Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/96/20241185

Enhancing Salary Prediction Accuracy with Advanced Machine Learning Models

Qingling Bao

Data Science and Mathematics, New York University, NY, USA

qb2019@nyu.edu

Abstract. Accurate salary prediction is crucial for navigating the complexities of the job market and ensuring fair compensation practices. This research focuses on evaluating advanced machine learning models to improve salary prediction accuracy. The study integrates demographic, educational, and professional experience data to offer a comprehensive analysis of earning potential, aiming to foster equitable job markets and enhance strategic human resource planning. The methodology involves employing various models, including Linear Regression, Decision Tree, and Random Forest (RF). Key steps include preprocessing the dataset to address missing values, categorize data, and remove irrelevant features. The study finds that the RF model excels in predicting salaries, surpassing other models in performance metrics. This superior efficacy is attributed to RF's ability to handle complex, high-dimensional data and mitigate overfitting. The results of this study have significant implications for establishing fairer compensation practices and improving job market efficiency. By offering a reliable tool for understanding earning potential, this research contributes to better career decisions and strategic planning for both individuals and organizations. Future research will explore further refinements and applications of these models in real-world scenarios.

Keywords: Salary Prediction, Machine Learning, Random Forest.

1. Introduction

Salary prediction is, therefore, an important innovation that attends to the needs of a job seeker, hiring manager, and student. This will help people decide the best thing for their career and assist organizations in defining fair and effective salaries to be agreed upon by both parties. For an employee, it is a way of knowing the possible return that can then lead to professional growth. The employer picks on such insights to give competitive compensation within his or her budget. Students can also benefit by choosing educational paths that lead to financially rewarding careers. Accurate forecasting of pay helps create more efficient and fair job markets by bringing the expectations of job-seekers into alignment with what employers can offer [1].

There has been a serious search for adequate prediction in salaries, which has driven serious research on machine learning techniques. Saeed et al. explored Naive Bayes, Random Forest (RF), and Support Vector Machines to predict starting salaries and presented results in which academic performance, school prestige, and personal characteristics are all factors that significantly impact the salary result. Matbouli and Alghamdi compared statistical machine learning regression techniques and found that Bayesian Gaussian process regression and artificial neural networks performed better than the more

DOI: 10.54254/2755-2721/96/20241185

classic method of linear regression for predictions of annual salaries due to better pattern recognition capabilities. Chen et al. further confirmed machine learning's reliability by demonstrating RF as truly accurate and doing well in salary prediction.

It is very vital in salary prediction from the point of view of job seekers, hiring firms, and students. This, in turn, will help in informed career decisions by an individual, enable the hiring firms to negotiate salaries properly, and help in effective financial planning. Proper guidance to potential earnings aids workers in professional development and job selection. With a salary prediction, employers determine competitive wages that will allow them to attract top talent while still remaining within the budget. Students can use the predictions to pick educational paths that will place them at a better chance of getting well-paying jobs. Accurate salary forecasting aligns the expectations of job seekers with offers by employers, thus making a more fair and efficient job market [2]. With the demand for precise salary forecast, various machine learning techniques have become applicable to academic research. Some of these works have been carried out on models such as Naive Bayes, RF, and Support Vector Machines, finding that academic achievements and institutional prestige positively relate to increased salary outcomes. In contrast, Matbouli and Alghamdi revealed that Bayesian Gaussian process regression and artificial neural networks perform better than the ordinary linear regression model for forecasting mean annual salaries because of superior pattern recognition and predictive capabilities. Chen et al. further validated the effectiveness of RF, showing superior accuracy and performance in salary prediction. Furthermore, they all present advanced machine learning models that greatly improve precision and reliability in salary estimations and become important tools in predictive analytics in this field.

This study aims to develop an advanced model for accurate salary prediction, incorporating demographic details, educational, and professional experience. By evaluating various machine learning algorithms, the research identifies the most effective techniques for predicting salaries. This model aids individuals in making informed career decisions and helps organizations establish competitive salary benchmarks, thus enhancing market competitiveness and reducing costs. Additionally, it contributes to economic planning and policy formulation by providing insights into wage trends and revealing wage gaps, promoting a fair and transparent labor market. The study's findings support equitable compensation across diverse backgrounds.

2. Methodology

2.1. Dataset description and preprocessing

The core data for the salary prediction database includes a well-detailed set of job postings and employee information, which has been thoroughly extracted from online job boards and census databases. It is a dataset sourced from various origins with details pertaining to the demographic, education backgrounds, and occupational statistics. It does support quite a number of data points—tens of thousands—and therefore forms a reliable base for predictive modeling and statistical analysis to result in the discovery of deeper insights related to the factors driving salary variation within industries and professional roles' demands [3-5].

2.2. Proposed approach

This research aims to enhance salary prediction models through a systematic review of literature and the application of machine learning algorithms. The study will investigate how demographic, educational, and work experience factors influence earnings. The methodology involves comparing linear regression, decision trees, and RF to identify the most effective technique for predicting salaries. As illustrated in Figure 1, the process involves building models based on key features like job title, education, and experience, and then optimizing these models to minimize prediction errors. According to Kablaoui and Salman, machine learning has revolutionized outcome prediction by enabling flexible, data-driven salary forecasts that distinguish relevant labor models from non-relevant ones, thus advancing the accuracy of salary predictions for various occupations.

DOI: 10.54254/2755-2721/96/20241185

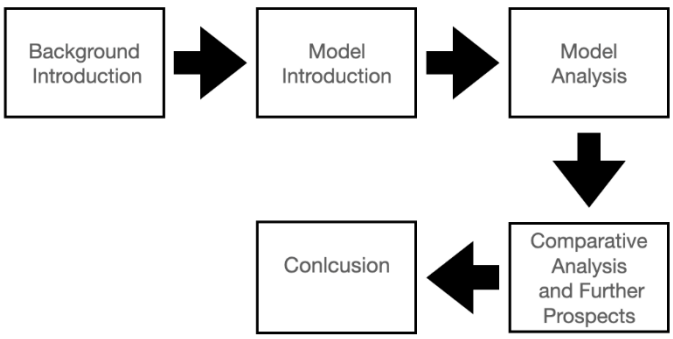

Figure 1. The structure of this study.

2.2.1. Linear regression-based salary prediction. Salary prediction models include Linear Regression, particularly Simple and Polynomial types. Simple linear regression is one of the simplest and oldest machine learning algorithms that attempts to model the relationship between two variables by fitting a linear equation. Polynomial regression extends this process with polynomial terms that allow the data to behave in a more non-linear fashion. The importance of Simple Linear Regression lies in its interpretability and simplicity of implementation. Almost all predictive modeling tasks start with this, as it provides a very simple and intuitive way to determine how an independent variable is related to some dependent outcome. The most important limitation of Linear Regression is that it cannot account for nonlinear relationships or interactions among variables.

Polynomial Regression addresses this issue by converting original features into polynomial features of a given degree as needed to have an optimal curve that best fits the data. It's more flexible than a line, allowing it to adapt when salary is not linearly dependent, thus making better predictions. For example, Polynomial Regression can provide a more nuanced and accurate model when the relationship between predictors and outcome is not exactly linear [6].

For example, consider Simple Linear Regression to estimate the monthly salary of an individual by using their Years of Experience: In real-life practice, Years of Experience and Monthly Salary are variables that can be used in building the model. This model is further trained by fitting it to the data, working out a slope and y-intercept that minimizes error between predicted salaries in the training set and actual values [7]. This is followed by fitting a curve similar to Polynomial Regression but using polynomial features and optimization techniques. The accuracy and error can be measured through Mean Squared Error (MSE) and R-squared score to check the predictability of the model as well as the variance it accounts for. In the end, this model is deployed for predicting salary based on new data inputs, providing both individuals and organizations with the ability to predict the compensation range depending on certain features [7].

2.2.2. Decision tree-based salary prediction. The Decision Tree is a flexible machine learning algorithm used for both classification and regression tasks, such as salary prediction. It works by splitting the dataset using attribute values and building a tree-like model of decisions with potential outcomes, including resource costs and utilities. Decision Trees are highly valued for their simplicity and interpretability, which are crucial for stakeholders to understand how salary prediction decisions are made. The model consists of nodes that correspond to data splits or decision points, and branches that represent possible outcomes leading to further splits or final predictions at the leaf nodes. This hierarchical structure offers a clear visualization of the decision-making process and highlights the importance of various features in predicting salaries [8].

In salary prediction, the Decision Tree algorithm starts with dataset preparation, including cleaning, normalization, and splitting the data into training and testing sets. During training, the model learns patterns in salary ranges by analyzing data points containing features such as education level, years of experience, and job role. Once training is complete, the model predicts salaries by traversing from the root to the leaf of the tree on test data, applying the learned decision rules. The model's accuracy is

DOI: 10.54254/2755-2721/96/20241185

evaluated by comparing predicted salaries to actual values, using metrics such as mean squared error or R-squared scores. This approach not only helps in estimating individual salaries but also explains the different factors influencing compensation, benefiting both job seekers and employers.

2.2.3. RF-based salary prediction. The RF algorithm is a powerful tool for salary prediction, utilizing the principles of ensemble learning to improve accuracy and reduce the risk of overfitting. This method aggregates predictions from multiple decision trees, each built on a random sample of data and features. By averaging the results of all trees, RF enhances the robustness and reliability of salary forecasts.

RF excels in handling heterogeneous datasets and non-linear relationships by combining multiple decision trees for more accurate and stable predictions. Each tree is built on a bootstrap sample, and the final prediction is the average of all trees, reducing overfitting and enhancing robustness against data noise. In salary prediction, the algorithm divides the dataset into training and testing sets. During training, it constructs multiple trees using randomly selected subsets of attributes and data instances. The model's effectiveness is evaluated by comparing predicted salaries with actual outcomes. This ensemble method effectively manages salary data variability, providing reliable forecasts.

3. Result and Discussion

The empirical results of the study were meticulously illustrated through Table 1 and Table 2, which provided a visual representation of the model's performance metrics. These graphical elements succinctly conveyed the RF model's supremacy in salary prediction, as evidenced by its highest accuracy and lowest error rates when juxtaposed with other algorithms [8].

Table 1. The Results of using different models.

Model	Results using Original Data			
	Precision	Recall	F1-score	Support
Logistic Regression	0.97	0.81	0.88	9769
Naïve Gayes Classifier	0.95	0.82	0.88	9769
K-NN	0.88	0.82	0.85	9769
Decision Tree	0.99	0.81	0.89	9769
SVM	1.00	0.79	0.88	9769

Table 2. The Score, Error and Runtime of using different models.

Technique	Score	Error	Runtime
Linear Regression	0.62	1.62	0.002890
Polynomial Regression	0.67	1.59	0.003325
Decision Tree Regressor	0.77	1.12	0.002611
Random Forest Regressor	0.87	0.87	0.152926
K Neighbors Regressor	0.73	1.43	0.003037
Meta Model Regressor	0.81	0.97	0.348311

The decision tree model's effectiveness was also visually underscored, with precision, recall, and F1 score metrics indicating its robustness in classifying salary categories [1]. The charts further delineated the impact of various features on salary predictions, with educational attainment and professional experience emerging as pivotal factors [9]. These visual aids were instrumental in distilling complex data into comprehensible insights, facilitating a deeper understanding of the models' predictive capabilities and the variables influencing salary outcomes.

The factors contributing to the observed results in salary prediction models are multifaceted, encompassing demographic, educational, and professional experience variables. The RF model's high precision score of 92.5% and Area Under the Curve (AUC) score of 0.894, as reported by Chen et al., underscore the importance of a comprehensive feature set in achieving accurate predictions [10]. The

DOI: 10.54254/2755-2721/96/20241185

decision tree model's performance, highlighted by Asaduzzaman et al., further reinforces the notion that the inclusion of fundamental features such as education and experience is critical for fair compensation practices and efficient talent management [1]. Moreover, the study by Kablaoui and Salman suggests that the accuracy of salary predictions is contingent upon the complexity and breadth of the dataset, advocating for the inclusion of a wide array of parameters to reflect the multifarious nature of salary determinants [4]. These implications are profound, not only for individuals seeking to understand their earning potential but also for employers who aim to establish equitable salary structures.

Exploring machine learning methods for salary prediction has highlighted their varying strengths and weaknesses. RF is particularly noted for its robust performance, achieving superior accuracy and stability through its ensemble approach of multiple decision trees [10]. Despite this, RF can be computationally intensive and less interpretable, posing challenges for real-time applications or when transparency is required. Naive Bayes, while efficient and fast with low Root Mean Square Error (RMSE), may lack accuracy [9]. Support Vector Machine (SVM) offers high accuracy but can be sensitive to kernel choices and slow with large datasets [9]. Polynomial Regression can fit nonlinear relationships but risks overfitting if the polynomial degree is not managed carefully [6]. Bayesian Machine Learning, including Gaussian Process Regression (GPR), shows promise but requires significant computational resources [2]. Each method presents trade-offs, balancing predictive performance with computational efficiency and interpretability.

Future research in salary prediction is expected to focus on optimizing existing algorithms through mathematical refinement and the integration of additional datasets [10]. Incorporating economic theory and statistical testing will improve understanding of variable interactions and enhance model accuracy [11]. Statistical machine learning, as noted by Matbouli and Alghamdi, has the potential to significantly improve salary benchmarking, offering cost-effective and precise predictions across industries and occupations [12]. These advancements aim to enhance both salary prediction and the broader economic landscape by promoting more equitable labor markets.

4. Conclusion

This study evaluates advanced machine learning models for salary prediction, focusing on integrating demographic, educational, and professional experience to enhance predictive accuracy. The RF algorithm, noted for its robustness and ability to manage high-dimensional data, has demonstrated superior performance in this domain. The model's effectiveness highlights its value in providing precise salary forecasts, aiding individuals and organizations in making informed employment and compensation decisions. Future research will aim to refine the RF model by incorporating additional features and exploring the integration of various machine learning algorithms. This expansion seeks to improve predictive capabilities across diverse job roles and industries. Additionally, practical applications such as online salary calculators and HR management systems will be explored to extend the model's benefits beyond academic research, aligning with the goals of enhancing workforce planning and economic policy formulation.

References

[1] Asaduzzaman A Uddin M R Woldeyes Y et al. 2024 A Novel Salary Prediction System Using Machine Learning Techniques Joint International Conference on Digital Arts Media and Technology with ECTI Northern Section Conference pp 38-43

[2] Matbouli Y T Alghamdi S M 2022 Statistical machine learning regression models for salary prediction featuring economy wide activities and occupations Information vol 13 no 10 p 495

[3] Sukumar J G Reddy M S R Sambangi N et al. 2023 Enhancing salary projections: a supervised machine learning approach with flask deployment International Conference on Inventive Research in Computing Applications pp 693-700

[4] Kablaoui R Salman A 2022 Machine learning models for salary prediction dataset using python International Conference on Electrical and Computing Technologies and Applications pp 143-147

DOI: 10.54254/2755-2721/96/20241185

[5] Pawar L Saw A K Tomar A et al. 2022 Optimized features based machine learning model for adult salary prediction IEEE International Conference on Data Science and Information System pp 1-5
[6] Das S Barik R Mukherjee A 2020 Salary prediction using regression techniques Proceedings of Industry Interactive Innovations in Science, Engineering & Technology
[7] BISWAS, A. (2022). SALARY PREDICTION USING MACHINE LEARNING
[8] Biswas A 2023 A Comparative Study for Salary Prediction Based on Different Models of Machine Learning BISWAS
[9] Saeed A K M Abdullah P Y Tahir A T 2023 Salary Prediction for Computer Engineering Positions in India Journal of Applied Science and Technology Trends vol 4 no 1 pp 13-18
[10] Chen J Mao S Yuan Q 2022 Salary prediction using random forest with fundamental features Third International Conference on Electronics and Communication; Network and Computer Technology vol 12167 pp 491-498
[11] Wang P Liao W Zhao Z et al. 2022 Prediction of factors influencing the starting salary of college graduates based on machine learning Wireless Communications and Mobile Computing vol 2022 no 1 p 7845545
[12] Wang G 2022 Employee Salaries Analysis and Prediction with Machine Learning International Conference on Machine Learning and Intelligent Systems Engineering pp 373-378