

Gap Estimation Exercise

The aim of this exercise is to predict tax assessed using the return and audit data, and proceed with the tax gap estimation. The exercise follows Part 2 of the Toolkit.

From Part 1 to Part 2

At the end of Toolkit Part 1, you have a cleaned and matched return-audit dataset. This dataset is the starting point for Toolkit Part 2. In this practical, we provide the cleaned and matched VAT return-audit dataset. The VAT is used for demonstration purposes.

Exercise 1 – Stata Environment setup & data import

In this exercise, you will try to implement the First Stage of the Toolkit Part 2. Ensure the random forest machine learning algorithm is installed in Stata, prepare the Stata environment, set up the global directory paths and import the data.

- a) Run the code lines 25 to 30 to prepare the Stata environment.
- b) Modify line 33 to your own file directory and run lines 34 & 35 to set the directory paths for the data and results. Open the vat_audit Stata data (.dta file)

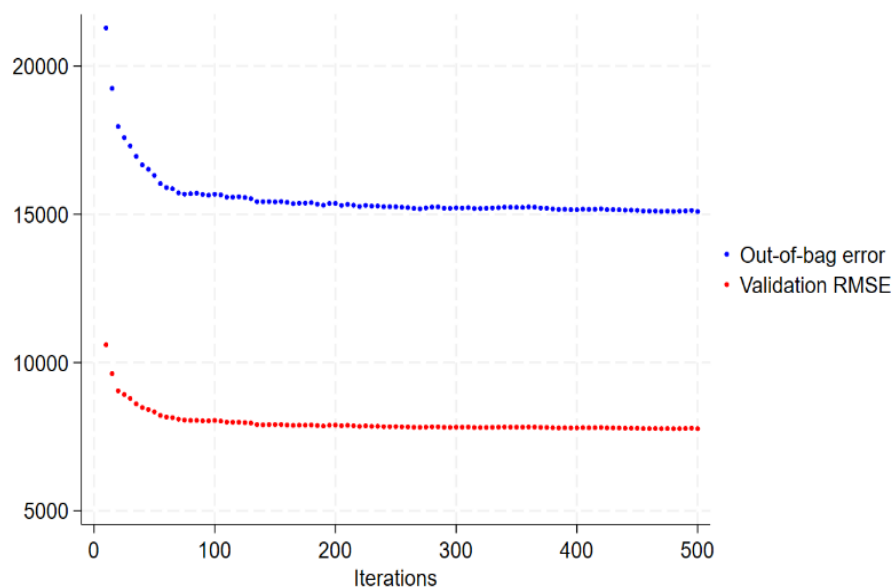
Exercise 2 – Choosing the Model of Machine Learning

In this exercise, you will try to implement the Second Stage of the Toolkit Part 2. Set up parameters, split the data, tune parameter one (number of iterations), and tune parameter two (number of variables).

- a) Create a dummy variable “reg_audit” that indicates if a taxpayer is audited or not (0/1).
- b) Create an “evaded” variable with the specific liability amount from the audit outcome. Keep observations with audit outcomes.
- c) Split the sample (50/50) into two groups. This forms the training sample and the testing sample.

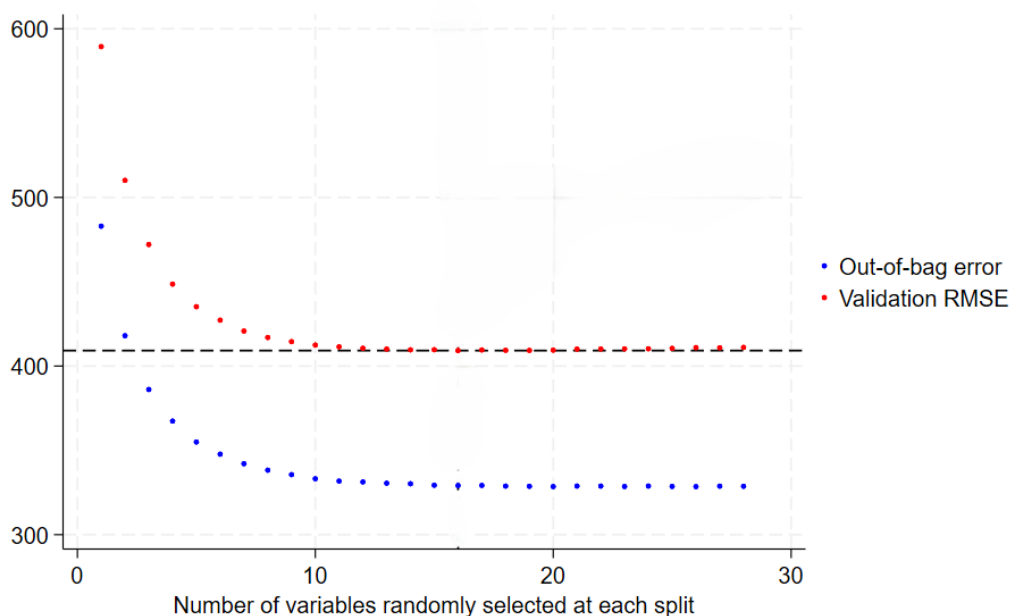
d) In machine learning, the “number of iterations” (the number of times model parameters are updated during training) is what’s called a hyperparameter. Setting this value appropriately is important to prevent overfitting, where the model learns the training data too well but performs poorly on new, unseen data. Typically, as iterations increase, the model’s error on validation data (like “Out-of-bag error” or “Validation RMSE”) will decrease significantly at first. However, after a certain point, the error will stabilize, and further iterations may offer no substantial improvement or could even lead to worse generalization. The provided figure illustrates how the model’s “out-of-bag error” and “Validation RMSE” change with an increasing number of iterations.

At approximately which number of iterations do both metrics level out? How many iterations should we train our model on?



- e) During model training, random forest will iteratively select a random subset of variables, often called “features”. While the model has a mechanism to assess and weigh the importance of each feature for its predictions, a challenge can emerge related to the sheer number of features used. Similar to how excessive iterations can be detrimental, including too many features, even if their importance is evaluated, can negatively affect the model via overfitting.

The following figure shows the model's performance over the number of variables. From the graph, select a number where it appears the Out-of-bag error and Validation RMSE stabilise. The value is used as the number of variables in the prediction model in the next exercise.

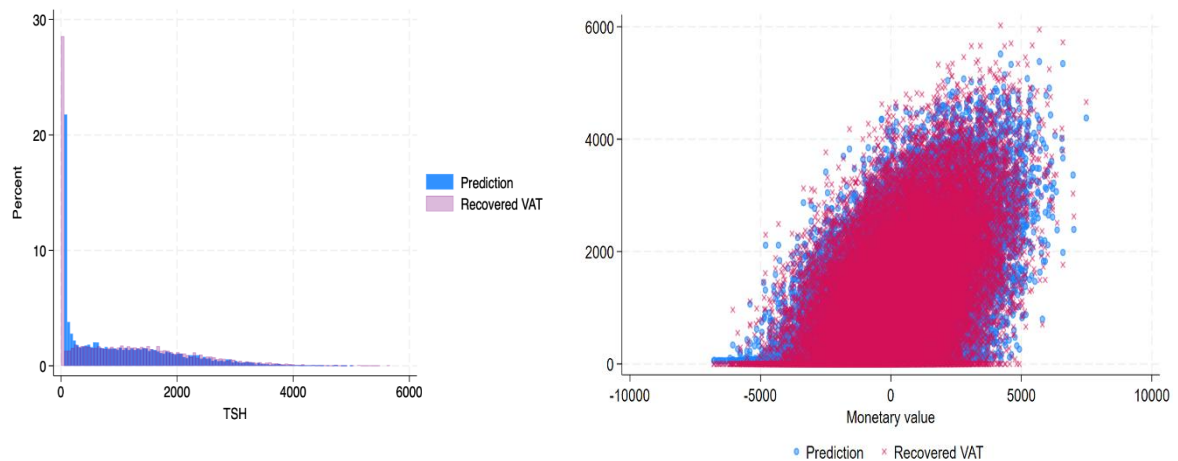


Exercise 3 – Testing the model

In this exercise, you will try to implement the Third Stage of the Toolkit Part 2. Before applying the model to the full dataset, the model is tested and checked.

- a) Run the random forest model with the parameters identified in Exercise 2 on the training sample.
- b) During model evaluation, the toolkit generates diagnostic plots that track prediction accuracy as model complexity changes. The following figure shows the model's accuracy by comparing predicted and actual assessment values.

What should you be looking for in these graphs to confirm that the prediction is accurate?



Exercise 4 – Estimating the VAT gap and disaggregation

In this exercise, you will try to implement the Fourth Stage of the Toolkit Part 2. The tested model is applied to the full dataset, and the tax gap is calculated.

- a) Repeat the steps in Exercise 2. Open the dataset, create a dummy variable "reg_audit" for all audits greater than zero. Create a new variable "evaded" from the "vat_evasion" variable.
- b) Create an indicator variable that flags records where the evaded variable is not missing, and reg_audit is equal to 1 (indicating audited regions). Create a new variable mimic_tax as a copy of the evaded variable, and replace the missing values in evaded with zero in mimic_tax.
- c) Run the random forest model with the parameters identified in Exercise 2 on the full sample.
- d) Predict the amount of undeclared tax (assessment) for unaudited records. Replace the assessment variable with the evaded amount for the audited records. The prediction is done! Save the dataset with the predictions to calculate the tax gap.

Exercise 5 – Estimating the tax gap

This is a continuation of the Fourth Stage of the Toolkit Part 2. The VAT gap is calculated

- a) Open the dataset. Collapse the assessed (vat_assessed) and VAT declaration (vat_declared) variables using the return year (rtn_year) variable to get year level estimates. Browse the dataset.
- b) Generate the potential VAT (pot_tax) by summing the assessed amounts (vat_assessed) and VAT declared (vat_declared) variables. Generate the tax gap variable. The tax gap formula is $\text{taxgap} = \text{assessed} / \text{potential tax}$ (multiply by 100 for a percentage).
- c) Plot the vat gap by the return year (rtn_year) and save the graph. What can you say about the VAT gap?
- d) You are asked for the monetary value of the tax gap. Plot the tax potential in monetary terms

Bonus Exercise: Investigating Compliance Challenges

The advantage of the bottom-up approach is disaggregated analysis to identify compliance challenges.

- a) Calculate the tax gap by region in 2005 and graph the VAT gap by region. Hint: follow the same steps as Exercise 5 but specify the year and region.
- b) Calculate the tax gap by sector in 2010 and graph the VAT gap by sector. Hint: follow the same steps as Exercise 5 but specify the year and sector.

Further disaggregation is possible to answer compliance questions in your own context. For example, firm sizes, sub sectors, etc.

Test your knowledge at home: Estimate CIT & PAYE Gaps

Repeat Exercise using CIT and PAYE datasets and variable list and parameters below:

CIT variables: rtn_year, fixed_assets, depreciation, investment, loans, provisions, dividends, reserves, debts, sales_turnover, wages, total_purchases, profits, operating_cost, taxes_paid, donations, total_profit, taxable_profits, cit_payment, city, region, isic, sector, tax_region.

CIT selection: 400 iterations ("iter(400)") and 20 variable replacements ("numvar(20)")

PAYE variables: rtn_year, number_employers, total_gross_payment, total_nontax_payment, total_tax_deductions, total_payroll_tax, total_chargeable_payments, paye_tax, paye_payment, city, region, isic, sector.

PAYE selection: 400 iterations ("iter(400)") and 20 variable replacements ("numvar(6)")