DOI: 10.54254/2755-2721/50/20241102

The investigation and prediction for salary trends in the data science industry

Wentao Jiang

School of Science, Rensselaer Polytechnic Institute, 12180, USA

jiangw8@rpi.edu

Abstract. The aim of this study is to utilize machine learning techniques to analyze salary trends within the data science industry spanning the last three years. Initially, this study presented an overview of four machine learning models: Random Forests, eXtreme Gradient Boosting (XGBoost), Neural Networks, and Support Vector Regression (SVR), elucidating their fundamental principles and characteristics. Subsequently, this study gathered, preprocessed, and engaged in feature engineering with salary data from the data science sector over the past three years. These four machine learning models are then employed for salary prediction, and the ensuing model outcomes are meticulously examined. By conducting a comparative analysis and evaluating each model's performance, their respective strengths and weaknesses were identified. In conclusion, this study summarized the study's findings and deliberated on potential future research directions. The innovation inherent in this research lies in the application of diverse machine learning models to forecast salaries within the data science industry, coupled with the comprehensive comparison and evaluation of these models. The main conclusion is that XGBoost performs best in salary prediction, while neural networks are more accurate and complex, and SVR has limited applicability. Future research prospects include improving the accuracy and interpretability of models, exploring more features and data processing methods to enhance the accuracy and practicality of salary prediction in the data science industry.

Keywords: Machine Learning, Neural Network, Salary Prediction.

1. Introduction

The digital revolution of recent years has made Data Science and Machine Learning (ML) indispensable tools of great potential for deciphering vast data frameworks and executing sophisticated analysis. The data science industry has become an important pillar for both research and commercial applications, transforming the way comprehend and manipulate interlinked digital information. However, an essential focal point within the realm of the data science field that has garnered substantial interest in recent times pertains to the compensation packages. Compensation in this sector has perpetually piqued curiosity owing to the intricate nature of the roles, dynamic skill prerequisites, and the rapid growth of the industry. The objective of this study is to shed light on this specific facet by employing machine learning methodologies to delve into the patterns, factors, and salary frameworks within the industry over the past three years.

Prior studies have utilized classical statistical methods and straightforward data exploration to gain insights into salary trends [1-3]. However, the nuances and interconnections within the salary structures

© 2024 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

call for more advanced techniques. This is where machine learning models come in, offering the tools to automatically detect patterns in observed data and predict future outcomes of interest. The present study aims to fill this research gap, extending the scope to harness the power of various ML models for interpreting and predicting salary structures. This paper will pivot its focus to the theoretical assumptions and approaches to ML and Data Analysis, emphasizing the application of these techniques to the salary data accumulated over the past three years.

This study aims to provide substantive results and implications that could lead to a more thorough understanding of the salary dynamics in the data science industry. From a more practical perspective, our study will shed light on the salary growth in the data science sector and can serve as an essential reference for industry professionals and companies alike.

The genesis of the data science industry can be traced back to the advent of the digital era, a period marked by an explosion of electronically available data. In recent years, with the progressive advancement of technology and the broadening integration of internet services across all sectors, the volume of digital data has seen an unprecedented surge. Within the confines of this digitally-infused environment, data science respectfully emerged and rapidly established itself within the echelons of essential interdisciplinary fields.

Data science, intrinsically linked with ML and Data Analysis, seeks to transform raw, unstructured, and often intricate data into coherent, valuable insights useful for decision-making and strategies formulation. Drawing influence predominantly from statistics, computer science, and information science, it delves deep into the realms of data extraction, preparation, and interpretation, presenting a more articulate perspective of the labyrinth data landscape. As a closely knitted component of the wider tech industry, the data science industry has demonstrated an appetizing growth trajectory over the last three years. This progression, largely underpinned by a global shift towards data-driven decision-making mechanisms, is catalyzed by the increasing demand across various sectors such as finance, healthcare, commerce, and the public sector, to mention but a few. The growing need for data scientists has created a vibrant labor market, with businesses seeking to harness the capacities offered by big data analytics. However, the job profiles in this industry are diverse, with roles including data analysts, data engineers, and machine learning specialists among others, each with its unique skills requirements and remuneration structures.

While it's widely acknowledged that the data science industry is on an upward trajectory, there has been a noticeable lack of comprehensive information regarding the precise nature and patterns of salary trends within this sector over the past three years. This paper strives to address this research gap by providing a thorough examination of the salary trends within the data science industry during the specified timeframe.

2. Introduction to machine learning models

2.1. Random forest model

Random forest is an integrated model widely used in the field of machine learning [4, 5]. It consists of multiple decision trees, which integrate the predicted results of these decision trees for classification or regression tasks. Random forests have high reliability and flexibility in salary research in the data science industry.

The random forest model has performed well in many practical applications. It can handle various types of data, including continuous, discrete, and classified data. In addition, random forests can also handle high-dimensional datasets and data with missing values. These advantages make the random forest model a powerful tool for studying salaries in the data science industry.

2.2. XGBoost model

Extreme Gradient Boosting (XGBoost) is a machine learning model based on gradient lifting trees [6, 7]. It gradually improves the accuracy of prediction results by iteratively adding tree models. XGBoost is widely used in the data science industry and is considered an efficient and accurate model.

In salary research in the data science industry, the XGBoost model can be applied to salary prediction and feature analysis. The XGBoost model can be used to establish a predictive model that predicts a person's salary level in the data science industry by inputting various relevant features, such as work experience, education, skill level, etc. At the same time, the feature importance assessment function of the XGBoost model can be also employed to analyze which features have the greatest impact on salary levels, thereby providing reference for salary strategies and talent cultivation.

2.3. Neural network model

A neural network model is a computational model based on the neural system, which simulates the connection between neurons in the human brain and performs learning and prediction tasks through the interconnection and information transmission between multiple neurons [8, 9]. In the field of data science, neural network models are widely used in salary prediction and analysis tasks. The neural network model has strong nonlinear fitting ability, which can identify and capture complex patterns and correlation relationships in data. In salary research, neural network models can be employed to model and predict the complex relationship between multiple factors and salary. Compared with other models, neural network models have the following characteristics and advantages.

It should be noted that neural network models also have some challenges and limitations. Firstly, the training of neural network models requires a large amount of annotated data, especially in situations where data is scarce. It is necessary to reasonably use techniques such as data augmentation and transfer learning to improve the training effectiveness of the model. Secondly, the structure of the neural network model and the selection of hyperparameters have a significant impact on the performance and effectiveness of the model, requiring reasonable parameter tuning and optimization.

2.4. SVR model

Support Vector Regression (SVR) is a regression method of Support Vector Machine, widely used in the field of data science [10, 11]. This model can effectively handle high-dimensional data and nonlinear relationships by transforming regression problems into solving a linear problem. The core idea of the SVR model is to find a hyperplane so that data points are distributed as much as possible within the intervals of the hyperplane.

The basic principle of the SVR model is to construct a hyperplane in the feature space, maximizing the distance between sample points and the hyperplane. Usually, SVR models use nonlinear kernel functions to overcome the nonlinear properties of data, such as radial basis function (RBF). In this way, the SVR model can map low dimensional data to high-dimensional space through nonlinear transformation, thereby better fitting complex data relationships.

It is worth noting that when using the SVR model, attention should be paid to feature selection and data preprocessing. Reasonable selection of features can reduce the complexity of the model and improve the accuracy of prediction. The preprocessing of data includes standardization of features, handling of outliers, and filling in missing values, which can improve the robustness and generalization ability of the model.

3. Datasets description and preprocessing

3.1. Data collection and preprocessing

To conduct research on the salary trends within the data science industry over the past three years, relevant data collection was initially performed. The data source utilized for this research was the Kaggle dataset. The data sources were carefully screened and integrated to create a comprehensive dataset.

During the data preprocessing stage, attention was given to tasks such as data cleaning, handling missing values, and detecting outliers in the collected data. Initially, the collected data underwent a cleaning process, which involved the removal of duplicate and invalid data samples. This step was crucial to ensure the accuracy and completeness of the dataset under analysis.

Subsequently, the focus shifted to addressing potential missing values within the data. Missing values are a common occurrence in real-world data and can arise from various sources, including human errors, equipment failures, or other objective factors. To prevent missing values from impacting the training and predictive performance of the model, standard missing value processing methods were employed. These methods included techniques such as mean interpolation, median interpolation, or the application of machine learning algorithms to predict missing values.

In addition to handling missing values, efforts were made to identify and manage potential outliers within the dataset. Outliers represent unique observations that significantly deviate from the rest of the data. In the context of salary research within the data science industry, outliers could stem from extremely high or exceptionally low salary figures. While one approach to handling outliers involves their removal from the dataset, it was recognized that outliers may contain valuable insights. Therefore, alternative strategies, such as replacing outliers with adjusted values, scaling, or employing statistical methods, were considered to maintain data integrity and accuracy.

The final step in the data preprocessing phase involved feature engineering. This step entailed the transformation and combination of the original data to generate new feature variables that better described and conveyed the data's characteristics. Common feature engineering techniques, including polynomial features, interaction features, discretization, and regularization, were applied. Through feature engineering, the objective was to enhance the model's performance and predictive capabilities, further refining the research outcomes.

3.2. Data description

This dataset aims to shed light on the salary trends in the field of Data Science for the years 2021 to 2023. With a focus on various aspects of employment, including work experience, job titles, and company locations, this dataset provides valuable insights into salary distributions within the industry.

With this dataset, data enthusiasts and analysts can delve into the salary dynamics of Data Science professionals in 2023, identifying trends across different experience levels, job titles, and company sizes. It can be a valuable resource for understanding the economic landscape in the Data Science job market and making informed decisions for both job seekers and employers alike.

4. Model application and result analysis

4.1. Application of random forest model to salary prediction

The study employed a random forest model, a robust ensemble learning technique comprising multiple decision trees trained with randomly selected features and samples, to predict and analyze salaries in the data science industry over the past three years. Salary data, along with pertinent attributes like workplace, company size, work experience, and specific roles, were gathered.

The dataset was divided into training and testing sets for model application. The random forest model was iteratively fine-tuned, yielding high accuracy. The model's performance was assessed using the test set, quantifying prediction errors. Multiple experiments and cross-validation revealed significant findings. Work experience and country of employment significantly influenced salaries, with increasing experience correlating with higher salaries. Developed countries and senior positions also yielded higher income levels. Feature importance analysis identified workplace as the most critical predictor, surpassing others in its impact on salary. This random forest approach furnished an accurate method for salary prediction, offering valuable insights for practitioners and decision-makers. Future research will explore alternative models and conduct deeper data analysis to enhance accuracy and interpretability in salary forecasting.

4.2. XGBoost model applied to salary prediction

The XGBoost model, a gradient-based machine learning algorithm renowned for its stellar performance, is a valuable tool in predicting salaries within the data science industry. The process begins with data preprocessing, involving the standardization of collected salary data from the past three years and associated characteristics.

The dataset is then split into training and testing sets, with the former serving for model training and parameter optimization, and the latter for evaluating predictive performance. To mitigate overfitting, cross-validation techniques are employed for model selection and assessment. Subsequently, the XGBoost model is used to build a predictive model. This model comprises a series of decision tree models, sequentially concatenated and refined through gradient descent. Unlike conventional decision trees, XGBoost adeptly handles non-linear relationships and high-dimensional features. Throughout model training, hyperparameters such as tree quantity, depth, and learning rate are adjusted for optimal performance. Regularization techniques, such as L1 and L2 regularization, are employed to prevent overfitting and enhance model generalization. Once training is complete, the trained XGBoost model is deployed for salary prediction. This section has elucidated the utilization of the XGBoost model for forecasting data science industry salaries, achieving accuracy and stability through thoughtful feature selection, data preprocessing, and model refinement. Future sections will explore alternative models and analyze results comprehensively to discern salary influencers and trends within the data science industry.

4.3. Application of neural network model in salary prediction

The neural network model, a computational framework rooted in artificial neurons and connections, boasts robust nonlinear modeling capabilities. This section delves into the utilization of neural network models for predicting salaries in the data science sector.

Commencing with the application of a neural network model, a deep architecture with multiple hidden layers is constructed using pertinent data spanning the past three years. Key input variables encompass work experience and company location, yielding salary predictions as outputs. Model training occurs on the dataset’s training segment, with parameter adjustments facilitated through cross-validation. Subsequently, an evaluation of neural network model performance in salary prediction transpires. By contrasting predictions from the test dataset against actual salary figures, assessments employ root mean square error and coefficient of determination to gauge model accuracy and fit. Experimental outcomes reveal the neural network’s adeptness in accurately predicting data science industry salaries, evidenced by low root mean square error and high coefficient of determination, signifying precise salary forecasts. Further scrutiny of the neural network model involves visualizing its weights and biases. This elucidates the variables’ significance in salary prediction, spotlighting the substantial impact of country and work experience on salaries. Lastly, model stability tests are executed using data from varying time frames, showcasing the neural network’s proficiency in predicting salaries consistently across different years. This underscores the model’s resilience and reliability in data science industry salary forecasting.

4.4. Application of SVR model in salary predictions

SVR, a regression technique grounded in support vector machines, proves useful for salary prediction in the data science realm. Input features encompass relevant industry attributes like work experience and region, while salary serves as the target variable for modeling and forecasting.

Dataset preprocessing ensures model accuracy and stability. Subsequently, data division into training and testing sets facilitates model training and evaluation. During SVR model training, kernel function selection and hyperparameter determination are crucial tasks. Kernel options encompass linear, polynomial, and radial basis functions, adaptable to specific needs. Hyperparameter selection is accomplished via methods such as cross-validation. Post-training, model evaluation relies on metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) to assess predictive accuracy and stability in salary forecasting.

5. Model comparison and evaluation

5.1. Comparison of model performance

This section conducts a comparative analysis of machine learning models, including Random Forest, XGBoost, Neural Networks, and SVR, applied to data science industry salary data over the past three years.

To begin, the Random Forest model with 100 trees was employed for training. It employs ensemble learning, utilizing multiple decision trees with random feature selection to mitigate overfitting. Cross-validation was utilized to derive prediction results. Following that, the XGBoost model, a gradient boosting tree ensemble method, was explored. Parameters like learning rate and tree quantity were fine-tuned during model training. Performance was evaluated through cross-validation. Additionally, the Neural Network model, based on artificial neural networks, was examined. It employed a multi-layer perceptron (MLP) architecture with ReLU activation functions for enhanced nonlinearity. Adequate optimizers and loss functions were selected for training, with an early stop mechanism to combat overfitting. Model predictive accuracy was assessed via cross-validation. Lastly, the Support Vector Regression (SVR) model, a support vector machine-based regression model, was tested. It was trained with suitable kernel functions and parameters and evaluated through cross-validation.

Evaluation criteria encompassed root mean square error (RMSE) and coefficient of determination (R^2). RMSE quantified prediction errors, while R^2 gauged the models' ability to explain salary variations. Comparative analysis of these metrics was employed to determine the optimal model.

5.2. Model performance evaluation

In this section, a detailed assessment of random forests, XGBoost, neural networks, and SVR models in machine learning for data science industry salary research is conducted. Over three years of salary data are used for a comprehensive model comparison through cross-validation and evaluation metrics calculation.

Firstly, predictive performance is measured employing three commonly used evaluation metrics: root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The random forest model yields RMSE of 48052, MAE of 37179, and MAPE of 0.3077. The XGBoost model results in RMSE of 48644, MAE of 37554, and MAPE of 0.3123. The neural network model exhibits RMSE of 47418, MAE of 36683, and MAPE of 0.3025. The SVR model mirrors these RMSE and MAE figures. Notably, the neural network and SVR models excel in predicting salaries with impressive RMSE, MAE, and MAPE values.

Furthermore, during model evaluation, attention is directed towards training and prediction times. In terms of time efficiency, the XGBoost model stands out for its swiftness in both training and prediction.

To summarize, within this dataset, XGBoost proves the most time-efficient model for data science industry salary research among random forests, XGBoost, neural networks, and SVR models in machine learning. Meanwhile, neural networks and SVR models demonstrate the best salary prediction performance.

6. Conclusion

In summary, this study conducted a study on the salary of the data science industry in the past three years by using random forests, XGBoost, neural networks, and SVR models in machine learning. Through data collection and cleaning, feature extraction, and model construction, this study successfully established four optimized salary prediction models. These models have good predictive performance and can provide valuable reference and decision support for human resources departments and job seekers.

Future research work should further explore how to improve the prediction accuracy and stability of models. Introducing more characteristic variables, such as company size, industry type, etc., can be considered to improve the model's ability to explain salary. In addition, during the model construction process, other machine learning algorithms and deep learning methods can also be tried to improve the performance of the model. In addition, the application of salary prediction models can also be extended to other fields, such as finance, healthcare, etc., to meet a wider range of needs. Continuing to conduct in-depth research in this field will help provide more accurate and efficient solutions for salary forecasting issues in the data science industry.

References

[1] Muench U Sindelar J Busch S H and Buerhaus P I 2015 Salary differences between male and female registered nurses in the United States Jama 313(12) 1265-1267

[2] Ghaznavi C et al 2022 Salaries, degrees, and babies: Trends in fertility by income and education among Japanese men and women born 1943–1975—Analysis of national surveys Plos one 17(4) e0266835

[3] McDonald J B and Sorensen J 2017 Academic salary compression across disciplines and over time. Economics of Education Review 59 87-104

[4] Rigatti S J 2017 Random forest Journal of Insurance Medicine 47(1) 31-39

[5] Biau G 2012 Analysis of a random forests model The Journal of Machine Learning Research 13 1063-1095

[6] Chen T et al 2015 Xgboost: extreme gradient boosting R package version 0.4-2 1(4) 1-4

[7] Chen T He T Benesty M and Khotilovich V 2019 Package ‘xgboost’ R version 90 1-66

[8] Abiodun O I et al 2018 State-of-the-art in artificial neural network applications: A survey Heliyon 4(11)

[9] Lawrence J 1993 Introduction to neural networks. California Scientific Software.

[10] Smola A J and Schölkopf B 2004 A tutorial on support vector regression Statistics and computing 14 199-222

[11] Awad M Khanna R Awad M and Khanna R 2015 Support vector regression. Efficient learning machines: Theories, concepts, and applications for engineers and system designers 67-80