Automated Income Classification Feasibility Study

1. Executive Summary

This study examines whether an automated income classification system could be used by a mid-sized Kenyan financial institution to support credit scoring and customer segmentation. The U.S. Census Adult Income dataset was used as a benchmark to test data quality, income patterns, and predictive models before applying similar methods to Kenyan data.

The analysis showed that the dataset contains several data quality problems, including missing values and redundant variables. However, these issues can be handled using a clear and reproducible data cleaning process. Exploratory analysis showed that age and education are strongly related to income, although these relationships cannot be directly transferred to Kenya due to the large informal (Jua Kali) sector.

Two predictive models were compared. The Random Forest model performed better than Logistic Regression across multiple performance measures, but it is harder to interpret.

Final Recommendation: Automated income classification should be used as a **decision-support tool**, not as a fully automated system. To comply with the Kenyan Data Protection Act (2019), complex models should be combined with explainability methods and human review, especially when rejecting loan applications.

2. Business Intelligence Insights: Income Patterns & Institutional Relevance

2.1 Purpose of BI Analysis

Before using predictive models, financial institutions need to understand the main factors that influence income. Exploratory Data Analysis (EDA) was carried out to identify demographic and economic variables associated with higher income and to consider how these patterns might apply in a Kenyan context.

2.2 Key Income-Related Patterns Observed

2.2.1 Age & Income

Pattern:
Income tends to increase with age, peaking in middle adulthood (around 35–55 years), and then stabilising or declining near retirement.

Institutional Meaning:
This suggests that standard credit scoring systems may disadvantage young applicants who currently earn less but may have strong future earning potential. Special products may be needed for younger customers.

2.2.2 Education & Income

Pattern:
Education emerged as one of the strongest predictors of income. Individuals with higher levels of education were more likely to earn above the 50K threshold.

Institutional Meaning:
In formal employment settings, education can act as a useful indicator of income stability. However, it should not be treated as the only measure of creditworthiness.

2.2.3 Workforce Sector / Occupation

Pattern:
Higher income levels were concentrated in certain occupations, particularly among incorporated self-employed individuals and senior professional roles.

Institutional Meaning:
Self-employment should not automatically be treated as high risk. Successful entrepreneurs may have stronger cash flows than salaried employees.

2.2.4 Working Hours & Income

Pattern:
Working longer hours did not always result in higher income. Extremely long working hours were often associated with lower-income roles.

Institutional Meaning:
This highlights the presence of a “working poor” group. Credit assessments should focus on income level and stability rather than effort alone.

2.3 Kenyan Institutional Context

Although the dataset is from the U.S., several risks arise if these patterns are applied directly to Kenya:

- **Education Limitation:** Many individuals in Kenya's informal sector earn stable incomes without formal qualifications. Relying heavily on education could exclude creditworthy customers.

- **Youth Population:** Kenya has a young population. Models that penalise youth could result in high rejection rates for young entrepreneurs.

- **Informal Employment:** Most Kenyan businesses are not formally incorporated. Treating business registration as a strict requirement would exclude much of the market.

These differences show the need to adapt models to local economic realities.

3. Predictive Modelling: Managerial Interpretation & Trade-offs

3.1 Purpose of Predictive Modelling

Predictive models were used to convert observed income patterns into a system that can support loan pre-qualification and customer segmentation.

3.2 Model Comparison (Non-Technical)

Two models were compared:

- **Logistic Regression:** Simple and easy to explain, but limited in capturing complex relationships.

- **Random Forest:** More powerful and able to model non-linear relationships, but less transparent.

The Random Forest model performed significantly better than Logistic Regression across key performance metrics, including accuracy and AUC.

3.3 Interpretation of Model Performance

- **Precision:** The Random Forest model showed relatively high precision, meaning that when it predicted a customer as high income, it was usually correct. This reduces the risk of issuing loans to unsuitable customers.

- **Recall:** The model identified a majority of high-income individuals but still missed some, showing that automated models should not be the only screening method.

3.4 Model Selection Recommendation

Although Logistic Regression is easier to explain, its performance was not strong enough for practical use.

Recommendation: The Random Forest model is more suitable due to its higher predictive performance. However, it should be used alongside explainability techniques (such as feature contribution analysis) and human review to reduce risk and ensure transparency.

4. Ethical, Fairness & Governance Considerations

4.1 Algorithmic Bias Risks

Sensitive attributes such as gender and race were excluded from the model, which helps reduce direct discrimination. However, some variables may still act as indirect proxies for sensitive characteristics, requiring regular fairness checks.

4.2 Governance Challenges

- **Data Minimisation:** Collecting unnecessary personal details increases privacy risk.

- **Automated Decisions:** Fully automated rejections can create accountability and legal issues.

4.3 Kenyan Regulatory Alignment

The Kenyan Data Protection Act (2019) gives individuals the right to understand how decisions are made. Automated systems must therefore be transparent and include human oversight.

4.4 Mitigation Strategies

- Regular bias testing
- Monitoring changes in data patterns over time
- Ensuring humans review automated decisions

5. Conclusion & Strategic Recommendations

The study shows that automated income classification can support credit scoring in Kenya if used carefully. The Random Forest model provides strong performance, but automation should assist—not replace—human judgment. This approach supports compliance with Kenyan regulations while accounting for the informal nature of the economy.