energies

MDPI

Research

Предсказание отключений электроэнергии на ЛЭП 110 кВ на основе параметров самих ЛЭП

Vadim Bolshev [4,*],

[4] Laboratory of Power Supply and Heat Supply, Federal Scientific Agroengineering Center VIM, 109428 Moscow, Russia; schkolamolen@gmail.com

[5]

* Correspondence: vadimbolshev@gmail.com (V.B.); Tel.: +7 499 174 85 95 (V.B.).

Abstract:

В рамках данного исследования предложено использование алгоритмов машинного обучения для прогнозирования отключений электрической энергии на линиях электропередачи 110 кВ на основе данных по параметрам самих линий. В качестве алгоритмов были использованы 5 классификаторов: машина опорных векторов, логистическая регрессия, случайный лес, градиентные бустинги, основанные на деревьях решений LightGBM Classifier и CatBoostClassifier. Для автоматизации процесса преобразования данных и устранения возможности их утечки использовался пайплайн (Pipeline) и компоновщик разнородных признаков Column Transformer, данные для моделей подготавливались методами горячего кодирования One-Hot Encoding и стандартизацией данных. Разбиение данных на обучающую и валидационную выборки выполнено через кросс-валидацию со стратифицированным разделением. Настройка гиперпараметров классификаторов осуществлена методами оптимизации случайных параметров RandomizedSearchCV и сеточного поиска GridSearchCV. Наилучшее качество прогнозирования отключений удалось достичь модели логистической регрессии с метриками качества ROC AUC, равной в 0,78, и AUC-PR – в 0,68. На последнем этапе исследования произведен анализ вероятности влияния параметров ЛЭП 110 кВ на вероятность их отказов за счет определения важности признаков различных моделей, в том числе оценки вектора коэффициентов регрессии.

Keywords:

электрические сети, линии электропередачи, надежность электроснабжения, перерывы в электроснабжении, отключения электроэнергии, отказы ЛЭП, машинное обучение, оценка важности признаков, машина опорных векторов, логистическая регрессия, случайный лес, градиентный бустинг.

1. Введение

Снабжение потребителей электрической энергией осуществляется через сложную распределительную систему, которая включает в себя многочисленные воздушные и подземные ЛЭП различного напряжения, трансформаторы, опоры, линейную аппаратуру различных типов и другое оборудование. Надежность системы электроснабжения (СЭС), то есть вероятность бесперебойного снабжения потребителей электрической энергией, увеличивается с использованием новых, инновационных решений, включающие в себя внедрение современных средств защиты,

мониторинга и управления электрическими сетями [1,2]. Несмотря на общую высокую надежность СЭС, отказы на линиях электропередачи являются относительно частым явлением, и составляет 35-50 % от всех отказов в системах электроснабжения напряжением 35-750 кВ [3]. Такой объем отключений связан с большой территориальной протяженностью ЛЭП и их подверженностью влиянию климатическим воздействиям [4,5]. Несмотря на то, что неблагоприятные погодные условия являются самыми часто встречаемыми причинам неисправности, необходимо учитывать и другие причины, в том числе износ и устаревание инфраструктуры электрической сети [6]

Анализ данных об отключениях электрической энергии позволяют определить факторы, оказывающие наибольшее влияние на вероятность отказа. Существует большое количество работ, посвященных статистическому анализу отказов на ЛЭП, как например, в работах [7–9], посвящённых определению причин отказов на ЛЭП и основных направлений снижения количества отключений в электрических сетях или в работе [10] посвященной анализу отказов на ЛЭП в результате гололедных явлений. В работе [11] дополнительно к статистический обработке данных, применен регрессионный анализ временных рядов отказов в СЭС протяженностью в 314 км, позволившие определить основные причины аварийных отключений, параметры и критерии качества трендовых обратных моделей частоты среднемесячных отказов.

Несмотря на то, что в некоторых работах по статистическому анализу аварийных отключений были определены причины отказов с указанием определенных элементов, вышедших из строя, в них не рассматривалось влияние типов этих элементов на вероятность отказов ЛЭП. Одним из эффективных методов определения вероятности отключения на основе таких факторов является прогностическое моделирование, построенное на методах машинного обучения (ML – machine learning) [12]. Так, в работе [13] предлагается применение трехмерного метода опорных векторов (Support Vector Machine - SVM) для прогнозирования отключений компонентов энергосистемы, а в работе [14] этот же метод используется для определения места повреждения в системе посредством измерения величины и угла падения напряжения на первичной подстанции распределительной системы. В статье [15] представлен подход к выявлению неисправностей оборудования в распределительных системах, в результате которого решается задача бинарной классификации, в которой отключения делятся на два класса: отказы оборудования и отказы, не связанных с оборудованием. В качестве классификаторов используются три алгоритма: дерево решений, логистическая регрессия и наивный байесовский классификатор. Применение искусственных нейронных сетей представлено в работе [16] для многоклассовой классификации неисправностей систем электроснабжения на основе значений тока и напряжения на всех линиях электропередачи (ЛЭП). Большое количество работ посвящено применению методов машинного обучения для прогнозирования отключений в электрораспределительных сетях во время неблагоприятных погодных условий [17,18], в частности ураганов, тропических штормов, дождей и ветровых штормов, а также прибрежных наводнений, в том числе на основе данных США [18], Франции [19], Пуэрто-Рико [20], Китая [21]. В случае возникновения тайфуна для определения количества отключений в работе [22] используется алгоритм случайного леса, решающей задачу многоклассовой классификации, в работе [23] – алгоритм градиентного бустинга, в работе [21] - ансамбль, состоящий из 2 уровней, причем последний - градиентный бустинг XGBoost. В статье [24] предлагается подход для прогнозирования отключений в распределительных системах, вызванных факторами окружающей среды, за счет использования глубоких нейронных сетей. В отличии от классических регрессионных нейронных сетей со скрытыми полно связанными слоями [25], в данном подходе имплементированы промежуточные звенья с

независимыми блоками, объединенными в ансамбль. Проводя обзор существующей литературы по прогнозированию отключений электрической энергии, было установлено, что прогнозирование отказов ЛЭП на основе параметров самих линий не проводился ни в каком виде, что говорит об актуальности проводимого исследования.

Таким образом, проанализировав существующие работы по прогнозированию отключений электрической энергии, было обнаружено, что прогнозирование отказов ЛЭП на основе параметров самих линий не проводилось ни в каком виде, поэтому актуальность данного исследования подтверждена.

Целью исследования – разработать модель машинного обучения для прогнозирования возможных отключений электроэнергии на линиях электропередачи на основе характеристик самих ЛЭП.

2. Методология и материалы

Материалы для исследования

Настоящее исследование построено на данных по отключениям электрической энергии в электрических сетях Орловской области. В предыдущем исследовании [] рассматриваемые данные были проанализированы методами математической статистики, в рамках которого были проведены обработка пропущенных значений, удаление дубликтов, создание синтетических параметров, включая целевой признак, выявление выбросов и аномалий, выбор наиболее подходящих признаков для обучения ML модели. Итоговым результатом стала таблица из 9 признаков, включая целевой, и 395 объектов. Таблица содержит следующие признаки:

a) 3 признака с категориальными значениями:
- Факт отключения (целевой признак);
- Проводник, тип, сечение;
- Отношение ЛЭП к транзиту.

b) 7 признаков с количественными значениями:
- Индекс состояния, %;
- Проводник, тип, сечение;
- Протяженность воздушных участков, км;
- Переэксплуатация, бр.;
- ЖБ Опоры, %;
- Протяженность по лесу, %;
- Протяженность по населенной местности, %.

Методология исследования

Начало исследования заключалось в проверке качества подготовленных данных методами разведочного анализ данных (Exploratory Data Analysis), включающих в себя статистический анализ распределения количественных и категориальных переменных в разрезе целевого признака, а также исследование корреляционной зависимости между переменными с помощью коэффициента корреляции phi ϕ_k.

Так как цель работы является прогнозирование вероятности отключении ЛЭП на основе её параметров, то алгоритмами машинного обучения решается задача бинарной классификации, поэтому в качестве классификаторов были выбраны 5 ML моделей, основанных на следующих алгоритмах:

- машина опорных векторов (Support Vector Machine - SVM),
- логистическая регрессия регрессия (LogisticRegression - LR),
- случайный лес (RandomForestClassifier - RFC),
- градиентные бустинги, основанные на деревьях решений, LightGBM Classifier и CatBoostClassifier.

Правильная настройка гиперпараметров классификаторов осуществлена посредством метода оптимизации случайных параметров RandomizedSearchCV, а лучшие найденные параметры уточнены алгоритмом сеточного поиска GridSearchCV. Разбиение данных на обучающую и валидационную выборки выполнено через кросс-валидацию со стратифицированным разделением. Подготовка данных для ML моделей осуществлена методом горячего кодирования One-Hot Encoding для категориальных переменных и методом стандартизацией данных Standard Scaler для количественных. Для автоматизации процесса преобразования данных и обучения модели, а также для устранения возможности утечки данных использовался пайплайн (Pipeline) и компоновщик разнородных признаков Column Transformer. Выбор лучшей обученной ML модели производился метрикой качества ROC AUC. Кроме метрики ROC AUC, оценка качества моделей производилась дополнительными метриками [26], такими как AUC-PR, Accuracy, Precision, Recall, F1-score. На последнем этапе произведен анализ вероятности влияния параметров ЛЭП 110 кВ на вероятность их отказов за счет определения важности признаков различных моделей, в том числе оценки вектора коэффициентов регрессии.

Применяемое программное обеспечение

В данном исследовании обработка и анализ данных производились на языке программирования Python в среде разработки Jupyter notebook (ver. 7.0.6) программного комплекса Anaconda Python. Для работы с табличными данными применялась библиотека Pandas (ver. 2.1.4), для математических обработки массивов данных – NumPy (ver. 1.26.3), для визуализации данных - Matplotlib (ver. 3.8.0) и Seaborn (ver. 0.12.2), для корреляционного анализа Phik (ver. 0.12.4). Для выстраивания алгоритмов обучения, преобразования данных, пригодных для ML задач, а также работой с основными классическими моделями машинного обучения использовалась библиотека Scikit-learn (ver. 1.2.2). Дополнительно к ней, применялись фреймворки градиентных бустингов из библиотек LightGBM (ver. 4.1.0) и CatBoost (ver. 1.2.2). Для работы с несбалансированным набором данных применялись инструменты из библиотеки Imbalanced-learn.

3. Результаты и обсуждение

Разработка модели машинного обучения, способную выполнять поставленную задачу, включает в себя правильные выбор и предобработку данных для обучения, подбор наиболее подходящего алгоритма машинного обучения, а также точную настройку гиперпараметров модели для обеспечения оптимальной производительности за счет её тестирования [27]. Каждый из этих этапов является важным звеном в построении качественной модели в соответствии с измеряемой метрикой.

3.1 Быстрый анализ данных

Перед началом разработки моделей машинного обучения изучим подготовленные данные. Выведем гистограммы для категориальных признаков и график плотности распределения для количественных, при этом последние будем строить с ядерной оценкой плотности [28,29], в качестве весовой функции которой воспользуемся Гауссовским ядром [30].

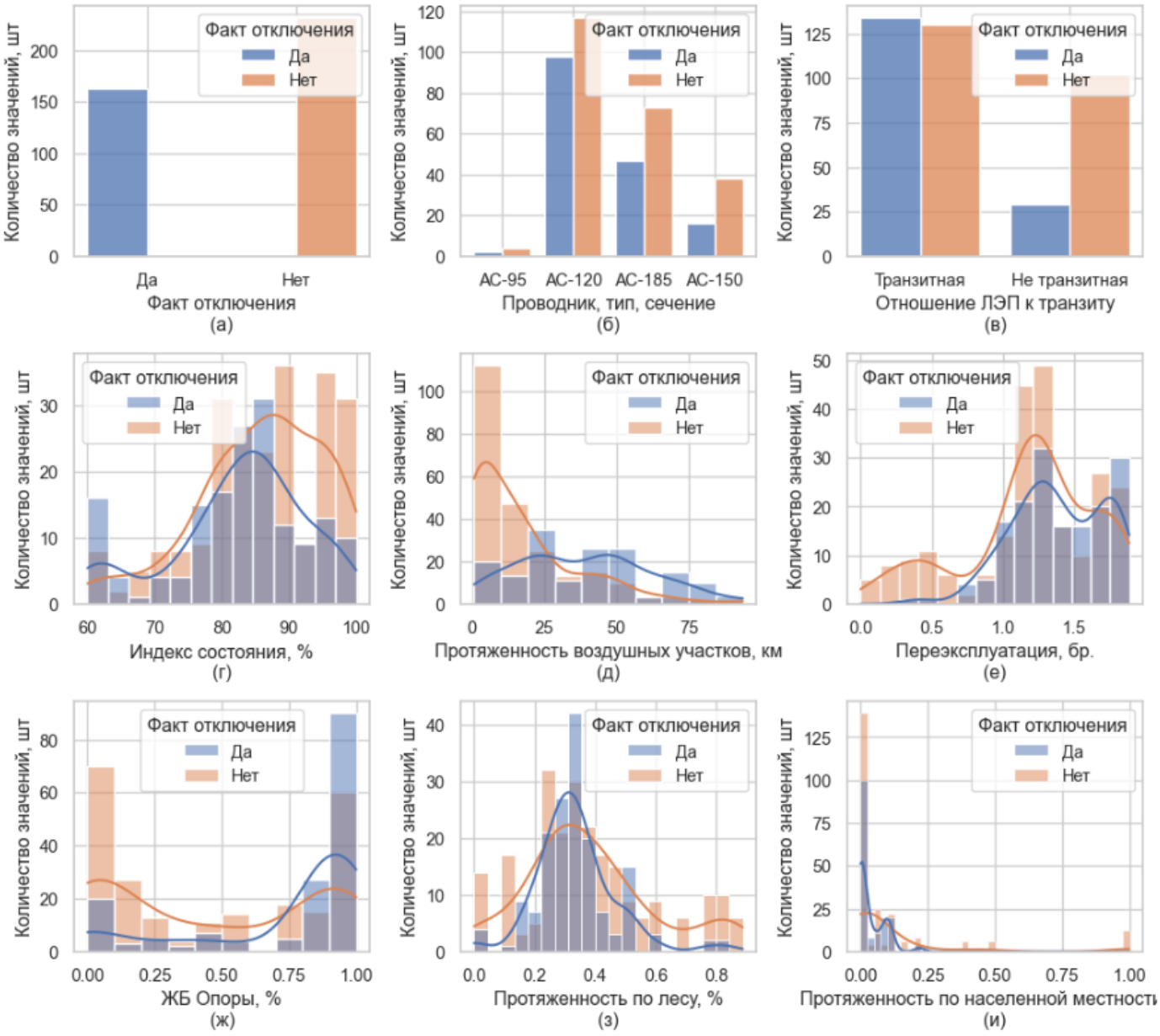


Рисунок 1 – Распределение значений параметров с категориальными значениями в разрезе целевой переменной

Из приведённых гистограмм можно сделать следующие выводы:

1. Рассматривая гистограммы категориальных признаков видно, что имеется разное распределение значений для *ЛЭП* с отказами и без, что сигнализирует об их влиянии на целевую переменную. Также необходимо отметить низкую кардинальность категориальных признаков (4 значения у признака «Проводник, тип, сечение» и 2 – у «Отношение *ЛЭП* к транзиту»), поэтому при подготовке данных к непосредственному обучению ML моделей, наиболее оптимальным кодированием категориальных переменных будет выступать метод горячего кодирования «OneHotEncoding».
2. Графики плотностей распределения количественных признаков между линиями электропередачи, на которых наблюдались отключения и нет,

различаются по своей форме, поэтому можно предположить, что имеется влияние этих признаков на целевую переменную.

3. Целевой признак «Факт отключения» немного несбалансирован. Количество ЛЭП с отказами составляет 163 шт., без отказов – 232, разница в которых примерно находится в районе 20 %. Дисбаланс классов может вызывать проблемы при обучении моделей машинного обучения, если последние являются не-вероятностными, например, метода опорных векторов (англ. SVM, Support Vector Machine) или при решении задач многоклассовой классификации.

Воспользуемся методом корреляционного анализа для нахождения степени взаимосвязи между различными переменными. В качестве меры корреляции воспользуемся коэффициентом корреляции phi ϕ_k, позволяющему проанализировать не только количественные переменные, но и категориальные [31]. Степень корреляции при таком методе будет лежать в пределах 0…1, где 0 означает отсутствие взаимосвязи между признаками, а 1 – об ее максимальной степени [32]. На рисунке 2 отображена тепловая карта, графически отображающую матрицу корреляции, полученную при определении взаимосвязи между рассматриваемыми переменными с помощью метода расчета коэффициента корреляции ϕ_k.

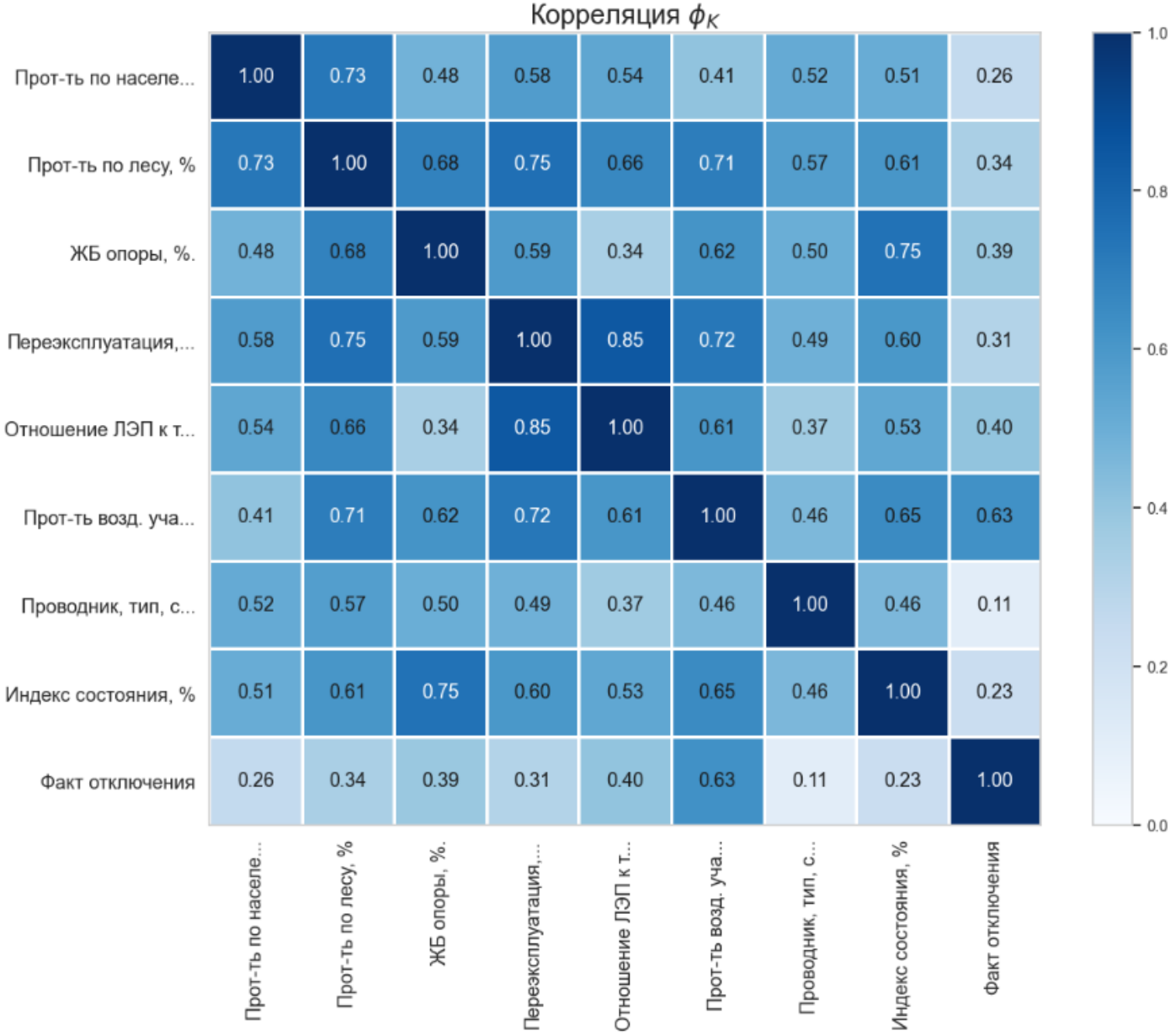


Рисунок 2 – Тепловая карта матрицы корреляции ϕ_k между параметрами ЛЭП.

Корреляционный анализ показал, что:

- Замена абсолютных значений количества железнобетонных и металлических опор, а также протяженности по лесу и по населенной местности на относительные величины позволило справиться с проблемой мультиколлеарности этих признаков (данная проблема обсуждалась на предыдущем этапе исследования []), что должно благотворно повлиять на обучение ML моделей;
- Целевой признак «Факт отключения» имеет не сильную, но достаточную корреляционную зависимость с рассматриваемыми переменными. Самую большую корреляцию показывает признак «Протяженность воздушных участков, км» (0,63), наименьшую - категориальный признак «Проводник, тип, сечение» (0,11);
- Выявлена достаточно сильная корреляция (0,85) между сроком эксплуатации линии электропередачи и фактом, является ли ЛЭП транзитной или нет. Данный вопрос обсужден на предыдущем этапе по подготовки данных к машинному обучению [], в котором сделан вывод, что транзитные ЛЭП по представленной статистике имеют больший срок эксплуатации, чем не транзитные ЛЭП. Также решено оставить оба признака, так как эти параметры отражают абсолютно разные значения.

3.2 Алгоритм обучения и настройки гиперпараметров ML моделей

Алгоритм обучения ML моделей посредством метода оптимизации случайных параметров RandomizedSearchCV представлен на рисунке 1. В соответствии с алгоритмом, рандомизированый набор гиперпараметров модели выбирается из распределения по возможным значениям параметров (сетка параметров) и инициализируется для обучения (блок 4). Количество вариаций гипермараметров моделей методом RandomizedSearchCV выбрано равным 100 итераций (значение “n_iter” - блок 2 рисунка 1). В этом же блоке выбирается количество фолдов для кросс-валидации (значение “cv” - блок 2 рисунка 1). Значение “scor_iter” (значение метрики качества модели) приравнивается к 0 и используется в дальнейшем для поиска лучшей модели (блоки 11, 12, 13). Как только количество итераций превысиит значение “n_iter” (блок 3), процесс подбора гиперпараметров остановится и будет выдана ML модель с гиперпараметрами, соответствующими лучшей метрики качества (блок 14).

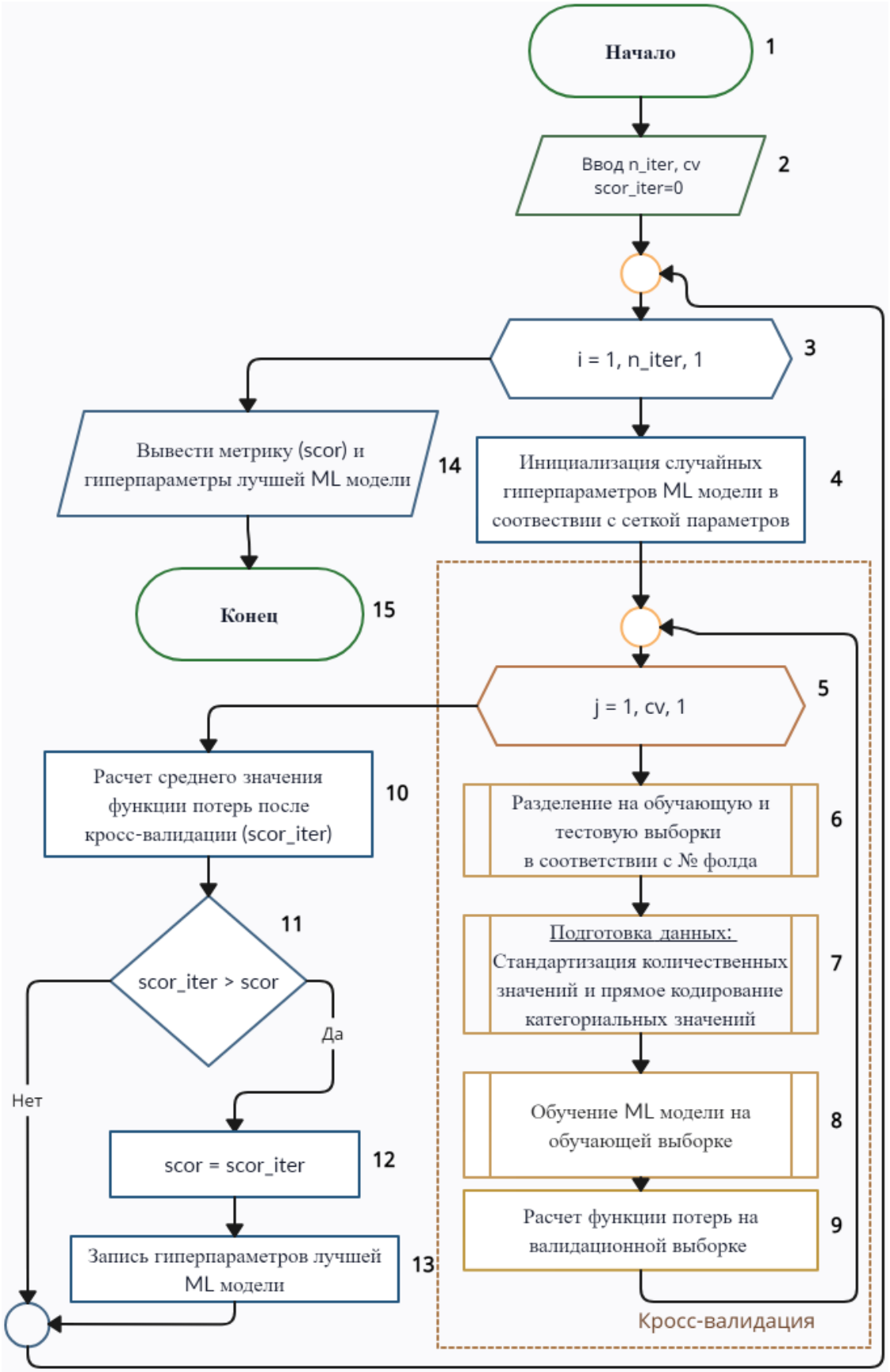


Рисунок 1 - Алгоритм настройки гиперпараметров ML моделей

3.4 Кодирование и масштабирование признаков

Решающее значение в разработке и применении моделей машинного обучения, в том числе нейронных сетей, имеет целенаправленный и тщательный подход к подготовке данных. Успех процесса обучения ML моделей во многом зависит от качества и пригодности данных, подаваемых в модели. На предыдущем этапе была выполнена важная часть обработки – это обработка пропущенных значений и дубликатов, а также определение подходящих признаков для поставленной задачи классификации []. Теперь требуется проведение как преобразования категориальных переменных в числовые представления, так и нормализация количественных переменных: масштабирование и центрирование числовых элементов для облегчения сходимости.

Как было отмечено выше, категориальные переменные имеют низкую кардинальность, поэтому решено воспользоваться кодированием этих переменных методом горячего кодирования One-Hot Encoding. Для нормализации количественных переменных был использован метод стандартизации данных Standard Scaler (блок 7 рисунка 1). Для автоматизации процесса преобразования данных и обучения модели, а также для устранения возможности утечки данных применен пайплайн (Pipeline) и компоновщик разнородных признаков Column Transformer.

Для борьбы с дисбалансом классов целевого признака решено воспользоваться двумя методами: методом взвешивания классов, реализованных внутри моделей из библиотеки scikit-learn, и методом синтетической избыточной выборки меньшинства для номинальных и непрерывных значений (Synthetic Minority Over-sampling Technique for Nominal and Continuous - SMOTE-NC), реализованного через Pipeline библиотеки imblearn для автоматизации процесса и устранения проблемы утечки признаков. Оба метода применены при обучении каждой модели и затем полученные результаты представлены в итоговой таблице.

3.5 Разбиение датасета на обучающую и тестовую выборки

Еще одним важным моментом при построении моделей машинного обучения является проверка её качества на независимых данных, то есть на тех данных, которые модель не видела. С этой целью из исходных данных была выделяется тестовая выборка обычно в размере 10-20 % от обучающей. Данный подход позволяет определить и соответственно устранить основную проблему обучения ML моделей – переобучение, то есть явление, когда обученные ML модели заучивают ответы на обучающей выборке, но плохо определяют закономерности на сторонних данных. Однако, так как объем данных составляет 395 объектов, то было решено отделить тестовую выборку в размере 20%, а к обучающей выборке применить метод кросс-валидации (на рисунке 1 блоки 5-9,10), подразумевающий разделение данных на несколько частей (фолдов), при этом каждый фолд на своем этапе обучения должен выступать в качестве валидационной выборки, остальные – в качестве обучающих. В нашем случае была использована кросс-валидация на 5 фолдов со стратифицированным разделением, гарантирующим одинаковое соотношение классов на всех выборках, что особенно важно при несбалансированных данных. Таким образом, количества итераций обучения на одном наборе гиперпараметров модели зависит от количества фолдов (значение “cv” блок 5 рисунка 1). Согласно алгоритму, как только обучение проведется на всех фолдах, будет произведен расчет среднего значения метрики качества (блок 10), которое далее будет сравнено с лучшей метрикой в блоке условия 11.

3.6 Обучаемые ML модели и используемая сетка гиперпараметров

3.6.1 Метод опорных векторов (Support Vector Machine - SVM)

Метод опорных векторов (SVM) — это алгоритм обучения с учителем, используемый для классификации и регрессионного анализа [33]. Алгоритм SVM работает путем преобразования входных данных в многомерное пространство с помощью функции ядра, затем поиска гиперплоскости, которая лучше всего разделяет точки данных на разные классы. Проще говоря, SVM пытается найти лучшую линию (в двух измерениях) или гиперплоскость (в нескольких измерениях), которая разделяет точки данных, принадлежащие разным классам. Точки, ближайшие к гиперплоскости, называются опорными векторами и используются для определения гиперплоскости. Основными гиперпараметрами, от которых зависит обучения SVM модели, являются коэффициент регуляризации C, тип используемого ядра и, соответственно, его коэффициент. Сетка значений для этих гиперпараметров представлена в таблице 1.

3.6.2 Логистическая регрессия (LogisticRegression - LR),

Регрессионное моделирование — один из наиболее популярных статистических подходов, позволяющий определить взаимосвязи между целевой переменной и набором независимых предикторов. Модели регрессии подразделяются на логистические и линейные, при это линейная регрессия не может использоваться для определения дихотомической (бинарной) переменной, поскольку результатом такой модели является непрерывные значения, в том числе с отрицательным направлением [34]. Поэтому для предсказания вероятности бинарного значения используется расширенная версия линейной регрессии - логистическая регрессия, предсказывающая вероятностное распределение события (да/нет или 1/0) через функцию логит-связи. В процессе обучения были выбраны нерегулируемые гиперпараметры алгоритма оптимизации (“saga”) и вида регуляризации (“Elastic-Net”), позволяющий использовать оба вида регуляризации (l1 и l2) одновременно. Приемлемое соотношение между этими двумя видами регуляризации определялось изменением гиперпараметра “l1_ratio” (Таблица 1). Также регулировалась сила регуляризации C.

3.6.3 Случайный лес (RandomForest),

Случайный лес является одним из наиболее широко используемых алгоритмов машинного обучения, применяемого как в задачах регрессии, так и классификации [35]. Алгоритм случайного леса основан на методе ансамблевого обучения, результаты которого строятся за счет объединения независимых прогнозов леса - множества решающих деревьев, обученных на выборках, полученных с помощью метода бутстрап (англ. bootstrap) [36]. В случае задачи классификации финальный ответ модели будет основан за счет голосования решающих деревьев, в случае регрессии – за счет усреднения ответов [37]. Чем больше решающих деревьев, то в большинстве случаев качество модели будет выше, поэтому при обучении случайного леса использовалось до 1000 деревьев. Известно, что глубина решающих деревьев также влияет на качество предсказания модели – чем она больше, тем качество выше, поэтому обучение производилось на глубине от 1 до 21 объекта. В дополнение к этому, также перебирались все критерий расщепления, ограничение на число объектов в листьях и минимальное число объектов, при котором выполняется расщепление (Таблица 1).

3.6.4 Алгоритмы градиентного бустинга LightGBM и CatBoost

Также применен метод градиентного бустинга над решающими деревьями, суть которого заключается в обучении некоторого числа моделей (в нашем случае решающих деревьев) с учетом ошибок, полученных на предыдущих моделях. Градиентный бустинг позволяет строить аддитивную функцию в виде суммы решающих деревьев итерационно, по аналогии с методом градиентного спуска [38]. Таким

образом, такой подход позволяет достичь более высокой точности предсказания. В качестве алгоритмов градиентного бустинга воспользуемся LightGBM Classifier и CatBoostClassifier, имеющие схожие функциональные возможности для поддержки автоматической обработки категориальных функций[39].

CatBoost — это библиотека градиентного бустинга, представленная Яндексом в 2017 г., с открытым исходным кодом для контролируемого машинного обучения, содержащая две инновации: упорядоченную целевую статистику и упорядоченное повышение [40]. Отличительной особенностью CatBoost является возможность работы с гетерогенными наборами данных с разными типами данных. CatBoost применяется для задач регрессии, классификациия и ранжирования.

LightGBM —библиотека градиентного бустинга, представленная Microsoft DMKT также в 2017 году. Благодаря быстроте и высокой производительности [41,42], данная модель широко используется при решении регрессии, классификации и других задач ML. Как и CatBoost, LightGBM имеет встроенную поддержку кодирования категориальных переменных.

Алгоритмы градиентного бустинга достаточно быстро переобучаются, поэтому требуется производить точный подбор гипперпараметров для получения качественной модели. Также как и со случайным лесом, важное значение имеет количество решающих деревьев, их глубина и количество листьев (терминальных узлов). Однако, если для случайного леса большая глубина деревьев в большинстве случаев оказывает положительную роль на качество предсказаний, то для градиентных бустингов рекомендуется использовать среднюю глубину, что позволяет достичь баланса между обученностью и обобщенностью. Также важным параметром является скорость обучения, являющейся степенью вклада каждого дерева в прогнозирование модели, типа алгоритма бустинга (gbdt, dart, goss) и степени регуляризации.

Таблица 1 - сетка гиперпараметров обучаемых ML моделей

Название ML модели	Гиперпараметр	Расшифровка гиперпараметра	Значения гиперпараметра
SVM	C	Коэффициент регуляризации C	от 0 до 10 с шагом 0,5
	kernel	Тип ядра	‘rbf’, ‘poly’, ‘sigmoid’.
	gamma	Коэффициент ядра	от 0 до 1 с шагом 0,01
Logistic Regression	solver	Алгоритм оптимизации	“saga”
	penalty	Вид регуляризации	“elasticnet”
	l1_ratio	Соотношение между l1 и l2 регуляризациями	от 0 до 1 с шагом 0,1
	C	Сила регуляризации C	от 0 до 5 с шагом 0,1
RandomForest Classifier	n_estimators	Число решающих деревьев	от 10 до 1000 с шагом 100
	min_samples_split	Минимальное число объектов, при котором выполняется расщепление	от 2 до 50 с шагом 10
	min_samples_leaf	Ограничение на число объектов в листьях	от 2 до 50 с шагом 10
	max_depth	Глубина решающих деревьев	от 1 до 21 с шагом 1
	criterion	Критерий расщепления	“gini”, “entropy”, “log_loss”
LGBM Classifier	learning_rate	Скорость обучения	0,0001, 0,001, 0,01
	max_depth	Глубина решающих деревьев	от 1 до 21 с шагом 1
	n_estimators	Количество решающих деревьев	от 10 до 1000 с шагом 10
	num_leaves	Количества листьев в дереве	от 2 до 50 с шагом 1
	boosting_type	Алгоритм бустинга	“gbdt”, “dart”, “goss”
	reg_alpha	Коэффициент регуляризации l1	от 0 до 1 с шагом 0,1
	reg_lambda	Коэффициент регуляризации l2	от 0 до 1 с шагом 0,1
CatBoost	depth	Глубина решающих деревьев	от 1 до 10 с шагом 1
	learning_rate	Скорость обучения	0,0001, 0,001, 0,01

	iterations	Количество решающих деревьев	от 10 до 1000 с шагом 10
	l2_leaf_reg	Коэффициент регуляризации l2	от 1 до 15 с шагом 1
	max_leaves	Максимальное количество листьев в дереве	от 2 до 50 с шагом 1

3.7 Оценка качества модели ML моделей

Эффективность предсказания дихотомической переменной в ML моделях оценивается различными метриками, рассчитываемые на основе матрицы ошибок (рисунок). Матрица ошибок (матрица неточностей) представляет собой классификацию результатов предсказания моделей на истинно-положительные TP (True Positive), истинно-отрицательные TN (True Negative), ложноположительные FP (False Positive) и ложноотрицательные FN (False Negative), как показано на рисунке 1.

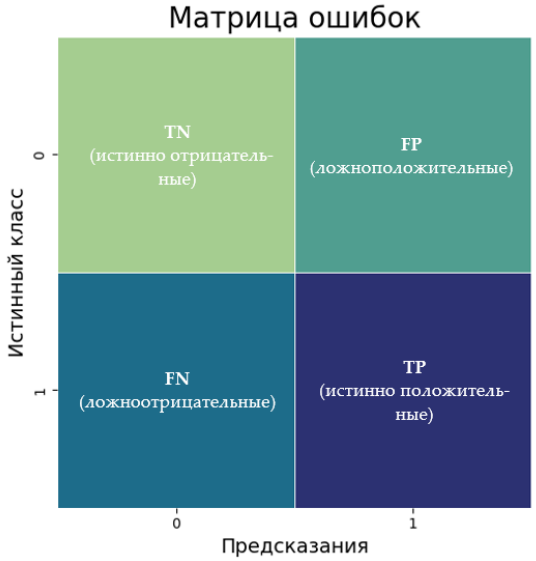


Рисунок 1 – Матрица ошибок

Так как в исследуемых данных имеется небольшой дисбаланс классов, то качество модели была оценена метрикой ROC AUC, невосприимчивой к этой проблеме. Метрика ROC AUC является площадью под кривой ошибок (ROC-кривой), отображающей соотношение между количеством верно и ошибочно классифицированных ответов при варьировании порога решающего правила [43]. Значения метрики находятся в пределах от 0 до 1, при этом 1 – говорит о высоком качестве модели, 0,5 – о случайном предсказании ответов модели, 0 - о ложном предсказании, когда предсказания модели противоположны истинным значениям. Верно классифицированные TPR (True Positive Rate) и ошибочно классифицированные FPR (False Positive Rate) ответы находятся по формулам 1 и 2.

$$TPR = \frac{TP}{TP + FN}$$	(1)
$$FPR = \frac{FP}{FP + TN}$$	(2)

Кроме метрики ROC AUC, взятую в качестве функции потерь при обучении моделей, рассчитаны дополнительные метрики качества для оценки работы моделей [26], такие как:

- Accuracy («правильность») - доля правильных ответов:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

- Precision (Точность) - доля истинно положительных предсказаний среди всех положительных предсказаний модели:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

- Recall (Полнота) - доля положительных предсказаний среди всех положительных случаев:

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

- F1-score (F-1 мера) - среднее гармоническое полноты и точности, принимающее значение от 0 до 1 и позволяющая оценить качество модели при наличии несбалансированного набора данных:

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision * Recall} \quad (5)$$

- AUC-PR - площадь под PR-кривой, отображающей соотношение между метриками полноты (Precision) и точности (Recall). AUC-PR в отличии от ROC AUC восприимчива к небалансу классов.

4. ОБСУЖДЕНИЕ

4.1 Результаты обучения моделей

Результаты обучения моделей сведены в таблице 2, отсортированной по метрике ROC AUC, гиперпараметры лучших моделей представлены в таблице 3. Для проверки на адекватность, в дополнение к рассмотренным моделям была добавлена константная модель DummyClassifier со стратегией классификации “uniform”, генерирующей предсказания случайным образом с равной вероятностью для каждого класса.

Таблица 2 – Метрики качества обученных моделей

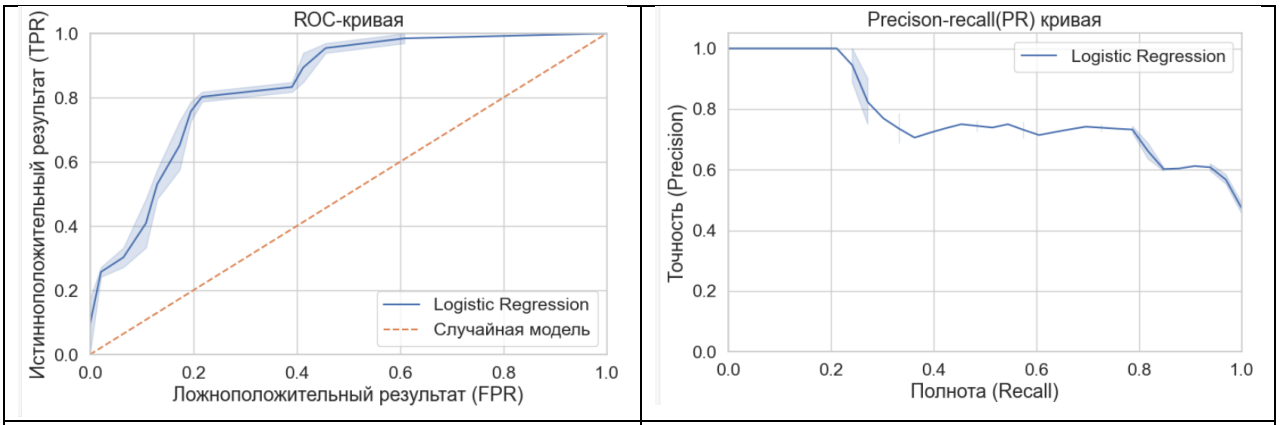
ML модель	ROC-AUC	AUC-PR	Accuracy	Recall	Precision	F1
Logistic Regression	0,779	0,683	0,69	0,685	0,611	0,644
CatBoost Classifier	0,776	0,655	0,693	0,708	0,618	0,657
Logistic Regression (SMOTE)	0,772	0,676	0,69	0,678	0,612	0,641
Random Forest Classifier	0,772	0,661	0,687	0,693	0,607	0,644
LightGBM Classifier (SMOTE)	0,772	0,652	0,69	0,686	0,609	0,642
LightGBM Classifier	0,771	0,647	0,715	0,762	0,631	0,687
CatBoost Classifier (SMOTE)	0,771	0,64	0,712	0,692	0,641	0,664
Support Vector Machine	0,768	0,672	0,687	0,685	0,606	0,642
Support Vector Machine (SMOTE)	0,761	0,671	0,687	0,67	0,61	0,637
Random Forest Classifier (SMOTE)	0,761	0,67	0,671	0,67	0,592	0,626
Dummy Model	0,5	0,411	0,519	0,523	0,431	0,472

Таблица 3 – Гиперпараметры лучших моделей

ML модель	Гиперпараметры
Support Vector Machine	kernel: 'rbf'; gamma: 0,04; C: 9,0; class_weight: 'balanced'
Support Vector Machine (SMOTE)	kernel: 'sigmoid'; gamma: 0,03; C: 6,0
Logistic Regression	penalty: 'elasticnet'; solver: 'saga'; l1_ratio: 0,8; C: 1; class_weight: 'balanced'
Logistic Regression (SMOTE)	penalty: 'elasticnet'; solver: 'saga'; l1_ratio: 0,6; C: 1

Random Forest Classifier	n_estimators: 300; min_samples_split: 12; min_samples_leaf: 32; max_depth: 12; criterion: 'log_loss'; class_weight: 'balanced'
Random Forest Classifier (SMOTE)	n_estimators: 100; min_samples_split: 32; min_samples_leaf: 22; max_depth: 1
LightGBM Classifier	reg_lambda: 0,2; reg_alpha: 0,6; num_leaves: 31; n_estimators: 490; max_depth: 1; learning_rate: 0,01; boosting_type: 'gbdt'; class_weight: 'balanced'
LightGBM Classifier (SMOTE)	reg_lambda: 0,7; reg_alpha: 0,9; num_leaves: 47; n_estimators: 865; max_depth: 15; learning_rate: 0,001; boosting_type: 'goss'
CatBoost Classifier	max_leaves: 16; learning_rate: 0,001; l2_leaf_reg: 14; iterations: 720; depth: 4; class_weight: 'balanced'
CatBoost Classifier (SMOTE)	max_leaves: 16; learning_rate: 0,001; l2_leaf_reg: 14; iterations: 720; depth: 4

Согласно таблице 2, все обученные ML модели показали результаты на порядок выше предсказаний константной модели. Наилучшим методом борьбы с несбалансированной выборкой оказался метод взвешивая классов. По метрике ROC AUC лучшей моделью выбрана логистическая регрессия с методом взвешивания классов в качестве борьбы с дисбалансом классов (0,78), что выражается в простоте самой модели и хорошей аппроксимации. Данная модель показала также лучшие результаты по метрике AUC-PR (0,68), что указывает на успешное прогнозирование целевого признака в условиях несбалансированной выборки. Модель была проверена на тестовой выборке и показала отличные результаты -ROC AUC со значением в 0.84. Значение метрики оказалось выше, чем на валидационной выборке, что указывает на отсутствие переобучения модели, однако, большие расхождения между метриками на разных выборках сигнализирует о проблеме недостаточного количества данных (в тестовой выборке было всего 79 объектов). Проблема малой выборки наглядно видна на ROC и PR кривых, на которых имеются резкие «переломы» из-за недостаточности данных (рисунок 1).



(а) (б)

а – ROC кривая, б – Precision-recall кривая

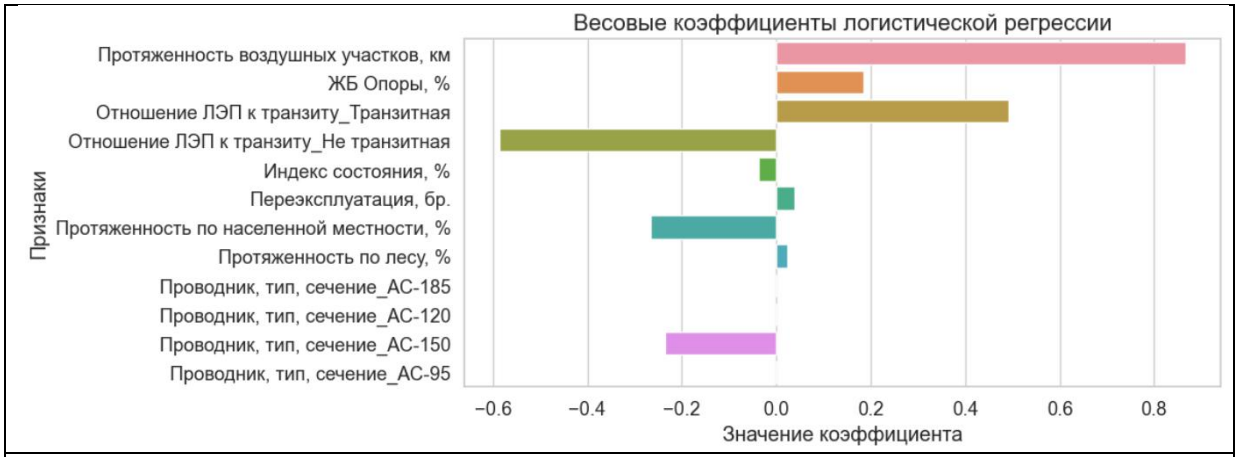
Рисунок 1 - ROC и PR кривые на тестовой выборке для лучшей логистической регрессии

Несмотря на это, модель справляется с поставленной задачей, а графическое отображение ROC кривой (Рисунок 1 а) позволяет утверждать, что при изменении порога классификации можно достичь высокого результата. Например, если необходимо, чтобы в предсказание модели попадали все 100 % ЛЭП, на которых наблюдаются отказы (отсутствие ошибок I рода), то согласно графику ROC кривой (Рисунок 1 а), к этому классу будет отнесено до 60% ЛЭП, на которых отказов не наблюдалось (ошибки II рода). И наоборот, если нужна точность предсказаний для обоих классов, то можно вычислить, что при пороге классификации приблизительно в 0,4

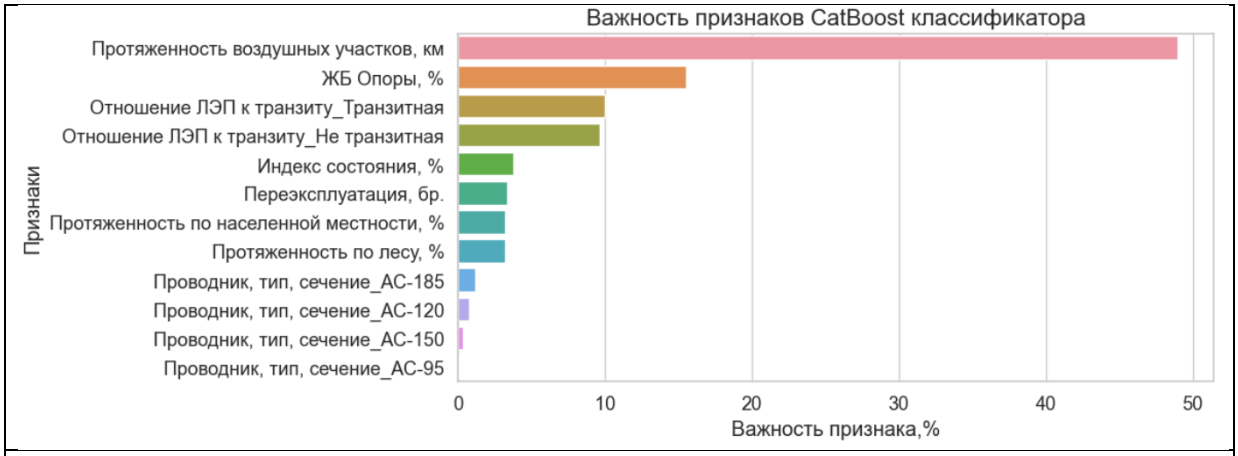
истинно-положительные предсказания TPR равны 0,8, тогда как ложноположительные предсказания составят всего лишь 0,2. На PR кривой (рисунок 1 б) видно, что точность (Precion) стабилизируется на значении 0,72…0,75 при полноте (Recall) от 0,35 до 0,78 при соответствующих порогах классификации 0,45…0,58. Таким образом, знание рассмотренных выше метрик при определенных уровней порога классификации позволяет регулировать количество ошибок I и II рода.

4.2 Анализ важности признаков с помощью встроенных методов в модели.

В статье [], а также в пункте 3.1 настоящей статьи проведен анализ корреляционной зависимости между признаками с помощью коэффициента корреляции phi ϕ_k, по сути являющимся одним из статистических методов (методов фильтрации). В ходе такого анализа были определены степень влияния признаков на целевую переменную - «факт отказа на ЛЭП», и соответственно, должным образом признаки были обработаны для достижения наилучшей корреляции. В тоже самое время, многие ML модели во время обучения сами определяют степень влияния признаков на итоговый результат предсказания, и соответственно, оптимизируют их влияние для получения наилучшего результата. Анализ таких данных позволяет достичь еще лучшего понимания важности признаков в поставленной задаче. Отбор признаков в соответствии с параметрами моделей называются встроенными методами оценки важности признаков [44]. Касательно регрессионных моделей к таким методам можно отнести анализ весовых коэффициентов для каждого признака (степень регуляризации), а для моделей, основанных на решающих деревьях, - анализ показателей важности каждого признака, рассчитанных на этапе обучения с помощью критерия Джини. На рисунке 1 отображена важность признаков для обученных в настоящей статье лучших ML моделей: логистической регрессии и CatBoost классификатора.



(a)



(б)

а - Логистическая регрессия; б - CatBoost классификатор

Рисунок 1 – Важность признаков встроенными методами в ML модели.

Ожидаемо, наибольший вклад в прогнозирование обоих моделей вносит признак «Протяженность воздушных участков ЛЭП», при этом влияние этого признака на предсказание CatBoost классификатора достигает 48 %. Таким образом, еще раз подтверждается факт, что с увеличением протяженности ЛЭП увеличивается количество отказов на ней, поэтому для оценки эффективности работы энергоснабжающих организаций часто используется параметр потока отказов - количество отказов на определенную длину ЛЭП.

Признаки, определяющие протяженность ЛЭП по населенной местности и лесу были переведены в относительные единицы от общей протяженности ЛЭП, поэтому их влияние на результаты моделей оказались не такими существенными. Несмотря на это, интересным фактом является наличие обратной зависимости отказов ЛЭП от относительной протяжённости по населённой местности (весовой коэффициент признака LR равен -0,27 при важности у CatBoost классификатора в 3%), то есть чем больше протяженность ЛЭП по населенной местности, тем ниже вероятность отказов на ней. В свою очередь, протяжённость по лесу имеет важность у CatBoost классификатора в 3 % (Рисунок 1 б), а весовой коэффициент у LR модели всего лишь 0.02, что достаточно мало и говорит о том, что древесно-кустарниковая растительность хорошо расчищается в охранной зоне ЛЭП 110 кВ и их влияние на отказы не проявляется.

В предыдущем исследовании по EDA [] установлено, что количество опор явно коррелирует с протяженность ЛЭП, поэтому в этом исследовании использовался синтетический признак, отражающий относительное число ЖБ опор от всего числа опор (ЖБ плюс металлические). Выявлено, что большее содержание ЖБ опор относительно металлических опор влечет за собой повышение вероятности отказов на ЛЭП. Так важность этого признака у CatBoost классификатора равна 15,5 %, весовой коэффициент логистической модели – 0,2.

Сильное влияние на вероятность отключений на ЛЭП 110 кВ дает факт транзитности линии. Данный признак является категориальным и в предложенном в данной работе алгоритме по подготовке данных этот признак был обработан методом горячего кодирования (One Hot Encoding), в результате которого образовались 2 признака: факт транзитности линии и факт не тразитности, причем оба признака влияют на предсказание CatBoost модели с важностью почти в 10 %. При этом весовые коэффициенты логистической регрессии также подтверждают сильное влияние этого признака на предсказание факта отключения электроэнергии на ЛЭП, причем факт транзитности имеет прямую зависимость с целевой переменной (весовой

коэффициент равен 0,49), а факт не тразитности – обратную (весовой коэффициент равен -0,59).

Срок эксплуатации, который выражается через признак «Переэксплуатация», оказывает довольно слабое влияние на целевую переменную (важность признака равна 3,3 %, весовой коэффициент – 0,04), что говорит о том, что линии такого напряжения всегда стараются поддерживать в прекрасном состоянии и постоянно модернизируются.

Довольно неочевидно выделяется влияние индекса состояния ЛЭП на целевую переменную. Логически рассуждая, можно предположить, что чем хуже состояние линии, тем больше вероятность отказа этой линии. Однако, этот факт при анализе важности признаков не наблюдается - так важность этого признака для предсказания CatBoost классификатора равна всего лишь 3,8 %, а весовой коэффициент LR модели – минус 0,04. Такое слабое влияние может быть объяснено искусственным управлением данным параметром в отчетных документах электросетевых компаний в зависимости от необходимости проведения планово-предупредительных работ.

Рассматривая важность типа проводника ЛЭП на предсказание модели, необходимо отметить, что весовые коэффициенты для проводников типа АС-185, АС-120, АС-95 равны нулю, что говорит об отсутствии их влияния на отказы ЛЭП. Однако, имеется достаточная отрицательная зависимость влияния проводника АС-150 на целевую переменную (-0,24). Данный факт объясняется малой выборкой и случайной спецификой электрических сетей в Орловской области – проводники типа АС сечением 150 и 185 чаще являются транзитными (в процентном соотношении в 78% и 80% случаях, соответственно), в то время как проводники сечением 150 – только в 56% случаях.

5. Выводы

В рамках данного исследования было предложено использование моделей машинного обучения, а именно классификаторов SVM, LogisticRegression, RandomForestClassifier, LightGBM Classifier и CatBoostClassifier, для прогнозирования отключений электрической энергии на линиях электропередачи 110 кВ на основе данных по параметрам самих линий. Данные для моделей предложено подготавливать методом горячего кодирования One-Hot Encoding для категориальных переменных и методом стандартизацией данных Standard Scaler для количественных. Для автоматизации процесса преобразования данных и устранения возможности их утечки используется пайплайн (Pipeline) и компоновщик разнородных признаков Column Transformer. Настройка гиперпараметров классификаторов осуществлена методами оптимизации случайных параметров RandomizedSearchCV и сеточного поиска GridSearchCV.

По метрике ROC AUC лучшей моделью стала логистическая регрессия с методом взвешивания классов в качестве борьбы с дисбалансом классов, показавшая результат в 0,78, что выражается в простоте самой модели и хорошей аппроксимации. Данная модель показала также лучшие результаты по метрике AUC-PR (0,68), что указывает на успешное прогнозирование целевого признака даже в условиях несбалансированной выборки. На тестовой выборке логистическая регрессия также показала отличный результат - ROC AUC со значением в 0,84, что подтвердило отсутствие переобучения модели, но показало влияние проблемы недостаточного количества данных, выраженного в значительном превышении метрик на тестовой выборке в сравнении с валидационной.

Использование встроенных методов для оценки важности признаков для логистической регрессии и CatBoost классификатора позволило степень влияния параметров ЛЭП на факт отключения электрической энергии. Ожидаемо, наибольший

вклад в прогнозирование обоих моделей вносит признак «Протяженность воздушных участков ЛЭП» (до 48 % у CatBoost классификатора достигает). Протяженность ЛЭП по населенной местности показывает наличие обратной зависимости с целевой переменной, то есть чем больше протяженность ЛЭП по населенной местности, тем ниже вероятность отказов на ней. В тоже самое время протяженность по лесу имеет довольно слабую корреляцию с фактом отказов. Касательно типов опор, выявлено, что большее содержание ЖБ опор относительно металлических опор влечет за собой повышение вероятности отказов на ЛЭП. Так важность этого признака у CatBoost классификатора равна 15,5 %, весовой коэффициент логистической модели – 0,2. Сильное влияние на вероятность отключений на ЛЭП 110 кВ дает факт транзитности линии - важность этого признака у CatBoost классификатора равна 10 %, причем факт транзитности имеет прямую зависимость с целевой переменной (весовой коэффициент логистической регрессии равен 0,49), а факт не тразитности – обратную (весовой коэффициент равен -0,59). Срок эксплуатации и состояния ЛЭП оказывают довольно слабое влияние на целевую переменную.

Результаты данного исследования показывают возможность прогнозирования отключений электрической энергии на линиях электропередачи 110 кВ на основе данных по параметрам самих линий, что видно из полученных метрик качества ML моделей. Однако, из-за ограниченного набора данных (395 объектов) не удалось достичь постоянного результата (наблюдаются значительные расхождения между метриками качества на тестовой и валидационной выборках), поэтому в рамках будущего исследования необходимо расширить набор данных за счет включения линий других регионов и/или анализа дополнительных периодов фиксации отказов на ЛЭП.

Author Contributions:

Funding: This research received no external funding.

Data Availability Statement: Data are available on request to abrar0613@gmail.com

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Long L. Research on status information monitoring of power equipment based on Internet of Things // Energy Reports. Elsevier, 2022. Vol. 8. P. 281–286.
2. Sun B. et al. Distributed optimal dispatching method for smart distribution network considering effective interaction of source-network-load-storage flexible resources // Energy Reports. Elsevier, 2023. Vol. 9. P. 148–162.
3. Базан Т.В., Галабурда Я.В., Иселёнок Е.Б. Анализ отключений воздушных линий 35-750 кВ // Актуальные проблемы энергетики. Электроэнергетические системы. Минск, республика Беларусь: БНТУ, 2020. P. 114–116.
4. Yang L., Teh J. Review on vulnerability analysis of power distribution network // Electric Power Systems Research. Elsevier, 2023. Vol. 224. P. 109741.
5. Shakiba F.M. et al. Real-Time Sensing and Fault Diagnosis for Transmission Lines // International Journal of Network Dynamics and Intelligence. 2022. P. 36–47.
6. Latka M., Hadaj P. Technical and statistical analysis of the failure of overhead lines and its impact on evaluating the quality of the power supply // 2016 Progress in Applied Electrical Engineering, PAEE 2016. Institute of Electrical and Electronics Engineers Inc., 2016.
7. Ильдиряков С.Р., Вафин Ш.И. Статистический анализ провалов напряжения в системе электроснабжения ОАО" Казаньоргсинтез" // Известия высших учебных заведений. Проблемы

энергетики. Федеральное государственное бюджетное образовательное учреждение высшего …, 2011. № 3–4. P. 73–81.

8. Виноградов А.В. et al. Анализ времени перерывов в электроснабжении сельских потребителей и методы его сокращения за счет мониторинга технического состояния линий электропередачи // Вестник ВИЭСХ. Федеральное государственное бюджетное научное учреждение" Федеральный …, 2017. № 2. P. 3–11.
9. Ланин А.В., Полковская М.Н., Якупов А.А. Статистический анализ аварийных отключений в электрических сетях 10Кв // Актуальные вопросы аграрной науки. Федеральное государственное бюджетное образовательное учреждение высшего …, 2019. № 30. P. 45–52.
10. Ратушняк В.С., Ильин Е.С., Вахрушева О.Ю. Статистический анализ аварийных отключений электроэнергии из-за гололедообразования на проводах ЛЭП на территории РФ // Молодая наука Сибири: электрон. науч. журн. 2018. № 1. P. 12.
11. Сбитнев Е.А., Жужин М.С. Анализ аварийности сельских электрических сетей 0, 38 кВ Нижегородской энергосистемы // Вестник НГИЭИ. Государственное бюджетное образовательное учреждение высшего …, 2020. № 11 (114). P. 36–47.
12. Sood S. Power Outage Prediction Using Machine Learning Technique // 2023 International Conference on Power Energy, Environment & Intelligent Control (PEEIC). IEEE, 2023. P. 78–80.
13. Eskandarpour R., Khodaei A. Leveraging accuracy-uncertainty tradeoff in SVM to achieve highly accurate outage predictions // IEEE Transactions on Power Systems. Institute of Electrical and Electronics Engineers Inc., 2018. Vol. 33, № 1. P. 1139–1141.
14. Gururajapathy S.S. et al. Fault location in an unbalanced distribution system using support vector classification and regression analysis // IEEJ Transactions on Electrical and Electronic Engineering. John Wiley & Sons, Ltd, 2018. Vol. 13, № 2. P. 237–245.
15. Doostan M., Chowdhury B.H. Power distribution system equipment failure identification using machine learning algorithms // IEEE Power and Energy Society General Meeting. IEEE Computer Society, 2018. Vol. 2018-January. P. 1–5.
16. Warlyani P. et al. Fault classification and faulty section identification in teed transmission circuits using ANN // International Journal of Computer and Electrical Engineering. IACSIT Press, 2011. Vol. 3, № 6. P. 807–811.
17. Alqudah M., Obradovic Z. Enhancing Weather-Related Outage Prediction and Precursor Discovery Through Attention-Based Multi-Level Modeling // IEEE Access. Institute of Electrical and Electronics Engineers Inc., 2023. Vol. 11. P. 94840–94851.
18. Allen M. et al. Application of hybrid geo-spatially granular fragility curves to improve power outage predictions // J Geogr Nat Disast. 2014. Vol. 4, № 127. P. 2167–2587.
19. Lair W. et al. Windy Smart Grid; Forecasting the Impact of Storms on the Power System // Book of Extended Abstracts for the 32nd European Safety and Reliability Conference. Singapore: Research Publishing Services, 2022. P. 905–912.
20. Montoya-Rincon J.P. et al. On the Use of Satellite Nightlights for Power Outages Prediction // IEEE Access. Institute of Electrical and Electronics Engineers Inc., 2022. Vol. 10. P. 16729–16739.
21. Hou H. et al. Prediction of user outage under typhoon disaster based on multi-algorithm Stacking integration // International Journal of Electrical Power & Energy Systems. Elsevier, 2021. Vol. 131. P. 107123.
22. Li M. et al. Prediction of Power Outage Quantity of Distribution Network Users under Typhoon Disaster Based on Random Forest and Important Variables // Math Probl Eng. Hindawi Limited, 2021. Vol. 2021.

23. Taylor W.O. et al. Community power outage prediction modeling for the Eastern United States // Energy Reports. Elsevier, 2023. Vol. 10. P. 4148–4169.
24. Das S., Kankanala P., Pahwa A. Outage Estimation in Electric Power Distribution Systems Using a Neural Network Ensemble // Energies (Basel). 2021. Vol. 14, № 16. P. 4797.
25. Onaolapo A.K. et al. Event-Driven Power Outage Prediction using Collaborative Neural Networks // IEEE Trans Industr Inform. IEEE Computer Society, 2023. Vol. 19, № 3. P. 3079–3087.
26. Chokr B. et al. Feature extraction-reduction and machine learning for fault diagnosis in PV panels // Solar Energy. Pergamon, 2023. Vol. 262. P. 111918.
27. Maraden Y. et al. Enhancing Electricity Theft Detection through K-Nearest Neighbors and Logistic Regression Algorithms with Synthetic Minority Oversampling Technique: A Case Study on State Electricity Company (PLN) Customer Data // Energies 2023, Vol. 16, Page 5405. Multidisciplinary Digital Publishing Institute, 2023. Vol. 16, № 14. P. 5405.
28. Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин). Москва: MachineLearning, 2011. 141 p.
29. Boutaba R. et al. A comprehensive survey on machine learning for networking: evolution, applications and research opportunities // Journal of Internet Services and Applications. Springer, 2018. Vol. 9, № 1. P. 1–99.
30. Вершинин Д.С., Браништи В.В. РАСЧЕТ ОПТИМАЛЬНОГО ПАРАМЕТРА РАЗМЫТОСТИ ДЛЯ ОЦЕНКИ РОЗЕНБЛАТТА–ПАРЗЕНА С ГАУССОВСКИМ ЯДРОМ // Актуальные проблемы авиации и космонавтики. Федеральное государственное бюджетное образовательное учреждение высшего …, 2021. Vol. 2. P. 109–111.
31. Baak M. et al. A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics // Comput Stat Data Anal. North-Holland, 2020. Vol. 152. P. 107043.
32. Barnard G.A. Introduction to Pearson (1900) On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to have Arisen from Random Sampling. Springer, New York, NY, 1992. P. 1–10.
33. Jang H.S. et al. Solar Power Prediction Based on Satellite Images and Support Vector Machine // IEEE Trans Sustain Energy. Institute of Electrical and Electronics Engineers Inc., 2016. Vol. 7, № 3. P. 1255–1263.
34. Das A. L Logistic Regression. 2021.
35. Jogunuri S. et al. Random forest machine learning algorithm based seasonal multi-step ahead short-term solar photovoltaic power output forecasting // IET Renewable Power Generation. The Institution of Engineering and Technology, 2024.
36. Villegas-Mier C.G. et al. Optimized Random Forest for Solar Radiation Prediction Using Sunshine Hours // Micromachines 2022, Vol. 13, Page 1406. Multidisciplinary Digital Publishing Institute, 2022. Vol. 13, № 9. P. 1406.
37. Дружков П.Н., Золотых Н.Ю., Половинкин А.Н. Реализация параллельного алгоритма предсказания в методе градиентного бустинга деревьев решений // Вестник Южно-Уральского государственного университета. Серия: Математическое моделирование и программирование. Федеральное государственное бюджетное образовательное учреждение высшего …, 2011. № 37 (254). P. 82–89.
38. Салахутдинова К.И., Лебедев И.С., Кривцова И.Е. Алгоритм градиентного бустинга деревьев решений в задаче идентификации программного обеспечения // Научно-технический вестник информационных технологий, механики и оптики. Федеральное государственное автономное образовательное учреждение высшего …, 2018. Vol. 18, № 6. P. 1016–1022.

39. Hancock J.T., Khoshgoftaar T.M. CatBoost for big data: an interdisciplinary review // J Big Data. Springer Science and Business Media Deutschland GmbH, 2020. Vol. 7, № 1. P. 1–45.
40. Prokhorenkova L. et al. CatBoost: unbiased boosting with categorical features // Adv Neural Inf Process Syst. 2018. Vol. 31.
41. Rufo D.D. et al. Diagnosis of Diabetes Mellitus Using Gradient Boosting Machine (LightGBM) // Diagnostics. Multidisciplinary Digital Publishing Institute (MDPI), 2021. Vol. 11, № 9.
42. Singh N.K., Fukushima T., Nagahara M. Gradient Boosting Approach to Predict Energy-Saving Awareness of Households in Kitakyushu // Energies 2023, Vol. 16, Page 5998. Multidisciplinary Digital Publishing Institute, 2023. Vol. 16, № 16. P. 5998.
43. Nusinovici S. et al. Logistic regression was as good as machine learning for predicting major chronic diseases // J Clin Epidemiol. Pergamon, 2020. Vol. 122. P. 56–69.
44. Кудрявцева А.С. Применение встроенных методов отбора признаков для оптимизации модели референциального выбора // Компьютерная лингвистика и интеллектуальные технологии. 2017. Vol. 16, № 23. P. 1–9.