IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 4, Issue 1, September 2024
Impact Factor: 7.53

To Examine the Impact of Key Data Characteristics on the Performance of Machine Learning Techniques in E-Commerce

Richa Mishra and Dr. Rita K. Saini
Research Scholar, Himalayiya University, Dehradun, Uttarakhand, India
Supervisor, Himalayiya University, Dehradun, Uttarakhand, India
richamishra767@gmail.com

Abstract: *The accelerating growth of online retail has made machine learning (ML) a central instrument for tasks such as purchase-intent prediction, customer-churn detection, recommendation and fraud screening. While a large body of work compares classifiers on e-commerce data, comparatively little attention is paid to a more fundamental question: how do the intrinsic characteristics of the data itself govern which algorithm performs best? This study systematically examines the impact of five key data characteristics—dataset size, class imbalance, feature dimensionality, missing values and label noise—on the predictive performance of eight widely used ML techniques: Logistic Regression (LR), Naïve Bayes (NB), K-Nearest Neighbours (KNN), Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN) and Extreme Gradient Boosting (XGBoost). Using a controlled experimental design built on a representative e-commerce purchase-prediction task, each characteristic is varied independently while all other factors are held constant, and performance is measured with accuracy, precision, recall, F1-score and the area under the ROC curve (AUC). The results show that ensemble methods—particularly XGBoost and Random Forest—are the most robust across nearly every data condition, that distance-based learners such as KNN degrade sharply under high dimensionality, and that class imbalance is the single most damaging characteristic, collapsing minority-class F1 for linear and probabilistic models far more than for boosted trees. A sensitivity ranking is derived to guide practitioners in matching algorithms to the data conditions they actually face. The findings reinforce that, in e-commerce analytics, careful characterisation and conditioning of the data frequently yields larger gains than the choice of algorithm alone.*

Keywords: Machine learning; e-commerce analytics; data characteristics; class imbalance; dimensionality; data quality; missing values; label noise; XGBoost; Random Forest; model robustness; predictive analytics

Abbreviations. Acronyms used throughout this paper.

Acronym	Meaning	Acronym	Meaning
ML	Machine Learning	RF	Random Forest
LR	Logistic Regression	ANN	Artificial Neural Network
NB	Naïve Bayes	XGB	Extreme Gradient Boosting
KNN	K-Nearest Neighbours	AUC	Area Under the ROC Curve
DT	Decision Tree	F1	F1-score (harmonic mean of P/R)
SVM	Support Vector Machine	RFM	Recency–Frequency–Monetary

DOI: 10.48175/IJARSCT-19600F