

Вестник Череповецкого государственного университета. 2022. № 3 (108). С. 61–76.
Cherepovets State University Bulletin, 2022, no. 3 (108), pp. 61–76.

Научная статья

УДК 519.257

<https://doi.org/10.23859/1994-0637-2022-3-108-5>

Система прогнозирования участников государственных закупок по федеральным законам № 44-ФЗ и № 223-ФЗ

Мария Сергеевна Цыганова^{1✉}, Дарья Александровна Чернушенко²,
Сергей Владимирович Буреш³

^{1, 2}Тюменский государственный университет,
Тюмень, Россия,

¹m.s.cyganova@utmn.ru[✉], <https://orcid.org/0000-0002-9832-7882>

²d1ash2a@yandex.ru, <https://orcid.org/0000-0003-0529-7280>

³ООО «Электронный Эксперт»,
Тюмень, Россия,

³sv@eetmn.ru, <https://orcid.org/0000-0002-3899-6283>

Аннотация. Анализ государственных закупок может быть весьма полезен предприятию-поставщику для понимания конкурентной среды размещаемых заказов и, как следствие, принятия решений, направленных на повышение эффективности закупочной деятельности. Данная статья посвящена созданию модуля прогнозирования, позволяющего получать оценки вероятности участия в электронных государственных закупках по федеральным законам № 44-ФЗ и № 223-ФЗ предприятий-поставщиков, имеющихся в информационной базе системы поиска тендеров. В рамках проекта были реализованы сценарии очистки исходных данных о закупках, конструирования входных признаков для модели прогнозирования, анализа значимости и отбора этих признаков; выполнено построение и обучение модели логистической регрессии. В дальнейшем предполагается интеграция разработанного модуля с системой поиска и анализа тендеров MetaTender.

Ключевые слова: электронные государственные закупки, бинарная классификация, аналитическая платформа, прогнозирование вероятности

Благодарность. Работа выполнена при поддержке благотворительного фонда Владимира Потанина по гранту № ГСГК-30/19 от 30.09.2019 на создание магистерской программы «Внедрение и разработка систем управления производственными процессами нефтегазопереработки».

Для цитирования: Цыганова М. С., Чернушенко Д. А., Буреш С. В. Система прогнозирования участников государственных закупок по федеральным законам № 44-ФЗ и № 223-ФЗ // Вестник Череповецкого государственного университета. 2022. № 3 (108). С. 61–76. <https://doi.org/10.23859/1994-0637-2022-3-108-5>.

Predicting system for public procurement participants under
Federal Laws No. 44 and No. 223

Mariya S. Tsyganova^{1✉}, Darya A. Chernushenko², Sergey V. Buresh³

^{1, 2}University of Tyumen,
Tyumen, Russia,

¹m.s.cyganova@utmn.ru[✉], <https://orcid.org/0000-0002-9832-7882>

²d1ash2a@yandex.ru, <https://orcid.org/0000-0003-0529-7280>

³ООО “Elektronnyi Ekspert” (Electronic Expert LLC),
Tyumen, Russia,

³sv@eetmn.ru, <https://orcid.org/0000-0002-3899-6283>

Abstract. Public procurement analysis might be quite useful for a supplier enterprise to understand the competitive environment of placed orders and, consequently, to make decisions aimed at improving the efficiency of its procurement activities. This work considers the development of the prediction module that allows obtaining estimates of the involvement in electronic public procurement under Federal Laws 44 and 223 of supplier enterprises, listed in information base of tender search system. Within the project, scripts for original procurement data clearing, constructing input features for the prediction model, analyzing significance and selecting these features were developed; logistic regression model was also built and trained. In the future, integration of the developed module with the system of tender search and analysis MetaTender is assumed.

Keywords: digital public procurements, binary classification, analytical platform, probability prediction

Acknowledgments. The work was supported by the Vladimir Potanin Charitable Foundation and carried out under grant № GSGK-30/19 of 30.09.2019 for the development of the master’s program “Implementation and Development of Control Systems for Oil and Gas Processing Production Processes”.

For citation: Tsyganova M. S., Chernushenko D. A., Buresh S. V. Predicting system for public procurement participants under Federal Laws No. 44 and No. 223. *Cherepovets State University Bulletin*, 2022, no. 3 (108), pp. 61–76. (In Russ.). <https://doi.org/10.23859/1994-0637-2022-3-108-5>.

Введение

На сегодняшний день проведение конкурсных торгов является наиболее выгодным способом выбора поставщика товаров или услуг. По сравнению с традиционными формами закупок электронные торги имеют ряд преимуществ: эффективный контроль закупочных процессов на всех этапах, их автоматизация, оперативность торгов, а также возможность серьезно расширить круг потенциальных поставщиков. Кроме того, данный вид снабжения является необходимым условием товарооборота на государственном уровне.

За 2020 г. Управлением государственных закупок Тюменской области было организовано закупок за счет средств областного бюджета и внебюджетных источников

финансирования на общую сумму 52 млрд 385 млн руб.¹ Кроме того, для осуществления закупок привлекались финансовые средства федерального и муниципального бюджетов, реализуемых в регионе. Согласно статистической информации официального сайта Единой информационной системы в сфере закупок, в 2020 г. было заключено 27005 контрактов на общую сумму 87,52 млрд руб. в рамках федерального закона № 44-ФЗ и 30366 контрактов на сумму 120,29 млрд руб. по закону № 223-ФЗ². Количество коммерческих организаций – потенциальных поставщиков государственных заказов, запись о которых внесена в Единый государственный реестр юридических лиц (кроме юридических лиц, прекративших свою деятельность), в Тюменской области на начало 2020 г. составляло 34591 предприятие³.

Анализ государственных закупок может быть весьма полезен для определения стратегии предприятия-поставщика. Например, наличие на соответствующем рынке значительного количества тендеров при отсутствии высокой конкуренции дает стимул к изучению вопроса выхода на данное направление. В случае, когда рынок перегружен предложениями, а средняя стоимость снижения цены контракта невелика, получить маржинальный доход не представляется возможным. В такой ситуации решение об участии в закупках может приниматься только с целью формирования положительного имиджа компании при условии жесткого контроля затрат.

В результате проведенного нами исследования⁴ был определен ряд показателей для оценки эффективности участия в электронных государственных закупках. Их значения вычислялись на основе данных электронных торгов, полученных из первичных источников. Предварительная обработка информации и расчет показателей выполнялись с помощью сценариев, реализованных на базе аналитической платформы Loginom⁵. Продолжением исследования стало создание представленной в данной работе системы прогнозирования участников государственных закупок, проводившихся в рамках федеральных законов № 44-ФЗ и № 223-ФЗ.

Основная часть

Постановка задачи

Важность роли электронных торгов в современном ведении бизнеса послужила предпосылкой к появлению рынка программных продуктов поддержки закупочной

¹ Официальный портал органов государственной власти Тюменской области. Отчет о результатах деятельности за 2020 год. URL: <https://gzto.admtymen.ru/OIGV/gz/actions/ongoing/more.htm?id=11901575@cmsArticle> (дата обращения: 05.07.2021).

² Единая информационная система в сфере закупок. URL: <https://zakupki.gov.ru/epz/main/public/home.html> (дата обращения: 05.07.2021).

³ Федеральная налоговая служба. URL: https://www.nalog.ru/rn72/related_activities/statistics_and_analytics/forms/8376083/ (дата обращения: 05.07.2021).

⁴ Цыганова М. С., Буреш С. В., Чернушенко Д. А. Разработка инструментария анализа данных о государственных закупках по федеральным законам № 44-ФЗ и № 223-ФЗ (на базе аналитической платформы Loginom) // Вестник Череповецкого государственного университета. 2020. № 6 (99). С. 59–72.

⁵ Аналитическая платформа Loginom. URL: <https://loginom.ru/> (дата обращения: 05.07.2021).

деятельности. Функционал таких систем обычно включает обслуживание основных операций закупочных процедур (создание, размещение и сопровождение заявок, подбор поставщиков, поиск и анализ тендеров и т. п.). Со стороны компаний-поставщиков наибольший интерес до недавнего времени вызывали услуги по получению электронной подписи и аккредитации на выбранной торговой площадке; размещению и сопровождению заявки на участие в закупочных процедурах; обеспечению банковской гарантии для заключения контракта. Однако на сегодняшний день наблюдается смещение приоритета в сторону сервисов по поиску и анализу тендеров. Это связано с тем, что, как уже отмечалось выше, анализ электронных закупок может быть весьма полезен компании-поставщику для понимания конкурентной среды размещаемых заказов и, как следствие, принятия обоснованных решений, направленных на повышение эффективности своей закупочной деятельности. Результативность в данном случае может оцениваться по двум основным направлениям:

- снижение затрат на закупочную деятельность посредством выявления действительно перспективных для данной компании тендеров и фокусирования своих усилий именно на этих закупках;
- повышение вероятности победы в торгах, что также является следствием правильной оценки ожиданий от участия в той или иной процедуре.

Аналитический функционал имеющихся на рынке систем автоматизации закупочных процедур, как правило, включает возможности формирования отчетов в различных разрезах (проектов и программ, товаров, работ и услуг, ответственных лиц и т. п.). При этом, как показало изучение требований пользователей системы MetaTender¹, выполненное компанией «Электронный эксперт»², одним из пяти основных запросов авторизованных пользователей стало отображение прогноза возможных участников тендера. На сегодняшний день данная опция только начинает появляться в отдельных программных продуктах, представленных на рынке. Для ее реализации необходимо создание модуля прогнозирования, позволяющего получать оценки вероятности участия известных системе компаний-поставщиков в электронных государственных закупках в рамках федеральных законов № 44-ФЗ и № 223-ФЗ. В настоящей работе поставлена задача разработки такого модуля.

Доступными исходными данными, на основании которых можно пытаться прогнозировать активность поставщиков, являются сведения о государственных тендерах, выгруженные из Единой информационной системы в сфере закупок. В ходе анализа, выполненного нами в 2020 г.³, было установлено, что исходные данные могут содержать пропуски и дубликаты, также возможно наличие ошибочных записей. Кроме того, имеющиеся характеристики закупок и поставщиков (в том виде, в кото-

¹ MetaTender. URL: <https://metatender.ru/> (дата обращения: 05.07.2021).

² Электронный эксперт. URL: <http://eetmn.ru/company> (дата обращения: 05.07.2021).

³ Цыганова М. С., Буреш С. В., Чернушенко Д. А. Разработка инструментария анализа данных о государственных закупках по федеральным законам № 44-ФЗ и № 223-ФЗ (на базе аналитической платформы Loginom) // Вестник Череповецкого государственного университета. 2020. № 6 (99). С. 59–72.

ром они хранятся в исходном наборе) непригодны для их использования в качестве входных признаков модели прогнозирования.

Следовательно, в ходе разработки модуля необходимо решить ряд задач:

- осуществить первичный анализ исходных данных о закупках; выявить возможные проблемы (пропуски, ошибки и т. п.) и определить способы решения выявленных проблем;
- реализовать сценарии очистки первичных данных;
- сконструировать входные признаки для модели прогнозирования (feature engineering¹), реализовать соответствующие сценарии;
- выполнить анализ значимости полученных признаков, реализовать их отбор для подачи на вход модели прогнозирования;
- построить и обучить модель прогнозирования участников закупок, протестировать ее;
- спроектировать механизмы интеграции модуля с системой поиска тендеров.

Формирование и очистка исходного набора данных

В качестве первичного источника использовалась база данных, полученная в результате парсинга официального сайта Единой информационной системы в сфере закупок. Она содержит 25 таблиц с различной информацией, однако информация о закупках находится только в трех таблицах, по которым и был сформирован набор данных для последующей обработки. Структура полученного набора представлена в табл. 1.

Сформированный набор данных содержал 4607 записей, относящихся к закупкам в период с января 2018 года по апрель 2021 года.

С целью реализации процедур очистки данных (выявления и обработки пропусков, дубликатов и противоречий) использовалась аналитическая платформа Loginom. Базовые принципы, положенные в основу соответствующих сценариев, в целом аналогичны опубликованным ранее². По этой причине достаточно отметить, что с помощью реализованных процедур очистки было обработано 1042 записи с пропущенными значениями и 142 записи с дубликатами. Противоречий в имеющемся наборе данных не выявлено, однако в сценарий заложена проверка на противоречивость, поскольку база данных будет пополняться новыми записями, качество данных в которых заранее неизвестно.

¹ Zheng A., Casari A. Feature Engineering for Machine Learning. Principles and Techniques for Data Scientists. Sebastopol, CA: O'Reilly Media, 2018. 218 p.

² Цыганова М. С., Буреш С. В., Чернушенко Д. А. Разработка инструментария анализа данных о государственных закупках по федеральным законам № 44-ФЗ и № 223-ФЗ (на базе аналитической платформы Loginom) // Вестник Череповецкого государственного университета. 2020. № 6 (99). С. 59–72.

Таблица 1

Структура исходного набора данных

Описание поля	Наименование	Формат
Реестровый номер извещения о закупке (идентификатор)	purchaseNumber	Строковый
Дата публикации закупки	publicationDateTime	Дата / время
Наименование закупки	title	Строковый
Закон закупки (№ 44-ФЗ или № 223-ФЗ)	purchaseType	Целый
Классификатор закупки по ОКВЭД	okved	Строковый
Дата окончания подачи заявок	closeDateTime	Дата / время
Заказчик (ИНН как идентификатор)	customerInn	Целый
Цена тендера	price_tender	Вещественный
Номер лота закупки	lot_num	Целый
Участник лота (ИНН как идентификатор)	providerInn	Целый
Победитель лота	winner	Логический
Начальная максимальная цена контракта (НМЦК) лота	nmck	Вещественный
Итоговая цена лота	price_lot	Вещественный
Итоговая цена контракта	price_contract	Вещественный
Дата публикации контракта	contacDateTime	Дата / время

Особо следует отметить процесс обработки значений в поле «Классификатор закупки по ОКВЭД»: путем применения регулярного выражения из строки исключались буквы и выбирались первые две пары цифр в формате **.**. Такой способ представления данных в указанном поле применялся на последующих этапах обработки для определения отрасли закупки.

Конструирование входных признаков для модели прогнозирования

Среди полей исходного набора данных (см. табл. 1) несколько полей являются идентифицирующими (не содержат полезной информации для построения прогноза). Значения некоторых полей на момент публикации закупки неизвестны («Итоговая цена контракта», «Дата публикации контракта»), следовательно, эти признаки не могут быть использованы для прогнозирования участия в закупке. В качестве отправных точек для целей прогнозирования можно рассматривать такие поля, как «Наименование закупки», «Классификатор закупки по ОКВЭД», «Дата публикации закупки», «Цена тендера», «НМЦК лота», «Дата окончания подачи заявок». Следует обратить внимание на то, что значения всех этих полей могут как-то характеризовать закупку, для которой требуется прогнозирование участия, но сами по себе ничего не говорят о предполагаемых участниках торгов. Чтобы оценивать вероятность участия в торгах того или иного поставщика, заявленного в системе, необходимо сформировать дополнительные признаки, характеризующие активность каждого участника в

предшествующих закупочных процедурах. В этой связи следует отметить, что все указанные признаки можно условно разбить на 2 группы:

- признаки, характеризующие закупки;
- признаки, отражающие активность поставщиков.

Общий перечень выделенных нами признаков представлен в табл. 2.

Таблица 2

Перечень сформированных признаков закупочной деятельности

Характеристики закупки	Характеристики участников торгов
1. Количество уникальных участников по классификатору закупки. 2. Месяц публикации закупки. 3. Сезон (осень, зима, весна, лето) публикации закупки	1. Количество уникальных классификаторов закупок, в которых принимал участие поставщик. 2. Общее количество побед. 3. Дата первого (по времени) участия в торгах. 4. Дата последнего (по времени) участия в торгах. 5. Интервал времени (количество месяцев) между датами первой и последней (по времени) закупок, в которых принимал участие поставщик. 6. Минимальная стоимость лота по каждому классификатору закупки. 7. Максимальная стоимость лота по каждому классификатору закупки. 8. Количество участий поставщика по каждому классификатору закупки. 9. Количество побед поставщика по каждому классификатору закупки

Для формирования дополнительных признаков были разработаны сценарии на базе платформы Loginot. Все признаки (кроме второго и третьего из первой группы) рассчитывались посредством различных операций агрегации записей исходного набора. Алгоритм вычислений предполагает соблюдение следующего условия: все агрегированные значения (общее количество побед, количество участий и др.) должны определяться не на основе общего объема имеющихся данных, а с учетом периода времени, предшествующего дате закупки, для которой необходимо выполнить прогнозирование. В процессе реализации использовались стандартные компоненты: «Группировка», «Цикл» и др. После вычисления всех указанных показателей соответствующие столбцы были присоединены к набору данных, полученному на предыдущем этапе.

Для того чтобы можно было использовать набор данных для обучения модели прогнозирования участия в закупках, необходимо обеспечить наличие в нем информации об участии / неучастии каждого из поставщиков во всех процедурах торгов за рассматриваемый промежуток времени. С этой целью выполнялось полное слияние записей с уникальными значениями ИНН поставщиков, с записями о закупках, а также создавался дополнительный столбец-индикатор участия поставщика в конкретной процедуре торгов. При совпадении ИНН поставщика с одним из участников

закупки в столбец-индикатор вносилось значение True, в противном случае – False. Реализация действия была выполнена с помощью компонентов «Слияние» и «Калькулятор».

Результатом всех описанных процедур обработки исходных данных стал набор, содержащий почти 3 млн записей с 29 столбцами (включая исходные и сконструированные признаки).

Анализ значимости признаков

Процесс конструирования входных признаков для моделей прогнозирования обычно отражает точку зрения исследователей на решаемую задачу; полученный набор производных столбцов фактически представляет собой комплекс гипотез относительно того, какие конкретно факторы способны оказать влияние на значения прогнозируемых показателей. Очевидно, что среди таких признаков могут оказаться действительно полезные, малозначимые и даже шумовые. Проблема выделения значимых признаков играет важную роль и с точки зрения эффективности обучения модели Machine Learning, и в аспекте понимания исследуемых зависимостей и формирования методик последующего применения полученных моделей.

Все вышесказанное имеет непосредственное отношение к набору признаков, представленному в табл. 2. Прежде чем переходить к построению модели машинного обучения, необходимо выполнить исследование значимости признаков сформированного набора данных.

Одним из известных методов оценки значимости признаков¹ в задачах бинарной классификации является вычисление Информационного индекса (коэффициента IV)² для предполагаемых входных столбцов модели прогнозирования. Этот показатель, в свою очередь, рассчитывается на основе коэффициентов WoE ³. Каждому наблюдению, содержащему набор признаков, ставится в соответствие бинарная выходная переменная (Событие или Не-событие). После этого весь диапазон изменения значений каждого признака разбивается на несколько начальных классов, для каждого из которых вычисляется коэффициент WoE :

$$WoE_i = \ln \frac{\frac{N_i}{P_i}}{\frac{N}{P}},$$

где i – индекс начального класса; P_i – число Событий, попавших в класс; P – общее число Событий; N_i – число Не-событий, попавших в класс; N – общее число Не-событий. Если соотношение Событий и Не-событий в классе такое же, как во всей выборке (принадлежность классу не меняет вероятности Событий), то значение ко-

¹ Паклин Н. Б., Афанасьев В. В. Оптимальное квантование для повышения качества бинарных классификаторов // Искусственный интеллект. 2013. № 4. С. 392–399.

² Коэффициент IV. URL: <https://wiki.loginom.ru/articles/coefficient-iv.html> (дата обращения: 05.07.2021).

³ Коэффициент WoE . URL: <https://wiki.loginom.ru/articles/coefficient-woe.html> (дата обращения: 05.07.2021).

эфициента WoE будет близко к нулю. Отрицательные значения коэффициентов WoE указывают на более высокую вероятность появления Событий в классе, чем в выборке в целом; положительные – на более высокую вероятность Не-событий. Таким образом, коэффициенты WoE характеризуют значимость отдельных классов значений данного признака для прогнозирования результата.

Информационный индекс является агрегированной величиной, характеризует предсказательную способность признака в целом и определяется по формуле

$$IV = \sum_{i=1}^k \left(\left(\frac{N_i}{N} - \frac{P_i}{P} \right) \cdot WoE_i \right).$$

Коэффициенты IV всегда положительны; на основе их величины значимость признаков оценивается по следующей шкале: при $IV < 0,02$ значимость признака отсутствует; при $0,02 \leq IV < 0,1$ значимость низкая; при $0,1 \leq IV < 0,3$ – средняя; при $IV \geq 0,3$ – высокая.

Следует отметить, что, помимо оценки значимости входных признаков модели бинарной классификации, коэффициенты WoE и IV используются для определения оптимального квантования значений непрерывных и дискретных полей (т. е. формирования конечных классов) перед подачей их на вход модели прогнозирования. Это позволяет повысить эффективность ее обучения как за счет максимизации значимости входных признаков, так и в контексте борьбы с выбросами.

Вычисление коэффициентов IV столбцов набора данных и WoE -анализ реализованы нами в стандартном компоненте платформы Loginom, называемом «Конечные классы»¹. Данный обработчик преобразует значения входных столбцов в последовательность интервалов (конечные классы) с определенными метками; для всех столбцов рассчитывается значение информационного индекса и определяется уровень значимости согласно приведенной выше шкале. Применение соответствующего узла в сценарии обработки данных о закупках на первом этапе продемонстрировало практически нулевые значения коэффициентов IV для всех входных столбцов. Такой результат легко объяснить, если обратить внимание на соотношение числа Не-событий и Событий в анализируемом наборе: 2925125 к 3423 соответственно. Такое соотношение отражает тот факт, что для каждого конкретного предприятия-поставщика доля закупок, в которых он принимал участие, в общем числе тендеров, проведенных за рассматриваемый промежуток времени, очень невелика. Другими словами, в задаче бинарной классификации, подлежащей решению в рамках указанного проекта, имеется ярко выраженный дисбаланс классов.

В ходе решения задач классификации с несбалансированным набором данных, как правило, реализуются дополнительные процедуры для преодоления проблемы дисбаланса. В противном случае процесс обучения алгоритма классификации может завершиться получением константной модели, выдающей для всех объектов метку преобладающего (мажоритарного) класса.

¹ Конечные классы. URL: <https://help.loginom.ru/userguide/visualization/coarse-classes/> (дата обращения: 05.07.2021).

Решение проблемы дисбаланса классов и отбор признаков

Традиционными методами решения проблемы дисбаланса классов считаются методы повторной выборки, которые можно разделить на две категории¹: недостаточная выборка для мажоритарного класса и избыточная выборка для миноритарного класса.

К первой категории относятся следующие методы:

- Random Under-sampling (RUS) – набор наблюдений балансируется путем удаления случайно выбранных объектов мажоритарного класса;
- Neighborhood cleaning rule (NCR) – классификация всех наблюдений осуществляется по правилу трех ближайших соседей с последующим удалением наблюдений мажоритарного класса из числа получивших верную метку и неверно классифицированных соседей миноритарного класса.

Ко второй категории методов повторной выборки относятся следующие:

- SMOTE – последовательный выбор одного наблюдения из миноритарного класса и построение нового объекта путем случайной модификации одного из ближайших соседей выбранного объекта;
- Random Over-sampling (ROS) – генерация синтетических наблюдений миноритарного класса случайным образом.

В ходе данной работы исследовалась применимость методов RUS и ROS к решению проблемы дисбаланса в рассматриваемой задаче. Для сравнительного анализа эффективности указанных методов в Logitom-сценарий были добавлены узлы, запускающие работу Python-скриптов, реализующих соответствующие алгоритмы. Сбалансированные наборы данных подавались на вход узла «Конечные классы», выполняющего построение оптимального квантования и оценку значимости входных признаков. На рис. 1 представлены результаты оценки значимости, полученные для наборов данных, сформированных с помощью методов ROS и RUS.

Анализ результатов позволяет сделать вывод о том, что выборки, сформированные путем применения каждого из обозначенных методов, дают возможность выполнить отбор значимых признаков, причем получаемые оценки в целом совпадают. По этой причине окончательный выбор в пользу одного из методов можно будет сделать после построения и обучения модели классификации – на основании качества предсказаний. На данном же этапе было принято решение использовать в качестве входных полей модели классификации только те признаки, значимость которых признана средней или высокой (значение коэффициента IV составило не менее 0,1). Это решение нашло отражение в сценарии обработки данных.

¹ Рюмина Е. В., Карпов А. А. Сравнительный анализ методов устранения дисбаланса классов эмоций в видеоданных выражений лиц // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20, № 5. С. 683–691.

ab Имя столбца	ga Информационный индекс	ab Значимость столбца	ga Информационный индекс	ab Значимость столбца
winner	10,42	Высокая	6,22	Высокая
time_between	0,06	Низкая	0,05	Низкая
month	0,01	Отсутствует	0,01	Отсутствует
season	0,01	Отсутствует	0,00	Отсутствует
publicationDateTime	0,04	Низкая	0,06	Низкая
okved_new	1,16	Высокая	0,78	Высокая
customerInn	0,09	Низкая	0,10	Низкая
date	0,04	Низкая	0,03	Низкая
count_uniq_prov_on_okved	0,02	Низкая	0,03	Низкая
count_part_prov	0,03	Низкая	0,02	Низкая
count_uniq_okv_on_prov	0,02	Отсутствует	0,03	Низкая
count_win	0,35	Высокая	0,34	Высокая
count_win_on_okved	0,00	Отсутствует	0,00	Отсутствует
min_price_on_okved	4,36	Высокая	2,74	Высокая
max_price_on_okved	4,33	Высокая	2,72	Высокая
lot_num	0,00	Отсутствует	0,00	Отсутствует
price_lot	9,60	Высокая	6,97	Высокая
COL2	1,16	Высокая	0,78	Высокая
min_date_prov	0,09	Низкая	0,08	Низкая
max_date_prov	0,05	Низкая	0,06	Низкая
purchaseNumber	5,50	Высокая	4,93	Высокая
okved	2,24	Высокая	1,48	Высокая
purchaseType	0,00	Отсутствует	0,00	Отсутствует
title	4,78	Высокая	3,88	Высокая
time_between_pur_contract	0,03	Низкая	0,02	Низкая
contacDateTime	0,05	Низкая	0,05	Низкая

Метод ROS

Метод RUS

Рис. 1. Оценки значимости признаков, полученные после применения методов ROS и RUS

Особо следует рассмотреть результаты WoE-анализа, отображаемые в визуализаторе узла «Конечные классы». Он выполнялся по всем признакам набора данных, однако мы рассмотрим только некоторые аспекты изучения двух признаков, значимость которых оценена как высокая. На рис. 2 представлены коэффициенты WoE, характеризующие влияние признака «Общее количество побед поставщика» (поле count_win) на вероятность участия этого поставщика в торгах. При отсутствии побед относительная частота неучастия поставщика в торгах заметно больше, чем по всей выборке. В целом увеличение количества побед ведет к повышению вероятности участия (см. верхнюю диаграмму на рис. 2), но при этом изменение индекса IV не является монотонным: рассматриваемый признак имеет наибольшую предсказательную силу в диапазонах «менее 1» и «более 17», в то время как в диапазоне «от 1 до 16» его влияние на результат невелико (см. нижнюю диаграмму на рис. 2).

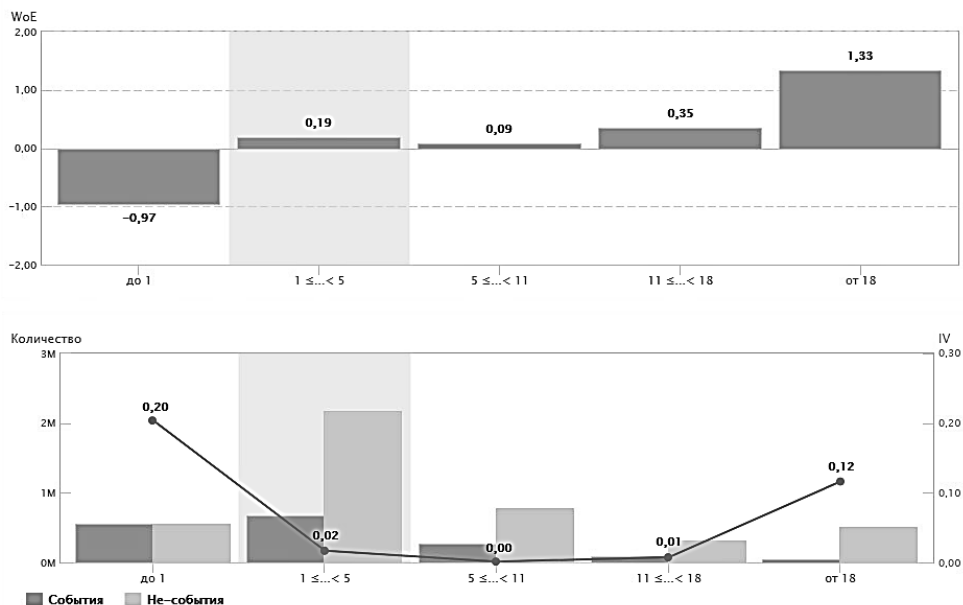


Рис. 2. Влияние общего количества побед поставщика на вероятность участия в торгах

На рис. 3 графически изображены итоги анализа влияния стоимости лота (поле price_lot) на вероятность участия. Для некоторых значений стоимости (см. последний интервал на рис. 3) эта вероятность близка к 1. Для остальных интервалов влияние стоимости на результат намного слабее.

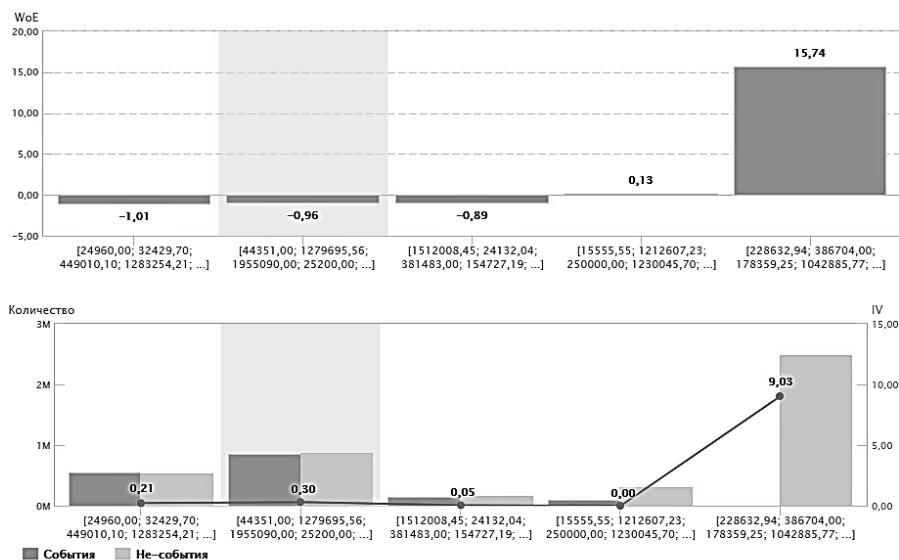


Рис. 3. Влияние стоимости лота на вероятность участия в торгах

Модель машинного обучения

Прогнозирование вероятности участия в торгах на основе подготовленного набора данных с отобранными признаками выполнялось с помощью алгоритма логистической регрессии. Использовался стандартный компонент Loginom – Логистическая регрессия¹. Разбиение набора на обучающую и тестовую выборки выполнялось в соотношении 70/30; при разбиении учитывалась упорядоченность данных во времени.

Модель обучалась на наборах данных, полученных после применения двух методов устранения дисбаланса классов: Random Over-sampling и Random Under-sampling. В обоих случаях были использованы одни и те же настройки алгоритма машинного обучения (для корректности сравнения качества обученных моделей). Оценка качества полученных моделей по метрике AUC ROC представлена в табл. 3.

Таблица 3

Оценка качества классификации после устранения дисбаланса классов

Метод устранения дисбаланса классов	Значение метрики AUC ROC	
	На обучающем множестве	На тестовом множестве
Random Over-sampling	0,8123	0,8120
Random Under-sampling	0,8479	0,8505

Соотношение числа ошибок первого и второго рода, допущенных классификаторами, зависело от выбранного порога отсечения. В целом можно отметить, что по количеству ошибок, как и по значению AUC ROC, метод Random Under-sampling показал более высокие результаты. По этой причине было принято решение отдать предпочтение данному методу и в дальнейшем использовать его в модуле прогнозирования.

Пути интеграции модуля с информационной системой поиска тендеров

На сегодняшний день рассматривается вопрос об интеграции разработанного модуля прогнозирования с системой поиска и анализа тендеров MetaTender. Для интеграции планируется создание web-сервисов, обрабатывающих REST-запросы системы и возвращающих прогноз участников закупки. Реализацию web-сервисов планируется осуществлять на базе платформы Loginom.

Выводы

В ходе проведенного нами исследования разработан модуль прогнозирования участия предприятий-поставщиков в электронных государственных закупках по федеральным законам № 44-ФЗ и № 223-ФЗ. Функционал модуля включает сценарии очистки первичных данных о закупках; сценарии формирования входных признаков для модели прогнозирования; инструментарий анализа значимости полученных признаков и формирования конечных классов для загрузки данных в модель логистической регрессии; модель прогнозирования вероятности участия в закупках известных

¹ Логистическая регрессия. URL: <https://help.loginom.ru/userguide/processors/datamining/logistic-regression> (дата обращения: 05.07.2021).

системе предприятий-поставщиков.

Качество классификации обученной модели по метрике AUC ROC на тестовой выборке составило 0,85.

Представленный нами модуль разработан на базе аналитической платформы Loginom. Его интеграция с системой поиска и анализа тендеров возможна посредством web-сервисов на базе этой же платформы.

Список источников

- Аналитическая платформа Loginom. URL: <https://loginom.ru/> (дата обращения: 05.07.2021).
Госзаказ.ру. URL: <https://gozakaz.ru> (дата обращения: 05.07.2021).
Единая информационная система в сфере закупок. URL: <https://zakupki.gov.ru/epz/main/public/home.html> (дата обращения: 05.07.2021).
Конечные классы. URL: <https://help.loginom.ru/userguide/visualization/coarse-classes/> (дата обращения: 05.07.2021).
Коэффициент IV. URL: <https://wiki.loginom.ru/articles/coefficient-iv.html> (дата обращения: 05.07.2021).
Коэффициент WoE. URL: <https://wiki.loginom.ru/articles/coefficient-woe.html> (дата обращения: 05.07.2021).
Логистическая регрессия. URL: <https://help.loginom.ru/userguide/processors/datamining/logistic-regression> (дата обращения: 05.07.2021).
Официальный портал органов государственной власти Тюменской области. Отчет о результатах деятельности за 2020 год. URL: <https://gzto.admtumen.ru/OIGV/gz/actions/ongoing/more.htm?id=11901575@cmsArticle> (дата обращения: 05.07.2021).
Паклин Н. Б., Афанасьев В. В. Оптимальное квантование для повышения качества бинарных классификаторов // Искусственный интеллект. 2013. № 4. С. 392–399.
Рюмина Е. В., Карпов А. А. Сравнительный анализ методов устранения дисбаланса классов эмоций в видеоданных выражений лиц // Научно-технический вестник информационных технологий, механики и оптики. 2020. Т. 20, № 5. С. 683–691.
Федеральная налоговая служба. URL: https://www.nalog.ru/rn72/related_activities/statistics_and_analytics/forms/8376083/ (дата обращения: 05.07.2021).
Цыганова М. С., Буреш С. В., Чернушенко Д. А. Разработка инструментария анализа данных о государственных закупках по федеральным законам № 44-ФЗ и № 223-ФЗ (на базе аналитической платформы Loginom) // Вестник Череповецкого государственного университета. 2020. № 6 (99). С. 59–72.
Электронный эксперт. URL: <http://eetmn.ru/company> (дата обращения: 05.07.2021).
MetaTender. URL: <https://metatender.ru/> (дата обращения: 05.07.2021).
Zheng A., Casari A. Feature Engineering for Machine Learning. Principles and Techniques for Data Scientists. Sebastopol, CA: O'Reilly Media, 2018. 218 p.

References

- Analiticheskaja platforma Loginom* [Loginom Analytical Platform]. Available at: <https://loginom.ru/> (accessed: 05.07.2021).
Goszakaz.ru [State order.ru]. Available at: <https://gozakaz.ru> (accessed: 05.07.2021).
Edinaia informatsionnaja sistema v sfere zakupok [Unified information system in the field of procurement]. Available at: <https://zakupki.gov.ru/epz/main/public/home.html> (accessed: 05.07.2021).

Konechnye klassy [Final classes]. Available at: <https://help.loginom.ru/userguide/visualization/coarse-classes/> (accessed: 05.07.2021).

Koeffitsient IV [Coefficient IV]. Available at: <https://wiki.loginom.ru/articles/coefficient-iv.html> (accessed: 05.07.2021).

Koeffitsient WoE [Coefficient WoE]. Available at: <https://wiki.loginom.ru/articles/coefficient-woe.html> (accessed: 05.07.2021).

Logisticheskaya regressiya [Logistic regression]. Available at: <https://help.loginom.ru/userguide/processors/datamining/logistic-regression> (accessed: 05.07.2021).

Ofitsial'nyi portal organov gosudarstvennoi vlasti Tiimenskoi oblasti. Otchet o rezul'tatakh deiatel'nosti za 2020 god [The official portal of state authorities in Tyumen region. 2020 Statement of performance]. Available at: <https://gzto.admtymen.ru/OIGV/gz/actions/ongoing/more.htm?id=11901575@cmsArticle> (accessed: 05.07.2021).

Paklin N. B., Afanas'ev V. V. Optimal'noe kvantovanie dlia povysheniia kachestva binarnykh klassifikatorov [Optimal quantization to improve the quality of binary classifiers]. *Iskusstvennyi intellekt* [Artificial Intelligence], 2013, no. 4, pp. 392–399.

Riumina E. V., Karpov A. A. Sravnitel'nyi analiz metodov ustraneniia disbalansa klassov emotsii v videodannykh vyrazhenii lits [Comparative analysis of methods for imbalance elimination of emotion classes in video data of facial expressions]. *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki* [Scientific and Technical Journal of Information Technologies, Mechanics and Optics], 2020, vol. 20, no. 5, pp. 683–691.

Federal'naya nalogovaya sluzhba [Federal Tax Service]. Available at: https://www.nalog.ru/rn72/related_activities/statistics_and_analytics/forms/8376083/ (accessed: 05.07.2021).

Tsyganova M. S., Buresh S. V., Chernushenko D. A. Razrabotka instrumentarii analiza dannykh o gosudarstvennykh zakupkakh po federal'nykh zakonam № 44-FZ i № 223-FZ (na baze analiticheskoi platformy Loginom) [Data analysis instrumentation development for public procurements under 44 and 233 Federal Laws (based on Loginom analytical platform)]. *Vestnik Cherepovetskogo gosudarstvennogo universiteta* [Cherepovets State University Bulletin], 2020, no. 6 (99), pp. 59–72.

Elektronnyi ekspert [Electronic expert]. Available at: <http://eetmn.ru/company> (accessed: 05.07.2021).

MetaTender. Available at: <https://metatender.ru/> (accessed: 05.07.2021).

Zheng A., Casari A. *Feature Engineering for Machine Learning. Principles and Techniques for Data Scientists*. Sebastopol, CA: O'Reilly Media, 2018. 218 p.

Сведения об авторах

Мария Сергеевна Цыганова – кандидат технических наук, доцент; <https://orcid.org/0000-0002-9832-7882>, m.s.cyganova@utmn.ru, Тюменский государственный университет (д. 6, ул. Володарского, 625003 г. Тюмень, Россия); **Mariya S. Tsyganova** – Candidate of Technical Sciences, Associate Professor; <https://orcid.org/0000-0002-9832-7882>, m.s.cyganova@utmn.ru, University of Tyumen (6, ul. Volodarskogo, 625003 Tyumen, Russia).

Дарья Александровна Чернушенко – магистрант; <https://orcid.org/0000-0003-0529-7280>, d1ash2a@yandex.ru, Тюменский государственный университет (д. 6, ул. Володарского, 625003 г. Тюмень, Россия); **Darya A. Chernushenko** – Student in Master's programme; <https://orcid.org/0000-0003-0529-7280>, d1ash2a@yandex.ru, University of Tyumen (6, ul. Volodarskogo, 625003 Tyumen, Russia).

Сергей Владимирович Буреш – директор; <https://orcid.org/0000-0002-3899-6283>, sv@eetmn.ru, ООО «Электронный эксперт» (оф. 621, д. 59, ул. Республики, 625000 г. Тюмень,

Россия); **Sergey V. Buresh** – Managing Director; <https://orcid.org/0000-0002-3899-6283>, sv@eetmn.ru, ООО “Elektronnyi Ekspert” (Electronic Expert LLC) (ofis 621, 59, ul. Respubliki, 625000 Tyumen, Russia).

Заявленный вклад авторов: все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Contribution of the authors: the authors contributed equally to this article. The authors declare no conflicts of interests.

Статья поступила в редакцию 14.07.2021; одобрена после рецензирования 14.10.2021; принята к публикации 26.10.2021.

The article was submitted 14.07.2021; Approved after reviewing 14.10.2021; Accepted for publication 26.10.2021.