Detailed Project Report

Customer Personality Analysis

REVISION NUMBER – 1.2

Last date of Revision: 06/12/2022

Authored by: Aditya Jain
Aditya Sharma

Document Version Control

Date	Version	Description	Author
04/12/2022	1.0	Abstract	Aditya Sharma/Aditya Jain
		Introduction	
		General Description	
05/12/2022	1.1	Technical Requirements	Aditya Sharma/Aditya Jain
		Data Requirements	
		Data Preprocessing	
		Design Flow	
06/12/2022	1.2	Data from User and its validation	Aditya Sharma/Aditya Jain
		Rendering the Results	
		Deployment	
		Conclusion	

Contents

Abstract

As the legal cannabis industry emerges from its nascent stages, there is increasing motivation for retailers to look for data or strategies that can help them segment or describe their customers in a succinct, but informative manner. While many cannabis operators view the state-mandated traceability as a necessary burden, it provides a goldmine for internal customer analysis. Traditionally, segmentation analysis focuses on demographic or RFM (recency- frequency-monetary) segmentation. Yet, neither of these methods has the capacity to provide insight into a customer's purchasing behavior. With the help of 4Front Ventures, a battle-tested multinational cannabis operator, this report focuses on segmenting customers using cannabis-specific data (such as flower and concentrate consumption) and machine learning methods (K-Means and Agglomerative Hierarchical Clustering) to generate newfound ways to explore a dispensary's consumer base. The findings are that there are roughly five or six clusters of customers with each cluster having unique purchasing traits that define them. Although the results are meaningful, this report could benefit with exploring more clustering algorithms, comparing results across dispensaries within the same state, or investigating segmentations in other state markets.

1. Introduction

1.1 Why this DPR Document ?

The main purpose of this DPR documentation is to add the necessary details of the project and provide the description of the machine learning model and the written code. This also provides the detailed description on how the entire project has been designed end-to-end.

Key points-:

- Describes the design flow
- Implementations
- Software requirements
- Architecture of the project
- Non-functional attributes like:
  - Reusability
  - Portability
  - Resource utilization

2. General Description

2.1 Problem Perspective

Customer Personality Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers.

2.2 Problem Statement

Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer In the company'

database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment.

The main objective here is -

1. What people say about your product: what gives customers' attitude towards the product.

2. What people do: which reveals what people are doing rather than what they are saying about your product.

2.3 Proposed Solution

To solve the problem, we have created a User interface for taking the input from the user to predict the **Customer Behaviors** using our trained ML model after processing the input and at last the output (predicted value) from the model is communicated to the User.

2.4 Further Improvements

We also analysis the data used for training the ML model by considering different labels like Income, Kidhome, Teenhome, Age, Partner, Education Level.

3. Technical Requirements

As technical requirements, we don't need any specialized hardware for virtualization of the application. The user should have the device that has the access to the web and the fundamental understanding of providing the input.

3.1 Tools Used

- Python 3.9 is employed because the programming language and frameworks like NumPy, Pandas, Scikit - learn and alternative modules for building the model.
- Jupyter-Notebook is employed as IDE.
- For Data visualizations, seaborn and components of matplotlib are getting used.
- For information assortment prophetess info is getting used.
- Front end development is completed with streamlit by python
- Flask is employed for each information and backend readying.
- GitHub is employed for version management.

4. Data Requirements

The Data requirements is totally supported the matter statement and also the dataset is accessible on the Kaggle within the file format of (.csv).

4.1 Data Collection

The data for this project is collected from the Kaggle Dataset, the URL for the dataset-https://raw.githubusercontent.com/amankharwal/Website-data/master/marketing_campaign.csv

4.2 Data Description

The dataset contain 29 column and 2240 row which give us identity and other detail of customer which help us to predict the customer behavior and the dataset column income contain 24 null values.
First give a brief introduction to all attributes in the dataset we use:

- People
  - ID: Customer's unique identifier
  - Year_Birth: Customer's birth year
  - Education: Customer's education level
  - Marital_Status: Customer's marital status
  - Income: Customer's yearly household income
  - Kidhome: Number of children in customer's household
  - Teenhome: Number of teenagers in customer's household
  - Dt_Customer: Date of customer's enrollment with the company
  - Recency: Number of days since customer's last purchase
  - Complain: 1 if customer complained in the last 2 years, 0 otherwise

- Products
  - MntWines: Amount spent on wine in last 2 years
  - MntFruits: Amount spent on fruits in last 2 years
  - MntMeatProducts: Amount spent on meat in last 2 years

- MntFishProducts: Amount spent on fish in last 2 years
- MntSweetProducts: Amount spent on sweets in last 2 years
- MntGoldProds: Amount spent on gold in last 2 years

- Promotion
  - NumDealsPurchases: Number of purchases made with a discount
  - AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise
  - AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
  - AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
  - AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise
  - AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise
  - **Response**: 1 if customer accepted the offer in the last campaign, 0 otherwise

- Place
  - NumWebPurchases: Number of purchases made through the company's web site
  - NumCatalogPurchases: Number of purchases made using a catalogue
  - NumStorePurchases: Number of purchases made directly in stores
  - NumWebVisitsMonth: Number of visits to company's web site in the last month

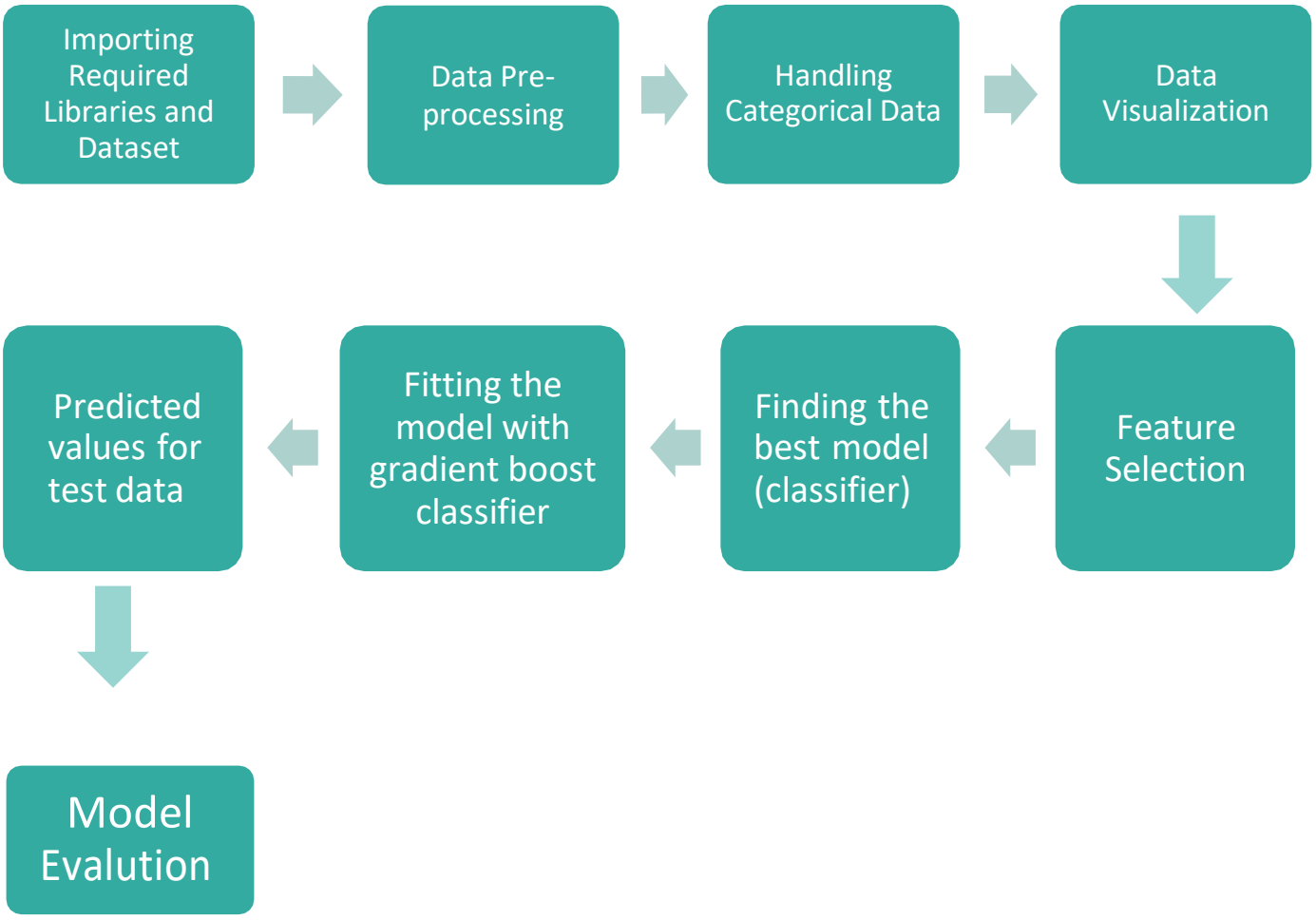
#	Column	Non-Null Count	Dtype
1	ID	2240 non-null	int64
2	Year_Birth	2240 non-null	int64
3	Education	2240 non-null	object
4	Marital_Status	2240 non-null	object
5	Income	2216 non-null	float64
6	Kidhome	2240 non-null	int64
7	Teenhome	2240 non-null	int64
8	Dt_Customer	2240 non-null	object
9	Recency	2240 non-null	int64
10	MntWines	2240 non-null	int64
11	MntFruits	2240 non-null	int64
12	MntMeatProducts	2240 non-null	int64
13	MntFishProducts	2240 non-null	int64
14	MntSweetProducts	2240 non-null	int64
15	MntGoldProds	2240 non-null	int64
16	NumDealsPurchases	2240 non-null	int64
17	NumWebPurchases	2240 non-null	int64
18	NumCatalogPurchases	2240 non-null	int64
19	NumStorePurchases	2240 non-null	int64
20	NumWebVisitsMonth	2240 non-null	int64
21	AcceptedCmp3	2240 non-null	int64
22	AcceptedCmp4	2240 non-null	int64
23	AcceptedCmp5	2240 non-null	int64
24	AcceptedCmp1	2240 non-null	int64
25	AcceptedCmp2	2240 non-null	int64
26	Complain	2240 non-null	int64
27	Z_CostContact	2240 non-null	int64
28	Z_Revenue	2240 non-null	int64
29	Response	2240 non-null	int64

5. Data Preprocessing

- Checked for info of the Dataset, to verify the correct datatype of the Columns.
- Checked for Null values, because the null values can affect the accuracy of the model.
- Converted all the desired columns into Datetime format.
- Performed One – Hot encoding on the desired columns.
- Checking the distribution of the columns to interpret its importance.
- Now, the info is prepared to train a Machine Learning Model.

6. Design Flow

6.1 Modelling Creation and Evaluation



6.2 UI Integration

We used streamlit for web page and UI for the prediction the customer behaviour.

6.3 Deployment Process

6.4 Logging

In logging, at each if an error or an exception is occurred, the event is logged into the system log file with reason and timestamp. These helps the developer to debug the system bugs and rectifying the error.

7. Data from User

The data from the user is retrieved from the created streamlit web page.

8. Data Validation

The data provided by the user is then being processed by streamlit file and validated. The validated data is then sent to the prepared model for the prediction.

9. Rendering the Results

The data sent for the prediction is then rendered to the web page.

10. Deployment

The tested model is then deployed to Heroku. So, users can access the project from any internet devices.

11. Conclusion

Regardless of the information provided, the results provide actionable ways for retailers to employ a marketing campaign or similar segmentation for their consumers. Despite the usefulness of the analysis as-is, there are numerous routes for improvement and growth. While there was motivation to keep the number of features low, adding a separate feature to account for the recency of the consumer would provide clearer details on whether certain purchase profiles are more common now than in the store's past. On a similar note, finding ways to cluster a customer quicker (such as in one or two visits rather than three) could generate insights into not only the evolutionary aspect of the clustering but potentially also the leakage of customers. Finally, attempting the same analysis with numerous other clustering algorithms such as Gaussian Mixture Models or deep learning would bring about insight into the stabilityof cluster formation.

12. Frequently Asked Questions (FAQs)

Q1) What's the source of data?

The data for training is provided by the client in multiple batches and each batch contain multiple files.

Q2) What was the type of data?

The data was the combination of numerical and Categorical values.

Q3) What's the complete flow you followed in this Project?

Refer Page no 12 for better Understanding.

Q4) After the File validation what you do with incompatible file or files which didn't pass the validation?

Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.

Q5) How logs are managed?

We are using different logs as per the steps that we follow in validation and modeling like File validation log, Data Insertion, Model Training log, prediction log etc.

Q6) What techniques were you using for data pre-processing ?

- Removing unwanted attributes.
- Visualizing relation of independent variables with each other and output variables.
- Checking and changing Distribution of continuous values.
- Removing outliers
- Cleaning data and imputing if null values are present.
- Converting categorical data into numeric values.

Q7) How training was done or what models were used?

- Before dividing the data in training and validation set, we performed pre-processing over the data set and made the final dataset.
- As per the dataset training and validation data were divided.
- Algorithms like Linear regression, SVM, Decision Tree, Random Forest, XGBoost were used based on the recall, final model was used on the dataset and we saved that model**.**

Q8) How Prediction was done?

The testing files are shared by the client. We Performed the same life cycle on the provided dataset. Then, on the basis of dataset, model is loaded and prediction is performed. In the end we get the accumulated data of predictions.

Q9) What are the different stages of deployment?

- First, the scripts are stored on GitHub as a storage interface.
- The model is first tested in the local environment.
- After successful testing, it is deployed on Heroku.