41% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Detection Groups

27 AI-generated only 38%
Likely AI-generated text from a large-language model.

4 AI-generated text that was AI-paraphrased 3%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (it may misidentify writing that is likely AI generated as AI generated and AI paraphrased or likely AI generated and AI paraphrased writing as only AI generated) so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.

Frequently Asked Questions

How should I interpret Turnitin's AI writing percentage and false positives?
The percentage shown in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was either likely AI-generated text from a large-language model or likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

False positives (incorrectly flagging human-written text as AI-generated) are a possibility in AI models.

AI detection scores under 20%, which we do not surface in new reports, have a higher likelihood of false positives. To reduce the likelihood of misinterpretation, no score or highlights are attributed and are indicated with an asterisk in the report (*%).

The AI writing percentage should not be the sole basis to determine whether misconduct has occurred. The reviewer/instructor should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in accordance with their school's policies.

What does 'qualifying text' mean?
Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be likely AI-generated will be highlighted in cyan in the submission, and likely AI-generated and then likely AI-paraphrased will be highlighted purple.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

༄༅།། ཤེས་རབ་རི་མཐོ་རིམ་སློབ་གྲྭ།

འབྲུག་རྒྱལ་འཛིན་གཙུག་ལག་སློབ་སེ།།

SHERUBTSE COLLEGE

ROYAL UNIVERSITY OF BHUTAN

Predictive Analysis of Salary

BY

MR. SONAM TSHOCHEN

MS. PASANG GEM

MS. KARMA TSHOMO

MR. TASHI WNAGCHUK

AN UNDERGRADUATE RESEARCH

BSC IN STATISTICS

DEPARTMENT OF MATHEMATICAL SCIENCE

SHERUBTSE COLLEGE

ROYAL UNIVERSITY OF BHUTAN ACADEMIC YEAR 2025

SHERUBTSE COLLEGE

ROYAL UNIVERSITY OF BHUTAN

BSC IN STATISTICS

RESEARCH

BY

MR. SONAM TSHOCHEN

MS. PASANG GEM

MS. KARMA TSHOMO

MR. TASHI WANGCHUK

ENTITLED

PREDICTIVE ANALYSIS OF SALARY

Was approved on ________

Supervisor:	MR. UGYEN SAMDRUP TSHERING
	MR. EBA RAJ DAHAL

Predictive Analysis of Salary

Sonam Tshochen[1], Tashi Wangchuk[1], Karma Tshomo[1], Pasang Gem[1], Ugyen Samdrup Tshering[2*], Eba Raj Dahal[2*]

1. Department of Mathematical Science (Statistics program), Sherubtse College, Royal Government of Bhutan, Kanglung-42002, Trashigang, Kingdom of Bhutan.
2. School of Data Science & Data Analytics, Sherubtse College, Royal Government of Bhutan, Kanglung-42002, Trashigang, Kingdom of Bhutan.

[*] *Corresponding author:utshering.sherubtse@rub.edu.bt*

Abstract

This capstone project aims to compare the effectiveness of four machine learning algorithms for salary prediction: linear regression, random forest, gradient boosting, and neural networks. The main goal is to find the model that produces the most accurate, effective, and comprehensible predictions based on independent factors like age, gender, years of experience, job title, and educational level. The project begins with data exploration, cleaning, and preprocessing, followed by model development and evaluation using performance metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and R-squared (R²). The study assesses the models' effectiveness as well as their interpretability in order to offer practical insights into the variables affecting salary projections. This study will advance predictive analytics by determining the top-performing model, providing useful advice to companies, job seekers, and policymakers. Finding the most accurate and effective model for predicting salaries with knowledge of how various factors affect pay is the anticipated result. This study might significantly benefit the community by influencing economic analysis, workforce planning, and wage negotiations.

Keywords: Predictive Analysis, Salary, linear regression, random forest, Neural Networks, Gradient Boosting, Cross-Validation.

1. Introduction

In the contemporary context of the digital age economy, information has emerged as a key factor that shapes decision-making across different industries. Organizations are leveraging knowledge extracted from information to enhance effectiveness in operations and policy formulation. One of the most promising areas where predictive data analysis has worked well is the prediction of salaries (Matbouli & Alghamdi, 2022). Estimating salaries accurately according to personal and professional characteristics is not only valuable for organizations seeking to ensure competitiveness and fairness in their compensation strategies but also for prospective employees looking to make extremely well-informed career choices (Meng et al.,2018) In a wider context, sound salary prediction is of strategic interest to policymakers by illuminating the dynamics of labor market trends and possible imbalances in income distribution (Varian, 2014). Inspite of its significance, salary prediction is a challenging issue because of the multidimensional and commonly nonlinear character of the determinants of compensation (Feng et al.2023; Jiang, 2024). This Capstone project attempts to solve this issue by applying contemporary machine learning (ML) algorithms to develop models that can successfully predict salaries from organized employee data (Li et al., 2024).

Conventional salary estimation methods employ heuristic-based decision processes, descriptive statistical inference, or regression analysis assuming linear relationships between variables (James et al.,2021). These approaches, although providing an initial insight, fail to fully capture the complexity of the interaction that exists between diverse personal and work-related variables (Saraswathi et al., 2023). For example, the influence of education level on salary can change considerably in relation to factors such as age, gender, years of experience, education level, and job title. With these weaknesses, machine learning offers itself as a flexible alternative. Machine learning algorithms excel at picking up on faint underlying patterns within data sets without shying away from making severe parametric assumptions (Hastie et al., 2009). This work utilizes machine learning approaches not just to boost the accuracy of predictions, but also to improve the interpretability and real-world applicability of models for salary prediction.

The key objective of this Capstone project is to critically assess and contrast the efficacy of four different machine learning algorithms for salary prediction. The models in focus are Linear Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Neural Network (NN). Each of these algorithms possesses some advantages. LR model provides transparency and simplicity of interpretation, while RF and GB models exhibit skill in identifying nonlinear relationships and in overfitting reduction by employing ensemble learning techniques (Hussain, 2024; Li et al., 2024). Despite the fact that NN are characterized by their complexity and lack of transparency, they provide a powerful framework for modeling complicated interactions among variables (Feng et al., 2023). The objective of this project is to systematically explore which of these methods produces the highest accuracy and reliability when predicting continuous salary data.

In order to perform this analysis, a dataset of anonymized employee data is used. The dataset has attributes like age, gender, years of experience, education level, and job title. The presence of these variables allows for a comprehensive analysis of the impact each factor has on salary levels. Preprocessing operations like missing value filling, feature encoding, outlier identification, and normalization are performed to make the data consistent and of high quality (James et al., 2021). Furthermore, feature selection and dimensionality reduction techniques are explored with the aim of improving the performance of the models, furthermore to reducing possible challenges such as multicollinearity (Hastie et al., 2009). Training and validation of each model is done using cross-validation techniques with the performance tested against conventional regression evaluation metrics, which are the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination (R^2) (Saraswathi et al., 2023; Hussain, 2024).

Apart from the predictive accuracy, usability and trustworthiness of the model are also essential features in this research. Though the ability to make precise predictions is the core goal, it is important that the outcomes are reported in a manner that facilitates sound decision-making (Meng et al., 2018). Interpretability of the outcomes is particularly significant in applications such as salary prediction, where decision-makers may be interested in knowing the basis for various salary ranges. Less complex models such as LR possess interpretability as an inherent characteristic, while ensemble models such as RF and GB give feature importance values that offer grounding intuition into the effect of variables. These aspects form the foundation for the project's objective of developing not only accurate but also decision-useful models for application in salary planning and analysis (Matbouli & Alghamdi, 2022; Li et al., 2024).

This research is important because it has the potential to provide several benefits to several stakeholders. To employers, the research provides an evidence-based justification for developing competitive and fair salary structures. This is particularly pertinent given heightened scrutiny related to pay equity and increasing demands for transparency in corporate compensation policies. For workers, access to realistic salary projections can enhance bargaining power and career planning, thereby improving job satisfaction and reducing turnover (Jiang, 2024). For policymakers and labor economists, the models and insights developed in this research can aid in the detection of wage inequities, monitoring labor market trends, and developing evidence-based employment policy (Varian, 2014). Scholarship-wise, the project adds to the accumulating literature on the application of machine learning in the discipline of human capital management, thus spanning theoretical development and practical implementation (Meng et al., 2018; Hussain, 2024).

The project scope is definitively bounded to heighten manageability and concentration. The analysis only addresses structured data, such as demographic and occupational factors pertinent to the assessment of salary. Extrinsic macroeconomic determinants, for instance, inflation rates, geographic living costs, or industry conventions, are excluded from this analysis because of the unavailability of data (Li et al., 2024). Moreover, the models developed are intended to be widely applicable and are not tailored to specific organizations or sectors. While this limitation diminishes the precision of the predictions, it enhances the generality of the findings.

The work was conducted over a planned period of three months, divided into phases to enable orderly progress. The initial phase consisted of topic selection, literature review, and proposal formulation, where the research problem was refined and project objectives were consolidated. Data collection, exploration, and preprocessing were addressed in the second phase. Significant emphasis was given to upholding data integrity, which included missing value identification and handling, outlier handling, and encoding categorical variables. In the third phase, model selection and training processes were carried out. Each model was implemented via the use of Python programming with relevant machine learning libraries, which include Scikit-learn (Pedregosa et al., 2011). To further improve model performance, hyperparameter tuning was conducted using strategies such as grid search and random search (James et al., 2021). Phase four focused on the evaluation and interpretation of the models, whereby the models were tested on holdout datasets and the corresponding performance metrics were computed. Finally, the fifth phase involved the documentation and presentation of the outcomes, which included report writing, visual aid development, and submission of the final outputs.

For maintaining the smoothness of the project flow and enhanced efficiency, various tools and platforms were employed. Python was employed as the primary programming language due to its comprehensive data science libraries. Jupyter Notebooks were utilized for developing code and visualization in order to enable interactive analysis and debugging steps. Preprocessing of data and exploratory analysis was conducted using pandas and seaborn libraries, while model development was implemented using Scikit-learn libraries (Pedregosa et al., 2011). Gantt charts and project timeline visualizations were created using Microsoft Excel and Google Sheets. Collectively, these software packages made for an efficient, well-organized, and well-documented research project.

A feasibility analysis was performed before the initiation phase to evaluate the project's viability from several angles, encompassing technical, operational, economic, and scheduling feasibility. From a technical viewpoint, the project utilized open-source tools and platforms that are well-regarded within the data science community. All essential computational tasks could be executed using freely accessible resources, thereby negating the necessity for costly software or hardware. Operational feasibility was determined to be high as the project had a clear workflow with realistic milestones. Economic feasibility was assessed as positive, considering that the project involved no direct costs as free tools and publicly available data were utilized. The time for each task of the project was properly calculated so that the workload would be within reasonable bounds given the time constraints of the academic year. The possible problems, including overfitting and multicollinearity, were pre-empted and tackled using strict model validation procedures and regularization methods Hastie et al., (2009). In summary, the feasibility study demonstrated that the project was not only achievable but also well placed to be successfully executed given the available resources and time frame.

This Capstone project is an exhaustive exploration of the application of ML models in predicting salaries. It encompasses stringent data preprocessing, relative comparisons of several models, and interpretability analysis to provide valuable insights for various stakeholders. The systematic approach, underpinned by a viable strategy and suitable tools, guarantees academic integrity as well as applicability in practical situations. Through the identification of the best model to predict salaries and the exploration of determinants with the largest impact on compensation, the project seeks to make an important contribution to the areas of predictive analytics, human resource management, and labor market analysis Meng et al., (2018). Generally, the project seeks to illustrate the promise of machine learning as a transformative technology for facilitating data-driven decision-making in intricate, real-world problems like employee compensation (Li et al., 2024; Hussain, 2024).

Literature Review

Salary prediction has become popular for data-driven decisions and actions making, particularly in applications like labor market analysis, economics, and human resources. A number of studies have attempted to establish the effect of variables on earnings using statistical techniques like simple and multiple linear regression. Such traditional models are used because they are uncomplicated and easy to understand. For example, more advanced models such as LR, RF, support vector machines (SVM), NN, and GB have been introduced in machine learning, offering improved accuracy by identifying complex, nonlinear patterns in data. The capabilities of machine learning have been greatly increased by these models, allowing it to tackle a variety of challenging issues. Employers can manage budget, recruit talent, and provide fair compensation by using salary prediction, which also helps workers understand their market value (Biswas et al., 2023).

LR is a statistical modelling technique used to establish a relationship between a dependent variable and one or more independent variables (Gurpreet et al., 2023). The author has expanded this interest with the proposal of models that are LR-based with a focus on software engineers primarily focusing on years of experience as the key predictor. The model fails to incorporate more important features such as age, gender, job title and education level. However, such models are not able to capture complex and nonlinear interactions. Similarly, (Lothe et al., 2021) use LR predictive model with suitable algorithm using key features required to predict the salary of employees. They and found out that years of experience have the highest correlation with salary. They enhance the model by utilizing Polynomial Transformation to decrease error and increase accuracy. It was done mainly to concentrate on academic experimentation with intricate performance measurements like as correlation analysis and MSE**.** To address these limitations in the existing literature or body of knowledge, our project aims to use a larger array of features to investigate and contrast sophisticated regression techniques such as GB, RF, and NN, and LR for creating a more precise and generalizable salary prediction system.

Salary prediction is frequently used to help with career planning, support salary negotiations, and determine appropriate remuneration (Bao, 2024). Research by Das and Barik (2020) supports on using linear regression and polynomial regression models to predict an individual salary based on their past salary growth. The system was able to accurately predict salary and produce a clear graphical representation of salary growth. It enabled users to track patterns and make reasonably accurate salary estimates over time. However, using only linear and polynomial regression may not be sufficient to capture the complex patterns found in salary data may offer less accurate predictions. More sophisticated models such as RF and GB provide better accuracy by better managing non linearities and variable interactions (Hussain & Ahmad, 2024).

ML methods such as RF and decision trees have been investigated to get around the drawbacks of LR model. Monteiro (2023) developed a salary prediction model using the RF algorithm, not just based on individual data, but also on macroeconomic factors such as average wage, GDP per capita, CPI, PPI, and inflation. This model offers a broader and more accurate way to predict salaries. The author concluded that RF was more accurate than conventional regression techniques. Similarly, Ren et al. (2023) has used RF to predict salary and improve accuracy. However, in order to verify the effectiveness of the proposed technique, they compare it with other approaches including decision tree, logistic regression, naive bayes, (NN) and KN neighbor. Thus, they found out that RF Model with 84.49 % has the highest effectiveness and best performance among other models. Same like that our project aims to compare four model and see whether RF predicts the salary well.

GB is an ensemble method that focuses on minimizing errors to enhance model accuracy. Erpapalemlah (2023) used GB model in order to facilitate better decision making for both job seekers and employers. The predictive accuracy of the models was 0.321 MAE and mean absolute percentage error (MAPE) of 2.856 %. The author note that GB is a better model to predict salary which can help job seekers and employers in making informed decisions instead of comparing it with different models. Likewise, one of the study have used GB Regressor and RF Regressor models for employee salary prediction with a promising accuracy rate of 99% with regression models (Hussain & Ahmad, 2024). They found out RF model achieved a higher accuracy of 96% compared to gradient boosting model. Nevertheless, a number of important features are still not fully examined in both studies, even if the RF model and GB model shows encouraging performance. With this limitation, our project aims to compare four different advanced model with important feature and select the best model for the salary prediction.

A comparative analysis of models helps in selecting the optimal approach for predicting salaries. Biswas et al. (2023) compared various model including LR and RF model to predict the salaries of employees based on various factor. Their study was mainly to predict the most preferred algorithm for salary and to find which model offer high accuracy and low error value. LR model, being a lightweight and interpretable yielded decent performance indicating constraints in modeling complex patterns in the data (Biswas et al., 2023). However, RF model did significantly better compared to LR model. As it has a capacity to handle non-linear interactions and interactions between features, but at higher computational expense. Similarly, Yang (2023) discovered that GB techniques achieved even better predictive performance than RF model. The author further highlights the importance of using robust, non-linear models when dealing with multifactor salary data. These studies collectively emphasize that simpler models like LR are useful for interpretability. And also, the more advanced models like RF and GB are better suited for achieving higher predictive accuracy in real-world salary prediction tasks.

Devi and Neelambika (2023) have demonstrated the effectiveness of individual regression models like RF, Decision Tree (DT) regressor, and LR, whereby reaching an accuracy of 95.68% on the training dataset and 95.33% on the testing dataset**.** These findings show how ML methods, in particular LR, may be used to accurately forecast graduate salaries. They did not clearly mention how LR performs better to predict salaries. To overcome these challenges, our project aims to investigate which model performs better to predict salary using ML method. NN model have also become a viable method for predicting salaries, especially when working with extremely complicated salary determinants (Dsouza, 2024). The author investigated how NN model was used to predict technology industry salaries. The NN model reveal non-linear interactions and hidden patterns that conventional models are unable to identify. However, despite their high predictive capability, NN model require large data and lots of computing power (Dsouza, 2024).

Matbouli and Alghamdi (2022) illustrated that employers can use salary prediction models to develop fair pay plans, job seekers can use them to make educated career decisions, and politicians can use them to assess compensation trends and address income. Five models are investigated and RF show the highest accuracy among all. This approach is further supported by Bao (2024), who explained that advanced ML models, particularly Random Forest, achieved high salary prediction accuracy when combined with careful data preprocessing and feature selection. The author emphasized the importance of integrating demographic, educational, and professional variables to enhance the predictive strength of models. It aligns with the need for broader and more dynamic feature sets in salary prediction research. Moreover, they highlighted that rigorous cross-validation and tuning practices were key to minimizing prediction errors, reinforcing the necessity of these techniques for creating reliable and generalizable models.

3.3 Methodology

3.3.1 Methods Used to Provide Structure and Cohesiveness

The project followed a systematic data science workflow to ensure structural integrity and cohesion across all phases of development. The approach employed ensured replicability, traceability, and data-driven decision-making, utilizing best practices in data preprocessing, model selection, evaluation, and tuning.

3.3.1.1 Workflow from Inception to Completion

The life cycle of the project was segregated into a series of structured phases to ensure a proper and systematic process. It began with data acquisition of the dataset, followed by preprocessing of the data, where missing values were handled, categorical fields like education level and job role were encoded, and numerical attributes like year of experience and age were normalized. The second phase was Exploratory Data Analysis (EDA) employing Python libraries such as Seaborn and Matplotlib to find trends, patterns, and correlations in the data. In model construction, multiple predictive models such as LR, RF, GB and NN were used to assess and compare their performance. The models were evaluated on significant measures such as RMSE, MAE, and R^2, and cross-validation was used to generate solid and generalizable results. Finally, the model with the lowest error and greatest value of R^2 was found to be the best model and used to predict salaries.

3.3.1.1.1 Data Preprocessing

In the process of preprocessing, several crucial steps were performed in cleaning the data and preparing it for quality modeling. Firstly, missing values for the dataset were examined using the isnull().sum() function. It was observed that in columns such as Age, Gender, Education Level, Job Title, Years of Experience, and Salary, there were few missing values. The issue was therefore addressed through the application of appropriate imputation strategies. Median values has been used to fill the missing entries in numerical columns like Age, Years of Experience, and Salary, as the median is less prone to outliers and preserves the central tendency of the data. For categorical variables such as Gender, Education Level, and Job Title, mode (mode being the most frequent category) was used for imputation to preserve category frequency consistency.

In addition, cleaning processes on the data were conducted to address inconsistencies in the categorical values. The data set had variations in naming within the "Education Level" column, such as "Bachelor's Degree" and "Master's Degree" alongside "Bachelor's" and "Master's". They were standardized using the replace() function to consolidate them into uniform categories such as "Bachelor's", "Master's", and "PhD". Minor anomalies such as "phD" were also corrected during the process. For the "Gender" column, entries labeled as "Other" were removed from the dataset because of their small count, enabling a more distinct binary division between "Male" and "Female.".

Categorical and numerical columns were then identified separately for further processing. The categorical features included Gender, Education Level, and Job Title, while the numerical ones included Age, Years of Experience, and Salary. A check for duplicate values was implied, and overall, the dataset was thoroughly cleaned, imputed, and standardized to ensure that it was ready for the next stages of exploratory data analysis and model building.

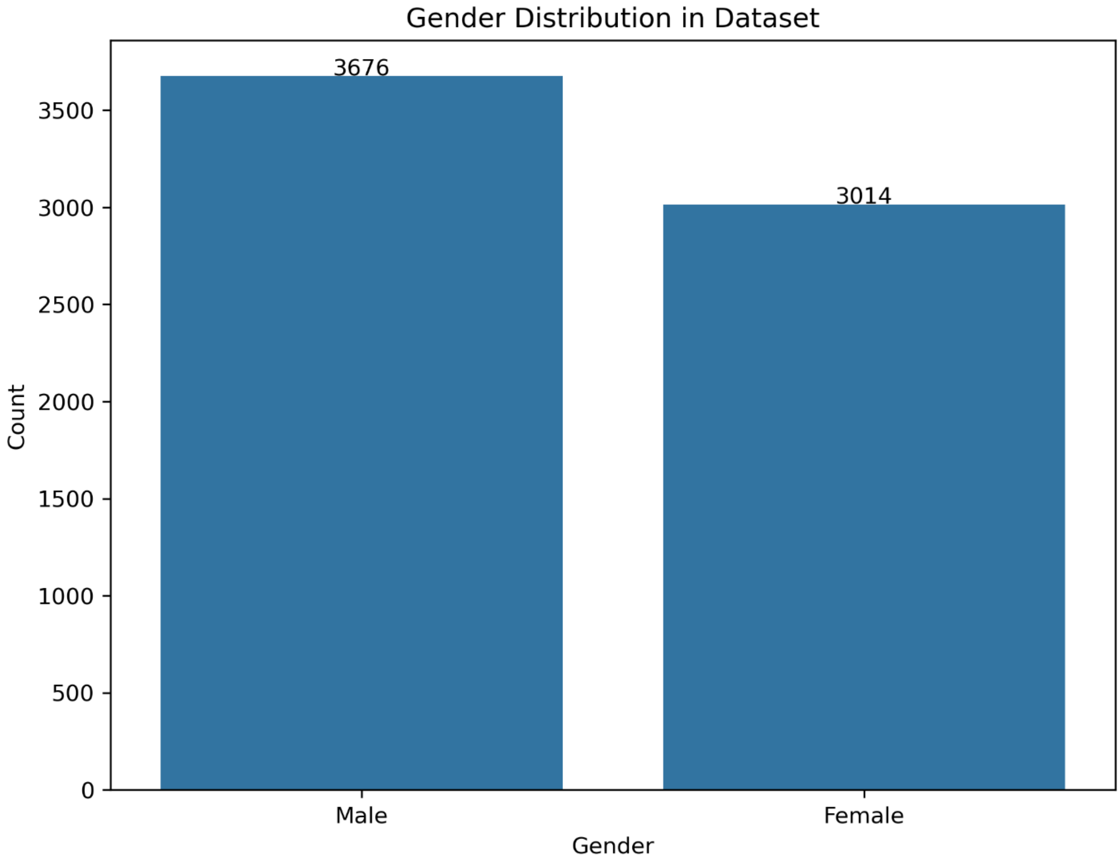
Additional important steps were undertaken to ensure data quality and model readiness by addressing multicollinearity and outliers. To mitigate the risk of multicollinearity—two or more independent variables with high correlation—the Variance Inflation Factor (VIF) was determined for each numerical feature. Excessive multicollinearity may distort model coefficient interpretation and predictive performance. Any of the variables with VIF greater than 5 were considered to have significant multicollinearity and were removed to preserve model stability and interpretability. This step was meant to enhance the robustness of the subsequent modeling by avoiding redundancy among the predictors.

Outliers in numerical features were also addressed to prevent them from skewing the analysis and negatively impacting model performance. The Interquartile Range (IQR) method was used to detect and remove outliers. By calculating the 25th (Q1) and 75th (Q3) percentiles for each numerical column, values lying outside 1.5 times the IQR from these quartiles were considered outliers and were excluded from the dataset. This ensured that the data retained its representative central distribution without being disproportionately influenced by extreme values.

3.3.1.1.2 EDA

Gender distribution in dataset

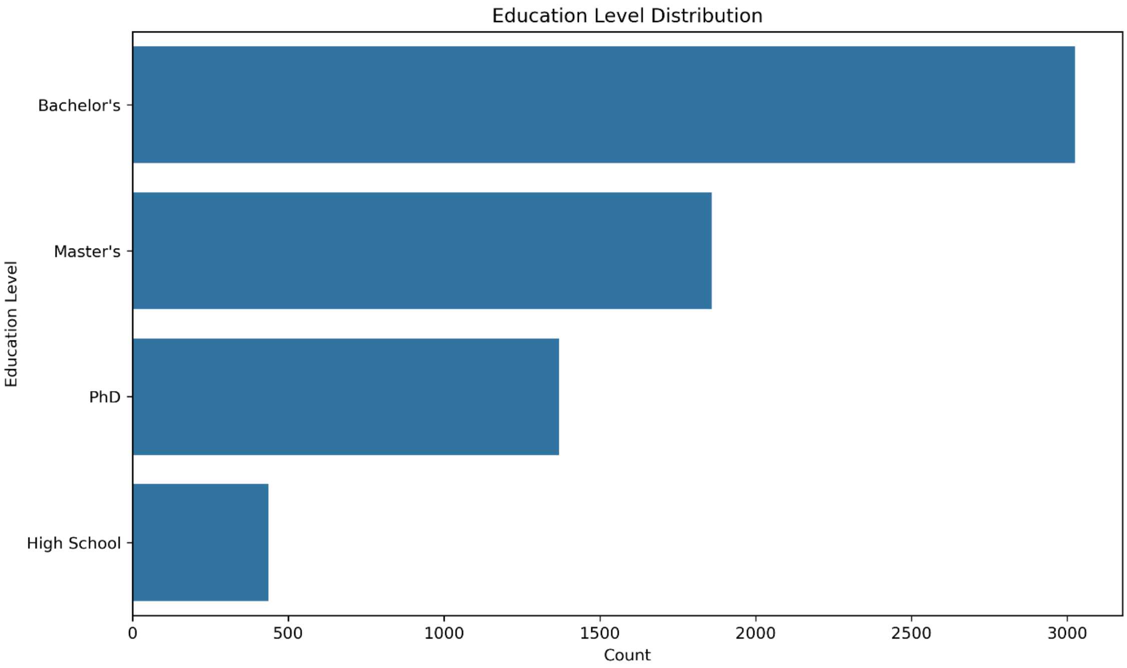
The gender distribution in the dataset indicates that male workers 3,676 outnumber female workers 3,014. When salary figures are distinguished by gender, there is noticeable difference in earnings. Male employees on average earn a higher mean salary of 121,383 compared to female employees with 107,891. The median salary also indicates the same trend, with males earning 120,000 and females earning 105,000. Both groups have the same standard deviations, indicating equal salary variation within each group. These results suggest a probable gender pay gap where male employees tend to earn more than female employees. This result is worth exploring to determine if other variables such as the job title, level of education, or years of service affect the gap.



```
   Salary Statistics by Gender:
                 mean                median       std              count

    Gender

    Female       107891.357996       105000.0     52715.019086     3014

    Male         121382.917845       120000.0     52064.795192     3676
```

Education Level Distribution in Dataset

The education level distribution within the dataset indicates that majority of the individuals have a Bachelor's degree, followed by Master's degree, PhD, and High School education, respectively. Examining the salary data in terms of education level yields a distinct positive correlation between education and earnings. Those with a PhD hold the highest mean salary of 165,651, a median salary of 170,000, indicating that they earn high and consistent incomes. They are followed by those with a Master's degree, who have a mean salary of 130,070. Bachelor's degree holders have much lower incomes with a mean of 95,111. High School education holders only earn the lowest mean salary of 34,416. The standard deviation in salaries also tends to decline with higher education levels, particularly among PhDs, indicating more consistent salary levels. These findings strongly suggest that a higher education level coefficient with higher and steady salaries, supporting the value of higher academic credentials in the job market.

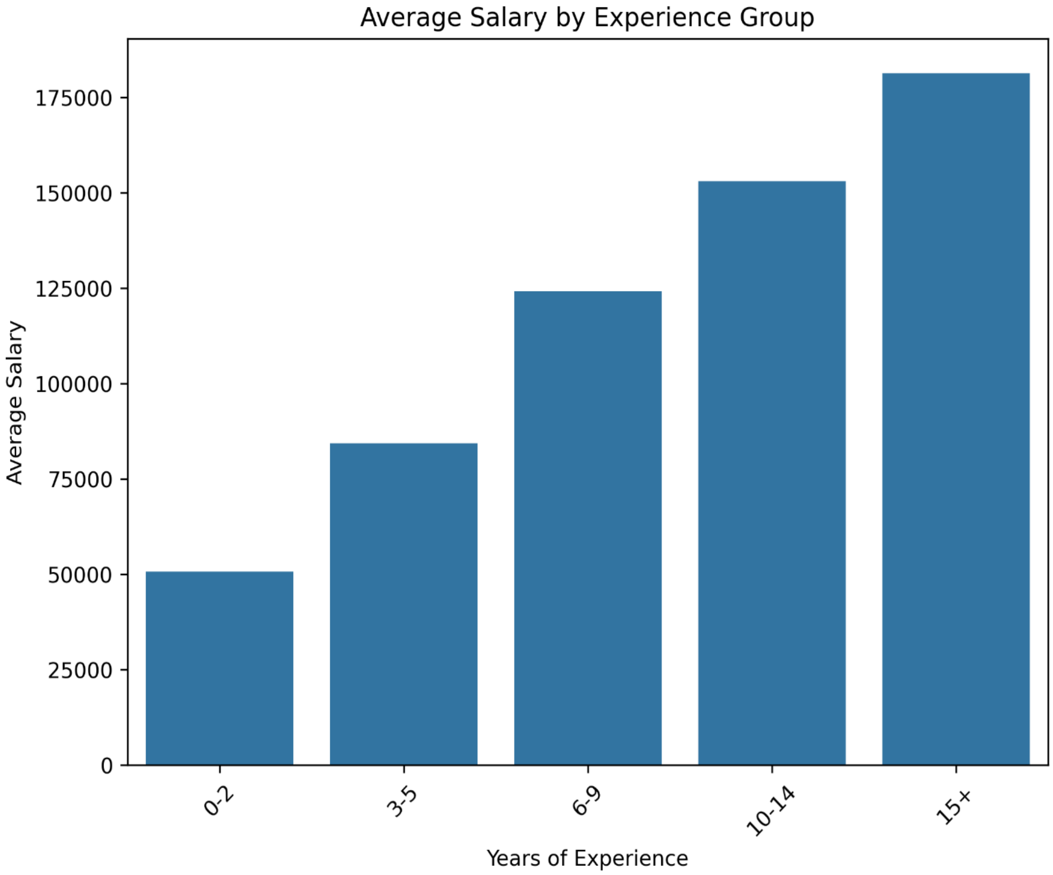


Salary Statistics by Education Level:

```
                        mean          median        std           count

 Education Level

 PhD                   165651.457999  170000.0     34339.751664    1369

 Master's              130070.273803  130000.0     40640.287946    1859

 Bachelor's            95110.861533   80000.0      44061.370318    3026

High School            34415.612385   30000.0      16563.414595    436
```

Salary by Year of Experience

The experience groups analysis of salary shows a clear increasing trend and reveals that salary rises with a function of years of work experience. People with 0–2 years of experience earn about an average salary of 50,783 and those with 3–5 years earn approximately 84,408, showing a significant jump in initial career growth. The trend led by the individuals with 6–9 years of experience, on average earning 124,242, followed by those with 10–14 years with an average earning of around 152,987. The individuals with 15+ years of experience is the category with the highest salaries, averaging 181,280 and median 185,000. Again, the consistent rise proves that the strong positive correlation between experience and salary, further confirming the significance of long-term career development and experience-accumulation in increasing earning potential.

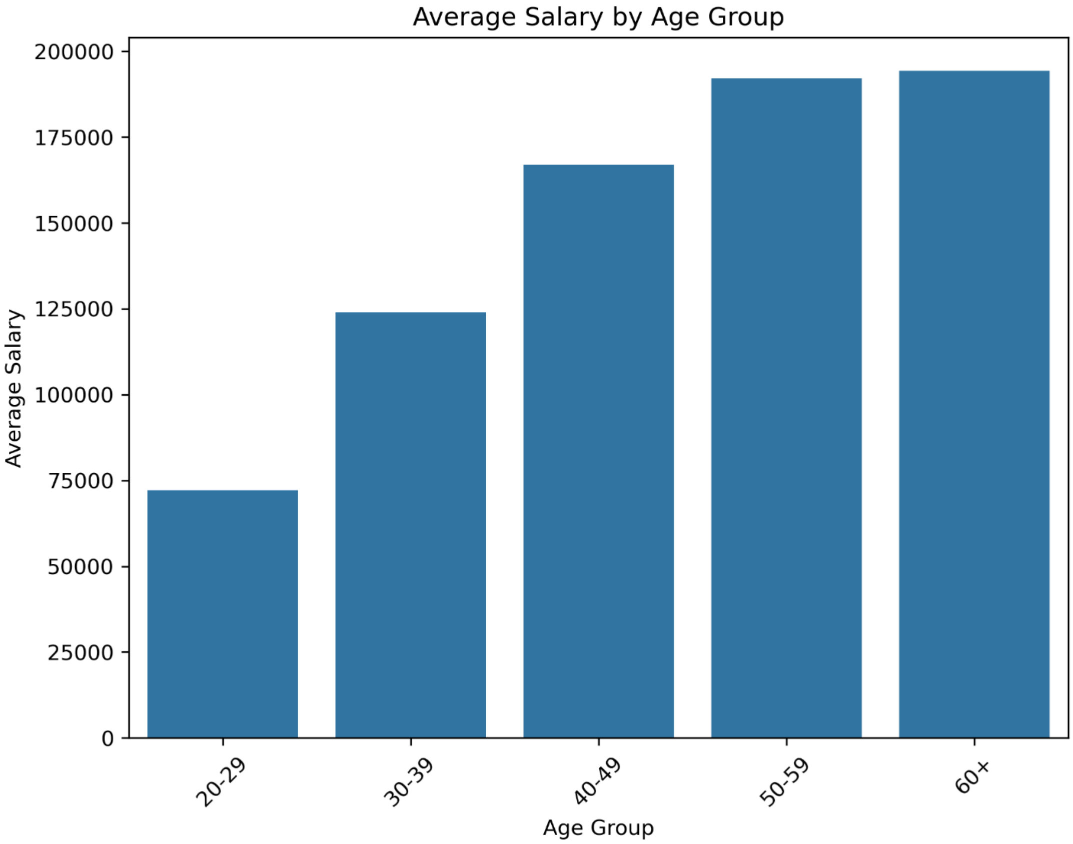


Salary Statistics by Experience Group:

```
                      mean          median        std            count
Experience Group
  15+             181279.810241   185000.0    22274.313123     996
  10-14           152986.992891   160000.0    28061.282170     1266
  6-9             124242.425266   120000.0    33665.791437     1599
  3-5              84407.885752    75000.0    31347.206843     1523
  0-2              50782.655199    50000.0    22754.308240     1279
```

Salary by Age

There is a clear rise in average salary with rising age, indicating that salary generally rise with seniority and experience. The best-paid categories are age above 60s, as well as the 50–59 years old group, with mean wages of around 194,222 and 192,099 respectively. This would be indicative of career experience and higher job titles development, as well as likely more years in the industry. The standard deviation is relatively low in the age 60 plus group (8,251), showing less salary variation, maybe due to retirement adjustments or equalization of earnings. In contrast, age range 50–59 shows greater variation (17,223), which may be due to varying types of jobs or varying seniority levels amongst this range. These results indicate that age (as a proxy measure for experience) is a strong predictor of earnings potential, though growth appears to taper off slightly in the later stages of a career.



Salary Statistics by Age Group:

```
                  mean           median      std            count

Age Group

    60+           194221.833333   200000.0    8251.159046    12

  50-59           192098.559846   191510.0   17223.186281    259

  40-49           166899.199484   170000.0   27131.802931    1163

  30-39           123985.392185   120000.0   42143.352246    2815

  20-29           72175.879558     60000.0    36524.304004    2441
```

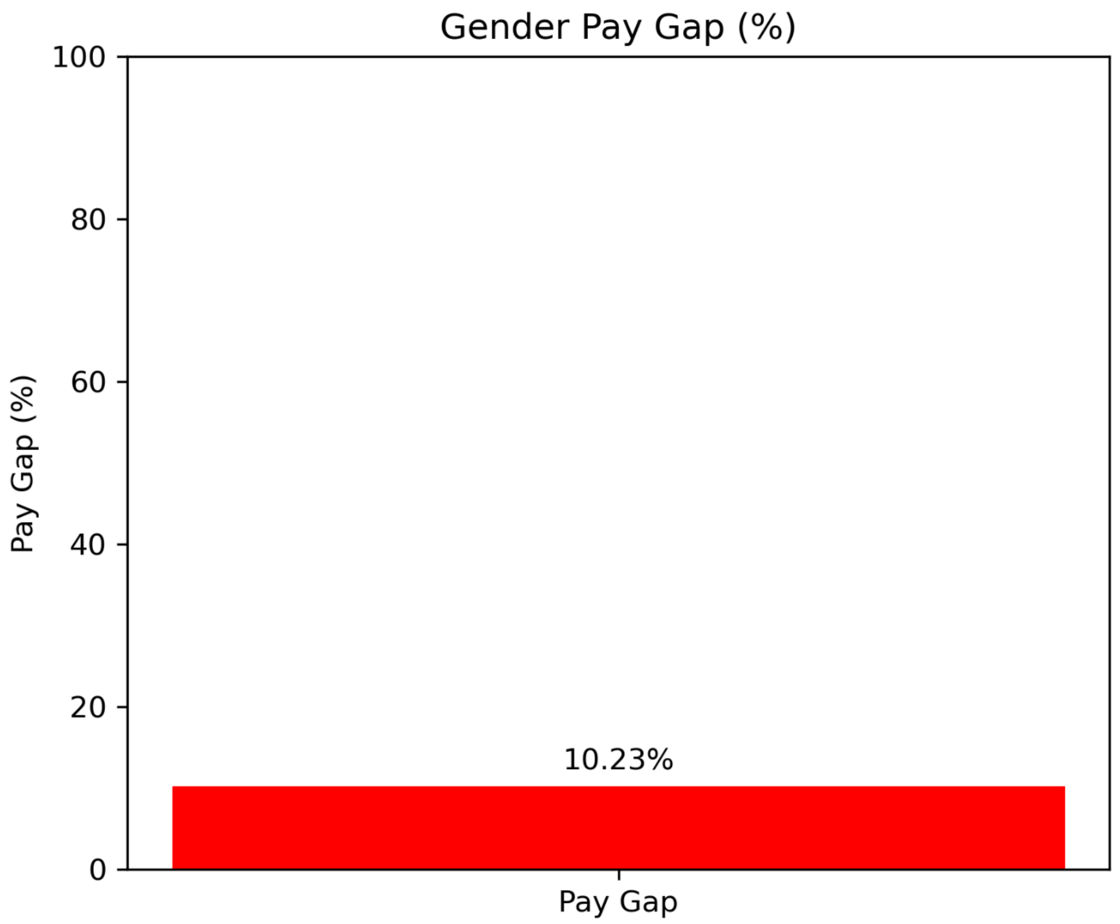
Salary by Job Title

The analysis reflects significant variations in salaries between job titles, and it points out distinct earning tiers within the workforce. At the top of the pay scale are executive positions, and their highest paid are the CEO, Chief Technology Officer, and Chief Data Officer, receiving the highest salaries (averaging 250,000 and 220,000 respectively). Mid-range salaries (centered around the median of 220,000 respectively). Mid-range salaries (centered around the median of 95,000) includes positions like Digital Marketing Manager, Financial Advisor, and Product Marketing Manager, offering consistent, and competitive salaries. On the lower side are administrative and entry-level jobs such as Junior Business Operations Analyst and Receptionist that recorded the lowest reported salary (averaging 17,675 and 17,675 and 25,000 respectively). Notably, executive and "chief"-level jobs dominate the high-paid category, reinforcing the leadership premium. The database features 193 unique job titles, with Software Engineer appearing most frequently (518 times), suggesting its top popularity in the job market. These findings emphasize the influence of job role and experience on pay variations, which are helpful for employers who create pay bands and working professionals who evaluate career opportunities.

Gender gap analysis

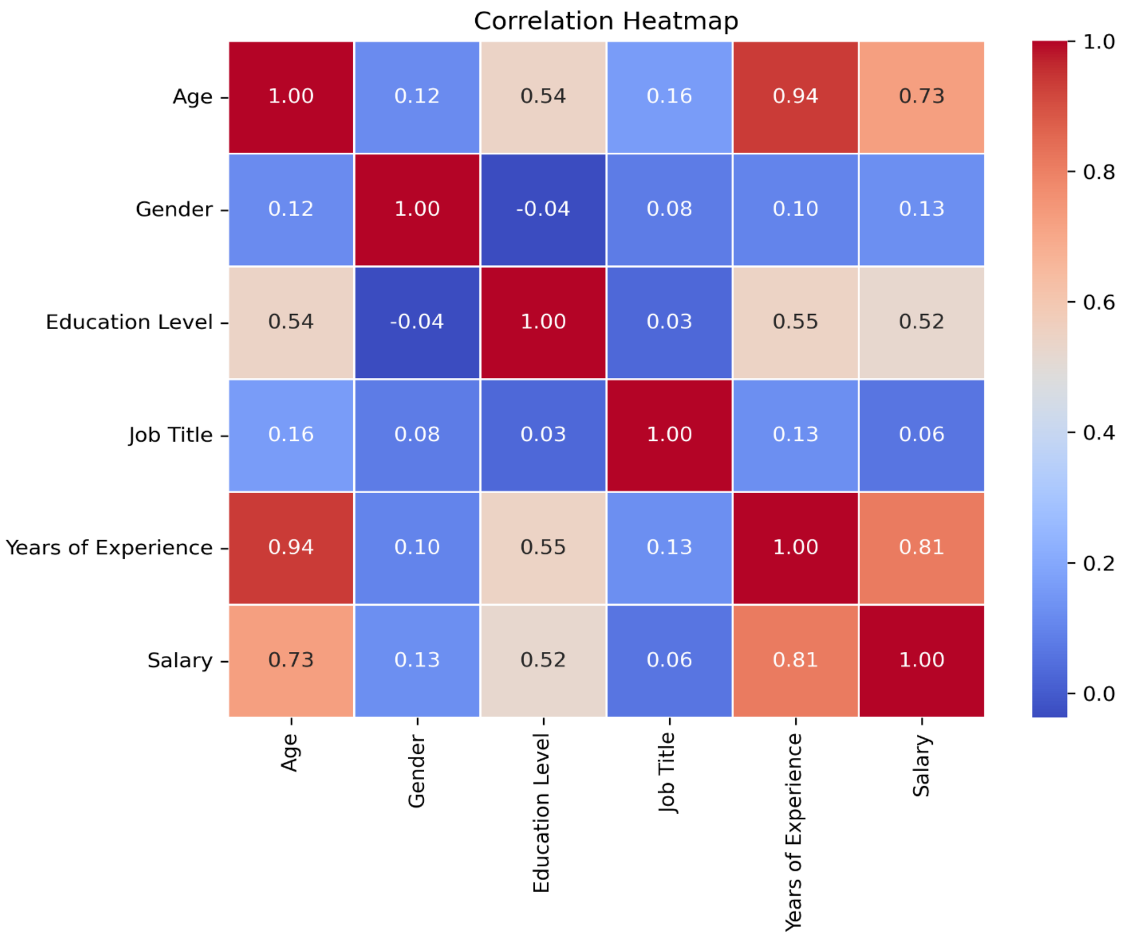
The analysis reveals a statistically significant gender pay gap. The t-test value of 9.482 and p-value of less than 0.00001 enable us to reject the null hypothesis that there is no difference in average salaries for gender. The graph shows that females earn approximately 10.23% lower than males, on average.

This finding identifies an important gap in earnings between genders in the sample that must be investigated further into possible causes such as variation in job titles, experience levels, or potential pay practice bias. Filling such gaps is of utmost importance to the cause of creating fairness and equality in the workplace.



T-test Statistic: 9.482, P-value: 0.00000

Significant gender pay gap exists (p < 0.05)



The correlation heatmap provides valuable insights on the nature of the relationship between the key variables that determine salary prediction. Years of Experience is the most significant variable, with a very strong positive relationship with salary (0.81), as hypothesized that individuals with greater experience will have higher salary. Age is also highly correlated with pay (0.73), though its high correlation with experience (0.94) which raises multicollinearity as an issue if both are to be included in linear models, producing biased estimates of their individual effects. Education Level has moderate correlated with salary (0.52), as expected, in reaffirming education qualifications do contribute substantially to earnings but is a lesser one than the professional experience. Conversely, Gender and Job Title only have a slight or minimal relation to the salary (0.13 and 0.06, respectively), implying that they are not strong standalone predictors in this data set. That being said, their relevance may come when considered in interaction with other variables. Most interestingly, Gender also shows close to zero or negatively directed correlations with predictors like job title and education level, indicating possible imbalance or reduce effect of gender in income determination within this specific dataset. These findings reinforce the importance of placing emphasis on experience and education while practicing caution against redundant predictors in salary modeling.

Data Standardization

Data standardization is also a preprocessing technique used to transform features to have a mean of zero and a standard deviation of one. Standardization is commonly used when the features have different units or ranges. Standardization ensures that each feature contributes equally to the analysis, helping improve the performance and convergence speed of the model.

The standardization formula is as follows

$$Z = \frac{x - \mu}{\delta}$$

Where:

z: Standardized Value.

x: The original value of the feature.

μ: The mean value of the feature.

σ: The Standard Deviation of the feature.

3.3.2 Subject or Participants of the Project

The project used a structured dataset composed of 6,704 entries sourced from multiple publicly available platforms. These entries represent anonymized salary data points with features including age, gender, education level, job role, and years of experience. No human subjects were directly involved in the data collection.

3.3.3 Data Collection Approaches and Strategies

Data was aggregated from diverse sources including online job portals, surveys, and open data repositories. The dataset was compiled by Mohith Sai Ram Reddy and collaborators and published on Kaggle, ensuring that it met open-source standards for accessibility and use in academic projects.

3.3.3.1 Advantages, Limitations, and Ethical Issues

- **Advantages**:

  One of the key advantages of conducting a predictive salary analysis is the depth and diversity of the dataset upon which it is based. The data encompasses broad range of job titles, qualifications, and year of experience, enabling the model to learn from a wide range of patterns and relationships. This diversity ensures that the model is not biased towards prediction for a specific group, making its prediction more equitable and reliable among different segments of the workforce. Also, the data set includes data from diverse sources, further enhancing the model's generalizability. Through analysis of trends and drivers of salary in different sectors, the prediction model is strengthened and adaptable to generate precise salary prediction. This broad coverage range not only improves model performance but also increases the value of the analysis, making it a powerful tool for job seekers, employers, and policy-makers who wish to understand and predict salary movements.

  - The data was rich and diverse, capturing different roles, educational backgrounds, and experience levels.
  - It covered multiple industries, allowing the model to generalize better.

- **Limitations**:

This project is limited by the dataset used, which will not necessarily reflect all of the potential salary-influencing variables (e.g., company-specific policy or general economic conditions). Additionally, while the models selected offer valuable insights, some techniques, like neural networks, may lack interpretability. Also, missing data and potential multicollinearity of the predictors can limit the models' performance. The generalizability of the results is also limited by the scope of the study in the dataset used.

  - Being secondary data, it may carry biases from source platforms (e.g., overrepresentation of certain job sectors).
  - Missing or imbalanced demographic categories (e.g., skewed gender distribution).

- **Ethical Considerations**:

Processing to salary, it requires careful attention to ethical concerns to ensure fairness, privacy, and transparency in the analytical process. Privacy and confidentiality are prioritized by anonymizing any personally identifiable information (PII) to protect individuals' privacy so that the dataset is free of sensitive information that can be used for identification of users. To address bias and fairness, a thorough analysis is carried out in order to detect any potential biases in salary predictions. If biases related to gender or other demographic variables are identified, the corrections in the form of removing biased features or applying fairness algorithms are implemented to ensure an unbiased and ethical modeling process. In addition, there is a presumption that informed consent has been achieved in data collection since the dataset is publicly accessible on Kaggle. Ethical considerations of data collection and usage are carefully examined to ensure moral compliance. Lastly, transparency is maintained by documenting the methodology, assumptions, and analysis findings in depth. Model limitation transparency and potential consequences ensure that stakeholders understand the findings and their significance.

  - Data was anonymized, reducing privacy risks.

- The use of publicly available data adheres to ethical standards for non-invasive research.

3.3.4 Data Analysis Approaches and Software

The data analysis and modeling for the predictive salary analysis project were done using Python, benefiting from a robust collection of libraries well-tailored for data science and machine learning. **Pandas** and **NumPy** were fundamental in the initial steps of data preparation and provided robust data structures and functions for efficient data cleaning, transformation, and manipulation. Such tools made it easy to handle missing values, categorical encoding, and raw data transformation into a structured format suitable for modeling.

Scikit-learn was heavily used in model development and evaluation. **Scikit-learn** was used for large-scale preprocessing tasks such as scaling numerical attributes along with splitting data into training and test sets. **Scikit-learn** was also the site used to train various machine learning models—including linear regression, random forest, and gradient boosting—and evaluate their performance using cross-validation techniques and error metrics like RMSE and R-squared.

To gain insights and an in-depth understanding of the data, **Seaborn** and **Matplotlib** were used for data visualization. Using these libraries, creation of detailed and informative plots such as correlation heatmaps, histograms, box plots, and scatter plots were generated to uncover relationships between variables and to detect outliers or trends.

3.3.4.1 Coding Method, Analysis of Interviews/Recordings

No interviews or recordings were used. All analysis was conducted through code written in Jupyter Notebooks. The methodology focused on predictive coding, structured around modular and reusable scripts for each phase: preprocessing, visualization, modeling, and evaluation.

3.3.4.2 Statistical Analysis

Descriptive statistics were used as a foundational step for summarizing and interpreting the key features of the dataset. This included measures of central tendency— such as mean, median, and mode—and measures of dispersion like standard deviation and range. These statistical numbers helped decide the range and average values in characteristics such as age, years of experience, and salary, where the mean salary provided an average benchmark, and the standard deviation measured how spread out salaries for different positions and education levels were. This stage also involved deciding outliers and skewness, which are crucial in deciding appropriate modeling techniques and transformations.

Correlation analysis was conducted to quantify the strength and direction of the linear relationships between numerical variables. It helped us in identifying which variables had a significant impact on salary. It is typically represented by a value between -1 and +1, where +1 denotes a perfect positive correlation (both variables increase simultaneously), -1 indicates a perfect negative correlation (one variable increases while the other declines), and 0 suggests no correlation (no predictable relationship between the variables).

The formula is as follows:

$$r = \frac{n\sum(xy) - \sum x \sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

LR was one of the primary models used to predict salary based on features like experience, education level, and job title. The model assumes a linear relationship between the dependent variable (salary) and one or more independent variables. The equation for a simple linear regression is:

$$y^{\wedge} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Where:

y_hat is the predicted salary,

beta_0 is the intercept,

beta_i are the coefficients for each predictor xi.

To model more complex, nonlinear relationships in the data, GB, RF, and NN were applied. GB is a very powerful ensemble method that builds a sequence of decision trees, where each tree is trained to correct the residual errors of its predecessors. It minimizes a loss function—typically MAE for regression—by using gradient-based updates, making it well-suited to learn subtle patterns in the data and improve prediction accuracy over time.

RF, another ensemble learning method, constructs multiple decision trees separately and then aggregates their outputs to produce final predictions. The method reduces overfitting and enhances overall generalization by introducing randomness in data sampling and feature selection. It is appropriate for modeling nonlinear interactions between features and it provides stable performance on various kinds of data types.

NN model proved very useful in identifying slight patterns in high-dimensional data. A basic feed forward neural network is a simple one that consists of input layers, hidden layers, and an output layer. Every neuron computes a weighted sum of inputs and subsequent application of an activation function to introduce nonlinearity. Back propagation and gradient descent are used to minimize a loss function such as MAE for training the network. NN function is best and effective when the relationships among variables are highly nonlinear and complex.

To accommodate more complex model, nonlinear relationships in the data, RF and NN were applied. RF is an ensemble learning method that generates a large number of decision trees and merges their predictions to have improved predictive accuracy and reduced overfitting. RF identifies interactions and non-linear effects by repeatedly splitting the data according to feature values.

To evaluate model performance, three primary metrics were used:

- **RMSE**: Measures the average magnitude of error in predictions, giving higher weight to larger errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **MAE**: Measures the average absolute difference between actual and predicted values.

$$MAE = \frac{1}{n}\sum_{1=1}^{n}|y_i - \hat{y}_i|$$

- **R^2** : Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

- These metrics provided a comprehensive view of how well each model performed, allowing for a fair comparison between linear models and more complex approaches like RF and NN.

3.4 Results, Findings, and Discussions

As part of the exploratory data analysis, a correlation heatmap was generated to explore the linear relationships between the most significant numerical variables, such as Age, Years of Experience, Education Level, Job Title, and Salary. The visual tool helped in assessing whether there was any multicollinearity among predictors. From the heatmap, one of the interesting findings was the exceptionally high correlation (0.94) between Age and Years of Experience. This strong correlation was a clear sign of multicollinearity that could negatively impact the performance and interpretability of machine learning models, particularly linear regression, by inflating the variance of estimates of coefficients and distorting the contribution of individual predictors.

To mitigate this issue and improve the reliability of models, we decided to leave out the 'Age' variable from the prediction model. We made this choice because Years of Experience had a more direct and higher correlation with Salary (0.81) compared to Age (0.73), and it has greater practical relevance in the context of determining salary. Years of Experience is more a measure of professional maturity and accumulation of skills, which employers are likely to consider when determining compensation, and thus a meaningful and greater predictor for our analysis.

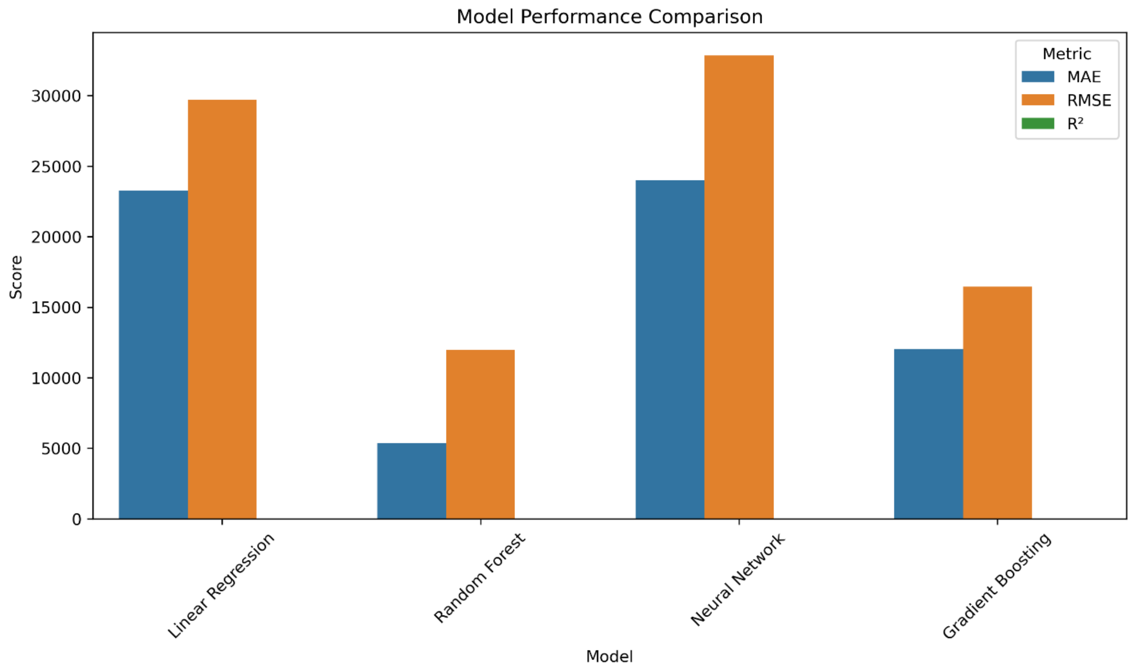
In addition to the heatmap results, we also confirmed this decision using Variance Inflation Factor (VIF) analysis, quantitatively confirming multicollinearity. The VIF greater than 5 between Experience and Age suggested scaling down the variables. By removing the redundant feature (Age), we enhanced model stability as well as avoided the risk of overfitting. These models were then retrained on the cleaned feature set to obtain improved or comparable performance on measures of evaluation such as RMSE, MAE, and R^2. This strategic fine-tuning made the prediction more accurate while the interpretability and generalizability of the model were maintained.

The predictive analysis of salary using various machine learning algorithms yielded insightful results, particularly in identifying the most influential variables through correlation analysis. After training and validating four models, the performance metrics (RMSE, MAE, and R^2) revealed distinct patterns:

Based on the model performance metrics—Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R-squared (R^2)—the Random Forest model clearly outperforms the others in predicting salary. It has the lowest MAE (5363.862741) and RMSE (11970.152908), indicating that its predictions are consistently closer to the actual values with less error. Furthermore, its R^2 score of 0.94 means it explains 94% of the variance in the salary data, making it the most accurate and reliable model among those evaluated.

In comparison, while Gradient Boosting also performed well, with an R^2 of 0.90, its error metrics were notably higher than Random Forest. Linear Regression and Neural Network models showed significantly larger errors and lower R^2 values, suggesting that they were less capable of capturing the complexity of the data, particularly any nonlinear relationships.

Therefore, Random Forest is the best fit for the given salary dataset, due to its superior ability to generalize and handle diverse feature interactions effectively.



Model Performance Comparison:

	MAE	RMSE	R²
Random Forest	5363.862741	11970.152908	0.948029
Gradient Boosting	12041.235483	16458.743371	0.901746
Linear Regression	23272.031490	29720.158419	0.679623
Neural Network	24014.423570	32833.633545	0.608982

Key findings:

Experience and job position were the strongest features in influencing the prediction of salary, highlighting their critical role in determining earning potential. Education level was also a factor in salary determination but was notably small influence relative to experience. Contrary to expectations, age had a less salient correlation with salary when adjusting for years of experience, which suggests that age as a predictive metric is not highly potent when isolated from professional experience. These findings highlight the importance of job-related characteristics over simple demographic variables in real-world salary prediction applications. They also validate the efficacy of ensemble approaches like RF, which are particularly capturing up on complex patterns and correlations among the variables and hence also well-suited for modeling multifaceted relations in salary data.

3.5 Metadata and Data Management Plans (DMPs)

3.5.1 Information Concerning How Data and Materials Are Collected, Organized, and Stored

The dataset used in this project were obtained from Kaggle. It was downloaded as a CSV file with labeled data for each observation. The data was collected from multiple sources including job posting platforms and surveys, and organized into structured columns with five main variables: Age, Experience, Education Level, Job Role, and Salary. For security and accessibility, the dataset was stored in a secure local environment and backed up through the use of GitHub and Google Drive. Jupyter Notebooks were employed for development to ensure version control and traceability throughout the project. Additionally, raw and processed data were kept separate, and all transformation steps were well documented to maintain reproducibility and transparency in the analysis.

3.5.2 Findability, Accessibility, Interoperability, and Reusability (FAIR) of the Project Data

The data used in the project conforms to the FAIR principles of Findability, Accessibility, Interoperability, and Reusability for it to be of use to researchers and developers. The dataset achieve **Findability** as the dataset is openly available in a public repository on Kaggle, a well-known platform for data science resources, and thus easily accessible through search functions and community submissions. **Accessibility** is offered as the dataset is downloadable for free without restrictive permissions so that research and collaboration are open. The data is stored in a **CSV format**, which enhances **Interoperability** through easy integration with common data analysis tools such as Python, R, and Excel. The structured format and standardized variables also facilitate easy data manipulation on different platforms. **Reusability** is supported by the use of clear column naming conventions, proper documentation, and an open licensing framework, allowing other researchers to understand and make use of the dataset for different projects. To further ensure and support reproducibility, all Jupyter Notebooks and scripts used for processing include detailed comments and are reasonably structured in a way that other researchers can easily repeat or follow the process of analysis workflow and take ownership of the code. The compliance with FAIR principles in this way makes the dataset transformed into a useful, open, and reusable asset for the broader data science community.

3.6 Implications, Recommendations, and Applications

The findings from this study underscore the significant influence of **job title** and **year of experience** in predicting salaries, suggesting that these factors may be more overwhelming than age and education in modern compensation frameworks. This shift indicates a potential change in how organizations value skills and tenure compared to traditional demographic factors. To employers, these results provide a data-backed foundation for structuring equitable pay system, and job seekers can utilize them to better negotiate market-level salaries aligned with the industry standards. However, the stronger correlation between year of experience and salary also raises questions of whether education is undercompensated or age-based discrimination is present, warranting further investigation into fair pay practices in evolving job markets.

To promote fair and competitive compensation strategies, organizations must adopt **data-driven salary assessment tools** that minimize the influence of subjective biases and unite pay systems with measurable factors like role requirements and level of experience. Additionally, future datasets must also incorporate a greater **demographic diversity**—including gender, ethnicity, and geographic location—to evaluate and correct any potential differences in salary distributions. Policymakers and HR professionals could use such enriched data to develop more balanced wage frameworks. Expanding data collection to include non-traditional education paths (e.g., certifications, bootcamps) would similarly increase the relevance of salary prediction models in today’s dynamic workforce.

The predictive models developed in this study have practical applications in **HR analysis** and **career guidance**. Companies can implement them in employees' management systems to offer automated salary recommendations for new hires or promotions, with a promise of fairness and equity. Educational institutions and career guidance counsellors can use the models to advise students on well-paying careers according to salary expectations as a function of job titles and levels of study. Furthermore, recruitment platforms can incorporate similar algorithms to provide instant salary estimates, empowering applicants with data-driven negotiation tools. By incorporating these insights within education and workforce planning systems, stakeholders can bridge the gaps between skill development, career choice, and equitable compensation.

References

Bao, Q. (2024, November). Enhancing salary prediction accuracy with advanced machine learning models. *Applied and Computational Engineering*, 20-150. doi:10.54254/2755-2721/96/20241185

Biswas, A., Saha, A., Sarkhel, A., Sengupta, A., & Koner, D. (2023). A comparative study for salary prediction based on different models of machine learning. *Department of Computer Application Institute of Engineering & Management*, 6-22.

Chen, J., Mao, S., & Yuan, Q. (2021). Salary prediction using random forest with fundamental features. *Third International Conference on Electronics and Communication; Network and Computer Technology, 12167*, 2-8. doi: 10.1117/12.2628520

Das, S., & Barik, R. (2020). Salary prediction using regression techniques. *International Conference on Industry Interactive Innovations in Science, Engineering and Technology*, 1-5.

Devi, A., & Neelambika, P. (2023). Employee salary prediction system using machine learning. *International Journal of Engineering Science and Advanced Technology, 23*(9), 1-4.

Dsouza, O. D. (2024). Predictive model of employee attrition based on stacking ensemble learning. *International Journal of All Research Education and Scientific Methods, 12*(1), 2-13

Erpapalemlah, M. A. (2023). Predicting salary in the field of data science of gradient descent Algorithm.4-18

Feng, Z., Liu, Z., & Yin, Y. (2023). Comparison of deep-learning and conventional machine learning algorithms for salary prediction. *Applied and Computational Engineering, 6*, 643– 651

Gurpreet, M., Bistania, K., & Mahendroo, S. (2023, May). Salary Prediction System using Machine Learning. *viii*(05).

Hastie, T., Tibshirani, R., & Friedman, J. (2009*).* The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.).

Hussain, J., & Ahmad, a. (2024, April 2). Employee salary prediction in HRMS Using regression models. *Journal of Innovative Computing and Emerging Technologies*, 1-14.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R (2nd ed.).

Jiang, W. (2024). The investigation and prediction for salary trends in the data science industry. Applied and Computational Engineering*, 50*, 8–14.

Li, S., Liu, Z., & Zhang, J. (2024). Unlocking the potential of LSTM for accurate salary prediction with MLE, Jeffreys prior, and advanced risk functions. *BioMed Research International*. https://doi.org/10.1155/2024/8845563

Lothe DM, Tiwari P, Patil N, Patil S, Patil V (2022) Salary prediction using machine learning. Int J Adv Sci Res Eng Trends. https://doi.org/10.51319/2456-0774.2021.5.0047

Matbouli, Y. T., & Alghamdi, S. M. (2022). Statistical machine learning regression models for Salary prediction featuring economy wide activities and occupations. *13*(10), 495.

Meng, Q., Zhu, H., Xiao, K., & Xiong, H. (2018). Intelligent Salary Benchmarking for Talent Recruitment: A Holistic Matrix Factorization Approach. *IEEE International Conference on Data Mining (ICDM)* (337-346).https://doi.org/10.1109/ICDM.2018.00047

Monteiro, C. (2023). Salary estimation using random forest based on economic indicators. 116-129.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,& Duchesnay,É. (2011). *Scikit-learn: Machine learning in Python*. *Journal of Machine Learning Research, 12*, 2825–2830.

Ren, Y. M., Luo, J., Chen, S., & Abdullah, F. (2022). A Tutorial Review of Neural Network Modeling Approaches for Model Predictive Control. 1-71.

Saraswathi, M., Akhila, J., & Sireesha, K. (2023). Predictive Insights: Using Machine Learning to Determine Your Future Salary. *International Journal of Soft Computing and Engineering, 13*(2), 1–7.

Yang, S. (2023). Automated employee salary prediction algorithm based on machine learning. *International Conference on Computer Vision, Application, and Algorithm, 12613*, 2-6. doi: 10.1117/12.2673738

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3–28.

References

Aitchison, J. (2021). Statistical Prediction Analysis.

Bao, Q. (2024, November). Enhancing Salary Prediction Accuracy with Advanced Machine Learning Models. *Applied and computational Engineering*, 20-150. doi:10.54254/2755-2721/96/20241185

Bjerre, L. (2016, january 25). *Machine Learning Models vs. Statistical Models: Choosing the Right Approach for Your Predictive Analytics*. Retrieved from https://infomineo.com/data-analytics/machine-learning-models-vs-statistical-models/

Cam Thao Trang Hyunh. (2021, october). Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance. Retrieved from https://www.researchgate.net/publication/355835606_Using_Decision_Trees_and_Random_Forest_Algorithms_to_Predict_and_Determine_Factors_Contributing_to_First-Year_University_Students'_Learning_Performance

Changanl, V. (2021). A Comprehensive Guide to K-Fold Cross Validation. Retrieved from https://www.datacamp.com/tutorial/k-fold-cross-validation

Chen, J., Mao, S., & Yuan, Q. (2021). Salary prediction using random forest with fundamental features. *Third International Conference on Electronics and Communication; Network and Computer Technology, 12167*, 2-8. doi: 10.1117/12.2628520

Doohee, C. (2023, april 1). Predictive model of employee attrition based on stacking ensemble learning. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S095741742202382X

Erpapalemlah, M. A. (2023, april 1). Predicting Salary in the field of data science of Gradient Descent Algorithm. *INFORMATICS ENGINFORMATICS ENGINEERING DEPARTMENT*. Retrieved from file:///C:/Users/USER/Downloads/salary%20prediction%20using%20gradient%20boosting.pdf

Hasan, S. A. (2024). *Random Forest Algorithm Overview*. Retrieved from https://www.researchgate.net/publication/382419308_Random_Forest_Algorithm_Overview

Hyun, C. T. (2021, October). Using Decision Trees and Random Forest Algorithms to Predict and Determine Factors Contributing to First-Year University Students' Learning Performance. Retrieved from https://www.researchgate.net/publication/355835606_Using_Decision_Trees_and_Random_Forest_Algorithms_to_Predict_and_Determine_Factors_Contributing_to_First-Year_University_Students'_Learning_Performance

Matbouli, Y. T. (2022, september). Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations. Retrieved from https://www.mdpi.com/2078-2489/13/10/495

Monteiro, C. (2023). Salary estimation using random forest based on economic indicators. 116-129. Retrieved from https://www.wcdanm2024.uevora.pt/wp-content/uploads/2024/10/BoA_WCDANM_2024.pdf#page=116

Ms. Gurpreet, K. B. (2023, May). Salary Prediction System using Machine Learning. *viii*(05). Retrieved from https://scispace.com/pdf/salary-prediction-system-using-machine-learning-2xz2thyn.pdf

Rubery, J. (2003, November). Pay equity, minimum wage and equality at work. *International Labour Office*, 1-80. Retrieved from https://www.ilo.org/sites/default/files/wcmsp5/groups/public/@ed_norm/@declaration/documents/publication/wcms_decl_wp_20_en.pdf

Vinod Chuganl. (2021, june). A Comprehensive Guide to K-Fold Cross Validation. Retrieved from https://www.datacamp.com/tutorial/k-fold-cross-validation

Zahidi, S. (2024). *Future of Jobs Report.* World economic forum. Retrieved from https://reports.weforum.org/docs/WEF_Future_of_Jobs_Report_2025.pdf