

# Предсказание вероятности покупки товара, показанного клиенту рекомендательной системой

Работу выполнили: Парфенов П. А.

Тимофеева А. А.

Куратор практики: Кислинский В. Г.

# Цель и задачи

## Цель:

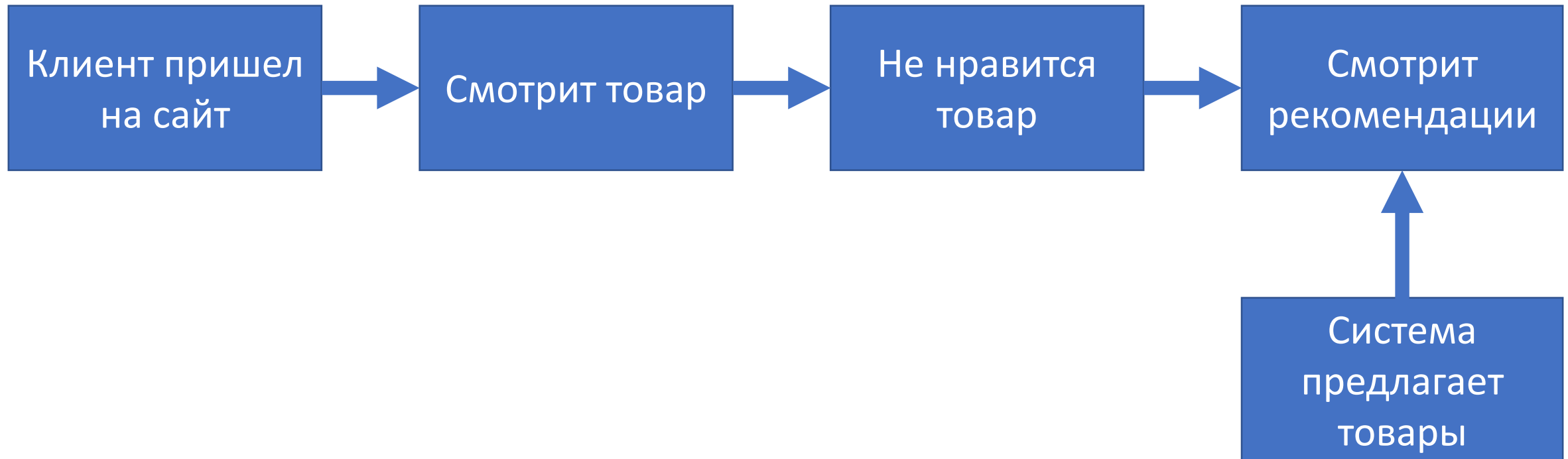
Улучшение существующей рекомендательной системы с помощью подсчета вероятности покупки товаров, предложенных системой.

## Задачи:

- Разработка алгоритма предсказания вероятности покупки;
- Анализ методов машинного обучения;
- Реализация дополнительных метрик для анализа результатов.

# Задача рекомендательной системы:

Увеличить продажи с помощью ранжирования рекомендаций



# Постановка задачи

$$X = \{c, i, r\}_{k=1}^N$$

$c$  — клиент;

$i$  — товар, который смотрит клиент;

$r$  — рекомендованный товар.

Задача ранжирования:

$$(c, i) \rightarrow r$$

$$r = \{r_1 \succcurlyeq r_2 \succcurlyeq r_3 \succcurlyeq \dots \succcurlyeq r_n\}$$

# Исходные данные

- Действия пользователей за август
- Данные о сессиях
- Текстовые описания товаров
- Целевая таблица

# Обучающая выборка

clientid	itemid	jointitemid	label
7833842	31499843	138176581	1
19548158	147389610	148381589	0
32943407	6261257	4490956	0
10185243	148455169	148455173	0
30552232	152440009	152440052	0

clientid – клиент;

itemid – товар, который смотрит клиент;

jointitemid – рекомендованный товар;

label – реакция пользователя

(1-добавил в корзину, 0 - просмотрел товар).

# Признаки

$$f(c, i, r) = \textit{feature}$$

- Признаки популярности и новизны товаров;
- Коллаборативная фильтрация;
- Признаки, основанные на пользовательских сессиях;
- Схожесть по текстовому описанию.

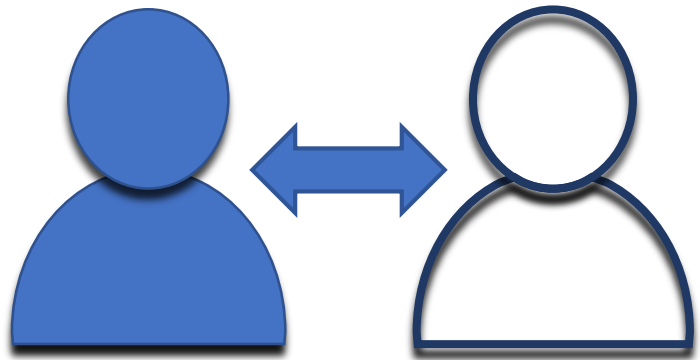
# Признаки популярности и новизны товаров

- Популярность товара по просмотрам/добавлению в корзину;
- Новизна - дата первого просмотра товара;
- Количество просмотров/добавлений в корзину в последний день;
- Спрос на товар в последние 7 дней наблюдений.

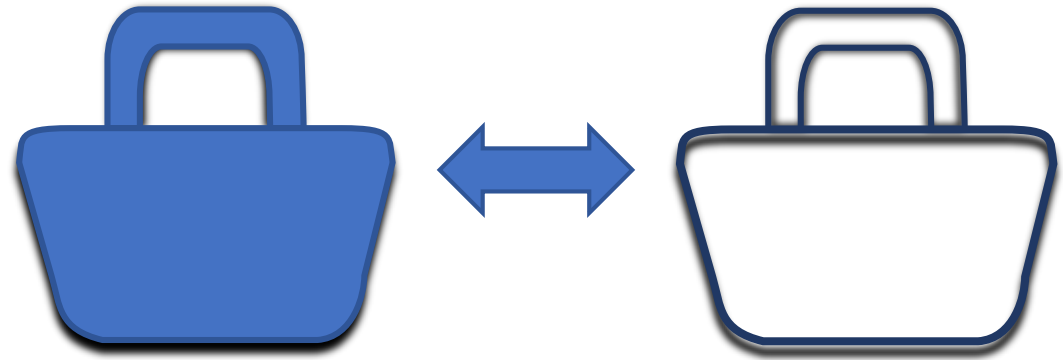


# Коллаборативная фильтрация

User-Based подход



Item-Based подход



# User-Based подход

Взаимодействие пользователя с товаром

R =

clientid	itemid				
	0	1	2	3	4
0	1	0	0	1	0
1	0	1	0	1	1
2	0	1	0	1	0
3	1	1	0	0	1

Похожесть клиентов

$S =$

		clientid			
		0	1	2	3
clientid	0	1	0.4	0.5	0.4
	1	0.4	1	0.8	0.33
	2	0.5	0.8	1	0.4
	3	0.2	0.33	0.4	1

Требуется найти:

$$f(cr) = f(c_0 \cdot r_0) = \langle c_0, r_0 \rangle$$

$$s_{ij} = \text{sim}(\text{client}_i, \text{client}_j)$$

# Признаки, основанные на пользовательских сессиях

## Пример пользовательской сессии

Cart Add:



[iRobot Roomba 895](#)  
[робот-пылесос](#)

Product View:



[Philips FC9174/02](#)  
[пылесос](#)



[Philips FC8671/01](#)  
[PowerPro Active, Red пылесос](#)



[Philips PowerPro Expert FC9734/01, Purple пылесос](#)



[Philips FC8021/03](#)  
[мешок для сбора пыли](#)



[Робот-пылесос iRobot Roomba 616](#)



[iRobot Roomba 895](#)  
[робот-пылесос](#)

## Пример матрицы сессии

		sessionid			
itemid		0	1	2	3
	0	1	0	1	0
	1	1	0	0	0
	2	0	1	1	0
	3	1	1	1	1
	4	0	0	1	0

# Пример расчета схожести товаров

Найдем похожесть товаров:

Телефон + наушники

$$1) \text{sim} = T_0 \cdot T_1 = 1 + 1 = 2$$

$$2) \text{cossim} = \frac{T_0 \cdot T_1}{\|T_0\| \cdot \|T_1\|} = \frac{2}{\sqrt{3} \cdot \sqrt{3}} = 0.667$$

Телефон + свечи

$$1) \text{sim} = T_0 \cdot T_2 = 1_0 + 1_2 = 2$$

$$2) \text{cossim} = \frac{T_0 \cdot T_2}{\|T_0\| \cdot \|T_2\|} = \frac{2}{\sqrt{3} \cdot \sqrt{4}} = 0.577$$

		sessionid					
		itemid	0	1	2	3	4
$T_0 \rightarrow$		1	0	1	1	0	
$T_1 \rightarrow$		1	1	1	0	0	
$T_2 \rightarrow$		1	1	1	0	1	

# Основы nlp для текста

- Токенизация
- Регулярные выражения
- Лемматизация текста
- Удаление стоп-слов
- Расчет TF-IDF

# Расчет TF-IDF

$$Tf - idf(t, d, D) = tf(t, d) \times \log\left(\frac{D}{df_t}\right)$$

$tf(t, d)$  – частота слова  $t$  в документе  $d$ ;

$df$  – количество документов, содержащих слово  $t$ ;

$D$  – общее количество документов.

# Обучение модели



CatBoost

# Используемые метрики

## Mean average precision

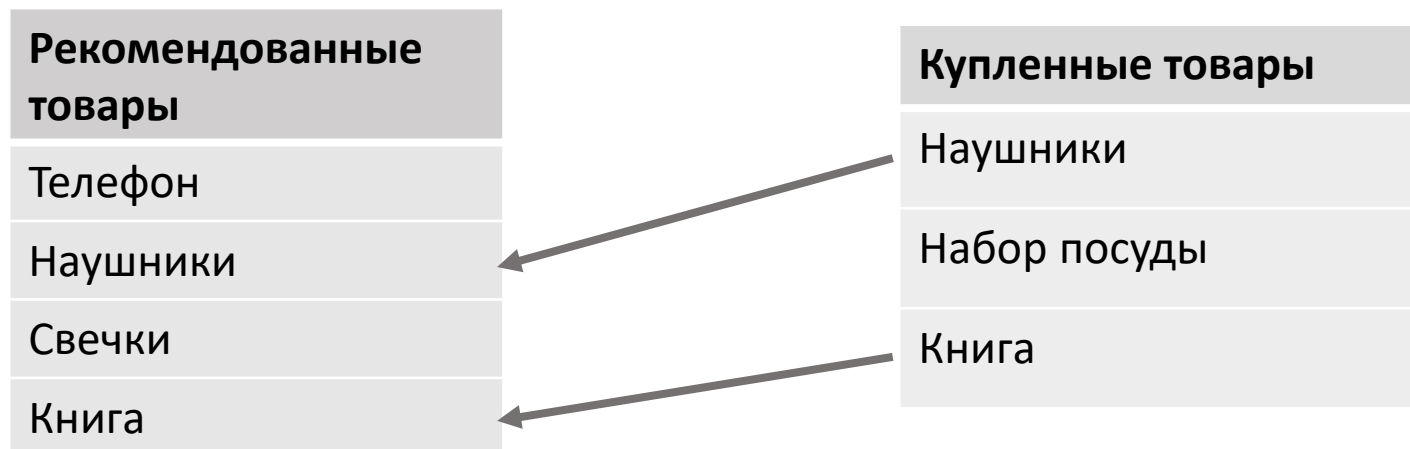
- *Precision at K*  
$$p@k = \frac{\sum_{k=1}^K r^{true}(\pi^{-1}(k))}{K} = \frac{\text{купленное из рекомендованного}}{K}.$$
- *Average precision at K*  
$$ap@k = \frac{1}{K} \sum_{k=1}^K r^{true}(\pi^{-1}(k)) \cdot p@k.$$
- *Mean average precision at K*  
$$map@k = \frac{1}{N} \sum_{j=1}^N ap@K_j.$$



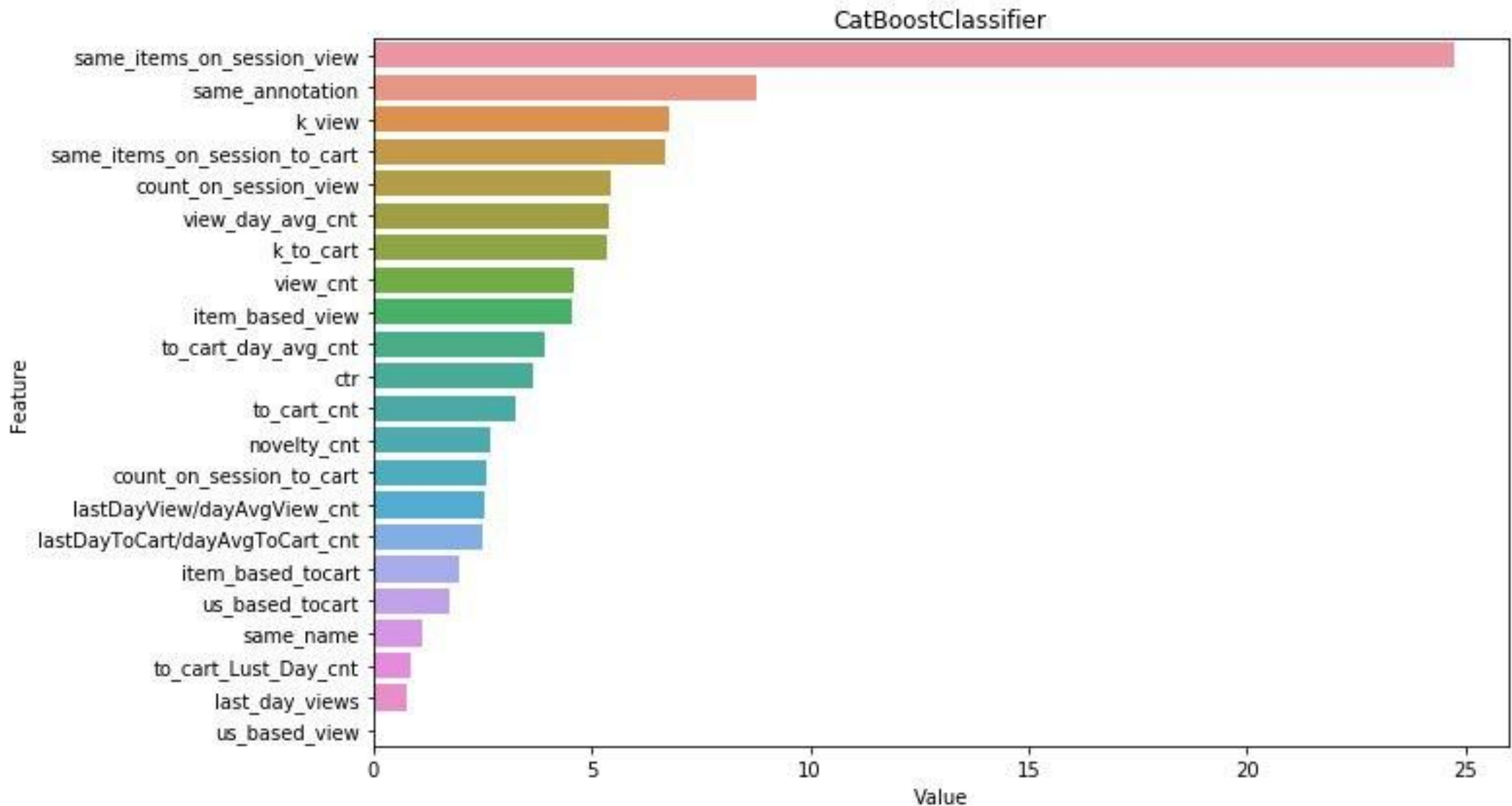
# Используемые метрики

## Recall

$$Recall@k = \frac{\text{рекомендованные } k \text{ товаров, которые релевантны}}{\text{общее количество релевантных товаров}}$$



# Важность признаков



# Анализ результатов. Сравнение с Baseline

Метрики	Baseline RandomForest	CatboostClassifier
AUC	0.55339	0.6345
Map@3	0.11874	0.1368
Recall@3	0.47102	0.5372

# Выводы

- Мы разработали и реализовали алгоритм предсказания вероятности покупки;
- Провели анализ методов машинного обучения;
- Реализовали дополнительные метрики для анализа результатов
- Обучили модель и сравнили результаты