ISSN 0005-1179 (print), ISSN 1608-3032 (online), Automation and Remote Control, 2024, Vol. 85, No. 8, pp. 785–801.

CONTROL IN SOCIAL ECONOMIC SYSTEMS

The Study of the Strategic Consequences of a Scoring Model Disclosure

G. M. Kryukov[*,a] **and M. S. Sandomirskaia**[**,b]

[*] *New Economic School, Moscow, Russia*
[**] *HSE University, Moscow, St. Petersburg, Russia*
e-mail: [a] *gkryukov@nes.ru,* [b] *msandomirskaya@hse.ru*

Received January 25, 2024
Revised July 1, 2024
Accepted July 10, 2024

Abstract—In this paper, the disclosure of information about the scoring model is investigated. Some of the company's customers find out their internal rating in the company. Such customers can change their behavior to increase their internal rating. The customers who are aware of the leakage are represented as players who can choose a strategy: whether to increase their internal rating and, if so, how much. The main goal is to find the Bayesian-Nash equilibrium in this game and find out how it depends on various parameters, such as the scale of the leakage, the distribution of ratings.

Keywords: scoring model, Bayesian game, manipulation, information disclosure

DOI: 10.31857/S0005117924080046

1. INTRODUCTION

Machine learning methods are being implemented by companies around the world to solve business problems. In particular, companies often use Data Science methods to calculate internal user parameters (for example, the probability of repaying a loan for a bank or attractiveness for an online dating service). A scoring model assigns a certain rating (score) to each user. This rating can be either a discrete random variable (for example, dividing users into several clusters) or continuous (a neural network often returns real values from the interval [0,1]). These models can be useful for solving a wide variety of business problems, such as: "select m users with the greatest propensity for spontaneous purchases to participate in a promotion" or "select m users from a cluster the company needs and offer them a trial version of a new product."

The problem is that information leakages happen from time to time. As a result of such leakages, users can find out what their internal rating is or what internal cluster they belong to. For example, in 2016, the internal rating of users in the dating app Tinder leaked [1]. In addition to the leaked final rating, users can find out how exactly the machine learning model works, as well as the ratings/cluster distribution of other users. Then, users may be motivated to cheat the algorithm: users can change their behavior so that the machine learning algorithm improves their internal rating or assigns them to a better cluster. However, changing behavior and other characteristics related to the algorithm is not free of charge: the user needs to spend time and sometimes money to change their rating (e.g., visit certain pages on the site, apply for a certain product). In addition, the costs of changing the algorithm's assessment may vary for different users. Accordingly, each user decides individually whether to change their behavior to improve their internal rating or leave everything as is.

This paper addresses the question of how users will behave when information about the company's internal user ranking leaks. How does their behavior depend on the values of the scoring

model parameters? To do this, a market is modeled where the company wants to select m clients from n for a certain activity (issuing loans, participating in a profitable promotion, testing a new product, etc.). It is assumed that all clients value this activity equally and that taking part in it is better than not taking. Some users receive information about how the machine learning algorithm works, which ranks/distributes clients into clusters. These users are faced with a strategic choice. On the one hand, they can incur certain costs in order to increase their rating or move to a better cluster, thereby increasing their chances of getting into the m clients who will receive a certain bonus. On the other hand, they can consider their current internal rating as acceptable and not change anything. Thus, clients who have gained access to internal information about the company's machine learning models become players who choose the optimal strategy for themselves. This situation is represented as a Bayesian game, and its equilibrium is further investigated. The relationship between strategic consequences and the type of cost distribution function can be used to analyze the robustness of the scoring model to data leakages. Different scoring models can include parameters that are more or less susceptible to manipulation by an informed client, so that a firm, when choosing a specific scoring model, can predict in advance the potential scale of scoring errors in the case of small or large leakages. This can improve the security and robustness of the scoring models used.

1.1. About Scoring Models in Literature

The focus of this paper is on a machine learning model that scores customers. Customer scoring models are widely used in the industry and are actively evolving. The canonical example is credit scoring models that assess the creditworthiness of customers. Data science methods are actively used to assess credit ratings, including interpretable deep learning [2] and genetic hierarchical networks [3]. The authors in [4] found that after implementing the advanced methods mentioned above, in which customers are not aware of the detailed mechanism of the scoring model, they become less sophisticated in the steps they should take to improve their credit rating. This paper suggests that if the information about the credit scoring model is leaked, customers will obtain a better understanding of how to influence their credit rating, and therefore customers may be willing to manipulate their rating.

It should be noted that companies can evaluate users not only for credit scoring. Advanced machine learning methods are actively used to predict whether a user will perform a certain action, such as buying a certain product [5]. This data are also used to make business decisions about promotion strategies and distribution of personal discounts. Therefore, information about scoring models can also be used by customers to gain additional benefits.

It is worth noting that machine learning algorithms are also used to cluster users into categories [6]. In this case, users also have incentives to manipulate the model when it is disclosed in order to get into a more attractive cluster for themselves.

1.2. About the Game-Theoretic Context in the Model

To model the situation of leakage of the information about the scoring model, a game-theoretic approach will be used. Its effectiveness for studying empirical problems is noted, for example, in [7]. The game component arises from the fact that clients, i.e. players in the model, compete for a finite amount of the good, so that an increase in the rating of an individual client reduces the chances of being selected for others. Having an idea of the scoring model mechanism, the client is able to calculate her own rating, but the exact rating of other clients is unknown to her, and she can only rely on some generally known distribution of ratings in the population. Also, clients act independently and do not observe the actions taken by others, so their interaction can be considered as simultaneous. Thus, a Bayesian game [8] is constructed.

It is worth noting that the process of strategic manipulation is elaborated in game-theoretic models, including optimal control models. For example, [9–12] study optimal control mechanisms in systems with active elements. The most complete review of their ideas is presented in [13]. This direction elaborates the problem of mechanism design, in which the key idea is to disclose information about the type of agents and adequately take into account the agent's behavior in the objective function of the center. Most of the works are devoted to the manipulation problem and to the identification of preference classes for which the control procedure is resistant to a manipulation. Manipulation is understood as the distortion of agent declared preferences, and this distortion does not usually require explicit costs from agents. In this paper, the attention is focused on the associated costs of rating distortion, which is more typical, for example, for works on reputation manipulation [14, 15]. Another important feature of this work is that as a result of the leakage, a certain random subsample of agents becomes active, but not all agents, and inside this subsample, due to heterogeneity in costs, there will be those who prefer to manipulate, and those who prefer to leave their true characteristics. Thus, even under an exogenous narrowing of the set of active agents, the problem of manipulability does not disappear, and its analysis requires joint consideration of the behavior of both active agents and those who are formally present in the system with fixed characteristics.

Since the goal of the scoring model is to select a limited set of customers, the analysis will yield structures typical of works in the field of finite markets [16]. [17] studies the price competition in markets with a rare product and private information about buyers' valuation of the product. The authors consider a market with two sellers, each has one unit of an identical product. Sellers simultaneously choose prices, then buyers choose which seller to go to for the product or not to anyone. The conditions derived when solving the buyer's problem are similar to those that arise in this work when analyzing a game with a single candidate selected by the scoring model (the winner). For a model with several winners, the formulas with a binomial coefficient obtained here are conceptually similar to the results in [18] on Bertrand oligopoly with constraints on the production capacities of firms.

2. MODEL

Let's formalize a game model in which users who have received information about the scoring algorithm decide whether to use the received information to increase their rating.

Let a firm produces a certain product (e.g., a bank that issues credit cards). Let there be n customers who apply for the product (n customers who want to get credit cards). The firm has a limit, e.g., on the amount of plastic for cards, so they are willing to sell only $m < n$ units of the product. Assume that all customers value the product equally. Owning the product brings utility 1.

The firm decides to use a scoring model to classify users. In this article, it is assumed that the scoring model classifies customers into only two categories: good (one) and bad (zero). The scoring model's mechanism is based on some machine learning methods and is essentially a "black box" due to the complexity of its work, but the firm knows the input parameters of this box and, if necessary, can estimate their weight in the final classification result by running the model on a sufficiently wide sample of its customers whose characteristics are known to the firm. Also, using this sample, the firm can approximate the distribution of customer costs to change some of their individual parameters important for the scoring model to values sufficient to be classified as "good."

The game considers a situation where a leakage of information about a scoring model happens. The user who has access to this information learns what class the model assigns her to and what parameters it uses. She also learns the distribution of the rating on the set of all clients (there are many clients, the rating is determined for everyone, not just for n who apply for a credit card or

another reward). However, the user does not know for sure what characteristics other users who learn about the leakage have, i.e. the strategic user makes further decisions under the incomplete information.

Let a customer is classified as good by the scoring algorithm with probability p. The customer finds the disclosed data with probability α. If the customer learns that the model classifies her as a good type, she has no incentive to change her usual behavior. In other words, changing the behavior so that the algorithm classifies her as bad is a weakly dominated strategy and therefore will not be used in the further analysis.

If a customer learns that the algorithm classifies her as a bad type, she has a strategy to change her behavior, bear the costs c_i and mimic a good one. We assume that c_i can be different for different players within the support of the distribution function, each player knows her own value of costs c_i. Indeed, someone needs to do very little to fulfill the conditions of the algorithm for a "good consumer," while other needs to change their behavior greatly and, accordingly, bear large costs. Let the distribution function of the costs of "mimicry" for users whom the model classifies as a bad type be equal to $F(x) = P[c_i \leqslant x]$; for consistency with the normalization of utility to 1, we also assume that the distribution function of costs is defined on $[0; 1]$. If some customers had costs of mimicry higher than the utility from owning the firm's product, then they do not mimic with certainty and can be excluded from consideration. Suppose that $F(x)$ is known to all strategic players. It is appropriate to consider the value of costs as a relative value, i.e. the share of utility from owning the product. Then all further calculations in the model are linearly scaled by the required size of the "prize." As a possible example of a scoring model parameter, whose costs of changing are high for most clients, we can consider the presence and size of a mortgage loan for a borrower (in this example it is easy to see that the costs may exceed the size of the "prize," so such a parameter in the model is unlikely to be subject to distortion). On the contrary, an example of a parameter whose distortion is low-cost is the presence of a completely filled out open profile in a social network.

Note that the detailed knowledge of the internal scoring mechanism is not important for the current analysis. It is sufficient to know only the probability p and the function $F(x)$, which relate both to the scoring mechanism and to the characteristics of the customer population to which scoring is to be applied.

With probability $1 - \alpha$, the customer knows nothing about the leakage. She is not a strategic player, and the information leakage does not affect her internal rating and her behavior. Thus, the only strategic players in this game are agents who are aware of the leakage, which are assigned to the bad class by the scoring model. In this case, the Bayesian type of player i is her costs c_i.

The payoff of a strategic agent from the strategy "NOT to mimic" is equal to the probability of getting the product if the scoring model classifies this player as bad. This is possible only in the situation where there are more products than good customers determined by the algorithm, so that the remaining ones from the bad category must be randomly selected. The payoff of a strategic agent from the strategy "to mimic" is equal to the difference between the probability of getting the product if the model classifies the player as good and the costs of the mimicry. Note that all strategies are chosen by the players simultaneously and independently.

The last issue is how the firm determines exactly m winners. Let k agents be classified by the scoring model as good (they can be either initially good or strategic agents who decided to change their data and cheat the model). If $k > m$, the winners are random m out of k good clients according to the model. If $k \leqslant m$, all good clients according to the model are winners, and $m - k$ winners are additionally randomly selected from the bad ones.

Thus, in the resulting simultaneous game, each client who learned about the leakage and was classified as bad makes a binary choice and decides whether she is ready to invest in improving the

rating in accordance with her costs, or leave everything as is. When a "bad" client, in accordance with the original work of the scoring algorithm, decides to mimic a "good" client, this distorts the results of the model classification and can lead to an incorrect choice of the winner by the firm. The problem is to find the equilibrium strategies for the agents who learned about the leakage, i.e. to determine for each client the optimal choice of the strategy "to mimic" or not, depending on the number of "prizes," the selectivity of the scoring model p, the parameters it uses and, as a result, the distribution of the costs of mimicry of these parameters in the population, the scale of the leakage, and the value of her own costs of improving her type, which the client knows precisely. This choice is not obvious, since an individual client cannot rely solely on its own strategic activity, but must correctly predict the behavior of other active agents that will affect the final probability of being selected. In other words, being among the winners is probabilistic, while the costs of mimicry are deterministic and non-refundable.

By examining the equilibrium strategy of all clients, the firm can predict what proportion of informed clients will decide to mimic and, consequently, how this will distort the scoring results. Since a scoring model can include various parameters, the cost of manipulating which may be different for users, then under the same scale of data leakage, different models may show a greater or lesser degree of the distortion. Accordingly, at the stage of choosing a scoring model, the firm can take into account the further risks from such leakages.

3. NASH EQUILIBRIUM IN A SINGLE-WINNER GAME

Let's consider a special case of the model at $m = 1$. For example, a firm can choose one client who will be the face of a new product. Some of the clients (bad according to the algorithm) will definitely not suit the firm. Among the good ones, the firm can choose arbitrary, since everyone is suitable enough.

We are looking for a symmetric Bayes–Nash equilibrium. Let's denote by y the probability that a bad client will mimic a good one in equilibrium. Note that this probability in the future will consist of the fact that agents of some costs types will deterministically mimic, while other types will not. However, since an individual player does not know the actual type of other agents (for her, it is a random variable), their behavior will also look like random, or probabilistic. Define q as the a posteriori probability that the player will be classified by the algorithm as a good type, taking into account the strategic mimicry of the proportion of consumers who learned about the leakage, then

$$q = p + (1 - p)\alpha y.$$

If the player does not mimic, then they have a chance to win only if everyone else also turns out to be bad *ex-post*. The expected payoff from the "do not mimic" strategy is given by:

$$u_i(0) = \frac{1}{n}(1 - q)^{n-1}.$$

Let's conduct the formula for the expected payoff when choosing the "mimic" strategy. Regardless of the outcome, the player will have to pay c_i. If exactly k of competitors are classified as good, the probability of winning is $\frac{1}{k+1}$. The random variable "the number of competitors to be classified as good" has a binomial distribution: $\mathrm{Bin}(n - 1, q)$. In total, the expected gain from the "mimic" strategy is expressed by the formula

$$u_i(1) = -c_i + \sum_{k=0}^{n-1} \frac{1}{k+1} \binom{n-1}{k} q^k (1 - q)^{n-1-k}.$$

Let's use the following identity proved in [17]:

$$\sum_{k=0}^{n-1} \frac{1}{k+1}\binom{n-1}{k} q^k (1-q)^{n-1-k} = \frac{1-(1-q)^n}{nq}.$$

The expected payoff from the "mimic" strategy is rewritten as follows:

$$u_i(1) = -c_i + \frac{1-(1-q)^n}{nq}.$$

The player will choose the one of the two strategies that will bring her the greatest expected payoff. Note that the expected payoff from the "do not mimic" strategy does not depend on c_i, while the expected payoff from the "mimic" strategy decreases with an increase in c_i. This means that the optimal strategy of player i is monotonic (threshold). In other words, there is such a threshold value of costs c^* that for all $c_i < c^*$ the player i mimics and for all $c_i > c^*$ i the player does NOT mimic. With $c = c^*$, the expected payoffs from both strategies are the same.

Then the probability that the bad type will mimic the good one in equilibrium is equal to the probability that the costs will be lower than the threshold level, and is expressed in terms of c^* as

$$y = P[c_i \leqslant c^*] = F(c^*).$$

The condition for c^* is determined from the equality of expected payoffs from the two pure strategies:

$$c^* = -\frac{1}{n}(1-q)^{n-1} + \frac{1-(1-q)^n}{nq}.$$
$$c^* = \frac{1-(1-q)^{n-1}}{nq}.$$

It is important to clarify that this formula is an implicit expression, because q depends on y, which is uniquely defined through c^*:

$$q(c^*) = p + (1-p)\alpha y = p + (1-p)\alpha F(c^*).$$

Let's introduce the function $f(q)$:

$$f(q) = \frac{1-(1-q)^{n-1}}{nq}.$$

Then the condition for c^* is written as $c^* = f(q(c^*))$.

Theorem 1. *If the costs distribution function $F(c)$ is continuous, then the threshold value c^* exists and is the only one in the optimal monotonic strategy.*

Proof. First, note that $f(q)$ decreases monotonously with the growth of q at $n > 2$ and is condtant at $n = 2$. Let's prove it. Consider the partial derivative

$$\frac{\partial f}{\partial q} = \frac{-1+(1-q)^{n-2}((n-2)q+1)}{nq^2}.$$

The denominator is positive. Consider the numerator. Note that for $n = 2$ the numerator is zero.

The following inequality holds:

$$(1-q)^n(nq+1) > (1-q)^{n+1}((n+1)q+1).$$

Indeed,

$$(1-q)^n(nq+1) - (1-q)^{n+1}((n+1)q+1) = (n+1)q^2(1-q)^n > 0.$$

The monotonicity follows from this statement: the numerator is $-1+(1-q)^{n-2}((n-2)q+1)$ and decreases by n, while it is equal to 0 for $n=2$. It follows that $-1+(1-q)^{n-2}((n-2)q+1) < 0$ for $n > 2$, and from this the partial derivative is negative.

$$\begin{cases} \dfrac{\partial f}{\partial q} = 0, & \text{if } n = 2, \\ \dfrac{\partial f}{\partial q} < 0, & \text{if } n > 2. \end{cases}$$

Now let's prove that $\frac{1}{n} \leqslant c^* \leqslant \frac{n-1}{n}$. To do this, it is enough to show that $\frac{1}{n} \leqslant f(q) \leqslant \frac{n-1}{n}$ for $q \in (0,1)$. Due to the monotonicity, it is enough to calculate the values of the function at the points $q=0$ and $q=1$.

To calculate the limit at the point $q=0$, use the Lopital rule.

$$\lim_{q\to 0} f = \lim_{q\to 0} \frac{1-(1-q)^{n-1}}{nq} = \lim_{q\to 0} \frac{(n-1)(1-q)^{n-2}}{n} = \frac{n-1}{n}.$$

Also calculate the value at the point $q=1$: $f(1) = \frac{1}{n}$.

Then both the function $f(q)$ and the equilibrium value c^* lie in the desired interval.

Finally, it remains to consider the following function:

$$g(c) = c - f(q(c)).$$

The solution of the equation $g(c)=0$ will be the solution of the original equation by c^*. It follows from the properties of the function f that $g(c)$ increases monotonously with c. Also $g(c)$ is continuous.

At the boundaries of the interval one has $g(0) \leqslant -\frac{n-1}{n} < 0$ and $g(1) \geqslant 1 - \frac{1}{n} > 0$. Then, according to the intermediate value theorem, there is a point where $g=0$. The theorem has been proved.

Analysis of the Results in a Single-Winner Game

Let's study the properties of the equilibrium in the game with $m=1$.

Corollary 1. *In a two-player game, $c^* = \frac{1}{2}$ holds for any cost distribution function $F(c)$.*

Indeed,

$$c^* = \frac{1-(1-q)^{2-1}}{2q} = \frac{1}{2}.$$

This means that for $n=2$, a player's decision does not depend on the proportion of good types, the proportion of strategic players, and the distribution of mimicry costs.

In the proof Theorem 1, it was additionally obtained the following.

Corollary 2. *For any values of the model parameters, $\frac{1}{n} \leqslant c^* \leqslant \frac{n-1}{n}$.*

It follows from this that for any values of the model parameters, there will be types of agents for whom it is optimal to mimic and those who will not mimic. If the costs are small $c_i < \frac{1}{n}$, then it is optimal for the player to mimic regardless, for example, of the scale of the leakage or the cost distribution function of other players. Conversely, at high costs $c_i > \frac{n-1}{n}$ the player never mimics.

Statement 1. *The threshold equilibrium value c^* decreases monotonously with the growth of p, α at $n > 2$. Also, c^* has a monotonic dependence on the distribution parameter λ if the distribution function F monotonically depends on λ.*

Proof. Since c^* is not expressed analytically, we will use the implicit function theorem. Consider this equation:

$$g = c^* - f(q(c^*)) = 0.$$

To analyze the dependence of the threshold value on the fraction of p, we determine the sign of the derivative:

$$\frac{\partial c^*}{\partial p} = -\frac{\frac{\partial g}{\partial p}}{\frac{\partial g}{\partial c^*}}.$$

The numerator is:

$$\frac{\partial g}{\partial p} = -\frac{\partial f}{\partial p} = -\frac{\partial f}{\partial q}\frac{\partial q}{\partial p} > 0,$$

since it was previously shown that $\frac{\partial f}{\partial q} < 0$ if $n > 2$, and $\frac{\partial q}{\partial p} = 1 - \alpha F(c^*) > 0$.

The denominator is:

$$\frac{\partial g}{\partial c^*} = 1 - \frac{\partial f}{\partial c^*} = 1 - \frac{\partial f}{\partial q}\frac{\partial q}{\partial c^*} \geqslant 1 > 0,$$

since $\frac{\partial f}{\partial q} < 0$ if $n > 2$, and $\frac{\partial F}{\partial c^*} \geqslant 0$ by definition of the cumulative distribution function, so $\frac{\partial q}{\partial c^*} = (1-p)\alpha\frac{\partial F}{\partial c^*} \geqslant 0$.

It is also true that $\frac{\partial c^*}{\partial p} < 0$.

This result has a natural interpretation. The more likely it is that agents who are really strong according to the model will appear, the more difficult it is for weak, albeit strategic, players to compete with them. This means that fewer bad agents will try to mimic.

Similarly, we prove the dependence on α for $n > 2$.

$$\frac{\partial c^*}{\partial \alpha} = -\frac{\frac{\partial g}{\partial \alpha}}{\frac{\partial g}{\partial c^*}} < 0$$

due to the fact that

$$\frac{\partial g}{\partial \alpha} = -\frac{\partial f}{\partial \alpha} = -\frac{\partial f}{\partial q}\frac{\partial q}{\partial \alpha} = -\frac{\partial f}{\partial q}((1-p)F(c^*)) > 0.$$

This result stems from a similar logic of strategic behavior of agents. The more likely it is that a weak agent who knows about information leakages will appear, the more likely it will be necessary to compete with them too. This means that fewer weak agents, according to the initial scoring model, will try to mimic.

Finally, consider the parametric family of distribution functions $F(c, \lambda)$. Let's determine the possible nature of the dependence on the parameter λ using the implicit function theorem:

$$\frac{\partial c^*}{\partial \lambda} = -\frac{\frac{\partial g}{\partial \lambda}}{\frac{\partial g}{\partial c^*}}.$$

Transform the numerator

$$\frac{\partial g}{\partial \lambda} = -\frac{\partial f}{\partial \lambda} = -\frac{\partial f}{\partial q}\frac{\partial q}{\partial \lambda} = -\frac{\partial f}{\partial q}(1-p)\alpha\frac{\partial F}{\partial \lambda}.$$

It follows that $\operatorname{sgn}(\frac{\partial g}{\partial \lambda}) = \operatorname{sgn}(\frac{\partial F}{\partial \lambda})$, which means

$$\operatorname{sgn}\left(\frac{\partial c^*}{\partial \lambda}\right) = -\operatorname{sgn}\left(\frac{\partial F}{\partial \lambda}\right).$$

In particular, if F monotonously increases (decreases) with the growth of the parameter λ, then the threshold value of c^* monotonously decreases (increases) with the growth of the parameter λ. Thus, the growth of the parameter λ in the cost distribution function means that the cost of mimicry is reduced in society (for example, the new scoring model uses more obvious parameters that can be more easily distorted), and this leads to a decrease in the threshold value at which the player stops trying to mimic.

Statement 1 has been proven.

It became clear how the threshold value changes with a change in the λ parameter, and this result does not seem intuitive until it is found out how the proportion of those strategic agents who decide to mimic changes. This value corresponds to the value of $F(c^*)$. We have a distribution function $F(c,\lambda)$. Denote by $F^{(1,0)}$, $F^{(0,1)}$ the derivatives of F in the first and second variables, respectively. By definition of the distribution function $F^{(1,0)} \geqslant 0$. Let the distribution function depend monotonically on the parameter λ, i.e. the sign $F^{(0,1)}$ is the same for all parameter values. Let's examine the sign of $\frac{\partial F(c^*(\lambda),\lambda)}{\partial \lambda}$. Make some calculations

$$\frac{\partial F(c^*(\lambda),\lambda)}{\partial \lambda} = \frac{\partial c^*}{\partial \lambda}F^{(1,0)} + F^{(0,1)} = -\frac{-\frac{\partial f}{\partial q}\frac{\partial q}{\partial F}F^{(0,1)}}{1-\frac{\partial f}{\partial q}\frac{\partial q}{\partial F}F^{(1,0)}}F^{(1,0)} + F^{(0,1)},$$

that gives

$$\frac{\partial F(c^*(\lambda),\lambda)}{\partial \lambda} = \left(\frac{\frac{\partial f}{\partial q}\frac{\partial q}{\partial F}F^{(1,0)}}{1-\frac{\partial f}{\partial q}\frac{\partial q}{\partial F}F^{(1,0)}} + 1\right)F^{(0,1)}.$$

It was shown above that $\frac{\partial f}{\partial q} < 0$, $\frac{\partial q}{\partial F} > 0$, $F^{(1,0)} \geqslant 0$ for $n > 2$. It is also easy to see that $\frac{x}{1-x} > -1$ for $x \leqslant 0$. It follows from these statements that the expression in large brackets is positive. From here we conclude:

$$\operatorname{sgn}\left(\frac{\partial F(c^*(\lambda),\lambda)}{\partial \lambda}\right) = \operatorname{sgn}(F^{(0,1)}).$$

Let's look at this effect in more details using the example of an exponential distribution. Let's fix $n = 3$, $m = 1$, $p = 0.2$, $\alpha = 0.5$, $F(c,\lambda) = 1 - \exp(-\lambda_i c)$, where $\lambda_1 = 0.5$, $\lambda_2 = 2$.

We solve the equation $c^* - f(q(c^*)) = 0$ numerically and find c^* with an accuracy of four decimal places.

$$\begin{aligned} c^*(\lambda_1) &= 0.5671,\\ c^*(\lambda_2) &= 0.5143,\\ F(c^*(\lambda_1),\lambda_1) &= 0.2469,\\ F(c^*(\lambda_2),\lambda_2) &= 0.6425. \end{aligned}$$

In the example considered, an increase in λ entailed a slight decrease in the threshold value, but significantly increased the proportion of mimics.

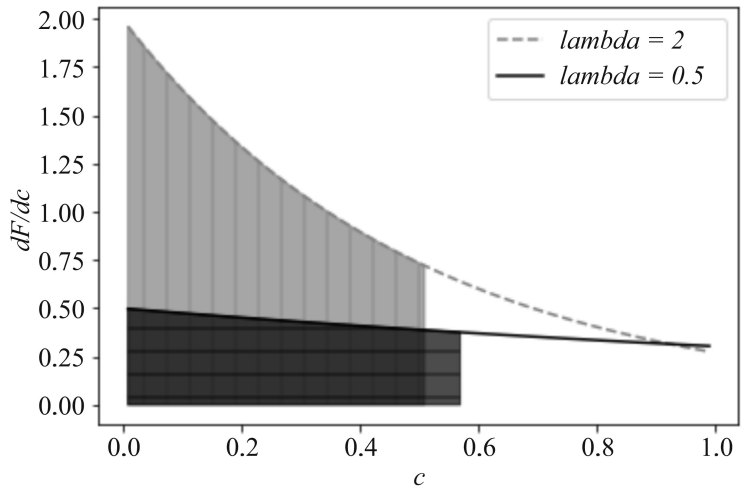

Fig. 1. The equilibrium threshold value and the proportion of mimicking players with exponential cost distribution.

Figure 1 shows a graph with distribution densities at $\lambda_1 = 0.5$, $\lambda_2 = 2$ and the values of the proportion of mimics in equilibrium as the area under the graphs at $c \in [0, c^*]$.

Let's summarize the results. Suppose the cost distribution has changed in such a way that $F(c)$ has become larger for all values of c. This is equivalent to simplifying the scoring model and increasing its vulnerability. Then the threshold value in the equilibrium of c^* will decrease, but not so much: the probability that the consumer will mimic will increase, as well as the proportion of mimics. In addition, as it was found out earlier, c^* decreases with the growth of p and α. Since these parameters do not affect the distribution, the probability that the consumer will mimic also decreases with the growth of p and α.

4. NASH EQUILIBRIUM IN A GAME WITH m WINNERS

Now let $m \leqslant n$ good players get access to the product according to the scoring model. We will look for a symmetric Bayesian–Nash equilibrium. Let y be the probability that the bad one will mimic the good one in equilibrium. Similarly to the case of a single winner, we define q as the probability that the player will be classified by the algorithm as a good type. This is calculated using the full probability formula:

$$q = p + (1 - p)\alpha y.$$

Let's find the expected utilities of the ith player, who is a strategic agent, from the "mimic" and "do not mimic" strategies. Let k competitors be classified as a good type. As in the case of a single winner, the random variable "the number of competitors who will be classified as good" has a binomial distribution: $\mathrm{Bin}(n-1, q)$.

If k competitors are classified as a good type and the ith player chooses the "do not mimic" strategy, then if $m \leqslant k$ all places will be occupied by players who will be classified as good. If $m > k$, then there remains $m-k$ of products for bad players according to the model, the probability of getting them is $\frac{m-k}{n-k}$.

If k competitors are classified as a good type and the ith player chooses the "mimic" strategy, then if $m > k$ all the players who are classified as good by the model are guaranteed to receive the goods. This means that the ith player will receive the product with probability 1. If $m \leqslant k$, then $k+1$ good players according to the model will compete for m places. Therefore, the probability of receiving the product is $\frac{m}{k+1}$.

Therefore, the expected gain from the "do not mimic" strategy is

$$\sum_{k=0}^{n-1} \max\left(0, \frac{m-k}{n-k}\right) \binom{n-1}{k} q^k (1-q)^{n-1-k}.$$

The expected gain from the "mimic" strategy is

$$-c_i + \sum_{k=0}^{n-1} \min\left(1, \frac{m}{k+1}\right) \binom{n-1}{k} q^k (1-q)^{n-1-k}.$$

Note that the expected gain from the "do not mimic" strategy does not depend on c_i, while the expected gain from the "mimic" strategy decreases with an increase in c_i. This means that the optimal strategy of the player i is monotonous in the same meaning as in Section 3: there is a c^* such that for all $c_i < c^*$ ith player mimics and for all $c_i > c^*$ ith player does not mimic. With $c = c^*$, the expected gains from both strategies are the same.

The probability that a bad type will mimic a good one in equilibrium is expressed in terms of c^*:

$$y = P[c_i \leqslant c^*] = F(c^*).$$

By equating the expected gains from pure strategies, we obtain a condition for the threshold value of c^*:

$$c^* = \sum_{k=0}^{n-1} \left(\min\left(1, \frac{m}{k+1}\right) - \max\left(0, \frac{m-k}{n-k}\right)\right) \binom{n-1}{k} q^k (1-q)^{n-1-k},$$

where

$$q = p + (1-p)\alpha F(c^*).$$

Let's transform the resulting expression. If $k < m$, then

$$\min\left(1, \frac{m}{k+1}\right) - \max\left(0, \frac{m-k}{n-k}\right) = 1 - \frac{m-k}{n-k} = \frac{n-m}{n-k}.$$

If $k \geqslant m$, then

$$\min\left(1, \frac{m}{k+1}\right) - \max\left(0, \frac{m-k}{n-k}\right) = \frac{m}{k+1} - 0 = \frac{m}{k+1}.$$

Then we get the following expression for the threshold value:

$$c^* = \sum_{k=0}^{m-1} \frac{n-m}{n-k} \binom{n-1}{k} q^k (1-q)^{n-1-k} + \sum_{k=m}^{n-1} \frac{m}{k+1} \binom{n-1}{k} q^k (1-q)^{n-1-k}.$$

For the convenience of the equilibrium analysis, we introduce an additional function

$$\tilde{f}(m,q) = \sum_{k=0}^{m-1} \frac{n-m}{n-k} \binom{n-1}{k} q^k (1-q)^{n-1-k} + \sum_{k=m}^{n-1} \frac{m}{k+1} \binom{n-1}{k} q^k (1-q)^{n-1-k}.$$

Then the condition is rewritten as

$$c^* = \tilde{f}(m, q(c^*)),$$
$$q(c^*) = p + (1-p)\alpha F(c^*).$$

Analysis of the Results of the Game with m Winners

Let's start by analyzing the dependence of the equilibrium threshold on q. This dependence reflects the marginal increase in the equilibrium threshold with the a posteriori probability of mimic. Technically, this is an important object, since the effect of other parameters on the equilibrium ultimately depends on the sign of df/dq.

Statement 2. *The function $\tilde{f}(m,q)$ is monotone in q if and only if $m=1$ or $m=n-1$.*

Proof. First of all, note that $\tilde{f}(m=1,q)=\tilde{f}(m=n-1,1-q)$.

As it was proved in the analysis of the game with one winner, $\tilde{f}(m=1,q)$ strictly decreases with the growth of q. Then it follows that $\tilde{f}(m=n-1,q)$ strictly increases with the growth of q.

Now we show that for $1<m<n-1$, the function $\tilde{f}(m,q)$ is not monotonic with respect to q. To do this, calculate the limits of the derivative at $q\to 0$ and $q\to 1$.

$$\frac{\partial\tilde{f}}{\partial q}=\sum_{k=0}^{m-1}\frac{n-m}{n-k}\binom{n-1}{k}q^{k-1}(1-q)^{n-2-k}(k-(n-1)q)$$
$$+\sum_{k=m}^{n-1}\frac{m}{k+1}\binom{n-1}{k}q^{k-1}(1-q)^{n-2-k}(k-(n-1)q).$$

Let's use the fact that $\lim\limits_{q\to 0}q^a(1-b)^b=1$ for $a=0,\ b>0$.

$$\lim_{q\to 0}\frac{\partial\tilde{f}}{\partial q}=\begin{cases}-(n-1)\dfrac{n-m}{n}+\dfrac{m}{2}\binom{n-1}{1}, & \text{if } m=1,\\ -(n-1)\dfrac{n-m}{n}+\dfrac{n-m}{n-1}\binom{n-1}{1}, & \text{if } m>1,\end{cases}$$
$$=\begin{cases}-\dfrac{(n-1)(n-2)}{2n}, & \text{if } m=1,\\ \dfrac{n-m}{n}, & \text{if } m>1.\end{cases}$$
$$\lim_{q\to 1}\frac{\partial\tilde{f}}{\partial q}=\begin{cases}-\dfrac{n-m}{n-(n-2)}(n-1)+\dfrac{m}{n}(n-1), & \text{if } m=n-1,\\ -\dfrac{m}{n-1}(n-1)+\dfrac{m}{n}(n-1), & \text{if } m<n-1,\end{cases}$$
$$=\begin{cases}\dfrac{(n-1)(n-2)}{2n}, & \text{if } m=n-1,\\ -\dfrac{m}{n}, & \text{if } m<n-1.\end{cases}$$

Thus, three cases are possible:

1. $m=1$. In this case, $\frac{\partial\tilde{f}}{\partial q}<0$.

2. $m=n-1$. In this case, $\frac{\partial\tilde{f}}{\partial q}>0$.

3. $1<m<n-1$. In this case, $\lim\limits_{q\to 0}\frac{\partial\tilde{f}}{\partial q}>0$ and $\lim\limits_{q\to 1}\frac{\partial\tilde{f}}{\partial q}<0$. Then it follows from the continuity of $\frac{\partial\tilde{f}}{\partial q}$ that the function reaches a maximum at $q\in[0,1]$ at $q^*\in(0,1)$.

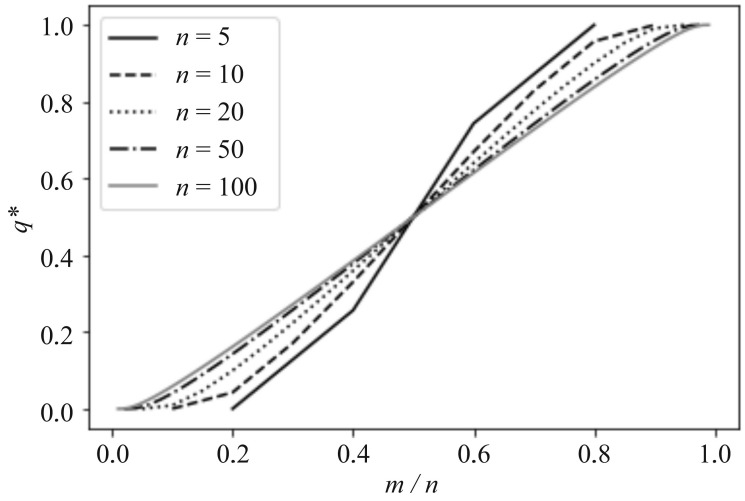

Fig. 2. The maximum point of $\tilde{f}(q)$ depending on n and $\frac{m}{n}$.

So, it is shown that if $1 < m < n-1$, then $\tilde{f}(m,q)$ is not monotonic in q. Apparently, the sums used to describe $\tilde{f}(m,q)$ are too complex and cannot be simplified in the form of elementary functions. Therefore, it is not possible to analytically express q^*, which maximizes $\tilde{f}(m,q)$ at a fixed m.

Statement 2 has been proven.

Let's consider a remarkable special case $n = 2m$, $m > 1$. Let's suppose $k \geqslant m$. Then, for $n = 2m$, $n-1-k < m$ holds. Also note that for $n = 2m$,

$$\frac{n-m}{n-(n-1-k)}\binom{n-1}{n-1-k} = \frac{m}{k+1}\binom{n-1}{k}.$$

In addition, when $q = \frac{1}{2}$, it is true that $q^{k-1}(1-q)^{n-2-k} = 2^{3-n}$.

These facts allow us to calculate the partial derivative at $q = \frac{1}{2}$.

$$\frac{\partial \tilde{f}}{\partial q}\left(n = 2m, q = \frac{1}{2}\right) = \sum_{k=m}^{n-1} \frac{m}{k+1}\binom{2m-1}{k} 2^{3-2m}(k + (2m-k-1) - (2m-1)) = 0.$$

In general, it is not easy to show that the given point is the maximum point. For example, consider $n = 4, m = 2$. The second-order condition is:

$$\frac{\partial^2 \tilde{f}}{\partial q^2}(n = 4, m = 2) = -1.$$

Thus, $\tilde{f}$ is a concave function with respect to q at $n = 4,\ m = 2$, which means that at the point $q = 0.5$ it reaches a global maximum.

An interesting question is what the maximum point of the function $\tilde{f}$ tends to (denote by $q^*(n,m)$) for large m. To do this, we find q^* for fixed n, m by numerical methods.

We note that the larger n, the more the dependence $q^*(m)$ is similar to a linear one (already at $n = 100$ we note $R^2 > 99.9\%$ for regression $q^* = \beta_0 + \beta_1 m$; R^2 grows with the growth of n). In addition, with the growth of n $q^*(m=2)$ tends to 0, and $q^*(m = n-2)$ tends to 1. This allows us to hypothesize that for large n and $1 < m < n-1$, the maximum point of $\tilde{f}(q)$ tends to $q^* = \frac{m-2}{n-4}$ (Fig. 2) The verification of this hypothesis is interesting for the further research; in the case of truth, the intuition of the result is interesting.

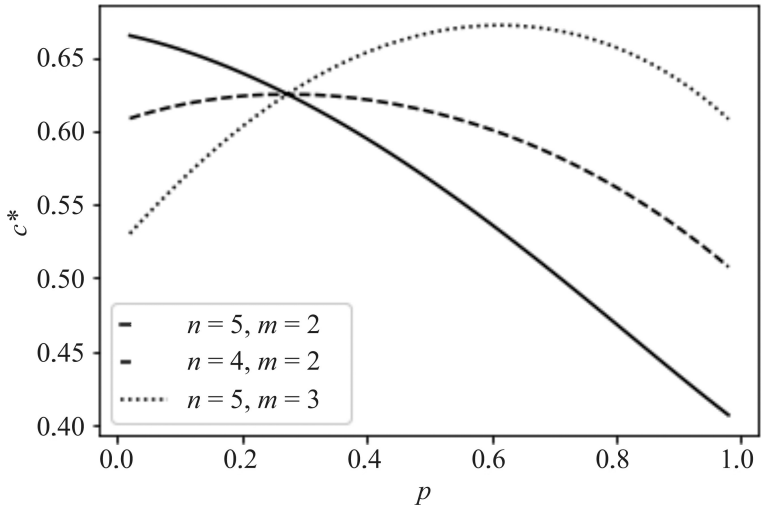

Fig. 3. Dependence of c^* on p at $\alpha = 0.5$, $F(c) = c$.

Statement 3. *In general, the threshold value of c^* is non-monotonic with respect to p and α.*

Proof. Let's calculate the derivatives of the function given implicitly, as was done earlier in Statement 2.

$$\tilde{g} = c^* - \tilde{f}(q(c^*)) = 0.$$

$$\operatorname{sgn}\left(\frac{\partial c^*}{\partial p}\right) = \operatorname{sgn}\left(-\frac{\frac{\partial \tilde{g}}{\partial p}}{\frac{\partial \tilde{g}}{\partial c^*}}\right) = \operatorname{sgn}\left(\frac{\partial \tilde{f}}{\partial q}\right).$$

$$\operatorname{sgn}\left(\frac{\partial c^*}{\partial \alpha}\right) = \operatorname{sgn}\left(-\frac{\frac{\partial \tilde{g}}{\partial \alpha}}{\frac{\partial \tilde{g}}{\partial c^*}}\right) = \operatorname{sgn}\left(\frac{\partial \tilde{f}}{\partial q}\right).$$

Partial differential transformations are similar to the transformations in statement 1, when these derivatives were calculated for the case of $m = 1$. In Statement 2, the nonmonotonicity of $\tilde{f}$ by q was proved in the general case. It follows that $\operatorname{sgn}\left(\frac{\partial \tilde{f}}{\partial q}\right)$ is not a constant value. But then $\operatorname{sgn}\left(\frac{\partial c^*}{\partial p}\right), \operatorname{sgn}\left(\frac{\partial c^*}{\partial \alpha}\right)$ are not constant values for $1 < m < n-1$. The case of $m = 1$ was considered earlier, for $m = n-1$ all conclusions are opposite to the case of $m = 1$ (i.e. $\frac{\partial c^*}{\partial p} > 0$, $\frac{\partial c^*}{\partial \alpha} > 0$ for $m = n-1$). Moreover, it is obvious that the maxima for p and for q are also uniquely related.

Statement 3 has been proven.

Let's demonstrate nonmonotonicity with examples. Numerically calculate the value of c^* for fixed parameter values from the conditions $c^* = \tilde{f}(m, q(c^*))$, $q = p + (1-p)\alpha F(c^*)$. Let's fix that the costs are uniformly distributed over the interval $[0, 1]$, and find several solutions to the equations for different parameter values with an accuracy of up to the fourth decimal place.

$$c^*(n = 4, m = 2, p = 0.1, \alpha = 0.5) = 0.6175,$$
$$c^*(n = 4, m = 2, p = 0.3, \alpha = 0.5) = 0.6248,$$
$$c^*(n = 4, m = 2, p = 0.5, \alpha = 0.5) = 0.6132.$$

Thus, there is generally no monotonicity of c^* by p (Fig. 3). Intuitively, we explain this fact: with increasing probability p, the initial number of good agents increases, so that under a small proportion of them, it makes sense to actively mimic and fight for a relatively large number of

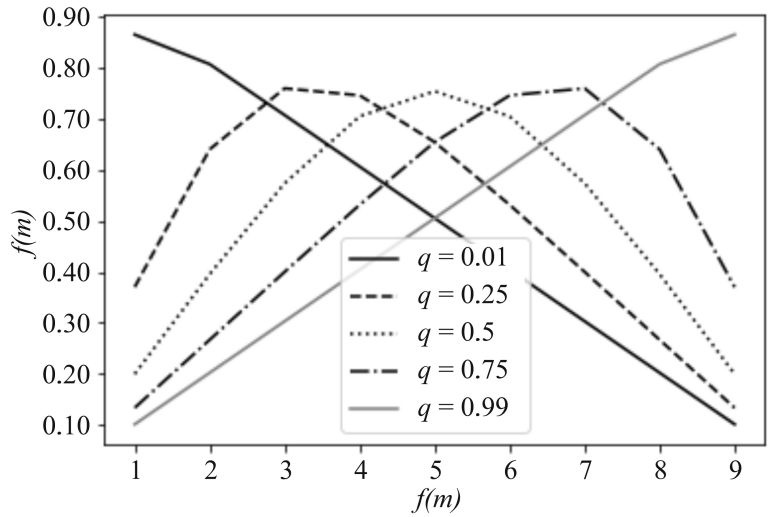

Fig. 4. Dependence of c^* on m at $n = 20$, $\alpha = 0.2$, $F(c) = c$.

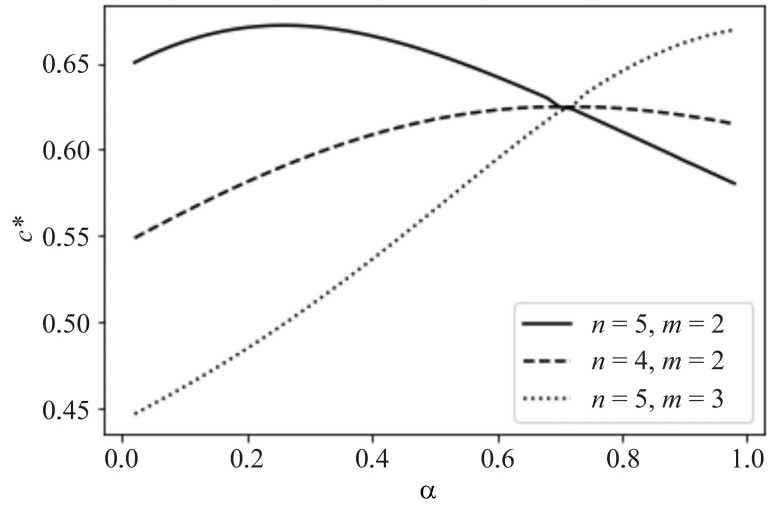

Fig. 5. Dependence of c^* on α at $p = 0, 1$, $F(c) = c$.

"prizes." However, if their share is already high, then it is difficult to compete with them, the probability of winning is too low, so the proportion of mimicking players (and the corresponding threshold value) falls.

Similarly, let's look at the dependence on the scale of the leakage α.

$$c^*(n = 4, m = 2, p = 0.1, \alpha = 0.5) = 0.6175,$$
$$c^*(n = 4, m = 2, p = 0.1, \alpha = 0.7) = 0.6248,$$
$$c^*(n = 4, m = 2, p = 0.1, \alpha = 0.9) = 0.6198.$$

Thus, there is generally no monotonicity of c^* by α (Fig. 4). A similar logic applies here: with increasing leakage, more and more weak people will potentially want to mimic the strong ones and up to some level it will be beneficial for them. However, the further competition will become too tough and therefore the desire to mimic will decrease.

Statement 4. *The threshold value c^* is non-monotonic with respect to m in the general case.*

As in Statement 3, we calculate c^* for various parameter values with an accuracy of up to four decimal places and show the non-monotonicity of the threshold value in equilibrium by the number of winners.

$$c^*(n = 4, m = 1, p = 0.5, \alpha = 0.5) = 0.3910,$$
$$c^*(n = 4, m = 2, p = 0.5, \alpha = 0.5) = 0.6132,$$
$$c^*(n = 4, m = 3, p = 0.5, \alpha = 0.5) = 0.5045.$$

In this case, the explanation is that with the number of winners growth close to the number of all participants, it may become pointless to try and spend money on mimicry, since even without this, the probability of entering the number of m winners is high. However, as can be seen from Fig. 5, for special scoring models that generate a very high or very low probability of classifying a player as good, there is a monotonous dependence of the proportion of mimics on the number of prizes.

5. CONCLUSION

The article models a situation in which some of the firm's clients learn their internal rating in the company and can change their behavior to increase their internal rating. In the considered framework, the scoring model splits users into "good" and "bad" (binary classification of agent types).

It is proven that equilibrium exists, is unique and is a profile of monotonic strategies. The presence of an internal threshold value indicates that not all agents, obtaining access to the mechanism of the scoring model, use this information for manipulation.

We believe that there is great potential for further research. In the case of a binary distribution of types, it is interesting to study the discovered non-monotonic dependencies in more details, in particular, to find the extreme points analytically depending on the parameter values. It may be useful to extend the binary classification to a discrete one, since some scoring models distribute users across several clusters, and to a continuous one, since many models return the rating as a real number, often from the interval $[0, 1]$.

In conclusion, it should be noted that the considered problem of behavior manipulation is an example of the manifestation of the property of natural intelligence to continuously adapt and search for complex and non-trivial strategies to improve its utility. In this sense, automated models are unlikely to be able to fully resist all the possibilities potentially available to humans. However, if one predicts the strategic behavior of users in advance, she can more effectively assess both the stability and prospects of the original model, which can affect the optimal choice of the model, or formulate more adequate criteria for improving the model. Another related problem is the manipulation of behavior during model training, which will certainly affect its further functioning. This leads to the proposal that the optimal trajectory of the development of the machine learning and artificial intelligence should include an adequate game-theoretic element, taking into account that the data for these models are generated by strategic people.

FUNDING

This work was supported by HSE University (Basic Research Program).

REFERENCES

1. Carr, A., I found out my secret internal tinder rating and now I wish I hadn't, https://www.fastcompany.com/3054871/whats-your-tinder-score-inside-the-apps-internal-ranking-system (Accessed: 22.01.2024)

2. Albanesi, S. and Vamossy, D., Predicting consumer default: A deep learning approach, *National Bureau of Econom. Res.*, 2019, no. w26165, 72 p.

3. Pławiak, P., Abdar, M., Plawiak, J., et al., DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring, *Inform. Sci.*, 2020, vol. 516, pp. 401–418.

4. Hurley, M. and Adebayo, J., Credit scoring in the era of big data, *Yale JL & Tech.*, 2016, vol. 18, no. 148.

5. Martínez, A., Schmuck, C., Pereverzyev, S., Pirker, C., and Haltmeier, M., A machine learning framework for customer purchase prediction in the non-contractual setting, *Eur. J. Oper. Res.*, 2020, vol. 281, no. 3, pp. 588–596.

6. Syakur, M.A. et al., Integration k-means clustering method and elbow method for identification of the best customer profile cluster, *IOP conference series: materials science and engineering*, IOP Publishing, 2018, vol. 336, no. 012017.

7. Roth, A., Game theory as a part of empirical economics, *The Econ. J.*, 1991, vol. 101, no. 404, pp. 107–114.

8. Harsanyi, J., Games with incomplete information played by "bayesian" players, I–III part I. The basic model, *Management Sci.*, 1967, vol. 14, no. 3, pp. 159–182.

9. Burkov, V.N., *Osnovy matematicheskoi teorii aktivnyh sistem* (Foundations for mathematical theory of active systems), Moscow: Nauka, 1977.

10. Burkov, V.N. and Novikov, D.A., Teorii aktivnyh system 50 let: istoriia razvitiia (50 years to the theory of active systems), *Matherials of international scientific-practical conference "Theory of active systems – 50 years"*, Moscow: ICS, 2019, pp. 10–57.

11. Enaleev, A.K., The optimality of correlated mechanisms of funtioning in active systems, *Upravlenie bolshimi sistemami*, 2011, no. 33, pp. 143–166.

12. Enaleev, A.K., The optimal correlated mechanism in the system with several active agents, *Problemy upravleniia*, 2015, no. 3, pp. 20–28.

13. Burkov, V.N., Enaleev, A.K., and Korgin, N.A., Incentive Compatibility and Strategy-Proofness of Mechanisms of Organizational Behavior Control: Retrospective, State of the Art, and Prospects of Theoretical Research, *Autom. Remote Controle*, 2021, no. 7, pp. 5–37.

14. Dellarocas, C., Strategic manipulation of internet opinion forums: Implications for consumers and firms, *Management Sci.*, 2006, vol. 52, no. 10, pp. 1577–1593.

15. Dini, F. and Spagnolo, G. Buying reputation on eBay: Do recent changes help?, *Int. J. Electron. Business*, 2009, vol. 7, no. 6, pp. 581–598.

16. Wright, R. et al., Directed search and competitive search equilibrium: A guided tour, *J. Econ. Lit.*, 2021, vol. 59, no. 1, pp. 90–148.

17. Sandomirskaia, M. and Shavshin, R., Price Competition in Finite Markets with a Rare Good and Private Consumer Valuations, *Higher School Econom. Res. Paper*, 2021, vol. 248, 29 p.

18. Peters, M., Bertrand equilibrium with capacity constraints and restricted mobility, *Econometrica*, 1984, vol. 52, no. 5, pp. 1117–1127.

This paper was recommended for publication by D.A. Novikov, a member of the Editorial Board