Инженерный вестник Дона, №4 (2024)
ivdon.ru/ru/magazine/archive/n4y2024/9160

Сравнительная характеристика библиотек машинного обучения для внедрения искусственного интеллекта в CRM-систему

В.А. Передерий, М.Л. Рысин

МИРЭА - Российский технологический университет, Москва

Аннотация: Рассматривается производительность решения задачи классификации средствами различных библиотек искусственного интеллекта и машинного обучения с открытым исходным кодом в области маркетинга и управления взаимоотношениями с клиентами, по результатам экспериментов производится выбор наилучшей библиотеки с целью внедрения искусственного интеллекта в отечественные CRM-системы на основе численных показателей производительности.
Ключевые слова: искусственный интеллект, машинное обучение, большие данные, классификация, маркетинг, управление взаимоотношениями с клиентами, импортозамещение, открытый исходный код, artificial intelligence, machine learning, big data.

Введение

В последние годы наблюдается тренд значительного расширения применения искусственного интеллекта и машинного обучения. Искусственный интеллект используется для решения прикладных задач в самых разных областях. Примерами могут выступать генерация и распознавание текста, классификация объектов на изображениях, обработка обращений клиентов, предотвращение происшествий [1 – 4].

Одной из областей применения машинного обучения является маркетинг и системы управления взаимоотношениями с клиентами (от англ. Customer Relationship Management, далее CRM) [5]. Алгоритмы машинного обучения хорошо подходят для выявления скрытых закономерностей в данных, что делает их эффективными в данной области. Сейчас машинное обучение применяется в маркетинге и CRM для предсказания спроса на товары и услуги, формирования персонализированных предложений, анализа рынка, противодействия оттоку клиентов и т.д. [6].

Машинное обучение тесно связано с понятием больших данных. Большие данные – это гигантские массивы информации, скорость

накопления которых превышает 150 гигабайт в сутки [7]. Массивы такого объёма физически невозможно обработать и анализировать без применения компьютерных средств и специализированных методов. Методы машинного обучения – математический аппарат, позволяющий анализировать большие данные, обнаруживать и использовать скрытые в них закономерности. Существуют различные методы машинного обучения. Одна из основных задач, которую можно решить методами машинного обучения – классификация [8].

Классификатор позволяет путём оптимизации параметров математической модели над набором данных и гиперпараметров построить частную модель, позволяющую предсказывать принадлежность экземпляра выборки определённой группе (классу), даже если экземпляр не использовался при обучении модели. Применительно к маркетингу классификатор, например, может помочь предсказать заинтересованность клиента в продукте или услуге, предсказать отток клиентов [9].

На фоне событий и тенденций последних лет, в том числе мирового экономического кризиса 2007-2009 г. [10], важным является разработка отечественного программного обеспечения с применением перспективных технологий для обеспечения технологического суверенитета, а также конкурентоспособности отечественных решений на мировом рынке. В данной статье мы рассмотрим несколько библиотек машинного обучения, на базе которых возможно построение модуля машинного обучения, применимого в проектах интеграции отечественных систем в сфере больших данных и автоматизации маркетинга. Путём проведения экспериментов на наборах данных из открытых источников определим показатели каждой библиотеки и произведём выбор наилучшей альтернативы с применением многокритериального метода анализа альтернатив.

Применимость открытого программного обеспечения

Как отметил в 2022 г. вице-президент и исполнительный директор Кластера информационных технологий фонда «Сколково» Константин Паршин, более половины продуктов в реестре отечественного программного обеспечения (далее ПО) написаны с использованием открытого программного обеспечения [11].

У открытого ПО в задачах импортозамещения есть несколько весомых преимуществ:

1. Трудоёмкость – код уже разработан и поддерживается сообществом, не требуется разработка решения с нуля.

2. Безопасность – открытое программное обеспечение широко применяется и тестируется различными методами, в результате чего обнаруживаются и устраняются дефекты, влияющие на информационную безопасность. Открытая библиотека почти всегда будет безопаснее самостоятельно написанного кода.

3. Свобода применения – в зависимости от специфик лицензии, открытое ПО можно применять в составе других систем, а также при необходимости расширять.

У открытого ПО также есть и недостатки. Обычно открытое ПО лишено гарантий качества и технической поддержки. Также существует перспектива смены лицензии проекта на более ограничивающую или вовсе закрытую. Поэтому стоит учитывать, что применение отечественных разработок обычно несёт в себе значительно меньше рисков, чем открытого ПО. На наш взгляд, преимущества открытого программного обеспечения превосходят недостатки, поэтому мы будем рассматривать для модуля машинного обучения не только отечественные, но и открытые программные библиотеки.

Перечень тестируемых библиотек и наборов данных

В качестве библиотек для проведения экспериментов были выбраны следующие:

- Yandex Catboost – библиотека градиентного бустинга на деревьях решений с открытым исходным кодом, разрабатываемая и поддерживаемая российской компанией Yandex [12];
- XGBoost – библиотека градиентного бустинга на деревьях решений с открытым исходным кодом, разрабатывается и поддерживается сообществом, спонсируется в том числе иностранными компаниями NVIDIA (USA, California), Intel (USA, California) [13];
- Scikit-learn – открытая библиотека машинного обучения, реализующая широкий спектр моделей машинного обучения: метод опорных векторов, метод K-ближайших соседей, метод случайного леса, градиентный бустинг и т.д. [14];
- Keras – открытая библиотека машинного и глубокого обучения, разрабатываемая и поддерживаемая иностранной компанией Google (USA, California) [15].

Эксперименты были проведены на следующих наборах данных:

- UCI Machine Learning Repository – Bank Marketing – предсказание отклика клиентов на предложение открыть банковский вклад [16].
- Maven Analytics Telecom Customer Churn Prediction – предсказание оттока клиентов оператора сотовой связи [17].
- Credit Score Classification – кредитный скоринг клиентов банка [18].

Выбор гиперпараметров моделей машинного обучения

Показатели модели машинного обучения на наборе данных во многом зависит от выбора гиперпараметров. На рис. 1 проиллюстрировано изменение показателя F-меры в зависимости от глубины деревьев решений в ансамбле для алгоритма Catboost на наборе данных Maven Analytics Telecom Customer Churn Prediction при зафиксированных значениях остальных гиперпараметров модели.

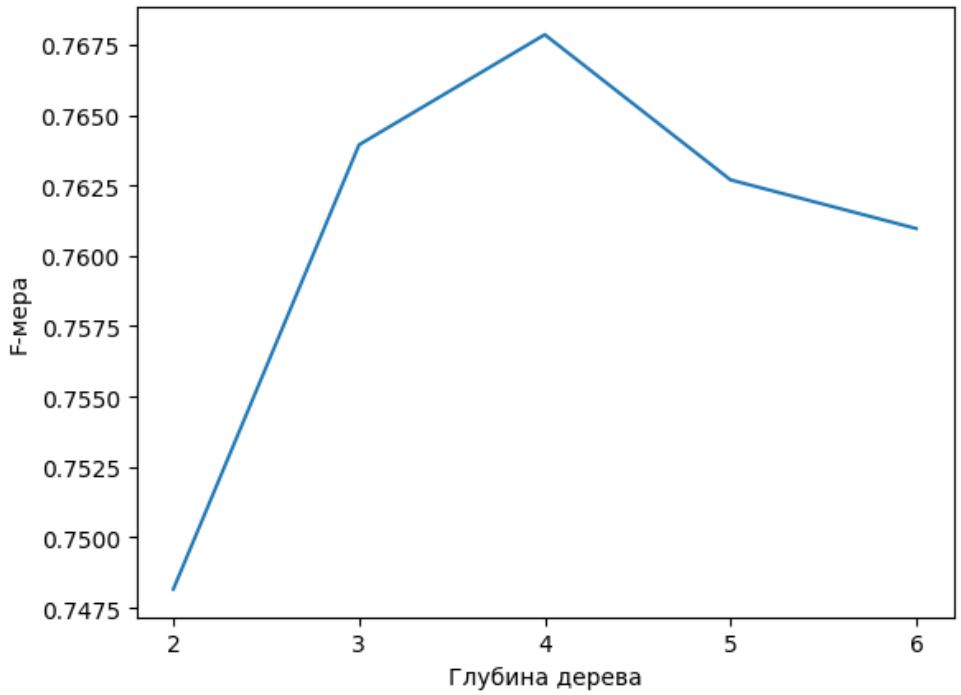

Рис. 1 – График значений F-меры для различных значений глубины дерева

При этом для различных наборов данных оптимальными могут выступать разные наборы параметров. Одним из способов нахождения оптимальных гиперпараметров моделей машинного обучения является метод поиска по сетке, при котором производительность модели оценивается для всех заданных комбинаций гиперпараметров, затем выбираются гиперпараметры модели, показавшей наилучшую производительность [19]. Произведём поиск наилучших гиперпараметров моделей для трёх рассматриваемых наборов данных с использованием метода перебора по

сетке. В таблице №1 приведены наилучшие параметры по результатам автоматизированных расчётов.

Таблица № 1

Наилучшие гиперпараметры моделей машинного обучения

Модель\Набор данных	Bank Marketing	Telecom Customer Churn Prediction	Credit Score Classification
CatboostClassifier	learning_rate=0.04 depth=4	learning_rate=0.03 depth=4	learning_rate=0.07 depth=9
XGBClassifier	learning_rate=0.06 max_depth=6	learning_rate=0.03 max_depth=4	learning_rate=0.08 max_depth=9
LinearSVC	C=1	-	-
KNN	algorithm='kd_tree' leaf_size=20 n_neighbors=9 p=1 weights='uniform'	-	-
RandomForestClassifier	criterion='gini' max_depth=15 max_features='log2' min_samples_leaf=1 min_samples_split=5	criterion='entropy' max_depth=15 max_features='log2' min_samples_leaf=1 min_samples_split=10	criterion='gini' max_depth=15 max_features='log2' min_samples_leaf=1 min_samples_split=2
HistGradientBoostingClassifier	learning_rate=0.08 max_depth=5	learning_rate=0.06 max_depth=4	learning_rate=0.07 depth=10
keras.Sequential	Dense(32, relu) Dense(1, sigmoid) optimizer='adam', loss='binary_crossentropy' batch_size=15	Dense(32, relu) Dense(6, relu) Dense(1, sigmoid) optimizer='adam', loss='binary_crossentropy' batch_size=15	Dense(32, relu) Dense(6, relu) Dense(1, sigmoid) optimizer='adam', loss='binary_crossentropy' batch_size=15

Из результатов автоматизированных расчётов можно сделать вывод, что оптимальные гиперпараметры алгоритмов могут значительно отличаться для разных наборов данных. В частности, в алгоритмах градиентного бустинга на деревьях решений для разных наборов данных оптимальными являются различные значения глубины деревьев.

Результаты экспериментов

Произведём оценку качества классификации по результатам экспериментов. Для экспериментов будем брать модели, показавшие наилучшую производительность для набора данных на предыдущем шаге. Применим показатели эффективности классификации, предлагаемые стандартом ISO IEC TS 4213-2022:

- Точность (Precision);
- Полнота (Recall);
- F1-мера.

Приведённые величины принимают значения от нуля до единицы, при этом большие значения соответствуют лучшему качеству классификации.

Объём обучающей выборки для всех наборов данных был равен 80 ± 0,0001% от общего объёма набора данных, тестовой выборки - 20 ± 0,0001%. Погрешность связана с тем, что наборы данных не всегда можно разделить ровно в соотношении 80/20. Результаты приведены в таблице №2.

По результатам экспериментов мы видим, что наибольшую эффективность в задачах маркетинга показывают классификаторы, основанные на построении ансамблей деревьев решений – CatboostClassifer, XGBClassifier, HistGradientBoostingClassifier, RandomForestClassifier. Классификаторы, основанные на нейронных сетях (библиотека Keras), показали средний результат. Классификаторы LinearSVC и KNN показали наихудшие результаты и предъявляли более строгие требования к данным,

поэтому не рассматривались для наборов данных Telecom Customer Churn Prediction, Credit Score Classification.

Таблица № 2

Показатели эффективности классификации по результатам экспериментов

Набор данных	Библиотека	Модель	Precision	Recall	F1-мера
Bank Marketing	Yandex Catboost	CatboostClassifier	0.665746	0.463462	0.546485
	XGBoost	XGBClassifier	0.623134	0.481731	0.543384
	Scikit-learn	LinearSVC	0.200000	0.002885	0.005687
		KNN	0.569767	0.282692	0.377892
		RandomForestClassifier	0.666667	0.415385	0.511848
		HistGradientBoostingClassifier	0.663235	0.433654	0.524419
	Keras	Sequential	0.443813	0.375961	0.407079
Telecom Customer Churn Prediction	Yandex Catboost	Catboost Classifier	0.831715	0.713889	0.768311
	XGBoost	XGBClassifier	0.825243	0.708333	0.762332
	Scikit-learn	RandomForestClassifier	0.858156	0.672222	0.753894
		HistGradientBoostingClassifier	0.860068	0.700000	0.771822
	Keras	Sequential	0.674931	0.676796	0.675862
Credit Score Classification	Yandex Catboost	Catboost Classifier	0.909690	0.916227	0.912947
	XGBoost	XGBClassifier	0.916241	0.914408	0.915324
	Scikit-learn	RandomForestClassifier	0.887309	0.913862	0.900390
		HistGradientBoostingClassifier	0.856446	0.910042	0.882431
	Keras	Sequential	0.827896	0.887859	0.856830

Для ранжирования библиотек будем применять метод TOPSIS, как наиболее подходящий для рассматриваемого числа альтернатив [20]. Для применения метода необходимо выбрать численные показатели для каждой альтернативы [21]. Так как в предметной области маркетинга количество контактов с клиентом ограничено правилами контактной политики [22], качество работы итоговой модели чувствительно не только к ложно-

отрицательным, но и к ложно-положительным предсказаниям, что следует учитывать. По этой причине будем применять в качестве показателя качества наилучшие показатели F-меры моделей из библиотеки на трёх наборах данных. F-мера является гармоническим средним между точностью и полнотой и чувствительна как к ложно-положительным, так и ложно-отрицательным предсказаниям. На рис. 2 изображен график показателей F-меры всех моделей машинного обучения.

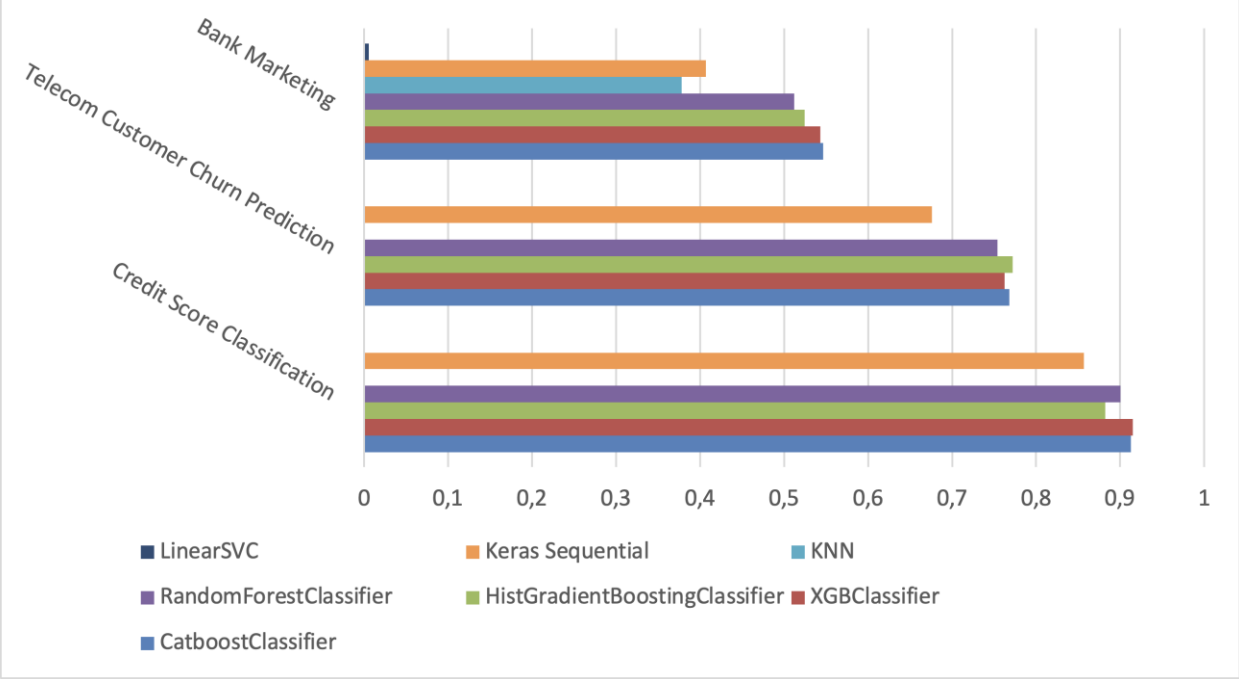

Рис. 2 – Показатели F-меры моделей машинного обучения

Наилучшие показатели F-меры, которых удалось достигнуть с применением моделей из библиотек, приведены в таблице №3.

Таблица № 3

Наилучшие показатели F-меры моделей из библиотек

Библиотека\Набор данных	Bank Marketing	Telecom Customer Churn Prediction	Credit Score Classification
Yandex Catboost	0.546485	0.768311	0.912947
XGBoost	0.543384	0.762332	0.915324
Scikit-learn	0.524419	0.771822	0.900390
Keras	0.407079	0.675862	0.856830

Результаты расчётов по методу TOPSIS представлены в таблице №4. Меньшее значение показателя «расстояние от идеального решения» соответствует более предпочтительному решению.

Таблица № 4

Результаты расчета по методу TOPSIS

Библиотека машинного обучения	Расстояние от идеального решения	Место по TOPSIS
Yandex Catboost	0.000892	1
XGBoost	0.002342	2
Scikit-learn	0.007871	3
Keras	0.052327	4

По результатам расчётов мы видим, что наименьший показатель расстояния от идеального решения показала библиотека Yandex Catboost. Соответственно, она является наилучшим решением. Визуализация ранжирования представлена на рис. 3. На рисунке стрелка, идущая от одной библиотеки к другой, означает превосходство первой библиотеки над второй.

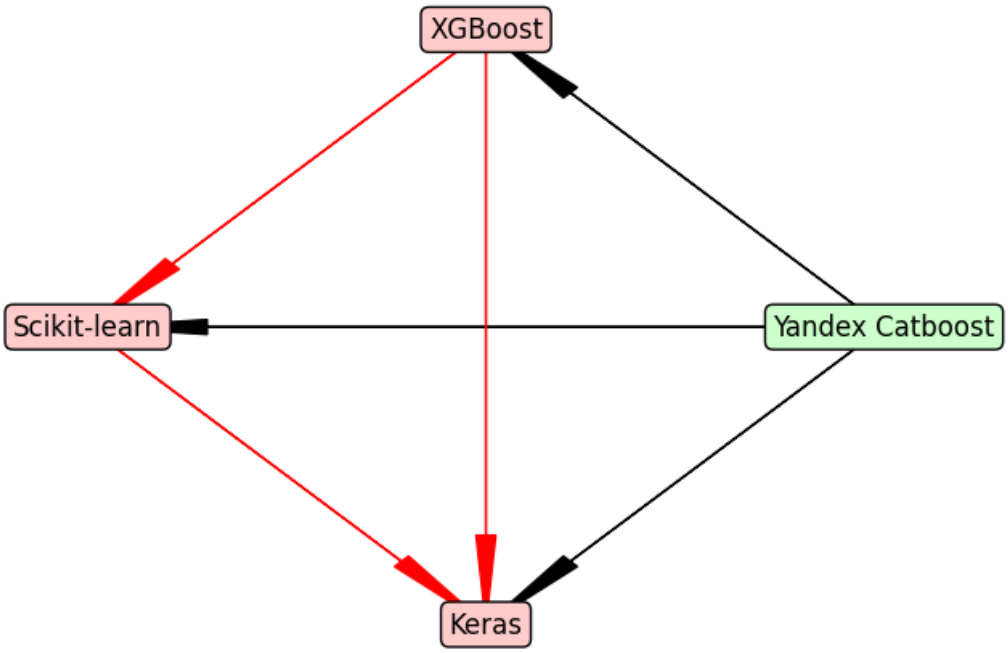

Рис. 3 – Визуализация ранжирования библиотек

Выводы

В результате проведения экспериментов на трёх различных наборах данных, относящихся к предметной области маркетинга, было выявлено, что наилучшие показатели демонстрирует библиотека градиентного бустинга на деревьях решений Yandex Catboost. Также следует отметить, что, так как библиотека разрабатывается и поддерживается отечественной компанией, её применение несёт меньшие риски для российских предприятий. Соответственно, при построении отечественных решений в сфере машинного обучения в маркетинге рационально применять именно эту библиотеку.

Литература

1. Ахмедова Милена Расуловна, Перова Анастасия Евгеньевна Специфика использования технологий искусственного интеллекта в IT-отрасли // Журнал прикладных исследований. 2021. №5. С. 17-22.

2. Кузьменко Е.А., Донченко Д.С., Рагозин В.О. Анализ данных для прогнозирования вероятности дорожно-транспортных происшествий с участием пешеходов // Инженерный вестник Дона. 2020. №6 URL: ivdon.ru/magazine/archive/N6y2020/6506

3. 20+ задач, которые решаются искусственным интеллектом // Белов М., блог VC.RU, 2023. URL: vc.ru/u/1380581-maksim-belov/590134-20-zadach-kotorye-reshayutsya-iskusstvennym-intellektom (Дата обращения – 25.02.2024).

4. Золотарев О.В., Юрчак В.А. Инструменты решения проблем распознавания и кластеризации данных из документов методами машинного обучения // Инженерный вестник Дона. 2023. №2. URL: ivdon.ru/magazine/archive/n2y2023/8215

5. Рябова В.А. Применение машинного обучения в маркетинге // Инновации и инвестиции. 2022. №4. С. 74-75.

6. Машинное обучение в маркетинге: зачем и как использовать // Курченок Н., блог Perfluence. URL: perfluence.net/blog/article/mashinnoe-obuchenie-marketing (Дата обращения - 25.02.2024).

7. Что такое Big Data и как они устроены // Яндекс Практикум, 2022. URL: practicum.yandex.ru/blog/chto-takoe-big-data/ (Дата обращения – 20.02.2024).

8. Классификация // MachineLearning.ru, 2011. URL: machinelearning.ru/wiki/index.php?title=Классификация (Дата обращения – 20.02.2024).

9. Камалходжаева Н., Шиков А.Н. Исследование и анализ алгоритмов машинного обучения для прогнозирования оттока клиентов в телекоммуникационной компании // Международный научно-исследовательский журнал. 2022. №7-1 (121). С. 108-111.

10. Комолов О.О. Деглобализация: новые тенденции и вызовы мировой экономике. Вестник Российского экономического университета имени Г. В. Плеханова. 2021; 18(2). С. 34-47.

11. Конференция Open Source Day 2022 // TAdviser, 2022. URL: tadviser.ru/index.php/Конференция:Конференция_Open_Source_DayOpen_Source_Day_2022 (Дата обращения – 25.02.2024).

12. CatBoost // Яндекс, 2024. URL: yandex.ru/dev/catboost/ (Дата обращения – 25.02.2024).

13. XGBoost // Github, 2024. URL: github.com/dmlc/xgboost (Дата обращения - 25.02.2024).

14. Pedregosa F. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, p. 2825-2830, 2011.

15. Что такое Keras: вводный гид по библиотеке Keras // Simplilearn, 2024. URL: simplilearn.com/tutorials/deep-learning-tutorial/what-is-keras (Дата обращения - 25.02.2024).

16. Набор данных «маркетинг в банке» (Bank Marketing) // UC Irvine Machine Learning Repository, 2012. URL: archive.ics.uci.edu/dataset/222/bank+marketing (Дата обращения – 25.02.2024).

17. Introducing the Maven Churn Challenge // Maven Analytics, 2022. URL: mavenanalytics.io/blog/maven-churn-challenge (Дата обращения - 25.02.2024).

18. Набор данных “Credit score classification” // Роан П., Kaggle, 2022. URL: kaggle.com/datasets/parisrohan/credit-score-classification (Дата обращения - 25.02.2024).

19. Оптимизация гиперпараметров классификатора // Scikit-learn.org, 2024. URL: scikit-learn.org/stable/modules/grid_search.html

20. Подоплелова Е.С. Анализ методов многокритериального принятия решений на примере задачи ранжирования // Известия ЮФУ. Технические науки. 2023. №3 (233). С. 118-125.

21. Halicka K. Technology Selection Using the TOPSIS Method. Foresight and STI Governance, 14 (1), pp. 85-96, 2020. DOI: 10.17323/2500-2597.2020.1.85.96.

22. EFMA Customer Intelligence and CRM 2012. Записки очевидца // Тулубьев П., блог FutureBanking, 2012. URL: futurebanking.ru/post/1973 (Дата обращения – 25.02.2024).

References

1. Ahmedova Milena Rasulovna, Perova Anastasiya Evgen'evna. Zhurnal prikladnyh issledovanij. 2021. №5. pp. 17-22.

2. Kuz'menko E.A., Donchenko D.S., Ragozin V.O. Inzhenernyj vestnik Dona. 2020. №6. URL: ivdon.ru/magazine/archive/N6y2020/6506

3. 20+ zadach, kotorye reshajutsja iskusstvennym intellektom [20+ tasks solved by artificial intelligence], Belov M., VC.RU blog, 2023. URL:

vc.ru/u/1380581-maksim-belov/590134-20-zadach-kotorye-reshayutsya-iskusstvennym-intellektom

4. Zolotarev O.V., Yurchak V.A. Inzhenernyj vestnik Dona. 2023. №2 URL: ivdon.ru/magazine/archive/n2y2023/8215

5. Ryabova V.A. Innovacii i investicii. 2022. №4. pp. 74-75.

6. Mashinnoe obuchenie v marketinge: zachem i kak ispol'zovat' [Machine Learning in Marketing: why and how to use it]. Kurchenok N., blog Perfluence. URL: perfluence.net/blog/article/mashinnoe-obuchenie-marketing

7. Chto takoe Big Data i kak oni ustroeny [What is Big Data and how it works]. Yandex Praktikum, 2022. URL: practicum.yandex.ru/blog/chto-takoe-big-data/

8. Klassifikaciya [Classification]. MachineLearning.ru, 2011. URL: machinelearning.ru/wiki/index.php?title=Классификация

9. Kamalhodzhaeva N., Shikov A.N. Mezhdunarodnyj nauchno-issledovatel'skij zhurnal. 2022. №7-1 (121). pp. 108-111

10. Komolov O.O. Vestnik Rossijskogo jekonomicheskogo universiteta imeni G. V. Plehanova. 2021; 18(2) pp. 34-47.

11. Konferenciya Open Source Day 2022 [Open Source Day 2022 Conference] TAdviser, 2022. URL: tadviser.ru/index.php/Конференция:Конференция_Open_Source_DayOpen_Source_Day_2022

12. CatBoost. Yandex, 2024. URL: yandex.ru/dev/catboost/

13. XGBoost. Github, 2024. URL: github.com/dmlc/xgboost

14. Pedregosa F. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, p. 2825-2830, 2011.

15. What Is Keras: The Best Introductory Guide To Keras, Simplilearn, 2024. URL: simplilearn.com/tutorials/deep-learning-tutorial/what-is-keras

16. Bank Marketing Dataset UC Irvine Machine Learning Repository, 2012. URL: archive.ics.uci.edu/dataset/222/bank+marketing

17. Introducing the Maven Churn Challenge, Maven Analytics, 2022. URL: mavenanalytics.io/blog/maven-churn-challenge

18. “Credit score classification” Dataset, Rohan P., Kaggle, 2022. URL: kaggle.com/datasets/parisrohan/credit-score-classification

19. Tuning the hyper-parameters of an estimator, Scikit-learn.org, 2024. URL: scikit-learn.org/stable/modules/grid_search.html

20. Podoplelova E.S. Izvestiya YUFU. Tekhnicheskie nauki. 2023. №3 (233). pp. 118-125.

21. Halicka K. Technology Selection Using the TOPSIS Method. Foresight and STI Governance, 14 (1), pp. 85-96, 2020. DOI: 10.17323/2500-2597.2020.1.85.96.

22. EFMA Customer Intelligence and CRM 2012. Zapiski ochevidca [EFMA Customer Intelligence and CRM 2012. Summary from an attender], Tulub'ev P., FutureBanking blog, 2012. URL: futurebanking.ru/post/1973

Дата поступления: 3.03.2024

Дата публикации: 16.04.2024