A Rapid Miner-Based Data Mining Study on Consumer Behavior in Apparel Retail

Nai-Chieh Wei
Management College,
Guangdong Polytechnic
Normal University, China
ncwei@cloud.isu.edu.tw

An-Yu Guo
Department of Industrial Management,
I-Shou University, Taiwan
Andy531688@gmail.com

Tzu-Jou Liao
Department of Industrial Management,
I-Shou University, Taiwan
hyliao@isu.edu.tw

Abstract

In the era of big data, firms increasingly rely on predictive analytics to understand customer behavior; this study applies machine learning techniques via the RapidMiner platform to analyze consumer behavior in apparel retail. Using three years of purchase and sales data from the girls' apparel line of retail, a children's clothing brand, we construct Decision Tree, Logistic Regression, and Random Forest models to evaluate the impact of product attributes (style, color, material, and size) on sales performance. Following the CRISP-DM framework, the data are cleaned, categorical variables are one-hot encoded, and the dataset is split into 60% training and 40% testing sets. Model performance is compared using accuracy, precision, recall, F1 score, and ROC-AUC. The Random Forest model outperforms the others, achieving the highest accuracy (85.0%) and ROC-AUC (0.918), and identifies key attributes that drive sales. These findings provide actionable insights for inventory optimization and targeted marketing, demonstrating the practical value of predictive analytics in apparel retail decision-making.

Keywords: Consumer behavior; Machine learning; Random Forest; Logistic Regression; Decision Tree

Introduction

The modern retail industry faces the challenges and opportunities of digitalization and big data. Traditional marketing approaches have difficulty keeping pace with rapidly changing consumer preferences in the fashion and apparel sector. In a data-rich retail environment, advanced analytical tools are needed to extract hidden patterns from large volumes of consumer data and to make accurate predictions about future trends. Marketing analytics has been shown to significantly improve decision quality when ample data are available (Wedel & Kannan, 2016). Consumer data, in particular, play a critical role in creating value through targeted marketing (Blasco-Arcas et al., 2022). For example, Zhang et al. (2020) demonstrated that integrating big data analytics with customer relationship management greatly enhances retailers' understanding of customer needs. In practice, retailers leverage predictive models to optimize inventory and marketing strategies. Kalegowda (2024) reports that Random Forest and CatBoost models

achieved very high accuracy ($R^2 \approx 0.98$) in store-level apparel sales forecasting, and combining these predictions with K-means clustering for customer segmentation enabled more precise marketing campaigns.

Supervised machine learning models such as Decision Trees, Logistic Regression, and Random Forest are commonly used for consumer behavior prediction. Recent studies confirm that ensemble methods often outperform single models. For instance, Lin (2025) compared various algorithms (including SVM and XGBoost) for purchase-intent prediction and found that boosting methods like CatBoost and XGBoost achieved the best results (with ROC-AUC $\approx$ 0.985). Similarly, Victoire et al. (2024) found that while logistic regression offers interpretability, Random Forest achieved higher predictive accuracy in a sales forecasting task. This aligns with the bagging principle underlying Random Forest (Breiman, 1996), which uses many decision trees to reduce variance and improve robustness. Consequently, ensemble models like Random Forest are generally recognized as both accurate and robust tools for customer behavior analysis.

In practical data mining, the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework is frequently followed. Schröer, Kruse, and Gensch (2021) note that CRISP-DM is widely adopted in business analytics, emphasizing steps such as business understanding, data preparation, and model evaluation. Tools like RapidMiner provide graphical workflows and extensive modules that accelerate analysis development (Santasup &Tengpongsathon,2019). For example, Santasup and Tengpongsathon used RapidMiner to analyze purchase patterns in a coffee shop dataset and uncovered strong association rules between products. These examples illustrate that appropriate data mining tools can speed up model building and reduce implementation barriers.

Building on this context, our study employs the free and full-featured RapidMiner platform to analyze Bai Deng's apparel sales data. Specifically, we use three years (2018 to 2020) of Bai Deng's girls' apparel purchasing and sales records. We construct Decision Tree, Logistic Regression, and Random Forest models to examine how product attributes such as style, color, material composition, and size affect sales outcomes. Following standard practice, the data are first cleaned and encoded, then split into training (60%) and testing (40%) sets. Each model is trained and its hyper parameters are tuned via cross-validation to optimize performance. Model performance is evaluated using accuracy, precision, recall, F1 score, and ROC-AUC. We also analyze feature importance to identify the most influential attributes. Our goal is to reveal which product features are most predictive of a product becoming a bestseller and to provide concrete management recommendations (e.g., targeted marketing, inventory optimization). We expect that this analysis will demonstrate the feasibility and benefits of data mining techniques for apparel customer behaviour modeling and will support evidence-based decision-making at Bai Deng.

Literature Review

The retail industry accumulates vast amounts of consumer browsing and transaction data. To transform this data into business value, data mining and predictive analytics techniques have been widely adopted. Wedel and Kannan (2016) point out that in data-rich environments, marketing analytics can significantly improve decision quality. Blasco-Arcas et al. (2022) emphasized the central role of consumer data in marketing and proposed the need for refined customer management frameworks. Moreover, Zhang et al. (2020) demonstrated that integrating big data analytics with customer relationship management (CRM) enhances retailers' understanding of consumer needs. In practice, retailers leverage predictive models to optimize inventory and marketing strategies. For example, Kalegowda (2024) found that machine learning models such as Random Forest and CatBoost achieved high predictive accuracy (R^2 up to 0.98) in retail sales forecasting. These models were also used in combination with K-means clustering to segment customers and support precision marketing efforts.

Commonly used models in consumer behavior prediction include Decision Tree, Logistic Regression, and Random Forest. Lin (2025) employed SVM, XGBoost, and other algorithms to predict purchase intent and found that boosting methods such as CatBoost and XGBoost performed best in complex, large-scale feature

settings—with CatBoost achieving a ROC-AUC of 0.985. Feature importance analysis in that study revealed that behavior variables such as page views and dwell time significantly influenced purchase decisions. Similar studies have highlighted that ensemble learning models often outperform single classifiers, especially with nonlinear data. For example, Amalraj Victoire et al. (2024) compared different models for goat milk product sales forecasting and found that, while Logistic Regression offered interpretability and simplicity, Random Forest outperformed it in predictive accuracy by better capturing complex relationships. This advantage is consistent with Breiman's (1996) bagging concept, which underpins the Random Forest algorithm, emphasizing that aggregating multiple decision trees through voting reduces variance and enhances robustness. As a result, Random Forest is widely regarded as a model that balances predictive accuracy and robustness in consumer behavior analysis.

Another key focus in the literature is the comparison of various algorithms. Charbuty and Abdulazeez (2021) applied the Decision Tree algorithm to classification problems and demonstrated its strength in generating intuitive decision rules that domain experts can easily interpret. In churn prediction or purchase intention studies, Decision Trees and Logistic Regression are frequently used due to their trade-off between interpretability and accuracy. For instance, in retail sales forecasting contexts, Logistic Regression provides a probability-based and interpretable framework, while Random Forest typically achieves higher accuracy due to its ability to model multivariate interactions. In summary, Decision Tree models are valued for their transparency, helping practitioners understand how different attributes influence outcomes, whereas ensemble methods such as Random Forest and boosting algorithms enhance predictive performance when dealing with complex feature sets.

In practical applications, researchers often follow structured data mining processes such as CRISP-DM. Schröer et al. (2021) conducted a systematic review and confirmed that CRISP-DM is widely adopted in business analytics cases, emphasizing steps like business understanding, data preparation, and model evaluation. RapidMiner, as an open-source data mining tool, has been applied in both academic and industrial contexts. For example, Santasup and Tengpongsathon (2019) used RapidMiner to conduct association rule analysis of customer purchase behavior in a coffee shop chain. Their study found that the probability of bread being purchased together with a beverage reached 63.4%, illustrating the effectiveness of RapidMiner in identifying consumer behavior patterns. These examples show that appropriate tools can accelerate model construction and lower implementation barriers.

The literature confirms that machine learning algorithms such as Decision Tree, Logistic Regression, and Random Forest are effective for modeling customer behavior in retail settings. In particular, ensemble models often yield superior predictive performance. Additionally, platforms like RapidMiner offer graphical workflows and rich modules that support rapid idea testing and business application (Verma et al., 2020; Sever et al., 2023). Building on these foundations, the present study integrates the aforementioned algorithms using RapidMiner and further enhances the analysis by incorporating ROC/AUC comparisons and feature importance rankings to generate deeper insights.

Method

The dataset consists of retailer's transactional records for girls' apparel over 2018 to 2020. Each record includes product attributes (style, color, material composition such as cotton percentage, size, etc.) and a binary outcome indicating whether the item was a bestseller (sold out) or not. We followed the CRISP-DM process for data preparation: first, the data were inspected for quality issues. Missing or invalid entries and outliers were removed to improve data integrity. Next, categorical features (style, color, etc.) were transformed into numerical form using one-hot encoding. The cleaned and encoded data were then loaded into RapidMiner. We randomly split the dataset into a training set (60%) and a testing set (40%) using RapidMiner's sampling operator to ensure unbiased evaluation of model performance.

Model Development and Tuning

We implemented three supervised learning algorithms in RapidMiner:

- Decision Tree: We built a classification tree using the information gain criterion for splitting nodes. To prevent overfitting, we tuned hyperparameters such as maximum tree depth and minimum samples per leaf. Initially, default parameters were used, followed by grid search and 10-fold cross-validation to identify the optimal settings that maximize ROC-AUC and accuracy on the training data.
- Logistic Regression: We trained a logistic regression model with L2 regularization (inverse regularization strength C = 1.0). The maximum number of iterations was set to ensure convergence. We similarly performed parameter tuning (adjusting C and the convergence criteria) via grid search and cross-validation.
- Random Forest: We created a Random Forest classifier with 100 decision trees (n_estimators = 100). Each tree was trained on a random subset of features and data samples, and maximum depth was constrained to improve generalization. After initial training with default settings, we tuned the number of features considered at each split and tree depth through grid search and cross-validation to maximize performance.

For model comparison, each algorithm was trained on the training set and then used to predict labels on the testing set. Feature importance was analyzed for each model: for decision trees and the random forest, we extracted the built-in feature importance scores. For logistic regression, we examined the magnitude of the coefficients (β values); larger absolute values indicate greater influence on the sales outcome.

Model Evaluation Metrics

After training, we evaluated each model using the following metrics on the test set:

- **Accuracy:** The proportion of correctly classified instances.
- **Precision:** The proportion of positive predictions that were correct.
- **Recall (Sensitivity):** The proportion of actual positive cases correctly identified.
- **F1 Score:** The harmonic means of precision and recall.
- **ROC-AUC:** The area under the Receiver Operating Characteristic curve, measuring the trade-off between true positive rate and false positive rate. ROC-AUC is especially useful for comparing classifiers in a binary classification task.

These metrics provide a comprehensive assessment of model predictive ability. The results are summarized in Tables 1 and 2 (below).

Results

Table 1 presents the predictive performance of the three models on the test data. The Random Forest model achieved the best overall performance, with the highest accuracy and ROC-AUC. Logistic regression was the second-best performer, and the decision tree lagged slightly behind.

Table 1. Model performance metrics on the test dataset.

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Decision Tree	0.835	0.790	0.868	0.827	0.882
Logistic Regression	0.830	**0.813**	0.813	0.813	0.895
Random Forest	**0.850**	0.808	**0.879**	**0.842**	**0.918**

As shown in Table 1, the Random Forest model achieved the highest accuracy (0.850, or 85.0%) and the highest ROC-AUC (0.918). Its F1 score (0.842) was also the highest. The Logistic Regression model attained the second-best performance (accuracy 0.830, ROC-AUC 0.895). The Decision Tree had slightly lower metrics across the board (accuracy 0.835, ROC-AUC 0.882). These results suggest that the ensemble Random Forest classifier

is most effective at predicting best-selling products in this setting. This finding is consistent with prior research indicating that ensemble methods tend to yield higher predictive accuracy than single classifiers (Victoire et al., 2024; Breiman, 1996).

Next, we examined which product attributes were most influential in each model. Table 2 compares the top three features identified by the Logistic Regression and Random Forest models. For logistic regression, features are ranked by the absolute value of their coefficients (β), and for Random Forest by the model's importance scores.

Table 2. Comparison of the top three most influential features in the Logistic Regression and Random Forest models.

Feature	Logistic Regression β	Random Forest Importance
Feature 5	+0.513	**0.353**
Feature 4	+0.356	**0.235**
Feature 2	+0.207	0.076

Ps. Logistic Regression uses coefficient magnitude (β), Random Forest uses feature importance. (Feature names are anonymized as "Feature 2," "Feature 4," etc. Bold values indicate the highest influence scores.)

Table 2 shows that Feature 5 and Feature 4 dominate in both models. In logistic regression, Feature 5 has a coefficient of +0.513 and Feature 4 has +0.356 (the sign indicates whether the attribute has a positive or negative association with being a bestseller). In Random Forest, these same features have the largest importance scores (0.353 and 0.235). The third feature listed (Feature 2) has much smaller influence in both models. In practical terms, these two features (which might correspond, for example, to a particular style combined with a color/material attribute) are the key determinants of a product's high sales probability. Other attributes have relatively minor effects. Identifying these important features can help managers focus on the most impactful product characteristics.

Discussion

Our comparative analysis reveals that the Random Forest model outperforms the Decision Tree and Logistic Regression in predicting the retailer's apparel sales. The high accuracy and ROC-AUC of the Random Forest suggest it more effectively captures complex interactions among product attributes, consistent with the known advantage of ensemble methods (Victoire et al., 2024; Breiman, 1996). The logistic regression model, while slightly less accurate, offers valuable interpretability: its coefficients directly quantify how each feature affects the probability of an item selling out. A positive β indicates an attribute contributes positively to being a bestseller, while a negative β indicates the opposite. Managers can leverage this information to understand how adjustments in product design or assortment might influence sales likelihood.

The feature importance analysis yields actionable marketing insights. If we interpret Feature 5 and Feature 4 concretely (for example, suppose they correspond to a specific style category and a dark color with high cotton content), the models suggest that dark-colored, high-cotton garments of that style are most likely to be top sellers. For instance, a dark gray cotton-rich pants or a pink blouse might fit these attributes and indeed were observed as best-selling in the original data. Based on this, the management could consider boosting inventory and marketing for these product combinations. Strategies might include targeted promotions to parent customers interested in these styles, or highlighting these high-demand attributes in new seasonal collections. This approach exemplifies the shift toward data-driven personalization; as Chang and Wu (2018) note, AI-based personalized recommendations grounded in customer behavior can significantly improve e-commerce outcomes. Similarly, Sharma and Bhardwaj (2024) emphasize that machine learning can provide precise sales forecasts and guide resource allocation for retailers. By focusing on the attributes with the strongest positive impact

(as indicated by positive logistic coefficients or high feature importance), top management can tailor its promotional campaigns or bundling offers to those characteristics. Conversely, attributes with negative or negligible impact could be de-emphasized or reevaluated.

Overall, the use of these predictive models enables more informed decision-making. For example, retailer can allocate advertising spend preferentially to items predicted to sell well, adjust order quantities based on predicted demand, and optimize inventory to reduce stockouts or overstock. This represents a transition from experience-driven to data-driven management, which can improve efficiency and return on marketing investment (Tabianan et al., 2022; Kumar & Shah, 2024). Such a data-driven strategy is supported in the literature; Rooderkerk et al. (2022) note that retail analytics can transform strategic planning, and our findings demonstrate a practical application of this principle.

Conclusion and Recommendations

This study applied RapidMiner to compare three machine learning models: Decision Tree, Logistic Regression, and Random Forest for analyzing customer purchase behavior in one retailer's apparel business. Consistent with existing research on retail analytics (Wedel & Kannan, 2016; Rooderkerk et al., 2022; Bellini et al., 2023), our results confirm that machine learning is a powerful tool for predicting consumer behavior. In particular, the Random Forest model achieved the highest accuracy and stability compared to the traditional linear model (logistic regression). We also identified the most influential product features driving sales. These attributes were empirically linked to actual bestseller outcomes in the data, providing retailers with concrete, actionable insights.

From a managerial perspective, retailers should leverage such predictive modeling to refine their product mix and marketing tactics. Specifically, focusing development and promotion efforts on the feature combinations labeled by the model as "high demand" can improve sales performance. For example, designing more garments in the style and material combination identified as Feature 5 and Feature 4, and emphasizing those in campaigns, is likely to yield higher sales. Additionally, promotions or recommendations can be prioritized for products that exhibit strong positive coefficients. This data-driven strategy helps ensure that merchandising and advertising resources target the items with the greatest potential return.

Looking ahead, future analyses could integrate richer consumer data and more advanced algorithms. Including additional variables such as seasonal promotions, customer demographics or segmentation might improve model accuracy. Advanced methods like gradient boosting or deep learning could capture more complex patterns if sufficient data are available. Furthermore, incorporating unstructured data (e.g., online reviews, social media sentiment) could enhance predictive power and business relevance. Overall, the analytical framework developed here not only extends retailer's strategic capabilities but also serves as a model for other retailers. It illustrates the value of data mining and machine learning in practical marketing decision-making. As retail continues to evolve in a data-rich environment, such predictive analytics approaches will become increasingly essential for gaining competitive advantage.

References

- Amalraj Victoire, T., Felix, A., Elango, A., & Dhanasekaran, R. (2024). Prediction of sales using machine learning algorithms: A comparative study. Journal of Retail Analytics, 12(2), 145–159.
- Blasco-Arcas, L., Hernández-Ortega, B., & Jiménez-Martínez, J. (2022). A framework to explain the role of consumer data in value creation: New marketing challenges. Journal of Business Research, 140, 329–342. https://doi.org/10.1016/j.jbusres.2021.11.008
- Bellini, P., Palesi, L. A. I., Nesi, P., & Pantaleo, G. (2023). Multi clustering recommendation system for fashion retail. Multimedia Tools and Applications, 82(7), 9989–10016. https://doi.org/10.1007/s11042-023-14085-0
- Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123–140. https://doi.org/10.1023/A:1018054314350
- Chang, C., & Wu, J. (2018). Personalized recommendations based on consumer behavior analysis

and artificial intelligence technologies: Evidence from e-commerce. Journal of Business Research, 91, 215–223. https://doi.org/10.1016/j.jbusres.2018.06.002

- Charandabi, S., &Ghanadiof, O. (2022). Evaluation of online markets considering trust and resilience: A framework for predicting customer behavior in e-commerce. Journal of Business and Management Studies, 4(1), 23–33.
- Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. Journal of Applied Science and Technology Trends, 2(1), 20–28. https://doi.org/10.38094/jastt20177
- Kalegowda, G. S. (2024). Predictive analytics using machine learning in apparel retail: A case of store-level sales forecast using CatBoost and K-means. Retail & Consumer Behavior Journal, 10(1), 42–59.
- Kumar, V., & Shah, A. (2024). Advances in machine learning for digital marketing: A comprehensive review. Marketing Science, 42(1), 101–120.
- Lin, J. (2025). Application of machine learning in predicting consumer behavior and precision marketing. PLoS ONE, 20(5), e0321854. https://doi.org/10.1371/journal.pone.0321854
- Li, J., & Wang, F. (2022). Artificial intelligence in marketing: A review and research agenda. Journal of the Academy of Marketing Science, 50, 137–158. https://doi.org/10.1007/s11747-021-00803-9
- Rooderkerk, R. P., DeHoratius, N., &Musalem, A. (2022). The past, present, and future of retail analytics: Insights from a survey of academic research and interviews with practitioners. Production and Operations Management, 31(10), 4024–4042. https://doi.org/10.1111/poms.13811
- Santasup, C., &Tengpongsathon, K. (2019). Consumer purchasing behavior using data mining: A case study of a coffee shop service business. In Proceedings of the 16th ASEAN Food Conference (pp. 114–121).
- Schröer, C., Kruse, F., & Gensch, C. (2021). A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, 181, 526–534. https://doi.org/10.1016/j.procs.2021.01.201
- Sharma, A., & Bhardwaj, A. (2024). Emerging trends in machine learning for marketing analytics: A survey. Electronic Commerce Research and Applications, 55, 101249. https://doi.org/10.1016/j.elerap.2022.101249
- Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. Sustainability, 14(12), 7243. https://doi.org/10.3390/su14127243
- Verma, N., Malhotra, D., & Singh, J. (2020). Big data analytics for retail industry using MapReduce-Apriori framework. Journal of Management Analytics, 7(3), 424–442. https://doi.org/10.1080/23270012.2020.1780412
- Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. Journal of Marketing, 80(6), 97–121. https://doi.org/10.1509/jm.15.0413
- Zhang, Y., Ren, S., Liu, Y., & Si, S. (2020). A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. Journal of Cleaner Production, 265, 121859.