Journal of Information Technology and Digital World (ISSN: 2582-418X)
www.irojournals.com/itdw/

Predicting Consumer Shifts to Sustainable Products using Machine Learning Models

Zaobiya Khan[1], Neha Vora[2]

[1]PG student, [2]Research Scholar, Department of Information Technology, SVKM's Usha Pravin Gandhi College of Arts, Science and Commerce, Mumbai, India

E-mail: [1] zaobiyakhan3@gmail.com, [2]nehavora2501@gmail.com

Abstract

This study evaluates the extent to which consumers are willing to switch to sustainable products and identifies the strategic measures to establish sustainable brands. Consumer preferences and behaviours are analysed to identify the early adopters of sustainable products The study employs several machine learning algorithms including Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN), to determine the likelihood of consumers switching towards sustainable choices. SMOTE (Synthetic Minority Over-sampling Technique) was applied to address the class imbalance in the data. The models were evaluated using metrics such as accuracy, precision, recall, and F1 score. The results indicated that the Random Forest and SVM outperformed the other models in predicting consumer willingness to adopt sustainable products. This study demonstrates the potential of machine learning techniques in understanding customer behaviour, thereby supporting marketers in promoting sustainable brands.

Keywords: Sustainable Products, Consumer Behaviour, Logistic Regression, SVM, Gradient Boosting, Random Forest, KNN, SMOTE.

1. Introduction

The increasing awareness about sustainability issues has changed the approach of both consumers and companies toward products and services. Over the past decade, the number of consumers inclined to choose sustainable products that align with their eco-friendly beliefs has

DOI: https://doi.org/10.36548/jitdw.2024.3.008
Received: 12.08.2024. received in revised form: 10.09.2024, accepted: 21.09.2024, published: 14.10.2024

increased. This, in turn, has prompted brands to adopt green initiatives in their business practices and product promotions. However, a significant gap still exists between the original intentions of consumers and their actual purchase behavior, regardless of rapid changes in consumer consciousness towards the environment. Though many people have the intention of purchasing goods from these sustainability-based brands, factors, such as price, availability, and brand information, often hinder their purchasing decision.

The purpose of this study is to explore the factors that make customers switch to sustainable products. This is achieved by employing machine learning models that measure the probability of a customer choosing to move from conventional products to sustainable alternatives. The models employed in this analysis include logistic regression, support vector machines, random forests, gradient boosting, and K-nearest neighbors. Such models allow an understanding of the most important features, trends, and preferences of the different consumer market segments. SMOTE was applied to address the class imbalance in the dataset, which improved the prediction accuracy. Using these machine learning models, the research aims to determine the factors that influence the adoption of such sustainable products. Among all the models, random forest and support vector machines worked more accurately in predicting consumer behavior in sustainability contexts.

2. Related Work

Consumer attitudes and preferences toward buying sustainable products are also very important [1]. The initial scholars in this area used conventional survey approaches to understand consumer attitudes and explain the cause of sustainable purchases. Sustainable production and consumption are an important strategy, as their implementation can help manufacturers achieve overall development plans [2]. However, understanding the consumer attitude toward purchasing and analyzing an accurate outcome is a challenging task.

Consumer satisfaction over a product also plays an important role in shopping [3]. Predicting consumer behavior with certain modeling techniques and machine learning has been employed to understand the pattern more accurately. Predictive modeling, logistic regression, and decision tree models have been used to research how price, brand reputation, and environmental effects influence a consumer's purchasing decision. The purchase patterns of the consumers have been analyzed in the dataset. For instance, logistic regression and decision tree

models have been used to research how price, brand reputation, and environmental impact influence a consumer's buying decision.

Also, KNN is used for customer segmentation and to understand the customer's shopping pattern [4]. Some other works include Big Data in promoting sustainable consumption behavior with the help of bibliometric analysis [5]. Studies using these models have indicated that, although environmental concern plays a role, it is often outweighed by practical considerations such as product cost and convenience.

Advanced techniques such as SVM, gradient boosting, and random forest have been used to understand more clearly the complex relationships between consumer demographics and psychographics about purchasing decisions. Whenever there is a class imbalance, it is addressed with Synthetic Minority Oversampling Technique (SMOTE), as consumers who actively buy sustainable products represent a small minority compared to the general population. These advanced methods have been applied effectively to generate much more precise consumer behavior predictions, highlighting the role of targeted marketing in promoting sustainable products.

Overall, with the help of machine learning, it became easy to predict consumer willingness to switch towards sustainable products, which in turn provides insight that will help different brands make proper decisions in their marketing strategies to match the changing consumer preferences.

3. Proposed Work

3.1 Dataset Collection

Because no public dataset is available for consumer switching behavior to sustainable products and brands, a unique dataset was created for this study. Finally, the dataset was obtained by conducting a short survey with the help of Google Forms, targeting people from different backgrounds. The respondents were asked about their care for sustainability, their current purchasing habits, and whether or not there was a willingness to switch out traditional items for more sustainable options.

The Survey Google Form contained the following questions:

1. Age
2. Gender

3. Location
4. Income level
5. How often do you shop for non-essential products?
6. Where do you usually shop for these products?
7. What factors influence your purchasing decisions the most?
8. How familiar are you with sustainable products?
9. Which aspects of sustainability matter most to you when buying a product?
10. Do you currently prefer buying from brands that promote sustainability?
11. Which sustainable products have you purchased in the last 6 months?
12. What prevents you from buying sustainable products more often?
13. Would you switch to a sustainable brand if it offered products similar in quality and price to your current brand?
14. What type of information would help you make more sustainable purchasing choices?
15. How likely are you to recommend a sustainable brand to others?
16. Any additional comments or suggestions on how brands can better promote sustainability?

All the questions contained options in which the respondent selected their appropriate choices, and a total of 1,500 responses were collected and compiled into a dataset for analysis.

3.2 Data Cleaning and Preprocessing

The Missing values of "Willingness to Switch" or "Awareness of Sustainability" were addressed through imputation. Mean and mode imputation were used for continuous and categorical variables. Rows containing excessive missing values were removed.

Outliers in numerical fields like income were detected using the interquartile range (IQR) method using the boxplot visualization and addressed by capping extreme values to a reasonable range. Values beyond 1.5 times the IQR, that is, income above $100,000, were considered outliers, avoiding undue influence on the analysis. The categorical variables, such as "Gender," were encoded to make the data suitable for machine learning models.

3.3 Data Augmentation

Since the number of respondents who were very willing to switch was largely outnumbered by those who were neutral or less likely, the SMOTE (Synthetic Minority Oversampling Technique) oversampling method was used, which helped in balancing the dataset.

3.4 Feature Extraction

Demographic information like age, gender, income, shopping preferences, sustainability awareness, and product familiarity were selected and considered as the features for this analysis. Further the most important feature of this study, the consumer shift towards sustainability was considered in predicting the likelihood.

3.5 Data Splitting

The dataset was split mainly for training and testing. 70% of the data was used for training the model and getting consumer behavior insights. 30% of the data was used as a test dataset. This splitting strategy ensured a reliable dataset to analyze consumer trends toward sustainability without the need for additional validation set.

3.6 Experimental Setup

The machine learning models were employed and evaluated using Python programming language and its libraries such as pandas for data preprocessing, Seaborn and Matplotlib for visualization and EDA, Scikit-learn for model implementation, SciPy for statistical analysis, and Jupyter Notebook as the development environment. The Flowchart in Figure 1 depicts the steps in the analysis.

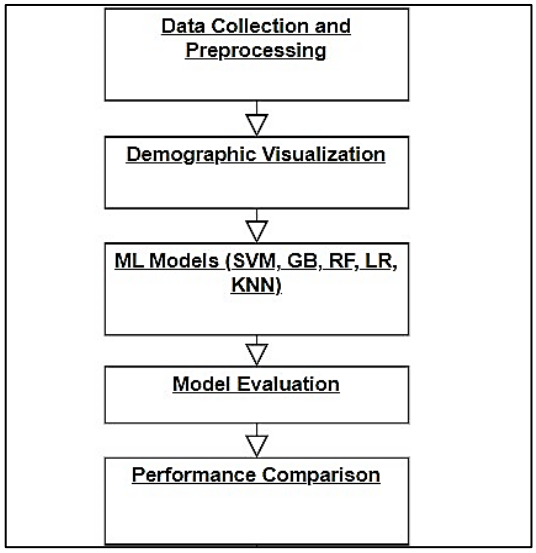

Figure 1. A Flowchart of Steps Included in the Analysis

3.7 Models Used

A total of five machine learning models were trained and tested that is, Support Vector Machine, Logistic Regression, Random Forest, Gradient Boosting, and K-Nearest Neighbours. All the models were trained on a 70% training set, with the remaining 30% used for testing.

- **Logistic Regression**

Logistic Regression is used when the dependent variable is dichotomous and contains decision-making factors [6]. Binary logistic regression is one method that is particularly appropriate for analysing survey data in the widely used cross-sectional and case–control research designs [7]. Logistic regression yielded a basic linear model for consumers' chances of going green. More importantly, it suggested that people who exhibit sustainability behavior are such to a much greater extent because of factors like income, awareness, or previous purchases rather than altruistic commitment.

For the the parameters, max_iter = 2000 was set to ensure convergence during the training phase with the 'lbfgs' solver which specifies the optimization algorithm to be used in the model. It performed adequately with logistic regression but struggled to accurately predict negative cases, i.e., users who would never switch.

- **Support Vector Machine (SVM)**

Support Vector Machines are supervised learning methods and can be utilized for classifications and regressions [8]. It selects from the training samples a set of characteristic subsets so that the classification of the character subset is equivalent to the division of the entire dataset [9]. SVM drew this hyperplane as a decision boundary to separate the two classes (agree/intent switch) and managed non-linear relationships through the Radial Basis Function (RBF) kernel.

Default parameters, such as regularization parameter C=1, were set to handle non-linear relationships in data. The SVM model outperformed the logistic regression regarding accuracy and recall, particularly for consumers willing to switch to sustainable products.

- **Random Forest**

Random Forest is a classification algorithm that can be used for large data sets and has high prediction and accuracy [10]. The Random Forest model is all in one and combines multiple decision tree outputs to improve performance. It can improve predictions for many supervised methods, especially decision trees. These trees are grown in-depth without a pruning phase [11].

Parameters such as random_state = 42 which helps control the randomness involved in generating decision trees, the number of trees (n_estimators = 100) and (max_depth = 10) were set to balance the model complexity. The parameter class_weight = 'balanced' was set to adjust the weights of frequency-based classes in the data to handle the class imbalance. RF yields the

highest precision, accuracy, and recall, which addressed the class imbalance correctly and provided the most reliable results.

- **Gradient Boosting**

Gradient Boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models [12]. Gradient Boosting consecutively fits new models to provide a more accurate estimate of the response variable [13]. Gradient boosting is a type of ensemble model that builds decision trees one after the other to continuously decrease prediction error.

The model was trained with 100 boosting stages (n_estimators = 100) with the learning rate = 0.1 for optimizing the balance between accuracy and overfitting. It performed better than Logistic Regression but did worse than Random Forest Classifier and SVM as far as overall accuracy. Finally, we applied a gradient boosting model, which had the best results with precision and recall of positive cases (ready to switch) compared to other models, but it was the most overfitted as well.

- **K-Nearest Neighbours (KNN)**

KNN requires all the training data instances to be stored, and then, for each unseen case and every training instance, it performs a pairwise computation of a certain distance or similarity measure [14]. A KNN classification algorithm is a query-triggered yet improvisational learning procedure, in which they are carried out only when test data is predicted that sets a suitable K value and searches the K nearest neighbors from the whole training sample space [15]. KNN uses the classification of data points by comparing the nearest neighbors in the feature space.

The model was trained with 5 neighbors (n_neighbors = 5), which helped in determining classification based on proximity to the nearest data points. KNN is a simple and non-parametric model that performed quite well but struggled with the complex data. KNN showed the lowest accuracy and struggled with the imbalanced dataset and hence provided poor predictions on the test set.

Table 1. Dataset with Pre-processing and Without Pre-processing

Aspect	Before Preprocessing	After Preprocessing
Missing Values	0	0(after dropping)
Duplicates	1516	0
Columns	16	15
Data Rows	975	975
Outliers in Income	No filtering applied	Income values filtered < 100000
Gender Unique Values	Male, Female	male, female
Location Unique Values	Mumbai, Chennai, Delhi, Hyderabad, Ahmedabad, Pune	mumbai, chennai, delhi, hyderabad, ahmedabad, pune
Frequency of Non-Essential Products	Varies	Cleaned data, maintained frequencies
Frequency of Preferred Shopping Factors	Various frequencies	Cleaned and plotted
Frequency of Barriers to Sustainability	Various frequencies	Cleaned and plotted
Gender Preference for Sustainability	Not visualized	Plotted by gender
Chi-Square Test p-Value	Not tested	0.711 (for Gender vs. Switching to Sustainability)
Machine Learning Models	Not applied	Applied with LR, SVM, RF, GB, KNN

Table 1 describes the dataset before preprocessing, which contained a lot of duplicates, missing values, outliers, etc. and the cleaned dataset which was after preprocessing.

4. Results and Discussion

The dataset was first prepared for preprocessing; in this step, all the missing values and outliers were handled, and the cleaned dataset was extracted.

With the demographic data, various groups, factors, and aspects that affected the consumers' shopping behavior were studied. From Figure 2, both genders, especially females, and different age groups showed interest in the survey.

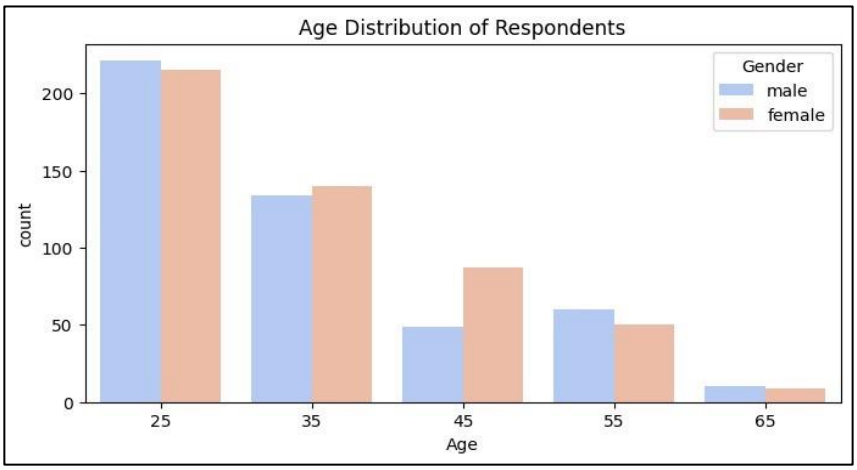

Figure 2. Age Distribution of Respondents

From Figure 3, the frequency of shopping for non-essential products was visualized, and it was concluded that most of the consumers shop for non-essential products monthly, and some of them a few times a year rather than daily or rarely.

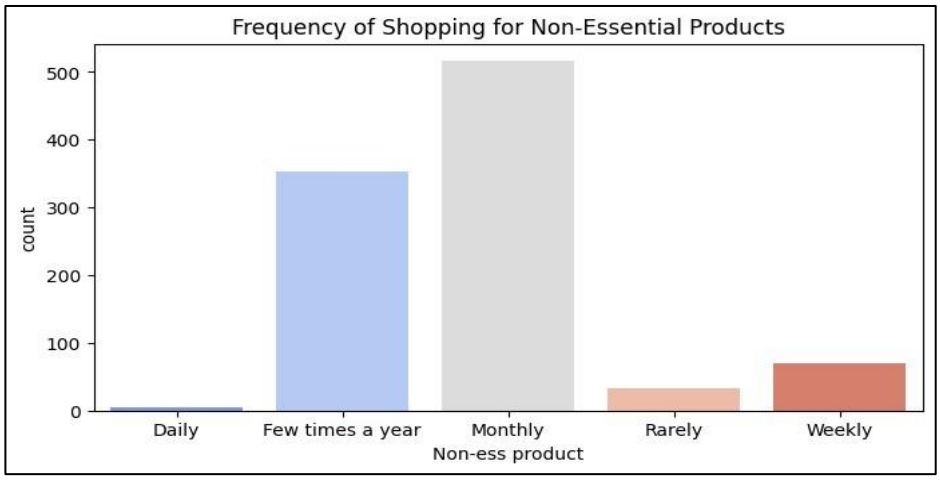

Figure 3. Frequency of Shopping

Considering the important aspects of buying sustainable products in Figure 4, most consumers prefer environmentally friendly materials and the recyclability of the products over other aspects.

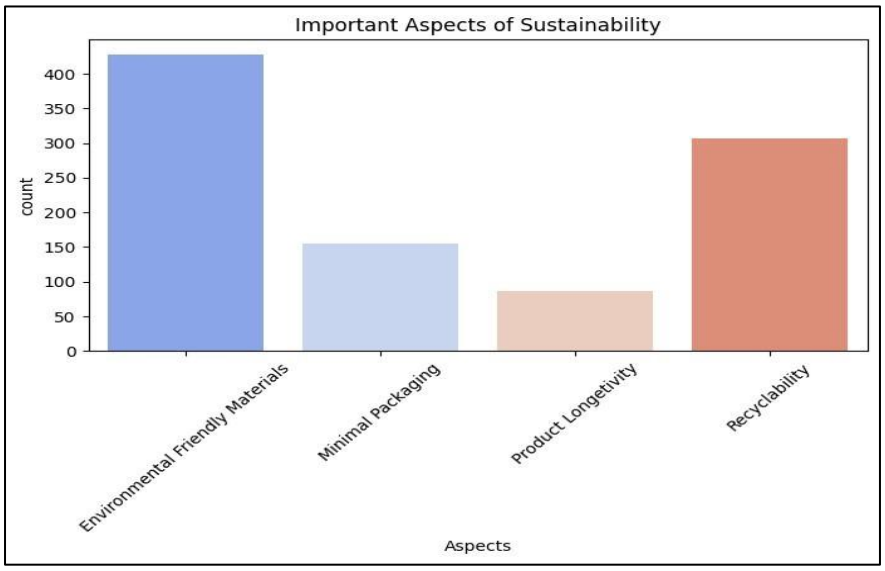

Figure 4. Aspects of Sustainability

Different places for shopping, such as physical retail stores, online marketplaces, brand-specific online stores, and local markets, were analyzed for highly preferred shopping for sustainable products. From Figure 5, it was understood that consumers prefer online shopping more than others.

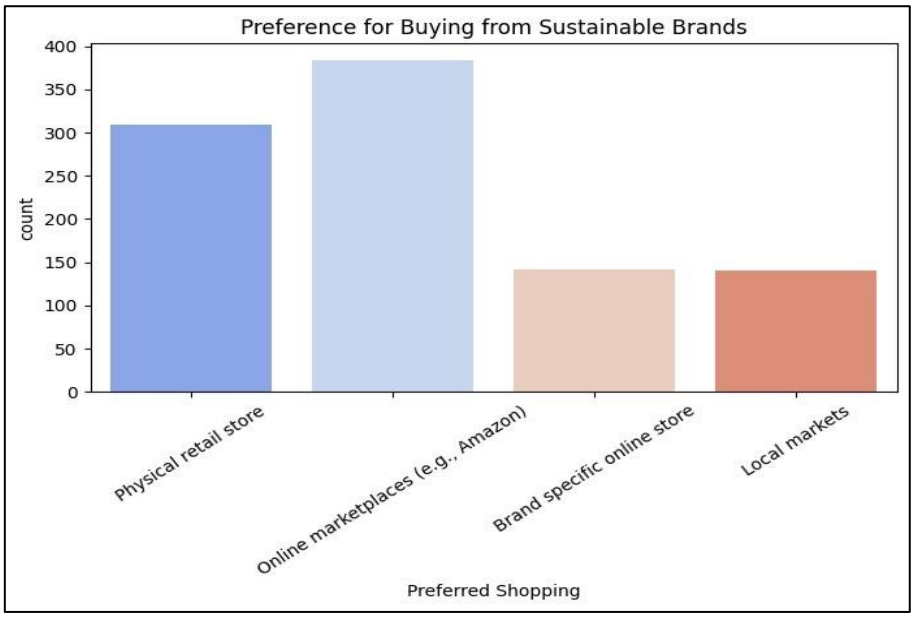

Figure 5. Shopping Preferences

Due to the limited availability of sustainable products in Figure 6, consumers find it difficult to opt for sustainability fully, and hence, the awareness of adopting sustainable products completely is necessary in today's day-to-day life.

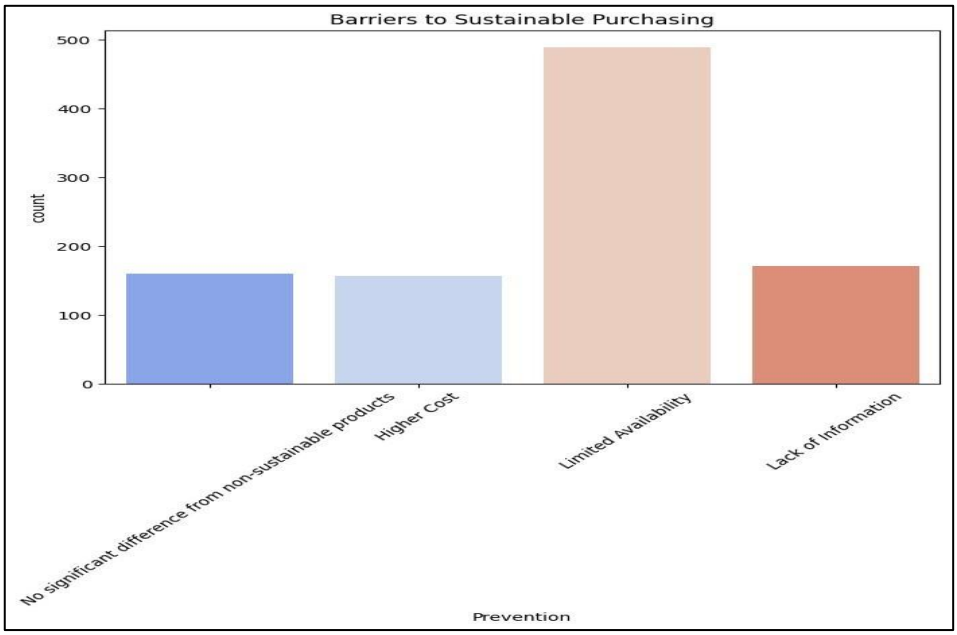

Figure 6. Barriers in Purchasing

Finally, in Figure 7, the willingness to switch completely towards sustainable products according to gender was examined, and it was found that almost everyone considered switching, most of them being women.

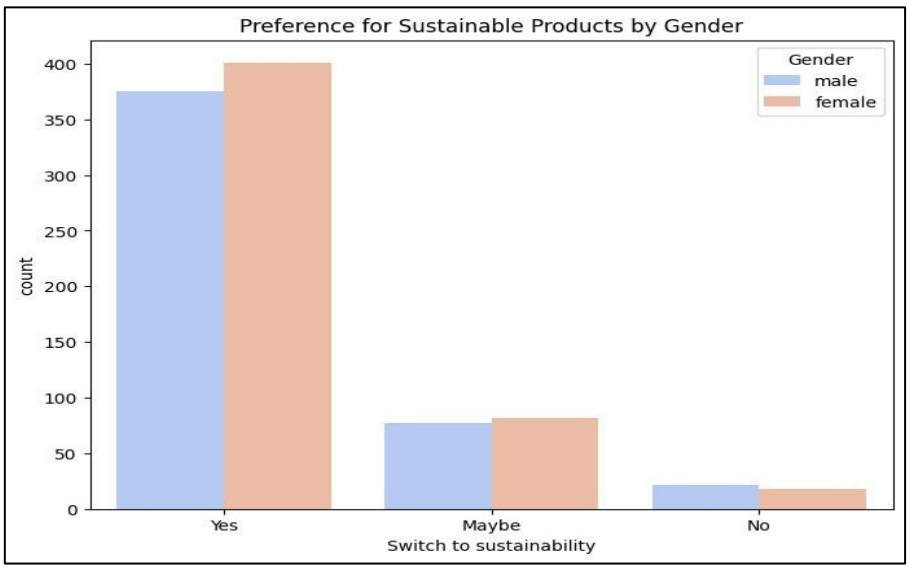

Figure 7. Shopping Preferences

The application of machine learning models on the dataset for predicting the likelihood of switching towards sustainable products gave an insightful result. Various models were evaluated on the dataset, and among all, the Random Forest and Support Vector Machine performed well in predicting the consumer switch toward sustainability.

From Table 2 below, the Random Forest model achieved an accuracy of 98.5%, whereas SVM showed a 90.3% accuracy. Both the models highlighted strong precision and recall, with the help of the SMOTE, which addressed the class imbalance in the dataset, leading to more reliable results, especially for minority classes such as those consumers who are actively purchasing sustainable products. SMOTE improved the model's ability to predict accurately.

On the other hand, gradient boosting showed 88.7% accuracy with quite good precision, recall, and F1 score. Logistic regression and KNN showed less accuracy, which is 75.2%, and 71.3% compared to other models.

Table 2. Performance Scores for all Tested Models

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	98.5%	0.98	1.00	0.99
Support Vector Machine	90.3%	0.92	0.97	0.94
Gradient Boosting	88.7%	0.91	0.95	0.93
Logistic Regression	75.2%	0.86	0.84	0.85
K-Nearest Neighbours (KNN)	71.3%	0.75	0.72	0.73

5. Conclusion

This study showed that machine learning models, especially Random Forest and Support Vector Machines, can be used quite effectively to predict consumer behavior concerning sustainable product adoption. Since younger, high-income consumers emerge as the strongest factor in the modeled factor for purchasing, the models offered insights into why consumers engage and purchase eco-friendly products.

The research also saw the use of advanced models that would address the issue of class imbalance regarding the dataset; hence, accurate predictions could be made and used for brands. However, the dataset was of a specific target demographic group, which perhaps would not apply to wider target audiences.

In future work, considering adding more heterogenicity about consumer segments in the dataset and utilizing real-time data, for example, social media sentiment, could further improve the prediction using these models and provide deeper knowledge about the adoption of sustainable products.

References

[1] Kumar, Bipul, Ajay K. Manrai, and Lalita A. Manrai. "Purchasing behaviour for environmentally sustainable products: A conceptual framework and empirical study." Journal of retailing and consumer services 34 (2017): 1-9.

[2] Ren, Shan, Yingfeng Zhang, Yang Liu, Tomohiko Sakao, Donald Huisingh, and Cecilia MVB Almeida. "A comprehensive review of big data analytics throughout product lifecycle to support sustainable smart manufacturing: A framework, challenges and future research directions." Journal of cleaner production 210 (2019): 1343-1365.

[3] Moon, Nazmun Nessa, Iftakhar Mohammad Talha, and Imrus Salehin. "An advanced intelligence system in customer online shopping behavior and satisfaction analysis." Current Research in Behavioral Sciences 2 (2021): 100051.

[4] Tabianan, Kayalvily, Shubashini Velu, and Vinayakumar Ravi. "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data." Sustainability 14, no. 12 (2022): 7243.

[5] Chandra, Shobhana, and Sanjeev Verma. "Big data and sustainable consumption: a review and research agenda." Vision 27, no. 1 (2023): 11-23.

[6] Panda, Nihar Ranjan. "A review on logistic regression in medical research." National Journal of Community Medicine 13, no. 04 (2022): 265-270.

[7] Harris, Jenine K. "Primer on binary logistic regression." Family medicine and community health 9, no. Suppl 1 (2021).

[8] Janan, Farhatul, and Sourav Kumar Ghosh. "Prediction of student's performance using support vector machine classifier." In Proc. Int. Conf. Ind. Eng. Oper. Manag, vol. 11, no. 12021, pp. 7078-7088. .

[9] Abdullah, Dakhaz Mustafa, and Adnan Mohsin Abdulazeez. "Machine learning applications based on SVM classification a review." Qubahan Academic Journal 1, no. 2 (2021): 81-90.

[10] Touw, Wouter G., Jumamurat R. Bayjanov, Lex Overmars, Lennart Backus, Jos Boekhorst, Michiel Wels, and Sacha AFT van Hijum. "Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?." Briefings in bioinformatics 14, no. 3 (2013): 315-326.

[11] Aria, Massimo, Corrado Cuccurullo, and Agostino Gnasso. "A comparison among interpretative proposals for Random Forests." Machine Learning with Applications 6 (2021): 100094.

[12] Nie, Peng, Michele Roccotelli, Maria Pia Fanti, Zhengfeng Ming, and Zhiwu Li. "Prediction of home energy consumption based on gradient boosting regression tree." Energy Reports 7 (2021): 1246-1255.

[13] Natekin, Alexey, and Alois Knoll. "Gradient boosting machines, a tutorial." Frontiers in neurorobotics 7 (2013): 21.

[14] Maillo, Jesus, Sergio Ramírez, Isaac Triguero, and Francisco Herrera. "kNN-IS: An Iterative Spark-based design of the k-Nearest Neighbors classifier for big data." Knowledge-Based Systems 117 (2017): 3-15.

[15] Zhang, Shichao, and Jiaye Li. "KNN classification with one-step computation." IEEE Transactions on Knowledge and Data Engineering 35, no. 3 (2021): 2711-2723.

Author's biography

Zaobiya Khan holds a bachelor's degree in Information Technology (BSc.IT) with a CGPA of 9.2 and is currently pursuing a master's degree in Information Technology (MSc.IT) at SVKM's Usha Pravin Gandhi College. With a strong background in full-stack development, Zaobiya is proficient in technologies such as Python, Django, Java Spring Boot, and AWS. Her experience spans creating REST APIs for seamless communication between client and server applications.. In addition to her technical expertise, Zaobiya has worked on various projects, including a student placement prediction system and a food ordering platform, demonstrating her ability to build innovative solutions.

Neha Vora is currently pursuing her Ph.D. in Computer Science and holds a master's in computer applications (MCA). She is qualified in NET, SET, and GATE, and brings over 9 years of teaching experience, along with 1 year of industry experience. Her primary research

areas include computer vision, image processing, machine learning, object detection, and artificial intelligence