НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Факультет компьютерных наук,
Прикладная математика и информатика.

ЧАТ БОТ РЕКОМЕНДУЕТ

Попов Илья
Никифоров Алексей
Кудрявцева Софья

Руководитель: Ляпина Светлана Юрьевна

Москва, 2020

ПОСТАНОВКА ЗАДАЧИ

Цель работы – Создание алгоритма персонализированных рекомендаций подарков для пользователей, учитывая вводимое пользователем описание человека, контекста подарка и желательную стоимость товара.

Задачи работы:

1. Поиск, сбор и подготовка данных;
2. Разработка алгоритма подбора подарка;
3. Разработка инфраструктуры;

ПОДГОТОВКА ДАННЫХ

	23 февраля	8 марта	IT	age	man	woman	Бокс	Видеоигры	...	91076	91078	91082	91112	91113
0	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	...	0	0	0	0	0
1	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	...	0	0	0	0	0
2	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	...	0	0	0	0	0
3	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	...	0	0	0	0	0
4	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	...	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16072	1.0	0.0	0.0	109.0	0.0	1.0	0.0	0.0	...	0	0	0	0	0
16073	1.0	0.0	0.0	110.0	1.0	0.0	0.0	0.0	...	0	0	0	0	0
16074	1.0	0.0	0.0	120.0	0.0	0.0	0.0	0.0	...	0	0	0	0	0
16075	1.0	0.0	0.0	120.0	0.0	0.0	0.0	0.0	...	0	0	0	0	0
16076	1.0	0.0	0.0	120.0	0.0	1.0	0.0	0.0	...	0	0	0	0	0

Признаки модели:

- Хобби – категориальный признак
- Пол – категориальный признак
- Возраст – Численный признак
- Повод для подарка – категориальный признак

Категориальные признаки человека и контекста закодированы One Hot encoding

Целевая переменная:

- Категория товара для подарка, закодированная One Hot encoding, одному человеку сопоставлено несколько категорий товаров

ОБУЧАЮЩАЯ ВЫБОРКА

Google Forms

Создан опрос в Google forms с полями: “пол”, “возраст”, “хобби”, “повод” и “социальная связь”. Опрос прошло 180 человек, однако часть анкет была отсеяна.

- 114 строк
- Вся необходимая информация о клиенте имеется
- Известно только название товара, категорию необходимо восстанавливать вручную

Mywishlist

Mywishlist - это сервис, в котором пользователи могут составлять списки желаемых подарков. Пользователи указывают свой пол и возраст, а товары объединены в категории.

- 8500 строк
- Нет информации о хобби и поводе для подарка
- Для каждого товара известна категория

КОРРЕКТИРОВКА ВЫБОРКИ

Дополнение выборки:

- Добавление хобби для данных из Mywishlist на основе категории
- Замена названий товаров категориями для данных из опроса в Google Forms

Масштабирование выборки:

- Добавление шума в возраст
- Добавление случайного повода для подарка

Объединение пользователей:

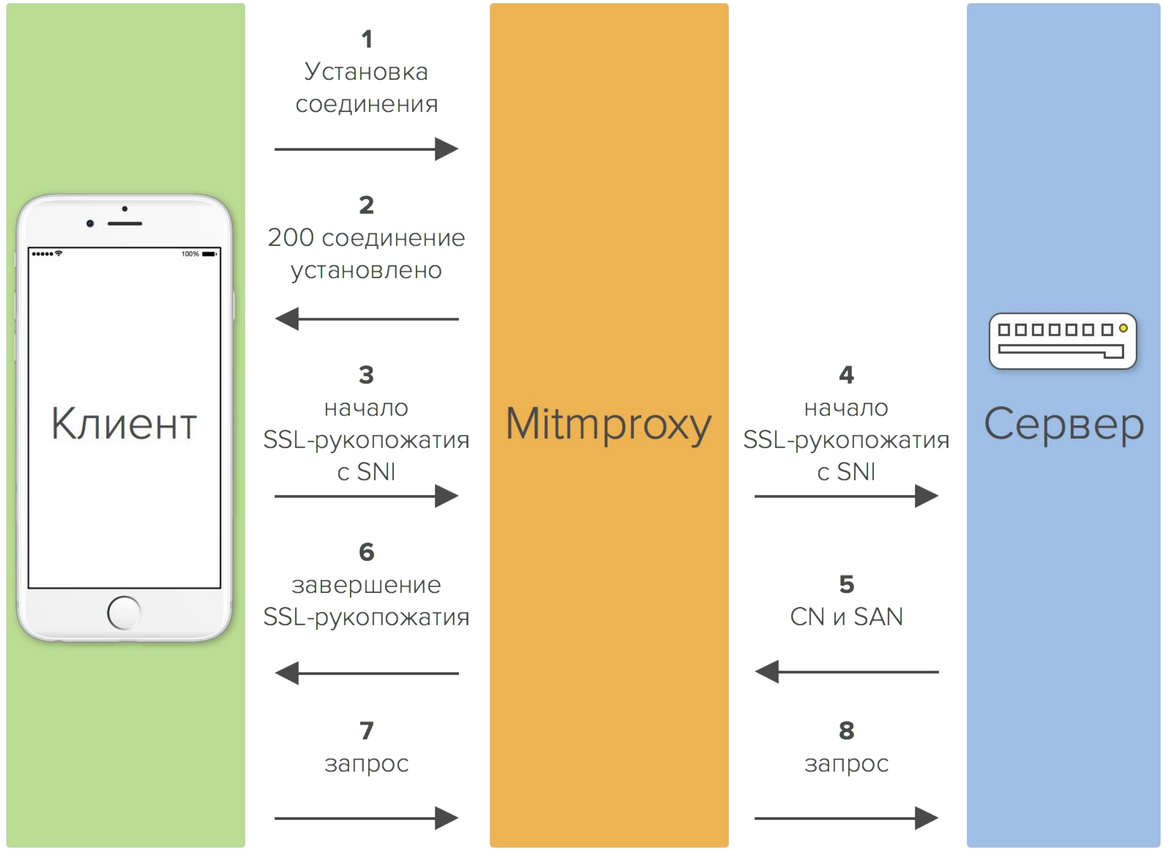
- Для каждого уникального клиентского портрета рассчитаем вероятность выбора каждой категории товара
- Сгенерируем для каждого уникального клиентского портрета по несколько подарков в соответствии с найденным распределением

Данные о товарах

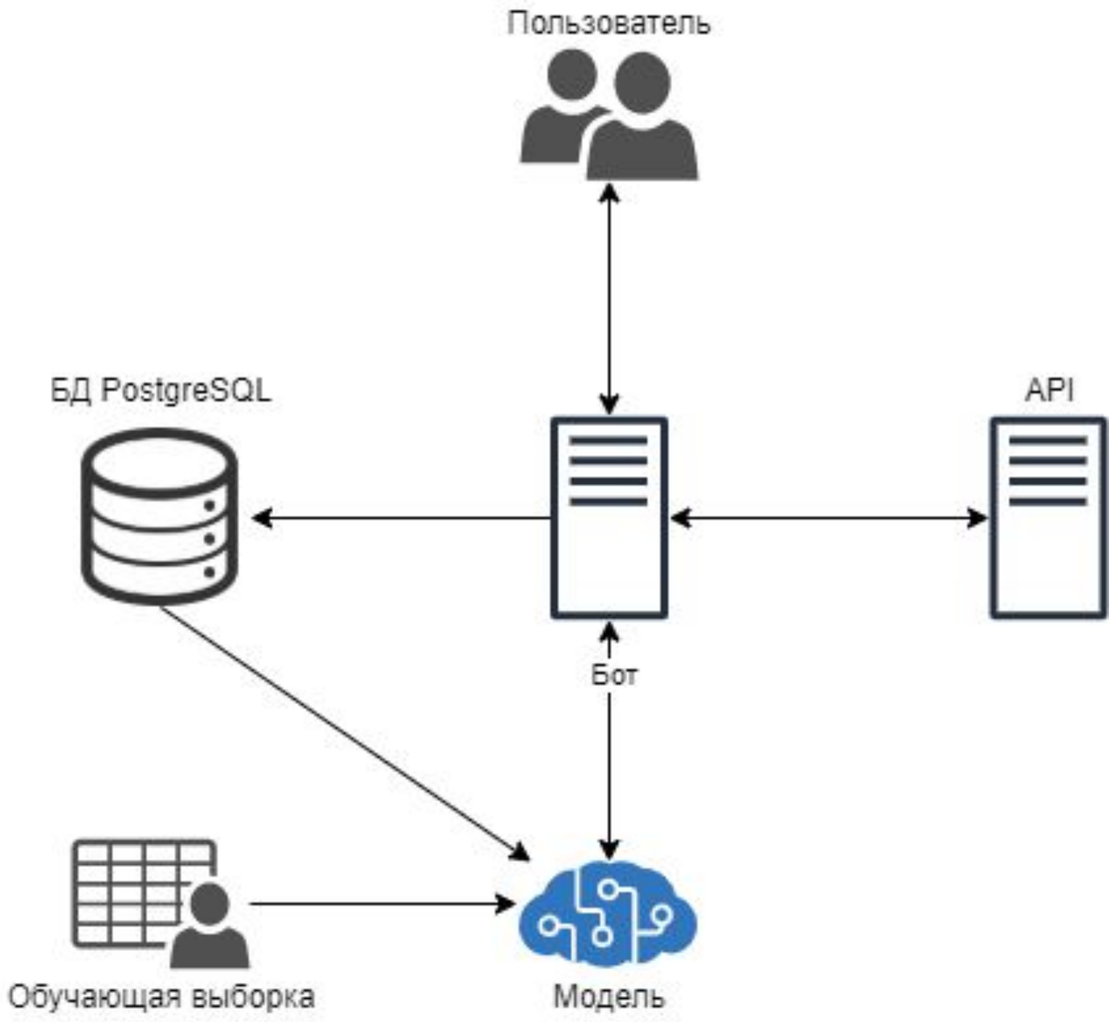
Источник данных	Минусы и сложности
Готовая БД из интернета	• Невозможность актуализации данных • Замкнутость внутри ассортимента одного/нескольких магазинов
Получение данных с помощью парсера	• Необходимость разработки парсера как отдельного сервиса • Необходимость в дополнительной инфраструктуре • Проблема матчинга товаров из разных интернет-магазинов
Получение данных от стороннего сервиса (например, Яндекс.Маркет)	• Стоимость • Отказоустойчивость

Получение данных с помощью MITM

- С помощью программы MITMproxy были обнаружены запросы мобильного приложения Яндекс.Маркет, их принцип действия был изучен.
- Выяснилось, что некоторые методы, которые предлагаются для коммерческого использования, и документация к которым есть в открытом доступе, также доступны и в мобильном API.
- С помощью запросов был получен полный список категорий сервиса Яндекс.Маркет.
- Были написаны функции для получения наиболее популярных товаров в категории, а также - для получения детальной информации о товаре.



Итоговая структура проекта



Модель

Постановка задачи:

В результате работы алгоритма мы хотим получить k категорий товаров, товары которых наиболее подходят в качестве подарка данному человеку при определенном контексте.

Задача:

Multilabel classification - Многоклассовая классификация с пересекающимися классами

Классические методы решения multilabel задачи:

- методы преобразования задачи
  - сведение к K бинарным классификациям (метод one-vs-all). В нашем случае K = 476
  - сведение к K * (K - 1) / 2 бинарным классификациям (метод all -versus- all), то есть к 113050 классификациям.
- методы преобразования алгоритма (модели) под задачу

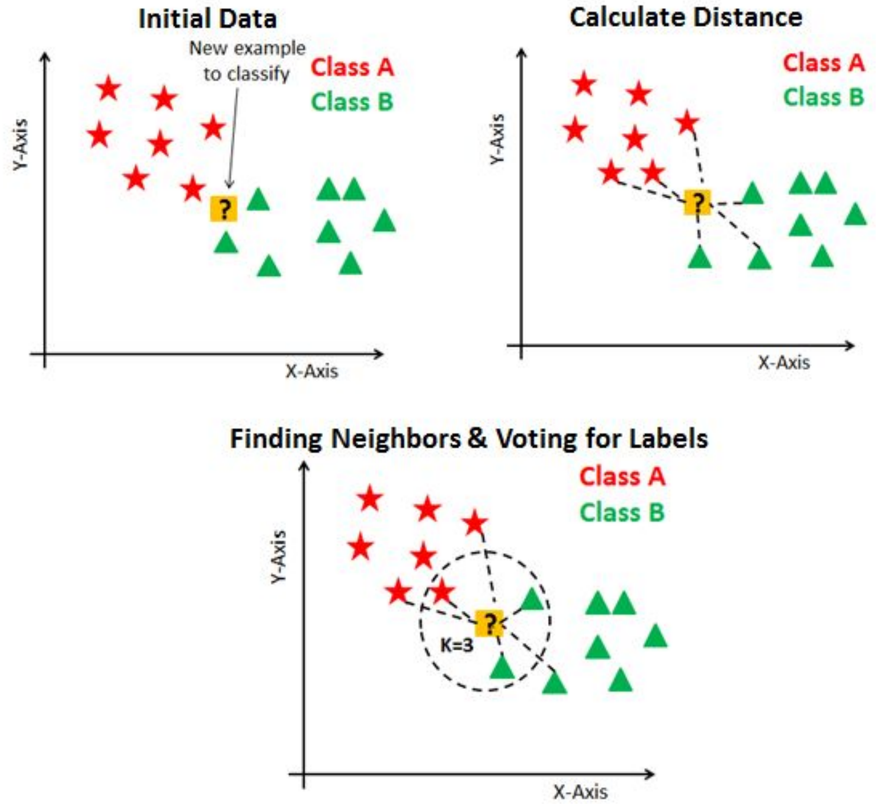
Метрики

- Coverage error $$coverage(y, \hat{f}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \max_{j:y_{ij}=1} \text{rank}_{ij}$$
- Discounted Cumulative Gain (DCG) $$\sum_{r=1}^{\min(K,M)} \frac{y_{f(r)}}{\log(1+r)}$$
- Разнообразие категорий $$diversity(x) = \frac{1}{l} \sum_{n=1}^{l} \frac{2}{(k+1) * k} \sum_{i,j=1}^{k} \rho(i,j)$$

- Покрытие возможных пользователей
- Покрытие каталога товаров
- Среднее количество товаров, которые мы рекомендуем пользователю

Простейшие классификаторы

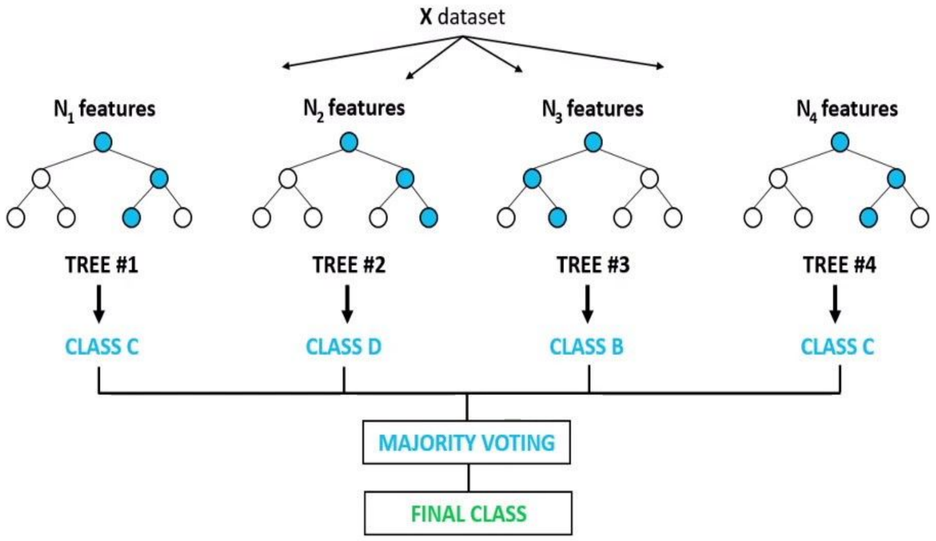
Метод ближайших соседей



Алгоритм	Время обучения	Время предсказания категории	Время предсказания товара	Сложность алгоритма
Brute Force	0.115 секунды	0.28 секунды	6.17 секунды	$O(DN^2)$
KD-Дерево	0.160 секунды	0.26 секунды	5.89 секунды	$O(DNlog(N))$
Ball Tree	0.12 секунды	0.24 секунды	5.66 секунды	$O(DNlog(N))$

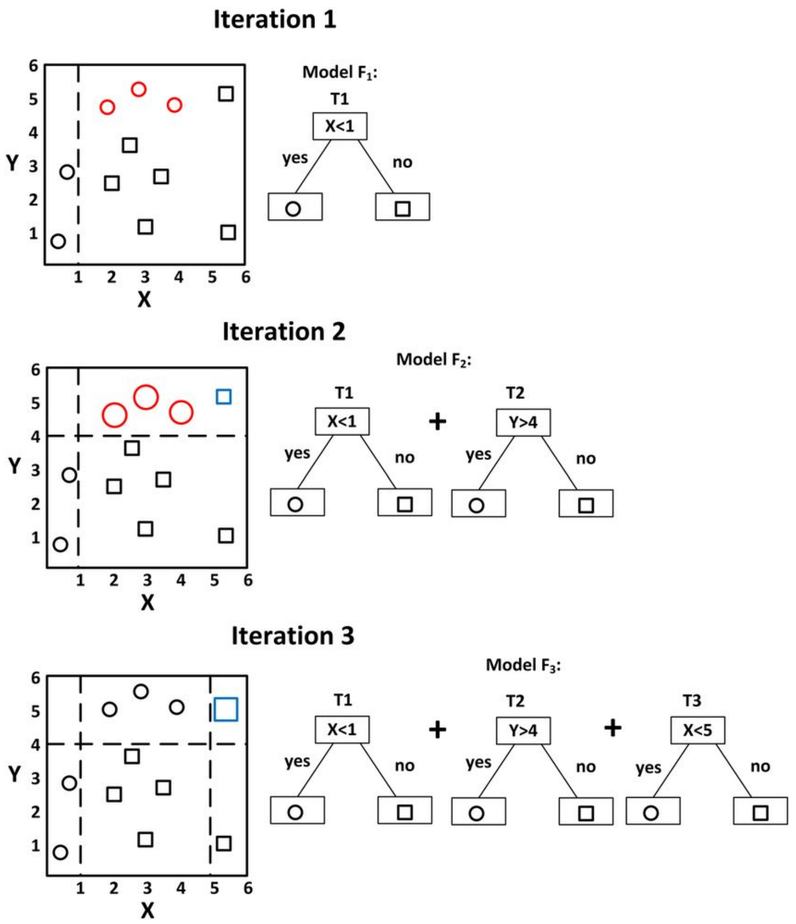
Древовидные модели

Случайный лес



Кросс-энтропия	nDCG@k	coverage error	label ranking average precision score
14.64	0.81	119.89	0.68
Разнообразие предсказаний	Среднее количество предсказаний на пользователя	Полнота покрытия каталога	Полнота покрытия пользователей
2.06	2.2	0.66	1.0

Градиентный бустинг



LGBMClassifier

Разнообразие предсказаний	Среднее количество предсказаний на пользователя	Полнота покрытия каталога	Полнота покрытия пользователей	Кросс-энтропия
2.4	2.45	0.55	0.96	32.57

XGBoost

Разнообразие предсказаний	Среднее количество предсказаний	Полнота покрытия каталога	Полнота покрытия пользователей	Кросс-энтропия
1.83	1.83	0.38	0.93	11.6

CatBoost

Разнообразие предсказаний	Среднее количество предсказаний на пользователя	Полнота покрытия каталога	Полнота покрытия пользователей	Кросс-энтропия
2.0	2.06	0.63	0.98	12.91

Нейронная сеть

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
digits (InputLayer)          (None, 26)                0
_________________________________________________________________
dense_1 (Dense)              (None, 75)                2025
_________________________________________________________________
dense_2 (Dense)              (None, 200)               15200
_________________________________________________________________
dropout_49 (Dropout)         (None, 200)               0
_________________________________________________________________
dense_3 (Dense)              (None, 500)               100500
_________________________________________________________________
dense_4 (Dense)              (None, 800)               400800
_________________________________________________________________
dropout_50 (Dropout)         (None, 800)               0
_________________________________________________________________
predictions (Dense)          (None, 476)               381276
=================================================================
Total params: 899,801
Trainable params: 899,801
Non-trainable params: 0
_________________________________________________________________
```

Кросс-энтропия	nDCG@k	coverage error	label ranking average precision score
106.6	0.6	464.2	0.3
Разнообразие предсказаний	**Среднее количество предсказаний на пользователя**	**Полнота покрытия каталога**	**Полнота покрытия пользователей**
3.0	5.0	0.42	1.0
Время обучения		**Время предсказания**	
120 секунд		0.01 секунда	

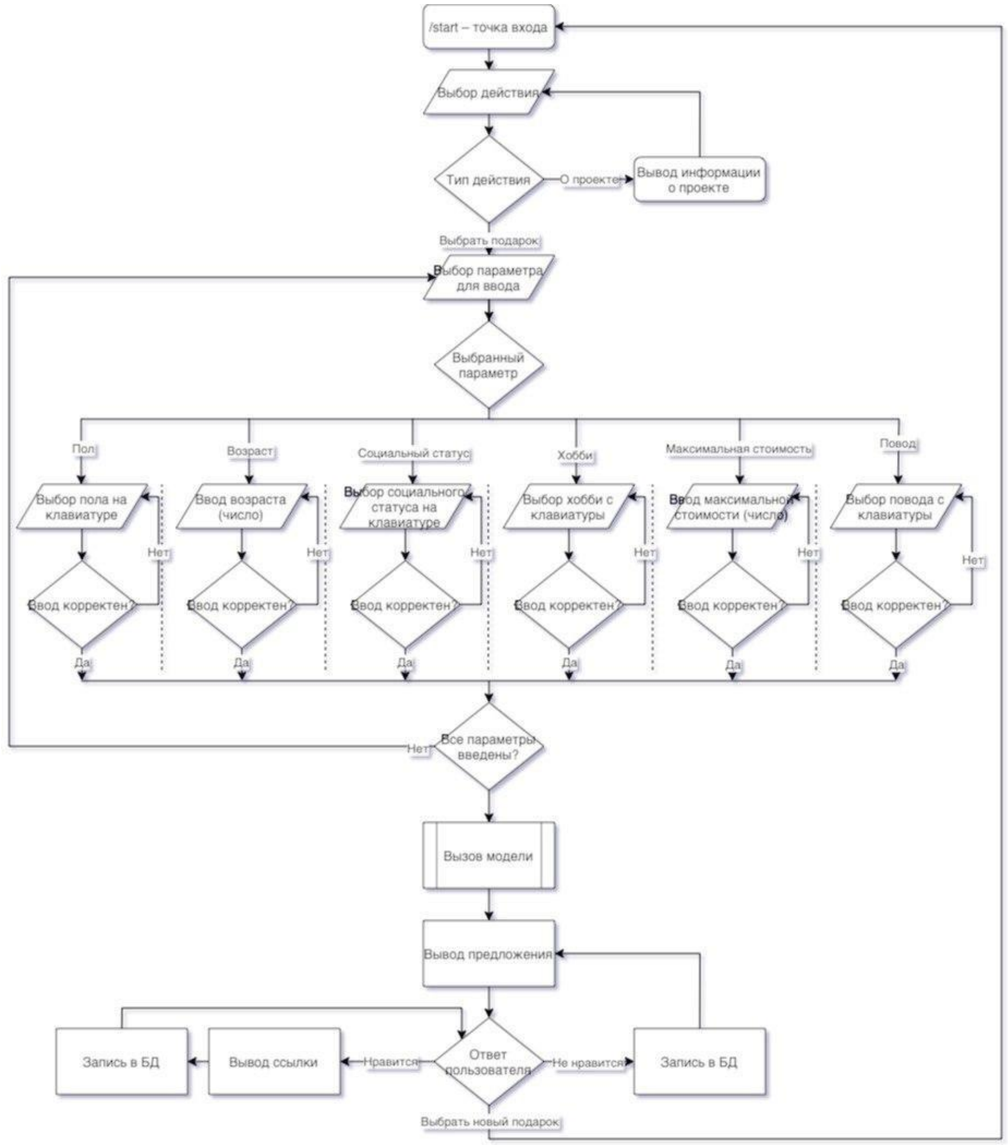
Сравнение алгоритмов

Алгоритм	Время обучения	Время предсказания
KNN	0.1 секунды	0.24 секунды
MLP	120 секунд	0.01 секунды
Случайный лес	28.1 секунды	0.92 секунды
Градиентный бустинг	7 минут	0.73 секунды

Алгоритм	Полнота покрытия каталога	Полнота покрытия пользователей	Разнообразие предсказаний
KNN	12%	100%	3.24
MLP	42%	100%	3.0
Случайный лес	4,6 %	100%	3.43
Градиентный бустинг	38%	93%	1.83

ИНТЕРФЕЙС

- В качестве интерфейса взаимодействия с конечным пользователем был выбран Telegram-бот.
- Для реализации были использованы библиотеки Telebot и pyTelegramBotAPI.
- После того, как пользователю предлагается выбранный моделью подарок, действия пользователя записываются в таблицу в БД для использования при дополнительном обучении модели.



ИТОГИ РАБОТЫ

- Анализ и поиск необходимых данных для построения предсказательной модели. Преобразование данных в нужный формат и синтетическое дублирование данных.
- Исследование различных методов машинного обучения. Разработана тактика дальнейшего улучшения качества моделей.
- Разработан сервис на основе Telegram бота, позволяющий ввести данные о человеке, которому необходимо подобрать подарок, и о стоимости подарка, а затем просмотреть возможные варианты подарка, а также при желании сразу приобрести понравившийся.

Дальнейшее развитие проекта:

- обновление обучающей выборки
- тестирование моделей на реальных пользователях для выбора наилучшей (например AB-тестирование)

Решение проблем:

- учитывание социального статуса человека
- добавление хобби, которых нет в обучающей выборке

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ