УДК 001.817

МЕТОДОЛОГИЯ РАСЧЕТА ВЫСОКОЧАСТОТНЫХ СУБ-ИНДЕКСОВ ИПЦ С ПРИМЕНЕНИЕМ ВЕБ-СКРЕЙПИНГА И КЛАССИФИКАЦИИ ДАННЫХ

Т. Т. Сафиуллин

Белорусский государственный университет,
Минск, Беларусь, SafiullinTT@bsu.by

В статье представлена методика расчета высокочастотных суб-индексов потребительских цен на основе данных интернет-ритейла Беларуси. Описаны разработка и применение ETL-пайплайна для веб-скрейпинга товаров с онлайн-платформ, их очистки и предварительной обработки. Основное внимание уделено задаче автоматической классификации товаров по категориям потребительской корзины НСК РБ с использованием методов машинного обучения. Результаты работы формируют основу для создания оперативной системы мониторинга инфляции.

Ключевые слова: веб-скрейпинг; индекс потребительских цен; машинное обучение; классификация товаров.

METHODOLOGY FOR CALCULATING HIGH-FREQUENCY CPI SUB-INDEXES USING WEB SCRAPING AND DATA CLASSIFICATION

T. T. Safiullin

Belarusian State University,
Minsk, Belarus, SafiullinTT@bsu.by

The abstract should be informative. The article presents a methodology for calculating high-frequency consumer price sub-indexes based on data from online retailers in Belarus. It describes the development and application of an ETL pipeline for web scraping product data from online platforms, including its cleaning and preprocessing. The primary focus is on the task of automatic product classification according to the categories of the consumer basket defined by the National Statistical Committee of Belarus. The results form the basis for creating a real-time inflation monitoring system.

Keywords: web scraping; consumer price index; machine learning; product classification.

1. Введение

Современные экономические исследования все чаще опираются на анализ веб-данных, позволяющий получать актуальную и детализированную информацию о рыночных процессах. Одним из ключевых инструментов сбора таких данных является веб-скрейпинг – автоматизированное

извлечение информации с веб-сайтов. Как отмечают исследователи, этот метод становится все более популярным среди национальных статистических служб для расчета официальных ценовых показателей [1]. Особую ценность представляет возможность ежедневного мониторинга изменений, что позволяет получать высокочастотные данные для оценки воздействия экономической политики и поддержки принятия решений [2].

Актуальность веб-скрейпинга в экономических исследованиях подтверждается его успешным применением в международной практике. Исследования демонстрируют устойчивую связь между онлайн- и оффлайн-ценами [3–5], что делает веб-данные ценным источником для построения суб-индексов потребительских цен (ИПЦ) [6], прогнозирования статистических показателей [7] и выявления изменений инфляционных тенденций [8]. Достоинством суб-индексов является возможность получения оценки инфляционных процессов до появления официальных значений индексов цен, рассчитываемых национальными статистическими органами. Например, в статье [2] показано, как веб-скрейпинг данных крупных интернет-магазинов и торговых сетей помогает выявлять расхождения между официальной и реальной инфляцией в Аргентине. В работе [9] авторы используют скрейпинг для подсчета фактического ИПЦ и анализа ценовой реакции на экономические санкции в России. Эти примеры демонстрируют, что автоматизированный сбор данных с торговых онлайн-платформ обеспечивает более оперативный и точный мониторинг экономических показателей по сравнению с традиционными методами.

В Беларуси, как показано в работе В. И. Малюгина [10], применение веб-скрейпинга и методов наукастинга позволяет осуществлять мониторинг потребительской корзины на основе веб-данных о ценах в реальном времени. Автор демонстрирует, что использование веб-данных в сочетании со статистическими методами и машинным обучением дает возможность не только отслеживать текущую стоимость продовольственной корзины и бюджета прожиточного минимума, но и прогнозировать их динамику для различных категорий населения. Однако ручной сбор данных затруднен из-за большого объема информации и ее постоянного обновления. Основная проблема при использовании веб-скрейпинга заключается в получении неструктурированных данных, требующих сложной обработки перед анализом.

Цель данной работы состоит в представлении результатов выполнения первых этапов вычисления суб-индексов на основе веб данных, включая: автоматизированный сбор и обработка данных о товарах, представленных в торговых сетях, классификация товаров по категориям потребительской корзины в Беларуси.

В рамках исследования решаются следующие задачи:

- разработка парсера для сбора веб-данных с белорусских торговых онлайн-платформ;
- очистка и предварительная обработка неструктурированных данных;
- разметка данных для обучения моделей классификации;
- классификация товаров по категориям Национального статистического комитета Республики Беларусь (НСК РБ) с использованием методов машинного обучения.

2. Методология сбора данных

Начальным этапом исследования является сбор структурированных данных с популярных белорусских онлайн-платформ e-dostavka.by и green-dostavka.by, выбранных благодаря их рыночной значимости и репрезентативному товарному ассортименту. Для реализации процесса сбора данных разработана специализированная ETL-система на языке Python.

При работе со статическим контентом используется стек технологий, включающий библиотеки Requests для HTTP-запросов и BeautifulSoup для парсинга DOM-структуры. BeautifulSoup обеспечивает гибкий поиск элементов через CSS-селекторы, позволяя точно извлекать ключевые атрибуты: наименование товара, цену, категорию и уникальный артикул, служивший первичным ключом для интеграции данных.

Анализ сайта e-dostavka.by показывает необходимость обработки динамически генерируемого контента. Для решения этой задачи применяется библиотека Selenium WebDriver в headless-режиме Chrome, что обеспечило корректное исполнение клиентских скриптов и доступ к динамическим элементам DOM.

В ходе реализации пайплайна решаются несколько критических проблем: битые ссылки и нестабильности DOM-структуры. Для обработки битых ссылок внедряется система валидации URL с экспоненциальным backoff, где невалидные ссылки маркировались специальным флагом. Проблема нестабильности DOM-структуры решается через алгоритм избыточного поиска элементов с использованием комбинации CSS-селекторов. Для товаров со скидками реализуется двухэтапная верификация ценовых данных с исключением некорректных записей из финального набора.

Особое внимание уделяется обеспечению воспроизводимости процесса сбора данных. Реализованная система логирования и контрольных точек позволяет отслеживать прогресс сбора, возобновлять процесс после прерываний и гарантировать целостность данных при сетевых сбоях. В результате получается устойчивый пайплайн, обеспечивающий стабильное поступление актуальных рыночных данных для последующего анализа. Собранный корпус данных включает более 23 000 товарных позиций

с еженедельным обновлением и создает надежную основу для моделирования потребительской корзины.

3. Предварительная обработка и категоризация данных

На этапе первичной обработки данных последовательно выполняются ключевые процедуры очистки и нормализации. Первоначально из выборки исключаются записи с отсутствующими ключевыми признаками (название товара, цена), поскольку их включение может привести к искажению аналитических результатов. Далее устраняются дубликаты, возникающие из-за особенностей категориальной структуры исходных платформ, где один товар может относиться к нескольким пересекающимся категориям. Эта процедура позволяет сформировать уникальный и непротиворечивый набор данных.

Важным этапом становится фильтрация нерелевантных категорий, не соответствующих целям исследования. В частности, исключаются сезонные товары (новогодние украшения, праздничная атрибутика и т.д.), поскольку их включение нарушает репрезентативность данных для анализа устойчивых потребительских тенденций.

Процесс категоризации включает несколько последовательных этапов:

1. автоматическое сопоставление категорий магазинов с официальной классификацией НСК РБ на основе векторных представлений и расчета косинусного сходства;
2. создание словаря соответствий для однозначно идентифицируемых категорий;
3. разработка правила для обработки неоднозначных случаев на основе анализа ключевых слов в названиях товаров.

Результатом предварительной обработки становятся два независимых набора данных:

- green-dostavka.by: 13 096 записей;
- e-dostavka.by: 9 994 записей.

Каждый набор содержит единообразную структуру:

- название товара;
- цена товара;
- исходная категория товара в магазине;
- соответствующая категория НСК РБ.

Анализ распределения категорий (рис. 1) выявляет существенный дисбаланс классов, что требует применения специальных метрик оценки качества на этапе классификации. В исследовании используются F1-мера (гармоническое среднее precision и recall) и ROC-AUC (площадь под характеристической кривой), которые демонстрируют устойчивость к

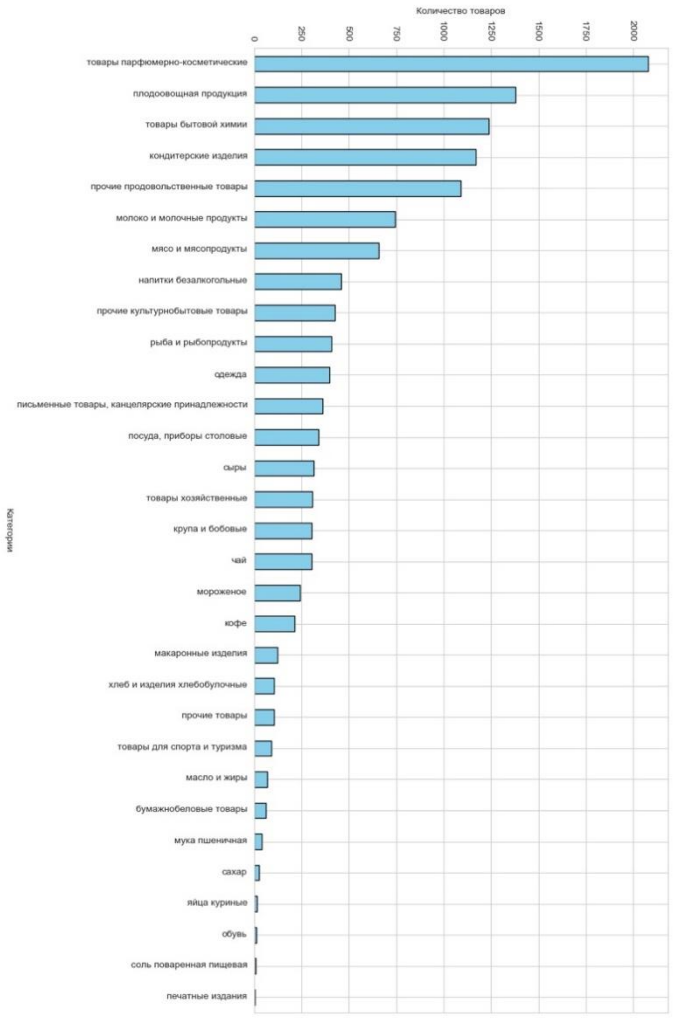

Рис. 1. Распределение категорий НСК РБ

дисбалансу классов в отличие от стандартной метрики accuracy. Эти метрики позволяют получить объективную оценку эффективности модели, минимизируя влияние доминирующих категорий на итоговые показатели.

4. Экспериментальное исследование эффективности алгоритмов классификации товаров

В рамках исследования эффективности различных подходов к классификации товаров был проведен комплексный анализ на данных белорусских онлайн-ритейлеров. Первоначальный эксперимент использовал данные платформы green-dostavka.by, где в качестве входных признаков применялось комбинированное текстовое представление, объединяющее название товара и его магазинную категорию. Этот признак подвергался TF-IDF векторизации перед разделением на стандартные обучающую и тестовую выборки в соотношении 80 на 20.

Тщательная оптимизация гиперпараметров методом решетчатого поиска проводилась для пяти различных алгоритмов машинного обучения: логистическая регрессия, случайный лес, градиентный бустинг, SVM с RBF-ядром и ансамблевая модель, включающая SVM и XGBoost. Сравнительный анализ показал, что все модели продемонстрировали исключительно высокое качество классификации по метрике ROC-AUC (рис. 2), превышающее 0,99, что свидетельствует об отличной разделяющей способности всех рассмотренных алгоритмов для данной задачи.

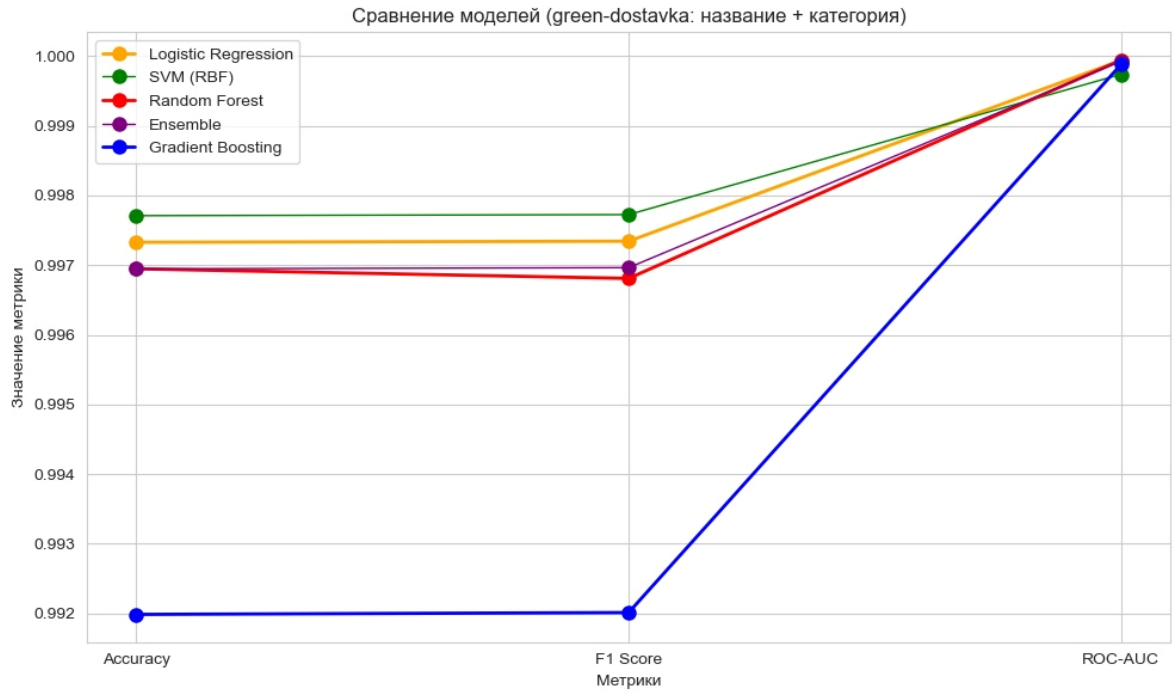

Рис. 2. График сравнения моделей, обученных на названиях и категориях

Наиболее сбалансированные результаты по комплексной метрике F1 и accuracy продемонстрировали SVM, логистическая регрессия и ансамблевая модель. Примечательно, что градиентный бустинг, несмотря на высокий ROC-AUC, показал относительно более низкие значения F1-меры, что может объясняться его повышенной чувствительностью к дисбалансу классов в обучающей выборке. Это наблюдение подчеркивает важность выбора соответствующих метрик оценки в зависимости от специфики задачи.

Дальнейшее исследование было сосредоточено на разработке универсального классификатора, использующего только названия товаров без привязки к конкретной категориальной структуре магазина. Такой подход показал несколько сниженные, но все равно весьма высокие показатели качества (F1 > 0,85), при этом обеспечивая существенно более широкую применимость решения. В этом сценарии SVM и ансамблевая модель

вновь подтвердили свою эффективность, тогда как случайный лес показал более скромные результаты, особенно по F1-мере (рис. 3).

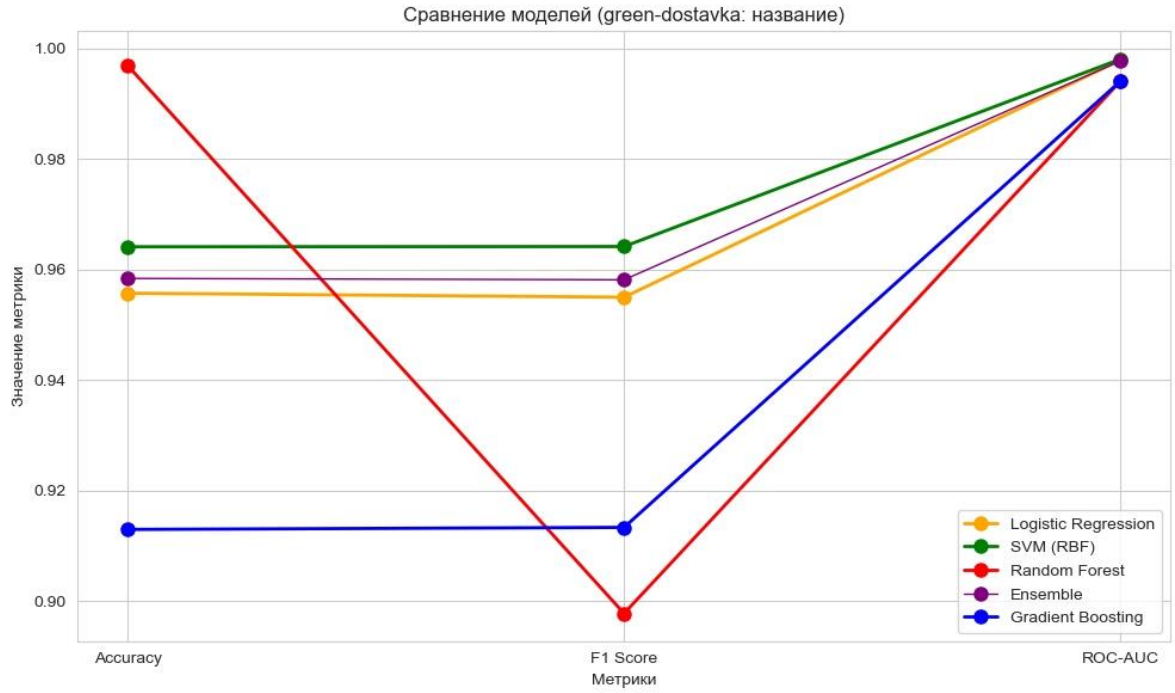

Рис. 3. График сравнения моделей обученных на названиях

Критически важным этапом стала валидация лучших моделей на независимых данных с платформы e-dostavka.by. Устойчивость показателей качества (ROC-AUC > 0,99) для всех протестированных моделей свидетельствует о хорошей обобщающей способности разработанных решений и их применимости для различных онлайн-ритейлеров (рис. 4). Особенно впечатляющим оказалось сохранение высокого качества при переходе к универсальной классификации только по названиям товаров, что открывает перспективы создания кросс-платформенного решения.

5. Заключение

Проведенное исследование позволяет сделать следующие ключевые выводы:

- веб-скрейпинг в сочетании с методами машинного обучения доказал свою эффективность для автоматической классификации товаров потребительской корзины в Беларуси;
- наилучшие результаты показали SVM и ансамблевые модели, продемонстрировавшие высокую точность и устойчивость к дисбалансу классов;

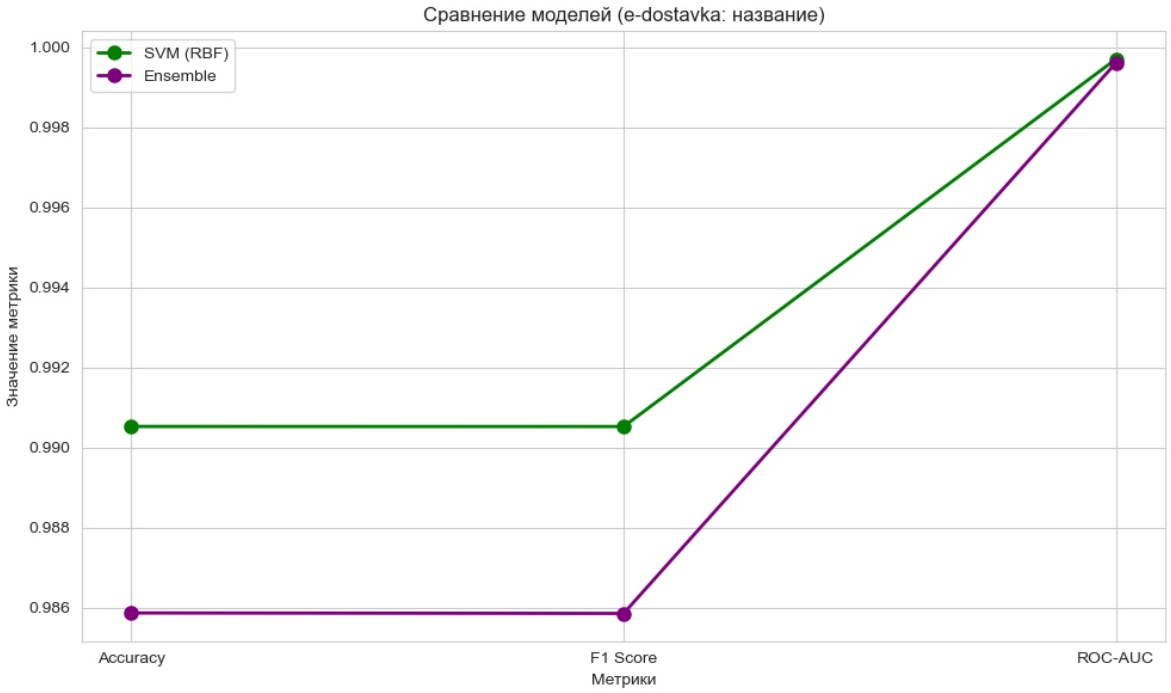

Рис. 4. Проверка лучших моделей на новом наборе данных

- универсальные классификаторы, работающие только с названиями товаров, обеспечивают кросс-платформенную применимость при сохранении высокого качества классификации.

На основе полученных результатов можно сформулировать следующие практические рекомендации:

- для максимально точной классификации в рамках одной платформы рекомендуется использовать SVM с комбинированными признаками (название + категория);
- для кросс-платформенных решений оптимальным выбором являются ансамблевые модели, работающие только с названиями товаров;
- особое внимание при внедрении следует уделить обработке дисбаланса классов и регулярному обновлению обучающих данных.

Важным следующим этапом данного исследования станет расчет индекса потребительских цен на основе собранных и классифицированных веб-данных. Это позволит:

- создать оперативную систему мониторинга инфляционных процессов;
- обеспечить альтернативный источник для верификации официальной статистики;
- реализовать методологию наукастинга для краткосрочного прогнозирования ценовой динамики.

Полученные результаты имеют важное значение для совершенствования систем мониторинга потребительской корзины и могут быть использованы государственными органами статистики, аналитическими центрами и ритейл-компаниями для повышения качества экономического анализа и прогнозирования.

Библиографические ссылки

1. *Eurostat*. Practical Guidelines on Web Scraping for the HICP. 2020. URL: https://ec.europa.eu/eurostat/documents/272892/12032198/Guidelines-web-scraping-HICP-11-2020.pdf/ (date of access: 09.09.2025).

2. *Cavallo A.* Online and official price indexes: Measuring Argentina's inflation // Journal of Monetary Economics. 2013. Vol. 60, iss. 2. P. 152–165.

3. *Cavallo A.* Are online and offline prices similar? Evidence from large multi-channel retailers // American Economic Review. 2017. Vol. 107, iss. 1. P. 283–303.

4. *Cavallo A.* Scraped data and sticky prices // Review of Economics and Statistics. 2018. Vol. 100, iss. 1. P. 105–119.

5. E-commerce and price setting: evidence from Europe / G. Strasser [et al.] // Occasional Paper Series. No. 320. European Central Bank, 2023. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4514760 (date of access: 09.09.2025).

6. *Harchaoui T. M., Janssen R. V.* How can big data enhance the timeliness of official statistics? // International Journal of Forecasting. 2018. Vol. 34, iss. 2. P. 225–234.

7. *Aparicio D., Bertolotto M. I.* Forecasting inflation with online prices // International Journal of Forecasting. 2020. Vol. 36, iss. 2. P. 232–247.

8. *Cavallo A., García Zavaleta G.* Detecting structural breaks in inflation trends: A high-frequency approach / HBS Working Paper, Harvard Business School, 2023.

9. *Benchimol J., Palumbo L.*, Sanctions and Russian Online Prices // Journal of Economic Behavior & Organization. 2024. Vol. 225. P. 483–521.

10. *Малюгин В. И., Якубович А. В.* Анализ и прогнозирование стоимости потребительской корзины в режиме реального времени // Экономика. Моделирование. Прогнозирование. 2020. Вып. 14. С. 235–241.