2021

A Customer segmentation model using logistic regression: a case of Telkom Kenya.

Omonge, Jevans
Faculty of Information Technology
Strathmore University

A CUSTOMER SEGMENTATION MODEL USING LOGISTIC REGRESSION: A CASE OF TELKOM KENYA

Jevans Omonge

091345

A Thesis Submitted to the Faculty of Information Technology in partial fulfillment of the requirements for the award of Master of Science in Information Technology.

Strathmore University

2020

Declaration and Approval

I, Omonge Jevans Ochieng, declare that this research has not been submitted to any other University for the award of a Degree in Masters of Science in Information Technology.

Student Name: Omonge, Jevans Ochieng

Sign.....  .....Date:.....

Supervisor's Name: Dr. Bernard Shibwabo

Sign  .....Date: 10th November 2020

Abstract

Market segmentation is a marketing strategy that has been widely used by many companies globally. With the ever increasing volume of client data, many companies are now unable to clearly cluster their clients into their respective segments, subsequently providing them with products and services that are best suited for them. Telkom Kenya is currently the third largest telecommunications company in Kenya. Currently, telecommunications companies do not have well defined marketing plans for their customers based on their daily expenditure. Some companies, for instance, may provide their customers with additional voice airtime even when such customers spend significant amounts of credit on data bundles rather than the actual voice airtime. One way of overcoming this challenge is by enhancing the current state of market segmentation in telecommunication companies in general. In this study, we present an approach that incorporates business intelligence, big data and machine learning in order to achieve customer segmentation. The study is based on data collected from the spending patterns of Telkom Kenya customers. When designing the customer segmentation model, the fundamental steps in the designing of any machine learning model were followed. To begin with, data was collected from the CRM department of the company. Key trends and inferences from the data were obtained from extensive data visualization that was performed on the data. The data was then formatted to ensure that it was consistent before performing feature engineering with the primary purpose of improving the quality of the features. Thereafter, the data was split into training and testing sets. Finally, the processed data was fed into the actual machine learning models. The main classification algorithms evaluated in this study are Logistic Regression Classifier, Linear Discriminant Analysis, K-Nearest Neighbor, Decision Tree Classifier and the Guassian Naive Bayes Classifier. Of the five, Logistic Regression Classifier was found to have the cross-validation accuracy and was thus embraced for the customer segmentation process. The results of this study therefore show yet another potential application of machine learning in marketing in general through customer segmentation. As seen from the results, the machine learning has been able to categorize customers into their respective categories with 71% accuracy. Through the classification, Telkom Kenya is in a position of marketing their products and services to the right group of customers, thereby ensuring that their marketing strategies are effective.

Acknowledgment

I would like to acknowledge God for His grace, strength and good health as I undertook this research. My sincere gratitude to: my supervisor, Dr. Bernard Shibwabo for his continued commitment to guide and support this research, staff from the marketing department of Telkom Kenya, Ian Some and Adam Kipkemoi for assisting me with extraction and big data analytics and my boss Anne Wagikuyu for giving me ample time to study.

Table of Contents

Declaration and Approval ..... **ii**
Abstract ..... **iii**
Acknowledgment ..... **iv**
Dedication ..... **ix**
Chapter 1: Introduction **1**
1.1 Background Information 1
1.2 Problem Statement 3
1.3 Research Objectives 3
1.4 Research Questions 3
1.5 Justification of the Study 4
Chapter 2: Literature Review **5**
2.1 Overview 5
2.2 Market Segmentation 5
2.3 Challenges Facing Market Segmentation 7
2.4 Existing Market Segmentation Techniques 8
2.5 The Impact of Big Data on Market Segmentation 10
2.6 The Use of Machine Learning in Market Segmentation 12
Chapter 3: Research Methodology **15**
3.1 Research Design 15
3.2 Model Development 15
3.2.1 Data Collection 15
3.2.2 Data Preprocessing 15
3.2.3 Exploratory Data Analysis 16
3.2.4 Feature Engineering 16
3.2.5 Model Creation and Evaluation 17
3.3 System Development Methodology 18
3.4 Research Quality 19
3.5 Ethical Considerations 20
Chapter 4: System Design and Architecture **21**
4.1 Requirement Analysis 21

4.1.1 Functional Requirements 21
4.1.2 Non-functional Requirements 21
4.2 System Architecture 22
4.3 Use-case Diagram 23
4.4 System-sequence Diagram 25
Chapter 5: Implementation and Testing 27
5.1 Implementation 27
5.1.1 Data Collection 27
5.1.2 Data Preprocessing 27
5.1.3 Exploratory Data Analysis 28
5.1. 4 Feature Engineering 28
5.2 Model Testing 29
5.3 Marketing Algorithm 29
Chapter 6: Results and Discussion 31
6.1 Results 31
6.1.1 Age Distribution 31
6.1.2 Service Expenditure Distribution 32
6.1.3 Revenue Contribution of Different Technologies 33
6.1.4 Expenditure by Age 34
6.1.5 Revenue Contribution of Different Services in the Respective Technologies 36
6.1.6 Model Factory Results 38
6.1.7 Logistic Regression Classification Results 39
6.2 Discussion 40
6.2.1 EDA Results 40
6.2.2 Machine Learning Model Results 42
Chapter 7: Conclusion, Future Work and Recommendation 43
7.1 Conclusion 43
7.2 Future work 43
7.3 Recommendation 43
References 45
Appendix 48

List of Figures

Figure 2. 1: A general overview of customer segmentation (Jobber, 2009) 6
Figure 2. 2: A representation of how big data incorporated different customer attributes to be later used in segmentation processes (Higgins, 2017) 12
Figure 3. 1: An overview of DDM models in machine learning (s0f, 2009) 19
Figure 4. 1: Overall system architecture of the customer segmentation and targeted marketing 23
Figure 4. 2: Use-case diagram for the customer segmentation and targeted marketing system 24
Figure 4. 3: System diagram for the targeted-marketing system 26
Figure 4. 4: The overall data flow diagram of the system 26
Figure 5. 1: Conditional statement used in the targeted marketing based on predictions made by the logistic regression classifier and the EDA conducted 30
Figure 6. 1: Distribution of clients by age in the company 32
Figure 6. 2: Percentage contribution of different services to the total revenue of the company 33
Figure 6. 3: Percentage contribution of different technologies to the total revenue of the company 33
Figure 6. 4: Service consumption by the youth 34
Figure 6. 5: Service consumption by adults 35
Figure 6. 6: Service consumption by very old 35
Figure 6. 7: Service consumption by very old 36
Figure 6. 8: Revenue contribution of different bundles in 4G technology 36
Figure 6. 9: Service consumption by very old 37
Figure 6. 10: Distribution of revenue in 2G technology 37
Figure 6. 11: Cross-validation scores from the different models used in the model factory 38
Figure 6. 12: Classification report for the logistic regression classifier 39

List of Tables

Table 4. 1: Use-case diagram for the system administrator ..... **Error! Bookmark not defined.**

Table 4. 2: Use-case diagram for the marketing analyst ..... **Error! Bookmark not defined.**

Dedication

To my Dad Elisha, my dearest wife Dolphine Omonge and my loving daughters Mich and Mor,

Thank you for your continued support, understanding and prayers.

Chapter 1: Introduction

1.1 Background Information

Market segmentation is an approach commonly employed by organizations to cluster their clients into different categories based on the common attributes possessed by such clients. Market segmentation can be performed based on factors such as the purchasing traits, gender, geographical location and many more (Wedel & Kamakura, 2012). The concept of segmentation is essential since it enables most organizations to appropriately plan and meet the demands of their clients.

Customer segmentation is essentially a common practice in many large organizations. One noticeable aspect of large-scale customer segmentation is that it is implemented by organizations having a significantly large customer-base. Such client-bases makes it possible for organization to identify intrinsic properties among the clients that can be used to cluster them into given groups and subsequently deliver services to them in a manner corresponding to the specific needs of the mentioned groups (Tianyuan, 2018). Among other factors, customer segmentation has been found to increase the loyalty of clients to the companies, subsequently implying that the organizations are able to sustain their operations profitably. Customer segmentation, like every other marketing aspect in the present age, has been affected by technological changes (Melnic, 2016). The impact of technology ranges from the incorporation of social media in marketing strategies to the use of novel technologies such as machine learning to perform customer segmentation.

From customer segmentation practices implemented, organizations can implement precision marketing strategies. The primary purpose of segmentation is to cluster clients based on their similarities in terms of purchasing trends and other demographic factors. With this information at hand, companies can proceed to design specially tailored products and services that appeal

to unique segments (Cross et al., 2015). This is the basic principle of targeted marketing (Maji et al., 2019). This approach has several benefits to an organization. For instance, it presents an organization with an easier way of analyzing the customer trends at a more holistic level since different clusters will have different performances. Also, the approach makes the entire advertising process economical since the organization does not produced generalized advertisements but rather specific advertisements whose impact on the profitability of the organization is considerable.

The constantly evolving volume of customer data has however posed serious challenges to organizations which intend to successfully implement market segmentation. With increasing data, it is evident that traditionally used methods of data analysis and modeling tend to be unreliable. Therefore, more complex techniques have to be embraced by such organizations to ensure that they are able to leverage into the rich information contained within the data (Wang et al., 2018). Business intelligence, is a modern approach which has been embraced by most top organizations to mitigate this challenge. Driven by modern technology, the core mandate of business intelligence is to provide organizations with a tool that can be used to analyze and generate insights from the large data collected from their clients (Fan et al., 2015). By itself, business intelligence is a wide concept that involves aspects such as data mining, online analytical processing and business reporting. Historical or archived data plays a crucial role in determining the success of business intelligence since it is from such data that trends can be generated.

While business intelligence is mainly associated with generating insights from organizational data, it is important to consider the fact that predictive analysis is also of essence for organizations. Machine learning has in the recent past been extensively used in performing predictive analysis in many fields, marketing included. Therefore, when it comes to targeted marketing, incorporating the insights obtained from business intelligence with the predictive

analysis done using machine learning, organizations can be assured of top notch systems that can improve the efficiency of their marketing processes.

1.2 Problem Statement

Currently, telecommunication companies in Kenya lack a well-defined customer segmentation strategy (Kyengo et al., 2016). Existing customer segmentation models tend to leverage on a very narrow aspect of their customer-base consumption traits thus lowering the quality of service delivery rendered. For instance, some telecommunication companies may offer voice airtime offers to individuals who are significant users of data bundles rather than providing them with data bundle offers. This disparity in marketing offers therefore reduces the general quality of service delivery of the telecommunication companies. This study therefore proposes a model that takes into consideration a wide range of customer data to categorize the customers, and subsequently develop an algorithm that can be used to market telecommunication offers with better precision. With the proposed model, it is expected that companies like Telkom Kenya should be able to effectively target their clients, ultimately, this improved service delivery is expected to improve the competence of the company in the telecommunication industry.

1.3 Research Objectives

i. To investigate the factors and challenges relating to customer segmentation.
ii. To analyze the existing approaches and techniques of customer segmentation.
iii. To develop and test a customer segmentation model for telecommunication firms.
iv. To validate the performance of the proposed model.

1.4 Research Questions

i. What factors can be used to segment customers into given categories?
ii. What are the current techniques and approaches used for customer segmentation?

iii. How can a customer segmentation model be designed and developed using logistic regression?

iv. How can machine learning models in customer segmentation be validated?

1.5 Justification of the Study

Different studies on marketing have asserted the fact that customer segmentation is crucial for any organization that intends to satisfy the needs of its client base effectively (Yankelovich & Meer, 2006). Over time, traditional methods of customer segmentation have been used to achieve this process. However, with the rapidly growing volume of data from clients, traditional approaches have proven to be significantly ineffective when performing customer segmentation. Therefore, more novel techniques such as machine learning have been proposed to take over from the traditional methods due to the suitability of the latter to effectively handle large volumes of data effectively (Abad-Grau et al., 2009). The implementation of the method developed in this project would therefore be instrumental in improving the state of marketing currently being done in Telkom. By reaching out to the right customers, it is anticipated that Telkom Kenya's revenue will increase significantly.

Chapter 2: Literature Review

2.1 Overview

In this chapter, the concept of customer segmentation will be reviewed based on previous studies. Studies focusing on how the approach has been implemented in the past and the current trends in customer segmentation will also be reviewed. Finally, previous investigations done on the application of machine learning in customer segmentation will equally be considered in this section.

2.2 Market Segmentation

The customer-base of every organization consists of clients who have different purchasing trends. The said trends are as diverse as the different individual attributes and other aspects that can be used to distinguish people. Nevertheless, common trends can be mapped out from the individual traits. From such trends, it is then possible to categorize customers into unique categories based on the commonalities that such customers have. This segmentation forms the foundational principle of market segmentation. Yankelovich and Meer (2006) established that market segmentation is a tool that can be used by organizations to ensure that they address their customer needs in an effective manner (Yankelovich & Meer, 2006). The authors explained that companies can leverage on the concept of segmentation to market their products and services to target groups which are most likely going to purchase their products. In their findings, the duo introduced the concept of psychographics.

Psychographics refers to information on customers' which companies can use to understand their purchasing patterns. However, the authors reported that by itself, psychographics are ineffective when it comes to predicting the purchasing trends of clients since it mostly focuses on aspects such as personality and individual lifestyle. In spite of this shortcomings, the authors caution against the idea of completely disregarding the concept of market segmentation. Their

findings indicate that if properly implemented, market segmentation has the potential of identifying the right target groups which organizations can focus on to ensure that they achieve their intended marketing objectives. Therefore, to ensure that an organization implements an effective market segmentation strategy, the article provided minimum considerations which have to be addressed (Yankelovich & Meer, 2006). To begin with, the segmentation strategy must be in line with the company's strategy. Secondly, the market segments obtained from the segments should be an indicator of the source of revenue of the organization. In addition, the segments must clearly bring out the values, attitudes and beliefs of the customers to whom the services or products are being marketed to. The segments developed from the strategy should also be flexible enough to allow for any changes that may occur among the clients or the organization.

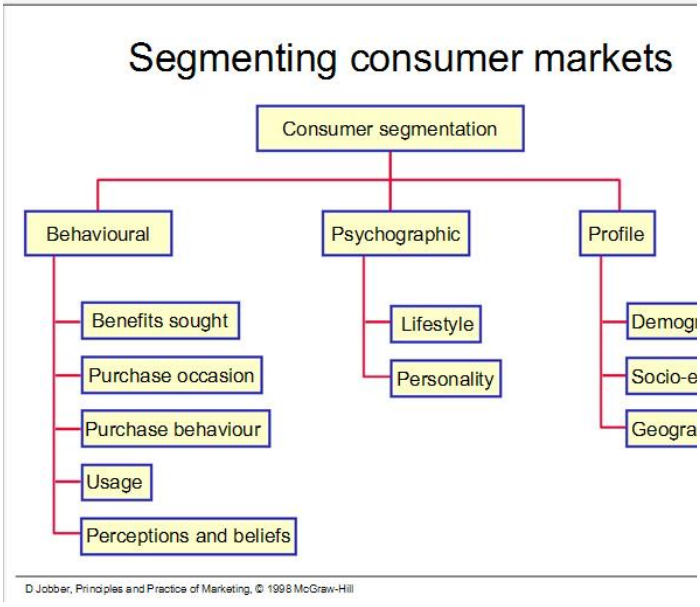

Figure 2. 1: A general overview of customer segmentation (Jobber, 2009)

The present age is characterized by a high number of people using devices such as mobile phones than previously before. The utilization of these devices implies that the

telecommunication industry has one of the highest number of clients globally. Telecommunication companies are therefore among the leading institutions that actively implement market segmentation. However, for such companies, the process of customer segmentation is not trivial. Many telecommunication companies are shifting from the monolithic approach of market segmentation to a new paradigm where different types of market segmentation can be implemented in the same organization concurrently. The primary objective of having different segmentation models existing concurrently stems from the primary need of categorizing customers into clusters as a means of solving existing business problems.

Judy Bayer, (2010) explains that over 10 tactical segmentation approaches can be employed by a telecommunications company. In her report, the author narrows down on four primary segmentation schemes commonly used in telecommunication companies (Bayer, 2010). These are customer value segmentation, customer behavior segmentation, customer life cycle segmentation and finally customer migration segmentation. From these segmentation techniques, telecommunication companies are able to extract information such as the likelihood of customer churn as well as the profitability of their customers. With these information, precise targeting can be done by the companies subsequently leading to improved business planning strategies being implemented in an organization.

2.3 Challenges Facing Market Segmentation

While market segmentation may have significant benefits to an organization, it is essential to note that there are several challenges associated with it as well. These challenges range factors arising from the internal mechanisms of an organization to external factors which are beyond an organization's control. Dibb (2017) notes that among other factors, successful implementation of customer segmentation systems is hindered by factors such as ethical concerns, insufficient organizational resources and lack of proper expertise on how to

implement the segmentation process in the organization. Dibb focused the findings of the study on organizations in the social marketing field. Nevertheless, the results can still be generalized to a wider category of industries since most of the aspects that contribute to the unavailability of effective market segmentation techniques are universal. For instance, the study noted that in most segmentation processes, technological advances were not being used appropriately to perform the segmentation. Despite the fact that the organizations had sufficient data at their disposal, they had no data analytics strategies in place to help in identification of key customer trends in the data. The author therefore emphasized on developing segmentation tools that leverage on advanced technologies that can be used to make significant inferences from the large volume of customer data available.

2.4 Existing Market Segmentation Techniques

The primary purpose of this case study is to present Telkom Kenya with a market segmentation approach that is built on machine learning. However, to effectively justify the need of such a model, it is important to evaluate some of the preexisting models that have been and still are in use. Specifically, the methods used in implementing the segmentation models shall be reviewed in this section. The Taguchi method is one approach that has been considered to be effective in implementing market segmentation. Hong applied the Taguchi clustering technique on tea-beverage customers in an attempt of investigating the effectiveness of the approach. The author employed an empirical approach where several factors were taken into consideration. These include the location of the customer, the occupation of the customer, customer preferences, customer's gender and customer experience. These five factors were considered to be the primary control factors of the experiment. The study also involved other uncontrolled factors such as the health of the customer, the prior experience of the customer and the weight of the customer in a given sales period.

The Taguchi model employed in the study relied significantly on interclass inertia to establish a performance measure. The measure was primarily used to show how close each category was to the m-dimensional space of numerical attributes in the variables identified in the study. An important factor to note is that the Taguchi model is mostly statistical in nature (Hong, 2012). The nature of this approach can be attributed to the fact that the study conducted was empirical in nature and therefore various statistical attributes such as t-tests had to be provided to evaluate the performance of the model.

Another statistical approach that has been used in performing market segmentation is the CHAID approach. Díaz-Pérez and Bethencourt-Cejas, (2016) developed a Chi-square Automatic Interaction Detection algorithm that was primarily aimed at segmenting a tourism market (Díaz-Pérez & Bethencourt-Cejas, 2016). The process of performing market segmentation depends on various factors. The diversity of the factors therefore implied that multivariate statistical tools would be the best approach in developing a market segmentation plan. Some of the techniques encompassed within the multivariate technique include cluster analysis, multiple correspondence analysis and discriminant analysis. These models have however been found to be less sophisticated as compared to the CHAID approach proposed by the authors. It was therefore hypothesized that the CHAID approach might be more effective than the multivariate approaches based on the former's level of sophistication. Therefore, the study by the duo was aimed at comparing the effectiveness of the CHAID algorithm and the multivariate techniques. Specifically, the study compared the CHAID model to the discriminant analysis approach. The study revealed that indeed the CHAID approach tends to be more effective than the multivariate techniques when it comes to customer segmentation. The superiority of the CHAID approach stems from the fact that it considers the contribution of individual variables based on the perspective of the analysis being conducted. Based on this logic, the results showed that while some variables may play an important role when considered

in the multivariate analysis scope, the same variables may be dropped when implementing the CHAID algorithm due to their misalignment with objectives of the analysis being undertaken.

2.5 The Impact of Big Data on Market Segmentation

Big Data is extremely huge sets of data that can be analyzed computationally to reveal trends, patterns and associations that can be used to reveal human behaviors. Most of the works reviewed in the previous sections were based on studies undertaken in industries with considerably less massive customer base. In the telecommunications industry however, the number of clients is known to be growing exponentially over the years. This growth implies that the volume of data obtained from clients is also increasing in an equal manner. The concept of big data therefore stems from the large volume of information that is collected from clients. It is important to note that big data has been, in this context, localized to data regarding customer trends and demographics. The term big data however has a wider meaning in a broader context. The increasing volume of customer data has presented organizations with challenges and opportunities in equal measure. To begin with, the existence of such data initially overwhelmed organizations since no platforms had been dedicated to handle and process such data. Therefore, performing different processes such as market segmentation among many more proved to be difficult since the existing approaches were not designed to handle such volumes.

Big data on the other hand has also provided organizations with an opportunity of improving the quality of the products and services offered to their clients. With the increasing volume of data, companies are now able to extract useful insights from their customers and subsequently adjust their operations to meet their clients' expectations. Galbraith explains that through enhanced data analytics capabilities, companies are now in a position of effectively handling the large volume of information collected from their clients to make well-informed

decisions that have the potential of increasing the profitability of the companies in general (Galbraith, 2014). Sun et al., (2014) also reported their findings on the possibility of improving business intelligence through big data analytics. The latter focused on big data analytics service-oriented architecture and how this architecture can be applied to the overall concept of business intelligence (Sun et al., 2014). The authors begin by acknowledging the fact that business intelligence has played a major role in enhancing the competence of organizations in different sectors. However, in spite of its relevance, there are still numerous challenges facing the concept of business intelligence. For instance, the evolution of data technologies has proven to be a major concern since with the increasing development of big data, organizations are compelled to seek better data analytics methodologies to improve their business intelligence systems. Therefore, to overcome this challenge, the authors developed an ontology that incorporates both big data and data analytics to effectively come up with a big data analytics model that organizations can incorporate into their working to sustain their business intelligence systems. In the proposed system, a combination of big data and data analytics can be split into three distinct components. These are big data descriptive analytics, big data predictive analytics and big data prescriptive analytics. Big data descriptive statistics, according to the article, refers to the use of large volumes of data to explain the trends and relationships of different attributes in the data. Big data predictive analytics on the other hand refers to the use of the attributes contained within the big data to foretell future trends based on given probabilities. Big data prescriptive analytics involves using the attributes contained in the big data to provide suggestions on what measures should be taken to address given situations in an organizational setting. These three can finally be morphed together into a general big data analytics model which can then be integrated into the business intelligence system of an organization. The resulting big data analytics is based on emerging trends in fields such Information communication technology and machine learning (Singh et al., 2014). The

primary objective of a business intelligence system in an organization is to provide individuals in the marketing field with an opportunity of making proper decisions that will transform the performance of an organization positively. To achieve this mandate, it is essential to have a support system which can guide an organization towards making such decisions. Therefore, Sun et al believe that organizations will stand to benefit through the incorporation of the proposed approach into their business intelligence systems.

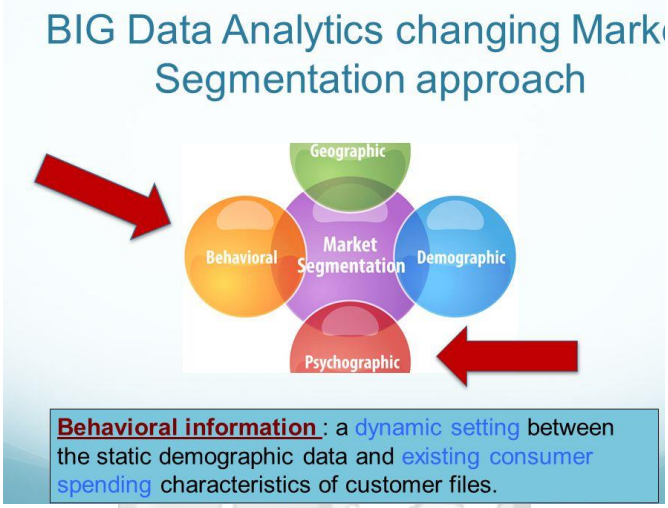

Figure 2. 2: A representation of how big data incorporated different customer attributes to be later used in segmentation processes (Higgins, 2017)

2.6 The Use of Machine Learning in Market Segmentation.

Machine learning has witnessed a significant surge in applications in the past decade. Machine learning models have been incorporated in different fields with the results being the performance in the respective companies have been increasing. This section therefore evaluates how machine learning in general has also been incorporated in performing market segmentation.

Abad-Grau et al., (2009) employed three different machine learning models with the objective of performing market segmentation for performing art audiences. In their study, the authors employed decision tree classifiers, Bayesian classifiers and finally k-nearest neighbors

classifier to categorize the audiences of performing arts into different clusters. When conducting their study, the authors acknowledged the fact that optimal decision making in marketing relies on the combination of different methods, among them machine learning techniques (Abad-Grau et al., 2009). The classification model employed in this study was based on data collected from attendees and non-attendees of Opera and Ballet performances. For each of the three models mentioned above, additional variations were included. For instance, the Bayesian classification was further split into the Bayesian Inference model and the Naive Bayes with ML estimation, while the K-nearest neighbor was further divided into k=1 and k=5. The decision tree classifier used in this study was found to have the best performance with a predictive accuracy of 83.61% for Opera attendees and non-attendees and an accuracy of 82.39% for Ballet audience. K-nearest neighbor was found to have the least accuracy, with the Opera audience having an accuracy of 80% and the Ballet audience recording an accuracy of 79% (Abad-Grau et al., 2009).

Machine learning methodologies have been embraced in many data intensive institutions due to their abilities to handle different kinds of data. Under normal conditions, data tends to contain outliers and other inconsistencies which if not addressed, would significantly affect the subsequent applications of the data. Machine learning provides organizations with tools which can be used to handle the outliers thus ensuring that any subsequent application of the data is not affected by the outliers. Wang, (2009) demonstrated the prowess of machine learning in handling outliers in a study aimed at identifying outliers and performing market segmentation using kernel-based clustering techniques (Wang, 2009). The model proposed in the study was based in the concept of possibilistic C means which had been structured to identify single clusters rather than multiple clusters. The flexibility of the kernel approach in detecting outliers was achieved by proposing the use of kernel possibilistic clustering method that was designed to identify outliers in feature space by designing nonlinear boundaries of the probability of the

outliers. The author compared the effectiveness of the fuzzy C means and the kernel fuzzy clustering methods in outlier identification using the Iris dataset and CRM data obtained from an automobile company. The kernel fuzzy clustering methods were found to be better as compared to the fuzzy C means due to the flexibility of the former when addressing overlapping clusters (Wang, 2009). The results by Wang therefore demonstrate the significance of machine learning models when addressing market segmentation in instances where customer traits overlap.

Clustering methods have proven to be the most popular machine learning approaches when conducting market segmentation. Support vector clustering and k-means clustering are among some of the most commonly used clustering techniques used in the studies evaluated. Nevertheless, other machine learning frameworks have also proven to be equally effective when performing customer segmentation. It is therefore anticipated that the results obtained from the classification models implemented in this study will be compared to the existing results obtained in this review and evaluate the consistency of the results with what has been recorded in literature.

Chapter 3: Research Methodology

In this section, the research design and model development will be discussed. The development of the model will follow the following sequence.

i. Data Collection
ii. Data Preprocessing
iii. Model Development
iv. Model validation

3.1 Research Design

This research is essentially an applied research in the sense that it shows how fundamental concepts of machine learning can be applied in the field of marketing. A non-experimental approach is used to meet the objectives of the study due to the nature of the research.

3.2 Model Development

3.2.1 Data Collection

The data used in this study was obtained from Telkom Kenya's CRM department. This data, which is mostly collected and organized using the Tableau tool was extracted and later stored in .csv format before being used for the development of the model.

3.2.2 Data Preprocessing

The data preprocessing step involved activities such as handling missing values, re-scaling the data to remove outliers, fixing the typo errors in the data and ensuring that both continuous and categorical data in the datasets are well handled. Data cleaning and preprocessing was done using different libraries provided by Python. Primarily, NumPy (Walt et al., 2011) and Pandas (Wedel & Kamakura, 2012) were the most widely used libraries in this step. Columns in the data files containing a significant number of missing values were entirely dropped using the Pandas dropna function. On the other hand, for columns containing a small number of missing

value, the “ffill” method of filling the data was used. This approach was selected based on the fact that the data had a chronological order and therefore using the “ffill” method would preserve the general trend of the data.

3.2.3 Exploratory Data Analysis

The cleaned data was then subjected to exploratory data analysis in order to identify any trends the data. Visualization played a crucial role in obtaining insights from the data and therefore plotting tools such as Matplotlib (Hunter, 2007) and Seaborn (Waskom et al., 2017) were extensively used in this process. Among the different trends that were investigated include the distribution of different age groups among the company’s clients, the contribution of different age groups to the revenue of the company, the contribution of different services to the total revenue of the company and how each of the services offered by the company was distributed among the age groups.

3.2.4 Feature Engineering

Feature engineering was then performed on the data to transform it into a form that can be easily fed into any machine learning algorithm. To begin with, the independent and dependent variables were defined. For this study, the kind of technology used by the individual client was the dependent variable, while other attributes such as the expenditure and device type were the independent variables. Categorical features in the data were label encoded in order to convert them into a form that could be easily understood by the models used without losing their significance. The get dummies functionality of Pandas was used to perform this process. Finally, the last feature engineering step undertaken was to scale the data in order to normalize the distribution. This scaling was achieved using the Scikit-Learn’s MinMax scaler (Pedregosa et al., 2011). The primary objective was to ensure that any outliers in the data were removed

and that a normal bell-shaped distribution is achieved. The equation for the scaler is provided below as:

$$scaler = \frac{x - min(x)}{max(x) - min(x)}$$

where x is the variable(s) being scaled (Kramer, 2016). In this study, customers were segmented based on the technology they used; that is whether they were 2G, 3G or 4G users. Therefore, the technology column was the target variable while other columns were the predictor columns.

3.2.5 Model Creation and Evaluation

A machine learning 'factory' was then created with the primary objective of comparing the performance of different classification models. This process was considered to be essential since different models tend to perform differently. This being a classification problem, no regression techniques were used. Instead, classifiers were used to categorize the clients into either 2G, 3G or 4G users. Five primary classifiers were used. These are the Linear Discriminant Analysis, Logistic Regression, K-Nearest Neighbor classifier, Decision Tree Classifier and finally the Gaussian Naive Bayes Classifier. The accuracy of each model in the factory was evaluated based on the K-fold cross validation technique. Here, a K-value of 10 and a random state of 7 was used to compute the accuracy of each of the classifiers. A generalized formula of the K-fold cross validation technique is shown below (Kramer, 2016).

$$CV = \frac{1}{N}\sum_{1}^{N} L\left(y_i, f^{(-k(i))}x_i\right)$$

After identifying the most accurate model from the factory, a classification model was then implemented on the provided training data and used to make predictions on the provided test data. These predictions were then used to segment the customers into their respective technology classes. The segments were then used as a basis of creating a marketing plan for the clients.

3.3 System Development Methodology

A data-driven modeling approach was embraced when developing the system in this study. The DDM methodology is ideal for this study based on the fact that the success of the classification model depends on the continued availability of data from the clients of the company (Abrahart et al., 2008). Through the frequently collected data, it is possible to understand the different trends and insights and possibly map out any evolution in the purchasing trends of the clients.

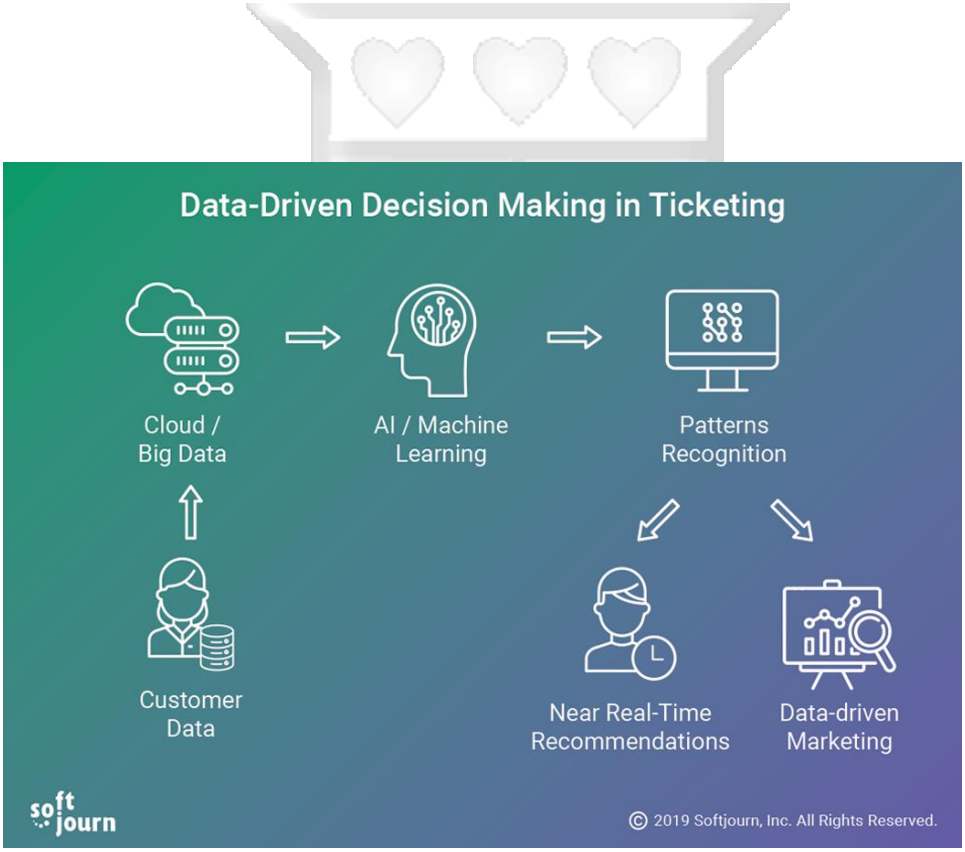

Figure 3. 1: An overview of DDM models in machine learning (s0f, 2009)

3.4 Research Quality

The primary intent of performing any research is to add on to the existing knowledge regarding a specific subject. To satisfy this objective, it is essential to ensure that all processes undertaken

when performing the research are not only of high quality but also specifically geared towards the main objective of the study. Therefore, to achieve quality in this research, the following measures were taken:

i. Only data obtained from Telkom Kenya was used in the entire procedure outlined. Focusing on this data ensured that whatever information that was captured was a true representation of Telkom Kenya's customers.

ii. Any assumptions made in the development of the classification model are clearly explained during the implementation of the model. This move helps in reducing the ambiguity of the entire process.

iii. The research has employed an analytical approach which can be evaluated to ascertain its effectiveness. Based on the nature of the scope and intended objectives of the research, an experimental approach could not be employed. This approach tends to be easy to evaluate based on its reproduce-ability.

In terms of evaluating the performance of the model being used, the cross-validation score of the respective machine learning models used was computed. This approach is widely embraced due to its effectiveness in determining the bias of given models in new sets of data.

3.5 Ethical Considerations

The following ethical considerations were made when conducting the research and developing the model:

i. Telkom Kenya was fully aware of the researcher's intention to use the customer data for the study. Therefore, there were no irregularities in obtaining the data.

ii. Personal information which may be used to reveal personal attributes of individuals such as their names, precise locations and other aspects were omitted from the data. Instead, the data only contained customer purchasing trends and the types of devices used.

Chapter 4: System Design and Architecture

In this section, the overall design of the system being developed will be discussed in addition to its architecture. This section structured as follows:

i. Requirement analysis
ii. Functional Requirements
iii. Non-functional requirements
iv. System architecture
v. Use-case diagram
vi. System-sequence diagram
vii. Flowchart

4.1 Requirement Analysis

The implementation of the classification system was based on two general requirements. These were functional requirements and non-functional requirements.

4.1.1 Functional Requirements

i. The designed model should be able to classify customers into their respective segments based on their purchasing trends.
ii. The designed model should help the company in performing targeted marketing based on the customer-segmentation achieved through the classification model.
iii. The designed model should have an accuracy that is reliable for commercial purposes.

4.1.2 Non-functional Requirements

i. Usability - The developed model should be in a state that can be used by Telkom Kenya to assist in the implementation of targeted marketing.

ii. Scalability - Based on the increasing volume of client data, the developed model should be easily scalable to handle the ever increasing customer data collected by Telkom Kenya.

iii. The output of the developed model should be in a form that can be easily stored and retrieved by the facilities within Telkom Kenya.

4.2 System Architecture

System architecture is mainly the conceptual model which defines the behavior, the views and the structure of a system. An architecture description refers to a description that is formal and a system representation that is organized in a way that supports the reasoning about the behaviors and structures of the system. A system architecture may comprise sub-systems developed and the system components, which work together to implement the whole system.

The system architecture depicts the application layout which implements the algorithm for market segmentation and targeted marketing. Supervised machine learning techniques were implemented in developing the segmentation and thereafter assist in performing targeted marketing. Specifically, multi-class classification models were tested on the company's customer data to achieve the segmentation and targeted marketing. Supervised machine learning techniques were used to aid in targeted marketing. These techniques use subscribers' training and testing datasets and the selected features that characterize those subscribers. Finally, based on the categories in which the customers belonged to, various service plans were marketed to them. The system architecture is summarized in the schematic below

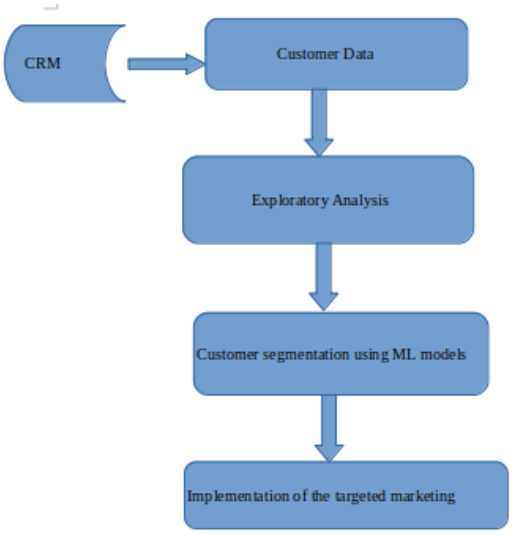

Figure 4. 1: Overall system architecture of the customer segmentation and targeted marketing

4.3 Use-case Diagram

The main users in this system are the administrator, the CRM database and experts from the marketing department. The administrator in this system is responsible for the fetching of data from the CRM database and performing all the data cleaning, feature engineering, model development and model evaluation. Further, the administrator is also responsible for designing the targeted marketing algorithm based on the advice of the marketing expert. The marketing expert on the other hand will be developing marketing plans based on inferences obtained from the administrator's work. The marketing suggestions from the marketing analyst are then fed to the clients based on their identified features. The entire process is summarized in the Figure 4.2

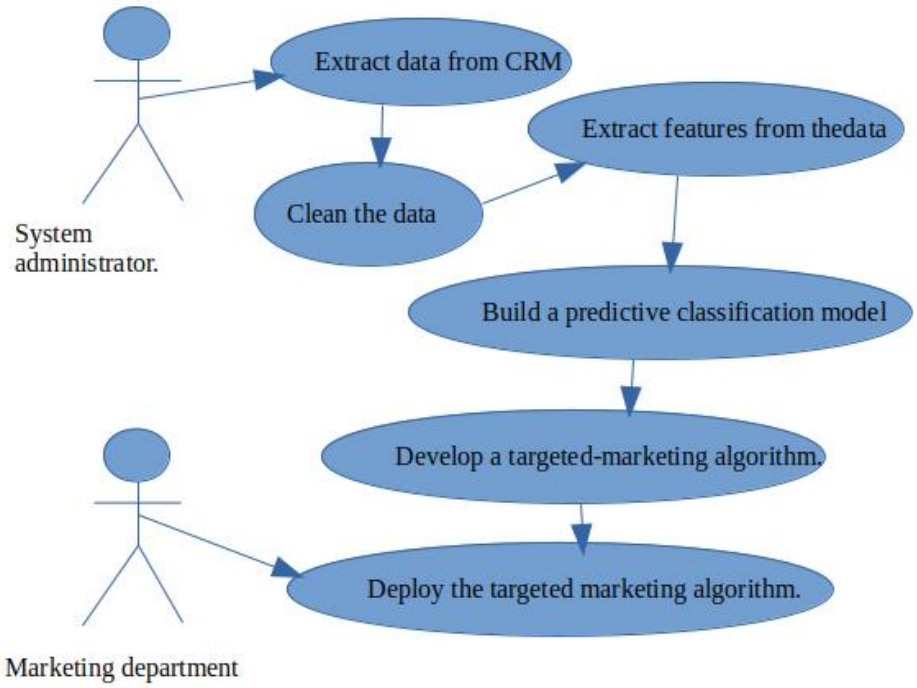

Figure 4. 2: Use-case diagram for the customer segmentation and targeted marketing system

The use-case diagram for the system analyst is provided in Table 4.1 as shown.

Table 4. 1: Use-case diagram for the system administrator

Use Case: Data Acquisition, Data Pre-processing and Extracting Features	
Primary Actors: System Administrator	
Pre-condition: System administrator has access to customer data stored in the company's CRM.	
Post-condition: System admin carries out data preprocessing, feature extraction and finally develops a classification model based on the technology used by the clients.	
Main Success Scenarios	
Actor Intention	**System Responsibility**
1. Admin obtains raw data from the CRM of the company.	
2. Admin performs data preprocessing, feature engineering and finally fits the processed data into classification algorithms.	
	3. The system identifies the features in the raw data.
	4. The system succesfully groups the customers according to their respective technologies used.

The use-case diagram for the marketing analyst is also provided in Table 4.2

Table 4. 2: Use-case diagram for the marketing analyst

Use Case: Performing targeted marketing	
Primary Actors: Marketing analyst	
Pre-condition: Receives data containing customers clustered according to the technologies they use.	
Post-condition: Performing targeted marketing to the customers based on their categories.	
Main Success Scenarios	
Actor Intention	**System Responsibility**
1. Marketing analyst receives structured data containing clustered customers.	
2. Marketing analyst then develops appropriate marketing techniques for each cluster of customers.	
	3. System automatically prescribes to the clients offers based on their respective clusters.

4.4 System Sequence Diagram

A system sequence diagram is used to describe the sequence of operations that the algorithm undertakes from beginning to end. It is also an illustration that shows, for a specific scenario of a use case, the events that external actors do generate, their order, and the possible inter-system events. System sequence diagrams usually are visual depictions of individual use cases. The sequence diagram for this research is depicted in Figure 4.3

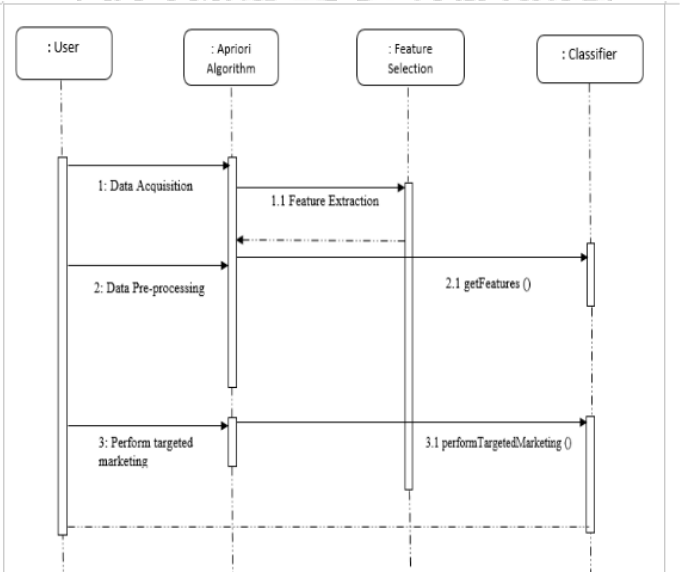

Figure 4. 3: System diagram for the targeted-marketing system

Finally, the data flow diagram of the entire system is provided as follows.

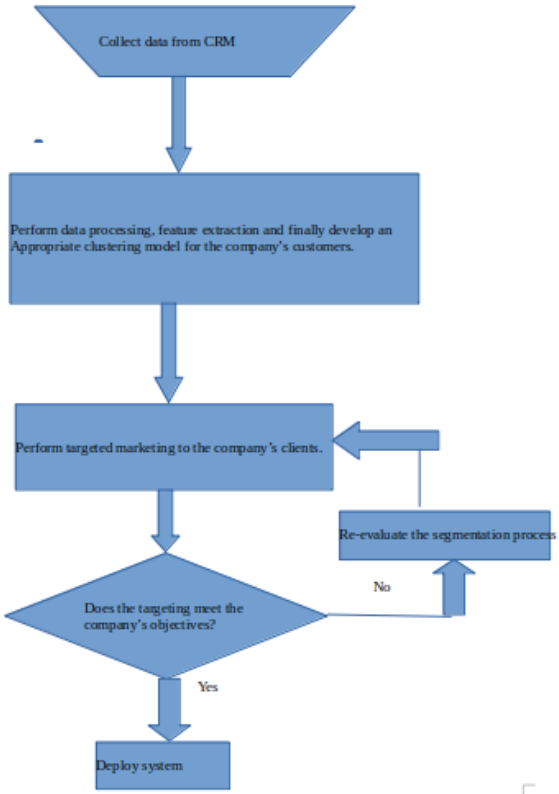

Figure 4. 4: The overall data flow diagram of the system

Chapter 5: Implementation and Testing

This chapter outlines the implementation of the customer segmentation system architecture outlined in chapter 4 above. Having developed the classification model, this section outlines the steps followed in implementing the system based on the architecture provided in chapter 4 above. The main sections in this chapter are the implementation and testing sections.

5.1 Implementation

5.1.1 Data Collection

For the customer segmentation model, the data used was collected for two months. The first month data was used for the actual training and testing of the model using a split ratio of 0.7:0.3 respectively. The segmentation was performed from predictions made on the second month data.

5.1.2 Data Preprocessing

The first step in the implementation of the system was to clean the collected the data. This was accomplished in various ways. To start with, missing data in the provided data was handled either by dropping the data or filling it using a predetermined approach. Dropping missing data was performed for features which were considered to have little or no impact on the overall outcome of the system. In addition, dropping was performed when the missing data was founded to be significantly large in comparison to the non-missing data. For missing data which was significantly less compared to the entire data set, the ffill method was used to fill in the missing values. This approach was selected due to the fact that the data was sequentially arranged and therefore implementing the ffill method implied that the general structure of the data was maintained. Any form of typing errors was also corrected in the data to ensure that the dataset was in a form that could be easily used in the subsequent processes. The age values in the data were found to be abnormally distributed and it was therefore necessary to scale them

down to between 0 and 100 years. This range was considered appropriate since it would make it possible to easily categorize the clients into different ages. The final preprocessing step undertaken was to categorize the data into their respective genders. Here, three instead of 2 genders were defined. Apart from the conventional male and female genders, the system also included company as part of the gender since some organizations and businesses were also found to be significant consumers of the company's services.

5.1.3 Exploratory Data Analysis

Insights from data are well generated when the data is visualized using different visualization tools. For this system, it was necessary to explore the data to identify the trends that would be used to guide the targeted marketing explained in the introduction section of this chapter. The results of this step are the graphs provided in the results and analysis section.

5.1. 4 Feature Engineering

In its raw form, data cannot be fed into any machine learning model and yield the required results. Instead, this data has to be fine-tuned in a process commonly known as feature engineering. In this system, the following activities were conducted to engineer the data features appropriately. To begin with, columns which contained categorical data were label encoded into numerical formats that maintained the categorical nature of the data. Such features include the gender of the clients, the devices used by the customers, the different age groups and finally the type of technology used by the customers. It is important to note that the technology type was not part of the predictive features but rather, the variable that had to be predicted for the customer segmentation process. The data was then scaled using the MinMax scaler to ensure that a normal distribution of the data was achieved. Without performing this process, the data may contain a lot of outliers which would affect the quality of the model being tested. The MinMax scaler ensured that a bell shaped distribution of the data was achieved.

5.2 Model Testing

Different machine learning models exist which could be used to perform the classification exercise in this project. It is also important to acknowledge that each of these models have different performances. It was therefore necessary to create a model factory that could be used to identify the most accurate model and therefore use it for the problem at hand. In this problem, five different models were used and their performance accuracy was compared based on their respective cross validation scores. These were the Logistic Regression classifier, the Linear Discriminant Analysis classifier, K-Nearest Neighbor Classifier, the Decision Tree Classifier and finally the Gaussian Naive Bayes Classifier. The model factory used in this project is shown below. Note that the model used 10 folds for the cross validation process to determine the most accurate model of the five that were used.

After identifying the most appropriate model, the next procedure was then to perform the actual customer segmentation based on the technology that the customers used. Here, 30% of the data was used for training while the rest was used for testing the data.

5.3 Marketing Algorithm

Having successfully performed the classification using machine learning algorithms, the next step was to design an algorithm that could be used to help in the targeted marketing. A conditional statement was used to determine the correct type of services to be offered to customers based on the EDA that had been conducted earlier on. Essentially, this section brought together all the previous steps, starting from the EDA to the final classification. It is after the implementation of this step that the marketing algorithm can now be deployed by the company to be tested out on its clients. The evaluation of the deployment and the results of the deployment are however beyond the scope of this project. The snippet below shows the conditional statement used in performing the targeted marketing based on the inferences made from the EDA.

```
def Suggested_Offers(row):
    if row['Predicted Technology'] == 0:
        return "Suggest Voice Plans"
    if row['Predicted Technology'] == 1:
        return "Suggest voice and data plans"
    if row['Predicted Technology'] == 2:
        return "Strongly Suggest Data Plans. Voice plans should also
be considered"
new_data = X_test.assign(Offers = X_test.apply(Suggested_Offers, axis
= 1))
```

Figure 5. 1: Conditional statement used in the targeted marketing based on predictions made by the logistic regression classifier and the EDA conducted

Chapter 6: Results and Discussion

In this section, the results obtained from the data visualization process as well as the classification process will be outlined. Thereafter, the results will be discussed in detail. The section is structured as follows.

i. Age distribution among the customers
ii. Service expenditure distribution
iii. Revenue contribution of different technologies
iv. Expenditure by age
v. Revenue contribution of different services in the respective technologies
vi. Discussion

6.1 Results

6.1.1 Age Distribution

To properly understand the customer base, it was necessary to first categorize the clients according to their respective ages and thereafter find the age distribution. The customers were categorized as either youth, adults, the elderly or very elderly. Their distribution is shown below:

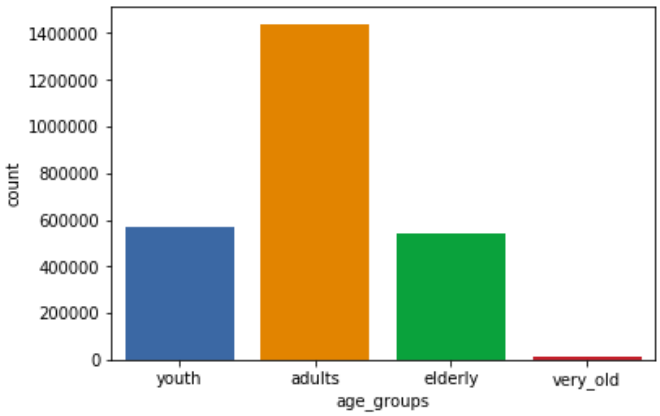

Figure 6. 1: Distribution of clients by age in the company

Adults are the majority of clients in the company. It is nevertheless important to acknowledge that the data contained inconsistencies in the ages and therefore rescaling was necessary. The representation above should therefore not be considered as an actual representation of the distribution of the company's clients.

6.1.2 Service Expenditure Distribution

The company mainly offers voice, SMS and data bundle offers. Therefore, evaluating the contribution of each of these services to the total revenue of the company was necessary. Below is a representation of the contribution of each service to the total revenue of the company.

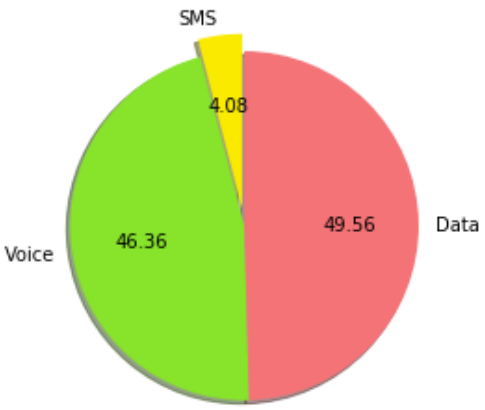

Figure 6. 2: Percentage contribution of different services to the total revenue of the company

Data bundles are noticeably the largest contributors to the total revenue of the company. This is expected since among most internet consumers in the country, Telkom's data bundles are known to be affordable. Similarly, voice is equally significant to the company's revenue. Compared to other telecommunication companies, Telkom's voice tariffs have been found to last longer for a cheaper amount.

6.1.3 Revenue Contribution of Different Technologies

After splitting the customers into 2G, 3G and 4G users, it was necessary to evaluate the contribution of each cluster to the total revenue of the company. The following results show how each category contributes to the revenue of the company.

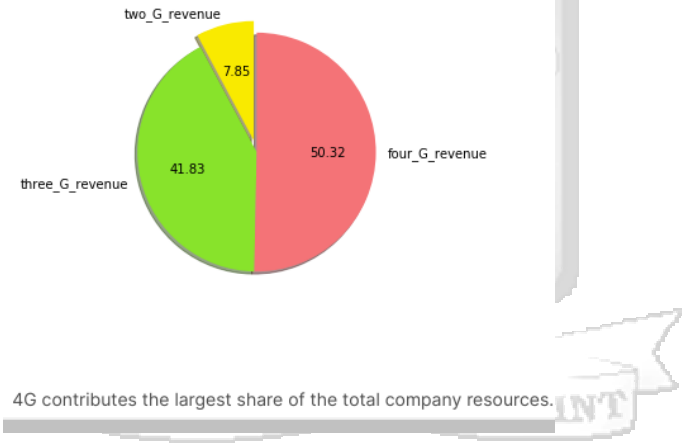

Figure 6. 3: Percentage contribution of different technologies to the total revenue of the company

4G technology is seen to be the largest contributor to the company's revenue. 4G is commonly associated with internet services and therefore these results are in agreement with what was identified in the previous section. Most clients use 4G technology primarily for internet access services.

6.1.4 Expenditure by Age

It is well understood that the purchasing trends of clients changes significantly with their ages. Therefore, to demonstrate the validity of this statement, the consumption trends of individual age groups were investigated as shown below.

When investigating the expenditure of the youth, results showed that this age group mainly purchases internet bundles followed closely by voice bundles. This trend is expected since most youths are known to spend a significant amount of their time on online platforms.

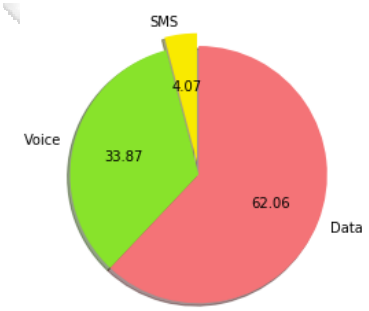

Figure 6. 4: Service consumption by the youth

When considering adults, both voice and internet bundles were found to be the most purchased services. This trend can be attributed to the fact that among the adults, phone conversations are a preferred mode of communication. Additionally, adults have also been found to be active in social media platforms and this explains why data expenditure is a significant contributor in this age group.

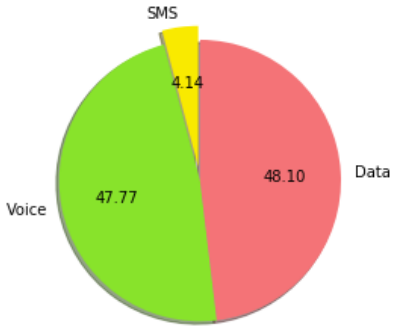

Figure 6. 5: Service consumption by adults

As seen from the two charts below, both the old and very old are mostly purchasers of voice bundles. Unlike the other two categories, these category of clients tend to use their devices primarily for communication primarily through phone calls.

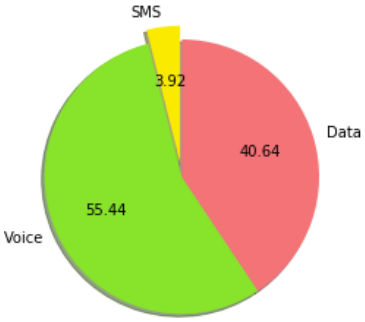

Figure 6. 6: Service consumption by very old

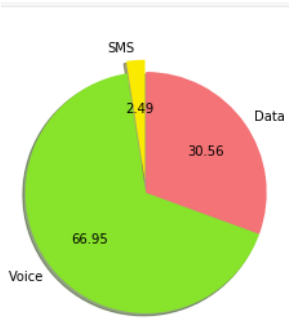

Figure 6. 7: Service consumption by very old

From the three distribution of services among the different age groups, it is evident that voice bundles tend to be closely associated with data bundles. Again, this observation is expected since in most data offers provided by Telkom, there are corresponding free Telkom call offers as well. This explains why both voice calls and data bundles are the most purchased services in the company.

6.1.5 Revenue Contribution of Different Services in the Respective Technologies

Understanding the revenue dynamics of the individual technologies revealed that data bundles are the most consumed products in 4G technology closely followed by voice bundles and finally SMS bundles.

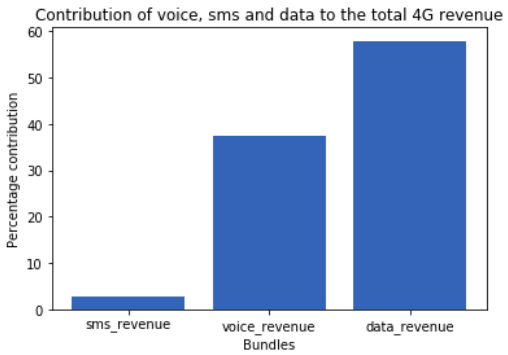

Figure 6. 8: Revenue contribution of different bundles in 4G technology

In the 3G category, voice and bundles were found to be the highest purchased products while SMS bundles were the least purchased.

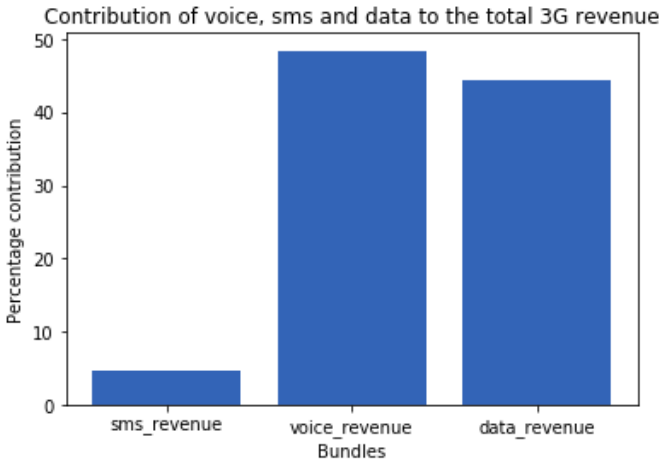

Figure 6. 9: Service consumption by very old

Finally, in the 2G category, voice bundles were the predominant bundles purchased. Both SMS and data bundles were found to have a significantly low contribution to the revenue of the 2G category.

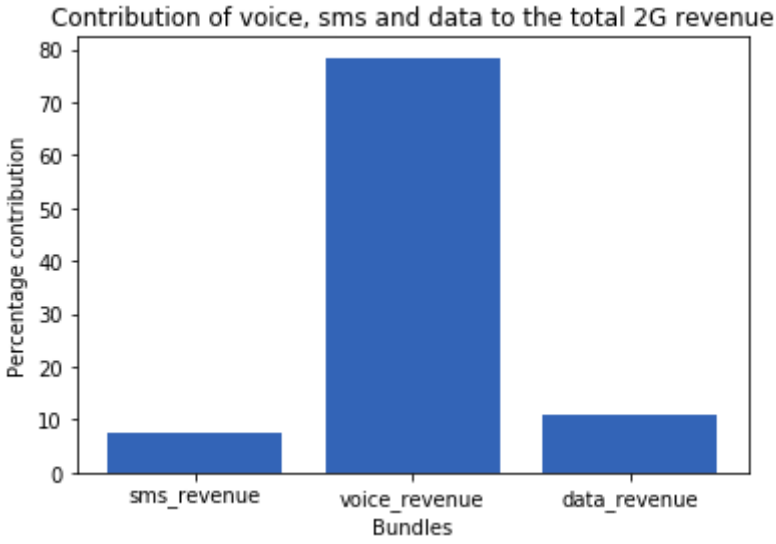

Figure 6. 10: Distribution of revenue in 2G technology

By convention, these results can be considered to be an actual representation of the company's customer base. Most individuals using 4G technology tend to have devices which have enhanced internet access features. Therefore, this explains why the technology's revenue

mainly comes from internet bundles. As stated earlier, most Telkom data bundles have an associated voice bundle offer. This explains why voice bundles are also a major contributor of the technology's revenue. The same rationale can be applied to the 3G category.

Phones operating on 2G technology tend to be basic feature phones whose internet access capabilities are significantly poor. Therefore, these devices are exclusively used for communication by either voice calls or SMS. This explains why this category's revenue predominantly comes from voice bundles.

From the results presented above, it is evident that the type of technology being used plays a crucial role in the revenue collected from clients. Therefore, grouping the clients into clusters based on the technology they used would be an ideal starting point for achieving targeted marketing in the organization. With this in mind, a multi-class classification model factory was designed to help in classifying the customers into either 2G, 3G or 4G users. Using cross-validation score as the metrics of performance, the following results were obtained.

6.1.6 Model Factory Results

```
LogisticRegressor: 0.674950 (0.357638)
LDA: 0.583370 (0.260618)
KNN: 0.657520 (0.241425)
CART: 0.592340 (0.176006)
NB: 0.643270 (0.330489)
```

Figure 6. 11: Cross-validation scores from the different models used in the model factory

Logistic regression classifier was found to have the highest cross-validation score and was therefore selected to be used in the classification problem. The results from the model factory are in contrast with different sources of literature since most studies found clustering techniques such as the KNN to have higher classification performances. The difference can be attributed to factors such as difference in data and how the data was treated prior to being fed into the model.

6.1.7 Logistic Regression Classification Results

Having identified the right model to use, the next step involved performing the actual classification and evaluating different metrics from the logistic regression classifier. From this process, the following classification report was obtained.

	precision	recall	f1-score	support
0	0.78	0.80	0.79	20047
1	0.56	0.54	0.55	9789
2	0.00	0.00	0.00	164
accuracy			0.71	30000
macro avg	0.45	0.45	0.45	30000
weighted avg	0.70	0.71	0.71	30000

Figure 6. 12: Classification report for the logistic regression classifier

From the report, the overall accuracy of the logistic regression classifier is seen to be 0.72. Additionally, other metrics such as precision, recall and f1-score were also evaluated. The target classes were label encoded such that 0 represented 2G technology, 1 represented 3G technology and 2 represented 4G technology. A striking observation can be made from the precision and recall scores of 4G since they are all 0.0. This result is a major downside of the logistic regression. As a sigmoid function, the classifier works well when predicting values between 0 and 1. Therefore, this implies that for binary classes, the model works perfectly well. This classification was however a multi-class problem and therefore the 0 precision, recall and f1 scores were observed.

6.2 Discussion

6.2.1 EDA Results

To begin with, it is important to consider some of the findings obtained from the EDA conducted in the data. To begin with, the revenue collected from data bundles was seen to be the largest contributor of the company's total revenue. From the visualization done, it was realized that data bundles contributed to 49.63% of the total revenue of the organization. This contribution is however expected since most people are known to use Telkom services mostly for internet bundles due to their affordable data plans. After categorizing the customers into their respective ages, it was realized that a bulk of the company's customers were adults followed closely by the youth. However, this approximation is crude considering that the initially provided ages were mainly incorrect and therefore scaling had to be done to obtain a proper age range. Nevertheless, the results still had some significant implications as seen from further preprocessing. For instance, when considering the distribution of consumption of services among the age groups, it was realized that data bundles contribute a major share of the youth's expenditure. 62.06% of the revenue collected from the youth results from data bundles alone. This results are in agreement with what would be expected from the society since most young people at this age have devices that have constant access to the internet and therefore having sufficient internet credit is a necessity.

Among adults, both data bundles and voice bundles were found to be the most significant contributors of the total revenue of the company. Again, this trend is expected since just like the youth, adults during this age are also active users of the internet. However, unlike the youth, adults also make a significant number of calls for either personal or business reasons and therefore this explains why calls are a major contributor to this category as well.

Among the elderly and very old, voice bundles are the most commonly purchased services from the company. The rationale behind this trend can be thought of as having resulted from

the fact that many old people only use their phones for basic communication purposes. Therefore, whenever they purchase airtime, it is primarily for calling people. Data expenditure in these two categories can also be seen to be significantly high. This observation can be explained by the fact that most voice call offers in Telkom Kenya are accompanied by data offers. Therefore, by purchasing such voice offers, individuals also purchase data bundles offers. This duality explains why the data expenditure for the two categories is significantly high. In general, adults were found to be the highest contributors of the company's revenue followed closely by the youth, the elderly and the very old. This trend can be understood through the distribution of customers as seen in the appendix section.

From the EDA, it is also important to acknowledge how different technologies contribute to the total revenue of the company. 4G technology was found to be the highest contributor of the company's revenue with over 50% coming from this technology. In terms of the products offered, data bundles were found to be the highest contributor of the revenue generated by 4G. The trend can be explained by the fact that most devices that operate on 4G are used for internet access services hence the high data expenditure. 3G technology was found to be the second highest contributor to the company's total revenue. With 41.83% of the company's total revenue coming from this technology, voice bundles were found to be the most significant contributor of the total revenue of this technology followed closely by internet bundles. Finally, 2G was the least contributor to the total company revenue generating close to 8% of the company's total revenue. This revenue was predominantly as a result of voice bundles purchased by the clients. Similarly, the trends in the last two technologies were expected since most phones, both smartphones and basic phones operate using 3G technology. Therefore, purchase of internet bundles and voice bundles in this category of devices is significantly high. For 2G devices, the only expected use would be for making calls since these are basic feature

phones with limited internet access. Therefore, the dominance of voice bundles in this category of technology users is understandable.

6.2.2 Machine Learning Model Results

From the model factory created in the implementation section, it was realized that the Logistic Regression Classifier had the highest cross validation accuracy of the five models that were tested. K-Nearest Neighbor was found to be the least accurate of the five. The performance of the Logistic Regression classifier can be attributed to the fact that the fundamental working of the model is based on a structure similar to that of neural networks. Based on this underlying principle, the architecture of the classifier therefore makes it possible to effectively map out different parameters that determine the features being predicted. As a result, the classifier ends up having a better cross validation score when compared to the other models being used.

The classification report of the classifier also showed significant results. The average accuracy of the model was found to be 71%. The average precision and recall of the classifier was found to be 45%. These values are low due to the fact that by nature, the Logistic Regression classifier relies on a sigmoid function for its output. This function is well suited for binary classification problems since the results of the function lie between 0 and 1. However, the problem in this case involved three classes. Therefore, although the precision and recall scores of the first two classes were high, the precision and recall of the last category was 0 and therefore this reduced the average. Nevertheless, the model was fine tuned to yield an accuracy of 71%. This was quite commendable since it implied that there were no instances of over-fitting or under-fitting.

Chapter 7: Conclusion, Future Work and Recommendation

7.1 Conclusion

The primary objective of this project was to show how business intelligence, big data and machine learning can be cumulatively used to implement targeted marketing based on customer segmentation. By using a case study of Telkom Kenya, this study has implemented a model that can be used to achieve the stated objective. As seen from the results, the machine learning has been able to categorize customers into their respective categories with 71% accuracy. Through the classification, Telkom Kenya is in a position of marketing their products and services to the right group of customers, thereby ensuring that their marketing strategies are effective. This study therefore lays a stepping stone for future studies on precision targeting using machine learning.

7.2 Future work

This work has only addressed the implementation of a customer segmentation model that assists in precision targeted marketing. Although the performance of the model was seen to be commendable, it is important to note that the model had not been deployed and therefore no real world results can be used to justify its effectiveness. Therefore, future work can be done on evaluating the effectiveness of applying machine learning in customer segmentation and targeted marketing in the company. In addition, future projects can be done on how to best deploy the model proposed in this study to achieve the intended organizational goals.

7.3 Recommendation

This study focused on using demographic data to predict the type of technology used by the company's clients as a means of clustering the customers into individual groups. However, future work can be conducted on how the company can further implement geographical segmentation as a means of further improving its marketing strategies. It is anticipated that by

focusing on the geographic distribution and purchasing trends of its clients, the company will be in a better position of understanding how well to market its products to different regions in the country at large.

References

Abad-Grau, M. M., Tajtakova, M., & Arias-Aranda, D. (2009). Machine learning methods for the market segmentation of the performing arts audiences. International Journal of Business Environment, 2(3):356–375.

Abrahart, R. J., See, L. M., and Solomatine, D. P. (2008). Practical hydroinformatics: computational intelligence and technological developments in water applications, volume 68. Springer Science & Business Media.

Bayer, J. (2010). Customer segmentation in the telecommunications industry. Journal of Database marketing & customer strategy management, 17(3-4):247–256.

Cross, J. C., Belich, T. J., and Rudelius, W. (2015). How marketing managers use market segmentation: An exploratory study. In Proceedings of the 1990 Academy of Marketing Science (AMS) Annual Conference, pages 531–536. Springer.

Díaz-Pérez, F. M. and Bethencourt-Cejas, M. (2016). Chaid algorithm as an appropriate analytical method for tourism market segmentation. Journal of Destination Marketing & Management, 5(3):275–282.

SoftJourn. (2019). Data-driven decision making. Softjourn Inc. Retrieved September 4, 2020, from https://softjourn.com/blog/article/data-driven-decision-making

Dibb, S. (2017). Changing times for social marketing segmentation. In Segmentation in social marketing (pp. 41-59). Springer, Singapore.

Fan, S., Lau, R. Y., and Zhao, J. L. (2015). Demystifying big data analytics for business intelligence through the lens of marketing mix. Big Data Research, 2(1):28–32.

Galbraith, J. R. (2014). Organizational design challenges resulting from big data. Journal of Organization Design, 3(1):2–13.

Higgins, P. (2017). Big data, analytics: a gis approach on market segmentation - ppt video online download.

Hong, C.-W. (2012). Using the taguchi method for effective market segmentation. Expert systems with applications, 39(5):5451–5459.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment https://doi. org/10.1109/MCSE. 2007.55 Comput. Sci.

Jobber, D. (2009). Principles and practices of marketing. 6th edn mcgraw-hill higher education.

Kramer, O. (2016). Scikit-learn. In Machine learning for evolution strategies, pages 45–53. Springer.

Kyengo, J., Ombui, K., & Iravo, M. A. (2016). Influence of competitive strategies on the performance of telecommunication companies in Kenya. International Academic Journal of Human Resource and Business Administration, 2(1), 1-16.

Maji, G., Dutta, L., & Sen, S. (2019). Targeted marketing and market share analysis on pos payment data using dw and olap. In Emerging Technologies in Data Mining and Information Security (pp. 189-199). Springer, Singapore.

McKinney, W. (2010, June). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51-56).

Melnic, E. L. (2016). How to strengthen customer loyalty, using customer segmentation?. Bulletin of the Transilvania University of Brasov. Economic Sciences. Series V, 9(2), 51.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Singh, A., Rumantir, G., South, A., and Bethwaite, B. (2014). Clustering experiments on big transaction data for market segmentation. In Proceedings of the 2014 International Conference on Big Data Science and Computing, pages 1–7.

Sun, Z., Strang, K. D., and Yearwood, J. (2014). Analytics service oriented architecture for enterprise information systems. In Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services, pages 508–516.

Tianyuan, Z. (2018). Telecom customer segmentation and precise package design by using data mining (Doctoral dissertation).

Walt, S. V. D., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. Computing in science & engineering, 13(2), 22-30.

Wang, C.-H. (2009). Outlier identification and market segmentation using kernel-based clustering techniques. Expert Systems with Applications, 36(2):3744–3750.

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, 3-13.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Lukauskas, S., Gemperline, D. C., ... & de Ruiter, J. Brian, Chris Fonnesbeck, Antony Lee, and Adel Qalieh. mwaskom/seaborn: v0. 8.1 (september 2017), September 2017. URL https://doi.org/10.5281/zenodo, 883859.

Wedel, M. and Kamakura, W. A. (2012). Market segmentation: Conceptual and methodological foundations, volume 8. Springer Science & Business Media.

Yankelovich, D. and Meer, D. (2006). Rediscovering market segmentation. Harvard business review, 84(2):122.

Appendices

Appendix A: Code

```
#!/usr/bin/env python
# coding: utf-8

# In[1]:


# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load in

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import warnings
warnings.filterwarnings("ignore", category=FutureWarning)
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder, MinMaxScaler
from sklearn.preprocessing import OneHotEncoder
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn import model_selection

# Input data files are available in the "../input/" directory.
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under
the input directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# Any results you write to the current directory are saved as output.


# In[2]:


#Data loading
train_data = pd.read_csv("../input/patora/Dec_2018.csv")
```

```
test_data = pd.read_csv("../input/patora/Feb_2019.csv")
# In[3]:


train_data.head()


# **DATA PREPROCESSING**

# In[4]:



train_data.columns


# In[5]:


#renaming the columns to reduce typing errors
train_data.columns = ['CRM', 'Age', 'Data Revenues', 'Data Traffic', 'Device_Type', 'ETP',
       'Gender', 'IVR', 'MFS', 'MPT', 'Month-Year', 'Msisdn','Number of Records', 'SMS',
       'Site Name', 'Sms Revenues', 'Sms Traffic', 'Technology',
       'Total Revenues', 'USS', 'Vas Revenues', 'Voice Revenues',
       'Voice Traffic']


# In[6]:


test_data.columns


# In[7]:


test_data.columns = ['CRM', 'Age', 'Data_Revenues', 'Data_Traffic', 'Device_Type', 'ETP',
       'Gender', 'IVR', 'MFS', 'MPT', 'Month-Year', 'Msisdn', 'SMS',
       'Site_Name', 'Sms_Revenues', 'Sms_Traffic', 'Technology',
       'Total_Revenues', 'USS', 'Vas_Reveues', 'Voice_Revenues',
       'Voice_Traffic']
```

In[8]:

```
#eliminating columns with high missing values
train_data = train_data.drop(['CRM', 'ETP', 'IVR', 'MFS', 'MPT', 'SMS', 'Site Name'], axis = 1)
test_data = test_data.drop(['CRM', 'ETP', 'IVR', 'MFS', 'MPT', 'SMS', 'Site_Name'], axis = 1)
```

In[9]:

```
#filling out missing values in both train and test datasets
train_data['Age'].fillna(method = 'ffill', inplace = True)
train_data['Gender'].fillna(method = 'ffill', inplace = True)
train_data['Technology'].fillna(method = 'ffill', inplace = True)
train_data['Device_Type'].fillna(method = 'ffill', inplace = True)
#filling out missing values in testset
test_data['Age'].fillna(method = 'ffill', inplace = True)
test_data['Gender'].fillna(method = 'ffill', inplace = True)
test_data['Technology'].fillna(method = 'ffill', inplace = True)
test_data['Device_Type'].fillna(method = 'ffill', inplace = True)
```

In[10]:

```
genders = train_data['Gender'].unique()
devices = train_data['Device_Type'].unique()
Technologies = train_data['Technology'].unique()
print('Customer categories in the data are : ' + str(genders) )
print ('Devices used : ' + str(devices) )
print('Technologies used are : ' + str(Technologies))
```

In[11]:

```
#the genders can be specified into male, gender or business as follows:
gender_dict = {'F':'Female', 'FEMALE':'Female', 'Female':'Female', 'female':'Female',
'FeMale':'Female',
            'm':'Male', 'M':'Male', 'MALE':'Male', 'Male':'Male', 'male':'Male',
             'BUSINESS':'Businesses', 'COMPANY':'Businesses', 'SCHOOL':'Businesses'}
train_data['Gender'] = train_data['Gender'].map(gender_dict)
test_data['Gender'] = test_data['Gender'].map(gender_dict)
```

In[12]:

```
train_data['Gender'].unique()
```

In[13]:

```
#in this segment, we make an assumption that the most recent technology will be used in instances
#where other earlier technologies are present.
#for example, where 2G, 3G and 4G will be in use, 4G then is more commonly used.
#this assumption however has several setbacks.
#for example, in Nairobi, some people use basic phones which cannot operate with 4G network.
#However, compared to the total population, basic phones account only for 20% of the total number of devices.
def set_value(row_number, assigned_value):
   return assigned_value[row_number]
technology_dict = {'2G':'2G', '2G_3G':'3G', '2G_3G_4G':'4G'}
train_data['Usedtech'] = train_data['Technology'].apply(set_value, args=(technology_dict, ))
#since this is the target variable, this procedure will not be done for the test data
```

In[14]:

```
#looking at the age distribution
minimum_age = train_data['Age'].min()
maximum_age = train_data['Age'].max()
print('Minimum age is :  ' +str(minimum_age) + '\n maximum age is : '+ str(maximum_age))
```

In[15]:

```
#rescaling the age to fit between reasonable age brackets
train_data = train_data[~(train_data['Age'] <= 0)]
train_data = train_data[~(train_data['Age'] >=100)]
test_data = test_data[~(test_data['Age']<=0)]
test_data = test_data[~(test_data['Age'] >=100)]
```

In[16]:

```
#the customers can now be binned as follows:
#binning the age to get contributions for different age groups
bins = [0, 25, 45, 75, 100]
labels = ['youth', 'adults', 'elderly', 'very_old']
train_data['age_groups'] = pd.cut(train_data['Age'], bins = bins, labels = labels)
```

```
test_data['age_groups'] = pd.cut(test_data['Age'], bins = bins, labels = labels)
```

EXPLORATORY DATA ANALYSIS

In[17]:

```
#We begin by analyzing the total contributions of voice, data and sms revenue to the total revenue of the company.
#to supplement the statistical results obtained, various data visualizations tools will be used.
'''SMS contribution'''
SMS_revenue = train_data['Sms Revenues'].sum()
'''Voice revenues'''
Voice_revenues = train_data['Voice Revenues'].sum()
'''Data Revenues'''
Data_revenues = train_data['Data Revenues'].sum()
'''total revenues'''
total_Revenue = train_data['Total Revenues'].sum()
print("The contribution of SMS revenues to the total revenue of the company is : {:.2f}".format((SMS_revenue/total_Revenue)*100) + "%")
print("The contribution of Voice revenues to the total revenue of the company is : {:.2f}".format((Voice_revenues/total_Revenue)*100) + "%")
print ("The contribution of Data Revenue to the total revenue of the companu is {:.2f}".format((Data_revenues/total_Revenue)*100) + "%")
```

In[18]:

```
#plotting the results
x = SMS_revenue
y = Voice_revenues
z = Data_revenues
labels = "SMS", "Voice", "Data"
sizes = [x, y, z]
colors =  ['gold', 'yellowgreen', 'lightcoral']
explode = (0.1, 0, 0)  # explode 1st slice
plt.pie(sizes, explode=explode, labels=labels, colors=colors,
autopct='%1.2f', shadow=True, startangle=90)

plt.axis('equal')
plt.show()
```

> The pie chart above corresponds with the previous calculations above. Data revenues are the highest contributors of the total revenue while SMS revenues have the least contributions.

In[19]:

```
#checking out the distribution of age groups among the clients
sns.countplot(x = 'age_groups', data = train_data)


# The company's clients are mostly adults

# In[20]:


#having a clear understanding of the individual categories of the clients in terms of their preferred services i.e sms, data or voice calls
youth = train_data[train_data['age_groups'] == 'youth']
adults = train_data[train_data['age_groups'] == 'adults']
elderly = train_data[train_data['age_groups'] == 'elderly']
very_old = train_data[train_data['age_groups'] == 'very_old']


# In[21]:


#analyzing the revenues of the youth by the different contributor
SMS_revenue_youth = youth['Sms Revenues'].sum()
Voice_revenue_youth = youth['Voice Revenues'].sum()
Data_revenues_youth  = youth['Data Revenues'].sum()
total_revenues_youth = youth['Total Revenues'].sum()
youth_sms_revenue  = (SMS_revenue_youth/total_revenues_youth)*100
youth_voice_revenue = (Voice_revenue_youth/total_revenues_youth)*100
youth_data_revenue = (Data_revenues_youth/total_revenues_youth)*100
labels = "SMS", "Voice", "Data"
sizes = [SMS_revenue_youth, Voice_revenue_youth, Data_revenues_youth]
colors =  ['gold', 'yellowgreen', 'lightcoral']
explode = (0.1, 0, 0)  # explode 1st slice
plt.pie(sizes, explode=explode, labels=labels, colors=colors,
autopct='%1.2f', shadow=True, startangle=90)

plt.axis('equal')
plt.show()


# > The youth are heavy consumers of data

# In[22]:


#adults
SMS_revenue_adults = adults['Sms Revenues'].sum()
Voice_revenue_adults = adults['Voice Revenues'].sum()
```

```
Data_revenues_adults  = adults['Data Revenues'].sum()
total_revenues_adults = adults['Total Revenues'].sum()
adults_sms_revenue  = (SMS_revenue_adults/total_revenues_adults)*100
adults_voice_revenue = (Voice_revenue_adults/total_revenues_adults)*100
adults_data_revenue = (Data_revenues_adults/total_revenues_adults)*100
#plotting the results
labels = "SMS", "Voice", "Data"
sizes = [SMS_revenue_adults, Voice_revenue_adults, Data_revenues_adults]
colors =  ['gold', 'yellowgreen', 'lightcoral']
explode = (0.1, 0, 0)  # explode 1st slice
plt.pie(sizes, explode=explode, labels=labels, colors=colors,
autopct='%1.2f', shadow=True, startangle=90)

plt.axis('equal')
plt.show()


# Adults consume both voice and data services significantly
#

# In[23]:


#elderly
SMS_revenue_elderly = elderly['Sms Revenues'].sum()
Voice_revenue_elderly = elderly['Voice Revenues'].sum()
Data_revenues_elderly  = elderly['Data Revenues'].sum()
total_revenues_elderly = elderly['Total Revenues'].sum()
elderly_sms_revenue  = (SMS_revenue_elderly/total_revenues_elderly)*100
elderly_voice_revenue = (Voice_revenue_elderly/total_revenues_elderly)*100
elderly_data_revenue = (Data_revenues_elderly/total_revenues_elderly)*100
#plotting the results
labels = "SMS", "Voice", "Data"
sizes = [SMS_revenue_elderly, Voice_revenue_elderly, Data_revenues_elderly]
colors =  ['gold', 'yellowgreen', 'lightcoral']
explode = (0.1, 0, 0)  # explode 1st slice
plt.pie(sizes, explode=explode, labels=labels, colors=colors,
autopct='%1.2f', shadow=True, startangle=90)

plt.axis('equal')
plt.show()


# In[24]:


#the very old
SMS_revenue_very_old = very_old['Sms Revenues'].sum()
Voice_revenue_very_old = very_old['Voice Revenues'].sum()
```

```
Data_revenues_very_old  = very_old['Data Revenues'].sum()
total_revenues_very_old = very_old['Total Revenues'].sum()
very_old_sms_revenue  = (SMS_revenue_very_old/total_revenues_very_old)*100
very_old_voice_revenue = (Voice_revenue_very_old/total_revenues_very_old)*100
very_old_data_revenue = (Data_revenues_very_old/total_revenues_very_old)*100
#plotting the results
labels = "SMS", "Voice", "Data"
sizes = [SMS_revenue_very_old, Voice_revenue_very_old, Data_revenues_very_old]
colors =  ['gold', 'yellowgreen', 'lightcoral']
explode = (0.1, 0, 0)  # explode 1st slice
plt.pie(sizes, explode=explode, labels=labels, colors=colors,
autopct='%1.2f', shadow=True, startangle=90)

plt.axis('equal')
plt.show()
```

The elderly and very old are heavy users of voice bundles

In[25]:

```
train_data.groupby('age_groups').sum()[['Total Revenues']]
```

In[26]:

```
#we can then proceed to find the contributions of the different technologies to the total revenue
of the company first.
two_G = train_data[train_data['Usedtech'] == '2G']
three_G = train_data[train_data['Usedtech'] == '3G']
four_G = train_data[train_data['Usedtech'] == '4G']
```

In[27]:

```
#plotting their contributions
two_G_revenue = (two_G['Total Revenues'].sum()/train_data['Total Revenues'].sum())*100
three_G_revenue = (three_G['Total Revenues'].sum()/train_data['Total Revenues'].sum())*100
four_G_revenue = (four_G['Total Revenues'].sum()/train_data['Total Revenues'].sum())*100
#plotting the results
labels = "two_G_revenue", "three_G_revenue", "four_G_revenue"
sizes = [two_G_revenue, three_G_revenue, four_G_revenue]
colors =  ['gold', 'yellowgreen', 'lightcoral']
explode = (0.1, 0, 0)  # explode 1st slice
plt.pie(sizes, explode=explode, labels=labels, colors=colors,
autopct='%1.2f', shadow=True, startangle=90)
```

```
plt.axis('equal')
plt.show()


# 4G technology has the highest contribution to the company's revenue at 50.32%

# In[28]:


#we can further explore how different bundles contribute to the total revenue of the individual revenues.
#2G

sms_revenue = (two_G['Sms Revenues'].sum()/two_G['Total Revenues'].sum())*100
voice_revenue = (two_G['Voice Revenues'].sum()/two_G['Total Revenues'].sum())*100
data_revenue = (two_G['Data Revenues'].sum()/two_G['Total Revenues'].sum())*100
#visualizing the contributions
print(sms_revenue)
print(voice_revenue)
print(data_revenue)
plt.bar(['sms_revenue', 'voice_revenue', 'data_revenue'], [sms_revenue, voice_revenue, data_revenue])
plt.xlabel("Bundles")
plt.ylabel("Percentage contribution")
plt.title("Contribution of voice, sms and data to the total 2G revenue")
plt.show()


# In[29]:


#3G
sms_revenue = (three_G['Sms Revenues'].sum()/three_G['Total Revenues'].sum())*100
voice_revenue = (three_G['Voice Revenues'].sum()/three_G['Total Revenues'].sum())*100
data_revenue = (three_G['Data Revenues'].sum()/three_G['Total Revenues'].sum())*100
#visualizing the contributions
print(sms_revenue)
print(voice_revenue)
print(data_revenue)
plt.bar(['sms_revenue', 'voice_revenue', 'data_revenue'], [sms_revenue, voice_revenue, data_revenue])
plt.xlabel("Bundles")
plt.ylabel("Percentage contribution")
plt.title("Contribution of voice, sms and data to the total 3G revenue")
plt.show()


# In[30]:
```

```
#4G
sms_revenue = (four_G['Sms Revenues'].sum()/four_G['Total Revenues'].sum())*100
voice_revenue = (four_G['Voice Revenues'].sum()/four_G['Total Revenues'].sum())*100
data_revenue = (four_G['Data Revenues'].sum()/four_G['Total Revenues'].sum())*100
#visualizing the contributions
print(sms_revenue)
print(voice_revenue)
print(data_revenue)
plt.bar(['sms_revenue', 'voice_revenue', 'data_revenue'], [sms_revenue, voice_revenue, data_revenue])
plt.xlabel("Bundles")
plt.ylabel("Percentage contribution")
plt.title("Contribution of voice, sms and data to the total 4G revenue")
plt.show()
```

Final Data Preprocessing

In[31]:

```
train_data = train_data.drop(['Month-Year','Msisdn', 'Number of Records','USS', 'Vas Revenues'], axis = 1)
```

In[32]:

```
test_data = test_data.drop(['Month-Year','Msisdn','USS', 'Vas_Reveues' ], axis = 1)
```

In[33]:

```
train_data.info()
```

Machine learning models work with numerical data. It's important to convert the objects above into a form that can be easily understood for easier classification

In[34]:

```
Genders = pd.get_dummies(train_data['Gender'])
train_data = train_data.join(Genders)
```

In[35]:

```
Devices = pd.get_dummies(train_data['Device_Type'])
train_data = train_data.join(Devices)
```

In[36]:

```
ages = pd.get_dummies(train_data['age_groups'])
train_data = train_data.join(ages)
```

In[37]:

```
Ages = pd.get_dummies(test_data['age_groups'])
test_data = test_data.join(ages)
```

In[38]:

```
genders = pd.get_dummies(test_data['Gender'])
test_data = test_data.join(genders)
```

In[39]:

```
devices = pd.get_dummies(test_data['Device_Type'])
test_data = test_data.join(devices)
```

In[40]:

```
train_data = train_data.iloc[:100000]
```

In[41]:

```
X = train_data.drop(['Technology', 'Device_Type', 'Gender', 'age_groups', 'Usedtech'], axis = 1)
y = train_data['Usedtech']
le = LabelEncoder()
y = le.fit_transform(y)
train_data['Label_Technologies'] = y
```

```
train_data.Label_Technologies.head()


# 1 represents 3G
# 2 represents 4G
# 0 represents 2G

# In[42]:


#creating a model factory
models = []
models.append(('LogisticRegressor', LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('CART', DecisionTreeClassifier()))
models.append(('NB', GaussianNB()))
#models.append(('SVM', SVC()))
results = []
names = []
scores = 'accuracy'
for name, model in models:
    kfold = model_selection.KFold(n_splits = 10, random_state = 7)
    cross_validation_results = model_selection.cross_val_score(model, X, y, cv = kfold, scoring
= scores)
    results.append(cross_validation_results)
    names.append(name)
    message = "%s: %f (%f)" % (name, cross_validation_results.mean(),
cross_validation_results.std())
    print(message)
```

LogisticRegression proves to have the best cross_validation score. this approach is however not reliable because is has used a significantly less amount of the available data. the constraints are brought about by the fact that the training of the dataset requires powerful computational resources which are not availale

MAKING THE PREDICTIONS USING LogisticRegression

In[43]:

```
train_data.columns


# In[44]:


predictor_values = ['Data Revenues', 'Data Traffic', 'Sms Revenues', 'Sms Traffic',
```

```
                'Total Revenues', 'Voice Revenues', 'Voice Traffic', 'Basic Phone', 'Camera',
                'Feature Phone', 'Mobile broadband PCI card', 'Phablet', 'Router', 'Smartphone',
      'Tablet', 'USB Modem', 'youth', 'adults', 'elderly', 'very_old']
x = train_data[predictor_values]
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 42)
scaler = MinMaxScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)


# In[45]:


model = LogisticRegression(penalty = 'l2', C=0.1)
model.fit(x_train, y_train)
predictions = model.predict(x_test)
scores = classification_report(y_test, predictions)
print(scores)


# PREPARING THE TEST SET

# In[46]:


test_data.columns


# In[47]:


test_data.columns = ['Age', 'Data Revenues', 'Data Traffic', 'Device Type', 'Gender',
      'Sms Revenues', 'Sms Traffic', 'Technology', 'Total Revenues',
      'Voice Revenues', 'Voice Traffic', 'age roups', 'youth', 'adults',
      'elderly', 'very old', 'Businesses', 'Female', 'Male', 'Basic Phone',
      'Camera', 'Feature Phone', 'IoT Device', 'Mobile broadband PCI card',
      'Phablet', 'Router', 'Smartphone', 'Tablet', 'USB Modem', 'Wearable']
test_features = ['Data Revenues', 'Data Traffic', 'Sms Revenues', 'Sms Traffic', 'Total Revenues',
            'Voice Revenues', 'Voice Traffic', 'Basic Phone', 'Camera', 'Feature Phone',
      'Mobile broadband PCI card', 'Phablet', 'Router', 'Smartphone',
      'Tablet', 'USB Modem', 'youth', 'adults', 'elderly', 'very old']
X_test = test_data[test_features]
# In[48]:


'''Filling the missing values with ffil, essence is to minimize NaN data'''

X_test['youth'].fillna(method = 'ffill', inplace = True)
X_test['adults'].fillna(method = 'ffill', inplace = True)
```

```
X_test['elderly'].fillna(method = 'ffill', inplace = True)
X_test['very old'].fillna(method = 'ffill', inplace = True)


# In[49]:


X_test.info()


# MAKING THE PREDICTIONS****

# In[50]:


new_predictions = model.predict(X_test)


# Offers for the Different Clients
# #remember 0 represents 2g: 1 represents 3g and 2 represents 4g

# In[51]:


X_test['Predicted Technology'] = new_predictions
def Suggested_Offers(row):
    if row['Predicted Technology'] == 0:
        return "Suggest Voice Plans"
    if row['Predicted Technology'] == 1:
        return "Suggest voice and data plans"
    if row['Predicted Technology'] == 2:
        return "Strongly Suggest Data Plans. Voice plans should also be considered"
new_data = X_test.assign(Offers = X_test.apply(Suggested_Offers, axis = 1))


# In[52]:


new_data.head()


# In[ ]:
```

Appendix B: Ethical Approval

Strathmore UNIVERSITY

3rd December 2019

Mr Omonge, Jevans
jevans.omonge@strathmore.edu

Dear Mr Omonge,

RE: An Algorithm for Targeted Marketing: A Case of Telkom Kenya

This is to inform you that SU-IERC has reviewed and **approved** your above research proposal. Your application approval number is **SU-IERC0568/19**. The approval period is **3rd December, 2019 to 2nd December, 2020.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-IERC.
iii. Death and life threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-IERC within 72 hours of notification
iv. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-IERC within 72 hours
v. Clearance for export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days upon completion of the study to SU-IERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI) https://oris.nacosti.go.ke and also obtain other clearances needed.

Yours sincerely,

Dr Virginia Gichuru,
Secretary; SU-IERC

Cc: Prof Fred Were,
Chairperson; SU-IERC

STRATHMORE UNIVERSITY INSTITUTIONAL ETHICS REVIEW COMMITTEE (SU-IERC)
03 DEC 2019
TEL: +254 (0)703 034 000
P. O. Box 59857 - 00200
NAIROBI - KENYA

Appendix C: Turnitin Report

A CUSTOMER SEGMENTATION MODEL USING LOGISTIC REGRESSION A CASE OF TELKOM KENYA.docx

ORIGINALITY REPORT

11%	7%	4%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Indian School of Business Student Paper	1%
2	Submitted to University College London Student Paper	1%
3	pythonspot.com Internet Source	<1%
4	Submitted to UT, Dallas Student Paper	<1%
5	Submitted to Ghana Technology University College Student Paper	<1%
6	Submitted to Istanbul Aydin University Student Paper	<1%
7	Submitted to University of Leeds Student Paper	<1%

8	www.slideshare.net Internet Source	<1%
9	machinelearningmastery.com Internet Source	<1%
10	mafiadoc.com Internet Source	<1%
11	stackoverflow.com Internet Source	<1%
12	Submitted to The University of the South Pacific Student Paper	<1%
13	Submitted to University of Southampton Student Paper	<1%
14	Submitted to Universiti Teknologi Petronas Student Paper	<1%
15	Submitted to Nottingham Trent University Student Paper	<1%
16	www.coursehero.com Internet Source	<1%

	Student Paper	<1%
16	www.coursehero.com Internet Source	<1%
17	Puneet Mathur. "Machine Learning Applications Using Python", Springer Science and Business Media LLC, 2019 Publication	<1%
18	eprints.uthm.edu.my Internet Source	<1%
19	Submitted to Strathmore University Student Paper	<1%
20	"Emerging Technologies in Data Mining and Information Security", Springer Science and Business Media LLC, 2019 Publication	<1%
21	Submitted to RDI Distance Learning Student Paper	<1%
22	Flora Ma Díaz-Pérez, Carlos G. García-González, Alan Fyall. "The use of the CHAID algorithm for determining tourism segmentation: A purposeful outcome", Heliyon, 2020 Publication	<1%

23	Submitted to University of Sheffield Student Paper	<1%
24	vitela.javerianacali.edu.co Internet Source	<1%
25	Submitted to University of Central Lancashire Student Paper	<1%
26	www.docstoc.com Internet Source	<1%
27	talisman-intl.com Internet Source	<1%
28	link.springer.com Internet Source	<1%
29	Submitted to University of Stirling Student Paper	<1%
	Submitted to Technische Universiteit Delft	

30	Student Paper	<1%
31	Submitted to Universiti Teknologi MARA Student Paper	<1%

32	Zhaohao Sun, Huasheng Zou, Kenneth Strang. "Chapter 16 Big Data Analytics as a Service for Business Intelligence", Springer Science and Business Media LLC, 2015 Publication	<1%
33	Submitted to City University of Hong Kong Student Paper	<1%
34	monkeysandbox.com Internet Source	<1%
35	www.irantahgig.ir Internet Source	<1%
36	"Artificial Neural Networks", Springer Science and Business Media LLC, 2021 Publication	<1%
37	lib.dr.iastate.edu Internet Source	<1%
38	Submitted to University of Witwatersrand Student Paper	<1%
39	Prasad Kasibhatla, Tomás Sherwen, Mathew J. Evans, Lucy J. Carpenter et al. " Global impact of nitrate photolysis in sea-salt aerosol on NO , OH, and O in the marine boundary layer ", Atmospheric Chemistry and Physics, 2018 Publication	<1%

40	1hj0ztcvh9.download2.org Internet Source	<1%
41	cmsexternal.nt.gov.au Internet Source	<1%
42	"Intelligent Computing, Networking, and Informatics", Springer Science and Business Media LLC, 2014 Publication	<1%
43	library.kic.ae Internet Source	<1%

44	app.dtmsys.com Internet Source	<1%
45	digitalcommons.fiu.edu Internet Source	<1%
46	www.learntechlib.org Internet Source	<1%
47	docplayer.net Internet Source	<1%

48	Maria M. Abad Grau. "Machine learning methods for the market segmentation of the performing arts audiences", International Journal of Business Environment, 2009 Publication	<1%
49	ecommons.usask.ca Internet Source	<1%
50	lnu.diva-portal.org Internet Source	<1%
51	es.slideshare.net Internet Source	<1%
52	Submitted to University of Cincinnati Student Paper	<1%
53	aramse02.wordpress.com Internet Source	<1%
54	"Social Vulnerability in Europe", Springer Science and Business Media LLC, 2010 Publication	<1%
55	Marcelo R.P. Ferreira, Francisco de A.T. de Carvalho. "Kernel fuzzy c-means with automatic variable weighting", Fuzzy Sets and Systems, 2014 Publication	<1%

56	Wang, C.H.. "Outlier identification and market segmentation using kernel-based clustering techniques", Expert Systems With Applications, 200903 Publication	<1%
57	www.tandfonline.com Internet Source	<1%
58	"Universal Access in Human-Computer Interaction. Applications and Services", Springer Science and Business Media LLC, 2009 Publication	<1%
59	"Competitiveness in Emerging Markets", Springer Science and Business Media LLC, 2018 Publication	<1%
60	Flora Ma Díaz-Pérez, Ma Bethencourt-Cejas. "CHAID algorithm as an appropriate analytical method for tourism market segmentation", Journal of Destination Marketing & Management, 2016 Publication	<1%

61	creativecommons.org Internet Source	<1%
62	open.library.ubc.ca Internet Source	<1%
63	Uday Kamath, John Liu, James Whitaker. "Deep Learning for NLP and Speech Recognition", Springer Science and Business Media LLC, 2019 Publication	<1%
64	"Social Media: The Good, the Bad, and the Ugly", Springer Science and Business Media LLC, 2016 Publication	<1%
65	"Information and Communication Technology", Springer Science and Business Media LLC, 2015 Publication	<1%
66	"Advances in Communication and Computational Technology", Springer Science and Business Media LLC, 2021 Publication	<1%

Exclude quotes On Exclude matches Off