CustomerSegmentation Using MachineLearning

Our Team:

39110081 - Arjun Singh

39110246 - D.V.V.Manoj

39110177 - B.Achyuth

1. INTRODUCTION:

1.1 Overview:

➢ The primal task of Management is to identify potential customers from the rest. This will be simplified with the help of Machine Learning models to classify the customers into segments based on various attributes.

➢ The AI helps the business to build such models to analyze the customers and their products. To improvise the business process and to improve the revenue.

➢ We will be using classification algorithms such as H-clustering, k-means clustering Decision tree, Randomforest, KNN, and xgboost.

➢ From this best model is selected and saved in pkl format. Once the model is saved, we integrate it with the flask application.

1.2 Purpose:

➢ Segmenting allows you to more precisely reach a customer or prospect based on their specific needs and wants.

➢ You'll able to understand the unsupervised learning methods such as k-means clustering.

➢ You'll be able to understand the problem to classify if it is a regression or a classification kind of problem.

➢ You will be able to know how to build a web application using the Flask framework.

2. LITERATURE SURVEY:

2.1 Existing Problem:

Over the years, as there is very strong competition in the business world, the organizations have to enhance their profits and business by satisfying the demands of their customers and attract new customers according to their needs. The identification of customers and satisfying the demands of each customer is a very complex and tedious task. This is because customers may be different according to their demands, tastes, preferences and so on. Instead of "one-size-fits-all" approach, customer segmentation clusters the customers into groups sharing the same properties or behavioural characteristics. According to customer segmentation is a strategy of dividing the market into homogenous groups. The data used in customer segmentation technique that divides the customers into groups depends on various factors like, data geographical conditions, economic conditions, demographical conditions as well as behavioural patterns. The customer segmentation technique allows the business to make better use of their marketingbudgets, gain a competitive edge over their rival companies, demonstrating the better knowledge of the needs of the customer. It also helps an organization in, increasing their marketing efficiency, determining new market opportunities, making better brand strategy, identifying customers retention.

2.2 Proposed Solution:

➢ In this segmentation we segment customers of based on marital status, age, income…
➢ So we use Demographic segmentation for this.
➢ Demographic segmentation groups customers and potential customers together by focusing on certain traits such as age, gender, income, occupation & family status.

➢ Advantages of Demographic segmentation is it is Readily Available Data, Increased Customer Loyalty and Retention and Helps Develop more Effective Marketing Strategies.

➢ The word demography is derived from the Greek word "demos," meaning people and the English word "graph" which means "the study of." When these two words are combined, they mean "the study of people."

CLUSTERING AND K-MEANS ALGORITHM:

Clustering algorithms generates clusters such that within the clusters are similar based on some characteristics. Similarity is defined in terms of how close the objects are in space.K-means algorithm in one of the most popular centroid basedalgorithm. Suppose data set, D, contains n objects in space. Partitioning methods distribute the objects in D into k clusters, $C_1,...,C_k$, that is, $C_i \subset D$ and $C_i \cap C_j = \emptyset$ for $(1 \leq i, j \leq k)$. A centroid-based partitioning technique uses the centroid of a cluster, C_i , to represent that cluster. Conceptually, the centroid of a cluster is its center point. The difference between an object $p \in C_i$ and c_i , the representative of the cluster, is measured by $dist(p,c_i)$, where $dist(x,y)$ is the Euclidean distance between two points x and y.

Algorithm: The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster. Input: k: the number of clusters, D: a data set containing n objects. Output: A set of k clusters. Method: (1) arbitrarily choose k objects from D as the initial cluster centers; (2) repeat (3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster; (4) update the cluster means, that is, calculate the mean value of the objects for each cluster; (5) until no change.

METHODOLOGY:

The data set used to implement clustering and Kmeans algorithm was collected from Smart bridge. The data set contains 8 attributes and has 2000entries, representing the data of 2000 customers. The attributes in the data set has CustomerId, gender, age, Marital status, Education,income, Occupation, Settlementsize on the scale of(1-1999).

ELBOW METHOD:

The elbow method is based on the observation that increasing the number of clusters can help to reduce the sum of within-cluster variance of each cluster. This is because having more clusters allows one to capture finer groups of data objects that are more similar to each other.To define the optimal clusters, Firstly, we use the clustering algorithm for various values of k. This is done by ranging k from 1 to 10 clusters. Then we calculate the total intra-cluster sum of square. Then, we proceed to plot intra-cluster sum of square based on the number of clusters. The plot denotes the approximate number of clusters required in our model.The optimum clusters can be found from the graph where there is a bend in the graph.
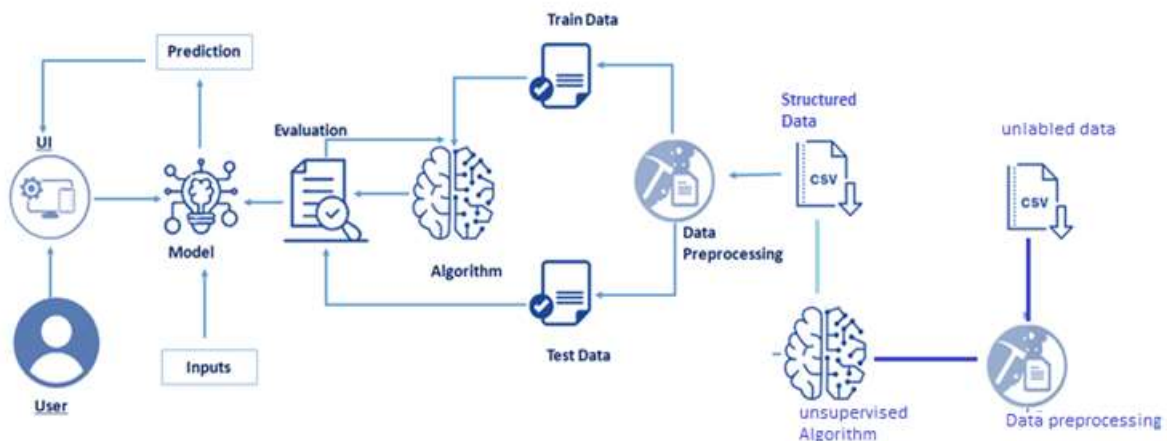
```python
wcss = []
for i in range(1,11):
    kmeans = cluster.KMeans(n_clusters=i,init='k-means++',random_state=0)
    kmeans.fit(data)
    wcss.append(kmeans.inertia_)
```

```python
plt.pyplot.plot(range(1,11),wcss)
plt.pyplot.title('elbow method')
plt.pyplot.xlabel('no.of clus')
plt.pyplot.ylabel('wcss')
plt.pyplot.show()
```

elbow method
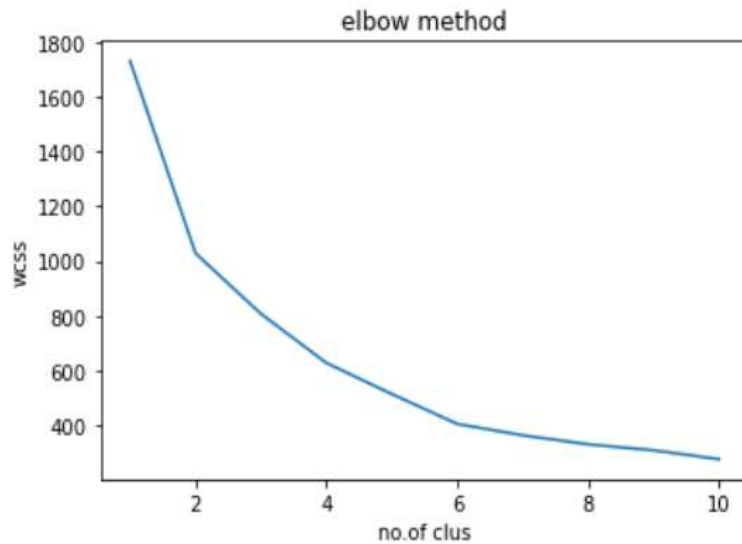
3.THEORITICAL ANALYSIS:

3.1 Block Diagram:
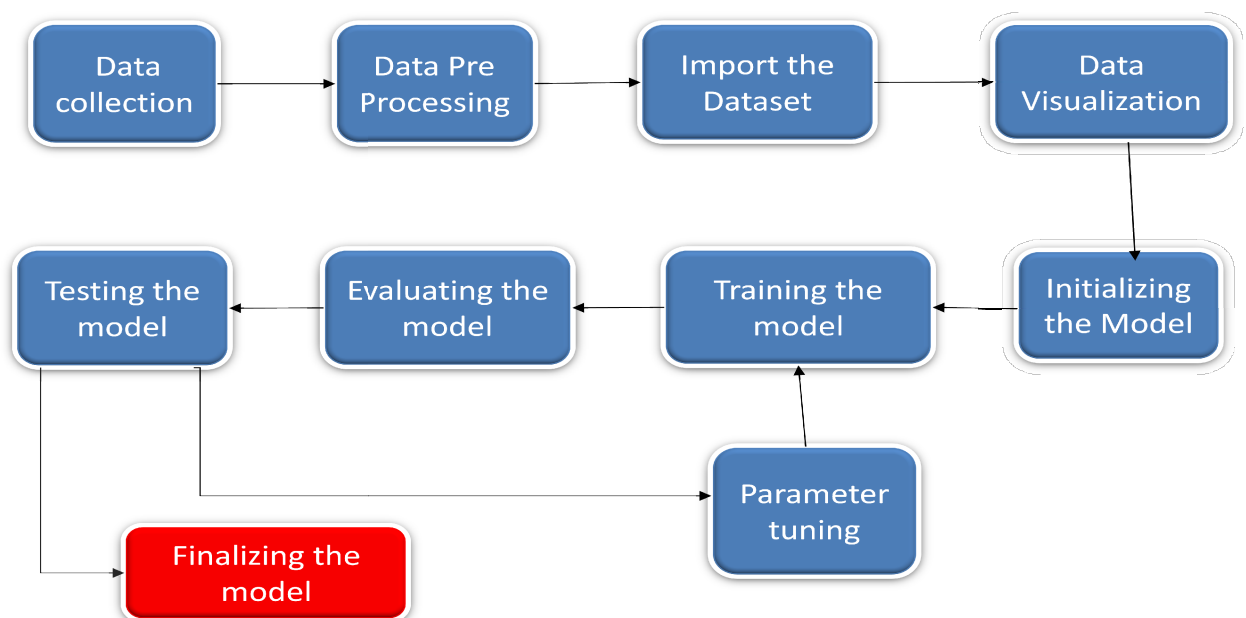

3.2 Hardware / Software requirements:

➢ Anaconda Navigator

➢ Jupyter Notebook

➢ Spyder

➤ OS: Windows 10

➤ CPU: Intel or AMD processor with 64-bit support

➤ Disk Storage: 4 GB of free disk space

➤ Internet: Internet connection required for software activation

4. EXPERIMENTAL INVESTIGATIONS:

➤ Customer segmentation research is research that is used to help a firm identify segments in a market, with the end goal of developing different strategies and tactics for the different segments.

➤ While working on this project I have known that there are 3 more methods for customer segmentation they are psychographic, behavioral and geographic segmentation.

5.FLOWCHART:

```
Data         →   Data Pre      →   Import the    →   Data
collection       Processing        Dataset           Visualization
                                                         ↓
Testing the  ←   Evaluating the ←  Training the  ←   Initializing
model            model             model             the Model
   ↓                                  ↑
   |                                  |
   |              Finalizing the  →  Parameter
   |              model              tuning
```

6. RESULT:

> ➤ It will display all the input parameters and the prediction text will display the output value of the data given by the user.

> ➤ If customer is highly potential it will redirect to the following page.

Customer Segmentation

Please enter the following details

Sex: Male
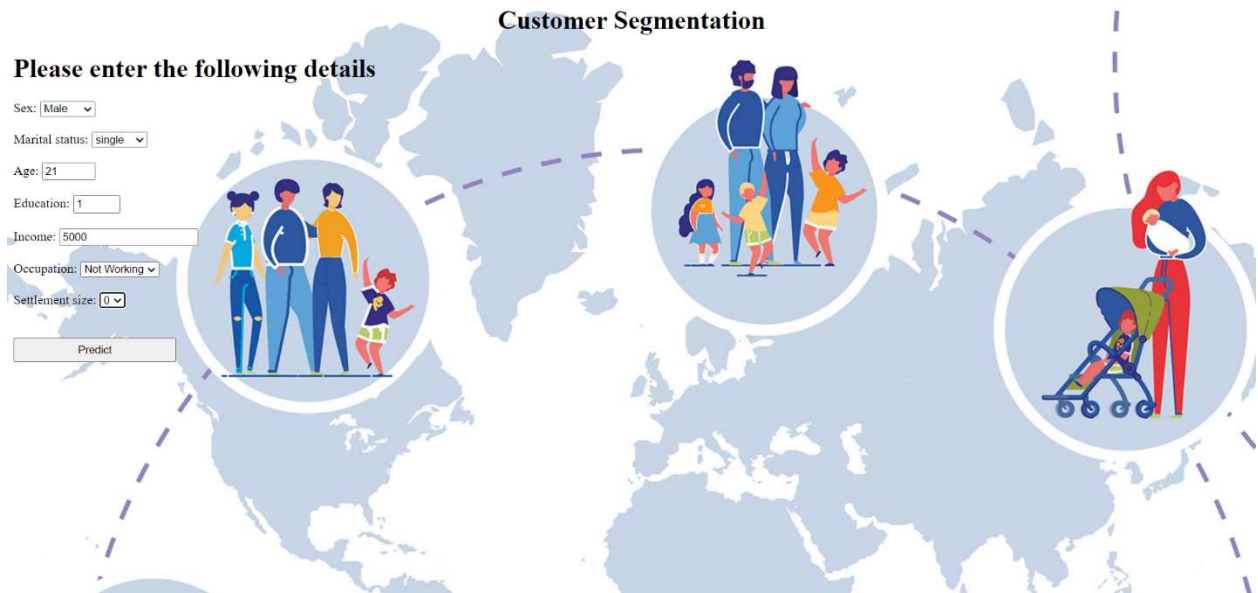Marital status: single
Age: 21
Education: 1
Income: 5000
Occupation: Not Working
Settlement size: 0

Predict

> If the customer is not potential it will redirecting to this page.

Predicted person is not a potential customer.

7. ADVANTAGES:

1. Data is easy to obtain
2. Straight forward targeting and analysis
3. Cost – effective
4. Easy to measure
5. Ideal for monitoring trends and social shifts

DISADVANTAGES:

1. Based on assumptions
2. Demographic data is too indefinite
3. Misinterpreting data
4. Change

8. APPLICATIONS:

➢ Customer segmentation enables a company to customize its relationships with the customers, as we do in our daily lives.
➢ When you perform customer segmentation, you find similar characteristics in each customer's behavior and needs. Then, those are generalized into groups to satisfy demands with various strategies.
➢ Targeted marketing activities to specific groups
➢ Launch of features aligning with the customer demand
➢ Development of the product roadmap
➢ There are different products/solutions available in the market from packaged software to CRM products. Today, I will apply an unsupervised machine learning algorithm with Python.

9.CONCLUSION& FUTURE SCOPE:

We approached customer segmentation problem from aspect with the demographic featureslike Age, Education, Income, Occupation, Settlement size of each customer. Use of 5 features helped us with the understandability and visualization of the model.

All in all, the dataset was apt to perform an unsupervised machine learning problem. At first, we only had customers data with demographic information and did not know if they belonged to any group. With the K-means clustering, patterns in the data were found and extended further into groups. We carved out strategies for the formed groups, making meaning out of a dataset that is a dust cloud initially.

Finally, I would like to thank Smart Internz for providing this wonderful opportunity to work with real-world data. This helped me gain valuable experience and helped me use and improve my skills.

10. BIBILOGRAPHY

[1] I. S. Dhillon and D. M. Modha, "Concept decompositions for largesparse text data using clustering," Machine Learning, vol. 42, issue 1,pp. 143-175, 2001.

[2] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R.Silverman, and A. Y. Wu, "An efficient K-means clusteringalgorithm," IEEE Trans. Pattern Analysis and Machine Intelligence,vol. 24, pp. 881-892, 2002.

[3] MacKay and David, "An Example Inference Task: Clustering,"Information Theory, Inference and Learning Algorithms, CambridgeUniversity Press, pp. 284-292, 2003.

[4] Jiawei Han, Micheline Kamber, Jian Pei "Data Mining Concepts and Techniques", Third Edition.

[5] D. Aloise, A. Deshpande, P. Hansen, and P.Popat, "The Basis Of Market Segmentation"Euclidean sum-of-squares clustering," Machine Learning, vol. 75, pp.245-249, 2009.

APPENDIX

A.SOURCE CODE

```python
import numpy as np
import pickle
import pandas
import os
from flask import Flask, request, jsonify, render_template


app = Flask(__name__)
model = pickle.load(open(r"C:\Users\hp\GradientBoostingClassifier.pkl", 'rb'))

@app.route('/')# route to display the home page
def home():
    return render_template('index.html') #rendering the home page

@app.route('/predict',methods=["POST","GET"])# route to show the predictions in a web UI
def predict():
    #  reading the inputs given by the user
    input_feature=[float(x) for x in request.form.values() ]
    features_values=[np.array(input_feature)]
    names = [['Sex', 'Marital status', 'Age', 'Education', 'Income', 'Occupation',
        'Settlement size']]
    data = pandas.DataFrame(features_values,columns=names)

     # predictions using the loaded model file
    prediction=model.predict(data)
    print(prediction)

    if (prediction == 0):
        return render_template("notimp.html",prediction_text ="Not a potential customer")
    elif (prediction == 1):
        return render_template("imp.html",prediction_text = "Potential customer")
    else:
        return render_template("moreimp.html",prediction_text = "Highly potential customer")
     # showing the prediction results in a UI
if __name__=="__main__":

    # app.run(host='0.0.0.0', port=8000,debug=True)    # running the app
    port=int(os.environ.get('PORT',5000))
    app.run(port=port,debug=True,use_reloader=False)
```