

Сравнительный анализ методов машинного обучения для решения задачи прогнозирования уровня безработицы в регионах России

И.А. Сироткин^а, Л.А. Геращенко^б

МИРЭА – Российский технологический университет, проспект Вернадского, 78, Москва, Россия

^а sirotkinia@inbox.ru, ^б geracshenko@mirea.ru

^а <http://orcid.org/0009-0000-4828-4653>, ^б <http://orcid.org/0000-0002-1766-2735>

Статья поступила 12.01.2026, принята 09.02.2026

Данная научная статья посвящена решению задачи прогнозирования уровня безработицы в регионах Российской Федерации с применением методов машинного обучения. Актуальность исследования обусловлена значительной пространственной неоднородностью социально-экономического развития регионов России, а также ограниченной эффективностью традиционных эконометрических подходов, таких как регрессионный анализ и модели временных рядов, при работе с нелинейными зависимостями и большими массивами данных. Целью работы является определение наиболее эффективного алгоритма машинного обучения для прогнозирования региональной безработицы. В рамках исследования был проведён сравнительный анализ четырёх популярных методов: линейной регрессии, дерева решений, случайного леса и градиентного бустинга. Информационной базой послужили агрегированные данные Росстата за период с 2000 по 2024 гг. Для моделирования были отобраны ключевые группы признаков, такие как валовой внутренний продукт (ВВП), валовой региональный продукт (ВРП), показатели промышленного производства и деятельности компаний. Практическая часть исследования выполнена на языке программирования Python с использованием библиотек scikit-learn, pandas, numpy и matplotlib. Предварительно была проведена предобработка данных, включающая в себя удаление выбросов, обработку пропусков и формирование обучающих и тестовых выборок. Прогнозируемым показателем выступал уровень безработицы в регионе на следующий год. Для оценки качества моделей использовались метрики RMSE (среднеквадратичная ошибка), MAE (средняя абсолютная ошибка) и R² (коэффициент детерминации). Результаты эксперимента показали, что наилучшую прогностическую способность продемонстрировали ансамблевые методы. Модель градиентного бустинга достигла наименьших ошибок и наибольшего значения R², что свидетельствует о её высокой точности и объясняющей способности. Случайный лес также показал достойные результаты, в то время как более простые модели – дерево решений и линейная регрессия – оказались существенно менее точными из-за неспособности адекватно учитывать сложные нелинейные взаимосвязи между признаками. Таким образом, исследование подтверждает перспективность и эффективность методов машинного обучения, в частности градиентного бустинга, для задачи прогнозирования уровня безработицы на региональном уровне. Полученные результаты могут быть использованы органами государственного управления для повышения точности планирования и разработки более эффективных мер регулирования рынка труда.

Ключевые слова: безработица; анализ данных; машинное обучение; линейная регрессия; дерево решений; ансамблевое обучение.

A comparative analysis of machine learning methods for solving the problem of forecasting unemployment rates in Russian regions

I.A. Sirotkin^а, L.A. Geraschenko^б

MIREA – Russian Technological University; 78, Vernadsky Ave., Moscow, Russia

^а sirotkinia@inbox.ru, ^б geracshenko@mirea.ru

^а <http://orcid.org/0009-0000-4828-4653>, ^б <http://orcid.org/0000-0002-1766-2735>

Received 12.01.2026, accepted 09.02.2026

The research paper examines the problem of forecasting unemployment rates in the regions of the Russian Federation using machine learning methods. The relevance of the study stems from the significant spatial heterogeneity of the socioeconomic development of Russian regions, as well as the limited effectiveness of traditional econometric approaches, such as regression analysis and time series models, when working with nonlinear relationships and large data sets. The aim of the study is to identify the most effective machine learning algorithm for forecasting regional unemployment. The study includes a comparative analysis of four popular methods: linear regression, decision tree, random forest, and gradient boosting. Aggregated Rosstat data for the period from 2000 to 2024 serves as the data base. Key groups of indicators, such as gross domestic product (GDP), gross regional product (GRP), industrial production indicators, and company performance indicators, are selected for modeling. The practical part of the study is conducted in Python using the scikit-learn, pandas, numpy, and matplotlib libraries. Data preprocessing is performed, including removing outliers, handling missing data, and generating training and test samples. The predicted indicator is the unemployment rate in the region for the following year. The metrics RMSE (root mean square error), MAE (mean absolute error), and R² (coefficient of determination) are used to evaluate the quality of the models. The experimental results show that ensemble methods demonstrate the best predictive ability. The gradient boosting model achieves the lowest errors and the highest R² value, demonstrating its high accuracy and explanatory power. The random forest also demonstrates perfect results, while simpler models – decision tree and linear regression – are significantly less accurate due to their inability to adequately account for complex nonlinear relationships between features. Thus, the study confirms the potential and effectiveness of machine learning methods, particularly gradient boosting, for forecasting unemployment rates at the regional level. The results can be used by government agencies to improve planning accuracy and develop more effective labor market regulation measures.

Keywords: unemployment; data analysis; machine learning; linear regression; decision tree; ensemble learning.

Введение. Прогнозирование социально-экономических показателей является одной из ключевых задач в сфере государственного планирования и регионального развития. Среди таких показателей уровень безработицы занимает особое место, выступая не только как индикатор состояния рынка труда, но и как важный фактор социальной стабильности и экономического благополучия. В современных условиях развития экономики России безработица выступает как одна из серьёзных глобальных проблем, оказывающих значительное влияние на базисные аспекты отраслей экономики, снижая экономическую активность домашних хозяйств, фирм и государственного сектора, вызывая проблему резкого падения деловой активности и покупательной способности, что замедляет инвестиционное развитие и ускорение оборачиваемости оборотного капитала для обеспечения расширенного воспроизводственного процесса в реальном секторе экономики [1].

Безработица в России характеризуется существенными различиями в региональном разрезе, что обусловлено территориальной протяжённостью и историческими особенностями своего развития [2]. В условиях высокой пространственной неоднородности регионов России, обусловленной различиями в отраслевой структуре, демографической ситуации и инвестиционной привлекательности регионов, задача прогнозирования региональной безработицы приобретает повышенную актуальность.

Традиционные подходы к прогнозированию, такие как регрессионный анализ и модели временных рядов (ARIMA, GARCH), часто сталкиваются с ограничениями при работе с нелинейными зависимостями и большими объёмами данных. В связи с этим всё более востребованными становятся методы машинного обучения, демонстрирующие высокую эффективность в задачах прогнозирования за счёт способности выявлять нелинейные последовательности и адаптироваться к нестационарным данным.

Несмотря на растущий интерес к применению машинного обучения для решения задач в сфере экономики, вопрос об эффективности различных алгоритмов для прогнозирования уровней безработицы в регионах России остаётся недостаточно изученным.

Актуальность работы заключается в оценке применимости методов машинного обучения для прогнозирования уровня безработицы в регионах России и определении наиболее подходящего метода для решения данной задачи.

Постановка целей и задач. Цель исследования – определение наиболее эффективного метода машинного обучения из рассматриваемых в контексте решения задачи прогнозирования уровня безработицы в регионах России.

Задачи исследования:

- определение перечня факторов, связанных с уровнем безработицы;
- поиск и предобработка набора данных, содержащего вышеупомянутые факторы;
- обучение моделей машинного обучения с применением нескольких методов;
- расчёт и сравнение метрик качества полученных моделей.

Информационная база. В качестве информационной базы исследования выступили данные о различных социально-экономических показателях в регионах России, собранные Росстатом и агрегированные проектом «Если быть точным». Набор данных охватывает период с 2000 по 2024 гг. и содержит значения 1117 социально-экономических показателей. Единица наблюдения в наборе данных – это страна, федеральный округ или регион в конкретный год. Для каждого значения показателя приводятся сведения о единице измерения и методические комментарии, если они содержались в сборнике Росстата [3].

Данные о ВВП России по годам взяты напрямую из Росстата.

Определение перечня данных. Чтобы выполнить анализ данных при помощи методов машинного обучения, следует сначала определить набор признаков, оказывающих наибольшее влияние на целевой показатель (в данном случае – уровень безработицы в регионе на следующий год). Если признаков будет слишком мало, то это приведёт к тому, что модель не сможет в полной мере описать закономерности в данных, а если в наборе данных будут лишние признаки, то они могут ухудшить качество модели, так как они могут стать источником «шума».

В целях определения перечня признаков для итогового набора данных рассмотрены теоретические данные, связанные с безработицей и социально-экономическими показателями, которые могут быть с ней связаны.

Так как безработица является одним из социально-экономических показателей, отражающих экономическую обстановку и благосостояние населения, принято решение обратить внимание на показатели, описывающие схожие социально-экономические аспекты обстановки в стране и регионах.

По мнению многих учёных, в настоящих условиях основным и главным фактором в социально-экономическом развитии страны является человек. В большей степени от эффективного использования трудового потенциала зависит устойчивый экономический рост страны и как следствие увеличение его валового внутреннего продукта (ВВП) [4]. Следовательно, ВВП стоит включить в перечень анализируемых признаков, так как он является одним из важнейших социально-экономических показателей страны.

Валовой региональный продукт (ВРП) является показателем, схожим с ВВП, однако связан не со страной в целом, а с отдельным её регионом. Рассматривая и анализируя региональное развитие территорий, невозможно не принимать во внимание ВРП, так как он идентифицирует уровень регионального производства и связан с ростом региональной эффективности, которая является инструментом повышения качества жизни, благосостояния населения [5]. В связи с этим ВРП так же, как и ВВП, стоит включить в набор данных, подлежащий дальнейшему анализу.

При изучении социально-экономических показателей страны и регионов стоит учитывать показатели, связанные с промышленным производством, так как промышленность является ключевой отраслью экономики, от роста которой зависит экономическое развитие [6].

Стоит отметить, что промышленные показатели с большой вероятностью тесно связаны с уровнем безработицы в регионе, так как недостаток рабочей силы может стать причиной упадка промышленного производства в регионе, а недостаточное развитие промышленности может привести к малому количеству рабочих мест. Из этого следует, что при анализе данных стоит учитывать показатели промышленного производства в регионах.

Показатели деятельности компаний в регионах имеют большую корреляцию с уровнем безработицы, так как любая компания зависит от наличия рабочей силы. Территориальное развитие городов представляет собой комплексный процесс, включающий социальные, экономические, экологические и инфраструктурные аспекты. В сфере территориального развития городов (следовательно, и регионов) в последние годы внимание уделяется роли бизнеса в развитии городов, что проявляется через инвестиции, создание рабочих мест, внедрение технологий и устойчивое развитие [7]. По этим причинам показатели деятельности компаний нельзя оставлять без внимания в рамках данного исследования.

В результате выделены следующие группы признаков:

- ВВП и ВРП;
- показатели промышленного производства;
- показатели деятельности компаний.

Однозначно определить перечень признаков, который приведёт к наилучшим результатам обучения моделей, представляется очень сложной задачей, решение которой может быть найдено лишь эмпирическим путём. Исследуемый набор признаков с высокой вероятностью не является оптимальным, однако его достаточно для проведения сравнительного анализа методов машинного обучения.

Исследование методов машинного обучения для прогнозирования с применением языка программирования Python. В рамках данного сравнительного анализа рассмотрены следующие методы машинного обучения:

- линейная регрессия;
- дерево решений;
- случайный лес;
- градиентный бустинг.

Для оценки качества моделей использованы следующие метрики качества:

- корень средней квадратичной ошибки (Root Mean Square Error, RMSE);
- средняя абсолютная ошибка (Mean Absolute Error, MAE);
- коэффициент детерминации (R2).

Сравнительный анализ выполнен при помощи языка программирования Python с применением следующих библиотек:

- pandas;
- numpy;
- matplotlib;
- sklearn.

Перед этапом обучения моделей и оценки их качества, была проведена загрузка и обработка данных.

Импортируем нужные библиотеки, заранее определяем функцию для удаления выбросов, как показано на рис. 1–2.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import warnings

from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score
from sklearn.metrics import root_mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn import preprocessing
from sklearn.ensemble import RandomForestRegressor
from sklearn.ensemble import GradientBoostingRegressor
from matplotlib.ticker import MaxNLocator

warnings.filterwarnings('ignore')
```

Рис. 1. Импорт необходимых библиотек и модулей

```
def remove_outliers_iqr(df, columns=None):
    if columns is None:
        columns = df.select_dtypes(include=[np.number]).columns

    df_clean = df.copy()

    for col in columns:
        if col in df.columns:
            Q1 = df[col].quantile(0.25)
            Q3 = df[col].quantile(0.75)
            IQR = Q3 - Q1
            lower_bound = Q1 - 1.5 * IQR
            upper_bound = Q3 + 1.5 * IQR

            mask = (df[col] >= lower_bound) & (df[col] <= upper_bound)
            df_clean = df_clean[mask]

    return df_clean
```

Рис. 2. Определение функции для удаления выбросов

Загруженные данные нужно привести к нужному формату, подходящему для дальнейшего анализа. Из набора данных с показателями рынка труда требуется извлечь один показатель – уровень безработицы. Также нужно взять только те показатели, что относятся к регионам России – данные о стране в целом не представляют интереса в рамках данного исследования. Из набора данных с показателями рынка труда были взяты только нужные столбцы (наименование объекта, год, значение показателя) и выполнена фильтрация по наименованию показателя (Уровень безработицы: Уровень безработицы) и типу объекта (регион), после чего добавлен столбец со значением уровня безработицы на следующий год и выполнена очистка набора данных от пропусков и выбросов.

Наборы данных с показателями ВРП, промышленными показателями и показателями деятельности компаний имеют следующую структуру:

- наименование объекта;
- год;
- наименование показателя;
- значение показателя.

Для дальнейшего использования этих наборов данных созданы столбцы, соответствующие каждому отдельному показателю. Такая структура приемлема для применения методов машинного обучения. Из данных по ВРП взяты только нужные столбцы (наименование

объекта, год, наименование показателя, значение показателя), выполнена фильтрация по типу объекта (регион) и исключены показатели, привязанные к отдельным временным периодам (т. е. не покрывающим весь временной период данных). Столбцы, соответствующие каждому отдельному показателю, добавлены при помощи функции `pivot()`, позволяющей создать сводную таблицу на основе входных данных. Наборы данных с промышленными показателями и показателями деятельности компаний претерпели аналогичные преобразования.

После обработки полученные наборы данных объединены в один набор, который затем был очищен от пропусков. Столбец с уровнем безработицы на текущий год был исключён из набора, так как он слишком тесно связан с уровнем безработицы на следующий год, что может внести существенные помехи в работу моделей.

Объединённый набор данных затем был поделён на набор признаков (X) и набор целевых значений (y). В рамках данного исследования целевым показателем является уровень безработицы в регионе на следующий год. Набор признаков и набор целевых значений затем были поделены на обучающую и тестовую выборки.

Предобработка данных завершена, можно приступать к обучению моделей. Первый метод, рассматриваемый в рамках данного исследования – линейная регрессия. Линейная регрессия является фундаментальным методом статистического моделирования, позволяющим оценивать и анализировать отношения между зависимой переменной и одной или несколькими независимыми переменными [8].

На рис. 3 представлен код для обучения и оценки модели линейной регрессии и значения метрик качества полученной модели.

```
model = LinearRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

rmse = root_mean_squared_error(y_test, y_pred)
print(f'RMSE: {rmse}')

mae = mean_absolute_error(y_test, y_pred)
print(f'MAE: {mae}')

r2 = r2_score(y_test, y_pred)
print(f'R^2: {r2}')
```

RMSE: 2.1596740181330842
MAE: 1.6927010409345098
R^2: 0.25392297220121673

Рис. 3. Обучение и оценка модели линейной регрессии

Для наглядности построена визуализация, отображающая расхождение фактических значений целевого показателя с предсказанными моделью значениями, как показано на рис. 4. Для данной визуализации использованы показатели, относящиеся к Хабаровскому краю. В дальнейших визуализациях в рамках данного исследования будут использоваться эти же данные.

```
test_data = data[data['object_name']=='Хабаровский край']
y_test_2 = test_data[['year', 'unemployment_next_year']]
X_test_2 = test_data.drop(columns=['object_name', 'year', 'unemployment_next_year'])
y_pred_2 = model.predict(X_test_2)

fig = plt.figure(figsize=(8,6))
ax = fig.gca()
ax.plot(y_test_2['year'], y_test_2['unemployment_next_year'])
ax.plot(y_test_2['year'], y_pred_2)

ax.xaxis.set_major_locator(MaxNLocator(integer=True))
plt.show()
```

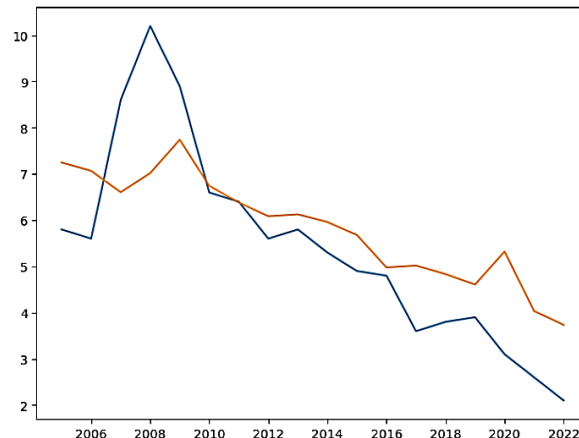


Рис. 4. Визуализация фактических и предсказанных значений целевого показателя (Линейная регрессия)

Перейдём к следующему методу машинного обучения – дереву решений. Дерево решений является гибким алгоритмом машинного обучения, используемым для решения задач классификации и регрессии. Они строятся путём разделения данных на основе признаков с целью минимизации неопределённости или увеличения чистоты классов в каждом листе дерева. Математическая модель дерева решений представляет собой древовидную структуру, состоящую из узлов и листьев. Каждый узел представляет собой тест на определённое условие, а каждый лист представляет собой прогнозируемый класс или значение [9].

Этот метод имеет гиперпараметры, настройка которых может оказать существенное влияние на качество итоговой модели. В целях подбора оптимальных гиперпараметров была выполнена кросс-валидация, как показано на рис. 5.

Суть использованного метода кросс-валидации заключается в том, что при помощи цикла происходит обучение и оценка множества моделей с разными гиперпараметрами, метрики качества которых затем сравниваются друг с другом, и на основе наилучших значений метрик качества выбирается оптимальный набор гиперпараметров. В примере на рис. 5 приведено 3 набора гиперпараметров, каждый из которых был выбран на основе наилучшего значения одной из трёх метрик качества. В двух из трёх случаев наилучшим значением гиперпараметра `max_depth` является 7, и во всех трёх случаях наилучшим значением гиперпараметра `min_samples_split` является 2. Следовательно, именно эти значения выбраны при обучении модели дерева решений. Гиперпараметры для методов машинного обучения, рассматриваемых в дальнейшем, подобраны аналогичным образом.

```

rmse_tree = []
mae_tree = []
r2_tree = []
for i in range(2,31):
    for j in range(2,6):
        tree = DecisionTreeRegressor(
            max_depth = i,
            min_samples_split = j,
            random_state = 42
        )
        tree.fit(X_train, y_train)
        y_pred = tree.predict(X_test)
        rmse = root_mean_squared_error(y_test, y_pred)
        mae = mean_absolute_error(y_test, y_pred)
        r2 = r2_score(y_test, y_pred)
        rmse_tree.append(rmse)
        mae_tree.append(mae)
        r2_tree.append(r2)

        if rmse == min(rmse_tree):
            rmse_best_params = [i, j]
        if mae == min(mae_tree):
            mae_best_params = [i, j]
        if r2 == max(r2_tree):
            r2_best_params = [i, j]

print(rmse_best_params)
print(mae_best_params)
print(r2_best_params)

```

```

[7, 2]
[10, 2]
[7, 2]

```

Рис. 5. Кросс-валидация (Дерево решений)

На рис. 6 представлен код для обучения и оценки модели Дерева решений и значения метрик качества полученной модели.

```

tree = DecisionTreeRegressor(
    max_depth=7,
    min_samples_split = 2,
    random_state = 42)
tree.fit(X_train, y_train)

y_pred = tree.predict(X_test)

rmse = root_mean_squared_error(y_test, y_pred)
print(f'RMSE: {rmse}')

mae = mean_absolute_error(y_test, y_pred)
print(f'MAE: {mae}')

r2 = r2_score(y_test, y_pred)
print(f'R^2: {r2}')

```

```

RMSE: 1.8240261092949899
MAE: 1.3059538598351663
R^2: 0.4678067498964815

```

Рис. 6. Обучение и оценка модели Дерева решений

Построена визуализация, отображающая расхождение фактических значений целевого показателя с предсказанными моделью значениями, как показано на рис. 7.

```

y_pred_2 = tree.predict(X_test_2)

fig = plt.figure(figsize=(8,6))
ax = fig.gca()
ax.plot(y_test_2['year'], y_test_2['unemployment_next_year'])
ax.plot(y_test_2['year'], y_pred_2)

ax.xaxis.set_major_locator(MaxNLocator(integer=True))
plt.show()

```



Рис. 7. Визуализация фактических и предсказанных значений целевого показателя (Дерево решений)

Следующий метод машинного обучения – Случайный лес. Этот метод относится к методам ансамблевого обучения. Этот алгоритм устраняет сложности, возникающие из-за нелинейных взаимосвязей и многомерных данных, которые невозможно адекватно смоделировать с помощью линейной регрессии. Благодаря построению множества деревьев решений и объединению их прогнозов, метод случайного леса обеспечивает надёжную прогностическую модель, которая отражает сложные взаимодействия между рабочими параметрами [10]. Этот метод особенно эффективен для повышения точности прогнозов в условиях неопределённости и изменчивости, которые характерны для социально-экономических показателей вроде уровня безработицы из-за постоянного изменения обстановки в стране и мире.

На рис. 8 представлен код для обучения и оценки модели Случайного леса и значения метрик качества полученной модели.

```

rf = RandomForestRegressor(
    n_estimators=100,
    random_state=42,
    max_depth = 17,
    max_features = 'sqrt'
)

rf.fit(X_train, y_train)

predictions = rf.predict(X_test)

rmse = root_mean_squared_error(y_test, predictions)
print(f'RMSE: {rmse}')

mae = mean_absolute_error(y_test, predictions)
print(f'MAE: {mae}')

r2 = r2_score(y_test, predictions)
print(f'R^2: {r2}')

```

```

RMSE: 1.3527219327189555
MAE: 1.0377109673938318
R^2: 0.7072989205000095

```

Рис. 8. Обучение и оценка модели Случайного леса

Построена визуализация, отображающая расхождение фактических значений целевого показателя с предсказанными моделью значениями, как показано на рис. 9.

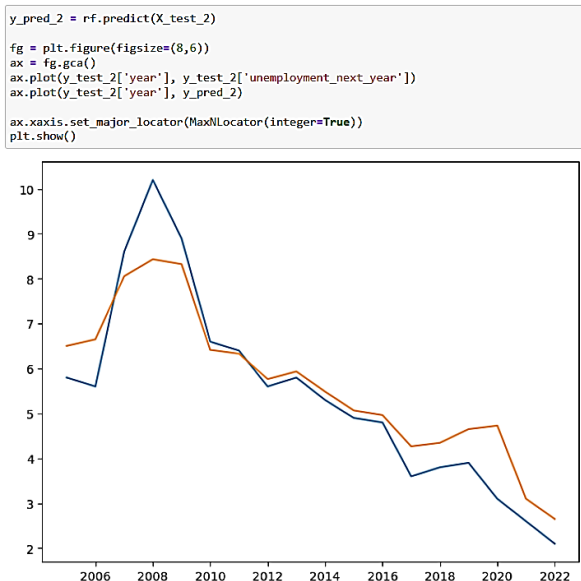


Рис. 9. Визуализация фактических и предсказанных значений целевого показателя (Случайный лес)

Последний метод машинного обучения, рассматриваемый в рамках данного исследования – Градиентный бустинг. Он, как и Случайный лес, относится к методам ансамблевого обучения. Принцип работы Градиентного бустинга заключается в совокупности слабых моделей прогнозирования, которые представляют собой Деревья решений [11]. На каждой итерации алгоритм подбирает новую модель к остаткам предыдущих моделей, а затем объединяет их для формирования окончательного прогноза [12].

На рис. 10 представлен код для обучения и оценки модели Градиентного бустинга и значения метрик качества полученной модели.

```
gb = GradientBoostingRegressor(
    random_state=42,
    learning_rate = 0.15,
    max_depth = 6,
    min_samples_split = 5
)
gb.fit(X_train, y_train)

y_pred = gb.predict(X_test)

rmse = root_mean_squared_error(y_test, y_pred)
print(f'RMSE: {rmse}')

mae = mean_absolute_error(y_test, y_pred)
print(f'MAE: {mae}')

r2 = r2_score(y_test, y_pred)
print(f'R^2: {r2}')
```

RMSE: 1.1583126853325352
MAE: 0.8655250355895914
R^2: 0.785385586400108

Рис. 10. Обучение и оценка модели Градиентного бустинга

Построена визуализация, отображающая расхождение фактических значений целевого показателя с предсказанными моделью значениями, как показано на рис. 11.

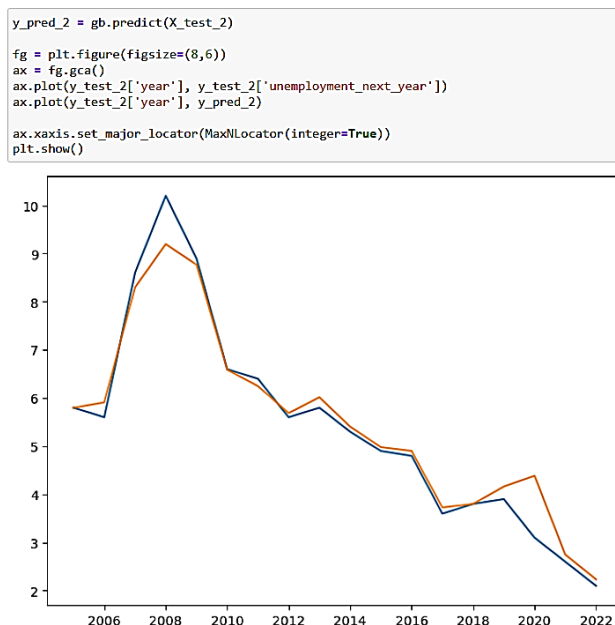


Рис. 11. Визуализация фактических и предсказанных значений целевого показателя (Градиентный бустинг)

В табл. 1 представлены метрики качества обученных моделей, округлённый до трёх знаков после запятой. На основе полученных метрик качества можно сделать вывод, что метод Градиентного бустинга является наиболее эффективным для решения задачи прогнозирования уровня безработицы в регионах РФ относительно других рассмотренных методов машинного обучения, так как этот метод выигрывает по всем рассчитанным метрикам качества.

Таблица 1. Метрики качества обученных моделей

Метод машинного обучения	RMSE	MAE	R ²
Линейная регрессия	2,160	1,693	0,254
Дерево решений	1,824	1,306	0,468
Случайный лес	1,353	1,038	0,707
Градиентный бустинг	1,158	0,866	0,785

Выводы. Эффективность методов машинного обучения для прогнозирования уровня безработицы в регионах России подтверждена экспериментально. Все рассмотренные модели показали различную степень точности, при этом наилучшие результаты продемонстрировали ансамблевые методы.

Градиентный бустинг оказался наиболее эффективным методом среди рассмотренных. Он показал наименьшие значения ошибок (RMSE = 1,158, MAE = 0,866) и наибольшее значение коэффициента детерминации (R² = 0,785), что свидетельствует о его высокой прогностической способности.

Случайный лес также показал достойные результаты (RMSE = 1,353, MAE = 1,038, R² = 0,707), уступая градиентному бустингу, но значительно превосходя простые модели.

Дерево решений и Линейная регрессия показали относительно низкую точность прогнозирования, что объясняется их ограниченной способностью учитывать сложные нелинейные зависимости и взаимодействия признаков.

Важность предобработки данных и формирования информативного набора признаков была подтверждена

в ходе исследования, однако оптимальный набор признаков может быть уточнён в дальнейших работах.

Таким образом, применение методов машинного обучения, в частности градиентного бустинга, открывает новые возможности для повышения качества прогнозирования социально-экономических показателей вроде уровня безработицы на региональном уровне.

Литература

1. Бурковский П., Мамий С.А., Бочкова Т.А. Безработица и её типы. Проблемы безработицы в России // МНИЖ. 2023. № 12 (138). URL: <https://cyberleninka.ru/article/n/bez-rabotitsa-i-ee-tipy-problemy-bezrabotitsy-v-rossii> (дата обращения: 13.11.2025).
2. Кузнецов М.Ю. Региональные особенности безработицы // Альманах «Крым». 2023. № 40. URL: <https://cyberleninka.ru/article/n/regionalnye-osobennosti-bezrabotitsy> (дата обращения: 13.11.2025).
3. Социально-экономические показатели регионов России // Росстат; обработка «Если быть точным», 2023. URL: https://tochno.st/datasets/regions_collection (дата обращения: 13.11.2025).
4. Жминько А.Е., Самойленко А.Д., Брусенская М.О., Федосова Ю.В. Исследование уровня занятости населения в Российской Федерации // ЕГИ. 2025. № 3 (59). URL: <https://cyberleninka.ru/article/n/issledovanie-urovnya-zanyatosti-naseleniya-v-rossiyskoy-federatsii> (дата обращения: 13.11.2025).
5. Баринов М.А. Использование индикаторов социальной направленности для мониторинга уровня регионального развития // Beneficium. 2024. № 4 (53). URL: <https://cyberleninka.ru/article/n/ispolzovanie-indikatorov-sotsialnoy-napravlenosti-dlya-monitoringa-urovnya-regionalnogo-razvitiya> (дата обращения: 13.11.2025).
6. Голоднюк Р.А. Состояние и тенденции в развитии угольной отрасли Донецкой народной республики // МНИЖ. 2024. № 9 (147). URL: <https://cyberleninka.ru/article/n/sostoyanie-i-tendentsii-v-razviti-ugolnoy-otrasli-donetskoj-narodnoy-respubliki> (дата обращения: 13.11.2025).
7. Пьянкова С.Г., Ергунова О.Т. Специфика развития городов под влиянием деятельности крупных промышленных компаний в России // Проблемы развития территории. 2025. № 1. URL: <https://cyberleninka.ru/article/n/spetsifika-razvitiya-gorodov-pod-vliyaniem-deyatelnosti-kрупnyh-promyshlennyh-kompaniy-v-rossii> (дата обращения: 13.11.2025).
8. Евсеева К.В., Ватлашов С.А., Литвиненко А.Н., Конников Е.А. Компаративный анализ факторной специфики финансовых показателей энергетических корпораций средствами апплицированных методов машинного обучения // Вестник Академии знаний. 2024. № 3 (62). URL: <https://cyberleninka.ru/article/n/komparativnyy-analiz-faktornoy-spetsifiki-finansovyh-pokazateley-energeticheskikh-korporatsiy-sredstvami-aplitsirovannyh-metodov> (дата обращения: 13.11.2025).
9. Кондратов Д.В., Володин Д.Н. Математическое моделирование алгоритмов машинного обучения // Математическое моделирование, компьютерный и натурный эксперимент в естественных науках. 2023. № 2. URL: <https://cyberleninka.ru/article/n/matematicheskoe-modelirovanie-algoritmov-mashinnogo-obucheniya> (дата обращения: 13.11.2025).
10. Астраханцева И.А., Котенев Т.Е., Горев С.В. Астраханцев Р.Г., Грименицкий П.Н. Интеграция методов линейной регрессии и случайного леса в системы интеллектуальной поддержки управления криогенными транспортными системами // Известия ВУЗов ЭФУ. 2024. № 3 (61). URL: <https://cyberleninka.ru/article/n/integratsiya-metodov-lineynoy-regressii-i-sluchaynogo-lesa-v-sistemy-intellektualnoy-podderzhki-upravleniya-kriogennymi> (дата обращения: 13.11.2025).

11. Антропова Е.М., Конкин Н.А. Выбор ансамблевых моделей машинного обучения для прогнозирования полосы когерентности трансионосферных каналов связи // Всероссийские открытые Армандовские чтения: Современные проблемы дистанционного зондирования, радиолокации, распространения и дифракции волн. 2023. № 1. URL: <https://cyberleninka.ru/article/n/vybor-ansamblevyh-modeley-mashinnogo-obucheniya-dlya-prognozirovaniya-polosy-kogerentnosti-transionosfernnyh-kanalov-svyazi> (дата обращения: 13.11.2025).
12. Макеев К.А., Греченева А.В., Котов Я.С., Голбан А.Н. Смыслов Д.М. Сравнение эффективности алгоритмов машинного обучения в задачах прогнозирования урожайности сельскохозяйственных культур // Известия ТулГУ. Технические науки. 2023. № 2. URL: <https://cyberleninka.ru/article/n/sravnenie-effektivnosti-algoritmov-mashinnogo-obucheniya-v-zadachah-prognozirovaniya-urozhaynosti-selskohozyaystvennyh-kultur> (дата обращения: 13.11.2025).

References

1. Burkovsky P., Mamiy S. A., Bochkova T. A. Unemployment and its types. Unemployment problems in Russia // MNIZH. 2023. No. 12 (138). URL: <https://cyberleninka.ru/article/n/bez-rabotitsa-i-ee-tipy-problemy-bezrabotitsy-v-rossii> (date of access: 13.11.2025).
2. Kuznetsov M. Yu. Regional Features of Unemployment // Almanac "Crimea". 2023. No. 40. URL: <https://cyberleninka.ru/article/n/regionalnye-osobennosti-bezrabotitsy> (date of access: 13.11.2025).
3. Socio-economic indicators of Russian regions // Rosstat; processed "If to Be Precise", 2023. URL: https://tochno.st/datasets/regions_collection (date of access: 13.11.2025).
4. Zhminko A.E., Samoylenko A.D., Brusenskaya M.O., Fedosova Yu.V. Study of the level of employment of the population in the Russian Federation // EGI. 2025. No. 3 (59). URL: <https://cyberleninka.ru/article/n/issledovanie-urovnya-zanyatosti-naseleniya-v-rossiyskoy-federatsii> (date of access: 13.11.2025).
5. Barinov M.A. Using socially oriented indicators to monitor the level of regional development // Beneficium. 2024. No. 4 (53). URL: <https://cyberleninka.ru/article/n/ispolzovanie-indikatorov-sotsialnoy-napravlenosti-dlya-monitoringa-urovnya-regionalnogo-razvitiya> (accessed: 13.11.2025).
6. Golodnyuk R.A. State and trends in the development of the coal industry of the Donetsk people's republic // MNIZH. 2024. No. 9 (147). URL: <https://cyberleninka.ru/article/n/sostoyanie-i-tendentsii-v-razviti-ugolnoy-otrasli-donetskoj-narodnoy-respubliki> (date of access: 13.11.2025).
7. Pyankova S.G., Ergunova O.T. Specificity of city development under the influence of the activities of large industrial companies in Russia // Problems of Territorial Development. 2025. No. 1. URL: <https://cyberleninka.ru/article/n/spetsifika-razvitiya-gorodov-pod-vliyaniem-deyatelnosti-kрупnyh-promyshlennyh-kompaniy-v-rossii> (date of access: 13.11.2025).
8. Evseeva K.V., Vatlashov S.A., Litvinenko A.N., Konnikov E.A. Comparative analysis of factor specificity of financial indicators

- tors of energy corporations using applied machine learning methods // Bulletin of the Knowledge Academy. 2024. No. 3 (62). URL: <https://cyberleninka.ru/article/n/komparativnyy-analiz-faktornoy-spetsifiki-finansovyh-pokazateley-energeticheskikh-korporatsiy-sredstvami-aplitsirovannyh-metodov> (accessed: 13.11.2025).
9. Kondratov D.V., Volodin D.N. Mathematical modeling of machine learning algorithms // Mathematical Modeling, Computer and Natural Experiments in Natural Sciences. 2023. No. 2. URL: <https://cyberleninka.ru/article/n/matematicheskoe-modelirovanie-algoritmov-mashinnogo-obucheniya> (date of access: 11/13/2025).
 10. Astrakhantseva I.A., Kotenev T.E., Gorev S.V., Astrakhansev R.G., Grimenitsky P.N. Integration of linear regression and random forest methods into intelligent control support systems for cryogenic transport systems // Bulletin of Universities of Economics and Management. 2024. No. 3 (61). URL: <https://cyberleninka.ru/article/n/integratsiya-metodov-lineynoy-regressii-i-sluchaynogo-lesa-v-sistemy-intellektualnoy-podderzhki-upravleniya-kriogennymi> (accessed: 13.11.2025).
 11. Antropova E.M., Konkin N.A. "Selection of Ensemble Machine Learning Models for Predicting the Coherence Bandwidth of Transionospheric Communication Channels" in: All-Russian Open Armandov Readings: Modern Problems of Remote Sensing, Radar, Wave Propagation and Diffraction. 2023, no. 1. URL: <https://cyberleninka.ru/article/n/vybor-ansamblevyh-modeley-mashinnogo-obucheniya-dlya-prognozirovaniya-polosy-kogerentnosti-transionosfernyh-kanalov-svyazi> (accessed: 11/13/2025).
 12. Makeev K.A., Grecheneva A.V., Kotov Ya.S., Golban A.N., Smyslov D.M. Comparison of the efficiency of machine learning algorithms in the problems of forecasting agricultural crops yield // Bulletin of TulSU. Technical sciences. 2023. No. 2. URL: <https://cyberleninka.ru/article/n/sravnenie-effektivnosti-algoritmov-mashinnogo-obucheniya-v-zadachah-prognozirovaniya-urozhaynosti-selskohozyaystvennyh-kultur> (date of access: 13.11.2025).