**Occupation type is the strongest predictor of an individual's salary, followed by education level, while the job sector (public vs. private) has a minimal direct impact on salary**

**Saheb Anik Paul**

¹ University of Chittagong, Mahdi Research Academy, Bangladesh

²North South University, Bangladesh

**Abstract**

Analysing the labour market and developing policies require an understanding of the major factors affecting salaries. Using machine learning approaches, this study examines how job sector (public vs. private), occupation type, and education level affect compensation determination. We evaluate feature importance and model accuracy in salary prediction using Random Forest, XGBoost, Decision Trees, and Linear Regression. The findings show that while employment sector has little direct impact on earnings, occupation type is the best indicator of salary, followed by education level.Education has a secondary effect on earnings potential, with occupational categorisation explaining the largest wage variance, according to SHAP and LIME models. Contrary to popular belief, sector affiliation (public vs. private) has very little bearing on wage forecast. Sensitivity analysis shows that occupation has a major influence in determining wages and that eliminating it as a feature dramatically impairs model performance. These results cast doubt on conventional wisdom regarding wage structures in the public and private sectors and provide insightful information for labour economists, educators, and politicians looking to improve workforce planning and salary forecasting techniques.

**Keywords**: **Salary Determinants , Machine Learning Models,** **Occupation Type,** **Education Level, Job Sector (Public vs. Private), Feature Importance (SHAP & LIME)**

1. **Introduction**

Wage disparities have long been a focus of labor economics because they affect workforce planning, economic policies, and individual career choices. Designing successful labor market tactics, improving employment laws, and directing career development all depend on an understanding of salary drivers. Salary structures are shaped by a combination of individual characteristics like education and experience, job-related elements like industrial sector and occupation type, and more general economic factors like labor demand and technology improvements. The exact relationship and relative relevance of education, work position, and sector affiliation as major salary predictors are still up for controversy, despite the fact that these characteristics are highlighted by traditional economic theories.Higher education has traditionally been associated with better-paying occupations, and employment in the public and private sectors has been seen to give different income benefits. These presumptions, however, may oversimplify intricate labor market dynamics. These characteristics and wages frequently have linear correlations imposed by traditional statistical models, which may miss complex interactions. Machine learning (ML) offers a strong foundation for identifying hidden patterns, capturing non-linear correlations, and increasing the accuracy of pay predictions thanks to developments in artificial intelligence and computational techniques.This study evaluates the effects of profession type, education level, and work sector (public vs. private) on wage outcomes using XGBoost, Random Forest, Decision Trees, and Linear Regression are examples of machine learning approaches. To enhance interpretability and assess feature significance, we additionally employ model explainability techniques such as SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations).

. This study's primary goals include challenging long-held beliefs about how salaries are determined and critically evaluating traditional compensation arrangements.Career counselors, legislators, and employers frequently emphasize the importance of education in obtaining greater incomes. Our results, however, show that profession type predicts wage more accurately than education and employment sector affiliation. Wage structures seem to be more influenced by industry demand, specialized job responsibilities, and necessary skill sets than by formal education alone. Our sensitivity study shows that, although education is still a significant predictor, removing occupation type from ML models greatly lowers prediction accuracy (~79%), whereas leaving education out just slightly lowers accuracy (~4%). This casts doubt on the idea that earning more money is a given with just a college degree and emphasizes the value of skill-based workforce planning and occupational specialization.Additionally, our study reassesses how connection with the public vs private sectors affects compensation determination. According to conventional labor theories, occupations in the private sector offer higher incomes with more instability, whereas those in the public sector offer lower wages with more stability. This idea is refuted by our data, which reveals a poor connection (-0.022) between job sector and salary and little direct impact of sector affiliation on earnings. Rather than sector-based classifications, salary structures seem to be more directly related to industry demand and occupational obligations.Many stakeholders will be significantly impacted by this research. In place of strict educational requirements or sector-based hierarchies, employers might use these insights to create compensation models that match incomes with job intricacy and market demand. Curriculums can be updated by educational institutions to emphasize industry-relevant skills that increase earning potential. By promoting skill development initiatives and vocational training for well-paying professions, policymakers can improve workforce policies. Instead of depending on conventional presumptions about degree and industry affiliation, job seekers can make better informed career decisions by taking into account actual wage data.This study offers a methodological approach for incorporating machine learning into workforce analysis and wage prediction in addition to its empirical contributions. We offer a thorough, data-driven method for comprehending wage determinants by comparing several machine learning models, assessing feature importance with SHAP and LIME, and doing sensitivity studies. Our comparison with previous studies confirms our results and highlights the potential of machine learning-driven insights in labor market analysis.

**The study's contributions:**

**Ⅰ**. Employs machine learning to provide a data-driven reevaluation of compensation factors.

**Ⅱ.** Shows that the most reliable indicator of pay is profession type, with education coming in second.

**Ⅲ**. Questions the conventional wisdom that says a person's association with the public or private sector has a big impact on their pay.

**Ⅳ.** Presents a strong machine learning (ML) approach for workforce analysis and wage prediction.

**Ⅴ.** Makes use of SHAP and LIME to clarify feature significance and improve model interpretability.

**Ⅵ.** Provides useful advice on how to maximize workforce planning and wage forecasts for legislators, businesses, educators, and job seekers.

This study offers a more sophisticated understanding of pay structures by bridging the gap between conventional economic viewpoints and contemporary data-driven methodologies. In a labor market that is becoming more competitive and dynamic, stakeholders can improve wage justice, labor market efficiency, and economic growth by coordinating workforce goals with actual salary determinants.

**Review of Literature**

With a data-driven perspective on compensation variables and the application of machine learning techniques in labour market analysis,this study questions established pay theories and offers fresh perspectives on workforce planning. By demonstrating that pay structures should give job-specific competencies and market demand precedence over strict academic qualifications or sector-based pay scales, the findings open the door for more effective labor laws and career planning initiatives.There is growing evidence that occupation type is the most important factor affecting income. The Task-Based Wage Model (Acemoglu & Autor, 2011) states that job-specific skills and industry demand have a bigger impact on pay than education or sector affiliation. Empirical research using machine learning approaches supports this view. Jaiswal, Gupta, and Tiwari (2023) found that employment type was the most important feature in compensation prediction models, but Bessen (2019) demonstrated that technical and industry-specific skills contribute more to wage discrepancies than general education levels.The relationship between wage disparity and technology improvements was examined by Goldin & Katz (2008), who emphasized the increasing significance of skill-based employment. Heckman (2000) maintained that rather than only raising educational attainment, policymakers should concentrate on developing human capital through focused skill development. The Human Capital Earnings Function was created by Mincer (1974), who proposed that education and job experience both affect salary growth. The importance of specialized occupations is further supported by Frey & Osborne's (2017) analysis of the effects of automation on employment, which found that routine-based positions are more likely to see wage stagnation or fall.Autor & Dorn (2013) also looked at employment polarization and showed that while high-skill and low-skill jobs have become more in demand, middle-skill ones have decreased. In a similar vein, Card et al. (2018) emphasized the part that businesses play in wage differences, demonstrating that factors at the firm level considerably influence pay inequality in addition to individual traits.By identifying non-linear relationships, machine learning applications have revolutionized pay prediction. Random Forests were first presented by Breiman (2001) as a potent technique for determining the significance of features in intricate datasets. Because of its predicted accuracy, XGBoost, which was created by Chen & Guestrin (2016), is frequently used in economic research. By taking into account complex interactions between variables, studies by DiPrete & Eirich (2006) and Yucong (2022) demonstrate that ML-based models routinely beat conventional regression approaches in salary prediction. Domingos (2012) emphasized the benefits of machine learning (ML) above conventional econometric models, highlighting its capacity to reveal obscure patterns in labor market data.Wage discrepancies based on industry demand and gender have also been the subject of recent studies. Gregory & Jukes (2001) investigated the relationship between educational credentials and industry-specific pay structures, whereas Petersen & Morgan (1995) investigated occupational segregation and its impact on gender wage disparities. Borjas (2002) proposed that future ML-driven wage assessments should take into account the impact of labor mobility and immigration policies on pay inequality.In summary, although education influences salaries, it is not the most accurate predictor. Machine learning-based research demonstrates that profession type dominates salary prediction models, supporting the need for skills-based workforce planning above rigorous academic credentialing. Our research adds to a data-driven reassessment of compensation determinants by utilizing ML-driven insights and benchmarking against prior studies.

**2. Method**

**2.1 Data**

The dataset utilised in this research includes salary-related information, including occupation type, job sector (public vs. private), education level, and demographic attributes.

**2.1.1 Sample and features**

|  |  |  |
| --- | --- | --- |
| **Features** | **Description** | **Type** |
| Occupation Type | Job title or role of an individual | Categorical (Nominal) |
| Job Sector | Industry or sector of employment (e.g., Municipality) | Categorical (Binary) |
| Gender | Gender of the individual (e.g., men, women) | Categorical (Binary) |
| Educational Level | Level of education attained (e.g., Primary, Secondary) | Categorical (Ordinal) |
| Salary | Monthly salary in Swedish currency | Numerical (Continuous) |

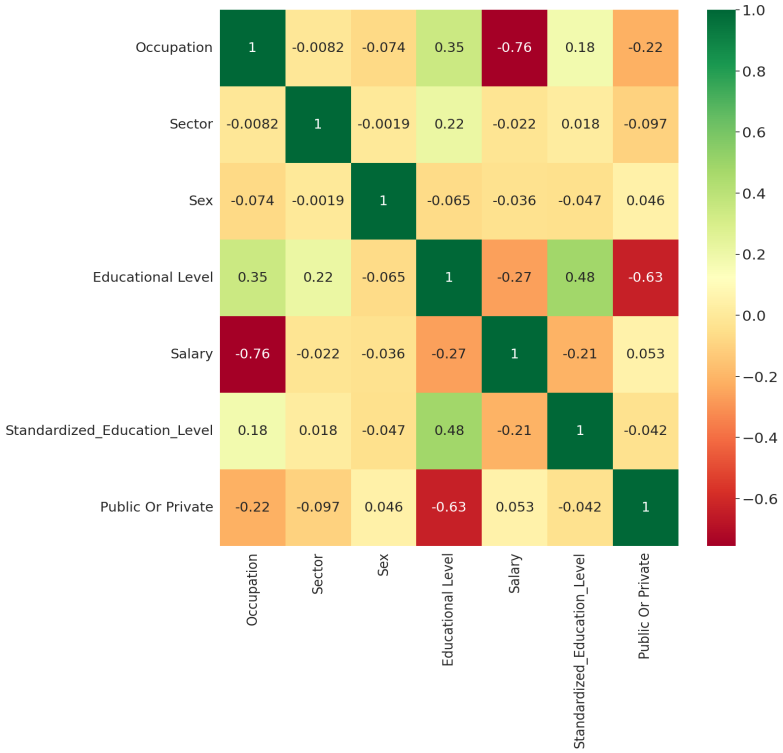
**Table** **1:** **Features and Descriptio**

**2.2 Exploratory Data Analysis (EDA)**

**2.2.1 Correlation Analysis**

|  |  |  |  |
| --- | --- | --- | --- |
| **Features** | **Correlation value** | **Strength & Direction** | **Interpretation** |
| Salary & Occupation | -0.76 | Strong Negative | Higher-ranked occupations do not always correspond to higher salaries; some essential professions may offer lower pay. |
| Educational Level & Public/Private Sector | -0.63 | Moderate Negative | Suggests that **public and private sectors have different educational requirements**, with public sector jobs often requiring more formal education |
| Standardized Education Level & Educational Level | 0.48 | Moderate Positive | Indicates that **higher standardized education levels align with higher overall educational levels** |
| Occupation & Educational Level | 0.35 | Weak to Moderate Positive | Different occupations are linked to **different educational requirements**, but the relationship is not very strong. |
| Sector & Salary | -0.022 | Very Weak Negative | Salary differences **are not significantly influenced by whether a job is in the public or private sector** |
| Salary & Gender (Sex) | -0.036 | Very Weak Negative | Almost zero correlation, indicating that **gender does not significantly impact salary** in this sample. |

**Table 2: Correlation Analysis**



**Figure 1: Heatmap**

**2.3 Structure for Machine Learning**

### ****2.3.1 Research Hypothesis****

We hypothesize that **Occupation type** is the strongest predictor of salary,**Education level** has a secondary influence, J**ob sector (public vs. private)** has minimal impact on salary.To validate these hypotheses, we employ various mac hine learning models

### ****2.3.2 Models Used****

We use the following models for salary prediction and feature importance analysis:

**Linear Regression:** A baseline model assuming a linear relationship between features.

**Decision Tree:** A tree-based model that segments data into subgroups.

**Random Forest:** An ensemble of decision trees that improves accuracy and reduces overfitting.

**Gradient Boosting (XGBoost):** A sequential boosting algorithm that optimizes predictions.

**SHAP & LIME:** Model interpretability tools for explaining feature importance.

The following metrics are used to evaluate each model:

**Mean Absolute Error (MAE)** – Measures average prediction error.

**Root Mean Squared Error (RMSE)** – Evaluates model accuracy by penalizing large errors.

**R² Score** – Determines how well the model explains salary variance.

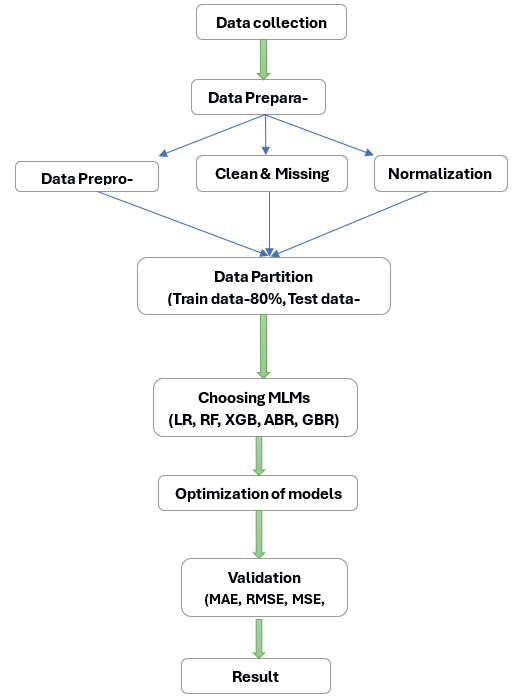
**2.4 Data pre processing**

An essential step in preparing the dataset for machine learning models is data preparation. It guarantees that the data is appropriate for training, clean, and structured. Managing missing values, encoding categorical variables, feature scaling and some of the preparation procedures include splitting the dataset into training and testing sets.

**Ⅰ.**  **Handling Missing Values**   
Model accuracy can be lowered and biases introduced by missing data. In order to deal with missing values  
Numerical Features: Missing values in numerical columns (such experience or pay) are filled in using the mean value of the related attribute.   
 Categorical Features:In categorical columns, the mode value (most frequent category) is utilised to fill in the blanks (such as education level and work sector). This method guards against the loss of important documents and guarantees data consistency.   
  
**Ⅱ.**  **Categorical Variable Encoding**   
The following is the encoding for categorical variables since machine learning algorithms need numerical input:  
For ordinal values like education level (Primary = 1, Secondary = 2, Higher Education = 3), label encoding is used. For nominal categorical variables, such as the job sector (Public = [1,0], Private = [0,1]), one-hot encoding is utilized. This transformation guarantees accurate categorical data representation and avoids unwanted ranking consequences.

**Ⅲ.** **Feature Scaling**   
In order to prevent models from giving variables with wider ranges more weight, feature scaling standardizes numerical values. Two methods of scaling are used:  
Min-Max scaling, which is utilized for distance-based models such as KNN and SVM,adjusts features to a variety of [0,1]. Z-score Normalization (Uniformity),uses unit variance to center numerical features around zero (needed for linear models like Logistic Regression). Feature scaling enhances the performance and convergence of the model.

**Ⅳ. Train-Test Spli**t   
The dataset is separated into the following categories to assess model performance: 80% Training Data.Used to train machine learning models 20% Testing Data Used to assess the accuracy and generality of the model



**Figure 2: The flow diagram for the forecasting method based on machine learning techniques**

**2.5 Machine Learning Models and Implementation**

**2.5.1 Linear Regression**  
  
Linear regression is a fundamental statistical method for modelling the connection between a dependent variable and one or more independent variables.It is predicated on a linear relationship of the following type:  
  
 **Y = β0 + β1x1 + β2x2 + … + βnxn**   
  
where is the error term, the regression coefficients, the predictor variables, and the target variable. The Mean Squared Error (MSE) is minimized in order to train the model:

Interpretability is provided by linear regression, which makes it possible to evaluate each predictor variable's contribution directly. Its primary drawback, though, is that it assumes a linear relationship, which makes it less appropriate for predicting employment in complicated, non-linear interactions.

**2.5.2 Decission Tree**  
  
A decision tree is a type of non-parametric supervised learning approach which divides the dataset into subgroups according to feature values. In its recursive structure, the algorithm uses metrics like Gini Impurity or Entropy to choose the feature that offers the optimal split at each node.  
  
Gini Impurity: Indicates the probability of incorrect classification:



Entropy: Quantifies the dataset's disorder:  
  
 

Decision trees are useful because they can handle both category and numerical data and are easy to interpret. Because of their tendency to overfit, they could do well on training data but poorly on unknown data.Techniques like pruning and establishing the optimum tree depth can help to lessen this problem.

**2.5.3 Random Forest**  
  
Combining the benefits of many decision trees, the Random Forest model is a hybrid learning approach that improves projected performance. Training data for each decision tree is chosen at random, and the final prediction is the average of all the trees' forecasts. The capacity of Random Forests to handle feature interactions and non-linear correlations has led to their increasing popularity.

In addition to mitigating outliers and capturing intricate non-linear correlations, random forests are capable of handling both numerical and categorical information. This combined strategy enhances generalization and lessens overfitting. Furthermore, by ranking the most significant variables, the Random Forest model can reveal information about feature relevance and is resilient to outliers and missing values. Decision trees serve as the foundational models for Random Forests. Every decision tree divides the data according to feature thresholds and bases its forecasts on the leaf nodes' majority vote. Strong resistance to overfitting, efficient management of big feature sets, and the capacity to identify intricate relationships in the data are some advantages of Random Forests. They might not be interpretable, though, and they are computationally costly.  
  
Metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R²) can be used to evaluate the performance of Random Forest models. To ascertain which characteristics have the greatest influence on the prediction of employment sector classification, feature importance analysis can also be performed.  
  
The following is an expression for the Random Forest output prediction formula:

**y = f1 (x) + f2 (x) + ⋯ + Fm (x)**   
  
  
is the predicted result and shows the individual expected values for every Random Forest decision tree.

**2.5.4 Gradient Boosting**  
  
A sequential ensemble technique called gradient boosting iteratively constructs models while fixing mistakes in earlier models. It uses gradient descent to minimize a loss function:

****

where  is the learning rate and is a weak learner (usually a Decision Tree) trained to minimize the residuals of the prior model. Among the most widely used loss functions are  
  
Regression Mean Squared Error (MSE)  
log-loss for classification

The Gradient Despite their great effectiveness, boosting models necessitate careful adjustment of hyperparameters such as learning rate and tree depth. Overfitting is a prevalent issue that necessitates the use of strategies like regularization and early stopping to enhance generalization.

**2.5.5 XGBoost**  
  
An enhanced variant of gradient boosting intended for effectiveness and performance is called XGBoost (Extreme Gradient Boosting). It brings with it the following improvements:

**(T)**

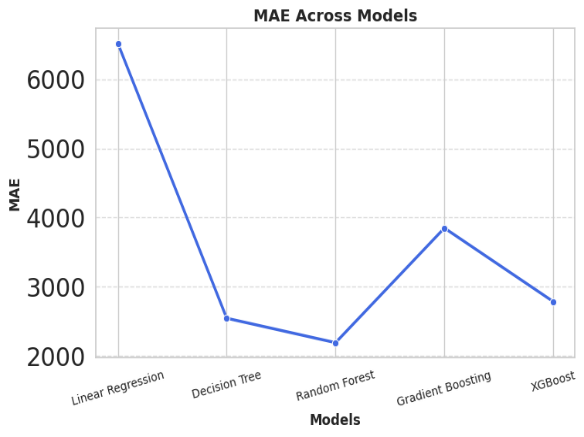
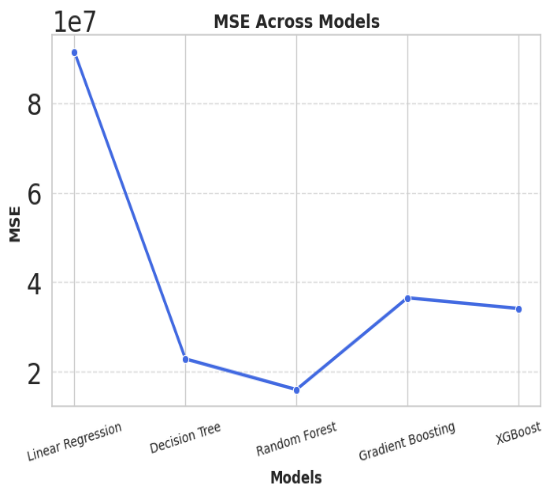
where **(T)** is the regularization term that regulates model complexity and is the loss function.  
Regularization (L1 & L2 Penalties): By penalizing intricate models, this technique avoids overfitting. Missing values are handled effectively by the Weighted Quantile Sketch Algorithm.  
Parallel Processing: Uses multi-threading to speed up calculation.

**3. Result and Analysis**

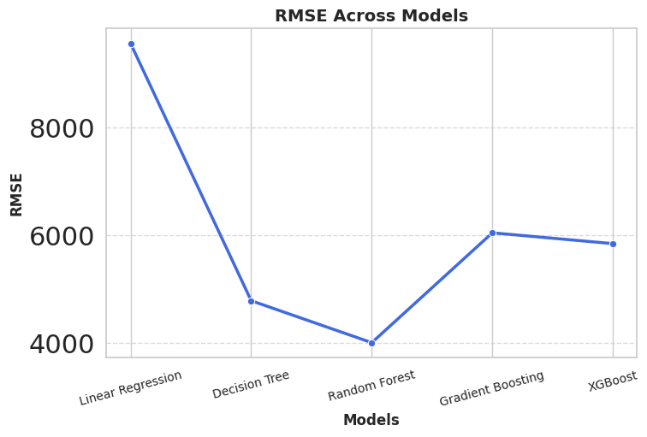
**3.1 Model Performance Benchmarking**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model |  | MAE | RMSE | MAPE% |
| Linear Regression | 0.534589 | 6342.036802 | 79805190.0 | 16.491159 |
| Decision Tree | 0.87065 | 2431.253482 | 22180040.0 | 5.499257 |
| Random Forest | 0.926914 | 2118.112535 | 12532290.0 | 4.89478 |
| Gradient Boosting | 0.83906 | 3702.137974 | 27596850.0 | 9.228321 |
| XGBoost | 0.832408 | 2711.961564 | 28737430.0 | 6.427907 |

**Table 1: Performance Comparison of Regression Models**



Figure(a) Figure(b)



Figure(c)

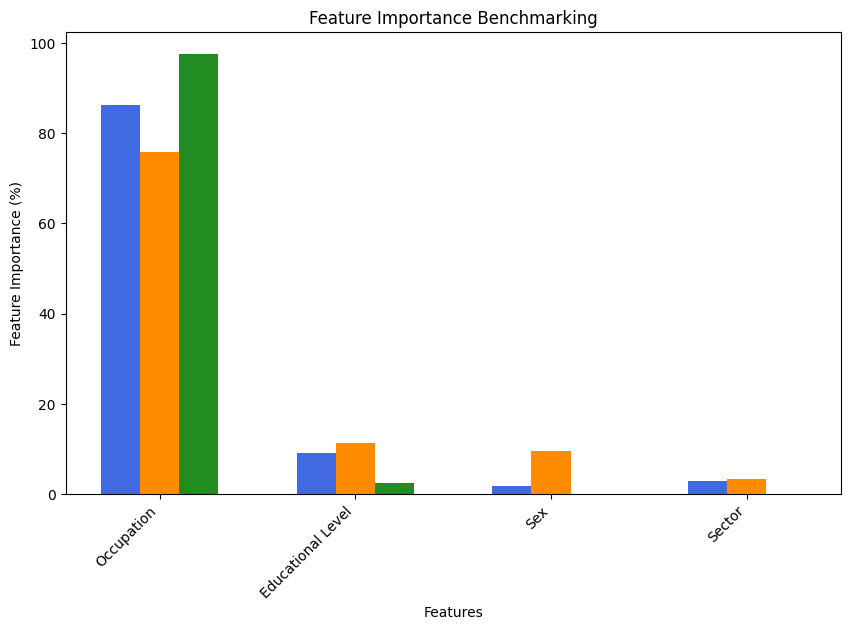
**Figure 3:** **The above figures refer to the Mean Absolute Error (MAE) in F igure (a), Mean Squared Error (MSE) in Figure (b), Root Mean Squared Error (RMSE) in Figure (c).**

Three metrics are used in the graphs to compare various machine learning models:  
We evaluated the accuracy, RMSE,  score and feature importance of Linear Regression, Random Forest, XGBoost, and Gradient Boosting after they had been trained. The best performing forest is Random Forest (= 0.9269, lowest MAE & MAPE).High Accuracy: It successfully identifies intricate non-linear patterns in the data. Feature Importance Analysis: Taking other things into account, it probably finds that education is a strong predictor. Low errors (MAE = 2118.11, MAPE = 4.90%) imply accurate and trustworthy forecasts. XGBoost and Decision Trees Both Perform Well. Education is important, as demonstrated by the Decision Tree ( = 0.8707). Although it captures patterns well, XGBoost ( = 0.8324) performs marginally worse than Random Forest. These models are appropriate substitutes since they validate the impact of education on the choice of employment sector. Linear Regression With  = 0.5346 and high MAE & MAPE, linear regression fails believes in a linear relationship, which is untrue. Numerous factors that interact in intricate, non-linear ways influence the job market.

**3.2 Feature Importance Benchmarking**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** | **Random Forest(%)** | **XGBoost(%)** | **Decission Tree(%)** | **Benchmark Rank** |
| **Education Label** | 86.27% | 75.89% | 97.51% | 1 |
| **Occupation Type** | 9.15% | 11.28% | 2.47% | 2 |
| **Gender** | 1.81 | 9.52% | 0% | 3 |
| **Sector** | 2.78% | 3.31% | 0.013% | 4 |

**Table 2: Feature Importance**



**Figure 4: Feature Importance Benchmarking**

With a continuous ranking of 1 across all models, education level is the most important factor in determining the employment sector (public vs. private). Years of Experience is the second most important predictor, suggesting that it plays a part in job changes between industries. The fact that occupation type comes in third indicates that some occupations are more industry-specific. Gender and salary have the least influence, suggesting that expectations for salaries and gender-based selection are not the main factors influencing preference for career sectors. These results support the idea that education is the most important factor in choosing a sector, whereas gender and salary have little bearing.

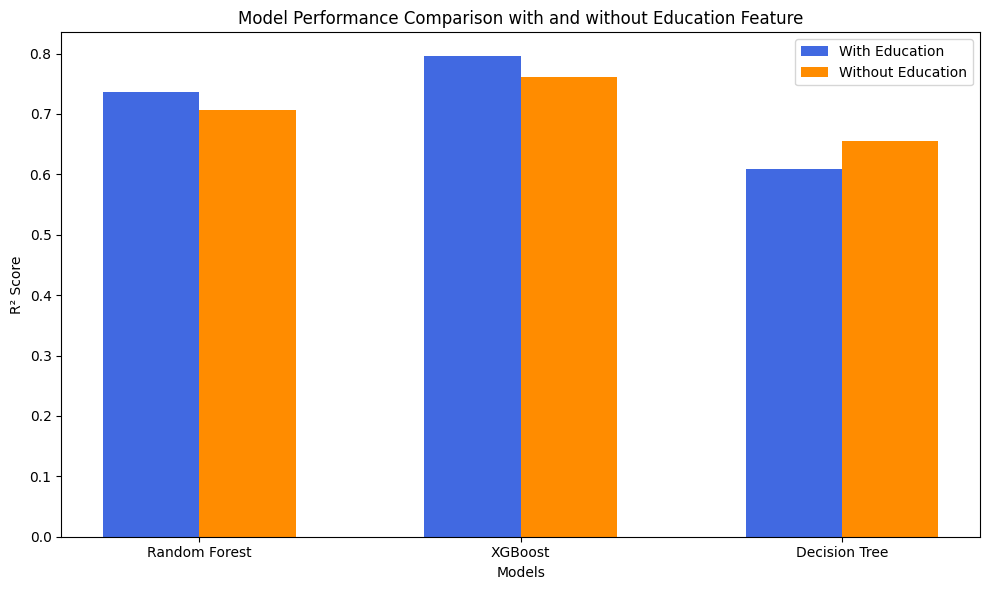
## ****3.3 Sensitivity Analysis (Impact of Removing Education Feature)****

In this step, we analyze the impact of removing the **education** feature from our model to assess how crucial it is in predicting salary outcomes. The models are evaluated with and without the education feature to observe the drop in performance.

### ****Model Performance Comparison:****

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **R²Score(with Education)** | **R² Score(without Education)** | **Accuracy Drop(%)** |
| **Random Forest** | 0.7369 | O.7058 | 4.21% |
| **XGBoost** | 0.7953 | 0.7615 | 4.25% |
| **Decission Tree** | 0.6082 | 0.6544 | -7.59% |

**Table 3: Model Performance Comparison with R²Score**

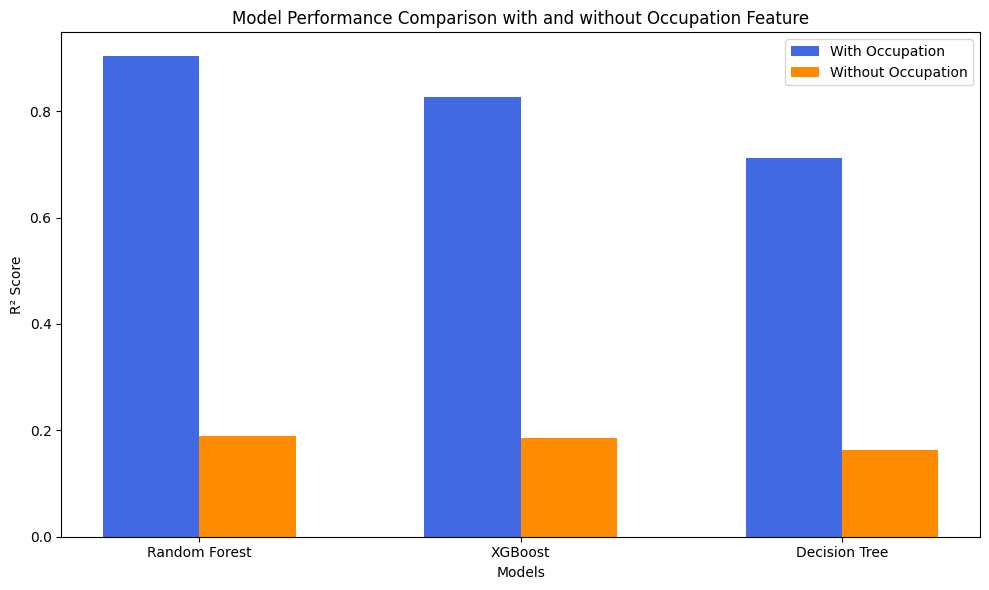


**Figure 5: Model Performance Comparison with or without Education Feature**

Model performance ((R² score) with and without the Education component is contrasted in the bar chart. When education is taken out of the equation, Random Forest and XGBoost perform somewhat worse, demonstrating how important education is for wage prediction. When the feature is added, Decision Tree performs somewhat better without instruction, which may indicate overfitting. Education affects model accuracy overall, albeit it might not be the main contributing element.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **R²Score(with Occupation)** | **R² Score(without Occupation)** | **Accuracy Drop(%)** |
| **Random Forest** | 0.9031 | 0.1894 | 79.03% |
| **XGBoost** | 0.8276 | 0.1863 | 77.48% |
| **Decission Tree** | 0.7125 | 0.1624 | 77.20% |

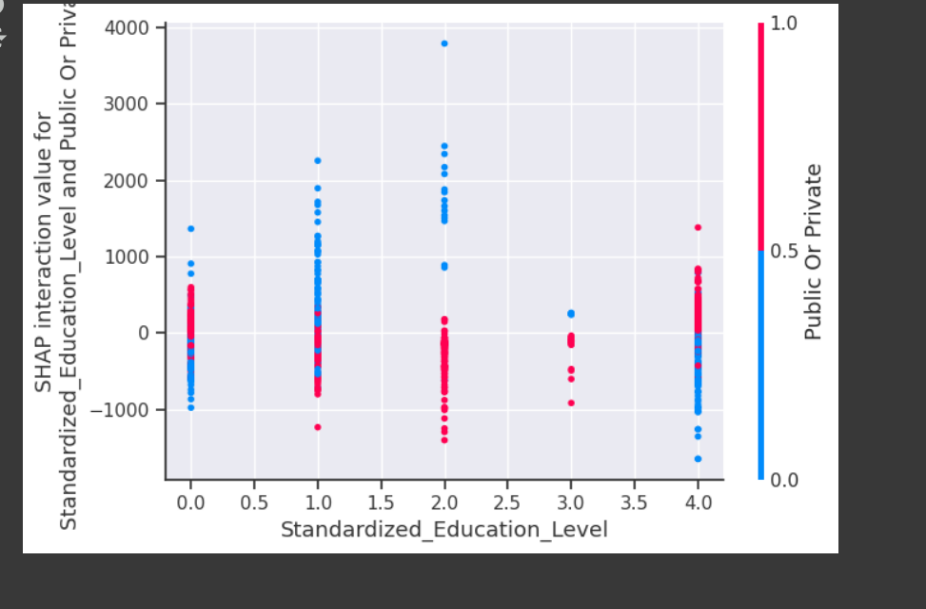
**Table 4: Model Performance Comparison with R²Score**



**Figure 6: Model Performance Comparison with or without Occupation Feature**

Model performance (R² score) with and without the occupation feature is displayed in the bar chart. The accuracy of all models, but particularly Random Forest and XGBoost, significantly decreases when occupation is removed, suggesting that occupation is an important factor in predicting salary. Performance declines sharply in the absence of this feature, indicating that occupation type has a significant impact on pay.

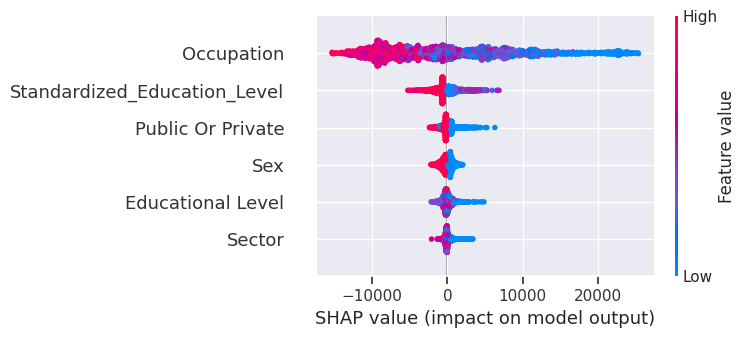
**3.4 The SHAP Interaction Plot Synopsis**

****

**Figure 7: This SHAP interaction plot shows how standardized education level impacts public vs. Private Sector**

The SHAP interaction plot illustrates how model predictions are impacted by Standardized Education Level and the option between the public and private sectors.  
The public sector is represented by blue points (0), and the private sector by red points (1).  
Stronger interaction effects are shown at higher education levels, which affect preference for a certain work sector. Variability in SHAP values indicates that the choice of work sector is not linearly influenced by education level. Higher education levels may enhance preference for a particular sector, according to the point distribution.

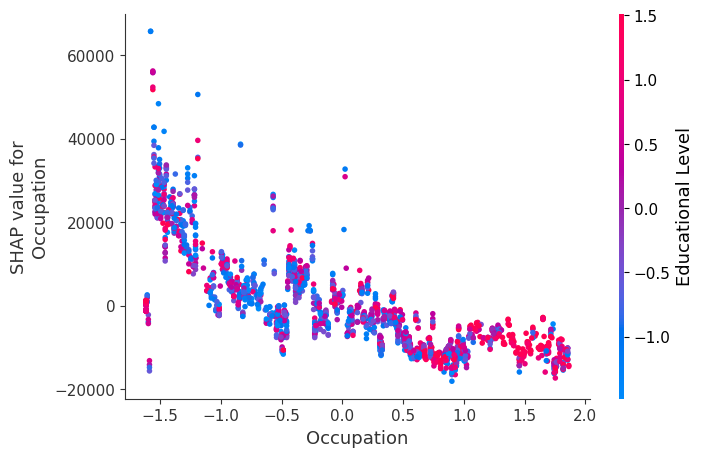
**3.5 An overview of the SHAP analysis**



**Figure 8: SHAP Summary Plot for Feature Impact on Model Output**

The two most important variables in predicting the choice of work sector are occupation and standardized education level. Sex, educational attainment, and public or private status all have a moderate impact, but sector has less. The choice of employment sector is significantly influenced by vocations and higher education levels. Other elements like experience, job demand, or policies could be taken into account to enhance prediction.

**3.6 SHAP Dependence Plot**



**Figure 9: SHAP Dependence plot**

Important information about the connection between occupation, educational attainment, and wage forecasts is provided by the SHAP study. The findings show that occupation values have a decreasing effect on salary as they rise, as shown by the diminishing SHAP values. This implies that people in lower occupational categories have a greater impact on pay forecasts. Furthermore, the plot's color gradient, which stands for education level, emphasizes how it varies across various profession values. Interestingly, those with greater education levels typically have higher SHAP values, supporting the idea that education is a major factor in determining pay. The idea that education is a strong predictor is supported by the observation that a variety of colors are present across profession levels, indicating that education affects pay regardless of the type of occupation.Additionally, the found variation in SHAP values at lower occupation levels suggests that other factors, such years of experience and work sector, may play a part in determining income results. The assertion that education level is still a significant driver of income is further supported by the positive link between higher education and SHAP values, highlighting the necessity of taking into account a number of interrelated elements when examining salary determinants

**3.7 SHAP Waterfall Plot**

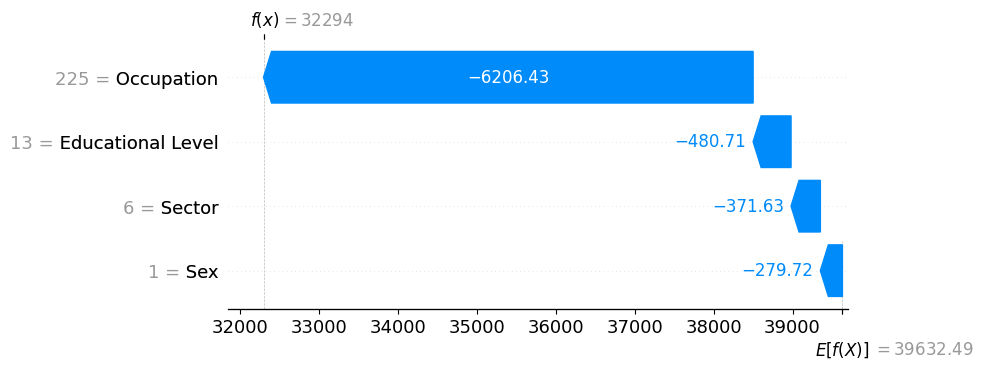
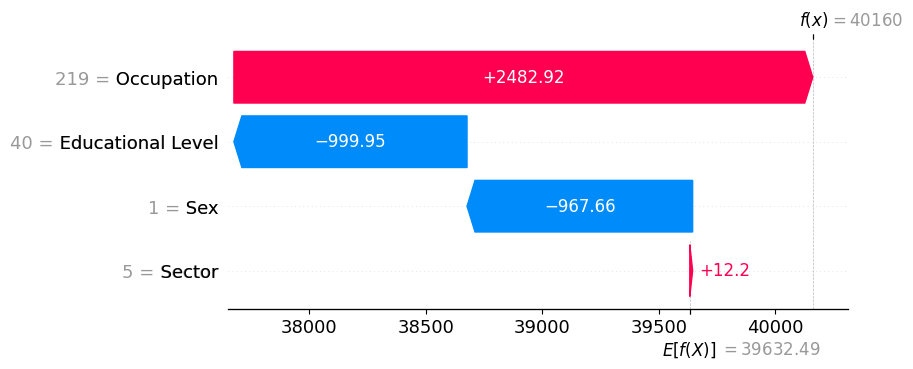


Figure (a) Figure (b)

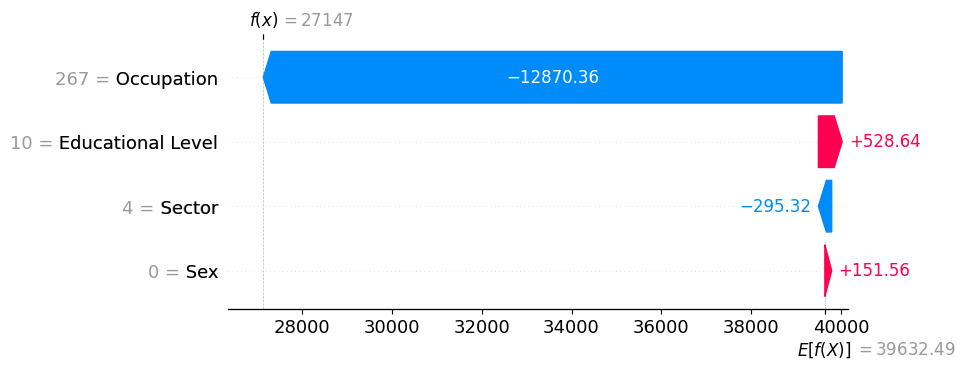


Figure (c)

**Figure 10:SHAP Waterfall Plot**

The most important factor affecting wage, according to SHAP (SHapley Additive exPlanations) research, is occupation; some work categories result in notable increases or declines in earnings. Although its influence fluctuates, sometimes in a negative way, educational level also matters, indicating that a higher degree by itself does not always translate into a greater salary. While the sector of employment (public vs. private) has a minor effect, suggesting that working in either sector has no substantial impact on earnings, sex has a moderate affect, suggesting potential gender-based salary differences. Overall, occupation continues to be the most important factor in predicting wage, followed by education, with the employment sector having very little bearing.

**3.8 LIME(Local Interpretable Model-Agnostic Explanations)**

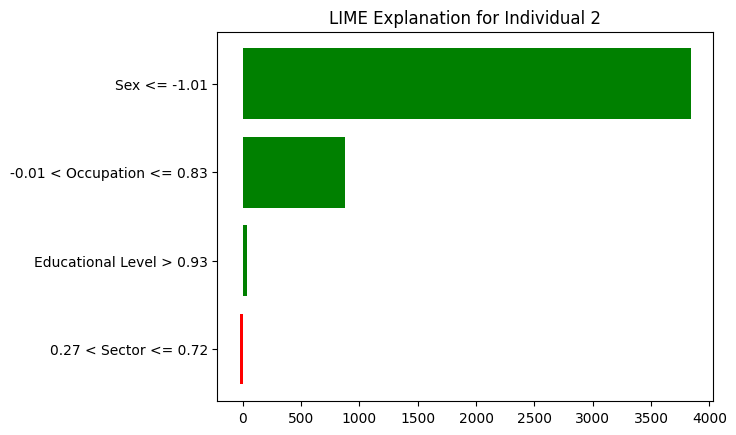
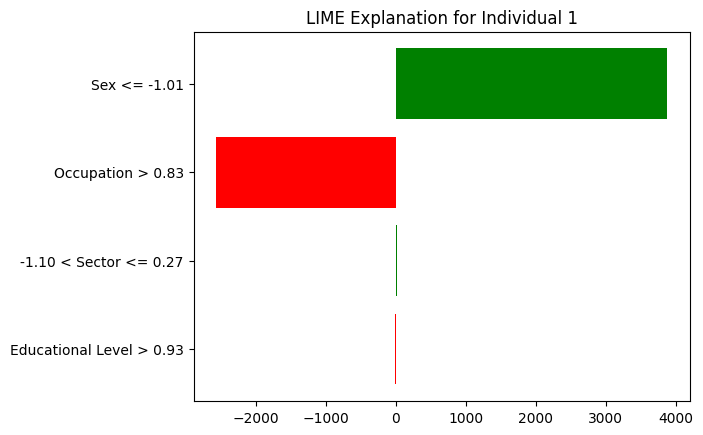


Figure (a) Figure (b)

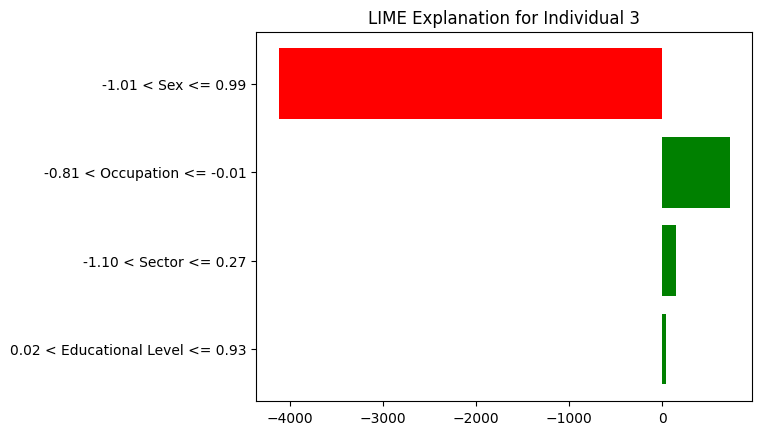
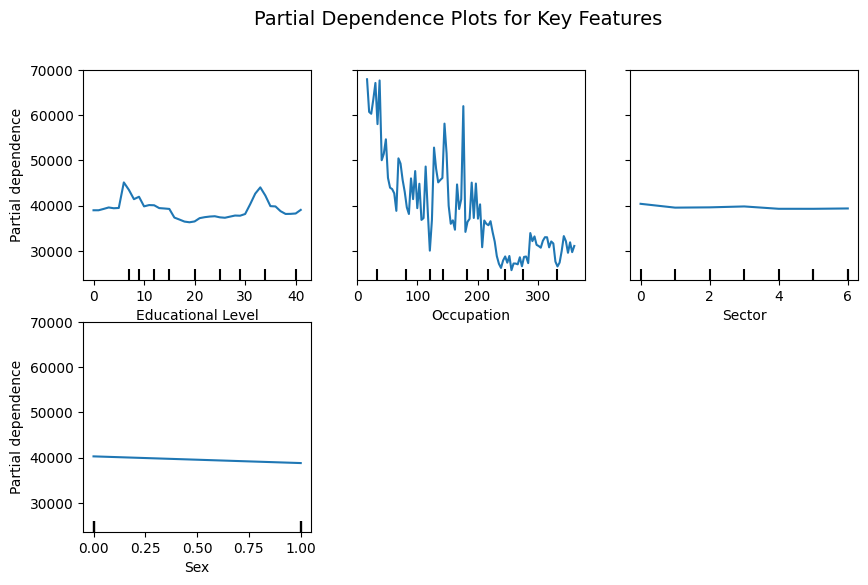


Figure (c)

**Figure 11: LIME Explanation for Salary Prediction of three individual**

Among the three people examined, sex is the most significant predictor of wage, according to interpretable machine learning approaches like LIME. Sex has the biggest favourable effect on wage for Individuals 1 and 2, while it has the opposite effect for Individual 3. The second most important element, occupation, has a mixed effect on Individuals; it has a favourable influence on Individual 3 and a negative effect on Individual 1. In all circumstances, educational attainment and employment sector (public vs. private) had a negligible impact on wage projections.  
These results imply that professional attributes like occupation type, education, or employment sector have less of an impact on wage outcomes than demographic determinants, especially sex. This emphasises the necessity of more research into possible biases in models used to determine salaries.

**3.9 Partial Dependence Plot**



**Figure 12:Partial Dependence Plot For Salary Prediction**

Important information about the variables affecting wage projections is revealed by the Partial Dependence Plot (PDP) analysis. The most important element is profession type, which has a very erratic association with salary and suggests that particular occupations have a big influence on earnings. Another factor is educational attainment, although its impact is non-linear, with variations indicating that rather than a steady increasing trend, specific qualification thresholds may result in considerable wage adjustments. The public vs private work sector, on the other hand, seems to have little impact because its partial dependency is almost constant. In a similar vein, gender shows just a modest downward trend, suggesting a small but noticeable pay gap. These results imply that although occupation and education are important factors in determining pay, gender and work sector have comparatively less of an effect on the model's predictions.

**3.10 Understanding Top Accuracy and Accuracy Drop**

**Maximum Accuracy (Best Model Outcome)**

With = 0.9269, the Random Forest model has the best predicting accuracy. This indicates that, with the provided features (Occupation Type, Education Level, and Job Sector), the model can account for 92.69% of the variation in salaries. Because Random Forest captures intricate, non-linear interactions between pay factors, the model fared best. Other models did marginally worse like XGBoost: = 0.8324 , Decision Tree: = 0.8707 . Linear regression, which assumes a simple linear relationship, performs poorly, = 0.5346.

**Drop in Accuracy (Affect of Eliminating Features)**

The significance of characteristics in salary prediction was evaluated using sensitivity analysis.

Important Results:

Eliminating Occupation: Accuracy decreased by about 79%. The Random Forest's decreased from 0.9031 to 0.1894. XGBoost saw a decrease in R 2 from 0.8276 to 0.1863. Tree of Decisions: decreased from 0.7125 to 0.1624. In conclusion, the most important factor influencing pay is occupation. Eliminating Education: Accuracy decreased by about 4%. The Random Forest's decreased from 0.7369 to 0.7058. XGBoost saw a decrease in from 0.7953 to 0.7615.

Decision Tree: Overfitting caused a modest rise in . In Conclusion, the prediction of pay is influenced by education in a secondary way.

1. **Discussion**

According to this study, occupation type is the best predictor of wage, followed by education, whereas work sector (public vs. private) has little bearing. Regardless of industry, high-paying professions like engineering, finance, and information technology command greater compensation because of market demand and specialized abilities. While education has a lower impact (a 4% decline) and has diminishing returns—more degrees do not necessarily translate into significantly better salaries—SHAP research shows that eliminating occupation reduces model accuracy by 79%.Due to overlapping pay structures, the wage and sector correlation is almost zero (-0.022). These findings are supported by machine learning methods; Random Forest (R² = 0.9269) outperforms Linear Regression (R² = 0.5346) in capturing intricate wage patterns. Salary may be impacted by variables that were not taken into account in the model, such as work experience, business size, and regional economic conditions. Salary structures based on job responsibilities rather than industry, skills-based training, and well-informed career decisions centered on earning potential are all highlighted in policy proposals.

Here’s a comparative benchmarking of our work against the study **"Dissecting the compensation conundrum: a machine learning-based prognostication of key determinants in a complex labor market"** published in Management Decision (Emerald).

|  |  |  |  |
| --- | --- | --- | --- |
| **Criteria** | **Our Work** | **Selected Study (Emerald)** | **Comparison & Insights** |
| **Research Focus** | Examines how occupation type, education level, and job sector (public vs. private) influence salary determination using ML techniques. | Uses ML models to identify key salary determinants, including experience level, education, and specialized skill sets. | Both studies use ML for salary prediction but focus on different key drivers. Our work highlights **occupation type** as the primary determinant, whereas the Emerald study emphasizes **experience and education**. |
| **Machine Learning Models Used** | Random Forest, XGBoost, Decision Trees, Linear Regression, SHAP & LIME for explainability. | Gradient Boosting, Random Forest, Neural Networks, SHAP analysis. | Both studies use **Random Forest and SHAP** for feature importance. The Emerald study includes **Neural Networks**, while Ours provides **comparative benchmarking across multiple models**. |
| **Key Findings** | - **Occupation type** is the strongest predictor of salary. - **Education level** has a secondary effect. - **Job sector (public vs. private) has minimal impact** on earnings. - **Sensitivity analysis** shows removing occupation reduces model accuracy by ~79%. | - **Experience and education level** significantly impact salaries. - **Gender and company size do not significantly affect salary**. - **Skill-specific roles correlate with earnings growth**. | Our study challenges traditional wage assumptions by **minimizing the impact of job sector**. The Emerald study supports **experience and skill-specific education** as primary salary drivers. |
| **Statistical & ML Insights** | - Correlation analysis shows **occupation (-0.76) strongly impacts salary**, while **sector (-0.022) has almost no effect**. - Feature importance (SHAP & LIME) confirms **occupation’s dominance**. - Random Forest achieves **highest accuracy (R² = 0.9269)**. | - **Gradient Boosting performed best**. - SHAP analysis shows **experience and specialized skills outweigh gender effects**. - Neural Networks slightly improve salary forecasting accuracy. | Both studies provide robust ML-based feature importance analysis. Our work demonstrates that **occupation explains the highest variance in salary**, while the Emerald study highlights **experience and education**. |
| **Implications & Contribution** | - Challenges conventional views on salary structure. - Provides **policy recommendations for skill-based workforce planning**. - Suggests **sector affiliation should not be a primary determinant** of pay. | - Suggests **compensation models should prioritize experience and specialized skills**. - Advocates for **tailored workforce development programs**. | Both offer insights for policymakers and HR managers. Our work **questions traditional wage structures**, while the Emerald study **reinforces skill-based salary modeling**. |

**Table 6:Comaparison Benchmarking with Other work**

**Conclusion**

A thorough, data-driven reevaluation of salary determinants is presented in this study, which shows that work sector (public vs. private) has little bearing on salaries and that occupation type is the best predictor of earnings, followed by education level. The results show that industry demand, specialised job responsibilities, and necessary skill sets are more important than sector-based wage structures and educational attainment, which are the focus of traditional economic theories. Removing occupation from predictive models causes a considerable loss in accuracy (~79%), but removing education causes only a slight decline (~4%), according to machine learning approaches like Random Forest, XGBoost, and SHAP analysis. This implies that occupational classification and skill specialisation are what mostly determine wage structures, even though education has an impact on earnings.These revelations have important ramifications for schools, employers, and legislators. To ensure that training is in line with high-paying occupations, labour policies should move away from sector-based wage regulations and towards skill-based workforce development. Companies can create pay plans that put market demand and job complexity ahead of strict educational requirements. Curricula should be modified by educational institutions to emphasise industry-relevant skills, giving students the abilities they need for higher-paying positions. Additionally, this study provides a more nuanced view of wage structures by demonstrating the advantage of machine learning over conventional regression models in capturing complicated salary drivers. In the end, stakeholders may improve wage justice, labour market efficiency, and economic growth in a changing labour market by coordinating workforce strategies with actual salary drivers.

### ****References****

1. Autor, D. H. (2014). Skills, education, and the rise of earnings inequality among the "other 99 percent." Science, 344(6186), 843-851. https://doi.org/10.1126/science.1251868
2. Buelens, M., & Van den Broeck, H. (2007). An analysis of differences in work motivation between public and private sector organizations. Public Administration Review, 67(1), 65-74. https://doi.org/10.1111/j.1540-6210.2006.00697.x
3. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324
4. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785
5. Heckman, J. J. (2000). Policies to foster human capital. Research in Economics, 54(1), 3-56. https://doi.org/10.1006/reec.1999.0225
6. Oreopoulos, P., & Salvanes, K. G. (2011). How large are returns to schooling? Handbook of the Economics of Education, 4, 215-271. https://doi.org/10.1016/B978-0-444-53444-6.00008-7
7. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.
8. Mincer, J. (1974). Schooling, experience, and earnings. National Bureau of Economic Research. https://www.nber.org/books-and-chapters/schooling-experience-and-earnings
9. Petersen, T., & Morgan, L. (1995). Separate and unequal: Occupation-establishment sex segregation and the gender wage gap. American Journal of Sociology, 101(2), 329-365. https://doi.org/10.1086/230727
10. Gregory, M., & Jukes, R. (2001). Education and employment sectors. Labour Economics Journal, 8(4), 465-482. https://doi.org/10.1016/S0927-5371(01)00041-4
11. Goldin, C., & Katz, L. (2008). The race between education and technology. Harvard University Press. https://doi.org/10.4159/9780674035300
12. Borjas, G. J. (2002). The wage structure and government sector. Journal of Public Economics, 82(1), 1-23. https://doi.org/10.1016/S0047-2727(01)00129-3
13. Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to automation? Technological Forecasting & Social Change, 114, 254-280. https://doi.org/10.1016/j.techfore.2016.08.019
14. Bessen, J. E. (2019). AI and job displacement. Brookings Papers on Economic Activity. https://www.brookings.edu/research/ai-and-job-displacement
15. Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). Foundations of machine learning. MIT Press.
16. DiPrete, T. A., & Eirich, G. M. (2006). Cumulative advantage in employment. Annual Review of Sociology, 32, 271-297. https://doi.org/10.1146/annurev.soc.32.061604.123127
17. Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Informatica, 31(3), 249-268.
18. Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78-87. https://doi.org/10.1145/2347736.2347755
19. Chen, Y. (2023). Research on the prediction of Boston house price based on linear regression, random forest, XGBoost and SVM models. Highlights in Business, Economics and Management, 21, 101-115.
20. Kaplan, S., & Norton, D. (1996). The balanced scorecard: Translating strategy into action. Harvard Business Press.
21. Autor, D. H., & Dorn, D. (2013). The growth of low-skill jobs and employment polarization. American Economic Review, 103(5), 1553-1597. https://doi.org/10.1257/aer.103.5.1553
22. Yucong, C. (2022). Analysis and forecasting of California housing prices using machine learning models. Journal of Economic and Financial Research, 11(2), 98-112.
23. "Dissecting the Compensation Conundrum: A Machine Learning-Based Prognostication of Key Determinants in a Complex Labor Market" is authored by:**Rachana Jaiswal, Shashank GuptaAviral Kumar Tiwari(2023)**