Автоматизированная система выдачи банковских гарантий на основе прогнозирования исполнения государственных контрактов

С.А. Корчагин[1], Е.П. Догадина[1], В.В. Мелентьев[2], П.В. Никитин[1], Д.В. Сердечный[1]

[1]Финансовый университет при Правительстве Российской Федерации, Москва
[2]Энгельсский технологический институт (филиал) Саратовского государственного технического университета имени Гагарина Ю.А.

Аннотация: Для информационного обеспечения поддержки принятия решений по выдаче банковских гарантий на исполнение контракта в сфере госзакупок, банкам важно получить исторически накопленную информацию по исполнению госконтрактов. Это необходимо для оценки возможности исполнения поставщиком его будущего контракта. Сделать это можно при помощи сбора и агрегирования сведений о контрактах из Единой информационной системы в сфере закупок. В работе предлагается использовать ИТ-технологии и анализ данных для построения прогноза исполнения контракта и выявления добросовестных поставщиков. В работе сформирована выборка первичных данных о контрактах для моделирования при помощи парсинга FTP-сервера Единой информационной системы в сфере закупок, а также произведена предобработка распарсенных данных для применения в моделях машинного обучения.
Ключевые слова: информационная система, анализ данных, государственный контракт, парсинг данных, машинное обучение.

Введение

В последние годы наблюдается тенденция роста денежного объема выданных банками гарантий, на которую значительное влияние оказывает увеличение выдач гарантий участникам госзакупок. В свою очередь, среди последних большую часть составляют гарантии, выданные в качестве обеспечения исполнения контракта по 44-ФЗ. В связи с наиболее широким распространением, именно этот подвид банковских гарантий выбран для дальнейшего изучения в работе. Благодаря гарантийному бизнесу, банки генерируют стабильную часть своих комиссионных доходов. По словам представителей банков [1], в большинстве случаев гарантии обеспечивают более высокую рентабельность, нежели кредитование. Также, по мнению представителей рейтинговых агентств [1], в последние годы гарантийный бизнес для многих стал альтернативой традиционному кредитованию,

поскольку часто этот бизнес имеет более низкий порог для входа на рынок и несет умеренные риски.

В последние годы практикуется выдача гарантий в удаленном режиме и сведение времени рассмотрения заявок к минимуму. Это помогает банкам получить конкурентное преимущество, которое способствует сохранению или увеличению их доходов.

Из сказанного следует, что сокращение времени рассмотрения заявок на банковскую гарантию является значимой и актуальной задачей, стоящей перед банками. Добиться ее решения возможно автоматизацией процесса принятия решения по заявке. На принятие решения по выдаче гарантии влияет вероятность ее раскрытия, которая, в случае гарантии на исполнение контракта, напрямую зависит от вероятности расторжения контракта. Поэтому, для принятия решения по заявке на выдачу гарантии этого вида следует учитывать прогноз результата исполнения контракта.

Целью работы является разработка автоматизированной системы прогнозирования результатов исполнения контрактов в сфере государственных закупок на основе парсинга данных FTP-сервера [2] Единой информационной системы в сфере закупок (ЕИС) и построения модели машинного обучения на основе распарсенных данных о контрактах.

Материалы и методы

На сайте портала государственных закупок в разделе «Форматы информационного взаимодействия по 44-ФЗ» [3] содержится документация, в которой описываются принципы формирования сведений в XML-документе на FTP-сервере ЕИС, порядок предоставления опубликованных документов, а также их структура. На FTP-сервер ЕИС госзакупок данные о муниципальном или государственном контракте выгружаются в виде упакованных в ZIP-архивы XML-файлов в кодировке UTF-8. Наименования ZIP-архивов и XML-файлов также записываются в кодировке UTF-8. Имя

каждого архива включает в себя указание типа выгрузки и промежутка времени, данные за который этот архив содержит.

Все заархивированные XML-файлы по 44-ФЗ находятся на FTP-сервере ftp://free:free@ftp.zakupki.gov.ru в каталоге fcs_regions, внутри которого применена следующая структура подкаталогов:

<Наименование региона>

contracts

currMonth

prevMonth

В результате для получения первичных данных был реализован парсер с использованием библиотек os, re, json, ftplib, zipfile, datetime, lxml. На вход скрипту парсера подается ссылка на каталог FTP-сервера с историческими файлами контрактов. В программном коде скрипта реализованы несколько этапов работы с данными: получение архивов данных с FTP-сервера ЕИС, обработка заархивированных XML-файлов с учетом изменений их версий схем, сохранение необходимой информации в виде csv-файлов с разделителем '\t'. Выходным результатом работы программы-парсера являются четыре csv-файла (contracts.csv, suppliers.csv, products.csv и contract_procs.csv) с данными по базе контрактов, содержащие в себе выборку первичных данных для моделирования. Далее производилась предобработка данных: необходимые преобразования типов данных; проверка контрактов, находящихся на стадии исполнения. В результате, вид итоговой выборки приведен на рис. 1.

В итоговой выборке были сформированы дополнительные признаки, характеризующие контракт. Этими признаками стали день даты заключения контракта (sign_day), месяц даты заключения контракта (sign_month), день даты окончания исполнения контракта (end_day), месяц даты окончания исполнения контракта (end_month), длительность исполнения контракта

(duration) и признак, содержащий подстроку из первых двух цифр ИНН заказчика (customer_inn_sub2).

price_rur	sign_day	end_day	end_month	duration	product_okpd2_sub2_cnt	sup_cust_same_reg	termination_result	cnt_end_90	sum_end_90	...	85
40000.00	1	31	12	364	1	1	0	45	1.388162e+08	...	0
30226.95	9	31	12	356	1	0	0	1	3.022695e+04	...	0
125545.00	5	31	12	360	1	1	0	1	1.255450e+05	...	0
229009.20	12	28	2	47	1	1	0	1	2.290092e+05	...	0
200000.00	12	31	12	353	1	1	0	3	2.292632e+06	...	0
...	...	...	...	...	...	...	...	...	...	...	...
52171.67	8	10	10	185	1	0	1	20	5.133523e+05	...	0
21959000.00	3	31	12	331	1	1	0	1	2.195900e+07	...	0
9771963.18	28	27	9	122	1	0	0	1	9.771963e+06	...	0
706689.96	11	31	12	234	1	0	0	24	1.698041e+06	...	0
372840.12	29	31	12	277	1	1	0	15	5.587407e+06	...	0

Рис. 1. – Выборка данных для моделирования

Признаки, которые не несли в себе информацию, используемую для дальнейшего построения выборки данных для моделирования, были исключены. Описание выборки представлено в таблице 1.

Полученный набор данных о контрактах был разделен на две части: обучающую и тестовую выборки. В обучающую выборку попало 77% данных, в тестовую – 33%.

Поскольку контракты с меткой 1 (расторгнутые) составляют всего 6,5% от общего объема, при разделении выборки на обучающую и тестовую была применена опция stratify модуля train_test_split [4] библиотеки scikit-learn. При использовании этой опции разделение производится таким образом, что внутри обучающей и тестовой выборок сохраняется соотношение классов.

После того, как выборка для обучения моделей была сформирована, на ее основе был обучен классификатор, входящий в библиотеку scikit-learn: RandomForestClassifier [5 – 7].

Таблица № 1

Выборка данных для моделирования

Наименование поля	Описание поля
price_rur	Цена контракта в рублевом эквиваленте
sign_day	День даты заключения контракта
end_day	День даты окончания исполнения контракта
end_month	Месяц даты окончания исполнения контракта
duration	Длительность исполнения контракта
product_okpd2_sub2_cnt	Количество уникальных первых разрядов кодов ОКПД2 на номер контракта
sup_cust_same_reg	Признак совпадения регионов поставщика и заказчика по контракту
termination_result	Результат исполнения контракта (1 – контракт расторгнут, 0 – контракт исполнен)
cnt_end_90	Количество исполненных контрактов поставщиком за последние 90 дней до заключения текущего контракта
sum_end_90	Сумма цен исполненных контрактов поставщиком за последние 90 дней до заключения текущего контракта
cnt_new_90	Количество заключенных контрактов поставщиком за последние 90 дней до заключения текущего контракта
sum_new_90	Сумма цен заключенных контрактов поставщиком за последние 90 дней до заключения текущего контракта
01…100	Фиктивная переменная, соответствующая первому разряду кода ОКПД2, равному 01…100

Случайный лес строится на базе деревьев решений, число которых определяется пользователем. Дерево решений базируется на модифицированном алгоритме CART [8].

Ответ классификатора случайного леса – это результат классификации набора деревьев решений, определенных «голосованием».

Положим, есть некоторое множество деревьев из леса решений, среди которых каждое относит объект x из множества X к одному из классов c, принадлежащих множеству классов Y. По деревьям применяется метод простого голосования, при котором подсчитывается число деревьев для

каждого класса c, относящих обозначенные объекты к этому классу. Количество деревьев в таком можно определить по формуле

$$G_c(x) = \frac{1}{T_c}\sum_{t=1}^{T_c} f_t(x), c \in Y, \quad (1)$$

где $f_t(x)$ – решающее дерево, T_c – общее количество деревьев в случайном лесу.

Итоговым ответом классификатора леса решений, считается класс, который определили большинством деревьев из леса: $a(x) = \arg\max_{c \in Y} G_c(x)$. Основой качественного распознавания объектов при этом подходе служит независимость ошибок классификации [9].

Результаты работы

В данной работе сначала обучение классификатора производилось с параметрами по умолчанию. Затем для улучшения качества работы обученных методов, с помощью модуля GridSearchCV [10] на обучающей выборке были подобраны параметры, с которыми метод показал более высокое значение метрики качества. GridSearchCV принимает на вход модель и различные значения ее параметров для подбора (сетку параметров). Модуль производит оптимизацию параметров за счет подсчета метрики качества и выявления наилучшего значения. В результате работы метода, выбирается ряд параметров с наилучшим значением метрики качества или наименьшей ошибкой, если она задана.

Для оценки качества работы обученных моделей была выбрана метрика ROC-AUC, поскольку эта метрика устойчива в случае дисбаланса классов в выборке данных. Кривая ошибок представляет собой линию от точки с координатами (0,0) до точки с координатами (1,1), где по вертикальная ось задается параметром, называемым True Positive Rate (TPR), а горизонтальная – параметром, именуемым False Positive Rate (FPR).

TPR и FPR вычисляются следующим образом:

$$TPR = \frac{TP}{TP + FN}, \quad (2)$$

$$FPR = \frac{FP}{FP + TN}, \quad (3)$$

где TP – количество объектов, принадлежность которых к классу с меткой 1 была определена алгоритмом классификации верно; FP – количество объектов, принадлежность которых к классу с меткой 1 была определена классификатором неверно; TN и FN – аналоги для класса с меткой 0.

До подбора параметров классификатор показал результат 0,78.

После подбора параметров результат несколько улучшился: 0,80.

Визуализация кривой ошибок для случайного леса после подбора параметров приведена на рис. 2. Визуализация была произведена при помощи библиотеки matplotlib и вычисление метрик roc_curve и auc, вычисленных при помощи библиотеки scikit-learn.

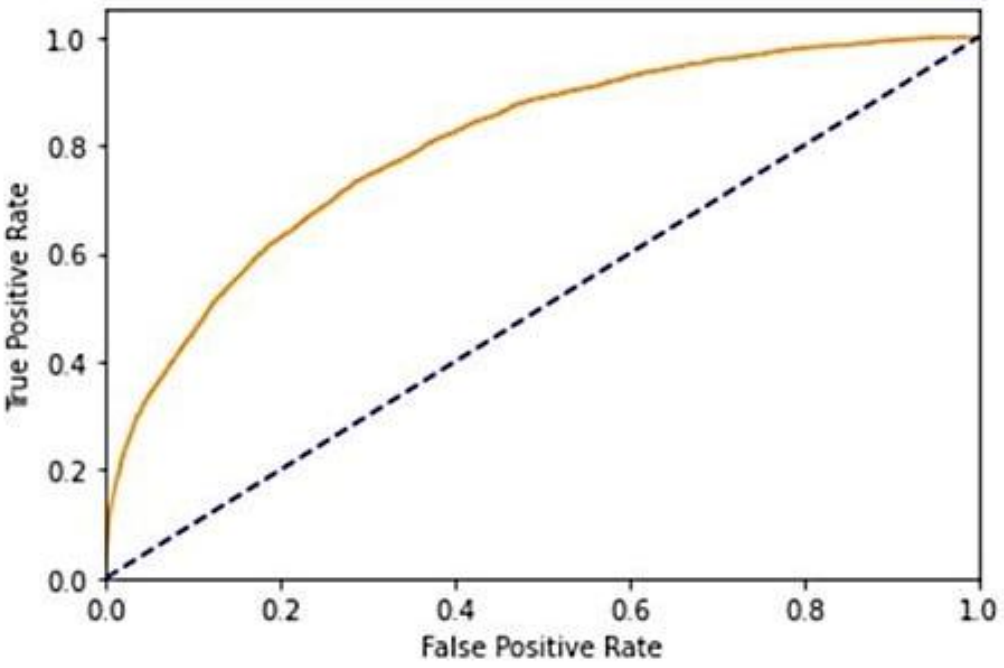

Рис. 2. – Кривая ROC-AUC

Заключение

В работе при помощи технологий парсинга был реализован сбор необходимой информации о контрактах с FTP-сервера портала ЕИС, на котором производится консолидация открытой информации по госзакупкам со всех торговых площадок.

Парсинг предоставляет возможность быстро собирать большие объемы данных, находящиеся в открытом доступе, структурируя их в удобную для анализа форму. Он имеет явные преимущества перед ручным способом сбора информации: технологии парсинга позволяют сократить время и снизить трудозатраты на выполнение сбора необходимых сведений.

После подготовки данных, при помощи методов машинного обучения, была реализована модель прогнозирования результата исполнения государственного контракта на основе случайного леса. Наилучшее качество прогнозирования показал классификатор случайного леса после подбора параметров. Величина метрики ROC-AUC для него составила 0,80. Данный классификатор был включен в систему поддержки принятия решения по выдаче банковских гарантий.

Таким образом, разработанная система включила в себя три модуля: модуль парсинга для получения первичной выборки данных, модуль предобработки данных для получения выборки данных для моделирования и модуль моделирования.

По итогам проделанной работы можно заключить, что разработанная система на основе технологий парсинга и методов машинного обучения позволит банкам быстрее проводить оценку результата исполнения госконтракта, применять эту оценку при принятии решения по предоставлению банковской гарантии на исполнение госконтаркта и предоставлять эти гарантии.

Статья подготовлена по результатам исследований, выполненных за счет бюджетных средств по государственному заданию Финуниверситета.

Литература

1. Клиентов взяли на гарантию. Развитию бизнеса помогли самоизоляция и дистанционные сервисы. kommersant.ru/doc/4451187 (дата обращения: 21.06.2023)

2. Официальный сайт Единой информационной системы в сфере закупок. zakupki.gov.ru/ (дата обращения: 15.06.2023)

3. Форматы информационного взаимодействия по 44-ФЗ. zakupki.gov.ru/epz/main/public/document/view.html?searchString=&sectionId=432&strictEqual=false (дата обращения: 15.06.2023)

4. Матюнина О. Е., Бадаев Ю. Л., Федосеев Н. А. Управление проектами с использованием методов машинного обучения // Современные тенденции развития науки и мирового сообщества в эпоху цифровизации: Сборник материалов XI Международной научно-практической конференции, Москва, 20 января 2023 года. – Москва: Общество с ограниченной ответственностью "Издательство АЛЕФ", 2023. – С. 336-342.

5. Yifter T. T., Razoumny Yu. N., Orlovsky A. V., Lobanov V. K. Monitoring the spread of Sosnowskyi's hogweed using a random forest machine learning algorithm in Google Earth Engine // Computer Research and Modeling. – 2022. – Vol. 14, No. 6. – P. 1357-1370. – DOI 10.20537/2076-7633-2022-14-6-1357-1370.

6. Гушанский С.М., Буглов В.Е. Разработка гибридной нейросети для классификации изображений // Инженерный вестник Дона, 2023, №1. URL: ivdon.ru/ru/magazine/archive/n1y2023/8150.

7. Горлатов Д.В. Машинное обучение прогнозных моделей на несбалансировнных данных по опасным астероидам // Инженерный вестник Дона, 2023, №5. URL: ivdon.ru/ru/magazine/archive/n5y2023/8394

8. Karminsky A. M. Comparative analysis of methods for forecasting bankruptcies of Russian construction companies // Business Informatics. – 2019. – Vol. 13, No. 3. – pp. 52-66. – DOI 10.17323/1998-0663.2019.3.52.66

9. Исаев Д. В. Стратегия поиска эффективного алгоритма машинного обучения на примере кредитного скоринга // Проблемы экономики и юридической практики. – 2020. – Т. 16, № 6. – С. 132-138.

10. Судаков В. А. Методы машинного обучения при расчёте скоринга клиентов банка // Международный журнал информационных технологий и энергоэффективности. – 2023. – Т. 8, № 3(29). – С. 22-25.

References

1. Klientov vzyali na garantiyu. Razvitiyu biznesa pomogli samoizolyaciya i distancionny`e servisy`. [The clients were taken on the guarantee. Self-isolation and remote services helped business development]. URL: kommersant.ru/doc/4451187 (date accessed: 21.06.2023).

2. Oficial`ny`j sajt Edinoj informacionnoj sistemy` v sfere zakupok. [Official website of the Unified Information System for Procurement]. URL: zakupki.gov.ru/ (date accessed: 15.06.2023)

3. Formaty` informacionnogo vzaimodejstviya po 44-FZ. [Formats of information interaction under 44-FZ]. URL: zakupki.gov.ru/epz/main/public/document/view.html?searchString=&sectionId=432&strictEqual=false (date accessed: 15.06.2023)

4. Matyunina O. E., Badaev Yu. L., Fedoseev N. A. Sovremenny`e tendencii razvitiya nauki i mirovogo soobshhestva v e`poxu cifrovizacii [Modern trends in the development of science and the world community in the era of digitalization]. Sbornik materialov XI Mezhdunarodnoj nauchno-prakticheskoj konferencii, Moskva, 20 yanvarya 2023 goda. Moskva: Obshhestvo s ogranichennoj otvetstvennost`yu "Izdatel`stvo ALEF"", 2023. pp. 336-342.

5. Yifter T. T., Razoumny Yu. N., Orlovsky A. V., Lobanov V. K. Komp`yuterny`e issledovaniya i modelirovanie, 2022. Vol. 14, №6. pp. 1357-1370.

6. Gushanskij S.M., Buglov V.E. Inzhenernyj vestnik Dona, 2023, №1. URL: ivdon.ru/ru/magazine/archive/n1y2023/8150.

7. Gorlatov D.V. Inzhenernyj vestnik Dona, 2023, №5. URL: ivdon.ru/ru/magazine/archive/n5y2023/8394

8. Karminsky A. M. Biznes informatika, 2019. Vol. 13, №3. pp. 52-66. DOI 10.17323/1998-0663.2019.3.52.66

9. Isaev D. V. Problemy` e`konomiki i yuridicheskoj praktiki, 2020. Vol. 16, № 6. pp. 132-138.

10. Sudakov V. A. Mezhdunarodny`j zhurnal informacionny`x texnologij i e`nergoe`ffektivnosti, 2023. Vol. 8, № 3(29). pp. 22-25.