Salary Prediction for Employees Using Regression

Meher Afrroz Binu
Department of Computer Science and Engineering,
Bangladesh University of Business and Technology, Dhaka, Bangladesh.
Email: 2122413077@cse.bubt.edu.bd

***Abstract*—This study underscores the growing importance of salary prediction for both personal career planning and organizational budgeting. Accurate salary forecasting is essential for individuals to make informed career decisions and for organizations to effectively plan compensation structures. Salary prediction was performed using multiple regression techniques to analyze and model the key factors influencing salary levels. Four regression models were applied to the analysis: Linear Regression, Support Vector Regression (SVR), Random Forest, and XGBoost. Among these models, the results revealed that XGBoost achieved the highest accuracy of 97.96%, demonstrating its ability to capture complex, non-linear relationships in the data, followed closely by Random Forest with an accuracy of 97.64%. These two models outperformed the others, effectively handling high-dimensional data and complex feature interactions. The application of multiple regression techniques has proven instrumental in assessing feature importance and understanding the influence of various factors on salary predictions. The study's findings highlight the potential of advanced machine learning models to provide accurate salary predictions, offering valuable insights that could enhance workforce management, improve compensation strategies, and support informed decision-making in organizations.**

***Keywords*—Salary Prediction, Linear regression, SVR, XGBoost, Random Forest, advanced machine learning models.**

I. Introduction

Salary prediction plays a crucial role in modern career planning and organizational budgeting. It is the process of estimating the potential income of an individual based on various factors, such as experience, education, location, and job industry. As businesses and individuals strive for better financial management and career development, accurate salary predictions have gained prominence in helping both employers and employees make data-driven decisions. With the increasing availability of data and advancements in machine learning techniques, predictive modeling has emerged as a powerful tool to estimate salaries based on features such as experience, education, location, and industry. However, the complexity of these relationships necessitates the use of robust regression models to capture nonlinear patterns and interactions effectively.As the demand for accurate and data-driven insights grows, salary prediction systems continue to evolve, providing essential support in the ever-competitive job market.

A. Terminology

The following table presents definitions of key terms and concepts used in this study. It aims to provide clarity and help readers better understand the regression models and techniques applied in salary prediction, as well as the related terminology discussed throughout the paper.

Term	**Definition**
XGBoost	Extreme Gradient Boosting.
SVR	Support Vector Regression.
RF	Random Forest.
SVM	support vector machine.
R²	Coefficient of Determination
MSE	Mean Squared Error.
RMSE	Root Mean Squared Error.
MAE	Mean Absolute Error.
ANN	Artificial neural network.
MLR	Multiple Linear Regression.
TBR	Tree Based Regression.
DNN	Deep Neural Network.
MLP	Multi-Layer Perceptron
CNN	Convolutional Neural Network
KNN	K-Nearest Neighbors
GPR	Gaussian Process Regression

TABLE I: Key Terminology and Definitions Used in the Study

B. Existing Application

Salary prediction models are extensively used by leading companies and platforms across industries to provide tailored solutions for recruitment, compensation planning, and financial decision-making. Here's a detailed look at how some prominent platforms and companies use these models:

In [1], today's competitive job market, understanding salary trends is crucial for both job seekers and employers. Platforms like Glassdoor leverage user-submitted data on salaries and job roles, combined with machine learning algorithms, to provide average and range-based salary predictions. These insights not only offer a clear understanding of market expectations but also promote transparency, helping build trust and improve negotiation strategies.

Amazon [2], uses predictive models for salary forecasting as part of its recruitment and HR processes. By analyzing market trends and workforce data, Amazon ensures competitive pay packages to attract and retain talent in a highly dynamic industry.

Workday [3], is a leading HR software provider that integrates predictive analytics to forecast salary trends and determine optimal compensation levels for employees. Organizations use Workday's tools to analyze market conditions and ensure equitable pay structures, improving employee satisfaction and retention.

Upwork [4], utilizes salary prediction models to suggest pricing for freelancers and clients. By analyzing job categories, skill sets, and market demand, Upwork provides freelancers with competitive rate recommendations, helping them set fair prices for their services while ensuring affordability for clients.

The adoption of salary prediction models across platforms and organizations highlights their practical utility

C. Objective

This study uses four regression models: Linear Regression, Support Vector Regression (SVR), Random Forest, and XGBoost, each chosen for its strengths in handling diverse data and relationships. Linear Regression was selected for its simplicity and interpretability, serving as a baseline model. SVR was included for its ability to capture nonlinear patterns, making it effective for more complex data. Random Forest, an ensemble model, was chosen for its robustness in handling high-dimensional data and feature interactions while preventing overfitting. XGBoost was selected for its superior performance in managing complex data, nonlinearity, and missing values.

By comparing these models, the study aims to provide insights into their varying complexities, interpretabilities, and predictive abilities. The goal is to help practitioners select the most appropriate model for salary prediction, improving decision-making in human resources, job market analytics, and financial planning.

D. Contribution

This paper presents the significant contributions both technically and in real-world applications. By leveraging advanced regression techniques, this model addresses the challenges of accurate salary forecasting while ensuring practical utility in various domains. Below, we outline the Some significant contributions of the model:

- **HR Efficiency:** Assists human resource departments in streamlining compensation packages, reducing manual effort, and ensuring equitable pay structures.
- **Support for Job Seekers:** Helps job seekers evaluate fair salary expectations based on skills, location, and industry trends, empowering them during job negotiations.
- **Educational Guidance:** Helps students and professionals choose career paths by providing insights into salary trends based on qualifications and skills.
- **Cost Optimization for Businesses:** Enables businesses to optimize costs by accurately benchmarking salaries, avoiding overpayment or underpayment.
- **Regional and Industry Insights:** Offers salary insights specific to industries and locations, enabling local governments and policymakers to address wage disparities.
- **Freelancer and Gig Economy Utility:** Assists freelancers and gig workers in setting competitive rates for their services based on market standards.
- **Startup Resource Planning:** Helps startups allocate resources efficiently by predicting salary benchmarks for key roles, aiding in sustainable growth.
- **Economic Insights for Policymakers:** Supports government agencies in understanding wage trends to develop better labor policies and improve overall economic planning.
- **Global Workforce Management:** For multinational organizations, salary prediction models help to standardize compensation across different regions. They ensure that employees in various countries receive competitive pay relative to their location and industry standards, simplifying global HR management.

the model offers valuable insights for businesses, employees, and policymakers, enabling informed, data-driven decisions that enhance efficiency, fairness, and economic development across various sectors.

II. Related Research

Over the years, numerous research studies and projects have focused on developing and refining regression-based prediction systems across various domains. These studies have significantly contributed to advancing predictive modeling techniques, with applications in fields such as finance, healthcare, marketing, and human resources. The continuous improvement of regression models, along with the integration of new algorithms and data processing methods, has enhanced prediction accuracy and efficiency, paving the way for more sophisticated models capable of delivering precise forecasts in diverse real-world applications.

In [5], K. Rathan et al. proposed a method for predicting cryptocurrency prices using decision trees and linear regression models. Data was sourced from Quandl.com ('BITSTAMPUSD'). key methods used for prediction and forecasting are decision trees, linear regression, and machine learning. The main contribution is comparing decision trees and regression models for predicting Bitcoin prices, with experimental results showing that linear regression outperforms decision trees. However, the author fails to provide more transparency on the dataset regarding weather data, extraction methods, validation techniques, and missing values, which affects its real-time applicability, despite the methodology including feature selection. This method has practical implications for Bitcoin price prediction, informing investment decisions and providing a framework for future research on cryptocurrency forecasting.

In [6], Y. Kim et al. proposed a study on tensile strength prediction of BFRP and GFRP using Multiple Regression Analysis. This method used Multiple Regression Analysis, Polynomial Regression Analysis, and Artificial Neural Networks for prediction. The main contribution is demonstrating that ANN models outperform MRA in predictive performance. This paper analyzes critical factors affecting GFRP and BFRP strengths and compares the accuracy of MRA, PRA, and ANN using

MAE, RMSE, and MAPE values. However, some limitations include the scalability of MRA and PRA experimental results and the need for diverse environmental conditions in durability models. This paper shows that ANN offers superior prediction accuracy in real-life contexts, with temperature and exposure time significantly affecting tensile strength. For GFRP, diameter plays a key role, but not for BFRP. Models should account for diverse environmental conditions for improved accuracy. .

In [7], Y. T. Matbouli et al. proposed Statistical Machine Learning Regression Models for Salary Prediction using Artificial Neural Networks. This study explores Bayesian Gaussian process regression (GPR), artificial neural networks (ANN), tree-based regression (TBR), support vector regression (SVR), and multiple linear regression (MLR). Key contributions include the applicability to all job titles and economic activities, the use of machine learning for improved salary estimation, and adaptability to various international job markets. However, limitations include the impact of limited survey data on accuracy, potential inaccuracies of MLR models, and reduced performance with smaller datasets. Real-life implications include assisting institutions in assessing alumni outcomes, promoting salary transparency, and emphasizing the importance of data cleansing and transformation for accuracy.

In [8], P. Viroonluecha et al. developed a Salary Prediction System for Thailand's labor workforce using Deep Neural Networks (DNNs) for regression. The dataset includes over 1.7 million users from a job search website. This study uses Deep Neural Networks for regression, compares Random Forest and gradient-boosted trees, and applies eleven feature selection algorithms including ACO, PSO, and HSA. The main contribution is the emphasis on personal factors influencing compensation, with multiple algorithms integrated to offer unique advantages. However, some limitations include the data being limited to job seekers in Thailand with bachelor's degrees or higher, and some outdated information affecting accuracy. The real-world implications of this method are: that automatic feature extraction reduces irrelevant parameters, and feature selection improves both accuracy and runtime. The system also provides insights into salary prediction factors.

In [9], B. Sravani et al. proposed a study on predicting student performance using supervised learning to identify at-risk students. This paper works under the following rules: linear regression, data pre-processing, training set selection, using multiple variables in regression, and comparing with SVM and decision trees. The key contribution is focusing on supervised learning to identify students at risk of dropping out and comparing various machine learning methods. However, Limitations include a small sample size of 100 students and variability in predictions based on student attributes. The author's advice proposed can be incorporated into education, enhancing performance prediction, improving learning outcomes, and guiding course design through effective data preprocessing and personalized analytics.

In [10], K. Brubakk et al. studied how hospital work environments affect the patient safety climate using logistic and linear regression. This study uses the Work Environment Survey (WES) and Safety Attitude Questionnaire (SAQ) as datasets. This study uses a longitudinal design with surveys in Norwegian clinical units, applying regression models to assess work environment and safety climate factors. This contributes to the importance of a strong safety culture in promoting safe behaviors and improving patient safety. The author fails to address potential gaps in the dataset and the limited applicability of findings beyond the studied units. Real-life implications include the impact of the work environment on patient safety, the need to prioritize staff safety, hospital management support, staff buy-in for safety culture, and the link between a positive safety climate and effective patient care.

In [11], J. Artin et al. proposed a novel method for traffic prediction using Ensemble Learning. The methods here include ensemble learning for traffic prediction, NAS for model optimization, linear regression for accuracy, and a combination of deep learning with regression models. The main contribution is introducing a traffic prediction method that improves accuracy by factoring in climate conditions through NAS and linear regression. However, prediction accuracy is limited by factors such as traffic infrastructure, road capacity, regulations, weather, and accidents. Future improvements may incorporate convolutional techniques and enhanced data collection. The real-life implication is this method enhances urban traffic congestion predictions, with the NAS algorithm optimizing performance method enhancing urban traffic predictions

In [12], F. D. G. Gámez et al. conducted a study using regression to identify variables predicting teachers' attitudes toward ICT in higher education, focusing on teaching and research. This Ex post facto methodology used online surveys for data collection, followed by descriptive and inferential analyses, with predictions from multiple linear regression. The research contributes to the ACB model to analyze teachers' attitudes and identify predictive variables influencing these attitudes, emphasizing their multidimensional nature. However the author fails to provide qualitative research for deeper insights, and the subjective nature of attitudes may influence data interpretation. The real-life implications highlight the importance of understanding attitudes toward ICT in higher education, with the ACB model offering a valuable tool for analysis.

In [13], B. Sravani et al. proposed a study on predicting student performance using supervised learning to identify at-risk students. This paper works under the following rules: linear regression, data pre-processing, training set selection, using multiple variables in regression, and comparing with SVM and decision trees. The key contribution is focusing on supervised learning to identify students at risk of dropping out and comparing various machine learning methods. However, Limitations include a small sample size of 100 students and variability in predictions based on student attributes. The author's advice proposed can be incorporated into education, enhancing performance prediction, improving learning outcomes, and guiding course design through effective data preprocessing and personalized analytics.

In [14], X. Huang et al. synthesized the method of Groundwater Recharge Prediction using Linear Regression. This paper works under the following rules: linear regression for predictions, MLP for analysis, and LSTM for forecasting. The primary contribution is the comparison of linear regression, MLP, and LSTM models to predict groundwater recharge, identifying key factors for management decisions. The author fails to address non-spatial data, short prediction timeframes, and no uncertainty quantification. The practical implication lies in the enhanced method that improves prediction accuracy, with the LSTM model outperforming others, aiding water resource management and climate adaptation strategies.

In [15], T. Le proposed Improving Electric Energy Consumption Prediction Using CNN and Bi-LSTM. The IHEPC dataset was utilized for validation. This model works by first using CNNs to extract features from the IHEPC dataset. Then, Bi-LSTM layers capture time-series trends, and fully connected layers predict future electric energy consumption. This study contributes to the EECP-CBL model, combining CNN and Bi-LSTM, to improve energy prediction using the IHEPC dataset, with potential enhancements from evolutionary algorithms and more data. Limitations are the need for improved prediction model performance and additional datasets for verification. Practical implications of the EECP-CBL model improve energy prediction accuracy, outperform current models, aid power management, support energy policy, and can be enhanced with evolutionary algorithms and more datasets. .

In [16], D. Tanouz et al. proposed credit card fraud detection using machine learning techniques. This paper works in the following rules: logistic regression or classification implements the random forest algorithm for fraud detection and applies Naive Bayes and decision tree classifiers for predictions. Key contributions of this research article include the development of machine learning algorithms aimed at identifying fraudulent transactions in credit card data. However, there are some boundaries to be considered: The heavy imbalance in the dataset affects prediction accuracy, leading to high rates of false positives and negatives. Finally, the real-life implications of the proposed method are: The study develops machine learning algorithms for credit card fraud detection, which improve classification accuracy and reduce false positives and negatives.

In [17], S. Hills et al. developed factors linked to non-adherence to social distancing in North London using logistic regression. The key contributions of this research article are as follows: psychological and political influences on non-adherence, offering guidance for public health messaging. However, there are some boundaries to consider: The study's design restricts causal interpretation, and a convenience sample limits generalizability, with an overrepresentation of females and an underrepresentation of BAME participants. Finally, the real-life implications of the proposed method are: The findings support targeted public health policies, address psychological barriers, and emphasize clear communication to improve adherence.

In [18], N. Abdelaziz et al. developed the International Roughness Index prediction model for flexible pavements. The method follows the rules by using MLR and ANNs for modeling, extracting data from the LTPP database, conducting hypothesis tests, and developing regression models for initial IRI estimation. The key contribution of the project is the development of IRI prediction models using MLR and ANNs, utilizing the Long-Term Pavement Performance Database, identifying key roughness parameters, comparing model performance, and addressing missing initial IRI values. The author fails to address the lack of initial IRI measurements in GPS sections and the issue of IRI and pavement distress being measured at different times. Practical implications include developing accurate IRI prediction models, improving estimates with MLR and ANNs, identifying effective maintenance, understanding pavement age-IRI relationships, and utilizing the LTPP database.

In [19],J. W. BAEK et al. developed a model for Predicting Depression Risk Using Multiple Regression. This study uses data from the Korea National Health and Nutrition Examination Survey. This method uses regression analysis, context-DNN for prediction, data pre-processing from health surveys, variable selection, multiple regression for depression risk, and evaluates performance with accuracy and recall metrics. The key contribution is the article-DNN model for depression risk prediction, using multiple regression, context integration, and performance evaluation. However, some limitations should be noted: explanatory power of independent variables, social stigma affecting accurate mental health assessment, challenges in selecting relevant context variables, handling incomplete data, limited prediction accuracy, and the complexity of data preprocessing. The real-life implication is improving mental health management by accurately predicting depression risk and guiding targeted interventions.

In [20], C. M. Liyewet et al. developed machine learning techniques for rainfall prediction. Raw data was collected from the Bahir Dar City meteorological station. This paper works in the following manner: machine learning techniques for rainfall prediction involved MLR, RF, and XGBoost, with relevant variables selected by Pearson correlation and performance measured by RMSE and MAE. The key contributions include improving rainfall prediction for agriculture, water management, and flood risk reduction, with machine learning techniques enhancing accuracy through relevant environmental features. However, the study has limitations: it did not incorporate sensor data and was restricted to specific environmental features for rainfall prediction, failing to measure correlations of all atmospheric factors. The real-life implications of this research are significant, Accurate rainfall predictions benefit agriculture and water management, with future improvements possible through sensor data integration.

In [21], R. Bhardwaj proposed a Predictive Model for the Evolution of COVID-19. This model uses the following methods: fitting infection growth rates with exponential decay, validated by data from various countries, and optimized using least-squares fitting. The key contributions include the model using logistic regression with an exponential decay to predict COVID-19 trends, validated with data from China and

South Korea, and identifying peak infections, total cases, and weak weather correlation. However, some limitations include assuming a constant population, inability to predict fatalities or recoveries, neglecting recovery effects on growth rate, and not accounting for mitigation measures. The real-life implications of this research are aids in predicting COVID-19 infection peaks, informing government policies, enhancing over time with more data, highlighting lockdown impacts, and supporting healthcare planning.

In [22], M. Banal et al. proposed a study on a comparative analysis of five machine learning algorithms using the UMIST, ORL, and Yale datasets. The paper works in the following rules —KNN classifies based on nearest samples, GA optimizes solutions through natural selection, SVM identifies optimal class separation, DT splits features for decisions, and LSTM manages sequential data and long-term dependencies. The key contributions of this research article are as follows: This study highlights the novel applications of these algorithms and compares their performance, discussing their origins and methodologies while emphasizing the future potential of machine learning and AI. However, it also addresses several limitations, including the risk of premature convergence in genetic algorithms, overfitting in decision trees, and high computational costs in KNN. Finally, this research improves decision-making and predictive analytics in healthcare, finance, and transportation, advancing AI applications in these sectors.

In [23], F. L. Huang et al. analyzed the study of binary outcomes using the logistic regression model. This paper includes logistic regression, linear probability, and modified Poisson models, with Monte Carlo simulations to assess their effectiveness. The key contributions of this research article are as follows: this paper explores alternatives to logistic regression, using Monte Carlo simulations to assess bias and power. However, Some limitations are: this study is limited to experimental conditions and excludes continuous predictors or nested models. Finally, the Real-World Implications are that linear and modified Poisson models are easier to interpret and suitable for experiments.

In [24], A. Mehbodniya et al. proposed a study on credit card fraud detection using machine learning, utilizing an imbalanced European credit card dataset. This paper operates under the following principles: Logistic Regression, Naive Bayes, Decision Tree, KNN, SVM, Random Forest, LSTM, and Multilayer Perceptron. This research article makes the following key contributions: Compares CNN with other algorithms and enhances fraud prediction in imbalanced data. However, a few constraints should be considered: Bayesian methods need better anomaly detection; neural networks are computationally intensive; decision trees lack real-time analysis. Finally, the real-life implications of this study include: findings improve fraud detection in healthcare by favoring robust models like Random Forest.

III. Proposed Methodology

The methodology for this research concentrated on designing a robust and precise machine learning model to predict salaries. The process began with comprehensive data collection, gathering a detailed salary dataset. This dataset encompassed a variety of influential characteristics such as job title, years of experience, education level, gender, and age, which play crucial roles in determining salary levels. Ensuring the dataset's integrity was a top priority, as any inconsistencies or missing values could adversely affect the model's performance. To address this, data preprocessing techniques were applied, including imputation—where missing values were replaced using statistical measures such as the mean, median, or predictive models—or removal, where rows containing incomplete data were excluded. These steps ensured a high-quality, consistent dataset suitable for accurate analysis.

Following preprocessing, the data was split into two subsets: a training dataset and a testing dataset. The training dataset was used to train the regression models, while the testing dataset was reserved for validation purposes. This approach was critical to ensure that the model's performance could be evaluated on unseen data. Four regression models were selected for training: Linear Regression, Random Forest, XGBoost, and Support Vector Regressor (SVR). These models were chosen for their unique strengths in handling data of varying complexity. Linear Regression served as a simple yet effective baseline for performance comparison. Random Forest and XGBoost, both ensemble models, were employed for their ability to handle high-dimensional data and capture complex feature interactions. SVR was included for its capability to model non-linear relationships effectively, offering a diverse range of predictive techniques.

Once the models were trained, they were subjected to rigorous evaluation using multiple performance metrics, including MSE, RMSE, MAE and R^2 (Coefficient of Determination). These metrics provided a comprehensive understanding of each model's performance, assessing how well they captured the underlying data patterns and their ability to deliver precise salary predictions. The model with the best performance across these metrics was identified as the optimal regressor and was subsequently utilized to predict salaries on new or unseen data. This model not only demonstrated superior accuracy but also provided actionable insights into the key factors influencing salaries. The final model offered practical applications for various stakeholders. Businesses and HR departments could leverage it to refine compensation strategies, align salary structures with industry standards, and make data-driven decisions. For individuals, the model provided valuable insights to guide career planning and salary negotiations.

To conclude, the methodology underscored the critical importance of meticulous data preparation, thoughtful model selection, and comprehensive evaluation. This systematic approach ensured the development of a reliable and accurate salary prediction model with real-world utility, supporting informed decision-making in areas such as human resources, financial planning, and labor market analysis. By employing advanced machine learning techniques, the study aimed to enhance the salary forecasting process and contribute meaningfully to both organizational and personal decision-making contexts.

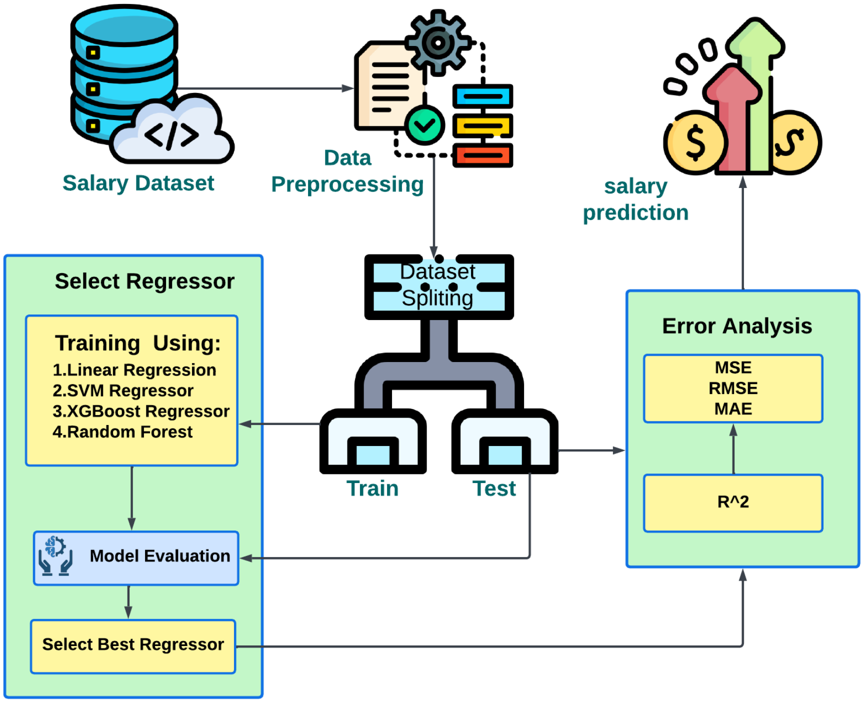

Fig. 1: salary prediction methodology diagram

IV. Result Analysis

This analysis reveals noticeable differences in the performance of the four models used for salary prediction, with XGBoost emerging as the leader. Achieving an impressive accuracy of 97.96% and demonstrates a strong ability to make reliable predictions. This high accuracy suggests that the model consistently predicts salary values with great precision, making it a valuable tool for salary forecasting. In addition to its accuracy, XGBoost also shows the lowest values for key performance indicators such as MSE and RMSE, and MAE, all of which are essential for measuring prediction accuracy. The low MSE and RMSE values suggest that XGBoost's predictions are very close to the actual salary values, minimizing prediction errors. This makes it the most effective and reliable model for salary prediction in this case. Overall, XGBoost stands out as the optimal model for salary prediction, showing robustness and precision in forecasting salaries. This superior performance makes XGBoost highly effective in providing actionable insights for decision-making in salary forecasting and human resource planning. Random Forest Regressor and Support Vector Regression (SVR) follow closely behind, both achieving a high accuracy of 97.64%. These models show strong performance, though slightly behind XGBoost in terms of accuracy. However, their error metrics—MSE, RMSE, and MAE—are still relatively low, indicating that these models also make precise predictions. While they don't perform quite as well as XGBoost in terms of accuracy, they are still reliable choices for salary prediction tasks. The similarity in performance between Random Forest and SVR suggests that both are effective at capturing the underlying relationships in the data and providing valuable insights into salary predictions.

In contrast, Linear Regression performs significantly worse than the other models. With an accuracy of only 69.09%. The model also shows the highest MSE, RMSE, and MAE values, reflecting that its predictions are much further from the actual salary values. This indicates that Linear Regression struggles to capture the complex patterns in salary data, which are likely non-linear and influenced by various factors. Linear Regression's poor performance in this task underscores the limitations of using linear models for complex prediction tasks such as salary forecasting.

Overall, XGBoost is clearly the best performing model, offering the highest accuracy and the lowest error metrics, making it the most suitable choice for salary prediction. Random Forest and SVR are also strong contenders, providing similar levels of performance, though slightly less accurate than XGBoost. On the other hand, Linear Regression demonstrates substantial limitations, making it less effective for this type of prediction. Thus, while simpler models like Linear Regression may be easier to implement, more sophisticated models like XGBoost, Random Forest, and SVR provide superior performance and more accurate predictions for complex tasks such as salary forecasting.Ultimately, the choice of model hinges on the balance between accuracy, complexity, and computational efficiency, with advanced models often providing better results for complex predictions.

Model	**Accuracy (%)**	**MSE**	**RMSE**	**MAE**
Linear Regression	69.09	878,872,341.67	29,645.78	23,538.24
XGBoost	97.96	57,887,409.68	7,608.38	3,572.10
Random Forest Regressor	97.64	67,024,886.39	8,186.87	3,167.45
Support Vector Regression	97.64	67,024,886.39	8,186.87	3,167.45

TABLE II: Performance Comparison of Different Models

The Fig.2 shows the comparison plot between actual and predicted values for XGBoost, the model that performed the best in terms of prediction accuracy.he x-axis represents the actual values of the target variable, while the y-axis displays the predicted values.Each point in the scatter plot represents a data point, and the dashed red line indicates the ideal scenario where the predicted values perfectly match the actual values. XGBoost's predicted values align very closely with the actual values, demonstrating its ability to make accurate predictions. The strong correlation between the actual and predicted values emphasizes the model's effectiveness in salary prediction.

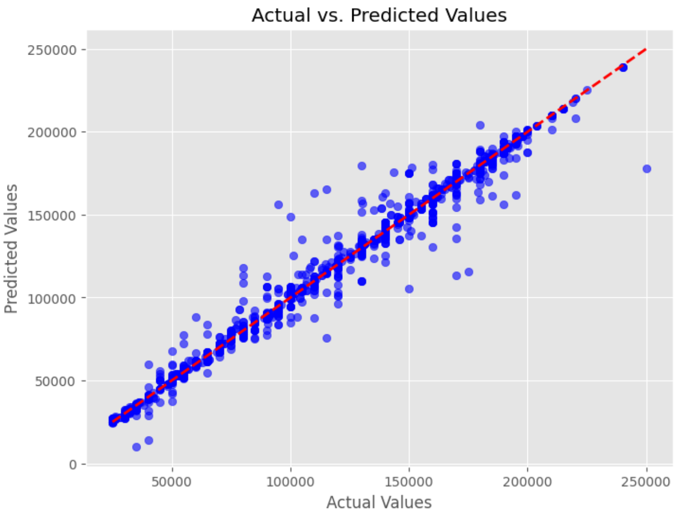

Fig. 2: Actual Vs Predicted graph

The Fig.3 shows a bar chart comparing the accuracy of four salary prediction models: Linear Regression, XGBoost, RandomForest, and Support Vector Regression (SVR). XGBoost stands out with the highest accuracy, closely followed by RandomForest and SVR, both achieving strong results around 97%. Linear Regression, however, lags behind in terms of accuracy, indicating its limitations in handling complex datasets. While RandomForest and SVR show good performance, the chart confirms that while all models performed well, XGBoost emerged as the top model for salary prediction. It provided the best combination of high accuracy and minimal error, making it the most reliable model in this analysis.

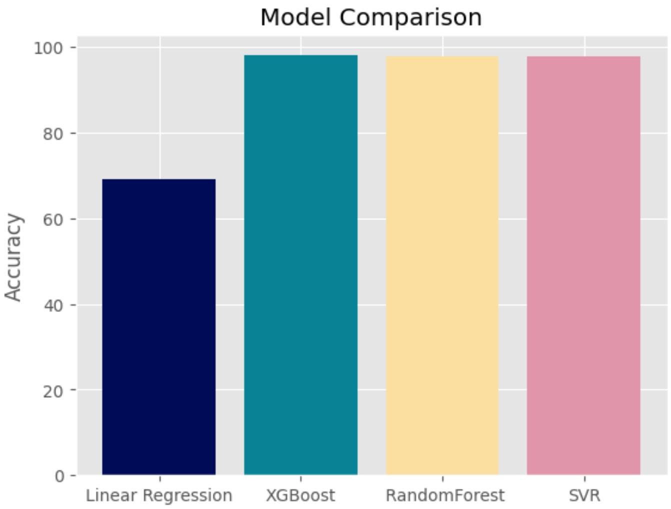

Fig. 3: Model Accuracy Comparison

The Fig.4 shows line plot compares the actual values with the predicted values from four machine learning models: Linear Regression, XGBoost, Random Forest, and Support Vector Regression (SVR). The x-axis represents the index, which could refer to data points or time steps, while the y-axis displays the actual values of the target variable. The plot provides a visual comparison of homodel's predictions align with the actual values. This comparison helps illustrate the relative performance of the models in predicting the target variable.

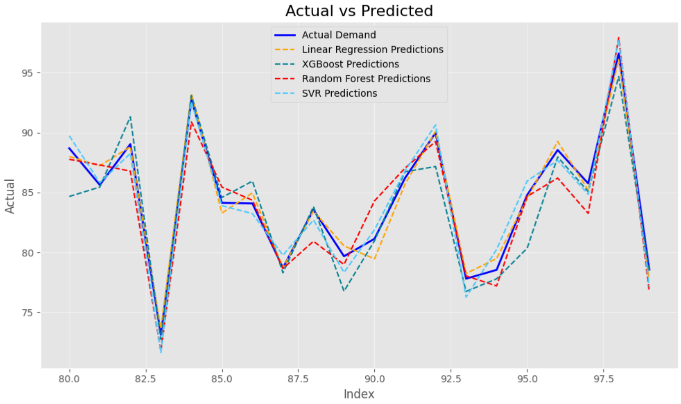

Fig. 4: Model Accuracy Comparison

V. Conclusion

We have proposed an integrated solution for salary prediction using advanced machine learning algorithms such as Linear Regression, XGBoost, Support Vector Regression (SVR), and Random Forest. By utilizing real-world datasets and sophisticated modeling techniques, this framework enables accurate and reliable salary predictions for individuals across various sectors. The models were evaluated based on metrics such as accuracy, Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), with XGBoost emerging as the most accurate and efficient model, achieving an accuracy of 97.96%. The application of these algorithms provides valuable insights for organizations to optimize compensation strategies, enhance workforce planning, and ensure fair remuneration practices.

The integration of advanced machine learning techniques, such as ensemble learning with Random Forest and boosting techniques with XGBoost, ensures high performance and robustness in salary prediction, even when dealing with complex and nonlinear data patterns. This solution not only supports accurate salary forecasting but also aids in identifying potential discrepancies, assisting HR departments in making data-driven decisions. Furthermore, the application of models like SVR provides a flexible approach to salary prediction, proving effective in capturing intricate data relationships.

Moving forward, the scalability and adaptability of this framework can be explored by incorporating additional variables, such as experience, education, and market trends, to further enhance the predictive power of the models. Future research could also address potential challenges such as model interpretability, feature selection, and dealing with imbalanced datasets to improve accuracy. By continuously advancing these methodologies, we can build a dynamic, intelligent system that benefits businesses and employees alike, ensuring fair and equitable salary structures in the future.

Acknowledgment

I would like to acknowledge the support of the Bangladesh University of Business & Technology and the Data Mining lab for their suggestion and resource sharing.

References

[1] www.glassdoor.com, [Accessed 06-12-2024].

[2] "Salary.com – Unlock the Power of Pay — salary.com," www.salary.com, [Accessed 06-12-2024].

[3] "Workday Platform | HR, Finance, Planning, Spend — workday.com," https://www.workday.com/, [Accessed 06-12-2024].

[4] https://www.upwork.com/, [Accessed 06-12-2024].

[5] K. Rathan, S. V. Sai, and T. S. Manikanta, "Crypto-currency price prediction using decision tree and regression techniques," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*. IEEE, 2021, pp. 190–194.

[6] Y. Kim and H. Oh, "Comparison between multiple regression analysis, polynomial regression analysis, and an artificial neural network for tensile strength prediction of bfrp and gfrp," *Materials*, vol. 14, no. 17, p. 4861, 2021.

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

[8] P. Viroonluecha and T. Kaewkiriya, "Salary predictor system for thailand labor workforce using deep learning," in *2018 18th International Symposium on Communications and Information Technologies (ISCIT)*. IEEE, 2021, pp. 473–478.

[9] B. Sravani and M. M. Bala, "Prediction of student performance using linear regression," in *2020 International Conference for Emerging Technology (INCET)*. IEEE, 2020, pp. 1–5.

[10] K. Brubakk, M. V. Svendsen, E. T. Deilkås, D. Hofoss, P. Barach, and O. Tjomsland, "Hospital work environments affect the patient safety climate: A longitudinal follow-up using a logistic regression analysis model," *PloS one*, vol. 16, no. 10, p. e0258471, 2021.

[11] J. Artin, A. Valizadeh, M. Ahmadi, S. A. Kumar, and A. Sharifi, "Presentation of a novel method for prediction of traffic with climate condition based on ensemble learning of neural architecture search (nas) and linear regression," *Complexity*, vol. 2021, no. 1, p. 8500572, 2021.

[12] F. D. Guillén-Gámez and M. J. Mayorga-Fernández, "Identification of variables that predict teachers' attitudes toward ict in higher education for teaching and research: A study with regression," *Sustainability*, vol. 12, no. 4, p. 1312, 2020.

[13] M. Seyedan and F. Mafakheri, "Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities," *Journal of Big Data*, vol. 7, no. 1, p. 53, 2020.

[14] X. Huang, L. Gao, R. S. Crosbie, N. Zhang, G. Fu, and R. Doble, "Groundwater recharge prediction using linear regression, multi-layer perception network, and deep learning," *Water*, vol. 11, no. 9, p. 1879, 2019.

[15] T. Le, M. T. Vo, B. Vo, E. Hwang, S. Rho, and S. W. Baik, "Improving electric energy consumption prediction using cnn and bi-lstm," *Applied Sciences*, vol. 9, no. 20, p. 4237, 2019.

[16] D. Tanouz, R. R. Subramanian, D. Eswar, G. P. Reddy, A. R. Kumar, and C. V. Praneeth, "Credit card fraud detection using machine learning," in *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE, 2021, pp. 967–972.

[17] S. Hills and Y. Eraso, "Factors associated with non-adherence to social distancing rules during the covid-19 pandemic: a logistic regression analysis," *BMC Public Health*, vol. 21, pp. 1–25, 2021.

[18] M. Mamun, A. Farjana, M. Al Mamun, and M. S. Ahammed, "Lung cancer prediction model using ensemble learning techniques and a systematic review analysis," in *2022 IEEE World AI IoT Congress (AIIoT)*. IEEE, 2022, pp. 187–193.

[19] J.-W. Baek and K. Chung, "Context deep neural network model for predicting depression risk using multiple regression," *IEEE Access*, vol. 8, pp. 18 171–18 181, 2020.

[20] C. M. Liyew and H. A. Melese, "Machine learning techniques to predict daily rainfall amount," *Journal of Big Data*, vol. 8, pp. 1–11, 2021.

[21] R. Bhardwaj, "A predictive model for the evolution of covid-19," *Transactions of the Indian National Academy of Engineering*, vol. 5, no. 2, pp. 133–140, 2020.

[22] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of k-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning," *Decision Analytics Journal*, vol. 3, p. 100071, 2022.

[23] F. L. Huang, "Alternatives to logistic regression models in experimental studies," *The Journal of Experimental Education*, vol. 90, no. 1, pp. 213–228, 2022.

[24] A. Mehbodniya, I. Alam, S. Pande, R. Neware, K. P. Rane, M. Shabaz, and M. V. Madhavan, "[retracted] financial fraud detection in healthcare using machine learning and deep learning techniques," *Security and Communication Networks*, vol. 2021, no. 1, p. 9293877, 2021.