

СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ ДЛЯ ПРОГНОЗИРОВАНИЯ ПОПУЛЯЦИОННЫХ РЯДОВ

А.А. Кондорова¹, К.В. Галанов¹

¹Самарский университет, г. Самара
e-mail: kondorovanastya@gmail.com, galanov.kv@ssau.ru

Аннотация

В работе проведён сравнительный анализ моделей машинного обучения для прогнозирования популяционных рядов на примере официальной статистики числа родившихся в РФ (Росстат). Оценка качества выполнена с использованием метрики RMSE. Установлено, что наилучшие результаты показали ансамблевые модели, в частности CatBoost, что подтверждает их эффективность при работе с данными сложной структуры.

Ключевые слова: машинное обучение, прогнозирование, популяционный ряд, MLR, SARIMAX, Prophet, Random Forest, XGBoost, LightGBM, CatBoost.

Введение

Целью работы является сравнительный анализ различных классов моделей машинного обучения для прогнозирования популяционных рядов на примере официальной статистики числа родившихся в Российской Федерации. Научная новизна заключается в проведении единого сравнительного исследования классических моделей временных рядов и современных ансамблевых методов на реальных демографических данных с использованием единого протокола оценки.

В рамках данной работы исследуются подходы для прогнозирования популяционных рядов. Было выбрано несколько классов моделей: регрессионные модели, модели для временных рядов и ансамблевые модели, а именно: MLR, SARIMAX, Prophet, Random Forest, XGBoost, LightGBM и CatBoost. Такой выбор обусловлен тем, что каждая из этих групп решает задачу прогнозирования с разной стороны, что позволяет сравнить подходы и выявить их сильные и слабые стороны.

1. Описание моделей машинного обучения

Множественная линейная регрессия (МЛР) – это статистический метод, используемый для моделирования взаимосвязи между зависимой переменной и одной или несколькими независимыми переменными [1].

Модели SARIMAX и Prophet широко применяются для прогнозирования временных рядов, включая популяционные данные, за счёт способности учитывать как внутреннюю структуру ряда, так и внешние факторы. Модель SARIMAX представляет собой расширение ARIMA и сочетает компоненты авторегрессии, скользящего среднего и интегрирования с возможностью учёта сезонности и экзогенных переменных [2]. Модель Prophet реализует аддитивный подход, в рамках которого

временной ряд представляется как сумма тренда, сезонных компонентов и эффектов значимых событий [3].

Ансамблевые методы машинного обучения представляют собой класс алгоритмов, основанных на объединении множества базовых моделей с целью повышения точности и устойчивости прогнозирования. К ним относятся как методы бэггинга, так и бустинга. В частности, алгоритм Random Forest реализует подход бэггинга, формируя совокупность независимых деревьев решений и агрегируя их предсказания [4]. В свою очередь, модели XGBoost, LightGBM и CatBoost относятся к методам градиентного бустинга, в которых обучение происходит последовательно: каждая последующая модель направлена на коррекцию ошибок предыдущих [5, 6, 7].

2. Постановка эксперимента

В рамках работы проведено сравнительное исследование моделей машинного обучения по качеству прогнозирования. Тестирование осуществляется на одном и том же наборе данных с равным набором выбранных гиперпараметров, что обеспечивает корректность сопоставления результатов.

Для оценки точности прогнозов используется метрика среднеквадратичной ошибки (RMSE), вычисляемая по формуле (1):

$$RMSE = \sqrt{\frac{1}{h} \sum_{i=1}^h (y_i - f_i)^2}, \quad (1)$$

где h – это количество прогнозируемых точек, y_i – реальное значение в момент времени i , f_i – это предсказанное моделью значение в тот же момент времени i [8].

В качестве исходных данных использовался временной ряд, характеризующий число родившихся в Российской Федерации по данным Росстата за период с 2006 по 2025 год в течение каждого месяца [9]. Целью исследования является построение прогноза на 4 года вперёд. В качестве признаков использовались значения самого временного ряда, преобразованные в лаговые признаки, а также календарные характеристики (месяц, год). Экзогенные переменные в рамках данного исследования не рассматривались.

Разделение выборки на обучающую и тестовую осуществлялось по временной структуре без перемешивания данных: последние 48 наблюдений использовались в качестве тестовой выборки. Для подбора гиперпараметров применялась кросс-валидация с сохранением временного порядка. Пропущенные значения в исходном ряде отсутствовали.

С целью повышения качества моделей применяется оптимизация гиперпараметров с использованием метода Grid Search. Для модели SARIMAX подбор параметров осуществлялся аналитически: $p, d, q, P, Q \in [0; 1; 2]$, с фиксированным периодом сезонности $s = 12$, соответствующим месячной частоте данных. Для модели Prophet варьировались параметры, определяющие гибкость тренда и сезонности: $changepoint_prior_scale \in [0,001; 0,01; 0,05; 0,1; 0,5]$, $seasonality_prior_scale \in [0,01; 0,1; 1; 5; 10]$, а также рассматривались различные варианты диапазона точек изменения тренда $changepoint_range \in [0,8; 0,85; 0,9; 0,95; 1,0]$. Для ансамблевых моделей осуществлялся подбор ключевых гиперпараметров. В частности, для CatBoost варьировались глубина деревьев ($depth \in [4; 6; 8; 10; 12]$), скорость обучения ($learning_rate \in [0,01; 0,03; 0,1; 0,2; 0,3]$) и количество итераций ($iterations \in [50; 100; 200; 300; 500]$). Для модели LightGBM подбирались параметры скорости

обучения и количества деревьев ($n_estimators \in [50; 100; 200; 300; 500]$). Аналогичные параметры, включая глубину деревьев, скорость обучения и количество базовых моделей, оптимизировались для алгоритмов Random Forest и XGBoost в соответствии с заданными диапазонами.

Реализация эксперимента выполнена на языке программирования Python с применением специализированных библиотек машинного обучения.

3. Результаты эксперимента

Ниже в таблице 1 представлены результаты сравнения алгоритмов.

Таблица 1 – Сравнение моделей

Метод	Значение RMSE
MLR	9891,67
SARIMAX	15981,22
Prophet	9358,36
Random Forest	11328,70
XGBoost	12861,91
LightGBM	9839,87
CatBoost	9122,44

По результатам проведённого эксперимента установлено, что наименьшее значение метрики RMSE демонстрирует модель CatBoost (9122,44), что свидетельствует о её наилучшей способности к аппроксимации исследуемого временного ряда. Близкие результаты показали модели Prophet (9358,36) и LightGBM (9839,87), что указывает на их высокую эффективность при прогнозировании данных с выраженной временной структурой. Модель множественной линейной регрессии (9891,67) уступает более сложным алгоритмам за счёт ограниченной способности учитывать нелинейные зависимости, но всё же демонстрирует качество, сопоставимое с моделью LightGBM, что свидетельствует о наличии устойчивой и достаточно регулярной структуры временного ряда. В то же время методы Random Forest (11328,70) и XGBoost (12861,91) показали более низкую точность, что может быть связано с особенностями настройки гиперпараметров и спецификой признакового пространства. Наихудший результат продемонстрировала модель SARIMAX (15981,22), что указывает на её недостаточную эффективность в рамках рассматриваемой задачи. Полученные результаты демонстрируют, что на реальных демографических данных ансамблевые методы могут превосходить классические модели временных рядов, такие как SARIMAX и Prophet, даже без использования сложного признакового описания.

Заключение

В работе проведён сравнительный анализ моделей машинного обучения для прогнозирования популяционных рядов. Установлено, что наилучшие результаты демонстрируют ансамблевые методы, в частности CatBoost, благодаря способности учитывать сложные зависимости в данных. Модели Prophet и LightGBM также показали высокую точность, тогда как классические подходы уступают им по качеству

прогнозирования. Полученные результаты подтверждают эффективность современных алгоритмов машинного обучения для решения задач анализа временных рядов.

Список использованных источников

1. Multiple Linear Regression (MLR) in Python – URL: <https://eminebozkus.medium.com/multiple-linear-regression-mlr-in-python-cf22b37ea5c5>.
2. Болсуновская, М. В. Разработка модели внутреннего туристического потока на основе модели SARIMAX / М.В. Болсуновская, Т.Р. Иванов, С. А. Рубцова // Интеллектуальная инженерная экономика и Индустрия 5.0: сборник трудов X международной научно-практической конференции. – 2024. – С. 433–436.
3. Орлова, И. В. Использование пакета Prophet в прогнозировании временных рядов / И. В. Орлова // Фундаментальные исследования. – 2021. – № 3. – С. 94–102.
4. What is random forest? – URL: <https://www.ibm.com/think/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems>.
5. XGBoost – URL: <https://www.nvidia.com/en-us/glossary/xgboost/>.
6. LightGBM – URL: <https://en.wikipedia.org/wiki/LightGBM>.
7. Smart Aspects of CatBoost Algorithm – URL: <https://python.plainenglish.io/smart-aspects-of-catboost-algorithm-2720a6de4da6>.
8. Выбор метрики качества в задачах прогнозирования спроса – 2022. – URL: <https://datanomics.ru/artciles/vybor-metriki-kachestva-v-zadachah-prognozirovaniya-sprosa/>.
9. Число родившихся за год – URL: <https://showdata.rosstat.gov.ru/finder/>.

COMPARATIVE ANALYSIS OF MODELS FOR FORECASTING POPULATION SERIES

A.A. Kondorova¹, K.V. Galanov¹

¹Samara University, Samara

e-mail: kondorovanastya@gmail.com, galanov.kv@ssau.ru

Abstract

This paper presents a comparative analysis of machine learning models for forecasting population time series using official birth statistics in the Russian Federation (Rosstat). Quality assessment was performed using the RMSE metric. It was found that ensemble models, particularly CatBoost, demonstrated the best results, confirming their effectiveness when working with complex data.

Keywords: machine learning, forecasting, population series, MLR, SARIMAX, Prophet, Random Forest, XGBoost, LightGBM, CatBoost.