**ABSTRACT:**

We live in a world where a large and vast amount of data is collected daily. Analysing such data is an important need. In the modern era of innovation, where there is a large competition to be better than everyone, the business strategy needs to be according to the modern conditions. The business done today runs on the basis of innovative ideas as there are a large number of potential customers who are confused about what to buy and what not to buy. The companies doing the business are also not able to diagnose the target potential customers. This is where machine learning comes into picture; the various algorithms are applied to identify the hidden patterns in the data for better decision making. The concept of which customer segment to target is done using the customer segmentation process using the clustering technique.

In this paper, the clustering algorithm used is K-means algorithm which is the partitioning algorithm, to segment the customers according to the similar characteristics. To determine the optimal clusters, the elbow method is used.

**INTRODUCTION:**

Over the years, the competition amongst businesses has increased and the large historical data that is available has resulted in the widespread use of data mining techniques in extracting the meaningful and strategic information from the database of the organisation.

Data mining is the process where methods are applied to extract data patterns in order to present it in the human readable format which can be used for the purpose of decision support. According to, Clustering techniques consider data tuples as objects.They partition the data objects into groups or clusters so that objects within a cluster are similar to one another and dissimilar to objects in other clusters.Customer Segmentation is the process of division of customer base into several groups called as customer segments such that each customer segment consists of customers who have similar characteristics. The segmentation is based on the similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

The customer segmentation has the importance as it includes, the ability to modify the programs of market so that it is suitable to each of the customer segment,

support in business decision; identification of products associated with each customer segment and to manage the demand and supply of that product; identifying and targeting the potential customer base, and predicting customer defection, providing directions in finding the solutions. The thrust of this paper is to identify customer segments using the data mining approach, using the partitioning algorithm called K-means clustering algorithm. The elbow method determines the optimal clusters.

**PROBLEM STATEMENT:**

Customer segmentation is the method of distributing a customer base into collections of people based on mutual characteristics so organizations can market to group efficiently and competently individually.

The purpose of segmenting customers is to determine how to correlate to customers in multiple segments to maximize customer benefits. Perfectly done customer segmentation empowers marketers to interact with every customer in the best efficient approach.

**DATA DESCRIPTION:**

The data description phase starts with an initial data collection and proceeds with activities in order to get familiar with the data. Identifying data quality problems, discovering first insights into the data and detecting interesting subsets to form hypotheses from hidden information are activities of this step. Datawhich contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.It has 541909 rows and 8 columns.

**DATASET PREPARATION:**

The dataset is UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers. It contains 8 features and 541909 observations of a complete year I.e. 01/12/2010 and 09/12/2011. Below Table shows the data features.

**Data-set description**

| **Feature Name**  InvoiceNo  StockCode  Description  Quantity  InvoiceDate  UnitPrice  CustomerID  Country | **Type**  object  object  object  int64  object  Float64  Float64  object |
| --- | --- |

**FEATURE BREAKDOWN:**

***InvoiceNo****: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter ‘c’, it indicates a cancellation.*

***StockCode****: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.*

**Description**: Product (item) name. Nominal.

**Quantity**: The quantities of each product (item) per transaction. Numeric.

**InvoiceDate**: Invoice Date and time. Numeric, the day and time when each transaction was generated.

**UnitPrice**: Unit price. Numeric, Product price per unit in sterling.

**CustomerID**: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.

**Country**: Country name. Nominal, the name of the country where each customer resides.

**EXPLORATORY DATA ANALYSIS:**

If we want to explain EDA in simple terms, it means trying to understand the given data much better, so that we can make some sense out of it. Using univariate frequency analysis was conducted to describe key characteristics of each feature including, minimum and maximum value, average, standard deviation and others. It was also used to produce a value distribution and identify missing values, and outliers.

EDA is a process of examining the available dataset to discover patterns, spot anomalies, test hypotheses, and check assumptions using statistical measures. In this chapter, we are going to discuss the steps involved in performing top notch exploratory data analysis

In statistics, A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing tasked in Python uses data visualization to draw meaningful patterns and insights

* **DATA ANALYSIS:**

This is one of the most crucial steps that deals with descriptive statistics and analysis of the data. The main tasks involve summarizing the data, finding the hidden correlation and relationships among the data, developing predictive models, evaluating the models, and calculating the accuracies. Some of the techniques used for data summarization are summary tables, graphs, descriptive statistics, inferential statistics, correlation statistics, searching, grouping, and mathematical models.

* **DATA SOURCING**

Data Sourcing is the process of finding and loading the data into our system. Broadly there are two ways in which we can find data.

1. Private Data
2. Public Data

Data collected from several sources must be stored in the correct format and transferred to the right information technology personnel within a company. As mentioned previously, data can be collected from several objects on several events using different types of sensors and storage tools.

* **DATA PREPROCESSING:**

A dataset may contain noise, missing values, and inconsistent data, thus, pre-processing of data is essential to improve the quality of data and time required in the data mining.

* **DATA CLEANING**

After completing the Data Sourcing, the next step in the process of EDA is Data Cleaning. It is very important to get rid of the irregularities and clean the data after sourcing it into our system.

Irregularities are of different types of data.

* Missing Values
* Incorrect Format
* Incorrect Headers
* Anomalies/Outliers
* **DATA TRANSFORMATION:**

Data transformation is the process of normalizing and aggregating the data to further improve the efficiency and accuracy of data mining.

* **DATA DEDUPLICATION:**

It is very likely that your dataset contains duplicate rows. Removing them is essential to enhance the quality of the dataset.

* **MISSING VALUES:**

There is a representation of each service and product for each customer. Missing values may occur because not all customers have the same subscription. Some of them may have a number of services and others may have something different. In addition, there are some columns related to system configurations and these columns may have null values but in our orange telecom data set there are no null values present

If there are missing values in the Dataset before doing any statistical analysis, we need to handle those missing values.

There are mainly three types of missing values.

1. MCAR (Missing completely at random): These values do not depend on any other features.
2. MAR (Missing at random): These values may be dependent on some other features.

MNAR (Missing not at random): These missing values have some reason for why they are missing.

* **DROPPING MISSING VALUES:**

One of the ways to handle missing values is to simply remove them from our dataset. We have know that we can use the isnull() and notnull() functions from the pandas library to determine null values

* **HANDLING OUTLIERS:**

Outliers are data points that diverge from other observations for several reasons. During the EDA phase, one of our common tasks is to detect and filter these outliers. The main reason for this detection and filtering of outliers is that the presence of such outliers can cause serious issues in statistical analysis.

There are two types of outliers:

* **UNIVARIATE OUTLIERS:**

Univariate outliers are the data points whose values lie beyond the range of expected values based on one variable.

* **MULTIVARIATE OUTLIERS:**

While plotting data, some values of one variable may not lie beyond the expected range, but when you plot the data with some other variable, these values may lie far from the expected value.

* **MEASURES OF CENTRAL TENDENCY:**

The measure of central tendency tends to describe the average or mean value of datasets that is supposed to provide an optimal summarization of the entire set of measurements. This value is a number that is in some way central to the set. The most common measures for analysing the distribution frequency of data are the mean, median, and mode.

* **MEASURES OF DISPERSION**:

The second type of descriptive statistics is the measure of dispersion, also known as a measure of variability. If we are analyzing the dataset closely, sometimes, the mean/average might not be the best representation of the data because it will vary when there are large variations between the data. In such a case, a measure of dispersion will represent the variability in a dataset much more accurately.

Multiple techniques provide the measures of dispersion in our dataset. Some commonly used methods are standard deviation (or variance), the minimum and maximum values of the variables, range, kurtosis, and skewness.

* **STANDARDIZING VALUES:**

To perform data analysis on a set of values, we have to make sure the values in the same column should be on the same scale. For example, if the data contains the values of the top speed of different companies’ cars, then the whole column should be either in meters/sec scale or miles/sec scale.

* **UNIVARIATE ANALYSIS:**

If we analyse data over a single variable/column from a dataset, it is known as Univariate Analysis. Univariate analysis looks at one feature at a time. When we analyse a feature independently, we are usually mostly interested in the distribution of its values and ignore other features in the dataset

Univariate analysis is the simplest form of analysing data. It means that our data has only one type of variable and that we perform analysis over it. The main purpose of univariate analysis is to take data, summarize that data, and find patterns among the values. It doesn't deal with causes or relationships between the values. Several techniques that describe the patterns found in univariate data include central tendency (that is the mean, mode, and median) and dispersion (that is, the range, variance, maximum and minimum quartiles (including the interquartile range), and standard deviation).

* **BIVARIATE ANALYSIS:**

If we analyse data by taking two variables/columns into consideration from a dataset, it is known as Bivariate Analysis.

* **a)Numeric-Numeric Analysis:**

Analysing the two numeric variables from a dataset is known as numeric-numeric analysis. We can analyse it in three different ways.

* Scatter Plot
* Pair Plot
* Correlation Matrix
* **b) Numeric - Categorical Analysis:**

Analysing the one numeric variable and one categorical variable from a dataset is known as numeric-categorical analysis. We analyse those mainly using mean, median, and box plots.

* **MULTIVARIATE ANALYSIS:**

Multivariate analysis is the analysis of three or more variables. This allows us to look at correlations (that is, how one variable changes with respect to another) and attempt to make predictions for future behaviour more accurately than with bivariate analysis.

One common way of plotting multivariate data is to make a matrix scatter plot, known as a pair plot. A matrix plot or pair plot shows each pair of variables plotted against each other. The pair plot allows us to see both the distribution of single variables and the relationships between two variables

* **CORRELATION AMONG VARIABLES**:

In words, the statistical technique that examines the relationship and explains whether, and how strongly, pairs of variables are related to one another is known as correlation. Correlation answers questions such as how one variable changes with respect to another. If it does change, then to what degree or strength? Additionally, if the relation between those variables is strong enough, then we can make predictions for future behaviour

* **GRAPHICAL REPRESENTATION OF THE RESULTS:**

This step involves presenting the dataset to the target audience in the form of graphs, summary tables, maps, and diagrams. This is also an essential step as the result analysed from the dataset should be interpretable by the business stakeholders, which is one of the major goals of EDA. Most of the graphical analysis techniques include Line chart, Bar chart, Scatter plot, Area plot, and stacked plot Pie chart, Table chart, Polar chart, Histogram, Lollipop chart etc.

**ALGORITHMS:**

**1. LINEAR REGRESSION:**

Linear regression is a supervised machine learning model majorly used in forecasting. Supervised machine learning models are those where we use the training data to build the model and then test the accuracy of the model using the loss function.

Linear regression is one of the most widely known time series forecasting techniques which is used for predictive modelling. As the name suggests, it assumes a linear relationship between a set of independent variables to that of the dependent variable (the variable of interest).

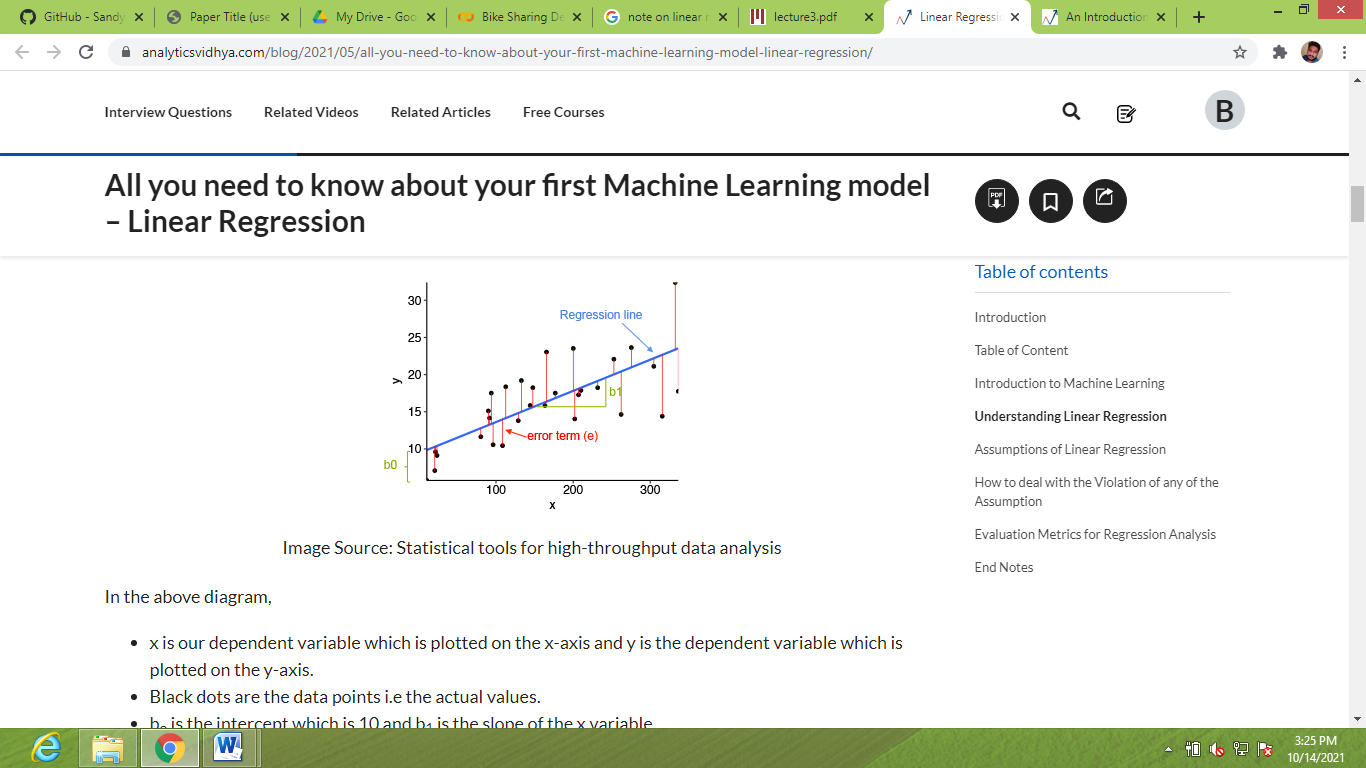
We’re going to fit a line

y = β0 + β1x

to our data. Here, x is called the independent variable or predictor variable, and y is called the dependent variable or response variable. Before we talk about how to do the fit, let’s take a closer look at the important quantities from the fit:

• β1 is the slope of the line: this is one of the most important quantities in any linear regression analysis

• β0 is the intercept of the line.



**2. RIDGE REGRESSION:**

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

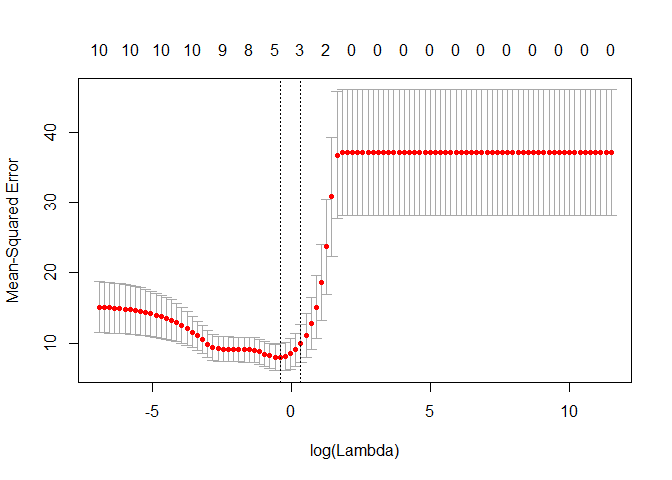
We have concluded that we would like to decrease the model complexity, that is the number of predictors. We could use the forward or backward selection for this, but that way we would not be able to tell anything about the removed variables' effect on the response. Removing predictors from the model can be seen as setting their coefficients to zero. Instead of forcing them to be exactly zero, let's penalize them if they are too far from zero, thus enforcing them to be small in a continuous way. This way, we decrease model complexity while keeping all variables in the model. This, basically, is what Ridge Regression does.



**3. LASSO REGRESSION:**

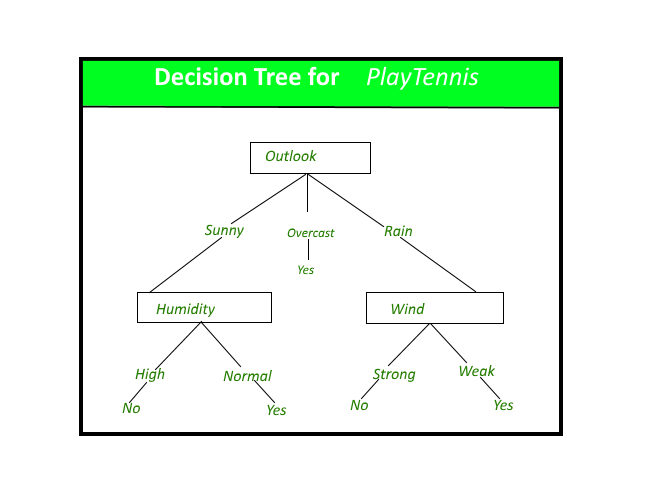
Lasso, or Least Absolute Shrinkage and Selection Operator, is quite similar conceptually to ridge regression. It also adds a penalty for non-zero coefficients, but unlike ridge regression which penalizes sum of squared coefficients (the so-called L2 penalty), lasso penalizes the sum of their absolute values (L1 penalty). As a result, for high values of λ, many coefficients are exactly zeroed under lasso, which is never the case in ridge regression.

The only difference in ridge and lasso loss functions is in the penalty terms. Under lasso, the loss is defined as:

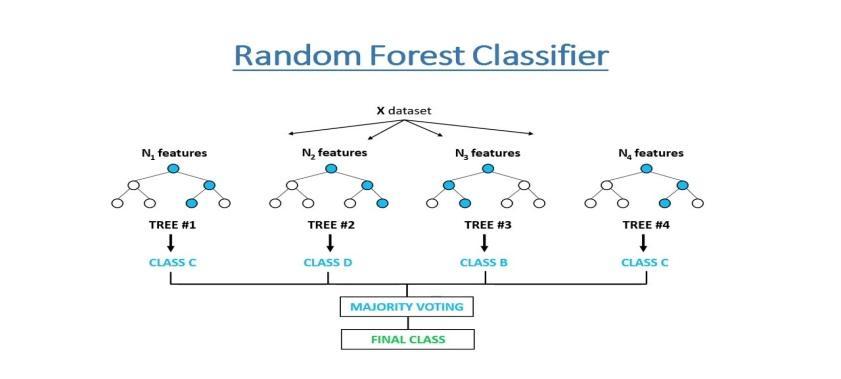


**4.DECISION TREE:**

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. A tree can be *“learned”* by splitting the source set into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called*recursive partitioning*. Decision trees classify instances by sorting them down the tree from the root to some leaf node, which provides the classification of the instance. An instance is classified by starting at the root node of the tree, testing the attribute specified by this node, and then moving down the tree branch corresponding to the value of the attribute as shown in the above figure. This process is then repeated for the subtree rooted at the new node.



**5. RANDOM FOREST:**

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicted out of all.

**6. GRADIENT BOOSTING:**

The term gradient boosting consists of two sub-terms, gradient and boosting. We already know that gradient boosting is a boosting technique. Let us see how the term ‘gradient’ is related here.

Gradient boosting re-defines boosting as a numerical optimization problem where the objective is to minimize the loss function of the model by adding weak learners using gradient descent. Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. As gradient boosting is based on minimising a loss function, different types of loss functions can be used resulting in a flexible technique that can be applied to regression, multi-class classification, etc

**CONCLUSIONS:**

We approached customer segmentation problems from a behavioural aspect with the number of products ordered, average return rate and total spending for each customer. Use of 3 features helped us with the understandability and visualization of the model.

All in all, the dataset was apt to perform an unsupervised machine learning problem. At first, we only had customer data with order information and did not know if they belonged to any group. With the K-means clustering, patterns in the data were found and extended further into groups. We carved out strategies for the formed groups, making meaning out of a dataset that is a dust cloud initially.

**REFERENCES :**

* Analytics Vidhya
* Towards Data Science
* greeks for greeks