Corpus Intellectual
ISSN PRINT 2811-3187 ONLINE 2811-3209 Volume 3 NO 3 2024 Conf. Edition

Development Of A Machine Learning Model For Brand And Audience Segmentation Using Demographic Data

Lead Author

Gbeminiyi Falowo

Affiliation:

Department of Mass communication Redeemer's University Ede, Osun State

Abstract

The expansion of the global business landscape, a high-impact factor in eCommerce, has resulted in identifying potential customers and their positive reactions to products or services offered by companies that use the internet to promote their electronic business. With a high increase in audience using social media, there is a need for brand and audience segmentation and targeting for profit-making; thus, this study developed a machine learning model for brand and audience segmentation using the Social Media Advertising Dataset. The dataset includes comprehensive data on social media advertising campaigns across Facebook, Instagram, Pinterest, and Twitter, featuring ad impressions, clicks, spending, demographic targeting, and conversion rates. With 16 columns and 300,000 rows, the dataset offered substantial data for analysis. The study compared the performance of a Naive Bayes model with a Random Forest algorithm in two existing literature; the Naive Bayes model achieved an accuracy of 35%, the Random Forest model achieved an accuracy of 89.6%, and the Random Forest model in the current study's model reached 97% accuracy. The Random Forest model's superior performance in both studies demonstrates its effectiveness in consumer group segmentation, indicating its practical utility in optimizing marketing strategies and improving customer targeting. An implementation of the developed model of the study was in Python and deployed on a website using the Flask framework, providing an accessible tool for practical applications.

Co-Authors: Blessing Oluwatobi Olorunfemi; Inwang Emmanuel Inwang Computer Science Department, Redeemer's University Ede, Osun State.

Keywords: Brand Segmentation, Audience Segmentation, Machine Learning, Demographic Data, Social Media Advertising, Naive Bayes, Random Forest, Marketing Strategies, Customer Targeting

1. Introduction

Customer data forms the foundation for successful business strategies. Exploring data to uncover customer insights and support decision-making enhances business interest. Rather than applying marketing strategies uniformly to all customers, clustering customers allows businesses to identify target segments, enabling a deeper understanding of each segment's characteristics and the development of tailored business strategies (Dawane et al., 2021). Consequently, applying clustering methods to identify potential customers is a leading trend in today's tech space. Combining machine learning (ML) algorithms with user data exemplifies customer segmentation and supports businesses in identifying segments of customers that are difficult to detect through intuition and manual information inspection (Kumar, 2023). The combination of these ML models further results in market segmentation, which is the dividing of a market into distinct sub-groups of customers with different needs, characteristics, or behaviours who may require separate products or respond differently to various marketing efforts (Durojaye & Obunadike, 2022). In today's business landscape, companies face the challenge of identifying potential customers most likely to respond positively to a product or offer. Here, data mining techniques become crucial. With the growing amount of available data, data mining has become essential for direct marketing efforts, enabling companies to create prediction response models based on past client purchase data (Kasem et al., 2023). Companies must understand client demands and provide tailored products and services to secure ample profits. This understanding can be achieved through segmentation via machine learning. Applying the right marketing tactics to the correct customer segments increases the probability of profit maximization and enhances cost efficiency by avoiding the expenditure of resources on unlikely customer bases (Yadegaridehkordi et al., 2021). Demographic data such as Gender, age, familial and marital status, income, education, occupation, and geographical information are crucial for segmentation. Depending on the company's scope, this geographical information could range from specific towns or counties

to broader regions such as cities, states, or countries (Thalkar, 2021). In this context, machine learning, a subfield of artificial intelligence, is focused on developing algorithms and techniques that enable computers to learn from data (Kasem et al., 2023) and make essential predictions, especially for marketing products. Despite several studies on the role of intelligence in marketing (Boisena et al., 2018), AI-driven segmentation to predict customer behaviours remains underexplored. The segmentation in this research specifically focuses on social media platforms, enabling brands and audiences to be categorized into their desired categories for more targeted marketing efforts.

2. Review Of Literature

The dynamic interactions between brands and audiences within social media are critical in forming digital marketing strategies. As Zhang and Daugherty (2018) identified, electronic businesses (eBusinesses) utilize social media platforms to communicate with their target audience, advertise products or services, and establish brand awareness. However, they only focus on Pinterest, which may not generalize to other social media platforms. Consequently, brands (eBusinesses) that want to customize their marketing strategies and cultivate brand loyalty must comprehend their target audience's varied demographics, preferences, and behaviours (Suryakanthan et al., 2024). The segments in Suryakanthan et al. (2024) provided valuable insights for tailored marketing campaigns, product suggestions, and enhanced customer experiences but were limited to K-means clustering.

On the other hand, social media platform audiences are made up of a wide range of people with different interests, demography, and levels of involvement (Nguyen, 2021). Furthermore, Amutha and Khan (2023) stated that through shares, clicks, comments, and other types of engagement, audiences actively consume information, engage with companies, and add to the online conversation. By implication, from the statement above, audiences look for real connections, pertinent material, and tailored experiences from brands on social media. In Zote (2024), an audience can be divided into groups of various interests, such as pastimes and activities. This enables the dissemination of messages primarily to relevant audiences. Thus, Kubade et al. (2023) combined and compared the analysis of the Support Vector Machine, Random Forest algorithm, and KNN model for audience segmentation, with Random Forest outperforming the others in accuracy, yet the proposed model in this study had a better performance as projected by Sruthi (2024).

3. Methodology

This section details the methodology for developing the Machine Learning Model for Brand and Audience segmentation using demographic data. It comprises data collection, pre-processing, feature engineering, model selection, system design, and evaluation metrics.

3.1 Method of Data Collection

This study uses the Social Media Advertising Dataset as obtained from the Kaggle Repository. The file in the dataset is named Social_Media_Advertising.csv. The dataset's access link https://www.kaggle.com/datasets/jsonk11/social-media-advertising-dataset

3.1.1 Data Description

The Social Media Advertising dataset is a comprehensive collection of data related to various social media advertising campaigns. It includes ad impressions, clicks, spending, demographic targeting, and conversion rates. The dataset encompasses multiple social media platforms such as Facebook, Instagram, Pinterest, and Twitter, providing diverse advertising campaign data. This dataset contains 16 columns and 300,000 rows, offering substantial data for analysis. Table 3.1 explains the data attributes.

Table 3.1: Description of the Dataset

S/N	Attribute Name	Attribute Description
1	Campaign_I.D.	A unique identifier for each advertising campaign.
2	Target Audience	The specific demographic or audience segment targeted by the ad campaign.
3	Campaign Goal	The main objective of the campaign (e.g., brand awareness, lead generation, sales conversion).
4	Duration	The length of time the ad campaign ran was typically measured in days.

5	Channel Used	The social media platform where the ad was displayed (e.g., Facebook, Instagram, Pinterest, Twitter).
6	Conversion Rate	The percentage of persons who finished a desired action after engaging with the ad.
7	Acquisition Cost	The cost incurred to acquire a customer through the ad campaign.
8	ROI	Return on Investment: a way to measure the ad campaign for profit making.
9	Location	The geographical region targeted by the ad campaign.
10	Language	The language used in the ad campaign.
11	Clicks	The actual number of times users press the mouse button on the ad.
12	Impressions	The number of times the ad was shown to users.
13	Engagement Score	A metric indicating user engagement with the ad (e.g., likes, shares, comments).
14	Customer Segment	The segment of customers targeted by the ad campaign; this serves as the target or label for analysis.
15	Date	The date when the ad campaign was run.
16	Company	The company or brand running the ad campaign.

3.2 System Architecture

The model employs a Random Forest classifier, a powerful ensemble learning method that builds multiple decision trees and merges their predictions to improve accuracy and control overfitting. The dataset is divided into training and testing sets, with 80% used for training the Random Forest classifier and 20% reserved for testing. The training process is conducted using Jupyter, which allows for rapid iterations and model improvements.

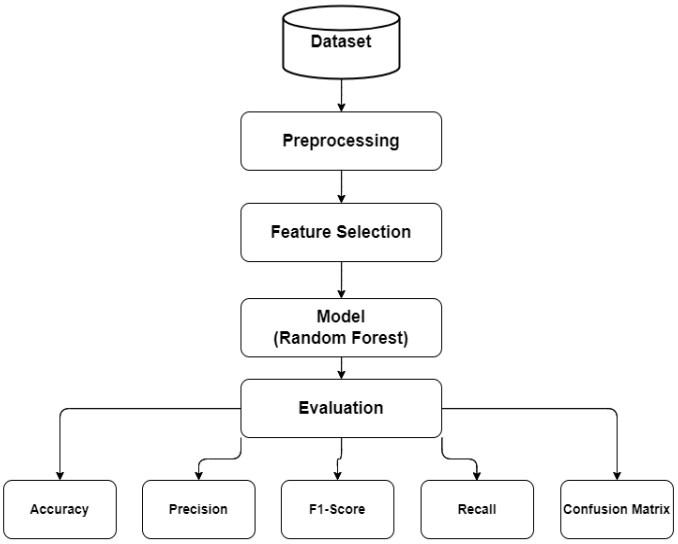


Figure 3.1. System Architecture

3.2 System Design

The flowchart, as depicted in Figure 3.3 below, illustrates the logical flow and relationships between various components, which helps to comprehend the design of the System, spot possible bottlenecks, and make sure all required processes are taken into account.

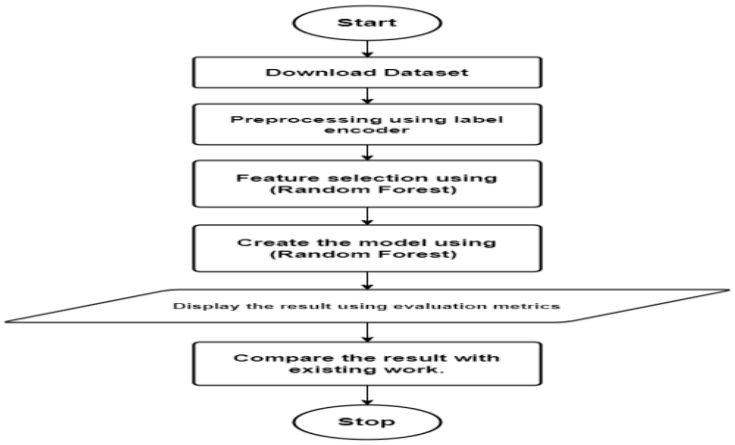


Figure 3.2. System Flowchart

Also, as seen in Figure 3.3, the use case diagram for "Customer Segment Prediction" depicts the interactions between various actors: User, System, and Admin and the System to forecast customer segments. The process begins with the User inputting customer details into the System. Once the input is provided, the User submits the information, initiating the prediction process. The System then takes over, processing the submitted details to predict the customer segment. After the prediction is made, the System displays the prediction to the User, allowing them to view the results.

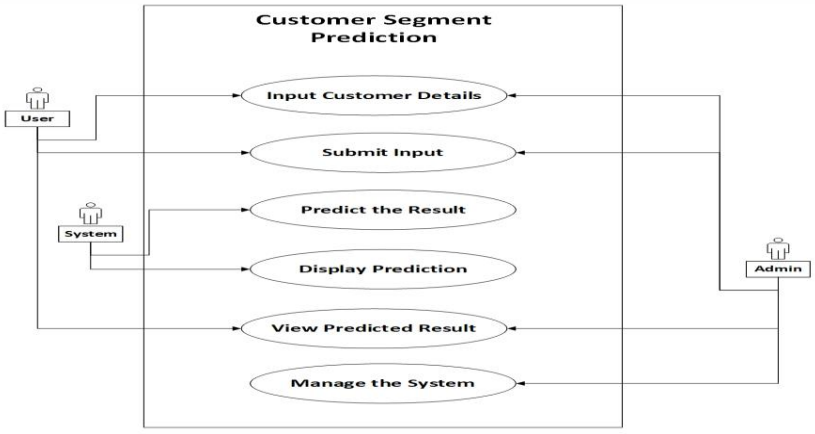


Figure 3.4. System Use Case Design

System Implementation

Several tools and platforms were used to design the model. Python was employed for data collection and pre-processing, with libraries like Pandas and NumPy utilized for numerical operations and data manipulation. Matplotlib and Seaborn were used for exploratory data analysis and visualization. Scikit-learn, a powerful package that efficiently implements numerous techniques, is used to build the machine learning model. Scikit-learn's pre-processing module will handle missing values and feature scaling, ensuring that the data is clean and prepared for analysis. All missing data were filled in using imputation techniques, and feature engineering was performed to enhance the dataset's informative value. The model training and evaluation were conducted using Jupyter Notebook, which provides an interactive environment ideal for data exploration, visualization, and iterative model development. Additionally, evaluation metrics such as accuracy, precision, and recall were calculated using Scikit-learn's metrics module to comprehensively assess the Random Forest model's performance.

4.1 Dataset Pre-processing and Analysis

In the initial phase of the study, as illustrated in Figure 4.1, the necessary library, pandas, was imported for data manipulation and analysis in Python. The path to the CSV file containing the dataset was specified, and the dataset was stored on the System. The dataset, named "Social_Media_Advertising.csv," was read into a panda Data Frame using the pd.read_csv function, which allowed the data to be loaded into a structured format suitable for analysis. To verify that the data was loaded correctly and to gain an initial understanding of its structure, the first five rows of the Data Frame were displayed using the data.head() function.

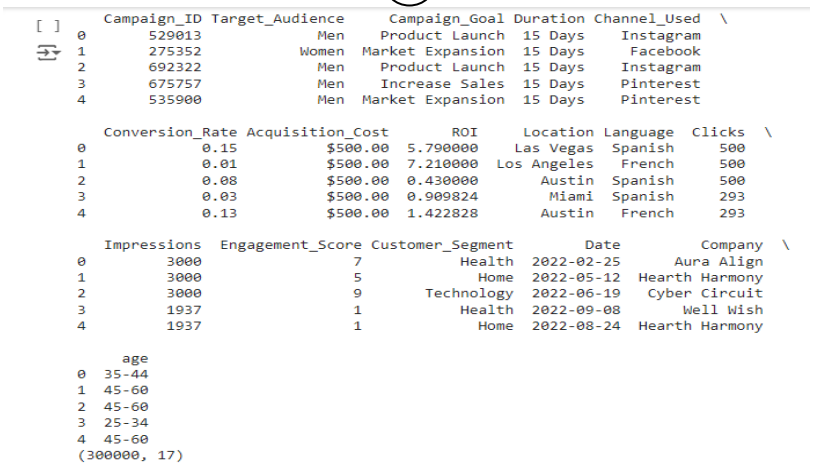
	Campaign_ID	Target_Audience	Campaign_Goal	Duration	Channel_Used	\
0	529013	Men 35-44	Product Launch	15 Days	Instagram	
1	275352	Women 45-60	Market Expansion	15 Days	Facebook	
2	692322	Men 45-60	Product Launch	15 Days	Instagram	
3	675757	Men 25-34	Increase Sales	15 Days	Pinterest	
4	535900	Men 45-60	Market Expansion	15 Days	Pinterest	

	Conversion_Rate	Acquisition_Cost	ROI	Location	Language	Clicks	\
0	0.15	$500.00	5.790000	Las Vegas	Spanish	500	
1	0.01	$500.00	7.210000	Los Angeles	French	500	
2	0.08	$500.00	0.430000	Austin	Spanish	500	
3	0.03	$500.00	0.909824	Miami	Spanish	293	
4	0.13	$500.00	1.422828	Austin	French	293	

	Impressions	Engagement_Score	Customer_Segment	Date	Company
0	3000	7	Health	2022-02-25	Aura Align
1	3000	5	Home	2022-05-12	Hearth Harmony
2	3000	9	Technology	2022-06-19	Cyber Circuit
3	1937	1	Health	2022-09-08	Well Wish
4	1937	1	Home	2022-08-24	Hearth Harmony

Figure 4. 1: Overview of the Dataset First Rows.

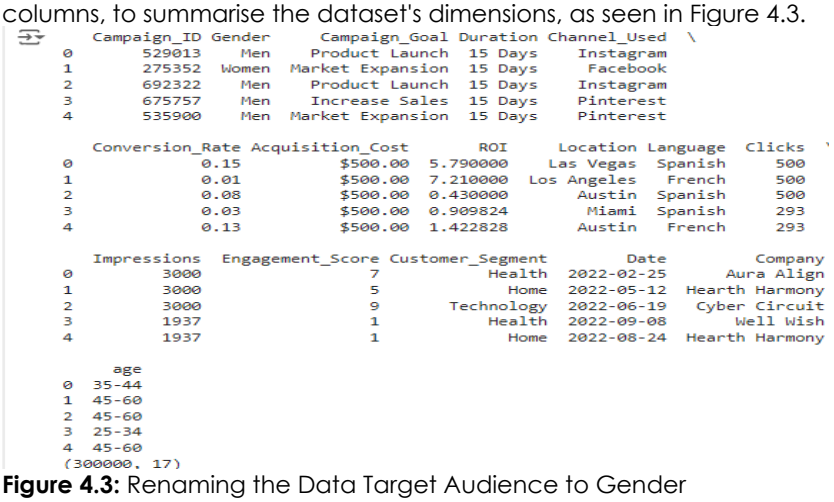
As the next step in data pre-processing and analysis, the target audience information was split into two columns using the first space in the string. Specifically, the Target Audience column was divided at the first space, and the portion following the first space was extracted into a new column named age. This was achieved by using the str.split(' ') method and selecting the elements after the first split. To ensure that the age data was recorded correctly, the extracted list of strings for the age column was then transformed back into a single string using the apply method using a lambda function. Concurrently, the initial Target Audience column was modified to preserve solely the initial segment of the string, so the target audience data is divided into two distinct and significant columns for additional examination, as illustrated in Figure 4.2.



```
[ ]    Campaign_ID Target_Audience     Campaign_Goal Duration Channel_Used  \
0       529013            Men    Product Launch  15 Days     Instagram
1       275352          Women  Market Expansion  15 Days      Facebook
2       692322            Men    Product Launch  15 Days     Instagram
3       675757            Men    Increase Sales  15 Days     Pinterest
4       535900            Men  Market Expansion  15 Days     Pinterest

   Conversion_Rate Acquisition_Cost       ROI     Location Language  Clicks  \
0             0.15          $500.00  5.790000    Las Vegas  Spanish     500
1             0.01          $500.00  7.210000  Los Angeles   French     500
2             0.08          $500.00  0.430000       Austin  Spanish     500
3             0.03          $500.00  0.909824        Miami  Spanish     293
4             0.13          $500.00  1.422828       Austin   French     293

   Impressions  Engagement_Score Customer_Segment        Date         Company  \
0         3000                 7           Health  2022-02-25      Aura Align
1         3000                 5             Home  2022-05-12  Hearth Harmony
2         3000                 9       Technology  2022-06-19   Cyber Circuit
3         1937                 1           Health  2022-09-08       Well Wish
4         1937                 1             Home  2022-08-24  Hearth Harmony

     age
0  35-44
1  45-60
2  45-60
3  25-34
4  45-60
(300000, 17)
```

Figure 4.2: Overview of the Data after Splitting

In this step of data pre-processing, the column name Target Audience was changed to Gender to reflect better the data it represents. The original column name and the new column name were specified in a dictionary format utilizing the rename method from pandas to achieve this. The inplace=True option made sure that the modifications were applied to the DataFrame without requiring its creation. Print(data.head()) was used to show the first five rows of the revised DataFrame following the column renaming in order to verify the modifications. In addition, print(data.shape) was used to display the DataFrame's shape, which contains the number of rows and columns, to summarise the dataset's dimensions, as seen in Figure 4.3.



```
   Campaign_ID Gender     Campaign_Goal Duration Channel_Used  \
0       529013    Men    Product Launch  15 Days     Instagram
1       275352  Women  Market Expansion  15 Days      Facebook
2       692322    Men    Product Launch  15 Days     Instagram
3       675757    Men    Increase Sales  15 Days     Pinterest
4       535900    Men  Market Expansion  15 Days     Pinterest

   Conversion_Rate Acquisition_Cost       ROI     Location Language  Clicks
0             0.15          $500.00  5.790000    Las Vegas  Spanish     500
1             0.01          $500.00  7.210000  Los Angeles   French     500
2             0.08          $500.00  0.430000       Austin  Spanish     500
3             0.03          $500.00  0.909824        Miami  Spanish     293
4             0.13          $500.00  1.422828       Austin   French     293

   Impressions  Engagement_Score Customer_Segment        Date         Company
0         3000                 7           Health  2022-02-25      Aura Align
1         3000                 5             Home  2022-05-12  Hearth Harmony
2         3000                 9       Technology  2022-06-19   Cyber Circuit
3         1937                 1           Health  2022-09-08       Well Wish
4         1937                 1             Home  2022-08-24  Hearth Harmony

     age
0  35-44
1  45-60
2  45-60
3  25-34
4  45-60
(300000, 17)
```

Figure 4.3: Renaming the Data Target Audience to Gender

In this data pre-processing step, duplicated rows were removed from the DataFrame to ensure data integrity and accuracy. Using the inplace=True option, the drop duplicates method from pandas was used to apply the modifications directly to the DataFrame. Print(data.head()) was used to show the top five rows of the revised DataFrame to confirm the alterations after the duplicates were eliminated. In addition, when the duplicates were eliminated, print(data.shape) was used to print the DataFrame's shape, which shows the number of rows and columns, to summarise the dataset's dimensions.

Following this, specific columns deemed unnecessary for the analysis were dropped from the DataFrame. Using the drop technique and the columns parameter, the columns Campaign_ID, Acquisition_Cost, ROI, Duration, Date, and Campaign_Goal was eliminated. Once more, these modifications were applied straight to the DataFrame using inplace=True. Print(data.head()) was used to show the first five rows of the revised DataFrame to verify that the designated columns had been removed. As seen in Figure 4.4, the shape of the DataFrame was printed again using print(data.shape) to display the updated dataset dimensions following the removal of redundant columns.

```
  Gender Channel_Used  Conversion_Rate     Location Language  Clicks  \
0    Men    Instagram             0.15    Las Vegas  Spanish     500
1  Women     Facebook             0.01  Los Angeles   French     500
2    Men    Instagram             0.08       Austin  Spanish     500
3    Men    Pinterest             0.03        Miami  Spanish     293
4    Men    Pinterest             0.13       Austin   French     293

   Impressions  Engagement_Score Customer_Segment         Company    age
0         3000                 7           Health      Aura Align  35-44
1         3000                 5             Home  Hearth Harmony  45-60
2         3000                 9       Technology   Cyber Circuit  45-60
3         1937                 1           Health       Well Wish  25-34
4         1937                 1             Home  Hearth Harmony  45-60
(300000, 11)
```

Figure 4.4: Dataset after Dropping some Columns

Finding the target variable and the feature set was the first step. Customer_Segment was declared as the goal variable, or the variable that has to be forecasted. The dataset's remaining columns were all regarded as features. A list of feature columns that did not include the target variable was produced to do this. Since machine learning models usually require numerical input, label encoding was used to translate categorical variables into numerical values. The LabelEncoder from the sklearn.pre-processing module was used for this procedure. Gender, Channel_Used, Location, Language, Company, and Age were the categorical columns that were

recommended for encoding. Every categorical column was encoded iteratively after a LabelEncoder object was initialized. This process assigned a unique numerical value to each category within a column, transforming the categorical data into a format suitable for the Random Forest model. To make sure the target variable Customer_Segment was in the appropriate numerical format for model training, it was also encoded using the LabelEncoder if it was categorical. Next, the dataset was divided into testing and training sets. After separating the features (X) and target (y), the data was split into 80% training and 20% testing sets using the train_test_split method from sklearn.model_selection, with a random state of 42 for repeatability. Using the trained Random Forest model, feature importance was computed to determine each feature's importance in the model's predictions. After being extracted, the feature importances were saved in a DataFrame and sorted by importance.

The "Company" variable, which has a significantly higher importance score than other features, is the most important component impacting the model's predictions, according to the feature importance chart in Figure 4.5. After that, "Clicks" and "Impressions" both provide a significant but far less contribution than "Company." While features like "age," "language," "channel_used," and "gender" have little bearing on the model's predictions, features like "conversion_rate," "engagement_score," and "location" demonstrate considerable value. This implies that the business linked to the data points significantly influences the result, outweighing other factors in terms of predictive ability. When evaluating the dataset's applicability for machine learning tasks and making decisions on how best to handle class distributions during model development, this graphical representation in Figure 4.6 was essential.

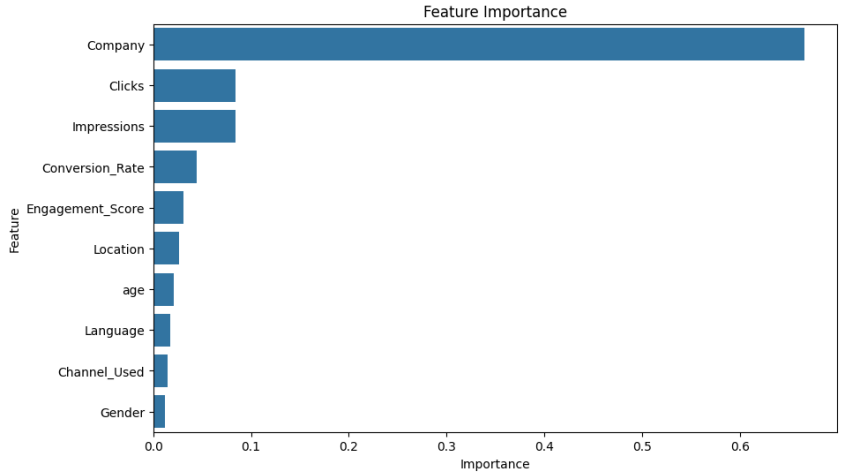


Figure 4.5: Feature Importance of the Dataset Class

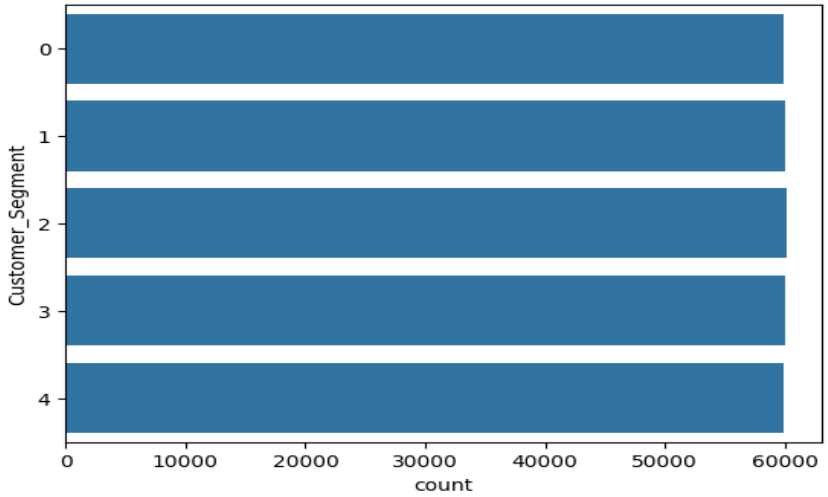


Figure 4.6: Description of the Balanced Training Dataset Class

4.2 Implementation

The described implementation presents a streamlined approach to predicting customer segments using a trained Random Forest Classifier through an interactive interface built with widgets. Users can input key features influencing predictions, such as categorical features (Gender, Channel_Used, Location, Language, Company, and age) through dropdown menus and numerical features (Conversion_Rate,

Clicks, Impressions, and Engagement_Score) via sliders. This setup allows for dynamic adjustment of input values within specified ranges. The process begins with data pre-processing, where categorical inputs are transformed into numerical representations using pre-fitted LabelEncoder instances to ensure compatibility with the machine learning model. These inputs are then structured into a data frame and fed into the trained classifier, which predicts the customer segment based on learned patterns. The predicted segment is displayed immediately, providing users with quick insights. This interactive approach is valuable for businesses needing rapid experimentation and scenario analysis to tailor marketing strategies, optimize resources, and enhance customer engagement, thereby improving decision-making processes through actionable insights derived from predictive analytics, as shown in Figure 4.7. and 4.8, respectively.

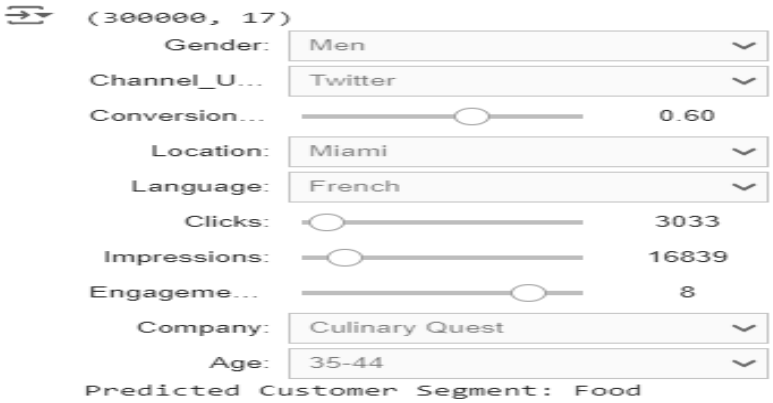


Figure 4.7: The Model Implementation on Python.

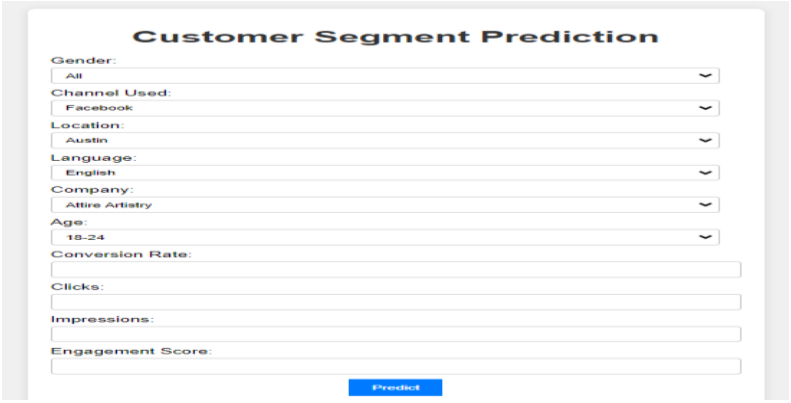


Figure 4. 8: Description of the Home Page

4.3 Testing

Testing the customer segmentation prediction system involves using test data to evaluate its performance. After the User inputs the test data, they can click on the submit button to initiate the prediction process. The description of some sample test data is presented in Figure 4.9.

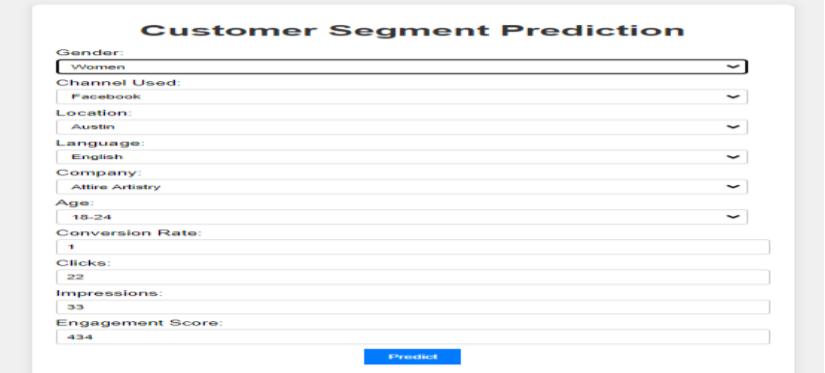


Figure 4.9: Testing the system

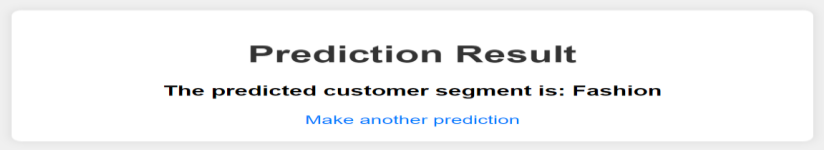


Figure 4.10: Testing Result

In the above figure, the result page displays the predicted outcome of the customer segmentation process. After the User inputs the test data and clicks the predict button, the System processes the information and presents the results.

Results

The achieved metrics accuracy, precision, recall, and F1 Score all stand at 0.97, highlighting the classifier's consistent capacity to generate correct predictions across a range of assessment criteria, as shown in Figure 5.1. Accuracy served as a foundational metric, indicating that 97% of the classifier's predictions align correctly with the actual customer segments in the test data. Precision, which measured the proportion of correctly predicted positive instances

(customer segments) out of all the cases predicted as positive, stands at 0.97. This signifies that the classifier maintains a high precision rate, minimizing false positives and ensuring the identified customer segments are reliably accurate. Recall that quantifying the proportion of correctly predicted positive instances out of all actual positive instances also stands at 0.97. This indicates the classifier's ability to effectively capture most of the true positive customer segments in the dataset. F1 Score, a combined metric of precision and recall, further reinforces the classifier's strong performance with a score of 0.97. This harmonic mean reflects a balanced assessment of the model's predictive power, highlighting its ability to identify positive instances and avoid misclassifications accurately.

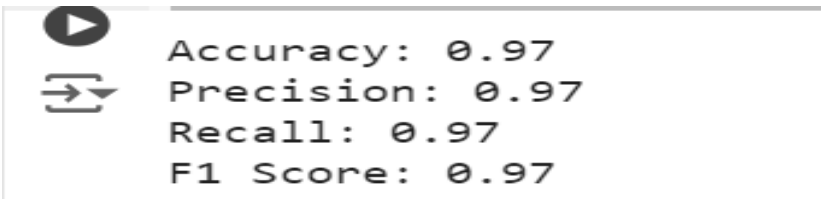


Figure 5.1: Random Forest Result.

According to the confusion matrix in Figure 5.2, the model demonstrated strong performance across all classes, with the highest number of correct predictions occurring along the diagonal. Specifically, the model correctly classified 11,769 instances of class 0, 11,591 instances of class 1, 11,834 instances of class 2, 11,772 instances of class 3, and 11,424 instances of class 4. There are relatively few misclassifications. For instance, class 0 had minor misclassifications into other classes, with the highest being 180 instances misclassified as class 1. Similarly, for class 1, the most notable misclassification was 270 instances classified as class 3. Overall, the matrix indicates that the model is effective at correctly predicting the majority of instances, with only a small number of instances being incorrectly classified. The confusion matrix of the Random Forest model is presented in Figure 5.2.

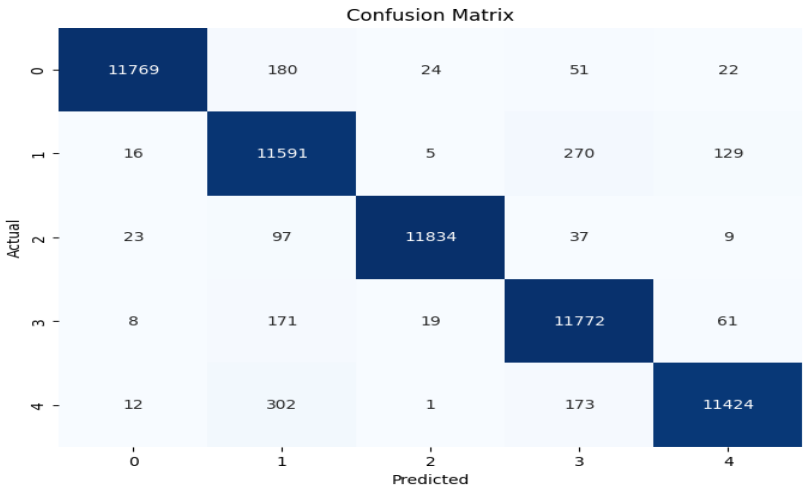


Figure 5.2: Random Forest Confusion Matrix.

The results for the Naive Bayes model were reported as follows: the accuracy was 0.35, indicating that the model correctly classified 35% of the instances. The precision was 0.41, meaning that 41% of the cases predicted as positive were positive. The recall was 0.35, showing the model identified 35% of the actual positive instances. Finally, the F1 score was 0.36, which is the harmonic mean of precision and recall; summarizing the balance between these two metrics and the confusion matrix is presented in Figure 5.3.

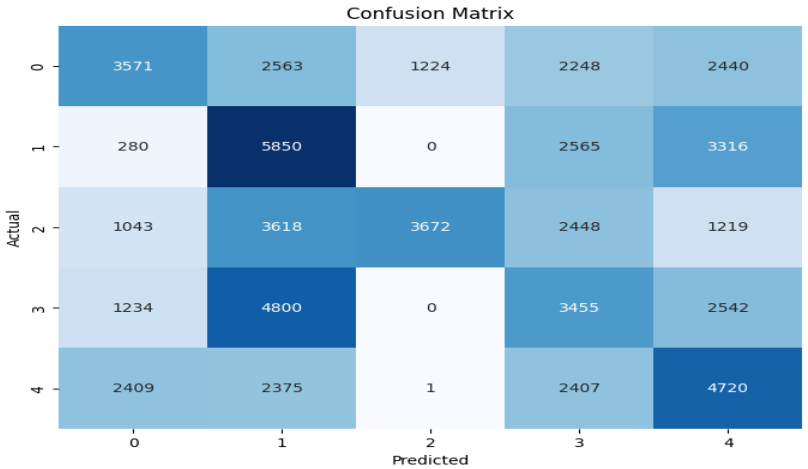


Figure 5.3: Naive Bayes Confusion Matrix.

Discussion

Comparing the Naive Bayes model results with the results achieved by the Random Forest algorithm in the study by Kubade et al. (2023) and the model from the current study highlights a significant performance gap, as shown in Table 4.1. The Naive Bayes model achieved an accuracy of 0.35, precision of 0.41, recall of 0.35, and an F1 score of 0.36. In contrast, the Random Forest algorithm in the study by Kubade et al. (2023) achieved an accuracy of 89.6% for customer segmentation, while the current study's model achieved an accuracy of 97%. The Naive Bayes model's accuracy of 35% was substantially lower than the Random Forest model's 89.6% and the current study's model's 97%. This suggests that the Naive Bayes model did not capture the underlying patterns and relationships within the data as effectively as the other two models. The significantly higher accuracy rates of the Random Forest model and the current study's model indicate its superior ability to forecast consumer groups, making them more favourable for practical applications in optimizing marketing strategies and enhancing customer targeting with greater confidence and precision.

Table 6.1 Comparison of the models' accuracy

S/N	Model	Accuracy
1	Naive Bayes	35%
3	Random Forest	97%

Table 6.2: Results Comparison with existing studies

S/N	Author(s)	Methods	Accuracy	Precision	Dataset	Overall Best Method
1	Kubade et al. (2023)	Support Vector Machine Random Forest Classifier K-Nearest Neighbour	75.3% 89.6% 83.2%	66.7% 67.5% 75.3%	E-commerce Customer Data	RF achieved the best result with the dataset

2	Current Study	Naive Bayes Random Forest	35% 97%	41% 97%	Social Media Advertising Dataset	RF achieved the best result with the dataset

Conclusion

In conclusion, developing and evaluating a Random Forest Classifier for customer segmentation demonstrated remarkable performance, achieving a 97% accuracy across various metrics, including Precision, Recall, and F1 Score. This performance was far better than Kubade et al. (2023) found in comparison research, where a model identical to this one attained an accuracy of 89.6%. Thus, the model's resilience and efficacy in forecasting client categories were emphasized. It was also observed that adding an interactive prediction interface improved the model's usefulness. This interface allows stakeholders to input and analyze critical criteria impacting consumer segmentation in real-time. As a result, companies can quickly make well-informed decisions, optimize marketing plans, and enhance consumer interaction techniques through trustworthy predictive insights.

References

Amutha, R., & Khan, A. A. (2023). Customer segmentation using machine learning techniques. Tuijin Jishu/Journal of Propulsion Technology, 44(3), 2051.

Boisena, M, K Terlouwb, P Grootea and O Couwenberga (2018). Reframing place promotion, place marketing, and place branding — moving beyond conceptual confusion. Cities, 80, 4–11

Dawane, V., Waghodekar, P., & Pagare, J. (2021). RFM Analysis Using K-Means Clustering to Improve Revenue and Customer Retention. In Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021). 1982-1989.

Durojaye, D. I., & Obunadike, G. N. (2022). Analysis and visualization of market segmentation in the banking sector using KMeans machine learning algorithm. FUDMA Journal of Sciences (FJS), 6(1), 387-393. https://doi.org/10.33003/fjs-2022-0601-910

Kasem, M. S., Hamada, M., & Taj-Eddin, I. (2023). Customer profiling, segmentation, and sales prediction using AI in direct marketing. arXiv. https://arxiv.org/abs/2302.01786

Kubade, H., Gharde, P. J., Fulbandhe, T. D., Pandey, A. A., Rehpade, K. S., & Hedaoo, S. R. (2023). Customer segmentation: Types of models and clustering techniques. International Journal of Advanced Research and Innovative Ideas in Education (IJARIIE), 9(2), 2706. https://ijariie.com

Kumar, A. (2023). Customer Segmentation of Shopping Mall Users Using K-Means Clustering. In Advancing SMEs Toward E-Commerce Policies for Sustainability (pp. 248-270). IGI Global

Nguyen, S. P. (2021). Deep customer segmentation with applications to a Vietnamese supermarket's data. Soft Computing, 25, 7785-7793. https://doi.org/10.1007/s00500-021-05796-0

Sruthi, E. R. (2024). Understand random forest algorithms with examples (updated 2024). Retrieved from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Suryakanthan, M., Vimal, K., Sanjay Raj, R., & Thiruselvan, P. M. E. (2024). Customer segmentation for enhancing business strategy. International Journal of Research Publication and Reviews, 5(1), 4735-4740. https://www.ijrpr.com

Thalkar, V. R. (2021). Customer segmentation using machine learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 7(6), 28-37. https://doi.org/10.32628/CSEIT217654

Yadegaridehkordi, E., Nilashi, M., Nasir, M. H. N. B. M., Momtazi, S., Samad, S., Supriyanto, E., & Ghabban, F. (2021). Customer segmentation in eco-friendly hotels using multi-criteria and machine learning techniques. Technology in Society, 65, 101528. https://doi.org/10.1016/j.techsoc.2021.101528

Zhang, Y., & Daugherty, T. (2018). Data-driven visual content marketing: Understanding consumer engagement through Pinterest. Journal of Retailing and Consumer Services, 43, 205-216. https://doi.org/10.1016/j.jretconser.2018.02.006

Zote, J. (2024). Social media target audience: How to find and engage yours. Retrieved from https://sproutsocial.com/insights/social-media-target-audience/