IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

NICHE PREDICTION FOR AUTOMOBILE MANUFACTURERS

Mrs. V. Tamilselvi M.E.,
Department of Computer Science and Engineering
Velalar College of Engineering and Technology
Thindal, Erode - 638012

Meenachi.S
Department of Computer Science and Engineering
Velalar College of Engineering and Technology
Thindal, Erode - 638012

Janaranjani.C
Department of Computer Science and Engineering
Velalar College of Engineering and Technology
Thindal, Erode

Kavipriya.E
Department of Computer Science and Engineering
Velalar College of Engineering and Technology
Thindal, Erode - 638012

***Abstract*—** Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Customer segmentation is the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately. In business-to-business marketing, a company might segment customers according to a wide range of factors. Creating a software application for customer segmentation for segmenting new customers into the right group for an automobile company, so the company can adopt the specific proven marketing strategy to each of them and be more successful in the business using classification algorithm.

Keys - Customer segmentation, Prediction, classification models.

I INTRODUCTION

Earlier automobile companies used to make marketing for all people without considering the number of sales and demand. For any manufacturer to determine whether to increase or decrease the production of several units, data regarding the demand for products on the market is required. Companies can face losses if they fail to consider these values while competing on the market. Different companies choose specific criteria to determine their demand and sales [1].

In today's highly competitive environment and ever-changing consumer landscape, accurate and timely segmentation of customer, also known as niche prediction, or customer segmentation, can offer valuable insight to companies engaged in the manufacture, distribution or retail of goods[2].

Customer Segmentation means grouping the customers based on marketing groups which shares the similarity among customers. Alternatively, the abundance of sales data and related information can be used through Machine Learning techniques to automatically develop accurate customer segmentation models. This approach is much simpler. It is not prejudiced by a single sales manager's particularities and is flexible, which means it can adapt to data changes. For example, once companies used to produce the products without taking into consideration the number of sales and demand as they

faced several problems. Since they don't know how much to sell, for any manufacturer to decide whether to increase or decrease the number of units, data regarding the consumer demand for products is essential. If companies do not consider these principles when competing in the market, they will face losses.

There are several ways of customer segmentation in which companies focused on various statistical models such as xgboost, decision tree, random forest, linear regression, svm. XGBoost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models.

Linear regression is a mathematical tool used to forecast past values. It can help to determine the underlying trends and address cases involving overstated rates[5][6].A decision tree is a fundamental principle behind a model of random forests. The decision tree approach is a technique used in data mining to forecast and classify data. The decision tree approach does not provide any conceptual understanding of the issue itself. Random forest is the more sophisticated method that allows and merges many trees to make decisions. The random forest model results in more accurate forecasts by taking out an average of all individual tree decision predictions.

II LITERATURE REVIEW

The concept of customer segmentation using machine learning is proposed by V.Vijilesh, A.Harini, M.Hari Dharshini, R.Priyadharshini[10]. In this paper, they implemented using "k-means", an unsupervised clustering machine learning algorithm. The model has partitioned customers into mutually exclusive groups, three clusters in our case. Dr.C.K.Gomathy, Kuncham Pavan Kumar Reddy, Kondakandla Srikar ,K.Siva Sankar[14] has proposed. Machine learning approaches are an incredible instrument for dissecting customer information and tracking down bits of knowledge and examples. Misleadingly wise models are useful assets for chiefs. They can exactly recognize client fragments, which is a lot harder to do physically or with ordinary logical techniques. There are many machine learning algorithms, each reasonable for a particular sort of issue. One extremely normal AI calculation that is appropriate for client division issues is the k-means clustering algorithm. Yossi Hadad, Baruch Keren[18] has proposed. This paper proposes a DMSS module for achieving segmentation and a customer ranking matrix. The module is based on an objective multi-criteria method for evaluating the relative value and cost to serve each customer. The DMSS module was designed on the basis of a survey conducted among 39 managers. This module can be considered as an expert system for customer segmentation. Six significant advantages of the proposed module were presented in the introduction section. The main advantage of the module is its ability to segment customers objectively according to quantitative criteria that can be extracted from the organizational CRM. Segmentation and a full ranking of the customers can be derived automatically at any moment with up-to-date data. This allows a quicker response to changes in customer profiles. Syedreza Baharisaravi[16] has proposed. The first objective of our research was to perform automatic customer segmentation based on usage behavior, without the direct intervention of a human specialist. The second part of the research was focused on profiling customers and finding a relation between the profile and the segments. The customer segments were constructed by applying Gustafson-Kessel Algorithm. The clustering algorithms used selected and preprocessed data from the Vodafone data warehouse. This led to solutions for the customer segmentation with respectively four segments and six segments. The customer's profile was based on personal information of the customers. A novel data mining technique, called Support Vector Machines was used to estimate the segment of a customer based on his profile. Balmeet Kaur, Pankaj Kumar Sharma[17] has proposed. This paper proposes a study on integrated novel approach based on clustering using K-means and associative mining using Apriori technique. After identification of targeted customers and their associative buying pattern, the business managers take the strategic profitable decisions accordingly. This

integrated model could be directly brought into implementation for providing better profitable margins from sales.

III DATASET

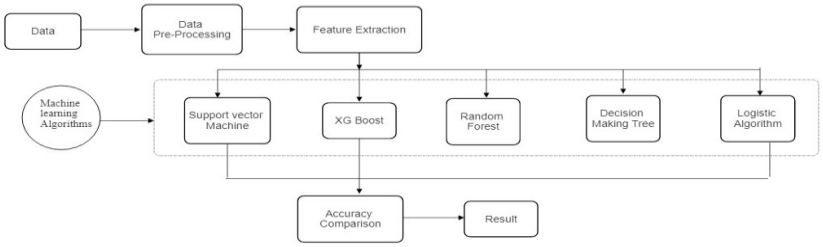
Variables Name	Description
ID	Unique ID
Gender	Male/Female
Martial_Status	customer married (yes/no)
Age	customer age
Graduated	customer graduated (yes/no)
Profession	customer profession
Work Experience	customer work experience
Spending_Score	customer spending score
Var_1	Anonymised Category for the customer
Family_Size	customer family size
Segmentation	Segmentation of customer

Table-1: Dataset Variables and their Description

The dataset includes 10577 rows and a total 11 variables or columns. In which 10 are independent variables and 1 is dependent variable. There are also some null values in the dataset. These values are filled by using the mean and mode method and also the label encoder is used to convert the string values by 1 and 0. The Segmentation attribute is the target variable.

IV METHODS

Machine Learning algorithms are used to divide customers into groups based on common characteristics. Organizations can then choose how to connect with clients in every class to advance the worth of every client to the business.



During the process, a machine learning model can able to segment the new customers into the right group for an automobile company . The procedure has six steps:

- Data Collection
- Data Preprocessing
- Feature Extraction
- Encoding Categorical Values
- Model buliding
- Prediction

A. Data Collection

In this thesis, there is labeled sales data from different automobile model from different outlets that provide information such as item type, item price, outlet type, etc. These data were extracted from various sources and will be used to train and improve the model for Machine Learning. In the data set being analyzed there are 10577 instances and 11 attributes. The data set has been properly divided into training and testing data that can be described in the sections below.

B. Data Preprocessing

```
[ ] data.isnull().sum()

ID                   0
Gender               0
Ever_Married       190
Age                  0
Graduated          102
Profession         162
Work_Experience   1098
Spending_Score       0
Family_Size        448
Var_1              108
Segmentation         0
dtype: int64
```

Fig 2. Data Preprocessing

Before applying Machine Learning algorithms some of the missing values have been found which can impact the model's output so this should be handled. In above shown figure, have more than one attribute of missing values. To make the dataset more efficient, these missing values will be replaced by the most promising values. There's more correlation between two of the different attributes with similar work. Removing one of the attributes will make the work better.

C. Feature Extraction

Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.[7]

When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set offeatures (also named afeature vector). Determining a subset of the initial features is called feature selection.[8] The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

D. Encoding Categorical Values

Categorical data contains label values that are considered nominal values. Each value has categories of different types. Besides, a few of the groups have a normal relationship with each other is known as natural ordering. The categorical data can be converted into numerical data to improve the efficiency of the Machine Learning model[9].

Get Dummy

In regression analysis, a dummy variable represents subgroups of the sample numerically. In the simplest case, a dummy variable where a category is given a value of 0 if the category is in the control group or 1 if the category is in the treated group. These variables are useful since a single regression equation can be applied to multiple groups.

A dataset may contain various type of values, sometimes it consists of categorical values. So, in-order to use those categorical value for programming efficiently we create dummy variables. A dummy variable is a binary variable that indicates whether a separate categorical variable takes on a specific value.

```
[ ] data
```

	ID	Gender	Ever_Married	Age	Graduated	Profession	Work_Experience	Spending_Score	Family_Size	Var_1	Segmentation
0	462809	Male	No	22	No	Healthcare	1.0	Low	4.0	Cat_4	D
1	462643	Female	Yes	38	Yes	Engineer	NaN	Average	3.0	Cat_4	A
2	466315	Female	Yes	67	Yes	Engineer	1.0	Low	1.0	Cat_6	B
3	461735	Male	Yes	67	Yes	Lawyer	0.0	High	2.0	Cat_6	B
4	462669	Female	Yes	40	Yes	Entertainment	NaN	High	6.0	Cat_6	A
...	...	...	...	...	...	...	...	...	...	...	...
2622	467954	Male	No	29	No	Healthcare	9.0	Low	4.0	Cat_6	B
2623	467958	Female	No	35	Yes	Doctor	1.0	Low	1.0	Cat_6	A
2624	467960	Female	No	53	Yes	Entertainment	NaN	Low	2.0	Cat_6	C
2625	467961	Male	Yes	47	Yes	Executive	1.0	High	5.0	Cat_4	C
2626	467968	Female	No	43	Yes	Healthcare	9.0	Low	3.0	Cat_7	A

10695 rows × 11 columns

Fig 3. Before Get dummy

	Segmentation	Male	Yes	Yes	Doctor	Engineer	Entertainment	Executive	Healthcare	Homemaker	...	High	Low	30-40	40-50	50-60	60+	10+	5-10	3-6	6+
0	D	1	0	0	0	0	0	0	1	0	...	0	1	0	0	0	0	0	0	1	0
1	A	0	1	1	0	1	0	0	0	0	...	0	0	1	0	0	0	0	0	1	0
2	B	0	1	1	0	1	0	0	0	0	...	0	1	0	0	0	1	0	0	0	0
3	B	1	1	1	0	0	0	0	0	0	...	1	0	0	0	0	1	0	0	0	0
4	A	0	1	1	0	0	1	0	0	0	...	1	0	0	1	0	0	0	0	0	1

Fig 4. After Get dummy

As you can see three dummy variables are created for the three categorical values of the temperature attribute. We can create dummy variables in python using get_dummies() method.

E. Model Building

The proposed system for customer segmentation using machine learning algorithms. There are many machine learning algorithms, each reasonable for a particular sort of issue. Furthermore going to use classification algorithm like xgboost algorithm, Random forest algorithm, Logistic regression, Support vector machine. All algorithm check against one another and the one xgboost which is best among all will be chosen for model building. After the model is build then that model will be deployed in Flask framework.

F.Prediction

The proposed feature extraction method consists of eight features, they are gender, married, graduated, profession, spending score, age, family count, experience. And giving input based on the above features to the model, then the model which has been deployed in flask framework work into it and segment the new customers into the right group for an automobile company, so the company can adopt the specific proven marketing strategy to each of them and be more successful in the business.

V RESULT

Comparative Study of Different Algorithms

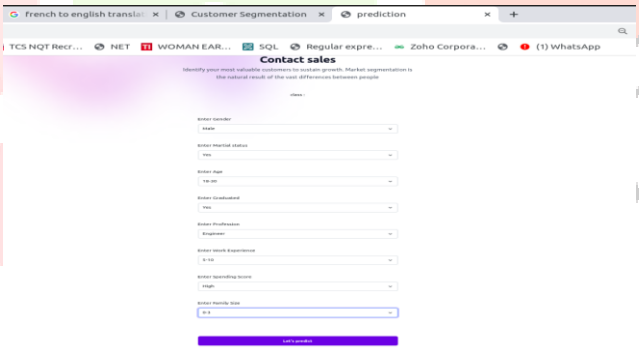
Model for building the Customer segmentation using the machine learning gives the accurate result. To make sure the model is accurate and dependable in segment the customer, it is crucial to assess its performance on test data once the model has been trained. The effectiveness of the model may be assessed using metrics such as accuracy score. The performance and accuracy of the model may be increased by tweaking the hyper parameters or utilizing other feature selection methods. The Xgboost algorithm gave the best accuracy (49.18%).

Algorithm name	Accuracy
decision tree	45.58
Random forest	45.48
SVM	48.48
logistic regression	47.68
KNN	40.06
Xgboot	**49.18**

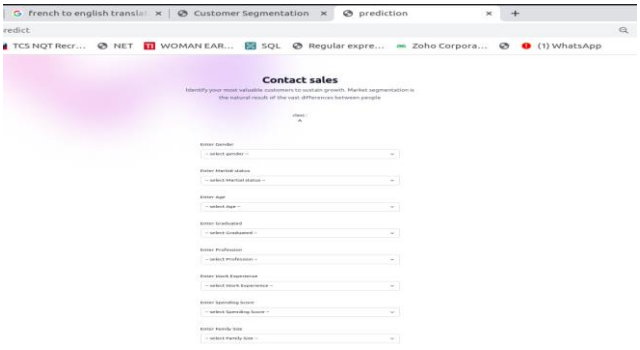
Table-2: Comparison of Algorithms

Implementation Output:

First, we have prediction page where the business/organization user can fill the form with customer demographic information to segment the new customers into the right group for an automobile company, so the company can adopt the specific proven marketing strategy to each of them and be more successful in the business.



In the same prediction page where it shows the result of segmentation of customers.



VI CONCLUSION

We approached customer segmentation problem from a demographical aspect with the number of family members, gender, graduated, profession, married, spending score, age, experiences for each customer and found out the users info which might help the automobile company to expand their business. At first, we had segment customer with demographic information using K means algorithm and did not know if they belonged to the right group. But, using xgboost algorithm for classifying the new customers into the right group for an automobile company, so the company can adopt the specific proven marketing strategy to each of them and be more successful in the business

VII FUTURE WORK

The below enhancements can be made in future.

1. Not only Focussing on demographic segmentation , in future we can focus on other a segmentation too.
2. Dealing with more categories of classification categories so that we can get better results.
3. Expand work on any other platforms other than automobile company.

ACKNOWLEDGMENT

We would acknowledge our guide for development of the Niche Prediction For Automobile Manufactures with full support and guidance. We would like to thank all the supporters who contribute their data to study.

REFERENCES

[1]Guido Van Rossum. (2007), “Python programming language”, In USENIX annual technical conference, volume 41, page 36.

[2] Travis E Oliphant. (2006), “A guide to NumPy, volume 1”, Trelgol Publishing USA.

[3] Wes McKinney. (2015), “Pandas, python data analysis library”, see http://pandas. pydata. org.

[4] Niyazi Ari and Makhamadsulton Ustazhanov (2014), "Matplotlib in python", In 2014 11th International Conference on Electronics, Computer and Computation (ICECCO), pages 1–6. IEEE.

[5] Raul Garreta and Guillermo Moncecchi. (2013), "Learning scikit-learn: machine learning in python", Packt Publishing Ltd.

[6] Michael Waskom. (2020) ” seaborn documentation”, https://seaborn.pydata.org/ introduction. html.

[7] Sarangi, Susanta. Sahidullah, Md. Saha, Goutam (2020), "Optimization of data-driven filter bank for automatic speaker verification", Digital Signal Processing .

[8] Alpaydin. Ethem. (2017) “Introduction to Machine Learniing” London: The MIT Press. p. 110 . ISBN.

[9] Kedar, Potdar. Taher, S, Pardawala. and Chinmay, D, Pai. (2017) “ A comparative study of categorical variable encoding techniques for neural network classifiers”, International journal of computer applications.

[10] Vijilesh, V. Harini, A. Hari Dharshini, M. Priyadharshini R. (2021) "CUSTOMER SEGMENTATION USING MACHINE LEARNING", International Research Journal of Engineering and Technology Volume: 08 Issue: 05.

[11] Zarah, Shibli. Wajd, Alturki. Mashael, Alhassan. Lama. (2021) "Customer Alzahrani, Segmentation: Unsupervised Machine Learning Algorithms In Python", Towards Data Science.

[12] Natassha. (2021) "Customer segmentation with Python", Natassha Selvaraj.

[13] Mrinal, sigh, Walia. (2021) "How To Solve Customer Segmentation Problem With Machine Learning", Analytics Vidya.

[14] Gomathy, C.K. Kuncham, Pavan, Kumar Reddy Kondakandla, Srikar. Siva, Sankar, K. (2022) "Customer Segmentation Techniques", International Research Journal of Engineering and Technology.

[15] Jay Prakash Bind. (2022) "Loan Prediction", Github.

[16] Bahari saravi, seyed Reza. (2012) ”New approach clustering algorithm for customer segmentation in automobile retailer industry”, International Research Journal of Applied and Basic Sciences.

[17] Balmeet kaur. Pankaj kumar sharma. (2019) “Implementation of Customer Segmentation using Integrated Approach”, International Journal of Innovative Technology and Exploring Engineering (IJITEE).

[18] Yossi Hadad. Baruch keren. (2022) “A decision-making support system module for customer segmentation and ranking”, Expert Systems.