Как отпугнуть клиента и получить прибыль: фильтрация в системе массового обслуживания

София Сурова[1], **Кирилл Фурманов**[2]

[1] Национальный исследовательский университет “Высшая школа экономики”, г. Москва, Россия
[2] Центральный экономико-математический институт РАН г. Москва, Россия

Информация о статье
Поступила в редакцию: 02.03.2024
Принята к опубликованию: 23.04.2024

УДК 303.4

JEL C60

Ключевые слова: фильтрация, система массового обслуживания, нетерпеливые заявки

Keywords: filtration, queueing systems, balking

Аннотация
Рассматривается система массового обслуживания с ограничением на длину очереди и разнородным потоком входящих заявок (клиентов), которые делятся на два типа. Заявки первого типа терпеливы: они встают в очередь, если это позволяет ёмкость системы и дожидаются обслуживания. Заявки второго типа нетерпеливы: они отказываются вставать в очередь, если время ожидания обслуживания оказывается слишком велико. Показывается, что если заявки второго типа приносят меньшую прибыль, чем заявки первого типа, то увеличение прибыли может достигаться за счёт замедления обслуживания, отпугивающего нетерпеливых клиентов.

How to Scare a Customer Away and Get Profit: Filtration in a Queueing System

Sofia V. Surova, Kirill K. Furmanov

Abstract
We consider a simple exponential queueing system with a finite capacity and heterogeneous customers. Type I customers are patient, they join the queue if the system capacity allows it. Type II customers are impatient, so that they refuse to join the queue if the waiting time is too large (so-called wait-based balking). The system makes profit servicing customers, and the profit depends on the type of the customer. We show that if type II customers bring less profit than type I customers then the administration may, in certain cases,

DOI: https://doi.org/10.24866/2311-2271/2024-1/1113

increase profit by slowing down the service (decreasing the service rate). It makes the system unattractive for relatively unprofitable type II customers who stop forming the queue and thus leave the place for more profitable customers who otherwise would not have joined the queue due to capacity restriction.

Введение

В ряде экономических моделей особое внимание уделяется фильтрации — действиям экономических агентов, косвенным образом препятствующих неблагоприятному отбору. Так, работодатель может поставить барьер для трудоустройства в виде требования образования, которое может не пригодиться на работе непосредственно, но благодаря этому требованию отсеиваются малопроизводительные кандидаты, для которых издержки получения образования относительно высоки [1–3]. Как правило, фильтрация моделируется в рамках теории игр: работодатель делает ходы так, чтобы выигрышной стратегией желательных кандидатов было поведение, отличающее их от нежелательных.

Настоящая статья представляет попытку описания аналогичного процесса в рамках теории массового обслуживания. Мы рассматриваем систему, в которую поступают заявки (клиенты) двух типов, причём заявки первого типа терпеливы и встают в очередь, если это позволяет ёмкость системы, а заявки второго типа нетерпеливы: они отказываются вставать в очередь, если предполагаемое время ожидания в очереди слишком велико. Мы показываем, что, если заявки второго типа приносят меньшую прибыль, чем заявки первого типа, администрация системы в некоторых случаях может увеличивать прибыль за счёт замедления обслуживания и искусственного создания очереди. При этом нетерпеливые и относительно невыгодные клиенты будут “фильтроваться” — отказываться присоединяться к очереди, освобождая место для относительно выгодных и терпеливых.

Описание системы

Рассмотрим механизм фильтрации в простейшей системе массового обслуживания (СМО) — системе M/M/1/2 согласно нотации Кендалла. Это система с простейшим потоком входящих заявок, показательно распределённым временем обслуживания, одним каналом обслуживания и ёмкостью в две заявки (т.е. очередь вмещает не более одной заявки, в это время вторая заявка обслуживается). Система представлена схематически на рис. 1.

Каждая из поступающих заявок независимо от прочих принадлежит к одному из двух типов. Заявки первого типа при поступлении входят в систему, если очередь не переполнена. Каждая такая заявка приносит единицу прибыли (допустима также случайная величина прибыли — тогда её математическое ожидание равно единице). Заявки второго типа отличаются нетерпеливостью типа balking [4–5]: они входят в систему, если это позволяет ёмкость — т.е. очередь не переполнена — и, если математическое ожидание времени нахождения в очереди не превышает допустимый для этих заявок предел θ. Каждая за-

явка второго типа приносит прибыль ***r***, если входит в систему. Заявки, потерянные из-за ограничения ёмкости или из-за нетерпения, прибыли не приносят.

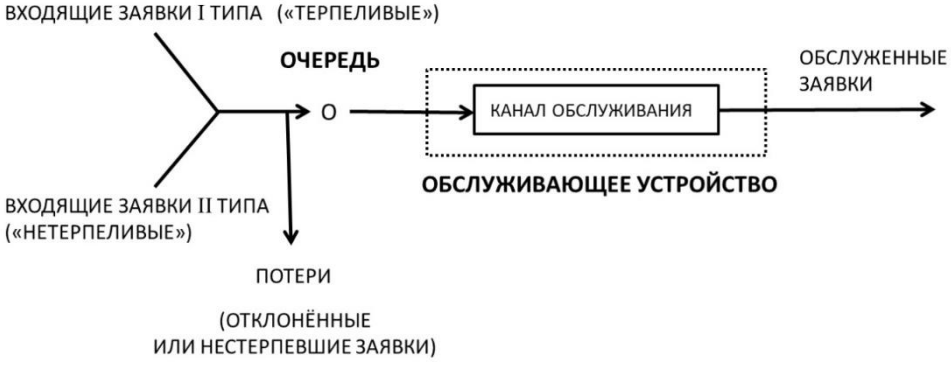


Рис. 1. Схема рассматриваемой СМО

Как пример такой системы можно представить себе кафе быстрого питания с одним окном заказа и выдачи для автомобилистов, где на подъезде к окну есть место только для одной машины. Часть клиентов заказывает себе полноценный обед, а другие отвлекают кассира, чтобы в конце концов купить себе малостоящую закуску. Для администрации кафе может быть выгоднее, чтобы доля последних была пониже, так как они занимают место в очереди, из-за чего выгодные клиенты не могут заехать и сделать заказ.

Обозначим интенсивность входящего потока заявок λ, а долю заявок первого типа (терпеливых) ***s***. Тогда среднее числа заявок первого типа, поступающих за единицу времени, равно λs, а заявок второго типа — $\lambda(1 - s)$. Так как мы свободны в выборе единицы измерения времени, положим без ограничения общности $\lambda = 1$. В этом случае приведённая интенсивность входящего потока (отношение интенсивности входящего потока к интенсивности обслуживания) совпадает со средним временем обслуживания — обозначим эту величину ρ. Будем считать, что время обслуживания не зависит от типа заявки.

Стационарное распределение числа заявок в системе

Система описывается процессом размножения и гибели с тремя возможными состояниями — по числу заявок в системе: 0, 1 и 2. Выражения, описывающие стационарное распределение, зависят от соотношения среднего времени обслуживания ρ и максимально допустимого ожидаемого времени в очереди для нетерпеливых заявок θ.

Время в очереди, для только что вставшей в очередь заявки, имеет то же распределение, что и остаточное время, которое потребуется для завершения работы над обслуживаемой в текущий момент заявкой. По свойству отсутствия последействия показательного распределения это время имеет показательное распределение с математическим ожи-

данием ρ. Таким образом, нетерпеливые заявки будут присоединяться к очереди в случае $\rho \le \theta$ и уйдут без обслуживания при $\rho > \theta$.

Случай 1: $\rho \le \theta$. Все заявки входят в систему, если позволяет ёмкость. Это классическая система M/M/1/2 с хорошо известным распределением числа заявок (см. например [6, 7]):

$$p_0 = p_1 = p_2 = \frac{1}{3}, \rho = 1;$$
$$p_j = \frac{(1-\rho)\rho^j}{1-\rho^3}, j = 0,1,2, \rho \neq 1.$$

Здесь p_j — вероятность пребывания системы в состоянии j в стационарном режиме, совпадает с долей времени в долгосрочном периоде, в течение которого в системе находится j заявок.

В дальнейшем удобнее будет использовать единое выражение для вероятностей, доступное, в нашем случае, благодаря точно известной ёмкости системы в две заявки:

$$p_j = \frac{\rho^j}{1+\rho+\rho^2}, j = 0,1,2. \quad (1)$$

Случай 2: $\rho > \theta$. Заявки второго типа не соглашаются стоять в очереди и входят в систему, только когда она свободна. Это также вариант процесса размножения и гибели, но в отличие от классической системы M/M/1/2 здесь интенсивности "рождений" зависят от текущего состояния системы — графы интенсивностей переходов для обоих случаев представлены на рис. 2. Благодаря нормировке $\lambda = 1$, интенсивность обслуживания ("гибели" заявки) обратна приведённой интенсивности входящего потока ρ.

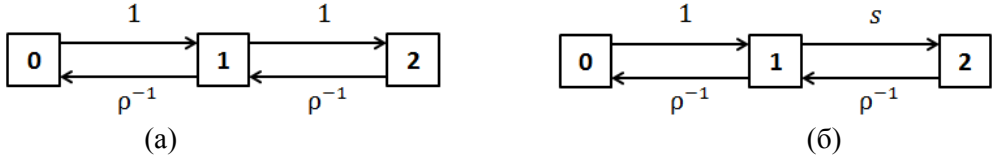


Рис. 2. Графы интенсивностей переходов между состояниями СМО в случаях (а) $\rho \le \theta$ и (б) $\rho > \theta$

Пользуясь выражениями для стационарного распределения процессов размножения и гибели, получаем:

$$p_1 = p_0\rho, p_2 = p_1 s\rho = p_0 s\rho^2.$$

Найдём p_0 из условия $p_0 + p_1 + p_2 = 1$:

$$p_0(1+\rho+s\rho^2) = 1,$$

$$p_0 = \frac{1}{1 + \rho + s\rho^2} \tag{2}$$

Отсюда получаем стационарные вероятности остальных состояний:

$$p_1 = \frac{\rho}{1 + \rho + s\rho^2}, \tag{3}$$

$$p_2 = \frac{s\rho^2}{1 + \rho + s\rho^2}. \tag{4}$$

Ожидаемая прибыль

Прибыль приносят только те заявки, которые входят в систему — не теряются из-за ограничения ёмкости или нетерпения. Как следствие, заявки первого типа приносят прибыль тогда, когда по прибытии застают систему без очереди (0 или 1 заявка в системе). По свойству PASTA простейшего потока (Poisson Arrivals See Time Averages — см., например [7]) это происходит с вероятностью $\boldsymbol{p_0 + p_1}$, так что среднее число не потерянных заявок первого типа за единицу времени составляет $\boldsymbol{\lambda s(p_0 + p_1)}$ и совпадает с ожидаемой прибылью за единицу времени от заявок первого типа. При $\boldsymbol{\lambda = 1}$ получаем выражение для этой прибыли:

$$\pi_1 = s(p_0 + p_1).$$

Индекс “1” здесь соответствует типу заявок.

Заявки второго типа обязательно приносят прибыль, если поступают в пустую систему (что происходит с вероятностью $\boldsymbol{p_0}$), и не приносят, если система заполнена (вероятность $\boldsymbol{p_2}$). Если канал обслуживания занят, но очереди нет, то заявка второго типа создаст очередь и принесёт прибыль при $\boldsymbol{\rho \le \theta}$ и покинет систему при $\boldsymbol{\rho > \theta}$. Учитывая, что каждая такая заявка приносит прибыль $\boldsymbol{r}$, получаем выражение для ожидаемой прибыли от заявок второго типа за единицу времени:

$$\pi_2 = \begin{cases} r(1-s)(p_0 + p_1), & \theta \ge \rho; \\ r(1-s)p_0, & \theta < \rho. \end{cases}$$

Чтобы избежать ошибок при истолковании формул для ожидаемой прибыли, надо помнить, что вероятности $\boldsymbol{p_0}$, $\boldsymbol{p_1}$, $\boldsymbol{p_2}$ и сами зависят от соотношения “предела терпения” $\boldsymbol{\theta}$ и среднего времени обслуживания $\boldsymbol{\rho}$.

Случай 1: $\boldsymbol{\rho \le \theta}$. Распределение числа заявок задаётся формулой (1), так что выражения для ожидаемой прибыли приобретают вид:

$$\pi_1 = s(p_0 + p_1) = \frac{s(1+\rho)}{1 + \rho + \rho^2},$$

$$\pi_2 = r(1-s)(p_0 + p_1) = \frac{r(1-s)(1+\rho)}{1 + \rho + \rho^2}.$$

Случай 2: $\boldsymbol{\rho > \theta}$. Распределение числа заявок задаётся формулами (2)–(4). Выражения для ожидаемой прибыли:

$$\pi_1 = s(p_0 + p_1) = \frac{s(1+\rho)}{1+\rho+s\rho^2},$$
$$\pi_2 = r(1-s)p_0 = \frac{r(1-s)}{1+\rho+s\rho^2}.$$

Общая прибыль за единицу времени естественно получается сложением прибылей от заявок первого и второго типа. Будем использовать для неё два обозначения: π_P для случая $\boldsymbol{\rho \le \theta}$ (индекс P от слова “patience” — терпение, так как в этом случае все заявки терпеливо ждут в очереди) и $\boldsymbol{\pi_{IP}}$ для случая $\boldsymbol{\rho > \theta}$ (ImPatience — “нетерпение”). В обоих случаях суммарная прибыль выражается с помощью ранее выведенных формул следующим образом:

$$\pi_P = \frac{(1+\rho)(s+r-rs)}{1+\rho+\rho^2}, \quad (5)$$
$$\pi_{IP} = \frac{s(1+\rho)+r(1-s)}{1+\rho+s\rho^2}. \quad (6)$$

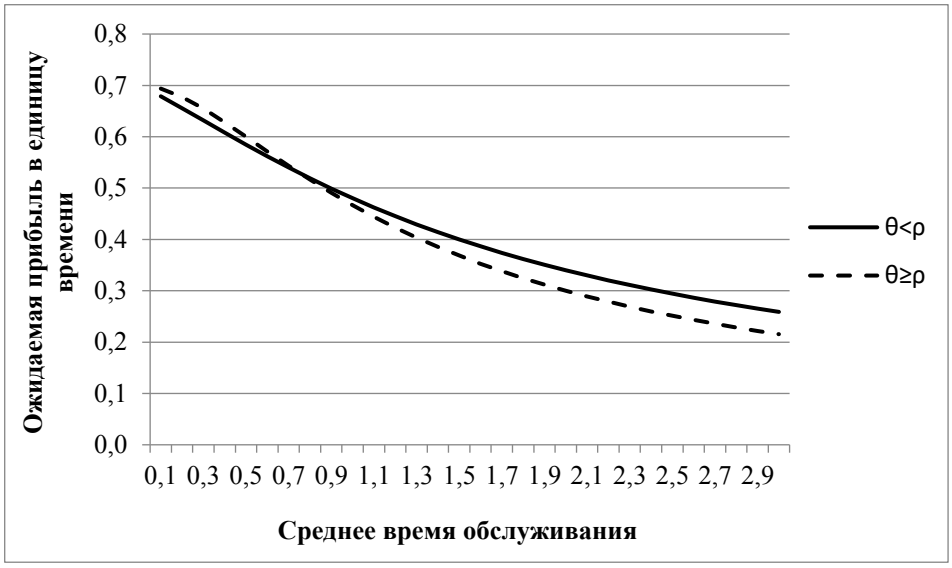
Как отпугнуть клиента: прибыль и время обслуживания

Из формул (5) и (6) следует, что величины $\boldsymbol{\pi_P}$ и $\boldsymbol{\pi_{IP}}$ убывают с ростом $\boldsymbol{\rho}$ (для наглядности соответствующие графики приведены на рис. 3, использованные значения прочих параметров: $\boldsymbol{r = 0{,}4, p = 0{,}5}$). Таким образом, ожидаемая прибыль отрицательно зависит от времени обслуживания на участках $\boldsymbol{\rho \in (0;\theta]}$ и $\boldsymbol{\rho \in (\theta;+\infty)}$. Это естественно: увеличение времени обслуживания приводит к постоянной загруженности системы и росту доли потерянных заявок. Интересно другое: как видно из рис. 3, начиная со средней длительности обслуживания $\boldsymbol{\rho \approx 0.8}$ ожидаемая прибыль в “нетерпеливом” случае $\boldsymbol{\rho > \theta}$ оказывается выше. Таким образом, если предел ожидаемого времени в очереди для заявок второго типа превышает 0,8, а среднее время обслуживания немного меньше этого предела, то администрация системы может увеличить прибыль, замедлив обслуживание.

Действительно, хотя ожидаемая прибыль убывает на множествах $\boldsymbol{\rho \in (0;\theta]}$ и $\boldsymbol{\rho \in (\theta;+\infty)}$, в точке $\boldsymbol{\rho = \theta}$ происходит скачок, который может как понизить, так и повысить прибыль (система переключается с прерывистой линии, изображённой на рис. 3, на сплошную). Этот скачок изображён на рис. 4, где для примера взято предельное время в очереди $\boldsymbol{\theta = 2}$ (прочие параметры — как и для рис. 3).

Когда среднее время обслуживания становится настолько велико, что заявки второго типа отказываются стоять в очереди, очередь освобождается для более прибыльных заявок первого типа, что может увеличить прибыль, если только заявки первого типа будут поступать до-

статочно интенсивно, чтобы компенсировать потерю нетерпеливых заявок.



Источник: расчёты авторов по формулам (5), (6).

Рис. 3. Ожидаемая прибыль за единицу времени в случаях $\boldsymbol{\theta \geq \rho}$ (прерывистая линия) и $\boldsymbol{\theta < \rho}$ (сплошная линия) в зависимости от среднего времени обслуживания.

Источник: расчёты авторов по формулам (5), (6).

Рис. 4. Скачок прибыли при увеличении времени обслуживания

Обобщение для произвольной ёмкости системы

Введённое ранее фиксированное ограничение на длину очереди (не более одной заявки) удобно для рассмотрения примера фильтрации, но не обязательно. В настоящем разделе приведены более общие

результаты — основные формулы для системы M/M/1/K в нотации Кендалла с разнородными заявками.

Такая система может вместить K заявок, из которых одна будет обслуживаться, а остальные ждать в очереди. Пусть $\boldsymbol{n}$ — наибольшее число заявок в системе, при котором нетерпеливые заявки присоединяются к очереди. Из ранее сделанных предпосылок следует, что $n = \min\left(K-1, \left[\frac{\theta}{\rho}\right]\right)$, где $\left[\frac{\theta}{\rho}\right]$ — целая часть от деления.

Динамика системы описывается процессом размножения и гибели с состояниями $0, \dots, K$, где интенсивности "гибели" (обслуживания) постоянны и равны $\boldsymbol{\rho^{-1}}$, а интенсивности "рождений" (поступления заявок) зависят от текущего состояния. Пока в системе не более $\boldsymbol{n}$ заявок, новые клиенты присоединяются с интенсивностью 1. Когда в системе более $\boldsymbol{n}$ заявок, присоединяются к очереди только терпеливые клиенты I типа, так что интенсивность притока новых заявок равна $\boldsymbol{s}$. Как следствие, стационарные вероятности состояний связаны друг с другом формулами

$$p_j = p_{j-1}\rho = p_0\rho^j, j \le n+1,$$
$$p_j = p_{j-1}\rho s = p_n(\rho s)^{j-n-1} = p_0\rho^j s^{j-n-1}, n+1 < j \le K.$$

В сумме вероятности всех состояний должны равняться единице, что позволяет выразить вероятность простоя системы $\boldsymbol{p_0}$:

$$p_0 = \left(1 + \sum_{j=1}^{n+1}\rho^j + \sum_{j=n+2}^{K}\rho^j s^{j-n-1}\right)^{-1}.$$

Следовательно, вероятность произвольного состояния $\boldsymbol{j}$ можно найти по следующей формуле:

$$p_j = \begin{cases} \dfrac{\rho^j}{1 + \sum_{j=1}^{n+1}\rho^j + \sum_{j=n+2}^{K}\rho^j s^{j-n-1}}, & j \le n+1, \\ \dfrac{\rho^j s^{j-n-1}}{1 + \sum_{j=1}^{n+1}\rho^j + \sum_{j=n+2}^{K}\rho^j s^{j-n-1}}, & n+1 < j \le K. \end{cases} \quad (7)$$

Заметим, что в ранее рассмотренном случае $K = 2$ формула (7) сводится к формуле (1) при $n = 1$ (т.е. $\boldsymbol{\rho} \le \theta$) и к формулам (2)–(4) при $\boldsymbol{n} = 0$ ($\rho > \theta$).

Суммы в формуле (7) имеют разное выражение в зависимости от того, являются ли слагаемые постоянными или образуют геометрическую прогрессию:

$$1 + \sum_{j=1}^{n+1}\rho^j = \begin{cases} n+2, & \rho = 1, \\ \dfrac{1-\rho^{n+2}}{1-\rho}, & \rho \ne 1. \end{cases}$$

$$\sum_{j=n+2}^{K} \rho^j s^{j-n-1} = \begin{cases} (K-n-1)\rho^{n+1}, & \rho s = 1, \\ \dfrac{\rho^{n+2}(1-(\rho s)^K)}{1-\rho s}, & \rho s \neq 1. \end{cases}$$

Наконец, получим выражение для ожидаемой прибыли в единицу времени. Заявки первого типа приносят прибыль, если система не заполнена (в ней меньше $\boldsymbol{K}$ заявок). Они поступают с интенсивностью $\boldsymbol{s}$ заявок в единицу времени, и каждая приносит единичную прибыль, так что ожидаемая прибыл за единицу времени находится следующим образом:

$$\pi_1 = s \sum_{j=0}^{K-1} p_j = s(1-p_K).$$

Заявки второго типа поступают с ринтенсивностью $1-\boldsymbol{s}$, приносят $\boldsymbol{r}$ единиц прибыли и входят в систему, если в ней не более $\boldsymbol{n}$ заявок, так что

$$\pi_2 = (1-s)r \sum_{j=0}^{n} p_j.$$

Выражение для суммарной ожидаемой прибыли за единицу времени:

$$\pi = \pi_1 + \pi_2 = s(1-p_K) + (1-s)r \textstyle\sum_{j=0}^{n} p_j. \qquad (8)$$

Рис. 5. Скачки прибыли в системе M/M/1/6

На рис. 5 представлена зависимость ожидаемой прибыли $\boldsymbol{\pi}$ от времени обслуживания $\boldsymbol{\rho}$ для случая $\boldsymbol{K} = 6, \boldsymbol{r} = 0.2, \boldsymbol{s} = 0.5, \boldsymbol{\theta} = 11$. Обратим внимание, что скачков в этом случае несколько. При увеличении времени обслуживания сначала “отпугиваются” те клиенты II типа, которые застают систему почти заполненной (пять заявок из мак-

симальных шести), затем к системе перестают присоединяться те нетерпеливые клиенты, кто застаёт четыре заявки и так далее — в каждом случае прибыль совершает скачок.

Заключение

Как следует из примеров, изображённых на рис. 4 и 5, администрация системы массового обслуживания может увеличить прибыль за счёт фильтрации относительно неприбыльных заявок при увеличении времени обслуживания. Отметим характеристики системы, которые делают фильтрацию возможной.

1. Относительно невыгодные заявки должны быть нетерпеливы.

2. Если прибыль, которую приносят заявки, положительна, то у системы должна быть ограниченная ёмкость — иначе все заявки первого типа будут обслуживаться вне зависимости от того, соглашаются ли заявки второго типа стоять в очереди. Ограничение ёмкости можно заменить нетерпеливостью заявок первого типа, но тогда они должны быть терпеливее невыгодных заявок второго типа.

3. Если прибыль, которую приносят нетерпеливые заявки, отрицательна, то выгодная фильтрация возможна и в системах с неограниченной очередью — “отпугивание” заявок второго типа само по себе увеличивает прибыль, при этом терпеливые заявки первого типа не будут теряться.

Мы рассматривали и другой вариант нетерпения — reneging, уход из очереди заявок, не дождавшихся обслуживания [8, 9], в отечественной литературе известность получил частный случай reneging, при котором каждая заявка имеет ограничение на время пребывания в очереди [10, 11]. При однородном входящем потоке заявок вывести стационарное распределение и ожидаемую прибыль легко, для разнородных заявок дело усложняется тем, что вероятность ухода заявки из очереди начинает зависеть не только от длины очереди, но и от её состава (числа нетерпеливых заявок). Мы не выводили аналитического решения, но имитационным моделированием обнаружили скачки прибыли, аналогичные приведённым на рис. 4.

На практике определить удачный момент для увеличения времени обслуживания вряд ли возможно, но на предлагаемую модель можно посмотреть с иной стороны: она показывает, почему ускорение обслуживания может иметь эффект ниже желаемого и даже приводить к снижению прибыли.

Список источников

1. Stiglitz J.E. The Theory of “Screening”, Education, and the Distribution of Income // The American Economic Review. 1957. Vol. 65. No. 3. P. 283–300.
2. Аистов А.В. О фильтрующей роли образования в России // Экономический журнал Высшей школы экономики. 2009. Т. 13. № 3. С. 452–481.
3. Grubb W.N. Further tests of screening on education and observed ability // Economics of Education Review. 1993. Vol. 13. No. 2. P. 125–136.

4. Haight F.A. Queueing with balking // Biometrika. 1957. Vol. 44. No. 3. P. 360–369.
5. Liu L.Q. Service systems with balking based on queueing time // PhD Thesis. University of North Carolina at Chapel Hill. 2007. — URL: https://research.tue.nl/en/publications/service-systems-with-balking-based-on-queueing-time
6. Shortle J.F., Thompson J.M., Gross D. [et al.]. Fundamentals of Queueing Theory. — 5th ed. — Wiley, 2018.
7. Ross S.M. Introduction to Probability Models. — 10th ed. — Elsevier, 2010.
8. Pazgal A.I., Radas S. Comparison of customer balking and reneging behavior to queueing theory predictions: An experimental study // Computers & Operations Research. 2008. Vol. 35. No. 8. P. 2537–2548.
9. Choudhury A., Medhi P. Balking and reneging in multiserver Markovian queueing system // International Journal of Mathematics in Operational Research. 2011. Vol. 3. No. 4. P. 377–394.
10. Гнеденко Б.В., Коваленко И.Н. Введение в теорию массового обслуживания. — М.: Наука, 1966. — 432 с.
11. Кирпичников А.П. Методы прикладной теории массового обслуживания. — М.: ЛЕНАНД, 2018. — 224 с.

Сведения об авторах

Сурова София Валерьевна, студент факультета компьютерных наук, Национальный исследовательский университет “Высшая школа экономики”. 101000, Россия, г. Москва, ул. Мясницкая, 20. E-mail: *svsurova@edu.hse.ru*.

Sofia V. Surova, student, Faculty of Computer Science, National Research University Higher School of Economics. 20, Myasnitskaya St., Moscow, 101000, Russia. E-mail: *svsurova@edu.hse.ru*.

Фурманов Кирилл Константинович, кандидат экономических наук, старший научный сотрудник отделения Эконометрики и прикладной статистики, Центральный экономико-математический институт РАН. 117418, Россия, г. Москва, Нахимовский проспект 47. E- mail: *kfurmanov@hse.ru*.

Kirill K. Furmanov, PhD in Economics, Senior Researcher, Department of Econometrics and Applied Statistics, Central Economic and Mathematical Institute of the Russian Academy of Sciences. 47, Nakhimovsky pr., Moscow, 117418, Russia. E-mail: *kfurmanov@hse.ru*.

Ссылка для цитирования

Сурова С.В., Фурманов К.К. Как отпугнуть клиента и получить прибыль: фильтрация в системе массового обслуживания // Известия Дальневосточного федерального университета. 2024. № 1. С. 80–90. — DOI: https://doi.org/10.24866/2311-2271/2024-1/1113.