IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Customer Credit and Buying Profile in E-commerce

Arnav Gupta, R. Jatin Kumar, Kumari Puja SharmaChowdam Likitha, Viraj M.

Abstract

The rise of e-commerce has posed challenges for businesses regarding returns, cancellations and varied pay- ment preferences while ensuring customer satisfaction. This paper proposes an innovative methodology that utilizes big data, Machine Learning Techniques and statistical analysis to construct customer profiles while up- holding individual privacy. Through the aggregation and anonymization of customer data, the system generates personalized credit and purchasing profiles. This data-driven approach enables tailored delivery benefits and EMI options for customers. Notably, the system categorizes customers based on their credit and purchasing behavior, limiting Cash on Delivery for users with unfavorable buying profiles and withholding EMI choices for those with less favorable credit profiles. Prioritizing privacy protection, this solution optimizes e-commerce op- erations while respecting customer confidentiality, striking a balance between service customization and privacy preservation.

1 Introduction

The research addresses the dynamic challenges in e-commerce by proposing a sophisticated system for creating personalized credit and purchasing profiles while prioritizing customer privacy. Leveraging advanced technologies like big data analytics, artificial intelligence, and statistical analysis, the system aims to provide e-commerce platforms with profound insights into customer behaviors. The key focus is on tailoring services based on individualized profiles to enhance the shopping experience. Crucially, the system emphasizes privacy preservation through anonymization and aggregation techniques. The research underscores ethical considerations, emphasizing transparency, consent,and adherence to privacy regulations. It recognizes the technical intricacies of harnessing technology and seeks to navigate them effectively within legal frameworks. The ultimate goal is to offer a comprehensive roadmapfor deploying a privacy-respecting yet highly effective system for customer profiling and service customization in e-commerce.

2 Literature Review

2.1 Credit Scoring:

- Wang and Yang (2020)[1]: Their unspecified ML model achieves high accuracy (85.71%) in credit scoring forbank credit cards. While the specific algorithm remains unknown, this result suggests ML effectiveness in financial risk assessment.
- Zhang et al. (2020)[2]: This study focuses on using multiple data sources (personal information + credit card data) to build a credit scoring model, but lacks accuracy reports. Exploring multi-source data could enhance model accuracy and capture a more holistic financial picture.
- Maheshwari et al. (2019)[3]: This paper delves into profiling e-commerce customer credit behavior but doesn't mention ML techniques or accuracy. Analyzing e-commerce purchase data for credit profiling holds immense potential but requires exploration of advanced ML algorithms.

2.2 Customer Segmentation:

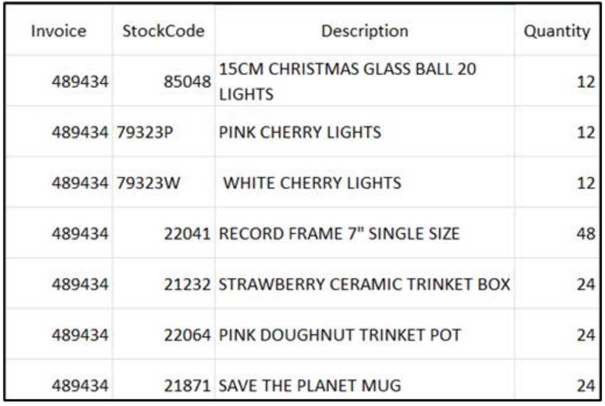
- Wu et al. (2022)[4]: They introduce an improved k-medoids clustering algorithm for e-commerce customer segmentation, achieving moderate clustering quality (silhouette coefficient = 0.585). This shows that customized clustering approaches can offer better results than standard algorithms.
- Dursun and Caber (2016)[5]: Though lacking accuracy metrics, their application of RFM and hierarchical clustering to segment hotel customers demonstrates the effectiveness of traditional methods in specific contexts.

- Christy et al. (2021)[6]: While they utilize RFM ranking for segmentation, the unspecified ML methods and lack of accuracy data limit our understanding of their approach. RFM remains a valuable tool for initial customer grouping, but integrating ML can refine segmentation further.
- Dogan et al. (2018)[7]: Similar to Dursun and Caber (2016), this study uses RFM and k-means for retail customer segmentation without mentioning accuracy. This reinforces the applicability of these methods in offline retail environments.
- Cheng and Chen (2009)[8]: They achieve an impressive accuracy of 85.6% by combining RFM with rough set theory for customer value segmentation. This highlights the potential of hybrid approaches for improved segmentation and customer value assessment.
- Brahmana et al. (2020)[9]: Their comparison of k-means, k-medoids, and DBSCAN for RFM-based segmentation concludes that k-means performs best, offering valuable insights into choosing the right clustering algorithm based on data characteristics.
- Kabasakal (2020)[10]: Similar to Dursun and Caber (2016), this study focuses on RFM and k-means for e-retail segmentation without accuracy metrics. This further solidifies the prevalence of these methods in e-commerce customer analysis.
- Firdaus and Utama (2021)[11]: Though lacking details on ML techniques and accuracy, their development of a bank customer segmentation model using RFM+B (Behavior) expands the traditional RFM approach, suggesting the importance of incorporating broader behavioral data for customer understanding.
- Hossain et al. (2023)[12] explore the role of technology in boosting e-commerce adoption for small and medium enterprises (SMEs). Analyzing ICT adoption, internet connectivity, and business data management, they findthese factors significantly promote e-commerce success. The paper calls for further research on organizational and environmental factors influencing SME e-commerce adoption, offering valuable insights for SMEs and policymakers alike on leveraging technology for growth in the digital marketplace.

3 Description of the Data

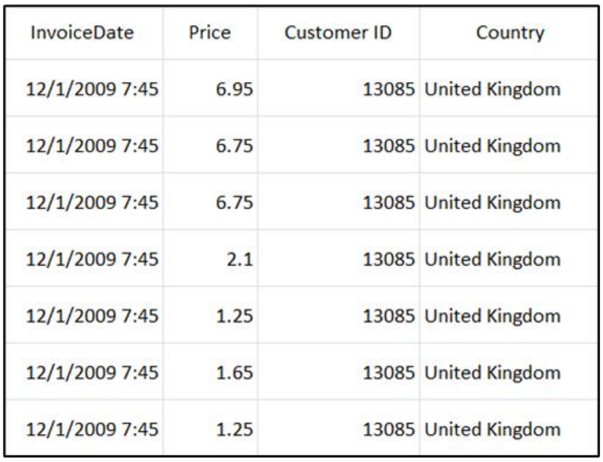
The initial dataset comprises eight columns providing comprehensive information on retail transactions as shownin fig.[1],[2]. The key features include:

- Invoice: Unique identification code assigned to each purchase transaction.
- Stock Code: Distinct 5-digit code associated with each product in the inventory.
- Description: Product name and brief description.
- Quantity: Number of units of the product involved in each transaction, accounting for both purchases and cancellations.
- Invoice Date: Date and time of the transaction recorded in ”mm/dd/yyyy HH:MM” format.
- Price: Cost or amount associated with the purchased items.
- Customer ID: Unique 5-digit identification number assigned to each customer.
- Country: Name of the country where the customer resides.
- The dataset is organized with records such as Invoice ID, Stock Code, Description, Quantity, Invoice Date, Price, Customer ID, and Country. This comprehensive dataset initially encompasses a vast and diverse collection of over one million records, offering valuable insights into retail transactions for further analysis and model training.



Invoice	StockCode	Description	Quantity
489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12
489434	79323P	PINK CHERRY LIGHTS	12
489434	79323W	WHITE CHERRY LIGHTS	12
489434	22041	RECORD FRAME 7" SINGLE SIZE	48
489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24
489434	22064	PINK DOUGHNUT TRINKET POT	24
489434	21871	SAVE THE PLANET MUG	24

Figure 1: Initial Dataset(a)



InvoiceDate	Price	Customer ID	Country
12/1/2009 7:45	6.95	13085	United Kingdom
12/1/2009 7:45	6.75	13085	United Kingdom
12/1/2009 7:45	6.75	13085	United Kingdom
12/1/2009 7:45	2.1	13085	United Kingdom
12/1/2009 7:45	1.25	13085	United Kingdom
12/1/2009 7:45	1.65	13085	United Kingdom
12/1/2009 7:45	1.25	13085	United Kingdom

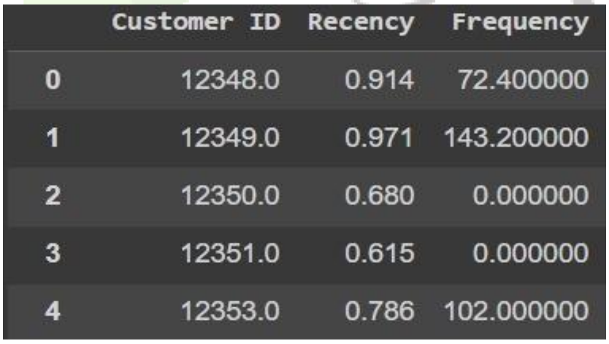
Figure 2: Initial Dataset(b)

The features in the sample dataset were the best fit for our methodology. We applied various algorithms and mathematical calculations to extract information and transform them into our desired feature set suitable for application in credit score calculation formula.

The final dataset after modification contains the following feature set as shown in fig.[3],[4]:

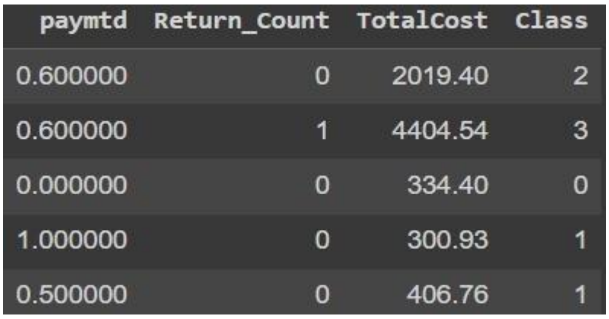
- Customer ID: A unique 5-digit identification number assigned to each customer.
- Recency: Represents how recently the customer made their last purchase, with values ranging between 0 and 1, where higher values indicate more recent transactions.
- Frequency: Reflects the frequency of customer purchases, indicating how often they engage in transactions.
- paymtd: Combines customer payment method (0 - cash on delivery / 1- prepayment or online payment) and normalizes the value in range [0,1]. It represents the number of times customer made transaction using online or prepayment.
- Return_Count: The total number of returns made by the customer.
- TotalCost: Represents the total amount spent by the customer in overall purchases.
- Class: Categorizes customers into different classes (0, 1, 2, 3) based on their calculated credit score (C_S). The credit score is derived from a formula considering recency, frequency, payment method, return count, and total cost.

Each row in the dataset corresponds to a unique customer, and the features provide valuable insights into their transactional behavior and creditworthiness. The ”Class” column categorizes customers into different groups, aiding in further analysis or targeted actions based on their credit scores.



	Customer ID	Recency	Frequency
0	12348.0	0.914	72.400000
1	12349.0	0.971	143.200000
2	12350.0	0.680	0.000000
3	12351.0	0.615	0.000000
4	12353.0	0.786	102.000000

Figure 3: Final Dataset(a)



paymtd	Return_Count	TotalCost	Class
0.600000	0	2019.40	2
0.600000	1	4404.54	3
0.000000	0	334.40	0
1.000000	0	300.93	1
0.500000	0	406.76	1

Figure 4: Final Dataset(b)

4 Data Analysis and Preparation Methodology

4.1 Data Preprocessing Methods

- Concatenation: The data from two different CSV files (d1.csv and d2.csv) is read into separate dataframes (d1 and d2). These dataframes are then concatenated along the rows using pd.concat, forming the combineddataframe df.
- Handling Missing Values: The df dataframe is checked for missing values using df.isnull().sum(). Rows containing missing values are removed with df = df.dropna().
- Date Conversion and Time-Based Feature Engineering: The 'InvoiceDate' column is converted to datetime format using pd.to_datetime. New features related to recency, last purchase date, purchase period, and frequency are created based on date information.

4.2 Feature Extraction Methods

- Grouping and Aggregation: Data is grouped by 'Customer ID', 'Invoice', 'DaysSinceLastPurchase', and 'Frequency'. Aggregation is performed on 'TotalCost' using sum().
- Return Count Calculation: The total return count for each customer is calculated and merged back into the original dataframe.
- Payment Method Frequency: 'paymtd' is calculated as the frequency of the payment method for each customer.
- Credit Score Calculation (c_s): Recency, TotalCost, Frequency, paymtd, and Return_Count are used to calculate the credit score ('c_s') for each customer.
- Categorization into Classes: 'Class' is derived by categorizing customers based on quantiles of the credit score ('c_s').

4.3 Outlier Detection and Removal

Z-Score and IQR Methods: Z-Score and IQR methods are applied to identify outliers in columns: 'Recency', 'Frequency', 'paymtd', 'Return_Count', 'TotalCost'. Rows identified as outliers are removed from the dataframe tocreate the final processed dataframe df4.

5 Experimental Setup

In our pursuit of developing an effective customer credit scoring system for e-commerce, we embarked on a comprehensive experimental setup involving various machine learning models. In our experimental setup, we applied a variety of machine learning models to predict customer credit scores in the context of an e-commerce platform. The dataset, initially preprocessed and transformed into a refined form (df4), was subjected to several classification algorithms, including k-Nearest Neighbors (KNN), Decision Tree, Support Vector Machine (SVM), Random Forest, and Random Forest with Polynomial Features. The features used for prediction included Recency, Frequency, TotalCost, paymtd, and Return_Count. To evaluate the performance of each model, we employed metrics such as accuracy, precision, recall, F1 score, and Cohen's Kappa.

6 Results

- KNN Classifier (with outliers): The KNN classifier demonstrated an accuracy of 80.15%. Precision, recall, F1score, and Cohen's Kappa were observed to be 80.22%, 80.15%, 80.18%, and 72.04%, respectively.
- Decision Tree Classifier (with outliers): The Decision Tree model exhibited superior performance with an accuracy of 92.26%. Precision, recall, F1 score, and Cohen's Kappa were recorded at 92.29%, 92.26%, 92.27%, and 89.10%, respectively.
- SVM (with outliers): The SVM model achieved an accuracy of 87.30%, along with precision, recall, F1 score,and Cohen's Kappa scores of 87.54%, 87.30%, 87.29%, and 81.93%, respectively.
- Random Forest (with outliers): The Random Forest model yielded an accuracy of 94.62%, showcasing robust performance. Precision, recall, F1 score, and Cohen's Kappa stood at 94.70%, 94.62%, 94.63%, and 92.39%, respectively.
- KNN Classifier (without outliers): After removing outliers, the KNN classifier's accuracy decreased to 78.74%. Precision, recall, F1 score, and Cohen's Kappa were observed at 78.92%, 78.74%, 78.78%, and 66.90%, respectively.

- Decision Tree Classifier (without outliers):The Decision Tree model, sans outliers, maintained a high accuracy of 92.52%. Precision, recall, F1 score, and Cohen's Kappa were reported at 92.59%, 92.52%, 92.53%, and 88.33%, respectively.
- Random Forest (without outliers): The Random Forest model, excluding outliers, demonstrated an accuracy of 95.44%. Precision, recall, F1 score, and Cohen's Kappa reached 95.45%, 95.44%, 95.44%, and 92.83%, respectively.
- SVM (without outliers): Without outliers, the SVM model's accuracy significantly dropped to 41.46%. Preci-sion, recall, F1 score, and Cohen's Kappa showed scores of 31.33%, 41.46%, 24.86%, and 1.08%, respectively.
- Random Forest with Polynomial Features: With the incorporation of Polynomial Features, the Random Forest model achieved an outstanding accuracy of 96.64%. Precision, recall, F1 score, and Cohen's Kappa were measured at 96.64%, 96.64%, 96.64%, and 94.72%, respectively.

7 Conclusion and Future Work

In conclusion, our research on predicting customer credit scores in the e-commerce domain reveals the effective-ness of diverse machine learning models. The Decision Tree and Random Forest models consistently perform well, showcasing their suitability for credit scoring tasks with accuracies of 92.26% and 94.62%, respectively. The SVM model's effectiveness with outliers highlights the importance of data preprocessing, while the introduction of Polynomial Features in the Random Forest model demonstrates the potential benefits of feature engineering, resultingin an impressive accuracy of 96.64%. Sensitivity to outliers observed in the KNN classifier underscores the need for careful model selection and understanding the dataset's nature. Our findings emphasize the significance of model selection and data preprocessing in developing reliable credit scoring systems for e-commerce platforms. Future work should explore advanced feature engineering and ensemble models to further enhance robustness and adaptability in dynamic e-commerce environments.

References

[1] Maoguang Wang and Hang Yang. Research on customer credit scoring model based on bank credit card. In *Intelligent Information Processing X: 11th IFIP TC 12 International Conference, IIP 2020, Hangzhou, China, July 3–6, 2020, Proceedings 11*, pages 232–243. Springer, 2020.

[2] Haichao Zhang, Ruishuang Zeng, Linling Chen, and Shangfeng Zhang. Research on personal credit scoring model based on multi-source data. In *Journal of Physics: Conference Series*, volume 1437, page 012053. IOP Publishing, 2020.

[3] Kirti Maheshwari, Ria Khapekar, Anmol Bahl, and Kunal Bhatia. Credit profile of e-commerce customer. 2019.

[4] Zengyuan Wu, Lingmin Jin, Jiali Zhao, Lizheng Jing, and Liang Chen. Research on segmenting e-commerce customer through an improved k-medoids clustering algorithm. *Computational Intelligence and Neuroscience*, 2022, 2022.

[5] Aslıhan Dursun and Meltem Caber. Using data mining techniques for profiling profitable hotel customers: An application of rfm analysis. *Tourism management perspectives*, 18:153–160, 2016.

[6] A Joy Christy, A Umamakeswari, L Priyatharsini, and A Neyaa. Rfm ranking–an effective approach to customer segmentation. *Journal of King Saud University-Computer and Information Sciences*, 33(10):1251–1257, 2021.

[7] Onur Dogan, Ejder Ayçin, and Zeki Bulut. Customer segmentation by using rfm model and clustering methods:a case study in retail industry. *International Journal of Contemporary Economics and Administrative Sciences*,8, 2018.

[8] Ching-Hsue Cheng and You-Shyang Chen. Classifying the segmentation of customer value via rfm model and rs theory. *Expert systems with applications*, 36(3):4176–4184, 2009.

[9] RW Sembiring Brahmana, Fahd Agodzo Mohammed, and K Chairuang. Customer segmentation based on rfm model using k-means, k-medoids, and dbscan methods. *Lontar Komput. J. Ilm. Teknol. Inf*, 11(1):32, 2020.

[10] İnanç Kabasakal. Customer segmentation based on recency frequency monetary model: A case study in e-retailing. *Bilişim Teknolojileri Dergisi*, 13(1):47–56, 2020.

[11] Uus Firdaus and D Utama. development of bank's customer segmentation model based on rfm+ b approach. *Int. J. Innov. Comput. Inf. Cont*, 12(1):17–26, 2021.

[12] Md Billal Hossain, Nargis Dewan, Aslan Amat Senin, and Csaba Balint Illes. Evaluating the utilization of technological factors to promote e-commerce adoption in small and medium enterprises. *Electronic Commerce Research*, pages 1–20, 2023.