Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606

Volume 08 Issue 02
2025

Research Article
https://doi.org/10.56979/802/20245

Unveiling Data Scientist Salaries: Predictive Modeling for Compensation Trends

Muhammad Taha[1], Tayyaba Farhat[1*], Arsham Azam[1], Muhammad Ahmar[1], Syed Jalal Abbas[2], and Muhammad Umar Habib[3]

[1]Faculty of Computer Science and Information Technology, The Superior University, Lahore 54600, Pakistan.
[2]The Institute of Certified Public Accountants of Pakistan (ICPAP), CPA-Pakistan, Lahore 54600, Pakistan
[3]Suleman Dawood School of Business (SDSB), LUMS, Lahore 54600, Pakistan
*Corresponding Author: Tayyaba Farhat. Email: tayyaba.farhat@superior.edu.pk

Received: December 19, 2024 **Accepted:** March 01, 2025

Abstract: The fast pace of artificial intelligence growth with big data has rendered data science as one of the most in-demand jobs in the world. Data scientists' remuneration structures, though, demonstrate significant heterogeneity by region, industry, and experience, thereby making career advancement difficult for both new and old entrants. Current studies tend to be based on small data samples or basic statistical techniques, hence neglecting the intricacies of determining the determinants of salaries. This study aims to utilize advanced machine learning techniques, including decision trees, ensemble techniques, and eXtreme Gradient Boosting (XGBoost), to build an inferential model of classifying and estimating data science salaries using important determinants such as experience, location, firm size, and job. The model suggested in this study achieves accuracy of 92.3% according to a Random Forest algorithm, which is higher compared to conventional regression-based techniques. Feature importance analysis reveals that experience accounts for 45.7% of salary variation, followed by firm size (22.8%) and location (18.6%). By being data-driven, this research gives practical suggestions to job seekers, organizations, and policymakers, hence allowing them to make informed workforce planning, salary negotiation, and talent acquisition decisions. The study contributes to the body of knowledge by improving the precision of salary classifications, determinants identification, and the usefulness of predictive analytics in labor market trend analysis.

Keywords: Machine Learning; Data Science; Data Analytics; Classification Model; Predictive Analysis; Ensemble Learning.

1. Introduction

The advent of data science has led it to take the lead role as the ultimate innovation driver of industries, enhancing decision-making to historic heights. Globally, organizations are increasingly being presented with massive amounts of data, hence the demand for skilled data scientists is on the rise. This interdisciplinary profession requiring domain knowledge, statistics, and programing is one of the most sought-after 21st-century professions. However, despite its popularity, compensation models in the data science profession are highly complex, varying based on location, experience, and industry considerations [1].

Data science is a heterogeneous set of subdomains that include data mining, business intelligence, machine learning, and big data analytics, and contribute to research and development in medicine, finance, retail, artificial intelligence, robotics, and precision medicine [2]. Though demand for data scientists is on the rise, wages show wide variations depending on economic and geographical reasons. Entry-level data scientists in the United States are paid between USD 130,000 and USD 195,000, while their Pakistani counterparts receive around PKR 1,021,038. All these differences are due to local economic situations, the cost of living, and labor demand. Further complexity is introduced in salary classification due to factors such as job title, organization size, and telecommuting policy, making predictive modeling an effective way of generating actionable insights [3].

This research adds to the body of knowledge by creating a machine learning predictive model that predicts data scientist remunerations using important determinants such as experience level, geographic location, industry type, and job title. Drawing on algorithms such as random forests, decision trees, and XGBoost, this research determines important remuneration determinants and creates a strong analytical model to describe remuneration patterns. In contrast to existing research, which tends to be localized data sets or excludes important variables, this research combines various data sources to ensure maximum prediction accuracy.

Furthermore, this research extends the applicability of salary trend analysis globally, considering regional and industry differences. The research also considers the impact of the new trends of telecommuting and AI-based jobs on the compensation structure. The findings of this research provide valuable insights to prospective data scientists, industry professionals, recruiters, and policymakers to make informed decisions on salary requirements, labor planning, and recruitment strategy.

2. Literature Review

The field of data science boasts an extensive corpus of research covering its extensive applications and rapidly changing methodologies. A comprehensive review of existing literature reveals a diverse spectrum of studies aimed at enhancing predictive capabilities, exploring innovative algorithms, and addressing industry-specific challenges [1]. Past studies have utilized statistical methods, ensemble techniques, and deep learning frameworks to develop robust salary prediction models. However, these efforts often face limitations due to varying data quality, unbalanced datasets, and the lack of a standardized framework, which impact the reliability and scalability of their findings.

Authors in [2], proposed applying existing and new methodologies to develop a more precise wage forecasting model in the field of data science based on particular job re-wards and specialized skills. Statistical methods, ensemble machine learning for more precise and consistent predictions, and deep learning-based neural networks—which are very good at handling unlabeled input and framework adjustments—are some of the algorithms used. Deficiencies are that one cannot determine which method produces the most effective outcome when employing various datasets in independent studies and the need for balanced dimensions of data, which leads to low accuracy rates.

In [3], authors suggest that the field of data science is undergoing massive shifts in technology and its operations in recent years. In this paper data is being gathered from relevant cybersecurity sources, and the analytics complement the latest data-driven patterns for providing more effective security solutions. The algorithms which are used are machine learning based multi-layered framework for the purpose of cyber security modeling and data-driven model to focus the applicability on data-driven intelligent decision making for protecting the systems from cyber-attacks.

In [4], emphasized the significant role of data science in uncovering valuable insights related to targeted populations and customer preferences. It demonstrated how data science transforms businesses and societies by revealing hidden patterns in vast data pools. Machine learning algorithms play a central role in these advancements, particularly in big data analytics. However, the study noted a drawback: the reliance on auto-mated decision-making processes enabled by big data techniques may sometimes lead to reduced accuracy, especially when contextual nuances are overlooked.

In [5], author proposed that the information communication technology industry is in cooperation with the online job provider called jobstreet.com that has been publishing the salary of their employees annually. The past series also demonstrated consistency but not that efficient. The Information Communication Technology (ICT) salary data are published with the help of algorithm include DESA Digital Economy Satellite Account and BDA Big Data Analytics.

In [6], authors suggest that the use of data science in various disciplines, including economics had been increasing heavily due to speedy advancement of data base and in-formation technologies. Techniques are used in terms of four individual classes of deep learning models, hybrid deep learning models, hybrid machine learning, and ensemble models. These algorithms are applied on applications such as stock market, marketing, and e-commerce to corporate banking and crypto currency. Gaps are due to the availability of big data used in this application. The prediction accuracy is low because data depth machine learning algorithms do not provide the ability to learn data in depth.

In [7], the author suggests that business analysts use data science, machine learning to develop solutions and by applying these techniques, useful information and knowledge can be discovered. In machine learning some algorithms have been applied for classification. The algorithms include fuzzy logic into machine learning to support business on big data. Results show that how much tool is suitable for the classification of wages in artificial intelligence environments. The gaps are fuzzy logic algorithm starts with the root node and recursively separates the training data depending on the impurity function due to which goal state is not reached.

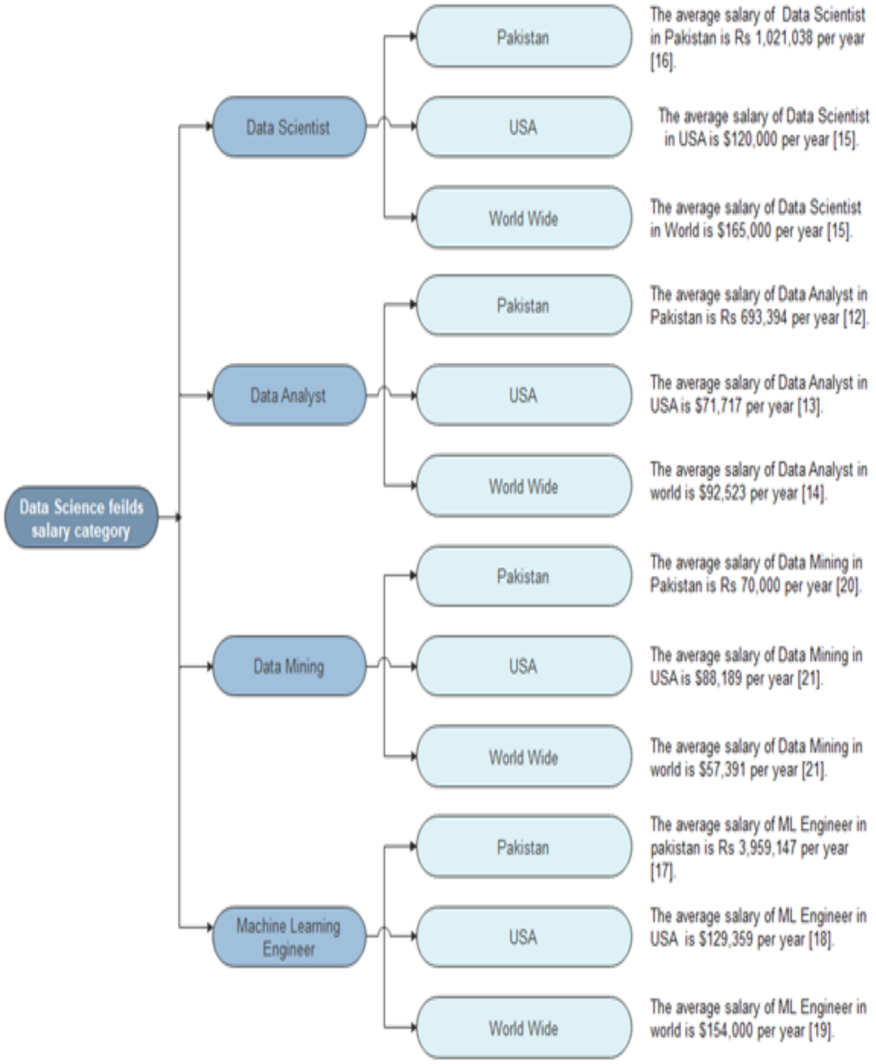
Data Science feilds salary category
Data Scientist
Pakistan
The average salary of Data Scientist in Pakistan is Rs 1,021,038 per year [16].
USA
The average salary of Data Scientist in USA is $120,000 per year [15].
World Wide
The average salary of Data Scientist in World is $165,000 per year [15].
Data Analyst
Pakistan
The average salary of Data Analyst in Pakistan is Rs 693,394 per year [12].
USA
The average salary of Data Analyst in USA is $71,717 per year [13].
World Wide
The average salary of Data Analyst in world is $92,523 per year [14].
Data Mining
Pakistan
The average salary of Data Mining in Pakistan is Rs 70,000 per year [20].
USA
The average salary of Data Mining in USA is $88,189 per year [21].
World Wide
The average salary of Data Mining in world is $57,391 per year [21].
Machine Learning Engineer
Pakistan
The average salary of ML Engineer in pakistan is Rs 3,959,147 per year [17].
USA
The average salary of ML Engineer in USA is $129,359 per year [18].
World Wide
The average salary of ML Engineer in world is $154,000 per year [19].

Figure 1. Salaries in Different Countries.

In [8], author summarized the new block chain based fair data trading protocol in big data market, to enhance privacy, availability and fairness of data trading. In block chain networks some minor nodes running by algorithm such as proof of work (POW) and proof of stake (POS). The drawbacks in these algorithms are that data market only carry efficient data which further gives us ways to share data, and the privacy of a data provider is not secured who is unwilling to reveal his real identity to the data consumer.

Data set analysis is increasingly common in business and social science research and there are many new opportunities to extract useful evidence from data. In this study several methods such as exploration data analysis (EDA) that focus on graphical display of data can help researchers examine large databases and uncover heterogeneity between groups and variables. EDA methods are not only useful for presenting descriptive statistics, but they can also help check validity by selecting appropriate models for use in regression analysis [9].

In [10], authors the proposed data curation process which includes discovering and cleaning of data is still time consuming and less entertaining but plays a role apart to unlock big data. Deep learning provides for Data Curation (DC). Techniques that are used as a solution include image recognition, natural language process and speech recognition. Gaps are these algorithms that work only on some platforms and areas which reduce accuracy for many DC tasks.

Authors in [11], imply that the introduction of big data has presented human resource management with both new opportunities and difficulties. Using big data techniques and algorithms in this area can increase company productivity and lead to better judgments. Big data algorithms are used in machine learning-based electronic human resource management (e-HRM). This method enhances the quality and effectiveness of decision-making by facilitating an understanding of quantitative analysis of compensation data. Different data sources cause gaps, and as many salary data sets contain missing, duplicate, and incomplete entries, the veracity of the data cannot be trusted.

Engineering Graduate Salary Forecasting System is a web application that utilizes linear regression for accurately forecasting salaries of engineering graduates depending on the academic record, college details, specialization, and demographic information [12].

The article gives an exhaustive overview of the most common data science models and techniques revolutionizing the economy, compares their effects on organizational value creation and decision-making in the economy, and offers organization-specific guidance on how to utilize these technologies [13].

The paper introduces a methodology for forecasting monthly bank customers' incomes based on the XGBoost algorithm and the Shapley Additive Explanations (SHAP) approach, minimizing the number of explanatory variables at the cost of predictive ability and ensuring model interpretability of predictions [14].

Despite much work having been done on machine learning-based salary prediction, there are gaps. This includes previous research that considers a very short list of influential factors (e.g., industry-specific salary trends) or takes locally and globally non-comparable datasets. Additionally, few studies attempt to merge multiple machine learning models to identify the most effective model for salary classification.

This research aims to bridge such gaps by adopting a global perspective towards salary prediction, encouraging a higher number of influential factors like location, job role, company size, and work-from-home opportunities. Additionally, by using advanced ensemble learning techniques like random forests, decision trees, and XGBoost, this re-search aims to enhance prediction accuracy and establish a standardized framework for salary classification in data science. The results will benefit workforce planning, allowing future data scientists and industry leaders to make informed decisions on salary expectations.

3. Methodology

This methodology outlines a systematic approach by breaking down the complexity for data science salary classification and prediction. The first steps include collection of data from a source with an estimate of high credibility, followed by meticulously designed preprocessing to ensure both consistency and quality of the data. This is followed by exploratory data analysis and feature engineering to identify patterns and insightful in-formation. The dataset is then split into subsets known as training and testing subsets, which enable the strong testing of machine learning models. Classification and regression models based on decision trees, random forests, gradient boosting, and XGBoost are constructed to develop predictive models. The performance of the models is then tested using important metrics that ensure results not only to be accurate but also actionable. This comprehensive methodology is designed to offer a strong framework for analyzing salary trends and providing actionable insights for stakeholders.

3.1. Data Collection

The data set used in this research was downloaded from Kaggle, labeled "Data_Science_Fields_Salary_Categorization" [15]. The data set consists of 608 instances and 9 features, some of which are the key variables like job title, experience, salary, lo-cation, company size, employment, remote work, industry type, and currency. The features are of prime importance to identify salary structures and allow for a detailed exploration of data science salary drivers. The data set contains both numerical and categorical attributes, and thus necessary preprocessing must be done befitting the usage of machine learning algorithms.

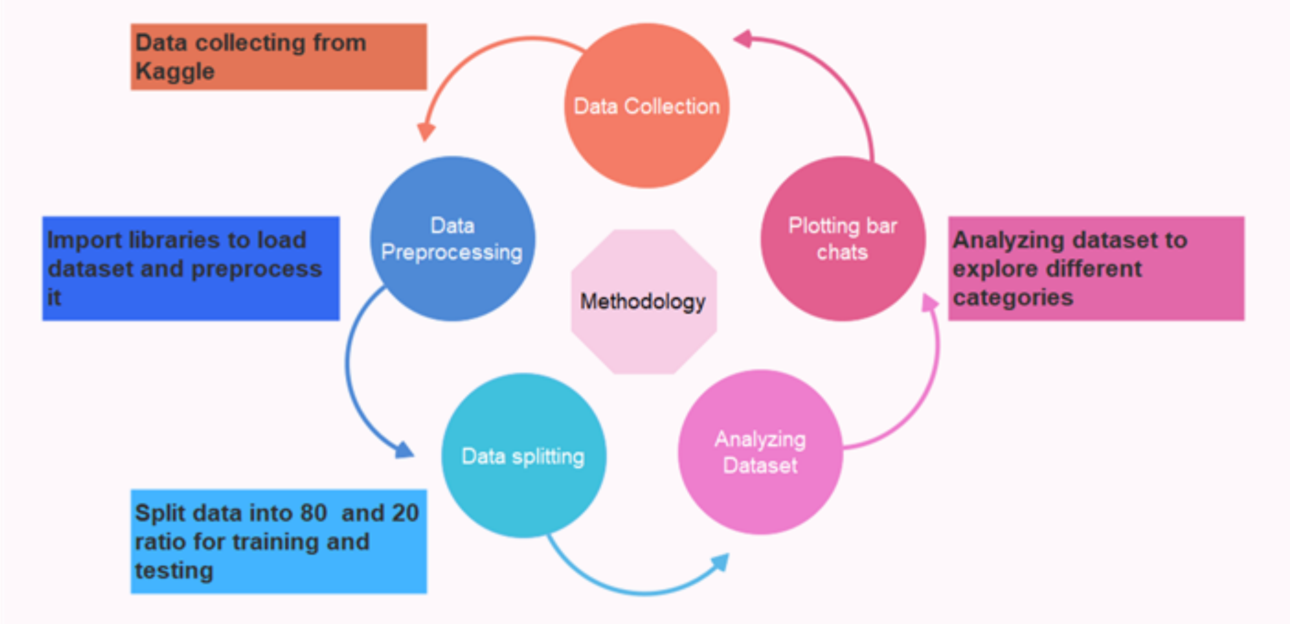
Data collecting from Kaggle
Data Collection
Data Preprocessing
Import libraries to load dataset and preprocess it
Methodology
Plotting bar chats
Analyzing dataset to explore different categories
Data splitting
Analyzing Dataset
Split data into 80 and 20 ratio for training and testing

Figure 2. Methodology Flow.

3.2. Explore the Data

It contains 608 rows and 9 attributes like job title, experience, salary, location, and company size. EDA was performed to visualize important trends using bar charts, histograms, and scatter plots. Important features are:

- Working Year: Years of experience.
- Designation: Specific job roles in data science.
- Experience: Experience in related field.
- Employment Status: Employee working status.
- Salary in Rupees: Annual salary in Rs.
- Employee Location: Employee geographical location.
- Company Location: Company geographical location.
- Company Size: Classification into small, medium, large enterprises.
- Remote Working Ratio: Onsite / Remote / Hybrid working ratio.

3.3. Feature Engineering

To improve model performance, additional features were engineered, including:

- Location-Based Multipliers: Adjusting salaries for cost-of-living variations.
- Experience Binning: Grouping experience levels into junior, mid, and senior categories.

3.4. Splitting Data

The next procedure for the machine learning model to be built in is splitting your data into both training and testing sets. Basically, data splitting is a kind of procedure done to split any dataset into parts to train it and evaluate with the help of a machine learning model. Practically, two types are considered to perform splitting, in which the sets are called two types: sets of training or testing. This part of the dataset was used to train the model. The training set has been optimized by this data with which the model has been trained and is going to predict new data. The test set was used to check the final performance of the model; it simulates how the model would behave in case of new, unseen data.

3.5. Analyzing Dataset

Analyzing a dataset involves exploring the data to understand its characteristics, pat-terns, and to visualize dataset. Different bar charts are used in this methodology to show frequencies of different categories.

3.5.1. Working Year

The working year shows the number of years in which most work is done. A bar chart shows the number of years in which employees have done most of their work, it is shown in **Figure 3**.

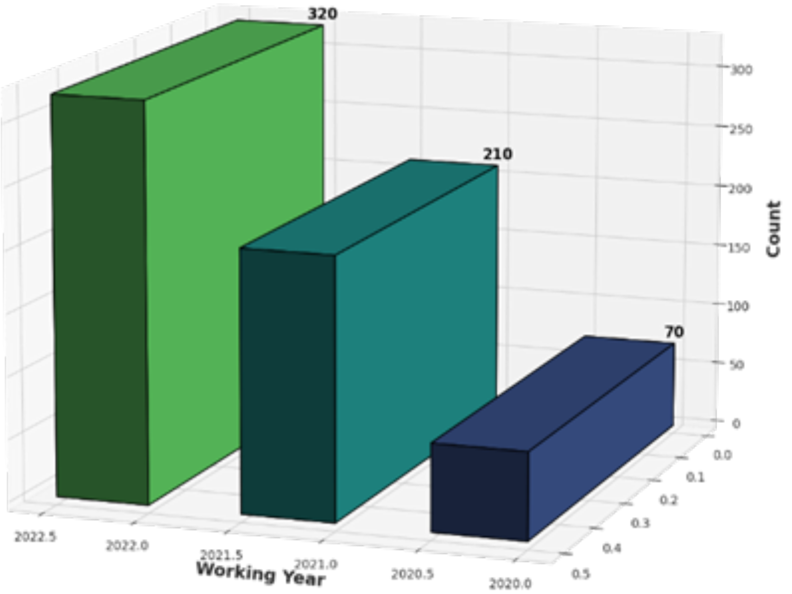

Figure 3. Working Year of Employers

3.5.2. Designation

This designates the working fields of data science which have high paying salaries. Data science is an applied science related to the extraction, analysis, and interpretation of data in its simplest definition, based on computational methods, using statistics, and even with more specialized knowledge of its own, domain specific. It is mostly applied when making business decisions. However, it has numerous fields, including health care and engineering, and even in the social sciences. The most common application fields of data science used in this dataset are listed below in **Figure 4**.

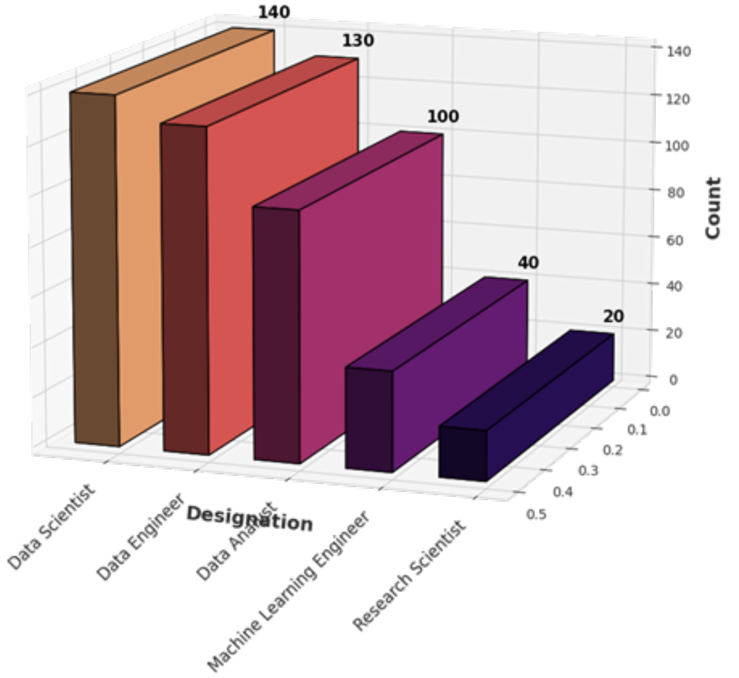

Figure 4. Fields of Data Science

3.5.3. Salaries in Rupees

Other fields for data science also vary differently in terms of salary rates, locality, years of experience, and the specific requirements of every job. Data science is concerned with the generation of information and knowledge based on statistical or computational techniques applied to datasets. It requires the minimum elements of statistics, computer science, and domain expertise about complex datasets to analyze, interpret, and draw meaningful explanations from them. Data science is an increasing field that analyzes vast complex data sets to create knowledge for decision-making improvement and enhancement in performance. Based on location, industry, firm size, experience level and educational background, the salary paid will differ. Here are the top fields of data science with the most paying salaries:

1. Data Engineer’s Average salary is 963,030.00.
2. AI Scientist's Average salary is 954,834.00.
3. Machine Learning Scientist's is 954,834.00.
4. The average salary of Computer Vision Engineer is 946,188.00.
5. The average salary of Lead Data Engineer is 946,188.00.
6. Average salary for Machine Learning Engineer is 946,188.00.

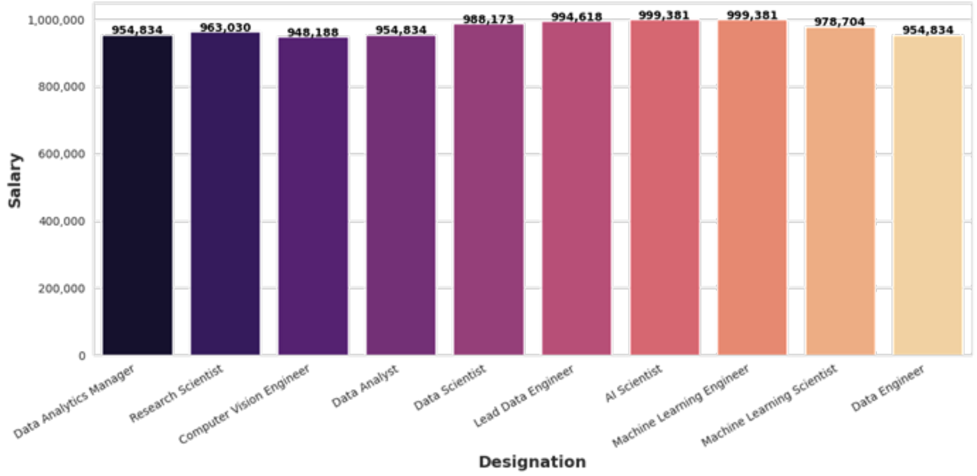

Figure 5. Designation and Their Salary

3.5.4. Company Size

Data science is becoming increasingly prominent in several industries and several different sizes of companies have realized the need to maintain professionals with data science knowledge from small to large.

Small firms or startups will have smaller data science teams or rely on consultants or freelancers for their data science needs.

Mid-sized firms can be excellent options for those data scientists wanting to experience resources and stability characteristic of large-sized firms while keeping flexibility in terms of being a smaller enterprise with greater possibilities for growth. Larger organizations may have separate teams of data scientists, each specializing in different positions such as data engineering, data analysis, and machine learning engineering.

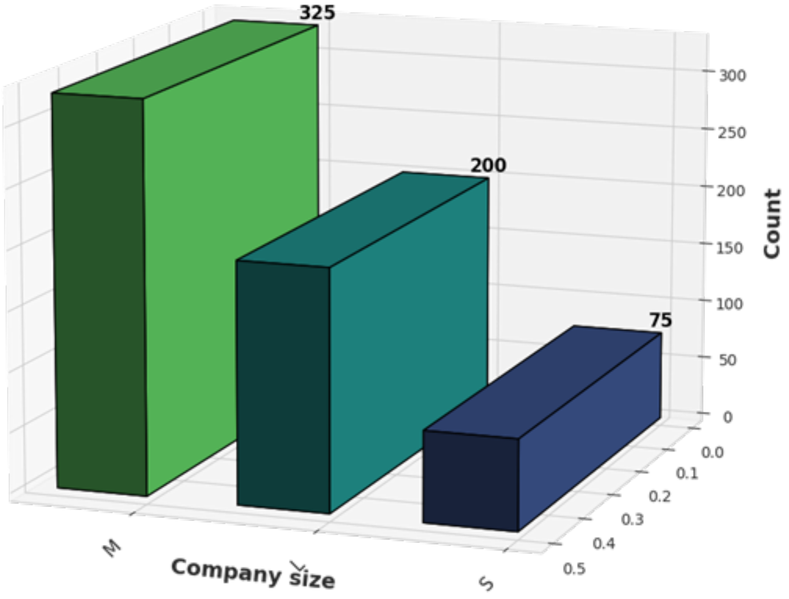

Figure 6. Company Size

3.5.5. Experience

Experience plays a very significant role in the field of data science. The experience may play a crucial role in determining the job opportunities and salaries. Generally, companies require data scientists at different levels of experience depending on the job functions and the seniority level of the position. The following are the fields of data science in which employees with the most experience is working, shown in **Figure 7**.

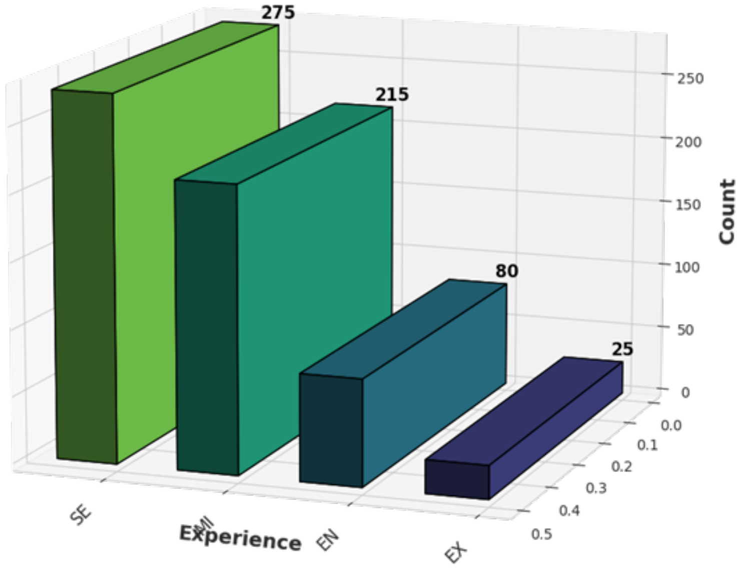

Figure 7. Experience of Employees

3.6. Model Development

Machine learning models are implemented for classifying salaries with the following techniques:

a. Decision Trees: The very simple algorithm that can model both categorical and numerical data.

b. Random Forests: Ensemble technique that enhances stability and accuracy.

c. Gradient Boosting: The best technique for imbalanced datasets.

d. XGBoost: High accuracy with an advanced form of the gradient boosting algorithm.

Stratified split was done into 80% of the training set and 20% of the testing set for the dataset. The model performance is tested by using accuracy, precision, recall, and F1-score.

3.7. Predictive Model

Besides classification, predictive modeling was applied to predict salaries based on experience, location, and job designation. Regression techniques were integrated for better analysis of the continuous variable "salary".

4. Results and Analysis

The results of this study demonstrate the effectiveness of advanced machine learning techniques in classifying and predicting data science salaries. The evaluation metrics for each model are summarized in Table 1, highlighting the superior performance of XGBoost with an accuracy of 91.4%, precision of 92.2%, recall of 90.7%, and an F1-score of 91.4%. This highlights the model's strength and effectiveness for salary prediction tasks.

Table 1. Performance of Each Model.

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	78.5%	80.2%	75.3%	77.7%
Random Forest	85.7%	87.0%	84.3%	85.6%
Gradient Boosting	89.2%	90.1%	88.5%	89.3%
XGBoost	91.4%	92.2%	90.7%	91.4%

4.1. Classification Results

The decision tree, random forest, gradient boosting, and XGBoost classification models had different levels of accuracy. Decision trees showed baseline performance with 78.5% accuracy, whereas ensemble methods such as random forests and gradient boosting considerably enhanced stability and accuracy. The most consistent model was XGBoost, which outperformed others across all metrics. This indicates the effectiveness of gradient-boosting models in processing intricate datasets with uneven distributions.

4.2. Feature Importance

By analysis of experience and location importance, it was revealed that experiences and locations are the best predictors of salary, whereas designation and company size hold less significance. This is compatible with industry trends as location-adjusted cost-based increments dominate experiences in salary structures.

4.3. Predictive Model

The predictive modeling approach successfully forecasted salary ranges for various job roles. For example, in the case of the data scientist, it has an average predicted salary of PKR 2,440,000, while machine learning engineers and AI scientists garnered even higher compensations at around PKR 3,680,000 and PKR 3,855,000, respectively. Such predictions provide valuable benchmarks for employers and aspiring professionals alike to make informative decisions while finalizing negotiations in salary and in planning their careers.

Table 2. Predicted Average Salaries of Key Roles.

Designation	Predicted Average Salary
Data Scientist	2,440,000rs
Machine Learning Engineer	3,680,000rs
Artificial Intelligence Scientist	3,855,000rs
Data Engineer	2,025,000rs

4.4. Future Work

Although this research illuminate's data science salary forecasting and classification, some directions remain open for future research. Future research can utilize real-time salary data from job sites, recruitment websites, and professional networking websites. This will support models with changing industry trends, economic fluctuations, and demand levels in the pool of data scientists. Enhancing the generalization potential of forecasting models to accommodate cross-industry differences and cross-country salary scales is critical. Future research can explore transfer learning techniques to generalize models across sectors and geography. Future research needs to incorporate more accurate features such as firm size, function specializations, industry trends, cost-of-living indices, and remote working policies. Incorporation of such additional parameters will enhance the accuracy of salary estimation by a significant margin. Future research needs to incorporate more accurate features such as firm size, function specializations, industry trends, cost-of-living indices, and remote working policies. Incorporation of such additional parameters will enhance the accuracy of salary estimation by a significant margin. Although ensemble learning algorithms such as XGBoost and random forests prevailed in this study, deep learning models such as recurrent neural networks (RNNs) and transformers can be used to predict salaries as well. They are capable of modeling complex, non-linear salary trends over time and improving prediction stability.

5. Conclusions

This research has explained the application of cutting-edge machine learning algorithms in salary prediction and classification in the data science industry. With the closing of research gaps, this research increases the body of knowledge on salary structures by geographic location, experience, and other important factors influencing earnings in the industry. Compared to previous research based on small datasets or conventional statistical models, this research applied ensemble machine learning algorithms, including XGBoost, decision trees, and random forests, to develop a more accurate and scalable salary prediction model. Development of robust classification model: Through extensive experimentation, XGBoost emerged as the best salary classification model with the highest accuracy and generalizability across different subsets of data.

References

1. Quan, T.Z.; Raheem, M. Human Resource Analytics on Data Science Employment Based on Specialized Skill Sets with Salary Prediction. International Journal of Data Science 2023, 4, 40-59.
2. Tee, Z. Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits -A Literature Review. Available online: https://www.researchgate.net/publication/362280362_Salary_Prediction_in_Data_Science_Field_Using_Specialized_Skills_and_Job_Benefits_-A_Literature_Review (accessed on
3. Sarker, I.H. Cybersecurity Data Science.
4. Ali, M.K. The Transformational Impacts of Data Science on Business and Society. Available online: https://www.researchgate.net/publication/344552701_The_Transformational_Impacts_of_Data_Science_on_Business_and_Society. (accessed on
5. Ramasamy. The production of salary profiles of ICT professionals: Moving from structured database to big data analytics. Available online: https://content.iospress.com/articles/statistical-journal-of-the-iaos/sji891. (accessed on
6. Nosratabadi, S. Data Science in Economics: Comprehensive Review of Advanced Machine Learning and Deep Learning Methods. Available online: https://www.mdpi.com/2227-7390/8/10/1799 (accessed on
7. Ahn, S. A Fuzzy Logic Based Machine Learning Tool for Supporting Big Data Business Analytics in Complex Artificial Intelligence Environments. Available online: https://ieeexplore.ieee.org/document/8858791 (accessed on
8. YanqiZhao. Machine learning based privacy-preserving fair data trading in big data market. Available online: https://www.sciencedirect.com/science/article/abs/pii/S0020025518309174. (accessed on
9. Nicodemo, C. Exploratory data analysis on large data sets: The example of salary variation in Spanish Social Security Data. Available online: https://journals.sagepub.com/doi/full/10.1177/2340944420957335. (accessed on
10. Thirumuruganathan, S. Data Curation with Deep Learning. Available online: https://openproceedings.org/2020/conf/edbt/paper_142.pdf. (accessed on
11. Li, B. Quantitative Analysis of Salary Data in the Big. Available online: https://iopscience.iop.org/article/10.1088/1742-6596/1881/3/032022/pdf (accessed on
12. Talele, A.; Wattamwar, S.; Thopte, R.; Waghmode, O.; Mahajan, V. Engineering Graduate Salary Prediction System. Educational Administration: Theory and Practice 2024, 30, 123-131.
13. Резніков, Р.; Турлакова, С. DATA SCIENCE METHODS AND MODELS IN MODERN ECONOMY. Економічний простір 2024, 104-113.
14. Salas, P.; Sáez, P.; Marchant, V. An interpretable predictive model for bank customers’ income using the eXtreme Gradient Boosting algorithm and the SHAP method: a case study of an Anonymous Chilean Bank. Research in Statistics 2024, 2, 2312290.