EasyChair Preprint

№ 5573

Salary Forecasting Using Neural Network Methods

Natalya Shklyaeva

May 20, 2021

ПРОГНОЗИРОВАНИЕ ЗАРАБОТНОЙ ПЛАТЫ С ПОМОЩЬЮ НЕЙРОСЕТЕВЫХ МЕТОДОВ

Н.В.Шкляева
НГТУ
natalyashkl@yandex.ru

Аннотация

В статье описано применение нейронной сети для прогнозирования диапазона заработной платы специалистов в сфере программирования и разработки для Новосибирской области. В качестве инструмент анализа данных предложено использовать аналитическую платформу Deductor.

Представлена основная информация об архитектуре нейронной сети, а также определены наиболее значимые параметры, оказывающие наибольшее влияние на ожидаемую зарплату.

Ключевые слова: нейронная сеть, прогнозирование, заработная плата, интернет.

Введение

В современных реалиях рынок труда является одним из показателей, по состоянию которого можно судить о благополучии, стабильности, эффективности политики и развития регионов и страны в целом. А быстрый рост использования интернета для поиска и размещения предложений по трудоустройству предоставляет прекрасную возможность для мониторинга рынка труда в режиме реального времени. Все это способствовало введению термина «анализ рынка труда» (LMI), который относится к использованию и разработке алгоритмов и структур ИИ (искусственного интеллекта) для анализа данных о рынке труда для поддержки принятия решений [1]. Анализ вакансий в Интернете, действительно, представляет конкурентное преимущество для участников рынка труда в отношении классического анализа на основе опросов. Поскольку он является богатым источником информации для понимания динамики и тенденций рынка труда.

В связи с этим в последнее время появилось множество различных исследований в данной области, в которых авторы изучали различные подходы к анализу рынка труда и выявления ключевых факторов, оказывающих наибольшее влияние. Чаще всего среди таких факторов выделяют следующие [2-3]:

- Уровень заработной платы;
- Политика государства;
- Демографическая структура населения;
- Число квалифицированных работников;
- Социальное законодательство.

Но если исследование всех этих факторов имеют практическую ценность в основном для государства, то изучение заработной платы будет также интересно и для большей части населения.

Знания человека о собственной ценности, не только с точки зрения компании, но и с точки зрения рынка труда, будет полезна при поиске работы, при продвижении по карьерной лестнице внутри компании и при переходе в другую компанию.

Помимо прочего на рынке труда существует проблема непонимания реальной ценности конкретных компетенций. Для соискателя это может привести к неправильно выстроенным зарплатным и должностным ожиданиям. А обладая информацией о ценности компетенций, можно не только правильно написать резюме, но также понять какие навыки стоит развить в себе для повышения собственной ценности.

Компаниям также необходимо знать и учитывать тенденции рынка, так как отсутствие такого анализа может повысить вероятность оттока ценных кадров, снижения мотивации сотрудников и т.д.

Исходя из выше сказанного, наша задача в данном исследовании состоит в том, чтобы определить, как компетенции влияют на уровень заработной платы. Для решения этой задачи воспользуемся самым популярным методом для прогнозирования на основе онлайн-данных - нейронными сетями.

Обзор методов прогнозирования

В общем смысле прогнозирование является процессом предопределения будущего на основании различных исходных параметров (например, опыта, выявленных закономерностей, тенденций, связей, возможных перспектив и т. п.). Сейчас прогнозирование используется в самых различных областях жизнедеятельности человека: экономике, социологии, демографии, политологии, метеорологии, генетике и многих других. В свою очередь, эффективное использование прогнозов на научной основе требует применения определенных методик, включающих в себя целый ряд методов прогнозирования.

Учитывая сложную структуру рынка труда, для описания его функционирования используют методы машинного обучения, такие как метод опорных векторов (SVM), алгоритм случайного леса (RFs) и нейронные сети [4-5]. Рассмотрим их более подробнее.

Метод опорных векторов (SVM – Support Vector Machine) относятся к категории универсальных сетей прямого распространения, как многослойный персептрон и сети на основе радиальных базисных функций. Основная идея метода опорных векторов заключается в построении гиперплоскости, разделяющей объекты выборки оптимальным способом. Алгоритм работает в предположении, что чем больше расстояние (зазор) между разделяющей гиперплоскостью и объектами разделяемых классов, тем меньше будет средняя ошибка распознавания. Основным преимуществом методов SVM является то, что они особенно хорошо подходят для классификации сложных, но небольших или средних наборов данных. Среди минусов данного подхода следует отметить, что для классификации используется не все множество образцов, а лишь их небольшая часть, которая находится на границах.

Следующий алгоритм - алгоритм случайного леса (RFs). Идея данного алгоритма состоит в том, чтобы объединить несколько решающих деревьев, взять выходные значение каждого и получить некоторое среднее значение, тем самым снизив уровень ошибки алгоритма.

Решающие деревья обучаются на разных подвыборках и используют для разбиения разные признаки. За счёт этого получаются разные решающие деревья, а соответственно и

разные результаты. В результате случайные леса имеют тенденцию к большей устойчивости к изменениям в данных, они очень устойчивы к шуму (то есть, переменным, которые имеют небольшое влияние на целевую переменную). Одним из недостатков является то, что при реализации Random Forest в различных пакетах алгоритмов машинного обучения практически не поддерживается работа с категориальными признаками (как представленными в виде строк, так и в виде целых чисел).

Самым популярным методом являются искусственные нейронные сети, который мы и будем использовать в рамках данного исследования.

Нейронные сети - это очень мощный и гибкий механизм прогнозирования. В связи с этим нейросетевое прогнозирование является одной из наиболее интересных и имеющих практическое применение областей использования нейросетевых подходов.

Интерес к использованию нейронных сетей в задачах прогнозирования связан с их способностью решать неподдающиеся строгой формализации задачи, выявлять внутренние и скрытые закономерности, а также проводить глубокий анализ данных.

Построение нейросетевой системы включает в себя обработку входных данных, разработку архитектуры и обучение сети. На данный момент не существует единого алгоритма реализации каждого из перечисленных этапов, так как конфигурация системы зависит от множества факторов связанных с конкретной задачей. Поэтому при разработке нейронной сети важно учитывать природу исходных данных, период прогнозирования, объем входных данных и т.д. Важнейшим преимуществом нейронных сетей с точки зрения прогнозирования заключается в гибкости и отсутствии строгой формализации. Все это дает широкий спектр возможностей исследования, усовершенствования и адаптации существующих моделей нейронных сетей и повышения точности прогноза.

Искусственная нейронная сеть представляет собой совокупность нейронов, взаимодействующих друг с другом, в результате чего происходит обработка информации и обмен данными между собой. Обмен данными происходит через синаптические связи, соединяющие выход одного нейрона с входом другого. Каждая связь характеризуется весовым коэффициентом, благодаря которому, входная информация изменяется, когда передается от одного нейрона к другому. По мере того как происходит обработка информации, нейрон суммирует входные сигналы, вычисляет от полученной суммы функцию и после чего передает результирующее значение на выход [6].

Существуют различные виды искусственных нейронных сетей, которые отличающихся друг от друга строением и процессом обучения. Однако наиболее распространенной и успешно применяемой в прогнозировании является многослойный персептрон. Эта сеть, в которой помимо входного и выходного слоя присутствуют и скрытые слои нейронов, число которых зависит от решаемой задачи.

Многослойный персептрон относится к классу сетей, обучение которых осуществляется методом обратного распространения ошибки [7]. Для обучения системы используется обучающая выборка, состоящая из обучающих примеров, формирование которых основывается на исходных данных и специфике решаемой задачи.

При решении задачи прогнозирования нейросетевая система строится следующим образом: входной слой содержит несколько нейронов, на которые подаются значения исследуемого временного ряда, а последний слой состоит из единственного нейрона, на выходе которого получается прогноз.

Прогнозирование заработной платы с помощью нейросетевых методов

Для построения нейронной сети будем использовать программу аналитической платформы Deductor, которая является основой для принятия конечных прикладных решений в области анализа данных. Нейронная сеть представляет собой многослойный персептрон с одним скрытым слоем из 10 нейронов, так как такое количество обладает высокой скоростью обучения и при этом имеет низкий уровень ошибки.

Количество входных нейронов в сети – 9, что соответствует количеству входных параметров - компетенций претендентов на должность в сфере программирования и разработки (наиболее часто встречаемые в объявлениях в интернет-бирже труда – HeadHunter требования к кандидатам):

- Опыт использования реляционных СУБД или NoSQL-хранилищ;
- Опыт работы в команде с использованием гибких методологий (Scrum, Kanban);
- Уверенные знания, понимание принципов ООП и паттернов;
- Знание английского языка;
- Высшее образование;
- Знание нескольких языков программирования;
- Опыт в коммерческих и промышленных разработках;
- Опыт работы в Linux из командной строки;
- Опыт работы.

Входной параметр такой как «Опыт работы» принимает оценку данного навыка от 0 до 4, где:

- 1 – без опыта работы,
- 2 – требуемый опыт от 1 года до 3 лет,
- 3 - требуемый опыт от 3 до 6 лет,
- 4 – требуемый опыт более 6 лет.

Все остальные входные параметры принимают значение 1, если такой навык требуется от кандидата, в противном случае - 0.

Поскольку в регионах наблюдаться значительные различия в уровне платы труда то, было принято решение анализировать только одни регион, в связи, с чем исследование проводилось только по вакансиям Новосибирской области.

Описанные выше данные были получены путем обращения к публичному API сайта HeadHunter, в результате чего для исследования было собрано 1334 вакансий. Обязательным условием при сборе данных было наличие заработной платы в вакансии. Для выявления ключевых навыков из объявления был использован текстовый анализ. Полученные данные были разбиты на обучающее и тестовое множество в отношении 70/30, что в переводе на числа означает 934 примеров в обещающем и 400 примеров в тестовом множестве.

Так как в объявлениях чаще всего указывают диапазон заработной платы, то на выходе у сети будем использовать 2 нейрона, которые соответствуют началу диапазона прогнозируемых зарплат и концу данного диапазона. В ходе многочисленных испытаний, было выявлено, что наиболее удачной функцией активации (нелинейная функция, которая вычисляет выходной сигнал формального нейрона) для исследуемых данных является сигмоида. А для бучения сети лучше всего использовать метод Resilient Propagation (Rprop) [8]. Данный алгоритм использует так называемое «обучение по эпохам», когда

коррекция весов происходит после предъявления сети всех примеров из обучающей выборки. Преимущество этого метода заключается в том, что обучение сети происходит в 4-5 раз быстрее, чем стандартный алгоритм Backprop.

В результате мы получили многослойную нейронную сеть, которая способна прогнозировать диапазон заработной платы в зависимости от требуемых навыков кандидата на должность с уровнем ошибки в 0,0049. На тестовой выборке средняя ошибка составляет 0.007, что также довольно неплохой результат.

Оценить качество построенной модели можно также и с помощью диаграммы рассеивания (рис. 1-2), на которой видно, что прогнозируемые значения имеют небольшое от реальных. Большая часть точек сосредоточена на небольшом расстоянии от линии идеальных оценок и находится в пределах заданного «коридора» погрешности.

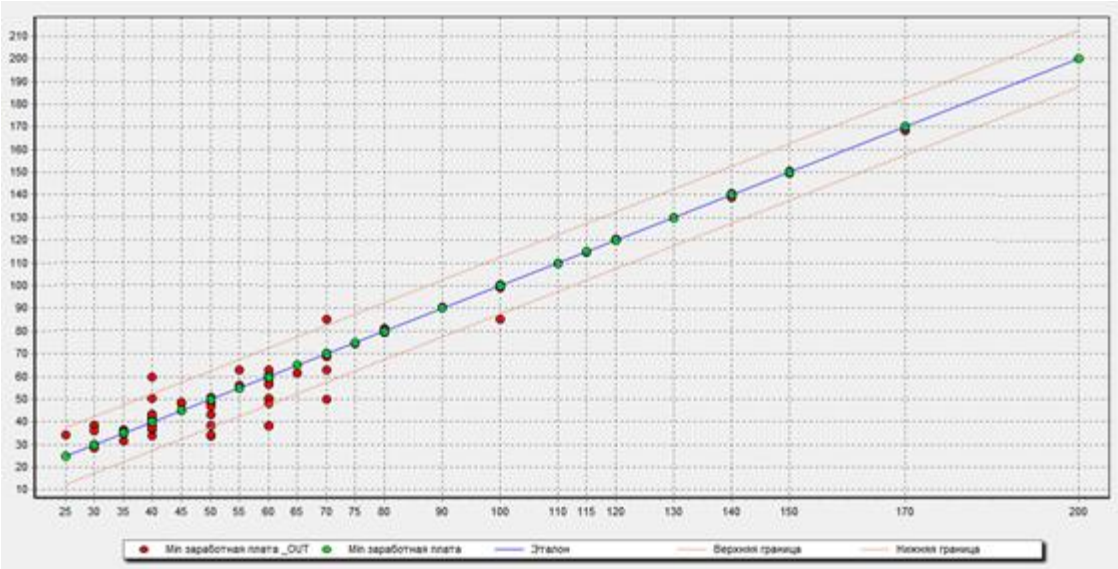

Рис. 1. Диаграмма рассеяния (отклонения) прогнозируемых значений от реальных для минимальной з/п (тыс. руб)

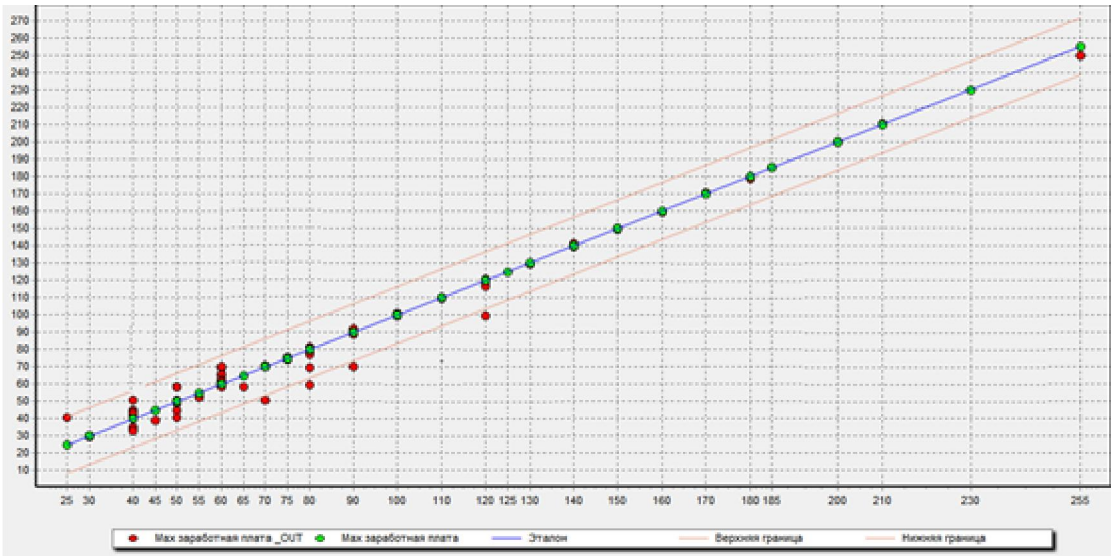

Рис. 2. Диаграмма рассеяния (отклонения) прогнозируемых значений от реальных для максимальной з/п (тыс. руб)

С помощью поочередной активации входных нейронов и наблюдением за результатом сети можно определить степень влияния входных параметров модели на минимум и максимум диапазона прогнозируемой заработной платы. Наиболее значимыми компетенциями претендентов на должность в сфере программирования и разработки оказались «Знание нескольких языков программирования», «Опыт работы» и «Опыт в коммерческих и промышленных разработках». Менее значимыми навыками являются «Опыт работы в команде с использованием гибких методологий», «Опыт использования реляционных СУБД или NoSQL-хранилищ» и «Высшее образование». На рисунке 3 представлена гистограмма значимости входных параметров на выходные значения.

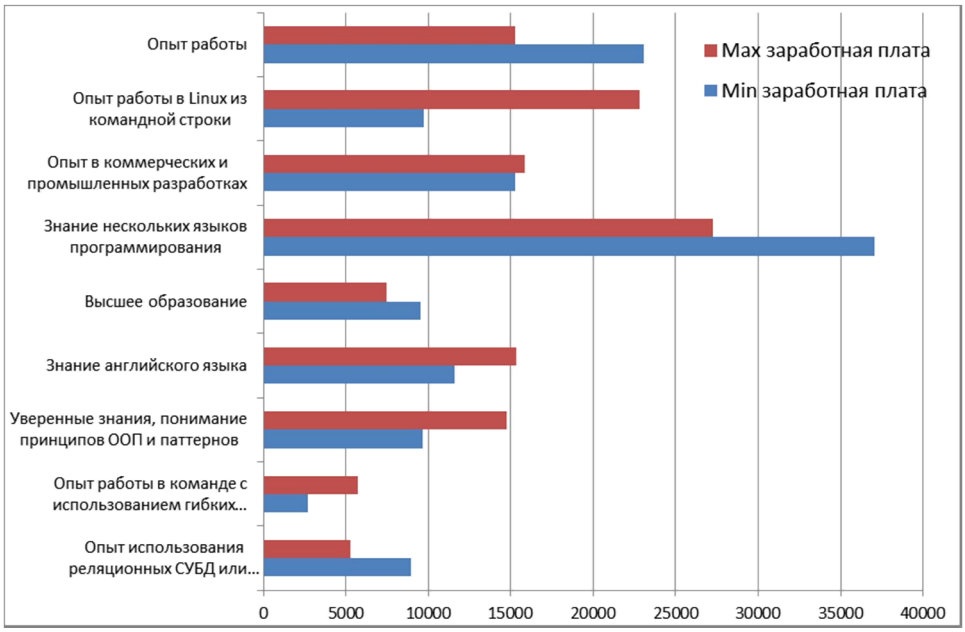

Рис.3. Значимость входных параметров

Высокая значимость такого навыка, как «Знание нескольких языков программирования» объясняется тем, что все чаще работодатели хотят видеть «универсальных» сотрудников, способных решать различные задачи. «Опыт работы» и «Опыт в коммерческих и промышленных разработках» также обладают высокой ценностью, так как являются доказательством применения своих умений и возможностей на практике соискателем на должность.

Заключение

В рамках исследования была построена нейросетевая модель, позволяющая выполнять прогнозирование заработной платы претендентов на должность в сфере программирования и разработки для Новосибирской области. А также определить наиболее значимые параметры, оказывающие наибольшее влияние на ожидаемую зарплату.

Несмотря на то, что полученная модель обладает достаточно высокой точностью, у нее имеется недостаток. Так как данные были взяты из вакансий, в которых обычно требуется определенный специалист с определенным набором навыков в рамках какой-нибудь конкретной экосистемы для решения какой-либо конкретной задачи. И в этих случаях заработная плата будет формировать не только исходя из навыков, но и исходя из тех задач и типа проекта, на которые идет специалист, что невозможно учесть при прогнозировании.

Литература

[1] Boselli R. et al. Using machine learning for labour market intelligence //Joint European Conference on Machine Learning and Knowledge Discovery in Databases. – Springer, Cham, 2017. – С. 330-342.
[2] Кобец Е. А. Факторы, влияющие на рынок труда //Инновационная наука. – 2016. – №. 8-1.
[3] Лехтянская Л. В., Римская Т. Г. Факторы, влияющие на формирование и развитие рынка труда //Российское предпринимательство. – 2016. – Т. 17. – №. 5.
[4] Boselli R. et al. Labour Market Intelligence for Supporting Decision Making //SEBD. – 2017. – С. 74.
[5] Dawson N. et al. Adaptively selecting occupations to detect skill shortages from online job ads //2019 IEEE International Conference on Big Data (Big Data). – IEEE, 2019. – С. 1637-1643.
[6] Шагалова П.А., Ляхманов Д.А. НЕЙРОСЕТЕВЫЕ ТЕХНОЛОГИИ В РЕШЕНИИ ЗАДАЧ ПРОГНОЗИРОВАНИЯ // Современные проблемы науки и образования. – 2014. – № 6.;URL: http://science-education.ru/ru/article/view?id=16494
[7] Хайкин С. Нейронные сети: полный курс / Саймон Хайкин. – М.: Издательский дом «Вильямс», 2006. – 104с.
[8] Saputra W. et al. Analysis resilient algorithm on artificial neural network backpropagation //Journal of Physics: Conference Series. – IOP Publishing, 2017. – Т. 930. – №. 1. – С. 012035.

Авторы

Шкляева Наталья Васильевна, НГТУ

SALARY FORECASTING USING NEURAL NETWORK METHODS

N.V.Shklyaeva
NTSU
natalyashkl@yandex.ru

The article describes the use of a neural network to predict the salary range of specialists in the field of programming and development for the Novosibirsk region. It was proposed to use the Deductor analytical platform as a data analysis tool. Basic information about the architecture of the neural network is presented, and the most significant parameters that have the greatest impact on the expected salary are identified.

Keywords: neural network, forecasting, salary, Internet.

Reference for citation:

Reference

[1] Boselli R. et al. Using machine learning for labour market intelligence //Joint European Conference on Machine Learning and Knowledge Discovery in Databases. – Springer, Cham, 2017. – C. 330-342.
[2] Kobec E. A. Faktory, vliyayushchie na rynok truda //Innovacionnaya nauka. – 2016. – №. 8-1.
[3] Lekhtyanskaya L. V., Rimskaya T. G. Faktory, vliyayushchie na formirovanie i razvitie rynka truda //Rossijskoe predprinimatel'stvo. – 2016. – T. 17. – №. 5.
[4] Boselli R. et al. Labour Market Intelligence for Supporting Decision Making //SEBD. – 2017. – C. 74.
[5] Dawson N. et al. Adaptively selecting occupations to detect skill shortages from online job ads //2019 IEEE International Conference on Big Data (Big Data). – IEEE, 2019. – C. 1637-1643.
[6] Shagalova P.A., Lyahmanov D.A. NEJROSETEVYE TEKHNOLOGII V RESHENII ZADACH PROGNOZIROVANIYA // Sovremennye problemy nauki i obrazovaniya. – 2014. – № 6.;URL: http://science-education.ru/ru/article/view?id=16494
[7] Hajkin S. Nejronnye seti: polnyj kurs / Sajmon Hajkin. – M.: Izdatel'skij dom «Vil'yams», 2006. – 104s.
[8] Saputra W. et al. Analysis resilient algorithm on artificial neural network backpropagation //Journal of Physics: Conference Series. – IOP Publishing, 2017. – T. 930. – №. 1. – C. 012035.

Прогнозирование заработной платы с помощью нейросетевых методов