ISSN 1561-8323 (Print)
ISSN 2524-2431 (Online)

ИНФОРМАТИКА

INFORMATICS

УДК 004.89
https://doi.org/10.29235/1561-8323-2025-69-2-101-108

Поступило в редакцию 10.02.2025
Received 10.02.2025

В. И. Бегунков, член-корреспондент М. Я. Ковалёв

Объединенный институт проблем информатики Национальной академии наук Беларуси, Минск, Республика Беларусь

КЛАССИФИКАЦИЯ ЗАЙМА С ИСПОЛЬЗОВАНИЕМ АЛГОРИТМА СЛУЧАЙНОГО ЛЕСА И СРАВНИТЕЛЬНЫЙ АНАЛИЗ С ДРУГИМИ КЛАССИФИКАТОРАМИ

Аннотация. Целью исследования является анализ использования алгоритма случайного леса для решения задачи классификации займа и проведение сравнительного анализа с результатами, полученными при использовании логистической регрессии, нейронной сети прямого распространения и глубокой нейронной сети прямого распространения. В результате исследований определены лучшее максимальное количество входных показателей и лучшее количество деревьев в ансамбле при использовании алгоритма случайного леса, исследовано воздействие альтернативного разбиения данных на тренировочный и тестовый наборы на точность прогнозирования модели при использовании алгоритма случайного леса. В заключение предложена стратегия решения задачи классификации займа на основе исследованных ранее классификаторов.

Ключевые слова: классификация займа, скоринг, машинное обучение, алгоритм случайного леса, сравнительный анализ классификаторов

Для цитирования. Бегунков, В. И. Классификация займа с использованием алгоритма случайного леса и сравнительный анализ с другими классификаторами / В. И. Бегунков, М. Я. Ковалёв // Доклады Национальной академии наук Беларуси. – 2025. – Т. 69, № 2. – С. 101–108. https://doi.org/10.29235/1561-8323-2025-69-2-101-108

Uladzimir I. Behunkou, Corresponding Member Mikhail Ya. Kovalyov

United Institute of Informatics Problems of the National Academy of Sciences of Belarus, Minsk, Republic of Belarus

LOAN CLASSIFICATION USING RANDOM FOREST ALGORITHM AND COMPARATIVE ANALYSIS WITH OTHER CLASSIFIERS

Abstract. The study aims to analyze the application of the random forest algorithm in addressing the loan classification issue. Furthermore, it intends to perform a comparative analysis by juxtaposing the outcomes with those derived from logistic regression, feedforward neural network, and deep feedforward neural network models. The research determined the ideal maximum number of input indicators and the ideal number of trees in the ensemble when utilizing the random forest algorithm. Additionally, it explored the impact of alternative data partitioning into training and test sets on the accuracy of model forecasting with the random forest algorithm. In conclusion, a strategy for addressing the loan classification issue using the classifiers studied has been proposed.

Keywords: loan classification, scoring, machine learning, random forest algorithm, comparative analysis of classifiers

For citation. Behunkou U. I., Kovalyov M. Ya. Loan classification using random forest algorithm and comparative analysis with other classifiers. *Doklady Natsional'noi akademii nauk Belarusi = Doklady of the National Academy of Sciences of Belarus*, 2025, vol. 69, no. 2, pp. 101–108 (in Russian). https://doi.org/10.29235/1561-8323-2025-69-2-101-108

Введение. В [1; 2] отмечались важность и актуальность поиска решения стоящей перед финансовыми институтами задачи классификации займа, которая представляется как бинарная с делением заемщиков на хороших (без дефолта) и плохих (дефолт). Так как в [3] рекомендуется

использовать алгоритм случайного леса (RF) в качестве эталона для сравнения при новых исследованиях алгоритмов классификации, то данный алгоритм также необходимо рассмотреть при решении обозначенной задачи классификации. При этом стоит отметить, что RF относится к классу однородных ансамблевых классификаторов.

Цель работы – исследование возможности эффективного использования алгоритма случайного леса для решения задачи классификации займа и сравнение результатов со значениями, полученными при использовании логистической регрессии, нейронной сети прямого распространения и глубокой нейронной сети прямого распространения.

Описание данных. Для решения задачи все используемые данные можно разделить на две группы: входные данные и выходные данные.

Входные данные. При проведении исследований с рассматриваемым алгоритмом используются исторические данные по выданным на платформе для кредитования LendingClub займам как описано в [1], которые состоят из 2260668 строк. Набор входных показателей и принцип преобразования входных данных аналогичны тем, которые были описаны ранее [1; 2]. Таким образом окончательный набор входных данных состоит из $m = 1221731$ позиций и $n = 73$ входных показателей.

Предполагается, что значения данных показателей были известны до принятия решения о выдаче соответствующего займа. Обозначим значение показателя j в займе i из исходного набора данных через элементы $x_j^{(i)}$ матрицы X размером m на n, где $i = 1, \ldots, m$, $j = 1, \ldots, n$. Определим через x_j столбец матрицы X, а через $x^{(i)}$ – строку матрицы X, которая содержит значения независимых показателей в отдельной позиции (займе) i набора данных.

Еще в качестве исходных данных используются целевые значения $y^{(i)}$ (итоговый результат по займам, где $i = 1, \ldots, m$), которые определены в поле loan_status исходного набора данных и могут быть также представлены в виде вектора Y. Показатель $y^{(i)}$ принимает два значения:

1. Возвратный займ (со значением Fully Paid). Данные займы были погашены. Соответствует значению $y^{(i)} = 1$.

2. Невозвратный займ (Charged Off или Default). Кредиты, по которым был объявлен дефолт или погашение займа просрочено более чем на 180 дней. Соответствует значению $y^{(i)} = 0$.

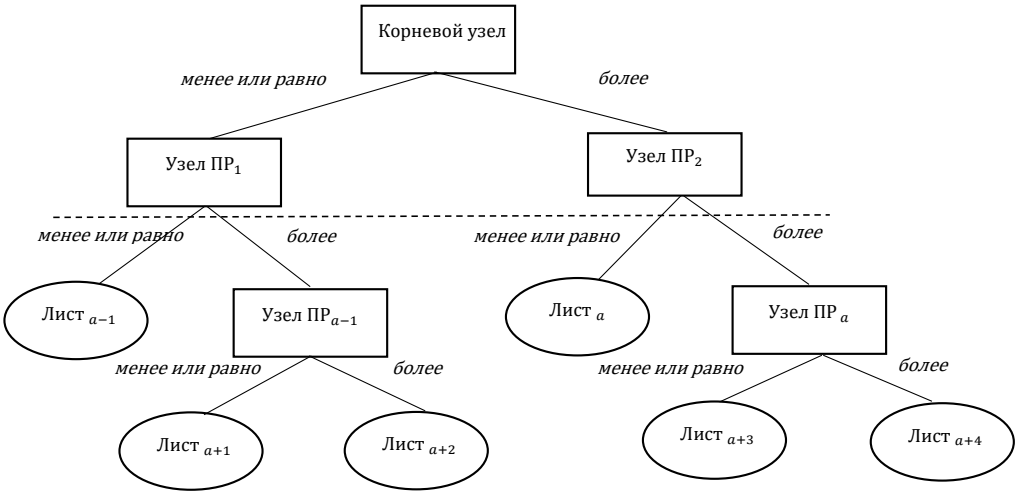
Займы со значениями Current, In Grace period, Late (16–30 days) и Late (31–120 days) исключаются из анализа, так как однозначно нельзя понять, были такие кредиты возвратными или невозвратными.

Выходные данные. Выходными данными изучаемой бинарной задачи классификации (т. е. определения нового кредита как возвратного или невозвратного) являются величины $\hat{y}^{(i)} \in \{0, 1\}$, где 1 соответствует возвратному, а 0 – невозвратному займу i, $i \in \{1, \ldots, m\}$.

Постановка задачи при использовании деревьев решений и метода случайного леса. Одним из вариантов решения задачи классификации займа является использование деревьев решений [4]. Данный метод контролируемого машинного обучения является весьма популярным при решении задач классификации. Как и при использовании других методов, при данном подходе также используется набор данных X с элементами $x_j^{(i)}$ и вектор Y с целевыми значениями $y^{(i)}$, как отмечено в описании данных. При этом обучение с помощью данного метода начинается с корня дерева и продолжается вниз на каждом из узлов как отражено на рисунке.

Построение данной модели осуществляется рекурсивно следующим образом [4]:

1. Для корневого узла выбирается входной показатель j и определяется пороговое значение этого показателя x_j^0, которые наиболее уменьшают (улучшают) неоднородность при разделении исходного тренировочного набора данных на два подмножества: позиции со значениями больше порогового значения обрабатываются далее по одной ветви дерева, а остальные позиции по другой ветви дерева. В результате, данные в каждом из дочерних подмножеств становятся менее неоднородными (т. е. в одном подмножестве начинают доминировать «возвратные» займы, а в другом «невозвратные»). Для этого, согласно алгоритму CART [5], для всех входных показателей j по отдельности определяется величина уменьшения примеси с помощью примеси Джини или энтропии [6]. Для текущего узла выбирается входной показатель и его пороговое значение с наибольшей величиной уменьшения примеси (неоднородности данных).



Дерево решений

Decision tree

2. Далее, для каждого узла принятия решения повторяется алгоритм аналогично шагу 1, но на основе данных, которые ранее были разделены для данной ветви дерева. Таким образом, для каждого узла выбирается входной показатель j и определяется пороговое значение данного показателя x_j^0, которые обеспечивают наибольшую величину уменьшения примеси (на базе примеси Джини или энтропии). Стоит отметить, что для разных узлов дерева может быть выбран одинаковый либо различные показатели, а также одинаковое либо разное пороговое значение.

3. Построение дерева прекращается на отдельном листовом узле, если количество позиций для разделения менее 2 или если все позиции в каждом из листьев принадлежат одному классу («возвратный» или «невозвратный») или если более не происходит улучшение разделения данных (величина уменьшения примеси меньше или равна 0). При этом данный терминальный узел (лист) прогнозирует (определяет) категорию, которая доминирует в этом узле, т. е. если в множестве позиций этого листа большую долю занимают позиции со значением $y^{(i)} = 1$, то терминальному узлу присваивается категория займа «возвратный», а при $y^{(i)} = 0$ – «невозвратный». Также во всех листах определяются вероятности каждой из категорий.

Недостатком метода деревьев решений является то, что построенная с помощью данного метода модель часто показывает хорошие результаты прогнозирования на тренировочных данных, но точность существенно снижается на тестовых значениях, т. е. модель подвержена переобучению. В таком случае применяется более сложный подход для решения задачи классификации займа: алгоритм случайного леса [7]. Он представляет собой множество отдельных деревьев решений, которые в дальнейшем агрегируются с целью получения более точной модели классификации. В то время как отдельные деревья решений подвержены переобучению, при использовании ансамбля из множества отдельных решений можно усреднить индивидуальные ошибки деревьев и, тем самым, снизить риск переобучения. При этом важно, чтобы отдельные деревья отличались друг от друга. Это достигается внедрением случайной вариации при построении отдельных деревьев решений двумя способами:

1. Для построения каждого из деревьев из исходного тренировочного набора данных формируется выборочная совокупность объектов с повторениями, которая равна по количеству позиций с исходным тренировочным набором данных, но при этом отдельные позиции могут отсутствовать, а некоторые повторяться несколько раз.

2. При построении дерева выбирается случайное количество входных показателей из множества всех n ($n = 73$) показателей для каждого из отдельных деревьев решений.

В дополнение, при построении модели случайного леса необходимо определить количество отдельных деревьев, на основе которых строится модель. Обычно, увеличение количества отдельных деревьев ведет к увеличению точности прогнозирования, но при этом повышается сто-

имость компьютерных вычислений с точки зрения увеличения времени вычислений и использования оперативной памяти.

После того как модель, основанная на методе случайного леса, обучена, она осуществляет прогнозирование значения следующим образом: вначале осуществляется прогнозирование категории (в текущей задаче $\hat{y}^{(i)} = 1$ или 0) на уровне каждого дерева в конкретном листе, а далее на основе спрогнозированных категорий и их вероятностей, рассчитанных при обучении в отдельном листе для каждого из деревьев в ансамбле, определяется средняя оценка вероятности по всем деревьям. Категория с наибольшей вероятностью является прогнозируемым значением для конкретного случая (займа).

Для сравнения эффективности решения задачи классификации займа с использованием дерева решений и случайного леса с рассмотренными ранее методами используется одинаковый набор входных данных с разделением на тренировочный и тестовый наборы данных в пропорции 70 на 30 %. Далее определяется лучшее (с точки зрения точности прогнозирования) разделение входных данных на тренировочный и тестовый наборы. Однако стоит отметить, что при данных исследованиях входные данные не нормализованы, так как при использовании методов деревьев решений и случайного леса отсутствует необходимость в такой нормализации [7].

После этого вычисляются следующие метрики, которые определяют качество прогнозирования: Accuracy (A), Precision (P), Recall (R) и мера F_1. В завершение осуществляется сравнение результатов со значениями, которые были получены при использовании логистической регрессии, нейронной сети прямого распространения и глубокой нейронной сети прямого распространения.

Классификация займов с использованием деревьев решений. Для анализа эффективности решения задачи классификации займа с использованием деревьев решений рассматриваются деревья с оценочной функцией на основе примеси Джини и энтропии без ограничений на глубину дерева решений. Для расчетов используется класс DecisionTreeClassifier[1] со значениями по умолчанию, кроме параметра criterion, которому было присвоено значение «entropy» при использовании энтропии в качестве оценочной функции. По итогу проведения эксперимента были получены результаты, отраженные в табл. 1 и 2.

Т а б л и ц а 1. **Результаты исследования при использовании метода деревьев решений на основе примеси Джини и энтропии**

T a b l e 1. **Results of the study using the decision trees method based on Gini impurity and entropy**

Результат Result	Примесь Джини Gini impurity	Энтропия Entropy
Длительность обучения алгоритма (с)	57,19	63,81
Accuracy training (%)	100	100
Accuracy testing (%)	70,23	70,76

Т а б л и ц а 2. **Ключевые метрики при использовании метода деревьев решений на основе примеси Джини и энтропии**

T a b l e 2. **Key metrics when using the decision trees method based on Gini impurity and entropy**

Класс Class	Примесь Джини Gini impurity			Энтропия Entropy		
	Precision	Recall	Мера F_1 Measure F_1	Precision	Recall	Мера F_1 Measure F_1
Невозвратные займы, %	28,25	30,43	29,30	28,77	29,95	29,35
Возвратные займы, %	81,96	80,35	81,14	82,00	81,14	81,57
Средневзвешенное, %	71,07	70,23	70,63	71,21	70,77	70,98

[1] DecisionTreeClassifier // Sklearn Tree. – URL: https://scikit-learn.org/dev/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier (date of access: 01.10.2024).

Из полученных результатов можно сделать вывод, что использование энтропии как оценочной функции является лучшим выбором для оценочной функции, так как при таком варианте коэффициент эффективности A и F_1-мера были выше, чем при использовании примеси Джини. Поэтому в дальнейших исследованиях как основа оценочной функции используется энтропия. При этом стоит отметить, что в двух вариантах полученная модель была подвержена переобучению, так как коэффициент эффективности на тренировочных данных был равен 1, а на тестовых – около 0,7. Для решения данной проблемы, как отмечено в постановке задачи, целесообразно применять алгоритм случайного леса.

Классификация займов с использованием алгоритма случайного леса с лучшим максимальным количеством входных показателей. Так как алгоритм случайного леса реализуется на основе множества деревьев решений, его применение позволяет уменьшить проблему переобучения, что приводит к более точному и стабильному прогнозированию категории займа. Это достигается с помощью двух особенностей данного метода: для обучения каждого дерева решений в ансамбле определяется набор данных с замещениями (т. е. некоторые позиции исходного набора данных из m позиций могут не попасть в выборку, а иные попасть несколько раз), а также для каждого дерева используется выбранное случайным образом подмножество входных показателей из множества всех n (n = 73) входных показателей. Таким образом достигается уникальное формирование отдельного дерева решений в ансамбле. В результате ансамбль состоит из отличающихся друг от друга деревьев, что позитивно влияет на устранения переобучения. При выборе произвольным образом только одного из входных показателей для каждого из деревьев в случайном лесу приведет к тому, что все деревья окажутся очень разные, и в таком случае модель, скорее всего, не сможет определить взаимосвязь между входными показателями и выходной переменной с высокой точностью как на тренировочном, так и на тестовом наборе данных. С другой стороны, при выборе максимального количества входных показателей n все деревья в случайном лесу могут быть похожи, а полученная модель становится ближе к описанной ранее с сопутствующей проблемой переобучения. Поэтому необходимо провести исследование и определить лучшее максимальное количество входных показателей, при котором коэффициент эффективности A максимален на тестовых данных. В результате исследования были получены результаты, отраженные в табл. 3 и 4.

Из результатов видно, что применение метода случайного леса привело к улучшению коэффициента эффективности A на 9,40 процентных пункта по сравнению с использованием отдельного дерева решений. При этом лучшее максимальное количество входных показателей, которое используется в дальнейших экспериментах, равняется 23.

Классификация займов с использованием алгоритма случайного леса с лучшим количеством деревьев в ансамбле. При использовании алгоритма случайного леса по умолчанию[1] используется количество деревьев, равное 100. Однако данное количество может быть недостаточным

Т а б л и ц а 3. **Результаты исследования при использовании метода деревьев решений с лучшим количеством входных показателей, лучшим количеством деревьев в ансамбле и использованием альтернативного разделения данных**

T a b l e 3. **Results of the study using the decision trees method with the ideal number of input features, the ideal number of trees in the ensemble and alternative data separation**

Результат Result	С лучшим количеством входных показателей With the ideal number of input features	С лучшим количеством деревьев With the ideal number of trees	С использованием альтернативного разделения данных With alternative data separation
Длительность обучения алгоритма (с)	62,28	1057,95	1519,76
Лучшее максимальное количество входных показателей	23	23	23
Лучшее количество деревьев в ансамбле	–	1950	1950
Доля лучшего тренировочного набора данных (%)	–	–	94
Accuracy training (%)	99,99	100	100
Accuracy testing (%)	80,16	80,27	80,55

[1] RandomForestClassifier // Sklearn Ensemble. – URL: https://scikit-learn.org/dev/modules/generated/sklearn.ensemble.RandomForestClassifier.html#randomforestclassifier (date of access: 15.10.2024).

Т а б л и ц а 4. **Ключевые метрики при использовании метода деревьев решений с лучшим количеством входных показателей, лучшим количеством деревьев в ансамбле и использованием альтернативного разделения данных**

T a b l e 4. **Key metrics when using the decision trees method with the ideal number of input features, the ideal number of trees in the ensemble and alternative data separation**

Класс Class	С лучшим количеством входных показателей With the ideal number of input features			С лучшим количеством деревьев With the ideal number of trees			С использованием альтернативного разделения данных With alternative data separation		
	Precision	Recall	Мера F_1 Measure F_1	Precision	Recall	Мера F_1 Measure F_1	Precision	Recall	Мера F_1 Measure F_1
Невозвратные займы, %	56,10	9,93	16,87	59,31	8,55	14,94	60,49	8,82	15,40
Возвратные займы, %	81,06	98,03	88,74	80,90	98,51	88,84	81,16	98,56	89,02
Средневзвешенное, %	76,00	80,17	74,17	76,53	80,27	73,86	77,02	80,56	74,25

для поиска лучшего решения конкретной задачи. При этом использование слишком большого количества деревьев может не привести к улучшению коэффициента эффективности A, но при этом потребуется больше вычислительных мощностей. Поэтому необходимо определить лучшее количество деревьев в ансамбле при использовании метода случайного леса. При этом диапазон для количества деревьев составляет от 1 до 10000. С целью лучшего использования вычислительных мощностей данное исследование предпочтительно осуществить с помощью двухшагового подхода: на первом шаге предполагается выбрать лучшее количество деревьев из множества (5, 50, 500 и 5000), т. е. на основе логарифмической шкалы. По итогу реализации первого шага определяется лучший отрезок (например, при лучшем количестве деревьев равном 5, лучшим отрезком становится диапазон от 1 до 10, а при 50 – значения от 10 до 100 и так далее). На втором шаге исследуется лучший отрезок с шагом 1 при выбранном значении на первом шаге равном 5 или 50 и с шагом 10 во всех остальных случаях. Соответственно, этот подход первично определяет диапазон значений, в котором может находиться лучшее решение, а на втором шаге более детально его исследует. Полученные результаты отражены в табл. 3 и 4.

Как следует из полученных значений, 1950 является оптимальным количеством деревьев в ансамбле при задействованном методе поиска лучшего количества деревьев в ансамбле. При этом значение коэффициента эффективности улучшилось на 0,11 процентных пункта.

Применение альтернативного разделения данных на тренировочный и тестовый наборы при использовании алгоритма случайного леса. Как было видно в эксперименте при использовании глубокой нейронной сети, изменение процентного отношения тренировочных и тестовых данных положительно влияло на точность прогнозирования. Поэтому при применении метода случайного леса целесообразно провести аналогичный эксперимент. Для этого доля тренировочного набора данных также изменяется с 70 до 99 % с шагом 1 %. Соответственно, доля тестового набора данных уменьшается с 30 до 1 % с тем же шагом. Итоги исследования представлены в табл. 3 и 4.

Результат эксперимента показывает, что изменение доли тренировочного набора с 70 до 94 % привело к улучшению коэффициента эффективности A модели на основе деревьев решений в рамках рассматриваемой задачи до 80,55 %. Как отмечалось ранее при исследовании глубокой нейронной сети, выбор пропорции разбиения тренировочного и тестового наборов данных зависит от выбранного пользователем критерия для оптимизации, так как разные соотношения тренировочных и тестовых данных по-разному влияют на обозначенные в постановке задачи метрики.

Сравнение результатов при использовании логистической регрессии, нейронной сети прямого распространения, глубокой нейронной сети прямого распространения и алгоритма случайного леса для решения задачи классификации займа. Как следует из результатов исследования, при использовании алгоритма случайного леса наибольшее значение коэффициента эффективности A на тестовых данных было 80,55 %. Целесообразно сравнить полученные результаты со значениями, рассчитанными при использовании логистической регрессии [1], нейронной сети прямого распространения [2] и глубокой нейронной сети прямого распространения.

Из табл. 5 видно, что с помощью глубокой нейронной сети прямого распространения получено наивысшее (лучшее) значение коэффициента эффективности A. При этом важно подчеркнуть, что средняя длительность обучения глубокой нейронной сети прямого распространения существенно выше аналогичных величин при использовании других алгоритмов, указанных в табл. 5.

Т а б л и ц а 5. **Сравнение лучших результатов при применении логистической регрессии, нейронной сети прямого распространения, глубокой нейронной сети прямого распространения и алгоритма случайного леса**

T a b l e 5. **Comparison of the ideal results when applying logistic regression, feed-forward neural network, deep feed-forward neural network and random forest algorithm**

Результат Result	При алгоритме случайного леса With random forest algorithm	При глубокой нейронной сети прямого распространения With a deep feed-forward neural network	При нейронной сети прямого распространения With a feed-forward neural network	При логистической регрессии With logistic regression
Средняя длительность обучения (с)	1520	151681	17191	2063
Среднее значение стоимостной функции	Не применимо	0,4426	0,4435	0,4571
Accuracy training (%)	100	80,40	80,43	79,93
Accuracy testing (%)	80,56	80,65	80,32	80,04

Значения метрик F_1, P и R при лучшем коэффициенте эффективности A сравниваемых алгоритмов по невозвратным и возвратным займам, а также при расчете средневзвешенной величины отражены в табл. 6.

Т а б л и ц а 6. **Мера F_1, Precision и Recall метрики при применении логистической регрессии, нейронной сети прямого распространения, глубокой нейронной сети прямого распространения и алгоритма случайного леса**

T a b l e 6. **Measure F_1, Precision and Recall metrics when applying logistic regression, feed-forward neural network, deep feed-forward neural network and random forest algorithm**

Класс Class	Алгоритм случайного леса Random forest algorithm			Глубокая нейронная сеть прямого распространения Deep feed-forward neural network			Нейронная сеть прямого распространения Feed-forward neural network			Логистическая регрессия Logistic regression		
	Мера F_1 F_1 measure	Precision	Recall	Мера F_1 F_1 measure	Precision	Recall	Мера F_1 F_1 measure	Precision	Recall	Мера F_1 F_1 measure	Precision	Recall
Невозвратные займы, %	15,40	60,49	8,82	18,21	59,79	10,74	20,96	56,45	12,87	15,20	54,78	8,83
Возвратные займы, %	89,02	81,16	98,56	89,03	81,43	98,19	88,76	81,48	97,48	88,69	80,89	98,15
Средневзвешенное, %	74,25	77,02	80,56	74,82	77,09	80,65	75,02	76,41	80,33	73,79	75,60	80,04

Как следует из табл. 6, средневзвешенные значения метрик P и R оказались выше для глубокой нейронной сети прямого распространения при использовании в качестве критерия оптимизации коэффициента эффективности A. При этом, как подчеркивалось выше, результат может быть иным, если в качестве критерия оптимизации использовать другой критерий.

Заключение. В данном исследовании изучено использование алгоритма случайного леса для решения задачи классификации займа. Вначале было установлено, что поиск лучшего максимального количества входных показателей привел к улучшению прогнозирования модели. Также выявлено, что от количества деревьев в алгоритме случайного леса зависит точность прогнозирования, и целесообразно проводить поиск лучшего количества деревьев в ансамбле при решении конкретной задачи, что в итоге привело к улучшению прогнозируемых результатов. В завершение поиск альтернативного разделения исходных данных на тренировочный и тестовый наборы привел к увеличению коэффициента эффективности A на тестовых данных.

Из результатов сравнительного анализа следует, что итоговая точность прогнозирования при использовании глубокой сети прямого распространения оказалась выше полученной при использовании нейронной сети прямого распространения и логистической регрессии, рассмотренных в предыдущих исследованиях [1; 2], а также выше значений, полученных при использовании алгоритма случайного леса. Однако стоит отметить, что решение задачи классификации займа с использованием алгоритма случайного леса требует существенно меньше производительных ресурсов и времени, а точность прогнозирования лишь несколько уступает полученной с помощью глубокой нейронной сети прямого распространения и выше, чем точность, полученная при использовании как логистической регрессии, так и нейронной сети прямого распространения. Исходя из полученных результатов исследования, сформулируем лучшую стратегию решения задачи классификации займа: в начале следует использовать и исследовать алгоритм случайного леса, так как при существенно меньших требуемых ресурсах (вычислительных мощностей и времени обучения на обозначенные ранее) точность прогнозирования уступает лишь полученной с помощью глубокой нейронной сети. После того как разработан и функционирует инструмент классификации займа на основе алгоритма случайного леса, целесообразно продолжить исследования классификации займа с использованием глубокой нейронной сети прямого распространения, что в среднесрочной перспективе позволит улучшить точность прогнозирования.

Список использованных источников

1. Бегунков, В. И. Классификация займов с использованием логистической регрессии / В. И. Бегунков, М. Я. Ковалев // Информатика. – 2023. − Т. 20, № 1. – С. 55–74. https://doi.org/10.37661/1816-0301-2023-20-1-55-74

2. Бегунков, В. И. Классификация займа с использованием нейронной сети прямого распространения / В. И. Бегунков // Информатика. – 2024. − Т. 21, № 1. – С. 83–104. https://doi.org/10.37661/1816-0301-2024-21-1-83-104

3. Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research / S. Lessmann, B. Baesens, H.-V. Seow, L. C. Thomas // European Journal of Operational Research. – 2015. – Vol. 247, N 1. – P. 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

4. Quinlan, J. R. Induction of Decision Trees / J. R. Quinlan // Machine Learning. – 1986. – Vol. 1, N 1. – P. 81–106. https://doi.org/10.1007/bf00116251

5. Classification and Regression Trees / L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. – New York, 1984. – 368 p. https://doi.org/10.1201/9781315139470

6. Hastie, T. The elements of statistical learning: Data mining, inference, and prediction / T. Hastie, R. Tibshirani, J. Friedman. – New York, 2009. – 3d ed. – P. 308–310. https://doi.org/10.1007/978-0-387-84858-7

7. Breiman L. Random Forests / L. Breiman // Machine Learning. – 2001. – Vol. 45. – P. 5–32. https://doi.org/10.1023/a:1010933404324

References

1. Behunkou U. I., Kovalyov M. Y. Loan classification using logistic regression. *Informatics*, 2023, vol. 20, no. 1, pp. 55–74 (in Russian). https://doi.org/10.37661/1816-0301-2023-20-1-55-74

2. Behunkou U. I. Loan classification using a feed-forward neural network. *Informatics*, 2024, vol. 21, no. 1, pp. 83–104 (in Russian). https://doi.org/10.37661/1816-0301-2024-21-1-83-104

3. Lessmann S., Baesens B., Seow H.-V., Thomas L. C. Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. *European Journal of Operational Research*, 2015, vol. 247, no. 1, pp. 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

4. Quinlan J. R. Induction of Decision Trees. *Machine Learning*, 1986, vol. 1, pp. 81–106. https://doi.org/10.1007/bf00116251

5. Breiman L., Friedman J. H., Olshen R. A., Stone C. J. *Classification and Regression Trees*. New York, 1984. 368 p. https://doi.org/10.1201/9781315139470

6. Hastie T., Tibshirani R., Friedman J. *The elements of statistical learning: Data mining, inference, and prediction*. 3d ed. New York, 2009, pp. 308–310. https://doi.org/10.1007/978-0-387-84858-7

7. Breiman L. Random Forests. *Machine Learning*, 2001, vol. 45, pp. 5–32. https://doi.org/10.1023/a:1010933404324

Информация об авторах

Бегунков Владимир Иванович – магистр технических наук. Объединенный институт проблем информатики НАН Беларуси (ул. Сурганова, 6, 220012, Минск, Республика Беларусь). E-mail: vbegunkov@gmail.com.

Ковалёв Михаил Яковлевич – член-корреспондент, д-р физ.-мат. наук, профессор. Объединенный институт проблем информатики НАН Беларуси (ул. Сурганова, 6, 220012, Минск, Республика Беларусь). E-mail: kovalyov_my@newman.bas-net.by.

Information about the authors

Behunkou Uladzimir I. – Master of Sciences (Engineering). United Institute of Informatics Problems of the National Academy of Sciences of Belarus (6, Surganov Str., 220012, Minsk, Republic of Belarus). E-mail: vbegunkov@gmail.com.

Kovalyov Mikhail Y. – Corresponding Member, D. Sc. (Physics and Mathematics), Professor. United Institute of Informatics Problems of the National Academy of Sciences of Belarus (6, Surganov Str., 220012, Minsk, Republic of Belarus). E-mail: kovalyov_my@newman.bas-net.by.