

Faculty of Industrial Engineering and Management

Master thesis

To obtain the academic degree
Master of Engineering (M. Eng.)

Development of a lead scoring model for simple and advanced use cases using the Mautic software as an example

Company: Leuchtfeuer Digital Marketing GmbH
Immengarten 16-18
30177 Hanover

First examiner: Prof. Dr. Ing. Peter Kraus

Second examiner: Prof. Dr. Robert Kuttler

Submitted on: 30.03.2024

Created by: Jonas Ludwig

Explanation

I hereby declare that I have written this thesis independently, have not used any sources and aids other than those stated and have marked literal and analogous quotations as such. This thesis has not been submitted to any other body for a similar purpose. This declaration also applies to graphical representations and to any software included or used as a basis.

Place, date

Jonas Ludwig

Abstract

Digital marketing gives companies the ability to generate leads online and store them in their databases. Since not all of these leads necessarily show a clear interest in making a purchase, it is necessary to prioritize them so that only qualified leads are allocated further resources. One method to do this is lead scoring. Traditionally, each lead is assigned a score based on their characteristics and behavior. If the score exceeds a certain threshold, the contact is considered qualified and forwarded to the sales team. In addition to the traditional approach, there is also the predictive approach, which utilizes machine learning algorithms to qualify leads.

The objective of this work is to develop a lead scoring model for both simple and advanced use cases, using the software Mautic. To achieve this, a generic model for developing both traditional and predictive lead scoring systems is created and then adapted to the Mautic software. In addition, the implementation of advanced use cases such as account-based scoring and product-based lead scoring in Mautic will be examined.

The research results indicate that the functionalities of Mautic are sufficient to implement both the traditional and predictive lead scoring approaches. The implementation of product-based lead scoring is also feasible. However, the current capabilities of Mautic are not sufficient to create an account-based scoring system. In addition, a data-driven lead scoring system requires the use of external data analysis software. Real-world tests reveal that the predictive lead scoring approach enables more accurate predictions about the quality of a lead than the traditional approach.

Table of contents

Explanation	II
Abstract.....	III
Table of contents	IV
List of Figures	VI
List of tables.....	VIII
List of abbreviations.....	IX
1 Introduction	1
1.1 Problem definition and objective.....	1
1.2 Methodology.....	2
2 Basics	3
2.1 Lead.....	3
2.2 Customer Journey	3
2.3 Sales Funnel.....	3
2.4 Lead management.....	5
2.5 Marketing Automation.....	5
3 Classification of the term lead scoring.....	8
3.1 Traditional lead scoring.....	8
3.2 Goals of lead scoring	12
3.3 Challenges in lead scoring	13
4 Advanced approaches and use cases of lead scoring	20
4.1 Predictive lead scoring	20
4.2 Product-based scoring.....	27
4.3 Account Based Scoring	28
4.4 Lead scoring without a sales team	29
4.5 Further use cases	30
5 Alternative and complementary methods to lead scoring.....	30
5.1 RFM analysis.....	30
5.2 Product recommendation systems	31
6 Development of a generic process model for lead scoring.....	34
6.1 Traditional lead scoring.....	35
6.2 Predictive lead scoring	41
7 Application of the process model to the Mautic software.....	42

7.1	Introduction of the Mautic software.....	42
7.2	Traditional lead scoring in Mautic.....	45
7.3	Predictive lead scoring in Mautic	55
7.4	Product-based lead scoring in Mautic.....	58
7.5	Account Based Scoring in Mautic.....	58
8	Practical implementation and results of lead scoring in Mautic.....	59
8.1	Traditional lead scoring.....	59
8.2	Predictive lead scoring.....	69
9	Summary and outlook.....	71
	Bibliography.....	X
	Appendix.....	XVII
	Appendix 1: Campaign to generate sales feedback.....	XVII
	Appendix 2: Python code for preparing the lead data.....	XX
	Appendix 3: SQL statements for downloading the data of converted leads from Mautic	XXII
	Appendix 4: SQL statements for downloading the data of non-converted leads from Mautic.....	XXIII
	Appendix 5: Python code for analyzing the lead data	XXIV
	Appendix 6: Python code for determining and adjusting the threshold value in lead scoring.....	XXVI
	Appendix 7: Results of predictive lead scoring with the support vector machine algorithm.....	XXVIII
	Appendix 8: Results of predictive lead scoring with the decision tree algorithm	XXIX
	Appendix 9: Results of predictive lead scoring with the logistic regression algorithm	XXX
	Appendix 10: Python code to develop a machine learning model for lead scoring	XXXI
	Appendix 11: Python code for applying the machine learning model.....	XXXIII

List of Figures

Figure 1: Traditional sales funnel.....	4
Figure 2: Lead scoring matrix	12
Figure 3: Content-based product recommendation systems	32
Figure 4: Collaborative filter systems.....	33
Figure 5: Differences between the traditional and the predictive process model for creating a lead scoring system	35
Figure 6: User interface of the Mautic software	43
Figure 7: Point Groups in Mautic	47
Figure 8: Point Actions in Mautic	48
Figure 9: Campaign action "Adjust contact points"	49
Figure 10: Campaign for calculating the explicit score	50
Figure 11: Campaign to implement a forfeiture model.....	51
Figure 12: Campaign to reset negative implicit scores	51
Figure 13: Campaign for the transfer of leads	53
Figure 14: Campaign for feedback evaluation	54
Figure 15: Campaign to reactivate leads	55
Figure 16: Importing the predictions of a predictive lead scoring model into Mautic	57
Figure 17: Campaign for calculating the implicit score in account-based scoring	59
Figure 18: Correlation analysis with the lead status	64
Figure 19: Comparison of the threshold values of converted and non-converted leads.....	66
Figure 20: Confusion matrix of the preliminary traditional lead scoring system.....	67
Figure 21: Confusion matrix of the final traditional lead scoring system.....	68
Figure 22: Confusion matrix for predictive lead scoring with the random forest algorithm	70
Figure 23: Campaign to generate sales feedback	XVII
Figure 24: Email for generating sales feedback.....	XVIII
Figure 25: JavaScript code for sending the sales feedback	XIX
Figure 26: Python code for preparing the lead data.....	XXI
Figure 27: SQL statement for downloading the data of converted leads from Mautic.....	XXII
Figure 28: SQL statement for downloading the data of non-converted leads from Mautic	XXIII
Figure 29: Python code for analyzing the lead data.....	XXV
Figure 30: Python code for determining and adjusting the threshold value in lead scoring	XXVII
Figure 31: Results of predictive lead scoring with the support vector machine algorithm	XXVIII
Figure 32: Results of predictive lead scoring with the decision tree algorithm	XXIX
Figure 33: Results of predictive lead scoring with the logistic regression algorithm	XXX
Figure 34: Python code for developing a machine learning model for lead scoring	XXXII



Figure 35: Python code for the application of the machine learning model XXXIII

List of tables

Table 1: Relationship between the components of marketing automation software and lead management.....	7
Table 2: Explicit lead scoring parameters.....	8
Table 3: Positive and negative implicit lead scoring parameters.....	10
Table 4: Implicit scorecard in lead scoring.....	11
Table 5: KPIs in lead scoring.....	19
Table 6: Differences in traditional and predictive lead scoring.....	21
Table 7: Confusion matrix.....	27
Table 8: Differences in the B2C and B2B markets.....	29
Table 9: RFM table for awarding points per purchase.....	31
Table 10: Calculation of the RFM score.....	31
Table 11: DataFrame of converted leads.....	60
Table 12: Lead overview DataFrame.....	62
Table 13: Statistics table for data analysis.....	63
Table 14: Points system in the traditional lead scoring practice example.....	65
Table 15: Table for analyzing the score percentiles.....	66

List of abbreviations

ABS	Account Based Scoring
B2B	Business-to-Business
B2C	Business-to-Consumer
CRISP-DM	Cross-Industry Standard Process for Data Mining
CRM	Customer Relationship Management
CTA	Call-to-action
DF	DataFrame
ERP	Enterprise Resource Planning
FN	False Negative
FP	False positive
KPI	Key Performance Indicator
MAS	Marketing automation software
MQL	Marketing Qualified Lead
SAL	Sales Accepted Lead
SLA	Service Level Agreement
TN	True Negative
TP	True Positive

1 Introduction

1.1 Problem definition and objective

In the context of digital marketing, companies have the opportunity to generate leads online and store them in their databases. However, not all of these leads show a concrete interest in buying. Consequently, as part of the efficient use of sales resources, leads must first be qualified so that only high-quality leads can be processed further, for example by handing them over to the sales team (cf. Duncan and Eklan 2015: 1752). One way of doing this is lead scoring. In the traditional lead scoring approach, each lead receives a score. Points are added to or subtracted from this score based on the characteristics and behavior of the lead. If the score of a contact exceeds a predefined threshold value, it is considered qualified and is passed on to the sales team or processed with special marketing measures for qualified leads (cf. Schüller and Schuster 2022: 162-168). In addition to the traditional lead scoring approach, which uses point systems, the predictive approach also exists in practice. This uses machine learning algorithms and data mining models to qualify leads for forwarding to the sales department (cf. Wu et al. 2023: 1).

At the time of writing, there are several academic research papers available that examine the practical application of traditional lead scoring (cf. Monat 2011: 178-194; Naveen and Hariharanath 2021: 1302-1309; Verma et al. 2016: 220-237) and various machine learning algorithms in the context of predictive lead scoring (cf. Duncan and Eklan 2015: 1751-1758; Buckinx and van den Poel 2005: 252-268; Bohanec et al. 2015: 338-352; D'Haen et al. 2016: 69-78; Espadinha-Cruz et al. 2021: 1-14; Bohanec et al. 2017: 416-428; D'Haen and van den Poel 2013: 544-551; Kazemi and Babaei 2011: 37-45; Gokhale and Joshi 2018: 279-291; Jadli et al. 2022: 433-443; Nygard and Mezei 2020: 1439-1448; Kim and Street 2004: 215-228). Some of these papers, including Bohanec et al. (cf. 2017: 416-428) Gokhale and Joshi (cf. 2018: 279-291), Jadli et al. (cf. 2022: 433-443) and Nygard and Mezei (cf. 2020: 1439-1448) deal, among other things, with the choice of the best machine learning algorithm in predictive lead scoring. Wu et al. (cf. 2023: 1-30) also examine the current state of lead scoring and its influence on sales results. To this end, the results of 44 research papers in which different traditional and predictive lead scoring systems are developed are compared and the results of the individual papers are evaluated.

Although the available research suggests that predictive lead scoring leads to better results than traditional lead scoring, at the time of writing there is no research comparing the results of both approaches using the same practical example. This fact makes it difficult to derive insights into the differences in the performance of the two approaches. Furthermore, there is no research to date that examines the implementation of the different approaches and use

cases of lead scoring using the example of the Mautic software. In the current situation, there is therefore a lack of a methodologically sound guide for users to establish an effective lead scoring system. Due to a lack of research on lead scoring in Mautic, it is also unclear to what extent the software's functionalities are sufficient to enable effective lead scoring.

Given the identified research gaps, this research aims to develop a lead scoring model for simple and advanced use cases using the Mautic software as an example. This model will be tested on a practical example to validate its performance. This will enable users of the software to align their lead scoring system with a tried-and-tested guideline. In addition, the research gap is closed because the results of traditional and predictive lead scoring have not yet been compared in scientific research using the same application example. A further aim of this thesis is to evaluate the functionalities of Mautic in the area of lead scoring. Based on this evaluation, recommendations are made for the further development of the software in order to support the lead scoring process more effectively and to increase the value of the software from the user's perspective.

1.2 Methodology

At the beginning of this research paper, a comprehensive literature review is conducted to specify terms associated with lead scoring. The traditional approach to lead scoring is then discussed, along with its objectives and challenges. To conclude the theoretical part, extended approaches and use cases of lead scoring are described and alternative or complementary methods to lead scoring are considered.

In the practical part, generic process models for the development of robust traditional and predictive lead scoring systems are created on the basis of the previously identified challenges. The Mautic software is then presented and the previously developed process models are applied to this software. In addition, extended use cases of product-based and account-based lead scoring in Mautic are examined. Subsequently, traditional and predictive lead scoring are implemented in Mautic using a real practical example. Finally, the results of the practical study are evaluated, interpreted and compared with the findings of the literature research. The focus here is particularly on the evaluation of the individual use cases and the differences in the results of the traditional and predictive lead scoring approach. Furthermore, possibilities for the targeted improvement of the Mautic software are developed in order to enable a more targeted implementation of lead scoring.

2 Basics

2.1 Lead

In order to explain the term "lead scoring" appropriately in the marketing context, a precise definition of the term "lead" is necessary. In this paper, we refer to the definition by Todor (cf. 2016: 90) is referred to. He defines leads as potential prospects or existing customers who show interest in a newly offered product or service.

2.2 Customer Journey

Leads go through what is known as the customer journey on their way to making a purchase and beyond. The customer journey encompasses all of a lead's interactions with a company across all touchpoints. The touchpoints are interfaces that enable leads to make contact with a company. The fundamental goal within the customer journey is to provide the prospect or customer with the right information at the right time and in the desired form, thereby supporting the purchasing process. The strategic design of the customer journey is therefore essential in order to generate more sales (cf. Hannig 2021: 246; Koerner 2021: 67-69).

2.3 Sales Funnel

A sales funnel is used to visualize the qualification process of a lead within the customer journey (cf. Hannig 2021: 247). In this work, a traditional sales funnel model is used, which is shown in Figure 1 and divides leads into four qualification stages: Lead, Marketing Qualified Lead (MQL), Sales Accepted Lead (SAL) and Sales Qualified Lead (SQL). Once a lead has passed through these four stages, the aim is to reach a conclusion or "close". The individual phases of the sales funnel are explained in the following sections.

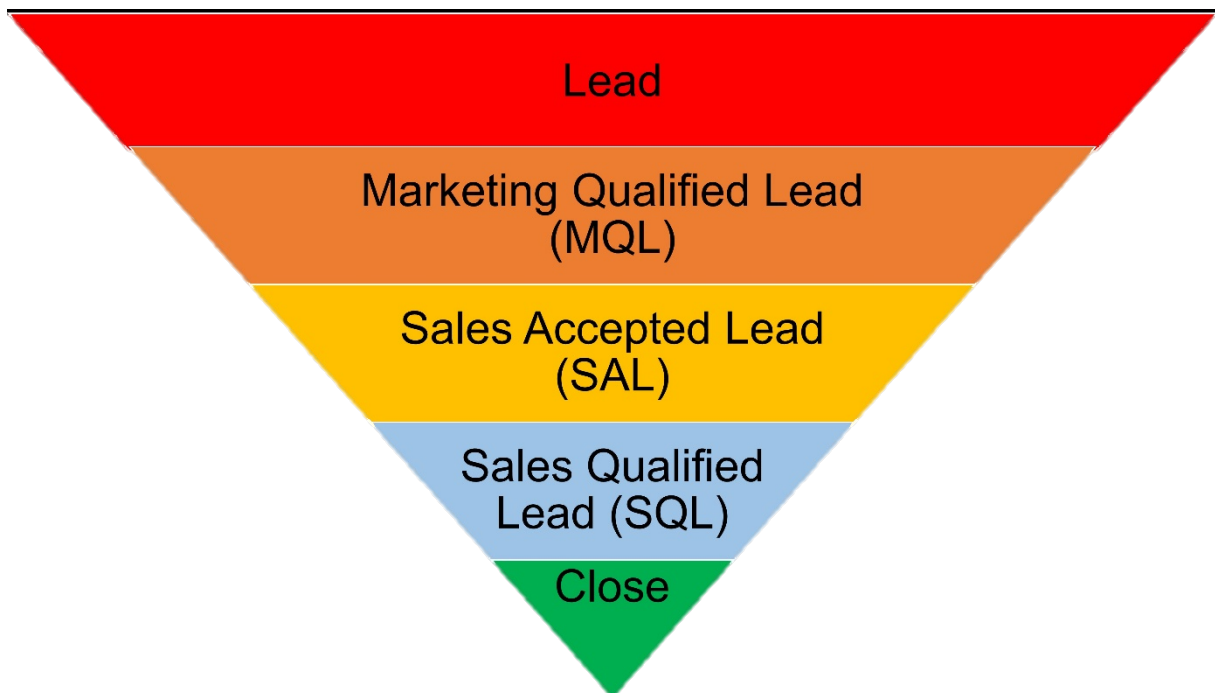


Figure 1: Traditional sales funnel

Source: Based on Schüller and Schuster 2022: 148 and Schuster 2022: 89

Lead

Leads are potential prospects or existing customers who show interest in a new product or service (see chapter 2.1).

Marketing Qualified Lead

These leads are managed by the marketing team and processed with marketing measures until they are classified as ready for sale. If this is the case, they are given MQL status and handed over to the sales team (cf. Philipp 2021: 207-208).

Sales Accepted Lead

Following the handover, the leads are checked more closely by the sales staff and either accepted or rejected. If a lead meets the parameters jointly defined by Marketing and Sales and is accepted by Sales, its status changes to SAL (cf. Schüller and Schuster 2022: 149).

Sales Qualified Lead

The contacts now go through a sales process in order to lead them to the final phase in the sales funnel, the closing (cf. Schüller and Schuster 2022: 149). In particular, this involves qualifying the leads from a sales perspective. If the sales team classifies them as a lead with high potential, they are assigned the status of Sales Qualified Lead. To classify a contact as SQL, the following BANT criteria must be met (cf. Hannig 2020: 218-219):

- Budget (B): A sufficient budget is available.
- Authority (A): The contact is authorized to make the purchase decision or has influence on the purchase decision.

- Need (N): The contact has a need that matches the sales offer.
- Time (T): The contact intends to buy in the foreseeable future.

2.4 Lead management

Lead management is a systematized process that aims to bring leads through the entire sales funnel to the sales team or to the close (cf. Koerner 2021: 72). Lead management can be divided into the following areas (cf. Jörvinen and Taiminen 2016: 170-172):

1. Lead generation and identification: The first step in lead management is lead generation and identification. This involves a lead providing their contact details to the company, e.g. by filling out a form. The behavior of existing contacts can then be recorded using various methods such as IP address tracking, cookies or website logins.
2. Lead nurturing: Lead nurturing refers to marketing processes that aim to convert leads into MQLs. The aim is to bring contacts closer to the purchase decision through meaningful and relevant content. The content is personalized based on a contact's profile information such as company, industry or position and their online behaviour.
3. Lead qualification: As not all interested leads proactively contact the sales team and not all leads have the same customer potential, lead qualification systems are used. These use collected information about leads in order to prioritize them. By awarding points based on behavior and characteristics, leads with higher scores are preferentially forwarded to the sales team.
4. Lead transfer: Qualified leads are automatically transferred to the Customer Relationship Management (CRM) system and distributed to lead queues that assign incoming leads to the appropriate sales teams. These queues can be categorized by geographic location and business unit, for example. Each sales team is responsible for processing at least one queue.
5. Closing: After the handover, the company takes measures to persuade the qualified leads to close the deal. The relevant information is recorded in the CRM system, regardless of whether a contact becomes a customer or not. In an optimal scenario, the systematic organization enables the entire marketing and sales process to be tracked, starting with the recording of contact information through to the purchase.

2.5 Marketing Automation

Marketing automation refers to the use of technologies and software solutions to automate tasks in lead and existing customer management (cf. Biegel 2009: 203; Schüller and Schuster 2022: 71; Heinzlbecker 2021b: 141-142). This includes processes such as customer segmentation, data integration and campaign management. Through the efficient

use of marketing automation software (MAS), more relevant content can be sent and leads can be converted into paying customers more effectively. As a result, the progress of leads through the various stages of the sales funnel is accelerated. At the same time, personnel expenses and the associated costs are reduced. However, successful automation requires strategic planning, adaptability and a well thought-out alignment of campaigns and the customer journey in order to reach the target group with relevant offers at the right time (cf. Bagshaw 2015: 84-85). When examining the possible functions and components of marketing automation software, the following were identified, among others:

Interfaces and CRM integration

One basis for efficient marketing automation is the connection of the MAS to other systems to ensure that the MAS can access data from these systems and transfer data to these systems at the same time. In particular, the connection of the MAS to the CRM system is necessary so that marketing and sales can work together efficiently, for example when transferring MQLs to sales (cf. Griebisch 2021).

Segmentation and dynamic content

In the course of segmentation, contacts are divided into different target groups or customer groups, also known as segments, based on their characteristics or behavior. The individual segments then receive personalized content or advertising messages. Segmentation can also be used to generate dynamic content for the individual contacts. This means that different content is displayed on websites and other platforms depending on which segment a contact is in (cf. Griebisch 2021; Teiu 2021: 330-331; DMEXCO n.d.).

Campaigns

Marketing campaigns ensure that contacts receive the right marketing content at the right time. In campaigns, various marketing activities are linked together and automatically sent to leads in a predefined sequence. The content is based on the interests and behavior of potential customers in order to ensure an individual and relevant approach (cf. Teiu 2021: 330-331).

E-mail marketing

In the area of email marketing, it is necessary for leads to first agree to receive content from a company by email as part of the opt-in process. After this consent, the leads can be provided with relevant content by email as part of the lead nurturing process. With the help of an MAS, the emails sent can be personalized by integrating recipient data such as name, company name and telephone number to make communication more individual (cf. Teiu 2021: 330).

Landing pages

Landing pages are websites that are used to present information or allow contact details to be entered in the form of forms (cf. Teiu 2021: 331).

Forms

Forms offer leads the opportunity to request information, register for events or ask to be contacted by the company. If a lead fills out a form and sends it off, this is a clear sign for marketing that the contact is interested in a product, service or event (cf. DMEXCO n.d.; Teiu 2021: 331).

Social Media Marketing

Lead nurturing can be carried out via social media. Forms can also be placed there to identify contacts or add them to marketing campaigns (cf. Teiu 2021: 331).

Scoring system

Another possible component of an MAS is a lead scoring system. In this system, contacts are evaluated and prioritized based on their behavior and characteristics (cf. Teiu 2021: 331-332).

Analytics + Reporting

Data is generated as part of the individual campaigns and marketing measures. This data is stored in the MAS in order to create reports or analyze patterns. This gives the company a better understanding of its leads and enables it to optimize its marketing measures for the future (cf. Teiu 2021: 332).

In Table 1 the previously identified components of lead management are assigned to the individual steps in lead management:

Lead generation and identification	Lead nurturing	Lead qualification	Lead routing	Closing
<ul style="list-style-type: none"> • Landing pages • Forms • Social media marketing 	<ul style="list-style-type: none"> • Campaigns • Email marketing • Social media marketing • Landing pages • Segmentation and dynamic content 	<ul style="list-style-type: none"> • Scoring system 	<ul style="list-style-type: none"> • CRM integration 	<ul style="list-style-type: none"> • Reporting tool

Table 1: Relationship between the components of marketing automation software and lead management
Source: Own representation

3 Classification of the term lead scoring

This section discusses the traditional lead scoring approach, the objectives of lead scoring and the challenges of lead scoring.

3.1 Traditional lead scoring

As described in chapter 2.4 lead scoring is a component of lead management that enables leads to be assessed according to their sales maturity using predefined criteria. As part of lead scoring, leads receive points based on their characteristics and actions. As soon as a lead's score reaches a predefined threshold, it is classified as MQL and passed on to the sales team (cf. Hannig 2020: 217).

Implicit and explicit lead scoring

When selecting the parameters relevant for lead scoring, a distinction is made between implicit and explicit parameters. Explicit scoring focuses on the evaluation of characteristics of the prospective customer. On the one hand, information about the person themselves and the target person's company is taken into account. Explicit scoring is used to determine how well a prospective customer matches the company's offer (cf. Schüller and Schuster 2022: 164-165). A collection of possible explicit parameters taken from the works of Schüller and Schuster (cf. 2022: 164-165) Adobe (cf. 2019: 18) and Patel (cf. n.d.) can be seen in Table 2.

Explicit parameters for the lead	Explicit parameters for the lead's company
<ul style="list-style-type: none"> • Current position • Number of years in the current position • Department • Interests • Management affiliation (Yes/No) • Role in the decision-making process or buying center • Type of email address (company address or Gmail, Yahoo, etc.) • Lead source • Products already bought 	<ul style="list-style-type: none"> • Number of employees • Revenue • Industry affiliation • Country • Budget availability • Time of the planned realization of the solution • Year of establishment • Organizational structure • Website traffic • Financial condition of the company • Revenue growth

Table 2: Explicit lead scoring parameters
Source: Own representation

It should be noted that explicit parameters can also have a negative impact on the score. For example, if a contact fills a position or is employed in an industry that does not match the company, points may be deducted (cf. Adobe 2019: 22).

In contrast to explicit scoring, implicit or action-based scoring focuses on evaluating the behavior of prospective customers. Even if a contact receives a high score in explicit scoring, it should not automatically be passed on to the sales department. This is because the explicit score only indicates that the lead's profile fits the company, but not whether the lead is ready for sales. Sales readiness is only achieved when the lead actively interacts with the company through actions. These actions are evaluated with the help of implicit lead scoring (cf. Schüller and Schuster 2022: 166). Within implicit lead scoring, it should also be noted that there are negative implicit parameters. These indicate that leads have a low level of interest in the company. Points are therefore deducted from the lead score for negative parameters (cf. Naveen and Hariharanath 2021: 1306). Examples of positive and negative implicit parameters in lead scoring, which were taken from the literature examined (cf. Adobe 2019: 19-21; Schüller and Schuster 2022: 166; Naveen and Hariharanath 2021: 1306-1307; Braun 2021: 163) can be seen in Table 3.

Positive implicit parameters	Negative implicit parameters
<ul style="list-style-type: none"> • Newsletter registration • Form submission • Blog page visit • Price page visit • Product page visit • Number of website visit • Website dwell time • Participation in webinars • Requesting informational material • Downloading informational materials • Reading reviews • Participation in forum or blog • Clicking links in emails • Email opens • Response to telemarketing campaigns • Attending an event • Social media engagement • Interaction with customer service • Utilizing a free trial • Visiting a trade fair 	<ul style="list-style-type: none"> • Joining a do-not-call list • Long periods of website inactivity • Newsletter unsubscribe request • Unopened emails • Visiting non-commercial pages, such as the career page, indicating no purchasing intent • Spam complaint

*Table 3: Positive and negative implicit lead scoring parameters
Source: Own representation*

Determining the scoring

In traditional lead scoring, experts with an understanding of the customer journey first determine which implicit and explicit parameters are relevant for their own lead scoring. A scorecard is then defined for both the implicit and explicit parameters. This lists all relevant parameters and assigns them point values, which are added to or subtracted from the implicit

or explicit score if the parameter is fulfilled (cf. Duncan and Eklan 2015: 1752; Schüller and Schuster 2022: 164-166). Table 4 shows an example of an implicit scorecard.

Activity	Points
Initial contact via requesting a content offer	30
Link-click in a lead nurturing email	10
Form submission	10
Website or product page visit	5
Response to content offer Whitepaper A	20
Response to content offer Whitepaper B	35
Response to content offer Whitepaper C	25
Participation in a webinar	40
Trade fair attendance	45
Newsletter subscription	10
...	

Table 4 Implicit scorecard in lead scoring
Source: Based on Schüller and Schuster 2022: 166

Defining a threshold value

Before the scoring system can be used in practice, the correct threshold values for the implicit and explicit score must be defined. These are the point values above which the leads are transferred to sales. At the same time, the MQLs can be prioritized using a lead scoring matrix. The individual scores are divided into different categories with threshold values depending on their level. For example, the implicit score can be divided into categories 1, 2 and 3 depending on the level and the explicit score into categories A, B, C and D (cf. Naveen and Hariharanath 2021: 1306; Schüller and Schuster 2022: 167). Accordingly, the leads can then be processed in order of priority, with, for example, the leads in category A1 from Figure 2 have the highest priority.



Figure 2: Lead scoring matrix
Source: Based on Naveen and Hariharanath 2021: 1306

3.2 Goals of lead scoring

In order to classify lead scoring in more detail, the positive effects of lead scoring are discussed below. Among other things, three main objectives were identified.

Increase in sales

In the marketing and sales process, it can happen that leads are contacted in the sales process before they are ready to make a purchase decision. This not only leads to frustration among the leads, but can also result in the loss of potential customers. Lead scoring makes it possible to identify and prioritize mature and ready-to-buy leads (cf. Schüller and Schuster 2022: 147). As a result, leads are only handed over to sales employees when they are ready to buy. This improved timing means that more leads are converted, which increases sales.

Improving collaboration between marketing and sales

Collaboration between marketing and sales is crucial for the successful conversion of leads into paying customers. If leads are handed over too early, this can lead to conflicts and inefficiencies in the collaboration between the two departments. On the one hand, sales lacks confidence in the quality of the leads handed over and may reject them prematurely,

while on the other hand, marketing feels that the leads are not being properly followed up by sales. By implementing a transparent lead scoring system, these conflicts can be avoided as the criteria for qualifying and prioritizing leads are clearly defined (cf. Hannig 2020: 218; Schüller and Schuster 2022: 147). This improves cooperation and the relationship between marketing and sales.

Cost savings and increased efficiency

A central goal of lead scoring is to streamline the sales process in order to maximize time and resource efficiency. If marketing forwards too many leads to sales, they can no longer be processed (cf. Lontzek 2022; Duncan and Eklan 2015: 1752). Lead scoring solves this problem as it ensures that only the most relevant and qualified leads are passed on to sales (cf. Michiels 2008: 4). This enables more targeted and efficient processing, as only those contacts that are actually interested in making a purchase are forwarded to the sales department. Lead scoring can also be useful if there are no capacity bottlenecks. As uninteresting contacts are automatically filtered out, the sales department can invest more time and resources in promising leads. This not only increases the chances of more successful sales deals, but also reduces the costs of lead processing. This leads to an overall more resource-efficient way of working in sales and increases the efficiency of the entire process (cf. Auerochs 2021).

3.3 Challenges in lead scoring

When analyzing the literature on lead scoring, several challenges were identified that can affect the success of lead scoring projects. These are outlined below.

Ensuring the completeness and quality of the data

As already described in the current chapter, lead scoring is based on the collection, storage and use of personal data. This makes lead scoring more difficult if the required data is not available in the required quantity and quality (cf. Monat 2011: 188; Lontzek 2022). According to the study "Benchmarking Marketing Automation: The Shift Toward Next Generation Lead Scoring & Segmentation" by Decision Tree Labs, one of the two main reasons for the failure of lead scoring projects from the perspective of the marketing experts surveyed is the availability of incomplete or inconsistent data on leads (cf. Lattice 2014: 6). The lack of relevant data for lead scoring can be due to several reasons.

One of these is limitations in lead tracking. Tracking methods are necessary to successfully analyze the customer journey of a lead and thus collect behavior-based data for lead scoring. The most common tracking method is the use of cookies (cf. Flocke and Holland 2014: 216). Cookies are small text files that are stored on the visitor's device when they visit a website. They help to uniquely identify individual visitors and store information about them. The

cookies are stored directly in the user's browser (cf. Gradow and Greiner 2021: 5-6). In the literature reviewed, some limitations of cookie tracking were identified that can have a negative impact on the acquisition of action-related lead data:

- **Necessity of consent:** If a cookie is not absolutely necessary to ensure the functionality of a website, users must actively consent to the use of cookies in accordance with the General Data Protection Regulation. If consent is not given, users must still be able to use the website to its full extent (cf. Gradow and Greiner 2021: 10-12). Therefore, no data for implicit lead scoring may be collected from users who have not consented to the use of cookies.
- **Deleting cookies:** In addition to the need for users to actively consent to cookies, users can delete cookies at any time within the browser (cf. Flocke and Holland 2014: 218) for example by deleting the browser history or clearing the cache. In addition, browsers such as Safari, Google Chrome or Mozilla Firefox offer the option of completely preventing the storage of cookies via the settings (cf. Mozilla n.d.; Apple n.d.; Google n.d.). Within Safari, cookies are even automatically deleted by default after seven days, even if the user has consented to the storage of a cookie on a website (cf. MM editorial office 2021). Actions that users carry out after deleting cookies can therefore no longer be clearly assigned and used for lead scoring.
- **Use of different browsers and devices:** As already described, cookies are stored in the browser of an end device. Therefore, users cannot be identified if they work with a different browser or device. For example, if a prospective customer fills out a form at work and then accesses the website on their own computer at home, the new data is not added to the previously generated lead profile (cf. Adobe 2019: 39).

In addition to the tracking of website activities, there are also problems with the tracking of emails. In email tracking, a small image, also known as a tracking pixel, is embedded in the text of the email. This pixel can then be used to determine whether an email has been opened (cf. Hu et al. 2019: 366). Adobe mentions two sources for pixel tracking in emails that can lead to incorrect tracking (cf. Adobe 2019: 39):

- **Email opening by bots:** Some email providers now use bots that automatically open emails to check them for spam. This incorrectly records that a contact has opened an email.
- **Blocked HTML:** Email recipients also have the option of blocking the loading of HTML parts in emails. In this case, the tracking pixel is not loaded. As a result, the email opening is not recorded in the lead scoring.

There are several alternative lead tracking technologies that can be used instead of cookies and pixels. These include, for example, tracking via a virtual fingerprint, IP address or website login. These alternative tracking methods solve some of the problems associated with the use of cookies. However, the alternative technologies also have weaknesses and therefore do not guarantee complete lead tracking. In practice, a combination of different tracking methods is therefore usually used as part of customer journey tracking (cf. Flocke and Holland 2014: 218-225; Jörvinen and Taiminen 2016: 170).

In addition to the restrictions on lead tracking, the data economy of the leads poses a further challenge for the completeness and quality of the data. New leads are often unwilling to disclose an excessive amount of personal data. The collection of too much data in forms can therefore lead to more interested parties deciding not to provide their data, thus reducing the conversion rate of opt-in forms, for example (cf. Schuster 2021: 110-111; Schüller and Schuster 2022: 154). Schüller and Schuster (cf. 2022: 154-156) therefore recommend the use of progressive profiling techniques. Progressive profiling in the context of lead scoring refers to a step-by-step data collection strategy in which companies gradually collect additional information from leads in several interactions instead of requesting all the necessary information in a single data collection process. This method aims to increase conversion rates by minimizing the initial collection effort while building a more comprehensive profile of the prospect over time. This step-by-step approach improves the quality of lead scoring and enables more accurate qualification of potential customers. Another option, which is described in the lead scoring guide from Adobe (cf. 2019: 11) is to use a service to complete missing fields in the lead database. This guarantees a complete database and at the same time achieves a high conversion rate, as only a small amount of information is requested in forms.

In addition, the use of data collected before lead identification poses a further challenge. A large part of the customer journey can already take place before the lead has been clearly identified by the company, for example by filling out an opt-in form. In principle, this data can be tracked anonymously before the lead is captured and assigned to the lead profile after identification. However, this enrichment of the lead profile with data prior to registration is not permitted in Germany. Therefore, the data may only be used after opt-in (cf. Schüller and Schuster 2022: 165-166).

Cooperation between marketing and sales

The efficient implementation of lead scoring requires close cooperation between the marketing and sales departments. Shaping this collaboration can be difficult, especially if both sides have different ideas about the lead scoring process (cf. Hannig 2020: 214). This can lead to the sales department not being satisfied with the leads handed over by the

marketing department (see chapter 3.2). In order to ensure smooth cooperation between marketing and sales, it is therefore necessary for marketing and sales to jointly define the framework conditions for their collaboration in the context of lead scoring. If the expectations of both sides are aligned, cooperation can be more efficient as both sides are acting in the interests of the other. A recommended procedure for defining these framework conditions in writing is the use of a Service Level Agreement (SLA). An SLA is a binding agreement between two parties on recurring services. In this case, the SLA is a kind of contract between the sales and marketing departments to ensure optimal lead management. This document clarifies various aspects such as the service levels to be met, the scope of services, the manner in which services are to be provided and the time frame. As the SLA is a critical factor for lead scoring, the document should always be maintained and regularly updated. For an SLA in the area of lead scoring, it is particularly recommended to document the answers to the following questions (cf. Schüller and Schuster 2022: 181-182):

- Which target group is selected for lead scoring?
- What is an ideal lead, or what does the ideal customer profile look like?
- How is an MQL defined?
- What is the definition of an SAL?
- How is an SQL defined?
- What data is transferred between the MAS and the Enterprise Resource Planning (ERP) and CRM system?
- What stages do prospective customers go through as part of the customer journey?
- What data is collected in the lead-nurturing process and how?
- Which method of lead evaluation is chosen (e.g. with points or grades in the form of letters)?
- What threshold values are defined for lead scoring?
- What information about the lead is transferred to the sales department?
- What happens after the lead handover?
- What does the sales team report back after the handover to the marketing department?

Integration of the marketing automation software

Another prerequisite for effective marketing automation and effective lead scoring is the integration of other systems that are used along the customer journey. Integration allows data to be automatically transferred to the MAS and used for data analysis as part of lead scoring. If this is not the case, there is a risk that important data will not be available for lead scoring (cf. Rahimi 2020; Schoepf 2021: 284; Koerner 2021: 76; Griebisch 2021). Data can

also be passed on to other systems. For example, data relevant to sales can be transferred to the CRM software (cf. Michiels 2008: 16).

Identifying the purchase probability factors

The importance of the individual implicit and explicit parameters varies from company to company. It is therefore a challenge to identify and weight the decisive criteria that have an impact on lead quality (cf. Monat 2011: 187). According to the study "Benchmarking Marketing Automation: The Shift Toward Next Generation Lead Scoring & Segmentation" mentioned earlier in this chapter, the second main reason for the failure of lead scoring projects from the perspective of the marketing experts surveyed, in addition to a lack of data, is a lack of insight into which implicit and explicit data actually provide indications of a willingness to buy (cf. Lattice 2014: 9). In traditional lead scoring, values and scores are determined on the basis of expert assessments. However, these are only subjective assumptions and there is no guarantee that they actually correspond to reality (cf. Jadli et al. 2022: 434, 2022: 434; Duncan and Eklan 2015: 1752). For this reason, it is recommended to base the evaluation system on data (cf. Bohanec et al. 2017). It is therefore advisable to use statistical methods to base the lead scoring system on data.

Sorting out irrelevant leads

In practice, not all contacts in the MAS are actually relevant. It cannot be ruled out that the contacts recorded there are research students, competitors or similar people who are researching without any interest in buying. It is therefore important to prevent these leads from reaching the sales team. Indications of this can be terms such as "student", "professor" or "unemployed", for example, which are specified as a position in forms. In these cases, the status of the lead must automatically be set to "unqualified" (cf. Gooding 2022: 236).

Consideration of non-linear effects

When lead scoring with a scorecard, it is not possible to take into account complex or non-linear relationships between the individual criteria. For example, if a lead attends several webinars, their score increases by the specified number of points for each webinar. However, it may be that the purchase probability of a lead only increases slightly after a certain number of webinars. For example, if the highest quality prospects have attended between one and three webinars, further webinar visits are not necessarily a signal of a higher propensity to buy. It may even be the case that attending many webinars is a negative signal. This could, for example, indicate the behavior of students or competitors who inform themselves about the company's offering (cf. Duncan and Eklan 2015: 1752).

Dependence on behavior-based data

Traditional lead scoring models rely heavily on implicit data. While this data can be a good indicator of leads' interest in a company's offering, it can prevent the early identification of

high-quality leads. Potential leads are not identified until they have taken enough actions to achieve a high score. Leads that are ready to buy from the outset are therefore not identified (cf. Duncan and Eklan 2015: 1752). As a solution, Gooding suggests (cf. 2022: 235) suggests developing a so-called "fast track path" as part of the lead evaluation. This involves defining actions or combinations of actions that indicate that a lead is interested in buying from the outset. If a lead performs these special actions, it is forwarded directly to the MQL and sales.

Consideration of the time component

One challenge in lead scoring is the consideration of the time component. If past actions, such as a visit to a product website two years ago, are evaluated in the same way as current actions, this can lead to inefficiencies in lead scoring. The reason for this is that past actions do not provide any information as to whether the prospective customer is still interested. It is therefore recommended to set past actions in relation to time in order to enable an accurate evaluation of the leads (cf. Kumar and Reinartz 2018: 151; Rahimi 2020).

Determining the threshold value

Another challenge is setting the threshold at which leads are handed over to the sales team. If this is set too low, too many unqualified leads will be passed on. If it is set too high, there is a risk that qualified leads will not be identified (cf. Lontzek 2022). month (cf. 2011: 191) and Adobe (cf. 2019: 26) recommend retrospectively calculating the lead scores of SALs at the time of handover and comparing these with the scores of leads that were not handed over or were rejected by sales. On this basis, the threshold value that would have achieved the best results in retrospect can then be selected.

Handover of leads to the right sales team

In order to ensure targeted lead scoring, leads must be assigned to the right sales team or the right sales employees. This assignment process requires consideration of the employees' individual knowledge and skills to ensure that the right person handles the contact (cf. Schüller and Schuster 2022: 182). To ensure an efficient process, it is therefore advisable to integrate the assignment step automatically into lead scoring. Integrating the CRM system can help here by automatically assigning the leads to the appropriate sales team or the right sales employees based on predefined rules (cf. Jörvinen and Taiminen 2016: 172).

Updating and optimizing the lead scoring process

If lead scoring systems are not updated regularly, it can happen that leads are no longer scored correctly. For this reason, it is necessary to review the system regularly and take into account new data and suggestions for improvement from sales staff (cf. Wu et al. 2023: 16). Adobe (cf. 2019: 32-33) recommends carrying out the review process at least every three months and integrating the following measures:

- Analyze the scores of all leads and their statistical distribution
- Checking for outliers and disqualified leads and adjusting the parameters in the system based on this
- Checking the behavior of current SALs and adjusting the parameters in the system based on this
- Incorporating new materials such as newly created landing pages into the system

Evaluation of the success of the lead scoring system

After implementing a lead scoring system, it is important to check its success. After all, it must be determined whether lead scoring actually leads to positive results. Various key performance indicators (KPIs) can be used for this purpose. Wu et al. (cf. 2023: 8) have identified the most common KPIs used in lead scoring practice as part of their research. These are listed in Table 5 where the Count column indicates how often the respective KPI was used in the literature examined.

KPI	Count
Lead conversion rate	12
Cost reduction / monetary savings	10
Number of MQLs	9
Hit rate on number of customers who buy	8
Annual revenue	7
Profit / financial gains	7
Density of profitable customers in the list	2
Response percentage	2
Customer value matrix	1
Overall customer satisfaction	1
Average time needed to qualify a lead	1
Activity level (e.g., website visits, log-ins)	1
Equilibrium percentage	1
Gain curve/score	1

Table 5 KPIs in lead scoring
Source: Based on Wu et al. 2023: 8

The most common lead scoring KPIs can be derived from this:

- Conversion rate
- Cost savings
- Number of MQLs

- Turnover
- Customer hit rate
- Profit

Other selected KPIs that Adobe (cf. 2019: 46-47) recommends are as follows:

- Sales productivity in the form of sales per sales employee
- Proportion of MQLs accepted by the sales team
- Success rate of leads accepted by sales
- Duration of the sales cycle
- Turnover per contract

4 Advanced approaches and use cases of lead scoring

This chapter will first discuss predictive lead scoring, which represents an alternative approach to traditional lead scoring. In addition, product-based lead scoring and account-based scoring are presented as extended use cases. Finally, some specific use cases found in the analyzed literature are mentioned.

4.1 Predictive lead scoring

An alternative approach to conventional lead scoring is predictive lead scoring, which is assigned to the predictive analytics application area. Predictive analytics combines various mathematical and statistical techniques that are used to recognize patterns in data and make predictions for future events on this basis. Predictive lead scoring relies on machine learning and statistical models to forecast the probability of a lead being converted into an actual customer (cf. Swani and Tyagi 2017; Nygard and Mezei 2020: 1441). Current developments indicate that predictive lead scoring models are increasingly preferred over traditional models. The existing literature also suggests that the integration of machine learning into the lead scoring process leads to improved results (cf. Wu et al. 2023: 1; Duncan and Eklan 2015: 1751-1758).

Differences between traditional and predictive lead scoring

This section explains the differences between the traditional and predictive lead scoring approaches in more detail, as well as the inefficiencies in traditional lead scoring that can be addressed by implementing predictive lead scoring.

	Traditional Lead Scoring	Predictive lead scoring
Rules	Subjective rules established by expert marketers	Detected by machine learning algorithms
Supervision	Requires manual supervision and regular adjustments and updates	Minimal supervision
Data size	Small datasets and limited processing power	Large datasets (accuracy increase with training data size)
Result	Lead scores	Conversion probability

Table 6 Differences in traditional and predictive lead scoring
Source: Based on Jadli et al. 2022: 435

Predictions instead of lead scores

Like shown in Table 6 a key difference between predictive and traditional lead scoring is that there is no score at the end of the process. The result at the end of the lead scoring process depends on the choice of machine learning algorithm (cf. Wu et al. 2023: 9-13). The following possible predictions were identified in the literature examined:

- Predicting lead conversion: By using classification algorithms, leads can be divided into two categories - those that are likely to convert and those that will not convert (cf. Bohanec et al. 2015: 338-352).
- Prioritization of leads: The use of machine learning makes it possible to classify leads based on their likelihood of conversion and then process them according to their priority (cf. Duncan and Eklan 2015: 1751-1758; D'Haen et al. 2016: 69-78).
- Conversion probability: By using algorithms in the "regression" category, the probability of individual contacts becoming customers can be predicted (cf. Espadinha-Cruz et al. 2021: 1-14).
- Feature Importance Output: Some algorithms also allow the display of Feature Importance, which shows which features or variables have the greatest influence on the predictions. This can be useful to understand which factors have the strongest influence on the likelihood of conversion (cf. Bohanec et al. 2017: 416-428). It is therefore possible to use this information to develop the scoring system in a traditional lead scoring system or to make the procedure within a predictive system more comprehensible.

Less updating effort

As described in chapter 3.3 time must be regularly invested in the maintenance of traditional lead scoring systems. In contrast, predictive systems require less maintenance (see Table 6).

Data-based decision-making instead of subjective expert assessments

As already discussed, the decisive parameters in lead scoring vary depending on the company. The traditional approach relies on expert assumptions to determine the individual parameters and their weighting. This approach means that lead scoring is not based on actual data, which in turn can lead to inaccurate results. In contrast, predictive lead scoring is based entirely on empirical data, which means that no subjective assessments are incorporated into the model (cf. Duncan and Eklan 2015: 1751-1758). This allows the challenge of identifying purchase probability factors to be overcome more efficiently.

In addition, predictive lead scoring models can recognize correlations that are too complex to be recognized by experts (cf. Wu et al. 2023: 15). In addition, non-linear effects can also be taken into account that cannot be mapped by the scorecard in traditional lead scoring (cf. Duncan and Eklan 2015: 1752).

Predictive lead scoring models are also characterized by a lower dependency on the amount of behaviour-based data compared to the traditional approach. In traditional lead scoring, potential leads are often only identified once enough actions have been carried out to achieve a high score. With predictive models, however, the dependency on the amount of behavior-based data can be reduced, allowing a more accurate assessment of the likelihood of purchase even for contacts who are ready to buy from the outset (cf. Duncan and Eklan 2015: 1752).

Suitability for evaluating large amounts of data

Predictive models offer the advantage of being particularly effective when analyzing and evaluating large amounts of data (see Table 6). In comparison, experts reach their limits when developing scorecards for traditional models, as the consideration of numerous parameters and their weighting can become too complex with large amounts of data.

Less transparency with predictive models

One disadvantage of predictive lead scoring algorithms is that the underlying models act as a "black box" and it is therefore not possible to understand what happens within such a model (cf. Wu et al. 2023: 17). This lack of transparency can have a negative impact on cooperation between marketing and sales, as the models do not offer any room for discussion. If sales lacks confidence in the quality of the leads submitted on the basis of the model, the efficiency of the process can be significantly impaired. However, there are methods to make the procedure within the machine learning algorithms more comprehensible and thus reduce the problem of intransparency (cf. Bohanec et al. 2017: 416-428).

Development of a predictive lead scoring model

This section examines the individual steps involved in creating a predictive lead scoring model. Predictive lead scoring is a specific use case from the field of predictive analytics and machine learning and therefore falls within the area of data mining (cf. Elkan 2013: 7; Swani and Tyagi 2017: 5-10; Jo 2021: 19). For this reason, the Cross-Industry Standard Process for Data Mining (CRISP-DM) is used in this thesis to develop a predictive lead scoring system. The CRISP-DM is a generic and cross-industry methodology for data mining that offers both beginners and experts a guideline for the implementation of a data mining project. It was developed as early as mid-1996 (cf. Shearer 2000: 14-19) but is still regarded as the standard in the field of data mining (cf. Martinez-Plumed et al. 2021: 3048; Abbasi et al. 2016: 13). The process is divided into six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment (cf. Shearer 2000: 14-19).

Business Understanding

In this phase, project goals are defined, business objectives are understood and a plan is developed. Key steps include determining business objectives, assessing the situation, setting data analysis goals and creating a project plan (cf. Shearer 2000: 14-19).

Data Understanding

This phase begins with the initial data collection and includes steps such as the description of the data, exploration through queries and visualizations as well as checking the data quality (cf. Shearer 2000: 14-19):

- Collecting the available data: In this step, the necessary data is acquired.
- Data Description: The properties of the previously acquired data are examined, obtaining information such as the data format, quantity, number of records and fields in each table.
- Data exploration: Detailed insights into the data are gained through queries, visualization and reporting. Initial findings and hypotheses are recorded on this basis.
- Data quality check: The data quality check is carried out to ensure that the data is complete and does not contain any missing values. The plausibility of the data is also checked.

Data Preparation

This phase includes all activities to build the final dataset or data that will later be incorporated into the model. These are described below:

- **Data selection:** The decision on which data to use for analysis is based on criteria such as relevance to the objectives of the data analysis, quality and technical limitations (cf. Shearer 2000: 16).
- **Data cleansing:** Data cleansing is an essential step in data mining, as the quality of the results depends on the cleanliness of the data (cf. Shearer 2000: 16). In the context of machine learning, there are models that can deal with missing data, while others are dependent on complete data. Therefore, it is necessary to handle missing values in the data by either deleting the corresponding rows or replacing missing values with suitable techniques such as the addition of the mean or median of a column (cf. Kumar and Reinartz 2018: 139; Elkan 2013: 19-20; Shearer 2000: 16). In addition to the treatment of missing values, data cleansing also includes the identification and elimination of outliers and the correction of incorrect values (cf. Kumar and Reinartz 2018: 139-140; Cleff 2019: 24; Kuhn and Johnson 2013: 33-35). In the area of lead scoring, for example, care could be taken to sort out leads with unusually high activity, as these may originate from system testers and could falsify the analysis results. This helps to improve the quality of lead scoring, as extreme values do not influence the results. Another aspect of data cleansing in machine learning is data reduction. Here, an attempt is made to reduce the amount of data that flows into the model. This can be done by removing columns that do not provide any relevant information for the model. An example of this is the elimination of redundant variables that correlate strongly with other variables and could therefore negatively influence the analysis results (cf. D'Haen and van den Poel 2013: 548; Kuhn and Johnson 2013: 35).
- **Data construction:** After data cleansing, new data sets or attributes derived from existing data sets may be developed as part of data construction. The purpose of these derived attributes is to facilitate the modeling process or to support the modeling algorithm (cf. Shearer 2000: 16). In principle, text-based variables are not suitable for making predictions in the field of machine learning. Therefore, text-based data is coded in the data construction process by converting it into numerical values. Various coding techniques are used for this (cf. Kuhn and Johnson 2013: 47-48; D'Haen and van den Poel 2013: 548; Uhlemann 2015: 10). An example in the context of lead scoring would be the coding of the "Action" column in a list containing all actions of all leads. Assuming that the "Action" column contains the four unique actions "Visit price page A", "Visit price page B", "Download PDF A" and "Download PDF B", one possible approach would be to create two columns for each customer:

one with the number of visits to price pages and one with the number of PDF downloads.

- **Data integration:** In data integration, information from several tables or data sets is merged to create a table that contains all the information required for the model. Data integration also includes aggregations. Aggregations refer to operations where new values are calculated by combining information from multiple datasets and/or tables. An example of an aggregation could be the transformation of a table of customer purchases in which there is one record for each purchase made. This table could be transformed into a new form where there is one record for each customer. The fields of this new table could contain information such as the number of purchases made, the average purchase amount, the percentage of orders paid for by credit card and other aggregated values (cf. Shearer 2000: 16-17).
- **Data formatting:** In connection with data mining, it may be necessary to adapt the format or structure of data. These adjustments can be simple, such as removing invalid characters from strings or truncating to a maximum length, or more complex, such as restructuring information. Such changes are sometimes necessary to make the data suitable for the application of a particular modeling tool (cf. Shearer 2000: 17). When developing machine learning models, it is important to divide the data into test and training data as part of the data formatting process in order to enable an accurate assessment of the model's performance at a later stage. Models are first trained on the basis of training data. The test data is then used to evaluate the performance of the trained models by validating them against unknown data. This division makes it possible to recognize so-called overfitting and ensure that a model can correctly predict not only the training data, but also previously unknown data. This increases the reliability of the model in real-life situations (cf. Kuhn and Johnson 2013: 60-61). It can also be useful to scale and center the data. When centering, for example, the mean value can be subtracted from each variable so that the new mean value is zero. When scaling, the individual variables can be divided by the standard deviation within the column. This procedure can improve the prediction of models, as the individual columns now lie on a common scale (cf. Cleff 2019: 20; Kuhn and Johnson 2013: 30-31).

Modeling

In this phase, various models are selected and applied, with the parameters of each model being calibrated to optimum values. The following steps are carried out for this purpose:

- **Model selection:** In this step, one or more models are selected to be used in the project. Predictive lead scoring algorithms can come from the categories of

classification, regression and clustering, with classification algorithms being used most frequently. Regression models provide numerical values based on parameters that are entered into the model. This can be, for example, the probability with which a lead converts. Classification models, on the other hand, assign the objects entered into the model to predefined categories, such as the "lead will convert" group and the "lead will not convert" group. Clustering models work in a similar way to classification models, but they work without predefined categories, instead forming the individual groups themselves (cf. Wu et al. 2023: 9-13). Machine learning algorithms such as linear regression, logistic regression, decision tree, random forest, support vector machine and neural networks were used in the literature examined. The best results were achieved using the random forest algorithm, which is a model in the classification category (cf. Bohanec et al. 2017: 416-428; Gokhale and Joshi 2018: 279-291; Jadli et al. 2022: 433-443; Nygard and Mezei 2020: 1439-1448).

- Model development: The model selection is followed by the model development step. In machine learning, this step includes hyperparameter tuning. In hyperparameter tuning, different settings that influence model performance are systematically tried out for each selected model. The hyperparameters that achieve the best performance are determined. The main goal of this process is to adjust the hyperparameters so that the model delivers good results on previously unknown data (cf. Kuhn and Johnson 2013: 63, 73).
- Model evaluation: The evaluation of machine learning models requires a precise analysis of their performance, using the confusion matrix as a tool. This matrix provides a detailed overview of the predictions of a model compared to the actual results. The confusion matrix contains four categories: True Positives (TP) for correctly predicted positive outcomes and True Negatives (TN) for correctly predicted negative outcomes. False positives (FP) and false negatives (FN) describe the number of incorrectly predicted positive or negative results. An example of a confusion matrix can be found in Table 7. One metric for evaluating the performance of a machine learning model that is derived from the confusion matrix is accuracy. This indicates the proportion of correct predictions by dividing the number of TP and FP predictions by the total number of predictions. Precision indicates the proportion of correct predictions in relation to all positive predictions. Sensitivity indicates the proportion of correct positive predictions in relation to the number of actual positive values. The 1-specificity indicates the proportion of false positive predictions in relation to all false predictions (cf. Nygard and Mezei 2020: 1444; Von der Hude 2020: 149-152).

		Predicted	
		Positive	Negative
Truth	Positive	True Positive	False Negative
	Negative	False positive	True Negative

Table 7: Confusion matrix
Source: Based on Elkan 2013: 49

Evaluation

Before final deployment of the model, it is important to evaluate the model and review the design process. The most important steps are to evaluate the results in terms of achieving the business objectives, reviewing the process for weaknesses and determining the next steps (cf. Shearer 2000: 17-18).

Deployment

As a rule, the project is not completed with the creation and evaluation of the model. The knowledge acquired must be organized and processed so that it can be used. The most important steps in this context are the planning of the application, the monitoring and maintenance and the preparation of the final report (cf. Shearer 2000: 18).

4.2 Product-based scoring

Product-based scoring is a form of lead scoring that enables companies to measure the interest of potential customers in various products. This approach goes beyond general behavior-based scoring, which focuses on interest in the company as a whole. Instead, product-based scoring creates different scores for different products to gather more detailed information. If a lead exceeds the threshold for a product score, it can be forwarded to the sales team responsible for that product. With product-based scoring, product scores can be created not only for individual products, but also for higher-level product groups (cf. Adobe 2019: 29). Scoring can therefore theoretically also be based on individual business units. The appropriate structure can be agreed in advance with the individual sales teams. Product-based scoring can be used, for example, in a company that sells ERP software, CRM software and supply chain management software. If only one score is used in this case, it is not possible to measure the interest of the leads exclusively in relation to the CRM software product and then transfer the MQLs for CRM software to the designated sales team. The option of product-based scoring is offered by several providers of MAS and CRM systems (cf. Oracle n.d.; Salesforce n.d.; InvestGlass 2023).

In practice, it is advisable to start with lead scoring on a small scale and with reduced complexity and to gradually expand the system (cf. Schoepf 2021: 283; Auerochs 2021). Product scoring can also help here, as it enables individual product areas to develop isolated scoring systems instead of having to set up a complex company-wide system. If necessary, the lead scoring system can first be tested in an individual department before it is used company-wide.

4.3 Account Based Scoring

In addition to product-based scoring, account-based scoring (ABS) is another advanced application of lead scoring. ABS is part of account-based marketing and evaluates the companies superior to the leads instead of individual leads. The aim is therefore not to qualify individual leads as MQLs, but to qualify a company or an account as a Marketing Qualified Account (MQA) (cf. Day and Wei Shi 2020: 18-19).

Scoring at company level can be more effective, as larger transactions in the business-to-business (B2B) sector usually involve several people in the purchasing process. Therefore, looking at a single lead no longer provides sufficient insight to infer the company's buying interest (cf. Day and Wei Shi 2020: 16; Schuster 2022: 203). For example, it may be that none of the company's individual leads exceeds the defined score, but the leads together have a very high score, which is a clear indication of the company's interest in buying (cf. Adobe 2019: 30).

To develop an ABS model, high-quality companies that have been accepted by sales in the past are first examined. On this basis, criteria can then be derived that indicate high account quality. The scorecards for the ABS are then created. Explicit parameters can be the following, for example (cf. Day and Wei Shi 2020: 20):

- Number of employees
- Turnover
- Location
- Industry
- Technology profile
- Settings information
- Product information
- Financing
- Web ranking
- Presence in social media

The implicit score results from the sum of the points for the activities that all leads assigned to the company have collected. If the defined thresholds are exceeded, the company receives MQA status and is transferred to sales (cf. Day and Wei Shi 2020: 19-20).

A predictive lead scoring approach can also be used in the course of ABS (cf. Heinzlbecker 2021a: 394). However, it should be noted that a smaller number of accounts are scored in ABS (see Table 8). As sufficient data must be available for the development of a predictive lead scoring system (cf. Elkan 2013: 8) it must therefore be ensured that sufficient data from the past is available to train the algorithm.

	B2C	B2B
Target audience	End User	Enterprise
Target market size	Large	Smaller
Sales volume	Low	High
Decision making	By the consumer	By a committee
Risk	Low	High
Purchasing process	Short	Longer
Consumer decision	Emotional	Rational
Usage of mass media for promotion	Common	Often avoided

Table 8: Differences in the B2C and B2B markets
Source: Based on Saha et al. 2014: 295-297

4.4 Lead scoring without a sales team

Within the scientific literature, handover to a sales team is almost exclusively mentioned as an action when a score threshold is exceeded. Only Zumstein et al. (cf. 2023: 35) mention the implementation of special communication measures based on the lead score, indicating that MQLs do not necessarily have to be handed over to sales. As companies without a sales team also exist in practice, for example in the area of e-commerce, search engines were used to research other options for implementing measures based on the score in lead scoring. The following were identified:

- Lead segmentation: One option is to dynamically assign leads to segments based on their scores, for example "cold leads", "lukewarm leads", "warm leads" and "hot leads". This not only enables better structuring of the leads, but also allows customized marketing measures to be implemented at a later date (cf. Brevo 2023).
- Sending personalized lead nurturing content, offers and Calls To Action (CTAs): In companies without a dedicated sales force, special offers and CTAs can be sent to MQLs to convert them into customers after their score exceeds the threshold. In

addition, the lead score can be used as part of lead nurturing to ensure that leads receive the right content at the right time. This makes it possible, for example, to provide warm leads with special content to turn them into hot leads (cf. Hufford 2021; Ghorbel 2023). In product-based lead scoring, content, offers and CTAs can also be tailored to the product groups that a lead is interested in (cf. faraday.ai n.d.).

- Targeting in social media: Leads that exceed a predefined score can also be addressed via social media to increase the likelihood of a purchase (cf. Brevo 2023).

4.5 Further use cases

In addition to the methods presented in this chapter, which have been examined in several scientific publications, there are also more specific use cases that have been examined in a few publications. For example, predictive lead scoring has been proposed and used to identify existing customers at risk of churning (cf. Buckinx and van den Poel 2005: 264; Simmoleit 2023). In addition, Kim and Street (cf. 2004: 215-228) constructed a model that calculated the profit for leads that could be achieved by sending direct mail. This meant that direct mail was only sent to leads that were predicted to make a high profit in order to maximize profit.

5 Alternative and complementary methods to lead scoring

5.1 RFM analysis

RFM analysis is a method from the field of CRM. RFM stands for "Recency", "Frequency" and "Monetary". The method is used to segment customers based on their most recent transactions (Recency), the frequency of their transactions (Frequency) and the monetary value of their purchases (Monetary). The aim is to divide customers into groups according to their quality in order to address them with more targeted marketing strategies and activities (cf. Kumar and Reinartz 2018: 103-111). To this end, a table (see Table 9), in which all of the contacts' purchases are listed, points are awarded for each entry. The points awarded are determined in advance. In the example, 100 points are awarded for purchases made in the last two months, 50 points for purchases made in the last six months, 15 points for purchases made in the last nine months and five points for purchases made in the last twelve months. Frequency points are calculated by awarding six points for each purchase made in the last twelve months, up to a maximum of 30 points. Monetary points in this example amount to ten percent of the purchase value, with a maximum of 75 points.

Customer	Recency Points	Frequency Points	Monetary Points
John	100	6	12
John	50	6	36
John	5	6	18
Smith	15	6	75
Mary	100	6	27
Mary	50	6	21
Mary	25	6	24
Mary	5	6	12

Table 9: RFM table for awarding points per purchase
Source: Based on Kumar and Reinartz 2018: 111

The points of the individual purchases per contact are then added up. If the recency score, the frequency score and the monetary score are now added together, the RFM score is obtained (see Table 10).

Customer	Recency Score	Frequency Score	Monetary Score	RFM score
John	165	18	66	249
Smith	15	6	75	96
Mary	180	24	84	288

Table 10: Calculation of the RFM score
Source: Based on Kumar and Reinartz 2018: 103-111

RFM analysis and lead scoring therefore differ in their areas of application. RFM analysis aims to evaluate existing customers and therefore cannot qualify leads that have not yet purchased. Lead scoring, on the other hand, can also evaluate new leads.

5.2 Product recommendation systems

Product recommendation systems are powerful tools that are used by e-commerce companies in particular to suggest relevant products to users (cf. Hu and Zhang 2012: 1). The systems work in two steps. First, the activities and interests of the users are analyzed. Secondly, it tries to find a group of items that could be of interest to users (cf. Sharma et al. 2021: 1).

For this purpose, mainly content-based systems, collaborative filter systems and hybrid approaches of both systems are used. As shown in Figure 3 content-based systems analyze the properties and features of the products themselves. They use information such as keywords and categorizations to recommend products that are similar to the products users

are interested in. For example, a content-based system could recommend other formal fashion accessories to users searching for formal clothing (cf. Sharma et al. 2021: 3).

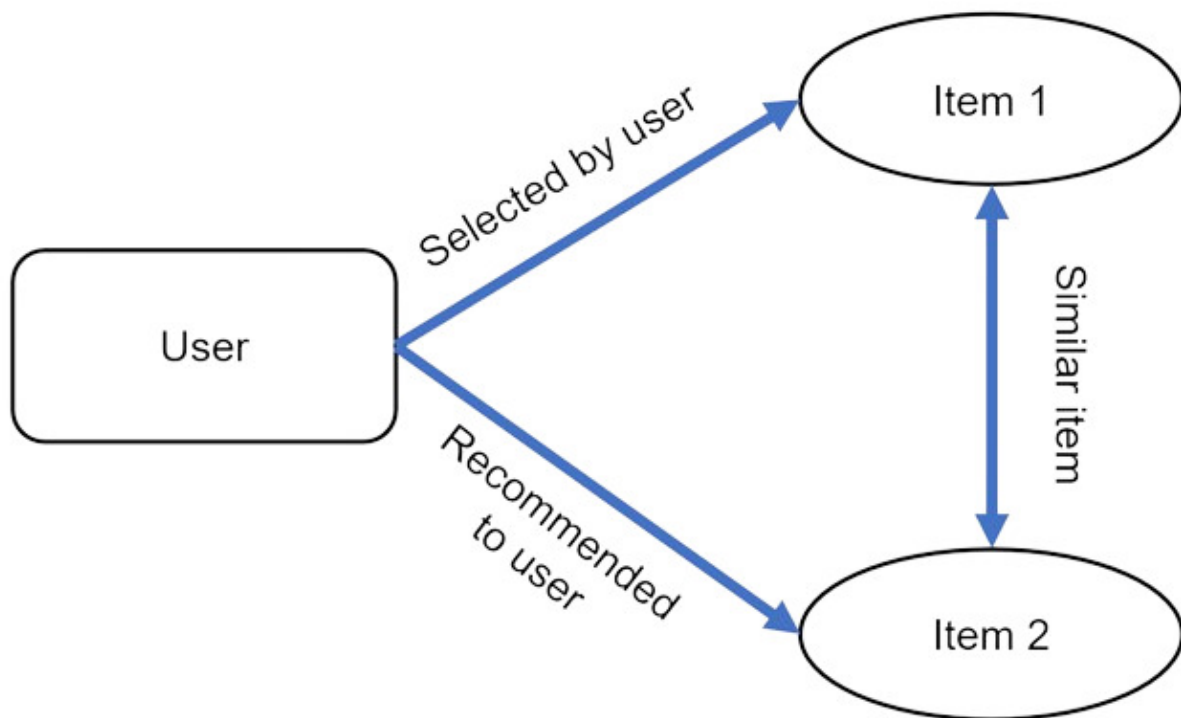


Figure 3: Content-based product recommendation systems
Source: Based on Sharma et al. 2021: 3

In collaborative filter systems, recommendations are created on the basis of interactions and ratings from similar users. For this purpose, the activities of users are analyzed to identify similar user groups. If user A has similar preferences to user B, products that user B liked can also be recommended to user A (see Figure 4).

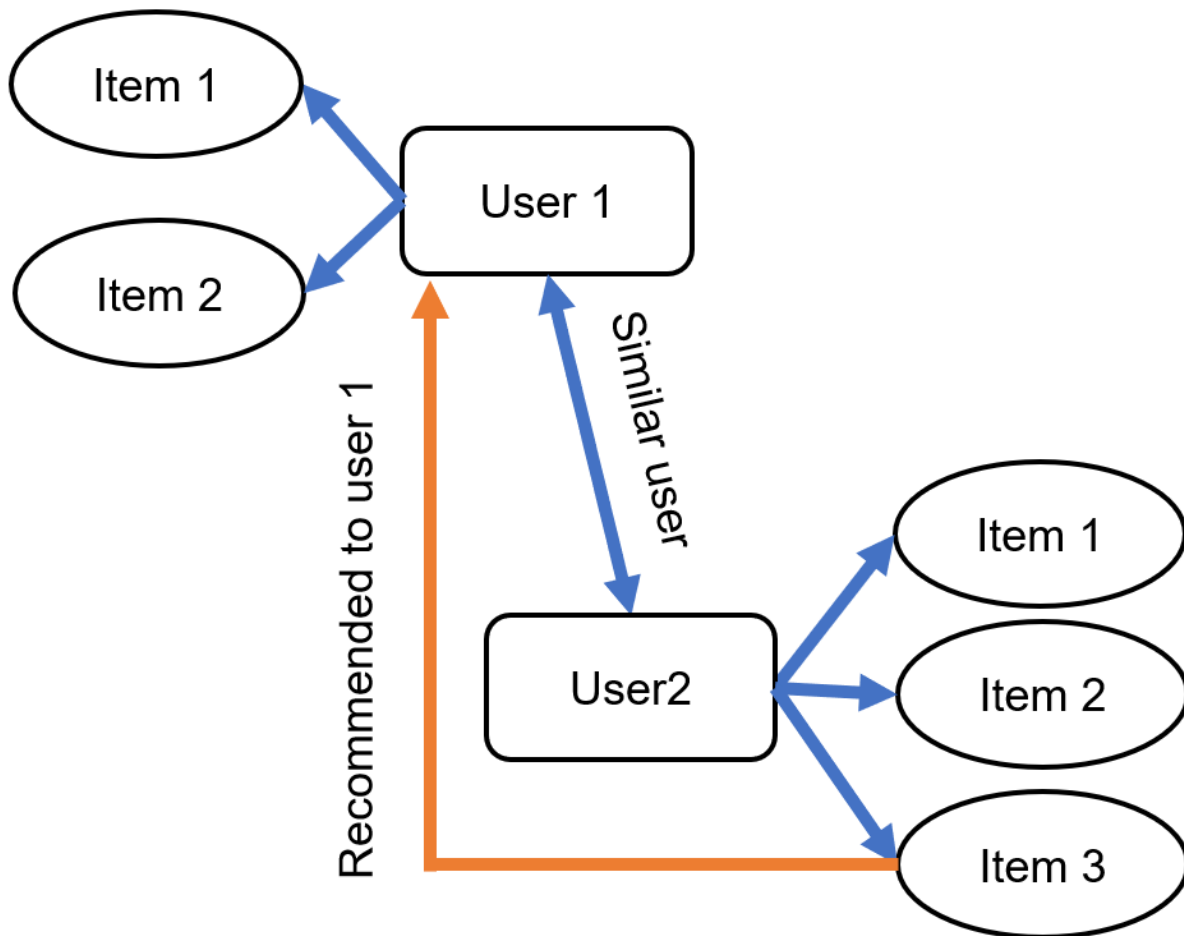


Figure 4: Collaborative filter systems
Source: Based on Sharma et al. 2021: 4

The hybrid approach is a combination of content-based systems and collaborative filter systems. As a result, users are recommended articles that match their interests as well as articles that are preferred by customers with similar purchasing behavior (cf. Sharma et al. 2021: 3-4).

Algorithms that can be considered for the implementation of the aforementioned systems include, for example, K-Nearest Neighbor and Matrix Factorization (cf. Sharma et al. 2021: 4-5). In addition, approaches such as bipartite projection, spanning tree and the application of cosine similarity are relevant methods in the context of product recommendation systems (cf. Hu and Zhang 2012: 2).

In contrast to lead scoring, which aims to evaluate customers in terms of their likelihood to buy, product recommendation systems therefore aim to suggest the most relevant products to customers. Despite these differences, product recommendation systems can pursue similar goals to product-based lead scoring, as both offer the possibility of identifying the products that are particularly relevant for a specific lead from a range of different products.

6 Development of a generic process model for lead scoring

Based on the previously identified challenges of lead scoring, this chapter presents a generic process model for the development of a traditional and a predictive lead scoring system. These two process models are illustrated in Figure 5 with the differences between the models highlighted in light blue.

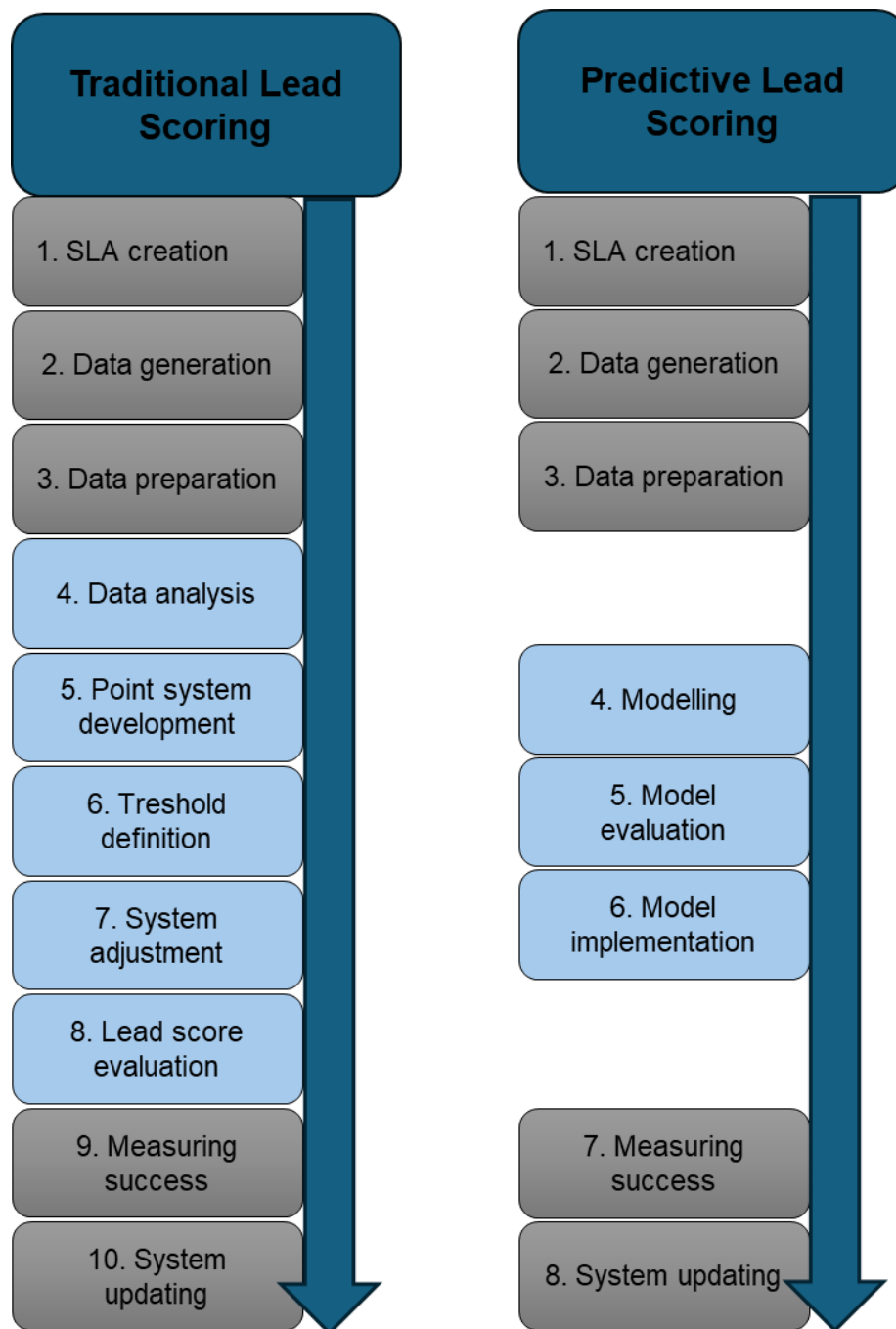


Figure 5: Differences between the traditional and the predictive process model for creating a lead scoring system
Source: Own representation

6.1 Traditional lead scoring

Creation of a service level agreement

As described in chapter 3.3 it is recommended that marketing and sales clearly define the framework conditions of the lead scoring project in the SLA as a first step. This should clarify the following questions, among others:

- Which target groups is lead scoring applied to?
- What does the ideal lead profile look like?

- How are MQLs, SALs and SQLs defined?
- What stages do prospective customers go through as part of the customer journey?
- What data should be collected?
- What data needs to be transferred between the MAS and other systems?
- Which lead evaluation method is chosen?
- When is a lead ready to be handed over to sales?
- What information about the leads is transmitted when they are handed over to the sales department?
- What processes need to be followed after the handover of marketing and sales?
- When and how does the lead scoring process end?
- Which KPIs are used to measure the results of the lead scoring process?

Data generation

The first step in lead scoring is the generation of historical data. This involves collecting data that includes all potentially important lead characteristics, i.e. implicit and explicit parameters, as well as the final disposition of the leads (cf. Monat 2011: 188). This data is then used to answer the following questions (cf. Adobe 2019: 16):

- Which touchpoints do leads go through to become SALs or customers?
- Which touchpoints do leads go through that are rejected by sales or do not become MQLs?
- What properties do MQLs have?
- What are the characteristics of leads that are rejected by sales or never reach MQL status?
- How many touchpoints do leads pass through on average before they receive MQL status?
- How long does it take on average for a lead to become an MQL?

To answer the relevant questions, it is advisable to generate the following data (cf. Nygard and Mezei 2020: 1443):

- Lead ID for unique identification of a lead
- Date of lead capture
- Relevant explicit lead data
- List of lead activities with timestamp
- Lead status ("SAL", "Rejected", "No transfer")
- If available, date of lead handover to the sales department
- Date of the last lead activity

In addition to the automatically generated data, it is also advisable to obtain feedback from the sales team when collecting data. In particular, problems with the leads currently submitted to the team should be discussed. This can be used to determine which data is currently frequently missing or what the most common reasons are for leads being rejected (cf. Adobe 2019: 16). As described in chapter 3.3 it is also advisable to develop a strategy to ensure the completeness and quality of the data. For example, data supplementation services or progressive profiling can be used for this purpose.

Data preparation

Several of the texts examined criticize the fact that the scorecard in traditional lead scoring is not based on data. Therefore, a data analysis is carried out in this step in order to create a generic lead scoring model. As traditional lead scoring is also a form of data mining, the following steps of data preparation or data preparation from the CRISP-DM are carried out here:

- **Data selection:** The first step is to select the data to be used for the analysis. This decision is based on criteria such as relevance to the analysis objectives, data quality and technical limitations.
- **Data cleansing:** In this step, incomplete data is either removed or replaced by techniques such as averaging or median formation. In addition, outliers are removed and obviously incorrect values are corrected, as these can have a negative impact on the quality of the data analysis. Data that is not relevant for the data analysis is also deleted.
- **Data construction:** After data cleansing, new data sets or attributes derived from existing data sets may be created as part of data construction in order to improve the results of the data analysis.
- **Data integration:** In data integration, information from several tables or data sets is merged to create a table that contains all the information required for data analysis. Data integration also includes aggregations. Aggregations refer to operations where new values are calculated by combining information from multiple datasets and tables.
- **Data formatting:** In connection with the data mining product, it may be necessary to adapt the format or structure of the data. These adjustments can be simple, such as removing invalid characters from character strings or shortening to a maximum length, or more complex, such as reorganizing information.

Data analysis

Once the data has been processed, it is analyzed. For this purpose, first a univariate and then a bivariate analysis is carried out. In the univariate analysis, each variable is considered

individually. Statistical tools such as mean, median, standard deviation, minimum, maximum, spread and histograms can be used here (cf. Cleff 2019: 26-52). An example from lead scoring would be that a mean value, a standard deviation and a histogram are formed for the number of visits to price pages per contact. This makes it easier to understand the variable.

Bivariate analysis is a statistical approach in which two different variables are compared with each other in order to identify relationships, patterns or correlations between them (cf. Cleff 2019: 71-110). In lead scoring, for example, the correlation between the number of visits to price pages by a lead and the achievement of MQL status can be determined.

Developing the points system

After the data analysis, a data-supported scoring model can be created. The points are awarded by experts, but they use the results of the data analysis to validate their assessments. This counteracts the problem that traditional lead scoring systems are based exclusively on subjective expert assessments. In order to map the scoring model, points are assigned to the individual actions and characteristics in an explicit and an implicit scorecard. These are added to or subtracted from the respective implicit or explicit score upon fulfillment.

As described in chapter 3.3 the chronological sequence also represents a challenge in lead scoring. Actions that were carried out before a period that goes far beyond the usual sales cycles are not accurate indicators that there is still interest in buying. One method that providers of MAS and CRM systems recommend to incorporate older actions with lower value into lead scoring is expiration models. Three expiry models were identified as part of the research. These are the expiry of points after a predetermined period of time, halving the score after a longer period of inactivity and reducing the score after a longer period of inactivity (cf. ActiveCampaign n.d.; Encharge.io 2021; ConstantContact 2023):

- Points expire after a specified time: The time after which the points awarded expire and are thus deducted from the lead's score is specified here for each point allocation (cf. ActiveCampaign n.d.).
- Halving the points after a longer period of inactivity: With this expiry method, the last activity of the contact is checked. If this exceeds a certain period of time, the lead score is halved (cf. ConstantContact 2023).
- Reduction of points after prolonged inactivity: This expiry method also checks the contact's last activity. If this exceeds a certain period of time, a fixed number of points are deducted from the score. For example, 15 points are deducted after 30 days of inactivity, 30 points after 60 days of inactivity and 50 points after 90 days of inactivity (cf. Encharge.io 2021).

However, it should be noted that although expiry models are recommended by various MAS and CRM system providers, no scientific research has yet been conducted to determine whether an expiry model actually achieves a better result. However, the questions of how a decay model affects the success of lead scoring, how the right decay model can be selected and how it can be successfully configured are beyond the scope of this research.

Defining the threshold value

Once it has been determined how the individual parameters are evaluated, the threshold value is defined. The threshold value is the score from which a lead is classified as MQL and transferred to sales. The threshold value is determined by retrospectively calculating the lead scores of past SALs at the time of transfer to sales. This makes it possible to determine how high the score must be for qualified leads to be transferred. Rejected and disqualified leads are also analyzed. This allows you to determine how low the threshold value may be without bad or unqualified leads being passed on.

Customizing the lead scoring system

In principle, it should be noted that several runs are necessary to develop a robust lead scoring model (cf. Hannig 2021: 250). It therefore makes sense to adjust the threshold value, the score and the point decay after the first run and to test whether more leads are correctly predicted as a result. The point values and threshold value with which the model achieves the best results are then selected.

Evaluating the lead scores

Various paths are recommended when evaluating leads:

Path to success

Leads collect points on this path until the previously defined implicit and explicit thresholds are exceeded. They then become MQLs and are automatically transferred to sales. As described in chapter 3.3 the lead scoring system relies on a sufficient amount of behavioral data. For this reason, it is a challenge to identify leads that are ready to buy from the outset. Therefore, parameters are determined from the data analysis that indicate an immediate interest in buying and an accelerated path in lead scoring is developed accordingly, along which the leads are immediately forwarded to sales (cf. Gooding 2022: 235).

Disqualification path

On the disqualification path, leads that have been rejected by sales are removed from the lead scoring system. In addition, there are contacts for which it is clear from the outset that they are research students, competitors or similar contacts. To prevent such leads from reaching sales, rules are first defined on the basis of data analysis to identify them.

Workflows are then developed to automatically remove these leads from the lead scoring system (cf. Gooding 2022: 236).

Recycling path

On the recycling path, leads that have interacted with the sales team but are not yet ready to buy are returned to marketing. Contacts that were initially classified as SAL but then lost in the sales process should also be identified and returned to marketing. One way to design the recycling path is to give sales staff the opportunity to select reasons for not closing a deal. Depending on these reasons, marketing can automatically take predefined measures to sort out these contacts or re-qualify them through marketing. In this way, leads that were already considered lost can be kept warm until there is concrete interest (cf. Gooding 2022: 236).

Measuring success

Once a lead scoring model has gone live, it is important to check how successful it is. Finally, it is necessary to find out whether lead scoring is actually producing positive results. This can be done using the methods described in 3.3 can be used for this purpose. These include

- Conversion rate
- Turnover
- Profit
- Cost savings
- Number of leads that receive the status MQLs
- Sales productivity in the form of sales per sales employee
- Proportion of MQLs accepted by the sales team
- Success rate for SALs
- Duration of the sales cycle
- Turnover per contract

Updating the lead scoring system

Lead scoring systems should be reviewed at least quarterly to ensure that they are still in line with reality. The following actions are carried out for this purpose (see chapter 3.3):

- Checking the KPIs
- Analyzing the score of the leads
- Adjust scoring by analyzing disqualified leads, MQLs and outliers
- Incorporating new marketing materials into the lead scoring system
- Adjusting the scoring and the threshold value

The changes in the lead scoring process are then documented. To make updating the lead scoring system even more efficient, it is also advisable to integrate a feedback mechanism.

Here, the sales team provides feedback on the quality of each individual MQL. This feedback can then be used to improve the system or scoring (cf. D'Haen and van den Poel 2013: 544-551).

6.2 Predictive lead scoring

In addition to the generic process model for creating a traditional lead scoring system, a process model for creating a predictive system was also developed. In this model, some steps are identical, but the steps "Data analysis", "Developing the scoring system", "Determining the threshold value", "Adjusting the lead scoring system" and "Evaluating the lead scores" are replaced by the steps "Modeling", "Model evaluation" and "Model implementation" (see Figure 5). The individual steps of the process model were developed by combining the necessary steps for creating a lead scoring system identified in the literature and the steps of the CRISP-DM.

Creation of a service level agreement

The creation of an SLA is similar to the "Business Understanding" step of the CRISP-DM (see chapter 4.1). The objectives and framework conditions of lead scoring are defined between marketing and sales.

Data generation

As with traditional lead scoring, data is also collected here that includes all potentially important implicit and explicit parameters as well as information on the final disposition of the leads.

Data preparation

As with the traditional process model, the predictive process model also goes through the steps of data selection, data cleansing, data construction, data integration and data formatting. As part of data formatting, the data in predictive scoring is also divided into training and test data and scaled in order to develop a more robust model.

Modeling

In the modeling step, several algorithms are first selected, which are then tested. Common algorithms for predictive lead scoring are linear regression, logistic regression, decision tree, random forest, support vector machine and neural networks. These algorithms are then fed with the previously prepared training data to develop a model. Different hyperparameters are tested for each algorithm. The combination of algorithm and hyperparameter settings that provides the most accurate predictions when applied to the test data is then selected.

Model evaluation

In the evaluation phase, weaknesses in the model and its design process are sought. If any are identified, the model is adapted.

Model implementation

In this phase, the model is integrated into the MAS. In addition, all automations are created so that the MQLs predicted by the machine learning model are automatically transferred to the sales team.

Measuring success

After commissioning, the success of the system is measured using KPIs. If the KPIs deteriorate, the machine learning model may need to be retrained.

Updating the lead scoring system

Just like traditional lead scoring systems, predictive systems also need to be updated regularly. To do this, it is advisable to identify the weaknesses of the current model and then train a new model. In addition, the process of data preparation and modeling can be adapted to eliminate the weaknesses of the current system.

7 Application of the process model to the Mautic software

This chapter first introduces the MAS Mautic. Subsequently, the previously created generic process model for developing a traditional and a predictive lead scoring system is applied to Mautic. It also examines how extended use cases can be mapped within the software.

7.1 Introduction of the Mautic software

Mautic is an open source MAS. The term open source software refers to programs whose source code is publicly accessible and enables a global community of developers to work together on improving the program (cf. Wu and Lin 2001: 33). The software is used by over 200,000 companies and supported by over 1,000 volunteers (cf. mautic.org n.d.). As described in chapter 2.5 a MAS must support the following lead management tasks and accordingly have the necessary components:

- Lead generation and identification
 - Landing pages
 - Forms
 - Social Media Marketing
- Lead Nurturing
 - Campaigns

- E-mail marketing
- Social Media Marketing
- Segmentation and dynamic content
- Lead qualification
 - Lead scoring system
- Lead Routing
 - CRM integration
- Conclusion and beyond
 - Reports

These individual components within the Mautic software are described below. A breakdown of the individual items can be found in the software menu (see Figure 6).

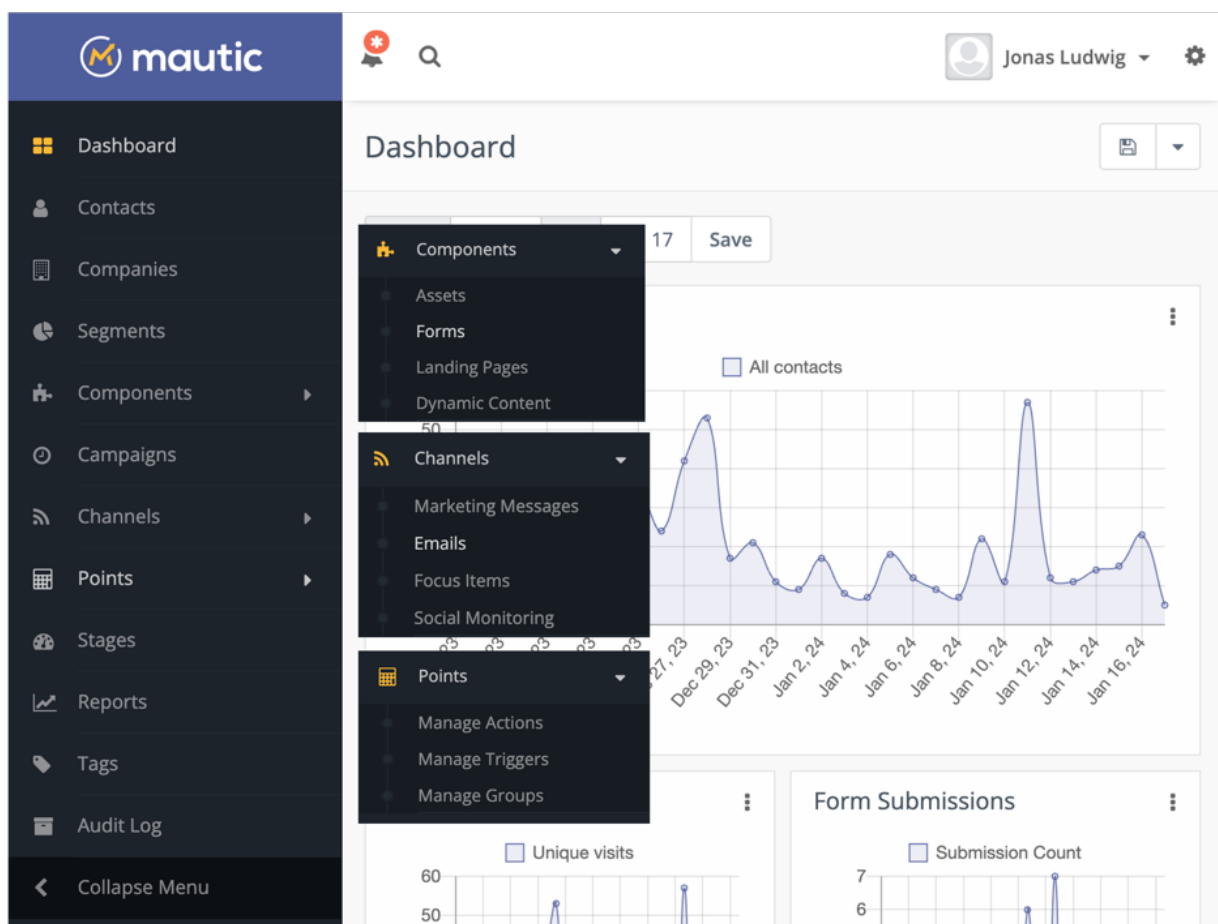


Figure 6: User interface of the Mautic software
Source: Own representation

Dashboard

The dashboard in Mautic is the central point of contact for users to get an overview of their marketing activities. It displays important key figures such as visitor numbers, email performance and other relevant statistics.

Contacts

The "Contacts" menu item in Mautic allows you to manage contacts. Leads can be added, edited and filtered according to various criteria here.

Companies

Companies can be used in Mautic to conduct account-based marketing. Leads are assigned to superordinate companies for this purpose. The company information can be managed under "Companies".

Segments

The "Segments" item in Mautic enables the targeted grouping of contacts based on shared characteristics or behavior. Segments enable precise targeting and can be used for personalized marketing campaigns. Users can create and customize segments according to their specific requirements to ensure effective management and analysis of contacts.

Components

The "Components" menu item in Mautic combines various marketing elements:

- **Assets:** Files such as PDFs are stored here, which can then be offered for download.
- **Forms:** This section allows you to create and manage forms that can be integrated into websites or emails to collect information from contacts.
- **Landing pages:** Here, a drag-and-drop builder can be used to design landing pages that serve to convert visitors into leads by prompting them to take a specific action.
- **Campaigns:** Campaigns allow individual marketing activities to be combined into comprehensive campaigns. This creates automated processes that combine various actions and events.

Channels

Channels comprise various communication channels for marketing activities:

- **Marketing Messages:** Marketing Messages in Mautic makes it possible to create content and make it available via various channels such as email, SMS, browser notifications, mobile notifications and tweets. Mautic then sends the messages via the preferred channels of the individual contacts and switches to an alternative channel if required.
- **Emails:** This area enables the creation, management and automation of emails via a drag-and-drop builder.
- **Focus items:** Here, targeted elements such as pop-ups or banners are created to attract the user's attention.

- Social monitoring: This area enables the monitoring of activities in social media in order to gain insights into the behavior and interests of the target group.

Points

The Points menu item is intended for lead scoring and contains three sub-items:

- Manage Groups: Several scores can be created under this item. This also enables more complex use cases such as product-based scoring or the use of an implicit and an explicit score.
- Manage Actions: The points system is mapped in this area by selecting specific actions and determining the associated number of points. The score that is increased or reduced by the actions is also selected.
- Manage Triggers: The actions that are executed as soon as a score exceeds a threshold value are selected here.

Stages

In Mautic, "stages" enable the definition of different phases in marketing campaigns. These can be stages in a sales funnel, for example, to which leads are assigned.

Reports

In Mautic, "Reports" enables the creation and customization of reports that provide an overview of KPIs and marketing activities. Users can configure reports according to their specific requirements to obtain the data and metrics relevant to their analysis.

Tags

In Mautic, "tags" are used to assign contacts to individual categories. These tags enable effective segmentation and targeting of contacts.

7.2 Traditional lead scoring in Mautic

In the following, the steps of the generic process model for creating a traditional lead scoring system from Chapter 6.1 are adapted to the Mautic software.

Creation of a service level agreement

The first step in the process of developing a traditional lead scoring system in Mautic is to define an SLA. This is drawn up jointly by Sales and Marketing and sets out the objectives and framework conditions of the project.

Data generation

In order to answer the questions relevant to lead scoring and to develop a robust lead scoring system, the following data must be generated:

- Lead ID for unique identification of a lead

- Date of lead capture
- Relevant explicit lead data
- List of lead activities with timestamp
- Lead status ("SAL", "Rejected", "No transfer ")
- Date of lead handover
- Date of the last lead activity

All this data can be automatically recorded within Mautic and stored in the database.

Data preparation

To prepare the data for data analysis, the following data preparation steps are carried out:

- Data selection: First, all data to be analyzed is selected from the generated data.
- Data cleansing: As part of data cleansing, data gaps are reduced or eliminated and outliers or incorrect values are removed.
- Data construction: Data construction creates new features from existing data sets, enabling improved data analysis. In addition, text-based values are converted into numerical values to enable effective analysis.
- Data integration: This involves creating a table from several tables or data records that contains all the data relevant for lead scoring. The data is also aggregated in a meaningful structure.
- Data formatting: In some cases, the format or structure of the data is changed at the end of data preparation.

Data analysis

The "Dashboard" and "Reports" menu items in Mautic can be helpful for data analysis. However, it was found that the data analysis functions are not sufficient to answer the questions relevant to lead scoring. In addition, Mautic offers only a few statistical tools for analyzing the collected data. For this reason, in this process model the relevant data is exported from the software database using SQL commands and analyzed using external tools outside of Mautic. Python code is used for this purpose in this thesis, but it is also possible to use external programs for data analysis. The data is processed there and then analyzed using univariate and bivariate data analysis. In particular, the question of how the activities of accepted, rejected and unsubmitted leads differ is answered. To answer this question, the characteristics and touchpoints passed through are examined for all accepted leads. For example, it can be determined how many pricing pages SALs visited on average before the handover. In addition, histograms, standard deviations and other statistical tools are used to gain more detailed insights into the relationship between reaching SAL status and visiting pricing pages and other actions. Then, the same analysis is performed for

declined and non-passed leads to find out how the activities and characteristics of SALs and unqualified leads differ. Other questions to be answered by the data analysis are as follows:

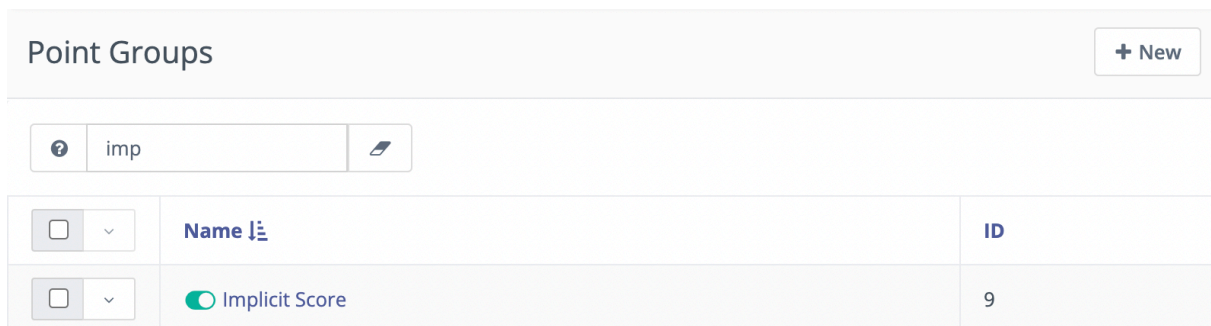
- How many touchpoints do leads go through before they become MQLs?
- How long does it take on average for leads to become MQLs?

Developing the points system

Once sufficient insight into the data has been gained to develop a realistic point model, a scorecard is developed using data-supported expert estimates. The score model is then created in Mautic. This is done via the menu item "Manage Actions".

Creating the scores

Point groups can be created in Mautic under the "Manage Groups" menu item. These are additional scores that are created alongside the predefined standard score. For reasons of efficiency, the standard score is used for the explicit score and the "Implicit Score" point group is created for the implicit score (see Figure 7).



Point Groups		+ New
<input type="text" value="imp"/>		
<input type="checkbox"/> Name ↓	ID	
<input type="checkbox"/> Implicit Score	9	

Figure 7: Point Groups in Mautic
Source: Own representation

Determining the implicit scoring

Once all the required scores have been created, actions that increase or decrease the implicit score can be defined under "Manage Actions". To do this, the specific actions, the number of points to be deducted or added and the corresponding score are selected. In the example in Figure 8 ten points are added to the implicit score when a contact completes and submits the "Lead Scoring Test Form".

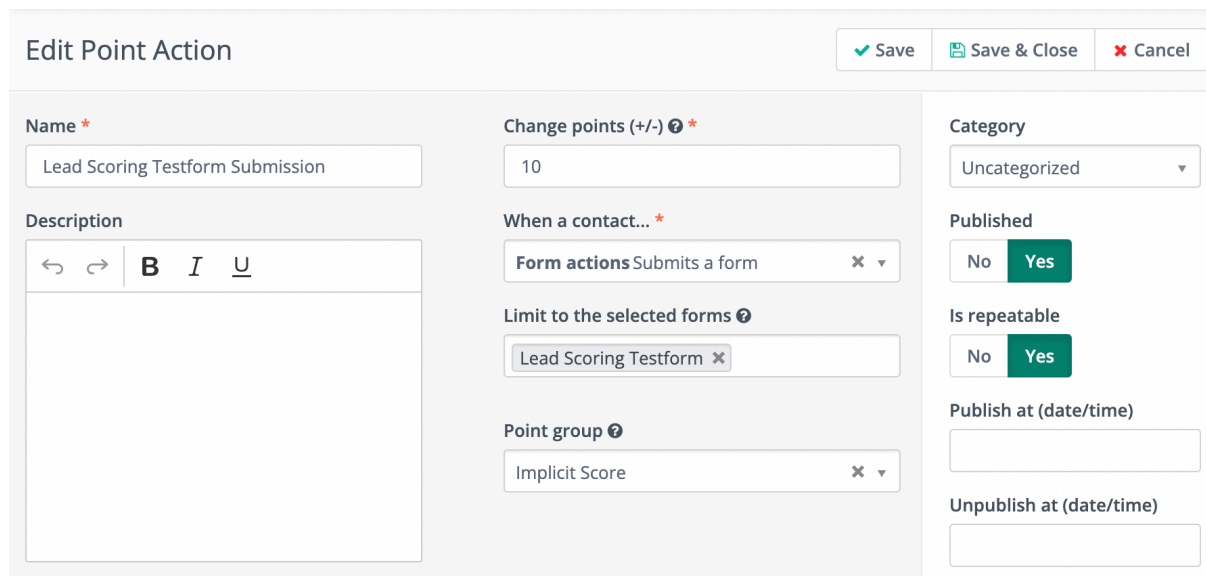
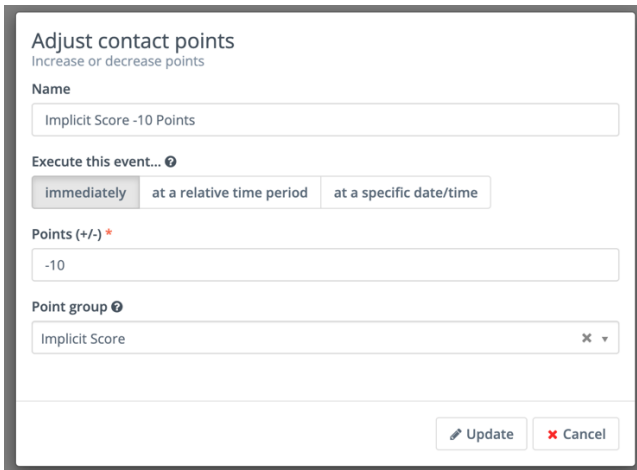


Figure 8: Point Actions in Mautic
Source: Own representation

Under "Manage Actions", the following actions can be selected to increase or decrease the implicit score:

- Download an asset
- Receipt of an e-mail
- Opening an e-mail
- Sending a form
- Visit to a landing page or URL

Consequently, not all actions relevant to the implicit score (cf. chapter 3.1) can be taken into account. For example, the unsubscription of a contact from the newsletter cannot be taken into account under "Manage Actions". However, this action can still be intercepted by using campaigns. To do this, all contacts who have unsubscribed are first added to the "Do Not Contact" segment using a segment filter. A campaign is then started for this segment. In this campaign, the desired number of points is deducted from each contact in the segment using the "Adjust contact points" campaign action (see Figure 9). In addition, the campaign settings specify that the campaign can only be run once per contact to ensure that points are not deducted multiple times for a single unsubscribe from the newsletter.



Adjust contact points
Increase or decrease points

Name
Implicit Score -10 Points

Execute this event... ⓘ
 immediately
 at a relative time period
 at a specific date/time

Points (+/-) *
-10

Point group ⓘ
Implicit Score ✕ ▾

Figure 9: Campaign action "Adjust contact points"
Source: Own representation

Determining the explicit scoring

As scores within the "Manage Actions" menu item can only be increased on the basis of actions, the explicit score, which evaluates the characteristics of leads, must be adjusted via campaigns. It must also be updated regularly to take account of changes in the demographic and company-related characteristics of a contact.

For this purpose, all contacts that are taken into account in lead scoring are first included in the "scoring segment". Based on this segment, a campaign is started to calculate the explicit score (see Figure 10). After the campaign is started, the explicit score of the contacts is reset to zero. The new explicit score is then calculated based on the lead properties. After the explicit lead score has been calculated in this way, the system waits 24 hours before calculating it again.

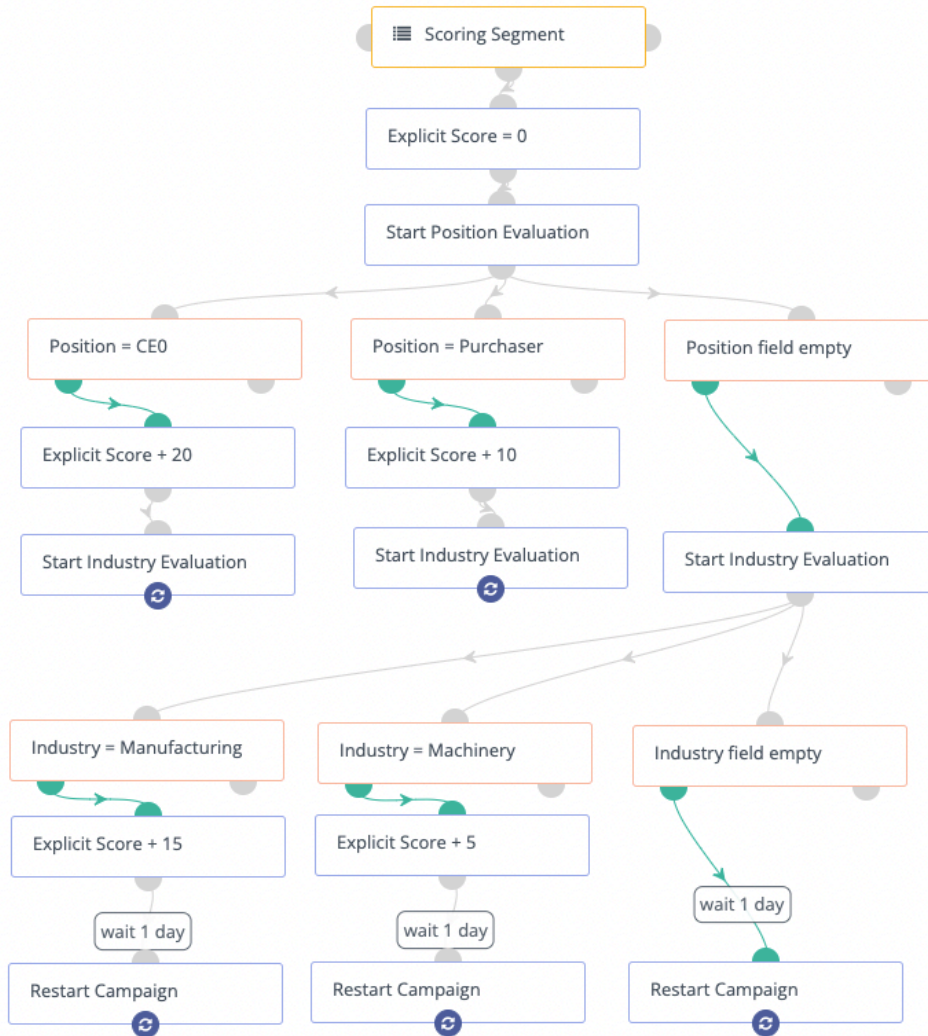


Figure 10: Campaign for calculating the explicit score
Source: Own representation

Creating a suppression segment

To ensure that irrelevant contacts are not passed to the sales team, a suppression segment called "Do Not Score" is created. All contacts that should not be included in lead scoring are added to this segment, for example employees, competitors, students or contacts that have been disqualified by the sales team. Contacts who are in the segment can still collect points later, but no actions are triggered when the threshold value is reached.

Creating a decay model

In chapter 6 three expiry models were determined, which ensure that only leads with current interest have high implicit Scores. Of these three models, only the reduction of a contact's points after a longer period of inactivity is technically possible at the current state of the Mautic software. As shown in Figure 11 the first step is to check when a lead was last active. If the last activity was 30, 60 or 90 days ago, points are deducted. The system then waits one

day so that the campaign is not restarted on the same day and the points for inactivity are deducted multiple times.

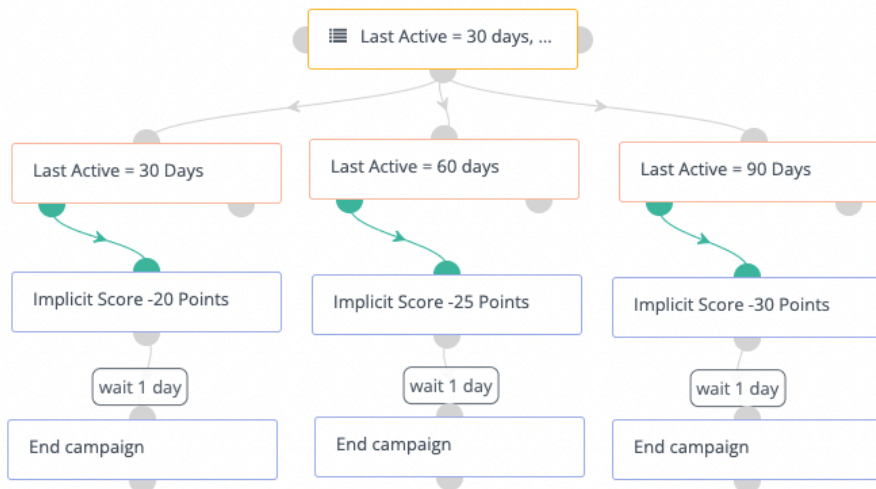


Figure 11: Campaign to implement a forfeiture model
Source: Own representation

Create a campaign to reset negative scores

A campaign to reset negative scores is also being developed as part of an expiry model. This ensures that contacts who have had points deducted due to prolonged inactivity do not receive a negative score. Otherwise, there is a risk that a renewed increase in interest will not be recognized, as additional points must be collected to reach the threshold value in order to compensate for the negative score.

The campaign is started on the basis of a segment to which leads with a negative implicit score are added. After the start, one point is iteratively added to the implicit score and checked again to see whether the score is still negative. If this is the case, another point is added until the score is zero (see Figure 12).

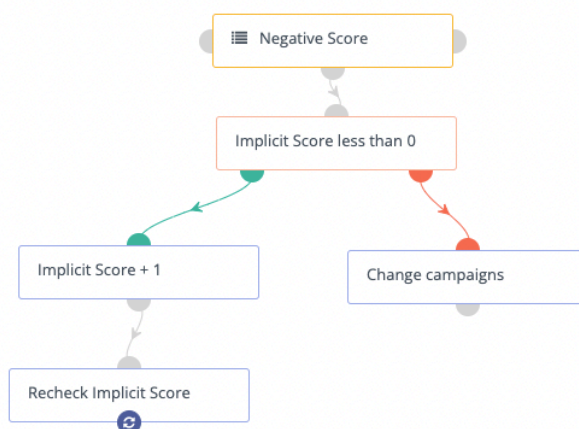


Figure 12: Campaign to reset negative implicit scores

Source: Own representation

Defining a threshold value

After the scoring system has been created in Mautic, implicit and explicit threshold values are determined, from which leads become MQLs. The ideal threshold values for the implicit and explicit score are determined retrospectively using Python code. This involves calculating the scores of past accepted and rejected leads at the time of handover. The score of leads that were not handed over at the time of their last activity is also determined. After an analysis, the implicit and explicit thresholds are selected for which the highest number of leads is correctly predicted.

Customizing the lead scoring system

Once the model has been created, it is adjusted iteratively. This involves adjusting the threshold values, the scoring and, if necessary, the expiry model and retrospectively determining whether previous MQLs are predicted more accurately. After several runs, the parameters with which the model achieves the best results are selected.

Evaluation of the lead scores

Once the lead scoring system has been completed, leads that exceed the threshold value are evaluated. The following steps are carried out for this purpose.

Lead handover

As soon as a lead exceeds the implicit and explicit thresholds, it is transferred to the sales team. For this purpose, a start segment is created in which leads that have exceeded both thresholds are transferred. Based on this, a campaign is triggered to hand over the lead. The responsible sales team is informed by email that a new MQL has been received. The sales team responsible for the contact is determined via synchronization with the CRM software. The CRM software uses characteristics of the individual leads, such as region and industry, to automatically determine which sales team is responsible. Integration with the CRM software also allows the data collected during lead scoring to be transferred to the CRM software so that it is available to the sales team after the handover. In addition, the contacts are tagged with "Routed" and automatically added to the feedback loop campaign (cf. Figure 13).

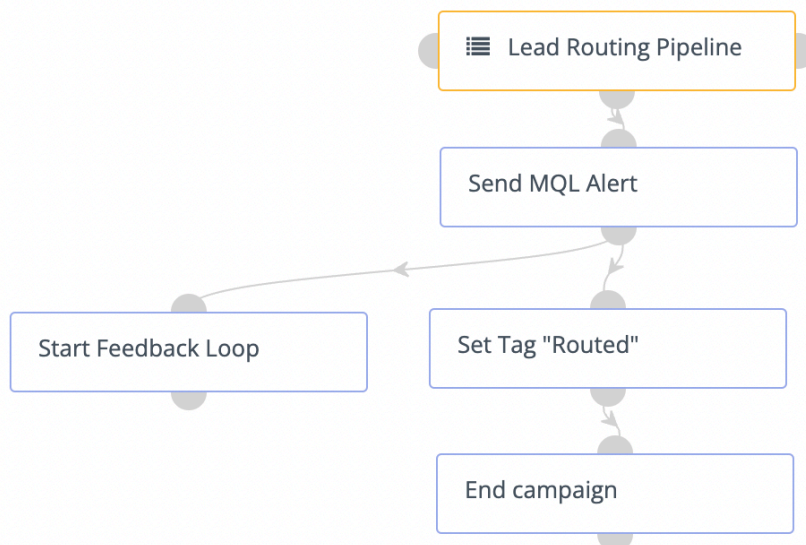


Figure 13: Campaign for the transfer of leads
Source: Own representation

It should be noted that in addition to the handover to the sales team, alternative actions can be executed in Mautic when the threshold value is reached. Some of the actions available in Mautic are as follows:

- Sending an e-mail to the contact
- Remove or add the contact from campaigns
- Removing or adding the contact from segments
- Removing or adding tags
- Transfer of the contact to an integration

Obtaining sales feedback

Once the leads have been handed over to the sales department, feedback is obtained from the sales staff on the quality of the individual leads. This feedback supports the creation of KPIs and at the same time enables the lead scoring system to be optimized. Sales employees can indicate whether a lead is accepted and receives SAL status, whether the lead is not yet ready for the sales process and should therefore be recycled, or whether the lead is rejected. It is advisable to define a feedback process and integrate it into the CRM system that is synchronized with the marketing automation software. Alternatively, feedback can also be generated through a feedback loop campaign in Mautic, which is discussed in more detail in Appendix 1.

Evaluation of sales feedback

As soon as the sales feedback is available, the contact is added to a campaign to evaluate the feedback. There are three paths for this (see Figure 14):

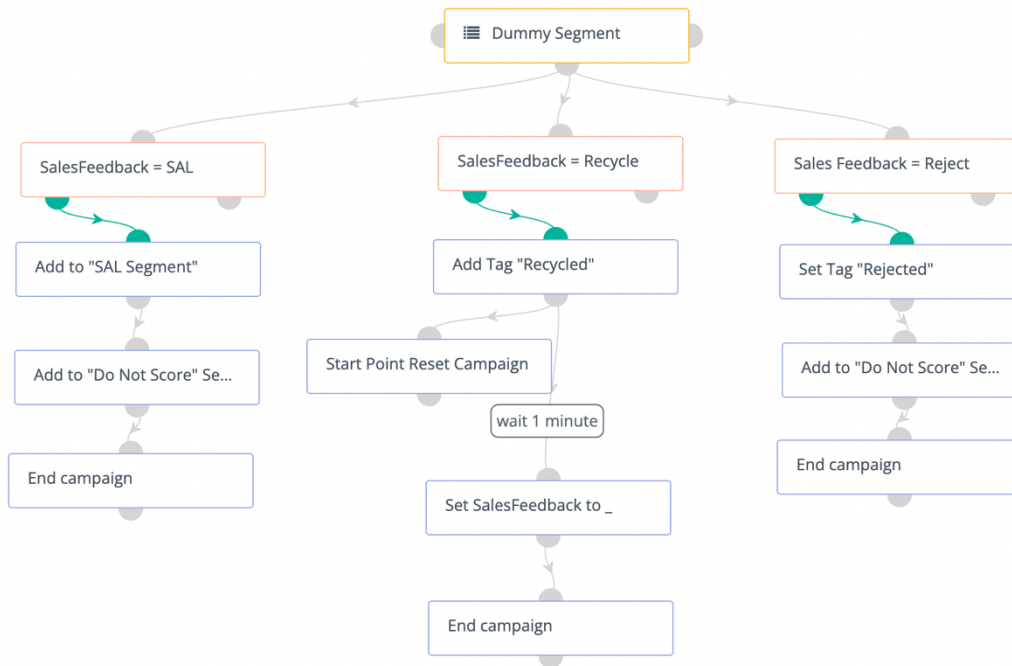


Figure 14: Campaign for feedback evaluation
Source: Own representation

- Success path: The success path is taken when the "SAL" feedback is given. In this case, the lead is first added to the "Do Not Score" segment to prevent it from being passed again. The lead is also added to a segment for SALs. Since not every SAL converts to a customer, another campaign is developed that integrates SALs that cannot be converted in the sales process back into the lead scoring process. For this purpose, a field called "Reactivate_SAL" is created in Mautic, which is synchronized with the CRM software. This field can have the values "Yes" or "No". If it is now set to "Yes" by a sales employee, a campaign is started to reintegrate the SALs into the lead scoring process. The tag "Reactivated_SAL" is first added to the contact. The implicit score is then reset to a previously defined value. The feedback from Sales is also reset. In addition, the lead is removed from the "SAL segment" and the "Do Not Score segment" (see Figure 15).
- Recycling path: If "Recycle" is given as feedback, the lead is currently not ready for processing by the sales team. In this case, the lead is marked with the "Recycled" tag and returned to Marketing. In addition, the lead's points are reset to a previously defined value and the content of the "SalesFeedback" field is emptied.
- Disqualification path: If the "Reject" feedback is submitted, the lead is completely rejected by Sales. Rejected leads are added to the "Do Not Score Segment" and receive the "Rejected" tag.

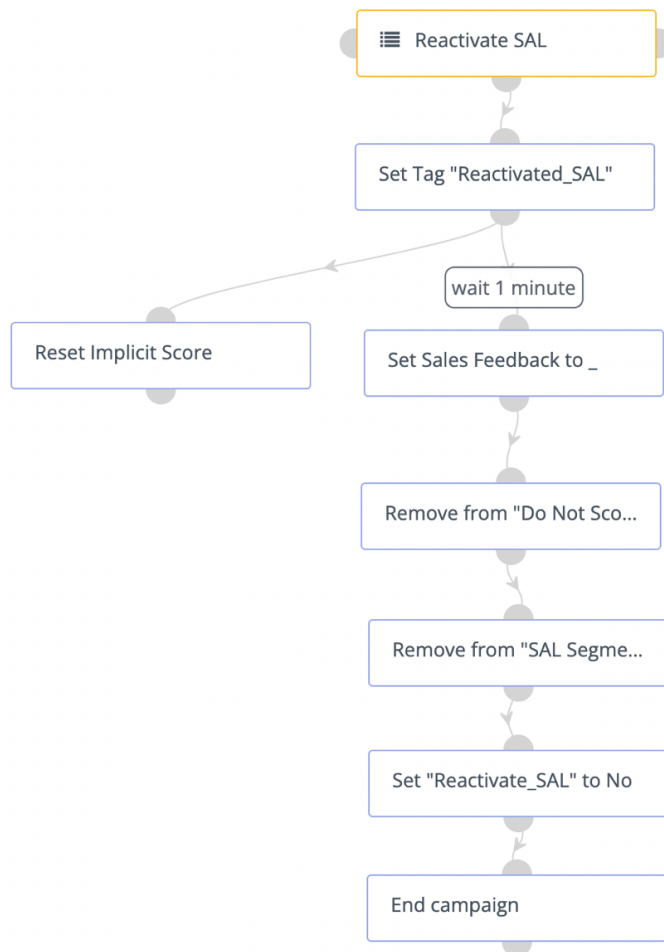


Figure 15: Campaign to reactivate leads
Source: Own representation

Measuring success

Once the system has been put into operation, KPIs are used to measure its success. If the KPIs deteriorate, this may be a sign that the lead scoring system needs to be updated or improved.

Updating the lead scoring model

The lead scoring model is updated at least every three months. For this purpose, the steps for data preparation and data analysis are carried out again and the system is updated accordingly.

7.3 Predictive lead scoring in Mautic

In this chapter, the steps of the generic process model for creating a predictive lead scoring system from chapter 6.2 are adapted to the Mautic software.

Creation of a service level agreement

The first step in the process of developing a predictive lead scoring system in Mautic is to define an SLA. This involves defining the objectives and framework conditions of the lead scoring system between marketing and sales.

Data generation

This step ensures that all the required data is available in the Mautic database. This involves the following data:

- Lead ID for unique identification of a lead
- Date of lead capture
- Relevant explicit lead data
- List of lead activities with timestamp
- Lead status ("SAL", "Rejected", "No transfer")
- Date of lead handover
- Date of the last lead activity

Data preparation

In order to prepare the data for the development of a machine learning model, the data is exported from Mautic via CSV export and prepared there using Python code. This involves going through the steps of data selection, data cleansing, data construction, data integration and data formatting.

Modeling

After data preparation, a robust machine learning model is developed. Before the actual model development, various models or machine learning algorithms are first selected to be tested in the context of lead scoring. In the context of predictive lead scoring, four algorithms are used that are increasingly used in the literature on predictive lead scoring. These are the algorithms Logistic Regression, Decision Tree, Random Forest and Support Vector Machine. A grid search is carried out with the algorithms to develop the model. This means that the model is trained and tested with different combinations of the defined hyperparameters in order to find the best settings. Once the grid search has been completed for each model, the best model with the best hyperparameters is saved for later use so that it can be used to predict new data in the future without having to train a model again. The best model here is calculated by the accuracy of the models, which indicates what percentage of a model's predictions are correct.

Model evaluation

This evaluation phase looks for weaknesses in the model and its development process. It also assesses whether the model meets the objectives set out in the SLA. The process of providing the model is also planned.

Model implementation

The individual steps for applying the model to Mautic are described below:

Export of lead data from Mautic

In the first step of the model application, all leads that have not been converted and are not in the suppression segment are exported from Mautic as a CSV.

Data preparation

The lead data is then prepared using the Python code created in the data preparation step.

Creating the predictions

The previously saved machine learning model is then loaded to make predictions for each lead. The predictions are saved together with the corresponding lead ID as a CSV file.

Transferring the model predictions to Mautic

Once the predictions have been made, the CSV file with the model predictions is imported into Mautic. For this purpose, the lead_id column is assigned to the contact ID in Mautic during import and the column with the predictions is assigned to a previously created database field with the name "Scoring Predictions" (see Figure 16).

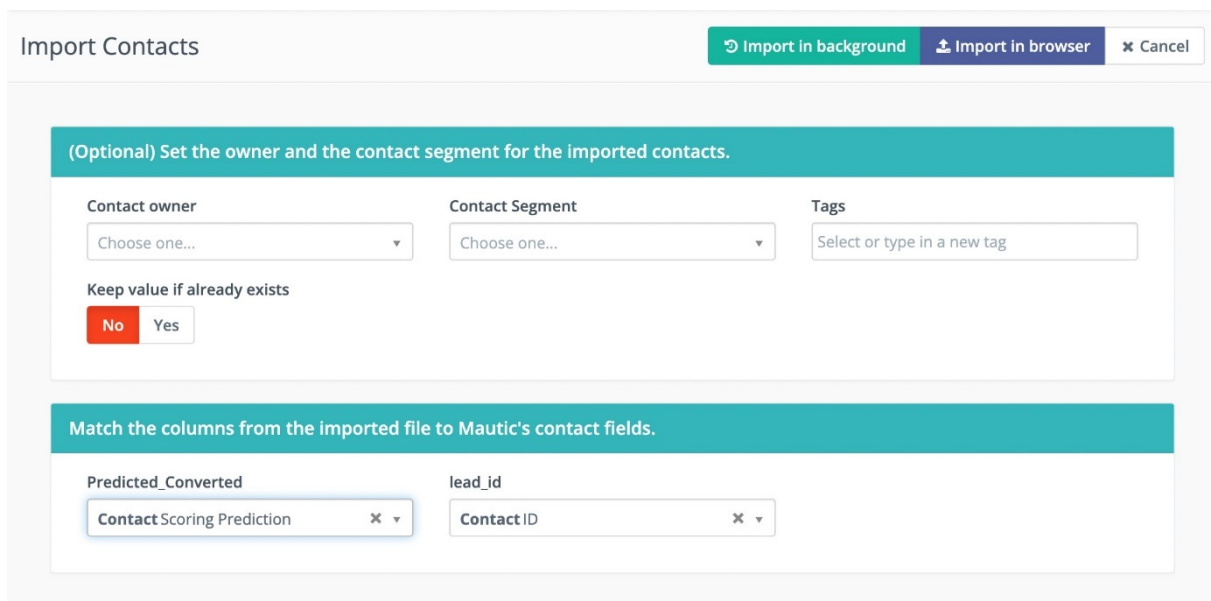


Figure 16: Importing the predictions of a predictive lead scoring model into Mautic
Source: Own representation

Lead handover

Contacts classified as MQLs by the machine learning model are then automatically forwarded to the sales team via a campaign.

Evaluating the leads

As soon as a lead has been transferred, the mechanism for obtaining feedback on the quality of the transferred lead from the sales team starts. This mechanism can be implemented either in the CRM system or in the MAS (see Appendix 1).

A campaign is then launched to evaluate the feedback. Depending on the feedback from Sales, the SAL path, the recycling path or the disqualification path is selected. An additional campaign is created for the SAL path, which can be used to reintegrate SALs that could not be acquired as customers into the lead scoring process for the corresponding product group.

To ensure that recycled and reactivated leads are not immediately predicted as MQLs again, it is advisable to wait for a predetermined period of time before recycling or reactivating them. This changes their activity profile, providing the machine learning model with updated data for lead scoring.

Measuring success

After commissioning, the success of the system is measured using KPIs. If the KPIs deteriorate, the machine learning model may need to be retrained.

Updating the lead scoring system

Just like traditional lead scoring systems, predictive lead scoring models also need to be updated regularly. To do this, it is advisable to identify weaknesses in the current model and then train a new model. If necessary, a modified data preparation and modeling process can be used to eliminate the weaknesses of the old model.

7.4 Product-based lead scoring in Mautic

In the course of this research work, the use case of product-based lead scoring was also implemented in Mautic. It should be noted that a separate lead scoring system must be developed for each product group. Depending on the preferred approach, the creation of these individual systems can either follow the steps for designing a traditional lead scoring system (see chapter 7.2) or the steps for developing a predictive lead scoring system (see chapter 7.3). As several product areas are integrated into lead scoring, several SLAs may need to be created with different sales teams.

7.5 Account Based Scoring in Mautic

In addition to traditional lead scoring and product-based scoring, ABS was also examined in Mautic. The traditional scoring system can be mapped here with the help of campaigns (see

Figure 17). However, when testing the ABS functionalities of the software, it was found that there is no option within Mautic to execute actions based on the score of an account. Even with predictive ABS, it is possible to export and evaluate all properties and actions assigned to a company and import the model predictions, but here too the problem is that it is not possible to start actions at company level based on the predictions of the model. Therefore, the ABS functionalities in Mautic are not sufficient to enable an automated scoring process.

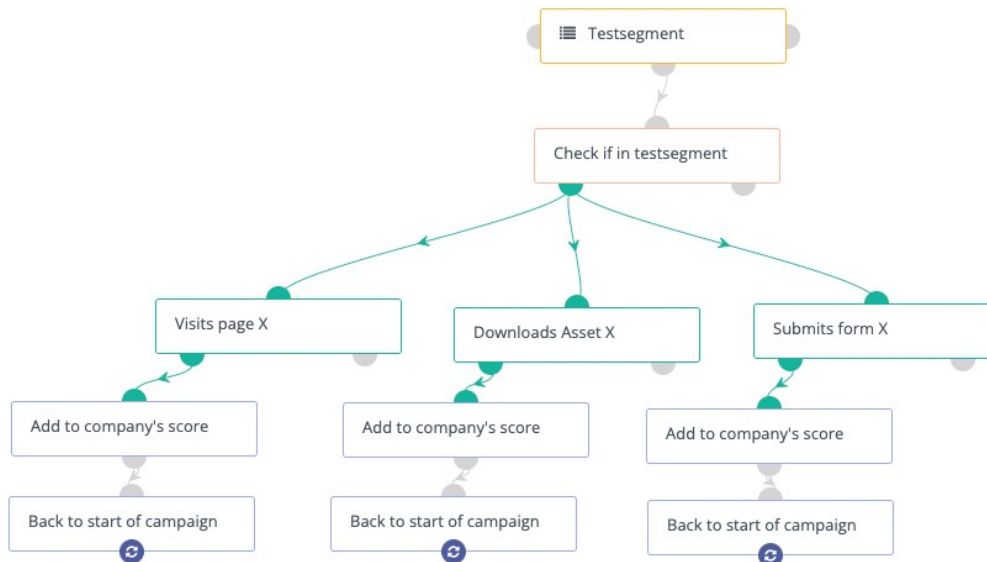


Figure 17: Campaign for calculating the implicit score in account-based scoring
Source: Own representation

8 Practical implementation and results of lead scoring in Mautic

In this part of the thesis, the application of the previously determined process model for the development of a lead scoring system is applied to a real online store that uses MAS Mautic. Both a traditional and a predictive lead scoring system are developed and their results are compared using the online store as an example.

8.1 Traditional lead scoring

Creation of a service level agreement

As there is no contact with the online store's marketing and sales team, the creation of an SLA was not carried out as part of the practical example.

Data generation

The data that is already in the Mautic database of the online store is used as the database for lead scoring. This includes all website visits including timestamps and, if a contact has

unsubscribed from the newsletter, the date of unsubscription. In addition, the segment affiliations of the contacts can be used to determine whether they are converted or non-converted leads.

Data preparation

Once the data has been generated, it is processed within Python. The Python code from Appendix 2 is used for this. The individual steps of data preparation are described below.

Data selection

In this step, the data of converted and non-converted leads is first exported as CSV files using two SQL statements (see Appendix 3 and 4). An anonymized version of the table for converted leads is shown in Table 11 and contains the following columns:

- `lead_id`: The `lead_id` for is used to uniquely identify a contact.
- `end_date`: The `end_date` represents the time at which a contact is converted to a customer or the time of the last activity for non-converted leads.
- `url`: The website visited is specified in the URL column for each activity entry.
- `date_hit`: The `date_hit` is the time stamp for visiting a website.
- `start_date`: The `start_date` describes the time at which a lead was recorded in Mautic.
- `confirmed_optout_date`: The date of unsubscription from the newsletter is saved here, if an unsubscription has taken place.

lead_id	end_date	url	date_hit	start_date	confirmed_optout_date
84654	2021-09-08 06:10:04	https://www.mywebsite.com/.....	2022-01-01 02:40:14	2021-03-01 22:22:37	
84654	2021-09-08 06:10:05	https://www.mywebsite.com/.....	2022-01-01 02:40:15	2021-03-01 22:22:38	
84654	2021-09-08 06:10:06	https://www.mywebsite.com/.....	2022-01-01 02:40:16	2021-03-01 22:22:39	
84654	2021-09-08 06:10:07	https://www.mywebsite.com/.....	2022-01-01 02:40:17	2021-03-01 22:22:40	
84654	2021-09-08 06:10:08	https://www.mywebsite.com/.....	2022-01-01 02:40:18	2021-03-01 22:22:41	
84654	2021-09-08 06:10:09	https://www.mywebsite.com/.....	2022-01-01 02:40:19	2021-03-01 22:22:42	
21544	2021-09-08 06:10:10	https://www.mywebsite.com/.....	2022-01-01 02:40:20	2021-03-01 22:22:43	
21544	2021-09-08 06:10:11	https://www.mywebsite.com/.....	2022-01-01 02:40:21	2021-03-01 22:22:44	
21544	2021-09-08 06:10:12	https://www.mywebsite.com/.....	2022-01-01 02:40:22	2021-03-01 22:22:45	
21544	2021-09-08 06:10:13	https://www.mywebsite.com/.....	2022-01-01 02:40:23	2021-03-01 22:22:46	
21544	2021-09-08 06:10:14	https://www.mywebsite.com/.....	2022-01-01 02:40:24	2021-03-01 22:22:47	
84587	2021-09-08 06:10:15	https://www.mywebsite.com/.....	2022-01-01 02:40:25	2021-03-01 22:22:48	
84587	2021-09-08 06:10:16	https://www.mywebsite.com/.....	2022-01-01 02:40:26	2021-03-01 22:22:49	
84587	2021-09-08 06:10:17	https://www.mywebsite.com/.....	2022-01-01 02:40:27	2021-03-01 22:22:50	
84587	2021-09-08 06:10:18	https://www.mywebsite.com/.....	2022-01-01 02:40:28	2021-03-01 22:22:51	
84587	2021-09-08 06:10:19	https://www.mywebsite.com/.....	2022-01-01 02:40:29	2021-03-01 22:22:52	
84587	2021-09-08 06:10:20	https://www.mywebsite.com/.....	2022-01-01 02:40:30	2021-03-01 22:22:53	
84587	2021-09-08 06:10:21	https://www.mywebsite.com/.....	2022-01-01 02:40:31	2021-03-01 22:22:54	
84587	2021-09-08 06:10:22	https://www.mywebsite.com/.....	2022-01-01 02:40:32	2021-03-01 22:22:55	
84588	2021-09-08 06:10:23	https://www.mywebsite.com/.....	2022-01-01 02:40:33	2021-03-01 22:22:56	2022-01-01 02:40:35

Table 11: DataFrame of converted leads
Source: Own representation

After reading the data into Python, a column with the name "Covered" is first added to both tables or DataFrames (DFs). A "1" is added here for the entries from the table of converted leads and a "0" for non-converted leads. The two tables are then merged.

Data cleansing

As the focus in the application example is on predicting conversion to new customers, actions that took place after the first purchase are not relevant. Therefore, website visits and unsubscribes from the newsletter after the purchase date are removed.

Data construction

In order to enable a meaningful data analysis, new data is constructed in this step. The `confirmed_optout_date` column either contains a date if a lead has unsubscribed from the newsletter or is empty if a lead has not unsubscribed. As part of the coding, all data is converted to a "1" and all empty values are converted to a "0". The `salescycle_duration` column is also created, in which the duration is determined for each lead until it receives the MQL status. The difference between the `start_date` and the `end_date` is calculated for this. For non-converted leads, the period between the capture and the last activity is calculated.

Data integration

As part of the analysis of the URL column, it was determined that a total of 2327 different websites were visited by the contacts. In order to reduce the complexity of lead scoring and guarantee an accurate evaluation, the URLs visited are assigned to different URL categories. For this purpose, the categories "Basket", "Browsing 1", "Browsing 2", "Browsing 3", "Checkout", "Checkout Registration", "Contact", "Faq", "Account", "Registration", "Newsletter Registration", "Price", "Product" and "Other Page" were created after analyzing the values in the URL column. The URL column is divided into URL categories using keywords. As soon as each URL has been replaced by a category, a column is created in a new DF for each URL category in which it is recorded how often a lead has visited the URL category. In addition, another column is created that aggregates the total website visits per lead.

The total website visits, the website visits per URL category, the `salescycle_duration`, the "Converted" column and the newsletter unsubscription status are then merged together with the lead ID in the DF "Lead overview", the first lines of which are shown in Table 12 can be seen in Table 12.

lead_id	total_pagehits	Basket	Browsing 1	Browsing 2	Browsing 3	Checkout	Checkout-Registration	Contact	Faq	Account	Newsletter-Registration	Other Page	Price	Product	Registration	salescycle_duration	Converted	opted_out
1	6	0	0	0	0	0	0	0	0	0	0	0	0	6	0	1018,62	0	0
2	20	1	0	0	0	2	0	0	1	0	0	2	0	14	0	811,45	1	0
3	4	0	0	0	0	0	0	0	0	0	0	0	0	4	0	796,98	0	0
4	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	796,96	0	0
5	12	2	0	0	0	4	0	0	0	0	0	0	0	6	0	796,95	0	1
6	4	0	0	0	0	0	0	0	0	0	0	0	0	4	0	796,94	0	0
7	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	796,94	0	0
8	4	0	0	0	0	0	0	0	0	0	0	0	0	4	0	796,89	0	0
9	765	44	6	37	10	77	13	6	4	61	0	84	13	406	4	796,88	0	0
10	14	0	0	1	0	0	0	0	0	2	0	6	0	5	0	796,88	0	0
11	3	0	0	0	0	0	0	0	0	0	0	0	0	2	1	796,87	0	0
12	32	0	0	0	0	0	0	0	0	0	0	0	0	32	0	796,87	0	0
13	8	0	0	0	0	0	0	0	0	0	0	0	0	8	0	796,85	0	0
14	117	7	0	0	7	22	6	0	0	11	0	21	0	43	0	796,84	0	0
15	25	2	0	0	0	0	0	0	0	0	0	0	0	23	0	796,84	0	0
16	4	0	0	0	0	0	0	0	0	0	0	0	0	4	0	796,84	0	0
17	4	0	0	0	0	0	0	0	0	0	0	0	0	4	0	796,83	0	0
18	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	796,83	0	0
19	52	7	0	6	0	9	1	0	0	4	1	4	1	19	0	796,82	0	0
20	80	9	0	1	0	25	0	0	0	3	0	4	0	38	0	796,81	0	0
21	15	0	0	0	0	0	0	0	0	0	0	0	0	15	0	796,8	0	0
22	10	0	0	0	6	0	0	0	0	0	0	1	1	2	0	796,8	0	0
23	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	796,79	0	0
24	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	796,79	0	0
25	4	0	0	0	0	0	0	0	0	4	0	0	0	0	0	796,78	0	0
26	41	2	2	2	0	2	0	0	0	1	0	0	1	31	0	796,77	0	0
27	25	0	2	0	0	0	0	0	0	0	0	2	0	14	0	796,77	0	0

Table 12: Lead overview DataFrame
Source: Own representation

As part of the data cleansing process, outliers are subsequently removed from the DF lead overview to prevent them from influencing the analysis. These are leads that have an excessively high level of activity and have made more than 300 website visits. In addition, leads with fewer than three website visits are deleted, as their value is considered low in the context of the data analysis.

Data formatting

It is not necessary to format the data for this application.

Data analysis

To analyze the data, the lead overview DF is divided into two DFs for converted and non-converted leads in order to examine them separately as part of the analysis.

For each column, apart from the "opted_out" column in the two DFs, the average, standard deviation, median and maximum are calculated. The resulting values are summarized in a statistics table. In addition, the "opted_out" column is used to determine the percentage of contacts who have unsubscribed from newsletters, both for converted and non-converted leads. This is added to the statistics table (see Table 13).

		Converted	Unconverted
Basket	Average	1,67	0,78
	Stdev	3,77	2,81
	Median	0	0
	Max	42	42
Browsing 1	Average	0,29	0,13
	Stdev	1,55	0,69
	Median	0	0
	Max	31	12
Browsing 2	Average	1,9	0,76
	Stdev	4,74	3,56
	Median	0	0
	Max	53	73
Browsing 3	Average	0,84	0,62
	Stdev	2,14	2,11
	Median	0	0
	Max	27	50
Checkout	Average	2,34	1,32
	Stdev	5,19	4,96
	Median	0	0
	Max	58	88
Checkout-Registration	Average	0,22	0,09
	Stdev	0,79	0,51
	Median	0	0
	Max	8	9
Contact	Average	1,68	0,55
	Stdev	3,15	3,72
	Median	0	0
	Max	15	67
Faq	Average	0,09	0,04
	Stdev	0,41	0,28
	Median	0	0
	Max	5	8
MyAccount Page	Average	2,81	0,9
	Stdev	6,16	4,56
	Median	0	0
	Max	71	81
Newsletter Registration	Average	0,03	0,08
	Stdev	0,29	0,48
	Median	0	0
	Max	4	10
Other Page	Average	7,97	3,68
	Stdev	10,41	11,16
	Median	4	0
	Max	93	134
Price	Average	0,24	0,4
	Stdev	1,26	1,36
	Median	0	0
	Max	29	30
Product	Average	7,93	12,84
	Stdev	11,46	21,28
	Median	5	6
	Max	139	225
Registration	Average	0,28	0,08
	Stdev	1	0,54
	Median	0	0
	Max	8	12
salescycle_duration	Average	230,26	657,55
	Stdev	142,79	187,24
	Median	258,5	763,8
	Max	811,45	1018,62
total_pagehits	Average	28,29	22,27
	Stdev	31,89	38,76
	Median	18	8
	Max	282	294
opted_out	Percentage opted out	0,22	3,14

Table 13: Statistics table for data analysis
Source: Own representation

Looking at the statistics table, it can be seen that there are no actions that clearly indicate an increased likelihood of purchase. However, converted leads tend to carry out more actions. URLs in the categories "Basket", "Browsing 1-3", "Checkout", "Checkout Registration",

"Contact", "Faq", "Other Page" and "Registration" tend to be accessed more frequently by converted leads. URLs in the categories "Newsletter registration", "Price" and "Product" tend to be accessed more frequently by customers who do not convert. In addition, almost exclusively customers who do not convert unsubscribe from the newsletter. In addition, the likelihood of a contact becoming a customer decreases if they were registered a long time ago.

In addition to the statistics table, a correlation analysis of the individual columns from the lead overview DF with the conversion to the customer was carried out. The individual correlations are Figure 18 can be seen in Figure 18.

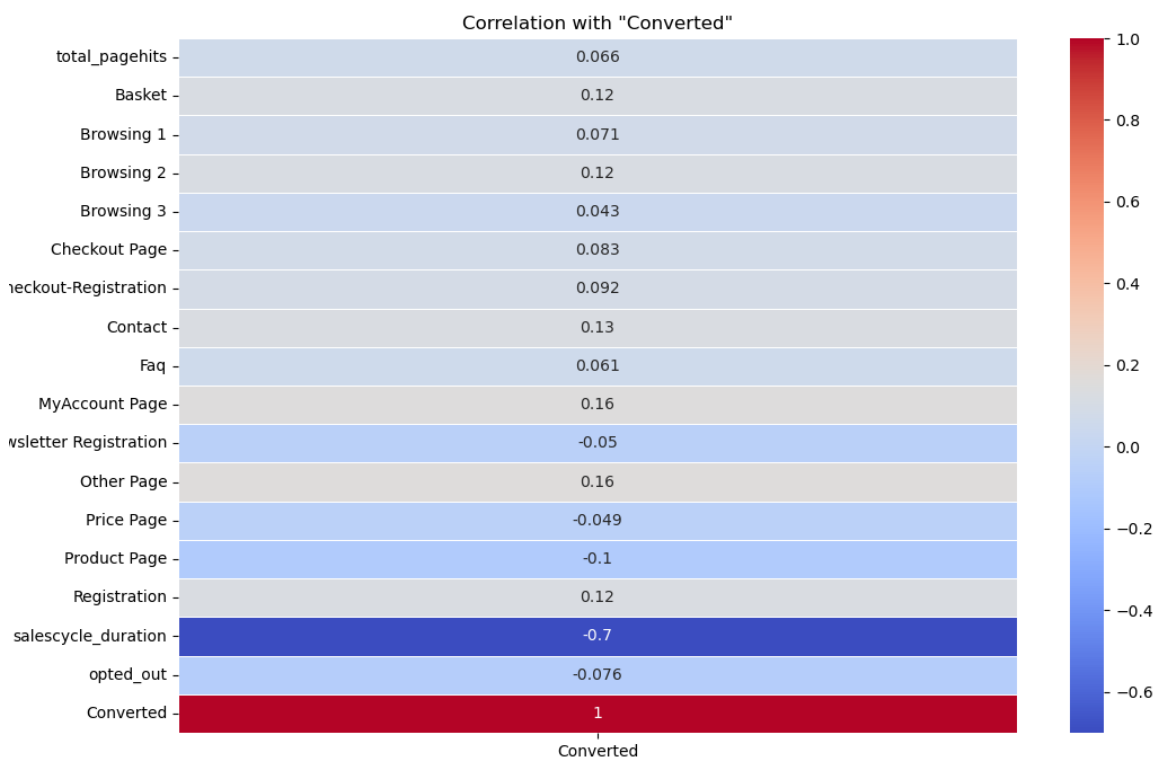


Figure 18: Correlation analysis with the lead status
Source: Own representation

The results from the correlation analysis confirm the conclusions drawn from the statistics table. The code for the data analysis can be found in Appendix 5.

Development of the points system

The following steps are carried out to create the points system:

Creating the scores

For scoring in the business-to-consumer (B2C) or e-commerce sector, it should be noted that, compared to the B2B sector, other explicit parameters such as gender, age and place of residence are generally relevant (cf. Wuttke n.d.). However, as these parameters are not available in the online store's database, only an implicit score is created.

Determining the implicit scoring

When determining the scoring, a decision is first made based on the data analysis as to whether points should be added or deducted for an action. Points are then awarded for the individual actions on a scale of one to ten. Points that are indicators of an immediate interest in buying or no interest in buying were also multiplied by a factor of 2.5. Website visits of the type "checkout" and "checkout registration" were selected as indicators of a current interest in purchasing. Although these do not emerge from the data analysis as clear purchase indicators, they are the last step in the purchase process after a customer clicks on "Buy" in the shopping cart. They therefore indicate that there is an interest in buying and that the contact was already close to making a purchase. One action that indicates that there is no interest in buying is unsubscribing from the newsletter. The points system selected is shown in Table 14 can be seen in Table 14.

Pagehit Type	Stats	Converted	Unconverted	Ratio (Converted / Unconverted)	Positive / Negative	Points	Indicator for immediate purchase interest / non purchase interest	Final Points
Basket	Average	1,67	0,78	2,14	+	10	No	10
Browsing 1	Average	0,29	0,13	2,23	+	1	No	1
Browsing 2	Average	1,9	0,76	2,50	+	1	No	1
Browsing 3	Average	0,84	0,62	1,35	+	1	No	1
Checkout	Average	2,34	1,32	1,77	+	10	Yes	25
Checkout-Registration	Average	0,22	0,09	2,44	+	10	Yes	25
Contact	Average	1,68	0,55	3,05	+	3	No	3
Faq	Average	0,09	0,04	2,25	+	2	No	2
Account	Average	2,81	0,9	3,12	+	4	No	4
Newsletter Registration	Average	0,03	0,08	0,38	+	3	No	3
Other Page	Average	7,97	3,68	2,17	+	1	No	1
Price	Average	0,24	0,4	0,60	+	5	No	5
Product	Average	7,93	12,84	0,62	+	4	No	4
Registration	Average	0,28	0,08	3,50	+	3	No	3
Optout	% opted out	0,22	3,14	0,07	-	-10	Yes	-25

Table 14: Points system in the traditional lead scoring practice example
Source: Own representation

Creating a suppression segment

A suppression segment is created to ensure that no irrelevant contacts are classified as MQLs.

Choice of forfeiture model

As it has not yet been scientifically investigated whether the use of a decay model actually improves results, a decay model is not used in the application example. As there is no expiry model that can lower the score to a negative range, no campaign to reset negative scores is required.

Defining the threshold value

Using the code from Appendix 6, the score at the time of the first purchase is calculated retrospectively for converted leads from the online store and the current score for non-

converted leads. Two histograms were created to analyze the threshold values (see Figure 19), which compare the threshold values of converted and non-converted leads. This shows that converted leads tend to have higher scores at the time of handover than leads that do not convert.

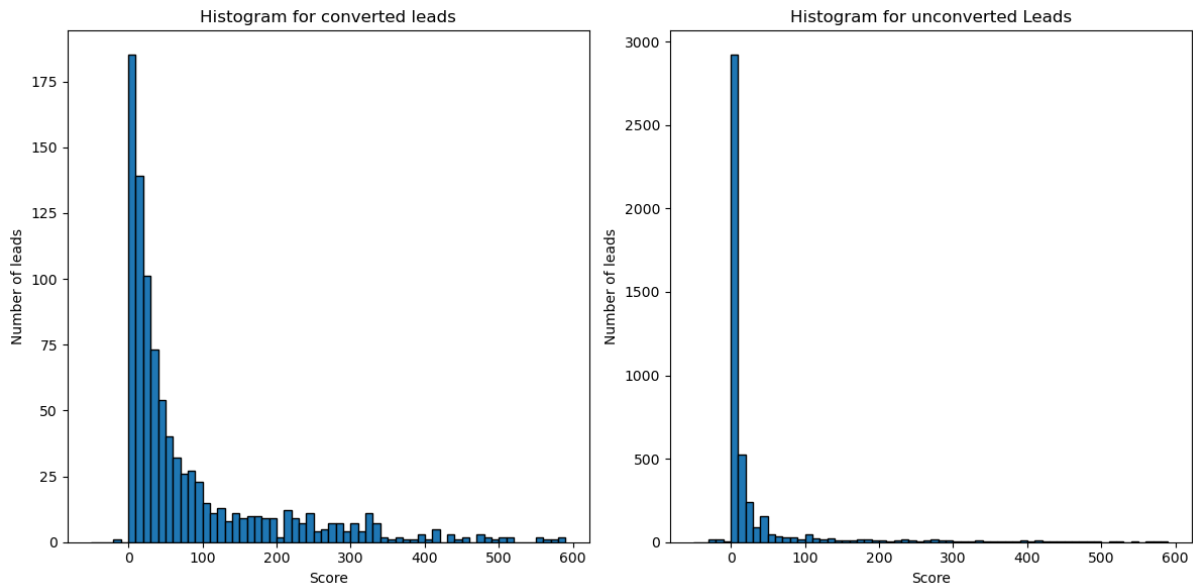


Figure 19: Comparison of the threshold values of converted and non-converted leads
Source: Own representation

In addition, a table was created to analyze the percentiles (cf. Table 15). This table shows that at a threshold value of 54 points, 50 percent of converted leads are correctly predicted and less than 25 percent of non-converted leads are incorrectly predicted as converting leads.

	Percentile	Converted_0	Converted_1
0	0.00	-23.0	1.0
1	0.05	4.0	5.0
2	0.10	8.0	9.0
3	0.15	8.0	13.0
4	0.20	8.0	17.0
5	0.25	8.0	18.0
6	0.30	9.0	24.0
7	0.35	12.0	31.0
8	0.40	15.0	36.0
9	0.45	16.0	48.0
10	0.50	18.0	54.0
11	0.55	21.0	69.0
12	0.60	24.0	86.0
13	0.65	32.0	103.0
14	0.70	40.0	123.0
15	0.75	52.0	156.0
16	0.80	74.0	207.0
17	0.85	130.0	273.0
18	0.90	207.0	350.0
19	0.95	410.0	574.0

Table 15: Table for analyzing the score percentiles
Source: Own representation

The threshold value was then adjusted step by step and the value at which the best results were achieved was determined. A threshold value of 52 points was determined, at which the status of the conversion is predicted with an accuracy of 71.73 percent. The results at a threshold value of 52 are Figure 20 can be seen in Figure 20.

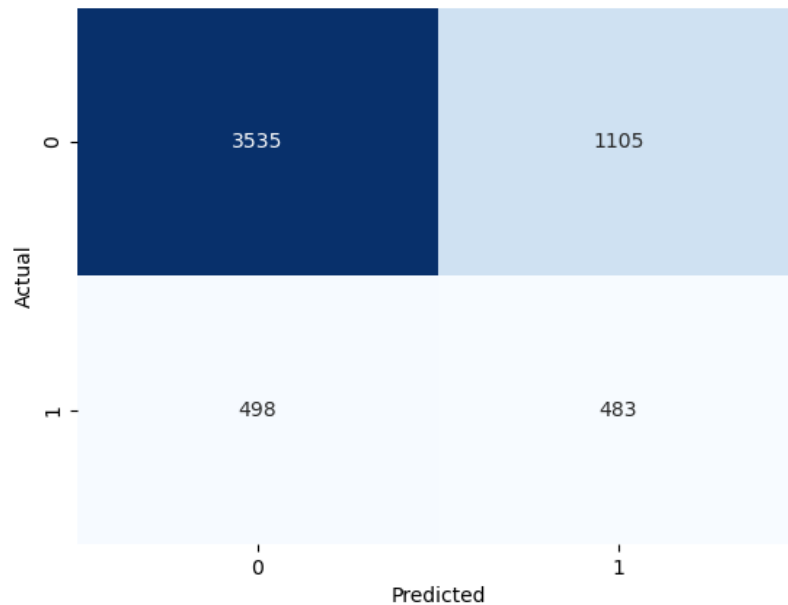


Figure 20: Confusion matrix of the preliminary traditional lead scoring system
Source: Own representation

Customizing the lead scoring system

When adapting the system, the scoring was changed several times and the appropriate threshold value was determined. The best predictions with 76.55 percent accuracy were achieved by adapting the scoring more to the results of the data analysis. For this purpose, the points for visiting a page in the "Price" category were reduced to two points and the points for visiting a page in the "Product" category were reduced to one point. A threshold value of 34 was selected. The results of the confusion matrix of the adapted system are Figure 21 can be seen in Figure 21.

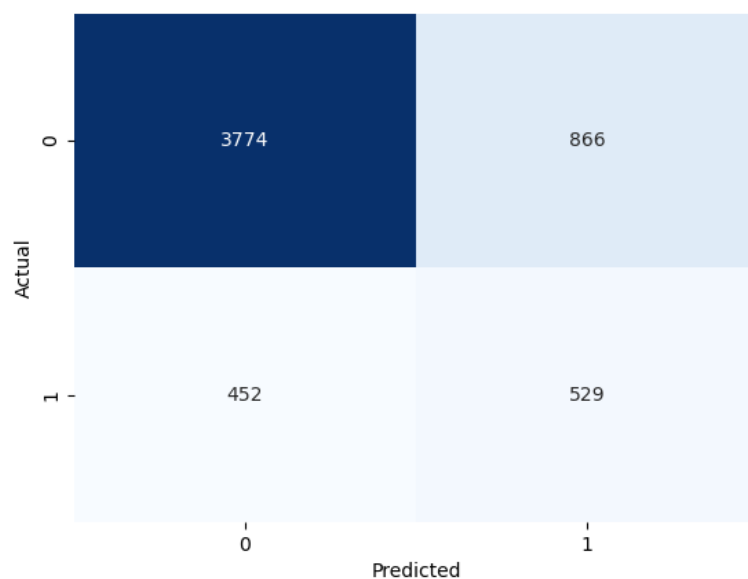


Figure 21: Confusion matrix of the final traditional lead scoring system
Source: Own representation

Sending marketing measures for MQLs

As there is no sales team to further qualify and contact MQLs, there is no handover to the sales team once the threshold has been reached. Alternative actions for handing over to the sales team for MQLs are creating MQL segments, sending special offers, CTAs and marketing content for MQLs or addressing MQLs via social media.

Evaluating the lead scores

Since no feedback can be requested from a sales team, the lead evaluation step differs from that described in section 7.2. However, it is possible to record which MQLs convert to customers. It is also possible to create a recycling path. Leads that do not convert within a predefined time after receiving the marketing measures for MQLs can be reintegrated into the lead scoring process.

Measuring success

Once the system has been put into operation, success is measured using KPIs. Several KPIs are defined for this purpose, for example those described in chapter 6 and reviewed regularly. If the KPIs deteriorate, the scoring system may need to be updated.

Updating the lead scoring system

The lead scoring system is updated at least every three months. This involves checking the KPIs, analyzing the lead scores, integrating new marketing materials into the system, adjusting the score by analyzing the converted MQLs and redefining the threshold value.

8.2 Predictive lead scoring

Creation of a service level agreement

As there is no contact with the online store's marketing and sales team, no SLA can be created for the predictive lead scoring system.

Data generation

The data from the online store's database is used as the database. As with traditional lead scoring, predictive lead scoring also provides the data of all website visits including time stamps, the date of newsletter unsubscriptions, the date of lead capture, the date of conversion and the date of the last activity. In addition, the segment affiliation of the contacts can be used to determine whether they are converted or non-converted leads.

Data preparation

The data preparation is carried out within Python. First, the SQL statements from Appendix 3 and 4 are used to export the data of converted leads and non-converted leads from Mautic. The steps of data selection, data cleansing, data construction, data integration and data formatting are then carried out as in the traditional system. This results in the lead overview DF from Table 12. In addition, the data is scaled and divided into training and test data.

Modeling

Once the data has been processed, a robust machine learning model is developed. For this purpose, the algorithms Logistic Regression, Decision Tree, Random Forest and Support Vector Machine were initially selected, which are common machine learning algorithms in predictive lead scoring (see chapter 4.1). For model development, a grid search is carried out for the previously identified algorithms. This means that the models are trained and tested with different combinations of the identified hyperparameters in order to find the best settings. In addition, the accuracy of the models with the best hyperparameters is calculated, which indicates what percentage of a model's predictions are correct. The following accuracies were achieved:

- Random Forest: 94.93 percent
- Decision Tree: 94.76 percent
- Support Vector Machine: 92.18 percent
- Logistic regression: 88.0 percent

Consequently, the random forest model was selected with the parameters determined from the grid search. The results of the random forest model in the form of a confusion matrix, which compares the prediction with the actual status "converted" (1) and "not converted" (0), are as follows Figure 22 can be seen in Figure 22. The confusion matrices for the Logistic Regression, Decision Tree and Support Vector Machine algorithms can be found in

Appendices 7, 8 and 9. Following the grid search, the best model is saved so that it can be used later to make predictions without having to train the model again. The code for modeling and saving the model can be found in Appendix 10.

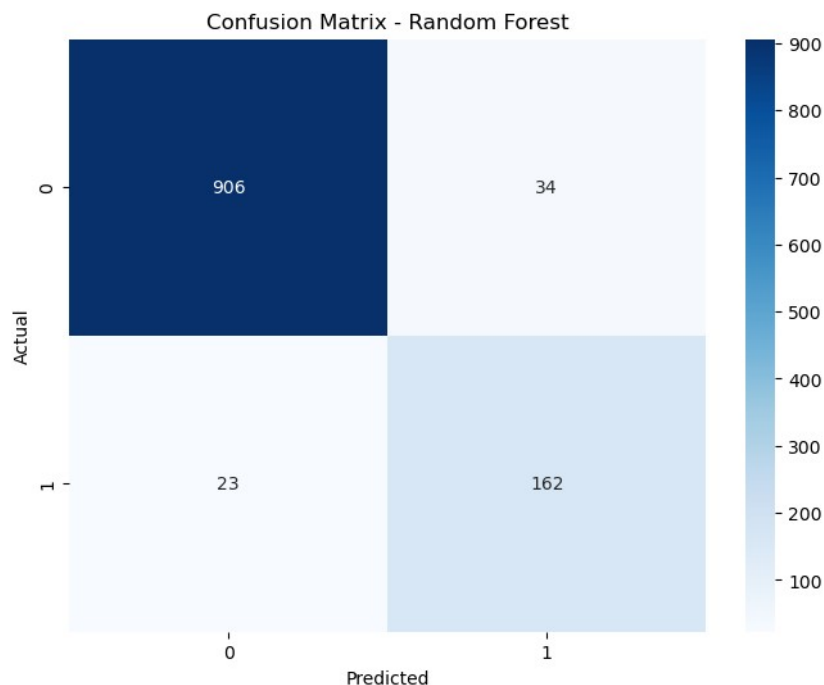


Figure 22: Confusion matrix for predictive lead scoring with the random forest algorithm
Source: Own representation

Model evaluation

In this evaluation phase, weaknesses in the model and its construction process were sought. Some adjustments were made, which were already taken into account in the data preparation process.

Model implementation

The following steps describe the application of the previously saved model to Mautic:

Creating the predictions

To create the predictions, all leads that are not in the suppression segment are exported from Mautic and processed. The previously saved machine learning model is then loaded and used to make predictions for the individual leads. The predictions are saved together with the associated lead ID as a CSV file. The code for this can be found in Appendix 11.

Transferring the model predictions to Mautic

Once the predictions have been generated, the CSV file with the model predictions is imported into Mautic. For this purpose, the lead_id column is assigned to the Mautic Contact ID during import and the column with the predictions is assigned to a database field previously created for the predictions.

Sending marketing measures for MQLs

As there is no sales team to further qualify and contact MQLs, there is no handover to the sales team once the threshold has been exceeded. Alternative measures for handing over MQLs to the sales team include creating MQL segments, sending special offers, CTAs and marketing content for MQLs or contacting MQLs via social media.

Evaluating the leads

Since no feedback can be requested from a sales team, the lead evaluation step differs from that described in section 7.2. However, it is possible to record which MQLs convert to customers. It is also possible to create a recycling path. Leads that do not convert within a predefined time after receiving the marketing measures for MQLs can be reintegrated into the lead scoring process.

Measuring success

Once the system has been put into operation, its success is measured using KPIs. Several KPIs are defined for this purpose, for example those described in chapter 6 and reviewed regularly. If the KPIs deteriorate, the machine learning model may need to be updated.

Updating the lead scoring system

The lead scoring system is updated at least every three months. This involves checking the KPIs, analyzing the results in more detail and, if necessary, training a new machine learning model.

9 Summary and outlook

As part of this work, a lead scoring model for simple and advanced use cases was developed using the Mautic software as an example. For this purpose, a generic process model for the development of a robust traditional and predictive lead scoring system was first developed following a literature review. The model was then adapted to the Mautic software, whereby the use cases of product-based lead scoring and ABS were also taken into account. Finally, the traditional and predictive lead scoring approaches and their results were examined using a practical example, whereupon the research results were analyzed and compared with the findings from the literature review. The individual approaches and use cases of lead scoring, their practical implementation, the resulting findings, their interpretation, the identified optimization potential for the Mautic software, the limitations of the research and recommendations for further research are summarized below.

Traditional lead scoring

In the traditional lead scoring approach, each lead receives a score. Points are added to or subtracted from this score depending on the characteristics and behavior of the lead. If a

contact's score exceeds a predefined threshold, it is considered qualified and is passed on to the sales team or receives special marketing measures for MQLs. To build a data-based, traditional lead scoring system in Mautic, the data is exported from the software's database and analyzed externally, for example with Python code. Based on the data analysis, experts create an implicit and an explicit scorecard. The latter defines the points that are awarded when a lead performs predefined actions or fulfills certain criteria. Leads that have exceeded the implicit and explicit thresholds are then passed on to the responsible sales team or provided with marketing materials for MQLs. Finally, feedback is generated on the individual MQLs in order to optimize the system or to re-integrate contacts that are not yet ready to buy into the scoring process.

Predictive lead scoring

In addition to the traditional lead scoring approach, there is also the predictive approach, which uses machine learning algorithms to predict leads that have a high purchase probability. In predictive scoring, the lead data is exported from Mautic, processed externally using Python code and then split into training and test data. The training and test data is then used to train various machine learning models and find out which algorithm delivers the best results. The model that makes the most accurate predictions is then saved to be used for future predictions. To make predictions for new leads, their data is first imported from Mautic and processed. The saved model is then applied to the data. The model's predictions are then imported into Mautic. Leads that are predicted to convert to customers by the machine learning algorithm are now passed on to the sales team or provided with marketing materials for MQLs. Feedback is also collected here in order to improve the model or to reintegrate leads that are not yet ready for a purchase into the scoring process.

It should be noted that both the traditional and predictive approaches proposed in this thesis are based on existing data. If this data is not available, it is possible to develop a model based on subjective expert assessments until sufficient data is available for a data analysis.

Product-based lead scoring

Product-based lead scoring is an application of lead scoring that aims to evaluate a lead's interest in a company's individual products, product groups and business areas rather than interest in the company as a whole. For this purpose, a separate scoring system is developed for each product. To build a product-based scoring system in Mautic, both the traditional approach and the predictive approach can be followed.

Account Based Scoring

In the ABS lead scoring use case, companies are scored instead of individual leads. In the traditional approach, an explicit and an implicit score is calculated for each company. The

explicit score is made up of company characteristics such as size, industry or technological status. The implicit score is made up of the sum of the actions carried out by the individual leads assigned to a company. As the functionalities of account-based marketing in Mautic are not yet advanced enough, automated ABS cannot currently be operated there. This is the case when using both the traditional and the predictive approach in the context of ABS.

Summary and interpretation of the results from the practical example

In traditional lead scoring, it was possible to predict which leads would become customers with an accuracy of 76.55%. In contrast, predictive lead scoring examined four different machine learning algorithms, with the random forest algorithm achieving the best results with an accuracy of 94.93 percent. These results support the findings of other research that identifies predictive lead scoring as a more effective alternative to the traditional approach. The fact that the performance of both approaches was tested on the same application example confirms this thesis.

These results are due to the fact that predictive lead scoring covers some of the weaknesses of traditional lead scoring. For example, non-linear effects can be taken into account in predictive systems. In addition, subjectivity is removed from the evaluation process as the system is based entirely on data. Another advantage of predictive scoring is the reduction of dependency on the amount of behavior-based data, which is why contacts who show an interest in buying from the outset are identified more quickly.

The finding that the random forest algorithm achieves the best results in predictive lead scoring is confirmed in several of the research studies examined. However, it remains uncertain whether increased accuracy of the models necessarily leads to improved results. Increased accuracy does lead to a higher percentage of qualified leads being passed on to the sales team, which increases the conversion rate of MQLs to customers. Nevertheless, lead scoring also measures other KPIs such as the impact on revenue or profit. It is uncertain whether higher accuracy automatically leads to improvements in these KPIs. For example, with an accuracy of 95 percent, the focus could be predominantly on MQLs who would buy anyway. This would mean that only a few leads with purchase potential but a lower probability of conversion would be passed on. One solution to this potential problem is to define several groups based on lead quality. For example, leads can be divided into the following groups to train a machine learning model:

- Leads that were not transferred
- Leads that have been passed but not purchased
- Leads that have been transferred and converted,

As a result, the model classifies the leads into three different categories after training, which can be processed according to their priority. Alternatively, an evaluation of the leads by the sales team on a scale of one to ten is also conceivable in order to develop the model. An analogous procedure can be used in traditional lead scoring by using a lead scoring matrix, as shown in Figure 2 can be implemented. Here, leads are divided into different categories based on the level of the implicit and explicit score in order to then work through them step by step.

It should also be noted that both the traditional and predictive approaches proposed in this paper are based on existing data. If this data is not available, it is possible to provisionally develop a model based on subjective expert assessments and adapt it as soon as sufficient data is available for effective data analysis.

Recommendations for the further development of the Mautic software

It has been found that problems can occur when using multiple scores, for example in product-based scoring or when working with an implicit and explicit score. This is due to the fact that all point groups that are created in addition to the main score only have a provisional solution in the form of a dedicated campaign (see Figure 11), which increases the complexity of the scoring process. Furthermore, there is no way to set the value of a point group by importing a CSV file. This would be particularly useful if the lead scoring system is implemented with existing customers and their scores are to be determined retrospectively and imported into Mautic. Consequently, the software needs to be further developed so that the same functions can be applied to point groups as to the standard score of a lead predefined in Mautic.

In Mautic, part of the points system must also be mapped through campaigns instead of implementing the logic exclusively under the "Manage Actions" menu item provided for this purpose. On the one hand, this applies to explicit parameters, as points can currently only be awarded on the basis of actions and not on the basis of a lead's properties. In addition, some actions are not included in the "Manage Actions" menu item, which is why some of the implicit evaluation must be carried out via campaigns. An example of this is when a contact unsubscribes from a newsletter. Therefore, another recommendation for optimizing the software is to integrate all actions and properties that can influence the score under "Manage Actions".

The proposed model works with suppression segments and multiple scores. However, since only actions based on a single score can be started under the "Manage Triggers" menu item, which is intended for triggering actions when score thresholds are exceeded, it is not possible to implement multiple scores and suppression segments in the logic. Alternatively,

leads must be added to segments with filters, on the basis of which campaigns can then be started. It would therefore make sense to extend the "Manage Triggers" menu item so that combined point triggers can also be used. These could take into account threshold values of several scores in combination with other conditions such as segment memberships.

Furthermore, although it is possible to score companies at account level in Mautic, it is not possible to carry out automated actions based on this. In order to enable the development of an account-based scoring system, this thesis recommends that the software be further developed to better support account-based marketing and scoring.

In the course of this work, Python code was used in both traditional and predictive lead scoring. Therefore, integrating Python into Mautic could provide additional value to users by automating predictive lead scoring and other complex analytics without manual effort and in real-time. A more user-friendly alternative to integrating Python is to integrate additional data analysis and machine learning functions directly into the software. This allows the model presented in this thesis to be used without the need for programming skills.

Limitations of the research

The accuracy of the predictions of the different approaches was determined as part of the practical examples examined. However, it was not possible to implement the proposed model in the company and measure the change in the individual KPIs as part of this work. Therefore, it would be of scientific importance to implement both a traditional and a predictive lead scoring system in a company and to examine and compare their results over a longer period of time.

Furthermore, ABS was not tested in the practical example, as this use case cannot currently be mapped in Mautic and no company data was available, as the online store from the practical example operates in the B2C sector.

The test of the product-based use case of lead scoring using a practical example was also omitted. However, as several lead scoring systems are set up simultaneously as part of product-based scoring, it can be assumed that these will achieve similar results to the traditional and predictive approaches, depending on the approach chosen.

Recommendations for further research

As part of the investigation, it was found that predictive lead scoring systems are models whose procedure is not comprehensible to users. Although ways have now been researched to make the procedure within the models more comprehensible, the exact procedure of the model is still not transparent. This issue could be addressed through further research by developing a hybrid approach of traditional and predictive lead scoring that uses machine learning to develop the scoring system and threshold for a traditional lead scoring system.

This would make the model fully data-driven and comprehensible at the same time. In addition, manual process steps such as adjusting the score and threshold could be automated, saving time when updating the system.

The research also identified several decay models that ensure that the score of leads with longer periods of inactivity is lowered. In this way, only leads with current interest have a high score. Although expiry models are recommended by various providers of MAS and CRM systems, there is no scientific research on whether they actually lead to better results. The questions of the extent to which an expiry model affects the success of lead scoring, how the right expiry model is selected and how it is successfully configured therefore provide material for a separate scientific paper.

In addition, chapter 5 identified two alternative or complementary methods to lead scoring: the RFM model and recommendation systems. Research that compares the application of alternative methods with lead scoring and draws conclusions as to which method is more suitable in which cases and when the application of a hybrid approach is recommended would therefore also be of scientific importance.

Bibliography

- Abbasi, Ahmed; Sarker, Suprateek; Chiang, Roger H.L. (2016): Big Data Research in Information Systems: Toward an Inclusive Research Agenda, in: *Journal of the Association for Information Systems*, Jg. 17, Nr. 2, S. 1–32.
- ActiveCampaign (o. D.): *Lead Scoring Best Practices (The Only Framework You Need to Get Started)*, [online] <https://www.activecampaign.com/learn/guides/lead-scoring-best-practices-the-only-framework-you-need-to-get-started> [20.11.2023].
- Adobe (2019): *DER ULTIMATIVE LEITFADEN FÜR LEAD SCORING.*, [online] <https://business.adobe.com/de/resources/guides/lead-scoring.html> [20.11.2023].
- Apple (o. D.): *Erlauben von Cookies in Safari auf dem Mac*, [online] <https://support.apple.com/de-lu/guide/safari/ibrw850f6c51/mac> [20.11.2023].
- Auerochs, Jessica (2021): *Lead Scoring - Definition und Umsetzung in der B2B Praxis*, [online] <https://www.marconomy.de/lead-scoring-definition-und-umsetzung-in-der-b2b-praxis-a-1eac3dcecab184cf86a6d9104a2147f3/> [20.11.2023].
- Bagshaw, Anthony (2015): What is marketing automation?, in: *J Direct Data Digit Mark Pract*, Jg. 17, Nr. 2, S. 84–85.
- Biegel, Bruce (2009): The current view and outlook for the future of marketing automation, in: *J Direct Data Digit Mark Pract*, Jg. 10, Nr. 3, S. 201–213.
- Bohanec, Marko; Kljajić Borštnar, Mirjana; Robnik-Šikonja, Marko (2015): Integration of machine learning insights into organizational learning. A case of B2B sales forecasting, in: *Proceedings of 28th Bled eConference*, Bled, S. 338–352.
- Bohanec, Marko; Kljajić Borštnar, Mirjana; Robnik-Šikonja, Marko (2017): Explaining machine learning models in sales predictions, in: *Expert Systems with Applications*, Jg. 71, S. 416–428.
- Braun, Simone (2021): Valide Kundendaten – Das Fundament für Omni-Channel Marketing, in: Uwe Hannig (Hrsg.), *Marketing und Sales Automation*, Wiesbaden: Springer, S. 159–175.
- Brevo (2023): *What is a Lead Scoring Model and How to Make One*, [online] <https://www.brevo.com/blog/lead-scoring-model/> [06.12.2023].

- Buckinx, Wouter; van den Poel, Dirk (2005): Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting, in: *European Journal of Operational Research*, Jg. 164, Nr. 1, S. 252–268.
- Cleff, Thomas (2019): *Applied Statistics and Multivariate Data Analysis for Business and Economics. A Modern Approach Using SPSS, Stata, and Excel*, Cham: Springer International Publishing.
- ConstantContact (2023): *Understanding Lead Score Decay*, [online] https://knowledgebase.constantcontact.com/lead-gen-crm/articles/KnowledgeBase/50268-Understanding-Lead-Score-Decay?lang=en_US [16.02.2024].
- D’Haen, J.; van den Poel, D.; Thorleuchter, D.; Benoit, D. F. (2016): Integrating expert knowledge and multilingual web crawling data in a lead qualification system, in: *Decision Support Systems*, Jg. 82, S. 69–78.
- D’Haen, Jeroen; van den Poel, Dirk (2013): Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework, in: *Industrial Marketing Management*, Jg. 42, Nr. 4, S. 544–551.
- Day, Daniel G.; Wei Shi, Savannah (2020): Automated and Scalable: Account-Based B2B Marketing for Startup Companies, in: *JBTP*, Jg. 8, Nr. 2, S. 16-23.
- DMEXCO (o. D.): *Marketing Automation*, [online] <https://dmexco.com/de/topic/marketing-automation/> [09.03.2024].
- Duncan, Brendan; Eklan, Charles (2015): Probabilistic Modeling of a Sales Funnel to Prioritize Leads, in: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York: Association for Computing Machinery, S. 1751–1758.
- Elkan, Charles (2013): *Predictive analytics and data mining*, University of California, San Diego.
- Encharge.io (2021): *10 Must-Know Lead Scoring Best Practices to Win Your Leads in 2022*, [online] <https://encharge.io/lead-scoring-best-practices/> [20.11.2023].
- Espadinha-Cruz, P.; Fernandes, A.; Grilo, A. (2021): Lead management optimization using data mining: A case in the telecommunications sector, in: *Computers & Industrial Engineering*, Jg. 154, S. 1–14.
- faraday.ai (o. D.): *Predictive lead scoring: Best practices for B2C lead scoring using machine learning*, [online] <https://faraday.ai/blog/predictive-lead-scoring-b2c> [20.11.2023].

- Flocke, Louisa; Holland, Heinrich (2014): Die Customer Journey Analyse im Online Marketing, in: *Dialogmarketing Perspektiven 2013/2014*, Wiesbaden: Springer Fachmedien, S. 213–242.
- Ghorbel, Amine (2023): *Everything you Need to Know About Lead Scoring in 2023*, [online] <https://lagrowthmachine.com/lead-scoring/> [06.12.2023].
- Gokhale, Prasad; Joshi, Pratima (2018): A Binary Classification Approach to Lead Identification and Qualification, in: A. V. Deshpande, Aynur Passi U.K., Dharm S.M. Nayak und Bharat P.S. Pathan (Hrsg.), *Smart Trends in Information Technology and Computer Communications*, Singapur: Springer, S. 279–291.
- Gooding, Resa (2022): *Empowering Marketing and Sales with HubSpot. Take your business to a new level with HubSpot's inbound marketing, SEO, analytics, and sales tools*, 1. Aufl., Birmingham: Packt Publishing Limited.
- Google (o. D.): *Cookies in Chrome löschen, zulassen und verwalten*, [online] <https://support.google.com/chrome/answer/95647> [20.11.2023].
- Gradow, Lisa; Greiner, Ramona (2021): *Quick Guide Consent-Management*, Wiesbaden: Springer.
- Griebsch, Laura (2021): *Marketing Automation im B2B – Grundlagen und Erklärungen und 8 Tipps für den Einsatz*, [online] <https://www.marconomy.de/marketing-automation-im-b2b-grundlagen-und-erklaerungen-und-8-tipps-fuer-den-einsatz-a-1001337/> [30.01.2024].
- Hannig, Uwe (2020): Marketing-Automaton - automatisch mehr Markterfolg, in: Marcus Stumpf (Hrsg.), *Die 10 wichtigsten Zukunftsthemen im Marketing*, 2. Aufl., Freiburg: Haufe-Lexware GmbH & Co. KG, S. 207–230.
- Hannig, Uwe (2021): Lead Management Automation vereint Marketing und Vertrieb, in: Uwe Hannig (Hrsg.), *Marketing und Sales Automation*, Wiesbaden: Springer, S. 243–257.
- Heinzelbecker, Klaus (2021a): Account-based Marketing mit CRM und Marketing Automation, in: Uwe Hannig (Hrsg.), *Marketing und Sales Automation*, Wiesbaden: Springer, S. 387-410.
- Heinzelbecker, Klaus (2021b): CRM, CXM und Marketing Automation, in: Uwe Hannig (Hrsg.), *Marketing und Sales Automation*, Wiesbaden: Springer, S. 135–147.
- Hu, Hang; Peng, Peng; Wang, Gang (2019): *Characterizing Pixel Tracking through the Lens of Disposable Email Services*, 2019 IEEE Symposium on Security and Privacy (SP), San Francisco: IEEE, S. 365–379.

- Hu, Jianfeng; Zhang, Bo (2012): *Product Recommendation System*, Stanford University.
- Von der Hude, Marlis (2020): *Predictive Analytics und Data Mining*, Wiesbaden: Springer Fachmedien Wiesbaden.
- Hufford, Brendan (2021): *Create an Effective Lead Scoring System in 7 Steps*, [online] <https://www.activecampaign.com/blog/create-effective-lead-scoring-rules> [06.12.2023].
- InvestGlass (2023): *Die 4 besten Lead-Scoring-Modelle im Jahr 2023*, [online] <https://www.investglass.com/de/the-4-best-lead-scoring-models-in-2023-examples/> [25.01.2024].
- Jadli, Aissam; Hamim, Mohammed; Hain, Mustapha; Hasbaoui, Anouar (2022): TOWARD A SMART LEAD SCORING SYSTEM USING MACHINE LEARNING, in: *INDJCSE*, Jg. 13, Nr. 2, S. 433–443.
- Jo, Taeho (2021): *Machine Learning Foundations. Supervised, Unsupervised, and Advanced Learning*, 2. Aufl., Cham: Springer International Publishing.
- Jörvinen, Joel; Taiminen, Heini (2016): Harnessing marketing automation for B2B content marketing, in: *Industrial Marketing Management*, Jg. 54, S. 164–175.
- Kazemi, Abolfazl; Babaei, Mohammad Esmail (2011): Modelling Customer Attraction Prediction in Customer Relation Management using Decision Tree: A Data Mining Approach, in: *Journal of Optimization in Industrial Engineering*, Jg. 4, Nr. 9, S. 37–45.
- Kim, YongSeog; Street, W.Nick (2004): An intelligent system for customer targeting: a data mining approach, in: *Decision Support Systems*, Jg. 37, Nr. 2, S. 215–228.
- Koerner, Alexander (2021): Marketing und Sales Automation, in: Uwe Hannig (Hrsg.), *Marketing und Sales Automation*, Wiesbaden: Springer, S. 61–78.
- Kuhn, Max; Johnson, Kjell (2013): *Applied Predictive Modeling*, New York: Springer New York.
- Kumar, V.; Reinartz, Werner (2018): *Customer Relationship Management*, Berlin, Heidelberg: Springer Berlin Heidelberg.
- Lattice (2014): *The Evolution From Traditional to Predictive Lead Scoring: A how-to guide for considering predictive scoring*, [online] <https://www.demandgenreport.com/industry-resources/ebooks/2892-the-evolution-from-traditional-to-predictive-lead-scoring> [20.11.2023].
- Lontzek, Nicole (2022): *Lead Scoring - Das Punktesystem für echte Marketingspezialisten*, [online] <https://www.marconomy.de/lead-scoring-das-punktesystem-fuer-echte-marketingspezialisten-a-1087780/> [20.11.2023].

- Martinez-Plumed, Fernando; Contreras-Ochando, Lidia; Ferri, Cesar; Hernandez-Orallo, Jose; Kull, Meelis; Lachiche, Nicolas; Ramirez-Quintana, Maria Jose; Flach, Peter (2021): CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories, in: *IEEE Trans. Knowl. Data Eng.*, Jg. 33, Nr. 8, S. 3048–3061.
- mautic.org (o. D.): *About Mautic*, [online] <https://github.com/mautic/mautic> [08.12.2023].
- Michiels, Ian (2008): *Lead Prioritization and Scoring. The Path to Higher Conversion*, [online] <https://silo.tips/download/lead-prioritization-and-scoring> [20.11.2023].
- MM-Redaktion (2021): *Das nervt SEO-Agenturen: Apple schränkt Cookie-Tracking ein*, [online] https://magazinmedien.de/cookie-tracking_eingeschraenkt/ [20.11.2023].
- Monat, Jamie P. (2011): Industrial sales lead conversion modeling, in: *Marketing Intelligence & Planning*, Jg. 29, Nr. 2, S. 178–194.
- Mozilla (o. D.): *Cookies blockieren*, [online] <https://support.mozilla.org/de/kb/Cookies-blockieren> [20.11.2023].
- Naveen, Kumar G.; Hariharanath, K. (2021): Designing a Lead Score Model for Digital Marketing Firms in Education Vertical in India, in: *Indian Journal of Science And Technology*, Jg. 14, Nr. 16, S. 1302–1309.
- Nygaard, Robert; Mezei, Jozsef (2020): Automating Lead Scoring with Machine Learning: An Experimental Study, in: *Proceedings of the 53rd Hawaii International Conference on System Sciences*, Honolulu, S. 1439–1448.
- Oracle (o. D.): *Lead scoring*, [online] <https://docs.oracle.com/en/cloud/saas/marketing/eloqua-user/Help/LeadScoring/LeadScoring.htm> [25.01.2024].
- Patel, Neil (o. D.): *Wie man Lead Scoring richtig nutzt, um mehr Einnahmen zu erzielen*, [online] <https://neilpatel.com/de/blog/lead-scoring-richtig/> [05.12.2023].
- Philipp, Martin (2021): Vom E-Mail-Marketing zum Lead Management, in: Uwe Hannig (Hrsg.), *Marketing und Sales Automation*, Wiesbaden: Springer, S. 193–213.
- Rahimi, Omid (2020): *Mit 5 Tipps zum effektiven B2B-Lead-Scoring*, [online] <https://www.marconomy.de/mit-5-tipps-zum-effektiven-b2b-lead-scoring-a-d2678f31de7c5524c335634a2e7e7df8/> [20.11.2023].
- Saha, Swapan K.; Aman, Ashraful; Hossain, Md. shawkat; Islam, Aminul; Rodela, Ripa S. (2014): A Comparative Study On B2B Vs. B2C Based On Asia Pacific Region, in: *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, Jg. 3, Nr. 9, S. 294–298.

Salesforce (o. D.): *Scoring Categories*, [online]

https://help.salesforce.com/s/articleView?id=sf.pardot_leadqual_scoring_categories.htm&type=5 [25.01.2024].

Schoepf, Alex (2021): Best Practices für Marketing-Automation-Einsteiger, in: *Wirtsch Inform Manag*, Jg. 13, Nr. 4, S. 280–289.

Schüller, Anne M.; Schuster, Norbert (2022): *Marketing-Automation. Neukundengewinnung, Up-Selling, Cross-Selling, Bestandskunden- Management: Mehr Umsatz mit der Wasserlochstrategie®*, 2. Aufl., Freiburg: Haufe Group.

Schuster, Norbert (2021): Marketing Automation ändert den Vertrieb, in: Uwe Hannig (Hrsg.), *Marketing und Sales Automation*, Wiesbaden: Springer, S. 105–119.

Schuster, Norbert (2022): *Digitalisierung in Marketing und Vertrieb. Richtige Strategien entwickeln und Potentiale der Digitalisierung für mehr Umsatz nutzen*, 2. Aufl., Freiburg, München, Stuttgart: Haufe Group.

Sereda, Evgeni (2021): *So gehst du mit URL-Parametern um und organisierst sie SEO-freundlich*, [online] <https://www.semrush.com/blog/de/url-parameter/> [17.02.2024].

Sharma, Jatin; Sharma, Kartikay; Garg, Kaustubh; Sharma, Avinash Kumar (2021): Product Recommendation System a Comprehensive Review, in: *IOP Conf. Ser.: Mater. Sci. Eng.*, Jg. 1022, Nr. 1, S. 1–8.

Shearer, Colin (2000): The CRISP-DM Model: The New Blueprint for Data Mining, in: *Journal of Data Warehousing*, Jg. 5, Nr. 4, S. 13–22.

Simmoleit, Rainer (2023): *Künstliche Intelligenz im Lead Scoring*, [online] <https://www.marconomy.de/kuenstliche-intelligenz-im-lead-scoring-a-5e21a051bb63bff701d8c39e72a2fef3/> [20.11.2023].

Swani, Lakshay; Tyagi, Prakita (2017): Predictive Modelling Analytics through Data Mining, in: *International Research Journal of Engineering and Technology*, Jg. 4, Nr. 9, S. 5–11.

Teiu, Codrin (2021): Marketing Automation Systems as Part of the Management Information Systems Evolution, in: *Organizations and Performance in a Complex World*, Cham: Springer, S. 325–333.

Todor, Raluca D. (2016): Marketing Automation, in: *Bulletin of the Transilvania University of Braşov*, Jg. 9, Nr. 2, S. 87–94.

Uhlemann, Ingrid Andrea (2015): *Einführung in die Statistik für Kommunikationswissenschaftler*, Wiesbaden: Springer Fachmedien Wiesbaden.

- Verma, Rakesh; Koul, Saroj; Pai, Sushanth S. (2016): Identifying profitable clientele using the analytical hierarchy process, in: *International Journal of Business and Systems Research*, Jg. 10, 2-4, S. 220–237.
- Wu, Migao; Andreev, Pavel; Benyoucef, Morad (2023): The state of lead scoring models and their impact on sales performance, in: *Information technology & management*, S. 1–30.
- Wu, Ming-Wei; Lin, Ying-Dar (2001): Open source software development: an overview, in: *Computer*, Jg. 34, Nr. 6, S. 33–38.
- Wuttke, Laurenz (o. D.): *Lead Scoring: Definition, Prozess und Unterschiede zwischen B2B vs. B2C*, [online] <https://datasolut.com/lead-scoring> [20.11.2023].
- Zumstein, Darius; Gasser, Marc; Thüring, Urs; Völk, Klaus; Wicki, Alexander; Oswald, Carmen; Merdzanovic, Adis; Hannich, Frank (2023): *Marketing Automation Report 2023*, ZHAW Zürcher Hochschule für Angewandte Wissenschaften, Winterthur.

Appendix

Appendix 1: Campaign to generate sales feedback

As an alternative to implementing the logic for obtaining sales feedback in the CRM system, a campaign in Mautic can also be used for this purpose (see Figure 23).

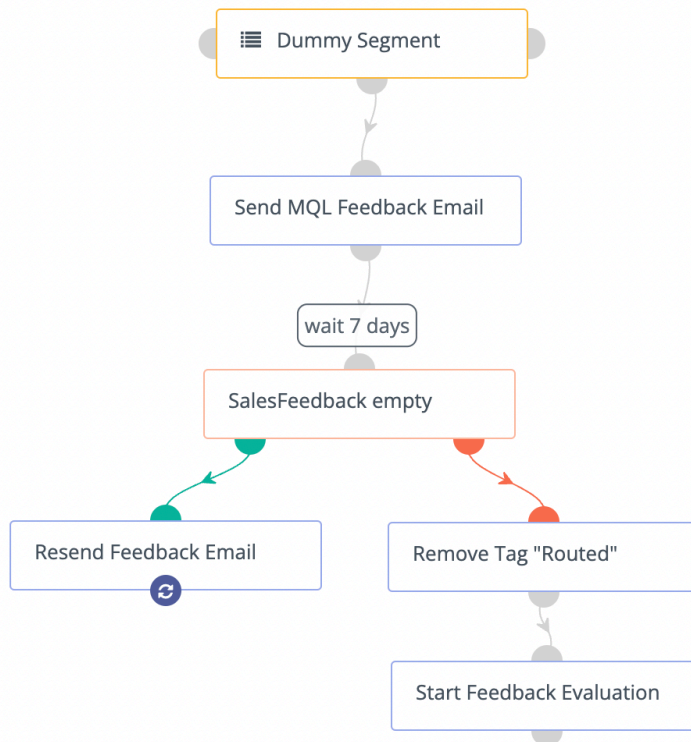


Figure 23: Campaign to generate sales feedback
Source: Own representation

At the start of this campaign, an email is sent to the responsible sales team asking for feedback on the quality of a lead that has been handed over (see Figure 24). Within this email, the sales employees can select whether the lead is an SAL, whether the lead is not yet ready for the sales process and should be recycled or whether the lead should be rejected. If a sales employee clicks on one of the three feedback buttons in the email, the feedback is automatically saved in the Mautic database.

Hey [Owner First name],

you've recently been sent a MQL alert for [Email]. Could you please provide feedback on the quality of the lead?

SAL The lead has high quality and was accepted by the sales team

Recycle The lead is not ready yet and has to be handed back to marketing for further nurturing

Reject The lead has no future potential

Figure 24: Email for generating sales feedback
Source: Own representation

Depending on which of the three buttons is clicked, different URL parameters are attached to the link of a form. URL parameters are character strings that are appended to the end of a URL in order to transmit information. The actual URL is separated from the parameters by a question mark (cf. Sereda 2021). If the names of the individual URL parameters match the field names in the form, Mautic automatically transfers them to the form fields. For example, in the following URL, the lead's email address is entered in the form using the variable {contactfield=email} and the feedback "SAL".

"https://meinformular?email={contactfield=email}&salesfeedback=SAL"

Once this has been done, the following Java script from Figure 25 the form with the sales feedback is sent automatically. If no feedback is received from the sales team within seven days, the email is sent again. This process is repeated until feedback on the MQL is available.

```
<script>
function autosubmit() {
  var form = document.querySelector("form[id="mauticform_autosubmitfeedbackform"]");
  if (form) {
    form.submit();
  }
}
window.addEventListener("load", autosubmit);
</script>
```

*Figure 25: JavaScript code for sending the sales feedback
Source: Own representation*

Appendix 2: Python code for preparing the lead data

```

import os
import pandas as pd

# Import Mautic data
os.chdir('working_folder_path')
df1 = pd.read_csv('converted.csv')
df2 = pd.read_csv('not_converted.csv')

# Add column "Converted" and merge both DataFrames
df1['Converted'] = 1
df2['Converted'] = 0
df = pd.concat([df1, df2])

#Convert date columns into datetime
df['date_hit'] = pd.to_datetime(df['date_hit'])
df['end_date'] = pd.to_datetime(df['end_date'])
df['start_date'] = pd.to_datetime(df['start_date'])
df['confirmed_optout_date'] = pd.to_datetime(df['confirmed_optout_date'], errors='coerce')

# Delete pagehits and optouts which took time after the conversion
df = df[df['date_hit'] <= df['end_date']]
df.loc[df['confirmed_optout_date'] > df['end_date'], 'confirmed_optout_date'] = pd.NaT

# Replace empty optout dates with 0 and filled optout dates with 1
df['confirmed_optout_date'] = df['confirmed_optout_date'].apply(lambda x: 0 if pd.isnull(x) else 1)

#Calculate Sales Cycle Duration Column for unconverted and unconverted leads
df['time_diff'] = (df['end_date'] - df['start_date']).dt.total_seconds() / (24 * 60 * 60)

#Build Url Categories
df['url'] = df['url'].astype(str)
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Product'
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Checkout-Registration'
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Registration'
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Price'
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Basket'
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Checkout'
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Contact'
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Newsletter Registration'
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Browsing 1'
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Browsing 2'
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Browsing 3'
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Faq'
df.loc[df['url'].str.contains('Keyword1|Keyword2|Keyword3|...'), 'url'] = 'Account '
df = df[~df['url'].str.contains('confirm')]
df.loc[df['url'].str.contains('my_website'), 'url'] = 'Other Page'

```



```
df = df[~df['url'].str.contains('http')]

# Create DataFrame with consolidated data
#Count total pagehits for each contact
total_pagehits = df.groupby('lead_id')['url'].count().reset_index()
total_pagehits = total_pagehits.rename(columns={'url': 'total_pagehits'})
#Count number of pagehits in each category for each contact
pagehit_group = df.groupby(['lead_id', 'url']).size().unstack(fill_value=0).reset_index()
#Calculate salescycle duration for each contact
salescycle_duration = df.groupby('lead_id')['time_diff'].mean().round(2).reset_index()
salescycle_duration = salescycle_duration.rename(columns={'time_diff': 'salescycle_duration'})
#Collect conversion status for each contact
conversion_df = df.groupby('lead_id')['Converted'].mean().reset_index()
#Collect optout status for each lead
optout_df = df.groupby('lead_id')['confirmed_optout_date'].mean().reset_index()
optout_df = optout_df.rename(columns={'confirmed_optout_date': 'opted_out'})
#Combine the data in one DataFrame
merged_df = total_pagehits.merge(pagehit_group, on='lead_id', how='left')
merged_df = merged_df.merge(salescycle_duration, on='lead_id', how='left')
merged_df = merged_df.merge(optout_df, on='lead_id', how='left')
merged_df = merged_df.merge(conversion_df, on='lead_id', how='left')

# Remove entries in which the number of total page visits is either less than 3 or higher than 300 (outliers)
merged_df = merged_df[(merged_df['total_pagehits'] <= 300) & (merged_df['total_pagehits'] >= 3)]
```

Figure 26: Python code for preparing the lead data
Source: Own representation

Appendix 3: SQL statements for downloading the data of converted leads from Mautic

```
SELECT
  lead_lists_leads.lead_id,
  lead_lists_leads.date_added AS end_date,
  page_hits.url,
  page_hits.date_hit,
  leads.date_added AS start_date,
  leads.confirmed_optout_date
FROM lead_lists_leads
JOIN leads ON lead_lists_leads.lead_id = leads.id
JOIN page_hits ON lead_lists_leads.lead_id = page_hits.lead_id
WHERE lead_lists_leads.leadlist_id = 4
  AND (leads.date_added LIKE '2021%' OR leads.date_added LIKE '2022%' OR leads.date_added LIKE '2023%')
  AND page_hits.url LIKE '%mywebsite%';
```

Figure 27: SQL statement for downloading the data of converted leads from Mautic
Source: Own representation

Appendix 4: SQL statements for downloading the data of non-converted leads from Mautic

```
SELECT
  lead_lists_leads.lead_id,
  page_hits.url,
  page_hits.date_hit,
  leads.last_active AS end_date,
  leads.date_added AS start_date,
  leads.confirmed_optout_date
FROM leads
JOIN page_hits ON leads.id = page_hits.lead_id
JOIN lead_lists_leads ON leads.id = lead_lists_leads.lead_id
WHERE (leads.date_added LIKE '2021%' OR leads.date_added LIKE '2022%' OR leads.date_added LIKE '2023%')
AND lead_lists_leads.leadlist_id IN (1, 2, 3)
AND page_hits.url LIKE '%mywebsite%'
AND DOES NOT EXIST (
  SELECT 1
  FROM lead_lists_leads
  WHERE lead_lists_leads.lead_id = leads.id
  AND lead_lists_leads.leadlist_id = 4
);
```

Figure 28: SQL statement for downloading the data of non-converted leads from Mautic
Source: Own representation

Appendix 5: Python code for analyzing the lead data

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Split df into df_converted and df_unconverted
df_converted = merged_df[merged_df['Converted'] == 1]
df_unconverted = merged_df[merged_df['Converted'] == 0]

# Create df_stats
exclude_columns = ['lead_id', 'Converted', 'opted_out']
all_columns = df_converted.columns.difference(exclude_columns)
df_stats = pd.DataFrame(index=pd.MultiIndex.from_product([all_columns, ['Average', 'Stdev', 'Median', 'Max']]),
columns=['Converted', 'Unconverted'])
for col in all_columns:
    # Calculate stats
    avg_converted = df_converted[col].mean()
    stdev_converted = df_converted[col].std()
    median_converted = df_converted[col].median()
    max_converted = df_converted[col].max()
    avg_unconverted = df_unconverted[col].mean()
    stdev_unconverted = df_unconverted[col].std()
    median_unconverted = df_unconverted[col].median()
    max_unconverted = df_unconverted[col].max()
    # Round stats (2 digits)
    avg_converted = round(avg_converted, 2)
    stdev_converted = round(stdev_converted, 2)
    median_converted = round(median_converted, 2)
    max_converted = round(max_converted, 2)
    avg_unconverted = round(avg_unconverted, 2)
    stdev_unconverted = round(stdev_unconverted, 2)
    median_unconverted = round(median_unconverted, 2)
    max_unconverted = round(max_unconverted, 2)
    # Add to df_stats
    df_stats.at[(col, 'Average'), 'Converted'] = avg_converted
    df_stats.at[(col, 'Stdev'), 'Converted'] = stdev_converted
    df_stats.at[(col, 'Median'), 'Converted'] = median_converted
    df_stats.at[(col, 'Max'), 'Converted'] = max_converted
    df_stats.at[(col, 'Average'), 'Unconverted'] = avg_unconverted
    df_stats.at[(col, 'Stdev'), 'Unconverted'] = stdev_unconverted
    df_stats.at[(col, 'Median'), 'Unconverted'] = median_unconverted
    df_stats.at[(col, 'Max'), 'Unconverted'] = max_unconverted

# Add percentage opted out to df_stats
percentage_opted_out1 = (df_converted['opted_out'] == 1).mean() * 100
percentage_opted_out2 = (df_unconverted['opted_out'] == 1).mean() * 100

```

```
percentage_opted_out1 = round(percentage_opted_out1, 2)
percentage_opted_out2 = round(percentage_opted_out2, 2)
df_stats.at[('opted_out', 'Percentage opted out'), 'Converted'] = percentage_opted_out1
df_stats.at[('opted_out', 'Percentage opted out'), 'Unconverted'] = percentage_opted_out2

# Build correlation table
df_corr = merged_df.drop("lead_id", axis=1).copy()
correlation = df_corr.corr()
# Extract correlation with 'Converted'
correlation_with_converted = correlation["Converted"].to_frame()
# Create correlation heatmap
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_with_converted, annot=True, cmap='coolwarm', linewidths=.5)
plt.title('Correlation with "Converted"')
plt.show()
```

Figure 29: Python code for analyzing the lead data
Source: Own representation

Appendix 6: Python code for determining and adjusting the threshold value in lead scoring

```
import pandas as pd
import os
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import accuracy_score

# Import data
os.chdir('working_folder_path')
df = pd.read_excel('lead_overview.xlsx')

# Define point system
column_weights = {
    'Basket': 10,
    'Browsing 1': 1,
    'Browsing 2': 1,
    'Browsing 3': 1,
    'Checkout Page': 25,
    'Checkout registration': 25,
    'Contact': 3,
    'Faq': 2,
    'MyAccount Page': 4,
    'Newsletter Registration': 3,
    'Other Page': 1,
    'Price Page': 5,
    'Product Page': 4,
    'Registration': 3,
    'opted_out': -25
}

# Calculate score for each lead
df['weighted_sum'] = df[list(column_weights.keys())].multiply(list(column_weights.values())).sum(axis=1)

# Create df for converted and nonconverted lead scores
converted_1 = df[df['Converted'] == 1]['weighted_sum']
converted_0 = df[df['Converted'] == 0]['weighted_sum']

# Create histogram for converted and non converted lead scores
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
plt.hist(converted_1, bins=range(-50, 600, 10), edgecolor='black')
plt.title('Histogram for converted leads')
plt.xlabel('Score')
plt.ylabel('Number of leads')
plt.subplot(1, 2, 2)
```

```
plt.hist(converted_0, bins=range(-50, 600, 10), edgecolor='black')
plt.title('Histogram for unconverted Leads')
plt.xlabel('Score')
plt.ylabel('Number of leads')
plt.tight_layout()
plt.show()

# Create percentile table
percentiles = np.arange(0, 1, 0.05)
# For nonconverted leads
table_converted_0 = pd.DataFrame({
    'Percentile': percentiles,
    'Value': [df[df['Converted'] == 0]['weighted_sum'].quantile(p) for p in percentiles]
})
# For converted leads
table_converted_1 = pd.DataFrame({
    'Percentile': percentiles,
    'Value': [df[df['Converted'] == 1]['weighted_sum'].quantile(p) for p in percentiles]
})
# Combined table
table_combined = pd.DataFrame({
    'Percentile': percentiles,
    'Converted_0': [df[df['Converted'] == 0]['weighted_sum'].quantile(p) for p in percentiles],
    'Converted_1': [df[df['Converted'] == 1]['weighted_sum'].quantile(p) for p in percentiles]
})
# Print combined percentile table
print("\nCombined table:")
print(table_combined)

# Define treshold
threshold = 52

# Create 'Prediction' column
df['Prediction'] = np.where(df['weighted_sum'] > threshold, 1, 0)

# Create heatmap
confusion_matrix = pd.crosstab(df['Converted'], df['Prediction'], rownames=['Actual'], colnames=['Predicted'])
sns.heatmap(confusion_matrix, annot=True, fmt='d', cmap='Blues', cbar=False)
plt.show()

# Calculate accuracy for chosen treshold
accuracy = accuracy_score(df['Converted'], df['Prediction'])
print(f'Accuracy: {accuracy}')
```

Figure 30: Python code for determining and adjusting the threshold value in lead scoring
Source: Own representation

Appendix 7: Results of predictive lead scoring with the support vector machine algorithm

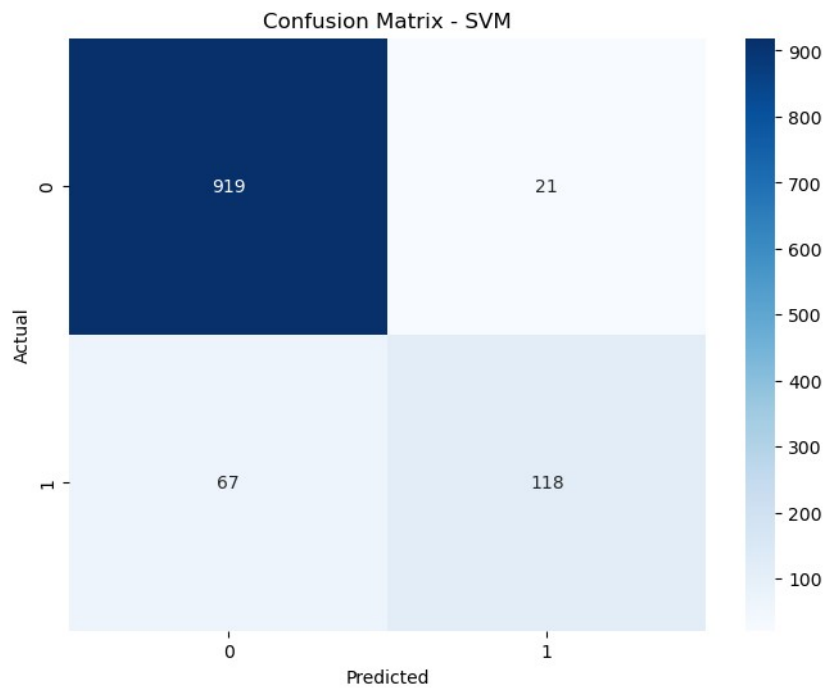


Figure 31: Results of predictive lead scoring with the support vector machine algorithm
Source: Own representation

Appendix 8: Results of predictive lead scoring with the decision tree algorithm

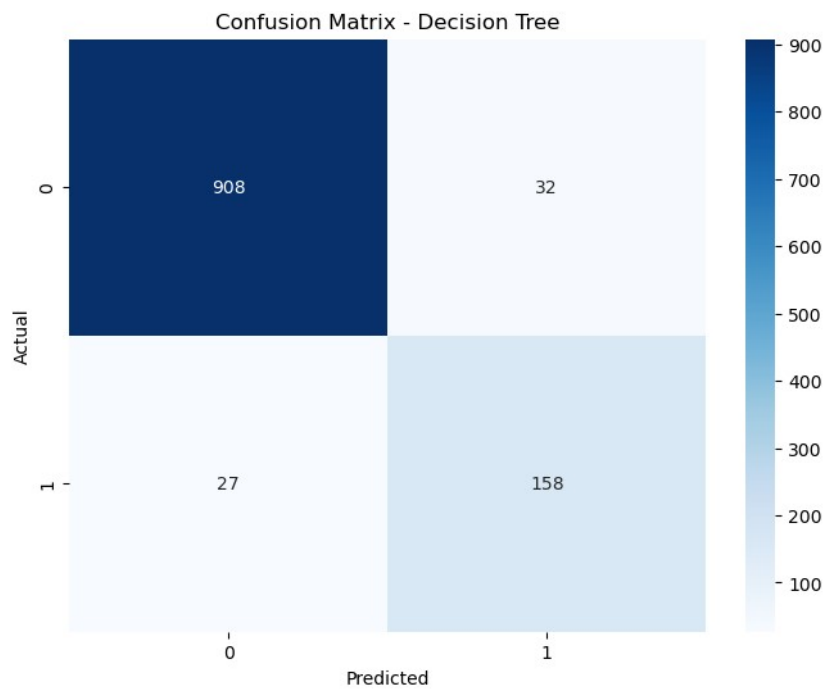


Figure 32: Results of predictive lead scoring with the decision tree algorithm
Source: Own representation

Appendix 9: Results of predictive lead scoring with the logistic regression algorithm

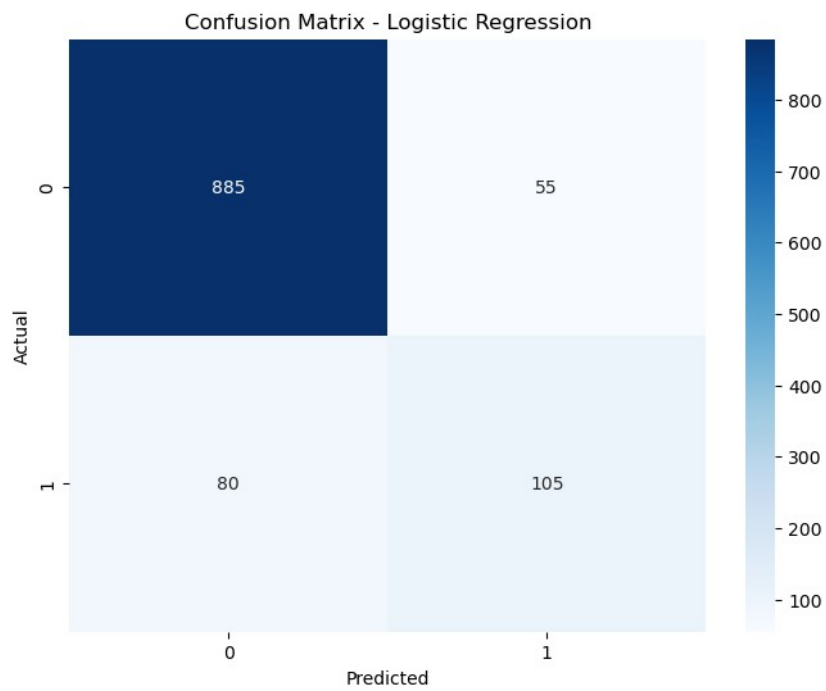


Figure 33: Results of predictive lead scoring with the logistic regression algorithm
Source: Own representation

Appendix 10: Python code to develop a machine learning model for lead scoring

```
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, confusion_matrix
import joblib

# Import data
os.chdir('working_folder_path')
df = pd.read_excel('lead_overview.xlsx')

# Split into features (X) and target variable (y)
X = df.drop(["Converted", "lead_id"], axis=1)
y = df["Converted"]

# Scaling and splitting into training and test data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)

# Define models and hyperparameters
models = {
    'Logistic Regression': (LogisticRegression(), {'C': [0.1, 1, 10]}),
    'Decision Tree': (DecisionTreeClassifier(), {'max_depth': [None, 10, 20, 30]}),
    'Random Forest': (RandomForestClassifier(), {'n_estimators': [50, 100, 200]}),
    'SVM': (SVC(), {'C': [0.1, 1, 10], 'kernel': ['linear', 'rbf']})
}

# Executing GridSearchCV for every model
best_model = None
best_accuracy = 0.0

for name, (model, params) in models.items():
    grid_search = GridSearchCV(model, params, cv=5, scoring='accuracy')
    grid_search.fit(X_train, y_train)

#Displaying the best hyperparameters
print(f "Best parameters for {name}: {grid_search.best_params}")
```

```

# Evaluation on the test data
y_pred = grid_search.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f "Accuracy for {name}: {accuracy}")

if accuracy > best_accuracy:
    best_model = grid_search.best_estimator_
    best_accuracy = accuracy

#Confusion Matrix
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues", xticklabels=['0', '1'], yticklabels=['0', '1'])
plt.title(f "Confusion Matrix - {name}")
plt.xlabel("Predicted")
plt.ylabel("Actual")

# Saving plots as png
plot_filename = f"{name}_confusion_matrix.png"
plt.savefig(plot_filename, format='png', bbox_inches='tight')

#Save best model and used scaler
joblib.dump(scaler, 'scaler.joblib')
joblib.dump(best_model, 'best_model.joblib')

```

Figure 34: Python code for developing a machine learning model for lead scoring
Source: Own representation

Appendix 11: Python code for applying the machine learning model

```
import pandas as pd
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
import joblib
import os

# Load model and scaler
os.chdir('working_folder_path')
loaded_model = joblib.load('best_model.joblib')
scaler = joblib.load('scaler.joblib')

# Load the data on which a prediction has to be made
df_new = pd.read_excel('lead_overview.xlsx')
X_new = df_new.drop(["Converted", "lead_id"], axis=1)
X_new_scaled = scaler.transform(X_new)

# Make predictions
predictions = loaded_model.predict(X_new_scaled)

# Create df to be exported
df_predictions = pd.DataFrame({
    'lead_id': df_new['lead_id'],
    'Predicted_Converted': predictions
})

# Export df as csv
df_predictions.to_csv('new_predictions.csv', index=False)
```

Figure 35: Python code for the application of the machine learning model
Source: Own representation