Применение метод машинного и глубинного обучения для оценки дохода по вакансиям у интернет рекрутера - HeadHunter

Оглавление:

Постановки и описание задачи.

Поиск работы - необходимая потребность человека. Каждый из нас когда либо искал работу на той или иной площадке. Сейчас, поиск работы происходит на двух популярных площадках: Avito и Head Hunter. Нацелены эти площадки на разный сегмент рабочих.
Avito - включает в себя синие воротнички и нацелена только на них. Работа синих воротничков заключается в физическом труде и не требующая высшего образования, поэтому зарплаты там минимальны. Такие работы обычно оплачиваются почасово или по проделанному объему работ. В данной работе фокус направлен на рекрутинговый сайт Head Hunter.
Head Hunter, он же HH, главная площадка для поиска и найма белых воротников. Белые воротнички - это офисные рабочие, которые тесно связаны с умственном трудом. Так как таких профессий становится с каждым годом все больше, а указывать доход на вакансиях стали все меньше - решено научиться предсказывать доход по вакансиям, основываясь на их описания, параметры и других информативные поля.
Делается это для того, чтобы каждый пользователь, ищущий работу, мог ориентировочно понимать на какой доход стоит рассчитывать и не обесценивал или не завышал предполагаемый доход на собеседовании.

Поэтому, целью данной работы является:

a) Скачать / распарсить данные из Head Hunter. Для этого у них существует готовый фреймворк - https://dev.hh.ru/
b) Провести EDA (Exploratory Data Analysis), или же разведывательный анализ данных, где требуется сгенерировать дополнительные

переменные, заполнить пропуски если они имеются, исследовать целевой признак и выбрать метрику качества.

c) Решить задачу регрессии - так как таргетом является доход на вакансии.

d) Выбрать наилучшую модель::
   i) LinearRegression
   ii) KNeighborsRegressor
   iii) RandomForestRegressor
   iv) CatBoostRegressor
   v) LGBMRegressor
e) Написать векторное представление описания вакансии с помощью глубинного обучения.
f) Обернуть в сервис, где пользователь мог бы вставить ссылку на вакансию, а модель выдать вердикт по возможной заработной плате.

Обзор существующих подходов.

(позже напишу)

1) https://www.researchgate.net/publication/362280362_Salary_Prediction_in_Data_Science_Field_Using_Specialized_Skills_and_Job_Benefits_-A_Literature_Review
2) https://www.atlantis-press.com/journals/ijcis/25899235/view

Сборка и описание данных:

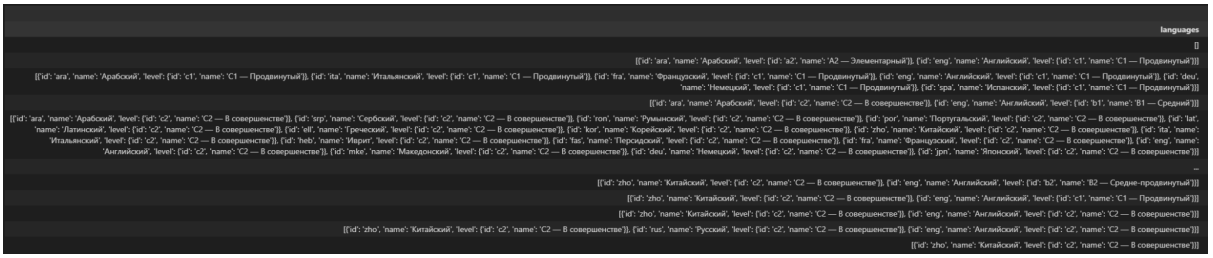
Для сбора данных вакансий Head Hunter написано несколько скриптов на языке программирования Python. Первый скрипт находил существующие профессии в Head Hunter. Делалось это для того, чтобы обойти их ограничение в лимите 2 тыс. вакансий на один запрос. Так как первый скрипт нашел 174 профессии, а одна профессия равна одному запросу, то, в лучшем случае, возможно собрать выборку из 348 тыс. вакансии, по 2 тыс. на каждую профессию. В нашем случае, найдено 228 тыс. объявлений из-за того, что не каждая профессия имеет 2 тыс. вакансий. После того, как была собрана выборка из 228 тыс. вакансий, она отправлялась на вход второму скрипту, который по уникальному ключу доставал все данные по вакансии и сохранял отдельный json файл.

Какие бесплатные, доступные через API HH, данные может содержать вакансия:

- company_vacancies_url - ссылка на компанию, предоставляющая данную вакансию. С помощью нее можно собрать статистику по другим ее объявлениям, например минимальную и максимальную заработную плату на вакансиях, где они указали доход;
- has_premium - премиум вакансия. Гарантируется больше всего откликов;

- billing_type - тариф у объявления. Может принимать три значения: Премиум, Стандарт, Стандарт Плюс;
- schedule - график работы;
- name - название вакансии, заголовок;
- area - город;
- salary_from - нижний лимит заработной вилки;
- salary_to - верхний лимит заработной вилки;
- salary_currency - валюта заработной платы;
- is_gross - учтен ли налог на доход;
- address - адрес работы. Включает в себя также словарь ближайших метро;
- allow_messages - разрешение на отправку сообщений при успешном отклике;
- experience - требуемый опыт работы;
- department - департамент. Проставляется только у крупных компаний;
- description - описание вакансии;
- key_skills - хэштеги требуемых навыков. Может быть неограниченное количество на вакансии;
- accept_handicapped - допускаются люди с инвалидностью;
- accept_kids - допускаются люди с детьми;
- professional_roles - профессия;
- employer - название компании;
- published_at - дата публикации;
- accept_temporary - временная работа;
- languages - требуемое знание языков.

Генерация дополнительных переменных:
В большинстве случаев, поля, приходящие через api.hh, находятся в сыром виде. Например переменная languages:



Поэтому была проведена тщательная фильтрация и генерация таких переменных:

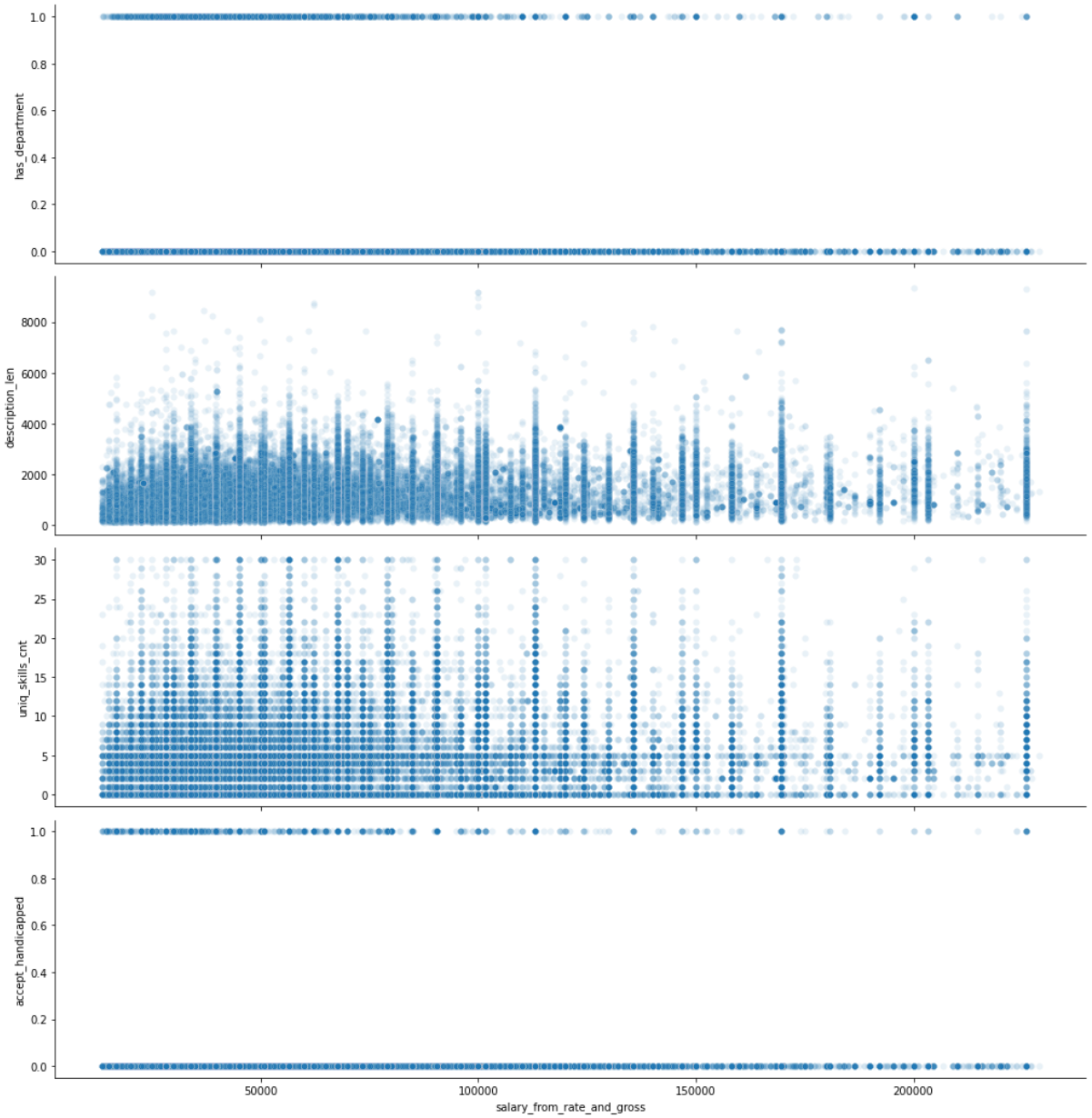
- address, переменная station_embedding. Так как adress содержал массив словарей ближайших станций метро, то был сгенерирован эмбеддинг станций, где существовало только 2 значения - 0 или 1.

- address, переменная line_embedding. Так как adress содержал массив словарей ближайших линий метро, то был сгенерирован эмбеддинг линий метро, где существовало только 2 значения - 0 или 1.
- address, переменная lat. Широта указанного адреса.
- address, переменная lng. Долгота указанного адреса.
- description. Удалена html разметка внутри текста. Текст приведен к нижнему регистру, удалены знаки препинания.
- key_skills, переменная key_skills_embedding. Так как уникальных навыков составляло 29 тыс., было принято решение отобрать топ-500 самых встречающихся и сделать из них эмбеддинг. Также были созданы переменные uniq_skills_cnt - количество навыков указанных в вакансии и uniq_popular_skills_cnt - количество навыков указанных в вакансии и входящие в топ-500.
- dollar_rate. Для каждой вакансии был найден курс доллара на дату публикации published_at, так как зарплата сильно скоррелированы с курсом доллара.
- languages. Подсчитано количество языков, требуемых на вакансии cnt_lang, и сгенерированы несколько флагов:
    - is_eng - требуется английский;
    - is_chi - требуется китайский;
    - is_ger - требуется английский;

Сборка таргета.
В нашей задаче у вакансий зарплата может быть указана в разной валюте, с учетом налогов или без. Поэтому заработная плата была приведена к единому виду: доход переведен в рубли по курсу доллара/евро на дату публикации объявления, плюс добавлен налог 13%, если переменная is_gross = False.

EDA.
После сбора данных и агрегации переменных, был проведен разведывательный анализ данных, где из рис.1 можно увидеть, что:

a) У департаментов стабильно зарплаты до 100 тыс.руб;
b) Людям с инвалидностью платят в разы меньше;
c) Видно, что компании с маленькой зарплатой пытаются как можно больше написать в описании вакансии, чтобы завлечь соискателей. Далее идет нисходящий тренд вниз по количеству символов в описании, и после резкий подъем, когда пытаются найти на крупную зарплату. Скорее всего, в области крупных зарплат компании ищут грамотных специалистов, поэтому в описание расписывают о бонусах, привилегиях и тд.
d) Тоже самое касается и количества указанных навыков в вакансии.

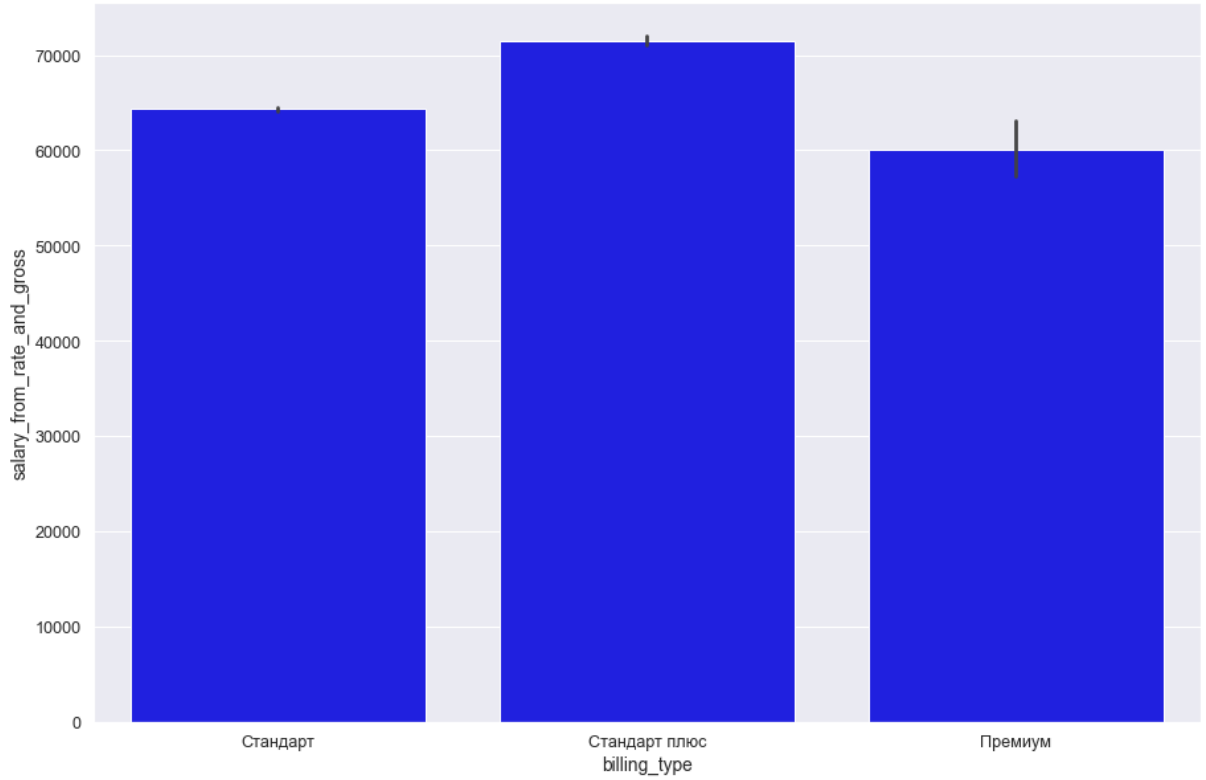


(рис.1)

Ни на одной переменной не прослеживается линейная зависимость с таргетом. Поэтому рациональнее всего будет использовать нелинейные модели для данной задачи.

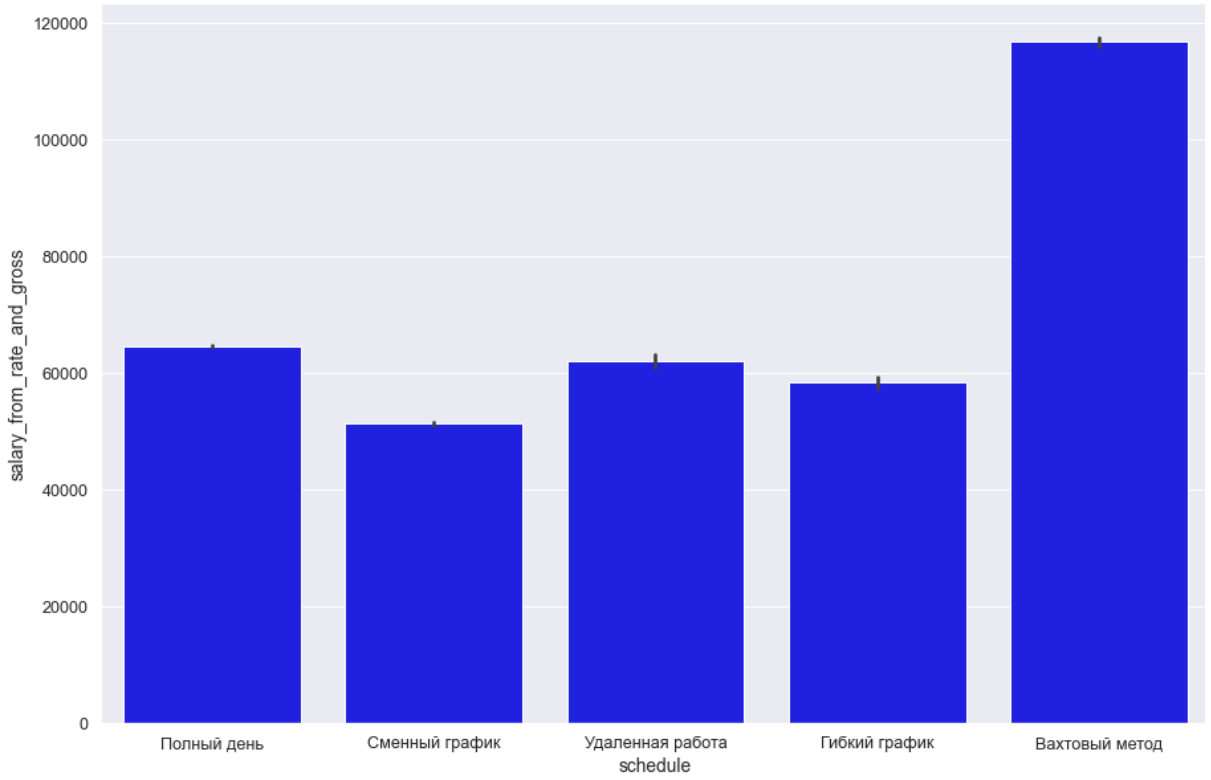
Выше были проанализированы вещественные данные, но взглянем также на категориальные:

1) Скорее всего, премиумом пользуются только тогда, когда вакансия не особо релевантна на рынке профессий, например по зарплате, из-за чего компании начинают платить больше самой площадке, чтобы та поднимала их вакансию выше, тем самым накручивая просмотры вакансии и, соответственно, отклики. .

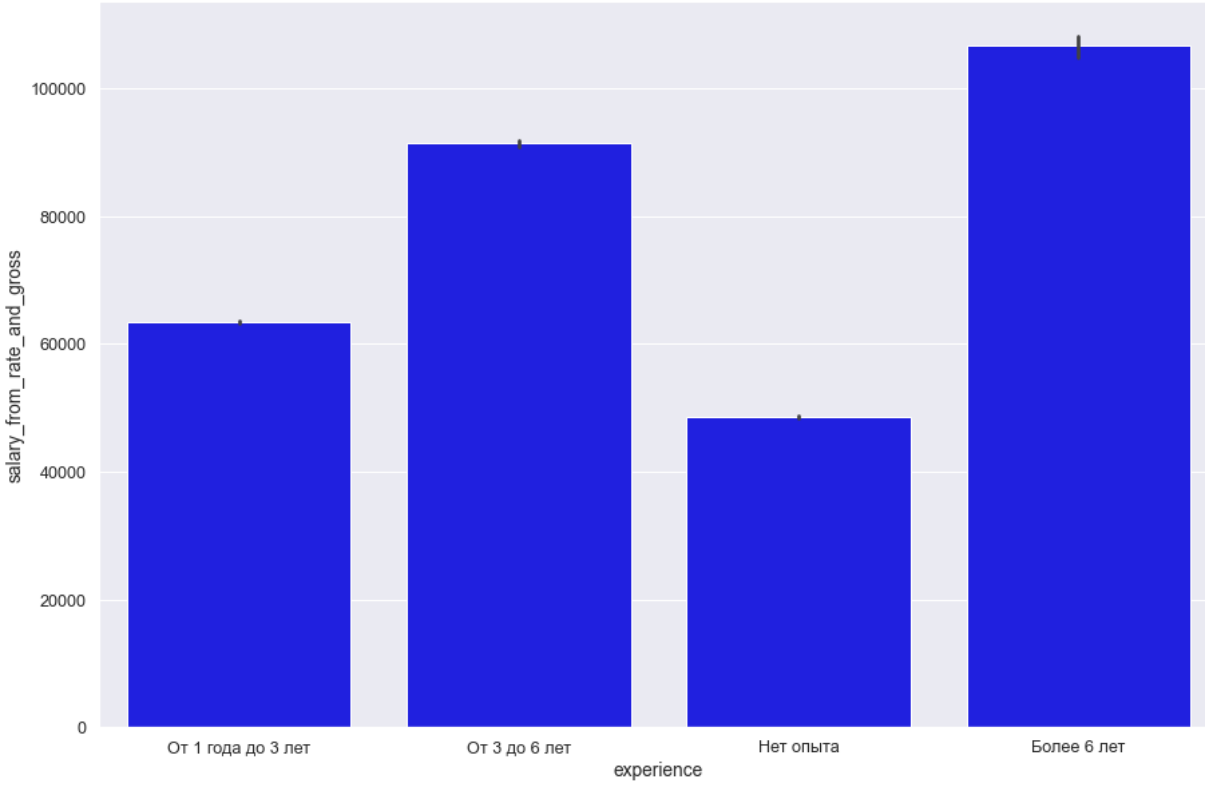


2) Вахтовикам платят гораздо больше остальных, но и работа у них, обычно, связано с тяжелым трудом и требует командировки в другой

город. Остальные же предлагают почти одинаковый доход.

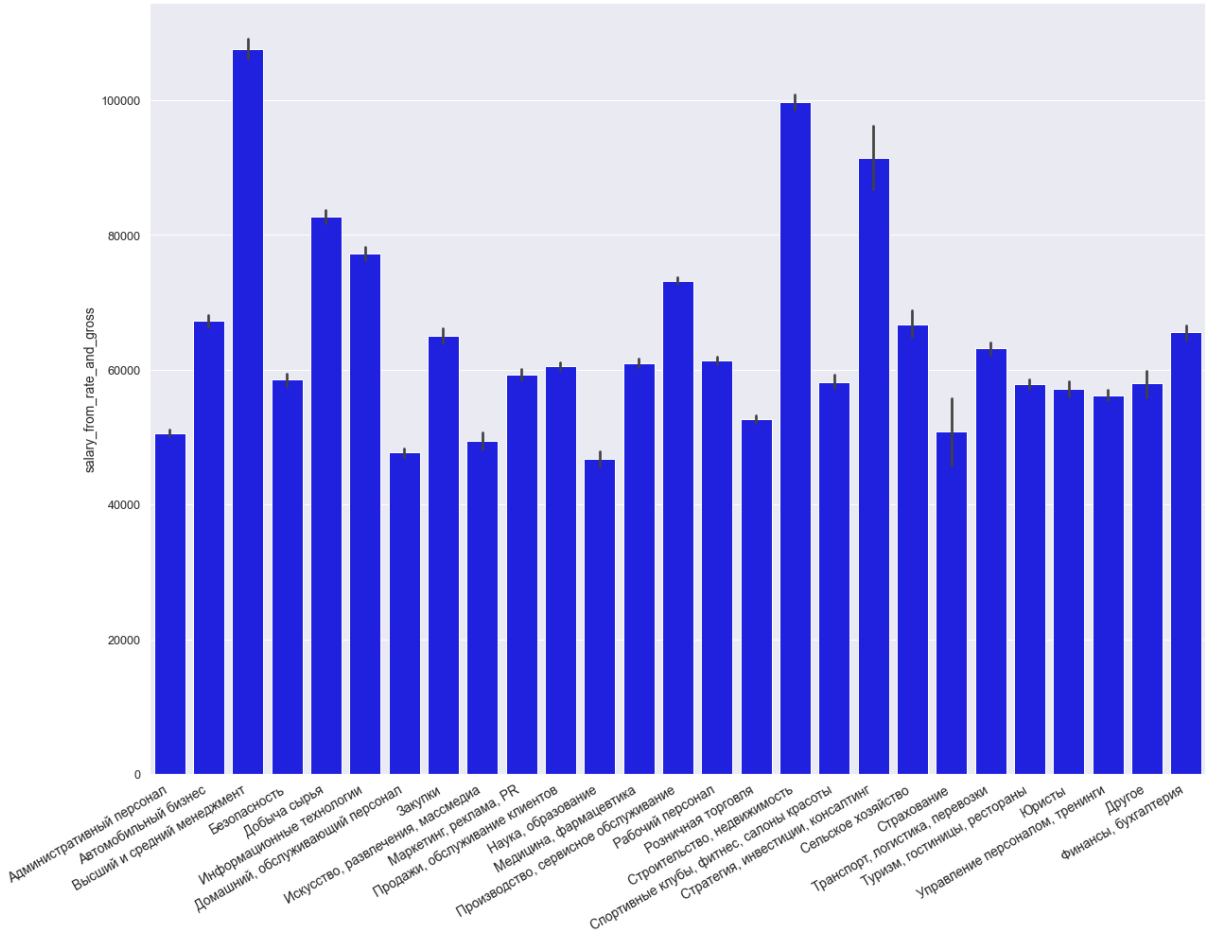


3) С опытом работы всё предельно логично. Чем больше опыт, тем выше зарплату предлагают компании

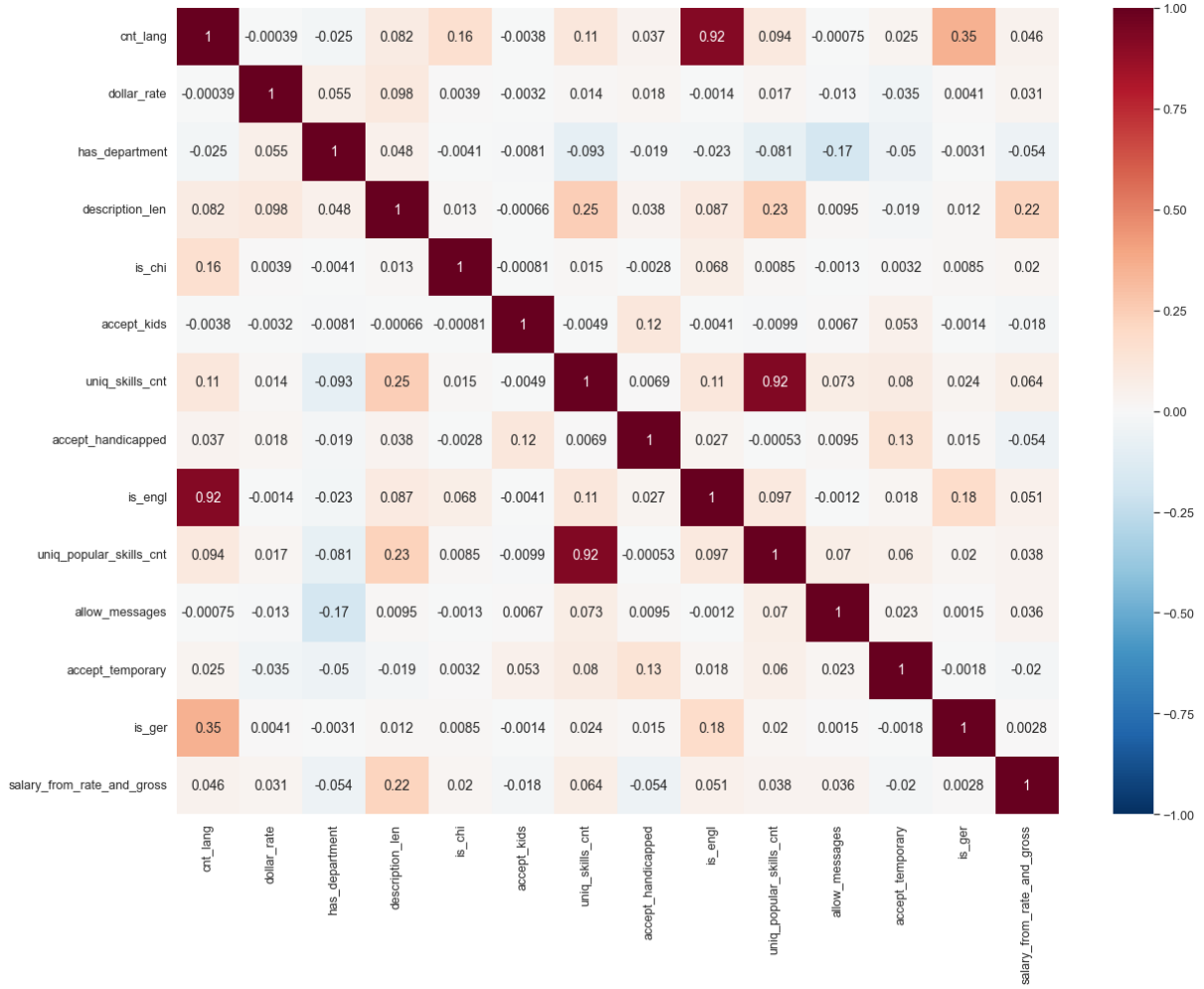


4) На данном графике можно увидеть, что самые высокие зарплаты предлагают строителям, менеджменту и инвестициям, консалтингу. Меньше всего за труд предлагают домашнему персоналу и образованию,

как бы грустно это не звучало.



Также было решено посмотреть корреляции переменных, чтобы пренебречь себя от возможных ликов в данных или схожесть с таргетом



Но сильных корреляций с таргетом salary_from_rate_and_gross выявлено не было. Можно наблюдать подтвержающий факт, описанный выше, при построении диаграммы рассеиваний, что длинна символов в описании вакансии влияет на зарплату на вакансии.