Hungary: Predictive modelling to forecast the fulfilment of the VAT return obligation

In view of the fact that the highest budget revenue is generated on VAT tax and the reduction of the VAT tax gap is of core importance, the central risk analysis unit of the National Tax and Customs Administration has initiated the development of an individual taxpayer risk model.

The purpose of the model is to anticipate changes in taxpayer behaviour at the time the first changes in the taxpayers' compliance occur, that is to say, at the earliest time when a change in the former behaviour of submitting tax returns is expected. We assumed that in the event of a change predicted sufficiently early, compliance can be maintained by addressing the taxpayer.

We expected external experts to develop a forecasting model based on statistical and data mining methods, which could predict which taxpayers we should contact directly to support the timely submission of the tax return.

At the beginning, risk analysts collected tax administration data and information that could characterize this form of behaviour, based on their professional experience. During the work, 206 professional and behavioural variables were developed, based on the data of taxpayers' employees, online invoicing, cash register operation, current account and representation data. In addition, 22 technical variables have been set up, based on the taxpayers' tax return habits. 64 other variables were used, which were developed for other modelling tasks in support of enforcement activities, but were expected to be useful explanatory variables in this case as well. 29 data sources were used during the development of indicators. A software analysis determined the strength of the indicators used to explain whether the taxpayer is significantly late or completely fails to submit the VAT return.

Preparation of the data was followed by the determination of the taxpayer population, which - according to our knowledge - previously had that type of behaviour (late submission for the first time or missing return) in the period 2016-2019. From these parameters, the model could "learn" about the relevant behavioural characteristics of the group. The behaviour was determined based on a data table containing the data on the tax return standards and filings. We found that it was necessary to clarify the definition, because it included a large proportion of public administration institutions, companies under liquidation and enterprises with no tax performance, which were not supposed to be included in the model.

Taxpayers were finally included in the machine learning stock in the following breakdown:

Learning File	Reporting frequency	Number of taxpayers		
		Missing/Late	Timely performers	Total
Non-submissions	Monthly	6 627	26 444	33 071
	Quarterly	13 563	54 245	67 808
	Annual	6 458	25 828	32 286
Late submissions	Monthly	5 010	20 040	25 050
	Quarterly	4 343	17 372	21 715
	Annual	2 148	8 592	10 740

The next step in the model development was the creation of homogeneous segments in terms of the characteristics of the taxpayers determining the risk within the given taxpayer group. Segmentation analyses showed that the main criterion is the reporting frequency (monthly, quarterly and annual) and the

time of submission of the return, as provided by the late submissions and by the number of defaulters. On this basis, six models were developed.

Based on preliminary results, models have significant distinction value in the segments. The distinction index is interpreted in the range of 0-1 and is considered very strong above 0.7. The following table shows the pre-estimated distinction power of the models developed in the segments.

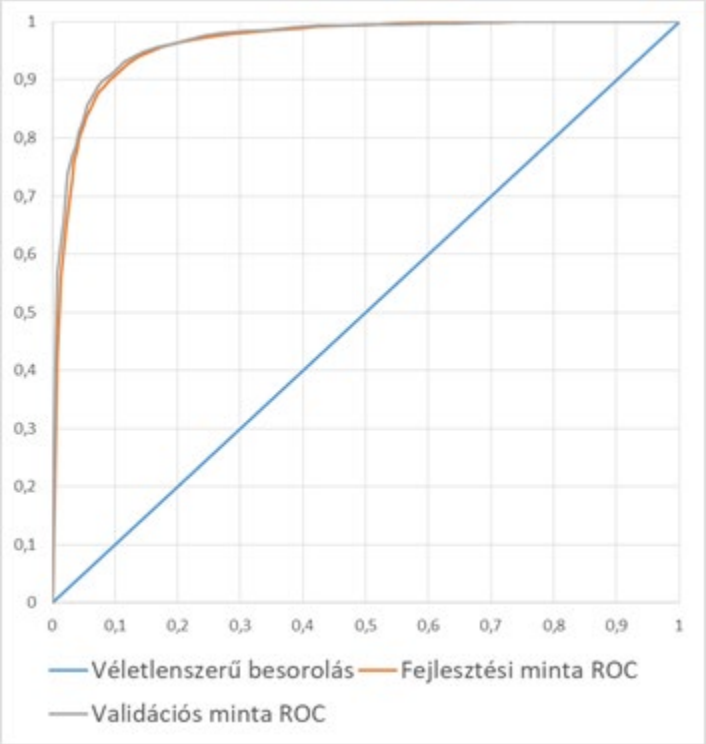
Model	Segment		
	Monthly	Quarterly	Annual
Non-submissions	**0.93**	**0.88**	**0.79**
Late submissions	**0.80**	0.67	0.55

In the entire modelling exercise, we finally worked with the 6 sub-models developed during segmentation. Of the explanatory variables, in total 32 were selected into the models by the algorithms, which originated from 15 data sources and one variable could appear in several models. Of the variables selected, 16 were used as explanatory in one model. The most important were those describing the previous tax return discipline.

We will present two sub models in detail, the strongest and the weakest ones.

1. During the modelling, the model developed for defaulting taxpayers with monthly returns was the strongest. The separation power of the model is shown in the ROC curve below on the development and validation data. The curve illustrates how much more likely we are able to choose defaulting taxpayers based on the model (marked with orange) than by random selection (marked with blue):

Figure 1. Sub-model 1

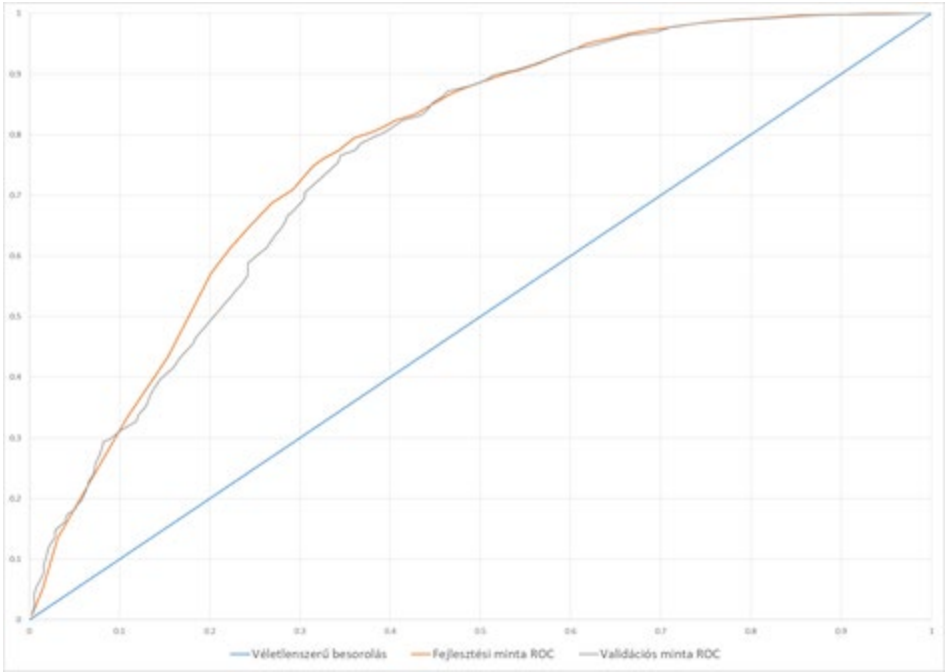


The following 11 variables have been used for the model:

- Intentional tax avoiders at the end of the tax return period according to the Information System for Especially Risky Taxers (15.95%)
- The number of days covered by tax return with a given code within 1 year after the VAT return period considered (15.14%)
- Business form code at the end of the VAT return period considered (14.40%)
- Tax account balance in HUF one month before the last month of the VAT return period considered (10.87%)
- Sales proportion of the two return periods before the VAT return period considered (8.26%)
- The way the company was formed (7.12%)
- Amount of outstanding fine, default penalty and self-assessment allowance established and not yet paid in 1 year ending at the end of the VAT return period under consideration (6.49%)
- Difference between the procurement in the two VAT return periods before the return period considered and the number of online invoices (6.45%)
- Whether there is an underlying obligation for limited companies at the end of the VAT return period under consideration (5.21%)
- Whether the taxpayer is on the list of taxpayers free of public debt at the end of the VAT return period considered (5.10%)
- Number of private persons with missing tax returns compared to the theoretical number of employees in the month before the last month of the VAT return period considered (5.02%)

2. The weakest model was the one for late taxpayers with annual tax returns. The separation power of the model is shown in the following ROC curve, also on both development and validation data. The curve illustrates how much more likely we are able to choose defaulting taxpayers based on the model (marked with orange) than by random selection (marked with blue):

Figure 2. Sub-model 2



The following 7 explanatory variables have been used for the model:

- Taxpayer's current account - opening amount for the month preceding the last month of the VAT return period considered (27.78%)
- Taxpayers' current account - balance of VAT for the month preceding the last month of the VAT return period considered (19.06%)
- Number of days covered by timely VAT returns for 370 days ending at the end of the VAT return period considered (17.93%)
- VAT returns due for 370 days (15.52%)
- Number of days covered by a VAT return with the given code in 1 year ending at the end of the VAT return period considered (12.70%)
- Rate of use of foreign services before the last two return periods before the VAT return period considered (4.17%)
- Proportion of sales in the last two return periods before the VAT return period under review (2.82%)

In the first back-testing, as illustrated in the graph below, the monthly return taxpayers of the September period were re-measured. Even during this period, the model had a good separation of risky taxpayers, bringing them together in a narrow range, with high risk scores.

Figure 3. Distribution of taxpayers with high-risk points

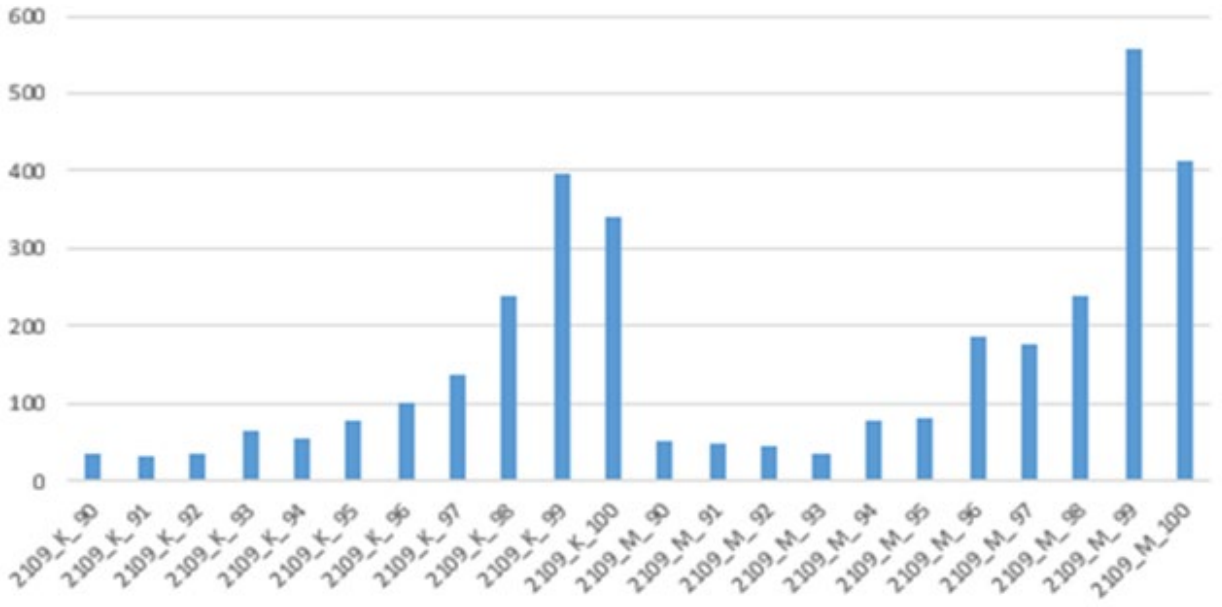
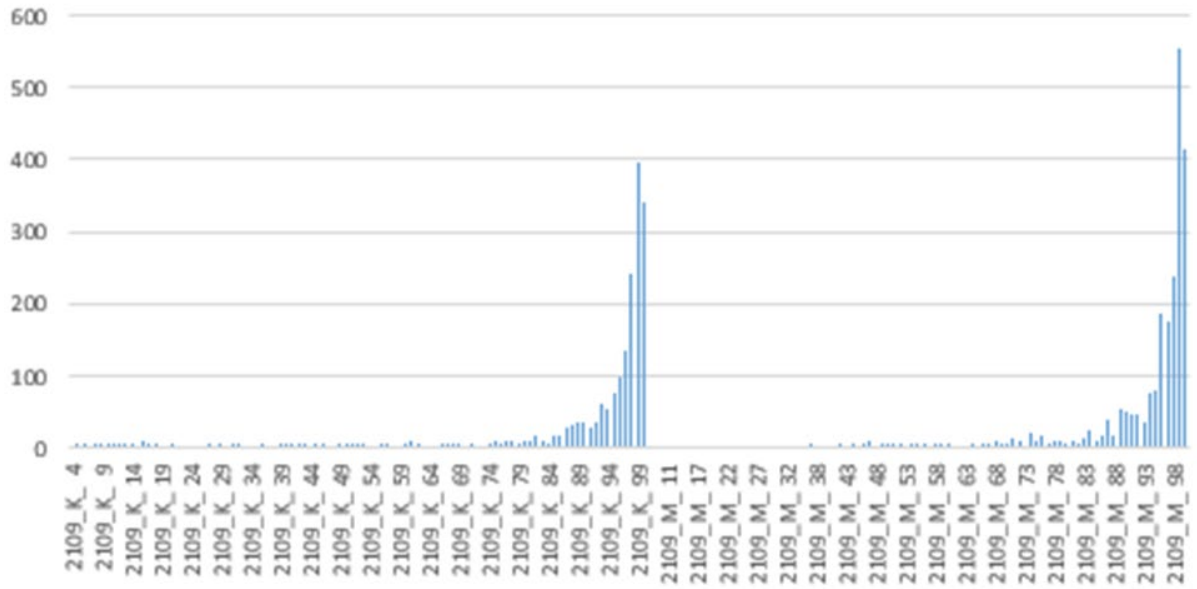


Figure 4. Distribution of the taxpayer population with scoring



The model scoring for taxpayers with monthly return rates was similar in October and November as well.

However, based on the model results so far, we believe that there is a need to further clarify the models, as the economic operators targeted by the modelling exercise represent approximately 0.1 % of the total number of economic operators concerned. Due to the highly unbalanced nature of the stock, it is considered appropriate to test other modelling methods, further refine the models that may require the training, programming and testing of additional variables.