Comprehensive Report on Customer Segmentation Analysis

By : Kevin Marakana

Project Overview

In this practical, we will predict customer segments (A, B, C, or D) for potential new customers using machine learning classification models. An automobile company plans to expand into new markets with their existing products (P1-P5), and wants to classify 2,627 new potential customers into segments based on demographic and behavioral attributes.

1. Data Exploration Findings

1.1 Overview of the Dataset

- The dataset contains customer information such as demographics, spending behavior, and segmentation.
- Initial analysis of data types and structure is done using `df.info()`, `df.describe()`, and `df.head()`.

Dataset Information

The dataset contains various customer attributes:

Variable	Definition
ID	Unique ID of the customer
Gender	Gender of the customer
Ever_Married	Marital status of the customer
Age	Age of the customer
Graduated	Whether the customer is 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 (including customer)
Var_1	Anonymized category of the customer
Segmentation	Target Variable - Customer segment (A, B, C, or D)

1.2 Checking for Missing Values

- Missing values are identified using `df.isnull().sum()`.

- Replace Nan with Mean and Mode value .
- Percentage of missing data is calculated to assess the impact.

1.3 Data Summary

- `df.nunique()` is used to check the diversity of categorical variables.
- Distribution of numerical features is visualized using histograms.
- `df.describe()` is use for get overall summary about mean , mode , etc .
- `df.info()` is use for getting breif of dataset like datatype , not null value etc .
- `df.shape()` for geting shape of dataset.
- Different many more EDA's .

.

Project Tasks

Task	Description
1. Data Exploration	• Load the dataset and display basic statistics • Identify the number of records, missing values, and data types • Visualize class distribution of Segmentation
2. Handling Missing Values	• Identify missing values and impute or drop them as required • Use appropriate strategies such as mean/median for numerical data and mode for categorical data
3. Exploratory Data Analysis	Perform at least 5 analyses: • Univariate Analysis (Distribution plots for numerical features) • Bivariate Analysis (Comparison between variables) • Correlation Heatmap • Bar plot for categorical variables • Outlier detection using box plots
4. Model Building	• Convert categorical variables into numerical representations • Perform train-test split (e.g., 80-20 split) • Train at least four classification models: 1. Logistic Regression 2. Random Forest Classifier 3. Support Vector Machine (SVM) 4. XGBoost
5. Model Evaluation	Compare models using performance metrics: • Accuracy • Precision, Recall, and F1-Score • Confusion Matrix
6. Model Saving	• Save the best-performing model
7. Prediction	• Load the saved model and make predictions on the new dataset • Display the predicted segments for new customers

Let's begin our analysis!

---

2. Exploratory Data Analysis (EDA) Visualizations

2.1 Data Distributions

- Histograms are used to visualize distributions of `Age`, `Work_Experience`, and `Family_Size`.
- Box plots help detect outliers.

2.2 Categorical Feature Analysis

- Bar charts display distributions of `Gender`, `Ever_Married`, `Graduated`, `Profession`, `Spending_Score`, and `Var_1`.

2.3 Feature Relationships

- Pair plots visualize relationships between numerical variables.
- A correlation heatmap highlights dependencies among features.

2.4 Used Graph

- Boxplot
- Barchart
- Histogram
- Heatmap .

---

3. Handling Missing Values Strategy

- **Numerical Features:** Missing values are replaced using mean or median imputation.
- **Categorical Features:** Mode imputation is used for missing categorical values.
- **Column Removal:** Features with excessive missing data (>40%) are dropped.

---

4. Performance Evaluation of Models

4.1 Machine Learning Models Used

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)

4.2 Model Evaluation Metrics

- **Accuracy Score:** Measures overall correctness.
- **Precision & Recall:** Evaluates class-specific performance.
- **F1 Score:** Harmonic mean of precision and recall.
- **Confusion Matrix:** Visualizes classification results.

4.3 Chooesing Best model

- Choosing Best model from this all the model base on accuracy.

---

5. Predictions for New Customers

- First make new data according to save model
- Removing Nan value or replacing Nan Value with mean and mode.
- Convert Data into machine understandable format then it is feed to save model.
- The trained model is used to classify new customers.
- Predictions are mapped to respective segmentation labels.
- The results are saved as a CSV file for further use.

Conclusion

This report provides insights into customer demographics, spending behaviors, and segment distributions. The trained model successfully predicts customer segments, aiding in better marketing and business decisions.