DOI 10.37882/2223–2966.2021.06.22

# ОЦЕНКА ТОЧНОСТИ РАБОТЫ АНСАМБЛЕВЫХ АЛГОРИТМОВ ДЛЯ КРЕДИТНОГО СКОРИНГА

## ASSESSMENT OF THE ACCURACY OF THE ENSEMBLE ALGORITHMS FOR CREDIT SCORING

***I. Kanishchev***

*Summary.* The relevance of the development, implementation and use of a digital customer scoring system model for credit risk management is beyond doubt. Modern information technologies for the implementation of customer credit scoring allow you to concentrate on yourself most of the technical work on the collection and processing of initial data and the implementation of machine learning algorithms. The volume of banking information about clients will increase in the next few years, and the requirements for the quality and speed of its processing will become more stringent.

This study compares the performance of ensemble algorithms, i.e., random forest, XGBoost, LightGBM, CatBoost, and Stacking, in terms of area under the curve (AUC), Brier rating (BS), and model runtime. In addition, analysis of three popular basic classifiers, i.e. decision tree (DT), logistic regression (LR) and linear discriminant analysis (LDA), are considered benchmarks in credit scoring.

Experimental evidence shows that ensemble learning is better than basic classifiers. In addition, Stacking stands out from the rest of the models.

*Keywords:* ensemble methods, staking, credit scoring, classification.

***Канищев Илья Сергеевич***
*Аспирант, Вятский государственный университет, Киров*
*kanishchev.ilya@gmail.com*

*Аннотация.* Актуальность разработки, внедрения и использования модели цифровой системы скоринга клиентов для управления кредитными рисками сегодня не вызывает сомнения. Современные информационные технологии для реализации кредитного скоринга клиентов позволяют сконцентрировать на себе большую часть технической работы по сбору и обработке исходных данных и реализации алгоритмов машинного обучения. Объемы банковской информации о клиентах в ближайшие несколько лет будут возрастать, а требования к качеству и скорости ее обработки ужесточаться.

В этом исследовании проводится сравнительная оценка производительности ансамблевых алгоритмов, то есть случайного леса, XGBoost, LightGBM, CatBoost и Stacking, с точки зрения площади под кривой (AUC), рейтинга Бриера (BS) и времени работы модели. Кроме того, анализ трёх популярных базовых классификаторов, то есть, дерево решений (DT), логистическая регрессия (LR) и линейный дискриминантный анализ (LDA), которые считаются эталонами в кредитном скоринге.

Экспериментальные данные показывают, что ансамблевое обучение лучше, чем базовые классификаторы. Кроме того, Stacking выделяется среди остальных моделей.

*Ключевые слова:* ансамблевые методы, стекинг, кредитный скоринг, классификация.

## Введение

Качественный и количественный рост рынка кредитования и банковского сектора ставит перед финансовыми институтами задачу улучшения качества обслуживания, скорости обработки кредитных заявок и снижения издержек. Ключевыми задачами являются разработка адекватных и современных инструментов оценки и управления рисками кредитных и финансовых организаций, а также изучение закономерностей, которые характеризуют современные финансовые рынки.

В настоящее время на данных рынках высокий спрос, что порождает и высокую конкуренцию на рынке кредитования. Банки и другие кредитные институты столкнулись с проблемой обработки и анализа больших объемов информации в сжатые сроки.

В общем кредитный скоринг представляет собой технологию, которая позволяет банкам и финансовым институтам решить вопрос о предоставлении кредита клиенту с учетом его характеристик, таких как возраст, пол, образование, семейное положение, доход и др.

Со времени новаторской работы Бивера [1] и Альтмана [2] кредитный скоринг стал основным предметом исследований ученых и финансовых организаций. Впоследствии многие типы моделей кредитного скоринга были предложены и разработаны с использованием статистических методов, таких как линейный дискриминантный анализ (LDA) и логистическая регрессия (LR) [3, 4, 5, 6]. Однако в современном мире объем данных стремительно растет, когда речь идет об огромных объемах данных, эластичность классических моделей статистического анализа низкая. В результате некото-

рые допущения в этих моделях не могут быть установлены, что, в свою очередь, влияет на точность прогнозов. С прорывом технологий, таких как нейронные сети (NN) [7, 8], случайный лес (RF) [9] и Naïve Bayes (NB) [10] может дать такие же или лучшие результаты по сравнению со статистическими моделями. Однако LDA и LR по-прежнему имеют широкий спектр применения из-за их высокой точности и простоты интерпретации.

Машинное обучение относится к категории искусственного интеллекта и является одним из наиболее эффективных методов интеллектуального анализа данных, который может предоставить аналитикам более продуктивную информацию об использовании больших данных [11]. Модели машинного обучения обычно подразделяются на индивидуальное машинное обучение (NN и LR), ансамблевое обучение (бэггинг, бустинг и стекинг) и интегрированное ансамблевое машинное обучение (RS-бустинг и мульти-бустинг) [12]. Подходы к ансамблевому обучению считаются современным решением для многих задач машинного обучения [13]. После Чена и Гестрина [14], предложенный XGBoost в 2016 году, ансамблевое обучение, то есть XGBoost и LightGBM, стало выигрышной стратегией для широкого спектра задач. Соответственно, все больше и больше ученых вводят модели ансамблевого обучения для решения задач кредитной оценки [15, 16, 17, 18]. Ансамблевое обучение — это метод достижения отличных результатов путем создания и объединения нескольких базовых учащихся с определенными стратегиями. Его можно использовать для решения задач классификации, задач регрессии, выбора признаков, обнаружения выбросов и т.д.

## 1. Выбор ансамблевых методов

Ансамблевые методы — это парадигма машинного обучения, в которой несколько моделей (часто называемых слабыми учениками или базовыми моделями) обучаются для решения одной и той же проблемы и объединяются для повышения производительности.

Основная гипотеза состоит в том, что, если правильно объединить слабых учеников, то получится более точные и/или надежные модели.

В задачах классификации простейший пример ансамбля — комитет большинства:

$$a(x) = mode(b_1(x), \dots, b_n(x))$$

В отличие от голосования большинством, Stacking также является методом стратегий интеграции, который объединяет учащихся низкого уровня с алгоритмом обучения высокого уровня (т.е. метаклассификатором) [19]. Однако исследований по Stacking в области кредитного скоринга мало [20].

В теории машинного обучении — метод построения ансамбля моделей, в котором обучение базовых моделей производится параллельно [21]. При этом каждая модель обучается на отдельной выборке, сформированной из исходного набора данных с помощью алгоритма бутстрэпа. Выход ансамбля определяется путем усреднения выходов базовых моделей.

Метод позволяет улучшить точность и устойчивость работы алгоритмов машинного обучения, уменьшить дисперсию ошибки и уменьшить эффект переобучения. Хотя изначально метод был разработан для классификаторов на основе деревьев решений, он может использоваться для любых видов моделей.

Метод был предложен Лео Брейманом в 1994 году для улучшения точности классификаторов на основе деревьев решений.

Случайный лес — один из примеров объединения классификаторов в ансамбль. Итоговый классификатор случайного леса, состоящего из $N$ деревьев на основе обучающей выборки $X$:

$$a(x) = \frac{1}{N}\sum_{i=1}^{N} t_i(x)$$

Для задачи классификации выбирается решение по большинству результатов, выданных классификаторами. Таким образом, случайный лес — бэггинг над решающими деревьями, при обучении которых для каждого разбиения признаки выбираются из некоторого случайного подмножества признаков.

RF считается одним из лучших алгоритмов в настоящее время, который не чувствителен к мультиколлинеарности, а результаты относительно устойчивы к отсутствующим и несбалансированным данным [22].

По сравнению с параллельным построением базовых учащихся в Bagging, Boosting последовательно устанавливает набор базовых классификаторов, основная идея которых состоит в том, что, во-первых, на обучающей выборке создается слабый классификатор.

В соответствии с результатом классификатора каждой выборке должен быть присвоен вес в обучающем наборе, и вес будет относительно небольшим, если выборка классифицирована правильно; в противном случае ему будет присвоено относительно большое число. Затем, чтобы точно классифицировать выборки с боль-

шим весом, весь вес каждой выборки рассматривается для построения второго слабого классификатора. Повторяя этот процесс, будет создано несколько слабых классификаторов, чтобы добиться лучших результатов классификации.

Параметры модели каждого слабого классификатора получаются путем минимизации функции потерь предыдущей модели на обучающей выборке.

Окончательная модель, полученная с помощью алгоритма Boosting, представляет собой линейную комбинацию нескольких базовых классификаторов, взвешенных по их собственным результатам.

Технология eXtreme Gradient Boosting (XGBoost) была предложена Ченом и Гестрином в 2016 году [14]. Она предлагает множество улучшений по сравнению с традиционными алгоритмами повышения градиента и признана усовершенствованной оценкой со сверхвысокой производительностью как в классификации, так и в регрессии.

Light Gradient Boosting Machine (LightGBM) — это структура градиентного бустинга, основанная на алгоритме дерева решений, предложенном Microsoft Research [23]. LightGBM похож на XGBoost в том, что он аппроксимирует остаток (как первого, так и второго порядка) с помощью разложения функции потерь Тейлора и вводит термин регуляризации, чтобы справиться со сложностью модели. В отличие от XGBoost, который использует предварительно отсортированную идею точного жадного алгоритма для поиска точек разделения, LightGBM может уменьшить использование памяти и повысить скорость обучения, используя алгоритм дерева решений на основе гистограммы.

CatBoost — это библиотека градиентного бустинга, созданная Яндексом [24]. Она использует небрежные деревья решений, чтобы вырастить сбалансированное дерево. Одни и те же функции используются для создания левых и правых разделений на каждом уровне дерева. По сравнению с классическими деревьями, небрежные деревья более эффективны при реализации на процессоре и просты в обучении.

Stacking — это одна из стратегий агрегирования для нескольких учащихся, которая объединяет несколько моделей в двухуровневую структуру. Процесс стекирования заключается в создании нескольких классификаторов на первом уровне в качестве учащегося базового уровня, затем выходные данные этого уровня принимаются как новые функции для повторного обучения нового мета-уровня. Опыт показал, что обучение на мета-уровне сложными моделями может легко привести к проблеме переобучения. Поэтому во многих случаях во втором слое предпочтительны более простые модели, такие как линейная регрессия [25]. Примечательно, что учащиеся базового уровня не ограничиваются слабыми учащимися; по сути, это обычно модели с хорошей производительностью, такие как RF, NN, SVM и т.д.

## 2. Практическое применение ансамблевых методов

Для анализа данных использована выборка по кредитным договорам клиентов банка, включающая 213 201 записей о кредитных заявках, включающая положительные решения и отказы в предоставлении кредита. Общее количество признаков данных — 71. Данные представлены за период с января 2014 года по апрель 2020.

Набор данных разделен на две группы: обучающая выборка (80% выборки) и тестовая выборка (20% выборки), которые используются для обучения модели и оценки производительности соответственно. Для обеспечения сопоставимости различных экспериментов обучающий набор и набор для тестирования одинаковы для разных классификаторов.

Для построения модели скоринга полученные данные требуют первичной обработки, с целью формирования поля признаков модели.

Прежде чем приступать к построению моделей, необходимо проанализировать исходные данные по клиентам.

Алгоритмы машинного обучения не работают с выборками, имеющими пропущенные значения. Поэтому возникает необходимость перейти к данным, не имеющим пропусков [26].

Наилучшим вариантом в случае наличия пропусков в небольшом количестве признаков является удаление таких признаков из выборки [27, 28]. Такой метод применяется только в том случае, когда малая часть объектов выборки имеет пропущенные значения.

Для формирования первичных гипотез была отобрана группа признаков, которые наилучшим образом поддаются интерпретации.

Признаки делятся на два типа: количественные и категориальные.

Количественные признаки в статистике преобладают над другими видами признаков — они наиболее информативны [29].

Таблица 1. Пространство поиска гиперпарамеров

| Классификатор | Пространство поиска |
|---|---|
| LR | C ∈ (–10, 10) |
| DT | max_depth ∈ (1, 10), min_samples_leaf ∈ (1,7) |
| RF | n_estimators ∈ (100, 1000), max_features ∈ (2, 8) |
| XGBoost | n_estimators ∈ (20, 500), gamma ∈ (0,1, 0,5), colsample_bytree ∈ (0,3, 1), max_depth ∈ (2, 8) |
| LightGBM | n_estimators ∈ (20, 500), num_leaves ∈ (8, 128), colsample_bytree ∈ (0,3, 1), max_depth ∈ (2, 8) |
| CatBoost | iterations ∈ (100, 1000) |

Многие классические методы машинного обучения предполагают, что все признаки $X^j = \mathbb{R}$. Однако в некоторых задачах признаки могут принимать значения из множеств, не совпадающих с множествами вещественных чисел. Так, например, признаки могут принимать значения из конечного неупорядоченного множества, например: пол, семейное положение и т.п. Для таких признаков в настоящей работе будет использоваться прием Dummy-кодирования [30].

Еще одним недостатком такого подхода является сильно увеличивающаяся размерность пространства объектов. Многие алгоритмы не способны обрабатывать полученные матрицы данных во многих реальных задачах. В связи с этим описание объектов приходится хранить в разреженном формате и использовать приспособленные методы.

Преимущественно выбирались данные, которые могут являться количественными признаками, так как они поддаются большей интерпретации и легче в подготовке данных.

Чтобы обеспечить полный контраст между моделями ансамблевого обучения и базовыми классификаторами, для сравнения рассматриваются 3 показателя: площадь под кривой (AUC) [31], коэффициент Gini (Gini) [32] и оценка Брайера (BS) [33].

Для оценки классификатора используются методы, основанные на ранжировании элементов (элементы упорядочены по убыванию и прогнозируемая вероятность (оценка) является положительной). Этот ранжированный список сравнивается с истинным классом тестовых элементов и демонстрируется графически с помощью ROC кривой.

На производительность классификаторов напрямую влияют гиперпараметры. Классификаторы в этом исследовании, такие как LR, DT, RF, XGBoost и LightGBM, имеют несколько гиперпараметров, которые необходимо значительно изменить.

Пусть дано пространство гиперпараметров $\lambda = \lambda_1 \times ... \times \lambda_i$, где $\lambda_i$ — пространство $i$-го параметра. Входные данные $D$.

Необходимо найти:

$$\alpha^* = argmin_{\alpha \in \lambda} E_{(D_{train}, D_{valid}) \sim D} L(M_\alpha, D_{train}, D_{valid}),$$

Где $L$ — функция потерь модели $M_\alpha$, обученной при гиперепараметрах $\alpha$ на $D_{train}$ и проведена валидация на $D_{valid}$. В байесовской оптимизации используются вероятностные критерии.

В качестве основного инструмента использовался фреймворк для оптимизации гиперпараметров Optuna [34]. Optuna автоматически находит оптимальные значения гиперпараметров, используя различные семплеры, такие как поиск по сетке, случайные и байесовские алгоритмы.

В таблице 1 приведено пространство поиска для некоторых моделей. Для остальных выбраны с стандартными гиперпараметрами.

Для Stacking после многократных экспериментов с комбинациями выбраны три базовых алгоритма (RF, XGBoost, CatBoost). Поскольку простые линейные модели хорошо работают в классификаторе второго уровня [35]; поэтому мы выбрали классификатор второго уровня LR.

## 3. Результаты работы алгоритмов

Цель этого исследования — провести сравнительную оценку эффективности ансамблевого обучения, т.е. CatBoost, случайного леса, Stacking, XGBoost и LightGBM, в отличие от пяти отдельных моделей, то есть LDA, LR и DT. Сравнение основано на четырех аспектах, а именно: Gini, площадь под кривой (AUC), оценка Бриера (BS) и время работы.

В таблице 2 представлены результаты ансамблевого обучения и традиционных индивидуальных уча-

Таблица 2. Результаты обучения

| Метрика | Ансамблевые алгоритмы | | | | |
|---|---|---|---|---|---|
| | RF | CatBoost | XGBoost | LigthGBM | Stacking |
| AUC | 0,869 | 0,847 | 0,850 | 0,854 | 0,871 |
| Gini | 0,738 | 0,695 | 0,700 | 0,708 | 0,742 |
| BS | 0,038 | 0,033 | 0,035 | 0,038 | 0,034 |
| Time | 02:39 | 00:08 | 00:14 | 00:02 | 04:12 |

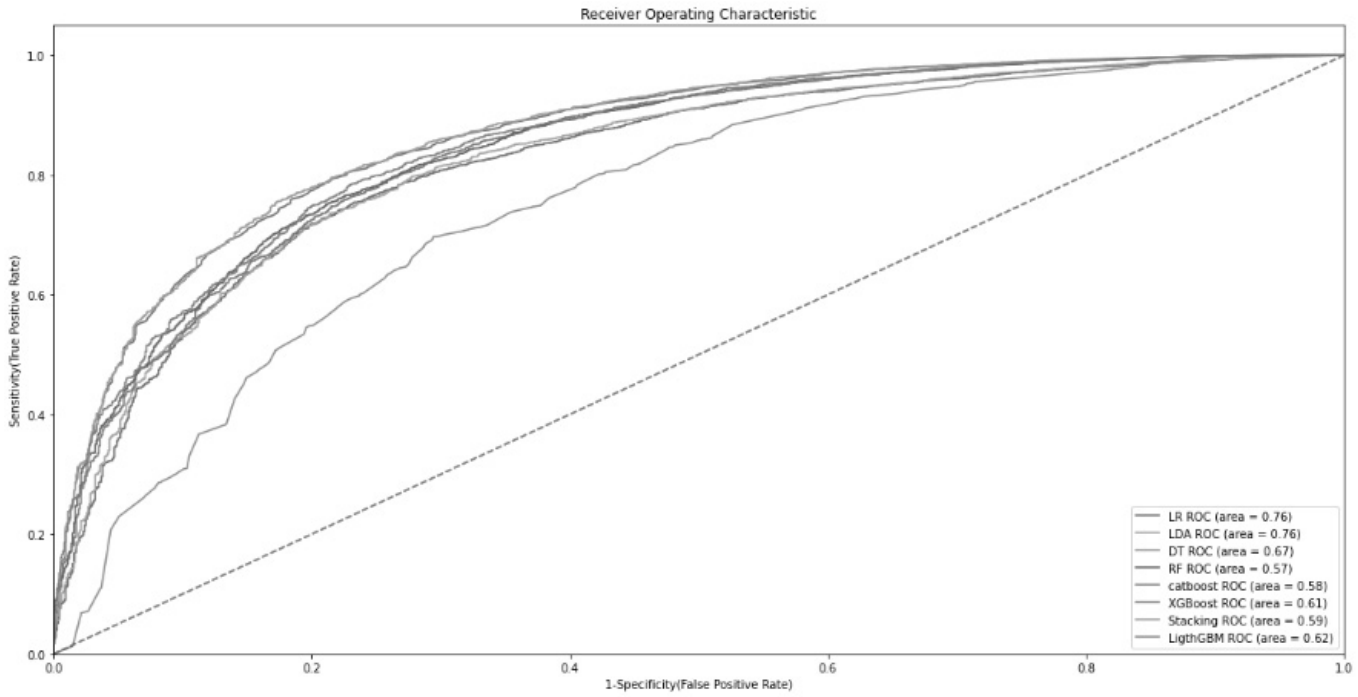


Рис. 1. ROC-кривая по результатам работы каждой модели

щихся (жирным шрифтом показаны лучшие результаты по разным показателям),

Согласно таблице, Stacking достигает наилучших результатов по всем метрикам качества модели. В целом модели ансамблевого обучения превосходят традиционные модели.

В свете изложенных выше экспериментальных данных можно сделать следующие выводы:

1. По сравнению с традиционными моделями, ансамблевое обучение принесло несколько улучшений
2. Таким образом, Stacking (RF+XGboost+CatBoost) является относительно лучшим выбором для кредитного скоринга, и это согласуется с предыдущими исследованиями.
3. RF, XGBoost, LightGBM и CatBoost должны быть идеальным выбором для финансовых учреждений с точки зрения кредитного скоринга и интерпретации результатов.

Рис. 1 показывает кривую ROC каждой модели. Можно видеть, что кривая ROC RF и Stacking лежат выше всех других кривых по всем пороговым значениям, она также имеет выпуклую форму круга по сравнению с другими кривыми, что подразумевает более низкую частоту ложноотрицательных и ложноположительных ошибок. Это означает, что Stacking и RF является лучшим по всем значениям чувствительности и специфичности.

## Заключение

Основная идея этой статьи — экспериментальное исследование о предпочтительных моделях для прогнозирования кредитного риска. Выполняется сравнительная оценка пяти ансамблевых алгоритмов, то есть RF, CatBoost, XGBoost, LightGBM и Stacking, и трёх традиционных моделей, то есть LDA, LR, DT.

Все эксперименты были реализованы на реальном наборе кредитных данных, полученном от кредитной организации. Экспериментальные результаты показывают, что ансамблевое обучение дает явно более высокую производительность, чем отдельные модели. Кроме того, Stacking достигает лучших результатов по трём критериям производительности, т.е. AUC, Gini, BS. По времени работы модели лидером является LigthGBM. Кроме того, в работе учтены временные затраты. Время работы Stacking зависит от выбора базовых моделей. В целом, RF, XGBoost, LightGBM и Stacking могут быть лучшим выбором для финансовых учреждений в период кредитного скоринга при ограничении определенного времени и оборудования.

ЛИТЕРАТУРА


1. Beaver W.H. Financial ratios as predictors of failure. J. Account. Res. 1966, 4, 71–111.
2. Altman E.I. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. J. Financ. 1968, 23, 589–609.
3. Orgler Y.E. A credit scoring model for commercial loans. J. Money Credit Bank. 1970, 2, 435–445.
4. Grablowsky B.J.; Talley W.K. Probit and discriminant functions for classifying credit applicants-a comparison. J. Econ. Bus. 1981, 33, 254–261.
5. Eisenbeis R.A. Pitfalls in the application of discriminant analysis in business, finance, and economics. J. Financ. 1977, 32, 875–900.
6. Desai V.S.; Crook J.N.; Overstreet G.A., Jr. A comparison of neural networks and linear scoring models in the credit union environment. Eur. J. Oper. Res. 1996, 95, –37.
7. West D. Neural network credit scoring models. Comput. Oper. Res. 2000, 27, 1131–1152.
8. Atiya A.F.; Parlos A.G. New results on recurrent network training: Unifying the algorithms and accelerating convergence. IEEE Trans. Neural Netw. 2000, 11, 697–709.
9. Verikas A.; Gelzinis A.; Bacauskiene M. Mining data with random forests: A survey and results of new tests. Pattern Recognit. 2011, 44, 330–349.
10. Hsieh N.-C.; Hung L.-P. A data driven ensemble classifier for credit scoring analysis. Expert Syst. Appl. 2010, 37, 534–545.
11. Ma X.; Sha J.; Wan D.; Yu, Y.; Yang Q.; Niu X. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. Electron. Commer. Res. Appl. 2018, 31, 24–39.
12. Zhu Y.; Xie C.; Wang G.J.; Yan X.G. Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. Neural Comput. Appl. 2017, 28, 41–50.
13. Sagi O.; Rokach L. Ensemble learning: A survey. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2018, 8, 1–18.
14. Chen T.; Guestrin C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
15. Liang W.; Luo S.; Zhao G.; Wu H. Predicting hard rock pillar stability using GBDT, XGBoost, and LightGBM algorithms. Mathematics 2020, 8, 765.
16. Xia Y.; Liu C.; Liu N. Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. Electron. Commer. Res. Appl. 2017, 24, 30–49.
17. Ala'raj M.; Abbod M.F. Classifiers consensus system approach for credit scoring. Knowl.-Based Syst. 2016, 104, 89–105.
18. Li Y.; Chen W. Entropy method of constructing a combined model for improving loan default prediction: A case study in China. J. Oper. Res. Soc. 2019, 1–11.
19. Wang G.; Hao J.; Ma J.; Jiang H. A comparative assessment of ensemble learning for credit scoring. Expert Syst. Appl. 2011, 38, 223–230.
20. Wolpert D.H. Stacked generalization. Neural Netw. 1992, 5, 241–259.
21. Breiman L. Bagging predictors. Mach. Learn. 1996, 24, 123–140.
22. Breiman L. Random forests. Mach. Learn. 2001, 45, 5–32.
23. Ke G.; Meng Q.; Finley T.; Wang T.; Chen W.; Ma W.; Ye Q.; Liu T.Y. LightGBM: A highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 2017, 2017, 3147–3155.
24. Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin. CatBoost: unbiased boosting with categorical features. NeurIPS, 2018.
25. Zhang H., Su J. Learning probabilistic decision trees for AUC //Pattern Recognition Letters. — 2006. — Т. 27. — № . 8. — P. 892–899.
26. Mayr A., Hofner B., Schmid M. The importance of knowing when to stop //Methods of Information. — 2012. — Т. 51. — № . 02. — P. 178–186.
27. Finlay S. Multiple classifier architectures and their application to credit risk assessment //European Journal of Operational Research. — 2011. — Т. 210. — № . 2. — P. 368–378.
28. Ефимова М.Р. Общая теория статистики. Учебник. — 2000.
29. Wood S.N. mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. Retrieved 7 July 2014. — 2013.
30. Джини коэффициент / В.Г. Минашкин // Большая российская энциклопедия: [в 35 т.] / гл. ред. Ю.С. Осипов. — М.: Большая российская энциклопедия, 2004–2017.

31. Brier (1950). Verification of Forecasts Expressed in Terms of Probability. Monthly Weather Review. 78: 1–3.
32. Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019.
33. Optuna: A Next-generation Hyperparameter Optimization Framework. In KDD.
34. Witten I.H.; Frank E. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed.; Morgan Kaufmann: San Francisco, CA, USA, 2005.
35. Chawla Nitesh V.; Herrera Francisco; Garcia Salvador; Fernandez Alberto (2018–04–20). "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary". Journal of Artificial Intelligence Research. 61: 863–905


Журнал «Современная наука: актуальные проблемы теории и практики»

Вятский государственный университет