Improving Reliability Estimation for Individual Numeric Predictions: A Machine Learning Approach

Gediminas Adomavicius   Yaqiong Wang
Department of Information and Decision Sciences
Carlson School of Management, University of Minnesota
gedas@umn.edu, wang6285@umn.edu

Abstract: Numerical predictive modeling is widely used in different application domains. While many modeling techniques have been proposed, and a number of different aggregate accuracy metrics exist for evaluating the overall performance of predictive models, other important aspects, such as the reliability (or confidence, uncertainty) of individual predictions, have been underexplored. We propose to use *estimated absolute prediction error* as the indicator of individual prediction reliability, which has the benefits of being intuitive and providing highly interpretable information to the decision makers as well as allowing for more precise evaluation of reliability estimation quality. As importantly, the proposed reliability indicator allows to reframe reliability estimation itself as a canonical numeric prediction problem, which makes the proposed approach general-purpose (i.e., can work in conjunction with any outcome prediction model), alleviates the need for distributional assumptions, and enables the use of advanced, state-of-the-art machine learning techniques to learn individual prediction reliability patterns directly from data. Extensive experimental results on multiple real-world datasets show that the proposed machine-learning-based approach can significantly improve individual prediction reliability estimation as compared to a number of baselines from prior work, especially in more complex predictive scenarios.

Keywords: Numeric prediction, reliability of individual predictions, machine learning

1. Introduction and Motivation

Many critical decisions in real world rely on predictions, e.g., investors forecast returns, doctors diagnose diseases, producers predict sales. Facilitated by continuous improvements in data processing and storage technologies, this has spurred development and improvement of machine learning and, more generally, predictive modeling techniques. However, these automated predictions are often imperfect because they are made from noisy, limited data or using simplified computational or probabilistic reasoning.

For numeric prediction tasks, predictive models focus primarily on providing *individual* prediction outcomes; for example, a diabetes risk estimation model would output the risk score of diabetes for each potential patient. Meanwhile, the quality of predictive models is commonly evaluated using *aggregate* prediction accuracy metrics, such as mean absolute error or root mean squared error, calculated on some test set of data. The issue of *individual prediction reliability* (IPR), i.e., the magnitude of error or level of uncertainty of any *specific* individual prediction, has not been explored as comprehensively. When applying properly trained models, i.e., models with best possible aggregate accuracy, to real-world data, the ability to provide reliability estimation for any specific prediction is undoubtedly important, especially for the purpose of facilitating decision support. As an example, let's assume that, when estimating the severity of Parkinson's disease for two individual patients using Parkinson's Disease Rating Scale (Tsanas et al. 2010), both patients are predicted to have the same rating score of 123, i.e., the same predicted disease severity. At the same time, the prediction reliability could be highly different for numerous reasons, e.g., because these two patients belong to highly different age groups for which different amounts of data are available. For example, it is possible that the prediction of 123 for a younger patient means that the true disease rating value likely is 123 ± 30 (i.e., between 93 and 153), while the same prediction for an older patient might be much more reliable, i.e., 123 ± 5. The diagnosis reliability information is important for deciding on individualized treatment, yet is not captured by the predicted outcome (i.e., 123) alone.

As another simple illustration of the research context, consider the stylized, synthetically generated data[1] in Fig. 1, where X axis represents the input variable and Y axis represents the outcome to be predicted. Specifically, the black dots represent data points (x, y), and the solid red line represents the estimated linear regression model $\hat{y} = f(x)$ that is used for prediction. Although the linear regression model represents the most accurate predictive model for this dataset (as this dataset was generated with this purpose in mind), it is easy to see that the predictions for $x \in$ [-0.5, 0.5] are much less reliable in a given setting, i.e., prediction errors $e = |y - \hat{y}| = |y - f(x)|$ for individual data points in this area are typically much higher than for $x \notin$ [-0.5, 0.5].

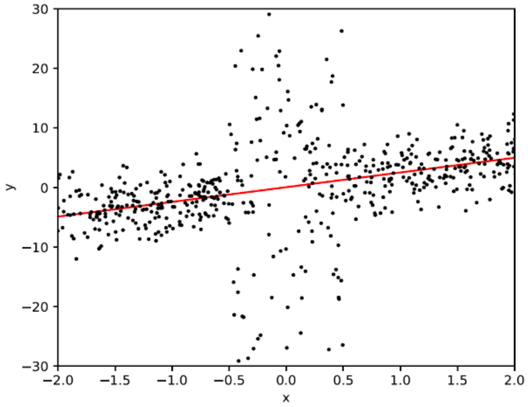


Figure 1. Synthetic Data Example for Prediction Reliability Issue.
(X axis: input variable. Y axis: outcome variable. Dots: data points. Solid line: predictive model based on linear regression.)

In other cases, where the true data generating process cannot be accurately recovered, the prediction errors can result not only from the random noise, which typically leads to the variance of outcome predictions, but also from the misfit of the models which leads to systematic bias of outcome predictions (Domingos 2000, Geman et al. 1992). It is also important to reiterate that individualized prediction reliability estimates are not captured by traditional aggregate accuracy metrics, yet this knowledge can be critical in many real-world numeric prediction applications, for example, in risk-sensitive areas where decision based on an individual prediction would entail health or financial consequences. For example, when predicting the 10-year risk score of a cardiovascular disease for a specific patient or the 3-year return of a stock portfolio for a specific customer, it would be important to know not only the actual prediction but also the estimated reliability of such a prediction before making a final decision about medical treatment or financial

[1] 2000 points were created by generating their x values uniformly at random from [-2, 2], and their corresponding y values were generated using function $y = 2.5x + \varepsilon$, $\varepsilon \sim N(0,\sigma^2)$. In particular, $\sigma = 2$ for $x \in [-2,-0.5] \cup [0.5,2]$; $\sigma = 10$ for $x \in [-0.5, 0.5]$.

investment. In summary, by design, the goal of individual prediction reliability is not to be another metric that needs to be balanced together (e.g., as part of the machine learning loss function) with the overall model accuracy, but rather to provide diagnostic information to decision makers who use a given outcome prediction model, i.e., providing not only the model's prediction for a given input, but also the indication of how reliable each specific prediction is expected to be.

It should be mentioned that prediction reliability has been referred to in different ways in previous literature: prediction risk, prediction uncertainty, prediction confidence, etc. We draw on (Bosnić and Kononenko 2008a) to use *prediction reliability* for the sake of terminological consistency. Reliability estimation has been used for two main purposes. One line of research uses estimated reliability as an additional criterion (e.g., in conjunction with accuracy-based metrics) for model evaluation, where models with higher prediction reliability are typically more preferred. Different methods have been proposed for estimating prediction reliability for this purpose, e.g., cross-validation, bootstrapping, Bregman divergence, covariance-based (Efron 2004, Shao 1996, Taylor and Ye 2012). Similar to aggregate accuracy metrics mentioned before, reliability estimated in this type of work is still used as an aggregate model evaluation tool. The other line of research uses prediction reliability for individual prediction explanation or description, which is directly aligned with the focus of this paper. Those studies fall into three finer-grained groups based on the type of outcome to be predicted, i.e., reliability for a single example in classification, probability estimation, or numeric prediction. In this paper, we focus specifically on reliability of numeric prediction models (as will be discussed in the next section), which has been significantly underexplored in research literature, as compared to reliability estimation for other types of outcomes. For example, reliability of probability estimation is often measured by Brier score (Brier 1950) which is calculated as the squared difference between actual outcome (binary or categorical) and predicted probability assigned to that outcome. There have been numerous studies investigating individual classification reliability. For some classifiers, like logistic regression or naïve Bayes (Hand and Yu 2001, Walker and Duncan 1967), the posterior probability of an individual predicted class can be viewed as confidence (reliability) of its prediction. Most related studies propose more general (model-agnostic) approaches, e.g., transductive reliability estimation (Kukar and Kononenko 2002, Tzikas et al. 2007) drawing on transduction based confidence estimation (Ho and Wechsler 2003, Proedrou et al. 2002, Saunders et al. 1999) or the *typicalness* framework (Melluish et al. 2001, Nouretdinov et al. 2001).

Even though the reliability estimation for numeric prediction models has been significantly underexplored in research literature, it undoubtedly represents an increasingly important issue due to the needs for more fine-grained understanding of predictive model performance, as will be discussed in next section. Therefore, going beyond the evaluation of the overall (i.e., aggregate) accuracy performance of numeric prediction models, in this study we focus on providing a *general-purpose*, data-driven approach to *individual prediction reliability* (**IPR**)[2] estimation. In particular, we propose to use a simple IPR indicator based on expected *absolute prediction errors*, which has the benefits of being intuitive and providing highly interpretable information to the decision makers as well as allowing for more precise evaluation of reliability estimation quality. Even more importantly, the proposed IPR indicator also allows us to reframe reliability estimation itself as a canonical numeric prediction problem (of the absolute prediction error), which makes the proposed approach general-purpose (i.e., can work in conjunction with any outcome prediction model), alleviates the need for any statistical/distributional assumptions, and enables the use of advanced, state-of-the-art machine learning techniques to learn IPR patterns directly from data. Advantages of the proposed approach are demonstrated using comprehensive computational experiments on several real-world datasets and in comparison to multiple techniques from prior work.

2. Related Work

Given the popularity of (and reliance on) predictive modeling techniques in many aspects of everyday life, in general a more comprehensive and nuanced understanding of predictive model performance represents an increasingly important issue. Ability to provide IPR estimates is an important aspect for both application and interpretation of predictive models (Briesemeister et al. 2012, Bosnić and Kononenko 2008a, Shrestha and Solomatine 2006). In particular, for a given

[2] We use acronym IPR to refer to “individual prediction reliability” throughout the paper.

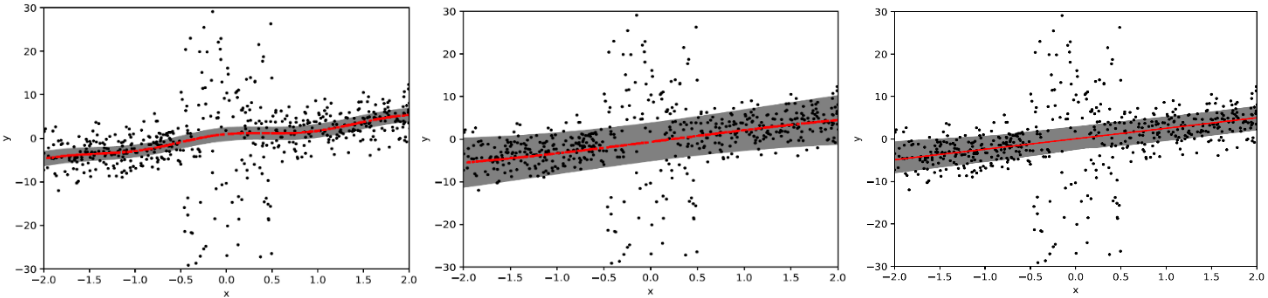
predictive model, IPR estimates would provide better understanding for which data points the model is expected to perform better vs. worse (i.e., have higher vs. lower reliability). This connects well to the topic of *error analysis*, which helps to find opportunities for substantial increase in predictive performance. For example, in biomedical informatics, the error models of individual cells can discern new subpopulations within complex mixtures of cells and derive more robust measures for cell classification (Kharchenko et al. 2014). In medical diagnosis, analyzing inaccurate predictions are important to find out what cases can confuse machine learning models even when the overall predictive performance is impressive (Choi et al. 2016). In biological natural language processing (Hakala et al. 2013), analyzing inaccurate predictions helps diagnosing whether false predictions of the event type (e.g., gene expression, transcription, etc.) is due to missing or incorrectly constructed features. In speech recognition (Qian et al. 2018), error analysis is used to identify top types of errors (substitution, deletion, etc.) that the system makes under different noise contexts, which is valuable in informing prediction application as well as system adaptation. In online recommender systems, examining rating prediction errors at individual level can inform designing of meta-learning algorithms for different users or user groups (Collins et al. 2018). IPR estimates are also relevant to the important research topic of algorithmic bias (Datta et al. 2015, Simoiu et al. 2016, Hosanagar 2019, Johndrow and Lum 2019), as they could provide early detection signals of potential systematic bias of predictive models. Finally, as mentioned earlier, IPR provides extra information, which is important for facilitating better decision making across a broad array of applications in chemical and pharmaceutical research (Briesemeister et al. 2012, Liu et al. 2018, Toplak et al. 2014, Cortés-Ciriano and Bender 2018), financial markets (Dash et al. 2015, Huang et al. 2018, Solares et al. 2019), medical diagnosis (Lebedev et al. 2014, Iorio et al. 2015, Tomassetti et al. 2016), and many others.

In terms of methodologies for the IPR representation and calculation, traditional approaches could be summarized into two broad categories: (i) *distribution-based*, i.e., estimating an entire distribution of the outcome variable predictions for any given input value *x*, which can then be provided to the decision makers directly or in some aggregate form (such as confidence interval) as information about prediction reliability or confidence, and (ii) *indicator-based*, i.e., providing a simple, single-numeric-value-based indicator of IPR for given *x*, often based on some heuristic.

Distributional, or confidence-interval-based (Wonnacott and Wonnacott 1990), approaches are rooted in statistical properties of prediction models, especially regression models, and represent an intuitive way to indicating IPR – predictions with wider confidence intervals (for a given confidence level) indicate higher model uncertainty. Distributional approaches also tend to be model-specific, i.e., designed specifically for a particular outcome prediction model, and rely on certain statistical assumptions. In both least-squares-based and likelihood-based learning of regression models, generation of confidence intervals or other confidence metrics is based on the assumption of independent and identical distribution of errors across the input space (i.e., homoscedasticity) (Halpe 1963, Knafl et al. 1985). However, this homoscedasticity assumption is usually violated in many real-world settings which is explicitly the focus of this study (reflecting situations similar to the one illustrated in Fig. 1), and thus, the derived confidence intervals would fail to reflect actual IPR. More sophisticated regression-based distributional approaches draw on the flexible Gaussian process (Rasmussen 2004), which allows to incorporate information on similarity between data points into the model building. Although the probabilistic Gaussian process regression model facilitates the derivation of predictive distribution for the regression outcome, the key characteristic of this modeling technique is that the variance of the distribution for new observation *x* (i.e., the indicator of its prediction reliability) *only depends on the input features of x* and, in particular, on the relative location (e.g., distance calculated using feature values) of *x* to other observations in the training data, and *not on the observed target (outcome) values* (Rasmussen 2004).[3] Because of the latter fact, it is unlikely to capture the magnitude of error in the prediction that is due to variability in the outcomes, which makes it a less informative measurement of IPR. Figure 2a emphasizes this by presenting the 95% prediction intervals of Gaussian process regression learned from the synthetic dataset used in Fig. 1 – the widths of

[3] More detailed discussion of this issue can be found in Appendix A of the Online Supplement.

individual prediction intervals are similar across the input (x) space, not reflecting the actual variability in the outcomes. There have been other distributional approaches that extend certain specific learning techniques to make predictions together with corresponding probabilistic reliability estimates (Khosravi et al. 2010, Papadopoulos et al. 2001). For example, Hwang et al. (1997) use an asymptotic approach to build confidence intervals for neural networks; however, similarly to what has been discussed above, due to traditional statistical assumptions on errors (e.g., homoscedasticity) and model parameters, the prediction intervals generated by this approach are not designed to reflect the variability in the actual outcomes (but rather the variability in model predictions), as illustrated in Figure 2b on the same stylized dataset.



(a) Gaussian Process-based prediction interval (b) Neural Networks-specific prediction interval (c) Bootstrapping-based prediction interval

Figure 2. Individual Prediction Reliability Representation Based on 95% Confidence / Prediction Interval
(X axis: input variable. Y axis: outcome variable. Dots: data points. Solid line: predictive model based on different techniques.)

Another standard approach to construct prediction distributions and corresponding prediction intervals is bootstrapping (Efron 1979). Such an approach has some clear benefits in that it can be used with all kinds of predictive models (i.e., it is not model-specific) and generate distributions without having to rely on statistical assumptions; however, heteroscedasticity still poses a significant challenge for bootstrapping-derived IPR representation in certain situations. To illustrate, in Fig. 2c we plot confidence intervals obtained from this approach on data presented in Fig. 1.[4] Obviously, due to the similarity of data patterns in the bootstrap samples, the predictions of *all* linear models across the entire input space would be similar. This means that, across the entire range of input values (x), the width of point-wise confidence intervals derived from the

[4] Each bootstrap sample was generated from original training data by randomly sampling (with replacement) 500 data points, on which a linear regression model was learned and then applied to make predictions for test data. We repeat this process 100 times – resulting in 100 predictions for each data point, from which 95% prediction interval is empirically constructed.

prediction variance would be similar too and not reflective of the actual underlying variability of data, as indicated in Fig. 2c.

To summarize, the existing distributional approaches (many of which are model-specific and rely on restrictive statistical assumptions) have been designed mainly to reflect the distribution of model predictions and, therefore, are less well-suited for capturing the actual underlying variability of ground truth data (and, hence, actual prediction errors), i.e., for capturing IPR in heteroscedastic environments. As an alternative, a number of prior studies addressed this issue by turning to simpler, yet more flexible, indicator-based approaches to IPR estimation, which we discuss next.

Indicator-based approaches typically represent general-purpose (i.e., applicable with any outcome prediction model, free from statistical assumptions) IPR estimators that provide a simple numeric value as an indicator of IPR for any individual input value. Among these approaches, early work focused on using nonparametric bootstrapping techniques (Carney et al. 1999, Heskes 1997) and summarizing the individual prediction variability across samples (e.g., by using confidence/prediction interval widths or prediction variance) as reliability indicators, or estimating errors based on the covariance among data points (Efron 2004). Several other methods are based on heuristics that try to exploit local information of individual data points in order to directly capture the actual variability of underlying data, e.g., using prediction errors (Briesemeister et al. 2012), prediction variance of the nearest neighbors of the focal data point (Clark 2009), or the density of the input space in close proximity to the focal data point (Bosnić and Kononenko 2008a), as surrogates of IPR. These approaches are based on intuition that the uncertainty of individual predictions should be higher around data points with high prediction errors or high prediction variance, or for points around which there is not much training data available. One can see that some of these heuristics – in particular, density-based – would not be very useful in heteroscedastic settings (such as the one illustrated by Fig. 1). Somewhat similarly, Shrestha and Solomatine (2006) propose to partition the input space into different clusters and then construct prediction intervals based on the empirical distributions of the errors associated with instances in the same cluster. In terms of specific IPR indicators, Briesemeister et al. (2012) designed two statistics based on the local properties of training data, while another related study (Bosnić and Kononenko 2008b) proposed several empirical measures based on sensitivity analysis.

For our computational experiments, we use nine commonly used indicator-based reliability approaches as baselines for comparison: VarBag (Breiman 1996), VarA, MSE (Briesemeister et al. 2012), VarP, AvgDiff (Bosnić and Kononenko 2008a), AvgDist (Sheridan et al. 2004), LCV (Demut 2010), SAV and SAB (Bosnić and Kononenko 2008b). The relevant notation and the formal definitions of these approaches are provided in Tables 1 and 2, respectively; note that all measures are calculated for a given individual data point (x, y), where x represents an input feature vector and y is an outcome (target) value. We narrowed down our choice to this particular set of approaches as most promising baseline candidates due to their potential flexibility for capturing IPR in heteroscedastic environments and for their advantageous performance reported in prior studies and observed in our pilot experiments.

Finally, evaluation is necessary to test and compare the effectiveness of different methods for IPR estimation. As observed in prior literature, for an IPR indicator to be meaningful and useful, the estimates that it produces should be "aligned" with actual individual prediction errors; i.e., predictions estimated to be more reliable should exhibit smaller errors (and vice versa). Based on this intuition, previous studies typically use the correlation coefficient (between the reliability estimates and actual prediction errors) as the measure of "alignment" to evaluate the performance of proposed IPR indicators for numeric prediction models (Bosnić and Kononenko 2009, Briesemeister et al. 2012), where higher correlation indicates better IPR estimation performance.

In summary, the general structure and contribution of many existing individual prediction reliability estimation studies can be outlined as: (i) defining some reliability indicator; (ii) demonstrating how it can be computed/derived; and (iii) showing its quality by showing that its values are well "aligned" with the actual outcome prediction errors (using some "alignment" measure, typically correlation coefficient). Our study follows the same general structure to provide further improvements to the current state of the art in this area, as discussed in the next section.

Table 1. Common Notations for Describing Reliability Estimation Methods

Symbol	Definition
$\boldsymbol{x}$	input vector (of different features) of a given example
$\boldsymbol{y}$	outcome value of a given example
$\boldsymbol{x_i}$	input vector of ith nearest neighbor in heuristic-based methods
$\boldsymbol{y_i}$	actual outcome of ith nearest neighbor in heuristic-based methods
$\boldsymbol{\hat{y}_i}$	predicted outcome of ith nearest neighbor in heuristic-based methods
$\boldsymbol{\varepsilon_i}$	$\varepsilon_i = y_i - \hat{y}_i$ prediction error of ith nearest neighbor in heuristic-based methods
$\boldsymbol{m}$	number of random samples in bootstrapping-based methods
$\boldsymbol{M_j}$	prediction for x made by the model learned from the jth sample in bootstrapping-based methods
$\boldsymbol{n}$	number of nearest neighbors selected in heuristic-based methods
$\boldsymbol{d(x_i, x)}$	distance between the example x and its ith nearest neighbor in heuristic-based methods
$\boldsymbol{\hat{y}_{-i}}$	leave-one-out prediction of ith nearest neighbor in heuristic-based methods
$\boldsymbol{\tau}$	sensitivity parameters ($\tau \in [0,1]$)
$\boldsymbol{S}$	set of sensitivity parameters. An example of S = {0.01, 0.1, 0.5, 1}
$\boldsymbol{t_{max}\,/\,t_{min}}$	Maximum/minimum value of outcome in the training data
$\boldsymbol{\hat{y}_\tau}$	predicted outcome of x using models trained using training data (X, Y) plus augmented sample of $(\, x,\, y + \tau * (t_{max} - t_{min}))$ in sensitivity based methods
$\boldsymbol{\hat{y}_{-\tau}}$	predicted outcome of x using models trained using training data (X, Y) plus augmented sample of $(\, x,\, y - \tau * (t_{max} - t_{min}))$ in sensitivity based methods

Table 2. Description of Baseline Reliability Estimation Methods

Baseline	Calculation and Description
VarBag	$\frac{1}{m}\sum_{j=1}^{m}(M_j - \hat{y})^2$,$\hat{y} = \frac{\sum_{j=1}^{m} M_j}{m}$. Variance of example x's predictions M_j s made by models learned from different random samples.
VarA	$\frac{1}{n}\sum_{i=1}^{n}(\bar{y} - y_i)^2$,$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Variance of example x's nearest neighbors' actual values (y_is).
VarP	$\frac{1}{n}\sum_{i=1}^{n}(\bar{\hat{y}} - \hat{y}_i)^2$,$\bar{\hat{y}} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i$. Variance of example x's nearest neighbors' predictions ($\hat{y}_i$s).
AvgDiff	$\lvert \frac{\sum_{i=1}^{n} y_i}{n} - \hat{y} \rvert$. Difference between the average of nearest neighbors' actual values (y_is) and the example x's prediction ($\hat{y}$).
MSE	$\frac{1}{n}\sum_{i=1}^{n} {\varepsilon_i}^2$,$\varepsilon_i = y_i - \hat{y}_i$. Mean squared error of example x's nearest neighbors' predictions ($\hat{y}_i$s).
AvgDist	$\frac{1}{n}\sum_{i=1}^{n} d(x_i, x)$. Average distance between the example x and its nearest neighbors (x_is).
LCV	$\frac{\sum_{i=1}^{n} d(x_i,x)*E_i}{\sum_{i=1}^{n} d(x_i,x)}$,$E_i = \lvert y_i - \hat{y}_{-i} \rvert$. Weighted average errors of nearest neighbors' leave-one-out predictions ($\hat{y}_{-i}$).
SAV	$\frac{\sum_{\tau \in S}(\hat{y}_\tau - \hat{y}_{-\tau})}{\lvert S \rvert}$. Average difference between sensitivity predictions $\hat{y}_\tau$ and $\hat{y}_{-\tau}$ over different sensitivity parameters in set S.
SAB	$\frac{\sum_{\tau \in S}(\hat{y}_\tau - \hat{y}) + (\hat{y}_{-\tau} - \hat{y})}{2 * \lvert S \rvert}$. Average difference between sensitivity predictions ($\hat{y}_\tau$ or $\hat{y}_{-\tau}$) and original prediction ($\hat{y}$).

3. Machine Learning Approach to Individual Prediction Reliability Estimation

3.1 Proposed Individual Prediction Reliability Indicator: General Overview

In this study, we propose a novel indicator-based approach to IPR representation and calculation. The key motivation for the specific proposed method was the observation that, while the existing IPR indicators have been defined in a variety of different ways (e.g., as variance or density of certain data, etc.), their performance is always judged by how well the IPR estimates are aligned with actual errors of the outcome prediction model. Therefore, we propose to use a simple and intuitive reliability indicator that is designed to be directly related to errors of the outcome prediction model, more specifically, an indicator based on *expected absolute prediction error* for a given individual prediction.

The proposed idea provides a number of significant benefits. First and foremost, it provides a way to reframe IPR estimation as a canonical numeric prediction problem (of the absolute prediction error). That is, while many IPR indicators from prior work are "constructed" based on heuristics and/or distributional assumptions, and their performance is validated empirically afterwards, in our case the individual actual outcome prediction errors serve as ground truth (targets/labels) to the wide variety of existing advanced machine learning techniques that are able to directly learn complex relationships between input features and prediction reliability.

More specifically, prediction uncertainty can come from different sources that are often hard to disentangle, e.g., random noise/variability, inappropriate model selection, or suboptimal model parameters. The actual prediction errors, i.e., the discrepancies between observed y and prediction $\hat{y}$ from some predictive model, provide the most reliable signals of the level of prediction uncertainty. Higher prediction error typically indicates lower IPR, and prediction errors could arguably be used in at least two distinct ways under different contexts.

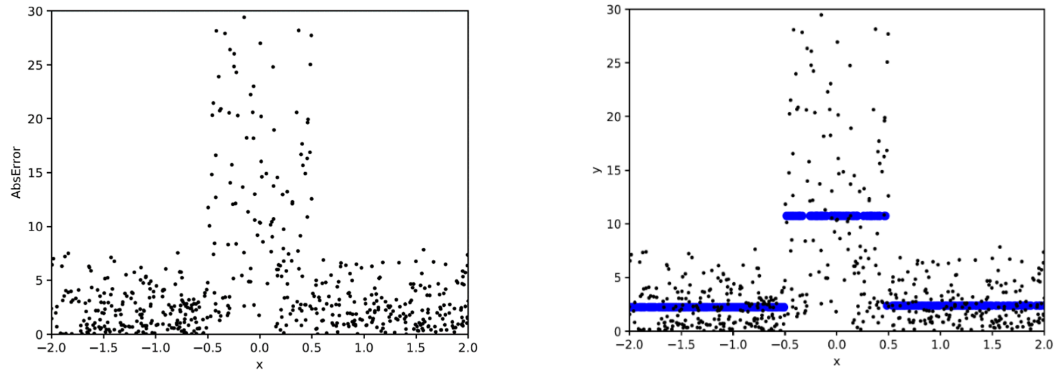
In particular, one could use *absolute* prediction errors, i.e., $e = |\hat{y} - y|$, vs. *direct* prediction errors, i.e., $e = \hat{y} - y$; the latter would reflect not only the absolute magnitude of discrepancy, but also its *direction*, in other words, whether y is overestimated or underestimated by the outcome prediction model. In this study, we focus on the absolute-error-based IPR indicator for the following key reason. The situations where the model's prediction errors are highly imbalanced (model under-predicts and over-predicts in numerous portions of the data space) typically reflect the fact that the outcome prediction model poorly represents the underlying generative process of the data (i.e., the model is biased, poorly fit), and the first goal typically is to improve the overall model fit. These situations often can be readily diagnosed with standard, traditional aggregate model evaluation metrics; of course, such situations could be remedied by looking at direct errors as well, and there are entire machine learning approaches dedicated to that.[5] However, once the outcome prediction model fit is improved using detected systematic direct errors (no significant over- or under-prediction), the remaining patterns of direct errors would be impossible to learn (essentially being random noise of different magnitudes), yet the key IPR problem as stated in the paper would still be highly relevant (as motivated by Fig. 1). And, insightful and actionable IPR information can still be mined from data.

Thus, abstracting away from the directionality of errors, we propose to view the reliability of a given individual prediction as the expected *absolute* prediction error. As mentioned earlier, this allows us to address the IPR estimation problem as a canonical, *meta-algorithmic* numeric prediction problem, i.e., it can to use any advanced machine learning technique for reliability estimation. More specifically, IPR represented by absolute prediction error, i.e., $e = |\hat{y} - y|$, could be directly modeled as a function of input variables x, i.e., as $e = F(x)$, to capture the structural relationships between the input space and the prediction reliability for *any* given outcome prediction model. Building machine learning model F (i.e., the reliability estimator) does require labeled training data $\{(x, e)\}$. It is important to point out that this data usually is readily available, because the outcome prediction models (i.e., models predicting $\hat{y}$) are typically evaluated on some *hold-out test data* $\{(x, y)\}$ which can then be straightforwardly reused to construct the ground truth for reliability estimation; i.e., every data point (x, y) together with

[5] In cases where prediction errors are not balanced (i.e., when the outcome prediction model is significantly biased), there are substantial opportunities to improve the outcome predictive models themselves first, before performing reliability estimation. In fact, some boosting-based machine learning techniques, e.g., XGBoost (Chen and Guestrin 2016), use this idea: they build an ensemble of models sequentially one-by-one and take advantage of the unbalanced direct prediction errors from models learned in earlier stages iteratively to improve the ultimate outcome prediction performance of the entire ensemble model.

corresponding outcome prediction $\hat{y}$ can be converted to (x, e), where $e = |\hat{y} - y|$.

Taking the data shown in Fig. 1 as an example, the absolute prediction error (and, hence, the IPR) of the best outcome prediction model is consistently higher in certain areas. This is illustrated in Fig. 3a, where x axis still represents the input features, while vertical axis now represents absolute error (i.e., e) of the outcome prediction model. As the figure shows, the absolute prediction error is much higher within interval $x \in$ [-0.5, 0.5] than elsewhere, which can be learned by machine learning techniques. For example, using the data plotted in Fig. 3a, a regression tree model can learn to predict e from x, and we show the pointwise prediction errors estimated from this regression tree in Fig. 3b. Each blue dot in Fig. 3b represents an estimated absolute prediction error for given x, which shows that the prediction of the errors, i.e., the IPR indicators ($\hat{e}$), are quite informative. As can be seen in Fig. 3b, estimated reliability is able to accurately differentiate the levels of outcome model's prediction uncertainty across different intervals, i.e., the uncertainty is lower for $x \in$ [-2.0, -0.5] and $x \in$ [0.5, 2.0] and higher for $x \in$ [-0.5, 0.5].



(a) Actual absolute prediction error

(b) Regression-tree-based error estimation

Figure 3. Pointwise Prediction Error Estimation of Linear Regression Model from Fig. 1.
(X axis: input variable. Y axis: absolute prediction error. Black dots: actual abs. prediction error. Blue dots: estimated abs. prediction error.)

It is important to reiterate that reframing IPR estimation as a data-driven numeric prediction problem makes the proposed approach *general-purpose* (i.e., reliability estimation can be done for any outcome prediction model) and alleviates the need for distributional modeling assumptions. An added benefit of the proposed IPR indicator is its clear *interpretability* to end-users and decision makers, which may not be the case with some existing approaches that require probabilistic assumptions (e.g., distribution-based approaches) and non-intuitive quantifications (e.g., density-based heuristic indicators). Specifically, the reliability score of a given prediction simply represents the expected absolute error for this prediction, along the lines of "for given *x*, the outcome prediction model is expected to be off by this much, on average".

Finally, the proposed approach also allows for a more precise and informative evaluation. As mentioned earlier, a popular reliability evaluation metric has been the *correlation* between IPR values and actual prediction errors. Even though correlation coefficient is not a very precise measure in the sense that it captures only very high-level patterns (general trends), it has been widely used largely because the existing IPR indicators have been defined in highly differing ways (as variance, density, or average of certain data, etc.) – a more direct comparison to actual prediction errors was not feasible. In other words, even with high correlation, it is possible that the magnitude of IPR estimates might be significantly different than the one of the actual errors, thus, reducing diagnosticity (or interpretability) of IPR indicators. In contrast, the proposed IPR indicator (i.e., expected absolute prediction error) is, by design, "on the same scale" as the ground truth (i.e., actual absolute prediction error) against which the reliability is judged. This allows for an even more precise performance measurement (going well beyond correlation coefficient), e.g., using canonical numeric accuracy measures such as *root mean squared error* (RMSE).

3.2 Estimating and Evaluating the Proposed Reliability Indicator: ML-Based Framework

In this subsection, we more formally describe the details of machine-learning-based framework, which can be used for estimating and evaluating the proposed absolute-prediction-error-based IPR indicator. We also use this framework for the computational experiments in our study.

As a quick summary, the proposed framework follows a two-stage process. Because IPR estimation is done for some given outcome prediction model, the overarching goal of Stage 1 is to use the outcome prediction model (build it first, if necessary) and produce the data about its errors, which is typically achieved by deploying the outcome prediction model on a representative, hold-out data sample (i.e., test data). This data then serves as the ground truth for Stage 2, where the actual IPR estimation is done – i.e., the actual absolute prediction errors from Stage 1 are used as outcome variables to build an error prediction model using best machine learning practices. The overview of the entire framework is depicted in Fig. 4.

In terms of data, as depicted in the first row of Fig. 4 and as is typical in predictive modeling, we assume the existence of two datasets (often based on a random split of an underlying database with known outcome values) $\{(x_o, y_o)\}$ and $\{(x_{test}, y_{test})\}$, where the former is used for outcome prediction model learning and the latter for outcome model evaluation. In both datasets, for each data point, x represents the input feature vector, and y represents the corresponding outcome variable. In Stage 1, given $\{(x_o, y_o)\}$, outcome prediction model f is built using any desired numeric prediction modeling technique, e.g., Neural Network, Regression Tree, Random Forest, etc., as is shown in the second row of Fig. 4.[6] This is a standard model learning process where the best machine-learning practices and procedures, such as cross-validation method (Kohavi 1995, Picard and Cook 1984), can be used to properly build and fine-tune outcome prediction models. In our experiments (discussed later in the paper), we use multiple different machine learning techniques to explore the effectiveness of the proposed prediction reliability approach in conjunction with various outcome prediction models.

After model f is trained, naturally it can be deployed for outcome prediction purposes, i.e., to make outcome predictions for any input x as $f(x)$. Stage 1 concludes by deploying f on outcome evaluation data $\{(x_{test}, y_{test})\}$, i.e., prediction for each observation (x_{test}, y_{test}) is constructed as $\hat{y}_{test} = f(x_{test})$, and corresponding (absolute) prediction error is derived as $e = |\hat{y}_{test} - y_{test}|$. This newly generated data $\{e\}$ – the set of actual prediction errors of f on the outcome evaluation dataset – has traditionally been used for the final, authoritative evaluation of model f performance on hold-out data. However, it also carries specific information about the performance on individual predictions (i.e., actual errors) by model f and, thus, we use this information in the form of labeled error learning dataset $\{(x_{test}, e)\}$ in Stage 2 for building models for IPR estimation.

Stage 2 represents our proposed machine-learning approach to IPR estimation. As discussed earlier, we propose to use machine-learning techniques to estimate absolute prediction errors

[6] In our study, during Stage 1, we use outcome prediction models built using *machine learning* techniques, as is done in many advanced real-world applications. However, *any* outcome-predicting model can be used in this framework, e.g., an already existing rule-based expert system or some black-box approach which may not require separate outcome learning data at this point.

(representing IPR) of any given outcome prediction model directly as a function of input features x. The ground truth labels for this machine learning task are obtained from Stage 1, which results in the labeled error learning dataset $\{(x_{test}, e)\}$, as discussed above and shown in Fig. 4. Following standard machine-learning practices, dataset $\{(x_{test}, e)\}$ is randomly split into training dataset $\{(x_t, e_t)\}$ and validation dataset $\{(x_v, e_v)\}$. Based on training dataset $\{(x_t, e_t)\}$, to encapsulate the underlying relationships between input feature vector x_t and the absolute error e_t, error prediction model f_e is constructed as $\hat{e}_t = f_e(x_t)$, where $\hat{e}_t$ denotes the model prediction. Model f_e could be produced by any available machine learning technique using best model-building and fine-tuning practices (e.g., cross-validation), and the best choice of the technique ultimately will depend on the context of each specific prediction problem, e.g., complexity of underlying relationships in data, availability of the data, etc. In our experiments, we use numerous machine learning techniques to explore their performance under different contexts.

Once error prediction model f_e is trained, it can be deployed for reliability estimation purposes, and Stage 2 concludes by deploying f_e to error validation data $\{(x_v, e_v)\}$, as shown in Fig. 4. In particular, the actual absolute prediction error e_v of outcome prediction model f for any data point (x_v, y_v), i.e., $e_v = |\hat{y}_v - y_v|$, would be estimated by f_e as $\hat{e}_v = f_e(x_v)$. In other words, $f_e(x_v)$ represents the proposed IPR indicator for the corresponding individual outcome prediction $f(x_v)$, for any input x_v. In summary, based on the proposed approach and framework, *for any given input x, the two models (f and f_e) would be able to provide the essential prediction-related information: the outcome prediction as $f(x)$ and the estimated reliability of this prediction as $f_e(x)$.*

Finally, error validation dataset $\{(x_v, e_v)\}$ can also be used to properly evaluate the performance of error prediction model f_e, as this data has been used to build neither outcome prediction nor IPR estimation models. Thus, as shown in Fig. 4, the final evaluation of reliability estimation performance is done by comparing the obtained IPR estimates $\{\hat{e}_v\}$ with actual prediction errors $\{e_v\}$. As discussed earlier, for an IPR indicator to be meaningful, the IPR estimates ideally should be "aligned" with actual errors of the outcome prediction model; thus, one relevant and widely used IPR estimation performance metric is correlation coefficient, i.e., $corr(\{e_v\}, \{\hat{e}_v\})$; a higher correlation value indicates better performance. Importantly, as the proposed IPR indicator has been designed to be "on the same scale" as the ground truth, it allows to use more precise numeric prediction accuracy measures as well, such as root mean squared error, i.e., $RMSE(\{e_v\}, \{\hat{e}_v\}) = \sqrt{\sum_v (e_v - \hat{e}_v)^2 / |\{e_v\}|}$), where a lower RMSE value indicates better performance.

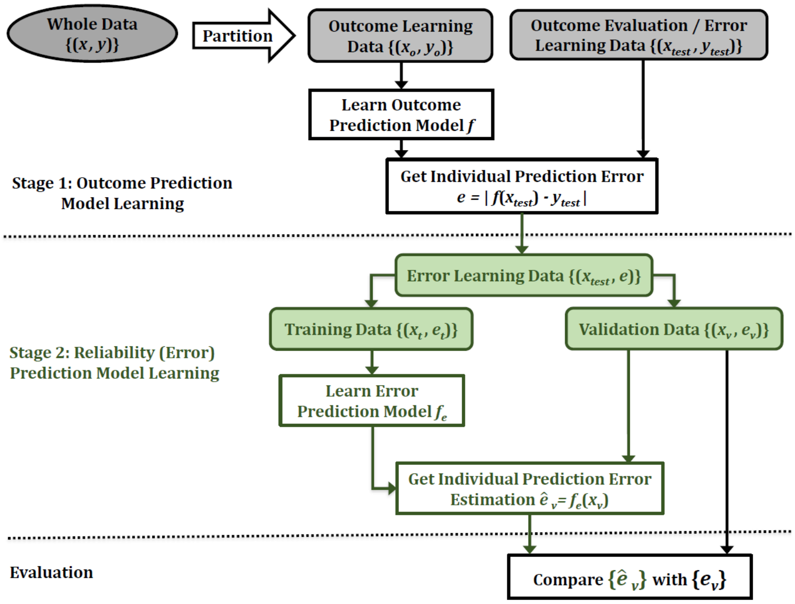

Whole Data $\{(x, y)\}$
Partition
Outcome Learning Data $\{(x_o, y_o)\}$
Outcome Evaluation / Error Learning Data $\{(x_{test}, y_{test})\}$
Learn Outcome Prediction Model f
Stage 1: Outcome Prediction Model Learning
Get Individual Prediction Error $e = |f(x_{test}) - y_{test}|$
Error Learning Data $\{(x_{test}, e)\}$
Training Data $\{(x_t, e_t)\}$
Validation Data $\{(x_v, e_v)\}$
Stage 2: Reliability (Error) Prediction Model Learning
Learn Error Prediction Model f_e
Get Individual Prediction Error Estimation $\hat{e}_v = f_e(x_v)$
Evaluation
Compare $\{\hat{e}_v\}$ with $\{e_v\}$

Figure 4. Two-Stage Machine-Learning-Based Framework for Prediction Reliability Estimation

4. Experiments

4.1 Experimental Setup

We demonstrate the effectiveness of our approach through comprehensive computational experiments following the general two-stage framework described in Section 3.2 and Fig. 4.

Seven public data sets from UCI Machine Learning Repository[7] (as summarized in Table 3) are used to test the performance of the proposed approach. Selected data sets vary by application domain, size (number of records), and complexity (number of input attributes). For Stage 1, each data set is randomly split into two parts, i.e., outcome learning data and outcome evaluation data, with percentages of 40% and 60%, respectively. For Stage 2, the latter part is used as error learning data and is further randomly split into equal-sized (i.e., 50%-50%) error training and error validation datasets. Note that we do the performance evaluation 30 times for each dataset (by generating a different random split into outcome learning, error learning, and error validation

[7] https://archive.ics.uci.edu/ml/datasets.html

datasets). All results are based on the average performance of the 30 runs, and all techniques (ML-based and baselines) were evaluated on the same evaluation data within each run.

For Stage 1, i.e., to build outcome prediction models, we chose seven machine learning techniques widely used for predicting numeric outcomes, i.e., KNN (k nearest neighbors), NN (neural network), LR (linear regression), RT (regression tree), RF (random forest), SVR (support vector regression), and XGB (extreme gradient boosting). The use of different predictive modeling techniques highlights the general-purpose applicability of the proposed reliability estimation approach for use in conjunction with a wide variety of outcome prediction models. For Stage 2, i.e., to build absolute error prediction models, we used the same set of machine learning techniques to explore whether some of them might be more advantageous for the reliability estimation task.

Table 3. Overview of Data Sets Used in Computational Experiments

Dataset	#Obs	#Attributes	Output description	Output range
Power Plant	9568	4	Hourly electrical energy output.	[420, 495]
ISE	536	7	Istanbul Stock Exchange 100 index.	[-8.5, 10]
Housing	506	13	Value of houses in $1000s.	[6, 50]
Bike Rental	17389	16	Daily count of rental bikes.	[1, 977]
Parkinsons	5875	26	Parkinson's disease symptom score.	[7, 29]
Posts Comments	40949	54	Log of number of Facebook posts comments.	[0, 8]
News Popularity	39797	58	Log of number of total shares of news.	[3, 13]

To benchmark the proposed approach, we use nine baseline algorithms for comparison (summarized in Table 2): one bootstrapping-based and eight heuristic-based reliability estimators. For parameter setting, specifically, in our experiments, we set m = 20 (number of random samples generated from original data in bootstrapping) and n = 20 (number of nearest neighbors chosen to calculate those heuristic-based baselines). The reliability estimation performance of different approaches is evaluated using both correlation-based and predictive-accuracy-based metrics, i.e., correlation coefficient and RMSE; as mentioned earlier, higher correlation and lower discrepancy between actual and estimated prediction errors indicate better reliability estimation.

For expositional completeness, we first present the predictive accuracy measured by RMSE of different *outcome* prediction models (i.e., models built in Stage 1), as shown in Table 4, where each row and column represents each outcome prediction model and dataset, respectively. Note that, for each technique, the results represent the best performance of each model achieved after optimizing model parameters, e.g., the number of nearest neighbors in KNN, depth of the tree in RT and RF, number of neurons and hidden layers in NN, number of estimators and size of subsample in XGB, length scale and gamma parameters of kernel functions in SVR, and many other parameters. Best performance, i.e., lowest RMSE, on each dataset is highlighted in bold, which shows that XGB generally tends to perform well on different data sets, followed by RF and NN. The one exception is the simple ISE dataset, where arguably the simplest model – linear regression – is sufficient to capture predictive relationships in the data.

Table 4. Predictive Accuracy (RMSE) of Different Outcome Prediction Models.
(Average performance based on 30 runs; best performance on each data set is shown in bold.)

Model	Power Plant	ISE	Housing	Bike Rental	Parkinsons	Posts Comments	News Pop
KNN	3.985	1.548	6.040	0.896	0.329	0.701	0.882
LR	4.578	**1.404**	5.600	1.072	0.378	0.814	0.874
NN	4.285	1.431	4.914	0.431	0.364	0.647	0.869
RF	3.775	1.524	4.643	0.369	0.277	0.509	0.864
RT	4.415	1.688	5.696	0.501	0.382	0.555	0.888
SVR	4.535	1.526	7.095	0.906	0.400	0.634	0.889
XGB	**3.574**	1.514	**4.601**	**0.341**	**0.226**	**0.493**	**0.852**

The next two subsections discuss the results of IPR estimation (i.e., Stage 2) experiments.

4.2 Experimental Results: Performance Comparison Based on Correlation

We first focus on the performance comparisons in terms of correlation coefficient – the widely used metric for evaluating IPR indicators, as discussed earlier. We first show the detailed performance of each machine-learning-based method for our proposed absolute-error-based IPR indicator as well as each baseline method, and then compare the effectiveness of these two classes of methods using summarized results.

In particular, Table 5 compares the reliability estimation performance among seven machine learning techniques. The bold numbers represent best performance for a given outcome prediction model in terms correlation coefficient. A closer look at the results shows that the XGB approach produced the best (or near best) reliability performance among the ML-based approaches. Specifically, in the majority (33 out of 49) of settings that were explored XGB outperforms other techniques, followed closely by RF which performs the best in the rest (15 out of 49) of the settings. An interesting pattern observed from the results is that RF is better than or competitive with XGB only on data with simpler structure (having fewer input features), e.g., Housing, ISE, and Power

Plant; that is, XGB consistently has the edge over all approaches on more complex datasets.

Table 5. Reliability Estimation Performance of Machine-Learning-Based Methods (Correlation Coefficient)
(Average performance based on 30 runs; best result for each outcome prediction model on each data shown in bold.)

	Outcome Prediction / **Reliability Estimation**	**KNN**	**LR**	**NN**	**RF**	**RT**	**SVR**	**XGB**
Power Plant	KNN	0.253	0.336	0.304	0.215	0.240	0.337	0.202
	LR	0.071	0.089	0.104	0.074	0.077	0.121	0.062
	NN	0.094	0.194	0.156	0.099	0.105	0.201	0.094
	RF	**0.291**	**0.413**	**0.374**	0.219	**0.311**	**0.416**	0.191
	RT	0.059	0.182	0.137	0.068	0.097	0.175	0.046
	SVR	0.160	0.286	0.271	0.157	0.185	0.265	0.155
	XGB	0.278	0.407	0.373	**0.229**	0.301	0.415	**0.210**
ISE	KNN	0.122	0.099	0.089	0.135	0.182	0.150	0.123
	LR	0.015	0.026	0.020	0.040	0.000	0.000	0.015
	NN	0.043	0.052	0.047	0.107	0.115	0.080	0.046
	RF	**0.123**	**0.163**	**0.129**	**0.173**	**0.231**	**0.183**	0.160
	RT	0.045	0.075	0.112	0.062	0.157	0.097	**0.199**
	SVR	0.058	0.035	0.100	0.028	0.103	0.083	0.023
	XGB	0.110	0.100	0.066	0.133	0.199	0.141	0.111
Housing	KNN	0.419	0.384	0.333	0.327	0.339	0.454	0.306
	LR	0.381	0.326	0.340	0.308	0.293	0.442	0.326
	NN	0.490	0.484	0.357	0.310	0.321	0.635	0.341
	RF	**0.538**	0.500	0.457	**0.433**	**0.424**	**0.638**	0.440
	RT	0.398	0.409	0.360	0.320	0.298	0.518	0.350
	SVR	0.301	0.318	0.279	0.253	0.264	0.398	0.227
	XGB	0.522	**0.505**	**0.459**	0.431	0.417	0.636	**0.453**
Bike Rental	KNN	0.523	0.490	0.392	0.460	0.408	0.563	0.447
	LR	0.460	0.382	0.342	0.398	0.349	0.527	0.396
	NN	0.557	0.567	0.345	0.405	0.357	0.617	0.400
	RF	0.742	0.858	0.482	0.499	0.462	0.840	0.488
	RT	0.684	0.798	0.365	0.410	0.361	0.801	0.419
	SVR	0.479	0.441	0.370	0.443	0.383	0.541	0.434
	XGB	**0.748**	**0.865**	**0.494**	**0.505**	**0.480**	**0.844**	**0.498**
Parkinsons	KNN	0.508	0.478	0.476	0.484	0.538	0.560	0.419
	LR	0.267	0.205	0.235	0.257	0.279	0.285	0.237
	NN	0.285	0.232	0.255	0.277	0.307	0.345	0.226
	RF	0.496	0.377	0.430	0.498	0.456	0.404	0.401
	RT	0.240	0.209	0.239	0.283	0.254	0.248	0.242
	SVR	0.424	0.223	0.255	0.246	0.196	0.428	0.283
	XGB	**0.637**	**0.653**	**0.664**	**0.625**	**0.714**	**0.697**	**0.575**
Comments	KNN	0.528	0.519	0.478	0.428	0.438	0.495	0.420
	LR	0.489	0.556	0.449	0.374	0.395	0.456	0.368
	NN	0.539	0.599	0.451	0.460	0.459	0.506	0.458
	RF	0.634	0.672	0.622	0.532	0.541	0.611	0.525
	RT	0.574	0.589	0.566	0.496	0.512	0.551	0.491
	SVR	0.549	0.526	0.492	0.435	0.448	0.491	0.427
	XGB	**0.640**	**0.676**	**0.629**	**0.540**	**0.549**	**0.618**	**0.527**
News Pop	KNN	0.185	0.197	0.201	0.204	0.182	0.183	0.209
	LR	0.203	0.229	0.227	0.226	0.200	0.204	0.233
	NN	0.205	0.226	0.232	0.232	0.207	0.206	0.237
	RF	0.213	0.227	0.236	0.243	0.221	0.215	0.246
	RT	0.175	0.181	0.194	0.199	0.179	0.171	0.202
	SVR	0.052	0.175	0.198	0.179	0.179	0.181	0.181
	XGB	**0.229**	**0.242**	**0.251**	**0.258**	**0.236**	**0.233**	**0.262**

Similarly, in Table 6 we show the comparison among nine baseline techniques. Although no one baseline predominantly outperforms others, MSE (heuristic-based) and VarBag (bootstrapping-based) tend to have higher correlation coefficient with actual errors in 20 and 12 (out of 49) settings, respectively.

Table 6. Reliability Estimation Performance of Heuristic-Based Methods (Correlation Coefficient)
(Average performance based on 30 runs; best result for each outcome prediction model on each data shown in bold.)

	Outcome Prediction Reliability Estimation	KNN	LR	NN	RF	RT	SVR	XGB
Power Plant	VarBag	**0.232**	0.030	0.018	**0.187**	0.102	0.123	**0.184**
	VarA	0.173	0.104	0.139	0.133	0.136	0.150	0.117
	VarP	0.058	-0.001	0.002	0.096	0.070	0.025	0.091
	AvgDiff	0.025	0.023	0.015	0.014	0.044	0.009	0.018
	MSE	0.187	**0.231**	0.173	0.124	**0.160**	**0.225**	0.098
	AvgDist	0.028	0.002	-0.007	-0.010	0.018	0.045	-0.010
	LCV	-0.026	-0.045	0.141	-0.015	-0.027	-0.016	-0.009
	SAV	-0.003	0.012	0.132	-0.006	0.024	0.103	-0.078
	SAB	0.017	0.002	**0.180**	-0.001	0.012	0.001	0.111
ISE	VarBag	0.166	**0.342**	**0.321**	0.253	0.250	0.227	**0.296**
	VarA	0.107	0.093	0.084	0.112	0.152	0.100	0.122
	VarP	0.084	0.153	0.132	0.126	0.174	0.167	0.137
	AvgDiff	-0.024	-0.023	-0.001	-0.010	0.024	0.008	0.005
	MSE	0.100	0.075	-0.087	0.066	0.078	0.073	0.071
	AvgDist	**0.273**	0.252	0.261	**0.284**	**0.318**	**0.315**	0.287
	LCV	0.040	0.024	0.094	-0.016	-0.005	0.013	-0.009
	SAV	0.007	0.259	0.007	0.082	0.181	0.311	0.078
	SAB	0.020	-0.064	0.020	0.023	-0.003	0.004	-0.046
Housing	VarBag	**0.396**	0.342	0.404	**0.418**	**0.385**	0.266	**0.437**
	VarA	0.373	0.326	0.349	0.306	0.280	0.397	0.306
	VarP	0.272	0.130	0.329	0.294	0.256	0.108	0.305
	AvgDiff	-0.316	-0.323	**-0.417**	-0.360	-0.268	-0.318	-0.373
	MSE	0.368	**0.353**	0.245	0.303	0.232	**0.452**	0.271
	AvgDist	0.178	0.204	0.130	0.071	0.027	0.174	0.082
	LCV	-0.148	0.006	0.240	-0.061	-0.074	-0.358	-0.033
	SAV	0.007	0.143	0.106	0.012	0.162	0.255	0.122
	SAB	-0.123	-0.037	-0.002	-0.002	0.132	0.053	0.217
Bike Rental	VarBag	0.445	0.052	0.295	0.376	**0.323**	0.299	**0.376**
	VarA	**0.523**	0.461	0.242	0.241	0.180	0.505	0.261
	VarP	0.395	0.038	0.234	0.221	0.177	0.390	0.253
	AvgDiff	0.149	0.049	0.172	0.249	0.161	0.162	0.275
	MSE	0.479	**0.463**	**0.305**	**0.384**	0.265	**0.522**	0.272
	AvgDist	0.052	0.025	0.172	0.233	0.238	0.004	0.202
	LCV	-0.014	0.014	0.035	0.072	0.073	0.078	0.011
	SAV	-0.014	0.022	0.188	-0.002	0.233	0.169	0.225
	SAB	0.052	-0.012	-0.131	-0.021	-0.003	0.046	0.212
Parkinsons	VarBag	0.470	0.053	0.055	0.293	0.022	0.102	0.399
	VarA	0.531	0.413	0.446	0.315	0.423	0.434	0.361
	VarP	0.276	-0.012	0.134	0.149	0.116	0.124	0.263
	AvgDiff	0.120	0.013	-0.026	0.184	0.162	0.013	0.192
	MSE	**0.557**	**0.552**	**0.452**	**0.532**	**0.615**	**0.570**	**0.445**
	AvgDist	-0.085	-0.045	-0.080	-0.048	-0.010	-0.100	-0.069
	LCV	0.112	-0.007	-0.107	0.059	0.164	0.188	0.054
	SAV	-0.002	-0.053	-0.047	-0.157	-0.163	0.132	-0.026
	SAB	-0.082	0.001	-0.043	-0.048	0.014	0.064	0.034
Comments	VarBag	0.384	0.321	0.263	0.310	0.283	0.238	0.309
	VarA	0.470	0.379	**0.404**	0.330	0.353	**0.410**	**0.318**
	VarP	0.346	0.276	0.346	0.292	0.310	0.297	0.279
	AvgDiff	-0.213	-0.130	-0.137	-0.188	-0.201	-0.178	-0.178
	MSE	**0.473**	**0.493**	0.393	0.326	**0.383**	0.379	0.317
	AvgDist	0.230	0.361	0.214	0.159	0.171	0.257	0.149
	LCV	-0.083	0.006	0.043	-0.065	-0.051	-0.02	-0.075
	SAV	-0.045	0.365	0.320	**0.429**	0.251	0.312	0.279
	SAB	-0.095	-0.047	0.227	0.067	0.100	-0.062	0.237
News Pop	VarBag	0.135	0.063	0.050	0.107	0.047	0.077	0.130
	VarA	**0.137**	0.139	0.155	**0.146**	**0.130**	**0.134**	0.150
	VarP	0.125	0.132	0.129	0.111	0.068	0.119	0.125
	AvgDiff	-0.085	-0.116	-0.129	-0.125	-0.106	-0.087	-0.128
	MSE	0.135	**0.142**	0.142	0.128	0.130	0.115	0.103
	AvgDist	0.106	0.125	0.119	0.120	0.106	0.108	0.116
	LCV	-0.016	0.001	**0.172**	-0.007	-0.035	-0.042	0.148
	SAV	-0.005	0.083	0.080	0.143	0.026	0.05	**0.182**
	SAB	0.068	0.008	0.007	0.050	-0.011	0.046	0.095

Another observation is that VarBag is more competitive on data with less complex structure (having fewer input features), e.g., Power Plant, ISE, Housing, and Bike Rental, while on data sets like Parkinsons, Posts Comments, and News Popularity, heuristic-based estimators like MSE and VarA generally perform better.

We summarize the comparison of correlation coefficient results between ML-based reliability estimators and baselines in Table 7. Specifically, we compare the *best* baseline (BL) technique (chosen among the nine baseline techniques discussed earlier) and the *best* machine learning (ML) model (chosen from the seven ML techniques used earlier) in terms of correlation. The bold and red numbers represent significantly higher correlation coefficients from ML-based methods, and the bold and blue numbers represent better results from baselines. The results show that, in 42 out of 49 (85.7%) predictive task configurations in our experiments, the best ML-based estimator is a better IPR indicator (exhibiting higher correlation with actual prediction errors), and in 39 out of these 42 cases the advantage is statistically significant, emphasizing the advantages of using the proposed approach over baselines for IPR estimation.

Table 7. Comparison of Reliability Estimation Performance (Correlation Coefficient)
(Average performance based on 30 runs; better result on each data set is shown in bold; red bold: machine learning technique is significantly better; blue bold: baseline is significantly better)

Outcome prediction model	**KNN**		**LR**		**NN**		**RF**		**RT**		**SVR**		**XGB**	
Prediction reliability estimator	**Best BL**	**Best ML**	**Best BL**	**Best ML**	**Best BL**	**Best ML**	**Best BL**	**Best ML**	**Best BL**	**Best ML**	**Best BL**	**Best ML**	**Best BL**	**Best ML**
Power Plant	0.23	**0.29*****	0.23	**0.41*****	0.18	**0.37*****	0.19	**0.23*****	0.16	**0.31*****	0.23	**0.42*****	0.18	**0.21***
ISE	**0.27*****	0.12	**0.34*****	0.16	**0.32*****	0.13	**0.28*****	0.17	**0.32*****	0.23	**0.32*****	0.18	**0.30*****	0.16
Housing	0.40	**0.54*****	0.36	**0.51*****	0.42	**0.46*****	0.42	**0.43**	0.39	**0.41**	0.45	**0.64*****	0.44	**0.45**
Bike Rental	0.52	**0.75*****	0.46	**0.87*****	0.31	**0.50*****	0.38	**0.51*****	0.32	**0.48*****	0.52	**0.84*****	0.38	**0.50*****
Parkinsons	0.56	**0.64*****	0.55	**0.65*****	0.45	**0.66*****	0.53	**0.63***	0.62	**0.71*****	0.57	**0.70*****	0.45	**0.58*****
Comments	0.47	**0.64*****	0.50	**0.68*****	0.40	**0.63*****	0.43	**0.54*****	0.38	**0.55*****	0.41	**0.62*****	0.32	**0.53*****
News Pop	0.14	**0.23*****	0.14	**0.24*****	0.17	**0.25*****	0.15	**0.26*****	0.13	**0.24*****	0.13	**0.23*****	0.18	**0.26*****

*** $p<0.001$, ** $p<0.01$, * $p<0.05$

Also note that ML-based IPR estimators were outperformed by baselines only on the ISE dataset, i.e., the simple dataset (500+ observations and 7 input features) with significantly less complex predictive relationships, where a simpler outcome prediction model like linear regression was sufficient to guarantee high prediction accuracy, as mentioned earlier. In other words, the heuristic-based IPR estimators may be sufficient for simpler datasets; however, more sophisticated approaches are advantageous when more complex predictive settings must be considered.

4.3 Experimental Results: Performance Comparison Based on RMSE

As discussed in Section 3.1, while correlation coefficient is able to capture general variability patterns, it is not designed to reflect the situations where the magnitude of IPR estimates might be significantly different than that of the actual errors, reducing our ability to make more precise judgements about IPR estimation performance. The proposed absolute-prediction-error-based IPR indicator provides for more precise and informative performance evaluation, due to being "on the same scale" as the ground truth, which allows us to bring in standard numeric prediction accuracy measures – specifically, *root mean squared error* (RMSE) – and provide a much clearer picture of true IPR estimation performance, as discussed below. Here we follow the same structure as in the previous subsection, where we first show the detailed performance of each machine-learning-based method for our proposed absolute-error-based reliability indicator, followed by detailed performance of each baseline method, and then compare the effectiveness of these two classes of methods using summarized results.

In Table 8, we compare the IPR estimation performance in terms of RMSE among machine learning methods, bold numbers representing best performance for each outcome prediction model. The results show similar patterns as in the correlation coefficient comparisons. In particular, XGB still performs the best, i.e., exhibits lower discrepancy with actual prediction errors, among all machine learning techniques in most cases. Specifically, in 29 out of 49 cases, XGB produces most accurate reliability estimation, followed by RF and KNN which perform better in the rest 14 and 6 cases, respectively. Also, detailed results show that RF and KNN tend to outperform XGB on datasets with fewer input features, i.e., Power Plant, ISE, and Housing, while XGB is more advantageous on more complex datasets.

As mentioned earlier, RMSE should be calculated when IPR indicator and actual prediction error are on the same scale. However, this is not the case for heuristic-based IPR indicators, and computing RMSE based on raw values of heuristic-based indicators would put them at a significant disadvantage in terms of their performance comparison with the proposed approach.

Table 8. Reliability Estimation Performance of Machine-Learning-Based Methods (RMSE)
(Average performance based on 30 runs; best result for each outcome prediction model on each data shown in bold.)

	Outcome Prediction Reliability Estimation	KNN	LR	NN	RF	RT	SVR	XGB
Power Plant	KNN	2.606	2.613	2.552	**2.520**	2.806	2.680	**2.407**
	LR	2.680	2.759	2.658	2.566	2.875	2.902	2.446
	NN	2.683	2.728	2.642	2.565	2.862	2.790	2.442
	RF	**2.590**	**2.535**	**2.498**	2.530	**2.756**	2.594	2.417
	RT	2.693	2.739	2.654	2.576	2.890	2.824	2.456
	SVR	2.684	2.672	2.599	2.566	2.858	2.839	2.450
	XGB	2.598	2.550	2.501	2.524	2.768	**2.595**	2.415
ISE	KNN	**1.014**	**0.909**	**0.954**	1.018	1.133	1.043	**1.019**
	LR	1.049	0.931	0.973	1.043	1.179	1.075	1.051
	NN	1.037	0.925	0.962	1.024	1.146	1.053	1.034
	RF	1.033	0.912	0.958	**1.014**	**1.115**	**1.031**	1.020
	RT	1.073	0.942	0.981	1.054	1.161	1.081	1.055
	SVR	1.024	0.918	0.961	1.026	1.146	1.039	1.025
	XGB	1.059	0.940	0.995	1.047	1.150	1.061	1.054
Housing	KNN	4.292	3.750	3.429	3.346	4.023	5.025	3.372
	LR	4.377	3.869	3.433	3.397	4.125	5.011	3.395
	NN	4.076	3.536	3.342	3.367	4.033	4.308	3.324
	RF	**3.924**	**3.477**	**3.206**	**3.200**	**3.906**	**4.179**	**3.164**
	RT	4.408	3.700	3.451	3.395	4.086	4.868	3.276
	SVR	4.503	3.919	3.522	3.425	4.171	5.169	3.527
	XGB	4.046	3.526	3.295	3.275	4.041	4.315	3.259
Bike Rental	KNN	0.532	0.580	0.269	0.247	0.327	0.527	0.228
	LR	0.553	0.615	0.275	0.255	0.335	0.541	0.234
	NN	0.518	0.550	0.275	0.255	0.334	0.501	0.234
	RF	0.417	0.342	0.256	0.240	0.317	0.339	0.222
	RT	0.454	0.402	0.272	0.254	0.334	0.376	0.231
	SVR	0.558	0.604	0.276	0.254	0.338	0.546	0.233
	XGB	**0.414**	**0.335**	**0.254**	**0.240**	**0.314**	**0.337**	**0.221**
Parkinsons	KNN	0.188	0.210	0.205	0.144	0.198	0.217	0.131
	LR	0.207	0.229	0.223	0.159	0.219	0.244	0.140
	NN	0.206	0.228	0.223	0.158	0.217	0.239	0.141
	RF	0.188	0.216	0.207	0.139	0.206	0.234	0.130
	RT	0.207	0.228	0.222	0.157	0.220	0.246	0.140
	SVR	0.201	0.233	0.225	0.160	0.227	0.238	0.139
	XGB	**0.169**	**0.180**	**0.176**	**0.124**	**0.173**	**0.188**	**0.115**
Comments	KNN	0.442	0.526	0.407	0.333	0.362	0.400	0.324
	LR	0.455	0.471	0.415	0.343	0.372	0.408	0.333
	NN	0.443	0.517	0.462	0.329	0.360	0.427	0.318
	RF	0.401	0.459	0.362	0.311	0.338	0.360	0.303
	RT	0.425	0.496	0.382	0.320	0.346	0.382	0.311
	SVR	0.440	0.495	0.412	0.337	0.364	0.393	0.327
	XGB	**0.399**	**0.462**	**0.360**	**0.310**	**0.336**	**0.358**	**0.301**
News Pop	KNN	0.593	0.577	0.574	0.569	0.581	0.614	0.566
	LR	0.591	0.573	0.570	0.566	0.579	0.611	0.563
	NN	0.591	0.574	0.570	0.565	0.578	0.613	0.562
	RF	0.589	0.573	0.569	0.563	0.576	0.609	0.561
	RT	0.594	0.579	0.574	0.569	0.581	0.615	0.566
	SVR	0.604	0.586	0.584	0.575	0.591	0.621	0.575
	XGB	**0.587**	**0.571**	**0.567**	**0.561**	**0.574**	**0.607**	**0.558**

Therefore, we take a broader view of the heuristic-based indicators by observing that some of them are calculated by aggregating (e.g., as variance) a certain set of discrepancies (errors), and we aggregated these discrepancies by averaging their absolute values to provide the best-effort estimation of an absolute prediction error. In particular, only five (i.e., VarBag, VarA, VarP, AvgDiff, MSE) out of nine baseline IPR indicators could be converted to estimates of absolute prediction errors[8] (i.e., VarBag.AE, VarA.AE, VarP.AE, AvgDiff.AE, MSE.AE) and, thus, could be used for RMSE comparisons. The other baselines are heuristics that provide a numeric index indicating the degree of prediction reliability but have no direct connection to prediction errors.

As a result, in Table 9 we provide the comparison of reliability estimation performance in terms of RMSE among the four aforementioned baselines. As with performance comparisons based on correlation coefficient, no single heuristic-based indicator dominates all others, but MSE.AE and VarA.AE provide best performance in 24 and 21 (out of 49) settings, respectively.

Table 9. Reliability Estimation Performance of Heuristic-Based Methods (RMSE)
(Average performance based on 30 runs; best result for each outcome prediction model on each data shown in bold.)

	Outcome Prediction / Reliability Estimation	KNN	LR	NN	RF	RT	SVR	XGB
Power Plant	VarBag.AE	3.977	4.569	4.366	3.741	4.078	4.535	3.609
	VarA.AE	2.822	2.877	2.815	2.964	2.967	2.894	2.941
	VarP.AE	2.865	3.149	3.151	**2.798**	3.241	3.310	**2.783**
	AvgDiff.AE	3.301	3.345	3.446	3.072	3.358	3.391	3.260
	MSE.AE	**2.641**	**2.646**	**2.791**	3.038	**2.858**	**2.778**	2.993
ISE	VarBag.AE	2.090	2.087	2.087	2.088	2.089	1.534	2.088
	VarA.AE	**1.021**	0.936	0.961	**1.020**	**1.091**	**1.022**	**1.019**
	VarP.AE	1.185	1.003	1.033	1.099	1.297	1.191	1.081
	AvgDiff.AE	1.031	0.915	1.154	1.164	1.176	1.166	1.170
	MSE.AE	1.031	**0.915**	**0.961**	1.164	1.176	1.066	1.215
Housing	VarBag.AE	6.011	5.693	5.356	4.699	4.981	7.085	4.647
	VarA.AE	**4.436**	**3.804**	4.261	4.557	5.111	**5.036**	4.657
	VarP.AE	4.439	3.859	4.124	4.377	5.184	6.033	4.546
	AvgDiff.AE	4.858	4.026	4.095	4.408	5.263	5.773	4.274
	MSE.AE	4.536	3.833	**3.690**	**3.953**	**4.564**	5.239	**4.238**
Bike Rental	VarBag.AE	0.902	1.072	**0.235**	0.367	0.414	0.902	0.343
	VarA.AE	0.552	**0.578**	0.862	0.915	0.882	0.587	0.921
	VarP.AE	0.605	0.823	0.826	0.895	0.849	0.626	0.895
	AvgDiff.AE	0.758	0.809	0.695	0.739	0.738	0.716	0.734
	MSE.AE	**0.542**	0.584	0.282	**0.302**	**0.361**	**0.538**	**0.278**
Parkinsons	VarBag.AE	0.330	0.378	0.272	0.275	0.381	0.399	0.229
	VarA.AE	**0.170**	**0.182**	**0.181**	0.184	0.175	**0.305**	0.181
	VarP.AE	0.281	0.326	0.251	0.260	0.356	0.390	0.197
	AvgDiff.AE	0.718	0.702	0.702	0.913	0.718	0.692	0.184
	MSE.AE	0.173	0.187	0.210	**0.136**	**0.173**	0.398	**0.131**
Comments	VarBag.AE	0.701	0.863	0.536	0.504	0.525	0.608	0.493
	VarA.AE	**0.446**	**0.452**	0.451	0.491	0.496	0.631	0.495
	VarP.AE	0.521	0.598	0.449	0.459	0.473	**0.515**	0.462
	AvgDiff.AE	0.498	0.523	0.452	0.500	0.482	0.575	0.462
	MSE.AE	0.447	0.529	**0.435**	**0.349**	**0.365**	0.629	**0.339**
News Pop	VarBag.AE	0.883	0.872	0.903	0.859	0.878	0.882	0.849
	VarA.AE	**0.601**	**0.582**	**0.579**	**0.574**	**0.586**	0.841	**0.571**
	VarP.AE	0.808	0.744	0.665	0.687	0.751	0.840	0.658
	AvgDiff.AE	0.605	0.733	0.590	0.690	0.723	0.850	0.714
	MSE.AE	0.602	0.585	0.585	0.590	0.589	**0.839**	0.605

Finally, we summarize the comparison of error estimation accuracy (RMSE) between ML-based estimators and baselines in Table 10. Similar to Table 7, we compare RMSE of the *best*

[8] Formal calculations of these estimators can be found in Appendix B of the Online Supplement.

baseline (BL) technique chosen among the four baselines discussed earlier and the *best* machine learning (ML) model of the seven machine learning techniques used earlier. Significantly lower RMSEs from machine learning based methods are highlighted in bold and red, while significantly lower RMSEs from baselines are highlighted in bold and blue. The results show that, in 40 out of 49 (81.6%) predictive settings in our experiments, ML approaches constitute better IPR indicators, i.e., exhibit lower discrepancy with actual prediction errors as measured by RMSE. Furthermore, in 35 out of 49 cases, best ML-based IPR indicators provide statistically significantly better performance than the heuristic approaches. In contrast, only in 1 out of 49 settings, baselines were statistically significantly better than ML-based approaches. Even on the simpler ISE dataset (where heuristic-based approaches demonstrated better correlation performance), with a more precise performance evaluation using RMSE no statistically significant performance differences are observed between ML-based approaches and baselines. In aggregate, all the experimental results indicate substantial advantages of using machine learning techniques to estimate IPR.

Table 10. Comparison of Reliability Estimation Performance (RMSE)
(Average performance based on 30 runs; better result on each data set is shown in bold; red bold: machine learning technique is significantly better; blue bold: baseline is significantly better)

Outcome prediction model	**KNN**		**LR**		**NN**		**RF**		**RT**		**SVR**		**XGB**	
Prediction reliability estimator	**Best BL**	**Best ML**	**Best BL**	**Best ML**	**Best BL**	**Best ML**	**Best BL**	**Best ML**	**Best BL**	**Best ML**	**Best BL**	**Best ML**	**Best BL**	**Best ML**
Power Plant	2.64	**2.59**	2.65	**2.54*****	2.79	**2.50*****	2.80	**2.52*****	2.86	**2.76****	2.78	**2.60****	2.78	**2.41*****
ISE	1.02	**1.01**	0.92	**0.91**	0.96	**0.95**	1.02	1.01	**1.09**	1.12	**1.02**	1.03	1.02	1.02
Housing	4.44	**3.92***	3.80	**3.48***	3.70	**3.21***	3.95	**3.20*****	4.56	**3.91***	5.04	**4.18*****	4.24	**3.16*****
Bike Rental	0.54	**0.41*****	0.58	**0.34*****	**0.24*****	0.25	0.28	**0.24*****	0.35	**0.31*****	0.54	**0.34*****	0.27	**0.22*****
Parkinsons	0.17	0.17	0.18	0.18	0.18	0.18	0.14	**0.12***	0.17	0.17	0.31	**0.19*****	0.13	**0.12*****
Comments	0.45	**0.40*****	**0.45**	0.46	0.44	**0.36*****	0.35	**0.31*****	0.37	**0.34*****	0.52	**0.36*****	0.34	**0.30*****
News Pop	0.60	**0.59*****	0.58	**0.57*****	0.58	**0.57*****	0.57	**0.56*****	0.59	**0.57*****	0.84	**0.61*****	0.57	**0.56*****

*** $p<0.001$, ** $p<0.01$, * $p<0.05$

5. Conclusions

Estimating individual prediction reliability (IPR) is important for both interpretation and application of predictive models and could be used for several purposes. It provides extra information on the error of individual predictions and, thus, gives practitioners more confidence in making decisions. Going beyond global prediction performance, it also gives a finer-grained evaluation even for presumably well-trained predictive models. For example, even when the outcome prediction model is relatively accurate in general, it may be important to know that, under some circumstances, some predictions objectively can be expected to be much worse than others. More generally, IPR can be used as part of the criteria for identifying most advantageous data points (e.g., points not only with most advantageous predicted outcomes, but also with most reliable predictions) among many candidates for subsequent actions or analysis.

While the awareness of how reliable the specific individual predictions are can be important in many complex real-world numeric predictive modeling applications, this issue has been under-explored in research literature. In this study, we propose to estimate IPR for any given numerical outcome prediction model by using machine learning techniques. Specifically, we reconceptualize the reliability estimation problem to a numeric prediction problem by proposing to use *absolute prediction error* as a simple IPR indicator due to its merits of higher interpretability and easy evaluation. The study also describes a general-purpose framework for implementing the proposed reliability estimation approach, which takes can take advantage of any state-of-the-art machine learning methods to directly learn the relationships between input features of a given data point and absolute prediction errors (i.e., reliability indicators) obtained from the outcome prediction model. In addition to providing an intuitive reliability indicator, the proposed machine-learning-based approach is *general-purpose* (i.e., reliability estimation can be done for *any* outcome prediction model), reduces the need for statistical modeling assumptions that some distributional approaches require, and allows for more precise and informative performance evaluation.

The general-purpose framework was also used in comprehensive computational experiments designed to test the proposed approach. Specifically, we observed that machine learning methods can significantly improve IPR estimation, especially in more complex settings, i.e., on datasets that are larger both in the number of examples and input features. We compared the proposed approach with numerous heuristic approaches used in prior work on seven different public datasets based on two different evaluation metrics. The performance advantages of the proposed machine-learning-based approach (over heuristic-based indicators) can be observed across different outcome prediction models, which further emphasizes the generality of the proposed approach.

In addition to introducing a machine-learning-based approach to estimating IPR and demonstrating its effectiveness, this study provides a number of directions for future research. One such direction would be to understand the impact of *dataset characteristics* on the performance of simpler (heuristic-based) vs. more complex (machine-learning-based) reliability estimators. Another direction would be to explore the impact of different *sources* of prediction uncertainty, e.g., whether low reliability of an individual prediction is due to noisy data, model misfit, etc. Revisiting the possibilities of designing additional, more sophisticated and accurate reliability indicators of different types (indicator-based vs. distribution-based) and levels of applicability (general-purpose vs. building specifically on the strengths of some specific outcome prediction model) also represent important direction for follow-up investigations. Advancing our understanding of these issues should not only make reliability estimation increasingly relevant and valuable in real-world predictive modeling applications, but should also lead to deeper, more significant developments of reliability estimation theory.

References

Bosnić Z, Kononenko I (2008a) Comparison of approaches for estimating reliability of individual regression predictions. *Data Knowl. Eng.* 67(3):504–516.

Bosnić Z, Kononenko I (2008b) Estimation of individual prediction reliability using the local sensitivity analysis. *Appl. Intell.* 29(3):187–203.

Bosnić Z, Kononenko I (2009) An overview of advances in reliability estimation of individual predictions in machine learning. *Intell. Data Anal.* 13(2):385–401.

Breiman L (1996) Bagging predictors. *Mach. Learn.* 24(2):123–140.

Brier GW (1950) Verification of forecasts expersses in terms of probaility. *Mon. Weather Rev.* 78(1):1–3.

Briesemeister S, Rahnenführer J, Kohlbacher O (2012) No Longer Confidential: Estimating the Confidence of Individual Regression Predictions. *PLoS One* 7(11).

Carney JG, Cunningham P, Bhagwan U (1999) Confidence and prediction intervals for neural network ensembles. *IJCNN'99. Int. Jt. Conf. Neural Networks*. 1215–1218.

Chen T, Guestrin C (2016) XGBoost. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD '16*:785–794.

Choi, E., Schuetz, A., Stewart, W. F. and Sun, J. (2016). Using recurrent neural network models for early detection of heart failure onset. *J. of the American Medical Informatics* Association, 24(2), 361-370.

Clark RD (2009) DPRESS: Localizing estimates of predictive uncertainty. *J. Cheminform.* 1(1):1–16.

Collins, A., Tkaczyk, D. and Beel, J. (2018). One-at-a-time: A Meta-Learning Recommender-System for Recommendation-Algorithm Selection on Micro Level. *ArXiv preprint*:1805.12118.

Cortés-Ciriano, I. and Bender, A. (2018). Deep confidence: a computationally efficient framework for calculating reliable prediction errors for deep neural networks. *Journal of Chemical Information and Modeling*, 59(3), 1269-1281.

Dash, R., Dash, P. K. and Bisoi, R. (2015). A differential harmony search based hybrid interval type2 fuzzy EGARCH model for stock market volatility prediction. *International Journal of Approximate Reasoning*, 59, 81-104.

Datta A., Tschantz M.C., Datta A. (2015). Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, (1), 92-112.

Demut IR (2010) Reliability of Predictions in Regression Models. *PhD Conf.* 11–18.

Domingos P (2000) A Unified Bias-Variance Decomposition. *Proc. 17th Int. Conf. Mach. Learn.* 231–238.

Efron B (1979) Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* 7(1):1–26.

Efron B (2004) The estimation of prediction error: Covariance penalties and cross-validation. *J. Am. Stat. Assoc.* 99(467):619–632.

Geman S, Doursat R, Bienenstock E (1992) Neural Networks and the Bias Variance Dilemma. *Neural Comput.* 4:1–58.

Hakala, K., Van Landeghem, S., Salakoski, T., Van de Peer, Y. and Ginter, F. (2013). EVEX in ST'13: Application of a large-scale text mining resource to event extraction and network construction. *In Proceedings of the BioNLP Shared Task 2013 Workshop*, 26-34.

Halpe M (1963) Confidence Interval Estimation in Non-linear Regression. *J. of the Royal Statistical Society: Series B (Methodological)* 25(2):330–333.

Hand DJ, Yu K (2001) Idiot's Bayes---Not So Stupid After All? *Int. Stat. Rev.* 69(3):385–398.

Heskes T (1997) Practical conndence and prediction intervals. *Adv. Neural Inf. Process. Syst.* 176–182.

Ho SS, Wechsler H (2003) Transductive confidence machine for active learning. *Proc. Int. Jt. Conf. Neural Networks 2003* 2:1435–1440.

Hosanagar K. (2019). A Human's Guide to Machine Intelligence: How Algorithms are Shaping Our Lives and how We Can Stay in Control.

Huang, J., Zhu, L., Fan, B., Chen, Y., Jiang, W. and Li, S. (2018). Large-Scale Price Interval Prediction at OTA Sites. *IEEE Access*, 6, 807-817.

Hwang JTG, Ding AA, Hwang JTG, Ding AA (1997) Prediction Intervals for Artificial Neural Networks. *Am. Stat. Assoc.* 92(438):748–757.

Iorio, A., Spencer, F. A., Falavigna, M., Alba, C., Lang, E., Burnand, B. and Wolff, R. (2015). Use of GRADE for assessment of evidence about prognosis: rating confidence in estimates of event rates in broad categories of patients. *British Medical Journal Publishing Group*, 350, 1-8.

Johndrow, J. E. and Lum, K. (2019). An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1), 189-220.

Kharchenko, P. V., Silberstein, L. and Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, 11(7), 740.

Khosravi A, Nahavandi S, Creighton D (2010) Construction of optimal prediction intervals for load forecasting problems. *IEEE Trans. Power Syst.* 25(3):1496–1503.

Knafl G, Sacks J, Ylvisaker D (1985) Confidence bands for regression functions. *J. Am. Stat. Assoc.* 80(391):683–691.

Kohavi R (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *Int. Jt. Conf. Artif. Intell.* 1137–1143.

Kukar M, Kononenko I (2002) Reliable classifications with machine learning. *Eur. Conf. Mach. Learn. ECML 2002*. 1–8.

Lebedev, A. V., Westman, E., Van Westen, G. J. P., Kramberger, M. G., Lundervold, A., Aarsland, D. and

Vellas, B. (2014). Random Forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *NeuroImage: Clinical*, 6, 115-125.

Liu, R., Glover, K. P., Feasel, M. G. and Wallqvist, A. (2018). General approach to estimate error bars for quantitative structure–activity relationship predictions of molecular activity. *Journal of Chemical Information and Modeling*, 58(8), 1561-1575.

Melluish T, Saunders C, Nouretdinov I (2001) Comparing the Bayes and typicalness frameworks. *Eur. Conf. Mach. Learn. ECML 2001*. 360–371.

Nouretdinov I, Melluish T, Vovk V (2001) Ridge regression confidence machine. *Proc. Eighteenth Int. Conf. Mach. Learn.* 385–392.

Papadopoulos G, Edwards PJ, Murray AF (2001) Confidence estimation methods for neural networks: A practical comparison. *IEEE Trans. Neural Networks*. 1278–1287.

Picard RR, Cook RD (1984) Cross-Validation of Regression Models. *J. Am. Stat. Assoc.* 79(387):575–583.

Proedrou K, Ilia N, Volodya V, Alex G (2002) Transductive Confidence Machines for Pattern Recognition. *Eur. Conf. Mach. Learn. ECML2002*. 381–390.

Qian, Y., Tan, T., Hu, H. and Liu, Q. (2018). Noise robust speech recognition on aurora4 by humans and machines. *In IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 5604-5608.

Rasmussen CE (2004) Gaussian processes in machine learning. *Adv. Lect. Mach. Learn.* 63–71.

Saunders C, Gammerman A, Vovk V (1999) Transduction with confidence and credibility. *IJCAI Int. Jt. Conf. Artif. Intell.* 2(4):722–726.

Shao J (1996) Bootstrap model selection. *Am. Stat. Assoc.* 91(434):655–665.

Sheridan RP, Feuston BP, Maiorov VN, Kearsley SK (2004) Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* 44(6):1912–1928.

Shrestha DL, Solomatine DP (2006) Machine learning approaches for estimation of prediction interval for the model output. *Neural Networks* 19(2):225–235.

Simoiu, C., Corbett-Davies, S., and Goel, S. (2016). Testing for racial discrimination in police searches of motor vehicles. *SSRN* abs, 2811449.

Solares, E., Coello, C. A. C., Fernandez, E. and Navarro, J. (2019). Handling uncertainty through confidence intervals in portfolio optimization. *Swarm and Evolutionary Computation*, 44, 774-787.

Taylor P, Ye J (2012) On Measuring and Correcting the Effects of Data Mining and Model Selection On Measuring and Correcting the Effects of Data Mining and Model Selection. *Am. Stat. Assoc.* 93(441):120–131.

Tomassetti, S., Wells, A. U., Costabel, U., Cavazza, A., Colby, T. V., Rossi, G. and Tantalocco, P. (2016). Bronchoscopic lung cryobiopsy increases diagnostic confidence in the multidisciplinary diagnosis of idiopathic pulmonary fibrosis. *American J. of Respiratory and Critical Care Medicine*, 193(7), 745-752.

Toplak, M., Mocnik, R., Polajnar, M., Bosnić, Z., Carlsson, L., Hasselgren, C. and Stalring, J. (2014). Assessment of machine learning reliability methods for quantifying the applicability domain of QSAR regression models. *Journal of Chemical Information and Modeling*, 54(2), 431-441.

Tsanas A, Little MA, McSharry PE, Ramig LO (2010) Accurate telemonitoring of parkinsons disease progression by noninvasive speech tests. *IEEE Trans. Biomed. Eng.* 57(4):884–893.

Tzikas D, Kukar M, Likas A (2007) Transductive Reliability Estimation for Kernel Based Classifiers. *Adv. Intell. Data Anal.* 4723:37–47.

Walker SH, Duncan DB (1967) Estimation of the Probability of an Event as a Function of Several Independent Variables. *Biometrika* 54(1):167–179.

Wonnacott TH, Wonnacott RJ (1990) Confidence Intervals. *Introd. Stat.* 254–281.