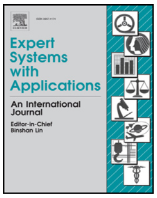
Expert Systems With Applications 71 (2017) 416–428

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Explaining machine learning models in sales predictions

Marko Bohanec[a,b,*], Mirjana Kljajić Borštnar[b], Marko Robnik-Šikonja[c]

[a] Salvirt Ltd., Dunajska cesta 136, Ljubljana 1000, Slovenia
[b] University of Maribor, Faculty of Organizational Sciences, Kidričeva cesta 55a, Kranj 4000, Slovenia
[c] University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, Ljubljana 1000, Slovenia

ARTICLE INFO

Article history:
Received 13 April 2016
Revised 19 October 2016
Accepted 5 November 2016
Available online 9 November 2016

Keywords:
Machine learning
Prediction explanation
Intelligent system
Black-box models
B2B Sales forecasting

ABSTRACT

A complexity of business dynamics often forces decision-makers to make decisions based on subjective mental models, reflecting their experience. However, research has shown that companies perform better when they apply data-driven decision-making. This creates an incentive to introduce intelligent, data-based decision models, which are comprehensive and support the interactive evaluation of decision options necessary for the business environment.

Recently, a new general explanation methodology has been proposed, which supports the explanation of state-of-the-art black-box prediction models. Uniform explanations are generated on the level of model/individual instance and support what-if analysis. We present a novel use of this methodology inside an intelligent system in a real-world case of business-to-business (B2B) sales forecasting, a complex task frequently done judgmentally. Users can validate their assumptions with the presented explanations and test their hypotheses using the presented what-if parallel graph representation. The results demonstrate effectiveness and usability of the methodology. A significant advantage of the presented method is the possibility to evaluate seller's actions and to outline general recommendations in sales strategy.

This flexibility of the approach and easy-to-follow explanations are suitable for many different applications. Our well-documented real-world case shows how to solve a decision support problem, namely that the best performing black-box models are inaccessible to human interaction and analysis. This could extend the use of the intelligent systems to areas where they were so far neglected due to their insistence on comprehensible models. A separation of the machine learning model selection from model explanation is another significant benefit for expert and intelligent systems. Explanations unconnected to a particular prediction model positively influence acceptance of new and complex models in the business environment through their easy assessment and switching.

1. Introduction

Data-driven decision-making is the practice of basing decisions on the data analysis, rather than on intuition (Provost & Fawcett, 2013). In the use of data-driven decision-making, companies in the top third of their industry are, on average, 5% more productive and 6% more profitable than their competitors (Brynjolfsson, Hitt, & Kim, 2011; McAfee & Brynjolfsson, 2012). The research into the acceptance of decision support systems (DSS) shows that users are more likely to adhere to recommendations when an explanation facility is available (Arnold, Clark, Collier, Leech, & Sutton, 2006; Gönül, Önkal, & Lawrence, 2006). However, the top performing, non-transparent black-box machine learning (ML) models, such as random forests, boosting, support vector machines (SVM), and neural networks achieve significantly better predictive performance than simple, interpretable models such as decision trees, naïve Bayes, or decision rules (Caruana & Niculescu-Mizil, 2006). This is one of the reasons for low usage and acceptance of predictive ML models in areas where transparency and comprehensibility of decisions are required.

This motivated the authors of the present paper, which focuses on how to address users' needs in a complex business environment and effectively explain state-of-the-art, incomprehensible ML predictive models, and their recommendations. Our goal is to enable business users to apply top performing (black-box) ML models of their choice and to obtain comprehensive interactive explanations in a uniform representation, regardless of the chosen model. Such an approach enables decision makers to deal with different decision-making tasks in a similar way and select the best performing ML model. This paper introduces a novel approach to building an intelligent system, which combines top-performing

[*] Corresponding author.
E-mail addresses: marko.bohanec@salvirt.com (M. Bohanec), mirjana.kljajic@fov.uni-mb.si (M. Kljajić Borštnar), marko.robnik@fri.uni-lj.si (M. Robnik-Šikonja).

http://dx.doi.org/10.1016/j.eswa.2016.11.010
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

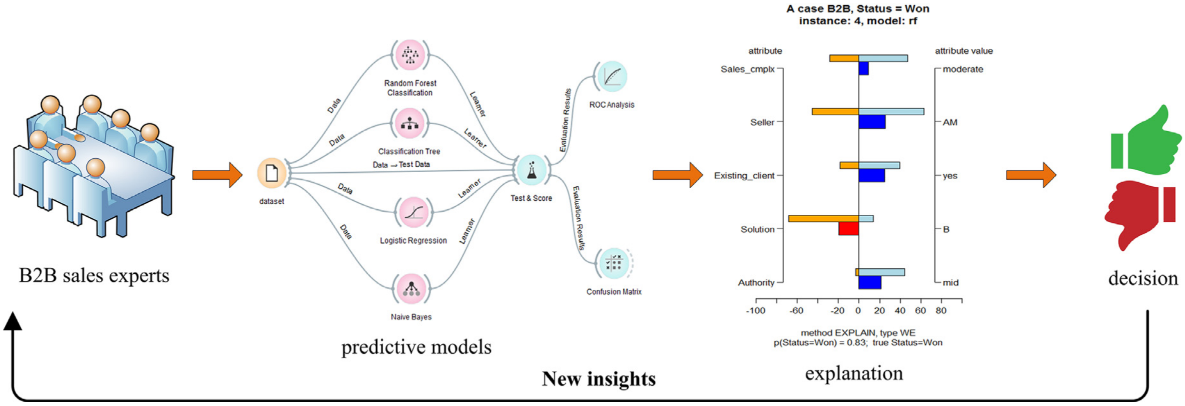


Fig. 1. High-level overview of presented intelligent system.

models, general explanation methodology and human consultants which initially help to introduce the methodology and overcome resistance to changes. In such a way top performing ML models can be introduced to domains which require comprehensibility and might currently use lower performing but interpretable models, such as decision trees or classification rules.

Recently, two general methods for explaining classification models and their predictions have been introduced (Robnik-Šikonja & Kononenko, 2008; Štrumbelj & Kononenko, 2010). Both methods are based on the idea that the importance of an attribute or a group of attributes in a specific model can be estimated by simulating the lack of knowledge about the values of the attribute(s). The methods, therefore, contrast a model's output using all attributes with the output obtained using only a subset of attributes. The methods output is a decomposition of the ML models' predictions into the individual contributions of attributes. The generated explanations closely follow the learned model and enable the visualization of the decision of each instance separately. As discussed by Robnik-Šikonja and Kononenko (2008), there are two levels of explanation: the *domain level* and the *model level*. The domain level would give the true causal relationship between the dependent and independent variables and is unreachable unless in regard to artificial problems in which relations and probability distributions are known in advance. The model level explanation aims to make transparent the prediction process of a particular model. This level of explanation is feasible for both human mental models and machine-learned models. Unsurprisingly, a strong correlation between the performance of the ML models and the similarity between the model level and domain level explanations was observed on artificial data, i.e., for well performing models, the explanation methodology produces explanations that are close to true causal dependencies.

To demonstrate the power and usability of the explanation methodology, we present a challenging real-world case of B2B sales forecasting, frequently done judgmentally. A high-level business process leveraging the proposed intelligent prediction system is presented in Fig. 1. A group of sales experts is collecting historical B2B sales cases with known outcomes to support them in a prediction task for new sales opportunities. The collected data is processed by various ML techniques in the next step, resulting in the statistically validated prediction model. The explanation methodology provides explanations for the past and new cases and enables a cognitive evaluation of the model by the users. Based on the new insights, they have an opportunity to update the data set, retrain and re-evaluate the models before the next use, practicing human-in-the-loop in ML (Holzinger, 2016). As we are dealing with a real-world problem, its true structure and causal dependencies are unknown; therefore, we model them using ML classifiers whose performance we evaluate with standard statistical measures (classification accuracy and AUC). The good performance of the models obtained gives credibility to the generated model level explanations. This is analogous to the model validation approach in predictive ML, in which the generalization performance of models is assessed through statistical evaluation.

The rest of the paper is organized as follows. Section 2 gives an overview of related work. Section 3 describes the explanation methodology and introduces an explanatory toy example. A real world use-case is presented in Section 4. Conclusions are put forward in Section 5.

2. Related work

This paper combines different research fields. We first review related work on explanation methods for ML models, followed by ML applications in DSS and discussion of the B2B knowledge gap. We outline research in sales forecasting and end the review with a description of selected ML applications in B2B sales.

2.1. Explanation of ML models

We first present criteria for the evaluation of model explanation methods, followed by several existing explanation techniques, which we compare with the two methods used in this work.

Andrews, Diederich, and Tickle (1995) and Jacobsson (2005) introduced a taxonomy and the criteria for evaluation of rule extraction methods from artificial neural networks (ANN). Despite its limited original scope, the taxonomy reveals significant aspects of explanation methods, so we extend it here to cover arbitrary ML models and then classify the methods we use. According to the taxonomy, explanation techniques can be classified according to:

1. *expressive power* describing the language of extracted knowledge: propositional logic (i.e., if-then rules), nomograms, non-conventional logic (e.g., fuzzy logic), first-order logic, finite state machines (deterministic, non-deterministic, stochastic);
2. *translucency* describing the degree to which an explanation method looks inside the model: decompositional (using model-specific representation, e.g., rules extracted from individual neurons of ANN), pedagogical (treating the model as a black box), and eclectic methods (combining both compositional and pedagogical types);
3. *portability* describes how well the explanation technique covers the plethora of currently available ML models;
4. *quality* of the extracted explanation, which can be further split into several criteria: accuracy, generalization ability; only for explanations which can be used in a predictive setting, e.g., de-

cision rules), fidelity (how well the explanation reflects behavior of the method), consistency (similarity of explanations for different models trained on the same task), and comprehensibility (readability and size of the explanation);

5. *algorithmic complexity of computing explanations.*

The two explanation methods we use, EXPLAIN (Robnik-Šikonja & Kononenko, 2008) and IME (Štrumbelj & Kononenko, 2010), exhibit the following properties according to the above criteria.

1. Expressive power: extracted explanations are in the form of contributions of individual attributes, which speak in favor/against the given class; these contributions can be seen as a limited form of propositional logic.
2. Translucency: both methods are pedagogical (the model is treated as a black-box, only the causal relationship between input and output is considered).
3. Portability: the approaches are general and applicable to an arbitrary ML model; its uniformity does not require users to be repeatedly trained.
4. Quality of extracted knowledge: while accuracy and consistency are not appropriate criteria for the methods used (accuracy can be applied to, e.g., rule-based methods), the produced explanations exhibit high fidelity as they are closely following the learned model (Robnik-Šikonja & Kononenko, 2008). Concerning comprehensibility, the methods EXPLAIN and IME can be considered very good due to their simplicity and graphical visualization.
5. Algorithmic complexity: If a is a number of attributes, then for a single instance and for a particular model we need $O(a)$ classifications with the model for the method EXPLAIN (one for each attribute). For the method IME, which uses sampling over the set of attributes, we also require $O(a)$ classifications but with a larger constant factor. In practice, the time required to generate individual explanations is negligible for EXPLAIN, and for IME it depends on the quality of requested approximation, which is a parameter of the implementation.

Below, we discuss a selection of model explanation methods. While simple symbolic prediction models, such as decision trees and decision rules, are self-explanatory if they are small enough, more complex models require either model-specific explanation techniques or general explanation methods.

Neural networks are examples of completely black box models. Due to their early use and good predictive capabilities, they received significant attention from researchers attempting to make this type of models more transparent. A review of explanation techniques for neural models is given in Jacobsson (2005). Typical approaches extract rules or decision trees from trained neural networks.

Some less complex non-symbolic models enable the explanation of their decisions in the form of weights associated with each attribute. Weight can be interpreted as the proportion of the information contributed by the corresponding attribute value to the final prediction. Such explanations can be easily visualized. For logistic regression, a well-known approach is to use nomograms, first proposed by Lubsen, Pool, and van der Does (1978). In Jakulin, Možina, Demšar, Bratko, and Zupan (2005), nomograms were developed for SVM, but they work only for a restricted class of kernels and cannot be used for general non-linear kernels. SVMs can also be visualized using projections into lower dimensional subspaces (Caragea, Cook, & Dianne Honavar, 2003; Poulet, 2004) or using self-organizing maps from unsupervised learning (Hamel, 2006). These methods concentrate on visualization of the separating hyperplane and decision surface and do not provide explanations for individual decisions. The Naive Bayesian (NB) classifier can explain its decisions as the sum of information gains (Kononenko, 1993). A straightforward visualization of NB was used in Becker, Kohavi, and Sommereld (1997), while in Možina, Demšar, Kattan, and Zupan (2004) nomograms were developed for the visualization of NB decisions. In the EXPLAIN method we use, the NB list of information gains was generalized to a list of attribute weights for any prediction model. The IME method uses the same explanation format as the EXPLAIN.

Madigan, Mosurski, and Almond (1997) considered the case of belief networks and used each (binary or multi-valued discrete) attribute as a node in the graph. By computing “evidence flows” in the network, it is possible to explain its decisions. In the ExplainD framework (Poulin et al., 2006), weights of evidence and visualizations similar to the EXPLAIN method are used, but the explanation approach is limited to (linear) additive models, while EXPLAIN and IME can be used for all probabilistic models.

Visualization of decision boundaries is an important aspect of model transparency. Barbosa et al. (2016) presented a technique to visualize how the kernel embeds data into a high-dimensional attribute space. With their Kelp method, they visualize how kernel choice affects neighborhood structure and SVM decision boundaries. Schulz, Gisbrecht, and Hammer (2015) proposed a general framework for visualization of classifiers via dimensionality reduction. Goldstein, Kapelner, Bleich, and Pitkin (2015) presented another useful visualization tool for classifiers that can produce individual conditional expectation plots, graphing the functional relationship between the predicted response and the attribute for individual instance.

In the context of attribute subset selection, attributes are evaluated in Lemaire and Clérot (2004) as the difference between the correct and perturbed output, which is similar to the EXPLAIN approach for a model level explanation. Lemaire, Féraud, and Voisine (2008) extended their approach to instance level explanations and applied it to a customer relationship management system in the telecommunications industry. Both, the EXPLAIN and the method of Lemaire are limited as they cannot detect disjunctive concepts and redundancies in the model. These deficiencies are ameliorated in the IME method as described in Section 3.2.

The IME method has been successfully applied to real-world medical problems (Štrumbelj, Bosnić, Kononenko, Zakotnik, & Kuhar, 2010) and to assess the impact of enterprises’ organizational quality on their financial results (Pregeljc, Štrumbelj, Mihelcic, & Kononenko, 2012); however, it has not been tested in the context of expert or intelligent systems and business problems, where interactive use and what-if analysis is needed.

The method that successfully deals with high-dimensional text data is presented in Martens and Provost (2011). Its idea is based on general explanation methods EXPLAIN and IME and offers explanations in the form of a set of words that would change the predicted class of a given document. Bosnić et al. (2014) adapt the same general explanation methodology to data stream scenarios and show the evolution of attribute contributions through time. This is used to explain the concept drift in their incremental model.

All presented explanations are related to statistical sensitivity analysis and uncertainty analysis (Saltelli, Chan, & Scott, 2000). In that methodology, the sensitivity of models is analyzed with respect to the models’ input, which is called “model-level explanation” in this paper. The presented visualizations of averaged explanations can, therefore, be viewed as a form of sensitivity analysis. In a related sensitivity based approach, called inverse classification, Aggarwal, Chen, and Han (2010) try to determine the minimum required change to a data point in order to reclassify it as a member of a different class. A SVM model based approach is proposed by Barbella et al. (2009). Another sensitivity analysis-based approach explains contributions of individual attributes to a particular classification by observing (partial) derivatives of the classifiers prediction function at the point of interest (Baehrens et al., 2010). A limitation of this approach is that the classification function has to be first-order differentiable. For classifiers not satisfying this criterion (for example, decision trees) the original classifier is first fitted with a Parzen window-based classifier that mimics the original one and then the explanation method is applied to this fitted classifier. The method was shown to be practically useful with a kernel-based classification method to predict molecular attributes (Hansen, Baehrens, Schroeter, Rupp, & Müller, 2011).

2.2. The application of ML in DSS and B2B sales forecasting

In a recent survey of the application of ML in DSS, Merkert, Mueller, and Hubl (2015) argued that the factors of ML design affecting the usefulness of decision support are still understudied. Their findings, based on 52 relevant papers from 1993 to 2013, suggested that ML usefulness depends on the task, the phase of decision-making and the applied technologies. The first two phases of the decision-making process described by Simon (1960), namely intelligence and design, are mostly supported by ML, using different prediction techniques (e.g., SVM, NN), while the third phase, choice, is less supported. They recommend that future research focus on organizational and people-related evaluation criteria. Our methodology supports all three phases of Simon's model, for example, choice is supported with what-if analysis.

An example of ML approach to improving dynamic decision-making is introduced by Meyer et al. (2014). The approach of PROCEDO (PRediction of Control Errors in Dynamic Contexts) attempts to improve feedback control strategies, which guide decision-making in a complex dynamic context. ML is used to identify possible improvement areas to prevent undesired or suboptimal outcomes. They validated their approach on a medical and two manufacturing use cases. To secure user acceptance in a real-world dynamic decision context, the choice of ML methods was limited by the requirement of outcome interpretability. The authors, therefore, chose decision trees and as future work proposed to use more powerful predictive models. Similarly, Florez-Lopez and Ramon-Jeronimo (2015) combined decision trees with more powerful ML techniques in a business environment, proposing the correlated-adjusted decision forest (CADF) to produce both accurate and comprehensible models. Our methodology does not need combinations of weaker and stronger models to assure their comprehensibility. With our approach, a choice of a suitable ML model is driven by the nature of problems and can be supported with the statistical model selection, as interpretability is secured by the general explanation methodology.

Lilien (2016) observed that B2B problems receive much less research attention than B2C problems do, and attributes this deficit to difficulties in understanding the specifics of the B2B field. Complexity and heterogeneity in the problem domain, the fact that data for B2B research are less voluminous and more difficult to collect than the data from consumer sources are, and lack of domain knowledge on the part of researchers are among the contributors to this gap. This author also perceived great potential in B2B customer analytics to address business problems, but a lack of tools and guidance currently prevent the realization of that potential. With our work, we aim to support B2B practitioners and academic researchers with a documented real-world application using ML models coupled with explanations. We also make our B2B sales forecasting data set publicly available to spur further research.

Armstrong, Green, and Graefe (2015) reviewed the academic work in the field of sales forecasting and concluded that due to sophisticated statistical procedures and despite major advances in forecasting methods, the forecasting practice has seen little improvement. Our paper addresses this concern by presenting a practical guideline on how to use data-based ML models whose decisions are presented in a simple and uniform way in order to reduce subjective biases in sales forecasting.

Yan et al. (2015) proposed a win-propensity score for a given future time window based on sellers' interaction with the system and personalized profiles for new sales opportunities, showing that their ML method outperforms subjective ratings. When sellers were provided with scorings for their resource allocation decision, their results improved, indicating the regenerative effect between prediction and action. D'Haen and Van der Poel (2013) proposed an iterative three-phased automated ML model designed to help acquire clients in a B2B environment. Their goal is to generate a high-quality list of prospective clients, which are likely to become sales opportunities and ultimately clients; they emphasize a need for an extensive documentation on decisions made, steps taken, etc., to incrementally improve client acquisition. We move beyond numeric win-propensity score to further improve results, offering context-dependent guidance to develop sales leads from an opportunity to an established client.

In previous work (Bohanec, Kljajić Borštnar, & Robnik-Šikonja, 2016), we used several ML techniques and used their native explanation mechanisms to better understand the task of B2B sales forecasting. We use different visualizations (e.g., rules, trees, parallel coordinates), but do not use the EXPLAIN and IME general explanation methods, and do not explain individual predictions of ML models. The data set presented here is also novel and significantly improved over previous versions.

In short, in contrast to the existing work we contribute to the application of ML in DSS by addressing all three decision-making phases (Merkert et al., 2015), removing the limitation that only weaker interpretable models, such as decision trees, can be used (Florez-Lopez & Ramon-Jeronimo, 2015; Meyer et al., 2014), introducing a real-world application of ML models to B2B practitioners (Lilien, 2016), allowing reduction of personal bias in sales forecasting (Armstrong et al., 2015), moving beyond the win-propensity score (Yan et al., 2015), and supporting sellers to develop sales leads from the opportunities to established clients (D'Haen & Van der Poel, 2013).

3. Explanation methodology

The general explanation methods EXPLAIN (Robnik-Šikonja & Kononenko, 2008) and IME (Štrumbelj, Kononenko, & Robnik-Šikonja, 2009) can be applied to any prediction model, which makes them a useful tool both for interpreting models and comparing different types of models. The key technique they use is sensitivity analysis: changing the inputs of the model and observing changes in the model's output.

The basic idea is that the contribution of a particular input value (set of values) can be captured by "hiding" the input value (set of values) and observing how the output changes. As such, the key component of general explanation methods is the expected conditional prediction: the prediction for which only a subset of the input variables is known. Let Q be a subset of the set of input variables $Q \subseteq S = \{X_1, \dots, X_a\}$. Let $p_Q(y_k|x)$ be the expected prediction for class value y_k given the vector of independent variables x, conditional to knowing only the input variables represented in Q:

$$p_Q(y_k|x) = \mathbb{E}(p(y_k)|X_i = x_{(i)}, \forall X_i \in Q). \tag{1}$$

Therefore, $p_S(y_k|x) = p(y_k|x)$ (a prediction using all variables).

Computing an input attribute's contribution by using the expected conditional prediction is common to all general explanation methods. However, the methods differ in how many and which subsets of attributes they take into account and how they combine the conditional predictions. The two methods applied use the two extreme examples: (a) the EXPLAIN method computes an input attribute's contribution by omitting just that attribute, and

(b) the IME method considers all subsets of attributes. The former is more efficient but does not give desirable results in some cases.

For complex real-world problems and state-of-the-art models, the true classification function of the model is not known and, similarly to the relation between domain-level and model-level explanation, we can only obtain it for simple models and artificial problems with known dependencies. In practice, we can only compute the predictions of classification functions for any vector of input values, and thereby obtain a suitable approximation. Therefore, for practical problems, the exact computation of Eq. (1) is not possible and sampling-based approximations are presented below.

3.1. The EXPLAIN method

The straightforward characterization of the ith input variable's importance for the prediction of instance x is the difference between the model's prediction for that instance and the model's prediction if the value of the ith variable is not known: $p(y_k|x) - p_{S\setminus\{i\}}(y_k|x)$. If this difference is large, then the ith variable is important. If it is small, then the variable is less important. The sign of the difference reveals whether the value contributes towards or against class value y_k.

To improve human comprehensibility, in Robnik-Šikonja and Kononenko (2008) the difference in information or the difference in log-odds ratios (i.e., weight of evidence (Good, 1950)) is proposed instead of the difference in probabilities. In this paper, we adopt the log-odds ratio approach:

$$\mathrm{WE}_i(k,x) = \log_2\left(\frac{p(y_k|x)}{1-p(y_k|x)}\right) - \log_2\left(\frac{p_{S\setminus\{i\}}(y_k|x)}{1-p_{S\setminus\{i\}}(y_k|x)}\right)\ \text{[bits]}. \tag{2}$$

In general, the expected conditional prediction $p_{S\setminus\{i\}}(y_k|x)$ in Eq. (2) cannot be computed, so an approximation is proposed, based on perturbing the ith attributes values and computing the weighted sum of perturbations, with each element weighted according to the probability of that value.

3.2. The IME method

The main disadvantage of the one-variable-at-a-time approaches is that disjunctive concepts and other redundancies captured by the model might result in unintuitive contributions. Therefore, in certain situations, a method that considers such interactions is more desirable. For example, observe the simple disjunction of two binary variables $1\vee 1$. Changing either of the two values to 0 would not change the value of the expression and both variables would be assigned a zero importance. Both have to be changed simultaneously to observe a change in the value of the expression and assign a non-zero contribution to the variables, which are clearly important.

A solution was proposed in Štrumbelj and Kononenko (2010), which, if used without sampling, requires 2^a steps and results in an exponential time complexity:

$$\varphi_i(k,x) = \sum_{Q\subseteq\{1,2,\ldots,a\}\setminus\{i\}} \frac{1}{a\binom{a-1}{a-|Q|-1}}\left(\Delta(Q\cup\{i\})(k,x) - \Delta(Q)(k,x)\right), \tag{3}$$

where $\Delta(Q)(k,x) = p_Q(y_k|x) - p_\varnothing(y_k|x)$.

For a larger number of attributes a, the computation of Eq. (3) becomes infeasible, so an alternative formulation was used:

$$\begin{aligned}\varphi_i(k,x) &= \frac{1}{a!}\sum_{\mathcal{O}\in\pi(a)}\left(\Delta(Pre^i(\mathcal{O})\cup\{i\})(k,x) - \Delta(Pre^i(\mathcal{O}))(k,x)\right)\\ &= \frac{1}{a!}\sum_{\mathcal{O}\in\pi(a)}\left(p_{Pre^i(\mathcal{O})\cup\{i\}}(y_k|x) - p_{Pre^i(\mathcal{O})}(y_k|x)\right),\end{aligned} \tag{4}$$

where $\pi(a)$ is the set of all permutations of a elements and $Pre^i(\mathcal{O})$ is the set of all input variables that precede the ith variable in the permutation $\mathcal{O}\in\pi(a)$. This approach can be efficiently implemented and is called the IME method.

The method iteratively samples the space of attribute combinations (one iteration for each attribute set permutation $\mathcal{O}$). In each iteration, a part of attribute values is randomly generated (the attributes in $Pre^i(\mathcal{O})$) and the remaining attribute values are taken from the original instance x. The difference in prediction using either the original value or a randomly generated value for the ith attribute is evidence for the importance of an ith attribute in the interactions with other attributes. Details of this estimation procedure and its convergence can be found in Štrumbelj and Kononenko (2010).

The problem of assigning credit to individual attributes can be viewed from the point of coalitional game theory. In this context, Eqs. (3) and (4) represent a Shapley value for the coalitional game of a players with Δ as the characteristic function. The contributions to individual attributes are, therefore, fair according to all interactions in which they are taking part (Štrumbelj & Kononenko, 2010).

3.3. Practical details

In principle, the predictive performance and correctness of explanations at model level are independent of each other. However, empirical observations on both artificial and real-world data show that better models (with higher prediction performance) enable better explanations (Štrumbelj et al., 2009).

The generated model level explanations are of two types: *instance explanations* and *model explanations*. An instance explanation explains the prediction of a single instance with the given model. A model explanation is an average over instance explanations for many training instances. This averaging provides information about the impact of attributes and their values in the model. The averaging over many instances enables the identification of the different roles attributes play in the prediction model and classification of instances. To avoid loss of information, we collect and present evidence for and against each decision separately. In this way we can, for example, see that a particular value of an attribute supports the decision for a specific outcome for one instance but not for another.

Both presented explanations can treat numerical and discrete attributes, they handle missing values, and work in classification as well as regression settings. In generating instance explanations, for the EXPLAIN method, numeric attributes are discretized as a pre-processing step and centers of intervals are used as proxies. Several discretization techniques can be used; both class-blind methods, such as equal-frequency and equal-width approaches, as well as discretizations with attribute evaluation techniques, such as ReliefF (Robnik-Šikonja & Kononenko, 1995). Instead of discretization and discrete intervals, it would also be possible to use fuzzy approaches like the one described in Costea and Bleotu (2012). For the IME method, the handling of numeric values is natural. In Eq. (4), random numeric values from the domain of ith attribute are generated, and the difference with the actual value is assessed for a given random permutation of attributes. To obtain model explanations, for both methods the instance explanations are averaged. For numeric attributes, the explanations are displayed as sensible

Table 1
Description of attributes for a toy example (with value frequency in brackets).

Attribute	*Description*	*Values*
Authority	Authority level at a client side.	low(24), mid(37), high(39)
Solution	Which solution was offered?	A(51), B(49)
Existing_client	Selling to existing client?	no(47), yes(53)
Seller	Seller name (abbreviation).	RZ(35), BC(29), AM(36)
Sales_complexity	Complexity of sales process.	low(31), moderate(53), high(16)
Status	An outcome of sales opportunity.	lost(45), won(55)

intervals obtained with a given discretization. Further details can be found in Robnik-Šikonja and Kononenko (2008) for the EXPLAIN method and Štrumbelj and Kononenko (2010) for the IME method. The complete source code of the implementation is available as the R package ExplainPrediction (Robnik-Šikonja, 2015).

3.4. A toy example of the proposed methodology

We present a simple toy example to introduce the use of the proposed explanation methodology and B2B sales problem. The data set contains a basic description of B2B sales events of a fictional company.

Let us assume that the company is selling two complex solutions, A and B, on B2B markets. Their primary sales audience is business managers, and a certain level of sales complexity is expected (e.g., several business units are involved, the alignment of their expectations is not ideal, etc.). The company is successfully selling their initial Solution A, but recently Solution B was added to the sales portfolio. Ideally, the company cross sells the Solution B to existing clients. The sales personnel are not focusing their efforts on simple opportunities; rather they pursue deals in which they can offer complex solutions together with the company's deployment consultants. For a successful sale, the sales team attempts to engage senior managers at clients, with authority to secure funding and participate in the definition of requirements. Given its maturity, we expect sales of Solution A to be more successful than B. The company collects the data described in Table 1. For example, the attribute *Sales_complexity* has three values with the following meaning: *high* (different clients' business units have diverse expectations), *moderate* (only one unit is involved, but the expectations have yet to be clarified), and *low* (one unit with well-defined expectations).

The data set consists of 100 instances. We randomly take 80% of the instances as a training set, and the remaining 20% of the instances is used as a testing set. To build a classifier, we use the ensemble learning method Random Forest (RF) (Breiman, 2001), broadly recognized as a robust and well-performing ML method (Verikas, Gelzinis, & Bacauskiene, 2011). RF constructs a set of decision trees, called a forest, using a bootstrap sampling of training instances. The splits in tree nodes are selected from a small set of randomly selected attributes. During the classification, each tree votes independently. Votes are counted, and the majority decision is taken as a prediction for the given instance. Due to the complex structure of the RF model (hundreds of randomized trees) and the voting mechanism employed, RF can be considered a black-box model, as there is no simple way to explain its predictions.

We take the RF model as the input to the EXPLAIN or IME explanation methods. Fig. 2a introduces an example of an explanation for a specific case (the sales opportunity named *instance 4*), where the sale was discussed with mid-level managers at an existing client, seller AM offered Solution B and was experiencing moderate complexity in the sales effort. The left-hand side of Fig. 2a outlines the attributes, with the specific values for the selected instance on the right-hand side. For this instance, the probability returned by the model for the outcome "Status = Won" is 0.83, and "Won" is the true outcome of this instance. The impact of attributes on the outcome expressed as the weight of evidence (WE) is shown as horizontal bars. The length of bars corresponds to the impact of the attribute values on the outcome predicted by the model. Right-hand sidebars show positive impacts on the selected class value ("Status=Won" in this case), and left-hand side bars correspond to negative impacts. The thinner bars (light gray) above the explanation bars (dark gray) indicate the average impact overall training instances for a particular attribute value. For the given *instance 4* in Fig. 2a, we can observe that while "Solution=B" is not in favor of closing the deal, the rest of the attributes' values (in dark gray) are supportive of a positive outcome. The thinner bars show that on average, these values can support positive outcomes, but the opposite outcome is also possible.

To understand the problem on the level of the model, all explanations for the training data are combined. A visualization of the complete model showing all attributes and their values (separated by dashed lines) is shown in Fig. 2b. We see that the impact indicators for attributes (dark gray) spread across the horizontal axis, which indicates that both positive and negative impacts are possible for each attribute. The dark bars representing attributes are weighted averages of the impact of their values that are shown above them (light gray bars). For each attribute value, an average negative and positive impact is presented. Specific attribute values often contain more focused information than the whole attribute. For example, moderate sales complexity or dealing with mid-level managers indicate a stronger tendency towards positive outcomes than towards negative outcomes. The specific value "yes" for the attribute *Existing_client* has a prevailing positive impact on the positive outcome. However, the value "no" can also have a positive impact. Note also the scale of the horizontal axis on Figs. 2a and b. While on Fig. 2b original values of WE are shown, we normalized the sum of contributions to 100 in Fig. 2a. Such normalization can be useful if we compare several decision models or if we want to assess the impact of attributes in terms of percentages.

In practice, sellers are interested in explanations of forecasts for new (open) cases, for which the outcome is still unknown. Fig. 3a visualizes an explanation of such a case. The predicted probability of a successful sale is 0.72. The attribute values of the new case are printed on the right-hand side of the graph. The explanation reveals a weak negative influence of the fact that the sale is not discussed with an existing client. The *Seller* AM (thin bars indicate his large variation in sales performance) seems to have a marginal positive impact in this case. The attributes *Solution*, *Sales_complexity*, and *Authority* contribute towards a positive outcome.

Assume that this sales opportunity is critical for the company. The relatively low predicted probability triggers a discussion about the actions needed to enhance the likelihood of winning the contract. Some attributes cannot be changed (e.g., *Existing_client* ="no" is a fact). However, what effect on the prediction would a change of seller from "AM" to "BC" cause, with all other attribute values left the same? Fig. 3b shows the implications of this change. The likelihood of winning the deal rises to 0.92. The explanation bars indicate the strong positive impacts of all the attribute values.

Equipped with such an insight provided by the introduced explanation methodology as part of an intelligent system, decision makers have better grounds to make analytically supported and transparent decisions and to discuss possible alternatives. In addition, they can challenge their prior assumptions.

4. Using explanations in a real-world business example

The toy example in the previous section introduced the explanation methodology and the B2B sales forecasting domain. In this

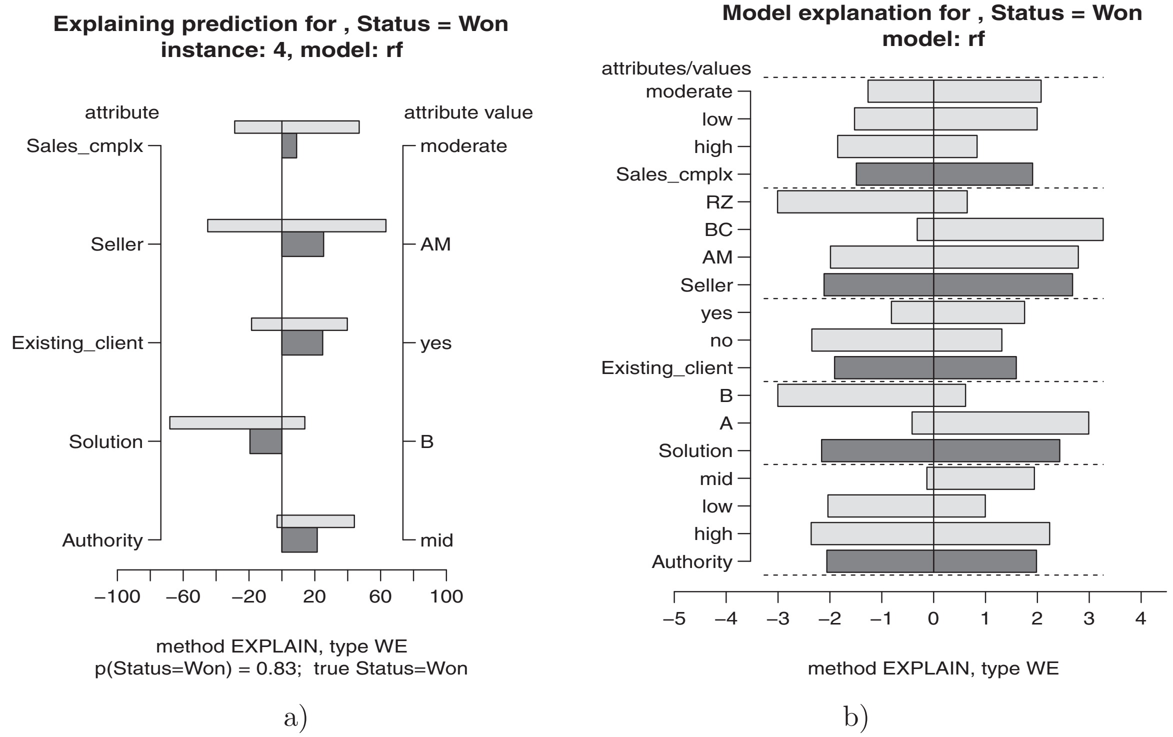


Fig. 2. The EXPLAIN-based explanation for a) an individual instance (normalized to 100) and b) the whole model.

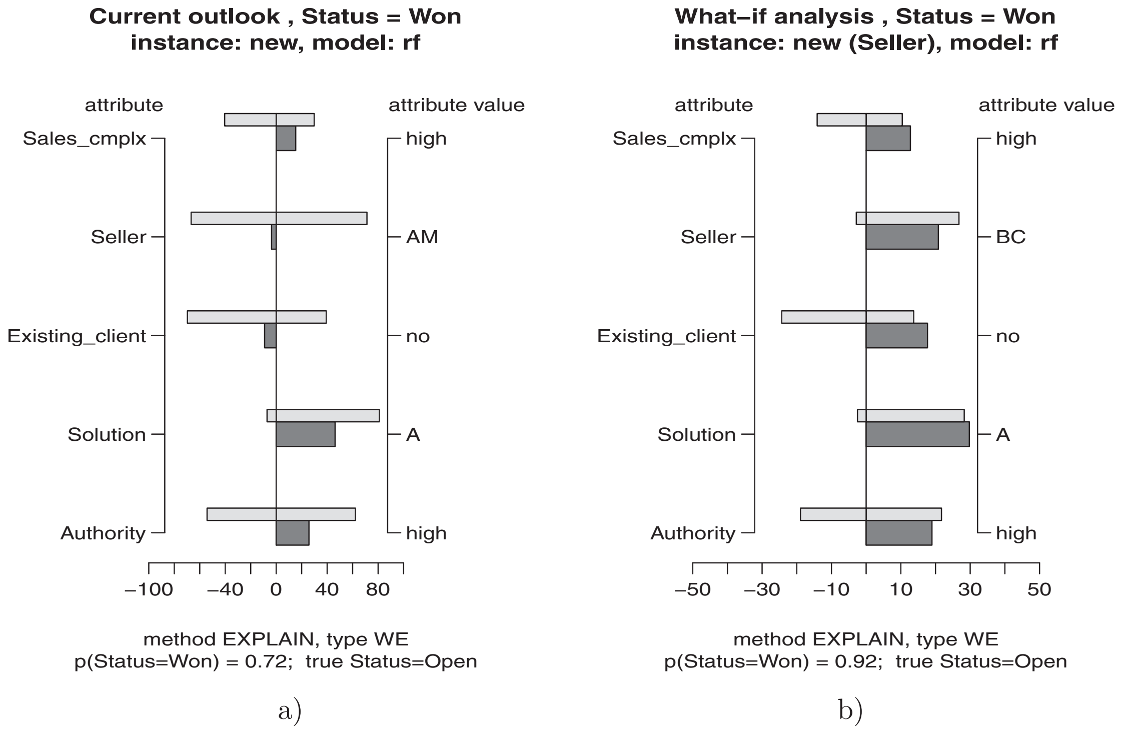


Fig. 3. Explanation for a new, open case (a) and its variation (b).

section, we describe the use of explanation techniques in a real-world case of a medium-sized company providing software solutions to clients in international B2B markets. Its internal analytical team constitutes a general manager, two sales leaders and a data scientist. The participation in the research is motivated by (a) improvements in understanding factors influencing the outcome of their sales process and (b) improving the sales performance for new clients. Approximately 15 new (open) sales opportunities per month are analyzed with the help of an external consultant, using the described methodology.

First, a set of essential attributes, describing characteristics of the B2B sales process was defined. The company's CRM system did not provide sufficient information to allow the application of ML techniques, so we proposed a list of attributes, describing the B2B

Table 2
Description of attributes for the real-world B2B sales forecasting study.

Attribute	*Description*	*Values*
Product	Offered product.	e.g. ERP, A, B, etc.
Seller	Seller's name.	Seller's name
Authority	Authority level at a client side.	Low, Mid, High
Company_size	Size of a company.	Big, Mid, Small
Competitors	Do we have competitors?	No, Yes, Unknown
Purchasing_department	Is the purchasing department involved?	No, Yes, Unknown
Partnership	Selling in partnership?	No, Yes
Budget_allocated	Did the client reserve the budget?	No, Yes, Unknown
Formal_tender	Is a tendering procedure required?	No, Yes
RFI	Did we get Request for Information?	No, Yes
RFP	Did we get Request for Proposal?	No, Yes
Growth	Growth of a client?	Fast Growth, Decline, etc.
Positive_statements	Positive attitude expressed?	No, Yes, Unknown
Source	Source of the opportunity.	e.g. Referral, Web, etc.
Client	Type of a client.	New, Current, Past
Scope_clarity	Implementation scope defined?	Clear, Few questions, etc.
Strategic_deal	Does this deal have a strategic value?	No, Yes, Normal
Cross_sale	Do we sell a different product to existing client?	No, Yes
Up_sale	Increasing existing products?	No, Yes
Deal type	Type of a sale.	Consulting, Project, etc.
Needs_defined	Is client clear in expressing the needs?	Info gathering, etc.
Attention_to_client	Attention to a client.	First deal, Normal, etc.
Status	An outcome of sales opportunity.	Lost, Won

sales process. The company added a few additional attributes relevant in their sales context. The consultant organized a workshop to develop a final list of attributes. Particular attention was dedicated to defining the meaning of each attribute and clarifying distinctions between the values of attributes. For example, the attribute *Authority* was defined as "a capacity of a head person in a client's team to define the scope, secure the budget, and officially authorize the order". The meaning of three possible values *low, mid,* and *high* for this attribute was discussed and agreed upon by the team. Finally, 22 attributes (+ 1 class variable), perceived by the team as important, were selected. The list is shown in Table 2 and does not include some standard fields from CRM (e.g., the *Opportunity ID* field contains identifiers that do not contribute to learning). Some other attributes, e.g. *Probability of closure* or *Sales stage*, were also omitted due to their strong correlation with the outcome and low explanation value. Such attributes contain subjective judgments based on individuals' mental models, which can be helpful for prediction but not for objective analysis of the problem and its understanding.

The company's CRM was updated with fields reflecting the chosen attributes and the company provided 448 cases (51% Won, 49% Lost) from its sales history, which constitute our machine learning data set. The data set is publicly available (Bohanec, 2016). We randomly took 80% of instances as a training set, and the remaining 20% of instances were used as a testing set. The process was repeated 30 times, and the results were averaged. To compare the performance of the ML models used, we used the classification accuracy (CA) and AUC (area under the receiver operator characteristic curve). The AUC describes the predictive behavior of a classifier independent of class distribution or error costs, so it decouples predictive performance from these factors. As in real-world data sets, the class distributions are not known in advance; Provost, Fawcett, and Kohavi (1998) showed that the AUC is a more appropriate measure than the CA in this context. We trained several prediction models and compared them according to AUC and CA. Table 3 shows the results.

As a final model, we selected the classifier known for its robust performance, Random Forest (RF), which achieved the best performance in both CA and AUC. We ran a Wilcoxon signed rank test for significance of the difference in AUC between two top performing ML methods: RF and NB. The differences are highly significant (*p=2.08e-05*). Additionally, we tested naïve Bayes (NB), decision tree (DT), support vector machine (SVM) and neural network (NN) classifiers, but omitted those results from further discussion.

Table 3
The CA and AUC average performance on the business data set.

ML model	*CA*	*AUC*
RF	0.782	0.85
NB	0.777	0.83
DT	0.742	0.76
NN	0.702	0.70
SVM	0.567	0.59

4.1. Model level explanations

We first present results and explanations for the model as a whole. This gives insight into the average behavior of the model and, therefore, reveals information extracted from the whole data set.

All the data set attributes are included in the trained model. To secure compact and comprehensible explanations, the methodology supports a user-defined threshold for minimal attribute impact, with which we can display only the most relevant attributes. For our data set, the model explanations of methods EXPLAIN and IME slightly differ, as shown in Fig. 4. As IME can capture attribute interactions involving disjunctive and redundant explanations, in the remainder of the paper we report only IME results. Nevertheless, both IME and EXPLAIN identify the same most influential B2B sales attributes: *Up_sale, Client,* and *Competitors*. To obtain more understanding about the contribution of attribute values, we can drill into a subset of attributes in the model explanations. Fig. 5a shows the average impact of values for the attributes *Competitors, Client, Up_sale* and *Posit_statm* as individual bars. For example, attribute *Competitors* with value "No" reveals a positive impact on the outcome vs. value "Yes", for which the impact is negative. If sellers do not know about the existence of competition (value "Unknown"), this negatively impacts the outcome, but less significantly. Fig. 5b shows two attributes, *Product* and *Seller*, with several values. Looking at the attribute *Product* as a whole in the

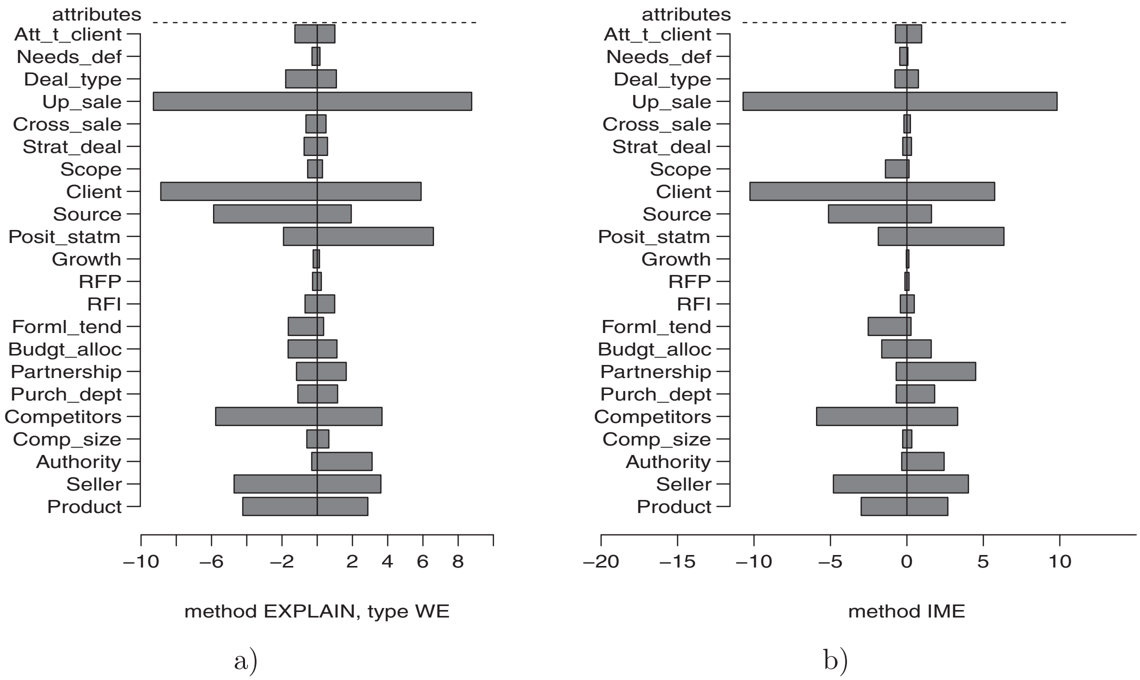


Fig. 4. Business case - EXPLAIN and IME model level explanations for RF model.

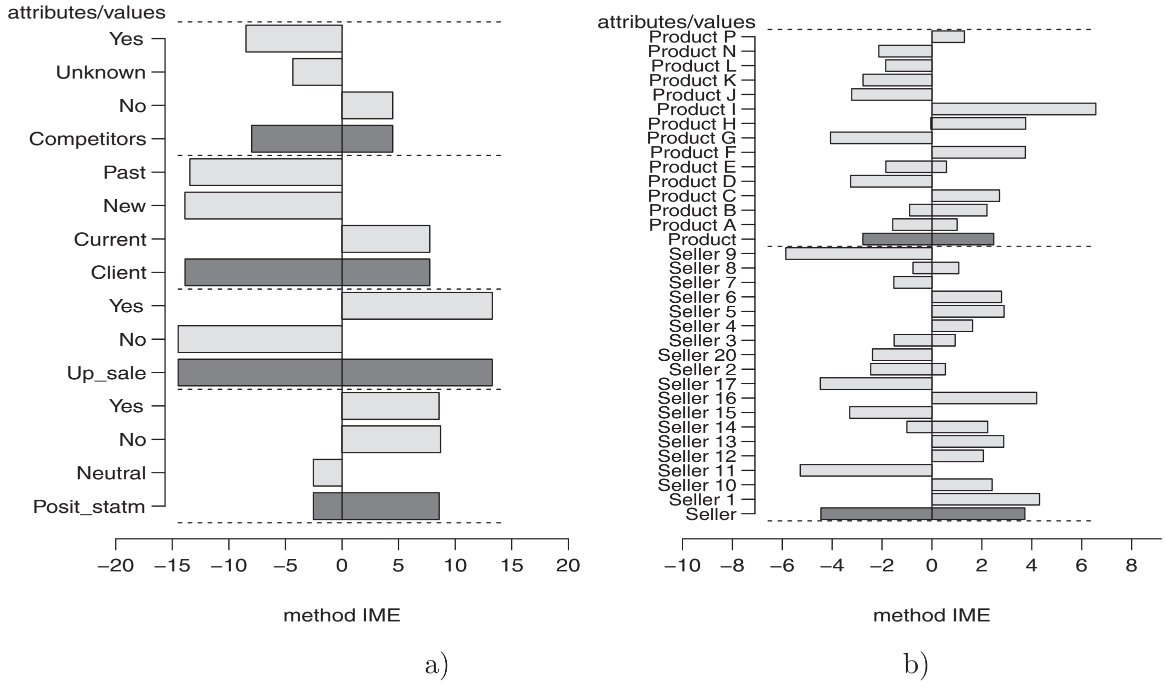


Fig. 5. Drilling into the model to visualize selected attributes and their values.

model explanation, we see that its positive and negative impact are approximately the same; therefore, one cannot draw a general conclusion about its impact. However, when looking at specific values, a better understanding can be formed. For example, products *P, I, H, F*, and *C* have a positive impact, while other products mostly do not contribute positively to the outcome. Products *A* and *B* have a mixed impact. Similar reasoning can be applied to the performance of sellers.

4.2. Supporting human learning and reflection

Sales teams often perform so-called “post-mortem” discussions with the purpose of uncovering what was done right and what did not go well for a particular deal. In this way, teams learn and potentially update their beliefs to understand where to focus in the future. The proposed explanation methodology can contribute to these discussions by making reasoning of ML models transparent. Participants can analyze the model, test their beliefs and exper-

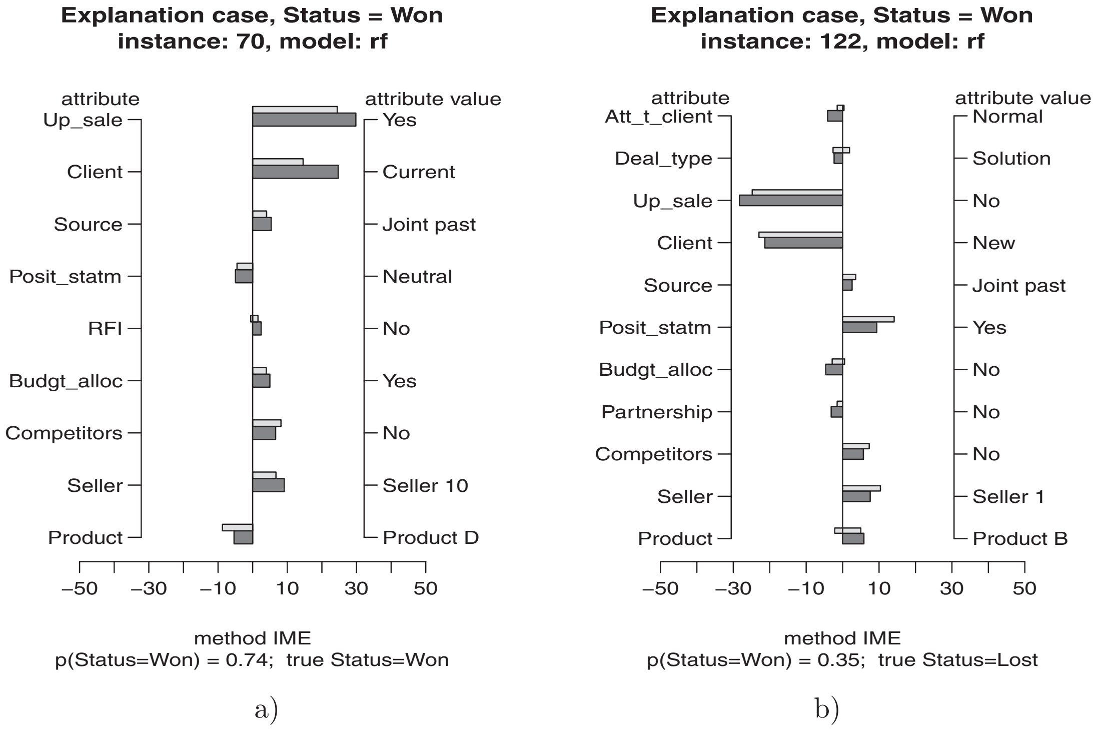


Fig. 6. Instance explanations for one Won and one Lost deal.

iment with different scenarios, all of which supports the notion of interpretability. Fig. 6 shows two cases of explanations, a) for a won deal and b) for a lost deal. Only the attributes with impacts larger than the minimum impact threshold (2) are shown. The explanation for a won deal shows that most of the attributes are positively impacting the outcome of the deal, i.e., the client is an existing one, it is buying up-sale, has a budget, there are no competitors, etc. Two attributes (*Product* and *Positive statements*) have a negative impact; however, their impact is not strong. Fig. 6b provides possible reasons for a negative outcome: a new client, no opportunity yet for up-sale, no allocated budget, and the absence of a partnership. Furthermore, seller's attention to the client was normal, without additional sales efforts. The favorable factors for a successful sale are “joint past” (i.e. worked or studied together), clearly expressed positive statements about the selling company for *Product B* and no competition. Finally, *Seller 1* has a positive impact; however, in the context of the deal, this was not enough to win.

In the presented business case, the CA of the RF model is 78.2%. Although this is a good score for the difficult sales forecasting domain, there are some missed predictions, which can lead to incorrect explanations. Fig. 7 shows two examples, a) the deal is predicted as Won, but it was actually Lost, and b) the deal is predicted as Lost, and in reality it was Won. This situation triggers two types of discussions. First, do all attribute values correctly reflect a state of an opportunity? If not, the values should be corrected and the model re-run. Second (provided the values are correct), what information is missing, can we provide it in the form of additional attributes and include it in the model? These questions require the sales team to re-assess the opportunity, reflect on its context, identify potential new attributes or rethink (redesign) the values of certain attributes. A guided workshop can initially facilitate such an analysis and establish the human-in-the-loop process for iterative refinements of ML models. If the sales team finds a new operational attribute that might significantly affect the outcome, this attribute should be added to the attribute list. This is a part of the model review and contributes to better understanding of the problem and data. It also stimulates refinement of the B2B sales domain expertise.

To demonstrate a what-if analysis, Fig. 8a shows the status of a new opportunity, with values of all 22 attributes shown (threshold = 0). During the team's discussion, it was observed that value of the attribute *Partnership* was recorded incorrectly and should be corrected to “Yes”. Furthermore, according to the new information provided by *Seller 1*, the value for budget allocation attribute should be updated to “Yes”. The effect of these updates is visible in Fig. 8b, where the likelihood of a successful outcome increases from 0.29 to 0.52. We show only the most relevant attributes (Threshold 2).

The participating company wanted to know how to address a slowdown in the acquisition of new clients. To respond to this request, from the initial business data set, only instances related to new clients were selected (158 instances). This new data set was assessed with the same approach as the initial business data set. By selecting only instances involving new clients, we intentionally biased the learning outcome. The resulting model and its explanations are not generally applicable but can help us to distinguish successful and unsuccessful deals involving new clients. The model explanation is presented in Fig. 9 with a threshold (4) applied to discard values with marginal impact. The strongest positive impact comes from the attribute *Partnership* with value “Yes”, which indicates a recommendation to form a partnership with companies when bidding for new business. When a sales opportunity comes from the participation at an event (e.g., conference presentation), then a positive impact is also secured. Among *Product* values, I, F and C are indicators of the most positive impact, while value D has the strongest negative impact. The rest of the values have impact below the threshold. Such a compact view enables a more focused discussion when building a company's sales strategy.

5. Conclusions

This study demonstrates a novel application of a general explanation methodology for ML models to a complex real-world

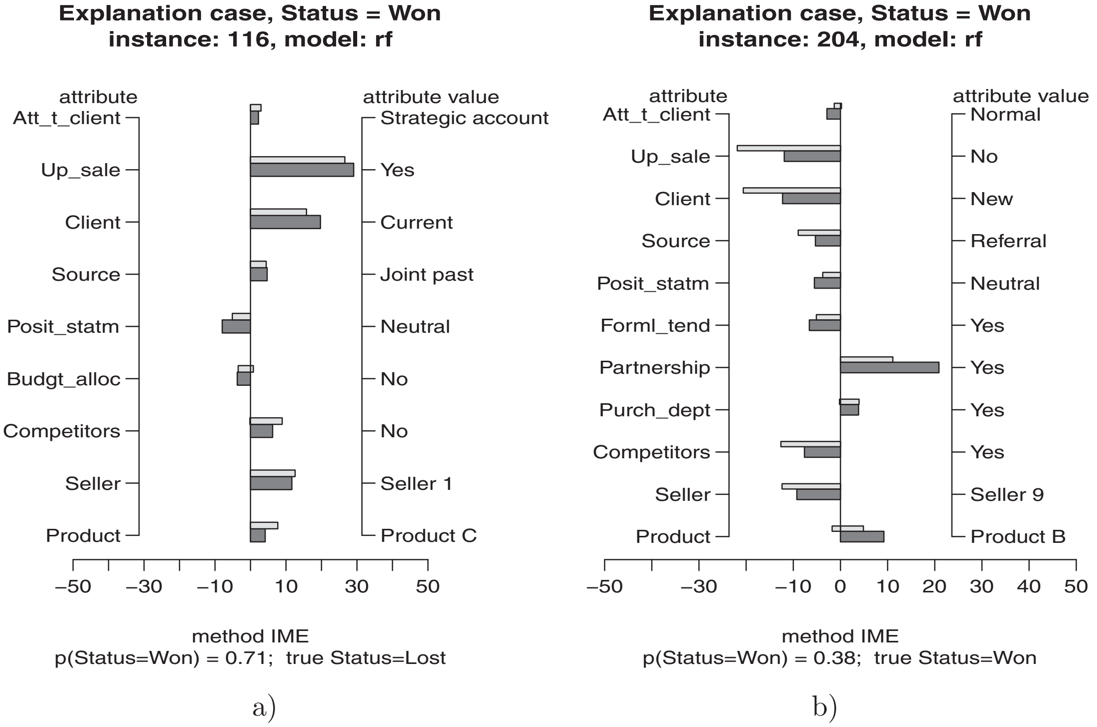


Fig. 7. Two examples of wrong predictions.

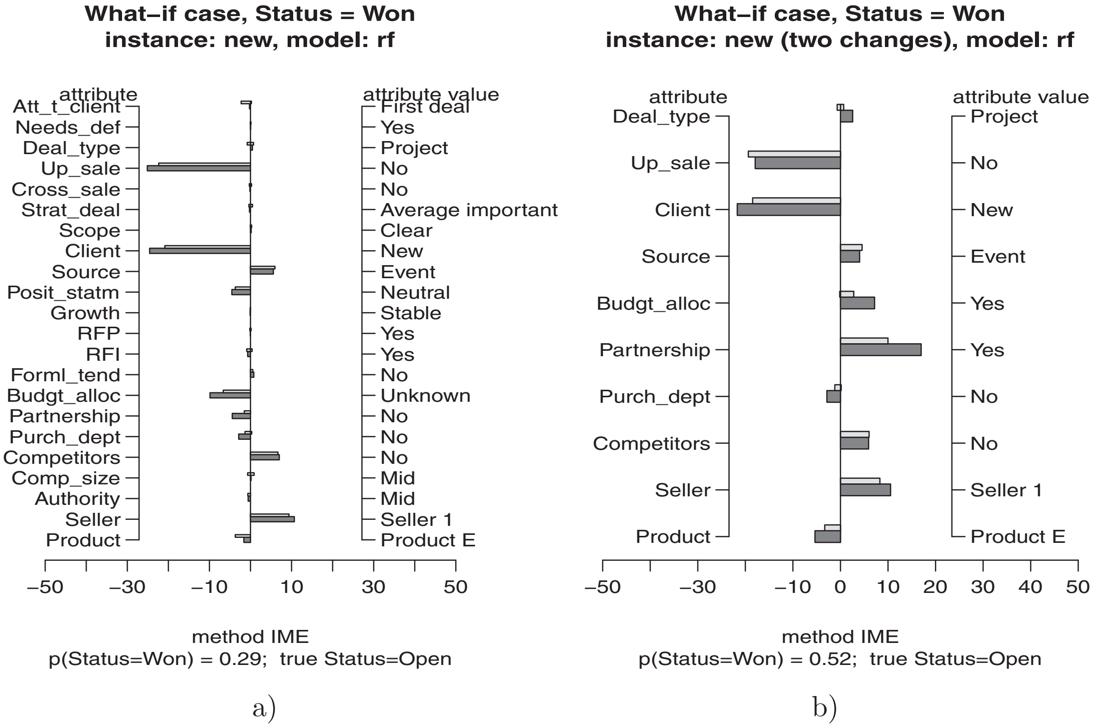


Fig. 8. Initial explanation (a) and explanation after updates (b).

business problem of B2B sales forecasting. We show how powerful black-box ML models can be made transparent and help domain experts to iteratively evaluate and update their beliefs.

To our knowledge, the proposed approach is the first application of the general explanation methods EXPLAIN and IME to a complex real-world business problem. We explained the design of the data set with descriptive attributes reflecting the sales process and sales history. We evaluated several popular well-performing black-box ML models and selected the best performing model. On the B2B sales forecasting problem, we presented several modes of explanations: from individual decisions to the whole model. For new (open) cases, we demonstrated interactive support for decision makers, assessing various scenarios with explanatory what-if analysis. We presented an adaptation of the methodology to address a specific business request (weak performance in one segment). The explanations of the prediction models and what-if

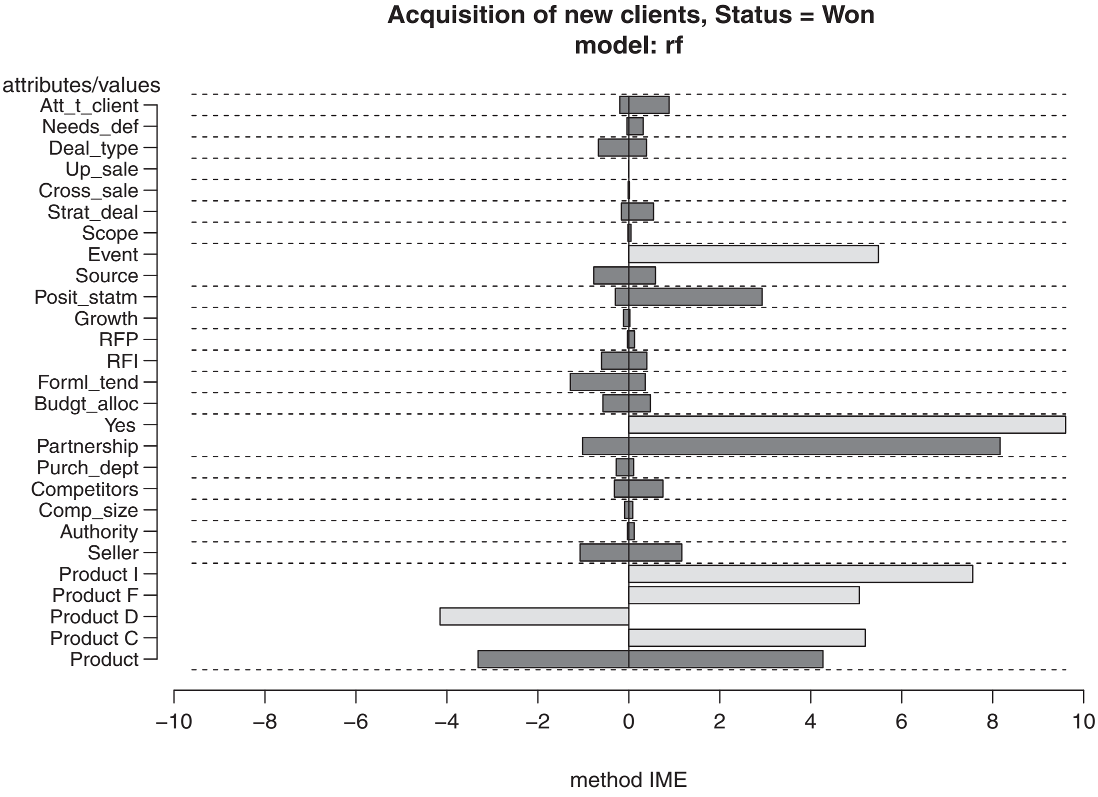


Fig. 9. Explanation of drivers for the acquisition of new clients.

analysis proved to be an effective support for B2B sales predictions. The presented methodology enhanced the team's internal communication and improved reflection on the team's implicit knowledge.

The presented intelligent system was tested for a longer period in a real-world company. The performance indicators confirm that forecasts created on the basis of the provided explanations outperform initial sales forecasts, which is in line with the intuition that explanations based on data better facilitate unbiased decision-making than the individual mental models of sellers. To foster the usage of the proposed methodology, the external consultant initially supported and trained the business users on how to apply the predictive ML models and the presented explanations. These activities have been proved to increase adoption speed; namely, a focused effort is needed to overcome resistance when users question/rebuilt their existing perspectives (Nonaka & Takeuchi, 1995). Organizations aiming to apply the presented approach are advised to follow the steps outlined in Section 4: a) identify informative and available descriptive attributes, reflecting the context of their decision-making process, b) select the best-performing ML prediction model, and c) use the presented explanations to expose insights.

The presented approach is affected by some weaknesses and limitations. The EXPLAIN method cannot capture disjunctively expressed dependencies in the prediction model. This is resolved by the method IME; however, for large data sets, this method might be slow and would have to be precomputed in order to be used interactively in a discussion session. The interactions between attributes are captured but not expressed explicitly in the visualization; therefore, the user has to manually discover the type of interdependencies with interactive analysis. The explanations closely follow the prediction model; if the model is wrong or performs poorly, the explanations will reflect that. An attribute-based data set representing historical instances of the decision problem is a necessary condition to use the presented approach. In B2B sales, forecasting slippages (delays in deals, i.e., deals not being finished in the observed month) are frequent. They reflect overly optimistic forecasts (Armstrong et al., 2015), which is a weakness requiring additional research.

The favorable properties of the proposed methodology include a comprehensive format of explanations for an arbitrary ML model and support for model's parsimony via threshold for significance of attributes. The use of descriptive attributes enables explanations in the language of a specific domain, where explanations are on the level of each prediction as well as on the level of the whole model. The methodology enables advanced what-if analysis for evaluation of decision options. For the IME method, the explanations in the form of attribute-value contributions have a theoretical guarantee that the computed contributions to the final prediction are fair in the sense that they represent Shapley values from coalitional game theory (only the IME). The software implementation of the explanation methodology is available as the open-source R package *ExplainPrediction* (Robnik-Šikonja, 2015). The real-world B2B sales forecasting data set is publicly accessible (Bohanec, 2016).

We see several opportunities to extend the work presented in this study. Analyses of additional decision-making processes shall further improve the visualization and presentation of explanations. Further practical use shall establish recommended parameters for different needs and application scenarios. We believe that it is possible to identify relevant attribute interactions with modification of the presented explanation methods. In B2B sales forecasting, further work is needed to identify slippages; we believe that deal closure time could be treated as a regression problem.

Acknowledgment

We are grateful to the company Salvirt, ltd., for funding the research and development presented in this paper. Mirjana Kljajić Borštnar and Marko Robnik-Šikonja were supported by the

Slovenian Research Agency, ARRS, through research programmes P5-0018 and P2-0209, respectively.

References

Aggarwal, C. C., Chen, C., & Han, J. (2010). The inverse classification problem. *Journal of Computer Science and Technology, 25*(3), 458–468.

Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems, 8*(6), 373–384.

Armstrong, J. S., Green, K. C., & Graefe, A. (2015). Golden rule of forecasting: Be conservative. *Journal of Business Research, 68*(8), 1717–1731.

Arnold, V., Clark, N., Collier, P. A., Leech, S. A., & Sutton, S. G. (2006). The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. *Mis Quarterly*, 79–97.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K.-R. (2010). How to explain individual classification decisions. *Journal of Machine Learning Research, 11*(June), 1803–1831.

Barbella, D., Benzaid, S., Christensen, J. M., Jackson, B., Qin, X. V., & Musicant, D. R. (2009). Understanding support vector machine classifications via a recommender system-like approach. . In R. Stahlbock, S. F. Crone, & S. Lessmann (Eds.), *Proceedings of international conference on data mining* (pp. 305–311).

Barbosa, A., Paulovich, F., Paiva, A., Goldenstein, S., Petronetto, F., & Nonato, L. (2016). Visualizing and interacting with kernelized data. *IEEE Transactions on Visualization and Computer Graphics, 22*(3), 1314–1325.

Becker, B., Kohavi, R., & Sommereld, D. (1997). Visualizing the simple bayesian classier. *KDD workshop on issues in the integration of data mining and data visualization*.

Bohanec, M. (2016). Anonymized B2B sales forecasting data set http://www.salvirt.com/research/b2bdataset/.

Bohanec, M., Kljajić Borštnar, M., & Robnik-Šikonja, M. (2016). Integration of machine learning insights into organizational learning: A case of B2B sales forecasting. In F. D'Ascenzo, M. Magni, A. Lazazzara, & S. Za (Eds.), *Blurring the boundaries through digital innovation: Individual, organizational, and societal challenges*: 19. Springer.

Bosnić, Z., Demšar, J., Kešpret, G., Rodrigues, P. P., Gama, J., & Kononenko, I. (2014). Enhancing data stream predictions with reliability estimators and explanation. *Engineering Applications of Artificial Intelligence, 34*, 178–192.

Breiman, L. (2001). Random forests. *Machine Learning Journal, 45*, 5–32.

Brynjolfsson, E., Hitt, L. M., & Kim, H. H. (2011). Strength in numbers: How does data-driven decision-making affect firm performance? Available at SSRN 1819486,.

Caragea, D., Cook, & Dianne Honavar, V. (2003). Towards simple, easy-to-understand, yet accurate classifiers. In *Third IEEE international conference on data mining, ICDM 2003.* (pp. 497–500).

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on machine learning ICML '06* (pp. 161–168). New York, NY, USA: ACM.

Costea, A., & Bleotu, V. (2012). A new fuzzy clustering algorithm for evaluating the performance of non-banking financial institutions in Romania. *Economic Computation and Economic Cybernetics Studies and Research, 46*(4), 179–199.

D'Haen, J., & Van der Poel, D. (2013). Model-supported business-to-business prospect prediction based on an iterative customer acquisition framework. *Industrial Marketing Management, 42*, 544–551.

Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. a correlated-adjusted decision forest proposal. *Expert Systems with Applications, 42*(13), 5737–5753.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics, 24*(1), 44–65.

Gönül, M. S., Önkal, D., & Lawrence, M. (2006). The effect of structural characteristics of explanations on use of a DSS. *Decision Support Systems*, 1481–1493.

Good, I. J. (1950). *Probability and the weighing of evidence*. C. Griffin, London.

Hamel, L. (2006). Visualization of support vector machines with unsupervised learning. In *Proceedings of 2006 IEEE symposium on computational intelligence in bioinformatics and computational biology*.

Hansen, K., Baehrens, D., Schroeter, T., Rupp, M., & Müller, K.-R. (2011). Visual interpretation of kernel-based prediction models. *Molecular Informatics, 30*(9), 817–826.

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics, 3*(2), 119–131.

Jacobsson, H. (2005). Rule extraction from recurrent neural networks: A taxonomy and review. *Neural Computation, 17*(6), 1223–1263.

Jakulin, A., Možina, M., Demšar, J., Bratko, I., & Zupan, B. (2005). Nomograms for visualizing support vector machines. . In R. Grossman, R. Bayardo, & K. P. Bennett (Eds.), *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 108–117). ACM.

Kononenko, I. (1993). Inductive and bayesian learning in medical diagnosis. *Applied Artificial Intelligence, 7*(4), 317–337.

Lemaire, V., & Clérot, F. (2004). An input variable importance definition based on empirical data probability and its use in variable selection. In *Proceedings of IEEE international joint conference on neural networks* (pp. 1375–1380 vol.2). doi:10.1109/IJCNN.2004.1380149.

Lemaire, V., Féraud, R., & Voisine, N. (2008). Contact personalization using a score understanding method. In *Proceedings of international joint conference on neural networks*.

Lilien, G. L. (2016). The B2B knowledge gap. *International Journal of Research in Marketing, 33*(3), 543–556.

Lubsen, J., Pool, J., & van der Does, E. (1978). A practical device for the application of a diagnostic or prognostic function. *Methods of Information in Medicine, 17*, 127–129.

Madigan, D., Mosurski, K., & Almond, R. G. (1997). Graphical explanation in belief networks. *Journal of Computational and Graphical Statistics, 6*(2), 160–181.

Martens, D., & Provost, F. (2011). Explaining documents classifications. *Technical Report*. Center for Digital Economy Research, New York University, Stern School of Business. Working paper CeDER-11-01

McAfee, A., & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review, 90*(10), 61–67.

Merkert, J., Mueller, M., & Hubl, M. (2015). A survey of the application of machine learning in decision support systems. *European conference on information systems (ECIS), completed research papers. Munster, Germany*.

Meyer, G., Adomavicius, G., Johnson, P. E., Elidrisi, M., Rush, W. A., Sperl-Hillen, J. M., & O'Connor, P. J. (2014). A machine learning approach to improving dynamic decision making. *Information Systems Research, 25*(2), 239–263.

Možina, M., Demšar, J., Kattan, M. W., & Zupan, B. (2004). Nomograms for visualization of naive bayesian classifier. . In J.-F. Boulicaut, F. Esposito, F. Giannotti, & D. Pedreschi (Eds.), *Knowledge discovery in databases: PKDD 2004* (pp. 337–348). Springer.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford University Press.

Poulet, F. (2004). SVM and graphical algorithms: A cooperative approach. In *Fourth IEEE international conference on data mining (ICDM'04)* (pp. 499–502).

Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D. S., … Anvik, J. (2006). Visual explanation of evidence with additive classifiers. In *Proceedings of AAAI'06*. AAAI Press.

Pregeljc, M., Štrumbelj, E., Mihelcic, M., & Kononenko, I. (2012). Learning and explaining the impact of enterprises organizational quality on their economic results. In R. Magdalena-Benedito, M. Martnez-Sober, J. M. Martnez-Martnez, P. Escandell-Moreno, & J. Vila-Francs (Eds.), *Intelligent data analysis for real-life applications: Theory and practice* (pp. 228–248). Information Science Reference, IGI Global.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. *Big Data, 1*(1), 51–59.

Provost, F. J., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms.. In *Proceedings of international conference on machine learning, ICML'98: 98* (pp. 445–453).

Robnik-Šikonja, M. (2015). Explainprediction: Explanation of predictions for classification and regression. R package version 1.1.2.

Robnik-Šikonja, M., & Kononenko, I. (1995). Discretization of continuous attributes using ReliefF. In F. Solina, & B. Zajc (Eds.), *Proceedings of electrotehnical and computer science conference (ERK'95)* (pp. B149–152). Slovene section of IEEE, Ljubljana.

Robnik-Šikonja, M., & Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering, 20*(5), 589–600.

Saltelli, A., Chan, K., & Scott, E. M. (2000). *Sensitivity analysis*. New York: Wiley.

Schulz, A., Gisbrecht, A., & Hammer, B. (2015). Using discriminative dimensionality reduction to visualize classifiers. *Neural Processing Letters, 42*(1), 27–54.

Simon, H. A. (1960). The new science of management decision.,.

Štrumbelj, E., Bosnić, Z., Kononenko, I., Zakotnik, B., & Kuhar, C. G. (2010). Explanation and reliability of prediction models: The case of breast cancer recurrence. *Knowledge and Information Systems, 24*(2), 305–324.

Verikas, A., Gelzinis, A., & Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition, 44*(2), 330–349.

Štrumbelj, E., & Kononenko, I. (2010). An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research, 11*, 1–18.

Štrumbelj, E., Kononenko, I., & Robnik-Šikonja, M. (2009). Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering, 68*(10), 886–904.

Yan, J., Zhang, C., Zha, H., Gong, M., Sun, C., Huang, J., … Yang, X. (2015). On machine learning towards predictive sales pipeline analytics. In *Twenty-ninth AAAI conference on artificial intelligence* (pp. 1945–1951).