Estimating Shortfall in Social Security Contribution Tax Revenue : Highlighting Treatment Bias in the Audit

Gabriel Fiastre
Paris Dauphine, PSL University
gabriel.fiastre@dauphine.eu

***Abstract*— This paper presents an estimate of the tax shortfall revenue through different methods, highlighting a bias in the selection process of auditing certain companies for potential fraud on Social Security Contribution. After defining the theoretical framework, we first propose an estimate of tax shortfall through different combination of propensity score matching and classical Machine Learning methods. The propensity score matching shall take eventual treatment bias into account and supposedly cancel it out, making comparison with classical results interesting. We will then compare the results in order to discuss the existence of a bias in the auditing selection process of the companies, showing that although there is no obvious selection bias based on the probability of being fraudulous, a selection bias on the potential adjutment amount can be highlighted. The full paper is available in french here.**

I. INTRODUCTION

For this research project a control agency provided us with data from $n = 187,000$ companies ($m = 265$ variables). The long process of data processing/engineering is not detailed in this paper. Some companies have been audited: we then know if they have committed fraud, and if so, the associated amount. We can therefore divide the database into three nested subclasses: All the firms, the set of audited companies, the set of audited and fraudulent companies.
We wish to estimate how much could the inspection agency have recovered if all companies had been inspected?
Therefore, we need to be able to :
- Predict whether an individual belongs to the fraudulent group We will use traditional classification methods, and compare it with a Propensity score matching approach where each uncontrolled individual is associated with a controlled individual and potential fraud deducted
- Predict the adjustment amount for a fraudulent observation. For this part we will consider a classical linear regression approach.

II. THEORETICAL FRAMEWORK

Let M_i^* be the variable denoting the potential amount of adjustment for firm $i \in 1, ...n$. Let F_i^* be the binary variable indicating for an uncontrolled individual whether he is potentially fraudulent or not. We make the following assumption :

$$M_i^* = \begin{cases} x_i\beta + \epsilon_i & \text{if} \quad F_i^* = 1 \\ 0 & \text{else} \end{cases} \qquad \forall i = 1, ...n \qquad (1)$$

with β a vector of p parameters and ϵ_i noise terms

We want to estimate the Shortfall, denoted as MAG, with :

$$MAG = \sum_{i=1}^{n} MAG_i$$

where, with C_i is the binary variable indicating past Audit :

$$MAG_i = M_i^* \times \mathbb{1}_{(C_i=0)} \times \mathbb{1}_{(F_i^*=1)} \qquad \forall i = 1, ..., n$$

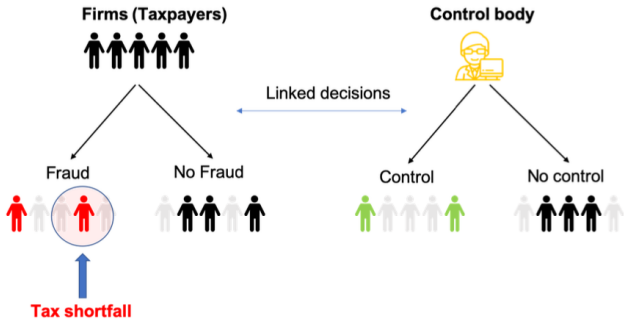

Fig. 1. Tax Shortfall definition, control scheme

III. ADRESSING CLASSIFICATION STEP : PREDICT COMPANIES FRAUD

A. Classical approach : Random Forest classification

The first step consists in estimating F_i^*. We first considered using traditional classification algorithms. After trying different models, it quickly appeared through cross-validation tests that a random forest model was better suited for this task given the few data available (no computational issue).
However there was a need for balancing the data as the fraudulant compagnies were a clear minority class. We implemented oversampling with the *SMOTE* technique in order to balance the dataset.
The model was implemented on R using the *randomForest* package. After implementing cross-validation we chose to use Entropy for building the trees. We then implemented *K-Fold cross-validation* in order to adress overfitting issues with $K = 5$. For model-selection we chose to maximize the *ROC-AUC* (Area Under ROC Curve), which is defined as following.
The *ROC* (Receiving Operator Curve) is the curve $(FPR(s), TPR(s))$ with :

$$FPR(s) = 1 - \frac{TP(s)}{TP(s) + TN(s)}$$

$$TPR(s) = 1 - \frac{FN(s)}{FP(s) + FN(s)}$$

with TP, TN, FP, FN respectively the True Positives, True Negatives, False Positives and False Negatives among test predictions.
After tuning the model, we have a final set of hyperparameters with $max\_features = 5$, $n_{tree} = 600$, and a classification threshold of $c = (0.58, 0.42)$
Finally, we get a final model with an AUC of 0.85 which indicates a good predictive capacity for the model.
We retain a final result of $n_{F^*} = \sum_{i=1}^{n} F_i^* = 49,833$ potential fraudeurs designated for the Random Forest Model. It represents approximatively 27% of the compagnies which weren't audited, which is almost the same proportion of fraudulent companies among the audited ones ($\approx 29.5\%$)

B. Propensity Score Matching

We are now considering for this task an approach based on *Propensity Score Matching* techniques. We considered our data divided in two groups : the *Control Group*(companies which were not audited) and the *Treatment Group*(companies which were audited). This approach focuses on a potential bias in the attribution of a group for observations [1] : here the selection of companies to be audited, or in other words companies suspected of frauds. It is obvious to suggest that a control agent, when doing his job correctly, should audit the companies (assign them to treatment group) with the higher probablity of being fraudulent. The propensity score matching is supposed to consequently cancel that bias out. The idea is to use the "unbiased" matching for classification purpose in order to predict F_i^*.

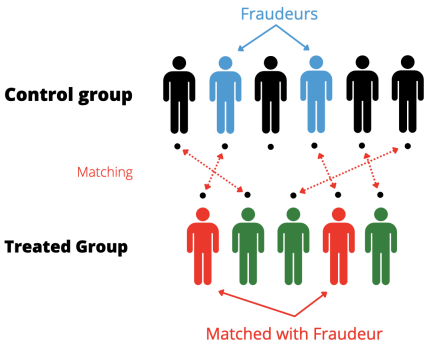

Fig. 2. Matching used for classification

Let us define :

$$B_{Tr} = \{x_i \in B,\ x_i \text{ was not audited}, i \in (1, ..., n)\}$$

$$B_{ctrl} = \{x_i \in B,\ x_i \text{ was audited}, i \in (1, ..., n)\}$$

We can write the propensity score as the following :

$$p_{score}(x_i) = \mathbb{P}[C_i = 1 \mid x_i]$$

We estimate the propensity scores for all observations in both B_{Tr} and B_{ctrl} using logistic regression with *LASSO* Regularization [4], and obtain for any observation x_i : $\widehat{p_{score}}(x_i)$
We define the matching distance d_{PSM}: $B_{Tr} \times B_{ctrl} \to \mathbb{R}$

$$\forall x \in B_{Tr},\ y \in B_{ctrl}, d_{PSM}(x, y) = |p_{score}(x) - p_{score}(y)|$$

We can eventually define the Propensity score matching as the function $\Psi : B_{Tr} \to B_{ctrl}$ such that :

$$\forall x \in B_{Tr},\ \Psi(x) = \underset{y \in B_{ctrl}}{\text{argmin}}\ d_{PSM}(x, y)$$

We apply the propensity score matching model to our data after two step of feature selection :
- We want to keep only significant features in the model in order to remove noise for the matching. We therefore use Coefficient significance test in the process of the LASSO regression for estimating the propensity score vector. We choose a level of significance of $\alpha = 0.05$.
- The propensity score is an equilibrium score: if it is accurate, the distributions of the features must, conditionally to it, be relatively similar between the treatment and control groups [2]. Thus, to evaluate and refine its quality further, we stratify the vector of propensity scores and for each stratum, we are interested in the distribution of features between the two groups. A good balance will thus indicate an accurate propensity score. Therefore we want to minimize the *SSMD* norm of the propensity score vectors by stratification. For this step we choose $L = 5$ strata as it is sufficient to remove 90% of the bias[5].
After those steps we keep a model comporting 22 features for the matching. We get a final model predicting $n_{F^*} = 53,174$ potential fraudulent companies which also represents approximatively 29.4% of the total number of the control group.
We conclude from this number and the proximity with the random forest approach that a potential bias on the assignment to treatment group can't be inferred from those results and without further analysis of the observations predicted as fraudulous.

IV. Regression problem : estimating the tax shortfall

Finally we estimate for each company which was predicted as fraudulous the potential adjustment amount M_i^*. Therefore (after making assumption 1) we apply a linear regression to the results of both the random forest and the matching.
For the choice of the model we decided to minimize the *BIC* criterion. We had to address multicolinearity issues using VIF scores [3], and select the features using the *regsubset* R package. We kept a final model with 9 features. The model was relatively poor considering the data which

was hardly gaussian, with an *adjusted R-squared* around 0.3, but we could still use it to compare the results between the two approaches (RF-classification and matching).

Let $\widehat{\beta} \in \mathbb{R}^p$ the coefficients vector estimated by the linear model.

$$\forall x_i \in B_{Tr},\ i \in \{1, ..., n_{Tr}\}, \quad \widehat{M_i^*} = x_i\widehat{\beta}$$
$$\implies \widehat{MAG_i} = \begin{cases} x_i\widehat{\beta} & \text{si} \quad F(\Psi(x_i)) = 1 \\ 0 & \text{sinon} \end{cases}$$

We then compute :

$$\widehat{MAG} = \sum_{i=1}^{n} \widehat{MAG_i}$$

For the two approaches we get very different results : After combining the linear model with the random forest classification, we get an estimate $\widehat{MAG}^1 = 141,140,106$ whereas the Propensity score matching metghod gives us a different estimate $\widehat{MAG}^2 = 86,258,571$.

V. CONCLUSIONS

From these results we conclude that there is surpisingly no obvious bias on the probability of being potentially fraudulous, as the numbers of companies being predicted as fraudulous was very similar between the two approaches and even significantly similar to the effective fraud ratio among Treatment group.

It appears that the differences between the two results lie in the companies which were predicted as fraudulous, as the estimation of shortfall using the same model gave such extreme differences. After further analysis of the observations and model predictions, we observe that the random forest method assigns $F_i^* = 1$ to companies with a relatively large adjustment amount, which also happen to be quite large companies. On the contrary, the propensity score matching method designation as fraudulous seems to have no significant link with the size of the companies, and thus potentially fraudulous companies have on average smaller adjustment amounts attributed.

We conclude with the existence of a selection bias (treatment bias) influing on the process on selecting companies for audit. But unlike we expected this bias does not rely only on high probability of being fraudulous, but more on high probability to have large adjustment amounts attributed. In other words, the control agents tend not to select the companies which are the most suspicious, but the ones which, if they reveal fraudulous, would have huge amounts of adjustment to pay : they don't try to get the most fraudulous companies as they can, but the highest amount of adjustment money. Maybe one could thus interrogate this strategy and arbitrate between the two in order to optimize shortfall revenue.

REFERENCES

[1] Peter C. Austin. *An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies*. 2011.

[2] Peter C. Austin. *The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies*. 2009.

[3] Alboukadel Kassambara. *Machine learning essentials*. 2018.

[4] Yating Liu. *Cours d'apprentisatge statistiques*. 2021-2022.

[5] Paul R. Rosenbaum and Donald B. Rubin. *The central role of the propensity score in observational studies for causal effects*. 1983.