УДК 004.67

Лысенко Н. А.
студент, Белгородский государственный национальный исследовательский университет,
Россия, г. Белгород
Огородников Л. О.
студент, Белгородский государственный национальный исследовательский университет,
Россия, г. Белгород
Научный руководитель: Путивцева Н. П., к.т.н.
доцент кафедры прикладной информатики и информационных технологий
Белгородский государственный национальный исследовательский университет,
Россия, г. Белгород

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ В ЭКОНОМЕТРИЧЕСКИХ ИССЛЕДОВАНИЯХ ДЛЯ РЕШЕНИЯ ПРОБЛЕМЫ ПРОПУЩЕННЫХ ДАННЫХ

Аннотация: Проблема пропущенных данных является одной из основных в эконометрике в связи с ограничением точности и достоверности результатов исследований. В данной статье рассматривается применение методов машинного обучения для решения проблемы пропущенных данных в эконометрических исследованиях. Акцент делается на эффективности машинного обучения в вопросах восстановления данных, классификации, регрессии и ансамблевых методов в эконометрике.

Ключевые слова: Машинное обучение, Эконометрика, Пропущенные данные, Импутация данных, Классификация, Регрессия, Ансамблевые методы.

Lysenko N. A.
student, Belgorod State National Research University,
Russia, Belgorod
Ogorodnikov L. O.
student, Belgorod State National Research University,
Russia, Belgorod
Research Supervisor: Putivtseva N. P., Ph.D.
Associate Professor at the Department of Applied Informatics and Information Technologies
Belgorod State National Research University,
Russia, Belgorod

APPLICATION OF MACHINE LEARNING IN ECONOMETRIC RESEARCH TO ADDRESS MISSING DATA ISSUES

Abstract: The issue of missing data remains a key challenge in econometrics, constraining the accuracy and reliability of research findings. This article explores the application of machine learning methods to address the problem of missing data in econometric studies. Emphasis is placed on the effectiveness of machine learning in the realms of data recovery, classification, regression, and ensemble methods within econometrics.

Keywords:

Machine Learning, Econometrics, Missing Data, Data Imputation, Classification, Regression, Ensemble Methods.

В условиях стремительного развития технологий и роста объемов данных в экономике, проблема пропущенных данных в эконометрических исследованиях становится более актуальной и сложной. Традиционные методы обработки пропусков не всегда эффективны, и именно в этом контексте машинное обучение предоставляет перспективные решения. Цель

данной статьи заключается в рассмотрении применения методов машинного обучения для решения проблемы пропущенных данных в эконометрике.

Пропущенные данные в эконометрике представляют серьезное препятствие для корректной оценки параметров моделей и влияют на статистическую мощность и достоверность результатов. Традиционные методы, такие как удаление неполных наблюдений, часто нежелательны из-за потери ценной информации и смещения результатов. Именно здесь методы машинного обучения становятся неотъемлемым инструментом для эффективной импутации данных.

Традиционные методы, такие как метод k-ближайших соседей или линейная регрессия, были широко использованы в эконометрике для восстановления пропущенных значений. Однако эти методы могут оказаться недостаточно гибкими для обработки сложных взаимосвязей в экономических данных. В контексте переменных с высокой степенью взаимосвязи и нелинейных зависимостей традиционные методы часто не справляются с задачей восстановления.

Алгоритмы машинного обучения, такие как случайные леса и градиентный бустинг, предоставляют эффективные средства для борьбы с пропущенными данными в эконометрических моделях. Эти методы позволяют учесть сложные структуры данных, выявлять нелинейные взаимосвязи и адаптироваться к изменениям в данных. Процесс обучения моделей на основе имеющихся данных и последующее использование их для предсказания пропущенных значений становится ключевым элементом успешной импутации данных.

Случайные леса, по сути, осуществляют предсказания для объектов на основе меток похожих объектов из обучения. Схожесть объектов при этом тем выше, чем чаще эти объекты оказываются в одном и том же листе дерева. [2]

Рассмотрим задачу регрессии с квадратичной функцией потерь. Пусть $T_n(x)$ — номер листа $n(x)$-го дерева из случайного леса, в который попадает объект x. Ответ объекта x равен среднему ответу по всем объектам обучающей выборки, которые попали в этот лист $T_n(x)$. Это можно записать в виде формулы:

$$b_n(x) = \sum_{i=1}^{l} w_n(x, x_i) y_i \,,$$

где

$$w_n(x, x_i) = \frac{[T_n(x) = T_n(x_i)]}{\sum_{j=1}^{l} [T_n(x) = T_n(x_j)]},$$

N – количество деревьев,

i – счетчик для деревьев,

b – решающее дерево,

x – сгенерированная на основе данных выборка.

Тогда ответ композиции равен:

$$a_n(x) = \frac{1}{N} \sum_{n=1}^{N} \sum_{i=1}^{l} w_n(x, x_i) y_i$$

Ответ случайного леса представляет собой сумму ответов всех объектов обучения с некоторыми весами. Отметим, что номер листа $T_n(x)$, в который попал объект, сам по себе является ценным признаком. Достаточно неплохо работает подход, в котором по выборке обучается композиция из небольшого числа деревьев с помощью случайного леса или градиентного бустинга, а потом к ней добавляются категориальные признаки $T_1(x), \ldots, T_n(x)$. Новые признаки являются результатом нелинейного разбиения пространства и несут в себе информацию о сходстве объектов. [3]

В контексте решения проблемы пропущенных данных в эконометрике, классификация и регрессия находят применение для построения моделей, предсказывающих отсутствующие значения на основе имеющихся данных. Классификация подходит, когда пропущенные значения являются категориальными, а регрессия применяется, когда нужно предсказать непрерывные переменные.

Алгоритмы классификации, такие как Support Vector Machines и Decision Trees, а также алгоритмы регрессии, включая линейную регрессию и градиентный бустинг, обладают способностью улавливать сложные зависимости в данных. Это особенно важно в эконометрике, где переменные могут взаимодействовать многими способами. [2]

Важным этапом при использовании классификации и регрессии для восстановления данных является оптимизация моделей и подбор гиперпараметров. Это обеспечивает наилучшую производительность моделей и предотвращает их переобучение или недообучение.

Ансамблевые методы, такие как Random Forest и Gradient Boosting, могут быть особенно полезными в контексте восстановления пропущенных данных. Их способность объединять прогнозы нескольких моделей повышает устойчивость и качество предсказаний. [2]

Хорошим примером ансамблей считается теорема Кондорсе «о жюри присяжных» (1784). Если каждый член жюри присяжных имеет независимое мнение, и если вероятность правильного решения члена жюри больше 0.5, то тогда вероятность правильного решения присяжных в целом возрастает с увеличением количества членов жюри и стремится к единице. Если же вероятность быть правым у каждого из членов жюри меньше 0.5, то вероятность принятия правильного решения присяжными в целом монотонно уменьшается и стремится к нулю с увеличением количества присяжных. [1]

$$\mu = \sum_{i=m}^{N} C_N^i p^i (1-p)^{N-i}$$

N – количество присяжных,

p – вероятность правильного решения присяжного,

μ – вероятность правильного решения всего жюри,

m — минимальное большинство членов жюри, $m = floor\left(\frac{N}{2}\right) + 1$,

C_N^i— число сочетаний из N по i.

Если $p > 0.5$, то $\mu > p$.

Если $N \to \infty$, то $\mu \to 1$.

Использование алгоритмов классификации и регрессии для восстановления пропущенных данных представляет собой эффективный подход в современной эконометрике. Эти методы не только повышают точность предсказаний, но и обеспечивают более гибкое моделирование сложных взаимосвязей в экономических данных. Однако, необходимость тщательного подбора моделей и параметров подчеркивает важность методологического подхода к данной задаче.

Оценка качества восстановленных данных является ключевым этапом при применении методов машинного обучения в решении проблемы пропущенных данных в эконометрике. Стандартные метрики, такие как Mean Squared Error (MSE) и R-squared, предоставляют количественные оценки точности восстановления. Однако, в эконометрике важно также учитывать особенности моделей и специфику данных. [3]

При оценке качества восстановленных данных в эконометрике необходимо учитывать специфику используемых моделей. Например, при восстановлении данных для временных рядов экономических показателей, важно оценивать не только точность предсказаний, но и устойчивость временных зависимостей.

Для более точной оценки качества восстановленных данных в эконометрике предлагается разработка новых метрик, учитывающих особенности экономических моделей. Например, метрика, уделяющая внимание точности восстановления в периодах экономической нестабильности, может быть более релевантной для определенных задач.

Процесс валидации моделей важен для обеспечения их способности обобщения на новые данные. Кросс-валидация и разделение выборки на обучающую и тестовую позволяют оценить, насколько хорошо модель восстанавливает пропущенные значения на новых данных.

Не менее важным этапом является интерпретация результатов оценки качества восстановленных данных. Анализ ошибок моделей и понимание того, как они могут повлиять на исследовательские выводы, предоставляет дополнительный уровень информации для принятия решений.

Оценка качества восстановленных данных в эконометрике требует комплексного подхода, учитывающего не только стандартные метрики точности, но и специфику экономических моделей. Разработка новых метрик и акцент на статистической значимости позволят более точно определить эффективность применяемых методов в конкретных эконометрических задачах.

В реальных эконометрических исследованиях методы машинного обучения успешно применялись для прогнозирования макроэкономических показателей. Например, отбор переменных для анализа и прогнозирования нестабильности с помощью моделей градиентного бустинга для предсказания ВВП национальной экономики на основе имеющихся данных, включая индексы потребительских цен, инвестиции и торговый баланс. [4] Полученные результаты сравнимы с традиционными эконометрическими моделями, но при этом методы машинного обучения позволяют учесть более сложные зависимости в данных.

В финансовых исследованиях машинное обучение успешно применяется для анализа временных рядов и прогнозирования рыночных трендов. Используются методы классификации, такие как Support Vector Machines, для предсказания направления движения ценных бумаг. [1] Это позволяет трейдерам и инвесторам принимать информированные решения на основе моделей машинного обучения.

В банковской сфере применение методов машинного обучения широко используется для анализа кредитного риска. Модели машинного обучения позволяют учесть более широкий спектр факторов и взаимосвязей, что улучшает точность оценок риска.

Применение методов машинного обучения в эконометрических исследованиях для решения проблемы пропущенных данных является актуальной и перспективной областью исследований. Предложенные методы позволяют эффективно восстанавливать пропущенные значения и улучшать качество моделей. Однако, необходимо учитывать особенности данных и выбирать подходящую модель для каждого конкретного случая. Дальнейшие исследования в этой области помогут разработать новые методы и подходы для решения проблемы пропущенных данных и повысить точность и надежность результатов эконометрических исследований.

Использованные источники:

1. Ануфриева Е.В. Предсказание индекса Мосиржи при помощи метода опорных векторов // Экономические исследования. - 2019. - №4. - С. 34-42.

2. Радченко В. Открытый курс машинного обучения [Электронный ресурс] // Open Data Science (дата публикации 27.03.2017). - URL: https://habr.com/ru/companies/ods/articles/324402/ (дата обращения: 05.01.2024).

3. Хасти Т., Тибришани Р., Фридман Д. Основы Статического обучения: интеллектуальный анализ данных, логический вывод и прогнозирование, 2-е изд.: Пер. с англ. - СПб.: ООО «Диалектика», 2020. - 764 с.

4. Шульгин С.Г. Отбор переменных для анализа и прогнозирования нестабильности с помощью моделей градиентного бустинга // Ежегодник. - Волгоград: Учитель, 2018. - С. 115-153.