Smart E-Commerce Recommendation Engine using Data Engineering and Machine Leaning

|  |  |  |
| --- | --- | --- |
| Muskan Sureka  *Department of ECE*  *Haldia Institute of Technology*  Haldia, West Bengal, India  muskan240305@gmail.com | Aadi Poddar School of Computer Engineering KIIT Deemed UniversityBhubaneswar, Odisha, INDIA aadipoddarmail@gmail.com | Dr. Saurabh *School of Computer Engineering KIIT Deemed to be University* Bhubaneshwar, Odisha, India saurabh@gmail.com |
| Mahendra Kumar Gourisaria  *School of Computer Engineering KIIT Deemed to be University* Bhubaneshwar, Odisha, India mkgourisaria2010@gmail.com | Muskan Sureka  *Department of ECE*  *Haldia Institute of Technology*  Haldia, West Bengal, India  muskan240305@gmail.com | Aadi Poddar School of Computer Engineering KIIT Deemed UniversityBhubaneswar, Odisha, INDIA aadipoddarmail@gmail.com |

Abstract – Product recommendation in E-Commerce websites with the customers’ purchase patterns over the years through online shopping is of the highest demand in the industries. These recommendations are based on their previous purchases over the years with their specific price point, location and the season of the purchase. This helps in increasing customer satisfaction further leading in maximizing the profits of the E-Commerce businesses. This paper draws attention to the product recommendation and prediction system based on the dataset including detailed customers purchase patterns over the years. Different ML classification techniques like Support Vector Machine, Random Forest Classifier, K-Nearest Neighbor, Logistic Regression, Gradient Boosting, XG Boost and Decision Tree were tested and trained on the dataset and comprehensively evaluated with F1-score, accuracy, precision and recall. It is found out that Random Forest Classifier outperforms other models giving the best results in F1-score with a value of 0.99 and an Accuracy of 99.98%.

Keywords—Online Shopping, Product Recommendation, Machine Learning, Recommendation Modeling, Classifier Analysis

# Introduction

Product recommendation (PR) is a model or system that makes predictions about products based on a user's past searches and purchases. It might not be totally true, although they can be slightly tailored to the user's preferences. Any kind of recommendation may be made in light of the user's data. For instance, when you watch a movie or song on YouTube, you will receive recommendations that are similar to the movie when it ends [1]. Similarly, when shopping online, one will receive recommendations that are similar to what they have previously purchased based on your past purchases.

Several scientists performed research works and methods of getting better ways of product recommendation on online shopping platforms. This process is based on many techniques. User preference models that are developed from browsing history, orders, shopping cart additions, searches, likes, favourites, comments, and other behaviours are typically used to recommend products [2]. These models are computed using map reduce and trained offline using logistic regression (LR). User client reporting monitors users' clicking, browsing, and shopping cart behaviours to give them real-time feedback. Storm or Spark streaming computations are used to build real-time customer preferences, even though handling product and user relationships and the related demands on storage space and access performance for online services are extremely difficult.

In the present research, various machine learning classification models — Support Vector Machine (SVM), Random Forest Classifier (RFC), K-Nearest Neighbor (KNN), Logistic Regression (LR), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), and Decision Tree (DT) — are considered and compared to determine the best method for recommending products. The main goal is to effectively pair products with the correct target audience according to past purchase behaviors, thus enhancing customer satisfaction, scalability, and profitability for e-commerce companies.

The recommender system being proposed here makes use of three essential elements that were used to create this system: datasets, user-based prediction ranking, and cosine similarity. The user is the one who rates, and the number of accurate ratings they provide determines how accurately they rate anything. The results are then arranged according to cosine similarity [3]. To reduce time looking for clients, the proposed method for recommending research papers matches a user with the best articles in their field depending on how similar their searches are to those of other users. This is done using a collaborative filtering mechanism.

Websites with mass scale production use advance models in the production pipelines to efficiently sort the defected and accepted produced goods. They take advantage of various machine learning models, trained on several datasets to effectively sort the damaged products. This solves a major problem of getting the target audience for the right products also it saves the manual calculation time taken to find product probability. To solve this problem, data scientists and machine learning engineers around the world are working to make more efficient models to accurately predict the products for the users giving them a personalized experience. This is done just by collecting the user history and algorithms of their shopping techniques.

Every product page on an e-commerce platform includes a number of fields for summarizing the item and its attributes.   
Textual elements like titles, descriptions, categories, and tags are especially crucial in these systems because they offer valuable information for information retrieval (IR) services and have a significant impact on customer behavior, which in turn influences sales and revenue success. These textual components are crucial for e-commerce search and product recommendation engines to assist users locate what they're looking for fast or learn about new concepts and goods [4].

ML algorithms can also be used in different other fields like maize leaf disease diagnosis [5], rice disease prediction [6], diabetes prediction [7] etc. The rest of the paper is organized like: Section II. Literature Review, covering significant publications on the subject; Section III. tells about Methods and Materials, providing details on the classifiers and strategies employed; Section IV covers Implementation and Results, contrasting the performance of various regression models to determine the most effective model; and Section V. Conclusion and Future Work, summarizing findings and proposing future research directions followed by references followed by reference

# Related Work

This section provides the review of various research works performed by various scientist and machine learning engineers from over the world in the field of Online Shopping and Purchasing Priorities on customers. The study of D. Koehn *et.al (2020)* [8] targeted marketing initiatives in real-time and forecasts online purchase behavior using clickstream data. Such AI-powered targeting has been shown to increase store sales and save enormous sums of money on marketing expenses. The study suggests a technique that uses the recurrent neural network (RNN) framework to fully exploit clickstream data. On every metric, only the two RNN-based designs are able to surpass the naïve model. The GRU appears to perform somewhat better than the LSTM among these two designs with an R2 Score of 0.4337. Malik et al. (2022) [9] proposed EPR-ML, an integration of ML and NLP for tag-based product recommendation with textual descriptions. They experimented with various classifiers, and L-SVM showed the highest accuracy (96%) in predicting trending video duration. Tiwari et al. (2025) [10] performed sentiment analysis on more than 11,000 Amazon reviews using SVM, Random Forest, and Naive Bayes for sentiment classification. Their joint models had accuracy levels of 0.85, 0.87, and 0.85 and provided insights for enhancing e-commerce prediction and user experience.

Noman et al. (2025) [11] examined Bangladeshi product reviews on e-commerce to determine customer satisfaction, for which Random Forest had 94% predictive accuracy in determining consumer happiness. This paper provides insights to improve user experience and strategies in the emerging Bangladeshi market. Kumar et al. (2025) [12] researched restructuring websites by user behavior, interaction, and preferences with data mining and predictive modeling. Based on two datasets, Random Forest performed well in predicting purchase intention (precision 89.8%, TP rate 90.3%), whereas Bayes Net performed well in predicting evolving shopping behavior (precision and TP rate 99.7%). Youbi et al. (2023) [13] proposed a dynamic pricing model using machine learning, where the GBM model, optimized through hyperparameter tuning, achieved strong results (MSE = 0.012, R² = 0.92), outperforming other techniques. Singh et al. (2020) [14] developed ML algorithms to predict e-commerce sales, with gradient boosting delivering superior train accuracy (99.99%) compared to Random Forest, enabling better insights into current and future sales. A constructive and detailed analysis of all the literature reviews alongside their performance score i.e., accuracy is shown in Table.1

Table I. Analysis of Literature Review

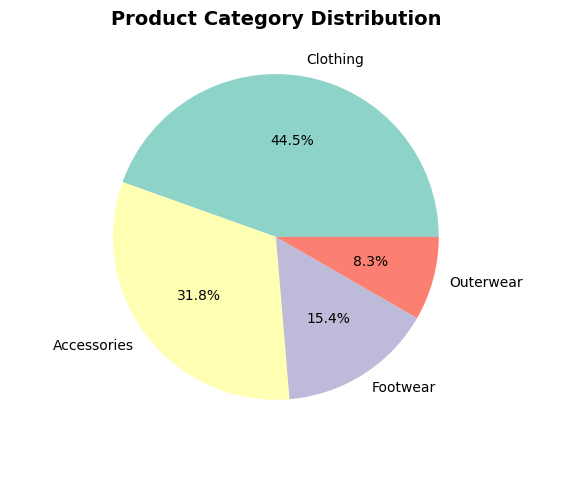
|  |  |  |  |
| --- | --- | --- | --- |
| Year | Model | Dataset | Results |
| [8] 2020 | Support Vector Machine | Supplementary Raw Research Data | 0.4337 R2-Score |
| [9] 2022 | Least Squares Support Vector Machine (LSSVM) | Kaggle Dataset | 96.00% Accuracy |
| [10] 2025 | Random Forest Classifier | Amazon product Feedback over 11,000 datasets | 87.00% Accuracy |
| [11] 2025 | Random Forest Classifier | Bangladesh E-Commerce 6,964 samples | 94.00% Accuracy |
| [12] 2025 | Random Forest  Bayes Net | Online shoppers intention1  Ecommerce Customer Behavior2 | 90.30%  99.70%  Accuracy |
| [13] 2023 | Gradient Boosting | 10,000 rows of real-world transaction data | 0.92  R2-Score |
| [14] 2020 | Gradient Boosting | Kaggle Brazilian ECommerce Public Dataset by Olist Store | 99.99%  Accuracy |

# PROPOSED METHODS

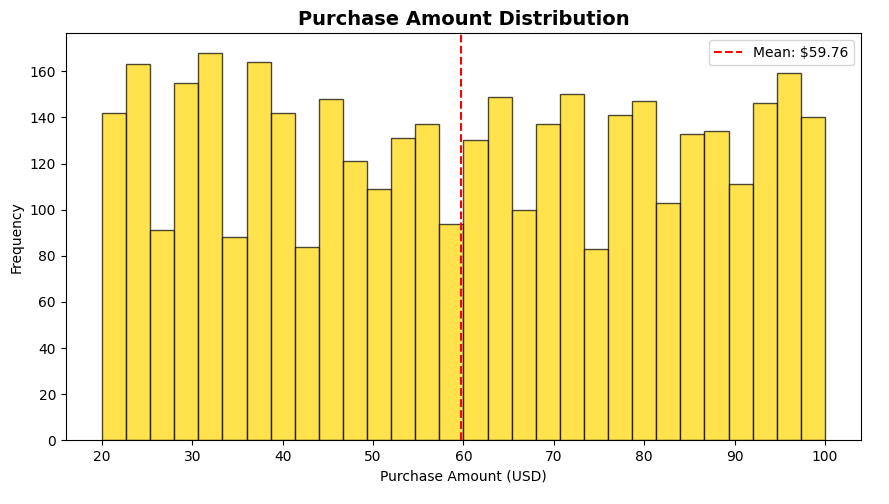
Several machine learning classifier and regression models work differently on different proposed datasets and it is entirely up to us as to which model should be trained and tested accordingly to obtain the best results in the most efficient time possible with the highest accuracy. This paper uses sampling technique like Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. The results are tested and passed through different machine learning classifier techniques which includes algorithms like Support Vector Machine, Random Forest Classifier, K-Nearest Neighbor, Logistic Regression, Gradient Boosting, XG Boost and Decision Tree. These classification algorithms are discussed further, along with the results obtained after training and testing the models on the dataset. The subsection A. described the Dataset Used, B. Hardware and Software used, C. Preprocess Data, D. talk about Splitting the Data, E. Handling Dataset Imbalance F. Model Workflow Diagram, G. Classification Models Used, and H. Comprehensive Evaluation Metrices.

## Dataset Used

The dataset consists of over 3900 records of customer behaviors in online purchasing in E-Commerce websites over several years. The dataset is an open source and free dataset from Kaggle [15]. Comprehensive information on the preferences, inclinations, and trends of customers during their purchasing experiences may be found in the Consumer Behavior and purchasing Habits Dataset. Numerous characteristics are included in this dataset, such as demographic data, past purchases, product preferences, frequency of purchasing, and online and offline shopping habits. With the use of this extensive data set, analysts and researchers may explore the complexities of consumer decision-making, which helps companies develop more focused marketing campaigns, improve their product lines, and raise customer satisfaction levels. The product Category Distribution and Purchase Amount Distribution is shown in Fig-1 and Fig-2 respectively.



1. Product Category Distribution



1. Purchase Amount Distribution

## Hardware and software

All models were trained and tested on a workstation with an Intel 12th Gen Intel(R) Core (TM) i7-1255U 1700 MHz CPU, 10 cores, 12 logical processors, and 16 GB of RAM running Windows 11 Home Single Language OS. The dataset was trained and tested using Python 3 on Google Collab, Scikit-Learn and PyTorch for machine learning classifier models for prediction and processing.

A chart with different colored squares

AI-generated content may be incorrect.

1. Feature Correlation Matrix of Dataset

## Preprocess Data

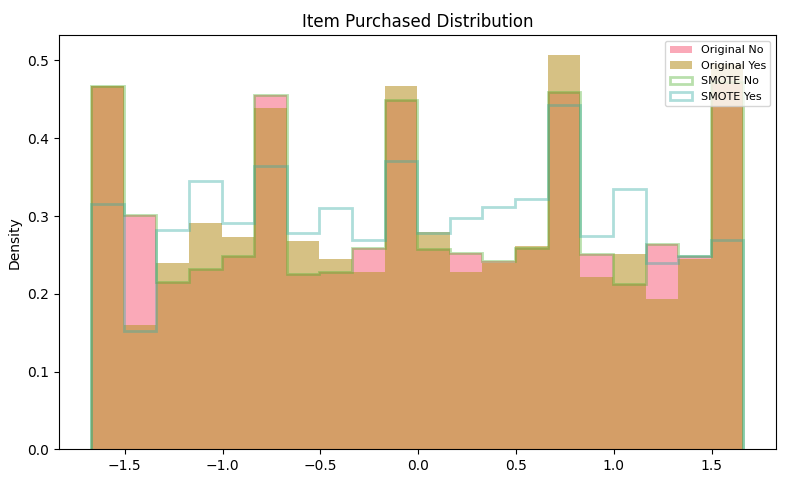
The dataset is loaded with the CSV file with over 3900 records of customer data. The data is preprocessed by converting all 17 features in the format of text and string variables to numerical data which is understood by the machine. The other features are scaled between 0 and 1 according to their unique values like Location has 50 unique values and Color has 25 unique values. Fig-3 shows the Feature Correlation Matrix which is the importance of the features and attributes of the dataset.

## Data Splitting

The dataset used is already split into test and train sections with the training set of data consists of 70% of the data and the testing set includes 30% of the data. The dataset as a whole has 3900 records of consumers purchase history. The testing set of the classification models evaluates performance on unseen data. Dependent variables are predicted using the trained algorithm, and results are compared against original values to minimize error.

## Dataset Imbalance Handling

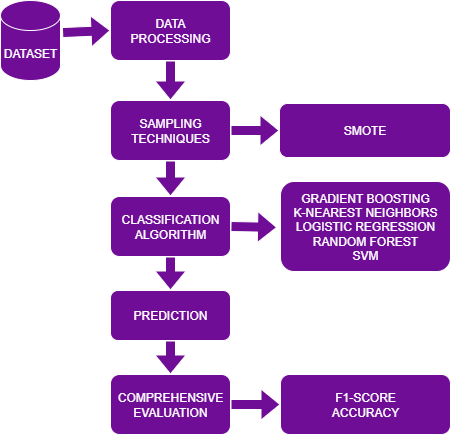
It is inferred from the dataset that it is imbalanced with the ratio 2.70:1 in each test and train dataset. Sampling technique such as Synthetic Minority Oversampling Technique (SMOTE) is used in this paper to handle such imbalances. The different sampling techniques are set up as specified below. One effective technique for addressing datasets with class imbalance is the Synthetic Minority Over-Sampling Technique. In order to address this problem and achieve a balanced class distribution, SMOTE creates samples of minority classes. SMOTE creates artificial examples in the minority class's feature space [16]. The Item purchased distribution of original vs SMOTE is shown in Fig-4.



1. Item Purchased Distribution Before and After SMOTE

## Model Workflow Diagram

A well-thought-out pipeline is developed that applies the classifier models after first normalizing the features and preprocessing the dataset. The data is passed through sampling strategies like SMOTE to handle class imbalance. The results from the sampling techniques applied on the dataset and are passed through the different classifier machine learning models. Finally, the results are compared to find out the most accurate and effective model. To create a good machine learning model, the pipeline should be framed from the dataset to determining the scores and outcomes of the algorithms. Fig-5 shows the flowchart of all the steps needed to create a model and test it on a processed dataset.



1. Model Workflow Diagram

## Classification Models

We have used Machine Learning Classification models on the dataset like Gradient Boosting, K-Nearest Neighbors, Support Vector Machine, Logistic Regression, Decision Tree, XGBoost and Random Forest Classifier. A brief description of all the models is given below.

### Gradient Boosting

A set of potent machine-learning methods known as gradient boosting provide remarkable outcomes in practical settings. For example, they may be easily tailored to the particular needs of the dataset. Gradient boosting is a powerful machine learning technique that combines many "weak" learners, often decision trees, to create a prediction model that is more accurate and dependable. The method is guided by the gradient descent optimization approach, which helps to minimize a chosen loss function and enhances overall performance [17]. After training, the learning rate η, which can range from 0 to 1, is multiplied by each tree's predictions to lower them. Predictions are produced by summing the contributions of each tree once they have all been trained. The following formula provides the final prediction in Eq-1.

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

In the Equation *r₁, r₂, …, rₙ* are the errors predicted by each tree.

### K-Nearest Neighbors (KNN)

A supervised machine learning technique called K-Nearest Neighbors is frequently used for classification problems, while it may also be used for regression tasks. It creates predictions based on the majority class (for classification) or the average value (for regression) after identifying the "k" data points that are most similar to a given input. KNN is an instance-based, non-parametric learning technique as it does not assume anything about the distribution of the underlying data. thus, instead of learning from the training set of data, it saves the dataset and then acts upon it during classification. A lazy learner algorithm is a common term for it [18]. The straight-line distance between two locations in a plane or space is known as the Euclidean distance that is shown in Eq-2.

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

### Logistic Regression

One supervised machine learning technique for categorization issues is logistic regression. It predicts the likelihood that an input belongs to a certain class, as opposed to linear regression, which predicts continuous values. Using the sigmoid function, it transforms inputs into a probability value ranging from 1 to 0. The sigmoid function is used to transform the value linear regression function output into a categorical value output [19].

### Support Vector Machine (SVM)

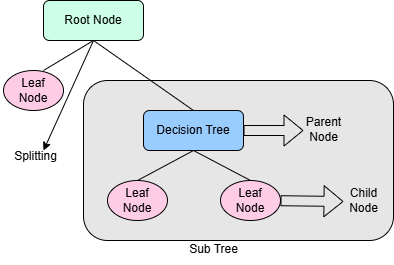
A supervised machine learning tool utilized in regression and classification applications is the support vector machine. It searches for the best boundary, or hyperplane, between the several classes of data. The basic idea behind the SVM algorithm is to find the hyperplane that optimally separates two classes by maximizing their margin. The margin is the separation on each side between the nearest data points (support vectors) and the hyperplane [20].

### XGBoost

eXtreme Gradient Boosting, or XGBoost for short, is a cutting-edge machine learning technique that is optimized for effectiveness, speed, and high performance. Conventional machine learning models, such as random forests and decision trees, are simple to understand yet frequently exhibit poor accuracy when applied to intricate datasets. Boosting is the technique by which each new tree is trained to fix the mistakes produced by the one before it [21].

### Decision Tree

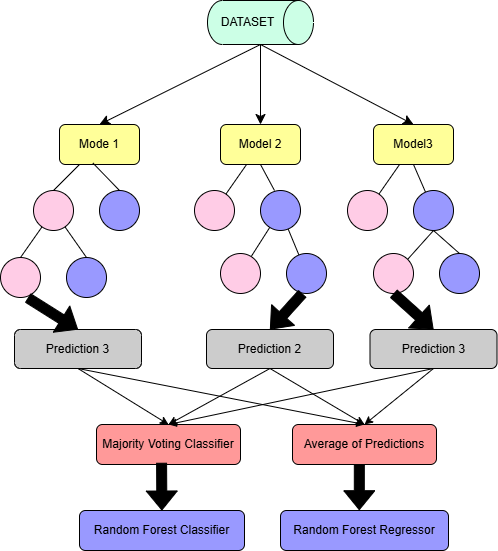
This is a sort algorithm where data is split according to a certain parameter [22]. Feature are selected by Information gain. Decision Tree does not have a good time complexity for execution purpose. Decision Tree Regression also uses Entropy, Information Gain, Gini Impurity to calculate the further results. Fig-6 shows the functioning of the Decision Tree regressor in detail with proper explanation of the root and leaf nodes.



1. Decision Tree

### Random Forest

The main problem with decision trees is that overfitting occurs when they are built without a hyperparameter [23]. Reducing the high volatility of decision trees is the main objective of Random Forest. As a result, the models replicate datapoints, records, and characteristics. All of the models will be quite good at predicting the specific data that they were trained on. The Random Forest Regressor's operation is described in detail in Fig-7 using several decision tree models [24].



1. Random Forest Classifier

## Comprehensive Evaluation Metrices

Analyzing and testing datasets is the foundation of machine learning. provides the forecasts once the model has been processed using certain methods. The regression model's accuracy, or how well it anticipated the outcomes, is computed. To evaluate the model's correctness and behavior, a comparison between the expected and actual values was shown. Any incorrect predictions the model makes are represented as errors. The models are evaluated based on certain comprehensive metrices like F1-Score, Accuracy, Precision and Recall value that are explained below.

### F1-Score

A statistic which access a classification model's performance, particularly when working with unbalanced datasets, is the F1-score. It provides a balanced metric taking into account both false negatives and false positives by representing the harmonic mean accuracy.

### Accuracy

Accuracy is the percentage of accurate predictions made by a model taking into consideration all the forecasts. It is an essential indicator for assessing how well a model performs, particularly in classification jobs. But accuracy by itself can be deceptive, especially when datasets are unbalanced and one class has a disproportionately high number of instances.

# Results And Discussion

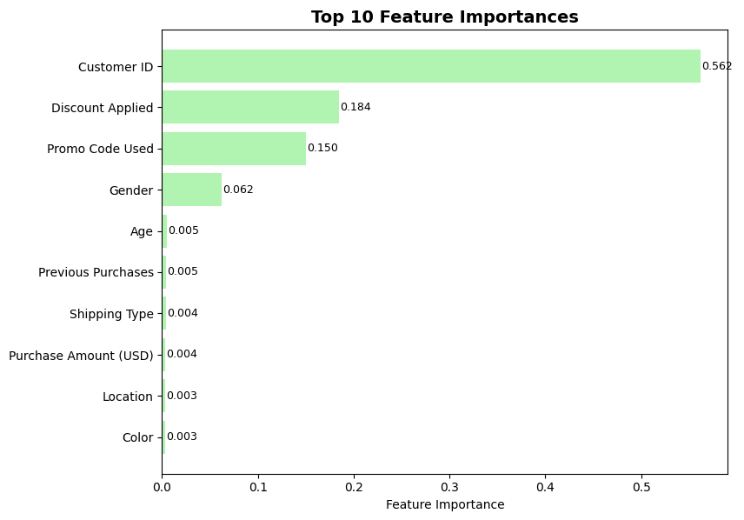
Machine learning classifier models were evaluated and tested on the dataset based on their efficiency and performance to observe the robustness and effectiveness of the model. The results and judging of the best model can be verified by several ways such as by checking the F1-Score, Accuracy, Precision and Recall. F1-Score is considered to be the best evaluation matrix for imbalanced datasets. The Random Forest Classifier on original dataset with SMOTE applied to balance the dataset outperforms as the best classifier model by giving the best results in F1-Score and Accuracy. Table 2 shows the performance of the overall best machine learning classifier models with SMOTE applied on the dataset used in this study, along with their evaluation scores like F1-Score, Accuracy percentage, precision and recall value.

Table II. Analysis of related work

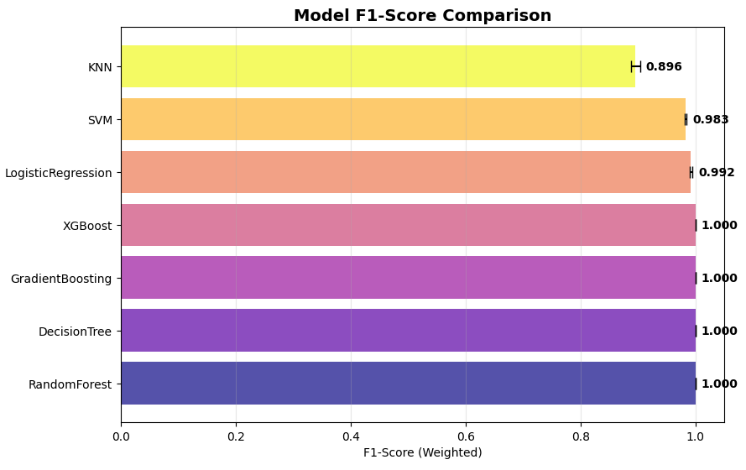
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Classifier | F1-Score | Accuracy | Precision | Recall |
| Random Forest | 0.9998 | 99.98% | 0.9998 | 0.9998 |
| Decision Tree | 0.9998 | 99.98% | 0.9998 | 0.9998 |
| Gradient Boosting | 0.9998 | 99.98% | 0.9998 | 0.9998 |
| XGBoost | 0.9998 | 99.98% | 0.9998 | 0.9998 |
| Logistic Regression | 0.9920 | 99.20% | 0.9922 | 0.9920 |
| SVM | 0.9827 | 98.27% | 0.9831 | 0.9827 |
| KNN | 0.8956 | 89.65% | 0.9107 | 0.8965 |

The best overall F1-Score, Accuracy, Precision and Recall values were achieved with Random Forest Classifier with SMOTE applied. It outperforms all other machine learning classifier models with a F1-Score of 0.9998, an Accuracy of 99.98%, Precision of 0.999824715 and a Recall value of 0.999824407 with 5 folds satisfied in Cross Validation. The least performing model was KNN with 89.65% of Accuracy and an F1-Score of 0.8956.

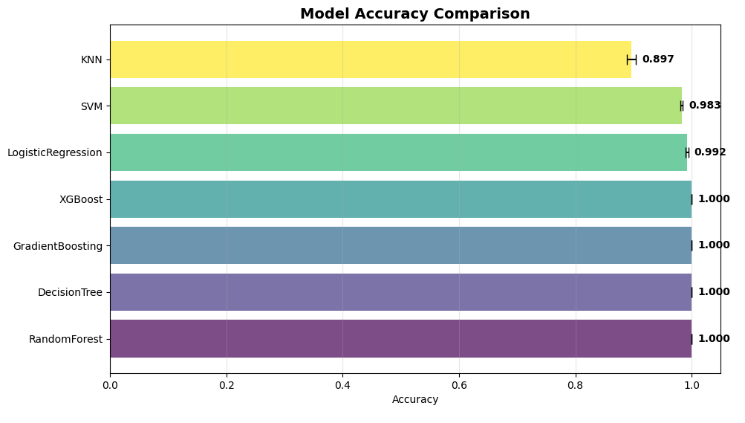
Several useful business insights were retrieved from the model training like; the best target segment of consumers was between the age group of 46 to 55 which was 30.4%. The top category of product sold was Footwear with was 28.5%. The average amount spent by the consumers was 59.49 dollars. The customer retention rate was greater than average with a percentage of 26.1% of subscribers from previous purchases. Fig-8 shows the Feature Importance of the top 10 features of the dataset. Fig-9 shows the F1-Score comparison of all the machine learning models used in this paper. Fig-10 shows the model accuracy comparison of all the machine learning models used in this paper.



1. Top 10 Feature Importance of the Dataset



1. F1-Score Comparison of all Classifier Models



1. Accuracy Comparison of all Classifier Models

# Conclusion And Future Work

This paper uses consumer based records based on their purchases on E-Commerce websites based on location, category and other attributes with the purchase price and the discount retrieved. This dataset is used to train and test the machine learning model for accurately predicting and recommending accurate products based on the consumer purchases. The pipeline provides a solid foundation for industrial product recommendation with state-of-the-art feature extraction and robust machine learning techniques. This research identifies the overall best model from the combined results and save the updated results and best model pipeline. Random Forest outperforms all models that were tested on the dataset.

For further research, the researchers can train bigger real world datasets for better results in product recommendation. While the original feature set with SMOTE applied on it yielded the highest F1 score, the reduced/selected feature sets provide viable alternatives that could offer computational benefits in terms of training time and memory usage without a significant drop in performance. Further an automated machine learning can be deployed to the E-Commerce websites for accurately recommending the products to the consumers based on their past purchases which further leads to maximize business profits and higher customer satisfaction.

##### REFERENCES

1. A.N. Sodhar, U.A. Khan, I.N. Sodhar, A.H. Buller, J. Sodhar. Product Recommendation Using Machine Learning a Review of the Existing Techniques. IJCSNS;22(5):523, 2022.
2. K. Xu, H. Zhou, H. Zheng, M. Zhu, Q. Xin. Intelligent classification and personalized recommendation of e-commerce products based on machine learning. arXiv preprint arXiv:2403.19345. 2024.
3. S. Priyanka, N. Abinaya, S. Keerthika, S. Santhiya, P. Jayadharshini, B. Vinothini. Product Recommendation System Using Machine Learning. In2024 2nd International Conference on Disruptive Technologies (ICDT) (pp. 1515-1518). IEEE, 2024.
4. V. Malik, R. Mittal, S.V. SIngh. EPR-ML: E-Commerce Product Recommendation Using NLP and Machine Learning Algorithm. In2022 5th International Conference on Contemporary Computing and Informatics (IC3I) (pp. 1778-1783). IEEE, 2022.
5. P. Theerthagiri, A. U. Ruby, J. G. C. Chandran, T. H. Sardar, & A. Shafeeq BM, Deep SqueezeNet learning model for diagnosis and prediction of maize leaf diseases. Journal of Big Data, 11(1), 112, 2024.
6. S. Das, T. H. Sardar, & D. S. Sahana, An ensemble technique using genetic algorithm and deep learning for the prediction of rice diseases. In Machine Learning Hybridization and Optimization for Intelligent Applications, pp. 289-303, CRC Press, 2025.
7. A. Raj, J. J. Jena, D. Ghosh, S. Jain, T. H.Sardar, & M. K.Gourisaria, Early-Stage Diabetes Risk Prediction Using Machine Learning Techniques. In 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), pp. 2289-2295, 2025
8. D. Koehn, S. Lessmann, M. Schaal. Predicting online shopping behaviour from clickstream data using deep learning. Expert Systems with Applications. 2020.
9. V. Malik, R. Mittal, SV. SIngh. EPR-ML: E-Commerce Product Recommendation Using NLP and Machine Learning Algorithm. In2022 5th International Conference on Contemporary Computing and Informatics (IC3I) (pp. 1778-1783). IEEE, 2022.
10. N.K. Tiwari, H. Arora. Sentiment Analysis and Forecasting for Improved Business Performance in E-Commerce using Machine Learning Algorithms. In2025 International Conference on Electronics and Renewable Systems (ICEARS) (pp. 1487-1491). IEEE, 2025.
11. A. Noman, M.D. Rafi, P.S. Antara, M.M. Tanzim, M.W. Payel, M.H. Bijoy. Predicting Customer Sentiment from Bangladeshi E-Commerce using Machine Learning Techniques. In2025 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-6). IEEE, 2025.
12. S. Kumar, S. Mishra, D. Rathour. Customer Behaviour Prediction for E-Commerce Sites Using Web Mining and Machine Learning. In2025 3rd International Conference on Communication, Security, and Artificial Intelligence (ICCSAI) (Vol. 3, pp. 1-6). IEEE, 2025.
13. R.E. Youbi, F. Messaoudi, M. Loukili. Machine learning-driven dynamic pricing strategies in E-commerce. In2023 14th International Conference on Information and Communication Systems (ICICS) (pp. 1-5). IEEE, 2023.
14. K. Singh, P.M. Booma, U. Eaganathan. E-commerce system for sale prediction using machine learning technique. InJournal of Physics: Conference Series (Vol. 1712, No. 1, p. 012042). IOP Publishing, 2020.
15. S. Aslam, Consumer Behaviour and Shopping Habits Kaggle Dataset, 2023, Accessed: 2025-08-10. <https://www.kaggle.com/datasets/zeesolver/consumer-behavior-and-shopping-habits-dataset>
16. N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. and Kegelmeyer, SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357, 2002.
17. C. Bentéjac, A. Csörgő and G. Martínez-Muñoz. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), pp.1937-1967, 2021
18. H. Wang, D. Bell, B. Yaxin, Guo, Gongde and K. Greer. "KNN model-based approach in classification." In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", pp. 986-996. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003.
19. P. Michael, LaValley, Logistic regression. Circulation, 117(18), pp.2395-2399, 2008.
20. S. Suthaharan. "Support vector machine." In Machine learning models and algorithms for big data classification: thinking with examples for effective learning, pp. 207-235. Boston, MA: Springer US, 2016.
21. T. Chen, C. Guestrin. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794) 2016.
22. Y.Y. Song, Y. Lu. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry. 2015.
23. S. J. Rigatti. Random forest. Journal of Insurance Medicine, 47(1), 31-39, 2017.
24. G. Biau, and E. Scornet, A random forest guided tour. *Test*, *25*(2), pp.197-227, 2016.