УДК 004.8
DOI 10.17513/snt.39692

СИСТЕМА ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПО ВЫДАЧЕ БАНКОВСКИХ ГАРАНТИЙ НА ОСНОВЕ ПРОГНОЗИРОВАНИЯ ИСПОЛНЕНИЯ КОНТРАКТОВ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ И ТЕХНОЛОГИЙ ПАРСИНГА

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

[1]ФГОБУ ВО «Финансовый университет при Правительстве Российской Федерации», Москва, e-mail: epdogadina@fa.ru;

[2]ФГБОУ ВО «Энгельсский технологический институт (филиал) Саратовского государственного технического университета имени Гагарина Ю.А.», Энгельс, e-mail: melen2004@inbox.ru

Эффективная деятельность любого банка должна включать в себя сокращение времени рассмотрения заявок на банковскую гарантию. Эффективность работы банка является актуальной задачей, требующей постоянного совершенствования. Существуют различные способы автоматизации процесса принятия решения по заявкам. Но наиболее значимым параметром при принятии решения по выдаче гарантии является вероятность расторжения контракта. Получение прогноза вероятности расторжения контракта может быть достигнуто с помощью различных методов машинного обучения с учетом различных метрик. В статье предлагается построение автоматизированной системы и модели машинного обучения на основе распарсенных данных о контрактах, полученных на основе парсинга данных FTP-сервера Единой информационной системы в сфере закупок. Объектом исследования служит информационное обеспечение поддержки принятия решения по выдаче гарантии банка на исполнение контракта в сфере госзакупок. Предметом исследования выступает система прогнозирования результата исполнения госконтракта как инструмент информационного обеспечения. Практическая значимость работы состоит в увеличении скорости принятия решения по выдаче банковской гарантии исполнения обязательств по госконтракту посредством внедрения разработанной автоматизированной системы поддержки принятия решения по выдаче банковских гарантий.

Ключевые слова: информационная система, анализ данных, государственный контракт, парсинг данных, машинное обучение

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

DECISION SUPPORT SYSTEM ON ISSUANCE OF BANK GUARANTEES ON THE BASIS OF FORECASTING THE PERFORMANCE OF CONTRACTS USING MACHINE LEARNING METHODS AND PARSING TECHNOLOGIES

[1]Korchagin S.A., [1]Dogadina E.P., [2]Melentiev V.V., [1]Nikitin P.V., [1]Serdechny D.V.

[1]Financial University under the Government of the Russian Federation, Moscow, e-mail: epdogadina@fa.ru;

[2]Engels Technological Institute (branch) of Yu.A. Gagarin Saratov State Technical University, Engels, e-mail: epdogadina@fa.ru

The efficient operation of any bank should include the reduction of the time for processing applications for a bank guarantee. The efficiency of the bank's work is an urgent task that requires constant improvement. There are various ways to automate the decision-making process on applications. But the most significant parameter when making a decision on issuing a guarantee is the likelihood of contract termination. Obtaining a prediction of the probability of contract termination can be achieved using various machine learning methods, taking into account various metrics. The article proposes the construction of an automated system and a machine learning model based on parsed data on contracts obtained by parsing data from the FTP server of the Unified Procurement Information System. The object of the study is information support for decision-making on issuing a bank guarantee for the execution of a contract in the field of public procurement. The subject of the study is a system for predicting the result of the execution of a state contract as an information support tool. The practical significance of the work is to increase the speed of decision-making on the issuance of a bank guarantee for the fulfillment of obligations under a state contract through the introduction of a developed automated decision support system for issuing bank guarantees.

Keywords: information system, data analysis, government contract, data parsing, machine learning

The article was prepared based on the results of research carried out at the expense of budgetary funds under the state assignment of the Financial University.

Онлайн-гарантии сейчас выдаются только в секторе государственного заказа. Система госзакупок прозрачная и открытая. Информация об участниках тендеров является публичной, что дает банку возможность из открытого источника получить информацию о поставщике по контракту, об опыте его участия в подобных госзаказах, качестве их исполнения и, имея эти сведения, быстро провести нужную оценку [1].

Однако, несмотря на то что портал Единой информационной системы (ЕИС) в сфере закупок по своей сути представляет собой огромную открытую базу данных, она недостаточно проста в использовании и навигации. Поиск данных на портале имеет ограничения, связанные с количеством строк для выгрузки csv-файлов. Кроме этого, для сбора комплексной информации о контракте необходимо переключаться по нескольким вкладкам внутри страницы с карточкой контракта, что неудобно ни для ручной сборки данных, ни для написания алгоритма парсинга сайта.

Путем к решению проблемы трудоемкого извлечения данных с официального сайта Единой информационной системы служит FTP-сервер портала системы. На FTP-сервере содержится структурированная версия данных сайта в виде архивируемых файлов формата XML (текстовые файлы, которые используют пользовательские теги для описания структуры документа). Сведения на FTP-сервере содержат полную региональную выгрузку информации, опубликованной на официальном сайте ЕИС. В полную региональную выгрузку включаются сведения о контракте и его изменении, а также информация об исполнении / прекращении действия контракта.

Конкретный адрес FTP-сервера зависит от Федерального закона, согласно которому осуществляются закупки. Для 44-ФЗ адресом FTP-сервера является строка ftp://zakupki.gov.ru (логин и пароль: free).

Согласно официальной документации ЕИС [2], каждый календарный день на FTP-сервер выгружаются действующие редакции документов, опубликованные за предыдущий календарный день (от 00:00:00 до 24:00:00 предыдущего календарного дня до момента выполнения выгрузки). Также каждый календарный месяц выгружаются действующие редакции документов, опубликованные за предыдущий календарный месяц.

В ежедневной и ежемесячной выгрузках всегда присутствуют все типы документов, опубликованных за прошедший календарный день или календарный месяц соответственно. Если на момент формирования выгрузки за истекший период не было ни одного опубликованного документа какого-нибудь типа, то XML-файл с данным типом документов выгружается пустым.

Таким образом, производя загрузку данных ежедневной или ежемесячной выгрузки данных в зависимости от нужд организации с FTP-сервера в контур банка, можно получать актуальные данные по контрактной базе.

В литературе [3; 4] было обнаружено описание создания моделей, способных прогнозировать результат исполнения контракта на основе данных Единой информационной системы в сфере закупок. Однако цели разработки этих моделей не касались банковского сектора, а инструменты парсинга данных из ЕИС описаны не были. Кроме этого, в открытых источниках информации не было найдено сведений о том, какие методы моделирования и технологии сбора данных для построения своих моделей используют банки для автоматизации процесса принятия решения по выдаче банковской гарантии.

Таким образом, научная новизна и цель работы заключаются в:

– разработке автоматизированного способа получения комплексной информации о заключенных контрактах (на примере контрактов по 44-ФЗ) при помощи технологий парсинга XML-файлов, расположенных на FTP-сервере ЕИС, с использованием библиотеки lxml языка Python;

– обоснованном выборе наиболее применимого метода машинного обучения для применения в задаче прогнозирования результата исполнения госконтракта, решаемой для обеспечения поддержки принятия решения по выдаче банковской гарантии.

Материал и методы исследования

Для получения исторических данных о контрактах конкретного региона в виде csv-файлов, пригодных для автоматизированной обработки, необходимо произвести следующие действия:

1) извлечь архивы из каталога contracts, лежащего внутри каталога соответствующего региона на FTP-сервере;

2) распарсить XML-файлы в соответствии с их структурой из собранных архивов при помощи библиотеки lxml;

3) сохранить распарсенные данные XML-файлов в csv-файлы с разделителем ‘\t’ (применение этого разделителя более надежно, чем применение классической запятой).

Для реализации второго и третьего шага были более подробно изучены документы, которые помещаются внутрь каталога contracts. Из подкаталога contracts для работы интерес представляют следующие два типа документов: информация

о заключенном контракте (его изменении) и информация об исполнении (исполнение обязательств по предоставленной гарантии качества, расторжение, возврат переплаты по контракту, признание контракта недействительным) контракта.

В результате отработки скрипта парсинга было сформировано четыре csv-файла, которые включили в себя 556 314, 571 878, 791 953 и 2 119 707 строк соответственно.

Данные по контракту с течением времени могут изменяться. При их корректировке на FTP-сервере ЕИС госзакупок публикуются обновленные XML-файлы с информацией по контракту, при этом исторически загруженные XML-файлы также на FTP-сервере сохраняются. В результате такого дублирования публикаций сведений о контракте в распарсенных csv-файлах возникли случаи появления нескольких строк с данными, выгруженными на FTP-сервер в разное время, на один номер контракта.

Во избежание переобучения модели прогнозирования результата исполнения контракта на задвоенных исходных данных по контракту, приведенное дублирование информации по одному контракту было устранено путем отбора последних актуальных данных, выгруженных на FTP-сервер ЕИС.

Далее все четыре предобработанных csv-файла были соединены в одну таблицу при помощи объединения по номеру контракта, присутствующему в каждой таблице. В таблицу с результатом объединения был добавлен еще один признак, характеризующий контракт – признак того, что регионы поставщика и заказчика совпадают (sup_cust_same_reg), который был заполнен значением 1, если первые две цифры в ИНН поставщика и заказчика совпали, в противном случае было проставлено значение 0.

Также в результирующей таблице был сформирован целевой признак, обозначающий результат исполнения контракта (termination_result), который был рассчитан на основе поля termination_date: если дата расторжения контракта была указана, то контракт принимался расторгнутым и в поле termination_result проставлялось значение 1, иначе termination_result заполнялся значением 0.

На основе полученной таблицы были собраны и дополнительно добавлены в выборку четыре агрегата, характеризующие историю исполнения контрактов поставщиком.

Этими агрегатами стали количество исполненных контрактов поставщиком за последние 90 дней до даты заключения рассматриваемого контракта (cnt_end_90), сумма исполненных контрактов поставщиком за последние 90 дней до даты заключения рассматриваемого контракта (sum_end_90), количество и сумма новых контрактов, заключенных поставщиком за аналогичный период (cnt_new_90 и sum_new_90 соответственно).

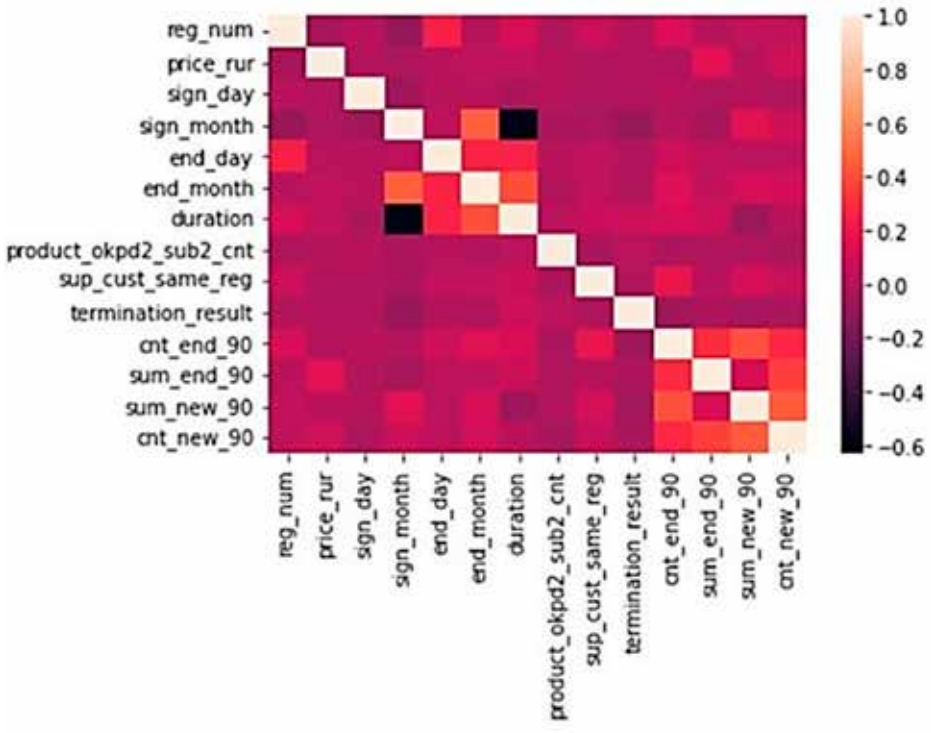

Рис. 1. Корреляционная матрица признаков

После составления итоговой выборки признаки в ней были проверены на корреляцию. Степень корреляции между атрибутами приведена на рисунке 1, сделанном при помощи библиотеки seaborn.

Выяснилось, что месяц даты заключения контракта сильно коррелирует с его длительностью, поэтому признак sign_month был исключен из дальнейшего рассмотрения. В результате размер выборки для моделирования сократился до 153 108 строк, уникальным идентификатором для которых является пара значений номера контракта и ИНН поставщика. Количество записей с расторгнутыми контрактами составило 9 901. На вход моделям машинного обучения будет подаваться 92 признака-предиктора и 1 целевой признак.

Для модели прогнозирования результата исполнения контракта в сфере госзакупок пригодны такие методы машинного обучения, как логистическая регрессия, дерево решений и случайный лес, поскольку эти методы хорошо интерпретируемы.

Модель логистической регрессии (логит-модель) относится к классу моделей бинарного выбора, в которых выходная переменная принимает только два возможных значения: $y \in \{0,1\}$. Вероятность наступления события, при котором $y = 1$, рассчитывается с помощью логистической функции, график которой приведен на рисунке 2:

$$p = P\{y = 1|x_1, x_2, ..., x_n\} = \frac{1}{1+e^{-z}}, \quad (1)$$

где p – вероятность наступления события, при котором $y = 1$ (принимает значения от 0 до 1), $z = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$ – интегральный показатель (принимает значения от –∞ до +∞), x_i – предикторы модели, β_i – параметры, которые требуется оценить, $i = 1, ..., n$ [5].

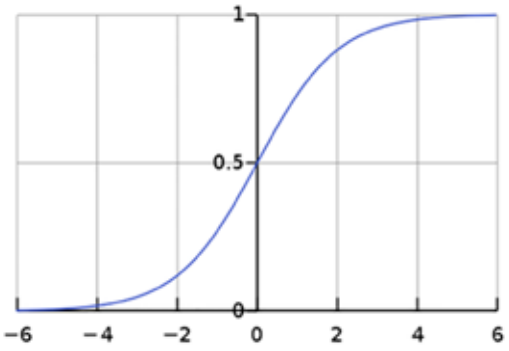

Рис. 2. График логистической функции

Для определения значения бинарной переменной y применяют порог отсечения. При значении p меньше выбранного порога прогнозируемое значение выходной переменной считается равным нулю, в противном случае – равным единице. Обычно, если отсутствуют априорные предположения о данных, пороговое значение принимают равным 0,5 [5].

Алгоритм CART (Classification And Regression Tree) – алгоритм построения бинарного дерева решений. В этом алгоритме каждый узел дерева имеет только двух потомков. На каждом шаге построения дерева формируемое в узле правило делит обучающую выборку на две части: часть, в которой правило выполняется, и часть, в которой оно не выполняется [6].

Библиотека scikit-learn для языка Python использует оптимизированную версию алгоритма CART. Визуализация работы этого алгоритма представлена на рисунке 3.

В процессе обучения дерева узлами дерева становятся наиболее информативные предикаты. Существует формула оценки информативности условия, размещенного в вершине дерева:

$$I = \frac{|L|}{|L|+|R|} \times H(L) + \frac{|R|}{|L|+|R|} \times H(R), \quad (2)$$

где L и R – множества примеров, попадающих в результате разбиения по условию в левый и правый узлы дерева соответственно.

Оценка же качества разбиения дерева производится по функциям $H(L)$ и $H(R)$, которые оцениваются с помощью индекса Джини, энтропии Шеннона или ошибки классификации.

Индекс Джини вычисляется по формуле

$$H(S) = 1 - \sum_{k=1}^{K} p^2(k|S), \quad (3)$$

где $p(k|S)$ – доля экземпляров класса k в S, k – количество классов, S – некоторое множество обучающих объектов.

Энтропия Шеннона рассчитывается по формуле

$$H(S) = -\sum_{k=1}^{K} p0(k|S)\log_2\left(p(k|S)\right). \quad (4)$$

Ошибка классификации находится по формуле

$$H(S) = 1 - \max_{k \in K} p(k|S). \quad (5)$$

Случайные леса состоят из определенного пользователем числа деревьев решений, которые строятся с помощью модифицированного алгоритма CART.

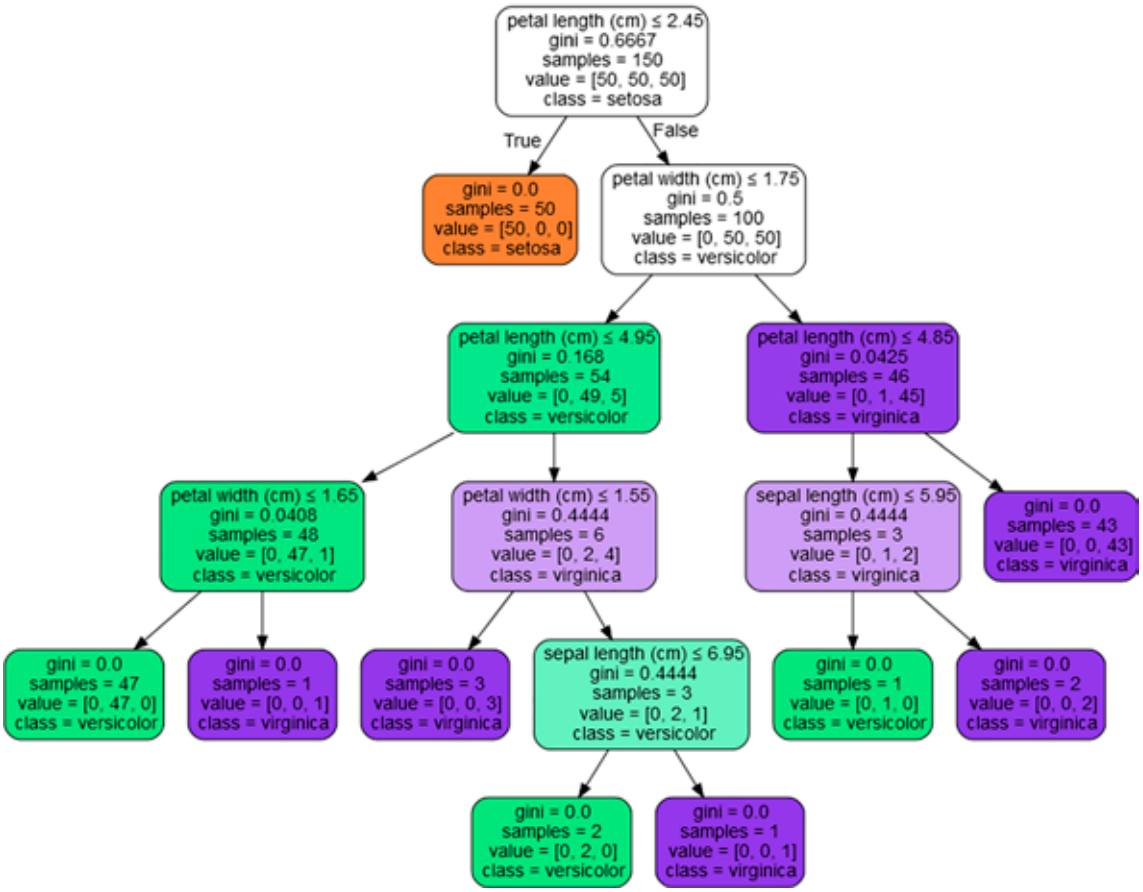

Рис. 3. Алгоритм CART

В алгоритме использованы два подхода: 1) каждое дерево обучается на собственной подвыборке исходных данных (bootstrapped data); 2) при построении деревьев решений используются различные подмножества факторов. То есть сначала строятся деревья решений, которые затем «голосуют» за принадлежность объекта к определенному классу [7–9].

Осуществим прогнозирование результата исполнения контракта в сфере госзакупок тремя представленными выше методами машинного обучения, а также определим наиболее эффективный метод машинного обучения для прогнозирования.

Результаты исследования и их обсуждение

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 [10] 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 было обучено три классификатора, также входящие в библиотеку scikit-learn: LogisticRegression, DecisionTreeClassifier и RandomForestClassifier.

Для улучшения качества работы обученных методов, с помощью модуля GridSearchCV на обучающей выборке были подобраны параметры, с которыми каждый из методов показал более высокое значение метрики качества. За основу метрики качества была взята метрика ROC-AUC. Определение лучшего набора параметров производилось при помощи 5-кратной кросс-валидации.

Визуализация работы кросс-валидации приведена на рисунке 4.

Значение метрики качества работы модели, найденное на кросс-валидации, представляет собой среднее значение вычисленных в цикле значений метрики на каждой из k итераций.

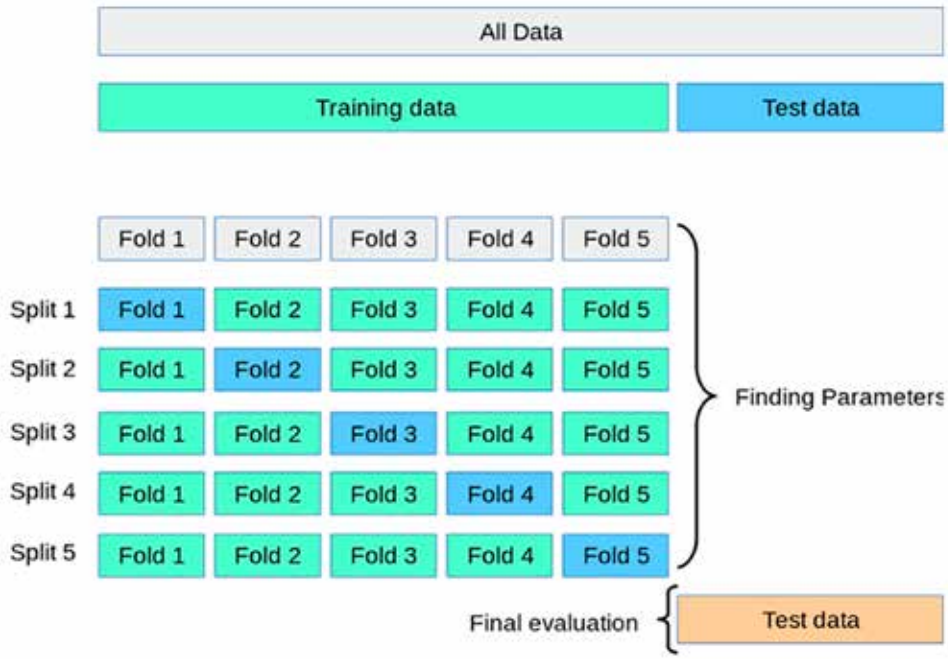

Рис. 4. Кросс-валидация

GridSearchCV принимает на вход модель и различные значения ее параметров для подбора (сетку параметров). Модуль осуществляет так называемый поиск по сетке: для каждого возможного сочетания значений параметров рассчитывается метрика качества, в конце выбирается такое сочетание параметров, при котором метрика качества получает наибольшее значение (или наименьшую ошибку, если она задана).

Для модели логистической регрессии при помощи поиска по сетке подбирались параметры C (обратный коэффициент регуляризации) и solver (алгоритм выбора параметров оптимизации, который определяет метод оптимизации для функции потерь логистической регрессии). Для дерева решений – criterion (критерий измерения качества разделения дерева) и max_depth (максимальная глубина дерева). Для случайного леса – n_estimators (количество деревьев в лесу), criterion (критерий измерения качества разделения дерева) и max_depth (максимальная глубина дерева).

В результате лучшими параметрами для логистической регрессии оказались C=0,001; solver='liblinear'. Для дерева решений – criterion='entropy'; max_depth=10. Для случайного леса – criterion='entropy'; max_depth=20; n_estimators=200.

На практике метрика ROC-AUC была вычислена для каждого из трех классификаторов на тестовой выборке данных после подбора их параметров при помощи модуля roc_auc_score библиотеки scikit-learn.

Классификаторы показали следующие результаты: для логистической регрессии площадь под кривой ошибок составила 0,48; для дерева решений – 0,73; для случайного леса – 0,80.

Визуализация кривой ошибок для случайного леса приведена на рисунке 5.

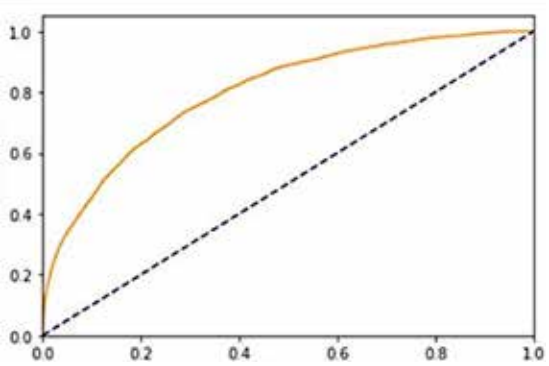

Рис. 5. Кривая ROC-AUC (ось x – False Positive Rate, ось y – True Positive Rate)

Итак, наиболее качественной моделью признается классификатор случайного леса, потому что имеет максимальное значение метрики ROC-AUC среди полученных значений.

Заключение

По итогам проделанной работы можно заключить, что разработанная автоматизированная система на базе методов машин-

ного обучения и парсинга данных сможет эффективно осуществлять прогнозирование исполнения госконтракта.

Увеличившаяся скорость принятия решения позволит банку удовлетворить потребности своих клиентов в скорости предоставления услуги, стать более привлекательным для них, тем самым сохранить или приумножить свою прибыль.

При этом процесс выдачи банковских гарантий в ускоренном режиме наиболее актуален для применения на большом количестве отдельных гарантий на небольшие суммы, поэтому такой подход имеет смысл применять для сегмента малого и среднего бизнеса.

Список литературы

1. 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.

2. Официальный сайт Единой информационной системы в сфере закупок. [Электронный ресурс]. URL: https://zakupki.gov.ru/ (дата обращения: 15.06.2023).

3. Голощапова Л.В., Чуприна О.И. Перспективы развития цифровизации банковских гарантий в сегменте государственных и муниципальных закупок // Вестник Южно-Российского государственного технического университета (НПИ). Серия: Социально-экономические науки. 2021. Т. 14, № 3. С. 149-153.

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

5. Пак К.И. Прогнозирование банкротства компаний на основе модели логистической регрессии // Экономика глазами молодых: материалы Международной конференции студентов, аспирантов и молодых ученых (г. Томск, 29–30 апреля 2021 г.). Томск: Издательский Дом Томского государственного университета, 2021. С. 113-116.

6. Осечкин А.И., Зубкова Л.Н. Методы классификации в задаче скоринговой оценки кредитоспособности заемщика // Вестник научных конференций. 2019. № 4-3 (44). С. 98-100.

7. Karminsky A.M. Comparative analysis of methods for forecasting bankruptcies of Russian construction companies // Business Informatics. 2019. Vol. 13, Is. 3. P. 52-66. DOI: 10.17323/1998-0663.2019.3.52.66.

8. Полин Я.А., Зудилова Т.В., Ананченко И.В., Войтюк Т.Е. Деревья решений в задачах классификации: особенности применения и методы повышения качества классификации // Современные наукоемкие технологии. 2020. № 9. С. 59-63.

9. Карминский А.М., Бурехин Р.Н. Сравнительный анализ методов прогнозирования банкротств российских строительных компаний // Бизнес-информатика. 2019. Т. 13, № 3. С. 52-66.

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.