Journal of Siberian Federal University. Engineering & Technologies 2024 17(6): 758–776
~ ~ ~

EDN: FFDERV

УДК 621.316.13:621.315.1:621.315.668: 621.3.019.3

Analysis of the Influence of 110 kV Power Line Parameters on the Probability of Their Failures

Vadim E. Bolshev* and Alina V. Vinogradova
Federal Scientific Agroengineering Center VIM
Moscow, Russian Federation

Received 19.06.2024, received in revised form 13.08.2024, accepted 29.08.2024

Abstract. When considering power supply system equipment, it should be noted that most often power supply interruptions occur due to damage to power lines. It is important for power grid companies to promptly diagnose faults and restore power supply in order to minimize losses. Therefore, determining in advance the probability of power outages based on the parameters of the power lines themselves can help power grid companies modernize much more efficiently and build new power grids with less failure rate. Aim of the article is analysis of the influence 110 kV power line parameters on the probability of their failures. This study consists of preparing a data on power outages of 110 kV power lines by by clustering, removing unnecessary information, creating synthetic parameters and processing them by imputing missing values and removing duplicates. Then exploratory data analysis was carried out, including analysis of statistical characteristics for considered parameters, identification of outliers and anomalies, and correlation analysis. The research was carried out in the Python programming language in the Jupyter Notebook development environment, using the Pandas, NumPy, Matplotlib, Seaborn, Phik libraries. The data for analysis was prepared using grouping and merging methods, the least significant parameters were removed, the "power line service life" parameter was standardized, and the target attribute was synthesized – "The fact of power line outage." The final result was a table containing 10 parameters, including the target feature, and 395 rows. During the analysis of categorical parameters, an imbalance of classes of the target feature, the influence of the type of wire and the transit fact on 110 kV power line failure were identified. An analysis of the distribution of quantitative variables confirmed that a decrease in outage probability is observed with an improvement in technical condition of lines, a decrease in line length, in the number of supports and in the service life of power lines. Correlation analysis allows establishing the absence of a strong correlation with the target feature and the presence of multicollinearity between all parameters reflecting the length of power lines and the number of reinforced concrete supports.

* Corresponding author E-mail address: vadimbolshev@gmail.com
ORCID: 0000-0002-5787-8581 (Bolshev)
ResearcherID: M-8440-2018 (Bolshev)
Scopus ID: 57201923106 (Bolshev)

Keywords: electrical networks, power lines, ETL, power line parameters, power supply reliability, power supply interruptions, power supply outages, power line failures, exploratory data analysis, correlation analysis.

Citation: Bolshev V. E., Vinogradova A. V. Analysis of the influence of 110 kV power line parameters on the probability of their failures. J. Sib. Fed. Univ. Eng. & Technol., 2024, 17(6), 758–776. EDN: FFDERV

Анализ влияния параметров ЛЭП 110 кВ на вероятность их отказов

В. Е. Большев, А. В. Виноградова
Федеральный научный агроинженерный центр ВИМ
Российская Федерация, Москва

Аннотация. Рассматривая оборудование системы электроснабжения, необходимо отметить, что наиболее часто перерывы в электроснабжении случаются из-за повреждений линий электропередачи. Для сетевых компаний важно своевременно диагностировать неисправности и восстанавливать электроснабжение, чтобы минимизировать возникающие потери. Поэтому заблаговременное определение вероятности отключений электроэнергии на основе параметров самих линий может помочь электросетевым компаниям гораздо эффективнее модернизировать и строить новые, менее подверженные отказам электрические сети. Цель исследования – анализ влияния параметров линий электропередачи напряжением 110 кВ на вероятность их отказов. Настоящее исследование заключается в подготовке данных по отключениям ЛЭП 110 кВ путем группирования, удаления ненужной информации, создания синтетических параметров и их обработки путем заполнения пропущенных значений и удалений дубликатов. Затем производился разведочный анализ данных, включающий в себя анализ статистических характеристик рассматриваемых параметров, выявление выбросов и аномалий и корреляционный анализ. Исследование производилось на языке программирования Python в среде разработки Jupyter Notebook, задействованы библиотеки Pandas, NumPy, Matplotlib, Seaborn, Phik. Данные для анализа были подготовлены с помощью методов группировки и слияния, удалены наименее значимые параметры, параметр «срок эксплуатации ЛЭП» стандартизирован, синтезирован целевой признак – «Факт отключения на ЛЭП». Итоговым результатом стала таблица, содержащая 10 параметров, включая целевой признак, и 395 строк. В ходе анализа категориальных параметров были выявлены дисбаланс классов целевого признака, влияние типа провода и транзитности линии на отказ ЛЭП 110 кВ. Анализ распределения значений количественных переменных подтвердил, что снижение вероятности отключений наблюдается с улучшением технического состояния ЛЭП, уменьшением длины линий, количества опор и снижением срока эксплуатации ЛЭП. Корреляционный анализ позволил установить отсутствие сильной корреляции с целевой переменной и наличие мультиколлинеарности между всеми параметрами, отображающими протяжённость ЛЭП и количество ЖБ опор.

Ключевые слова: электрические сети, линии электропередачи, ЛЭП, параметры ЛЭП, надежность электроснабжения, перерывы в электроснабжении, отключения электроэнергии, отказы ЛЭП, разведочный анализ данных, корреляционный анализ.

Цитирование: Большев В. Е. Анализ влияния параметров ЛЭП 110 кВ на вероятность их отказов / В. Е. Большев, А. В. Виноградова // Журн. Сиб. федер. ун-та. Техника и технологии, 2024, 17(6). С. 758–776. EDN: FFDERV

1. Введение

Передача электроэнергии от электростанции до потребителя осуществляется через широкую и сложную распределительную систему, включающую в себя линии электропередачи (воздушные и кабельные), подстанции и множество контрольно-измерительных приборов. Чем более развита, постоянно совершенствуется и модернизируется система, тем больше вероятность надежного снабжения электроэнергией каждого потребителя [1, 2].

Однако, несмотря на использование новых, инновационных решений, полностью исключить возникновение аварий невозможно. Рассматривая оборудование системы электроснабжения, необходимо отметить, что наиболее часто перерывы в электроснабжении случаются из-за повреждений линий электропередачи (ЛЭП) вследствие их территориальной протяженности и подверженности влиянию климатических воздействий [3, 4]. Наиболее часто применяемый параметр надежности электроснабжения, поток отказов, для ЛЭП в несколько раз превышает поток отказов для трансформаторов и другого оборудования энергосистем. Так, доля отказов высоковольтных линий электропередачи составляет 35–50 % от всех отключений электрического оборудования в системах электроснабжения 35–750 кВ [5]. Для сетевых компаний важно своевременно диагностировать неисправности и восстанавливать электроснабжение, чтобы минимизировать возникающие потери. Электросетевые компании ведут учёт отказов в работе всех элементов сети и на основе определения участков с более частыми отказами выбирают первоочередность ремонта систем электроснабжения. Большинство мер, используемых сегодня, являются ретроспективными, то есть меры применяются после повреждения оборудования, а все превентивные решения принимаются исключительно на основе прошлого опыта.

Поэтому одной из задач внедрения новых технологий в строительстве электрических сетей является сокращение времени перерывов в электроснабжении независимо от причины отключения электроэнергии. Сбор и анализ данных об аварийных отключениях может помочь определить причины, оказывающие наибольшее влияние на изменение уровня отказов. В свою очередь, заблаговременное определение вероятности отключений электроэнергии на основе этих данных за счет анализа влияния параметров ЛЭП на их отказы может помочь электросетевым компаниям гораздо эффективнее как модернизировать существующие электрические сети, так и строить новые, менее подверженные отключениям электрической энергии.

Цель исследования – анализ влияния параметров линий электропередачи напряжением 110 кВ на вероятность их отказов.

2. Обзор литературы

Существует достаточно большое количество исследований, посвящённых анализу отключений электрической энергии в системах электроснабжения от 0,4 кВ до 750 кВ, которые можно разделить по типу исследуемых данных и методам их обработки. Важное значение в исследованиях отказов электрических сетей относится к применению методов математической статистики. Так, в работах [6–8] проведён анализ отключений электроэнергии по причинам отключений и их вероятности возникновения, количеству и времени перерывов электроснабжения. В каждой работе авторы предлагают мероприятия по снижению влияния перерывов электроснабжения на потребителей электроэнергии. В статье [9] представлены результаты статистического анализа гололедных аварий на линиях электропередачи. В работе [10] в допол-

нение к статистическому анализу временных рядов аварийных отключений в нижегородской энергосистеме проведен регрессионный анализ посредством реализации метода наименьших квадратов.

Много работ посвящено применению методов искусственного интеллекта для прогнозирования отказов [11]. В работах [12, 13] предлагается применение алгоритма опорных векторов (Support Vector Machine – SVM) для прогнозирования отключений компонентов энергосистемы, а также для определения места повреждения в системе. Для точной классификации неисправностей систем электроснабжения в статье [14] используются глубокие нейронные сети. В [15, 16] применяются методы искусственного интеллекта для прогнозирования отключений в электрораспределительных сетях во время неблагоприятных погодных условий.

Проводя обзор существующей литературы по прогнозированию отключений электрической энергии, было установлено, что прогнозирование отказов ЛЭП на основе параметров самих линий не осуществлялось ни в каком виде, что говорит об актуальности проводимого исследования. Необходимо отметить, что существуют работы, рассмотренные выше, по статистическому анализу отказов в электрической сети, в ходе которых определялись причины отключений с указанием определенных элементов, вышедших из строя, например, повреждения изоляторов, обрыв провода и т.д. Однако в этих работах не рассматривалось влияние типов этих элементов на вероятность отказов линии.

Настоящая работа посвящена обработке, выбору и анализу данных, являющихся наиважнейшими составляющими для построения модели машинного обучения, способной с определенной точностью идентифицировать линии электропередачи с наибольшей вероятностью отказов. Результатом работы будет как анализ влияния параметров ЛЭП на вероятность отказов, так и рекомендации по выбору параметров ЛЭП для построения модели машинного обучения.

3. Методология и материалы

Настоящее исследование построено на данных по отключениям электрической энергии в электрических сетях Орловской области и данных по параметрам определенных ЛЭП 110 кВ, находящихся на балансе сетевых компаний [17, 18].

Данное исследование состоит из трех этапов. Первый этап заключается в ознакомлении с данными, их подготовке путем группирования, удаления ненужных показателей, в создании синтетических параметров и их обработке путем заполнения пропущенных значений и удалений дубликатов. Второй и третий этапы представляют собой разведочный анализ данных, включающий в себя анализ статистических характеристик данных, выявление выбросов и аномалий и корреляционный анализ показателей.

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 вычислений, включая методы статистической обработки, – NumPy (ver. 1.26.3), для статистической визуализации данных – Matplotlib (ver. 3.8.0) и Seaborn (ver. 0.12.2), для проведения методов корреляционного анализа – Pandas и Phik (ver. 0.12.4).

4. Результаты и обсуждение

4.1. Ознакомление с данными и их подготовка

Ознакомление с данными

Набор данных по отключениям электрической энергии представлен четырьмя таблицами с информацией об отказах с 1 января 2018 по 31 декабря 2023 гг. При конкатенации таблиц было образовано довольно-таки существенное число дубликатов в количестве 5778 шт., после их удаления количество строк объединенной таблицы получилось 53768 шт. Информация, представленная в объединенной таблице, составила:

– напряжение сети;
– дата и время возникновения события;
– вид технологического отключения;
– наличие автоматического повторного включения;
– успешность ручного повторного включения;
– дата и время прекращения электроснабжения потребителей, восстановления электроснабжения потребителям, восстановления нормальной (доаварийной) схемы;
– количество обесточенных потребителей, трансформаторных подстанций, населённых пунктов, населения, социально значимых объектов;
– длительность перерывов электроснабжения;
– причина отключения и описание работы релейной защиты.

Для проведения анализа влияния параметров линий электропередачи 110 кВ на вероятность их отказов нам понадобились характеристики самих этих ЛЭП. Была обработана таблица с данными по ЛЭП напряжением 110 кВ, согласно которой в районных электрических сетях Орловской области содержится 66 линий, а их характеристики представлены следующими параметрами:

– индекс состояния ЛЭП, %;
– проводник, тип, сечение ЛЭП;
– год, с которого ЛЭП в эксплуатации;
– срок эксплуатации, лет;
– протяженность воздушных участков ЛЭП, протяженность по цепям, по трассе, по лесу, по населенной местности, длина кабельных участков ЛЭП, км;
– количество ЖБ опор, металлических опор, опор из чистого дерева, опор на ЖБ пасынках, шт.;
– отношение ЛЭП к транзиту;
– класс напряжения, кВ.

Группировка данных по аварийным отключениям и параметрам ЛЭП 110 кВ

Следующим этапом было соединение полученных таблиц и группировка данных с целью определения целевого признака для поставленной задачи. Были проведены следующие шаги:

– все данные по аварийным отключениям были отсортированы по классу напряжения (“110 кВ») и по виду технологического отключения («Аварийное»), так как в поставленной задаче необходимо прогнозировать именно аварийные отключения;

– далее методом слияния таблиц по общему ключу (метод «merge» библиотеки «Pandas») эти данные были объединены с данными по характеристикам ЛЭП 110 кВ так, чтобы в итоговую таблицу вошли все ЛЭП, для которых не были зафиксированы отключения за 2018–2023 гг., а значит, изначально не были представлены в исходной таблице. Поэтому для этих ЛЭП вся информация по отключениям электроэнергии была приравнена к нулю;

– далее полученная таблица была сгруппирована таким образом, чтобы для каждой ЛЭП в каждый рассматриваемый год были посчитаны количество отключений электроэнергии и факт этих отключений (были ли отключения в определенном году – да или нет);

– таким образом, получилась таблица с информацией по отключениям электроэнергии, привязанной ко всем ЛЭП напряжением 110 кВ и тем годам, в которые происходили эти отключения (данные доступны за 2018–2023). Те года, которые не указаны в таблице, означали, что отключения электроэнергии в это время для определенных ЛЭП не происходили. А значит, можно было добавить эти года в таблицу, зафиксировав количество отключений, равным нулю. При этом в предоставленных данных была линия электропередачи, введённая в эксплуатацию в 2019 году, поэтому эта ЛЭП за 2018 г. не учитывалась. Как итог – получилась таблица, содержащая 395 строк.;

– затем методом слияния таблиц по общему ключу к каждой ЛЭП были присоединены данные по характеристикам ЛЭП 110 кВ.

В итоге получилась таблица, содержащая информацию по всем ЛЭП 110 кВ, распределенную по годам, с информацией о факте отключения, которую в последующем будем использовать в качестве целевого признака. Далее была проведена работа с обработкой параметров, полученных на предыдущем этапе, содержащая шаги до разведочного анализа данных, а именно:

– была исправлена информация по параметру «срок эксплуатации ЛЭП», что было необходимо из-за того, что эти данные относились к моменту выгрузки таблицы с параметрами ЛЭП, то есть 2022 г, а отключения фиксировались в течение 2018–2023 гг. Таким образом, этот параметр рассчитывался как разница между годами возникновения аварийного события и ввода в эксплуатацию ЛЭП;

– затем был создан синтетический параметр «Переэксплуатация» из существующего «Срок эксплуатации». Признак показывает, вышла ли ЛЭП за установленный нормативный срок службы (35 лет) или нет. Значение выше единицы говорит, что срок фактической эксплуатации превышает нормативный, ниже – не превышает. В дальнейшем этот параметр был проверен методом корреляционного анализа и показал более высокую взаимосвязь с целевым признаком, чем параметр «Срок эксплуатации», поэтому последний был удален;

– после всех преобразований и выделения целевого признака были удалены параметры, содержащие избыточную для машинного обучения информацию, например диспетчерское наименование ЛЭП. Были удалены параметры, отражающие длину кабельных участков и тип линий, так как первый параметр был представлен только одним значением (только одна ЛЭП частично была выполнена кабельной линией), а второй – только одним уникальным значением «ВЛ», что сделало невозможным их использование для дальнейшего анализа. Параметры, содержащие информацию о количестве опор на ЖБ пасынках и из чистого дерева, также были удалены, так как ЛЭП напряжением 110 кВ в Орловской области строятся только на ЖБ и металлических опорах;

– в случае несоответствия типов данных представленной информации они были заменены, а оставшиеся показатели были проверены на отсутствие пропусков и дубликатов. Итоговым результатом стала таблица, содержащая 10 параметров, включая целевой признак, и 395 строк.

4.2. Анализ распределения количественных и категориальных переменных в разрезе целевой переменной, выявление выбросов и аномалий

Разведочный анализ (от англ. exploratory data analysis – EDA) представляет собой анализ данных для выявления общих закономерностей их статистических характеристик, уделяя особое внимание пяти ключевым аспектам, таким как меры центральной тенденции (включающие среднее значение, моду и медиану), меры разброса (включающие стандартное отклонение и дисперсию), форма распределения и наличие выбросов, а также корреляционный анализ [19, 20].

Рассмотрим основные статистические характеристики параметров в разрезе целевой переменной, что позволит определить влияние каждого на отключения электроэнергии на ЛЭП напряжением 110 кВ. Посмотрим на распределение значений категориальных признаков, включая целевой признак (рис. 1).

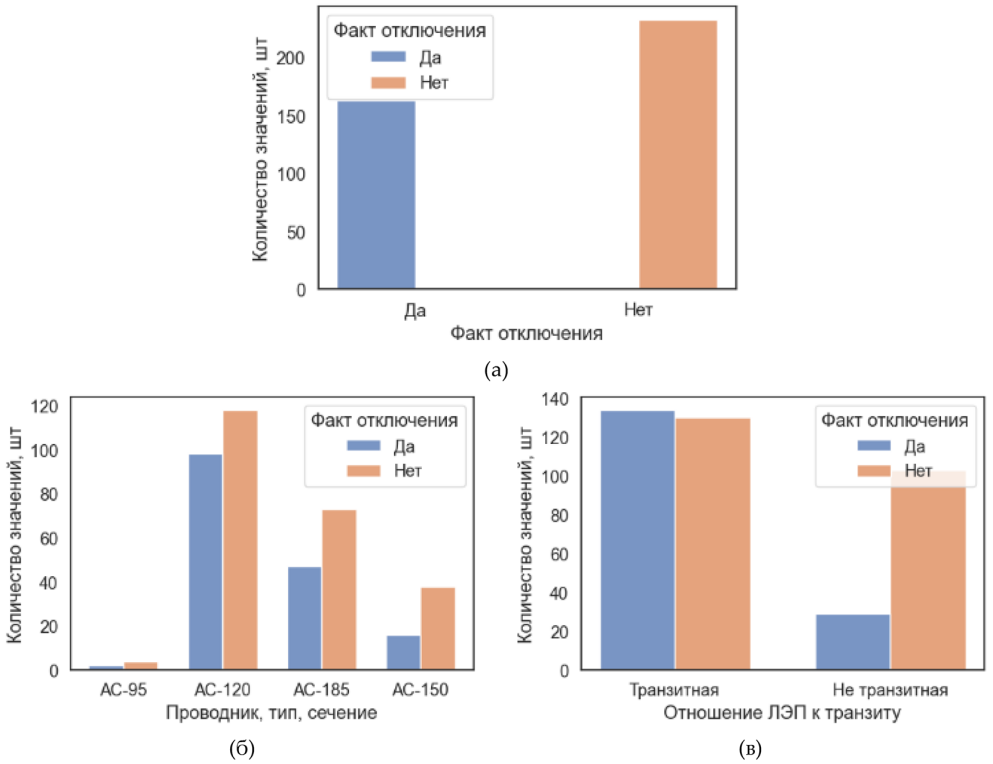


Рис. 1. Распределение значений категориальных параметров в разрезе целевой переменной: (а) гистограмма распределения целевого признака «Факт отключения», (б) гистограмма распределения признака «Проводник, тип, сечение», (в) гистограмма распределения признака «Отношение ЛЭП к транзиту»

Fig. 1. Values distribution of categorical parameters in the context of the target variable: (a) Histogram of distribution of the target attribute “Disconnection Fact”, (б) Histogram of distribution of the attribute “Conductor, type, section”, (в) Histogram of distribution of the attribute “Relationship of power lines to transit”

Рис. 1а показывает небольшой дисбаланс классов целевого признака «Факт отключения» в районе 20 %, что может вызвать определенные трудности в случае, если понадобится проводить классификацию с помощью алгоритмов машинного обучения, восприимчивых к этой проблеме. Также этот факт стоит учитывать при анализе распределения параметров ЛЭП в разрезе целевого признака, так как количество значений параметров для ЛЭП, не подвергавшихся отключениям электроэнергии, всегда будет превышать количество значений ЛЭП с отказами.

На рис. 1б представлено распределение типов проводов по ЛЭП напряжением 110 кВ, из которого видно, что наиболее популярным является неизолированный сталеалюминиевый провод с жилой сечением 120 мм^2 (216 линий), далее следует провод сечением 185 мм^2 (120 линий) и 150 мм^2 (54 линии). Провод АС-95 используется крайне редко – только в 6 случаях. Можно отметить, что вероятность отключения электроэнергии увеличивается (до 45 %) при использовании АС сечением 120 мм^2, тогда как для провода АС-150 она составляет 29 %, для АС-185–39 %. Данный факт может быть связан с отношением ЛЭП к транзиту, так как провода типа АС сечением 150 и 185 чаще являются транзитными (в процентном соотношении в 78 и 80 % случаях соответственно), чем АС-150 (56 % случаев). На рис. 1в видна явная зависимость отношения ЛЭП к транзиту на вероятность отключения электроэнергии. Из 132 не транзитных ЛЭП только на 29 линиях были зафиксированы отказы, тогда как на транзитных ЛЭП в количестве 264 штук отключения электроэнергии наблюдались уже более чем у половины линий (134 шт.).

Также рассмотрим основные статистические характеристики количественных параметров в разрезе целевой переменной. Для этого построим два вида графиков: первый – распределение значений параметра, второй – диаграмма размаха. Диаграмма размаха («ящик с усами») представляет собой график, на которой в графическом виде будут отображены медиана, нижний и верхний границы интерквартильного расстояния – Interquartile range (IQR), минимальное и максимальное значения выборки и выбросы (значения, выходящие за края статистически значимой выборки) [21, 22]. Подсчет выбросов будем производить методом интерквартильных расстояний [23], то есть все значения выше максимальных или ниже минимальных будут рассматриваться в качестве выбросов. Максимальные и минимальные значения рассчитываются по формулам:

$$Lim_{max} = Q3 + 1{,}5 \times IQR, \tag{1}$$

$$Lim_{min} = Q1 - 1{,}5 \times IQR, \tag{2}$$

где $Q3$ – третий квартиль (75-ый перцентиль); $Q1$ – первый квартиль (25-ый перцентиль); IQR – интерквартильное расстояние.

Интерквартильное расстояние (размах) определяется по формуле:

$$IQR = Q3 - Q1. \tag{3}$$

В свою очередь, гистограммы для сглаживания распределения будем строить с графиком распределения плотности, а именно с ядерной оценкой плотности – kernel density estimate (kde), рассчитываемой методом окна Парзена-Розенблатта [24]. Ядерная оценка плотности распределения для выборки $x_1 \ldots x_n$ определяется следующей функцией [25]:

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \quad (4)$$

где K – неотрицательная функция (ядро), в нашем исследовании используется Гауссовское ядро; h – сглаживающий параметр (неотрицательное число), в нашем исследовании $h = 1$.

На рис. 2–4 представлены графики распределения значений (а) и диаграммы размаха (б) для всех характеристик количественных параметров линий электропередачи, рассмотренных отдельно для ЛЭП с отказами и без.

На рис. 2 представлено распределение значений индекса состояния ЛЭП и переэксплуатации в зависимости от факта, наблюдалось ли отключение на ЛЭП или нет, из которого можно сделать следующие выводы:

1. Как видно из графиков, ЛЭП с индексом состояния ниже 60 % не зафиксированы в обоих случаях, выбросов и аномалий в данных нет. Индекс состояния ЛЭП имеет небольшое влияние на вероятность отключения электроэнергии, так, средние (медианные) значения индекса состояния для линий, на которых фиксировались отключения, составляют 82 % (84 %), на которых не фиксировались – 86 % (87 %). Нижняя граница интерквартильного расстояния – IQR (25-ый перцентиль) примерно одинакова (78 и 79 % для ЛЭП с отказами и без соответственно), но есть очевидное различие в верхней границе IQR (75-ый перцентиль) – (78 и 79 % соответственно). То есть те ЛЭП, которые находятся в более хорошем техническом состоянии, менее подвержены отключениям, что в принципе очевидно.

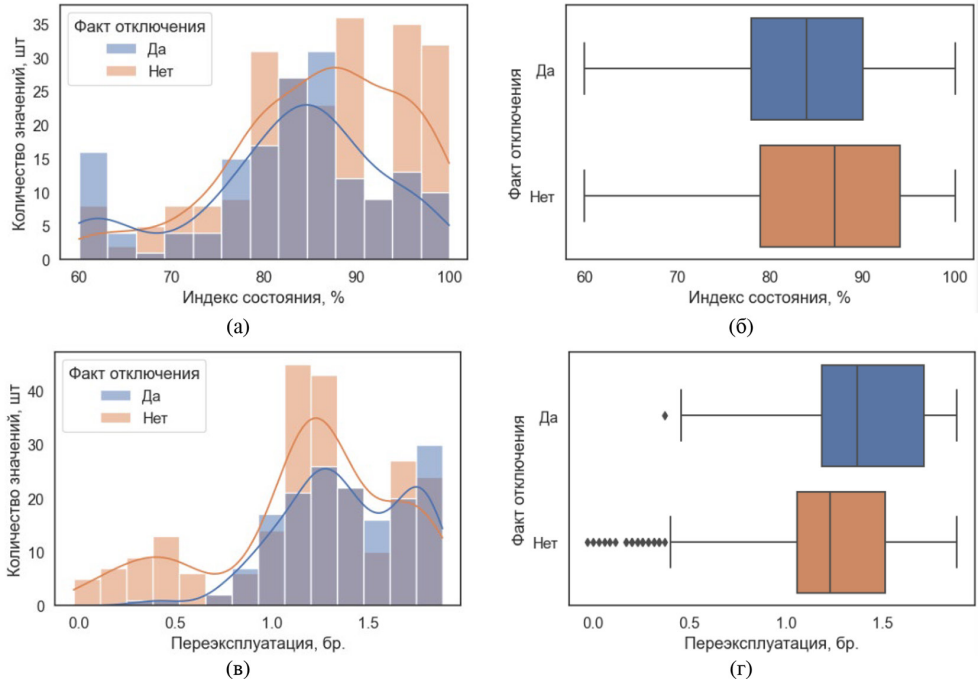


Рис. 2. Распределение значений параметров «Индекс состояния, %», «Переэксплуатация, бр.» в разрезе целевой переменной: (а, в) графики плотности распределения, (б, г) диаграммы размаха

Fig. 2. Value distribution of parameters “Condition index, %”, “Overexploitation” in terms of the target variable (а, в) Distribution density graphs, (б, г) Boxplots

2. Имеется небольшая зависимость отключения электрической энергии от переэксплуатации ЛЭП, параметра, синтезированного из срока эксплуатации ЛЭП. Несмотря на то что графики плотности для обоих значений целевой переменной похожи друг на друга, наблюдается сильная вероятность работы ЛЭП без отказов, если значение переэксплуатации меньше 0,6, то есть срок эксплуатации ЛЭП менее 20 лет. Линий с такими параметрами немного, поэтому на диаграмме размаха они отмечены как выбросы. С увеличением значения переэксплуатации, начиная приблизительно от 1,3, наблюдается тенденция к увеличению аварийных ситуаций на ЛЭП. Касательно статистических характеристик, среднее и медианное значения для ЛЭП с отказами составляют одинаково 1,38, 25-ый и 75-ый перцентили – 1,19 и 1,71 соответственно. Эти значения для ЛЭП, не подвергавшихся отключениям электроэнергии, составили 1,19, 1,22, 1,05 и 1,51 соответственно.

На рис. 3 представлено распределение параметров ЛЭП, характеризующих распределение длины участков ЛЭП в зависимости от факта отключения электроэнергии. Из представленных графиков можно сделать следующие выводы:

1. Подтверждается давно известный факт, что количество отключений электроэнергии зависит от длины ЛЭП, поэтому для оценки эффективности работы энергоснабжающих организаций часто используется параметр потока отказов – количество отказов на определенную длину ЛЭП. Несмотря на то что в рассматриваемом случае отключения наблюдаются при любой длине линий, 25-ый перцентиль длин ЛЭП, подвергавшихся отказам (21,4 км), начинается примерно в том же месте, где закачивается 75-ый перцентиль не подвергавшихся (24,9 км). Средние (медианные) значения длины ЛЭП с отключениями составляют 38,9 (40,0) км, без – 17,4 (13,0) км. Так как основная часть ЛЭП, не подвергавшихся отказам, в основном составляют относительно короткие линии, то наблюдаются выбросы в количестве 10 шт., начиная от 60,97 км.

2. Похожая картина наблюдается при изучении распределения протяженности ЛЭП по лесу. Средние (медианные) значения протяженности ЛЭП с отключениями составляют 12,5 (13,4) км, без – 5,6 (3,3) км. Наблюдаются выбросы для ЛЭП, на которых не фиксировались отказы, в количестве 16 шт., начиная от протяженности по лесу в 17,34 км.

3. Наблюдается некоторое влияние прохождения ЛЭП по населённой местности на возможность отказа, так, нижняя граница интерквартильного расстояния значений протяженности ЛЭП по населённой местности в обоих случаях равна нулю. Верхняя граница IQR (75-ый перцентиль) в случае, если были отказы, составляет 2,37 км, среднее значение – 1,27 км, медиана – 0,45 км. В случае, если отказы не фиксировались, эти параметры равны 0,865 км, 0,73 км и 0,14 км соответственно. Выбросы наблюдаются в обоих случаях: для ЛЭП, на которых фиксировались отказы, в количестве 6 шт., для ЛЭП, на которых не фиксировались отказы,– 28 шт.

На рис. 4 представлено распределение ЛЭП, имеющих ЖБ и металлические опоры. Необходимо отметить, что в изначально предоставляемых данных были ошибки касательно количества опор, так, некоторые ЛЭП имели нулевые значения по всем типам опор, что невозможно. Поэтому при проведении анализа данных признаков не учитывались строки, если сумма значений количества ЖБ и металлических опор равнялось нулю, которых оказалось 60 шт.

Так как количество опор прямо пропорционально протяженности воздушных линий, то ожидаемо малое количество опор будет соответствовать меньшей вероятности возникнове-

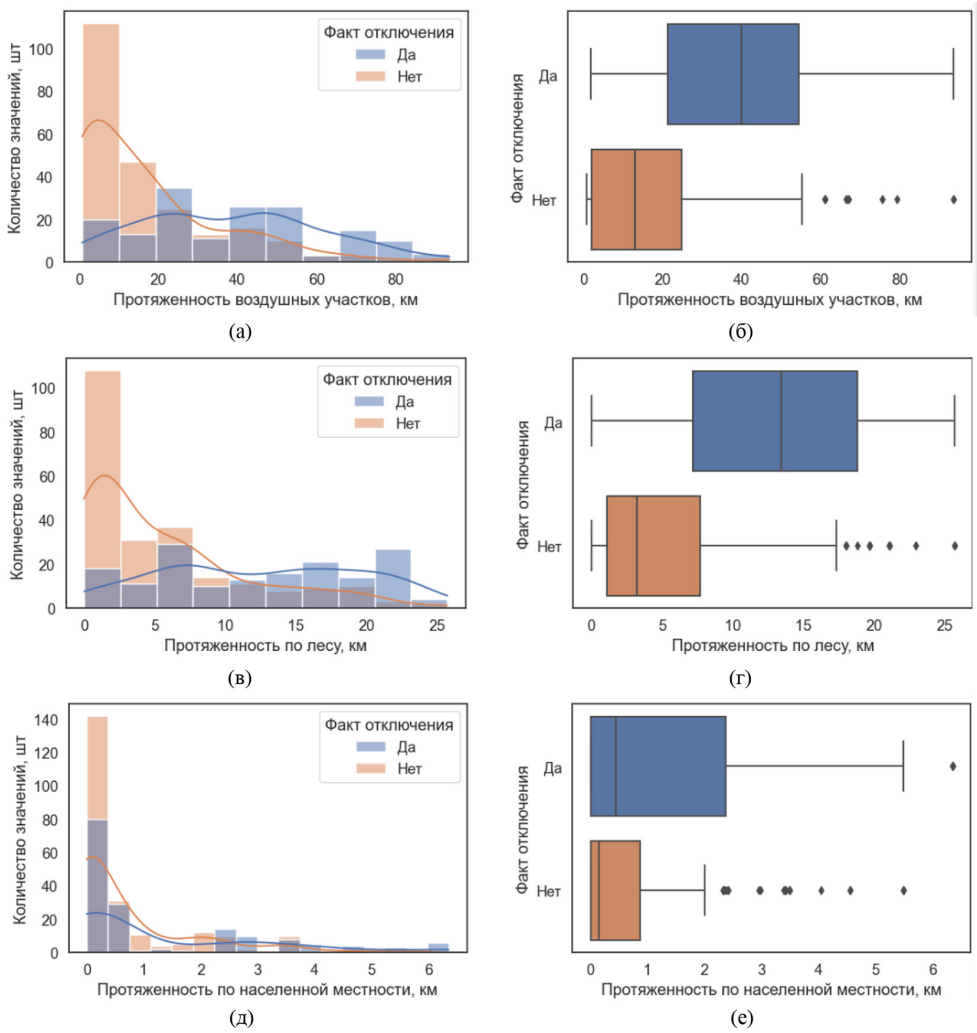


Рис. 3. Распределение значений параметров «Протяженность воздушных участков, км», «Протяженность по лесу, км», «Протяженность по населенной местности, км» в разрезе целевой переменной: (а, в, д) графики плотности распределения, (б, г, е) диаграммы размаха

Fig. 3. Value distribution of parameters “Air section length, km”, “Length through forest, km”, “Length through populated areas, km” in the context of the target variable (а, в, д) Distribution density graphs, (б, г, е) Boxplots

ния отказов. Среднее количество опор на ЛЭП, на которых не фиксировались отключения электроэнергии, составляет около 87 шт., медианное значение – 64 шт., 25-ый перцентиль – 7 шт., 75-ый перцентиль –137 шт., тогда как для ЛЭП с отказами эти значения составляют 174, 184, 74, 272 шт. соответственно. Выбросы имеются только на ЛЭП без отказов в количестве 35 шт.

В свою очередь, количество металлических опор не сильно влияет на вероятность возникновения отказа ЛЭП. Так, в разрезе целевой переменной среднее количество металлических опор на одной линии составляет 21 шт. при условии, что не было зафиксировано отключение электроэнергии, медианное значение составило 14 шт., 25-ый и 75-ый перцентили – 6 и 25 шт. соответственно. Эти статистические значения для ЛЭП с отказами равнялись 28, 20, 9, 28 шт. соответственно, то есть среднее значение и 75-ый перцентиль равны, что говорит о том, что

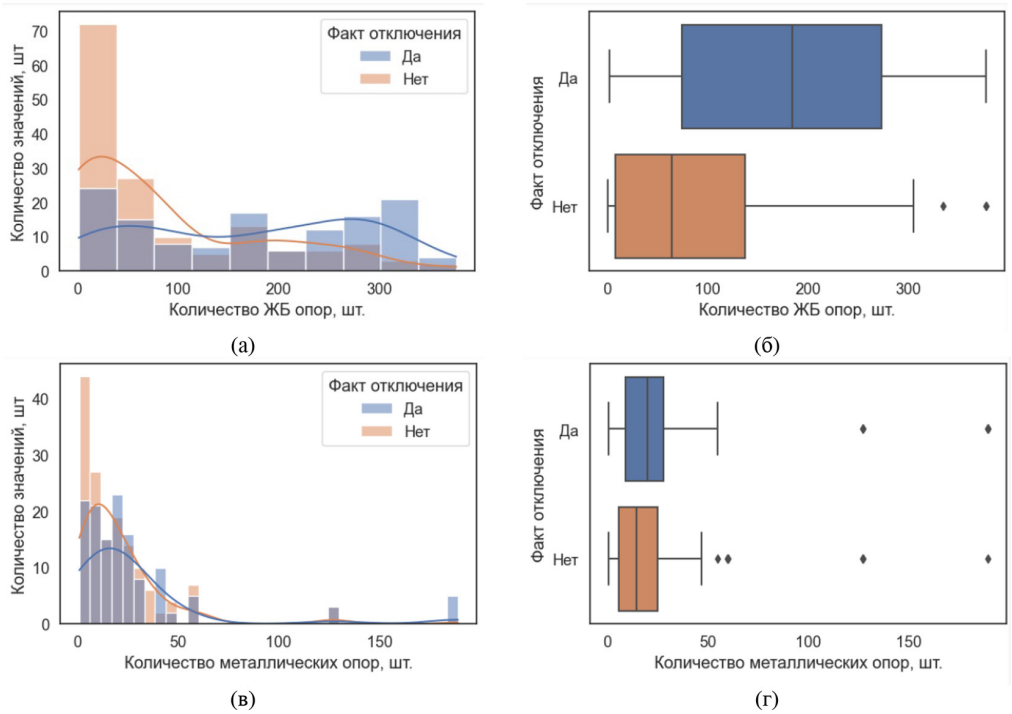


Рис. 4. Распределение значений параметров «Количество ЖБ опор, шт.», «Количество металлических опор, шт.» в разрезе целевой переменной: (а, в) графики плотности распределения, (б, г) диаграммы размаха

Fig. 4. Value distribution of parameters “Number of reinforced concrete supports, pcs.”, “Number of metal supports, pcs.” in terms of the target variable: (а, в) Distribution density graphs, (б, г) Boxplots

наблюдается смещение количества металлических опор в сторону уменьшения. Имеются выбросы в количестве 8 и 10 шт. для ЛЭП с отказами и без соответственно.

Рассмотрим параметры на наличие выбросов, что позволит определить вероятность их появления в зависимости от целевой переменной (см. табл. 1).

Согласно таблице, если на ЛЭП фиксировались отказы, то выбросы наблюдаются в незначительном количестве только для трех параметров: протяженность ЛЭП по населенной местности, количество металлических опор и переэксплуатация. В то же время, рассматривая ЛЭП, на которых фиксировались отказы, видно, что практически для всех их параметров были зафиксированы выбросы в достаточно большом количестве, кроме индекса состояния, что связано с тем, что этот параметр является своего рода искусственным параметром, который сотрудники сетевой компании предпочитают держать в определенных пределах (не менее 60 %) на практике или, по крайней мере, в отчетных документах. Таким образом, можно сделать вывод, что если факт отключения не фиксировался, то такие ЛЭП будут иметь меньший размер статистически значимой выборки и значения будут стремиться быть в границах интерквартильного расстояния. Проанализировав выбросы, было принято решение не удалять выбросы, так как они не являются ошибками ввода или случайными событиями, а линии электропередачи с такими параметрами действительно существуют, поэтому они должны быть включены в дальнейшее исследование.

Таблица 1. Количество выбросов по параметрам ЛЭП в разрезе целевой переменной

Table 1. Number of outliers by power line parameters in the context of the target variable

Признак	Количество выбросов	
	для ЛЭП с отказами	для ЛЭП без отказов
Индекс состояния, %	0	0
Протяженность воздушных участков, км	0	10
Протяженность по лесу, км	0	16
Протяженность по населенной местности, км	6	28
Количество ЖБ опор, шт. для ЛЭП	0	25
Количество металлических опор, шт.	8	10
Переэксплуатация, бр. для ЛЭП	1	19

4.3. Корреляционный анализ

Одним из основных этапов разведочного анализа данных является корреляционный анализ, позволяющий определить взаимосвязи между различными параметрами, что в конечном итоге позволяет сделать выводы о степени влияния одних параметров на другие. Коэффициент корреляции позволяет измерить степень линейной зависимости между двумя переменными. Самой распространенной мерой для определения корреляции между переменными является коэффициент корреляции Пирсона, представляющий собой статистику, которая принимает значения от –1 до +1, при этом значение выше нуля сигнализирует о наличии положительной линейной связи между рассматриваемыми переменными, значение ниже нуля – о наличии отрицательной линейной связи, значение 0 – об отсутствии какой-либо взаимосвязи [26]. Однако коэффициент Пирсона по своей конструкции работает только для интервальных переменных, что делает неудобным его использование при работе со смешанными типами данных [27]. Поэтому если имеются категориальные переменные, то рекомендуется использовать другие меры корреляции, например коэффициент корреляции ϕ_k. Корреляция ϕ_k следует единообразному подходу для интервальных, порядковых и категориальных переменных, поскольку ее определение инвариантно относительно порядка значений каждой переменной. По сути, каждая переменная рассматривается так, как если бы ее тип был категориальным. Суть метода основана на подсчете статистики χ^2 (хи-квадрат) Пирсона и ее приведении к значениям на отрезке 0…1, где 1 сигнализирует о максимальной корреляции между переменными, 0 – об её отсутствии [27].

Так как в рассматриваемых параметрах ЛЭП имеются категориальные значения, то воспользуемся коэффициентом корреляции ϕ_k. Однако, чтобы установить направления связи (положительные или отрицательные) между количественными значениями, воспользуемся расчётом коэффициента Пирсона. Отобразим все значения корреляции на тепловых картах, графических представлениях матрицы данных, где цветовая шкала показывает степень взаимосвязи между переменными, при этом построим на рис. 5 тепловую карту для матрицы корреляции ϕ_k, а на рис. 6 для матрицы корреляции Пирсона.

Из полученных тепловых карт можно сделать несколько выводов о взаимосвязи параметров:

1. Ни один из параметров ЛЭП не имеет высокую корреляцию с целевой переменной, то есть прямое построение линейной регрессии между двумя параметрами не представляется возможным. Наименьшее влияние на факт отказа ЛЭП оказывают типы использованных проводов, наибольшее влияние – протяженность как всех воздушных участков ЛЭП, так и участков, проходящих по лесу.

2. Небольшое влияние на целевую переменную имеет индекс состояния ЛЭП, что кажется странным, так как если логически предположить, что чем хуже состояние линии, тем больше шанс появления повреждений оборудования на ней, то есть больше вероятность отказа линии. Это может быть объяснено искусственным завышением данного показателя в отчетных документах. Также стоит отметить, что индекс состояния имеет высокую корреляцию с параметром «переэксплуатации», причем имеет отрицательную связь.

3. Имеется высокая положительная корреляция между всеми параметрами, отображающими протяжённость ЛЭП, а также с количеством ЖБ опор. Это означает, что между рассмотренными независимыми переменными имеется мультиколлинеарность, которая, скорее всего, приведет к сложности и переобучаемости моделей машинного обучения, поэтому на дальней-

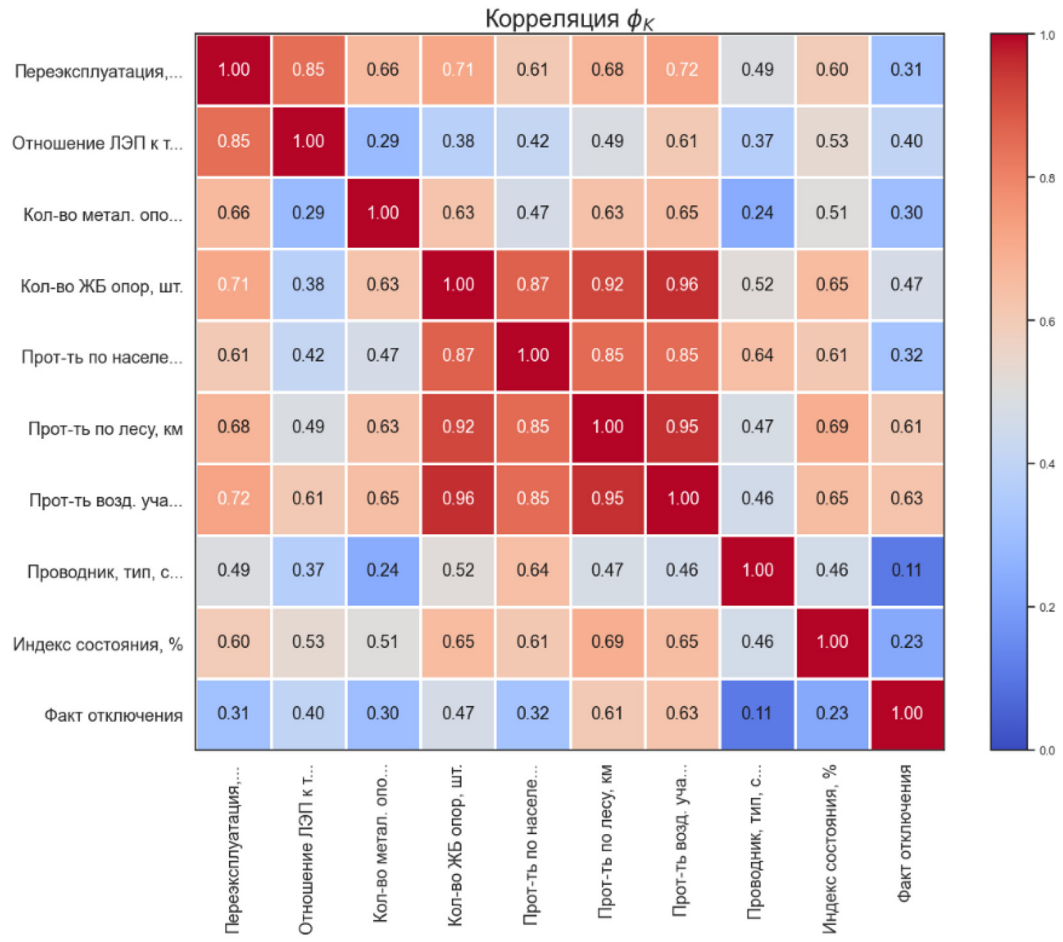


Рис. 5. Тепловая карта матрицы корреляции ϕk между всеми параметрами ЛЭП

Fig. 5. Heat map of ϕk correlation matrix between all power line parameters

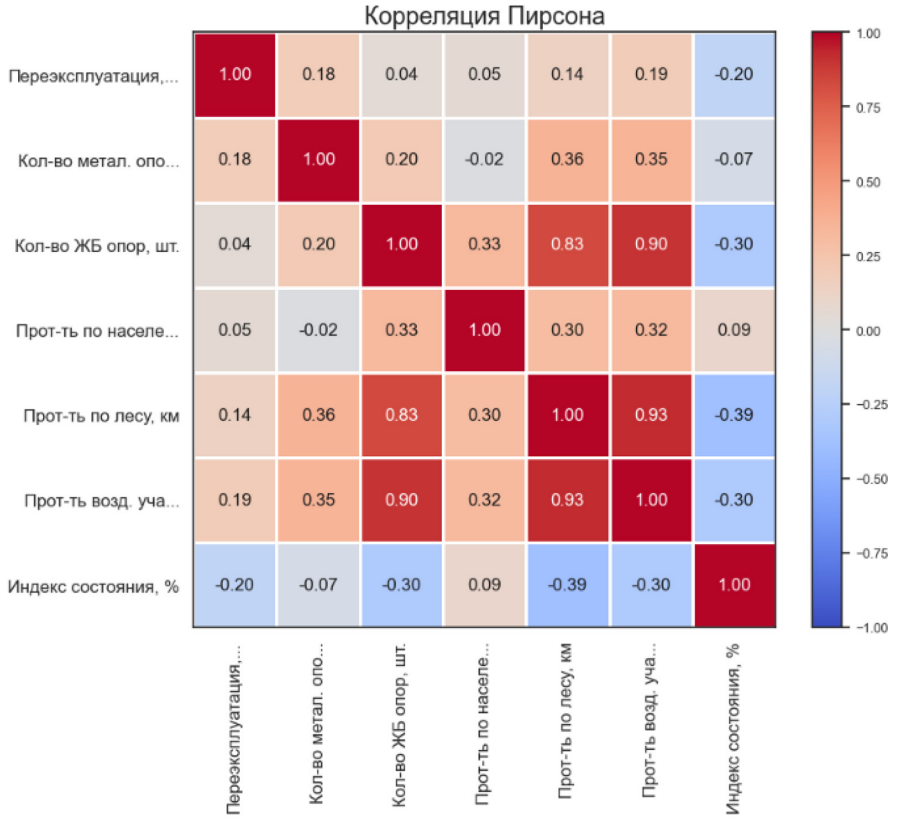


Fig. 6. Heat map of Pearson correlation matrix between power line parameters with quantitative values

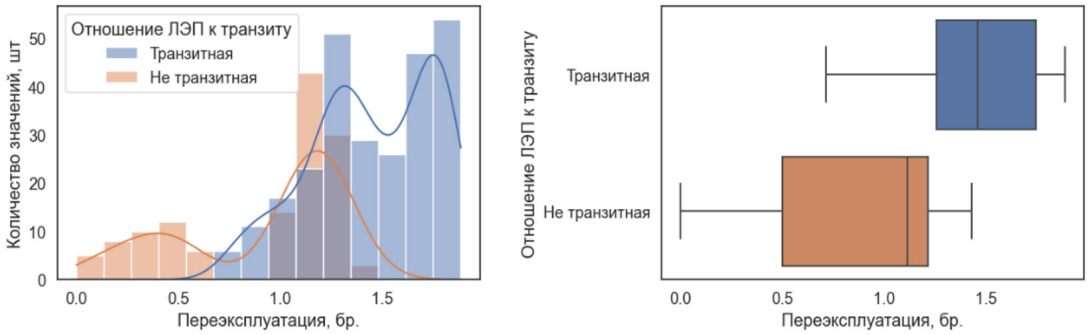
Рис. 6. Тепловая карта матрицы корреляции Пирсона между параметрами ЛЭП с количественными значениями

ших этапах следует исключить переменные, оставив только один, наиболее сильно коррелирующий с целевым признаком.

4. Выявлена достаточно сильная степень корреляции между сроком эксплуатации линии электропередачи и фактом, является ли ЛЭП транзитной или нет, которая равна 0,85 по степени корреляции ϕ_k. На рис. 7 представлены график плотности распределения значений и диаграмма размаха для параметра «Переэксплуатация, бр» в разрезе параметра «Отношение ЛЭП к транзиту», из которых видно, что статистически значимая выборка переэксплуатации не транзитных ЛЭП смещена влево по отношению к транзитным ЛЭП, что говорит о том, что срок эксплуатации транзитных линий больше, чем у не транзитных. Таким образом, можно сделать вывод, что при вводе в эксплуатацию ЛЭП она с большей вероятностью будет являться не транзитной, но со временем при увеличении количества потребителей к ней подключат дополнительные линии и её статус изменится на «транзитную».

5. Рекомендации и будущие исследования

В данном исследовании были подготовлены и обработаны данные по доступным параметрам ЛЭП, которые будут использованы в рамках следующего этапа изучения по построению



(а) (б)

Рис. 7. Распределение значений параметра «Переэксплуатация, бр.» в разрезе параметра «Отношение ЛЭП к транзиту»: (а) график плотности распределения, (б) диаграмма размаха

Fig. 7. Value distribution of parameters “Overexploitation, br.” in the context of the parameter “Relationship of power lines to transit”: (a) Distribution density graph, (б) Boxplots

модели машинного обучения, способной с определенной точностью идентифицировать линии электропередачи с наибольшей вероятностью отказов и рассчитать в количественном значении степень влияния параметров ЛЭП на вероятность отказа.

Для более точного предсказания моделей машинного обучения предлагается изменить параметры с высокой мультиколлинеарностью. Поэтому необходимо будет заменить абсолютные значения количества разных видов опор (железобетонные, металлические) на относительные (выражая в % от общего количества опор). При этом нужно учитывать, что для избежания мультиколлинеарности между ЖБ и металлическими опорами нужно оставить только один их тип, например, железобетонные как наиболее часто встречаемый параметр. То же самое необходимо будет провести с параметрами протяженности по лесу и по населенной местности по отношению к общей протяженности воздушных участков.

Несмотря на то, что выявлена сильная корреляция между сроком эксплуатации линии электропередачи и фактом, является ли ЛЭП транзитной или нет, удаление какого-либо из них не представляется правильным решением, так как у этих параметров разная природа.

6. Выводы

В рамках данного исследования были обработаны данные по отключениям электрической энергии и данные по параметрам ЛЭП напряжением 110 кВ. С помощью методов группировки и слияния табличные данные были подготовлены для анализа. На этом этапе также были удалены параметры, не имеющие отношения к анализу, параметр «Срок эксплуатации ЛЭП» стандартизирован путем преобразования в параметр «Переэксплуатация», синтезирован целевой признак – «Факт отключения на ЛЭП». Итоговым результатом стала таблица, содержащая 10 параметров, включая целевой признак, и 395 строк.

В ходе анализа установлено, что в целевом признаке наблюдается несбалансированность классов – количество фактов отключения меньше на 20 % фактов их отсутствия, что необходимо учитывать при дальнейшем анализе. Выявлено увеличение вероятности отключения электроэнергии при использовании типа провода АС сечением 120 мм^2 (до 45 %) и наличии

факта транзитной линии. Анализ распределения значений индекса состояния ЛЭП подтвердил, что снижение вероятности отключений наблюдается с улучшением технического состояния ЛЭП, с уменьшением длины линий, уменьшением количества опор. Сильное снижение вероятности отказов наблюдается, если значение переэксплуатации становится меньше 0,6, то есть срок эксплуатации ЛЭП менее 20 лет.

Рассматривая наличие выбросов по методу интерквартильных расстояний, сделан вывод о их взаимосвязи с фактом отказа линии. Так, если факт отключения не фиксировался, то такие ЛЭП будут иметь меньший размер статистически значимой выборки и значения будут стремиться быть в границах интерквартильного расстояния, поэтому у них наблюдается большее количество выбросов.

Корреляционный анализ проводился по двум коэффициентам Пирсона и ϕ_k, что позволило установить силу и направления связи между значениями параметров. Было выявлено отсутствие сильной корреляции между целевой переменной и другими параметрами, а также наличие мультиколлинеарности между всеми параметрами, отображающими протяжённость ЛЭП и количество ЖБ опор.

Список литературы / References

[1] Long L. Research on status information monitoring of power equipment based on Internet of Things*, Energy Reports*, 2022, 8, 281–286. DOI: 10.1016/j.egyr.2022.01.018

[2] Sun B., Jing R., Zeng Y., Li Y., Chen J., Liang G. Distributed optimal dispatching method for smart distribution network considering effective interaction of source-network-load-storage flexible resources, *Energy Reports*, 2023, 9, 148–162. DOI: 10.1016/j.egyr.2022.11.178

[3] Yang L., Teh J. Review on vulnerability analysis of power distribution network, *Electric Power Systems Research*, 2023, 224, 109741. DOI: 10.1016/j.epsr.2023.109741

[4] Shakiba F.M., Shojaee M., Azizi S.M., Zhou M. Real-Time Sensing and Fault Diagnosis for Transmission Lines, *International Journal of Network Dynamics and Intelligence*. 2022, 36–47. DOI: 10.53941/ijndi0101004

[5] 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, 114–116. [Bazan T.V., Galaburda Y.V., Iselenok E.B. Analysis of outages of 35–750 kV overhead lines. Aktual'nyye problemy energetiki. Elektroenergeticheskiye sistemy. Minsk, Republic of Belarus: BNTU, 2020, 114–116. (In Rus.)]

[6] Ильдиряков С.Р., Вафин Ш.И. Статистический анализ провалов напряжения в системе электроснабжения ОАО» Казаньоргсинтез», *Известия высших учебных заведений*, 2011, 3–4, 73–81. EDN: NUZVNR [Ildiryakov S.R., Vafin Sh.I. Statistical analysis of voltage dips in the power supply system of OJSC Kazanorgsintez, *Izvestiya vysshikh uchebnykh zavedeniy*, 2011, 3–4, 73–81. EDN: NUZVNR (In Rus.)]

[7] Виноградов А.В., Васильев А.Н., Семенов А.Е. Синяков А.Н., Большев В.Е..Анализ времени перерывов в электроснабжении сельских потребителей и методы его сокращения за счет мониторинга технического состояния линий электропередачи, *Вестник ВИЭСХ*, 2017, 2, 3–11. EDN: ZEOGKZ [Vinogradov A.V., Vasilyev A.N., Semenov A.E. Sinyakov A.N., Bolshev V.E. Analysis of the time of interruptions in power supply to rural consumers and methods

for reducing it by monitoring the technical condition of power lines, *Vestnik VIESKh*, 2017, 2, 3–11. EDN: ZEOGKZ (In Rus.)]

[8] Ланин А.В., Полковская М.Н., Якупов А.А. Статистический анализ аварийных отключений в электрических сетях 10 кВ, *Актуальные вопросы аграрной науки*, 2019, 30, 45–52. EDN: ZBHPXF [Lanin A.V., Polkovskaya M.N., Yakupov A.A. Statistical analysis of emergency shutdowns in 10 kV electrical networks, *Aktual'nyye voprosy agrarnoy nauki*, 2019, 30, 45–52. EDN: ZBHPXF (In Rus.)]

[9] 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, 12. EDN: ZZKEMP [Ratushnyak V.S., Ilyin E.S., Vakhrusheva O. Yu. Statistical analysis of emergency power outages due to ice formation on power transmission lines in the Russian Federation, *Molodaya nauka Sibiri: elektron. nauch. zhurn*, 2018, 1, 12. EDN: ZZKEMP (In Rus.)]

[10] Сбитнев Е.А., Жужин М.С. Анализ аварийности сельских электрических сетей 0, 38 кВ Нижегородской энергосистемы, *Вестник НГИЭИ*, 2020, 11(114), 36–47. EDN: KEJHDX [Sbitnev E.A., Zhuzhin M.S. Analysis of accident rates of rural electrical networks 0.38 kV of the Nizhny Novgorod energy system, *Bulletin of NGIEI*, 2020, 11(114), 36–47. EDN: KEJHDX (In Rus.)]

[11] Sood S. Power Outage Prediction Using Machine Learning Technique, *2023 International Conference on Power Energy, Environment & Intelligent Control (PEEIC)*. IEEE, 2023, 78–80. DOI: 10.1109/PEEIC 59336.2023.10451753

[12] Eskandarpour R., Khodaei A. Leveraging accuracy-uncertainty tradeoff in SVM to achieve highly accurate outage predictions, *IEEE Transactions on Power Systems*, 2018, 33(1), 1139–1141. DOI: 10.1109/TPWRS.2017.2759061

[13] Gururajapathy S.S., Mokhlis H., Illias H.A., Abu Bakar A.H., Awalin L.J. Fault location in an unbalanced distribution system using support vector classification and regression analysis, *IEEJ Transactions on Electrical and Electronic Engineering*, 2018, 13(2), 237–245. DOI: 10.1002/tee.22519

[14] Warlyani P., Jain A., Thoke A.S., Patel R.N. Fault classification and faulty section identification in teed transmission circuits using ANN, *International Journal of Computer and Electrical Engineering*, 2011, 3(6), 807–811. DOI: 10.7763/IJCEE.2011.V3.424

[15] Hou H., Zhang Z., Yu S., Huang Y., Zhang Y., Dong Z. Damage prediction of transmission lines under typhoon disasters considering multi-effect, *J Smart Environ Green Computing*. 2021, 2, 90–102. DOI: 10.20517/jsegc.2020.04

[16] Alqudah M., Obradovic Z. Enhancing Weather-Related Outage Prediction and Precursor Discovery Through Attention-Based Multi-Level Modeling, I*EEE Access*, 2023, 11, 94840–94851. DOI: 10.1109/ACCESS.2023.3303110

[17] Виноградова А.В., Лансберг А.А., Виноградов А.В. *Энергосистема Орловской области: обзор статистической информации*. Орел: изд-во «Картуш», 2023. 360. [Vinogradova A.V., Lansberg A.A., Vinogradov A.V. *Energy system of the Oryol region: review of statistical information*. Orel, Russia: publishing house "Kartush", 2023. 360. (In Rus.)]

[18] Виноградов А.В., Лансберг А.А., Сорокин Н.С. Характеристика электросетевых компаний по количеству и протяженности линий электропередачи, мощности подстанций, *Электротехнологии и электрооборудование в АПК*. 2022, 2 (47), 31–41. DOI: 10.22314/2658–

4859–2022–69–2–31–41 [Vinogradov A. V., Lansberg A. A., Sorokin N. S. Characteristics of electric grid companies by the number and length of power transmission lines, substation capacity, *Electrical technologies and electrical equipment in the agro-industrial complex*. 2022, 2(47), 31–41. DOI: 10.22314/2658–4859–2022–69–2–31–41 (In Rus.)]

[19] Peng J., Wu W., Lockhart B., Bian S., Yan J. N., Xu L., Chi Z., Rzeszotarski J. M., Wang J. Dataprep. EDA: Task-centric exploratory data analysis for statistical modeling in Python, *Proceedings of the 2021 International Conference on Management of Data*. 2021, 2271–2280. DOI: 10.1145/3448016.3457330

[20] Singh J., Singh J., Singh G., Kaur N. Exploratory Data Analysis for Interpreting Model Prediction using Python, *International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON 2022),* Bangalore, India, 2022, 1–6. DOI: 10.1109/ SMARTGENCON 56628.2022.10083533

[21] Guha Majumder M., Dutta Gupta S., Paul J. Perceived usefulness of online customer reviews: A review mining approach using machine learning & exploratory data analysis, *J Bus Res*, 2022, 150, 147–164. DOI: 10.1016/j.jbusres.2022.06.012

[22] Nongthombam K., Sharma D. Data analysis using python, *International Journal of Engineering Research & Technology (IJERT).* 2021, 10, 7. C. IJERTV10IS 070241. DOI: 10.17577/ IJERTV10IS 070241

[23] Сальникова К. В. Анализ массива данных с помощью инструмента визуализации «ящик с усами», *Universum: экономика и юриспруденция*, 2021, 6(81), 11–17. EDN: APSOIG [Salnikova K. V. Analysis of a data array using a “box with a mustache” visualization tool, *Universum: ekonomika i yurisprudentsiya*, 2021, 6(81), 11–17. EDN: APSOIG (In Rus.)]

[24] Лапко А.В., Лапко В. А. Анализ отношения средних квадратических отклонений ядерной оценки плотности вероятности в условиях независимых и зависимых случайных величин, *Измерительная техника*, 2021, 3, 9–14. EDN: NGLNOE [Lapko A. V., Lapko V. A. Analysis of the ratio of standard deviations of the kernel probability density estimate in conditions of independent and dependent random variables, *Izmeritel’naya tekhnika*, 2021, 3, 9–14. EDN: NGLNOE (In Rus.)]

[25] Ушаков В.Г., Ушаков Н. Г. Об одной ядерной оценке плотности, *Информатика и её применения*, 2011, 5(3), 67–73. EDN: OEDYBN [Ushakov V. G., Ushakov N. G. On one kernel density estimate, *Informatika i yeyo primeneniya*, 2011, 5(3), 67–73. EDN: OEDYBN(In Rus.)]

[26] Buda A., Jarynowski A. *Life time of correlations and its applications*. Warsaw, Poland: Wydawnictwo NiezaleĹŀne, 2010. 231.

[27] Baak M., Koopman R., Snoek H., Klous S. A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics, *Comput Stat Data Anal. North-Holland*, 2020, 152, 107043. DOI: 10.1016/j.csda.2020.107043