Measuring the Success of Recommender Systems: A PLS-SEM Approach

MAXIMILIAN GOTTHARDT AND VITALIY MEZHUYEV, (Senior Member, IEEE)

Institute of Industrial Management, FH Joanneum University of Applied Sciences, 8605 Kapfenberg, Austria

Corresponding author: Vitaliy Mezhuyev (vitaliy.mezhuyev@fh-joanneum.at)

ABSTRACT Recommender systems, which suggest relevant products to internet users, have become an integral part of our daily lives. The factors responsible for their success from the different stakeholder perspectives, however, have never been thoroughly investigated. This study proposes a novel model for measuring the success of recommender systems that consolidates different success factors. The model is a modified version of the DeLone and McLean Information Systems Success Model with trust as an additional latent variable. The model was evaluated in an empirical study with PLS-SEM. The proposed model exhibits a high predictive power and all structural paths were significant. The integration of trust is an important contribution as the path between information quality and trust yielded the highest path coefficient. The proposed model can be used by recommendation system providers to explain and predict the successful use of the systems and to improve business processes.

INDEX TERMS Information systems success model, PLS-SEM, recommender systems, success factors.

I. INTRODUCTION

Over the recent years, recommender systems have found their way into everyday lives. These ubiquitous tools and techniques suggest items[1] of interest to internet users. The items can have a different nature, for example, it can be a product to buy or a movie, streaming on Netflix. The main goal of a recommender system is to ease decision-making in the case of several alternatives, by providing appropriate information and relevant selection options [1]. In many cases, suggestions are personalized, e.g., book recommendations on e-Commerce platforms like Amazon.com, which take into account user preferences in the best way [2].

The business reasons for implementing recommender systems are selling more items, diversifying items sold, improving user satisfaction, increasing user fidelity, evaluating user preferences among others. All of these may be attributed to an increase in business value. Both researchers and industry representatives have shared interests in the domain, which resonates with data accessibility and the steady improvement of algorithms [1]. Recommender systems are, in fact, a good example of large-scale use of machine learning techniques in commercial applications [3].

The effect of recommender systems is difficult to judge, let alone the determination of whether such a recommender system was successful. The success may include achievements from an economic point of view but is likely not limited to this aspect. Currently, only a few companies like booking.com [4] or Airbnb [5] publish their findings and metrics on which they evaluate the success (e.g. economic feasibility) of a recommender system. Hence, it appears that the evaluation of the actual success or the determination of key contributions to the success of recommender systems are oftentimes neglected.

Consequently, a clear research gap can be identified. There is no specific model that assesses the success factors of recommender systems. To the best of our knowledge, there is currently no study that gathered different factors and metrics from literature and puts them to the test in a success model.

This study aimed to examine the different success factors of recommender systems. In essence, it proposed a model that is based on the DeLone and McLean information system success model [6], [7] and modified it for the application to the recommender system domain.

The proposed model was examined at the interviews with recommender systems experts and next in the empirical study with PLS-SEM. The result is a comprehensive model that portrays the success of recommender systems.
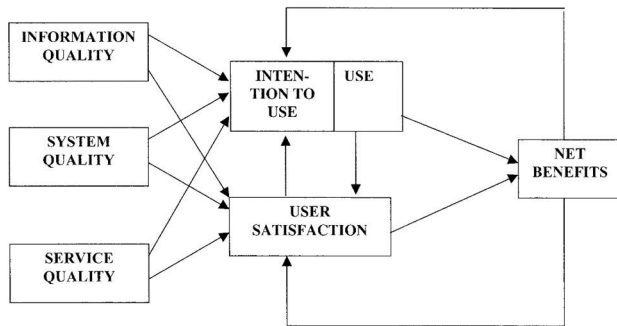

The associate editor coordinating the review of this manuscript and approving it for publication was Nikhil Padhi.

[1] An item refers to an entity that is being suggested to the user [1].

FIGURE 1. DeLone and McLean model for information systems success.

This paper is organized as follows. The first chapter provides a short introduction to the research topic, formulates a problem statement and the research goals.

Chapter 2 introduces the underlying DeLone and McLean Information Systems Success Model. It also discusses the reasons for using PLS-SEM and its possible implications.

Chapter 3 provides a comprehensive methodological overview of the study. It discusses stages of the model development and evaluation. The methodology of the empirical study is explained, ranging from the questionnaire formulation to the process of conducting the survey. The measures and implications of data preparation are also described.

Chapter 4 is the results and discussion section. First, the data evaluation process is discussed and descriptive statistics are reported. Then, the proposed model is analysed and discussed.

Chapter 5 is the conclusion section, which summarizes the main findings of the study.

References list finalizes the study.

II. THEORY
A. THE DeLone AND McLean INFORMATION SYSTEMS SUCCESS MODEL
The model that was applied in this research is based on the DeLone and McLean information systems success model (D-M model). The D-M model was introduced in 1992 in an effort to model information systems success [6]. 10 years later, in 2002, the model was updated and revised by DeLone and McLean [7]. The model was updated since empirical research has failed to account for the interdependent and multidimensional nature of information systems success. And more, the model was not applied uniformly as studies tested different relationships between the constructs [8]. The updated by DeLone and McLean model is depicted in Figure 1.

One of the main changes was the introduction of the service quality dimension, which represents the significance of service and support in e-Commerce [9]. A second change of the original model included the "intention to use" construct, which represents the attitude of users, as an alternative to the "use" [9]. In the context of recommender systems, adding "intention to use" allows us to highlight and measure the reasons why users utilize internet services.

The third change was the combination of organizational impact and individual impact to a single net benefits construct [9]. DeLone and McLean argue that the use of net benefits might be more parsimonious but likely the most accurate representation of the final success variable. Net benefits regard the fact that no result is entirely positive or negative. Different stakeholders have different views on what is a benefit to them. Thus, one of the main advantages is that the construct allows to assess benefits at various levels.

The model allows scholars to define the research context and set the focus to whatever is deemed more relevant. Urbach and Müller proposed a way to interpret the model [9]. An information system may be evaluated by considering information quality, system quality and service quality. These characteristics influence the use or intention to use but also the user satisfaction. Benefits will be gained by using the system. These net benefits (either positive or negative) in turn influence the user satisfaction and use of the system [9].

The enhanced D-M model was shown to be a reliable framework for IS success measurements. It is commonly used to assess and understand the dimensions that make up IS success. Hence, applying this model to understand recommender systems success is a reasonable approach that is also backed up by previous literature.

B. PLS-SEM
This section argues and explains why Partial Least Squares - Structural Equation Modelling (PLS-SEM) was chosen for this study and discusses the implications of using PLS-SEM.

1) INTRODUCTION TO PLS-SEM
The statistical basis for PLS-SEM was developed between the mid-1970s and mid-1980s. The approach was originally designed for social and behavioural sciences but since has gained considerable popularity in business management research. PLS-SEM approximates partial model structures, which are defined by a path model. It combines both principal components analysis and conventional least squares regressions. PLS-SEM is often viewed as an alternative to CB-SEM, a covariance-based technique, due to the less restrictive demands regarding data distribution, types of variables and actual sample size [10], [11].

The basic principles deserve further elaboration. PLS-SEM is a multivariate approach that applies statistical methods to analyse multiple variables simultaneously. SEM allows researchers to use unobservable variables.[2] These variables can be, for example, abstract concepts like trust or satisfaction which cannot be assessed directly. Instead, a set of indicators are measured which in turn are used to estimate the unobservable variables. They can be considered proxy variables that describe an aspect of a much larger, but abstract concept. Thus, combining multiple items may be used to assess a construct [12].

---

[2]The unobservable variables are often called latent variables or constructs.

2) CHARACTERISTICS OF PLS-SEM

Compared to its alternatives, PLS-SEM is applicable when the underlying theory is not thoroughly developed. Rigdon provides arguments supporting this claim [13]. In the context of our study, we assumed that PLS-SEM can handle the exploratory components and the associated uncertainties well.

Another benefit is the higher statistical power achieved with PLS-SEM in comparison to factor-based SEM. Consequently, PLS-SEM has a higher chance of discovering significant effects when they are indeed significant [14]. According to Chin [15], higher statistical power allows applications in exploratory research where the theoretical knowledge is limited, and where the priority might lie in identifying substantial effects [14], [15].

Some key data characteristics should be emphasized. PLS-SEM is a non-parametric method. Thus, no assumptions on the distribution of available data are made, which allows for analysing non-normally distributed data. PLS-SEM also handles small sample sizes well and still reaches high levels of statistical power without identification issues. Small sample sizes generally work well and do not have an impact on the biases in most cases, e.g. when examining effects like multi-collinearity or misspecification. PLS-SEM also handles missing values effectively if they do not exceed a reasonable level (e.g. below 5% for each indicator). Then, missing value treatments like mean replacement can be applied with a little impact on the results [12].

Another aspect is the possibility to evaluate item weights. Individual item weights can provide valuable insights as one can determine the relative importance of an item's contribution to the composite variable within a specific context. In other words, researchers can use the item weights to describe the relationship between an item and other composites in the model. When measuring a construct like user satisfaction, scholars can determine which items are relevant for user satisfaction [12]. However, it must be accepted that the proxies are not equivalent to the constructs they replace. These proxies are weighted composites of other variables that create the model together. They are approximations and act as a stand-in for the construct itself [13].

When the weights are determined, the algorithm shows a specific score for every composite and every observation [17]. Then, ordinary least squares regression is used under the premise of minimizing the error associated with the endogenous constructs. Thus, PLS-SEM approximates the path model relationships, i.e., the coefficients which maximize the R^2 of each endogenous construct. According to Hair *et al.*, it is a reason why PLS-SEM is a viable method for theory development and explanation of variance, i.e., prediction of constructs [12]. It is also a reason why PLS-SEM is often referred to as a variance-based approach [12]. These features contributed to choosing PLS-SEM over its alternatives for this study.

Further, PLS-SEM can handle complex models, i.e. models that incorporate a large number of structural model relations. Both reflective and formative measurement models can be implemented. The number of items describing a construct can be quite diverse as well. PLS-SEM can manage constructs with multiple items and single-item constructs. However, in its original form PLS-SEM cannot handle circular relationships [12].

Another limitation of PLS-SEM is that there is no widely established goodness-of-fit measure. Thus, its application for testing and confirming theories is generally limited. Literature advises caution when using goodness-of-fit measures to validate PLS-SEM models [14]. Some authors have, nevertheless, have sought to address this issue. For example, Henseler *et al.* applied the standardized root mean square residual (SRMR) measure to approve or validate the underlying measurement Model [18]. It determines the squared difference between the correlations implied by the model and the observed correlations [18]. A more detailed description can be found in the work of Hu and Bentler [19].

III. DATA AND METHODS

The exploratory character of the study demands a clear framework that serves as a general guideline. This chapter serves as a comprehensive methodological reference for the study. The application of SEM shapes the approach taken throughout the model design, instrument development and data analysis phases.

A. MODEL DESIGN

As previously discussed, the proposed model is based on the DeLone and McLean (D-M) information systems success model. To serve the purpose of research, the original D-M model was modified.

Certain changes were made because of PLS-SEM constraints: due to the inability to handle circular relationships [12], the reverse impact of net benefits on usage intentions/use and user satisfaction was removed. Further, the relationship of usage intentions/use and user satisfaction is bidirectional, which cannot be handled by PLS-SEM as well.

There are samples of the D-M model modifications, connecting usage intentions/use to the user satisfaction construct [20], i.e. with the path arrow, which points from the use towards user satisfaction. For the purpose of this study, a connection from user satisfaction to use was established. The rationale behind this is that the satisfaction, triggered by the use of a recommender system, causes an increase in the use. Further, it could be argued that the indicators for the use construct reflect the intention to use aspect better. This choice is also supported by the results of the empirical study - the path coefficient is higher when user satisfaction is connected to the use and not the other way around.

As compared with the original model, the service quality dimension is omitted. In the case of recommender systems, direct contact between users and the support team of the recommender systems provider is rather rare. Commonly,
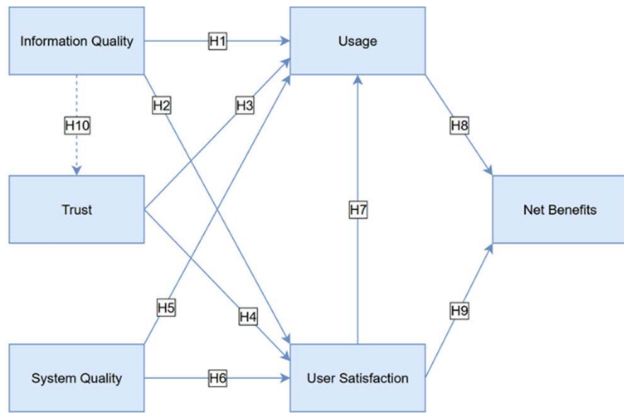
FIGURE 2. Proposed model with research hypotheses.

Hypotheses	
H1	Information Quality has a significant positive effect on Usage.
H2	Information Quality has a significant positive effect on User Satisfaction.
H3	Trust has a significant positive effect on Usage.
H4	Trust has a significant positive effect on User Satisfaction.
H5	System Quality has a significant positive effect on Usage.
H6	System Quality has a significant positive effect on User Satisfaction.
H7	User Satisfaction has a significant positive effect on Usage.
H8	Usage has a significant positive effect on Net Benefits.
H9	User Satisfaction has a significant positive effect on Net Benefits.
H10	Information Quality has a significant positive effect on Trust.

recommender system providers do not have dedicated support for their recommendation engine.

Another significant change is that trust was added as a construct. The reasons for that supported by relevant literature references are provided in the next chapters.

1) HYPOTHESES

The hypotheses under investigation are listed in Table 1. They essentially describe whether the constructs have a significant positive effect on others constructs.[3] The hypotheses correspond to the paths in the proposed model.

Figure 2 shows the proposed model with the respective hypotheses.

Since trust is a newly introduced latent variable and was not a part of the original D-M model, underlying hypotheses need to be discussed. First, it is assumed that trust has a significant effect on both usage (H3) and user satisfaction (H4). The idea behind H3 is that when users trust the system, it has a positive effect on the usage of the system. For example, users tend to use the system more often if they trust it. Secondly, H4 assumes that trust has a significant positive effect on user satisfaction. For instance, if users trust the system, they are more likely to appreciate the assistance of the system. The path between information quality and trust (H10) proposes that information quality affects trust in a positive way. For example, when the recommendation is accurate, up-to-date and easy to grasp, then the trust in the system is higher.

2) MEASURES OF THE CONSTRUCTS

The following section elaborates on the constructs and items of the model. The items for the constructs were chosen based on a systematic review of applying the D-M model literature sources.

a: SYSTEM QUALITY

System quality encompasses the preferred characteristics of an e-Commerce system. Thus, integrating measures of the

---

[3]The significance of effects is evaluated by examining the p-values in SmartPLS bootstrapping algorithm. The effect is considered significant when the p-values are below 0.05, whereas values exceeding this threshold are deemed not significant.

system itself to form the construct is a viable approach. System quality measures typically tend to focus on measures describing usability and performance [9]. The relative importance of the underlying measures is subject to change depending on the environment and context. In the e-Commerce domain, users are most likely customers and not employees. In that case, using the service is a voluntary decision. Poor usability or usefulness would deter customers from using the service. In addition, if the service quality is insufficient many benefits may not be realized [21].

Table 2 shows the proposed measures that reflect the system quality of recommender systems.

Multiple criteria were used to evaluate whether a measure fits into the model. Firstly, it should portray an important feature or functionality of a recommender system. Secondly, the measure should align with that applied in the D-M model. For example, the works of DeLone and McLean [7] and Urbach and Müller [9] give a comprehensive list of measures that can and have been used for system quality. Also, the measure should be used in the recent and relevant literature, preferably in the context of recommender systems and PLS-SEM. This includes studies conducted by Nilashi *et al.* [23], Ali *et al.* [22], and Ramadhanti and Slamet [20].

b: INFORMATION QUALITY

Information quality includes preferable characteristics of the information system's output [9]. For recommender systems, it includes the generated information, when a user looking for an item on an e-Commerce website. Information quality does not only include the quality of retrieved information; it also encompasses the usefulness of information [9].

In the e-Commerce domain, the website is a key factor when it comes to information quality. It supplies information on different products and services which helps customers to buy products [22]. This is also a front-end where recommender systems come into play. Recommender systems generate information, describing items to be presented to the users.

TABLE 2. Service quality measures.

Index	Measure	Source
SYQ_1	Response Time	DeLone and McLean (2004) [21], Urbach and Müller (2012) [9], Ali et al. (2018) [22]
SYQ_2	Ease of Use	DeLone and McLean (2004) [21], Urbach and Müller (2012) [9], Nilashi et al. (2016) [23]
SYQ_3	Navigation	DeLone and McLean (2004) [21], Urbach and Müller (2012) [9], Ramadhanti and Slamet (2020) [20]
SYQ_4	Accuracy	Urbach and Müller (2012) [9], Jannach and Gediminas (2016) [24]
SYQ_5	Conversion	Jannach and Jugovac (2019) [25]
SYQ_6	Learning	Elahi et al. (2016) [26], Tiwari et al. (2018) [27], Zhang et al. (2019) [28]

TABLE 3. Information quality measures.

Index	Measure	Source
INQ_1	Up-to-date information	Ramadhanti and Slamet (2020) [20]
INQ_2	Level of detail	Nilashi et al. (2016) [23], Ali et al. (2019) [22]
INQ_3	Understandability	DeLone and McLean (2004) [21], Urbach and Müller (2012) [9], Nilashi et al. (2016) [23], Ramadhanti and Slamet (2020) [20]
INQ_4	Accuracy	DeLone and McLean (2004) [21], Urbach and Müller (2012) [9], Nilashi et al. (2016) [23], Ali et al. (2018) [22], Ramadhanti and Slamet (2020) [20]

Presented in Table 3 measures characterise the quality of information that recommender systems provide. They align with the items provided by DeLone and McLean [7] and are used in the recent and relevant literature. For all given in Table 3 information quality measures, supportive studies both in recommender system and applied PLS-SEM were found.

c: TRUST

Trust is an important matter of discussion in the recommender system domain. In fact, the literature covers it extensively in different contexts. Some studies closely relate trust to security and privacy [29]–[31]. For one, recommenders may have access to personal user information and thus there is a possibility of data leaks. In addition, attacks on recommenders can affect recommendations, e.g. shilling attacks can result in biased recommendations [29]. Trust can also be used to generate more accurate recommendations, e.g. movie recommendations relying on trust ratings of users in social networks [32]. Another study investigated the effects of trust on the acceptance of recommendations [33].

Trust and privacy were introduced as an addition to the D-M model in the study by Ali *et al.* [22]. The study evaluated e-Commerce success using a modified D-M model. The evaluation was performed with PLS-SEM. The study concluded that trust and privacy affect user satisfaction in addition to system and service quality [22]. Authors [23] assumed that trust is one of the primary factors that influence the success of recommender systems in the e-Commerce domain. They

TABLE 4. Trust measures.

Index	Measure	Source
TRU_1	Trustworthiness	Nilashi et al. (2016) [23], Ali et al. (2018) [22]
TRU_2	Trust in benefits	
TRU_3	Explanations	Nilashi et al. (2016) [23]
TRU_4	Privacy	Calandrino et al. (2011) [34], Ali et al. (2018) [22], Sharma and Aggarwal (2019) [35]

applied a trust model on two e-Commerce sites where trust is influenced by website quality, recommendation quality and transparency. The study was evaluated using PLS-SEM and reported many intriguing findings. For example, it found that solely focusing on recommendation quality is not enough; and to increase the adoption of recommendations, factors influencing trust also need to be considered.

As the importance of trust and its consideration in literature has been established, the items can be examined more closely. The items to measure trust are shown in Table 4.

Since trust is not a part of the original D-M model, further clarification of the measures may be required. Trustworthiness refers to what extent users consider a recommender system as reliable and faithful. If a recommender is not deemed as trustworthy, users may refrain from using it. The trust in benefits measure assesses the user's beliefs in the benefits that users may potentially gain from using the system. The explanations measure ascertains if explanations help users to build trust in the recommender system. Explanations can, for example, be a clarification of why these specific items were recommended to a user. In this context, some studies also relate transparency to explanations [23]. Privacy is another factor affecting trust. For the purpose of this study, the privacy measure focuses on the confidentiality of handling users' data. In other words, it ascertains if the prevention of third-party access or inferring user data should be considered a priority.

d: INTENTION TO USE/USE

The intention to use/use construct describes the possibility and nature of the usage of an information system (IS). The use of an IS is a very wide construct, which can be assessed with a variety of different measures. When using an IS, e.g. a recommender system, on a voluntary basis, its usage can be representative of its success. On its own, the timespan a system is used is not a success measure. As there are problems in interpreting the use as a dimension, DeLone and McLean proposed the intention to use it as an alternative [7].

The technology acceptance model (TAM) proposed by Davis [36] provides a thorough concept that can be applied to describe the use of an IS. The TAM variables are independent and include perceived ease of use, attitude toward use, intention to use, actual use and perceived usefulness [36]. It covers the dimension of system use to a large extent [16].

The measures, adapted to the domain of recommender systems, are presented in Table 5.

TABLE 5. Intention to use/use measures.

Index	Measure	Source
USE_1	User activity	DeLone and McLean (2004) [21], Urbach and Müller (2012) [9], Petter et al. (2013) [37], Jannach and Jugovac (2019) [25]
USE_2	Coverage	Avazpour et al. (2014) [38], Gunawardana and Shani (2015) [39], Jannach and Adomavicius (2016) [24]
USE_3	Retention	DeLone and McLean (2002) [7], Urbach and Müller (2012) [9], Tehseen et al. (2019) [40]
USE_4	Decision making	Petter et al. (2013) [37], Jannach and Adomavicius (2016) [24]

Note that while the intention to reuse is a part of the use construct in Urbach and Müller [9], DeLone and McLean [7] have assigned customer retention to the net benefits construct. In the context of recommender systems, arguments for fitting retention of users either construct can be made. For this study, retention was placed in the use/intention to use construct as the original definition of the construct entails both nature and amount of use. However, one can argue that from the provider perspective it would make sense to place customer retention in the net benefits section if it is an important element of the business model.

Petter *et al.* list the purpose of use as a valid measure of system use [37]. Ultimately, the purpose of using recommender systems includes easing the decision-making process as they are often designed to help users to cope with information overload [41].

e: USER SATISFACTION

As the name implies, user satisfaction measures the level of satisfaction that is achieved when using the information system [37]. It is, in fact, considered as one of the most important constructs in the D-M model. User satisfaction is particularly important when the use of the system is mandatory or involuntary and when the amount of use is considered as an inappropriate indicator of system success [9]. For the recommender systems, their use is often involuntary, but rather the use of the service is voluntary. Thus, user satisfaction could be one of the most important measures as the recommender system can deter users from using the service at all (for example, when the recommender constantly provides irrelevant results). Table 6 outlines the items that were used to measure user satisfaction in this study.

System satisfaction refers to whether the system performance and functionality meet user expectations. Appreciation assesses if the recommender system is generally welcomed by its users. The overall satisfaction is intended to assess the general level of satisfaction that the recommender system provides. It serves as a way of identifying to what extent delivering satisfaction to the user is important for a recommender. For example, a recommender system can provide good results without triggering a feeling of satisfaction in the user. A user might not even notice the presence of a recommender system and, as a consequence, will not

TABLE 6. User satisfaction measures.

Index	Measure	Source
USS_1	System Satisfaction	Urbach and Müller (2012) [9], Ali et al. (2018) [22], Yang et al. (2020) [42]
USS_2	Appreciation	Yang et al. (2020) [42]
USS_3	Overall Satisfaction	Urbach and Müller (2012) [9], Ali et al. (2018) [22], Yang (2020) [42]
USS_4	Persuasion	Yang (2020) [42]

associate the results with the underlying recommender system. Lastly, information persuasion aims to describe whether users believe that going for the suggestions is a good decision. The measure determines to what level the recommender can convince users of their propositions and ultimately influence user satisfaction.

f: NET BENEFITS

Net benefits describe to what degree information systems influence the success of the various stakeholders. As mentioned previously, the updated D-M model combines the constructs of individual and organizational impact to a single construct. The choice, which impact should be examined ultimately depends on the study purpose and depth of analysis. Some researchers evaluate net benefits as investments, through financial measures like market share, profitability, ROI etc. Others avoid these quantifiable measures as in many cases benefits may not be quantifiable using numerical measures. However, a significant number of studies assess the benefits using both individual and organizational dimensions [9].

Table 7 provides an overview of the net benefits measures used in this study.

User engagement describes the intensity of involvement that users experience when using a recommender system. For example, users can actively apply the recommender system options, or the system may run in the background and users even might not notice it. Profit refers to the income that the provider can achieve with a recommender system.

Further, customer acquisition refers to the increase in the number of users induced by the recommendations. Nevertheless, a large increase of users is not necessarily a long-term benefit. A sudden increase in users, e.g. due to a sales promotion can be temporary. Thus, a measure that indicates whether the users will remain active and keep on using the services is needed. Customer loyalty (or customer retention) was added to the list of measures as it represents if users keep on using the service. It may be used to indicate how the recommender system influences loyalty as well. Productivity describes all aspects related to user output.

Competitive advantage is a measure outlining the level of differentiation the provider can achieve considering competitors. Jannach and Adomavicius [24] discuss the recommender's ability to differentiate the providers' service from its competitors as an important purpose of a recommender.

TABLE 7. Net benefits measures.

Index	Measure	Source
NEB_1	User engagement	Jannach and Adomavicius (2016) [24]
NEB_2	Profit	DeLone and McLean (2004) [21], Ramadhanti and Slamet (2020) [20]
NEB_3	Customer acquisition	DeLone and McLean (2004) [21]
NEB_4	Customer loyalty	DeLone and McLean (2004) [21], Ramadhanti and Slamet (2020) [20]
NEB_5	Productivity	DeLone and McLean (2004) [21], Urbach and Müller (2012) [9]
NEB_6	Competitive Advantage	DeLone and McLean (2004) [21], Urbach and Müller [9]
NEB_7	Learning User Preferences	Jannach and Adomavicius (2016) [24]
NEB_8	Willingness-to-pay	Adomavicius et al. (2017) [43]

Further, a competitive advantage can be achieved by increasing the switching costs, which are costs incurred when a user is trying to switch to a competitor. The study by Sharma and Aggarwal discusses different factors that allow e-businesses to gain a competitive advantage over other businesses [35].

The ability of a recommender system to collect, provide and then apply the user-related information is reflected in the "learning user preferences" measure. The ability to understand user preferences is obviously important for the provider. A provider can, for example, determine user preferences directly by allowing users to express their preferences [24]. Finally, willingness-to-pay represents the recommender systems effect on the user's decision to pay for the services or to buy a proposed item. A study by Adomavicius et al. discusses the measure in more detail [43].

B. EMPIRICAL STUDY

Data gathering was conducted by the means of an empirical study (survey). The steps of the process are elaborated in this section.

1) QUESTIONNAIRE

The questionnaire was created using Google Forms and contains a short introduction, a demographic part and the model evaluation-related questions. The questionnaire incorporated a total of 36 questions, of which 30 questions were the items for the model evaluation. The time to answer the questionnaire was assumed to be around 12-15 minutes.

After allocation of the items from the literature, they were checked in expert interviews. Three independent experts in the recommender systems domain were consulted via e-mail and Skype calls. With the help of expert feedback, both the model and the questionnaire were significantly improved.

The choice of survey candidates was a key task. The candidates should have a certain level of knowledge in the domain of the recommender systems to reasonably assess the importance of their success factors. An ideal way to find experts would be to contact researchers, who have recently published on the topic of recommender systems.

The demographic part (which asks for gender, age, education level and business domain) was moved to the end of the survey. This ensures that the candidates do not fatigue early in the survey process and keep a high level of concentration when answering the questions related to the research model.

Because of PLS-SEM, the measurement scales used in the questionnaire require clarification. The demographic questions, where respondents were asked e.g. to indicate their profession, have a nominal scale. For the questions that were used to evaluate the model, a five-point Likert scale was used. As Hair et al. noted, both ratio scales and interval scales can be used with multivariate analysis [12]. However, the Likert scale is an ordinal one and, consequentially, an equidistance between the scale points needs to be assured. A five-point Likert scale with the categories of strongly disagree (1), disagree (2), neutral (3), agree (4), strongly agree (5) assumes that the distances between strongly disagree and disagree and also neutral and agree are the same.

A Likert scale for PLS-SEM should have symmetry of items around the middle (neutral) category and clear linguistic qualifiers for the respective category. Because of the symmetric scaling, the equidistance between items is more likely to be perceived. When the Likert scale can be considered as both equidistant and symmetrical it can be compared to an interval scale. Thus, the ordinal Likert scale can be considered as an approximated interval scale that can be used for PLS-SEM [12].

2) DATA COLLECTION

The main source for collecting contact data was research papers, published by the target audience. There are important criteria for papers selection. The source should be recent and relevant to the domain of recommender systems. A recent publication implies that survey candidates have recently been active in the domain. Hence, it can be assumed that the answers of respondents reflect the current state of the domain. Therefore, the publication timeframe for literature sources was primarily limited to the period between 2017 and 2020.

During this study, more than 3300 papers were examined. Contact data were gathered from the papers using a PDF scraper, programmed in python. The approach greatly increased the efficiency of the data collection process as the time to collect e-mails of responders was significantly reduced. To ensure that the e-mail addresses are unique, duplicate e-mail addresses were removed from the database.

3) CONDUCTING THE SURVEY

At the first stage, an invitation to participate in the survey was sent out. The invitation informs candidates why they were contacted. The purpose of the survey was briefly explained and how the input of responders will be evaluated. The estimated survey time was mentioned and then the survey

link was provided. It was mentioned as an incentive, that the research results will be provided to responders if they request it. Candidates were also invited to provide feedback and to contact the authors in case they have any questions. The survey also outlines confidentiality and reminds potential responders that data will be processed anonymously and treated as strictly confidential. After one week, a reminder was sent out in order to increase the response rate.

C. DATA PREPARATION

Data, gathered during the survey process, are not flawless. For example, some values could be missing, the survey could have accidentally been submitted multiple times etc. This section briefly discusses the steps taken in the context of data preparation.

Recommendations on how to properly conduct data preparation for PLS-SEM are provided by Hair *et al.* [12]. The problems and solutions that were applied for this study are shown in Table 8.

In addition to the rules of thumb, outlined in Table 8, other options to deal with missing data are also available in SmartPLS 3. Case-wise deletion would remove certain observations if values were missing. Pairwise deletion, on the other hand, uses all valid data and ignores missing values. It is advisable to do this when large amounts of data are missing, i.e. when mean-replacement is not reasonable. Lastly, regression approaches can be used but they are generally advised against.

Suspicious response patterns can be identified using visual examination, mean, variance, and check of the values. A typical response pattern would be straight-lining, where users choose the same answers for the majority of their responses. Such responses should be removed.

Outliers are usually extreme responses and should be interpreted in the study's context. If a researcher can explain the reasons for outliers, they should be kept in the study. Nevertheless, the impact of outliers on the study should always be considered.

Another point to examine was the kurtosis and skewness of the data. While PLS-SEM can handle non-normally distributed data, it should not be extremely non-normal. Extreme non-normal distributions can lead to problems when it comes to determining the parameter's significances [12].

IV. RESULTS AND DISCUSSION

A. DATA EVALUATION AND FURTHER PREPARATION

Out of the 6289 candidates contacted, 133 responses were received, from which 120 were selected as valid. The response rate of valid responses is about 1.9 %. The low response rate can be partially attributed to the fact that the survey was conducted during the holiday season. Three empty responses were removed. Verification of duplications was done by the analysis of timestamps of responses in Google Sheets. The timestamps revealed that the identical entries were in fact consecutive submissions and therefore the duplicate entries were removed.

TABLE 8. Considerations regarding data preparation.

Problem	Solution
Missing data	Discard an observation, if more than 15% of data is missing. Mean value replacement, if less than 5% of indicator values are missing.
Suspicious Response patterns	Identify and remove if necessary
Outliers	Identify and either correct them or keep them
Data Distribution	Skewness within -1 and +1 Kurtosis within -1 and +1 for normal distributions

Suspicious response patterns, e.g. straight-liners, were identified by checking the standard deviation of the responses (cases, when the standard deviation is zero). Further, a visual examination was also applied. If only one or two answers differed from the rest of the answers, then they could also be considered as straight-lining as well. In this study, 8 unique answers exhibit straight-lining and were deleted following Hair *et al.* advice on the "garbage in, garbage out" rule, meaning that inexplicable research results are often the consequence of improper data [12].

The number of missing values in the remaining set of observations was remarkably low. In the selected as valid 120 responses, a total of only 10 missing values occurred. Out of the 30 model related questions, 22 do not have any missing value. Also, no responder left out more than one model related question in the questionnaire. Regarding the indicators, two missing answers occurred only for one indicator. This corresponds to 1,67% of missing values. Following the criterion of Hair *et al.* [12], a mean replacement can be used in SmartPLS 3 as it is below 5%.

B. DEMOGRAPHIC DATA EVALUATION

Almost 72 % of the respondents were male and more than 28% of respondents were female. Most respondents were between 30-39 years old (38.3%), followed by the group respondents between 20-29 years of age (30%). These two subgroups form the majority of responses. Then, the number of responses decreases as the age groups increase. All respondents have completed a form of higher education starting with a bachelor's degree.[4] In fact, the number of valid responses increases with higher levels of education. Almost 62% of responders have received doctorate degrees followed by close to 31% of master's degrees with the rest carrying bachelor's degrees.

C. MODEL EVALUATION

1) REFLECTIVE MEASUREMENT MODEL EVALUATION

a: CONVERGENT VALIDITY AND INTERNAL CONSISTENCY RELIABILITY

The first step includes the evaluation of the indicator loadings and their impact on the average variance explained (AVE).

---

[4]The fact that all responders have a high educational background can be easily explained by the allocation of e-mail addresses from research publications.

The review of indicator loadings and the AVE are part of the convergent validity concept. It describes to what extent the indicators correlate with other remaining indicators of the particular construct. Higher outer loadings mean that the indicators are quite similar [46]. Note, that reflective measurement implies that the indicators are similar because the aim is a maximization of the overlap in-between indicators [12].

Loadings above 0.70 are generally accepted and recommended as it implies that 50% of the variance of the indicator is explained. This is because the indicator reliability, i.e. amount of explained variance, is the square of its loading [10]. Examination of the loadings in Table 9 shows that all indicators fulfil this requirement.

There are many other measures for assessing construct validity, including Cronbach's alpha and composite reliability. Higher values imply higher reliability. For composite reliability, values between 0.6 and 0.7 are deemed acceptable when performing exploratory research. Further, values ranging from 0.7 to 0.9 reach from ''satisfactory to good''. Cronbach's alpha relies on similar threshold values but generally exhibits lower values. While the values of Cronbach's alpha are unweighted and therefore less precise, composite reliability exhibits higher reliability because indicators are weighted based on individual indicator loadings of the construct. The composite reliability might be too permissive, whereas Cronbach's alpha is likely too conservative. The actual reliability may lie somewhere in between those two measures [10], [46].

b: DISCRIMINANT VALIDITY

There is a lot of discussion on the topic of establishing discriminant validity. The two traditional measures are the assessment of cross-loadings and the evaluation of the Fornell-Larcker criterion. For cross-loadings, the cross-loadings for the latent variable itself should be higher than the ones on the other latent variable. A matrix is evaluated where the non-diagonal elements are the correlations between the constructs. The diagonal is the square root of the AVE. Discriminant validity is established when the square root of each AVE value, i.e. the values in the diagonal, are greater than the values below the diagonal. These are the correlations of the construct to the other remaining constructs. Nevertheless, both criteria are not considered reliable when it comes to the determination of discriminant validity [47], [12].

Instead, authors like Henseler *et al.* have proposed the Heterotrait-Monotrait ratio of correlations (HTMT) for the verification of discriminant validity. It measures the indicator correlations throughout the latent variables relative to the correlations within the latent variable [47]. It can either be used as a criterion, where values are compared to a threshold or as a statistical test that relies on bootstrapping. If it is used as a criterion, the threshold for the HTMT is subject to debate. In many cases, either 0.85 or 0.9 are proposed. For example, when the threshold is 0.9 and the HTMT is below 0.9, then discriminant validity is established. When the

HTMT is used as a statistical test, then bootstrapping is used to determine confidence intervals for the HTMT. The intervals should not contain the value one to confirm discriminant validity. In their simulation study, Henseler *et al.* used a one-tailed 90% bias-corrected confidence interval and examined whether it includes the value one [47]. A recent study by Hair *et al.* indicated that this approach is contemporary and suggest the same criteria and threshold values [48].

All variants were found to reliably detect discriminant validity and their main difference is their specificity. The choice often depends on sample size as the inferred HTMT does not perform as well for larger sample sizes [47]. For the evaluation of the proposed model, an HTMT threshold of 0.9 was assumed. Initially, discriminant validity was not supported given as the values exceeded 0.9 twice (for usage->system quality and user satisfaction->system quality). For handling problems regarding HTMT authors follow the guidelines proposed by Henseler *et al.* [47]. To decrease the HTMT we consider the elimination of items that have a strong correlation to the other construct. Another option would be reassigning the problematic indicators to the other latent variable in case the underlying theory allows it [47].

Following these recommendations, the SYQ_2 measure was removed from the system quality construct. It also had the lowest outer loading of the remaining indicators. The HTMT results for evaluating discriminant validity are shown in Table 10.

The system quality->information quality HTMT is high but still below the threshold. Two further cases (trust -> information quality and user satisfaction -> system quality) are approaching 0.9 too. Still, all the values are below the assumed threshold and thus discriminant validity is established.

2) STRUCTURAL MODEL EVALUATION

a: COLLINEARITY ASSESSMENT

The first step is the review of collinearity metrics to ensure that the results are unbiased. The evaluation was performed using the variance inflation factor (VIF) metric. While values above 5 indicate possible collinearity, it can still occur for VIF values between 3 and 5. Thus, VIF values should ideally be below 3. [10]

Table 11 shows a summary of the inner VIF values. In fact, all values are below 3, which ensures that the results are unbiased.

b: PATH MODEL COEFFICIENTS

The next step is the assessment of the path model coefficients. These coefficients portray the hypothesized relationships between the constructs. The values are typically between $+1$ and -1, where plus indicates a positive relationship and minus a negative relationship. The closer the values are to 1, the stronger the relationship. Consequently, the closer the values get to 0, the weaker the relationship. Bootstrapping was used to obtain the values and their confidence intervals for assessing the statistical significance [10], [12].

TABLE 9. Summary of inner VIF values.

Construct	Indicator	Convergent Validity		Internal Consistency Reliability	
		Loading	AVE	Cronbach's Alpha	Composite Reliability
Information Quality	INQ_1	0.718	0.583	0.759	0.847
	INQ_2	0.846			
	INQ_3	0.711			
	INQ_4	0.771			
Net Benefits	NEB_2	0.713	0.502	0.837	0.876
	NEB_3	0.766			
	NEB_4	0.711			
	NEB_5	0.680			
	NEB_6	0.735			
	NEB_7	0.678			
	NEB_8	0.674			
System Quality	SYQ_4	0.898	0.675	0.535	0.805
	SYQ_5	0.737			
Trust	TRU_1	0.797	0.528	0.697	0.813
	TRU_2	0.862			
	TRU_3	0.601			
	TRU_4	0.608			
Usage	USE_2	0.515	0.553	0.609	0.781
	USE_3	0.835			
	USE_4	0.836			
User Satisfaction	USS_1	0.782	0.590	0.767	0.851
	USS_2	0.783			
	USS_3	0.842			
	USS_4	0.655			

TABLE 10. Summary of convergent validity and internal consistency metrics.

	Information Quality	Net Benefits	System Quality	Trust	Usage	User Satisfaction
Information Quality						
Net Benefits	0.279					
System Quality	0.895	0.529				
Trust	0.872	0.453	0.681			
Usage	0.634	0.679	0.812	0.743		
User Satisfaction	0.700	0.498	0.851	0.738	0.815	

The recommendations of Hair *et al.* [12] for bootstrapping were applied. 5000 subsamples were used for the procedure and parallel processing was selected. For the path coefficients, basic bootstrapping is sufficient, however, for a more extended evaluation including HTMT or internal consistency measures, complete bootstrapping can be run. The Bias Correct and Accelerated Bootstrap and the two-tailed test type were selected. The significance level was set at 0.05.

Note, that both t-values and p-values can be used for the assessment. For two-tailed tests, the t value of 1.96 corresponds to a p-value of 0.05. There are examples of studies in the e-Commerce domain applying the D-M model where either t-values [20] or p-values [22] were reported. This study interpreted the p-values to evaluate the significance levels. Additionally, confidence intervals were analysed. If an interval does not include a zero, then a significant effect is

TABLE 11. Summary of HTMT values.

	Inform. Quality	Net Benefits	System Quality	Trust	Usage	User Satisfact.
Inform. Quality					2.143	2.109
Net Benefits						
System Quality					1.778	1.519
Trust					1.905	1.740
Usage		1.564				
User Satisfact.		1.564			1.844	

TABLE 12. Summary of path coefficients and significance.

		Path coef.	95% conf. Interval	t-value	p-value	p<0.05
H1	INQ -> USE	0.018	[-0.184,0.229]	0.171	0.864	No
H2	INQ -> USS	0.134	[-0.092,0.330]	1.246	0.213	No
H3	TRU -> USE	0.250	[0.057,0.439]	2.551	0.011	Yes
H4	TRU -> USS	0.300	[0.113,0.476]	3.238	0.001	Yes
H5	SYQ -> USE	0.197	[0.000,0.393]	1.978	0.048	Yes
H6	SYQ -> USS	0.375	[0.231,0.506]	5.311	0.000	Yes
H7	USS -> USE	0.338	[0.144,0.541]	3.375	0.001	Yes
H8	USE -> NEB	0.456	[0.256,0.610]	5.222	0.000	Yes
H9	USS -> NEB	0.151	[-0.010,0.301]	1.881	0.060	No

assumed. SmartPLS 3 also allows evaluating bias-corrected confidence intervals [12].

The results of the bootstrapping including the path coefficients and bias-corrected 95% confidence intervals are shown in Table 12.

The evaluation shows that three paths have p values higher than the selected significance of 0.05 ($p \geq 0.05$). These paths are INQ->USE, INQ->USS and USS->NEB in descending order of p-values. The p-value of INQ -> USE is extremely high at 0.864 and the path coefficient is close to zero (0.018). This means that a fundamental change in the model is necessary, e.g. the path needs to be deleted or the information quality construct needs to be re-evaluated (e.g. replaced by a more suitable construct, redefined or removed). The fact that the p-value of INQ->USS is 0.213 also supports the idea of reviewing the information quality construct. USS->NEB has a p-value of 0.060 indicating that minor changes are required to get a significant result ($p < 0.05$). Two of the problematic paths lead to or from the USS construct and thus attention must be paid to this construct as well.

A review of the confidence intervals shows that the paths, which did not pass the significance test, contain a zero in their confidence intervals. In addition, the lower bound of the 95% confidence interval of the SYQ->USE path is exactly 0, so the path is likely insignificant. In fact, rerunning the

TABLE 13. Summary of R2 values.

	R^2	R^2 Adjusted
Net Benefits	0.314	0.303
Usage	0.446	0.427
User Satisfaction	0.458	0.444

bootstrapping could lead to a p-value higher than 0.05 as it is already 0.048.

In conclusion, the results call for the removal of the path between INQ->USE or significant changes regarding the information quality construct. Such significant changes are, however, reserved for the final model development, which is outside of the scope of this paper. Because this is an exploratory study, the results are satisfying as the majority of the path coefficients exhibited significant results.

c: COEFFICIENT OF DETERMINATION – R^2

In short, the coefficient of determination is a measure to establish the variance explained for every endogenous construct. Therefore, it is a measure of the explanatory power of the model [10], [44]. R^2 values are between 0 and 1, where higher numbers indicate higher explanatory power. In general, values of 0.75, 0.50 and 0.25 for endogenous constructs fall under the categories of substantial, moderate or weak respectively [49]. The threshold values are extremely dependent on the study context or domain. For example, some domains consider 0.10 as satisfactory, such as stock returns. A value of 0.20 is considered high for consumer behaviour studies. For success driver research, examining concepts like customer satisfaction or loyalty, values of 0.75 or higher are expected. Furthermore, R^2 values are affected by the number of predictor constructs, i.e. the R^2 is higher when the amount of predictor constructs is higher [10], [12].

The adjusted coefficient of determination is given for the sake of completeness. It is generally used for complex models to avoid biased results. Since the model in this study does not contain many exogenous constructs, i.e. is parsimonious [12], evaluating the R^2 is considered reasonable.

Table 13 outlines the R^2 for the endogenous latent variables. While this study examines success drivers, the observations were not based on the answers of customers after purchasing a specific product. Further, the study does not only examine consumer behaviour. Thus, following the general threshold recommendation, the values of R^2 lie in the weak and moderate regions.

The R^2 values of usage and user satisfaction are close to 0.45, thus the explanatory power can be considered moderate. The net benefits coefficient of determination is slightly above 0.3 indicating a weak predictive relevance.

d: f^2 EFFECT SIZE

The f^2 effect size measures in what way the omission of a predictor construct changes the R^2 of an endogenous construct. In general, effect sizes of 0.02, 0.15 and 0.35 designate

TABLE 14. Summary of Q2 values (D = 9).

	Q²
Information Quality	-
Net Benefits	0.133
System Quality	-
Trust	-
Usage	0.209
User Satisfaction	0.258

TABLE 15. Summary of q2 effect size (D = 9).

	Inform. Quality	Net Benefits	System Quality	Trust	Usage	User Satisf.
Inform. Quality					-0.003	0.003
Net Benefits						
System Quality					0.009	0.071
Trust					0.009	0.040
Usage		0.069				
User Satisf.		0.005			0.030	

small, medium and large effects respectively. Values smaller than 0.02 suggest the lack of an effect. f^2 effect sizes are sometimes viewed as dispensable as they provide the same information as the sizes of the path coefficients. In other words, the rank order of the f^2 effect sizes is oftentimes identical to the rank order of path coefficient sizes. Thus, these values should be only reported when differences in the rank order occur [12], [10]. After examining the f^2 effect sizes, it was concluded that their rank order corresponds to the path coefficient rank order. Consequently, no in-depth analysis was made.

e: PREDICTIVE RELEVANCE – Q²

The Q^2 value measures the in-sample explanatory power and out-of-sample predictive power of the model. High predictive power implies that the model can predict data that is not used in the model estimation. Q^2 values larger than zero for an endogenous reflective construct suggest predictive relevance for the underlying structural model of the respective construct. Q^2 values of 0, 0.25 and 0.5 correspond to small, medium and large predictive power [10], [12].

Blindfolding is used to determine the Q^2 values. The omission distance D indicating the omission of every Dth data point of the indicators serves as an input. These values are then treated as missing values by the algorithm, for example, resulting in mean value replacement. The recommendations concerning D values vary depending on the source [12]. A D value between 5 and 10 was recommended by Hair *et al.* [49]. Further, dividing the number of samples by the selected D value must not result in an integer [49].

Table 14 summarizes the Q^2 values for an omission distance of 9. Referring to the rule of thumb above, the net benefits construct has a small predictive relevance. User satisfaction has moderate predictive power and the usage construct is close to exhibiting a moderate predictive power. Therefore, data not used for the model estimation can be approximated with moderate predictive power.

f: q² EFFECT SIZE

The q^2 effect size for examining Q^2 works like the f^2 effect size for the R^2. The relative effect on the predictive power of an exogenous latent variable on an endogenous construct can be computed and compared using the q^2 metric. The rule of thumb assumes 0.02, 0.15, 0.35 for small, medium and large effect sizes. Since SmartPLS 3 does not support

this measure, the q^2 effect size (Table 15) was calculated manually by (1). [12]

$$q^2 = \frac{Q^2_{included} - Q^2_{excluded}}{1 - Q^2_{included}} \tag{1}$$

It can be concluded that effect sizes are only detectable for some of the relationships. These include Usage -> Net Benefits, User Satisfaction -> Usage, System Quality -> User Satisfaction and Trust -> User Satisfaction. Note that the endogenous latent variables trust and system quality affect the predictive power of user satisfaction, even though the effects are small. In turn, user satisfaction affects the predictive power of usage and usage affects the predictive relevance of net benefits. Thus, a chain of small q^2 effect sizes from the trust or system quality to net benefits could be identified for the model.

g: PREDICTIVE POWER

The R^2 only gives information about the in-sample predictive power but not out-of-sample predictive power. The PLSpredict algorithm addresses this issue and is used to assess the out-of-sample predictive power. It uses different statistics including the mean absolute error (MAE) and the root-mean-squared error (RMSE). [10]

In terms of evaluation, the primary endogenous construct should be the focus. In this study, it is the net benefits construct. For the first step, the $Q^2_{predict}$ statistic should be evaluated to ensure that the values are better than the naïve benchmark. A positive $Q^2_{predict}$ means that the path model's prediction error is less than the one of the naïve benchmarks [45].

Then, the RMSE or the MAE need to be reviewed, where RMSE gives a larger weight to higher errors. According to Shmueli *et al.* [45], both metrics can be used to choose the models based on a good balance of model fit and predictive power. The RMSE should be used for evaluation of the out-of-sample predictive power, however, the MAE may be used in case the distribution of the prediction error turns out to be highly non-symmetric. The evaluation consists of a comparison of the values with the output of a naïve benchmark, which is the outcome of a linear regression model (LM).

For the evaluation, the guidelines of Shmueli *et al.* were applied [45]. Ten folds (k = 10) were used and ten repetitions

TABLE 16. Summary of PLSpredict.

	$Q^2_{predict}$	RMSE (PLS)	RMSE(LM)	ΔRMSE	PLS<LM
NEB_4	0.092	0.823	0.889	0.066	Yes
NEB_6	0.058	0.758	0.811	0.053	Yes
NEB_3	0.058	0.844	0.902	0.058	Yes
NEB_8	0.001	0.951	0.974	0.023	Yes
NEB_7	0.061	0.836	0.869	0.033	Yes
NEB_2	0.005	0.875	0.881	0.006	Yes
NEB_5	0.185	0.762	0.76	-0.002	No

(r = 10) were set for the PLSpredict run. The results are presented in Table 16.

The focus lies on evaluating the data from the net benefits construct. All $Q^2_{predict}$ values are larger than 0, meaning that predictive relevance could be established for every indicator in the net benefits construct. The value of NEB_8 is 0.001 is close to zero, which reason could be the high standard deviation of this item.

The next step encompasses a review of the RMSE values, i.e. a comparison between the PLS and LM values. All indicators must have a lower RMSE (PLS) than the benchmark case (LM) for claiming that the model has a high predictive power [45]. The RMSE is lower for the PLS values for every indicator, except NEB_5. Hence, the predictive relevance is very close to being considered high according to the PLSpredict simulation results.

V. CONCLUSION

The study develops a model that explains and predict the success of recommender systems. The proposed model is based on the DeLone and McLean information systems success model and includes trust as an additional latent variable.

Through an extensive literature review, more than 180 factors that could affect the success of recommender systems were identified. The factors were carefully analysed by their aligning with commonly used measures in the D-M model. Expert evaluation leads to a total of 30 indicators and six constructs forming the research model.

The proposed model was critically evaluated in the survey, which was sent to more than 6200 researchers in the recommender system domain. 133 responses were received, of which 120 were deemed valid.

Relations between constructs were hypothesised and the proposed success model was then evaluated with PLS-SEM. The results of path analyses are satisfying as the majority of the path coefficients exhibited significant results.

The R^2 values obtained for usage and user satisfaction constructs are close to moderate predictive relevance. The R^2 of net benefits indicates a weak predictive power.

Concerning the Q^2 values, net benefits have small and usage is close to having a moderate predictive relevance. On a more positive note, the user satisfaction construct has moderate predictive power. Studying the q^2 effect size revealed a chain of small effects through the model, starting from trust to user satisfaction, then from user satisfaction via usage to the

net benefits construct. Q^2 effect size from user satisfaction to usage suggests that there is indeed a relationship. Thus, the initial assumption that a path leads from user satisfaction to usage is confirmed.

The predictive power of the model is very close to being considered high according to the PLSpredict simulation results.

Future research will improve the proposed model of recommender system success and analyse results by decomposition of the general model into provider and consumer perspectives.

REFERENCES

[1] F. Ricci, L. Rokach, and B. Shapira, *Recommender Systems Handbook*, 2nd ed. New York, NY, USA: Springer, 2015.

[2] J. A. Konstan and J. Riedl, "Recommender systems: From algorithms to user experience," *User Model. User-Adapted Interact.*, vol. 22, nos. 1–2, pp. 101–123, Apr. 2012.

[3] X. Amatriain, "Mining large streams of user data for personalized recommendations," *in: ACM SIGKDD Explor. Newslett.*, vol. 14, no. 2, pp. 37–48, 2013.

[4] L. Bernardi, T. Mavridis, and P. Estevez, "150 successful machine learning models: 6 lessons learned at booking. com," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Jul. 2019, pp. 1743–1751.

[5] M. Haldar, M. Abdool, P. Ramanathan, T. Xu, S. Yang, H. Duan, Q. Zhang, N. Barrow-Williams, B. C. Turnbull, B. M. Collins, and T. Legrand, "Applying deep learning to airbnb search," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Jul. 2019, pp. 1927–1935.

[6] W. H. DeLone and E. R. McLean, "Information systems success: The quest for the dependent variable," *Inf. Syst. Res.*, vol. 3, no. 1, pp. 60–95, 1992, doi: 10.1287/isre.3.1.60.

[7] W. H. DeLone and E. R. McLean, "Information systems success revisited," in *Proc. 35th Annu. Hawaii Int. Conf. Syst. Sci.*, Big Island, HI, USA, 2002, pp. 238–249.

[8] A. Jeyaraj, "DeLone & McLean models of information system success: Critical meta-review and research directions," *Int. J. Inf. Manage.*, vol. 54, pp. 1–15, Oct. 2020.

[9] N. Urbach and B. Müller, "The updated DeLone and McLean model of information systems success," in *Information Systems Theory*, Y. K. Dwivedi, M. R. Wade, S. L. Schneberger, Eds. New York, NY, USA: Springer, 2012, pp. 1–18.

[10] J. F. Hair, J. J. Risher, M. Sarstedt, and C. M. Ringle, "When to use and how to report the results of PLS-SEM," *Eur. Bus. Rev.*, vol. 31, no. 1, pp. 2–24, Jan. 2019.

[11] G. Mateos-Aparicio, "Partial least squares (PLS) methods: Origins, evolution, and application to social sciences," *Commun. Stat. Theory Methods*, vol. 40, no. 13, pp. 2305–2317, Apr. 2011.

[12] J. F. Hair, G. T. M. Hult, C. M. Ringle, and M. Sarstedt, *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 2nd ed. Los Angeles, CA, USA: Sage, 2017.

[13] E. E. Rigdon, "Rethinking partial least squares path modeling: In praise of simple methods," *Long Range Planning*, vol. 45, nos. 5–6, pp. 341–358, Oct. 2012.

[14] M. Sarstedt, C. M. Ringle, and J. F. Hair, "Partial least squares structural equation modeling," in *Handbook of Market Research*, C. Homburg, M. Klarmann, and A. Vomberg, Eds. Cham, Switzerland: Springer, 2017, pp. 1–40.

[15] W. W. Chin, "How to write up and report PLS analyses," in *Handbook of Partial Least Squares: Concepts, Methods and Applications*, V. Esposito, C. Vincenzo, W. Wynne, J. Henseler, and H. Wang, Eds. Berlin, Germany: Springer, 2010, pp. 655–690.

[16] V. Mezhuyev, M. Al-Emran, M. Arfian Ismail, L. Benedicenti, and D. A. P. Chandran, "The acceptance of search-based software engineering techniques: An empirical evaluation using the technology acceptance model," *IEEE Access*, vol. 7, pp. 101073–101085, 2019, doi: 10.1109/ACCESS.2019.2917913.

[17] J. M. Becker, A. Rai, and E. Rigdon, "Predictive validity and formative measurement in structural equation modeling: Embracing practical relevance," in *Proc. Int. Conf. Inf. Syst. (ICIS)*, vol. 4, 2013, pp. 3088–3106.

[18] J. Henseler, T. K. Dijkstra, M. Sarstedt, C. M. Ringle, A. Diamantopoulos, D. W. Straub, D. J. Ketchen, J. F. Hair, G. T. M. Hult, and R. J. Calantone, "Common beliefs and reality about PLS: Comments on Rönkkö and Evermann (2013)," *Organizational Res. Methods*, vol. 17, no. 2, pp. 182–209, Apr. 2014.

[19] L. T. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," *Struct. Equ. Model.*, vol. 6, no. 1, pp. 1–55, 1999.

[20] A. Ramadhanti and A. S. Slamet, "Developing E-commerce success model by measuring website quality of Indonesian MSMEs," in *Proc. 5th Sriwijaya Econ., Accounting, Bus. Conf. (SEABC)*, Palembang, Indonesia, 2020, pp. 16–23.

[21] W. H. DeLone and E. R. McLean, "Measuring E-commerce success: Applying the DeLone & McLean information systems success model," *Int. J. Electron. Commerce*, vol. 9, no. 1, pp. 31–47, Oct. 2004.

[22] N. I. Ali, S. Samsuri, M. S. A. Seman, I. A. Brohi, and A. Shah, "Measuring E-commerce success in Malaysia: Modified Delone Mclean model with trust and privacy," *Int. J. Eng. Technol.*, vol. 7, no. 4, pp. 524–529, 2018.

[23] M. Nilashi, D. Jannach, O. B. Ibrahim, M. D. Esfahani, and H. Ahmadi, "Recommendation quality, transparency, and website quality for trust-building in recommendation agents," *Electron. Commerce Res. Appl.*, vol. 19, pp. 70–84, Sep. 2016.

[24] D. Jannach and G. Adomavicius, "Recommendations with a purpose," in *Proc. 10th ACM Conf. Recommender Syst.*, Boston, MA, USA, Sep. 2016, pp. 7–10.

[25] D. Jannach and M. Jugovac, "Measuring the business value of recommender systems," *ACM Trans. Manage. Inf. Syst.*, vol. 10, no. 4, pp. 1–23, Dec. 2019.

[26] M. Elahi, F. Ricci, and N. Rubens, "A survey of active learning in collaborative filtering recommender systems," *Comput. Sci. Rev.*, vol. 20, pp. 29–50, May 2016.

[27] S. Tiwari, A. Jain, P. Kothari, R. Upadhyay, and K. Singh, "Learning user preferences for recommender system using Youtube videos tags," in *Proc. 18th Int. Conf. Comput. Sci. Appl.*, Melbourne, VIC, Australia, 2018, pp. 464–473.

[28] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM Comput. Surv.*, vol. 52, no. 1, pp. 1–38, Feb. 2019.

[29] S. K. Lam, D. Frankowski, and J. Riedl, "Do you trust your recommendations? An exploration of security and privacy issues in recommender systems," in *Proc. Int. Conf. Emerg. Trends Inf. Commun. Secur. (ETRICS)*, Freiburg, Germany, 2006, pp. 14–29.

[30] B. Mobasher, R. Burke, R. Bhaumik, and C. Williams, "Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness," *ACM Trans. Internet Technol.*, vol. 7, no. 4, pp. 1–38, 2007.

[31] A. Friedman, B. P. Knijnenburg, K. Vanhecke, L. Martens, and S. Berkovsky, "Privacy aspects of recommender systems, in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. Boston, MA, USA: Springer, 2015, pp. 649–688.

[32] J. Golbeck, "Generating predictive movie recommendations from trust in social networks," *Proc. 4th Int. Conf. Trust Manage. (iTrust)*, Berlin, Germany, 2006, pp. 93–104.

[33] H. Cramer, "The effects of transparency on trust in and acceptance of a content-based art recommender," *User Model. User-Adaption Interact.*, vol. 18, no. 5, pp. 455–496, 2008.

[34] J. A. Calandrino, A. Kilzer, A. Narayanan, E. W. Felten, and V. Shmatikov, "You might also like:' privacy risks of collaborative filtering," in *Proc. IEEE Symp. Secur. Privacy*, Oakland, CA, USA, May 2011, pp. 231–246.

[35] H. Sharma and A. G. Aggarwal, "Finding determinants of e-commerce success: A PLS-SEM approach," *J. Adv. Manage. Res.*, vol. 16, no. 4, pp. 453–471, Oct. 2019.

[36] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quart.*, vol. 13, no. 3, pp. 319–340, 1989.

[37] S. Petter, W. DeLone, and E. R. McLean, "Information systems success: The quest for the independent variables," *J. Manage. Inf. Syst.*, vol. 29, no. 4, pp. 7–62, Apr. 2013.

[38] I. Avazpour, T. Pitakrat, L. Grunske, and J. Grundy, "Dimensions and metrics for evaluating recommendation systems," in *Recommendation Systems in Software Engineering*, M. P. Robillard, W. Maalej, and R. J. Walker, T. Zimmermann, Eds. Berlin, Germany: Springer, 2014, pp. 245–273.

[39] A. Gunawardana and G Shani, "Evaluating recommender systems," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. Boston, MA, USA: Springer, 2015, pp. 265–308.

[40] S. Tehseen, Z. H. Qureshi, F. Johara, and T. Ramayah, "Assessing perceived business success as a reflective-formative (type II) second-order construct using PLS-SEM approach," *J. Sustainability Sci. Manage.*, vol. 14, no. 5, pp. 84–114, 2019.

[41] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: A survey," *Decision Support Syst.*, vol. 74, pp. 12–32, Jun. 2015.

[42] K.-H. Yang, Y. Song, C.-Y. Wu, P.-T. Yang, and C.-Y. Wang, "An exploratory study of the promotion effectiveness of recommender systems," in *Proc. the 3rd Int. Conf. Comput. Manage. Bus.*, Jan. 2020, pp. 135–140.

[43] G. Adomavicius, J. C. Bockstedt, S. P. Curley, and J. Zhang, "Effects of online recommendations on consumers' willingness to pay," *Inf. Syst. Res.*, vol. 29, no. 1, pp. 84–102, 2018.

[44] G. Shmueli and O. R. Koppius, "Predictive analytics in information systems research," *MIS Quart.*, vol. 35, no. 3, pp. 553–572, Sep. 2011.

[45] G. Shmueli, M. Sarstedt, J. F. Hair, J.-H. Cheah, H. Ting, S. Vaithilingam, and C. M. Ringle, "Predictive model assessment in PLS-SEM: Guidelines for using PLSpredict," *Eur. J. Marketing*, vol. 53, no. 11, pp. 2322–2347, Nov. 2019.

[46] M. Al-Emran, V. Mezhuyev, and A. Kamaludin, "PLS-SEM in information systems research: A comprehensive methodological reference," in *Proc. Int. Conf. Adv. Intell. Syst. Inform.*, Cham, Switzerland, 2019, pp. 644–653.

[47] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," *J. Acad. Marketing Sci.*, vol. 43, no. 1, pp. 115–135, 2015.

[48] J. F. Hair, M. Sarstedt, and C. M. Ringle, "Rethinking some of the rethinking of partial least squares," *Eur. J. Marketing*, vol. 53, no. 4, pp. 566–584, Apr. 2019.

[49] J. F. Hair, M. Sarstedt, C. M. Ringle, and J. A. Mena, "An assessment of the use of partial least squares structural equation modeling in marketing research," *J. Acad. Marketing Sci.*, vol. 40, no. 3, pp. 414–433, 2012.

[50] S. Hussain, Z. Fangwei, A. F. Siddiqi, Z. Ali, and M. S. Shabbir, "Structural equation model for evaluating factors affecting quality of social infrastructure projects," *Sustainability*, vol. 10, no. 5, pp. 1415–1439, 2018.

MAXIMILIAN GOTTHARDT received the master's degree in geothermal engineering from the University of Leoben, Austria, and the master's degree in international industrial management from FH Joanneum, Kapfenberg, Austria, where he focused on the research field of recommender systems. He is currently a Drilling Engineer with geothermal consulting company Erdwerk. His current research interest includes modeling recommender systems using PLS SEM.

VITALIY MEZHUYEV (Senior Member, IEEE) received the Ph.D. degree in educational technology from Kyiv National Pedagogical University, Ukraine, in 2002, and the Sc.D. degree in information technology from Odesa National Technical University, Ukraine, in 2012. He was a Professor of informatics with Berdyansk State Pedagogical University, Ukraine, from 2004 to 2014. He was a Professor of informatics at University Malaysia Pahang, from 2014 to 2019. Since 2019, he has been with the Institute of Industrial Management, FH Joanneum, Kapfenberg, Austria. He was a Visiting Professor at six international universities. He has participated in multiple international scientific and industrial projects, devoted to the design, development, and formal verification of computer systems. He has published over 130 scientific articles in peer-review journals, including highly reputed venues, such as IEEE Access, *Complexity*, *Computers and Education*, *Cybernetics and Systems*, *Information Management*, and *Technology in Society*.

• • •