## Interim Report (Task 1–2)

**Project**: Credit Risk Probability Model for Alternative Data (BNPL)
**Client**: Bati Bank (analytics engineering)
**Dataset**: Xente e-commerce transaction data (Uganda)

### 1) Executive Summary

This interim submission covers **Task 1 (Business Understanding)** and **Task 2 (EDA)**. Since the dataset does not include a direct loan repayment outcome, we are preparing the ground for a proxy-based supervised learning approach: first by understanding regulatory/modeling constraints (Basel II context), then by exploring the transaction data to identify quality issues, dominant patterns, and hypotheses for feature engineering.

### 2) Task 1 — Credit Scoring Business Understanding (What we delivered)

We updated `README.md` with a “Credit Scoring Business Understanding” section that addresses:

- **Basel II and interpretability/documentation**: why the model must be explainable, auditable, validated, and well-documented in a regulated environment.
- **Why a proxy target is required**: because we have no observed “default” label, but supervised modeling still needs a target; proxy-based modeling introduces business and model risk.
- **Trade-offs (simple vs complex models)**: interpretable scorecard/logistic regression vs higher-performing gradient boosting, and why governance and explainability matter.

### 3) Task 2 — Exploratory Data Analysis (EDA) Summary

#### 3.1 Dataset overview
- **Rows / Columns**: **95,662 rows** × **16 columns**
- **Missing values**: **0 total missing values**
- **Duplicate rows**: **0 duplicates**
- **Time range**: from **2018-11-15** to **2019-02-13** (UTC timestamps)
- **Geography/currency**:
- `CountryCode` is constant (**256**) and `CurrencyCode` is constant (**UGX**) → these do not add predictive signal in this dataset.

#### 3.2 Key numerical behavior (Amount/Value)
- `Amount` includes **negative values** (credits/refunds) and positives (debits); min = **-1,000,000**, max = **9,880,000**
- Strong relationship: **Corr(Amount, Value) ≈ 0.99**, which indicates `Value` is almost redundant with `Amount` magnitude.
- Very large spread (heavy tails/outliers): standard deviation is much larger than the median, suggesting skewness and outliers.

#### 3.3 Fraud label distribution
- `FraudResult` is **highly imbalanced**:
- 0 (non-fraud): **95,469**
- 1 (fraud): **193**

This implies that if fraud is later used as a feature/label in modeling, we must treat imbalance carefully and avoid leakage (depending on the final business objective).

#### 3.4 Categorical distributions (high-level)
Highly dominant categories:
- `ProductCategory`: top category is **financial_services** (~45k rows)
- `ChannelId`: top channel is **ChannelId_3** (~56.9k rows)
- `ProviderId`: top provider is **ProviderId_4** (~38k rows)

These skews suggest some categories will dominate learning; encoding strategy and grouping rare categories may matter later.

### 4) Top 3–5 Most Important Insights (From EDA)

1. **Data quality is strong (no missing/duplicates)**
This simplifies preprocessing and allows us to focus on feature engineering rather than heavy cleaning.

2. **`CountryCode` and `CurrencyCode` are constant**
They likely provide no predictive value and can be dropped (or kept for schema consistency but expected to have zero importance).

3. **`Amount` and `Value` are nearly redundant**
With correlation ≈ 0.99, using both may add multicollinearity/redundancy; we can consider keeping one (or using both only if a model benefits).

4. **Transactions are heavy-tailed with extreme outliers**
Scaling/robust transformations (or clipping/winsorization) may improve model stability.

5. **Fraud label is extremely imbalanced**
Any future modeling that uses fraud must handle imbalance and be explicit about the target definition to avoid misleading performance metrics.

### 5) Interim Deliverables Checklist

- **Task 1**: `README.md` includes “Credit Scoring Business Understanding” ■
- **Task 2**: `notebooks/eda.ipynb` created and executed for EDA ■
- **Interim report**: This document ■

### 6) Next Steps (Planned)

For the final submission, we will:
- Engineer customer-level features and time-derived features (Task 3)
- Create proxy target `is_high_risk` using RFM + clustering (Task 4)
- Train and compare multiple models with MLflow tracking (Task 5)
- Deploy the best model behind an API with CI/CD and Docker (Task 6)