0162779
SEMINAR REPORT

BIG DATA ANALYSIS FOR CUSTOMER BEHAVIOR

Authored by:
A.Radhika Patali
17841A0552
CSE-4A

CONTENTS

1. Abstract

Although there are many systems that have implemented customer behavior analytics, it's still an upcoming and unexplored market that has greater potential for better advancements. Big data is one of the most rising technology trends that have the capability for significantly changing the way business organizations use customer behavior to analyze and transform it into valuable insights.

Big data analytics is where advanced analytic techniques operate on big data sets. Hence, big data analytics is really about two things—big data and analytics—plus how the two have teamed up to create one of the most profound trends in business intelligence (BI) today. Let's start by defining advanced analytics, then move on to big data and the combination of the two. Even decision trees can be used efficiently for analyzing data.

At the end of the paper a real time use case is explained with a decision tree algorithm.

---

2. Introduction

Big data

Collection of data that is huge in volume yet growing exponentially with time.Data with so large in size and complexity that none of traditional data and management tools can store it or process it efficiently.Big data is a collection of unstructured data that has very large volume, comes from variety of sources like web ,business organizations etc. in different formats.

The first definition of Big Data comes from Merv Adrian: "Big data exceeds the reach of commonly used hardware environments and software tools to capture, manage, and process it within a tolerable elapsed time for its user population".

Another good definition is given by the McKinsey Global Institute: "Big data refers to data sets whose size is beyond the ability of typical database software tools to capture, store,

manage and analyze".

Types of Big data

Following are the types of Big Data:

1. Structured
2. Unstructured
3. Semi-structured

Structured: Any data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data. The data is stored in RDBMS ,i.e, in tables which have rows and columns.

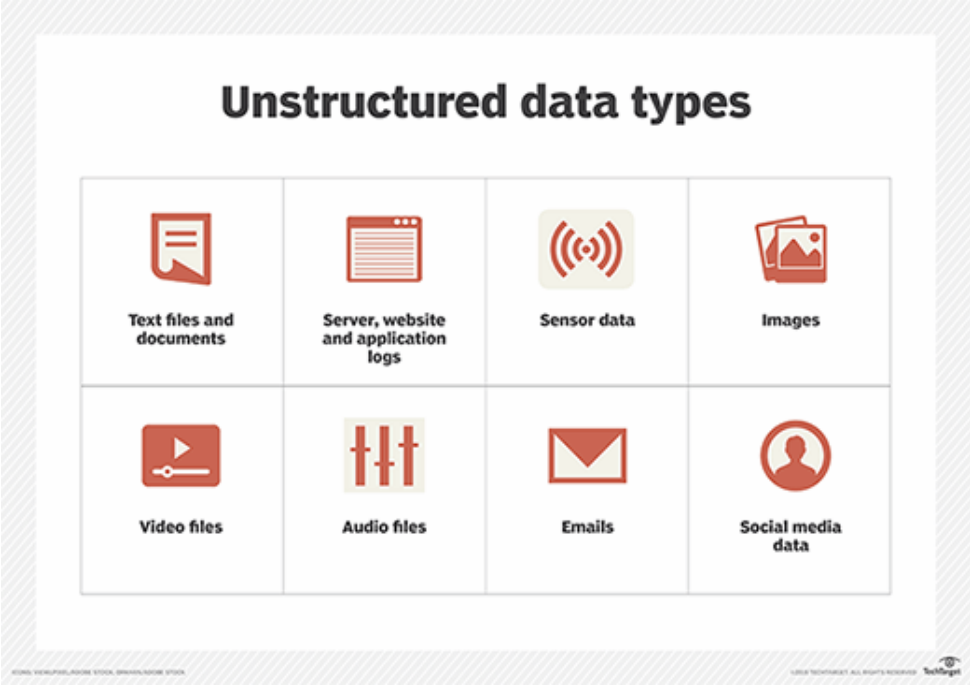
Examples Of Structured Data

An 'Employee' table in a database is an example of Structured Data

college_id	Employee_Name	Gender	Batches	Area
17841A0502	A.Manya Reddy	Female	B1	Uppal
17841A0511	G.Nischala Gangeya	Female	B1	Nagole
17841A0516	J.vaishnavi Devi	Female	B2	Punjagutta
17841A0525	N.Sushitha	Female	B2	Uppal
17841A0543	G.Likitha	Female	B3	Kothapet

Unstructured: Any data with unknown form or the structure is classified as unstructured data.The data is in the form of audio,video,images e.t.c.

Examples Of Unstructured data



Semi-structured: Semi-structured data can contain both the forms of data.the data is in the form of JSON and XML files.

Examples Of Semi-structured Data

Personal data stored in an XML file-

```
<rec><name>Prashant Rao</name><sex>Male</sex><age>35</age></rec>
<rec><name>Seema R.</name><sex>Female</sex><age>41</age></rec>
<rec><name>Satish Mane</name><sex>Male</sex><age>29</age></rec>
<rec><name>Subrato Roy</name><sex>Male</sex><age>26</age></rec>
<rec><name>Jeremiah J.</name><sex>Male</sex><age>35</age></rec>
```

The 5 v's of Big data

Volume, velocity, variety, veracity and value are the five keys to making big data a huge business.

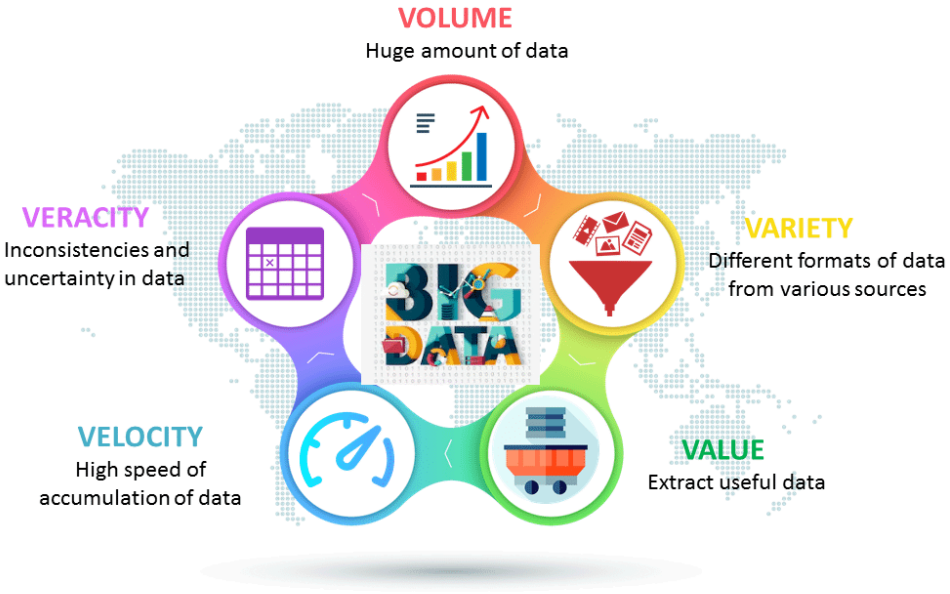
Velocity- Velocity refers to the speed at which the data is generated, collected and analyzed.

Volume- Big data volume defines the 'amount' of data that is produced. The value of data is also dependent on the size of the data.

Value- Although data is being produced in large volumes today, just collecting it is of no use. Instead, data from which business insights are garnered add 'value' to the company.

Variety- While the volume and velocity of data are important factors that add value to a business, big data also entails processing diverse data types collected from varied data sources.

Veracity/Validity- The Veracity of big data or Validity, as it is more commonly known, is the assurance of quality or credibility of the collected data.



Big data Analytics

Earlier, the term 'Analytics' indicated the study of existing data to research potential trends and to analyze the effects of certain decisions or events that can be used for business intelligence to gain various valuable insights. Today's biggest challenge is how to discover all the hidden information through the huge amount of data collected from a varied collection of sources. There comes Big Data Analytics into picture. One of them is the customer behavior analysis which is referred to as customer analytics.

Customer analytics helps to turn big data into big value by allowing the organizations to predict the buyer behavior thereby improving their sales, market optimization, inventory planning, fraud detection and many more applications. A wide range of approaches are available and can be implemented but the one that stands out is the use of decision trees for the purpose of classification that can be efficiently used in customer analytics.

Types of Big data analytics

Here are 4 types of big data analytics:

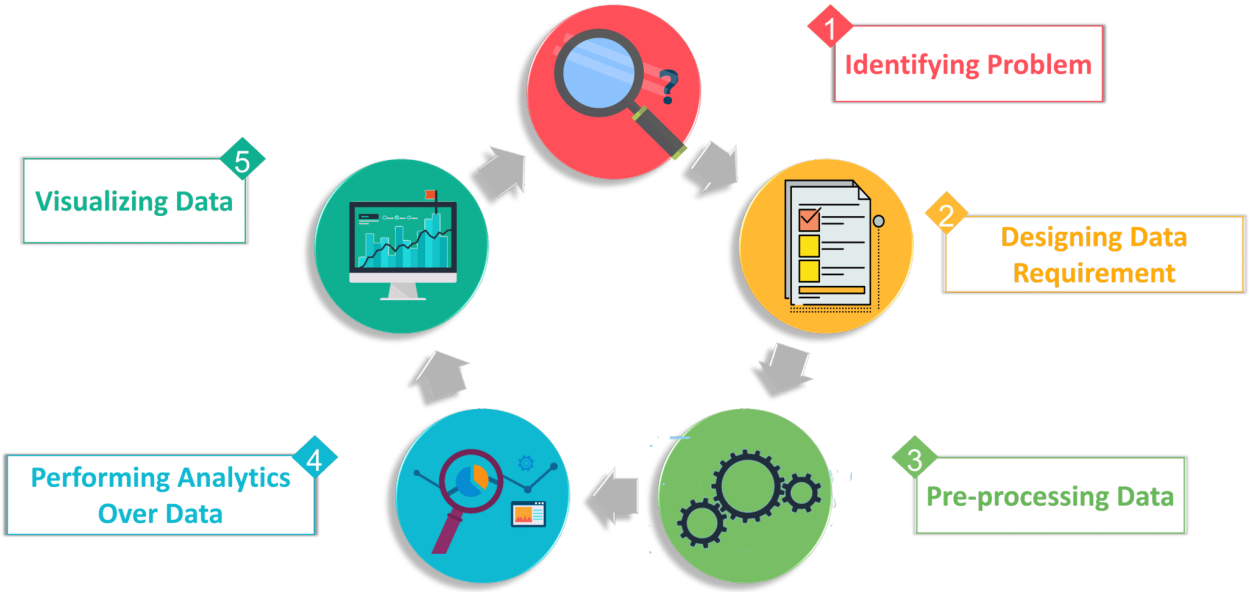
Descriptive Analytics(What has happened?): This technique is the most time-intensive and often produces the least value; however, it is useful for uncovering patterns within a certain segment of customers.

Predictive Analytics(What could happen?): The most commonly used technique; predictive analytics use models to forecast what might happen in specific scenarios.

Prescriptive Analytics(What should we do?): It helps to determine the best solution among a variety of choices, given the known parameters and suggests options for how to take advantage of a future opportunity or mitigate a future risk.

Diagnostic Analytics(Why did this happen?): Data scientists turn to this technique when trying to determine why something happened.

Stages in Big data Analytics



Identifying problem: It begins with a well-defined business case that presents a clear understanding of the justification, motivation and goals of carrying out the analysis.

Designing Data Requirement: This stage is dedicated to identify the datasets required for the analysis project and their sources.

Pre-Processing Data: The data is gathered from all of the data sources that were identified during the previous stage. The acquired data is then subjected to automated filtering for the removal of corrupt data or data that has been deemed to have no value to the analysis objectives.

Performing Analytics Over The Data: Depending on the type of analytic result required, this stage can be as simple as querying a dataset to compute an aggregation for comparison. On the other hand, it can be as challenging as combining data mining and complex statistical analysis techniques to discover patterns and anomalies or to generate a statistical or mathematical model to depict relationships between variables.

Visualizing Data: The Data Visualization stage , is dedicated to using data visualization techniques and tools to graphically communicate the analysis results for effective interpretation by business users

Customer Behavior:

Customer behavior is the study of individuals and organizations and how they select and use products and services. It is mainly concerned with psychology, motivations, and behavior.

The study of customer behavior includes:

- How consumers think and feel about different alternatives (brands, products, services, and retailers)
- How consumers reason and select between different alternatives
- The behavior of consumers while researching and shopping
- How consumer behavior is influenced by their environment (peers, culture, media)
- How marketing campaigns can be adapted and improved to more effectively influence the consumer

These considerations are influenced by three factors:

Personal factors – A person's interests and opinions. These will be affected by demographics such as age, gender, culture, profession, background and so on.

Psychological factors – Everybody's response to a particular marketing campaign will be based on their perceptions and attitudes. A person's ability to comprehend information, their perception of their need, their attitude, will all play a part.

Social factors – Peer groups, from family and friends to social media influence. This factor also includes social class, income, and education level.

Collecting consumer behavior data- As the motivations that influence consumer behavior are so wide, a research mix including a variety of data will be the most robust. Some are more cost effective than others.

- Customer reviews
- Q&A sites
- Surveys

- Focus groups
- Google analytics
- Google trends
- Government data
- Social media
- Blog comments

---

3. History

Traditional Analytical Systems For Customer behavior:

In the late 1970s, there were two approaches for constructing Database Management Systems (DBMS's). The first approach was based on the hierarchical data model, typified from IBM, in response to the enormous information storage requirements generated by the Apollo space program. The second approach was based on the network data model, which attempted to create a database standard and resolve some of the difficulties of the hierarchical model, such as its inability to represent complex relationships DBMSs. However, these two models had some fundamental disadvantages like the complex programs had to be written to answer even simple queries. Also there was minimal data independence .

Many experimental relational DBMS were implemented thereafter, with the first commercial products appearing in the 1970's and early 1980's. Relational DBMS used extensively in the 80's and 90's was limited in meeting the more complex entity and data needs of companies, as their operations and applications became increasingly complex. In response to the increasing complexity of database applications, two new data models had emerged, the Object-Relational Database Management Systems (ORDBMS) and Object-Oriented Database Management Systems (OODBMS), which subscribes to the relational and object data models respectively. The OODBMS and ORDBMS have been combined to represent the third generation of Database Management Systems.

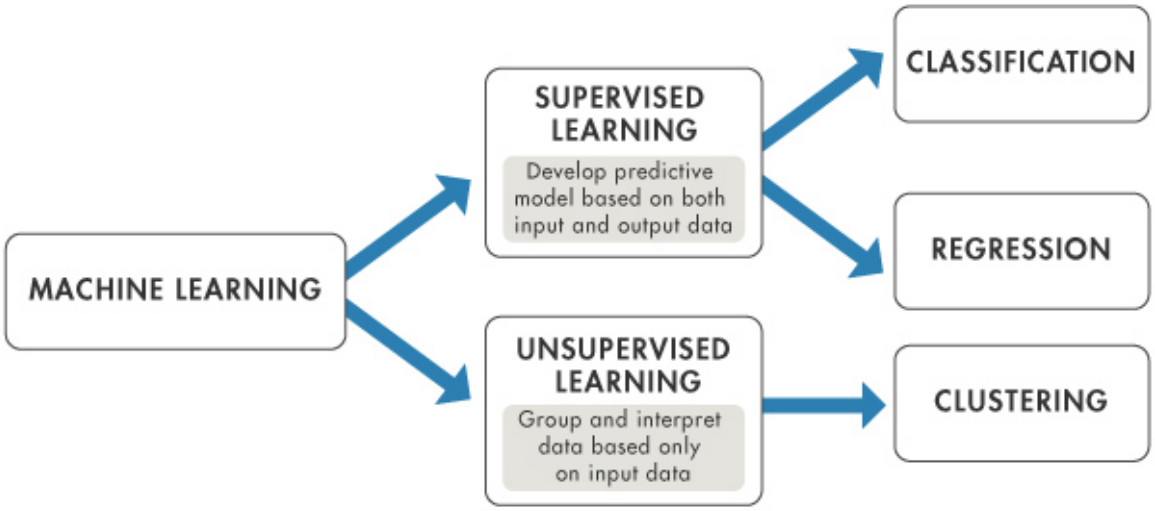
Big Data is emerging from the realms of science projects at companies to help telecommunication giants understand exactly which customers are happy with their service and what processes caused the dissatisfaction, and predict which customers are going to change the service. To obtain this information, billions of loosely-structured bytes of data in different locations need to be processed until the required data is found out. This type of analysis enables executive management to fix faulty processes or people and may be able to reach out to retain at-risk customers . Big data is becoming one of the most important technology trends that have

the potential for dramatically changing the way organizations use customer behaviour to analyze and transform it into valuable insights.

---

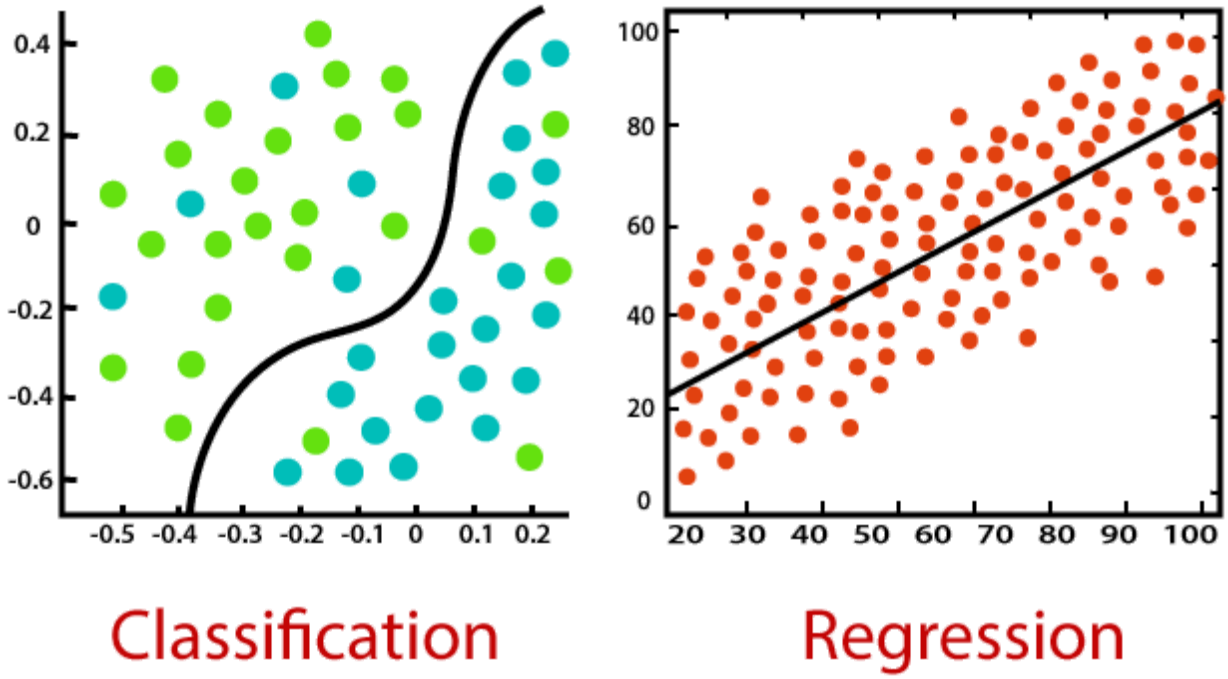
4. Customer behavior analysis using Big data

To sync customer behavior with data analytics there should be a medium connecting,here comes machine learning techniques.There are various algorithms to perform the given task, but the common of all is decision trees.



- **Classification** is the process of finding or discovering a model or function which helps in separating the data into multiple categorical classes i.e. discrete values. In classification, data is categorized under different labels according to some parameters given in input and then the labels are predicted for the data.

- **Regression** is the process of finding a model or function for distinguishing the data into continuous real values instead of using classes or discrete values. It can also identify the distribution movement depending on the historical data.



Types of classification models:

1. Artificial neural networks
2. Support vector machine
3. Decision trees
4. Bayes naive
5. K-nearest

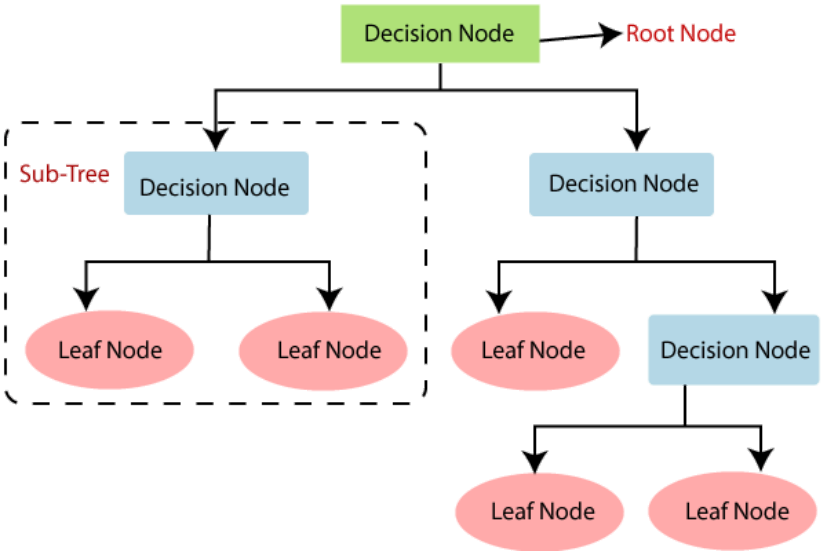
We cannot decide which algorithm is best for what problem,it totally depends on the types of datasets.so you just need to hit and try every algorithm to find which works perfectly to the specific problem.

Decision Trees

It is a type of classification algorithm which comes under the supervised learning technique.
It is a graphical representation of all possible solutions of taking a decision.
It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome
In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**

- **Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.**



- **The logic behind the decision tree can be easily understood because it shows a tree-like structure.**

How does the Decision Tree algorithm Work?

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of the root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

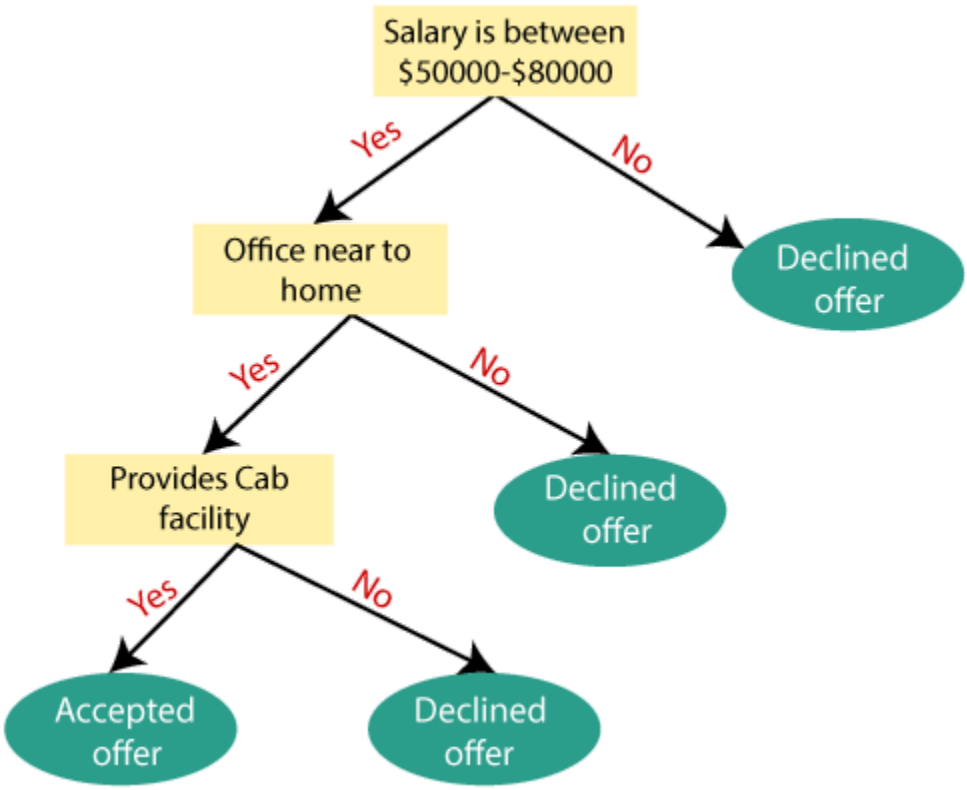
For the next node, the algorithm again compares the attribute value with the other sub-nodes and moves further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

- **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
- **Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**
- **Step-3:** Divide the S into subsets that contain possible values for the best attributes.
- **Step-4:** Generate the decision tree node, which contains the best attribute.

- **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Example:

Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram:



5. Tools

Below are the tools used to analyze customer behavioral data using Big Data analytics

- Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides massive storage for any kind of data, enormous processing power and the ability to handle virtually limitless concurrent tasks or jobs.

- Apache Pig is a high-level platform for creating programs that run on Apache Hadoop. The language for this platform is called Pig Latin.

- HBase is an open-source non-relational distributed database modeled after Google's Bigtable and written in Java.

- Apache Spark is an open-source distributed general-purpose cluster-computing framework. Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.

- Apache Hive is a data warehouse software project built on top of Apache Hadoop for providing data query and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop.

---

6. Domains using Customer analytics

- E-commerce
- Aviation
- Transportations and logistics
- Banking and securities
- Communications
- Media and entertainments
- Healthcare
- Education
- Manufacturing
- Natural resources
- Government
- Insurance

- Retail and wholesale trade

---

7. Advantages & Disadvantages

Advantages

- Creation of personalized products and services
- Informing of the marketing strategy by analyzing search engines' trends
- Determination of effective content
- Clear definition of ideal customer profile (ICP)
- Advancement of lead scoring methods
- Cuts down the costs
- Increases efficiency
- Improves pricing
- Increases sales

Disadvantages

- Limited buyer interest
- Social and cultural influences
- Incomplete data
- Privacy
- Inconsistency

---

8. Conclusion

Eighty percent of the world‟s data is unstructured, and most businesses do not even attempt to use this data to their advantage. The trend of Big Data is generating new opportunities and new challenges for businesses across industries.

The churches have a lot of unstructured data. The data should be stored and analyzed by Big Data techniques and methods. This approach is helpful for better analysis.

Hadoop is a scalable platform for ingesting Big Data and preparing it for analysis. Using Hadoop in Big Data can reduce time to analysis by hours or even days.

Let's make one thing clear: big data is not a sure way to eliminate bad marketing decisions and increase revenues overnight. It is, however, excellent for decreasing the probability of human error in decision making, validating suggestions, and proposing trends.Another thing: you need to give those large volumes of data a human face by making the connection between numbers and customer experience.Once you've learned to leverage the power of using big data, you will create highly-targeted campaigns that will increase your profits and eliminate inefficiencies.

---

9. References

1. *Big Data Now*, O‟Reilly Media, Sebastopol, 2012, 3
2. Tom white, ―Hadoop - The Definitive Guide‖,3rd.Edition, O'Reilly Media, Inc.,Sebastopol, CA 95472,2012.
3. R. Halenar, Appl. Mech. Mater., **229-231** (2012) 2125-2129.
4. J.R. Quinlan,―C4.5: programs for machine learning‖, Morgan Kaufmann,1993
5. K. Davis and D. Patterson, *Ethics of Big Data*, O‟Reilly Media, city California, 2012, 1.
6. google.