https://doi.org/10.31891/2307-5732-2023-317-1-276-279
УДК 004.93, 004.8

КРИВЕНЧУК Юрій
Національний університет "Львівська політехніка"
https://orcid.org/0000-0002-2504-5833
e-mail: Yurii.P.Kryvenchuk@lpnu.ua
ГОРІШНА Надія
Національний університет "Львівська політехніка"
e-mail: nadiia.horishna.knm.2019@lpnu.ua

АНАЛІЗ ТА ПРОГНОЗУВАННЯ ЗАРПЛАТ ЗА ДОПОМОГОЮ МАШИННОГО НАВЧАННЯ

В роботі наведено результати дослідження теми прогнозування заробітної плати людини за параметрами посади, досвіду й рівня освіти та розроблення системи для вирішення такої задачі, з метою покращення якості ринку заробітних плат та впровадження таких систем у відповідні сфери в Україні. Виділено та описано такі основні етапи: визначення тестових наборів даних,навчання нейронної мережі, передбачення зарплати. Після проведення аналізу результатів було виявлено, що створення системи передбачення заробітної плати за посадою, досвідом та рівнем освіти є актуальним та доцільним завданням на сьогодні, а найбільш ефективним інструментом для цього є використання нейронних мереж.

Ключові слова: передбачення зарплати, нейронні мережі.

KRYVENCHUK Yurii, HORISHNA Nadiia
Lviv Polytechnic National University

CREATION OF SALARY PREDICTION SYSTEM

Salary prediction is an important problem for many organizations as it directly affects the financial wellbeing of employees and the competitiveness of the organization. Accurately predicting employee salaries enables organizations to make informed decisions about compensation and benefits packages, leading to a more equitable distribution of compensation, improved organizational performance and increased employee satisfaction. In today's highly competitive job market, accurate salary prediction is increasingly important as organizations seek to attract and retain the best talent, maintain a motivated and productive workforce, and stay ahead of the competition. The prediction of salaries for employees can be challenging as it requires taking into account a wide range of factors including years of experience, education, skills, job responsibilities, industry trends and more.

Salary prediction systems can also help to identify and eliminate any potential salary disparities between different groups of employees. By providing a more objective and data-driven approach to determining salaries, these systems can help to promote fairness and equity in the workplace. Additionally, salary prediction systems can be useful for companies to budget and plan their finances, by having a better understanding of the potential salary range for a given job, the company can budget accordingly. The practical value of this work is the salary prediction system which can provide valuable insights and help to make more informed decisions about compensation, which can benefit both employers and employees

Moreover, with the advent of machine learning and artificial intelligence, it has become possible to develop sophisticated algorithms and models to analyze vast amounts of data and make accurate salary predictions. These systems use deep learning techniques, such as neural networks, to analyze large amounts of data and predict salaries based on a wide range of factors. This technology is becoming increasingly widespread and is poised to play a major role in the future of human resources and compensation decision-making.

Keywords: salary prediction, neural networks.

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Системи прогнозування заробітної плати можуть надати цінну інформацію, яка допоможе як роботодавцям, так і працівникам приймати більш обґрунтовані рішення щодо винагороди, сприяти чесності та справедливості на робочому місці. Люди, що у пошуках роботи, зможуть використовувати систему прогнозування зарплати, щоб знайти роботу, яка відповідає їхнім кваліфікаціям і очікуваній зарплаті. Це може допомогти їм у вирішенні питання про те, на яку посаду претендувати, а також скоротить час і зусилля, витрачені на роботу, яка не підходить. Крім того, дана система може допомогти визначити найважливіші характеристики, які впливають на заробітну плату, що можна використовувати для покращення процесу роботи з персоналом. Забезпечуючи більш об'єктивний підхід до визначення заробітної плати, який базується на даних, компанії можуть використовувати систему прогнозування, щоб краще складати бюджет і планувати свої фінанси. Маючи краще розуміння потенційного діапазону зарплати для даної роботи, компанія може відповідним чином скласти бюджет.

Аналіз досліджень та публікацій

У статті [1] обговорюється передбачення зарплати та візуалізація вакансії, пов'язані з їх майбутньою кар'єрою. Для вирішення проблеми було застосовано підхід лінійної регресії та методи візуалізації. Дані, які використовувалися для дослідження, були попередньо оброблені, після цього автори

розробили модель і працювали з реальними даними. Для визначення якості моделі використовувалася середня абсолютна похибка (MAE). У даному випадку лінійна регресія була непридатна для перевірки точності, оскільки даний проект передбачає зарплату на основі необхідного багаторічного досвіду. Отже, у цій роботі сказано, що було доцільніше з'ясувати якість даних і різницю між фактичними та прогнозованими даними, використовуючи MAE, як найпростіший показник для розуміння. Результати надійності вказують на позитивну кореляцію з фактичними значеннями. Недоліком я би відокремила використання обмеженої інформації на веб-сайті Jobstreet як єдиного джерела даних. У статті [2] розв'язувалась задача порівняння ефективності двох методів регресії, а саме алгоритмів простої лінійної регресії (SLR) і множинної лінійної регресії (MLR) у двох випадках: прогнозування зарплати працівників через певні роки та прогнозування цін на нерухомість. Набір даних, використаний у цьому експерименті, є набором даних із відкритим кодом від KaggleInc. Алгоритми порівнювалися з використанням таких параметрів, як значення R-квадрат, середня абсолютна похибка (MAE), середня квадратична похибка (MSE), показник дисперсії та середньоквадратична похибка (RMSE). Результати показали, що MLR забезпечує кращу ефективність порівняно з SLR. У даній науковій публікації самі автори стверджують, що покращити експеримент рекомендується таким чином, що брати на опрацювання великий набір даних, щоб побудувати кращу модель передбачення. У статті [3] прогнозування зарплати розглядається як проблема порядкової регресії та використовуються методи глибокого навчання для побудови моделі прогнозування зарплати для визначення відносного порядку між різними рівнями зарплати. Зокрема, щоб вирішити цю проблему, автори використовують особисту інформацію студентів, оцінки та сімейні дані як вхідні дані функції та багатовихідну глибоку нейронну мережу для фіксації кореляції між рівнями заробітної плати під час навчання. Щоб покращити продуктивність моделі, використовується агрегація відсіву та початкового завантаження. Проте дане дослідження не може бути застосовано більш широко, адже тут для аналізу беруться такі вхідні значення, як оцінки студентів, тому запропоноване вирішення підійде тільки для вузького кола людей, а саме студентів. У статті [4] дослідження проводяться для прогнозування вмісту веб-сайту на основі даних про відвідувачів із підходом інтелектуального аналізу даних. У цій роботі для аналізу та передбачень використовують два алгорими для порівняння: Random forest та k-NN. Оцінка алгоритму Random forest має значення точності 71 відсоток, тоді як алгоритм k-NN має вищі значення точності, а саме 84,88 відсотка. Підсумовуючи цей експеримент, можна зробити висновок, що алгоритм k-NN виконує прогнозування процесу обробки даних щодо ефективніше, ніж Random forest у даному випадку. З недоліків варто виокремити складність обчислень та відносно низьку точність. У статті [5] автор зосередився на вивченні прогнозування зарплати за допомогою різних моделей глибокої нейронної мережі, включаючи TextCNN, Bi-GRU-LSTM-CNN і Bi-GRU-CNN з різними попередньо підготовленими вставками слів у набір даних про роботу в IT. Крім того, він запропонував просту та ефективну модель ensemble, що поєднує різні моделі глибоких нейронних мереж. Результати його експериментів показали, що запропонована модель досягла результату з показником точності F 72,71%. Для покращення даного досдіження, можна працювати над отриманням кращих результатів з вищим показником точності експерименту. У статті [6] описано, як проводити власні дослідження зарплати, як інтерпретувати результати та як організації можуть застосовувати результати. Автор стверджує, що дане дослідження було проведено, щоб використовувати результати аналізу для виявлення окремих випадків, коли зарплата може не відповідати прогнозам. Для цього було використано модель заробітної плати, щоб передбачити заробіток кожної особи, а потім порівнювання цих прогнозованих даних з фактичними заробітками, які отримують працівники. Першим кроком у цьому процесі було виконання регресійного аналізу, а після цього використання результатів цієї моделі, щоб передбачити, якою була б зарплата кожної особи, в залежності від від статі чи кольору шкіри. Недоліком даної роботи є те, що дослідження більше фокусується на різних вхідних варіаціях даних, ніж на якості регресійного аналізу. У статті [7] виконано дослідження того, які сфери роботи мають більший вплив на зарплату, як вони взаємопов'язані та як це можна передбачити. Формулювання проблеми було поставлено, як прогнозування заробітної плати як задачі класифікації для кращої точності шляхом зосередження на дискретних діапазонах замість безперервних значень заробітної плати. У даній роботі було проведено порівняння кількох класифікаторів, включаючи SVM, MLP, random forest, AdaBoost і їх ансамблі, щоб знайти модель з найкращою точністю прогнозування діапазону зарплати. Цю модель автор запропонував використовувати на веб-сайтах з пошуку роботи, щоб забезпечити автоматичну класифікацію вакансій за діапазоном зарплати, навіть якщо реальна пропонована зарплата відсутня, або використовувати як алгоритм у системі рекомендацій щодо роботи. Дане дослідження проведено масштабно, з використаннях різних методів та також було зроблено аналіз використаних методів з точки зору точності та оцінки F1. Класифікатори на основі ансамблів дерев рішень, а також на основі ансамблів голосування досягають найкращої точності. Їхня середня точність становить ≈ 0,84. KNN досягає середньої точності ≈ 0,79, а всі решта моделей (LR, MLP і SVM) поводяться істотно гірше. Для LR автор це пояснює нелінійністю проблеми, тоді як для MLP і SVM дефіцит даних, ймовірно, є найбільшою перешкодою. У статті [8] досліджуються алгоритми машинного навчання та підходи до прогнозування доходу випускників. У цій роботі проведено поглиблений аналіз, щоб визначити, чи можна підвищити точність традиційних алгоритмів за допомогою наукового підходу. Автор для даного дослідження використав набір даних, який містить значення доходу та великий набір незалежних демографічних ознак студентів. Результати показують, що моделі машинного навчання перевершили параметричні моделі лінійної та логістичної

регресії у прогнозуванні поточного доходу зі статистично значущими результатами у трьох різних завданнях. Крім того, пізніші методи виявилися найточнішими для прогнозування першого доходу випускника після закінчення навчання. Проте тут така ж проблема як у статті [5], автори використовують для вхідних даним саме інформацію про студентів. У статті [9] за допомогою статистичного машинного навчання (ML) розроблено та перевірено цілісну професійну та економічну структуру для прогнозування зарплати. Використані моделі розроблялися на основі професійних особливостей та організаційних характеристик. П'ять різних керованих алгоритмів ML навчали на основі даних опитування ринку праці Саудівської Аравії для оцінки середньої річної зарплати в різних видах економічної діяльності та основних професійних групах. У прогнозуванні середньої заробітної плати за видами економічної діяльності регресія байєсівського процесу Гаусса ML показала помітне покращення порівняно з множинною лінійною регресією. Крім того, були отримані нижчі похибки результатів. Проте прогноз зарплати для основних професійних груп показав, що штучні нейронні мережі показали найкращі результати. У статті [10] автор пропонує розширену нейронну мережу з кооперативною структурою, а саме мережу композиції заробітної плати та навичок (SSCN), щоб відокремлювати професійні навички та вимірювати їх цінність на основі масових оголошень про роботу. Експерименти показують, що SSCN може не тільки призначити значущу цінність професійним навичкам, але й перевершує порівняльні моделі для прогнозування зарплати. У даній роботі робиться наголос саме на навички людей, в залежності від яких буде проводитись оцінка заробітної плати . Тут автор стверджує, що використовував класичні моделі регресії, включаючи лінійну регресію (LR), опорну векторну машину (SVM) і дерево рішень із посиленням градієнта (GBDT). Оскільки ці методи обробляють структуровані вектори ознак фіксованого розміру, було об'єднано характеристики навичок і контекст роботи як їхні вхідні дані. Проте у цій роботі не було описано прогнозування й аналіз заробітної плати використовуючи вказані методи.

Формулювання цілей статті

Метою роботи є: створення сучасної та якісної системи передбачення заробітної плати людей за параметрами посади, досвіду й рівня освіти для застосуванні на ринку заробітних плат.

Виклад основного матеріалу

Огляд наборів даних

Проаналізувавши роботи у численних джерелах, було прийнято рішення для навчання моделей нейронної мережі обрати кілька наборів даних та об'єднати їх частини. У таблиці 1 наведено основні характеристики проаналізованих датасетів.

Таблиця 1

Огляд наборів даних, що містять інформацію про заробітну плату людей з вказаними параметрами

База даних	*Кількість прикладів*	*Локація*	*Параметри*
Jobs Dataset from Glassdoor	741	USA	Job title, salary estimate, job description. Rating, company name, location, headquarters, size, founded
indeed-job-site-software-jobs-dataset	7289	worldwide	Job position, company, requirements, rating, experience, average yearly salary, work category, education level, job title, state
Glassdoor Job Postings : Data Science	3324	USA	Job title, company, state, city, min salary, max salary, job description, industry, rating, date posted, job type
Salary prediction dataset	366918	USA	Title, full description, location, company, category, salary

Етапи процесу передбачення заробітної плати

Процес прогнозування заробітної плати складається з наступних етапів: збір даних, підготовка даних, виокремлення ознак, вибір моделі, навчання моделі та оцінка моделі. *Збір даних*. Перший етап складається з пошуку даних з різних джерел та вибір ключових характеристик, які впливатимуть на заробітну плату, як-от посада, багаторічний досвід, рівень освіти та географічне розташування. *Підготовка даних*. Даний етап включає очищення та попередню обробку даних, щоб забезпечити їх точність і якість, наприклад, видалення даних поза вибіркою та обробка відсутніх значень. *Розробка функцій*. На цьому кроці створюються нові функції з існуючих даних, наприклад таку функцію, яка поєднує освіту та багаторічний досвід. *Вибір моделі*. Тут потрібно вибрати найбільш відповідний алгоритм машинного навчання для завдання, а саме нейронну мережу. Загалом FNN, MLP і RNN є найбільш часто використовуваними нейронними мережами для прогнозування зарплати. Проаналізувавши численну кількість робіт, було прийнято рішення обрати кілька архітектур рекурентних та нейронних мереж прямого зв'язку та натренувати їх, після чого на основі швидкодії та коректності вибрати найоптимальнішу. *Навчання моделі.* Наступним етапом буде навчання моделі, використовуючи підготовлені навчальні дані, налаштовуючи гіперпараметри та оцінюючи продуктивність моделі за допомогою таких показників, як середня квадратична помилка або середня абсолютна помилка. *Оцінка моделі*. Тут оцінюється продуктивність розгорнутої моделі на нових, невідомих даних, щоб визначити її точність і внести необхідні покращення.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

В результаті проведеної роботи було створено систему прогозування заробітної плати людини. Ця розробка може мати значний вплив на процес прийняття рішень у багатьох організаціях. Нейронні мережі, зокрема методи глибокого навчання, показують багатообіцяючі результати в цій галузі, пропонуючи високий рівень точності та здатність обробляти складні та нелінійні зв'язки в даних. Процес прогнозування заробітної плати включає кілька етапів, включаючи збір даних, підготовку даних, розробку функцій, вибір моделі, навчання моделі та оцінку моделі. Поєднуючи технічну експертизу зі знаннями предметної області, організації можуть розробити ефективні та результативні системи прогнозування заробітної плати, які допоможуть приймати обґрунтовані рішення щодо винагороди та пільг для працівників. Оскільки ця галузь недостатньо досліджена в Україні, то дана розробка є доцільною та актуальною.

Література

1. Abu Samah K. A. F. A linear regression approach to predicting salaries with visualizations of job vacancies: a case study of jobstreet malaysia / K. A. F. Abu Samah, N. S. Dinnie Wirakarnain, R. Hamzah, [et al.] // IAES International Journal of Artificial Intelligence (IJ-AI). — 2022. — Vol. 11, No. 3. — P. 1130.
2. Bansal U. Empirical analysis of regression techniques by house price and salary prediction / U. Bansal, A. Narang, A. Sachdeva, [et al.] // IOP Conference Series: Materials Science and Engineering. — 2021. — Vol. 1022, No. 1. — P. 012110.
3. Kuo J.-Y. Building graduate salary grading prediction model based on deep learning / J.-Y. Kuo, C.-H. Liu, H.-C. Lin // Intelligent Automation & Soft Computing. — 2021. — Vol. 27, No. 1. — P. 53–68.
4. Iskandar I. D. Popular content prediction based on web visitor data with data mining approach / I. D. Iskandar, N. C. Basjaruddin, D. Supriadi, [et al.] // Journal of Physics: Conference Series. — 2020. — Vol. 1641, No. 1. — P. 012105.
5. Van Huynh T. Job prediction: from deep neural network models to applications / T. Van Huynh, K. Van Nguyen, N. L.-T. Nguyen, A. G.-T. Nguyen. — Ho Chi Minh, Vietnam : IEEE, 2020.
6. Taylor L. L. How to do a salary equity study: with an illustrative example from higher education / L. L. Taylor, J. N. Lahey, M. I. Beck, J. E. Froyd // Public Personnel Management. — 2020. — Vol. 49, No. 1. — P. 57–82.
7. Martín I. Salary prediction in the it job market with few high-dimensional samples: a spanish case study: / I. Martín, A. Mariello, R. Battiti, J. A. Hernández // International Journal of Computational Intelligence Systems. — 2018. — Vol. 11, No. 1. — P. 1192.
8. Gomez-Cravioto D. A. Supervised machine learning predictive analytics for alumni income / D. A. Gomez-Cravioto, R. E. Diaz-Ramos, N. Hernandez-Gress, [et al.] // Journal of Big Data. — 2022. — Vol. 9, No. 1. — P. 11.
9. Matbouli Y. T. Statistical machine learning regression models for salary prediction featuring economy wide activities and occupations / Y. T. Matbouli, S. M. Alghamdi // Information. — 2022. — Vol. 13, No. 10. — P. 495.
10. Sun Y. Market-oriented job skill valuation with cooperative composition neural network / Y. Sun, F. Zhuang, H. Zhu, [et al.] // Nature Communications. — 2021. — Vol. 12, No. 1. — P. 1992.

References

1. Abu Samah K. A. F. A linear regression approach to predicting salaries with visualizations of job vacancies: a case study of jobstreet malaysia / K. A. F. Abu Samah, N. S. Dinnie Wirakarnain, R. Hamzah, [et al.] // IAES International Journal of Artificial Intelligence (IJ-AI). — 2022. — Vol. 11, No. 3. — P. 1130.
2. Bansal U. Empirical analysis of regression techniques by house price and salary prediction / U. Bansal, A. Narang, A. Sachdeva, [et al.] // IOP Conference Series: Materials Science and Engineering. — 2021. — Vol. 1022, No. 1. — P. 012110.
3. Kuo J.-Y. Building graduate salary grading prediction model based on deep learning / J.-Y. Kuo, C.-H. Liu, H.-C. Lin // Intelligent Automation & Soft Computing. — 2021. — Vol. 27, No. 1. — P. 53–68.
4. Iskandar I. D. Popular content prediction based on web visitor data with data mining approach / I. D. Iskandar, N. C. Basjaruddin, D. Supriadi, [et al.] // Journal of Physics: Conference Series. — 2020. — Vol. 1641, No. 1. — P. 012105.
5. Van Huynh T. Job prediction: from deep neural network models to applications / T. Van Huynh, K. Van Nguyen, N. L.-T. Nguyen, A. G.-T. Nguyen. — Ho Chi Minh, Vietnam : IEEE, 2020.
6. Taylor L. L. How to do a salary equity study: with an illustrative example from higher education / L. L. Taylor, J. N. Lahey, M. I. Beck, J. E. Froyd // Public Personnel Management. — 2020. — Vol. 49, No. 1. — P. 57–82.
7. Martín I. Salary prediction in the it job market with few high-dimensional samples: a spanish case study: / I. Martín, A. Mariello, R. Battiti, J. A. Hernández // International Journal of Computational Intelligence Systems. — 2018. — Vol. 11, No. 1. — P. 1192.
8. Gomez-Cravioto D. A. Supervised machine learning predictive analytics for alumni income / D. A. Gomez-Cravioto, R. E. Diaz-Ramos, N. Hernandez-Gress, [et al.] // Journal of Big Data. — 2022. — Vol. 9, No. 1. — P. 11.
9. Matbouli Y. T. Statistical machine learning regression models for salary prediction featuring economy wide activities and occupations / Y. T. Matbouli, S. M. Alghamdi // Information. — 2022. — Vol. 13, No. 10. — P. 495.
10. Sun Y. Market-oriented job skill valuation with cooperative composition neural network / Y. Sun, F. Zhuang, H. Zhu, [et al.] // Nature Communications. — 2021. — Vol. 12, No. 1. — P. 1992.