**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

Факультет компьютерных наук

Образовательная программа «Машинное обучение и высоконагруженные системы»

СОГЛАСОВАНО УТВЕРЖДАЮ

|  |  |
| --- | --- |
| Руководитель, | Академический руководитель |
| доцент департамента | образовательной программы |
| программной инженерии | «Машинное обучение и высоконагруженные системы», |
|  |  |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_М. К. Горденко | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Е. О. Кантонистова |
| “15” января 2023 г. | “\_\_\_”\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2023 г. |
|  |  |

**Выпускная квалификационная работа**

(проектно-исследовательская)

на тему: **«Оценка дохода по вакансиям интернет – рекрутера HeadHunter»**

по направлению подготовки XX.XX.XX «Машинное обучение и высоконагруженные системы»



ВЫПОЛНИЛ

студент группы XXX

образовательной программы

XX.XX.XX «Машинное обучение и высоконагруженные системы»

\_\_\_\_\_\_\_\_\_\_\_\_\_\_М. А. Королев

«2» апреля 2023 г.

**Москва 2023**

Оглавление

[Описание задачи 3](#_Toc131367803)

[1. Сбор данных 3](#_Toc131367804)

[2. Построение модели 5](#_Toc131367805)

[2.1 Разделение выборки на части. 5](#_Toc131367806)

[2.2 Сбор переменных и таргета для обучения модели 5](#_Toc131367807)

[2.3 Baseline модель без использования методов машинного обучения. 6](#_Toc131367808)

[2.4. Baseline модель с использованием методов машинного обучения. 6](#_Toc131367809)

[3. Дальнейшие планы исследования. 8](#_Toc131367810)

# Описание задачи

Часто для соискателей работы проблемой является отсутствие в описание вакансии предлагаемой заработной платы, это может значительно замедлять поиск работы, ввиду необходимости контакта с каждым работодателем.

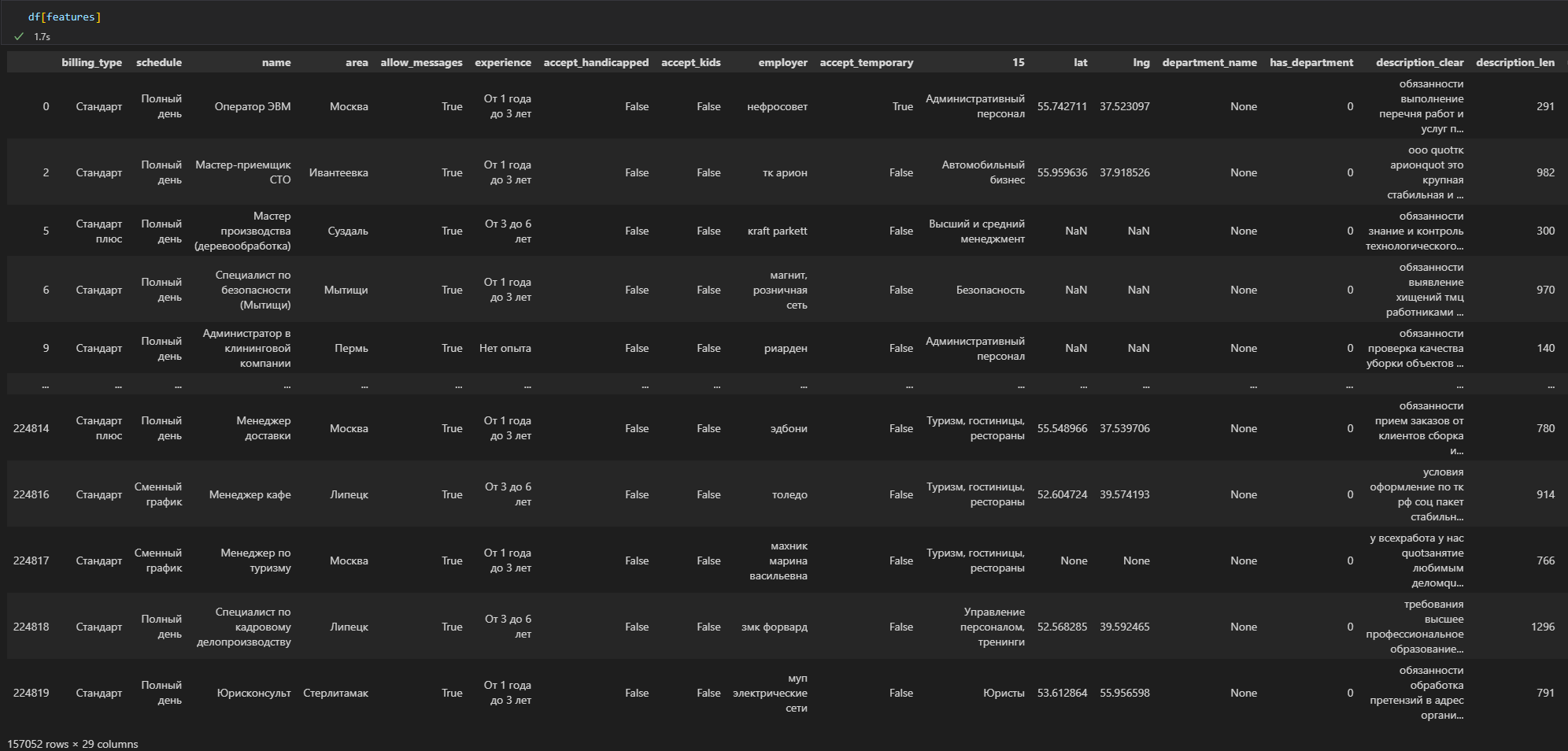
В данном исследовании была поставлена задача предсказать предлагаемый доход для вакансий, где он отсутствует на основе описания вакансии: описание компании, задачи, которые необходимо решать, требования, предъявляемые для соискателя.

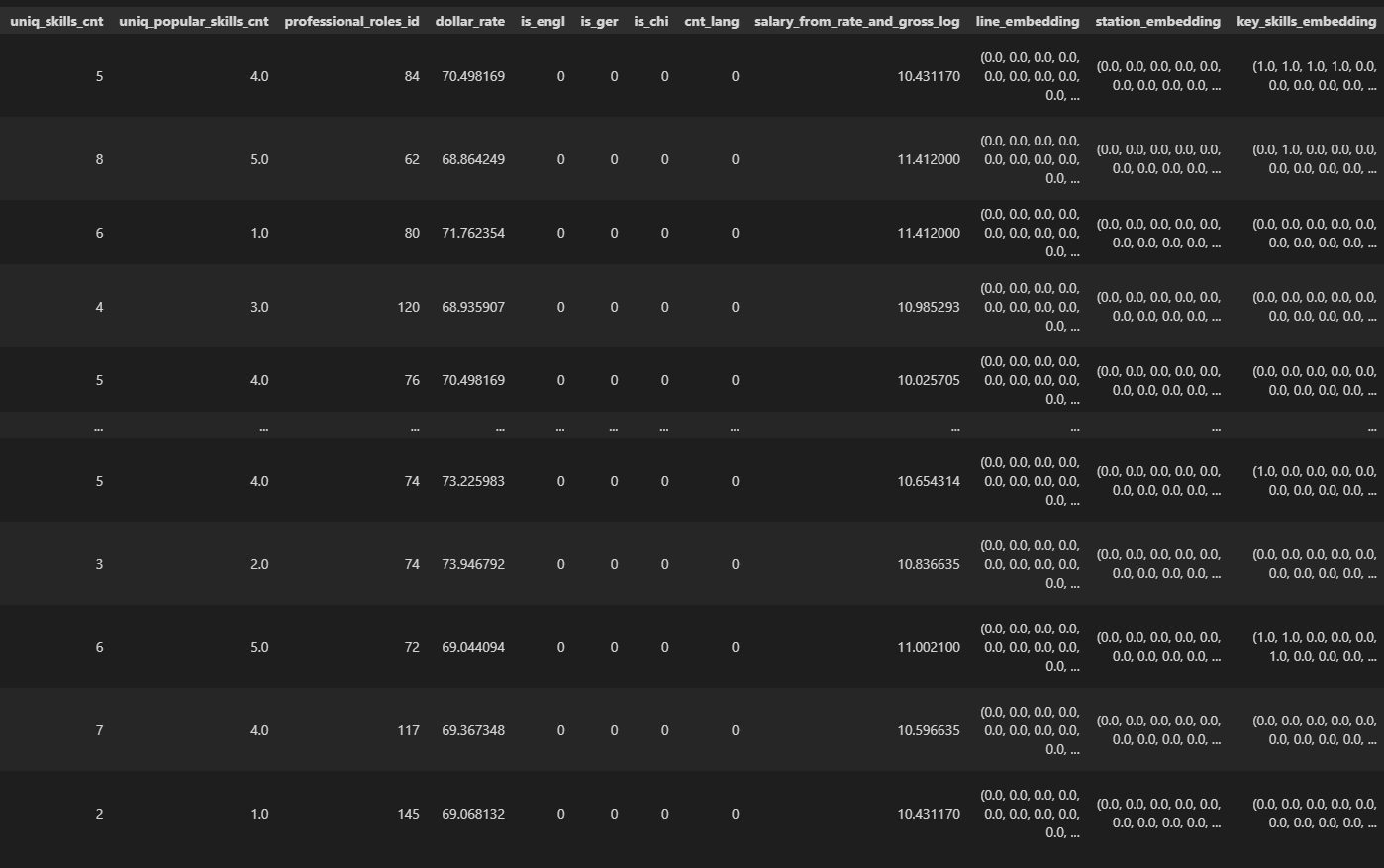
# Сбор данных

Для исследования необходимо самостоятельно собрать данные с сайта hh.ru, для этого было написано несколько программ. Первая программа необходима для сбора всех возможных профессий в HH для того, чтобы расширить выборку обучения модели. Расширение выборки оказалось необходимым выборку из-за того, что у API hh.ru имеются ограничения на выгрузку данных, а именно, 10 тыс. объявлений из одного условия. Отправляя запрос по каждой профессии, нами было собрано 250 тыс. объявлений (по 10 тыс. на каждую профессию).

Диапазон публикаций 4 месяца, начиная с 2022-12-01 и заканчивая 2022-03-01.

Вторая написанная программа брала url объявления и парсила json файл. Так как данные парсились через самую тяжелую ручку, в которой содержится вообще вся информация по объявлениям, потребовалось приобрести proxy ключи и настроить многопоточное программирование. Таким образом, были собраны следующие переменные (рис.1, рис.2).

 (рис.1)

 (рис.2)

# Построение модели

## 2.1 Разделение выборки на части.

Для корректного обучения, валидации и тестирования модели необходимо собранную выборку разбить на части. В нашем случае данные были разделены на 3 выборки: train, validation, test в пропорции 70,15,15% соответственно.

## 2.2 Сбор переменных и таргета для обучения модели

Для обучения модели было создано 4 эмбеддинга:

- эмбеддинг местоположения компании, включая станции метро;

- эмбеддинг необходимых навыков;

- эмбеддинг описания вакансии.

Длина одного вектора навыков составила 500 значений. Для эмбеддинг - описания была обучена BERT модель, однако по качеству она почти не превзошла TF-IDF, поэтому она будет подвержена доработкам в будущем исследование.

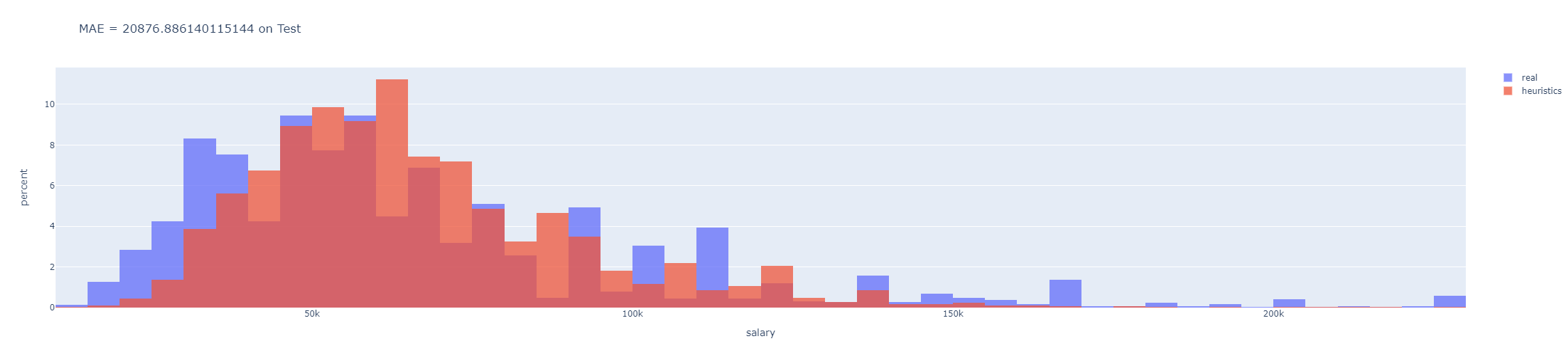
Также в качестве переменной быд собран курс доллара на дату публикации.

Помимо этого, использовались непрерывные и категориальные переменные.

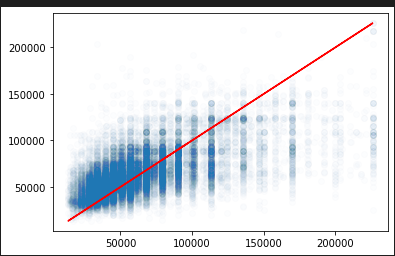
Целевая переменная (таргет) для всех объявлений был приведен к общему виду: заработная плата в гросс в рублях.

## 2.3 Baseline модель без использования методов машинного обучения.

Для того, чтобы оценить качество используемой ML модели, необходима стартовая точка сравнения - эвристика. Для этого была придумана следующая стратегия: разбить объявления на рынки, по рынкам посчитать среднее и применить на все наблюдения. Рынком является группа, состоящая из графика работы, профессии, города. Ошибка по метрики MAE составила 21 тыс.руб (рис.3, рис.4).



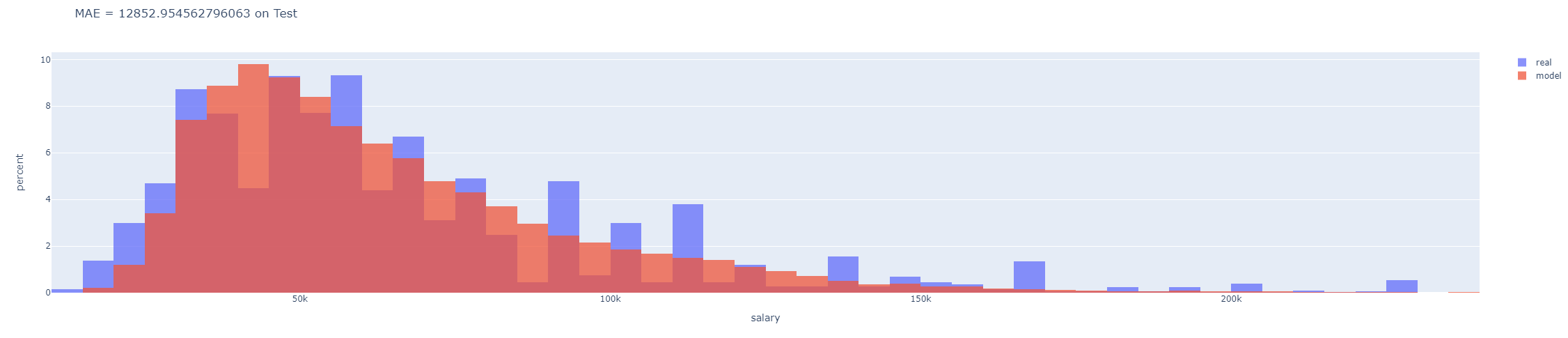
(рис.3)



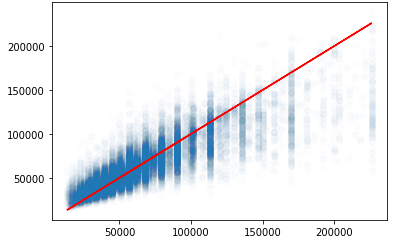
(рис.4)

## 2.4. Baseline модель с использованием методов машинного обучения.

За основу был взят градиентный бустинг, так как не все переменные зависят линейно от таргета. Для обучения модели использовался catboost.CatBoostRegressor. Функцией потерь была выбрана MAE, так как она меньше всего штрафует за большие ошибки. Результаты модели сократили ошибку эвристики в 2 раза (рис.5, рис.6).

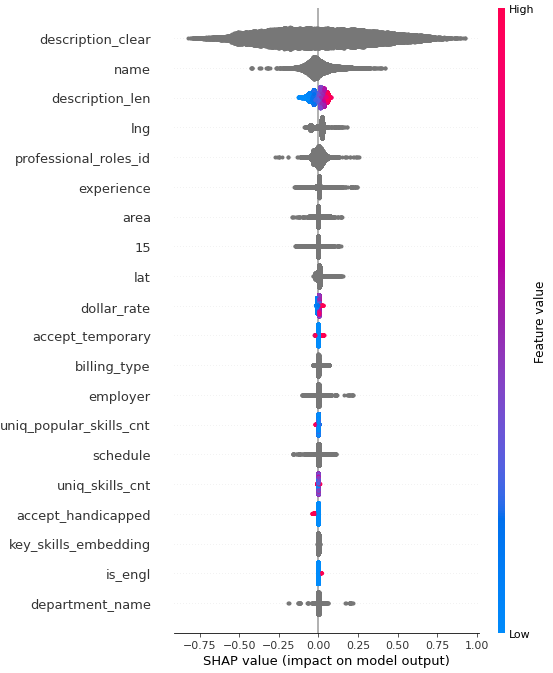


(рис.5)



(рис.6)

Порядок вхождения переменных:



(рис.7)

# Дальнейшие планы исследования.

1. Дописать BERT модель для описания вакансии;
2. Распарсить статистику по рынкам из hh.ru- <https://stats.hh.ru/far_eastern_federal_district#hhindex%5Bactive%5D=true&vacancies%5Bactive%5D=true&resumes%5Bactive%5D=true&dynamicVacancies%5Bactive%5D=true&dynamic-vacancies%5Bdynamic-vacancies%5D=month&dynamicResumes%5Bactive%5D=true&structureResumes%5Bactive%5D=true&hhindexProf%5Bactive%5D=true>
3. Построить эмбеддинг для города, для этого распарсить сайт - <https://superresearch.ru/?id=825> и эксель файлы росстата - <https://rosstat.gov.ru/folder/11109/document/13259>
4. Собрать статистику похожих объявлений через алгоритм HH
5. Собрать статистику зарплат и вакансии от компании, которая выставила объявление
6. Возможно, построить эмбеддинг отзывов о компании - <https://dreamjob.ru/employers/27953> (BERT модель)