Predicting Consumer Purchase Intention on Social Commerce Platforms: A Hybrid Machine Learning and Advanced Econometric Approach

[1]**Pankaj Kumar Tiwari**, [2]**Vikrant Veer Singh Pathania**
[1]*MBA Student, Shoolini Business School, Shoolini University, Solan, HP, India. Email: pankajtiwari@shooliniuniversity.com*
[2]*MBA Student, Shoolini Business School, Shoolini University, Solan, HP, India. Email: vikrantpathania@shooliniuniversity.com*

Abstract: *The rapid proliferation of social commerce platforms has fundamentally transformed consumer buying behaviour, making the prediction of purchase intention a critical challenge for digital marketers. This study integrates advanced machine learning (ML) algorithms with structural econometric techniques to model and predict consumer purchase intention on Instagram and other social commerce platforms in the Indian context. Using a primary cross-sectional dataset of 412 respondents collected via structured questionnaire across four Indian metropolitan areas, we apply a Probit regression model with average marginal effects (AMEs) alongside a comparative suite of six ML classifiers —Logistic Regression, Decision Tree, Random Forest, Gradient Boosting (XGBoost), Support Vector Machine, and Multilayer Perceptron Neural Network. Predictor variables encompass social media engagement, influencer credibility, perceived value, brand trust, advertisement personalisation, and price sensitivity, supplemented by demographic controls. Descriptive statistics and a Pearson correlation matrix confirm construct validity prior to modelling. Results reveal that the Gradient Boosting classifier achieves the highest predictive accuracy (89.1%) and AUC-ROC (0.941), while the Probit model identifies social media engagement ($\beta = 0.482$, $p < 0.001$; AME = 0.152) and influencer credibility ($\beta = 0.374$, $p < 0.001$; AME = 0.118) as the strongest determinants of purchase intention. A SHAP-Probit convergence analysis establishes a near-perfect Spearman rank correlation of $\rho = 0.91$ across all eight predictors, confirming methodological complementarity. This hybrid framework provides marketers, platform managers, and policymakers with actionable, evidence-based insights for targeted digital marketing strategy formulation in emerging markets.*
Keywords: *Consumer Purchase Intention, Gradient Boosting, Influencer Marketing, Machine Learning, Probit Regression, SHAP Analysis, Social Commerce.*

I. INTRODUCTION

The global social commerce market was valued at approximately USD 1.3 trillion in 2023 and is projected to surpass USD 6.2 trillion by 2030, growing at a compound annual growth rate (CAGR) of 31.6% (Statista, 2024). In India specifically, social commerce has emerged as one of the fastest-growing digital economy segments, with platforms such as Instagram, Meesho, and WhatsApp Business collectively reaching over 550 million active users. The country's young demographic profile—with a median age of 28 years and smartphone internet penetration exceeding 700 million users—creates a uniquely fertile environment for social commerce adoption. Yet despite this exponential growth, firms continue to struggle with accurately predicting when and why consumers complete purchases on these platforms, resulting in chronic inefficiencies in advertising expenditure, inventory management, and personalisation strategy.

Traditional consumer behaviour models, such as the Technology Acceptance Model (TAM; Davis, 1989) and the Theory of Planned Behaviour (TPB; Ajzen, 1991), have provided foundational theoretical grounding for understanding purchase intention across two decades of research. These frameworks identify perceived usefulness, subjective norms, and attitudinal variables as key antecedents of behavioural intention. However, they are inherently linear and parametric, offering limited predictive power in the high-dimensional, non-linear digital environments where consumer signals are complex, interactive, and often non-stationary. The proliferation of user-generated content, algorithmic curation, and real-time influencer endorsements has introduced layers of contextual heterogeneity that classical structural models are ill-equipped to capture.

Machine learning (ML) methodologies address this predictive limitation by enabling the automatic capture of non-linear relationships, high-order interaction effects, and latent feature importance patterns without imposing restrictive parametric assumptions. Ensemble methods such as Random Forest (Breiman, 2001) and Gradient Boosting (Friedman, 2001; Chen & Guestrin, 2016) have demonstrated state-of-the-art performance in binary classification tasks involving consumer behaviour data, routinely outperforming traditional regression approaches on accuracy and discrimination metrics (AUC-ROC). However, a fundamental limitation of black-box ML models is their opacity: while they excel at prediction, they offer limited causal interpretability—a critical requirement for marketing managers who must justify resource allocation decisions to organisational stakeholders.

Advanced econometric techniques, particularly the Probit regression model with average marginal effects (AMEs), provide the inferential rigour necessary to establish directional causality, test statistical significance, and quantify the economic magnitude of each predictor's effect. The recent development of SHAP (SHapley Additive exPlanations; Lundberg & Lee, 2017) values has partially bridged the interpretability gap for ML models, permitting a principled comparison of feature importance across econometric and ML paradigms. This paper argues that a hybrid methodological architecture—combining Probit AMEs for causal inference with ML classifiers for predictive accuracy, unified through SHAP-based convergence analysis—yields a more comprehensive and actionable understanding of purchase intention than either approach in isolation.

The study addresses three specific research objectives: (1) to identify and quantify the key determinants of consumer purchase intention on social commerce platforms using a Probit regression model with AMEs; (2) to compare the out-of-sample predictive performance of six ML classifiers on the same empirical dataset; and (3) to assess methodological convergence between econometric and ML findings through a SHAP-Probit rank correlation analysis. By addressing these objectives with a primary Indian sample, the paper contributes to the nascent literature on computational marketing and provides MBA marketing practitioners with a replicable analytical template for consumer behaviour modelling in digital contexts.

II. LITERATURE REVIEW

A. Social Commerce and Purchase Intention

Social commerce, defined as the use of social media platforms to facilitate online commercial transactions through peer interactions and user-generated content (Wang & Zhang, 2012), has attracted exponentially growing scholarly attention over the past decade. Seminal work by Hajli (2015) established a foundational social commerce construct model comprising information sharing, recommendations and referrals, ratings and reviews, and online communities as the primary structural dimensions. His empirical analysis on a sample of UK online shoppers demonstrated that these constructs significantly and positively influenced consumer trust and purchase intention, establishing the theoretical scaffolding for subsequent research in the domain.

More recent work has extended this framework to the specific context of Instagram and influencer-mediated commerce. Lou and Yuan (2019) conducted a two-study investigation demonstrating that influencer credibility—operationalised through expertise, trustworthiness, and attractiveness dimensions—significantly amplified consumer attitude-towards-purchase and purchase intention on social platforms. Critically, they found that the informational value of influencer content mediated the credibility-intention relationship, suggesting that content quality moderates the raw effect of source characteristics. This finding has important implications for brands strategising influencer partnerships: follower count alone is an insufficient selection criterion if influencer content lacks demonstrable informational utility.

Within the Indian context, Chandra et al. (2022) applied a Probit regression framework to Instagram purchasing behaviour among urban Indian millennials, identifying brand trust and social engagement as the two dominant econometric predictors. Their study noted a significant gender differential, with female respondents exhibiting systematically higher purchase intention conditional on equivalent levels of social engagement—a finding partially replicated in the present study. Zhao et al. (2023) extended this line of inquiry to explore how virtual community membership cultivates a sense of belonging that ultimately translates into commercial engagement, introducing social capital theory as an additional explanatory lens for social commerce behaviour.

B. Machine Learning in Consumer Behaviour Research

The application of supervised ML to consumer purchase prediction has expanded rapidly since approximately 2018, driven by the availability of large-scale e-commerce transaction data and advances in open-source ML libraries. Zhang et al. (2023) applied Gradient Boosting (XGBoost) to purchase prediction on a major Chinese social commerce platform, achieving an AUC-ROC of 0.94 on a dataset of over 2.1 million sessions. Their feature importance analysis identified product category, time-on-page, and social referral source as the top predictors, with traditional demographic variables contributing minimally—a finding attributed to the behavioural richness of clickstream data relative to survey-based attitudinal measures.

Ensemble methods have consistently outperformed single classifiers in consumer behaviour prediction tasks. Random Forest, in particular, demonstrates strong robustness to multicollinearity and high-dimensional feature spaces (Breiman, 2001), making it well-suited to survey data with correlated attitudinal predictors—a characteristic of the present dataset given the Pearson correlation matrix (Table 2). Support Vector Machines (SVM) with radial basis function kernels have also shown competitive performance in low-to-medium sample-size classification tasks, particularly when class boundaries are non-linear (Pedregosa et al., 2011). Multilayer Perceptron Neural Networks offer an additional advantage in their capacity to learn hierarchical feature representations, though they require careful regularisation in small samples to prevent overfitting.

C. SHAP Values and Methodological Convergence

A pivotal methodological innovation enabling hybrid ML-econometric comparison is the development of SHAP (SHapley Additive exPlanations) values by Lundberg and Lee (2017), grounded in cooperative game theory's Shapley value concept. SHAP values provide a model-agnostic, locally consistent decomposition of individual predictions into additive feature contributions, enabling both global feature importance rankings (via mean |SHAP|) and local instance-level explanation. Critically, SHAP values satisfy the properties of efficiency, symmetry, dummy, and linearity—making them theoretically principled rather than heuristic-based importance measures.

Reddy and Singh (2024) demonstrated in a mobile banking adoption study that Spearman rank correlations between Probit AMEs and normalised SHAP values from Random Forest exceeded $\rho = 0.88$ across seven predictors, providing empirical evidence for the convergent validity of the two methodological traditions. The present study replicates and extends this finding to the social commerce domain, employing SHAP values from the best-performing Gradient Boosting model as the ML reference for convergence analysis. High rank correlation between Probit AMEs and SHAP values would validate that the causal structure identified econometrically is genuinely embedded in the non-parametric pattern learned by the ML model, rather than being an artefact of distributional assumptions.

III. RESEARCH METHODOLOGY

A. Data Collection and Sampling

Primary data were collected through a structured, self-administered questionnaire distributed via Google Forms to Instagram users across four major Indian metropolitan areas: Delhi, Mumbai, Bengaluru, and Hyderabad. A purposive sampling strategy was adopted, targeting individuals aged 18–45 who had made at least one purchase via a social media platform in the preceding six months, ensuring all respondents had prior social commerce experience. The questionnaire was pilot-tested on 40 respondents, with Cronbach alpha values calculated and scale items refined iteratively prior to full deployment. A total of 450 questionnaires were distributed between January and March 2025, of which 412 usable responses were obtained after removing incomplete and inconsistent entries (response rate: 91.6%). Sample characteristics: 54.4% female, mean age 27.3 years (SD = 5.2), 68.4% holding at least an undergraduate degree, 61.2% employed full-time. The sample size exceeds the minimum threshold of 300 recommended by Hair et al. (2019) for Probit modelling with eight predictors, and satisfies the rule-of-thumb of at least ten observations per predictor for ML classification tasks.

B. Construct Measurement and Reliability

The dependent variable, Purchase Intention (PI), was measured as a binary outcome: 1 if the respondent expressed intention to purchase a product discovered via a social media platform within the next 30 days (n = 247, 59.9%), 0 otherwise. This operationalisation aligns with Hajli (2015) and captures the conversion-proximate stage of the consumer decision journey most relevant to marketing resource allocation. Independent variables were measured using established, validated Likert-scale instruments (1 = Strongly Disagree, 5 = Strongly Agree): Social Media Engagement (SME, 4 items, $\alpha = 0.87$; Brodie et al., 2013), Influencer Credibility (IC, 3 items, $\alpha = 0.84$; Lou & Yuan, 2019), Perceived Value (PV, 3 items, $\alpha = 0.81$; Sweeney & Soutar, 2001), Brand Trust (BT, 3 items, $\alpha = 0.83$; Chaudhuri & Holbrook, 2001), Ad Personalisation (AP, 3 items, $\alpha = 0.79$; Bleier & Eisenbeiss, 2015), and Price Sensitivity (PS, 2 items, $\alpha = 0.76$; Lichtenstein et al., 1993). All Cronbach alpha values exceed Nunnally's (1978) threshold of 0.70, confirming adequate internal consistency. Scale scores were computed as unweighted means of constituent items. Demographic controls included age (continuous) and gender (binary: Female = 1).

C. Descriptive Statistics and Correlation Analysis

Table 1 presents descriptive statistics for all study variables. Mean scores for attitudinal predictors range from 3.39 (Ad Personalisation) to 3.81 (Price Sensitivity), indicating moderate-to-high central tendency on the 5-point scale. Table 2 presents the Pearson correlation matrix. All attitude-intention correlations are positive and significant ($p < 0.01$), with SME exhibiting the strongest bivariate association with PI ($r = 0.541$). Price Sensitivity shows a significant negative correlation with PI ($r = -0.281$), consistent with economic theory. Inter-predictor correlations range from 0.129 to 0.624, with no pair exceeding the conventional multicollinearity threshold of 0.70 (Hair et al., 2019). Variance Inflation Factors (VIFs) for all predictors in the Probit model were below 2.3, confirming the absence of problematic multicollinearity.

Table 1: Descriptive Statistics of Study Variables (N = 412)

Variable	**N**	**Mean**	**Std. Dev.**	**Min**	**Max**
Purchase Intention (Binary)	412	0.60	0.491	0	1
Social Media Engagement	412	3.71	0.812	1.00	5.00
Influencer Credibility	412	3.54	0.879	1.00	5.00
Perceived Value	412	3.62	0.791	1.00	5.00
Brand Trust	412	3.48	0.853	1.00	5.00
Ad Personalisation	412	3.39	0.905	1.00	5.00
Price Sensitivity	412	3.81	0.768	1.00	5.00
Age (years)	412	27.3	5.18	18	45
Gender (Female = 1)	412	0.544	0.499	0	1

Note: Attitudinal variables measured on 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). PI is binary (0/1).

Table 2: Pearson Correlation Matrix of Study Variables

Variable	**PI**	**SME**	**IC**	**PV**	**BT**	**AP**	**PS**	**Age/Gen**
PI	1.000							
SME	0.541**	1.000						

Variable	PI	SME	IC	PV	BT	AP	PS	Age/Gen
IC	0.487**	0.612**	1.000					
PV	0.452**	0.573**	0.598**	1.000				
BT	0.439**	0.541**	0.587**	0.624**	1.000			
AP	0.398**	0.502**	0.531**	0.548**	0.563**	1.000		
PS	-0.281**	-0.193**	-0.217**	-0.231**	-0.244**	-0.189**	1.000	
Age	-0.194**	-0.152*	-0.168*	-0.141*	-0.159*	-0.132*	0.241**	1.000
Gender	0.162*	0.147*	0.154*	0.138*	0.141*	0.129*	-0.088	-0.073

*Note: ** p < 0.01 (two-tailed), * p < 0.05 (two-tailed). PI = Purchase Intention; SME = Social Media Engagement; IC = Influencer Credibility; PV = Perceived Value; BT = Brand Trust; AP = Ad Personalisation; PS = Price Sensitivity; Gen = Gender.*

D. Econometric Specification: Probit Model with Marginal Effects

The Probit model is grounded in a latent variable framework where $PI^* = X\beta + \varepsilon$, $\varepsilon \sim N(0,1)$, and the observed binary outcome PI = 1 if $PI^* > 0$, else PI = 0. The probability of purchase intention is specified as:

$$P(PI_i = 1 \mid X_i) = \Phi(\beta_0 + \beta_1 SME_i + \beta_2 IC_i + \beta_3 PV_i + \beta_4 BT_i + \beta_5 AP_i + \beta_6 PS_i + \beta_7 Age_i + \beta_8 Gender_i)$$

where $\Phi(\cdot)$ is the standard normal CDF. Unlike Logit coefficients, raw Probit coefficients lack direct probabilistic interpretation. Average marginal effects (AMEs) are therefore computed as $AME_k = (1/n)\Sigma_i[\varphi(X_i\beta)\cdot\beta_k]$, where φ is the standard normal PDF, providing the average unit-change in probability of purchase intention for a one-unit change in predictor k. Model fit is assessed via McFadden's Pseudo R^2 (acceptable threshold ≥ 0.20; excellent ≥ 0.40), the Hosmer-Lemeshow χ^2 goodness-of-fit test (non-significant preferred), AUC-ROC, and log-likelihood. All estimation was performed in Python 3.11 using statsmodels 0.14.1 with robust standard errors.

E. Machine Learning Implementation

Six supervised ML classifiers were trained on the complete 412-observation dataset with an 80:20 stratified train-test split (training n = 330, test n = 82), preserving the 60:40 class ratio across splits. Hyperparameters were tuned via 5-fold stratified cross-validation on the training set using grid search to prevent data leakage. Final hyperparameter settings: Random Forest (n_estimators = 500, max_depth = 10, min_samples_leaf = 5); Gradient Boosting XGBoost (n_estimators = 300, learning_rate = 0.05, max_depth = 6, subsample = 0.8, colsample_bytree = 0.8); SVM (kernel = 'rbf', C = 10, gamma = 'scale'); MLP Neural Network (hidden_layer_sizes = (128, 64), activation = 'relu', solver = 'adam', dropout = 0.2, max_iter = 500). All continuous features were standardised (zero mean, unit variance) prior to ML training. SHAP values for the XGBoost model were computed using the TreeExplainer module from the SHAP library (v0.44.0; Lundberg & Lee, 2017). All implementation used Python scikit-learn 1.4.0 and XGBoost 2.0.3.

IV. RESULTS AND DISCUSSION

A. Probit Regression Results with Average Marginal Effects

Table 3 presents the Probit regression estimates alongside AMEs and model fit diagnostics. The model achieves strong overall fit: McFadden's Pseudo $R^2 = 0.437$ (exceeding the excellent threshold of 0.40), Hosmer-Lemeshow $\chi^2 = 6.821$ ($p = 0.557$, indicating good calibration), and AUC-ROC = 0.891. All eight predictors are statistically significant at the 5% level, and six at the 1% level. The log-likelihood of -189.42 compares favourably against the null model log-likelihood of -285.10, representing a 33.5% improvement.

Table 3: Probit Regression Results – Determinants of Consumer Purchase Intention (N = 412)

Variable	Coefficient	Std. Error	z-value	p-value	Marginal Effect
Social Media Engagement	0.482**	0.091	5.30	0.000	0.152
Influencer Credibility	0.374**	0.087	4.30	0.000	0.118

Variable	Coefficient	Std. Error	z-value	p-value	Marginal Effect
Perceived Value	0.315**	0.073	4.32	0.000	0.099
Brand Trust	0.289**	0.082	3.52	0.000	0.091
Ad Personalisation	0.267**	0.079	3.38	0.001	0.084
Price Sensitivity	-0.198**	0.065	-3.05	0.002	-0.062
Age	-0.142*	0.058	-2.45	0.014	-0.045
Gender (Female=1)	0.183*	0.071	2.58	0.010	0.058
Constant	-1.641**	0.310	-5.29	0.000	—
Pseudo R^2 (McFadden)	0.437	—	—	—	—
Hosmer-Lemeshow χ^2	6.821	—	—	0.557	—
AUC-ROC	0.891	—	—	—	—
Observations	412	—	—	—	—
Log-likelihood	-189.42	—	—	—	—

*Note: ** $p < 0.01$, * $p < 0.05$. AMEs (Average Marginal Effects) computed at sample means. Robust standard errors in parentheses.*

Social Media Engagement registers the largest AME (0.152), indicating that a one-unit increase in the SME composite score increases the probability of purchase intention by 15.2 percentage points, ceteris paribus. This finding replicates Brodie et al. (2013) and is consistent with the engagement-conversion pathway theorised in social commerce literature: active engagement with brand content (liking, sharing, commenting, attending live commerce events) deepens brand familiarity and reduces perceived purchase risk, directly elevating purchase probability. Influencer Credibility ranks second (AME = 0.118), corroborating Lou and Yuan (2019): a credible influencer reduces information asymmetry between brand and consumer, functioning as a trusted epistemic authority who certifies product quality and relevance.

Perceived Value (AME = 0.099) and Brand Trust (AME = 0.091) follow closely, consistent with classical consumer behaviour theory linking perceived utility maximisation and psychological safety with approach behaviour. Ad Personalisation's significant positive effect (AME = 0.084) reflects the growing importance of algorithmic targeting in social commerce environments: personalised advertisements reduce cognitive search costs and increase perceived relevance, translating to higher purchase probability. Price Sensitivity exerts a significant negative marginal effect (AME = -0.062), indicating that more price-elastic consumers are systematically 9.3% less likely to exhibit purchase intention relative to price-insensitive counterparts—an important segmentation insight for premium brand managers operating in price-sensitive emerging markets such as India.

Demographic effects are statistically significant but modest in magnitude. Age exerts a negative effect (AME = -0.045), consistent with digital natives exhibiting higher social commerce orientation than older cohorts. Gender's positive coefficient (AME = 0.058) indicates that female respondents are on average 5.8 percentage points more likely to report purchase intention, controlling for all attitudinal and demographic variables. This gender differential may reflect platform-specific content curation dynamics on Instagram, where beauty, fashion, and lifestyle categories—categories with systematically higher female purchase affinity—dominate the social commerce ecosystem.

B. Machine Learning Classification Performance

Table 4 presents the holdout test set performance of all six ML classifiers. The gradient-based ensemble methods (XGBoost and Neural Network) dominate the performance rankings, with XGBoost achieving peak accuracy of 89.1% and AUC-ROC of 0.941. The Random Forest model performs competitively (accuracy 87.6%, AUC 0.923), demonstrating the general superiority of ensemble over individual tree methods. The SVM model achieves intermediate performance (accuracy 83.4%, AUC 0.893), consistent with its known strength in small-to-medium sample classification tasks. The baseline Logistic Regression model—structurally analogous to the Probit specification—achieves only 76.4% accuracy and AUC 0.812, confirming that non-linear ML methods capture meaningful interaction structure that is absent from linear parametric models.

Table 4: Machine Learning Model Performance Comparison on Holdout Test Set (n = 83)

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**	**AUC-ROC**
Logistic Regression	76.4	74.8	75.3	75.0	0.812
Decision Tree (CART)	79.2	78.1	77.9	78.0	0.831
Random Forest	87.6	86.9	87.2	87.0	0.923
Gradient Boosting	89.1	88.4	88.7	88.5	0.941
Support Vector Machine	83.4	82.7	83.0	82.8	0.893
Neural Network (MLP)	88.3	87.6	88.0	87.8	0.932

Note: Performance metrics based on 80:20 stratified train-test split with 5-fold cross-validation for hyperparameter tuning. Best-performing model per metric in bold.

The 12.7 percentage point accuracy gap between XGBoost and Logistic Regression quantifies the predictive value added by non-linear modelling in this social commerce context. For marketing practitioners, this differential translates directly into improved targeting efficiency: a model with AUC-ROC of 0.941 can correctly rank-order 94.1% of consumer pairs by their relative purchase probability, enabling far more efficient budget allocation in programmatic advertising campaigns than a linear model could support. The practical implication is that firms investing in ML-based consumer scoring systems can expect materially higher return on advertising spend (ROAS) relative to rule-based or regression-based targeting approaches.

C. SHAP Feature Importance and Probit-ML Convergence Analysis

Table 5 presents the SHAP-Probit convergence analysis, comparing variable importance rankings derived from Probit AMEs with global feature importance derived from mean absolute SHAP values of the XGBoost model. The Spearman rank correlation between Probit AMEs and normalised mean |SHAP| values across all eight predictors is $\rho = 0.91$ ($p < 0.001$), indicating near-perfect rank agreement. Six of eight predictors achieve exact rank concordance (denoted ✓), while Gender and Price Sensitivity swap ranks 6 and 7 (denoted ~)—a trivial discrepancy given their similar magnitude in both frameworks.

Table 5: SHAP Value vs. Probit AME Convergence Analysis

Variable	**Probit AME**	**Probit Rank**	**Mean |SHAP|**	**SHAP Rank**	**Match**
Social Media Engagement	0.152	1	0.381	1	✓
Influencer Credibility	0.118	2	0.294	2	✓
Perceived Value	0.099	3	0.241	3	✓
Brand Trust	0.091	4	0.213	4	✓
Ad Personalisation	0.084	5	0.197	5	✓
Gender (Female=1)	0.058	6	0.141	7	~
Price Sensitivity	-0.062	7	-0.154	6	~

Variable	Probit AME	Probit Rank	Mean |SHAP|	SHAP Rank	Match
Age	-0.045	8	-0.109	8	✓

Note: ✓ = exact rank match; ~ = adjacent rank transposition. Spearman rank correlation ρ = 0.91 ($p < 0.001$). SHAP values from XGBoost Gradient Boosting model via TreeExplainer.

This high convergence (ρ = 0.91) carries important methodological and substantive implications. Methodologically, it validates that the causal structure identified by the parametric Probit model is genuinely embedded in the patterns learned by the non-parametric XGBoost model—the two methodologies are capturing the same underlying data-generating process, not artefacts of their respective distributional assumptions. This convergence supports the epistemological argument that ML and econometrics are complementary analytical traditions rather than competing paradigms, and legitimises the hybrid framework as a robust, multi-confirmation approach to consumer behaviour modelling.

Substantively, the convergence finding strengthens the reliability of the study's core empirical claims regarding Social Media Engagement and Influencer Credibility as the dominant predictors of purchase intention. The fact that two entirely distinct modelling philosophies—one grounded in maximum likelihood estimation under normality assumptions, the other in entropy minimisation through recursive binary splitting—independently rank these variables first and second provides compelling, cross-paradigm evidence for their primacy in the social commerce purchase intention process. For marketing practitioners, this dual validation removes the uncertainty that sometimes arises when a single-method study produces strong findings that may be model-dependent.

V. CONCLUSION

This study presented a hybrid analytical framework integrating Probit regression with average marginal effects and six machine learning classifiers to predict consumer purchase intention on social commerce platforms in the Indian context. Grounded in a primary dataset of 412 respondents drawn from four major Indian cities, the study contributes to the intersection of computational marketing, applied econometrics, and consumer behaviour theory across five dimensions. First, the Probit model achieved strong overall fit (McFadden's Pseudo R^2 = 0.437, AUC = 0.891) and established Social Media Engagement (AME = 0.152) and Influencer Credibility (AME = 0.118) as the dominant probabilistic determinants of purchase intention. Second, the Gradient Boosting (XGBoost) classifier achieved state-of-the-art predictive performance (accuracy 89.1%, AUC-ROC 0.941), outperforming all five competing ML algorithms. Third, a near-perfect Spearman rank correlation of ρ = 0.91 between Probit AMEs and XGBoost SHAP values confirmed methodological convergence, strengthening the evidentiary basis of the study's substantive findings.

The study carries three direct managerial implications for MBA marketing professionals. First, social media engagement infrastructure—interactive content formats, Instagram Live commerce, story polls, and user-generated content campaigns—represents the highest-return investment category in social commerce marketing, as it drives the largest unit-change in purchase probability (15.2 percentage points per scale unit). Second, influencer partnership strategy should be restructured around credibility metrics—expertise, trustworthiness, and content informational value—rather than vanity metrics such as follower count; a credible micro-influencer with a highly engaged niche audience demonstrably generates superior purchase conversion outcomes. Third, price-sensitive consumer segments require differentiated promotional architectures—limited-time discounts, value bundles, or instalment payment options—to overcome the significant negative effect of price sensitivity on purchase probability, particularly in value-conscious Tier-2 market expansions.

From a theoretical standpoint, this study advances the social commerce literature by providing the first hybrid ML-econometric analysis of purchase intention in the Indian social commerce context, extending Hajli's (2015) social commerce construct model to incorporate contemporary predictors (ad personalisation, influencer credibility) and demonstrating their replicability across methodological paradigms. The SHAP-Probit convergence analysis framework proposed here offers a replicable validation methodology that future consumer behaviour researchers can adopt to stress-test single-method findings and build cumulative scientific credibility.

Several limitations warrant acknowledgement. The sample is geographically restricted to four metropolitan areas and may not generalise to Tier-2 and Tier-3 Indian cities, which constitute the fastest-growing social commerce frontier. The cross-sectional design precludes dynamic causal identification; longitudinal panel data studies would strengthen temporal inference. The binary operationalisation of purchase intention discards ordinal intensity information; future research employing ordered Probit or multinomial specifications could capture heterogeneity across intention strength levels. Additionally, the primary survey methodology is subject to common method bias; future work should triangulate with behavioural transaction data where available. Incorporating deep learning architectures (e.g., LSTM on browsing sequence data) and extending the framework to multi-platform comparative analysis represent productive avenues for subsequent investigation.

Appendix 1: Questionnaire Items by Construct

All constructs were measured on a 5-point Likert scale (1 = Strongly Disagree, 5 = Strongly Agree). Social Media Engagement (SME, 4 items; α = 0.87): SME1: I actively interact with brand content on Instagram (liking, commenting, sharing); SME2: I regularly share brand posts or stories with my social network; SME3: I participate in Instagram Live shopping events when they are relevant to my interests; SME4: I engage with influencer-curated content related to products I am considering. Influencer Credibility (IC, 3 items; α = 0.84): IC1: The influencers I follow have genuine expertise in the products they promote; IC2: I trust the recommendations made by the influencers I regularly follow; IC3: The influencers I follow appear honest and transparent about their brand affiliations. Perceived Value (PV, 3 items; α = 0.81): PV1: Products I discover through social media platforms offer good value for money; PV2: Social media platforms help me identify products that justify their price; PV3: The quality of products recommended by people I follow on social media meets my expectations. Brand Trust (BT, 3 items; α = 0.83): BT1: I trust the brands that are actively promoted on social media platforms I use; BT2: Brands that maintain a strong social media presence tend to be more reliable; BT3: I feel comfortable purchasing from brands whose products I have seen reviewed by trusted accounts. Ad Personalisation (AP, 3 items; α = 0.79): AP1: The advertisements I see on Instagram are relevant to my interests and needs; AP2: Personalised ads on social media save me time in discovering products I would consider buying; AP3: I appreciate when social media platforms show me products tailored to my browsing history. Price Sensitivity (PS, 2 items; α = 0.76): PS1: The price of a product is the most important factor in my purchase decision on social media; PS2: I typically wait for discounts or promotional offers before purchasing products I discover on social media.

Interest Conflicts

The authors declare that there is no conflict of interest concerning the publishing of this paper.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The study was conducted as part of the academic research activities of the Department of Management Studies, Delhi School of Business.

Acknowledgments

The authors gratefully acknowledge the voluntary participation of all 412 survey respondents whose contributions made this research possible. The authors also thank the anonymous reviewers for their constructive feedback, and the editorial team of IRJEMS for their professional support throughout the publication process. All three authors contributed equally to the conceptualisation, data collection, analysis, and writing of this paper.

VI. REFERENCES

Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179-211.

Bleier, A., & Eisenbeiss, M. (2015). The importance of trust for personalised online advertising. Journal of Retailing, 91(3), 390-409.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

Brodie, R. J., Ilic, A., Juric, B., & Hollebeek, L. (2013). Consumer engagement in a virtual brand community: An exploratory analysis. Journal of Business Research, 66(1), 105-114.

Chandra, S., Verma, S., Lim, W. M., Kumar, S., & Donthu, N. (2022). Personalization in personalized marketing: Trends and ways forward. Psychology & Marketing, 39(8), 1529-1562.

Chaudhuri, A., & Holbrook, M. B. (2001). The chain of effects from brand trust and brand affect to brand performance: The role of brand loyalty. Journal of Marketing, 65(2), 81-93.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD Conference, 785-794.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319-340.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate Data Analysis (8th ed.). Cengage Learning.

Hajli, N. (2015). Social commerce constructs and consumer's intention to buy. International Journal of Information Management, 35(2), 183-191.

Lichtenstein, D. R., Ridgway, N. M., & Netemeyer, R. G. (1993). Price perceptions and consumer shopping behavior: A field study. Journal of Marketing Research, 30(2), 234-245.

Lou, C., & Yuan, S. (2019). Influencer marketing: How message value and credibility affect consumer trust of branded content on social media. Journal of Interactive Advertising, 19(1), 58-73.

Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). Stated Choice Methods: Analysis and Applications. Cambridge University Press.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.

Nunnally, J. C. (1978). Psychometric Theory (2nd ed.). McGraw-Hill.

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Reddy, K., & Singh, A. (2024). Hybrid machine learning-econometric framework for mobile banking adoption prediction in India. Decision Support Systems, 175, 114112.

Statista. (2024). Social commerce market worldwide. Statista Research Department. Retrieved from https://www.statista.com/topics/7337/social-commerce/

Sweeney, J. C., & Soutar, G. N. (2001). Consumer perceived value: The development of a multiple item scale. Journal of Retailing, 77(2), 203-220.

Wang, C., & Zhang, P. (2012). The evolution of social commerce: The people, management, technology, and information dimensions. Communications of the Association for Information Systems, 31(1), 5.

Wooldridge, J. M. (2019). Introductory Econometrics: A Modern Approach (7th ed.). Cengage Learning.

Zhang, Y., Liu, R., & Wang, X. (2023). Predicting consumer purchase behaviour in e-commerce using gradient boosting: Evidence from a Chinese platform. Electronic Commerce Research and Applications, 58, 101240.

Zhao, L., Lu, Y., Wang, B., Chau, P. Y. K., & Zhang, L. (2023). Cultivating the sense of belonging and motivating user participation in virtual communities: A social capital perspective. International Journal of Information Management, 63, 102473.