Predictive Fraud Analytics: B-tests

Sergey Afanasiev
Executive Director, Head of Antifraud, Bank "Renaissance Credit", Moscow, Russia.

Anastasiya Smirnova
Head of Data Science, Bank "Renaissance Credit", Moscow, Russia.

Figures and Tables:

Contents

Abstract

In the banking sector, machine learning methods are applied in a wide variety of business areas: in assessing a client's risk profile (application & behavior scoring), to form targeted sales (x-sell, up-sell), when choosing collection strategies (collection scoring), etc. Also bank anti-fraud is not an exception, where with the help of machine learning methods effective anti-fraud tools are developed. This article deals with B-tests - methods by which it is possible to identify internal fraud among employees and partners of the bank at an early stage.

Key words and phrases: B-test, Benford's Law, Fraud detection, Predictive fraud analytics, Fraud modeling

1. Introduction.

In this article, we illustrate the method of trigger developing for internal fraud detection, which uses the concept of digital analysis based on the Benford's law.

As Bolton and Hand [1] noted, fraud detection methods may be divided into both supervised and unsupervised methods.

In supervised methods the samples with marked classes are used- "fraud" and "non-fraud". Thus, these methods allow us to identify fraudulent schemes that were used before.

In unsupervised methods there is no class markup, and fraudulent schemes are identified by analyzing abnormalities (outliers). Unsupervised methods can detect new fraudulent schemes, but the downside of these methods is that there is a large number of non-fraudulent cases (false positives) among the outliers.

To minimize the proportion of false positives, Lu, Boritz and Covvey [2] offer to use reinforcement learning methods, which allow to combine supervised learning and unsupervised learning approaches.

The methods for outliers detection suggested by Mark Nigrini [8] are based on the Benford's Law. In turn, the Benford's distribution is used only for the analysis of digits and has a number of limitations, such as the requirement of completeness (records must be complete) and the requirement of geometric distribution [6].

To eliminate the completeness requirement, Lu and Boritz [3] offer to use the adaptive Benford's Law method, which can be used to analyze anomalies in incomplete samples. At the same time, the requirement of the geometric distribution remains as a limitation for applying adaptive Benford's Law method for detecting anomalies.

The reason of our approach is that we broaden the typical methods of detecting outliers using the comparison component of segmented and verified samples, which do not have to obey the geometric distribution, as required by the Benford's Law. In addition, the samples can contain not only digital data, but also textual data (categorical and nominal), on which classic methods of digital analysis are not applicable.

We also show the benefits of the target variable "potential loss", created specifically to detect internal fraud. This target variable allows to configure operational triggers that detect fraud at the early stages when the organization has incurred small losses. In the typical classification for "fraud" and "non-fraud," trigger configuration algorithms give large amounts of samples (to meet the criteria of statistical significance). Therefore the triggers fire too late, when the company incurred large losses from fraudsters.

In this article, we explain the algorithm of our methodology and test our new fraud detection technology using real bank data. Bank secrecy and all personal data were excluded.

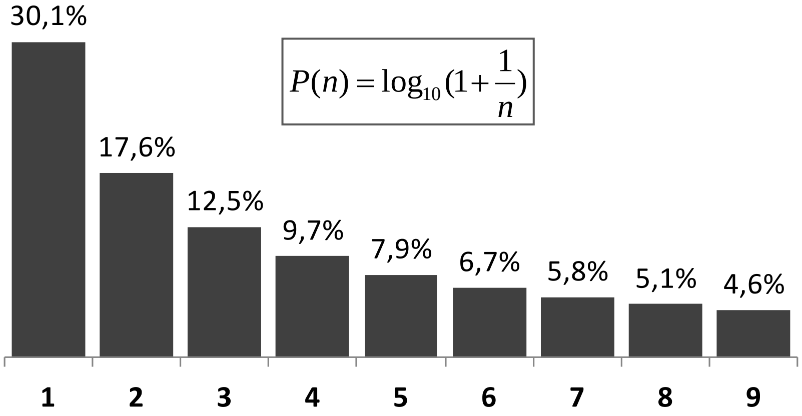
2. The Newcomb-Benford phenomenon

It is difficult to deny that the application of machine learning methods is economically profitable practically for any business activity: from increasing sales and cost reduction to increasing customer loyalty and forming an internal corporate culture. Just 100 years ago some of these methods did not exist at all, and some were just mathematical curiosities and were not used in practice. In the eighties of the XIX century American astronomer Simon Newcomb found out that the values of logarithms beginning with 'one' occur more often in the logarithmic directory than the values beginning with 'two' or any other number. Counting the frequencies of the first figures, Newcomb deduced the regularity, that later became known as the "the first-digit law". In 1881, Newcomb published a short article in the mathematical journal describing this law, which can be briefly formulated as follows: if any number is chosen randomly from a logarithmic reference, the probability that it will start from 'one' is 30.1%, with 'two' - 17, 6%, with 'three' - 12.5%, etc. Numbers starting with 'nine' will be least likely to occur – in about 4.6% of cases. In his article, Newcomb also gave a table of frequencies of the "second digit" where the differences were not so significant. Moreover, Newcomb suggested that the frequencies of digits at later positions would tend to be uniformly distributed [4].

The article published by Newcomb was ignored by the scientific community, until 1938 American physicist engineer Frank Benford drew public attention to this phenomenon. Benford checked the law of the "first digit" on the data of 20 different directories: the area of river basins, physical constants, the masses of molecular compounds, records of mathematical reference books, street numbers, etc [5].

In his study Benford analyzed more than 20 thousand observations and showed that everywhere the same law is observed: 'one' is more frequent than 'two'; 'two' is more frequent than 'three', etc. Later this regularity was called the Benford's law (Fig. 1), which was in its turn a confirmation of another phenomenon - Stigler's law of eponymy. It states that no scientific discovery is named after its original discoverer.

Figure 1. *Benford's Law*



Since the law of the "first digit" discovery (1881), more than 300 scientific works devoted to this phenomenon have been published. In 1997, an outstanding Russian mathematician Vladimir Arnold made a mathematical generalization of this law. In his report, he gave a mathematical proof that the Benford's law is satisfied for almost all geometric progressions, except progressions with the common ratio 10. Thus, Arnold succeeded to give mathematical explanation why the law of the "first digit" is often observed in socio-economic processes and physical phenomena, which, in their turn, often obey geometric distributions [6].

More than 100 years, the Newcomb-Benford phenomenon could not find practical application and it was attributed to the category of mathematical curiosities. In 1972, American economist Hal Varian suggested that the Benford's law could be used to detect fraud in the financial and economic spheres. Varian explained his assumption as follows: fraudsters subconsciously tend to fit the data to uniform distributions, so it's enough to compare the first figures of the statistical reports with the Benford's law and identify the anomalies [7].

Following this idea, the Assistant Professor of Accounting Mark Nigrini showed that the Benford's law can be used for forensic accounting and audit of financial statements. In the nineties, Nigrini conducted a number of studies, among which was the verification of tax declarations in the administrations of several states. Based on the results of this test, Nigrini revealed embezzlement of officials in the Treasury Department of Arizona, where about $ 2 million was stolen from the state budget in 1993. Due to this high-profile case, auditors started to use the tests developed by Nigrini to check financial reports and identify fraud [8].

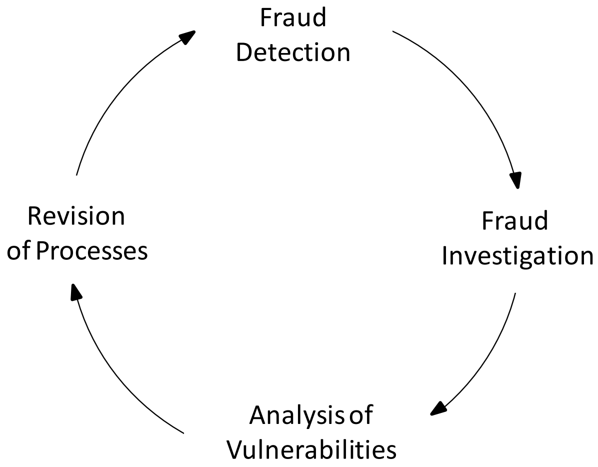
3. Fraud detection

3.1. Scheme

The scope of bank fraud can be considered highly adaptive, which means that fraudsters constantly come up with new ways how to get round the bank security. That is why the fight against bank fraud is compared to "confrontation of armor and projectiles." With such a constant arms race, bank anti-fraud requires an integrated, system approach.

One of the main principles of the system approach is the arrangement of anti-fraud processes in the form of a cyclic scheme according to the principles of the "scientific method" (Fig. 2). This scheme shows how anti-fraud technologies are developed and improved. One of the main ideas of the scheme is that fraud and counteraction to fraud (anti-fraud) constantly influence each other [9]. It means that, according to the scheme, the bank should regularly review its anti-fraud processes, modifying and improving them.

Figure 2. *Arrangement of anti-fraud processes in the bank according to the principles of the "scientific method"*



The development of predictors for fraud detection is also built according to a cyclic scheme and consists of four stages:

1) development of predictors;

2) verification of predictors on historical data;

3) introduction of predictors into the anti-fraud model;

4) fraud detection.

The cyclic scheme allows to refine and improve constantly the fraud detection process. It does not matter what stage to begin with. You can first develop a predictor, and then use it to detect fraudsters. Or you can, on the contrary, first identify fraud, and then, based on the findings, develop a predictor that will detect similar fraud cases. Thus, the development of predictors can be carried out either "top-down" ("predictor-case") or "bottom-up" ("case-predictor").

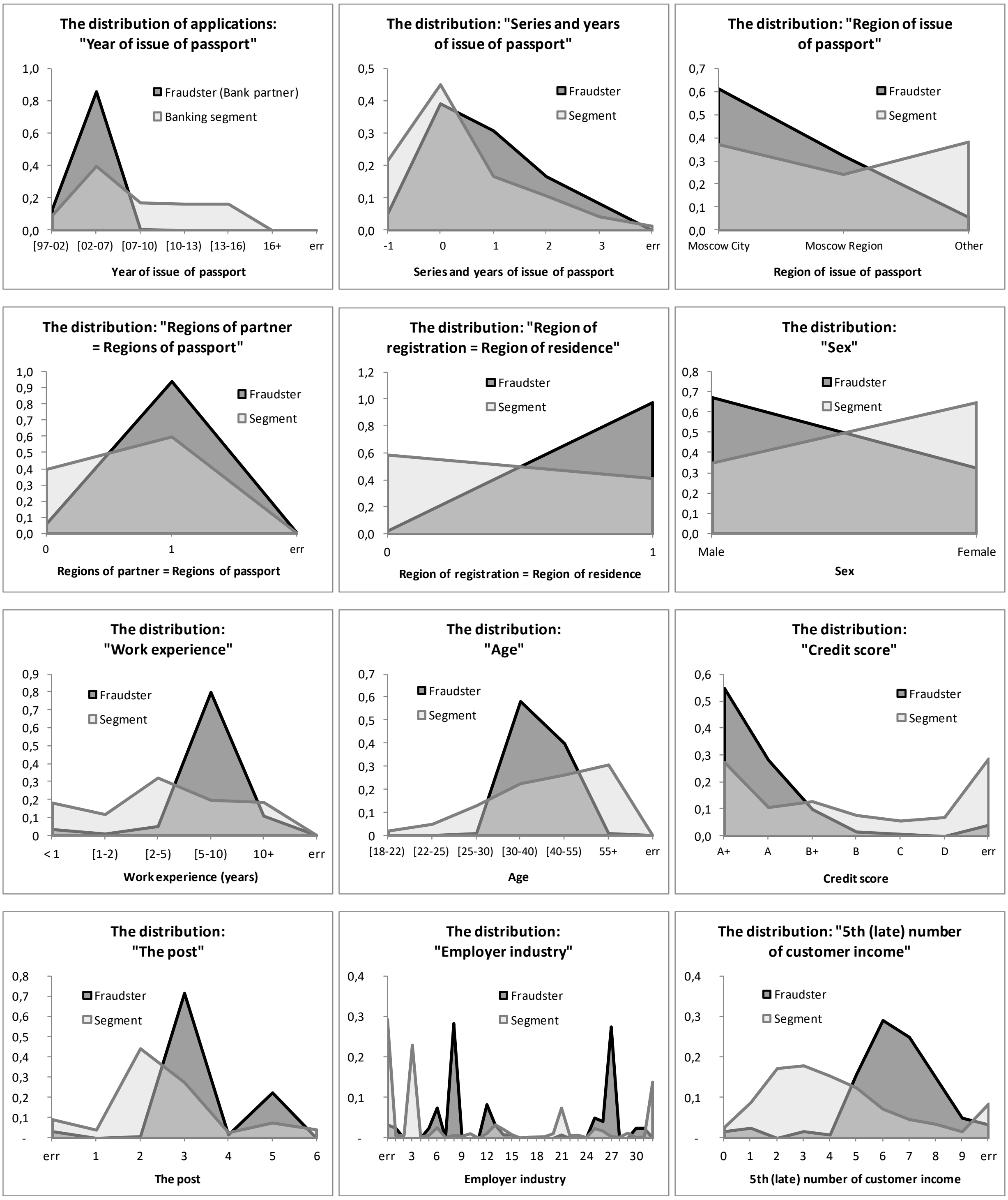
3.2. Fraud of POS-partner

Acting according to the "predictor-case" scheme, we tried to use Benford's law in the development of the predictor for internal fraud detection. We use the logic scheme that if an employee (or a partner of the bank) starts fraud applications, then he invents client data, specifically the client income. Therefore, the distribution of the first figures of salaries in applications, entered by a fraudster, should significantly differ from Benford's law. Conversely, the distribution of the first figures of salaries in applications entered by a conscientious employee should not be very different from Benford's law.

After testing the predictor on historical data, we saw that the differences are significant both at the request of fraudsters and at the requests of conscientious employees and partners of the bank. As a result, we did not manage to separate fraudsters from conscientious employees and partners, and we indefinitely abandoned attempts to use Benford's law. Later we learned that our failure was due to the fact that the first figures of salaries do not submit to the Benford's law. This is explained by the fact, that most customers have salaries out of a narrow range, and therefore the highest frequency is reached on the first digits of this particular range, and not necessarily on 'one'.

We returned to the idea of using Benford's principles after several years, when a fraud of a POS-partner was detected in the bank. He made out loans using the "passport scan-copy generator" (fraudulent software that can create a scan of a passport with the generated data). On these fake scans, the fraudster partner managed to make out 120 loans for 100 thousand dollars in 2 months. After analyzing fraudulent applications, we saw that almost all 120 passports were issued in a short period of time - from 2004 to 2007. This distribution was very different from the general banking distribution. Analyzing the other data, we saw that the distributions of many fields of fraudulent applications differ significantly from general banking distributions (Fig. 3).

Figure 3. *Comparison of fraudulent applications of the POS-partner with clients' data of the banking segment "region + type of goods"*



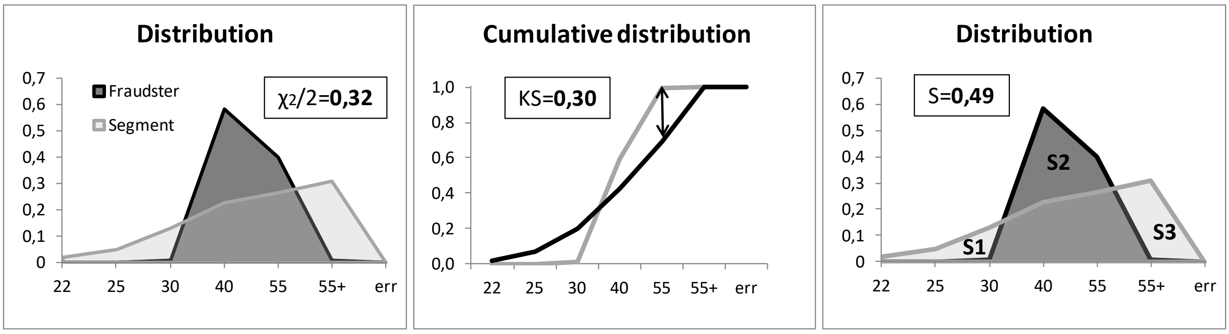
Based on the results of the "bottom-up" analysis ("case-predictor"), we understood that it is possible to build predictors by comparing the distributions of various client data from the applications of employees or partners to general banking distributions. At the same time, in order to reduce the impact of socio-demographic factors, it is necessary to set general banking distributions according to the segments with a similar client profile. For example, in the segment of cosmetic products and services among clients with predominance of women, in the segment of mobile phones with predominance of young people, etc. As part of our activities, we set general banking distributions according the product type (POS, CASH, CARD), the borrower region and the commodity group (for POS-loans).

We called the comparisons of distributions for the applications of an employee (or a partner) with general banking distributions "B-tests" - according to the first letter of the name of Frank Benford, the popularizer of the “first-digit law”.

4. Metrics

In order to digitize the results, it is necessary to select the metric, which shows the differences between the two distributions. For our calculations, we studied and compared three most popular metrics: chi-square statistics, Kolmogorov-Smirnov statistics and square statistics (S-statistics).

Figure 4. *An example of metric calculation of Chi-square, Kolmogorov-Smirnov and S-statistics.*



4.1. Chi-Square Statistics

The chi-squared statistics for two distributions $\{a_i\}$ and $\{b_i\}$ are calculated by the formula (1) and is equal to the sum of the normalized quadratic value differences of the two distributions [10]:

$$\chi^2 = \sum_{i=1}^{n} \frac{(a_i - b_i)^2}{a_i + b_i} \qquad (1)$$

Values of the chi-square statistics lie in the range [0; 2], therefore the normalized statistics $\chi^2 / 2$ are taken for convenience. Thus, we can assume that for $\chi^2 / 2 = 0$ the distribution of $\{a_i\}$ and $\{b_i\}$ completely identical (the difference is 0%), and for the distribution they do not intersect (100% difference). Chi-square statistics works well for all types of distributions, however, when interpreting the results, it should be taken into account that nonlinear statistics and values $\chi^2/2$ will be more often concentrated near 'zero' and less often near 'one'.

4.2. Kolmogorov-Smirnov Statistics

The Kolmogorov-Smirnov statistics is calculated by the formula (2) and is equal to the maximum difference in values of cumulative distributions [11]:

$$KS = \max_i |F(a_i) - F(b_i)| \quad (2)$$

The values of KS-statistics lie in the range [0; 1] and vary linearly. Thus, we can assume that for KS=0 the distributions are completely identical (the difference is 0%), at KS=0.5 the distributions intersect half (50% difference), and for KS=1 the distributions do not intersect (100% difference). However, this interpretation is not always correct, because KS-statistics are sensitive to outliers, or so-called "non-standard" distributions, and fraudulent distributions are often related to them. This feature of KS-statistics has the following mathematical interpretation: if the "non-standard" (fraudulent) distribution has three or more disjoint areas with a "standard" (general banking) distribution, then the cumulative distributions of these samples will form several "petals"[1]. And since KS-statistics is calculated as the maximum difference between two cumulative distributions, the more "petals" will form two cumulative curves, the smaller will be the value of KS-statistics, while the distributions themselves will differ significantly from each other (Figure 4 with an example of two petals). Thus, it can be concluded that the values of KS-statistics will often be underestimated for fraudulent distributions, which may lead to incorrect interpretation and incorrect results of predictor configuring.

4.3. Square Statistics (S-statistics)

The square statistics is calculated by the formula (3) and is equal to the normalized sum of absolute values of the difference between the values of two distributions:

$$S = \sum_{i=1}^{n} \frac{|a_i - b_i|}{2} \quad (3)$$

S-statistics is linear and works well on different types of distributions (we'll show the advantages of linearity when analyzing the calibration of triggers). The values of S-statistics lie in the range [0; 1], that is, we can assume that even at S=0 the distributions coincide (the difference is 0%), at S=0.5 the distributions intersect half (the difference is 50%), for S=1 the distributions do not intersect (100% difference). It is also possible to prove mathematically that S-statistics is a generalization of KS-statistics in the sense that S is always greater than KS or equal to KS (Table

[1] This feature has strict mathematical proof, which is beyond the scope of this article

1)[2]. If two cumulative distributions form one "petal", then S = KS. Thus, it can be concluded that the S-statistics are not understated and have a simple geometric interpretation: the value of S-statistics is the sum of squares under disjoint distribution areas (normalized to 2, since the sum of squares under the graphs of two normalized distributions is 2).

Based on the above advantages and disadvantages, we chose S-statistics as a metric for calibration of B-tests. Table 1 shows a list of B-tests and the results of calculating three metrics built on data of fraudulent POS-partner applications.

Table 1. *List of B-tests and values of metrics built on data of fraudulent POS-partner applications.*

#	B-tests	χ2/2	KS	S
1.	Branch of client's employer	0,70	0,48	**0,80**
2.	The fifth digit of the income (from the end)	0,43	0,55	**0,61**
3.	Work experience of a client	0,39	0,53	**0,60**
4.	Client's position	0,40	0,54	**0,59**
5.	Match of addresses (actual/registration)	0,37	0,56	**0,56**
6.	Year according to passport series	0,31	0,50	**0,50**
7.	Year passport issue	0,33	0,50	**0,50**
8.	Client's age	0,32	0,30	**0,49**
9.	Client's rating	0,25	0,24	**0,45**
10.	Region according to client's passport	0,16	0,34	**0,34**
11.	Region according to passport series of a client	0,16	0,33	**0,33**
12.	Client's gender	0,11	0,32	**0,32**
13.	Year according to passport series minus Year passport is	0,09	0,23	**0,24**
14.	The fourth digit of the income (from the end)	0,07	0,20	**0,22**

Among the alternative metrics, we can distinguish Anderson-Darling test [12], Z-test [13], MAD-test (mean absolute deviation test) [14] and others. It should be noted that in the described technique, the considered metrics are not used to test statistical hypotheses (as they are used in mathematical statistics) but are intended to determine the "closeness" value of two distributions. The threshold of this value is determined empirically by means of machine learning methods.

5. Target Variable

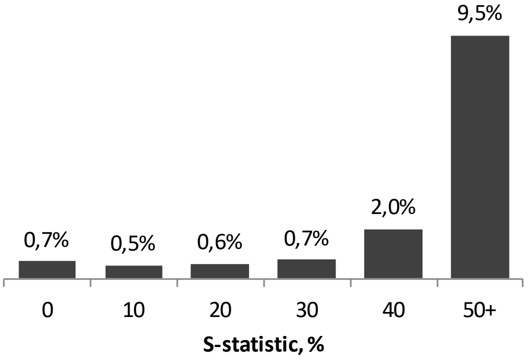
Before starting to calibrate the B-tests, the target variable should be selected, which evaluates the effectiveness of the predictors. In problems of fraud-analytics two types of target variables are usually used:

[2] The mathematical proof of this feature is based on changing the signs of the differences in the values of the two distributions.

1) based on the results of investigations - when flags are displayed on the analyzed object (application/employee/partner): "no fraud," "soft fraud" and "hard fraud";
2) based on financial results - when the financials (default rate, potential loss, etc.) are calculated according to the analyzed object.

As a target variable of the first type (based on the results of investigations) percentage of blocking of employees or partners can be taken. Since the blocking is the final positive result of the investigation (when fraud or critical violations are approved), this indicator will reflect the fraud level in the segment. For example, for the B-test "Year of issue of passport" there is a strong correlation of blocking level from the value of S-statistics: as the value of S-statistics grows, percentage of blocking of employees or partners increases, i.e. the probability of internal fraud increases (Fig. 5). That means that the predictor "Year of issue of passport" is highly effective.

Figure 5. *Predictor "Year of issue of passport", calculated using the target variable "Percentage of blocking".*

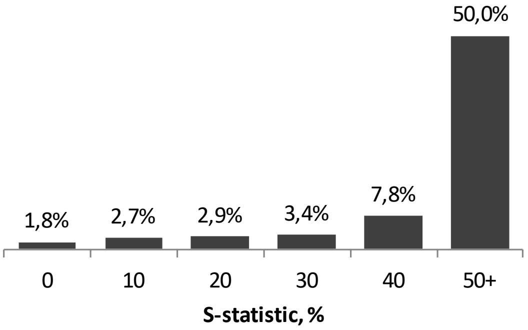


Sampling options:

Analysis horizon	1 year
Analysis period (for one partner)	30 days
Lock period	90 days
Number of groups "Year of issue of passport"	7
Minimum number of applications	15
Number of partners	91 032
Number of locking partners	552

As the financial target variable, the delinquency more than 30 days on the 3rd month of bucket (30+mob3) can be taken. This default is quite mature and does not strongly correlate with the level of the first payments on loans paid by fraudsters (which has already become quite common). It can be seen that for the B-test "Year of issue of passport" there is a strong correlation of the default on the values of S-statistics: the higher the value of S-statistics, the higher the indicator 30+mob3, that is, the higher the probability of fraud (Fig. 6).

Figure 6. *Predictor "Year of issue of passport", calculated using the target variable "Default level 30+mob3".*



Sampling options:

Analysis horizon	1 year
Analysis period (for one partner)	30 days
Number of groups "Year of issue of passport"	7
Minimum number of applications	15
Number of contracts	857 933
Number of default contracts	19 418

Different types of target variables have their pros and cons, and the choice of the target variable depends on the specific task. For example, the advantage of the target variable "Level of blockings" is that it takes into account only confirmed facts of fraud and excludes the social default. On the other hand, the results of investigations are subjective expert conclusions and do not reflect real financial losses (not always the issuance of a fraudulent employee or partner is 100% loss-making, and often even vice versa - the loss is much less than 100%).

The financial target variable "default of 30+mob3" reflects real losses for the bank, which makes this variable more objective from the point of view of influence on the bank's financial indicators. This target variable is well suited for the development of predictors that detect fraudulent applications online, with the next-following setting of refusals for these applications in the bank scoring system. However, for predictors detecting internal fraud, such a target variable has a number of shortcomings. In addition to the non-fraudulent component (social default), the 30+mob3 default indicator (or any other parameter of default) shows the already occurred loss for the bank. Using such a target variable leads to the fact that when tuning a predictor, the parameters are selected according to the level of statistical significance (the number of applications), and not by the immediacy degree of detecting internal fraud. That is, the triggering of predictors occurs when the real losses for the bank are already large. This negates one of the main advantages of predictive fraud analytics - the identification of internal fraud at the early stages, when the losses for the bank are still insignificant.

The conclusion is that to adjust the predictors that detect internal fraud, we need a more complex target variable - one that shows the potential loss after blocking a suspicious employee or partner (i.e., after the predictor has triggered). This objective function can be constructed through the indicator parameter of default and the breakeven point of this indicator ("zero-target") by the following formula (4)[3]:

$$TV_{(T_l - T_a)} = S * \frac{D - D_0}{1 - D_0} \tag{4}$$

where $TV_{(T_l - T_a)}$ - target variable, showing the losses in a period $(T_l - T_a)$ after the predictor has been triggered;

T_l - the period of losses in question (under study);

T_a - the analyzed period on which the predictor is triggered;

$T_l - T_a$ - the period after the predictor is triggered;

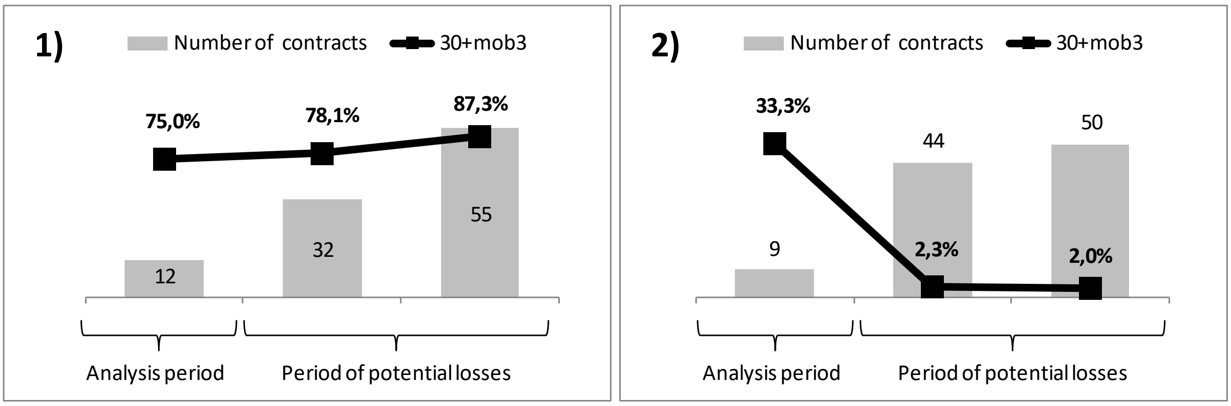
D - the default level in the issuance of an employee (or partner) for the period $(T_l - T_a)$;

D_0 - "zero-target" of default in the period $(T_l - T_a)$;

S - the employee's (or partner's) credit portfolio for the period $(T_l - T_a)$.

With the help of the target variable of potential loss, it's possible to tune up fast predictors, that allow to detect internal fraud at the early stages (Fig. 7.1). In addition, predictors tuned with such a target variable will have a low percentage of false positives, that is, they will rarely will be triggered on employees (or partners) who have a high default level in a short analyzed period, and low default level in the next-following period. Such false positives are distinguishing if the predictor was tuned with the target variable "Default level 30+mob3" (Fig. 7.2).

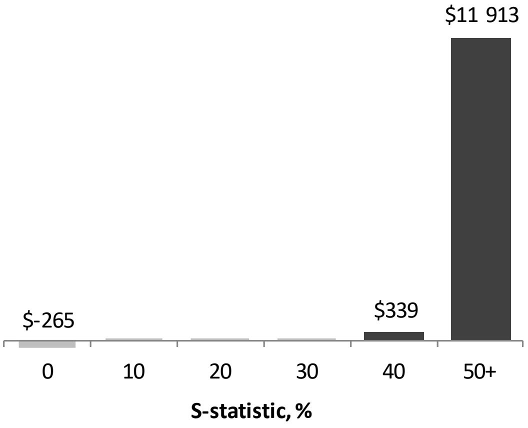
Figure 7. *An example of correct predictor tuning with the target variable "Potential loss" (1) and incorrect tuning using the target variable "Default level 30+mob3"(2).*



[3] The derivation of this formula is based on the assumption that the level of collection by default contracts is 0%, i.e. the entire amount of the default contracts is written off at a loss. This is permissible if default contracts are fraudulent and refunds for such contracts are impossible

If we select the analyzed period of 30 days, and the loss period of 90 days and calculate the value of the target variable of potential loss for the developed B-tests, we can see how much bank can save (or lose) averagely from each blocked employee (or partner) in a certain group of "S-statistics" for 60 days after the predictor is triggered. So, for the B-test "Year of issue of passport", each partner who got into the group with S-statistics “≥ 50%”, brought the bank averagely $11 913 loss for 60 days after the triggering of this B-test (Fig. 8). In total, the group "S≥50%" got 6 partners, which means that the total loss for this group was $71 477 for 60 days after the triggering of the predictor (on the horizon of 1 year).

Figure 8. *Predictor "Year of issue of passport", calculated using the target variable "Average potential loss per partner".*



Sampling options:

Analysis horizon	1 year
Analysis period (for one partner)	30 days
Period of loss	90 days
Period of potential losses	60 days
Number of groups "Year of issue of passport"	7
Minimum number of applications	20
Number of partners in S-group "40-50"	113
Number of partners in S-group "50+"	6

It should be noted that when tuning predictors using the target variable described by formula (4), we should consider the entire potential loss for the period, and not the average loss for any period (day, month, year). This is due to the fact that different types of fraud can have a different period of time, which in turn depends on many factors, including internal anti-fraud processes of the bank (in some cases, fraud is disclosed in a month, in others in six months). Thus, the most valuable will be the predictor, which maximizes the entire potential loss: it is more profitable to catch in time an employee who can steal $50 000 for 5 months than an employee who stole $20 000 for 1 month and quitted after that (although the average monthly loss for the second one is higher).

6. Tuning Scheme for Predictors

After choosing the metrics to compare two distributions and the target variable for predictors, you can start tuning B-tests on the historical sample. The tuning consists of three steps (Fig. 9):

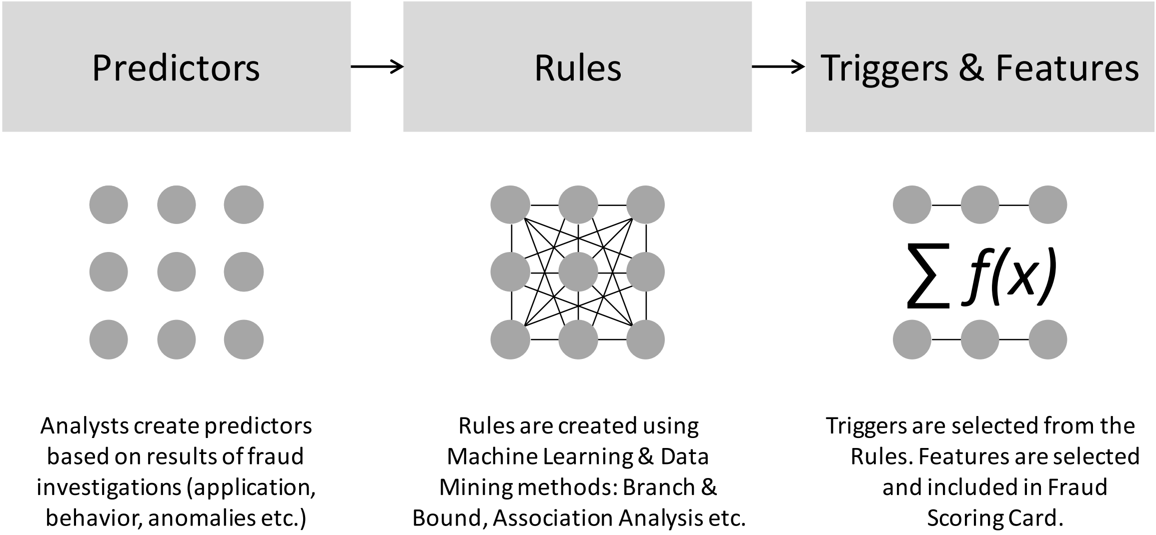
1. In the first step, predictors are developed - basic algorithms for fraud detection.
2. In the second step, the rules are built by enumerating parameter values that affect the predictive ability of these predictors.
3. In the third step, effective triggers and features are selected from the rules to develop a scorecard.

Triggers are rules with a high predictive ability. They allow us to identify a group of objects (customers, employees, partners) with a high fraud concentration. Triggers work as independent algorithms.

Features are rules with low predictive power, but with high Information Value. Using features, effective scorecards can be built.

Taken together, the triggers and the fraud-scorecards form a Fraud Scoring Model, which allows to detect fraud in an automatic mode very quickly.

Figure 9. *Scheme of the development of triggers and features for Fraud Scoring Model*


Predictors
Rules
Triggers & Features
∑ f(x)
Analysts create predictors based on results of fraud investigations (application, behavior, anomalies etc.)
Rules are created using Machine Learning & Data Mining methods: Branch & Bound, Association Analysis etc.
Triggers are selected from the Rules. Features are selected and included in Fraud Scoring Card.

According to the described scheme, it is possible to develop triggers and scorecards both for internal fraud detection among employees or partners, and for identifying fraudulent applications in order to use these triggers and scoring online, that is, at the decision-making stage. The main difference between the development of predictors for detecting internal fraud and predictors for detecting fraudulent applications is that in the first case, employees or partners of the bank are considered as objects, and in the second - loan applications or borrowers. However, most predictors that can be effectively tuned to detect internal fraud can also be effectively tuned to detect fraudulent applications (and vice versa). In turn, B-tests are an exception, since the calculation of metrics (S-statistics) requires the data of two distributions: the general banking distribution (by segment) and the distribution of the investigated object (employee or partner). In other words, for the B-tests discussed above, it is not possible to build distributions for a single application or a customer. Therefore, these B-tests can only be used to detect internal fraud. However, it should be noted that for certain types of data, it is possible to develop B-tests and to identify client fraud. For example, analyzing bank card transactions using of B-tests can detect anomalous behaviour of card clients.

Next, we will consider the methods for trigger developing on the example of the B-test "Year of issue of passport." These methods are common to all predictors, so the tuning results of the rest of B-tests will be omitted.

7. Parameters

The effectiveness of fraud predictors is determined by the target variable and depends on many factors. On the one hand, the fraud predicator should be tuned so that fraud could be detected at an early stage. On the other hand, the percentage of false positives should be low, since resources of analysts and security officers are being wasted on the investigation (and not every employee will take responsibly into account the investigation, in which fraud is rarely confirmed). Maximization of the target variable of potential loss allows to select the optimal ratio "efficiency/ percentage of false positives, resulting in effective triggers. To maximize the target variable, special tuning parameters are used, which determine the effectiveness of fraud detection and the percentage of false positives.

The first important parameter affecting the efficiency of detecting internal fraud is the analyzed period, during which the employee or partner issued loan applications. The shorter the analyzed period, when fraud is detected, the greater the potential loss that can be avoided. On the other hand, the percentage of false positives on a short analyzed period may be high. Therefore, the analyzed period must be included in the tunable parameter list.

Analyzing various fraudulent schemes, you can notice that depending on various factors (segment, product, business process, etc.) for the same period of time, fraudsters issue a different number of applications. So we can conclude that the second important parameter affecting the effectiveness of internal fraud detection is the minimum number of analyzed applications per an employee or partner: the fewer applications are used in the predictor, the faster the fraud will be revealed. On the other hand, with a small number of applications, the proportion of false positives can be high. Therefore, this parameter must also be included in the tunable parameter list.

The third parameter affects the percentage of false positives and is related to the specifics of calculating B tests. In order to calculate the metrics for the B-test (in our case, S-statistics), it is necessary to construct two frequency distributions: bank-segment sample and employees/partners sample. In the first sample of data, as a rule, it is enough to build a stable form of distribution. In the second sample (for the employee or partner) can occur a small number of applications, that is why the distribution will be unstable in the sense that for law-abiding employees and partners, the values of S-statistics may be overestimated, which means that the B-test will show a high level of false positives. The stability of the distribution form can be improved in two ways:

1) to increase the average number of applications in each distribution category by increasing the minimum number of analyzed applications in the entire sample (the number of categories in the distribution does not change);
2) to reduce the number of categories in the distribution, thereby increasing the average number of applications in each category (the minimum number of analyzed applications remains unchanged).

The first method reduces the proportion of false positives, but at the same time reduces the responsiveness of the B-test. The second method does not affect the efficiency, but it reduces the proportion of false positives. That means that "Number of categories in distribution" is the key parameter that affects the effectiveness of the B-test, and therefore is included in the tunable parameter list.

As already mentioned, the value of the S-statistics ranges from 0 to 1, where the value 0 is reached for completely coinciding distributions, and the value 1 for non-crossing distributions. The question arises: which S-statistics threshold value should be chosen to separate good employees from scammers? The S-statistics threshold value depends on various factors. In some cases, a predictor with the value $S \geq 0.2$ will distinguish a high-risk group of employees or partners, in other cases profitable employees or partners can take the values $S = 0.6$. So, for example, if we consider the correlation with the parameter "Number of analyzed applications", then the smaller the number of such applications for an employee or partner, the higher the S-statistic threshold value should be, so that the percentage of false positives is low. That is, through the value of S-statistics, it is possible to regulate the indicator of the effectiveness of internal fraud detection. Another peculiarity of S-statistics threshold value is that it depends on the bank-segment sample: the less accurately the segment sample describes the client thread with which fraudsters work, the higher the S-statistic threshold value, so that the percentage of false positives is low. Thus, the S-statistics threshold value also affects the efficiency of the B-test, and the percentage of false positives. Therefore, the S-statistics should also be included in tunable parameter list.

The target variable of "potential loss" is calculated in the period $(T_l - T_a)$, where T_a is the analysed period, and T_l is the period of the fraudulent losses under consideration. Different fraudulent schemes have different duration. That means that the period, generally speaking, should not be fixed. For example, there may be cases when a partner has several employees, one of them turned out to be a fraud. For some time this employee issued fraudulent loans, and after his dismissal the partner continued to work and became profitable again. Then, if after the dismissal of the fraudster, the newly issued credits get into the period T_l, then the value of the target variable of "potential loss" will be understated. That means that the target variable depends on the period T_l, i.e. this period should also be included in tunable parameter list.

Table 2. *List of key parameters for tuning B-tests.*

Analyzed period	Minimum number of applications per an employee/partner	Number of categories in distributions	S-statistics threshold value	Period of analyzed losses
7 days	10	3	>= 5%	60 days
14 days	15	6	>= 10%	90 days
30 days	20	10	>= 20%	
60 days	30		>= 30%	
	50		>= 40%	
			>= 50%	
			>= 60%	
			>= 70%	
			>= 80%	
			>= 90%	

The list of key parameters is presented in Table 2, where the range values of these parameters are also indicated for B-tests tuning. This list is not exhaustive and can be expanded to dozens of parameters, such as:

- Credit product: POS-loans, cash loans, credit cards, car loans, etc. (we considered just POS-loans);
- Sales channel: bank offices, POS-partners, web-channels, courier delivery, etc. (we considered just POS-channel);
- Region of the country;
- Type of settlement (city/village);
- Social and demographic parameters of the employee: age, duration of work, marital status, etc.;
- POS-partner parameters: type of goods, the period of presence on the market, the number of outlets, the size of retail spaces, etc.;
- Social and demographic parameters of the client: age, marital status, education, etc.;
- Commodity group: mobile phones, computers, household appliances, furniture, building materials, cosmetics, tourism, services, etc.;
- Range of average loan amount;

and etc.

If the above parameters affect the predictor's efficiency, then the question arises: why did not we include them in tunable parameter list? This question touches on another important step of tuning B-test - choosing a tuning method, which in fact needs to be determined before the tunable parameter list is formed.

B-test tuning can be carried out by various methods, each of them will have different computational complexity and time complexity. For the purposes of this article, we used the Brute Force Method, the essence of which is to search through all possible combinations of the values of the selected parameters (on which the target variable is calculated). In our case (see Table 2), the number of all possible combinations is: 4 × 5 × 3 × 10 × 2 = 1200 variants. Calculation of the target variable for all 1200 variants is acceptable from the point of view of computational complexity. If we take the tuning step equal to one (for 1 day, 1 application, 1%, etc.) for the selected parameters, then the number of all possible combinations will be about 70 million variants. If we add other parameters to the list, then the number of combinations can increase to 10^9–10^{15}. For such tasks Brute Force Method is not optimal. Among the methods that allow to solve such problems, it is possible to use the Branch and Bound Algorithm [15], Association Analysis [16], Backtracking [17], etc. These methods allow to find effective groups of combinations and to cut ineffective, thereby reducing the amount and time of calculations.

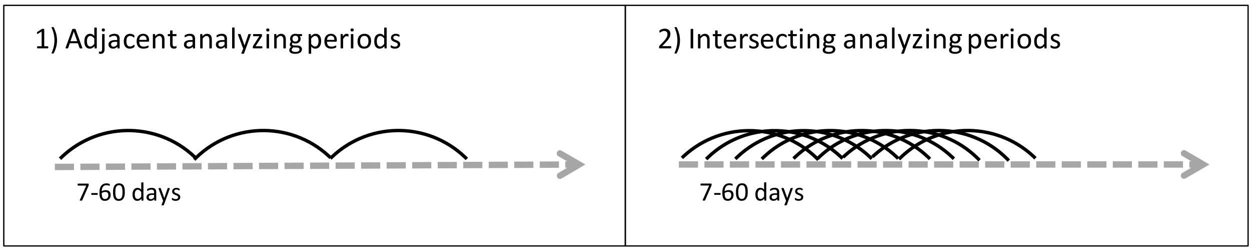
8. Triggers

8.1. Tuning

The final stage in the of B-tests development is choosing the effective triggers (tuning) and checking their stability. For tuning the predictors, we should choose the data analysis horizon (training sample). Since internal fraud refers to "black swan" event (that comes as a surprise, has a major effect) [18], then for the problems of fraud analytics the minimum data analysis horizon is usually about 1-2 years. For tuning the B-test "Year of issue of passport" we took a horizon equal to 1 year. To reduce the number of calculations, we used the method of covering the horizon with adjacent analyzing periods (Fig. 10.1). For more accurate tuning of predictors, the method of covering the horizon with intersecting periods with the step of 1 day is used (Fig. 10.2)[4] It is worth noting that using an adjacent or intersecting cover, applications for one employee (or partner) can fall into several different horizon periods. Therefore, when tuning a B-test, one employee (or partner) may have several different target variable values, each of them must be taken into account when tuning.

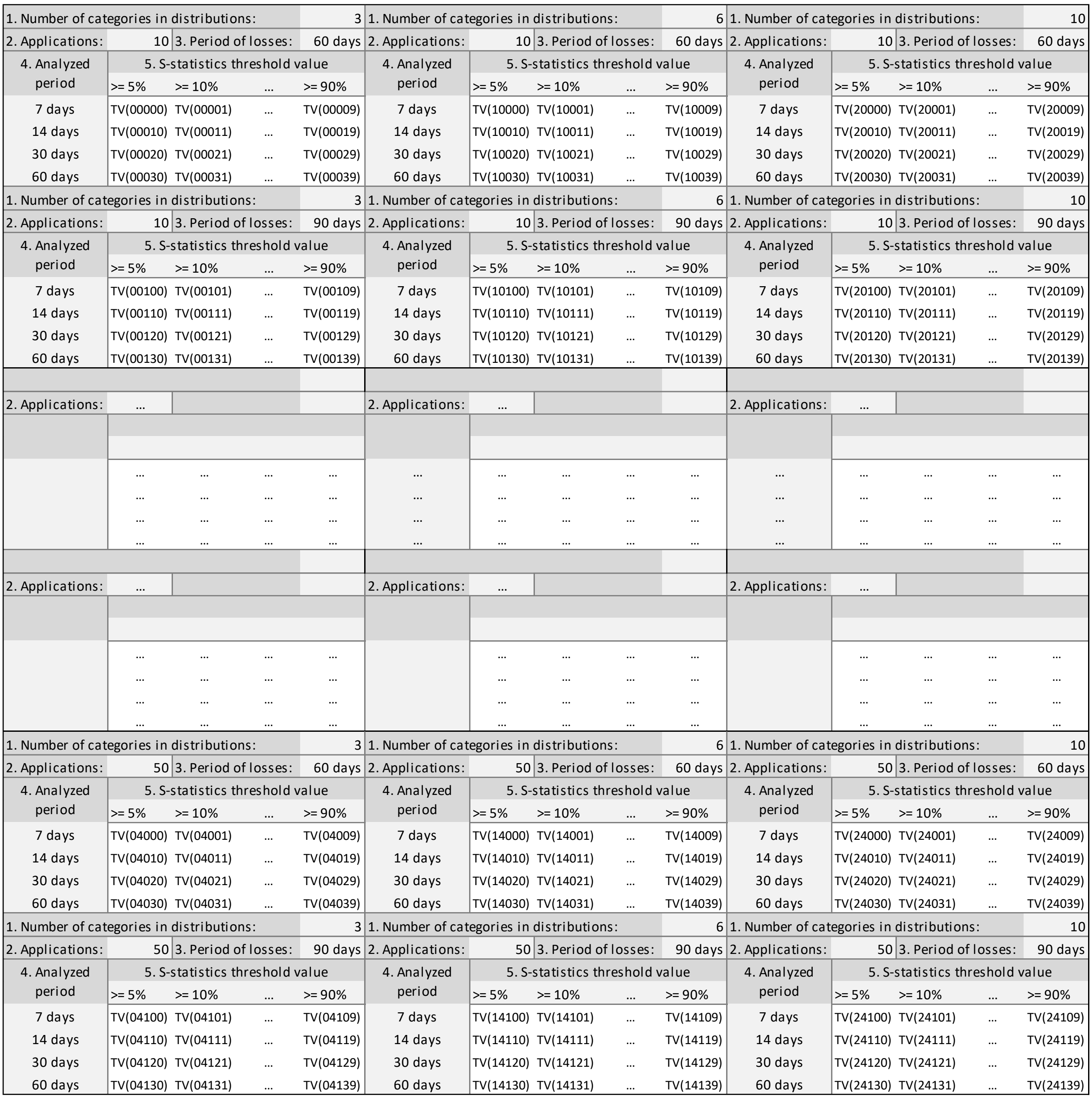
[4] Covering with intersecting horizons significantly increases the number of calculations. For example, for selected periods [7 days; 14 days; 30 days; 60 days], the number of adjacent periods on the horizon "1 year" will be 96, and the number of intersecting periods with the step of 1 day is 1349.

Figure 10. *Options for covering the data analysis horizon.*


1) Adjacent analyzing periods
7-60 days
2) Intersecting analyzing periods
7-60 days

After the parameters list is formed and the values ranges are chosen, the triggers tuning is carried out by Brute Force method. The tuning process can be visualized by building the *N*-dimensional matrix of the target values, where *N* is the number of tuning parameters (Fig. 11). Each cell of the given matrix corresponds to a set of values of the tuning parameters (explanatory variables), and the cell themselves contain the values of "potential loss" (target variable). Target values are calculated as the ratio of the sum of all losses for employees/partners included in the matrix cell with the specified parameters to the number of employees/partners that generated this loss. As already noted, losses for an employee or partner can be accounted for in one point several times due to the fact that the data analysis horizon is covered by several adjacent periods (in our case, 96 periods). That means that the total number of calculations of the target values for such a matrix is 1200 × 96 = 115 200 iterations.

Figure 11. *Brute Force matrix for tuning of the B-test "Year of issue of passport" (in parentheses the index of the target variable TV is indicated).*



1. Number of categories in distributions:				3	1. Number of categories in distributions:				6	1. Number of categories in distributions:				10
2. Applications:	10	3. Period of losses:		60 days	2. Applications:	10	3. Period of losses:		60 days	2. Applications:	10	3. Period of losses:		60 days
4. Analyzed period	5. S-statistics threshold value				4. Analyzed period	5. S-statistics threshold value				4. Analyzed period	5. S-statistics threshold value			
	>= 5%	>= 10%	…	>= 90%		>= 5%	>= 10%	…	>= 90%		>= 5%	>= 10%	…	>= 90%
7 days	TV(00000)	TV(00001)	…	TV(00009)	7 days	TV(10000)	TV(10001)	…	TV(10009)	7 days	TV(20000)	TV(20001)	…	TV(20009)
14 days	TV(00010)	TV(00011)	…	TV(00019)	14 days	TV(10010)	TV(10011)	…	TV(10019)	14 days	TV(20010)	TV(20011)	…	TV(20019)
30 days	TV(00020)	TV(00021)	…	TV(00029)	30 days	TV(10020)	TV(10021)	…	TV(10029)	30 days	TV(20020)	TV(20021)	…	TV(20029)
60 days	TV(00030)	TV(00031)	…	TV(00039)	60 days	TV(10030)	TV(10031)	…	TV(10039)	60 days	TV(20030)	TV(20031)	…	TV(20039)
1. Number of categories in distributions:				3	1. Number of categories in distributions:				6	1. Number of categories in distributions:				10
2. Applications:	10	3. Period of losses:		90 days	2. Applications:	10	3. Period of losses:		90 days	2. Applications:	10	3. Period of losses:		90 days
4. Analyzed period	5. S-statistics threshold value				4. Analyzed period	5. S-statistics threshold value				4. Analyzed period	5. S-statistics threshold value			
	>= 5%	>= 10%	…	>= 90%		>= 5%	>= 10%	…	>= 90%		>= 5%	>= 10%	…	>= 90%
7 days	TV(00100)	TV(00101)	…	TV(00109)	7 days	TV(10100)	TV(10101)	…	TV(10109)	7 days	TV(20100)	TV(20101)	…	TV(20109)
14 days	TV(00110)	TV(00111)	…	TV(00119)	14 days	TV(10110)	TV(10111)	…	TV(10119)	14 days	TV(20110)	TV(20111)	…	TV(20119)
30 days	TV(00120)	TV(00121)	…	TV(00129)	30 days	TV(10120)	TV(10121)	…	TV(10129)	30 days	TV(20120)	TV(20121)	…	TV(20129)
60 days	TV(00130)	TV(00131)	…	TV(00139)	60 days	TV(10130)	TV(10131)	…	TV(10139)	60 days	TV(20130)	TV(20131)	…	TV(20139)
2. Applications:	…				2. Applications:	…				2. Applications:	…			
	…	…	…	…	…	…	…	…	…	…	…	…	…	…
	…	…	…	…	…	…	…	…	…	…	…	…	…	…
	…	…	…	…	…	…	…	…	…	…	…	…	…	…
	…	…	…	…	…	…	…	…	…	…	…	…	…	…
2. Applications:	…				2. Applications:	…				2. Applications:	…			
	…	…	…	…		…	…	…	…		…	…	…	…
	…	…	…	…		…	…	…	…		…	…	…	…
	…	…	…	…		…	…	…	…		…	…	…	…
	…	…	…	…		…	…	…	…		…	…	…	…
1. Number of categories in distributions:				3	1. Number of categories in distributions:				6	1. Number of categories in distributions:				10
2. Applications:	50	3. Period of losses:		60 days	2. Applications:	50	3. Period of losses:		60 days	2. Applications:	50	3. Period of losses:		60 days
4. Analyzed period	5. S-statistics threshold value				4. Analyzed period	5. S-statistics threshold value				4. Analyzed period	5. S-statistics threshold value			
	>= 5%	>= 10%	…	>= 90%		>= 5%	>= 10%	…	>= 90%		>= 5%	>= 10%	…	>= 90%
7 days	TV(04000)	TV(04001)	…	TV(04009)	7 days	TV(14000)	TV(14001)	…	TV(14009)	7 days	TV(24000)	TV(24001)	…	TV(24009)
14 days	TV(04010)	TV(04011)	…	TV(04019)	14 days	TV(14010)	TV(14011)	…	TV(14019)	14 days	TV(24010)	TV(24011)	…	TV(24019)
30 days	TV(04020)	TV(04021)	…	TV(04029)	30 days	TV(14020)	TV(14021)	…	TV(14029)	30 days	TV(24020)	TV(24021)	…	TV(24029)
60 days	TV(04030)	TV(04031)	…	TV(04039)	60 days	TV(14030)	TV(14031)	…	TV(14039)	60 days	TV(24030)	TV(24031)	…	TV(24039)
1. Number of categories in distributions:				3	1. Number of categories in distributions:				6	1. Number of categories in distributions:				10
2. Applications:	50	3. Period of losses:		90 days	2. Applications:	50	3. Period of losses:		90 days	2. Applications:	50	3. Period of losses:		90 days
4. Analyzed period	5. S-statistics threshold value				4. Analyzed period	5. S-statistics threshold value				4. Analyzed period	5. S-statistics threshold value			
	>= 5%	>= 10%	…	>= 90%		>= 5%	>= 10%	…	>= 90%		>= 5%	>= 10%	…	>= 90%
7 days	TV(04100)	TV(04101)	…	TV(04109)	7 days	TV(14100)	TV(14101)	…	TV(14109)	7 days	TV(24100)	TV(24101)	…	TV(24109)
14 days	TV(04110)	TV(04111)	…	TV(04119)	14 days	TV(14110)	TV(14111)	…	TV(14119)	14 days	TV(24110)	TV(24111)	…	TV(24119)
30 days	TV(04120)	TV(04121)	…	TV(04129)	30 days	TV(14120)	TV(14121)	…	TV(14129)	30 days	TV(24120)	TV(24121)	…	TV(24129)
60 days	TV(04130)	TV(04131)	…	TV(04139)	60 days	TV(14130)	TV(14131)	…	TV(14139)	60 days	TV(24130)	TV(24131)	…	TV(24139)

The choosing of effective triggers is iteratively performed using the forward selection method [19]:

1) At the first iteration, the target variable is calculated for all combinations of parameters in the sample and the matrix cell with the highest value of the target variable is selected. The parameter values in this matrix cell determine the first effective trigger;
2) At the second iteration, all applications for employees (or partners) lying in the first effective trigger are removed from the general sample. The Target variable is calculated on the updated sample and again the matrix cell with the maximum loss is chosen. The next trigger is selected and trigger applications are excluded from the sample.

Iterations continue until the rules with the Target variable below the given threshold are left. In our example, a threshold value of 0 was set for the target variable. The selection algorithm stopped at the 19th iteration, when all matrix points assumed "no-loss values". As a result, 18 triggers were selected for the B-test "Year of issue of passport" (Table 3).

Table 3. *Triggers selected using the forward selection method, for the predictor "Year of issue of passport."*

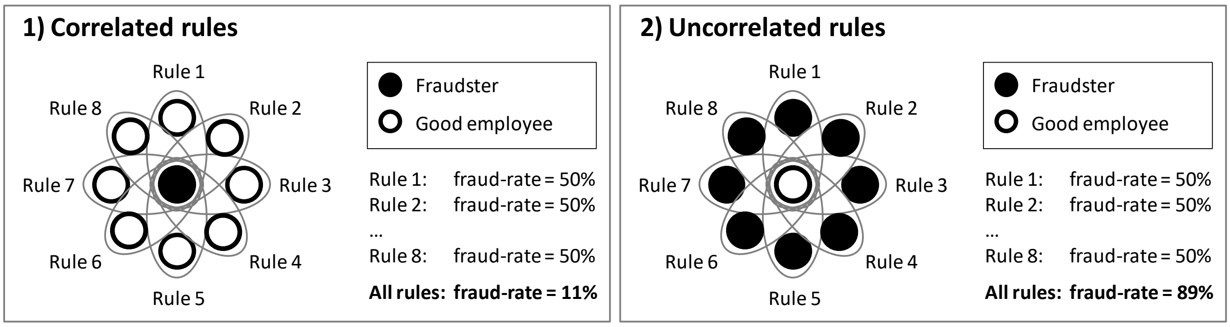
	Parameters					Average potential	Number of	Hit-rate, %
#	Analyzed period	Number of applications	Number of categories	S-statistics threshold	Period of losses	loss per one employee/partner	employees /partners	employees /partners
Trigger 1	14	10	10	>=90%	90	$65 694	1	0,0007%
Trigger 2	30	50	3	>=30%	90	$5 284	13	0,0474%
Trigger 3	7	15	10	>=70%	60	$1 797	1	0,0011%
Trigger 4	7	20	3	>=50%	60	$1 020	3	0,0051%
Trigger 5	14	20	3	>=40%	90	$964	30	0,0401%
Trigger 6	30	15	6	>=50%	90	$541	19	0,0243%
Trigger 7	14	10	3	>=50%	90	$509	63	0,0444%
Trigger 8	7	15	6	>=60%	60	$258	8	0,0088%
Trigger 9	7	10	10	>=70%	90	$156	35	0,0222%
Trigger 10	14	20	10	>=50%	90	$159	67	0,0907%
Trigger 11	30	20	6	>=40%	90	$113	79	0,1203%
Trigger 12	7	10	3	>=60%	90	$91	13	0,0084%
Trigger 13	14	15	3	>=40%	60	$70	70	0,0716%
Trigger 14	14	20	6	>=40%	90	$65	127	0,1751%
Trigger 15	14	50	10	>=40%	90	$1 005	3	0,0161%
Trigger 16	30	50	6	>=30%	90	$272	20	0,0763%
Trigger 17	7	20	3	>=40%	90	$140	26	0,0456%
Trigger 18	14	30	10	>=40%	90	$42	74	0,1815%

Forward selection method allows to reduce multicollinearity of triggers, that is, with the help of this method effective low correlated triggers are selected.

Multicollinearity triggers has a strong influence on the whole efficiency of the model: the larger the amount of highly correlated triggers, the higher the percentage of false positives. This property can be illustrated by the following synthetic example: let's say that training sample got 9 employees, 1 of which turned out to be a fraudster. Suppose that when tuning the rules, 8 triggers were selected, each of them identified 2 employees - one fraudster and one good employee (in Figure 12 fraudsters are shown by black circles, and good employees by white circles). Thus, each trigger identify a group of employees with a 50% share of fraudsters. At the same time, together all the triggers identified a group of 9 employees, only one of which turned out to be a fraudster. That is, the total percentage of fraudsters according to the trigger model was 11% (this is 4.5 times less than the percentage for each trigger individually). If we apply forward selection method, then for this sample only one trigger will be selected, and the whole efficiency of the model will be 50% (that is, it will be equal to the efficiency of the selected trigger).

The second example, illustrated in Figure 12, shows the opposite situation when 9 employees got into the training sample, of which 8 employees were fraudsters. In this example, each trigger identified 2 employees - a fraudster and a good one, i.e. with a 50% share of fraud. At the same time, the total percentage of fraud in all 8 triggers was 89%. Under such conditions, forward selection method allows to identify all 8 triggers, each of them gives an increase to the efficiency of the whole model.

Figure 12. *An example of multicollinearity influence on the general trigger model: 1) correlated rules; 2) uncorrelated rules.*



8.2. Testing

As already demonstrated, triggers for detecting internal fraud are selected from several tens of thousands and even millions of rules. The more rule combinations are used during the tuning of triggers, the more likely that any of the selected triggers will appear randomly in the list of effective triggers. Therefore, after tuning it is necessary to check the selected triggers stability. There are two popular methods for testing trigger stability :

1) on out-off-time test sample (stability test);
2) on a training sample (numerical stability test).

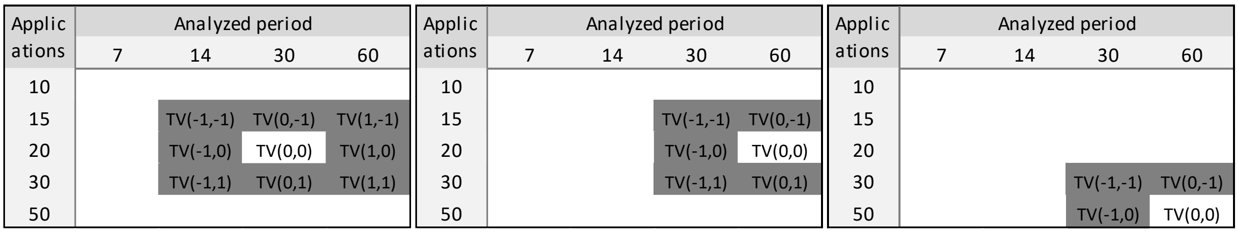
The first method is classic for modeling tasks. On the test sample different from the training sample for each selected trigger, the target value is calculated, which is compared with the threshold value (in our case, the threshold is 0). If the calculated target value is positive (i.e. shows a loss), then the tested trigger is considered stable. If the target value is negative, then the trigger is considered unstable. Since the training sample for predictive fraud analytics problems is almost always formed as a continuous segment (according to the time of loan applications issuing), the test sample in this case will always be out-off-time[5].

[5] That is, test applications are taken from another time interval, different from the applications issuing time and from the training sample.

The second method is to check the numerical stability [20]. The numerical stability for the trigger is defined as follows.

Suppose an effective trigger is defined by the parameters vector $\overline{x}^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, x_4^{(0)}, x_5^{(0)})$ with the target value $TV(\overline{x}^{(0)})$. This vector determines the coordinates of point in the five-dimensional space. If we change any parameter (coordinate) of the vector $\overline{x}^{(0)}$ by one value upward or downward, then we get a set of points $\{\overline{x}_i^{(-1,0,+1)}\} \setminus \{\overline{x}_i^{(0)}\}$ describing the deleted neighborhood of the point with coordinates $\{x_i^{(0)}\}$. A trigger with a target value $TV(\overline{x}^{(0)})$ is considered stable if the total target value of all rules (points) forming a deleted neighborhood of a point $\{x_i^{(0)}\}$ in a certain metric is close to the target value of the trigger itself. Since we have chosen parameter values with a large step, the trigger can be considered stable if the total target value for the points of the deleted neighborhood is positive, that is, the neighboring rules will be unprofitable. If the total target value for the points of the deleted neighborhood is negative (profitable), then the trigger is considered unstable. Figure 13 shows examples of different types of deleted neighborhoods for two-dimensional vector.

Figure 13. *Examples of deleted neighborhoods for two-dimensional trigger with a target value TV (0,0).*



Applications	Analyzed period			
	7	14	30	60
10				
15		TV(-1,-1)	TV(0,-1)	TV(1,-1)
20		TV(-1,0)	TV(0,0)	TV(1,0)
30		TV(-1,1)	TV(0,1)	TV(1,1)
50				

Applications	Analyzed period			
	7	14	30	60
10				
15			TV(-1,-1)	TV(0,-1)
20			TV(-1,0)	TV(0,0)
30			TV(-1,1)	TV(0,1)
50				

Applications	Analyzed period			
	7	14	30	60
10				
15				
20				
30			TV(-1,-1)	TV(0,-1)
50			TV(-1,0)	TV(0,0)

Having calculated stability on out-off-time test sample and numerical stability on the training sample, we can make a conclusion about general stability of trigger. Usually, if the trigger is both time stable and stable in the neighborhood, it is considered stable in a general sense and included in the model. Table 4 presents the results of calculating target values for stability and numerical stability for 18 selected triggers. All the triggers showed positive stability in the neighborhood, however, the trigger 17 turned out to be unstable on the out-off-time sample, so we excluded it from the final model.

Table 4. *Indicators of stability and numerical stability of triggers "Year of issue of passport."*

	Target variable (training sample - 12 months)		Numerical stability (training sample - 12 months)		Stability (out-off-time sample - 6 months)		General stability
#	Average loss	Number of POS	Average loss	Number of POS	Average loss	Number of POS	
Trigger 1	$65 694	1	$32 503	57		0	stable
Trigger 2	$5 284	13	$117	38 764	$1 101	7	stable
Trigger 3	$1 797	1	$94	2 406	$260	4	stable
Trigger 4	$1 020	3	$90	6 382	$15	2	stable
Trigger 5	$964	30	$53	65 712	$3 817	13	stable
Trigger 6	$541	19	$49	71 446	$849	16	stable
Trigger 7	$509	63	$51	33 876	$637	30	stable
Trigger 8	$258	8	$58	21 158	$290	6	stable
Trigger 9	$156	35	$52	2 005	$43	15	stable
Trigger 10	$159	67	$50	32 907	$199	24	stable
Trigger 11	$113	79	$45	141 956	$422	47	stable
Trigger 12	$91	13	$35	3 056	$1 189	5	stable
Trigger 13	$70	70	$32	194 611	$740	41	stable
Trigger 14	$65	127	$17	231 967	$1 018	60	stable
Trigger 15	$1 005	3	$72	11 871	$230	4	stable
Trigger 16	$272	20	$27	111 110	$855	8	stable
Trigger 17	$140	26	$18	44 118	-$162	9	unstable
Trigger 18	$42	74	$8	65 295	$304	23	stable

In addition to the target values and stability of the selected triggers, it is also necessary to control the statistical significance indicators, namely:

- The number of employees/partners included in the calculation;
- The number of loan applications included in the calculation;
- The percentage of employees/partners where the trigger responded, from all employees/partners included into the test sample (hit-rate for employees/partners);
- The percentage of applications where the trigger responded, from all applications included into the test sample (hit-rate for applications);
- The number and percentage of unprofitable employees/partners where the trigger responded;
- The number and percentage of profitable employees/partners where the trigger responded (false negative).

After all the fraud-predictors are calibrated and effective triggers are selected, the total test parameters of the trigger model efficiency are calculated on the test sample: the sum of the actual losses, the sum and the percentage of potential losses.

9. Conclusion

In the above scheme of developing triggers for detecting internal and external fraud in retail lending, the following key points can be highlighted:

1. B-tests are a generalization of Nigrini tests - auditor predictors, built on the Benford's law. At the same time for the development of B-tests the distributions do not have to correspond

the Benford's law, it is enough to select a segment as close as possible to the tested one and build a "reference distribution" on it;

2. To compare the distributions in fraud analytics problems, it is better to use metrics insensitive to the outliers. These metrics include S-statistics and Chi-square statistics. The Kolmogorov-Smirnov statistics are sensitive to the outliers and may underestimate the result of comparing two distributions in fraud analytics problems;
3. For tuning predictors that detect internal fraud, it is best to use the target variable of "potential loss", which shows how much money can be saved averagely per employee or partner. Maximizing the target variable of "potential loss" allows to get operational predictors (detecting internal fraud in the early stages);
4. Tuning of fraud predictors allows to build effective triggers. In this case, a large number of tuning parameters gives millions and even billions of different combinations (rules). With this amount of tuning rules, it is better to use combinatorial optimization methods: Association Analysis, Branch and Bound Algorithm, etc.;
5. For tuning triggers it is better to use the forward selection method or stepwise selection. These methods can reduce the multicollinearity of the selected triggers;
6. With a large number of search options, there is a high probability that any rule is mistakenly (accidentally) can be taken as effective. To minimize the level of such errors, it is necessary to check the selected triggers for stability (on out-off-time test sample) and numerical stability (on training sample);
7. Internal fraud refers to events such as the "black swan" (occurs rarely and has significant consequences), so the data analysis horizon for tuning of fraud predictors should be at least 1-2 years. The same requirement applies to test samples, on which stability is checked;
8. In problems of fraud analytics, the best method for developing predictors is the "bottom-up" ("case-algorithm") analysis, that is, on the basis of the results of investigations, predictors are developed that make it possible to identify such schemes in the future. Therefore, for the development of effective predictors, there must be obligatory expertise by fraud analysts and security officers.

And, probably, the last important question, which we would like to try to answer: Why do B tests work? Why fraudsters, inventing data, cannot "fit" them to market distributions? A simple answer to this question is because it is difficult. The scientific justification for this "simple" answer was given by psychologists-economists Daniel Kahneman and Amos Tversky. Starting their large research on cognitive distortions in risk assessment and potential benefits, Kahneman and Tversky conducted surveys like the following:

> *Someone describes his neighbor: "Steve is very shy and unsociable, always ready to help, but has little interest in surrounding and reality. He is quiet and tidy, loves order and systematic and very attentive to detail. Who is Steve likely to work: a farmer or a librarian?"*

Almost all survey participants noted Steve's similarity to a typical librarian. And practically none of the respondents knew that in USA there were 20 farmers per 1 librarian. The amount of farmers was so much more than librarians, that "quiet and tidy" were more often at the wheel of a tractor than at a librarian table. Carrying out similar experiments, Kahneman and Tversky described about 20 biases arising in the formation of judgments using heuristics. Anyone, including fraudsters, makes errors associated with these biases. Kahneman and Tversky explain this by the fact that emotional thinking has a strong effect on the rational and the person is not able to operate unerringly with complex interrelationships in statistical data [21]. That's why B-tests work.

References

1. Richard J. Bolton and David J. Hand. Statistical Fraud Detection: A Review. Statistical Science, 17(3): 235–255, 1999.
   doi: 10.1214/ss/1042727940

2. Lu. F., J. Efrim. Boritz and Dominic Covvey, “Adaptive Fraud Detection Using Benford’s Law” in Luc Lamontagne and Mario Marchand (eds.) Advances in Artificial Intelligence: Proceedings of the 19 th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2006, Quebec City, Canada, June 2006, Springer 2006, pp. 347-358.
   doi: 10.1007/11766247_30

3. F. Lu and J. E. Boritz. Detecting Fraud in Health Insurance Data: Learning to Model Incomplete Benford’s Law Distributions. In 16th European Conference on Machine Learning, pages 633–640, Porto, Portugal, 2005. Springer.
   doi: 10.1007/11564096_63

4. Simon Newcomb (1881). "Note on the frequency of use of the different digits in natural numbers". American Journal of Mathematics 4 (1): 39–40.
   doi: 10.2307/2369148

5. Frank Benford (March 1938). "The law of anomalous numbers". Proc. Am. Philos. Soc. 78 (4): 551–572.
   doi: 10.2307/984802

6. Arnold, V. I. (1999) "Anti-Scientific Revolution and Mathematics", Vestnik RAS, 1999, No. 6, 553–558.

7. Hal Varian (1972). "Benford's Law (Letters to the Editor)". The American Statistician 26 (3): 65.
   doi:10.1080/00031305.1972.10478934

8. Mark J. Nigrini (May 1999). "I've Got Your Number: How a mathematical phenomenon can help CPAs uncover fraud and other irregulaities". Journal of Accountancy.

9. Walter Wallace. "The Logic of Science and Sociology". Chicago: Aldine-Atherton, 1971.
doi: 10.4324/9781315132976

10. Joseph F. Healey (2002), "Statistics: A Tool for Social Research, Sixth Edition", Wadsworth/Thomson Learning, 2002

11. Greene E, Wellner JA. (2016), "Finite sampling inequalities: an application to two-sample Kolmogorov-Smirnov statistics.", Stoch Process Their Appl. 2016 Dec;126(12):3701-3715.
doi: 10.1016/j.spa.2016.04.020.

12. Anderson, T.W.; Darling, D.A. (1954). "A Test of Goodness-of-Fit". Journal of the American Statistical Association. 49: 765–769.
doi:10.2307/2281537.

13. Sprinthall, R. C. (2011). Basic Statistical Analysis (9th ed.). Pearson Education. ISBN 10: 0205052177. ISBN 13: 9780205052172

14. Kader, Gary (1999). “Means and MADS” Mathematics Teaching in the Middle School, 4(6):398–403.

15. A. H. Land and A. G. Doig (1960). "An automatic method of solving discrete programming problems". Econometrica. 28 (3). pp. 497–520.
doi: 10.1007/978-3-540-68279-0_5

16. Tan Pang-Ning; Michael Steinbach; Vipin Kumar (2005). "Introduction to Data Mining". Addison-Wesley. ISBN-13: 978-0321321367. ISBN-10: 0321321367

17. Donald E. Knuth (1968). The Art of Computer Programming. ISBN 0-201-03801-3

18. Nassim Nicholas Taleb (2010), "The Black Swan: The Impact of the Highly Improbable.", New York: Random House and Penguin. 2007. ISBN 978-1-4000-6351-2. expanded 2nd ed, 2010

19. Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression" Biometrics, 32.
doi: 10.2307/2529336

20. Nicholas J. Higham (1996). "Accuracy and Stability of Numerical Algorithms." Philadelphia: Society of Industrial and Applied Mathematics. ISBN-13: 978-0898713558. ISBN-10: 0898713552.

21. Kahneman, D. (2011). "Thinking, Fast and Slow", Farrar, Straus and Giroux, ISBN 978-0374275631

APPENDIX

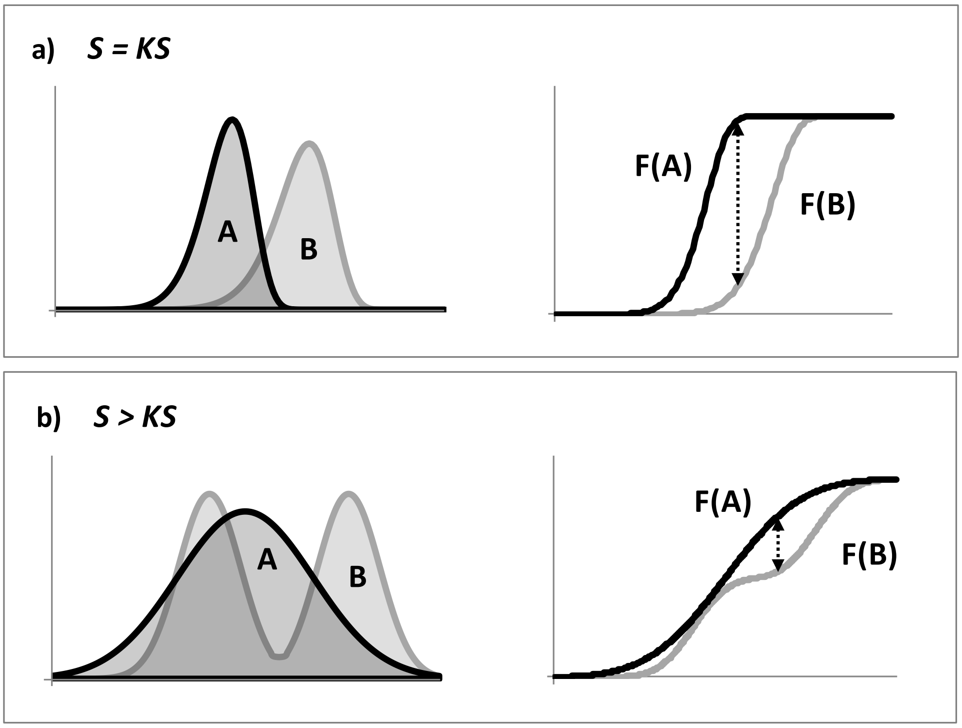
Metrics comparison

Let A and B be two independent samples from a statistical population, then:

$$S(A,B) \geq KS(A,B)$$

where $KS(A,B)$ is the Kolmogorov–Smirnov statistic, $S(A,B)$ – is the S- statistic.

Figure 14. *Metrics comparison: a) S(A,B)=KS(A,B); b) S(A,B)>KS(A,B)*



The proof

Let n be the number of possible outcomes in the sample space, S- statistic for two samples is:

$$S(A,B) = \frac{1}{2}\sum_{i=1}^{n} |a_i - b_i\ |$$

where $a_i = \frac{n_i}{n_a}$ is the relative frequency of the elements x_i in the sample A; b_i - is the relative frequency of the elements x_i in the sample B, sum over all possible outcomes.

Kolmogorov–Smirnov statistic is:

$$KS(A,B) = \max_i |F_a(x_i) - F_b(x_i)|$$

where $F_a(x_i)$ and $F_b(x_i)$ are the empirical distribution functions.

From the definition

$$F_a(x_i) = \sum_{k=1}^{i} a_k$$

Therefore: $a_i = F_a(x_i) - F_a(x_{i-1})$

S-statistics will take the form:

$$S = \frac{1}{2}\sum_{i=1}^{n}\left|(F_a(x_i)\ -\ F_a(x_{i-1})\ ) - \left(F_b(x_i)\ -\ F_b(x_{i-1})\right)\right| =$$

$$= \frac{1}{2}\sum_{i=1}^{n}|(F_a(x_i) -\ F_b(x_i)\ ) - (F_a(x_{i-1}) -\ F_b(x_{i-1})\ )\ |$$

Let Δ_i be equal: $\Delta_i =\ (F_a(x_i) -\ F_b(x_i)\ )$

By the definition of empirical distribution functions and Δ_i:

- $-1\ \leq\ \Delta_i\ \leq 1$
- $\Delta_n = 0$

Kolmogorov–Smirnov statistics and S-statistics will take the form:

$$KS =\ \max_i |\ \Delta_i|$$

$$S =\ \frac{1}{2}\sum_{i=2}^{n}|\Delta_i - \Delta_{i-1}\ | + \frac{1}{2}|\Delta_1|$$

To prove the theorem, it is necessary to open the module. Let $\Delta_i - \Delta_{i-1}$ change sign at points $\{x_1,\ x_2, \cdots, x_m\}$

S-statistics will take the form:

$$\frac{1}{2}\sum_{i=2}^{n}|\Delta_i - \Delta_{i-1}\ | + \frac{1}{2}|\Delta_1| =$$

$$= \frac{1}{2}\ \Big((\Delta_1 - \Delta_2 +\ \Delta_2 - \Delta_3 +\ \cdots +\ \Delta_{k_1-1} - \Delta_{k_1})$$

$$+\ \left(\Delta_{k_1+1} - \Delta_{k_1} +\ \Delta_{k_1+2} - \Delta_{k_1+1} +\ \cdots +\ \Delta_{k_2} - \Delta_{k_2-1}\right) +\ \cdots\Big) + \frac{1}{2}|\Delta_1|$$

Open the brackets and get the expression (1):

$$\frac{1}{2}\sum_{i=2}^{n}|\Delta_i - \Delta_{i-1}\ | + \frac{1}{2}|\Delta_1| =\ \frac{1}{2}\ (\Delta_1 -\ \Delta_{k_1}\ ) +\ \frac{1}{2}\ (\Delta_{k_2} -\ \Delta_{k_1}\ ) + \cdots + \frac{1}{2}|\Delta_1| =$$

$$=\ (-\ \Delta_{k_1}\ ) +\ \Delta_{k_2} +\ \ (-\ \Delta_{k_3}\ ) + \cdots + \frac{1}{2}\Delta_1 + \frac{1}{2}|\Delta_1| =$$

$$= \sum_{j=1}^{m} sgn\left(\Delta_{k_j} -\ \Delta_{k_j-1}\right)\ \Delta_{k_j} + \frac{|\Delta_1|}{2}\ (1 -\ sgn(\Delta_2 -\ \Delta_1)sgn(\Delta_1))$$

Remark: $\Delta_n\ = 0.$ *if* $\Delta_i - \Delta_{i-1}$ does not change sign, then $\Delta_i - \Delta_{i-1} > 0\ with\ \Delta_1\ < 0$ or $\Delta_i - \Delta_{i-1} < 0\ with\ \Delta_1\ > 0$, then $\sum_{j=1}^{m} sgn\left(\Delta_{k_j} -\ \Delta_{k_j-1}\right)\ \Delta_{k_j} = 0$, and $sgn(\Delta_2 -\ \Delta_1)sgn(\Delta_1) = (-1)$.

In this case: $S =\ |\Delta_1|$

Considering monotonic function Δ_i: $|\Delta_1| =\ \max_i |\ \Delta_i| = KS$, therefore $KS =\ S$

Now we show that in the general case $S\ \geq KS$

Let KS be equal: $KS =\ \max_i |\ \Delta_i| =\ |\ \Delta_s|$, then Δ_k is maximum or minimum of a function Δ_i

If Δ_s is maximum then $\Delta_\mathrm{s} > 0$ because $\Delta_n\ = 0$

If Δ_s is minimum then $\Delta_s <\ 0$

Therefore $sgn(\Delta_s -\ \Delta_{s-1})\Delta_s =\ |\ \Delta_s| =\ KS$, and in point Δ_s difference $\Delta_i - \Delta_{i-1}$ change sign.

S-statistics will take the form:

$$S = KS + \sum_{j=1, j\neq s}^{m}\ sgn\left(\Delta_{k_j} -\ \Delta_{k_j-1}\right)\ \Delta_{k_j} + \frac{|\Delta_1|}{2}\ (1 -\ sgn(\Delta_2 -\ \Delta_1)sgn(\Delta_1))$$

$$S - KS = \sum_{j=1, j\neq s}^{m}\ sgn\left(\Delta_{k_j} -\ \Delta_{k_j-1}\right)\ \Delta_{k_j} + \frac{|\Delta_1|}{2}\ (1 -\ sgn(\Delta_2 -\ \Delta_1)sgn(\Delta_1))$$

$S - KS$ divided into 2 parts.

$\frac{|\Delta_1|}{2}\left(1 - sgn(\Delta_2 - \Delta_1)sgn(\Delta_1)\right) > 0$, as

$$\frac{|\Delta_1|}{2}\left(1 - sgn(\Delta_2 - \Delta_1)sgn(\Delta_1)\right) = \begin{cases} 0, & sgn(\Delta_2 - \Delta_1)sgn(\Delta_1) = 1 \\ |\Delta_1|, & sgn(\Delta_2 - \Delta_1)sgn(\Delta_1) = (-1) \end{cases}$$

Now we show that:

$$\sum_{j=1, j\neq s}^{m} sgn\left(\Delta_{k_j} - \Delta_{k_j-1}\right)\Delta_{k_j} \geq 0$$

Consider the following sum separately: $\sum_{j=1,j\neq s}^{m} sgn\left(\Delta_{k_j} - \Delta_{k_j-1}\right)\Delta_{k_j}$

This sum contains only those points Δ_i where difference $\Delta_i - \Delta_{i-1}$ change sign. In this sum there can be negative terms, if:

- $\Delta_i > 0$, for which $sgn(\Delta_i - \Delta_{i-1}) = (-1)$ (the local minimum is greater than zero);
- $\Delta_i < 0$, for which $sgn(\Delta_i - \Delta_{i-1}) = 1$ (the local maximum is less than zero);

In the first case:

Let $\Delta_k > 0,\ sgn(\Delta_k - \Delta_{k-1}) = (-1),\ sgn(\Delta_{k+1} - \Delta_k) = 1$

$\Delta_n = 0$, then there is $\Delta_j > 0, j > k, |\Delta_j| > |\Delta_k|$: $sgn(\Delta_j - \Delta_{j-1}) = 1$, $sgn(\Delta_{j+1} - \Delta_j) = (-1)$

Otherwise $\Delta_n > \Delta_k > 0$

Then Δ_j is the points where $\Delta_j - \Delta_{j-1}$ change sign and Δ_j is contained in the sum

$$\sum_{j=1, j\neq s}^{m} sgn\left(\Delta_{k_j} - \Delta_{k_j-1}\right)\Delta_{k_j}$$

Wherein $sgn(\Delta_j - \Delta_{j-1})\Delta_j > 0$, $|\Delta_j| > |\Delta_k|$

In the second case:

Let $\Delta_k < 0,\ sgn(\Delta_k - \Delta_{k-1}) = 1,\ sgn(\Delta_{k+1} - \Delta_k) = (-1)$

$\Delta_n = 0$, then there is $\Delta_j < 0\ j > k, |\Delta_j| > |\Delta_k|$: $sgn(\Delta_j - \Delta_{j-1}) = (-1)$, $sgn(\Delta_{j+1} - \Delta_j) = 1$

Otherwise $\Delta_n < \Delta_k < 0$

Then Δ_j is the points where $\Delta_j - \Delta_{j-1}$ change sign and Δ_j is contained in the sum

$$\sum_{j=1, j\neq s}^{m} sgn\left(\Delta_{k_j} - \Delta_{k_j-1}\right)\Delta_{k_j}$$

Wherein $sgn(\Delta_j - \Delta_{j-1})\Delta_j > 0$, $|\Delta_j| > |\Delta_k|$

It follows that for any negative term of the sum $\sum_{j=1,j\neq s}^{m} sgn\left(\Delta_{k_j} - \Delta_{k_j-1}\right)\Delta_{k_j}$ here is a positive term that is greater in modulus.

Then $S - KS \geq 0$ as the sum of two nonnegative parts.

Then $S \geq KS$