International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Forecasting Customer Actions: Exploring Machine Learning Techniques for Behavior Analysis and Prediction

Sandeep Rajani[1], Dr. Birendra Goswami[2]

[1]Research Scholar, Sai Nath University, Ranchi
[2]Professor, Sai Nath University, Ranchi

Abstract
Modern modelling of today's consumer behaviour is highly dependent on data mining which mines customers' data in order to answer specific questions in time. This complexity and uncertainty relating to the consumer behaviour makes it difficult for us to predict. Therefore, selection of appropriate methods is very important. Predictive models can be designed so as to identify specific group or individual actions and thus making them useful marketing tools. However, many models are overly simplistic for they leave out some of the essential factors hence leading to inaccurate forecasting outcomes. Consequently, instead of concentrating on customer behaviour-firm capital structure relationships only a strong association rule mining model should be created using online store data that effectively estimates customer response. Knowing their business clients is crucial when it comes to profitability by aligning the company's earnings with their concerns. Artificial intelligence (AI) optimizes product location by clustering techniques targeting the right customers. It examines customer behaviour and buying trends utilizing Kaggle's client membership card records such as Customer ID, age, gender, annual income, spending score among others. This includes basic market analytics based on demographic information like age or income split into groups, however big data superiority over traditional methods cannot be underestimated including machine learning applications such as Customer Segmentation Automated Methodology (CSAM), k-means clustering for user segmentation among others. This approach facilitates focused advertising campaigns and development and launch of new products aimed at particular consumer sections

Keywords: Customer Prediction, Machine learning, Business decision.

1 Introduction

Dividing customers is one vital part of marketing tactics. it allows companies to develop more personalized marketing strategies by breaking a target market down into separate groups based on common characteristics. But conventional approaches to consumer segmentation are time-consuming, biased and error-prone. When such cases arise, businesses are at a point where AI can transform their customer segmentation approaches

Benefits of AI and ML in Customer Segmentation

Many companies with abundant data can leverage AI-driven tools whilst having some of the richest stores when it comes to consumers. Thus, acquiring regarding consumers that has never been available is possible

for them. Using AI in customer segmentation helps organizations to achieve the following:"

1. **Effectively identify target groups and personalized marketing campaign**: Technology that drives algorithms that depend on this sorts of thing involves such clients which will include similar characteristics with each one such as their age groups and even where they live across various places or even how much money people spend daily. Knowing who might purchase your product is exact science where these organizations should select their exact customers thanks these sophisticated approaches when trying to plan their strategies beforehand. Using AI companies can now refine their marketing messages hence strategies hence meeting different groups' unique needs through enhanced client division. Such customization doesn't only improve satisfaction levels overall; it also boosts the effectiveness of promotions.
2. **Improve decision-making and enhancing data accuracy**: Having access to the information about the preferences of real people helps businesses make better decisions in respect of designing new goods and positioning them in the market [Sharp, Taylor, and Wright 714-727]. Besides that, incorrect classification of customer data into distinct groups can also be avoided when using this technique.
3. **Heighten allocation of resources**: Businesses should allocate their resources effectively by focusing more on areas of their customer base that offer the highest return on investment (ROI) so that they can identify these different groups and optimize resource allocation.

Machine learning along with deep learning can help businesses analyze huge amounts of data in order to gain insights and recognize patterns which may otherwise be difficult to detect.

To find divisions of customers within an enterprise, this paper is going to use data mining methods. The set of business clients that belong to one customer base and share similar market characteristics is called a customer segment.[3] For forecasting consumer segments I would suggest using CRM data mining in the right way. CRM enables businesses to gather, store and analyze consumer-related data effectively and make it available for all staff in the organization involved in corporate business [1]. Analytical Customer Relationship Management (ACRM) is used for customer analysis within CRM whereby stored data about customers is examined through machine learning so as to identify interesting patterns among them. In ACRM, ML techniques are employed during consumer analysis. The main aim of this work is to determine which among machine learning algorithms can be considered best for solving the problem on customer analysis more effectively. The dataset used in this investigation was sourced from Kaggle repository.[2] During the experimental process several different ML algorithms were tried out and each was assessed based on its performance using various validity scores.

2. Review of the literature

2.1 Determination of Customers

In this sector, there is more competition among firms who are working towards increasing their numbers of customers through meeting their demands and desires. [4] It may not be possible to satisfy each and every customer's needs and wants since they often have different requirements, tastes, preferences based on demography, size among other things. Thus treating all consumers equally does not make business sense. The strategy of customer segmentation has been embraced to solve this problem – it involves dividing the client base into smaller groups with common traits or behaviors which makes it easier to serve them appropriately. [5]

2.2 Massive Data Repository with K-Means Data Clustering

Big data exploration has been growing over the past years which means dealing with huge amounts of

structured and unstructured information that cannot be analyzed by traditional methods. Companies gather data on their customers, vendors as well as operational activities; millions of sensors in vehicles and mobile phones are also interconnectedly collecting information. Data is important for research in various fields such as natural sciences, social sciences, humanities and business. The dataset used for this study was obtained from Kaggle repository where details about mall customers like ID's, gender, annual income and spending score were recorded. Clustering involves grouping data together based on common features and there are many algorithms used to achieve this process. Since there is no one-size-fits-all clustering algorithm it is important to select appropriate strategies for clustering depending on what you want to achieve in your analysis. This article utilises Jupyter notebook during exploratory data analysis while algorithmic hypotheticals are justified here too. K-means a popular classification algorithm is applied in segmenting customers so as to expose hidden patterns within the data thereby helping decision making processes.

3. Methodology

Customer analysis is done through many machine learning methods. The KNN method is one of the classifiers used in this process. The resulting performance can be measured using different metrics such as precision, recall, sensitivity and specificity. Finally, the best model performing at its maximum capability is selected.

3.1 Amassed Data

The phrase "amassed data" refers to the knowledge which is collected over an extended period or from various origins so as to form a complete set of information. Many sources or cases of data need to be collected, registered and arranged during this process. In other words, amassed data is what you get when you put together different events or bits of information.

This stage of data is known as the data medication phase. It's objective is to increase the efficacy and productivity of clustering algorithms by updating every datum at a fixed pace.[8] The dataset contains records on mall customers such as their client id number, age, annual income (in thousands of dollars), gender and spending score. We read in the initial data stored in the file "shop_data.csv" into pandas' DataFrame constructor function.

Clients = pd.read_csv("shop_data.csv")

We can see that we've ID, Gender, Age, Annual Income expressed as price x1000, and the spending score as we anticipated.

Clients.head( )

Table 1. shop_data.csv (first 5 rows)

ID	Ever_Married	Spending_Score	Profession	Family_Size	Gender	Age	Annual Income (k$)	Spending
462141	No	Low	Artist	1	Female	33	21	35
466286	Yes	Average	Artist	5	Female	47	22	33
462358	Yes	Average	Doctor	2	Male	34	24	33
459713	No	High	Healthcare	3	Male	27	24	30
467749	No	Low	Healthcare	9	Male	26	22	29

Duplicate rows can also be found. Luckily, we do not have duplicates. The last thing we do is consider the representation of each variable in the DataFrame. You can't work directly with categorical variables. K-means depends on distances. Different sets of category variables require different conversion methods.

3.2 Techniques for classifying customers

Methods of categorizing customers There are various ways for grouping things, and each has its own purpose, data needs and level of strictness. Research on particle identification, artificial neural networks and advanced ensemble forms was not sufficiently represented for exclusion.
We can now look at how the variables are distributed. Let's define two functions for now. The first one would receive the descriptive statistics of variables while the second one might be used to plot variable distribution.
The descriptive data will be presented to us; if a variable is not numerical we will find the counts within each category.
spending = Clients["Spending Score"]
getting descriptive statistics of spending :- statistics(spending)

Table 3 Spending Scored descriptive statistics

	Spending_Score
Low	1590
Average	639
High	369

graph_pie(spending)

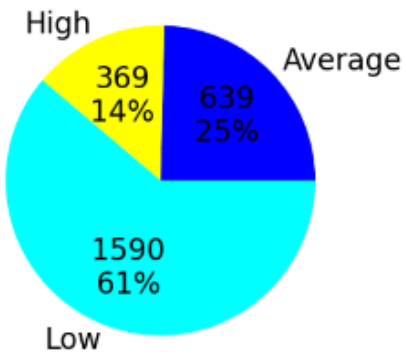


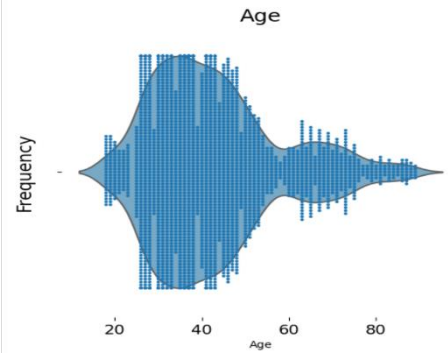
Fig. 1. Graph representing percentage of spendings

After that another important aspect i.e. Age, we'll evaluate it.
age = Clients["Age"]
statistics(age)

Table 3: mean, standard deviation, median, and variance for the Age descriptive statistics

Variable	Mean	Standard Deviation	Median	Variance
Age	42.89184	14.94404	40.0	223.324329

Fig. 2. Violin graph representing frequency chart of Age using above statistics

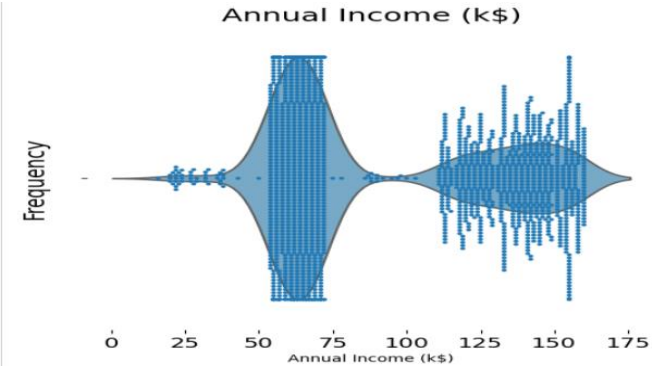


With this now we will explore the most important variable i.e. **Annual Income**
inc = Clients["Annual Income (k$)"]
statistics(inc)

Table 4: mean, standard deviation, median, and variance for the Annual Income descriptive statistics

Variable	Mean	Standard Deviation	Median	Variance
Annual Income (k$)	88.574673	37.383203	68.0	1397.503854

Fig. 3. Violin graph representing histogram of Annual Income($) in Violin form



Thinking about the customer retention we need to know the Gender statistics too because Female behaviour as per Male is totally different.
Gender = Clients["Gender"]
statistics(Gender)

	Gender
Male	1353
Female	1245

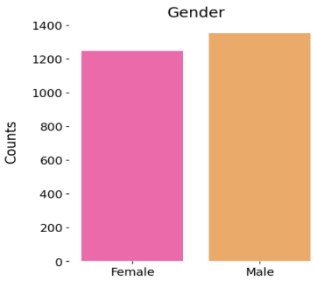


Fig. 4. Graph representing histogram of Gender

The connection between the numerical parameters will be explored and we intend to use the seaborn function pairplot for this. We'll check if there's any gender difference which may necessitate changing some aspects of the charts such as point coloration or shape in order to distinguish between males & females. In order to differentiate between female and male data points, the adjustment of the hue parameter will enable us compare them with each other using different colours.

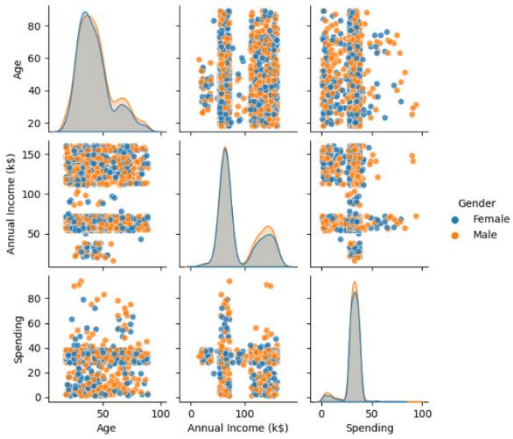


Fig. 5. Graph representing gender difference using pair plot seaborn function

The dataset contains variables that follow the normal distribution. The difference between them is small, except for age, as it varies less than other variables. After we have verified this fact we can use the k-means method then do Principal Component Analysis (PCA) to determine the dimensions that best explain the variance among these features. In the case of (0 & 1), we will dichotomize the categorical variable into two dummy variables.

Table 6: Categorial variable gender converted to 0 & 1 for present of absence

	Age	Annual Income (k$)	Spending	Male	Female
0	33	21.0	35	1	0
1	47	22.0	33	1	0
2	34	24.0	33	0	1
3	27	24.0	30	0	1
4	26	22.0	29	0	1

In order to apply Principal Component Analysis , we are going to use the function from sklearn module.
print(pca.components_)

```
[ [3.39028045e-03  -9.00378776e-04  -9.99993807e-01  -2.02365073e-04  2.02365073e-04]
 [ 9.99214467e-01  -3.94785342e-02   3.42330420e-03  -2.96710452e-04  2.96710452e-04] ]
```

```
print(pca.explained_variance_)
   [919.843159   287.3063123]
```

The Vectors that are used are made by these seemingly intangible units. The variance determined explains the squared length of vector and the component tell us about the direction of vectors.

I think a biplot, which is basically a scatter plot, could help us visualize this. Each point on the plot represents a primary component score and serves as a representation of that specific point. Additionally, the biplot can be useful in uncovering any connections between the original variables and the major components.

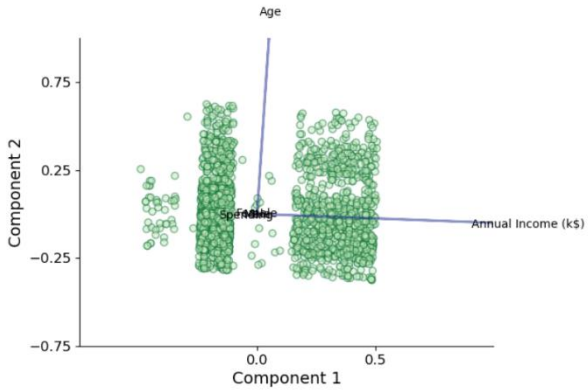


Fig. 6. Biplot representing score regarding the principal components.

We can see that the two most crucial factors are Annual Income and Spending Score.

3.3 K-Means confront and Centroids Clustering

Often, the K-means clustering algorithm is used to reveal patterns and dissimilarities within a database, as cited in [9]. It's widely employed in marketing for segmentation and understanding the interrelationships between such groups. Usually, Euclidean distance is measured by the K-means clustering algorithm to determine distances between two data points. At first, we will fix up the number of clusters that will be needed. Several methods can be used to find out the best number of clusters, including elbow and silhouette methods. The elbow method involves plotting total within-cluster sum of squares (WSS), which we want to minimize. We'll run K-means algorithm several times for different k values; in this case k = 5 and compute total WSS for each k. After that we will make a plot of WSS versus number of clusters and identify the location of an inflection point or 'elbow' which denotes optimal number of cluster(s). Furthermore, using hue parameter one may color differently those belonging to males and females.

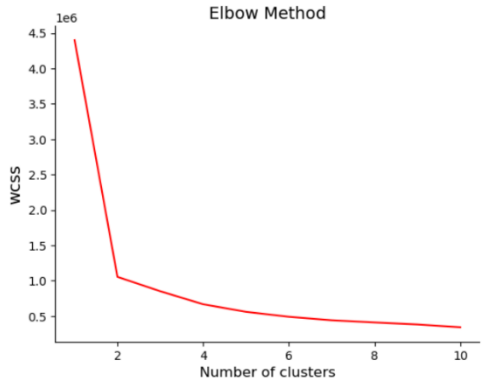


Fig. 7. Intra-cluster variation Elbow Method

In order to detect the arbitrary state, we use the K-means technique, assuming that there are 5 clusters at this time. We then execute the procedure 10 times with different centroid seeds. Our clusters appear to be:

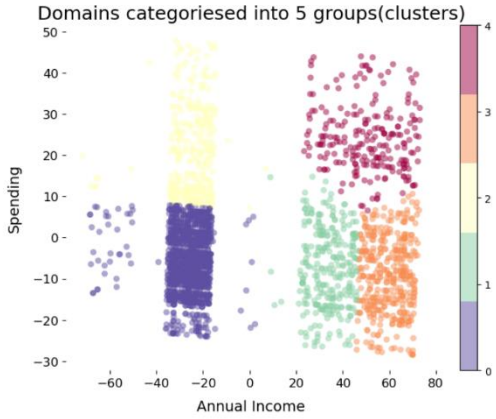


Fig. 8: Domains grouped into 5 clusters

Table 7: Centroids

	Age	Annual Income	Spending	Male	Female	ClusterID
0	35.921053	62.444361	32.403759	0.503759	0.496241	0
1	37.257951	122.660777	29.243816	0.459364	0.540636	1
2	62.467391	62.698370	32.581522	0.461957	0.538043	2
3	36.717949	148.664103	29.297436	0.456410	0.543590	3
4	69.629956	137.889868	29.431718	0.427313	0.572687	4

The score for annual income and spending appears to be the most important factor. Section 0 has individuals who are low paid yet they spend relatively the same way. Section 1: High earning people with large purchases. Segment 2 of customers who earn average but spend equal amounts. Thus, section 4 accommodates clients with high revenues of money but also the highest buying value. Group five comprises individuals who make little money but spend much. Let's assume that tomorrow we have a new member. Now, let us say, we need to know which group this person is in. This could be predicted using Kmeans predict feature as such: what if age,annual income, spending and gender(1 for male,0 for female) is equal to73,88,74,1,0

```
newClient = np.array([[43, 76, 56, 0, 1]])
new_client = kmeans.predict(newClient)
print(f"The new customer belongs to segment {new_client[0]}")
```

Prediction will be that: The new customer belongs to segment 2

4. Conclusion

AI-enabled customer segmentation can help businesses identify target groups more efficiently, create focused marketing campaigns, and make data-based decisions. In consumer segmentation AI has several benefits such as predictive analytics, improving data quality, optimizing resources; enhancing personalization and targeting; increasing accuracy as well as efficiency.

In this paper, demonstration unsupervised learning at work is done and it offer recommendations to a possible client using real world data. Today most businesses gather large amounts of information about

their clients and customers and are eager to find out important connections hidden within their clientele base. Learning such information will enable companies to develop future products and services aptly meeting the needs or wishes of its customers in its consumer base. Based on these findings, this paper gives recommendations for additional potential buyers in this section.

References

1. Dalla Pozza, I., Goetz, O., & Sahut, J. M. (2018). Implementation effects in the relationship between CRM and its performance. Journal of Business Research, 89, 391–403. https://doi.org/10.1016/j.jbusres.2018.02.004
2. Blanchard, Tommy. Bhatnagar, Pranshu. Behera, Trash. (2019). Marketing Analytics Scientific Data: Achieve your marketing objectives with Python's data analytics capabilities. S.l: Packt printing is limited
3. Puwanenthiren Premkanth, - Market Classification and Its Impact on Customer Satisfaction and Special Reference to the Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Publisher Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1
4. Puwanenthiren Premkanth, - Market Classification and Its Impact on Customer Satisfaction and Special Reference to the Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Publisher Research: Global Magazenals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue 1
5. T.Nelson Gnanaraj, Dr.K.Ramesh Kumar N.Monica. AnuManufactured cluster analysis using a new algorithm from structured and unstructured data. International Journal of Advances in Computer Science and Technology. 2007. Volume 3, No.2
6. McKinsey Global Institute. Big data. The next frontier is creativity, competition and productivity. 2011. Accessed at: www.mckinsey.com/mgi on July 14, 2015
7. Sulekha Goyat. "The basis of market segmentation: a critical review of the literature. European Journal of Business and Management www.iiste.org. 2011. ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online). Vol 3, No.9, 2011
8. A.K. Jain, M.N. Murty and P.J. Flynn.‖ Data Integration: A Review‖. ACM Computer Research. 1999. Vol. 31, No. 3
9. Vishish R. Patel1 and Rupa G. Mehta. MpImpact for External Removal and Standard Procedures for JCSI International International Science Issues Issues, Vol. 8, Appeals 5, No 2, September 2011 ISSN (Online): 1694-0814
10. V.K.G.Kalaiselvi, A. Ponmalar , Hariharan Shanmugasundaram, Bhanuprasad A , Mamathibala V, Swetha Sri M “CUSTOMER SEGMENTATION USING MACHINE LEARNING” , Vol. 7 Issue. 9 (September-2022) EPRA International Journal of Research & Development (IJRD).