# Customer Segmentation

**Yitong Wang**

DSI, Brown University

GitHub[1]

## Introduction

Customer segmentation is the process of categorizing a company's customers into distinct groups based on various characteristics, such as age, gender, and marital status. This practice enables companies to tailor their marketing strategies to the specific needs and preferences of each segment, fostering customer attraction, retention, and ultimately, revenue generation[2].

The dataset[3], acquired from Kaggle and originated from an automobile company's hackathon, contains 8068 observations and 11 features (Table 1). The target variable defines four segments - A, B, C, and D - classified by the company's sales team. These segments correspond to distinct outreach and communication strategies. Prior models built on this dataset achieved an accuracy of approximately 50%[4] on the training data. For simplicity, the ID column has been excluded since its values are all unique.

| **Feature** | **Description** |
|---|---|
| ID | Unique ID |
| Gender | Gender of the customer |
| Ever_Married | Marital status of the customer |
| Age | Age of the customer |
| Graduated | Is the customer a graduate? |
| Profession | Profession of the customer |
| Work_Experience | Work Experience in years |
| Spending_Score | Spending score of the customer |
| Family_Size | Number of family members for the customer (including the customer) |
| Var_1 | Anonymised Category for the customer |
| Segmentation | (target) Customer Segment of the customer |

Table 1. Dataset features and descriptions

The primary objective of this project is to develop a robust multi-class classification model empowering the company to effectively perform customer segmentation for their new customers.

## EDA

The exploratory data analysis (EDA) serves as a foundational step in gaining insights into both the target variable and the relationships between features and the target variable. We start with exploring the target variable, we observe a well-balanced customer segmentation, where each class represents approximately 25% of the target variable (Figure 1).

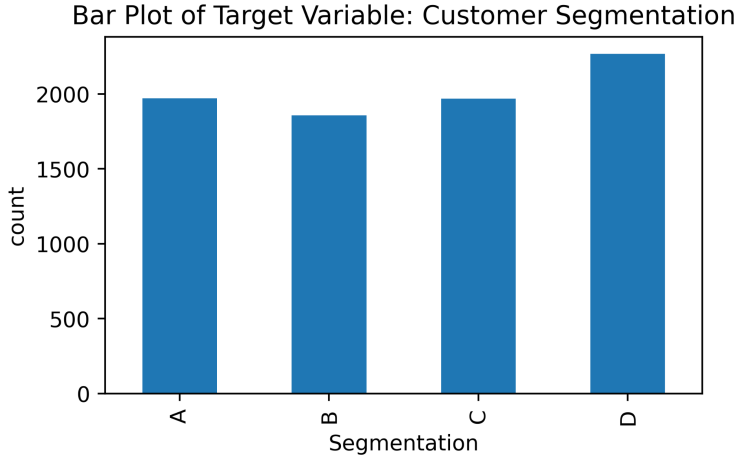


Figure 1. The 4 classes of the target variable are balanced

Moving on to explore the connection between individual features and the target variable, one interesting implication emerges: the company does not employ markedly distinct marketing strategies based on gender. Figure 2 illustrates that the four segments are proportionally distributed among both Female and Male categories, indicating that no specific segment is predominantly targeted towards either gender.

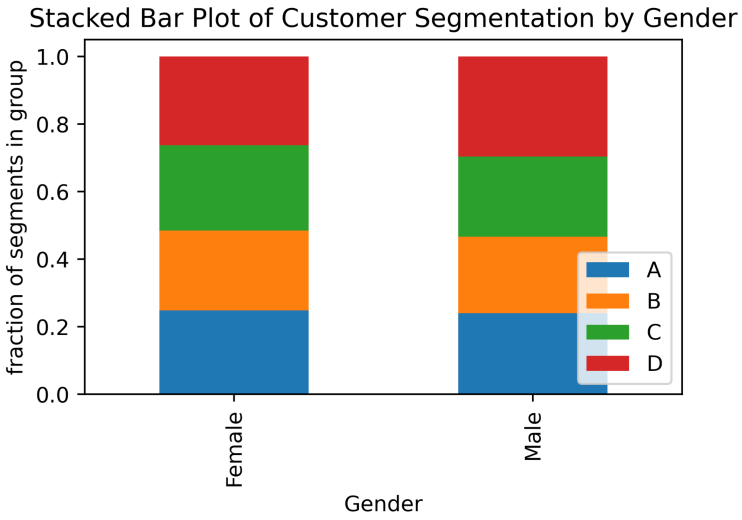


Figure 2. Stacked bar plot comparing the fraction of segments in Female and Male

Later we will show that age is the most important global feature of the model with the highest accuracy. Initial insights from the boxplot of age segmented by customer categories (Figure 3) reveal that Strategy D is particularly tailored to attract the relatively younger demographic within the company's customer base.

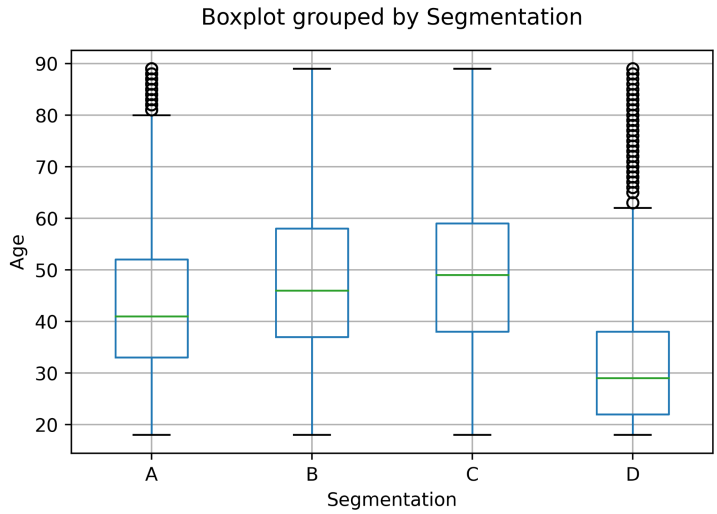


Figure 3. The 4 segments are strategies targeting customers of various ages. The y-axis represents the age of the customers, and the x-axis represents the 4 segments of the target variable.

## Methods

Given the modest size of the dataset and the balanced nature of the target variable, the classic splitting strategy is employed, allocating 60% of the data for training, 20% for validation, and 20% for testing. This is executed using the train_test_split method, with 80% of data points allocated to the (training+validation) set and the remaining 20% to the testing set. For Logistic Regression, Random Forest, and KNN models, a 4-fold KFold with GridSearchCV is applied to the (training+validation) set, as the data lacks distinct group structures. For XGBoost models, the data is split into 60% training, 20% validation, and 20% testing using train_test_split.

For Logistic Regression, Random Forest, and KNN models, a pipeline is constructed for continuous features, incorporating IterativeImputer for imputing missing values through Linear Regression. In contrast, XGBoost models handle missing values in continuous columns internally. Pipelines for categorical and ordinal features are developed, integrating SimpleImputer to treat missing values as a separate category, followed by OneHotEncoder and OrdinalEncoder, respectively. A final StandardScaler is applied to normalize coefficients for subsequent feature importance analysis. After data preprocessing, we pass in the algorithms as a parameter to our pipeline function to train the models (Table 2).

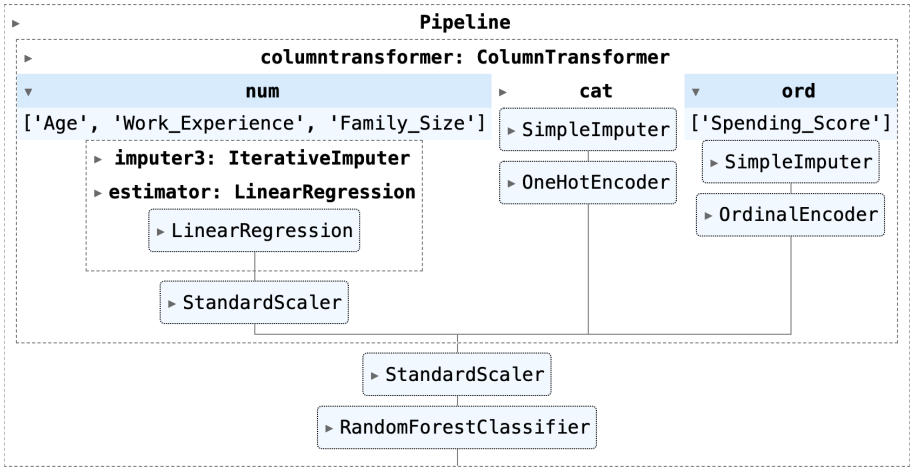


Table 2. CV pipeline (preprocessing) when passing RandomForestClassifier as an algorithm parameter

Four machine learning algorithms (Logistic Regression, Random Forest, KNN, and XGBoost) are implemented with GridSearchCV to execute cross-validation for hyperparameter tuning (Table 3).

| **ML algorithm** | **Parameters** | **Values** |
|---|---|---|
| Logistic Regression | C | [1e-3, 3.16e-3, 1e-2, 3.16e-2, 0.1, 0.316, 1, 3.16, 10] |
| Random Forest | n_estimators<br>max_depth<br>max_features<br>min_samples_leaf | [100, 150, 170]<br>[ 9, 10, 11]<br>[0.2, 0.3, 0.5]<br>[1, 10] |
| KNN | n_neighbors | [30, 50, 70, 100, 300] |
| XGBoost | max_depth<br>learning_rate<br>n_estimators | [3, 5, 7]<br>[0.01, 0.05, 0.1]<br>[80, 100, 120] |

Table 3. ML algorithms with corresponding tuned parameter values

To account for uncertainties arising from both splitting and non-deterministic ML methods such as random forest, training and testing are performed over 5 random states. The metric chosen for evaluating model performance is accuracy, given the balanced nature of the target variable, with a focus on true positives and true negatives. The impact of false positives and false negatives is considered less significant, as the specific misclassification of customers into segments holds lower priority.

## **Results**

All of the models perform better than the baseline accuracy of 0.28, with the Random Forest algorithm emerging as the most predictive, while the KNN algorithm performs relatively less well (Figure 4). Specifically, the mean accuracy of Logistic Regression exceeds the baseline by 27.769 standard deviations, Random Forest by 22.683 standard deviations, KNN by 15.528 standard deviations, and XGBoost by 24.946 standard deviations.

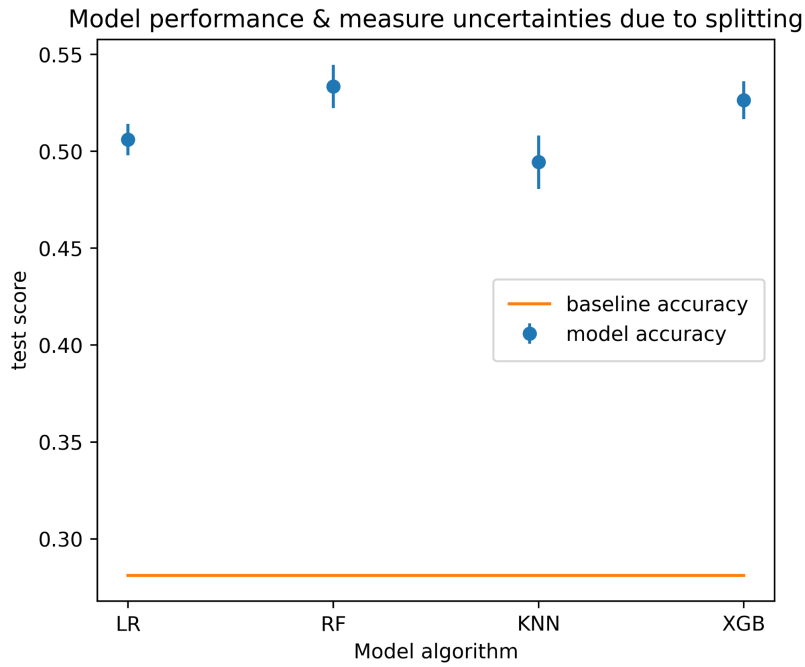


Figure 4. The RF algorithm outperforms the others

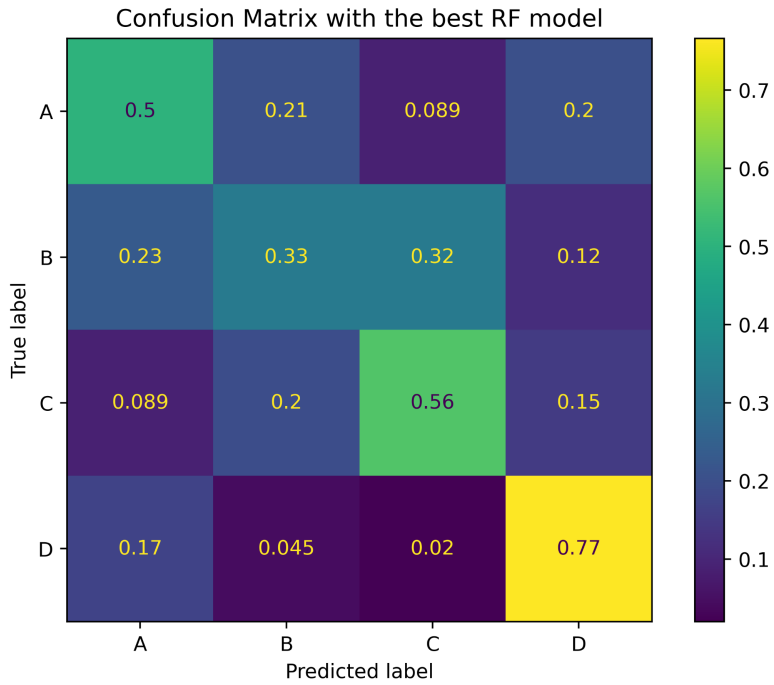


Figure 5. Normalized confusion matrix shows the highest accuracy when predicting class D

From the normalized confusion matrix of the best RF model, we can tell that the model identifies segment D with the highest accuracy (Figure 5). This may be attributed to the feature space of class D that differs the most from that of the other three classes. For instance, global feature importance plots consistently highlight Age as the most influential factor (Figure 6, 7, 8). Recall that the age range for strategy D differs the most from the other three classes, targeting relatively the youngest demographic of the automobile company's customers. Thus, class D's distinct feature space enables the model to identify segment D with the highest accuracy. On the other hand, the model struggles to distinguish class C from B, as their age ranges are more similar (Figure 3).

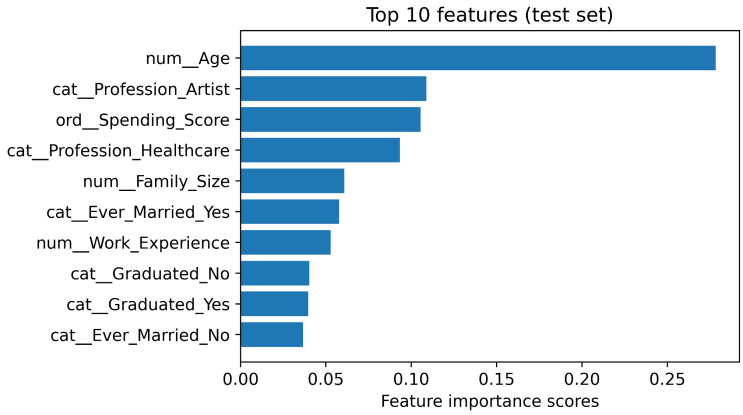


Figure 6. Top 10 feature importances of the best RF model according to RF's .feature_importances_: Age is the most important feature

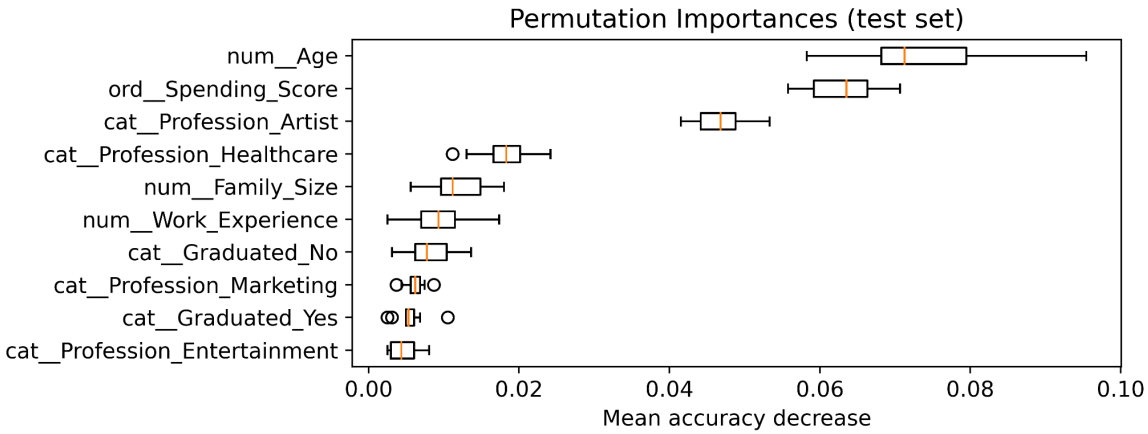


Figure 7. Top 10 features of the best RF model by permutation feature importance: Age is the most important feature; same set of top 5 features as figure 6 but slightly different rankings

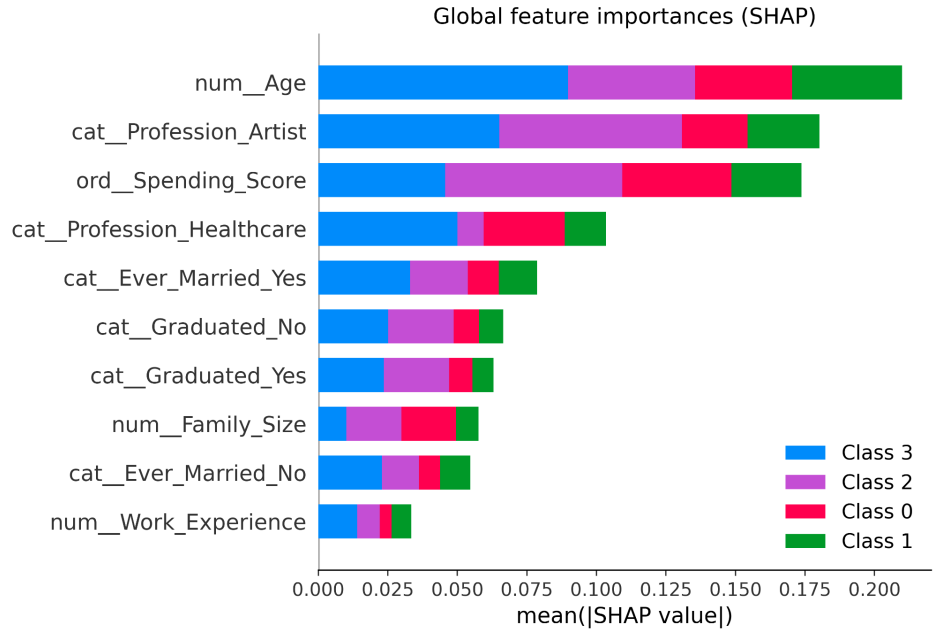


Figure 8. Top 10 feature importances of the best RF model by SHAP (Class 0 = Class A, Class 1 = Class B, , Class 2 = Class C, Class 3 = Class D): Age is the most important feature; Ever_Married_Yes appears for the first time in top 5 features

Beyond Age, Spending_Score consistently emerges among the top three global features. This aligns with industry norms where companies tailor outreach based on customers' spending history, providing VIP services to customers with higher spending scores and offering coupons to those with lower spending scores. Interestingly, Profession_Artist and Profession_Healthcare appear every time as the top four features, indicating the company employs unique marketing strategies targeting individuals in these professions. The features created from the missing values in the categorical features are the least important features.


Figure 9. SHAP local feature importance of the point at index 0, predicted probability for classes ABCD respectively (from top to bottom)

Figure 9 provides plots where features with arrows pointing to the right positively contribute to predictions, while those pointing to the left contribute negatively. The length of the arrow indicates the strength of the contribution. For the customer of index 0, the predicted probabilities are 0.57 for class A, 0.13 for class B, 0.07 for class C, and 0.21 for class D. The predicted class is A, aligning with the true label. Despite Age being the most crucial global feature, Spending_Score contributes the most to the predicted probability and the contribution is positive, it pushes the probability above the base value.

## Outlook

Exploring alternative strategies, a reduced-features model can be applied to address missing values in the continuous features to assess its impact on model performance. Moreover, incorporating feature interactions into the pipelines may enhance the accuracy. Beyond the existing tuned parameters, further hyperparameter tuning in Random Forest and XGBoost may unlock additional improvements in the accuracy of customer segmentation. Furthermore, an expanded dataset comprising more customer data could be collected to bolster model performance. Trying these approaches in the future may improve model performance in customer segmentation.

## References

[1] GitHub repository - https://github.com/Yitong001/DATA1030-project.git
[2] https://www.forbes.com/advisor/business/customer-segmentation/
[3] Kaggle - Janatahack : Customer Segmentation
[4]https://www.kaggle.com/datasets/vetrirah/customer/code?datasetId=848479&sortBy=voteCount