Detail Project Report

Customer Personality Analysis

Revision Number – 1.0

# Last Date of Revision: 12 – 05 -2024

Harshal Zarikar

**Document Version Control**

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Description | Author |
| 12– 05- 2024 | 1.0 | First Draft | Harshal Zarikar |

## **Contents**

Document Version Control ……………………………………………………………………………….. 2

Abstract …………………………………………………………………………………………………... 4

1. Introduction …………………………………………………………………………………………………. 5
   1. Why this DPR Document ? ………………………………………………………………….. 5
2. General Description ……………………………………………………………………………………… 5
   1. Problem Perspective …………………………………………………………………………. 5
   2. Problem Statement …………………………………………………………………………… 5
   3. Proposed Solution …………………………………………………………………………….. 5
   4. Further Improvements ……………………………………………………………………… 6
3. Technical Requirements ………………………………………………………………………………. 6
   1. Tools Used ………….…………………………………………………………………………….. 6
4. Data Requirements ……………………………………………………………………………………… 6
   1. Data Collection ……….…………………………………………………………………………. 6
   2. Data Description …..…………………………………………………………………………… 7
   3. Exploratory Data Analysis …………………………………………………………….. 7
5. Data Preprocessing ………………………………………………………………………………………. 8
6. Design Flow ………………………………………………………………………………………..………… 8
7. Logging …………………………………………………………………………………………..…. 9
8. Data Validation ……………………………………………………………………………………….…….. 9
9. Rendering the Results …………….…………………….……………………………………………….... 9
10. Deployment ……………………………………………………………………………………………….. 9
11. Conclusion …………………………………………………………………………………………………. 10

**Abstract:**

Customer personality analysis is the process of identifying and understanding the unique characteristics and traits that make up an individual customer's personality. This information can be used by companies to tailor their marketing and sales efforts to better target and serve each customer's specific needs and preferences. The project aims to segment the customer data into clusters to better understand customer preferences towards a product.

This detailed project report will provide an overview of the project, its scope, objectives, methodology, and implementation details. The report will also discuss the challenges faced during the project, the solutions implemented, and the results obtained.

**1.Introduction**

The Customer Personality Analysis project is a machine learning-based solution to segment customers based on their information. The project aims to help companies to tailor their marketing and sales efforts to better target a product towards certain group of customers. The project will be developed using a unsupervised learning algorithm, which will be trained on a dataset of historical customer data.

**1.1 Why this DPR Document?**

The purpose of this detailed project report (DPR) is to provide a comprehensive overview of the Customer Personality Analysis project. This report will cover the entire project, from the problem statement to the proposed solution, and will also highlight the technical details, data analysis, and key performance indicators (KPIs). The document will serve as a guide for stakeholders to understand the project's progress and goals.

**2.General Description**

**2.1 Problem Perspective**

Understanding the customer’s interests and finding patterns in their purchasing behaviour is of key importance to many companies. Traditionally, customer personality analysis has been done manually by marketing and sales teams, who would use their expertise and experience to identify common patterns and trends among customers. However, with the advent of data mining and machine learning, it is now possible to automate this process using algorithms that can analyse large amounts of data and identify common patterns and traits among customers.

**2.2 Problem Statement**

The aim of this project is to segment the customers data based on the income, number of children, expenses and their marital status. The project uses unsupervised learning algorithms for customer personality analysis, and how they can be used by companies to better understand and serve their customers.

**2.3 Proposed Solution**

The proposed solution for this project is to use machine learning algorithms to develop customer data segmentation model. The model will be trained on historical customer data to learn the relationship between various factors that allow them to group customers in certain categories. Once the model is trained, it can be used to target customers with products.it also can be used to check to which cluster a new user belongs to.

**2.4 Further Improvements**

There are several opportunities for improvement in this project, including:

Enhancing the accuracy of the model by incorporating more data sources.

Implementing the model in a real-time environment to facilitate on-demand cluster segmentation and classification.

**3.Technical Specification**

This project will use Python as the primary programming language, along with several libraries such as Pandas, NumPy, Matplotlib, and Scikit-learn. The project will also use a Jupyter notebook for data analysis and model training. The final model will be deployed using Streamlit, a web application framework.

**4.Data Requirements:**

Historical data on customers purchasing information on a product

Data on customers including year of birth, income, marital status, number of children, expenses, education

Data on the type of expenses made including expenses on fruits, beverages, sweets, gold, meat etc.

Data on the purchases made on websites, through catalogues, store purchases

Data on response of customer on the product including frequency of them accepting product on the first instance or subsequent instances.

**4.1Data Collection**

The data has been collected from Kaggle:

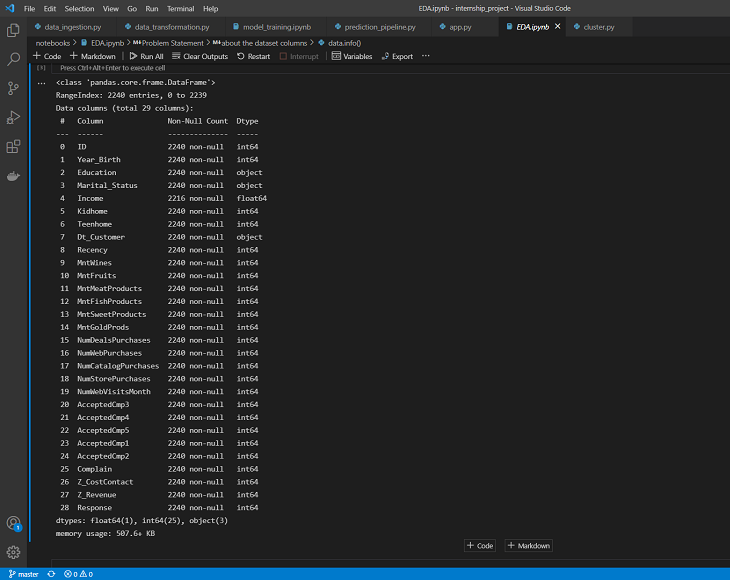
https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

**4.2 Data Description:**

The data should be in a structured format and should include all the required fields. The data should be clean and free from any errors or inconsistencies.

**4.3 Exploratory Data Analysis:**

Data cleaning: Check for missing values, outliers, and data errors that could impact the modeling process. Remove or impute missing values, remove outliers, and correct any data errors.



Data visualization: Visualize the data to get insights and identify potential issues. Use histograms, scatter plots, box plots, and other visualizations to explore the relationships between features and the target variable.

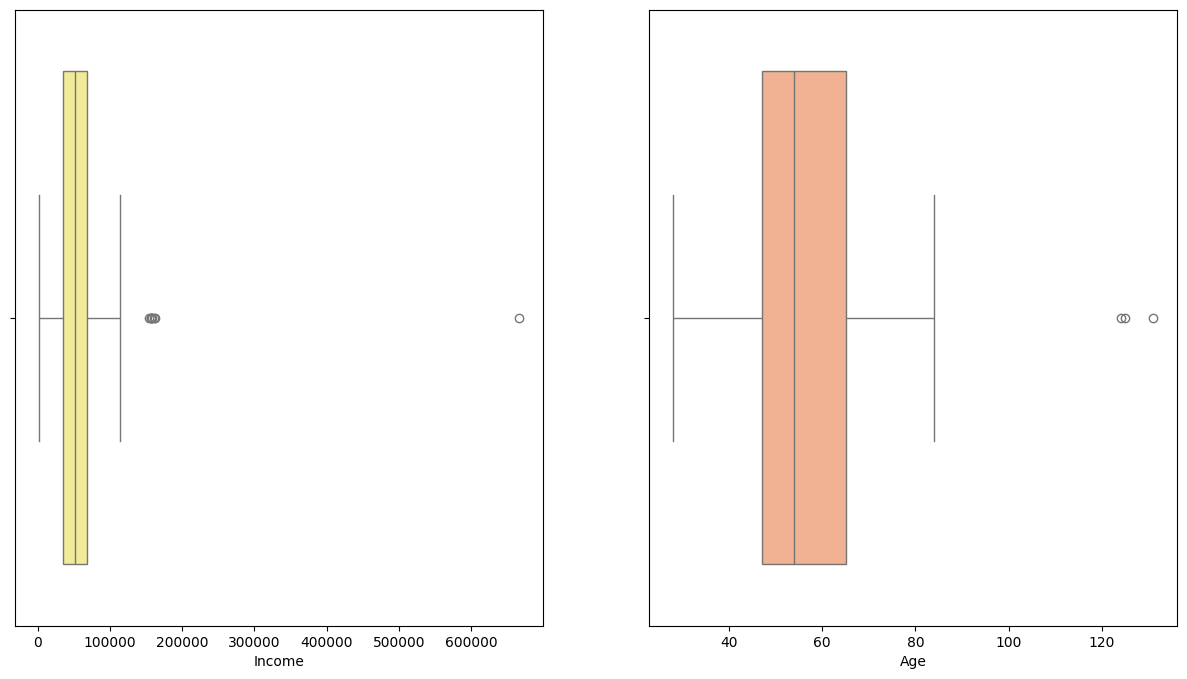


Fig 1. Plotting distribution of income and age to check for outliers

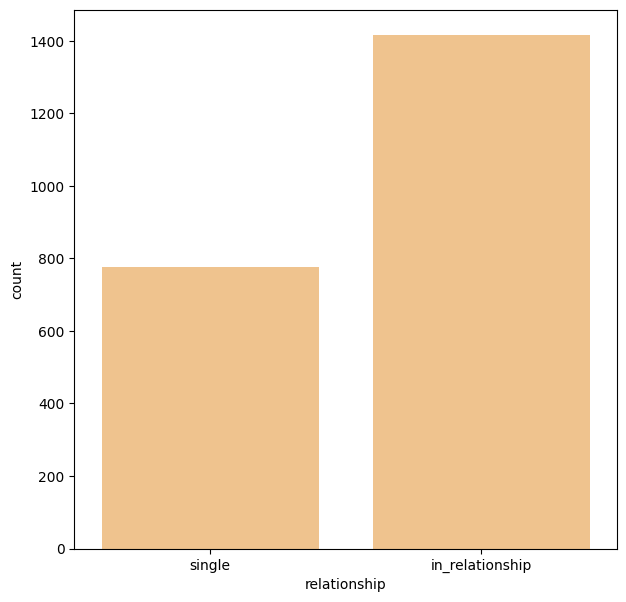


Fig 2. Plotting distribution of single and people in relationship

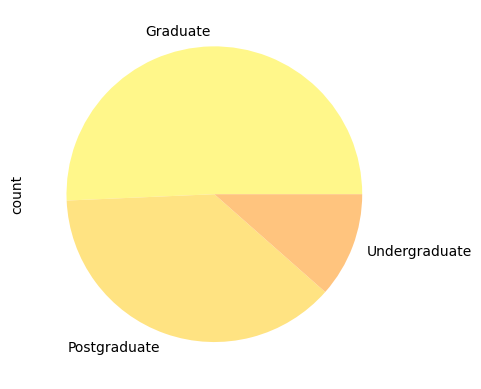


Fig 3. graph for distribution of education level

**5.Data Preprocessing**

Data cleaning refers to identifying incomplete, incorrect, inaccurate, or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Steps taken for preprocessing:

1. The income column of the data frame has 24 missing values. The missing numbers are being dropped because it is fewer.

2. We are creating a feature that displays age of the customer using the date of birth column..

3. In order to improve the clarity and consistency of the data, I added a new column called "relationship” which has the same meaning as an existing column with a different name. The same approach was used for the "Education" data as well.

4. Add a new feature called "expenses" that displays the customer's overall spending across all categories over a two-year period.

5. Dropping some of the redundant features.

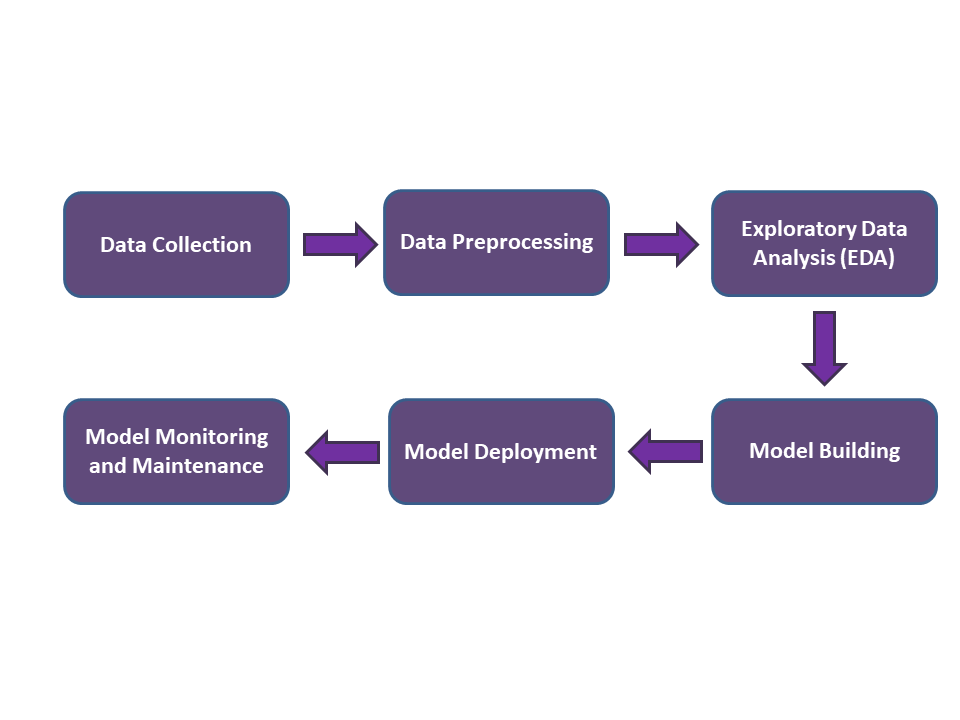
6. After plotting for “Income” and “Age”, outliers are present which will be deleted

7. Doing ordinal encoding for categorical feature i ‘Education’ and one hot encoding for feature ‘relationship’.

9. Dropping the columns of deals accepted and promotions, then scaling the remaining features using “Standard scaler”.

The data is quite clean, and the new features have been included.

**6.Design Flow**



**7. Logging**

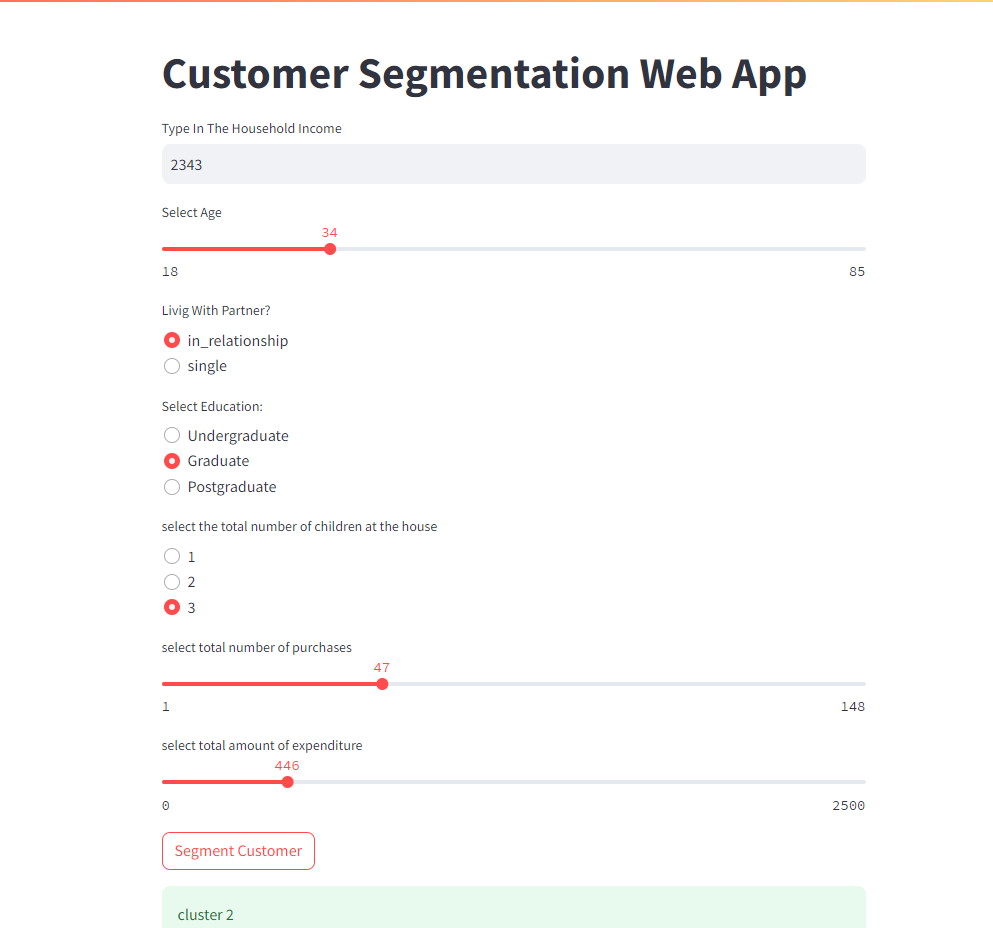
Logging will be used to capture and store relevant data for future analysis and improvement. The logging will capture key performance indicators (KPIs), such as the accuracy of the model, the time taken for the model to train, and the time taken for the model to make predictions.

**8.Data Validation**

The training set is used to train the model, while the testing set is used to evaluate the model's accuracy. Ensure that the data is randomly split to avoid bias.

**9.Deployment**

The first step in deploying an Customer Personality Analysis model is to prepare the model for deployment. This includes ensuring that the model is properly trained and validated, and that it is compatible with the deployment environment. Once the model is prepared, the next step is to create a web application using sreamlit that can be used to interface with the model. front-end are ready, the next step is to deploy it on the AWS.



**11.Conclusion**

In conclusion, the Customer Personality Analysis project has the potential to aid companies in targeting their products to specific group of customers and improve the customer experience by providing accurate and personalized products.