Journal of Consumer Behavior, Marketing, and Management (JCBMM)

ISSN: 3070-5673 (ONLINE)

VOLUME 2 ISSUE 1 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

Journal of Consumer Behavior Marketing, and Management (JCBMM)
Volume 2 Issue 1, Year 2026
ISSN: 3070-5673 (Online)
DOI: https://doi.org/10.54536/jcbmm.v2i1.6573
https://journals.e-palli.com/home/index.php/jcbmm

Predicting Customer Satisfaction Under Severe Class Imbalance in Online Retail Data

Felix Reichel[1*]

Article Information

Received: September 14, 2025

Accepted: December 29, 2025

Published: July 23, 2026

Keywords

Customer Satisfaction, E-Commerce Analytics, Imbalanced Classification, Logistic Regression; Random Forests, Support Vector Machines

ABSTRACT

This study examines the prediction of post-purchase customer satisfaction in an e-commerce setting using order-level transactional and behavioral data. Customer ratings from one to five stars are grouped into three classes representing low, medium, and high satisfaction. The resulting classification task shows a strong skew toward high ratings. Three supervised learning models are evaluated: a support vector machine with a radial basis function kernel, multinomial logistic regression with spline-expanded numeric predictors, and a random forest classifier. Input variables include demographic attributes, order characteristics, browsing behavior, and delivery information. Model performance is assessed using accuracy, macro-averaged precision, recall, F1-score, ROC-AUC, and confusion matrices. The random forest achieves the highest accuracy by predicting the dominant class, while the support vector machine and spline-based logistic regression show lower accuracy with more balanced class-wise results. The findings indicate that distinguishing dissatisfied and moderately satisfied customers remains difficult with the available feature set and label design

INTRODUCTION

Post-purchase ratings summarize customer perception of the overall shopping experience and influence future purchasing behavior (Oliver, 1997; Rust *et al.*, 2000; Kotler & Keller, 2016). This imbalance complicates model training and evaluation, as standard loss functions and accuracy-based metrics tend to favor the dominant class (He and Garcia, 2009).

The objective of this work is to study how standard supervised learning models behave under such conditions. Three commonly applied methods are selected: support vector machines, logistic regression, and random forests. These models represent different inductive biases and levels of model flexibility, which allows a comparison of their behavior on a skewed three-class target.

LITERATURE REVIEW

Customer satisfaction prediction has been examined across marketing research, operations management, and information systems. Foundational empirical work commonly applied linear and logistic regression to relate transactional features and customer demographics to satisfaction measures. These approaches were favored because model parameters could be directly interpreted and estimation relied on well-established statistical procedures (Kotler & Keller, 2016, Oliver, 1997).

The development and commercialization of machine learning techniques broadened the methodological toolkit available to managerial economists and researchers. Support vector machines introduced margin-based classification combined with kernel functions, enabling the modeling of complex, nonlinear class boundaries beyond the scope of traditional regression (Cortes & Vapnik, 1995). Subsequent research explored ensemble learning techniques, with random forests receiving particular attention due to their capacity to model interaction effects and non-linear patterns directly from data, while reducing the need for extensive feature specification by analysts (Breiman, 2001).

Related recent streams of work addressed practical challenges encountered in customer analytics. Studies on imbalanced data highlighted the impact of skewed class distributions on predictive performance, a frequent issue in satisfaction and churn datasets (He & Garcia, 2009). Other contributions incorporated relational and network-based information to enhance predictive accuracy in customerrelated outcomes, demonstrating the relevance of social structure alongside individual-level attributes (Verbeke, Martens, & Baesens, 2014). Collectively, these advances reflect a progression from parametric statistical models toward more flexible data-driven methods in the study of customer satisfaction and related behaviors (Rust, Zeithaml, & Lemon, 2000).

METERIALS AND METHODS

Research Questions

RQ1: How do an RBF support vector machine, spline-based multinomial logistic regression, and a random forest compare on a three-level customer satisfaction task?

RQ2: How does class imbalance affect prediction results across these models?

RQ3: Which model offers the most balanced performance across satisfaction levels?

Data

The dataset consists of order-level observations from an online retail environment. Each record contains customer

[1] University of Linz, Linz, Austria
[*] Corresponding author's e-mail: felix.reichel@jku.at

demographics, transaction details, browsing behavior, and delivery outcomes. Identifiers are removed prior to analysis. Time stamps are decomposed into calendar-based features. Missing numeric values are replaced with feature-wise medians, while missing categorical values are replaced with the most frequent category. Numeric variables are standardized, and categorical variables are encoded using one-hot representations.

Models and Algorithms

Three classifiers are examined. The first is a support vector machine with a radial basis function kernel and class-weighted loss. The second applies cubic spline transformations to numeric predictors followed by multinomial logistic regression. The third is a random forest classifier with 200 trees, Gini impurity, and inverse-frequency class weights.

Experimental Setup

Data are split into training and test sets using an 80/20 stratified procedure. All preprocessing
steps and model estimation are performed within pipeline structures to prevent information leakage. Performance is evaluated on the held-out test set using accuracy, macro- and weighted precision, recall, F1-score, macro ROC-AUC, and confusion matrices.
Formal Mathematical Specification

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with $x_i \in \mathbb{R}^p$ and $y_i \in \{1,2,3\}$ denote the dataset.

Support Vector Machine. The class-weighted SVM solves

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} w_{y_i}\xi_i$$

subject to margin constraints using the RBF kernel

$$K(x, x') = \exp\left(-\gamma\|x - x'\|^2\right).$$

Multinomial Logistic Regression with Splines. Class probabilities are modeled as

$$P(Y = k \mid x) = \frac{\exp\left(\eta_k(x)\right)}{\sum_{j=1}^{3} \exp\left(\eta_j(x)\right)},$$

where η_k (x) includes cubic spline basis expansions of numeric predictors.
Random Forest. Each tree minimizes Gini impurity

$$G(S) = 1 - \sum_{k=1}^{3} p_k^2$$

and predictions are aggregated by majority vote.
Evaluation Metrics. Macro F1-score is defined as

$$F1_{macro} = \frac{1}{3}\sum_{k=1}^{3} \frac{2TP_k}{2TP_k + FP_k + FN_k}$$

RESULTS AND DISCUSSION

Quantitative Results

The test set contains 3,410 observations, with approximately 13% low, 16% medium, and 71% high satisfaction. The random forest reaches an accuracy close to 0.71, while the support vector machine and spline-based logistic regression achieve accuracies near 0.35. Macro F1-scores remain around 0.30 for all models, indicating limited separation between classes when equal weight is assigned to each satisfaction level.

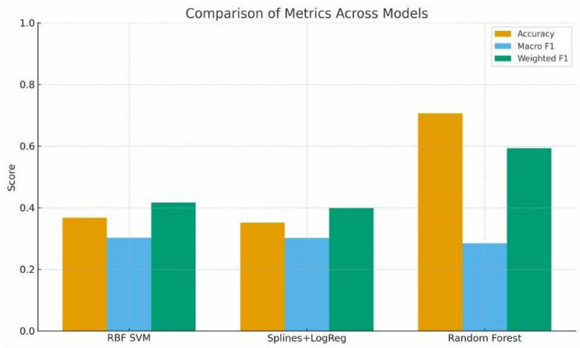

Figure 1: Comparison of accuracy, macro F1, and weighted F1 across models.

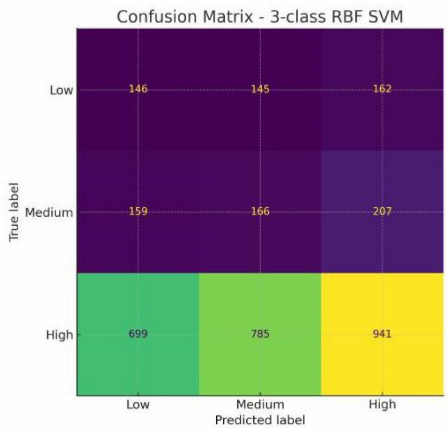

Figure 2: Confusion matrix for the RBF support vector machine.

Qualitative Analysis

Confusion matrices reveal different prediction patterns. The support vector machine and logistic regression distribute predictions across all three classes, with frequent confusion between adjacent satisfaction levels. The random forest assigns most observations to the high satisfaction class.

Discussion

Similar challenges have been reported in business research on customer churn and credit default, where rare outcomes are hard to identify without richer behavioral histories (Verbeke *et al.*, 2014).

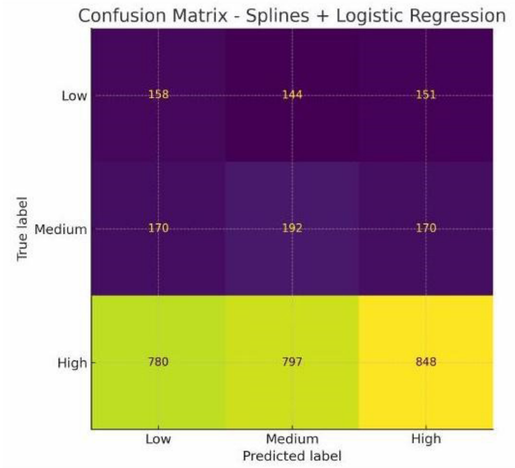

Figure 3: Confusion matrix for the spline-based logistic regression model.

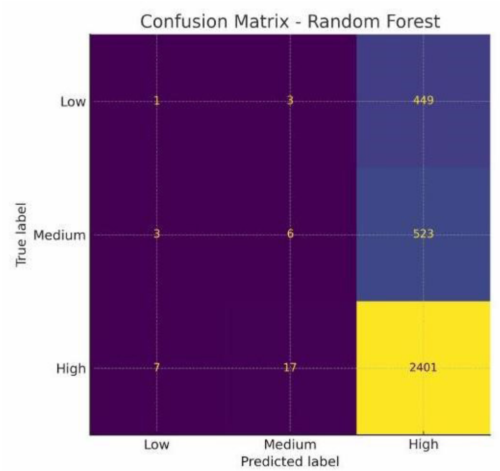

Figure 4: Confusion matrix for the random forest classifier.

CONCLUSION

This paper compared three supervised learning models for predicting customer satisfaction in a highly imbalanced e-Commerce dataset. The random forest achieved the highest overall accuracy, while the support vector machine and spline-based logistic regression showed more even treatment of minority classes. None of the models achieved strong macro-level performance. Future studies may examine alternative target definitions, longer customer histories, and cost-sensitive training approaches, with attention to managerial decision-making contexts.

REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. http://dx.doi.org/10.1023/A:1010933404324

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Kotler, P., & Keller, K. L. (2016). Marketing management (15th ed.). Pearson Education.

Oliver, R. L. (1997). Satisfaction: A behavioral perspective on the consumer. McGraw-Hill.

Rust, R. T., Zeithaml, V. A., & Lemon, K. N. (2000). Driving customer equity. Free Press.

Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. *Applied Soft Computing, 14*, 431–446. https://doi.org/10.1016/j.asoc.2013.09.017