Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 192 (2021) 3391–3402

Procedia Computer Science

www.elsevier.com/locate/procedia

25th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

Dataset shift assessment measures in monitoring predictive models

Aneta Becker[a*], Jarosław Becker[b]

[a]*West Pomeranian University of Technology, Faculty of Economics, Janickiego 31, 71-270 Szczecin, Poland*
[b]*The Jacob of Paradies University, Faculty of Technology, F. Chopina 52, 66-400 Gorzów Wielkopolski, Poland*

Abstract

The article presents the results of the study, which is a fragment of the work carried out under the project entitled "Hybrid system for intelligent diagnostics of prognostic models", co-financed through the National Centre for Research and Development from the European Regional Development Fund. They concerned the analysis of the phenomenon of dataset shift, also known as the shift of variable distributions. It is important to answer the question: has the distribution of current data for the implemented forecasting model changed significantly compared to the distribution of data used to develop it? If so, it could lead to incorrect operation. In the context of assessing and monitoring the stability of variable distributions of predictive models, the aim of the study was to compare the properties of two indicators, the Population Stability Index (PSI) and Population Accuracy Index (PAI). These measures were calculated for 78 controlled shifts of the distribution of the 3 explanatory variables of the hypothetical prognostic model. The research procedure was carried out in 2 scenarios. The first involved a comparison of PSI and PAI for the distributions of categorical variables. In scenario 2, an answer was sought to the question whether discretization of variables significantly influenced the assessment of the stability of their distributions using PSI compared to PAI, which does not require such a procedure? The results of the research proved that both indicators complement each other well, and when used together to assess the stability of the model's variable distributions, they compensate each other's shortcomings. PSI and PAI measure subtly different concepts of stability – PSI measures any change in the distribution of explanatory variables, and PAI only measures how this change affects the prognostic accuracy of the model – therefore they should be treated as complementary measures.

Keywords: Dataset shift, Population Stability Index (PSI), Population Accuracy Index (PAI), monitoring of predictive model;

* Corresponding author. Tel.: +48-91-449-68-50; fax: +48-91-422-46-52.
E-mail address: abecker@zut.edu.pl

1877-0509
10.1016/j.procs.2021.09.112

1. Introduction

A dataset shift takes place in a situation, where the joint distribution of inputs and outputs differs between the stages of training, testing, and using predictive models. Until recently, this phenomenon was relatively unnoticeable by researchers. Scientists dealing with, for example, machine learning have paid more attention to this subject. They focused on seemingly similar problems found in partially supervised or active learning. Today, in most practical applications, data set shifts are observed. Most often this is due to the subjectivism introduced by the experimental design or the inability to repeat the test conditions when training the models. Moreno-Torres et al. [1] list two, in their opinion, main reasons for the shift in the distribution of variables in classification studies: the error of sample selection and the occurrence of a non-stationary environment. In the first case, the shift is due to the fact that the training examples obtained do not reliably represent the operating environment in which the classifier would be implemented (which in terms of machine learning would constitute a test set). The second reason arises when the training environment differs from the test environment, whether it is due to a temporal or spatial change. According to Quionero-Candela et al. [2], dataset shift is one of the common problems in predictive modelling when the common distribution of input and output data differs between training and testing phases. According to Moreno-Torres et al. [1], a dataset shift occurs when a phenomenon occurs during data testing that leads to a change in the distribution of a single feature, combination of features, or class boundaries.

Most of the applications in use today must incorporate some form of change in their solutions. Therefore, research in this area is important and useful. However, their carrying out is difficult, for example due to the use of different concepts describing the same phenomenon in the literature. The issues related to the dataset shift are widely discussed in the literature related to the prediction of phenomena. In forecasting, the aim is to estimate the desired properties of future events, using generalizations obtained on the basis of previous experiences. Future events are expected to be almost identical to past events. Therefore, predictive models assume that the distribution of test data does not differ from the distribution of training data, which in fact is not always feasible. After implementing the predictive model, it is also important to monitor the distribution of new data. Too large deviations of this distribution in relation to the data distribution from the model construction phase may result in lower accuracy of forecasts and indicate the need to update or build a new model.

The aim of the article is to present the results of scientific research involving the analysis of the dataset shift, also known as the shift of the distribution of variables. In the practical part, the properties of two indicators, the *Population Stability Index* (PSI) and *Population Accuracy Index* (PAI) were examined in the context of assessing and monitoring the stability of the variable distributions of the predictive models. The issue of a dataset shift is so important that in a critical situation it leads to incorrect operation of the implemented model. These studies are part of the work carried out under the second stage of the project entitled "Hybrid system for intelligent diagnostics of prognostic models" (Measure 1.1. R&D projects of enterprises of the Intelligent Development Operational Program 2014-2020) [3], [4]. The hybrid system under development is used to model, evaluate and monitor prognostic models (validation of the existing models, creation of new "better" models) and managing their life cycle.

The article consists of 6 sections. The introduction defines the problem and research goal. Section 2 provides a brief overview of the *dataset shift* literature. Section 3 contains a formal description of both examined indicators (PSI and PAI). Section 4 describes the testing procedure and section 5 the results of the tests and comparative analyses obtained under the two scenarios. Section 5 provides a summary of the research and conclusions.

2. 2. Literature review on dataset shift

In the literature on the dataset shift, there is a disordered nomenclature, which makes it difficult to navigate the topic in question. Quiñonero-Candela et al. [2] used the term dataset shift for the first time and defined it as "cases where the overall distribution of inputs and outputs differs between training phase and test" [5]. Among other concepts that appear in the literature, and are intended to characterize the dataset shift in classification issues, the following can be distinguished: the concept of shift or the concept of drift [6], change of classification [7], changing surroundings (environment) [8], exploration of contrasts in learning classification [9], the concept of break points [10] and data brakes [1]. Moreno-Torres et al. [1] presented an extensive analysis of the different types of shifts identified in the actual classification tasks in their work. They discussed the problem of the covariate shift, defined for the first time by

Shimodair [11]. The mentioned type of shift is known as population drift [12], [13]. Another suggestion is an earlier probability shift and concerns changes in the class distribution [14]. On the other hand, the next approach, defined as a concept shift, is usually referred to in the literature as a "concept change" that occurs as a result of a changing context and may cause changes in the concepts of the target research [6].

The dataset shift problem, especially in the area of machine learning, has been fully described, among others, in [2], [13]. On the other hand, Storkey [5] classified various forms of dataset shift into six groups: covariate shift [15], [16], [17], prior probability shift [5], sample selection bias, [18], [19], [20] [21], [22], imbalanced data [5], domain shift [23], source component shift [5].

The occurrence of a dataset shift can to some extent be eliminated by using certain techniques. The literature proposes two main groups of methods based on: instances and distribution. Among the instance-based techniques, the methods for detecting outliers [24], [25], [26] deserve attention. Other proposals on this subject were presented by Kitchenham et al. [27] and Keung et al. [28], Turhan et al. [19] and Menzies et al. [20], Lin et al. [29]. The methods by which the relevance is filtered and the introduction of a controlled error of sample selection adapt to the samples of the test kit are presented by Turhan et al. [19] and Kocaguneli et al. [18] and Kocaguneli and Menzies [30], about the algorithms known as soft filtering can be found in [31]. Among the distribution-based methods, stratification used in *post hoc* analyses deserves attention, as well as the cost curve analysis in empirical studies of predictive models, which Drummond and Holte [32] as well as Jiang et al. [33] wrote about and taking into account the displacement of the source component of different origin [34], [5].

A comprehensive empirical study discussing how to combine dimension reduction and testing two samples to create a practical tool for detecting a distribution change in real machine learning systems is presented in Rabanser et al. [35]. In the literature related to machine learning, many probabilistic deep learning methods are proposed to quantify the predictive uncertainty. Noteworthy is the proposal for the practical application of deep neutral networks (DNN). Predictive distributions of these models, integrated with conventional approaches, for example, in medical diagnosis assisted by machine learning, in imaging [36] and in autonomous cars [37]. The software using DNN supports: vision systems in social networks [38], radiologists [39], Internet platforms [40], [41], speech recognition [42], [41]and translators [43].

Applications using DNN require point forecasts and qualification of predictive uncertainty. Safe implementation of machine learning requires model reliability in situations related to *out-of-distribution* (OOD) [44]. Probabilistic neural networks such as mixture density networks capture the intrinsic ambiguity of outputs for a given input [45]. Bayesian neural networks learn posterior decomposition on parameters that quantify parameter uncertainty, a kind of epistemic uncertainty that can be reduced by collecting additional data. Popular similar Bayesian approaches include the Laplace approximation [46], variational reasoning [47], [48], resignation-based variation inference [49], [50], expectation propagation [51] and MCMC stochastic gradient [52]. Non-Bayesian methods apply to training many probabilistic neural networks, for example with a bootstrap [53], [54].

The literature on the subject points out that machine learning-based software systems are sensitive and difficult to test [55]. Seemingly insignificant changes in data distributions may disrupt the operation of even the most modern classifiers [56], [57]. When decisions are made under conditions of uncertainty, even label shifts have an impact on their accuracy [58], [59]. Among the publications, the work by Lipton et al. [58] who discussed the Black Box Offset Detection (BBSD) method and proposed the use of an appropriate label classifier. Other literature proposals focus on anomaly detection, for example the article by Chandol et al. [24] and Markou and Singh [60]. Whereas Truong et al. [61] presented the classic problem of time series and detection of the change point in a data stream. A significant area of literature interest relates to the dataset shift in the context of domain adaptation. Note that shifts cannot be corrected without assumptions [62] and a covariate shift [63], [11], [17] or a label shift [64], [58], [65], [5], [59] is often assumed. In their work, Scholkopf et al. [66] presented a unified picture of these changes and combined the assumed invariants with the corresponding causal assumptions. Popular topics include outlier detection mechanisms, which are described in the literature as out of distribution (OOD) sample detection. More information on this was published by Hendrycks and Gimpel [67], Liang et al. [68] and Lee et al. [69]. Whereas Shafaei et al. [70] examined numerous techniques for the detection of OOD.

3. Measures of compliance of variable distributions – PSI and PAI

A popular measure to verify that the distribution of the current data has changed significantly compared to the distribution of the data used to develop the model is the *Population Stability Index* (PSI). The Population Accuracy Index (PAI) proposed by Taplin and Hunt [71] is an interesting alternative to the PSI. According to the authors, PAI can more accurately summarize the level of population stability and thus help risk analysts and managers determine if the model is working as intended.

The PSI population stability index is closely related to well-established entropy measures and is essentially a symmetric measure of the difference between two statistical distributions. Significant studies on PSI are the publications of Karakoulas [72] and Siddiqi [73]. For the purpose of determining the PSI, it is assumed that there are K mutually exclusive categories, numbered from 1 to K, and the mathematical formula can be written as:

$$PSI = \sum_{i=1}^{K}(O_i - E_i) \times ln\left(\frac{O_i}{E_i}\right), \quad (1)$$

where: O_i – the observed relative frequency of occurrences in category and during the review, E_i – the relative frequency of occurrences in the category and during the construction phase (the relative frequency during the survey is expected to be similar to the relative frequency during the model construction), i – the category taking values from 1 to K, $ln()$ – natural logarithm.

The PSI value equal to 0 means that the observed and expected distributions are identical, but as the two distributions are divergent, the value of the PSI index increases. According to Siddiqi [73], the PSI values can be classified as follows: below 10% there are no significant changes in stability, in the range of [10%, 25%] show little change requiring testing, above 25% show a significant change.

The Prediction Accuracy Index (PAI) is defined as the mean variable of the estimated mean response in a model review divided by the mean variance of the estimated mean response at the time of its development. The values of the explanatory variables (design space) are important, while the response values are irrelevant and are not required. A high PAI value occurs when, during the review, the explanatory variables assume values that cause the variance of the predicted response to be higher than the corresponding variance in the model development phase.

PAI measures the increase in the variance of the estimated mean response since the model was built. Taplin and Hunt [71] recommend using the following classification of PAI values: below 1.1 indicates no significant deterioration of the prognosis accuracy, in the range [1,1; 1,5] – deterioration requiring further investigation, above 1.5 – the prognostic accuracy of the model deteriorated significantly.

It is worth considering several situations when the PAI prognosis accuracy index can be determined.

1) For simple linear regression of the form (for $k = 1$):

$$\beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}, \quad (2)$$

where: $\beta_{0, \ldots,} \beta_k$ – estimated regression coefficients, $x_{i0}, \ldots, x_{ik}$ – values of the explanatory (numerical) variables for the i-th observation, the variance of the estimated mean response when the explanatory variable x_i is equal to z can be according to Ramsey and Schafer ([74], p. 187) written as:

$$MSE \times \left(\frac{1}{n} + \frac{(z-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right), \quad (3)$$

where: MSE – mean square error (of residuals) from the model development, $\bar{x}$ – mean value of the explanatory variable x_i at the development stage, n – sample size in the development stage.

PAI for a simple linear regression equals equation (3) averaged over all z values equal to the review data (denoted r_j; $j = 1, \ldots, N$) divided by equation (3) ave3raged over all z values equal to the development data (denoted x_i; $i = 1, \ldots, n$):

$$PAI = \frac{1}{2} \times \left(1 + \frac{\sum_{j=1}^{N}(r_j-\bar{x})^2/N}{\sum_{j=1}^{n}(x_i-\bar{x})^2/n}\right). \quad (4)$$

2) In the case of a multiple regression model:

$$\beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \quad (5)$$

estimated variance of the mean response when the explanatory variables $x_{i1}, x_{i2}, \ldots, x_{ip}$ take the values $z_{i1}, z_{i2}, \ldots, z_{ip}$ Johnson and Wichern ([75], p. 378) propose to be written as:

$$MSE \times z_j^T (X^T X)^{-1} z_j, \quad (6)$$

where: $z_j^T = z_{i1}, z_{i2}, \ldots, z_{ip}$ – vector of a row of explanatory variables, X – matrix of explanatory variables in the development stage, MSE – mean square error (residuals) from model development, T – transposition, $()^{-1}$ – matrix reverse.

The X columns are equal to the values of the explanatory variables of the development data (the rows are similar to z_j^T for each observation in the model development data). Equation (6) can be calculated by the formula:

$$z_j^T V z_j, \quad (7)$$

where $V = MSE \times (X^T X)^{-1}$ – the variance-covariance matrix of the estimated regression coefficients (1, 2, …, p).

PAI for multiple regression is defined as the mean of equation (7) calculated with the values of the explanatory variables z_j during the model review divided by the mean of equation (7) calculated with the values of explanatory variables z_j from the model development phase:

$$PAI = \frac{\sum_{j=1}^{N} r_j^T V r_j / N}{\sum_{j=1}^{n} x_i^T V x_i / n} \quad (8)$$

where: r_j – vector of explanatory variables for the j-th observation of the review data (j = 1, …, N), x_i – vector of explanatory variables for the i-th observation of the development data (i = 1, …, n) [71].

Equation (8) can be used for a multiple regression model and when one categorical variable has more than two feature variants, which requires constructing proxy variables to model the differences between these categories. PAI has the property of invariance, which means that any changes in the data related to its movement – from one category during the model development phase to another category during its review – do not affect the precision of the model, if both categories were estimated with the same precision in development phase. For each breakdown of the survey data, if they are evenly distributed among the categories during the construction phase of the model, the PAI will always be 1.

4. The course of the research procedure

The research procedure of the comparative analysis of population stability measures in the monitoring of prognostic models was carried out under 2 scenarios (1 – covered the dataset shift study for categorical variables, 2 – carried out for the corresponding continuous variables). Five main stages were distinguished in the research procedure. Stages 1 and 2 included the preparation of a special set of research data that allowed for the simulation of the phenomenon of shift in the distribution of the model variables in a fully controlled manner. The idea behind this procedure was to establish data distributions for a categorical variable, transform them into the form of distributions of a continuous variable in order to conduct comparative experiments. In order to compare the results of both research scenarios, the values of continuous variables were generated according to the order of observations in the set of categorical variables.

In stage 1 of the research procedure, a hypothetical explanatory variable x_{A1} o $n = 7$ with $(k_1, \ldots, k_n)$ categories was determined, and its individual distributions were described on the set containing $z = 100$ observations. The choice of the number of categories was deliberate due to the manual control of the shift in the distribution of variables and the proper illustration of the influence of persons migrating between the categories of observations on the shaping of the values of the compared indicators (PSI, PAI). The number of one hundred observations was indicated due to the ease of calculating the vector $C_r = [c_{r,i}, \ldots, c_{r,n}]$ of absolute frequencies of the observations (in terms of k_i, $i = 1, \ldots, n$ and for individual distributions $r = 1, \ldots, m$) on the relative values of the vector $E_r = [e_{r,i}, \ldots, e_{r,n}]$ during the

development of the model and the vector $O_r = [o_{r,i}, \dots, o_{r,n}]$ during the model review. In the next step of this stage, the vector of absolute frequency of occurrences was determined for the hypothetical categorical variable x_{A1} and for each k_i $(i = 1, \dots, 7)$ $C_{24} = [5,\ 10,\ 20,\ 30,\ 20,\ 10,\ 5]$. Then, 78 different data movements were determined against the C_{24} vector. These activities were performed using three different strategies. The first one concerned one-way, unbalanced in relation to the base C_{24} distribution, migration of observations between categories, from more frequent to less frequent (distributions C_{25}, …, C_{54}) and vice versa, from less frequent to more frequent categories (C_{23}, , C_1). The second strategy consisted in a bidirectional migration of observations in terms of C_{24} in extreme categories: $k_1, k_2\ i\ k_6, k_7$ from more to less frequent and in central categories: k_3, k_4, k_5 from less to more frequent (distributions C_{60}, …, C_{69}). These activities were repeated, changing the roles of the extreme categories with the central ones, and the distributions C_{59}, …, C_{55} were obtained. The third strategy included a one-way, C_{24} balanced migration of observations between categories. The samples were moved separately in extreme and central categories, always in one direction, from more to less frequent (C_{74}, …, C_{70}) and from less to more frequent (C_{75}, …, C_{79}). In total, 79 distributions of the categorical variable were established (C_1, …, C_{79}; tab. 1).

Stage 2 consisted in transforming m = 79 declared frequency vectors of the categorical variable C_r $(r = 1, \dots, m)$ to the form of vectors of values $V_r = [v_{r,1},\ \dots,\ v_{r,z}]$, where $z = 100$ observations. First, exemplary ranges of continuous variable values were determined for each category k_i $(i = 1, \dots, 7)$: p_1= ⟨30; 150), p_2= ⟨150; 160), p_3= ⟨160; 164), p_4= ⟨164; 167), p_5= ⟨167; 170), p_6= ⟨170; 180), p_7= ⟨180; 240). Then, for each interval p_i real numbers were drawn $c_{r,i}$, which together gave the one-element vectors V_r. The *Mersenne-Twister* generator available in the "R" programming package was used to determine the pseudo-random numbers.

In stage 3 of the research procedure, from among all the distributions C_1, …, C_{79}, three were selected that meet the condition of no full collinearity: C_{16}, C_{24}, C_{43} and were established with the distributions of 3 explanatory variables: x_{A1}, x_{A2}, x_{A3} of a hypothetical model in the phase of its development. The remaining distributions of categorical variables were treated in the PAI_A and PSI_A calculations as if they came from the model review phases. Scenario 2 was a follow-up study. It verified whether the discretization of variables can significantly affect the assessment of population stability using PSI_A compared to PAI_B, which does not require such a procedure. Therefore, the generated values of V_r vectors were established in PAI_B calculations with distributions of continuous variables: x_{B1}, x_{B2}, x_{B3}.

Stage 4 of the research procedure was required in the PAI_A calculations (scenario 1) and involved the quantification of qualitative factors (categorical variables) with binary variables. This was done by transforming n categories into $n - 1$ artificial variables (zero-one).

Stage 5 consisted in the development of algorithms in the "R" program and the performance of PSI_A, PAI_A calculations for categorical variables and PAI_B for continuous variables in accordance with (1) – (8) contained in section 5. In the final part of this stage, the results obtained in under both research scenarios.

5. Discussion of the obtained results

The results of the study, including a comparative analysis of univariate population stability assessment indicators, are summarized in Table 1 – PSI and PAI values for categorical variables (scenario 1) and PAI values for continuous variables (scenario 2). Blue marks the fields for situations where it was impossible to calculate the PSI value, i.e. for the distributions of categorical variables from the model review phase, for which the absolute frequency of observations in the k_i category was zero ($c_{r;i} = 0;\ r = 1, \dots, 79$), which meant the disappearance of the k_i category in the model review phase. The remaining fields are marked with colours that reflect the three interpretable ranges of values for PSI and PAI recommended in the literature [71], [76]:

1) safe area (green, normal conditions) – population stability (PSI < 0.1; PAI < 1.1),
2) hazard warning area (yellow, warning conditions) – slight deterioration of population stability (0.1 ≤ PSI ≤ 0.25 and 1.1 ≤ PAI ≤ 1.5),
3) critical area (red, risky conditions) – significant instability of the population (PSI > 0.25; PAI > 1.5).

Calculations in scenarios 1 and 2 were performed for $m - 1 = 78$ shifts of the base distribution (from the model development phase) for each variable. In Table 1, the lines with the base distributions are marked in grey. The values $PSI_A = 0$ and $PAI_A = PAI_B = 1$ included in them mean that the distributions from the model construction and review phase are identical.

Omitting 20 cases for which it was impossible to calculate the PSI (Table 1, blue), out of 78 different shifts of the distribution, i.e. for 58 cases and each explanatory variable x_{A1}, x_{A2}, x_{A3}, a comparative analysis was made to assess the degree of stability of the population of both indicators. The number of cases for which the population stability assessments using $\mathrm{PSI}_{A1}, \mathrm{PSI}_{A2}, \mathrm{PSI}_{A3}$ and $\mathrm{PAI}_{A1}, \mathrm{PAI}_{A2}, \mathrm{PAI}_{A3}$ were different, was as follows: for the explanatory variable x_{A1} it was 19 cases (32.8%), for x_{A2} 16 (27.6%) and for x_{A3} as many as 44 cases (75.9%). Comparing the colours of the fields representing the population stability classes defined by PSI_A and PAI_A for individual explanatory variables, it should be stated that signalling about the threat or instability of the population (yellow and orange) in many cases lies both with PSI_A and PAI_A.

In scenario 2, it was verified whether PAI_B calculated for the distributions of continuous variables, signals significantly differently about the degree of risk of population instability compared to PSI_A and PAI_A calculated for categorical variables (especially in relation to PSI_A, which requires data discretization).

For 58 cases of categorical variables and the corresponding continuous variables, a comparative analysis was performed to assess the degree of stability of the population of indicators: $\mathrm{PSI}_{A1}, \mathrm{PSI}_{A2}, \mathrm{PSI}_{A3}$ with: $\mathrm{PAI}_{B1}, \mathrm{PAI}_{B2}, \mathrm{PAI}_{B3}$. The number of cases for which the population stability assessments obtained using both indicators were different, was as follows: for variables x_{A1} and x_{B1} there were 18 cases (31%), for x_{A2} and x_{B2} 13 (22.4%) and for x_{A3} and x_{B3} as many as 46 cases (79.3%). The results of comparing PSI_A values with PAI_B are very close to the results of scenario 1 for PSI_A and PAI_A.

Table 1. Comparison of the values of the compliance assessments of the distributions of explanatory variables

C_r	The absolute frequency of the observations in categories k_i (i = 1, …, 7)							Indicator values for categorical variables						V_r	Indicator values for continuous variables		
								PSI_A			PAI_A				PAI_B		
	$c_{r,1}$	$c_{r,2}$	$c_{r,3}$	$c_{r,4}$	$c_{r,5}$	$c_{r,6}$	$c_{r,7}$	x_{A1}	x_{A2}	x_{A3}	x_{A1}	x_{A2}	x_{A3}		x_{B1}	x_{B2}	x_{B3}
C_1	0	0	20	60	20	0	0	-	-	-	0.5714	0.4637	1.3766	V_1	0.5095	0.5218	0.5171
C_2	0	0	21	58	21	0	0	-	-	-	0.5762	0.4680	1.3740	V_2	0.5092	0.5205	0.5171
C_3	0	0	22	56	22	0	0	-	-	-	0.5810	0.4724	1.3714	V_3	0.5105	0.5238	0.5181
C_4	0	0	23	54	23	0	0	-	-	-	0.5857	0.4768	1.3688	V_4	0.5106	0.5247	0.5178
C_5	0	0	24	52	24	0	0	-	-	-	0.5905	0.4812	1.3662	V_5	0.5113	0.5264	0.5183
C_6	0	0	25	50	25	0	0	-	-	-	0.5952	0.4856	1.3636	V_6	0.5108	0.5245	0.5182
C_7	0	0	26	48	26	0	0	-	-	-	0.6000	0.4900	1.3610	V_7	0.5099	0.5227	0.5175
C_8	0	0	27	46	27	0	0	-	-	-	0.6048	0.4944	1.3584	V_8	0.5116	0.5277	0.5184
C_9	0	0	28	44	28	0	0	-	-	-	0.6095	0.4987	1.3558	V_9	0.5112	0.5277	0.5177
C_{10}	0	1	27	44	27	1	0	-	-	-	0.6238	0.5345	1.3442	V_{10}	0.5129	0.5336	0.5182
C_{11}	0	2	27	42	27	2	0	-	-	-	0.6429	0.5746	1.3299	V_{11}	0.5148	0.5399	0.5189
C_{12}	0	3	26	42	26	3	0	-	-	-	0.6571	0.6103	1.3182	V_{12}	0.5158	0.5441	0.5190
C_{13}	0	4	26	40	26	4	0	-	-	-	0.6762	0.6504	1.3039	V_{13}	0.5175	0.5488	0.5200
C_{14}	0	5	25	40	25	5	0	-	-	-	0.6905	0.6861	1.2922	V_{14}	0.5239	0.5723	0.5218
C_{15}	0	6	25	38	25	6	0	-	-	-	0.7095	0.7262	1.2779	V_{15}	0.5282	0.5877	0.5231
C_{16}	**1**	**6**	**24**	**38**	**24**	**6**	**1**	0.2031	**0**	1.5999	0.7524	**1**	1.2724	**V_{16}**	0.6938	**1**	0.5754
C_{17}	1	7	24	36	24	7	1	0.1757	0.0042	1.4950	0.7714	1.0401	1.2581	V_{17}	0.6591	1.0508	0.5644
C_{18}	2	7	23	36	23	7	2	0.0957	0.0189	1.2500	0.8143	1.3139	1.2525	V_{18}	0.6364	0.9728	0.5562
C_{19}	2	8	23	34	23	8	2	0.0773	0.0307	1.1577	0.8333	1.3540	1.2382	V_{19}	0.7223	1.2721	0.5852
C_{20}	3	8	22	34	22	8	3	0.0382	0.0634	1.0050	0.8762	1.6278	1.2327	V_{20}	0.9486	2.0785	0.6542
C_{21}	3	9	22	32	22	9	3	0.0276	0.0821	0.9230	0.8952	1.6679	1.2184	V_{21}	1.0089	2.2888	0.6743
C_{22}	4	9	21	32	21	9	4	0.0088	0.1258	0.8114	0.9381	1.9417	1.2128	V_{22}	0.7413	1.3281	0.5954
C_{23}	4	10	21	30	21	10	4	0.0054	0.1510	0.7382	0.9571	1.9818	1.1985	V_{23}	0.7273	1.2929	0.5854
C_{24}	**5**	**10**	**20**	**30**	**20**	**10**	**5**	**0**	0.2031	0.6513	**1**	2.2556	1.1929	**V_{24}**	**1**	1.8745	0.7273
C_{25}	5	11	20	28	20	11	5	0.0033	0.2345	0.5859	1.0190	2.2957	1.1787	V_{25}	0.9219	1.9824	0.6464
C_{26}	6	11	19	28	19	11	6	0.0080	0.2937	0.5160	1.0619	2.5695	1.1731	V_{26}	0.9256	1.9899	0.6497
C_{27}	6	12	19	26	19	12	6	0.0177	0.3313	0.4576	1.0810	2.6096	1.1588	V_{27}	0.9682	2.1468	0.6607
C_{28}	7	12	18	26	18	12	7	0.0307	0.3967	0.4006	1.1238	2.8835	1.1532	V_{28}	1.3112	3.3561	0.7704
C_{29}	7	13	18	24	18	13	7	0.0468	0.4406	0.3489	1.1429	2.9236	1.1390	V_{29}	1.1943	2.9444	0.7328
C_{30}	8	13	17	24	17	13	8	0.0671	0.5120	0.3023	1.1857	3.1974	1.1334	V_{30}	1.2516	3.1396	0.7538
C_{31}	8	14	17	22	17	14	8	0.0897	0.5624	0.2568	1.2048	3.2375	1.1191	V_{31}	1.1739	2.8689	0.7276
C_{32}	9	14	16	22	16	14	9	0.1166	0.6394	0.2191	1.2476	3.5113	1.1135	V_{32}	1.5526	4.2142	0.8448
C_{33}	9	15	16	20	16	15	9	0.1460	0.6969	0.1797	1.2667	3.5514	1.0993	V_{33}	1.1409	2.7523	0.7172
C_{34}	10	15	15	20	15	15	10	0.1792	0.7795	0.1498	1.3095	3.8252	1.0937	V_{34}	1.6856	4.6806	0.8882

C_r	The absolute frequency of the observations in categories k_i (i = 1, …, 7)							Indicator values for categorical variables						V_r	Indicator values for continuous variables		
								PSI_A			PAI_A				PAI_B		
	$c_{r,1}$	$c_{r,2}$	$c_{r,3}$	$c_{r,4}$	$c_{r,5}$	$c_{r,6}$	$c_{r,7}$	x_{A1}	x_{A2}	x_{A3}	x_{A1}	x_{A2}	x_{A3}		x_{B1}	x_{B2}	x_{B3}
C_{35}	10	16	15	18	15	16	10	0.2158	0.8447	0.1166	1.3286	3.8653	1.0794	V_{35}	1.5271	4.1338	0.8331
C_{36}	11	16	14	18	14	16	11	0.2551	0.9330	0.0938	1.3714	4.1391	1.0738	V_{36}	1.5041	4.0260	0.8358
C_{37}	11	17	14	16	14	17	11	0.2997	1.0068	0.0669	1.3905	4.1792	1.0596	V_{37}	1.5654	4.2621	0.8478
C_{38}	12	17	13	16	13	17	12	0.3452	1.1010	0.0508	1.4333	4.4530	1.0540	V_{38}	1.4884	3.9817	0.8266
C_{39}	12	18	13	14	13	18	12	0.3989	1.1849	0.0305	1.4524	4.4931	1.0397	V_{39}	2.0815	6.0786	1.0140
C_{40}	13	18	12	14	12	18	13	0.4506	1.2853	0.0209	1.4952	4.7669	1.0341	V_{40}	1.7602	4.9361	0.9150
C_{41}	13	19	12	12	12	19	13	0.5151	1.3813	0.0079	1.5143	4.8070	1.0199	V_{41}	2.0665	6.0390	1.0041
C_{42}	14	19	11	12	11	19	14	0.5734	1.4884	0.0047	1.5571	5.0808	1.0143	V_{42}	1.7494	4.9076	0.9079
$\boldsymbol{C_{43}}$	**14**	**20**	**11**	**10**	**11**	**20**	**14**	0.6513	1.5999	**0**	1.5762	5.1209	**1**	$\boldsymbol{V_{43}}$	2.1074	6.1834	**1**
C_{44}	15	20	10	10	10	20	15	0.7167	1.7143	0.0033	1.6190	5.3947	0.9944	V_{44}	1.9079	5.4652	0.9590
C_{45}	15	21	10	8	10	21	15	0.8124	1.8467	0.0087	1.6381	5.4348	0.9801	V_{45}	2.1116	6.2032	1.0165
C_{46}	16	21	9	8	9	21	16	0.8856	1.9693	0.0188	1.6810	5.7086	0.9746	V_{46}	2.2983	6.8466	1.0820
C_{47}	16	22	9	6	9	22	16	1.0071	2.1325	0.0376	1.7000	5.7487	0.9603	V_{47}	2.7508	8.4497	1.2235
C_{48}	17	22	8	6	8	22	17	1.0891	2.2646	0.0550	1.7429	6.0226	0.9547	V_{48}	2.4405	7.3544	1.1249
C_{49}	17	23	8	4	8	23	17	1.2540	2.4805	0.0941	1.7619	6.0627	0.9404	V_{49}	2.3160	6.9199	1.0834
C_{50}	18	23	7	4	7	23	18	1.3464	2.6240	0.1196	1.8048	6.3365	0.9349	V_{50}	2.2915	6.8345	1.0752
C_{51}	18	24	7	2	7	24	18	1.6094	2.9607	0.1996	1.8238	6.3766	0.9206	V_{51}	2.0298	5.8886	1.0004
C_{52}	19	24	6	2	6	24	19	1.7143	3.1181	0.2345	1.8667	6.6504	0.9150	V_{52}	2.6990	8.2784	1.2026
C_{53}	19	25	6	0	6	25	19	-	-	-	1.8857	6.6905	0.9007	V_{53}	2.2899	6.8111	1.0815
C_{54}	20	25	5	0	5	25	20	-	-	-	1.9286	6.9643	0.8952	V_{54}	2.4663	7.4377	1.1360
C_{55}	0	15	25	20	25	15	0	-	-	-	0.8810	1.0871	1.1494	V_{55}	0.5435	0.6419	0.5278
C_{56}	1	14	24	22	24	14	1	0.1951	0.2230	1.0264	0.9048	1.3208	1.1581	V_{56}	0.6965	1.1836	0.5759
C_{57}	2	13	23	24	23	13	2	0.0925	0.1873	0.8269	0.9286	1.5545	1.1668	V_{57}	0.7939	1.5314	0.6054
C_{58}	3	12	22	26	22	12	3	0.0373	0.1761	0.7260	0.9524	1.7882	1.1755	V_{58}	0.8385	1.6898	0.6192
C_{59}	4	11	21	28	21	11	4	0.0087	0.1823	0.6728	0.9762	2.0219	1.1842	V_{59}	0.9281	2.0050	0.6481
C_{60}	6	9	19	32	19	9	6	0.0081	0.2372	0.6546	1.0238	2.4893	1.2017	V_{60}	0.9713	2.1539	0.6632
C_{61}	7	8	18	34	18	8	7	0.0316	0.2840	0.6796	1.0476	2.7231	1.2104	V_{61}	1.3529	3.5073	0.7822
C_{62}	8	7	17	36	17	7	8	0.0703	0.3436	0.7254	1.0714	2.9568	1.2191	V_{62}	1.2324	3.0732	0.7470
C_{63}	9	6	16	38	16	6	9	0.1247	0.4164	0.7926	1.0952	3.1905	1.2278	V_{63}	1.3350	3.4435	0.7766
C_{64}	10	5	15	40	15	5	10	0.1962	0.5037	0.8835	1.1190	3.4242	1.2365	V_{64}	1.7092	4.7626	0.8964
C_{65}	11	4	14	42	14	4	11	0.2877	0.6076	1.0032	1.1429	3.6579	1.2453	V_{65}	1.4432	3.8317	0.8087
C_{66}	12	3	13	44	13	3	12	0.4051	0.7319	1.1616	1.1667	3.8916	1.2540	V_{66}	1.5007	4.0288	0.8292
C_{67}	13	2	12	46	12	2	13	0.5605	0.8851	1.3815	1.1905	4.1253	1.2627	V_{67}	1.7473	4.8882	0.9118
C_{68}	14	1	11	48	11	1	14	0.7920	1.0915	1.7345	1.2143	4.3590	1.2714	V_{68}	1.9572	5.6369	0.9755
C_{69}	15	0	10	50	10	0	15	-	-	-	1.2381	4.5927	1.2801	V_{69}	1.9911	5.7614	0.9844
C_{70}	10	5	25	20	25	5	10	0.2015	0.5345	0.7420	1.1667	3.4680	1.2106	V_{70}	1.0446	2.4044	0.6897
C_{71}	9	6	24	22	24	6	9	0.1273	0.4390	0.6788	1.1333	3.2256	1.2071	V_{71}	1.1296	2.7221	0.7099
C_{72}	8	7	23	24	23	7	8	0.0714	0.3594	0.6397	1.1000	2.9831	1.2035	V_{72}	1.2050	2.9847	0.7352
C_{73}	7	8	22	26	22	8	7	0.0319	0.2940	0.6223	1.0667	2.7406	1.2000	V_{73}	0.9900	2.2169	0.6703
C_{74}	6	9	21	28	21	9	6	0.0081	0.2421	0.6259	1.0333	2.4981	1.1965	V_{74}	1.0436	2.4134	0.6846
C_{75}	4	11	19	32	19	11	4	0.0087	0.1775	0.7015	0.9667	2.0132	1.1894	V_{75}	0.8418	1.6920	0.6238
C_{76}	3	12	18	34	18	12	3	0.0369	0.1661	0.7833	0.9333	1.7707	1.1859	V_{76}	0.6534	1.0286	0.5633
C_{77}	2	13	17	36	17	13	2	0.0914	0.1715	0.9126	0.9000	1.5282	1.1824	V_{77}	0.7344	1.3150	0.5889
C_{78}	1	14	16	38	16	14	1	0.1924	0.2004	1.1402	0.8667	1.2857	1.1788	V_{78}	0.6649	1.0699	0.5666
C_{79}	0	15	15	40	15	15	0	-	-	-	0.8333	1.0432	1.1753	V_{79}	0.5388	0.6250	0.5265

In the next step, within 78 different shifts of the distribution of 3 variables, the following were compared: $\mathrm{PAI}_{A1}, \mathrm{PAI}_{A2}, \mathrm{PAI}_{A3}$ with: $\mathrm{PAI}_{B1}, \mathrm{PAI}_{B2}, \mathrm{PAI}_{B3}$. The number of cases for which the population stability assessments, calculated using PAI_A for categorical variables and PAI_B for continuous variables, differed and got summarized in Table 1. Fora x_{A1} and x_{B1} it was 16 cases (20.5%), for x_{A2} and x_{B2} 6 (7.7%), and for xd_{A3} and xd_{B3} 61 cases (78.2%). The question is, do differences in population stability assessments favour of continuous or categorical variables? The results indicate that a hypothetical model based on continuous variables x_{B1}, x_{B2}, x_{B3} could demonstrate greater prognostic accuracy, as these variables turned out to be less sensitive to the change of distribution compared to their categorical counterparts. It is clearly noticeable in the values of PAI_{A3} and PAI_{B3} (Table 1). The PAI_{A3} index for the

categorical variable x_{A3} in 57 cases out of 78 examined indicated the deterioration of the prognostic accuracy of the model, while PAI_{B3} for the continuous variable x_{B3} only in four: V_{47}, V_{48}, V_{52} *and* V_{54}.

The PSI assessment of population stability in models based on continuous variable distributions always requires discretization, which can significantly distort this assessment, which was confirmed by the results obtained in scenario 2. In this context, the PSI is a less reliable measure, as its values depend on the adopted method of discretization.

6. Final conclusions

PSI and PAI measure subtly different aspects of stability. The PSI reflects any change in the distribution of the explanatory variables, and the PAI only indicates how these changes affect the prognostic accuracy of the model. PAI in relation to the commonly used PSI index is a complementary solution. The results of the comparative studies showed that these indicators complement each other, and their simultaneous use to assess the stability of the model variable distributions compensates for their shortcomings (Table 2).

Table 2. Comparison of the properties of the PSI and PAI indicators

Properties	**PSI**	**PAI**
Measures the overall difference in the distribution of data from review and model development	+	-
Measures the predictability of the model on the basis of changes in the distribution of explanatory variables.	-	+
It is used to evaluate a shift in the distribution of a categorical variable.	+	+
It is applicable to distributions of equal frequency of observations in categories.	+	-
It can be calculated when a category is no longer present in the review data (the frequency of observations in the category is zero).	-	+
It is used to evaluate the shift of the continuous variable distribution (it does not require discretization of the continuous variable value)	-	+
It is used as a multivariate measure that is used to evaluate a shift in a multivariate distribution.	-	+

PAI is directly applicable to explanatory variables that are numeric or categorical (ordered or unordered). It can be determined even when the categories have frequencies close to zero in the development or review data. It can also be derived from the review data and without making any assumptions of linearity considered appropriate in designing the model. PAI's drawbacks include the use of variable segmentation to remove the effect of outliers.

If during the model construction phase a transformation of statistical data consisting in limiting the impact of outliers or minor observation errors was used, then when calculating PAI, these actions should be taken into account, because this indicator for the data from the review may be strongly influenced by several extreme outliers, which was confirmed in work [71]. However, if it is desired to monitor modelling decisions with such transformations in mind, it is recommended to calculate PAI with and without these activities. In cases of one-dimensional distribution for each variable, PAI or PSI values may expose the variables responsible for instability. The monitoring reports obtained in this way will allow to assess whether the model in the review phase meets the assumed goal or not.

The conducted research does not exhaust the issues related to detecting dataset shift in order to monitor statistical models. This phenomenon is so new and complex that it requires further research and experimental research. An important research problem and an important direction for further analyses is the assessment of the dataset shift in a multidimensional approach, e.g. testing the properties of the Multivariate Predictive Accuracy Index (MPAI), especially due to the fact that this issue is rarely present in the literature on the subject.

Acknowledgements

The project was co-financed through the National Centre for Research and Development with headquarters in Warsaw (contract no. POIR.01.01.01-00-0322/18-00) from the European Regional Development Fund. Contractor: BD Polska Sp. z o.o. with its registered office in Warsaw at 9/11 Wierzbowa street, Poland. Subcontractor: Jacob of Paradyz University with headquarters in Gorzów Wielkopolski, 25 Teatralna street, Poland.

References

[1] Moreno-Torres JG, Raeder T, Alaiz-Rodríguez R, Chawla NV, Herrera F. (2012) "A unifying view on dataset shift in classification." *Pattern Recognition* 45:521–30. https://doi.org/10.1016/j.patcog.2011.06.019.

[2] Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, editors. (2008) "Dataset Shift in Machine Learning." First edition. Cambridge, Mass: The MIT Press.

[3] Ziemba P, Radomska-Zalas A, Becker J. (2020) "Client evaluation decision models in the credit scoring tasks." Procedia Computer Science, 176:3301–9. https://doi.org/10.1016/j.procs.2020.09.068.

[4] Becker J, Radomska-Zalas A, Ziemba P. (2020) "Rough set theory in the classification of loan applications." Procedia Computer Science, 176:3235–44. https://doi.org/10.1016/j.procs.2020.09.125.

[5] Storkey A. (2009) "When Training and Test Sets Are Different: Characterizing Learning Transfer." in: Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence (eds), *Dataset Shift in Machine Learning.*" p. 3–28.

[6] Widmer G, Kubat M. (1996) "Learning in the Presence of Concept Drift and Hidden Contexts." *Machine Learning*, 23:69–101.

[7] Wang K, Zhou S, Fu CA, Yu JX. (2003) "Mining Changes of Classification by Correspondence Tracing". Proceedings of the 2003 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, p. 95–106. https://doi.org/10.1137/1.9781611972733.9.

[8] Alaiz-Rodríguez R, Japkowicz N. (2008) "Assessing the Impact of Changing Environments on Classifier Performance." in: Bergler S. (ed.), *Advances in Artificial Intelligence,* Berlin, Heidelberg: Springer, p. 13–24. https://doi.org/10.1007/978-3-540-68825-9_2.

[9] Yang Y, Wu X, Zhu X. (2008) "Conceptual equivalence for contrast mining in classification learning." *Data & Knowledge Engineering,* 67:413–29. https://doi.org/10.1016/j.datak.2008.07.001.

[10] Cieslak DA, Chawla NV. (2009) "A framework for monitoring classifiers' performance: when and why failure occurs?" *Knowl Inf Syst,* 18:83–108. https://doi.org/10.1007/s10115-008-0139-1.

[11] Shimodaira H. (2000) "Improving predictive inference under covariate shift by weighting the log-likelihood function." *Journal of Statistical Planning and Inference,* 90:227–44. https://doi.org/10.1016/S0378-3758(00)00115-4.

[12] Kelly MG, Hand DJ, Adams NM. (1999) "The impact of changing populations on classifier performance." Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, San Diego, California, USA: Association for Computing Machinery, p. 367–71. https://doi.org/10.1145/312129.312285.

[13] Hand DJ. (2006) "Rejoinder: Classifier Technology and the Illusion of Progress." *Statist Sci,* 21:30–34.

[14] Webb GI, Ting KM. (2005) "On the Application of ROC Analysis to Predict Classification Performance Under Varying Class Distributions." *Machine Learning* 58:25–32. https://doi.org/10.1007/s10994-005-4257-7.

[15] Bickel S, Brückner M, Scheffer T. (2009) "Discriminative Learning Under Covariate Shift." *J. of Machine Learning Research,* 10:2137–55.

[16] Huang J, Smola AJ, Gretton A, Borgwardt KM, Schölkopf B. (2006) "Correcting Sample Selection Bias by Unlabeled Data." in: Bernhard Schölkopf, John Platt, Thomas Hofmann (eds), *Advances in Neural Information Processing Systems 19,* Proceedings of the 2006 Conference, The MIT Press. https://doi.org/10.7551/mitpress/7503.003.0080.

[17] Sugiyama M, Suzuki T, Nakajima S, Kashima H, von Bünau P, Kawanabe M. (2008) "Direct importance estimation for covariate shift adaptation." *Ann Inst Stat Math,* 60:699–746. https://doi.org/10.1007/s10463-008-0197-x.

[18] Kocaguneli E, Gay G, Menzies T, Yang Y, Keung JW. (2010) "When to use data from other projects for effort estimation." Proceedings of the IEEE/ACM international conference on Automated software engineering, Antwerp, Belgium: Association for Computing Machinery, p. 321–4. https://doi.org/10.1145/1858996.1859061.

[19] Turhan B, Menzies T, Bener AB, Di Stefano J. (2009) "On the relative value of cross-company and within-company data for defect prediction." *Empir Software Eng,* 14:540–78. https://doi.org/10.1007/s10664-008-9103-7.

[20] Menzies T, Turhan B, Bener A, Gay G, Cukic B, Jiang Y. (2008) "Implications of ceiling effects in defect predictors. Proceedings of the 4th international workshop on Predictor models in software engineering, Leipzig, Germany: Association for Computing Machinery, p. 47–54. https://doi.org/10.1145/1370788.1370801.

[21] Bakir A, Turhan B, Bener A. (2010) "A new perspective on data homogeneity in software cost estimation: A study in the embedded systems domain." *Software Quality Journal,* 18:57–80. https://doi.org/10.1007/s11219-009-9081-z.

[22] Premraj R, Zimmermann T. (2007) "Building Software Cost Estimation Models using Homogenous Data." First International Symposium on Empirical Software Engineering and Measurement (ESEM 2007). https://doi.org/10.1109/esem.2007.34.

[23] Demirors O, Gencel C. (2009) "Conceptual Association of Functional Size Measurement Methods." *IEEE Software,* 26:71–8.

[24] Chandola V, Banerjee A, Kumar V. (2009) "Anomaly detection: A survey." *ACM Comput Surv,* 41:15:1-15:58.

[25] Briand LC, Melo WL, Wust J. (2002) "Assessing the applicability of fault-proneness models across object-oriented software projects." *IEEE Transactions on Software Engineering* 28:706–20. https://doi.org/10.1109/TSE.2002.1019484.

[26] Briand LC, Wüst J. (2002) "Empirical Studies of Quality Models in Object-Oriented Systems." in: Zelkowitz M.V. (ed.), *Advances in Computers,* vol. 56, Elsevier, p. 97–166. https://doi.org/10.1016/S0065-2458(02)80005-5.

[27] Kitchenham BA, Mendes E, Travassos GH. (2007) "Cross versus Within-Company Cost Estimation Studies: A Systematic Review." *IEEE Transactions on Software Engineering* 33:316–29. https://doi.org/10.1109/TSE.2007.1001.

[28] Keung J, Kitchenham B, Jeffery R. (2007) "Analogy-X: Providing Statistical Inference to Analogy-Based Software Cost Estimation." *Software Engineering, IEEE Transactions*, 34:471–84. https://doi.org/10.1109/TSE.2008.34.

[29] Lin J, Keogh E, Lonardi S, Lankford JP, Nystrom DM. (2004) "Visually mining and monitoring massive time series." Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD'04, Seattle, WA, USA: ACM Press; 2004, p. 460.

[30] Kocaguneli E, Menzies T. (2011) "How to Find Relevant Data for Effort Estimation?" in: 2011 International Symposium on Empirical Software Engineering and Measurement, p. 255–64. https://doi.org/10.1109/ESEM.2011.34.
[31] Zhang H, Sheng S. (2004) "Learning weighted naive Bayes with accurate ranking." in: Fourth IEEE International Conference on Data Mining (ICDM'04), p. 567–570. https://doi.org/10.1109/ICDM.2004.10030.
[32] Drummond C, Holte RC. (2006) "Cost curves: An improved method for visualizing classifier performance." *Machine Learning* 65:95–130.
[33] Jiang Y, Cukic B, Ma Y. (2008) "Techniques for evaluating fault prediction models." *Empirical Software Engineering* 13:561–95.
[34] Alpaydin E. (2010) "Introduction to Machine Learning." 2nd ed. The MIT Press.
[35] Rabanser S, Günnemann S, Lipton Z. (2019) "Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift." in: Wallach H, Larochelle H, Beygelzimer A, Alché-Buc F, Fox E, Garnett R (eds). *Advances in Neural Information Processing Systems 32,* Curran Associates, Inc.; 2019, p. 1396–1408. https://debug-ml-iclr2019.github.io/cameraready/DebugML-19_paper_20.pdf
[36] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. (2017) "Dermatologist-level classification of skin cancer with deep neural networks." *Nature* 542:115–118. https://doi.org/10.1038/nature21056.
[37] Bojarski M, Testa DD, Dworakowski D, Firner B, Flepp B, Goyal P, Jackel LD, Monfort M, Muller U, Zhang J, Zhang X, Zhao J, Zieba K. (2016) "End to End Learning for Self-Driving Cars." ArXiv:160407316 [Cs]. https://arxiv.org/pdf/1704.07911.pdf
[38] Stone Z, Zickler T, Darrell T. (2008) "Autotagging Facebook: Social network context improves photo annotation." in: 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, p. 1–8. https://doi.org/10.1109/CVPRW.2008.4562956.
[39] Lakhani P, Sundaram B. (2017) "Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks." *Radiology* 284:574–82. https://doi.org/10.1148/radiol.2017162326.
[40] Cheng H-T, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H, et al. (2016) "Wide & Deep Learning for Recommender Systems." ArXiv:160607792 [Cs, Stat]. https://arxiv.org/pdf/1606.07792.pdf%29/
[41] Covington P, Adams J, Sargin E. (2016) "Deep Neural Networks for YouTube Recommendations." in: Proceedings of the 10th ACM Conference on Recommender Systems, Boston, Massachusetts, USA: Association for Computing Machinery; 2016, p. 191–8. https://doi.org/10.1145/2959100.2959190.
[42] Graves A, Mohamed A, Hinton G. (2013) "Speech Recognition with Deep Recurrent Neural Networks." in: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings 38. https://doi.org/10.1109/ICASSP.2013.6638947.
[43] Sutskever I, Vinyals O, Le QV. (2014) "Sequence to Sequence Learning with Neural Networks." in: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, (eds), *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., p. 3104–12.
[44] Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. (2016) "Concrete Problems in AI Safety." ArXiv:160606565 [Cs]. https://arxiv.org/pdf/1606.06565.pdf
[45] Kendall A, Gal Y. (2017) "What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?" in: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, et al., (eds) *Advances in Neural Information Processing Systems 30,* Curran Associates, Inc., p. 5574–5584.
[46] Mackay DJC. (1992) "Bayesian methods for adaptive models." Dissertation (Ph.D.), California Institute of Technology. doi:10.7907/H3A1-WM07.
[47] Graves A. (2011) "Practical Variational Inference for Neural Networks." in: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ (eds). *Advances in Neural Information Processing Systems 24,* Curran Associates, Inc., p. 2348–2356.
[48] Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. (2015) "Weight uncertainty in neural networks." in:Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, Lille, France: JMLR.org, p. 1613–1622.
[49] Gal Y, Ghahramani Z. (2016) "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning." International Conference on Machine Learning, p. 1050–1059. https://arxiv.org/pdf/1506.02142.pdf
[50] Kingma DP, Salimans T, Welling M. (2015) "Variational Dropout and the Local Reparameterization Trick." in: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, (eds), *Advances in Neural Information Processing Systems 28,* Curran Associates, Inc., p. 2575–2583.
[51] Hernández-Lobato JM, Adams RP. (2015) "Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks." in: Proceedings of the 32nd International Conference on Machine Learning, PMLR 37:1861-1869.
[52] Welling M, Teh Y. (2011) "Bayesian Learning via Stochastic Gradient Langevin Dynamics." in: ICML , p. 681-688. Omnipress.
[53] Osband I, Blundell C, Pritzel A, Van Roy B. (2016) "Deep Exploration via Bootstrapped DQN." in: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R, (eds), *Advances in Neural Information Processing Systems 29,* Curran Associates, Inc., p. 4026–4034.
[54] Lakshminarayanan B, Pritzel A, Blundell C. (2017) "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles." in: NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, p. 6405–6416.
[55] Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. (2014) "Machine Learning: The High Interest Credit Card of Technical Debt." in: SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop).
[56] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. (2014) "Intriguing properties of neural networks." in: International Conference on Learning Representations, arXiv:1312.6199
[57] Zügner D, Akbarnejad A, Günnemann S. (2018) "Adversarial Attacks on Neural Networks for Graph Data." in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, United Kingdom: Association for Computing Machinery; 2018, p. 2847–56. https://doi.org/10.1145/3219819.3220078.
[58] Lipton Z, Wang Y-X, Smola A. (2018) "Detecting and Correcting for Label Shift with Black Box Predictors." in: arXiv:1802.03916.
[59] Zhang K, Schölkopf B, Muandet K, Wang Z. (2013) "Domain Adaptation under Target and Conditional Shift." in: International Conference on Machine Learning, 2013, p. 819–27.
[60] Markou M, Singh S. (2003) "Novelty detection: a review—part 1: statistical approaches." *Signal Processing* 83:2481–2497.
[61] Truong C, Oudre L, Vayatis N. (2018) "A review of change point detection methods." in: arXiv:1801.00718v2 [cs.CE].
[62] David SB, Lu T, Luu T, Pal D. (2010) "Impossibility Theorems for Domain Adaptation." in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, p. 129–136.

[63] Gretton A, Smola A, Huang J, Schmittfull M, Borgwardt K, Schölkopf B. (2008) "Covariate Shift by Kernel Mean Matching." in: Quiñonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, (eds), *Dataset Shift in Machine Learning,* The MIT Press, p. 131–60.
[64] Chan Y, Ng H. (2005) "Word Sense Disambiguation with Distribution Estimation." in" Proceedings of IJCAI, p. 1010–1015.
[65] Saerens M, Latinne P, Decaestecker C. (2002) "Adjusting the Outputs of a Classifier to New a Priori Probabilities: A Simple Procedure." *Neural Computation* 14:21–41. https://doi.org/10.1162/089976602753284446.
[66] Schölkopf B, Janzing D, Peters J, Sgouritsa E, Zhang K, Mooij J. (2012) "On causal and anticausal learning." in: Proceedings of the 29th International Coference on International Conference on Machine Learning, Edinburgh, Scotland: Omnipress, p. 459–466.
[67] Hendrycks D, Gimpel K. (2016) "A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks" in: arXiv:1610.02136 [cs.NE].
[68] Liang S, Li Y, Srikant R. (2018) "Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks." in: arXiv:1706.02690 [cs.LG].
[69] Lee K, Lee H, Lee K, Shin J. (2018) "Training Confidence-calibrated Classifiers for Detecting Out-of-Distribution Samples." in: arXiv:1711.09325 [stat.ML].
[70] Shafaei A, Schmidt M, Little JJ. (2018) "Does Your Model Know the Digit 6 Is Not a Cat? A Less Biased Evaluation of "Outlier" Detectors." in: arXiv:1809.04729v2 [cs.LG].
[71] Taplin R, Hunt C. (2019) "The Population Accuracy Index: A New Measure of Population Stability for Model Monitoring." *Risks* 7:53.
[72] Karakoulas G. (2004) "Empirical Validation of Retail Credit-Scoring Models." *The RMA Journal,* p. 56-60
[73] Siddiqi N. ed. (2015) "Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring." Wiley.
[74] Ramsey F, Schafer D. (2002) "The Statistical Sleuth: A Course in Methods of Data Analysis." 2nd edition, Daniel published by Duxbury Press Hardcover.
[75] Johnson RA, Wichern DW. (2007) "Applied multivariate statistical analysis." Upper Saddle River, N.J., Pearson Prentice Hall.
[76] Siddiqi N. (2005) "Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring." 1 edition. Hoboken, N.J, Wiley.