Project Documentation

Loan Default Risk Modelling

PROJECT OVERVIEW

- **Objective**: Estimate the probability (PD) that an individual loan will default (≥ 90 days past-due, write-off, or legal recovery) before maturity.
- **Business value**: Informs pricing, IFRS 9 / Basel III provisioning, and early-warning actions.
- **Scope**: Retail loans booked by Welford Bank from January 2020 to May 2025 (about 87 000 contracts).
- **Deliverables**: Clean code, baseline and neural-network models, performance report, and explainability material.

DATA DESCRIPTION

Files used:
• loans_welfordbank_en.csv – 87 135 rows – target table
• credit_history_welfordbank_en.csv – 65 812 rows – prior payment behaviour
• loan_metrics_welfordbank_en.csv – 87 135 rows – portfolio context
• clients_welfordbank_en.csv – 25 000 rows – client demographics

All datasets are synthetic; correlations may not match real portfolios.
Class imbalance: only about 2 % of loans are in default, so re-balancing is required.

PRE-PROCESSING PIPELINE

1. Merge tables and create features (credit-history aggregates, loan ratios, client attributes).
2. Parse date columns, one-hot encode categoricals, scale numeric variables.
3. Train/test split of 70 / 30 with stratification by target.
4. Two imbalance treatments tried: inverse-frequency class weights and SMOTE oversampling (train set only).

MODELLING WORKFLOW

• Logistic Regression with class weights – benchmark and interpretability.
• Logistic Regression with SMOTE – compare oversampling to weighting.
• MLP-A: two hidden layers (64 and 16 units) with 0.5 dropout, early-stopping.
• MLP-B: three hidden layers (128, 64, 16 units) with 0.4 dropout, higher capacity.

EVALUATION METRICS

AUC-ROC (discrimination), precision and recall for the default class, confusion matrix counts, and calibration curve.

RESULTS ON TEST SET

Model	AUC	Recall (Default)	Precision (Default)	F1
Logistic (weights)	0.66	0.56	0.05	0.09
Logistic (SMOTE)	0.64	0.59	0.05	0.09
MLP-A (2 layers)	0.68	0.79	0.05	0.09
MLP-B (3 layers)	0.69	0.83	0.07	0.13

Interpretation: neural nets lift recall up to 83 %, but precision remains below 7 %, yielding many false positives. The weighted logistic model is still useful as a regulatory benchmark; SMOTE offers little additional benefit.

Customer Segmentation with K-Means

PROJECT OVERVIEW

- Objective: Segment Welford Bank's retail customers into homogeneous groups based on sociodemographic attributes and relationship value so that product recommendations reach only the most relevant audiences.
- Business value: Increases cross-sell and up-sell effectiveness, reduces customer fatigue from irrelevant offers, and enables personalized pricing and retention.
- Scope: Synthetic retail-bank dataset (clients, accounts, transactions, cards) covering activity from January 2020 to May 2025.
- Deliverables: Clean, consolidated dataset; elbow + silhouette evaluation; baseline K-Means model (k = 3); cluster profiles; marketing action map.

DATA DESCRIPTION

- clients_welfordbank_en.csv ≈ 25 000 rows Demographics and CLV
- accounts_welfordbank_en.csv ≈ 40 000 rows Products and balances
- transactions_welfordbank_en.csv ≈ 1.3 million rows Transactional activity
- cards_welfordbank_en.csv ≈ 18 000 rows Card limits and usage

PRE-PROCESSING PIPELINE

1. Normalize column names and remove extra spaces.
2. Feature engineering
   • Demographics: age (from birth date), gender dummy, country/city
   • Relationship value: CLV, number of accounts, average balance
   • Behaviour: transaction count, average and standard-deviation amounts, total deposits and withdrawals
   • Credit usage: number of cards, average limit, utilization ratio
3. Merge all tables on client_id to obtain one row per customer.
4. One-hot encode categorical variables (gender) and drop the first level to avoid collinearity.

5. Scale numerical variables with StandardScaler.

MODELING FLOW

- Elbow method: k = 2…10 → SSE flattens sharply after k = 4.
- Silhouette: highest at k = 2 (0.22), very close at k = 3 (0.20), falls below 0.15 for k ≥ 4.
- Final choice: k = 3 (balance between interpretability and clear separation).
- Training: KMeans(n_clusters = 3, random_state = 12, n_init = "auto").

RESULTS WITH (k = 3)

Cluster 0 – Low-Engage (52 %)

- Key traits: 1 account; ≈ 55 transactions / year; credit utilization ~ 0 %
- Recommended actions: basic activation & cross-sell; no-fee card; cash-back on first 3 purchases; digital-banking reminders

Cluster 1 – High-Credit-Use (23 %)

- Key traits: 1 account; credit utilization ~ 29 %; CLV near average
- Recommended actions: responsible limit increase; pre-approved personal loans; financial-health alerts

Cluster 2 – Multi-Product-Active (25 %)

- Key traits: ≥ 2 accounts; ≈ 128 transactions / year; credit utilization ~ 17 %
- Recommended actions: premium / wealth up-sell; investment or interest-bearing account; insurance & pension products; VIP benefits (cash-back, dedicated advisor)

Global silhouette: 0.20

EVALUATION METRICS

- Inertia/SSE for elbow analysis
- Silhouette coefficient (overall and per cluster)
- Cluster balance (size distribution)

VISUALIZATIONS

1. Elbow and silhouette curves (k = 2–10)
2. PCA scatter plots for k = 2, 3, 4, colored by cluster

BUSINESS INTERPRETATION

- Segmented marketing: each group receives only relevant products, reducing attrition due to saturation.
- Resource allocation: retention budget toward Multi-Product-Active; risk monitoring on High-Credit-Use.

Fraud-Detection Mode

PROJECT OVERVIEW

- Objective: build a real-time classifier that flags fraudulent transactions with at least 95 % recall and no more than nine false alarms per genuine fraud (≈ 10 % precision).
-

  Business value: protects customer funds, reduces charge-backs, and feeds the fraud-ops queue with high-priority alerts instead of noise.

- Scope: 1.85 million retail-bank transactions generated between January 2020 and May 2025, joined with client and account context.

- Deliverables: a reproducible notebook, a requirements file, the trained XGBoost model, and SHAP explainability plots.

DATA DESCRIPTION

Files used – transactions_welfordbank_en (1.85 M rows), clients_welfordbank_en (25 k), accounts_welfordbank_en (33 k), fraud_detections_welfordbank_en (35 k confirmed cases).

All data are synthetic, so true correlations are limited and class imbalance is severe: only ≈ 2 % of rows are labelled as fraud.

PRE-PROCESSING PIPELINE

The tables are merged on Client_ID and Account_ID so every transaction carries segment, status, account type, and balance.
We extract year, month, weekday, and hour from the timestamp, convert the dotted IP string to a single integer, and add a simple behavioural feature—Tx_24h, the number of transactions the same client made in the previous twenty-four hours.
Numeric columns are standard-scaled, categoricals are one-hot encoded, and a stratified 75/25 split keeps the fraud ratio intact.

MODELLING WORKFLOW

1. Unsupervised baseline: a three-layer autoencoder trained on numeric features, using the 95th-percentile reconstruction error as the alert threshold.
2. Supervised baseline: Random Forest (250 trees, depth 12) with SMOTE oversampling.
3. Final model: XGBoost (600 trees, depth 8, learning-rate 0.05, scale_pos_weight) retrained after adding Tx_24h and evaluated both at the default 0.50 cut-off and at a higher threshold grid-searched for precision ≥ 20 % (no viable τ found).

Total training time on a mid-range laptop: about two minutes and thirty seconds.

EVALUATION METRICS

We report precision, recall, PR-AUC, and F1 for the fraud class, together with confusion-matrix counts. Accuracy is not considered informative under heavy imbalance.

RESULTS ON THE TEST SPLIT

Autoencoder – PR-AUC 0.03, recall ≈ 0.70, precision ≈ 0.03: plenty of coverage but 21 k false alarms.
Random Forest – PR-AUC 0.10, recall 0.93, precision 0.10: triples precision and keeps most frauds.
XGBoost – PR-AUC 0.099, recall 0.95, precision 0.10: matches the ten-to-one alert ratio while nudging recall up two points, thus hitting the target KPI.

A quick SHAP run confirms that amount, balance, Tx_24h, transfer flag, and night-time operations are the main drivers of the fraud score.

INTERPRETATION

The XGBoost model achieves the required 95 % coverage and keeps precision at about 10 %. That translates to roughly nine investigations per confirmed fraud—a workload that fraud-ops has deemed acceptable for a pilot. Raising the decision threshold to τ = 0.79 did not improve the trade-off: precision barely moved while recall fell below the target.