Will the customers be happy? Identifying unsatisfied customers from service encounter data

Lucas Baier, Niklas Kühl, Ronny Schüritz and Gerhard Satzger
Karlsruhe Service Research Institute, Karlsruhe Institute of Technology, Karlsruhe, Germany

Received 4 June 2019
Revised 8 November 2019
12 March 2020
1 May 2020
25 May 2020
Accepted 29 May 2020

Abstract

Purpose – While the understanding of customer satisfaction is a key success factor for service enterprises, existing elicitation approaches suffer from several drawbacks such as high manual effort or delayed availability. However, the rise of analytical methods allows for the automatic and instant analysis of encounter data captured during service delivery in order to identify unsatisfied customers.

Design/methodology/approach – Based on encounter data of 1,584 IT incidents in a real-world service use case, supervised machine learning models to predict unsatisfied customers are trained and evaluated.

Findings – We show that the identification of unsatisfied customers from encounter data is well feasible: via a logistic regression approach, we predict dissatisfied customers already with decent accuracy—a substantial improvement to the current situation of "flying blind". In addition, we are able to quantify the impacts of key service elements on customer satisfaction.

Research limitations/implications – The possibility to understand the relationship between encounter data and customer satisfaction will offer ample opportunities to evaluate and expand existing service management theories.

Practical implications – Identifying dissatisfied customers from encounter data adds a valuable methodology to customer service management. Detecting unsatisfied customers already during the service encounter enables service providers to immediately address service failures, start recovery actions early and, thus, reduce customer attrition. In addition, providers will gain a deeper understanding of the relevant drivers of customer satisfaction informing future new service development.

Originality/value – This article proposes an extendable data-based approach to predict customer satisfaction in an automated, timely and cost-effective way. With increasing data availability, such AI-based approaches will spread quickly and unlock potential to gain important insights for service management.

Keywords Customer satisfaction, Service encounter, IT service management, Machine learning, AI

Paper type Research paper

1. Introduction

Customer satisfaction is a crucial customer-centric metric for any business (Farris *et al.*, 2010) and the "ultimate goal of service providers" (Sureshchandar *et al.*, 2002). In turn, high customer satisfaction leads to higher customer loyalty as satisfied customers choose to reuse the service and to not move to competition (Oliver, 2014). In the long run, loyal customers translate into higher customer lifetime value, ultimately leading to revenue growth (Loveman, 1998). Since the 1980s, both academia and industry have investigated customer satisfaction to understand its drivers and its relationship to service quality (Sureshchandar *et al.*, 2002). Once dissatisfaction is successfully identified, service recovery can be initiated. For successful service recovery, it has been shown that customers increase their trust and commitment to the provider (Tax *et al.*, 1998). Additionally, customers are more likely to share their positive experiences with others (Maxham and Netemeyer, 2002). However, existing approaches to measure customer satisfaction still suffer from several drawbacks: As metrics based on *anonymous* questionnaires only reveal the satisfaction of the (small) sample group that is willing to participate, they typically involve a selection bias, and they do not allow for the identification of individual customers as recovery targets. Soliciting feedback from *individual, identifiable* customers, though, is mostly time-delayed and requires considerable

Journal of Service Management

1757-5818
DOI 10.1108/JOSM-06-2019-0173

resources and financial investment. Thus, today a lack of automation and scalability still impede to identify unsatisfied customers on a larger scale—as a prerequisite to systematically make service recovery decisions across the entire customer base.

However, the growing capabilities to exploit data and to generate insights from them create manifold novel opportunities for businesses (Chen and Zhang, 2014; Schüritz *et al.*, 2017) and may also help to solve the problem mentioned above: Typically, when customer satisfaction is shaped during service delivery, individual characteristics of the service encounter (e.g. waiting time) can be captured by the service provider and recorded as *encounter data* (Fromm *et al.*, 2012). If we are collecting such data during service encounters, should we then not be able to leverage it to systematically and automatically predict and identify unsatisfied customers, and thus allow for immediate service recovery actions?

We address this question by applying supervised machine learning techniques in an exemplary use case. We have access to 1,584 instances of an IT incident process of a major IT company as the basis for our analysis and are able to show that it is in fact possible to identify unsatisfied customers with a certain accuracy. Thus, we contribute to the body of knowledge by suggesting a new and extendable approach to predict and monitor customer satisfaction and by proving its feasibility. The application of machine learning models has several advantages compared to traditional approaches to elicit customer satisfaction: First, the cost of predictions by an automated machine learning mode may only be a *fraction of the cost* of traditional surveys. Second, a machine learning model can predict the satisfaction state for *each customer at an individual level* and is not restricted to a small sample of the entire customer base. Third, the proposed approach allows for *real-time identification* of dissatisfied customers enabling to launch recovery actions immediately. Finally, the approach also reveals *underlying influence factors* of customer dissatisfaction in detail.

The remainder of the article is structured as follows: In Section 2, we introduce the related work we drew upon for our research. In Section 3, we provide the context of the real-world case that we used to implement and evaluate our approach. In Section 4, we describe our methodology concerning the supervised machine learning approach before we present its results in Section 5. In Section 6, we summarize the results, acknowledge limitations, discuss potential implications for organizations (including data privacy and security concerns) and outline future research.

2. Foundations and related work

Before we suggest and apply machine learning to predict customer satisfaction, we will first review the related literature in customer satisfaction and service quality as well as the utilization of process data in that field.

2.1 Customer satisfaction and service quality

Customer satisfaction represents the customer fulfillment response (Oliver, 2014). It is a key benefit for businesses to gain insights into their customers' attitudes as well as their satisfaction levels and judgment toward the product or service being used (Fornell, 1992): Customer satisfaction directly impacts a customer's loyalty to a firm (Khan and Fasih, 2014), the likelihood of reusing a service and the inclination to recommend via word of mouth (Anderson *et al.*, 1994; Gremler and Brown, 1996). Harter *et al.* (2002) show the (positive) impact of customer satisfaction on share-of-wallet, i.e. the amount of money a provider can make with an existing customer. One may conclude that customer satisfaction is closely linked to a firm's long-term revenue growth and profitability (Loveman, 1998), and corresponding insights are indispensable for service firms (Woodside *et al.*, 1989). However, recent research has been critical about the relationship of customer satisfaction and financial

performance and has revealed that the relationship is much more complex than previous research has assumed (Keiningham *et al.*, 2014). In some cases, even negative ROIs of an increase of customer satisfaction have been identified. This illustrates that there is an ongoing discussion about the impact of customer satisfaction on business metrics, and as yet there is no academic consensus about this relationship. However, in light of the large number of articles that have observed a positive relationship, we assume the same for the remainder of this article.

It is generally accepted that customer satisfaction is the result of a subjective comparison between expected and perceived performance of a product or a service (Parasuraman *et al.*, 1985; Oh, 1999). In services, expectations are a customer's consideration of what may happen in the service encounter. If providers fail to meet these expectations, the customer is unsatisfied (Oliver, 2014).

Customer satisfaction of a service is influenced by various factors, such as service quality, price, situational and personal factors (Rust and Zahorik, 1993). In pure services (such as education and financial services), service quality is often the most crucial factor; it represents a customer's perception of reliability, assurance, responsiveness, empathy and tangibles (Parasuraman *et al.*, 1988). Other schools of thought suggest that perceived service quality is formed by technical and functional quality (Grönroos, 1984) or the service product, delivery and environment (Rust and Oliver, 1994). Brady and Cronin (2001) state that service quality is influenced by the interaction quality (e.g. attitude, behavior, expertise), the physical environment quality (e.g. ambient conditions, design, social factors) and the outcome quality (e.g. waiting time, tangibles, valence). It is generally accepted that a major portion of the customer perception of a service is formed in the *moment of truth*—the service encounter. During an encounter, the customer interacts with the service provider and experiences the service quality (Shostack, 1985).

Since the early 1970s, much research has been devoted to the measurement of customer satisfaction. Three major approaches have emerged: First, deriving customer satisfaction indirectly by measuring the perceived service quality; second, measuring customer satisfaction via the expectancy disconfirmation model; third, measuring it by directly asking customers about their satisfaction via a satisfaction survey (Churchill and Surprenant, 1982). A common method for the first, *indirect* approach is SERVQUAL (Parasuraman *et al.*, 1988), which measures the perceived service quality after using a service, which can be defined as a customer's judgment about a product or service's overall excellence or superiority (Anderson *et al.*, 1994). Customers are typically asked "Is the service delivered to you what you expected?" (Woodside *et al.*, 1989). Because the literature agrees on the correlation between a customer's perceived service quality and his satisfaction, SERVQUAL allows to indirectly measures customer satisfaction (Woodside *et al.*, 1989). The second approach, the *expectancy disconfirmation* model of satisfaction, directly measures customer satisfaction based on the disconfirmation paradigm (Oliver, 1980). It states that satisfaction relates to the size and direction of the disconfirmation experience, measured as the difference between the customer's initial expectations and the de facto service performance (Oliver, 2014). The third and most straightforward method to measure customer satisfaction is to *directly ask* customers using a few items or a single item that can be answered on an ordinal (Likert) scale, as often used in commercial studies (e.g. from 1 = unsatisfied to 5 = satisfied) (Hayes, 1998; Mittal and Kamakura, 2001). This may happen via various channels, for instance, through e-mail surveys, as part of popups in mobile phone applications, or with physical buttons at the exit of a service facility.

All these methods rely on customers actively stating their opinions. Thus, this requires an interaction with a customer via a medium (survey, hardware, popup) and a potentially unwanted response action by the customer. These methods are also subject to the widely known issue of response bias, as customers tend to be either too benign or overly critical

compared to their true opinions (Arndt and Crane, 1975). Further, most of these methods only yield a time-delayed picture of customer satisfaction. Also, they only serve as an after-the-fact measurement tool and cannot be used to identify unsatisfied customers during or shortly after a service delivery in order to initiate an instant service recovery process.

2.2 Utilizing digital process data from service encounters

Enterprises are constantly identifying data-driven ways to improve decision-making, optimize processes, drive efficiency, understand customer segments and conceive and develop new products and services (Kart *et al.*, 2013). Unsurprisingly, businesses that engage in big data and analytics outperform competitors and can expect high ROIs for big data technologies (Mithas *et al.*, 2012; Bughin, 2016).

In the service context, analytics—as an approach to exploit this data—has often been related to the notion of service systems (Böhmann *et al.*, 2014). Fromm *et al.* (2012, p. 139) define service analytics as "the process of capturing, processing and analyzing data taken from a service system—in order to improve, extend and personalize the service and create new value for both the provider and the customer." For this purpose, service analytics may utilize provider data, customer data as well as data created during a service encounter.

Service encounters may produce a wide dataset that can be used for different purposes: Organizations may analyze customer feedback, automatically aggregating the most important complaints and compliments (Ordenes *et al.*, 2014), analyzing the provider–customer interaction data to monitor the customer relationship (Habryn *et al.*, 2010), capturing customers' sentiments and emotions (Mattila and Enz, 2002; Gamon, 2004) or identifying customer needs (Eckstein *et al.*, 2016). For service encounter types, where there is written customer–provider communication (as, e.g. in incident tickets), sentiment analysis can be crucial. Written text is special compared to other communication forms, since facial expressions or vocal feedback are not available to detect the author's hidden intentions. However, in this case, sentiment analysis allows to still analyze a customer's underlying attitudes and emotions (Liu, 2015).

Despite the advances in the fields mentioned above, to our best knowledge no studies have yet sought to predict customer satisfaction based on data harvested in a service system. One can imagine that data from customer, provider or their service encounter(s) can be applied to predict the dimensions that impact customer satisfaction. For instance, metrics from the business processes such as a customer's waiting time or a service employee's reaction time can be used to estimate a service's interaction quality. Further, temperature, noise level or other proxies may serve as indicators to estimate a physical environment's quality. Customer characteristics such as street address, and customer history may indicate personal traits and can easily be retrieved from customer relationship management databases. By combining these attributes, we may be able to assess service quality and—eventually—predict customer satisfaction.

Closely related to the task of measuring and predicting customer satisfaction during the service encounter is the idea of predicting customer attrition. [1] Different authors point out various possibilities to enable attrition prediction, for instance with traditional regression approaches (Trubik and Smith, 2000), descriptive data mining (den Poel and Lariviere, 2004) or more modern neural net techniques (Tsai and Lu, 2009). Even though there is a strong relationship between customer satisfaction and customer attrition (Gustafsson *et al.*, 2005), many other reasons (apart from satisfaction) also exist for customer attrition (Woo and Fock, 2004).

3. Domain: IT service management

To demonstrate that data from a service information system can be exploited to identify unsatisfied customers, we choose IT service management as a domain in which customers

and providers frequently interact. Key service encounters are IT service incidents (Kieninger *et al.*, 2013), which occur whenever a customer is confronted with problems related to the provider's software product. In case of an incident, customers open incident tickets, usually via an online portal. Subsequently, the provider tries to resolve the problem. Interactions between the parties are documented in incident tickets—an initial problem description by the customer as well as the jointly achieved progress in solving the issue. In our view, the chosen case is a very suitable initial test scenario for our research, because (1) it represents a touchpoint in which interaction is controlled and standardized, (2) emotional reactions due to interpersonal challenges are limited and (3) other situational influence factors (e.g. weather) are limited.

We secure access to the incident ticket data for a specific software product of an international IT service provider, which has been collected throughout a full year (2014). We conduct detailed interviews with experts from the company in order to understand the large number of attributes in the dataset. We also use their knowledge to identify and discard irrelevant elements in the dataset. This data is combined with the results of an ex post customer satisfaction survey collected via standardized phone interviews by a professional third party. The survey renders results for 1,584 completed customer questionnaires capturing the perceived customer satisfaction. In combination with the underlying incident ticket data for these cases, this data can be used to train and test (supervised) machine learning models. A trained model will try to best approximate each known customer satisfaction outcome ("target variable") from the characteristics ("features") of the IT incident process as the service encounter. For any subsequent IT incident, the (unknown) customer satisfaction level can then be predicted from the (known) IT incident documentation.

Each of the 1,584 instances of our core dataset consists of 11 attributes relating to the incident itself (as shown in Table 1). These include, for instance, the priority and severity of the incident as specified by the customer when opening the incident ticket. The attributes also capture the time between ticket opening and first customer contact (*waiting time*) and the one between ticket opening and closure (*resolution time*). Other attributes refer to the product type as well as the error type discovered during the incident's resolution.

In addition, we can assign the customer satisfaction level that has been obtained from the survey to each instance as an additional attribute. Surveyed customers have ranked their satisfaction along a 5-point Likert scale between 1 (totally satisfied) and 5 (completely

Variable	Type	Description
Creation date	Date	The date and time of the initial creation of the incident ticket
Duration	Integer	The time between the opening and the closing of the incident ticket (in days)
Initial contact delay	Integer	The time between incident ticket creation and first contact between the provider and the customer (in minutes)
Contact misses	Integer	The number of failed attempts to contact the customer
Priority	Integer	The priority of the incident as specified by the customer (on a scale from 1 to 5)
Severity	Integer	The severity of the incident as specified by the customer (on a scale from 1 to 5)
Product	Integer	Dissemination of affected software product
Error type	Categorical	The type of identified error
Solution	Categorical	The type of solution provided (e.g. bug fix)
Description	String	The textual description of the problem by the customer
Satisfaction *(target)*	Binary	Describes whether the customer is unsatisfied (label = TRUE) or satisfied (label = FALSE)

Table 1. Available attributes in the incident ticket dataset

unsatisfied). For simplification reasons, we aggregate customers with a rating of 1 or 2 into a satisfied customer group, while all others are labeled as unsatisfied. This binary classification results in 1,419 (89.6%) satisfied and 165 (10.4%) unsatisfied customers.

4. Methodology

We use the available data to build and calibrate a machine learning model that infers customer satisfaction from the attributes of the incident case. The dataset of 1,584 instances with known customer satisfaction outcomes can be used to train and test different models helping us to classify satisfied vs unsatisfied customers for additional instances where customer satisfaction is not yet known but can be predicted. We follow the supervised machine learning process as outlined by Hirt *et al.* (2017).

First, we conduct the necessary data preprocessing steps such as data cleaning (Section 4.1). Then, we choose an evaluation metric and run pre-tests with different sampling strategies and classification algorithms (Section 4.2). This enables us to assess which type of model is most suitable for our use case (Section 4.3). In addition, we run an analysis to explore the sensitivity of the results to the size of the available set of IT incident tickets (Section 4.4.).

4.1 Data preparation

The data used to predict customer satisfaction comprise both unstructured (e.g. the problem description) and structured data (e.g. the time of incident ticket creation). Since not every incident ticket in the dataset contains information about each attribute, it is necessary to develop a strategy for handling missing values. This is crucial, because machine learning algorithms usually rely on complete data input. We decide to replace missing values in the structured data with a dummy value of −1. Categorical data are transformed such that each category value is turned into a separate binary feature column (one-hot-encoding). Before proceeding to the model training and evaluation, we generate additional features from the core features in Table 1. This includes simple conversions such as specifying the weekday of the incident ticket occurrence and higher-level transformations of existing features concerning the incident descriptions—like adding sentiment features obtained from textual descriptions:

As to the description text itself, we decide to not include the full text itself in the feature space of the prediction problem. After analyzing numerous examples, we observe that the description text has a very technical focus (e.g. "... requires Java 7 SR7 FP1," "... VMware vSphere that is running Exchange, 2013"). Thus, we do not apply a text analysis at the word level, such as a bag of words approach with word count (Cambria and White, 2014) or more sophisticated ones such as word2vec (Mikolov *et al.*, 2013). Due to the limited number of instances (1,584 incident tickets), we also refrain from training our own vector representation of the detected vocabulary.

Nonetheless, we expect helpful information to be contained in the textual descriptions that could be incorporated via a higher-level transformation compared to the methods mentioned above. As we expect emotional comments to be highly indicative of final customer satisfaction, we extract corresponding features from the IT incident descriptions by running a sentiment analysis on the description text. While many existing approaches are based on a lexicon of positive and negative words and sentences (Feldman, 2013), we resort to a more complex analysis applying the NLTK implementation of VADER (Hutto and Gilbert, 2014) as a sentiment analysis tool. VADER not only returns the polarity but also the intensity of sentiments expressed in texts. Thus, we are able to obtain intensity scores for the positive, neutral and negative sentiment in the text as well as a compound score as additional feature candidates. To gain a better understanding of the sentiment results, we scrutinize selected tickets with substantial positive or negative sentiment value: Tickets with a positive

sentiment use a more friendly language, expressed through text blocks like "thank you", "positive feedback" or the polite addressing of the employee, like "Dear (. . .)". On the other hand, tickets with a negatively labeled sentiment contain texts like "I've been struggling all day (. . .)" or "I am frustrated beyond belief!" or "no solution was provided". This plausibility check gives us confidence that the inclusion of sentiment values could inform our model, and thus add additionally generated sentiment features.

4.2 Pre-tests

Due to the huge variety of available classification algorithms, we run several pre-tests. The purpose of the pre-tests is to choose the most promising combination of sampling options and algorithms as a base for building the final machine learning model. We apply three different *sampling options* to deal with an unbalanced dataset and popular *classification algorithms* out of five groups (Michie *et al.*, 1994) summarized in Table 2.

The consideration of sampling options is required as our target is rather unbalanced: As stated above, our dataset includes 89.6% satisfied, but only 10.4% unsatisfied customers. This imbalance has to be catered for in the applied classification algorithms, since standard classification algorithms tend to be strongly influenced by the majority class, while the minority class is mostly ignored (Chawla *et al.*, 2004). Thus, we evaluate three sampling methods for the training set to address the class imbalance in the incident ticket dataset. All three sampling strategies pursue the objective to provide an equal number of unsatisfied and satisfied customers in the training set: *Oversampling* adds new instances to the minority class by randomly sampling from this class (Rahman and Davis, 2013). *Undersampling* randomly selects instances from the majority class to match the smaller number of instances in the minority class (Rahman and Davis, 2013). *Synthetic Minority Oversampling Technique (SMOTE)* as a third technique creates new synthetic instances for the minority class based on the distribution of attribute values in this class (Chawla *et al.*, 2002).

The sampling options can be used to prepare the dataset for use with different machine learning algorithms. In order to cover a broad set of algorithm groups, we implement key representatives of each group as shown in Table 2. Each implementation uses the standard parameter configuration in scikit-learn, a free software machine learning library for Python (Pedregosa *et al.*, 2011).

The different options for the pre-test (as combinations of sampling approaches and algorithms) have to be evaluated against a common metric. However, the class imbalance between unsatisfied and satisfied customers also affects the selection of a suitable evaluation metric for the classifier. This step is key, since the final model's performance will be judged by this value—and the hyperparameters of the final prediction model are also optimized with regard to the chosen metric. Standard metrics such as the simple accuracy clearly are inappropriate in our use case (Ling *et al.*, 2003): Accuracy would measure correctly predicted customer satisfaction instances as a share of all predictions. However, even a very simple naïve classifier that always predicts a customer to be satisfied would achieve a high accuracy of 89.6%—i.e. the share of all incident tickets relating to satisfied customers in our survey

Group	Algorithms
Traditional statistical methods	Logistic regression
Bayes classifiers	Gaussian Naïve Bayes
Support vector machines	Support vector machine with sigmoid kernel
Tree-based classifiers	Random forest
Neural networks	Multilayer perceptron (two hidden layers of size 100)

Table 2.
An overview over the applied classification algorithms

dataset. Despite the high accuracy, though, such a classifier would not generate any added value from a business perspective: We would not identify any unsatisfied customer and could not trigger any related service recovery activities.

Instead, we need an evaluation metric that balances precision (measuring correctness of results for each class) against recall (measuring the identification capability for instances of each class). In the above example, the naïve classifier is producing extreme and undesired results: While precision is high for satisfied customers and zero for unsatisfied ones, the recall is 100% for satisfied customers (as all of them are identified) and zero for unsatisfied ones. The F-measure is typically a good choice to address this problem (Kotsiantis *et al.*, 2006), since it considers the tradeoff between precision and recall and allows to include the economic impact of wrong classifications.

In our case, identifying unsatisfied customers is key, since the loss of a customer may be fatal: The costs of winning new customers are at least five times higher than the cost of converting unsatisfied customers into a satisfied state (Hart *et al.*, 1990). This ratio clearly indicates that the classifier must be adapted so that it detects most of the unsatisfied customers (requiring higher recall for this class). This would typically come at the expense of an increase in satisfied customers falsely classified as unsatisfied ones (leading to lower precision). In other words, we prefer to rather pay for a few unnecessary service recovery processes for already satisfied customers rather than missing out on converting an unsatisfied customer. This weighting concerning the outcome of the prediction model is considered by applying the F_2-score, which attaches more importance to recall than precision. In many other cases, too, the F_2-score is a widely used and accepted classification metric (Powers, 2011).

With these considerations at hand, we can now run our pre-tests regarding the performance evaluation for different sampling options (including the default option of not correcting the imbalance) and algorithms. In order to reduce the impact of the allocation of individual IT ticket instances to either train or test set, all pre-tests are carried out with a five-fold stratified cross-validation (Golub *et al.*, 1979). In particular, stratification ensures that the distribution of the target label is approximately equal across all folds. In addition, we scale the input data to have zero mean and unit variance.

Table 3 shows the mean F_2-scores for different combinations of sampling options and classification algorithms. The maximum values per column appear in italic. In general, Logistic Regression shows the best predictive performance, at an F_2-score of 0.343. Gaussian Naïve Bayes also achieves high F_2-scores, especially when applied with oversampling and SMOTE. The influences of the different sampling options vary significantly across the set of classification algorithms: The performance results for Logistic Regression are very similar across all sampling strategies, while the performance of the Multilayer Perceptron fluctuates heavily with the chosen sampling approach.

Looking for a benchmark to assess the quality of the result, we want to compare our results to appropriate benchmarks of alternative, already existing approaches. It may be noted that a comparison to customer satisfaction surveys is not adequate as the latter do not

Table 3.
The F_2-scores for the classification algorithms with different sampling options

Algorithm	Sampling			
	None	Oversampling	Undersampling	SMOTE
Random forest	0.030	0.064	0.293	0.050
SVM (sigmoid kernel)	0.007	0.058	0.032	0.045
Gaussian Naïve Bayes	0.248	0.325	0.230	*0.354*
Multilayer perceptron	0.022	0.294	*0.338*	0.281
Logistic regression	*0.343*	*0.342*	0.335	0.328

predict, but ex post determines the target variable (and, thus, by definition, are accurate). However, to our best knowledge, no study has so far described an automatable and scalable approach for the prediction of dissatisfied customers. Thus, we compare our results to a random guess that randomly assigns the label respecting the probability distribution in the training set (89.6% satisfied vs. 10.4% unsatisfied) to a new observation. Since this baseline classifier achieves an F_2-score of 0.098, our pre-test results (Table 3) clearly indicate that it is possible to build a classifier that significantly outperforms random guess as a "we do not know anything" situation.

4.3 Choosing and building the final model

With the results of the pre-tests, we now have to choose the model setup as a base for building the final model for customer satisfaction prediction. After a thorough analysis of the results in Table 3, we choose Logistic Regression as an algorithm for the subsequent performance estimation—for two reasons: First, it shows a very stable prediction performance independent of the chosen sampling strategy. The differences in F_2-scores between the two best-performing options (no sampling and oversampling) are both very small and within the statistical error range. Second, Logistic Regression has special characteristics, as we will now describe, which makes it a more suitable choice than the other two models with good performance (Gaussian Naïve Bayes and Multilayer Perceptron):

Compared to Gaussian Naïve Bayes, Logistic Regression allows for a hyperparameter optimization which permits to adjust algorithm parameters of a general machine learning algorithm to the specifics of the use case. Usually, this leads to a significant increase in prediction performance. While in Table 3 the Gaussian Naïve Bayes (with SMOTE) reaches its maximal predictive power at the F_2-score of 0.354, Logistic Regression may further benefit from hyperparameter optimization beyond its F_2-score of 0.343.

Compared to the Multilayer Perceptron model, Logistic Regression allows for more interpretability regarding the prediction results (Choi *et al.*, 2016). It is possible to directly analyze the weights of the model to deduct the importance of individual predictive features (Hastie *et al.*, 2009)—an important objective for us in understanding the influence factors on customer satisfaction.

Simultaneously, we have to make a decision on the sampling option. As the various sampling options do not significantly impact the Logistic Regression algorithm's performance, we choose not to apply these in the following model performance estimation. This choice decreases the solution space of the hyperparameter optimization and thus reduces the computational effort by a factor of four. Computational effort can also be an important factor in real-world business settings as service providers may frequently want to retrain their prediction model and monitor the satisfaction of their customers in real time.

After the selection of the machine learning algorithm as well as the suitable sampling strategy, we perform a hyperparameter optimization for the prediction model. The interested reader may find the details in Appendix.

4.4 Sensitivity to amount of training data

Based on the chosen prediction model setup (Logistic Regression before hyperparameter tuning and no sampling), we also analyze the influence of the number of training instances (incident tickets) on the prediction performance. This is an important analysis since it provides an estimate of the added value of additional labeled training data. In our use case, the generation of labeled data instances is expensive since it requires performing a structured customer satisfaction survey via telephone with customers linked to the incident management process. Therefore, the service provider should be able to upfront estimate the benefit of additional customer satisfaction information. Learning curves are an

appropriate tool to examine this relationship, since they depict the prediction scores for the training set as well as the test set based on training sets of different sizes (Perlich *et al.*, 2003). Figure 1 depicts the learning curves for the Logistic Regression for the training score (in red) as well as for the test score (in green). The training score refers to the prediction performance of the model on the utilized training set, whereas the test score relates to the prediction performance of this same model on unseen test data. With approximately 100 training instances that are randomly selected, the algorithm only achieves an F_2-score of 0.25 on the test set, whereas with 1,000 instances, the score improves to 0.34. This clearly demonstrates the positive effect of additional training instances on the overall prediction performance. Thus, we suspect that a larger training set with more incident tickets than the 1,584 tickets in this article would further increase the overall approach's prediction performance. However, the shape of the curves suggests that this increase will be less steep.

5. Results and evaluation

In this section, we present our results from the previously outlined methodology. We are interested in two aspects: First, *predicting* customer satisfaction and second, *understanding* customer satisfaction. Section 5.1 describes in detail the prediction performance of our approach, especially the different types of possible prediction errors. Section 5.2 aims at broadening the understanding of drivers of customer satisfaction. In this context, we analyze the importance of different incident ticket features with two different explainability methods.

5.1 Prediction of customer satisfaction

As discussed in Section 4, we implemented a logistic regression model without specific sampling strategy and optimized its hyperparameters for the particular use case. In this section, we discuss its robustness and predictive power.

Our five cross-validation folds, i.e. different models built on a different split of the data into a training and test set, produce robust values for the F_2-score. The scores are within the interval [0.363, 0.417], with a mean value of 0.379 and a standard deviation of 0.019. The low standard deviation of the scores indicates the model's stability. This means that we can expect the model to perform with similar results when applied to new unseen data; thus, the likelihood of an overfitted model is low. Overfitting describes the phenomenon that a machine

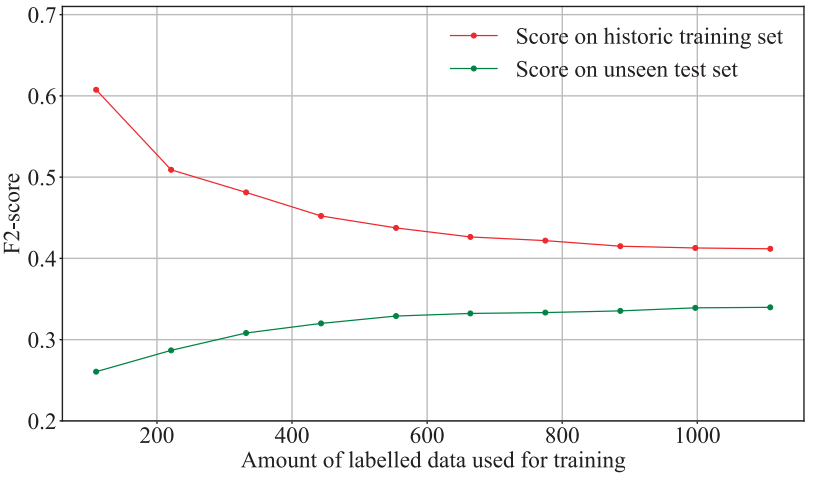

Figure 1.
The learning curve based on the number of training instances

learning model contains more parameters than necessary and thereby tends to memorize the original data and corresponding labels. This leads to high prediction performance on the training set, but poor generalization results. In our case, this would lead to a very high accuracy for the satisfaction prediction of already known customers included in the training set. However, this model would only identify very few new unsatisfied customers not included in the training process.

As described in Section 4, with a lack of available benchmarks, we also compare our results to a random guess, i.e. the situation where the provider does not yet know anything about the customer satisfaction during the service encounter. This baseline classifier, which randomly assigns the label to a new observation according to the probability distribution in the training set, achieves an F_2-score of 0.098. Thus, our model outperforms this baseline by 287%. The comparison to the chosen baseline clearly indicates that our model is able to predict customer satisfaction to a certain extent based on data from the incident management process. So far, this information has to be manually collected by surveys or questionnaires. The automatic retrieval of this information could be very valuable to any company that seeks to prevent customer churn after a service encounter. Despite outperforming a random guess, the prediction results are still far from perfect and may not be good enough to completely replace surveys at an initial stage when such a system is introduced. However, a prediction model can be a valuable extension by providing customer satisfaction information across the entire customer base. Over time, more labeled data can be collected through additional surveys, which can again be used to improve the prediction model (see Section 4.4).

Table 4 introduces the confusion matrix for the machine learning model based on the results after the hyperparameter optimization. This visualization allows to better interpret the implications of the achieved F_2-score. We round the results for each cell so as to improve interpretability. In our use case, approximately 10 out of 100 customers are unsatisfied (89.6% satisfied vs 10.4% dissatisfied customers). Using the F_2-score, we tune the approach to achieve high recall, which means that 9 out of 10 dissatisfied customers are identified correctly. However, at the same time, this high recall leads to the false identification of a large number of satisfied customers—72 in total—which is equivalent to very low precision. A company applying such a predictive model could define its own evaluation metric, calibrating the predictions differently. Nonetheless, it must still consider the tradeoff between precision and recall—a typical conflict in information retrieval tasks (Buckland and Gey, 1994).

To elaborate on our approach's tuning options, we run an additional decile analysis of the prediction results (Rud, 2001). The logistic regression model returns a probability score for each test instance. Since we perform a nested cross-validation, we receive a probability score for every data instance in our dataset (1,584 incident tickets). We order the data instances based on the probability score of the logistic regression function and split the dataset into deciles based on this score, as depicted in Table 5. Decile 1 contains the 10% of instances with the highest probability scores and decile 10 contains the 10% of instances with the lowest ones. For each decile, the number of unsatisfied customers is listed in absolute terms as well as a percentage. The column on the right shows the probability score intervals issued by the logistic regression.

If the prediction model had no predictive power, we would see a uniform distribution of unsatisfied customers across all deciles. Here, however, we can observe a significant decrease

Table 4. The confusion matrix after hyperparameter optimization

		True	
		Unsatisfied	Satisfied
Predicted	Unsatisfied	9	72
	Satisfied	1	18

Decile	No. of customers	No. of unsatisfied customers	% of unsatisfied customers	Probability interval
1	159	38	0.230	[0.521, 1]
2	158	20	0.121	[0.464, 0.521]
3	158	17	0.103	[0.427, 0.464]
4	159	12	0.073	[0.403, 0.427]
5	158	13	0.079	[0.372, 0.403]
6	158	14	0.085	[0.343, 0.372]
7	159	14	0.085	[0.315, 0.343]
8	158	14	0.085	[0.275, 0.315]
9	158	13	0.079	[0.219, 0.275]
10	159	10	0.061	[0, 0.219]
Total	1,584	165	1	–

Table 5.
An overview of predictive results per decile

in the de factor number and share of unsatisfied customers per decile. While the first decile includes 38 of unsatisfied customers, the last decile holds only 10. In contrast, a perfect model would allocate all unsatisfied customers to the highest two deciles. Thus, we can conclude that while the trained model can still be improved, it provides valuable information to a potential user, since the distribution of unsatisfied customers is not uniform.

Table 5 illustrates the target options for the company, for instance, it can target 23% of the unsatisfied customers by addressing all customers in the top decile, which is equivalent to an F_2-score of 0.61. Another option is to target 35% of the unsatisfied customers by considering the top two deciles.

These results indicate that companies could only focus on a smaller fraction of the entire predictions, for instance, considering only the data instances in the highest decile of the probability score issued by the logistic regression function. In this decile, the model's prediction performance is far better and allows to more specifically target unsatisfied customers. This is a *red flag approach*: Instead of trying to identify all the unsatisfied customers, a company can focus on customers where the algorithm is fairly sure that they belong to the target class of unsatisfied customers. This can save a significant amount of money by reducing the amount of unnecessary countermeasures. Based on the model's red flag predictions, a service recovery process including activities such as discounts for special customers can be triggered automatically after the termination of the service encounter.

While our model significantly outperforms the baseline, further improvement of the model performance seems possible, especially when more data are available. The low availability of instances in the minority class (in our case 10.4%) is typically a key issue in supervised machine learning. More incident tickets related to unsatisfied customers would surely improve the prediction performance because the machine learning model would obtain more information on unsatisfied customers and their IT incident context. However, our results clearly demonstrate the feasibility of our chosen machine learning approach.

5.2 Towards understanding drivers of customer satisfaction

While the previous part focused on predicting satisfaction, the following paragraphs aim at broadening the understanding of drivers of customer satisfaction.

Logistic regression allows to easily interpret the importance of different features in the input data. An additional analysis, though, reveals that some features in the dataset are highly correlated which is a problem for logistic regression as it can cause unstable parameter estimates (Midi *et al.*, 2010). As a consequence, we decide to remove the feature *priority* for feature importance analysis since it is highly correlated with *severity*. Furthermore, the four

sentiment scores (see Section 3) are also highly correlated (e.g. *neutral* with *compound score*). Therefore, we decide to keep only the *compound sentiment score* which depicts the customer sentiment on a scale from −1 (negative sentiment) to 1 (positive sentiment).

Since we apply five-fold cross-validation, we estimate not one single prediction model, but five. Thus, we average the weights of these five prediction models for a summarized representation of the coefficients. Figure 2 displays the results. In contrast to linear regression, the weights no longer linearly influence the prediction; nonetheless, the weights still allow us to interpret the importance and direction of a feature on the overall prediction. We model unsatisfied customers as 1 (label = TRUE) and satisfied ones as 0 (label = FALSE). Thus, an increase in an input feature with positive weights leads to a higher likelihood of a customer being unsatisfied.

As we standardized all input features for the logistic regression (see Section 4.1), we can derive the importance of different features directly from the magnitude of the corresponding regression coefficients. High values for the features *duration*, *initial contact delay*, number of *contact misses* and *product* increase the likelihood of a customer being unsatisfied. This corresponds to our expectations. A long overall service time (*duration*) in case of urgent IT incidents surely annoys customers. The same applies to the number of *contact misses* between service provider and customer. This feature is also slightly correlated (0.36) with *duration*. Surprisingly, *initial contact delay* is not correlated with the two previous features, but also has an impact on dissatisfaction. Presumably, customers quickly want to be reassured that the service provider has noticed their incident and has taken first actions to deal with it.

The weights for the remaining features are all negative which means that higher values for those features reduce the likelihood of dissatisfaction. An increase in *sentiment* therefore indicates a satisfied customer as we had expected. Interestingly, an increase in *severity* is likely related to a satisfied customer. This contradicts our expectations and depicts an interesting finding. Future research needs to further investigate this relationship. Potentially, the provider puts more effort in resolving severe incidents—or the customer tends to be more grateful for their resolution.

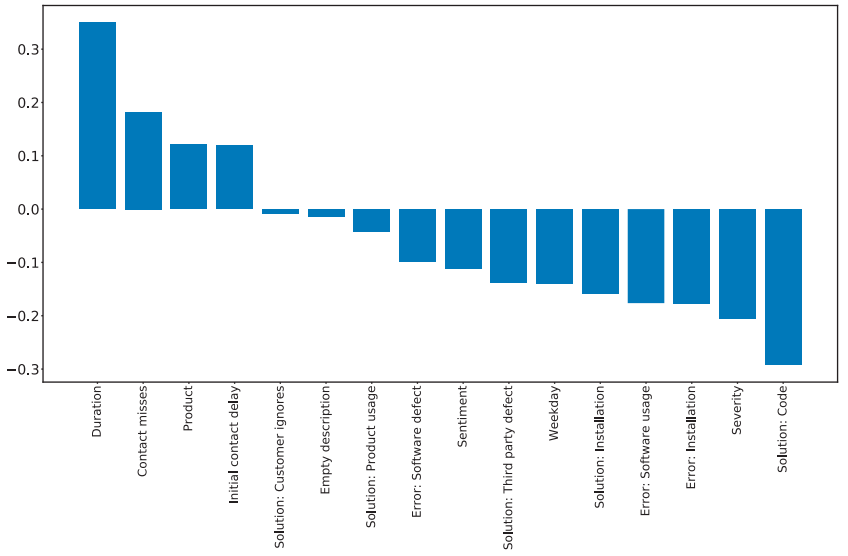

Figure 2.
An overview of the feature weights of logistic regression

The analysis of weights of the logistic regression above is a *global* explanation method as it explains the summarized importance of different features based on the whole training dataset. *Local* explanation methods, in contrast, focus on explaining predictions on an individual data instance level—in our case, the individual IT incident tickets. We use the concept of SHAP values (Lundberg and Lee, 2017) which are based on the game-theoretic approach of Shapley values for this local analysis. Methods based on Shapley values are often used in practice because they are less likely to suffer from problems regarding multicollinearity (Buoye *et al.*, 2014). Therefore, this analysis allows us to examine our previous results for consistency. The results for the SHAP values are depicted in Figure 3.

Features are displayed in decreasing importance. Each dot in the figure displays one data instance in the dataset. A larger vertical spread refers to accumulations of data instances. The color coding in the figure represents the value for that specific feature of a data instance. Again, a higher SHAP value for a feature is associated with a higher chance of identifying a dissatisfied customer. According to this analysis, the *solution type* is the most important feature: If the solution type is a change in the program code of the software product (value = 1, high value), the customer is likely to be more satisfied with the service, i.e. decreasing the probability for dissatisfaction. If the solution type is not a code repair (value = 0, low value), the SHAP value is positive and this increases the likelihood for identifying an unsatisfied customer. In general, the analysis confirms the results based on the logistic regression weights. High values for *duration*, *contact misses* and *initial contact delay* lead to increased SHAP values and are thereby an indicator for dissatisfied customers. These results seem plausible and are in line with previous results. For instance, the number of contact misses and the initial contact delay are proxies for the actual service waiting time which has been shown to be inversely related to customer satisfaction (Davis and Heineke, 1998).

Furthermore, we have also performed an additional analysis where sentiment is predicted based on a set of other features, similar to the hierarchical approach in commercial

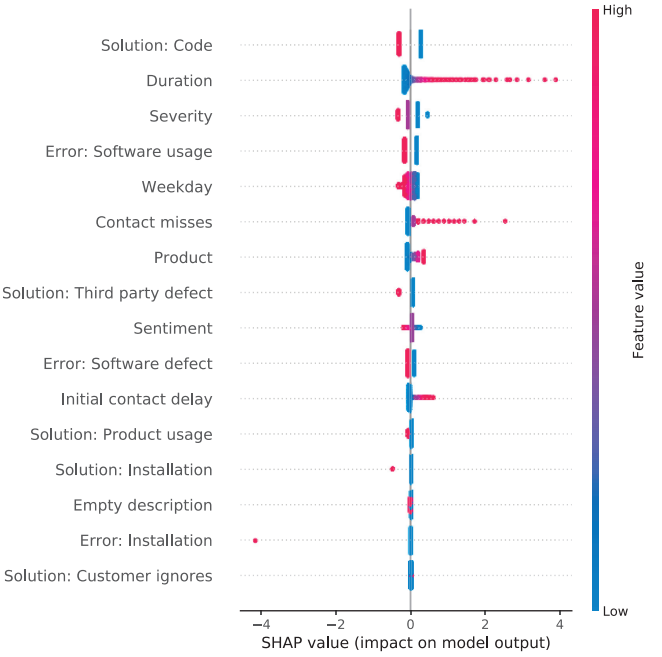

Figure 3.
SHAP values based on logistic regression for all data instances

satisfaction studies where drivers and corresponding subdrivers of satisfaction are derived (Buoye *et al.*, 2014). The interested reader finds this analysis in the Appendix.

6. Conclusion and outlook

We present the results of the application of supervised machine learning to an incident management process with the goal to identify unsatisfied customers. Analyzing a dataset of 1,584 records of IT service incidents, we build a predictive classification model for customer satisfaction. By applying a logistic regression model, we achieve a mean F_2-score of 0.379, which by far outperforms random guesses. Thus, this article demonstrates the potential feasibility to leverage data from service encounters in a service system, such as the incident management process, in order to automatically identify unsatisfied customers. This can be especially helpful if implemented as a red flag system for service recovery processes. An automatic retrieval of predicted customer satisfaction can be very valuable to any company that seeks to monitor its customer satisfaction on an aggregate portfolio level. Furthermore, our analysis shows the varying importance of different features included in the dataset. It seems that time-related features such as the overall incident duration or the time elapsed until the provider reaches out to the customer for the first time are important drivers of dissatisfaction.

6.1 Limitations

This study certainly has its limitations. First, the dataset is limited concerning volume and balance, since only 10.4% of the dataset instances represent unsatisfied customers. Currently, while the predictive model's performance is better than any random guess, it is still far from perfect. Thus, the developed approach with its current prediction performance may not be suitable to completely replace surveys. However, as shown in the decile analysis in Section 5, it can be a valuable support tool as a red flag indicator for a subset of the entire customer base. With more survey data being collected over time, the machine learning model's performance can be increased by considering the newly acquired surveys as additional training data. Second, we are limited to the data provided to us by the service provider and by its processes for collecting customer satisfaction. We assume that the survey results are a viable proxy for measuring the customers' de facto satisfaction. Further, only a subset of all the customers completed the customer survey that we use as the basis for the data analysis. We do not have any knowledge of whether the selected customers are representative of the whole customer base. Future studies may need to collect service quality and customer satisfaction with a designated focus on representativity. Third, the case shows applicability in a controlled and standardized environment. Future research should take the idea into a new service context and evaluate whether a predictive model can be derived that also performs adequately. Fourth, our model can predict customer satisfaction at the end of an incident management process. The dataset has no time simulating aspect and the satisfaction data are surveyed after the incident management process. With appropriate data collection during the incident process, the modeling of a classifier that predicts customer satisfaction in real-time may be possible; this could potentially change the entire service process. Finally, the level of customer satisfaction is always perceived by the individual customer. This obstacle could be overcome by, for instance, classifying different customer segments based on their preference structures.

6.2 Potential for organizations

The research into using machine learning to predict customer satisfaction shows great potential and is an interesting path for service research to rethink established methods. The application of prediction models compared to traditional approaches bears a number of advantages for service organizations:

First, the cost and the manual effort necessary to carry out a traditional survey for customer satisfaction by far exceeds the cost of prediction by an automated machine learning model. The model's current version does not yet deliver the necessary prediction performance, which should be close to perfect if it should completely substitute surveys. However, with the improvement of machine learning techniques and more data, such a scenario becomes more plausible in the foreseeable future. Second, from a provider perspective, the service incident may be regarded as successfully resolved upon termination. However, even after a successful fix of an incident, a customer may still be unsatisfied with the delivered service. Our approach would lay the foundation allowing to easily identify these customers. Third, methods such as customer satisfaction surveys usually only cover a small customer sample. In contrast, a machine learning approach could perform customer satisfaction predictions across the entire customer base, leading to an overall customer satisfaction score and a broad portfolio view. Fourth, service providers typically gather feedback via anonymized questionnaires, thereby only getting an aggregated view of customer satisfaction. The machine learning–based approach makes it possible to collect this information at an individual customer level. This allows to identify and target particular customers after a service process. However, organizations need to ensure that their actions comply with current legislation, e.g. data privacy laws. Fifth, established methods such as surveys will only identify unsatisfied customers with a significant time lag, as there is usually a time gap between the end of an incident and the questioning of a customer. A machine learning model can instead provide instant feedback to the service organization after completion of the incident ticket process. Sixth, supervised machine learning models such as logistic regression allow for a comprehensive analysis of their predictions. This property can be exploited by companies in order to identify the underlying reasons for a failure in service delivery. Finally, if we are able to reach close to perfect prediction models, other applications of the approach become plausible, for instance, awarding bonuses to managers who have effectively increased customer satisfaction. All these benefits should motivate companies to expand their efforts to digitally collect relevant data during service encounters (e.g. via the installation of new sensors).

However, companies that wish to take advantage from the aforementioned benefits also need to consider the necessary prerequisites for implementation. To set up a system that predicts unsatisfied customers, it is necessary to first conduct an extensive survey so as to get the necessary training data for the machine learning model. Further, one must account for the existing information systems and customer service management processes in place when implementing such a solution to allow it to work in real-world settings (Davenport and Ronanki, 2018). This includes typical challenges of the introduction of novel IT artifacts (Seebacher and Schüritz, 2019) regarding technical barriers such as compatibility (Dong *et al.*, 2017), missing standards (Hsu *et al.*, 2015) and/or organizational barriers such as internal resistance (Messerschmidt and Hinz, 2013) or organizational readiness (Chatterjee and Ravichandran, 2013). Furthermore, as soon as such a system is deployed in production, the system's validity must be ensured by closely monitoring the prediction results' robustness (Baier *et al.*, 2019). This can be achieved by frequent retraining or adaptations of the prediction model. Such a monitoring approach results in additional effort for a company and may be accompanied by the need to collect new survey data after some time.

Companies must also ensure that they respect current data security and privacy legislation. This can be especially challenging, since regional legal requirements can differ substantially, i.e. GDPR in Europe in contrast to US legislation (Veale *et al.*, 2018). Thus, a company may need to adapt its approach based on the country it is operating in. In this context, the explainability of decisions and fairness have key roles. Companies need to be aware that biases can still be a problem even if datasets have been pre-processed accordingly. For instance, it has repeatedly been shown that biases regarding gender or race are still included in machine learning models, even if the related attributes have been removed from

the corresponding dataset (Calders and Žliobaite, 2013). Thus, companies must train their employees accordingly. The relationships between a company's privacy policies and customers' attitudes and behaviors are very complex and depend on multiple factors, such as firm, customer and environmental characteristics (Beke *et al.*, 2018). Thus, any adaptations of privacy policies require a cautious approach. It has generally been shown that respecting customers' privacy improves customer–company relationships (Tucker, 2014).

Recent data breaches also emphasize the importance of advanced data security strategies. For instance, customer data can be protected by distributing sensitive information across separate systems, and access to private information must be restricted to a few specific employees (Wedel and Kannan, 2016). Data minimization is another strategy that requires companies to limit data collection to only necessary attributes. Any other data which is not needed should be disposed of. Additional security can be achieved by relying on anonymizing techniques (Verhoef *et al.*, 2016). Nonetheless, since even the most advanced techniques do no guarantee complete data security in all circumstances, companies should develop data breach response plans (Wedel and Kannan, 2016).

Despite these challenges, companies can greatly benefit from better understanding their customers' satisfaction. We have presented a first case to predict customer satisfaction with machine learning. We foresee many further application domains. For instance, hotels could leverage data that they are collecting during check-in, checkout and meals as well as data about room equipment and customer characteristics so as to automatically infer customer satisfaction. Such a process also seems applicable for hospitals and their patients. Emerging technologies and the rapid spread of sensors in many devices (Internet of Things) may even allow to establish a customer satisfaction prediction for shorter service encounters (e.g. customers visiting a restaurant or a shop). However, such predictions will be more challenging, because the amount of data collected via a standardized approach is lower compared to the domains mentioned above.

6.3 Theoretical implications and future work

The importance of a better understanding of the customer, his experiences and satisfaction are omnipresent in today's discussions within the service community. Our work has several implications for theory and future work within this field. In the following, we will discuss possible theoretical links to recent work in service research, namely *customer experience*, *big data analytics in services*, *customer expectations*, *personalized services*, *customer engagement* as well as methods for *capturing customer satisfaction* or *customer churn*. After these theoretical links, we will discuss precise possibilities for future modifications and iterations of our approach.

McColl-Kennedy *et al.* (2019) stress the importance of capturing customer's emotional and cognitive responses as part of the general customer experience. These responses could be captured through customer evaluations and sentiments, e.g. complaints and emotional comments in the incident ticket. They are directly integrated into our approach by considering the incident ticket data as input. Our work therefore shows one opportunity how to automatically analyze and act on the customers cognitive responses. Based on these features, we are also able to show the feasibility to automatically predict customer satisfaction with machine learning within our work. By doing so, we move from a "descriptive" toward a "predictive" paradigm in the analysis of customer encounter data. These results are also in line with the research directions mapped by Ostrom *et al.* (2015). They propose to leverage big data analytics in order to advance existing services as well as to develop novel service offerings.

Accurate predictions of customer satisfaction, either before, during or after the service encounter would allow a fruitful foundation for a broad set of opportunities. If we imagine a robust and highly accurate machine learning prediction model, different constructs and decision consequences could be measured within the application field, e.g. further analyzing

the complex relationship between customer expectations and satisfaction. As Habel *et al.* (2016) propose, it is worthwhile to explore if a controlled increase of customer expectation within a certain application might lead to significant changes in the satisfaction scores—or vice-versa. Another option in line with current research would allow to measure how the personalization of services influences the individual customer satisfaction (Rafaeli *et al.*, 2017). Also, as customer engagement changes in the "big data world", an accurate prediction of individual customer satisfaction levels could prove useful to specifically address and influence the engagement of that individual customer (Kunz *et al.*, 2017).

Additionally, the application of data-based machine learning approaches may shed new light on traditional methods and theories concerning service quality and customer satisfaction. For instance, it is well established that service quality has several components (Parasuraman *et al.*, 1988). By mapping different attributes in our dataset to these components, we would be able to measure these components' influences on overall service quality and potentially discover new influencing factors or even challenge the current understanding of service quality and customer satisfaction. Further, applying such an approach may drive implications for many established service processes that need rethinking, such as the service recovery process, thereby informing service design.

While we choose customer satisfaction as our target variable in this study, it would be worthwhile to similarly try to directly predict churn with the suggested methodology. On that basis, we could analyze key drivers of churn and satisfaction, and in a direct comparison to look for similarities and differences.

Regarding our current approach, we see further potential for enhancement in future iterations: we will improve the model's performance by increasing the amount of labeled data and by applying more sophisticated methods (e.g. word vectorization of the problem description). We will also address the topic of cost inclusion by comparing the costs for a company to implement such an approach into its information systems vs the financial benefits of automatically identifying unsatisfied customers, thereby preventing customer churn. Also, a promising idea is the application of the approach as an aggregate-level gauge for service delivery quality. This could be achieved by monitoring the movement of the overall satisfaction score averaged across all customers. Future work could collect satisfaction data over a longer period and could evaluate such an approach.

Furthermore, besides increasing the capabilities of the prediction itself, additional research should analyze the understanding of customer satisfaction as well as (novel) factors influencing it—as we only scratch the surface of the possibilities in this work. Future research should especially investigate the relationship between incident severity and satisfaction more precisely as the positive connection was rather surprising in our analysis.

Finally, we can conclude that the potential of data-based customer satisfaction prediction is promising. Being able to accurately predict the satisfaction of customers during a service encounter allows organizations to act faster and more efficiently. Therefore, they can retrieve unsatisfied customers earlier and care for an improved customer experience.

Note

1. Other typical terms for customer attrition are customer churn, customer turnover, and customer defection.

References

Anderson, E.W., Fornell, C. and Lehmann, D.R. (1994), "Customer satisfaction, market share, and profitability: findings from Sweden", *Journal of Marketing*, Vol. 58 No. 3, pp. 53-66, doi: 10.1177/002224299405800304.

Arndt, J. and Crane, E. (1975), "Response bias, yea-saying, and the double negative", *Journal of Marketing Research*, Vol. 12 No. 2, pp. 218-220, doi: 10.1177/002224377501200212.

Baier, L., Kühl, N. and Satzger, G. (2019), "How to cope with change? Preserving validity of predictive services over time", *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS-52)*.doi: 10.24251/HICSS.2019.133.

Beke, F.T., Eggers, F. and Verhoef, P.C. (2018), "Consumer informational privacy: current knowledge and research directions", *Foundations and Trends in Marketing*, Vol. 11 No. 1, pp. 1-71, doi: 10.1561/1700000057.

Bergstra, J. and Yoshua, B. (2012), "Random search for hyper-parameter optimization", *Journal of Machine Learning Research*, Vol. 13, pp. 281-305.

Böhmann, T., Leimeister, J.M. and Möslein, K. (2014), "Service-systems-engineering", Springer, *Business and Information Systems Engineering*, Vol. 6 No. 2, pp. 73-79, doi: 10.1007/s12599-014-0314-8.

Brady, M. and Cronin, J. (2001), "Some new thoughts on conceptualizing perceived service quality: a hierarchical approach", *Journal of Marketing*, Vol. 65 No. 3, pp. 34-49, doi: 10.1509/jmkg.65.3.34.18334.

Buckland, M. and Gey, F. (1994), "The relationship between recall and precision", Wiley Online Library, *Journal of the American Society for Information Science*, Vol. 45 No. 1, pp. 12-19, doi: 10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L.

Bughin, J. (2016), *Big Data: Getting a Better Read*, Vol. 2, McKinsey & Company Quarterly.

Buoye, A., Keiningham, T.L., Williams, L. and Aksoy, L. (2014), "Understanding what it takes to be number 1", in Kamdampully, J. (Ed.), *Customer Experience Management: Enhancing Experience and Value through Service Management*, Kendall Hunt Publishing Company, Dubuque, IA, pp. 327-345.

Calders, Toon and Žliobaite, I. (2013), "Why unbiased computational processes can lead to discriminative decision procedures", Custers, B., Calders, T., Schermer, B. and Zarsky, T. (Eds), *Discrimination and Privacy in the Information Society. Studies in Applied Philosophy, Epistemology and Rational Ethics*. Springer, Berlin, Heidelberg, pp. 43-57, doi: 10.1007/978-3-642-30487-3_3.

Cambria, E. and White, B. (2014), "Jumping NLP curves: a review of natural language processing research", *IEEE Computational Intelligence Magazine*, Vol. 9 No. 2, pp. 48-57, doi: 10.1109/MCI.2014.2307227.

Cawley, G.C. and Talbot, N.L.C. (2010), "On over-fitting in model selection and subsequent selection bias in performance evaluation", *Journal of Machine Learning Research*, Vol. 11, pp. 2079-2107.

Chatterjee, D. and Ravichandran, T. (2013), "Governance of interorganizational information systems: a resource dependence perspective", *Information Systems Research*, Vol. 24 No. 2, pp. 261-278, doi: 10.1287/isre.1120.0432.

Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002), "SMOTE: synthetic minority over-sampling technique", *Journal of Artificial Intelligence Research*, Vol. 16, pp. 321-357, doi: 10.1613/jair.953.

Chawla, N.V., Japkowicz, N. and Drive, P. (2004), "Editorial : special issue on learning from imbalanced data sets", *ACM SIGKDD Explorations Newsletter*, Vol. 6 No. 1, pp. 1-6, doi: 10.1145/1007730.1007733.

Chen, C.P. and Zhang, C.Y. (2014), "Data-intensive applications, challenges, techniques and technologies: a survey on big data", *Information Sciences*, Vol. 275, pp. 314-347, doi: 10.1016/j.ins.2014.01.015.

Choi, E., Bahadori, T.B., Sun, J., Kulas, J., Schuetz, A. and Stewart, W. (2016), "Retain: an interpretable predictive model for healthcare using reverse time attention mechanism", *Advances in Neural Information Processing Systems*, pp. 3504-3512.

Churchill, G.A.J. and Surprenant, C. (1982), "An investigation into the determinants of customer satisfaction", *Journal of Marketing Management*, Vol. 14 No. 11, pp. 491-504, doi: 10.1177/002224378201900410.

Davenport, T.H. and Ronanki, R. (2018), "Artificial intelligence for the real world", *Harvard Business Review*, Vol. 96 No. 1, pp. 108-116.

Davis, M.M. and Heineke, J. (1998), "How disconfirmation, perception and actual waiting times impact customer satisfaction", *International Journal of Service Industry Management*, Vol. 9 No. 1, pp. 64-73, doi: 10.1108/09564239810199950.

den Poel, D. and Lariviere, B. (2004), "Customer attrition analysis for financial services using proportional hazard models", *European Journal of Operational Research*, Vol. 157 No. 1, pp. 196-217, doi: 10.1016/S0377-2217(03)00069-9.

Dong, M.C., Fang, Y. and Straub, D.W. (2017), "The impact of institutional distance on the joint performance of collaborating firms: the role of adaptive interorganizational systems", *Information Systems Research*, Vol. 28 No. 2, pp. 309-331, doi: 10.1287/isre.2016.0675.

Eckstein, L., Kuehl, N. and Satzger, G. (2016), "Towards extracting customer needs from incident tickets in it services", *2016 IEEE 18th Conference on Business Informatics (CBI)*, pp. 200-207, doi: 10.1109/CBI.2016.30.

Farris, P.W., Bendle, N.T., Pfeifer, P.E. and Reibstein, D.J. (2010), *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance*, Pearson Education.

Feldman, R. (2013), "Techniques and applications for sentiment analysis", *Communications of the ACM*, Vol. 56 No. 4, pp. 82-89, doi: 10.1145/2436256.2436274.

Fornell, C. (1992), "A national satisfaction barometer: the Swedish experience", *Journal of Marketing*, Vol. 56 No. 1, pp. 6-21, doi: 10.1177/002224299205600103.

Fromm, H., Habryn, F. and Satzger, G. (2012), "Service analytics: leveraging data across enterprise boundaries for competitive advantage", in Bäumer, U., Kreutter, P. and Messner, W. (Eds), *Globalization of Professional Services*. Springer, Berlin, Heidelberg, pp. 139-149, doi: 10.1007/978-3-642-29181-4_13.

Gamon, M. (2004), "Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis", *Proceedings of the 20th international conference on Computational Linguistics*, pp. 841-847, doi: 10.3115/1220355.1220476.

Golub, G.H., Heath, M. and Wahba, G. (1979), "Generalized cross-validation as a method for choosing a good ridge parameter", *Technometrics*, Vol. 21 No. 2, pp. 215-223, doi: 10.1080/00401706.1979.10489751.

Gremler, D.D. and Brown, S.W. (1996), "Service loyalty: its nature, importance, and implications", in Edvardsson, B., Brown, S., Johnston, R. and Scheuing, E. (Eds), *Advancing Service Quality: A Global Perspective*, International Service Quality Association, Jamaica, NY, pp. 171-180.

Grönroos, C. (1984), "A service quality model and its marketing implications", *European Journal of Marketing*, Vol. 18 No. 4, pp. 36-44, doi: 10.1108/EUM0000000004784.

Gustafsson, A., Johnson, M.D. and Roos, I. (2005), "The effects of customer satisfaction, relationship commitment dimensions, and triggers on customer retention", *Journal of Marketing*, Vol. 69 No. 4, pp. 210-218, doi: 10.1509/jmkg.2005.69.4.210.

Habel, J., Alavi, S., Schmitz, C., Schneider, J.-V. and Wieseke, J. (2016), "When do customers get what they expect? Understanding the ambivalent effects of customers' service expectations on satisfaction", *Journal of Service Research*, Vol. 19 No. 4, pp. 361-379, doi: 10.1177/1094670516662350.

Habryn, F., Blau, B., Satzger, G. and Kölmel, B. (2010), "Towards a model for measuring customer intimacy in B2B services", in Morin, J., Ralyté, J. and Snene, M. (Eds), *Exploring Services Science. IESS 2010. Lecture Notes in Business Information Processing*, Springer, Berlin, Heidelberg, doi: 10.1007/978-3-642-14319-9_1.

Hart, C.W., Heskett, J.L. and Sasser, W.E. (1990), "The profitable art of service recovery", *Harvard Business Review*, Vol. 68 No. 4, pp. 148-156.

Harter, J.K., Schmidt, F.L. and Hayes, T.L. (2002), "Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: a meta-analysis", *Journal of Applied Psychology*, Vol. 87 No. 2, pp. 268-279.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Science & Business Media.

Hayes, B. (1998), *Measuring Customer Satisfaction - Survey Design, Use, and Statistical Analysis Methods*, McGraw-Hill, Milwaukee.

Hirt, R., Kühl, N. and Satzger, G. (2017), "An end-to-end process model for supervised machine learning classification: from problem to deployment in information systems", in Maedche, A., vom Brocke, J. and Hevner, A. (Eds), *Designing the Digital Transformation: DESRIST 2017 Research in Progress Proceedings of the 12th International Conference on Design Science Research in Information Systems and Technology*, pp. 53-63.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J. (2003), "A practical guide to support vector classification", Working Paper, National Taiwan University, Taipei, Taiwan, 15 April 2010.

Hsu, C., Lin, Y.-T. and Wang, T. (2015), "A legitimacy challenge of a cross-cultural interorganizational information system", *European Journal of Information Systems*, Vol. 24 No. 3, pp. 278-294, doi: 10.1057/ejis.2014.33.

Hutto, C.J. and Gilbert, E. (2014), "VADER: a parsimonious rule-based model for sentiment analysis of social media text", *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*.

Kart, L., Heudecker, N. and Buytendijk, F. (2013), *Survey Analysis: Big Data Adoption in 2013 Shows Substance behind the Hype*, Gartner.

Keiningham, T., Gupta, S., Aksoy, L. and Buoye, A. (2014), "The high price of customer satisfaction", *MIT Sloan Management Review*, Vol. 55 No. 3, pp. 37-46.

Khan, M. and Fasih, M. (2014), "Impact of service quality on customer satisfaction and customer loyalty: evidence from banking sector", *Pakistan Journal of Commerce and Social Sciences (PJCSS)*, Vol. 8 No. 2, pp. 331-354.

Kieninger, A., Straeten, D., Kimbrough, S., Schmitz, B. and Satzger, G. (2013), "Leveraging service incident analytics to determine cost-optimal service offers", *Wirtschaftsinformatik Proceedings 2013*.

Kotsiantis, S., Kanellopoulos, D. and Pintelas, P. (2006), "Handling imbalanced datasets : a review", *GESTS International Transactions on Computer Science and Engineering*, Vol. 30 No. 1, pp. 25-36.

Kunz, W., Aksoy, L., Bart, Y., Heinonen, K., Kabadayi, S., Ordenes, F.-V., Sigala, M., Diaz, D. and Theodoulidis, B. (2017), "Customer engagement in a big data world", *Journal of Services Marketing*, Vol. 31 No. 2, pp. 161-171, doi: 10.1108/JSM-10-2016-0352.

Ling, C.X., Huang, J. and Zhang, H. (2003), "AUC: a statistically consistent and more discriminating measure than accuracy", *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 519-524.

Liu, B. (2015), *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, Cambridge University Press.

Loveman, G.W. (1998), "Employee satisfaction, customer loyalty, and financial performance an empirical examination of the service profit chain in retail banking", *Journal of Service Research*, Vol. 1 No. 1, pp. 18-31, doi: 10.1177/109467059800100103.

Lundberg, S.M. and Lee, S.-I. (2017), "A unified approach to interpreting model predictions", *Advances in Neural Information Processing Systems*, pp. 4765-4774.

Mattila, A.S. and Enz, C.A. (2002), "The role of emotions in service encounters", *Journal of Service Research*, Vol. 4 No. 4, pp. 268-277, doi: 10.1177/1094670502004004004.

Maxham, J.G. and Netemeyer, R.G. (2002), "Modeling customer perceptions of complaint handling over time: the effects of perceived justice on satisfaction and intent", *Journal of Retailing*, Vol. 78 No. 4, pp. 239-252, doi: 10.1016/S0022-4359(02)00100-8.

McColl-Kennedy, J.R., Zaki, M., Lemon, K.N., Urmetzer, F. and Neely, A. (2019), "Gaining customer experience insights that matter", *Journal of Service Research*, Vol. 22 No. 1, pp. 8-26, doi: 10.1177/1094670518812182.

Messerschmidt, C.M. and Hinz, O. (2013), "Explaining the adoption of grid computing: an integrated institutional theory and organizational capability approach", *The Journal of Strategic Information Systems*, Vol. 22 No. 2, pp. 137-156, doi: 10.1016/j.jsis.2012.10.005.

Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (1994), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood Ltd.

Midi, H., Sarkar, S.K. and Rana, S. (2010), "Collinearity diagnostics of binary logistic regression model", *Journal of Interdisciplinary Mathematics*, Vol. 13 No. 3, pp. 253-267, doi: 10.1080/09720502.2010.10700699.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013), "Efficient Estimation of Word Representations in Vector Space", Working Paper, arXiv preprint arXiv:1301.3781.

Mithas, S., Tafti, A., Bardhan, I. and Goh, J.M. (2012), "Information technology and firm profitability : mechanisms and empirical evidence", *MIS Quarterly*, Vol. 36 No. 1, pp. 205-224, doi: 10.2307/41410414.

Mittal, V. and Kamakura, W.A. (2001), "Satisfaction, repurchase intent, and repurchase behavior: investigating the moderating effect of customer characteristics", *Journal of Marketing Research*, Vol. 38 No. 1, pp. 131-142, doi: 10.1509/jmkr.38.1.131.18832.

Oh, H. (1999), "Service quality, customer satisfaction, and customer value: a holistic perspective", *International Journal of Hospitality Management*, Vol. 18 No. 1, pp. 67-82, doi: 10.1016/S0278-4319(98)00047-4.

Oliver, R.L. (1980), "A cognitive model of the antecedents and consequences of satisfaction decisions", *Journal of Marketing Research*, Vol. 17 No. 4, pp. 460-469, doi: 10.1177/002224378001700405.

Oliver, R.L. (2014), *Satisfaction: A Behavioral Perspective on the Consumer*, Routledge.

Ordenes, F.V., Theodoulidis, B., Burton, J., Gruber, T. and Zaki, M. (2014), "Analyzing customer experience feedback using text mining: a linguistics-based approach", *Journal of Service Research*, Vol. 17 No. 3, pp. 278-295, doi: 10.1177/1094670514524625.

Ostrom, A.L., Parasuraman, A., Bowen, D.E., Patrício, L. and Voss, C.A. (2015), "Service research priorities in a rapidly changing context", *Journal of Service Research*, Vol. 18 No. 2, pp. 127-159, doi: 10.1177/1094670515576315.

Parasuraman, A., Zeithaml, V.A. and Berry, L.L. (1985), "A conceptual model of service quality and its implications for future research", *Journal of Marketing*, Vol. 49 No. 4, pp. 41-50, doi: 10.1177/002224298504900403.

Parasuraman, A., Zeithaml, V. and Berry, L.L. (1988), "SERVQUAL: a multiple-item scale for measuring consumer perceptions of service quality", *Journal of Retailing*, Vol. 64 No. 1, pp. 12-40.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E. (2011), "Scikit-learn: machine learning in Python", *Journal of Machine Learning Research*, Vol. 12, pp. 2825-2830.

Perlich, C., Provost, F. and Simonoff, J.S. (2003), "Tree induction vs. logistic regression: a learning-curve analysis", *Journal of Machine Learning Research*, Vol. 4, pp. 211-255.

Powers, D.M.W. (2011), "Evaluation: from precision, recall and F-measure to roc, informedness, markedness and correlation", *Journal of Machine Learning Technologies*, Vol. 2 No. 1, pp. 37-63.

Rafaeli, A., Altman, D., Gremler, D.D., Huang, M.-H., Grewal, D., Iyer, B., Parasuraman, A. and de Show, K. (2017), "The future of frontline research: invited commentaries", *Journal of Service Research*, Vol. 20 No. 1, pp. 91-99, doi: 10.1177/1094670516679275.

Rahman, M.M. and Davis, D.N. (2013), "Addressing the class imbalance problem in medical datasets", *International Journal of Machine Learning and Computing*, Vol. 3 No. 2, pp. 224-228.

Rud, O.P. (2001), *Data Mining Cookbook: Modeling Data for Marketing, Risk, and Customer Relationship Management*, John Wiley & Sons.

Rust, R.T. and Oliver, R.L. (1994), "Service quality: insights and managerial implications from the Frontier", in Rust, R.T. and Oliver, R.L. (Eds), *Service Quantity: New Directions in Theory and Practice*, Sage Publications, London, pp. 1-19, doi: 10.4135/9781452229102.n1.

Rust, R.T. and Zahorik, A.J. (1993), "Customer satisfaction, customer retention, and market share", *Journal of Retailing*, Vol. 69 No. 2, pp. 193-215.

Schüritz, R., Satzger, G., Seebacher, S. and Schwarz, L. (2017), "Datatization as the next Frontier of servitization – understanding the challenges for transforming organizations", *Proceedings of International Conference on Information Systems (ICIS) 2017: Transforming Society with Digital Innovation*.

Seebacher, S. and Schüritz, R. (2019), "Blockchain–just another IT implementation? A comparison of blockchain and interorganizational information systems", *Proceedings of the 27th European Conference on Information Systems (ECIS)*.

Shostack, G.L. (1985), "Planning the service encounter", in Czepiel, J.A., Solomon, M.R. and Surprenant, C.F. (Eds), *The Service Encounter*. Lexington Books, Lexington, MA, pp. 243-254.

Sureshchandar, G.S., Rajendran, C. and Anantharaman, R.N. (2002), "The relationship between service quality and customer satisfaction – a factor specific approach", *Journal of Services Marketing*, Vol. 16 No. 4, pp. 363-379, doi: 10.1108/08876040210433248.

Tax, S.S., Brown, S.W. and Chandrashekaran, M. (1998), "Customer evaluations of service complaint experiences: implications for relationship marketing", *Journal of Marketing*, Vol. 62 No. 2, pp. 60-76, doi: 10.1177/002224299806200205.

Trubik, E. and Smith, M. (2000), "Developing a model of customer defection in the Australian banking industry", *Managerial Auditing Journal*, Vol. 15 No. 5, pp. 199-208, doi: 10.1108/02686900010339300.

Tsai, C.-F. and Lu, Y.-H. (2009), "Customer churn prediction by hybrid neural networks", *Expert Systems with Applications*, Vol. 36 No. 10, pp. 12547-12553, doi: 10.1016/j.eswa.2009.05.032.

Tucker, C.E. (2014), "Social networks, personalized advertising, and privacy controls", *Journal of Marketing Research*, Vol. 51 No. 5, pp. 546-562, doi: 10.1509/jmr.10.0355.

Veale, M., Binns, R. and Van Kleek, M. (2018), "Some HCI Priorities for GDPR-Compliant Machine Learning", working paper, arXiv preprint arXiv:1803.06174.

Verhoef, P.C., Kooge, E. and Walk, N. (2016), *Creating Value with Big Data Analytics: Making Smarter Marketing Decisions*, Routledge.

Wedel, M. and Kannan, P.K. (2016), "Marketing analytics for data-rich environments", *Journal of Marketing*, Vol. 80 No. 6, pp. 97-121.

Woo, K. and Fock, H.K.Y. (2004), "Retaining and divesting customers: an exploratory study of right customers, "at-risk" right customers, and wrong customers", *Journal of Services Marketing*, Vol. 18 No. 3, pp. 187-197, doi: 10.1108/08876040410536495.

Woodside, A.G., Frey, L.L. and Daly, R.T. (1989), "Linking service quality, customer satisfaction, and behavioral intention", *Journal of Health Care Marketing*, Vol. 9 No. 4, pp. 5-17.

Appendix

1. Hyperparameter optimization

This paragraph describes the steps that we performed for the hyperparameter optimization for the machine learning approach. So far, we had only applied logistic regression with a fixed parameter configuration. Thus, we decided to optimize the C-value and the class weight parameter of the logistic regression. The C-value refers to the regularization applied to the model. High regularization forces the algorithm to select only a few features with high coefficients, influencing the overall prediction. However, an overly strong regularization will lower predictive performance since the resulting lack of

features will not allow for a meaningful prediction. Further, we optimized the class weight parameter that gives importance to the target classes based on their frequency in the dataset. In this use case, the class of dissatisfied customers will usually receive a higher weight owing to its lower overall frequency in the dataset.

As suggested by Hsu *et al.* (2003), at first, we used a broad range with exponentially growing values for the hyperparameter optimization; we then limited this range by only considering values with better performance.

Table A1 shows the parameter range we used for the final model tuning and performance estimation. We decided to use random search with 5,000 iterations for hyperparameter optimization, since it provided good results and is suitable for limited computational power (Bergstra and Yoshua, 2012). We performed the parameter optimization with a nested cross-validation with five outer folds and three inner folds so as to prevent overfitting (Cawley and Talbot, 2010). We used the three inner folds for the parameter optimization and the five outer folds for the performance estimation.

Table A1.
The final parameter range considered in the hyperparameter optimization

Parameter	Range
C (regularization parameter)	0.00001, 0.00002, . . . 0.001
Class weight	1:5, 1:6, . . . 1:25

2. SHAP values for submodel

Figure A1 shows the SHAP values for the submodel where we predict the compound sentiment score based on the feature incident duration, initial contact delay, contact misses and product. This way, we can estimate the impact of those features on satisfaction via the sentiment score. Since this is a regression problem—the compound sentiment score is distributed between −1 and 1—it is impossible to use logistic regression for this analysis. Therefore, we apply a boosting regression model. The SHAP values do not indicate an unambiguous influence of the different features on the compound sentiment score.

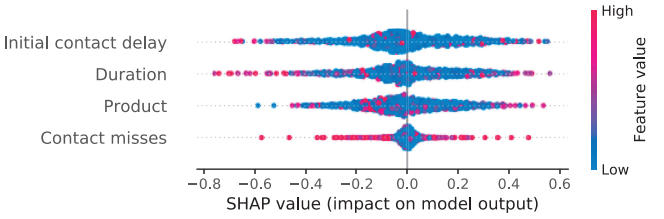

Figure A1.
SHAP values for prediction of compound sentiment

Corresponding author
Lucas Baier can be contacted at: lucas.baier@kit.edu