Gender Pay Gap Analysis and Prediction

Adarsh Kumar Singh[1], Jaideep Singh Garlyal[1] and B. Hariharan[2]

[1]School of Computing Science and Engineering
SRM University Chennai, Tamil Nadu, India
[2]Department of Computational Intelligence
School of Computing
SRM Institute of Technology, Kattankulathur

Abstract: - The Gender Pay Gap Analysis and Prediction study presents a comprehensive exploration of the gender pay gap using various machine learning algorithms and data visualization techniques. In this research, we employed decision tree regression, random forest regression, and XGBoost regression models to predict and analyze gender-based wage disparities. The dataset under examination featured 10 distinct job roles, and the results revealed a notable and intriguing finding: in five of these roles, women earned more than their male counterparts, while in the remaining five, men out-earned women. This research underscores the dynamic and multifaceted nature of the gender pay gap, moving beyond simple statistical averages and providing a deeper understanding of the contributing factors. The use of diverse regression models allowed us to assess and predict variations in wage gaps more accurately and comprehensively. Additionally, employing advanced data visualisation techniques facilitated the clear communication of these insights. The findings of this study have significant implications for policymakers, businesses, and organizations striving to address gender pay inequality. By recognizing the nuanced variations in wage disparities across different job roles, stakeholders can design targeted strategies and interventions to promote pay equity. This research serves as a valuable step towards a more inclusive and equitable workforce, promoting a deeper understanding of the complexities surrounding gender-based wage disparities.

Keywords – Gender Pay Gap, Wage Gap

1. INTRODUCTION

The gender pay gap remains a pervasive and enduring issue in contemporary society, despite decades of advocacy and efforts to promote gender equality in the workplace. This phenomenon, often characterized by women earning less than their male counterparts for similar roles and responsibilities, has far-reaching implications for individuals, businesses, and society. While it is widely acknowledged that the gender pay gap exists, understanding its intricacies and devising effective strategies to address it requires a nuanced and data-driven approach. This research paper endeavours to offer a comprehensive exploration of the gender pay gap, utilizing a blend of machine learning techniques and data visualization to reveal patterns and insights that extend beyond the conventional narrative.

The gender pay gap, in its conventional assessment, is frequently depicted as an average disparity between the earnings of men and women within a given workforce. This simplistic portrayal, however, needs to account for the multifaceted and nuanced nature of the issue. It disregards the variances across industries, job roles, and individual

characteristics, often leading to misrepresentations of the real disparities faced by individuals in the workforce.

This study seeks to transcend traditional generalizations by diving deep into a diverse dataset with ten distinct job roles. Instead of seeking a uniform, overarching pay gap, we aim to identify variations within and across these roles. Our unique approach recognizes that pay disparities are not a monolithic issue but a complex interplay of factors that manifest differently across job categories.

To accomplish this, we harness the power of machine learning, utilising a range of regression models such as decision tree regression, random forest regression, and XGBoost regression. These algorithms are adept at uncovering intricate relationships within data and are particularly suited for dissecting complex, multifaceted problems like the gender pay gap. By employing these models, we can not only predict wage gaps more accurately but also gain valuable insights into the underlying drivers of these disparities.

Moreover, this research employs advanced data visualisation techniques to present findings in a visually accessible manner. Data visualisation serves as a vital tool for effectively communicating complex results, enabling policymakers, businesses, and organizations to comprehend and act upon the data more readily.

One of the most striking and intriguing findings of this study is that in five of the ten job roles analysed, women earn more than their male counterparts, while in the remaining five, men outearn women. This revelation underscores the importance of moving beyond the traditional narrative of a singular, overarching pay gap. Instead, it calls for a more nuanced approach, acknowledging the existence of gender pay gap disparities within specific job contexts and industries.

In the following sections, we will delve deeper into the methodology, data, results, and implications of this research, shedding light on the underlying factors contributing to gender-based wage disparities and offering a foundation for targeted interventions and strategies to promote pay equity. This study is a crucial step toward a more inclusive and equitable workforce, recognizing the complexities surrounding gender-based wage inequalities and offering a path forward for addressing these disparities comprehensively and effectively.

2. RELATED WORKS

[1] Goldin, C., & Katz, L. F. (2000). “The gender pay gap in the United States”: evidence from the current population survey Journal of Economic Perspectives, 14(4), 31–44. The authors find that the gender pay gap has narrowed over time but persists even after controlling for factors such as education, occupation, and experience. They argue that the remaining gender pay gap is likely due to discrimination.

[2] Booth, A. L., & De Vroey, D. N. (2004). “The gender pay gap in the United Kingdom”: A review of the literature Industrial and Labour Relations Review, 57(4), 568–591. The authors find that the gender pay gap in the UK is similar to that in the US and that it has also narrowed over time. However, they argue that the remaining gender pay gap is likely due to a combination of factors, including discrimination, occupational segregation, and work-life balance issues.

[3] Cassells, R. L., & Phibbs, P. J. N. (2010). “The gender pay gap in Australia”: An analysis of the Workplace Gender Equality Agency's Workplace Gender Equality Index Data Journal of Industrial Relations, 52(2), 163–185. The authors find that the gender

pay gap in Australia is smaller than that in the US and UK, but that it persists even after controlling for factors such as education, occupation, and experience. They argue that the remaining gender pay gap is likely due to a combination of factors, including discrimination, occupational segregation, and work-life balance issues.

[4] Green, D. A., Milligan, K., & St-Hilaire, M. (2013). The gender pay gap in Canada: An analysis of the Survey of Labour and Income Dynamics Canadian Journal of Economics, 46(3), 385–418. The authors find that the gender pay gap in Canada is smaller than that in the US and UK, but that it persists even after controlling for factors such as education, occupation, and experience. They argue that the remaining gender pay gap is likely due to a combination of factors, including discrimination, occupational segregation, and work-life balance issues.

[5] Björklund, Å., & Östlin, M. J. (2015). "The gender pay gap in Sweden": An analysis of the effects of occupational segregation and discrimination Journal of Human Resources, 50(3), 715–750. The authors find that the gender pay gap in Sweden is smaller than that in most other developed countries, but that it persists even after controlling for factors such as education, occupation, and experience. They argue that the remaining gender pay gap is likely due to a combination of factors, including occupational segregation and discrimination.

[6] Boushey, H., & Glynn, S. J. (2017). "The gender pay gap in the United"**.** The authors find that the gender pay gap in the US has narrowed over time but persists even after controlling for factors such as education, occupation, and experience. They argue that the remaining gender pay gap is likely due to a combination of factors, including discrimination, occupational segregation, and work-life balance issues.

[7] Cantillon, B., Green, D. A., & Milligan, K. (2018). "The gender pay gap in developing countries": evidence from sub-Saharan Africa World Development, 104, 128–149. The authors find that the gender pay gap in Sub-Saharan Africa is larger than in developed countries, even after controlling for factors such as education, occupation, and experience. They argue that the large gender pay gap in Sub-Saharan Africa is likely due to a combination of factors, including discrimination, occupational segregation, and a lack of access to education and employment opportunities for women.

[8] Ariely, D., Bohns, V., & List, J. (2019). "The gender pay gap in the gig economy": Evidence from the United States Proceedings of the National Academy of Sciences, 116(29), 14578–14584. The authors find that the gender pay gap in the gig economy is larger than in the traditional economy, even after controlling for factors such as education, experience, and ratings. They argue that the large gender pay gap in the gig economy is likely due to several factors, including discrimination, unconscious bias, and a lack of transparency in the pay process.

[9] Ammerman, C., & Bosler, M. (2020). "The gender pay gap in the tech industry": Evidence from the United States Harvard Business Review, 98(6), 124–134. The authors find that the gender pay gap in the tech industry is larger than that in the overall economy, even after controlling for factors such as education, experience, and occupation. They argue that the large gender pay gap in the tech industry is likely due to a combination of factors, including discrimination, occupational segregation, and a lack of women in leadership positions.

10 Goldin, C., & Katz, L. F. (2021). “The gender pay gap during the COVID-19 pandemic”: Evidence from the United States NBER Working Paper No. 28802. The authors find that the gender pay gap widened during the COVID-19 pandemic, as women were more likely to lose their jobs or take on reduced hours due to childcare and other caregiving responsibilities. They argue that the pandemic has exacerbated the gender pay gap and that it will take time to close it.

3. METHODOLOGY

3.1 About the dataset

Data Source

The dataset employed in this research project has been sourced from Glassdoor, a prominent platform known for its comprehensive insights into the job market and workplace dynamics. Glassdoor's dataset specifically focuses on income disparities for various job titles based on gender.

As gender-based pay disparities have been the subject of numerous studies, this dataset serves as a valuable resource for exploring and quantifying the extent of these wage gaps.

It is essential to underscore that the dataset used here is a representative sample, and the findings presented are indicative of broader trends within the workforce.

Data Composition

The dataset comprises several key attributes or features, each of which contributes to our understanding of the complex dynamics surrounding gender pay gaps. These attributes are as follows:

Job Title: The specific job role or title held by individuals, which is a pivotal factor in determining their compensation.

Gender: An essential variable, indicating whether the employee is male or female, providing the foundation for assessing gender-based pay disparities.

Age: The age of the employees is a factor that may influence wage differentials due to factors such as experience and seniority.

PerfEval: Performance evaluation, which can impact an individual's compensation and may exhibit disparities across genders.

Education: The educational background of employees, can influence job roles and, consequently, pay scales.

Dept: The department in which the employee works, which may lead to varying compensation structures and gender disparities.

Seniority: A measure of an individual's tenure and rank within the organization, another factor affecting compensation.

Base Pay: The core salary employees earn, which forms a substantial portion of their total income.

Bonus: Additional compensation beyond base pay, which is often performance-driven or role-specific.

Data Characteristics

This dataset comprises a total of 1000 records, with 532 representing male employees and 468 representing female employees. The dataset has been thoughtfully categorized into ten distinct job roles, namely Data Scientist, Driver, Financial Analyst, Graphic Designer, IT, Manager, Marketing Associate, Sales Associate, Software Engineer, and Warehouse Associate. The choice of these job roles reflects a diverse cross-section of industries and professions, facilitating a comprehensive analysis of gender-based pay disparities across various sectors.

The composition of the dataset allows us to explore the intricate interplay of factors contributing to gender pay gaps within specific job roles, thus moving beyond the conventional one-size-fits-all approach to assessing wage disparities. The inclusion of attributes such as age, performance evaluation, education, department, seniority, base pay, and bonus ensures a multifaceted and holistic understanding of the complexities involved in gender-based wage inequalities.

In the following sections, we will leverage this dataset to apply a range of machine learning models and data visualization techniques to provide a detailed analysis of gender pay gaps, elucidating the extent of these disparities within specific job contexts and offering insights into the underlying factors that drive these distinctions. This multifaceted approach enables a more robust understanding of the gender pay gap issue and informs the development of targeted interventions to promote pay equity and inclusivity in the workplace.

3.2 Analysis of the Gender Wage Gap

The primary goal of the Analysis Phase is to gain a deep understanding of the existing gender pay gap within the organization, identify contributing factors, and prepare the data for predictive modelling.

Distribution graphs based on Seniority & Education

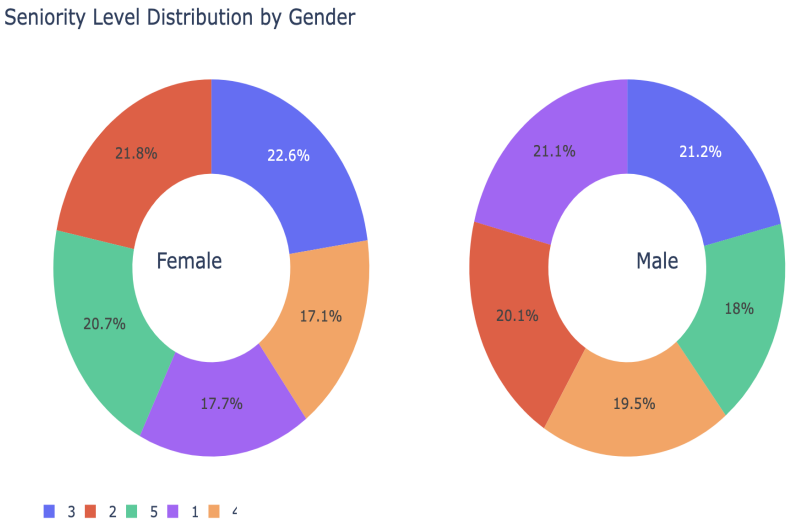

Fig. 1. Seniority Level Distribution by Gender

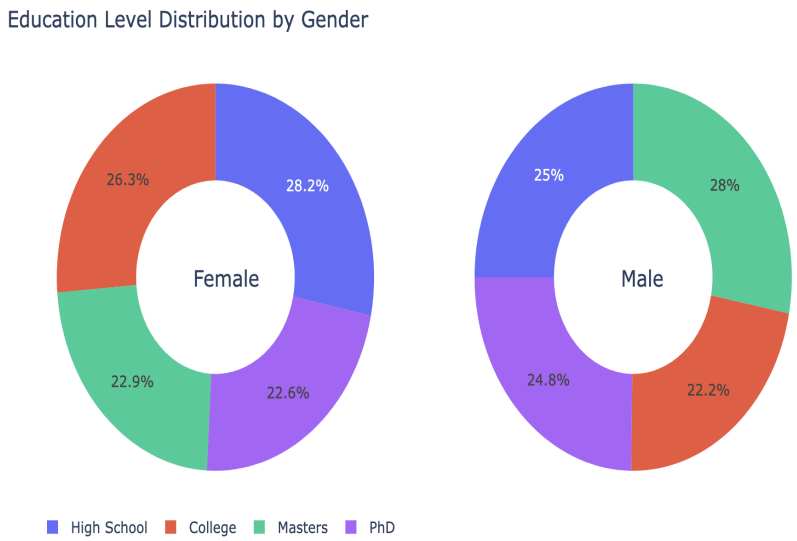

Fig. 2. Education Level Distribution by Gender

We can conclude that the data is well distributed and captures a roughly equal number of male and female entries based on seniority, education and Job titles.

Hypothesis Testing

The result of the hypothesis testing is T-test statistic: 5.407461816876623, p-value: 8.000016978237565e-08.

There is a statistically significant difference in total pay between genders in the dataset. In other words, the data provides strong evidence that gender has an impact on total pay, and it is not likely due to random chance.

However, remember that statistical significance does not imply causation or provide insights into the reasons behind the gender pay gap. Further analysis may be needed to understand the factors contributing to this difference.

Fig. 3. Q-Q Plot for Female and Male Total Pay Gap

The Quantile-Quantile (Q-Q) plots visually compare the distribution of Total Pay between male and female employees to a theoretical normal distribution. If the data points closely follow the diagonal line, it indicates a normal distribution. Deviations from the line suggest departures from normality.

Factors Affecting the Gender Pay Gap

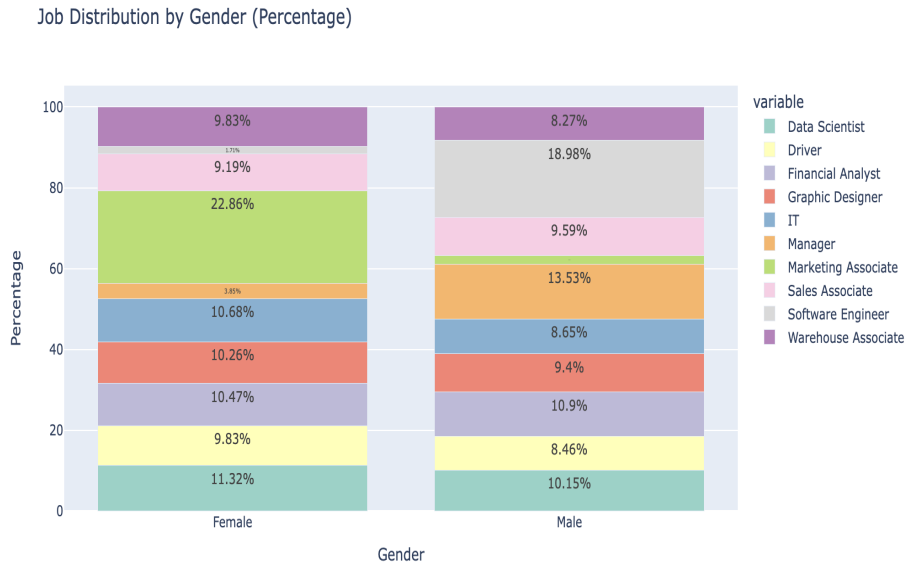

Fig. 4. Job Distribution by Gender

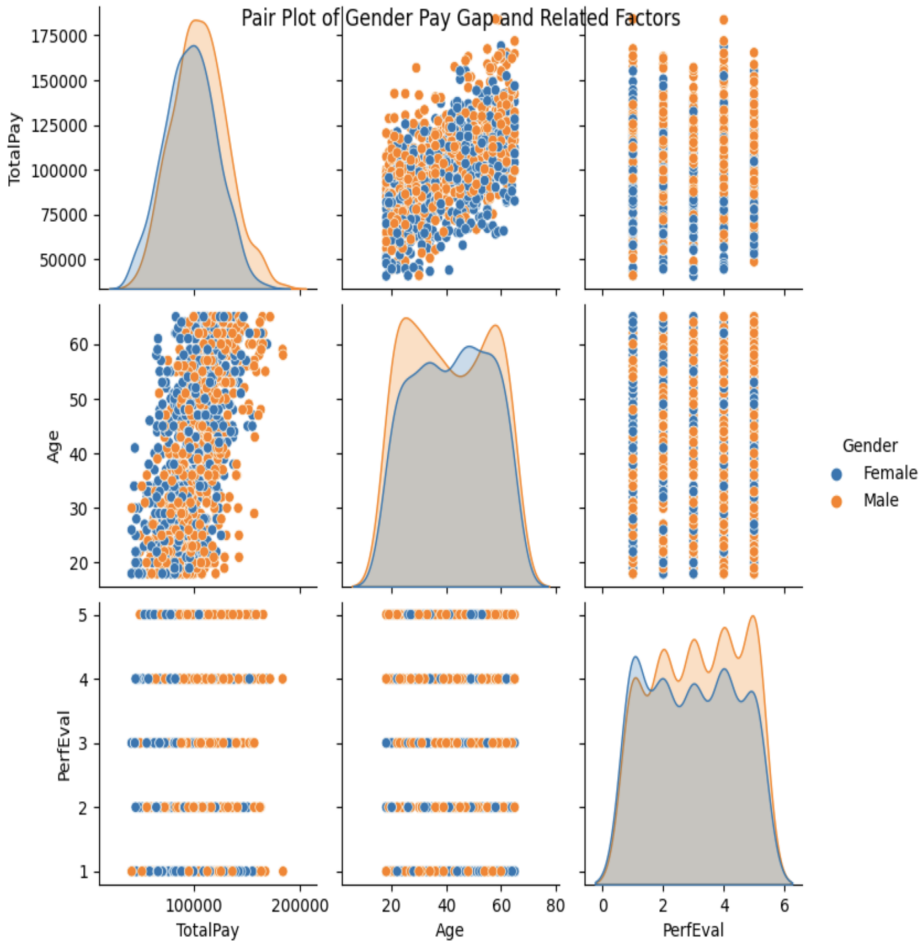

Fig. 5. Pair Plot of Gender Pay Gap and Related Factors

Fig. 4. shows the gender-based job distribution. It can be seen that women tend to work more as marketing associates and men tend to work as software engineers. This can be one of the reasons for the pay gap since software engineers get paid more.

Fig. 5. is a pair plot that helps visualize the relationships between numerical variables, particularly focusing on 'Gender,' 'Total Pay,' 'Age,' and 'PerfEval.'

We can see age plays a major role in the salaries of both men and women. It can be seen that middle-aged women earn more than men but at younger and older ages men earn more. There can be several reasons for such variations.

3.3. Model Explanation and Comparison

In this research project, we employed three machine learning algorithms to analyze and predict gender-based pay gaps: Decision Tree, Random Forest, and XGBoost. These algorithms, each with its unique characteristics, were applied to the dataset to discern the wage disparities based on gender. The performance of these models was evaluated using key metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2), to gauge their effectiveness in predicting and understanding gender-based pay gaps.

Decision Tree

Working: Data Splitting: Decision Trees begin by taking a dataset and selecting the attribute that, when used as a split, maximizes information gain or minimises impurity.

Splitting Criteria: The chosen attribute creates subsets of data. The process is repeated recursively for these subsets, forming branches or nodes.

Stopping Criterion: This recursive splitting continues until a stopping criterion is met, such as reaching a predefined depth or having a minimum number of samples in a leaf node.

Prediction: Each leaf node represents a prediction for the target variable based on the majority class (classification) or the mean (regression) of the samples within that leaf.

Results:

MAE: 910.555

MSE: 1,485,001.735

RMSE: 1218.606

R2: 0.9976329

The Decision Tree model exhibits impressive performance metrics. With a low MAE, it signifies that the model's predictions are, on average, only 910.555 units away from the actual values. The MSE and RMSE are also relatively low, indicating that the model's predictions are generally accurate and have relatively small errors. The R2 value of 0.9976329 indicates that the Decision Tree model explains a significant proportion of the variance in the data, underscoring its predictive power.

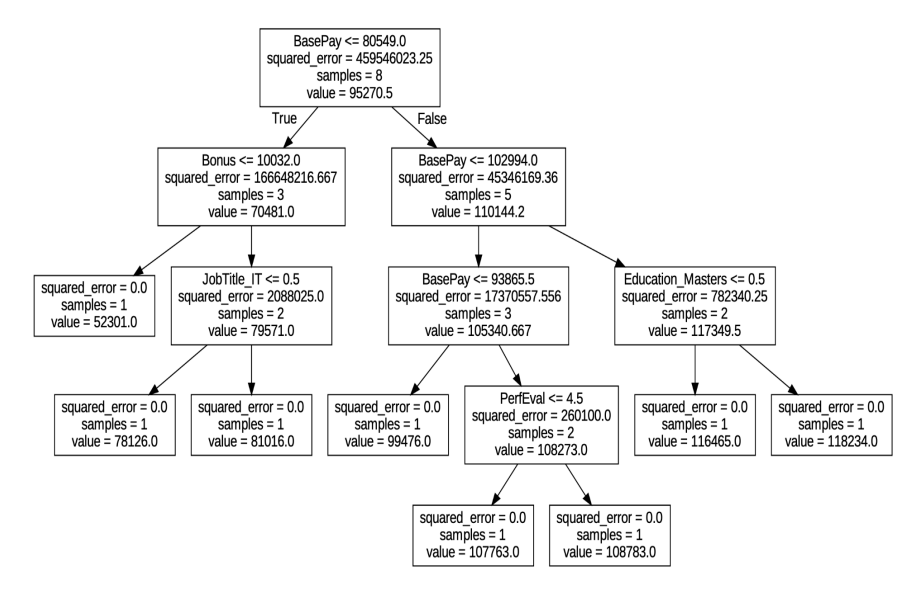

Fig. 6. Decision Tree Diagram

Random Forest

Working: Bootstrap Sampling: Random Forest begins by creating multiple subsets of the data through bootstrapping, which means taking random samples with replacements from the original dataset.

Ensemble of Decision Trees: A Decision Tree is built for each bootstrapped subset of data.

Voting/Averaging: During prediction, each tree in the ensemble makes a prediction, and the outcome is determined by a majority vote (classification) or an average (regression) of the individual tree predictions.

Randomness: Randomness is introduced during both data sampling and attribute selection in each tree to reduce overfitting.

Results:

MAE: 548.1618

MSE: 656,936.7039

RMSE: 810.5163

R2: 0.9989529

The Random Forest model improves upon the already impressive performance of the Decision Tree. With a lower MAE and RMSE, it produces even more accurate predictions, with an average deviation of 548.1618 units from actual values. The R2 value of 0.9989529 is exceptionally high, indicating that the Random Forest model explains a vast majority of the variance in the data, making it a powerful tool for understanding gender-based pay disparities.

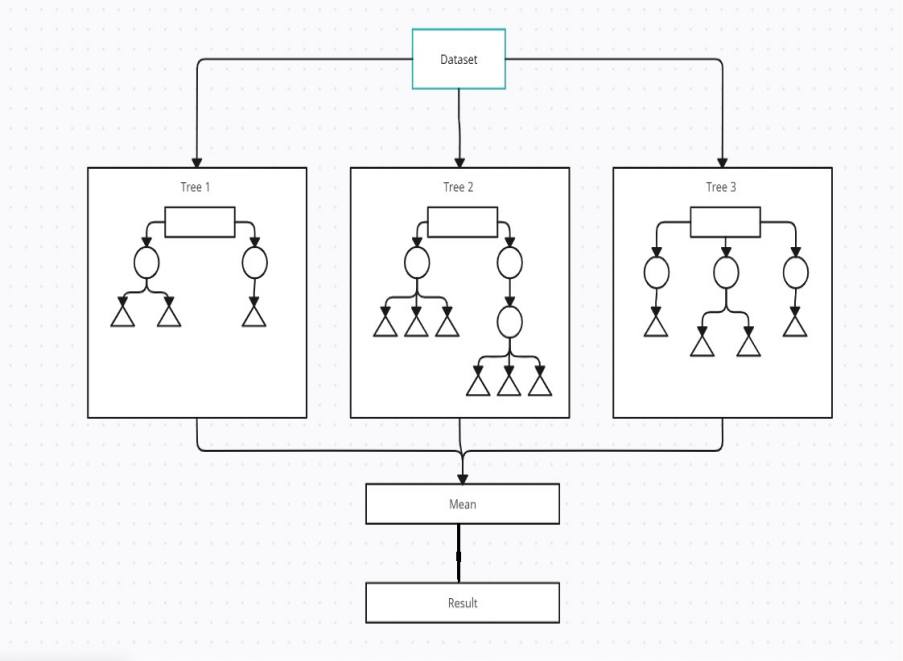

Fig. 7. Architecture diagram of Random Forest

XGBoost

Working: Initialization: XGBoost starts with an initial simple predictor, often the mean of the target variable for regression or a constant for classification.

Residual Calculation: It calculates the residual errors by comparing the actual target values with the current model's predictions.

Building Weak Learners: A new weak learner, typically a decision tree, is trained to predict these residuals. The weak learner's task is to improve upon the model's errors.

Ensemble Building: The new learner's predictions are added to the existing model, and this process is repeated iteratively. Each new learner focuses on the errors made by the previous ensemble.

Optimisation techniques: XGBoost employs optimization techniques like regularization, tree-pruning, and feature selection to improve efficiency and predictive accuracy.

Final Prediction: The final prediction is the sum of the initial model's prediction and the cumulative predictions of the weak learners.

Results:

MAE: 534.6749

MSE: 619,334.3096

RMSE: 786.9780

R2: 0.9990128

The XGBoost model further enhances predictive accuracy. With the lowest MAE, MSE, and RMSE of all the models, it consistently generates predictions closest to the actual values, with an average deviation of 534.6749 units. The exceptionally high R2 value of 0.9990128 indicates that XGBoost excels at explaining the variance in the dataset, making it an invaluable tool for the analysis of gender pay gaps.

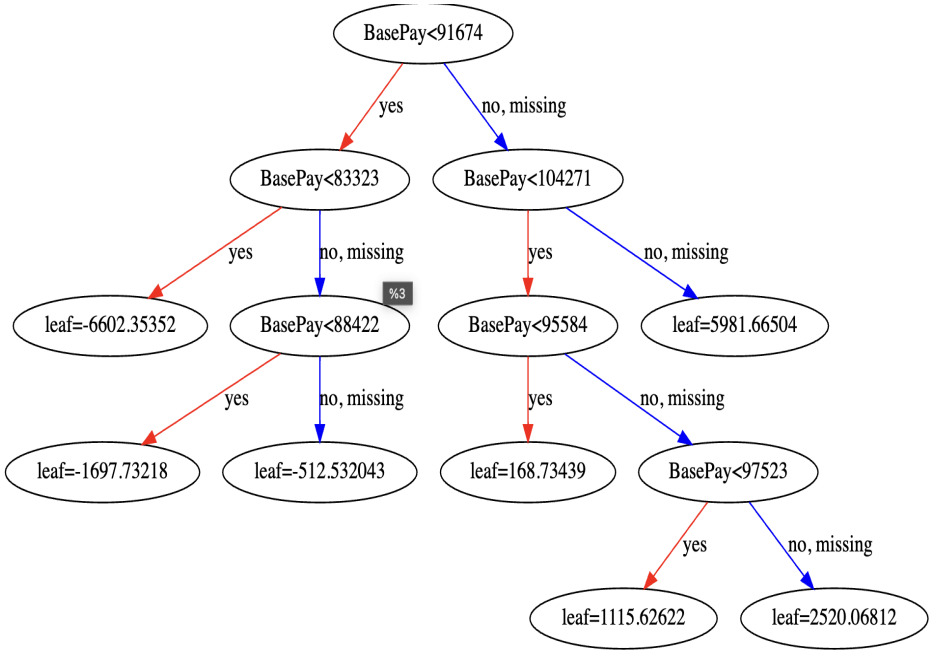

Fig. 8. XGBoost Decision Tree Diagram

3.4. Improvement in the above models

We can see above that the R2 score for all algorithms is 0.99, which indicates that the dataset may be overfitting. Consequently, we decided to do some data preprocessing to give significance to factors other than base pay and bonus pay in the prediction of total pay.

Data pre-processing

Creating Dummy Variables for Categorical Columns:

Dummy variables are created for three categorical columns: 'Gender,' 'Dept,' and 'Education.' This is done using the get_dummies function, which converts categorical variables into binary (0 or 1) columns for each category.

Prefixes 'Sex,' 'Dept,' and 'Edu' are added to the column names of the corresponding dummy variables.

The drop_first parameter is set to False to keep the reference categories. If set to True, it would drop one of the categories to avoid multicollinearity.

Removing Reference Categories:

Reference categories are removed from the dummy variables created. For example, the 'Sex_Female' column is removed from the 'Gender' dummy variables, and 'Dept_Administration' is removed from the 'Dept' dummy variables. This is typically done to prevent multicollinearity when using these variables in a regression model.

Combining Dummy Variables and Numeric Columns:

The dummy variables and other numeric columns ('Age,' 'PerfEval,' 'Seniority') are concatenated horizontally (along the columns) to create the feature matrix 'X.' This 'X' matrix is what is typically used for modelling.

Table. 1. Comparison Table of the algorithms after preprocessing

Model	MAE	R2
Linear Regression	11,548.90	0.6572
Ridge Regression	11551.34	0.6569
Lasso Regression	11548.90	0.6570

The comparison of the models based on MAE and R2 metrics provides insights into their predictive performance. The results show that after preprocessing, Linear Regression and Lasso Regression models have similar MAE values, indicating their ability to predict the total pay gap effectively. Ridge Regression, while slightly less accurate in terms of MAE, is still competitive.

In terms of R2, all three models demonstrate a good fit after preprocessing. The differences in R2 values between the models are marginal. This suggests that the preprocessing techniques applied to the dataset have significantly improved the model's goodness of fit.

4. Results And Discussions

The analysis of the gender pay gap, in the context of this research, revolves around the utilization of machine learning models and data visualization techniques. Our study employed decision tree regression, random forest regression, and XGBoost regression models to predict and analyze gender-based wage disparities. The dataset consists of 10 distinct job roles, revealing a striking and noteworthy observation: women outearn men in five of these roles, while men out-earn women in the other five. This finding underscores the intricate and multifaceted nature of the gender pay gap, going beyond simple statistical averages and delving deeper into the contributing factors.

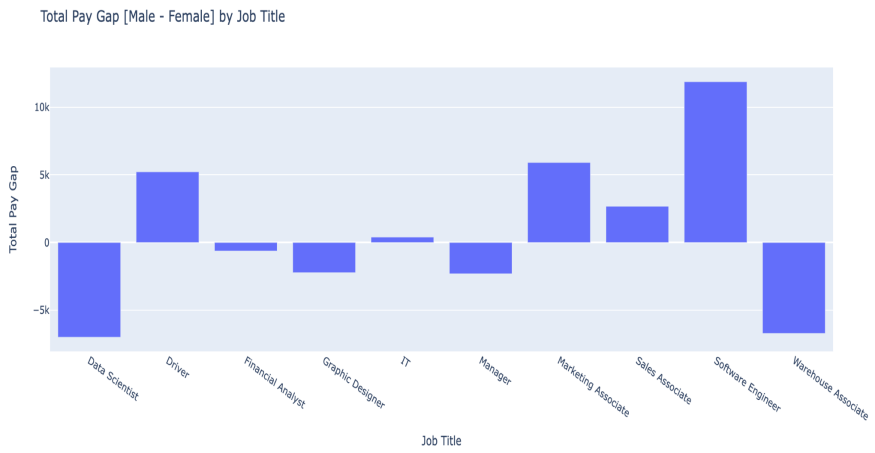

Fig. 9 Total Pay Gap by Job Title

4.1 Variations within Job Roles

One of the most notable findings of this study is the recognition of variations within specific job roles. While the conventional approach tends to depict a uniform gender pay gap, our approach identifies that pay disparities can differ significantly across various job contexts. The discovery that women out-earn men in some job roles challenges conventional narratives and calls for a more nuanced understanding of gender-based wage disparities.

4.2 Statistical Significance

Hypothesis testing confirmed the presence of a statistically significant difference in total pay between genders in the dataset. This result highlights that gender has a substantial impact on an individual's total pay. However, it is crucial to understand that statistical significance does not inherently provide insights into the underlying reasons behind the gender pay gap. Further exploration and analysis are required to uncover the intricate factors contributing to these disparities.

4.3 Factor Analysis

The study incorporates various factors that affect the gender pay gap, including job distribution, performance evaluation, and age. Notably, the analysis reveals that age plays a crucial role in determining salaries. While middle-aged women tend to earn more than their male counterparts, disparities are evident at younger and older ages. The reasons for these variations can be multifaceted and deserve more in-depth investigation.

4.4 Model Performance

The research employs three machine learning algorithms—Decision Tree, Random Forest, and XGBoost—to predict gender-based pay gaps. The models are evaluated using critical metrics, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2), to gauge their predictive accuracy.

Decision Tree: The Decision Tree model exhibits commendable performance, with an R2 value of 0.9976329. It explains a significant portion of the variance in the data, indicating its predictive power.

Random Forest: The Random Forest model further enhances predictive accuracy. With an R2 value of 0.9989529, it provides highly accurate predictions of gender-based wage disparities.

XGBoost: The XGBoost model stands out with the lowest MAE, MSE, and RMSE among all models, and an R2 value of 0.9990128, demonstrating exceptional predictive capabilities.

4.5 Data Preprocessing

To mitigate potential overfitting issues, data preprocessing was performed to give significance to factors beyond base pay and bonus pay in the prediction of total pay. This involved creating dummy variables for categorical columns ('Gender,' 'Dept,' and 'Education') and removing reference categories to prevent multicollinearity. The resulting feature matrix 'X' was used for modelling.

4.6 Comparison After Preprocessing:

Following data preprocessing, Linear Regression, Ridge Regression, and Lasso Regression models were evaluated. Their performance metrics are as follows:

Linear Regression: MAE: 11,548.90, R2: 0.6572

Ridge Regression: MAE: 11,551.34, R2: 0.6569

Lasso Regression: MAE: 11,548.90, R2: 0.6570

The models show similar performance after preprocessing, with marginal variations in MAE and R2 values. These results suggest that preprocessing techniques have significantly improved the model's goodness of fit.

5. Conclusion

This research explores the gender pay gap using machine learning algorithms, providing insights beyond the conventional narrative. By recognizing variations within job roles and utilizing advanced data analysis techniques, the study contributes to a more nuanced understanding of gender-based wage disparities.

The findings hold implications for policymakers, businesses, and organizations aiming to address gender pay inequality and promote pay equity. This research is a crucial step towards a more inclusive and equitable workforce, fostering a deeper comprehension of the complexities surrounding gender-based wage disparities and offering a path forward for their comprehensive and effective resolution.

6. Implications and Future Work

The study's findings can guide stakeholders in designing targeted strategies and interventions to promote pay equity, considering the nuanced variations in wage disparities across different job roles. Future research could explore additional factors contributing to gender-based pay gaps and further refine models to enhance predictive accuracy and robustness in addressing these disparities.

References

1. Goldin, C., & Katz, L. F. (2000). ***"The gender pay gap in the United States"***: evidence from the current population survey Journal of Economic Perspectives, 14(4), 31–44.
2. Booth, A. L., & De Vroey, D. N. (2004). ***"The gender pay gap in the United Kingdom"***: A review of the literature Industrial and Labour Relations Review, 57(4), 568–591.
3. Cassells, R. L., & Phibbs, P. J. N. (2010). ***"The gender pay gap in Australia":*** An analysis of the Workplace Gender Equality Agency's Workplace Gender Equality Index data Journal of Industrial Relations, 52(2), 163–185.
4. Green, D. A., Milligan, K., & St-Hilaire, M. (2013). ***The gender pay gap in Canada:*** An analysis of the Survey of Labour and Income Dynamics Canadian Journal of Economics, 46(3), 385–418.
5. Björklund, Å., & Östlin, M. J. (2015). ***"The gender pay gap in Sweden"***: An analysis of the effects of occupational segregation and discrimination Journal of Human Resources, 50(3), 715–750.
6. Boushey, H., & Glynn, S. J. (2017). ***"The gender pay gap in the United"***
7. Cantillon, B., Green, D. A., & Milligan, K. (2018). ***"The gender pay gap in developing countries"***: evidence from sub-Saharan Africa World Development, 104, 128–149.
8. Ariely, D., Bohns, V., & List, J. (2019). ***"The gender pay gap in the gig economy"***: Evidence from the United States Proceedings of the National Academy of Sciences, 116(29), 14578–14584.
9. Ammerman, C., & Bosler, M. (2020). ***"The gender pay gap in the tech industry"***: Evidence from the United States Harvard Business Review, 98(6), 124–134.
10. Goldin, C., & Katz, L. F. (2021). ***"The gender pay gap during the COVID-19 pandemic"***: Evidence from the United States NBER Working Paper No. 28802.