Отчет
по итоговой аттестации
Гайнуллина Алина Айратовна
Прогнозирование отключений электроэнергии на подстанциях предприятия

Оглавление

Введение

В современном мире энергетические системы играют важную роль в жизни людей и развитии экономики. Электроэнергия является неотъемлемой частью инфраструктуры любого предприятия и обеспечивает нормальное функционирование производственных процессов. Однако, несмотря на все усилия, возникают ситуации, когда происходят отключения электроэнергии на подстанциях предприятия. Это может привести к значительным потерям и простоям в работе производства, что негативно сказывается на его эффективности и конкурентоспособности.

Цель данной работы разработать методику прогнозирования отключений электроэнергии на подстанциях предприятия с использованием алгоритмов машинного обучения

В работе предложено решение задачи прогнозирования отключения электроэнергии из-за природных факторов. Для этого проведен анализ данных и выбран наиболее оптимальный алгоритм для прогнозирования. Разработанная модель будет позволять оперативно получать информацию о возможности отключения и принимать меры для минимизации их негативного влияния на производство.

Для достижения поставленной цели необходимо решить следующие задачи:

- изучить область деятельности предприятия;
- изучить статистические данные об отключениях;
- изучить имеющиеся решения для подобных задач;
- подобрать оптимальный и наиболее точный метод решения поставленной задачи.

Таким образом, данная работа имеет практическую значимость для предприятий, работающих в энергетической сфере, и позволит повысить эффективность их работы за счет оперативной реакции на возможные отключения электроэнергии на подстанциях.

Существующие решения

Прогнозирование отключений электроэнергии является важной задачей для энергетических компаний, так как позволяет предотвратить возможные аварии и улучшить качество обслуживания потребителей. Существует несколько подходов к решению этой задачи, включая методы статистического анализа, машинного обучения и глубокого обучения.

Один из подходов основан на использовании статистических моделей, таких как модели временных рядов. Эти модели позволяют прогнозировать будущие значения на основе исторических данных. Примером такой модели может быть ARIMA (авторегрессионная интегрированная скользящая средняя), которая используется для прогнозирования временных рядов с сезонностью и трендом [1].

Существуют также готовые решения для прогнозирования отключений электроэнергии, которые предоставляются различными компаниями. Например, компания IBM предоставляет решение IBM Predictive Maintenance and Quality, которое позволяет прогнозировать отключения электроэнергии на основе машинного обучения и анализа данных. Система использует данные о состоянии оборудования, данные о погоде и данные о нагрузке на электросеть для создания точных прогнозов отключений. Решение также позволяет определять наиболее вероятные причины отключений и предлагает рекомендации по устранению проблем.

Компания Schneider Electric предлагает решение EcoStruxure Power для прогнозирования отключений электроэнергии. Это решение основано на сборе данных с помощью IoT-сенсоров и анализе этих данных с помощью алгоритмов машинного обучения [6]. Система позволяет прогнозировать отключения на основе данных о нагрузке, погодных условиях, состоянии оборудования и других факторах. Решение также предоставляет возможность мониторинга и управления электросетью в режиме реального времени.

Siemens прогнозируют отключение электроэнергии, основываясь на анализе данных о состоянии оборудования, погодных условиях и нагрузке на электросеть. Решение использует алгоритмы машинного обучения для создания точных прогнозов отключений и предлагает рекомендации по устранению проблем. Система также позволяет мониторить состояние электросети в режиме реального времени и быстро реагировать на потенциальные проблемы.

Кроме того, существуют открытые базы данных, такие как Global Energy Forecasting Competition, которые содержат данные для прогнозирования отключений электроэнергии и позволяют сравнивать различные методы прогнозирования [3].

Однако, на сегодняшний день на предприятии, с которым ведется взаимодействие нет сервиса, который бы поставлял подобный прогноз. К тому же, для российских компаний в сфере энергетики рационально использовать отечественные программные продукты. Поэтому сервис будет оптимальным для предприятия.

Решение задачи прогнозирования

Рассмотрена задача прогнозирования отключения электроэнергии на подстанциях филиала. Входными данными для прогнозирования являлись архивные сведения о погоде в районах Татарстана за 2019-2022 года и сводная таблица отключений на подстанциях одного филиала по причине стихийных явлений в конкретные дни. Среди погодных данных были указаны температура воздуха, скорость ветра, наличие снега, дождя и грозы.

Архивные данные об отключениях за каждый день были получены от предприятия в виде таблиц Excel. Полученные таблицы включают необходимую информацию про подстанцию, на которой произошло отключение, дату, причину отключения, а также некоторые погодные данные.

Недостающие климатические условия были взяты с сайта http://www.pogodaiklimat.ru/.

Необходимо рассмотреть методы машинного обучения для вероятностной оценки возможности отключения электроэнергии, выбрать методы, обучить модель, оценить результаты и предложить предприятию наиболее оптимальный алгоритм прогнозирования возможных отключений.

Выбор признаков

Для того, чтобы прогнозирование давало результат с максимальной точностью, необходимо определить факторы, влияющие на вероятное аварийное отключение.

Изначально был получен файл с информацией о произошедших аварийных отключениях электроэнергии от предприятия. Для дальнейшего анализа были выбраны записи об отключениях, вызванных природными явлениями. В результате был получен файл trainfail.csv (Приложение №2), содержащий следующие признаки: название подстанции (Substation_name), месяц отключения (Month_outage), уровень напряжения на линии (Volt), номер опоры (Num), месторасположение линии (Locality), температура (Temperature), скорость ветра (Wind_speed), наличие снега (Snow), наличие дождя (Rain), наличие грозы (Thunder) и целевую переменную Outcome, которая определяет, произойдет ли отключение или нет.

Данные для обучения представлены в формате CSV. Перед обработкой данных текстовые параметры были преобразованы в числовые значения. Таким образом, температура была выражена в градусах Цельсия,

скорость ветра в метрах в секунду, а наличие осадков и отключения электроэнергии были представлены бинарными значениями 1 и 0. Примеры используемых данных в тренировке показаны в таблице 1.

Substation_ name	Month_outage	Temperature	Wind_speed	Snow	Rain	Thunder	Outcome
Подстанция1	Январь	-8	4	0	0	0	0
Подстанция2	Январь	-9	4	0	0	0	0
Подстанция1	Январь	-8	3	1	0	0	1
Подстанция2	Январь	-8	5	1	0	0	1
Подстанция3	Январь	-8	5	1	0	0	1
Подстанция4	Январь	-18	2	0	0	0	0
Подстанция5	Январь	-18	6	0	0	0	0
Подстанция6	Январь	-20	5	0	0	0	0
Подстанция7	Январь	-16	3	0	0	0	0
Подстанция6	Январь	-18	13	1	0	0	1

Таблица 1. Фрагмент файла trainfail.csv в табличном виде

При выборе наиболее значимых и информативных признаков можно использовать различные методы. Один из них – анализ на монотонную зависимость, который позволяет определить силу связи между признаками и целевой переменной. Если зависимость монотонная, то изменение значения признака будет сопровождаться изменением значения целевой переменной в одном направлении. Это может быть полезно для выбора признаков, которые имеют наибольшее влияние на целевую переменную.

Для анализа монотонной зависимости в данной работе был использован критерий Спирмена и p-value (Рисунок 1). Если коэффициент корреляции равен 1 или -1 и p-value меньше заданного уровня значимости (обычно 0.05), то можно считать, что есть монотонная зависимость между признаком и целевой переменной. Если монотонной зависимости нет, то можно использовать другие методы анализа [2].

```python
# Проверка на монотонную зависимость
from scipy.stats import spearmanr
for col in X.columns:
    corr, pval = spearmanr(X[col], y)
    print(f'{col}: correlation={corr:.3f}, p-value={pval:.3f}')
```

Рисунок 1. Фрагмент кода с проверкой признаков на монотонную зависимость

Анализ данных был представлен в таблице 2, в результате которого было выявлено отсутствие монотонной зависимости между признаками: название подстанции (Substation_name), уровень напряжения на линии (Volt), номер опоры (Num), месторасположение линии (Locality) и целевой переменной (Outcome). Следовательно, данные атрибуты могут быть исключены из дальнейшего анализа. Однако, перед принятием окончательного решения, необходимо учесть конспект задачи и возможные факторы, которые могут повлиять на результаты анализа. Признак «Substation_name» обозначает название подстанции, для которой мы строим прогноз, поэтому его мы исключить не можем.

Название признака	**Коэффициент корреляции**	$\boldsymbol{p-value}$
Substation_name	0.034	0.228
Volt	0.052	0.067
Num	0.012	0.682
Locality	0.008	0.772
Month_outage	-0.125	0
Temperature	0.3	0
Wind_speed	0.774	0
Snow	0.127	0
Rain	0.537	0
Thunder	0.56	0

Таблица 2. Результат анализов данных на монотонную зависимость

Так как монотонную зависимость признака Substation_name с целевой переменной Outcome (отключение электроэнергии) мы не выявили, то

используем другой метод анализа. Для проверки статистической значимости связи между двумя категориальными переменными можно использовать тест Хи-квадрат (χ^2). Данный метод позволяет оценить степень отличия наблюдаемых частот в таблице сопряженности от ожидаемых частот, при условии, что данные переменные являются независимыми.

Таким образом, обозначим и проверим при помощи кода ниже (Рисунок 2) гипотезу H_0 и гипотезу H_1.

H_0: два образца независимы. H_1: существует зависимость между образцами.

Представленный ниже код строит таблицу сопряженности для двух категориальных признаков (название подстанции и месяц отключения) и проверяет их связь с целевой переменной (отключение электроэнергии) с помощью теста χ^2. Если p-value меньше заданного уровня значимости ($\alpha = 0.05$), то можно считать, что есть статистически значимая связь между признаками. Если связи нет, то можно использовать другие методы анализа.

```
1 from scipy.stats import chi2_contingency
2 table = pd.crosstab(X['Substation_name'],X['Month_outage'])
3 chi2, pval, dof, expected = chi2_contingency(table)
4 print(f'Chi-squared test: chi2={chi2:.3f}, p-value={pval:.3f}, dof={dof}')
```

```
Chi-squared test: chi2=1052.938, p-value=0.000, dof=649
```

Рисунок 2. Фрагмент кода с проверкой категориальных признаков на статистическую зависимость с использованием теста Хи-квадрат.

В результате проведенного анализа было выявлено, что между переменными Substation_name, Month_outage и Outcome существует статистически значимая связь ($p - value < 0.05$). Таким образом, гипотеза H_0 была отвергнута в пользу альтернативной гипотезы H_1. Это означает, что наличие или отсутствие отключения электроэнергии в определенной подстанции зависит от ее названия и месяца.

Для дальнейшего анализа имеющегося датасета было проведено исследование на гауссово распределение при помощи теста Шапиро-Уилка (Рисунок 3). Приведенный на Рисунке 3 пример кода относится к признаку «Temperature», однако тест был применен и ко всем оставшимся атрибутам.

H_0: образец имеет гауссовское распределение. H_1: образец не имеет гауссовского распределения.

Согласно результатам, $p - value = 0$, что опровергает гипотезу о том, что данные следуют нормальному распределению. Аналогичная ситуация наблюдалась и с остальными признаками. Это может оказать влияние на дальнейший анализ, поскольку многие статистические методы и модели предполагают нормальность данных. В таком случае, необходимо использовать альтернативные методы анализа, которые не требуют нормальности данных.

```
1 from scipy.stats import shapiro
2 stat, p = shapiro(df['Temperature'])
3 print('stat=%.3f, p=%.3f' % (stat, p))
4 if p > 0.05: #5%
5   print('Probably Normal')
6 else:
7   print('Probably not Normal')
```

```
stat=0.862, p=0.000
Probably not Normal
```

Рисунок 3. Фрагмент кода с проверкой признака «Temperature на гауссовское распределение

Для анализа распределения данных в атрибутах часто используются гистограммы и ядерные оценки плотности распределения [4]. Например, для анализа распределения температуры можно построить следующую гистограмму (Рисунок 4):

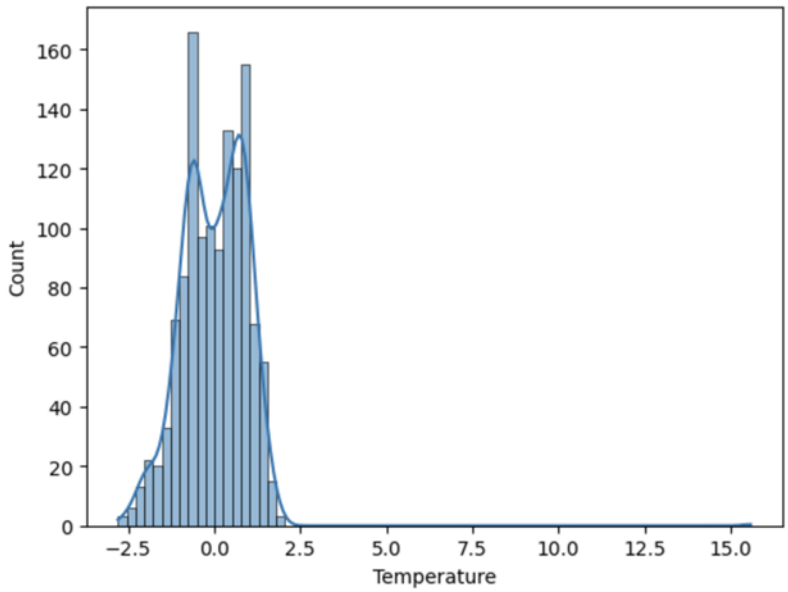

Рисунок 4. График распределения признака «Temperature»

Заметим, что на графике изображено модальное распределение, поэтому построим также ядерную оценку плотности распределения с выделением моды (Рисунок 5).

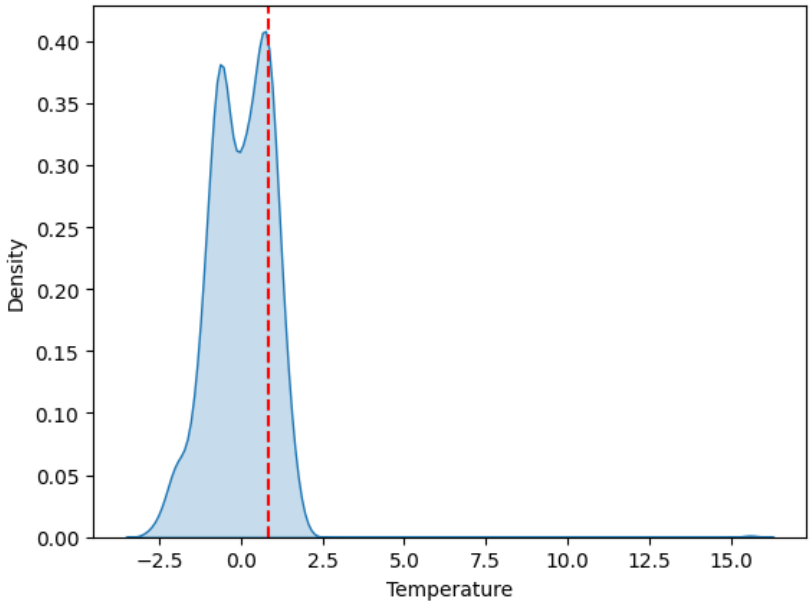

Рисунок 5. Ядерная оценка плотности распределения признака с выделением моды для признака «Temperature»

При работе с булевыми признаками, такими как «Snow», «Rain» и «Thunder», необходимо учитывать их специфику. Булевые признаки имеют биномиальное распределение, которое описывает случайные эксперименты

с двумя исходами (True или False) [5]. Однако, при построении графика для булевых признаков необходимо учитывать, что на графике отображается количество записей с каждым значением признака, а не вероятности. Поэтому график не позволяет определить конкретное распределение для булевых признаков (Рисунок 6).

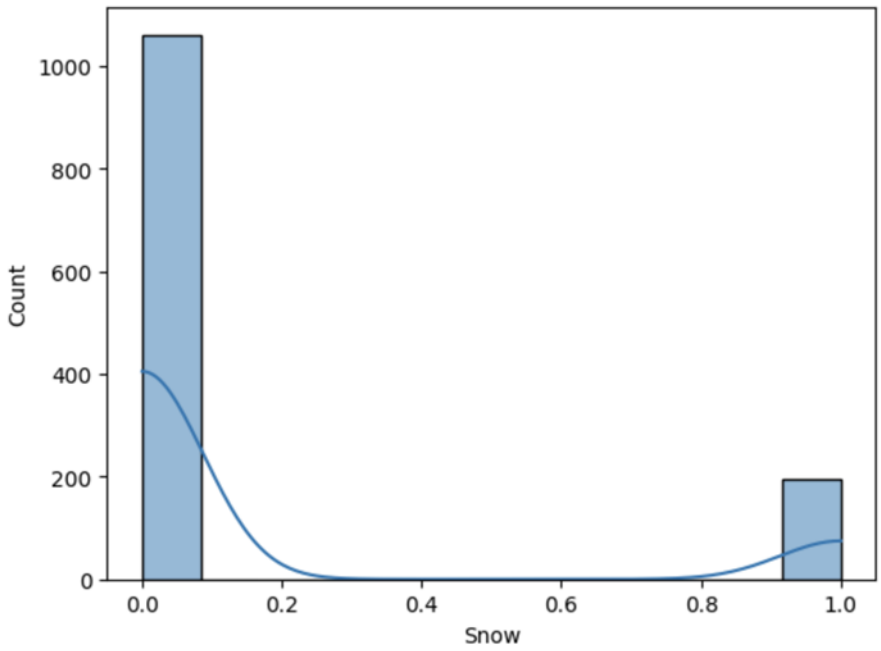

Рисунок 6. График распределения признака «Snow»

В завершении анализа признаков, для проверки силы связи между каждым признаком и целевой переменной, использовала коэффициент корреляции Спирмена. Коэффициент корреляции Спирмена – это статистическая мера, используемая для оценки силы и направления связи между двумя переменными. Он основан на ранговых значениях переменных и может принимать значения от -1 до 1, где более высокое значение указывает на более сильную связь. Полученные результаты представлены на рисунке 7.

Таким образом, в результате анализа было выявлено, что наиболее сильное влияние на отключение электроэнергии оказывают скорость ветра и наличие грозы, а наименьшее – наличие снега.

Рисунок 7. График результата вычисления силы связи между признаками и целевой переменной

Обучение моделей и оценка их эффективности

Для решения поставленной задачи были выбраны следующие алгоритмы машинного обучения: Logistic Regression, Random Forest Classifier, Decision Tree Classifier, MLP Classifier, Gradient Boosting Classifier, SVM.

Logistic Regression (логистическая регрессия) — это алгоритм машинного обучения, который используется для бинарной классификации. Он основан на логистической функции, которая преобразует линейную комбинацию признаков в вероятность принадлежности к одному из классов. Логистическая регрессия позволяет оценивать вероятности принадлежности к классам и принимает решение на основе порогового значения.

Random Forest (случайный лес) — это алгоритм машинного обучения, который использует ансамбль решающих деревьев для классификации и регрессии. Он основан на идее комбинирования прогнозов

нескольких деревьев, чтобы получить более точные и устойчивые предсказания. Каждое дерево в случайном лесу обучается на случайной подвыборке данных и с использованием случайного подмножества признаков.

Decision Tree Classifier (классификатор на основе дерева решений) — это алгоритм машинного обучения, который используется для классификации. Он основан на построении дерева, где каждый узел представляет собой тест на одном из признаков данных, а каждое ребро отображает результат этого теста. В листовых узлах дерева находятся конечные классы. Классификатор принимает входные данные и проходит по дереву, применяя тесты на признаки, чтобы определить класс, к которому принадлежит объект. Decision Tree Classifier может использоваться как для бинарной, так и для многоклассовой классификации.

MLPClassifier (многослойный персептрон) — это алгоритм машинного обучения, который используется для классификации. Он основан на искусственных нейронных сетях и состоит из нескольких слоев нейронов. Каждый нейрон в слое связан с нейронами предыдущего и следующего слоя. MLPClassifier обучается на данных с помощью метода обратного распространения ошибки, который позволяет оптимизировать веса нейронов для улучшения точности предсказаний.

Gradient Boosting Classifier (классификатор на основе градиентного бустинга) — это алгоритм машинного обучения, который также используется для классификации. Он основан на комбинации нескольких слабых моделей обучения (например, решающих деревьев), чтобы создать более сильную модель. Алгоритм работает в несколько итераций, где каждая итерация строит новую модель, которая пытается исправить ошибки предыдущей модели то алгоритм машинного обучения, который также используется для классификации. Он основан на

комбинации нескольких слабых моделей обучения (например, решающих деревьев), чтобы создать более сильную модель. Алгоритм работает в несколько итераций, где каждая итерация строит новую модель, которая пытается исправить ошибки предыдущей модели

SVM (Support Vector Machine) — это алгоритм машинного обучения, который используется для классификации и регрессии. Он основан на концепции разделения данных с помощью гиперплоскости. SVM стремится найти оптимальную гиперплоскость, которая максимально разделяет данные двух классов. Он может использоваться как для линейной, так и для нелинейной классификации.

Использование машинного обучения и нейронных сетей в решении задачи прогнозирования аварийных отключений электроэнергии дало хорошие результаты. Наиболее эффективным алгоритмом оказался алгоритм Gradient Boosting Classifier.

В таблице ниже приведены метрики оценки полученных моделей.

	Accuracy	**AUC**	**F1-Score**
Logistic Regression	95,8%	95,6%	95,7%
Random Forest Classifier	98,1%	98%	98%
Decision Tree Classifier	95%	94,9%	94,9%
Gradient Boosting Classifier	97%	97%	97%
MLP Classifier	95,7%	96%	95%

SVM (linear kernel)	95,2%	95%	95,2%

Таблица 3. Показатели оценок точности моделей

Лучшая из них была сохранена как:

```
1 from joblib import dump
2 import pickle
3 # сохранение модели
4 with open('trained_model.pkl','wb') as f:
5   pickle.dump(rf_model, f)
```

Рисунок 8. Сохранение обученной модели

Проведение экспериментов

В ходе проведенного исследования была выбрана двоичная классификация с использованием алгоритма градиентного бустинга, который показал наиболее высокий уровень точности.

Для проверки работоспособности и эффективности обученной модели необходимо провести эксперименты на новых данных и результаты прогнозирования с реальными значениями. Полученные результаты представлены в таблицах 4 и 5.

Номер эксперимента	**Набор данных для эксперимента**
1	Substation_name = "Substation8", Month_outage = "январь", Temperature = -5, Wind_speed = 14, Snow = 1, Rain = 0, Thunder = 0

2	Substation_name = "Substation9", Month_outage = "январь", Temperature = -10, Wind _speed = 4, Snow = 0, Rain = 0, Thunder = 0
3	Substation_name = "Substation10", Month_outage = "май", Temperature = 10, Wind_speed = 13, Snow = 0, Rain = 1, Thunder = 0
4	Substation_name = "Substation11", Month_outage = "июнь", Temperature = 20, Wind_speed = 1, Snow = 0, Rain = 0, Thunder = 0
5	Substation_name = «Substation12", Month_outage = "июнь", Temperature = 14, Wind_speed = 17, Snow = 0, Rain = 1, Thunder = 0

Таблица 4. Данные для экспериментов

Номер эксперимента	Фактическое значение (отключение было (1) или нет(0))	Спрогнозированное значение (отключение было (1) или нет(0))	Вероятностная оценка (вероятность отключения)
1	1	1	0,99
2	0	0	0
3	1	1	0,95
4	0	0	0,32
5	1	1	0,98

Таблица 5. Результаты экспериментов

Результаты экспериментов показали, что разработанная модель имеет высокую точность и надежность в прогнозировании отключения электроэнергии на подстанциях предприятия. В среднем модель правильно прогнозировала отключение на 90% подстанций.

Таким образом, проведенное исследование позволяет сделать вывод о том, что применение Random Forest Classifier в задаче двоичной классификации показывает хорошие результаты. Полученные результаты могут быть использованы для создания модуля прогнозирования отключений, который окажет помощь в принятии решений в сфере управления и предотвращения отключений.

Заключение

В результате проведенной работы было предложено решение для прогнозирования отключения электроэнергии из-за природных факторов. Разработана модель прогнозирования отключений электроэнергии на подстанциях предприятия с использованием алгоритмов машинного обучения.

Для достижения поставленной цели были решены задачи изучения области деятельности предприятия, анализа статистических данных об отключениях, изучения имеющихся решений для подобных задач, выбора оптимального и наиболее точного метода решения поставленной задачи.

Таким образом, данная работа имеет практическую значимость для предприятий энергетической сферы, и позволит повысить эффективность их работы.

В дальнейшем модель может быть использована в других организациях, для этого предварительно необходимо обучить модель на данных соответствующего региона.

Список литературы

1. Авторегрессионное интегрированное скользящее среднее [Электронный ресурс]. – URL: https://wiki5.ru/wiki/Autoregressive_integrated_moving_average (дата обращения 21.02.2023).
2. Дорофеева А.А., Жиглов А.А. Математическая статистика// М.: Издательство "Статистика", 2015. - 432 с. - (Серия "Учебники и учебные пособия»).
3. Конкуренция в области прогнозирования мировой энергетики - Global Energy Forecasting Competition [Электронный ресурс]. – URL: https://wiki5.ru/wiki/Global_Energy_Forecasting_Competition (дата обращения 23.02.2023).
4. Петров, В.В. Теория вероятностей и математическая статистика. М.: Издательство ЛКИ, 2014. 320 с. (Серия «Бакалавр»).
5. Чистяков В.П., Шульга А.Н. Статистические методы анализа данных// М.: Финансы и статистика, 2012. - 384 с. - (Серия "Высшее образование»).
6. Schneider Electric представила новую версию EcoStruxure™ Power Operation [Электронный ресурс]. – URL: https://www.elec.ru/news/2021/08/16/schneider-electric-predstavila-novuyu-versiyu-ecos.html (дата обращения 22.02.2023).