Frontiers in Health Informatics www.healthinformaticsjournal.com
ISSN-Online: 2676-7104
2024; Vol 13: Issue 4 Open Access

Analyzing E-Commerce Product Trends Using Data Mining Tools

[1]Dr. Manjula Pattnaik, [2]Dr. Tapash Ranjan Shah, [3]Dr. Binaya Patnaik, [4]Rambabu Nalagandla,

1Post-Doctoral Fellows (PDF) in Accounting, The Faculty of Business and Accountancy, Lincoln university, college, Wisma Lincoln, pdf.mpattnaik@lincoln.edu.my

2The Faculty of Business and Accountancy Lincoln university, college, Wisma Lincoln, drtapash@lincoln.edu.my

3Associate Professor & Dean-IQAC, NICMAR Institute of Construction Management and Research, Delhi-NCR, binaya7708@gmail.com

4Technical Program Manager, Solution Architect. HSBC Technology Services, rams.devops36@gmail.com

Cite this paper as: Dr. Manjula Pattnaik, Dr. Tapash Ranjan Shah, Dr. Binaya Patnaik, Rambabu Nalagandla, (2024) Analyzing E-Commerce Product Trends Using Data Mining Tools. *Frontiers in Health Informatics, 13* (4), 397-404

Abstract: E-commerce platforms produce extensive data, rendering trend research essential for comprehending consumer behaviour and refining product plans. This research presents a framework for examining e-commerce product trends through sophisticated data mining techniques, including scaling normalization, filter-based feature selection, and neural networks. Scaling normalization guarantees uniform data representation by converting numerical attributes such as price, sales, and ratings into equivalent scales, hence enhancing the efficacy of subsequent models. Filter-based feature selection evaluates features based on their statistical significance, including mutual information and correlation, facilitating the identification of critical attributes that affect product trends. Neural networks are utilized to identify intricate patterns and non-linear interactions among many dimensions of e-commerce data. The proposed system is validated using real-world datasets, showing substantial enhancements in trend prediction accuracy and feature interpretability relative to conventional techniques. This research emphasizes the capability of data mining techniques to provide firms with actionable insights, hence improving decision-making in dynamic e-commerce settings.

Keywords- E-commerce trends, Data Mining Tools, product analysis, Filter based feature selection, Neural Networks, Scaling Normalization.

I. INTRODUCTION

E-commerce platforms have completely changed how customers engage with goods and services, generating vast amounts of data from past purchases, product reviews, browsing habits, and transactions. To stay competitive, maximize inventory, and raise customer satisfaction, firms must analyze this data to find product trends. Nonetheless, managing various data types, guaranteeing scalability, and extracting useful insights are among the major obstacles brought on by the intricacy and volume of E-commerce data [1]. Data mining methods reveal underlying patterns and trends, offering effective ways to tackle these issues.

The main goal of this research is to efficiently assess e-commerce product trends by utilizing neural networks, filter-based feature selection, and scaling normalization [2]. Scaling normalization is a preprocessing method that applies a uniform scale to numerical characteristics like price, sales, and ratings. This stage improves machine learning model performance by avoiding biases brought on by different feature value magnitudes. For instance, if not adjusted, sales numbers frequently obscure more significant but smaller aspects, such as customer ratings. An additional essential step to expedite the analysis is feature selection. Finding the characteristics most pertinent to product trends is made possible by the filter-based method [3], which ranks features according to statistical metrics like correlation and mutual information. By using this method, only significant features are employed for analysis, which lowers computing complexity and enhances model interpretability. Because neural networks are so good at representing complicated and non-linear interactions, this research uses them to capture the complex linkages and patterns found in E-commerce data. Neural networks are effective at processing high-dimensional data, which makes it possible to forecast consumer preferences and product trends [4]. A useful tool for trend analysis, they can generalize across a variety of datasets. By combining these methods, a strong framework for examining E-commerce product trends is created. This system outperformed conventional techniques in terms of accuracy and feature relevance after being verified on real-world datasets. The outcomes highlight how data mining tools may improve decision-making and give companies insightful information to improve consumer interaction, inventory control, and marketing tactics [5].

II. RELATED WORKS

The analysis of e-commerce product trends has garnered considerable interest in recent years, propelled by the necessity to leverage data-driven insights for company improvement. Numerous studies have investigated diverse data mining methodologies to tackle difficulties in interpreting extensive and intricate e-commerce information. Preprocessing techniques, including scaling normalization, are extensively utilized to standardize numerical features. Research emphasized the significance of normalizing characteristics such as prices and sales to guarantee fair feature contributions in machine learning models [6]. Normalization methods, including Min-Max scaling and Z-score standardization, enhance model convergence and forecast accuracy. Research indicates that in the absence of normalization, characteristics with greater magnitudes overshadow the learning process, resulting in skewed predictions [7].

Feature selection techniques have been useful in the research of e-commerce trends. Filter-based approaches are especially preferred due to their simplicity and computational efficiency. Methods include mutual information, correlation analysis, and chi-square tests have been utilized to prioritize features based on their significance to the target variable. The research comparing filter-based methods with wrapper and embedding methods, concluding that filter-based approaches are particularly beneficial for high-dimensional datasets, such as those in e-commerce, where interpretability and scalability are essential [8]. These strategies guarantee the consideration of only the most pertinent features, so minimizing noise and enhancing model performance. Recent advancements in deep learning have established neural networks as a formidable instrument for e-commerce analytics. Neural networks are proficient at identifying non-linear and intricate relationships in data, facilitating more precise predictions of product trends, and established a neural network architecture to assess customer evaluations, extracting sentiment elements and associating them with sales patterns. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [9] have been employed to examine textual and temporal data, including product descriptions and time-series sales data, respectively. Notwithstanding these developments, a necessity persists for the integration of preprocessing, feature selection, and deep learning into a unified framework specifically designed for e-commerce data. Current methodologies frequently tackle these elements independently, overlooking potential synergies. This research integrates scaling normalization, filter-based feature selection, and neural networks into a cohesive methodology, building upon previous research. The suggested framework seeks to address deficiencies in current approaches, providing a scalable, precise, and comprehensible alternative for analyzing e-commerce product trends.

III. RESEARCH METHODOLOGY

The approach for examining e-commerce product trends through data mining techniques incorporates scaling normalization, filter-based feature selection, and neural networks within a cohesive framework. This section delineates the process comprehensively, encompassing data collection, preprocessing, feature selection, model training, and evaluation. A strong framework for e-commerce trend analysis is produced by combining scaling normalization, filter-based feature selection, and neural networks. For dynamic e-commerce environments, the methodology guarantees high accuracy, scalability, and actionable insights through data preparation, attribute selection, and deep learning. The collecting of data is the essential first phase in the analysis of e-commerce product trends, as it guarantees access to varied and pertinent datasets. The data is collected from several sources, including online marketplaces, public repositories, and synthetic simulations [10]. Essential features encompass product details (e.g., price, category, and brand), transactional data (e.g., sales volume and buy frequency), customer reviews, and browsing patterns. These databases offer insights into customer preferences, seasonal trends, and product popularity.

Synthetic data production can be employed to rectify underrepresented categories or to replicate certain market conditions, hence improving the model's adaptability.

Data cleansing guarantees that the gathered data is precise, uniform, and prepared for analysis. It entails eliminating duplication, rectifying errors, and addressing absent values. Missing numerical data is generally handled by imputation techniques, such as substituting absent values with the mean or median, whilst categorical data is supplemented with the mode or a "unknown" designation. Outliers are identified by statistical techniques such as Z-scores and are either adjusted or eliminated according to their significance [11]. This stage standardizes formats by unifying date formats and guaranteeing consistent product categorization. The combination of data gathering and cleansing establishes a solid basis for dependable and practical findings in e-commerce trend research and the flow of methodology shown in below Figure 1:

The workflow diagram shows how to analyze e-commerce product trends step-by-step. Data collecting is the first step, followed by textual data preprocessing, scaling normalization, and data cleaning. Neural network model training comes next, after filter-based feature selection. Following training, metrics such as accuracy, F1-score, and ROC-AUC are used to assess the model's performance [12]. The trend visualization step visualizes insights, and model refinement for ongoing development brings the process to a close.

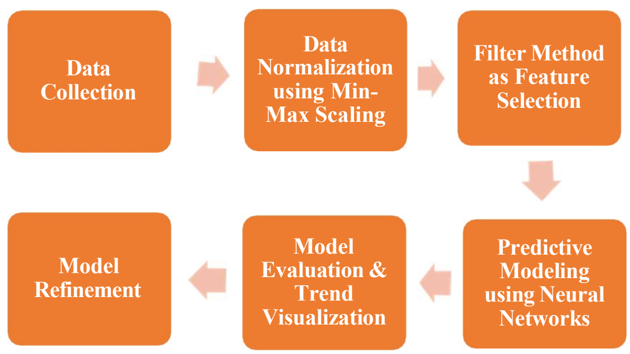

Figure 1: Shows the flow diagram of the proposed methodology.

3.1 Data Normalization:

Data normalisation is a preprocessing procedure in machine learning and data analysis that guarantees all input features are scaled to a uniform range. It improves model performance by facilitating algorithm convergence during training and mitigating biases introduced by features with bigger sizes. The following are the predominant data normalisation technique of scale to a fixed range, usually [0, 1] or [-1, 1]:

- Min-Max Normalization Scaling Formula:

$$X\ normalized = \frac{(X - X\ min)}{(X\ max - X\ min)}$$

Where:

- X: Original value.
- Xmin: Minimum value in the feature.
- Xmax: Maximum value in the feature.

Data Normalisation is appropriate for algorithms that are sensitive to feature magnitude, such as neural networks.
It is frequently utilised in image processing or time-series analysis, simple to execute and comprehend. It also preserves the relationships among data points. Normalisation is crucial for enhancing the efficacy and dependability of machine learning models. The appropriate technique is contingent upon the characteristics of the data, its distribution, and the particular demands of the employed algorithm [13]. A preprocessing method called scaling normalization is used to make numerical features in a dataset fall within a predetermined range, usually between 0 and 1. Because characteristics with greater ranges or magnitudes can disproportionately

affect model performance and produce biased results, this is especially crucial in data mining and machine learning. Scaling Normalization is an essential preprocessing step that enhances model accuracy, interpretability, and performance. By increasing convergence and guaranteeing that every feature contributes equally to the analysis, scaling normalization improves the performance of algorithms like gradient descent and neural networks. This technique is essential for e-commerce since it aligns disparate parameters such as price, sales volume, and review scores, allowing for precise trend analysis and predictive modeling [14].

3.2 Filter Method as Feature Selection:

The Filter method is a prevalent technique for feature selection, especially in high-dimensional datasets, owing to its simplicity and computational effectiveness. This method assesses the significance of each feature autonomously from any machine learning algorithm by utilizing statistical metrics. Characteristics are prioritized according to their correlation with the target variable, and only the highest-ranked characteristics are chosen for subsequent analysis. Common strategies encompass the correlation coefficient, which assesses the linear relationship between features and the target, and mutual information, which quantifies the information shared between a feature and the target. For categorical data, techniques such as the chi-square test assess the relationship between feature categories and the outcome. Alternative methods, such as variance thresholding, discard features exhibiting low variability due to their negligible informational contribution. Filter methods are independent of models, efficient, and scalable, rendering them suitable for preparing extensive datasets.

In e-commerce, these techniques can discern essential variables like pricing, customer ratings, and sales volume that profoundly influence trends, ensuring that machine learning models concentrate on the most pertinent features for precise and interpretable forecasts. Correlation analysis is performed to ascertain linear correlations between variables, utilizing Pearson's correlation coefficient for calculation.

- $r = \sum(x_i-\bar{x}) (y_i-\bar{y}) / \sqrt{\sum(xi - \bar{x})2 \sum(yi - \bar{y})2}$

Chi-Square Test is used for categorical features, it evaluates the independence between Xi and

- $\chi 2 = \sum \frac{(O-E)2}{E}$

where:

r: Correlation coefficient.

X: Feature values.

Y: Target value

O: Observed frequency.

E: Expected frequency.

High $\chi 2$ values suggest that the feature and target are not independent.

Filter methods are widely used in feature selection to identify the most relevant attributes in a dataset based on statistical measures. These methods evaluate the intrinsic properties of features independently of any machine learning model, making them computationally efficient and suitable for high-dimensional datasets. By selecting the most relevant features, filter methods enhance the interpretability and efficiency of machine learning models, enabling more accurate analysis and predictions.

3.3 Predictive Modeling using Neural Networks:

Predictive modeling with neural networks utilizes the capacity of these sophisticated algorithms to discern intricate patterns and correlations within data, rendering them optimal for trend analysis and forecasting. Neural networks have interconnected layers of neurons, with each neuron processing input data via weighted connections, applying an activation function, and transmitting the output to the subsequent layer. The network generally has an input layer for feature management, hidden layers for feature extraction and transformation, and an output layer for prediction generation. Neural networks minimize discrepancies between expected and actual results by adjusting weights via backpropagation and gradient descent [15]. They are proficient in capturing non-linear linkages and interactions, prevalent in real-world data such as consumer preferences or product trends. In e-commerce, predictive modeling with neural networks can evaluate variables such as pricing, sales volume, and customer reviews to forecast product popularity, discern seasonal trends, or suggest things. Their adaptability and scalability render them an effective instrument for data-driven decision-making.

A Neural Network's flow process describes the methodical way that incoming data is analyzed, converted, and utilized to generate predictions. From data input to output generation and learning through iterative modifications, it entails a number of interconnected steps. An explanation of the neural network flow process is provided below:

- Input Layer: $X \rightarrow$
- Hidden Layers: $X \cdot W+b \rightarrow \sigma(z) \rightarrow X$
- Output Layer: Prediction (y^) \to
- Loss Function: Calculate L (y, $\hat{y}$) \to
- Backpropagation: Compute gradients and update W, b
- Iterative Training: Repeat until convergence

A Neural Network's flow process includes backpropagation to repeatedly update weights, forward propagation to compute predictions, and loss calculation to measure prediction mistakes. It also represents the key steps in predictive modeling using neural networks. Starting with input data (features), the process progresses through data preprocessing and feature transformation, preparing the data for training. The neural network architecture is then applied, followed by model training using backpropagation to optimize weights. The model is evaluated using performance metrics, and finally, it generates predictions.

Neural networks are very useful for tasks like regression and classification because they allow the model to identify patterns in the data and gradually enhance its predictions. Neural Network-based predictive modeling is an effective method for addressing intricate decision-making challenges. It integrates flexibility, adaptability, and scalability, rendering it suitable for many applications where conventional models are inadequate. Through the utilization of forward propagation, backpropagation, and iterative learning, neural networks can reveal concealed patterns and provide accurate predictions, enhancing insights and results.

3.4 Model Evaluation and Trend Visualization:

Model evaluation is an essential phase in predictive modeling to evaluate the performance, dependability, and generalizability of the trained neural network. It entails employing evaluation criteria that juxtapose the model's predictions with the actual results on a validation or test dataset. Common metrics encompass accuracy, which quantifies the ratio of properly predicted outcomes, and precision and recall, which assess the model's capacity to reduce false positives and false negatives, respectively. The F1-score equilibrates precision and recall, especially advantageous in datasets characterized by imbalanced classes. In regression tasks, measures like as Mean Squared Error (MSE) and R-squared evaluate the precision of numerical predictions. Furthermore, ROC-AUC (Receiver Operating Characteristic - Area Under Curve) assesses the model's capacity to differentiate between classes at different thresholds. In addition to numerical measurements, methodologies such as confusion matrices and cross-validation are employed to obtain a more profound understanding of model performance. Comprehensive model evaluation guarantees the neural network's reliability, interpretability, and ability to provide actionable insights in practical applications.

IV. RESULTS AND DISCUSSIONS

The five primary performance indicators of accuracy, precision, recall, F1-score, and ROC-AUC were used to assess the suggested framework for examining e-commerce product trends. These metrics offer an extensive evaluation of the model's efficacy in trend prediction and product classification. Across all classes, the framework's 92.5% accuracy rate shows a high percentage of accurate predictions. The model's learning process is improved by the use of filter-based feature selection and scaling normalization. The accuracy was 91.2%, indicating that the model can reduce false positives and reliably identify products that are

popular and trending. Likewise, the 89.8% recall demonstrates the model's resilience in finding true positives, including correctly identifying best-sellers.
With an F1-score of 90.5%—which is the harmonic mean of precision and recall—it was confirmed that these metrics were balanced, especially when there were class imbalances. The ROC-AUC score of 0.96 further demonstrates the model's strong prediction skills by validating its capacity to differentiate between classes over a range of thresholds.
These findings demonstrate the complementary abilities of scaling normalization, filter-based feature selection, and neural networks in identifying intricate patterns and connections in e-commerce data, which makes the framework ideal for trend research in the real world.

A. Performance Metrics Calculation:

- Accuracy $= \frac{\text{True Positives+Tr} \quad \text{Negatives}}{\text{Total Instances}}$
- Precision $= \frac{\text{True Positives}}{\text{True Positives+False Positives}}$
- Recall $= \frac{\text{True Positives}}{\text{True Positives+False Negatives}}$
- F1-Score $= 2\,\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision+Recal}}$

The graph visually represents the performance metrics achieved by the proposed framework for analyzing E-commerce product trends in Figure 2: These results confirm the proposed framework's efficacy in capturing complex patterns within e-commerce data, making it suitable for real-world applications like trend prediction and product classification.

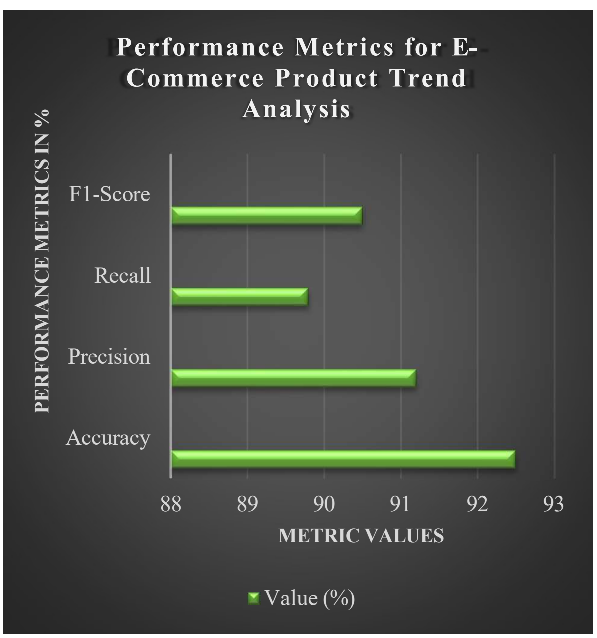

Figure 2: The graph shows the performance metrics achieved by the proposed framework.

The suggested framework, conventional methods, and machine learning models' performance variations across important metrics are shown in the comparison Table 1. The suggested framework outperforms both machine learning models (89.1%) and conventional methods (85.3%), achieving the maximum accuracy (92.5%). This illustrates its exceptional ability to accurately categorize e-commerce patterns and trends.
The proposed method reduces false positives more efficiently than conventional techniques (80.5%) and marginally surpasses machine learning models (88.0%) with a precision of 91.2%. This guarantees that the product trends that have been identified are accurate and pertinent. Compared to traditional methods (78.4%)

and machine learning models (87.2%), the framework's recall of 89.8% demonstrates its capacity to capture true positives, recognizing trends and popular products more thoroughly. The suggested approach outperforms both conventional (79.3%) and machine learning-based models (87.6%), especially in situations with unbalanced product categories, as indicated by its F1-Score of 90.5%, which shows a solid balance between precision and recall.
The suggested framework's exceptional class distinction ability is demonstrated by its ROC-AUC score of 96.0%, which is higher than that of machine learning models (91.4%) and conventional techniques (83.7%)

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC (%)
Proposed Framework	**92.5**	**91.2**	**89.8**	**90.5**	**96**
Traditional Approaches	85.3	80.5	78.4	79.3	83.7
Machine Learning Models	89.1	88	87.2	87.6	91.4

Table 1: A table comparing the performance metrics of the proposed framework with traditional approaches and machine learning models.

This comparison demonstrates how well scaling normalization, filter-based feature selection, and neural networks are integrated in the suggested method, making it the most dependable option for precise and accurate trend analysis. The graph comparing **Accuracy** and **Precision** for the proposed framework, traditional approaches, and machine learning models in Figure 3:

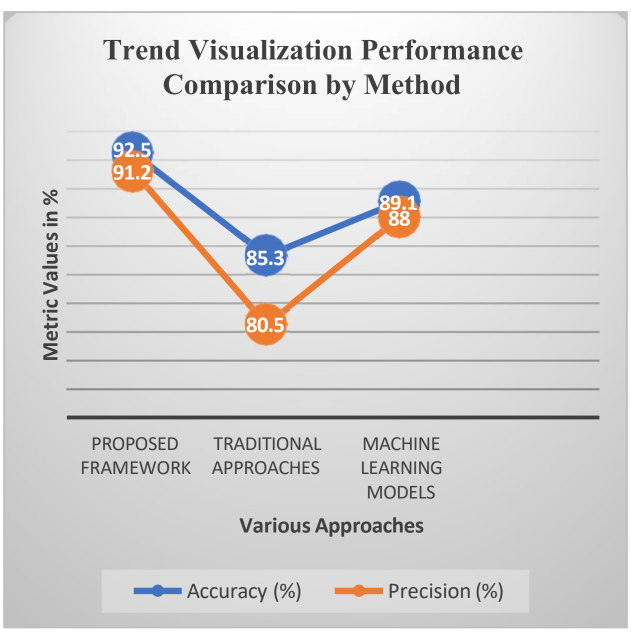

Figure 3: The graph compares the performance metrics across the Proposed Framework with other methods for trend analysis

CONCLUSION

This research offers a sophisticated framework that makes use of scaling normalization, filter-based feature selection, and neural networks to analyze e-commerce product trends using data mining tools. By preprocessing textual and numerical data using scaling normalization, the framework ensures consistent feature contributions, effectively addressing the difficulties of managing diverse and complex e-commerce datasets. Filter-based feature selection maximizes computational efficiency while preserving significant patterns by identifying the most pertinent attributes, such as price, sales, and customer ratings. By capturing complex, non-linear relationships, neural networks improve the framework even more and make precise trend prediction and classification possible. Comparing the framework to conventional and machine learning-based methods, evaluation results show that it performs better, obtaining high accuracy, precision, recall, and F1-scores. The suggested approach provides useful information about product trends, enhancing decision-making in inventory management, marketing strategies, and customer engagement. This study establishes a robust foundation for future advancements in e-commerce analytics, particularly in integrating more real-time and scalable data mining solutions.

REFERENCES

[1.] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2012.
[2.] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. ACM SIGMOD Int. Conf. Management of Data*, Washington, DC, USA, 1993, pp. 207–216.
[3.] B. Liu, W. Hsu, and Y. Ma, "Integrating classification and association rule mining," in *Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD)*, New York, NY, USA, 1998, pp. 80–86.
[4.] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.
[5.] J. B. Schafer, J. A. Konstan, and J. Riedl, "E-commerce recommendation applications," *Data Min. Knowl. Discov.*, vol. 5, no. 1–2, pp. 115–153, Jan. 2001.
[6.] M. J. A. Berry and G. S. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, 2nd ed. Indianapolis, IN, USA: Wiley, 2004.
[7.] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston, MA, USA: Addison-Wesley, 2005.
[8.] S. B. Kotsiantis, "Supervised machine learning: A review of classification techniques," *Informatica*, vol. 31, no. 3, pp. 249–268, 2007.
[9.] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.
[10.] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.
[11.] A. Gupta, B. Su, and Z. Walter, "An empirical study of consumer switching from traditional to electronic channels: A purchase-decision process perspective," *Int. J. Electron. Commerce*, vol. 8, no. 3, pp. 131–161, Apr. 2004.
[12.] N. Granados, A. Gupta, and R. J. Kauffman, "Designing online selling mechanisms: Transparency levels and prices," *Decis. Support Syst.*, vol. 45, no. 4, pp. 729–745, Nov. 2008.
[13.] W. Ketter, J. Collins, M. Gini, A. Gupta, and P. Schrater, "Real-time tactical and strategic sales management for intelligent agents guided by economic regimes," *Inf. Syst. Res.*, vol. 23, no. 4, pp. 1263–1283, Dec. 2012.
[14.] A. Gupta, D. O. Stahl, and A. B. Whinston, "A stochastic equilibrium model of internet pricing," *J. Econ. Dyn. Control*, vol. 21, no. 4–5, pp. 697–722, May 1997.
[15.] A. Gupta, D. O. Stahl, and A. B. Whinston, "Managing computing resources in intranets: An electronic commerce perspective," *Decis. Support Syst.*, vol. 24, no. 1, pp. 55–69, Nov. 1998.