CUSTOMER SEGMENTATION

Thesis/Dissertation submitted to Rchilli Inc for submission for fulfillment of AI/ML internship assessment

By

Faiz Hasan Zaidi

CSE-AI Department, Noida Institute of Engineering and Technology,

Greater Noida, Uttar Pradesh, India

zaidifaizy108@gmail.com

TABLE OF CONTENT

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Rchilli for providing this opportunity of internship assessment and endeavor on customer segmentation using Machine Learning. I am also grateful for various authors of multiple shinning research papers that filled my appetite of the project and working , while showcasing various other approaches and pathways instead of a single approach .

Finally, I would like to extend my heartfelt thank to my family and friends for their constant support and understanding during demanding times in my dissertation .

CODING COMPONENT

```
#import libraries

import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

#importing Amazon user dataset
data_set = pd.read_csv('Amazon.com Clustering Model.csv')
X = data_set.iloc[:, :-1].values
y = data_set.iloc[:, -1].values
print(X)
print(y)

#Handling Missing data from input
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(missing_values = np.nan, strategy = 'mean')

imputer.fit(X[:, 0:1])
X[: , 0:1] = imputer.transform(X[:, 0:1])
print(X)
imputer.fit(X[:, 4:5])
X[: , 4:5] = imputer.transform(X[:, 4:5])
print(X)

#Encoding independent variables and dependent variables
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder
ct = ColumnTransformer(transformers=[('encoder',OneHotEncoder(), [1])],
remainder= 'passthrough')
X = np.array(ct.fit_transform(X))
print(X)


from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
y = le.fit_transform(y)
print(y)

from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,
random_state = 42)
print(X_train)
print(X_test)
print(y_train)
print(y_test)
```

```
#Feature Scaling
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
X_train[:, 6:] = sc.fit_transform(X_train[:, 6:])
X_test[:, 6:] = sc.fit_transform(X_test[:, 6:])
print(X_train)
```

#K-Means Clustering
#Using Elbow method for Optimal number of Cluster formation , WCSS

```
from sklearn.cluster import KMeans
wcss = []
for i in range(1, 11):
  kmeans = KMeans(n_clusters = i, init = 'k-mean++', random_state = 42)
  kmeans.fit(X)
  wcss.append(kmeans.inertia_)
plt.plot(range(1.11),wcss)
plt.title("Clusters for the dataset with Frequencies")
plt.xlabel('Numbers of clusters')
plt.ylabel('WCSS')
plt.show()
```

#K-Means Model Training and Testing

```
kmeans = KMeans(n_clusters = 4, init = 'k-means++', random_state = 42)
y_means = kmeans.fit_predict(X)
print(y_means)
```

#Visualization of Clusters and final result from Customer Segmentation

```
plt.scatter(X[y_means == 0, 0], X[y_means == 0, 1], s = 100, c = 'magenta',
label = 'Cluster1 ')
plt.scatter(X[y_means == 1, 0], X[y_means == 1, 1], s = 100, c = 'blue',
label = 'Cluster 2')
plt.scatter(X[y_means == 2, 0], X[y_means == 2, 1], s = 100, c = 'red',
label = 'Cluster 3')
plt.scatter(X[y_means == 3, 0], X[y_means == 3, 1], s = 100, c = 'cyan',
label = 'Cluster 4')
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_center_[:, 1], s =
300, c = 'black', label = "Centroids")
plt.title('Cluster of Amazon users')
plt.xlabel('Annual Income in INR')
plt.ylabel('Purchase Rating')
plt.legend()
plt.show()
```

REPORT COMPONENT

AIM AND OBJECTIVES

The most successful companies today are the ones that know their customers so well that they can anticipate their needs. Data Scientists and analysts play a key role in unlocking these in dept insights and segmenting the customers to better serve them. This dissertation aims to see what is customer segmentation, segmentation factors and its advantages. We have to understand the approach to solve customer segmentation problem , first step is to create a business case. You don't want to go into this process blindly, otherwise the outcome may be messy and disorganized. The optimal approach would be to find the mist profitable customer groups within the entire pool of customers.

Customer segmentation is defined as dividing company's customers on the basis of demographic (age, gender, marital status) and behavioral (types of products ordered, annual income) aspects[1]. Since demographic characteristics does not emphasize on individuality of customer because same age groups may have different interests so behavioral aspects is a better approach for customer segmentation as its focus on individuality and we can do proper segmentation with the help of it[2].

Customer loyalty to a product or service is advantageous since customers will continue to look for the things they want[3].Every business owner must improve the quality of their current offering as well as client satisfaction[4].

LITERATURE REVIEW

The customer segmentation has the importance as it includes, the ability to modify the programs of market so that it is suitable to each of the customer segment, support in business decisions; identification of products associated with each customer segmentation and to manage the demand and supply of that products; identifying and targeting the potential customer base, and predicting customer defection, providing directions in finding the solutions[5].

Being fundamental aspect of marketing that involves dividing a diverse customer base into distinct groups based on similar characteristics, behaviors, or preferences[6].In this dissertation I have elaborated not only the methodology that I have used for customer segmentation but also the difference in the traditional and new machine learning approach.

Historically, customer segmentation relied on demographic, geographic, and psychographic factors. However, these static parameters often overlook the complexity of customer behavior and fails to adapt to evolving market dynamics[7].On the other hand, Machine learning algorithms offer the capability to process vast amounts of data and identify patterns that are beyond the scope of traditional methods[8].For this we have wide range of clustering algorithms, such as K-means, hierarchical clustering, and DBSCAN, have been widely used to attain parallel goals to segmentation of data.

For the literature review and coding implementation I have taken dataset of Amazon customers having details of there name, age, gender , purchase rate price and few other aspects. After the dataset only . I implemented K-means clustering algorithm to attain final goal. Also this should be noted that the data was preprocessed and then only clusters are made.

A major challenge faced in customer segmentation is dealing with high-dimensional data. Feature selection and dimensionality reduction techniques, including Principal Component Analysis(PCA) and t-Distribution Stochastic Neighbor Embedding, play a crucial role in enhancing the efficiency of

machine learning models[9]. These methods helps identify the most relevant features for accurate segmentation and visualization of complex datasets. It is still a issue that despite promises of machine learning in segmentation, overfitting , interpretability and data privacy can arise highly. Striking a balance between predictive accuracy and model transparency remains a focal point for research[10].Recent studies emphasize the importance of integrating diverse data sources, including social media, online behavior, and transaction history, to create comprehensive customer profiles[11]. This helps in enhancing the accuracy of customer preference as thoroughly performed and examined by myself and can be understood by the code provided. It is also observed that with the use of neural networks , specifically light weighted nodes in deep neural network can be more efficient with large data set and might also be smooth to work with national interest data.

METHODOLOGY

The Methodology used for customer segmentation in this assessment was with the help of data preprocessing and then performing K-means clustering on the dataset as resultant product.

1.Data Preprocessing: It is a step in data analysis process that transforms or changes the raw data into a format that can be easily understood and analyze by computer system. This raw data can not only be textual but can also be image, audio or video.

(a) Importing libraries: At first I imported the libraries that are essential for data preprocessing , which are numpy for numerical analysis of data , pandas, which is highly used for analyzing , exploring and extracting data and matplotlib for graphical representation.

(b) Importing Dataset: The dataset taken for the customer segmentation is amazon customer dataset. As prior to this assessment I have already worked on a project for Amazon customer segmentation therefore it was suitable for better performance and accurate outcome.

(c) Handling missing data: It was crucial aspect in the assessment , as incomplete information would have compromised the accuracy and performance of the model[12]. For this sklearn.impute was used which is a part of Simpleimputer class to provide basic strategies and work for imputing missing values. The missing values can also be imputed with a provided constant value , very useful when having minimum data.

Here , an advancement could also be done by applying K-nearest neighbor Algorithm however, It was not applied as the data was not that vast and requirements were minimum therefore, the probability of occurrence of overfitting is quite low.

(d) Encoding independent and dependent variable : As stated above the data is complex and avoid any knotted interconnection therefore using sklearn.compose from sklearn library which transforms the data and considers them in columnar form.

(e) Splitting data into Test set and Training set: This is a sub part of exploratory data analysis , where the data by using K-Folds or by input range (when the user defines the percentage taken for training and test data) . Generally this value in assessment is taken to be near to 80% for training and remaining 20% for testing set.

2. K-Means Clustering : After preprocessing of data , there is not further anomaly or chance of overfitting that may hinder the accuracy of model. As I performed this assessment in Google Colab, there are chances that one might have to import the libraries again however, It is recommended to perform this code in VS Code IDLE , Jupyter Notebook or Pycharm as it will be quite easy to initiate the steps , otherwise there is no harm in using Colab.

(a)Optimal Number of cluster via Elbow method: It is a common technique used to determine number of clusters in a dataset. The idea is to run the clustering algorithm respectively taken with various other clusters and plot he variation against the number of clusters[13].

I have imported cluster function from sklearn library and taking WCSS(which is the sum of squared points to the distance of centroid of the cluster. Then using n clusters and also *K-mean++ , which is used to deal with the problem of Random Initialization Trap* (Frequently occurs while forming clusters and is necessary to deal with).

(b)K-Means model Training and testing set: After the set are assigned earlier with respect to the percentages take , now these set are initiated and used for training and test model . At first , with 80% of the data training is performed and then only the testing is done. With there comparison we can identify the accuracy of the working model.

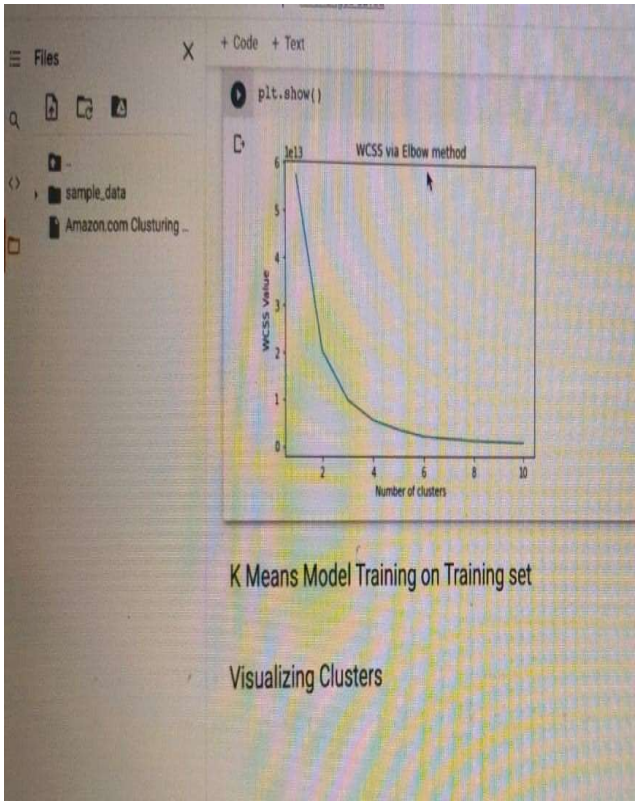

Figure 1: Optimal Number of Clusters,

Usage of Elbow Method

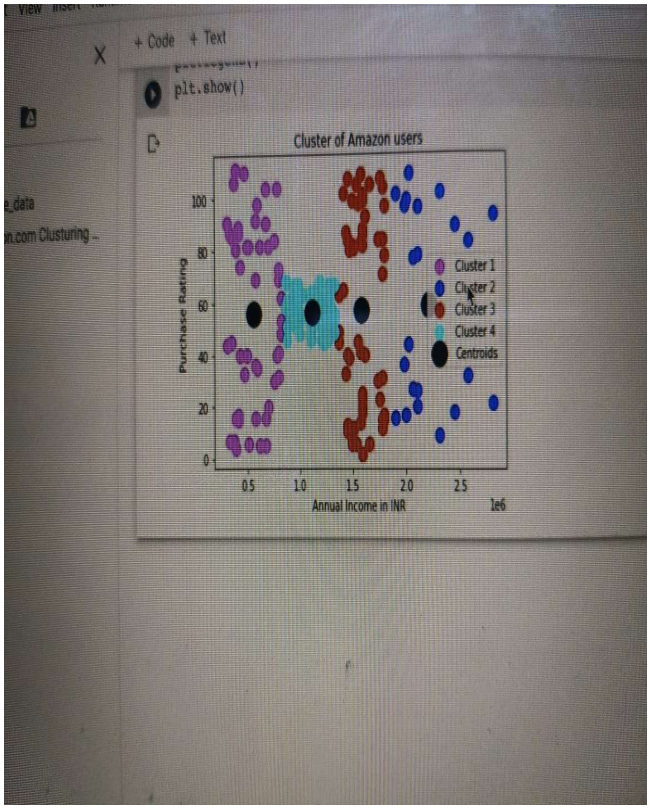

Figure 2: Final Amazon User Segmentation

CONCLUSION

The Integrational working and usages of machine learning and data analysis have revolutionized various sectors and applications have emerged, one of them is customer segmentation, offering unprecedented business insights into their customer preferences. As the technology continues to advance , addressing various challenges and ethical considerations will become essentially important to enable machine learning to its full extend. With the help of this dissertation I propose the application of Machine learning algorithms and various data analysis tools, in the face of customer segmentation of Amazon dataset, but it can also be used on larger scale . I also intend to use Deep Neural networks and Random Forest to make it close to 100% accuracy so that the model will also be applicable to wide range to data.

REFERENCES

[1]. Nikhil Patankar, Customer Segmentation using machine learning, Research gate, paper ID:10.3233/APC21022000.

[2]. Sharareh Rostam Niakan Kalhori, Towards the Application of Machine Learning in Emergency Informatics , 05/2022, researchgate.in,10.3233/SHTI220003

[3]. Md. Nasir Hussain, Ch. Harsha, Sameer Shaik, Usage of Deep Learning Techniques for personalized Recognition Systems , 07/2023, researchgate.in,ICESC

[4]. Riyo Hayat Khan, Explainable Customer Segmentation using K-Means Clustering, Paper ID:11.1109/4Cmcon53757, year 2021.

[5]. Varld R Thakur, Customer Segmentation using machine learning , International Journal of Scientific Research in Computer Science , 10.32628/cseit217654.

[6]. Akshay Bhamare, Machine Learning Segments and Marketable Application, SUS-DEV International Conference, Researchgate.com .

[7]. Hyatt Saleh,Books Machine Learning Fundamentals : sklearn for hottest developments, Packt Publications .

[8]. Iqbal.H.Sarker , Machine learning algorithms and real world Applications , springer.com, 21 Aug, 2021 .

[9]. Kaebling LP, Littman ML, Moore AW. Reinforcement learning: a survey. J Artif Intell Rs.1996;4:237-85.

[10]. Batta Mahesh, Machine Learning Algorithms – A Review, 01/2019, researchgate.net, DOI:10.21275/ART20203995

[11]. W. Richert, L. P. Coelho, “Building Machine Learning Systems with Python”, Packt Publishing Ltd

[12]. P. Harrimgton, “Machine Learning in action’, Manning Publications Co., Shelter Island, New York, 2012

[13]. Praba. R, Darshan. G, Roshanraj K. T

Surya Prakash. B, Study On Machine Learning Algorithms, International Journal of Scientific Research in Computer Science Engineering, researchagte.in, 10.32628/CSEIT2173105