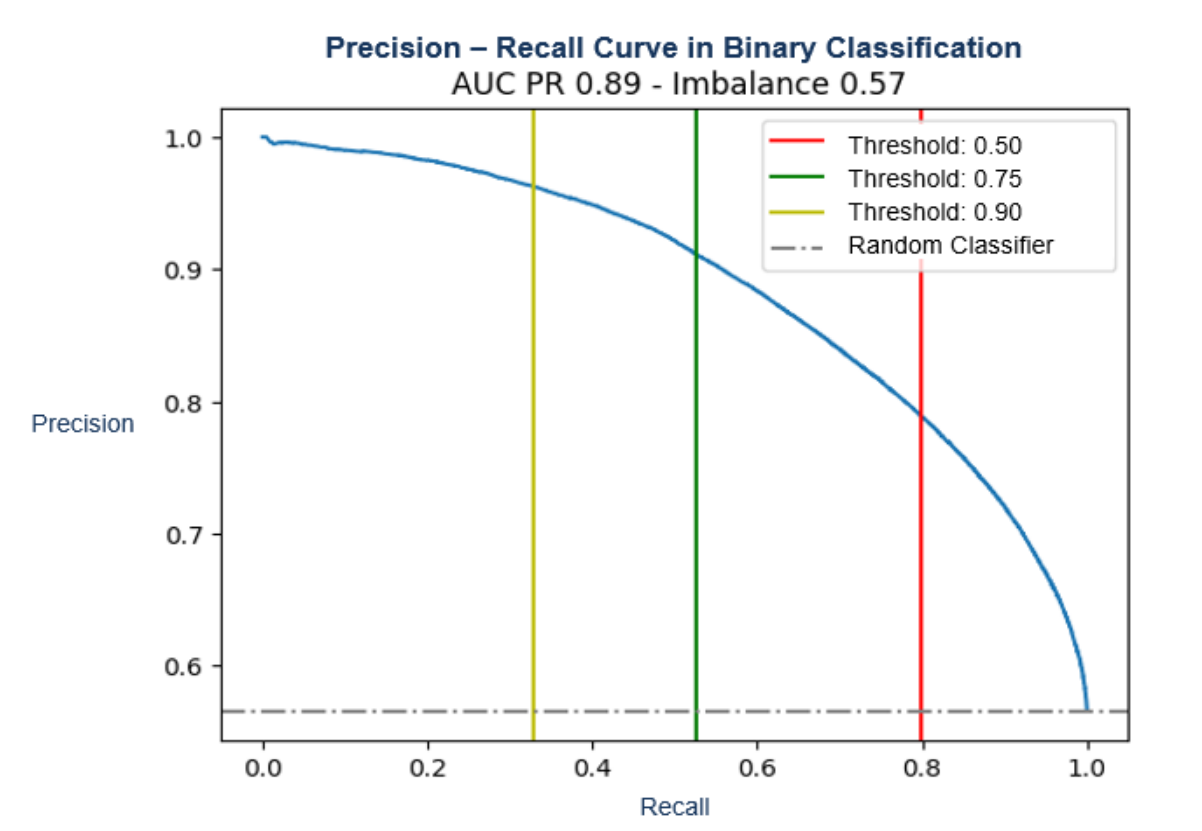
TRAINING RESULTS

Model metrics:

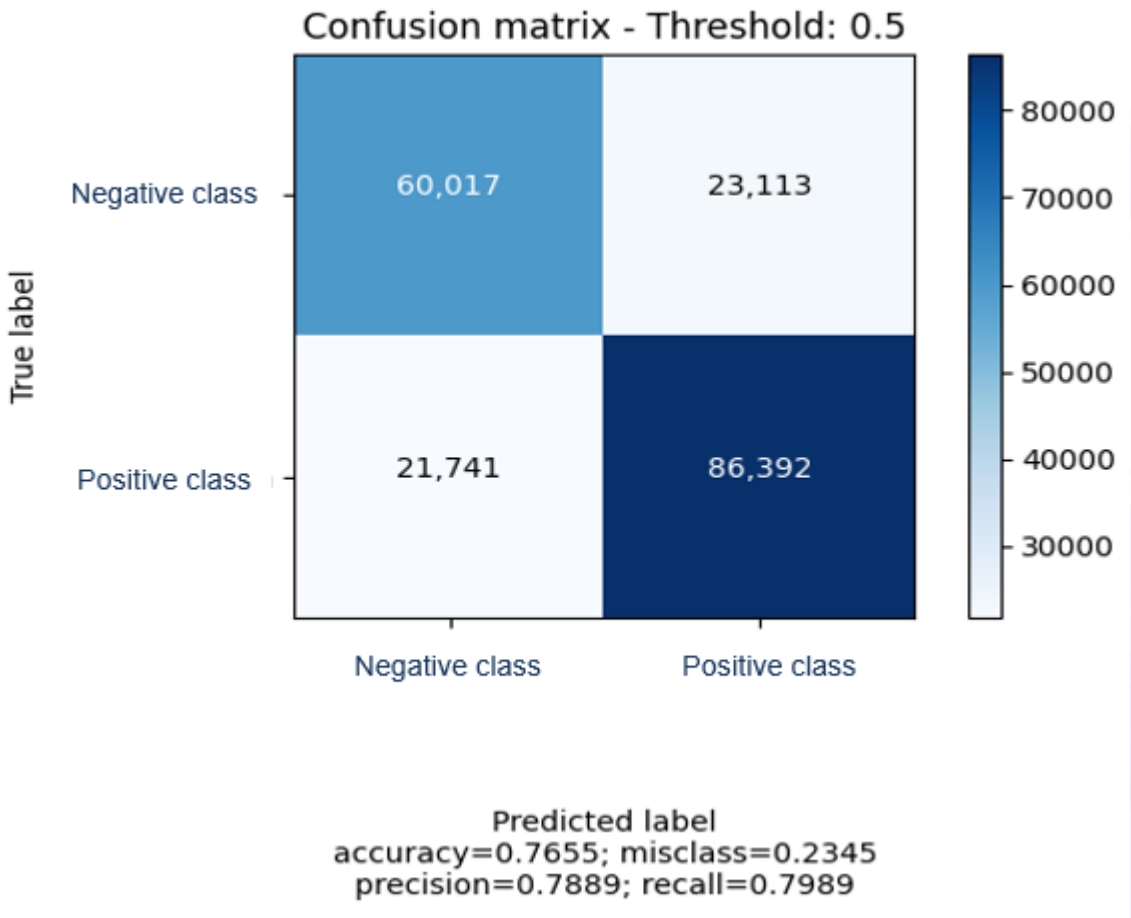
	Precision	Recall	F1 Score
Negative class	0.73	0.72	0.73
Positive class	0.79	0.8	0.79
macro avg	0.76	0.76	0.76
weighted avg	0.77	0.77	0.77



TRAINING RESULTS



TRAINING RESULTS



FORMER TAXPAYERS MODEL

Obtaining the probability of not filing on time from those taxpayers who are obliged to do so, but did file on time in the last income tax campaign.

The model is applied at the beginning of the tax campaign using the following data: income imputation data in the current campaign (N) and previous campaign (N-1), together with the reasons for mandatory declaration in the current campaign (N) and the data from the declaration of the previous campaign (N-1).

TRAINING

The best model (determined empirically) for this classification problem is a sequential ensemble of decision trees, XGBoost.

The optimization metric used is the area under the ROC curve. The proportion of both classes in the dataset is highly imbalanced, with only 1.57% of the positive class (non-filers). The use of XGBoost and the ROC-AUC metric aims precisely to counteract this significant imbalance, although undersampling and oversampling methods are being tested to see if they can improve the model's performance on the imbalanced dataset.

TAX LIABILITY MODEL

Obtaining the tax liability resulting from the declaration in the actual campaign

The model is applied at the beginning of the tax campaign using income imputation data in the current campaign (N)

TRAINING

The models currently under development for the regression problem are: on the one hand, a sequential ensemble of decision trees, XGBoost, and on the other hand, a linear regression to which restrictions or regularizations are applied on the coefficients, ElasticNet. The alternative that provides better results will be chosen between these two options.

The optimization metric used is the mean absolute error, MAE, a measure of error that is in the same units as the target variable and is robust to outliers. Other metrics that are also being considered include the coefficient of determination, R2 score (degree of fit of the model to the observations/data), and the root mean squared error, RMSE, which is more affected by outliers and is in the same units as the target variable.

SEARCH FOR THE OPTIMAL MODEL

What is described in the following paragraphs applies to all 3 models (the 2 binary classification models and the regression model)

In the model optimization process, the hyper-parameters of:
- Data post-processing
- The model itself

are being optimized

For each of these 2 stages, a hyper-parameter grid is defined. A GridSearch method is applied, trying all possible combinations and extracting the best result. Each possible combination of hyper-parameters is internally optimized by Cross-Validation.

The data post-processing hyper-parameters that are being optimized are:
- Application or not of scaling
- Application of dimensionality reduction methods (PCA, PLS)
- Establishment of minimum thresholds for variance and number of zeros
- Univariate feature selection (ANOVA method, Mutual Information method)
- Sampling methods (undersampling or oversampling).

The hyper-parameter tuning of the model is performed based on the chosen metric. This optimization metric is chosen considering whether the model is a classification or regression model and based on the balance of the data

Agencia Tributaria