Применение алгоритмов машинного обучения для автоматического анализа вакансий из портала HeadHunter.ru

Автор работы: Валерия Морозова
Научный руководитель: Александр Югай

1. Введение

Автоматический анализ кадровых документов представляет существенный интерес в текущей практике машинного обучения и применяется как работодателями и рекрутинговыми агентствами, так и соискателями. На сегодняшний день на портале kaggle.com опубликовано более 50 соревнований, посвященных задаче классификации вакансий, а также предсказания заработной платы по тексту вакансии.

В текущем исследовании будут рассматриваться методы автоматического анализа текстов, а также метаданных вакансий одной из крупнейших платформ для поиска работы в русскоязычном домене HeadHunter.ru.

1.1. Цель работы

Создание алгоритмов машинного обучения для следующих приложений:

1. предсказания зарплаты по тексту и/или метаданным вакансии;
2. генерации численного вектора вакансии для сравнения документов между собой (в том числе, для поиска наиболее релевантных или подобных вакансий.

1.2. Задачи

- сбор данных;
- разведочный анализ данных;
- формирование обучающей, валидационной и тестовой выборки;
- подбор архитектуры модели;
- обучение модели.

1.3. Данные

В настоящий момент датасет состоит из примерно 30 тысяч документов описаний и метаданных вакансий из портала HeadHunter.ru в формате json. Данные

получены из открытого API портала [1]. Общедоступная версия API допускает ограниченное количество запросов в сутки. Неограниченная версия доступна для корпоративных клиентов.

В текущем исследовании рассматриваются наиболее поздние вакансии, опубликованные в 2022-2023 годах. Узкий временной промежуток выбран с целью исключения из рассмотрения влияния на целевую экономических факторов, таких как изменяющийся спрос на специальности на рынке труда, насыщение рынка труда, инфляция и т.д.

Модель данных описана в официальной документации [2]. Она содержит в себе техническую информацию портала, метаданные локации вакансии, работодателя, категориальные признаки требований и ожидаемых навыков соискателя, необходимый опыт и текстовое описание вакансии.

Описание вакансии представлено в двух вариантах: в стандартном и брендированном оформлении. Стандартное описание присутствует во всех вакансиях, брендированное описание не обязательно к заполнению.

Зарплата описана в виде диапазона “от” и “до”, в котором одно или оба значения могут быть пропущены. Также указывается валюта суммы и учет НДФЛ.

2. Обзор литературы

В литературе описывается ряд подходов к решению задачи предсказания зарплаты по данным вакансии. В [3] авторы прогнозируют логарифм зарплаты в фунтах стерлингов по текстовому описанию вакансии. Они рассмотрели ряд методов, такие как решающие деревья, классические регрессионные модели и неглубокие нейронные сети. Наилучшую метрику на тестовой выборке показал случайный лес. По мнению авторов, итоговое качество прогноза (MAE) модели случайного леса превосходит прогноз человека.

В [4] решена смежная задача классификации названия позиции по ее описанию. Авторы применили методы глубокого машинного обучения, основанные на рекуррентных сверточных архитектурах, и достигли более 0.7 на F1-score.

3. Разведочный анализ данных

3.1. Предварительная обработка данных

Некоторые из полей требуют предварительной обработки для последующего анализа.

3.1.1. Зарплатная вилка

Формат представления зарплаты включает в себя следующие особенности:

1. Зарплата может быть указана как диапазон значений или не быть указана совсем.
2. Зарплата может быть указана в различных валютах.
3. Встречаются случаи указания зарплаты как до, так и после вычета налогов (это также отображено в структуре документа вакансии).

В результате обработки зарплатные вилки приведены в единый диапазон значений (в рублях после вычета налогов), а также добавлено поле, описывающее вилку одним значением — максимальным в вилке.

3.1.2. Текст вакансии

В структуре документа представлено два поля с описанием вакансии: обычное описание и брендированное. Они могут быть заполнены одновременно, но разным текстом.

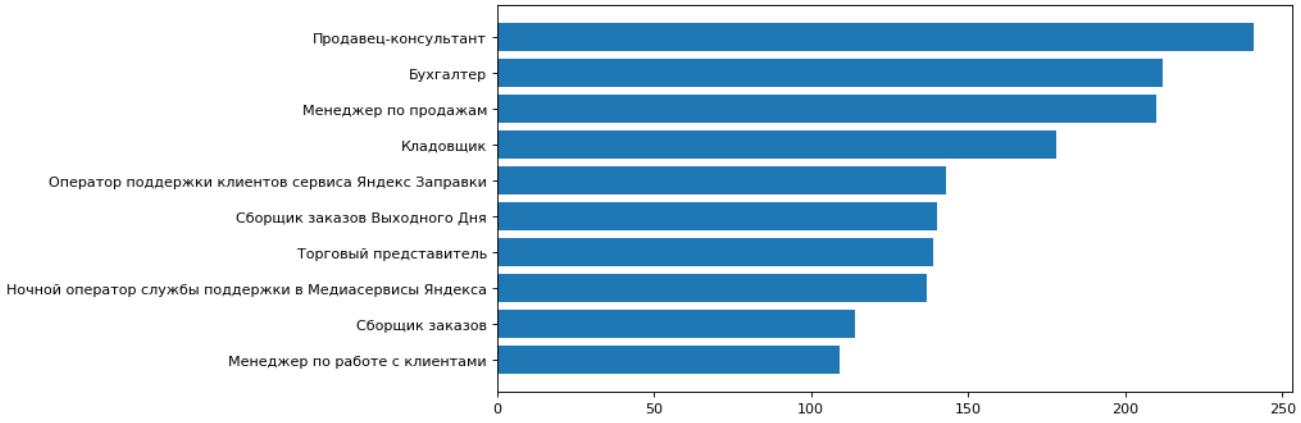
Данные поля содержат текст в html разметке. В результате обработки они были очищены от разметки для последующего текстуального анализа.

Была произведена предварительная токенизация, удаление специальных символов и пунктуации и лемматизация.

3.2. Анализ метаданных

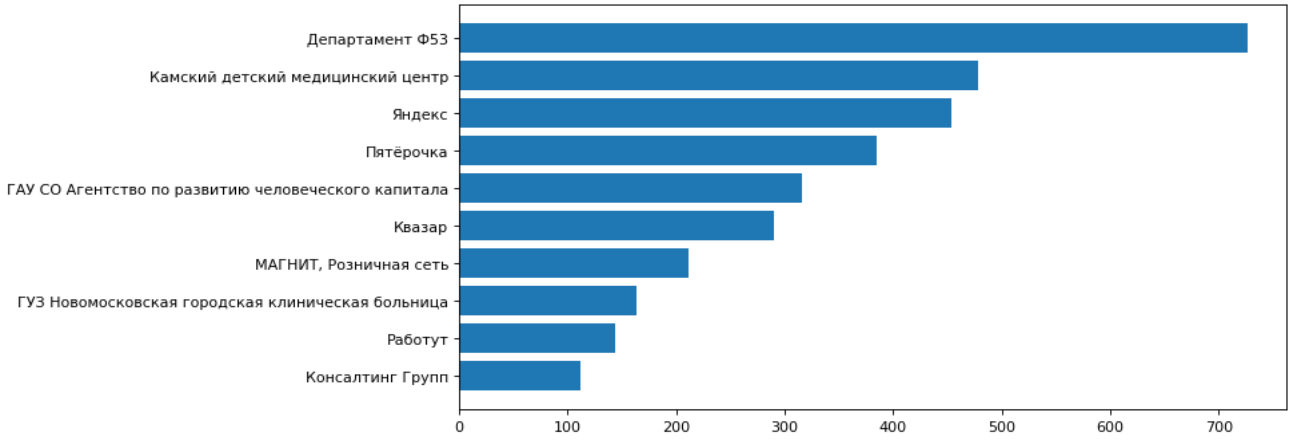
Наименование вакансии не представляет особенного интереса, так как частота их повторения невысока. Каждое второе вхождение уникально. Среди частотных наименований можно увидеть преимущественно вакансии, не требующие высокой квалификации.

Изображение 1. График частотности наименований вакансий



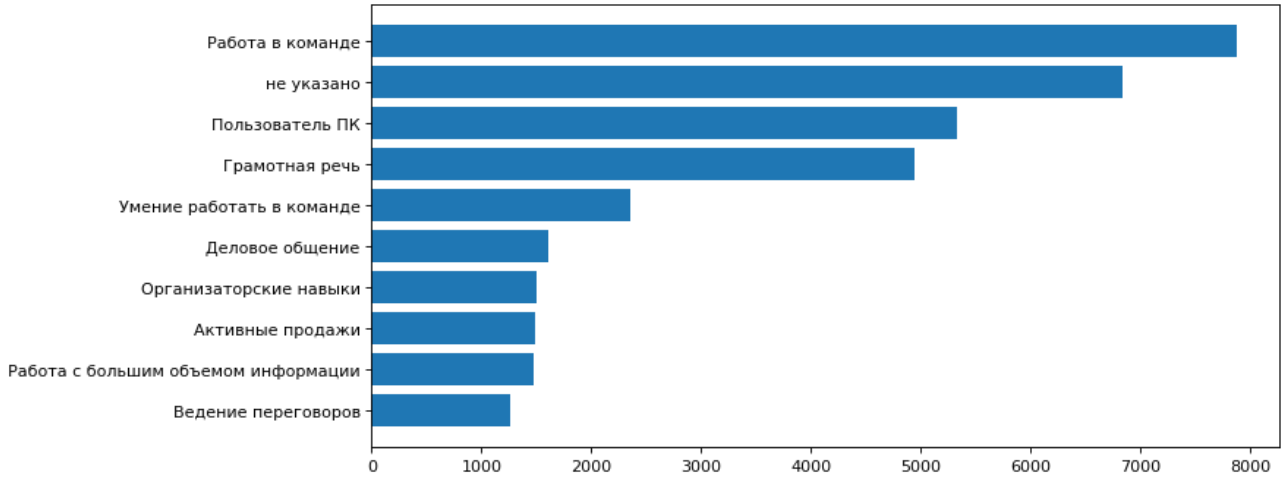
Большинство вакансий опубликовано в основном крупными российскими компаниями.

Изображение 2. График частотности наименований работодателей



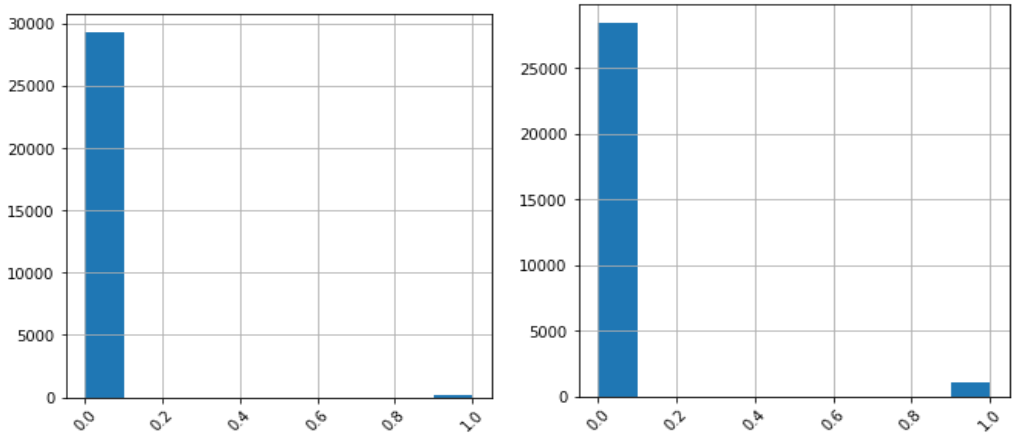
Самый популярные навыки также достаточно общие, которые можно отнести к широкому спектру специализаций.

Изображение 3. График частотности навыков

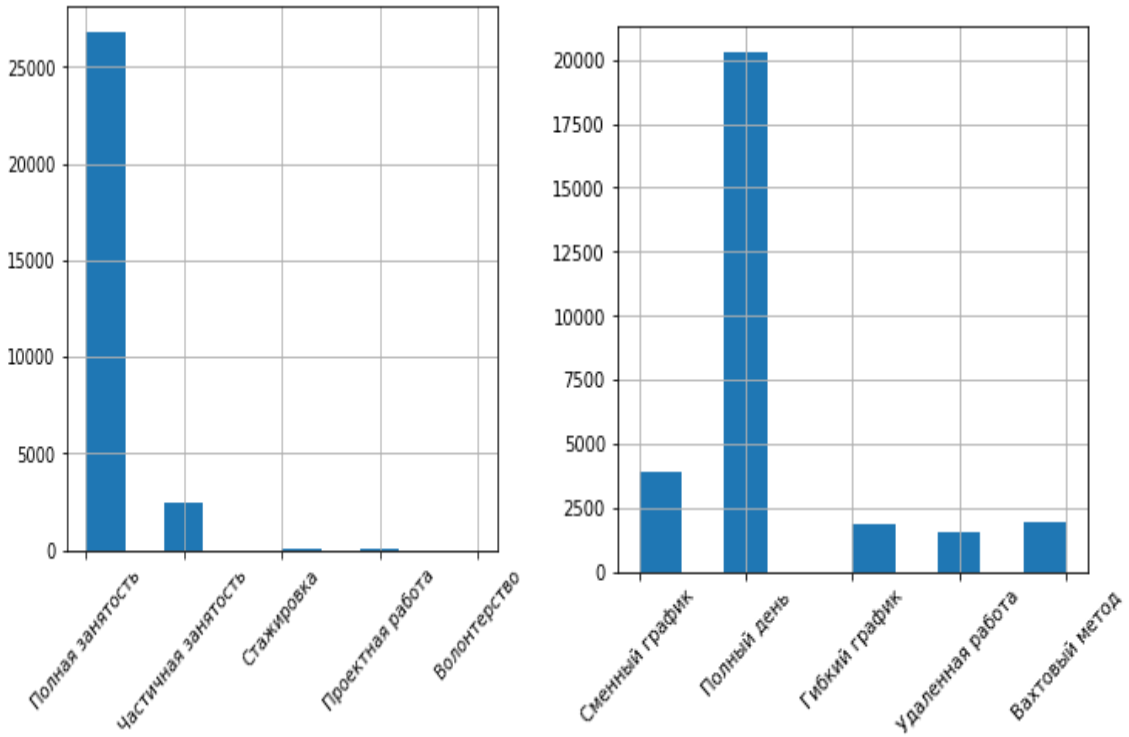


Большинство вакансий подразумевают полную занятость и стандартный график на полный день on site. Вакансии в основном недоступны для детей и для соискателей с инвалидностью.

Изображение 4. График частотности вакансий, доступных для детей (слева) и соискателей с инвалидностью (справа)

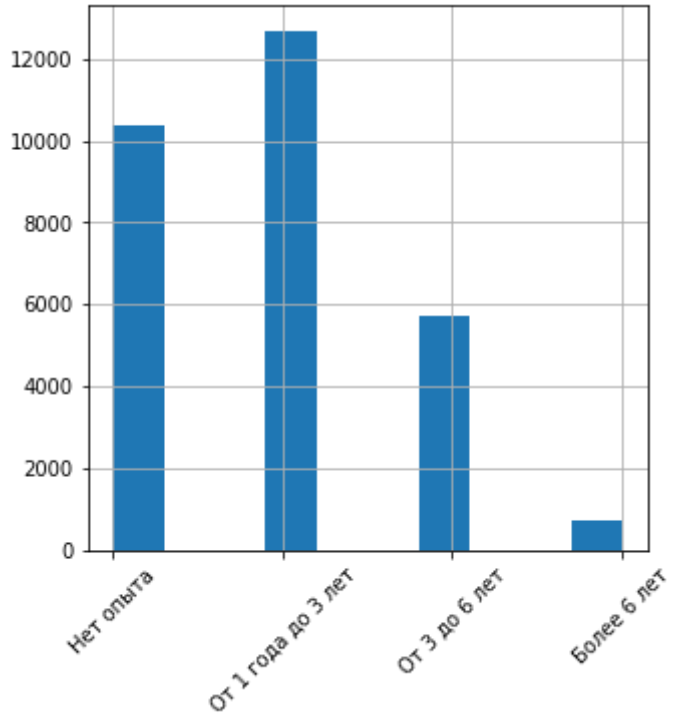


Изображение 5. График частотности по типам занятости (слева) и типам графика (справа)



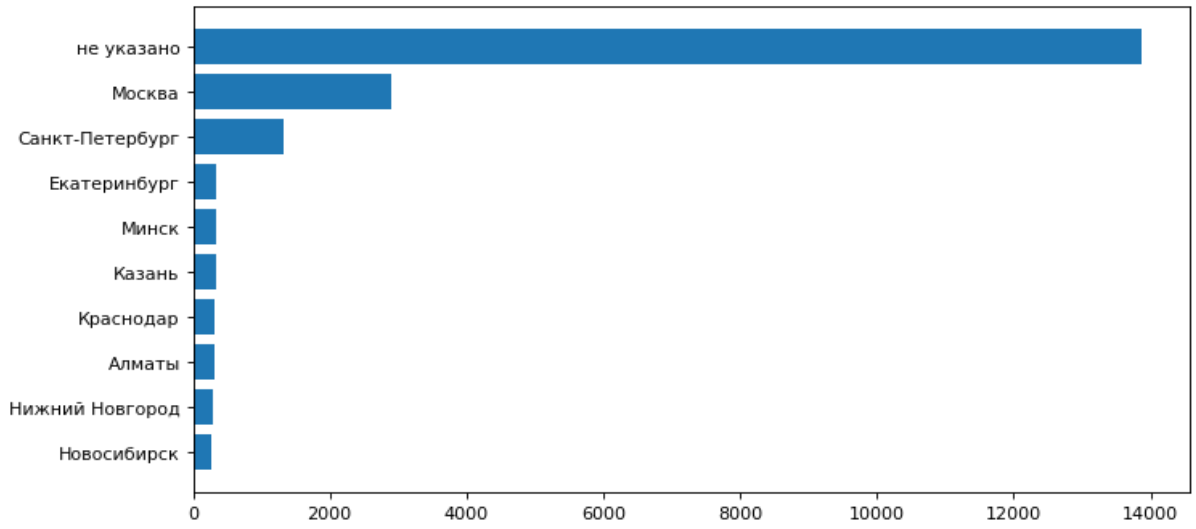
Большинство вакансий требуют минимального опыта работы, но также достаточно большая доля вакансий не требует никакого опыта.

Изображение 6. График частотности требуемого опыта



Большинство вакансий открыто в крупных городах России и ближнего зарубежья. Любопытно, что почти в половине выборки локация не указана, при этом, согласно графикам выше, только меньше 10% вакансий допускают удаленную работу. Я предполагаю, что локации не указаны для неспецифических вакансий не требующих высокой квалификации от крупных работодателей, такие как продавец, кассир и т.д.

Изображение 7. График частотности локаций вакансий

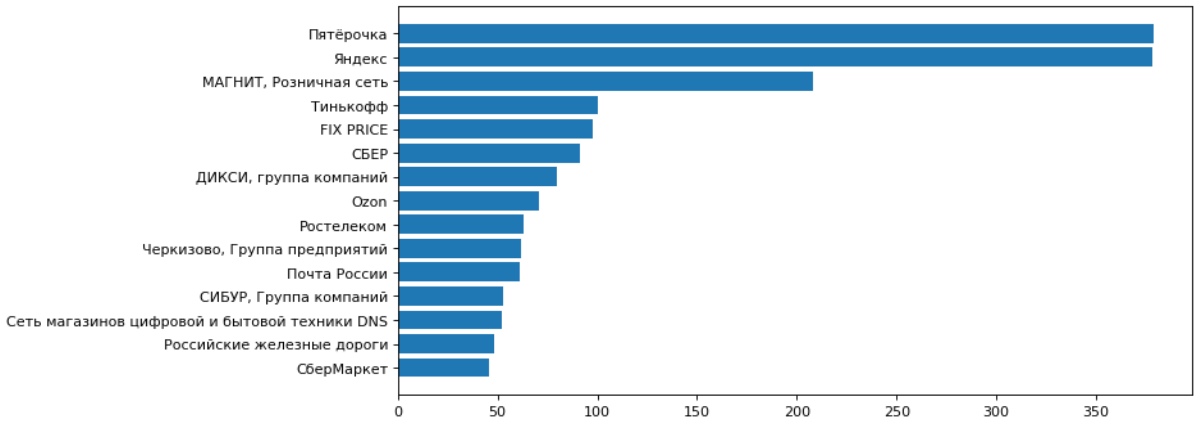


3.3. Анализ текстовых данных

Стандартное описание вакансии заполнено во всех случаях (скорее всего, по той причине, что это поле является обязательным для работодателя). Содержимое полей не совпадает, однако оба из них могут содержать информацию о вакансии. В то же время брендированное описание заполнено только для 18% полей.

На представленном графике можно увидеть, что брендированное описание заполнено в основном у крупных работодателей.

Изображение 8. График отношения наименования работодателя к числу вхождений брендированных описаний



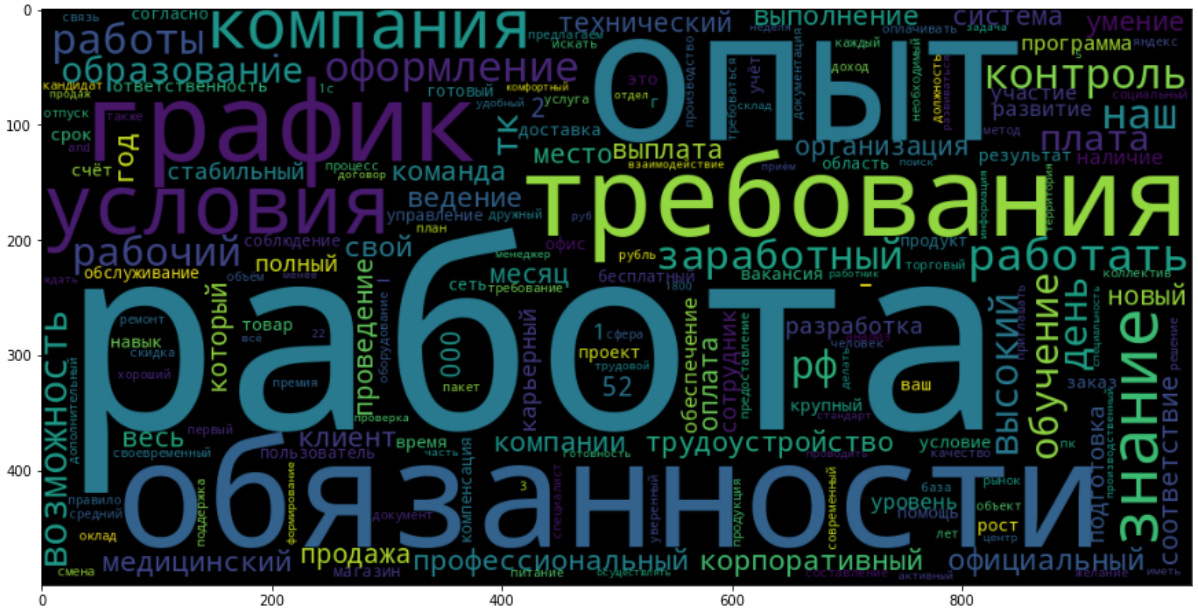
В таблице указаны основные текстовые метрики. Из значений следует, что брендированное описание, как правило, значительно объемнее стандартного.

Таблица 1. Основные текстовые метрики

	заполнено, %	средняя длина, символы	средняя длина, токены
брендированное описание	18	1928	235
стандартное описание	100	1211	151

Однако анализ частотности токенов в предобработанных текстах на корпусах брендированных и стандартных описаний показывает, то содержание этих полей отличается несущественно.

Изображение 9. Облако слов брендированных описаний

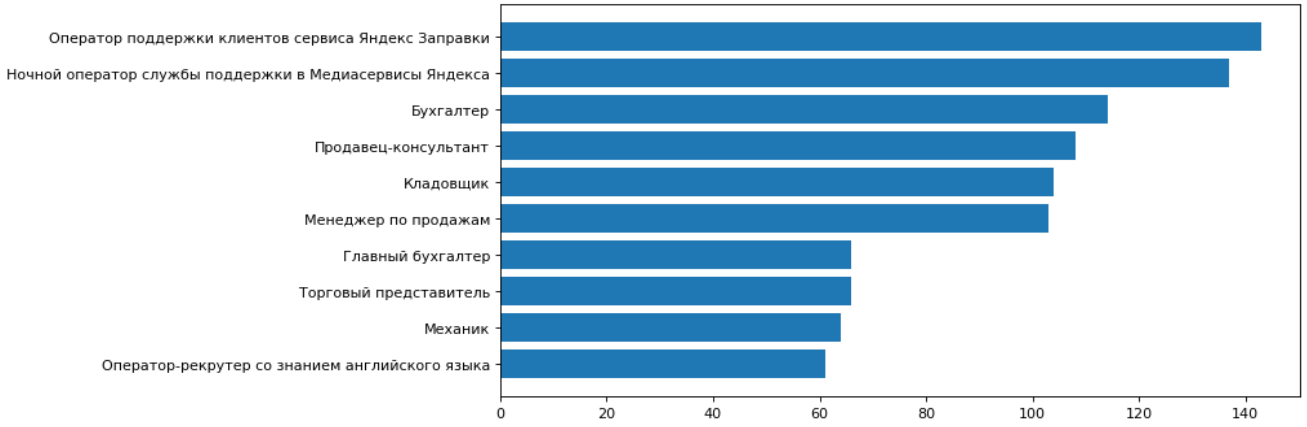


Изображение 10. Облако слов стандартных описаний

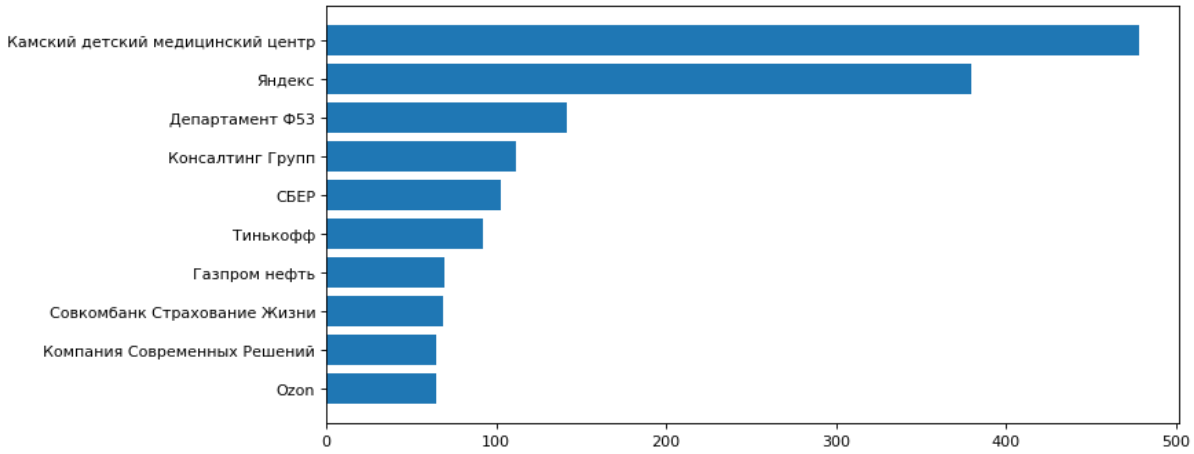
3.4. Анализ целевой переменной

В текущей выборке в 76% вакансий заполнено хотя бы одно из значений зарплатной вилки. На графиках ниже представлены наиболее частотные названия позиций и работодатели вакансий, в которых не указано никаких данных о зарплате. Видно, что превалируют крупные корпорации и низкоквалифицированные и универсальные позиции. Полагаю, что зарплата для подобных вакансий зависит от региона вакансии.

Изображение 11. График частотности наименования позиции к вакансиям без указанной заработной платы

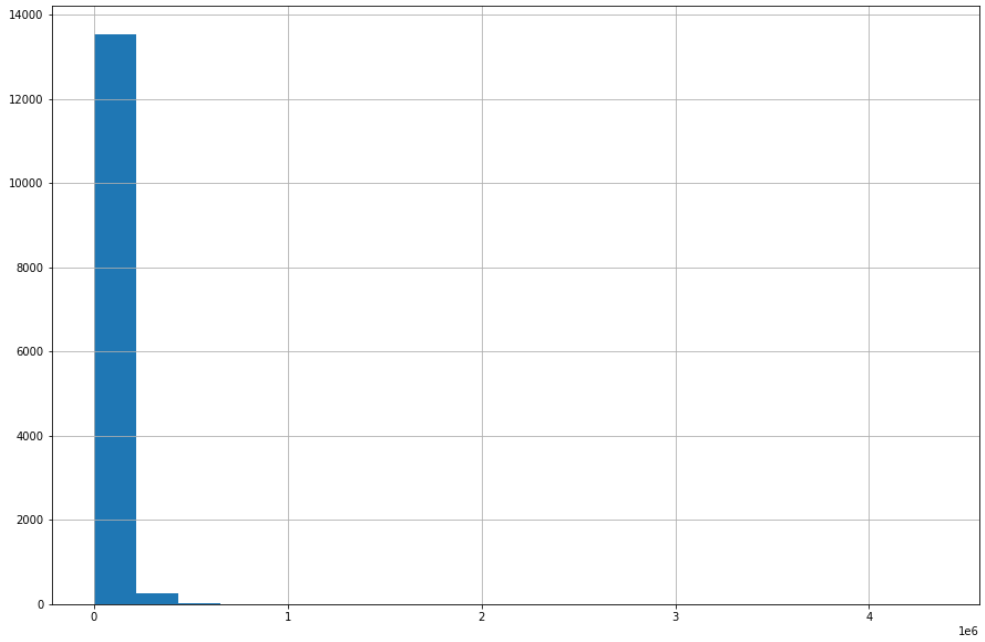


Изображение 12. График частотности наименования работодателя к вакансиям без указанной заработной платы

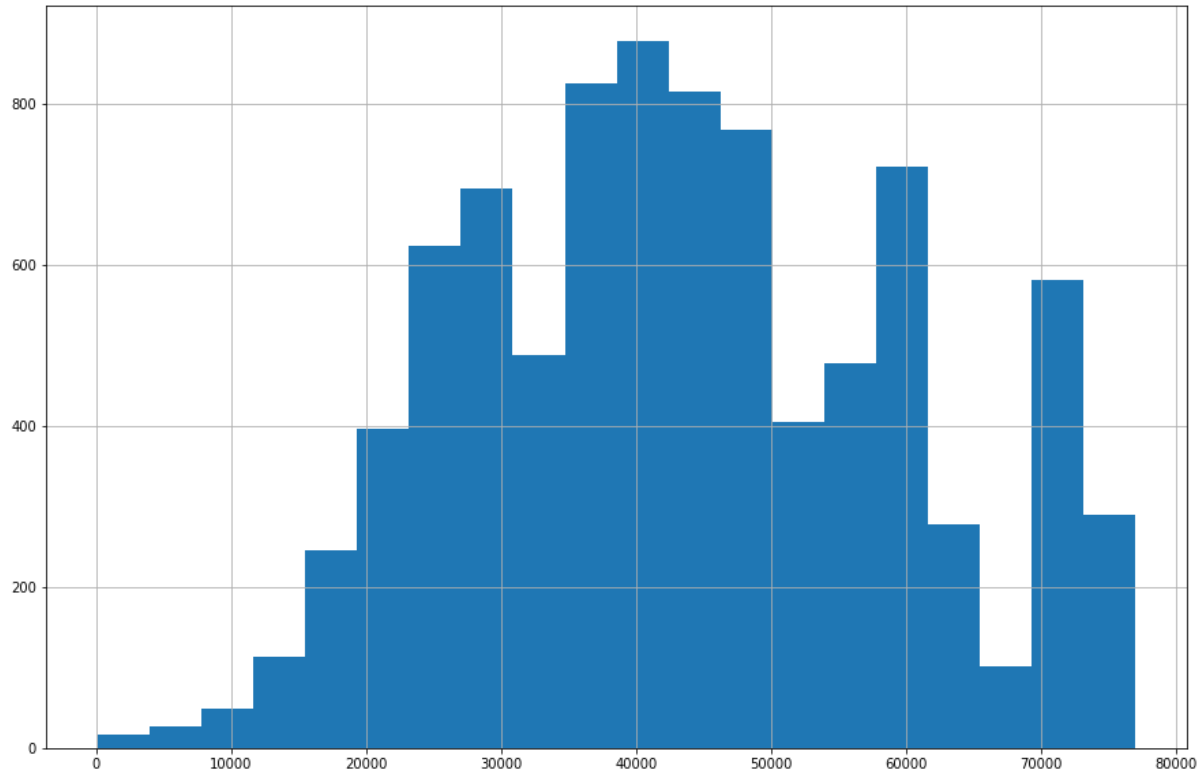


Средняя максимальная зарплата вилки среди заполненных вакансий — около 77 тысяч рублей, однако на графике ниже видно, что большинство вхождений сконцентрировано в районе 50 тысяч рублей, а также наблюдается крупный хвост выбросов. Распределение зарплат в зоне ниже среднего значения напоминает нормальное.

Изображение 13. График распределения максимального значения вилки заработной платы, полное множество

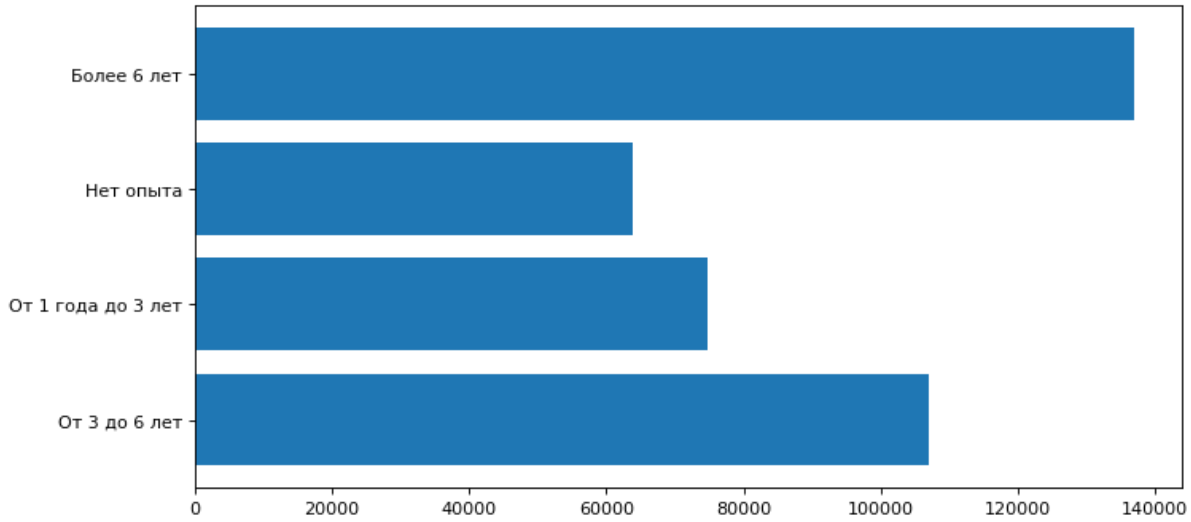


Изображение 14. График распределения максимального значения вилки заработной платы, < 77 тысяч рублей



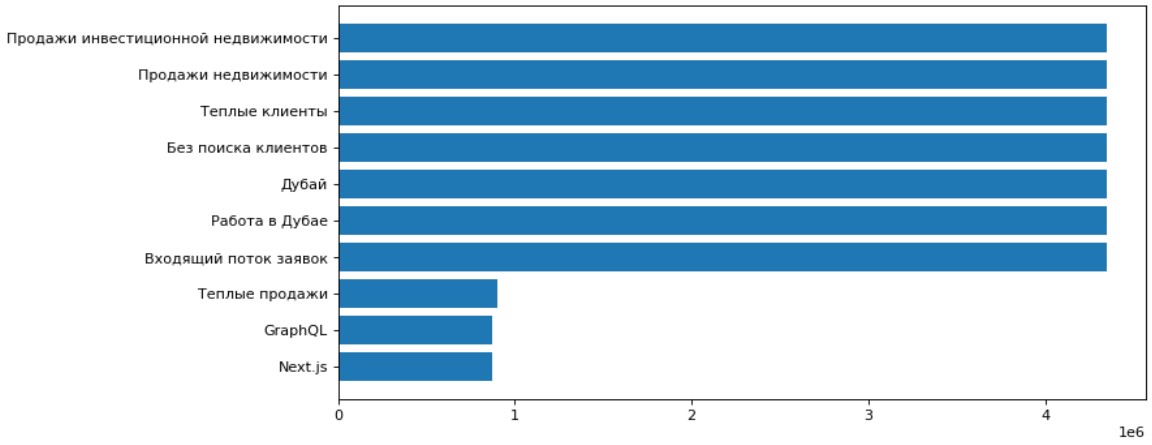
Опыт, согласно ожиданиям, положительно скоррелирован с целевой переменной.

Изображение 15. График отношения средней заработной платы к опыту

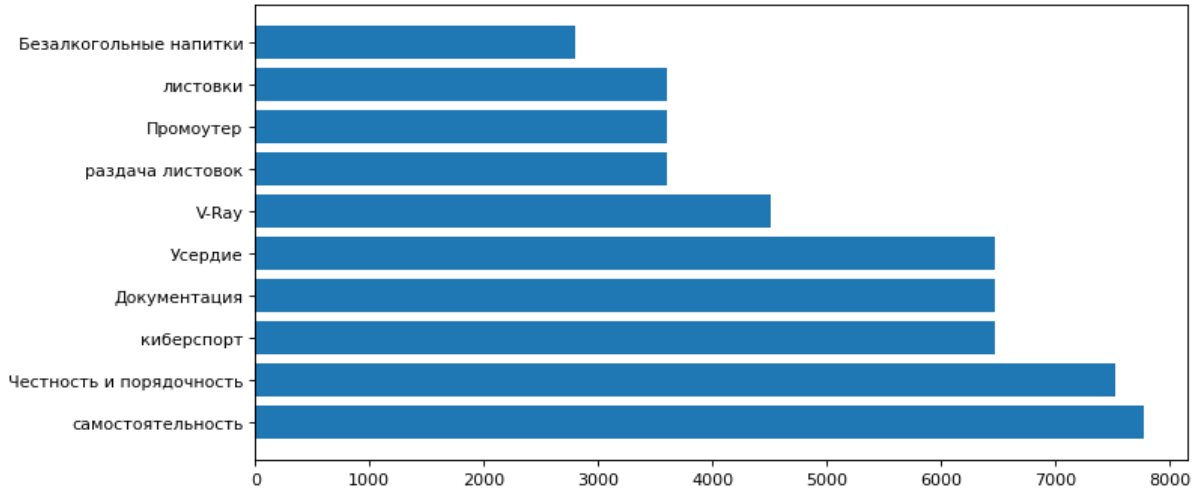


Наиболее оплачиваемые навыки связаны с продажами. Полагаю, это связано с тем, что в подобных вакансиях указывается высокое максимальное значение, зависящее от выполнения соискателями планов по продажам. Также в топе можно заметить навыки, связанные с IT. В тоже время наименее оплачиваемые навыки связаны с промоутерской деятельностью.

Изображение 16. График отношения максимальной заработной платы к навыкам, наиболее оплачиваемые навыки

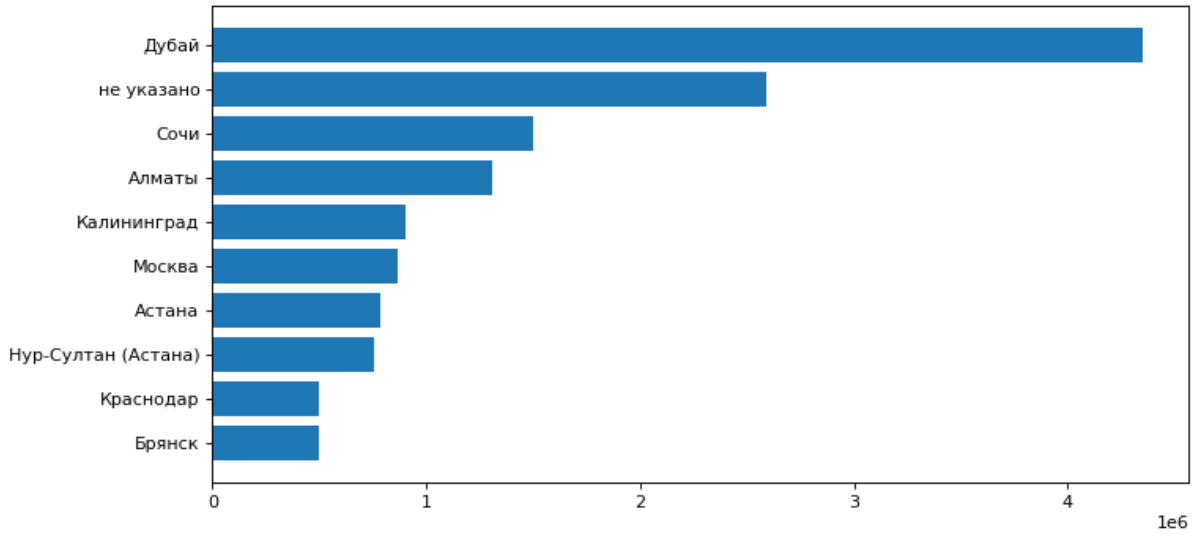


Изображение 17. График отношения максимальной заработной платы к навыкам, наименее оплачиваемые навыки

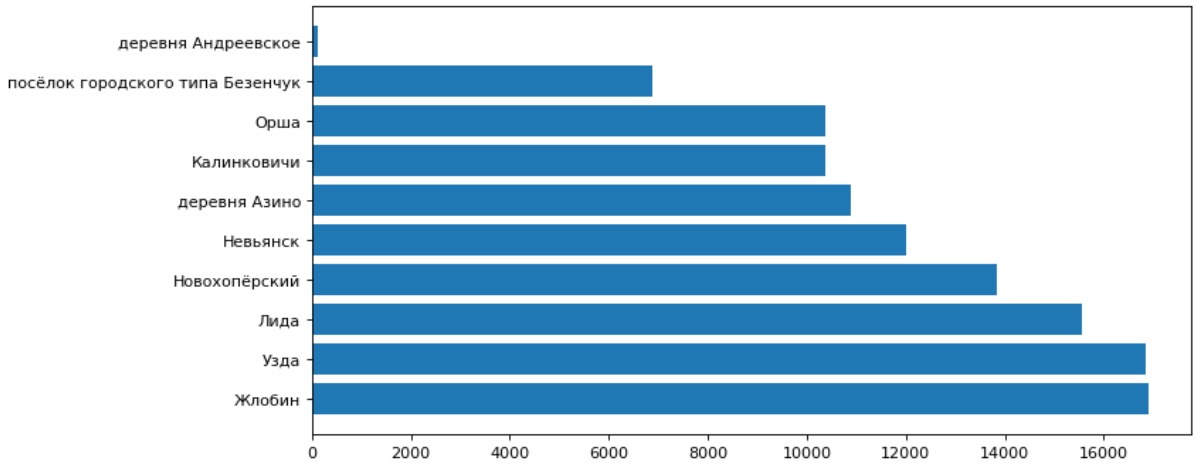


Вполне согласно ожиданиям, зарплаты в крупных городах выше, чем в небольших населенных пунктах. Однако интересно, что в топе оказались вакансии без указания локации, что сложно связать с предыдущими данными (низкая распространенность удаленной работы и превалирование низкоквалифицированных позиций).

Изображение 17. График отношения максимальной заработной платы к локации, наиболее оплачиваемые локации



Изображение 17. График отношения максимальной заработной платы к локации, наименее оплачиваемые локации



3.5. Промежуточные выводы

Анализ предварительной выборки позволяет сделать следующие выводы:

1. Целевая переменная — заработная плата — распределена нормально в диапазоне ниже среднего, но имеет огромный хвост выбросов. Для обучения моделей прогноза заработной платы следует прологарифмировать целевую переменную, как это было сделано в исследовании [3].

2. Два типа описаний, представленных в модели данных вакансий, очень схожи друг с другом по содержанию. К тому же, брендированные описания заполнены достаточно редко. Считаю, что будет целесообразно отказаться от брендированных описании в последующем анализе при помощи моделей машинного обучения.

3. Ряд полей метаданных достаточно сильно скоррелированы с целевой переменной, такие как навыки, локация и опыт работы. Возможно, будет целесообразно их включение в обучающую выборку.

4. Ссылки

[1] — Официальное API портала HeadHunter.ru: https://api.hh.ru/openapi/redoc

[2] — Модель данных документа вакансии API портала HeadHunter.ru: https://github.com/hhru/api/blob/master/docs/vacancies.md

[3] — Jackman, S., & Reid, G. (2013). Predicting Job Salaries from Text Descriptions [A]. doi: http://dx.doi.org/10.14288/1.0075767

[4] — T. Van Huynh, K. Van Nguyen, N. L. -T. Nguyen and A. G. -T. Nguyen. (2020). "Job Prediction: From Deep Neural Network Models to Applications," 2020 RIVF

International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh City, Vietnam, pp. 1-6, doi: 10.1109/RIVF48685.2020.9140760