NEXT BASKET PREDICTION

Финальный проект SKILL FACTORY

Цели проекта

Основная цель проекта - предсказание следующей покупки клиента. Для бизнеса это новые продажи.

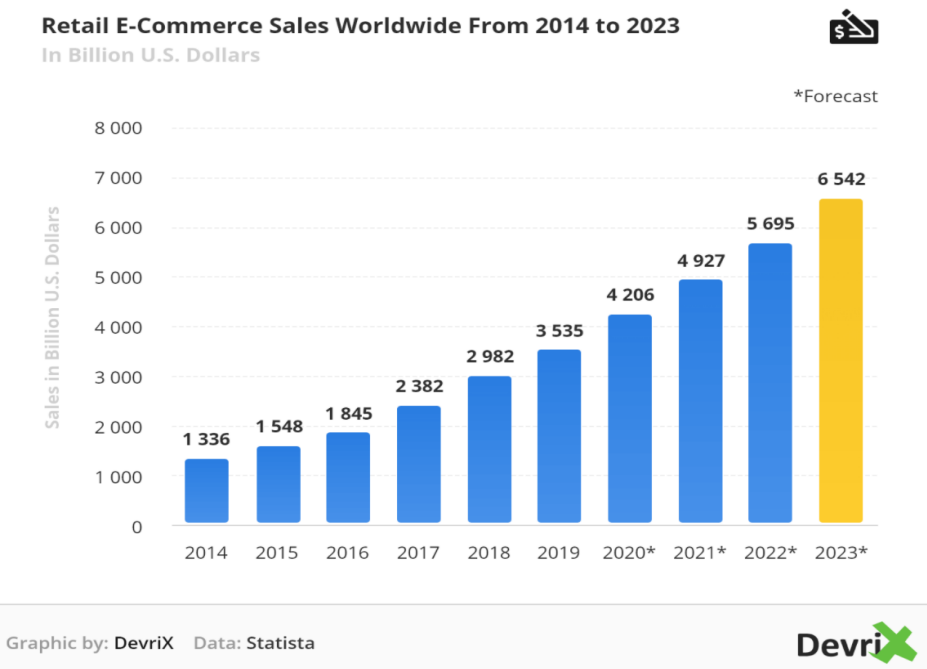
За основу берется реальная двухгодовая история клиентов в компании Ситилинк, по одной из когорт.

В данных содержится информация по заказам, товарам, стоимости и количеству, с делением на категории и брэнды. А так же в каких магазинах и регионах была осуществлена продажа.

SKILLFACTORY + СИТИЛИНК ЭЛЕКТРОННЫЙ ДИСКАУНТЕР =

Проблематика вопроса NBP

Эксперты в 2013 году прогнозировали что машинное обучение вскоре взорвет рынок и следующая покупка будет предсказываться максимально точно, электронная коммерция навсегда изменится. Прошло 7 лет, прогноз не сбылся, несмотря на трехкратный рост в E-Commerce.



Модели машинного обучения не могут прогнозировать достаточно точно, но все же шагнули далеко вперед. Амазон запустил эксперимент и доставляет товар до клиента, исходя из построенной модели и, если клиент согласен, то он оплачивает товар.

Выбор модели

Для предсказания NBP в основном используется две модели.

- Классификация – обычно обучаются оценивать одну выходную переменную. В этом случае можно спрогнозировать, будет ли в следующей корзине конкретный товар или нет. Недостатком является то, что модель строится для каждого продукта, что непрактично в плане ресурса т.к. сейчас продаются сотни тысяч наименований товаров.

- Рекомендательная система – применима в сценариях, когда многие пользователи взаимодействуют со многими элементами. Применяется не только в ритейле, но и в музыкальных стриминговых сервисах (Spotify, Yandex музыка), видеохостингах (Youtube), потоковое видео (Netflix).

Explicit vs Implicit Feedback

Explicit

В контексте рекомендательных систем явная обратная связь - это прямые и количественные данные, полученные от пользователей. Например, интернет магазины разрешают пользователям оценивать приобретенные товары по шкале от 1 до 10. Так же лайки и дизлайки. Однако проблема с явной обратной связью в том, что они редки.

Implicit

Неявная обратная связь собирается косвенно, в результате взаимодействия с пользователем. Например видео, которые вы смотрите на YouTube, используются в качестве неявной обратной связи для составления рекомендаций для вас, даже если вы не оцениваете видео явно. Другой пример неявной обратной связи включает в себя товары, которые вы просматривали в интернет магазине, которая используется, чтобы предложить вам другие похожие товары.

Метрика оценки качества модели

Важно! В метрике **map@k** мы сами определяем К элементов. Результат метрики должен быть максимальным, но это не 1, как например в **Precision.** Мы отталкиваемся от бейзлайна - минимальный результат, который мы получаем на первой модели. Обычно это самые популярные продукты. У меня был ALS из библиотеки Implicite. Дальше мы улучшаем модель по направлению качество/скорость.

Для того чтобы лучше понять рассмотрим с азов.

- **Precision at K (p@K)** – базовая метрика, точность на К элементах.

Алгоритм выдает оценки релевантности для каждого элемента.

Далее выбираем К элементов с наибольшей релевантностью.

- **Average precision at K (ap@K)** – это сумма p@k по индексам k от 1 до К только для релевантных элементов, деленому на К. Так, если из трех элементов мы релевантным оказался только находящийся на последнем месте, то:

  ap@3=1/3(0+0+1/3) ~0,11

  если угадали лишь тот, что был на первом месте, то:

  ap@3=1/3(1/3+0+0) ~0,33

- **Mean average precision at K (map@K)** — Берем среднее от ap@K, посчитанного для всех объектов (пользователей или поисковых запросов).

Поехали!

```
df = pd.read_csv
```

```
sep=';', engine= 'python'
```

	dtypes	nunique	isna
Клиент_Но	object	104837	0
Документ_Но	object	1832900	0
Дата_Создания	object	699	0
Товар_Но	object	79122	0
Наименование_Товара	object	76999	0
Бренд	object	939	0
Merdis_ПодГруппа	object	768	0
Merdis_Группа	object	67	0
Рег_Офис	object	90	0
РО_Отдел	object	189	0
Оборот	object	46003	0
Количество	object	99	0

```
df.shape
```

```
(2904805, 12)
```

Рекомендации для пользователей которые ранее не покупали

```
recommendations[5:100:20]
```

```
[['I0001384', '479889'],
 ['I0003236', '1071476'],
 ['I0005183', '596785'],
 ['I0005183', '1023977'],
 ['I0006351', '720592']]
```

Рекомендации для пользователей которые ранее покупали

```
reality.sample(5)
```

	Клиент_Но	Товар_Но	is_buyed
1335792	IC699567	475269	1
2309770	IE274850	1121277	1
2635663	IX144568	360859	1
2479827	IE692949	583251	1
521283	I1631145	1412080	1

Используемые ресурсы

SkillFactory
ML-9 Рекомендательные системы в МО
PROD – 1,2,3,4,5

https://stackoverflow.com

Evaluation Metrics for Recommender Systems
https://towardsdatascience.com/evaluation-metrics-for-recommender-systems-df56c6611093#:~:text=Mean%20Average%20Precision%20at%20K,insights%20into%20a%20model's%20performance

How to create a production-ready Recommender System
https://towardsdatascience.com/how-to-create-a-production-ready-recommender-system-3c932752f8ea

Метрики качества ранжирования
https://habr.com/ru/company/econtenta/blog/303458/

Deep Learning based Recommender Systems
https://towardsdatascience.com/deep-learning-based-recommender-systems-3d120201db7e

Building a Collaborative Filtering Recommender System with ClickStream Data
https://towardsdatascience.com/building-a-collaborative-filtering-recommender-system-with-clickstream-data-dffc86c8c65