**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
Руководитель,
доцент департамента
программной инженерии

_____М. А. Королев
“15” января 2023 г.

УТВЕРЖДАЮ
Академический руководитель
образовательной программы
«Машинное обучение и
высоконагруженные системы»,

_____Е. О. Кантонистова
“___”_____ 2023 г.

**Выпускная квалификационная работа**

(проектно-исследовательская)

на тему: «**Применение метод машинного и глубинного обучения для оценки дохода по вакансиям у 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**



# 1. 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

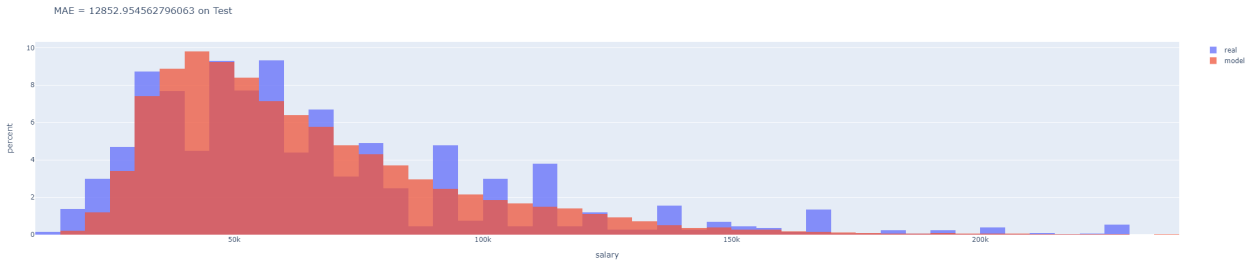
Для обучения модели было создано 4 эмбеддинга:
- эмбеддинг местоположения компании, включая станции метро;
- эмбеддинг необходимых навыков;
- эмбеддинг описания вакансии.

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.

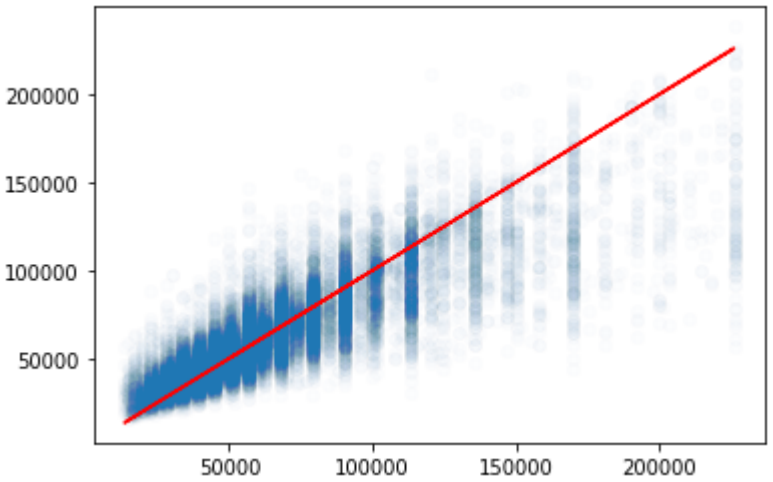
Целевая переменная (таргет) для всех объявлений был приведен к общему виду: заработная плата в гросс в рублях, также взят логарифм.

## 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. CatBoostRegressor может самостоятельно справляться с текстовыми и эмбеддинговыми переменными. 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).

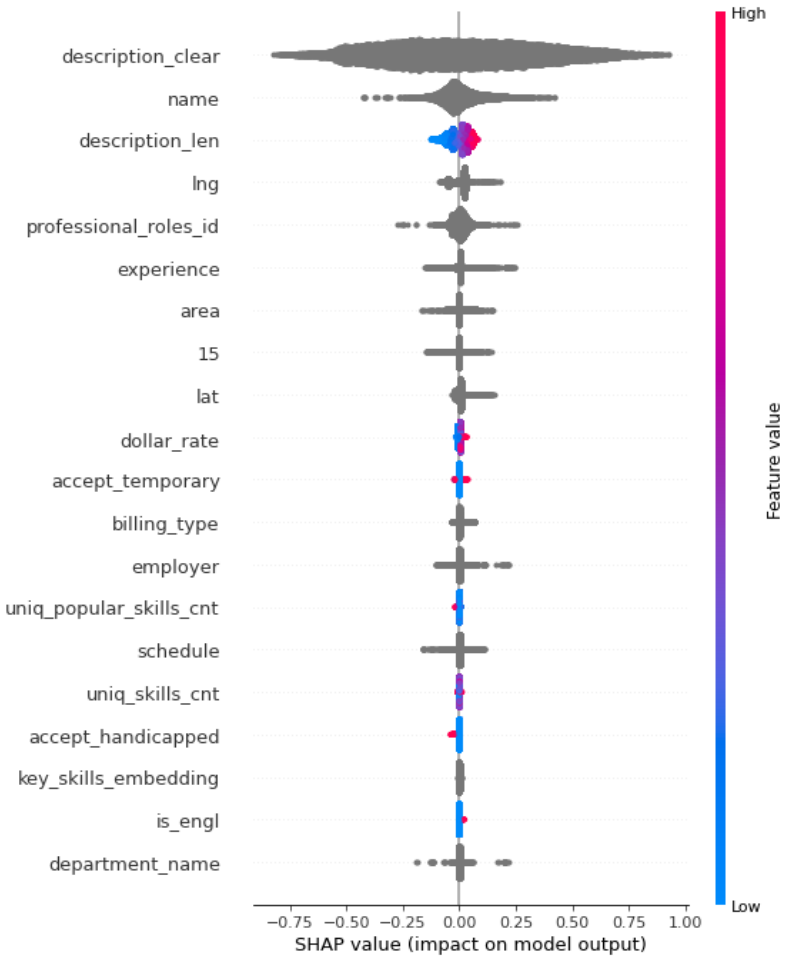



(рис.5)




(рис.6)

Порядок вхождения переменных:




(рис.7)

## 2. 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)