

**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. К. Горденко
“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: «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.

Оглавление

Описание задачи.....	4
1. Сбор данных.....	4
2. Построение модели	5
2.1 Разделение выборки на части.....	5
2.2 Сбор переменных и таргета для обучения модели.....	5
2.3 Baseline модель без использования методов машинного обучения.	6
2.4. Baseline модель с использованием методов машинного обучения.....	6
3. Дальнейшие планы исследования.....	8

Описание задачи

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

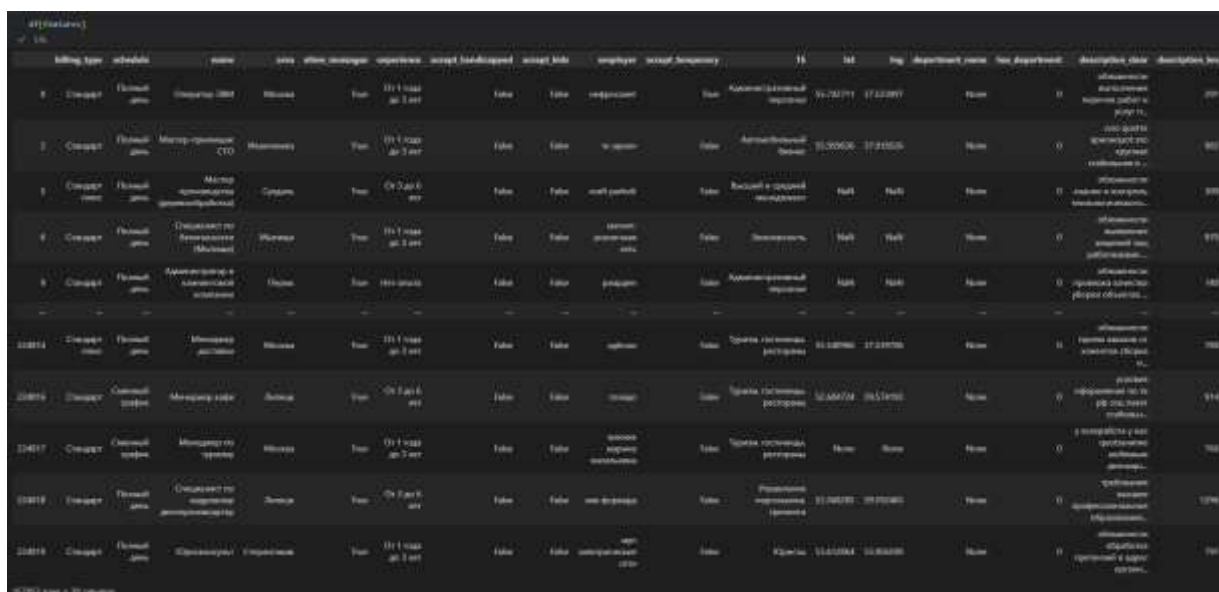
В данном исследовании была поставлена задача предсказать предлагаемый доход для вакансий, где он отсутствует на основе описания вакансии: описание компании, задачи, которые необходимо решать, требования, предъявляемые для соискателя.

1. 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 НН 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 проху ключи и настроить многопоточное программирование. Таким образом, были собраны следующие переменные (рис.1, рис.2).



id	salary_type	salary_min	salary_max	name	area	area_name	experience	accept_salary_min	accept_salary_max	company	company_salary	id	id	dep	department_name	dep_department	description_text	description_html
1	Средняя	10000	20000	Менеджер CRM	Москва	Тех	От 1 года до 3 лет	10000	20000	ИТ-компания	10000	123456789	123456789	Москва	ИТ	ИТ-компания	ИТ-компания	ИТ-компания
2	Средняя	10000	20000	Менеджер CRM	Москва	Тех	От 1 года до 3 лет	10000	20000	ИТ-компания	10000	123456789	123456789	Москва	ИТ	ИТ-компания	ИТ-компания	ИТ-компания
3	Средняя	10000	20000	Менеджер CRM	Москва	Тех	От 1 года до 3 лет	10000	20000	ИТ-компания	10000	123456789	123456789	Москва	ИТ	ИТ-компания	ИТ-компания	ИТ-компания
4	Средняя	10000	20000	Менеджер CRM	Москва	Тех	От 1 года до 3 лет	10000	20000	ИТ-компания	10000	123456789	123456789	Москва	ИТ	ИТ-компания	ИТ-компания	ИТ-компания
5	Средняя	10000	20000	Менеджер CRM	Москва	Тех	От 1 года до 3 лет	10000	20000	ИТ-компания	10000	123456789	123456789	Москва	ИТ	ИТ-компания	ИТ-компания	ИТ-компания
6	Средняя	10000	20000	Менеджер CRM	Москва	Тех	От 1 года до 3 лет	10000	20000	ИТ-компания	10000	123456789	123456789	Москва	ИТ	ИТ-компания	ИТ-компания	ИТ-компания
7	Средняя	10000	20000	Менеджер CRM	Москва	Тех	От 1 года до 3 лет	10000	20000	ИТ-компания	10000	123456789	123456789	Москва	ИТ	ИТ-компания	ИТ-компания	ИТ-компания
8	Средняя	10000	20000	Менеджер CRM	Москва	Тех	От 1 года до 3 лет	10000	20000	ИТ-компания	10000	123456789	123456789	Москва	ИТ	ИТ-компания	ИТ-компания	ИТ-компания
9	Средняя	10000	20000	Менеджер CRM	Москва	Тех	От 1 года до 3 лет	10000	20000	ИТ-компания	10000	123456789	123456789	Москва	ИТ	ИТ-компания	ИТ-компания	ИТ-компания
10	Средняя	10000	20000	Менеджер CRM	Москва	Тех	От 1 года до 3 лет	10000	20000	ИТ-компания	10000	123456789	123456789	Москва	ИТ	ИТ-компания	ИТ-компания	ИТ-компания

(рис.1)

uniq_skills_cnt	uniq_popular_skills_cnt	professional_rules_id	dollar_rate	is_engl	is_ger	is_chi	cnt_lang	salary_from_rate_and_gross_log	line_embedding	station_embedding	key_skills_embedding
5	4.0	64	70.458169	0	0	0	0	10.431170	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 1.0 1.0 1.0 0.0 0.0 0.0 0.0 0.0 -
8	5.0	62	60.866349	0	0	0	0	11.412003	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 1.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -
6	1.0	80	71.762354	0	0	0	0	11.412003	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -
4	3.0	120	66.985927	0	0	0	0	10.985293	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -
5	4.0	76	70.458169	0	0	0	0	10.025705	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -
-	-	-	-	-	-	-	-	-	-	-	-
5	4.0	74	73.225989	0	0	0	0	10.634314	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -
3	2.0	74	73.946792	0	0	0	0	10.836635	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -
6	5.0	72	69.866349	0	0	0	0	11.032100	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 1.0 0.0 0.0 0.0 1.0 0.0 0.0 0.0 -
7	4.0	117	69.367348	0	0	0	0	10.596635	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -
2	1.0	145	69.068132	0	0	0	0	10.431170	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -	0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 -

(рис.2)

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 эмбединга:

- эмбединг местоположения компании, включая станции метро;
- эмбединг необходимых навыков;
- эмбединг описания вакансии.

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

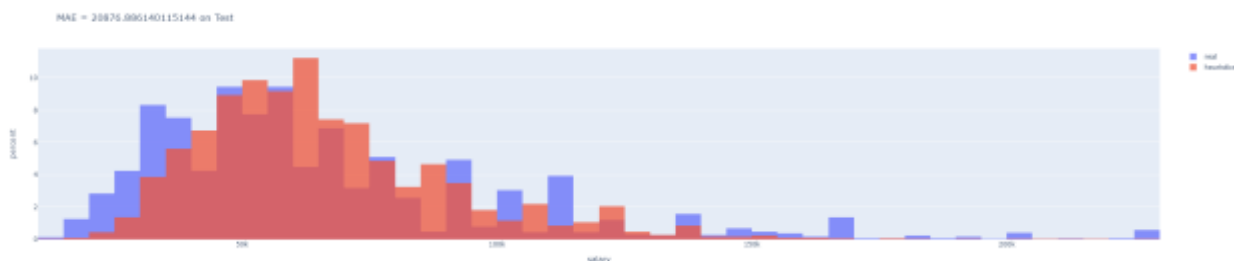
Также в качестве переменной был собран курс доллара на дату публикации.

Помимо этого, использовались непрерывные и категориальные переменные.

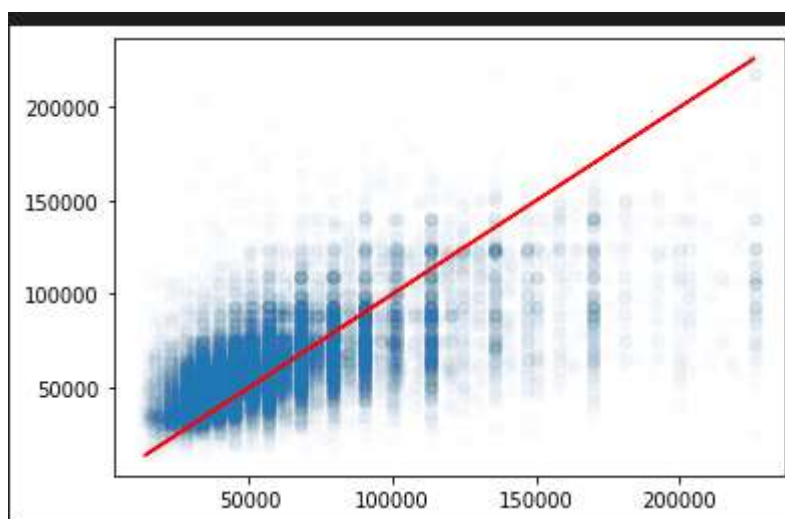
Целевая переменная (таргет) для всех объявлений был приведен к общему виду: заработная плата в гросс в рублях.

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).



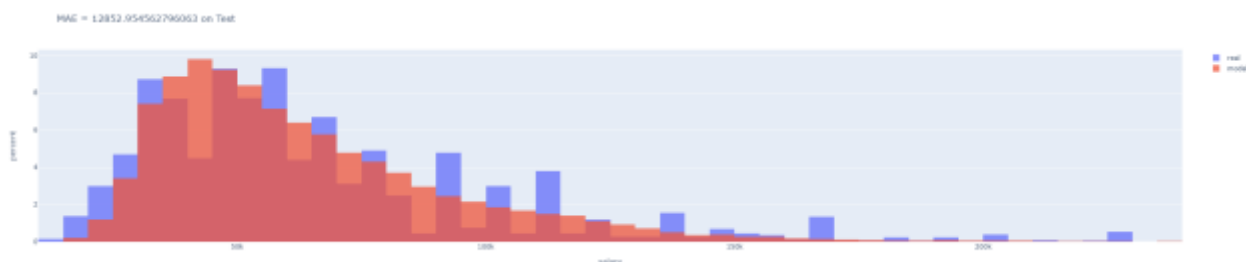
(рис.3)



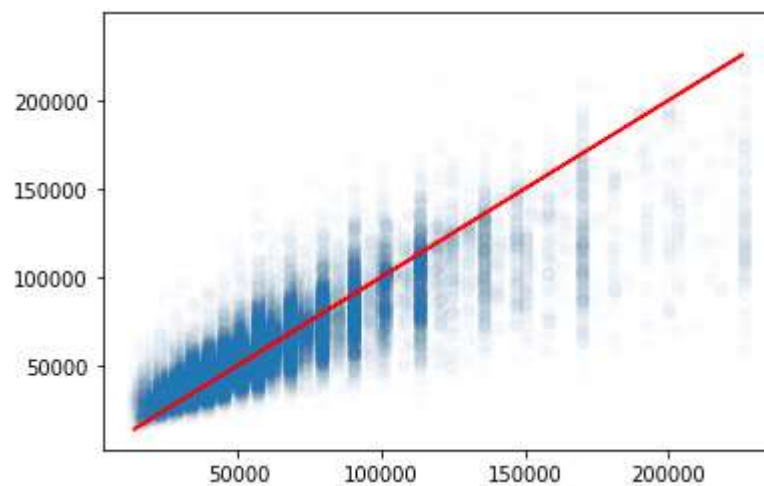
(рис.4)

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).

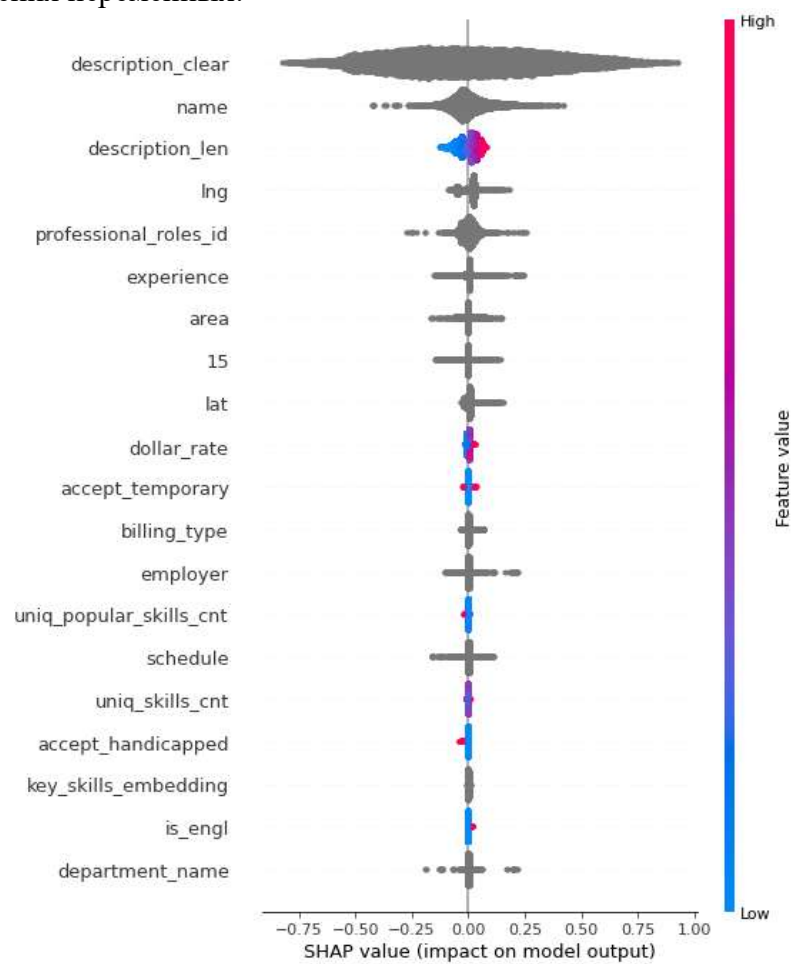


(рис.5)



(рис.6)

Порядок вхождения переменных:



(рис.7)

3. 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 эмбединг 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 НН
- 5) Собрать статистику зарплат и вакансии от компании, которая выставила объявление
- 6) Возможно, построить эмбединг отзывов о компании -
<https://dreamjob.ru/employers/27953> (BERT модель)