

УДК 004.9

ПРОГНОЗИРОВАНИЕ РОСТА ЧИСЛА ИНТЕРНЕТ-ПОЛЬЗОВАТЕЛЕЙ В РАЗЛИЧНЫХ РЕГИОНАХ МИРА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ**FORECASTING THE GROWTH OF INTERNET USERS IN VARIOUS WORLD REGIONS USING MACHINE LEARNING METHODS**

М. В. Пистер ИЭГУиФ «Институт экономики, государственного управления и финансов СФУ», г. Красноярск

Научный руководитель: Семёнова Анна Робертовна, кандидат физико-математических наук, доцент.

Аннотация

В статье представлены результаты анализа данных о количестве интернет-пользователей в различных регионах мира с использованием методов машинного обучения. На основе исторических данных построены модели полиномиальной регрессии, позволяющие прогнозировать рост числа интернет-пользователей на период 2025–2030 годов. Полученные результаты могут быть использованы для планирования цифровой инфраструктуры и разработки стратегий цифровой трансформации в государственном и муниципальном управлении.

Ключевые слова: машинное обучение, прогнозирование, интернет-пользователи, цифровизация, полиномиальная регрессия.

Abstract

The article presents the results of an analysis of data on the number of Internet users in various world regions using machine learning methods. Based on historical data, polynomial regression models were built to forecast the growth of Internet users for the period 2025–2030. The obtained results can be used for planning digital infrastructure and developing digital transformation strategies in state and municipal governance.

Keywords: machine learning, forecasting, Internet users, digitalization, polynomial regression.

Введение

Цифровизация стала ключевым фактором в повышении эффективности государственного и муниципального управления. Одним из важных аспектов цифровизации является анализ данных, которые помогают прогнозировать будущие тренды и принимать обоснованные решения. В данной статье мы использовали методы машинного обучения для анализа данных о количестве интернет-пользователей по регионам мира.

Данные и методы

Для анализа использовались данные с сайта Our World in Data, содержащие информацию о количестве интернет-пользователей по странам и годам. Данные были обработаны с использованием библиотеки `pandas` в Python.

Анализ

Мы построили три модели: линейную регрессию, дерево решений и случайный лес. Наилучший результат показал случайный лес (MSE: $6.05e+15$), что свидетельствует о его способности эффективно работать с нелинейными данными.

```
# Подготовка данных
# Используем 'Entity' (страна/регион), 'Year' (год) и 'num_internet_users' (количество пользователей)
X = df[['Year', 'Entity']] # Признаки: год и страна/регион
y = df['num_internet_users'] # Целевая переменная: количество интернет-пользователей

# Преобразуем категориальные данные (страны/регионы) в числовые с помощью one-hot encoding
X = pd.get_dummies(X, columns=['Entity'], drop_first=True)

# Разделение данных на обучающую и тестовую выборки
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Линейная регрессия
lr = LinearRegression()
lr.fit(X_train, y_train)
y_pred_lr = lr.predict(X_test)
mse_lr = mean_squared_error(y_test, y_pred_lr)

# Дерево решений
dt = DecisionTreeRegressor()
dt.fit(X_train, y_train)
y_pred_dt = dt.predict(X_test)
mse_dt = mean_squared_error(y_test, y_pred_dt)

# Случайный лес
rf = RandomForestRegressor()
rf.fit(X_train, y_train)
y_pred_rf = rf.predict(X_test)
mse_rf = mean_squared_error(y_test, y_pred_rf)

# Вывод результатов
print(f"MSE Linear Regression: {mse_lr}")
```

Рисунок 1 Код для анализа данных по числу интернет-пользователей в мире

Интерпретация результатов моделей

Линейная регрессия (MSE: $3.56e+17$): очень высокое значение среднеквадратичной ошибки (MSE) указывает на то, что линейная регрессия плохо справляется с прогнозированием количества интернет-пользователей. Это может быть связано с тем, что данные имеют сложную нелинейную структуру, которую линейная модель не может уловить.

Дерево решений (MSE: $8.34e+15$): дерево решений показало значительно лучший результат по сравнению с линейной регрессией. Это говорит о том, что модель смогла уловить некоторые нелинейные зависимости в данных.

Случайный лес (MSE: $6.05e+15$): случайный лес показал наилучший результат среди всех моделей. Это ожидаемо, так как ансамблевые методы, такие как случайный лес, обычно работают лучше на сложных данных благодаря комбинации множества деревьев.

Прогнозирование

Для прогнозирования применялась полиномиальная регрессия второй степени, которая позволяет учитывать нелинейные тренды в данных. При этом работает на относительно небольшом количестве данных, по сравнению со случайным лесом.

```
import pandas as pd
import requests

# Загрузка данных
df = pd.read_csv("https://ourworldindata.org/grapher/number-of-internet-users.csv?v=1&csvType=filtered&useColumnShortNames=true", storage_options = {'User-Agent': 'Our World In Data data fetch/1.0'})

# Загрузка метаданных
metadata = requests.get("https://ourworldindata.org/grapher/number-of-internet-users.metadata.json?v=1&csvType=filtered&useColumnShortNames=true").json()

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error

import pandas as pd
import requests
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
import pandas as pd
import requests
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import make_pipeline
import matplotlib.pyplot as plt
import numpy as np

# Загрузка данных
url = "https://ourworldindata.org/grapher/number-of-internet-users.csv?v=1&csvType=filtered&useColumnShortNames=true"
df = pd.read_csv(url, storage_options={'User-Agent': 'Our World In Data data fetch/1.0'})

# Проверка данных
print('Столбцы в данных:', df.columns) # Убедимся, что столбцы правильно загружены
print('Уникальные регионы:', df['Entity'].unique()) # Проверка названий регионов
```

Рисунок 2 Код для прогнозирования данных по числу интернет-пользователей в мире

```

from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import make_pipeline
import matplotlib.pyplot as plt
import numpy as np

# Загрузка данных
url = "https://ourworldindata.org/grapher/number-of-internet-users_cav?v=latest&type=filtered&useColumnsShortNames=true"
df = pd.read_csv(url, storage_options={'User-Agent': 'Our World In Data data fetch/1.0'})

# Проверка данных
print("Страны в данных:", df.columns) # Проверка, что страны правильно загружены
print("Уникальные регионы:", df['entity'].unique()) # Проверка названий регионов

# Список регионов для анализа
regions = ['Africa', 'Asia', 'Europe', 'North America', 'Oceania', 'South America']

# Функция для предсказания
def predict_internet_users(region_name):
    """
    # Загрузка данных по региону
    region_data = df[df['entity'] == region_name]

    # Проверка, что данные не пустые
    if region_data.empty:
        print(f"Нет данных для региона: {region_name}")
        return None

    # Удаление строк с пропущенными значениями в целевом столбце
    region_data = region_data.dropna(subset=['num_internet_users'])

    # Проверка, что после удаления пропущенных значений данные не пустые
    if region_data.empty:
        print(f"Нет данных после удаления пропущенных значений для региона: {region_name}")
        return None

    # Подготовка данных
    x = region_data['year'] # Годы: год
    y = region_data['num_internet_users'] # Целевая переменная

    # Проверка, что данные не пустые
    if len(x) < 2:
        print(f"Недостаточно данных для регрессии: {region_name}")
        return None

    # Создание модели (полиномиальная регрессия, степень 2)
    poly = make_pipeline(PolynomialFeatures(degree=2), LinearRegression())
    poly.fit(x, y) # Обучение на исходных данных

    # Прогнозирование на будущие годы (2025-2030)
    future_years = np.array([2025, 2026, 2027, 2028, 2029, 2030]).reshape(-1, 1)
    future_predictions = poly.predict(future_years)

    # Возвращаем исходные данные и прогнозы
    plt.figure(figsize=(10, 6))

    # Исходные данные
    plt.scatter(x, y, color='blue', label='Исходные данные')

    # Линия тренда (полиномиальная регрессия, степень 2)
    all_years = np.vstack([x, future_years]) # Объединяем исторические и будущие годы
    all_predictions = poly.predict(all_years) # Прогнозы для всех лет
    plt.plot(all_years, all_predictions, color='red', label='Тренд и прогноз')

    # Настройка графика
    plt.xlabel('год')
    plt.ylabel('количество интернет-пользователей')
    plt.title(f'Прогноз роста числа интернет-пользователей в {region_name} (2025-2030)')
    plt.legend()
    plt.grid()
    plt.show()

    return future_predictions

# Прогнозирование для каждого региона
for region_name in regions:
    print(f"Прогноз для {region_name}:")
    predictions = predict_internet_users(region_name)
    if predictions is not None:
        print(predictions)

```

Результаты

На основе исторических данных были сделаны прогнозы роста числа интернет-пользователей на 2025–2030 годы для различных регионов мира. Результаты представлены на графиках, которые демонстрируют исходные данные (синие точки) и линию тренда (красная линия), построенную с использованием полиномиальной регрессии второй степени:

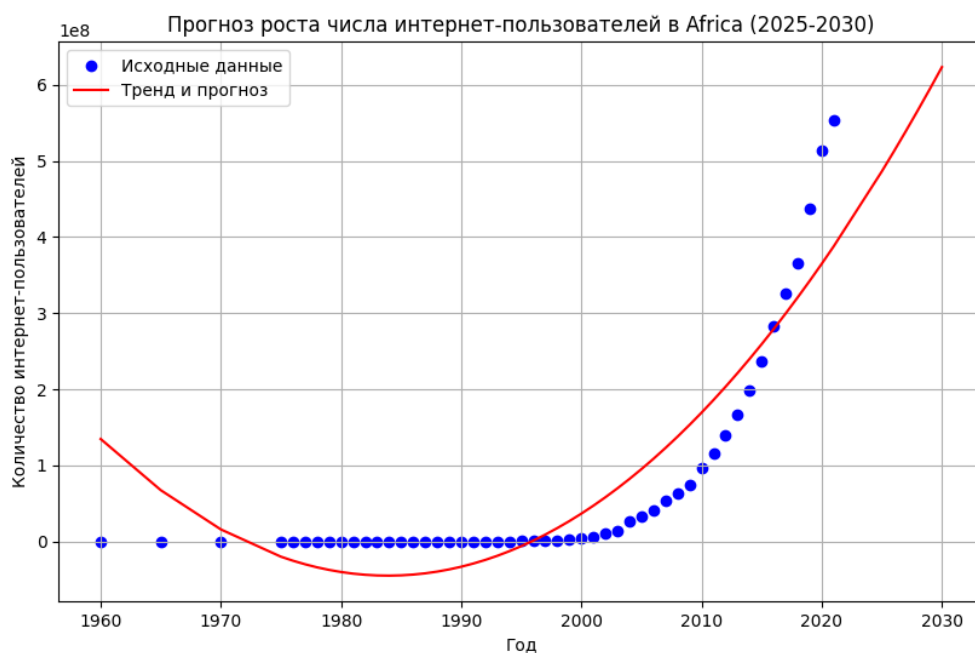


Рисунок 3 Прогноз роста числа интернет-пользователей в Африке

Африка: Прогнозируемое количество интернет-пользователей увеличится с 486 млн в 2025 году до 623 млн в 2030 году.



Рисунок 4 Прогноз роста числа интернет-пользователей в Азии

Азия: Прогнозируемое количество интернет-пользователей увеличится с 2.78 млрд в 2025 году до 3.52 млрд в 2030 году.

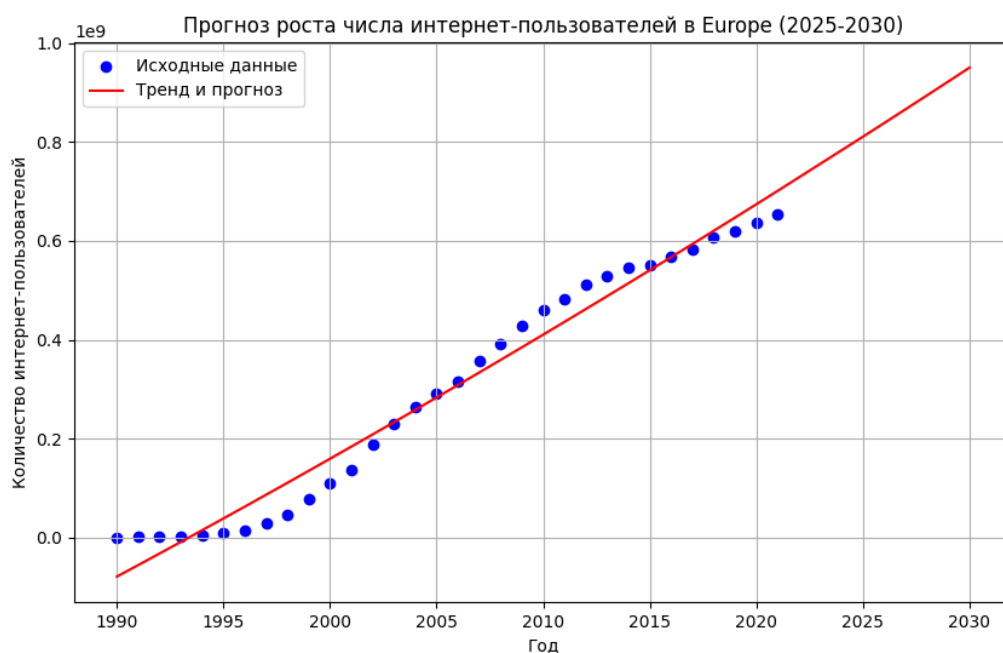


Рисунок 5 Прогноз роста числа интернет-пользователей в Европе

Европа: Прогнозируемое количество интернет-пользователей увеличится с 811 млн в 2025 году до 950 млн в 2030 году.

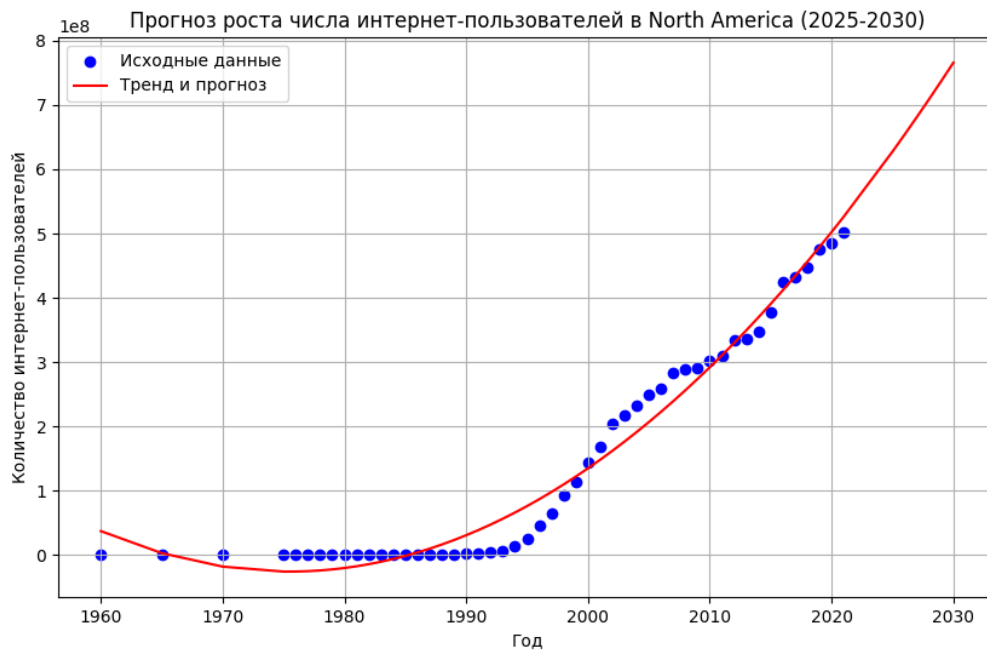


Рисунок 6 Прогноз роста числа интернет-пользователей в Северной Америке

Северная Америка: Прогнозируемое количество интернет-пользователей увеличится с 627 млн в 2025 году до 766 млн в 2030 году.

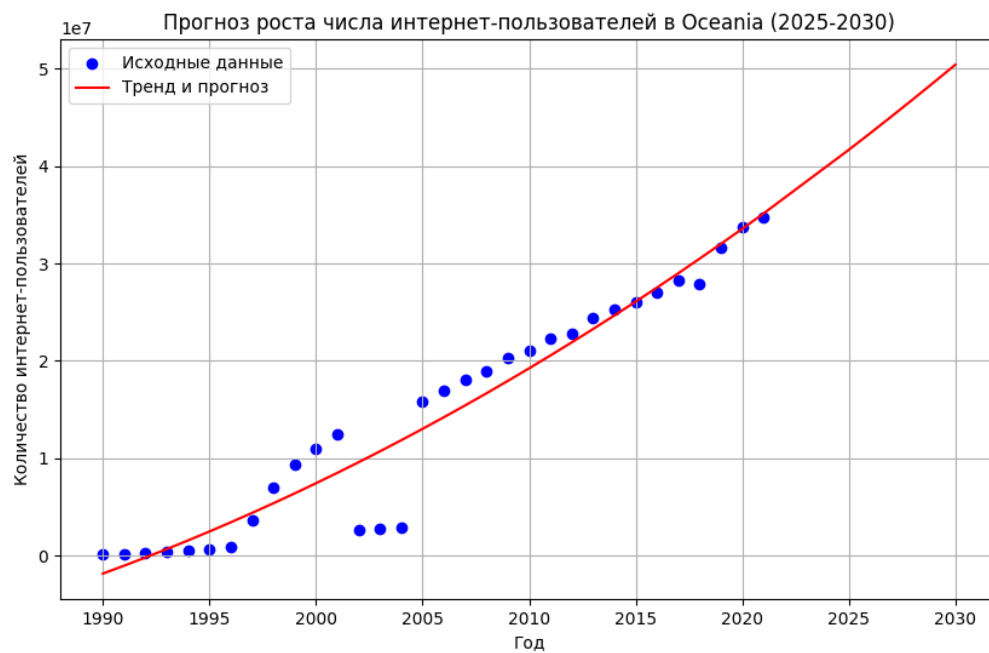


Рисунок 7 Прогноз роста числа интернет-пользователей в Океании

Океания: Прогнозируемое количество интернет-пользователей увеличится с 41.7 млн в 2025 году до 50.4 млн в 2030 году.

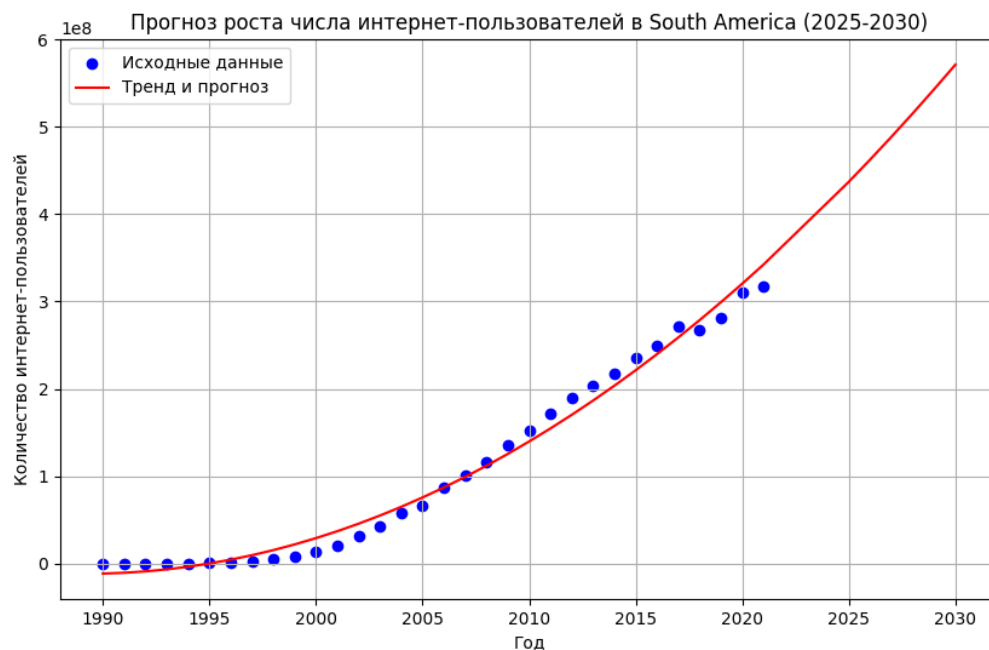


Рисунок 8 Прогноз роста числа интернет-пользователей в Южной Америке

Южная Америка: Прогнозируемое количество интернет-пользователей увеличится с 437 млн в 2025 году до 571 млн в 2030 году.

Также рассчитали прогноз для доли интернет-пользователей в России.



Рисунок 9 Прогноз доли интернет-пользователей в России

Прогнозируемая доля интернет-пользователей увеличится в 2025 году на 17,66%, а в 2030 году на 50,75% по сравнению с 2023 годом.

Результаты анализа данных о количестве интернет-пользователей и их прогнозирование с помощью машинного обучения могут использоваться государством для:

1. Прогнозирование спроса на цифровые услуги:

Государственные органы могут использовать подобные модели для прогнозирования роста числа интернет-пользователей. Это поможет планировать развитие цифровой инфраструктуры, например, расширение доступа к интернету в отдаленных регионах.

2. Оптимизация бюджетов:

Зная, как быстро растёт количество интернет-пользователей, государство может более эффективно распределять бюджет на цифровые проекты, такие как создание электронных государственных услуг (например, порталы "Госуслуги").

3. Улучшение доступности услуг:

Прогнозирование помогает понять, в каких регионах или странах необходимо ускорить цифровизацию. Это особенно важно для обеспечения равного доступа к государственным услугам для всех граждан.

4. Анализ эффективности цифровых инициатив:

Модели машинного обучения могут использоваться для оценки эффективности уже реализованных программ цифровизации. Например, можно сравнить прогнозируемое и фактическое количество интернет-пользователей после внедрения новых технологий.

5. Поддержка принятия решений:

Данные и прогнозы могут стать основой для принятия решений на уровне государственного и муниципального управления. Например, если модель предсказывает резкий рост числа интернет-пользователей, это может стать сигналом для ускорения разработки новых цифровых сервисов.

Заключение

Использование методов машинного обучения для анализа данных о количестве интернет-пользователей позволяет не только прогнозировать будущие тренды, но и разрабатывать стратегии цифровой трансформации. Полученные результаты могут быть использованы государственными органами для планирования развития цифровой инфраструктуры и обеспечения равного доступа к интернету для всех граждан.

Список литературы

1. Our World in Data. – URL: <https://ourworldindata.org/> (дата обращения: 25.02.2024).
2. Федоренко, К. П. Развитие таможенного менеджмента / К. П. Федоренко, О. В. Нетаев. – Текст: электронный // Российский внешнеэкономический вестник. – 2014. – № 4. – С. 78–96. – URL:

<https://cyberleninka.ru/article/n/razvitiie-tamozhennogo-menedzhmenta>
(дата обращения: 25.02.2024).

3. Горбуль, Ю. А. Политические режимы современных государств: монография / Ю. А. Горбуль; Кузбасский институт Федеральной службы исполнения наказаний. - Новокузнецк: Кузбасский институт ФСИН, 2019. - 63 с. - ISBN 978-5-91246-118-7. - Текст: непосредственный.
4. Радаев, В. В. Социальная стратификация: учебное пособие для вузов / Радаев Вадим, Шкаратан Овсей. – Москва : Аспект Пресс, 1996. – 318 с. – ISBN 5-7567-0119-2. – Текст: непосредственный.