JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

ANALYSIS ON CUSTOMER BEHAVIOR USING MACHINE LEARNING

Ankush Pandey, 1st Year MCA, JECRC University, Rajasthan

Dr. Prashant Dixit, Asst. Professor, Mewar University, Rajasthan

Abstract: Machine Learning (ML) and Artificial Intelligence (AI) have become an important part of many businesses and organisations. Digital Marketing has also benefited from modern Machine Learning techniques. The objective of this analysis 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 behaviour. Due to the complexity of both Digital Marketing and Machine Learning, the analysis 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 behaviour 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 analysis 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.

Introduction: Every person is different and has habits and personal traits. Customer behaviour is not an exception. We make a decision to buy, or no to buy or we may put the things in wish list, 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 that who the customers are.

Machine Learning (ML) comes in suitable in this case. With the extension of digital platforms and the digitalization of business of organisations, traditional marketing methods themselves became inefficient. This does not mean ML rewrites the fundamentals of marketing and clients' behaviour 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.

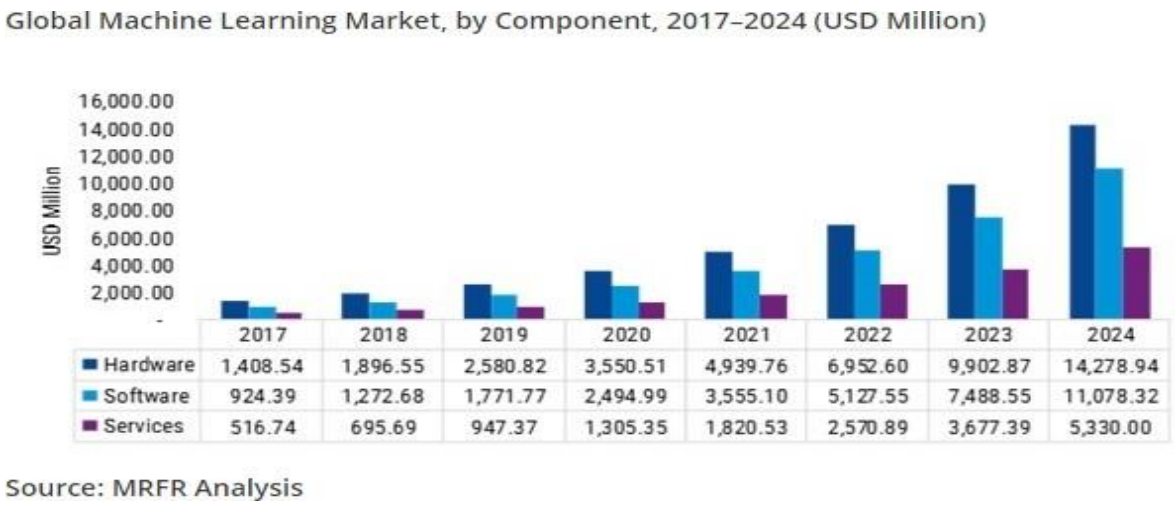

Figure 1: Global Machine Learning market

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.

The main objective of this analysis is to familiarize the viewer with the modern trends in Digital Marketing and ML, 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 analysis 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 analysis 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 following sub-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 behaviour 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 behaviour analysis?

Machine Learning (ML) and Artificial Intelligence

Machine learning (ML) is used continuously in different industries because its possibility to provide innovative solutions of an organization. [3]. In the broadest sense, AI refers to machines that can study, reason, and act for themselves. They can make their own result when faced with new situations, in the similar 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 Machine Learning. These algorithms use statistics to find patterns in massive amounts of data [4]

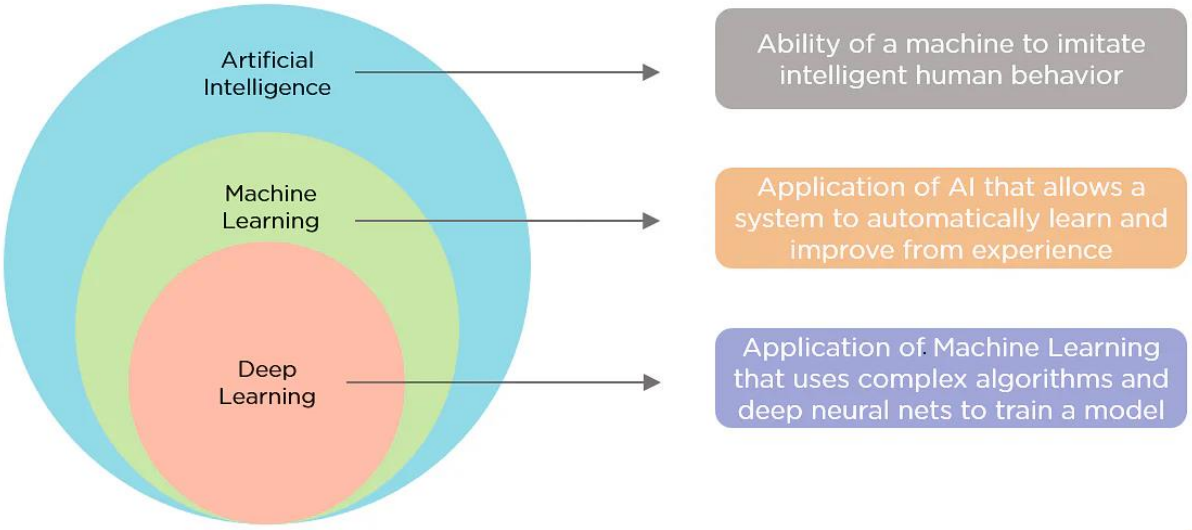

Figure 2: Structure of AI Source

Millions of the Netflix users use Machine Learning algorithms every day without even knowing that. ML algorithms are the main drivers of any recommendation machine on a website like Netflix, Hotstar, voice assistants such as Siri or Alexa, a search engine: Google or Yahoo [5].

It is clear from the picture that Machine Learning is a part (or one of the approaches) of Artificial Intelligence. Some viewer might be slightly confused at this point. The author of this study 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." [6]

Past of ML and AI:

Many people assume that Machine Learning 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 [7]:

Before 1950s - 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.

1950s - 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.

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

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

1980s - 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.

1990s - 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.

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

2010s - 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. TURKU UNIVERSITY OF APPLIED SCIENCES THESIS | Maksim Pokrovskii 10

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.

Data in Machine Learning:

Data is the fuel for any Machine Learning algorithm. Data brings information about users and their behaviour. 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 [8]. 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 scientists define four Vs of Big Data:
Volume - refers to the size of data sets that need to be processed and analysed

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 analysed.

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 labelled, categorized, and stored in a structured database. The majority of incoming data is unstructured and cannot 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 behaviour [10]

Algorithms of Machine Learning:

A ML model is defined as a computer-intensive mechanism and applies re-sampling and iterative methodologies for classification approaches. ML approaches are considered with optimal subset selection and eliminate the issues of classical classifiers like over-fitting as well as distributional demands of parameters. ML technologies that have emerged in computer science with logic and basic mathematics, statistics as ML approaches do not estimate the group features rather it is initialized with an arbitrary group separator and tunes frequently till satisfying the classification groups. ML examines the tuning variables and individual ML functions became unstable, which makes a suitable process. As the non-statistical nature is embedded, these approaches can apply the data in various formats like nominal data that generates maximum classification accuracies. [11]

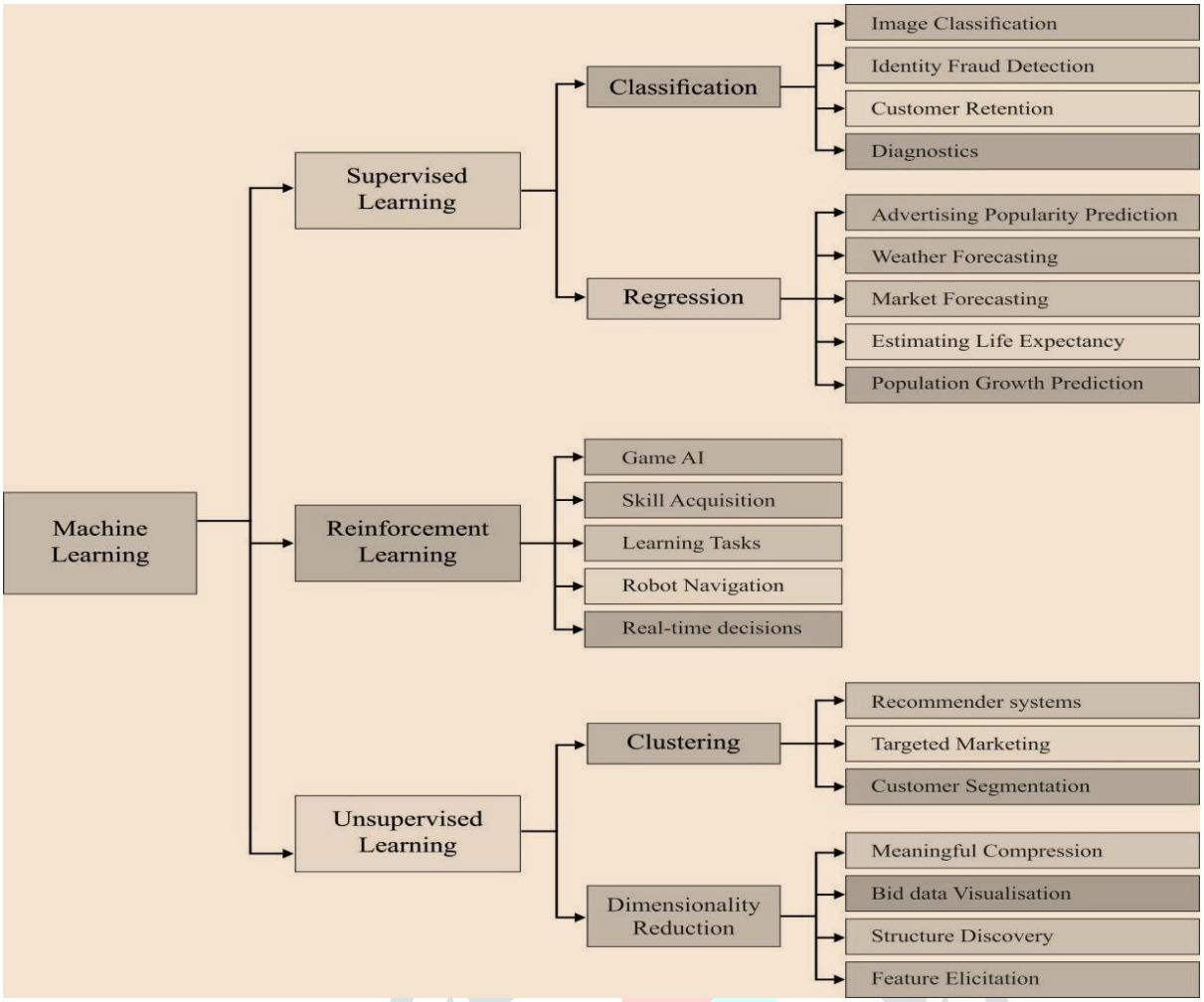

Figure 3: Summary of Machine Learning Algorithms (Adapted from Prashant Dixit, 2022) [12]

Marketing:

Marketing and Digital Marketing: According to Harvard dictionary Marketing - "A job that involves encouraging people to buy a product or service" [13]. 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. [14].

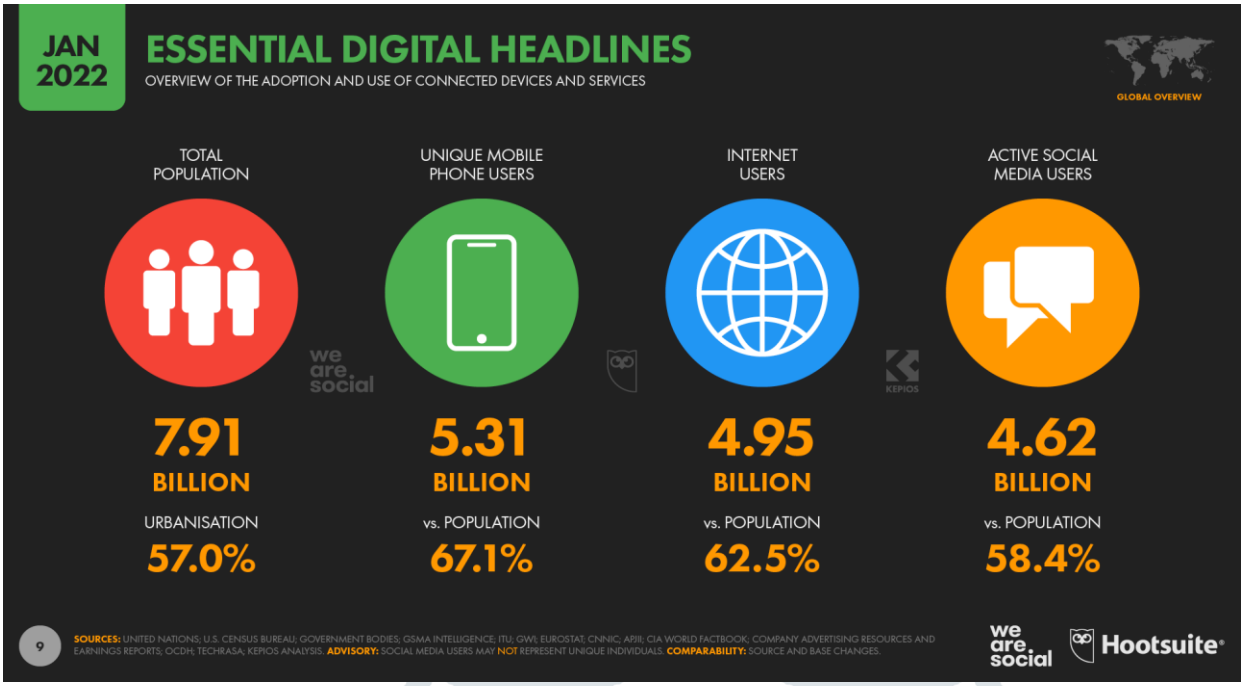

Figure 4: World Digitization Data Source

According to recent data, there are more than 5.31 billion mobile phones users and 4.95 billion internet users as of January 2022 (Fig.4). This chunk of customers cannot be ignored by companies and marketing experts.

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 marketingexperts Neil Patel [15].

And these categories are [16]:

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.

ML in Digital Marketing to Predict Customers' Behavior

Every business need customer 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.

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 [17]. 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**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." [18]

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 visit, 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.

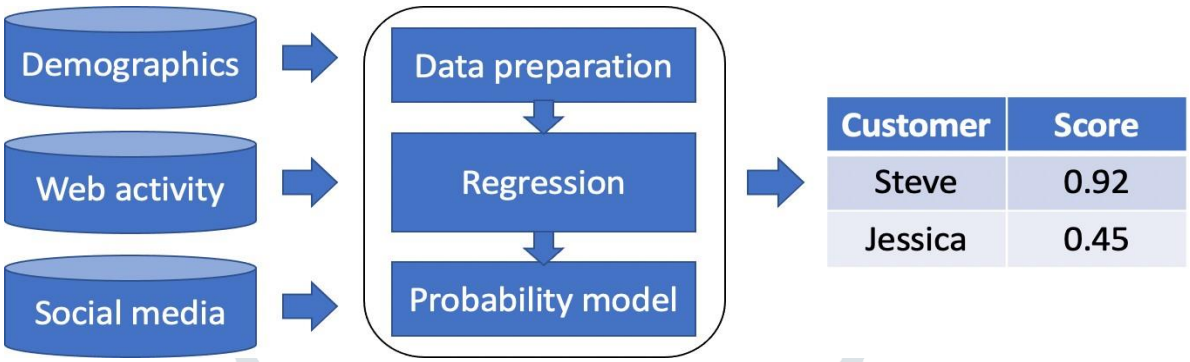

Figure 5: Finding High-Propensity prospects ML pipeline.

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.

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.

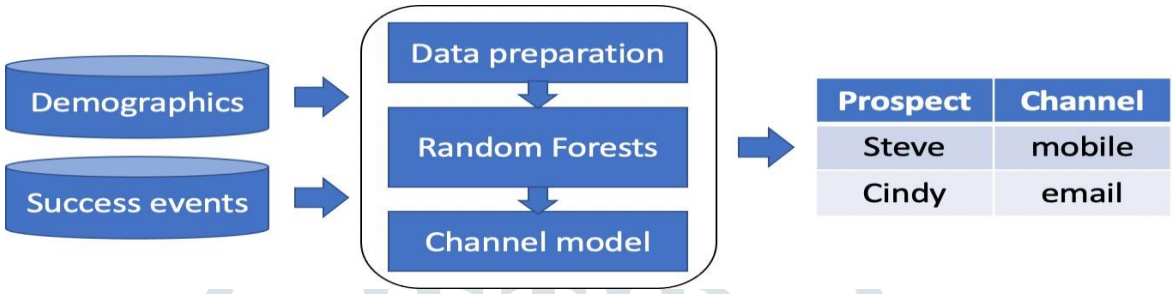

Figure 6: Finding best communication channel ML pipeline.

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)." [19] 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 number of resources involved in the process.

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." [20] 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.

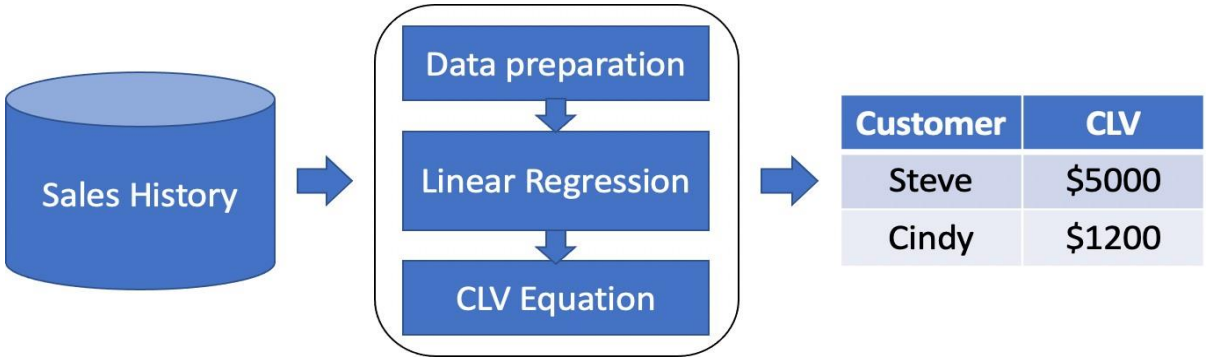

Figure 7: Predicting CLV ML pipeline.

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.

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." [21]

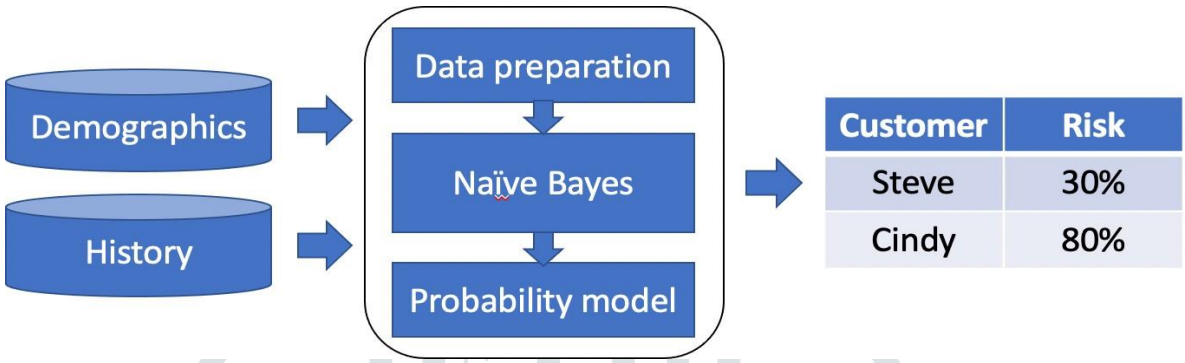

Figure 8: Customer attrition ML pipeline.

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

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 9:

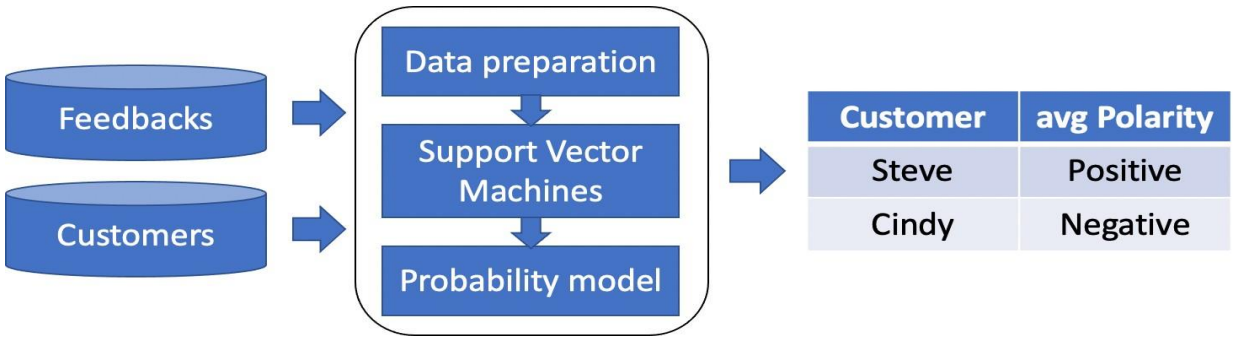

Figure 9: Sentiment Analysis ML pipeline.

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 8 showcases the process of data preparation for sentiment analysis:

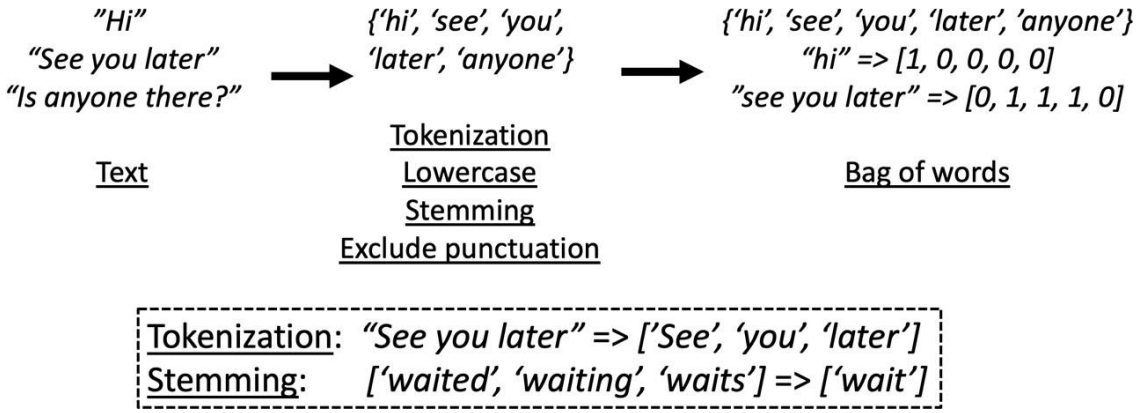

Figure 10: Data preparation for sentiment analysis.

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 client.

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.

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 11: 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
- klearn - 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 12: Function for dataset inspection.

As a result, the table showed in the Figure 13:

	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 13: 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.

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 14: Correlation analysis.

As a result, we received correlations presented in Figure 15:

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 15: 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.

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 16: Splitting data.

At this point, data is ready to be loaded into an ML algorithm. 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 17: 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.

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 18: Predicting for a new customer.

The output is in Figure 19:

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

Figure 19: Prediction output.

In this analysis, 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.

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-artificialintelligence-is-enhancing-traditional-marketing/ [Accessed 2 March 2021].

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

[3] Prashant Dixit, Dr. Harish Nagar, Dr. Sarvottam Dixit, "Student Performance Prediction Using Case Based Reasoning Knowledge Base System (CBR-KBS) Based Data Mining", International Journal of Information and Education Technology, Vol. 12, No. 1, January 2022

[4] 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-aflowchart-to-work-it-out/ [Accessed 4 March 2021].

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

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

[7] 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].

[8] Dictionary.cambridge.org. 2021. data. [online] Available at: https://dictionary.cambridge.org/dictionary/english/data [Accessed 12 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] Dr. Kumud, Dr. Prashant Dixit, Prof. Sarvottam Dixit, "Exploring Machine Learning in Higher Education: Prediction of Student Performance", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.9, Issue 10, page no. b387-b401, October-2022

[12] Prashant Dixit, Doctoral Thesis "Psychometric Analysis of Graduate Students using Machine Learning" (2022) http://hdl.handle.net/10603/409545

[13] 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].

[14] 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].

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

[16] 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].

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

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

[19] (//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%20 most%20popular [Accessed 2 April 2021].

[20] 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].

[21] hopify. 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].

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

[23] 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%20 most%20popular [Accessed 2 April 2021].

[24] 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].