RESEARCH ARTICLE **SPECIAL ISSUE**

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 05S May
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.S006

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

Customer Segmentation in Tourism Industry using Machine Learning Models

Vikram S [1], Gaurav Kumar[2], Vishwas T [1], Premsanth M [1], Vinodh N [1]

[1]Student, Department of Computer Science and Engineering, Dayananda Sagar University, Karnataka, India.
[2]Assistant Professor, Department of Computer Science and Engineering, Dayananda Sagar University, Karnataka, India

Email: vikramsathyamurthi@gmail.com

Article History

Received: 22 February 2023
Accepted: 6 March 2023

Keywords:

segmentation;
analysis;
bayesian;
regression;
unsupervised;
clustering;
propagation;
accuracy

Abstract

Manual segmentation of customers consumes a lot of time, in some cases months, even years to break down information and track down patterns in it. Customer Segmentation done through machine learning models result in quick identification of the ideal customers. This research paper focuses on the tourism industry to target the right customers for their business. By using the tourism dataset of customers, the research paper aims to produce a better decision making visualization patterns through histogram, pie charts, and heatmaps. Moreover, the use of Bayesian Inference Model, Descriptive Basic Analysis and Linear Regression Analysis only on the important attributes makes the decision making for the tourism business quite easy. Finally, the use of clustering unsupervised machine learning models on the dataset generates the primary, secondary, and tertiary group of customers that the company can target for the sale of their tourism packages. Clustering models will generate clusters as the output where each cluster showcases a group of customers. The clustering models employed under this research are K-means, DBSCAN, Affinity Propagation, Mini Batch K-means and Optics Algorithm. The result showed that the Mini Batch K-means algorithm had a better accuracy score for the segmentation than other algorithms used.

1. Introduction

Customer segmentation is the method involved with separating the client base into a few groups of people who offer market likenesses such as gender, age, interests and different ways of managing money. Organizations accept that every client has various prerequisites that require explicit promoting endeavors. Companies strive for deeper access to their target customers. Therefore, their purpose of use must be precise and adapted to the needs of each individual customer.

In addition, with the help of collected data, companies can gain a deeper understanding of customer preferences such as and requirements to find valuable segments that would bring maximum profits to companies. In this way they can more effectively strategize their promoting procedures and limit the risk of their investments. The customer segmentation technique relies on several key differentiators that divide the customers into target groups. Information related to demographics, geographic space, economic status and behavioural patterns play an essential part in deciding the direction of business to different segments.

With all this in consideration, the implementation of customer segmentation in the tourism industry opens an opportunity for the companies to target the primary, secondary and tertiary group of customers easily based on the data available with them. This paper opens up the opportunity for the product based companies to enhance their sales by targeting the required customers only.

1.1. Problem Definition

Today, we can modify everything. There is nobody size-fits-all methodology. However, for business, this is actually something incredible. This creates plenty of room for healthy competition and opens door for organizations to acquire innovative insights into customer acquisition and retention. One of the main steps to better personalization is customer segmentation. In the tourism industry, the services provided by the companies has a gap where all the tourism packages provided, doesn't offer a personalized approach for the customers. The packages available are the same for the customers with one-size-fits-all approach. This is where personalization starts, and the right segmentation illuminates choices about new features, new items, pricing, advertising, and even things like in-app recommendations for the customers in our industry. However, manual segmentation can be tedious. Using machine learning, we can solve this problem. Along with this, the primary solution of this research is to tackle the issue with a superior and more powerful machine learning model that has not been copied before, and make a suitable model for marketers in the tourism industry. The models will deliver clusters as the output where each cluster denotes the targeted group of customers.

1.2. Objectives

The first objective of this research is to analyze customer data from a tourism dataset of details provided by 2,773 customers over one year. The dataset contains demographic, behavioural, psychographic, and geographic data. The goal is to find similar characteristics in the groups that indicate good candidates for a marketing campaign among the population. Another goal is to use customer-personal analytics data for companies to customize their tourism product based on the target clients from various customer segments. The third goal is to develop a machine learning model that can use each person's information to classify new samples as good or bad candidates for the tourism marketing campaign. The fourth objective is to develop a machine learning model that helps the tourism companies to classify customers into a specific domain based on the degree of accuracy.

1.3. Scope

Our project identifies difficulties and offers solutions to improve the capabilities of a useful customer segment project. One of the biggest challenges in customer segmentation is data quality. Inaccurate information in source systems usually results in poor clustering (Pavithra, Prashar, and Abirami). One important aspect of data quality is the allocation of resources to manage client attributes. In addition, the research focuses on solving problems related to increasing the accuracy of targeting a segmented customer group. The different clustering methods was used namely K-means (Srijith, Kumar, and Philip), DBSCAN, Affinity Propagation, Optics Algorithm and Mini-Batch K means algorithm to achieve the best accuracy of the model.

2. Methodology

The methodology of this research helps to understand the distinctions between customer groups, it is easier to make strategic decisions about product growth and marketing. The segmentation (Regmi et al.) possibilities are endless in our research and mostly depend on how much customer data we have at our disposal. We analyze the data of customers from the travel database, which contains the information made by random customers during the year. The dataset may contain empty cells or redundant data, so the data cleaning process has to be performed. Then, the goal is to find similar characteristics in the groups, which means they are good candidates for a marketing campaign among the residents. Next, we use customer profile data for companies to customize the product based on target clients in distinct customer segments. The customer analysis includes Bayesian Inference Statistics, Descriptive Statistics and Linear Regression Analysis at a basic level. Moreover, visualization of the data by the use of histogram, pie charts and heatmap, paves the way for ideal decision making by the company. The final step includes the use of the unsupervised machine learning models namely K-means, DBSCAN, Affinity Propagation, Optics and Mini Batch K-means

algorithm to determine the clusters, and find the best fit algorithm that shows the better accuracy score among the others used.

2.1. Tourism Dataset

The tourism dataset used for this research consists of 2773 rows and 13 columns. The 13 parameters considered for this research based on the dataset includes name, age, gender, place, state, type of traveller, kind of traveller, average number of times you travel in a year, Are you a solo traveller or a group traveller?, Preferred Mode of Transportation most times, Preferred Accomodation most times, average expense of touring each year (in Rs.) and average money saved each year for touring (in Rs.). These attributes in the data helps to produce visualization patterns, statistical analysis and customer segments (Jayasen and Nandapala) easily.

2.2. Data Cleaning Procedure

After importing the dataset using the pandas library, it is important to find the existence of any missing values in the dataset. The data cleaning procedure involves removing the entire row having any missing value. By this procedure, the subsequent activities like visualization, statistics and models implementation can be performed efficiently with good accuracy. The box plots show the outliers but they are not removed because these are the true outliers present in the data.

2.3. Visualization Patterns

The visualization includes the use of histogram, box plots, pie charts and heatmap that show patterns between the multiple attributes. The python library used for this approach was 'seaborn'. The 'matplotlib' package was used to plot the graphs. The first pattern drawing includes the histogram on all the numeric values of the dataset. The histogram on attributes like age, average number of times you travel in a year, average expense of touring each year (in Rs.) and average money saved each year for touring (in Rs.). Then, the box plots were used to demonstrate the outliers in the multiple attributes of the dataset. The box plots on attributes like age, average number of times you travel in a year, average expense of touring each year (in Rs.) and average money saved each year for touring (in Rs.) were implemented. Thirdly, pie charts were used on the attributes like type of traveller, kind of traveller, Are you a solo traveller or a group traveller?, Preferred Mode of Transportation most times, and Preferred Accomodation most times. Lastly, heatmap was implemented on all the numeric values of the dataset namely age, average number of times you travel in a year, average expense of touring each year (in Rs.) and average money saved each year for touring (in Rs.). Based on the graphical analysis, a tourist company can make apt decisions for the sale of their products.

2.4. Statistical Analysis

The three types of statistical analysis implemented on the dataset were bayesian inferences, linear regression analysis and descriptive statistical analysis. The packages used for the implementation of bayesian inference model are pymc, sklearn and arviz. The bayesian inference statistics include the computation of mean and standard deviation by the use of normal distribution formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (1)$$

The symbols in equation (1)denotes: f(x) = probablity density function, μ = Mean, σ = Standard deviation and x = Normal random variable.

Secondly, the linear regression analysis was performed by the use of 'sklearn' package. The average expense of touring each year (in Rs.) was predicted based on age and average money saved each year for touring (in Rs.). Moreover, average money saved each year for touring (in Rs.) was predicted from the age attribute. All these analysis were plotted on the scatterplot graph.

Lastly, the descriptive statistical measure involved finding the mean, median, mode and bressel standard deviation of the numerical columns only. The attributes considered in this step were average number of times you travel in a year, average expense of touring each year (in Rs.) and average money saved each year for touring (in Rs.).

2.5. Machine Learning Models

The machine learning models that were used to find the clusters for our dataset included K-means, DBSCAN, Affinity Propagation, Optics and Mini Batch K-means algorithm. All these models are contained in the 'sklearn' package and directly can be implemented from it.

The K-means algorithm works well when some outliers are removed from the dataset. The outliers were removed based on the absolute z-score values. If the absolute z-score value is less than 3, then those values are filtered out as the outliers and are removed from the dataframe. The K-means assists with deciding the ideal number of clusters to be used for the model. Then, by using the cluster value, K-means is implemented on the filtered data. The accuracy of the K-means algorithm is calculated by using the 'silhouette score' function. The bar plot is used to visualize the clusters effectively and easily.

The DBSCAN (Wang et al.) algorithm can achieve better accuracy with the presence of outliers also. While implementing this algorithm, it is not mandatory to remove the outliers from the dataframe because it can work well in the presence of noise data. Here, the DBSCAN algorithm takes the epsilon value as 12.5 and minimum samples value as 1900. Then, it is fit for the unfiltered data containing the outliers. The algorithm is used to compute scatterplot clusters for 2 attributes average expense of touring each year (in Rs.) and average money saved each year for touring (in Rs.) based on the attribute age.

The Affinity Propagation algorithm has a parameter called preference and it is set to -50. This algorithm uses the blobs of points (make_blobs) data as the input. The samples value is set to 1900 and the data is fit to this algorithm. The scatterplot graphic represented clusters are generated for the attributes average expense of touring each year (in Rs.) and average money saved each year for touring (in Rs.) based on the attribute age.

The Optics algorithm works similarly to the K-means algorithm. It mandatorily uses the outlier removed data and the outliers are removed based on the absolute z-score values. If the absolute z-score value is less than 3, then those values are filtered out as the outliers and are removed from the dataframe. The filtered data is fit for the Optics algorithm and the cluster labels are expected as the output.

The Mini Batch K-means algorithm is used to compute large datasets easily. This algorithm breaks the dataset to batches where the number of batches is determined by the programmer. The quantity of clusters value is set to 6 and the batch size value is set to 15. The scatterplot cluster output is generated for the attribute average expense of touring each year (in Rs.) based on the attribute age.

3. Results and Discussion

The results show that Mini batch K-means algorithm is best suited for segmenting the customers of the tourism industry. It gives an accuracy score of 91%. The 6 cluster groups were generated where the lower valued clusters are the group of customers to be targeted. So, the cluster group numbered '0' is the primary customers that the company must target. The cluster group numbered '5' should be the least focused group of customers.

The histogram charts for the Avg. Expense of touring each year (Rs.) attribute (X-axis) VS the count of people present in the dataset (Y-axis) are displayed as shown in the FIGURE 1.

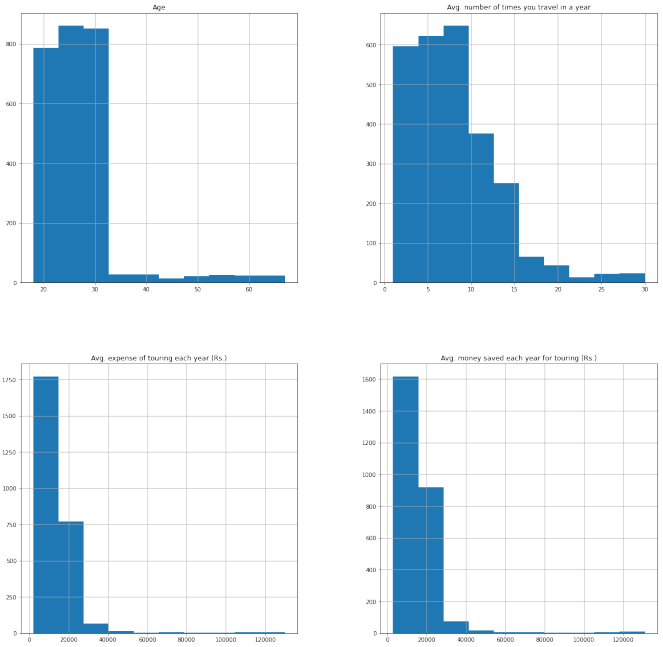

FIGURE 1. Histogram chart for Avg. Expense of touring each year (Rs.) attribute.

The box plots for the numerical values in the dataset are displayed as shown in the FIGURE 2.

Based on the graphical patterns, decision making can happen effectively in the companies. Just by visualizing the charts, the engineers can suggest the next steps for the company.

Before the data cleaning process, the dataset contained 2773 rows in it. After removing the rows having empty cells, the dataset was reduced to 2660 rows. Later, the outliers were removed based on the Z-score values resulting in the reduction of rows to 2553.

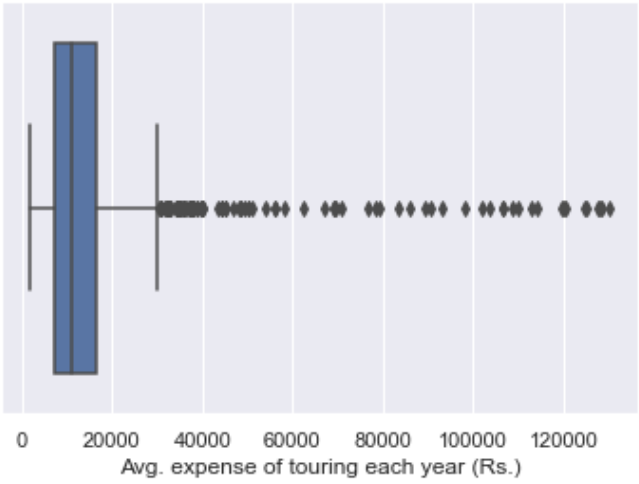

FIGURE 2. PIE chart for the Avg. Expense of touring each year (Rs.) attribute.

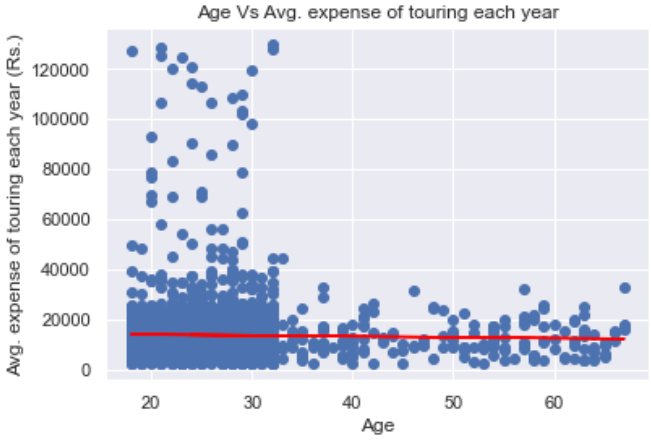

FIGURE 3. Linear Regression model predicting Avg. Expense of touring each year (Rs.) based on Age attribute.

Linear Regression model shows the results through scatterplot as shown in FIGURE 3.

Descriptive Statistics performed on the dataset will show the results of mean, median, mode and Bressel Standard deviation. The mean for the attributes average number of times you travel in a year, average expense of touring each year (in Rs.) and average money saved each year for touring (in Rs.) shows the value as 7.74, 13559.58 and 15778.38 respectively. The median for the attributes average number of times you travel in a year, average expense of touring each year (in Rs.) and average money saved each year for touring (in Rs.) shows the value as 7, 11000 and 13500 respectively. The mode for the attributes average number of times you travel in a year, average expense of touring each year (in Rs.) and average money saved each year for touring (in Rs.) shows the value as 7, 9500 and 12500 respectively. The Bressel Standard deviation for the attributes average number of times you travel in a year, average expense of touring each year (in Rs.) and average money saved each year for touring (in Rs.) shows the value as 5.24, 12227.86 and 12255.66 respectively.

The K-means algorithm shows bar chart clusters as shown in FIGURE 4.

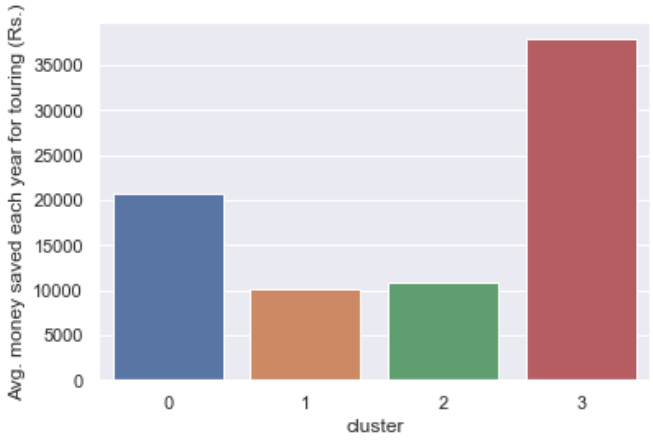

FIGURE 4. K-means generates clusters for the Avg. Money saved each year for touring (Rs.) attribute.

The primary group of customers are decided based on the lower value of the cluster. The primary group of customers are the ones present in the cluster '0'. The customers present in the cluster '3' are the least targeted customers. The 'silhouette score' generated for the K-means algorithm is 0.36. This means to say that it is an average model to be implemented for the tourism dataset used.

DBSCAN algorithm works on the noise dataset fluently and shows the clusters as shown in FIGURE 5.

The DBSCAN algorithm shows an average accuracy score than other models implemented. DBSCAN shows both the clusters and the true outliers present in the data. DBSCAN algorithm gives an accuracy score of 77% while operating with the tourism dataset.

Affinity Propagation algorithm shows different number of clusters when compared with other algorithms used. The scatterplot is shown in the FIGURE 6.

Affinity Propagation algorithm shows an accuracy score of 85%.

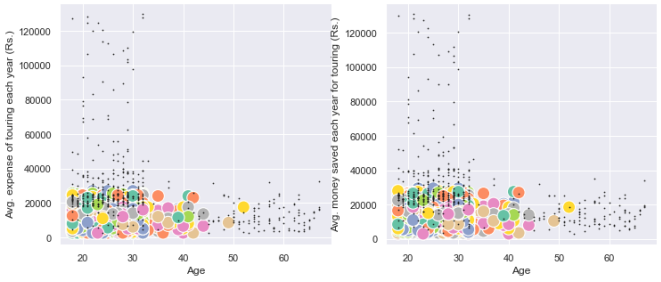

FIGURE 5. DBSCAN algorithm generates clusters for the Avg. Money saved each year for touring (Rs.) based on the age attribute.

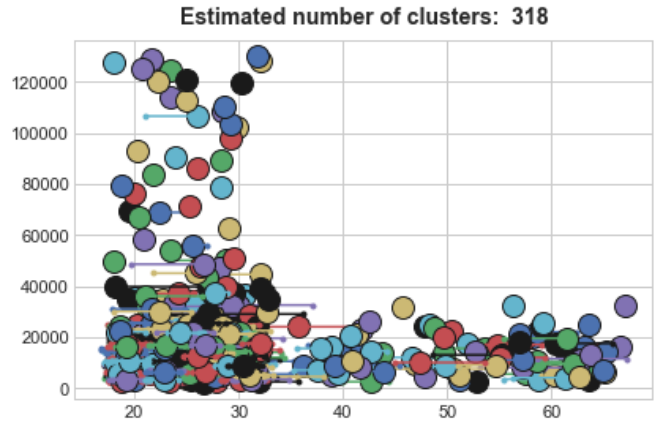

FIGURE 6. Affinity Propagation algorithm generates clusters for the Avg. Money saved each year for touring (Rs.) based on the age attribute. Estimated number of clusters generated is 318.

Optics algorithm works similarly to DBSCAN algorithm but it takes filtered data as the input. It cannot operate effectively in the presence of noise data. Optics algorithm produces cluster labels as the output. The cluster labels are stored in a one-dimesional array in the order of [0 6 4 ..... 11 9 8]. The accuracy score generated by the optics algorithm is 80%.

Lastly, the implementation of the Mini Batch K-means algorithm resulted in the best accuracy score when compared with other models used. The algorithm produced 6 cluster groups and indicated to which cluster group each individual in the dataset belonged. The scatterplot representation of the cluster group is shown in the FIGURE 7.

The mini batch K-means algorithm generated an accuracy score of 91% and is considered as the best fit algorithm while working on the tourism dataset. The mini batch K-means performs the best because it divides the dataset into fixed sizes of small multiple datasets. Each time, one of the small dataset is taken and K-means is applied on it to create the clusters. Here, each dataset is saved in the memory for the computational purposes. Moreover, the algorithm takes less computational time to generate the clusters. As a result, mini batch K-means gives the best results for our dataset.

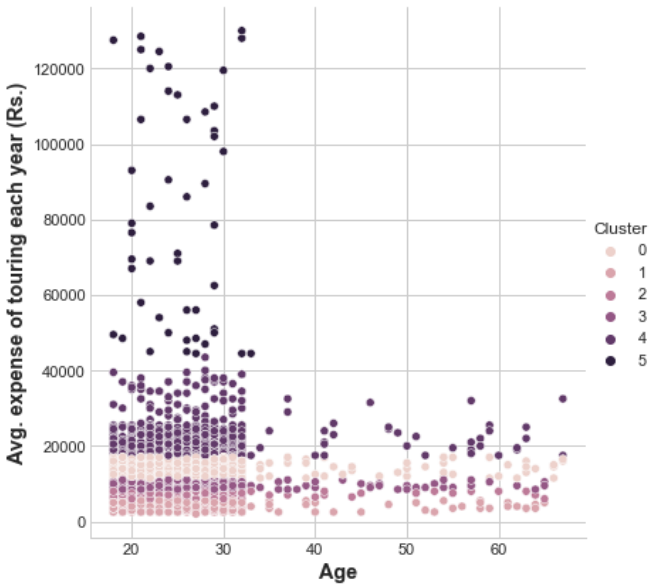

FIGURE 7. Mini Batch K-means algorithm generates clusters for the Avg. Expense of touring each year (Rs.) based on the age attribute. Totally 6 cluster groups were generated.

4. Conclusion

The outcomes obtained from the study demonstrates that the mini batch K-means algorithm is efficient and easy to implement while working on a large tourism datasets. The tourism company can target customers based on the cluster groups generated by the algorithm. The company can create multiple budget sized tour packages based on age, expenses and money saved for touring by their users. So, based on their budget the users can choose the tour package of their own choice. As a future work, the categorical data can be converted to numerical data and analysis can be drawn from it. Moreover, based on the cluster group the company can assign multiple marketing campaign for each of the cluster. By implementing this method, the company can acquire large number of customers and target specific users only for their desired marketing campaign.

Authors' Note

We declare that there is no conflict of interest

regarding the publication of this article. We confirm that the paper is free of plagiarism.

References

Jayasen, KPN and E.Y.L Nandapala. "The practical approach in Customers segmentation by using the K - Means Algorithm". *IEEE 15th International Conference on Industrial and Information Systems (ICIIS)* (2020). 10.1109/ICIIS51140.2020.9342639.

Pavithra, M, Ayushman Prashar, and Abirami. "Maximizing Strategy in Customer Segmentation Using Different Clustering Techniques". *2022 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)* (2022). 10.1109/SPICES52834.2022.9774200.

Regmi, Jasraj Swesh Raj, et al. "Customer Market Segmentation using Machine Learning Algorithm". *6th International Conference on Trends in Electronics and Informatics (ICOEI)* (2022). 10.1109/ICOEI53556.2022.9777146.

Srijith, J, Abin Oommen Kumar, and Philip. "Achieving Market Segmentation From B2B Insurance Client Data Using RFM & K-Means Algorithm". *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)* (2022). 10.1109/SPICES52834.2022.9774051.

Wang, Xuejin, et al. "Electricity Market Customer Segmentation Based on DBSCAN and k-Means : —A Case on Yunnan Electricity Market". *2020 Asia Energy and Electrical Engineering Symposium (AEEES)* (2020). 10.1109/AEEES48850.2020.9121413.

Embargo period: The article has no embargo period.

To cite this Article: , Vikram S, Gaurav Kumar, Vishwas T , Premsanth M , and Vinodh N . "**Customer Segmentation in Tourism Industry using Machine Learning Models**." International Research Journal on Advanced Science Hub 05.05S May (2023): 43–49. http://dx.doi.org/10.47392/irjash.2023.S006