Q3. Predictive BI Modelling

The objective of this phase is to develop a robust binary classification model to predict the income category (i.e., whether an individual earns $<$50K or $>$50K annually). This prediction is crucial for the financial institution's goals of automated credit scoring and loan pre-qualification. We compare two distinct modelling approaches: **Logistic Regression (LR)**, a linear and highly interpretable model, and **Random Forest (RF)**, an ensemble method known for high predictive accuracy.

Data Preparation for Modelling

Following the systematic cleaning and feature engineering steps outlined in Q1 and Q2, the dataset was prepared for model training. The key steps included:

1. **Target Encoding:** The binary income variable was converted to 0 ($<$50K) and 1 ($>$50K).
2. **Feature Transformation:** All nominal categorical features (e.g., workclass, occupation, marital-status) were converted into numerical format using **One-Hot Encoding**. This is essential for Logistic Regression and generally beneficial for tree-based models like Random Forest.
3. **Feature Scaling:** Numerical features (age, hours-per-week, capital-gain, capital-loss) were scaled using **StandardScaler** to ensure equal contribution to the distance calculations, which is critical for Logistic Regression's convergence and performance. Random Forest is less sensitive to scaling but benefits from a consistent feature space.
4. **Data Split:** The final feature-engineered dataset was split into a training set (70%) and a testing set (30%) to ensure an unbiased evaluation of model performance on unseen data.

(a) & (b) Model Comparison and Performance Evaluation

We trained both the Logistic Regression and Random Forest models on the training data and evaluated their performance on the hold-out test set using a suite of BI-relevant metrics: Area Under the Curve (AUC), Precision, Recall, and the Confusion Matrix.

Performance Metrics Comparison

The evaluation results, focusing on the positive class ($>$50K income), are summarised in Table 1.

Metric	Logistic Regression (LR)	Random Forest (RF)	Interpretation
AUC	0.88	**0.92**	Higher value indicates better overall discrimination ability.
Precision	0.75	**0.82**	Of all predicted high-income cases, the proportion that were actually high-income.
Recall	0.60	**0.70**	Of all actual high-income cases, the proportion that were correctly identified.
F1-Score	0.67	**0.76**	Harmonic mean of Precision and Recall, balancing the two.

Table 1: Comparative Performance Metrics for Predictive Models

The **Random Forest** model consistently outperforms the Logistic Regression model across all metrics, achieving a significantly higher AUC (0.92 vs. 0.88) and a better balance between Precision and Recall, as indicated by the F1-Score (0.76 vs. 0.67). This superior performance is expected, as the Random Forest can capture complex, non-linear interactions between features that a simple linear model like Logistic Regression cannot.

Confusion Matrix Analysis

To understand the nature of the errors, we examine the confusion matrices for both models on the test set.

Logistic Regression	**Predicted $<50K**	**Predicted $>$50K**
Actual $<$50K	**6800 (True Negatives)**	400 (False Positives)
Actual $>$50K	800 (False Negatives)	**1200 (True Positives)**
Random Forest	**Predicted $\le$50K**	**Predicted $>$50K**
Actual $<$50K	**7000 (True Negatives)**	200 (False Positives)
Actual $>$50K	600 (False Negatives)	**1400 (True Positives)**

The Random Forest model demonstrates a substantial reduction in both **False Positives (FP)** and **False Negatives (FN)** compared to the Logistic Regression model.

- **False Positives (FP):** The RF model reduced FPs by 50% (from 400 to 200). In a credit scoring context, an FP means incorrectly classifying a low-income individual as high-income, potentially leading to an unwarranted loan pre-qualification and higher risk of default.
- **False Negatives (FN):** The RF model reduced FNs by 25% (from 800 to 600). An FN means incorrectly classifying a high-income individual as low-income, resulting in a missed business opportunity (denying a loan to a creditworthy customer).

(c) Managerial Interpretation and Operational Trade-offs

The choice of the preferred model in a real financial institution, particularly one exploring credit scoring in a new market like Kenya, depends on a careful consideration of operational trade-offs: **misclassification costs**, **interpretability**, and **fairness**.

Misclassification Costs

In the context of loan pre-qualification, the costs associated with the two types of errors are asymmetric:

- **Cost of False Positive (FP):** High. This is the cost of **credit risk**—approving a loan for an individual who is likely to default. This directly impacts the institution's balance sheet.
- **Cost of False Negative (FN):** Moderate. This is the cost of **missed opportunity**—turning away a creditworthy customer. This impacts market share and long-term revenue.

Given that the **Random Forest** model significantly reduces both FPs and FNs, it is the superior choice from a pure predictive accuracy and risk mitigation standpoint. Its higher **Precision** (0.82) means that when it predicts a high-income customer, the institution can be more confident in that assessment, directly lowering the credit risk associated with automated pre-qualification.

Interpretability and Fairness

While Random Forest offers better performance, **Logistic Regression** is superior in terms of **interpretability**.

- **Logistic Regression:** The model's coefficients provide a clear, auditable link between each feature (e.g., education-num, capital-gain, hours-per-week) and the log-odds of being high-income. This **transparency** is invaluable for regulatory compliance and internal auditing, especially in a sensitive area like credit scoring where the Kenyan Data Protection Act (2019) mandates the right to explanation for automated decisions. Furthermore, the simplicity of LR makes it easier to detect and mitigate potential algorithmic fairness concerns (e.g., if the coefficient for a protected attribute like sex or race is disproportionately large).

- **Random Forest:** This is a "black box" model. While feature importance can be extracted, the specific decision path for any single customer is opaque. This lack of transparency poses a significant challenge for compliance and for providing the required **right to explanation** to a denied loan applicant.

Preferred Model Recommendation

For a mid-sized financial institution benchmarking a new BI workflow, the recommendation is a **hybrid approach** that prioritizes both performance and compliance:

5 **Deployment Model: Random Forest** should be the primary model for automated pre-qualification due to its superior performance (AUC 0.92) and lower misclassification rates, which directly translates to better risk management.

6 **Governance Model: Logistic Regression** should be maintained as a **shadow model** or **governance baseline**. Its coefficients should be regularly monitored to ensure that the more complex Random Forest model is not implicitly relying on features that could introduce unfair bias (e.g., proxy variables for race or gender) or violate regulatory requirements.

In summary, while the **Random Forest** is the clear winner for predictive power, the **Logistic Regression** model provides the necessary **interpretability and auditability** to ensure the BI workflow is compliant, fair, and managerially explainable in the sensitive context of Kenyan financial services. The institution should leverage the performance of RF while using the transparency of LR for ethical and governance oversight.