Managerial Report: Business Intelligence Analysis of Income Classification

1. Executive Summary

Project Objective & Context

- This report presents a managerial-level Business Intelligence (BI) analysis based on the Adult Census Income Dataset. The objective is to assess the feasibility of an automated income-classification system to support credit scoring, loan pre-qualification, and customer segmentation for a mid-sized Kenyan financial institution.

Key Analytical Insights

- Key findings indicate that education level, income class and working hours strongly influence income classification, highlighting the role of human capital and employment stability in income profiling.

Modelling & Decision Implications

- Predictive models demonstrate acceptable performance in distinguishing income categories, indicating potential operational value. However, the choice of model involves strategic trade-offs between predictive accuracy, interpretability, and governance requirements in regulated financial environments.

Risk, Governance & Recommendation Signal

- Ethical and governance risks related to algorithmic fairness and data protection require careful mitigation before deployment. Any implementation of automated income classification should therefore be accompanied by strong governance controls, transparency measures, and contextual adaptation to Kenyan regulatory and socio-economic conditions.

2. Business Context & Analytical Objective

Institutional Importance of Income Classification

- Income classification is a critical component in modern financial decision-making. It supports key institutional functions such as credit scoring, loan pre-qualification, customer segmentation, and risk profiling, particularly in data-driven financial environments.

Analytical Objective & Scope

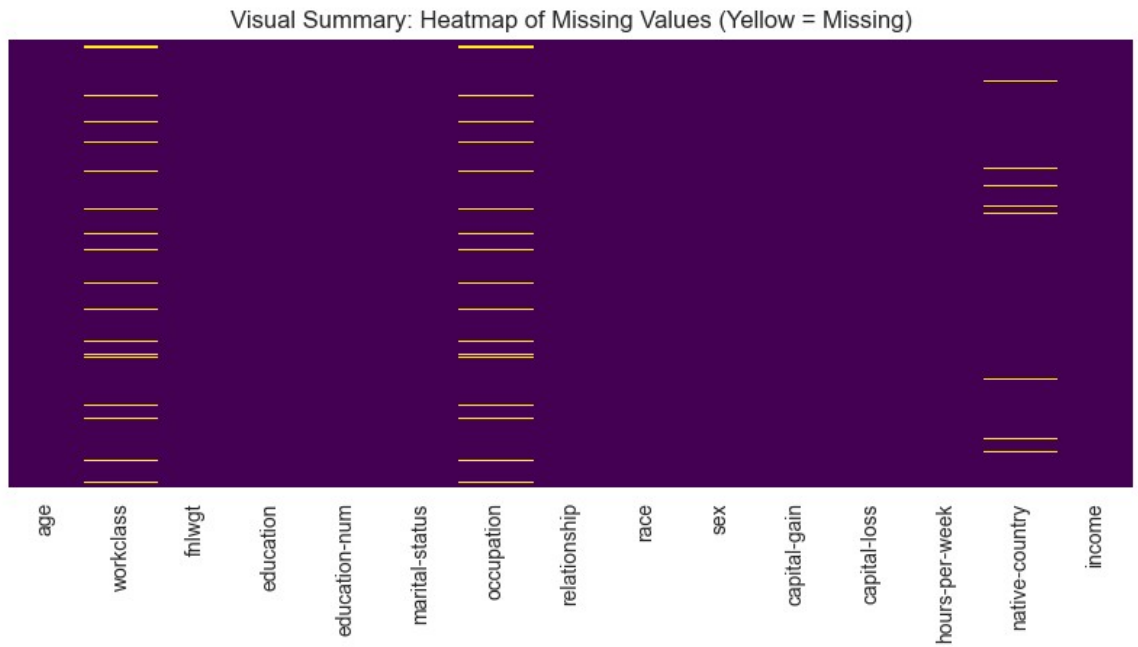
- This analysis was conducted to evaluate the feasibility of an automated income-classification system using a benchmarking dataset, with the objective of validating Business Intelligence workflows before applying similar approaches to Kenyan institutional data.

3. Data Quality Assessment & Cleaning Strategy

Data Quality Risks

- The dataset exhibited several forms of real-world data quality issues, including missing values, inconsistent categorical labels, redundant attributes, and formatting irregularities. If left unaddressed, such data quality problems can undermine the reliability of Business Intelligence outputs and lead to flawed institutional decision-making.

Missing Count	% of Total	
1836	5.64	Work class
1843	5.66	occupation
583	1.79	native-country



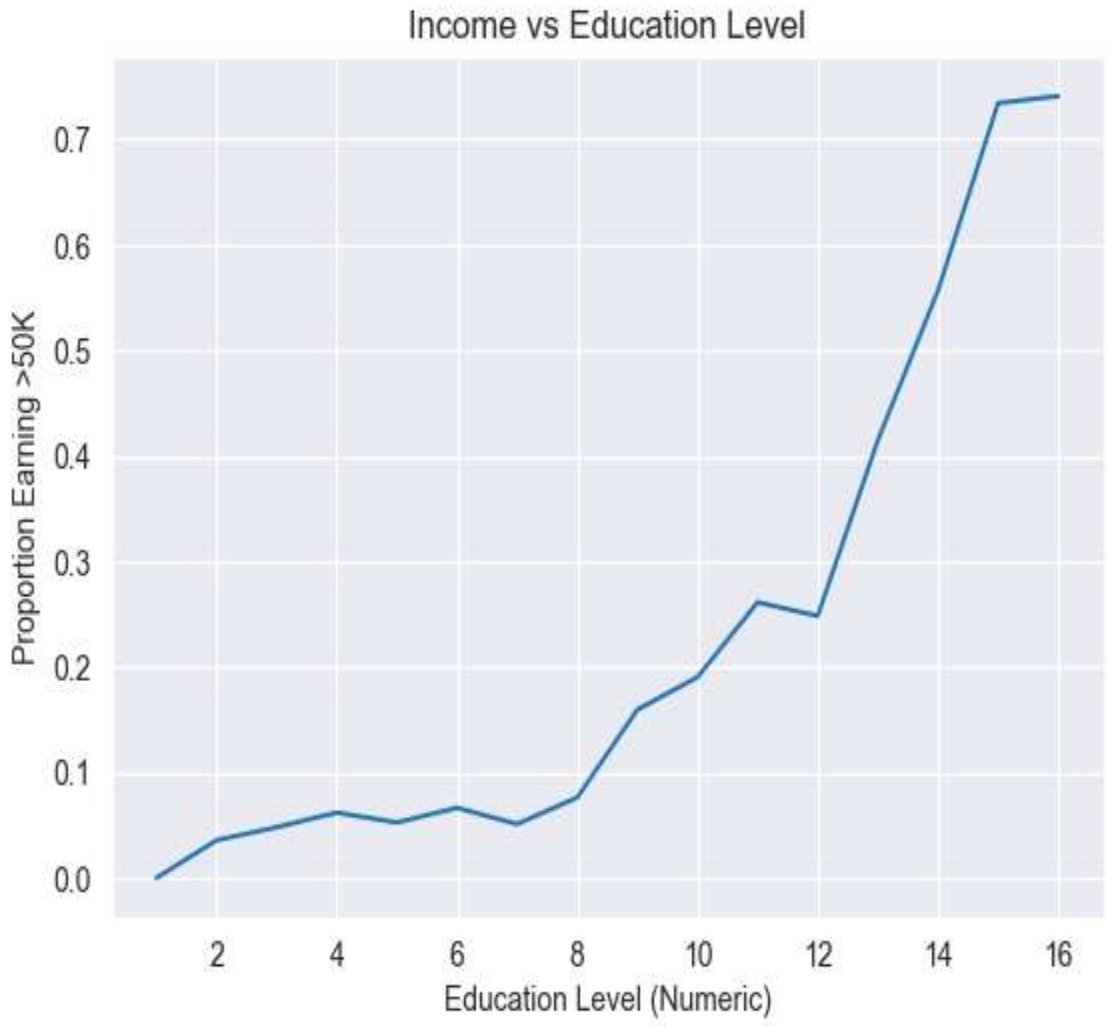
Cleaning Strategy & Governance Rationale

- To address these challenges, a systematic data-cleaning pipeline was implemented. This included standardizing categorical labels to ensure consistency, handling missing values using context-appropriate strategies, and removing irrelevant or weakly informative variables. These steps were designed to enhance analytical reliability, support reproducibility, and ensure that downstream BI insights are trustworthy for managerial use.

4. Exploratory Business Intelligence Insights

Education & Income

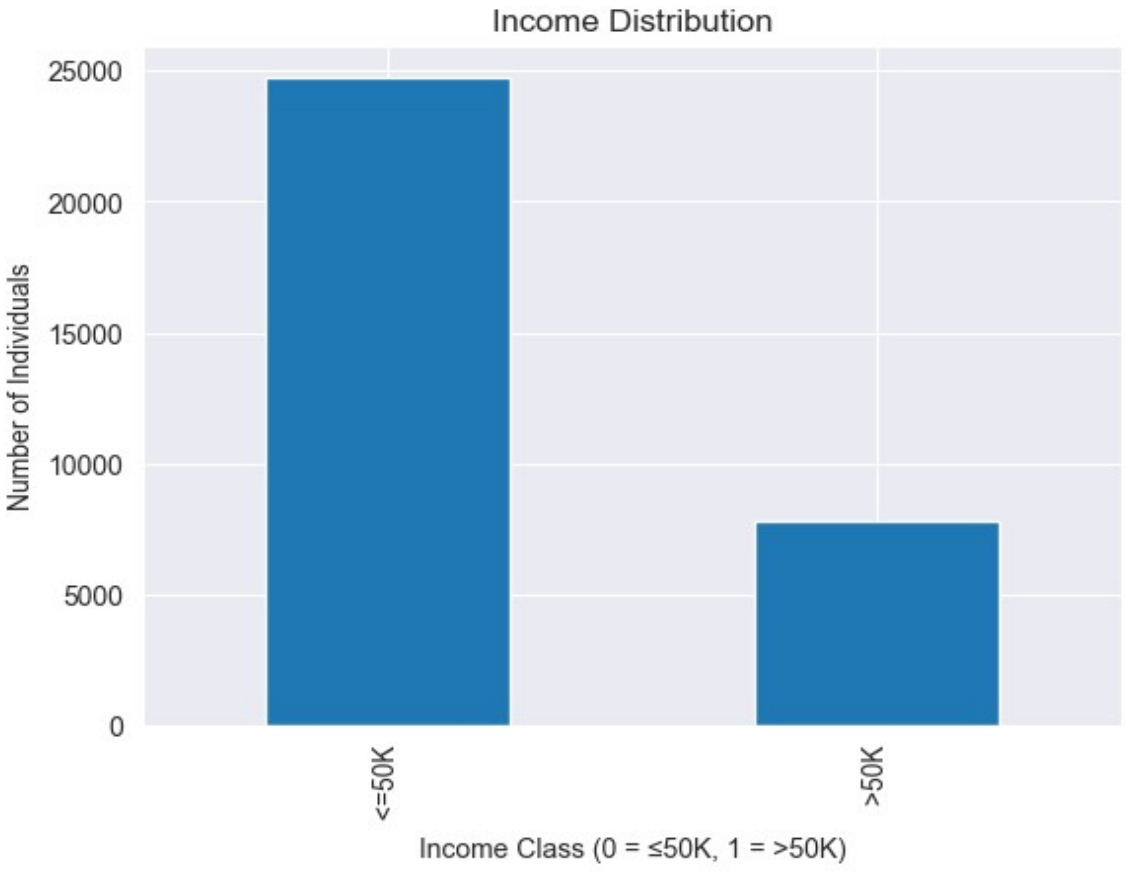
- Exploratory analysis revealed a strong relationship between education level and income category. Individuals with higher levels of formal education were significantly more likely to earn above the 50K threshold, indicating the role of human capital in income determination.



- From an institutional perspective, this has several implications. Firstly, Banks and SACCOs can have products that targets these people for example study-loans, savings accounts, good exchange rates e.t.c., Secondly, education can serve as risk indicator where higher education correlates with more stable income streams as shown by the visualizations. Thirdly, investment in TVET and higher education as a pathway to upward income mobility would work well as an idea. Finally, Income stratification by education supports evidence-based reporting on human capital returns.

Income Distribution Across Income classes

- Analysis of income distribution indicates that a substantially larger proportion of individuals fall within the lower income category, with relatively fewer individuals earning above the higher income threshold. This skewed distribution highlights the dominance of lower-income segments within the population and underscores the structural income inequality present in the dataset.



- From an institutional perspective, this distribution has several implications. First, it supports the prioritization of mass-market financial products, including low-fee transaction accounts, mobile-based micro-savings, and nano-loans, rather than an exclusive focus on premium banking services. Second, it emphasizes the limitations of income-based credit assessment alone, reinforcing the need for alternative credit scoring mechanisms such as mobile money usage patterns and repayment histories. Third, the observed income structure provides useful input for labor-market profiling and reinforces policy efforts aimed at expanding formal employment opportunities and social protection coverage. Finally, such income distribution insights support regulatory and statistical reporting requirements related to financial inclusion and income inequality, including those overseen by the Central Bank of Kenya (CBK) and the Kenya National Bureau of Statistics (KNBS).

Employment Intensity & Income

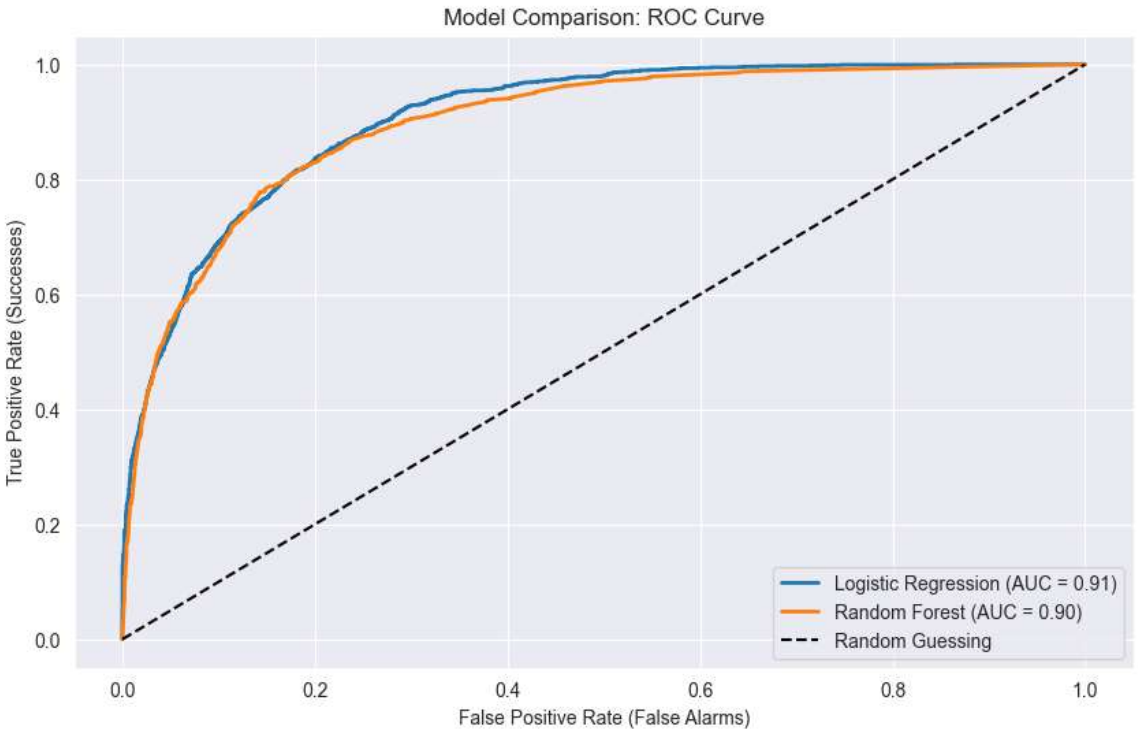
- Individuals working more than 40 hours per week exhibited a higher likelihood of earning above the income threshold, suggesting a positive association between employment intensity and income outcomes.

- This data allows for some key insights. Firstly, lenders can create cash-flow–aligned repayment products for high-hour earners such as traders, transport operators, or professionals with irregular but higher incomes. Secondly, we can use working hours as behavioural signal of income reliability while also supporting segmentation between informal low-hour workers and higher-intensity formal or entrepreneurial labour. Finally, working hours can highlight underemployment risks and inform labour inspections.

5. Predictive Modelling and Managerial Interpretation

Model Comparison

- Two predictive models were evaluated: Logistic Regression and Random Forest. The Random Forest model achieved higher predictive accuracy and AUC, while the Logistic Regression model offered superior interpretability



Strategic Trade-offs

- From a managerial perspective, interpretability and fairness are crucial in regulated environments. Choosing a more complex model like Random Forest may improve performance but could reduce transparency, making it harder to justify decisions to stakeholders or regulators.

Final Managerial Recommendation

- Given the importance of transparency and fairness, a simpler, more interpretable model such as Logistic Regression is recommended, even if it comes with slightly lower predictive performance.

6. Ethical, Governance Considerations and Recommendations

Fairness Risks

- Income vs Education Level**:** Higher-educated individuals are more likely to earn above 50K, potentially biasing predictions against lower-educated but capable individuals.

- Income vs Working Hours: Those working >40 hours/week earn more, which could unfairly classify casual workers or caregivers as low-income.

  Implication**:** Using these factors for automated credit scoring or HR decisions could systematically disadvantage certain groups despite accurate predictions.

Implications for Credit Scoring and HR Analytics in Kenya

- Credit scoring**:** Many reliable borrowers could be classified as low-income, particularly in the informal sector.
- HR analytics**:** Reliance on education and job type may reduce diversity and fairness.

Practical Mitigation Strategies

- Monitor prediction outcomes across education levels and working-hour groups.
- Review imputation and encoding choices to avoid favoring majority groups.
- Use models as support tools, not final decision-makers.
- Prefer interpretable models (e.g., Logistic Regression).
- Inform individuals about data use and allow review of automated decisions.

Governance & Compliance

- **Missing Values:** Mode imputation filled gaps in workclass, occupation, and native-country. Ensures completeness but may over-represent majority groups.
- **Redundant Variables:** Columns like education and education-num were simplified to avoid duplication, improving clarity and performance.
- **Categorical Encoding:** One-hot encoding made variables usable but increased complexity for auditing.

Alignment with the Kenyan Data Protection Act (2019)

Logistic Regression is interpretable and supports compliance:

Fair and transparent processing of personal data

Ability to explain automated decisions