Bachelor's thesis

Information and Communications Technology

2021

Maksim Pokrovskii

STUDY ON CUSTOMER BEHAVIOR ANALYSIS USING MACHINE LEARNING

TURKU AMK
TURKU UNIVERSITY OF APPLIED SCIENCES

BACHELOR’S THESIS | ABSTRACT

TURKU UNIVERSITY OF APPLIED SCIENCES

Degree Programme: Information and Communications Technology

2021 | 43

Maksim Pokrovskii

STUDY ON CUSTOMER BEHAVIOR ANALYSIS USING MACHINE LEARNING

Machine Learning and Artificial Intelligence have become an important part of many businesses. Digital Marketing has also benefited from modern Machine Learning techniques.

The objective of this thesis was to draw for a reader the landscape of modern marketing and Machine Learning worlds and discuss how Machine Learning can solve the problem of predicting customer behavior. Due to the complexity of both Digital Marketing and Machine Learning, the thesis starts with a brief introduction into them separately and then step by step introduces the problems in marketing that can be solved with Machine Learning. Customer behavior analysis is one of such problems that is successfully solved by Machine Learning algorithms.

A Machine Learning algorithm is created to show a real case scenario. The outcome of this thesis gives systematic knowledge to an IT expert on the application of ML for marketing. A marketing expert gains up-to-date knowledge in digital trends, Machine Learning, and algorithms. The Machine Learning algorithm in this thesis can be used by any small business or a marketing department as the first step into applied Machine Learning.

KEYWORDS:

Machine Learning, Marketing, Digital Marketing, Artificial Intelligence

CONTENTS

FIGURES

TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
CLV	Customer Lifetime Value
EDA	Exploratory Data Analysis
HCA	Hierarchical Cluster Analysis
IDE	Integrated Development Environment
LLE	Locally-Linear Embedding
ML	Machine Learning
NLP	Natural Language Processing
OS	Operating System
PC	Personal Computer
PPC	Pay Per Click
SEM	Search Engine Marketing
SEO	Search Engine Optimization
SMM	Social Media Marketing
SVM	Support Vector Machines
t-SNE	t-distributed Stochastic Neighbor Embedding
VR	Virtual Reality

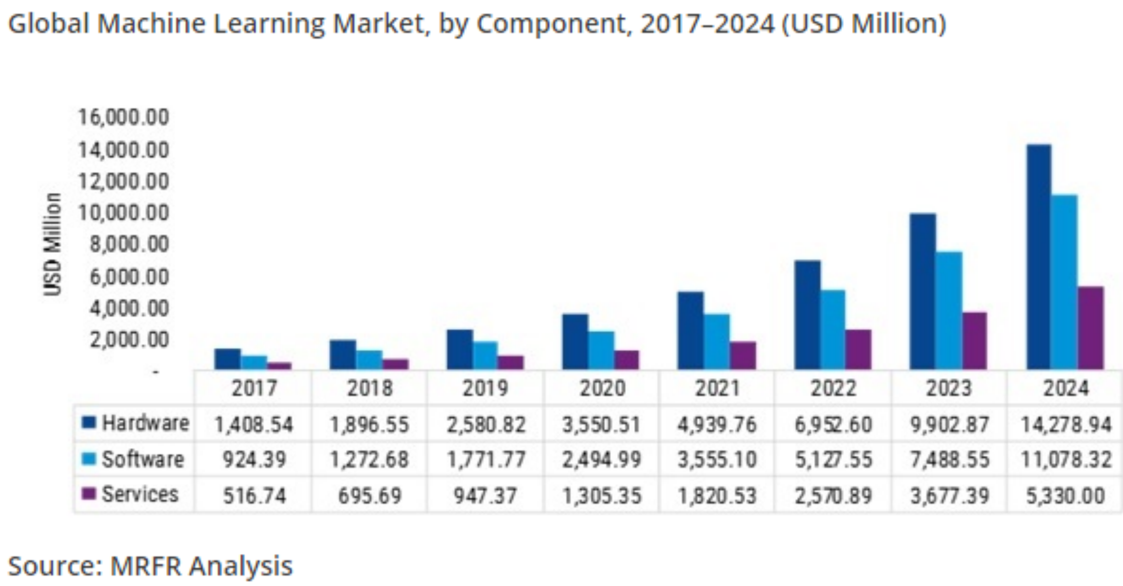
1 INTRODUCTION

Every individual is different and has habits and personal traits. Customer behavior is not an exception. We make a decision to buy, or no to buy, based on our lifestyle, experience, and feelings.

It does not matter if it is a small local bakery or a giant international network of supermarkets, it is good to know who the customers are.

Machine Learning(ML) comes in handy in this case. With the growth of digital platforms and the digitalization of business, traditional marketing methods themselves became inefficient. This does not mean ML rewrites the fundamentals of marketing and clients' behavior analysis, but it gives new tools and insights [1].

In recent years companies all over the world started actively adopting new ML tools, driven by data (Fig.1), to become more competitive in the race for clients. ML allowed businesses to significantly improve their customers' experience, thanks to a growing amount of data and wide access to high-performance computing and cloud services [2]. Figure 1 shows the growth of the global ML market over recent years.



	2017	2018	2019	2020	2021	2022	2023	2024
Hardware	1,408.54	1,896.55	2,580.82	3,550.51	4,939.76	6,952.60	9,902.87	14,278.94
Software	924.39	1,272.68	1,771.77	2,494.99	3,555.10	5,127.55	7,488.55	11,078.32
Services	516.74	695.69	947.37	1,305.35	1,820.53	2,570.89	3,677.39	5,330.00

Figure 1. Global ML market Source: datafloq.com.

The early adopters had to use large budgets, human resources, and expensive IT infrastructure to gain an advantage of ML models. However, with the development of cloud technologies and subscription-based services benefits of new digital trends became a reality for small businesses.

1.1 Research question

The main objective of this thesis is to familiarize the reader with the modern trends in ML and Digital Marketing, to highlight some marketing struggles that can be resolved with ML algorithms. Moreover, in the implementation part, a ML algorithm is created to show a real case scenario. The outcome of this thesis will give systematic knowledge to an IT expert on the application of ML for marketing. A marketing expert will gain up-to-date knowledge in digital trends, ML, and algorithms. The project showed in this thesis can be used by any small business or a marketing department as the first step into applied ML.

The thesis answers the main question: How is ML used in customer behavior analysis?

To give the topic a meaningful structure, the main research question has been broken down to the followingsub-questions.

Sub-questions:

Q1: What is ML, and its place in AI and IT landscape?

Q2: What are the key concepts of marketing and customer behavior analysis?

Q3: What are the problems marketers encounter in their profession?

Q4: How can those problems be addressed using ML?

Q5: What are the ML tools used in customer behavior analysis?

2 MACHINE LEARNING AND AI

2.1 Definition of AI and ML

In the broadest sense, AI refers to machines that can learn, reason, and act for themselves. They can make their own decisions when faced with new situations, in the same way that humans and animals can. As it currently stands, the vast majority of the AI advancements and applications refer to a category of algorithms known as ML. These algorithms use statistics to find patterns in massive amounts of data [3].

Millions of the Netflix customers use ML algorithms every day without even knowing that. ML algorithms are the main drivers of any recommendation machine on a website like Netflix, voice assistants such as Siri or Alexa, a search engine: Google or Yahoo [4].

The Figure 2 introduces the structure of AI:

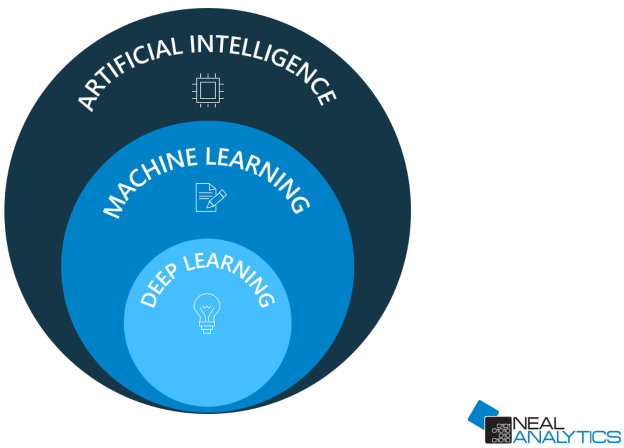


Figure 2. Structure of AI Source: nealanalytics.com.

It is clear from the picture that ML is a branch (or one of the approaches) of Artificial Intelligence. A reader might be slightly confused at this point. The author of this thesis prefers a simple definition from the Google ML dictionary: "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." [5]

In the meantime, ML evolved with the grown computational power of modern computers, something that was impossible just a decade ago has become reality.

2.2 Important milestones in AI and ML

Many people assume ML as a brand new technology, but the concept has been around for quite a while. There are several significant milestones that shaped the modern ML industry. We will take a look at some of them [6]:

<u>Before 1950s</u> - In data science, everything started from statistical methods. For decades it was the only way to analyze data. This period of time was called the Dark era.

<u>1950s</u> - The first ML researchers started in the 1950s with the world-famous "Turing test" by Alan Turing. In the same decade, the first neuro-computer was designed to recognize visual patterns, the algorithms to play checkers with the computer were invented, and the perceptron was invented.

<u>1960s</u> - The first Tic-tac-toe game was played utilizing reinforcement learning, and the nearest neighbor algorithm was created (The algorithm was used to map routes).

<u>1970s</u> - Ai winter - a period of reduced interest in AI and ML technologies.

<u>1980s</u> - LISP-based machines were developed and marketed, a program that learns to pronounce words was created, commercialization of ML on PCs started in this decade.

<u>1990s</u> - IBM's Deep Blue beat Gary Kasparov at chess, Sony introduced the first AI domestic robot AIBO, the first emotional AI machine demonstrated at MIT, random decision forests algorithm was invented.

<u>2000s</u> - The first challenge for autonomous vehicles was held, Ai-based recommendation engines were created, Google built a self-driving car.

<u>2010s</u> - In this decade Deep mind AI was developed, personal assistants became a mainstream, Oculus Rift VR headset was created, Boston dynamics created Atlas robot, Google Deep mind was invented, Deep Mind AlphaGo beat human Go player, art created by the neural network was sold for $400000, Google AI diagnosed lung cancer for the first time.

2020s - In this decade AI and ML were used for the COVID-19 fight, Mayflower autonomous ship project has started.

Even though the industry went far from its predecessor, it is important to remember where it all came from and how it has grown because it still has a long way to go. Studying the early concepts helps to understand and develop new approaches in ML and AI.

2.3 Data in ML

Data is the fuel for any ML algorithm. Data brings information about clients and their behavior. We use it to "load" the ML algorithm and to receive valuable facts. Let us start with a clear definition of data.

According to the Cambridge dictionary: data - information, especially facts or numbers, collected to be examined and considered and used to help decision-making, or information in an electronic form that can be stored and used by a computer [7]. Some sources tend to split definitions of information and data. They define information as an outcome of data processing and analysis, which brings new facts and conclusions. Meanwhile, most sources equalize concepts of data and information.

Data as a concept has been around for quite a while and well known. What is new in this concept is the idea of Big Data.

Big data - is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can not manage them. However these massive volumes of data can be used to address business problems individual would not have been able to tackle before [8]. This definition of Big Data clearly shows one of the major problems that arose with Big Data - volume.

Data scientists define four Vs of Big Data:

Volume - refers to the size of data sets that need to be processed and analyzed

Velocity - refers to the speed with which data is generated

Variety - the variety of data comes from various types of data (numeric, text, audio, video, and many others).

Veracity - refers to the quality of data that is being analyzed

Data can be categorized in many ways, but data science highlights two main types: structured and unstructured data [9]. Structured data is the data that has been labeled, categorized, and stored in a structured database. The majority of incoming data is unstructured and can not be used in any types of ML algorithms. This brings one of the main challenges of the Big Data era - turning unstructured data into structured data. This process requires a tremendous amount of computational power. Once data has been collected and structured, it can be used in ML algorithms to predict customer behavior [10].

2.4 Algorithms in ML

Data is an essential component for ML algorithms and client behavior analysis. This section discusses how to define an algorithm and its role in an ML pipeline. Let us dive into the topic.

A ML algorithm is the method by which the AI system conducts its task, generally predicting output values from given input data [11]. There are some variations on how to classify ML algorithms. The author of this thesis followed the common approach to divide algorithms by their purposes. The main categories are [12]:

- Supervised learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

2.4.1 Supervised learning

In supervised learning (Fig.3), the training data fed to the algorithm includes the desired solutions, called labels [13].

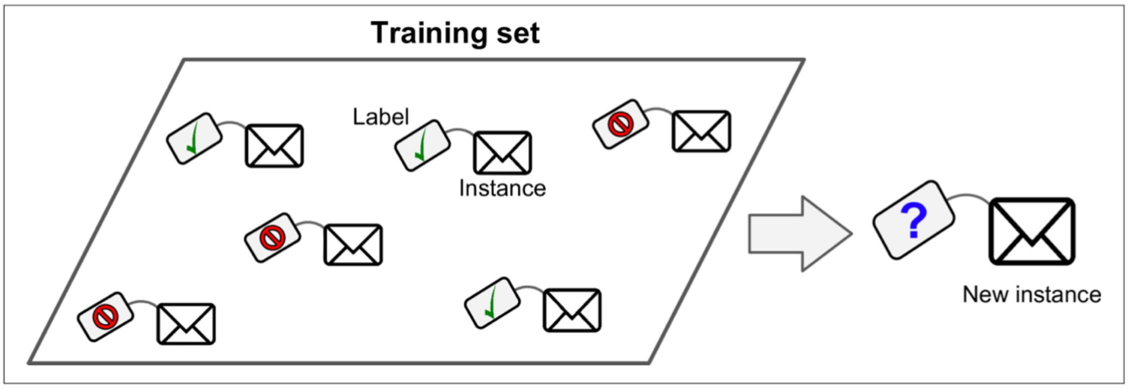


Figure 3. Supervised learning classification problem Source: Hands-on ML.

A classical supervised learning task is classification. A good examples is a spam filter. It is trained to recognize malicious emails by using many example emails. Every example email has a marker: spam or not spam. A supervised algorithm must learn how to classify new incoming emails [13].

The other typical task is to predict a numeric value, for example, a property price, based on its location, size, number of rooms, etc. This type of task is called regression. To gain the best results from training, the algorithm requires as many examples of properties as possible [13].

The following algorithms are the most common for supervised learning:

- Naive Bayes
- Support Vector Machines (SVM)
- Decision Trees
- Linear Regression
- Nearest Neighbor
- Neural Networks

2.4.2 Unsupervised Learning

In unsupervised learning (Fig.4), the input data has no predefined labels. This data is also called unlabeled. In other words, comparing to supervised learning, an algorithm tries to learn without a teacher.

For example, there is a large amount of data about website's visitors. In this case, a clustering algorithm can be utilized to detect groups of similar visitors. An algorithm finds groups of visitors by itself. For example, it can recognize, that 60% of visitors are females and they love comic books, while the other half are seniors and they prefer sci-fi, and so on [13].

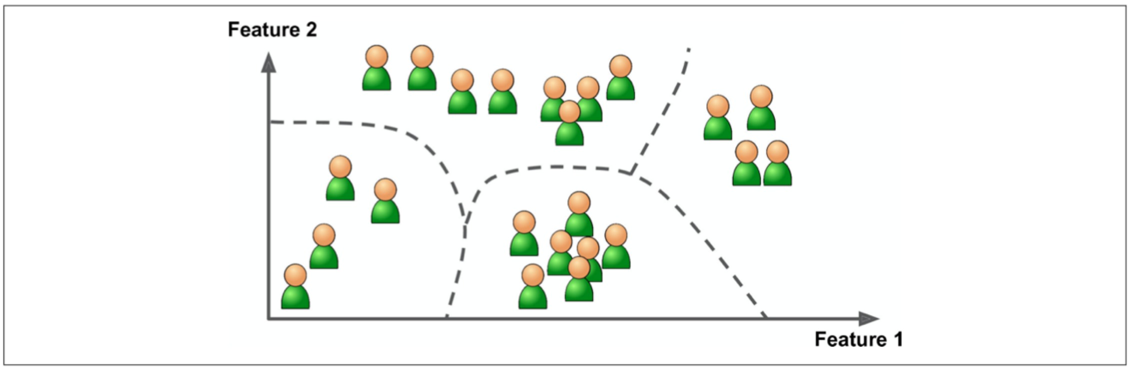


Figure 4. Unsupervised learning problem Source: Hands-on ML.

The following algorithms are the most common for unsupervised learning:

- K-Means
- Hierarchical Cluster Analysis (HCA)
- Isolation Forest
- Locally-Linear Embedding (LLE)
- t-distributed Stochastic Neighbor Embedding (t-SNE)

2.4.3 Semi-supervised Learning

In supervised and unsupervised algorithms, we either have labeled or unlabeled data. Semi-supervised algorithms are a mixture of the previous two. In some cases, a ML algorithm has to overcome some degree of uncertainty. The cost of labels might be high or even inappropriate since it requires highly skilled professionals or an enormous amount of working hours to process the data [14].

Some photo-hosting services, such as Google Photos, are good examples of this. Once family photos uploaded to the service, it automatically recognizes that the same person A shows up in photos 1, 5, and 11, while another person B shows up in photos 2, 5, and 7 [13].

2.4.4 Reinforcement Learning

Reinforcement Learning is very different compared to other types of ML algorithms. A Reinforcement Learning algorithm aims to interact with the dynamic environment, select, and perform the action. As a result of the action, the algorithm receives rewards or penalties.

For example, many robots implement Reinforcement Learning algorithms to learn how to walk. DeepMind's AlphaGo program is also a good example of Reinforcement Learning: it made the headlines in May 2017 when it beat the world champion Ke Jie at the game of Go. It learned its winning policy by analyzing millions of games, and then playing many games against itself [13].

The following algorithms are the most common for reinforcement learning:

- Q-Learning
- Temporal Difference (TD)
- Deep Adversarial Networks

2.5 ML pipeline

ML pipelines – "A sequence of data processing components. Pipelines are very common in Machine Learning systems, since there is a lot of data to manipulate and many data transformations to apply." [13] Most of the time components in a ML pipeline run asynchronously . Every pipeline component operates a tremendous amount of data. The data must be processed, split, stored, passed to other components. The somponents system makes maintenance and development process simpler. Moreover, if one of the components breaks down, it does not cause the whole system to be out of order.

Figure 5 illustrates a standard ML pipeline.

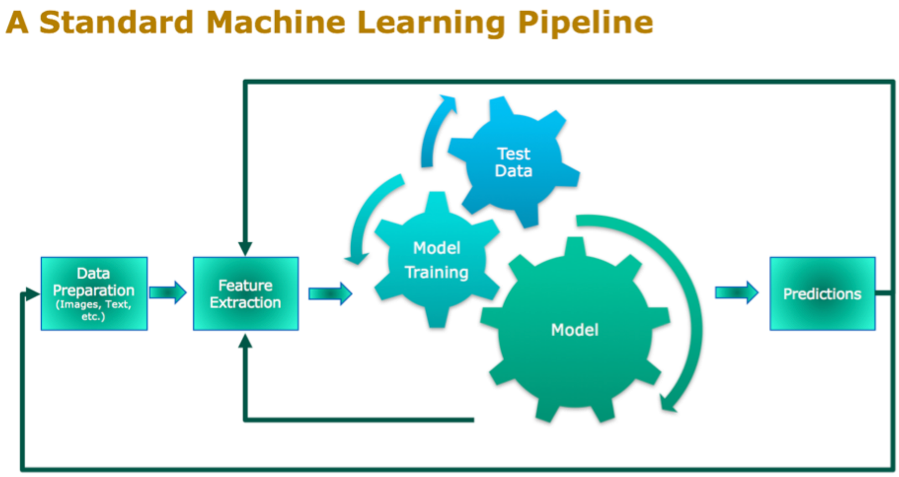


Figure 5. ML pipeline Source: opendatascience.com.

A typical machine learning pipeline would consist of the following processes:

- Formulating the problem
- Data collection and evaluation
- Data preparation
- Model training and evaluation
- Visualization and insights

2.5.1 Formulating the problem

The first questions to ask before starting the actual data collection are:

What exactly is a business goal?

How is a company going to use and benefit from the model?

These questions are crucial because they frame how to approach the problem, what data and algorithm to select, how to tweak the model, and then, measure the performance. Sometimes the best question is: Do we need to build any model at all? Receiving accurate answers to these questions directly affects revenue and resource distribution. ML can be costly, time-consuming, and require relevant research before the implementation phase [6].

2.5.2 Data collection and evaluation

In the typical environment, data stored in a database(s) and spread across multiple tables/files/objects.

There are multiple sources of data for ML models [15]:

- Google datasets search engine
- ".gov" datasets - data collected and stored by governmental authorities.
- Kaggle datasets
- Amazon datasets
- Own data, collected with research the topic of interest

These days information is diverse and accessible for analysis and further use. In most cases, to make a judgment to use, or not to use information, the data quality evaluation must be conducted. The consequences of using bad data go way beyond some incomplete rows or inconsistent records. The low data quality can cause a 10x cost increase.

There are 6 data quality characteristics to consider [15]:

1) Data is complete

This means that the main data attributes have no missing values. Missing values can skew the analysis.

2) Data is unique

Duplicates of data slow down the ML model, take storage space.

3) Data is up to date

It is essential to use data that reflects the most recent changes.

4) Data is consistent

Consistency relates to the absence of contradictions within a dataset or between various datasets.

5) Data is valid

Validity determines that data has a suitable type, format and fits the pre-defined standards. For instance, the use of data format as dd/mm/yyyy.

6) Data is accurate

Accuracy means the correctness of the data such as the date of birth or the number of units. It is crucial to understand that data accuracy is not the same as data validity. For instance, the height of the Eiffel Tower might be specified as 300 meters. The real height is 324 meters. In this case, 300 is inaccurate but valid data.

Obviously, having a perfect set of data is almost impossible. One of the solutions is to set a threshold data quality. It gives meaningful rules for data assessment and simplifies the selection process.

2.5.3 Data preparation

Data preparation is the next step after data collection in the ML pipeline and it is the process of cleaning and transforming the raw data collected in the previous step. Most ML algorithms can not work or perform poorly with raw input data, this is the reason why the data preparation phase is so important.

The first step on the way to prepared data is *Exploratory Data Analysis (EDA)*. This stage considers the following procedures [16]:

1) Feature and target values determination

In most situations using all available features is not the best idea. This approach will not result in the appropriate predictive model results. Features and target values must be chosen based on three main goals in mind [17] :

- Reduce computational cost
- Better model interpretation
- Maximize accuracy

2) Conversion of data types

From a ML perspective data can be classified into 4 basic types:

- numerical data - any numbers
- categorical data - characteristics, such as color or shape
- time-series data - sequence of numbers collected at regular intervals over some period of time
- text - words and phrases

3) Handling outliers and distribution analysis

Most common causes of outliers [18]:

- Data record errors (human errors)
- Measurement errors (tool errors)
- Experimental errors (experiment planning and execution errors)
- Intentional (dummy outliers made to test detection methods)
- Data processing errors (data manipulation and transformation errors)
- Sampling errors (wrong data sources)
- Natural (not an error, an anomaly)

The next step is *Data preprocessing*. This is the process in which data scientists spend most of their time. There is no universal step-by-step approach to go, but there are some helpful practices and essentials [18]:

1) Feature imputations - the process of handling missing values. There is a number of ways to deal with missing data, including replacing values with mean, median, mode, K-nearest neighbor, or deleting the entire row.

2) Feature encoding - encoding incoming data into numeric data. ML models can not process non-numeric values, and there are the following approaches to encode: label encoding and one-hot encoding

3) Feature normalization - the process of rescaling the values, for instance, conversion kilometers to meters.

4) Feature engineering - the method of transforming raw data into features that better depict the underlying problem that an individual is attempting to solve.

5) Feature selection - keeping the most relevant features of a dataset for an ML model

6) Dealing with data imbalances - in some cases, data can be imbalanced, causing a significant skew in the distribution. This problem can be tackled by collecting more data.

The last comes *Data Splitting*. Typical formula to split data for a model [16]:

1) Training set (70-80%) - data for model training

2) Validation set (10-15%) - data used for tunning hyperparameters

3) Test set (10-15%) - data for model performance evaluation

2.5.4 Model training and evaluation

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.”[19]

ML algorithm builds a prediction model by analyzing a large number of examples and trying to find a model to minimize losses. Loss is a universal indicator of the model's prediction accuracy. The better the model's prediction the lower the value of loss. The goal of the training process is to find a set of weights (indicating how important the input value is) and biases (bias reduces the variance and hence introduces flexibility and better generalization) that results in a minimum loss across all examples in a given algorithm [20].

To evaluate a model data scientists use validation and test datasets. Both datasets already hold target values (labels). Assessing the prediction accuracy of an ML model with the same data that was used for training is not useful, because it rewards models that can "remember" the training data, as opposed to generalizing from it. In other words, the model becomes biased. To avoid this behavior the validation dataset is used. This prevents bias and allows the tune of the model's hyperparameters to minimize losses. In the final step test dataset comes into play. The Test dataset provides the gold standard used to evaluate the model. It is only used once a model is completely trained [21].

2.5.5 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.”[20]

The main idea of data visualization is to present data, in our case - predictions and forecasts, in an easily consumable form. The language of images and shapes is one of the most accessible to humans. It also allows to learn more about the outcomes of the model predictions. While data science and ML go through exciting advances, it is essential to keep the user’s expectations and experiences in perspective [22].

Machine Learning is a wide topic to discuss. In the scope of this chapter, the author gave a brief introduction to ML and its pipeline. This chapter is vital for further understanding of this thesis.

3 MARKETING: REACHING CUSTOMERS WITH DIGITAL TECH

3.1 Marketing and digital marketing

According to Harvard dictionary Marketing - "a job that involves encouraging people to buy a product or service" [23]. The term marketing is quite all-encompassing and variable. In meantime, the term digital marketing appeared with the development of digital technologies such as internet radio, tv, smartphones, and the Internet. In simple words, digital marketing means everything related to the term marketing, but with the focus on digital solutions. Today, more and more people spend time online, on social networks, and on messengers, in respect to this transition, marketing industry just followed a customer to a new media. Although digital marketing is taking over, traditional marketing, such as print marketing, still an effective way of communication with a customer. [24].

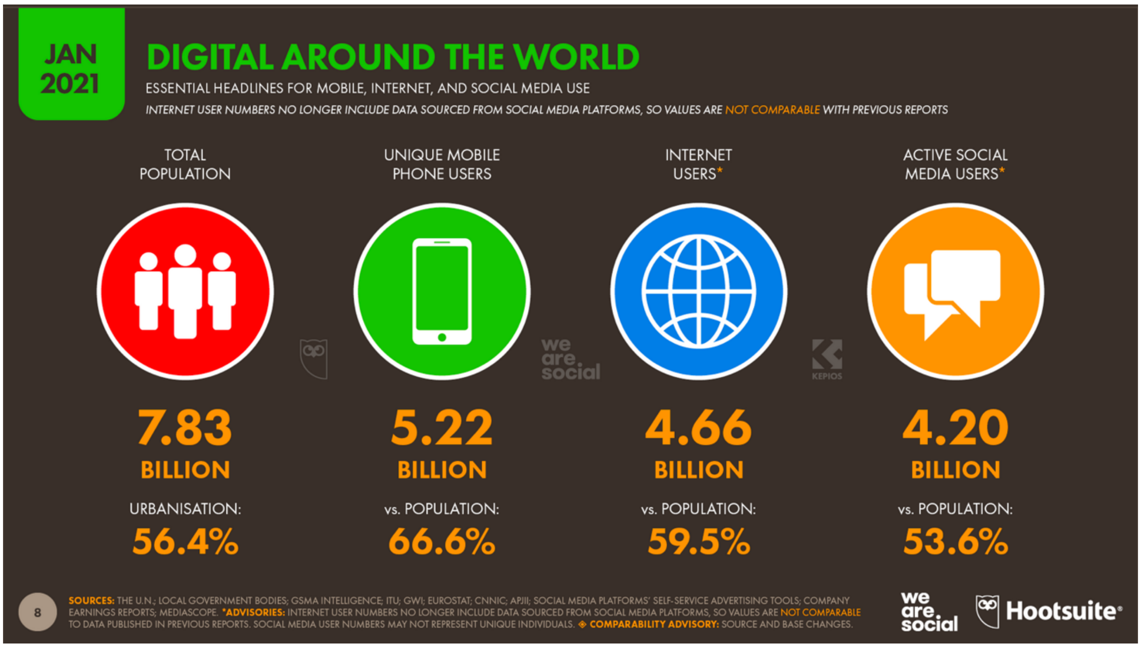


Figure 6. World digitalization data Source: developers.google.com.

According to recent data, there are more than 5.22 billion mobile phones users and 4.66 billion internet users as of January 2021 (Fig.6). This chunk of customers can not be ignored by companies and marketing experts.

3.2 The 11 categories of digital marketing

There are multiple approaches to categorizing digital marketing. In the scope of this thesis, the author follows the categorization made by one of the leading digital marketing experts Neil Patel[25].

And these categories are [26]:

SEO or Search Engine Optimization

In simple words, SEO means the process of optimization of a web page or blog to improve its visibility for search engines, such as Google. The better a webpage optimized the higher in the relevant search result it will appear. Gaining more attention brings more potential customers to a page. One of the core terms for SEO is keywords. A smartly crafted list of keywords helps search engines to distinguish the content of the website and show it in the relevant search request.

SEM or Search Engine Marketing

In various sources, the terms SEM and SEO are represented interchangeably, or SEM is highlighted as an umbrella term for SEO. Although, some authors define SEM as a set of paid strategies to promote a website and improve its visibility.

PPC or pay-per-click

PPC is an internet marketing strategy in which advertisers pay a fee each time one of their ads is clicked. It is a way to bring visits to a website not organically (by optimizing keywords and content), but by buying them.

SMM or Social Media Marketing

Social media marketing became vital after the rapid growth of social platforms such as Facebook, Instagram, and TicToc. Efficient work with social media increases sales, brand recognition and establishes a connection with the audience.

Content Marketing

This type of marketing is quite different compare to previous ones. It is not about direct advertising products or services to a customer, but rather creating quality and engaging content. The world biggest brands actively posting blogs, images, and videos, trying to attract attention.

Email marketing

It is a form of marketing that involves sending advertising emails, offers, other information to potential customers, utilizing mail lists. The most difficult part, in this case, is to send the right message to the right person, avoiding spam filters.

Influencer Marketing

This type of marketing involves endorsements and product mentions from influencers, who have a social following and some level of public attention. The other vital parameter of an influencer is a reputation as an expert in a particular niche and trust from viewers.

Viral Marketing

Some companies use this approach as a smart way to promote their products. This sales technique involves word-of-mouth information about a product or/and distinctive style of implementation. The home of viral videos is video platforms such as YouTube. The most popular viral videos gained millions of views.

Radio advertising

Although radio advertising (an audio message promoting, and aiming to market, a product or service) used to be exclusively based on radio waves, it became digital. This means, that radio advertising can be recognized as one of the fields of digital marketing.

Television advertising

TV advertising (a visual message promoting, and aiming to market, a product or service) is the other media that successfully stepped into the digital era. Digitalization allows ads to be targeted for particular viewers.

Mobile advertising

This form of marketing ultimately includes all the types we discussed before. With the rise of smartphones every form of digital marketing adopted to follow the customer. Mobile advertising can occur as text-based ads, banner advertisements, videos, or even as mobile games.

Digital marketing is all about attracting, analyzing, and communicating with clients with the adoption of modern digital innovations. In the next chapter, we will review the application of ML in digital marketing and customer analysis.

4 ML IN DIGITAL MARKETING TO PREDICT CUSTOMERS

Every business needs customers to survive. They are the source of revenue. The success of a business is directly proportional to its ability to acquire customers, nurture them, make them happy, solve their issues, and consequently make more money from them. However for that to happen, the business needs to identify the right potential customers. They have to figure out the who, what, why, and how. Who are the potential customers demanding their products? What do they want? Why do they want this particular product? And how are the customers making their buying decisions? How does a business go about doing this? Typically all businesses have customer-facing people, such as sales, marketing, and support who keep communicating to their customers. They become the front line of the company. However, it's impossible for a business to contact every potential and current customer individually from time to time to understand their needs. When the target markets are large, say, a million individuals or more, it is difficult to show one-on-one attention. Also, with most businesses going online, the business does not have any direct contact with the customers, they are scattered all over the world. The traditional barriers of geography and language are gone.

4.1 ML in the customer acquisition process

Customers today have more options for any product or service, and the barriers to switching to different vendors are becoming smaller [27]. This brings businesses to a situation where they need to understand and plan for what their customers might do in the future. At this point, ML and predictive customer analytics become handy. Predictive customer analytics uses customer data to build models. These models help to predict future behavior. It assists businesses to target prospects who will convert and identify additional products the customer might buy. When customers have problems, predictive customer analytics will serve businesses to identify the right resources to solve the problems.

How do businesses acquire customers? The first step is to identify markets and prospects. The next step is to find an efficient communication channel to reach out to prospects with appropriate advertisements and offers. In the case of the online store, the goal is to bring prospects to the website and convert them into loyal customers.

4.1.1 Finding High-Propensity prospects

The first big challenge any marketing department has is to identify prospective customers who have a higher likelihood to buy a product. The goal, in this case, is to generate a propensity score for each prospect identified by the marketing department. A propensity score is a decimal number in the range of 0 to 1 highlighting the probability.

Table 1. Propensity score.

Prospect	*Score*
Cindy	0.79
Steve	0.45

What data would we need? The first and the most widespread data for any marketing research is demographic data. "Demographic data is statistical data collected about the characteristics of the population, e.g. age, gender and income for example. It is usually used to research a product or service and how well it is selling, who likes it and/or in what areas it is most popular." [28]

Table 2 is an example of such data:

Table 2. Demographic data.

Demographic data	
Name	Steve
Age	40
Gender	Male
Employed	Yes
Income	40k
Marital status	Married
Children	1

Prospects might have a history of interaction with a company. Such interactions include commercial email responses, web sites visits, phone calls, tweets, etc. One way to store this data is to use binary flags (Y/N). An example of such data is in the Table 3:

Table 3. Interactions data.

Interactions	
Visited Website	Y
Received emails	N
Respond to emails	N

In the pipeline below, we are utilizing the ML pipeline reviewed earlier in this thesis which consists of data collection, data preparation, model training, visualization.

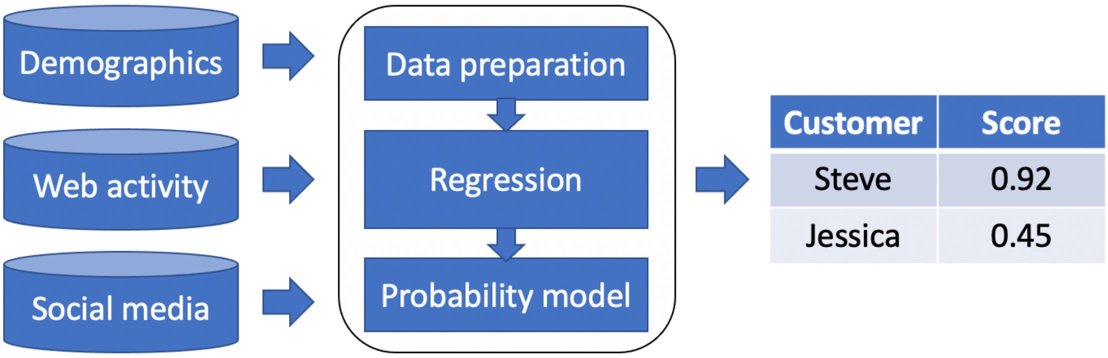


Figure 7. Finding High-Propensity prospects ML pipeline.

As we are looking for a numeric value we can consider this as a regression problem. The details on regression and supervised learning are given earlier in this thesis. As a result, a marketing expert gets a propensity score for every potential customer. All customers can be arranged in descending order to identify the ones with top scores. This information will be used for further marketing actions, such as special offers, or phone calls.

4.1.2 Identifying the best communication channel with a prospect

Once a list of top prospects was received, the next step is to identify the best channel to communicate with a prospect. With so many different mediums available, it is important to target customers in such a way that will receive the most attention and the highest return of investment. As a result of this step, we will get a Table 4 with prospects and a communication channel for each of them:

Table 4. Prospects and communication channels.

Prospect	Channel
Steve	Mobile
Cindy	email

What data do we need to succeed? We will again utilize demographic data, familiar from a previous step, and additionally, data about past successful events.

Table 5. Past Success Events.

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

In most cases, we are utilizing the same or similar ML pipelines, changing ML algorithm.

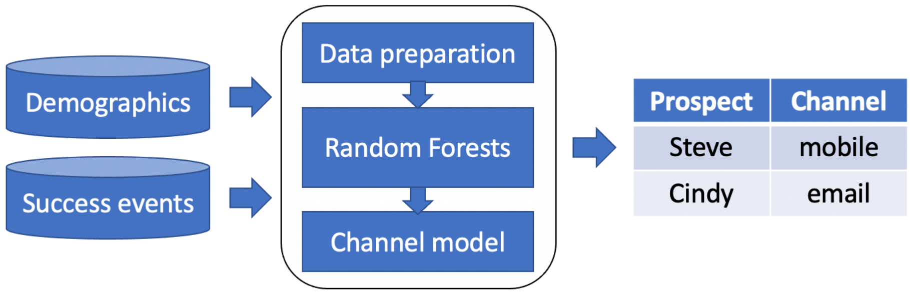


Figure 8. Finding best communication channel ML pipeline.

In this step of marketing analysis, we encountered a classification problem. To classify customers by different types of communication channels we utilize Random Forest as an algorithm. In the previous character, we have already reviewed different classification algorithms, including decision trees. Let us give a clear description of what is a Random Forest algorithm. "Random forest is a flexible, easy to use ML algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because of its simplicity and diversity (it can be used for both classification and regression tasks)." [29] Random Forest is just one example of classification algorithms that can be used in this case. It is considered good practice to try out different ML algorithms to define the best for the particular situation. From this point, we acquired enough information to perform a targeted marketing campaign. In these two examples, we used ML pipelines to increase customer analysis performance and to reduce the amount of resources involved in the process.

4.2 ML for predicting customer lifetime value (CLV)

Let us start off with a definition of CLV. "The lifetime value of a customer, or customer lifetime value (CLV), represents the total amount of money a customer is expected to spend in the business, or on products, during their lifetime." [30] There are multiple ways to calculate CLV. In the scope of this example, the formula itself does not affect the result. The only important aspect when it comes to CLV is the consistency of the formula throughout the whole dataset.

The goal of this example is to build a regression model that can predict the CLV for a new customer, based on his or her recent buying patterns and historical data from the other customers.

The Table 6 shows an example of a data record used for an ML prediction model:

Table 6. Monthly Sales.

Monthly Sales	
Name	Steve
1st Month	$3000
2nd Month	$0
3rd Month	$2000
CLV	$5000

In this case, we have a regression problem, and Linear Regression is one of the algorithms to go. We discuss this algorithm and some other algorithms for supervised learning earlier in the thesis.

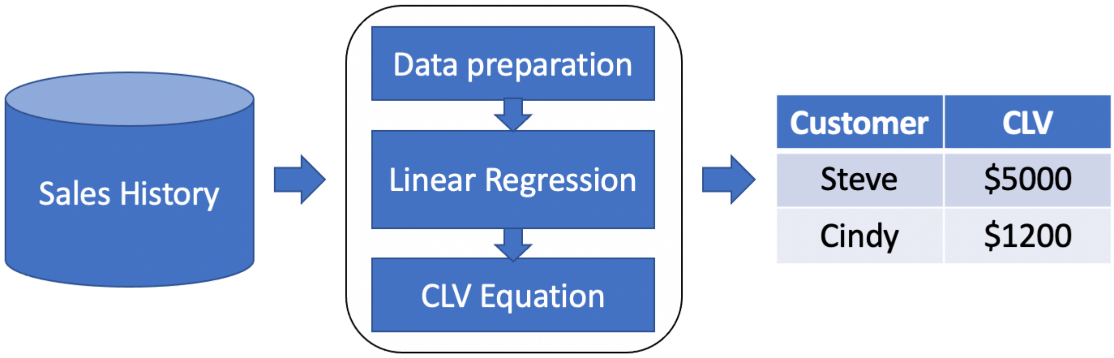


Figure 9. Predicting CLV ML pipeline.

The more data with the growing number of customers we get, the more accurate the resulting prediction is. With the change of existing customers' data, the CLV can be recalculated and the model retrained. Predictions of this model can be utilized in further marketing analysis to change the focus of a marketing campaign from one customer to another.

4.3 ML for predicting customers who might leave

When it comes to customer attrition, the best approach a business can take is to correctly identify customers who might leave and take preventative action to keep them. The goal of this case is to identify customers who might switch to competitors. The two possible ways to approach this problem are either to classify the customers (at risk / not at risk) or give each customer a risk score. After processing all the steps through the ML pipeline we will get a following result showed in Table 7:

Table 7. Customer attrition.

Customer	Risk
Steve	10%
Cindy	20%
Bob	81%

What data do we need to succeed? We will again utilize demographic data discussed in the previous chapters. The second data set is customer history records collected from the customer activity. For each customer, there will be one history record with a summary of different types of information. An example of such record in Table 8:

Table 8. Historical data.

History	
Tenure	2 years
Total value	$1200
Last Purchase	16.03.2020
Support Calls	6
Returns	1
Left?	Y

For this particular case, we use Naive Bayes as a classification algorithm to calculate probabilities based on historical data. A naive Bayes classifier is an algorithm that uses Bayes' theorem to classify objects. Bayes' theorem "is a mathematical formula for determining conditional probability. Conditional probability is the likelihood of an outcome occurring, based on a previous outcome occurring." [31]

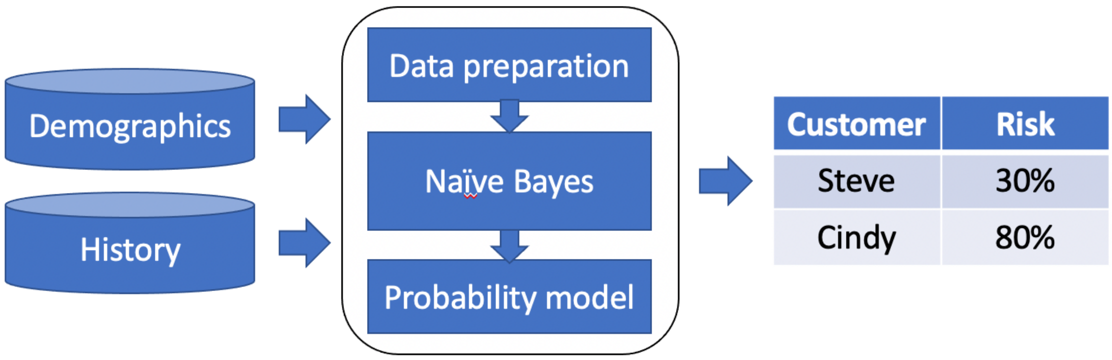


Figure 10. Customer attrition ML pipeline.

Based on this information marketing experts can target particular clients to reduce customer attrition and optimize the marketing budget.

4.4 How Sentiment Analysis revolutionized marketing

The idea behind sentiment analysis is some kind of social listening. Natural language processing (NLP), text analysis, computational linguistics, and even biometrics are involved in sentiment analysis. Manual processing of reviews and product feedbacks requires a tremendous amount of financial and human resources. The goal is to scale out this labor-intensive procedure, to be able to process millions of human written words in no time. As a result, we get a pipeline with the same accuracy as humans, but much faster and cheaper. The most common application of sentiment analysis in digital marketing is polarity assessment. In simple words, polarity assessment allows classifying positive, neutral, or negative feedbacks and comments. We use a pipeline for this classification problem as follows in Figure 11:

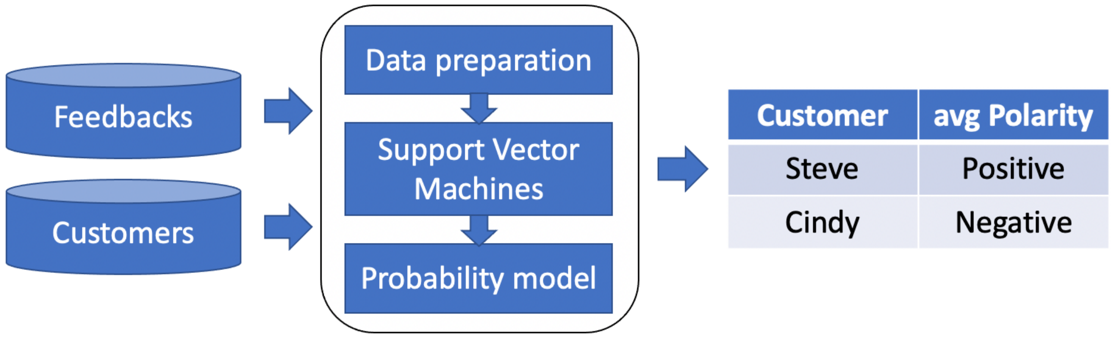


Figure 11. Sentiment Analysis ML pipeline.

However, in the case of Sentiment Analysis, data needs some special treatment. It is known that in ML algorithms only a numerical type of data can be used. If the initial data set consists of text data it is necessary to convert it into numerical data. Such conversion includes tokenization, stemming, punctuation exclusion, and bag-of-words. A Figure 12 showcases the process of data preparation for sentiment analysis:

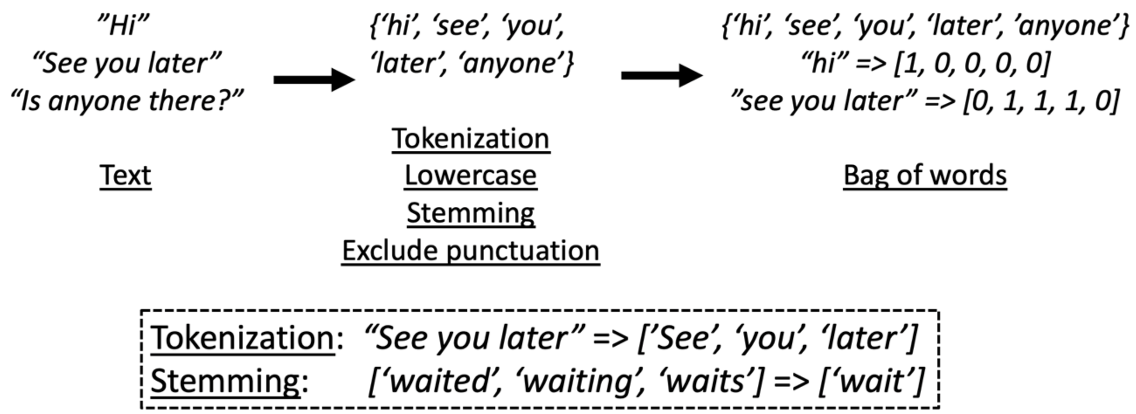


Figure 12. Data preparation for sentiment analysis.

In the final step, data gets the numerical form. The author of this thesis shows only one approach to converting data into numeric form out of many.

There are a few reasons for a marketing expert to do Sentiment Analysis. Number one is implementing a targeted marketing campaign. It is important to find people who are positive about the product but have not purchased it yet. The other reason is to find problems and complaints before they become major problems. In any case, utilizing Sentiment Analysis allows marketers to be proactive in communication with clients.

5 ML FOR CUSTOMER BEHAVIOR ANALYSIS: PREDICTING CLV USE CASE

In this chapter the author implements and reviews a CLV prediction ML program based on Python programming language, some extra libraries, and a dataset. The purpose is to show a use case implementation discussed in chapter 4.2, which can be further used by any digital marketing specialist. Most theoretical topics about ML and CLV are discussed in the earlier chapters, the author focuses on the practicalities in this part of the thesis.

5.1 Libraries setup and data preparation

The author uses a well-known data science distributive Anaconda. Anaconda includes Python 3.9.4 as the latest stable release and Jupiter Notebook Integrated Development Environment (IDE). All the information regarding preinstallation can be found on the corresponding websites. Let us start by importing libraries into the project using "import":

```
from pandas import Series, DataFrame
import pandas as pd
import numpy as np
import os
import matplotlib.pylab as plt
from sklearn.model_selection  import train_test_split
from sklearn.linear_model import LinearRegression
import sklearn.metrics

raw_data = pd.read_csv("history.csv")
```

Figure 13. Libraries setup.

The following libraries are used in the project:

- Pandas - library for data manipulation and structuring
- NumPy - library containing extra mathematical and statistical formulas, multi-dimensional array compatibility.
- os - operating system integration, files manipulations
- matplotlib - plotting library for graphical representation of data

- sklearn - free software ML library for python, featuring various regression, classification and clustering algorithms

Let us inspect the dataset using a function 'head()'

```
raw_data.head()
```

Figure 14. Function for dataset inspection.

As a result, the table showed in the Figure 15:

	CUST_ID	MONTH_1	MONTH_2	MONTH_3	MONTH_4	MONTH_5	MONTH_6	CLV
0	1001	150	75	200	100	175	75	13125
1	1002	25	50	150	200	175	200	9375
2	1003	75	150	0	25	75	25	5156
3	1004	200	200	25	100	75	150	11756
4	1005	200	200	125	75	175	200	15525

Figure 15. Function for dataset inspection output.

The table contains the first 5 lines of the data set utilized for this model. The first column CUST_ID contains a unique number of each customer in a data set. The following columns MONTH_1, MONTH_2, and the following columns contain the revenue for each month for every customer. The last column CLV contains the calculated CLV for each customer based on the purchasing history for the last 3 years. The total number of observations is 100. For demonstration purposes and to keep the data preparation part of the code compact in the scope of this model the author uses dummy data generated for this particular use case.

5.2 Correlation analysis

The next step is correlation analysis. Correlation analysis allows to identify features for a future ML algorithm. It is vital to distinguish and use features with strong correlations. More details on feature selection were given earlier in this thesis.

For correlation analysis, the following functions have been utilized:

```
cleaned_data = raw_data.drop("CUST_ID",axis=1)
cleaned_data .corr()['CLV']
```

Figure 16. Correlation analysis.

As a result, we received correlations presented in Figure 17:

```
MONTH_1     0.734122
MONTH_2     0.250397
MONTH_3     0.371742
MONTH_4     0.297408
MONTH_5     0.376775
MONTH_6     0.327064
CLV         1.000000
```

Figure 17. Correlation analysis output.

For each feature, a sufficient correlation to the target variable (CLV) can be observed. This means that all the features from this data set, except for customer ID, can be used for a prediction model.

5.3 Data split

Data split considers splitting data into training and testing datasets. Due to the small size of a data set data was split with the ratio 90:10 and without a validation dataset. The following code in Figure 18 does the job of splitting data:

```
predictors = cleaned_data.drop("CLV",axis=1)
targets = cleaned_data.CLV

pred_train, pred_test, tar_train, tar_test  =   train_test_split(predictors, targets, test_size=.1)
print( "Predictor - Training : ", pred_train.shape, "Predictor - Testing : ", pred_test.shape )
```

Figure 18. Splitting data.

At this point, data is ready to be loaded into an ML algorithm.

5.4 Build and test model

Now we are ready to build a model with an ML algorithm and use the data we prepared in previous steps.

```
#Build model on training data
model = LinearRegression()
model.fit(pred_train,tar_train)
print("Coefficients: \n", model.coef_)
print("Intercept:", model.intercept_)

#Test on testing data
predictions = model.predict(pred_test)
predictions

sklearn.metrics.r2_score(tar_test, predictions)
```

Figure 19. Build and test the model.

In the scope of this use case, a simple Linear Regression algorithm has been utilized. Due to the large variety of available algorithms, it is a good practice to start from the simplest (less computationally demanding) algorithms and then try out the others. It is important to measure the accuracy of the model on each step using a test dataset.

As a result, we received an accuracy of 0.8779757671388931, which equals to approx 88%. This a significant accuracy for such a small dataset. The model is ready to predict CLV for new incoming customers.

5.5 Predicting CLV for a new customer

The main purpose of any ML model is the ability to work with new data to make predictions. This is the reason why we created this model. Let us imagine that, there is a new customer, who has been purchasing products in our company for the last 3 months. The revenue month-by-month is 100, 0 and 50 euros respectively. In this case, we only have data for 3 months available. This situation is common if the client is new and does not have a long history of purchases. The code in Figure 20 makes a prediction with the mentioned parameters:

```
new_data = np.array([100,0,50,0,0,0]).reshape(1, -1)
new_pred=model.predict(new_data)
print("The CLV for the new customer is : $",new_pred[0])
```

Figure 20. Predicting for a new customer.

The output is in Figure 21:

```
The CLV for the new customer is : $ 4034.6871082013886
```

Figure 21. Prediction output.

By rounding up we get 4035 euros. By receiving more data from the client after some period of time it is possible to recalculate an estimated CLV. The more data we have from the client the more accurate the prediction will be.

In this chapter, we built a simple model based on a linear regression algorithm to simulate a real case scenario. The flexibility of Python and its libraries allows building ML prediction models in a short period of time. The code is flexible and can be modified for every scenario. External libraries extend the functionality and reduce the amount of code that needs to be written. The results of this model's prediction can be used by a marketing expert for a deeper analysis of customers' behavior.

6 CONCLUSION

In the scope of this thesis, various topics on ML for customer behavior analysis have been discussed. In the first two chapters, the most important theoretical aspects of ML and digital marketing were highlighted. Chapters 4 and 5 discuss more practicalities and show the example use case implementation in code. Since this thesis is targeted primarily at both IT students and marketing students, it introduces the vital marketing and ML topics. A marketing expert can also benefit from this thesis. With gaining first hands-on experience with ML, a reader sees a working prototype and a stimulus to continue learning about ML and digital marketing. This thesis also aimed to show how ML technologies evolved in time and have become accessible to more people. An individual does not have to be an expert in ML to gain the benefits of this powerful technology.

The high model accuracy (88%) mentioned in Chapter 5 indicates the validity and reliability of the conducted research.

There is still much that can be learned about ML and digital marketing. This thesis has brought up numerous topics to explore in further research some of which are the following:

1) Chatbots in digital marketing
2) Cost efficiency of implementing ML model in marketing
3) Future of digital marketing and ML
4) No code, low code solutions for ML
5) Data privacy problems raised by ML development
6) Transformation of marketing industry influenced by ML

REFERENCES

[1] Big Data Made Simple. 2021. *5 ways artificial intelligence is enhancing traditional marketing today*. [online] Available at: https://bigdata-madesimple.com/5-ways-artificial-intelligence-is-enhancing-traditional-marketing/ [Accessed 2 March 2021].

[2] Businessoverbroadway.com. 2021. [online] Available at: https://businessoverbroadway.com/2021/02/01/machine-learning-adoption-rates-around-the-world/ [Accessed 19 April 2021].

[3] Hao, K., 2021. *What is AI? We drew you a flowchart to work it out*. [online] MIT Technology Review. Available at: https://www.technologyreview.com/2018/11/10/139137/is-this-ai-we-drew-you-a-flowchart-to-work-it-out/ [Accessed 4 March 2021].

[4] Businessoverbroadway.com. 2021. [online] Available at: https://businessoverbroadway.com/2021/02/01/machine-learning-adoption-rates-around-the-world/ [Accessed 19 April 2021].

[5] Google Developers. 2021. *Machine Learning Glossary | Google Developers*. [online] Available at: https://developers.google.com/machine-learning/glossary#m [Accessed 12 March 2021].

[6] Doc.ic.ac.uk. 2021. *History of Machine Learning*. [online] Available at: https://www.doc.ic.ac.uk/~jce317/history-machine-learning.html [Accessed 2 May 2021].

[7] Dictionary.cambridge.org. 2021. *data*. [online] Available at: https://dictionary.cambridge.org/dictionary/english/data [Accessed 12 March 2021].

[8] Oracle.com. 2021. *What Is Big Data? | Oracle*. [online] Available at: https://www.oracle.com/big-data/what-is-big-data/ [Accessed 11 March 2021].

[9] dummies. 2021. *The 4 V's of Big Data - dummies*. [online] Available at: https://www.dummies.com/careers/find-a-job/the-4-vs-of-big-data/ [Accessed 10 April 2021].

[10] DATAVERSITY. 2021. *The Problem with Big Data: It's Getting Bigger - DATAVERSITY*. [online] Available at: https://www.dataversity.net/the-problem-with-big-data-its-getting-bigger/# [Accessed 2 March 2021].

[11] WhatIs.com. 2021. *What is machine learning algorithm? - Definition from WhatIs.com*. [online] Available at: https://whatis.techtarget.com/definition/machine-learning-algorithm [Accessed 12 March 2021].

[12] Sas.com. 2021. *A guide to the types of machine learning algorithms*. [online] Available at: https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html#:~:text=As%20new%20data%20is%20fed,%2Dsupervised%2C%20unsupervised%20and%20reinforcement [Accessed 2 March 2021].

[13] Geron, A., 2020. *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow*. O'Reilly.

[14] Osiński, B. and Budek, K., 2021. *What is reinforcement learning? The complete guide - deepsense.ai*. [online] deepsense.ai. Available at: https://deepsense.ai/what-is-reinforcement-learning-the-complete-guide/ [Accessed 11 March 2021].

[15] Medium. 2021. *Data Quality for Everyday Analysis*. [online] Available at: https://towardsdatascience.com/data-quality-for-everyday-analysis-d3aa1442c31 [Accessed 2 March 2021].

[16] Medium. 2021. *How To Prepare Your Data for Your Machine Learning Model*. [online] Available at: https://towardsdatascience.com/how-to-prepare-your-data-for-your-machine-learning-model-b4c9fd4e7ea [Accessed 6 March 2021].

[17] Medium. 2021. *The Art of Finding the Best Features for Machine Learning*. [online] Available at: https://towardsdatascience.com/the-art-of-finding-the-best-features-for-machine-learning-a9074e2ca60d [Accessed 6 April 2021].

[18] Medium. 2021. *A Brief Overview of Outlier Detection Techniques*. [online] Available at: https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561 [Accessed 2 April 2021].

[19] Docs.aws.amazon.com. 2021. *Training ML Models - Amazon Machine Learning*. [online] Available at: https://docs.aws.amazon.com/machine-learning/latest/dg/training-ml-models.html [Accessed 6 April 2021].

[20] Google Developers. 2021. *Descending into ML: Training and Loss | Machine Learning Crash Course*. [online] Available at: https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss [Accessed 2 April 2021].

[21] Medium. 2021. *About Train, Validation and Test Sets in Machine Learning*. [online] Available at: https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7 [Accessed 9 April 2021].

[22] SearchBusinessAnalytics. 2021. *What is data visualization and why is it important?*. [online] Available at: https://searchbusinessanalytics.techtarget.com/definition/data-visualization [Accessed 7 April 2021].

[23] Medium. 2021. *Visualising Machine Learning: How do we humanise the intelligence?*. [online] Available at: https://towardsdatascience.com/visualising-machine-learning-how-do-we-humanise-the-intelligence-e62658f1f6df [Accessed 2 May 2021].

[24] Dictionary.cambridge.org. 2021. *marketing*. [online] Available at: https://dictionary.cambridge.org/dictionary/english/marketing [Accessed 2 May 2021].

[25] Henneberry, R., 2021. *The Ultimate Guide to Digital Marketing | DigitalMarketer*. [online] Digitalmarketer.com. Available at: https://www.digitalmarketer.com/digital-marketing/ [Accessed 2 March 2021].

[26] Patel, N. 2021. *Neil Patel: Helping You Succeed Through Online Marketing!*. [online] Available at: https://neilpatel.com/ [Accessed 2 April 2021].

[27] (//www.bigwavemedia.co.uk), B., 2021. *3 Key Uses of Demographic Data - Bigwave media*. [online] Bigwavemedia.co.uk

[28] (//www.bigwavemedia.co.uk), B., 2021. *3 Key Uses of Demographic Data - Bigwave media*. [online] Bigwavemedia.co.uk. Available at: https://www.bigwavemedia.co.uk/blog/uses-of-demographic-data#:~:text=Demographic%20data%20is%20statistical%20data,areas%20it%20is%20most%20popular [Accessed 2 April 2021].

[29] Built In. 2021. *A complete guide to the random forest algorithm*. [online] Available at: https://builtin.com/data-science/random-forest-algorithm [Accessed 4 April 2021].

[30] Shopify. 2021. *Customer Lifetime Value (CLV) Definition - What is Customer Lifetime Value (CLV)*. [online] Available at: https://www.shopify.com/encyclopedia/customer-lifetime-value-clv [Accessed 2 May 2021].

[31] Investopedia. 2021. *Bayes' Theorem*. [online] Available at: https://www.investopedia.com/terms/b/bayes-theorem.asp [Accessed 2 March 2021].

[32] Revive Digital. 2021. *The 11 Types of Digital Marketing* | *Revive Digital*. [online] Available at: https://revive.digital/blog/the-11-types-of-digital-marketing/ [Accessed 2 April 2021].

https://www.bigwavemedia.co.uk/blog/uses-of-demographic-data#:~:text=Demographic%20data%20is%20statistical%20data,areas%20it%20is%20most%20popular [Accessed 2 April 2021].

[33] Data Science Society. 2021. *How Data Scientists Shape our Modern World – Data Science Society*. [online] Available at: https://www.datasciencesociety.net/how-data-scientists-shape-our-modern-world [Accessed 8 April 2021].