Business Analytics for Sales Pipeline Management in the Software Industry: A Machine Learning Perspective

Verena Eitle
Technische Universität Darmstadt
eitle@is.tu-darmstadt.de

Peter Buxmann
Technische Universität Darmstadt
buxmann@is.tu-darmstadt.de

Abstract

This study proposes a model designed to help sales representatives in the software industry to manage the complex sales pipeline. By integrating business analytics in the form of machine learning into lead and opportunity management, data-driven qualification support reduces the high degree of arbitrariness caused by professional expertise and experiences. Through the case study of a software provider, we developed an artifact consisting of three models to map the end-to-end sales pipeline process using real business data from the company's CRM system. The results show a superiority of the CatBoost and Random Forest algorithm over other supervised classifiers such as Support Vector Machine, XGBoost, and Decision Tree as the baseline. The study also reveals that the probability of either winning or losing a sales deal in the early lead stage is more difficult to predict than analyzing the lead and opportunity phases separately. Furthermore, an explanation functionality for individual predictions is provided.

1. Introduction

The high rate of business changes and the ongoing digital transformation in the global environment compel modern enterprises to remain agile and competitive by evolving their business processes accordingly. Based on the concept of dynamic capabilities, organizations can maintain and even strengthen their competitive advantage particularly in times of market uncertainty and fierce competition by creating, renewing and orchestrating their resources and assets [1, 2]. With the purpose of increasing business performance, companies have adopted business analytics on a large scale as data-driven decision-making procedures enhance business processes and enable the identification of market opportunities and threats [3, 4]. From the perspective of dynamic capabilities [1, 2], applying business analytics technologies in Customer Relationship Management (CRM) systems drives business value steadily as Information Technology (IT) resources and corporate assets such as organizational data are integrated and reorganized. Nam et al. [5] demonstrate in their research that the increase in CRM performance depends positively on the usage of business analytics, whereby data quality must be continuously improved. In general, CRM applications facilitate the process of managing and coordinating customer interactions with the primary goal of ensuring long-term customer value by improving customer acquisition and increasing customer retention [6, 7]. Therefore, converging CRM systems and business analytics technologies enables firms to analyze and incorporate valuable insights in their customer interactions and decision-making procedures to maximize customer value.

The study of Ngai et al. [8] presents that, besides statistical and mathematical approaches, the emergence of machine learning (ML) in the CRM context offers great potential for discovering and deriving insightful information from enterprise data. The increasing significance in customer centricity and the availability of customer data enable organizations to apply ML techniques, especially in the fields of customer identification, attraction, retention, and development. However, the majority of CRM literature focuses more on customer retention than on customer acquisition [9] as the establishment of long lasting customer relations and the associated cross and upsell potentials have a positive impact on corporate profitability [10, 11]. Nevertheless, since customer acquisition strategies are considered as a counterpart to customer retention, companies must also ensure a clear focus on gaining new customers on a consistent basis. Customer acquisition strategies are crucial for a company's success from the perspective of increasing market size in strategic industries, and exploiting new customer markets and product [12, 13]. Acquiring new customers involves significant effort and expenses as the sales pipeline process embraces several stages from the initial contact to the final sales deal. In general, the first phase of identifying and addressing prospects who

URI: https://hdl.handle.net/10125/59541
ISBN: 978-0-9981331-2-6

express first interest in purchasing a product is defined as lead management. The following phase of opportunity management includes all sales related activities that are tailored to the specific requirements of the sales prospect, and thus contribute to the successful closing of a sales deal [14, 15].

Since a data-driven decision-making process reduces the degree of human intuition through data analysis [16], this research paper proposes the integration of business analytics in the form of ML techniques in the lead and opportunity management phases. Despite the focus on applying ML methods in the CRM context such as in churn prediction [e.g. 17, 18] and the tremendous efficiency potential in sales procedures, the amount of academic contributions in the field of sales pipeline management have been insufficient up until now. To date, only a few scholars have dedicated their research to the development of ML models that facilitate the sales pipeline qualification process by predicting the likelihood of winning a sales deal [19, 20, 21]. In contrast to their rather narrow view on either the lead or the opportunity phase, we developed an artifact that takes the entire end-to-end sales pipeline process into consideration; from the initial lead phase, to the opportunity phase, and finally to the sales deal closing. Furthermore, we place more emphasis on the high number of categorical features arising from the sales pipeline management than existing state-of-the-art models by applying the CatBoost classifier that achieves superior results through its specialization on categorical data. By integrating an explanation model, we additionally increase the transparency of current black-box algorithms and enable salespeople to understand the impact on individual feature values. To reflect highly complex sales structures and long sales cycles, our study is based on a case study of a company, specializing in enterprise application software. The suitability and usability of the artifact can thus be tested on other business-to-business (B2B) case studies with similar convoluted sales structures. Therefore, this research aims to analyze the prediction of all three sales pipeline scenarios: 1) lead-to-opportunity, 2) opportunity-to-sales deal and 3) lead-to-sales deal to embrace the involvement of both marketing and sales. Thus, we investigate the following research questions:

RQ1: Can ML techniques be applied to the end-to-end sales pipeline process to predict the purchase probability in the lead and opportunity stage?

RQ2: Which ML techniques achieve the best predictive performance in the lead and opportunity qualification process?

Due to the strong profitability pressure in the license-driven software industry, the primary objective is the development of ML models that support salespeople in the qualification process of leads and opportunities. To reduce the level of arbitrariness in managing the sales pipeline, we propose a data-driven approach based on ML techniques. The remainder of this paper is structured as follows: First, the theoretical background of the sales process and the ML methods are outlined. After elaborating the research setting, the results of this study are presented. In the subsequent sections, we discuss our conclusions, highlight the limitations, and propose opportunities for future research.

2. Theoretical Background

2.1. Sales Pipeline Process

Despite the high technical maturity of CRM systems, to date no universally acknowledged definition exists amongst scholars and practitioners [22]. Most publications, however, share the common understanding that a CRM application embraces all touch points of a customer life cycle to ensure long-term customer value [23, 24]. Since CRM functions leverage business performance on a strategic, operational and analytical basis, the database is considered as a crucial corporate asset [7]. Combining the operational level of the lead and opportunity management with analytical CRM functions provides a central support for future sales potentials [25, 26].

Due to the large amount of hidden information in sales data, adopting a data-driven approach through predictive analytics helps salespeople to prioritize promising prospects [8]. In general, a sales pipeline process follows the structure of a sales funnel that consists of lead generation, opportunity management and the final sales deal [14, 15]. The lead stage comprises all marketing-related activities of identifying prospects that first express their interest in buying a product. After qualification and evaluation procedures conducted by marketing, the lead will be handed over to sales and converted into an opportunity. In this stage, salespeople take appropriate actions such as product demos and client meetings to maximize the likelihood of closing the sales deal. The primary goal is to ensure an increase in revenue and a growing customer base [27]. However, the qualification assessment is mainly influenced by personal judgement of the respective marketing or sales workforce. Relying on the professional competences and prior experiences leads to counterproductive effects as the personal bias might cause misjudgments within the sales pipeline [28]. For instance, salespeople tend to deliberately manipulate the sales pipeline to achieve their own sales quotas. Prospects can be either underrated to avoid additional management attention or overrated to

simulate the achievement of sales targets. In addition, sales negotiations may be intentionally postponed to upcoming quarters [20, 29]. In general, this qualification process requires a great effort as a recent appraisal states that "on average, sales reps spend 80 percent of their time qualifying leads and only 20 percent in closing" [30].

Taken these challenges into account, fostering automation in the lead and opportunity management is perceived as a significant benefit for organizations. According to Syam and Sharma [31], integrating ML techniques in the qualification assessment of leads and opportunities enables enterprises to simultaneously reduce subjective bias and to improve quality assurance. Due to these benefits, the development of ML models applicable for the sales pipeline is gaining importance in the research environment. For example, Yan et al. [20] present a win-propensity model based on ML algorithms that is built upon static features including company profile characteristics such as deal size, geography and industry as well as interaction sequences captured by the pipeline system. A relatively high accumulation of interaction activities including login, browsing, and updating of leads within a short period of time indicates a higher chance of winning the deal. The model developed by Megahed et al. [21] embraces the multi-stage sales pipeline by taking the diverse maturity levels of opportunities into account. As the focus rather lies on predicting the sales forecast generated by the opportunities, the sales pipeline growth towards the end of the target time period plays a crucial part. Another data-driven approach to prioritize prospects based on the likelihood of a purchase is presented by D'Haen and Van den Poel [19]. They propose a model that in the first phase applies unsupervised ML techniques to find similarities between existing customers and prospects and consequently rank them based on the sales probability. The second phase determines the actual probability of winning or losing the sales deal with the use of ML classifiers such as the logistic regression, decision trees, and neural networks. The third phase combines both approaches and therefore provides a ranked list of prospects. However, the prevalent black-box approach of ML models impedes the interpretation of findings as their complexity obfuscates the inner workings. This opacity makes it difficult for the recipient to understand how the output was achieved by the given input data [32, 33]. In order to create transparency, Bohanec et al. [34] present, in addition to the sales prediction, an explanation model that allows a deeper comprehensibility and transparent evaluation of the opportunity prediction. This model allows domain experts to evaluate the ML based results by incorporating the impact level of the given attributes.

2.2. Machine Learning Methods - Classification Techniques

The term machine learning describes a concept that enables computers to learn rather than being explicitly programmed [35]. In 1997, Tom Mitchell stated that "[a] computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" ([36], S. 2). Therefore, the goal of supervised learning is to learn a mapping function from input x to output y that correctly predicts the value of y when exposed to new data [37, 38]. Lead and opportunity management seems to be an appropriate field for the use of machine learning as organizations generally possess sufficient historical customer data. In the following, we would like to establish a common understanding of the supervised ML algorithms used in our artifact. To determine the best ML technique, we have set a traditional decision tree as the baseline.

Random Forest

As an advancement of decision trees, Random Forest is ideally suited to solve classification problems. The lack of robustness and the high instability of decision trees [39] led to the development of Random Forest introduced by Breiman [40]. As an ensemble approach, the algorithm generates a large number of decision trees on which the majority of votes determines the most popular class. In general, each tree is grown by using only a subset of randomly selected predictors that ultimately predict the final class. In addition to the robustness against outliers and noise, a major advantage of this classifier lies in the deeper interpretability of the black-box structure [40].

Support Vector Machine

Support Vector Machines (SVMs) were initially introduced by Cortes and Vapnik [41] with the purpose of solving binary classification tasks. In a binary context, a SVM defines an optimal hyperplane that maximizes the margin between two classes with the nearest data points defined as a support vector. To solve non-linearly separable problems, kernel functions such as sigmoid, polynomial and radial basis function (RBF) are used as remedies. The idea is to implicitly map the original feature space into a higher dimensional feature space to separate data linearly by a hyperplane [41]. A SVM differs from other linear classifiers as the optimal linear separator can even be found in feature spaces with multiple dimensions [37].

XGBoost

The eXtreme Gradient Boosting algorithm, shortened to XGBoost, developed by Chen and Guestrin [42] has recently gained popularity in ML competitions. The fundamentals are based on the gradient boosting

framework introduced by Friedman [43] that is built on the tree ensemble model, allowing to group several weak learners into a strong learner. By following an adaptive strategy, each successive tree is created to predict the residual of the prior tree that will be added to the final prediction. XGBoost outperforms other algorithms in scalability and model performance as parallel and distributed computing is enabled and missing data is handled automatically [42].

CatBoost

The CatBoost algorithm, recently launched by the company Yandex, is an implementation of gradient boosting that handles categorical data. As the ensemble of trees can generally only handle numeric features, converting categorical features to numbers requires major preprocessing efforts such as the one-hot-encoding technique that transforms each category into binary variables. Instead of these time-consuming preprocessing steps, CatBoost handles categorical data efficiently as after performing randomly permutation, an average label value is computed for each example when the same value was set before the permutation. In addition, overfitting is prevented by using multiple permutations for training different models [44].

3. Research Setting

While several approaches exist to predict sales deals through ML techniques [19, 20, 21, 34], these state-of-the-art models bear deficiencies in at least two aspects. First, these studies limit their scope of research to either the lead or the opportunity phase, and thus do not reflect the different maturity levels of the end-to-end sales pipeline process. Second, the existing prediction models lack transparency due to their black-box approaches. In order to address these gaps, we apply the Design Science Research (DSR) [45] to design an artifact for sales prediction along the end-to-end sales pipeline process. Since our objective is to develop a new prediction model for a known problem, the DSR contribution type is considered as an improvement [46]. To revise the artifact, we follow the iterative design cycle of Takeda et al. [47], comprising the DSR activities of awareness, suggestion, development, evaluation, and conclusion. In the first phase, we conducted a detailed literature research as presented in the previous chapter to identify the problem and specify the expectations. The second and third phases comprise model development activities including the definition of various sales pipeline scenarios and pre-processing steps such as class label verifications, feature selection techniques, and data cleansing, followed by the division of the data set into training and test sets, the application of undersampling techniques and hyperparameter methods. Besides these activities as described in the following section, we have also defined metrics to compare the prediction performances of the selected algorithms. For the evaluation phase the case study was chosen as the evaluation type presented by Peffers et al. [48] to test the artifact for its suitability and usability in a real-life situation. Details on the case study are presented in the section of data set description, followed by the predictive performance results of the artifact.

3.1. Model Development

Since our objective is to cover the entire end-to-end sales pipeline process, we developed three classification models to predict the following cases: 1) lead-to-opportunity, 2) opportunity-to-sales deal, 3) lead-to-sales deal as illustrated in Figure 1. The first model reflects the case when a lead is either converted in an opportunity or discontinued. To take the existing sales pipeline procedure of the respective company into account, the second model embraces both opportunities arising from this conversion and the opportunities created directly by a salesperson. Unlike the first two models, the results of the third model focus on leads that have been either won as a sales deal or lost, meaning that directly created opportunities are not considered in the results of the end-to-end process. In terms of feature selection, we have excluded variables from the original data set based on the following criteria: redundant features, amount of missing values that accounts for more than 50% of the data set as well as name- and team-based variables (to avoid performance benchmarking). In order to evaluate the various classification methods described above, we split the data set into a training and test set by randomly assigning 70% of the data to train the model and the remaining 30% to test the model on an unseen data sample. Due to the different phases along the sales process, the class labels refer either to the case where a lead will be converted or discontinued, or to the likelihood of winning or losing a sales deal. The average conversion rate of leads to sales deals of 10% in the B2B sector [49], however, leads to the presence of data imbalance. To reduce the risk of a class being favored by the presence of data imbalance, we use the technique of random undersampling on the training set, which eliminates random samples from the majority class. In addition, we apply the hyperparameter optimization method GridSearch along with a 10-fold cross-validation to determine the best combination of parameter values. Regarding SVM, we set the RBF kernel as the default kernel function and conducted a parameter search of the penalty parameter C and the kernel parameter gamma [50]. Tuning Random Forest refers to the optimal parameter selection of numbers of

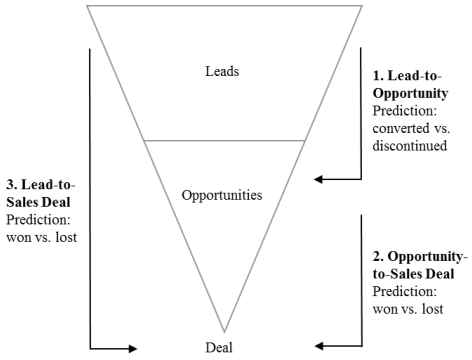

Figure 1. Sales pipeline

trees, max depth of trees, as well as minimum number of samples to split an internal node and to be at a leaf node, whereby the decision tree excludes the first mentioned parameter. The performance of XGBoost can be improved by finding the most favorable combination of the learning rate, the minimum sum of weights of all observations required in a child and the maximum depth of a tree. Finally, we tuned CatBoost by adjusting the learning rate and the tree depth.

3.2. Evaluation Metrics

To detect the best performing supervised classifier for the presented prediction task, appropriate evaluation metrics must be applied. The basis for these measures represents the confusion matrix which respectively denotes the true-positive and false-positive cases as TP and FP and describes true-negative and false-negative cases as TN and FN. For all three classification models, the Percentage Correctly Classified (PCC), also known as accuracy (Acc.), is calculated to indicate the ratio of correctly classified cases to the total number of classified records using the equation of (TP+TN)/ (TP+FP+TN+FN). To overcome the disadvantage of PCC's lack of robustness to data imbalance, the evaluation metrics are extended by the measures of sensitivity (Sens.), specificity (Spec.), precision (Prec.) and F1. Sensitivity refers to the true-positive rate as it reflects the proportion of positive cases that are correctly classified through the equation of TP/(TP+FN), whereby specificity measures the proportion of negatives that are correctly identified as negatives through the equation of TN/(TN+FP). In contrast, the precision calculates the probability of a sample classified as positive to be positive with the following equation TP/(TP+FP). However, since reaching good results with one of these measures does not necessarily imply good performance on the other, we use the evaluation metric F1 by calculating the equation of 2* (precision * sensitivity) / (precision + sensitivity) [51,38]. Furthermore, in contrast to the presented point-wise evaluation metrics, we additionally measure the area under the receiving operating curve (AUC) which plots sensitivity and 1-specificity at various threshold settings. Taking all thresholds into account, the AUC measure is ideally suited to compare the overall performance of the presented classifiers [52].

3.3. Data Set Description

For this study, we have gathered B2B sales data from a software company listed in the Fortune 500 to develop a ML classification model that supports sales representatives in their lead and opportunity qualification process by providing the probability of a purchase. To reflect the complex sales processes in the license-driven industry and to make the decision-making procedures in the sales pipeline less arbitrary, this provider of enterprise application software serves as a case study. By obtaining real business data from the company's internal CRM system, the artifact is developed on industry-specific sales conditions and peculiarities. Capturing lead and opportunity data in the period from January 2015 to July 2017 clearly represents the long and complex sales cycle of enterprise application software. Furthermore, the data set embraces all business regions of the software provider consisting of Middle and Eastern Europe (MEE), Middle East and Africa North (EMEAN), Europe, Middle East and Africa South (EMEAS), North America (NA), Latin America (LA), Asia Pacific Japan (APJ) and Greater China (GC). After applying feature selection techniques based on the mentioned specifications, the feature set contains 17 categorical and 19 numeric variables for the lead stage as well as 22 categorical and 20 numeric variables for the opportunity stage. Due to the compliance guidelines of the respective company, we can only outline the features in a broadly manner. Customer features refer, for example, to company size, industry, purchasing lifecycle, and location, whereby campaign features include campaign types, detailed descriptions as well as objectives. In addition to the sales channels and sales units being covered by the sales features, the product portfolio and deployment options are listed in the product features. Detailed information such as competitor, time and pipeline specifications are mentioned in lead-/opportunity-related features, which apply for both leads and opportunities. Furthermore, our assumption of unequal class label distribution is reflected in our data set, which leads to data imbalance. As shown in Table 1, the relatively high imbalanced class distribution differs across the software provider's business regions, leading to the assumption that

Table 1 Data imbalance

Region	1.Lead-Opportunity	2.Opportunity-Sales Deal	3.Lead-Sales Deal
MEE	60% / 40%	34% / 66%	20% / 80%
NA	73% / 27%	22% / 78%	8% / 92%
LA	55% / 45%	18% / 82%	10% / 90%
APJ	93% / 7%	13% / 87%	4% / 96%
GC	78% / 22%	14% / 86%	8% / 92%
EMEAS	78% / 22%	24% / 76%	13% / 87%
EMEAN	75% / 25%	19% / 81%	8% / 92%

regional specific procedures exist in handling the sales pipeline. By applying random undersampling on the training set, we ensure a balanced label class distribution for training the models. In summary, it must be noted that after verifying the sales pipeline procedure with the company we can ensure that the three models reflect the existing sales pipeline process.

4. Results of Predictive Performance

To evaluate and compare the prediction performances of the induced classifiers, we train and test the supervised algorithms on all three sales pipeline scenarios 1) lead-to-opportunity, 2) opportunity-to-sales deal, 3) lead-to-sales deal separately, using real-life business data from the company. As the data set reflects major regional differences in sales pipeline management, we must distribute the data records among the respective sales regions in order to reduce data bias. On the one hand, data bias might occur due to the conservative or likely lead and opportunity conversion procedures as well as the different CRM maintenance in each region. On the other hand, data bias might be caused by the behavior of the salesperson himself as his personal preferences and professional experiences could have influenced the decision in the lead or opportunity phase. By analyzing the model on a regional level, we were able to eliminate data bias caused by regional differences. However, the reduction of human intuition requires further research in non-standard ML approaches to solve the problems of subjectivity and noisy labels which is outlined in detail in the last chapter. Despite relatively similar results across the globe, we present the predictive performance of a particular sales region which remains anonymous due to compliance guidelines. This choice is based on the strong sales success and the high market share of this sales territory as well as the limited space of this research paper. After randomly dividing the data into the training and test set as well as eliminating data imbalance on the training set by using the random undersampling technique, we receive a total of 36929 unique leads for this sales region, splitted into 24170 records for training and 12759 for testing the first model. The second model is developed through the availability of 26216 unique opportunities, resulting in 16046 training samples and 10170 testing samples. Since the data imbalance of the end-to-end sales process in terms of won sales deals leads to an insufficient sample size, the third model is initially trained and tested on the basis of opportunity records. To ensure consistency with the lead data, these opportunities were selected based on an identical feature set and the involvement of a marketing campaign, as being a key feature of the lead phase. Subsequently, the classification model is then tested with historical lead data whose records resulted in either a closed or a lost sales deal. Therefore, for the second test series alone, we have a total of 10730 unique leads at our disposal that exhibit sales negotiation histories within this region. To avoid data redundancies in the third model, we ensure that the opportunity information arising from leads is eliminated in the initial data set for training and testing the third model, and that it is only used in the second testing phase. In general, all supervised algorithms including the baseline, Random Forest, SVM, XGBoost and CatBoost are applied on the test set for the selected sales region. Table 2 gives an overview of the predictive performances of all three classification models in terms of accuracy, sensitivity, specificity and F1. When comparing classification techniques, all four algorithms offer similar performances and exceed the baseline. Taking accuracy into account, CatBoost is with 78% and 79% the best classifiers in the first two models, whereby the same moderate results are also reached by SVM in the first and by XGBoost in the second. In the third model, Random Forest exceeds the results of the other algorithms with an accuracy of 71%. In terms of sensitivity and specificity, it is striking that the first two models show only minimal differences of just 0.05% between these evaluation metrics, indicating that no class is preferred. In contrast, the specificity of the third model far exceeds the sensitivity for all classifiers. These large discrepancies point out that the cases of losing the sales deals in the lead stage are more often correctly classified than the positive cases. Regarding F1, it can be observed that the relatively high results in the first and the second model indicate high performance and equality of sensitivity and precision. However, the relatively low F1 results of the third model are caused by the large discrepancies mentioned above. Since the best classifier cannot be clearly identified, with the given evaluation metrics, we also compare the AUC performance shown in Table 3. In terms of the lead-opportunity model, CatBoost outperforms the other classifiers with an AUC of 0.86, confirming the results

Table 2 Predictive performance results

Methods	1. Lead-Opportunity					2. Opportunity-Sales Deal					3. Lead-Sales Deal				
	Acc.	Sens.	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1	Acc.	Sens.	Spec.	Prec.	F1
Baseline	0.74	0.74	0.75	0.81	0.77	0.75	0.75	0.75	0.61	0.67	0.57	0.48	0.59	0.23	0.31
Random Forest	0.77	0.75	0.80	0.84	0.80	0.77	0.78	0.77	0.64	0.70	**0.71**	0.35	0.80	0.31	0.33
SVM	**0.78**	0.78	0.78	0.83	0.81	0.77	0.78	0.77	0.64	0.70	0.64	0.26	0.75	0.21	0.23
XGBoost	0.77	0.76	0.78	0.83	0.80	**0.79**	0.79	0.79	0.66	0.72	0.67	0.32	0.76	0.25	0.28
CatBoost	**0.78**	0.76	0.80	0.84	0.80	**0.79**	0.80	0.78	0.66	0.72	0.58	0.56	0.59	0.26	0.36

*Acc.=Accuracy, Sens.=Sensitivity, Spec.=Specificity, Prec.=Precision

of accuracy. With regard to the AUC of 0.88, the probability of winning or losing a sales deal is also best predicted with CatBoost, as the results of accuracy and F1 prove. As the best AUC of the third model yields 0.63, the Random Forest far exceeds the other results of 0.54 (SVM), 0.59 (XGBoost), 0.60 (CatBoost) and 0.59 (Baseline). In contrast to the other sales pipeline models, the Random Forest is therefore seen as the best performing supervised algorithm for predicting the probability of a sales deal in the initial lead phase. Examining the best results across the three sales pipeline models, it is obvious that the third model with a difference of 23-25% in AUC performs much worse than the pure lead and opportunity models. In addition to the evaluation metrics, the proposed artifact also provides an explanation model for a lead or an opportunity. Instead of showing salespeople only the accuracy, the implementation of a novel explanation technique, presented by Ribeiro et al. [53], allows to explain individual predictions by learning an interpretable model locally around them. Figure 2 depicts the explanation model of a randomly selected opportunity in relation to its feature importance using the Random Forest classifier. The prediction probabilities are displayed on the left, whereby the two graphs on the right assist salespeople to understand which feature values were most relevant for predicting the outcome. Considering this example, the values of product feature 3, customer features 4 and 1 positively influence the likelihood, while product feature 1, opportunity features 2 and 3 have the opposite effect.

Table 3 AUC metric

Methods	1. Lead-Opportunity	2. Opportunity-Sales Deal	3. Lead-Sales Deal
Baseline	0.84	0.83	0.59
Random Forest	0.85	0.86	**0.63**
SVM	0.85	0.85	0.54
XGBoost	0.85	0.87	0.59
CatBoost	**0.86**	**0.88**	0.60

This visualization allows salespeople to incorporate data-driven approaches in their qualification process.

5. Discussion

In this study, we propose three ML models as an artifact that support salespeople in their qualification process for the following sales pipeline scenarios 1) lead-to-opportunity, 2) opportunity-to-sales deal, 3) lead-to-sales deal. The results in accuracy and AUC of the first two classification models show that CatBoost clearly outperforms the other supervised algorithms. Due to this strong predictive performance, we would like to emphasize the attractiveness of this algorithm which refers to the sophisticated support of categorical features. Instead of converting each categorical value into binary values through the widely-used one-hot-encoding technique, CatBoost applies an efficient encoding method that leads to quality improvement by reducing overfitting. Since lead and opportunity data usually contain many categorical features such as in our case in marketing campaign, customer, sales and product data, this supervised algorithm is ideally suited to identify promising prospects. Predicting the sales probability in the early lead stage is best performed by Random Forest whose results significantly outperform SVM, XGBoost, CatBoost, and the baseline in terms of accuracy and AUC. Given the nature of Random Forest, our expectations regarding the strong predictive performance and the robustness to outliers and noise of this classifier were clearly met. To our knowledge, our study is among the first to demonstrate the high predictive performance of CatBoost in the lead and opportunity management through the excellent processing of categorical data. Despite the large presence of categorical data and the focus on supervised ML techniques, the study of D'Haen and Van den Poel [19] and Bohanec et al. [34] only apply standard ML algorithms such as decision tree, logistic regression, and neural networks. However, as our study shows, CatBoost is ideally suited for the lead and opportunity management which is characterized by its large amount of categorical data. Unlike existing

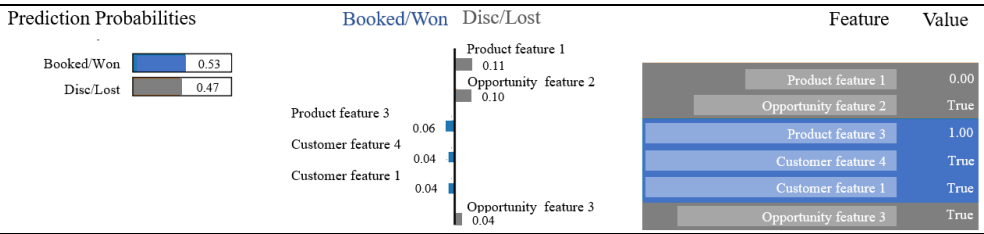

Figure 2 Explanation model

approaches, our artifact examines the end-to-end sales pipeline process by developing and comparing the predictive performances of the three sales pipeline models, covering the entire process of leads, opportunities and sales deals. In contrast to purely limiting the scope of research to either the lead [19] or the opportunity [21, 34] phase, our artifact takes the specific maturity levels of leads and opportunity into consideration. The marketing-oriented activities in the lead stage as well as the sales specific activities in the opportunity phase are clearly covered by the three models. In contrast, the model of Yan et al. [20], for example, does not consider crucial marketing-related information in the lead phase. Therefore, our study explicitly reflects the different phases along the sales funnels by carefully taking the maturity levels of leads and opportunities into account. Nevertheless, the large differences in performance between the three prediction scenarios also reflect the usability of the artifact. The very low AUC of the third model depicts that the likelihood of sales deals in the early lead stage can hardly be predicted. The large gap between sensitivity and specificity as well as the resulting poor F1 performance also point to the same assumption. Despite identical feature sets, a possible reason could be that the feature values of the opportunities, on which the model is initially trained, are more advanced along the sales cycle than the lead information available for testing the model. To give an example from our specific data set, the product information of opportunities is much more mature compared to leads as product requirements of enterprise application software, budget information and, general conditions are usually shared and communicated within the sales negotiations. Taking the results of this study into account, we can emphasize that mapping an end-to-end sales pipeline process into a single classification model does not yield the expected performance. Therefore, two separate lead and opportunity models, as presented in this study, are more suitable to predict whether a lead will be converted or discontinued, or a sales deal will be won or lost. This approach ensures that the different maturity levels of the lead and opportunities phases are reflected in the feature values. Furthermore, our artifact extends the existing state-of-the-art black-box prediction models [19, 20, 21] by applying the novel explanation technique by Ribeiro et al. [53]. Instead of just displaying the prediction performances, salespeople are able to analyze the impact of the individual feature values in order to follow the decision-making process based on ML techniques. Consequently, the first two models are highly recommended to assist sales representatives in qualifying their sales pipeline through data-driven decision support. In addition, it should be noted that our artifact is trained and tested using original real-life data extracted from the company's CRM, rather than pseudo tests and manually added attributes [19, 34]. Overall, by comparing the results of Random Forest, SVM, XGBoost, CatBoost, and the baseline across the lead and opportunity phases, we would like to emphasize that our research serves as a benchmark that has not yet been examined to this extent.

This research paper makes several contributions to research and practice. We designed a first version of an artifact for sales prediction along the end-to-end sales pipeline process whose applicability and suitability can be further tested and developed on other case studies with similar complex sales pipeline processes. By explicitly taking the lead and the opportunity phase into account, we were able to reflect the different maturity levels across these sales processes. After evaluating the artifact through the case study of an enterprise application software provider, we observed that mapping an end-to-end sales pipeline process into two separate lead and opportunity models yields superior results than a single prediction model. When dealing with categorical features, we were also able to prove that the CatBoost algorithm is ideally suited, whereby the other results can also be used as a sophisticated benchmark for other sales pipeline applications. Furthermore, instead of only displaying the predictive performance, our artifact helps even salespeople to understand the ML based decision-making process with its explanation model by demonstrating the most relevant feature values. Above all, the applicability of the models requires no human expertise about the algorithm running in the background. By providing the individual prediction probabilities and the explanation overview, the model can be used intuitively by sales representatives without extensive training.

6. Limitations and Future Research

While we firmly believe that this research paper adds value to the current literature, our study is affected by some limitations and therefore offers opportunities for further research. First, the presented artifact should be tested on other case studies with similar complex sales pipelines to prove its suitability and usability in industry-wide situations. Second, in view of the mentioned interpretation capabilities, it would make sense to extend the explanation model from individual to overarching predictions. Instead of looking at the success rate of a particular lead or opportunity, finding clusters of feature values such as certain industries coupled with specific marketing campaigns can be crucial for determining positive sales indicators. Third, through the availability of a larger data set and the associated higher degree of complexity, we are striving to apply deep learning approaches to improve performance of sales pipeline models. However, it should be noted that deep learning models offer only limited interpretability of predictions due to their black-box character. Fourth, since in a license-driven industry greater accuracy has a significant impact on a company's profitability, further research must clearly focus on enhancing the predictive performance through other methods. Incorporating non-standard ML approaches could be necessary, for example, to address the problem of subjectivity and noisy labels caused by different regional sales pipeline procedures, diverging professional backgrounds and work experiences. The ability to learn with noisy labels is required if the data set could be biased due to a salesperson's behavior who systematically discontinues leads as soon as a certain feature value occurs. To give an example, a sales representative may intentionally discontinue a prospect that belongs to a certain industry. In addition, counterfactual inference is also seen as a non-standard ML approach that should be further investigated. The underlying idea is to establish an understanding about the behavior of complex systems interacting with their environment to better predict the consequences of system changes. As part of sales pipeline management, the selection of marketing campaigns is ideal for a counterfactual analysis as the personal network of a salesperson could act as a confounder that chooses the marketing campaign to address the prospect. Based on the available historical data further research could conduct an experiment to assess a customer's potential response to winning or losing a sales deal if a marketing campaign N had been replaced by N´. These proposed methods could significantly improve the prediction of the purchase probability of leads and opportunities.

7. References

[1] D. J. Teece, G. Pisano, and A. Shuen, "Dynamic capabilities and strategic management", Strategic Management Journal, 18(7), 1997, pp. 509-533.

[2] F. Arndt, L. Pierce, and D. Teece, "The Behavioral and Evolutionary Roots of Dynamic Capabilities", Industrial and Corporate Change, In press, 2017.

[3] S. Akter, S. F. Wamba, A. Gunasekaran, R. Dubey, and S. J. Childe, "How to improve firm performance using big data analytics capability and business strategy alignment?", International Journal of Production Economics, (182), 2016, pp. 113-131.

[4] A. Popovič, R. Hackney, R. Tassabehji, and M. Castelli, "The impact of big data analytics on firms' high value business performance", Information Systems Frontiers, In press, 2016.

[5] D. Nam, J. Le, and H. Lee, "Business analytics use in CRM: A nomological net from IT competence to CRM performance", International Journal of Information Management, In press, 2018.

[6] T. W. Jackson, "CRM: From "art to science"", Journal of Database Marketing & Customer Strategy Management, 13(1), 2005, pp. 76-92.

[7] F. Buttle, "Introduction to customer relationship management", in Customer relationship management: Concepts and technologies, F. Buttler and S. Maklan (eds.), London, Taylor & Francis, 2009a, pp. 1-23.

[8] E. W. T Ngai, L. Xiu, and D. C. K Chau, "Application of data mining techniques in customer relationship management: A literature review and classification", Expert Systems with Applications, 36(2), 2009, pp. 2592-2602.

[9] F. Sohnchen, and S. Albers, "Pipeline management for the acquisition of industrial projects", Industrial Marketing Management, 39(8), 2010, pp. 1356-1364.

[10] W. Reinartz, and V. Kumar, "The impact of customer relationship characteristics on profitable lifetime duration", Journal of Marketing, 67(1), 2003, pp.77-99.

[11] A. T. Jahromi, S. Stakhovych, and M. Ewing, "Managing B2B customer churn, retention, and profitability", Industrial Marketing Management, 43(7), 2014, pp.1258-1268.

[12] L. Ang, and F. Buttle, "Managing for successful customer acquisition: An exploration", Journal of Marketing Management, 22(3-4), 2006, pp. 295-317.

[13] F. Buttle, "Managing the customer lifecycle: customer acquisition", in Customer relationship management: Concepts and technologies, F. Buttler and S. Maklan (eds.), London, Taylor & Francis, 2009b, pp. 1-23.

[14] T. M. Smith, S. Gopalakrishna, R. Chatterjee, "A three-stage model of integrated marketing communications at the marketing-sales interface", Journal of Marketing Research, 43(4), 2006, pp.564-579.

[15] D. Lippold, Akquisitionszyklen und -prozesse im B2B-Bereich, Wiesbaden, Springer Gabler, 2016.

[16] F. Provost, and T. Fawcett, "Data science and its relationship to big data and data-driven decision making", Big Data, 1(1), 2013, pp. 51-59.

[17] K. Coussement, and D. Van den Poel, "Improving customer attrition prediction by integrating emotions

from client/company interaction emails and evaluating multiple classifiers", Expert Systems with Applications, 36(3), 2009, pp. 6127-6134.
[18] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction", Simulation Modelling Practice and Theory, (55), 2015, pp. 1-9.
[19] J. D'Haen, and D. Van den Poel, "Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework", Industrial Marketing Management, 42(4), 2013, pp. 544-551.
[20] J. Yan, C. Zhang, H. Zha, M. Gong, C. Sun, J. Huang, S. Chu, and X. Yang, "On Machine Learning towards Predictive Sales Pipeline Analytics", Proceedings of the 29th AAAI Conference on Artificial Intelligence, 2015, pp. 1945-1951.
[21] A. Megahed, P. Yin, and H. R. M. Nezhad, "An Optimization Approach to Services Sales Forecasting in A Multi-Staged Sales Pipeline", IEEE International Conference on Services Computing, 2016, pp. 713-719.
[22] K. Paulissen, K. Milis, M. Brengman, J. Fjermestad, and N. C. Romano, "Voids in the Current CRM Literature: Academic Literature Review and Classification (2000-2005)", Proceedings of the 40th HICSS, 2007.
[23] I. J. Chen, and K. Popovich, "Understanding customer relationship management (CRM): People, process and technology", Business Process Management Journal, 9(5), 2003, pp. 672-688.
[24] V. Kumar, and W. Reinartz, Customer Relationship Management: Concept, Strategy, and Tools, Berlin Heidelberg, Springer Verlag, 2012.
[25] J. F. Tanner, M Ahearne, T. W. Leigh, C. H. Mason, and W. C. Moncrief, "CRM in Sales-Intensive Organizations: A Review and Future Directions", Journal of Personal Selling & Sales Management, 25(2), 2005, pp. 169-180.
[26] M. Torggler, "The Functionality and Usage of CRM Systems", World Academy of Science, Engineering and Technology International Journal of Computer and Systems Engineering, 2(5), 2008, pp. 771-779.
[27] B. Kawas, M. S. Squillante, D. Subramanian, and K. R. Varshney, "Prescriptive Analytics for Allocating Sales Teams to Opportunities", IEEE13th Conference on Data Mining Workshops, 2013 ,pp. 211-218.
[28] J. P. Monat, "Industrial sales lead conversion modeling", Marketing Intelligence&Planning, (29), 2011, pp. 178-194.
[29] X. Xu, L. Tang, and V. Rangan, "Hitting your number or not? A robust & intelligent sales forecast system", IEEE International Conference on Big Data, 2017.
[30] K. Porter, "Why AI won't replace (great) salespeople",https://www.entrepreneur.com/article/292162, 2017, Retrieved on 21.02.2018.
[31] N. Syam, and A. Sharma, "Waiting for a sales renaissance in the fourth industrial revolution: Machine learning and artificial intelligence in sales research and practice", Industrial Marketing Management. 2018.
[32] N. Diakopoulos, "Algorithmic Accountability Reporting: On the Investigation of Black Boxes", Columbia University Academic, 2014, pp. 1-33.
[33] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms", Big Data & Society, 3(1), 2016, pp.1-12.
[34] M. Bohanec, M. K. Borštnar, and M. Robnik-Šikonja, "Explaining machine learning models in sales predictions", Expert Systems With Applications, 71(1), 2017, pp. 416-428.
[35] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers", IBM Journal of Research and Development, 3(3), 1959, pp. 210-229.
[36] T. M. Mitchell, Machine Learning, New York, McGraw-Hill Inc, 1997.
[37] S. J. Russell, and P. Norvig, Artificial Intelligence - A Modern Approach, New Jersey, Pearson Inc, 2010.
[38] K. P. Murphy, Machine Learning: A Probabilistic Perspective, Cambridge, The MIT Press, 2012.
[39] T. Hastie, R. Tibshirani, and J. Friedman, The elements of statistical learning: Data mining, inference and prediction, New York, Springer-Verlag, 2001
[40] L. Breiman, "Random Forests", Machine Learning, 45(1), 2001, pp. 5-32.
[41] C. Cortes, and C. Vapnik, "Support-vector networks", Machine Learning, 20(3), 1995, pp. 273-297.
[42] T. Chen, and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", preprint arXiv: 1603.02754, 2016.
[43] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", The Annals of Statistics, 29(5), 2001, pp. 1189-1232.
[44] A. V. Dorogush, V. Ershov, and A. Gulin, "CatBoost: gradient boosting with categorical features Support", Conference NIPS, 2017.
[45] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design Science in Information Systems Research", Management Information Systems Quarterly, 28(1), 2004, pp. 75–105.
[46] S. T. March, and G. F. Smith, "Design and Natural Science Research on Information Technology", Decision Support Systems, 15, 1995, pp. 251–266.
[47] H. Takeda, P. Veerkamp, T. Tomiyama, and Y. Hiroyuki, "Modeling Design Processes", AAAI, 11(5), 1990, pp. 37-48.
[48] K. Peffers, M. Rothenberger, T. Tuunanen, and R. Vaezi, "Design Science Research Evaluation", Design Science Research in Information Systems, Advances in Theory and Practice, 2012, pp. 398–410.
[49] J. M. Coe, "The integration of direct marketing and field sales to form a new B2B sales coverage model", Journal of Interactive Marketing, (18), 2004, pp. 62-77.
[50] C. W. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Technical Report, Department of Computer Science and Information Engineering, Taiwan University, 2004.
[51] J. Han, M. Kamber, and J. Pei, Data mining: Concepts and techniques, Burlington, Morgan Kaufmann, 2011.
[52] J. Davis, and M. Goadrich, "The relationship between precision-recall and roc curves," Proceedings of the 23rd international conference on Machine learning, 2006, pp. 233-240.
[53] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?" Explaining the Predictions of Any Classifier", arXiv:1602.04938, 2016.