Оценка вакансий сотрудников для ранжирования по уровню заработной платы в поисковой выдаче

Осипов А.В., кандидат физико-математических наук, доцент кафедры «Прикладная информатика», ФГБОУ ВО «Московский политехнический университет», Москва, Россия

Суворов С.В., кандидат экономических наук, доцент кафедры «Прикладная информатика», ФГБОУ ВО «Московский политехнический университет», Москва, Россия

Жиляева И.А., кандидат экономических наук, доцент кафедры «Прикладная информатика», ФГБОУ ВО «Московский политехнический университет», Москва, Россия

Черников А.С., магистрант кафедры «Прикладная информатика», ФГБОУ ВО «Московский политехнический университет», Москва, Россия

Аннотация. Данная работа посвящена анализу вакансий с использованием технологий больших данных для улучшения поисковой выдачи организаций, занимающихся интернет-рекрутментом. Рассматривается проблема предложения соискателям наиболее релевантной поисковой выдачи с точки зрения учета заработной платы в условиях, когда работодатель не указывает уровень зарплаты в вакансиях, что является весьма распространённым явлением на рынке труда. В этой статье проведен анализ собранных реальных вакансий, предложен подход к сортировке вакансий по уровню заработной платы в отсутствие указанной вилки, проведена оценка качества такой сортировки. Предложенный метод позволяет повысить качество соискателей для работодателей и повысит шансы найти более высоко оплачиваемые вакансии для соискателей, что в свою очередь способствует развитию рынка труда.

Ключевые слова: искусственные нейронные сети, деревья решений, экспертные системы, рынок труда, вакансии, подбор кадров.

Assessment of employees of vacancies for ranking by level of wages in the search results

Osipov A.V., candidate of physico-mathematical sciences, Associate Professor of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia
Suvorov S.V., candidate of economic sciences, Associate Professor of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia
Zhilyaeva I.A., candidate of economic sciences, Associate Professor of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia
Chernickov A.S., magistrate of the «Applied Informatics» Department at the Moscow polytechnic University, Moscow, Russia

Annotation. This work is devoted to the analysis of vacancies using big data technologies to improve the search results of organizations involved in Internet recruitment. The article considers the problem of offering applicants the most relevant search results in terms of payroll accounting in conditions when the employer does not indicate the level of salary in vacancies, which is a very common phenomenon in the labor market. This article analyzes the collected real vacancies, proposes an approach to sorting vacancies according to the level of wages in the absence of the specified fork, and evaluates the quality of such sorting. The proposed method improves the quality of job seekers for employers and increases the chances of finding higher paid jobs for job seekers, which in turn contributes to the development of the labor market.

Keywords: artificial neural networks, expert systems, labor market, vacancy, recruitment, decision trees.

В наши дни, на рынке труда очень востребованным является автоматическая работа с вакансиями и резюме. Учитывая рост распространённости интернета в мире, растёт количество сайтов,

предоставляющих свои услуги для поиска работы и найма сотрудников, а также размер их аудитории. Влияние такого способа набора людей на рынок труда является активной темой исследований, например, этому посвящены работы [1-4]. С точки зрения рекрутёров ведутся исследования для улучшения их работы путём скрининга [6], разработка инструментов с адаптацией к обратной связи [7], развиваются инструменты для автоматической оценки резюме [8], разрабатываются системы для отбора кандидатов и оценки их деятельности [9].

С точки зрения нанимающихся сотрудников появляются системы для рекомендаций вакансий согласно резюме. Различные методы машинного обучения в работе с ванкансиями и резюме используются в работах. В работах рассматриваются несколько методов предсказания заработной платы с использованием алгоритмов машинного обучения. Работа акцентирует внимание на общих предсказаниях, а не на индивидуальных, на которых сосредоточена работа [10]. Такое предсказание требуется для размещающего вакансию человека, когда он не знает точно, какой уровень заработной платы по этой вакансии распространён на рынке труда, для сайта с вакансиями, чтобы иметь возможность ранжировать в поисковой выдаче вакансии без указания заработной платы согласно предсказанию, для работника, устраивающегося на работу, чтобы иметь представление о том уровне зарплаты, на который он может рассчитывать при устройстве на работу в которой нет указанной заработной платы.

Авторы этой работы провели исследование нескольких алгоритмов для предсказания заработной платы. Ими были рассмотрены такие алгоритмы, как SVM, MLP, AdaBoost, Random Forests, логистическая регрессия, метод ближайших соседей и их комбинации для достижения наилучших результатов. К сожалению, авторы этой работы не предоставили ни собранный ими датасет, ни программный код, при помощи которого они проводили свои исследования, поэтому с целью улучшения результатов, полученных в работе [10], был собран независимый датасет из 9851 вакансий на основе данных сайта headhunter, расположенного по адресу hh.ru, а также проведено исследование других

алгоритмов, не освященных в их работе. Все алгоритмы, показавшие хорошие результаты в этом исследовании относятся к алгоритмам вида «черный ящик», т.е. не дающим возможность проанализировать причины, по которым вакансия была отнесена к той или иной категории. Работа сконцентрирована на предсказании больше ли зарплата, чем медианная в регионе, что ограничивает количество предсказываемых групп до 2-ух и позволяет использовать менее точные алгоритмы с удовлетворительным результатом.

Для решения поставленной задачи были проведены следующие действия:

1. Сбор данных. При помощи обращения из разработанного для анализа приложения на языке Python к Headhunter API были собраны вакансии с указанием зарплат IT тематики, а именно те, в которых упомянуты наиболее распространённые языки программирования, такие как 1C, C, C++, Clojure, CSharp, Delphi, Go, Java, JavaScript, Kotlin, ObjectiveC, Pascal, Perl, PHP, Python, Ruby, Scala, Swift, TypeScript. Язык SQL был исключен из этого списка, чтобы не привлекать к рынку программистов системных администраторов, не смотря на это в собранный датасет вошли не только разработчики, но и другие профессии, работающие с этими технологиями, например, бухгалтера, использующие 1C.

2. Очистка данных. Дублированные вакансии были исключены для избавления от конфликтов и снижения проблемы переобучения.

3. Автоматическое вычисление существенных признаков, на основе тегов требуемых навыков, заполненных людьми при составлении вакансий. Поиск таких навыков в описании вакансий, вычисление статистически значимых.

4. Вычисление усреднённой зарплаты, удаление из выборки вакансий, в которых не удалось найти признаков. Сбор данных по средним зарплатам в регионах, приведение всей зарплаты к виду после вычета налогов с учётом курса рубля на момент сбора датасета.

5. Нормализация данных путём исключения экстремальных зарплат, а именно удаление 100 самых больших и 100 самых маленьких зарплат, для

избавления от опечаток и экстремальных значений (например, присутствовала зарплата разработчика на полный день в 120 рублей и зарплата в 600000 рублей для кассира). На этом этапе размер датасета сократился до 8227 вакансий с 9851 уникальной вакансии

6. Выбор модели. Был составлен список из нескольких алгоритмов машинного обучения, не вошедших в работы [10] и, таких как решающие деревья C4.5, кластеризация путём самоорганизующейся карты Кохена, g-means кластеризация, 2-ухи 3-ёх слойные нейронные сети.

7. Обучение модели и её валидация. Модели были обучены и была выполнена кросс-валидация, для проверки их работоспособности.

8. Сравнение моделей. Было проведено на основе данных кросс-валидации и таблиц сопряженности.

9. Наилучшая модель была адаптирована для решения частной задачи сортировки вакансий по уровню предсказанной заработной платы, была выполнена оценка качества такой сортировки

На этапе сбора данных все вакансии были сохранены в виде json файлов, пригодных для дальнейшей обработки, выполненной при помощи модуля json, входящего в поставку языка программирования python3. После этого на основе имеющихся в файлах тегах, были вычисленны наиболее распространённые, а также осуществлен поиск всех тегов в тексте описания вакансий, были отброшены признаки, встретившиеся менее 30 раз на выборку, что соответствует 0,3% принятыми за шум, в работе[10] за шум приняли 0,25% аналогичных признаков. Теги не были обобщены, как это сделали в работе[10], так как в теге может содержаться информация, которая косвенно указывает на особенности работодателя, например, указание вакансии C++ программиста тегом C/C++ вместо тегов C и C++, даёт стойкое снижение заработной платы. Итого получилось 249 признаков. Уровень заработной платы был рассчитан после вычета налогов (headhunter предоставляет флаг признака вычета налогов в API), после чего заработная плата была приведена к среднеарифметическому значению между максимальной и минимальной для каждой вакансии.

Заработная плата, указанная в разной валюте, приведена к рублям, согласно курсу Центробанка РФ, действовавшему на момент сбора вакансий, так как сам датасет собирался в России и по российскому рынку труда. Код региона для каждой вакансии был заменён на среднюю заработную плату по этому региону. Под регионом понимается город или область, указанные работодателями как место работы. Регионы, в которых было найдено меньше 30 вакансий, были исключены из набора данных. Было рассмотрено несколько вариантов представления данных по зарплате для наиболее точного предсказания, такие как представление зарплат в виде отклонений от средних по региону зарплатам в процентах, в рублях или просто представление зарплат в рублях. Забегая вперёд, можно с уверенностью сказать, что наилучший результат дало простое представление зарплат в рублях. Перевод вида занятости из категориального представления в представлении по часам не даёт однозначных результатов, так как нет строго определения графика труда и одна и та же категория может соответствовать разному количеству часов работы, что не просто не повысит точность предсказания, а окажет противоположный эффект, особенно в алгоритмах, хорошо работающих с категориальными данными. К сожалению, в отличии от датасета работы[10], в данных headhunter-а отсутствует или редко встречается информация по требованиям образования, что не дало возможность включить эти данные в признаки. Кроме нормализации данных для нейросетей, было обеспечено отсутствие нулей в обучающих наборах, так отсутствие признака было представлено за 1, а его наличие за 2 или же количество найденных тегов в тексте вакансии + 1, для обеспечения более быстрого обучения (некоторые реализации нейросетей используют логистическую сигмоиду, описываемую формулой $f(x) = \sigma(x) = \frac{1}{1+e^{-x}}$, производная от которой, используемая для настройки нейрона, так же принимает значение 0 в случае, когда $x = 0$, что не даёт возможность аппроксимации тождественной функции около начала координат).

Для улучшения предсказаний и уменьшения времени обучения, необходимо прибегнуть к кластеризации. Для выявления похожих вакансий с

целью улучшения результатов предсказания заработной платы были рассмотрены 3 алгоритма классификации, а именно k-means, g-means и самоорганизующаяся карта кохена. Для выявления признаков, по которым следует проводить классификацию, был проведён корреляционный анализ при помощи коэффициента корреляции Пирсона. Наибольшая корреляция была выявлена от опыта работы (0,469) и от средней заработной платы в регионе (0,435). Ближайшая к ним корреляция по тегу «oktell», относящемуся к навыкам, найденным в описании вакансий, составила (0,231). Учитывая уменьшение корреляции по вакансиям до 0,162 на 20 позиции от максимальной корреляции к минимальной, логичнее всего использовать для кластеризации признаки, относящиеся к зарплате, а именно: опыт работы, среднюю заработную плату по региону и тип занятости (коэффициент корреляции для него 0,090). Наилучшие результаты показали все алгоритмы предсказания вакансий при алгоритме кластеризации k-means с использованием всего 2-ух кластеров. Повышение количества кластеров приводило к ухудшению точности предсказания. Хуже всего показала себя кластеризация при помощи алгоритма g-means. Для определения количества кластеров был использован «Метод локтя», позволяющий оценить оптимальное количество кластеров на основе изменения разброса с изменением числа кластеров.

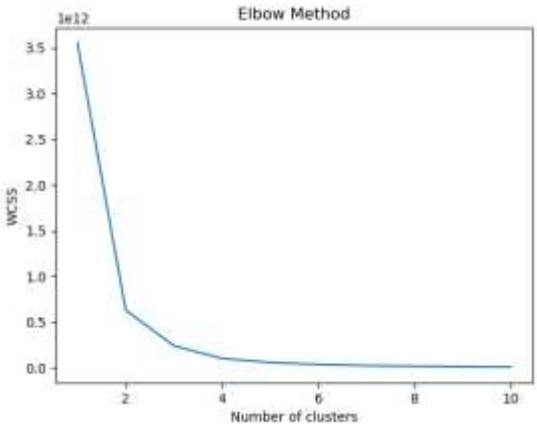

Рис. 1 – Изменение разброса с изменением числа групп разбиения

Кластеризация с использованием карт Кохена показала сопоставимое качество с кластеризацией по алгоритму k-means, но заняла гораздо большее время, однако, с её помощью возможно построить наглядные для аналитика

карты кластеров, по которым, например, просматривается сильная зависимость между предсказываемой зарплатой, средней зарплатой в регионе и опытом работы.

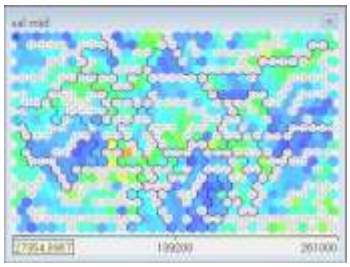

Рис. 2 – Карта Кохена для предсказываемой зарплаты

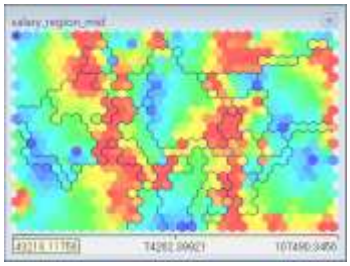

Рис. 3 – Карта Кохена со средними зарплатами по регионам

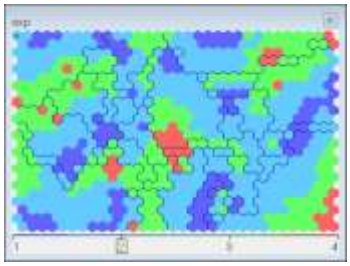

Рис. 4 – Карта Кохена с опытом работы

Для предсказания заработной платы, как правило, необходимо предсказать к какому интервалу относится заработная плата. Разбиение зарплаты на интервалы приводит задачу к задаче классификации и позволяет достигнуть большей точности[10]. Было принято разбить уровень зарплаты на 5 категорий, для чего необходимо воспользоваться квантованием данных:

1. Очень низкая (до 30000 рублей)
2. Низкая (от 30000 до 60000 рублей)
3. Средняя (от 60000 до 10000 рублей)
4. Высокая (от 100000 до 180000 рублей)
5. Очень высокая (от 180000 рублей)

Подобное разбиение помогает оценивать уровень заработной платы, согласно удобным для человека критериям. Такая оценка может использоваться для ранжирования выдачи вакансий без указания заработной платы по предполагаемому уровню, для помощи работодателю при составлении вакансии, для помощи нанимающемуся человеку, для налоговой службы с целью обратить внимание на должности, по которым заявлен более низкий уровень заработной платы, чем прогнозируемый (определение ухода от налогов путём выплаты «серой» зарплаты).

Для выявления оптимального алгоритма, для задачи сортировки данных, особенно в условиях, когда необходимо знать категории уровня средних заработных плат, необходимо уметь предсказывать уровень зарплаты соответствующей каждой вакансии. Как и у авторов работы [10], различные конфигурации многослойной искусственной нейронной сети показали посредственные результаты, не превышающие 55% точности предсказаний, что согласуется с их выводами, касательно актуальности применения нейросетей в данной задаче. Алгоритм C4.5 показал плохие результаты, в плане точности 48.36% и отличные результаты в плане производительности. Выбор k-means или карт Кохена для кластеризации не оказал значительного влияния на точность окончательного предсказания. Основные параметры точности такой модели представлены ниже.

Таблица 1

Сопряженность алгоритма C4.5

	Классифицировано					
Фактически	**До 30000**	**От 30000 до 60000**	**От 60000 до 100000**	**От 100000 до 180000**	**От 180000**	**Итого**
До 30000	419	184	64	14		678
От 30000 до 60000	104	1493	295	40	8	2089
От 60000 до 100000	49	319	1681	11	12	2525
От 100000 до 180000	21	130	291	2484	10	2545
От 180000	2	10	31		99	390
Итого	595	2136	2362	2549	129	8227

Оценка точности по методологии macro-averaging составляет 0,646 для данной модели, а полноты 0,771. Сбалансированная F-мера для данной модели, соответственно составляет 0,703

Хотя, такой алгоритм значительно проигрывает алгоритму Random Forests, показавшему отличные результаты в работе [10], дерево алгоритма C4.5 наглядно для понимания человеком, что позволяет аналитикам сказать, по какой причине, та или иная вакансия была отнесена к той или иной категории, что даёт больше возможностей для выявления значимых факторов.

Алгоритм Random Forest, показавший хорошие результаты в работах [10] показал наилучшие результаты из всех рассмотренных алгоритмов с точностью предсказаний на контрольной выборке в 75,63%. Точность получилась ниже, чем в работе [10] в первую очередь из-за большего числа групп разбиений в 5 интервалов против 4-ёх. Однако, в таком виде этот алгоритм не подходит для задачи ранжирования вакансий по предполагаемой средней зарплате в поисковой выдаче.

Для оценки качества поисковой выдачи был использован стандартный разброс, рассчитывающийся, как сумма отклонений элементов в списке от их идеальной позиции, делённая на количество элементов в списке. Очевидно, что в случае случайной неупорядоченной выборки эта величина будет стремиться к среднеквадратичному отклонению, приблизительно равному 0,34 для нормального распределения, а в случае идеального порядка следования эта величина примет значение 0. Таким образом, чем меньше эта величина, тем лучше, однако, следует понимать, что на практике она не может быть равной 0, хотя бы из-за того, что не учитывается порядок выдачи одинаковых заработных плат. На практике, для собранного датасета, в тестовой выборке, идеальная сортировка со случайным порядком следования одинаковых зарплат дала значение этой величины 0,01.

Было рассмотрено 3 способа ранжирования вакансий по предсказанным средним заработным платам при помощи алгоритма Random Forest. Первый способ – выполнить ранжирование по результатам работы алгоритма,

описанного в работах [10] и [5]. Второй способ – выполнить ранжирование по результатам работы алгоритма Random Forest не используя квантование, к которому прибегали в этих работах для повышения точности. Третий способ – комбинация первого и второго, как минимум, в 2 раза более затратная по ресурсам компьютера, но гарантирующая большую точность попадания вакансии в заданные диапазоны, т.е. сортировка внутри диапазонов согласно результатам предсказания модели, предсказывающей непрерывную заработную плату. С результатами такого ранжирования можно ознакомиться в таблице 2.

Таблица 2

Сравнение разных моделей сортировки вакансий по уровню заработной платы

	Стандартный разброс	**Точность**	**Средняя ошибка**
Модель 1	0,26416	75,63%	0,59
Модель 2	0,17714	60,54%	27817,49 рублей
Модель 3	0,17927	75,63%	-

Таким образом, модель в том виде, в котором она используется для предсказания диапазона зарплат плохо подходит для ранжирования вакансий. Между второй и третьей моделью небольшая разница в стандартном разбросе и вторая модель побеждает с незначительным отрывом, однако, она имеет пониженную точность, хотя это в большинстве случаев не критично. Таким образом, при высокой нагрузке, предпочтительнее использовать вторую модель, когда третью можно использовать в случае если необходима оценка шанса попадания вакансии в ту или иную категорию предсказанных зарплат. Удаление экстремально больших и экстремально малых вакансий из обучающей выборки положительно сказывается на точности предсказаний, даже если предположить, что все вакансии, убранные таким образом, будут классифицированы неправильно (прирост точности при удалении 2,4% вакансий составил 10%). Однако, это крайне негативно сказалось на точности нейросетей, упавшей с 54 до 45%, таким образом, логично предположить, что для их использования необходимо ещё больше данных, что невозможно для многих специальностей в связи с их востребованностью. Выводы авторов работы [10], касательно более хорошего поведения алгоритмов с

использованием решающих деревьев были подтверждены. Использование информации о среднем уровне зарплаты в регионе позволило улучшить точность предсказаний. Данная информация была установлена как линейно влияющая на уровень заработной платы и оказалась вторым по значимости признаком, незначительно уступившим опыту работы. Для более точного предсказания вакансий в регионах с незначительной информацией на рынке труда, вероятно, стоит использовать какие-либо методы предсказания средней заработной платы на основе соседних регионов, что в свою очередь заслуживает отдельного исследования.

Библиографический список

1. Anna B. Holm. E-recruitment: Towards an ubiquitousrecruitment process and candidate relationship man-agement.German Journal of Human Resource Man-agement, 26(3):241–259, 2012.
2. Lori Foster Thompson, Phillip W. Braddy, and Karl L.Wuensch. E-recruitment and the benefits of organi-zational web appeal.Computers in Human Behavior,24(5):2384 – 2398, 2008. Including the Special Issue:Internet Empowerment.
3. Peter Kuhn and Hani Mansour.Is internet jobsearch still ineffective?The Economic Journal,124(581):1213–1233, 2014.
4. Constantin Mang. Online job search and matchingquality. Technical report, Ifo Working Paper, 2012.
5. Archit Khosla "Job Salary Prediction" [Электронный ресурс] //University of California San Diego URL: http://cseweb.ucsd.edu/classes/sp15/cse190-c/reports/sp15/012.pdf (дата обращения: 26.09.2019).
6. Amit Singh, Catherine Rose, Karthik Visweswariah,Vijil Chenthamarakshan, and Nandakishore Kamb-hatla. Prospect: A system for screening candidates forrecruitment. InProceedings of the 19th ACM Inter-national

Conference on Information and KnowledgeManagement, CIKM ’10, pages 659–668, New York,NY, USA, 2010. ACM.

7. Evanthia Faliagka, Athanasios Tsakalidis, and Gian-nis Tzimas. An integrated e-recruitment system forautomated personality mining and applicant ranking.Internet Research, 22(5):551–568, 2012.

8. Evanthia Faliagka, Konstantinos Ramantas, Athana-sios K Tsakalidis, Manolis Viennas, Eleanna Kafeza,and Giannis Tzimas. An integrated e-recruitment sys-tem for cv ranking based on ahp. InWEBIST, pages147–150, 2011.

9. V Senthil Kumaran and A Sankar. Towards an auto-mated system for intelligent screening of candidatesfor recruitment using ontology mapping (expert).In-ternational Journal of Metadata, Semantics and On-tologies, 8(1):56–64, 2013.

10. Martín, Nacho & Mariello, Andrea & Battiti, Roberto & Hernández, José. (2018). Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study. International Journal of Computational Intelligence Systems. 11. 1192. 10.2991/ijcis.11.1.90.