© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718013

A Data-Driven Framework for Target Market Selection and Performance Evaluation of a Credit Card Product Using Statistical Analysis

INZAMAMAHMED MOHAMMEDMUSTAK SIDDIKI [1], GUIDE: PROF. RAKSHITHA B S [2]
[1]*PG - Research Scholar Department of M.Sc. Computer Science & IT JAIN (Deemed-to-be) University* Bengaluru, India
[2]*Assistant Professor Department of M.Sc. Computer Science & IT JAIN (Deemed-to-be) University Bengaluru*, India

Abstract- ***The Indian banking sector is highly competitive. Established banks leverage extensive customer data to maintain market dominance, making product launches by new entrants particularly challenging. Existing research has not produced a unified, end-to-end framework that integrates multi-source banking data analysis with rigorous experimental product validation. This paper addresses that gap through a two-phase data-driven framework. Phase 1 performs comprehensive data cleaning and exploratory data analysis (EDA) on a 40,000-record banking dataset comprising customer demographics, transaction history, and credit score information, with the objective of identifying the optimal target market segment. Phase 2 validates product performance through a statistically controlled A/B trial supported by hypothesis testing and confidence interval analysis. The framework introduces context-aware imputation techniques, a budget-constrained trial design methodology, and structured comparative tables to guide decision-making. Results confirm that targeting the 18–25 age group with a tailored credit card product leads to a statistically significant increase in average transaction amounts (95% CI: $226–$245). This study provides a reproducible, evidence-based blueprint for credit card product launches in emerging markets such as India.***

Index Terms—Exploratory Data Analysis, Credit Card Analytics, Customer Segmentation, A/B Testing, Hypothesis Testing, Statistical Power, Indian Banking Market

I. INTRODUCTION

A. Background of the Study

The Indian financial services sector is undergoing rapid transformation, compelling organisations to adopt data-driven decision-making. Incumbent banks possess extensive historical customer data, which they exploit to maintain competitive advantage. New entrants, lacking established brand equity and large capital reserves, must leverage multi-source data comprising customer demographics, transaction behaviour, and credit history to develop products that resonate with under-served segments.

The proliferation of banking data enables institutions to analyse customer behaviour, identify preferences, and pilot products with select customer groups before broad deployment. These techniques reduce financial risk and support informed product strategy, particularly in the dynamic Indian financial services market.

B. Problem Statement

Launching financial products based on intuition or simple surveys frequently results in low adoption rates and elevated financial risk. Traditional analytical methods are ill-suited to uncover non-linear, high-dimensional patterns in modern transaction data. A systematic framework is required that integrates target customer identification from combined data sources with statistically validated product performance testing prior to full-scale launch.

C. Motivation

This research is motivated by the need for rigorous, evidence-based product validation in emerging markets. The gap between descriptive analytics (EDA) and inferential statis-tics (hypothesis testing) creates ambiguity around product-market fit. This work aligns with contemporary industry practices such as A/B testing and evidence-based product

validation—practices notably absent from the academic banking analytics literature.

D. Objectives

The primary objectives of this study are:

- To conduct multi-source EDA on integrated customer, trans-action, and credit score datasets to identify patterns, trends, and inter-variable relationships.
- To identify the optimal target market segment through demographic and behavioural profiling.
- To design and execute a controlled A/B trial using statistical power analysis and experimental design principles.
- To evaluate credit card product performance through hypothesis testing and confidence interval analysis.

E. Contributions

This paper contributes a two-phase framework advancing from data cleaning and EDA to experimental product testing.

Unlike prior studies relying on synthetic or single-source data, this research employs a real-world banking dataset of 40,000 records. Key contributions include:

(i) context-aware, demographic-preserving missing value imputation;

(ii) a budget-constrained trial design approach; (iii) structured comparison tables justifying methodological choices; and (iv) workflow diagrams illustrating the dataset architecture and hypothesis testing process.

F. Organisation

The remainder of this paper is organised as follows. Section II reviews related literature. Section III presents the pro-posed methodology. Section IV describes the implementation. Section V discusses expected results. Section VI presents applications and use cases. Section VII concludes with future directions.

II. RELATED WORK

A. Thematic Classification of Literature

The reviewed literature is classified into four categories: traditional statistical methods, machine learning techniques, deep learning methods, and hybrid or survey-based studies. This taxonomy reflects the evolution of banking analytics from interpretable statistical models toward sophisticated data-driven frameworks.

1) *Traditional and Statistical Approaches:* Early banking analytics studies predominantly employed classical statistical methods. Logistic regression and chi-square analysis are widely used to identify factors influencing customer satisfaction—such as security, reliability, and service quality— particularly in emerging markets [1]. Correlation analysis and time-series visualisation transform raw financial data into actionable knowledge; however, traditional techniques struggle with the non-linear, high-dimensional patterns characteristic of modern financial data.

2) *Machine Learning Approaches:* Machine learning algorithms represent a significant advancement in customer segmentation and predictive modelling. Clustering algorithms including K-Means, Mean Shift, Agglomerative Clustering, and Self-Organising Maps (SOM) group customers by behavioural and financial profiles [10], [11], [13]. Comparative studies indicate that SOM often outperforms K-Means in segmentation quality [13]. Ensemble and boosting algorithms consistently outperform traditional classifiers in campaign response prediction [5]. AutoML further enhances performance through automated model selection and hyperparameter optimisation [12]. However, most ML studies remain model-centric, lacking structured product evaluation and decision-support systems.

3) *Deep Learning Approaches:* Deep learning techniques are primarily applied to fraud detection and risk analysis involving high-dimensional data patterns. Convolutional neural networks (CNNs) and recurrent models such as

Long Short-Term Memory (LSTM) networks outperform classical ML methods in fraudulent transaction classification [2], [3]. Hybrid architectures combine feature extraction with sequential modelling, and ensemble resampling methods address class imbalance [6]. A critical limitation is the reliance on synthetic or heavily preprocessed datasets, which constrains real-world generalisability.

4) *Hybrid and Recent Approaches:* Recent approaches combine ensemble learning and deep learning for improved predictive accuracy [9], [15]. Contextual modelling integrating customer, transaction, and behavioural data enables more comprehensive analytics. However, controlled experimental frameworks for financial product testing remain conspicuously absent.

B. Comparative Analysis of Existing Methods

Table I presents a structured comparison of representative banking analytics studies.

TABLE I COMPARATIVE ANALYSIS OF EXISTING METHODS IN BANKING ANALYTICS

Author	Year	Method	Dataset	Performance	Limitations
Abatsi et al.	2023	Statistical Analysis	Survey (Ethiopia)	Key satisfaction factors identified	Limited geography; no causal inference
Potluri et al.	2024	Customer Segmentation	Banking	Distinct segments identified	Dataset undisclosed; no product evaluation
Pandey et al.	2023	SOM, K-Means	TIC CRM	SOM improves quality	Outdated dataset; no validation
Bogireddy & Murari	2024	ML Campaign Prediction	UCI Bank	Boosting outperforms others	No controlled experiments
Met et al.	2023	AutoML	Proprietary KPI	High prediction accuracy	Not generalisable
Alarfaj et al.	2022	DL vs ML (Fraud)	European CC	DL outperforms ML	Transformed data; no testing framework
Damanik & Liu	2025	Ensemble (Fraud)	PaySim	Handles imbalance well	Synthetic data only
Almazroi, Ayub	2023	Hybrid DL	Synthetic	Accuracy improved	High complexity; synthetic
Van Acker et al.	2024	Survey (Fraud)	Multiple	Context improves performance	Multi-source integration underexplored
Kalid et al.	2024	Systematic Review	Multiple	DL ensembles recommended	No experimental design focus

C. Critical Review and Research Gaps

While machine learning and deep learning have substantially advanced predictive banking analytics, critical limitations persist. Most approaches remain model-centric with-out structured decision-making or experimental validation. Segmentation and prediction outcomes are rarely linked to product performance assessment, and formal statistical testing is largely absent. Many studies rely on synthetic, altered, or undisclosed datasets, constraining reproducibility and real-world applicability.

Furthermore, no existing framework integrates EDA, customer segmentation, and controlled A/B testing into a unified, actionable pipeline—particularly one tailored to the Indian banking context. Scalability and computational constraints of deep learning and ensemble methods further limit deployment in resource-constrained environments. These gaps motivate the proposed two-phase framework.

III. PROPOSED METHODOLOGY

A. System Overview

The proposed framework operationalises a two-phase pipeline. Phase 1 (Target Identification) integrates multi-source data and applies context-aware cleaning to identify the optimal customer segment. Phase 2 (Performance Validation) applies statistical experimental design and hypothesis testing to validate product efficacy.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718013

Table II contrasts the proposed data-driven approach with traditional methods across key decision points.

TABLE II TRADITIONAL VS. DATA-DRIVEN APPROACH: FRAMEWORK JUSTIFICATION

Decision Point	Traditional	Proposed	Justification
Target Market Selection	Surveys & intuition	Multi-source EDA	Reduces bias; improves reproducibility
Customer Profiling	Manual demographic analysis	Groupby binning, correlation	Reveals income-credit spending relationships
Null Value Treatment	Global imputation/deletion	Context-aware imputation	Preserves demographic variance
Outlier Treatment	Standard IQR/deletion	Modified IQR & domain rules	Retains valid extremes via business constraints
Sample Size Selection	Rule-of-thumb	Statistical power analysis	Statistically justified sample size
Control Group Design	Before-after comparison	Controlled experimental design	Enables causal inference; reduces bias
Performance Evaluation	Basic metrics	Hypothesis testing (Z-test, CI)	Ensures statistical significance
Launch Decision	Intuition-based	Evidence-based	Objective and defensible

B. Workflow

The framework follows a structured Input → Processing→ Output pipeline. In Phase 1, integrated customer data undergoes preprocessing (validation, missing value handling, outlier treatment), feature engineering, and EDA to identify the target segment. In Phase 2, the selected segment feeds into experimental validation via sample selection, controlled group design, and hypothesis testing, yielding a statistically validated launch decision.

C. Dataset Description

The framework operates on a real, non-synthetic banking dataset comprising approximately 40,000 records across three normalised tables:

- Customers Table: Customer ID, age, gender, location, occupation, marital status, annual income.
- Transactions Table: Transaction amount, payment type, platform, product category.
- Credit Score Table: Credit score, credit limit, credit utilisation ratio, outstanding debt.

IV. IMPLEMENTATION

This section presents the complete implementation of the two-phase framework using Python (pandas, NumPy, Matplotlib, Seaborn, SciPy, Statsmodels) in a Jupyter Notebook environment.

A. Phase 1: Data Cleaning and Exploratory Data Analysis

1) Cleaning annual_income: The customers dataset contained missing values in annual_income. A context-aware imputation strategy was applied: each missing value was replaced with the median income of the customer's occupation group, preserving demographic distributions and avoiding distortion from global imputation. Additionally, records with annual_income below $100 were treated as data-entry errors and replaced with the occupation-wise median.

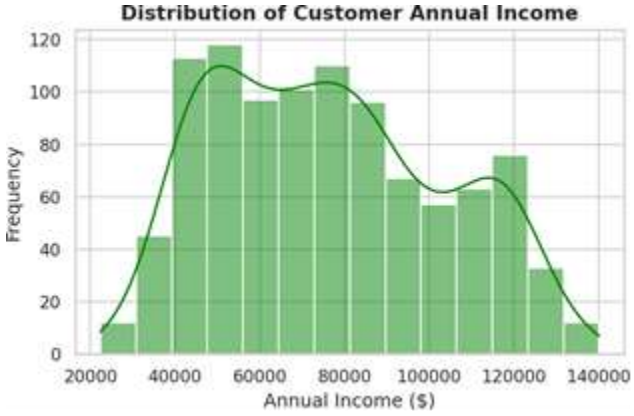

Fig. 1. Distribution of customer annual income after cleaning.

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718013

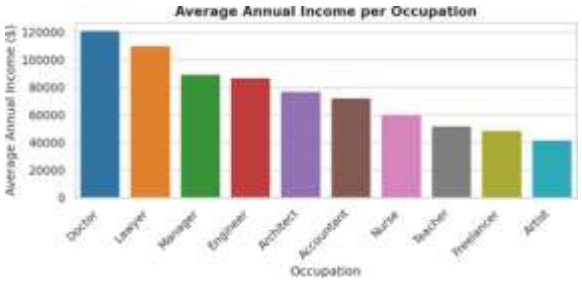

Fig. 2. Average annual income per occupation group.

2)Cleaning age: Age values outside the plausible range [15, 80] were replaced with the occupation-wise median age. Customers were then segmented into three groups using pd.cut(): 18–25 (Young Adults), 26–48 (Mid-Career), and 49–65 (pre-retirement).

3)Cleaning Credit Score Data: Duplicate cust_id entries were removed (retaining the last occurrence). Missing credit_limit values were filled using the modal credit limit within each credit score range bin (e.g., 700–749).

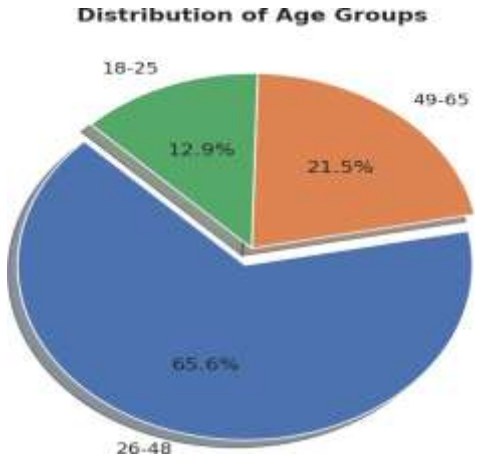

Fig. 3. Distribution of customer age groups.

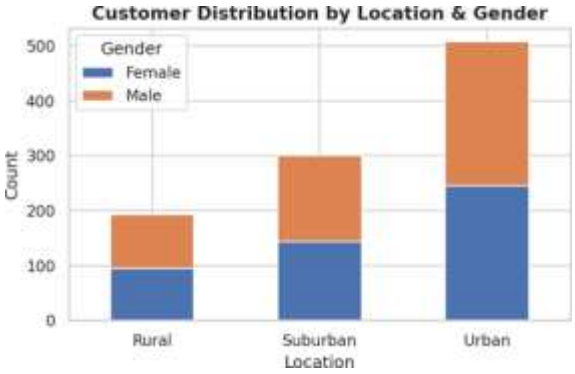

Fig. 4. Customer distribution by location and gender.

Outstanding debt values exceeding the credit limit were capped at the credit limit, enforcing domain logic.

Fig. 5. Credit score vs. credit limit after cleaning.

4) Cleaning Transactions Data: Missing platform values were filled with the mode. Zero-value transaction amounts were replaced with the median amount for the specific platform–category–payment type combination (Amazon, Electronics, Credit Card). A modified IQR rule (2× IQR) identified high-value outliers, which were replaced with the per-category mean, avoiding aggressive removal of valid large transactions.

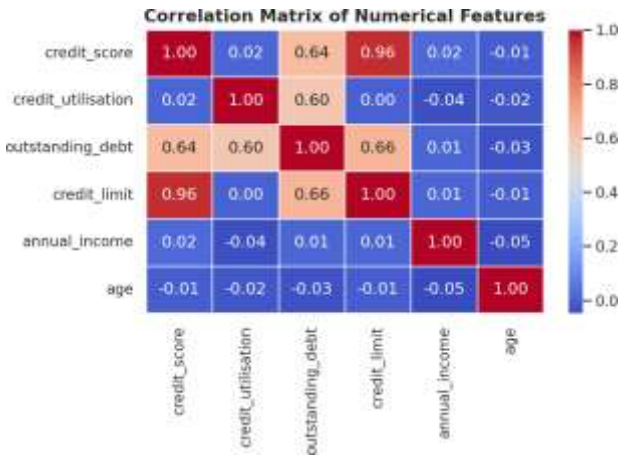

Fig. 6. Correlation matrix of key numerical features.

5) EDA: Cross-Dataset Insights: Following cleaning, the three datasets were merged on cust_id. Key visualisations of payment behaviour and financial profiles across age groups are presented below

© MAY 2026 | IRE Journals | Volume 9 Issue 11 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I11-1718013

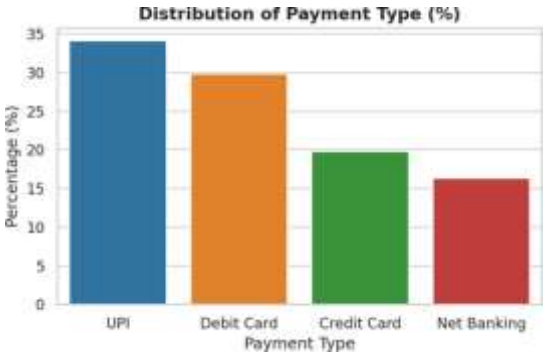

Fig. 7. Distribution of payment types across all transactions.

B. Target Segment Identification

EDA identified the 18–25 age group as the primary target segment based on five key findings:

1) This group constitutes approximately 25% of the total customer base.
2) Average annual income is below $50,000, indicating price sensitivity.
3) Credit scores and limits are lower due to limited credit history.
4) Credit card adoption is comparatively low—an untapped opportunity.

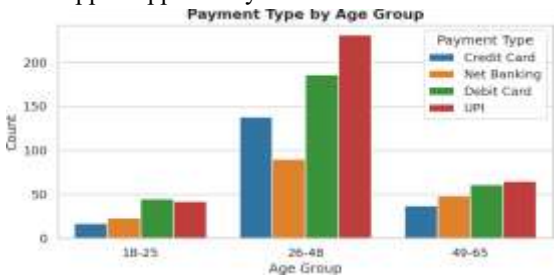

Fig. 8. Payment type usage by age group.

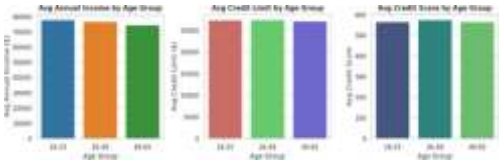

Fig. 9. Average annual income, credit limit, and credit score by age group.

5)Top product categories are Electronics, Fashion & Apparel, and Beauty & Personal Care.

These findings collectively justify targeting this segment with a tailored credit card product.

C.

D. Phase 2: A/B Testing and Statistical Validation

1)Sample Size Determination: Statistical power analysis was performed using tt_ind_solve_power (Statsmodels) at α = 0.05 and power = 0.80 across a range of effect sizes. Fig. 10 shows the required sample sizes. An effect size of 0.4 was selected, yielding 100 customers per group—within the available budget.

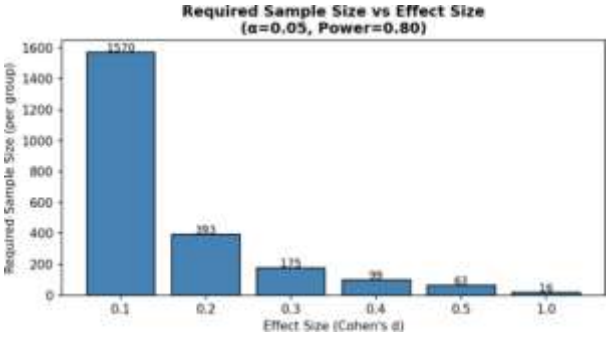

Fig. 10. Required sample size per group vs. effect size (α = 0.05, power= 0.80).

Of 246 eligible 18–25 customers, 100 were selected for the test campaign, run over a two-month period.

2)Campaign Design and Group Formation: The campaign achieved ≈40% conversion (40 of 100 customers adopted the card). A control group of 40 non-overlapping customers was formed for comparison. Daily average transaction amounts were recorded for all 80 participants over the campaign window (09-10-23 to 11-10-23).

3)Hypothesis Testing: Group transaction amount distributions were visualised prior to testing (Fig. 11).

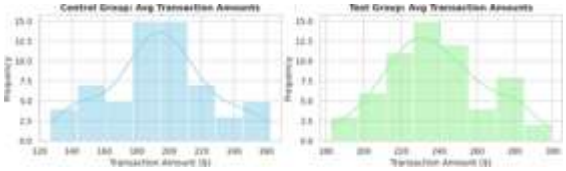

Fig. 11. Average transaction amount distributions: control group (left) vs. test group (right).

A two-sample Z-test was conducted:

- H0: μtest ≤ μcontrol (no campaign effect)
- H1: μtest > μcontrol (campaign increases spending)

Using the rejection region method, the computed Z-score exceeded Zcritical = 1.645 ($\alpha = 0.05$, one-tailed), rejecting H0. The p-value method confirmed $p < 0.05$. Both approaches were cross-validated using statsmodels.stats.ztest. The 95% confidence interval for the test group mean trans-action amount is $226–$245, directly supporting a full-scale product launch decision.

E. Summary of Implementation Results

Phase 1 produced a clean, analysis-ready dataset through context-aware imputation, domain-constrained outlier treatment, and multi-source integration. Phase 2 delivered a statistically validated product performance assessment via power-justified sampling, controlled experimental design, and hypothesis testing. The 18–25 segment shows a statistically significant and practically meaningful uplift in average transaction amounts following credit card adoption.

V. EXPECTED RESULTS AND DISCUSSION

A. Expected Outcomes

The proposed framework is expected to improve banking analytics decision-making by combining EDA with statistical validation. Multi-source data integration is anticipated to yield more accurate and meaningful customer segmentation than traditional single-source approaches. Context-aware preprocessing will enhance data quality and analytical reliability.

The framework is also expected to demonstrate robust-ness across diverse customer attributes (income, credit score, spending behaviour). Lightweight statistical techniques ensure scalability for real-world deployment in resource-constrained environments.

B. Comparative Evaluation

Framework effectiveness will be evaluated by comparison against traditional banking approaches and existing data-driven methods across three dimensions: accuracy of target segment identification, reliability of decision-making, and capacity to produce statistically validated results. Practical factors—implementation ease, computational efficiency, and result interpretability—will also be assessed.

C. Discussion

The proposed framework offers several advantages over existing approaches. By integrating data analysis, segmentation, and experimental validation into a single pipeline, it provides a systematic, evidence-based decision process. Hypothesis testing ensures statistical significance, while the lightweight design makes it suitable for real-world banking deployment. The framework bridges the gap between analytical modelling and business decision-making, enhancing both accuracy and applicability.

VI. APPLICATIONS AND USE CASES

A. Industry Application

The framework provides a reproducible blueprint for new-entrant banks in emerging markets to launch products with statistically validated, quantified risk. Operating entirely on existing data, it requires no additional data acquisition. The methodology extends directly to personal loans, savings ac-counts, and insurance product launches.

B. Social and Financial Inclusion

Data-driven identification of underserved urban segments promotes financial inclusion in markets such as India, where credit card penetration among young professionals remains significantly below developed-market levels. Targeted product design based on real data ensures products genuinely meet customer needs.

C. Policy and Regulatory Relevance

Regulators and policymakers can reference this framework to establish evidence-based standards for credit product launch validation. Requiring statistical evidence of product suitability before deployment reduces systemic risk and protects consumers from poorly designed financial products.

D. Academic Value

This framework contributes to academic literature by demonstrating that formal hypothesis testing and controlled experimental design should be integral to banking analytics research, identifying the absence of prospective A/B testing as a critical addressable gap,

and providing a replicable methodology for future researchers.

CONCLUSION

This paper proposed an integrated two-phase data-driven framework that operationalises the transition from multi-source data cleaning and EDA to prospective experimental validation. Applied to a real, non-synthetic banking dataset of approximately 40,000 records, the framework enables richer and more generalisable analytical outputs than existing synthetic-data-based approaches.

Context-aware cleaning—occupation-wise and credit-score-range-wise imputation, multi-filter median substitution, and modified IQR thresholds—ensures data integrity while pre-serving the demographic variance essential for meaningful segment differentiation. Phase 2 leverages the Central Limit Theorem to justify a two-sample Z-test on a sample of n = 40 within the large-scale population, providing robust statistical validation of product performance. The 95% confidence interval ($226–$245) translates the statistical result into a commercially interpretable spending range, directly supporting the full-scale launch decision.

The framework is affordable, reproducible, and adaptable to other financial product types. Future research will explore concept drift and temporal analysis to maintain framework accuracy under fluctuating market conditions, and will seek empirical validation across multiple banking institutions in emerging markets.

REFERENCES

[1] B. Ahatsi et al., “Study on electronic banking services and customer satisfaction,” 2023.
[2] F. Alarfaj et al., “Credit card fraud detection using machine learning and deep learning techniques,” 2022.
[3] A. Almazroi and N. Ayub, “Online payment fraud detection using machine learning models,” 2023.
[4] P. Beena et al., “Data science approach for mitigating credit card fraud,” 2021.
[5] C. Bogireddy and V. Murari, “Machine learning for predictive telemar-keting in banking,” 2024.
[6] R. Damanik and C. Liu, “Advanced fraud detection using SMOTE-based techniques and ensemble models,” 2025.
[7] A. Das, “Exploratory data analysis on financial stock data,” 2023.
[8] M. Haque et al., “Credit card fraud detection using supervised and ensemble learning,” 2024.
[9] N. Kalid et al., “Systematic review on fraud detection and payment defaults,” 2024.
[10] A. Khan et al., “Customer segmentation using K-means clustering,” 2022.
[11] L. Ling and K. Weiling, “Comparative study of clustering methods for customer segmentation,” 2025.
[12] A. Met et al., “Bank performance and target setting using AutoML and time series analysis,” 2023.
[13] R. Pandey et al., “Experimental analysis of customer segmentation using machine learning techniques,” 2023.
[14] S. Potluri et al., “Machine learning-based customer segmentation and personalised marketing,” 2024.
[15] T. Van Acker, “Survey of machine learning methods for fraud detection,” 2024.