In: Zhou J., Chen F. (Eds) Human and Machine Learning. pp. 159-175, Springer, Cham., 2018

Chapter 1
Perturbation-Based Explanations of Prediction Models

Marko Robnik-Šikonja and Marko Bohanec

Abstract Current research into algorithmic explanation methods for predictive models can be divided into two main approaches: gradient-based approaches limited to neural networks and more general perturbation-based approaches which can be used with arbitrary prediction models. We present an overview of perturbation-based approaches, with focus on the most popular methods (EXPLAIN, IME, LIME). These methods support explanation of individual predictions but can also visualize the model as a whole. We describe their working principles, how they handle computational complexity, their visualizations as well as their advantages and disadvantages. We illustrate practical issues and challenges in applying the explanation methodology in a business context on a practical use case of B2B sales forecasting in a company. We demonstrate how explanations can be used as a what-if analysis tool to answer relevant business questions.

1.1 Introduction

Machine learning models play an increasingly large role in many applications, products, and services. Their outcomes are part of everyday life (e.g., entertainment recommendations), as well as life-changing decisions (e.g., medical diagnostics, credit scoring, or security systems). We can expect that reliance on technology and machine learning will only increase in the future. It is only natural that those affected by various automated decisions want to get feedback and understand the reason-

Marko Robnik-Šikonja
University of Ljubljana, Faculty of Computer and Information Science
Večna pot 113, 1000 Ljubljana, Slovenia
e-mail: marko.robnik@fri.uni-lj.si

Marko Bohanec
Salvirt Ltd., Dunajska 136, 1000 Ljubljana, Slovenia
e-mail: marko.bohanec@salvirt.com

ing process and biases of the underlying models. Areas where model transparency is of crucial importance include public services, medicine, science, policy making, strategic planning, business intelligence, finance, marketing, insurance, etc. In these areas, users of models are just as interested to comprehend the decision process, as in the classification accuracy of prediction models. Unfortunately, most of the top performing machine learning models are black boxes in a sense that they do not offer an introspection into their decision processes or provide explanations of their predictions and biases. This is true for Artificial Neural Networks (ANN), Support Vector Machines (SVM), and all ensemble methods (for example, boosting, random forests, bagging, stacking, and multiple adaptive regression splines). Approaches that do offer an intrinsic introspection, such as decision trees or decision rules, do not perform so well or are not applicable in many cases [23].

To alleviate this problem two types of model explanation techniques have been proposed. The first type, which is not discussed in this chapter, is based on the internal working of the particular learning algorithm. The explanation methods exploit a model's representation or learning process to gain insight into the presumptions, biases and reasoning leading to final decisions. Two well-known models where such approach works well are neural networks and random forests. Recent neural networks explainers mostly rely on layer-wise relevance propagation [6] or gradients of output neurons with respect to the input [32] to visualise parts of images significant for particular prediction. The random forest visualisations mostly exploit the fact that during bootstrap sampling, which is part of this learning algorithm, some of the instances are not selected for learning and can serve as an internal validation set. With the help of this set, important features can be identified and similarity between objects can be measured.

The second type of explanation approaches are general and can be applied to any predictive model. The explanations provided by these approaches try to efficiently capture the causal relationship between inputs and outputs of the given model. To this end, they perturb the inputs in the neighbourhood of a given instance to observe effects of perturbations on the model's output. Changes in the outputs are attributed to perturbed inputs and used to estimate their importance for a particular instance. Examples of this approach are methods EXPLAIN [29], IME [35], and LIME [27]. These methods can explain the model's decision for each individual predicted instance as well as for the model as a whole. As they are efficient, offer comprehensible explanations, and can be visualised, they are the focus of this chapter. Other explanation methods are discussed in the background section.

Another aspect we try to address is how explanations of prediction models can be put into practical use. We are interested in the integration of explanations into a complex business decision process and their support of continuous organisational learning. Users of knowledge-based systems are more likely to adhere to automatic predictions, when, besides the predictive performance of models, explanations are also available [4]. In order to apply prediction models, users have to trust them first, and the model's transparency is a major factor in ensuring the trust. We illustrate an application of explanation methodology to a challenging real-world B2B sales forecasting [11]. A group of sales experts collected historical B2B sales cases to build

a machine learning prediction model. The explanations of past and new cases enabled cognitive evaluation of the model. Based on the new insights, provided by the explanations, experts can update the data set, propose new features, and re-evaluate the models. We discuss several issues arising and how they can be addressed with the explanation methodology.

The objectives of the chapter are twofold. First, to explain how general perturbation-based explanation methods work, and second, to demonstrate their practical utility in a real-world scenario. The first aim is achieved through an explanation of their working principle and graphical explanation of models' decisions on a well-known data set. Two types of explanations are demonstrated, individual predictions of new unlabelled cases and functioning of the model as a whole. This allows inspection, comparison, and visualisation of otherwise opaque models. The practical utility of the methodology is demonstrated on the B2B sales forecasting problem.

The structure of the chapter is as follows. In Section 1.2 we present a taxonomy of explanation methodologies and present background and related work on perturbation-based approaches. In Section 1.3 we present methods EXPLAIN, IME, and LIME, their similarity and differences. Explanations in a business context are discussed in Section 1.4 through B2B sales forecasting. In Section 1.5 we present conclusions.

1.2 Background and Overview of Perturbation Approaches

True causal relationships between dependent and independent variables are typically hidden except in artificial domains where all the relations, as well as the probability distributions, are known in advance. Therefore only explanations of the prediction process for a particular model is of practical importance. The prediction accuracy and the correctness of explanation for a given model may be orthogonal: the correctness of the explanation is independent of the correctness of the prediction. However, empirical observations show that better models (with higher prediction accuracy) enable better explanations [35]. We discuss two types of explanations:

- **Instance explanation** explains predictions with the given model of individual instances and provides the impact of input feature values on the predictions.
- **Model explanation** is usually an aggregation of instance explanations over many (training) instances, to provide top-level explanations of features and their values. This aggregation over many instances enables identification of different roles attributes may play in the classifications of instances.

Below we list several properties of machine learning explanations. They stem from criteria for evaluation of rule extraction methods from neural networks introduced by [2] and later extended by [18]. Some items were proposed by [27] and [21], and some are the result of our work.

1. *Expressive power* describes the language of extracted explanations: propositional logic (i.e. if-then rules), nonconventional logic (e.g., fuzzy logic), first-

order logic, finite state machines (deterministic, nondeterministic, stochastic), histograms, decision trees, linear models, a limited form of natural language etc.
2. *Translucency* describes the degree to which an explanation method looks inside the model. It can be decompositional (decomposes internal representation of the model, e.g., in neural networks meaning of individual neurons), pedagogical (treating the model as a black box), or eclectic (combining both compositional and pedagogical types).
3. *Portability* describes how well the technique covers the range of different models (e.g., limited to convolutional neural networks, suitable for additive models, general, etc.).
4. *Algorithmic complexity* deals with the computational complexity of algorithms producing explanations.

Quality of explanations is another very important aspect, which groups several properties of explanation methods:

5. *Accuracy:* the ability that explanation of a given decision generalises to other yet unseen instances. For example, if explanations are in the form of rules, are these rules general and do they cover unseen instances.
6. *Fidelity:* how well the explanations reflect the behaviour of the prediction model. *Local fidelity* expresses how well the explanations reflect the behaviour of the prediction model in the vicinity of predicted instances. Local fidelity does not imply general fidelity (e.g., features that are important in a local context may not be important in the global context of the model).
7. *Consistency:* the degree to which similar explanations are generated from different models trained on the same task. For example, while similar models may produce very similar predictions, the explanations of similar instances may vary due to the variance of certain explanation methods.
8. *Stability:* the degree to which similar explanations are generated for similar instances. Different to *consistency*, which covers several models, this criterion deals with explanations generated from the same model. As for consistency, while predictions of similar instances may be the same, the explanations may vary due to the variance of certain explanation methods.
9. *Comprehensibility:* readability of explanations (might depend on the audience, e.g., experts or the general public) and size of explanations (e.g., number of rules, number of items shown on a bar chart, number of words, number of factors in linear model etc.).
10. *Certainty:* are explanations reflecting certainty of a model about its predictions? For example, a classifier may be very certain of its prediction, but the explanation may or may not reflect it.
11. *Degree of importance:* are explanations reporting the degree of importance for each returned item (e.g., the importance of explained features, or importance of returned rules)?
12. *Novelty:* is a form of certainty and tells if explanations would reflect the fact that explained instance is from a new region, not contained or well represented in the training set (the model may be unreliable for such instances).

13. *Representativeness:* are explanations representative of the model? For example, a model explanation may cover behaviour of the whole model, or just a part of it.

In a typical data science problem setting, users are concerned with both prediction accuracy and the interpretability of the prediction model. Complex models have potentially higher accuracy but are more difficult to interpret. This can be alleviated either by sacrificing some prediction accuracy for a more transparent model or by using an explanation method that improves the interpretability of the model. Explaining predictions is straightforward for symbolic models such as decision trees, decision rules, and inductive logic programming, where the models give an overall transparent knowledge in a symbolic form. Therefore, to obtain the explanations of predictions, one simply has to read the rules in the corresponding model. Whether such an explanation is comprehensive in the case of large trees or large rule sets is questionable. [24] developed criteria for decision trees and performed a user study, which showed that the depth of the deepest leaf that is required when answering a question about a classification tree is the most important factor influencing the comprehensibility.

For non-symbolic models, there are no intrinsic explanations. A lot of effort has been invested into increasing the interpretability of complex models. For SVM, [16] proposed an approach based on self-organising maps that groups instances then projects the groups onto a two-dimensional plane. In this plane, the topology of the groups is hopefully preserved and support vectors can be visualised. Many approaches exploit the essential property of additive classifiers to provide more comprehensible explanations and visualisations, e.g., [19] and [25].

Visualisation of decision boundaries is an important aspect of model transparency. [9] present a technique to visualise how the kernel embeds data into a high-dimensional feature space. With their Kelp method, they visualise how kernel choice affects neighbourhood structure and SVM decision boundaries. [31] propose a general framework for visualisation of classifiers via dimensionality reduction. [15] propose another useful visualisation tool for classifiers that can produce individual conditional expectation plots, graphing the functional relationship between the predicted response and the feature for individual instance.

Some explanations methods (including the ones presented in Section 1.3) are general in a sense that they can be used with any type of prediction model that returns a numeric score (either probability of a class or numeric prediction) [20, 27, 29, 34]. This enables their application with almost any prediction model and allows users to analyse and compare outputs of different analytical techniques. [20] applied their method to a customer relationship management system in the telecommunications industry. The method which successfully deals with high-dimensional text data is presented in [22]. Its idea is based on general explanation methods EXPLAIN and IME and offers an explanation in the form of a set of words which would change the predicted class of a given document. [13] adapt the general explanation methodology to a data stream scenario and show the evolution of attribute contributions through time. This is used to explain the concept drift in their incremental model. In a real-life breast cancer recurrence prediction, [33] illustrate the usefulness of the visualisations and the advantage of using the general explanation

method. Several machine learning algorithms were evaluated. Predictions were enhanced with instance explanations using the IME method. Visual inspection and evaluation showed that oncologists found the explanations useful and agreed with the computed contributions of features. [26] used traditional modelling approaches together with data mining to gain insight into the connections between the quality of organisation in enterprises and the enterprises performance. The best performing models were complex and difficult to interpret, especially for non-technical users. Methods EXPLAIN and IME explained the influence of input features on the predicted economic results and provided insights with a meaningful economic interpretation. The interesting economic relationships and successful predictions come mostly from complex models such as random forests and ANN. Without proper explanation and visualisation, these models are often neglected in favour of weaker, but more transparent models. Experts from the economic-organisational field, which reviewed and interpreted the results of the study, agreed that such an explanation and visualisation is useful and facilitates comparative analysis across different types of prediction models.

Many explanation methods are related to statistical sensitivity analysis and uncertainty analysis [30]. In that methodology, the sensitivity of models is analysed with respect to models' input. A related approach, called inverse classification [1], tries to determine the minimum required change to a data point in order to reclassify it as a member of a different class. An SVM model based approach is proposed by [8]. Another sensitivity analysis-based approach explains contributions of individual features to a particular classification by observing (partial) derivatives of the classifiers' prediction function at the point of interest [7]. 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 used in practice with kernel-based classification method to predict molecular features [17].

Due to recent successes of deep neural networks in image recognition and natural language processing, several explanation methods specific to these two application areas emerged. Methods working on images try to visualise parts of images (i.e., groups of pixels) significant for a particular prediction. These methods mostly rely on the propagation of relevance within the network, e.g., layer-wise relevance propagation [6], or computation of gradients of output neurons with respect to the input [32]. In language processing, [5] applied layer-wise relevance propagation to a convolutional neural network and a bag-of-words SVM classifier trained on a topic categorisation task. The explanations indicate how much individual words contribute to the overall classification decision.

1.3 Methods EXPLAIN, IME, and LIME

General explanation methods can be applied to any classification model which makes them a useful tool both for interpreting models (and their predictions) and comparing different types of models. By modification of feature values of interest, what-if analysis is also supported. Such methods cannot exploit any model-specific properties (e.g., gradients in ANN) and are limited to perturbing the inputs of the model and observing changes in the model's output [20, 29, 34].

The three presented general explanation methods provide two types of explanations for prediction models: instance explanations and model explanations (see Section 1.2). Model explanations work by summarising a representative sample of instance explanations. All three methods estimate the impact of a particular feature on the prediction of a given instance by perturbing similar instances.

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

$$p_Q(y_k|x) = \mathbb{E}(p(y_k)|X_i = x_{(i)}, \forall X_i \in Q). \tag{1.1}$$

Therefore, $p_S(y_k|x) = p(y_k|x)$. The difference between $p_S(y_k|x)$ and $p_Q(y_k|x)$ is a basis for explanations. In practical settings, the classification function of the model is not known - one can only access its prediction for any vector of input values. Therefore, an exact computation of $p_Q(y_k|x)$ is not possible and sampling-based approximations are used.

In model explanations, to avoid loss of information due to summarisation of instance level explanations, in the presented visualisation the evidence for and against each class is collected separately. In this way, one can, for example, see that a particular value of an attribute supports specific class but not in every context.

1.3.1 EXPLAIN, One-Variable-at-a-Time Approach

The EXPLAIN method computes the influence of a feature value by observing its impact on the model's output. The EXPLAIN method assumes that the larger the changes in the output, the more important role the feature value plays in the model. The shortcoming of this approach is that it takes into account only a single feature at a time, therefore it cannot detect certain higher order dependencies (in particular disjunctions) and redundancies in the model. The EXPLAIN method assumes that the characterisation of the i-th input variable's importance for the prediction of the instance x is the difference between the model's prediction for that instance and the

model's prediction if the value of the i-th variable was not known, namely: $p(y_k|x) - p_{S\setminus\{i\}}(y_k|x)$. If this difference is large then the i-th 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. This approach was extended in [29] to use log-odds ratios (or weight of evidence) instead of the difference in predicted class probabilities.

To demonstrate behaviour of the method, an example of an explanation is given. We use a binary classification problem with three important (A_1, A_2, and A_3) and one irrelevant attribute (A_4), so the set of attributes is $S = \{1,2,3,4\}$. Let us assume that the learned model correctly expresses the class value as the parity (xor) relation of three attributes $C = A_1 \oplus A_2 \oplus A_3$. The correct model would classify an instance $x = (A_1 = 1, A_2 = 0, A_3 = 1, A_4 = 1)$ to class $C = 0$, and assigns it probability $p(C = 0|x) = 1$. When explaining classification for this particular instance $p(C = 0|x)$, method EXPLAIN simulates the lack of knowledge of a single attribute at a time, so one has to estimate $p_{S-\{1\}}(C = 0|x)$, $p_{S-\{2\}}(C = 0|x)$, $p_{S-\{3\}}(C = 0|x)$, and $p_{S-\{4\}}(C = 0|x)$. Without the knowledge about the values of each of the attributes A_1, A_2, and A_3, the model cannot correctly determine the class value, so the correct estimates of class probabilities are $p_{S-\{1\}}(C = 0|x) = p_{S-\{2\}}(C = 0|x) = p_{S-\{3\}}(C = 0|x) = 0.5$ The differences of probabilities $p_S(y_k|x) - p_{S-\{i\}}(y_k|x)$ therefore equal 0.5 for each of the three important attributes, which indicate that these attributes have positive impact on classification to class 0 for the particular instance x. The irrelevant attribute A4 does not influence the classification, so the classification probability remain unchanged $p_{S-\{4\}}(C = 0|x) = 1$. The difference of probabilities $p_S(C = 0|x) - p_{S-\{4\}}(C = 0|x) = 0$ so the explanation of the irrelevant attributes impact is zero.

The produced explanations, i.e. conditional probabilities of Eq. (1.1) computed for each feature separately with EXPLAIN method can be visualised with a form of quasi-nomograms. The positive and negative impacts of each feature for a given class value are presented separately. We present an example of this visualisation in Section 1.3.4.

1.3.2 IME, All-Subsets Approach

The one-variable-at-a-time approach is simple and computationally less-intensive but has some disadvantages. The main disadvantage is that disjunctive concepts or redundancies between input variables may result in unintuitive contributions for variables [35]. A solution was proposed in [34], where all subsets of values are observed. Such procedure demands 2^a steps, where a is the number of attributes, and results in the exponential time complexity. However, the contribution of each variable corresponds to the Shapley value for the coalitional game of a players. This allows an efficient approximation based on sampling.

1.3.3 LIME, Optimisation of Explanations

LIME (Local Interpretable Model-agnostic Explanations) [27] efficiently calculates explanations also for very large data sets in terms of a number of instances and number of features. It uses perturbations in the locality of an explained instance to produce explanations (e.g., in a fashion of locally weighted regression). It defines explanations as an optimisation problem and tries to find a trade-off between local fidelity of explanation and its interpretability. The search space is over explanations generated by interpretable models $g \in G$, where G is a class of interpretable models. These are not necessary input features but can be linear models, decision trees, or rule lists. Interpretability is quantified with the complexity of explanations $\Omega(g)$, where complexity measure Ω can be the depth of tree for decision trees or the number of non-zero weights for linear models. The model f being explained has to return numeric values $f : \Re^d \rightarrow \Re$, for example probability scores in classification. Locality is defined using a proximity measure π between the explained instance x and perturbed points z in its neighbourhood. Local fidelity $L(f,g,\pi)$ is a measure of how unfaithful the explanation model g is in approximating the prediction model f in the locality defined by $\pi(x,z)$. The chosen explanation then minimises the sum of local infidelity L and complexity Ω:

$$e(x) = \arg\min_{g \in G} L(f,g,\pi) + \Omega(g) \tag{1.2}$$

The approach uses sampling around explanation instance x to draw samples z weighted by the distance $\pi(x,z)$. The samples form a training set for a model g from an interpretable model class, e.g., a linear model. Due to locality enforced by π, the model g is hopefully a faithful approximation of f. In practice, [27] use linear models as a class of interpretable models G, the squared loss as a local infidelity measure, number of non-zero weights as complexity measure Ω, and choose sample points in the neighbourhood of explanation instance x according to the Gaussian distribution of distance between x and sampled point z.

To explain text classification tasks, LIME uses bag-of-words representation to output a limited number of the most locally influential words. In image classification, it returns a list of the most influential image areas (super-pixels) for particular prediction.

By presenting explanation as an optimisation problem, LIME avoids the exponential search space of all feature combinations which is solved by game-theory based sampling in IME. However, LIME offers no guarantees that the explanations are faithful and stable. Using neighbourhood around explanation instance, it may fall into a curse of dimensionality trap, which is fatal for neighbourhood-based methods like kNN in high dimensional spaces. The problem of feature interactions is seemingly avoided by using approximating function from a class of interpretable explanation but the problem is just swept under the carpet, as the interpretable explanation class may not be able to detect them (e.g., linear functions). Further investigation of this question is needed and we suggest a combination of IME and LIME compo-

nents as a further work. An idea worth pursuing seems to be integration of game theory based sampling from IME and explanations as optimisation used in LIME.

1.3.4 Presenting Explanations

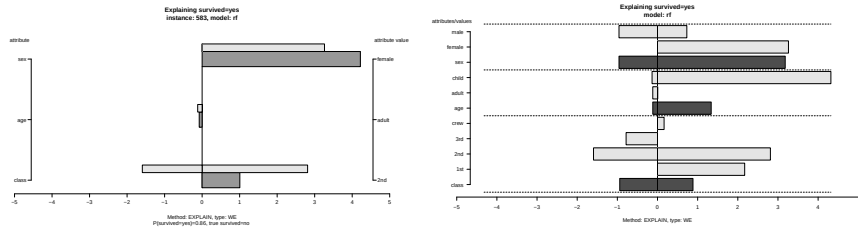
The explanations produced by EXPLAIN and their visualisation are illustrated on the well-known Titanic data set (we used the version accompanying the Orange toolkit [14]). The task is to classify survival of passengers in the disaster of the HMS Titanic ship. The three input variables report the passengers' status during travel (first, second, third class, or crew), age (adult or child), and gender (male or female). We have chosen this data set due to its simplicity but note the similarity of the problem with many business decision problems, such as churn prediction, mail response, insurance fraud, etc. As an example of an opaque prediction model, we use random forest (rf) classifier. This is an ensemble of many (typically hundreds), almost random, tree models. While this approach typically produces models with good predictive performance (on the Titanic problem the classification accuracy is 78%), the models are incomprehensible.

We demonstrate explanations extracted from the random forest model. Fig. 1.1(a) shows an example of an instance explanation for the prediction of the instance with id 583 (a second class adult female passenger). The text at the top includes the predicted value ("survived=yes"), instance id (583), and model name (rf). Below the graph, there is information on the explanation method (EXPLAIN, using the weight of evidence), the model's prediction (P("survived=yes") = 0.86), and the actual class value of the instance ("survived=no"). The input variables' names are shown on the left-hand side (sex, age, and class) and their values for the particular instance are on the right-hand side (female, adult, and second class). The thick dark shaded bars going from the centre to the right or left indicate the contributions of the instance's values for each corresponding input variable towards or against the class value "survived=yes", respectively. The longer the bars the stronger the contributions of the corresponding feature values. The scale of the horizontal axis depends on the explanation method. For the EXPLAIN method and weight of evidence (WE) shown in Fig. 1.1(a), the horizontal axis shows the log-odds transformed difference of probabilities (see Section 1.3.1). The thinner and lighter bars above the thick dark bars indicate average contributions of these values across all training instances. For the given instance, one can observe that both "sex=female" and "status=second class" speaks in favour of survival (therefore the model is pretty sure of survival with probability 86%), while being an adult has a tiny negative influence. Thinner average bars above them reveal that being a female is on average beneficial, while a second class can have both positive and negative impact. Being an adult has on average a tiny negative impact. Note that the same visualisation can be used even if some other classification method is applied.

A more general view of the model is provided by averaging the explanations over all training set instances. This summary form visualisation shows the average impor-

tance of each input variable and its values. An example of such model explanation for the Titanic data set is presented in Fig. 1.1(b). On the left-hand side, the input variables and their values are shown. For each value, the average negative and the average positive contributions across all instances is displayed. Note that negative and positive contributions would cancel each other out if summed together, so it is important to keep them separate. The lighter bars shown are equivalent to the lighter bars in the instance explanation on Fig. 1.1(a). For each input variable, the average positive and negative contributions for all values and instances are shown (darker bars). The visualisation reveals that travelling in first class or being a child or female has a strong positive contribution towards survival, travelling in third class has a predominately negative contribution, while other statuses have smaller or mixed effect in the random forest model. For more complex data sets with many attributes the visualisation of model explanation may become cluttered, so we can set the threshold and only visualise the most important values.

The presented visualisations are produced by the function explainVis from R package ExplainPrediction [28], which has many parameters controlling the computation of explanations and their visualisation. The most important parameters controlling computation of explanations are the type of explanation (EXPLAIN, IME), which class value shall be explained, and parameters specific for EXPLAIN (how the lack of information about certain feature is simulated) and IME (allowed error and the maximal number of iterations). The parameters controlling visualisation are the type of graphical output (e.g., jpg, eps, or png), the selection of attributes to be shown, the threshold of importance for displayed attributes, text shown on the graph, colours, etc.



(a) (b)

Fig. 1.1 An instance explanation (on the left-hand side) and a model explanation (on the right-hand side) for the random forest model classifying the Titanic data set.

1.4 Explanation in Business Context: A Case of B2B Sales Forecasting

[3] 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 practical use case demonstrates that this need not be the case. We show that explanations can successfully support data-based decision process in a real-world business context [11, 12].

We use a publicly available real-world data set describing the sales history of a medium-sized company providing software solutions to clients in international B2B markets [10]. The data set consists of 22 predictive attributes describing different aspects of the B2B sales process (e.g., a type of offered product, the authority of a contact person at the client, size of the company, seller's id, etc.). The class variable is boolean indicating if a deal was won or lost. The data set promotes research in understanding factors impacting the outcome of the sales process. To construct the dataset, the sales team analysed 448 open opportunities with the help of an external consultant. The predictions, as well as the final outcomes, were recorded and analysed with machine learning prediction models. To gain knowledge about the properties of the decision process, the sales team used general explanation methods EXPLAIN and IME. The analysis included explanations of individual decisions as well as the whole model. For new (open) cases, decision makers were supported with the explanations to assess various scenarios with explanatory what-if analysis. We discuss two interesting use cases, the effect of updates as a result of new information and adaptation of the model to the specifics of new customers.

Fig. 1.2(a) shows instance explanation for a new sale opportunity, with values of all 22 attributes shown (the importance threshold is not set). The prediction model explained is a random forest. During the sales team's discussion, it was observed that value of the attribute *Competitors* was recorded incorrectly and should be corrected to "No". Furthermore, the sales managers wanted to assess the impact of assigning a different seller with more expertise in Product D. This is reflected in the change for the attribute (note the change of *Seller* from "Seller 2" to "Seller 10"). The team could immediately investigate the effect of these two updates, which is visible in Fig. 1.2(b), where the likelihood of a successful outcome increases from 0.52 to 0.68. We show only the most relevant attributes by setting the appropriate importance threshold (to value 3).

The participating company wanted to get insight into 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, complete business data set. By selecting only instances involving new clients, the learning outcome was intentionally biased. The resulting model and its explanations are not generally applicable but can help in distinguishing successful and unsuccessful deals involving new clients.

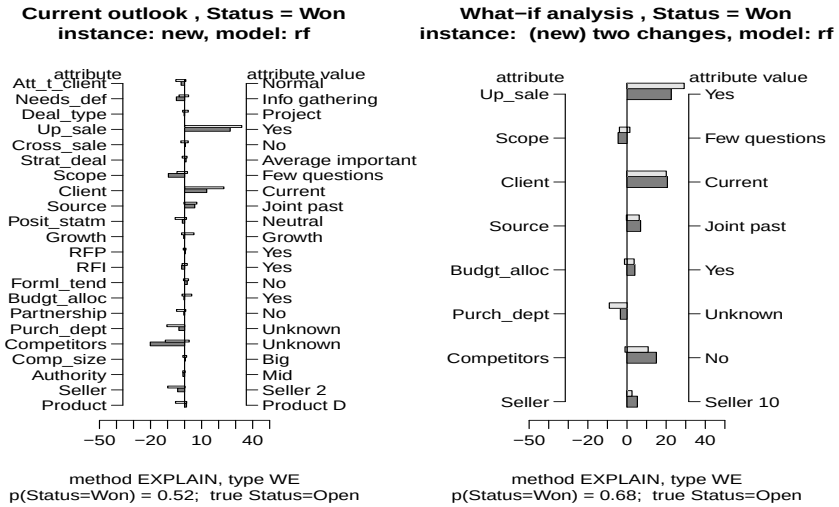


Fig. 1.2 Initial explanation (a) and explanation after updates (b).

The model explanation is presented in Fig. 1.3. We applied the importance threshold (value of 3) to discard features with low impact. The strongest positive impact comes from the attribute *Partnership* with value "Yes", which indicates a recommendation to form partnerships with companies when bidding for new business. Positive statements about the vendor have a positive impact, as well as when sales opportunities stem from participation in an event (e.g., booth at a conference). For this specific segment of new clients, some attributes have marginal or no impact (e.g. *Up_sale, Cross_sale*). This is in-line with reality – only existing clients qualify for *up sale* or *cross sale*. We can observe different impacts of products; some have positive and other have a negative impact. The rest of the values have impact below the set threshold of 3. Such a compact view enables a more targeted discussion when building a company's sales strategy.

One can conclude that the explanation methodology presented is a useful tool for many different problems. It is especially important for problems where prediction performance has to be supplemented with models transparency or knowledge discovery.

1.5 Conclusion

We presented three general methods for explanations of prediction models. The methods allow explanation of individual decisions as well as the prediction model as a whole. The methods can be efficiently computed and visualised, EXPLAIN and

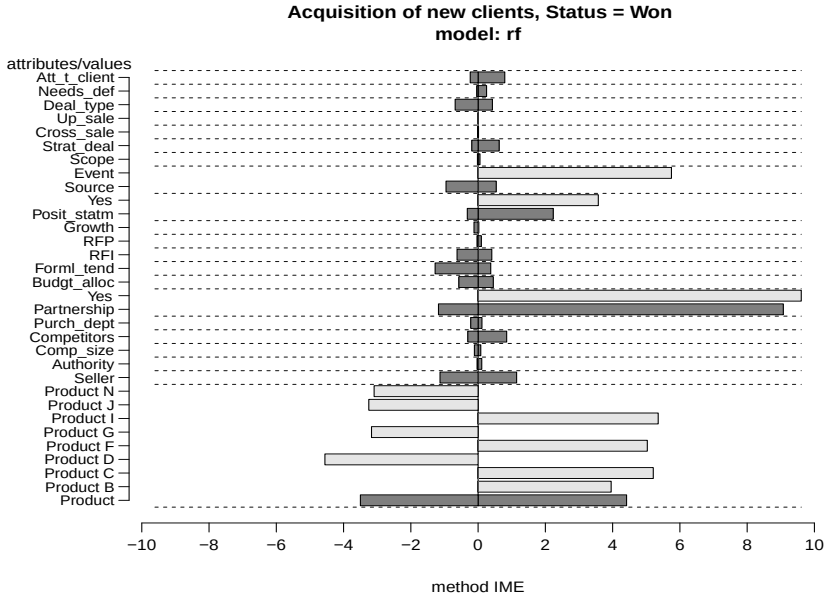


Fig. 1.3 Explanation of drivers for the acquisition of new clients.

LIME work efficiently even for very large data sets. The explanations reveal how the individual input variables influence the outcome of otherwise completely opaque models, thus making them transparent and comprehensible. The general methods allow users to compare different types of models or replace their existing model without having to replace the explanation method. The explanation methods EXPLAIN, IME, and LIME exhibit the following properties:

- *Instance dependency:* different instances are predicted differently, so the explanations will also be different.
- *Class dependency:* explanations for different classes are different, different attributes may have a different influence on different classes (for two-class problems, the effect is complementary).
- *Model dependency:* the methods explain a given model, so if the model is wrong for a given instance, the produced explanations will reflect that.
- *Capability to detect strong conditional dependencies:* if the model captures strong conditional dependencies, the explanations will also reflect that.
- *Visualisation ability:* the generated explanations can be graphically presented in terms of the positive/negative effect each attribute and its values have on the classification of a given instance.
- *Local fidelity:* the perturbation based approaches perturb instances in the neighbourhood of explanation instance, therefore they are sensitive to the model's functioning in the local context.

- *Efficiency:* methods EXPLAIN and LIME can be efficiently used with a large number of instances and features, while current implementation of IME is limited to a relatively low number of features (up to 100).
- *Fair contributions:* only 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.
- *Availability:* the software implementation of the explanation methodology is available as the open-source R package *ExplainPrediction* [28]. Furthermore, the real-world B2B sales forecasting data set is publicly accessible [10].

We can list the following limitations which can spur further improvements:

- EXPLAIN method is unable to detect and correctly evaluate the utility of attributes' values in instances where the change in more than one attribute value at once is needed to affect the predicted value. IME method samples the space of feature interactions and therefore avoids this problem.
- IME suffers from relatively large computational load required to reach probabilistic guarantees of its performance. The explanations would have to be precomputed in order to be used interactively in a discussion session and computations may be too slow for high dimensional problems.
- While efficient in high dimensional spaces, LIME offers no guarantees that the explanations are faithful, and ignores the required number and nature of obtained samples. By using uniform sampling in the proximity of explanation instance, it may be susceptible to problems of neighbourhood-based methods like kNN in high dimensional spaces. The problem of possible feature interactions is also not adequately solved.
- The interactions between attributes are captured but not expressed explicitly in the visualisations; therefore, the user has to manually discover the type of interdependencies with interactive analysis.

In a business context, we presented an extract from the successful grounded application of machine learning models coupled with general explanation methodology. On the complex real-world 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. For new (open) cases, we demonstrated interactive support for decision makers, assessing various scenarios with explanatory what-if analysis. We presented flexibility of the methodology to address a specific business request (weak performance in one segment). The explanations of the prediction models and what-if 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 simplicity and elegance of the perturbation-based explanations coupled with efficient implementations and visualisation of instance- and model-based explanations allow application of general explanation approaches to many areas. We expect that broader practical use will spur additional research into explanation mechanisms

and improvements in the visual design of explanations. Machine learning based automatic decisions have already spread to many areas of life and attracted attention of the general public and law-makers, who demand its better transparency. This makes model explanation a much needed and attractive research and application topic.

Acknowledgements We are grateful to the company Salvirt, ltd., for funding a part of the research and development presented in this paper. Marko Robnik-Šikonja was supported by the Slovenian Research Agency, ARRS, through research programme P2-0209.

References

1. Aggarwal, C.C., Chen, C., Han, J.: The inverse classification problem. Journal of Computer Science and Technology **25**(3), 458–468 (2010)
2. Andrews, R., Diederich, J., Tickle, A.B.: Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-Based Systems **8**(6), 373–384 (1995)
3. Armstrong, J.S., Green, K.C., Graefe, A.: Golden Rule of Forecasting: Be conservative. Journal of Business Research **68**(8), 1717–1731 (2015)
4. Arnold, V., Clark, N., Collier, P.A., Leech, S.A., Sutton, S.G.: The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. MIS Quarterly pp. 79–97 (2006)
5. Arras, L., Horn, F., Montavon, G., Müller, K.R., Samek, W.: What is relevant in a text document?: An interpretable machine learning approach. PloS ONE **12**(8), e0181,142 (2017)
6. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS ONE **10**(7), e0130,140 (2015)
7. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.R.: How to explain individual classification decisions. Journal of Machine Learning Research **11**(Jun), 1803–1831 (2010)
8. Barbella, D., Benzaid, S., Christensen, J.M., Jackson, B., Qin, X.V., Musicant, D.R.: 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 (2009)
9. Barbosa, A., Paulovich, F., Paiva, A., Goldenstein, S., Petronetto, F., Nonato, L.: Visualizing and interacting with kernelized data. IEEE transactions on visualization and computer graphics **22**(3), 1314–1325 (2016)
10. Bohanec, M.: Anonymized B2B sales forecasting data set (2016). URL `http://www.salvirt.com/research/b2bdataset`
11. Bohanec, M., Borštnar Kljajić, M., Robnik-Šikonja, M.: Explaining machine learning models in sales predictions. Expert Systems with Applications **71**, 416–428 (2017)
12. Bohanec, M., Robnik-Šikonja, M., Kljajić Borštnar, M.: Decision-making framework with double-loop learning through interpretable black-box machine learning models. Industrial Management & Data Systems **117**(7), 1389–1406 (2017)
13. Bosnić, Z., Demšar, J., Kešpret, G., Rodrigues, P.P., Gama, J., Kononenko, I.: Enhancing data stream predictions with reliability estimators and explanation. Engineering Applications of Artificial Intelligence **34**, 178–192 (2014)
14. Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., Zupan, B.: Orange: Data mining toolbox in python. Journal of Machine Learning Research **14**, 2349–2353 (2013)

15. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics **24**(1), 44–65 (2015)
16. Hamel, L.: Visualization of support vector machines with unsupervised learning. In: Proceedings of 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (2006)
17. Hansen, K., Baehrens, D., Schroeter, T., Rupp, M., Müller, K.R.: Visual interpretation of kernel-based prediction models. Molecular Informatics **30**(9), 817–826 (2011)
18. Jacobsson, H.: Rule extraction from recurrent neural networks: A taxonomy and review. Neural Computation **17**(6), 1223–1263 (2005)
19. Jakulin, A., Možina, M., Demšar, J., Bratko, I., Zupan, B.: 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 (2005)
20. Lemaire, V., Féraud, R., Voisine, N.: Contact personalization using a score understanding method. In: Proceedings of International Joint Conference on Neural Networks (IJCNN) (2008)
21. Lughofer, E., Richter, R., Neissl, U., Heidl, W., Eitzinger, C., Radauer, T.: Explaining classifier decisions linguistically for stimulating and improving operators labeling behavior. Information Sciences (2017)
22. Martens, D., Provost, F.: Explaining documents' classifications. Tech. rep., Center for Digital Economy Research, New York University, Stern School of Business (2011). Working paper CeDER-11-01
23. Meyer, D., Leisch, F., Hornik, K.: The support vector machine under test. Neurocomputing **55**, 169–186 (2003)
24. Piltaver, R., Luštrek, M., Gams, M., Martinčić-Ipšić, S.: What makes classification trees comprehensible? Expert Systems with Applications **62**, 333–346 (2016)
25. Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D.S., Fyshe, A., Pearcy, B., Macdonell, C., Anvik, J.: Visual explanation of evidence with additive classifiers. In: Proceedings of AAAI'06. AAAI Press (2006)
26. Pregeljc, M., Štrumbelj, E., Mihelčič, M., Kononenko, I.: Learning and Explaining the Impact of Enterprises Organizational Quality on their Economic Results. In: R. Magdalena-Benedito, M. Martínez-Sober, J.M. Martínez-Martínez, P. Escandell-Moreno, J. Vila-Francés (eds.) Intelligent Data Analysis for Real-Life Applications: Theory and Practice, pp. 228–248. Information Science Reference, IGI Global (2012)
27. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM (2016)
28. Robnik-Šikonja, M.: ExplainPrediction: Explanation of Predictions for Classification and Regression (2017). URL `http://cran.r-project.org/package=ExplainPrediction`. R package version 1.1.9
29. Robnik-Šikonja, M., Kononenko, I.: Explaining classifications for individual instances. IEEE Transactions on Knowledge and Data Engineering **20**(5), 589–600 (2008)
30. Saltelli, A., Chan, K., Scott, E.M.: Sensitivity analysis. Wiley, Chichester; New York (2000)
31. Schulz, A., Gisbrecht, A., Hammer, B.: Using discriminative dimensionality reduction to visualize classifiers. Neural Processing Letters **42**(1), 27–54 (2015)
32. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
33. Štrumbelj, E., Bosnić, Z., Kononenko, I., Zakotnik, B., Kuhar, C.G.: Explanation and reliability of prediction models: the case of breast cancer recurrence. Knowledge and information systems **24**(2), 305–324 (2010)
34. Štrumbelj, E., Kononenko, I.: An Efficient Explanation of Individual Classifications using Game Theory. Journal of Machine Learning Research **11**, 1–18 (2010)
35. Štrumbelj, E., Kononenko, I., Robnik-Šikonja, M.: Explaining instance classifications with interactions of subsets of feature values. Data & Knowledge Engineering **68**(10), 886–904 (2009)