ISSN: 2994- 9459

Research Article

Current Research in Statistics & Mathematics

A Robust Approach to Uncertainty Quantification in Deep Learning

Pierpaolo Massoli*

Department of Statistical Process Design (DCME), Italian National Institute of Statistics (ISTAT), Italy

***Corresponding Author**
Pierpaolo Massoli, Department of Statistical Process Design (DCME), Italian National Institute of Statistics (ISTAT), Italy.

Submitted: 2024, Nov 25; **Accepted:** 2024, Dec 24; **Published:** 2025, Jan 02

Citation: Massoli, P. (2025). A Robust Approach to Uncertainty Quantification in Deep Learning. *Curr Res Stat Math, 4*(1), 01-10.

Abstract

This study proposes a novel approach for quantifying the uncertainty of a deep learning model by investigating the coverage as well as the adaptivity of its prediction intervals in a Conformal Prediction context. The model investigated is designed to impute the equivalent household income by taking both specific household group characteristics and relevant features of the main income gainer into account as it is known that there are well-known correlations in literature. The imputation of such variable is critical as outliers occur or the required information for computing it is not entirely available. Due to the relevance of income in socio-economic policy contexts, the reliability of its imputation constitutes a key aspect. The Conformalized Quantile Regression is adopted in order to evaluate the prediction intervals of the model by incorporating this approach into the same. In this study an improved assessment of the model uncertainty is achieved by separating the aleatoric component from the epistemic one. For this purpose, an appropriate selection of training data is proposed. This non random selection introduces bias which may alter model estimates causing distortions which impair the uncertainty quantification approach. As a consequence, a correction of selection bias is integrated in the uncertainty evaluation process. A real-world case study is considered to demonstrate the potential of the proposed quantification approach.

Keywords: Uncertainty Quantification, Deep Learning, Income Imputation, Conformalized Quantile Regression, Heckman Selection

1. Introduction

In recent years the topic of income imputation has gained attention in economic research, especially when it comes to estimating the measurement of the equivalent household income [1]. This imputation is a challenging problem in general, due to the inherent uncertainty in socio economic data of which it is composed. This measure is designed for rendering the evaluation of household economic well-being more accurate by adjusting the total household income in accordance with the size of the household and relationships between its members, while keeping the economies of scale in mind as well as the different needs of the members [2]. Equivalent household income makes the comparison of different househod structures possible in that it detects various economic aspects which standard metrics might overlook [3]. Imputation techniques, such as regression-based or statistical matching methods, are not robust enough when it comes to managing incomplete or missing income data; therefore it is not insured that income distributions reflect realistic values in relation to known demographic groups [4]. Nevertheless, challenges persist, as the imputation process may introduce biases or inaccuracies that affect findings on income inequality, poverty rates, and related social policies [5]. As a result, accurate imputation remains critical for reliable socio-economic analyses and the development of equitable welfare policies. The inevitable existence of the aleatoric uncertainty pertaining to equivalent household income components may impair the point estimation process even in the case of a robust approach. An appropriate construction of input dataset for training the model is proposed [6,7] in order to enhance the epistemic aspect of the uncertainty under investigation. This specific construction may introduce a selection bias which results in distorted income estimates. As a consequence, in this study the Heckman correction is integrated into the aforementioned model [8]. The objective is to leverage deep learning methodologies in order to propose a robust procedure for evaluating the prediction intervals of a model for income imputation by means of the Conformalized Quantile Regression (CQR) from the Conformal Prediction framework [9-12]. A case study based on the statistical register by the name of ARCHIMEDE from ISTAT is considered in order to test for the potential of the uncertainty quantification technique proposed.

2. Basic Algorithms of a Deep Learning Model

Deep learning is a specific branch of machine learning which avails of artificial neural networks for complex data patterns modelling in large-scale datasets. Their general architecture provides multiple layers of connected neurons which manage complex problem solving in various fields of research.

The essential structures of the deep learning model under investigation are described below.

2.1. Denoising Autoencoder

The Denoising Autoencoder (DAE) is a type of artificial neural network designed for robust feature learning as well as removing noise from data. The DAE encodes both numeric vectors and vectors of categorical variables which has to be transformed into dummy variables so that the input data vectors become sequences of numbers included in the interval [1]. This model is trained to reconstruct the original input observation x from a corrupted version $\tilde{\mathbf{x}}$, of the same constituting an effective algorithm for data denoising and feature extraction. A pre-defined noise function $\eta \sim \mathcal{N}(0, 1)$ is applied to input data to disturb it, forcing the DAE to reconstruct x = *f* (x). as the data is being disturbed. A sketch of a simple one-layer architecture is shown in Fig. 1.

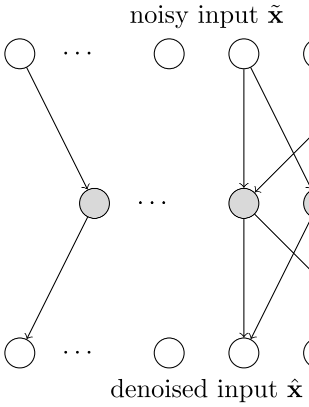
noisy input $\tilde{\mathbf{x}}$
denoised input $\hat{\mathbf{x}}$

Figure 1: Denoising Autoencoder

The encoder compresses each input into a smaller, more abstract representation, detecting essential features while filtering out irrelevant noise. This compressed, noisereduced rappresentation is subsequently reconstructed to obtain a noise-free version of the data, closely matching the uncorrupted original input. The training process is analogous to other popular neural network algorithms. In this case, the loss function is the *reconstruction error,* defined as follows:

$$\mathcal{L}(\tilde{\mathbf{x}}, \mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\mathbf{x}}_i)^2 \qquad (1)$$

where $\hat{\mathbf{x}}$ indicates the reconstructed input observation by the DAE subsequent to the noise application to the original observation x.

2.2. Multilayer Perceptron

The Multilayer Perceptron (MLP) is another type of artificial neural network rather similar to that described in the previous Section even though it is versatile in solving a great number of problems in data science. The basic structure is composed of multiple layers of neurons arranged in a feed-forward structure, which is highly effective for regression and classification problems. Its core architecture usually includes an input layer x, one or more hidden layers, and an output layer. Each layer consists of interconnected neurons that apply non-linear activation functions to model complex patterns in data. The MLP model *y* = *f* (x) is trained by adjusting the weights of connections between neurons based on the backpropagation of the error between predicted outputs

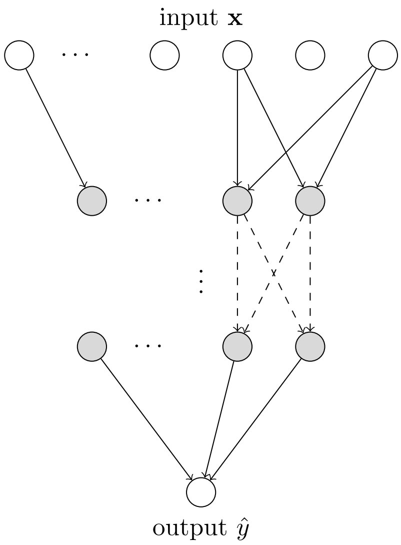

Figure 2: Multilayer Perceptron

$\hat{y}$ and actual outputs y. This error is usually optimized by implementing the stochastic gradient descent in order to minimize it. MLP training for regression problems usually requires the following loss function:

$$\mathcal{L}(\tilde{y}, y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2}$$

which is defined as the *Mean Square Error* (MSE) between the true value of the dependent variable y and its predicted value $\hat{y} = f(\mathrm{x})$ in a supervised learning perspective.

3. Setting up Training Data

The selection process of training data is one of the innovative aspects of this study. Complex data structures in current statistical registers require sophisticated approaches of analysis as is the case when detecting similar data structures in large scale datasets. To be more specific, different groups of individuals related one to another for kinship and utility reasons, *i. e.* household structures, deeply influence income dynamics. The equivalent household income of the OECD scale should take *economies of scale* into account without any disturbance of its predictors caused by aleatoric uncertainty present within them . As a result, aleatoric uncertainty in the data affects the estimation of the equivalent household income by altering the economies of scale and therefore hindering a reliable evaluation of the prediction intervals. As a consequence, an appropriate subset of the input dataset X is selected in order to facilitate this evaluation process. This subset is composed of pairwise similar households which are detected by means of the algorithm based on a Locality Sensitive Hashing (LSH) approach [[?], [?]]. Deep learning model training is accomplished by processing this specific subset so that the Conformalized Quantile Regression which is integrated in the model estimates the prediction intervals in two different scenarios as is described in this Section. Households pairs belonging to this subset share the same size as well as the same attributes of their members. Different pairs involve different sizes and different attributes of their members. The detection algorithm transforms all the households belonging to the input dataset into graphs and subsequently into sequences of integers which are elaborated by the LSH approach in order to detect the aforementioned pairs. These pairs are split into two datasets: A which contains the first household of every pair while B contains the second households of the same. The model in question $y = f(x)$ maps the changes in the equivalent household income y with respect to the variables x = $\{x_1, x_2, \ldots, x_n\}$, a set of features of the household *breadwinner* as well as specific household group characteristics subject to the pairwise similarity constraint. Data from both A and B sets is exchangeable and similarly distributed. The idea behind this study is to perform two independent uncertainty evaluation procedures by training the model with the same data (similar households) so that the contribution of the random component of the data is attenuated; this improves the accuracy of the uncertainty evaluation as well as increasing the robustness of the same by providing insights into model bias. In order to separate the aleatoric component of the uncertainty from the epistemic one, the DAE is trained on the entire dataset A for the subsequent training of the MLP by using a random 80% partition of the *encoded* dataset B and the remaining data for evaluating the model. This procedure is repeated by swapping A and B in order to compare the results for investigating epistemic uncertainty only.

4. Selection Bias Correction

This procedure known as *Heckman model* is a statistical approach designed to address selection bias, which arises when the dataset selected for model training is not a random sample in that it is subject to specific selection criteria yielding model estimates which may be biased. The Heckman model is implemented in two stages. The first stage consists of a *probit* model for estimating the probability of an observation being included into the sample; essential for understanding the selection process. This stage results in the *Inverse Mills Ratio* (IMR), a measure of the deviation of the selection process from being random. The IMR is defined as follows:

$$\mathrm{IMR} = \frac{\phi(\mathbf{x}'\beta)}{\Phi(\mathbf{x}'\beta)} \tag{3}$$

where $\phi(\bullet)$ denotes the probability density function of a standard normal distribution, $\Phi(\bullet)$ rappresents its cumulative distribution function, x is the vector of independent variables and β is the vector of coefficients estimated by the probit model. The IMR corrects the bias introduced by the non random selection process, an adjustment factor of the MLP regression model. After this calculation, in the next (second) stage the IMR is added to the training dataset as an additional feature in order to allow the neural network to improve the learning process and correcting the distortions caused by selection bias.

5. Uncertainty Quantification of the Model

The uncertainty quantification in model predictions is a key aspect of model fitting in machine learning. Conformal Prediction is a statistical framework for evaluating prediction intervals which are guaranteed to cover the true value with a pre-defined probability. The *Conformalized Quantile Regression* (CQR), is a method within this framework which combines the quantile regression and conformal prediction to compute valid adaptive prediction intervals. CQR is distribution-free as well as modelindependent, requiring no assumptions about the underlying data distribution. Classic regression models are used to predict point-estimates around an average value of the data while quantile regression models are used to estimate different quantiles, providing a range of estimates. The conformal framework adjusts the aforementioned estimates to ensure the specified coverage probability of the prediction intervals of the model. The CQR approach provides that given a dataset of N observations $\{(X = x_i, Y = y_i)\}$ $(i = 1, 2, \ldots, N)$ with features $X \in \mathbb{R}^d$ and response $Y \in \mathbb{R}$, CQR constructs a prediction interval

$C(X) = [L(X), U(X)]$ so that:

$$\Pr(Y \in \mathcal{C}(X)) \geq 1 - \alpha, \quad (4)$$

where $\alpha \in (0, 1)$ is the significance level. Prediction intervals are estimated by training a lower quantile regression model $\tau_{\alpha/2}$ and an upper quantile regression model $\tau_{1-\alpha/2}$ separately. Subsequent to the training of the models the quantiles $\hat{\tau}_{\alpha/2}$ and $\hat{\tau}_{1-\alpha/2}$ are estimated by using a *calibration* dataset $\{(X_{cal}, Y_{cal})\}$ containing n observations which are not used for training the model. By defining the score of each observation belonging to this dataset as follows

$$s(x, y) = \max[\hat{\tau}_{\alpha/2}(x) - y, y - \hat{\tau}_{1-\alpha/2}(x)] \quad (5)$$

the quantile $\hat{q} = \lceil (n + 1)\ (1 - \alpha) \rceil / n$ is computed in order to estimate the prediction *interval pertaining* each observation of $X_{test} = x$ belonging to the test dataset

$$\mathcal{C}(x) = [\tilde{\tau}_{\alpha/2}(x) - \hat{q}, \tilde{\tau}_{1-\alpha/2}(x) + \hat{q}] \quad (6)$$

where the quantiles $\tilde{\tau}$ are estimeted by using the true value of the test dataset $Y_{test} = y$. Smaller intervals correspond to simpler cases while larger intervals reveal more complicated cases. CQR approach can be adopted to evaluate the prediction intervals of any machine learning model by incorporating in it the *pinball loss* function:

$$L_\tau(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^{N} \max\left[\tau \cdot (y_i - \hat{y}_i),\ (\tau - 1) \cdot (y_i - \hat{y}_i)\right] \quad (7)$$

where y is the true value and $\hat{y}$ is the predicted one.

6. Proposed Approach: A Case Study

The proposed evaluation procedure for quantifying the uncertainty in a deep learning model for imputation purposes is presented in this Section and is applied to a real-world case study in order to evaluate its efficiency. Input data is from the ISTAT\ statistical register by the name of *ARCHIMEDE* regarding the resident population in Italy. The information gathered into the register is stored in socio-economic variables such as demographic variables, working activity, attained level of education, and income. Each record indicates an individual. People from the same household aregrouped by the same household identification number. The model being evaluated estimates the relationships between equivalent household income and relevant features of the household breadwinner in combination with specific household group variables pertaining to a specific region of the Italian territory only.

6.1. Input Data of the Model

An initial set X of 253286 households was selected by picking households with $n \geq 3$ members only. As it is reported in Sec. 3, these households are mathematically described as being fully connected undirected weighted graphs. The number of vertices is variable, so that there are graphs of different dimensions in the dataset. Every vertex is related to a sequence of K categorical variables (nodal attributes) $a^{(i)} = \{a^{(i)}_1, a^{(i)}_2, \ldots, a^{(i)}_K\}$ called the *profile* of the vertex v_i $(i = 1, 2, \ldots, n)$. The list of attributes in this case study is reported in Tab. 1: Subsequent to the application of the algorithm reported in Sec. 3, a dataset of $N = 254227$ pairs of *equal* households was selected from the X dataset. Selected pairs comprise 193803 distinct households of different sizes as well as different values of the attributes of each member. Pairwise similar households share the same size as well as the same attributes of their members so that the similarity value of every pair is equal to 1. The resulting input dataset is split into two partitions: dataset A contains the first household of every pair, while dataset B contains the second household of the same. These datasets are used to constitute two independent model training processes based on different but similarly distributed data, i.e., $A \longrightarrow B$ and $A \longleftarrow B$.

	Variable	Description	Number of classes
1	Gender	Gender of the household member	2
2	AgeClass	Age of the household member (in classes)	4
3	Citizn	Citizenship of the household member	2
4	EduLevel	Level of education of the household member	4
5	MainSourceIncome	Main source of income of the household member	7

Table 1: Attributes of the Household Members in the Input Dataset

6.2. Correction of the Selection Bias

The dataset $(A \cup B) \subset X$ is not selected at random. As a consequence, the deep learning model may be affected by bias. This selection bias is treated by using the 2-stage Heckman procedure as reported in Sec. 4. The list of the variables used in the probit model of the selection stage $sel \sim \text{probit}(predictors)$ is reported in Tab. 2: The dependent variable sel equals 1 if the observation belongs to the subset $A \cup B$

	Variable	Description
1	Ncomp	Number of household members (household size)
2	NForeigners	Number of household members who are foreigners
3	NChildren	Number of household members of age ≤ 14
4	Ncomp gender2	Number of household female members
5	NEduLev1	Number of household members who attend compulsory school
6	NEduLev2	Number of household members who attend high school
7	NMainSourceIncome1	Number of household members who are employers
8	NMainSourceIncome2	Number of household members who are self-employed
9	NRetired	Number of household members who are retired

Table 2: Predictors of the Probit Model

and 0 otherwise. Subsequent to this model fitting, the IMR is added as a feature in the predictive model of the multilayer perceptron to adjust it for bias.

6.3. Coverage and Adaptivity of the Prediction Intervals

The deep learning model in this case study is composed by a DAE which is stacked with a MLP in order to estimate the equivalent household income. The DAE transforms the input vector of the categorical variables into a smaller vector made up of numerical variables, the input of the MLP. This encoded input vector is augmented by adding the IMR variable in order to correct the model for the selection bias. The predictors of the probit model in Tab.2 are necessary for estimeting the probability of households of being selected in the training dataset. Variables used as predictors of the deep learning model (DAE and MLP) are listed in Tab.3. The categorical variables are transformed into dummy variables before being encoded by the DAE. Subsequent to this data pre-processing, the MLP is trained by using cross-validation. The numerical variables reported in Tab.3 are also considered as being categorical as a result of a top-coding of the aforementioned1. The uncertainty quantification of this model is carried out by integrating the loss function described in Eq.7 in the MLP in order to evaluate the prediction intervals as reported in Sec.5 in accordance with Angelopoulos and Bates [11]. The proposed approach for quantifying the epistemic component of the uncertainty provides two analoous procedures: 1) $A \rightarrow B$ and 2) $A \leftarrow B$ as is described in the diagram in Fig.3. The results of the application of the $A \rightarrow B$ and $A \leftarrow B$ are reported in Tab.4 and Tab.5, where coverage and adaptivity are respectively compared in order to investigate uncertainty. The empirical coverage measures the percentage of cases in which the prediction intervals contain the true value of the dependent variable. This property is also evaluated asymptotically by using an efficient caching of the non-conformity scores calculated as prescribed in the literature. Adaptivity measures the property of the prediction intervals to adapt to the cases covered by the model. Small intervals reveal good prediction ability of the model, while larger intervals suggest greater uncertainty in the predictions. In order

	Variable	Description	Number of classes
1	BW.IncomeClass	Income of the breadwinner	5
2	Ncomp	Number of household members (n ≥ 3)	6
3	BW.MainSourceIncome	Main source of income of the breadwinner	7
4	NChildren	Number of household members of age ≤ 14	6
5	BW.MaritalStatus	Marital status of the breadwinner	5
6	BW.EduLevel	Education level of the breadwinner	4
7	BW.Gender	Gender of the breadwinner	2
8	BW.Retired	Is the breadwinner retired?	2
9	AllSameCtzn	Citizenship of the whole household	3
10	BW.AgeClass	Age of the breadwinner	4
11	NRetired	Number of household members who are retired	5

Table 3: Predictors of the Imputation model

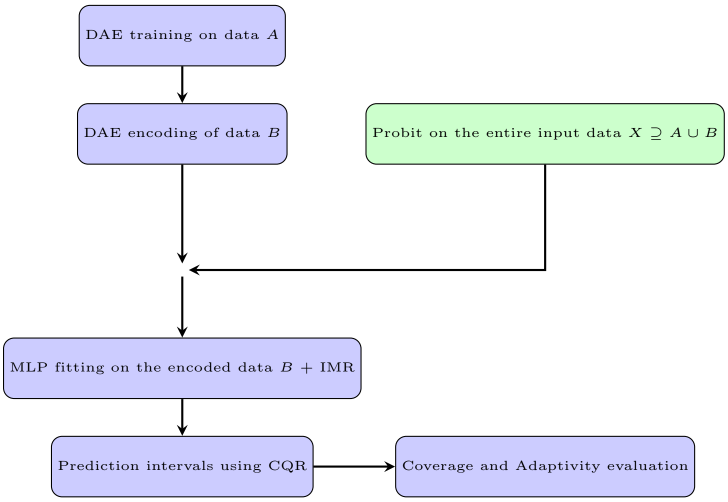
DAE training on data A
DAE encoding of data B
Probit on the entire input data $X \supseteq A \cup B$
MLP fitting on the encoded data B + IMR
Prediction intervals using CQR
Coverage and Adaptivity evaluation

Figure 3: Workflow diagram of the $A \rightarrow B$ procedure

Model	Empirical	Score caching
#1 ($A \rightarrow B$)	0.9545153	0.951681
#2 ($A \leftarrow B$)	0.9503557	0.9511803

Table 4: Coverage of the Prediction Intervals

Stats	#1 ($A \rightarrow B$)	#2 ($A \leftarrow B$)
Min	0.000001	0.0006018
1st Qu.	0.094100	0.1309214
Median	0.110540	0.1530824
Mean	0.142123	0.1946715
3rd Qu.	0.130294	0.1782837
Max	1.000000	1.0000000

Table 5: Adaptivity of the Prediction Intervals

to compare the evaluation procedures, the adaptivity was normalized by calculating as the ratio between the length of the prediction interval and the maximum value of all interval lengths as is reported in Tab. 5. In order to further investigate the uncertainty of the model, the boxplots of the distributions pertaining to the length of prediction intervals by the input categorical variables are reported in the following.

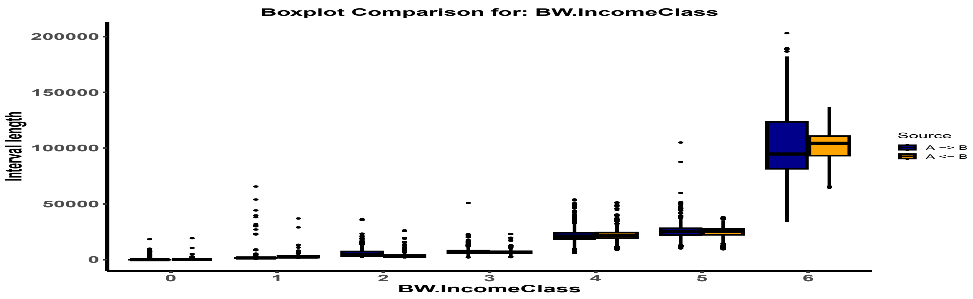

Figure 4: Prediction Interval Lengths: *BW. Income Class*

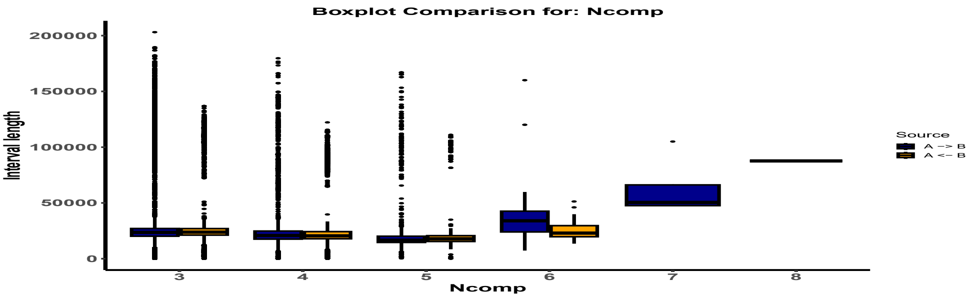

Figure 5: Prediction Interval Lengths: *Ncomp*

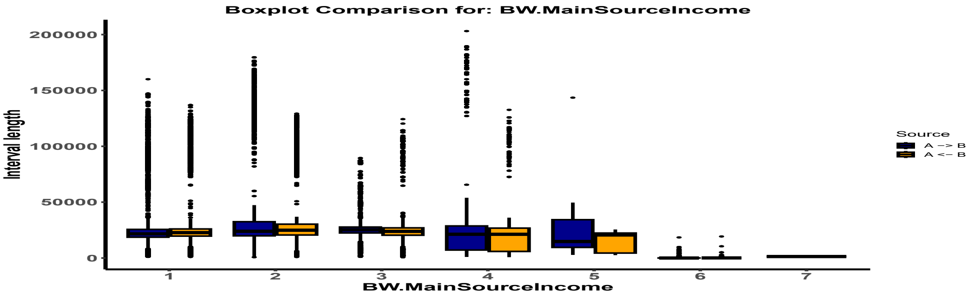

Figure 6: Prediction Interval Lengths: *BW.MainSourceIncome*

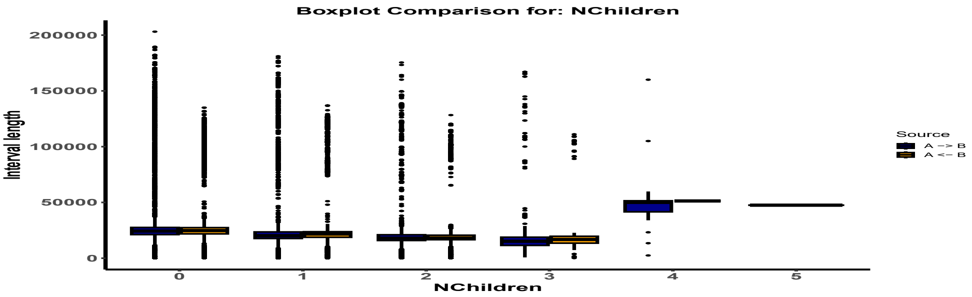

Figure 7: Prediction Interval Lengths: *NChildren*

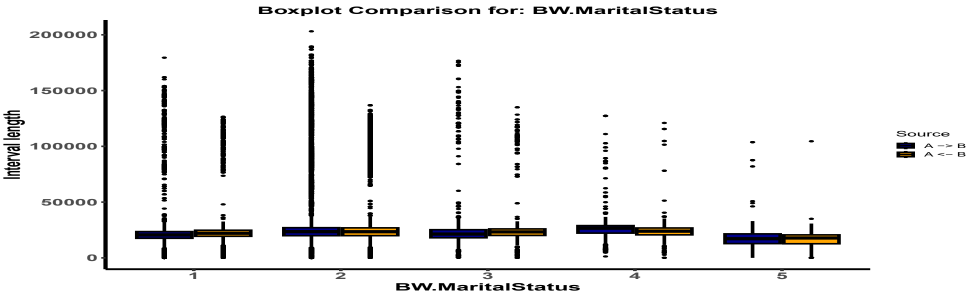

Figure 8: Prediction Interval Lengths: *BW.MaritalStatus*

7. Conclusions

This study proposes a general approach for evaluating the uncertainty of deep learning models as is the case of the Multilayer Perceptron stacked with the Denoising Autoencoder described in this paper. This evaluation approach can be extended to other machine learning models if they cater for the integration of robust techniques for quantifying uncertainty. The structure of the deep learning regression model being

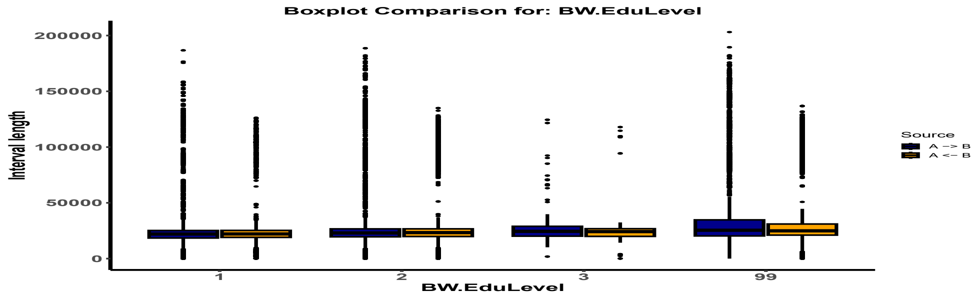

Figure 9: Prediction interval lengths: *BW.EduLevel*

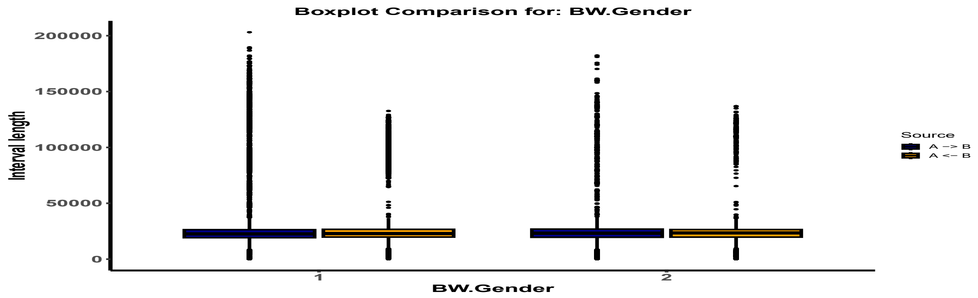

Figure 10: Prediction Interval Lengths: *BW.Gender*

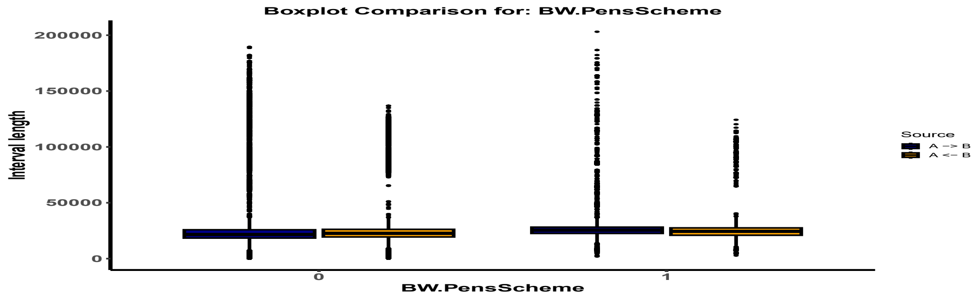

Figure 11: Prediction Interval Lengths: *BW.PensScheme*

considered combines a data pre-processing algorithm and a regression model. The proposed evaluation approach requires the selection of two similarly distributed datasets *A* ans *B*, focusing on the comparison of complementary evaluation procedures, i.e. *A* to- *B* and *B*-to-*A* which are carried out in order to analyze coverage and adaptivity of the model, measures for assessing the reliability of prediction intervals in rappresenting the underlying unknown data distribution. The integration of the Conformalized Quantile Regression approach in the deep learning model produces valid prediction

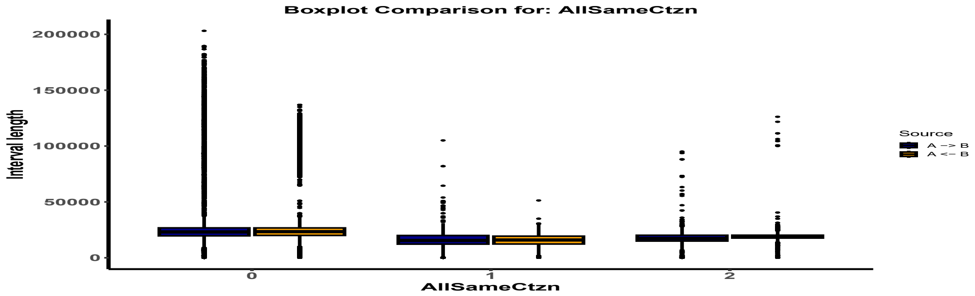

Figure 12: Prediction interval lengths: *AllSameCtzn*

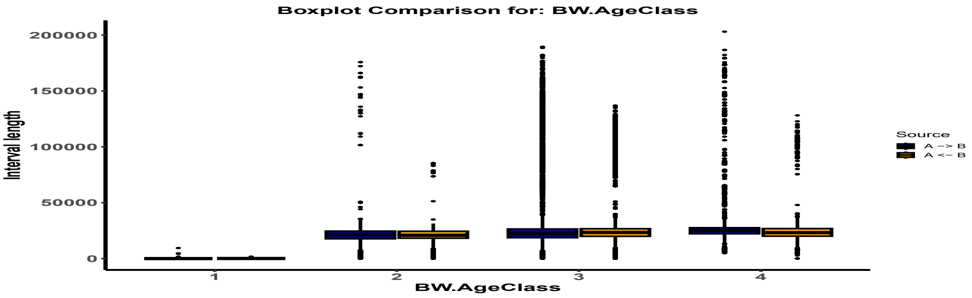

Figure 13: Prediction Interval Lengths: *BW.AgeClass*

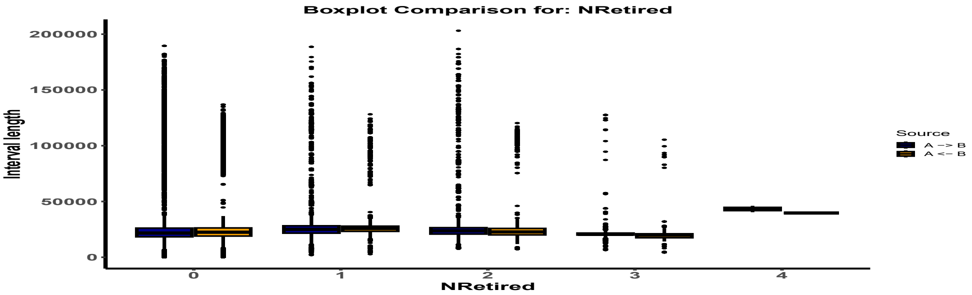

Figure 14: Prediction Interval Lengths: *NRetired*

intervals which provide an evaluation criterion of the reliability of the model. The two-sided evaluation approach proposed in this study provides consistent insights as a result of the enhancemwnt of the epistemic uncertainty coompared to the aleatoric one. This study does not explicitly separate the aleatoric and the epistemic components of the uncertainty. The aleatoric component is instead reduced by training the DAE on the selected datasets *A* and *B*. The subsequent data encoding in both the evaluation procedures equalize the aleatoric uncertainty as a result of the data similarity. The reduction of the aleatoric component contribution in the evaluation process of the uncertainty allow the prediction intervals to primarily reflect epistemic component while the equalization implies that differences between the evaluations are related to epistemic uncertainty. It is important to underline that, due to its probabilistic setup, the LSH algorithm assigns the pairs of similar households to *A* and *B* in a random order, minimizong potential biases during dataset creation. The detction algorithm resembles a split-plot design, where the primary variables act as main plot factors, and breadwinner specific variables are treated as subplot factors nested within similar households. In order to further address the bias introduced by the non random selection of training data, the Heckman correction is incorporated into the regression model. This is achieved by including the Inverse Mills Ratio as a feature to account for selection bias. The Heckman correction mitigates the concept drift resulting from a non rappresentative training sample, ensuring that the predictions generalize better to the target population. However, it assumes that the selection process is fully explained by observable variables and relies on the normality of errors, which may not hold universally. Including the IMR as a covariate also adds complexity to the interpretation of the model's coefficients. An in-depth investigation of the deep learning model under evaluation reports a higher occurrence of outliers on the left-hand side in the boxplots, corresponding to low values of the categorical variables under consideration. The probability of there being outliers is likely to increase as the number of observations which involve the aforementioned categories increases as well. The reduction of aleatoric uncertainty by using the DAE as a data pre-processing may not fully eliminate it for certain subgroups, resulting in higher frequency of outliers. This is particularly evident in households with similar structures and different breadwinner characteristics, where variations in prediction intervals reveal epistemic uncertainty pertaining to complex underrepresented groups. Differences in the lengths of prediction intervals between the categories indicate residual epistemic uncertainty steming from uneven data rappresentation. Categories pertaining to longer intervals reflect greater uncertainty, which arises from underrepresentationor complex data patterns. Outliers in these intervals suggest that the DAE does not correct input data anomalies or does not handle rare attributes, requiring a more complex pre-processing algorithm as is the case of architectures with a higher number of hidden layers or a higher number of DAEs arranged in a stack. The results of this study demonstrate the effectiveness of the comparison betwenn two opposite evaluation procedures based on noise-free similar data for smoothing the aleatoric component out from the uncertainty quantidication process while emphasize the epistemic component being reflected in the prediction intervals of the model. A further reduction of the residual uncertainty may be achieved by rcurring to data balancing in order to reduce the occurrence of rare observations or ensemble methods for combining both the evaluation proceduresin order to improvw the robustness of the prediction intervals [13-16].

References

1. Atkinson, A. B. (1995). Income Distribution in OECD Countries: Evidence from the Luxembourg Income Study.
2. Organisation for Economic Co-operation and Development. (2013). *OECD framework for statistics on the distribution of household income, consumption and wealth*. OECD

Publishing.
3. Jenkins, S. P., & Cowell, F. A. (1994). Modelling Household Income Distribution. LSE Research Online.
4. Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.
5. Creedy, J., & Kalb, G. (2005). Economics of Household Income Imputation: Concepts and Techniques. Springer.
6. Massoli, P. (2024). Detecting Similar Complex Data Structures in Large-Scale Datasets. *Curr Res Stat Math, 3*(2), 01-07.
7. Massoli, P. (2024). Assessing the Quality in the Detection of Similar Complex Data Structures in Large-Scale Datasets. *J Math Techniques Comput Math, 3*(8), 01-09.
8. Heckman, J. (1979). Sample selection bias as a specification error. Econometrica.
9. Koenker, R. (2005). Quantile regression. *Cambridge Univ Pr*.
10. Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research, 9*(3).
11. Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.
12. Romano, Y., Patterson, E., & Cand´es, E. J. (2019). Conformalized Quantile Regression. Advances in Neural Information Processing Systems (NeurIPS).
13. Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096-1103).
14. Zhou, S. K., & Chellappa, R. (2006). From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space. *IEEE transactions on pattern analysis and machine intelligence, 28*(6), 917-929.
15. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.
16. Yoon, J., Jordon, J., & Schaar, M. (2018, July). Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning* (pp. 5689-5698). PMLR.