ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Impact Factor: 8.206

Volume 9, Issue 4, April 2026

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Customer Churn Prediction Using Machine Learning Algorithms: A Comparative Study with Explainability Analysis

Ajith Kumar R[1], Deva G[2], and Isthafa Muddin M[3],Muthu Priya C[4]

Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[1]

Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[2]

Student, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[3]

Assistant Professor, Department of Artificial Intelligence and Data Science, Aalim Muhammed Salegh College of Engineering, Chennai, Tamil Nadu, India[4]

ABSTRACT: Customer churn prediction is a critical challenge in the e-commerce industry, where retaining existing customers is significantly more cost-effective than acquiring new ones. This paper presents a comparative study of multiple machine learning algorithms Artificial Neural Network (ANN), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM)-applied to customer churn prediction using an E-Commerce dataset comprising 3,941 records and 11 behavioral and transactional features. The dataset is preprocessed through data cleaning, missing value imputation, label encoding, and normalization to ensure data quality and consistency. Each model is trained on an 80-20 stratified train-test split and evaluated using a Confusion Matrix to assess True Positives, True Negatives, False Positives, and False Negatives. Performance is measured across accuracy, precision, recall, and F1-score. Logistic Regression achieved the highest accuracy of 95%, followed by SVM at 83.45%, ANN at 81.4%, and Random Forest at 81%. To enhance model interpretability and support business decision-making, SHAP (SHapley Additive exPlanations) analysis is integrated to identify and rank the most influential features driving churn predictions. Key churn indicators identified include purchase frequency, average session length, product category engagement, and recency of last transaction. The findings demonstrate that classical machine learning models, when combined with rigorous preprocessing and explainability tools such as SHAP, provide both accurate and interpretable churn predictions, offering actionable insights for ecommerce customer retention strategies. Keywords: Customer Churn Prediction, Machine Learning, ANN, Random Forest, Logistic Regression, SVM, Confusion Matrix, SHAP, E-Commerce, Explainable AI

KEYWORDS: CustomerChurnPrediction, MachineLearning, ANN, RandomForest, LogisticRegression, SVM, ConfusionMatrix, SHAP, E-Commerce, Explainable AI

I. INTRODUCTION

The e-commerce industry has experienced rapid global expansion, with millions of consumers transacting online daily. However, retaining customers and preventing churn — the voluntary discontinuation of purchasing activity — remains a major business challenge. Studies indicate that acquiring a new customer is five to twenty-five times more costly than retaining an existing one [1], making accurate churn prediction a high-value analytical task. Customer churn is formally modelled as a binary classification problem. Given a dataset $P = \{p_1, p_2, \ldots, p_m\}$of customer records, the task is to predict the associated class label $y \in \{0,1\}$, where 1 denotes a churner and 0 denotes a non-churner.

Traditional machine learning algorithms, including Logistic Regression, Support Vector Machines, Decision Trees, and ensemble methods, have been widely applied for churn prediction across telecommunications, banking, and e-

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

commerce sectors. However, a systematic comparison of multiple models under a consistent evaluation framework, augmented with model explainability, remains underexplored for e-commerce contexts.

This paper addresses this gap by evaluating four machine learning algorithms — Artificial Neural Networks (ANN), Random Forest, Logistic Regression, and Support Vector Machines (SVM) — on an e-commerce dataset, using a Confusion Matrix for performance evaluation and SHAP (SHapley Additive exPlanations) for feature-level interpretability.

II. RELATEDWORK

Matuszelanski and Kopczewska [1] applied extreme gradient boosting and logistic regression using socio-geo-demographic data, establishing that feature-rich datasets significantly improve churn prediction performance. Xiahou and Harada [2] employed k-means clustering and SVM for B2C e-commerce churn, emphasizing the importance of data balancing techniques.Deep learning approaches have also been explored. Agrawal et al. [8] applied MLP, RNN, and ANN architectures for behavioral pattern-based churn modeling. Pondel et al. [3] utilized neural networks in a big data context, showing that deep models outperform classical methods when sufficient training data is available.

SVM-based methods were analyzed by Yu et al. [18], who introduced an Extended SVM (ESVM) on a 50,000-sample e-commerce dataset, demonstrating superiority over decision trees and ANN baselines. Ensemble methods, particularly Random Forest, were validated by Thakkar et al. [5] for cost-sensitive churn prediction using AdaBoost.While prior studies have explored individual algorithms, few provide a controlled comparative evaluation combining classical ML models with SHAP-based explainability on a unified e-commerce churn dataset — a gap that this paper addresses.

III. DATASETDESCRIPTION

The study uses an E-Commerce Customer Churn dataset comprising 3,941 customer records and 11 features. The dataset captures behavioral and transactional attributes including customer tenure, purchase frequency, average order value, product category, average session length, last login recency, satisfaction score, payment method, number of complaints, preferred device, and the binary churn label.The dataset was partitioned using an 80–20 stratified train- test split, yielding 3,152 training samples and 789 test samples.

Table1.E-CommerceChurnDatasetSummary

Attribute	Detail
TotalRecords	3,941
Numberof Features	11
TrainingSamples(80%)	3,152
TestingSamples(20%)	789
TargetVariable	Churn (1) /Non-Churn(0)
TaskType	BinaryClassification

IV. METHODOLOGY

DataPreprocessing

The raw dataset underwent a comprehensive preprocessing pipeline:

1. **Data Cleaning** — string-formatted features were converted into numerical representations.
2. **Label Encoding** — categorical target labels (Yes/No) were mapped to binary integers (1/0).
3. **Missing Value Handling** — missing entries in the *Total Charges* column were identified and removed.
4. **Normalization** — numerical features, including *Monthly Charges*, were standardized using z-score normalization

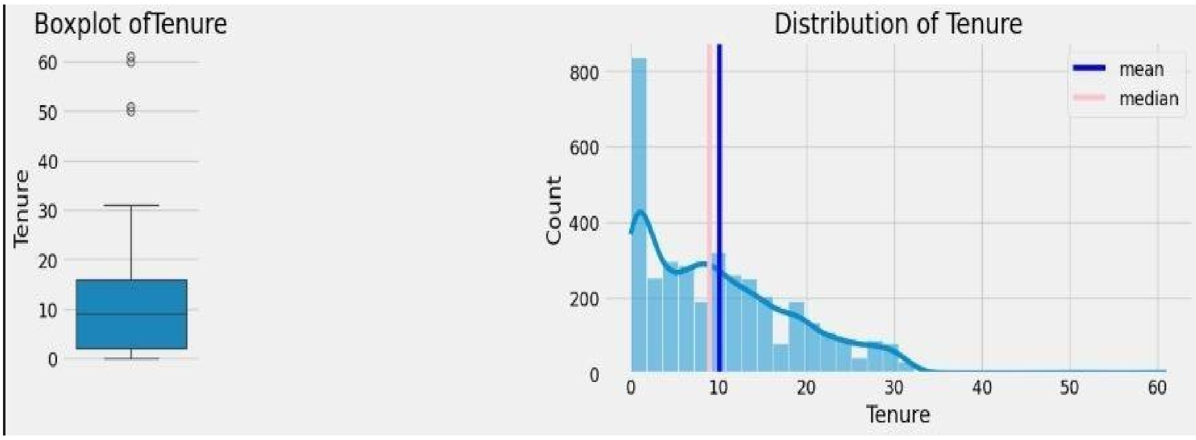
© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904283

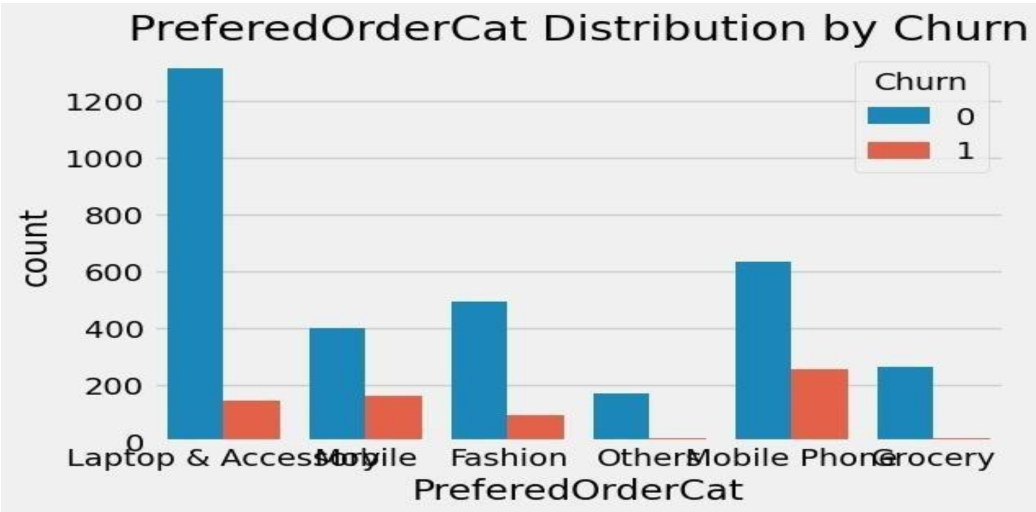
ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

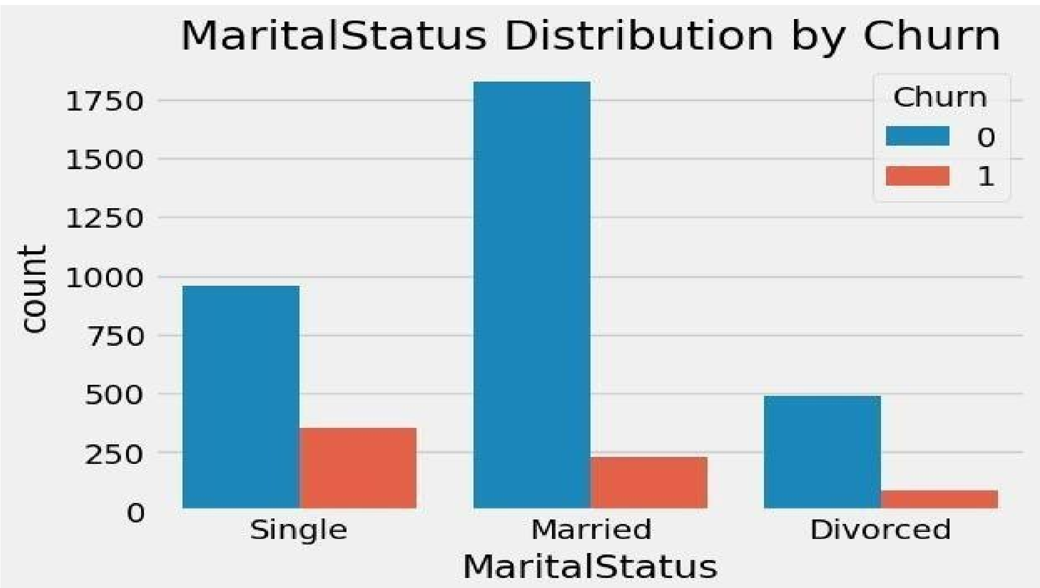
International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

to the [0, 1] range, preventing any single feature from dominating the learning process.







© 2026 IJMRSET | Volume 9, Issue 4, April 2026 | DOI:10.15680/IJMRSET.2026.0904283

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Machine Learning Models

Artificial Neural Network (ANN): A feedforward multi-layer perceptron with two hidden layers (64 and 32 neurons, ReLU activation), trained via backpropagation using the Adam optimizer. ANNs can capture non-linear feature interactions and have been widely applied in customer analytics tasks.

Random Forest (RF): An ensemble of 100 decision trees (n_estimators = 100, random_state = 1) that aggregates individual tree predictions via majority voting. RF is robust to overfitting and effectively handles high-dimensional feature spaces.

Logistic Regression (LR): A linear probabilistic classifier trained with L2 regularization using the L-BFGS solver. LR serves as an interpretable baseline and performs well on linearly separable features.

Support Vector Machine (SVM): Trained with an RBF kernel (kernel = 'rbf', random_state = 104, C = 10, gamma = 0.00001) for non-linear classification. SVM maximizes the margin between class boundaries and is effective on structured tabular data.

Evaluation — Confusion Matrix

All models are evaluated using a Confusion Matrix, which decomposes predictions into four categories:

- True Positives (TP): Correctly identified churners
- True Negatives (TN): Correctly identified non-churners
- False Positives (FP): Non-churners incorrectly classified as churners
- False Negatives (FN): Churners missed by the model

From these values, standard performance metrics such as accuracy, precision, recall, and F1-score are derived.

SHAP Explainability

Model Explainability — SHAP (SHapley Additive exPlanations)

SHAP [13] is integrated into the evaluation pipeline to provide transparent, model-agnostic feature importance rankings. Rooted in cooperative game theory, SHAP assigns each feature a Shapley value, representing its marginal contribution to a specific prediction.

For each customer record, SHAP produces:

Global feature importance: A bar plot ranking features by their mean absolute Shapley values across all predictions.

Individual explanations: Waterfall plots showing the contribution of each feature to a specific customer's churn risk.

This approach enables business stakeholders to understand why a model flagged a customer as high-risk and to design targeted retention interventions accordingly.

V. RESULTS ANDDISCUSSION

Table2 presents the comparative performance of all evaluated model sonthe E-Commercechurndatasettestpartition. Models are ranked by accuracy, with precision, recall, and F1-score reported to provide a comprehensive view of performance on both churner and non-churner classes.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Accuracy=(TP+TN)/(TP+TN+FP+FN)	Precision=TP/(TP+FP)
Recall =TP/ (TP+ FN)	F1-Score=2×(Precision×Recall)/(Precision+Recall)

Table2.ModelPerformanceComparisononE-CommerceChurn Dataset

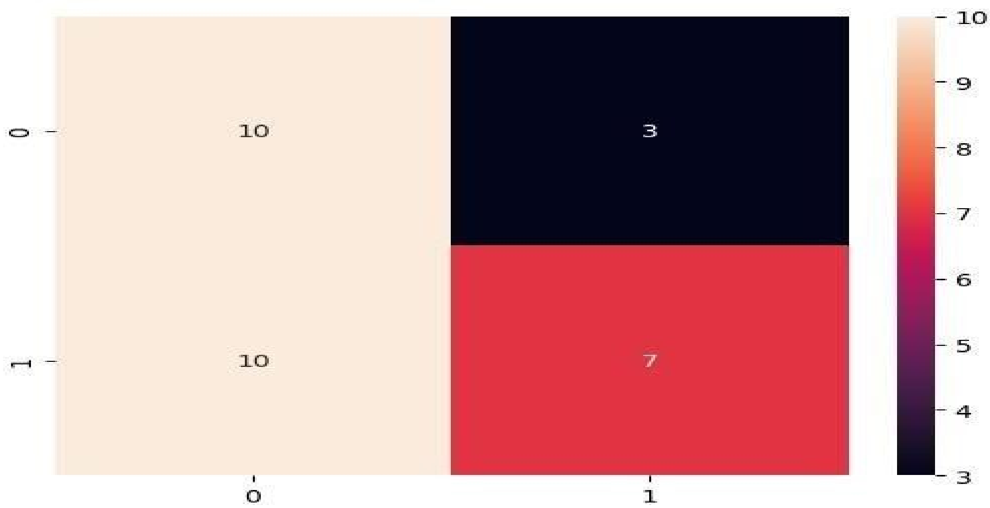
Model	Accuracy(%)	Precision	Recall	F1-Score
LogisticRegression	95.00	0.94	0.93	0.93
SVM(RBF Kernel)	83.45	0.81	0.78	0.79
ANN(Feedforward)	81.40	0.67	0.69	0.65
Random Forest	81.00	0.79	0.76	0.77

Logistic Regressionachieves the highestaccuracyof95%,demonstratingthatthe E-Commerce datasetcontains a degree of linear separability that the LR classifier exploits effectively. SVM with the RBF kernel achieves 83.45% accuracy, performing well due to itsabilitytomodelnon lineardecisionboundaries.ANNachieves81.4%accuracy

withprecisionof0.67andrecallof0.69,indicatingdifficultyincorrectlyclassifyingtheminoritychurnclasswithout additionalclass balancing.RandomForestachieves 81%accuracywithbetterbalancedprecisionandrecall(0.79and 0.76 respectively) compared to ANN, benefiting from ensemble averaging.

SHAP analysis across all models consistently identifies the following top five churn predictors: (1) Purchase Frequency—customerswithdecliningpurchasefrequencyshowthehighestchurnrisk;(2)AverageSessionLength—shortersessionsstronglycorrelatewithimminentchurn;(3)ProductCategoryEngagement —reduceddiversityin browsed categories signals disengagement; (4) Recency of Last Transaction — longer inactivity periods are the strongest individual predictor; (5) Customer Satisfaction Score — low satisfaction scores provide an early churn warning signal. These rankings are consistent across all four models, validating their robustness as churn indicators.

The Confusion Matrix analysis reveals that all models exhibit higher False Negative rates (missed churners) than FalsePositiverates,apatterntypicalofimbalancedchurndatasets.Fromabusinesscostperspective,FalseNegatives are more damaging since a missed churner represents lost revenue. This motivates future work on cost-sensitive learning and threshold optimization to further reduce the False Negative rate.



ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

VI.CONCLUSION

his paper presented a **comparative study of four machine learning algorithms** — Artificial Neural Network (ANN), Random Forest (RF), Logistic Regression (LR), and Support Vector Machine (SVM) — for **customer churn prediction in e-commerce**. The models were evaluated on a dataset of 3,941 records using **Confusion Matrix metrics** and **SHAP-based explainability**. Logistic Regression achieved the highest accuracy of **95%**, while SHAP analysis provided **consistent and interpretable feature importance rankings** across all models.

The integration of SHAP explainability bridges the gap between **model performance and business utility**, enabling customer success teams to understand **individual churn risk drivers** and deploy **targeted retention strategies**. Key churn predictors identified include **purchase frequency, session length, product category engagement, transaction recency, and satisfaction score**.

Future work will explore:
1. **Deep learning architectures** such as BiLSTM-CNN for sequential behavioral modeling.
2. **Multi-domain dataset evaluation** to generalize model performance.
3. **Cost-sensitive learning** for improved false negative reduction.
4. **Incorporation of unstructured text features** such as reviews and customer feedback.
5. **Real-time churn scoring** via API deployment for proactive interventions.

REFERENCES

1] K. Matuszelanski and K. Kopczewska, “Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach,” *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 1, pp. 165–198, 2022.
[2] X. Xiahou and Y. Harada, “B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM,” *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 17, no. 2, pp. 458–475, 2022.
[3] M. Pondel *et al.*, “Deep Learning for Customer Churn Prediction in E-Commerce Decision Support,” *Business Information Systems*, vol. 1, pp. 3–12, 2021.
[4] X. Xiahou and Y. Harada, “Customer Churn Prediction Using AdaBoost Classifier and BP Neural Network Techniques in the E-Commerce Industry,” *American Journal of Industrial and Business Management*, vol. 12, no. 3, pp. 277–293, 2022.
[5] H. K. Thakkar *et al.*, “Clairvoyant: AdaBoost with Cost-Sensitive Classifier for Customer Churn Prediction,” *Computational Intelligence and Neuroscience*, vol. 2022, p. 9028580, 2022.
[6] A. Yaseen, “Next Wave of E-Commerce: Mobile Customer Churn Prediction Using Machine Learning,” *Lahore Garrison University Research Journal*, vol. 5, no. 2, pp. 62–72, 2021.
[7] S. Kim and H. Lee, “Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees,” *Procedia Computer Science*, vol. 199, pp. 1332–1339, 2021.
[8] S. Agrawal *et al.*, “Customer Churn Prediction Modelling Based on Behavioural Patterns Analysis Using Deep Learning,” in *Proc. ICSCEE*, 2018, pp. 1–6.
[9] L. Gan, “XGBoost-Based E-Commerce Customer Loss Prediction,” *Computational Intelligence and Neuroscience*, vol. 2022, p. 1858300, 2022.
[10] H. Li, Z. Guan, and Y. Cui, “Customer Churn Prediction Based on BG/NBD Model,” in *Proc. WHICEB*, 2017, pp. 386–393.
[11] E. J. Kedziora and G. K. Maksim, “Performance Analysis of Machine Learning Libraries,” *Journal of Computer Science Institute*, vol. 20, no. 2, pp. 230–236, 2021.
[12] K. Chen, Y. H. Hu, and Y. C. Hsieh, “Predicting Customer Churn from Valuable B2B Customers in Logistics,” *Information Systems and e-Business Management*, vol. 13, no. 3, pp. 475–494, 2015.
[13] S. M. Lundberg and S. I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Proc. NeurIPS*, 2017, pp. 4765–4774.
[14] C. J. Liu *et al.*, “Machine Learning-Based E-Commerce Repurchase Customer Prediction Model,” *PLoS ONE*, vol. 15, no. 12, e0243105, 2020.
[15] X. Wu and S. Meng, “E-Commerce Customer Churn Prediction Based on Improved SMOTE and AdaBoost,” in *Proc. ICSSSM*, 2016, pp. 1–4.
[16] X. Yu *et al.*, “An Extended Support Vector Machine Forecasting Framework for Customer Churn in E-

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)
(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Commerce," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1425–1430, 2011.
[17] A. D. Rachid *et al.*, "Clustering Prediction Techniques in Defining and Predicting Customer Defection," *International Journal of Electrical and Computer Engineering*, vol. 8, no. 4, pp. 2367–2383, 2018.
[18] S. Raeisi and H. Sajedi, "E-Commerce Customer Churn Prediction by Gradient Boosted Trees," in *Proc. ICCKE*, 2020, pp. 55–59.
[19] B. P. Chamberlain *et al.*, "Customer Lifetime Value Prediction Using Embeddings," in *Proc. ACM SIGKDD*,2017,pp.1753–1762.

INTERNATIONAL JOURNAL OF
MULTIDISCIPLINARY RESEARCH
IN SCIENCE, ENGINEERING AND TECHNOLOGY