МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»**

Кафедра математического и компьютерного моделирования

Создание системы персонализированных рекомендаций в интернет - магазине с использованием машинного обучения

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 5 курса 561 группы

направление 09.03.03 — Прикладная информатика

механико-математического факультета

Болдырева Ивана Алексеевича

Научный руководитель
доцент, к.ф.-м.н., доцент Е.Ю. Крылова

Зав. кафедрой
зав. каф., д.ф.-м.н., доцент Ю.А. Блинков

Саратов 2025

Введение. Современные интернет-магазины всё чаще прибегают к использованию интеллектуальных технологий для удержания пользователей, повышения конверсии и увеличения среднего чека. Одним из ключевых инструментов в этом процессе стали рекомендательные системы — программные решения, направленные на автоматическое предложение релевантного контента или товаров на основе анализа поведения пользователей, их предпочтений и исторических данных.

Рекомендательные системы прочно вошли в технологический стек крупных IT-компаний: Amazon, Netflix, YouTube, Ozon, Яндекс.Маркет и других. Они позволяют не только увеличивать продажи за счёт точного предложения интересных товаров, но и существенно улучшать пользовательский опыт, повышать лояльность и вовлечённость аудитории.

Учитывая постоянный рост количества товаров и клиентов в электронной коммерции, становится очевидной необходимость в построении масштабируемых, интерпретируемых и легко встраиваемых моделей, способных учитывать как поведенческие, так и контентные признаки. Данная работа направлена на создание персонализированной рекомендательной системы для интернет-магазина с использованием современных подходов машинного обучения, с учётом требований к производительности, точности и реальной интеграции.

Цель: построение и внедрение гибридной рекомендательной системы в интернет-магазин, обеспечивающей персонализированную выдачу товаров пользователям на основе анализа их поведения и характеристик объектов.

Для достижения цели были поставлены следующие задачи:

- Изучить предметную область рекомендательных систем, их классификацию, архитектуры и метрики качества.
- Провести анализ и предобработку данных, предоставленных в рамках конкурса H&M Personalized Fashion Recommendations.
- Разработать двухэтапную модель рекомендаций:
    1. первый этап — генерация кандидатов с использованием модели имплицитной коллаборативной фильтрации (ALS);
    2. второй этап — ранжирование кандидатов с помощью градиентного бустинга на деревьях (CatBoost).

- Реализовать серверный компонент, обеспечивающий предоставление рекомендаций через REST API в режиме реального времени.
- Разработать клиентское приложение для визуализации выдачи и демонстрации персонализированного подхода.
- Контейнеризировать архитектуру и развернуть комплексную систему с помощью Docker и Docker Compose.
- Оценить качество рекомендательной системы по метрикам Recall@k и MAP@k.

Практическая значимость:

- разработанная система может быть внедрена в любой онлайн-магазин с минимальными изменениями в инфраструктуре;
- архитектура предусматривает масштабируемость, поддержку новых моделей и удобную отладку;
- в рамках работы была выполнена полная производственная интеграция: от сбора данных до визуального интерфейса и деплоя;
- демонстрационное приложение позволяет моделировать пользовательское поведение и отслеживать изменение рекомендаций.

Обзор используемых методов. Рекомендательные системы условно классифицируются на три подхода:

1. Коллаборативная фильтрация — строится на взаимодействиях пользователей с товарами.
2. Контентная фильтрация — использует характеристики объектов и профиля пользователя.
3. Гибридные модели — объединяют оба подхода для повышения качества рекомендаций.

В рамках данной работы реализована двухуровневая гибридная модель:

- ALS (Alternating Least Squares) — метод факторизации матрицы взаимодействий пользователя и товара, подходящий для неявных данных (например, клики, просмотры).
- CatBoost — алгоритм градиентного бустинга на деревьях решений с хорошей работой на категориальных признаках.

Валидация качества системы проводилась с использованием метрик:

- Recall@k — показатель полноты, определяющий долю релевантных объектов в топ-k.
- MAP@k — показатель усредненной точности с учетом порядка ранжирования элементов.

Предобработка и анализ данных. Для обучения рекомендательной модели использовался открытый набор данных H&M Personalized Fashion Recommendations, размещённый на платформе Kaggle. Этот датасет представляет собой реальную выборку транзакций крупного европейского интернет-магазина одежды и содержит обширную информацию о пользователях, товарах и истории их взаимодействия. В частности, в него входят следующие таблицы:

- customers - содержит информацию о зарегистрированных пользователях, включая пол, возраст, принадлежность к клубу лояльности и географическое положение;
- articles — каталог товаров, представленных в магазине. Включает уникальные идентификаторы товаров, категории, секции, цвет, индекс стиля, а также другие мета-признаки;
- transactions_train — журнал покупок, в котором отражены все покупки пользователей за последние месяцы с точностью до даты, артикула и цены.

Общий объём данных оказался значительным:

- более 1,4 млн уникальных пользователей;
- свыше 110 тыс. различных товаров;
- несколько десятков миллионов транзакций.

Однако, для эффективного обучения модели и получения устойчивых результатов, была проведена тщательная фильтрация, очистка и нормализация данных. Были выполнены следующие ключевые этапы предобработки:

- Удаление неактивных пользователей. Пользователи с крайне малой активностью были исключены из обучающей выборки. Это позволило сосредоточиться на наиболее ценной группе пользователей — тех, чья история содержит достаточно информации для построения устойчивых поведенческих профилей.

- Исключение непопулярных товаров. Эти позиции, как правило, не дают устойчивой статистики и могут вносить шум в обучение.
- Обработка и нормализация категориальных признаков. В таблице articles присутствует множество категориальных и индексных признаков, таких как:
  - индекс отдела (department_no);
  - секция (section_name);
  - группа (product_group_name);
  - код цвета (colour_group_code);
  - сезон, и др.
- Разделение данных по времени. Учитывая временную природу транзакций, была применена time-based сегментация:
  - тренировочная выборка — данные за более ранние недели;
  - валидационная выборка — последующие недели;
  - тестовая выборка — последние транзакции (эмулируется реальный прогноз).

В результате предварительной обработки:

- число пользователей было сокращено с 1.4 млн до примерно 11 тысяч активных;
- количество товаров — с 110 тысяч до около 21 тысячи уникальных артикулов.

Такое сужение выборки позволило:

- существенно снизить размерность латентных матриц, используемых в ALS;
- ускорить обучение моделей;
- повысить устойчивость к переобучению.

Подготовленные данные стали основой для построения рекомендательной системы, адаптированной под реальные пользовательские сценарии с учётом как массовых, так и индивидуальных особенностей поведения.

Построение и обучение моделей. Процесс обучения рекомендательной системы в данной работе основан на двухэтапной архитектуре: генерация кандидатов и их последующее ранжирование. Такой подход позволяет объединить сильные стороны коллаборативной и контентной фильтрации, а

также достичь хорошего баланса между точностью и масштабируемостью системы.

Первый этап — генерация кандидатов (ALS). Для генерации начального набора потенциально интересных товаров используется метод Alternating Least Squares (ALS) из библиотеки implicit.

Целевая метрика при обучении ALS — Recall@10 на валидационной выборке.

После обучения модель может для каждого пользователя предсказать top-N товаров, которые с наибольшей вероятностью могут его заинтересовать. Однако этот список не учитывает категориальные и контентные признаки, такие как группа товара, возраст покупателя и др., что и компенсирует второй этап.

Второй этап — ранжирование (CatBoost). После генерации top-N кандидатов они передаются на вход модели CatBoost для более точного ранжирования с учётом дополнительных признаков. Это позволяет повысить качество рекомендаций за счёт контекстной информации.

Процесс подготовки данных:

- Генерация кандидатов для каждого пользователя.
- Присвоение меток.
- Объединение с признаками пользователя и товара, включая:
    1. возраст, клубный статус, индекс, категорию;
    2. частоту покупки товаров определённого типа;
    3. товарные группы, секции, индексные признаки и т.д.

Метрики:

1. Основные метрики качества — Recall@10 и Mean Average Precision@10 (MAP@10).
2. MAP позволяет оценить порядок релевантных товаров: чем выше в списке находится нужный товар, тем лучше.

Преимущества второго этапа:

1. Учитываются индивидуальные характеристики пользователя.
2. Возможность внедрения бизнес-правил (например, запрет на определённые группы товаров).
3. Повышается точность предсказаний, особенно по метрике MAP.

Интеграция и логгирование моделей. Обе модели — ALS и CatBoost — логируются и версионируются с использованием MLflow:

- сохраняются параметры, метрики, веса модели;
- каждая модель получает уникальный идентификатор и может быть запрошена по API;
- используется MLflow Model Registry для организации доступа и обновления моделей.

Это позволяет:

- отслеживать качество моделей во времени;
- быстро переключаться между версиями;
- автоматизировать процесс обновления моделей в production.

Архитектура, инфраструктура и реализация системы. Рекомендательная система построена в виде набора микросервисов, каждый из которых реализует строго определённую функциональность. Такой подход обеспечивает модульность, масштабируемость и упрощает тестирование и деплой.

Вся система условно делится на три уровня:

1. Уровень моделей (машинное обучение):
    - MLflow — сервер логирования и версификации моделей. Обеспечивает:
        - сохранение метрик, гиперпараметров, артефактов моделей;
        - регистрацию версий моделей ALS и CatBoost;
        - централизованный доступ к последним продакшен-версиям.
    - ModelService — REST-сервис, реализующий логику двухэтапной модели:
        - генерация кандидатов с использованием ALS;
        - ранжирование с помощью CatBoost;
        - взаимодействие с MLflow Registry для загрузки последних моделей.
2. Сервисный уровень (бизнес-логика и API):
    - Backend-сервис (на FastAPI):
        - принимает запросы от клиентского приложения;
        - вызывает ModelService для получения рекомендаций;

- обращается к базе PostgreSQL за пользовательскими/товарными данными;
- загружает изображения из MinIO (S3);
- агрегирует всё в единый JSON-ответ.

3. Клиентский уровень (представление и взаимодействие):
   - Клиентское приложение:
     - отображает каталог товаров и блок персональных рекомендаций;
     - позволяет симулировать покупки и видеть изменение выдачи.
4. Инфраструктура и развёртывание:
   - MinIO (объектное хранилище):
     - хранение изображений товаров;
     - хранение артефактов моделей.
   - PostgreSQL (реляционная БД):
     - хранение структурированных данных: пользователи, товары, транзакции.
   - Docker и Docker Compose:
     - каждый компонент работает в собственном контейнере;
     - автоматический запуск всех микросервисов командой docker compose up;
     - независимость компонентов и простота масштабирования.

Архитектура backend - сервиса выглядит в соответствии с рисунком 1.

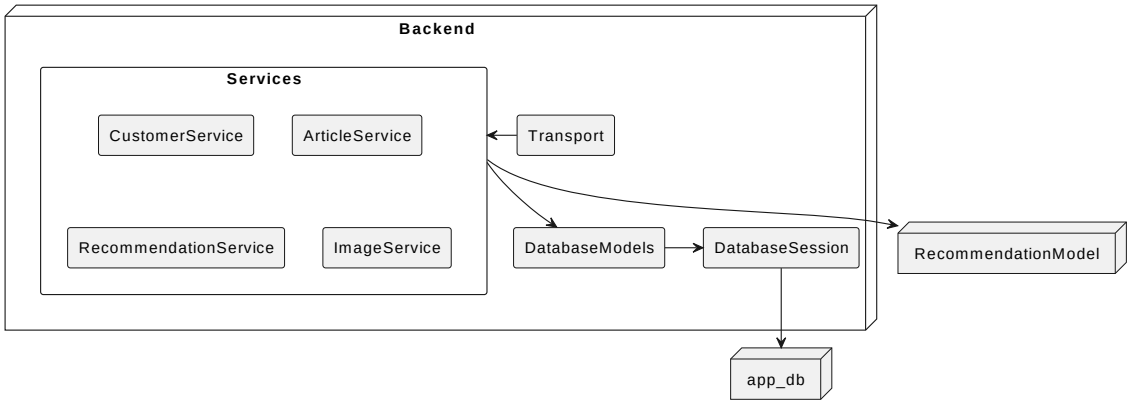

Рисунок 1 — Архитектура backend - сервиса

Полная архитектура системы выглядит в соответствии с рисунком 2.

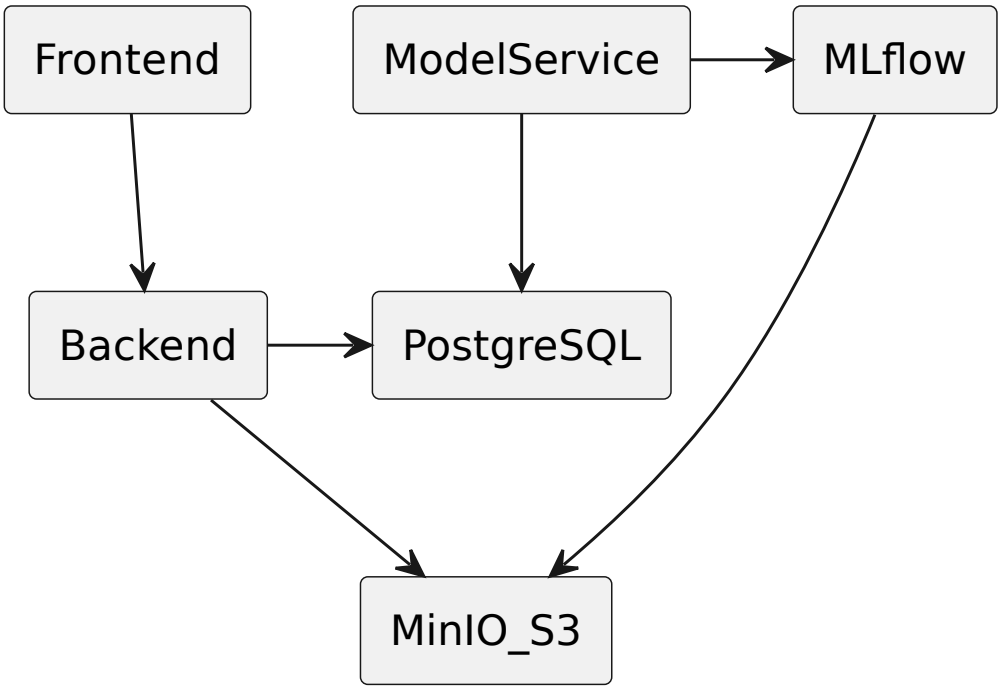

Рисунок 2 — Архитектура системы

Графический интерфейс клиентского приложения выглядит в соответствии с рисунком 3.

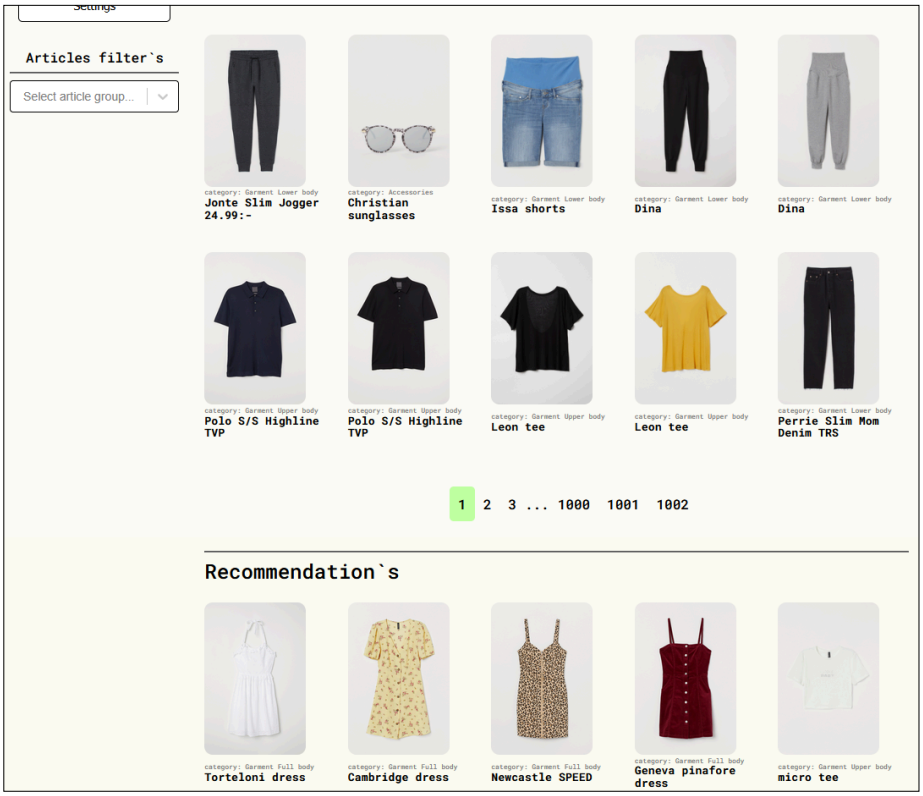

Рисунок 3 — Пользовательский интерфейс

Заключение. В рамках работы была успешно реализована рекомендательная система, охватывающая весь жизненный цикл:

- подготовка и очистка данных;
- обучение двухэтапной гибридной модели;
- построение и деплой REST API;
- создание клиентского интерфейса;
- логирование и контроль качества моделей.

Прототип рекомендательной системы, реализованный в рамках данной работы, может быть непосредственно использован в образовательных, медиаплатформах, e-commerce и других сферах, где требуется персонализированный подход. Архитектура допускает расширение с учётом новых бизнес-задач, включая real-time потоковые рекомендации, персонализированную e-mail-рассылку, сегментацию пользователей, баннерную рекламу и многое другое.