Proceedings of the 2nd International Conference on Financial Technology and Business Analysis
DOI: 10.54254/2754-1169/52/20230690

Whether the Salary of Data Scientists Can Be Predicted: New Evidence

Zehao Wang[1,a,*]

[1]*Bourns College of Engineering, University of California, Riverside, USA*
a. Zwang516@ucr.edu
**corresponding author*

Abstract: This study uses exploratory data analysis and an exploratory prediction model to examine data scientist salaries. The study examines salary determinants, salary trends, and a forecast model for data scientist salaries. The collection includes wage estimates, job descriptions, company evaluations, and industry data. Descriptive statistics and visualizations reveal variable distributions and trends. Linear regression is used to estimate salaries using geography, industry, firm rating, and job description. However, the original model has a large prediction error, requiring refining. The findings present significant implications for job seekers, companies, and policymakers, necessitating a thorough understanding and response from all these stakeholders. Addressing issues related to data availability and biases becomes imperative, as these could potentially distort the insights and the resulting decisions. Therefore, it's crucial to emphasize and encourage future research in this area to ensure a more equitable and comprehensive approach to employment and policy development.

Keywords: data science, machine learning, salary

1. Introduction

Technology, data, and need for data-based decision making have fueled tremendous growth and revolution in data science. Data scientists are essential across industries as the world grows data centric. employment seekers and professionals wishing to develop in data science must understand the data science employment market, including salary criteria. Many people have lost their jobs due to COVID-19. Thus, it is crucial to refine the job search process and equip candidates with the data science job market capabilities they require. This study examines data science employment salary determinants, salary ranges for distinct roles, and negotiating methods. this paper wish to equip data scientists with the knowledge to earn more.

"Data Scientist Jobs," a Glassdoor dataset with over 3,900 job advertisements, will help us reach our goals. This pickles-eat dataset comprises pay estimates, job descriptions, employer ratings, and locations. This dataset can reveal patterns and trends in the data science job market and pay drivers. In order to fully leverage the potential of big data, managers must master the art of recognizing exceptional data science expertise, enticing skilled professionals to join their organization, and effectively harnessing their abilities to drive productivity and innovation within the company.[1]. This paper will analyze the dataset and review data science and associated literature to provide a full picture of data science job pay. This paper will cover studies on data scientist employability,

DOI: 10.54254/2754-1169/52/20230690

predictive analytics models for contract renewals, data science solutions for social and economic analyses, and data science employment qualifications.

2. Literature Review

Data science has grown and become vital to comprehending the current work environment. Data-driven decision-making has been emphasized by the COVID-19 epidemic, increasing demand for data scientists across sectors. As the job market evolves, understanding data science trends and criteria helps job seekers and companies hiring skilled data experts. Job seekers and companies are interested in data science salary analysis and its components. Education, experience, geography, industry, and skill sets affect data science salaries. Knowing these criteria can help job seekers negotiate better pay and help businesses set competitive salaries to attract and retain top personnel.

Data scientists need technical, analytical, business, and soft skill [2]. In the digital workforce, data scientists require these talents. These skill clusters fit employment needs, making them significant. Large datasets and analytical models demand technical skills like Python, R, and data processing. Analytical skills include data analysis utilizing statistical and machine learning algorithms. Data-driven strategies involve communication and topic understanding. Soft skills like problem-solving, teamwork, and adaptability are necessary for working in interdisciplinary teams and adapting to fast-changing settings. To acquire appropriate data science skills, job seekers and enterprises must understand these skill clusters.

A software contract renewal predictive analytics approach has been developed [3]. This model predicts contract renewals using various methods, i.e., logistic regression, decision trees, random forest, gradient boosting, etc. These models are evaluated through various statistical metrics such as accuracy, precision, and others. Beyond contract renewals, data science employs predictive analytics. It improves HR hiring and predicts employee turnover. Contract renewal model predictive analytics can help data scientists negotiate contracts and determine job security.

COVID-19's economic and social impact can be analyzed using data science [4]. This system examines social media, news, government reports, and surveys utilizing natural language processing, sentiment analysis, subject modeling, and visualization. Data science solutions that solve social and economic concerns using evidence-based insights are useful. Data scientists can use these solutions to create new ways for real-world problems. Data science skills help solve social issues and make decisions.

Healthcare data scientist job advertisements were evaluated to identify relevant skills and qualifications [5]. Programming, statistics, machine learning, domain expertise (biology, medicine, and health informatics), and soft skills (communication, teamwork, and problem-solving) are needed by healthcare data scientists [5]. Programming, statistics, machine learning, business intelligence, and communication are needed for data science [6].

Job trends in data analytics and knowledge management were explored [7]. They identified employment needs and understood changing professions using the KSA framework. Job-specific knowledge, skills, and abilities are categorized under KSA. Data analytics and knowledge management specialists need these skills. Data visualization, programming, and statistics are included. Soft talents include communication, teamwork, problem-solving, and critical thinking combine technical and interpersonal skills. Knowledge management and data analytics share technical and soft skills. Programming and data manipulation are required. Knowledge management professionals organize, retrieve, and distribute information, whereas data analytics professionals analyze, model, and learn. To work in diverse teams and share knowledge, these professionals need soft skills like communication and teamwork [7].

Data science and big data education were investigated [8]. Online job postings were qualified using text mining and NLP. Natural language processing and text mining automate text analysis. This

DOI: 10.54254/2754-1169/52/20230690

survey found data science and big data skills in job ads. The researchers extracted education, experience, technical skills, domain knowledge, soft skills, and other traits from the text. The survey revealed data science and big data job credentials. Math, statistics, computer science, and related degrees are needed. Experience with Python, R, SQL, machine learning, data visualization, and big data technologies was highlighted. Marketing, finance, and healthcare knowledge were appreciated. These occupations required soft skills including communication, problem-solving, and critical thinking [8].

AI (Artificial Intelligence) in Human Resource Management (HRM) Problems and potential unveils future workforce solutions [9]. HR may alter with AI. AI technologies and algorithms automate and streamline HR procedures, improving efficiency and decision-making. AI may analyze resumes, candidates, and employee performance. However, HRM AI adoption presents challenges. Ethical concerns about prejudice and fairness, legal worries about privacy and data protection, talent gaps in understanding and deploying AI technology, and organizational change management are important difficulties. AI can recruit and manage data science jobs. AI systems can assess, and shortlist candidates based on abilities and qualifications, improving the hiring process. AI can identify skill gaps, provide personalized training, and help with data-driven performance reviews [9].

An Information and Communication Technology and data science book was written [10]. Trends and applications are covered. Data science employment market changes and ICT are important. Data science has changed due to cloud computing, big data analytics, AI, and ML. Data collection, storage, analysis, and interpretation are now possible because to these technologies. Data science and ICT are interdisciplinary; therefore, the book covers cloud computing, big data analytics, internet of things, artificial intelligence, and cybersecurity. To keep current, data science job aspirants must understand these technologies and their applications [10].

Skills and benefits predicted data science professional salary [11]. They discovered talents connected to higher incomes and studied salary-benefit correlations. They predicted data science salaries based on abilities using machine learning and statistical analysis. Deep learning, NLP, and cloud computing anticipated higher compensation. They evaluated how health insurance, retirement programs, and paid time off affected talents and wages. Financial and ESG performance of public enterprises were analyzed [12]. They used machine learning and logistic regression models to compare high and poor ESG firms' profitability, liquidity, solvency, and efficiency [12].

Predicting Chinese P2P loan failures and wage inequality that the Machine learning was utilized [13]. Using borrower attributes, loan details, and platform reputation, loan defaults are predicted by techniques such as logistic regression, random forest, support vector machine, neural networks, etc. Lenders and investors benefit from this study. Engineering wage inequality is assessed via predictive analytics [14]. Predictive methods like linear regression and decision trees use gender, race, ethnicity, education, and experience to determine salary discrepancies. Salary disparities can be identified and addressed using such analysis [14].

Construction Forecast and Job Market Analysis that the Purdue Index for Construction Analytics (PICA) was used to predict and forecast the construction industry [15]. Based on employment, spending, permits, confidence, and prices, they use time series analysis and machine learning to estimate construction industry health and outlook. Quan and Raheem use human resource analytics to predict job and income based on specialized skill sets in data science. To analyze employment market dynamics, they use machine learning models to assess data science skills demand and supply across industries and locales. Based on skills, experience, education, geography, sector, and company size, these algorithms predict data science wages.

The studied literature illuminates the market characteristics of data science. It stresses data scientist skill requirements and employability skills clusters. Predictive analytics could be used in contract renewals, social and economic analysis, loan default prediction, and pay inequality study. It also

DOI: 10.54254/2754-1169/52/20230690

shows the importance of forecasting models in construction and human resource analytics in data science employment and wage prediction. However, gaps in the literature present study opportunities. For instance, the long-term effects of the epidemic on the data science job market and the skills and certifications that have become important deserve further study. Research on the ethical implications of employing predictive analytics in hiring decisions and solutions to reduce wage inequities in different industries would be useful. Future study could also leverage online platforms and professional networking sites to collect more job market and compensation data. Natural language processing and sentiment analysis could improve job-related data extraction and analysis from unstructured data sources.

3. Data Collection and Preprocessing

The research analyzed the "Data Scientist Jobs" dataset. The collection includes over 3900 data scientist job advertisements with income estimates, job descriptions, business reviews, locations, and more. The Glassdoor dataset was useful for data science job market analysis. Data collecting is essential to data analysis. Picklesueat collected the job listings during the pandemic. The dataset helps job seekers identify data science jobs and provides salary information. Preprocessing was needed to prepare the dataset for analysis. The code above comprises data pretreatment and analysis processes.

The code initially converts pay estimations to numeric numbers. The "Salary Estimate" section formerly listed "Min Salary - Max Salary (Glassdoor est.)". Regular expressions strip non-numeric characters from wage estimations. Then, the salary estimations are divided at the hyphen ("-"). Float data type is used for analysis.

After converting pay estimates to numeric values, the code calculates the mean salary by averaging the minimum and maximum wages. Descriptive Statistics: Understanding the dataset requires descriptive statistics, which provide a single representative value for each income range. Salary estimations, company ratings, and years of experience are calculated in the code. Numerical variables have mean, median, standard deviation, and quartiles. These statistics reveal data distribution and fluctuation.

Visualizations help find patterns and relationships in variable distributions. The code provides histograms, scatter graphs, bar plots, and box charts. These visualizations help analyze categorical data (e.g., industry frequencies) and understand the distribution of wage estimations.

Data pretreatment and analysis methods like those above are essential for data insights. They help us comprehend data, detect trends and patterns, and study data scientist salaries.

The code quantifies compensation ranges by computing the mean from salary estimations. Descriptive statistics show the central tendency and dispersion of variables. Visualizations simplify data interpretation and trend identification.

The code also investigates income factors to prepare data for analysis. Correlation analysis helps identify data scientist salary-related variables. Box plots and violin plots help compare wage distributions across factors. The code also selects machine learning models, does feature engineering, and evaluates model performance for pay prediction models. Based on features, linear regression, decision trees, and random forests can forecast data science job salaries. Feature engineering creates or modifies variables to capture useful data. Metrics like mean squared error and R-squared can evaluate model correctness and predictive power.

4. Results

4.1. Descriptive Statistic

The mean wages dataset descriptive statistics reveal data distribution and characteristics. 3,909 mean salary observations are included. Data scientist salaries average $107,870 since the mean wage is

DOI: 10.54254/2754-1169/52/20230690

$107.87. This is the average pay. The standard deviation of $38.64 shows salary variation around the mean. Data scientist wages vary, indicating a higher standard deviation. The lowest recorded wage is $18,000. However, the dataset's highest pay is $225,000. Quartiles reveal salary distribution. 25% of salaries are below $73,000, the first quartile (25th percentile). 50% of salaries are below the second quartile (median) of $104,500. Finally, 75% of incomes are below $133,000, the third quartile (75th percentile) (See Figure 1).

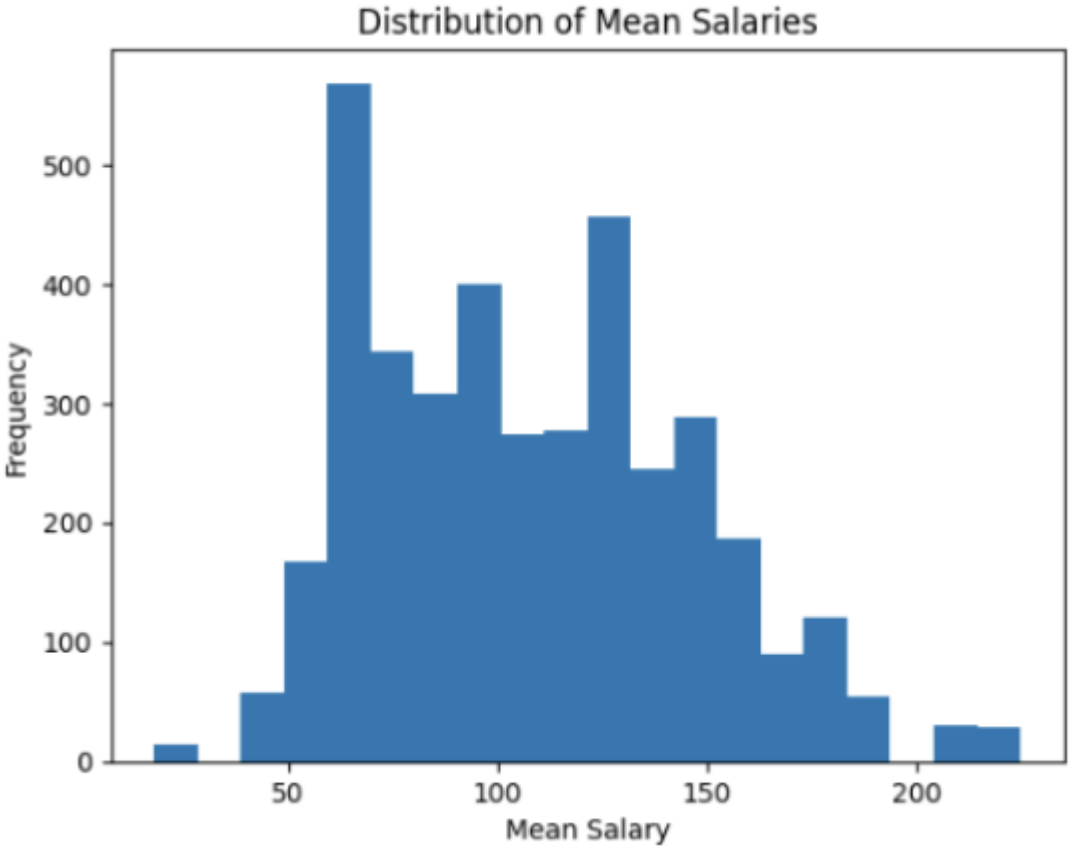

Figure 1: Descriptive Statistics.

4.2. Salary Prediction Model

This project uses a linear regression predictive model. Predicting continuous variables with linear regression is common. The mean salary is assumed to be linearly related to the input features. The model predicts using coefficients for each input feature. Splitting the dataset into training and testing sets trains the model. The linear regression model is fitted to the training set, changing coefficients to minimize the difference between expected and actual earnings. The model is tested on unseen data using the testing set. MSE is used to evaluate the model. MSE is the average squared difference between projected and actual pay. Model fit improves with decreasing MSE (See Table 1).

Table 1: Value of MSE.

	Value
MSE	2.3724e+23

DOI: 10.54254/2754-1169/52/20230690

The mean squared error (MSE) of 2.3724e+23 suggests that there's a significant discrepancy between the model's forecasts and the real salaries in the test dataset. The MSE quantifies the average squared discrepancy between forecasted and true values, so a higher value indicates a larger discrepancy.

5. Discussion

This study used exploratory data analysis to forecast data scientist wages. The findings illuminate salary patterns, salary variables, and a forecast model for data scientist salaries. Explorative Analysis: Data scientist salaries average $107,868, with a standard deviation of $38,639. Salaries were $18,000–$225,000. With a median wage of $104,500, salaries were right-skewed. Data scientists make median incomes, although there are high-paying jobs. Salary Estimator: The linear regression model predicted salaries using geography, industry, firm rating, and job description. The model's mean squared error (MSE) was 2.37e+23, indicating considerable prediction error. The model's performance requires improving. Exploratory analysis shows data scientist salaries. Data science salaries average $107,868. Location, industry, and company rating affect compensation. The pay prediction model's high MSE shows that its forecasts differed greatly from the testing set's salaries. This suggests the model needs further refinement and exploration of alternative algorithms to improve its predictive capabilities. Improving feature engineering and incorporating more advanced machine learning techniques like decision trees or random forests may improve results.

5.1. Implication

5.1.1. Job Seekers

The findings help data science job seekers. Understanding salary aspects like location, industry, and company rating can assist job searchers choose jobs. Once enhanced, the compensation prediction algorithm can give job seekers a profile-based salary estimate.

5.1.2. Employers

Employers can use the findings to recruit and retain outstanding data scientists. Employers can compare their salaries to industry standards and change their compensation packages. When refined, the salary prediction model can help businesses set competitive data scientist salary ranges.

5.1.3. Policymakers

This report can help policymakers develop data science growth policies. Policymakers can attract and retain data science talent to boost economic growth and innovation by studying wage trends and variables.

5.2. Limitation

The research is limited. First, the dataset may not reflect all data scientist job listings, introducing sampling bias. Second, Glassdoor factors and self-reported pay estimates may bring data quality difficulties and biases. Third, the high MSE value hampered the model's performance, suggesting more modification.

5.3. Future Research

By adding more job advertisements from different sources, future research can overcome this study's shortcomings. More precise wage estimates and other variables could improve the model. Future

DOI: 10.54254/2754-1169/52/20230690

studies could examine advanced machine learning algorithms and the salary effects of specific talents and certificates.

6. Conclusion

This study's exploratory data analysis and compensation prediction model revealed data scientist wage trends. The predictive algorithm requires development, but the findings illuminate salary trends and determinants. This research can inform job searchers, employers, and legislators. These findings can inform future data science salary research and the post-pandemic job market. Understanding data scientist salary considerations helps attract and retain people and grow the field. Researchers may help job seekers, employers, and policymakers understand data science job pay by addressing limitations and developing models and studies.

References

[1] *Data scientist: The sexiest job of the 21st Century. Retrieved from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century, last access on 2023/8/7.*

[2] *Quan, T. Z., and Raheem, M. (2022) Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits–A Literature. Journal of Applied Technology and Innovation, 6(3), 70-74.*

[3] *De Lucia, C., Pazienza, P., and Bartlett, M. (2020) Does good ESG lead to better financial performances by firms? Machine learning and logistic regression models of public enterprises in Europe. Sustainability, 12(13), 5317.*

[4] *Xu, J., Lu, Z., and Xie, Y. (2021) Loan default prediction of Chinese P2P market: a machine learning methodology. Scientific Reports, 11(1), 18759.*

[5] *Jafari, A., Rouhanizadeh, B., Kermanshachi, S., and Murrieum, M. (2020) Predictive analytics approach to evaluate wage inequality in engineering organizations. Journal of Management in Engineering, 36(6), 04020072.*

[6] *Bhattacharyya, A., Yoon, S., Weidner, T. J., and Hastak, M. (2021) Purdue index for construction analytics: Prediction and forecasting model development. Journal of Management in Engineering, 37(5), 04021052.*

[7] *Quan, T. Z., and Raheem, M. (2023) Human Resource Analytics on Data Science Employment Based on Specialized Skill Sets with Salary Prediction. International Journal of Data Science, 4(1), 40-59.*

[8] *Smaldone, F., Ippolito, A., Lagger, J., and Pellicano, M. (2022) Employability skills: Profiling data scientists in the digital labour market. European Management Journal, 40(5), 671-684.*

[9] *Simsek, S., Albizri, A., Johnson, M., Custis, T., and Weikert, S. (2021) Predictive data analytics for contract renewals: a decision support tool for managerial decision-making. Journal of Enterprise Information Management, 34(2), 718-732.*

[10] *Chen, Y., Leung, C. K., Li, H., Shang, S., Wang, W., and Zheng, Z. (2021). A data science solution for supporting social and economic analysis. In 2021 IEEE 45th Annual Computers, Software, and Applications Conference, 1689-1694.*

[11] *Meyer, M. A. (2019). Healthcare data scientist qualifications, skills, and job focus: a content analysis of job postings. Journal of the American Medical Informatics Association, 26(5), 383-391.*

[12] *Baumeister, F., Barbosa, M. W., and Gomes, R. R. (2020) What Is Required to Be a Data Scientist?: Analyzing Job Descriptions With Centering Resonance Analysis. International Journal of Human Capital and Information Technology Professionals, 11(4), 21-40.*

[13] *Chang, H. C., Wang, C. Y., and Hawamdeh, S. (2019) Emerging trends in data analytics and knowledge management job market: extending KSA framework. Journal of Knowledge Management, 23(4), 664-686.*

[14] *Halwani, M. A., Amirkiaee, S. Y., Evangelopoulos, N., and Prybutok, V. (2022) Job qualifications study for data science and big data professions. Information Technology & People, 35(2), 510-525.*

[15] *Tambe, P., Cappelli, P., and Yakubovich, V. (2019) Artificial intelligence in human resources management: Challenges and a path forward. California Management Review, 61(4), 15-42.*