International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 4, July-Aug-2021, ISSN (Online): 2395-566X

Comparative Analysis of Machine Learning Algorithms To Perform Customer Segmentation

Stuti Arora, Sunanda Mandal, Sasi Rekha Sankarl

Department of Software Engineering
SRM Institute of Science and Technology, Chennai
Kattankulathur, Chennai ,India

Abstract- Considering the current pandemic, companies are finding it difficult to reach out to their customers and target them. Given a dataset of customer details pertaining to any company, it would be possible to come up with a generalized machine learning algorithm (by comparing K-Means Clustering, Principal Component Analysis and DBSCAN) keeping in mind the factors - age, gender, annual income and spending score. Using this, we can identify the most profitable segment of customers, hence ensuring efficient utilization of funds, workforce and time and adding to the overall brand value of the company. Using the age group of the most profitable segment of customers, the concept of generational marketing can be applied in order to narrow down which marketing channel would be the most effective to target customers on the basis of their age.

Keywords- K-Means Clustering, Principal Component Analysis, DBSCAN, Generational Marketing, Segmentation

I. INTRODUCTION

Taking the current pandemic situation into consideration, people have started to develop an inhibition towards buying products (be it from stores or even online), and have also started to cut down on the frequency of usage of services (that require people to interact, in most cases). With the governments of various countries deciding to open up all companies, businesses and services, there are customers who are still very sceptical about accessing these. This in turn is resulting in major losses for the companies/businesses/service providers, because they still have to make ends meet in terms of costs including paying remuneration for the following -

- Their Rent,
- Their Employee Salaries,
- Their Electricity Bills,
- Their Transportation Costs, Etc.

The customers are unsure of the following factors, considering they may not have been informed about the same:

- About Whether Or Not The Companies/Businesses/Service Providers Are Open,
- About Whether Or Not They Are Following Required Norms (For Social Distancing And Hygiene),
- About Whether All Products And Services Are Available And,
- If There Are Any Offers That May Apply To The Customers

Hence, the most plausible solution for these companies/businesses/service providers would be to reach out to their customers using specific marketing channels and deliver required messages pertaining to the fact that they are open for work, they are following required norms and are also providing additional discounts and special services. But, one thing that cannot be taken out of the picture is the fact that it would require money in order to reach out to the customers.

Hence, it must be well understood by the company/business/service provider as to which customer group they should target for the same. This would mean that our aim is to derive the most revenue-generating group for the company/business/service provider in question and to then further evaluate which marketing channel would be the most appropriate to reach out to their customers. This would be necessary because the company/business/service provider would already have a lot of costs that would be involved in getting things back on track, hence it is important for them to invest in reaching out to their customers in a manner that would be the most beneficial to them while utilizing minimum costs for the same.

Based on previous data available pertaining to the customer base of the company/business/service provider, it would be possible to evaluate the most revenue generating customer group on the basis of factors such as:

- Age,
- Gender,
- Spending Score,
- Annual Income, Etc

and by applying any of the following segmentation algorithms:

- K-Means Clustering.
- PCA - Principal Component Analysis.
- DbScan.

The accuracy of these algorithms can be improved by applying the Adaboost classifier algorithm.

Once the customer segment has been identified, it would be possible for us to evaluate which marketing channel would be the most beneficial to reach out to them. In this case, age would play the most important factor - as age would be the distinction criteria with regard to which marketing channel one would prefer. The evaluation of this can be done on the basis of a real-time survey taken, which indicates the customer groups as per their age and the marketing channels they prefer. The survey would support the idea that technology is more accessible to the younger generation, and not so much to the older generation in comparison. Hence, applying this supporting data to the customer segment identified can help a company/business/service provider understand targeting which customers at such a time would be the most beneficial.

The aim would be to make the model generalized so that it can be adapted and used for any company/business/service provider, as required by them. Furthermore, it would be possible for the companies/businesses/service providers to understand which segments they could potentially gain new customers from, and also move further to understand which marketing channels would be the most beneficial in terms of reaching out to them, rather than following the generalized trial and error mechanism to reach out to prospective customers and convert each of them - which many a time requires a lot of time, money and effort.

II. REQUIREMENTS GATHERING METHODS

Requirement gathering methods used include:

- Surveys - to find out the age and the preferred marketing channel.
- Polling - to find out if people are sure about whether or not the rules of procedure are being adopted in nearby shops/showrooms to ensure safety and hygiene.
- Interviews - to find out the age, gender and which product an individual is most-to-least likely to buy among the industry.

III. FUNCTIONAL REQUIREMENTS

FR01	Use only relevant data for processing
FR02	Identify the most revenue generating segment of customers
FR03	Must take into account necessary factors for segmenting customers
FR04	Minimize the time and effort to segment data
FR05	Perform all functionalities within optimum costs
FR06	Improve Return on Investment
FR07	Choose the most effective marketing channel
FR08	Contribute to growth in customer base
FR09	Continually refine the most beneficial customer segment
FR10	Improve customer service

IV. NON-FUNCTIONAL REQUIREMENTS

- Correctness - The models give results with accuracies in the range of 88-97%, and the results helped identify the most beneficial customer segments and the most beneficial marketing channel to target them.
- Efficiency - The models are run on an online collaboration platform and hence, the space requirements for the same are minimised and maximum possible efficiency is achieved.
- Integrity - Only people who have access to the collaboration platform can make changes to the underlying code. The accesses can be changed, as per the requirement of the system, hence maintaining the integrity.
- Reliability - The model works with minimum possible time requirements in order to run the entire set of algorithms and produce reliable results.

Usability - Minimum knowledge is required to use the system. Clear indications of instructions are

- Made in the comments alongside the code, to ensure that any first time user is aware of the steps.
- Maintainability - The system would not have to be retrained each time the existing databases are updated. Whenever there is an inflow of more customer details, the process will be able to run on the basis of the prior training it has.
- Testability - Considering each of the test steps are followed and a comprehensive assessment of the

defects is done, the testability of the system can hence be accounted for.

- Flexibility – In case of a larger input of the same set of databases being feeded into the algorithms, the system will be able to adapt and produce results accordingly.
- Portability - Since the underlying code being used is in Python, it can easily be ported and used on any operating system by simply installing the required Python libraries. No specific hardware is required in order to run the system.
- Reusability - All modules of the system can be used for other projects with minor modifications in the code, hence making it highly reusable.
- Interoperability - With minimal changes made to the base code, the system can be interoperable on any platform that can support Python.

V. COST & EFFORT ESTIMATION

For costing, we turned to the COCOMO model (COnstructive Cost MOdel). COCOMO is a regression model based on the LoC (Number of Lines of Code). The key outcomes of performing COCOMO analyse is are - Time & Effort. According to Boehm's explanation, there can be 3 types of systems:

- Organic,
- Semi-detached,
- Embedded.

In this case, we consider it to be an organic system since the team size is small and the problem is well understood and has been solved before.

Effort (E) = a(KLOC)^b **1)**

Time = c(Effort)^d **2)**

Persons Required = Effort / Time

Using the values of a = 2.4, b = 1.05, c = 2.5, d = 0.38 (for Organic), and KLOC = 1

Effort (E) = a(KLOC)^b **3)**
= 2.4 * (1) ^ 1.05
= 2.4 Person-Months

Time = c(Effort)^d
= 2.5 (2.4) ^ 0.38
= 3.486 Months

Persons Req = Effort / Time **4)**
= 2.4 / 3.486
= 0.688
~= 1 Person

VI. RISK ANALYSIS

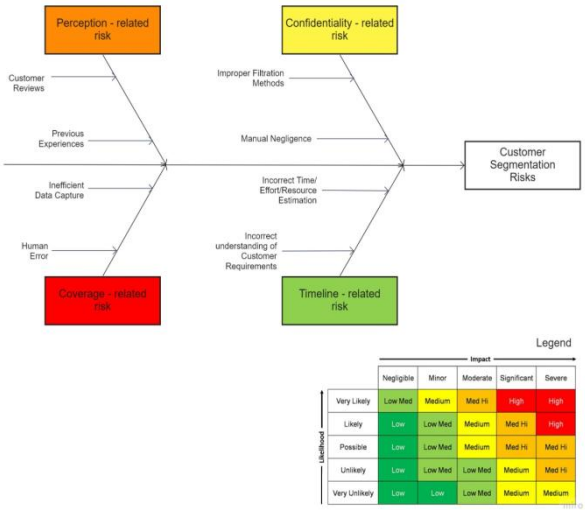


Fig.1. Fishbone analysis diagram

Risk Analysis was performed using the method of Fish-bone Analysis and the corresponding impacts were estimated. Risk mitigation strategies for each of the above are as follows:

1. **Confidentiality-related risk:** Once the customer's data is cleaned, each customer can be assigned a "de-identified" number - to be able to reference, without disclosing their actual identity.
2. **Timeline-related risk:** To ensure that effective time, effort and resources are allocated and an narrowed-down estimation is given to the client company/business/service provider (so that even if there are chances that some re-work may have to be done, the time to do that would also be accounted for as a buffer).
3. **Perception-related risk:** We can ensure that the promotional messages being sent to the customers emphasize the "quality" and "warranty period" of our products/services to ensure that the perception created about them makes them come across as reliable. Additionally, emphasis needs to be placed on the fact that standard rules of procedure are being adhered and there would be no compromise in the same.
4. **Coverage-related risk:** We can use a MECE approach - whereby before the analysis is performed, there is a comprehensive check for Mutually (M) Exclusive (E), Collectively (C) Exhaustive (E) approach which ensures that all the customer data is such that the different segments don't overlap within each other, and put together they represent the entire population of the customer base sufficiently.

International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 4, July-Aug-2021, ISSN (Online): 2395-566X

VII. SYSTEM ARCHITECTURE

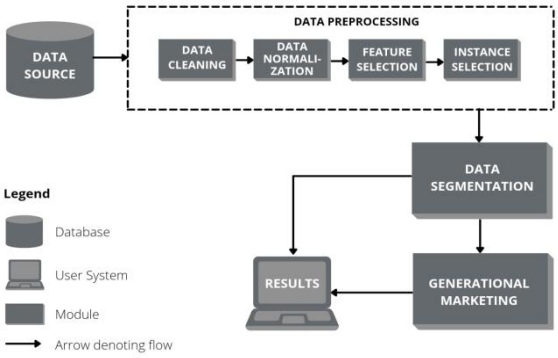
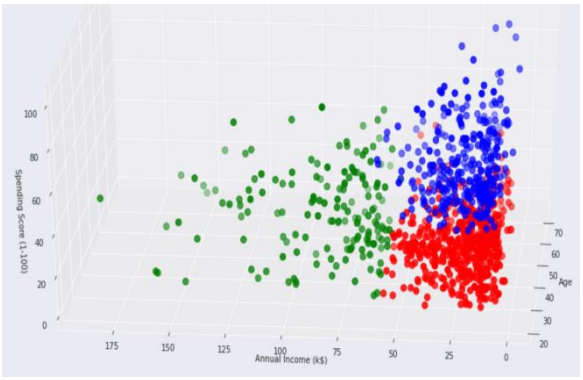
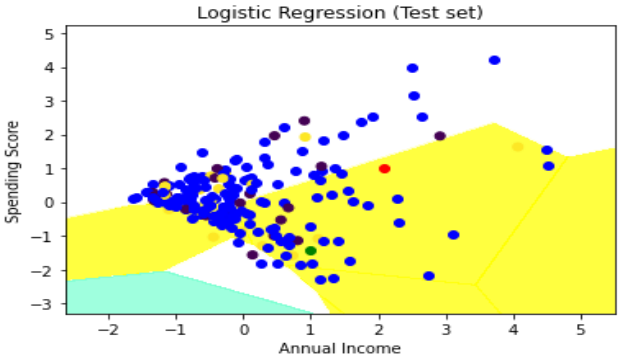


Fig. 2. System Architecture Diagram



The system architecture diagram includes the steps of data preprocessing - which in turn includes data cleaning, data



normalization, feature selection and instance selection. Post that, the steps of data segmentation and generational marketing are put into place, whereby the results are computed that give the most beneficial segment and the most ideal marketing channel to target them.

VIII. TOOLS & TECHNOLOGIES

- **Tools:** Google Collaborate - a web IDE for writing Python codes, Kaggle - for datasets.
- **Technologies:** K-Means Clustering, DbScan, Principal Component Analysis, Adaboost Classifier Algorithm.

IX. IMPLEMENTATION

Fig.3. Clusters formed using K-Means Clustering - a similar cluster is formed for all three datasets - Supermarket, Automobile, Banking Industry. Fig.3. Clusters formed using PCA - a similar cluster is formed for all three datasets - Supermarket, Automobile, Banking Industry.

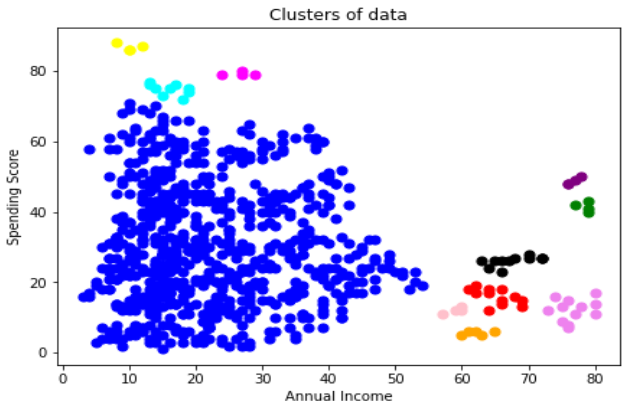


Fig.5. Clusters formed using DB Scan - a similar cluster is formed for all the three datasets - Supermarket, Automobile, Banking Industry.

The graphs to the left are the final outputs that are produced in the case of application of each of the three Machine Learning algorithms.The accuracies that each of these algorithms produced were tabulated before and after the application of Adaboost Classifier being used with them, and there was a significant increase in the accuracy of the algorithms that was seen after the application. Out of these, the K-Means Clustering method gave an optimal age group that could be the most beneficial segment in the case of each industry.

But, the problem posed by PCA was that the plot was being made in terms of a measurement of relativity of the factors of Annual Income and Spending Score, and hence could not help in narrowing down which segment of customers would ideally be the most beneficial to target in this case, Similarly, in the application of DBScan Algorithm, it was noticed that the most beneficial segment was being classified as a very major segment (spanning a major crowd) and hence, it was not narrowed down enough to be able to pick a specific category of customers - which ideally, would have defeated the purpose of the project as a whole. The most beneficial segment that was picked up from the above algorithms was checked against the below outputs for the most optimal marketing channel to target customers across different age groups. can be concluded that the most beneficial marketing channel to target customers across any of the age groups is email marketing. Ideally

International Journal of Scientific Research & Engineering Trends
Volume 7, Issue 4, July-Aug-2021, ISSN (Online): 2395-566X

of detail to any customer, irrespective of the company that is trying to target them - it would be the most ideal.

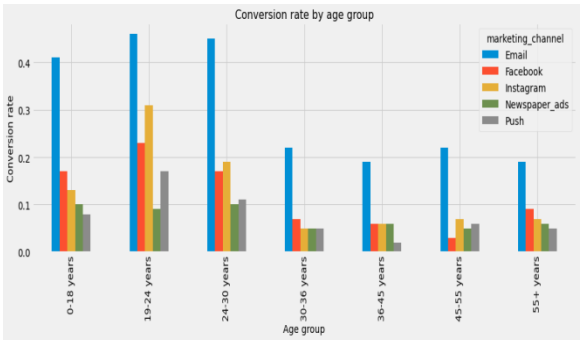


Fig.6. Conversion rate of customers for each marketing channel according to their age groups. From the above, it

X. TESTING

The idea of Unit Testing revolves around being able to individually test every functionality being performed within a piece of code. This can be done by mapping what the intended behaviour of a function within the code is vs what it actually ends up performing.

Integration testing is a phase in the process of software testing wherein individual components (as tested in Unit Testing) are combined and tested as a whole. The motive of performing integration testing is to evaluate whether or not the modules work synchronously when put together, and whether or not they are able to solve the motive that they were designed to solve. Ideally, the smallest individual units in a program are tested together in Unit Testing, after which they are grouped into larger aggregates and checked for their working.

User Acceptance Testing is a critical phase of any project and hence requires a significant level of detail to be looked into, in the case where the end user will be using the project. It is the type of testing that is done considering the perspective of the end user to ensure that the system is behaving as expected and the level of usability of the system is high.All of the above were performed on all units of the project - whereby comprehensive testing was performed and the results were in compliance with the estimated results. This would help ensure that the system is working as expected.

XI. COST AND EFFORT ANALYSIS TILL DATE

The updated lines of code were 1071 - which is 71 lines more than what was estimated. Hence, the updated calculation using the COCOMO model are as follows:

Effort (E) = a(KLOC)^b
= 2.4 * (1.071) ^ 1.05
= 2.58 Person-Months

Time = c(Effort)^d
= 2.5 * (2.58) ^ 0.38
= 3.58 Months

Persons Req = Effort / Time
= 2.58 / 3.58
= 0.72
~= 1 Person

In conclusion, the estimation and the actual values of Effort, Time and Persons required was similar. Hence, we can say that the estimation made was justified. Considering the nature of the project and the fact that open source resources were used to perform the analysis as a part of the project, there were no costs incurred in terms of labour or resources.

XII. RESULTS AND ANALYSIS

The accuracies for each of the industries are tabulated below. It can be observed that the accuracies of each of the algorithms (in the case of all the industries) increase when the Adaboost classifier algorithm is applied. The comparison can be visibly seen as below:
1. Automobile Industry:

Algorithm Name	Accuracy
K Means	88.88
Adaboost Classifier	91.11
PCA	93.33
Adaboost Classifier	96.67
DBSCAN	93.33
Adaboost Classifier	96.67

2. Banking Industry:

Algorithm Name	Accuracy
K Means	95.56
Adaboost Classifier	97.77
PCA	93.33

Adaboost Classifier	96.67
DBSCAN	86.67
Adaboost Classifier	96.67

3. Supermarket Industry:

Algorithm Name	Accuracy
K Means	88.88
Adaboost Classifier	91.11
PCA	95.55
Adaboost Classifier	97.77
DBSCAN	90.0
Adaboost Classifier	96.67

From the above tables, we can say that the average accuracy is highest in the case wherein the Principal Component Analysis (PCA) Algorithm was picked to be used along with the application of the Adaboost Classifier. Hence, for the purpose of analysis that can be applicable across multiple industries, this algorithm would be the ideal choice. We can also say that the most ideal marketing channel in the case of every age group was found to be Email Marketing - hence, together these results fulfil the purpose of the project.

REFERENCES

[1] Discovering Knowledge by Comparing Silhouettes Using K-Means Clustering for Customer Segmentation - International Journal of Knowledge Management, Volume 16, Issue 3, July-September 2020
[2] Customer Segmentation - A case of increase in customer satisfaction levels with segment-specific customer service strategies - International Journal of Computer Engineering & Technology (IJCET) Volume 8, Issue 6, Nov-Dec 2017
[3] Customer Segmentation with purchase channels and media touchpoints using single source panel data - Journal of Retailing and Consumer Services 41 – 2018
[4] Customer Segmentation from Massive Customer Transaction Data - International Research Journal of Engineering and Technology (IRJET), Volume 6, Issue 6 - June 2019
[5] The Use of Technology in Direct-Selling Marketing Channels: Digital Avenues for Dynamic Growth - Journal of Marketing Channels, Volume 24, Issue 1-2 - August 2017
[6] Customer churn prediction in telecom using machine learning in big data platform - Journal of Big Data, Springer - March 2019
[7] Domains of Digital Marketing Channels in the Sharing Economy - Journal of Marketing Channels, Volume 24, Issue 1-2 - August 2017
[8] The impact of demographic factors on supermarket shopping motivations in South Africa Department of Marketing and Retail Management, Journal of Business and Retail Management Research (JBRMR), Vol. 13 Issue 3 - April 2019
[9] Real Estate Market Segmentation: Hotels as Exemplar - Springer Journal - February 2018
[10] Starbucks Market Segmentation and Targeting - International Journal of Business and Management Invention (IJBMI), Volume 7 Issue 5 Ver. I - May 2018
[11] To study the customer satisfaction of Jio Telecom - SSRN Electronic Journal - May 2020
[12] Market Segmentation, Targeting and Positioning Strategy Adaptation for the Global Business of Vodafone Telecommunication Company - International Journal of Research and Innovation in Social Science (IJRISS), Volume 4, Issue 6 - June 2020
[13] Market Segmentation, Targeting and Positioning. In Travel Marketing, Tourism Economics and the Airline Product, Springer Journal - December 2017
[14] Retail business analytics: Customer visit segmentation using market basket data - Expert System with Applications Journal, Volume 100 - June 2018
[15] A study on factors influencing cosmetic buying behavior of consumers - International Journal of Pure and Applied Mathematics Volume 118 - 2018