Управление в социально-экономических системах

© 2024 г. Г.М. КРЮКОВ (gkryukov@nes.ru),
М.С. САНДОМИРСКАЯ, канд. физ.-мат. наук (msandomirskaya@hse.ru)
(Национальный исследовательский университет
Высшая школа экономики, Москва, Санкт-Петербург)

ИССЛЕДОВАНИЕ СТРАТЕГИЧЕСКИХ ПОСЛЕДСТВИЙ УТЕЧКИ СКОРИНГОВОЙ МОДЕЛИ[1]

В данной статье моделируется раскрытие информации о скоринговой модели. Некоторые клиенты компании узнают свой внутренний рейтинг в компании. Такие клиенты могут изменить свое поведение, чтобы повысить свой внутренний рейтинг. Клиенты, знающие об утечке информации, являются игроками, которые могут выбирать стратегию: повышать ли свой внутренний рейтинг и если да, то насколько. Главная задача – найти в этой игре равновесие Байеса–Нэша и выяснить, как оно зависит от различных параметров, таких как масштаб утечки, распределение рейтингов.

Ключевые слова: скоринговая модель, байесова игра, манипулирование, раскрытие информации.

DOI: 10.31857/S0005231024080041, **EDN:** WPGHKX

1. Введение

Компании по всему миру внедряют методы машинного обучения для решения задач бизнеса. В частности, компании нередко вычисляют с помощью методов Data Science внутренние параметры пользователя (например, вероятность вернуть кредит для банка или привлекательность для сервиса онлайн-знакомств). Скоринговая модель присваивает каждому пользователю определенный рейтинг (score). Этот рейтинг может быть как дискретной случайной величиной (например, разделение пользователей на несколько кластеров), так и непрерывной (например, нейронная сеть часто возвращает действительные значения из отрезка [0,1]). Эти модели могут быть полезны для решения самых разных бизнес-задач, таких как: “отобрать для участия в акции m пользователей с наибольшей склонностью к спонтанным покупкам” или “отобрать m пользователей из нужного компании кластера и предложить им пробную версию нового продукта”.

Проблема заключается в том, что периодически случаются утечки информации. В результате таких утечек пользователи могут узнать, какой у них

[1] Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ.

внутренний рейтинг либо к какому внутреннему кластеру они относятся. Например, в 2016 г. произошла утечка внутреннего рейтинга пользователей в приложении для знакомств Tinder [1]. Помимо утечки конечного рейтинга, пользователи могут узнать, как именно работает модель машинного обучения, а также рейтинги/распределение по кластерам других пользователей. Тогда у пользователей может появиться мотивация обмануть алгоритм: пользователи могут изменить свое поведение так, чтобы алгоритм машинного обучения улучшил им внутренний рейтинг либо отнес к более хорошему кластеру. Однако изменение поведения и иных видимых алгоритму характеристик небесплатно: пользователю нужно потратить время и иногда деньги, чтобы изменить свой рейтинг (например, посетить определенные страницы на сайте, оформить заявку на определенный продукт). Кроме того, для разных пользователей издержки на то, чтобы изменить оценку алгоритма, могут различаться. Соответственно каждый пользователь индивидуально решает задачу: менять свое поведение, чтобы улучшить внутренний рейтинг, либо оставить все как есть.

В данной работе ставится вопрос о том, как будут вести себя пользователи при утечке информации о внутреннем ранжировании пользователей от фирмы. Как их поведение зависит от значений параметров скоринговой модели? Для этого моделируется рынок, где компания хочет отобрать m клиентов из n для определенной активности (выдача кредитов, участие в выгодной акции, тестирование нового продукта и т.п.). Сделаем предположение, что все клиенты одинаково ценят эту активность и что принять в ней участие лучше, чем не принять. Часть пользователей получают информацию о том, как работает алгоритм машинного обучения компании, который ранжирует/распределяет по кластерам клиентов. Эти пользователи стоят перед стратегическим выбором. С одной стороны, они могут понести определенные издержки, чтобы увеличить свой рейтинг либо перейти в лучший кластер, тем самым увеличить свои шансы на попадание в m клиентов, которые получат определенный бонус. С другой стороны, они могут принять свой текущий внутренний рейтинг как приемлемый и ничего не менять. Таким образом, клиенты, получившие доступ к внутренней информации о моделях машинного обучения фирмы, становятся игроками, которые выбирают оптимальную для себя стратегию. Эта ситуация представляет собой байесовскую игру, и далее исследуется ее равновесие. Связь стратегических последствий и вида функции распределения издержек может быть использована для анализа устойчивости скоринговой модели к утечкам данных. Различные скоринговые модели могут включать параметры, в большей или меньшей степени сложные для манипулирования информированным клиентом, так что фирма, выбирая конкретную скоринговую модель, может заранее спрогнозировать потенциальные масштабы скоринговых ошибок при небольших или значительных утечках. Это может способствовать повышению безопасности и устойчивости применяемых скоринговых моделей.

1.1. О скоринговых моделях в литературе

Центральное место в данной работе занимает модель машинного обучения, которая производит скоринг клиентов. Модели скоринга пользователей используются в индустрии и активно развиваются. Каноничным примером являются модели кредитного скоринга, которые оценивают кредитоспособность клиентов. Для оценки кредитного рейтинга активно применяются методы науки о данных, в том числе методы интерпретируемого глубинного обучения [2] и генетические иерархические сети [3]. Авторы в [4] установили, что после внедрения продвинутых методов, указанных выше, в которых клиенты не осведомлены о детальном механизме работы скоринговой модели, они стали хуже понимать, какие шаги им следует предпринять, чтобы улучшить свой кредитный рейтинг. В этой работе предполагается, что при утечке информации о работе модели кредитного скоринга у клиентов станет больше понимания, как можно повлиять на свой кредитный рейтинг, и поэтому клиенты могут захотеть манипулировать своим рейтингом.

Отметим, что компании могут оценивать пользователей не только для кредитного скоринга. Так, продвинутые методы машинного обучения активно применяются для предсказания совершения пользователем определенного действия, например покупки определенного товара [5]. Эти данные также применяются для принятия бизнес-решений о стратегиях проведения акций и распределения персональных скидок. Поэтому информация о подобных моделях скоринга клиентов также может быть использована пользователями для получения выгоды.

Стоит отметить, что алгоритмы машинного обучения применяются также для кластеризации пользователей по категориям [6]. В этом случае пользователи также имеют стимулы манипулировать моделью при ее раскрытии, чтобы попасть в более привлекательный для себя кластер.

1.2. О теоретико-игровом контексте в модели

Для моделирования ситуации утечки информации о скоринговой модели будет использоваться теоретико-игровой подход, эффективность которого для изучения эмпирических задач отмечается, например, в [7]. Игровая составляющая возникает из-за того, что клиенты (игроки в модели) претендуют на конечное количество блага, так что повышение рейтинга отдельного клиента снижает шансы быть выбранными для остальных. Имея представление о механизме скоринговой модели, клиент способен рассчитать свой собственный рейтинг, однако рейтинг других клиентов ему неизвестен, и он может полагаться лишь на некоторое общеизвестное распределение рейтингов в популяции. Также клиенты действуют независимо и не наблюдают действия, предпринимаемые другими, так что можно считать их взаимодействие одновременным. Таким образом, будет построена байесовская игра [8].

Стоит отметить, что идея манипулирования нередко встречается в теоретико-игровых моделях, в том числе моделях оптимального управления.

Например, в [9–12] исследуются оптимальные механизмы управления в системах с активными элементами. Наиболее полный обзор работ представлен в [13]. По сути это направление исследует задачу дизайна механизмов, в которой ключевой идеей является раскрытие информации о типе агентов и адекватный учет поведения агента в целевой функции центра. С одной стороны, большинство работ так или иначе посвящено исследованию проблемы манипулирования и выделению классов предпочтений, для которых процедура управления устойчива к манипулированию. При этом под манипулированием понимается искажение агентами своих заявляемых предпочтений, и это искажение, как правило, не требует явных издержек от агентов. В данной работе внимание акцентируется на возникающих сопутствующих издержках на искажение рейтинга, что более характерно, например, для работ по манипулированию репутацией [14, 15]. Также важной особенностью этой работы является то, что в результате утечки активной становится некоторая случайная подвыборка агентов, а не все агенты, причем в этой подвыборке, вследствие гетерогенности по издержкам, найдутся и те, кто захочет манипулировать, и те, кто предпочтет оставить свои истинные характеристики. Таким образом, следует отметить, что даже при экзогенном сужении множества активных агентов проблема манипулируемости не исчезает, а ее анализ требует совместного учета поведения и активных агентов, и тех кто формально присутствует в системе с фиксированными характеристиками.

Поскольку целью скоринговой модели является отбор ограниченного множества клиентов, то при анализе будут возникать структуры, характерные для работ в области конечных рынков [16]. Одна из таких работ связана с исследованием ценовой конкуренции на рынках с редким товаром и частной информацией об оценках товара покупателями [17]. Авторы рассматривают рынок с двумя продавцами, у каждого из которых есть одна единица идентичного товара. Продавцы одновременно выбирают цены, после чего покупатели выбирают, к какому продавцу идти за товаром либо не идти ни к кому. Условия, выводимые при решении задачи покупателя, схожи с теми, что возникают в данной работе при анализе игры с единственным кандидатом, выбираемым скоринговой моделью (победителем). Если говорить о модели с несколькими победителями, то полученные здесь формулы с биномиальным коэффициентом идейно похожи на результаты в [18] об олигополии Бертрана с ограничениями на производственные мощности фирм.

2. Модель

Формализуем игровую модель, в которой пользователи, получившие информацию о скоринговом алгоритме, решают, пользоваться ли полученной информацией для увеличения своего рейтинга.

Пусть имеется фирма, которая производит определенный продукт (например, банк, выдающий кредитные карты). Пусть есть n потребителей, которые претендуют на продукт (n клиентов, которые хотят получить кредитные

карты). При этом у фирмы есть ограничение (например, на количество пластика для карт), поэтому они готовы продать только $m < n$ единиц товара. Предположим, что все покупатели ценят товар одинаково. Обладание товаром приносит единичную полезность.

Фирма решает использовать некоторую скоринговую модель для классификации своих пользователей. В этой статье предполагается, что скоринговая модель классифицирует потребителей только на две категории: хороший (единица) и плохой (ноль). Механизм работы скоринговой модели основан на некоторых методах машинного обучения и по сути представляет некоторый "черный ящик" ввиду сложности своей работы, однако фирма знает входные параметры этого ящика, при необходимости может оценить их вес в итоговом результате классификации прогонкой модели на достаточно широкой выборке своих клиентов, характеристики которых известны фирме. Также по этой выборке фирма может аппроксимировать распределение издержек клиентов на то, чтобы поменять какие-то из своих индивидуальных параметров, важных для скоринговой модели, на значения, достаточные чтобы быть классифицированными как "хорошие".

В игре рассматривается ситуация, когда произошла утечка информации о скоринговой модели. Пользователь, получивший доступ к этой информации, узнает, к какому классу его относит модель и какие параметры она использует. Также пользователь, получивший доступ к информации, узнает, какое распределение имеет рейтинг на выборке из всех клиентов (клиентов много, рейтинг определен для всех, а не только для n, которые хотят получить кредитную карту или другую награду). Однако пользователь достоверно не знает, какими характеристиками обладают другие пользователи, знающие об утечке, т.е. стратегический пользователь принимает дальнейшие решения в условиях неполной информации.

Пусть потребитель согласно скоринговому алгоритму является хорошим с вероятностью p. Клиент с вероятностью α нашел необходимые данные. Если покупатель узнает, что модель относит его к хорошему типу, то у него нет стимулов менять свое привычное поведение. Другими словами, изменение поведения таким образом, чтобы алгоритм классифицировал его как плохого, является слабо доминируемой стратегией и поэтому не будет использоваться в дальнейшем анализе.

Если потребитель узнает, что алгоритм относит его к плохому типу, у него появляется стратегия изменить свое поведение, понести издержки c_i и мимикрировать под хорошего. Полагаем, что c_i могут быть различны для разных игроков в рамках носителя функции распределения, каждый игрок знает свое значение издержек c_i. Действительно, кому-то надо сделать совсем немного, чтобы выполнить условия алгоритма для "хорошего потребителя", а кому-то надо сильно изменить свое поведение и соответственно понести большие издержки. Пусть функция распределения издержек на "мимикрирование" пользователей, которых модель относит к плохому типу, равна $F(x) = P[c_i \leqslant x]$;

для согласованности с нормировкой полезности к 1 предположим также, что функция распределения издержек определена на $[0; 1]$. Если бы какие-то потребители имели издержки на мимикрирование выше, чем полезность от обладания продуктом фирмы, то они гарантированно не станут мимикрировать и их можно исключить из рассмотрения. Предположим, что $F(x)$ известна всем стратегическим игрокам. Уместно считать величину издержек относительной величиной – долей полезности от обладания продуктом. Тогда все дальнейшие расчеты в модели линейно масштабируются на необходимый размер "приза". В качестве возможного примера параметра скоринговой модели, издержки на изменение которого высоки для большинства клиентов, можем рассмотреть наличие и размер ипотечного кредита у заемщика (на данном примере легко видеть, что издержки могут и превышать размер "приза", так что такой параметр в модели вряд ли будет подвержен искажению). Напротив, пример параметра, искажение которого малозатратно, – наличие полностью заполненного открытого профиля в социальной сети.

Отметим, что для поставленных здесь целей не важно детальное знание внутреннего механизма скоринга, а достаточно знать только вероятность p и функцию $F(x)$, которые относятся как к скоринговому механизму, так и к характеристикам популяции пользователей, на которой его планируется применять.

С вероятностью $1 - \alpha$ покупатель ничего не знает об утечке. Он не является стратегическим игроком, утечка информации не влияет на его внутренний рейтинг и поведение. Таким образом, стратегическими игроками в данной игре являются только агенты, знающие об утечке, которые относятся скоринговой моделью к плохому классу. При этом байесовым типом игрока i являются его издержки c_i.

Выигрыш стратегического агента при выборе стратегии "НЕ мимикрировать" равен вероятности получить товар, если скоринговая модель классифицирует игрока как плохого, т.е. должен допускать ситуацию, что товара больше, чем хороших, определяемых алгоритмом, так что придется случайно выбрать оставшихся из категории плохих. Выигрыш стратегического агента при выборе стратегии "мимикрировать" равен разности вероятности получить товар, если модель классифицирует игрока как хорошего, и издержек на мимикрирование. Отметим, что стратегии выбираются игроками одновременно и независимо.

Осталось обсудить, как фирма определяет ровно m победителей. Пусть k агентов классифицированы скоринговой моделью как хорошие (это могут быть как изначально хорошие, так и стратегические, которые решили изменить данные о себе и обмануть модель). Если $k > m$, то победителями становятся случайные m из k хороших по модели клиентов. Если $k \leqslant m$, то победителями становятся все хорошие по модели клиенты, а также из плохих дополнительно случайно выбирается $m - k$ победителей.

Таким образом, одновременная игра состоит в том, что каждый клиент из узнавших про утечку и классифицированных плохими делает бинарный выбор и решает, готов ли он вкладываться в соответствии со своими издержками в улучшение рейтинга или оставить все как есть. Когда "плохой", в соответствии с исходной работой скорингового алгоритма, клиент решает мимикрировать под "хорошего", это искажает результаты классификации модели и может приводить к неправильному выбору фирмы победителя. Задачей является поиск равновесных стратегий для агентов, узнавших об утечке, т.е. определение для каждого клиента оптимального выбора стратегии "мимикрировать" или нет в зависимости от числа "призов", избирательности скоринговой модели p, используемых ею параметров и, как следствие, распределения издержек на мимикрирование этими параметрами в популяции, масштаба утечки, а также от величины собственных издержек на улучшение своего типа, которые клиент знает с точностью. Этот выбор неочевиден, поскольку отдельный клиент не может полагаться исключительно на собственную стратегическую активность, а должен корректно предсказать изменение поведения и других активных агентов, которое повлияет на итоговую вероятность быть выбранным. При этом попадание в число победителей носит вероятностный характер, а издержки на мимикрирование детерминированы и невозвратны.

Исследовав равновесную стратегию всех клиентов, фирма сможет спрогнозировать, какая доля информированных клиентов решит мимикрировать и, следовательно, как это исказит результаты скоринга фирмы. Поскольку скоринговая модель может включать различные параметры, стоимость манипулирования которыми может быть различна для пользователя, то при одних и тех же масштабах утечек данных в разных моделях может наблюдаться большая и меньшая степень искажения. Соответственно, на этапе выбора скоринговой модели фирма может учесть дальнейшие риски от подобных утечек.

3. Равновесие Нэша в игре с одним победителем

Рассмотрим частный случай модели при $m = 1$. Например, компания может выбирать одного клиента, которого хотят сделать лицом нового продукта. Часть клиентов (плохие согласно алгоритму) ей точно не подойдут, а среди хороших можно брать любого, так как все подходят в достаточной степени.

Ищем симметричное равновесие Байеса–Нэша. Обозначим через y вероятность того, что плохой будет мимикрировать под хорошего в равновесии. Отметим, что данная вероятность в дальнейшем будет складываться из того, что агенты некоторых типов издержек будут детерминированно мимикрировать, а других типов – не будут, однако поскольку отдельный игрок не знает истинный тип других агентов (для него это случайная величина), то их поведение также будет выглядеть как случайное, вероятностное. Также определим q как апостериорную вероятность того, что игрок будет классифициро-

ван алгоритмом как хороший тип с учетом стратегического мимикрирования доли потребителей, узнавших об утечке,

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

Если игрок не мимикрирует, то у него есть шанс выиграть только в случае, когда все остальные также окажутся плохими *ex-post*. Ожидаемый выигрыш от стратегии "не мимикрировать":

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

Выпишем формулу ожидаемого выигрыша при выборе стратегии "мимикрировать". Независимо от исхода игроку придется заплатить c_i. В случае, если ровно k конкурентов будут классифицированы как хорошие, вероятность победы равна $\frac{1}{k+1}$. Случайная величина "число конкурентов, которые будут классифицированы как хорошие", имеет биномиальное распределение: $\mathrm{Bin}(n-1, q)$. Итого ожидаемый выигрыш от стратегии "мимикрировать", выражается формулой

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

Воспользуемся следующим тождеством, доказанным в [17]:

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

Ожидаемый выигрыш от стратегии "мимикрировать" переписывается в следующем виде:

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

Игрок выберет ту из двух стратегий, которая ему принесет наибольший ожидаемый выигрыш. Заметим, что ожидаемый выигрыш от стратегии "не мимикрировать" не зависит от c_i, в то время как ожидаемый выигрыш от стратегии "мимикрировать" уменьшается с ростом c_i. Значит, оптимальная стратегия игрока i является монотонной (пороговой). Иными словами, существует такое пороговое значение издержек c^*, что для всех $c_i < c^*$ игрок i мимикрирует и для всех $c_i > c^*$ i-й игрок НЕ мимикрирует. При $c = c^*$ ожидаемые выигрыши от обеих стратегий одинаковы.

Тогда вероятность того, что плохой будет мимикрировать под хорошего в равновесии, равна вероятности того, что издержки окажутся ниже порогового уровня, и выражается через c^* как

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

Условие для c^* определяется из равенства ожидаемых выигрышей от двух чистых стратегий:

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

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

Важно уточнить, что эта формула является выражением в неявном виде, ведь q зависит от y, который однозначно определен через c^*:

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

Введем функцию $f(q)$:

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

Тогда условие на c^* записывается в виде $c^* = f(q(c^*))$.

Теорема 1. Для непрерывной функции распределения оценок $F(c)$ пороговое значение c^* в оптимальной монотонной стратегии существует и единственно.

Доказательство. Сначала заметим, что $f(q)$ монотонно убывает с ростом q при $n > 2$ и постоянно при $n = 2$. Докажем это. Рассмотрим частную производную

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

Знаменатель больше нуля. Рассмотрим числитель. Заметим, что при $n = 2$ числитель равен нулю.

Выполняется следующее неравенство:

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

Действительно,

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

Из этого утверждения следует монотонность: числитель $-1 + (1-q)^{n-2}\times$ $\times((n-2)q+1)$ убывает по n, при этом равен 0 при $n = 2$. Отсюда следует, что $-1+(1-q)^{n-2}((n-2)q+1) < 0$ при $n > 2$, а отсюда следует, что частная производная отрицательная.

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

Теперь докажем, что $\frac{1}{n} \leqslant c^* \leqslant \frac{n-1}{n}$. Для этого достаточно показать, что $\frac{1}{n} \leqslant f(q) \leqslant \frac{n-1}{n}$ при $q \in (0,1)$. В силу монотонности достаточно посчитать значения функции в точках $q = 0$ и $q = 1$.

Для подсчета предела в точке $q = 0$ воспользуемся правилом Лопиталя.

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

Также вычислим значение в точке $q = 1$: $f(1) = \frac{1}{n}$.

Тогда и функция $f(q)$, и равновесное значение c^* лежат в искомом интервале.

Наконец осталось рассмотреть следующую функцию:

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

Решение уравнения $g(c) = 0$ будет являться решением исходного уравнения на c^*. Из свойств функции f следует, что $g(c)$ монотонно возрастает по c. Также $g(c)$ непрерывна.

На краях интервала имеем $g(0) \leqslant -\frac{n-1}{n} < 0$ и $g(1) \geqslant 1 - \frac{1}{n} > 0$. Тогда по теореме о промежуточном значении следует, что найдется такая точка, где $g = 0$. Теорема доказана.

Анализ результатов игры с одним победителем

Изучим найденное равновесие в игре с $m = 1$.

Следствие 1. В игре с двумя игроками для любой функции распределения издержек $F(c)$ имеем $c^* = \frac{1}{2}$.

Действительно,

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

Получается, что при $n = 2$ принятие решения игроком не зависит от доли хороших типов, доли стратегических игроков и распределения издержек на мимикрирование.

При доказательстве теоремы 1 попутно было доказано

Следствие 2. Для любых значений параметров модели $\frac{1}{n} \leqslant c^* \leqslant \frac{n-1}{n}$.

Из этого следует, что при любых значениях параметров модели найдутся типы агентов, которым оптимально мимикрировать и не мимикрировать. Если издержки маленькие $c_i < \frac{1}{n}$, то игроку оптимально мимикрировать независимо, например, от масштаба утечки или функции распределения издержек других игроков. И наоборот, при высоких издержках $c_i > \frac{n-1}{n}$ игрок никогда не мимикрирует.

Утверждение 1. Пороговое значение c^* в равновесии монотонно убывает с ростом p, α при $n > 2$. Также c^* имеет монотонную зависимость от параметра распределения λ, если функция распределения F монотонно зависит от λ.

Доказательство. Поскольку c^* не выражено аналитически, то будем пользоваться теоремой о неявной функции. Рассмотрим уравнение связки:

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

Для анализа зависимости порогового значения от доли p определим знак производной:

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

Числитель:

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

так как ранее было показано, что $\frac{\partial f}{\partial q} < 0$ при $n > 2$, а $\frac{\partial q}{\partial p} = 1 - \alpha F(c^*) > 0$.

Знаменатель:

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

так как $\frac{\partial f}{\partial q} < 0$ при $n > 2$, и $\frac{\partial F}{\partial c^*} \geqslant 0$ по определению функции распределения, так что $\frac{\partial q}{\partial c^*} = (1-p)\alpha\frac{\partial F}{\partial c^*} \geqslant 0$.

Тогда верно, что $\frac{\partial c^*}{\partial p} < 0$.

Данный результат имеет естественную интерпретацию: чем вероятнее появление действительно сильных по модели агентов, тем сложнее с ними конкурировать слабым, хоть и стратегическим игрокам, а значит, меньше плохих агентов будет пытаться мимикрировать.

Аналогично доказываем зависимость от α при $n > 2$. Получаем

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

в силу того, что

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

Этот результат обусловлен похожей логикой стратегического поведения агентов: чем вероятнее появление слабого агента, знающего про утечку информации, тем вероятнее и с ними тоже придется конкурировать, а значит,

меньше слабых согласно исходной скоринговой модели агентов будут пытаться мимикрировать.

Наконец, рассмотрим параметрическое семейство функций распределения $F(c, \lambda)$. Определим возможный характер зависимости от параметра λ, опять используя теорему о неявной функции:

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

Преобразуем числитель

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

Отсюда следует, что $\operatorname{sgn}(\frac{\partial g}{\partial \lambda}) = \operatorname{sgn}(\frac{\partial F}{\partial \lambda})$, а значит

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

В частности, если F монотонно возрастает (убывает) с ростом параметра λ, то пороговое значение c^* монотонно убывает (возрастает) с ростом параметра λ. Так, рост параметра λ в функции распределения издержек означает, что в обществе снижаются издержки на мимикрирование (например, новая скоринговая модель использует более очевидные параметры, поддающиеся более простому искажению), и это приводит к уменьшению порогового значения, при котором игрок перестает пытаться мимикрировать.

Утверждение 1 доказано.

Стало понятно, как изменяется пороговое значение с изменением параметра λ, и данный результат не кажется интуитивным до тех пор, пока не будет выяснено, как меняется при этом доля тех стратегических агентов, кто решит мимикрировать. Это величина соответствует значению $F(c^*)$. Имеем функцию распределения $F(c, \lambda)$. Обозначим через $F^{(1,0)}$, $F^{(0,1)}$ производные F по первой и второй переменной соответственно. По определению функции распределения $F^{(1,0)} \geqslant 0$. Пусть функция распределения монотонно зависит от параметра λ, т.е. знак $F^{(0,1)}$ одинаковый для всех значений параметра. Исследуем знак $\frac{\partial F(c^*(\lambda),\lambda)}{\partial \lambda}$. Преобразуем

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

что дает

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

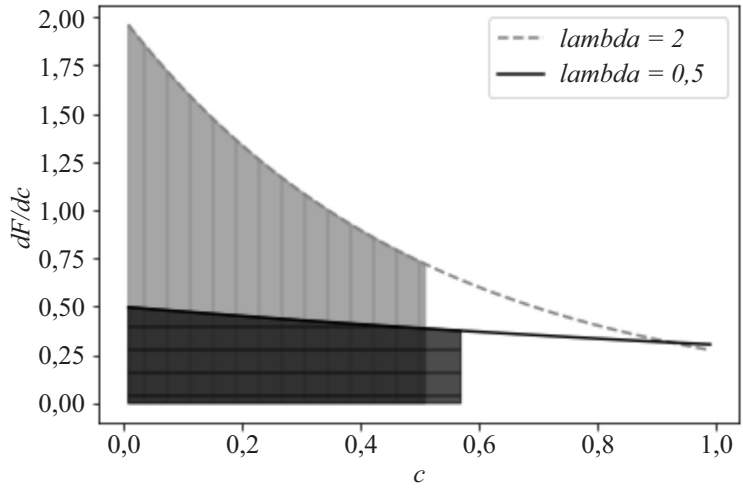

Рис. 1. Равновесные пороговое значение и доля мимикрирующих при экспоненциальном распределении издержек.

Выше было показано, что $\frac{\partial f}{\partial q} < 0$, $\frac{\partial q}{\partial F} > 0$, $F^{(1,0)} \geqslant 0$ при $n > 2$. Также легко видеть, что $\frac{x}{1-x} > -1$ при $x \leqslant 0$. Из этих утверждений следует, что выражение в больших скобках положительно. Отсюда заключаем:

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

Рассмотрим этот эффект подробнее на примере экспоненциального распределения. Зафиксируем $n = 3$, $m = 1$, $p = 0{,}2$, $\alpha = 0{,}5$, $F(c,\lambda) = 1 - \exp(-\lambda_i c)$, где $\lambda_1 = 0{,}5$, $\lambda_2 = 2$.

Численно решим уравнение $c^* - f(q(c^*)) = 0$ и найдем c^* с точностью до четырех знаков после запятой.

$$\begin{gathered} c^*(\lambda_1) = 0{,}5671, \\ c^*(\lambda_2) = 0{,}5143, \\ F(c^*(\lambda_1),\lambda_1) = 0{,}2469, \\ F(c^*(\lambda_2),\lambda_2) = 0{,}6425. \end{gathered}$$

В рассмотренном примере увеличение λ повлекло за собой небольшое уменьшение порогового значения, зато значительно увеличило долю мимикрирующих.

На рис. 1 приведен график с плотностями распределения при $\lambda_1 = 0{,}5$, $\lambda_2 = 2$ и значениями доли мимикрирующих в равновесии как площадь под графиками при $c \in [0, c^*]$.

Подведем итоги. Пусть распределение издержек изменилось таким образом, что $F(c)$ стало больше при всех значениях c. Это равносильно упрощению скоринговой модели, повышению ее уязвимости. Тогда пороговое значение в равновесии c^* уменьшится, однако не настолько сильно: вероятность

того, что потребитель будет мимикрировать, увеличится, как и доля мимикрирующих. Кроме того, как было выяснено ранее, c^* уменьшается с ростом p и α. Так как эти параметры не влияют на распределение, то вероятность того, что потребитель будет мимикрировать, также уменьшается с ростом p и α.

4. Равновесие Нэша в игре с m победителями

Пусть теперь $m \leqslant n$ хороших согласно скоринговой модели игроков получают доступ к продукту. Будем искать симметричное равновесие Байеса–Нэша. Пусть y – вероятность того, что плохой будет мимикрировать под хорошего в равновесии. Аналогично случаю с одним победителем q как вероятность того, что игрок будет классифицирован алгоритмом как хороший тип, вычисляется по формуле полной вероятности:

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

Найдем ожидаемые выигрыши i-го игрока, который является стратегическим агентом, от стратегий "мимикрировать" и "не мимикрировать". Пусть k конкурентов классифицированы как хороший тип. Как и в случае с одним победителем, случайная величина "число конкурентов, которые будут классифицированы как хорошие", имеет биномиальное распределение: $\mathrm{Bin}(n-1, q)$. Осталось посчитать ожидаемые выигрыши i-го игрока.

Если k конкурентов классифицированы как хороший тип и i-й игрок выбирает стратегию "не мимикрировать", то при $m \leqslant k$ все места будут заняты игроками, которые будут классифицированы как хорошие. Если $m > k$, то остается $m - k$ товаров для плохих по модели игроков, вероятность получить их равна $\frac{m-k}{n-k}$.

Если k конкурентов классифицированы как хороший тип и i-й игрок выбирает стратегию "мимикрировать", то при $m > k$ игроки, которые классифицированы моделью как хорошие, гарантированно получат товар, а значит i-й игрок с единичной вероятностью получит товар. Если $m \leqslant k$, то $k + 1$ хороших по модели игроков будут конкурировать за m мест. Поэтому вероятность получить товар равна $\frac{m}{k+1}$.

Итак, ожидаемый выигрыш от стратегии "не мимикрировать":

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

Ожидаемый выигрыш от стратегии "мимикрировать":

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

Заметим, что ожидаемый выигрыш от стратегии "не мимикрировать" не зависит от c_i, в то время как ожидаемый выигрыш от стратегии "мимикрировать" уменьшается с ростом c_i. Значит, оптимальная стратегия игрока i является монотонной в том же смысле, что и в разделе 3: существует c^* такое, что для всех $c_i < c^*$ i-й игрок мимикрирует и для всех $c_i > c^*$ i-й игрок не мимикрирует. При $c = c^*$ ожидаемые выигрыши от обеих стратегий одинаковы.

Вероятность того, что плохой будет мимикрировать под хорошего в равновесии, выражается через c^*:

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

Приравнивая ожидаемые выигрыши от чистых стратегий, получаем условие на пороговое значение c^*:

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

где

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

Преобразуем полученное выражение. Если $k < m$, то

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

Если $k \geqslant m$, то

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

Тогда получаем следующее выражение для порогового значения:

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

Для удобства анализа равновесия введем дополнительную функцию

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

и перепишем условие в виде

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

Начнем с анализа зависимости равновесного порога от q. Эта зависимость отражает предельный рост равновесного порога по апостериорной вероятности мимикрирования. Технически это важный объект, поскольку влияние других параметров на равновесие в итоге зависит от знака df/dq.

Утверждение 2. В общем случае функция $\tilde{f}(m,q)$ является монотонной по q тогда и только тогда, когда $m=1$ или $m=n-1$.

Доказательство. В первую очередь заметим, что $\tilde{f}(m=1,q)=\tilde{f}(m=n-1,1-q)$.

Как было доказано в исследовании игры с одним победителем, $\tilde{f}(m=1,q)$ строго убывает с ростом q. Тогда отсюда следует, что $\tilde{f}(m=n-1,q)$ строго возрастает с ростом q.

Теперь покажем, что при $1<m<n-1$ функция $\tilde{f}(m,q)$ не является монотонной по q. Для этого вычислим пределы производной при $q\to 0$ и $q\to 1$.

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

Воспользуемся тем, что $\lim\limits_{q\to 0}q^a(1-b)^b=1$ при $a=0,\ b>0$.

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

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

Таким образом, возможны три случая:

1. $m=1$. В этом случае $\frac{\partial\tilde{f}}{\partial q}<0$.

2. $m=n-1$. В этом случае $\frac{\partial\tilde{f}}{\partial q}>0$.

3. $1 < m < n-1$. В этом случае $\lim\limits_{q\to 0}\frac{\partial\tilde{f}}{\partial q} > 0$ и $\lim\limits_{q\to 1}\frac{\partial\tilde{f}}{\partial q} < 0$. Тогда из непрерывности $\frac{\partial\tilde{f}}{\partial q}$ следует, что функция достигает максимума на $q \in [0,1]$ при $q^* \in (0,1)$.

Итак, показано, что если $1 < m < n-1$, то $\tilde{f}(m,q)$ не является монотонной по q. По-видимому, используемые для описания $\tilde{f}(m,q)$ комбинаторные суммы слишком сложны и не упрощаются в виде элементарных функций. Поэтому аналитически выразить q^*, которое максимизирует $\tilde{f}(m,q)$ при фиксированном m, не удалось.

Утверждение 2 доказано.

Рассмотрим примечательный частный случай $n = 2m,\ m > 1$. Пусть $k \geqslant m$. Тогда при $n = 2m$ справедливо $n-1-k < m$. Также заметим, что при $n = 2m$:

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

Кроме того, при $q = \frac{1}{2}$ выполнено $q^{k-1}(1-q)^{n-2-k} = 2^{3-n}$.

Эти факты позволяют вычислить частную производную при $q = \frac{1}{2}$.

$$\frac{\partial\tilde{f}}{\partial q}\left(n = 2m, q = \frac{1}{2}\right) =$$

$$= \sum_{k=m}^{n-1}\frac{m}{k+1}\binom{2m-1}{k}2^{3-2m}(k + (2m-k-1) - (2m-1)) = 0.$$

В общем случае непросто показать, что найденная точка является точкой максимума. Для примера рассмотрим $n = 4, m = 2$. Условие второго порядка:

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

Таким образом, $\tilde{f}$ является вогнутой функцией по q при $n = 4,\ m = 2$, значит, в точке $q = 0{,}5$ она достигает глобального максимума.

Интересен вопрос о том, к чему стремится точка максимума функции $\tilde{f}$ (обозначим через $q^*(n,m)$) при больших m. Для этого находим q^* при фиксированных n, m численными методами.

Замечаем, что чем больше n, тем больше зависимость $q^*(m)$ похожа на линейную (уже при $n = 100$ отмечаем $R^2 > 99{,}9\%$ для регрессии $q^* = \beta_0 + \beta_1 m$; R^2 растет с ростом n). Кроме того, с ростом n $q^*(m = 2)$ стремится к 0, а $q^*(m = n-2)$ стремится к 1. Это позволяет выдвинуть гипотезу о том, что при больших n и $1 < m < n-1$ точка максимума $\tilde{f}(q)$ стремится к $q^* = \frac{m-2}{n-4}$ (рис. 2). Проверка этой гипотезы интересна для дальнейших исследований; в случае истинности интересна интуиция результата.

Утверждение 3. В общем случае пороговое значение c^* немонотонно по p и α.

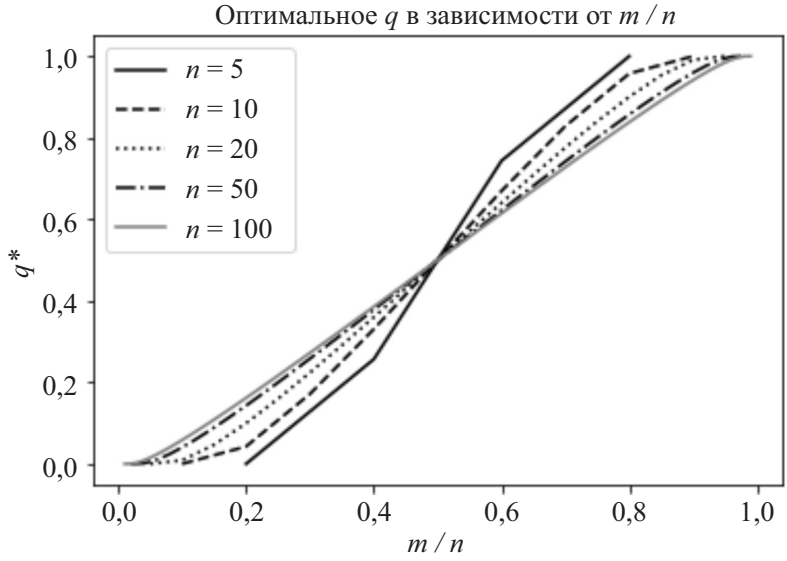

Рис. 2. Точка максимума $\tilde{f}(q)$ в зависимости от n и отношения $\frac{m}{n}$.

Доказательство. Вычислим производные функции, заданной в неявном виде, как было сделано ранее в утверждении 2.

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

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

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

Преобразования с частными производными аналогичны преобразованиям в утверждении 1, когда данные производные рассчитывались для случая $m = 1$. В утверждении 2 была доказана немонотонность $\tilde{f}$ по q в общем случае. Отсюда следует, что $\operatorname{sgn}\left(\frac{\partial \tilde{f}}{\partial q}\right)$ не является постоянной величиной. Но тогда и $\operatorname{sgn}\left(\frac{\partial c^*}{\partial p}\right), \operatorname{sgn}\left(\frac{\partial c^*}{\partial \alpha}\right)$ не являются постоянными величинами при $1 < m < n - 1$. Случай $m = 1$ был рассмотрен ранее, для $m = n - 1$ все выводы противоположны случаю $m = 1$ (т.е. $\frac{\partial c^*}{\partial p} > 0$, $\frac{\partial c^*}{\partial \alpha} > 0$ при $m = n - 1$). Более того, очевидно, что максимумы по p и по q также однозначно связаны.

Утверждение 3 доказано.

Продемонстрируем немонотонность наглядно. Численно вычислим значение c^* при фиксированных значениях параметров из условий $c^* = \tilde{f}(m, q(c^*))$, $q = p + (1 - p)\alpha F(c^*)$. Зафиксируем, что издержки распределены равномерно на отрезке $[0, 1]$, и найдем несколько решений уравнений для

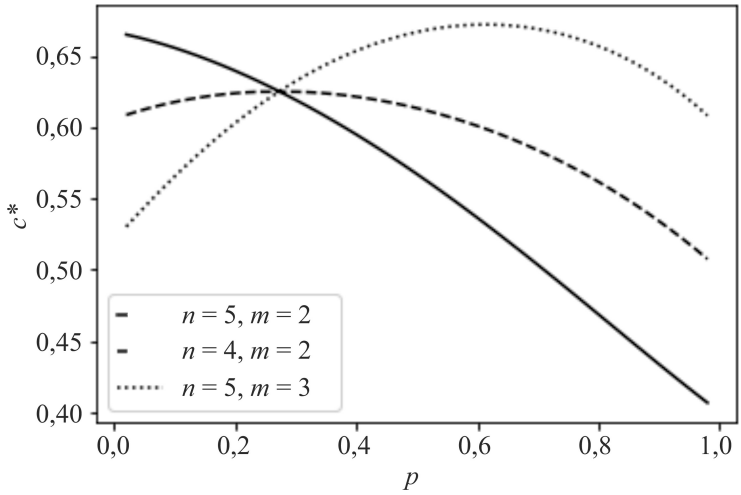

Рис. 3. Зависимость c^* от p при $\alpha = 0{,}5$, $F(c) = c$.

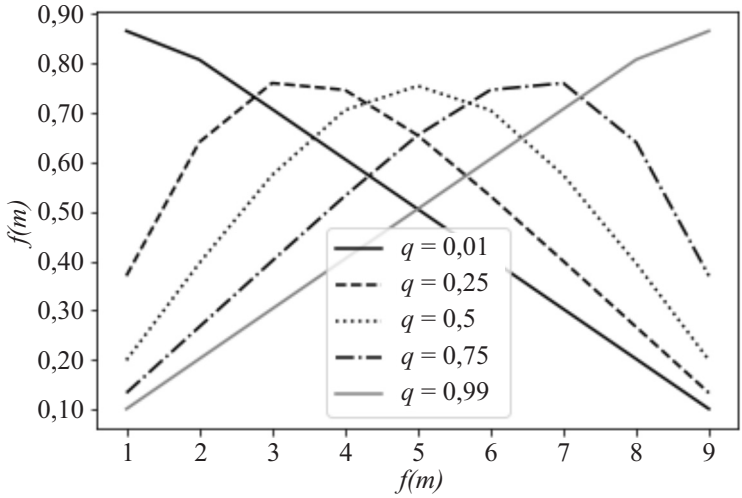

Рис. 4. Зависимость c^* от m при $n = 20$, $\alpha = 0,2$, $F(c) = c$.

разных значений параметра с точностью до четвертого знака после запятой.

$$c^*(n = 4,\ m = 2,\ p = 0{,}1,\ \alpha = 0{,}5) = 0{,}6175,$$
$$c^*(n = 4,\ m = 2,\ p = 0{,}3,\ \alpha = 0{,}5) = 0{,}6248,$$
$$c^*(n = 4,\ m = 2,\ p = 0{,}5,\ \alpha = 0{,}5) = 0{,}6132.$$

Таким образом, монотонности c^* по p в общем случае нет (рис. 3). Интуитивно объясним этот факт: с ростом вероятности p растет исходное число хороших агентов, таким образом, что при небольшой их доле есть смысл активно мимикрировать и бороться за сравнительно большое число “призов”. Однако если их доля уже высока, то конкурировать с ними сложно, вероятность победы слишком мала, так что доля мимикрирующих (и соответствующее пороговое значение) падает.

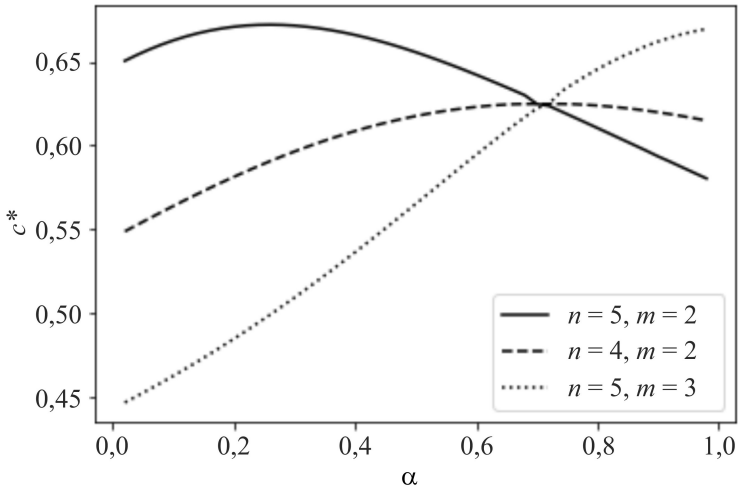

Рис. 5. Зависимость c^* от α при $p = 0,1$, $F(c) = c$.

Аналогично исследуем зависимость от масштаба утечки α.

$$c^*(n = 4,\ m = 2,\ p = 0{,}1,\ \alpha = 0{,}5) = 0{,}6175,$$
$$c^*(n = 4,\ m = 2,\ p = 0{,}1,\ \alpha = 0{,}7) = 0{,}6248,$$
$$c^*(n = 4,\ m = 2,\ p = 0{,}1,\ \alpha = 0{,}9) = 0{,}6198.$$

Таким образом, монотонности c^* по α в общем случае нет (рис. 4). Здесь действует аналогичная логика: с ростом утечки все больше слабых потенциально захочет мимикрировать под сильных и до какого-то уровня это будет для них выгодно, однако далее конкуренция станет слишком жесткой и поэтому желание мимикрировать снизится.

Утверждение 4. Пороговое значение c^* немонотонно по m в общем случае.

Как и в утверждении 3, вычислим c^* при различных значениях параметров с точностью до четырех знаков после запятой и покажем немонотонность порогового значения в равновесии по числу победителей.

$$c^*(n = 4,\ m = 1,\ p = 0{,}5,\ \alpha = 0{,}5) = 0{,}3910,$$
$$c^*(n = 4,\ m = 2,\ p = 0{,}5,\ \alpha = 0{,}5) = 0{,}6132,$$
$$c^*(n = 4,\ m = 3,\ p = 0{,}5,\ \alpha = 0{,}5) = 0{,}5045.$$

В этом случае объяснение заключается в том, что при росте числа победителей близко к числу всех участников может стать бессмысленно стараться и тратить средства на мимикрирование, так как и без этого вероятность войти в число m победителей высока. Однако, как видно из рис. 5, для особых скоринговых моделей, порождающих очень большую или очень маленькую вероятность классификации игрока как хорошего, имеет место монотонная зависимость доли мимикрирующих от числа призов.

5. Заключение

В статье смоделирована ситуация, при которой часть клиентов компании узнает свой внутренний рейтинг в компании и может изменить свое поведение, чтобы увеличить внутренний рейтинг. Рассмотрена постановка, в которой скоринговая модель распределяет пользователей на “хороших” и “плохих” (бинарная классификация типов агентов).

Доказано, что равновесие существует, единственно и представляет собой профиль монотонных стратегий. Наличие внутреннего порогового значения говорит о том, что отнюдь не все агенты, получив доступ к механизму работы скоринговой модели, будут использовать эту информацию в манипулятивных целях.

Полагаем, что существует большой потенциал для дальнейших исследований. В случае бинарного распределения типов интересно подробнее изучить найденные немонотонные зависимости, например найти точки экстремумов аналитически в зависимости от значений параметров. Может быть полезно расширить бинарное распределение до дискретного, так как некоторые модели распределяют пользователей по нескольким кластерам, а также до непрерывного, так как многие модели возвращают рейтинг как некоторое действительное число, часто на отрезке $[0, 1]$.

Подводя итог, хотелось бы отметить, что рассмотренная проблема манипулирования поведением является примером проявления свойства естественного интеллекта непрерывно адаптироваться и искать сложные и нетривиальные стратегии улучшения своей полезности. В этом смысле вряд ли автоматизированные модели способны в полной мере противостоять всем возможностям, потенциально доступным человеку. Однако если заранее спрогнозировать стратегическое поведение пользователей, то можно более эффективно оценить как устойчивость, так и перспективы исходной модели, что может повлиять на оптимальный выбор модели, либо сформулировать более адекватные критерии к улучшению модели. Другой примыкающей проблемой является манипулирование поведением при обучении модели, что безусловно повлияет на ее дальнейшее функционирование. Это приводит к тому, что оптимальная траектория развития науки о моделях машинного обучения и искусственного интеллекта должна включать адекватный теоретико-игровой элемент, учитывающий, что данные для этих моделей генерируются стратегическими людьми.

СПИСОК ЛИТЕРАТУРЫ

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., Vamossy D.* Predicting consumer default: A deep learning approach // National Bureau of Econom. Res. 2019. No. w26165. 72 p.

3. *Plawiak P., Abdar M., Plawiak J., et al.* DGHNL: A new deep genetic hierarchical network of learners for prediction of credit scoring // Inform. Sci. 2020. V. 516. P. 401–418.
4. *Hurley M., Adebayo J.* Credit scoring in the era of big data // Yale JL & Tech. 2016. V. 18. No. 148.
5. *Martinez A., Schmuck C., Pereverzyev S., Pirker C., Haltmeier M.* A machine learning framework for customer purchase prediction in the non-contractual setting // Eur. J. Oper. Res. 2020. V. 281. No. 3. P. 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. V. 336. No. 012017.
7. *Roth A.* Game theory as a part of empirical economics // The Econ. J. 1991. V. 101. No. 404. P. 107–114.
8. *Harsanyi J.* Games with incomplete information played by "bayesian" players, I–III part I. The basic model // Management Sci. 1967. V. 14. No. 3. P. 159–182.
9. *Бурков В.Н.* Основы математической теории активных систем. М.: Наука, 1977. Т. 255.
10. *Бурков В.Н., Новиков Д.А.* Теории активных систем 50 лет: история развития // Материалы международной научно-практической конференции «ТЕОРИЯ АКТИВНЫХ СИСТЕМ – 50 лет» (ТАС-50, Москва). М.: ИПУ РАН, 2019. С. 10–57.
11. *Еналеев А.К.* Оптимальность согласованных механизмов функционирования в активных системах // Управление большими системами: сборник трудов. 2011. № 33. С. 143–166.
12. *Еналеев А.К.* Оптимальный согласованный механизм в системе с несколькими активными элементами // Проблемы управления. 2015. № 3. С. 20–28.
13. *Бурков В.Н., Еналеев А.К., Коргин Н.А.* Согласованность и неманипулируемость механизмов организационного управления: текущее состояние проблемы, ретроспектива, перспективы развития теоретических исследований // А и Т. 2021. № 7. С. 5–37.
14. *Dellarocas C.* Strategic manipulation of internet opinion forums: Implications for consumers and firms // Management Sci. 2006. V. 52. No. 10. P. 1577–1593.
15. *Dini F., Spagnolo G.* Buying reputation on eBay: Do recent changes help? // Int. J. Electron. Business. 2009. V. 7. No. 6. P. 581–598.
16. *Wright R. et al.* Directed search and competitive search equilibrium: A guided tour // J. Econ. Lit. 2021. V. 59. No. 1. P. 90–148.
17. *Sandomirskaia M., Shavshin R.* Price Competition in Finite Markets with a Rare Good and Private Consumer Valuations // Higher School Econom. Res. Paper, 2021. V. 248. 29 p.
18. *Peters M.* Bertrand equilibrium with capacity constraints and restricted mobility // Econometrica. 1984. V. 52. No. 5. P. 1117–1127.

Статья представлена к публикации членом редколлегии Д.А. Новиковым.

Поступила в редакцию 25.01.2024

После доработки 01.07.2024

Принята к публикации 10.07.2024