Нейросетевые подходы к разработке рекомендаций по повышению продаж в товарных корзинах

(Neural network to up-sell recommendations on shopping carts)

Тип проекта: **исследовательский**

1 План работы над ВКР

- 1 декабря: формирование плана работы и обзор литературы.
- 20 января: разработка baseline-решений (TopPopular, TopPersonal, EASE, iALS, TIFU-KNN).
- 10 марта: основная работа по ВКР (двухуровневая архитектура с бустингом, модели SASRec, BERT4Rec, ReCANet, GPT, RAG).
- 10 мая: доработка ВКР.
- 20 мая – 10 июня: защита ВКР.

2 Обзор литературы

1. https://arxiv.org/abs/2401.16433 – рассмотрена модель Neural Pattern Associator (NPA) – модель для задачи рекомендаций внутри корзины (within-basket recommendation), предсказывающая товары, которые пользователь может добавить к уже имеющимся в корзине. **Преимущества**: специализированный подход для задачи дополнения корзины, учитывает контекст текущих товаров. Может выявлять сложные нелинейные зависимости между товарами. **Недостатки**: ограниченная применимость только к задаче within-basket рекомендаций, требует достаточного количества данных о совместных покупках.

2. https://arxiv.org/abs/2308.01308 – рассмотрена модель BTBR (Bi-directional Transformer Basket Recommendation) – модель на основе двунаправленных трансформеров для предсказания следующей корзины покупок на основе истории предыдущих корзин пользователя. **Преимущества**: механизм внимания позволяет находить долгосрочные зависимости, двунаправленность улучшает понимание контекста. Трансформеры показывают хорошие результаты на последовательных данных. **Недостатки**: высокая вычислительная сложность трансформеров, требовательность к объему данных для обучения. Модель может быть избыточной для простых сценариев покупок.

3. https://arxiv.org/html/2407.21191v2, https://arxiv.org/abs/2307.00457 – рассмотрен генеративный подход к рекомендациям, использующий языковые модели и трансформеры для формулирования задачи рекомендаций как задачи генерации последовательностей. **Преимущества**: единая архитектура для различных задач рекомендаций, возможность использования предобученных языковых моделей, интерпретируемость через генерацию объяснений. **Недостатки**: очень высокие требования к вычислительным ресурсам, необходимость больших объемов данных, сложность в оптимизации и контроле качества генерации. Может генерировать нерелевантные рекомендации.

4. https://www.ijcai.org/proceedings/2019/0389.pdf – рассмотрена модель Beacon (Basket Sequence Correlation Network) – модель для предсказания следующей корзины, учитывающая корреляции между последовательностями корзин (correlation-sensitive next basket recommendation). **Преимущества**: явное моделирование корреляций между товарами внутри корзин и между различными корзинами в последовательности. Использует внимание и свертки для захвата паттернов на разных уровнях (товар-товар, корзина-корзина). **Недостатки**: ограниченная гибкость архитектуры, может уступать более современным трансформер-подходам. Модель 2019 года, могут быть более свежие улучшения.

5. https://arxiv.org/pdf/1905.03375 – модель EASE – простая и эффективная модель для коллаборативной фильтрации на основе линейных автоэнкодеров, оптимизированная для разреженных данных. **Преимущества**: исключительная простота реализации и скорость работы, замкнутое решение без итеративной оптимизации,

отличное качество для бейзлайна. Легко интерпретируется и масштабируется. **Недостатки**: ограниченная выразительность из-за линейности, не учитывает временную динамику и последовательности, не использует дополнительные признаки товаров или пользователей. Чисто коллаборативный подход.

3 Особенности моего решения

1. **Использование свежего кросс-доменного датасета**

   В работе используется датасет T-ECD (T-Bank E-Commerce Dataset) – крупномасштабный датасет кросс-доменных данных российского ритейлера (Т-банк), включающий данные из marketplace, retail и offers доменов. Датасет содержит информацию о транзакциях, корзинах, товарах, брендах и отзывах пользователей.

   - Публикация на Habr: habr.com
   - Датасет на Hugging Face: huggingface.co/datasets/t-tech/T-ECD

2. **Комплексный бенчмаркинг моделей**

   В рамках исследования реализованы и сравнены различные подходы к задаче up-sell рекомендаций:

   - *Baseline-модели*: TopPopular, TopPersonal, EASE, iALS, TIFU-KNN
   - *SOTA-модели*: SASRec, BERT4Rec, ReCANet, GPT-based подходы
   - *Basket-специфичные модели*: NPA, BTBR, Beacon, GenRec

   Для каждой модели рассчитаны стандартные метрики качества рекомендаций: NDCG@k, Precision@k, Recall@k, что позволяет провести объективное сравнение подходов.

3. **Двухуровневая архитектура с бустингом**

   Предложена новая двухуровневая архитектура для задачи up-sell рекомендаций:

   - *Первый уровень (Candidate Generation)*: быстрая генерация множества потенциальных кандидатов из всего каталога товаров с использованием эффективных моделей (EASE, ALS)

- *Второй уровень (Ranking)*: точное ранжирование отобранных кандидатов с использованием сложных нейросетевых моделей (трансформеры)
- Интеграция бустинговых моделей (LightGBM, CatBoost) для финального ранжирования с учетом множества признаков

Такой подход позволяет сбалансировать качество рекомендаций и скорость инференса, что критично для production-систем.

4. **RAG-архитектура для up-sell рекомендаций**

   Разработана и реализована инновационная RAG (Retrieval-Augmented Generation) архитектура, адаптированная для задачи up-sell рекомендаций на корзинах:

   - *Retrieval-компонент*: извлечение релевантных товаров и контекстной информации (описания, категории, характеристики) на основе текущего содержимого корзины
   - *Generation-компонент*: использование языковых моделей для генерации персонализированных рекомендаций с объяснениями
   - Интеграция мультимодальной информации: текстовые описания товаров, категории, бренды, цены

   Данный подход позволяет не только генерировать точные рекомендации, но и предоставлять интерпретируемые объяснения для пользователей.

5. **Учет кросс-доменной специфики**

   Особое внимание уделяется анализу различий в поведении пользователей и эффективности моделей в различных доменах (marketplace vs retail vs offers), что позволяет разработать более гибкие и адаптивные решения.

6. **Практическая применимость**

   Все разработанные решения ориентированы на возможность внедрения в production-системы.