**MINISTRY OF SCIENCE AND EDUCATION
REPUBLIC OF AZERBAIJAN**

KHAZAR UNIVERSITY

SCHOOL OF SCIENCE AND ENGINEERING

Department: Engineering and applied sciences

Major: 060509 - Computer Science

Major: Informatics

MASTER THESIS

Title: **Customer Behavior Analysis Using Big Data Analytics and Machine Learning**

Student: **Zaman Karimov**

Supervisor: **PhD, Associate Professor Leyla Muradkhanli**

Baku – 2023

ABSTRACT

This thesis explores the use of big data analytics and Machine Learning (ML) for customer behavior analysis in the context of digital marketing. The primary objective is to uncover patterns and trends in customer behavior and leverage this information to drive data-driven decisions related to marketing strategy, product development, and customer service.

The thesis commences with an in-depth overview of the fundamental concepts of big data and ML, elucidating their applicability within the realm of digital marketing. This includes a comprehensive discussion of various types of ML algorithms and the ML pipeline employed to construct and deploy predictive models for customer behavior analysis.

Subsequently, the thesis delves into specific applications of ML in customer behavior analysis. It investigates the utilization of ML techniques to predict customer churn, identify high-potential prospects, and determine optimal communication channels for distinct customer segments. Furthermore, the thesis explores the integration of sentiment analysis in marketing, showcasing how ML can effectively assess customer sentiment and enhance the overall customer experience.

Throughout the thesis, real-world examples and compelling case studies are presented to exemplify the efficacy of ML in customer behavior analysis. These instances provide tangible evidence of how ML techniques have yielded actionable insights and facilitated decision-making processes within marketing contexts.

Concluding the thesis, an examination of the limitations and challenges associated with utilizing ML for customer behavior analysis is presented. Additionally, the thesis outlines prospective avenues for future research in this domain. By encompassing the various facets of ML in customer behavior analysis, this thesis aspires to serve as a comprehensive guide for professionals seeking to harness the potential of big data analytics and ML in their organizations.

By the conclusion of this thesis, readers will have acquired a profound understanding of the advantages offered by these technologies. Furthermore, they will be equipped with practical insights necessary for implementing and integrating big data analytics and ML methodologies effectively within their business frameworks.

List of Figures

List of Tables

List of acronyms and abbreviations

ML	Machine Learning
CE	Customer Engagement
CRM	Customer Relationship Management
IT	Information Technology
AI	Artificial Intelligence
NB	The Naive Bayes
BHFS	Bagging Homogeneous Feature Selection
RBF	Radial Basis Function
ANN	Artificial Neural Network
CLV	Customer Lifetime Value
CCB	Customer Citizenship Behavior
PLS	Partial Least Squares
BDA	Business Development Associate
EBL	Explanation-based Learning
API	Application Programming Interface
NLP	Natural Language Processing
SVM	Support Vector Machines
PCA	Principal Component Analysis
T-SNE	T-Distributed Stochastic Neighbor Embedding
SEO	Search Engine Optimization
SMM	Social Media Marketing
PPC	Pay-Per-Click
SEM	Search engine marketing
SERP	Search Engine Results Pages
ROI	Return on Investment
GG	Gamma-Gamma
BG-NBD	Beta Geometric Negative Binomial Distribution

Table of Contents

INTRODUCTION

In today's competitive marketplace, businesses face a growing need to understand and engage with their customers effectively. With the increasing amount of data being generated by customers through online and offline interactions, businesses have access to an unprecedented level of information about their customers. However, this data can be overwhelming and difficult to analyze, making it challenging for businesses to extract meaningful insights. Every person is unique and has various habits and personality features. Customer behavior is often consistent. We base our buying or not buying decisions on our way of life, past experiences, and emotions.

No matter if it is a tiny neighborhood bakery or a massive global network of supermarkets, it is wise to understand who the clients are.

ML is useful in this situation. Traditional marketing strategies themselves became ineffective because of the expansion of digital platforms and the digitization of business. This does not imply that ML rewrites the principles of marketing and customer behavior research, but it does provide new tools and insights.

The use of ML and big data analytics can help businesses overcome these challenges and gain a deeper understanding of customer behavior. ML algorithms can analyze large and complex datasets to identify patterns and relationships, allowing businesses to make predictions about future customer behavior. Additionally, big data analytics can provide businesses with real-time insights into customer preferences and needs, enabling them to make data-driven decisions that improve customer engagement (CE) and satisfaction.

One of the key benefits of using ML and big data analytics to analyze customer behavior is the ability to personalize customer experiences. ML algorithms can analyze customer data to understand individual preferences and behaviors, allowing businesses to tailor their products, services, and communications to meet the unique needs and preferences of each customer. This leads to higher levels of CE, loyalty, and satisfaction.

Another important application of ML and big data analytics in analyzing customer behavior is in customer relationship management (CRM). ML algorithms can analyze customer interactions and behavior data to identify patterns and trends, enabling businesses to identify areas for improvement in their CE and support processes. Additionally, big data analytics can help businesses understand the impact of their CRM strategies, enabling them to make data-driven decisions to optimize their customer relationships.

In latest years, businesses all over the world have aggressively begun using new ML techniques to increase their competitiveness in the client acquisition market. Because of the expanding amount of data and widespread access to high-performance computing and cloud services, ML has enabled businesses to greatly improve the customer experience.

To benefit from ML models, the early investors needed to invest heavily in pricey information technology (IT) infrastructure, human resources, and significant budgets. However, the advantages of modern digital trends for small enterprises became a reality with the introduction of cloud technology and subscription-based services.

Research Questions

This thesis' main goal is to introduce the reader to current ML and digital marketing trends and to certain marketing challenges that can be handled by ML algorithms. Additionally, an ML algorithm is developed in the implementation phase to demonstrate a real-world case situation. The results of this thesis will provide an IT specialist with organized information on the use of ML for marketing. A marketing specialist will learn the most recent information on digital trends, ML, and algorithms. Any small firm or marketing department can utilize the project presented in this thesis as a starting point for implementing applied ML.

RQ1: What is ML, and where does it fit in the IT and artificial intelligence (AI) scenery?

RQ2: What fundamental ideas underlie both marketing and the study of consumer behavior?

RQ3: What challenges do marketers face in their area of work?

RQ4: How can ML be used to solve those issues?

RQ5: What ML methods are used in the research of consumer behavior?

1.LITERATURE REVIEW

In recent years, there has been a growing interest in the use of big data analytics and ML in customer behavior analysis. In this literature review, we will examine some of the key research studies in this field.

One study by Kshetri and Voas (Kshetri, N., & Voas, J., 2018: 11-14) examined the use of big data analytics in CRM. The study found that big data analytics can help businesses to identify customer needs, preferences, and patterns, leading to more effective CRM strategies.

D. Prabha et al. (Prabha, D. 2021: 5105–5116) proposed the idea of understanding consumer behavior is crucial for the success of an organization, and ML techniques can be used for CRM and predicting customer behavior. The Naive Bayes (NB) model is a popular choice for predicting customer behavior, but it can perform poorly if there are irrelevant or noisy attributes in the dataset. The author proposes a new method called Bagging Homogeneous Feature Selection (BHFS) which uses ensemble data perturbation feature selection methods to select a relevant feature subset and improve the performance of the NB model. The BHFS-NB model is shown to outperform the standard NB model in terms of accuracy and running time complexity.

Chien-Chang Hsu et al. (Hsu, C.-C., & Deng, C.-W. 2004: 315-324) proposed an intelligent interface for customer behavior analysis in electronic commerce that contains three modules - the task editor, action supervisor, and behavior analyzer - and explain how each module works to provide an efficient and effective customer behavior analysis system. The task editor allows the system administrator to define business tasks and domain ontology. The action supervisor is responsible for monitoring customer operations, excluding unnecessary operations, and recognizing behavior patterns using interaction messages, Bayesian belief network, and Radial Basis Function (RBF) neural networks. The behavior analyzer generates customer behavior analysis information by measuring behavior patterns, constructing personalized domain ontology, and evaluating skill proficiency of the customer.

Stavros Anastasios Iakovou's et al. (Iakovou, S.A., et al., 2016) work presents a prediction model based on customer behavior using data mining techniques. The model utilizes data from a supermarket database and an additional database from Amazon to classify customers and products. The model is trained and validated with real data and is intended to be used as a tool for marketers to analyze and target consumer behavior.

Vahid Golderzahi et al. (Golderzahi, V., & Pao, H.-K., 2018: 56-71) proposed an idea that WIFI-based sensing can be used to analyze customer behavior, specifically group behavior, to improve store revenue forecasting. The proposed method involves collecting and analyzing WIFI signals from a coffee shop to predict revenue, number of WIFI-using devices, and number of sold products. The method achieves good performance in prediction accuracy and is further improved when group information or weather information is included.

Hamid Ahaggach (Ahaggach, H., 2023: 357-363). The main idea of author's paragraph is that the objective of the thesis is to conduct research on the use of data science and AI techniques to assist automotive dealership companies in decision-making processes and to use data-driven methods for computing these enterprises. The thesis proposes developing algorithms to extract relevant information from a diverse and multi-structured automotive environment, assess situations, suggest recommendations, develop marketing strategies, and automate manual tasks.

Meshal Alduraywish et al. (Alduraywish, M., Unhelkar, B., Singh, S., & Prasad, M., 2022: (pp. 558–567) discussed the competition in the fast fashion industry, specifically the emergence of online-only retailing and the use of AI and ML by these companies to enhance the online customer shopping experience. It also highlights the concern of multichannel companies regarding the performance of their physical stores and the potential for these stores to enhance customer experience. The paper aims to explore the application of AI in ecommerce by fast fashion companies and the impact of online-only business on physical stores, as well as the future role of physical stores in the industry.

Kailash Hambarde et al. (Hambarde, K., Silahtaroğlu, G., Khamitkar, S., Bhalchandra, P., Shaikh, H., Tamsekar, P., & Kulkarni, G., 2020: 45–50) conducted investigations on implementing data analytics on a Turkey-based e-commerce company's data to classify customer behavior patterns. They used an Artificial Neural Network (ANN) model to forecast customer purchasing patterns, which can benefit the marketing department in recognizing targeted customers for specific campaigns. The ANN model using the back-propagation technique showed high accuracy in predicting customer behavior. The study was conducted in R programming environment.

Kazuaki Tsuboi's et al. (Tsuboi, K., Shinoda, K., Suwa, H., & Kurihara, S., 2015: 353–361) paragraph major point is that it's crucial to comprehend consumer wants in order to promote the correct products and services, and that video image analysis has developed into a crucial tool for studying consumer behavior. The paragraph outlines how time-series data mining techniques are required since video data is sequential and how video data may be utilized to identify buying trends. A sequential pattern mining method based on collective intelligence is being developed to address this issue and has been successful in visualizing the relationship of goods that consumers continuously touch, according to the paragraph, which also notes that consumer behavior is frequently ambiguous.

Pin-Liang Chen et al. (Chen, PL. et al., 2018) proposed a study that aims to analyze the traits and possibility of making purchases of various customer groups in the Shimen shopping district in order to derive business value. The main idea of this paragraph is that the rising popularity of social networking services and mobile devices has brought new business challenges. Users' Facebook profiles and use data from a point-earning app were utilized in the study to conduct statistical analyses with a CBAS. The findings indicated that younger consumers made up the majority of the retail district's patrons and that different age and gender groupings had varying preferences. The study also showed that altering the assignments and promotions on the point-earning app based on the investigation's findings significantly increased the conversion rate.

Takao Terano et al. (Terano, T. et al., 2009: 244–251) proposed a study that introduces ABISS, an agent-based simulator that can be used to analyze customer wandering patterns and shopping patterns in a supermarket. The simulator enables "virtual experiments" to be carried out by altering different shop operations and retail business characteristics. The simulation model's construction is discussed in the article, which also demonstrates how the layout of the shop, and the placement of in-store advertising and recommendation systems affect consumer flow and sales.

Shalini et al. (Shalini, & Singh, D., 2018: 753–763) had an idea that businesses need to understand their consumers' preferences in order to increase customer satisfaction and retention since consumer behavior is continuously evolving. Customer clustering is a practical method that enables companies to see trends in consumer behavior and adjust their marketing plans appropriately. The goal of the study is to discover the strategy that is most effective for detecting consumer profiles and trends in a retail business. It does this by comparing the effectiveness of several data mining strategies for customer clustering.

Simon Denman (Denman, S., et al., 2012: 199–238) proposed an understanding how clients migrate between different geographic areas, how long they stay in each area, and where they are most likely to go is useful in business settings. These measurements are currently determined manually or by using hardware tags. In order to establish operational data on how individuals move around an area, the author suggests employing soft biometrics such as height, weight, gender, hair, skin, and clothing color. Those that stand out from the crowd are found using an unique average soft biometric, and as they are located throughout a dispersed camera network, operational statistics are gradually gathered about them.

José-Ramón Segarra-Moliner et al. (Segarra-Moliner, J.-R., & Moliner-Tena, M.-Á., 2022) This paragraph's main purpose is to present the purpose and methodology of a research that aims to examine the connection between customer lifetime value (CLV) and customer citizenship behavior (CCB) within the context of CE. The paper also explores how motivating factors and engagement may be used to forecast business performance using marketing analytics. In order to test hypotheses and assess the model's predictive ability, the study uses a data sample of telecom service users and applies prediction-oriented segmentation and a second-order partial least squares (PLS) model. According to the findings, deliberate loyalty plays a mediating role in attaining future financial business success, and CCBs, which are voluntary, discretionary, and extra-role customer behaviors, are predecessors of brand attitude-attachment, social value, and kindness CLV. The study intends to examine, from both theoretical and empirical angles, the effects of CE development through customer civic activity on CLV.

Dongyun Nie et al. (Nie, D., et al, 2022) suggested that CLV is a crucial metric for forecasting net profit from ongoing customer connections, but that prior research either concentrates on a conceptual model or presupposes that all necessary information are easily accessible. By creating a rigorous framework for CLV calculation using an actual insurance policy dataset and carrying out a thorough validation procedure to identify the best-performing models, the author of this research study seeks to close this gap. The framework includes all steps involved in CLV estimate, such as establishing a single customer record, categorizing customers into ranked groups, substituting for missing factors, and computing and confirming each customer's unique CLV values.

Bireshwar Ganguly et al. (Ganguly, B., & Ambhaikar, A., 2022: 393–403) proposed that organizations utilize data analytics and other resources to gain a better understanding of their customers, with the aim of minimizing marketing costs and targeting the appropriate audience. With online transactions in areas such as banking, shopping, and e-commerce now at a higher level, it is crucial for businesses to comprehend customer preferences and inclinations in order to offer a personalized shopping experience. The goal of the paper is to suggest a resolution that involves providing a personalized graphical interface on online portals that is tailored to the user's profile, thus improving the browsing experience.

Saurav Kumar et al. (Kumar, S., et al., 2023: 649–664) discussed how process digitization, particularly hyper-personalization, is an important aspect of the fourth industrial revolution. The use of data mining techniques, specifically a hierarchical recurrent neural network algorithm, is employed to examine the effect of hyper-personalization on customer behavior. Multiple methods and systems are utilized to gather data on customer behavior, which is then utilized to produce personalized customer offers based on their unique preferences and requirements.

B. V. R. Sai Teja et al. (Sai Teja, B. V. R., & Arivazhagan, N., 2021: 357–369) emphasized the growing significance of data mining in industries where data is constantly expanding, which can be utilized for various purposes like marketing analysis, fraud detection, and customer retention. The rising competition is impacting offline markets, but a proposed system aims to assist offline stores in retaining their customers by forecasting their purchasing patterns and managing their inventory. The paper further explores market basket analysis and text segmentation of customer data.

Manjula Ramannavar et al. (Ramannavar, M., & Sidnal, N. S., 2016: 291–306) discussed big data, which refers to data sets that are too large and complex for traditional methods and technologies to handle. To analyze big data in real-time and gain insights to support strategic decision-making, advanced analytics such as predictive and prescriptive analytics can be used. The author aims to explore new areas of big data and analytics, including a classification system for big data analytics, a comparison of different Business Development Associate (BDA) platforms, and a suggestion for creating a contextual model for BDA using advanced analytics.

D. Kalaivani et al. (Kalaivani, D., & Sumathi, P., 2019: 519–524) wrote about how businesses are adopting information technology and data-driven techniques like business intelligence and factor-based principle component analysis to gain insights into their customers and make informed decisions. The ultimate objective is to enhance their market position by scrutinizing enormous volumes of data, including sales demographics, economic trends, competitive data, consumer behavior, efficiency measures, and financial calculations. The key emphasis is on comprehending customer expectations and boosting sales, especially in online commerce.

2. MACHINE LEARNING AND ARTIFICIAL INTELEGENCE

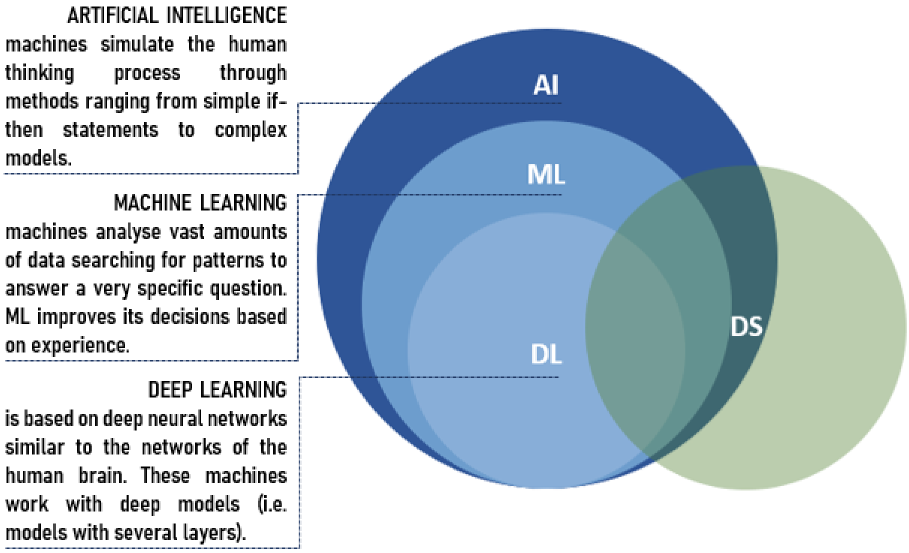
2.1 Definition of ML and AI

Machines with the ability to learn, think, and function independently are referred to as AI in its widest definition. They possess the same capacity for independent decision-making as both humans and animals when confronted with novel circumstances. The great majority of contemporary AI developments and applications pertain to a class of algorithms known as ML. In order to discover patterns in enormous volumes of data, these algorithms employ statistics.

Day after day, millions of Netflix users employ ML algorithms without ever realizing it. A recommender system is used by Netflix to propose movies and TV series to users. To produce recommendations, the system considers previous ratings and preferences. ML techniques are used by recommender systems to improve their predictions of an user 's needs. A recommender system can employ a variety of various methods of ML. The ideal algorithm for a given application will rely on the characteristics of the data because every algorithm has strengths and limitations of its own. The algorithm for linear regression is the most popular.

To forecast the result of future occurrences, a technique called linear regression establishes a linear connection between an independent variable and a dependent variable. The best linear approximation to a data set is found using the linear regression procedure. This algorithm is used in recommender systems to anticipate user ratings based on previous ratings.

Any recommendation engine, whether it be on a website like Netflix, a voice assistant like Siri, Alexa, Cortana, or a search engine like Google or Yandex, is mostly powered by ML algorithms (Businessoverbroadway.com, 2021). The Figure 2.1 describes the structure of AI:



ARTIFICIAL INTELLIGENCE
machines simulate the human thinking process through methods ranging from simple if-then statements to complex models.

MACHINE LEARNING
machines analyse vast amounts of data searching for patterns to answer a very specific question. ML improves its decisions based on experience.

DEEP LEARNING
is based on deep neural networks similar to the networks of the human brain. These machines work with deep models (i.e. models with several layers).

Figure 2.1 Structure of AI. Source: abrisconsult.com

The image makes it obvious that ML is a subset of AI. To make it clear to the reader, here is a simple Google ML dictionary definition of ML: “A program or system that builds (trains) a predictive model from input data. The system uses the learned model to make useful predictions from new (never-before-seen) data drawn from the same distribution as the one used to train the model. ML also refers to the field of study concerned with these programs or systems.” (. Google Developers. 2021).

While this was happening, ML developed along with the increased processing capacity of contemporary computers, making possible what was previously thought to be impossible.

2.2 Important milestones in AI and ML

Many people believe that ML is a recent technology, however the idea has really been for a long time. The development of the current ML sector was influenced by numerous important turning points. Let’s revise some of the most important milestone in the history of ML and AI development:

Before the 1950s - In data science, statistical approaches were the foundation of all work. It served as the exclusive method of data analysis for decades. The Dark Age was the name given to this historical period.

1950s - The "Turing Test" was developed by Alan Turing in 1950 to test if a computer had true intelligence. A computer must be able to deceive a person into thinking it is also human in order to pass the test (A. M. TURING, I, 1950: 433-460). During that same decade, the first neurocomputer was created with the ability to identify visual patterns, the development of the algorithms to play checkers with a computer was initiated, and the perceptron was also introduced (Rosenblatt, F., 1958: 386–408).

1960s - When the "nearest neighbor" technique was developed, computers could start making extremely simple pattern recognition decisions. This might be used to plan a route for salespeople who are on the road, starting in a random location and making sure they stop in each city during a brief tour(Marr, Bernard, 2016).

1970s - The "Stanford Cart," created by students at Stanford University, can autonomously negotiate obstacles in a space (Marr, Bernard, 2016). In addition, this decade also called “AI winter” is a time when interest in AI and ML technologies is at its lowest point.

1980s - Explanation-based learning (EBL) is a notion that Gerald Dejong presents. In EBL, a computer analyzes training data and develops a general rule it may adhere to by removing irrelevant data (Marr, Bernard, 2016).

1990s - The approach to ML is changing from one that is knowledge-driven to one that is data-driven. Scientists start developing computer systems that can process massive volumes of data and draw inferences, or "learn" from the outcomes. Also, in 1997 IBM’s Deep Blue beats the world champion Gary Kasparov at chess (Marr, Bernard, 2016).

2000s - Unsupervised ML techniques, such as Support-Vector Clustering and other kernel methods are becoming more popular (Hofmann, T., et al., 2008: 1171–1220). In 2009 was ImageNet was created (. Gershgorn, Dave, 2017). Fei-Fei Li from Stanford University conceptualized ImageNet, a sizable visual database, after realizing that even the strongest ML algorithms wouldn't perform effectively if the data didn't accurately represent the actual world. For many, ImageNet served as the impetus for the 21st-century AI boom.

2010s - As deep learning becomes practical, ML starts to play a bigger role in many commonly used software services and apps. In 2014 Google researchers describe their work on Sibyl (Canini K. et al., 2016), a private platform for massively parallel ML that Google uses internally to forecast user behavior and offer suggestions.

Even if the business has come a long way from its precursor, it is crucial to keep in mind where it all began and how it has developed since there is still a long way to go. Understanding and creating new methods in ML and AI are made easier by studying the early principles.

2.3 Data in ML

Any ML algorithm runs on data as its fuel. Data provides knowledge about customers and their behavior. We make use of it to "load" the ML algorithm and to gather important data. Start by giving data a precise definition.

The Cambridge Dictionary defines data as information, particularly facts or statistics, gathered for analysis, consideration, and use in support of decision-making, or information in an electronic format that can be utilized by a computer (Dictionary.cambridge.org, 2021). The use of data is essential in the field of ML. It speaks about the collection of data that may be analyzed or measured to train a ML model. The quantity and quality of data used for training and testing have a big impact on how well a ML model performs. Data can originate from a variety of sources, including databases, spreadsheets, and Application Programming Interface (API), and can take many different formats, including numerical, category, or time-series data. ML algorithms employ data to discover patterns and connections between input parameters and desired results, which may subsequently be used to tasks requiring categorization or prediction.

The idea of data has been around for a long and is well-known. The concept of Big Data is what is novel about it.

Big Data is a term used to describe large and complex data sets that cannot be processed or analyzed using traditional data processing tools. It refers to the vast amount of structured, semi-structured, and unstructured data that is generated by various sources, including social media, sensors, transactions, and other digital activities. The term is also used to describe the technologies and techniques used to store, process, and analyze these massive data sets to extract valuable insights and knowledge.

Data scientists define four Vs of Big Data:

- Volume: It refers to the vast amount of data generated from various sources such as social media, IoT devices, sensors, and transactions. Big Data is characterized by its large volume, which requires scalable storage solutions to manage and process it effectively.
- Velocity: It refers to the speed at which data is generated and needs to be processed. Big Data is generated in real-time or near real-time, and its velocity requires real-time or near real-time processing and analysis.
- Variety: It refers to the different types and formats of data generated from various sources, such as structured, semi-structured, and unstructured data. Big Data includes data from

different sources in different formats and requires diverse data processing and analysis techniques.

- Veracity: It refers to the accuracy and reliability of the data generated. Big Data is often incomplete, inaccurate, or inconsistent, and its veracity requires advanced data cleaning, filtering, and quality assurance techniques to ensure the data's accuracy and reliability.

There are various methods to organize data, but data science emphasizes two key categories: structured and unstructured data.

Structured data refers to the data that is organized in a fixed format with a predefined schema. Structured data can be easily managed, processed, and analyzed using traditional data processing tools and technologies. Examples of structured data in Big Data include transactional data, customer data, financial records, and other types of data that are typically stored in relational databases.

However, the volume of structured data generated in Big Data is often too large to be handled by traditional data processing tools. Therefore, specialized tools and technologies such as Hadoop, Spark, and NoSQL databases are used to manage and process structured data in Big Data. These technologies enable distributed processing of structured data across multiple nodes in a cluster, which can help to improve performance and scalability. Additionally, advanced analytics tools such as ML and data mining are used to extract insights and knowledge from structured data in Big Data. Overall, structured data plays a critical role in Big Data and provides valuable insights and knowledge to organizations that can help them make data-driven decisions.

Unstructured data refers to the data that does not have a predefined structure or format. It includes data such as text, images, videos, audio, social media posts, and other digital content. Unstructured data is generated in large volumes and is challenging to store, manage, process, and analyze using traditional data processing tools.

The rise of Big Data has led to an exponential increase in unstructured data, which presents new opportunities and challenges for businesses to extract insights and knowledge. Unstructured data can provide valuable information about customer behavior, sentiment analysis, and other aspects of business operations. However, analyzing unstructured data requires advanced techniques such as natural language processing (NLP), ML, and deep learning.

2.4 Algorithms in ML

For ML algorithms and client behavior analysis, data is a crucial component. The definition of an algorithm and its function in an ML pipeline are covered in this section. An algorithm in ML refers to a set of rules and instructions that are used to train a model to recognize patterns in data. It is a mathematical formula or a sequence of steps that enable a ML model to learn from data and make predictions or decisions based on that learning. In ML, algorithms are used to transform input data into a useful output by discovering relationships, patterns, and trends in the data.

There are four main types of algorithms in ML that can be divided by their purposes:

1. Supervised Learning
2. Unsupervised Learning
3. Semi-supervised Learning
4. Reinforcement Learning

2.4.1 Supervised Learning

Supervised learning is a type of ML in which a model is trained on a labeled dataset consisting of input features and their corresponding output or target variable. In supervised learning, the algorithm learns a mapping function from the input features to the output variable based on the labeled data provided during the training phase.

The training data used in supervised learning consists of pairs of inputs and their corresponding correct outputs. The objective is to learn a function that can accurately predict the output for new, unseen inputs. The goal is to minimize the difference between the predicted output and the actual output.

The process of supervised learning typically involves the following steps:

- Data collection and preprocessing: The first step is to collect and preprocess the training data, which involves cleaning, transforming, and preparing the data for use in the model.
- Feature extraction and selection: The next step is to extract relevant features from the input data and select the most important ones for use in the model.
- Model selection and training: The next step is to select an appropriate model architecture and train it using the labeled data. The model is trained by adjusting the weights and biases to minimize the difference between the predicted output and the actual output.
- Model evaluation: Once the model is trained, it is evaluated on a separate validation dataset to determine its accuracy and performance. The model is then adjusted and retrained as necessary.

Figure 2.2 shows detailed explanation of Supervised Learning model workflow.

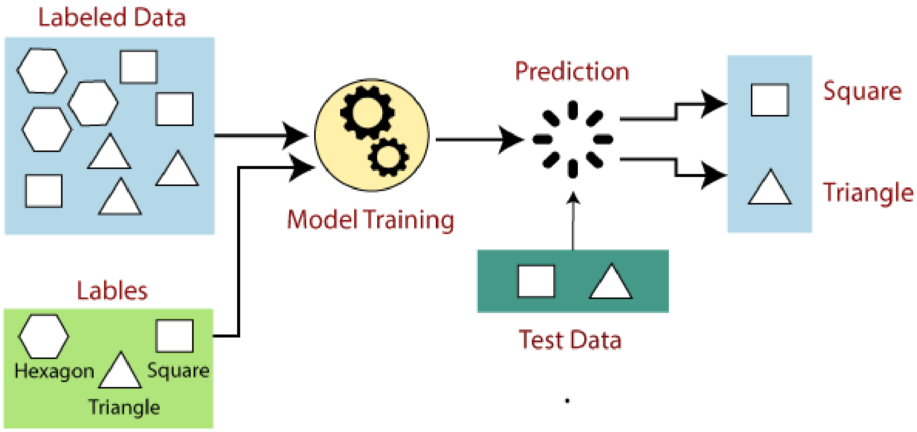


Figure 2.2 Workflow of Supervised Learning model. Source: javatpoint.com

Supervised learning algorithms are used in various applications such as classification, regression, and object detection. Examples of supervised learning algorithms include linear regression, logistic regression, decision trees, random forests, Support Vector Machines (SVM) and neural networks. The choice of algorithm depends on the type of data and the learning objective.

2.4.2 Unsupervised Learning

Unsupervised learning is a type of ML algorithm that is trained on unlabeled data, which means that there is no predefined output or target variable provided to the model during the training phase. The goal of unsupervised learning is to find patterns and relationships in the data without any guidance.

In unsupervised learning, the model must identify the underlying structure or distribution of the data on its own. The model is then used to make predictions or cluster the data based on similarities or patterns.

Unsupervised learning algorithms can be used for various applications, including:

- Clustering: This involves grouping similar data points together based on their features. For example, a clustering algorithm can be used to group customers with similar buying behavior together.
- Anomaly detection: This involves identifying data points that deviate significantly from the norm. For example, an unsupervised learning algorithm can be used to detect fraudulent transactions.
- Dimensionality reduction: This involves reducing the number of features or variables in the data while preserving the most important information. For example, an unsupervised

learning algorithm can be used to reduce the dimensionality of an image while preserving the most important features.

Figure 2.3 describes problem example of Unsupervised Learning model.

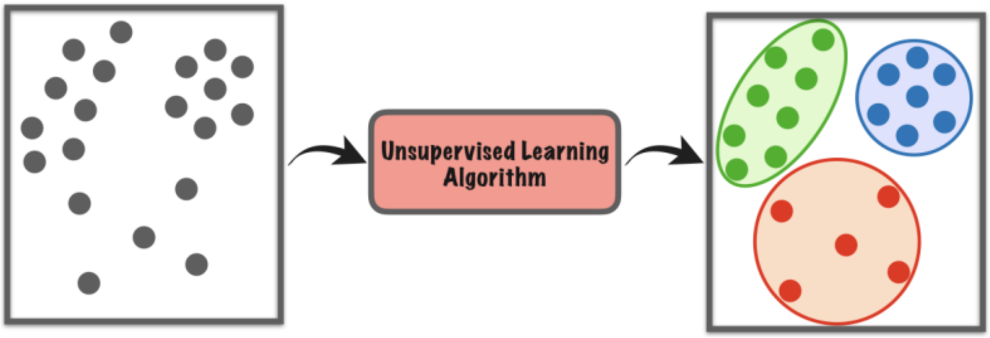


Figure 2.3 Unsupervised Learning problem. Source: towardsdatascience.com

Unsupervised learning algorithms include k-means clustering, hierarchical clustering, principal component analysis (PCA), and autoencoders. The choice of algorithm depends on the type of data and the learning objective.

2.4.3 Semi-supervised Learning

Semi-supervised learning is a type of ML algorithm that uses both labeled and unlabeled data to train a model. Unlike supervised learning, where the model is trained on only labeled data, semi-supervised learning uses the additional unlabeled data to improve the accuracy and generalization of the model.

The basic idea behind semi-supervised learning is that the unlabeled data can provide additional information about the data distribution and the relationships between the input features and the output labels. The model can then use this information to better generalize to new, unseen data.

Semi-supervised learning can be useful in situations where labeled data is expensive or time-consuming to obtain, but there is a large amount of unlabeled data available. For example, in a medical diagnosis task, it may be difficult to obtain labeled data for rare diseases, but there may be a large amount of unlabeled patient data available.

Some popular semi-supervised learning algorithms include self-training, co-training, and generative models such as autoencoders and variational autoencoders. These algorithms use the labeled data to initially train the model and then use the unlabeled data to fine-tune the model.

Overall, semi-supervised learning is a promising approach to ML that can help improve the accuracy and generalization of models while reducing the cost and effort of obtaining labeled data.

2.4.4 Reinforcement Learning

Reinforcement Learning is a type of ML algorithm where an agent learns to take actions in an environment to maximize a reward signal. The goal of reinforcement learning is to develop an optimal policy or decision-making strategy that maximizes the cumulative reward over time.

In reinforcement learning, the agent interacts with an environment through a series of actions and observations. The environment provides the agent with feedback in the form of a reward signal, which indicates the quality of the action taken by the agent. The agent's objective is to learn a policy that maximizes the expected cumulative reward over time.

Figure 2.4 describes basic diagram of Reinforcement Learning.

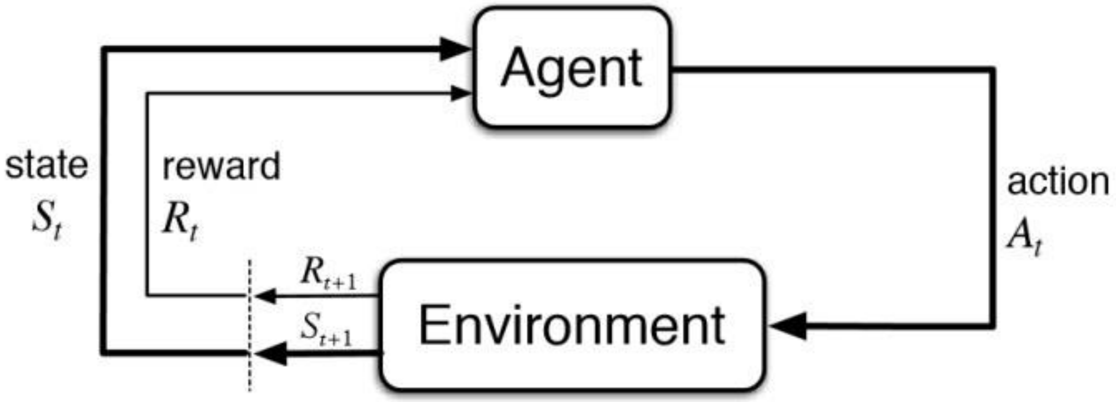


Figure 2.4 Basic diagram of Reinforcement Learning. Source: www.analyticsvidhya.com

Reinforcement learning is different from supervised learning in that the training data is not explicitly labeled. Instead, the agent learns from the feedback provided by the environment. The agent uses a trial-and-error approach to explore the environment and learn from its experience.

Reinforcement learning has various applications, including robotics, game playing, and recommendation systems. For example, in robotics, reinforcement learning can be used to train robots to perform complex tasks such as navigation, grasping, and manipulation. In game playing, reinforcement learning can be used to develop strategies for playing games such as chess and Go. In recommendation systems, reinforcement learning can be used to optimize personalized recommendations for users.

Reinforcement learning algorithms include Q-learning, SARSA, and policy gradient methods. The choice of algorithm depends on the type of environment, the reward signal, and the learning objective.

2.5 ML pipeline

ML pipelines – "A sequence of data processing components. Pipelines are very common in ML systems, since there is a lot of data to manipulate and many data transformations to apply." (Geron, A., 2020). A ML pipeline refers to the process of building, training, evaluating, and deploying a ML model. It is a sequence of steps that involve data preparation, feature engineering, model selection, and performance evaluation. The goal of a ML pipeline is to automate the entire process of building a ML model and to create a reproducible and scalable workflow.

Figure 2.5 illustrates a standard ML pipeline.

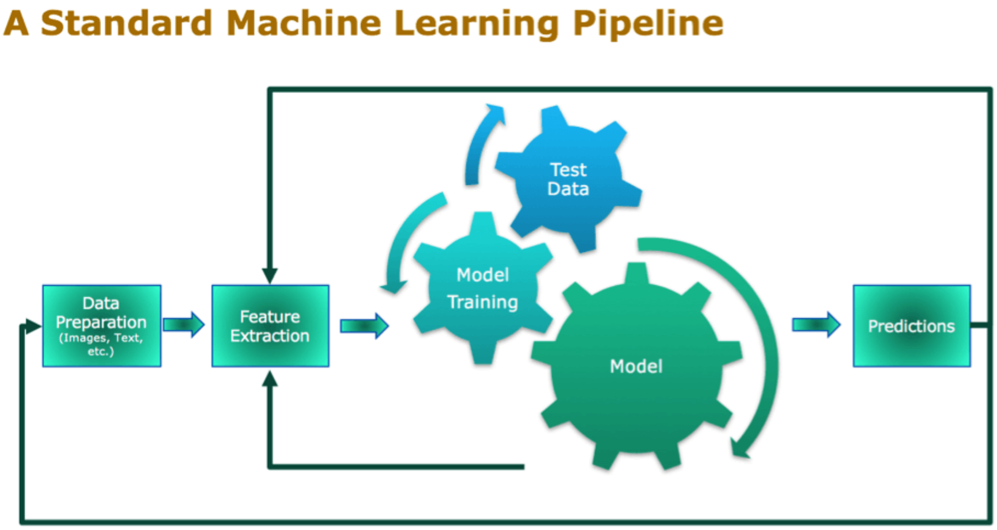

A Standard Machine Learning Pipeline
Data Preparation (Images, Text, etc.)
Feature Extraction
Test Data
Model Training
Model
Predictions

Figure 2.5. Standard ML pipeline. Source: blog.westerndigital.com

The following are the main steps involved in a typical ML pipeline:

- Formulating the problem. These questions are essential because they set the framework for how to approach the issue, choose the appropriate data and method, modify the model, and finally assess performance.
- Data collection: This involves gathering relevant data from various sources and storing it in a suitable format.

- Data preparation: This involves cleaning and transforming the raw data to make it suitable for analysis. It includes steps such as data cleaning, data normalization, feature scaling, and handling missing values.
- Feature extraction: This involves selecting and transforming the features in the data to improve the performance of the model. It includes steps such as feature selection, feature extraction, and feature scaling.
- Model selection: This involves selecting the most appropriate ML algorithm for the given task. It includes steps such as choosing the algorithm, setting hyperparameters, and optimizing the model.
- Model training: This involves training the selected ML algorithm on the preprocessed data to learn patterns and relationships.
- Model evaluation: This involves assessing the performance of the model on a test set to measure its accuracy, precision, recall, and other metrics.
- Prediction: This involves deploying the model in a production environment to make predictions on new, unseen data.
- Visualization and insights: These are crucial steps in the ML pipeline that help to understand the patterns and relationships in the data, and to interpret the results of the model

The ML pipeline can be automated using various tools and frameworks such as Apache Spark, TensorFlow, and Scikit-learn. Automation of the pipeline enables the entire process to be performed faster, with more accuracy, and with less human intervention.

2.5.1 Formulating the problem

Before beginning the real data collecting, the following questions should be addressed: Exactly what is a company goal? How will a business utilize and profit from the model? These questions are essential because they set the framework for how to approach the issue, choose the appropriate data and method, modify the model, and finally assess performance. The ideal query is occasionally: Do we even need to develop a model? Accurate responses to these inquiries have a direct impact on income and resource allocation [25].

2.5.2 Data collection

Data collection is the second step in the ML pipeline, where relevant data is gathered from various sources and stored in a suitable format for analysis. It involves identifying the data sources and collecting the required data to train the ML model.

Data collection can be a time-consuming process, and the quality of the data collected can significantly impact the accuracy of the model. Therefore, it is essential to carefully select the data sources and ensure that the data is reliable, relevant, and representative of the problem domain.

There are several methods for data collection, including:

- Web scraping: This involves extracting data from websites using tools such as BeautifulSoup, Scrapy, or Selenium.
- API calls: This involves collecting data from third-party APIs such as Twitter, Facebook, or Google Maps.
- Sensor data: This involves collecting data from sensors such as cameras, microphones, or accelerometers.
- Surveys: This involves collecting data by conducting surveys or questionnaires.
- Public data sources: This involves collecting data from public databases such as UCI ML Repository or Kaggle.

Once the data is collected, it is typically stored in a database or a file system. It is then preprocessed and transformed to prepare it for analysis using the subsequent steps in the ML pipeline.

2.5.3 Data preparation

Data preparation is a crucial step in the ML pipeline, as it involves cleaning, transforming, and organizing the raw data to make it suitable for analysis. Data preparation ensures that the data is of high quality, consistent, and relevant to the ML task at hand. The following are the main steps involved in data preparation:

- Data cleaning: This involves removing irrelevant or duplicate data, dealing with missing or null values, and handling outliers.
- Data transformation: This involves converting the data into a suitable format for analysis. It includes steps such as encoding categorical variables, normalizing or scaling numerical variables, and reducing the dimensionality of the data.
- Data splitting: This involves dividing the data into a training set and a test set. The training set is used to train the ML model, while the test set is used to evaluate its performance.
- Data augmentation: This involves generating new data from the existing data to increase the size and diversity of the dataset. It is especially useful when the dataset is small or imbalanced.

- Data validation: This involves checking the quality and consistency of the data. It includes steps such as cross-validation, which ensures that the model is robust to variations in the data, and outlier detection, which identifies data points that are significantly different from the rest of the data.

The quality of the data used to train a ML model has a significant impact on its performance. Therefore, it is important to devote sufficient time and resources to data preparation to ensure that the data is of high quality and suitable for the intended purpose.

2.5.4 Feature extraction

Feature extraction is a process in ML that involves selecting and transforming the most relevant features from the raw data to improve the performance of a ML model. Feature extraction is important because it helps to reduce the dimensionality of the data, while retaining the most important information.

Feature extraction involves the following steps:

- Feature selection: This involves selecting the most relevant features from the raw data. The selection of features is based on their relevance to the problem at hand, their correlation with other features, and their ability to discriminate between different classes or categories.
- Feature transformation: This involves transforming the selected features into a new representation that is more suitable for analysis. The transformation can be linear or nonlinear and can involve techniques such as scaling, normalization, PCA, and t-distributed stochastic neighbor embedding (t-SNE).
- Feature engineering: This involves creating new features from the existing ones to improve the performance of the model. Feature engineering can be done manually or automatically and involves techniques such as polynomial features, interaction terms, and feature cross.

Feature extraction can be done manually or automatically. Manual feature extraction is time-consuming and requires domain knowledge and expertise. Automatic feature extraction, on the other hand, uses ML algorithms to learn the most relevant features from the raw data. Automatic feature extraction is useful when the number of features is large or when domain knowledge is limited.

Feature extraction is a crucial step in ML as it helps to reduce the dimensionality of the data, while retaining the most important information. Feature extraction can be done manually or automatically, depending on the size and complexity of the dataset and the availability of domain knowledge.

2.5.5 Model selection

Model selection is the process of selecting the most appropriate ML algorithm for the given task. Model selection is a crucial step in the ML pipeline, as the choice of algorithm can have a significant impact on the performance of the model. The following are the main factors to consider when selecting a ML algorithm:

- Type of problem: The choice of algorithm depends on the type of problem, i.e., classification, regression, clustering, or anomaly detection. Each problem requires a different type of algorithm.
- Size and complexity of data: The choice of algorithm also depends on the size and complexity of the data. For example, decision trees and random forests work well for small datasets, while deep learning algorithms work well for large and complex datasets.
- Accuracy and interpretability: Some algorithms such as decision trees and logistic regression are highly interpretable, while others such as deep learning algorithms are less interpretable but more accurate.
- Training time and resource requirements: The choice of algorithm also depends on the available resources such as time, computational power, and memory. Some algorithms such as linear regression are fast and require less computational power, while others such as deep learning algorithms are slow and require a lot of computational power.
- Hyperparameters: Each algorithm has hyperparameters that need to be tuned to achieve optimal performance. The choice of algorithm also depends on the ease of tuning the hyperparameters.

Some commonly used ML algorithms include linear regression, logistic regression, decision trees, random forests, SVMs, k-nearest neighbors, and neural networks. The choice of algorithm depends on the specific requirements of the task at hand. It is common practice to experiment with multiple algorithms and compare their performance before selecting the best one.

2.5.6 Model evaluation and training

According to Amazon ML dictionary: “The process of training an ML model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model refers to the model artifact that is created by the training process (Docs.aws.amazon.com, 2021).

By examining a huge number of samples and attempting to identify a model that minimizes losses, the ML algorithm creates a prediction model. Loss is a constant measure of how well the model predicts. The value of loss decreases with improved model prediction. Finding a combination of weights (which indicate how significant an input value is) and biases (which minimize variance and so provide flexibility and improve generalization) that yields a particular algorithm's smallest loss for all samples is the aim of the training procedure (Google Developers. 2021).

Once the data has been prepared and the appropriate algorithm has been selected, the next step is to train the model on the training dataset and evaluate its performance on the test dataset. The following are the main steps involved in model training and evaluation:

- Model training: In this step, the selected algorithm is applied to the training dataset to learn the patterns in the data. The model parameters are adjusted iteratively to minimize the error between the predicted values and the actual values in the training dataset.
- Model validation: After the model has been trained on the training dataset, it is evaluated on the validation dataset to check its performance. The validation dataset is used to fine-tune the model parameters and prevent overfitting.
- Model evaluation: Once the model has been trained and validated, it is evaluated on the test dataset to check its performance on unseen data. The performance metrics such as accuracy, precision, recall, and F1 score are calculated to evaluate the performance of the model.
- Model refinement: If the performance of the model is not satisfactory, the model parameters are adjusted, and the training process is repeated until the desired level of performance is achieved.
- Model deployment: Once the model has been trained, validated, and evaluated, it is ready for deployment. The model can be deployed in various ways such as embedding it into a software application, serving it as a web service, or deploying it as an API.

The performance of the model depends on various factors such as the quality of the data, the choice of algorithm, the hyperparameters, and the training process. It is important to evaluate the model's performance carefully and fine-tune the parameters iteratively to achieve the best performance.

2.5.7 Prediction

Prediction is the final step in the ML pipeline, where the trained model is used to make predictions on new, unseen data. Once the model has been trained and evaluated on the test dataset, it can be used to make predictions on real-world data. The following are the main steps involved in prediction:

- Model application: Once the data has been prepared, the trained model is applied to the new data to make predictions. The model takes the input data as input and outputs the predicted target variable.
- Prediction evaluation: After making the predictions, the performance of the model is evaluated using various performance metrics such as accuracy, precision, recall, and F1 score. The performance of the model on the new data may differ from its performance on the test data, so it is important to evaluate the model's performance carefully.
- Model refinement: If the performance of the model is not satisfactory, the model parameters are adjusted, and the training process is repeated until the desired level of performance is achieved.
- Deployment: Once the model has been trained and evaluated on the new data, it can be deployed in various ways such as embedding it into a software application, serving it as a web service, or deploying it as an API.

Prediction serves as the fundamental and predominant application of ML across a multitude of domains encompassing finance, healthcare, marketing, and engineering, to name a few. The veracity and precision of these predictions are inherently reliant on the impeccable quality of the underlying data, the astute selection of an appropriate algorithm that aligns with the specific problem at hand, and the diligent tuning of hyperparameters to optimize the model's performance. Consequently, meticulous evaluation of the model's efficacy assumes paramount importance, necessitating a comprehensive analysis of its predictive capabilities vis-à-vis the expected outcomes. This rigorous evaluation enables practitioners to discern potential shortcomings, identify areas for improvement, and subsequently iterate upon the model, iteratively refining and honing it to attain optimal performance and unleash its full potential in delivering accurate and reliable predictions.

2.5.8 Visualization and insights

To gain a better understanding let us look at the definition: “Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from.” (Google Developers. 2021).

Data visualization's major goal is to convey information in a way that is simple to understand. One of the most understandable languages to humans is the language of visuals and forms. It also enables further understanding of the results of model predictions. Despite the amazing advancements in data science and ML, it is crucial to keep the user's expectations and experiences in mind (SearchBusinessAnalytics.com, 2021).

Visualization and insights are important for communicating the results of the ML model to stakeholders and decision-makers. The insights can help to identify new patterns and relationships in the data, to improve the accuracy of the model, and to guide decision-making in various fields such as finance, healthcare, marketing, and engineering. Here are some use cases and impoertance of visualization and insights:

Enhanced Interpretability. ML models, such as neural networks or ensemble models, can be complex and difficult to interpret. Visualization techniques, such as plots, charts, or interactive dashboards, provide a visual representation of the model's behavior and predictions. This enables stakeholders and decision-makers to grasp the underlying patterns, relationships, and decision boundaries of the ML model, leading to improved trust and confidence in the results.

Clear Communication of Findings. Visualization and insights allow for the clear and concise communication of the ML model's findings. Instead of presenting raw data or complex statistical metrics, visual representations enable stakeholders to quickly grasp the main takeaways, key trends, and significant patterns. Insights derived from the ML model can be visually conveyed through interactive charts, heatmaps, or infographics, making it easier for stakeholders to absorb and understand the implications of the results.

Decision Support and Validation. Visualization and insights provide decision-makers with a clear understanding of the impact and implications of ML models. By visualizing the predictions, trends, or anomalies identified by the ML model, decision-makers can validate the model's performance against their domain knowledge or business objectives. This facilitates data-driven decision-making, as stakeholders can assess the potential risks, opportunities, or areas for improvement based on the visualized insights.

Stakeholder Engagement and Collaboration. Visualizations and insights act as powerful tools for engaging stakeholders and promoting collaborative discussions. By presenting the ML results in a visually appealing and interactive manner, stakeholders from different backgrounds can actively participate in discussions, ask questions, and contribute their expertise. This fosters a collaborative environment where stakeholders can collectively make sense of the ML results, align on the next steps, and drive informed decisions.

Impact Assessment and Communication. Visualization and insights allow decision-makers to assess the impact of ML models on different aspects of the business or organization. Visualizations can highlight the predicted outcomes, identify influential factors, or showcase the potential return on investment. These visual representations make it easier to communicate the value and implications of ML models to stakeholders, supporting the decision-making process by providing a clear understanding of the potential benefits and risks.

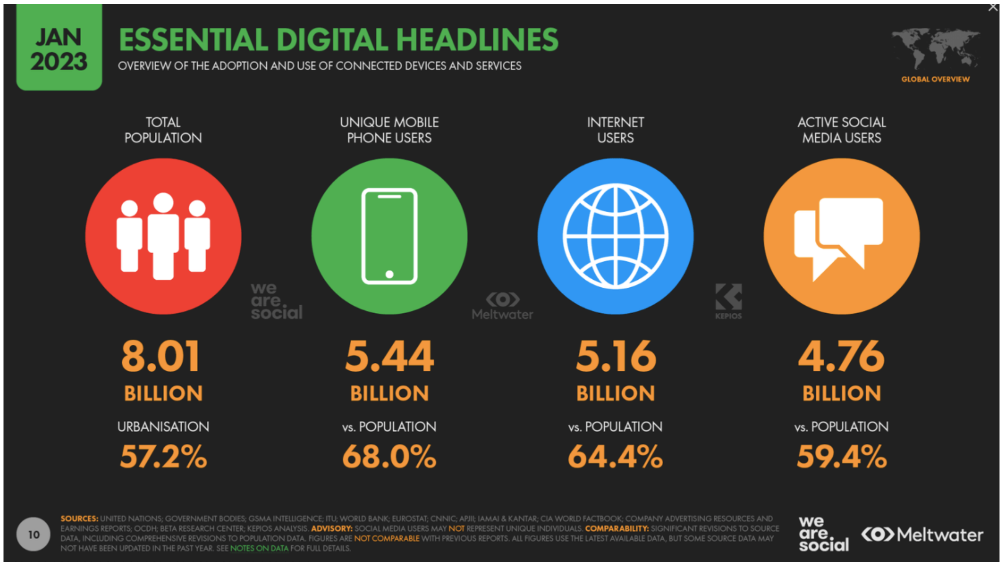
In summary, visualization and insights are crucial for effectively communicating the results of ML models to stakeholders and decision-makers. They enhance interpretability, enable clear communication of findings, provide decision support and validation, facilitate stakeholder engagement and collaboration, and assist in impact assessment and communication. By leveraging visualization and insights, organizations can ensure that ML results are understood, trusted, and utilized to drive informed decision-making and action.

The subject matter of ML is characterized by its vast expanse, encompassing a wide array of interconnected and multifaceted topics that contribute to its comprehensive understanding. However, within the confines of this specific chapter, the writer endeavors to present a succinct yet comprehensive overview of ML and its intricately woven pipeline, thereby establishing a foundation of knowledge that is essential for readers to delve into the depths of this thesis and acquire a profound and comprehensive comprehension of the subject matter at hand. In essence, this chapter assumes a pivotal role in not only providing an introduction to the fundamental principles and workings of machine learning but also in facilitating the reader's journey towards achieving a more nuanced, intricate, and profound understanding of the intricate and dynamic world of ML within the broader context of this thesis.

3. MARKETING: REACHING CUSTOMERS WITH DIGITAL TECHNOLOGY

3.1 Marketing and digital marketing

According to Harvard dictionary Marketing - "a job that involves encouraging people to buy a product or service" (towardsdatascience.com). The term digital marketing refers to the practice of promoting products or services through digital technologies, such as the internet, mobile phones, social media, search engines, and other digital channels. The goal of digital marketing is to reach a targeted audience and engage with them in a personalized and meaningful way, with the aim of converting them into customers and retaining their loyalty. Digital marketing encompasses a wide range of tactics and techniques, including search engine optimization (SEO), content marketing, social media marketing (SMM), email marketing, mobile marketing, and digital advertising. It is a rapidly evolving field that requires a deep understanding of consumer behavior, data analytics, and technology. As consumers spend more and more time online, on social networks, and on messengers nowadays, the marketing sector has simply followed its clients to new media.



SOURCES: UNITED NATIONS; GOVERNMENT BODIES; GSMA INTELLIGENCE; ITU; WORLD BANK; EUROSTAT; CNNIC; APJII; IAMAI & KANTAR; CIA WORLD FACTBOOK; COMPANY ADVERTISING RESOURCES AND EARNINGS REPORTS; OCDH; BETA RESEARCH CENTER; KEPIOS ANALYSIS. ADVISORY: SOCIAL MEDIA USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. COMPARABILITY: SIGNIFICANT REVISIONS TO SOURCE DATA, INCLUDING COMPREHENSIVE REVISIONS TO POPULATION DATA. FIGURES ARE NOT COMPARABLE WITH PREVIOUS REPORTS. ALL FIGURES USE THE LATEST AVAILABLE DATA, BUT SOME SOURCE DATA MAY NOT HAVE BEEN UPDATED IN THE PAST YEAR. SEE NOTES ON DATA FOR FULL DETAILS.

Figure 3.1 World digitalization data. Source: smartinsights.com

According to recent data, there are more than 5.44 billion mobile phones users and 5.16 billion internet users as of January 2023 (Fig.3.1). Companies and marketing professionals cannot overlook this segment of the client base.

Marketing has been a crucial aspect of any business and with the advancement of digital technology, marketers now have access to various tools and strategies to reach customers more effectively. This essay aims to discuss digital technologies that businesses can use to reach customers and examine how these tools have revolutionized the field of marketing. One of the most significant benefits of digital technology is the capability of businesses to target a broader audience. With the internet, businesses can market their products and services to people globally, regardless of location. Social media platforms such as Facebook, Twitter, and Instagram have millions of active users, which make them excellent marketing channels for businesses of all sizes.

Additionally, businesses have the option to utilize pay-per-click (PPC) advertising as a means of directing their marketing efforts to a specific audience. PPC is implemented by placing advertisements on social media platforms or search engine result pages and paying for each click the advertisement receives. This approach allows businesses to target customers who are actively searching for their products or services, thereby increasing the likelihood of conversion.

Another cutting-edge technology in marketing is AI. AI has the capability of personalizing marketing messages and recommending products based on a customer's behavior and preferences. This can result in greater engagement and conversions, as customers feel that the business comprehends their needs and interests.

To sum up, the field of marketing has been revolutionized by digital technology, enabling businesses to access an extensive range of tools and strategies to efficiently reach customers. These techniques include social media, SEO, email marketing, PPC advertising, and many more. As AI continues to progress, we can expect to witness even more inventive marketing techniques in the future.

3.2 The 11 categories of digital marketing

Digital marketing may be categorized in a variety of ways. And these categories are:

SEO:SEO is the process of enhancing a website and its content to improve visibility and attract organic (unpaid) traffic from search engines like Google, Bing, and Yahoo. The primary aim of SEO is to enhance the quality and relevance of a website's content, making it more appealing to both users and search engines (Wilson, R. F., & Pettijohn, J. B., 2006: 121–133) .

There are various techniques involved in SEO such as keyword research and optimization, on-page optimization, link building, and content creation. Keyword research and optimization involves identifying relevant search terms and phrases that people are searching for and integrating them into a website's content. On-page optimization entails optimizing the structure, content, and

HTML code of a website's pages to make them more search engine friendly. Link building involves obtaining backlinks from other websites to boost a website's authority and relevance in the eyes of search engines. Content creation involves generating high-quality and relevant content that can attract and engage visitors, encouraging them to spend more time on a website.

SEO is crucial for businesses and organizations as it can help them improve their online presence, target a broader audience, and drive more traffic to their websites. By optimizing their websites for search engines, businesses can increase their visibility and attract more potential customers, resulting in increased sales and revenue.

Search engine marketing (SEM): The practice of promoting websites by making them more visible in search engine results pages (SERPs) mostly through paid advertising is known as SEM. It is also frequently referred to as PPC advertising or sponsored search. Through the use of targeted keywords and phrases that potential consumers are likely to look for, SEM seeks to increase the number of visitors to a website (Santos, M. V. B., et al., 2022).

SMM: The process of marketing using social media sites like Facebook, Twitter, LinkedIn, Instagram, and others is known as SMM. To develop brand recognition, produce leads, improve traffic, and boost sales, it entails developing and distributing content as well as interacting with social media users. In order to analyze and enhance performance, SMM frequently includes paid advertising, content production, community management, influencer outreach, and social media analytics (Bilgin, Y., & Kethüda, Ö., 2022: 1091–1102).

Content marketing: Content marketing is a marketing technique that involves the creation and dissemination of valuable, relevant, and consistent content to attract and retain a clearly defined audience with the aim of driving profitable customer action. This content can take various forms, such as blog posts, videos, podcasts, social media updates, and infographics, and should be engaging and useful to the target audience (Barbosa, B., et al., 2023).

The goal of content marketing is to build a relationship with potential customers by providing them with informative and entertaining content that addresses their needs, solves their problems, or simply entertains them. By doing so, businesses can establish themselves as experts in their industry, earn trust and credibility, and become a go-to resource for their audience's needs.

Content marketing is typically used in combination with other marketing strategies, such as SEO, social media, and email marketing, to increase website traffic, generate leads, and ultimately convert leads into paying customers. Effective content marketing requires a thorough understanding of the target audience's interests and needs, as well as a dedication to producing high-quality content on a regular basis.

Email marketing: Email marketing refers to the practice of sending commercial messages to a group of people via email with the purpose of promoting a product or service, generating sales, building brand awareness, or fostering customer loyalty. The messages can take various forms, such as newsletters, promotional campaigns, event invitations, or follow-up emails. Email marketing is widely used by businesses of all sizes and industries due to its cost-effectiveness, scalability, and targeting capabilities (Zhang, J., & Liu-Thompkins, Y.,2023).

Mobile marketing: The technique of advertising goods, services, or brands to customers via their mobile devices, such as smartphones and tablets, is known as mobile marketing. This might involve a variety of strategies, including responsive design for mobile devices, mobile applications, SMS or MMS messaging, SMM, and advertising on mobile-friendly websites. Mobile marketing aims to connect with customers where they spend a lot of time and offer a convenient and tailored user experience (Barutçu, S., 2007: 26–38).

Video marketing: The term "video marketing" is the process of using videos to advertise and sell a commodity or service. The goal of this digital marketing approach is to raise brand recognition, generate traffic, and engage the target audience by producing and sharing films through a variety of online channels, including social media, websites, and video sharing platforms. The purpose of video marketing is to produce interesting and educational films that draw in viewers, hold their interest, and ultimately lead to lucrative consumer action (Dupin, C., 1988).

Influencer marketing: Influencer marketing is a sort of marketing approach in which a company partners with a person who has a sizable internet following in order to promote their goods or services. The influencer is picked with the intention of reaching a larger audience and boosting sales based on their fame, credibility, and capacity to impact the attitudes of their followers (Iwashita, M., 2020: 227–246).

Affiliate marketing: In a performance-based marketing strategy called affiliate marketing, companies pay affiliates for directing customers to their goods or services. Affiliates are people or businesses that market a company's goods or services in exchange for a commission for each sale or lead they create. Through a network of affiliates, it is a well-liked marketing technique utilized by companies to enhance sales and their consumer base (Akçura, M. T., 2010).

Display advertising: Online advertising known as "display advertising" uses visual components including photos, videos, and graphics to advertise a commodity or service. Display advertisements are often posted on social networking networks or other third-party websites with the aim of attracting visitors to a particular website or landing page. Display advertising may be an efficient approach to raise brand recognition and produce leads since it can be targeted based on a variety of criteria, such as demographics, hobbies, and browsing habits (Roels, G., & Fridgeirsdottir, K., 2009:452–466).

Remarketing and retargeting: Businesses employ remarketing and retargeting, two internet advertising strategies, to re-engage past consumers who have shown interest in their products or brand. Remarketing is the practice of displaying advertisements to users of a company's mobile app or website. The user may see these advertisements on other websites or online media. Remarketing aims to keep a company's name and goods in front of prospective consumers who have previously expressed interest (Piñeiro-Otero, T., & Martínez-Rolán, X., 2016:37–74).

Retargeting, on the other hand, is the technique of displaying advertisements to people who have had some sort of contact with a business but have not yet become consumers. Users who have added things to a shopping basket but never checked out, for instance, could see retargeting advertising. Retargeting aims to entice these prospective consumers back to the website or app and persuade them to carry out a desired activity, such completing a purchase or submitting a form.

The goal of digital marketing is to reach a targeted audience and engage with them in a personalized and meaningful way, with the aim of converting them into customers and retaining their loyalty. Digital marketing is about building relationships with customers and prospects through various digital channels such as search engines, social media, email, mobile devices, and other digital platforms. By utilizing various digital marketing tactics and techniques, businesses can increase their visibility, generate leads, drive sales, and ultimately, grow their brand and revenue. The ultimate goal of digital marketing is to create a strong online presence, build brand awareness, and establish a loyal customer base.

3.3 Big Data in digital marketing.

Big Data is changing the digital marketing landscape in profound ways, providing new opportunities for businesses to better understand their customers and create more effective marketing campaigns. In this section, the various ways in which Big Data is being used in digital marketing and how businesses can leverage this data to improve their marketing strategies will be explained.

One of the key ways in which Big Data is changing digital marketing is by providing businesses with access to vast amounts of data about their customers. With the explosion of digital channels, businesses can now collect data on everything from customer demographics and behavior to preferences and interests. This data can be used to gain valuable insights into customer behavior, allowing businesses to better understand what motivates their customers and how they can be reached effectively.

One of the key benefits of Big Data in digital marketing is the ability to create personalized marketing campaigns. By leveraging customer data, businesses can tailor their marketing efforts to specific individuals or groups of customers. This can include targeted advertising, personalized product recommendations, and customized messaging that is designed to resonate with specific audiences. Personalization can greatly increase the effectiveness of marketing campaigns, as customers are more likely to engage with content that is relevant to their interests and needs.

Another way in which Big Data is changing digital marketing is through the use of predictive analytics. Predictive analytics involves using historical data to make predictions about future behavior. This can be incredibly valuable in digital marketing, as businesses can use this data to identify trends and make more informed decisions about their marketing strategies. For example, a business might use predictive analytics to identify customers who are most likely to make a purchase or to identify the best channels for reaching specific customer segments.

Big Data is also being used to improve the customer experience. By collecting and analyzing data on customer behavior, businesses can identify pain points and areas where the customer experience can be improved. This might include identifying issues with website navigation or identifying areas where customers are abandoning the shopping cart. By addressing these issues, businesses can improve the overall customer experience and increase customer satisfaction.

Big Data is being used to measure the effectiveness of digital marketing campaigns. By collecting data on key metrics such as website traffic, conversion rates, and customer engagement, businesses can gain valuable insights into how their marketing efforts are performing. This data can be used to optimize marketing campaigns, identify areas for improvement, and make more informed decisions about future marketing strategies. With the right tools and strategies, businesses can leverage Big Data to stay ahead of the competition and drive growth in today's digital marketplace.

3.4 Big data technologies in digital marketing

Big data technologies have revolutionized digital marketing, enabling marketers to process, analyze, and make sense of vast amounts of data in real-time (Iwashita, M., 2020:227–246). Some of the most commonly used tools in big data in digital marketing include:

Hadoop: Hadoop is an open-source software framework used for distributed storage and processing of big data. It is used in digital marketing to store and analyze large volumes of unstructured data such as social media data, clickstream data, and customer data.

Apache Spark: Apache Spark is a distributed computing engine that is used for processing large data sets. It is often used in digital marketing for real-time processing of data and analytics.

NoSQL Databases: NoSQL databases are used to store unstructured data that cannot be stored in traditional relational databases. They are widely used in digital marketing for storing and processing customer data, social media data, and other unstructured data.

Tableau: Tableau is a data visualization tool that is used to create interactive and informative visualizations of large data sets. It is widely used in digital marketing for data analysis and visualization.

Apache Kafka: Apache Kafka is a distributed streaming platform that is used for real-time data processing and analytics. It is often used in digital marketing for real-time processing of customer data, clickstream data, and social media data.

Google Analytics: Google Analytics is a free web analytics tool that is used to track website traffic and user behavior. It is widely used in digital marketing for tracking website traffic, customer behavior, and other important metrics.

R and Python: R and Python are widely used programming languages in the field of data science and analytics. They are used in digital marketing for data analysis, machine learning, and predictive analytics.

The tools mentioned above are just a few examples of the many technologies that are used in big data in digital marketing. As the field of big data continues to evolve, new and innovative tools and technologies are sure to emerge, enabling marketers to gain even deeper insights into their customers and optimize their marketing strategies for better results.

4.ML IN DIGITAL MARKETING TO PREDICT CUSTOMERS BEHAVIUOR

Customers are the lifeblood of any business. Without customers, a business cannot survive. Customers are the source of revenue for a business, and the more customers a business has, the more revenue it can generate. Customers are also essential for business growth, as they provide valuable feedback and insights that businesses can use to improve their products or services.

Customers also play a critical role in establishing a business's reputation and brand image. Positive customer experiences can lead to word-of-mouth referrals and repeat business, while negative experiences can harm a business's reputation and lead to a loss of customers. Therefore, businesses need to focus on delivering exceptional customer service and building strong relationships with their customers to ensure their satisfaction and loyalty.

Customers are often communicated with by customer-facing employees in all firms, including those in sales, marketing, and customer support. They take on the role of the company's front line. However, it is impractical for a company to periodically get in touch with every single past, present, and future consumer to learn more about their needs. It is challenging to demonstrate one-on-one attention in huge target markets, such as those with a million or more people. Additionally, since most businesses now operate online, there is no longer any direct communication between the company and its clients, who are dispersed around the globe. Language and geographic restrictions are no longer an issue.

4.1 ML in the customer acquisition process

Customers now have more alternatives for any good or service, and there are less obstacles to switching vendors (www.bigwavemedia.co.uk, 2021). Businesses now find themselves in a position where they must anticipate and comprehend what their clients may do in the future. Predictive customer analytics and ML come in helpful at this time. ML can be used in the customer acquisition process to improve the effectiveness of marketing campaigns and increase the conversion rate of leads into customers. By analyzing customer data and behavior, ML algorithms can identify patterns and insights that can inform more targeted and personalized marketing efforts.

ML can be used to optimize lead generation and lead scoring, helping businesses to identify the most promising leads and prioritize them for follow-up. ML algorithms can also be used to analyze CE and response to different marketing messages, helping businesses to tailor their messaging for maximum impact.

ML can also be used to improve customer segmentation and targeting. By analyzing customer data, ML algorithms can identify different segments of customers and their unique characteristics, allowing businesses to tailor their marketing efforts to each segment.

So how exactly do businesses acquire customers? Finding markets and possibilities comes first. The next stage is to identify a productive communication avenue via which to approach potential customers with pertinent adverts and offers. The objective in the case of the online business is to attract potential clients to the website and turn them into devoted patrons. In the other perspective businesses can acquire customers through a combination of advertising, referrals, content marketing, email marketing, SMM, SEO, partnerships, and collaborations. By leveraging various customer acquisition channels, businesses can maximize their reach and acquire new customers more effectively.

4.1.1 Finding High-Propensity prospects

Finding high-propensity prospects involves identifying potential customers who are most likely to be interested in a business's products or services and have a higher likelihood of becoming paying customers. This process can be achieved by leveraging customer data and advanced analytics techniques such as ML.

Businesses can use customer data to create a profile of their ideal customer, including demographic information, purchase history, online behavior, and other relevant factors. This information can then be used to develop a predictive model that identifies high-propensity prospects based on their similarities to the ideal customer profile.

ML algorithms can be used to analyze customer data and identify patterns and insights that can help businesses to better understand their customers and predict their behavior. For example, businesses can use ML algorithms to analyze customer purchase history and identify products or services that are most likely to be of interest to specific customers.

By identifying high-propensity prospects, businesses can tailor their marketing efforts to these customers, increasing the likelihood of converting them into paying customers. This can help businesses to optimize their customer acquisition efforts and increase their Return on Investment (ROI) from marketing campaigns.

Finding potential clients who are more likely to purchase a product is the first significant problem every marketing department faces. In this situation, the objective is to create a propensity score for each prospect the marketing division has identified. A propensity score, which highlights likelihood, is a decimal value in the range of 0 to 1.

Table 4.1 Propensity table

Prospect	Score
Elnur	0.85
Samir	0.49

What information do we require? Demographic data is the earliest and most common type of data used in marketing research. Demographic data refers to statistical information about the characteristics of a population, such as age, gender, income, education, race, ethnicity, marital status, and other relevant factors. This type of data is often used in market research, social studies, and other fields to understand patterns and trends in different populations (Bigwavemedia.co.uk, n.d.). Table 4.2 is an example of such data:

Table 4.2 Demographic data.

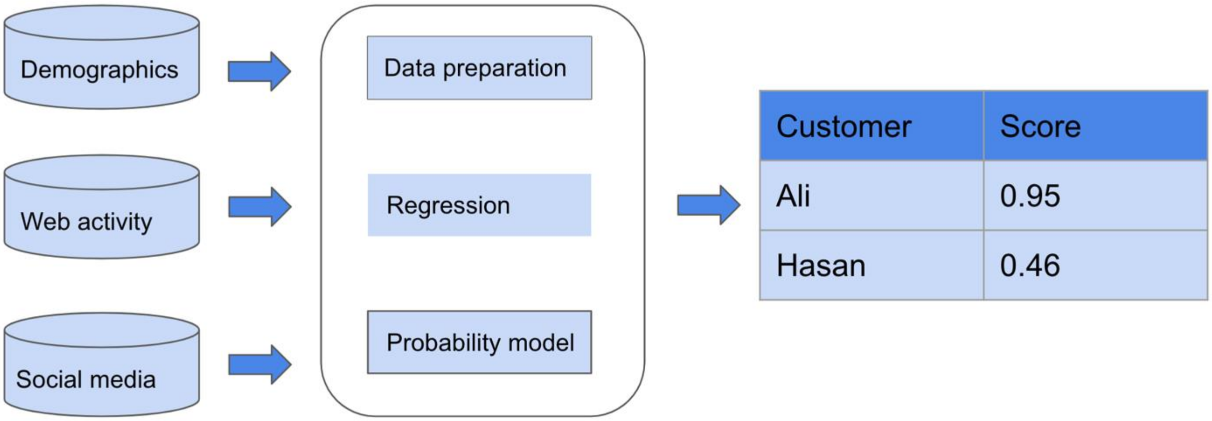
Demographic data	
Name	Emin
Age	28
Gender	Male
Employed	Yes
Income (annual)	36000 AZN
Marital Status	Married
Children	1

Prospects may have previously interacted with a business. These interactions include responding to business emails, visiting websites, making phone calls, tweeting, etc. The usage of binary flags (Y/N) is one method of storing this data.

Table 4.3 Interactions data.

Interactions	
Visited website	Y
Received emails	N
Respond to emails	N

We are using the ML pipeline discussed earlier in this thesis, which comprises of data collection, data preparation, model training, and visualization, in the pipeline shown below.



Customer	Score
Ali	0.95
Hasan	0.46

Figure 4.1 Finding High-Propensity prospects ML pipeline.

This may be thought of as a regression problem as we are seeking for a numerical number. Regression and supervised learning specifics are covered previously in this thesis. Therefore, each potential consumer is assigned a propensity score by a marketing specialist. To see which customers have the highest ratings, you may arrange them all in descending order. This data will be used for other marketing initiatives, such as phone calls or special offers.

4.1.2 Identifying the best communication channel with a prospect

Identifying the best communication channel with a prospect involves understanding their preferences and behavior. Different prospects may prefer different channels of communication, such as email, phone, social media, or direct mail.

To identify the best communication channel with a prospect, businesses can leverage data and analytics to gain insights into their behavior and preferences. For example, businesses can analyze a prospect's interaction history to understand which channels they have engaged with in the past, which channels they have responded to, and which channels have resulted in the highest conversion rates.

Businesses can also use customer surveys or other feedback mechanisms to gather information on their communication preferences. This information can be used to segment prospects based on their preferred communication channels and develop targeted marketing campaigns that are more likely to resonate with each segment.

By identifying the best communication channel with a prospect, businesses can improve the effectiveness of their marketing efforts and increase the likelihood of converting prospects into paying customers.

The following stage is to decide the best way to contact a prospect after receiving a list of the top candidates. With so many accessible channels, it's critical to target clients in a way that will garner the most interest and provide the most ROI. We will obtain a Table 4.4 containing prospects and a communication channel for each of them because of this step:

Table 4.4 Prospects and communication channels.

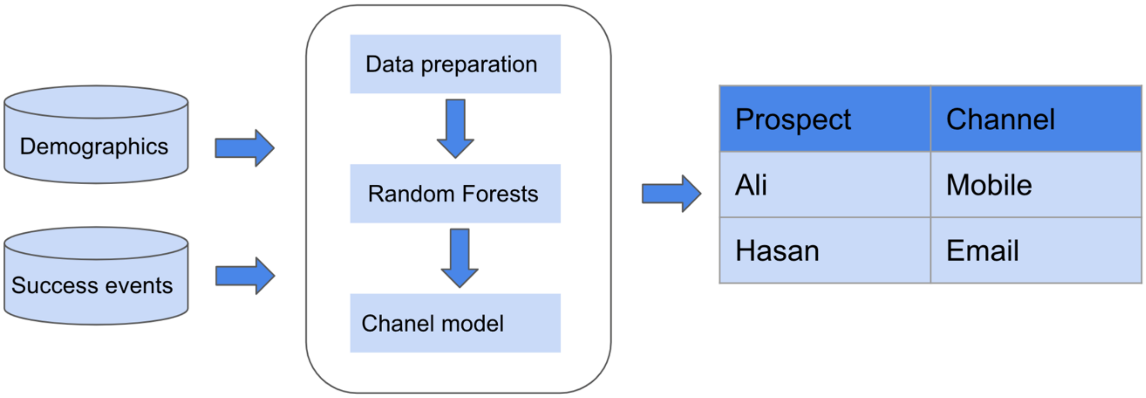
Prospect	Channel
Ali	Mobile
Hasan	Email

Which facts are necessary for success? We will once more use demographic information that you are familiar with from a prior phase, together with information on previous successful events.

Table 4.5 Past Success Events.

Past Success Events	
Opened emails	4
Clicked Pop-ups	33
Answered calls	1
Clicked Mobile Ads	12

Most of the time, we use the same or comparable ML pipelines while changing the ML method.



Prospect	Channel
Ali	Mobile
Hasan	Email

Figure 4.2. Finding best communication channel ML pipeline.

In this stage of the marketing study, a classification issue arose. We use the Random Forest algorithm to categorize clients according to various communication channels. We had discussed many categorization techniques, including decision trees, in the preceding paragraph. Here is the definition of Random Forest algorithm: “Random Forest is a robust ML algorithm that can be used for a variety of tasks including regression and classification. It is an ensemble method, meaning that a random forest model is made up of a large number of small decision trees, called estimators, which each produce their own predictions. The random forest model combines the predictions of the estimators to produce a more accurate prediction” (Thomas W., n.d.).

One of the possible categorization algorithms in this situation is Random Forest. It is regarded as best practice to test out many ML algorithms to see which one is best for a given circumstance. From this point on, we had gathered enough data to carry out a focused marketing effort. In these two instances, we employed ML pipelines to improve customer analysis performance and cut down on the required resources.

4.2 ML for predicting CLV

CLV is a marketing metric that represents the total amount of revenue a business can expect to generate from a single customer over the course of their relationship with the business. It takes into account not only the initial purchase but also the potential for repeat purchases and the length of the customer relationship.

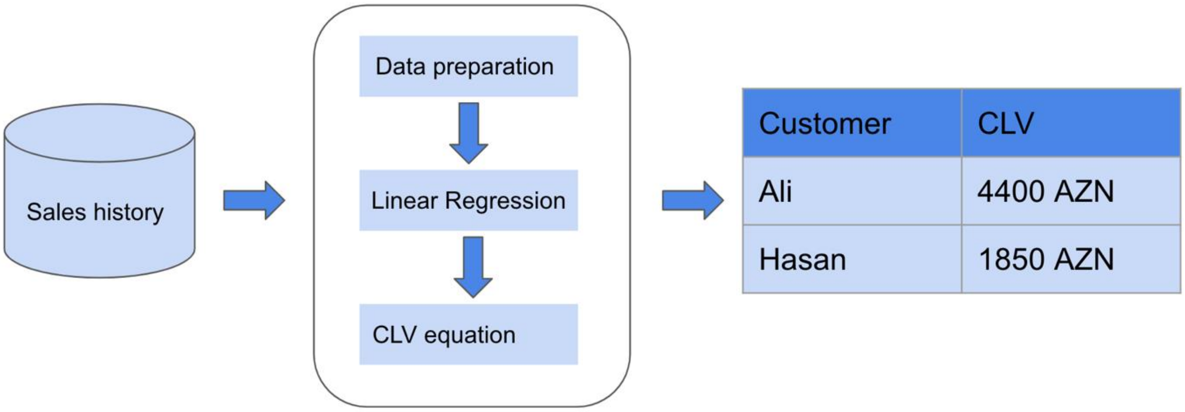
CLV is an important metric for businesses because it helps them to understand the long-term value of their customers and to make more informed decisions about customer acquisition and retention. By understanding the potential value of each customer, businesses can optimize their marketing efforts to focus on acquiring high-value customers and retaining them over time. This can help to improve customer loyalty and increase revenue and profitability over the long term.

There are several methods for calculating CLV. The formula itself has no bearing on the outcome in this case. The consistency of the formula over the whole dataset is the only thing that matters in terms of CLV. The objective of this example is to demonstrate how to create a regression model that can forecast the CLV for a new client based on that customer's recent purchasing trends and previous customer data. An example of a data record utilized for a ML prediction model is shown in Table 4.6:

Table 4.6 Monthly Sales.

Monthly Sales	
Name	Djafar
1st Month	2500 AZN
2nd Month	0 AZN
3rd Month	1900 AZN
CLV	4400 AZN

We have a regression problem here, and one of the strategies to use is linear regression. In a previous section of the thesis, we discussed this method as well as a few other supervised learning techniques.



Customer	CLV
Ali	4400 AZN
Hasan	1850 AZN

Figure 4.3. Predicting CLV ML pipeline.

The accuracy of the resultant forecast increases with the volume of data and the number of consumers. The CLV may be computed, and the model retrained with the updated customer data. This model's predictions may be used to shift the emphasis of a marketing effort from one consumer to another through further marketing study.

4.3 ML for predicting customers who might leave

ML can be used to predict customers who might leave or churn. By analyzing customer data, businesses can identify patterns and trends that indicate when a customer is becoming disengaged or dissatisfied with the product or service.

ML algorithms can be trained on historical data to identify these patterns and predict which customers are most likely to churn in the future. The algorithms can take into account a range of factors, such as the customer's purchase history, frequency of engagement, and online behavior.

By identifying customers who are at risk of churning, businesses can take proactive measures to retain them. This might involve reaching out to the customer with targeted marketing campaigns or offering incentives to encourage them to stay.

By leveraging ML to predict customer churn, businesses can improve their customer retention efforts and reduce the number of customers who leave. This can help to improve revenue and profitability over the long term by reducing the costs associated with customer acquisition and increasing CLV.

Finding potential clients for rivals is the aim of this instance. Giving each client a risk score or classifying the consumers as at risk or not at risk are the two alternative approaches to this issue. The following result, shown in Table 4.7, will be obtained when all stages have been processed via the ML pipeline:

Table 4.7 Customer attrition.

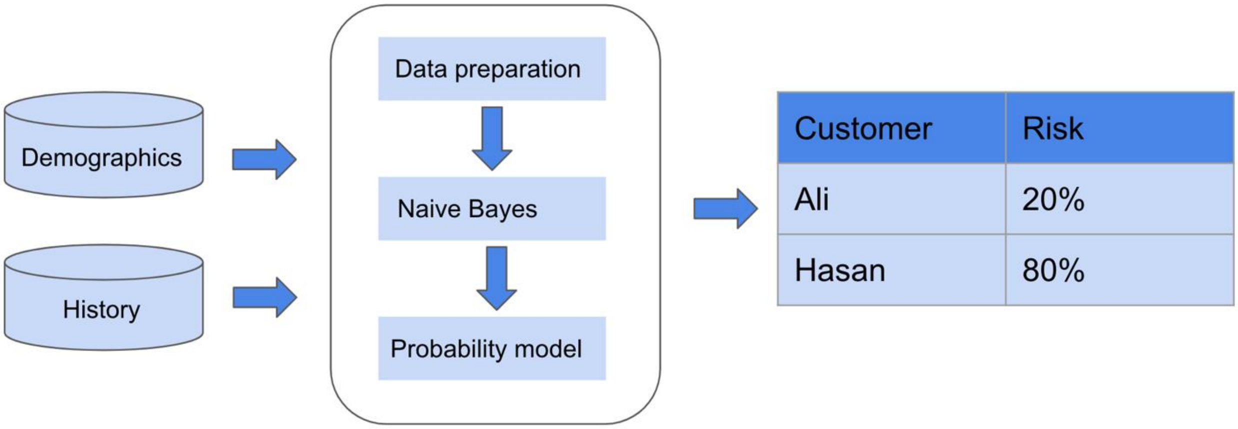
Customer	Risk
Ali	10%
Narmin	25%
Aytan	83%

Customer history records gathered from consumer activities make up the second data collection. There will be one history record per client that contains an overview of various sorts of data. Among the records in Table 4.8 is this one:

Table 4.8 Customer history record

History	
Tenure	2 years
Total value	800 AZN
Last purchase	31.10.2022
Support calls	5
Returns	1
Left?	Y

The author utilizes NB as a classification technique for this specific scenario to determine probability based on previous data. NB classifier algorithm is a probabilistic ML model that uses Bayes' theorem to classify data into different categories. It is based on the assumption that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature. This means that each feature is considered independently of the others. NB is commonly used in text classification, spam filtering, sentiment analysis, and recommendation systems (Frank, E., Trigg, L., Holmes, G. et al, 2000: 5-25).



Customer	Risk
Ali	20%
Hasan	80%

Figure 4.4 Customer attrition ML pipeline.

Marketing professionals may target specific customers based on this data to minimize customer attrition and maximize the marketing spend.

4.4 How Sentiment Analysis revolutionized marketing

Sentiment analysis, a concept rooted in the realm of NLP and ML, involves employing advanced computational techniques to meticulously scrutinize textual data, discerning and unraveling the underlying sentiment or emotional tone embedded within. This analytical methodology is frequently deployed to meticulously dissect an array of data sources such as customer feedback, social media posts, and diverse forms of user-generated content, facilitating an astute examination of the multifaceted landscape of customer sentiment and opinions, thereby yielding valuable and actionable insights for organizations operating in various domains.

The process of sentiment analysis involves using ML algorithms to classify text data as positive, negative, or neutral. This is typically done by analyzing the language used within the text and identifying key words or phrases that are associated with positive or negative sentiment. The algorithms may also take into account the context in which the text was written and the overall tone of the message.

Sentiment analysis can be used to monitor brand reputation, identify areas for improvement, and measure the effectiveness of marketing campaigns. By analyzing customer feedback and social media posts, businesses can gain insights into how their products or services are perceived and make informed decisions about how to improve the customer experience.

Sentiment analysis has a wide range of applications across different industries. Some of the main applications of sentiment analysis include:

Customer feedback analysis: Sentiment analysis can be used to analyze customer feedback from various sources such as online reviews, social media, and customer surveys. This helps businesses to understand how customers feel about their products and services, identify areas of improvement, and take corrective actions.

Brand reputation management: Sentiment analysis can help businesses to monitor their brand reputation online by analyzing social media posts, news articles, and customer reviews. This allows businesses to identify negative sentiment and respond proactively to prevent any damage to their brand reputation.

Market research: Sentiment analysis can be used to analyze customer feedback and preferences to gain insights into market trends, customer needs, and preferences. This helps businesses to make more informed decisions about product development, marketing, and sales strategies. By harnessing the potential of sentiment analysis, organizations can equip themselves with a comprehensive understanding of the sentiments expressed by their customers, enabling them to make astute and well-informed decisions pertaining to crucial aspects such as product development, marketing initiatives, and sales strategies. These insights, derived from sentiment analysis, serve as a compass guiding businesses towards targeted and personalized approaches that resonate with their customer base, fostering enhanced customer satisfaction, loyalty, and brand advocacy.

Customer service: Sentiment analysis can be employed to delve into and scrutinize the intricate dynamics of customer interactions with customer service representatives. By scrutinizing the sentiments expressed by customers during these interactions, businesses can discern and evaluate customer satisfaction levels, thereby gaining insights into areas that necessitate improvement. This analytical process subsequently equips businesses with invaluable knowledge to enhance their customer service and support infrastructure, resulting in heightened levels of customer satisfaction and, in turn, fostering increased customer retention rates.

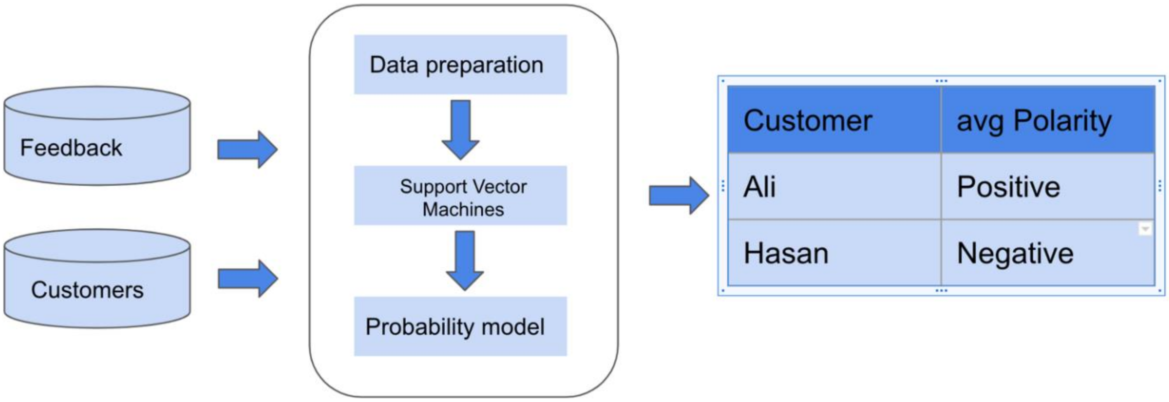
Political analysis: Sentiment analysis holds the capability to analyze vast volumes of data from social media platforms and news articles, enabling researchers and decision-makers to delve deep into the intricate realm of public opinion and sentiment surrounding political issues and candidates. By harnessing sentiment analysis, organizations and policymakers can gain valuable insights into the prevailing emotions, attitudes, and viewpoints of the general populace, thereby facilitating a comprehensive understanding of the dynamic landscape in which political actors operate. Through the systematic examination of sentiment expressed in social media posts and news articles, key patterns, trends, and shifts in public sentiment towards various political matters can be identified, empowering stakeholders to make informed decisions, devise targeted strategies, and respond promptly to the evolving political climate. Furthermore, sentiment analysis enables the detection and classification of sentiment polarities, ranging from positive to negative, neutral to highly emotive, which can be further leveraged to gauge the impact of political messages, campaign strategies, and policy initiatives on public perception. By harnessing the vast array of textual data available in the digital sphere, sentiment analysis serves as a vital tool in unraveling the complex dynamics that underpin public sentiment towards political issues and candidates, thereby enabling stakeholders to navigate the intricacies of democratic processes, engage with the electorate, and effectively shape public discourse.

Sentiment analysis has revolutionized marketing in several ways. Here are some of the key ways sentiment analysis has impacted marketing:

- Improved CE: Sentiment analysis has enabled businesses to understand their customers' needs and preferences, and tailor their marketing messages accordingly. This has led to more effective CE and improved customer satisfaction.

- Enhanced brand reputation management: Sentiment analysis has enabled businesses to monitor their brand reputation online and respond proactively to negative sentiment. This has helped businesses to protect their brand reputation and improve customer loyalty.
- Better market research: Sentiment analysis has enabled businesses to gain valuable insights into market trends and customer preferences. This has helped businesses to develop more effective marketing and sales strategies that are aligned with customer needs.
- More effective product development: Sentiment analysis has enabled businesses to gain insights into customer feedback and preferences, which has helped to inform product development decisions. This has led to the development of products that better meet customer needs and preferences.
- Increased ROI: Sentiment analysis has enabled businesses to optimize their marketing spend by targeting campaigns more effectively. By identifying high-propensity prospects and tailoring marketing messages to their needs, businesses can increase the ROI of their marketing campaigns.

As mentioned above Sentiment analysis functions as a form of social listening. Sentiment analysis incorporates NLP, text analysis, computational linguistics, and even biometrics. A significant amount of money and labor must be invested in the manual processing of reviews and customer feedback. The objective is to scale out this labor-intensive process such that millions of human-written words may be processed quickly. As a consequence, we have a pipeline that is faster and less expensive and has the same accuracy as people. Polarity assessment is the most typical use of sentiment analysis in digital marketing. Simply said, polarity assessment enables for the classification of favorable, unfavorable, or unfavorable remarks and feedback. For this categorization issue, we employ the pipeline shown in Figure 4.5:



Customer	avg Polarity
Ali	Positive
Hasan	Negative

Figure 4.5 Sentiment Analysis ML pipeline.

Sentiment analysis calls for some handling of the data, though. It is well known that only data of the numerical kind may be utilized in ML algorithms. It is important to transform text data into numerical data if the starting data collection consists only of text. Converting text data into numeric is a critical step in Sentiment Analysis, as most ML algorithms can only process numeric data. There are several methods for converting text data into numeric, including:

- Bag-of-Words: This is a simple method that involves creating a vocabulary of all the words in the text corpus and then counting the frequency of each word in each document. The resulting matrix can then be used as input for ML algorithms.
- Word Embeddings: Word embeddings are a more advanced method that represents words as dense vectors in a high-dimensional space. This method is more effective than bag-of-words as it captures the semantic meaning of words.
- Doc2Vec: Doc2Vec is an extension of word embeddings that assigns a vector representation to entire documents. This method is useful for sentiment analysis tasks that require understanding the sentiment of an entire document or sentence.
- Tokenization: is the process of breaking down a text document or a sentence into smaller units called tokens. These tokens could be words, phrases, or even sentences. It is an essential step in semantic analysis because it helps to extract meaningful information from unstructured text data. Tokenization is performed by separating words and punctuations with spaces or special characters, making it easier to analyze and process the text data.
- Stemming is a technique used in semantic analysis to reduce words to their base or root form. The purpose of stemming is to group together words that have the same base form but different endings, which can help improve the accuracy of text analysis algorithms. For example, the words "run," "running," and "ran" would all be reduced to the stem "run" through stemming. This can help reduce the dimensionality of the data and improve the performance of text analysis algorithms.

These methods can be used to convert text data into numeric features that can be used as input for ML algorithms. The choice of method depends on the specific requirements of the sentiment analysis task and the nature of the text corpus being analyzed.

Figure 4.6 illustrates the steps involved in getting data ready for sentiment analysis:

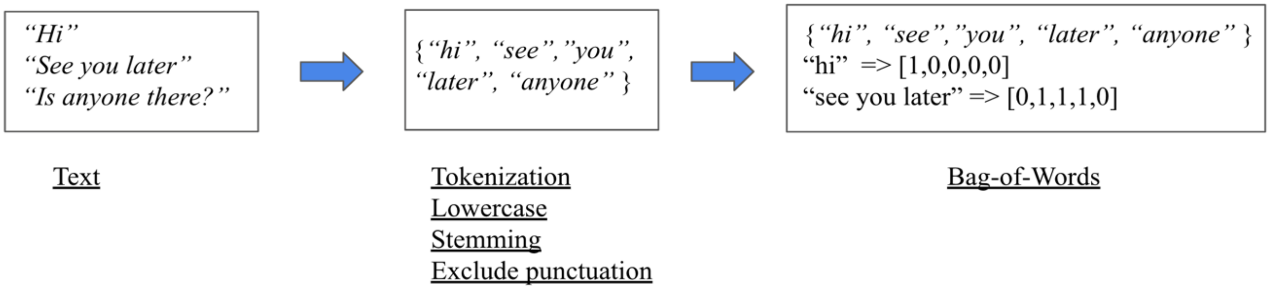


Figure 4.6 Data preparation for sentiment analysis.

The data is given a numerical representation in the last stage. One method out of several for transforming data into numeric form is presented in this thesis.

A marketing expert might want to do Sentiment Analysis for several reasons, including:

- Brand Monitoring: Sentiment Analysis can help a marketing expert monitor the sentiment around their brand or product in real-time. By analyzing social media posts, customer reviews, and other sources of feedback, a marketing expert can quickly identify potential issues and respond proactively.
- Competitive Analysis: Sentiment Analysis can also help a marketing expert analyze the sentiment around their competitors' products or services. By analyzing customer feedback, reviews, and social media posts, a marketing expert can gain insights into what their competitors are doing well and where they may be falling short.
- Product Development: Sentiment Analysis can also be used to gather customer feedback on new or existing products. By analyzing customer feedback, a marketing expert can identify areas for improvement and make data-driven decisions about product development.
- Customer Service: Sentiment Analysis can help a marketing expert identify potential issues with customer service. By analyzing customer feedback, a marketing expert can quickly identify areas where customers may be having difficulty and respond proactively.

Sentiment Analysis can help a marketing expert better understand their customers' needs and preferences, and make data-driven decisions about marketing strategy, product development, and customer service.

On a final note, the analysis has likely revealed that ML can be used to analyze vast amounts of data generated by digital channels such as social media, email, and web traffic to predict customer behavior and personalize marketing campaigns.

Additionally, the analysis likely highlights the importance of using ML algorithms to create predictive models that can identify patterns in customer behavior and anticipate future purchasing decisions. ML can be used to analyze customer data such as purchase history, demographics, and online behavior to create personalized marketing messages that are more likely to resonate with individual customers.

As digital channels continue to proliferate and customer data becomes more complex and abundant, ML will only become more important in helping businesses stay ahead of the curve and maximize their marketing ROI.

5.ML FOR CUSTOMER BEHAVIOR ANALYSIS: PREDICTING CLV USE CASE

In this chapter, the author develops and evaluates a ML algorithm for CLV prediction using the Python programming language, supplementary libraries, and a dataset. The goal is to demonstrate a use case implementation that was covered in chapter 4.2 and may be used further by any expert in digital marketing. The author concentrates on the practical aspects of ML and CLV in this section of the thesis because the majority of theoretical subjects are covered in the preceding chapters.

5.1 Libraries setup

The well-known data science distribution Anaconda is used by the author. Python 3.11.0—the most recent stable version—and the Jupiter Notebook Integrated Development Environment are both included in Anaconda (IDE). Let's begin by using "import" to add libraries to the project:

```
from sklearn.preprocessing import MinMaxScaler
from sqlalchemy import create_engine
import datetime as dt
import pandas as pd
import matplotlib.pyplot as plt
from lifetimes import BetaGeoFitter
from lifetimes import GammaGammaFitter
from lifetimes.plotting import plot_period_transactions
```

- **datetime**: A module in Python's standard library for working with dates and times. It provides classes for manipulating dates, times, and timedeltas (differences between two dates or times).
- **sklearn.preprocessing:** A module in the scikit-learn library that provides a range of functions for preprocessing input data before using it in ML models. Some common preprocessing techniques include scaling, normalization, and encoding categorical variables.
- **pandas**: A popular library for data manipulation and analysis. It provides data structures like DataFrame and Series that make it easy to work with tabular data, as well as functions for reading and writing data from a variety of file formats.

- **matplotlib.pyplot**: A module in the matplotlib library that provides a collection of functions for creating visualizations in Python. It allows users to create a wide range of plots, including line plots, scatter plots, histograms, and more.
- **lifetimes**: A library for analyzing CLV in Python. It provides functions for fitting probabilistic models to customer transaction data, and can be used to estimate CLV and other metrics like customer churn and purchase frequency. Lifetimes is a powerful tool for businesses looking to optimize their marketing and customer acquisition strategies by identifying their most valuable customers and predicting future purchase behavior.
- **BetaGeoFitter**: A class in the lifetimes library that provides a way to fit the Beta-Geometric (BG) model to customer transaction data. BetaGeoFitter is a probabilistic model used for analyzing customer purchase behavior over time. Specifically, it models the number of purchases made by a customer over time using a Poisson distribution, while modeling the probability that a customer is still "alive" (i.e., still making purchases) using a geometric distribution. The model assumes that customers vary in their purchase frequency and dropout rates, and estimates the underlying parameters that govern these variations. These parameters can then be used to predict future purchase behavior and estimate CLV. The model is useful for businesses looking to identify their most valuable customers and develop targeted marketing strategies to maximize their CLV.
- **GammaGammaFitter**: Another class in the lifetimes library that provides a way to fit the Gamma-Gamma (GG) model to customer transaction data. GammaGammaFitter is a statistical model used to estimate the average transaction value of customers in a business, based on historical transaction data. It is a variant of the Gamma distribution that is fitted to transaction value data, and it takes into account the heterogeneity of transaction values across customers. By estimating the parameters of the GammaGammaFitter model, we can calculate the expected average transaction value for each customer, which can then be used in conjunction with other models, such as the BG-NBD model, to estimate CLV. In essence, the GammaGammaFitter model helps to quantify the variance of transaction values across customers, which is a critical factor in understanding and predicting customer behavior..
- **plot_period_transactions**: A function in the lifetimes library that creates a visualization of the number of customers who made a certain number of transactions during a given period of time. This visualization can help identify any discrepancies between the model's predictions and the actual behavior of customers, and suggest areas for improvement in the

modeling process. The plot can also be useful in communicating insights to stakeholders, as it provides an intuitive visualization of customer behavior that can be easily understood by non-technical audiences.

5.2 Data Preparation

The author took the dataset that shows the sales of a British online store between 01/12/2010 and 09/12/2011 (www.kaggle.com/datasets/carrie1/ecommerce-data, 2017). Suppose that an e-commerce company wants to segment its customers and determine marketing strategies according to these segments by the start of next year (01/01/2012). These are the steps that should be executed in program:

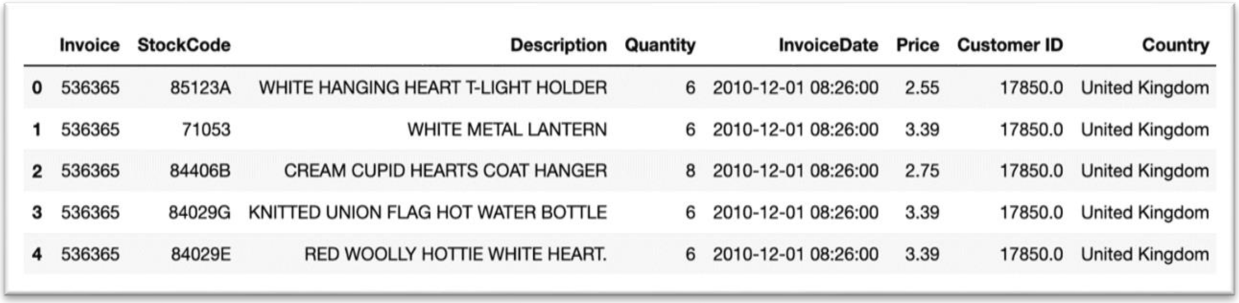
1. Data preprocessing
2. Expected Sales Forecasting with BG-NBD Model
3. Expected Average Profit with GG Model
4. Calculation of CLV with BG-NBD and GG Model
5. Creating Segments by CLV

5.2.1 Data preprocessing

Data preparation, also known as data preprocessing, is a critical step in data analysis using Python. It involves transforming raw data into a clean, consistent, and structured format that can be easily analyzed and used for modeling or visualization purposes. Data preparation ensures that the data is accurate, complete, and relevant for the analysis, leading to more reliable and meaningful insights.

```
def outlier_thresholds(dataframe, variable):
   quartile1 = dataframe[variable].quantile(0.01)
   quartile3 = dataframe[variable].quantile(0.99)
   interquantile_range = quartile3 - quartile1
   up_limit = quartile3 + 1.5 * interquantile_range
   low_limit = quartile1-1.5 * interquantile_range
   return low_limit, up_limit
def replace_with_thresholds(dataframe, variable):
   low_limit, up_limit = outlier_thresholds(dataframe, variable)
   dataframe.loc[(dataframe[variable] < low_limit), variable] = low_limit
   dataframe.loc[(dataframe[variable] > up_limit), variable] = up_limit
```

```
df_ = pd.read_excel('online_retail_II.xlsx')
df = df_.copy()
df.head(5)
```



	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

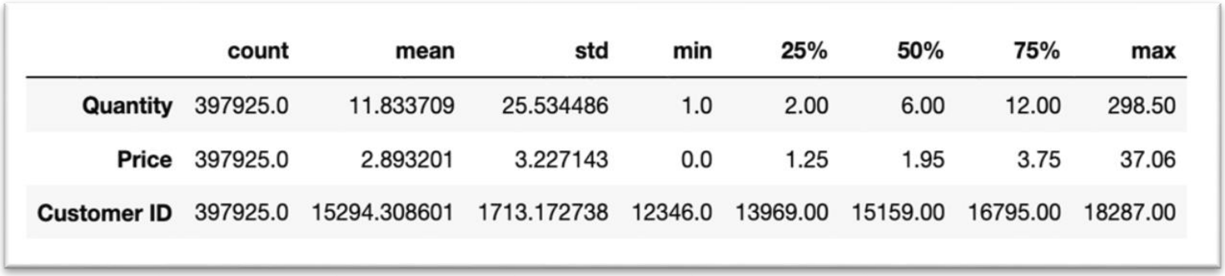
The ‘outlier_thresholds’ function in Python is typically used in data preparation to identify the upper and lower thresholds of outliers in a dataset. Outliers are observations that are significantly different from other observations in the dataset and can potentially distort the results of statistical analyses. The ‘outlier_thresholds’ function takes two arguments: the data argument is a pandas DataFrame containing the dataset to calculate the outlier thresholds from, and the variable argument is a string representing the name of the variable to calculate the outlier thresholds for.

Here is column names in dataset and their description:

- **InvoiceNo**: Invoice number. Nominal. A 6-digit integral number that is uniquely assigned to each transaction. If this code starts with the letter ‘c’, it indicates a cancellation.
- **StockCode**: Product (item) code. Nominal. A 5-digit integral number that is 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 a transaction was generated.
- **UnitPrice**: Unit price. Numeric. Product price per unit in sterling (£).
- **CustomerID**: Customer number. Nominal. A 5-digit integral number that is uniquely assigned to each customer.
- **Country**: Country name. Nominal. The name of the country where a customer resides.

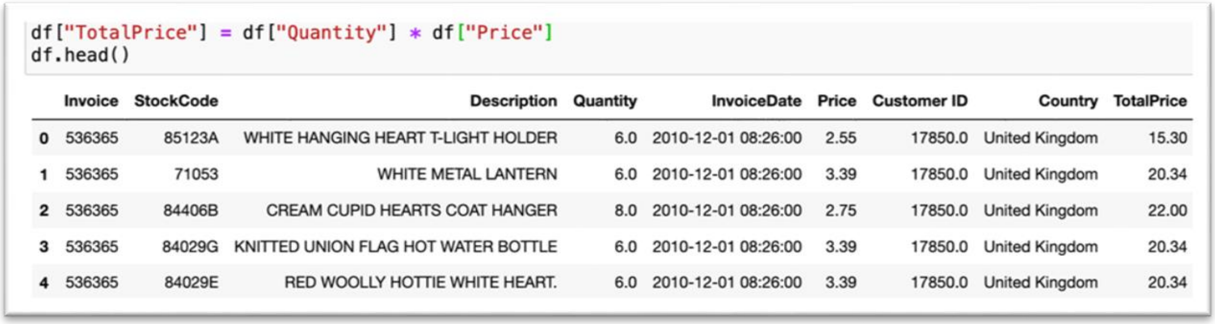
For focusing on calculation CLV, all missing values should be deleted. Next step is removing Cancellations from the data and after transactions data is described.

```
df.dropna(inplace=True)
df = df[~df["Invoice"].str.contains("C", na=False)]
replace_with_thresholds(df, "Quantity")
replace_with_thresholds(df, "Price")
df.describe().T
```



	count	mean	std	min	25%	50%	75%	max
Quantity	397925.0	11.833709	25.534486	1.0	2.00	6.00	12.00	298.50
Price	397925.0	2.893201	3.227143	0.0	1.25	1.95	3.75	37.06
Customer ID	397925.0	15294.308601	1713.172738	12346.0	13969.00	15159.00	16795.00	18287.00

To gain total price that each customer had spent, quantity of each product should be multiplied by its price:



```
df["TotalPrice"] = df["Quantity"] * df["Price"]
df.head()
```

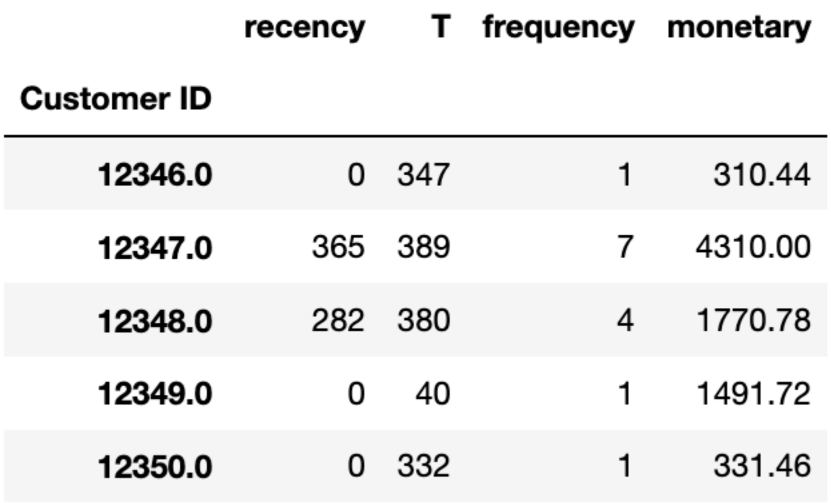
	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country	TotalPrice
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6.0	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	15.30
1	536365	71053	WHITE METAL LANTERN	6.0	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8.0	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	22.00
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6.0	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6.0	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	20.34

Then data is grouped by customers' ID and some features are calculated:

```
cltv_df = df.groupby('Customer ID').agg({'InvoiceDate': [lambda date:
(date.max() - date.min()).days,
 lambda date: (pd.Timestamp('2012-1-1') - date.min()).days],
 'Invoice': lambda num: num.nunique(),
 'TotalPrice': lambda TotalPrice: TotalPrice.sum()})
cltv_df.head(8)
```

By changing column names data comes to more understandable format:

```
cltv_df.columns = cltv_df.columns.droplevel(0)
cltv_df.columns = ['recency', 'T', 'frequency', 'monetary']
cltv_df.head(5)
```



	recency	T	frequency	monetary
Customer ID				
12346.0	0	347	1	310.44
12347.0	365	389	7	4310.00
12348.0	282	380	4	1770.78
12349.0	0	40	1	1491.72
12350.0	0	332	1	331.46

- Recency = Passed time since last purchase [Weekly];
- T = how long before the analysis date the first purchase was made [Weekly];
- Frequency = total number of repeat purchases;
- monetary_value = average earnings per purchase.

5.2.2 Expected Sales Forecasting with BG-NBD Model

The BG-NBD (Beta Geometric Negative Binomial Distribution) model is a popular probabilistic model used for CLV analysis. The model assumes that customer transactions follow a Poisson process, where the number of transactions a customer makes follows a negative binomial distribution, and the probability of a customer churning after a transaction follows a beta geometric distribution:

```
bgf = BetaGeoFitter(penalizer_coef=0.001)
bgf.fit(cltv_df['frequency'], cltv_df['recency'], cltv_df['T'])
```

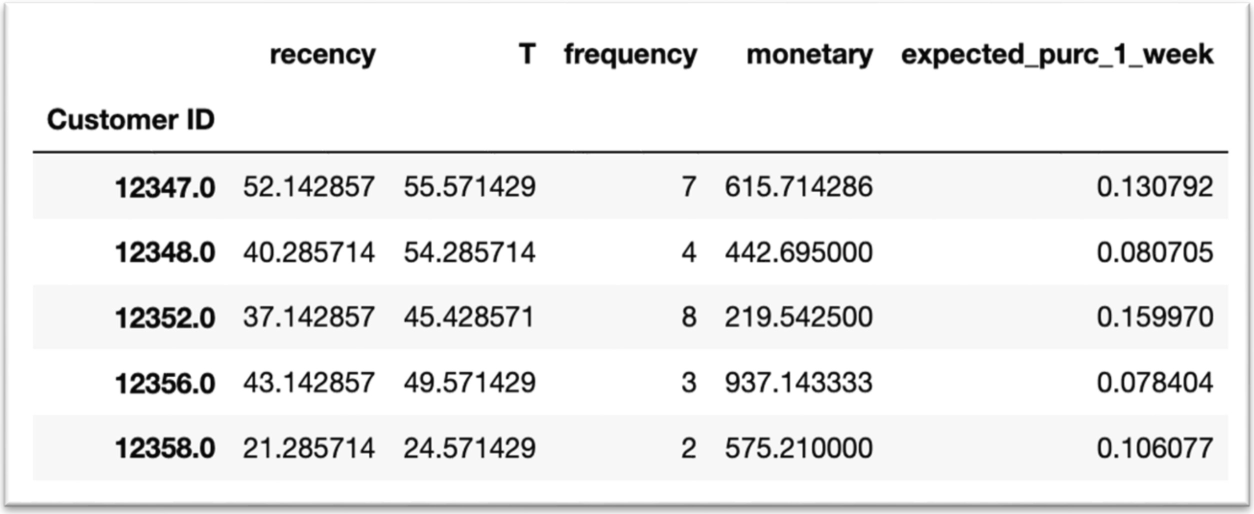
```
Output: <lifetimes.BetaGeoFitter: fitted with 2845 subjects, a: 0.22, alpha: 12.19, b: 3.08, r:
2.23>
```

Now that the model has been fitted, ML is complete. Data may be used to generate insightful models. Let's pose some inquiries and look for solutions. For example, the question ‘Who are the 5 customers we expect the most to purchase in a week?’ can be answered:

```
bgf.conditional_expected_number_of_purchases_up_to_time(1,cltv_df['frequency'],
cltv_df['recency'],cltv_df['T']).sort_values(ascending=False).head(5)
```

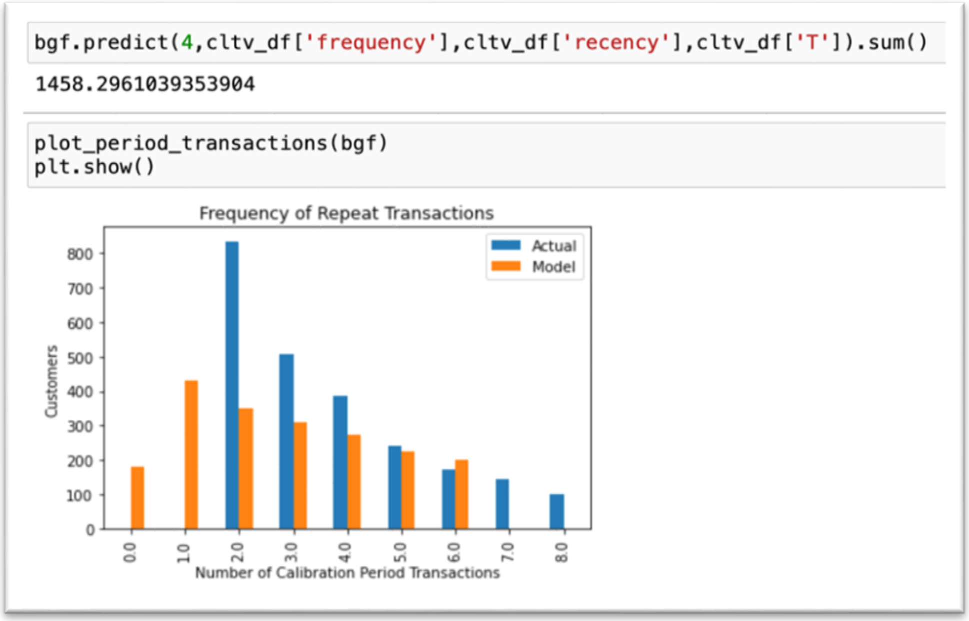
After doing some operations we can gain information about expected purchase and expected average profit by each customer:

```
cltv_df["expected_purc_1_week"] = bgf.predict(1,
                                              cltv_df['frequency'],
                                              cltv_df['recency'],
                                              cltv_df['T'])
cltv_df.head(5)
```



	recency	T	frequency	monetary	expected_purc_1_week
Customer ID					
12347.0	52.142857	55.571429	7	615.714286	0.130792
12348.0	40.285714	54.285714	4	442.695000	0.080705
12352.0	37.142857	45.428571	8	219.542500	0.159970
12356.0	43.142857	49.571429	3	937.143333	0.078404
12358.0	21.285714	24.571429	2	575.210000	0.106077

From this model we can predict the Expected Number of Sales of the Whole Company in 1 Month and visualize the Forecast Results:



```
bgf.predict(4,cltv_df['frequency'],cltv_df['recency'],cltv_df['T']).sum()
1458.2961039353904
```

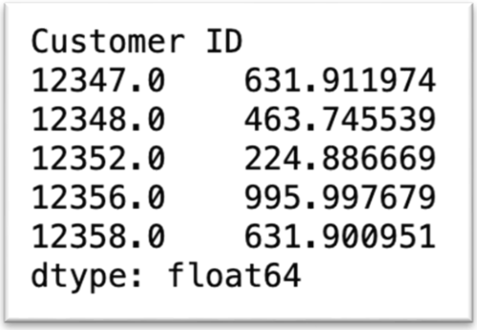
```
plot_period_transactions(bgf)
plt.show()
```

Frequency of Repeat Transactions
Actual
Model
Customers
800
700
600
500
400
300
200
100
0
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
Number of Calibration Period Transactions

5.2.3 Establishing the GG Model/Expected Average Profit with GG Model

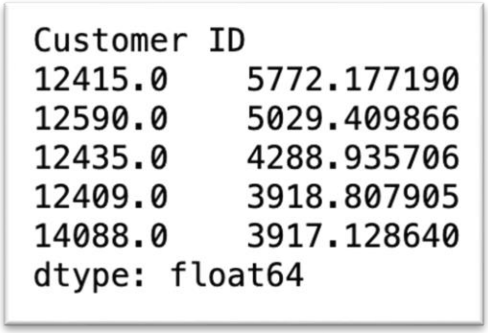
The GG model is a probabilistic model used in CLV analysis to estimate the average transaction value of customers. It assumes that the transaction values for each customer follow a Gamma distribution, and that the mean and variance of the Gamma distribution are related by a parameter called the "dispersion parameter".

```
ggf = GammaGammaFitter(penalizer_coef=0.01)
ggf.fit(cltv_df['frequency'], cltv_df['monetary'])
ggf.conditional_expected_average_profit(cltv_df['frequency'],
                                        cltv_df['monetary']).head(5)
```



```
Customer ID
12347.0    631.911974
12348.0    463.745539
12352.0    224.886669
12356.0    995.997679
12358.0    631.900951
dtype: float64
```

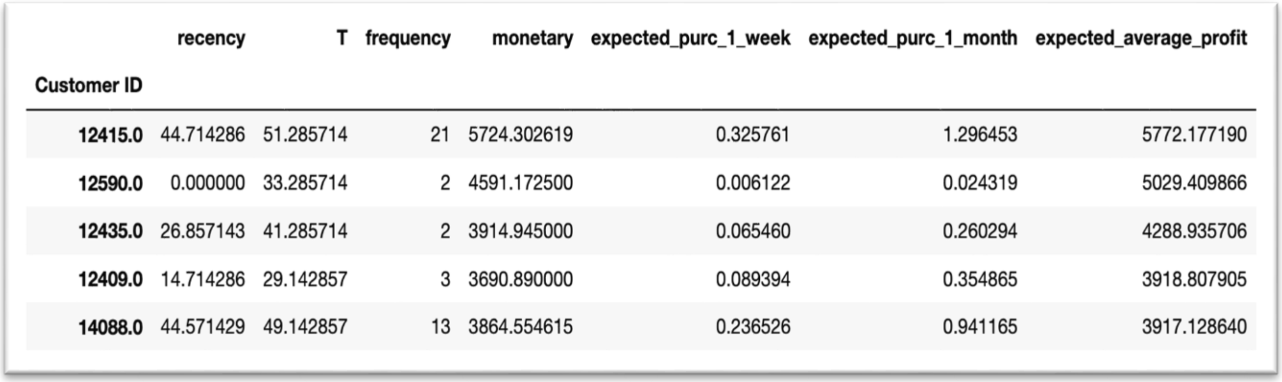
```
ggf.conditional_expected_average_profit(cltv_df['frequency'],
 cltv_df['monetary']).sort_values(ascending=False).head(5)
```



```
Customer ID
12415.0    5772.177190
12590.0    5029.409866
12435.0    4288.935706
12409.0    3918.807905
14088.0    3917.128640
dtype: float64
```

```
cltv_df["expected_average_profit"] =
ggf.conditional_expected_average_profit(cltv_df['frequency'],
cltv_df['monetary'])
cltv_df.sort_values("expected_average_profit",ascending=False).head(5)
```

Output of above script is:

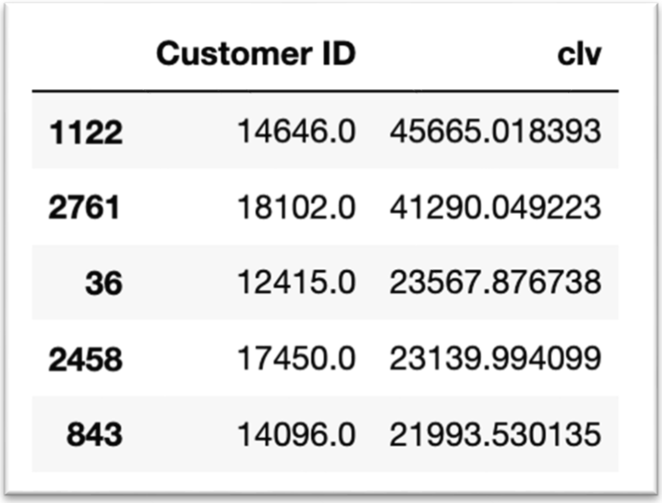


Customer ID	recency	T	frequency	monetary	expected_purc_1_week	expected_purc_1_month	expected_average_profit
12415.0	44.714286	51.285714	21	5724.302619	0.325761	1.296453	5772.177190
12590.0	0.000000	33.285714	2	4591.172500	0.006122	0.024319	5029.409866
12435.0	26.857143	41.285714	2	3914.945000	0.065460	0.260294	4288.935706
12409.0	14.714286	29.142857	3	3690.890000	0.089394	0.354865	3918.807905
14088.0	44.571429	49.142857	13	3864.554615	0.236526	0.941165	3917.128640

5.2.4 Calculation of CLV with BG-NBD and GG Model

We will now combine the two models to provide clean, evaluated CLV results.

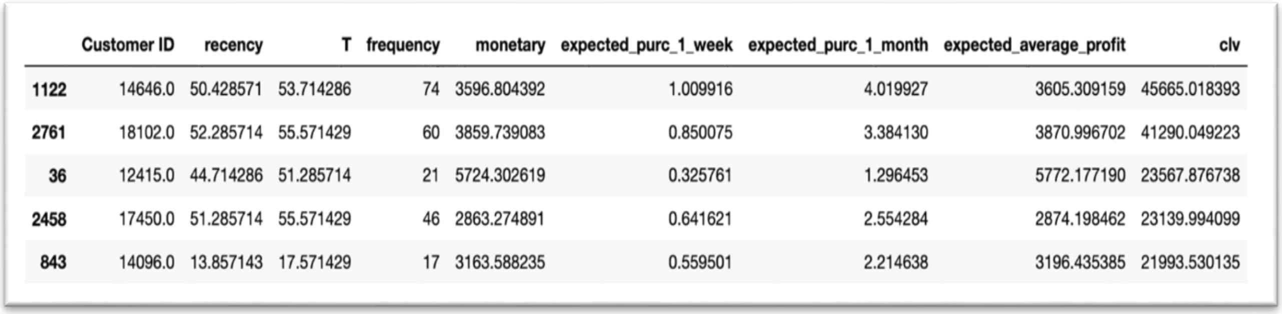
```
cltv = ggf.customer_lifetime_value(bgf,
 cltv_df['frequency'],cltv_df['recency'],
 cltv_df['T'],cltv_df['monetary'],
 time=3, freq="W",
 discount_rate=0.01)
cltv = cltv.reset_index()
cltv.sort_values(by="clv", ascending=False).head(5)
```



	Customer ID	clv
1122	14646.0	45665.018393
2761	18102.0	41290.049223
36	12415.0	23567.876738
2458	17450.0	23139.994099
843	14096.0	21993.530135

Throughout a 3-month period, we discovered the most significant customer list for me, listed from largest to smallest.

```
cltv_final = cltv_df.merge(cltv, on="Customer ID", how="left")
cltv_final.sort_values(by='clv', ascending=False).head(5)
```

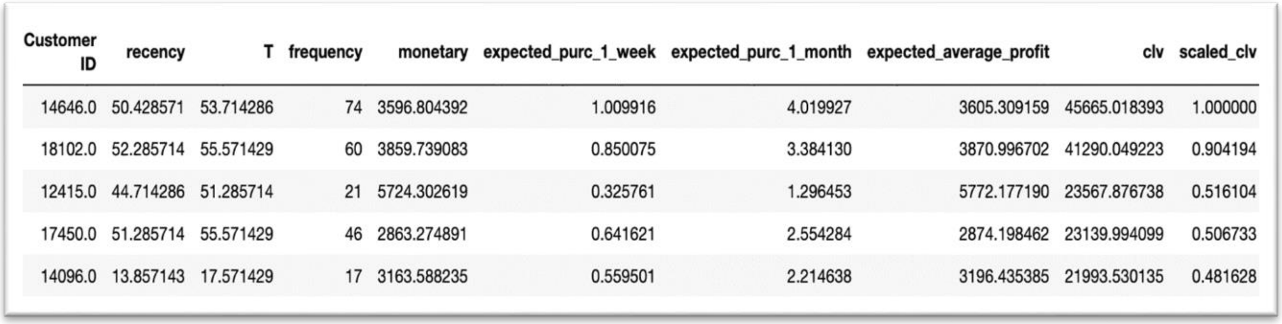


	Customer ID	recency	T	frequency	monetary	expected_purc_1_week	expected_purc_1_month	expected_average_profit	clv
1122	14646.0	50.428571	53.714286	74	3596.804392	1.009916	4.019927	3605.309159	45665.018393
2761	18102.0	52.285714	55.571429	60	3859.739083	0.850075	3.384130	3870.996702	41290.049223
36	12415.0	44.714286	51.285714	21	5724.302619	0.325761	1.296453	5772.177190	23567.876738
2458	17450.0	51.285714	55.571429	46	2863.274891	0.641621	2.554284	2874.198462	23139.994099
843	14096.0	13.857143	17.571429	17	3163.588235	0.559501	2.214638	3196.435385	21993.530135

Standardization of CLV

Standardization of CLV refers to the process of transforming CLV values into a common scale to make them more comparable and easier to interpret. This is particularly useful when you have CLV values for different segments of customers or across different time periods.

```
scaler = MinMaxScaler(feature_range=(0, 1))
scaler.fit(cltv_final[['clv']])
cltv_final['scaled_clv'] = scaler.transform(cltv_final[['clv']])
cltv_final.sort_values(by="scaled_clv", ascending=False).head()
```



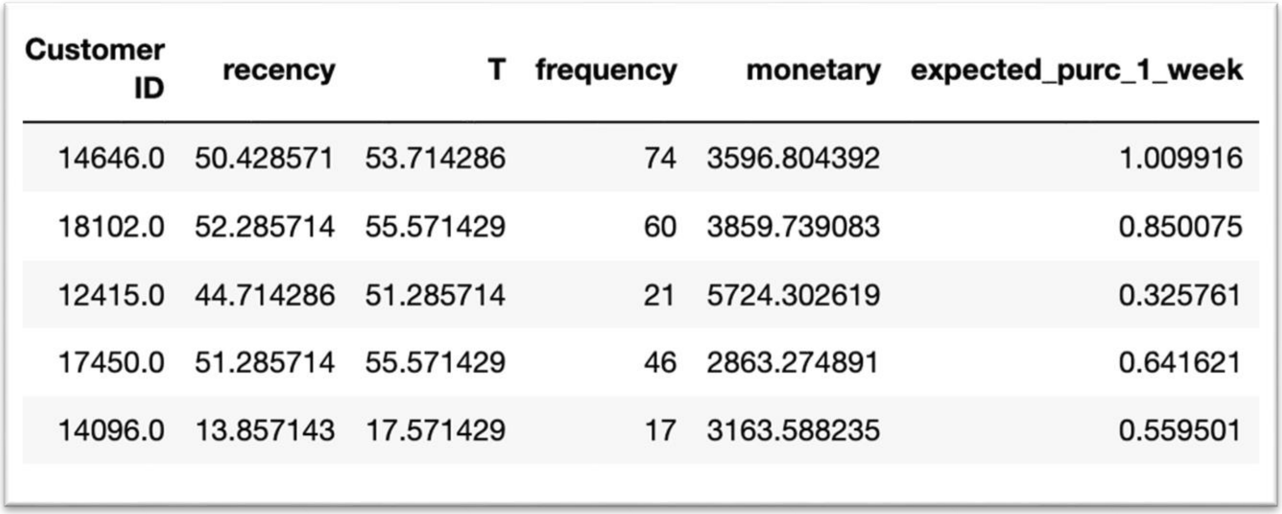
Customer ID	recency	T	frequency	monetary	expected_purc_1_week	expected_purc_1_month	expected_average_profit	clv	scaled_clv
14646.0	50.428571	53.714286	74	3596.804392	1.009916	4.019927	3605.309159	45665.018393	1.000000
18102.0	52.285714	55.571429	60	3859.739083	0.850075	3.384130	3870.996702	41290.049223	0.904194
12415.0	44.714286	51.285714	21	5724.302619	0.325761	1.296453	5772.177190	23567.876738	0.516104
17450.0	51.285714	55.571429	46	2863.274891	0.641621	2.554284	2874.198462	23139.994099	0.506733
14096.0	13.857143	17.571429	17	3163.588235	0.559501	2.214638	3196.435385	21993.530135	0.481628

5.2.5 Creating Segments by CLV

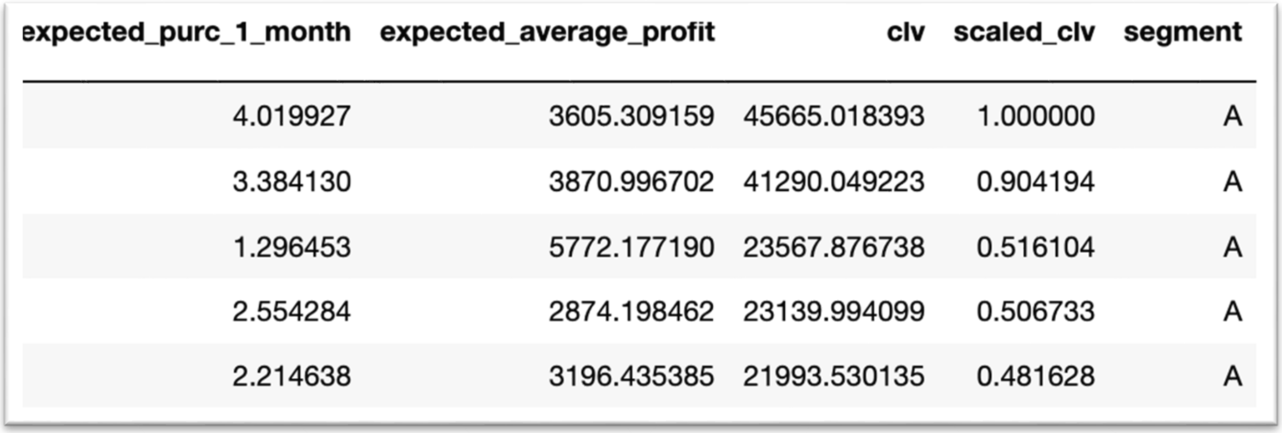
Creating segments by CLV is a useful technique to group customers based on their CLV values. This can help you identify high-value customers, low-value customers, and everything in between.

In the script provided below, a division of customers into 4 segments (A, B, C, D) was made:

```
cltv_final["segment"] = pd.qcut(cltv_final["scaled_clv"], 4, labels=["D", "C", "B", "A"])
cltv_final.sort_values(by="scaled_clv", ascending=False).head(5)
```



Customer ID	recency	T	frequency	monetary	expected_purc_1_week
14646.0	50.428571	53.714286	74	3596.804392	1.009916
18102.0	52.285714	55.571429	60	3859.739083	0.850075
12415.0	44.714286	51.285714	21	5724.302619	0.325761
17450.0	51.285714	55.571429	46	2863.274891	0.641621
14096.0	13.857143	17.571429	17	3163.588235	0.559501



expected_purc_1_month	expected_average_profit	clv	scaled_clv	segment
4.019927	3605.309159	45665.018393	1.000000	A
3.384130	3870.996702	41290.049223	0.904194	A
1.296453	5772.177190	23567.876738	0.516104	A
2.554284	2874.198462	23139.994099	0.506733	A
2.214638	3196.435385	21993.530135	0.481628	A

Calculating CLV and dividing clients into segments is a crucial step in CLV analysis. By identifying high-value customers, low-value customers, and everything in between, businesses can make more informed decisions about marketing, sales, and customer retention strategies.

To calculate CLV using the BG-NBD and GG model in Python, we first fit the BG-NBD model to the data to predict the expected number of transactions for each customer. We then fit the GG model to the data to predict the expected average profit per transaction for each customer. Using these two models, we calculate the CLV for each customer.

To standardize CLV values, we use z-scores. The resulting z-scores represent the number of standard deviations that each customer's CLV value is from the mean.

After standardizing the CLV values, we can divide customers into segments based on their z-score. For example, customers with z-score greater than 1 may be considered high-value customers, while customers with z-score less than -1 may be considered low-value customers.

6. RESULT AND FUTURE WORK

As a result of written script, I can make a conclusion that, the use of MinMaxScaler, pandas, and datetime libraries suggests preprocessed and formatted data to make it suitable for analysis. The use of BetaGeoFitter and GammaGammaFitter models implies modeled customer behavior using probabilistic approaches.

The BG-NBD model is particularly well-suited for analyzing customer purchase behavior, as it models the number of purchases made by a customer over time using a Poisson distribution, while modeling the probability that a customer is still "alive" (i.e., still making purchases) using a geometric distribution. The GammaGammaFitter model can then be used to estimate the average transaction value for each customer, which can be used in conjunction with the BG-NBD model to estimate the expected CLV for each customer.

The use of data segmentations and calculation of z-scores for customers is also noteworthy, as these techniques can help identify high-value customers and tailor marketing strategies accordingly. By segmenting customers based on their behavior and calculating their z-scores, businesses can identify customers who are most likely to make repeat purchases and are therefore likely to have a high CLV.

Overall, this project appears to be a well-designed and well-executed application of machine learning techniques to customer lifetime value analysis. The insights gained from this analysis can help businesses identify their most valuable customers and develop targeted marketing strategies to maximize their CLV.

By using this technique, businesses can better understand their customers and make more informed decisions about marketing, sales, and customer retention strategies.

In this thesis, we have presented a comprehensive analysis of customer behavior using big data analytics and ML techniques. However, there is always room for further research and improvement. This chapter will discuss some potential directions for future research in this field.

Firstly, one potential future work could be to explore the application of deep learning techniques in customer behavior analysis. Deep learning is a subset of ML that uses neural networks to model and solve complex problems. In recent years, deep learning has achieved significant success in image recognition, NLP, and speech recognition, among other fields. Applying deep learning techniques in customer behavior analysis could lead to more accurate and nuanced insights into customer behavior.

Secondly, a promising area for future research is to investigate the use of data visualization techniques in customer behavior analysis. Data visualization techniques such as heat maps, scatter plots, and line charts can help to uncover hidden patterns and insights in large datasets. By leveraging data visualization techniques, it may be possible to gain a better understanding of customer behavior and develop more effective marketing strategies.

Thirdly, an interesting direction for future research could be to explore the use of unsupervised learning techniques in customer behavior analysis. By using unsupervised learning techniques, it may be possible to uncover hidden relationships between customer behavior and other variables that were not previously considered.

Fourthly, another potential future work could be to investigate the impact of different data sources on customer behavior analysis. In this thesis, we used data from a single source, but in reality, customer data may be collected from multiple sources such as social media, customer feedback, and sales data. By examining the impact of different data sources, it may be possible to develop more accurate and comprehensive customer behavior models.

Another strategy is to apply real-time analysis. The ability to analyze customer behavior in real-time could enable businesses to respond quickly to changing customer needs and preferences. Future work could explore methods for performing real-time analysis on large volumes of customer data, such as using streaming data processing techniques and machine learning algorithms that are optimized for speed and scalability.

Lastly, another interesting area for future research could be to explore the ethical implications of customer behavior analysis. As customer behavior data becomes more prevalent, there is a growing concern about privacy and data protection. Therefore, it is important to investigate the ethical implications of customer behavior analysis and develop ethical guidelines for the collection, analysis, and use of customer data.

In conclusion, this thesis has presented a comprehensive analysis of customer behavior using big data analytics and ML. However, there are still many opportunities for further research in this field, such as the application of deep learning, the use of data visualization techniques, the use of unsupervised learning, the impact of different data sources, integrate real-time analysis and the ethical implications of customer behavior analysis.

6.1 Advantages and disadvantages

In this thesis, we have explored the use of big data analytics and ML in customer behavior analysis. Like any other approach, there are both advantages and disadvantages to using these techniques. In this section, we will discuss the pros and cons of using big data analytics and ML in customer behavior analysis.

Advantages:

- Large-scale analysis: Big data analytics allows for the analysis of large volumes of customer data, which can provide more accurate and comprehensive insights into customer behavior.
- Real-time analysis: With big data analytics, customer behavior can be analyzed in real-time, allowing for more timely and relevant marketing strategies.
- Customized marketing: ML algorithms can analyze customer data to identify patterns and preferences, which can be used to develop more personalized marketing strategies.
- Improved customer experience: By understanding customer behavior, businesses can develop more targeted and effective marketing campaigns, leading to improved customer experience and loyalty.
- Improved decision-making: Big data analytics and ML can provide businesses with data-driven insights, enabling more informed decision-making.

Disadvantages:

- Data quality: The accuracy of the insights generated by big data analytics and ML algorithms depends on the quality of the data. Poor data quality can lead to inaccurate insights and decisions.
- Data security and privacy: The use of customer data for analysis raises concerns about data security and privacy. It is important to ensure that customer data is collected and used in a responsible and ethical manner.
- Over-reliance on algorithms: ML algorithms are only as good as the data they are trained on. Over-reliance on algorithms can lead to biased or inaccurate insights, and it is important to combine ML with human expertise and judgement.
- High cost: The implementation of big data analytics and ML requires significant investments in hardware, software, and personnel, which can be a barrier for small and medium-sized businesses.
- Lack of interpretability: ML algorithms can generate complex models that are difficult to interpret, making it challenging to understand how the insights were derived and limiting their usefulness in decision-making.

The use of big data analytics and ML in customer behavior analysis has both advantages and disadvantages. While big data analytics and ML can provide businesses with valuable insights into customer behavior, it is important to address the potential issues with data quality, data security, over-reliance on algorithms, cost, and lack of interpretability. By recognizing these pros and cons, businesses can make more informed decisions about the use of big data analytics and ML in customer behavior analysis.

CONCLUSION

In conclusion, customer behavior analysis using big data analytics and machine learning techniques holds immense potential for businesses across various industries. This field of study has gained significant attention and traction in recent years due to its ability to extract valuable insights from large volumes of customer data, enabling businesses to make informed decisions and drive strategic initiatives.

Through the integration of big data analytics and machine learning, businesses can uncover patterns, trends, and hidden correlations in customer behavior that were previously challenging to identify using traditional methods. The combination of advanced analytics algorithms, scalable computing power, and vast amounts of available data has revolutionized the way customer behavior is understood and leveraged for business advantage.

By harnessing big data analytics, businesses can capture and process diverse data types, including transactional data, web browsing behavior, social media interactions, and customer feedback. These data sources enable a comprehensive understanding of customer preferences, needs, and sentiments. Machine learning algorithms play a pivotal role in transforming raw data into actionable insights, allowing businesses to predict customer behavior, personalize marketing campaigns, optimize pricing strategies, enhance customer experience, and ultimately, drive revenue growth.

The application of machine learning in customer behavior analysis has resulted in several significant advancements. For example, customer segmentation can now be performed with greater precision, allowing businesses to target specific customer groups and tailor their marketing efforts accordingly. Churn prediction models have become more accurate, enabling proactive retention strategies and minimizing customer attrition. Sentiment analysis techniques have provided deeper insights into customer sentiment, facilitating sentiment-based marketing and brand management.

Moreover, the use of machine learning in customer behavior analysis has demonstrated tangible benefits for businesses. Improved customer acquisition and retention rates, enhanced marketing campaign effectiveness, increased customer satisfaction, and higher customer lifetime value are just a few examples of the positive impacts observed in real-world implementations.

However, it is crucial to acknowledge that challenges and considerations exist in this field. Privacy concerns, ethical considerations, data quality issues, and the need for skilled data scientists are among the challenges that must be addressed to maximize the potential of customer behavior analysis using big data analytics and machine learning.

Looking ahead, the future of customer behavior analysis lies in continued advancements in big data technologies, machine learning algorithms, and artificial intelligence. As data sources grow in complexity and volume, businesses must adapt to effectively collect, store, process, and analyze this data. Integrating emerging technologies such as natural language processing, image analysis, and deep learning will further enhance the accuracy and depth of customer behavior insights.

In conclusion, customer behavior analysis using big data analytics and machine learning represents a transformative approach to understanding and predicting customer behavior. By leveraging the power of advanced analytics and large-scale data processing, businesses can gain a competitive edge, optimize their marketing strategies, and foster long-term customer relationships. As technology continues to evolve, the potential for customer behavior analysis will only expand, offering exciting opportunities for businesses to unlock the full value of their customer data.

REFERENCES

1. Kshetri, N., & Voas, J. (2018). Blockchain in Developing Countries. *IT Professional, 20(2), 11–14*. doi:10.1109/mitp.2018.021921645.
2. Prabha, D. (2021). RETRACTED ARTICLE: Customer behavior analysis using Naive Bayes with bagging homogeneous feature selection approach. *Journal of Ambient Intelligence and Humanized Computing, 12(5), 5105–5116*. doi:10.1007/s12652-020-01961-9
3. Hsu, C.-C., & Deng, C.-W. (2004). An intelligent interface for customer behaviour analysis from interaction activities in electronic commerce. *In Lecture Notes in Computer Science. Innovations in Applied Artificial Intelligence (pp. 315–324).* doi:10.1007/978-3-540-24677-0_33
4. Iakovou, S.A., et al. (2016). Customer Behaviour Analysis for Recommendation of Supermarket Ware. *In: Iliadis, L., Maglogiannis, I. (eds) Artificial Intelligence Applications and Innovations. AIAI 2016. IFIP Advances in Information and Communication Technology, vol 475. Springer, Cham*. https://doi.org/10.1007/978-3-319-44944-9_41
5. Golderzahi, V., & Pao, H.-K. (2018). Understanding customers and their grouping via WiFi sensing for business revenue forecasting. *In Lecture Notes in Computer Science. Machine Learning and Data Mining in Pattern Recognition (pp. 56–71)*. doi:10.1007/978-3-319-96133-0_5
6. Ahaggach, H. (2023). Data analytics and machine learning for smart decision making in automotive sector. *In Lecture Notes in Business Information Processing. Lecture Notes in Business Information Processing (pp. 357–363)*. doi:10.1007/978-3-031-26886-1_24
7. Alduraywish, M., Unhelkar, B., Singh, S., & Prasad, M. (2022). Application of artificial intelligence in recommendation systems and chatbots for online stores in fast fashion industry. *In Proceedings of the International Conference on Intelligent Vision and Computing (ICIVC 2021) (pp. 558–567)*. doi:10.1007/978-3-030-97196-0_46
8. Hambarde, K., Silahtaroğlu, G., Khamitkar, S., Bhalchandra, P., Shaikh, H., Tamsekar, P., & Kulkarni, G. (2020). Augmentation of behavioral analysis framework for E-commerce customers using MLP-based ANN. *In Advances in Data Science and Management (pp. 45–50)*. doi:10.1007/978-981-15-0978-0_4
9. Tsuboi, K., Shinoda, K., Suwa, H., & Kurihara, S. (2015). Collective intelligence-based sequential pattern mining approach for marketing data. *In Lecture Notes in Computer*

Science. Lecture Notes in Computer Science (pp. 353–361). doi:10.1007/978-3-319-15168-7_44

10. Chen, PL. et al. (2018). Social Network and Consumer Behavior Analysis: A Case Study in the Shopping District. In: Yen, N., Hung, J. (eds) Frontier Computing. FC 2016. Lecture Notes in Electrical Engineering, vol 422. Springer, Singapore. https://doi.org/10.1007/978-981-10-3187-8_84
11. Terano, T., Kishimoto, A., Takahashi, T., Yamada, T., & Takahashi, M. (2009). Agent-based in-store simulator for analyzing customer behaviors in a super-market. *In Lecture Notes in Computer Science. Knowledge-Based and Intelligent Information and Engineering Systems (pp. 244–251)*. doi:10.1007/978-3-642-04592-9_31
12. Shalini, & Singh, D. (2018). Comparative analysis of clustering techniques for customer behaviour. *In Advances in Intelligent Systems and Computing. Advances in Intelligent Systems and Computing (pp. 753–763)*. doi:10.1007/978-981-10-5699-4_71
13. Denman, S., Bialkowski, A., Fookes, C., & Sridharan, S. (2012). Identifying customer behaviour and dwell time using soft biometrics. *In Studies in Computational Intelligence. Studies in Computational Intelligence (pp. 199–238)*. doi:10.1007/978-3-642-28598-1_7
14. Segarra-Moliner, J.-R., & Moliner-Tena, M.-Á. (2022). Engaging in customer citizenship behaviours to predict customer lifetime value. *Journal of Marketing Analytics*. doi:10.1057/s41270-022-00195-2
15. Nie, D., Scriney, M., Liang, X., & Roantree, M. (2022). From data acquisition to validation: a complete workflow for predicting individual customer lifetime value. *Journal of Marketing Analytics*. doi:10.1057/s41270-022-00197-0
16. Ganguly, B., & Ambhaikar, A. (2022). Online E-commerce customer Point of View and sectional interest analysis for custom user experiences. *In Lecture Notes in Networks and Systems. Advances in Data and Information Sciences (pp. 393–403)*. doi:10.1007/978-981-16-5689-7_35
17. Kumar, S., Ashoka Rajan, R., Swaminathan, A., & Johnson, E. (2023). Hyper-personalization and its impact on customer buying behaviour. *In Data Intelligence and Cognitive Informatics (pp. 649–664)*. doi:10.1007/978-981-19-6004-8_50
18. Sai Teja, B. V. R., & Arivazhagan, N. (2021). Inventory prediction using market basket analysis and text segmentation—A review. *In Lecture Notes in Networks and Systems.*

Artificial Intelligence Techniques for Advanced Computing Applications (pp. 357–369). doi:10.1007/978-981-15-5329-5_34

19. Ramannavar, M., & Sidnal, N. S. (2016). Big data and analytics—A journey through basic concepts to research issues. *In Advances in Intelligent Systems and Computing. Proceedings of the International Conference on Soft Computing Systems (pp. 291–306).* doi:10.1007/978-81-322-2674-1_29
20. Kalaivani, D., & Sumathi, P. (2019). Factor based prediction model for customer behavior analysis. *International Journal of System Assurance Engineering and Management, 10(4), 519–524*. doi:10.1007/s13198-018-0739-4
21. Businessoverbroadway.com. 2021. [online] Available at: https://businessoverbroadway.com/2021/02/01/machine-learning-adoption-ratesaround-the-world/ [Accessed 19 April 2021].
22. Google Developers. 2021. ML Glossary | Google Developers. [online] Available at: https://developers.google.com/machine-learning/glossary#m
23. A. M. TURING, I.—COMPUTING MACHINERY AND INTELLIGENCE, *Mind, Volume LIX, Issue 236, October 1950, Pages 433–460*, https://doi.org/10.1093/mind/LIX.236.433
24. Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review, 65(6), 386–408*. doi:10.1037/h0042519
25. Marr, Bernard (19 February 2016). "A Short History of ML -- Every Manager Should Read". *Forbes. Archived from the original on 2022-12-05.* Retrieved 2022-12-25.
26. Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics, 36(3), 1171–1220*. doi:10.1214/009053607000000677
27. Gershgorn, Dave (26 July 2017). "ImageNet: the data that spawned the current AI boom — Quartz". *qz.com*. Retrieved 2018-03-30.
28. Canini K. et al. "Sibyl: A system for large scale supervised ML". *Jack Baskin School of Engineering. UC Santa Cruz*. Retrieved 8 June 2016.
29. Dictionary.cambridge.org. 2021. data. [online] Available at: https://dictionary.cambridge.org/dictionary/english/data
30. Geron, A., 2020. Hands-On ML with Scikit-Learn, Keras and TensorFlow. O'Reilly. *Available at: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632*
31. Docs.aws.amazon.com. 2021. Training ML Models - Amazon ML.

[online] *Available at: https://docs.aws.amazon.com/machine-learning/latest/dg/trainingml-models.html*

32. Google Developers. 2021. Descending into ML: Training and Loss | Machine Learning Crash Course. [online] *Available at: https://developers.google.com/machinelearning/crash-course/descending-into-ml/training-and-loss*
33. Google Developers. 2021. Descending into ML: Training and Loss | Machine Learning Crash Course. [online] *Available at: https://developers.google.com/machinelearning/crash-course/descending-into-ml/training-and-loss*
34. SearchBusinessAnalytics. 2021. What is data visualization and why is it important?. [online] *Available at: https://searchbusinessanalytics.techtarget.com/definition/data-visualization*
35. Visualising ML: How do we humanise the intelligence?. [online] *Available at: https://towardsdatascience.com/visualisingmachine-learning-how-do-we-humanise-the-intelligence-e62658f1f6df*
36. Wilson, R. F., & Pettijohn, J. B. (2006). Search engine optimisation: A primer on keyword strategies. *Journal of Direct Data and Digital Marketing Practice, 8(2), 121–133.* doi:10.1057/palgrave.dddmp.4340563
37. Santos, M. V. B., Mota, I., & Campos, P. (2022). Analysis of online position auctions for search engine marketing. *Journal of Marketing Analytics*. doi:10.1057/s41270-022-00170-x
38. Bilgin, Y., & Kethüda, Ö. (2022). Charity social media marketing and its influence on charity brand image, brand trust, and donation intention. *VOLUNTAS International Journal of Voluntary and Nonprofit Organizations, 33(5), 1091–1102*. doi:10.1007/s11266-021-00426-7
39. Barbosa, B., Saura, J. R., Zekan, S. B., & Ribeiro-Soriano, D. (2023). Defining content marketing and its influence on online user behavior: a data-driven prescriptive analytics method. *Annals of Operations Research*. doi:10.1007/s10479-023-05261-1
40. Zhang, J., & Liu-Thompkins, Y. (2023). Personalized email marketing in loyalty programs: The role of multidimensional construal levels. *Journal of the Academy of Marketing Science*. doi:10.1007/s11747-023-00927-5

41. Barutçu, S. (2007). Attitudes towards mobile marketing tools: A study of Turkish consumers. *Journal of Targeting Measurement and Analysis for Marketing, 16(1), 26–38*. doi:10.1057/palgrave.jt.5750061

42. Dupin, C. (1988). The video market in France: Economics of a new media. *Journal of Cultural Economics, 12(1)*. doi:10.1007/bf00220049

43. Iwashita, M. (2020). A framework of matching algorithm for influencer marketing. *The Review of Socionetwork Strategies, 14(2), 227–246*. doi:10.1007/s12626-020-00065-2

44. Akçura, M. T. (2010). Affiliated marketing. I*nformation Systems and E-Business Management, 8(4), 379–394*. doi:10.1007/s10257-009-0118-4

45. Roels, G., & Fridgeirsdottir, K. (2009). Dynamic revenue management for online display advertising. *Journal of Revenue and Pricing Management, 8(5), 452–466*. doi:10.1057/rpm.2009.10

46. Piñeiro-Otero, T., & Martínez-Rolán, X. (2016). Understanding Digital Marketing—Basics and Actions. In Management and Industrial Engineering (pp. 37–74). doi:10.1007/978-3-319-28281-7_2

47. Iwashita, M. (2020). A framework of matching algorithm for influencer marketing. *The Review of Socionetwork Strategies, 14(2), 227–246*. doi:10.1007/s12626-020-00065-2

48. www.bigwavemedia.co.uk, 2021. 3 Key Uses of Demographic Data - *Bigwave media*.

49. 3 Key Uses of Demographic Data -Bigwave media. [online] Bigwavemedia.co.uk. Available at:https://www.bigwavemedia.co.uk/blog/uses-of-demographicdata#:~:text=Demographic%20data%20is%20statistical%20data,areas%20it%20is%20most%20popular

50. Thomas W. (n.d.) https://deepai.org/machine-learning-glossary-and-terms/random-forest

51. Frank, E., Trigg, L., Holmes, G. et al. Technical Note: Naive Bayes for Regression. *Machine Learning 41, 5–25 (2000)*. https://doi.org/10.1023/A:1007670802811

52. https://www.kaggle.com/datasets/carrie1/ecommerce-data (2017)