

МАТЕМАТИЧЕСКИЕ, СТАТИСТИЧЕСКИЕ И ИНСТРУМЕНТАЛЬНЫЕ МЕТОДЫ ЭКОНОМИКИ

Научная статья

УДК 331.101.38:004.89

doi:10.46554/1993-0453-2026-6-260-137-148

Прогнозирование и интерпретация факторов производительности труда с применением ансамблевых методов машинного обучения

Иван Алексеевич Лабуткин

Национальный исследовательский университет ИТМО, Санкт-Петербург, Россия,
i.labutkin@outlook.com

Аннотация. Исследование посвящено применению нелинейных моделей машинного обучения в задаче прогнозирования региональной производительности труда и выявлению факторов, оказывающих наибольшее влияние на данный показатель. В ходе работы были использованы ансамблевые методы, среди которых наилучшие результаты продемонстрировала модель случайного леса. Для учета межрегиональной неоднородности выполнена кластеризация временных рядов производительности труда и введена категориальная метка принадлежности региона к группе, что позволило отразить структурные различия и повысить устойчивость и интерпретируемость прогнозов. Качество прогнозов сравнивалось с результатами линейных моделей, традиционно применяемых в отечественной литературе. Среди этих результатов было зафиксировано устойчивое преимущество нелинейных подходов. Интерпретация итоговой модели случайного леса с помощью метода Mean Decrease in Gini показала, что ключевыми факторами, влияющими на производительность труда, являются темпы роста реальной заработной платы и динамика инвестиций в основной капитал, что согласуется с результатами предыдущих исследований. Применение метода SHAP-values позволило дополнительно оценить направление и степень влияния признаков на прогнозируемый показатель, повысив интерпретируемость и прозрачность модели. Полученные результаты позволили подтвердить возможность перехода от традиционных эконометрических методов к методам машинного обучения при прогнозировании макроэкономических показателей.

Ключевые слова: прогнозирование производительности труда, драйверы производительности труда, случайный лес, градиентный бустинг, интерпретация моделей машинного обучения, кластеризация временных рядов, экономика регионов

Основные положения:

- ◆ применение ансамблевых методов машинного обучения (в частности, случайного леса) обеспечило более точное прогнозирование региональной производительности труда по сравнению с традиционными линейными моделями;
- ◆ ключевыми факторами, оказывающими наибольшее влияние на производительность труда, выступают темпы роста реальной заработной платы и динамика инвестиций в основной капитал;
- ◆ использование методов интерпретации (Mean Decrease in Gini и SHAP-values) повысило прозрачность модели и позволило оценить как важность признаков, так и направление их влияния;
- ◆ результаты подтверждают целесообразность перехода от классических эконометрических подходов к нелинейным методам машинного обучения при анализе и прогнозировании макроэкономических показателей.

Для цитирования: Лабуткин И.А. Прогнозирование и интерпретация факторов производительности труда с применением ансамблевых методов машинного обучения // Вестник Самарского государственного экономического университета. 2026. № 6 (260). С. 137–148. doi:10.46554/1993-0453-2026-6-260-137-148.

MATHEMATICAL, STATISTICAL AND INSTRUMENTAL METHODS IN ECONOMICS

Original article

Forecasting and interpreting factors of labor productivity using ensemble machine learning methods

Ivan A. Labutkin

ITMO University, Saint-Petersburg, Russia, i.labutkin@outlook.com

Abstract. This study explores the application of nonlinear machine learning models to forecasting regional labor productivity and identifying its key determinants. Several ensemble methods were employed, with the random forest model achieving the best predictive performance. To account for interregional heterogeneity, clustering of labor productivity time series was performed and a categorical label indicating each region's cluster membership was introduced, which captured structural differences and enhanced the sustainability and interpretability of the forecasts. The forecasting accuracy was compared with that of traditional linear models commonly used in domestic research, revealing a consistent superiority of nonlinear approaches. Interpretation of the final random forest model using the Mean Decrease in Gini criterion indicated that the most influential factors affecting labor productivity are the growth rate of real wages and the dynamics of investment in fixed capital, which aligns with previous empirical findings. Furthermore, the use of SHAP values enabled a more detailed assessment of the direction and magnitude of each feature's contribution, enhancing the interpretability and transparency of the model. The findings provide evidence supporting the transition from conventional econometric approaches to machine learning methods for forecasting macroeconomic indicators.

Keywords: labor productivity forecasting, labor productivity drivers, random forest, gradient boosting, machine learning model interpretation, time series clustering, regional economy

Highlights:

- ◆ the ensemble machine learning methods (particularly Random Forest) outperformed traditional linear models in forecasting regional labor productivity;
- ◆ the most influential drivers of labor productivity were identified as real wage growth rates and dynamics of fixed capital investment;
- ◆ the model interpretability was enhanced through Mean Decrease in Gini and SHAP values, revealing both feature importance and the direction of their impact;
- ◆ the study supports shifting from conventional econometric approaches to nonlinear machine learning methods for macroeconomic forecasting.

For citation: Labutkin I.A. Forecasting and interpreting factors of labor productivity using ensemble machine learning methods // Vestnik of Samara State University of Economics. 2026. No. 6 (260). Pp. 137–148. (In Russ.). doi:10.46554/1993-0453-2026-6-260-137-148.

Введение

Производительность труда является одним из ключевых показателей, определяющих конкурентоспособность экономики и уровень социально-экономического положения государ-

ства. Так, в Указе Президента РФ «О Стратегии экономической безопасности Российской Федерации на период до 2030 года» повышение производительности труда обозначено в числе приоритетных задач, направленных на обеспе-

чение устойчивого роста реального сектора экономики.

Для увеличения данного макроэкономического показателя необходимо как выявлять факторы, влияющие на производительность труда, так и осуществлять ее прогнозирование в целях выработки и корректировки экономической политики на федеральном и региональном уровнях. Традиционно в российской экономической науке прогнозирование производительности труда осуществлялось с помощью линейных эконометрических моделей. Однако в условиях усложнения социально-экономических процессов возникает потребность в применении более гибких инструментов анализа, улавливающих нелинейные взаимосвязи. Методы машинного обучения, способные выявлять сложные зависимости и работать с данными высокой размерности, представляют собой перспективное направление для прогнозирования макроэкономических индикаторов, в том числе производительности труда.

Целью исследования является построение нелинейных моделей прогнозирования производительности труда субъектов Российской Федерации с использованием методов машинного обучения и сопоставление их прогнозирующей способности с традиционными линейными моделями, а также определение факторов, оказывающих наибольшее влияние на динамику производительности труда.

Обзор литературы и гипотезы. Вопросы прогнозирования производительности труда и выявления факторов, влияющих на ее динамику в российской экономике, широко освещены в отечественных исследованиях. Наиболее распространенным подходом в данных работах является применение моделей линейной регрессии, что демонстрируется в трудах [1–3]. В работе [1] путем анализа матрицы корреляции и *t*-статистик Стьюдента было установлено, что ключевым фактором, определяющим динамику производительности труда, является динамика инвестиций в основной капитал. На основе данного показателя были построены модели полиномиальной регрессии для отдельных федеральных округов. В то же время в ряде исследований отмечается, что решающим фактором роста производительности труда является реальная заработная плата.

Так, в работе [4] на данных российских регионов с 2013 по 2019 г. с использованием модели *rooled*-регрессии и модели со случайными эффектами с учетом автокорреляции было установлено, что именно рост заработной платы оказывает влияние на производительность труда сильнее, чем инвестиции, несмотря на то что только эти переменные оказались значимыми в моделях. Аналогичные результаты были получены в исследовании [5], где была построена модель линейной регрессии.

Высокий уровень взаимосвязи между динамикой производительности труда и реальной заработной платы с точки зрения экономической теории подтвержден в статьях [6; 7]. Кроме того, в работе [8] была применена логистическая регрессия для прогнозирования роста региональной производительности труда. Полученные результаты показали высокую точность классификации направлений изменения показателя, однако данный подход также базируется на линейной модели. Таким образом, в большинстве случаев прогнозирование производительности труда в России осуществляется на основе эконометрических моделей преимущественно линейного типа. В исследованиях отмечается, что наиболее влияющими на производительность труда факторами являются заработная плата и динамика инвестиций в основной капитал.

Между тем в научных исследованиях активно развивается направление, которое базируется на применении методов машинного обучения для прогнозирования макроэкономических показателей. Так, в работе [9] было показано, что использование моделей машинного обучения, в частности градиентного бустинга, позволило повысить точность прогнозирования роста производительности труда в 40 странах ОЭСР на 35–40% по сравнению с традиционными моделями. В ряде исследований [10–12] рассматривались возможности применения методов машинного обучения для прогнозирования инфляции и проводилось их сопоставление с традиционными эконометрическими моделями. В статье [12], где в качестве данных были выбраны поквартальные измерения инфляции, было показано, что все использованные методы машинного обучения

(к-ближайших соседей, метод опорных векторов, деревья решений, случайный лес и гребневая регрессия) обеспечили более высокое качество прогнозирования по сравнению с классическими моделями авторегрессии (AR) и векторной авторегрессии (VAR).

В исследовании [11] прогнозировались темпы инфляции на российских данных 2002–2018 гг. и было установлено, что нелинейные методы машинного обучения (нейронные сети и метод опорных векторов) демонстрируют лучшие результаты, чем авторегрессия и гребневая регрессия, при прогнозировании на горизонте, превышающем один месяц. Анализ результатов данных работ позволяет заключить, что применение методов машинного обучения способно обеспечивать более высокую точность прогнозов макроэкономических показателей по сравнению с традиционными эконометрическими подходами, особенно в случае наличия нелинейных взаимосвязей и увеличения горизонта прогнозирования. В работе [13], посвященной изучению преимуществ методов машинного обучения над эконометрическими при предсказании значений макроэкономических показателей, отмечается, что основными факторами, обеспечивающими лучшее качество прогнозов моделей машинного обучения, являются большая устойчивость к данным высокой размерности, лучший контроль над переобучением, возможность улавливать нелинейные связи, возникающие во времена кризисов, а также устойчивость к выбросам и шуму.

Также в ряде работ применяется глубокое обучение для прогнозирования среднечастотных макроэкономических показателей. В некоторых трудах для прогнозирования макроэкономических показателей использовались глубокие рекуррентные нейронные сети. Для задач, связанных с временными рядами, успешно применяются рекуррентные нейронные сети, их модификации и гибриды [14; 15], однако их использование требует наличия высокочастотных данных, что ограничивает возможности анализа показателей, рассчитываемых ежегодно, таких как производительность труда.

Отдельное внимание в научных исследованиях уделяется интерпретации моделей ма-

шинного обучения. Наиболее распространенные подходы включают методы оценки важности признаков, такие как Mean Decrease in Gini (MDG) и Mean Decrease in Accuracy (MDA) [16]. В MDG оценивается вклад признака в снижение индекса Джини в деревьях решений, тогда как в MDA измеряется падение точности модели при случайном перемешивании значений признака и тем самым выявляется его значимость. Еще одним инструментом глобальной интерпретации являются Partial Dependence Plots (PDP) [17], которые показывают средний эффект изменения одного или двух признаков на прогноз модели, позволяя выявлять нелинейные зависимости. Для анализа отдельных прогнозов используются локальные методы. LIME (Local Interpretable Model-agnostic Explanations) [18] аппроксимирует сложную модель локально более простой, что позволяет понять вклад признаков в конкретное предсказание. Более универсальный подход представляет метод SHAP (SHapley Additive exPlanations) [19], основанный на значениях Шепли, который обеспечивает как локальную, так и агрегированную интерпретацию, позволяя количественно сравнивать влияние факторов.

По итогам обзора литературы были выдвинуты 2 основные гипотезы:

H1: использование нелинейных методов машинного обучения для предсказания динамики производительности труда существенно повышает качество прогноза относительно линейных моделей;

H2: динамика инвестиций в основной капитал и реальная заработная плата являются главными факторами, влияющими на рост индекса производительности труда.

Методы

Описание данных. В качестве исходных данных для построения моделей были взяты показатели, опубликованные на сайте Федеральной службы государственной статистики [20]. В качестве признаков были выбраны временные, которые чаще всего применяются в работах, посвященных прогнозированию индекса производительности труда:

♦ индексы потребительских цен на товары и услуги;

♦ реальная среднемесячная начисленная заработная плата работников (в процентах к предыдущему году);

♦ степень износа основных фондов на конец года по полному кругу организаций (в процентах);

♦ индекс физического объема инвестиций в основной капитал (процент, значение показателя за год);

♦ уровень инновационной активности организаций;

♦ коэффициент миграционного прироста на 10 000 человек населения.

В качестве целевой переменной использовался индекс производительности труда (в процентах к предыдущему году).

С целью повышения обучаемости моделей из всех индексов было вычтено 100, что позволило интерпретировать их как темпы роста или снижения соответствующих показателей.

Кроме того, в набор данных были включены значения признаков целевой переменной с лагом 1, поскольку производительность труда является изменяемым во времени показателем, и ее величина может зависеть от зна-

чений социально-экономических факторов в предшествующие периоды [9].

Прогнозирование проводилось для периода с 2007 по 2023 г., при этом лаговые значения учитывались с 2006 г. Анализ охватывает данные по всем регионам России, за исключением территорий, присоединенных с 2014 г., а также ЯМАО, ХМАО, НАО и Чеченской Республики, что связано с недостатком статистических данных по этим субъектам.

Обогащение данных. В исходном наборе данных не учитывается регион, по которому они были собраны. Однако прямое добавление такой переменной приводит к излишнему усложнению данных с точки зрения применяемых моделей. В связи с этим регионы были объединены по динамике производительности труда на всем рассматриваемом периоде, и в набор данных была добавлена категориальная переменная, отражающая принадлежность субъекта РФ к той или иной группе. Для выявления таких групп был проведен кластерный анализ временных рядов производительности труда методом k-средних. Оптимальное число кластеров определялось методом каменистой

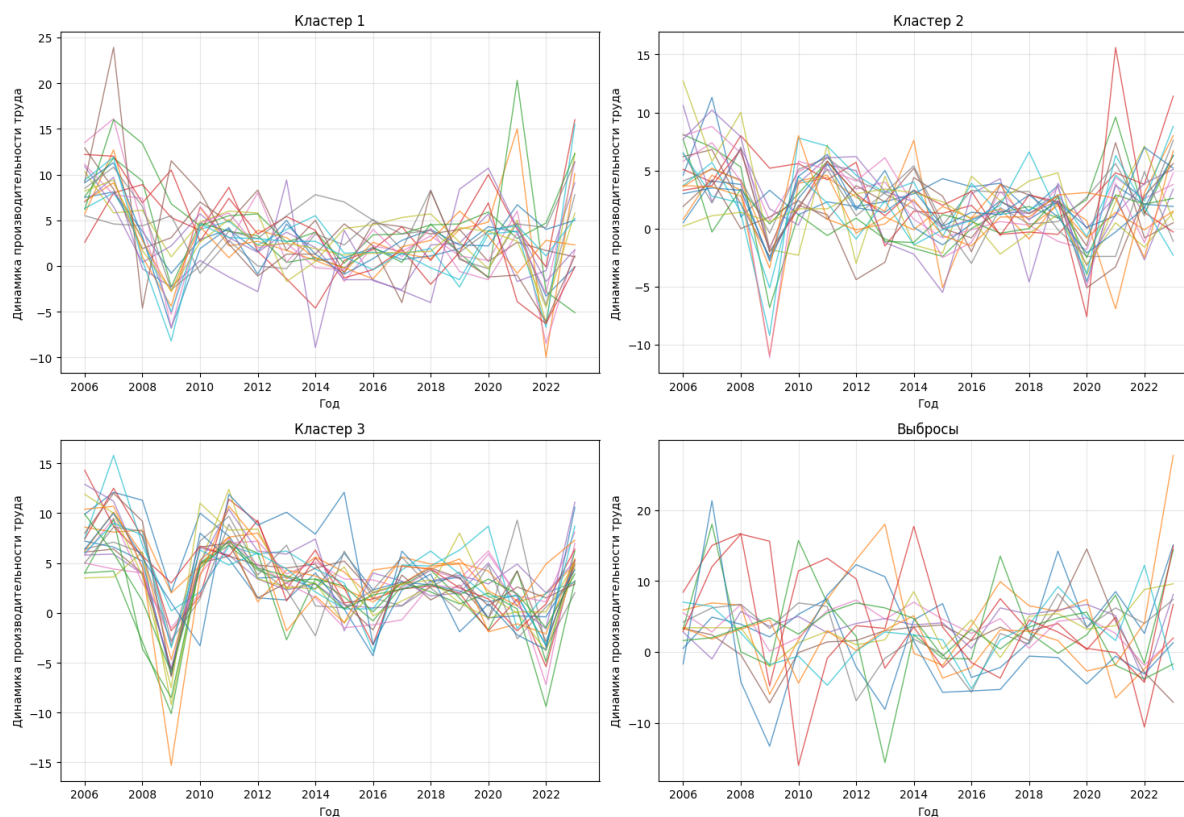


Рис. 1. Графики распределения временных рядов динамик производительностей труда регионов по кластерам

осыпи, а сравнение качества разбиений по группам между разными метриками проводилось путем сравнения индексов Силуэтта и визуальных схожестей составов кластеров. В итоге была использована модель с 8 кластерами и Евклидовой метрикой расстояния. Окончательное разбиение включало в себя 8 кластеров, из которых 3 группы оказались многочисленными, а оставшиеся 5 – малочисленными. Последние были объединены в один кластер с выбросами.

Графики распределений временных рядов по кластерам представлены на рис. 1.

Сформированный набор данных состоял из 1343 наблюдений, представляющих собой показатели производительности труда 79 регионов за 17 лет, и содержал в себе следующие признаки, представленные в табл. 1.

Используемые методы. Методология исследования включает следующие шаги:

♦ построение прогноза для целевой переменной «производительность труда» с помо-

щью ансамблевых методов машинного обучения;

♦ построение интерпретации модели с лучшими метриками – анализ вклада каждой входной переменной с помощью метода SHAP.

В качестве ансамблевых методов машинного обучения были выбраны случайный лес и градиентный бустинг, поскольку они относятся к классу нелинейных алгоритмов, способных уменьшать смещение при сохранении разброса. Также они демонстрируют высокие метрики качества на небольших наборах данных и обладают интерпретируемостью с точки зрения важности признаков.

1. *Случайный лес.* Данный метод представляет собой ансамбль деревьев решений, каждое из которых обучается на bootstrap-подвыборке данных и случайном подмножестве признаков, что снижает риск переобучения модели. Итоговое предсказание в случае регрессии представляет собой среднее арифметическое прогнозов базовых моделей.

Таблица 1

Признаки, используемые внутри набора данных, и способ их расчета

Признак	Как рассчитывался	Тип данных
Темп инфляции за текущий год	Индекс потребительских цен на товары и услуги - 100	Действительное число
Темп инфляции за предыдущий год	Индекс потребительских цен на товары и услуги с лагом 1 - 100	Действительное число
Рост заработной платы за текущий год	Реальная среднемесячная начисленная заработная плата работников - 100	Действительное число
Рост заработной платы за предыдущий год	Реальная среднемесячная начисленная заработная плата работников с лагом 1 - 100	Действительное число
Инновационная активность организаций за текущий год	Уровень инновационной активности организаций	Действительное число
Инновационная активность организаций за предыдущий год	Уровень инновационной активности организаций с лагом 1	Действительное число
Степень износа основных фондов за текущий год	Степень износа основных фондов на конец года по полному кругу организаций	Действительное число
Степень износа основных фондов за предыдущий год	Степень износа основных фондов на конец года по полному кругу организаций с лагом 1	Действительное число
Миграционный прирост за текущий год	Коэффициент миграционного прироста на 10 000 человек населения	Действительное число
Миграционный прирост за предыдущий год	Коэффициент миграционного прироста на 10 000 человек населения с лагом 1	Действительное число
Инвестиции в основной капитал за текущий год	Индекс физического объема инвестиций в основной капитал - 100	Действительное число
Инвестиции в основной капитал за предыдущий год	Индекс физического объема инвестиций в основной капитал с лагом 1 - 100	Действительное число
Индекс производительности труда за предыдущий год	Индекс производительности труда с лагом 1 - 100	Действительное число
Группа региона	Группа региона по итогам кластерного анализа	Категориальная переменная

Реализация алгоритма была использована из библиотеки `sklearn` [21].

2. *Градиентный бустинг*. Данный метод строит последовательный ансамбль решающих деревьев, где каждое последующее дерево минимизирует ошибку предыдущих, аппроксимируя градиенты функции потерь. Итоговое предсказание модели представляет собой взвешенную на шаг обучения сумму прогнозов базовых моделей. Реализация метода была взята из библиотеки `Catboost` [22]. Используемый в ней алгоритм позволяет эффективно обрабатывать категориальные признаки.

Результаты

Подбор оптимальных гиперпараметров обеих моделей был осуществлен с помощью перебора по сетке с 4-блочной кросс-валидацией.

Оптимизируемой метрикой была выбрана средняя абсолютная ошибка (MAE).

Случайный лес. Перебираемые в сетке гиперпараметры и их оптимальные значения представлены в табл. 2.

Градиентный бустинг. Перебираемые в сетке гиперпараметры и их оптимальные значения представлены в табл. 3.

Таблица 2

Перебираемые значения гиперпараметров случайного леса и полученные оптимальные значения

Гиперпараметр	Описание гиперпараметра	Значения гиперпараметров в сетке	Оптимальное значение гиперпараметра
<code>n_estimators</code>	Число деревьев	100, 200, 400, 600, 800, 1000	200
<code>max_depth</code>	Максимальная глубина деревьев	5, 10, 20, 25, 30, нет	Нет
<code>max_features</code>	Максимальное число признаков при обучении признаков	Квадратный корень от числа признаков, двоичный логарифм, отсутствие ограничений	Квадратный корень от числа признаков
<code>bootstrap</code>	Использование подвыборок с возвращением	Да, нет	Нет
<code>min_samples_split</code>	Минимальное число объектов для деления узла дерева	2,5,7,10	10
<code>min_samples_leaf</code>	Минимальное число объектов внутри листа	1,2,4,5	4

Таблица 3

Перебираемые значения гиперпараметров градиентного бустинга и полученные оптимальные значения

Гиперпараметр	Описание гиперпараметра	Значения гиперпараметров в сетке	Оптимальное значение гиперпараметра
<code>learning_rate</code>	Темп обучения	0.01, 0.05, 0.1	0.01
<code>depth</code>	Максимальная глубина деревьев	4, 8, 10, 12	8
<code>l2_leaf_reg</code>	Коэффициент L2-регуляризации	0, 0.5, 1	1
<code>iterations</code>	Число деревьев	500, 800, 1000	1000
<code>grow_policy</code>	Политика роста деревьев	SymmetricTree (Симметрично), Depthwise (В глубину)	Depthwise
<code>min_data_in_leaf</code>	Минимальное число объектов внутри листа	1, 3, 5	1

Таблица 4

Минимальные значения метрик MAE для рассматриваемых методов

Метод	MAE
Случайный лес	2.663
Градиентный бустинг	2.67

Минимальные значения MAE для каждого из методов, полученные в результате поиска по сетке, отображены в табл. 4. Случайный лес показал более высокое качество прогнозирования по сравнению с градиентным бустингом.

Построенная на основе модели случайного леса диаграмма SHAP-values изображена на рис. 2. Данная диаграмма позволяет проинтерпретировать поведение модели и выявить зависимость целевой переменной от значений признаков.

Определены следующие закономерности:

- ◆ высокие темпы роста заработной платы сопровождаются значительным увеличением производительности труда, однако ее низкие значения оказывают резкое негативное влияние на целевую переменную;

- ◆ при больших и средних значениях инвестиций в основной фонд наблюдается положительное влияние на производительность труда, тогда как низкие темпы инвестиций резко снижают показатель;

- ◆ высокий уровень производительности труда за предыдущий год способствует росту

текущих значений, а средние и низкие значения этого показателя относительно слабо влияют на прогноз;

- ◆ экстремальные значения инфляции за предыдущий год негативно влияют на производительность труда, тогда как ее умеренные темпы положительно влияют на целевую переменную.

Диаграмма важности признаков (рис. 3), полученная методом Mean Decrease in Gini, подкрепляет сделанные с помощью SHAP-values выводы.

Анализ диаграммы важности признаков показал, что наибольшее влияние на модель оказывают следующие признаки:

- ◆ рост заработной платы;
- ◆ рост инвестиций в основные фонды;
- ◆ значение производительности труда за предыдущие годы.

Кроме того, существенный вклад (выше 5%) оказывают темпы инфляции за текущий и предыдущий периоды, а также лаговые значения темпов роста заработной платы и инвестиций в основные фонды.

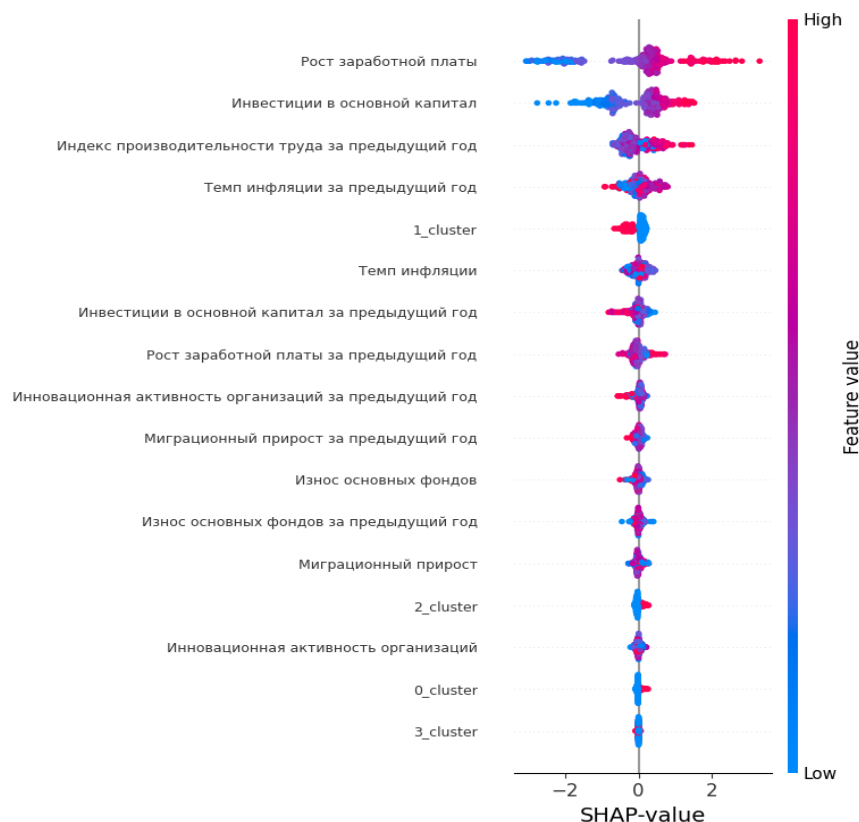


Рис. 2. Сводная диаграмма SHAP-values для модели случайного леса

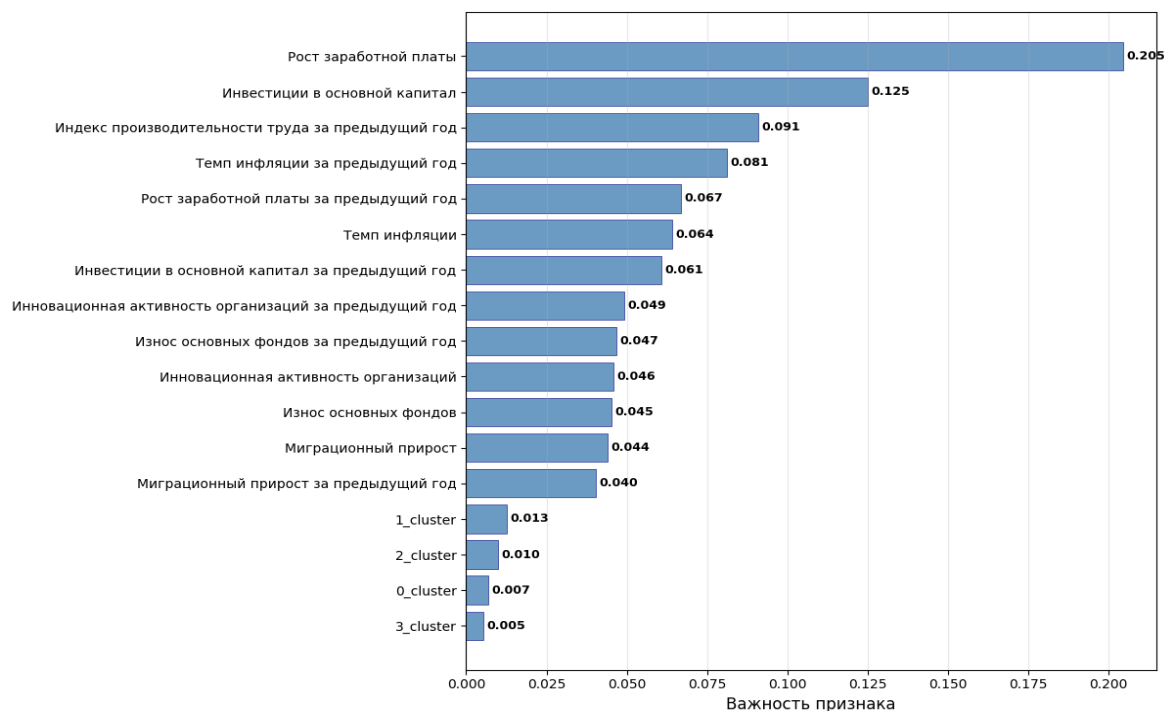


Рис. 3. Диаграмма важностей признаков, рассчитанных с помощью Mean Decrease in Gini

Обсуждение

Для оценки эффекта применения нелинейных методов прогнозирования производительности труда был проведен сравнительный анализ с линейными моделями, в том числе ранее используемыми в соответствующих исследованиях:

- ◆ линейная регрессия;
- ◆ гребневая регрессия;
- ◆ полиномиальная регрессия.

Для подбора оптимальных гиперпараметров линейных моделей была применена сетка поиска, а для оценки качества использовалась 4-блочная кросс-валидация. Все реализации методов были взяты из библиотеки `sklearn` [21].

Значения метрик для всех моделей приведены в табл. 5.

Результаты эксперимента показали, что нелинейные методы – случайный лес и градиентный бустинг – демонстрируют существенно лучшие показатели точности по сравнению с линейными моделями. Так, минимальное значение MAE, полученное для случайного леса, оказалось на 5,2% ниже, чем для гребневой регрессии. Самые низкие результаты относительно остальных моделей показала полиномиальная регрессия, несмотря на то что во мно-

гих статьях применение этого метода позволило получить модели с очень высоким коэффициентом детерминации. Это может быть объяснено переобучением: в ряде предыдущих исследований данный метод применялся к агрегированным данным без использования кросс-валидации, что повышало вероятность завышения оценки качества модели. При более строгой процедуре оценки полиномиальная регрессия не проявила удовлетворительной обобщающей способности.

Интерпретация итоговой модели на основе анализа важностей признаков показала, что ключевыми факторами роста производительности труда являются индекс заработной платы и динамика инвестиций в основной капитал. Вклад этих факторов превышает 10%, а остальные переменные (особенно без учета лагов) оказались менее значимыми, что совпадает с выводами, сделанными в предыдущих отечественных исследованиях, причем важность заработной платы более чем в 1,5 раза превысила значимость инвестиций, что совпадает с выводами ряда теоретических и практических исследований, где именно динамика реальной зарплаты рассматривалась как основной драйвер производительности

Таблица 5

Значения метрик для сравниваемых моделей

Модель	MAE	MSE	R ²
Случайный лес	2.6634	14.0296	0.2898
Градиентный бустинг	2.6703	13.9234	0.2957
Линейная регрессия	2.8287	15.0545	0.2293
Гребневая регрессия	2.8101	14.9706	0.2339
Полиномиальная регрессия	3.1271	19.6909	-0.0023

труда. В то же время лаг целевой переменной также продемонстрировал высокую значимость, заняв третье место по важности среди всех признаков, что соответствует выводу исследования, посвященному странам ОЭСР, где отмечается ключевая роль предшествующих значений производительности.

Применение метода SHAP-values продемонстрировало его преимущества по сравнению с более простыми подходами интерпретации, такими как Mean Decrease in Gini. SHAP позволяет не только оценить и ранжировать важность признаков, но и количественно определить направление и силу их влияния в зависимости от значений переменных. Особенно наглядно это проявилось при анализе лагового признака: высокие значения лага существенно повышали прогноз индекса производительности труда, тогда как средние и малые значения признака более слабо уменьшали выход модели. Соответственно, в случае, когда целевая переменная нелинейно зависит от признаков, применение SHAP существенно расширяет возможности интерпретации и повышает прозрачность модели.

Заключение

В данной работе было исследовано применение ансамблевых методов машинного обучения для прогнозирования индекса производительности труда. Для обучения использовались статистические данные по субъектам РФ за 2006–2023 гг., которые были обогащены метками о принадлежности региона к группе схожести по динамике производительности труда, выделенной с помощью кластерного анализа временных рядов. Для сравнения наряду с методами градиентного бустинга и случайного леса было обучено несколько ли-

нейных алгоритмов. Для всех моделей были подобраны оптимальные гиперпараметры, а оценка качества выполнялась на основе кросс-валидации.

Анализ метрик подтвердил гипотезу H1: нелинейные методы показали метрики на порядок лучше по сравнению с линейными моделями. Наилучший результат продемонстрировал случайный лес, для которого значение метрики MAE составило 2.663.

Интерпретация случайного леса с помощью методов Mean Decrease in Gini и SHAP-values позволила проверить гипотезу H2: количественные оценки важностей показали, что динамика инвестиций в основной капитал и реальная заработная плата являются главными факторами, влияющими на индекс производительности труда. При этом вклад заработной платы оказался существенно выше вклада инвестиций.

Кроме того, использование SHAP-values дало возможность не только ранжировать признаки по их важности, но и оценивать характер и силу их влияния на прогноз в зависимости от их значений, что обеспечило более высокую прозрачность модели по сравнению с традиционными методами интерпретации.

Полученные результаты подтверждают возможность перехода от традиционных эконометрических подходов к методам машинного обучения при прогнозировании макроэкономических показателей. Перспективным направлением дальнейших исследований может стать использование моделей глубокого обучения, которые способны обеспечивать более точные прогнозы, однако для их применения необходимы данные большей частотности, а также более высокого качества и достоверности.

СПИСОК ИСТОЧНИКОВ

1. Шумилина В.Е., Цвиль М.М. Статистическое моделирование и прогнозирование индекса производительности труда в Российской Федерации // Вестник евразийской науки. 2019. Т. 11, № 1. С. 46.
2. Шумилина В.Е., Цвиль М.М. Построение модели регрессии по временным рядам с целью прогнозирования индекса производительности труда в Российской Федерации // Вестник евразийской науки. 2020. Т. 12, № 1. С. 73.
3. Цвиль М.М., Нестерова А.В. Прогнозирование индекса производительности труда по Центральному федеральному округу // Инженерный вестник Дона. 2021. № 3. С. 120–129.
4. Дьячкова А.В., Карасс В.О. Оценка влияния заработной платы на производительность труда в России: эконометрический анализ // Управленческий учет. 2022. № 10-3. С. 710–715. doi:10.25806/uu10-32022710-715.
5. Баканач О.В., Лопухова Я.С. Статистическое моделирование индекса производительности труда в РФ // Наука XXI века: актуальные направления развития. 2020. № 1-1. С. 234–238.
6. Бахтизин А.Р., Сулакшин С.С., Колесник И.Ю. Заработная плата как фактор повышения производительности труда // Контуры глобальных трансформаций: политика, экономика, право. 2009. Т. 2, № 1. С. 79–87.
7. Печура О.В., Польшгалова Н.Ю. Заработная плата как фактор роста производительности труда // Аллея науки. 2020. Т. 1, № 5. С. 168–172.
8. Бурцева Т.А. Прогнозирование роста региональной производительности труда // Russian Journal of Labor Economics. 2023. Т. 10, № 3. doi:10.18334/et.10.3.117464.
9. Towards more timely measures of labour productivity growth / Y. Dorville [et al.]. OECD Publishing, 2025. doi:10.1787/436ecbb5-en.
10. Букина Т.В., Кашин Д.В. Прогнозирование региональной инфляции: эконометрические модели или методы машинного обучения? // Экономический журнал Высшей школы экономики. 2024. Т. 28, № 1. С. 81–107. doi:10.17323/1813-8691-2024-28-1-81-107.
11. Pavlov E. Forecasting inflation in Russia using neural networks // Russian Journal of Money and Finance. 2020. Vol. 79, No. 1. Pp. 57–73. doi:10.31477/rjmf.202001.57.
12. Chakraborty C., Joseph A. Machine learning at central banks. 2017. doi:10.2139/ssrn.3031796.
13. How is machine learning useful for macroeconomic forecasting? / Goulet Coulombe P. [et al.] // Journal of Applied Econometrics. 2022. Vol. 37, No. 5. Pp. 920-964. DOI: 10.1002/jae.2910.
14. Yurtsever M. Unemployment rate forecasting: LSTM-GRU hybrid approach // Journal for Labour Market Research. 2023. Vol. 57, No. 1. P. 18. doi:10.1186/s12651-023-00345-8.
15. Unemployment rate prediction using a hybrid model of recurrent neural networks and genetic algorithms / K. Mero [et al.] // Applied Sciences. 2024. Vol. 14, No. 8. P. 3174. doi:10.3390/app14083174.
16. Breiman L. Random forests // Machine learning. 2001. Vol. 45, No. 1. Pp. 5-32. doi:10.1023/A:1010933404324.
17. Friedman J.H. Greedy function approximation: a gradient boosting machine // Annals of statistics. 2001. Pp. 1189–1232. doi:10.1214/aos/1013203451.
18. Ribeiro M.T., Singh S., Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. Pp. 1135–1144. doi:10.1145/2939672.2939778.
19. Lundberg S.M., Lee S.I. A unified approach to interpreting model predictions // Advances in neural information processing systems. 2017. Vol. 30. doi:10.5555/3295222.3295230.
20. Федеральная служба государственной статистики : официальный сайт. URL: <https://rosstat.gov.ru/> (дата обращения: 05.10.2025).
21. Scikit-learn: machine learning in Python / F. Pedregosa [et al.] // The Journal of machine Learning research. 2011. Vol. 12. Pp. 2825–2830.
22. CatBoost: unbiased boosting with categorical features / L. Prokhorenkova [et al.] // Advances in neural information processing systems. 2018. Vol. 31.

References

1. Shumilina V.E., Tsvil M.M. Statistical modeling and forecasting of the labor productivity index in the Russian Federation // Bulletin of Eurasian Science. 2019. Vol. 11, No. 1. P. 46.
2. Shumilina V.E., Tsvil M.M. Building a regression model for time series to forecast the labor productivity index in the Russian Federation // Bulletin of Eurasian Science. 2020. Vol. 12, No. 1. P. 73.

3. Tsvil M.M., Nesterova A.V. Forecasting the labor productivity index for the Central Federal District // Engineering Bulletin of the Don. 2021. No. 3. Pp. 120–129.
4. Dyachkova A.V., Karass V.O. Assessing the impact of wages on labor productivity in Russia: econometric analysis // Managerial Accounting. 2022. No. 10-3. Pp. 710–715. doi:10.25806/uu10-32022710-715.
5. Bakanach O.V., Lopoukhova Ya.S. Statistical modeling of the labor productivity index in the Russian Federation // Science of the 21st Century: Current Directions of Development. 2020. No. 1-1. Pp. 234–238.
6. Bakhtizin A.R., Sulakshin S.S., Kolesnik I.Yu. Wages as a factor in increasing labor productivity // Contours of Global Transformations: Politics, Economics, Law. 2009. Vol. 2, No. 1. Pp. 79–87.
7. Pechura O.V., Polygalova N.Yu. Wages as a factor of labor productivity growth // Alley of Science. 2020. Vol. 1, No. 5. Pp. 168–172.
8. Burtseva T.A. Forecasting regional labor productivity growth // Russian Journal of Labor Economics. 2023. Vol. 10, No. 3. doi:10.18334/et.10.3.117464.
9. Towards more timely measures of labour productivity growth / Y. Dorville [et al.]. OECD Publishing, 2025. doi:10.1787/436ecbb5-en.
10. Bukina T.V., Kashin D.V. Forecasting regional inflation: econometric models or machine learning methods? // HSE Economic Journal. 2024. Vol. 28, No. 1. Pp. 81–107. doi:10.17323/1813-8691-2024-28-1-81-107.
11. Pavlov E. Forecasting inflation in Russia using neural networks // Russian Journal of Money and Finance. 2020. Vol. 79, No. 1. Pp. 57–73. doi:10.31477/rjmf.202001.57.
12. Chakraborty C., Joseph A. Machine learning at central banks. 2017. doi:10.2139/ssrn.3031796.
13. How is machine learning useful for macroeconomic forecasting? / Goulet Coulombe P. [et al.] // Journal of Applied Econometrics. 2022. Vol. 37, No. 5. Pp. 920-964. DOI: 10.1002/jae.2910.
14. Yurtsever M. Unemployment rate forecasting: LSTM-GRU hybrid approach // Journal for Labour Market Research. 2023. Vol. 57, No. 1. P. 18. doi:10.1186/s12651-023-00345-8.
15. Unemployment rate prediction using a hybrid model of recurrent neural networks and genetic algorithms / K. Mero [et al.]. // Applied Sciences. 2024. Vol. 14, No. 8. P. 3174. doi:10.3390/app14083174.
16. Breiman L. Random forests // Machine learning. 2001. Vol. 45, No. 1. Pp. 5-32. doi:10.1023/A:1010933404324.
17. Friedman J.H. Greedy function approximation: a gradient boosting machine // Annals of statistics. 2001. Pp. 1189–1232. doi:10.1214/aos/1013203451.
18. Ribeiro M.T., Singh S., Guestrin C. "Why should i trust you?" Explaining the predictions of any classifier // Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. Pp. 1135–1144. doi:10.1145/2939672.2939778.
19. Lundberg S.M., Lee S.I. A unified approach to interpreting model predictions // Advances in neural information processing systems. 2017. Vol. 30. doi:10.5555/3295222.3295230.
20. Federal State Statistics Service : official website. URL: <https://rosstat.gov.ru/> (date of access: 05.10.2025).
21. Scikit-learn: machine learning in Python / F. Pedregosa [et al.] // The Journal of machine Learning research. 2011. Vol. 12. Pp. 2825–2830.
22. CatBoost: unbiased boosting with categorical features / L. Prokhorenkova [et al.] // Advances in neural information processing systems. 2018. Vol. 31.

Информация об авторе

И.А. Лабуткин – инженер Национального исследовательского университета ИТМО.

Information about the author

I.A. Labutkin – engineer of the ITMO University.

Статья поступила в редакцию 27.12.2025; одобрена после рецензирования 18.01.2026; принята к публикации 09.06.2026.

The article was submitted 27.12.2025; approved after reviewing 18.01.2026; accepted for publication 09.06.2026.