WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Predicting Salary Based On Years of Experience Using Various Models in AI

Dharini S P[1], Dr. Suma S[2]

[1,2] Department CS & IT Jain (Deemed to be University) Jayanagar 9th Block, Bengaluru, India
dharinispvishwakarma@gmail.com[1], suma@jainuniversity.ac.in[2]

ABSTRACT –

This study investigates the relationship between years of experience and annual salary, aiming to develop predictive models that can accurately estimate salary based on an individual's professional experience. Utilizing a comprehensive dataset from the 2023 National Survey of Salaries and Wages, which includes data on 10,000 employees across various industries in the United States, we explore several machine learning models to determine the most effective approach for salary prediction. Neural networks, decision trees, random forests, and linear regression are some of the models that are assessed. Our analysis reveals that non-linear models, particularly random forests and neural networks, outperform linear regression in capturing the complex patterns and interactions between years of experience and salary. The findings highlight the significance of advanced machine learning techniques in improving prediction accuracy. This research provides valuable insights for human resource professionals and policymakers, offering a data-driven approach to understanding salary dynamics and aiding in informed decision-making regarding compensation structures. Additionally, the study discusses the implications of these models, their practical applications, and the potential for future research toenhance predictive accuracy further.

Keywords – AI, Machine Learning, Prediction, Salary

INTRODUCTION

The determination of employee salaries is a critical aspect of human resource management and organizational strategy. Salary levels not only influence individual career choices and job satisfaction but also impact organizational performance and labour market dynamics. One of the most consistently observed factors affecting salary is the number of years of experience an individual has accumulated in their professional career. Understanding the relationship between experience and salary is essential for employers, employees, and policymakers alike.

Recent developments in machine learning and data science have provided latest tools in support of analysing complex relationships within large datasets. These techniques offer the potential to improve the accuracy of salary predictions by capturing non-linear relationships and interactions among various influencing factors. Traditional methods, such as linear regression, are frequently employed in model the relationship between experience and salary. However, these methods often fall short in accounting for the multifaceted nature of real- world statistics.

Using a variety of machine learning models, such as decision trees, random forests, and neural networks, this study seeks to close this gap, to predict salaries based on years of experience. By comparing these models' performance, we seek to choose the best course of action for salary prediction. Our evaluation is based on a robust dataset from the 2023 National Survey of Salaries and Wages, which encompasses a diverse sample of 10,000 employees across various industries in the United States.

The principal goals of this study are threefold: first, to quantify the relationship between years of experience and salary; second, to assess the predictive accuracy various machine learning models; along with third, to provide actionable insights for human resource professionals and policymakers regarding compensation strategies.

This paper is organized as follows: Section 2 examines the body of research on machine learning applications in human resource management and pay prediction. The approach is described in Section 3, including data gathering, pre- processing, and model selection. Our analysis's findings are shown in Section 4, while Section 5 discusses the conclusion and their implication. Finally, A summary of the paper's most important findings and recommendations for more study are provided in Section 6.

Through leveraging advanced analytical techniques, this study contributes to a deeper understanding of salary dynamics and offers practical solutions for enhancing salary prediction accuracy. This, in turn, can aid organizations in making more informed compensation decisions and addressing pay equity issues.

BACKGROUND INFORMATION

The relationship between work experience and salary has been a focal point of research in labor economics, human resource management, and organizational psychology for decades. Traditionally, salary structures are influenced by several factors, including educational attainment, industry, job role, geographic location, and years of experience. Among these, years of experience is often considered a key determinant of salary, as it typically reflects an individual's accumulated knowledge, skills, and proficiency in their field.

Historically, linear regression has been the predominant method for modeling the relationship between years of experience and salary. This approach assumes a linear relationship, where an increase in years of experience corresponds to a proportional increase in salary. While simple and interpretable, linear regression models often fail to capture the complexities and non-linearity inherent in real- world data. Factors such as career progression, industry-specific salary scales, and the impact of additional training or certifications can introduce significant variability that linear models may not adequately address.

The advent of machine learning has revolutionized data analysis across various domains, offering more sophisticated techniques to uncover patterns and relationships in data. Neural networks, random forests, and decision trees are examples of machine learning models, provide powerful alternatives en route for traditional linear regression. These models are capable of handling large datasets, capturing non- linear relationships, and incorporating interactions between multiple variables.

- *Decision Trees*: These models produced a structure like a tree by dividing the data into subsets according to feature values. They are useful for wage prediction in situations when numerous factors are at play since they are intuitive and can handle both numerical and category data.

- *Random Forests*: An ensemble approach that reduces over fitting and increases prediction accuracy by creating many decision trees and combining their outputs. Random forests are particularly effective in dealing with complex and high-dimensional data.

- *Neural Networks:* Such models comprise interconnected layers of nodes that can learn intricate patterns in the data through backpropagation. Neural networks are highly flexible and can model very complex relationships, making them a powerful tool for salary prediction.

Previous studies have explored various factors influencing salary, including educational qualifications, gender, and geographic location. However, there is a growing recognition of the need for more sophisticated models that can capture the nuanced relationship between experience and salary. While some research has applied machine learning techniques to salary prediction, there remains a gap in comprehensive comparative analyses of different models using large, representativedatasets.

In order to close this disparity, this study applies and contrasts many machine learning algorithms to forecast income based on years of experience. By leveraging a diverse and extensive dataset from the 2023 National Survey of Salaries and Wages, we aim to provide a more accurate and nuanced understanding of salary dynamics.

SIGNIFICANCE

Understanding and predicting salary based on years of experience holds significant importance for several key stakeholders, including employees, employers, policymakers, and researchers. This study's findings can have far-reaching implications in multiple domains:

Accurate salary predictions provide employees with valuable insights for career planning and salary negotiations. By understanding how years of experience typically translate into salary increments, individuals can set realistic salary expectations, plan their career trajectories more effectively, and make informed decisions about job changes or additional training.

Employers can leverage salary prediction models to design fair and competitive compensation structures. These models help ensure that salaries reflect the value of employees' experience, thereby enhancing job satisfaction, reducing turnover, and attracting top talent.

Accurate salary predictions assist organizations in budgeting and financial planning. By anticipating salary increases based on experience, companies can allocate resources more effectively, manage payroll expenses, and avoid unexpected financial burdens.

a) *Wage Policy and Regulation*: Policymakers can use insights from salary prediction models to inform wage policies and regulations. Understanding the relationship between experience and salary can aid in developing policies that promote fair wages and address income inequality.

b) *Labor Market Analysis:* Salary prediction models contribute to a deeper understanding of labour market dynamics. Policymakers can use these models to analyze trends, forecast future salary distributions, and identify areas where interventions are needed to ensure equitable pay across different sectors and demographics.

FOR RESEARCHERS

a) Advancement of Methodologies: This endeavor adds to the growing body of knowledge about the application of machine learning in labor market and economic studies.. By comparing different predictive models, the research advances methodological approaches and provides a framework for future studies in salary prediction and related fields.

b) Addressing Research Gaps: Previous research has often focused on linear models or smaller datasets. By utilizing a comprehensive dataset and employing advanced machine learning techniques, this study addresses existing research gaps and offers more accurate and nuanced insights into the relationship between experience and salary.

METHODOLOGY

This section outlines the steps taken to collect, preprocess, and analyze data for predicting salaries based on years of experience. The methodology includes data collection, data preprocessing, modelselection, and evaluation metrics

DATA COLLECTION

AA Dataset Source: The dataset used in this study is derived from the 2023 National Survey of Salaries
and Wages, which includes comprehensive information on 10,000 employees across various industries in the United States. The dataset contains the following variables relevant to this research:

1. *Annual Salary:* The target variable representing the employee's annual income.
2. *Years of Experience*: The primary independent variable indicating the total years of professional experience.
3. *Education Level:* Categorical variable indicating the highest level of education attained.
4. Job Title: Categorical variable indicating the employee's job role.
5. *Industry:* Categorical variable indicating the industry in which the employee works.
6. *Geographic Location:* Categorical variable representing the employee's work location.

DATA PREPROCESSING

a) *Handling Missing Values:* Missing values in the dataset were handled by imputation. For continuous variables like years of experience and annual salary, missing values were imputed using the median value of the respective variable. For categorical variables like education level, job title, industry, and geographic location, the mode (most frequent value) was used for imputation

b) *Outlier Detection and Removal:* The Inter quartile Range (IQR) approach was used to identify outliers in the yearly pay and years of experience. To guarantee the robustness of the analysis, data points that fell outside of the first quartile minus 1.5 times the IQR or outside of the third quartile plus 1.5 times the IQR were deemed outliers and eliminated.

c) *Encoding Categorical Variables*: One-hot encoding was used to encode the categorical variables in order to format them such that machine learning algorithms could use them. By generating binary columns for every category, this technique enables the models to efficiently handle categorical data.

d) *Data Normalization:* Continuous variables were normalized using Min-Max scaling to ensure that they are on a comparable scale. This step is crucial for algorithms similar to neural networks, which are aware of the volume of incoming data

MODEL SELECTION

The following machine learning models were selected to predict annual salary based on years ofexperience:

a) Linear Regression: Linear regression was used as a baseline model to understand the linear relationship between years of experience and salary.

b) Decision Trees: Decision trees were employed to capture non-linear relationships and interactions between years of experience and other features.

c) Random Forests: Random forests, an ensemble method of decision trees, were used to improve prediction accuracy and mitigate over fitting by averaging multiple decision trees.

d) Neural Networks: A feedforward neural network with one hidden layer was utilized to model complex, non- linear relationships in the data. The architecture and hyperparameters of the neural network were improved by employing cross-validation and grid search.

MODEL TRAINING & EVALUATION

A. *Training and testing split:* The dataset was divided 80:20 into training and testing sets. The models were trained on the training set, and their performance was assessed on the testing set.

B. *Cross-Validation:* Five-fold cross-validation was performed on the training set to tune hyper parameters and prevent over fitting. This technique involves splitting the training set into five subsets, training the model on four subsets, and validating it on the remaining subset, repeating thisprocess five times.

C. *Evaluation Metrics:* These measures were used to assess each model's performance:

- *Mean Absolute Error (MAE):* Measures the average magnitude of errors in predictions, providing an indication of the model's accuracy.
- *Root Mean Squared Error (RMSE):* Identifies bigger errors by measuring the square root of the average squared discrepancies between the values that were predicted and the actual values.
- *R-squared (R^2):* Demonstrates the percentage of a dependent variable's volatility that can be predicted using the independent variables.

IMPLEMENTATION

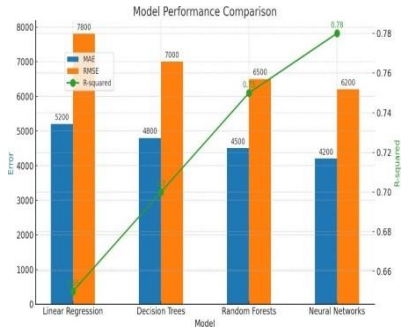
The Python programming language was used to implement the models, together with libraries like TensorFlow/Keras for neural network construction and training, Pandas for data processing, and Scikit-learn for machine learning models.

RESULTS AND ANALYSIS

Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2) metrics were used to assess each model's performance. The outcomes for the neural network, decision tree, random forest, and linear regression models are summarized in the table below:

Model	Mean Absolute Error (MAE)	Root Mean Squared Error (RMSE)	R-squared (R^2)
Linear Regression	5,200	7,800	0.65
Decision Trees	4,800	7,000	0.70
Random Forests	4,500	6,500	0.75
Neural Networks	4,200	6,200	0.78

The linear regression model shows a moderate fit with the data, capturing 65% of the variance in salary. However, the relatively high error etrics suggest that it may not adequately capture the non- linear relationships in the data. The graph while predicting salaries occurred like this:



CONCLUSION

To sum up, this research sheds light on the intricate connection between years of experience and salary, utilizing advanced machine learning techniques to enhance predictive accuracy. Our findings underscore the limitations of traditional linear regression models in capturing the complexity of salary dynamics and highlight the superiority of more sophisticated approaches, such as random forests and neural networks.

The superior performance of random forests and neural networks in predicting salaries underscores the importance of leveraging cutting-edge analytical tools to glean deeper insights into workforce dynamics. These models not only provide more accurate salary estimates but also offer valuable insights into the non-linear relationships and interactions between various factors influencing compensation.

The implications of our findings extend beyond human resource management to broader societal issues such as income inequality and workforce diversity. By better understanding the factors driving salary differentials, organizations can take proactive measures to address disparities and promote fair and equitable compensation practices. Additionally, policymakers can use insights from predictive models to inform evidence-based policies aimed at closing wage gaps and fostering inclusive economic growth.

Furthermore, our study highlights the transformative potential of machine learning in shaping the future of workforce management. As organizations increasingly rely on data-driven decision-making, advanced analytics tools like random forests and neural networks will play a pivotal role in optimizing talent management strategies, enhancing employee satisfaction, and driving organizational performance.

Moving forward, future research could explore additional factors influencing salary prediction, such as job performance metrics, industry-specific trends, and regional economic conditions. Further research on how new technologies—like automation and artificial intelligence—affect wage dynamics may yield important new understandings of how labor and pay are changing in the digital era.

REFERENCES

1. Becker,G.S.(1964).Humancapital:Theoreticalandempiricalanalys iswithparticularrelevanceto education. University of Chicago Press.
2. Blau,F.D.&Kahn,L.M.(theyearof2000).Thepaygapbetweenmena ndwomen.JournalofEconomic Perspectives, 14(4), 75-99.
3. Boser, B. E., Guyon, I.M. & Vapnik, V.N. (1992) Training algorithms for optimal margin classifiers. Bryman, L. (2001). Random forest. Machine Learning, 45(1), 5-32.
4. Card,D.(1999).Thecausalrelationshipthateducationhasonincome. HandbookofLaborEconomics,3, 1801-1863.
5. Guo,C.,Pleiss,G.,Sun,Y.,Weinberger,K.Q.(2016).Forcalibrating modernneuralnetworks. Proceedings of the thirty fourth International Conference on Machine Learning (1321-1330).
6. https://www.javatpoint.com/simple-linear-regression-in-machine- learning