Manoogian Simone Research Fund

Developing alternative methodologies of tax evasion identification

Final Report

This research has been implemented in the scope of Manoogian Simone College, financed by the Manoogian Simone Foundation (MSF) and in cooperation with AUA and the Armenian Government.

Contacts:

Principal Investigator: Vardan Baghdasaryan
40 Baghramyan Ave. 0019, Yerevan, Armenia
American University of Armenia
College of Business and Economics
Tel: +37460612567
E-mail: vbaghdasaryan@aua.am

September 2020

Yerevan

Contents

Preface

This is the Final Report of the MSRF "Developing alternative methodologies of tax evasion identification" Project carried out over the period of one year. Majority of deliverables, including the codes of the developed fraud identification models as well as training of SRC staff has been already implemented.

This report is submitted as a follow up to the Draft Final Report submitted on July 26. It mostly repeats the draft report but also briefly describes the results obtained as a result of analysis of importing firms. The team will transfer also these results together with all the remaining codes.

The results of the work will be presented at the conference on Tax compliance in Italy (online) at the end of October. Once the final paper will be edited and produced it will be shared with SRC and MSRF. It is important to highlight that the paper will summarize the work already conducted in a manner suitable for academic publication in the future. In this case full acknowledgement of MSRF financial support and SRC technical support will be recognized.

Finally, we remain at the disposal of SRC to discuss the results of actual audits to be conducted during 2020-2021. At the same time, it is important to highlight that major shocks caused by pandemic and war will most probably contaminate the results. In this regard, we believe that the results obtained so far, without even formal field experiments, suggest that the toolset developed can be useful for anomaly detection and should be considered for further deployment at SRC, in parallel with existing rule based approaches.

1. Background

This report summarizes the work carried out by MSRF Tax Evasion Research team on building and applying an alternative tax evasion detection model. To avoid reiterating all the research activities implemented, here we summarize shortly the previous stages of the work included:

a) Modelling object:

Our starting point is the universe of all profit tax payers registered in the SRC system as of 2016/2017. Later we divided all the taxpayers into two distinct groups: those who operate under VAT system and those who operate under turnover tax system. The reason for such a distinction was dictated by considerable difference in data collection formats.

As a result, we obtain the following two groups:

- 14,564 Profit tax and VAT tax payers (as well as profit tax payers)
- 58,373 Turnover tax payers

In addition to this we use also:
- month level aggregated tax receipts
- month level aggregated invoices

For the application of supervised learning techniques labelled data is required. For this part of the analysis we use data provided by the SRC on tax audits and results of those audits. The data on audit types is historically available since 2006. We categorize as fraudulent an enterprise which was fined as a result of such an audit. While in certain cases this approach is straightforward (e.g. whenever the taxpayer was fined for not printing and providing the customer tax receipt and was fined), it is not a priori clear whether fines resulting from other types of audits in fact represent a purposeful tax evasion. Hence we acknowledge the possible limitations of using labels provided by the SRC, but at the same time are constrained to use them due to absence of any other reliable alternative. **Broadly defined, what we label here as fraud is in fact misreporting.**

The overall approach is as follows: using data on audited entities we are trying to recover data-driven features that are associated with high fraud probability. Using that features we should be able to predict with a certain level of accuracy potentially fraudulent taxpayers also among those who were not audited.

After obtaining our final model (discussed below) we test it on 2019 data, which is completely unseen data. After validating the model, we run it for 2020 and obtain the predictions for the current year. When running the model for 2020, we incorporate the information from 2018 into the model. So the model we apply is adaptive, it evolves by incorporating recently revealed data.

b) Feature engineering:

It is by far one of the best methods of achieving competitive model performance. It is the process of creating/calculating new covariates based on the available independent variables. Various feature engineering techniques exist and were experimented on the dataset. The newly created features can be grouped into three categories: ratios, trend/growth features and clusters.

Ratios: Productivity, Profitability, Share of exempt and zero VAT turnover in total VAT turnover, Share of Admin costs in total costs, Share of Direct costs, Share of Indirect costs

Trend/growth features: Employee average growth over years, Tax receipt (number) y-o-y growth rates for 2016-2017 (based on monthly aggregated data), Tax receipt (sum) y-o-y growth rates for 2016-2017 (based on monthly aggregated data), Mean and standard deviations for tax receipts

c) Consideration of various supervised and unsupervised learning methods:

The target (dependent) variable in the supervised models is used from the 2018 dataset while all the independent variables represent the data as of 2017-year end. The above mentioned techniques were used for two different sets of tasks: fraud classification and audit classification. Fraud classification models utilize the data of taxpayers that were audited in 2018 and were found to be fraudulent or not. The covariates of the model, as mentioned, are reported as of 2017 to avoid reverse causality. The output of the model is the probability of being fraudulent for a given audited taxpayer. The audit classification models utilize the data of all taxpayers. The dependent variable shows whether a given record was audited in 2018 or not. The covariates again are extracted from the 2017 database for the same reason. The output of the model is the probability of being audited by SRC which we also interpret as the probability of being fraudulent from the SRC point of view (i.e. if SRC strongly believes that a given Taxpayer is not fraudulent, then it will most probably not be audited, where SRC belief is developed based on set of rules and approaches for fraud detection that they have).

The audit classification model provides us the opportunity to learn the audit rules. This is further used for two reasons. First, the audit classification model helps us to learn the existing rule set used for tax evasion detection and have a baseline model for comparison. In other words, for a given taxpayer, we can provide the estimated probability of being fraudulent (estimated by our fraud classification model) and compare it to the probability of being audited by SRC, which basically represents the probability of being fraudulent from the SRC point of view. Those two are compared by the lift score, which shows how many times our model is better than the existing SRC model. Second, the fraud classification model provides us rules that extract a segment of potentially fraudulent taxpayers. Yet, those rules are learnt on audited data only, which creates selection bias: first a record should be subject to audit and then only the fraud rules will be effective. Thus, learning audit rules helps us to have the complete set that takes information of a taxpayer input and tells whether it is a) subject to audit or not, and b) if yes, then whether it is potentially fraudulent or not. This kind of approach helps to increase the precision of audit as it creates stricter audit rule set and decreases the possibility of having false positives.

Logistic Regression: parametric technique used with categorical outcome (2 or more categories). The regression approach is used to analyze and predict probabilities of a variable belonging to a certain class or category, usually binary.

Decision Tree: non-parametric approach, which builds classification/regression models based on the tree structure, with a set of if-then-else decision rules which split the data to the most possible homogeneous segments.

Random Forest: non-parametric algorithm that is a collection of a large number of individual decision trees that operate as an ensemble, that is it uses multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone. Each individual decision tree provides a class prediction and at the end the class with the highest number of predictions from tree collection becomes the model's prediction.

Gradient Boosting: non-parametric boosting technique, which sequentially transforms weaker learners into stronger ones, by building each new predictor model based as the improvement of the previous one. The predictors can be chosen from a range of models like decision trees, regressors, classifiers etc.

d) Selection of the prediction model

Fraud Models: Overall 8 different fraud classifications methods with various specifications were experimented. Meanwhile, for the majority of the models the hyperparameters were chosen based on the grid search results, which is used to find the optimal hyperparameters of a model, that is the most 'accurate' predictions. Table 1 provides the summary of evaluation metrics of the experimented methods.

Table 1. Results of 2018 Fraud classification models

Model	ROC AUC on 2018 train	ROC AUC on 2018 test	ROC AUC on 2019	2018 test LIFT_1	2019 LIFT_1
Logistic Regression	0.705	0.686	0.672	1.263	1.593
Decision Tree	0.695	0.673	0.681	1.134	1.525
Random Forest	0.704	0.693	0.698	1.392	1.322
Gradient Boosting	0.717	0.709	0.685	1.366	1.492
XGB Ranking	0.720	0.706	0.708	1.263	1.322
Ensemble*	0.725	0.707	0.682	1.366	1.661

The winner model for 2018 (Gradient Boosting) shows 71% ROC AUC in general. While this is quite low, the predictive power is much higher for 1st decile of ranked probabilities (which have threshold probability of fraudulency higher than 80.4%).

Then the models were applied to 2019 unseen data, in order to check how general is the chosen approach and for having confidence for accurate predictions of 2020 fraud cases based on the updated 2019 model. As can be seen from the table, the results of 2018 test set and 2019 validation are quite similar, therefore we decided to retrain and update the model based on 2019 data set.

Table 2. Results of 2019 Fraud classification models

Model	Logit	DT	RF	GB	XGB Ranking	Ensemble
ROC AUC train	0.71	0.70	0.73	0.77	0.72	0.76

ROC AUC test	0.69	0.65	0.70	0.73	0.70	0.72
Precision train	0.62	0.65	0.59	0.68	0.60	0.67
Precision test	0.59	0.61	0.58	0.61	0.58	0.61
Recall train	0.68	0.52	0.69	0.64	0.70	0.66
Recall test	0.64	0.46	0.66	0.59	0.67	0.60
Accuracy train	0.66	0.65	0.64	0.70	0.65	0.69
Accuracy test	0.64	0.63	0.63	0.65	0.64	0.65
LIFT	1.49	1.49	1.62	1.85	1.59	1.62

2. Final audit/fraud prediction model

Given the limitation of having only financial information of taxpayers, the fraud models, as expected, are not very precise. Yet, we suggest a framework for implementing fraud models that is believed to provide the opportunity of accurate identification of potentially fraudulent taxpayers that otherwise would not be audited. Below are the steps for the suggested implementation framework.

1) **Identify threshold.** The first step is to learn the probability cutt-off value above which model results in confident and accurate predictions. As an estimate for this threshold we propose to use the minimum probability of top decile, i.e. only 10% of taxpayers have higher probability than the suggested value in the historical data. In our case, this value was 50.1%.

2) **Predict using a fraud model.** Secondly, use the fraud model to estimate the probability of tax evasion for all the taxpayers.

3) **Select those above cut-off.** Third, choose only taxpayers that have probability estimates above the calculated threshold value. This will be the set of taxpayers who the model is very confident in being fraudulent.

4) **[Only for validation] Check whether they would be audited otherwise.** To learn whether or not selected taxpayers would otherwise be audited if current rules are used, the audit model should be used.

The implementation of the above suggested framework on the historical data (2019 audits) showed that there are 2,504 taxpayers that the model selects at step 3) (i.e. it is confident in their fraudulency) but only 1,399 of them are captured as positives by audit model (i.e. only 1,399 would be considered for audit). It is also possible to implement the framework in a stricter approach by choosing a subset of taxpayers that intersects based on the expert knowledge recommendation or criterion.

a) Predicted TINS (comparison with SRC's TINS)

Using the model and the approach outlined above, we predict the probability of audit and fraud for about 25,000 taxpayers. Then we obtain the list of TINs that were identified by SRC risk management system for comprehensive audits in 2020 - 798 taxpayers. This list is compared with audit and fraud models. (see table below).

Table 3. TINs predicted as fraudulent for 2020 and comparison with SRC list.

Model probability Decile	Classified by Audit Model		Classified by Fraud Model	
	Count	Audit_prob	Count	Fraud_prob
1	580	0.677	326	0.595
2	138	0.265	95	0.47
3	49	0.097	21	0.436
4	23	0.04	49	0.405
5	7	0.019	0	0.394
6	0	0.01	19	0.391
7	1	0.008	38	0.373
8	0	0.008	101	0.336
9	0	0.007	89	0.273
10	0	0.007	60	0.183
TOTAL SRC Audit list 2020	798		798	

As it can be inferred from the table the Audit model captures quite well the current approach of the SRC. The fraud model instead provides an interesting outcome. In this regard, it is important to highlight how the models could be tested against the reality.

1. Each decile predicted by the fraud model includes around 2,500 taxpayers. One way to proceed would be to choose taxpayers from 1st decile and some lower deciles and audit them. The model would be validated if the fraud rate in the first decile taxpayers would be above current actual fraud rate as a result of audits in SRC (around 70%), whereas the fraud rate in lower deciles would be below that threshold. But obviously, this approach is costly and requires legal remedies, as currently the audits are conducted based on a specific legal framework.
2. Alternative approach is to utilize the current SRC audit framework, which we claim gives qualitatively similar results with a statistical caveat that we discuss below.

In particular, the fraud model places a bit less than half of “to be audited” in the first decile. At the same time, a considerable number of such taxpayers are also found in the 8th, 9th and 10th deciles. According to our fraud model, the latter should have considerably lower chances of being fraudulent. A viable validation of the fraud model would be an outcome, where the actual fraud rate (fraud rate = # fraudulent / # audited) in the first decile is higher than the one in the last 3 deciles. The only caveat here would be statistical power of these results as the number of experimental units in various deciles is not equal to each other.

3. Finally, we also provide a list of TINs to SRC that we believe have the highest chances of being fraudulent, but we keep in the list only relatively large taxpayers (List will be provided separately from this report).

3. Possible options to improve the fraud model.

We have sought expert opinion to verify the findings of the model. The audit and accounting professionals teaming up with us in this research project have reviewed the features we have included and have made a number of noteworthy recommendations/comments:

1) Starting January 1st 2018 the taxpayers subject to tax audits have been given the opportunity by the law to implement revisions to tax reports during the process of the audit. In such cases there might be certain irregularities unveiled by the tax auditors, which with the consent of the taxpayer, can be corrected as a part of audit process. As a result, no fines are applied to the taxpayer and hence in accordance with our approach to classify the taxpayers, they would not be considered as fraudulent. This might also include cases where the "mistakes" were not done by chance. Such a taxpayer in 2016 or 2017 would be classified as fraudulent. Hence these changes might hinder problems for the validity of our models when we move on predicting 2019 or 2020 frauds.
   The recommendation in this particular case is to incorporate the information on whether the taxpayer has implemented such corrections. There are two ways to proceed:
   a) One possibility would be to classify as fraudulent also the taxpayers who have implemented corrections. This would kind of assume that legal environment of 2016-2017 is persisting also for subsequent years. But this approach would create problems in the future, as it ignores the change in the "rules of the game", hence the model would not be that useful for predicting future outcomes.
   b) Second approach would be to incorporate the fact of carrying out corrections during the audit as additional features to the model. This can be just a binary variable or provide details on how many such corrections have been made.

2) According to the law, the taxpayer has the right to appeal the decision made by tax auditors. It is important to verify that the labels of fraud we are using are based on the final, post-appeal decisions to apply fines to the respective taxpayers.

3) Another avenue for improving the model would be to analyze for which type of taxes the fines are applied (this is relevant for budget audits, which comprise an important part of the audits in our models). There might be occasions, such as, for example, the case of environmental fees, when the taxpayers thought fined, should not be considered as fraudulent. Of course, this will depend on the final purpose of the model.

Another interesting observation made by the accounting experts pertains to the ratios we have included in the models:

1) We have included a ration on administrative costs (admin costs as a share of total eligible costs). Admin costs are defined by respective IFRSs, but the tax legislation in RA is vague on this. The classification of the costs as administrative is left to the accountant's discretion, so this indicator might largely differ from one taxpayer to another.

While from an accounting point of view the experts consider the inclusion of this feature problematic, we find it extremely useful for our modeling purposes. The admin ratio in our models have high contribution in the process of classification, hence in the future we suggest to utilize this type of features (which might be potentially used to mask some costs) more.

Currently most of the ratios we have constructed are utilizing the VAT form. Our audit experts suggest to pay attention also to the profit tax return. One particular suggestion is the following:

- Part 8 of the profit tax return form contains data on accounting and tax base. In particular, the decrease in assets and increase in liabilities, calculated on tax base should be compared with the eligible costs. At the same time the audit experts note that the taxpayers do not properly fill in this part of the return and there is a practice of not paying much attention to it by SRC tax auditors. So it is possible that some inconsistencies will be captured in there.

Audit experts have also commented on our other features and we have made sure that the ratios included in the main model do have economic meaning. As a result of this cross-check we have removed certain problematic features.

4. Additional initiatives

a) Taxpayers' actual geolocation determination

COVID-19 created specific challenges for SRC. During the lockdown period it was important to follow on economic activity in its spatial dimension. This requires an accurate geolocation of each taxpayer. Currently, in the SRC electronic system the address is self-declared and there is no unique format for inputting the addresses. SRC embarked on the task of refreshing the addresses of the taxpayers with the main objective of obtaining accurate locations for not only registered taxpayer, but its operational locations.

Our team was asked to provide assistance in this process, in particular to suggest ways machine learning could be applied in the process. SRC was able to obtain a roster of addresses that was intended to be used by State Cadastre. We were assured that those addresses were existing ones. Given the SRC database addresses, we implemented a multistep procedure for obtaining both the geographical measures and their accuracy estimates. First, we used the API (Application programming interface) of the Yandex Maps to obtain the geographical coordinates of those addresses. Second, we used Levensthein distance, which is a statistical measure for finding similarities between textual values. We used it to match SRC database addresses with the State Cadastre addresses to find structured version of user inputted address. Third, we obtained the geographic coordinates of those matched and structured addresses from the Yandex Maps API as well. Fourth, we calculated the distance between those two geographic coordinates. In case of perfect match and structured user inputs we would have distance between coordinates. Yet, given that user input in not fully structured, we use the aforementioned distance and the similarity between textual addresses as accuracy measures. Both the address with coordinates and accuracy measures were provided to SRC[1].

[1] In the draft report we had mentioned our readiness to help further with obtained correct addresses of operations. During a meeting with SRC we were notified that they have successfully completed identification of addresses.

b) Importers’ similarity identification

It has been observed that among major importing companies, some create small independent enterprises to divide their total earnings into several parts which will let them avoid surpassing the revenue threshold for the turnover tax. In order to address this issue of threshold-based splitting of earnings, we decided to implement product portfolio similarity-based analysis for the selected companies.

It was decided to limit the country list only to China, Turkey, and the United Arab Emirates and to select all of the companies who are importing from there. During the second step, we found all the companies who directly buy from these importing entities, based on the invoice data. The main assumption that we wanted to assess is based on the premise that these buyers are in reality sub-companies and instead of declaration based supplier-buyer relationships they have more parent-child kind of relationship.

In order to address this issue, we assume that if it's the same company behind two or more legal entities, then their product portfolio should be similar to each other, thus should have a very high intersection. As the number of unique products sold by different companies can differ in terms of volume, we normalize this intersection. More specifically, we divide it by the union of products and use Intersection Over Union as a similarity indicator (IoU). Higher the IoU, higher the risk that those 2 companies can have a common entity behind. IoU equal to one means that those 2 companies sell the same range of products. As a result, the pairs of companies having high IoU are flagged to be possibly connected.

We applied the above described approach. In particular we selected the importers and their direct counterparts (businesses acquiring from importers) from the above mentioned countries focusing on the ones which operate under the turnover law were selected. As a similarity measure, the intersection over the union (IOU) of sold products was applied. More specifically, companies which have a similar product portfolio will have higher (IOU) and thus will be considered more homogeneous. Afterwards, the turnover amount of the most similar partners was compared to see if their sum approaches the legal threshold. Unfortunately, this experiment demonstrates that our assumption was not supported and can be explained that actually companies split into more specialized smaller units rather than homogenous ones.

In addition, there are some limitations to this approach. First, for the product identification, we use ADG four-digit codes, which are more general and indicate group-wise belongings. Thus in the future, they can be replaced with lower-order code for more detailed and precise analysis. Moreover, instead of codes we can use the product names and implement natural language processing techniques, such as word2vec or 1d clustering on word vectors, for assessing similarity. Second, based on their total turnover companies are legally allowed not to specify what products they are selling. As a result, we are left with a very small number of companies for our analyses, which in reality can be not representative.

Another direction we followed was about identifying importer similarity based on the direct partner fraud or audit rate. More specifically, importers from China, Turkey, and United Arab Emirates (UAE) and their direct partners (companies they sold directly) were identified and selected. For each importer we calculate the average percentage of fraudulent partners separately for fraudulent importers and non-fraudulent importers. The results indicate that if the importer is fraudulent then on average 57% of his partners are fraudulent, else 44%. In order to validate the accuracy and to be sure that results are not biased, the similar analysis was implemented for the audit percentage. The results are presented in Table 4 and indicate that there is no difference in terms of audit rate. While these evidences are at descriptive level, this is an important result to be developed in future – looking at “fraud propagation” in network relationships. In this regard, further research will be required but at the moment the SRC can try to incorporate this information into their risk identification models.

Table 4. Importer Partner Fraud and Audit Rate[2]

Importer	Partner Fraud Rate	Partner Audit Rate
Fraudulent	57%	34%
Not-fraudulent	44%	34%

5. Capacity Building

As envisaged by our project proposal, we conducted a short training for SRC specialists who were interested in learning details on modeling or will be working with fraud detection models in the future. Overall training duration was 12 hours, which, upon consultation with SRC counterparts were divided into 6 two hours’ sessions and delivered over a duration of three weeks (see schedule below).

Date:	Topic:
Day 1 (June 23).	Intro to Python – data types and structures, loops and control flow (if/else)

[2] Corresponding code files, together with already provided ones, will be passed to SRC specialists.

Day 2 (June 25).	Intro to data analysis with Pandas – reading/writing data, filtering and slicing, calculating summary statistics and correlations, preliminary visualization
Day 3 (June 30.	Case study: data analysis with pandas – case study with descriptive analytics and visualizations using pandas (the top Python package for data manipulation)
Day 4 (July 2).	Intro to classification with Decision Trees – introduction to classification in Machine learning, Introduction to the theory and intuition of Decision trees, implementation in Python
Day 5. (July 7).	Performance evaluation and feature importance – measuring performance of machine learning model (with the example of Decision Trees), explaining the patterns learn by the model form data, understanding reasons behind predictions
Day 6. (July 14).	Models developed for tax evasion – case study on some of the approaches and techniques used by our team on PEK data to detect tax evasion.

Training was delivered via zoom platform and participants were joining either from SRC conference hall or from their home computers. We were using Google Colab to developing samples of python codes to avoid installation requirements. The trainings were video recorded and provided to SRC for those employees who would like to follow but were not able to due their working schedules.

The purpose of this short training was to introduce the fraud models that our research team has developed and to make sure that underlying principles are clarified in case the SRC decides to continue with the proposed framework.

6. Structural decisions for effective future deployment of Machine Learning in SRC

In order to deploy modeling in general and ML in particular the SRC will need to take some additional steps. We have formally separated those into two groups:

- Internal structures:
    1. As the current project demonstrated, a huge part of the modeling work pertains to data extraction, cleaning and cross-checks, which we dub data engineering. With the amount of data and externally managed databases, data engineering requires an internally dedicated task-force.

2. The SRC would benefit from setting up an analytical team that would start deploying modeling and ML into their daily operations. From our interactions with the relevant specialists, we believe that this can be accomplished with minimum additional hiring.

- External partnerships:
  3. To develop specific projects of interest SRC can team up with academic teams. The SRC possesses data which makes it extremely attractive for scientists from abroad as well. Such projects, with a bit more advanced capacity building element compared to current project, will also help to develop in-house capacities. But it is important to highlight that the benefit of these projects is real only if the SRC has got in-house modeling team that can benefit from the positive spillovers and cooperation.
  4. It is also possible to maintain some more institutionalized contacts with academic institutions by having joint research institutes. A relevant example would be the cooperation between the AUA and Central Bank of Armenia, and the jointly run Master of Science in Economics program.

7. Project Closure steps

1) **Transfer of the models, data files and other related materials to respective SRC specialists.**
   Once SRC will determine who will be accepting the models, we will transfer all the data on the relevant computers. If needed, additional on the spot assistance will be provided to run the models.
   Upon completion of the procedure the AUA will send a request to SRC to remove its stationary computer back to AUA premises (as per terms of the grant project).
2) **Development of academic paper draft**
   The work undertaken on the part of fraud prediction is conclusive and we believe can be developed into a short academic article. The paper will be presented at an online conference "Tax Compliance: new methodological and empirical approaches" in Italy on 29-30 of October, 2020.[3]
3) **The analysis of experimental results (based on 2020 audit program implementation by SRC).**
   The results of SRC 2020 comprehensive tax audit program will be available only in 2021 but we are interested in analyzing and comparing the actual outcomes with the ones predicted by our models. Though we will provide all the relevant data for the comparison to SRC so that they conduct their own comparative study, we will be available to conduct this analysis and discuss its outcomes with SRC IT leadership.

[3] The conference preliminary program is attached to this report.