Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits – A Literature Review

Tee Zhen Quan
School of Computing
Asia Pacific University of Technology
and innovation (APU)
Kuala Lumpur, Malaysia
TP048200@mail.apu.edu.my

Mafas Raheem
School of Computing
Asia Pacific University of Technology
and innovation (APU)
Kuala Lumpur, Malaysia
raheem@apu.edu.my

***Abstract*— This paper aims to review the recent and existing methodologies for building a more suitable salary prediction model based on specialized skills and given job benefits in the Data Science field. The knowledge discovery also includes identifying existing human resource problems in the data science field and the most demanded skillset for early exploration and determination of input variables. As data science involves a high dimension of positions and responsibilities, the experimental dataset was projected to include skill-based and job benefits factors for more accurate salary predictions. The reviewed benchmarking machine learning methodologies on related problems are categorized into three main categories with individual strengths under different situations and requirements. Statistical methods are better in presenting variable relationships with extraordinary parameter tuning potential if linearity is present. Ensemble machine learning methods like Random Forest that combines multiple classifiers for more stable and accurate prediction. Deep learning-based neural networks have a strong specialty in handling unlabeled data and framework modifications. Moreover, it was realized that huge datasets with appropriate variables and grid search tuning method achieves greater and more reliable performance. However, extraordinary research on data science-related job benefits could be conducted if sufficient studies were present during that period. Overall, further work is necessary to determine the project's objectives, scenario and experimental dataset to select suitable reviewed methodologies for the data science salary prediction model building.**

Keywords — data science, predictive model, specialized skill sets, machine learning, data mining

I. I. Introduction

The industry revolution 4.0 has already begun with the main objectives to improve organization work efficiency, effectiveness, maintainability and even growth through digitalization specifically during this Covid 19 pandemic (V.Sindhu et al., 2021). This is achieved by developing a cooperative ecosystem between different stakeholders like suppliers, manufacturers, customers and more to reach greater business or organizational objectives. Resolving manufacturing bottlenecks by identifying current issues to produce an even more optimized production line is one of the popular examples. Moreover, building predictive models to foresee sales demands and equipment maintenance deadlines are crucial in preventing excessive manufacturing and production delay due to hardware failure.

Internet of Things (IoT) and big data being the main focuses in industry revolution 4.0, enormous data science jobs are required to collect data, process it using machine learning or statistical techniques and most importantly report to the organization's management comprehensibly. Data science jobs are in serious shortage over the globe. As reported by McKinsey, there was already a shortfall of 250 000 data science-related jobs in the United States alone by 2016 (Jain, 2019). The same publisher has also mentioned that such a shortage is never seen before in other industries. Even with ready-to-use data science tools current days, the shortage in the United States has not been reduced based on recent reports from the same sources (DuBois, 2020). This leads to a few important reasons why data keeps generating at high speed requiring more data science professionals to fully utilize it for useful business insights. Furthermore, the success of Netflix and other leading e-commerce and social media companies has motivated others to develop their machine learning and predictive algorithms that can solve their specific problems best.

This has clearly shown a strong need for building a data science-specified salary prediction tool for the United States. An ordinary salary predictor allows professionals to estimate the salary and job opportunity based on common variables like job title, location, working experience and qualification. Unlike common jobs, data science is a profession with many specialized skill sets wanted specifically by different organizations. A salary prediction model for data science will accurately estimate the salary based on specialized variables for the field such as programming skills, data analytics tools and many more. Most importantly, allowing professionals and students to understand important and demanded skills. This does not only add value to personal salary and career development but also the whole growth of the data science industry.

Besides, this salary predictor will not only be useful for employees but also for the employers to study the current recruitment trend for better budget planning in human resource compensation and benefits. With the inclusion of job benefits as input variables, organizations can also understand how the bundle of employee benefits could affect the salary and even possibly reduce the recruitment cost. This strategy is significant to promote business expansion and success as mentioned in an Information Technology (IT) salary prediction research (More et al., 2021). Because accurate recruitment is always the main factor to improve an organization's market competitiveness and work productivity as employees are the company's biggest asset. Further, the predictive tool can also work as a motivator and guide for students to persuade their studies and skills in a particular field which can gradually resolve the mentioned problem of data science job shortage in the United States.

A salary prediction model based on specialized skills and given job benefits can effectively resolve the problems and needs in the data science field. This paper focuses on domain and knowledge discovery in related salary prediction studies including determining the most demanded data science skillsets for features selection and reviewing the existing methodologies. Finally, the conclusion session summarizes the findings of suitable benchmarking models with different techniques. Moreover, it identifies important data mining, visualization and analysis processes for effective model building and business objectives achievements.

II. II. Different Job and Skill Domains of Data Science

A. *Job Titles*

Data science is a massive domain that involves multiple components like data mining, data analytics, machine learning and many more. This requires multiple data science positions to perform specialized and individual tasks. To build the predictive model effectively, it is mandatory to understand the basic domains of data science including the common job titles and specialized skills. Data Scientist and Data Analyst are the two commonly known titles which both might have similar tasks but different responsibilities. Data scientist is the most commonly known title and has the biggest responsibility and strongest skillset. Based on the definition from CIO magazine, data scientist focuses on algorithm research and development while having comprehensive skills to deal with any project tasks (Olavsrud, 2020). The same author mentioned that data analyst specifically transforms data into useful information from data collection, analysis, and visualization to reporting for business insights. Metwalli (2020) stated that other job titles include data engineer to design, construct and maintain data pipelines and data architect to design and create the databases. Both positions prepare and maintain the environment for data scientists and data analysts to perform their tasks.

Data storyteller is another specialized role for data visualization which focuses on the end process of reporting. It works as a medium to convert technical information into a better and more understandable story for business purposes. Using this information, business intelligence developer comes out with business strategies for important decision-making. To dive more into technical job titles, machine learning scientists specifically research and develop new algorithms with a closer relation to academic work. Differently, machine learning engineer applies the developed algorithm and run comparison tests to obtain the most optimized model for the particular project condition. Overall, the comparisons of all mentioned titles have contrasted that data scientist is at the highest job level among other positions that are very specific to certain tasks.

A. *Specialized Skill Sets*

It requires a specific skill in a particular tool, programming language, library, or database to perform any task mentioned in section A. Each job title requires different skill sets to perform their data science-related tasks. Due to the huge availability of data science tools and techniques, it is important to identify the most demanded skills to narrow the input variables for building the data science salary prediction model. According to an analysis of 15 000 data science job listings, the most popular data science skills in 2021 include Python, SQL, R, Spark, AWS, Java, Tableau, Hadoop, TensorFlow, Scala, SAS, Azure, Scikit-learn and many more (Shin, 2021).

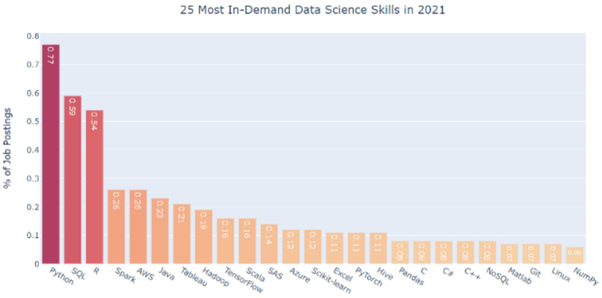

Fig. 1. 25 Top Data Science Skills of 2021 (Shin, 2021).

III. III. Existing Methodologies

B. *Statistical Methods*

Focusing on statistical techniques, Das et al. (2020) performed salary prediction using simple graph representation with regressions. Linear regression was first tested, but the straight line doesn't perfectly fit the scattered data points. Polynomial regression represented a curve line and was determined as the better solution to resolve some non-linearity between the salary and job position level. Despite these two variables having a linear relationship, the top few highest position level resulted in an enormous increase in salary value which is very common in a real industrial scenario. In addition, curve fitting was done to perform smoothing on the data for better prediction. The advantage of the statistical regression technique allows direct and clear observation of future position influence on salary through visualization. Although the prediction was straightforward, data mining software tools could be further used to obtain more accurate value and accuracy evaluation.

From an experiment carried out by Lothe et al. (2021), the linear regression base model was firstly created by measuring the input variable one by one against the target variable. The initial accuracy was extraordinary with 96-98% when the input work experience variable is tested followed by the job type. However, when the base model was transformed into polynomial regression to handle multiple input variables in one prediction, the final accuracy was reduced to 76% using Mean Square Error (MSE) evaluation. This showed that predicting salary using a single attribute with single linear regression could be much simple and more accurate but is highly unrealistic as salary generally depends on many variables in real cases. Also, cross-validation could be applied for a more accurate and reliable accuracy evaluation.

Besides performing quantitative analysis on salary factors, Pawha & Kamthania (2019) also proposed a statistical multiple linear regression model for salary prediction to handle multiple input variables for prediction. Differently, the dataset used consisted of candidates' examination scores upon the usual employment details. Even though the score did not show a linear relationship against the salary, the proposed model still achieved a Root Relative Squared Error (RRSE) score of around 82% through other useful variables like programing, language, quantitative and logical skills. This clearly showed the importance of skills development over examination scores. The accuracy improvement emphasized data cleaning and solving outliers, which is critical for regression problems. Nonetheless, the authors mentioned that more than half of the engineers were underpaid, this could reduce the dataset's veracity for predicting salary in other scenarios where underemployment is seldom.

C. Modern Machine Learning Methods

Dutta et al. (2018) proposed a modern tree-based prediction engine for salary prediction. With advanced use of data analysis, two uncorrelated variables such as contract type and contract time are removed. Since the dataset was left-skewed with much more low salary values, a new dataset is extracted using the log function which distributed the data more evenly. These operations effective performed noise reduction and unbalanced dataset, hence gradually increasing the prediction accuracy. The authors have proven that random forest produced higher accuracy of 87.3% compared to the ordinary decision tree which only has 84.8% accuracy. It is because this ensemble method combined multiple decision trees which predicted more reliably and accurately. However, the weakness of the tree-based model remained where the model could overly rely on the dataset instead of regression methods which specifics more on the variable relationships.

Similarly, Zhang & Cheng (2019) proposed a KNN classifier to predict salary specifically for Java back-end engineers with Java specialized skills as input variables. As KNN is non-perimetric, the authors focused mainly on distance measurement and K point selection, then finally determined the best K Value. The model obtained its best average accuracy of 88.1% when K Value is 7. This method showed a strong advantage in minimum parameter tuning, but the K value can increase with the size of the dataset which will require more experiments. The highest salary level prediction resulted in exceptional accuracy of 93.3% with an almost 20% decrease on the lowest salary level, this showed a certain bias in the proposed KNN model.

D. Deep Learning Neural Network

Sun, et al. (2021) proposed a different approach using an improved and cooperative neural network. It performed a two-step modelling where the skills were first passed to the main skill valuation model. Then, the estimated value was used to assist the cooperative salary prediction model. This modified framework allowed the main model to be retrained with the cooperative model's feedback. Exceptional from the previous approaches, this deep learning technique could extract information and perform predictions directly from unlabeled job listing data. Using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) evaluation methods, the proposed model resulted in the lowest error values compared to current exiting models like linear regression, Support Vector Machine (SVM), gradient boosting, text-mining methods and even the ordinary deep learning neural network designed with the exact variables. Even so, the authors were only able to evaluate the salary prediction model's accuracy due to a lack of factual validation for the skill valuation during the research period.

Wang et al. (2019) built a Neural Network model with the combination of Bidirectional Gated Recurrent Unit (GRU) and Conventional Neural Network (CNN). This proposed model also obtained the direct contextual data as input and performed Natural Language Processing (NLP) for continuous salary value prediction. The inclusion of GRU allowed the enhanced model to extract better information from single words when dealing with noisy data. The final features learning was then performed ordinary on the CNN layer with the assistance of ResNet. With comparisons of different deep learning neural network models, only the proposed model outperformed the dominant TextCNN with the least MAE while other combination models seemed to be less effective. However, with the increased complexity of the combined frameworks, training time was significantly more than conventional TextCNN.

Zhu (2021) proposed a Backpropagation Neural Network model (BPNN) using assisting gradient descent algorithm for salary prediction. The basic 3-layered BPNN model used got 1 input layer with 14 input neurons, 15 hidden layers and 1 output layer. The gradient descent algorithm was used to perform the backpropagation update for greater accuracy. The accuracy validation processes showed that the increase of neurons and hidden layers resulted in greater accuracy but only until a certain peak point. By also experimenting with different gradient descent techniques, Small-Batch Gradient Descent (SBGD) with Nadam optimization was determined as the best method to update the model's parameters. The final accuracy value obtained after validation was 89.98%. Since only 100 records were used for the validation where the model's accuracy evaluation might not be highly reliable.

By encoding the job data into graph representations, a new deep learning method called Graph Convolutional Network (GCN) was implemented by Chen et al. (2020). The semi-supervised model was designed to mainly work on unlabeled job posting data despite certain data being labelled manually to improve salary classification accuracy. The features extracted were categorized into seven metadata features for prediction and an additional relationship feature to represent the similarity score between two jobs. The authors realized that the inclusion of a spectral filter in the proposed GCN semi-supervised model further improved the accuracy to 77%. For comparison purposes, an exclusively labelled dataset was run on existing supervised learning classifiers and the random forest ensembled model outperformed again with 76% accuracy with metadata features. In this supervised learning comparison, deep learning neural networks were the most underperformed with the accuracy between 60% to 63% but were still the only suitable method to explore unlabeled data with great potential for accuracy improvement using a bigger and sufficient dataset. Even more, the simple GCN model could still be enhanced using Region-based Convolutional Neural Network (RCNN) techniques.

E. Methods Comparison

Mart´ın et al. (2018) implemented an accurate classification model to predict salary ranges with 4000 IT job offers. The comprehensive comparison was completed on several classification models such as logistic regression, KNN, Multilayer Perceptron Models (MLPs), Random Forests, SVMs, AdaBoost and voting classifiers using accuracy evaluation measures such as F1 score, ROC curve and Precision-Recall curve with cross-validation. As a result, the Random Forest model got the highest accuracy of 84% closely followed by voting classifiers. Besides, the study also discovered that the predictions were highly affected by 5 IT skill-based profiles and predicting the salary in ranges helps improve the accuracy. Other accuracy improvement techniques highlighted in the paper included preprocessing to reduce data dimensionality and perform optimization using the grid search technique on all experimental models' parameters. However, the authors did mention a weakness of insufficient records which is a critical factor for some classifiers like the regression models.

More et al. (2021) proposed a skill-based focused salary prediction model specifically for fresh graduates. The authors converted the original salary values into a classification problem for accuracy improvement purposes. Besides the ordinary technical skills, the user's verbal skills were also analyzed for salary prediction using sentiment analysis with the support of NLP. Since continuous values were used on technical and verbal skills scoring, MAE, RMSE, Relative Absolute Error (RAE) and Relative Squared Error (RSE) were used to evaluate the accuracy of experiment models. The linear regression model resulted the highest accuracy of 94.29% with the Random Forest at 91.07%. The obvious disadvantage of the proposed system was it got very few independent variables. Inclusion of additional information like recruitment location and position level could help to predict salary in a more realistic manner.

Using a self-collected dataset for the same salary prediction problem, Gomez-Cravioto et al. (2022) extensively compared the outcome of traditional statistical or parametric models and modern non-parametric models specifically the ensemble tree based. The tested parametric models include Quantile Regression, Linear Regression and Logistic Regression while Random Forest and Gradient Boosting were used as non-parametric models. Under the experiment on both classification and regression tasks, Gradient Boosting model performed the best, followed by Random Forest and Regression. Due to the ability to handle both linear and non-linear relationships, the Tree-Based Gradient Boosting and Random Forest models outperformed the high-dimensional dataset. Nevertheless, statistical models still allow better interpretation of the relationships between variables. The evidence of unfair comparison was present as the parametric and non-parametric models used different feature selection methods causing the individual models to be possibly different.

IV. Conclusions

Many studies highlighted the powerful influence of skill factors on salary prediction besides the data science field. Moreover, classifying salary prediction results in higher accuracy instead of predicting continuous salary values. A balanced data dimension and a good amount of input variables will help produce unbiased yet accurate predictions. With different datasets used in individual studies, it cannot be determined which particular method performs the best. However, reviews have shown that statical methods like regression models can represent variables' relationships better with great space to improve accuracy through intensive parameters tuning, but mostly under scenarios of independent variables with linear relationships and minimum outliers. For general cases where the dataset is highly dimensional, other machine learning models produce more accurate predictions by handling both linear and non-linear relationships, especially the tree-based models. Deep learning neural network techniques have shown their superiority in processing contextual data directly from job postings, allowing efficient data mining on a larger scale without labelling and structuring raw data. The comparisons presented that machine learning or non-perimetric models benefit from minimum parameter tuning which is the most time-intensive task in data mining. Further, the grid search technique appeared to be necessary and much more effective in all cases than the randomized search technique in model tuning as it explores all possible combinations with the guarantee to obtain the most optimized parameter configuration. With studies using thousands to hundred thousand records, a huge and historic dataset is critical to ensure the model's reliability for longer future prediction. Relying on sufficient information collected from this comprehensive study, the further work of building an accurate data science salary prediction model based on specialized skills and given job benefits can be effectively performed.

REFERENCES

Chen, L., Sun, Y., & Thakuriah, P. (2020). Modelling and Predicting Individual Salaries in United Kingdom with Graph Convolutional Network. *Hybrid Intelligent Systems: Advances in Intelligent Systems and Computing*, 61-74.

Das, S., Barik, R., & Mukherjee, A. (2020). Salary Prediction using Regression Techniques. *International Conference on Industry Interative Innovations in Science and Engineering*, 1-5.

DuBois, J. (2020). *The Data Scientist Shortage in 2020*. https://quanthub.com/data-scientist-shortage-2020/

Dutta, S., Halder, A., & Dasgupta, K. (2018). Design of a novel Prediction Engine for predicting suitable salary for a job. *Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 275-279.

Gomez-Cravioto, D. A., Diaz-Ramos, R. E., Hernandez-Gress, N., Preciado, J. L., & Ceballos, H. G. (2022). Supervised machine learning predictive analytics for alumni income. *Journal of Big Data*, 9(1), 1-31.

Jain, K. (2019). Big job opportunities in data science & machine learning. *Express Computer*, 1-3.

Lothe, P. D., Tiwari, P., Patil, N., Patil, S., & Patil, V. (2021). Salary Prediction using Machine Learning. *International Journal of Advance Sciencetic Research and Engineering Trends*, 6(5), 199-202.

Mart´ın, I., Mariello, A., Battiti, R., & andez, J. A. (2018). Salary Prediction in the IT Job Market with Few High-Dimensional Samples: A Spanish Case Study. *International Journal of Computational Intelligence Systems*, 11(1), 1192-1209.

Metwalli, S. A. (2020). *10 Different Data Science Job Titles and What They Mean*. https://towardsdatascience.com/10-different-data-science-job-titles-and-what-they-mean-d385fc3c58ae

More, A., Naik, A., & Rathod, S. (2021). PREDICT-NATION Skills Based Salary Prediction for Freshers. *Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021)*.

Olavsrud, T. (2020). *What is a data analyst? A key role for data-driven business decisions*. https://www.cio.com/article/217583/what-is-a-data-analyst-a-key-role-for-data-driven-business-decisions.html#:~:text=Data%20analysts%20work%20with%20data,predict%2C%20and%20improve%20business%20performance.

Pawha, A., & Kamthania, D. (2019). Quantitative analysis of historical data for prediction of job salary in India - A case study. *Journal of Statistics & Management Systems*, 22(2), 187-198.

Shin, T. (2021). *The Most In-Demand Skills for Data Scientists in 2021*. https://towardsdatascience.com/the-most-in-demand-skills-for-data-scientists-in-2021-4b2a808f4005

Sun, Y., Zhuang, F., Zhu, H., Zhang, Q., He, Q., & Xiong, H. (2021). Market-oriented job skill valuation with cooperative composition neural network. *Nature Communications*, 12, 1-11.

V.Sindhu, Anitha, G., & Geetha, R. (2021). Industry 4.0-A Breakthrough in artificial Intelligence the Internet of Things and Big Data towards the next digital revolution for high business outcome and delivery. *Journal of Physics: Conference Series*, 1937, 1-7.

Wang, Z., Sugaya, S., & Nguyen, D. P. (2019). Salary Prediction using Bidirectional-GRU-CNN Model. *The Association for Natural Language Processing*, 292-295.

Zhang, J., & Cheng, J. (2019). Study of Employment Salary Forecast using KNN Algorithm. *International Conference on Modeling, Simulation and Big Data Analysis*, 166-170.

Zhu, H. (2021). Research on Human Resource Recommendation Algorithm Based on Machine Learning. *Scientific Programming*, 1-10.