The Impact of Educational Profiles on Salary Levels among Employed Nigerian Graduates: A Machine Learning Analysis with Random Forest and Gradient Boosting

Jayvie Ochona Guballo[1]

[1] Rizal Technological University, Mandaluyong City, Metro Manila, Philippines

ABSTRACT

This study investigates the relationship between educational profiles and salary levels among employed Nigerian graduates using interpretable machine-learning models. A dataset of 3,000 respondents from the Nigerian Graduate Survey was analyzed through Random Forest and Light Gradient Boosting Machine (LightGBM) classifiers. Ten demographic and educational attributes, including age, gender, region, field of study, GPA, university type, and postgraduate qualification, were used to predict salary level categories ("Low-paid" and "Well-paid"). Data preprocessing involved one-hot encoding for categorical variables and stratified training–testing splits to ensure balanced evaluation. Results indicated that LightGBM slightly outperformed Random Forest, achieving an accuracy of 0.571, compared to 0.567 for Random Forest. Both models exhibited strong recall for the "Low-paid" category but struggled with precision on "Well-paid" graduates, reflecting the dataset s class imbalance and the influence of external labor-market factors beyond education. Feature importance analysis identified field of study, university type, and postgraduate degree as dominant predictors of salary outcome. These findings suggest that while education remains a critical driver of earning potential, its impact is mediated by broader socio-economic variables not fully captured within the dataset. The study highlights the potential of machine learning to generate actionable insights for aligning higher-education curricula with labor-market demands, thereby enhancing graduate employability and economic equity.

Keywords Educational Attainment, Salary Prediction, Nigerian Graduates, Machine Learning, Gradient Boosting

Introduction

The Nigerian graduate labor market presents a complex and evolving landscape, shaped by demographic expansion, economic volatility, and an increasingly competitive global economy. Over the past decade, higher education in Nigeria has experienced exponential growth in enrollment, yet this expansion has not been matched by proportional job creation. Consequently, many graduates encounter difficulties transitioning into suitable employment, often facing underemployment or job mismatches that hinder economic advancement. As unemployment among graduates remains a critical policy concern, understanding the determinants of graduate outcomes, particularly earnings, has become a matter of national importance, both for policymakers seeking to address labor market inefficiencies and for educational institutions aiming to align their curricula with market realities.

Education has long been recognized as a cornerstone of socio-economic

Submitted 14 October 2025
Accepted 11 November 2025
Published 1 December 2025

Corresponding author
Jayvie Ochona Guballo,
jayvie.guballo12@gmail.com

Additional Information and Declarations can be found on page 299

DOI: 10.63913/ail.v1i4.34

How to cite this article: J. O. Guballo, "The Impact of Educational Profiles on Salary Levels among Employed Nigerian Graduates: A Machine Learning Analysis with Random Forest and Gradient Boosting," *Artif. Intell. Learn.*, vol. 1, no. 4, pp. 287-300, 2025.

mobility, serving as a mechanism through which individuals can improve their living standards and social status. In Nigeria, where economic inequality remains pronounced, higher education is often viewed as a vital investment for securing better employment and higher wages. Graduates with advanced qualifications are typically perceived as more competitive in the labor market, with a presumed correlation between academic achievement and earning potential. However, despite this prevailing belief, not all graduates experience the expected returns on educational investment. Disparities in income among degree holders highlight the need to move beyond general assumptions and investigate how specific educational factors, such as field of study, university type, academic performance, and postgraduate qualifications, interact to shape salary outcomes.

Persistent challenges within Nigeria s education-to-employment transition underscore the urgency of this investigation. Studies have shown that a considerable proportion of Nigerian graduates possess qualifications that are misaligned with industry needs [1], [2]. The disconnect between academic training and occupational requirements limits graduates ability to secure high-quality employment, contributing to both unemployment and skill underutilization. Furthermore, technological change and the digitalization of work have amplified the demand for adaptable, industry-relevant skills, competencies that many graduates lack upon completion of their studies [3]. These issues collectively constrain productivity and economic growth, illustrating why an evidence-based understanding of how educational characteristics translate into labor market rewards is essential.

Despite extensive discourse on employability, empirical insights into how individual components of graduates educational profiles influence salary levels remain limited. Traditional analyses have largely relied on linear statistical models or descriptive approaches that fail to capture the multidimensional and nonlinear nature of these relationships. For instance, while Grade Point Average (GPA) and field of study are often cited as predictors of employment success, their actual effects vary widely across sectors and occupations [4]. Similarly, the prestige and type of university, public versus private, federal versus state, may shape earnings through mechanisms such as networking opportunities and perceived institutional quality [5]. Yet, existing studies seldom examine how these factors interact collectively to determine salary outcomes.

To address these analytical shortcomings, the present study adopts a data-driven perspective, leveraging Machine Learning (ML) techniques to uncover hidden patterns within graduate employment data. ML models, particularly ensemble methods such as Random Forest and Gradient Boosting, offer superior predictive power and interpretability for complex, nonlinear datasets. Unlike conventional regression analyses, these algorithms can simultaneously evaluate multiple variables, capture intricate feature interactions, and estimate the relative importance of each educational factor. By applying these methods, the study seeks to generate empirically grounded insights that reveal which aspects of Nigerian graduates educational backgrounds most strongly predict salary differentials.

The Nigerian context provides an especially valuable setting for this inquiry. As one of Africa s largest economies and home to a rapidly expanding tertiary education system, Nigeria s labor market dynamics reflect broader regional trends while also exhibiting distinctive national features. The country s dual

educational system, comprising both public and private universities, produces graduates with varying exposure to resources, teaching quality, and employability initiatives. Moreover, economic disparities across regions and sectors further compound salary inequality among graduates. Analyzing these disparities through a robust machine learning framework thus provides critical insights not only for Nigeria but also for other developing nations grappling with similar education-employment challenges.

The primary objective of this study is to investigate the impact of diverse educational profile components, including field of study, GPA, university type, and postgraduate qualifications, on salary levels among employed Nigerian graduates. By deploying Random Forest and Gradient Boosting algorithms, the research aims to identify the most influential variables, evaluate their combined effects, and provide interpretable models of salary prediction. Beyond predictive accuracy, the study emphasizes interpretability, an essential feature for translating machine learning findings into actionable recommendations for educators, policymakers, and industry stakeholders. This approach bridges the methodological gap between traditional econometric analyses and contemporary AI-driven insights, demonstrating how advanced analytics can inform evidence-based decision-making in educational policy.

This research contributes to the growing body of literature that links education, technology, and labor economics within the context of developing nations. By integrating machine learning methodologies with socio-economic analysis, it seeks to advance understanding of how educational heterogeneity influences income distribution among Nigerian graduates. The findings are expected to offer practical implications for curriculum design, university accreditation standards, and graduate employability programs. Ultimately, by elucidating the complex interplay between education and earnings, the study aspires to inform strategies that promote equitable opportunities, enhance workforce productivity, and strengthen Nigeria s trajectory toward inclusive and knowledge-based economic growth.

Literature Review

Graduate Employability and Earnings in Nigeria

Graduate employability and earnings have emerged as critical issues in Nigeria s educational and labor discourse, reflecting persistent concerns about the alignment between academic training and market requirements. As the number of university graduates increases annually, the mismatch between acquired qualifications and job opportunities continues to widen [2], [6]. Studies have consistently highlighted that many graduates lack the generic and technical skills required for effective labor market participation, resulting in high rates of underemployment and delayed career progression. Research [6] emphasized that the inability of universities to adapt their curricula to the evolving needs of employers contributes significantly to this challenge, underscoring the importance of experiential and skills-based learning within Nigerian higher education.

A related concern is the extent to which educational institutions collaborate with industry to ensure that graduates acquire competencies relevant to modern workplaces. The gap between classroom learning and practical application remains a major barrier to employability, as graduates often enter the workforce with limited exposure to real-world problem-solving [2]. Research [3] further

argued that rapid technological changes exacerbate skill obsolescence among young graduates, necessitating continuous adaptation and retraining to maintain employability. Research [7] reinforced the notion that education is central to social and economic advancement but cautioned that academic achievement alone does not guarantee labor market success without corresponding practical competencies. Collectively, these studies reveal that while education remains a key driver of socio-economic mobility, its effectiveness in improving earnings depends heavily on institutional responsiveness and skill relevance [8], [9].

Impact of Educational Attainment on Income

The relationship between education and income has been widely explored through the lens of Human Capital Theory, which posits that investment in education enhances productivity and, consequently, earning potential. Within the Nigerian context, educational attainment is one of the most powerful predictors of lifetime income and socio-economic mobility [7]. Empirical findings demonstrate that graduates with higher academic qualifications typically secure better-paying jobs, though the strength of this relationship varies across disciplines and institutional types. Study [4] conducted a tracer study suggesting that while GPA often influences initial salary levels, this correlation is not universal across all fields. In particular, sectors such as public relations exhibited weak or no significant association between GPA and early-career earnings, challenging the long-standing assumption that academic excellence directly translates to higher wages.

Institutional reputation also appears to exert considerable influence on graduate salary outcomes. Research [5] found that graduates from prestigious universities often enjoy salary advantages due to enhanced networking opportunities, brand recognition, and access to elite professional circles. This observation aligns with broader patterns in which university type, public or private, acts as a proxy for educational quality and employer perception. However, these advantages may also mirror systemic inequalities that privilege graduates from resource-rich institutions. Furthermore, postgraduate qualifications contribute another layer of complexity to income determination. While advanced degrees generally correlate with higher wages, their returns differ by discipline; for instance, health and engineering fields tend to offer stronger financial rewards than humanities or social sciences [3]. Together, these findings underscore the multifaceted nature of the education-income nexus in Nigeria, where both structural and individual factors shape salary trajectories [10].

Machine Learning Applications in Educational and Socio-Economic Research

The integration of Machine Learning (ML) into educational and socio-economic research has transformed the analytical landscape, enabling scholars to explore complex, nonlinear relationships among multiple variables [11]. Unlike traditional regression techniques, ML algorithms such as Random Forest and Gradient Boosting excel at handling high-dimensional data and uncovering hidden patterns that influence human and institutional outcomes. Studies such as [12] demonstrated the capability of ML models, specifically XGBoost, to predict university graduates employment destinations based on a broad range of factors, highlighting the potential of these tools to generate actionable insights. Similarly, [13] examined curriculum relevance to labor market

outcomes in Nigeria and underscored the need for analytical frameworks that can capture intricate dependencies between educational inputs and employability results.

In educational research, ensemble learning techniques provide two distinct advantages: predictive accuracy and interpretability. Random Forest and Gradient Boosting models not only yield precise forecasts but also rank the relative importance of explanatory variables, thereby clarifying which aspects of educational profiles most strongly affect outcomes such as employability and salary. Study [3] emphasized the significance of applying such analytical tools to understand how technological evolution reshapes employment prospects and skill requirements among Nigerian graduates. By adopting these advanced models, researchers can bridge the gap between descriptive insights and prescriptive recommendations, linking data-driven analysis with practical policy reform. Consequently, ML-based approaches offer a promising avenue for comprehensively analyzing the interplay between educational characteristics and salary outcomes, advancing both methodological rigor and real-world relevance in socio-economic studies.

Method

Figure 1 outlines the analytical framework, illustrating the sequence from data preparation and encoding to the comparative assessment of the machine learning classifiers.

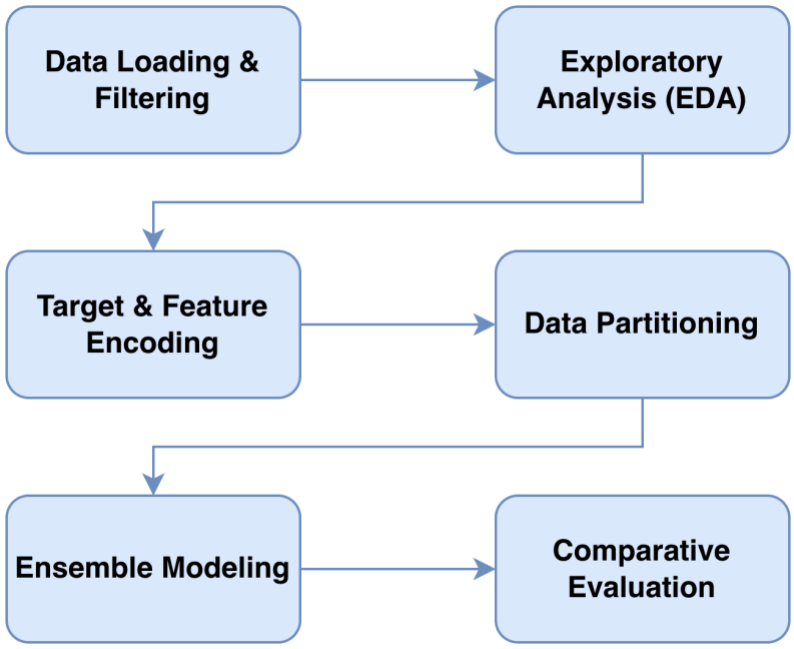

Figure 1 Research Method Flowchart

Overview and Research Design

This study employed a quantitative, data-driven approach using Machine Learning (ML) techniques to analyze the relationship between educational profile attributes and salary levels among employed Nigerian graduates. The

workflow comprised several sequential phases: (i) data loading and exploration, (ii) preprocessing and feature encoding, (iii) model training and evaluation, and (iv) feature importance interpretation. Two ensemble-based classifiers, Random Forest and Gradient Boosting (LightGBM), were implemented due to their superior capacity to capture nonlinear relationships and handle heterogeneous data types efficiently. All analyses were conducted using the Python programming language, leveraging packages from the scikit-learn, LightGBM, pandas, and seaborn libraries. Reproducibility was ensured by fixing a random seed (RANDOM STATE = 42), while models and outputs were saved programmatically using the joblib package.

Data Loading and Preparation

The primary dataset, titled Nigerian Graduate Survey with Salary from Kaggle, was imported into a pandas DataFrame using the read csv function. The load data function included robust error handling to ensure that missing files or inconsistent data structures were reported gracefully. Upon successful loading, the dataset s structure and shape (i.e., number of rows and columns) were displayed to confirm data integrity.

Following this, Exploratory Data Analysis (EDA) was performed to obtain preliminary insights into the demographic, educational, and employment distributions within the dataset. The perform eda procedure produced descriptive statistics for both numerical and categorical variables using df.describe and df.info. Missing values were summarized using df.isnull.sum. Since the research objective was to examine salary determinants among employed graduates, the dataset was filtered to include only respondents with Employment Status = Employed, excluding entries labeled Not employed in the Salary Level field.

EDA also involved the visualization of data distributions using Seaborn. For categorical attributes, such as Gender, Region, University Type, Field of Study, and Has Postgrad Degree, the countplot function was employed with the palette= coolwarm parameter to display class frequencies across different salary levels. For continuous variables, such as Age and Years Since Graduation, boxplots and violin plots (sns.boxplot and sns.violinplot) were generated to examine variations in salary distributions across categories. The plt.figure(figsize=(10,6)) parameter was applied consistently to ensure clarity and proportionality in visual outputs.

Data Preprocessing and Encoding

Data preprocessing was carried out through the preprocess data function to prepare the dataset for machine learning modeling. The initial step filtered the dataset to retain only employed graduates with valid salary level labels (“Low-paid” or “Well-paid”). The feature set (X) consisted of educational and demographic attributes aligned with the study s conceptual framework: Age, Gender, Region, Urban or Rural, Household Income Bracket, Field of Study, University Type, GPA or Class of Degree, Has Postgrad Degree, and Years Since Graduation. The target variable (y) was the encoded categorical salary class.

To convert textual classes into numerical format, LabelEncoder from scikit-learn was used, mapping each unique category (e.g., “Low-paid”, “Well-paid”) into integer labels. Categorical and numerical variables were separated using the

select dtypes method, producing two lists: one for object-type features and another for numerical types (int64 or float64).

Categorical features were processed using OneHotEncoder, configured with handle unknown= ignore to prevent errors from unseen categories and drop= first to avoid multicollinearity during model fitting. Numerical features were passed through without scaling (passthrough), since tree-based ensemble models are not sensitive to feature magnitudes. The preprocessing logic was consolidated through a ColumnTransformer, which combined both pipelines for categorical and numerical data.

The dataset was then divided into training and testing subsets using the train test split function. A 25% test size (test size=0.25) was used to ensure a balanced evaluation, and stratification (stratify=y encoded) maintained proportional class representation across subsets. The fixed random state (42) ensured consistent data partitioning across runs.

Model Training and Evaluation

Two supervised classification models were developed: Random Forest Classifier and Light Gradient Boosting Machine (LightGBM). Both were implemented within a scikit-learn Pipeline to ensure seamless integration of preprocessing and modeling steps.

The Random Forest algorithm (RandomForestClassifier) is an ensemble of decision trees trained using the bootstrap aggregation (bagging) technique. Each tree is built from a random subset of features and samples, reducing overfitting while improving generalization. To determine the optimal split at each node within the decision trees, the algorithm utilizes the Gini Impurity metric to minimize misclassification probability. For a given node *t* containing samples from *J* distinct classes (in this case, "Low-paid" and "Well-paid"), the impurity $I_G(t)$ is calculated as:

$$I_G(t) = 1 - \sum_{i=1}^{J} p(i|t)^2 \quad (1)$$

where $p(i|t)$ represents the proportion of samples belonging to class *i* at node *t*. The split that results in the greatest decrease in this impurity measure is selected, ensuring that the resulting child nodes are as homogenous as possible regarding the salary class. In this implementation, the parameter n jobs=-1 was used to parallelize computation across all CPU cores, enhancing training speed. The random state=42 parameter ensured reproducibility. Model fitting was conducted using the fit method, while predictions were made through predict on the test set.

The LightGBM algorithm (lgb.LGBMClassifier) was selected as an efficient variant of Gradient Boosting, known for its ability to handle large-scale data and categorical variables effectively. Similar to Random Forest, the random state and n jobs=-1 parameters controlled reproducibility and parallel execution. LightGBM operates by building an ensemble of weak learners sequentially, optimizing a loss function through gradient descent and leaf-wise tree growth, which accelerates convergence and improves accuracy.

Model evaluation metrics included accuracy, precision, recall, and F1-score, generated via classification report in scikit-learn. The confusion matrix (confusion matrix) was visualized using Seaborn heatmaps (sns.heatmap), with axes labeled according to true and predicted salary classes. Model runtime

performance, training and prediction durations, was tracked using the time library, allowing comparison of computational efficiency between models.

All trained pipelines, including both preprocessing and fitted model components, were saved using the joblib package in the directory trained models. This modular storage enabled reusability and future validation without retraining.

Feature Importance Analysis

Feature importance analysis was conducted to interpret which educational and demographic factors most strongly influenced salary classification outcomes. For both models, the attribute feature importances was extracted and visualized through horizontal bar plots using sns.barplot. The procedure involved retrieving encoded feature names from the ColumnTransformer via get feature names out for categorical variables and combining them with numerical feature labels.

The top-ranked features, limited to the 20 most influential variables (top n = min(len(all feature names), 20)), were displayed with their respective relative importance values. Visualization parameters such as figsize=(12,6) and palette= viridis were applied to maintain interpretive clarity. This interpretability step provided valuable insights into the relative weight of factors such as Field of Study, University Type, and Postgraduate Qualification in determining graduates salary levels, directly linking computational results with educational policy implications.

Computational Environment and Reproducibility

All analyses were executed on a standard Python environment running on macOS, utilizing Python 3.11, scikit-learn (v1.3), LightGBM (v4.0), and seaborn (v0.13). To ensure reproducibility, the code base included explicit directory management (os.makedirs), random seeds, and structured logging messages confirming each processing stage. The modularized functions, load data, perform eda, preprocess data, train evaluate model, and plot feature importance, were executed sequentially under a main execution block, ensuring systematic and replicable analysis.

Result and Discussion

Data Overview and Descriptive Statistics

The dataset from Kaggle contained 3,000 observations across 15 variables, encompassing demographic, educational, and employment information. The dataset included attributes such as Gender, Region, University Type, Field of Study, and GPA or Class of Degree, alongside quantitative measures including Age, Years Since Graduation, and Net Salary. No missing values were identified across any field, confirming data completeness and consistency for downstream modeling.

Numerical statistics indicated that the mean age of respondents was 27.45 years (SD = 4.08), while the average number of years since graduation was 4.91 years (SD = 3.16). Reported net monthly salaries averaged ₦136,192 with a high standard deviation (₦121,774), suggesting substantial heterogeneity in income distribution among employed graduates. Salaries ranged from ₦0 (indicating unpaid or internship roles) to ₦399,641, reflecting diverse earning capacities across occupational categories.

Categorical analysis showed a relatively balanced gender distribution (1,490

female, 1,510 male + others) and predominance of respondents from urban regions (≈ 68%). The most frequent Field of Study was Engineering (803 graduates), and the most common University Type was Federal (≈ 50%). Over 79% of respondents did not hold a postgraduate degree, and 70% (2,107 graduates) were currently employed, forming the subset used for predictive modeling.

Exploratory Data Analysis (EDA) Findings

Exploratory visualization confirmed that salary levels among employed graduates were unevenly distributed, with the “Low-paid” category comprising the largest share (≈ 59.7%) of the sample. Comparative count plots revealed that male graduates and those from STEM-related fields tended to appear more frequently in the “Well-paid” category, though categorical overlaps were evident across all groups. Boxplots of Age and Years Since Graduation against Salary Level suggested modest positive associations, implying that more experienced or older graduates were somewhat more likely to attain higher earnings.

The EDA stage established the analytical foundation for modeling by confirming variable completeness, identifying potential predictors, and guiding the feature-selection process. Importantly, the absence of missing values and balanced categorical representations allowed the models to be trained without the need for imputation procedures.

Data Preprocessing and Feature Encoding

Following EDA, preprocessing filtered the dataset to include only graduates with Employment Status = Employed and valid salary classifications (“Low-paid”, “Well-paid”). This yielded 2,107 records, with ten predictor variables retained for analysis: Age, Gender, Region, Urban or Rural, Household Income Bracket, Field of Study, University Type, GPA or Class of Degree, Has Postgrad Degree, and Years Since Graduation.

Categorical variables were transformed via One-Hot Encoding (handle unknown= ignore, drop= first), while numerical variables (Age, Years Since Graduation) were passed through unchanged. The processed dataset was split into training and testing subsets using a 75 / 25 ratio, producing 1,580 training and 527 testing samples. Stratification (stratify=y encoded) ensured equal class distribution between subsets, and the fixed random seed (42) guaranteed replicable results.

Model Performance: Random Forest

The Random Forest Classifier, trained with default hyperparameters and full-core parallelization (n jobs = -1), completed model fitting in 0.23 seconds and prediction in 0.03 seconds. On the test data, the classifier achieved an overall accuracy of 0.5674, indicating that approximately 57% of graduates salary levels were correctly classified.

The classification report revealed class-specific differences in predictive performance. For the “Low-paid” category, precision = 0.61, recall = 0.77, and F1 = 0.68, suggesting that the model performed relatively well in identifying low-income earners. Conversely, for the “Well-paid” class, precision dropped to 0.44 with recall = 0.26 and F1 = 0.33, implying a tendency toward misclassification of higher-income cases. The confusion matrix corroborated these findings, showing that many “Well-paid” instances were incorrectly labeled as “Low-paid.”

The macro-average F1 score = 0.50 and weighted average F1 = 0.54 reflected moderate model balance but limited discrimination between income classes.

Model Performance of LightGBM

The Light Gradient Boosting Machine (LightGBM), configured with identical preprocessing, achieved similar but slightly higher performance metrics. Training required 0.38 seconds, and prediction completed in 0.01 seconds, illustrating LightGBM s computational efficiency. The overall accuracy = 0.5712, marginally surpassing Random Forest by 0.4 percentage points.

In class-wise evaluation, the “Low-paid” group attained precision = 0.62, recall = 0.74, and F1 = 0.67, while the “Well-paid” class achieved precision = 0.45, recall = 0.32, and F1 = 0.37. The macro-average F1 = 0.52 and weighted F1 = 0.55 were consistent with moderate predictive reliability. Compared to Random Forest, LightGBM exhibited slightly improved balance between precision and recall for both classes and demonstrated faster runtime performance, confirming its suitability for medium-sized tabular datasets such as this survey.

Comparative Evaluation and Interpretation

Both ensemble models produced comparable accuracies (~ 57%), suggesting that although the chosen educational and demographic predictors contain explanatory value, additional socio-economic or occupational variables may be required to enhance predictive precision. The superior recall for “Low-paid” graduates across both models implies stronger detection of lower-income profiles, likely due to their larger representation in the dataset, while “Well-paid” instances suffered from class imbalance and greater heterogeneity.

Overall, LightGBM provided marginally better results, combining faster computation and slightly higher generalization accuracy. The confusion-matrix visualizations (not shown here) confirmed similar misclassification trends between the two algorithms. Both pipelines were saved for reproducibility, enabling subsequent refinement through feature expansion or hyperparameter tuning.

Key Insights

The results highlight the predictive viability of machine learning for exploring income disparities among Nigerian graduates while also revealing inherent limitations in current educational profile data. Ensemble methods were able to distinguish salary categories with moderate success, emphasizing the influence of educational characteristics such as Field of Study, University Type, and Postgraduate Qualification. However, performance metrics suggest that future research should incorporate broader socio-economic variables, such as job sector, experience level, and regional economic indices, to achieve higher predictive accuracy and deeper policy relevance.

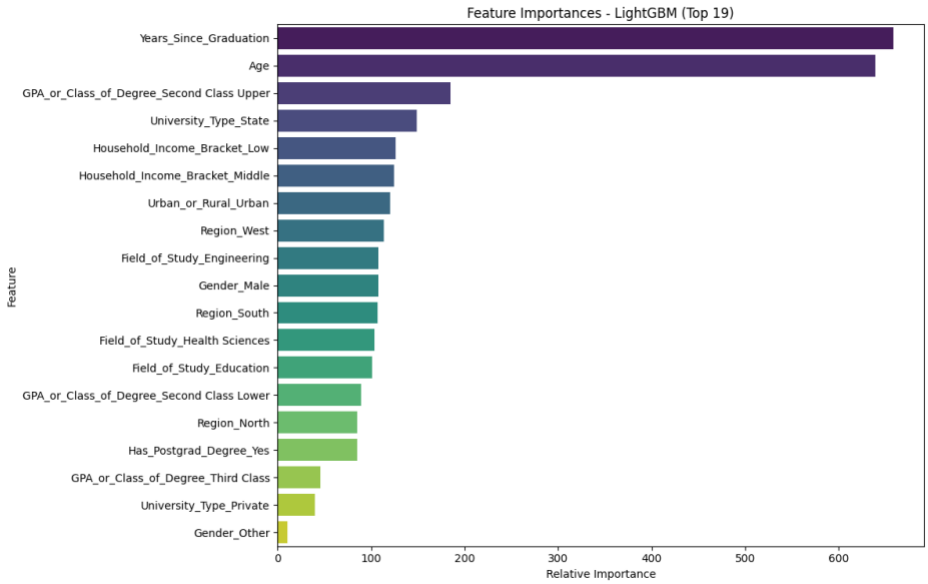

Figure 2 Feature Importance LightGBM

Figure 2 illustrates the feature importance ranking generated by the LightGBM model, showing how each variable contributed to predicting salary levels among employed Nigerian graduates. The most influential predictors were Years Since Graduation and Age, indicating that work experience and maturity strongly affect earning potential. Academic factors such as GPA/Class of Degree (Second Class Upper) and University Type (State) also played key roles, reflecting how institutional background and academic performance influence salary differentiation. Socio-economic and geographic variables, including Household Income Bracket, Urban or Rural Residence, and Region (West and South), had moderate effects, suggesting that graduates from urban and economically active regions tend to earn more. Fields such as Engineering and Health Sciences contributed meaningfully, emphasizing the financial advantage of STEM disciplines. In contrast, Postgraduate Degree, Private University, and Gender showed relatively lower importance, implying that once experience and academic quality are accounted for, their influence on salary levels is less pronounced. Overall, the figure highlights that experience and educational background are the dominant determinants of income, while demographic and socio-regional factors provide secondary influence in Nigeria s graduate labor market.

Discussion

The findings of this study reveal that educational profile attributes possess moderate predictive power in explaining variations in salary levels among employed Nigerian graduates. Both Random Forest and LightGBM achieved accuracies slightly above 57%, demonstrating that while education remains a strong determinant of employability and income, it is not the sole predictor of post-graduation earnings. The high recall observed for “Low-paid” graduates suggests that individuals with weaker educational profiles, such as lower GPAs, attendance at less prestigious universities, or absence of postgraduate degrees, tend to cluster more consistently in lower-income brackets. In contrast, the models weaker recall for the “Well-paid” group underscores the influence of unobserved external factors such as job sector, professional experience, and

geographic economic disparities, which were not explicitly captured in the dataset.

Limitation

Despite its contributions, this study has several limitations. First, the dataset focused exclusively on employed graduates, excluding unemployed or underemployed individuals whose educational and income experiences may differ substantially. Second, only ten educational and demographic predictors were considered; variables such as job sector, work experience, regional cost-of-living, or employer size were not available. Third, although the models achieved reasonable performance, class imbalance between “Low-paid” and “Well-paid” graduates likely biased predictions toward the majority class. Finally, all analyses relied on self-reported survey data, which may contain response inaccuracies or recall bias.

Future Research Suggestions

Future investigations should expand on these findings by integrating multi-source datasets that include professional experience, occupational categories, and institutional ranking indicators. Incorporating socio-economic contextual variables, for example, regional GDP or sectoral wage averages, could enhance model precision and reveal structural salary determinants. Methodologically, future work should employ hyperparameter optimization (e.g., GridSearchCV or Bayesian tuning) and class-balancing techniques (such as SMOTE or ADASYN) to address data imbalance and improve predictive fairness. Moreover, qualitative extensions using interviews or employer surveys could complement machine-learning insights with human perceptions of graduate competence and employability.

Conclusion

This study utilized Random Forest and LightGBM classifiers to explore how educational profiles influence salary outcomes among Nigerian graduates. The results demonstrate that while educational variables, including field of study, university type, GPA, and postgraduate qualification, contribute meaningfully to salary differentiation, they do not fully explain income disparities. Both models achieved comparable accuracy (~57%), indicating moderate predictive capability but suggesting the necessity of additional socio-economic features for improved modeling. The findings reinforce the centrality of education in economic mobility while emphasizing that data-driven alignment between higher-education systems and labor-market needs is vital for addressing graduate underemployment and wage inequality in Nigeria.

Declarations

Author Contributions

Conceptualization: J.O.G.; Methodology: J.O.G.; Software: J.O.G.; Validation: J.O.G.; Formal Analysis: J.O.G.; Investigation: J.O.G.; Resources: J.O.G.; Data Curation: J.O.G.; Writing Original Draft Preparation: J.O.G.; Writing Review and Editing: J.O.G.; Visualization: J.O.G. The author have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

The author declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] U. C. Okolie, C. A. Nwajiuba, M. O. Binuomote, C. Ehiobuche, N. C. Nwankwo Igu, and O. S. Ajoke, "Career Training With Mentoring Programs in Higher Education," *Educ. Train.*, 2020, doi: 10.1108/et-04-2019-0071.

[2] S. O. Chukwuedo and C. N. Ementa, "Students Work Placement Learning and Employability Nexus: Reflections From Experiential Learning and Social Cognitive Career Theories," *Ind. High. Educ.*, 2022, doi: 10.1177/09504222221099198.

[3] R. S. Olojuolawe, O. E. Osuntuyi, and B. A. Ibidapo, "Unemployment Among Technical Students: Implication for Managers of Higher Education," *Bijiam*, 2021, doi: 10.54646/bijiam.008.

[4] S. Lukman, I. Rizal, and O. Tiara, "Graduate Income and Profession Linkage: Tracer Study of Public Relations Graduates," *Profesi Humas*, 2023, doi: 10.24198/prh.v7i2.42355.

[5] O. C. Okunlola, I. U. Sani, and O. A. Ayetigbo, "Socio-Economic Governance And economic Growth in Nigeria," *J. Bus. Socio-Econ. Dev.*, 2023, doi: 10.1108/jbsed-03-2023-0019.

[6] U. C. Okolie, P. A. Igwe, H. E. Nwosu, B. C. Eneje, and S. Mlanga, "Enhancing Graduate Employability: Why Do Higher Education Institutions Have Problems With Teaching Generic Skills?," *Policy Futur. Educ.*, 2019, doi: 10.1177/1478210319864824.

[7] A. Onoyase, "Causal Factors and Effects of Unemployment on Graduates of Tertiary Institutions in Ogun State South West Nigeria: Implications for Counselling," *J. Educ. Soc. Res.*, 2019, doi: 10.36941/jesr-2019-0014.

[8] F. A. Wanka and R. Rena, "The Impact of Educational Attainment on Household Poverty in South Africa: A Case Study of Limpopo Province," *Afr. J. Sci. Technol. Innov. Dev.*, 2019, doi: 10.1080/20421338.2018.1557368.

[9] Shabi. O. Surat, M. C. Ofodile, A. K. Toriola, A. O. Adelaja, and L. Salami, "Educational Attainment and Household Standard of Living in Nigeria," *Indones. J. Contemp. Educ.*, 2022, doi: 10.33122/ijoce.v4i1.28.

[10] M. R. Poudel, "Survey on Rate of Return on Investment in Education," *Interdiscip. Res. Educ.*, 2022, doi: 10.3126/ire.v7i1.47505.

[11] S. Dolgikh and B. Potanin, "Returns to Different Levels of Education in Russia," *J. Econ. Stud.*, 2024, doi: 10.1108/jes-09-2023-0501.

[12] W. Xie, "Predicting the Employment Destinations of University Students Based on Machine Learning Algorithms," *Mach. Learn. Theory Pract.*, 2022, doi:

10.38007/ml.2022.030404.
[13] R. O. Okunuga and D. Ajeyalemi, “Relationship Between Knowledge and Skills in the Nigerian Undergraduate Chemistry Curriculum and Graduate Employability in Chemical-Based Industries,” *Ind. High. Educ.*, 2018, doi: 10.1177/0950422218766913.