Общая концепция модели

В концепции решения мы предлагаем создать отдельную модель для каждой специализации. При большом количестве параметром модель обучается хуже. Если мы разделим модель на несколько (каждая из которых будет предсказывать результат для определенной профессии), то мы добьемся увеличения точности модели, потому что она будет учитывать только навыки, релевантные только для данной специальности. Модель машинного обучения, примененная в данном решении основана на комбинации CatBoost и word2vec. Кроме того, в случае разных моделей мы можем изменять и модель, которая анализирует описание вакансии, обращая внимание на те или иные признаки.

Включение дополнительной информации о компании также важно. Основные параметры, которые надо добавить это размер компании (прибыль и количество работников), расположение офиса, деятельность компании и престижность компании. Размер компании может определять ее стабильность. Многим людям важна стабильность, поэтому они различают большие и небольшие компании. Деятельность компании и ее престиж тоже могут являться важными составляющими при выборе работы и оказывать влияние на заработную плату. В итоге, обучение модели на характеристиках компании может улучшить качетсво предсказанных данных.

Сбор данных

Для внедрения в реальные бизнес-процессы очень важна актуальность результатов, которые выдает модель. Рынок труда стремительно меняется. Меняются зарплаты, исчезают старые профессии и появляются новые буквально каждый год. Поэтому для повышения эффективности модели важно, чтобы она обучалась на актуальных данных. Именно поэтому мы уделили внимание алгоритмам, которые позволяют нам эффективно собирать данные из интернета.

Для начала, мы собрали данные о вакансиях с сайта HeadHanters по различным профессиям. Для сбора данных мы использовали библиотеку BeautifulSoup и анализировали html код. Это получалось намного быстрее и эффективней более сложных методов, потому что скорость сбора данных ограничивалась только скоростью интернета.

Сбор данных проходил в 3 основные фазы. Первая фаза - определение списка интересующих нас вакансий и корректная транслитерация на английский. После этого, формировались поисковые запросы, по которым собирались ссылки на вакансии. Эти ссылки сохранялись в разные файлы в зависимости от поискового запроса, что в дальнейшем упрощало определение профессии. Третья фаза заключалась в парсинге данных с конкретных вакансий. В конце, полученные ссылки сохранялись по отдельным файлам. Полученные таблицы с вакансиями в csv файлах. Более подробный алгоритм описан в приложенном ноутбуке.

При сборе данных в вакансиях важно было собрать все возможные данные. Основное, что было получено в результате поиска это зарплата, профессия, описание вакансии, ключевые навыки, график работы. Похожие данные были даны в тестовом датасете,

поэтому можно легко распространить построенную нами модель и на эти данные. В дальнейшем, помимо информации со страницы hh можно собирать информацию о компании. Например, размер компании может влиять на зарплату. Также, расположение офиса, репутация и род деятельности компании тоже могут быть важными факторами, которые в будущем можно будет использовать. Однако этой информации нет на hh, поэтому эти данные надо собирать со сторонних ресурсов. Таким образом, информация о компании может быть собрана с других источников и использована для улучшения предсказаний модели.

Обработка данных

Колонки, имеющие категориальные данные были приведены в вид, удобный для обучения модели: опыт, специализация, описание, расписания. В концепции, при разделении по профессиям, будет возможно более глубокое разделение по специальностям. Опыт и расписание были переведены в числовую характеристику. Таким образом, основные данные были разбиты по категориям и переведены в вид, удобный для обучения.

С помощью библиотеки gensim был проведен анализ описания и выделенны ключевые навыки. Сначала были выделенны ключевые навыки, которые встречаются в описаниях. Это довольно большой список из 30-40 значений. Однако, в этот список не были включены профессиональные навыки, потому что эта переменная и так учитывается в специализации. После, весь текст описания был разбит на слова, переведенные в формат, предназначенный для анализа и удалены знаки препинания. После этого искались семантические сходства между словами из описания и ключевыми навыками. Если сходство обнаруживалось, то мы считали, что мы нашли ключевой навык. В конце все навыки группировались по 8 основным категориям: креативность, продуктивность (результативность, выносливость), прилежность (внимательность, аккуратность), лидерские качества (инициативность, самостоятельность), компьютерные навыки, интеллект, коммуникабельность, языковые знания. Таком образом, составлялся вектор необходимых навыков, который можно использовать в обучении.

Однако анализ обучающей выборки не был проведен из-за того, что не хватило вычислительных мощностей, которые бы позволили обработать такое количество информации. Тем не менее, алгоритм работает, и его можно посмотреть в приложенном ноутбуке. В итоге, обработка данных заключалась в 3 основных шагах - приведение данных в категориальный вид, определение наиболее релевантных колонок и выделение ключевых навыков из текста описания.

Модель

В ходе разработки решения была предложена модель, которая использует регрессор CatBoost для обработки категориальных данных. Для текстовых данных было решено использовать word2vec для оценки схожести запрашиваемых навыков в тексте вакансии с набором данных, на которых была обучена модель, для более лучшей оценки заработной платы.

Внедрение

На базе этой модели можно создать несколько различных бизнес продуктов. Один из основных это помощь в выборе заработной платы. В маленьких компаниях при найме нового сотрудника трудно оценить зп, которую надо платить. Это ведет к большему сроку рекрутинга или финансовым потерям для фирмы. Создание же продукта, который бы советовал hr зп решает эту проблему.

Другое применение этой модели это помощь в определение желаемой зп для работников. Зачастую, человек, который не имеет опыта работы не представляет, какой должна быть его заработная плата. Это может приводить к тому, что работнику труднее найти работу, или же он соглашается на меньшую зп, нежели установилась на рынке. Модель, предсказывающая по его навыкам и специальности его зп поможет ему ориентироваться на рынке без самостоятельного анализа.

В итоге, улучшение информированности на рынке труда поможет быстрее находить работу или наоборот, работников и, возможно, снизит уровень естественной безработицы. Более конкретное применение возможно в виде создания веб-сервиса, интегрированного в различные сервисы по рекрутингу, такие как HeadHanters.