ADVANCED UNCERTAINTY QUANTIFICATION AND NOVELTY DETECTION FOR RANDOM FOREST MODELS

Janne Merilinna

VTT Technical Research Centre of Finland,
Espoo, Finland

ABSTRACT

In practical applications, model accuracy alone is insufficient; quantifying model uncertainty is crucial, particularly in mission-critical scenarios involving life, money, or reputation. In this paper, we propose a novel method called MACAU (Model-based AleatoriC and epistemic uncertainty Uantification) and implement it in the LightGBM gradient-boosting framework. MACAU enables the quantification of both aleatoric and epistemic uncertainties in Random Forest (RF). Additionally, MACAU offers enhanced noveltydetection capabilities, particularly valuable for identifying out-of-distribution (OOD) samples. We compare MACAU with other RF- or gradient boosted trees-based methods, including RF-native between-variance, quantile regression, inductive conformal prediction, exogeneous model for uncertainty estimation using the Gaussian negative log-likelihood method, Natural Gradient Boosting, and CatBoost. Our evaluation is conducted on both synthetic and real-world regression cases. The results demonstrate the effectiveness of MACAU in quantifying model uncertainty, as measured by the Continuous Ranked Probability Score, as well as detecting OOD samples, as measured by the ROCAUC.

KEYWORDS

Machine Learning, Epistemic and Aleatoric Uncertainty, Out-of-Distribution Detection

1. INTRODUCTION

In mission-critical applications where money, lives, or reputation are at risk, relying solely on point-wise predictions of machine learning (ML) models may not be sufficient. While achieving high average correctness through appropriate cross-validation schemes is important, it can be misleading and provide a false sense of security when there is covariate drift in the out-of-sample data, which is commonly encountered in real-life applications [1]. Moreover, in domains such as risk modelling, relying solely on point-wise predictions without proper confidence intervals is inadequate, particularly when dealing with asymmetric risks.

Therefore, there is a need for advanced uncertainty quantification techniques that provide actionable insights beyond average correctness. Estimating uncertainty can be highly valuable in various domains and applications as expressing uncertainty in model predictions enables better decision-making, risk assessment, and model interpretability. Additionally, it facilitates the identification of instances where the model exhibits uncertainty and where further data collection or model refinement is necessary.

David C. Wyld et al. (Eds): SIPR, NCWC, BIBC, SOFEA, DSML, ARIA, NLP, CSEN -2023
pp. 253-269, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.131920

The *Law of Total Variance* decomposes the total variance into *explained* and *unexplained variance* components, with the former representing the variance of a random variable Y in itself, and the latter representing the variance between the expectations of Y. These components can be interpreted as *aleatoric uncertainty* and *epistemic uncertainty*, respectively, which are fundamental elements of uncertainty in machine learning [2]. In regression, these components are typically combined to form confidence intervals without explicitly disentangling them, even though there are cases where a deeper analysis of the model behaviour would benefit from such disentanglement.

Various methods exist to express and, in some cases, disentangle the components of total uncertainty, particularly in the context of Artificial Neural Networks (ANNs), as discussed in the comparison presented in [3]. While ANNs have achieved remarkable breakthroughs and state-of-the-art performance in non-structured data domains, tree-based models remain highly relevant in structured, tabular data settings [4]. Furthermore, as concluded in [3], ensembles, commonly employed in tree-based models, appear to be the most suitable approach for capturing uncertainty. Therefore, tree-based models have an advantage over ANNs in uncertainty modelling, as ensembles are an integral part of their design rather than an afterthought.

In this paper, we investigate how uncertainty modelling can be conducted in tree-based models, with a specific focus on Random Forest (RF) [5]. We explore different approaches for modelling uncertainty and introduce our own method called MACAU (Model-based AleatoriC and epistemic uncertainty q Uantification). MACAU is designed to not only capture uncertainty but also disentangle it and detect out-of-distribution (OOD) samples, which can lead to risky extrapolations.

MACAU leverages the state-of-the-art LightGBM gradient boosting framework [6], offering a powerful and efficient platform for uncertainty quantification in tree-based models. As tree-based models are not naturally capable for dealing with OOD data, we augment RF with a capability of linear extrapolation which not only facilitates detecting OOD samples but also expands the range of applications for this model. MACAU is capable for disentangling aleatoric and epistemic uncertainties, offering enhanced tools for understanding predictions in both classification and regression tasks, leading to more reliable and trustworthy models. Additionally, MACAU incorporates capabilities for detecting OOD samples when the standard uncertainty modelling approaches fall short.

In this paper, we quantitatively compare MACAU to its relevant peers in a set of regression tasks to evaluate its uncertainty modelling and OOD sample detection capabilities. We assess the quality of confidence intervals using the Continuous Ranked Probability Score (CRPS) as a metric, which captures the accuracy of predictions and the width of the confidence intervals in a single measure [7]. Additionally, we use the Receiver Operating Characteristic (ROC) Area Under Curve (AUC) as a metric to evaluate the OOD sample detection capabilities. While this comparison is not comprehensive, the initial results demonstrate the compelling performance of MACAU, surpassing its peers in both uncertainty quantification and OOD detection.

This paper is structured as follows. First, we discuss related research on enabling tree-based models to extrapolate and capture uncertainty in their predictions. Second, we introduce MACAU and its capabilities in uncertainty modelling and OOD detection. Third, we conduct quantitative comparisons between MACAU and its closest peers using regression experiments on two synthetic datasets and one real dataset. Finally, we provide a brief discussion of the findings and conclude the paper.

2. RELATED RESEARCH

Breiman's Random Forest algorithm [5] aggregates predictions by computing the simple average of tree predictions, which has proven difficult to beat [9][10]. However, this approach limits the ability of RF to extrapolate and model trends, leading to non-smooth step-wise predictions. To address this limitation, researchers have proposed techniques such as piece-wise linear regression [11], treed regression [12] and fitting multivariate regressors in tree leaves [13]. In [14], piece-wise linear models are utilized also in the tree-growing phase with additive feature selection. The piece-wise regression is not limited to linear models as discussed in [15] where comparisons between linear and kernel-based leaf models have shown the superiority of kernel-based models in certain cases. These advancements in modelling techniques offer improved flexibility in capturing complex relationships in data.

In real-world applications, it is often necessary to estimate prediction intervals to assess the uncertainty associated with point-wise predictions. Several methods have been proposed for estimating uncertainty in RF regression, such as bootstrapping [16], subsampling, jackknife methods [17], and U-statistics on subsamples [18]. Gradient-Boosted Decision Trees have also been used for uncertainty modelling through ensembles or virtual ensembles, which are more suitable for larger datasets [19]. These methods can help disentangle aleatoric and epistemic uncertainty components when considering the law of variance.

Quantile regression forests [20] replace the mean-squared error loss function with a *pinball loss* to estimate different quantiles of interest. Generalized Random Forests [21], Local-Linear Forests [22], and NGBoost [23] allow for estimation of arbitrary quantiles. However, these methods focus primarily on modelling aleatoric uncertainty and do not explicitly disentangle uncertainty components.

Conformal Prediction (CP) [24] provides an approach to estimate aleatoric uncertainty based on transductive inference. However, it requires repeated training of the model, which may be intractable in real-world applications. Inductive Conformal Prediction (ICP) [25][26] offers an alternative that requires fitting the model and calculating conformance scores only once. ICP yields uniform confidence intervals that are independent of the input variables. Conformalized Quantile Regression [27] introduces a method that allows for locally varying confidence intervals, overcoming the fixed interval limitation of CP. While these methods estimate aleatoric uncertainty, they do not explicitly model epistemic uncertainty.

In the context of artificial neural networks, estimating uncertainty can be approached through the Gaussian Negative Log-Likelihood (GNLL) loss function [28]. This method can be extended to estimate the uncertainty of any model by minimizing the GNLL loss of an exogenous model to maximize the likelihood that the predictions follow a Gaussian distribution, providing both point-wise predictions and their associated standard deviation, representing aleatoric uncertainty.

For RF classifiers, Shaker et al. [29] propose an add-on method that models aleatoric and epistemic uncertainty using plausibility theory and tree introspection. The reported uncertainties are dependent on the leaf nodes in which a sample falls. However, it should be noted that if the out-of-sample data does not adhere to the independent and identically distributed (IID) assumptions of the training data, the reported uncertainties may fail to capture the true uncertainty.

3. MACAU: MODEL-BASED ALEATORIC AND EPISTEMIC UNCERTAINTY QUANTIFICATION METHOD

MACAU extends the LightGBM tree-based classification and regression framework [6], leveraging its tree-growing procedure without interfering with it. While LightGBM has its own functionality for fitting piece-wise linear trees, MACAU takes inspiration from previous research on fitting piece-wise models into tree leaves [11][12][15][22]. The primary goal of MACAU is to introduce smoothness in the decision function and enable extrapolation capabilities, which are not inherently present in RF.

Unlike approaches that solely focus on extrapolation, MACAU places emphasis on modelling the uncertainty of predictions to enhance the trustworthiness of the model. It expands the concept of fitting piece-wise models beyond regression tasks and incorporates it into classification settings, enabling extrapolation in both regression and classification scenarios. MACAU also provides the flexibility to use traditional random forest trees without the piece-wise linear models, ensuring compatibility with existing RF implementations.

In the subsequent sections, the uncertainty modelling capabilities of MACAU for both regression and classification tasks are discussed. As MACAU employs linear models within its leaves, it has inherent limitations in accurately capturing the complete extent of uncertainty when the model is required to extrapolate beyond the training data. To overcome this limitation, MACAU introduces the concepts of *novelty*, *conditional novelty*, and *inference novelty*. These concepts serve to identify situations where the model encounters atypical samples or when predictions significantly deviate from expectations. By incorporating these concepts, MACAU enhances the comprehension of the model's behaviour in extrapolation scenarios. Implementation details of MACAU are made available in GitHub [30].

3.1. Uncertainty Modelling

The *Law of Total Variance* (Equation 1), sometimes referred to as the *Law of Total Uncertainty*, is a fundamental concept in statistics that decomposes the total variance (or uncertainty) of a random variable Y into two distinct components: the explained, or conditional, variance (aleatoric uncertainty) and the unexplained, or variance of conditional expectation (epistemic uncertainty). This law provides a framework for understanding and quantifying different sources of uncertainty. In Equation 1, $Var\ (Y)$ represents the total variance (total uncertainty) of the random variable Y,$E[Var\ (Y|X)]$ represents the expectation of conditional variance of Y given a random variable X (aleatoric uncertainty), and $Var(E[Y|X])$ represents the variance of conditional expectation of Y given X (epistemic uncertainty).

$$\mathrm{Var}(Y) = E[Var(Y|X)] + Var(E[Y|X]) \tag{1}$$

RF, as an ensemble method, naturally incorporate both components of Equation 1 when we have access to individual tree predictions and know the leaves to which the samples belong in each tree. Epistemic uncertainty can be estimated by considering the variance among the predictions of the trees for that sample whereas aleatoric uncertainty can be computed as the average variance within each leaf. MACAU takes advantage of these characteristics by focusing on the leaves of the trees and seamlessly integrating with the LightGBM gradient boosting framework.

It is worth noting that the identification phase of MACAU builds upon the already established RF model, which includes typical hyperparameter tuning and cross-validation. Once the RF model is

deemed satisfactory, MACAU proceeds to its specific identification phase, leveraging the existing structure and information provided by the RF.

Similar to piece-wise linear models [11][12][13], MACAU operates within the leaves of the trees. Specifically, it fits a leaf-specific model, referred to as a *leaf-model*, to each leaf in each tree of the forest. When fitting the leaf-models, only the features of X that were utilized by the corresponding tree to reach the leaf are selected, taking advantage of the automatic feature selection capability of the RF model. Furthermore, only the samples that have fallen into each leaf are utilized for fitting the leaf-models. This approach is applicable for both regression and classification tasks.

3.1.1. Uncertainty Modelling In Regression

In the context of regression, MACAU employs Automatic Relevance Determination Regression (ARDRegression) [31] as the leaf-model. ARD Regression is a Bayesian linear regressor that samples coefficients from Gaussian distributions instead of using fixed values as in ordinary linear regression. This allows ARDRegression to capture and express its aleatoric uncertainty, which is a fundamental aspect utilized by MACAU. ARDRegression also provides automatic feature selection capabilities which vanilla Bayesian linear regression does not provide therefore it is chosen instead of Bayesian regressor.

During the model identification stage, the ARDRegression model within each leaf of the RF is trained using the samples that fell into that specific leaf. Only the features involved in the splits leading to the particular leaf are considered. This approach ensures that the samples within each leaf share common characteristics based on their features and the prediction task. Essentially, the RF acts as a conditional clustering algorithm, dividing the global feature space into sub-spaces defined by the leaves. Within each leaf, it is assumed that a linear model can effectively capture the underlying relationships between the features and the target variable. This formulation enables the modelling of local relationships within each leaf while leveraging the overall structure provided by the RF.

During the inference phase, the leaf to which a sample belongs in each tree is identified, and the corresponding leaf-model is used to predict the target variable, denoted as $\hat{Y}_{leaf,i}$ where I stands for sample index, and estimate the corresponding uncertainty represented by $\hat{\sigma}_{leaf,i}$ for each sample. In contrast to traditional random forest approaches that rely solely on observed mean values obtained during the model identification phase, MACAU combines the leaf-model predictions$E[\hat{Y}_{leaf,i}]$and uncertainties from the tree-models using Equation 1 to calculate the final uncertainty for each sample, as shown in Equation 2.

$$\boldsymbol{Var}(\hat{\boldsymbol{Y}}_{\boldsymbol{i}}) = \boldsymbol{E}[\hat{\boldsymbol{\sigma}}_{\boldsymbol{leaf,i}}] + \boldsymbol{Var}(\hat{\boldsymbol{Y}}_{\boldsymbol{leaf,i}}) \tag{2}$$

Additionally, MACAU offers a basic version where linear models are not fitted within the leaves. In this case, the leaf predictions are the same as those in the classic RF but accompanied by uncertainty measures. The uncertainty $\hat{\sigma}_{leaf,i}$ in this case is calculated as the variance $Var(Y_{leaf})$ observed during the model identification phase. Therefore, it is leaf-dependent rather than sample-specific, as it is in the piece-wise linear tree version of MACAU. This flexibility allows users to choose the appropriate version of MACAU based on their specific requirements, considering factors such as computational complexity and model interpretability.

3.1.2. Uncertainty Modelling in Classification

Similar to regression, both aleatoric and epistemic uncertainties can be modelled for predicted class probabilities. In MACAU, a logistic regression is employed as the leaf-model for classification, replacing the ARDRegressor used in regression. However, logistic regression does not inherently provide a means for estimating the variance of predicted probabilities. To overcome this limitation, MACAU utilizes the Delta Method [32] instead of techniques like bootstrapping, which primarily captures epistemic uncertainty. The Delta Method enables the construction of confidence intervals for the probabilities, similar to regression.

During the model identification phase, a logistic regressor is fitted in each leaf following a similar approach used in regression. After fitting a logistic regressor in a leaf, a covariance matrix is calculated using the training samples available in the leaf along with the selected features. This covariance matrix is later utilized during inference. The following equation is used to calculate the covariance matrix Cov_{leaf}:

$$Cov_{leaf} = (X^T V \cdot X)^{-1} \tag{3}$$

Here, X represents the design matrix of shape $(n_{samples}, n_{features})$ where the both are specific to the leaf, and V is a vector of shape $n_{samples}$ obtained by taking the element-wise product of the predicted probabilities of the training set available in the leaf.

During inference, the leaf-models are used to predict the dependent variable and estimate the prediction uncertainty in each leaf, similar to the regression approach. However, since the leaf-model itself does not naturally provide uncertainty estimates, the Delta Method is utilized to quantify uncertainty.

First, the gradient $\nabla_{leaf,i}$ of the predicted probabilities is calculated in each leaf-model. The gradient for a sample in a leaf is computed as follows:

$$\nabla_{leaf,i} = p_i \cdot X_i \tag{4}$$

Here, p_i is the predicted probability for the i-th sample obtained by taking the element-wise product of the predicted probabilities, and X_i represents the feature vector for the i-th sample considering only the selected features chosen by the tree for this particular sample. After computing the gradient, the uncertainty $\sigma_{leaf,i}$ is calculated for each sample:

$$\sigma_{leaf,i} = \sqrt{\nabla_{leaf,i} \cdot Cov_{leaf} \cdot \nabla_{leaf,i}} \tag{5}$$

Here, $\nabla_{leaf,i}$ represents the i-th element of the gradient vector. As a result, each leaf-prediction now includes the predicted probability along with its associated uncertainty expressed as σ. Once again, the *Law of Total Variance* (Equation 1) combines all tree predictions together and forms the collective prediction with associated uncertainties.

3.2. Novelty Modelling

While tree-models are effective at capturing and quantifying prediction uncertainties within their local sub-spaces, they have limitations when it comes to extrapolation, which involves making predictions beyond the observed data range. It is important to exercise caution when extrapolating with linear tree-models because they assume a linear relationship between variables

even in unobserved regions. However, in reality, the relationship between variables can be complex and non-linear beyond the observed range. Solely relying on linear models for extrapolation can lead to unreliable and potentially inaccurate predictions since the model assumptions may not hold true in the extrapolated region.

3.2.1. Conditional Novelty

To identify potential extrapolation in tree-models, MACAU employs the Mahalanobis distance to quantify the *novelty* of samples within each leaf in RF. This measure, known as *conditional novelty*, is calculated using leaf-specific covariance estimators obtained during the model identification phase where the covariance estimators focus on selected features by the trees in the forest and samples within those leaves, similar to uncertainty modelling. The use of the Oracle Approximating Shrinkage Estimator [33] improves the stability of the estimation process, particularly when the number of samples within each leaf is limited.

To facilitate the interpretation of the Mahalanobis distance, MACAU applies tree-wise quantile normalization, which transforms the distance into a standard normal distribution for better comparability and comprehension. During the model identification phase, a scaler is fitted for each tree using the Mahalanobis distances computed from the training set. During inference, the tree-specific scaler is utilized to calculate the z-score for each sample, replacing the original Mahalanobis distance. The resulting *conditional novelty* is computed as the mean of all the normalized *conditional novelties* across the trees.

A high normalized Mahalanobis distance (or z-score) for a sample indicates a significant difference between its feature values and the observed training data, suggesting that the sample resides in a novel or unfamiliar region within the feature space. Thus, the normalized Mahalanobis distance serves as a quantitative metric for detecting samples that may extrapolate beyond the observed range of the training data. It provides valuable information, indicating the potential for the corresponding tree-model to generate unreliable predictions due to extrapolation.

3.2.2 Inference Novelty

The concept of *conditional novelty*, derived from the covariance matrix computed within each leaf using the relevant samples and selected features, is not directly associated with the actual predicted value. It is possible for the Mahalanobis distance to be high without significantly affecting the predicted value, or for it to have a limited impact due to variations in feature contributions. To address this, MACAU incorporates *inference novelty*, which assesses the deviation of the actual predicted value from the observations made during the tree-model identification phase.

The calculation of *inference novelty* involves fitting a leaf-wise quantile normalization scaler using the predicted values of each leaf-model during the MACAU identification phase. This scaler enables the computation of the z-score for each prediction during inference. The final *inference novelty* for a sample is obtained by averaging the z-scores across all trees.

The resulting *inference novelty* provides valuable information about the typicality of the predicted value within its context. A value of 0 indicates that the predicted value is typical, while high or low values indicate deviations from the typical pattern, with directional information incorporated. The absolute value of the z-score represents the degree of atypicality in the prediction, thereby indicating the presence of extrapolation. This allows for a quantitative assessment of the novelty in the predicted values and helps identify cases where extrapolation may occur.

3.2.2. Novelty

In MACAU, the *novelty* feature aims to detect novelty in the input, regardless of whether the features were selected by the trees. The rationale behind this is to identify changes in features that may not have been initially selected by the trees but may have become relevant to the prediction task due to changes in the input data.

To compute *novelty* and specifically detect OOD samples, MACAU takes a localized approach compared to traditional methods that use a global covariance matrix. MACAU focuses on a specific subspace aligned with the prediction task, similar to its uncertainty modelling and other novelty modelling techniques. This tailored approach ensures that MACAU's notion of *novelty* is well-suited to the specific prediction task, unlike conventional approaches that rely solely on the data without considering its intended application.

This refinement is desirable because the reliable computation of the Mahalanobis distance assumes an elliptical data distribution, which may be overly restrictive when considering all the data. However, within the localized context of the leaves in MACAU's trees, it is more reasonable to assume that the data within each leaf follows an elliptical distribution.

Similar to the *conditional novelty* feature, the Mahalanobis distances are normalized using tree-wise quantile normalization. This normalization process converts the distances into standardized z-scores, allowing for a consistent measure of *novelty* across different samples and leaves. The final *novelty* value is obtained by calculating the mean of the normalized Mahalanobis distances across all trees in the forest.

4. QUANTITATIVE COMPARISON BETWEEN UNCERTAINTY MODELLING METHODS

In the context of regression tasks, there are several established methods for estimating uncertainty in tree-based models. However, in the case of classification settings, quantifying uncertainty is not as straightforward and lacks the same level of research and application. As a result, this section primarily focuses on evaluating uncertainty estimation methods in regression settings, where these methods have been more extensively studied and widely utilized.

Uncertainty estimation in regression can be approached through bootstrapping and ensembles. This involves generating multiple bootstrap samples from the training data and training multiple models, whose predictions are then averaged. As a baseline, we use Light GBM with the RF booster, representing the bootstrapping approach. We also include CatBoost, a gradient boosted trees-based method capable of capturing aleatoric and epistemic uncertainty [19].

Direct modelling of uncertainty is another approach, exemplified by Quantile Regression [20] using Light GBM with the RF booster. NGBoost models the distribution of confidence intervals directly, allowing any model as the primary model. Here, decision trees are used to create a RF model. GNLL [28] captures aleatoric uncertainty with a primary model and a separate exogenous model to model prediction uncertainty. In this comparison, a linear exogenous model is employed.

While the aforementioned methods provide sample-specific uncertainty estimates, ICP [25][26] can estimate global uncertainty. ICP is a straightforward yet effective approach, implemented here using LightGBM with the RF booster as the primary model.

Additionally, we evaluate MACAU in two forms: MACAU (basic) without piece-wise linear models, and MACAU (linear) with tree models capable of extrapolation. Both versions are implemented on the LightGBM framework.

For the evaluation, sane default parameters are used for all models without hyperparameter tuning. Consistent hyperparameters are employed whenever possible.

4.1. Synthetic Id Regression

The effectiveness of uncertainty modelling methods is demonstrated using a synthetic 1D regression non-linear function with Gaussian heteroscedastic noise, as described in Equation 6. Additionally, the methods are examined for their behaviour in situations requiring extrapolation, and their performance in such scenarios is assessed.

$$Y = \frac{1}{2}X + sin(X) + N(0,1) + N(X, 0.1^2) \tag{6}$$

In this experiment, all models consist of 100 trees. The RF models employ fully grown trees, while CatBoost is restricted to three depth trees, as commonly done in gradient boosting. To encourage the development of diverse trees capable of predicting smooth functions, a sub-sampling rate of 0.1 is used. The models are fitted with 2k samples, where 2k samples are reserved for the in-distribution test set, and 6k samples are used for OOD samples.

Figure 1 illustrates the performance of different models in capturing the noise present in the function. The RF model, which solely captures epistemic uncertainty, fails to effectively capture the noise. In contrast, the other methods exhibit varying degrees of success in capturing the noise. Quantile Regression and NGBoost tend to overestimate the variance, while CatBoost tends to underestimate it. On average, ICP successfully captures the noise in the function, while GNLL demonstrates the ability to capture the linearly increasing noise. Both versions of MACAU effectively capture the heteroscedastic noise and accurately identify it as aleatoric uncertainty. The quantitative results are presented in Table 1 (see 1DSynth(CRPS) column).

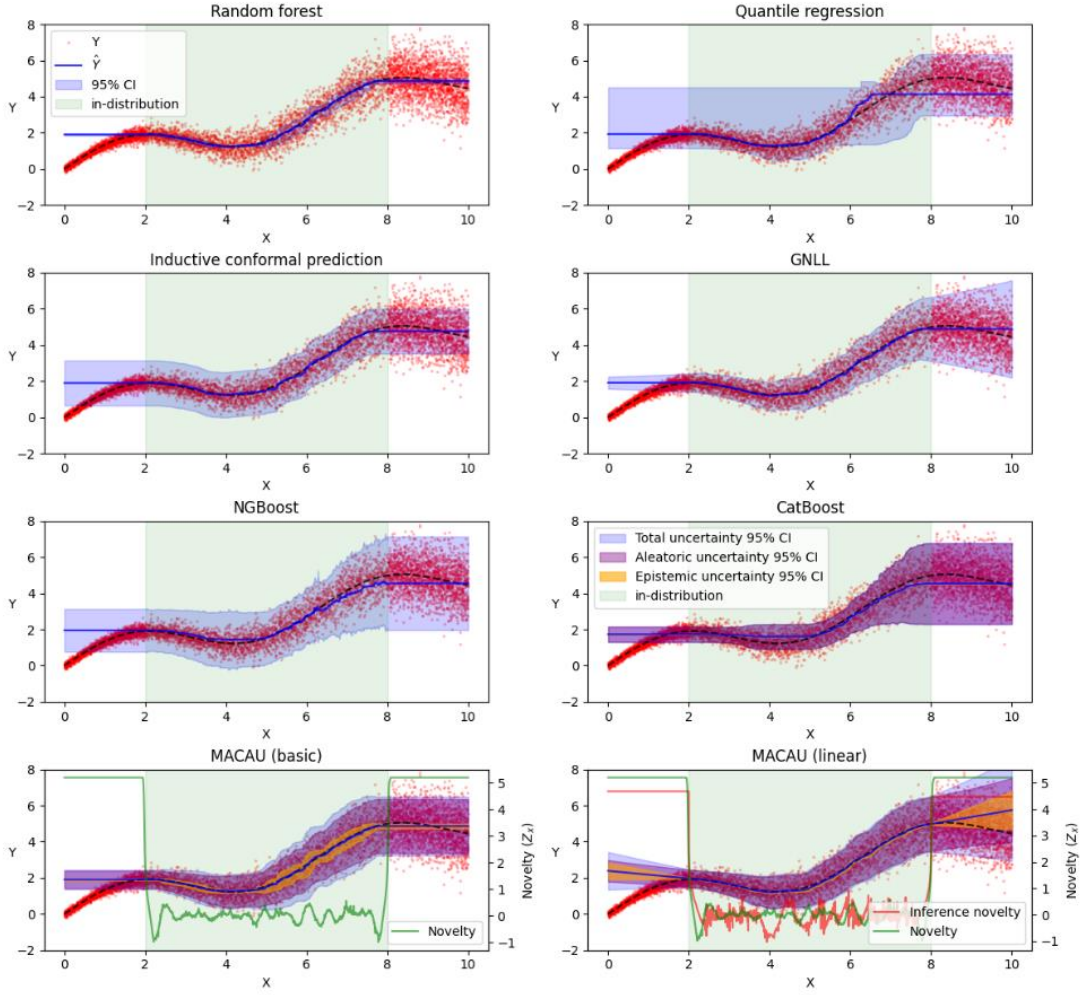

Figure 1. Comparison between uncertainty and OOD detection methods.

Detecting extrapolation presents a significant challenge for most models, but MACAU stands out with its specialized features designed precisely for this purpose. While CatBoost has capabilities for expressing epistemic uncertainty, it struggles to model trends and extrapolate due to its typical tree-based structure. As a result, CatBoost faces difficulties in effectively identifying OOD samples. Similarly, MACAU (basic), lacking extrapolation abilities like CatBoost, struggles to properly express its epistemic uncertainty in OOD scenarios. However, MACAU (linear), equipped with trend-modelling capabilities, demonstrates its prowess in handling extrapolation by exhibiting a linear increase in epistemic uncertainty when encountering OOD situations.

In addition to their uncertainty modelling capabilities, both MACAU models employ specialized techniques to detect extrapolation and identify abnormally high or low predictions. The concepts of *novelty* and *conditional novelty* are crucial in identifying covariate drift. Furthermore, the concept of *inference novelty* allows for the detection of deviations between the final predictions and what was observed during the model identification phase. Both MACAU models excel in capturing extrapolation by leveraging *novelty*. However, MACAU (linear) goes a step further by incorporating *inference novelty* to enhance its ability to identify extrapolation patterns. In contrast, MACAU (basic), being a tree-based model, lacks the necessary extrapolation capabilities and, therefore, cannot effectively utilize *inference novelty* for extrapolation detection. The quantitative results are as presented in Table 1 (see 1DSynth(AUC) column)

4.2. Multiple Regression

For multiple regression, two datasets are utilized. The first is a synthetic linear multivariate regression experiment with 10,000 samples and 10 features. Out of these features, five are relevant to the target variable, and the target variable itself is affected by noise. The second dataset is the California Housing Dataset, which contains eight input features and one target variable. In both cases, the datasets are divided into training and test sets using a 50/50 split.

All models in the evaluation use common hyperparameters. A sub-sampling rate of 0.8 is applied, with a minimum of 20 samples required in each leaf and a maximum of 31 leaves allowed. The random forest models are fully grown, while CatBoost is limited to a depth of three. In addition to the tree-based models, a linear regression model is included in the evaluation and serves as a baseline for comparison.

Table 1 presents the CRPS metrics of the models (see Synth(CRPS) column). In the synthetic dataset, the baseline model (inductive conformal linear prediction) performs the best in capturing the target variable. This result is not surprising, considering that the underlying relationship in the dataset is linear. Following the baseline model, MACAU (linear) ranks second, while the other models lag behind, sometimes by a significant margin. Interestingly, CatBoost struggles greatly in this dataset, contrary to expectations.

In the California Housing Dataset (see Cali(CRPS) column), MACAU (linear) emerges as the top performer in capturing the target variable. These results indicate that MACAU demonstrates fairly promising capabilities in modelling confidence intervals and capturing the underlying uncertainty in the predictions.

To evaluate the methods' ability to detect OOD samples, we use the same datasets. The OOD challenge is designed by dividing each feature, one at a time, into two groups based on feature values: the top 50% and the bottom 50%. Each group represents either in-sample or OOD data. For model training, we reserve 50% of the in-sample data, while the remaining 50% is used for evaluation alongside the OOD data. To ensure unbiased OOD detection, we repeat this process twice for each feature, switching between in-sample and OOD data. In total, we conduct experiments equal to twice the number of features in the dataset. The results are averaged to calculate the ROC AUC with 95% confidence intervals, as shown in Table 1 (see Synth(AUC) and Cali(AUC) columns). As a baseline for comparison, we utilize Isolation Forest [34], a powerful tree-based anomaly detection algorithm known for its effectiveness in detecting OOD samples.

The results indicate that MACAU and Isolation Forest are the only methods that consistently detect OOD samples in both the synthetic and California Housing dataset cases. MACAU's *novelty* capability performs exceptionally well in capturing OOD samples and surprisingly outperforms Isolation Forest in OOD detection. This may be because Isolation Forest considers the entire global feature space, while MACAU focuses on smaller, leaf-specific sub-spaces, which can make separating OOD samples easier.

Regarding MACAU's *conditional novelty* capability, it is evident that OOD detection is only possible when OOD is introduced to features relevant to the prediction task. This results in high variance in the OOD detection capabilities observed in these results. This behaviour is expected because *conditional novelty* is not aware of features that are not selected by the trees. Similarly, the *inference novelty* capability captures instances where extrapolation significantly influences the actual predictions. However, it is surprising that MACAU (linear) does not demonstrate effective OOD detection with its epistemic uncertainty. This contradicts the expectation that

epistemic uncertainty should indicate a lack of knowledge. In summary, MACAU's *novelty* emerges as the superior method for OOD detection compared to the other evaluated methods in these experiments.

Table 1. Confidence interval quality and OOD detection capabilities of the evaluated methods

Method	1DSynth(CRPS)	Synth(CRPS)	Cali(CRPS)	1DSynth(AUC)	Synth(AUC)	Cali(AUC)
CatBoost aleatoric uncertainty	0.32	857.6	0.46	0.55	0.5 [0.48;0.52]	0.5 [0.23;0.7]
CatBoost epistemic uncertainty	0.44	30.97	0.63	0.53	0.5 [0.44;0.54]	0.51 [0.09;0.77]
CatBoost total uncertainty	0.32	870.09	0.46	0.55	0.5 [0.49;0.51]	0.49 [0.36;0.7]
GNLL	**0.29**	30.28	0.43	0.5	0.5 [0.48;0.52]	0.37 [0.3;0.7]
Inductive conformal linear prediction	0.49	**5.99**	0.46	0.7	0.5 [0.44;0.55]	0.44 [0.37;0.73]
Inductive conformal prediction	0.3	30.34	0.44	0.66	0.5 [0.47;0.53]	0.47 [0.28;0.74]
Isolation forest				**1**	0.86 [0.77;0.94]	0.83 [0.72;0.98]
MACAU (basic) aleatoric uncertainty	**0.29**	30.37	0.43	0.44	0.5 [0.47;0.53]	0.49 [0.25;0.71]
MACAU (basic) conditional novelty				**1**	0.61 [0.49;1.0]	0.65 [0.4;0.95]
MACAU (basic) epistemic uncertainty	0.35	32.77	0.51	0.2	0.5 [0.45;0.54]	0.49 [0.26;0.74]
MACAU (basic) inference novelty				0.5	0.5 [0.48;0.52]	0.5 [0.28;0.72]
MACAU (basic) novelty				**1**	**1.0 [0.99;1.0]**	**0.93 [0.77;0.98]**
MACAU (basic) total uncertainty	**0.29**	32.87	0.43	0.32	0.5 [0.47;0.53]	0.49 [0.28;0.74]
MACAU (linear) aleatoric uncertainty	**0.29**	19.39	**0.4**	0.54	0.5 [0.34;0.67]	0.52 [0.27;0.81]
MACAU (linear) conditional novelty				**1**	0.61 [0.49;1.0]	0.65 [0.4;0.95]
MACAU (linear) epistemic uncertainty	0.39	20.89	0.46	0.95	0.5 [0.27;0.74]	0.57 [0.38;0.86]
MACAU (linear) inference novelty				**1**	0.6 [0.49;0.99]	0.61 [0.45;0.91]
MACAU (linear) novelty				**1**	**1.0 [0.99;1.0]**	**0.93 [0.77;0.98]**
MACAU (linear) total uncertainty	**0.29**	20.44	**0.4**	0.63	0.5 [0.31;0.7]	0.56 [0.36;0.83]
NGBoost	0.32	45.95	0.48	0.53	0.5 [0.46;0.54]	0.49 [0.21;0.76]
Quantile regression	0.31	37.02	0.45	0.28	0.5 [0.26;0.75]	0.48 [0.08;0.71]
Random forest	0.35	32.77	0.51	0.2	0.5 [0.5;0.5]	0.49 [0.5;0.5]

5. DISCUSSION

In this paper, we compared several methods for expressing uncertainty in tree-based models. While the evaluation datasets used were relatively easy to model, the results reveal the limitations of these methods. Expressing aleatoric uncertainty is not always straightforward. We quantitatively evaluated performance using the CRPS metric and found that certain methods struggle to accurately represent aleatoric uncertainty, whereas MACAU performs reasonably well compared to other evaluated methods.

Capturing epistemic uncertainty presents an even greater challenge. This is particularly true for detecting OOD samples. Reliable estimates of epistemic uncertainty are crucial for applications like active learning, where new samples are selectively collected to improve the model, and in domains where covariate drift is expected. In our comparison, MACAU is the only method capable of providing meaningful values for epistemic uncertainty, although it may not be sufficient for active learning or OOD detection. However, MACAU offers additional features such as *conditional novelty*, *inference novelty*, and *novelty* in particular to address these challenges, providing better tools for detecting risky predictions.

Although this study only scratches the surface of the potential uses for the additional information provided alongside the actual predictions, practitioners can leverage these outputs based on their specific requirements and use cases. Our hope is that the rich information provided by MACAU will contribute to the development of more trustworthy and reliable models for real-world applications.

6. CONCLUSIONS

In domains where the consequences of supervised machine learning models have significant impacts on matters of life, money, or reputation, relying solely on average correctness is insufficient. It is crucial to discern when to trust each prediction made by the model especially when real-world use cases often present challenges that defy the fundamental assumptions of supervised machine learning, such as data independence and identical distribution. These challenges can result in potentially misleading summary statistics and highlight the need for improved methods to comprehend predictions.

This paper introduces MACAU, a novel method that addresses the aforementioned challenges by leveraging tree introspection in random forest models. MACAU leverages the state-of-the-art LightGBM framework to provide valuable insights into the trustworthiness of predictions. By quantifying two types of uncertainty - aleatoric uncertainty and epistemic uncertainty - MACAU offers a deeper understanding of predictions.

Furthermore, MACAU tackles the inherent difficulties associated with detecting covariate drift in tree-based models by introducing additional features that can be used in conjunction with the uncertainty measures. The *conditional novelty* detects drift in the covariates that are relevant to the predictions, the *inference novelty* identifies significant discrepancies between the predicted values and the observed values in the model training data, helping to identify potential issues in the model's performance, and lastly, the *novelty* detects drifts in the data that may be relevant to the model but are not necessarily utilized by it, making it particularly useful for identifying OOD samples. By incorporating these features, MACAU equips decision-makers with additional information to evaluate the trustworthiness of predictions and make informed decisions.

To evaluate the performance of MACAU relative to other state-of-the-art methods designed for similar tasks, a comparison was conducted. The results demonstrate that expressing uncertainty in models is far from trivial. However, based on the evaluation presented in this paper, MACAU performs favourably against other methods when evaluated using the CRPS metric. Furthermore, in terms of OOD detection, MACAU surpasses all other evaluated methods, including a dedicated OOD detection method in terms of ROCAUC. These evaluation results, while not exhaustive, provide compelling evidence of the trustworthiness and reliability of MACAU's capabilities.

ACKNOWLEDGEMENTS

This work was labelled by ITEA3 and funded by Business Finland under grant agreement "ITEA-2019-18022-IVVES".

REFERENCES

[1] Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D., Dataset Shift in Machine Learning, The MIT Press, 2008, doi:10.7551/mitpress 9780262170055.001.0001.

[2] Hüllermeier, E., Waegeman, W., Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. Machine Learning 110, pp. 457-506, 2021, doi: 10.1007/s10994-021-05946-3.

[3] Valdenegro-Toro, M., Mori, D.S., A Deeper Look Into Aleatoric and Epistemic Uncertainty Disentanglement, IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 2022 pp. 1508-1516, doi: 10.1109/CVPRW56347.2022.00157

[4] Grinsztajn, L., Oyallon, E., Varoquaux, G., Why do tree-based models still outperform deep learning on typical tabular data?, 36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks, Nov 2022, New Orleans, United States.

[5] Breiman, L., Random forests, Machine Learning 45, 2001, pp. 5-32, doi:10. 1023/A:1010933404324.

[6] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., Lightgbm: A highly efficient gradient boosting decision tree, In Proceedings of the 31st International Conference on Neural Information Processing Systems, December 2017, pp. 3149-3157.

[7] Matheson, J.E., Winkler, R.L., Scoring rules for continuous probability distributions, Management Science, Vol. 22, No. 10, 1976, pp. 1087-1096.

[8] Gneiting, T., Raftery, A.E., Strictly proper scoring rules, prediction, and estimation, Journal of the American Statistical Association 102:477, pp. 359-378 doi:10.1198/016214506000001437.

[9] Busch, F., Kulessa, M., Loza Mencía, E., Blockeel, H., Combining predictions under uncertainty: The case of random decision trees, Arxiv abs/2208.07403, 2022.

[10] Bostrom, H., Estimating class probabilities in random forests, Sixth International Conference on Machine Learning and Applications (ICMLA 2007), Cincinnati, OH, USA, 2007, pp. 211-216, doi: 10.1109/ICMLA.2007.64.

[11] Breiman, L., and W. S. Meisel, General Estimates of the Intrinsic Variability of Data in Nonlinear Regression Models, Journal of the American Statistical Association, Vol. 71, No. 354, 1976, pp. 301-307, doi:10.2307/2285301.

[12] Alexander, W.P., Grimshaw, S.D., Treed regression, Journal of Computational and Graphical Statistics,Vol. 5, No. 2, June, 1996.

[13] Quinlan, J., Learning with continuous classes, In Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, Hobart 16-18 November 1992, pp. 343–348.

[14] Shi, Y., Li, J., Li, Z., Gradient boosting with piece-wise linear regression trees, In Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI Press, August 2019, pp. 3432–3438.

[15] Torgo, L., Functional models for regression tree leaves, In Proceedings of the Fourteenth International Conference on Machine Learning (ICML 1997), Nashville, Tennessee, USA, July 8-12, 1997.

[16] Coulston, J.W., Blinn, C.E., Thomas, V.A., Wynne, R.H., Approximating prediction uncertainty for random forest regression models, Photogrammetric Engineering & Remote Sensing, Volume 82, Issue 3, 2016, ISSN 0099-1112, doi:10.14358/PERS.82.3.189.

[17] Sexton, J., Laake, P., 2009. Standard errors for bagged and random forest estimators, Computational Statistics & Data Analysis, Volume 53, Issue 3, 2009, pp. 801-811, ISSN 0167-9473, doi: 10.1016/j.csda.2008.08.007.

[18] Mentch, L., Hooker, G., Quantifying uncertainty in random forests via confidence intervals and hypothesis tests, Journal of Machine Learning Research, Vol. 17, 2016, pp. 1-41.

[19] Malinin, A., Prokhorenkova, L., Ustimenko, A., Uncertainty in gradient boosting via ensembles, 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021, Open-Review.net.

[20] Meinshausen, N., Quantile regression forests, Journal of Machine Learning Research Vol. 7, 2006, pp. 983-999.

[21] Athey, S., Tibshirani, J., Wager, S., Generalized random forests, The Annals of Statistics, Vol. 47, No. 2, (2), April 2019, doi:10.1214/18-AOS1709.

[22] Friedberg, R., Tibshirani, J., Athey, S., Wager, S., Local linear forests, Journal of Computational and Graphical Statistics 30:2, pp. 503–517. doi:10. 1080/10618600.2020.1831930.

[23] Duan, T., Avati, A., Ding, D.Y., Thai, K.K., Basu, S., Ng, A., Schuler, A.,. Ngboost: Natural gradient boosting for probabilistic prediction, In Proceedings of the 37th International Conference on Machine Learning (ICML'20), Vol. 119. JMLR.org, Article 252.

[24] Gammerman, A., Vapnik, V., Vovk, V., Learning by transduction, In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI'98), July 1998, pp. 148-155.
[25] Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A., Inductive confidence machines for regression, Machine Learning: ECML 2002, Lecture Notes in Computer Science, Vol 2430, Springer, Berlin, Heidelberg, doi:10.1007/3-540-36755-1_29.
[26] Saunders, C., Gammerman, A., Vovk, V., Computationally efficient transductive machines, In: Arimura, H., Jain, S., Sharma, A. (eds) Algorithmic Learning Theory (ALT 2000), Lecture Notes in Computer Science, Vol 1968, Springer, Berlin, Heidelberg. Doi:10.1007/3-540-40992-0_25.
[27] Romano, Y., Patterson, E., Candes, E., 2019. Conformalized quantile regression, In Proceedings of the 33rd International Conference on Neural Information Processing Systems, December 2019, Article No.: 318, pp. 3543-3553.
[28] Nix, D., Weigend, A., Estimating the mean and variance of the target probability distribution, In Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94), Vol. 1., pp. 55–60, doi:10.1109/ICNN.1994.374138.
[29] Shaker, M.H., Hüllermeier, E., Aleatoric and Epistemic Uncertainty with Random Forests, In: Berthold, M., Feelders, A., Krempl, G. (eds) Advances in Intelligent Data Analysis XVIII, IDA 2020, Lecture Notes in Computer Science, vol 12080, Springer, doi:10.1007/978-3-030-44584-3_35.
[30] Merilinna, J., MACAU, GitHub, URL: https://github.com/jmerilinna/macau [visited at 7.9.2023].
[31] Tipping, M.E., Sparse Bayesian Learning and the Relevance Vector Machine, Journal of Machine Learning Research 1, 2001, pp. 211–244.
[32] Hoef, J.M.V., Who invented the delta method? The American Statistician Vol. 66, No. 2, 2012, pp.124-127.
[33] Chen, Y., Wiesel, A., Eldar, Y.C., Hero, A.O., Shrinkage algorithms for MMSE covariance estimation, IEEE Transactions on Signal Processing, Volume 58, Issue 10, October 2010, doi:10.1109/TSP.2010.2053029.
[34] Liu, F.T., Ting, K.M., Zhou, Z.H., Isolation forest, Eighth IEEE International Conference on Data Mining, 2008, pp. 413-422. doi:10.1109/ICDM.2008.17.