MANAGERIAL REPORT: BUSINESS INTELLIGENCE & ETHICAL FEASIBILITY STUDY

FROM: Business Intelligence Working Group [Group 3]
DATE: December 16, 2025
SUBJECT: Feasibility of Automated Income Classification: Kenyan Context & Ethical Governance Analysis

---

1.0 Introduction

This report provides a detailed interpretation of the Business Intelligence (BI) benchmarking exercise conducted using the Adult Census Income dataset. While the technical modeling confirmed the predictive power of machine learning algorithms (specifically Random Forest), the translation of these findings into the Kenyan financial ecosystem requires nuanced contextualization. This document specifically addresses **Question 2(b)** regarding the interpretation of insights within the Kenyan institutional context, **Question 3(c)** regarding the managerial interpretation of predictive models, and **Question 4**, detailing the ethical, legal, and governance implications of deploying such a system under the Kenyan Data Protection Act (2019).

2.0 Interpretation of Insights in the Kenyan Institutional Context (Q2b)

The exploratory analysis of the dataset revealed strong correlations between income levels and variables such as **Age**, **Education**, and **Hours Worked**. However, a direct transposition of these US-based patterns to the Kenyan market presents significant risks. Below is an analysis of how these variables must be re-interpreted for local deployment across four key banking domains.

2.1 Financial Product Design

The data indicates that income stability and growth are heavily biased toward mid-career individuals (30–50 years old) with formal education. In the US context, this supports "salary-check off" loan products.

Kenyan Interpretation:

In Kenya, the correlation between formal education and high income is less linear due to the dominance of the "Jua Kali" (informal) sector, which accounts for over 80% of employment.

- **Product Innovation:** A model relying heavily on "Education" or "Formal Job Titles" would systematically exclude high-net-worth informal entrepreneurs (e.g., wholesale traders, matatu fleet owners) who may lack formal degrees but possess significant cash flow.
- **Recommendation:** Financial products should not be designed solely around the "graduate salaried worker." Instead, product design should pivot to **Cash-Flow Based**

Lending, where income is verified via mobile money statements (M-Pesa) rather than educational certificates or standard 40-hour work weeks.

2.2 Microcredit and Lending Risk Assessment

The dataset shows a strict clustering of "safe" borrowers around the standard 40-hour work week. Deviations (working fewer or more hours) in the US context often signal instability.

Kenyan Interpretation:

- **The "Hustler" Economy:** In Kenya, working hours are highly variable. A micro-entrepreneur might work 70 hours one week and 20 the next due to seasonality or supply chain constraints.
- **Risk Scoring:** If our algorithm penalizes "irregular hours" (as suggested by the US data), we will falsely flag reliable Kenyan borrowers as "High Risk."
- **Strategic Shift:** Risk assessment for microcredits (e.g., Fuliza or Hustler Fund limits) must decouple "hours worked" from "reliability." The institution should prioritize **repayment history** and **transaction velocity** over static demographic inputs like Age or Hours Worked. The model must be retrained to view "variable hours" as a feature of entrepreneurship, not a bug of unemployment.

2.3 Labour-Market Profiling

The analysis highlighted "occupation" as a primary predictor of income. In the benchmark data, "Executive" and "Professional" roles guaranteed higher income.

Kenyan Interpretation:

- **Sectorial Fluidity:** The Kenyan labor market is characterized by "portfolio work," where an individual may hold a formal low-income job while running a lucrative side business (farming, rentals).
- **Profiling Strategy:** Standard labor profiling using checking accounts will miss this "hidden" income. The BI unit must develop a **"Total Economic Activity"** profile that aggregates data from multiple sources (banking, mobile wallets, and SACCO contributions) rather than relying on a single "Occupation" field provided at account opening.

2.4 Regulatory Reporting

The distinct patterns in Age and Gender found in the dataset have direct implications for Central Bank of Kenya (CBK) reporting.

Kenyan Interpretation:

- **Financial Inclusion:** The CBK mandates reporting on financial inclusion metrics. Our analysis shows that women and younger people are often under-represented in the "High Income" class in the training data.
- **Compliance Risk:** If we automate lending based on this data, our quarterly reports to the CBK will likely show a decline in lending to women and youth, potentially triggering regulatory audits for exclusionary practices. We must proactively adjust the model thresholds for these protected groups to ensure our aggregate lending portfolio remains balanced and compliant with financial inclusion goals.

3.0 Managerial Interpretation of Predictive Models (Q3c)

We evaluated two candidate models: Logistic Regression and Random Forest. This section outlines the managerial decision on which model to deploy, weighing performance against operational realities.

3.1 Model Selection Decision

We definitively recommend the **Random Forest** model for deployment.

- **Performance Gap:** The Random Forest model achieved an **AUC score of 0.90**, indicating outstanding discrimination ability. In contrast, the Logistic Regression model scored **0.58**, which is marginally better than random chance. Deploying the Logistic Regression model would result in high default rates and lost revenue.

3.2 Operational Trade-offs

While Random Forest is the superior statistical model, adopting it requires accepting specific trade-offs:

A. Interpretability vs. Accuracy

- **The Trade-off:** Logistic Regression is "glass-box"—it gives clear coefficients (e.g., "Age +5 years adds 10 points to score"). Random Forest is "black-box"—it uses thousands of decision trees, making it hard to explain *exactly* why a specific applicant was rejected.
- **Managerial Stance:** Given the 32% accuracy gap, we cannot afford the interpretable but inaccurate model. To satisfy the "Right to Explanation" (DPA 2019)[1], we will use secondary tools (like SHAP plots) to extract key reasons for rejection from the Random Forest model rather than relying on the model's native structure.

B. Misclassification Costs (Risk Appetite)

- **False Positives (Type I Error):** Predicting a borrower is "Rich" (Safe) when they are actually "Poor" (Risky).
  - *Cost:* Direct capital loss (Default).
- **False Negatives (Type II Error):** Predicting a borrower is "Poor" (Risky) when they are actually "Rich" (Safe).
  - *Cost:* Opportunity cost (Lost Interest Income).
- **The Random Forest Advantage:** The Random Forest displayed higher **Precision (0.75)** compared to Logistic Regression. This means it is far more conservative and accurate when approving loans. In the current volatile economic climate, prioritizing the reduction of Bad Debts (False Positives) is strategically preferred over aggressive lending.

4.0 Ethical & Governance Analysis (Q4)

The deployment of an automated income classification system introduces profound ethical risks. Our technical audit of the model revealed specific biases that, if left unmitigated, could expose the institution to legal action and reputational damage.

4.1 Algorithmic Fairness Concerns

We conducted a fairness audit by calculating the **False Negative Rate (FNR)** across demographic groups[2]. A "False Negative" occurs when the model incorrectly predicts a high-income applicant is "low income," effectively denying them a loan they deserved.

- **Gender Bias:** The benchmarking revealed that the model is more likely to reject qualified female applicants than male applicants. This stems from historical wage gaps present in the training data.
- **Racial/Tribal Bias:** While the US data contains "Race," in Kenya, this translates to "Ethnicity" or "Location." There is a high risk that the model could pick up proxies for ethnicity (e.g., zip codes or specific vernacular names if used) and systematically deny loans to specific communities, constituting **digital redlining**.

4.2 Governance Challenges with Socio-Economic Data

Handling this data presents unique governance hurdles:

- **Data Lineage & Quality:** The "Data Quality Assessment" (Q1) revealed that ~7% of records had missing values. In a Kenyan context, administrative data is often even "messier." Making automated decisions on incomplete data violates the principle of **Accuracy**[3].
- **Proxy Variables:** Even if we remove "Gender" from the model, other variables like "Occupation" (e.g., Nursing vs. Construction) can act as proxies for gender. Governance policies must define strictly which variables are admissible for credit scoring.

4.3 Alignment with the Kenyan Data Protection Act (2019)

The deployment of this system must strictly adhere to the Data Protection Act (DPA) 2019. We have identified three critical sections of the Act that impact this project:

A. Automated Individual Decision Making (Section 35)

- **The Law:** Section 35(1) states that a data subject has a right *not* to be subject to a decision based *solely* on automated processing, including profiling, which produces legal effects concerning or significantly affecting them[4].
- **Implication:** We **cannot** deploy a fully automated "Auto-Reject" system for loans.
- **Required Action:** We must implement a "Human-in-the-Loop" review process. Any customer rejected by the AI model must have the option to appeal to a human credit officer.

B. Processing of Sensitive Personal Data (Section 44)

- **The Law:** Data regarding an individual's race, marital status, and sex is classified as "Sensitive Personal Data."
- **Implication:** Our current model uses sex and marital_status as features. This is a high-risk practice.

- **Required Action:** We must strip these sensitive fields from the *deployment* model. While we use them for *auditing* fairness in the back office, they cannot be used to calculate the credit score itself.

C. Right to Explanation (Transparency)

- **The Law:** Data subjects have the right to be informed of the logic involved in automated decision-making.
- **Implication:** "Black Box" models like Random Forest (which we found to be most accurate) are difficult to explain.
- **Required Action:** We must use interpretation tools (like SHAP values) to generate simple reason codes for customers (e.g., "Your loan was declined due to low transaction volume," rather than "Error Code 404").

4.4 Practical Mitigation Strategies

To move forward safely, we propose the following mitigation framework[5]:

1. **Fairness through Unawareness:** Permanently drop Gender, Race, and Marital Status from the production model inputs to prevent direct discrimination.
2. **Bias Stress-Testing:** Before any model update is pushed to production, it must pass a "Fairness Unit Test" where we verify that the error rates (FNR) for men and women are within a 5% margin of difference.
3. **The "Shadow Mode" Phase:** For the first 3 months, the AI model will run in the background. It will generate a score, but the loan decision will still be made by a human. We will compare the AI's suggestion against the human's decision to verify safety before turning it on.
4. **Algorithmic Impact Assessment (AIA):** As required by the Office of the Data Protection Commissioner (ODPC), we will file an AIA detailing potential risks to customer privacy before processing begins.

5.0 Conclusion

While the automated income classification system demonstrates high technical accuracy (AUC 0.90), its immediate deployment in the Kenyan market is not recommended without significant localization. The differences in labor market structure (Formal vs. Informal) and the strict requirements of the Kenyan Data Protection Act (2019) demand a hybrid approach.

We recommend a **Semi-Automated** lending system where the AI model serves as a decision-support tool for credit officers, rather than a fully autonomous gatekeeper. This balances operational efficiency with ethical responsibility and legal compliance.